UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

EXPLORING DRUG-USE PROGRESSION THROUGH STABILITY ENHANCED CLUSTERING

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY

By

MATTHEW J. BEATTIE
Norman, Oklahoma
2022

EXPLORING DRUG-USE PROGRESSION THROUGH
STABILITY ENHANCED CLUSTERING

A DISSERTATION APPROVED FOR THE
GALLOGLY COLLEGE OF ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Charles Nicholson, Chair

Dr. Talayeh Razzaghi

Dr. Hairong Song

Dr. Randa Shehab

Acknowledgments

I would like to thank the members of my committee for their support in the development of this dissertation. My mom and dad, Janice and Phil, have made it clear that you're never too old to make your parents proud. I would like to thank my wife Mary-Pat, my daughter Alexandra, and my son Adam for their patience over the years as I can be a bit irritable when writing. Finally, and most importantly, this work, and indeed my entire body of research, is dedicated to the memory of my son, Jack. Maybe it will make a difference, Jack. I hope so. If not, I'll get it right next time. I will never stop trying.

Abstract

Background and aims: Drug use initiation sequences have been the subject of much research, and theories such as the Gateway Hypothesis have been created to explain patterns of progression from common to dangerous drugs. This study uses adult respondent observations from four years of the National Survey on Drug Use and Health (NSDUH) to uncover complex patterns associated with the age of first use (AFU) of drugs that are difficult to discern using a priori hypotheses. From these patterns, a classification study is conducted to determine what AFUs for quasi-legal drugs are most associated with subsequent illicit drug use. Associations of demographic features with the AFU patterns are explored as well. **Methods**: A modification to K-means clustering (KMC) is developed to improve the partition stability of survey data. This method, stability enhanced K-means clustering (SEKMC), builds partitions that are based upon relationships among observations that persist across multiple partitions of bootstrap samples of the NSDUH data. The computational complexity of the method is overcome through cluster computing and the development of an algorithm to calculate completely connected components in a graph in O(V) time. Classification of illicit drug use as a function of quasi-legal drug AFUs is conducted using decision trees and logistic regression. Descriptive techniques, including a χ^2 analysis are applied to the partitioned data to relate demographic features to AFU patterns. **Findings**: A partition of the data is extracted that contains 13 clusters, including ones of note – early age marijuana initiation, a set of clusters whose commonality is based upon illicit drug use, and one that indicates a link between prescription drug abuse and marijuana. Both the decision tree and logistic regression analyses demonstrate a strong association between early AFU of marijuana and subsequent illicit drug use. Non-Hispanic Asians are more likely than any other ethnicity to belong to a no-use cluster, and respondents with less than high school education are paradoxically more likely to belong to both the no-use and polyabuse clusters.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Substance abuse has been a problem in the United States since the 19th century and recent trends in mortality have increased its negative societal impact. Research into usage patterns has been looked to as a way to uncover information that can guide abuse mitigation policies and reduce substance-related deaths. The sequence by which subjects initiate the use of drugs[1] is of particular interest. Studying sequences involves difficulties associated with access to data and data complexity. Data access can be satisfied with the use of large, multiyear cross-sectional studies. However, researchers have constrained their analysis of these studies through the use of traditional statistical approaches, particularly hypothesis construction and testing. This study addresses the complexity of a massive drug-use dataset through cluster analysis, an unsupervised learning method, in an effort to allow the data to 'speak for itself.' Because the dataset is a survey-based representation of the US population, blind application of machine learning (ML)

[1] In this study the words 'drug' and 'substance' refer not only to the FDA Drug Schedule, but also alcohol and tobacco.

risks overfitting clusters to the survey. This study addresses this flaw through the presentation of a modification to KMC that produces clusters that are stable across multiple versions of the dataset. This method is used to explore the Gateway Hypothesis, a long standing theory of substance use initiation. It is also used to explore demographic variations among substance initiation clusters. This study's discussion flow is presented in Figure 1.1.

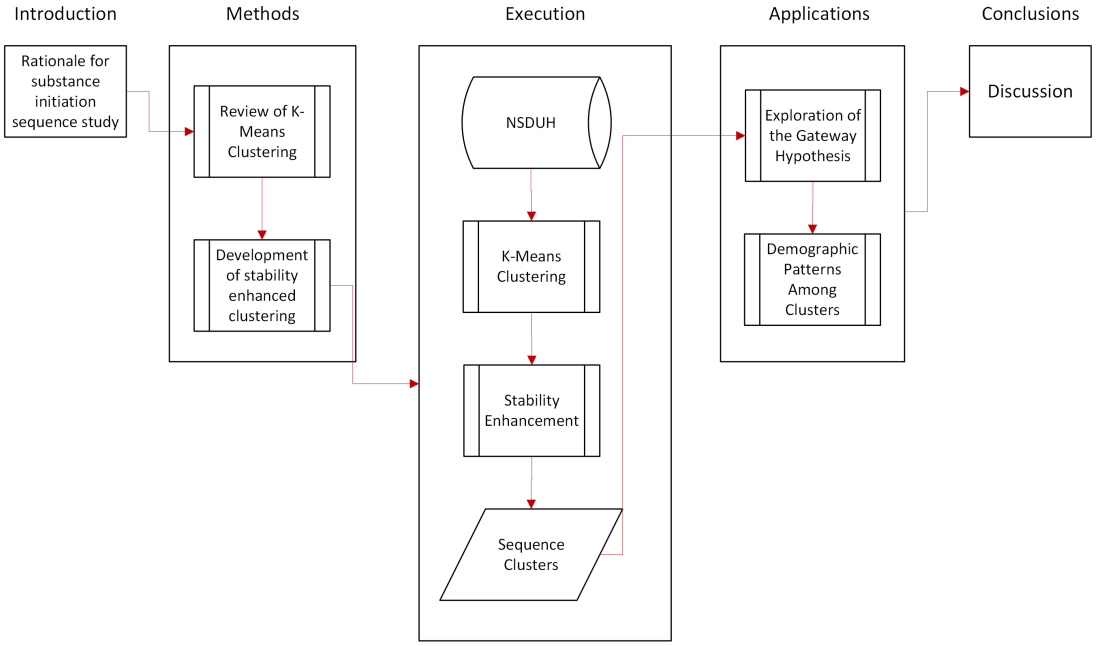


Figure 1.1: Discussion Flow

1.2 History of drug use prior to legalization

Throughout the 20th and 21st centuries, substance abuse has been a subject of ongoing concern and has caused great harm to the US population. It was recognized as a problem over 100 years ago, and local, state, and Federal governments have all pursued a variety of policies to mitigate it. One drug, heroin, was initially created as a medicine in 1898 and has a particularly long history of abuse. New

York's Bellevue hospital had its first admission for heroin addiction in 1910. The Federal government sought to eliminate some substance abuse, beginning with the Harrison Narcotic Act in 1914. Partly created to comply with international treaties, the act required the registration and taxation of the manufacture and distribution of narcotics, which at the time included opium, morphine, heroin, and coca products (US Drug Enforcement Administration, 2021a, p.12). By 1924, the US Congress banned all domestic manufacture of heroin (Scott, 1998). Public concern with substance abuse was not limited to less common drugs – during this era the Temperance Movement, which believed public health was endangered by alcohol, was able to pressure Congress into passing the Volstead Act. That legislation banned the sale and use of alcohol and launched Prohibition (Blocker, 2006).

The 18th Amendment to the US Constitution ended Prohibition, and the US entered a long period during which tobacco and alcohol were legal, albeit with increasing restrictions on age-of-use, and all other substances were banned. Narcotics use fell for a time, only to rise again after World War II with the introduction of synthetic alternatives to morphine such as dilaudid, amadone, and methadone. Cocaine use also grew during this period (US Drug Enforcement Administration, 2021a, p.20). The use of narcotics, including heroin, rose during the Vietnam War to the point where on July 14, 1969, President Richard M. Nixon messaged Congress that, "Within the last decade the abuse of drugs has grown from essentially a local police problem into a serious threat to the personal health and safety of millions of Americans (US Drug Enforcement Administration, 2021a, p.26)." Congress responded by passing of the Comprehensive Drug Abuse Prevention and Control Act on October 27, 1970. The act replaced previous legislation and provided a combined framework for treatment, rehabilitation, education, regula-

tion and enforcement. Title II of the act, commonly known as the Controlled Substances Act (CSA), includes five schedules that classify controlled substances according to their relative potential for abuse (US Drug Enforcement Administration, 2021a, p.27). Some examples from the current schedule are provided in Table 1.1 (US Drug Enforcement Administration, 2021b). Notably, alcohol and tobacco, despite their potential for abuse and addiction, are not covered in the schedule. On July 1, 1973, the Drug Enforcement Administration (DEA) was launched, and a long period during which the recreational use of alcohol and tobacco was tolerated[2], while all other substances, including marijuana, were illegal unless administered as medicine.

Table 1.1: FDA Drug Schedule

Schedule	Definition	Examples
I	No currently accepted medical use and a high potential for abuse	heroin, LSD, marijuana, ecstasy
II	High potential for abuse, with use potentially leading to severe psychological or physical dependence. These drugs are also considered dangerous.	hydrocodone, cocaine, methamphetamine, fentanyl
III	Moderate to low potential for physical and psychological dependence	codeine, ketamine
IV	Low potential for abuse and low risk of dependence	Xanax, Valium, Atavan
V	Lower potential for abuse than Schedule IV and consist of preparations containing limited quantities of certain narcotics. Generally used for antidiarrheal, antitussive, and analgesic purposes	cough syrups with codeine

The effect of these efforts has been arguable at best. The number of Americans with an illicit drug use disorder has been unchanged in recent years – it

[2]Alcohol and tobacco, while legal throughout most of the US, have been subjected by the states to increases in the legal age of use.

was 3.0% in 2003, and remained at 3.0% in 2019 (Substance Abuse and Mental Health Administration, 2019a). The number of patients admitted to care for substance abuse declined from 859 per 100,000 population in 2007 to 719 in 2018. However, this decline included a shift in the substances that caused admissions – opiate-related admissions grew 62.5%, and heroin-related admissions grew 67% (Substance Abuse and Mental Health Administration, 2019b).

1.3 Legalization of illicit substances

A recent shift in substance use has come about as a result of increased access to marijuana. In the 2000s, public sentiment began to drive the legalization of marijuana. California declared marijuana to be legal for medicinal use in 1996, an act that was soon followed by Alaska, the District of Columbia, Maine, Colorado, Oregon, Washington, Hawaii, and Nevada by 2000. Although marijuana was not legal for casual use, its prevalence increased steadily (Figure 1.2). In 2002, 6.2% of respondents to a national survey reported marijuana use within the past month. By 2014 that number increased to 8.4%. Marijuana first became legal for *recreational* use by people 21 or older in Colorado and Washington 2012, and increased usage followed immediately, especially among the young. In 2011, 8% of survey respondents in Washington reported past-month marijuana usage. By 2014, 10% reported usage, with the largest increase among 18-24 year olds, whose usage jumped from 15% to 21% (Campo et al., 2016). Legalization has since progressed to many other states, and it has been followed by a dramatic increase in marijuana use. As of 2019, national past month usage of marijuana among 12+ year olds has risen to 11.5%.

A dramatic increase in the potency of marijuana has accompanied its rise

in usage. Tetrahydrocannabinol (THC) is the primary psychoactive component of cannabis, and its concentration among sampled products rose from 8.9% in 2008 to 17.1% in 2017. So not only are there more Americans using marijuana, the substance they are using is much stronger, and these trends may indicate that people who use cannabis are at greater risk of harm than in previous years (Chandra et al., 2019).

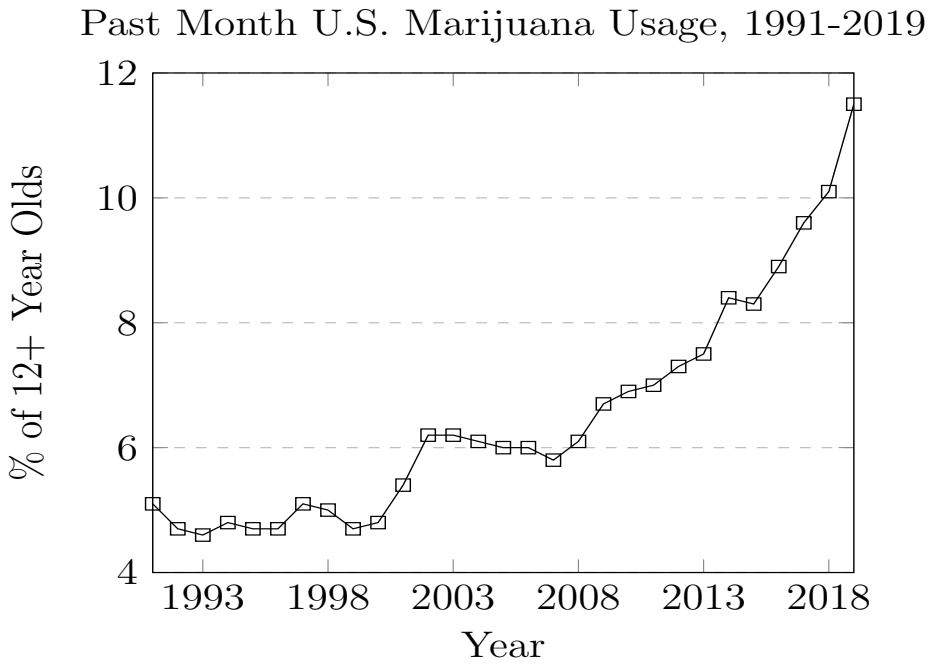


Figure 1.2: Past Month Marijuana Usage Among 12+ Year Olds (Substance Abuse and Mental Health Administration, 2019a)

1.4 An epidemic of fatal substance abuse

The most concerning recent aspect of substance use in the US has been the opioid epidemic. Since 1999, there has been a marked increase in substance-induced deaths across all races in the US (Figure 1.3). This rise has been well documented, and much of it occurs from opioid abuse. Since the middle of the last century, the

way opioid abusers initiate their habits has changed dramatically. In the 1960s, over 80% of users' first opioid abuse was with heroin. In the 2000s, over 70% of users initiated abuse with prescription opioids. This trend reversed somewhat in the 2010s as the heroin became the first opioid of abuse for 35% of users (Cicero et al., 2014).

The rise in opioid deaths has progressed in three overlapping phases (Figure 1.4). The first wave began with increased prescription of opioids such as oxycodone and hydrocodone in the 1990s. This increase occurred as attitudes shifted to help patients avoid pain despite a lack of objective studies to quantify the risks of an increase in opioid prescriptions (Wilkerson et al., 2016). As a result, deaths due to this activity have increased since 1999. Abuse deterrent formulas of opioids helped to reduce their misuse (Wilkerson et al., 2016), but other abuse patterns arose. The second wave began in 2010 as heroin usage increased and caused a steep increase in overdose fatalities. The third wave began in 2013 and has seen a dramatic rise in deaths due to synthetic opioids such as fentanyl. Fentanyl, which is approved for cases of extreme pain, is 50 to 100 times more potent than morphine. Fentanyl abuse presents an extreme challenge because it is often mixed with heroin or cocaine without the knowledge of the user.

Deaths due to other substances have increased as well. Somewhat masked by the opioid epidemic is an increase in deaths induced by alcohol. Following a period of decline from 1999 to 2005, the number of deaths due to alcohol per 100,000 persons has risen steadily (Figure 1.5).

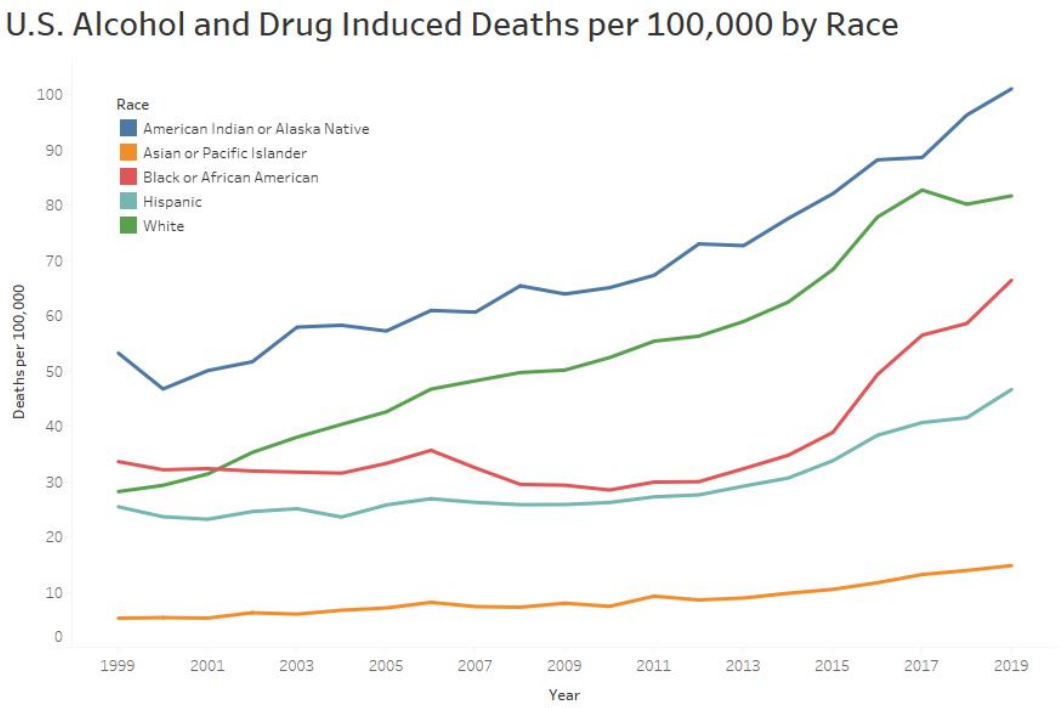


Figure 1.3: Substance Induced Deaths by Race
(*Underlying Cause of Death 1999-2019 on CDC WONDER Online Database, released in 2020. Data are from the Multiple Cause of Death Files, 1999-2019*, 2020)

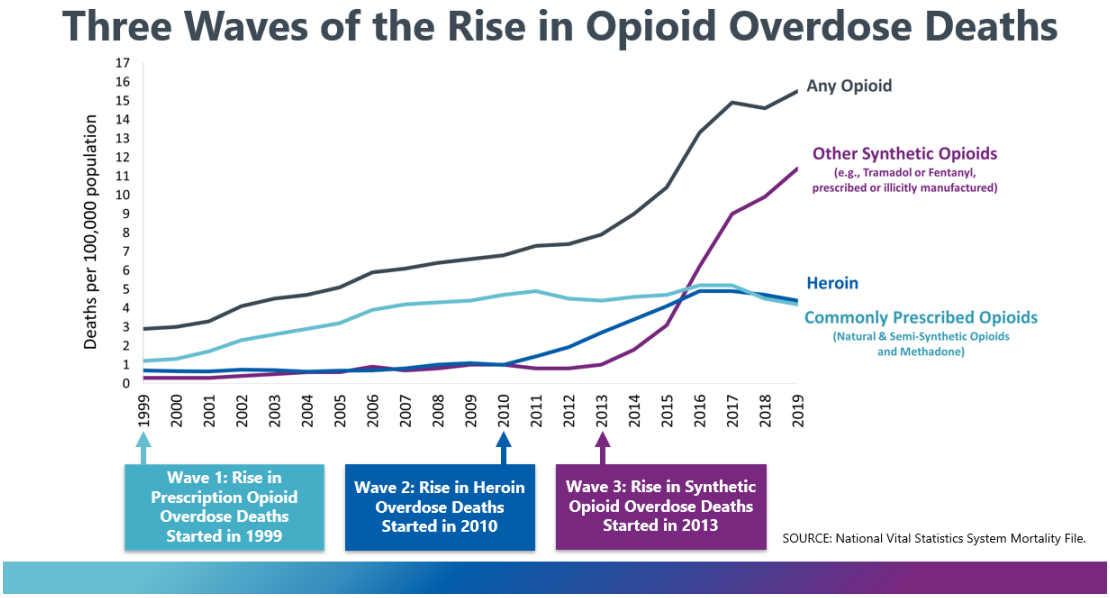


Figure 1.4: Phases of Opioid Abuse
(Centers for Disease Control and Prevention & Control, 2021)

1.5 Studying sequences of drug use stages

In 2002, the US Government estimated that the annual economic cost of drug abuse was $180.7 billion (Office of National Drug Control Policy, 2002), and this was *near the beginning* of the opioid epidemic. A recognition of the societal impact of substance abuse has driven a great deal of research in the hope of developing treatments and policies to improve the situation. Studies include investigation in virtually every discipline, including epidemiology, physiology, psychiatry, psychology, behavioral science, and statistics. One area of research has been on understanding the mechanisms and initiation of drug use among adolescents, a focus that can support prevention strategy and development (Zhang et al., 2021). Many studies focus on the fact that users necessarily initiate abuse of multiple substances in a sequence of stages. Identification of predominant sequences of drug use initiation can help identify populations at risk for progression from legal

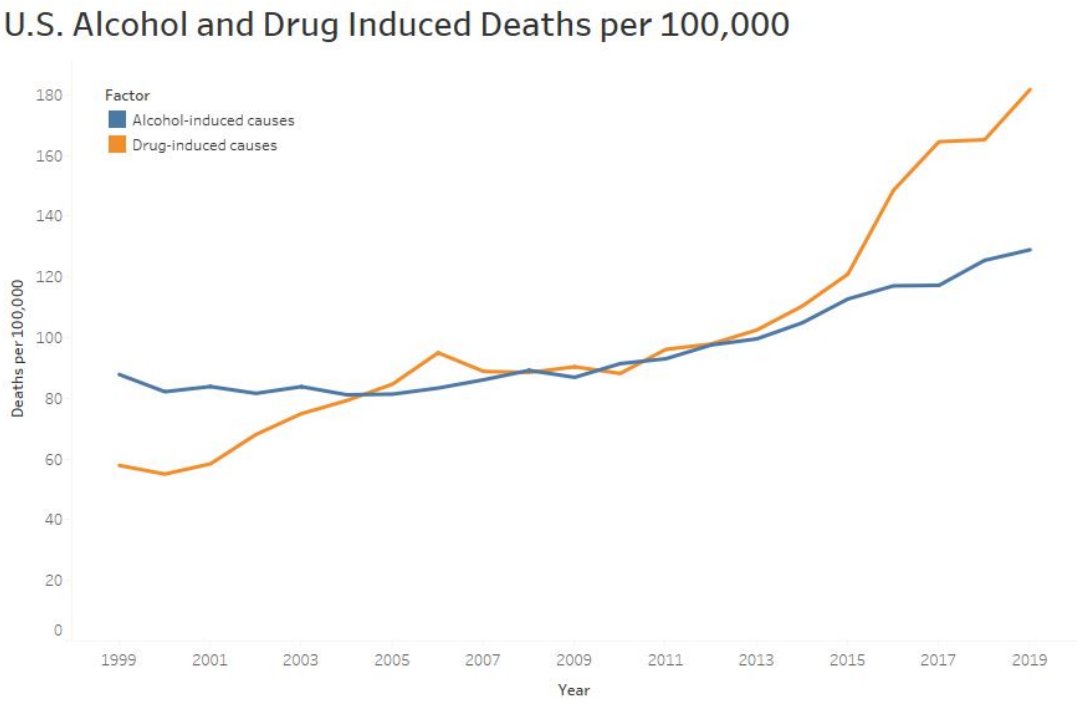


Figure 1.5: Alcohol and Drug Induced Deaths
(*Underlying Cause of Death 1999-2019 on CDC WONDER Online Database, released in 2020. Data are from the Multiple Cause of Death Files, 1999-2019*, 2020)

or less dangerous drugs to more harmful ones (Adler & Kandel, 1981).

In 1975, Denise Kandel published a foundational study that claimed a rigid progression in substance use (D. Kandel, 1975). She posited that substance use initiates with beer or wine, progresses to hard liquor and cigarettes, followed by marijuana, and eventually by other illicit drugs such as LSD or heroin. This behavior can be described as a progression through a sequence of substance abuse *stages*. The zero stage is no-use – a subject has never used a drug. The first stage occurs with the subject's first use of any drug, which D. Kandel (1975) maintained is likely to have been either beer or wine. The second stage continues with use of either hard liquor or tobacco, then by whichever of those has not yet been used. The third stage is the subject's first use of marijuana. Stage four involves use of other illicit drugs. The sequence is depicted in the flowchart from Kandel's original work (Figure 1.6). The numbers in the figure represent the probability of a respondent's transition from one stage to the next from the fall of 1971 to the spring of 1972. Kandel's idea would evolve into the *Gateway Hypothesis*, which has guided some abuse mitigation strategies. Prevention practices aimed at avoiding early stage substance use at young ages have been shown to lower the probability of other illicit drug use (G. Botvin et al., 2001; G. J. Botvin et al., 2000; Hawkins et al., 1992).

As shown in Sections 1.3 and 1.4, US substance use behavior is not static – it continues to evolve. Therefore, continued study of drug initiation sequences is required to support the development of new abuse mitigation strategies. Consider a hypothetical example in which a policy maker seeks to minimize the use of opioids. She develops a program that focuses on tobacco and alcohol prevention among adolescents because she has reviewed research that claims those are entry stage drugs strongly associated with later use of other substances. If in fact mari-

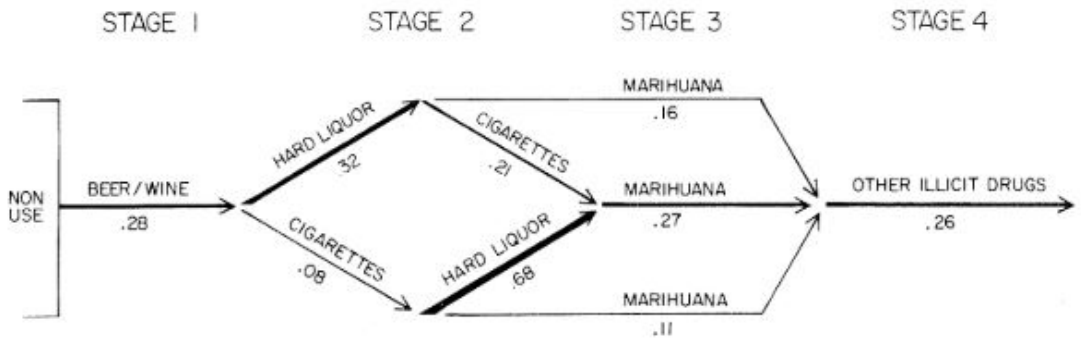


Figure 1.6: Kandel's Substance Initiation Stages
(D. Kandel, 1975)

juana has become a common first-use drug also associated with subsequent drugs, she would overlook a cohort of users that are not interested in tobacco or alcohol. Similarly, she knows that users often accidentally overdose on fentanyl disguised as heroin. If she was to learn that users have commonly abused cocaine prior to initiating heroin, she could include heroin prevention in a treatment regimen delivered to rehabilitation center patients who were admitted for cocaine use.

The study of drug use initiation sequences is difficult. Two of the biggest challenges faced by researchers are access to information and data complexity.

1.5.1 The challenge of information access

An investigator is not going to directly observe the drug using behavior of a significant number of subjects, and the chance of being present when a user first initiates use of a substance is negligible. Therefore, studies must rely on third party observations such as treatment data or on self-reported data from the user. Verified third party observations would be preferable, and there are datasets that contain usage information. One is the *Treatment Episode Dataset* (TEDS), administered by Center for Behavioral Health Statistics and Quality of the Sub-

stance Abuse and Mental Health Services Administration (SAMHSA), part of the US Centers for Disease Control. TEDS contains demographic, clinical, and substance use information associated with patients admitted to and discharged from substance treatment facilities that report data to state administrations. Unfortunately, TEDS only gathers age of first use (AFU) data regarding the substance(s) that drove the admission. Also, if a patient has multiple admissions to facilities, there is no way to tie those together due to patient confidentiality. Another way to gather verified information would be via electronic patient medical records (EMRs), which can be obtained with redacted identities through collaboration with insurance companies. However, EMRs will not contain a complete history of drug AFU information for a patient. Researchers are thus limited to gathering drug sequence information through surveys of users.

Longitudinal studies consist of multiple surveys administered to a stable set of respondents over a time horizon. When used to gather drug initiation sequences, these improve the chance that a user recently initiated a substance and can accurately recall their age when they did so. For example, if a respondent is surveyed in 10th grade and indicates initiation of alcohol within the past year, their reported AFU is likely to be accurate. When that same respondent is interviewed in 12th grade, his report of initiated substances since the last survey will also provide accurate AFUs. Longitudinal studies have been conducted on drug initiation sequences and have provided valuable insight into user behavior, especially among adolescents (Fergusson & Boden, 2008; Fergusson et al., 2006; D. B. Kandel et al., 1992). These surveys are obviously difficult to administer. They require identification and tracking of a set of respondents over many years. More importantly, the time required to track the likely drug initiation horizon of users is long. A user who started smoking at age 14 could first abuse prescription opioids at age

40.

A more common approach is a cross-sectional survey that gathers drug initiation data via recall by the respondents. In this case, the interviewer asks questions regarding multiple substances, such as, "Have you ever used oxycodone without having been prescribed it?", and "At what age did you first use oxycodone without having been prescribed it?" The number of drugs that can be investigated is only limited by the time available to the interviewer and respondent, so a rich set of AFU data can be gathered in this fashion. Cross-sectional surveys have the benefit of scale and repeatability. There are several surveys that have been administered to thousands of respondents over several decades, including the *National Survey on Drug Use and Health* (NSDUH), administered by SAMHSA, the *National Comorbidity Study* (NCS) from Harvard Medical School, and *Monitoring the Future* (MTF) from the University of Michigan. There have been many other smaller scale cross-sectional surveys of drug initiation sequences, including D. Kandel (1975), Adler and Kandel (1981), Fleming et al. (1989), Wagner and Anthony (2002), and Fuller et al. (2005).

All surveys regarding drug use are subject to measurement error. Both longitudinal and cross-sectional studies depend on the honesty of their respondents, who may be motivated to either minimize or embellish their drug use history. This error can be mitigated somewhat through carefully designed surveys that use combinations of questions that aim to reduce untruths through inconsistent answers. Cross-sectional surveys have the additional risk of faulty recall. As a respondent's age increases past age-of-first use, their proper recall of that age may drift. Fortunately, several studies have shown that during the years of young and middle adulthood, the age-of-first use reported by subjects does not deviate significantly with the number of years elapsed since the use occurred (Golub et al.,

2000; Prusoff et al., 1988; Wittchen et al., 1989).

1.5.2 The challenge of data complexity

The other major challenge to drug initiation sequence research is data complexity. There are many different substances that a subject may use, and she can first do so at any point in her life. This results in a great many potential sequences of substance use initiation. For example, if all substances are grouped into ten categories, and only the order in which they can be initiated is desired, there are approximately 3.6 million potential sequences to consider. If instead the AFU for each drug used by a respondent is taken into account, and the likely age of initiation horizon is 14-50 years of age, the number of possible AFU sequences becomes approximately 36^{10}. Of course, not all drug initiation patterns are equally likely, and the goal of research is to find those that will be observed in a significant number of subjects.

Many studies have approached this challenge by defining a priori hypothesized drug sequences, by ignoring AFU as a determining factor of sequence definition, limiting the substances under consideration, or some combination of the three. D. Kandel (1975) constructed sequences through observation, but did not consider AFU within them and limited the study to adolescents. Adler and Kandel (1981) mirrored this approach. Morrison and Plant (1991) pre-defined subgroups of usage based upon basic combinations of cannabis and other substances without regard to AFU and then measured AFU differences across the groups. Based upon prior research, D. B. Kandel et al. (1992) assumed a basic initiation sequence of alcohol, cigarettes, marijuana, and other illicit drugs, including prescribed psychoactive drugs. It then hypothesized slight modifications to this sequence and tested for them. This study did not consider AFU as a determining factor of the sequences.

Fleming et al. (1989) was limited to adolescents and a short list of substances with the goal of validating D. Kandel (1975). Blaze-Temple and Lo (1992) ignored AFU and examined combinations of a limited set of drugs to determine the most 'important' gateway drug. Golub and Johnson (1994) limited the number of substances in the sequences to normative early stage drugs and sought their role in later polydrug abuse. Wagner and Anthony (2002) did not develop preconceived sequences and did measure AFUs, but limited the substances under consideration and focused on the relationship between the AFU of a drug and the likelihood of becoming dependent upon it. Fuller et al. (2005) did not consider sequences but instead analyzed current usage patterns of different age cohorts to infer initiation sequences among adolescents. Fergusson et al. (2006) and Fergusson and Boden (2008) focused on a two-stage sequence, marijuana followed by other illicit drugs, to determine the relationship between levels of marijuana usage and adverse life events, including initiation of other substances. Degenhardt et al. (2009) did not define a priori sequences in an evaluation of the relationship between deviations from the assumed gateway normative progression and drug dependence problems. It did not consider AFU as a factor in determining the sequences. Keyes et al. (2016) studied age cohorts from the 1991-2008 Monitoring the Future (MTF) surveys to seek changing drug use patterns and relationships between adolescent smoking and later substance use. This study didn't seek explicit sequences and respondents were limited to 8th, 10th, and 12th graders. Barry et al. (2016) obtained AFUs and did not define a priori sequences in an effort to find the first drug used by eventual polydrug abusers. It limited the study to 8th, 10th, and 12th graders, which prevents consideration of sequences where drugs were initiated as adults. Fiellin et al. (2013) examined the dependence of prescription opioid abuse on the adolescent use of alcohol, tobacco, and marijuana. The study

did not address later use of other drugs. Lee and Petlakh (2020) ignored both sequences and AFUs in favor of a binary definition of adolescent marijuana use and its relationship to later use of other illicit drugs.

Some studies have looked for sequences without pre-definition, considered AFUs, and evaluated a great number of drugs. Darke et al. (2012) evaluated the AFUs of a great many drugs and sequences in an effort to see how they change among age cohorts. This study was limited to a very small (269) set of respondents. Zhang et al. (2021) built models that predicted the usage probability of a new drug as a function of the time lapsed from the initial use of a different drug. This study considered AFU, did not constrain sequences, and evaluated a great range of substances. The study limited its consideration to users between 14-17 years of age and left alcohol and tobacco out of the design.

1.6 Gaps in the literature

A summary of the aforementioned literature is presented in Table 1.2. For each study the data source and principal method is listed. The next three columns indicate whether AFUs were included in the study, whether a large number of substances are considered, and whether the study did not use a priori sequences. What is missing from this list is a study that combines an evaluation of many drugs, consideration of the AFUs for them, and most importantly, an exploration of the data unencumbered by an expectation of specific sequences. While there are many more works that discuss substance abuse sequences, one could not be found that met all of these criteria:

- Makes use of a very large and robust dataset
- Seeks to describe not just drug initiation sequences, but the AFUs associated

with those sequences

- Avoids the use of a priori hypothesized sequences
- Uses scalable machine learning (ML) techniques to explore data
- Ensures exploration does not overfit survey data, which is necessarily an approximation of a general population

This study meets these criteria, and by doing so, presents a set of prevalent substance initiation sequences, including the self-reported age of first use for each drug in a sequence.

1.7 Addressing the challenges of drug sequence study

This study addresses the challenges described in Sections 1.5.1 and 1.5.2 by applying unsupervised machine learning through cluster analysis on the NSDUH dataset.

1.7.1 National Survey of Drug Use and Health addresses the challenge of access to information

A proper exploration of substance use initiation patterns requires a dataset that is designed to adequately reflect a large population, in this case that of the United States. It must contain enough observations to not only ensure accurate representation of the average population, but it must include data that captures the behavior of small but interesting cohorts. For example, very few Americans are

Work	Data Source	Method	AFU?	Many drugs?	Open Sequences?
D. Kandel (1975)	Longitudinal study	Guttman scalogram analysis	N	N	N
Adler and Kandel (1981)	Cross-sectional survey	Descriptive statistics	N	N	N
Fleming et al. (1989)	Cross-sectional study of adolescents	Descriptive statistics, Guttman scalogram analysis	Y	N	Y
Morrison and Plant (1991)	Cross-sectional survey	Descriptive statistics	Y	Y	N
Blaze-Temple and Lo (1992)	Cross-sectional survey	Hazard rate calculation	N	N	Y
D. B. Kandel et al. (1992)	Longitudinal study	Modified Guttman analysis	N	N	N
Golub and Johnson (1994)	Cross-sectional survey	Descriptive statistics, logistic regression	Y	N	Y
Wagner and Anthony (2002)	Cross-sectional survey (NCS)	Survival analysis	Y	N	NA
Fergusson et al. (2006)	Longitudinal study	Descriptive statistics, logistic and multiple regression	Y	Y	NA
Fergusson and Boden (2008)	Longitudinal study	Descriptive statistics, logistic and multiple regression	Y	N	NA
Degenhardt et al. (2009)	Cross-sectional survey (NCS)	Logistic regression and discrete-time survival analysis	N	Y	N
Darke et al. (2012)	Cross-sectional survey	Descriptive statistics	Y	Y	Y
Fiellin et al. (2013)	Cross-sectional survey (NSDUH)	Logistic regression	Y	N	N
Keyes et al. (2016)	Cross-sectional survey (MTF)	Multiple regression	Y	N	N
Barry et al. (2016)	Cross-sectional survey (MTF)	Descriptive statistics, Tukey HSD test	Y	Y	NA
Lee and Petlakh (2020)	Cross-sectional survey	Propensity score matching	N	Y	N
Zhang et al. (2021)	Cross-sectional survey (NSDUH)	Life-table analyses and Cox regression models	Y	Y	Y

Table 1.2: Substance Sequence Literature Review

heroin users, so a small general survey of the population would be unlikely to interview any. The dataset must also be carefully constructed through stratification and weighting so that the responses can be extrapolated to the total population. NSDUH meets these requirements.

SAMHSA is an agency within the U.S. Department of Health and Human Services whose mission is to reduce the impact of substance abuse and mental illness on America's communities. One of SAMHSA's publications is the NSDUH – an annually produced dataset that is the result of a massive survey regarding substance abuse and mental health in the United States. Because it has been administered since 1971, NSDUH is a well-established source of substance use information among U.S. residents.

NSDUH explores the use of illicit drugs, alcohol, and tobacco among members of the U.S. civilian, non-institutionalized population aged 12 or older. The survey also includes several modules of questions that focus on physical and mental health issues. Surveys have been conducted periodically since 1971, with the most recent ones in 1979, 1982, 1985, 1988, and annually from 1990 through 2019 (Center for Behavioral Health Statistics and Quality, 2020b). Currently, public use files are available for surveys from 1979 onward. The present study uses data from the 2016-2019 surveys.

NSDUH's sampling methodology is designed to capture as many geographic and demographic sections of the United States as possible. Each observation in the resulting dataset contains a weight which researchers can use to extrapolate the observation to a section of the population. The sum of the weights of the observations is equal to the population of the United States as measured by the most recent census. The survey is large – in 2019, NSDUH contained 55,271 observations gathered from 67,901 interviews conducted by 700 field investigators.

There are some limitations to NSDUH:

- The data are comprised of self-reports of drug use, and their value depends on respondents' truthfulness and memory
- The survey is cross-sectional rather than longitudinal. That is, individuals were interviewed only once and were not followed for additional interviews in subsequent years.
- Because the target population of the survey is defined as the civilian, non-institutionalized population of the United States, a small proportion (approximately 3 percent) of the population is excluded.

NSDUH contains questions pertaining to drug usage history, physical and mental health history, and demographic factors. The dataset also contains variables that have been imputed by SAMHSA to reduce missingness and to improve the accuracy of results. For example, if a respondent skips a question regarding whether she has ever used heroin but subsequently answers a question regarding the last time she used heroin, an imputed heroin use flag will be positive. SAMHSA recommends that researchers use imputed variables rather than direct responses. There are thousands of variables in the dataset – in 2019, there were 2,741 variables for each observation.

The richness of the data in NSDUH allows researchers to explore a great number of topics. For example, data from 2005-2014 show that there has been an increase in binge drinking and alcohol use disorder among subjects aged 50 and over (Han et al., 2017). Another study correlated drug use to employment, finding that subjects who were unemployed following the 2008 recession were more likely to have been marijuana users prior to losing their jobs (Compton et al., 2014).

The AFU data in NSDUH is subject to respondents' faulty recall of when they initiated substance use. As discussed in Section 1.5.1, other studies have shown that respondents exhibit consistency in their survey answers through middle adulthood. We therefore accept recall inaccuracy as a necessary but minor source of potential error in our study.

1.7.2 Cluster analysis addresses the challenge of data complexity

A user's progression through a sequence of drug initiation stages can be gathered from NSDUH by extracting the AFUs a respondent provides for the substances covered by the survey. As mentioned in Section 1.5.2, the number of possible first-use sequences is huge. The researcher must determine how to group sequences together based upon their similarity into categories that represent common patterns of behavior without guessing at them a priori. This is an example of an *unsupervised learning* problem. In *unsupervised learning*, the researcher seeks structures, patterns, and relationships among observations without the benefit of labels that have been assigned to a subset of known data. In the case of the NSDUH dataset, the researcher wishes to determine a small number of groups into which the observations can be uniquely placed based upon the similarity of their drug AFU sequences. Unsupervised aggregation of sequence data into groups based upon similarity can be done with *cluster analysis* (Dong & Pei, 2007). Efficient algorithms such as *K-means clustering* (KMC) can quickly obtain clusters from large datasets based upon an inter-observation distance metric defined by the researcher.

Cluster analysis has been used in prior studies of substance abuse behaviors.

Young Mun et al. (2008) used the technique to categorize problem behavior among adolescents. Sevigny and Coontz (2008) applied hierarchical clustering to groups of individuals based upon arrest patterns and self-reported substance use. Harrington et al. (2012) used clustering to find patterns of alcohol and marijuana usage and frequency. Panlilio et al. (2020) used hierarchical and K-means clustering to determine patterns of drug test results. Cluster analysis has even been applied to NSDUH data. Wang et al. (2019), Ategbole et al. (2021), and Xie et al. (2022) conducted principal component clustering for feature reduction in studies that showed relationships between drug usage and demographic variables and the need for mental health services. Liew (2016) used cluster analysis to group military respondents with similar drinking and sociodemographic characteristics.

Despite efforts employed to make NSDUH highly representative of the US population, it remains a survey administered to a sample of Americans. A cluster analysis performed on NSDUH risks *overfitting* the survey – the groups may be very well suited for the survey studied, but less accurate when applied to a different population sample. A 'stable' clustering method is one in which similar observations are consistently grouped together despite perturbations in the data. Application of a stability-enhanced clustering method applied to NSDUH or other substance use initiation data was not found in the literature.

1.8 Contributions of this work

This work makes several contributions to existing knowledge. First, it uses unsupervised learning to explore a very large dataset in order to uncover, without prior bias, patterns of substance use initiation in the US. Furthermore, it considers age of first use as a determining feature of these sequences, where most prior work

has focused only on the order of substance initiation stages. Second, it applies stability-enhanced clustering methods in determining these patterns in order to minimize overfitting of the NSDUH study. Third, it presents novel algorithms and a computational approach to deal with the size and complexity of the data. Finally, it demonstrates the utility of patterns discovered in the study. Specifically, it considers the Gateway Hypothesis in light of the patterns discovered here, and it analyzes variation of membership in the sequence clusters with regard to demographic features.

Chapter 2

Methods

2.1 K-Means Clustering

There are many approaches to analyze and derive insights from large, complex data sets. The most common methodology in the drug abuse and addiction literature is to use a traditional and formal statistical hypothesis analysis. Researchers define specific hypotheses regarding drug use patterns, etc. and then analyze the related subsets of data to determine if there is sufficient evidence to explicitly reject one hypothesis in favor of another. The hypothesis serves to bound the data by reducing the number of variables considered in the study. NSDUH contains thousands of features, but a hypothesis test can gather important information efficiently. For example, one could form the hypothesis, "Heroin users initiate their drug use differently than non-heroin users." To test this, a group of heroin users and a group of non-heroin users can be selected from the database along with their associated distributions of initial drug use. A statistical test, such as the χ^2 test for independence, can be employed to evaluate the null hypothesis that the initial drug use distribution is independent of heroin usage. If there is

sufficient evidence to reject the null hypothesis, we conclude that heroin users do indeed differ in their drug initiation and we can gather from the test what their most common first drugs are. This traditional statistical hypothesis methodology is the gold standard for addressing specific, investigator-defined questions. However, a drawback of the traditional technique is the very limited nature of the hypothesis generating process. Specifically, such an approach is not amenable to broader-based and more general questions. Furthermore, the approach is limited to addressing only the a priori questions determined by the investigator – there is little room to let the "data speak for themselves" (Gould, 1981). For instance, if one seeks to answer the general question, "what drug initiation patterns exist in the United States", it is critical to not be limited to only preconceived notions of drug usage. Rather than construct a massive set of hypotheses, the researcher wants to use techniques that easily answer this general inquest.

Unsupervised learning is a modern approach to addressing such open and general questions. It is best understood through a comparison with its alternative, supervised learning. Supervised learning includes machine learning methods associated with using input data to predict or classify a data observation according to a known outcome or label. Unsupervised learning is set of techniques that attempt to identify patterns, structures, and/or relationships among unlabeled data, or at least data in which a particular outcome variable is of little interest. Unsupervised learning is often used to uncover aspects of data through exploration that can in turn be used for more detailed analyses. For example, *principal component analysis* is a method used to reduce the dimensionality of data by creating new features from linear combinations of the original input data in such a way as to preserve the information in a reduced feature set. These features can then be used for visualization or supervised techniques. Additionally, unsupervised learning can

be used to directly discover important but hidden structures in the data.

Clustering, an important type of unsupervised learning, is a set of techniques for finding subgroups, or clusters, in a data set. *K-means clustering* (KMC) is a common and well-explored unsupervised learning method. KMC partitions observations in a dataset into mutually exclusive and exhaustive sets called *clusters*. The algorithm seeks to partition the data such that all elements in the same cluster are similar (intracluster similarity) to one another and dissimilar to elements in other clusters (intercluster dissimilarity). The KMC algorithm was first developed by Stuart Lloyd at Bell Laboratories in 1957, not formally published until 1982 (Lloyd, 1982), and is often referred to as the *Lloyd Algorithm*. The formal optimization problem addressed by the Lloyd Algorithm was first described by MacQueen (1967).[1] The goal of the problem is to allocate n observations, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, each of which has p features, into K subsets. Let $C_1, \ldots, C_K$ be sets containing observations uniquely allocated to each cluster. Then each of the n observations is allocated to one and only one cluster:

$$C_1 \cup C_2 \cup \ldots \cup C_K = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$$
$$C_k \cap C_{k'} = \emptyset \text{ for all } k \neq k'$$

The allocation of n observations into K clusters is a *partition*, denoted by S. Let $\mathbf{S}$ denote the set of all possible partitions. Let $W(C_k)$ denote the intracluster variation among observations within a cluster k. The sum of all intracluster variations for S is given by $\sum_{k=1}^{K} W(C_k)$ and is known as *inertia*, denoted by I. For a given value of K, the optimal partition, denoted S^*, generates the minimum

[1] The mathematical notation for the following description of K-means clustering is taken from James et al. (2013, p. 385-389).

inertia, I^*, as described in Equation 2.1:

$$I^* = \min_{\mathbf{S}} \sum_{k=1}^{K} W(C_k) \tag{2.1}$$

Intracluster dissimilarity is determined based on a distance measure calculated between observations within the same cluster. Squared euclidean distance is commonly used. Let each observation $\mathbf{x}_i$, for $i = 1 \ldots n$, be represented by a real-value p-dimensional vector, $\mathbf{x}_i = [x_{i1}, \ldots, x_{ip}] \in \mathbb{R}^p$. The inertia of cluster k is computed according to Equation 2.2:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \tag{2.2}$$

Substitution of 2.2 into 2.1, leads to the complete KMC problem definition – find $S^* \in \mathbf{S}$ to minimize the inertia:

$$I^* = \min_{\mathbf{S}} \left[\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right] \tag{2.3}$$

Finding S^* becomes difficult as n increases in size. A partition of n observations into K clusters requires the evaluation of

$$\frac{1}{K} \sum_{k=1}^{K} (-1)^{K-k} \binom{K}{k} k^n$$

alternatives (Jain & Dubes, 1988, p.91). In fact, Drineas et al. (2004), Dasgupta (2007), Aloise et al. (2009) and Mahajan, Nimbhorkar, and Varadarajan (2012) proved that the problem is NP-hard. For large values of n, finding S^* is impractical. For this reason, the problem is commonly addressed by heuristics that find *local*, not *global* optimum solutions to KMC. That said, the methods are effi-

cient – Selim and Ismail (1984) prove finite convergence of Lloyd-type algorithms to either a local optimum or a Kuhn-Tucker point. The simplicity and ease of implementation of the basic KMC algorithm (Algorithm 1) make it popular.

Algorithm 1 Basic K-Means Clustering Algorithm

Require: Observations $\mathbf{X} : \mathbf{x_i} \in \mathbb{R}^p, i = 1, \ldots, n$, number of clusters $K \in \mathbb{Z}$

1: generate K random p-dimensional centroids, μ_k for $k = 1, \ldots, K$
2: **while** Cluster assignments do not change **do**
3: assign each observation to the nearest centroid to form the clusters $C_1, \ldots, C_K$
4: update the cluster centroids of each cluster, $\mu_k = 1, \ldots, \mu_K$ by computing the mean of all observations assigned to each cluster $C_1, \ldots, C_K$
5: **end while**

One challenge of KMC is that the practitioner must select the number of clusters, K, into which the observations will be partitioned. There are many methods for selecting K, including three popular ones, the *elbow method*, the *silhouette coefficient method*, and the *gap statistic method* (Yuan & Yang, 2019). The elbow method is a heuristic that is easy to apply to large datasets. In it, a sequence of KMC partitions is made, changing the value of K each time. The total inertia for each partition is calculated and plotted versus K. The plot is then examined to find an 'elbow' – a point at which the inertia drops significantly. This method is practical, but oftentimes the elbow is not readily apparent.

The silhouette coefficient method (Rousseeuw, 1987) seeks to maximize the *cohesion* and *resolution* of clusters. Cohesion is the similarity of an object to its cluster, and resolution is the separability between a cluster and its neighbors. For each value of K under consideration, KMC is run. The average dissimilarity of a data point $\mathbf{x}_i$ to all other members of the cluster C_k to which it is assigned is given by $a(\mathbf{x}_i)$. The average dissimilarity of $\mathbf{x}_i$ with all members of another cluster C_j is given by $d(\mathbf{x}_i, C_j)$. Define $b(\mathbf{x}_i) = \min_{C_j \neq C_k} d(\mathbf{x}_i, C_j)$. Then the silhouette s for

$\mathbf{x}_i$ is:

$$s(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{\max[a(\mathbf{x}_i), b(\mathbf{x}_i)]}$$

The best value for K is the one that maximizes the average $s(\mathbf{x}_i)$ across all data points. This method is superior to the elbow method in that it generates a quantitative optimum for K. However, it requires that the distances between all data points be calculated and stored. For large datasets this can be prohibitive.

Tibshirani et al. (2001) formalized an automatic approach to determining the optimal number of clusters through calculation of the *gap statistic*. The gap statistic compares the difference between within-cluster variation of a K cluster partition of the data with the variation expected under a K cluster partition of a reference null distribution of data, commonly the uniform or normal distribution. Like the silhouette method, the gap statistic provides a specific optimum for K, and there have been subsequent improvements on the method Yan and Ye (2007). Its drawback is that it is computationally expensive. For every value K, the algorithm develops multiple sets of Monte-Carlo simulated observations for the reference distribution and calculates $W(C_k)$ for each. This can be impractical for large datasets.

Since NSDUH is large, both in terms of the number of observations and features, the use of the silhouette coefficient and the gap statistic become impractical. So while both of these methods may provide more quantitative guidance for determining K, the elbow method is most appropriate for partitioning NSDUH.

2.2 Cluster Stability

Because it is a heuristic-based algorithm, KMC is not guaranteed to converge to a global optimal solution, and the final cluster assignment is sensitive to both the

initial random centroids chosen and the particular data sample being analyzed.[1]. Ideally, KMC would produce clusters whose members are the same regardless of starting criteria. This study defines cluster *stability* as a relative condition. A set of observations that is consistently grouped into the same cluster over multiple iterations of KMC is more *stable* than a grouping of observations whose membership varies.

Rand (1971) addresses the question of stability by examining how clusters changed when constructed from different samples of data. Levine and Domany (2001) defines a general method to assess the stability of a partition that forms a basis for other methods, including the approach developed in this research. It defines two useful concepts. The first is the *connectivity matrix*, $\mathbf{T}$, which for a given partition indicates pairwise memberships in the same cluster: $T_{ij} = 1$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ are in the same cluster, and T_{ij} is zero otherwise. For example, consider a set of six data elements $\{1, 2, 3, 4, 5, 6\}$. If $\{1, 2\}$ are in cluster C_1, $\{3, 5, 6\}$ are in cluster C_2, and $\{4\}$ is in its own cluster C_3, the connectivity matrix becomes:

$$\mathbf{T} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

The second is the *figure of merit*. Let V denote the set of centroids $\mu_1, \ldots, \mu_K$ for the initial clusters $C_1, \ldots, C_K$. Using V, KMC is first run on the complete dataset to generate a partition S with connectivity matrix T. KMC is then run, using the same V each time, on m samples of the dataset to generate a set of partitions $S^{(1)}, \ldots, S^{(m)}$ with connectivity matrices $T^{(1)}, \ldots, T^{(m)}$. The figure of

[1]For a review of methods proposed to improve initial selection of cluster centers, see Celebi et al. (2013)

merit, $0 <= M(V) <= 1$, measures the similarity of connectivity matrices, where 1 represents a perfect match in all elements of the compared matrices, and 0 represents no matches in any element of the matrices. This process is repeated for different sets V, and the best parameter set V^* maximizes $M(V)$.

Ben-Hur et al. (2002) specified a particular figure of merit computation for cluster similarity analysis. Let two partitions $S^{(1)}$ and $S^{(2)}$ of the same dataset be represented by their connectivity matrices, $\mathbf{T}^{(1)}$ and $\mathbf{T}^{(2)}$. Then their similarity is the number of common edges of the two matrices and is calculated via the dot product:

$$\mathbf{T}^{(1)} \cdot \mathbf{T}^{(2)} = \sum_{i,j} T_{ij}^{(1)} T_{ij}^{(2)}$$

which can be normalized into a correlation measure[2] that can be used for Levine's figure of merit:

$$M(V) = \text{cor}(\mathbf{T}^{(1)}, \mathbf{T}^{(2)}) = \frac{\mathbf{T}^{(1)} \cdot \mathbf{T}^{(2)}}{\sqrt{(\mathbf{T}^{(1)} \cdot \mathbf{T}^{(2)})(\mathbf{T}^{(1)} \cdot \mathbf{T}^{(2)})}}$$

The aforementioned methods are used for comparing the stability of complete partitions of a dataset. A complementary problem is to create only stable clusters in the first place. Tibshirani and Walther (2005) defines a way to quantify the predictive strength of a clustering method. It creates a training and test set from data and independently partitions each. The cluster centers from the training set are then used to partition the test data, and the memberships within this partition are compared to the original independent partition of the test set. Their method includes an evaluation of the average number of times pairs of observations appear in clusters drawn from cross-validation folds of a dataset. This measure

[2]Ben-Hur et al. (2002) also describe two other similarity measures, the *matching coefficient* and the *Jaccard coefficient*.

was adapted by Tseng and Wong (2005) into a method to determine stable clusters of data. The authors define an algorithm, referred to as *Algorithm A*, to identify candidates for stable clusters by applying KMC to B multiple subsets of the data. From these runs, clusters are formed by aggregating observations that occur in the same clusters at least $B(1-\alpha), 0 \leq \alpha < 1$ times, where $1-\alpha$ is the desired 'tightness'.[3] The process is repeated for successively decreasing values for K. A comparison is made across the cluster sets to find one that changes the least from $K = k$ to $K = k-1$. This cluster is defined as stable and is removed from the dataset. The process is repeated until a number of clusters chosen by the researcher is found.

This current study builds on *Algorithm A* from Tseng and Wong (2005) which is described as Algorithm 2 below. It requires $\mathbf{X}$, a set of N p-dimensional observations. The desired number of clusters in the partition is given by K. The number of partitions performed is given by B, a parameter chosen by the researcher. B must be large enough to allow variation of the partitions, but not so large as not to be practical. The fraction of observations in each subset of the data is given by $g, 0 < g \leq 1$. The number of partitions in which a pair of observations must occur in the same cluster is given by $B(1-\alpha)$. The authors settled on values for α, B, g appropriate for their application through experimentation.

There are some challenges associated with Algorithm 2. The simplest is that it doesn't specify in Step 7 whether set selection is with or without replacement. It therefore allows clusters which may contain non-unique members. Because Tseng adds an additional step that determines stability across candidate values for K, this distinction may not be necessary. More importantly, calculating $\bar{\mathbf{T}}$ becomes

[3] 'Tightness' and 'stability' are terms used by the authors, yet are unfortunately not defined by them.

Algorithm 2 Tseng and Wong *Algortihm A*

Require: Observations $\mathbf{X} : \mathbf{x}_i \in \mathbb{R}^p \; \forall i = 1, \ldots, N$; $K, B \in \mathbb{Z}$; $0 \leq g, \alpha \leq 1$

1: **for** $b = 1, \ldots, B$ **do**:
2: Take a random sample $X^{(b)} \in \mathbf{X}$, where $X^{(b)}$ has $g|\mathbf{X}|$ members
3: Partition $X^{(b)}$ using KMC into K clusters
4: Let $\mathbf{T}^{(b)}$ be the connectivity matrix for the cluster
5: **end for**
6: Let $\bar{\mathbf{T}} = \frac{1}{B} \sum_{b=1}^{B} \mathbf{T}^{(b)}$
7: Search for a set of observations $P_1 = \{\mathbf{x}_i\} \subset \mathbf{X}$ such that $\bar{\mathbf{T}}_{\mathbf{x}_i \mathbf{x}_j} \geq 1 - \alpha \; \forall \mathbf{x}_i, \mathbf{x}_j \in P_1$. Repeat to find $P_2, P_3, \ldots$ until no further sets can be found.

difficult as the number of observations in $\mathbf{X}$ increases. If $N = 10^5$, then $\bar{\mathbf{T}}$ contains 10^{10} elements, each of which is a floating-point value. As such, $\bar{\mathbf{T}}$ would require 4GB of RAM alone, and its calculation needs still more memory.

K-means clustering has been exhaustively studied, yet work on determining cluster stability is less well explored than other aspects of the method. Most stability studies have focused on determining if complete partitions of datasets are stable. Fewer studies have focused on the relative stability of individual clusters within the partitions. That is, few researchers have considered the value of a subset of clusters within a partition that are stable even if the remaining ones are less so. Perhaps the best of these is Tseng and Wong (2005), but it doesn't address the computational difficulty of finding relationships among observations that hold across multiple partitions of a dataset.

2.3 Stability enhanced K-means clustering through pair counting

2.3.1 Algorithm development: Unordered pair set generation

Optimal clustering is the unique solution to Equation 2.3 and results in sets whose members are fixed. As discussed in Section 2.2, KMC approximates the optimum partition, and the clusters it generates can change with different initial conditions for the algorithm. The goal of this study is to find clusters whose memberships are stable while leveraging the computational efficiency of KMC. The strategy of the proposed method is as follows:

1. Create a set of bootstrap samples from the dataset
2. Conduct KMC on each of the bootstraps
3. For each bootstrap, create a list of pairs, each of which contains two observations that occur in the same cluster.
4. Combine the pair lists from the bootstraps into one set and count the number of times each pair occurs in the combined list. Divide the count by the number of bootstraps to create a stability index for each pair.
5. Select a threshold value for the stability index and create new clusters from sets of pairs above the threshold.

This method is very similar to that of Tseng and Wong (2005). However, as pointed out in Section 2.2, that algorithm requires the generation and storage of a

connectivity matrix $\mathbf{T}$, which will necessarily be large even if it is sparse. To generate B partitions, one would have to store B connectivity matrices $\mathbf{T}^{(1)}, \ldots, \mathbf{T}^{(B)}$, a total of BN^2 data elements. A simple modification to the algorithm is to only store the necessary elements of $\mathbf{T}$, those above the diagonal, which number $BN(N-1)/2$. A dataset of observations from NSDUH 2016-2019 corresponding to 18+ year-olds contains 170,944 rows. Twenty partitions of it requires storage of $(20)(170{,}944)(170{,}943)/2 = 292{,}216{,}801{,}920$ elements, a massive amount of data.

Each element of $\mathbf{T}$ represents the relationship between a pair of observations. If observation $\mathbf{x}_i$ and $\mathbf{x}_j$ are allocated to the same cluster, their relationship can be represented as an unordered pair[4] of the form $(\mathbf{x}_i, \mathbf{x}_j)$. If $\mathbf{x}_i$ and $\mathbf{x}_j$ are in different clusters, the pair $(\mathbf{x}_i, \mathbf{x}_j)$ does not exist. So for every cluster C_k, there is a list of pairs $[(\mathbf{x}_i, \mathbf{x}_j)] : \mathbf{x}_i, \mathbf{x}_j \in C_k$. In the case where all observations are determined to be in the same cluster, the list size is equal to the number of elements in the upper triangle of $\mathbf{T}$ minus the diagonal, and no benefit over Tseng's method is gained. However, in the case where there are K clusters of equal size N/K, the total number of pairs becomes $B(N/K)(N/K-1)/2$. Eleven clusters of equal size from the dataset would thus have $(20)(15{,}540)(15{,}539)/2 = 2{,}414{,}760{,}600$ pairs. For large N, the potential ratio of minimum to maximum number of pairs is:

$$\frac{B(N/K)(N/K-1)/2}{BN(N-1)/2} \approx \frac{1}{K^2}$$

This study proposes a modification to Tseng's algorithm that uses pair lists instead of connectivity matrices to create stable clusters. Steps 1-4 are preserved from Algorithm 2, and 5-7 are introduced to reduce method complexity:

[4]Hereafter the term 'pair' will mean 'unordered pair' for conciseness.

1. Let the total set of N observations be $\mathbf{X}$, and each individual observation be $\mathbf{x}_i, 0 \leq i \leq N$.

2. Let B be the number of bootstrap samples from $\mathbf{X}$ to be partitioned

3. Let $X^{(b)}, 1 \leq b \leq B$ be a bootstrap of $\mathbf{X}$ with $gN, 0 \leq g \leq 1$ members.

4. Let K be the number of clusters into which $X^{(b)}$ will be partitioned.

5. Let $S^{(b)}$ indicate a specific KMC partition of $\mathbf{X}^{(b)}$, and let $C_k^{(b)}$ indicate the k^{th} cluster of S^b.

6. Let $\mathbf{l}^{(b)}$ be the labels that indicate the clusters to which the observations $\mathbf{X}$ are assigned in partition $S^{(b)}$, and $\mathbf{L}$ be the set of all labels from B partitions. Each element of $\mathbf{l}^{(b)}$ takes the form $l_i^{(b)} = k$, where $\mathbf{x}_i \in C_k^{(b)}$

7. Let $u_{ij} = (\mathbf{x}_i, \mathbf{x}_j)$ be a pair representing the membership of $\mathbf{x}_i$ and $\mathbf{x}_j$ in the same cluster k, and let $\mathbf{U}$ be the list of all pairs u.

Assume that an appropriate value for K is determined using the elbow method as described in Section 4.2. The complete *Stability Enhanced KMC Algorithm* is then given by Algorithm 3.

The sets $C_1, C_2, \ldots$ are candidates for stable clusters. Algorithm 3 is easier to implement then Tseng's method because it preserves memory. The use of label sets rather than connectivity matrices in the clustering steps 1-6 only requires storage of BN rather than BN^2 elements. Generation of the pair list $\mathbf{U}$ remains computationally expensive because it involves a nested loop with $N(N-1)/2$ comparisons. Storing $\mathbf{U}$ is also expensive because its size is $B(N/K)(N/K - 1)/2 \leq |\mathbf{U}| \leq BN(N-1)/2$. Likewise, creation and storage of $\mathbf{U}'$ is expensive, and its ultimate size is $(N/K)(N/K-1)/2 \leq |\mathbf{U}'| \leq N(N-1)/2$. Nevertheless, Algorithm 3 offers significant savings over Tseng's.

Algorithm 3 Stability Enhanced KMC Algorithm

Require: Observations $\mathbf{X} : \mathbf{x}_i \in \mathbb{R}^p, i = 1, \ldots, N$; $K, B \in \mathbb{Z}$; $0 \leq g, \alpha \leq 1$; $\mathbf{L} = \mathbf{U}' = \emptyset$; $\mathbf{U}$ is an empty list

for $b = 1, \ldots, B$ **do**:
    With replacement, create subset $X^{(b)} \subset \mathbf{X}, |X^{(b)}| \approx gN$
    Using K-means clustering, partition $X^{(b)}$ into K clusters with centroids $\{\mu\} = \{\mu_1, \mu_2, \ldots, \mu_K\}$
    Assign to each observation $\mathbf{x}_i \in \mathbf{X}$ label $l_i^{(b)} = d$ such that μ_d is the closest centroid to $\mathbf{x}_i$.
    $\mathbf{L} = \mathbf{L} \cup \mathbf{l^{(b)}}$
end for
for $b = 1, \ldots, B$ **do**:
    for $i, j = 1, \ldots, N$ **do**:
        if $l_i^{(b)} = l_j^{(b)}$ and $u_{ij}, u_{ji} \notin \mathbf{U}$ **then**:
            $\mathbf{U} = \mathbf{U} \cup \{u_{ij}\}$
        end if
    end for
end for
for each distinct $u_{ij} \in \mathbf{U}$ **do**:
    Let c be equal to the number of instances of u_{ij} in $\mathbf{U}$
    $u'_{ij} = (\mathbf{x}_i, \mathbf{x}_j : c)$
    $\mathbf{U}' = \mathbf{U}' \cup \{u'_{ij}\}$
end for
Search for a set of points $C_1 \subset \mathbf{X}$ such that $\forall \mathbf{x}_i, \mathbf{x}_j \in C_1, \exists u'_{\mathbf{x}_i\mathbf{x}_j} = (\mathbf{x}_i, \mathbf{x}_j : c), c \geq \alpha B$. Repeat to find $C_2 \subset \mathbf{X} - C_1$, $C_3 \subset \mathbf{X} - C_1 - C_2, \ldots$ until no further sets can be found.

2.3.2 Algorithm development: Cluster synthesis

Steps 1-6 in Algorithm 2 and steps 1-23 in Algorithm 3 both generate the set of all pairs from the B partitions. The creation of clusters from the pair list is achieved in step 7 in Algorithm 2, which is structurally the same as step 24 in Algorithm 3. This task is not insignificant as it requires the creation of sets from millions of pairs. Network science affords an elegant way to model this problem.

Consider an undirected graph G whose vertices are the N observations in $\mathbf{X}$. The edges among vertices are equivalent to the pairs in $\mathbf{U}'$. A weighted edge between two vertices $\mathbf{x}_i, \mathbf{x}_j$ with weight c is given by $u'_{i,j} = (\mathbf{i}, \mathbf{j} : c)$. Create a subgraph $G_{(c)}$ by limiting the edges allowed in the graph to those whose weight is equal to or greater than a given value. A graph including edges that occur in all partitions of the data is obtained by setting $c = B$ to create $G_{(c)} = G_{(B)}$. As an example, consider a six node graph G, with nodes 0, 1, 2, 3, 4 and 5. This graph is depicted in Figure 2.1. Subgraph $G_{(15)}$ contains nodes connected with edge weights of at least 15, nodes 0, 2 and 3. $G_{(15)}$ contains only one connected component, and it is a *completely* connected component because each node can reach every other node directly.

A connected component C within graph G is a subgraph of G such that each pair of vertices $(x_i, x_j) \in C$ is connected by a path. A connected component of subgraph $G_{(c)}$ contains only vertices joined by paths whose weights are at least c. As c decreases, the number of edges in $G_{(c)}$ increases, as does the average size of its connected components, while their number decreases. *In this graph model, a connected component of* $G_{(c)}$ *is equivalent to a subset of* $\mathbf{X}$ *whose members have been found to be in the same cluster at least* c *times.* Determining the clusters in $\mathbf{U}'$ for any given stability becomes a matter of selecting the parameter c and

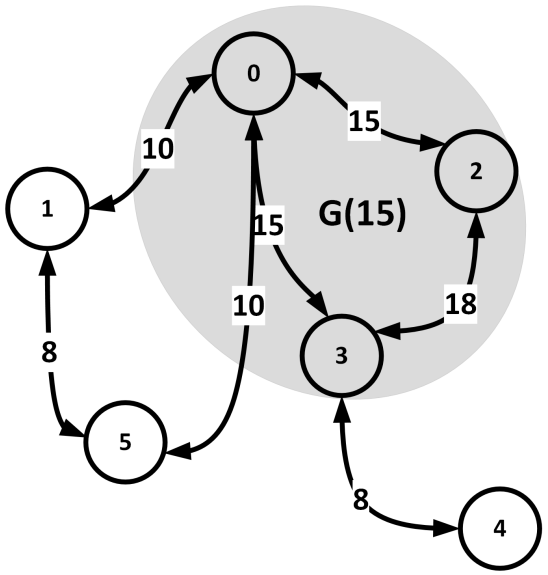


Figure 2.1: Subgraph $G_{(15)}$ within G

finding the connected components of $G_{(c)}$. This is a well understood problem that is solved using a *breadth first search* algorithm, which has complexity of $O(V + E)$ where V is the number of vertices in a graph, and E is the number of edges. In this case, $V = N$, and $E = |\mathbf{U}'|$, which for NSDUH data results in a large graph and BFS connected component discovery has a complexity of $O(N + |U'|)$. Fortunately, the network model of a partition as represented by $\mathbf{U}'$ from Algorithm 3 consists only of components that are complete. This permits great simplification of connected components discovery via an $O(N)$ algorithm. The derivation of this algorithm is presented in Appendix A.

In Algorithm 3, the parameter α is the level of stability required for pairs to be allocated to clusters. The level of stability α is a hyperparameter for the method, and it can be determined similarly to how K is chosen for KMC – heuristically. Generate all B subgraphs $G_{(c)}, 1 \leq c \leq B$ and measure how many connected components exist for each subgraph. The number of connected components is plotted versus stability and examination for an elbow in the curve is conducted.

The stability value $\frac{c}{B}$ at which this elbow occurs becomes the selection for α.

Chapter 3

Data preparation and implementation

Survey data regarding patterns of drug use is ingested and transformed to accommodate the methods described above. To deal with the complexity and compute requirements associated with the NSDUH dataset, specific cluster computing techniques are leveraged during algorithm implementation.

3.1 Ingestion

This study uses NSDUH survey data from 2016-2019 respondents whose age is 18 or over. This age restriction emphasizes individuals who have had enough time to initiate the use of multiple drugs. The number of resulting observations becomes 170,944. Only features that are directly related to AFU and demographic information are included. The complete list of NSDUH fields used in this study is enumerated in the Table 3.1. The prefix 'IR' is used by SAMHSA to represent the term 'imputed-revised'. For these variables, SAMHSA replaced missing values

with imputed values. The imputation-revised variable's value arises from one of three methods: interview responses (no imputation), logical assignment in the editing process, or statistical imputation. Because SAMHSA recommends the use of imputed variables wherever possible, they are included in this study instead of the direct responses in the survey.

NSDUH Field	Category	Description
RESPID	Identification	Respondent ID
IRCIGAGE	AFU	Cigarettes
IRCGRAGE	AFU	Cigars
IRSMKLSSTRY	AFU	Smokeless tobacco
IRALCAGE	AFU	Alcohol
IRMJAGE	AFU	Marijuana
IRCOCAGE	AFU	Cocaine
IRCRKAGE	AFU	Crack
IRHERAGE	AFU	Heroin
IRHALLUCAGE	AFU	Hallucinogens
IRINHALAGE	AFU	Inhalants
IRMETHAMAGE	AFU	Methamphetamine
IRPNRNMAGE	AFU	Pain relievers
IRTRQNMAGE	AFU	Tranquilizers
IRSTMNMAGE	AFU	Stimulants
IRSEDNMAGE	AFU	Sedatives
AGE2	Demographics	Age
SERVICE	Demographics	Military service
CATAG6	Demographics	Age category
IRSEX	Demographics	Gender
IRMARIT	Demographics	Marital status
NEWRACE2	Demographics	Race
EDUHIGHCAT	Demographics	Education
IRWRKSTAT	Demographics	Work status
GOVTPROG	Demographics	Govt assistance
INCOME	Demographics	Income
COUTYP4	Demographics	County type
AIIND102	Demographics	Native American area
ANALWT_C	Demographics	Person-level sample weight

Table 3.1: NSDUH Variables Included in the Study

The variable ANALWT_C is a metric constructed by SAMHSA to represent the weight of the observation. This feature can be interpreted as the total number of individuals in the target population represented by the observation. The sum of all the weights in the survey equals the size of the population targeted by NSDUH. ANALWT_C must be used to weight analyses using NSDUH to create unbiased estimates for survey outcomes. Because the dataset considered includes four years of survey information, ANALWT_C is divided by four to create YRWEIGHT, a factor appropriate to the annual population size.

The goal of this study is to consider drug use progression among categories similar to those explored by prior researchers. NSDUH gathers AFU data for a select category of drugs, some of which overlap.[1] Data preparation is begun by including the non-overlapping AFU data for: cigarettes, cigars, smokeless tobacco, alcohol, marijuana, hallucinogens, cocaine, crack, heroin, pain relievers, tranquilizers, stimulants, sedatives, methamphetamine, and inhalants.[2] For prescription drugs, AFU is defined as the first time a respondent uses the substance without instructions to do so from a doctor.[3] Unfortunately, NSDUH is not consistent in how it measures AFU across all drug categories. For most drugs the AFU is obtained directly. For example, the question for marijuana use is, 'How old were you the first time you used marijuana or hashish?.' For pain relievers, tranquilizers, stimulants, and sedatives, AFU data is only obtained in this way if the respondent used a drug from the category within the past 12 months (Center for Behavioral Health Statistics and Quality, 2020a, p. 173). This creates a significant number

[1] NSDUH gathers AFU data as integer values.

[2] The drugs included in the categories hallucinogens, pain relievers, tranquilizers, stimulants, sedatives and inhalants are listed in Center for Behavioral Health Statistics and Quality (2020b, p. 36-91).

[3] To be consistent, we simplify our description of improper use of prescription medicine as *use* throughout this study.

of missing values for these categories (Table 3.2).

Table 3.2: Missing AFU Values by Category

Category	Never Used	Used/AFU Known	Used/AFU Unknown
Pain relievers	150,839	1,458	18,647
Tranquilizers	161,894	1,082	7,968
Stimulants	160,689	1,084	9,171
Sedatives	166,752	195	3,997

To create a complete view of each respondent's drug initiation sequence, the missing values for these drugs must be replaced with imputed data. Troyanskaya et al. (2001) describe a method to impute missing value using a K-nearest-neighbors approach. To calculate a value for a missing feature g of observation $\mathbf{x}_i$, the mean of feature g from a fixed number F observations closest to $\mathbf{x}_i$ is taken. Closeness is defined in terms of the Euclidean distance between $\mathbf{x}_i$ and another observation $\mathbf{x}_j$. The distance is calculated as function of all features other than g from the two vectors. F is selected by the researcher and is a tuning parameter of the method. In this study, the $F = 5$ nearest neighbors of an observation are used for missing value imputation. To reduce the complexity of this study, tuning of this parameter was not performed.

Cigarettes, cigars, and smokeless tobacco are combined into one category, *tobacco*, represented by the AFU variable IRTOBAGE. Pain relievers, tranquilizers, stimulants, and sedatives are combined into one representing prescription drugs with the AFU variable IRSCRIPAGE. Cocaine and crack cocaine are also combined into one category with the AFU variable IRCOC2AGE. IRTOBAGE, IRSCRIPAGE and IRCOC2AGE are set as the minima of the sets of AFUs from which they are drawn:

$$\text{IRTOBAGE} = min(\text{IRCIGAGE, IRCGRAGE, IRSMKLSSTRY})$$

$$\text{IRSCRIPAGE} = min(\text{IRPNRNMAGE, IRTRQNMAGE, IRSTMNMAGE, IRSEDNMAGE})$$

$$\text{IRCOC2AGE} = min(\text{IRCOCAGE, IRCRKAGE})$$

With this consolidation, the final variables considered in this study are shown in Table 3.3.

NSDUH Field	Category	Description
RESPID	Identification	Respondent ID
IRTOBAGE	AFU	Tobacco
IRALCAGE	AFU	Alcohol
IRMJAGE	AFU	Marijuana
IRCOC2AGE	AFU	Cocaine
IRSCRIPAGE	AFU	Prescriptions
IRHALLUCAGE	AFU	Hallucinogens
IRINHALAGE	AFU	Inhalants
IRHERAGE	AFU	Heroin
IRMETHAMAGE	AFU	Methamphetamine
AGE2	Demographics	Age
SERVICE	Demographics	Military service
CATAG6	Demographics	Age category
IRSEX	Demographics	Gender
IRMARIT	Demographics	Marital status
NEWRACE2	Demographics	Race
EDUHIGHCAT	Demographics	Education
IRWRKSTAT	Demographics	Work status
GOVTPROG	Demographics	Govt assistance
INCOME	Demographics	Income
COUTYP4	Demographics	County type
AIIND102	Demographics	Native American area
YRWEIGHT	Demographics	Person-level sample weight

Table 3.3: Combined and Imputed NSDUH Variables

To cluster the AFU data a distance between observations must be defined. This in turn requires that the AFU fields be put into a data structure representing a point in multi-dimensional space. The combination of AFU variables is represented by a nine-dimension vector. Each position in the vector represents a drug (or drug category). For example, the first dimension of the vector is mapped to tobacco use. The value of the first dimension is the age in years at which the respondent first used the drug. The valid range of values is the set of integers from 0 through the maximum reported AFU by any respondent. If the respondent did *not* use a given substance, the value 991 is put in the corresponding position in the vector. The value 991 is chosen for two reasons:

- Never using a drug can be thought of as having an AFU equal to infinity. So a large value must represent no use. 991 is significantly higher than the maximum lifespan of a human.
- The raw data in NSDUH uses 991 for 'Never Used' in AFU values, so that convention is preserved.

To illustrate, consider a respondent who first used tobacco at age 15, alcohol at age 16, marijuana at age 20, and no other substances. The vector representing his usage progression would be

$$\mathbf{v}_{AFU} = [15, 16, 20, 991, 991, 991, 991, 991, 991]$$

3.2 Network data model

Efficient computation in the study requires use of alternative data structures. To create an appropriate data model, drug use progression is represented as traver-

sal across a network. The drug initiation sequence of the hypothetical subject described above can be represented as a path:

$$\text{start} \rightarrow 15 \text{ yrs} \rightarrow \text{tobacco} \rightarrow 1 \text{ yrs} \rightarrow \text{alcohol} \rightarrow 4 \text{ yrs} \rightarrow \text{marijuana}$$

This path can be thought of as a traversal through a network where each node is the use of a drug (plus one *start* node). The edges of this ten node network represent the progression between nodes. The node labels for the network representation of the drug initiation sequence are shown in Table 3.4. The time duration between nodes is represented by weights assigned to the edges. The

Table 3.4: Drug Use Progression Network

Node index	Drug
0	Start
1	Tobacco
2	Alcohol
3	Marijuana
4	Cocaine
5	Prescription drugs
6	Hallucinogens
7	Inhalants
8	Heroin
9	Methamphetamine

example respondent traverses this network as follows: initiate at the start node, travel along an edge with weight 15 to tobacco, travel next along an edge with weight 1 to alcohol, then along an edge with weight 4 to marijuana, and no further travel occurs (Figure 3.1).

A path can be represented by a 10x10 connectivity matrix $\mathbf{G}$. A non-zero element G_{ij} represents the time in years from the first use of drug i to the first use of drug j. Path progression requires that the AFU for drug j must be greater

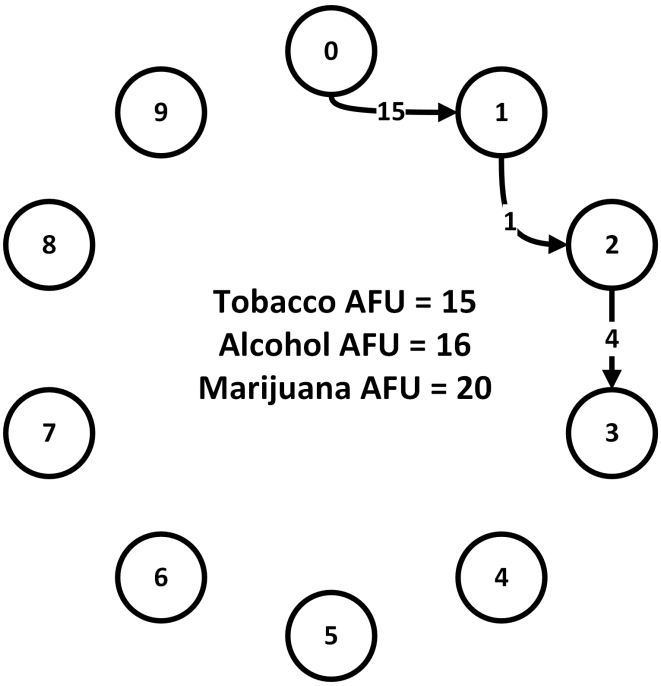


Figure 3.1: Network Model of an Example Drug Initiation Sequence

than that of drug i. Because NSDUH captures responses as integers, it is not uncommon for two drugs to have the same AFU value. To remove this situation, a random number between 0 and 1 is added to the AFU values, and the elements of $\mathbf{G}$ become real numbers. When $G_{ij} = 0$, there is no one-step progression from i to j. Note that drug use may proceed from i to j along an alternate path if $G_{ik} \neq 0$ and $G_{kj} \neq 0$. The connectivity matrix for the example respondent would be:

$$\mathbf{G} = \begin{pmatrix} 0.0 & \mathbf{15.1} & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & \mathbf{1.4} & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & \mathbf{4.2} & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{pmatrix}$$

In most cases, **G** will be sparse, making the storage in memory of 100 real numbers inefficient. An edgelist is a better way of representing a sparse graph. Each edge takes the form $e = (i, j : y)$, where i is the last drug initiated, j is the next drug initiated, and y is the number of years between AFU of i and j. The example respondent's progression is represented by the edgelist:

$$\mathbf{e} = [(0, 1 : 15.1), (1, 2 : 1.4), (2, 3 : 4.2)]$$

In this case the edgelist **e** requires the storage of only six integers and three real numbers, a great savings in memory over **G**. Network algorithm calculations become correspondingly more efficient as they are done via loops through the short edgelist rather than through matrix multiplication.

When presenting results, the names of the drugs replace their corresponding indices. This allows the reader to easily recognize a path. Using this convention, the example edgelist becomes:

$$\mathbf{e} = [(\text{start}, \text{tobacco} : 15.1), (\text{tobacco}, \text{alcohol} : 1.4), (\text{alcohol}, \text{marijuana} : 4.2)]$$

Each of these ways of representing drug use progression is used in this study. The AFU vector is used in generating clusters of common use patterns. The

edgelist notations are used to facilitate initial exploration of the data and in pair construction for stable cluster generation.

3.3 Implementation

Despite its improvements, the application of Algorithm 3 to the entire dataset of 170,944 observations is expensive. Steps 1-6 of Algorithm 3 are inexpensive and easy to perform using the entire dataset. Steps 7-23, the steps involving pair creation and summation, are memory and compute time intensive. Therefore, all 170,944 observations are used to generate the label sets, and a 30% random sample of the labeled observations is used for generation of the weighted pair set $\mathbf{U}'$. Clusters $C_1, C_2, \ldots, C_K$ are determined from $\mathbf{U}'$ via the connected component method described in Section 2.3.2. Each of the remaining 70% of the observations are then assigned to the cluster whose center is closest. This labeled dataset, appropriately weighted by the NSDUH survey weights, is then used for analysis. This process is illustrated in Figure 3.2.

The KMC hyperparameters are:

- Maximum iterations: 1,000
- Number of KMC starts per partition: 20
- Number of clusters[4] K: 11

Generation of $\mathbf{U}'$ presents two challenges. The first is that $\mathbf{U}'$ will have between 10,865,003 and 1,314,957,660 pairs (depending on cluster sizes), and assuming that each pair requires 16 bytes of RAM for two long integers, the list will take between 174MB and 21GB of RAM. The second is that identifying the pairs is a nested

[4]Determined using the elbow method on an initial partition of the entire dataset in Section 4.2

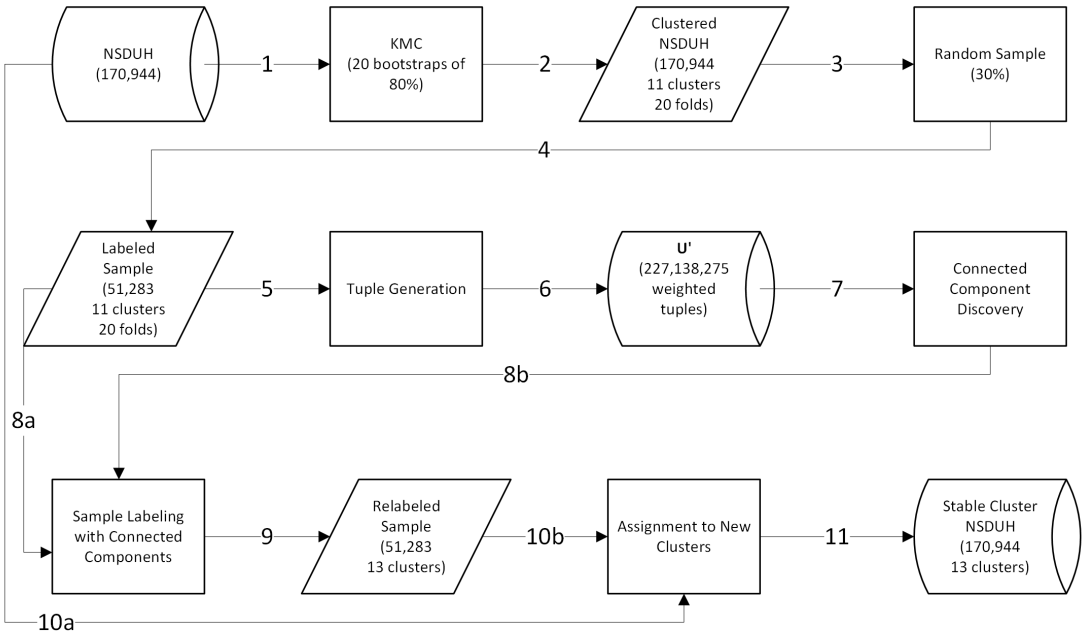


Figure 3.2: Stable Clustering Data Flow

loop, which is $O(N^2)$ complexity and requires 1,314,957,660 comparisons. The combination of memory and processing demands is infeasible on a desktop due to memory constraints. To process the data, a high-memory scalable Spark cluster with one driver node and a maximum of 8 worker nodes is provisioned. Each node has 60GB of RAM and 8 CPUs with clock speed of 2,294 MHz. The cluster is provisioned using Databricks, a cloud-based data warehouse solution running in an Azure cloud environment.

Choice of data structures and coding methodology greatly affects the performance of the routine. A first attempt at implementation of Algorithm 3 mixed use of Spark dataframes and Python lists. $\mathbf{L}$ was read in from a CSV file into a dataframe. The dataframe was queried to extract the observations for each combination of partition b and cluster label k. Next the observation labels (RESPID) were placed into a list which was passed into the nested loop (Algorithm 3 steps

8-17) to create a list of pairs. This list was appended to a master list, **U**, which was subsequently converted back to a Spark dataframe for further processing and storage. The code for this method is shown in Listing 3.1.

```
for b in range(0,B):
# Initiate tuplelist
tuplelist = []

# Populate tuplelist
clustset = 'labels_' + str(b)
for c in range(0,n_clusters):
        clustslice = dfclust[dfclust[clustset]==c]
        clustlist = clustslice.select('RESPID').rdd.
           flatMap(lambda x: x)\
                        .collect()
for i in range(0,len(clustlist)):
        for j in range(i+1,len(clustlist)):
        if clustlist[i] < clustlist[j]:
                tuplelist.append((clustlist[i],
                   clustlist[j]))
        else:
                tuplelist.append((clustlist[j],
                   clustlist[i]))

# Convert tuplelist to dataframe and insert into
   permanent table
columns = ['orignode', 'termnode']
df3 = spark.createDataFrame(tuplelist, columns)
df3.registerTempTable('tupletbl')
spark.sql("""
        INSERT INTO abuse_sequence.tuplecounts
        SELECT DISTINCT orignode, termnode, count(*) as
            tuplecount
        FROM tupletbl
        GROUP BY orignode, termnode
""")
```

Listing 3.1: Tuple Creation by Nested Loop

This method proved to be an inefficient means of executing the algorithm. Using brute iteration and lists (Algorithm 3 steps 7-18) omits the advantage of distributed processing made possible by Spark. In fact, those steps took 38,930 seconds to complete, used a maximum 60GB of RAM, and did not leverage the multitude of processors available in the clusters. To fix this problem, the data was left in dataframes and Spark SQL was used to generate the pairs. Replacing

the nested loop in Listing 3.1 with a dataframe JOIN causes Spark to conduct a broadcast nested loop join across the cluster. The clustering results **L** are again read into a dataframe. **L** is then queried to create small dataframes for each combination of b and k. These dataframes have only one column – the respondent identifier RESPID. Each single column dataframe is next joined to itself on unequal matches of RESPID to create a new dataframe with two columns of respondent identifiers. Each row of this dataframe represents a pair with the same cluster label from a partition. All of the resulting dataframes are concatenated to create one large dataframe with all of the pairs from the B partitions. Algorithm3 steps 19-22 are completed by performing a simple Spark SQL selection on this dataframe. The code for this revised approach is shown in Listing 3.2. This transformation resulted in a wall clock improvement in which construction of the large pair dataframe took only 2,628 seconds, a savings of 93%[5]

[5]Unfortunately, wall clock savings do not translate into reduced compute costs. The total amount of CPU time used by the Spark join is the same as nested loop execution, it is just spread across the cluster. Nevertheless, wall clock savings are obviously valuable in ways other than compute costs.

```
for b in range(0,B):
# Initiate tuplelist
clustset = 'labels_' + str(b)

# Generate pairs via a Spark join
for c in range(0,n_clusters):
        dfslice = df2[df2[clustset]==c].select('RESPID'
            )
        dfslice.createOrReplaceTempView('tblslice1')
        dfslice.createOrReplaceTempView('tblslice2')
        dftupleset = spark.sql("""
        SELECT tblslice1.RESPID AS orignode, tblslice2.
            RESPID
        AS termnode
        FROM tblslice1 JOIN tblslice2
        WHERE tblslice1.RESPID < tblslice2.RESPID
        """)

# Insert set into persistent table
dftupleset.createOrReplaceTempView('tblinsertslice')
spark.sql("""
        INSERT INTO abuse_sequence.sparktuples
        SELECT DISTINCT orignode, termnode, count(*) as
            tuplecount
        FROM tblinsertslice
        GROUP BY orignode, termnode
""")
```

Listing 3.2: Tuple Creation by Spark Join

To generate the stabilized clusters from the pair lists, $\mathbf{U}'$ is stored in a *GraphFrames* data structure and its connected components are found (*GraphFrames*, 2021). GraphFrames is a software package that distributes graph processing algorithms across a Spark cluster. As described above, the $B = 20$ subgraphs $G_{(c)}, 1 \leq c \leq B$ are examined, and for each the number of connected components and the sizes of the largest components are found. Running this process for the 20 subgraphs on the Spark cluster took 2,442 seconds of wall clock time.

Chapter 4

Results

4.1 Data exploration

4.1.1 Demographics

The working dataset is comprised of the curated AFU and demographic features. It contains 170,944 observations and 23 variables and represents a sampled population of 247,716,947 individuals.[1] Basic distributions for the sampled population are shown in Figure 4.1. Males make up 48.3% and females 51.2% of the population, and the largest age cohort is 50-64 years of age followed by 35-49 years of age. Non-Hispanic whites is the largest ethnic cohort, followed by Hispanics and Non-Hispanic blacks.[2]. Individuals tend to be either married or never married, with smaller cohorts of divorced and widowed individuals. The largest total family income bracket is $75k+, and the second largest is $20k-$49999. Highest educational attainment is somewhat balanced among high-school graduates with some college and college graduates, with high-school graduates (no college) next most common.

[1] Throughout the results discussion, the statistics corresponding to the sampled population, the respondent counts multiplied by their corresponding weights, are presented.

[2] We adopt the ethnic naming conventions used by NSDUH in this study

There is a smaller cohort of individuals who did not graduate from high-school. NSDUH accounts for population density by mapping the respondents' counties of residence into three categories. Large metropolitan counties have the most respondents, followed by small metropolitan ones. Non-metropolitan counties contain the fewest individuals.

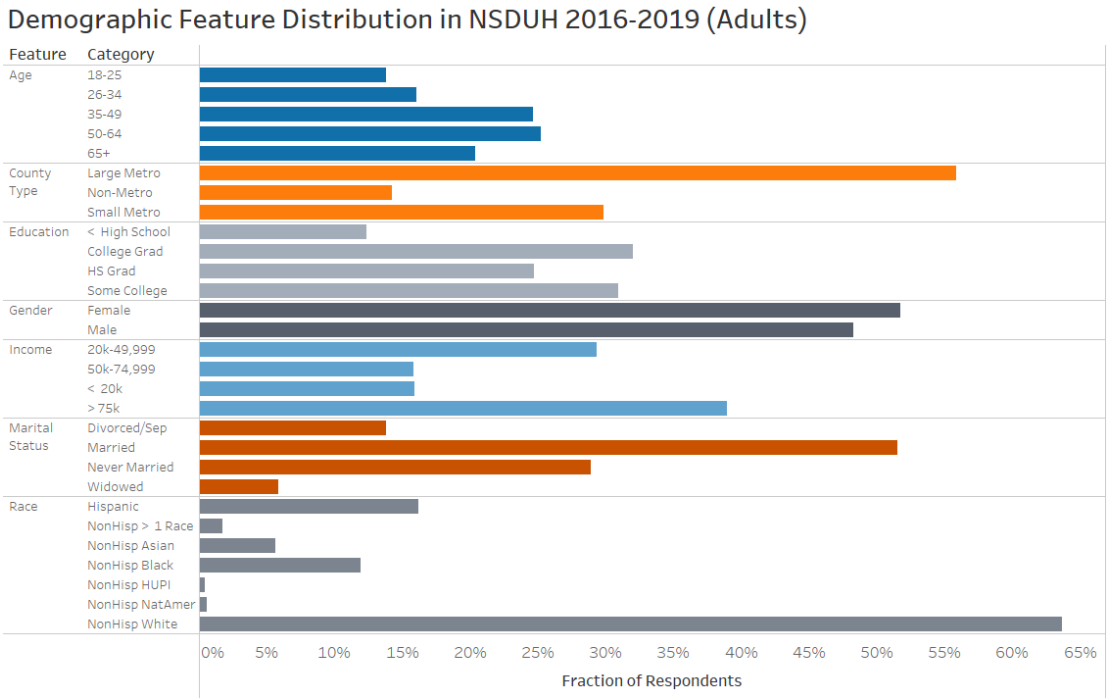


Figure 4.1: Demographic Distributions

4.1.2 Basic drug usage

The vast majority of the population has used at least one substance, with the long-time licit drugs, alcohol and tobacco, being the most common (Table 4.1). Marijuana, which is now legal for recreational use in many states, has been used by almost half of all individuals. Illicit drugs are less common, with hallucinogens most used among those. The least commonly used drug is heroin, with only 2.1% of the population indicating any use.

The average age of first use for most drugs is below 20 years, with tobacco having the earliest mean AFU at 17.4. Illicit drugs have higher AFUs, with prescription drugs having by far the highest at 35.6 years. Inhalants are a significant deviation from this pattern with an AFU of 17.9. Fraction of the population that has used a drug is plotted against its mean AFU in Figure 4.2.

Drug	**Fraction of Population**	**Mean AFU (years)**
Alcohol	85.9%	17.4
Tobacco	65.5%	16.4
Marijuana	48.2%	18.3
Hallucinogens	17.0%	19.6
Cocaine	16.2%	21.8
Prescriptions	13.4%	35.6
Inhalants	9.2%	17.9
Methamphetamine	6.1%	22.3
Heroin	2.1%	24.1

Table 4.1: Fraction of Population Used and Mean AFU by Drug

4.1.3 Drug use pathways

Drug initiation sequences without consideration of AFU are explored prior to expanding the analysis to include age. This enables the depiction of common progressions from drug to drug. Without age data, a drug use progression is modeled as a path through a simple graph with ten nodes, each one corresponding to a drug (plus one 'start' node), with unweighted edges. The edgelist notation is used to depict paths. Without weights, an edge takes the form $e_i = (\text{tobacco}, \text{alcohol})$, and a path is represented by an edgelist:

$$\mathbf{e} = [(\text{tobacco}, \text{alcohol}), (\text{alcohol}, \text{marijuana}), (\text{marijuana}, \text{cocaine})]$$

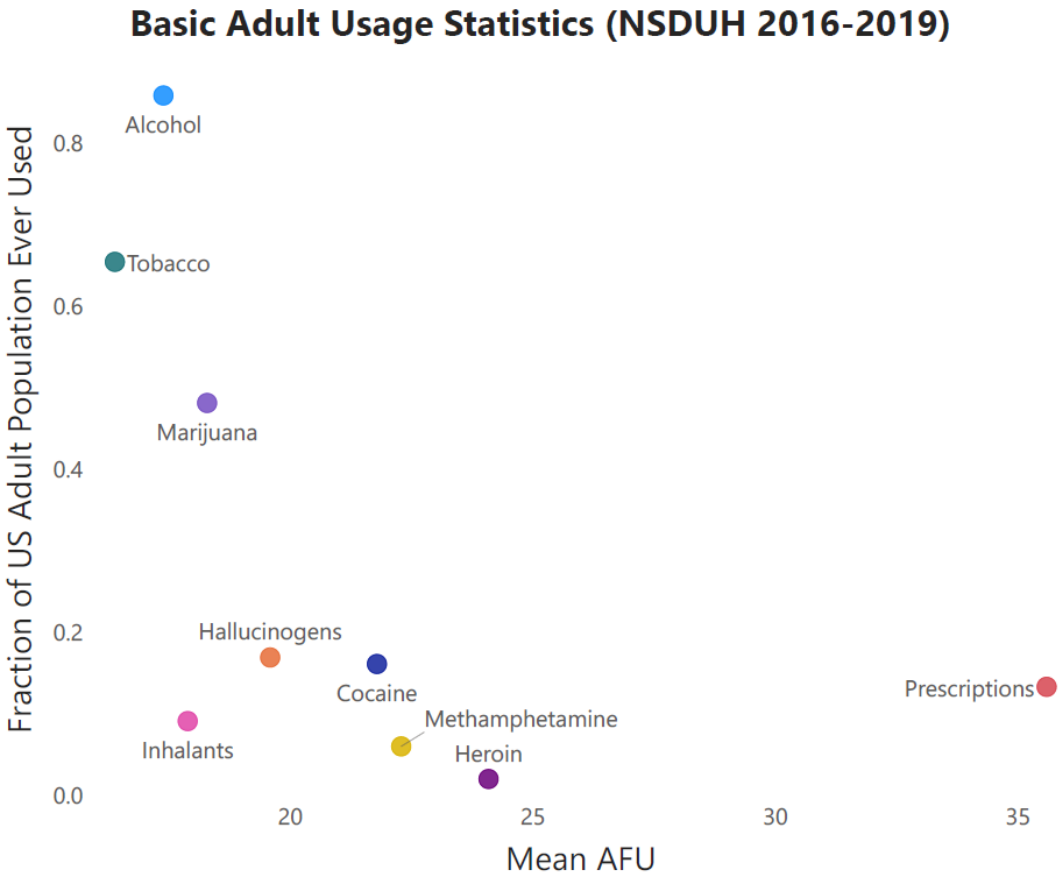


Figure 4.2: Basic Usage Statistics

For brevity, the path can be displayed simply as the drugs used in sequence:

[tobacco, alcohol, marijuana, cocaine]

There are a total of 247,716,947 paths through the graph, one for each member of the population represented by the survey. The fifteen most common paths, representing 71% of all paths, are shown in Table 4.2. With the exception of no-usage, the ten most common pathways all initiate with alcohol or tobacco. However, pathways 11-13 all initiate with marijuana, and those represent 3.7% of all pathways. These paths are not unexpected given marijuana legalization for both medical and recreational use.

The first completely illicit drug to occur in this list is cocaine. The second is improperly used prescriptions. This category was not highlighted in the literature, and it likely that this pattern is a recent effect of the opioid epidemic. Interestingly, the second most common pathway containing illicit drugs initiates with alcohol and jumps directly to prescriptions. According to D. B. Kandel and Yamaguchi (2002b), this would be considered a random deviation from a normative pattern. However, it could be that a jump from alcohol to prescription abuse has become a rare but non-random outcome, once again an effect of the opioid epidemic.

The pathways are portrayed as directed edges through the unweighted graph in Figure 4.3. The ten nodes associated with start and the nine drugs are shown in red. Edges are depicted in gray scale, with the darkness of the edge increasing with the frequency of the edge. Edges among tobacco, alcohol, and marijuana are most common, followed by edges connecting to prescriptions, hallucinogens, and cocaine. Edges to heroin are infrequent, but among those, the most common links are from cocaine and prescriptions.

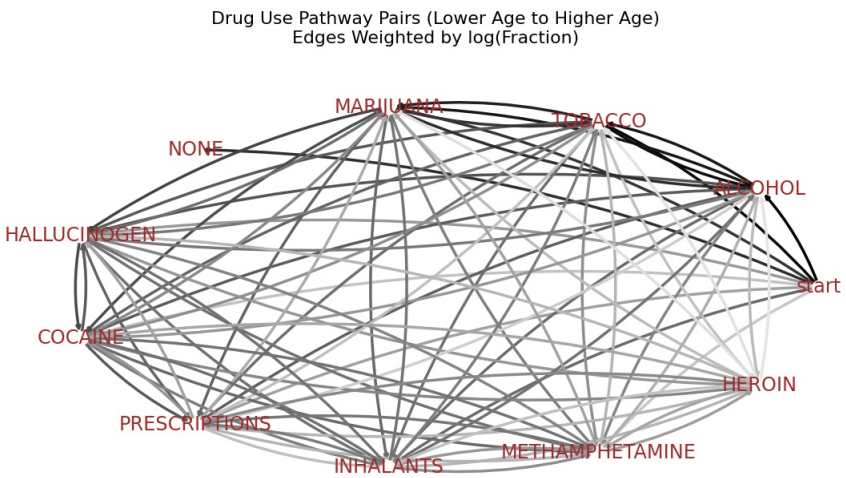


Figure 4.3: Drug Use Progression Graph

The normative pattern of first drug initiation with either alcohol or tobacco, followed by marijuana use, then by illicit drugs, dominates the data. There is a small set of pathways that likely are non-random and due to marijuana legalization

Path	**Fraction**
[start, alcohol]	16.2%
[start, tobacco, alcohol]	11.9%
[start, none]	10.3%
[start, alcohol, tobacco]	7.3%
[start, tobacco, alcohol, marijuana]	6.1%
[start, alcohol, tobacco, marijuana]	4.3%
[start, tobacco, marijuana, alcohol]	2.6%
[start, alcohol, marijuana]	2.4%
[start, tobacco]	2.3%
[start, alcohol, marijuana, tobacco]	2.3%
[start, marijuana, tobacco, alcohol]	1.4%
[start, marijuana, alcohol, tobacco]	1.3%
[start, marijuana, alcohol]	1.1%
[start, tobacco, alcohol, marijuana, cocaine]	0.7%
[start, alcohol, prescriptions]	0.6%

Table 4.2: Most Common Drug Use Pathways

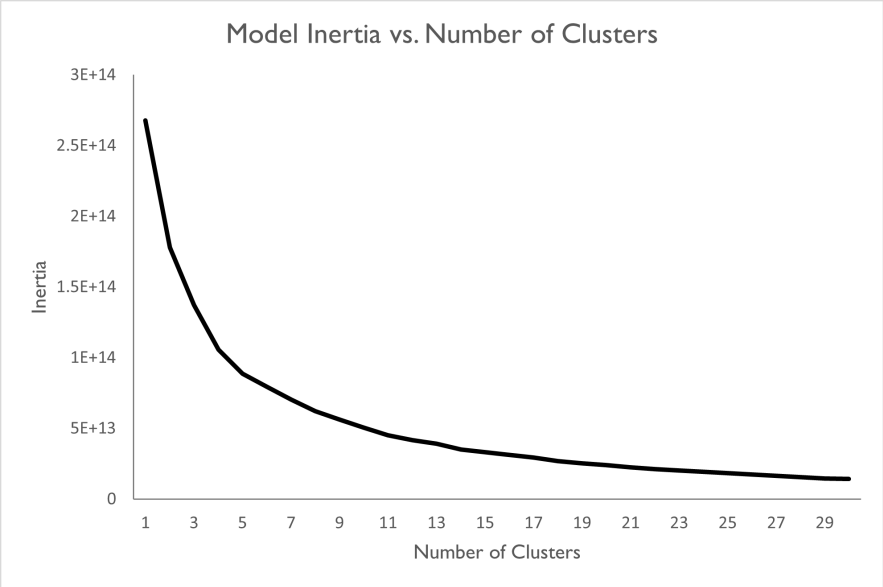
and the opioid epidemic. These patterns do not include AFU, which can add additional insight into drug initiation.

4.2 Initial clustering

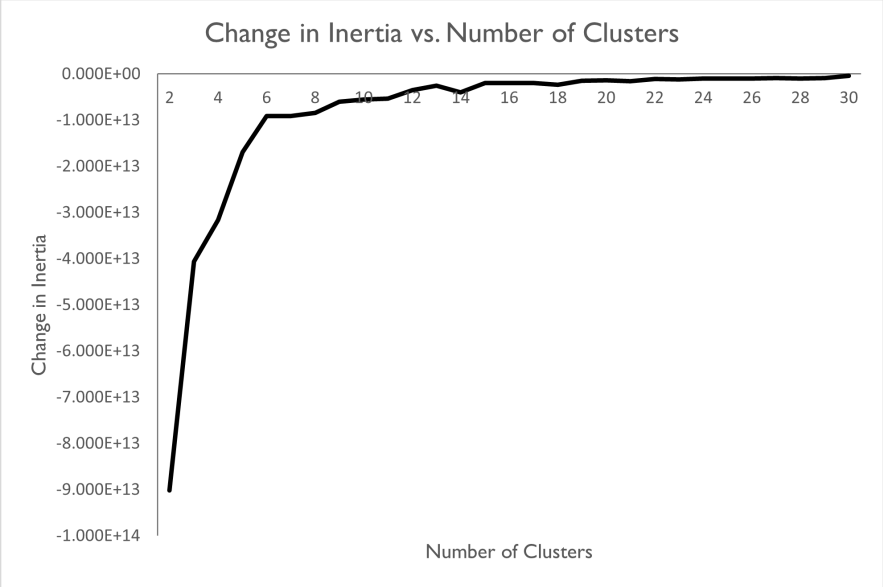
This study seeks to answer the question 'Are there identifiable patterns of progression of age of first use for drug categories?' This inquiry is addressed by clustering the NSDUH observations on the AFU vectors created in Section 3. To avoid overfitting the NSDUH data and to reduce variation in output clusters, stability methods are applied, but first a straightforward K-means cluster (KMC) analysis is performed on the data. Hyperparameters for KMC include the number of clusters, and the maximum iterations for the algorithm. Additionally, KMC is run multiple times, with random starting centroids, and the run with lowest inertia is chosen as the locally optimal clustering. Sci-Kit Learn (Pedregosa et al., 2011) is used to perform KMC with 1,000 iterations and 20 runs. To determine the best number of clusters, KMC is run multiple times with increasing numbers of clusters, $K = 2, 3, \ldots, 30$. Because the dataset is quite large, the elbow method is used to select K.

Figure 4.4(a) shows that model inertia decreases smoothly with an increasing number of clusters. The elbow at which change in inertia decreases abruptly is not apparent. A plot of the change in inertia versus the number of clusters (Figure 4.4(b)) provides better insight into the change in performance of the algorithm. The first point at which the change in inertia makes a significant reversal occurs at $K = 11$ clusters. Therefore this point is adopted as the best value for K.

Clusters can be described several different ways. One is to represent them



(a)



(b)

Figure 4.4: Model Inertia Change as a Function of K

in terms of their centroids (Table 4.3). The centroid, or geographic center, of each cluster is the mean of the nine drug AFU vectors across all of the cluster members. The centers include some well defined AFU values – those that are below 100. There are also some well defined no-use AFU values – those that are equal to or near the 'never used' value of 991. In between these two extremes are values over 100 but still less than 991. For example, in cluster **5**, the AFU center for Prescriptions is 369.0. Such data indicates the presence of observations whose majority never used prescriptions along with a smaller cohort of users who did.

The four largest clusters comprise over two thirds of the population and are similar to the most common individual paths found in Section 4.1.3. Cluster **0**, representing 21.0% of the population, is centered on early initiation of tobacco at 17.5, with a delay of one year before alcohol use at 18.4. Cluster **0** can be thought of as usage only of completely legal drugs initiated at an early age. Cluster **1**, with 19.4% of the population, is centered on tobacco initiation at 16.4 followed by alcohol at 16.7, and marijuana at 19.7. It is interesting to note that this cluster, which includes marijuana, involves earlier AFUs for tobacco and alcohol than Cluster **0**. Cluster **2** is centered on only the use of alcohol, and cluster **3** is centered on the absence of any drug use. These large clusters are in line with expectations of normative initiation sequences. However, several groups have very different patterns. Cluster **4**, representing 6.4% of the population, shows a pattern of multiple drug initiation around the age of 20, beginning with hallucinogens. Similarly, cluster **6** initiates with hallucinogens, followed by alcohol, marijuana, and tobacco. Cluster **7** represents 4.7% of the populations and initiates with marijuana at 20.4 followed by alcohol at 26.3 and then nothing else. A total of 25.7% of the population is represented by clusters that do not follow the tobacco $\rightarrow$ alcohol $\rightarrow$ marijuana $\rightarrow$ other drugs sequence.

It is important to point out that no respondent would likely have behavior exactly equal to a cluster center. Full interpretation of each cluster involves an analysis of the AFU distributions for each drug. Nevertheless, the *medoids* of the clusters can be identified. A medoid is the observation within a cluster that is closest to its geometric center. The medoid can be thought of as a 'typical' member of the cluster. Since a medoid likely only used a subset of the ten drugs, they can be described using pathway notation (Table 4.4). Most of the medoids are very similar to the cluster centroids.

	Cluster					
Drug	**0**	**1**	**2**	**3**	**4**	**5**
Tobacco	17.5	16.4	991.0	991.0	58.5	33.3
Alcohol	18.4	16.7	20.2	991.0	22.9	17.3
Marijuana	991.0	19.7	991.0	975.1	23.2	20.0
Cocaine	983.1	991.0	987.6	990.8	21.5	116.6
Prescriptions	945.5	991.0	951.4	966.4	609.1	369.0
Hallucinogens	984.0	991.0	985.3	988.7	19.8	56.0
Inhalants	973.0	933.6	972.6	977.8	991.0	18.3
Heroin	990.4	989.5	991.0	990.5	880.4	799.0
Methamphetamine	987.7	974.0	989.5	990.4	688.5	543.9
Fraction	21.0%	19.4%	17.3%	10.9%	6.4%	5.5%

	Cluster					
Drug	**6**	**7**	**8**	**9**	**10**	
Tobacco	31.2	991.0	31.7	17.4	26.8	
Alcohol	24.2	26.3	29.3	991.0	27.9	
Marijuana	28.5	20.4	31.9	850.9	36.8	
Cocaine	991.0	936.3	23.6	987.3	991.0	
Prescriptions	752.7	916.1	799.6	963.2	35.4	
Hallucinogens	20.4	911.5	991.0	983.0	991.0	
Inhalants	869.8	945.9	910.8	974.1	820.9	
Heroin	979.0	989.3	952.6	988.2	983.9	
Methamphetamine	929.9	982.4	839.8	983.4	958.5	
Fraction	4.7%	4.7%	4.4%	2.9%	2.8%	

Table 4.3: Basic KMC Geometric Centers

Cluster	Medoid Pathway
0	[(tobacco, 17), (alcohol, 18)]
1	[(tobacco, 16), (alcohol, 17), (marijuana, 20)]
2	[(alcohol, 20)]
3	[No use]
4	[(alcohol, 14), (marijuana, 15), (cocaine, 17), (hallucinogens, 17), (tobacco, 50)]
5	[(tobacco, 15), (alcohol, 17), (marijuana, 17), (cocaine, 18), (hallucinogens, 18), (inhalants, 18), (prescriptions, 65)]
6	[(alcohol, 19), (hallucinogens, 23), (tobacco, 29), (marijuana, 30)]
7	[(marijuana, 19), (alcohol, 26)]
8	[(alcohol, 26), (tobacco, 29), (marijuana, 30), (cocaine, 30)]
9	[(tobacco, 17)]
10	[(alcohol, 21), (tobacco, 25), (marijuana, 39), (prescriptions, 39)]

Table 4.4: Basic KMC Medoids

4.3 Identification of the Stable Clusters

As mentioned earlier, KMC is applied to multiple bootstrap samples from the complete dataset and pair generation is performed on a 30% sample of the labeled observations to form $\mathbf{U}'$. KMC is run on 80% folds of the data sample $B = 20$ times to create the label sets. Initial clustering (Section 4.2) determined the number of clusters for the folds: $K = 11$. Results from the pair formation stage are summarized in Table 4.5 below. Of the total possible pairs of observations, only 17.3% actually occur, and of those, 69.3% occur in all of the partitions. Figure 4.5 shows that only pairs with $\alpha = 1$ occur in significant numbers - 69.3%. The second largest set occurs at $\alpha = 0.05$ and contains 7.7% of the pairs. These results illustrate the instability inherent in KMC – varying partitions arise due to different starting conditions for KMC. KMC methods traditionally involve multiple partitions for a given value of K to determine the one with lowest inertia, but they do not evaluate how much the observation labels change among the multiple runs. As this analysis shows, only a fraction, albeit a large one, of the observations consistently coexist in the same clusters when KMC is repeated.

Measurement	**Value**
Total Observations (N)	51,283
Partitions (B)	20
Training Fold Size	80%
Possible Pairs	1,314,947,403
Actual Pairs	227,138,275
Highest Stability of any Pairs	1.00
Number of Highest Stability Pairs	157,484,658
Fraction of Highest Pairs	69.3%
Implementation Time (wall clock secs)	3,001

Table 4.5: Pair Creation Results

The 20 subgraphs $G_{(c)}, 1 \leq c \leq B$ are formed and the elbow method is used

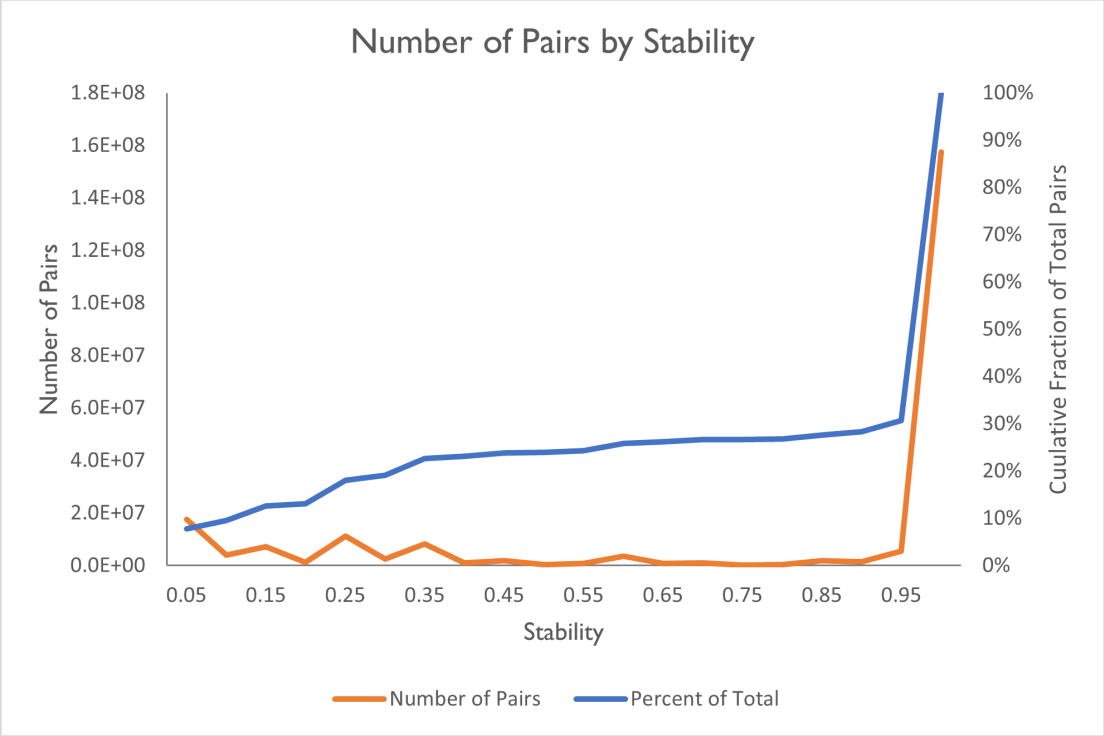
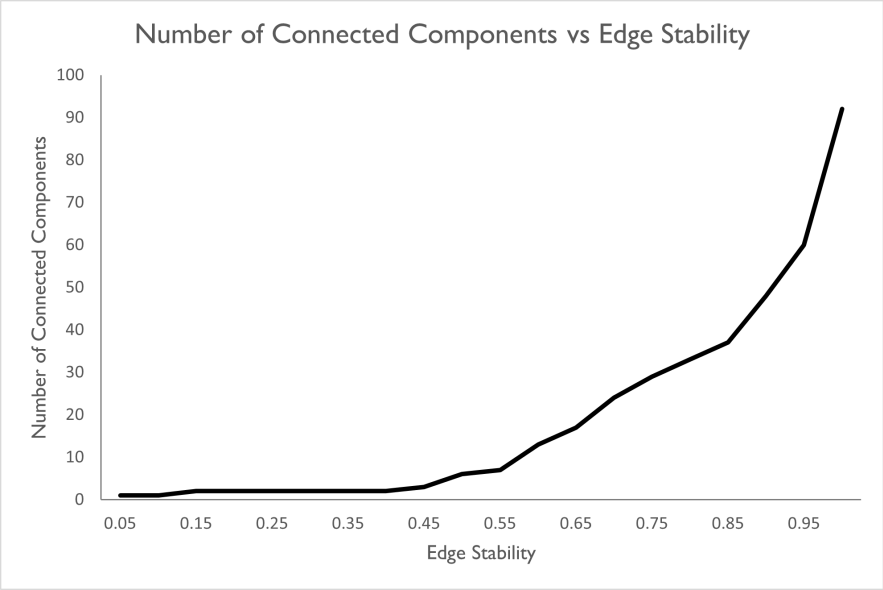


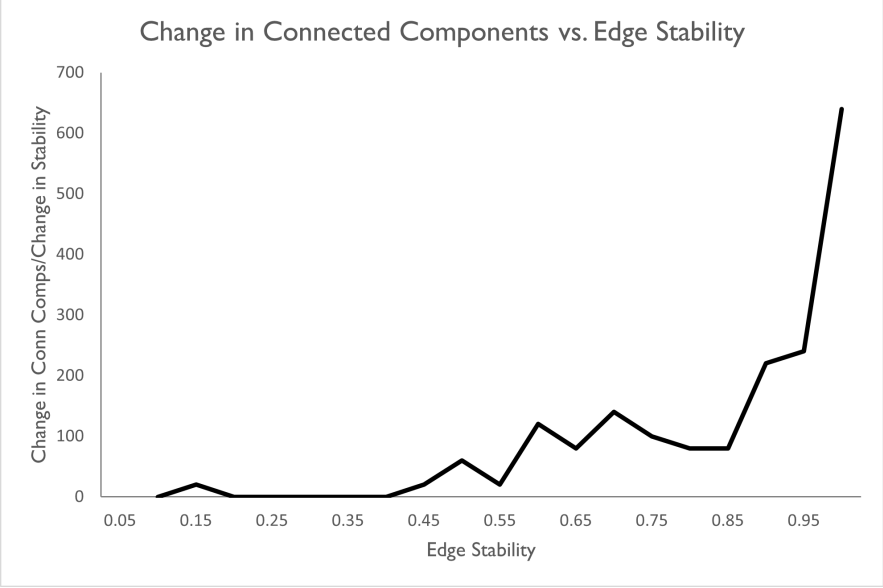
Figure 4.5: Number of Pairs vs Stability α

to determine the stability $\alpha = \frac{c}{B}$ at which the number of components changes most (Figure 4.6). The second subplot shows the first derivative of the connected components vs stability curve. Two significant shifts in direction occur – one at $\alpha = 0.60$, and one at $\alpha = 0.85$. At $\alpha = 0.60$ there are 13 connected components in the corresponding subgraph $G_{(12)}$. At $\alpha = 0.85$ there are 37 connected components in $G_{(17)}$. Selecting between these two parameters is somewhat arbitrary, but 37 components is an overly complex partition for this situation. It is therefore concluded that there are 13 connected components of observations.

The 13 components are the stable clusters for the 30% data sample. The size of the sample's represented population in each component is shown in Figure 4.7. The first three are significantly larger than the rest. The largest cluster contains 29.9% of the population, the second contains 19.8%, and the third 17.4%. Size



(a)



(b)

Figure 4.6: Change in Number of Connected Components vs Stability α

begins to decrease with the fourth, which has 10.7%, and the fifth only 5.4% of the population. The cumulative fraction of sample population in the stable cluster

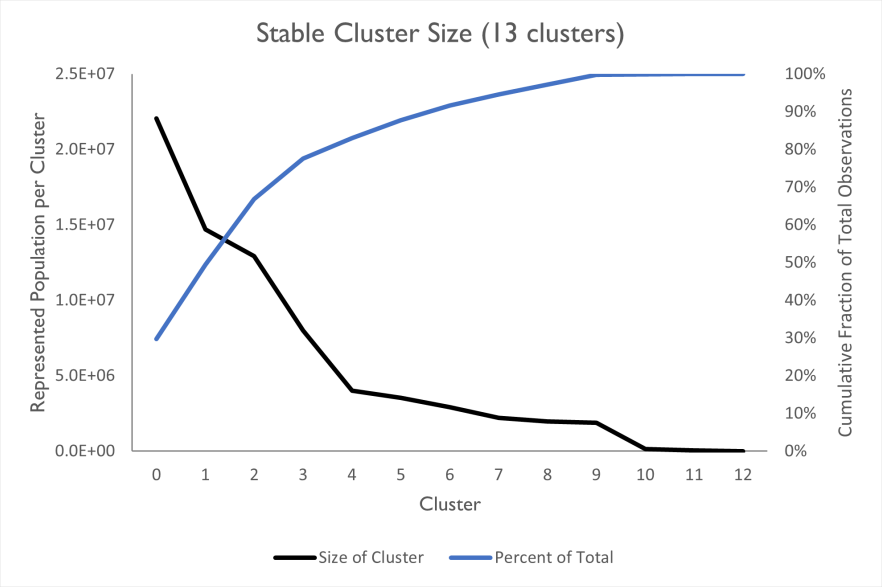


Figure 4.7: Size of Connected Component $\alpha = 0.60$

partitioning and in partitioning from the original KMC method are shown in Figure 4.8. The two curves are similar in shape, but the data is more concentrated in the stable cluster method than in the basic KMC one.

The entire NSDUH dataset is labeled by mapping each observation to the cluster whose center is nearest. This labeled dataset is the stable cluster partitioning of the four years of adult respondents to the NSDUH survey.

4.4 Characteristics of the Stable Clusters

The centers of the stable clusters are listed in Table 4.6. The three largest clusters from the original KMC and the stability methods are very similar with some differences in AFU and relative size. Cluster **0**, 21.1% of the population, is centered on initiation of tobacco at age 17 and alcohol at 18. This cluster is nearly identi-

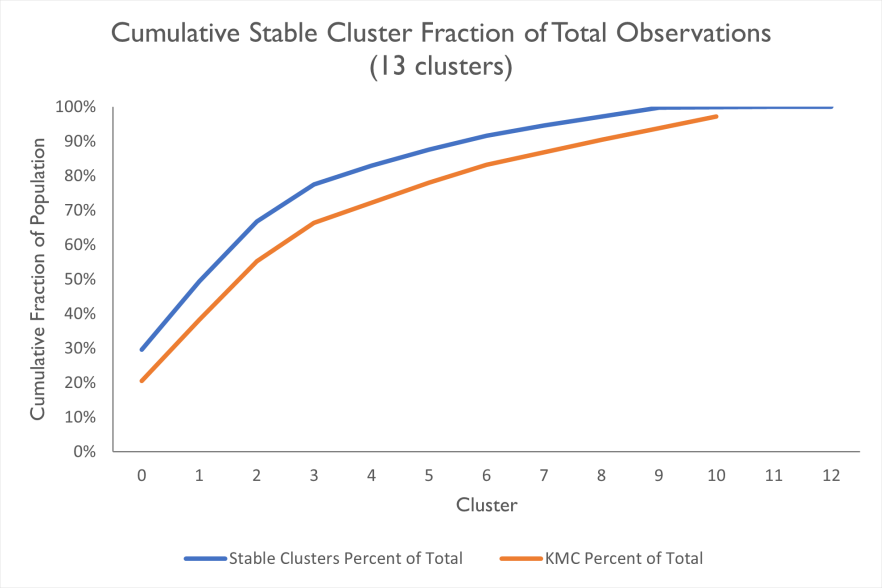


Figure 4.8: Comparison of Cluster Sizes: Stable vs KMC

cal to cluster **0** from standalone KMC. Cluster **1**, the largest cluster representing 25.2% of the population, is centered on initiation of tobacco at 16, marijuana at 19, and alcohol at 35. This cluster differs meaningfully from that of basic KMC, which was smaller, initiated alcohol prior to marijuana, and had a much lower mean AFU for alcohol initiation. The third large cluster, **2**, accounts for 17.3% of the population and is centered on initiation of alcohol at 20 and nothing else. Again, this cluster is nearly identical in size and structure to that of basic KMC. Cluster **3**, like that of basic KMC, represents no usage and represents 10.9% of the population.

Clusters **4**, **6**, **7**, and **8** all center on polyabuse. Three of the five, clusters **4**,**5** and **7**, have alcohol as the first used drug. Collectively, these polyabuse clusters represent 21.7% of the population. With the exception of **6**, all these polyabuse clusters are centered on some prescription drug abuse, which wasn't present in clusters **0-3**.

Cluster **9** is nearly identical to its basic KMC counterpart and centers on

initiation only of tobacco at age 17 and represents 2.9% of the population. The remaining clusters are very small, but cluster **10** is interesting in that it centers on initiation only of marijuana. This cluster was not seen in basic KMC, but the stable cluster partition does include more clusters, allowing for expression of patterns not identified in basic KMC.

We can think of each cluster's medoid (listed in Table 4.7), the observation closest to the geometric mean, as a 'typical' user in the group. Most of the medoids match the centers very well. For example, the center of **0** is [17.5, 18.4, 991.0, 980.5, 944.1, 983.1, 971.9, ,990.1, 986.9], and its medoid is [17, 18, 991, 991, 991, 991, 991, 991, 991]. The medoids for clusters **4**, **6**, **7**, and **12** differ somewhat from the cluster centers, but this is expected given the complexity of the polyabuse clusters.

	Cluster						
Drug	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Tobacco	17.5	16.3	991	991	31.8	991	65.4
Alcohol	18.4	35.0	20.2	991	17	17.6	26.2
Marijuana	991	19.3	991	991	18.7	20.3	26.3
Cocaine	980.5	888.2	988	990.8	179.9	926.2	25.9
Prescriptions	944.1	991	951.4	966.4	464.3	899.6	991
Hallucinogens	983.1	885.6	985.7	989.1	106	901	120.8
Inhalants	971.9	946.2	972.6	977.9	18.3	948	991
Heroin	990.1	989.8	991	990.5	823.9	989.1	895.6
Methamphetamine	986.9	977.7	989.5	990.4	600	976	659.4
Fraction	21.1%	25.2%	17.3%	10.7%	6.3%	4.8%	4.3%

	Cluster						
Drug	**7**	**8**	**9**	**10**	**11**	**12**	
Tobacco	34.7	40.3	17.6	985.7	17.3	81.4	
Alcohol	25.9	21.2	991.0	991.0	18.3	386.3	
Marijuana	33.4	21.1	986.1	19.7	18.2	845.5	
Cocaine	823.9	26.0	984.7	953.3	991.0	964.0	
Prescriptions	35.7	34.3	959.7	924.7	178.4	130.5	
Hallucinogens	984.1	99.6	985.2	929.0	19.9	19.6	
Inhalants	851.6	991.0	977.9	963.7	941.2	797.0	
Heroin	985.2	816.1	988.7	991.0	974.3	964.1	
Methamphetamine	964.4	616.1	985.2	988.7	719.8	965.9	
Fraction	3.5%	2.8%	2.5%	0.2%	1.3%	0.1%	

Table 4.6: Stable Cluster Geometric Centers

Cluster	Medoid Pathway
0	[(tobacco, 17), (alcohol, 18)]
1	[(tobacco, 14), (marijuana, 17), (alcohol, 35)]
2	[(alcohol, 20)]
3	[No use]
4	[(tobacco, 12), (alcohol, 14), (marijuana, 17), (hallucinogens, 17), (cocaine, 32), (inhalants, 34), (prescriptions, 61)]
5	[(alcohol, 18), (marijuana, 20)]
6	[(alcohol, 17), (marijuana, 17), (cocaine, 21), (hallucinogen, 40), (tobacco, 45)]
7	[(tobacco, 18), (alcohol, 23), (marijuana, 34), (prescriptions, 38)]
8	[(alcohol, 18), (marijuana, 18), (tobacco, 21), (cocaine, 34), (prescriptions, 57), (hallucinogens, 65)]
9	[(tobacco, 18)]
10	[(marijuana, 20)]
11	[(tobacco, 15), (alcohol, 16), (marijuana, 16), (hallucinogens, 17), (prescriptions, 69)]
12	[(alcohol, 21), (hallucinogens, 25), (tobacco, 35), (prescriptions, 61)]

Table 4.7: Stable Cluster Medoids

Chapter 5

Discussion

5.1 Review

Despite ongoing and various attempts to eliminate or reduce them, drug abuse and addiction have been persistent problems in the United States. Legalization of marijuana and decriminalization of other drugs have been accompanied by shifts in usage patterns. The opioid epidemic, in part driven by an increase in prescriptions of pain medications, includes the abuse of heroin, fentanyl, and other synthetic opioids, including prescription medicines. This epidemic has caused the death of tens of thousands of Americans and has placed a great economic burden on the nation.

The study of drug initiation sequences contributes to an understanding of usage patterns and can potentially inform intervention and mitigation strategies to prevent users' progression from less harmful drugs such as alcohol and marijuana, to dangerous ones such as heroin and methamphetamine. Initiation sequences have been studied in the past, but those studies have focused primarily on the sequence by which drugs are initiated rather than on the combination of sequence *and* age

of first use. Prior studies have also limited the number of drugs considered and have mostly sought to explore a priori hypotheses regarding initiation sequences. No studies have combined an open exploration of usage patterns, consideration of a wide variety of drugs, and use of a large dataset.

Two challenges complicate the study of age of first use sequences. The first is access to information. Researchers cannot practically observe drug initiation and must rely on self reported information from users. Fortunately, the National Survey of Drug Use and Health provides a vast amount of survey data gathered over many years. While it is not perfect, NSDUH is arguably the best resource available for researchers – it is designed to represent the majority of the US population, it records AFU values for drugs, and its stability over time allows combination of data from multiple years.

The second challenge is that of data complexity. There are many possible AFU sequences, and identifying patterns that can be generalized is computationally difficult. Prior studies have focused on searches for hypothesized sequences and have used traditional statistical techniques to evaluate them. This does not entail an open exploration of data that is unbiased by a priori supposition. Unsupervised machine learning techniques address this challenge by offering researchers ways to let patterns in the data emerge. Clustering, and K-means clustering in particular, is an efficient unsupervised technique capable of handling large datasets.

Because KMC is an heuristic method that identifies local, not global, optima, it is sensitive to starting conditions. When used on survey data that represents a larger population, KMC has another drawback – if it *did* arrive at a globally optimal solution, it risks overfitting the survey data. Hence a modification to KMC is required to stabilize its predictions while allowing for application to unforeseen data. Other researchers have proposed modifications to KMC to satisfy

the stability condition, but had not applied them at scale, and the methods would be unlikely to work practically in such situations.

This study builds on earlier proposals to develop a stability-enhanced KMC method (SEKMC). SEKMC conducts KMC on multiple bootstrap samples of a dataset and identifies persistent cluster co-membership pairs of data points. The pairs that exhibit at least a level of stability specified by the researcher are included in a set. This pair set is modeled as edges in a network. SEKMC uses list construction rather than connectivity matrices to limit the computational burden of the method. During implementation, it uses cluster computing to distribute pair set formation in order to make practical the execution of the algorithm. Connected components are identified in the constructed network. Because these components are completely connected due to the nature of network formation, SEKMC can benefit from an adjusted algorithm to determine them in $O(N)$ time. The consequent components are the clusters of drug use initiation sequences observed in the survey data.

5.2 Findings

Basic exploration of the NSDUH data shows that most of the US adult population has used alcohol and tobacco at least once, and that a substantial portion has used marijuana. Initiation of drugs predominantly occurs in the late teens and early twenties with the notable exception of prescription drug abuse, which tends to occur later. The most common initiation sequences begin with tobacco or alcohol, but marijuana has arisen as a common second, or less frequently, first drug in the sequence.

Initial unstabilized KMC determines that there are 11 clusters of AFU se-

quences in the NSDUH data. Multiple partitions of 11 clusters are performed on 20 bootstrap samples of the data. Of all the possible pairs generated by these partitions, only 17.3% actually occur, illustrating lack of randomness in usage patterns. Out of those pairs, only 69.3% occurred in every partition, demonstrating the variability of unstabilized KMC.

Results from SEKMC are similar to those of unstabilized KMC with a few deviations, notably the growth of a cluster where marijuana was the second initiated drug (after tobacco), and alcohol becomes the third: tobacco $\rightarrow$ marijuana $\rightarrow$ alcohol.[1] This largest cluster represents 25.2% of the US adult population and is centered on very early tobacco initiation at 16.3 followed by marijuana at 19.3. The late AFU for alcohol at 35 is likely due to the presence of a significant cohort of no-alcohol-use members. This cluster is worthy of further in-depth study, and its presence demonstrates the usefulness of this approach – it is unlikely that an a priori hypothesis would be formed that requires a search for early age tobacco and marijuana initiators with late or non-existent alcohol initiation.

The second largest cluster is centered on usage of only the traditionally legal drugs, tobacco and alcohol. The centered AFU for tobacco is 17.5 and for alcohol is 18.4, both below the current legal AFU for both drugs in most states. The largest cluster, containing more marijuana use, is centered on a lower AFU for tobacco than the tobacco/alcohol only cluster. This comparison again illustrates utility of the unsupervised approach – subsequent research can focus on the difference in AFU between tobacco/alcohol only users and tobacco/alcohol/marijuana or tobacco/marijuana users. Without unsupervised exploration, a researcher would have to develop this hypothesis through other means.

A significant portion of the population, 10.7%, did not initiate *any* drugs.

[1] In unstabilized KMC, the sequence is tobacco $\rightarrow$ alcohol $\rightarrow$ marijuana.

The aggregation of observations into clusters, including their accompanying demographic features, enables researchers to compare this cluster with the others to determine if any attributes are more common to non-users than users.

The third largest cluster is centered on alcohol use only, with an AFU of 20.2, which is near the common legal drinking age of 21. This cluster represents 17.3% of the population. Interestingly, the clusters centered on the two other most common drugs are much smaller. The tobacco-only centered clustered center represents only 2.5% of the population, and the marijuana-only centered a mere 0.2%. From this it is hypothesized that alcohol has been the intoxication drug of choice among non-polyabuse users, most likely due to its long history of legality. It is possible that the size of the marijuana-only cluster could grow as its legalization drives increased usage.

Polyabuse clusters show a great deal of variation, indicating that polyabuse itself, rather than the sequence of polyabuse, is a defining characteristic of drug initiation patterns. Nevertheless, some interesting observations regarding these clusters can be made. For example, there is no cluster centered on polyabuse that has a low AFU for tobacco, but there are polyabuse clusters whose centers include early AFU for alcohol and marijuana. This could indicate that polyabusers are focused on intoxication and tobacco initiation is not a precursor toward such behavior. Again, this is an hypothesis developed through unsupervised data exploration and is worthy of further study.

Prescription drug abuse, which has grown since the advent of Oxycontin, oxycodone and the like, is associated with the abuse of other drug combinations, including alcohol/marijuana, and alcohol/marijuana/cocaine. There is no cluster centered on only prescription drug use or on only prescription drugs and the traditionally legal drugs, tobacco and alcohol. This finding poses a particularly

interesting hypothesis: are users of only legal drugs less likely to become prescription drug abusers even later in life when they may experience a pain treatment regimen that includes licit opioid use? In light of the probable link between opioid over-prescription and the current epidemic, this is a fascinating question. Medical practitioners could be encouraged to use prior drug use history as a screening question to suggest other pain management techniques prior to opioid prescription.

5.3 Limitations of this study

This study uses a static combination of four years' of NSDUH data. If, as is surmised, drug use patterns are evolving with changes in legalization and destigmatization, a sequential analysis should be performed. This could be done by considering overlapping four year blocks, such as 2015-2018, 2016-2019, 2017-2020, and 2018-2021. To ensure the presence of observations containing the use of relatively rare drugs such as heroin and methamphetamine, four year, rather than one year, blocks of data are recommended. Such a sequential study would depict the transformation of usage through changes in the characteristics and sizes of stable clusters over time. With SEKMC already defined and demonstrated, such a study would be easy to conduct.

The most expensive portion of SEKMC is pair generation. The combined partitions dataset was limited to 30% of the total in this study to manage compute time and expense. Nevertheless, sensible stable clusters emerged and the time required to generate them was not prohibitive. Future studies, because they don't require the trial and error associated with algorithm development and coding, can expand to include larger samples or even the entire set of combined partitions.

There are other practical ways to limit compute costs. This study used unreserved clusters, which while flexible are more expensive to provision than reserved ones. Researchers who can spread costs across multiple studies can use reserved clusters and thereby greatly reduce the costs of SEKMC.

This study focused on the development of the SEKMC algorithm. Other than for the determination of K, it did not conduct hyperparameter tuning, which could improve performance. For example, the number of starting points for KMC used in generating partitions on folds of the data was not varied. It is possible that cluster stability can be improved by such tuning.

Finally, the emphasis of this study is on unsupervised learning and the discovery of patterns that can lead to further in-depth analysis. The AFUs of the observations within each cluster differ from those of the centroid. So interpretation of the clusters must be moderated. For example, it is not proper to say, "21.1% of US adults initiate tobacco at age 17.5 and alcohol at 18.4 with no further drug use." Instead, one would say, "21.1% of US adults have similarities that segregate them from other portions of the population, and on average, that cohort initiates tobacco use at age 17.5, alcohol at age 18.4, and is unlikely to use any other drug." Statements such as this are appropriately careful yet still provide insight into drug use patterns.

5.4 Conclusions

Exploration of large complex datasets can be conducted through unsupervised machine learning techniques such as cluster analysis. This study improves upon those techniques by increasing the stability of the underlying methods, in this case K-means clustering. This study employs those techniques on the NSDUH survey

to uncover patterns of drug initiation among the US adult population. These patterns provide unique insight into drug initiation sequences, including the ages at which drugs are first used. Future studies can build on these findings to explore novel hypotheses that can in turn guide drug abuse mitigation strategies. The subsequent chapters of this study provide two examples of the usefulness of stable cluster data. In the first, a review of the Gateway Hypothesis is conducted. In the second, demographic feature variation across the clusters is explored.

Chapter 6

Investigating the Gateway Hypothesis

6.1 Introduction

An individual can abuse zero, one, some, or many substances. Concern for the individual grows with the danger posed by the drug – a user of heroin risks immediate death far more than does a user of tobacco alone. Researchers have long known that a subject's first used drug is rarely a substance like cocaine or ecstasy, instead common substances like tobacco and alcohol tend to precede use of more dangerous or illicit drugs. Denise Kandel formalized a progression of drug use *stages* (D. Kandel, 1975). She placed drugs into categories: tobacco, alcohol[1], marijuana, and other illicit drugs. She claimed that drug use began with one legal drug, either alcohol or tobacco, then progressed to the other legal drug, then to marijuana, and from marijuana to other illicit drugs. A stage is defined by the last drug initiated by a user. For example, the first stage would be non-use,

[1]Kandel's original work further split alcohol into *beer/wine* and *hard liquor*

and the second would be the use of only tobacco or alcohol. Kandel maintained that the occupation of a stage does not demand progression to the next stage, but it is a required precursor to do so. In other words, a marijuana user will not necessarily use heroin, but before using heroin, a user must have used marijuana.

Kandel's theory evolved into the *Gateway Hypothesis*, which has influenced drug use prevention policy for decades. An assumption made by some policymakers is that if youths are prevented from using alcohol, they will not use marijuana. If they do not use marijuana, they will not use cocaine (or ecstasy, or heroin, etc.). This belief has been adopted by US Federal, state, and non-governmental organizations in efforts to combat drug abuse. For example, the Office for Substance Abuse Prevention (OSAP) was created by the Anti Drug Abuse Act of 1986 to lead efforts by the Federal Government to prevent substance abuse problems (Jansen, 1992, p. iii). Kandel's work guided OSAP to create 'mass media efforts [that focused] on preventing use of "gateway" drugs at early ages in an attempt to reduce the likelihood of developing alcohol and other drug problems and multi-drug use patterns.' (Jansen, 1992, p. 35)

In December 2010, the Office of Disease Prevention and Health Promotion (ODPHP) of the US Health and Human Services Administration (HHS) included reduction of initial alcohol and marijuana use by adolescents in its *Healthy People 2020* national objectives. The Gateway Hypothesis influences actions at the state level as well. In 2022, the Texas Health and Human Services Commission maintained that marijuana use can lead to 'dependence, addiction, and increased use of other drugs' (Texas Health and Human Services, n.d.). In the private sector, gateway theory has been adopted by non-profit advocacy groups and treatment facilities. In 2021, the Drug Abuse Resistance Education (D.A.R.E.) program opposed marijuana legalization in part because of 'increased risk of addiction and use

of other more lethal drugs' (Drug Abuse Resistance Education Program, 2021). In 2020, Advanced Recovery Systems, a behavioral health company with 750 employees and $110M of revenue dedicated an entire page to Gateway Drug education and explained that 'A gateway drug is a habit-forming drug that can lead to the use of other, more addictive drugs' (Advanced Recovery Systems, 2020).

As the popularity of the Gateway Hypothesis grew, so did its number of critics. Many point out that lower stage drug use is not a determinant of higher stage drug use, but that is a flawed critique because neither Kandel nor her supporters ever made such a claim – she said lower stage drug use was a necessary precursor for, but not a guarantee of higher stage use. Some have disagreed with the rigidity of the hypothesis – it could be reasonable to conclude that marijuana use can precede alcohol or tobacco use, especially given the drug's increasing popularity. Other debates have centered on 'which drug is *the* gateway drug?' This argument became public when legalization advocates falsely claimed that D.A.R.E. reversed its position about marijuana's status as a gateway drug. Barry et al. (2016) found that alcohol use precedes that of tobacco or marijuana among polyusers – individuals who have used many different drugs. Their conclusions were twisted in reporting by the Washington Post, which used the study to criticize Chris Christie's anti-legalization stance during his run for the Republican Presidential nomination (Ingraham, 2016).

This analysis focuses on the applicability of the Gateway Hypothesis to polyabuse, which is the use of many different drugs by a subject. This dissertation has shown that some respondents to the NSDUH survey can be partitioned into clusters whose most notable feature is polyabuse (Section 4.4). Building on this finding, a two-part analysis is conducted. First, the drug initiation sequences of the respondents in the polyabuse clusters is examined in more detail to determine

whether or not users' behavior adheres to the sequence dictated by the Gateway Hypothesis. Second, a classification analysis is conducted to determine the relationship between AFUs of the quasi-legal drugs (tobacco, alcohol and marijuana) and potential for use of illicit drugs.

6.2 Gateway Hypothesis literature review

The origin of the Gateway Hypothesis[2] is credited to Denise Kandel, who with other researchers proposed that there are developmental stages and sequences of involvement in drugs (Hamburg, Beatrin A.; Kraemer, Helena C.; Jahnke, 1975; D. Kandel, 1975). They described a progressive and hierarchical sequence of drug use stages that begins with one of the traditionally legal substances, tobacco or alcohol, progresses next to the other, then to marijuana, and then to other illicit substances such as cocaine, methamphetamine, and heroin. Involvement in various classes of drugs is not random, but instead follows specific pathways, and an individual who uses a lower level drug is at risk of progressing to a higher one. Kandel claims that these pathways adhere to the Guttman scale model (Guttman, 1944), in which presence in a stage of an ordered sequence indicates prior presence in all preceding stages. For example, a subject who has used heroin must have previously used marijuana and the historically legal substances, alcohol and tobacco. The user will not have progressed from alcohol to heroin without first using marijuana. Kandel's theory arose from a longitudinal survey-based study of New York state high school students (D. Kandel, 1975). It is important to note that neither Kandel nor other proponents of the Gateway Hypothesis maintain

[2]Kandel did not coin the term *Gateway Hypothesis*. Instead, the term *Gateway Drug* was popularized by Robert Dupont, who was the Director of the National Institute on Drug Abuse (Dupont, 1984). Description of the progression of drug use stages as the Gateway Hypothesis evolved from this term.

that presence in a lower stage guarantees progression to a higher stage. For example, most marijuana users will never use heroin. Also, a subject's progression can end at any stage. Kandel did however, claim that moving into a stage increases the risk of progression to subsequent stages.

Since its conception, the Gateway Hypothesis has been explored in subsequent research. Kandel's stages were observed internationally in the US, France, Israel (Adler & Kandel, 1981), Australia (Blaze-Temple & Lo, 1992), Japan (Oh et al., 1998), Spain (Adrados, 1995) and Scotland (Morrison & Plant, 1991). Other studies added an analysis of the age of initiation of drugs. Fleming et al. (1989) found that smoking cigarettes significantly increased the association with marijuana use in two years. Golub and Johnson (1994) showed that age-of-first-use of substances had decreased from the 1920s through the 1970s. Welte and Barnes (1985) considered age and found that cigarettes form an important step between alcohol and marijuana use for younger subjects.

The Gateway Hypothesis continued to be studied over several decades, and has been updated to consider deviations and links to other substance use. Fiellin et al. (2013) found an association between the use of alcohol, cigarettes, and marijuana with subsequent abuse of prescription opioids among young men. Keyes et al. (2016) studied twins and discovered that a prevalence of smoking in 8th and 10th grade is associated with marijuana and cocaine use in 12th grade. A broad analysis of the literature was done by Lynskey and Agrawal (2018), who concluded that the causal link between lower stage and higher stage drugs is heavily debated but not discounted. Noting some changes in stages among cultures, they suggest that drug use progression may be a factor of access and accessibility.

A subset of Gateway Hypothesis research operates with an assumption of the theory's validity and seeks to explain it. Some studies point to a pharmacological

basis – for example, Ellgren et al. (2007) found that intraperitoneal exposure to THC during the peri-pubertal period in rats was associated with an increased use of opioids. E. R. Kandel and Kandel (2014) observed that mice who were primed with nicotine exhibited an increased response to cocaine. Other studies focus on the effect of social factors on drug use stages. Fergusson et al. (2006) found that associations between regular or heavy cannabis use and use of other illicit drugs declines with age, and that exposure to the illicit drug market during cannabis use may increase accessibility to other substances. Wagner and Anthony (2002) claim that neighborhood and work environment impinge on risk of alcohol and drug use. Using an ongoing cohort study, Baggio et al. (2015) show that *positive* first use experience of cannabis amplifies the association with other illicit drug use.

A few studies discuss deviations from the normative drug initiation sequence and why they occur. Baggio et al. (2015) examined the US National Comorbidity Survey Replication to search for features that could explain deviations. They concluded that differing sequences are not predictive of later drug dependency, but they did find that adolescents with mental health problems are associated with sequence deviations. They also found that deviations from the gateway sequence are rare (5.2%), and that cannabis use has become more common in cohorts born since 2003. D. B. Kandel and Yamaguchi (2002a) consider the existence of different pathways among ethnic groups and use a log-linear analysis to conclude that the deviations among them are small, and that the alcohol/cigarette to marijuana to other illicit substances pathway is normative. They assign any deviations from this pattern to a latent class of users whose behavior is not normative.

A set of prior studies doesn't focus specifically on the Gateway Hypothesis, but seeks predictors of late stage drug abuse based upon other drug use and

demographic features. Lynskey et al. (2003) shows a relationship between early marijuana use and subsequent illicit drug that persists when other known risk factors are controlled. Fergusson et al. (2008) finds that environmental factors, when combined with early marijuana use, increase risk of illicit drug use. Jones (2013) identifies correlations between heroin and various demographic variables and drug use behaviors. This study is followed up in Jones et al. (2015), which describes relationships between cocaine, binge drinking, and marijuana use. Cerdá et al. (2015) uses hazard models to identify the link between prescription opioid abuse and subsequent heroin use. Two studies focus attention on age of first use. Wadekar (2020) correlates early marijuana use and mental illness with opioid use disorder. And Beattie and Nicholson (2021) finds that heroin use is strongly correlated to early use of marijuana and cocaine use.

6.3 Gaps in the literature

The Gateway Hypothesis is one of the most explored topics in drug use research. Nevertheless, opportunities for new insights remain. One is suggested by the presence of cluster **1** described in Section 4.4, whose center includes marijuana, rather than alcohol, as the second drug initiated. The Gateway Hypothesis rigidly maintains that alcohol and tobacco use must precede marijuana use. In light of recent legalization, it is reasonable to posit that marijuana may displace either alcohol or tobacco as a first or second drug in the initiation sequence. Deeper exploration of the stable clusters provides a means to check on the currency of the Gateway Hypothesis.

Existence of a set of clusters (**4**, **6**, **7**, **8**) whose common feature is polyabuse calls for further study of the link between early stage drug use and subsequent

illicit drug use. The literature has established correlation between factors, including early marijuana use, and later illicit drug use. However, no study focuses on the relationship of tobacco, alcohol, and marijuana AFUs to later stage drug use. Such a study can answer questions relevant to drug interdiction. For example, which drug, when used early, is the strongest determinant of later illicit drug use? Or is there a combination of early tobacco, alcohol, and marijuana use that is a strong determinant?

6.4 Methods

6.4.1 Evaluating drug initiation sequences in specific clusters

The stability enhanced K-means cluster analysis (SEKMC) conducted above sought patterns through unsupervised learning, considered age of first use, and was based upon a large multiyear survey with tens of thousands of respondents, a combined approach unique in the Gateway Hypothesis literature. The consequent partition in the study contains a cluster of particular interest. Cluster **1** represents 25.2% of the surveyed population and its centroid does not adhere to the Gateway Hypothesis: it's first drug used is tobacco at age 16.3, and instead of progressing to alcohol, it progresses to marijuana at age 19.3. The size of this cluster indicates the possible presence of a set of users that is too large to be considered a latent non-normative class. To determine if this is the case, the respondents in cluster **1** are studied apart from the rest of the NSDUH respondents. The AFU sequences for this cluster are evaluated with direct measurement of pathway frequencies.

6.4.2 Predicting later stage drug abuse based upon tobacco, alcohol, and marijuana use

Respondents in NSDUH are allocated into two classes. The positive class includes those who have used illicit drugs – those other than tobacco, alcohol, and marijuana. The negative class includes respondents who have used no drugs or those that have only used the quasi-legal drugs. Prediction of class membership is done using AFU of tobacco, alcohol, marijuana, and any combination of those three drugs. Two classification methods are employed, decision tree analysis and logistic regression. The methods are compared for efficacy and the importance of the AFU features is evaluated to determine what quasi-legal drug use pattern most strongly affects prediction of illicit drug use.

In both models, the AFU features are transformed to create nearly smooth value ranges, and the convention of using '991' for no-use is removed. The AFU for drug j of observation i is given by AFU_{ij}. The transformation of this feature is denoted by x_{ij}, where $j \in$ [tobacco, alcohol, marijuana] and is defined as follows:

$$x_{ij} = \begin{cases} \frac{1}{\text{AFU}_{ij}}, & \text{if } \text{AFU}_{ij} < 991, j \in \text{[tobacco, alcohol, marijuana]} \\ 0, & \text{otherwise} \end{cases}$$

Each observation in the NSDUH dataset is assigned into one of two classes, and the classification of an observation i is denoted as y_i. If a respondent has used any drug other than tobacco, alcohol, or marijuana, it is assigned to the positive

class ($y_i = 1$). Otherwise it is in the negative class ($y_i = 0$):

$$y_i = \begin{cases} 1, & \text{if } \exists \text{ AFU}_{ij} < 991, j \notin [\text{tobacco, alcohol, marijuana}] \\ 0, & \text{otherwise} \end{cases}$$

Decision tree classification aims to partition datasets into smaller, more homogeneous groups. Each node of the tree is a split point at which a set of the data is divided into two groups based upon the value of a feature selected by the algorithm. The split points are chosen to maximize the homogeneity of the resulting two subsets of data. The measure of homogeneity used in this study is the Gini index, which for a two-class problem is given by $2p_1p_2$, where p_i is the probability of an observation's membership in class i. The algorithm seeks to minimize the Gini index at each split point. The algorithm continues to build the tree until a stopping criterion is reached.

The performance of decision tree classification can be improved in various ways, including the use of random forests. The purpose of this study is to develop an explanatory model, and such methods are not used so as to preserve the interpretability of the basic tree model. The dataset is split into a training set containing 75% of the observations and a test set with the remaining 25%. Two tuning parameters are used, maximum tree depth and the minimum number of samples in a node required to proceed with a split. The model is trained using 5-fold cross-validation to maximize the area under receiver operating characteristic curve (AUROC). The trained model is applied to the test set and the AUROC and maximum F1-score are calculated. Using the threshold probability at which the F1-score is maximized, a confusion matrix for the test set is generated. The observation weights (YRWEIGHT) are used throughout the process to account

for the population represented by each observation.

Logistic regression is a general linear method that can be used to predict the probability of an observation's class membership. Because it is linear, combinations of features are modeled by inclusion in the model definition. For this study, the model is given by:

$$\log \frac{P(y_i = 1)}{1 - P(y_i = 1)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7}$$

$$x_{i4} = x_{i1} x_{i2}$$

$$x_{i5} = x_{i1} x_{i3}$$

$$x_{i6} = x_{i2} x_{i3}$$

$$x_{i7} = x_{i1} x_{i2} x_{i3}$$

The tuning parameters used for the logistic regression classification are whether or not to include the intercept β_0, and what type of penalization to use to reduce the number of significant features (none, L1, L2, both). As with the decision tree method, data is split into 75% training and 25% test sets, and the model is trained with 5-fold cross-validation in order to maximize AUROC. The trained model is applied to the test set to obtain AUROC, maximum F1-score, threshold probability to generate maximum F1-score, and the confusion matrix at that threshold.

6.5 Results

6.5.1 Drug initiation sequences

The most common pathways, those that make up 80% of the population represented by respondents allocated to cluster **1** are shown in Table 6.1. The two most common comprise 49.4% of all of the Cluster **1** pathways, and both adhere to the Gateways Hypothesis – they initiate with either tobacco or alcohol, progress to the other, and then to marijuana. The next two deviate from the Gateway sequence – marijuana is the second drug used in these pathways that make up 16.5% of those in the Cluster **1** population. Other deviations from the Gateway sequence are seen in the seventh and eighth most common pathways, and these combine to represent 5.3% of the Cluster **1** population.

Path	**Fraction**
[start, tobacco, alcohol, marijuana]	34.5%
[start, alcohol, tobacco, marijuana]	14.7%
[start, tobacco, marijuana, alcohol]	8.7%
[start, alcohol, marijuana, tobacco]	7.9%
[start, tobacco, alcohol, marijuana, cocaine]	5.1%
[start, tobacco, alcohol, marijuana, hallucinogen]	4.2%
[start, marijuana, tobacco, alcohol]	3.3%
[start, marijuana, alcohol, tobacco]	2.0%

Table 6.1: Most Common Drug Use Pathways in Cluster **1**

Drug use progression modeled as traversal through a network is shown in Figure 6.1. The strongest links in the network are among the early stage quasi-legal drugs with other heavily weighted links connecting marijuana to cocaine and both marijuana and alcohol to hallucinogens. Taken together, the common pathway list and network progression model show that while the centroid of Cluster **1** deviates from the sequence in the Gateway Hypothesis, the cluster's two most

common paths actually do adhere to the sequence. However, the cluster exhibits a multitude of paths in which marijuana is the first, or more commonly second, drug in the AFU sequence. The combined effect of these paths drives the cluster center to a low AFU for marijuana use. These observations indicate that the Gateway Hypothesis drug initiation sequence remains normative, but marijuana is frequently initiated earlier than that sequence would dictate.

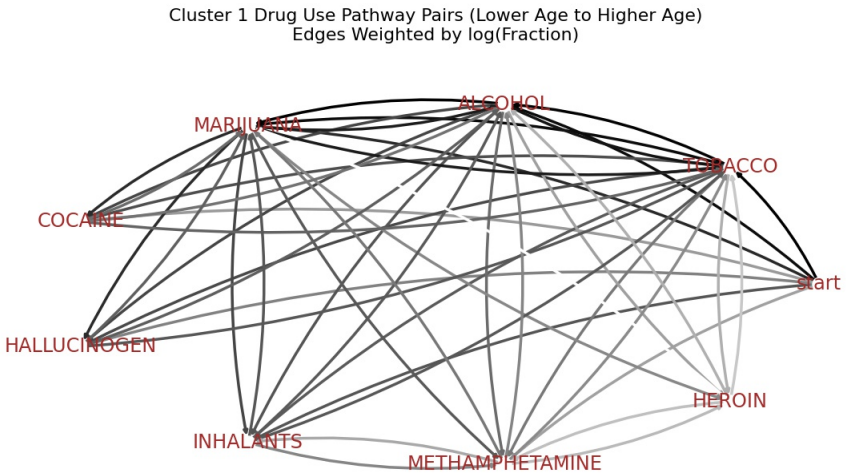


Figure 6.1: Cluster **1** Drug Use Progression Graph

6.5.2 Illicit drug use classification

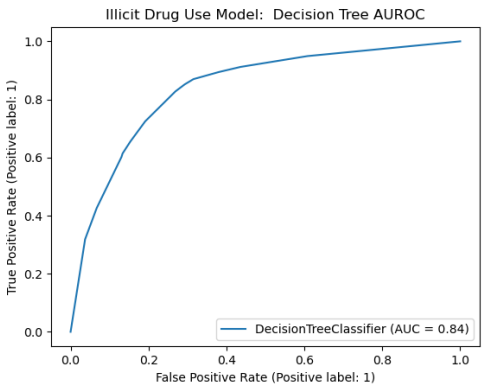
Results for the two classification models are summarized in Table 6.2 and in Figures 6.2 and 6.3. The performance of the two classification models was equivalent and high – both had a test set AUROC of 0.84 and maximum F1-score of 0.67. The best threshold probability for the decision tree model was 0.38, and that of the logistic regression model was 0.39. The decision tree confusion matrix is slightly more accurate than that of the logistic regression one. 83% of illicit drug

users were properly assigned to the positive class, and 73% of non users were properly assigned to the negative class. From these results it is concluded that both models effectively predict illicit drug use based upon AFUs of tobacco, alcohol, and marijuana, and interpretation of the models offers insights into clear links between the features and outcome variable.

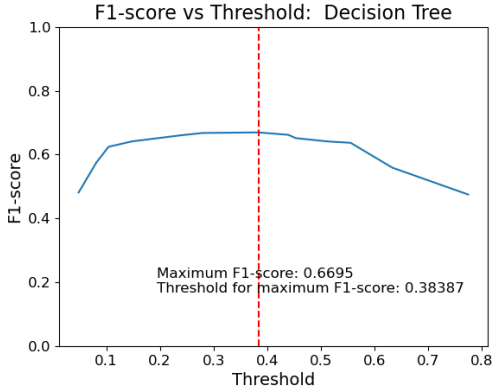
	Model	
Metric	**Decision Tree**	**Logistic Regression**
Best Model	Max depth = 4	Intercept included
	Min sample split = 2.5%	No penalty
AUROC	0.84	0.84
Max F1-score	0.67	0.67
Best Threshold	0.38	0.39

Table 6.2: Classification Model Performance

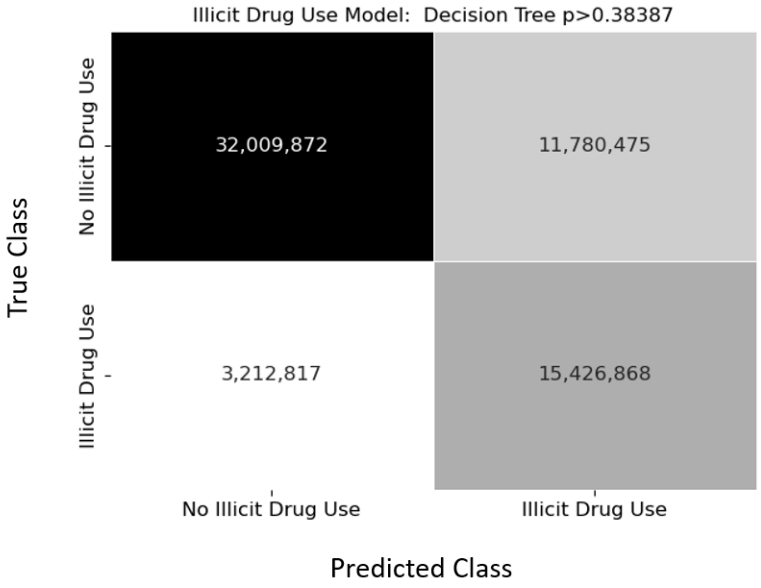
The decision tree classifier is shown in Figure 6.4. The first split (Node 0) partitions the dataset by whether or not a respondent first used marijuana at or above age 28.6. If true, the respondent is allocated to Node 1, and if false it is allocated to Node 2. 91.9% of the Node 1 respondents are in the negative class, meaning that they do not use other illicit drugs. In contrast, only 44.0% of Node 2 is in the negative class. Node 2 is then split on whether a respondent first used marijuana at or above age 16.4, and those whose marijuana AFU is below this amount are allocated to Node 6. Only 30.9% of Node 6 is in the negative class. The most homogeneous node dominated by the *positive* class is Node 24, which contains 32.0% of the illicit drug using population and is determined by marijuana AFU less than 16.4, alcohol AFU less than 15.4, and tobacco AFU less than 50. 77.5% of Node 24 is in the positive class. The most homogeneous node dominated by the *negative* class is Node 13, which contains 39.5% of the non-illicit drug using population and is determined by marijuana AFU $\geq$ 58.8, and alcohol AFU $\geq$ 19.6. 95.2% of Node 13 is in the negative class.



(a)

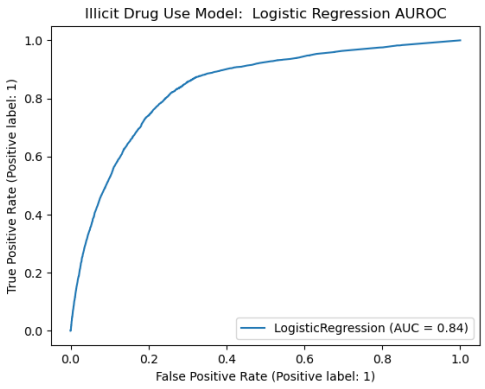


(b)

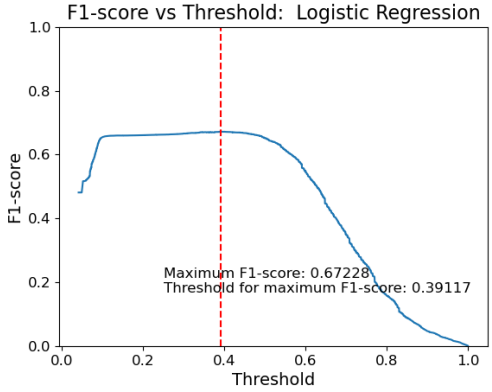


(c)

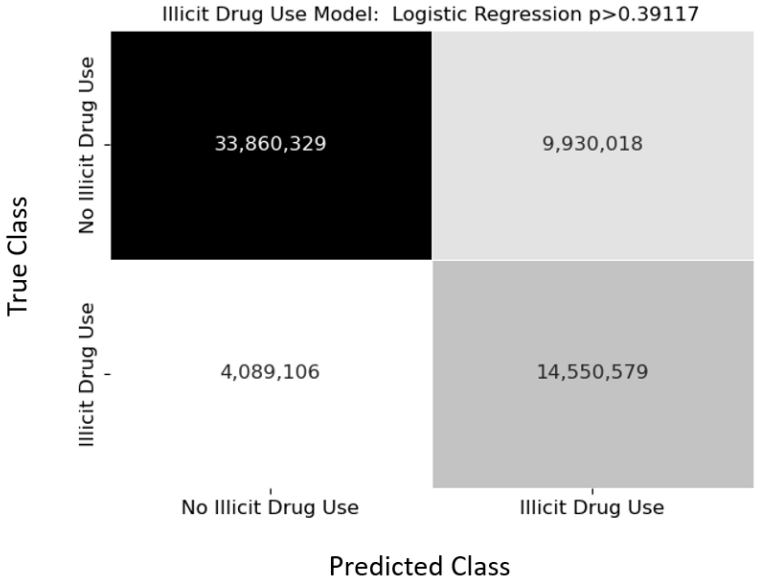
Figure 6.2: Decision Tree Model Performance



(a)



(b)



(c)

Figure 6.3: Logistic Regression Model Performance

Decision tree model variable importance scores summarize the comparative degree to which features contribute to dataset partitioning. Marijuana is by far the most important variable with a score of 0.93. Alcohol's is much lower at 0.06 and tobacco's is only 0.01.

The coefficients for the logistic regression model are shown in Table 6.3. Again, marijuana is the dominant individual AFU feature at 36.65, and the combination of tobacco and marijuana AFU has the highest coefficient at 114.83. The p-values for all coefficients except for alcohol/marijuana and tobacco/alcohol/marijuana are significant.

Feature	**Coefficient**	**p-value**
Intercept	-2.96	0.00
Tobacco	3.87	0.00
Alcohol	7.60	0.00
Marijuana	36.65	0.00
Tob/Alc	-15.69	$1.76e^{-5}$
Tob/Mar	114.83	0.00
Alc/Mar	-1.59	0.91
Tob/Alc/Mar	-247.17	0.09

Table 6.3: Logistic Regression Model Coefficients

6.6 Discussion

The presence of cluster **1**, whose centroid has marijuana as the second earliest drug of initiation suggests drug pathways that deviate from the sequence dictated by the Gateway Hypothesis and that their potential commonality could refute the rigidity of gateway theory. Upon deeper inspection, it appears that the Gateway Hypothesis sequence holds true, even for most of the pathways in cluster **1**. It is concluded that the gateway sequence remains valid for most of the US population. That said, controlling for age may lead to a different outcome. Marijuana legal-

Figure 6.4: Decision Tree Model

ization is a recent phenomenon. Consequently, marijuana may be a first or second drug in the initiation sequence for younger NSDUH respondents. This study focuses on adults – an investigation into the early stages of drug use among youths may lead to very different results. This investigation should be conducted and data should be segmented by survey year to evaluate whether the commonality of marijuana as a first or second stage drug is growing.

A set of clusters whose defining characteristic is polyabuse leads to a hypothesis that specific early stage drug initiation patterns may increase the likelihood of later illicit drug use. In particular, the AFUs for tobacco, alcohol, and marijuana can be used to predict such behavior. Classification of illicit drug use using AFUs for tobacco, alcohol, and marijuana provides overwhelming evidence that early marijuana use is *very highly* linked to use of later stage drugs. When present with early AFUs for alcohol and tobacco, early marijuana use is even more strongly correlated to subsequent illicit drug initiation. However, neither early tobacco AFU nor early alcohol AFU, or even a combination of the two, strongly predicts illicit drug use in the absence of early marijuana AFU. These findings are similar to those by other authors, but by excluding the presence of other demographic or environmental factors, this study definitively demonstrates that if indeed there is a 'gateway drug' to illicit drug use, it is marijuana used at an early age.

Future studies should focus on emerging patterns of drug initiation using youth responses from recent NSDUH studies. If it is shown marijuana use is supplanting tobacco or alcohol as a first or second initiated drug, there is a cause for concern – if marijuana AFUs are declining, it is likely that other illicit drug use will become more common. Practitioners should note these findings and end the debate over the identify of the 'gateway drug.' Ample evidence for the danger of early initiation of marijuana exists now. Policies to reduce potential use of dangerous

illicit substances should focus on restriction of marijuana access to youths. This is certainly difficult given the momentum behind marijuana legalization in the US. However, legalization does not preclude sentencing for marijuana possession by minors and for distribution of marijuana to teens. Education programs centered on responsible marijuana use by parents could also be beneficial. The dangers of youth access to their parents' marijuana products are higher than breaking into the liquor cabinet.

Chapter 7

Demographics and the Stable Clusters

The stable clusters simplify investigations into relationships between demographics and drug usage. Without them, researchers must ask very specific questions, such as "Does the initiation age of tobacco use differ among income categories?", and investigating drug use combinations becomes even more onerous. Partitioning the data into clusters enables the quick and easy determination of such relationships by examination of cluster distribution across demographic groups. If two or more groups have different distributions, and those differences are significant, a hypothetical relationship between drug initiation pathways and those groups exists. Such an analysis is conducted by creating cross-tabulation tables for each NSDUH demographic variable from Table 3.3 across the clusters.

The χ^2 test indicates statistical significance of differing frequencies within a cross-tabulation table. The influence of each cell provides insight into which particular combinations of factors most drive significance. Influence in a χ^2 table is a ratio of a cell's contribution to the overall χ^2 statistic divided by total χ^2. Let

a frequency table cell be denoted by c, let its expected frequency be denoted by f_c, its actual frequency by F_c, and its influence by I_c. χ^2 and influence are then calculated as follows.

$$\chi^2 = \sum_{c=1}^{N} \frac{(f_c - F_c)^2}{F_c}$$

$$I_c = \frac{\frac{(f_c - F_c)^2}{F_c}}{\chi^2}$$

Adult respondent NSDUH observations from 2016-2019 are partitioned into the 13 stable clusters by assigning each data point to the cluster whose mean is closest to its AFU vector. The frequency of each NSDUH demographic variable among clusters is shown in Tables B.1 through B.11 in Appendix B. The cross-tabulations are weighted according to each observation's representation of the sampled population. For ease of analysis, the tables are normalized so that population fractions, rather than raw counts, of observed frequencies are depicted. Every table exhibits a significant difference across the clusters for distributions of the demographic features, as is expected due to the great size of the sampled population.

Some interesting patterns are evident from a scan of the tables. Females, Non-Hispanic Asians, Pacific Islanders, and Hispanics are more likely to be included in the no-use cluster, **3** (Table B.5). Respondents aged 65+ are more likely to have only used tobacco and alcohol than any other cohort (Table B.2). Respondents with less than a high school education are least likely to be in the early marijuana cluster **1**, and most likely to be in no-use cluster **3**.

Such findings can guide researchers to more deeply explore patterns. For example, the prevalence of no-use among respondents with less than high school education may seem counter-intuitive, but it is also true that 18-25 year olds are

the most likely age cohort to be in the no-use cluster (Table B.2). This leads to a hypothesis that 18-25 year olds, who are more prevalent in the no-use cluster, are also less likely to complete high school due to their reduced age, and thus age, not education, is driving the increase in no-use among non-high school graduates. Examination of this hypothesis is easily accomplished through a cross-tabulation of no-use cluster **3** by age and education level . The tabulation shows that 22.5% of 18-25 year olds have not finished high school, but the proportions for other age cohorts are very similar: 19.3% of 26-34 year olds, 25.7% of 35-49 year olds, 25.7% of 50-64 year olds, and 27.0% of 65+ year olds. It is concluded that age does not explain the prevalence of no-use among low education respondents.

The complexity of the cross-tabulations makes direct interpretation difficult, and the data of most interest is that associated with high influence cells. The five most influential cells for each table are shown in Table B.12. Each row shows the feature table, feature value and cluster combination, influence and rank by influence of the combination in the χ^2 table. The table highlights some interesting findings, including:

- Military service respondents are more likely than expected to be in the tobacco/alcohol only (**0**) and no-use (**3**) clusters and less likely than expected to be in the alcohol only cluster (**2**), suggesting a prevalence of tobacco use in the military.

- Age 50-64 respondents show up in a cocaine cluster (**6**) more than expected, possibly due to popularity of cocaine in the 1980s.

- Females are more likely than expected to be in the alcohol only cluster (**2**) and males are less likely

- Non-Hispanic Asians are more likely than expected to be in the no-use cluster (**3**)

- Respondents who didn't finish high school and those in the lowest income category are more likely than expected to be in the no-use cluster(**4**)

- Those on government assistance are more likely than expected to be in the no-use and tobacco only clusters (**3** and **9**), but also in a polyabuse cluster (**5**)

This analysis demonstrates the ease by which meaningful hypotheses regarding the variation of drug initiation patterns as a function of age, education, and other demographic variables can be synthesized from a partition of complex data into stable clusters. Future research should expand this investigation to youth respondents and shifts in cluster frequencies by demographics as a function of time. One pattern that was *not* highlighted among the influential table cells was an obvious demographic predisposition to belong to the early marijuana use cluster, **1**. In fact, only respondents with more than $75,000 of income were more likely than expected to belong to it. Given the importance of marijuana's early AFU in classifying illicit drug users, it is somewhat disappointing not to find an obvious demographic pointer to early marijuana initiation. Again, a study that includes youth respondents, whose early stage drug initiations have occurred since marijuana legalization has become prevalent, may provide enlightening results.

Chapter 8

Conclusions

Given the continued harmful impact of drug abuse in the United States, it is important to maintain research to support the development of interdiction and mitigation methods. Drug use patterns also shift over time as new drugs are introduced to the population, attitudes towards use change, and most importantly, legalization of previously illicit substances occurs. These changes require research to be continually updated and reexamined to ensure the currency of conclusions used for policy formation.

A vast array of work has been done to understand drug using behavior. One area is a rich set of literature that examines the sequence in which drugs are initiated. The literature has produced grand theories such as the Gateway Hypothesis and has established links between the use of some drugs, such as alcohol and marijuana, with subsequent use of more dangerous drugs. Studying drug initiation sequencing faces two major challenges. The first, difficulty of information access, is partially addressed through the availability of large studies, conducted over many years and involving thousands of subjects, that are managed by institutions and governments. One of these, the National Survey on Drug Use and Health, is

well suited for drug initiation sequencing research.

Data complexity, the second challenge, arises from the existence of surveys like NSDUH, which contains hundreds of features and hundreds of thousands of observations. While most drug research starts with a priori hypotheses and uses data to test them, the richness of the NSDUH survey demands that the data 'speak for itself' through unsupervised machine learning techniques. One of these is the partitioning of the data into clusters to uncover patterns of drug initiation sequences.

By its very nature, data partitioning is computationally complex and particularly so for large datasets. Methods such as K-means clustering circumvent this complexity by approximating the optimal partitioning of data through heuristic techniques. However, partitions produced by KMC vary based upon starting conditions, and even if an optimal partition of a survey could be obtained, it risks overfitting the survey at the expense of accurately partitioning the represented population. This study produces a new method for stability enhanced K-means clustering (SEKMC) that creates partitions based upon observation relationships that persist across multiple applications of KMC to bootstrap samples of the data. This method generates millions of observation pairs and as such risks intractability. However, two practical techniques demonstrated here make SEKMC possible. One is the use of cluster computing to distribute expensive operations across multiple machines. Another is the development of a novel algorithm that leverages the unique structure of the observation pair data to create completely connected graph components that correspond to clusters in the data partition, in $O(V)$ time.

Applying SEKMC to the NSDUH data creates a partition that generates hypotheses for research. The idea that individuals who have used drugs beyond tobacco or alcohol may be more susceptible to prescription drug abuse is very rel-

evant to the opioid epidemic, which has been fueled in part due to over-prescription of pain killers. The existence of clusters whose common defining characteristic is the use of illicit drugs in addition to tobacco, alcohol, and marijuana, leads to a hypothesis that some combination of AFUs of those quasi-legal drugs may point to subsequent use of more dangerous ones. This is a form of the Gateway Hypothesis, which maintains that illicit drug use follows, and necessarily doesn't precede, use of quasi-legal drugs. This hypothesis is tested via two machine learning methods, decision tree classification and logistic regression. The methods both boldly point to a link between early AFU of marijuana and later illicit drug use. This finding reinforces other studies from the literature and together they demand the attention of health care policy makers.

Cross-tabulation and χ^2 analysis provide a simple but effective means of uncovering other hypotheses regarding drug use behavior. By examining the frequencies of demographic features across the NSDUH clusters generated by SEKMC, it is shown that some patterns worthy of further study are present in the data. For example, the lack of any drug use is more common among certain ethnicities and education levels. This finding should be explained through literature review and if necessary additional study.

Opportunities for further research have arisen from this work. The most important is to apply the SEKMC method to youth respondents of NSDUH. It is possible that given the state of legalization in the United States, marijuana may supplant tobacco or alcohol as an early AFU drug among users coming of age recently. Based upon aforementioned findings, an increase in early use of marijuana can certainly lead to elevated use of more dangerous drugs. Another area for consideration is to develop the AFU sequencing partitions into models based upon transition probabilities. Finally, the findings from this study should be

shared with health care practitioners, who can use them to create more effective mitigation policies.

Appendix A

Determining connected components in a graph whose components are complete

Define a graph G as a set of n vertices V and a set of m edges E. Each edge $e_{uv} \in E$ connects two vertices $u, v \in V$. A connected component C of G is a subset of the graph for which every node $c_i \in C$ is reachable by every other node c_j via a path of edges. The *breadth first search* (BFS) algorithm is used to find all of the connected components of a graph, and it executes with time complexity $O(V + E)$.

A component C is complete if every node in the component is connected to every other by a single edge. This construction of graph G as described in Section 2.2 results in a special case where all of the connected components of G are complete. In this case, the connected components can be determined with time complexity $O(N)$. First, a review of the BFS method of finding connected components of any graph is presented. Then the modified algorithm for the special

case for connected components that are all complete is developed.

The *breadth first search* (BFS) algorithm finds the distance between any two connected vertices of a graph. Cormen et al. (2009, p.595) provide an efficient version of BFS (Algorithm 4). The efficiency of BFS is due in part to the use of adjacency lists to store the graph. BFS also includes a concept of 'distance' between nodes. Distance is defined as the number of edges traversed from one node to another, and is discovered during algorithm execution. Distance from each node $u \in V$ to an initial node s is denoted as $u.d$. The algorithm uses colors to indicate whether a node has been examined, or *visited*, and whether or not all of its neighbors have been visited:

- u.color = WHITE: not visited
- u.color = GREY: visited but neighbors not yet visited
- u.color = BLACK: visited with all of its neighbors visited

For each visited node it also stores the node visited one step prior, $u.\pi$. When first visited, a node is placed into first-in, first-out queue Q. When all neighbors of a node are visited, it is removed from Q.

When BFS is presented with a graph G containing several connected components, it initiates with a node s_0 of one component, C_0. BFS then traverses C_0 and determines the distances from s_0 to each vertex $u \in C_0$. If BFS has not yet visited all the vertices of G, we select an unvisited node s_1 and explore its component C_1. This process continues until all vertices of G have been visited and all of its components have been listed and explored.

BFS runs in $O(V + E)$ time. Initialization is $O(1)$ for each vertex, totaling $O(V)$. As it runs, BFS scans the adjacency list for each vertex in the graph. The

Algorithm 4 Breadth First Search Algorithm

Require: $G = (V, E), s : s.\text{color} = \text{WHITE}, s.d = 0, s.\pi = \emptyset$

```
for u ∈ G.V − s do:
    u.color = WHITE
    u.d = ∞
    u.π = ∅
end for
Q = ∅
Q = Q + s
while Q ≠ ∅ do:
    Pop the node from beginning of Q and denote it u
    for v in the adjacency list of u do:
        if v.color = WHITE then:
            v.color = GREY
            v.d = u.d + 1
            v.π = u
            Add v to end of Q
        end if
    end for
    u.color = BLACK
end while
```

total length of all adjacency lists is equal to the number of edges in E, so the total complexity for BFS is $O(V + E)$. While BFS runs in linear time, it can become complex for a large graph with many edges. For a complete graph, $E \sim V^2$, forcing BFS to run at $O(V^2)$.

The special case algorithm is now developed. The minimum distance d between any pair of vertices (u, v) in a graph G with only completely connected components is either $d = 1$ or $d = \infty$ (u cannot be reached from v and vice versa). The completely connected component C of G necessarily consists of a vertex set for which the distance $d_{u,v} = 1, (u, v) \in C$. This means that when any node $u \in G$ is selected, its adjacency list completely defines the component to which u belongs. Therefore, all of the components of G are defined as follows. Consider all vertices of G as unvisited. Select a node u_0 from G and mark it as visited by setting a label u_0.component $= 0$. Similarly mark all nodes v in the adjacency list of u_0 as visited, v.component $= 0$. This set forms the first cluster C_0. Next select an unvisited node $u_1 \in G$ and repeat the process to determine C_1, each of whose members are labeled v.component $= 1$. Continue until there are no more unvisited nodes in G. This method is summarized in Algorithm 5 below. This algorithm initiates in $O(V)$, just like BFS. During progression, it visits every vertex exactly once. The execution takes $O(V)$ time, and the total complexity of the algorithm is the sum of initiation and execution: $O(2V) \sim O(V)$.

By its definition, the graph formed in Section 2.3.2 from the pairs of observations ($\mathbf{U}'$) contains only complete connected components. We can therefore use Algorithm 5 to find the clusters of observations from the pair set $\mathbf{U}'$.

Algorithm 5 Complete Connected Component Enumeration

Require: $G = (V, E), c = 0, Q = \emptyset$

for $u \in V$ **do**:
    u.component $= \emptyset$
    $Q = Q + u$
end for
while $Q \neq \emptyset$ **do**:
    Pop u from beginning of Q
    u.component $= c$
    for v in the adjacency list of u **do**:
        v.component $= c$
        $Q = Q - v$
    end for
    $c = c + 1$
end while

Appendix B

NSDUH Demographic χ^2 Results

Cluster Distribution of SVCFLAG

χ^2 = 4,133,495 ; p-value=0.0

Cluster	Military service	No military service
0	0.198615	0.335873
1	0.248375	0.290806
2	0.180107	0.103705
3	0.114473	0.034583
4	0.063040	0.065446
5	0.049440	0.027745
6	0.041456	0.054121
7	0.035161	0.029516
8	0.028108	0.022177
9	0.025267	0.023300
10	0.002336	0.001165
11	0.012950	0.011329
12	0.000673	0.000234

Table B.1: Cluster Distribution for Military Service

Cluster Distribution of CATAG6

χ^2 = 22,315,193 ; p-value=0.0

Cluster	18-25	26-34	35-49	50-64	65+
0	0.118700	0.159363	0.202303	0.186352	0.353664
1	0.233993	0.276418	0.271379	0.283595	0.183271
2	0.180370	0.170987	0.178017	0.157881	0.183904
3	0.157526	0.082501	0.094573	0.089176	0.131047
4	0.046682	0.081499	0.086081	0.074589	0.018608
5	0.092683	0.051303	0.035675	0.047876	0.027871
6	0.020029	0.039665	0.041543	0.070760	0.026564
7	0.056328	0.052712	0.034234	0.027827	0.014796
8	0.038530	0.046966	0.025142	0.027129	0.008469
9	0.023407	0.013544	0.016656	0.024125	0.046669
10	0.006694	0.001638	0.000935	0.002427	0.001006
11	0.024057	0.022401	0.012687	0.007798	0.003994
12	0.001000	0.001004	0.000775	0.000464	0.000136

Table B.2: Cluster Distribution for Age

Cluster Distribution of IRSEX

χ^2 = 8,795,318 ; p-value=0.0

Cluster	Male	Female
0	0.234781	0.188356
1	0.273990	0.231738
2	0.130683	0.213143
3	0.078396	0.134462
4	0.087121	0.040989
5	0.034923	0.059271
6	0.051216	0.034517
7	0.034528	0.034785
8	0.032216	0.023261
9	0.024791	0.025375
10	0.002242	0.002223
11	0.014339	0.011377
12	0.000774	0.000504

Table B.3: Cluster Distribution for Gender

Cluster Distribution of IRMARIT

χ^2 = 12,701,506 ; p-value=0.0

Cluster	Married	Widowed	Divorced/Sep	Never Married
0	0.244586	0.328982	0.198370	0.132500
1	0.253497	0.159164	0.283686	0.253353
2	0.185621	0.184273	0.138687	0.165837
3	0.103831	0.171052	0.072046	0.117852
4	0.053737	0.018872	0.077934	0.082157
5	0.039010	0.017831	0.042220	0.071229
6	0.038143	0.026106	0.072257	0.039601
7	0.030136	0.014249	0.034944	0.046718
8	0.017581	0.012274	0.037649	0.043694
9	0.022585	0.060003	0.028940	0.020678
10	0.001275	0.001765	0.001162	0.004546
11	0.009576	0.005368	0.011273	0.020803
12	0.000423	0.000061	0.000832	0.001033

Table B.4: Cluster Distribution for Marital Status

Cluster Distribution of NEWRACE2

$\chi^2 = 24{,}118{,}315$; p-value=0.0

Cluster	NonHisp White	NonHisp Black	NonHisp NatAmer	NonHisp HI/PI	NonHisp Asian	NonHisp> 1 Race	Hispanic
0	0.231037	0.150443	0.174794	0.186124	0.159756	0.155778	0.200745
1	0.276678	0.274540	0.314000	0.236976	0.121029	0.301913	0.177585
2	0.144649	0.193007	0.086919	0.162845	0.293975	0.107524	0.240124
3	0.061317	0.167364	0.108035	0.203263	0.304652	0.072369	0.177404
4	0.080858	0.015647	0.074935	0.037725	0.020092	0.096663	0.040645
5	0.041987	0.080289	0.034964	0.043009	0.029697	0.062296	0.050403
6	0.048554	0.028698	0.050194	0.026868	0.015393	0.069908	0.035934
7	0.040487	0.027728	0.039663	0.027419	0.010825	0.038782	0.024693
8	0.034131	0.013234	0.034687	0.020635	0.008766	0.042990	0.017189
9	0.023073	0.032910	0.055697	0.026224	0.028550	0.027729	0.024726
10	0.001302	0.006665	0.005020	0.009370	0.000970	0.002116	0.002814
11	0.015360	0.008524	0.018841	0.017190	0.005825	0.021416	0.007097
12	0.000567	0.000949	0.002252	0.002351	0.000471	0.000516	0.000640

Table B.5: Cluster Distribution for Ethnicity

Cluster Distribution of EDUHIGHCAT

χ^2 = 10,411,603 ; p-value=0.0

Cluster	<High School	HS Grad	Some Coll	Coll Grad
0	0.220598	0.221670	0.194191	0.214550
1	0.188000	0.242733	0.274426	0.262611
2	0.160339	0.156388	0.165487	0.199046
3	0.213918	0.127647	0.079564	0.077539
4	0.040498	0.056191	0.074157	0.066958
5	0.028564	0.039638	0.055201	0.053504
6	0.033596	0.046220	0.048438	0.037571
7	0.021607	0.028025	0.039824	0.039838
8	0.022095	0.028376	0.032881	0.023972
9	0.057037	0.038259	0.017245	0.010178
10	0.004020	0.003129	0.002048	0.001027
11	0.008655	0.011200	0.015886	0.012674
12	0.001073	0.000523	0.000652	0.000534

Table B.6: Cluster Distribution for Education

Cluster Distribution of IRWRKSTAT

χ^2 = 10,507,306 ; p-value=0.0

Cluster	Empl Full Time	Empl Part Time	Unemployed	Other
0	0.195913	0.172710	0.144428	0.256347
1	0.283926	0.247863	0.248617	0.206705
2	0.170883	0.184146	0.143851	0.176556
3	0.070877	0.114371	0.148107	0.154097
4	0.076212	0.068444	0.078365	0.039912
5	0.049781	0.068094	0.049781	0.035784
6	0.048404	0.039498	0.048231	0.034348
7	0.039834	0.039620	0.041360	0.024128
8	0.031952	0.029739	0.041882	0.018380
9	0.015187	0.017969	0.028232	0.042299
10	0.001620	0.002520	0.004743	0.002713
11	0.014789	0.014147	0.021243	0.008238
12	0.000620	0.000880	0.001160	0.000492

Table B.7: Cluster Distribution for Employment Status

Cluster Distribution of GOVTPROG

χ^2 = 2,183,514 ; p-value=0.0

GOVTPROG labels	Yes	No
0	0.172752	0.218838
1	0.250681	0.252438
2	0.137081	0.181051
3	0.138652	0.100764
4	0.073883	0.060995
5	0.044585	0.048144
6	0.049652	0.041073
7	0.035905	0.034397
8	0.037486	0.025479
9	0.040105	0.021903
10	0.003915	0.001875
11	0.013936	0.012566
12	0.001369	0.000478

Table B.8: Cluster Distribution for Government Assistance

Cluster Distribution of INCOME

χ^2 = 6,250,598 ; p-value=0.0

Cluster	<20k	20k-49999	50k-74999	>75k
0	0.183864	0.219218	0.217367	0.212691
1	0.214517	0.227123	0.257783	0.284005
2	0.151944	0.176179	0.178875	0.177703
3	0.171003	0.127580	0.095881	0.070942
4	0.057001	0.060876	0.063345	0.067556
5	0.044010	0.043625	0.047487	0.051898
6	0.043464	0.039838	0.042787	0.044190
7	0.034507	0.030224	0.032901	0.038779
8	0.033132	0.026526	0.028887	0.025583
9	0.048811	0.032260	0.020486	0.011891
10	0.003783	0.002847	0.001418	0.001467
11	0.012775	0.013192	0.012099	0.012816
12	0.001188	0.000513	0.000685	0.000479

Table B.9: Cluster Distribution for Income

Cluster Distribution of COUTYP4

χ^2 = 2,497,920 ; p-value=0.0

Cluster	Large Metro	Small Metro	Nonmetro
0	0.192035	0.220246	0.264282
1	0.246214	0.259700	0.259461
2	0.186398	0.164771	0.140155
3	0.115595	0.097469	0.096105
4	0.066896	0.062511	0.050533
5	0.053806	0.044314	0.029601
6	0.042266	0.043540	0.041774
7	0.033889	0.036756	0.033301
8	0.027056	0.028822	0.027053
9	0.020193	0.025733	0.042956
10	0.002499	0.002055	0.001559
11	0.012688	0.013268	0.012302
12	0.000465	0.000815	0.000917

Table B.10: Cluster Distribution for County Type

Cluster Distribution of AIIND102

χ^2 = 164,129 , p-value=0.0

Cluster	Amer Ind Area	Not Amer Ind Area
0	0.256769	0.210144
1	0.256870	0.252066
2	0.126620	0.173973
3	0.100353	0.107498
4	0.052869	0.063392
5	0.036400	0.047669
6	0.041077	0.042596
7	0.037427	0.034624
8	0.026097	0.027603
9	0.047704	0.024790
10	0.002645	0.002227
11	0.013728	0.012794
12	0.001440	0.000623

Table B.11: Cluster Distribution for Respondent Located in Indian Area

Top Five χ^2 Influencers by Demographic Field

TABLE	VALUE	CLUSTER	OBSERVED	EXPECTED	DIFF	INDCHI2	INFLUENCE	rank
SVCFLAG	No military service	0	4.484637e+07	4.758901e+07	-2.742643e+06	1.580636e+05	0.038240	5.0
SVCFLAG	Military service	0	7.362860e+06	4.620217e+06	2.742643e+06	1.628081e+06	0.393875	1.0
SVCFLAG	Military service	2	2.273377e+06	3.800010e+06	-1.526634e+06	6.133169e+05	0.148377	3.0
SVCFLAG	Military service	3	7.581186e+05	2.354449e+06	-1.596330e+06	1.082321e+06	0.261842	2.0
SVCFLAG	Military service	5	6.082135e+05	1.041713e+06	-4.334991e+05	1.803966e+05	0.043643	4.0
CATAG6	18-25	0	4.054974e+06	7.199927e+06	-3.144953e+06	1.373726e+06	0.061560	4.0
CATAG6	65+	0	1.787522e+07	1.065251e+07	7.222706e+06	4.897200e+06	0.219456	1.0
CATAG6	65+	4	9.405211e+05	3.196994e+06	-2.256473e+06	1.592643e+06	0.071370	2.0
CATAG6	18-25	5	3.166174e+06	1.623355e+06	1.542818e+06	1.466277e+06	0.065708	3.0
CATAG6	50-64	6	4.413909e+06	2.655834e+06	1.758075e+06	1.163787e+06	0.052152	5.0
IRSEX	Male	2	1.562361e+07	2.072415e+07	-5.100538e+06	1.255323e+06	0.142726	1.0
IRSEX	Female	2	2.731715e+07	2.221661e+07	5.100538e+06	1.170993e+06	0.133138	2.0
IRSEX	Male	3	9.372546e+06	1.284048e+07	-3.467932e+06	9.366124e+05	0.106490	5.0
IRSEX	Male	4	1.041561e+07	7.562129e+06	2.853482e+06	1.076728e+06	0.122421	3.0
IRSEX	Female	4	5.253237e+06	8.106718e+06	-2.853482e+06	1.004396e+06	0.114197	4.0
IRMARIT	Widowed	0	4.748924e+06	3.042386e+06	1.706539e+06	9.572340e+05	0.075364	2.0
IRMARIT	Never Married	0	9.474925e+06	1.507138e+07	-5.596452e+06	2.078130e+06	0.163613	1.0
IRMARIT	Never Married	5	5.093503e+06	3.398118e+06	1.695385e+06	8.458593e+05	0.066595	3.0
IRMARIT	Divorced/Sep	6	2.473313e+06	1.457354e+06	1.015960e+06	7.082523e+05	0.055761	4.0
IRMARIT	Widowed	9	8.661557e+05	3.622234e+05	5.039323e+05	7.010807e+05	0.055197	5.0
NEWRACE2	NonHisp Asian	2	4.104184e+06	2.420087e+06	1.684097e+06	1.171934e+06	0.048591	4.0
NEWRACE2	NonHisp White	3	9.668107e+06	1.693468e+07	-7.266575e+06	3.118045e+06	0.129281	2.0
NEWRACE2	NonHisp Asian	3	4.253250e+06	1.499462e+06	2.753788e+06	5.057378e+06	0.209690	1.0
NEWRACE2	Hispanic	3	7.098416e+06	4.297504e+06	2.800911e+06	1.825502e+06	0.075689	3.0
NEWRACE2	NonHisp Black	4	4.607117e+05	1.862432e+06	-1.401720e+06	1.054975e+06	0.043742	5.0
EDUHIGHCAT	< High School	3	6.550820e+06	3.289023e+06	3.261797e+06	3.234796e+06	0.310691	1.0
EDUHIGHCAT	Some Coll	3	6.094122e+06	8.226480e+06	-2.132358e+06	5.527214e+05	0.053087	5.0
EDUHIGHCAT	Coll Grad	3	6.148957e+06	8.517258e+06	-2.368301e+06	6.585274e+05	0.063249	4.0
EDUHIGHCAT	< High School	9	1.746638e+06	7.684262e+05	9.782119e+05	1.245271e+06	0.119604	2.0
EDUHIGHCAT	Coll Grad	9	8.071049e+05	1.989918e+06	-1.182813e+06	7.030672e+05	0.067527	3.0
IRWRKSTAT	Other	0	2.106192e+07	1.731654e+07	3.745377e+06	8.100838e+05	0.077097	4.0
IRWRKSTAT	Empl Full Time	3	8.707680e+06	1.319519e+07	-4.487514e+06	1.526145e+06	0.145246	2.0
IRWRKSTAT	Other	3	1.266090e+07	8.824458e+06	3.836443e+06	1.667898e+06	0.158737	1.0
IRWRKSTAT	Other	4	3.279233e+06	5.196978e+06	-1.917746e+06	7.076707e+05	0.067350	5.0
IRWRKSTAT	Other	9	3.475364e+06	2.061690e+06	1.413674e+06	9.693382e+05	0.092254	3.0
GOVTPROG	Yes	0	7.499065e+06	9.149031e+06	-1.649966e+06	2.975603e+05	0.136276	4.0
GOVTPROG	Yes	2	5.950583e+06	7.524844e+06	-1.574261e+06	3.293487e+05	0.150834	3.0
GOVTPROG	Yes	3	6.018804e+06	4.662320e+06	1.356485e+06	3.946644e+05	0.180747	1.0
GOVTPROG	Yes	8	1.627237e+06	1.197349e+06	4.298880e+05	1.543440e+05	0.070686	5.0
GOVTPROG	Yes	9	1.740943e+06	1.089274e+06	6.516683e+05	3.898665e+05	0.178550	2.0
INCOME	>75k	1	2.740434e+07	2.432865e+07	3.075691e+06	3.888367e+05	0.062208	5.0
INCOME	<20k	3	6.734951e+06	4.230081e+06	2.504870e+06	1.483275e+06	0.237301	1.0
INCOME	>75k	3	6.845412e+06	1.036364e+07	-3.518232e+06	1.194363e+06	0.191080	2.0
INCOME	<20k	9	1.922430e+06	9.882891e+05	9.341409e+05	8.829594e+05	0.141260	3.0
INCOME	>75k	9	1.147403e+06	2.421295e+06	-1.273892e+06	6.702203e+05	0.107225	4.0
COUTYP4	Large Metro	0	2.658303e+07	2.917529e+07	-2.592255e+06	2.303246e+05	0.092207	4.0
COUTYP4	Nonmetro	0	9.336549e+06	7.445781e+06	1.890767e+06	4.801379e+05	0.192215	1.0
COUTYP4	Nonmetro	2	4.951408e+06	6.123965e+06	-1.172556e+06	2.245095e+05	0.089879	5.0
COUTYP4	Nonmetro	5	1.045728e+06	1.678788e+06	-6.330598e+05	2.387227e+05	0.095569	3.0
COUTYP4	Nonmetro	9	1.517541e+06	8.864872e+05	6.310541e+05	4.492217e+05	0.179838	2.0
AIIND102	Amer Ind Area	0	8.419815e+05	6.911157e+05	1.508659e+05	3.293300e+04	0.200653	3.0
AIIND102	Amer Ind Area	2	4.152041e+05	5.684250e+05	-1.532209e+05	4.130119e+04	0.251639	2.0
AIIND102	Amer Ind Area	4	1.733655e+05	2.074152e+05	-3.404971e+04	5.589673e+03	0.034057	5.0
AIIND102	Amer Ind Area	5	1.193620e+05	1.558247e+05	-3.646270e+04	8.532208e+03	0.051985	4.0
AIIND102	Amer Ind Area	9	1.564266e+05	8.228353e+04	7.414309e+04	6.680800e+04	0.407046	1.0

Table B.12: Top Five χ^2 Influencers by Demographic Field

References

Adler, I., & Kandel, D. B. (1981, sep). Cross-cultural perspectives on developmental stages in adolescent drug use. *Journal of Studies on Alcohol*, *42*(9), 701–715. Retrieved from http://www.jsad.com/doi/10.15288/jsa.1981.42.701

Adrados, J.-L. R. (1995, jan). The Influence of Family, School, and Peers on Adolescent Drug Misuse. *International Journal of the Addictions*, *30*(11), 1407–1423. Retrieved from http://www.tandfonline.com/doi/full/10.3109/10826089509055840

Advanced Recovery Systems. (2020). *What are Gateway Drugs? Information and Prevention.* Retrieved 2022-09-01, from https://www.drugrehab.com/guides/gateway-drugs/

Aloise, D., Deshpande, A., Hansen, P., & Popat, P. (2009, may). NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, *75*(2), 245–248. Retrieved from http://link.springer.com/10.1007/s10994-009-5103-0

Ategbole, M., Su, B. B., Wang, N., Loudermilk, E., Xie, X., Acevedo, P., ... Wang, K. (2021, apr). Gender differences in the associations of early onset poly tobacco and drug use prior to age 18 with the prevalence of adult bronchitis in the United States. *Journal of Addictive Diseases*, *39*(2), 189–198. Retrieved from https://www.tandfonline.com/doi/full/10.1080/10550887.2020.1847992

Baggio, S., Henchoz, Y., Studer, J., Deline, S., N'goran, A., Mohler-Kuo, M., ... Gmel, G. (2015). Cannabis use and other illicit drug use: Do subjective experiences during first cannabis use increase the probability of using illicit drug? *Journal of Substance Use*, *20*(4), 234–238.

Barry, A. E., King, J., Sears, C., Harville, C., Bondoc, I., & Joseph, K. (2016). Prioritizing Alcohol Prevention: Establishing Alcohol as the Gateway Drug and Linking. *Journal of School Health*, *86*(1), 31–38. Retrieved from www.monitoringthefuture.org.

Beattie, M., & Nicholson, C. (2021). Feature Extraction for Heroin-Use Classification Using Imbalanced Random Forest Methods. *Substance Use and Misuse*, *56*(1), 123–130. Retrieved from https://doi.org/10.1080/10826084.2020.1843058

Ben-Hur, A., Elisseeff, A., & Guyon, I. (2002). A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 6–17.

Blaze-Temple, D., & Lo, S. K. (1992). Stages of drug use: a community survey of Perth teenagers. *British Journal of Addiction*, *87*(2), 215–225.

Blocker, J. S. (2006). Did prohibition really work? Alcohol prohibition as a public health innovation. *American Journal of Public Health*, *96*(2), 233–243.

Botvin, G., Griffin, K., Diaz, T., & Ifill-Williams, M. (2001). Drug Abuse Prevention Among Minority Adolescents: Posttest and One-Year Follow-Up of a School-Based Preventive Intervention. *Prevention Science*, *2*, 1–13.

Botvin, G. J., Griffin, K. W., Diaz, T., Scheier, L. M., Williams, C., & Epstein, J. A. (2000). Preventing illicit drug use in adolescents: Long-term follow-up data from a randomized control trial of a school population. *Addictive Behaviors*, *25*(5), 769–774.

Campo, J., Jetzer, K.-A., Mounts, T., Paterson, T., & Schmidt, J. (2016). *Monitoring Impacts of Recreational Marijuana Legalization: 2015 Update* (Tech. Rep.). Forecasting and Research Division, Washington State Office of Financial Management. Retrieved from https://ofm.wa.gov/sites/default/files/public/legacy/reports/marijuana{_}impacts{_}update{_}2015.pdf

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, *40*(1), 200–210. Retrieved from http://dx.doi.org/10.1016/j.eswa.2012.07.021

Center for Behavioral Health Statistics and Quality. (2020a). *2018 National Survey on Drug Use and Health (NSDUH) Methodological Resource Book, Section 10: Editing and Imputation Report* (Tech. Rep.). Rockville, Maryland: Substance Abuse and Mental Health Services Administration.

Center for Behavioral Health Statistics and Quality. (2020b). *2019 National Survey on Drug Use and Health Public Use File Codebook* (Tech. Rep.). Rockville, Maryland: Substance Abuse and Mental Health Services Administration.

Centers for Disease Control and Prevention, N. C. f. I. P., & Control. (2021). *Opioid Basics.* Retrieved from https://www.cdc.gov/opioids/basics/epidemic.html

Cerdá, M., Santaella, J., Marshall, B. D., Kim, J. H., & Martins, S. S. (2015). Nonmedical Prescription Opioid Use in Childhood and Early Adolescence Predicts Transitions to Heroin Use in Young Adulthood: A National Study. *Journal of Pediatrics*, *167*(3), 605–612.e2.

Chandra, S., Radwan, M. M., Majumdar, C. G., Church, J. C., Freeman, T. P., & ElSohly, M. A. (2019). New trends in cannabis potency in USA and Europe during the last decade (2008–2017). *European Archives of Psychiatry and Clinical Neuroscience*, *269*(1), 5–15. Retrieved from http://dx.doi.org/10.1007/s00406-019-00983-5

Cicero, T. J., Ellis, M. S., Surratt, H. L., & Kurtz, S. P. (2014). The changing face of heroin use in the United States a retrospective analysis of the past 50 years. *JAMA Psychiatry*, *71*(7), 821–826.

Compton, W. M., Gfroerer, J., Conway, K. P., & Finger, M. S. (2014, sep). Unemployment and substance outcomes in the United States 2002–2010. *Drug and Alcohol Dependence*, *142*, 350–353. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S037687161400920X

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to Algorithms, Third Edition* (3rd ed.). Cambridge, MA: The MIT Press.

Darke, S., Kaye, S., & Torok, M. (2012). Age-related patterns of drug use initiation among polydrug using regular psychostimulant users. *Drug and Alcohol Review*, *31*(6), 784–789.

Dasgupta, S. (2007). *The hardness of k-means clustering. Technical Report CS2007-0890* (Tech. Rep.). San Diego, CA: University of California.

Degenhardt, L., Chiu, W. T., Conway, K., Dierker, L., Glantz, M., Kalaydjian, A., ... Kessler, R. C. (2009). Does the gateway matter? Associations between the order of drug use initiation and the development of drug dependence in the National Comorbidity Study Replication. *Psychological Medicine*, *39*(1), 157–167.

Dong, G., & Pei, J. (2007). *Sequence Data Mining.* New York, NY: Springer.

Drineas, P., Frieze, A., Kannan, R., Vempala, S., & Vinay, V. (2004, jul). Clustering Large Graphs via the Singular Value Decomposition. *Machine Learning*, *56*(1-3), 9–33. Re-

trieved from http://link.springer.com/10.1023/B:MACH.0000033113.59016.96

Drug Abuse Resistance Education Program. (2021). *Marijuana legalization talking points.* Retrieved 2022-09-01, from https://dare.org/marijuana-legalization-talking-points/

Dupont, R. L. (1984). *Getting Tough on Gateway Drugs: A Guide for the Family.* Washington, D.C.: American Psychiatric Press.

Ellgren, M., Spano, S. M., & Hurd, Y. L. (2007). Adolescent cannabis exposure alters opiate intake and opioid limbic neuronal populations in adult rats. *Neuropsychopharmacology*, *32*(3), 607–615.

Fergusson, D. M., & Boden, J. M. (2008). Cannabis use and later life outcomes. *Addiction*, *103*(6), 969–976.

Fergusson, D. M., Boden, J. M., & Horwood, L. J. (2006). Cannabis use and other illicit drug use: Testing the cannabis gateway hypothesis. *Addiction*, *101*(4), 556–569.

Fergusson, D. M., Boden, J. M., & Horwood, L. J. (2008). The developmental antecedents of illicit drug use: Evidence from a 25-year longitudinal study. *Drug and Alcohol Dependence*, *96*(1-2), 165–177.

Fiellin, L. E., Tetrault, J. M., Becker, W. C., Fiellin, D. A., & Hoff, R. A. (2013). Previous use of alcohol, cigarettes, and marijuana and subsequent abuse of prescription opioids in young adults. *Journal of Adolescent Health*, *52*(2), 158–163. Retrieved from http://dx.doi.org/10.1016/j.jadohealth.2012.06.010

Fleming, R., Leventhal, H., Glynn, K., & Ershler, J. (1989). The role of cigarettes in the initiation and progression of early substance use. *Addictive Behaviors*, *14*(3), 261–272.

Fuller, C. M., Borrell, L. N., Latkin, C. A., Galea, S., Ompad, D. C., Strathdee, S. A., & Vlahov, D. (2005). Effects of race, neighborhood, and social network on age at initiation of injection drug use. *American Journal of Public Health*, *95*(4), 689–695.

Golub, A., & Johnson, B. D. (1994). The shifting importance of alcohol and marijuana as gateway substances among serious drug abusers. *Journal of Studies on Alcohol*, *55*(5), 607–614.

Golub, A., Johnson, B. D., & Labouvie, E. (2000). On Correcting Biases in Self-Reports of Age at First Substance Use with Repeated Cross-Section Analysis Author (s): Andrew Golub , Bruce D . Johnson and Erich Labouvie Source : Journal of Quantitative Criminology

, March 2000 , Vol . 16 , No . 1 , Speci. *Journal of Quantitative Criminology*, *16*(1), 45–68.

Gould, P. (1981, jun). Letting the data speak for themselves. *Annals of the Association of American Geographers*, *71*(2), 166–176. Retrieved from http://www.tandfonline.com/doi/abs/10.1111/j.1467-8306.1981.tb01346.x

GraphFrames. (2021). Retrieved from https://graphframes.github.io/graphframes/docs/{_}site/index.html

Guttman, L. (1944, apr). A Basis for Scaling Qualitative Data. *American Sociological Review*, *9*(2), 139. Retrieved from http://www.jstor.org/stable/2086306?origin=crossref

Hamburg, Beatrin A.; Kraemer, Helena C.; Jahnke, W. (1975, nov). A hierarchy of drug use in adolescence: behavioral and attitudinal correlates of substantial drug use. *American Journal of Psychiatry*, *132*(11), 1155–1163. Retrieved from http://psychiatryonline.org/doi/abs/10.1176/ajp.132.11.1155

Han, B. H., Moore, A. A., Sherman, S., Keyes, K. M., & Palamar, J. J. (2017, jan). Demographic trends of binge alcohol use and alcohol use disorders among older adults in the United States, 2005–2014. *Drug and Alcohol Dependence*, *170*, 198–207. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0376871616309978

Harrington, M., Baird, J., Lee, C., Nirenberg, T., Longabaugh, R., Mello, M. J., & Woolard, R. (2012). Identifying subtypes of dual alcohol and marijuana users: A methodological approach using cluster analysis. *Addictive Behaviors*, *37*(1), 119–123. Retrieved from http://dx.doi.org/10.1016/j.addbeh.2011.07.016

Hawkins, J. D., Catalano, R. F., & Miller, J. Y. (1992). Risk and Protective Factors for Alcohol and Other Drug Problems in Adolescence and Early Adulthood: Implications for Substance Abuse Prevention. *Psychological Bulletin*, *112*(1), 64–105.

Ingraham, C. (2016, jan). *The real 'gateway drug' is 100% legal.* Washington, D.C.. Retrieved from https://www.washingtonpost.com/news/wonk/wp/2016/01/06/the-real-gateway-drug-thats-everywhere-and-legal/

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data (Prentice Hall Advanced Reference Series : Computer Science)* (1st ed.). Pearson College Division.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical*

Learning (Vol. 103). New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-1-4614-7138-7

Jansen, M. A. (1992). *A Promising Future: Alcohol and Other Drug Problem Prevention Services Improvement*. Rockville, Maryland: U.S. Dept. of Health and Human Services, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, Office for Substance Abuse Prevention.

Jones, C. M. (2013, sep). Heroin use and heroin use risk behaviors among nonmedical users of prescription opioid pain relievers. United States, 2002 to 2004 and 2008 to 2010. *Drug and Alcohol Dependence*, *132*(1-2), 95–100. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0376871613000197

Jones, C. M., Logan, J., Gladden, . R. M., & Bohm, M. K. (2015). Morbidity and Mortality Weekly Report Vital Signs: Demographic and Substance Use Trends Among Heroin Users — United States, 2002–2013.

Kandel, D. (1975, nov). Stages in adolescent involvement in drug use. *Science*, *190*(4217), 912–914. Retrieved from https://www.science.org/doi/10.1126/science.1188374

Kandel, D. B., Ph, D., & Chen, K. (1992). Stages of Progression in Drug Involvement from Adolescence to Adulthood: Further Evidence for the Gateway Theory*. *Journal of Studies on Alcohol*, 447–457.

Kandel, D. B., & Yamaguchi, K. (2002a). Log Linear Sequence Analyses: Gender and Racial/Ethnic Difference in Drug Use Progression. In *Examining the gateway hypothesis: Stages and pathways of drug involvement* (chap. 9). Cambridge, UK: Cambridge University Press.

Kandel, D. B., & Yamaguchi, K. (2002b). Stages of Drug Involvement in the U.S. Population. In D. B. Kandel (Ed.), *Stages and pathways of drug involvement: Examining the gateway hypothesis* (chap. 9). Cambridge, UK: Cambridge University Press.

Kandel, E. R., & Kandel, D. B. (2014). A Molecular Basis for Nicotine as a Gateway Drug. *New England Journal of Medicine*, *371*(10), 932–943.

Keyes, K. M., Hamilton, A., & Kandel, D. B. (2016). Birth cohorts analysis of adolescent cigarette smoking and subsequent marijuana and cocaine use. *American Journal of Public Health*, *106*(6), 1143–1149.

Lee, J., & Petlakh, K. (2020). Progression to drug use from adolescent initiation of marijuana

among South Korean inmates: a propensity score matching technique. *International Journal of Comparative and Applied Criminal Justice*, *44*(3), 221–230.

Levine, E., & Domany, E. (2001, nov). Resampling method for unsupervised estimation of cluster validity. *Neural computation*, *13*(11), 2573–2593. Retrieved from `https://pubmed.ncbi.nlm.nih.gov/11674852/`

Liew, H. (2016, jan). Is There Something Unique About Marriage? The Relative Impact of Marital Status on Alcohol Consumption Among Military Personnel. *Journal of Divorce & Remarriage*, *57*(1), 76–85. Retrieved from `http://www.tandfonline.com/doi/full/10.1080/10502556.2015.1088126`

Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137.

Lynskey, M. T., & Agrawal, A. (2018). Denise Kandel's classic work on the gateway sequence of drug acquisition. *Addiction*, *113*(10), 1927–1932.

Lynskey, M. T., Heath, A. C., Bucholz, K. K., Slutske, W. S., Madden, P. A., Nelson, E. C., ... Martin, N. G. (2003). Escalation of drug use in early-onset cannabis users vs co-twin controls. *Journal of the American Medical Association*, *289*(4), 427–433.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, *1*(14), 281–297.

Mahajan, M., Nimbhorkar, P., & Varadarajan, K. (2012, jul). The planar k-means problem is NP-hard. *Theoretical Computer Science*, *442*, 13–21. Retrieved from `https://linkinghub.elsevier.com/retrieve/pii/S0304397510003269`

Morrison, V., & Plant, M. (1991). Licit and illicit drug initiations and alcohol-related problems amongst illicit drug users in Edinburgh. *Drug and Alcohol Dependence*, *27*(1), 19–27.

Office of National Drug Control Policy. (2002). *The Economic Costs of Drug Abuse in the United States: Estimates for States and Selected Metropolitan Areas, 2002* (Tech. Rep.). Washington, D.C.: The White House.

Oh, H., Yamazaki, Y., & Kawata, C. (1998, sep). [Prevalence and a drug use development model for the study of adolescent drug use in Japan]. *[Nihon koshu eisei zasshi] Japanese journal of public health*, *45*(9), 870–82. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/9847560`

Panlilio, L. V., Stull, S. W., Bertz, J. W., Burgess-Hull, A. J., Kowalczyk, W. J., Phillips, K. A., ... Preston, K. L. (2020, nov). Beyond abstinence and relapse: cluster analysis of drug-use patterns during treatment as an outcome measure for clinical trials. *Psychopharmacology*, *237*(11), 3369–3381. Retrieved from https://link.springer.com/10.1007/s00213-020-05618-5

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Prusoff, B. A., Merikangas, K. R., & Weissman, M. M. (1988). Lifetime prevalence and age of onset of psychiatric disorders: Recall 4 years later. *Journal of Psychiatric Research*, *22*(2), 107–117.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65.

Scott, I. (1998). A hundred-year habit. *History Today*, *48*(6), 6–8. Retrieved from http://web.b.ebscohost.com.ezproxy.lib.ou.edu/ehost/pdfviewer/pdfviewer?vid=3{&}sid=31afa7d5-ef8c-4e8f-83f6-a03ef5e8ed25{%}40sessionmgr102

Selim, S. Z., & Ismail, M. A. (1984, jan). K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(1), 81–87. Retrieved from http://ieeexplore.ieee.org/document/4767478/

Sevigny, E. L., & Coontz, P. D. (2008). Patterns of substance involvement and criminal behavior: A gender-based cluster analysis of Pennsylvania arrestees. *International Journal of Offender Therapy and Comparative Criminology*, *52*(4), 435–453.

Substance Abuse and Mental Health Administration. (2019a). *National Survey on Drug Use and Health* (Tech. Rep.). Rockville, Maryland: US Centers for Disease Control.

Substance Abuse and Mental Health Administration. (2019b). *Treatment Episode Data Set (TEDS)* (Tech. Rep.). Rockville, Maryland: US Centers for Disease Control.

Texas Health and Human Services. (n.d.). *Teachable Moments: Facts About Marijuana.* Retrieved 2022-09-01, from https://www.hhs.texas.gov/sites/default/

files/documents/services/health/drug-free-texas/marijuana-fact-sheet.pdf

Tibshirani, R., & Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, *14*(3), 511–528.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423. Retrieved from https://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00293

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525.

Tseng, G. C., & Wong, W. H. (2005, mar). Tight Clustering: A Resampling-Based Approach for Identifying Stable and Tight Patterns in Data. *Biometrics*, *61*(1), 10–16. Retrieved from https://onlinelibrary.wiley.com/doi/10.1111/j.0006-341X.2005.031032.x

Underlying Cause of Death 1999-2019 on CDC WONDER Online Database, released in 2020. Data are from the Multiple Cause of Death Files, 1999-2019. (2020). Retrieved 2021-10-16, from http://wonder.cdc.gov/ucd-icd10.html

US Drug Enforcement Administration. (2021a). *Drug Enforcement Administration: The Early Years.* Retrieved 2022-11-01, from https://www.dea.gov/sites/default/files/2018-05/EarlyYearsp12-29.pdf

US Drug Enforcement Administration. (2021b). *Drug Scheduling.* Retrieved 2022-09-01, from https://www.dea.gov/drug-information/drug-scheduling

Wadekar, A. S. (2020). Understanding Opioid Use Disorder (OUD) using tree-based classifiers. *Drug and Alcohol Dependence*, *208*(January), 15–19.

Wagner, F. A., & Anthony, J. C. (2002). From first drug use to drug dependence: Developmental periods of risk for dependence upon marijuana, cocaine, and alcohol. *Primary Care Companion to the Journal of Clinical Psychiatry*, *4*(1), 33.

Wang, N., Ouedraogo, Y., Chu, J., Liu, Y., Wang, K., & Xie, X. (2019, sep). Variable reduction for past year alcohol and drug use in unmet need for mental health services among US adults. *Journal of Affective Disorders*, *256*, 110–116. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0165032719305427

Welte, J. W., & Barnes, G. M. (1985, mar). Alcohol: The gateway to other drug use among

secondary-school students. *Journal of Youth and Adolescence*, *14*(6), 487–498. Retrieved from http://link.springer.com/10.1007/BF02139522

Wilkerson, R. G., Kim, H. K., Windsor, T. A., & Mareiniss, D. P. (2016, may). The Opioid Epidemic in the United States. *Emergency Medicine Clinics of North America*, *34*(2), e1–e23. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S073386271500108X

Wittchen, H. U., Ahmoi Essau, C., Hecht, H., Teder, W., & Pfister, H. (1989). Reliability of life event assessments: test-retest reliability and fall-off effects of the Munich interview for the assessment of life events and conditions. *Journal of Affective Disorders*, *16*(1), 77–91.

Xie, Z., Tanner, R., Striley, C. L., & Marlow, N. M. (2022, feb). Association of functional disability with mental health services use and perceived unmet needs for mental health care among adults with serious mental illness. *Journal of Affective Disorders*, *299*, 449–455. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0165032721013628

Yan, M., & Ye, K. (2007, dec). Determining the Number of Clusters Using the Weighted Gap Statistic. *Biometrics*, *63*(4), 1031–1037. Retrieved from https://onlinelibrary.wiley.com/doi/10.1111/j.1541-0420.2007.00784.x

Young Mun, E., Windle, M., & Schainker, L. M. (2008). A model-based cluster analysis approach to adolescent problem behaviors and young adult outcomes. *Development and Psychopathology*, *20*, 291–318. Retrieved from https://www.cambridge.org/core.

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*, *2*(2), 226–235.

Zhang, S., Wu, S., Wu, Q., Durkin, D. W., & Marsiglia, F. F. (2021). Adolescent drug use initiation and transition into other drugs: A retrospective longitudinal examination across race/ethnicity. *Addictive Behaviors*, *113*(September 2020).