Does Personality Pick the Poison?

A Story Crafted by Machine Learning Methods

Kelsey Cherland

A final project presented for
Math 533: Statistical Learning

Dr. Sam Behseta
California State University, Fullerton
10 December 2021

Abstract

In this project I will analyze the relationship between drug use and psychometric measures like the Five Factor Model. The data was obtained via the UCI Repository for Machine Learning. This data allows us to understand who is motivated to use drugs, and previous literature suggests that if we understand the personality traits of those who use drugs, we might be able to address drug addiction more directly. The objectives of this analysis are to identify the clusters of the data, create models for predicting whether someone is a user of a drug in the past decade, and identity traits associated with specific drugs. The methods used for this analysis were ridge regression and 2 Neural Networks. The ridge regression performed the best overall. The K-Means Clustering was especially insightful for identity traits associated with specific drugs.

1.0 Introduction

Drugs are chemicals, natural or synthetic, that can alter a person's biological function. Drugs can serve a medical purpose, but human beings also seek out drugs for no apparent medical reason. Human beings and a few other species have been known to seek out psychoactive drugs for recreational mind altering experiences (National Institute on Drug Addition 2006). Some research explores the psychological aspects of drug use and addiction in human beings. More specifically, there is research that examines the relationship between personality characteristics and drug use (Section 1.2 Literature Review).

The field of psychology has developed psychometric measures of individual personality traits. These can be normalized measures that help researchers determine whether someone is atypical compared to the general population. The Five Factor Model (FFM) is widely used as a comprehensive system for understanding personality differences for individual people. The FFM model measures Neuroticism (N), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C). More neurotic individuals may experience more negative emotions, like depression and anxiety, with less emotional resilience. Extroverted individuals are more assertive and sociable. Highly open individuals have a broad range of interests and are more interested in novel experiences and art. An agreeable person is more likely to be cooperative and polite than rude and antagonistic. And last, a conscientious person is less distracted and remains orderly in their organization (Soto and Jackson 2020). Combined, these five traits are known to encompass major aspects of a person's personality. There are other psychometric tests that focus on character traits but use a different system of evaluation and measurement. The Barratt Impulsiveness Scale (BIS-11) measures self-reported behavioral aspects of impulsiveness that involve acting without thinking, poor concentration, and a lack of planning. There is also the TCI-R questionnaire discussed in the Literature Review section, which measures Harm Avoidance, Novelty Seeking, Reward Dependence, Persistence, Self-Directedness, Cooperativeness, and Self-Transcendence. Previous research on the connection between trait measure and drug use suggests that certain personality traits are associated with high risk behaviors, including drug consumption and addiction.

Addiction and drug dependency research is a topic of public health concern because drug abuse can negatively impact individuals and communities. The U.S. Department of Health and Human Services estimates that thousands of people die from alcohol and drug overdoses each year. If a heavy user does not die, they will likely suffer long-term health effects like liver disease, cancer, and heart disease. Drug abuse also negatively affects society, causing higher health care costs, spread of infectious diseases, economic and psychological stress in families, and impulsive violence. It is estimated that one in seven people in the United States will develop a substance use disorder during their lifetime (USDHHS 2016). That means, for about 14.6 percent of the U.S. population, what may start out as an occasional fun activity will become an addiction.

These societal statistics suggests that for any given drug, there are people who become addicted, others who are occasional users, and others who remain uninterested in even trying it once. What individual traits might be able to explain these differences? That's what two psychology researchers from the United Kingdom were wondering when they conducted research via an online survey of respondents aged 18 and over from English-speaking countries. The questions probed an individuals personality attributes, demographic information, and illegal and legal drug use (Fehrman and Egan 2016). Together, with researchers in the Mathematics departments at the University of Leicester in Leicester, United Kingdom and the University of Salahaddin in Erbil, Iraq (Kurdistan Region), this data was was analyzed using machine learning techniques to try to identify how individual psychometric traits might coincide with drug use behavior (Fehrman et al. 2017). In this paper, I will use newly learned machine learning techniques to draw inference on Fehrman and Egan's data set.

1.1 Goals for the Project

There are several goals for this project. The goals below will be addressed via statistical learning methods.

0.1 EDA: Clustering (Goal 1)

I will uses the pleiades grouping proposed in Fehrman et al. 2017 to group together drugs with a similar modular structure. The individual drugs in the three pleiades groups are identified in Section 2.0 Features of the Data in Figure 1. The three pleiades are the Ecstasy Pleiade, the Heroin Pleiade, and the Benzodiazepine Pleiade. For the sake of a different comparison, I will also look at alcohol, tobacco / nicotine, and caffeine. The structure of the data for these three drugs is also discussed in Section 2.0 Features of the Data.

Before identifying relationships between traits and drug use, I will use clustering techniques to identify possibly distinct groups in the Fehrman and Egan 2016 data. I will explore PCA, K-Means Clustering, and hierarchical clustering via a Dendrogram. This preliminary step is meant to achieve the goal of reducing dimensions of the data but could provide insight into the natural clustering of variables around certain drug use. The conclusions of the clustering will be discussed in Section 3.0 Exploratory Data Analysis (EDA).

0.2 Goal 2-3: Modeling and Prediction

Goal 2 is to create a model or algorithm that adequately predicts whether an individual will be a drug user of any of the three pleiades and the three legal drugs, Alcohol, Caffeine, and Nicotine, within the past decades. This will be achieved through regression and neural networks. Goal 3 is to identify the directional relationship between character traits and the three pleiades. This will likely be accomplished with regression. The modeling techniques used to accomplish these goals is discussed more thoroughly in Section 4.0 Modeling. The results of these goals will be discussed in Section 5.0 Results and Section 6.0 Discussion.

In the next section, I will discuss the previous literature on character traits and drug use. I will draw comparisons between my modeling, the previous literature, and the analysis completed by hyperlinkFehrmanetalFehrman et al. 2017.

1.2 Literature Review

Etiopathogenesis is the cause and development of a disease or abnormal condition (Merriam-Webster 2021). The research concerning personality in the etiopathogenesis of drug addiction has grown considerably since the 1990s. Researchers have expressed hopes that their insight can better inform ideas to prevent addiction from happening. In this literature review, I will examine the previous research concerning the relationship between personality and drug use for the purpose of later comparing analysis on the data provided by Fehrman and Egan (Fehrman and Egan 2016).

Specific personality traits have emerged as possible contributing factors to the etiopathogenesis of drug addiction. Multiple studies have found that the etiopathogenesis of drug addiction typically coincides with a personality that is high in Neuroticism (N), low in agreeableness (A), low in conscientiousness (C), and high in impulsiveness (Dragan et al. 2012; Terracciano et al. 2008; Dash et al. 2019; Kroencke et al. 2021). There are two different hypotheses as to why these personality traits are more common with people who abuse drug substances. One hypothesis is that the genes that cause higher neuroticism, lower agreeableness, and lower conscientiousness will also increase risk for drug use (correlation not causation). There is some research on the genetic origin of these traits and behaviors (Slutskey et al. 2002) that will not be discussed here. The second hypothesis is that there is a direct causal relationship between these psychometric traits and drug use behaviors. That research is the topic of this brief literature review.

Research shows that different psychological reasons might motivate the choice between one drug over another. Psychoactive drugs, specifically opiates and sedatives, have been shown to relieve the unhappiness in people high in neuroticism (Cooper 1994; Simons et al. 1998). Indeed, high neuroticism was found to be associated with cannabis, sedatives, and stimulant and opioid abuse (Dash et al. 2021). Ecstasy, however, is a drug that produces feelings of increased pleasure and emotional warmth (NIDA 2020), so it has greater appeal in a pro-social environment for those with trait extraversion (Vreeker et al. 2020). Similarly, those who use psychedelic drugs (i.e. cannabis, hallucinogens), which in (Fehrman et al. 2017) were evaluated in the same group as the "ecstasy pleiades," might be seeking out the unique perceptual experiences these drugs induce because these individuals also tend to have greater trait openness to experience (Dash et al. 2019). The research done by Simons et al. 1998 suggests that individuals with greater trait openness to experience tend to seek new experiences that inspire the introspection provided by cannabis, ecstasy, and hallucinogen drugs. Someone who seeks novelty or ranks high in Sensation Seeking is typically using drugs at a higher rate than compared to the general population. Most researchers agree that these traits increase the risk for drug use and addiction. There are two studies that I will explore more deeply. One in Serbia, which compares the traits of drug users in different user groups, and another in Brazil, which uses similar methodology as (Fehrman et al. 2017).

In a study done in Belgrade, Serbia, 312 opiate addict and 100 alcohol addicts were studied and compared to 346 control participants (Dragan et al. 2012). The researchers hypothesized that because of the different nature and legality of the drugs, they would find personality differences between the two user groups. Opiates are an illegal heroin drug that relieve pain but when abused will lead to addiction. By contrast, alcohol is legal and typically used to reduce anxiety. Therefore, the researchers hypothesized that opiate addicts would rank more highly in impulsivity and general measures of anti-social traits, while alcohol users would rank higher in anxiety. The researchers were conscious of its unique context: Serbia is a society in transition, a developing country with easier access to illicit drugs because of sociopolitical instability. They suggested that this environment is optimal for identifying personality factors in one's choice of drugs precisely because it is less of a choice influenced by law enforcement(Dragan et al. 2012).

Their study found that Novelty Seeking appears to be a serious risk factor for addiction to either drug. Opiate addicts were considered impulsive, lacking in empathy for others, and lower in self-directedness. Alcohol addicts were more impulsive, anxious, and less fantastical in their thinking compared to opiate addicts. Novelty Seeking increases the risk of opiate addiction by a greater degree than compared to alcohol addiction. Opiate addicts seemed to already enjoy fantasy and day dreaming compared to both the general population and alcohol addicts, and this daydreaming, the researchers reported, is potentiated by opiates. Opiate addicts were also more likely to have a personality disorder. The prevalence of personality disorders in the general population is about 15-20 percent, but about 67 percent of opiate addicts were considered to have some kind of personality disorder. Meanwhile only 33 percent of the normal control and 36 percent of the alcohol addicts were identified as someone who could have a personality disorder (Dragan et al. 2012). This study in Serbia suggests that because not all drugs illicit the same response, individuals with particular trait characteristics are more likely to abuse one drug over another.

A different psychometric and drug-use survey in Brazil created a valid sample of over 8,000 participants. This Internet survey is similar to the research methodology of (Fehrman et al. 2017). Researchers found that men were often users and abusers of drugs, as were younger individuals, with the exception of Benzodiazepine drugs. Women and older adults were more likely to to be users of benzodiazepine drugs (Schneider Jr. et al. 2012). The study used the TCI-R, a 240-item self-report questionnaire, to measure one's Harm Avoidance (HA), Cooperativeness (CO), Persistence (PS), Reward Dependence (RD), Self-Directedness (SD), Novelty Seeking (NS) and Sensation Seeking (SS) / Self-Transcendence (ST), among other measures. Their study found some notable conclusions.

Their data findings largely agreed with previous findings in the literature on traits and drug use. Novelty Seeking (NS) is moderately associated with general substance use at all levels of severity, except for hallucinogens. Harm Avoidance (HA) scores were lower in individuals who use alcohol and cannabis at all and for occasional hallucinogen users. Harm Avoidance (HA) increased substantially with the severity of benzodiazepine use. For heavy hallucinogen users and those dependent on benzodiazepines, Self-Transcendence (ST) was higher. Additionally, lower Cooperativeness (CO) was found in participants with alcohol abuse and cocaine abuse. Lower Self-Directedness (SD) scores were found in those who abused alcohol, cannabis, and increasing benzodiazepine use. Higher Reward Dependence (RD) scores were found in those who used cannabis to any degree. Lower Persistence (PS) scores were found in occasional users of cocaine(Schneider Jr. et al. 2012).

The researchers also identified some differences between their findings and other literature. They found that other studies found higher Harm Avoidance (HA) scores in alcoholic inpatients and prisoners, which may be due to a more specific sample from the population at large that has higher alcohol morbidity and co-morbid disorders. Comparatively, internet users are more diverse in their psychometric measures. Other studies also found that higher Self-Directedness (SD) scores were found in heavy alcohol users compared to Schneider Jr. et al. 2012. In their research Harm Avoidance (HA) was more pronounced in benzodiazepine addiction than alcohol abusers. This research is one of the largest studies conducted on the relationship between character traits and drug use, and unlike most other studies, it also considered benzodiazepine use and traits measured using the TCI-R evaluation items (Schneider Jr. et al. 2012). The sample is necessarily biased towards people motivated to access an online survey and complete it accurately. This is a similar design and sample size as the Fehrman et al. 2017 data analyzed in this paper. Therefore it is important to recognize that these studies likely have similar bias toward internet users with an ability to complete a long survey correctly. In other words, large surveys may fail to capture data from individuals who will not sit through a survey on the internet. That is why research like the one conducted in Serbia remains important, despite smaller sample sizes.

2.0 Features of Data

In this section, I will discuss the features of the data provided by Fehrman and Egan (2016). The version of the data used for this analysis comes directly from the UCI Machine Learning Repository (UCI 2016). The data measures 32 variables total. This includes an ID, age, gender, education, country of residence, ethnicity, the five FFM traits from the FFM psychometric evaluation discussed earlier, Impulisivity measured by the BIS-11 (another psychometric evaluation), Sensation Seeking measured by ImpSS (a single psychometric scale with 19 items), and 19 drug-related variables.

In Table 1 below, the count for age, gender, education, country of residence, and ethnicity is shown in descending order. The data is largely composed of white individuals from the United Kingdom and the United States. Most individuals have some college education or more education (more than 50 percent). The data also skews toward younger adults under the age of 34 (more than 50 percent). There is, however, a very even sampling of male and female individuals (non-binary was not measured). Drawing inference on whether the Country or Ethnicity of an individual makes a difference in drug use will be very difficult because of the homogeneous nature of the sample. As we draw conclusions, it's important to remain mindful that our data skews young and educated.

In Table 2, please observe that cocaine is a drug in each of the three major pleiades and there some other repeat drugs. This is not by mistake. The repeated drugs reflect a "soft" clustering approach for distinguishing these pleiades discussed in Fehrman et al. 2017. Alcohol, caffeine, and nicotine do not exist in any pleiade. Semeron is a fictitious drug meant to catch individuals who overreport their drug use. VSA stands for volatile substance abuse, which is not discussed in their paper but is discussed further in Fehrman et al. 2017.

In Figure 1, we show the distribution of drug use in the past decade for the three pleiades defined in Table 1. These drug pleiades provide robust sample sizes and distributional characteristics for statistical analysis. In Figure 2, we have the classification of whether or not someone was a user of alcohol, nicotine, or caffeine in the past 10 years. We will use

Table 1: Non-Psychometric Traits (Count)

Age	Gender	Education	Country	Ethnicity
18-24 (643)	Male (943)	Some College (506)	U.K. (1044)	White (1720)
25-34 (481)	Female (942)	Undergrad. Degree (480)	U.S. (557)	Black/Asian (63)
35-44 (356)		Master's Degree (283)	New Zealand (118)	Black (33)
45-44 (294)		Certificate Diploma (270)	Canada (87)	Asian (26)
55-64 (93)		18 yrs (100)	Australia (54)	White/Asian (20)
65+ (18)		16 yrs (99)	Rep. of Ireland (20)	White/Black (20)
		PhD (89)	Other (5)	Other (3)
		17 yrs (30)		
		$\leq 15yrs(28)$		

Table 2: Drugs Sorted Into Pleiades

Heroin	Ecstasy	Benzodiazepine	Other
Cocaine	Cocaine	Cocaine	Alcohol
Methadone	Amphetamines	Amphetamines	Amyl Nitrite
Crack	Cannabis	Methadone	Caffeine
Heroin	Ketamine	Benzodiazepine	Chocolate
	Ecstasy		Nicotine
	Legal Highs		VSA
	Magic Mushrooms		Semeron
	LSD		

machine learning techniques on both sets of 3 and see how they perform. Rather than use my own definition of a drug user, I followed what the researchers said was a typical definition of user and non-user: A user is someone who used the drug within the past decade and a non-user is someone who used the drug over a decade ago or not even once. Therefore, when an individual has a value of 1 or greater for the drug pleiades, the user has used 1 or more drugs from the pleiade within the past decade. And when a user has a 1 for Alcohol, Caffeine, or Nicotine, they've used that drug within the past decade.

Figure 1: Number of Drugs Used in Each Pleiade (Past Decade)

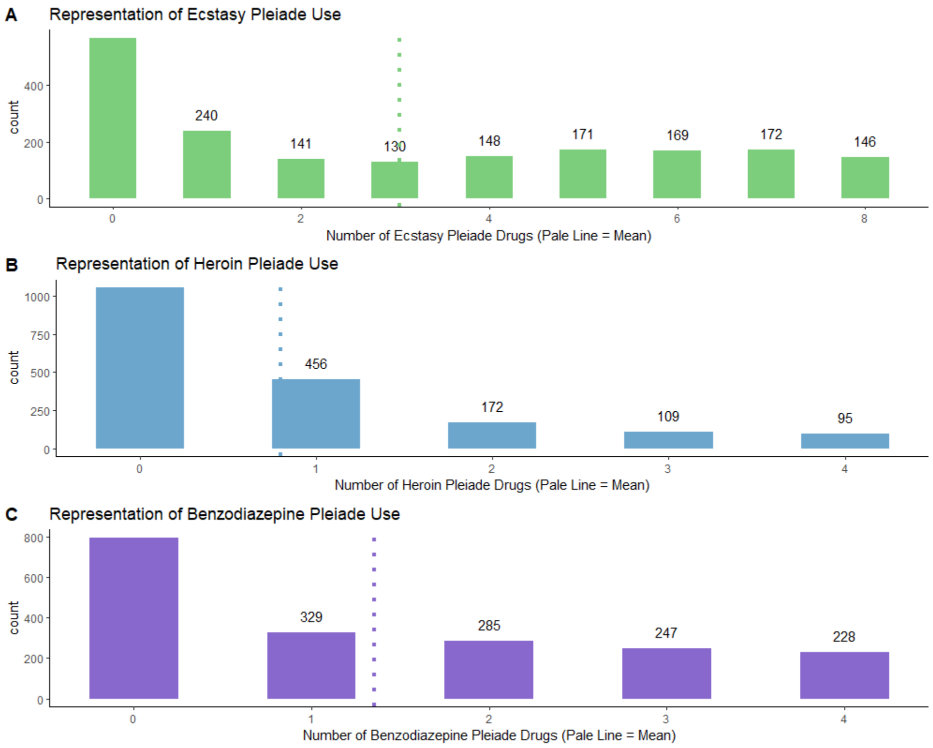
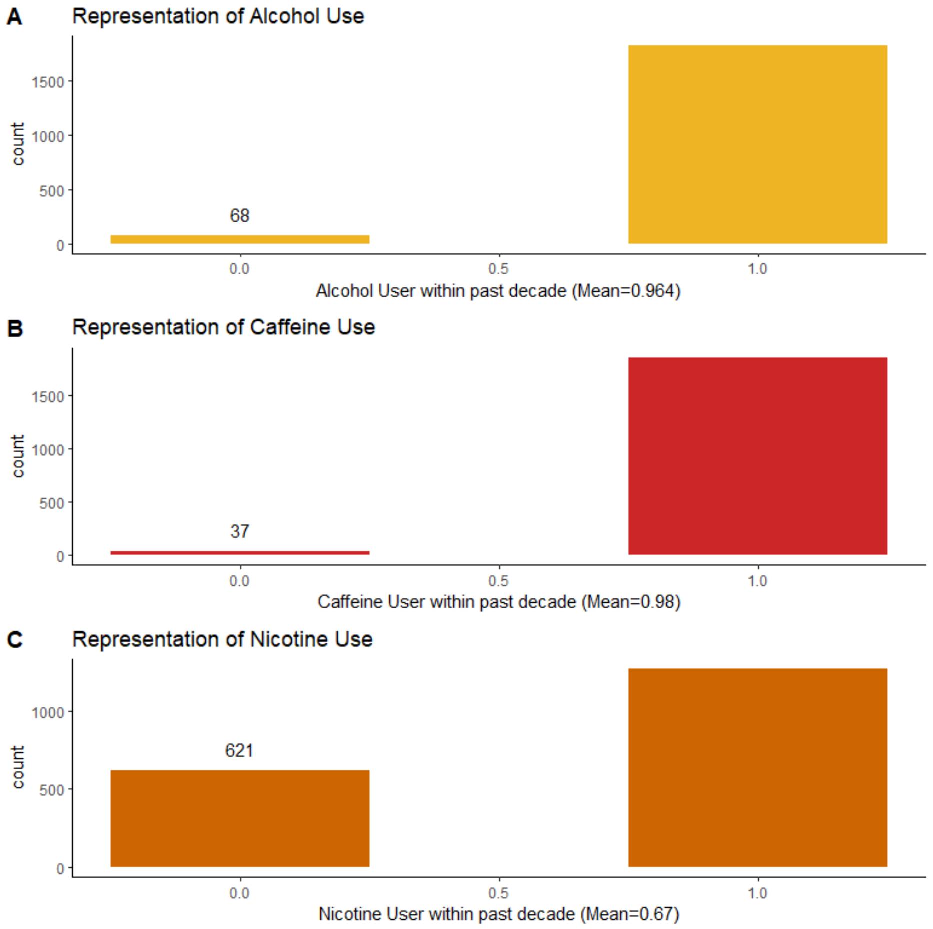


Figure 2: Decade Alcohol, Nicotine, Caffeine Users (1=User)



3.0 Exploratory Data Analysis (EDA)

The psychometric scores have been normalized with center 0 and standard deviation 1. This is helpful for comparing psychometric scores within this sample to identify whether the character traits are different for drug usage. In Figures 3-8, I show the correlation plots for the 6 groups I will be examining: the 3 drug pleiades (Heroin, Ecstasy, Benzodiazepine) and the three legal drugs (Alcohol, Caffeine, Nicotine). Sensation Seeking has the strongest positive correlation for all three drug pleiades. For the Heroin and Benzodiazepine Pleiades, the second most positively correlated trait was Impulsiveness and Openness was the third most positively correlated trait. Meanwhile for the Ecstasy pleiades, the correlation with Openness was greater than the other two pleiades but both Openness and Impulsiveness were in the top three positive correlations. This means that the population of drug users for the 3 pleiades has an above average score for Sensation Seeking, Openness, and Impulsiveness. Meanwhile, Age was the most negatively correlated trait for the three drug pleiades and Conscientiousness was the second most negatively correlated trait. All of these traits had an absolute value of their correlation greater than 0.2.

In the legal triad of drugs, Alcohol, Caffeine, and Nicotine, the top most positive correlation is Agreeableness but its' correlation is about 0.2 or less. There is negative correlation with Impulsiveness, which is about -0.3 or stronger. A distinguishing factor within these three is that Sensation Seeking is not a strong correlation for Alcohol or Caffeine but is a strong negative correlation for Nicotine. Meanwhile, Extraversion has a mildly negative correlation with all three drugs but is a stronger correlation for Alcohol and Caffeine. This means that we would expect the population of alcohol, caffeine, and nicotine users to score less than average on impulsiveness and extraversion but slightly greater than average for agreeableness. We might also expect the population of nicotine users to have a lower than average Sensation Seeking score. For all drug groups, being male, less educated, and younger is more strongly correlated with being a drug user.

It is important to note correlations between traits. In Table 3, we can see that higher Sensation Seeking and Impulsiveness scores are more strongly correlated with being male and Agreeableness is more strongly correlated with being female. The strongest correlation between all character traits is between Sensation Seeking and Age (-0.33). Indeed, the mean Sensation Seeking score for all individuals ages 18-24 is 0.40 (half a standard deviation), it's almost 0 (normalized average) for age 25-34, -0.3 for ages 35-44, around -.4 for ages 45-64, and almost -1 for ages 65 and older. This raises questions about the stability of Sensation Seeking over an individual's life-time. Age also has a moderate negative correlation with Openness and Impulsiveness. Meanwhile, having completed more educated is positively correlated with Conscientiousness. These directional relationships between traits are consistent with their correlations with drug use in Figures 3-8.

I also conducted Principal Component Analysis (PCA) on the variables to see if I could complete significant dimension reduction. The PCA-Biplot (Figure 9) shows the first two principal component loadings and the contribution of the psychometric traits to the variance. The amount of variance retained by each principal component is measured by the eigenvalue. It looks like the psychometric traits that most strongly contributes to the variance is Sensation Seeking and Extraversion (17.5 percent variance). The second contributors

Figure 3:

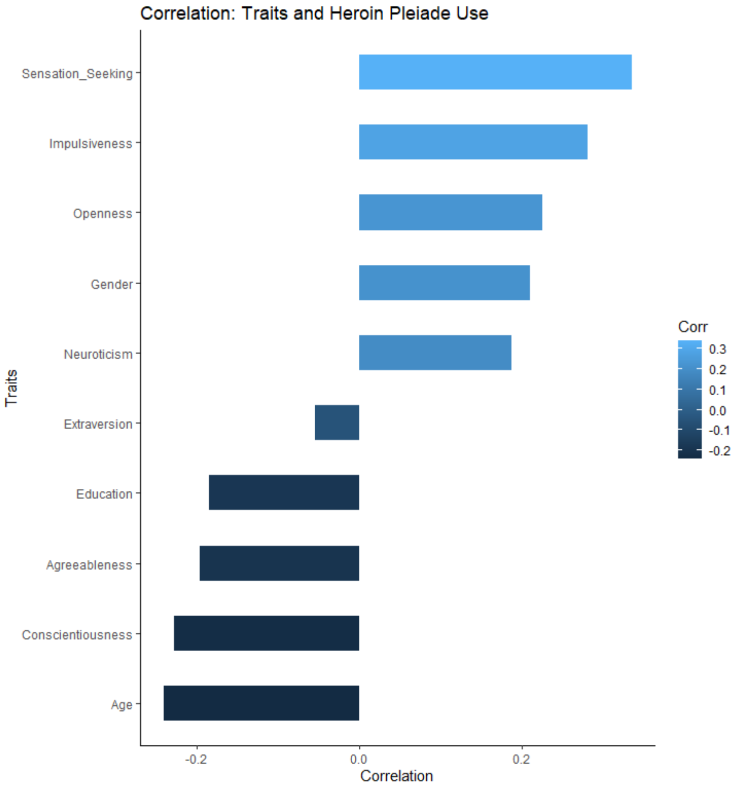


Figure 4:

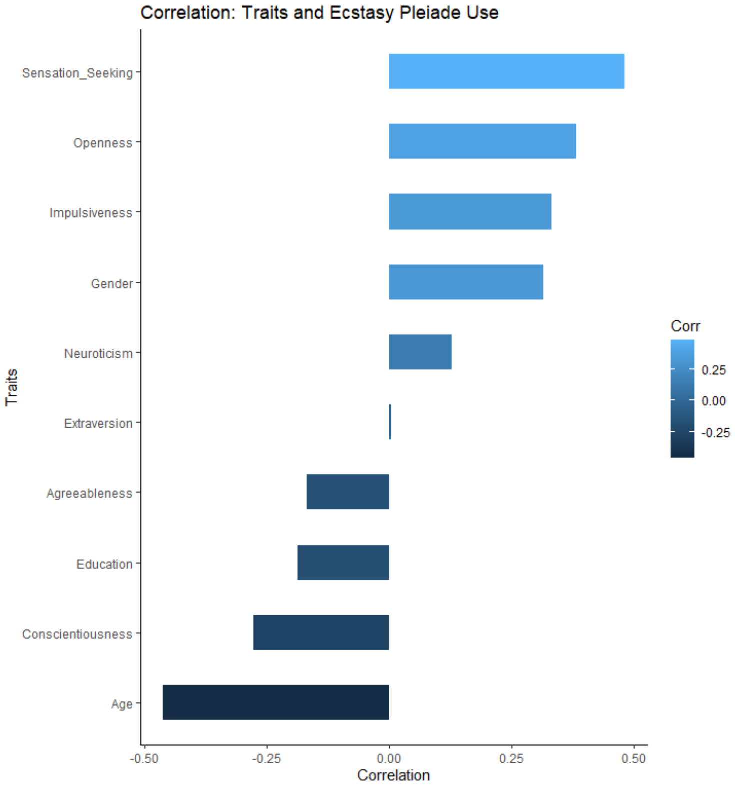


Figure 5:

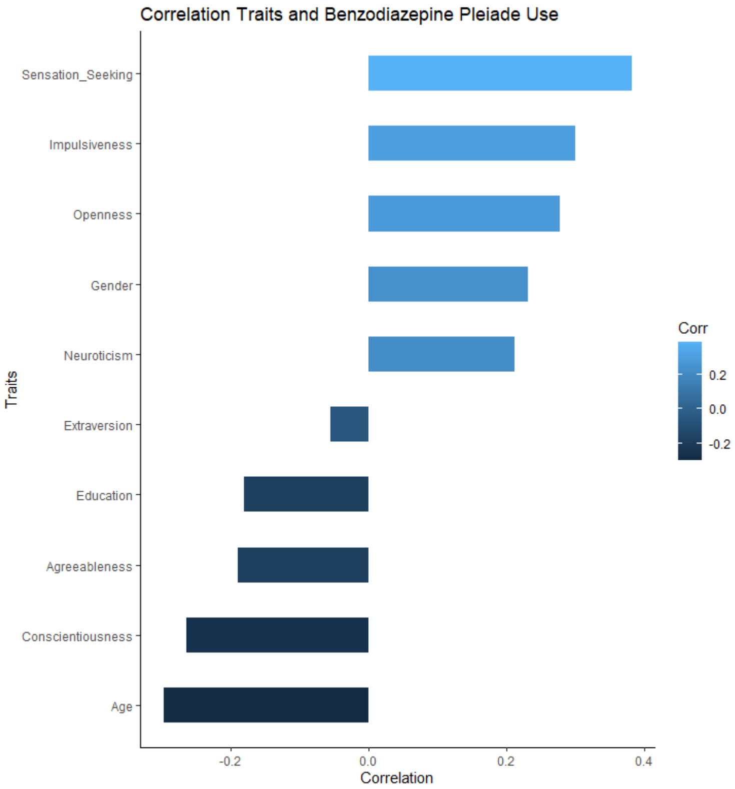


Figure 6:

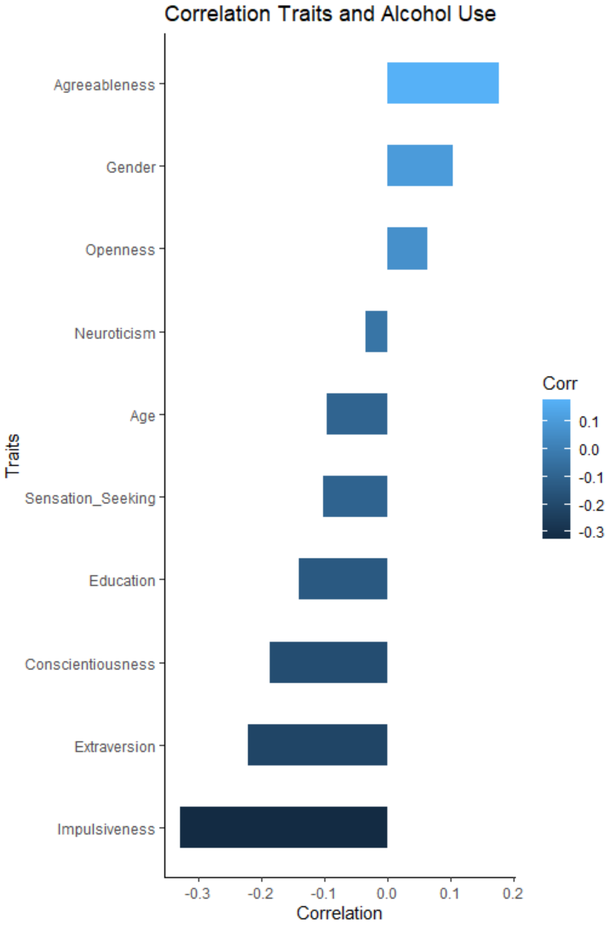


Figure 7:

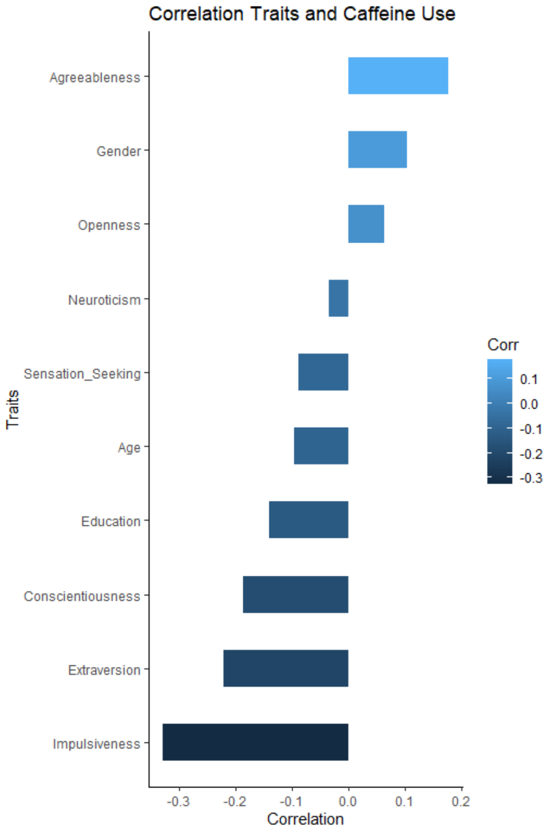


Figure 8:

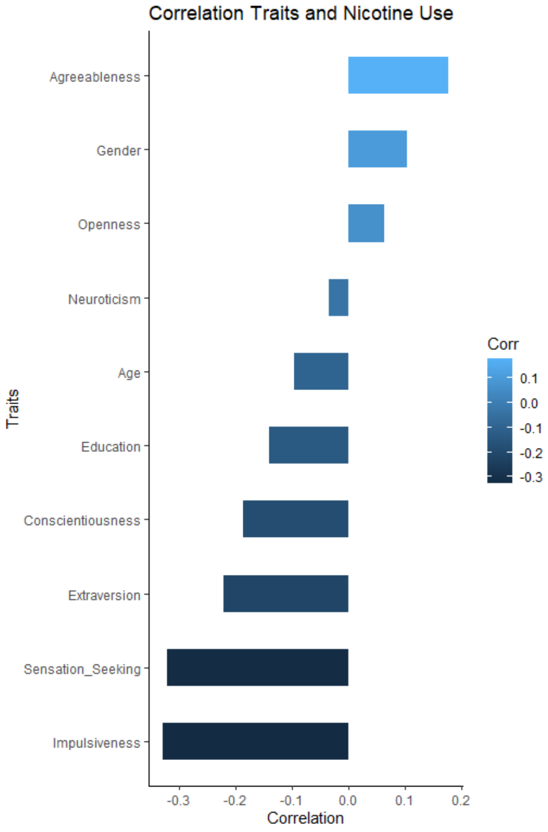


Table 3: Correlation Between Traits and Important Categorical Variables

Traits	Gender (0=Female)	Age	Education
Neuroticism	-0.0746	-0.1398	-0.0924
Extraversion	-0.0579	-0.0349	0.1125
Openness	0.1310	-0.2204	0.0726
Agreeableness	-0.2197	0.06292	0.0839
Conscientiousness	-0.1838	0.1774	0.2230
Impulsiveness	0.1675	-0.1865	-0.1150
Sensation Seeking	0.2442	-0.3276	-0.1114

are Impulsiveness and Neuroticism (about 15 percent variance). The third are Conscientiousness and Openness (13.5 and 12.5 respectively). And Agreeableness is last (about 9 percent variance). This suggests that there isn't overwhelming dimension reduction that can be done. The Scree Plot for the numerical variables confirms that we can explain about 88 percent of the variance with five (5) Principal Components out of the seven total. The dimension reduction would not streamline the models by much. Attempts were made to conduct PCA on the categorical variables because the original research (Fehrman et al. 2017) had done so and found that Country and Ethnicity failed to contribute a significant amount to the variance. I kept on encountering inconsistent results but the code is in the Appendix at the end of the paper. Rather than trying to wrangle non-ordinal categorical variables, I took another tip from the research and removed the Country and Ethnicity from modeling. In the section on the features of the data, it was notable that ethnicity was largely white and US/UK. Removing these variables will help us draw inference on traits that help distinguish individuals from one another.

I also conducted K-Means on the variables to try to identify clusters. I examined the k=3 K-Means clustering (see Figure 11) in order to find what generally distinguishes the data. In Figures 12-14, I examine the distributions in the data for the different character traits. The first cluster is quite normal: the approximate center of all the psychometric traits is approximately 0 (standardized mean). The second cluster is characterized by less Openness (mean = -0.5), more Conscientiousness (mean=0.5), less Sensation Seeking (mean = -0.5), and less Impulsiveness (mean = -0.5). The third cluster is characterized by more Openness (mean = 0.45), less Conscientiousness (mean=-0.5), more Sensation Seeking (mean=0.5), and more Impulsiveness (mean=0.5). This configuration of k-means is interesting because it seems to sort individuals into clusters that are also highly correlated with drug use.

At the risk of being too presumptuous, it looks like the first cluster are individuals who score very normal on psychometric measures, the second cluster includes individuals who are generally more cautious because they are less Impulsive, more Conscientious, and less Sensation Seeking / Open. And the third cluster looks like the profile we would expect for drug users. To investigate these hypotheses. I then examined the drug use of the three drug pleiades and the three legal drugs, Alcohol, Caffeine, and Nicotine. See Figures 15-17 for the histograms and their means (blue). Cluster 1 had higher Ecstasy use (mean = 4 drugs), rare

Figure 9:

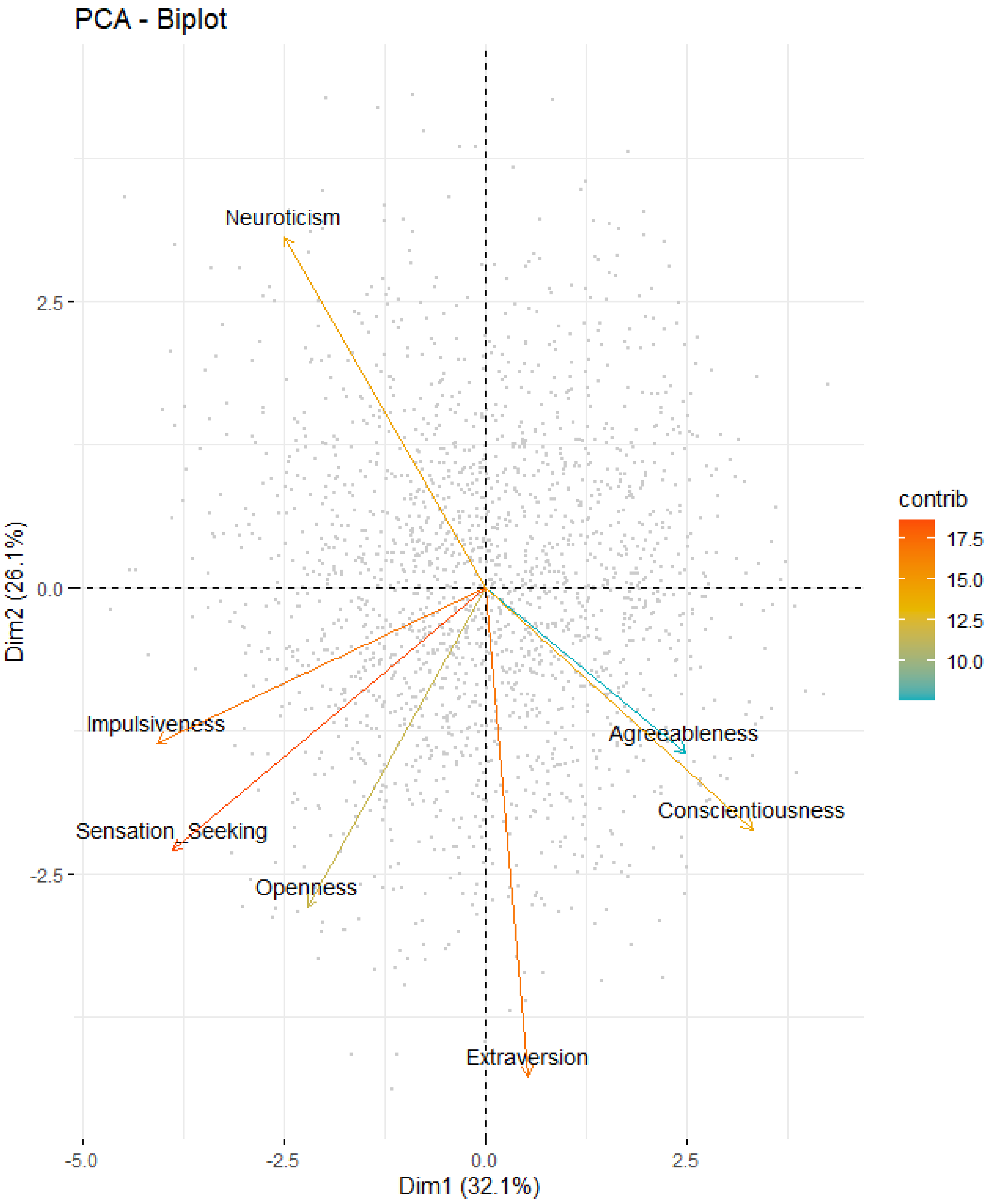

PCA - Biplot
Neuroticism
Impulsiveness
Sensation_Seeking
Openness
Agreeableness
Conscientiousness
Extraversion
contrib
17.5
15.0
12.5
10.0
Dim2 (26.1%)
2.5
0.0
-2.5
Dim1 (32.1%)
-5.0
-2.5
0.0
2.5

Figure 10:

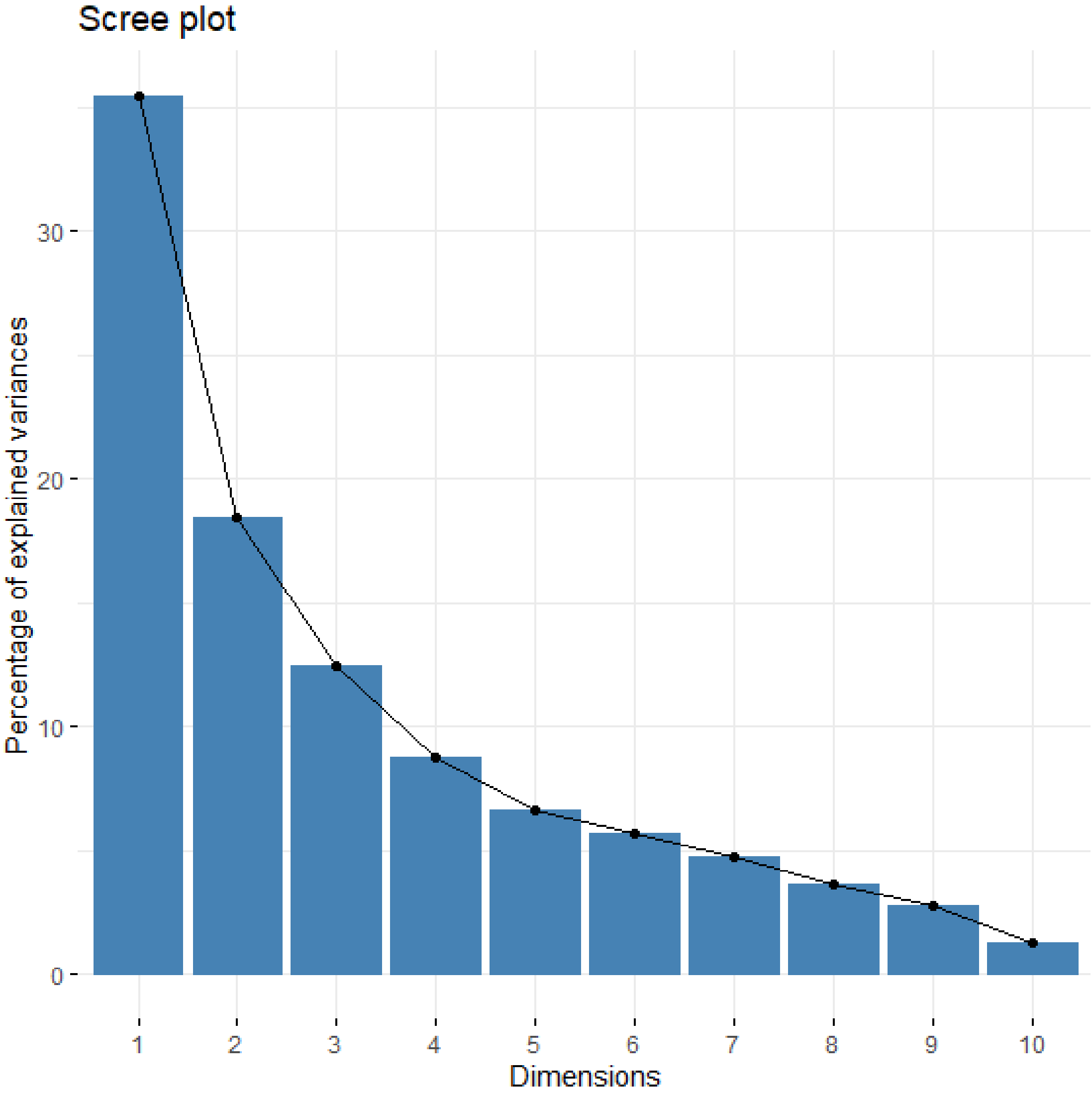

Scree plot
Percentage of explained variances
0
10
20
30
1
2
3
4
5
6
7
8
9
10
Dimensions

Figure 11:

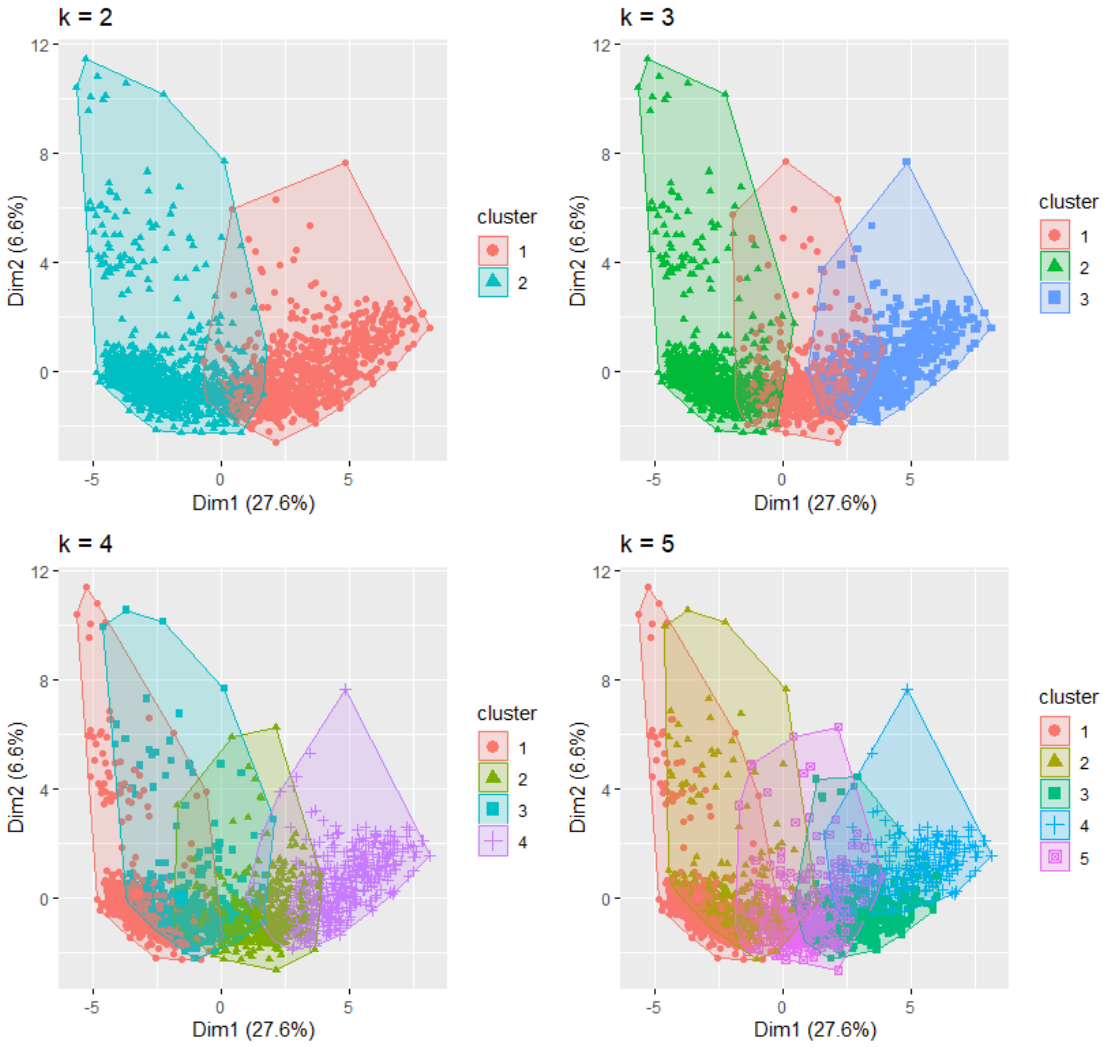


Figure 12: First Cluster Traits (K-Means)

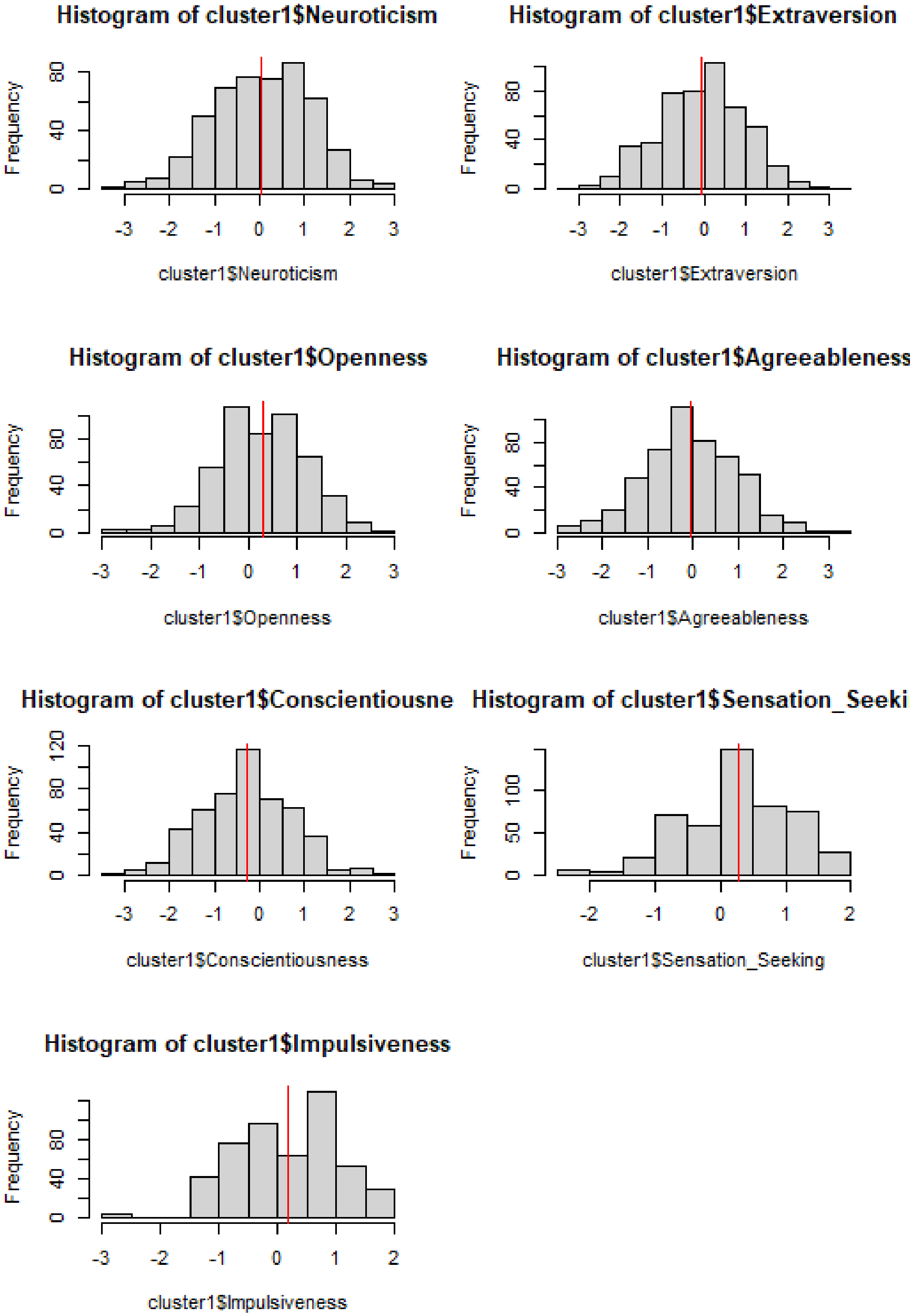


Figure 13: Second Cluster Traits (K-Means)

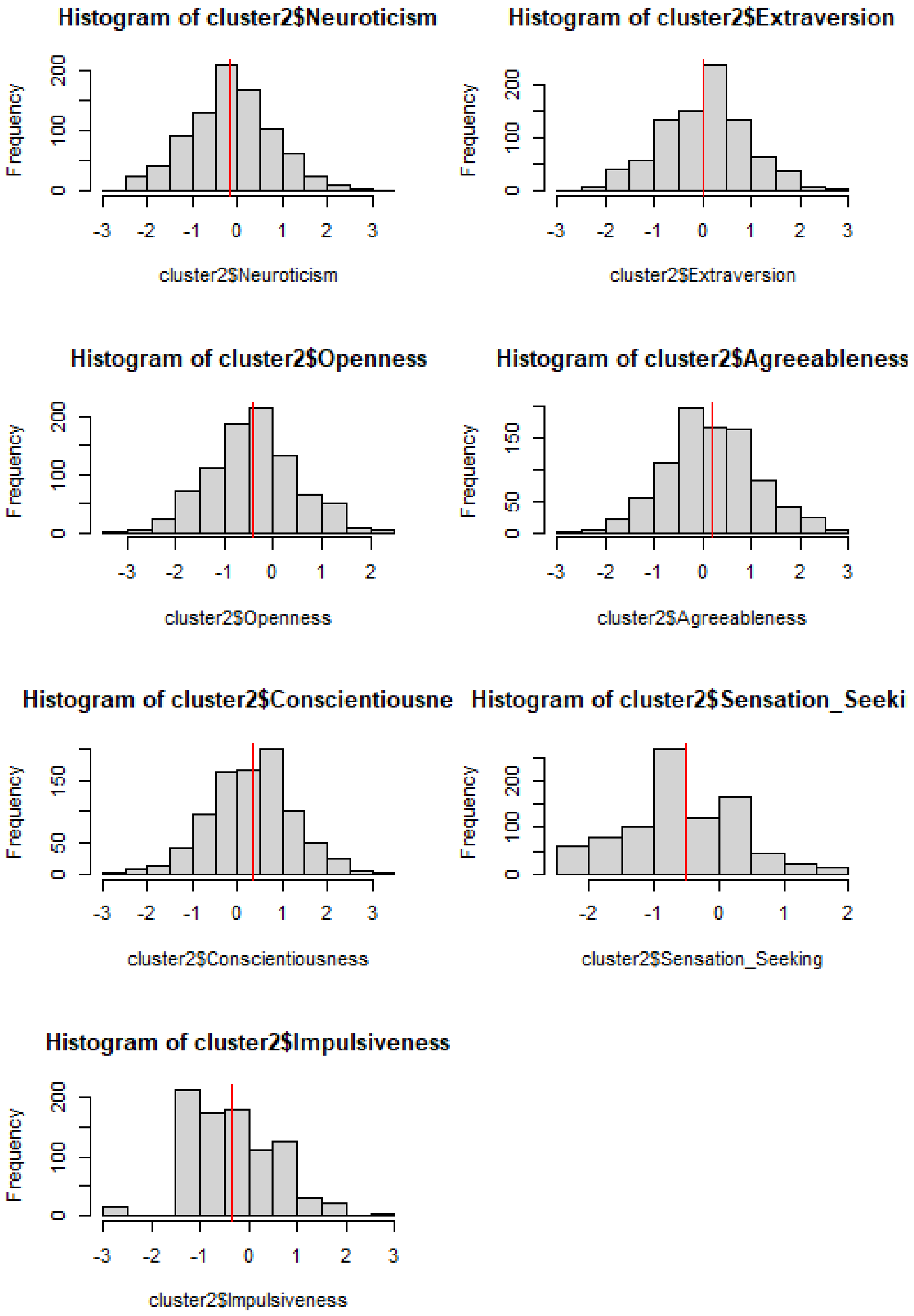
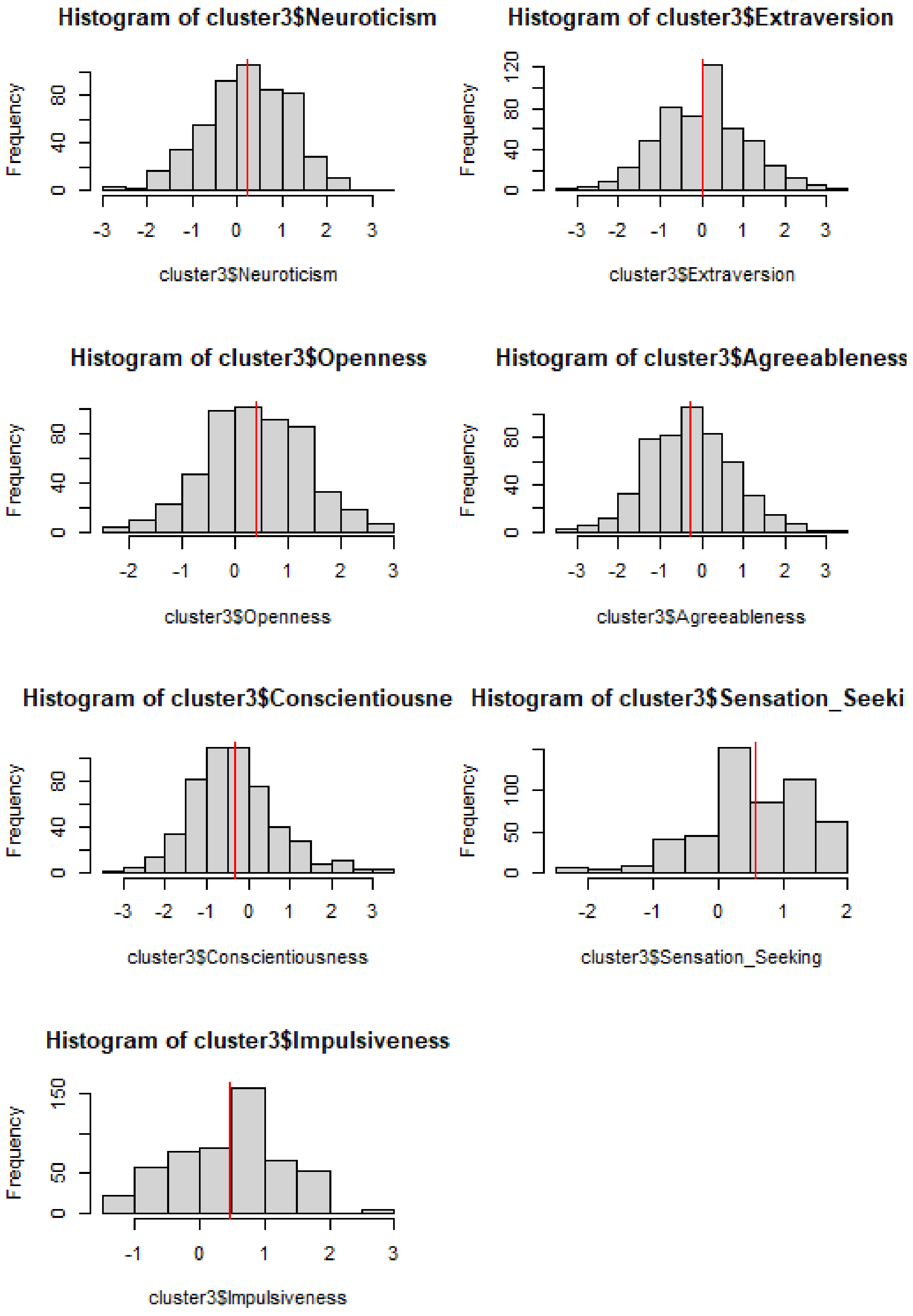


Figure 14: Third Cluster Traits (K-Means)



Heroin pleiade use (mean ¡ 1), mild Benzodiazepine use (mean = 1.5), and most individuals had used Alcohol, Caffeine and Nicotine within the past decade. Cluster 2 had almost no Heroin or Benzodiazepine use, very low Ecstasy use (mean = 0.5), low Nicotine use (mean = 0.42), and have used Alcohol and Caffeine within the past decade. Cluster 3 is characterized by the most Heroin use (mean = 2 drugs), the most Ecstasy use (mean = 6.7), the most Benzodiazepine use (mean = 3.2), and regular Alcohol, Caffeine, and Nicotine use (means are equal to or approximately 1). The k means cluster has identified not just clusters of traits but also clusters of drug use.

Previously, I conducted analysis of within group sum of squares for just the pleiades and it was difficult to identify an ideal number of clusters. After adding the Alcohol, Caffeine, and Nicotine use, the ideal number of cluster graph (Figure 18) became a clear "elbow" to demark a cutoff. It looks like about 4 clusters should be sufficient, although there is a second elbow. Just in case, I also created a dendrogram (Figure 19) and selected clusters with approximately unbiased (AU in red) p-values 95 percent or greater. That is, the elements in red are considered to be strongly supported by data. The Dendrogram suggests that alcohol and caffeine behavior are clustered together. It also suggests that all the three pleiades are clustered with Nicotine. This seems to reflect the insight found from K-Means Clustering where cluster 3 included heavy drug use. The dendrogram also suggests that Openness, Impulsiveness, and Sensation Seeking are clustered together, much like the K-Means analysis but also the insight found in the correlations. Additionally, the dendrogram suggests that Agreeableness, Extraversion, and Conscientiousness are clustered together. In what direction, it does not say, but this is similar insight as the K-Means clustering where cluster 2 had atypical means reflective of people who generally don't use drugs. Additionally, the age group 18-24 years old is clustered with some college, and Neuroticism is clustered with 17 years of education. These clusters with age were not previously found.

Figure 15: First Cluster Drug Use

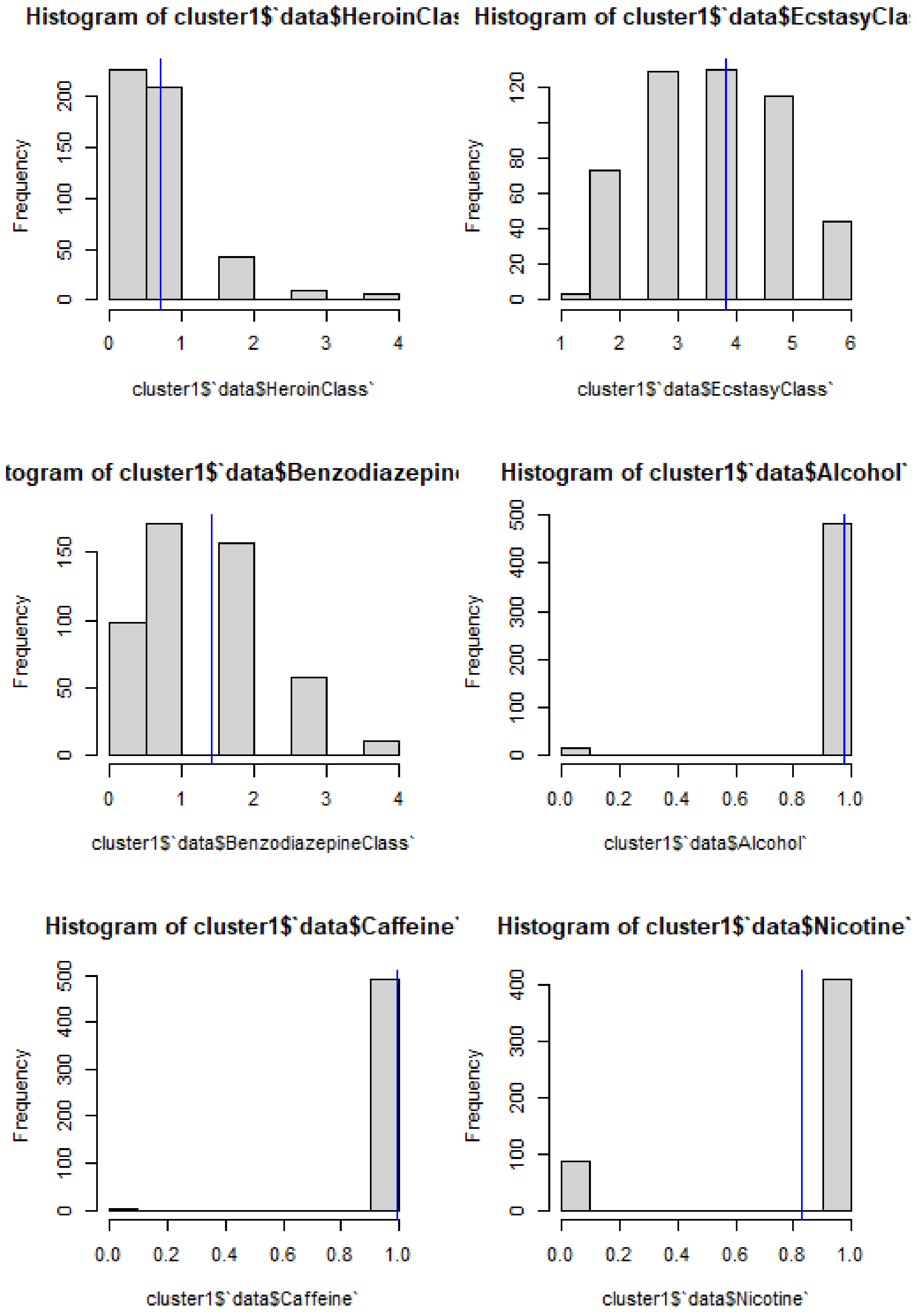


Figure 16: Second Cluster Drug Use

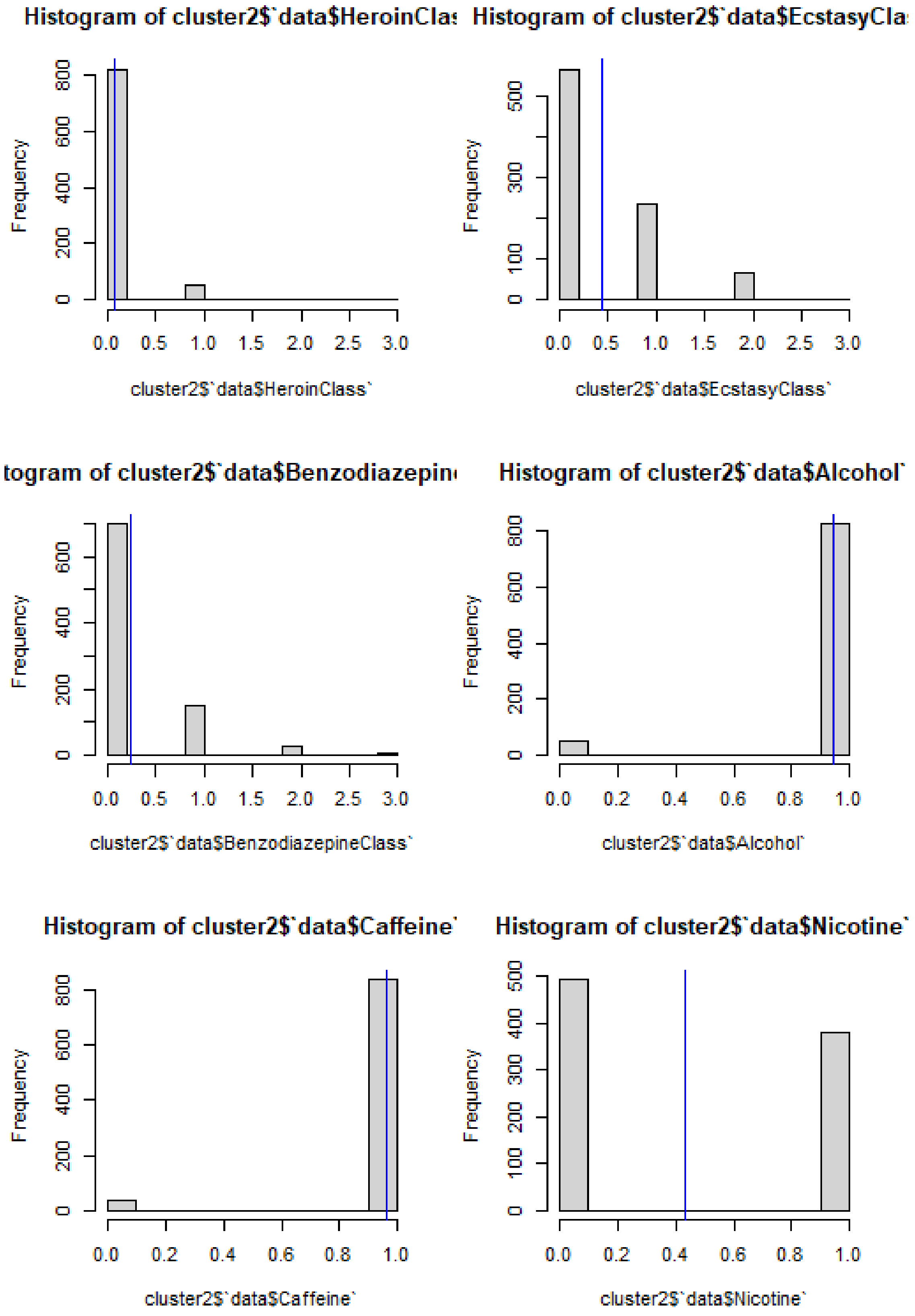


Figure 17: Third Cluster Drug Use

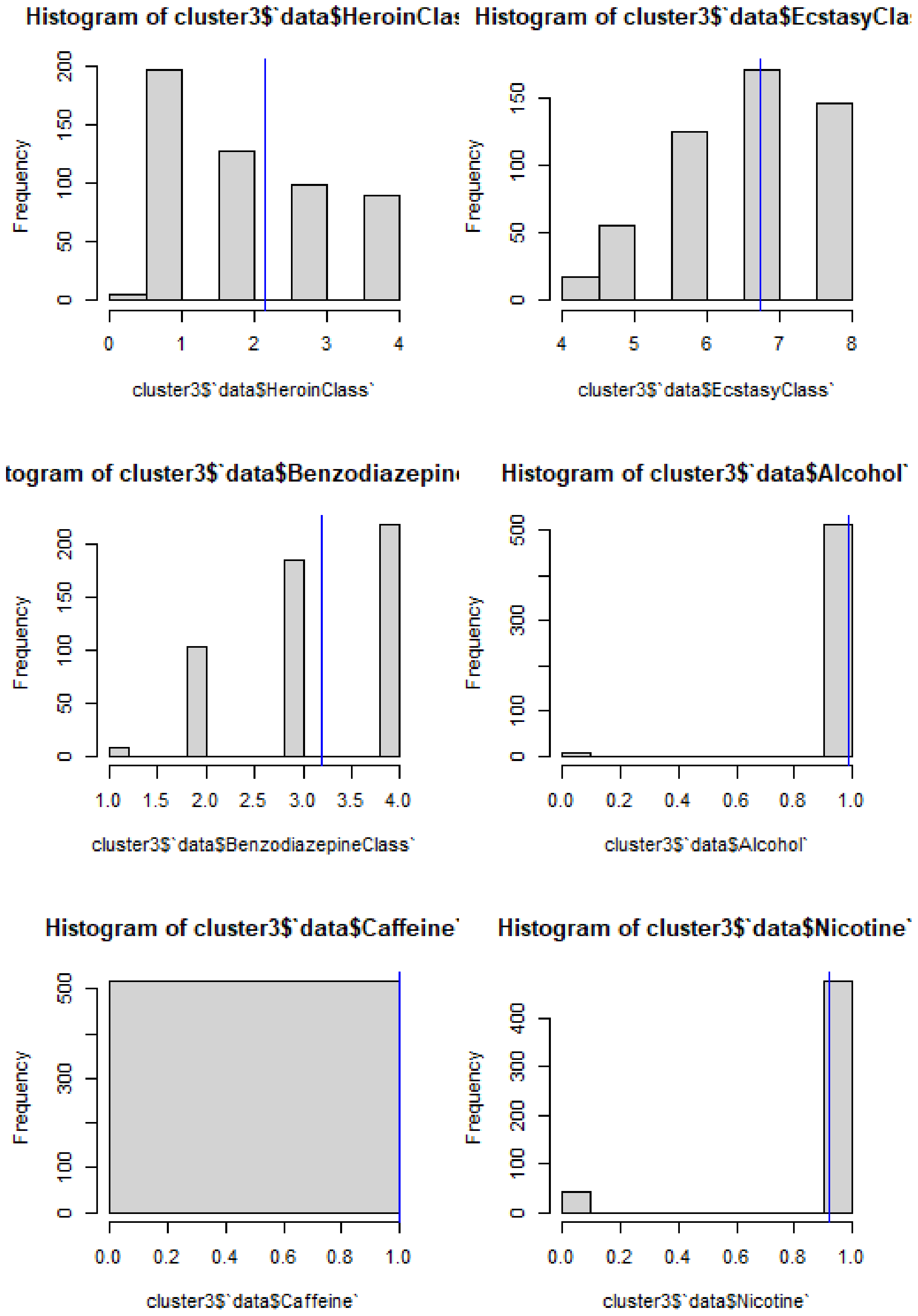


Figure 18: Ideal number of clusters

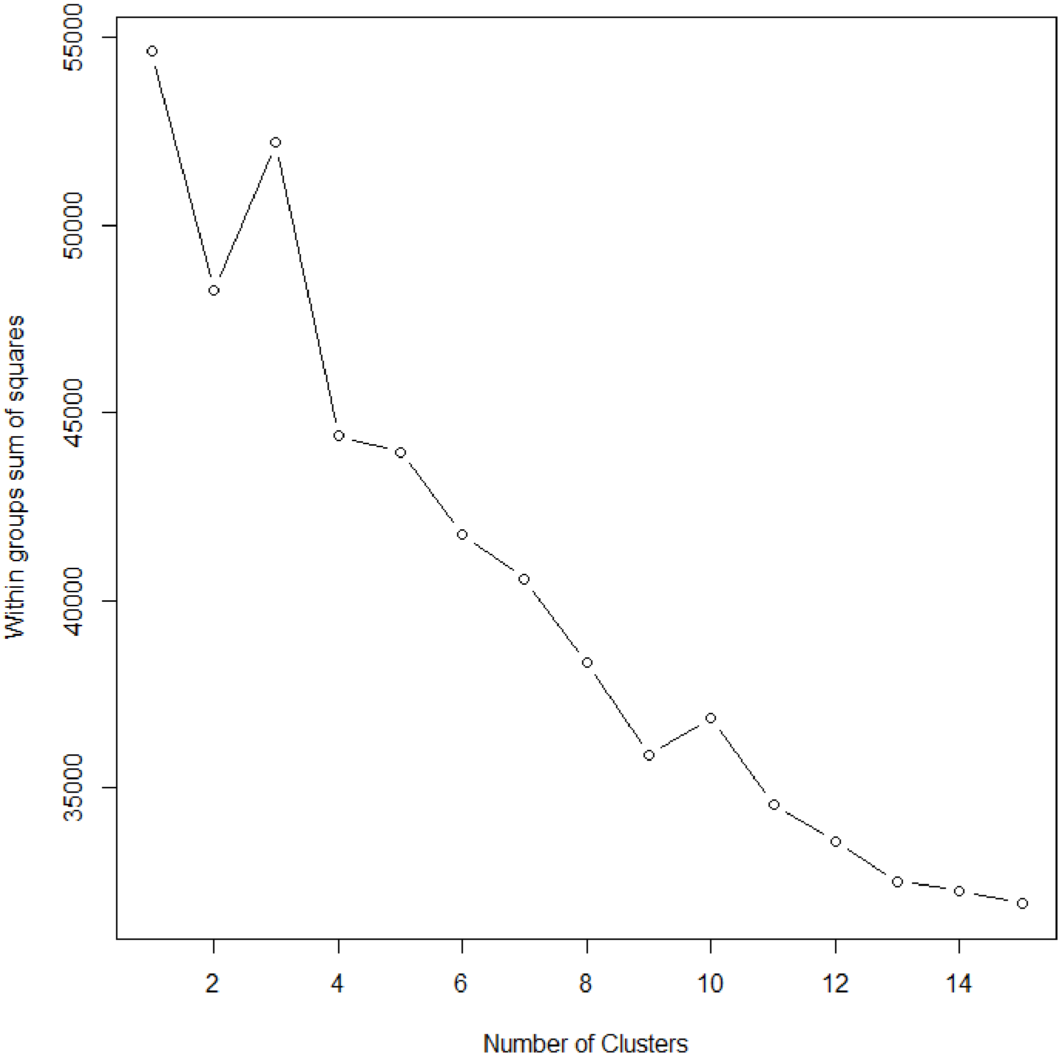


Figure 19:

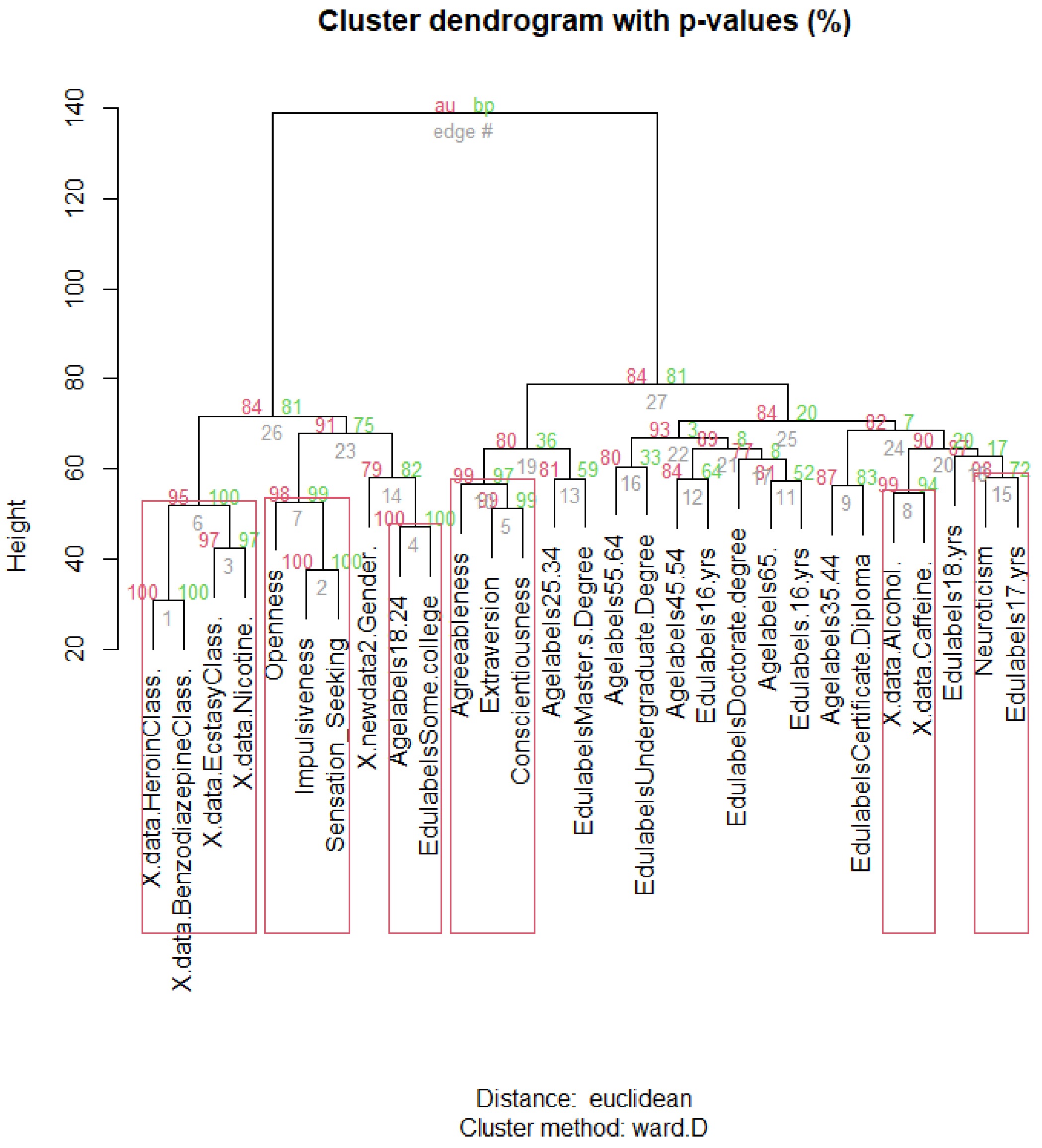

Cluster dendrogram with p-values (%)
au
bp
edge #
Height
20
40
60
80
100
120
140
X.data.HeroinClass.
X.data.BenzodiazepineClass.
X.data.EcstasyClass.
X.data.Nicotine.
Openness
Impulsiveness
Sensation_Seeking
X.newdata2.Gender.
Agelabels18.24
EdulabelsSome.college
Agreeableness
Extraversion
Conscientiousness
Agelabels25.34
EdulabelsMaster.s.Degree
Agelabels55.64
EdulabelsUndergraduate.Degree
Agelabels45.54
Edulabels16.yrs
EdulabelsDoctorate.degree
Agelabels65.
Edulabels.16.yrs
Agelabels35.44
EdulabelsCertificate.Diploma
X.data.Alcohol.
X.data.Caffeine.
Edulabels18.yrs
Neuroticism
Edulabels17.yrs
Distance: euclidean
Cluster method: ward.D

Table 4: Correlation Between Traits and Important Categorical Variables

Drug	Ridge Acc	NN1 Acc	NN2 Acc
Heroin Pleiade	0.97	0.65	0.65
Ecstasy Pleiade	0.81	0.82	0.84
Benzodiazepine Pleiade	0.70	0.72	0.68
Alcohol	0.97	0.97	0.74
Caffeine	0.99	0.99	0.77
Nicotine	0.70	0.71	0.71

4.0 Modeling

4.1 Goal 1: Clustering

See Section 3 EDA for discussion of insights from PCA, K-Means, and Dendrogram.

4.2 Goal 2: Prediction

For the purposes of prediction, I completed an optimized ridge regression with the best lambda and two neural networks (one simple and one with two dropout layers). The accuracy of these three models is shown in Figure 20-23, and also in Table 4. The general results are that ridge regression is an accurate predictor for all drugs but the neural networks ended up having mixed results. In a previous draft of this work, the neural network for the pleiades (without the additional drugs for analysis) worked better. If I had more time, I would definitely revise the neural networks again. Overall, with one of the three techniques, accuracy of predicting a user could be about 70 percent or better. The pitfall will be discussed in the discussion section.

4.3 Goal 3: Trait Association per Drug

Despite best efforts to discover what traits are associated with what drugs in the ridge regression, the best analysis I found comes from the EDA clustering and correlation. This will be discussed more thoroughly in the results section.

5.0 Results

With my analysis, I have been able to predict a drug user with at least 70 percent accuracy. However, there is a catch. After examining the confusion matrix for alcohol and caffeine, it's clear that our models, because of a paucity of non-users for these drugs, were completing the equivalent of "shooting fish in a barrel". That is, the models had high accuracy because the samples for training and testing were heavily skewed towards users. If I had more time,

Figure 20: Ridge Regression

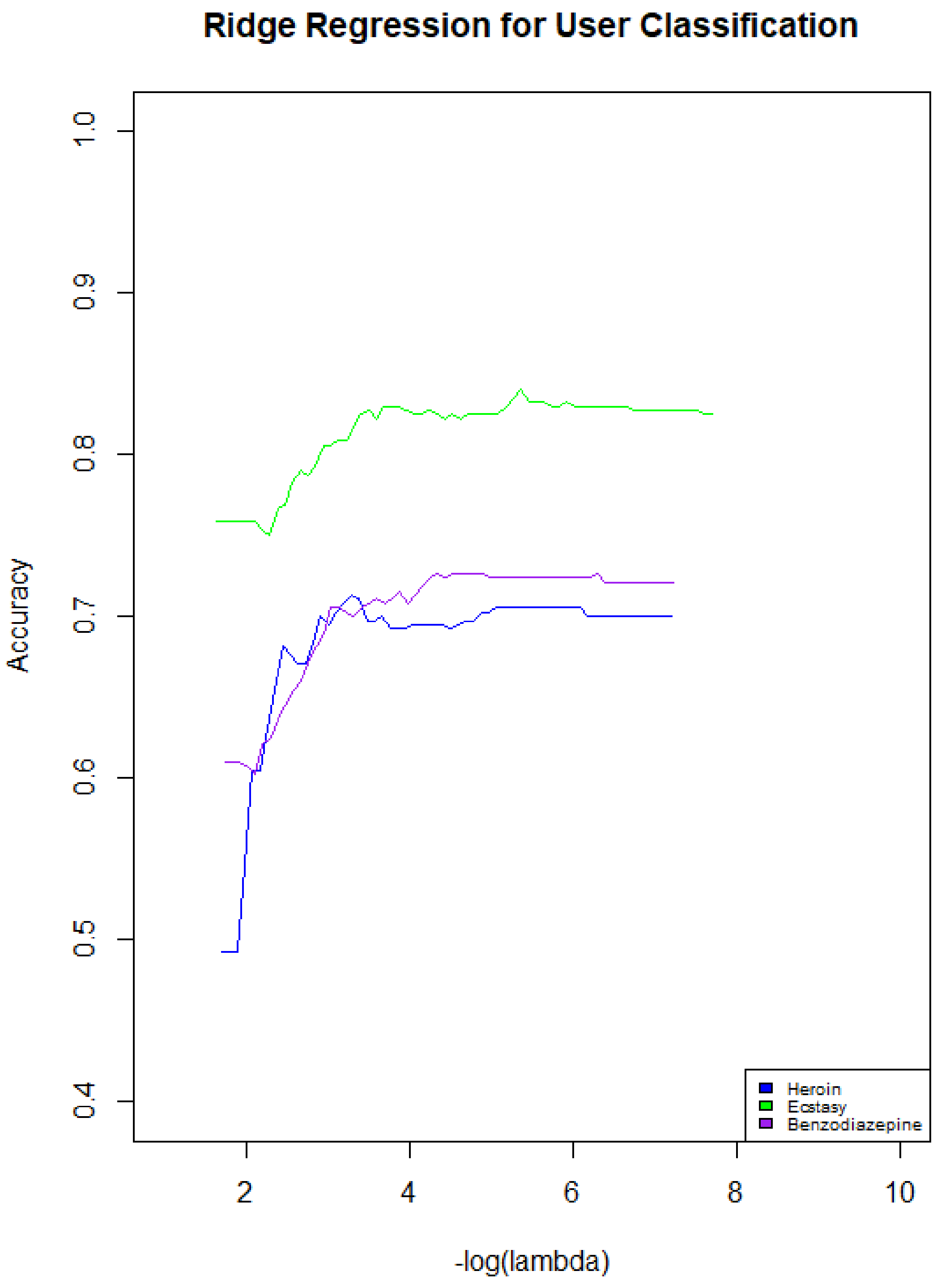


Figure 21:

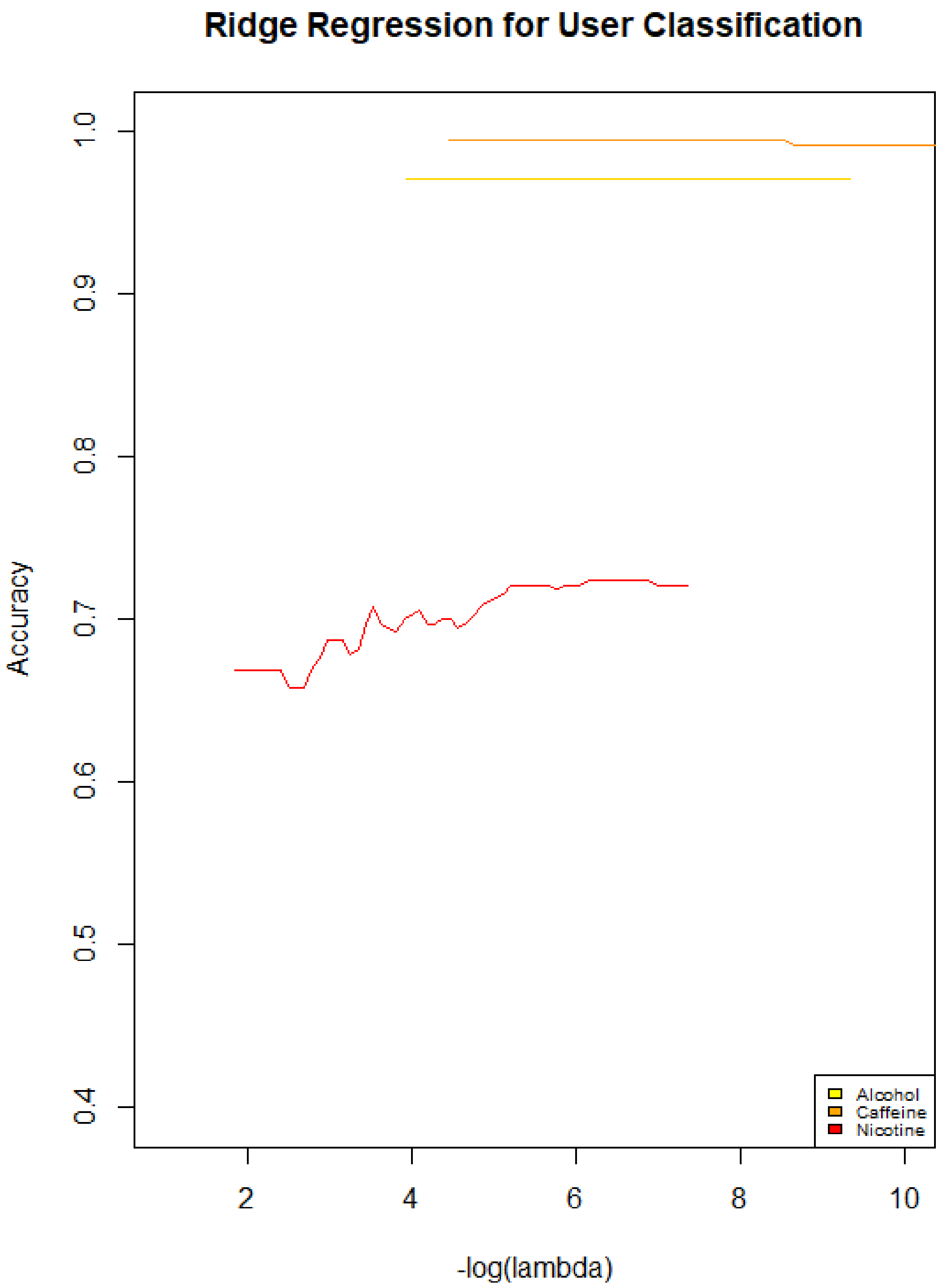

Ridge Regression for User Classification
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
2
4
6
8
10
-log(lambda)
Alcohol
Caffeine
Nicotine

Figure 22:

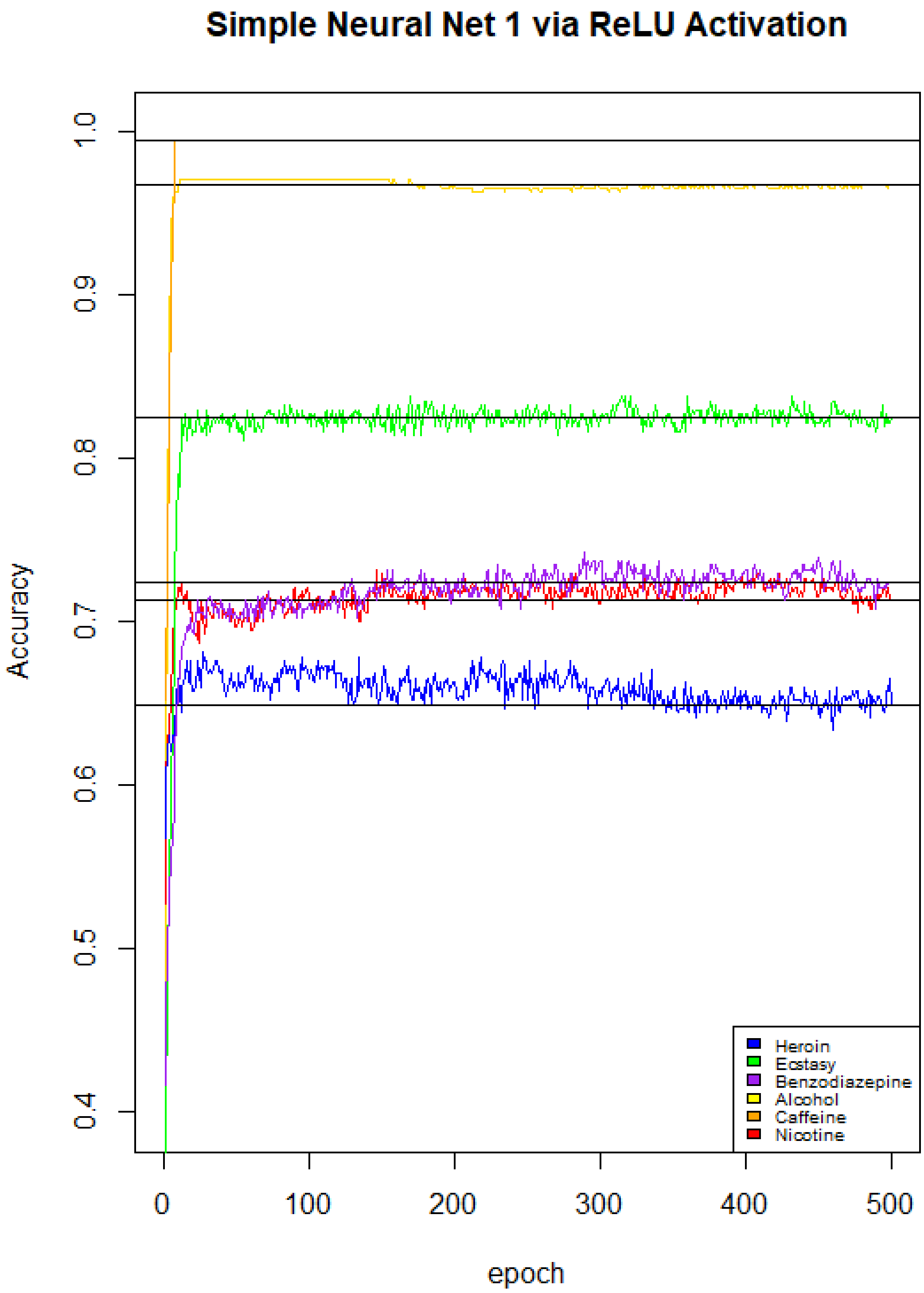
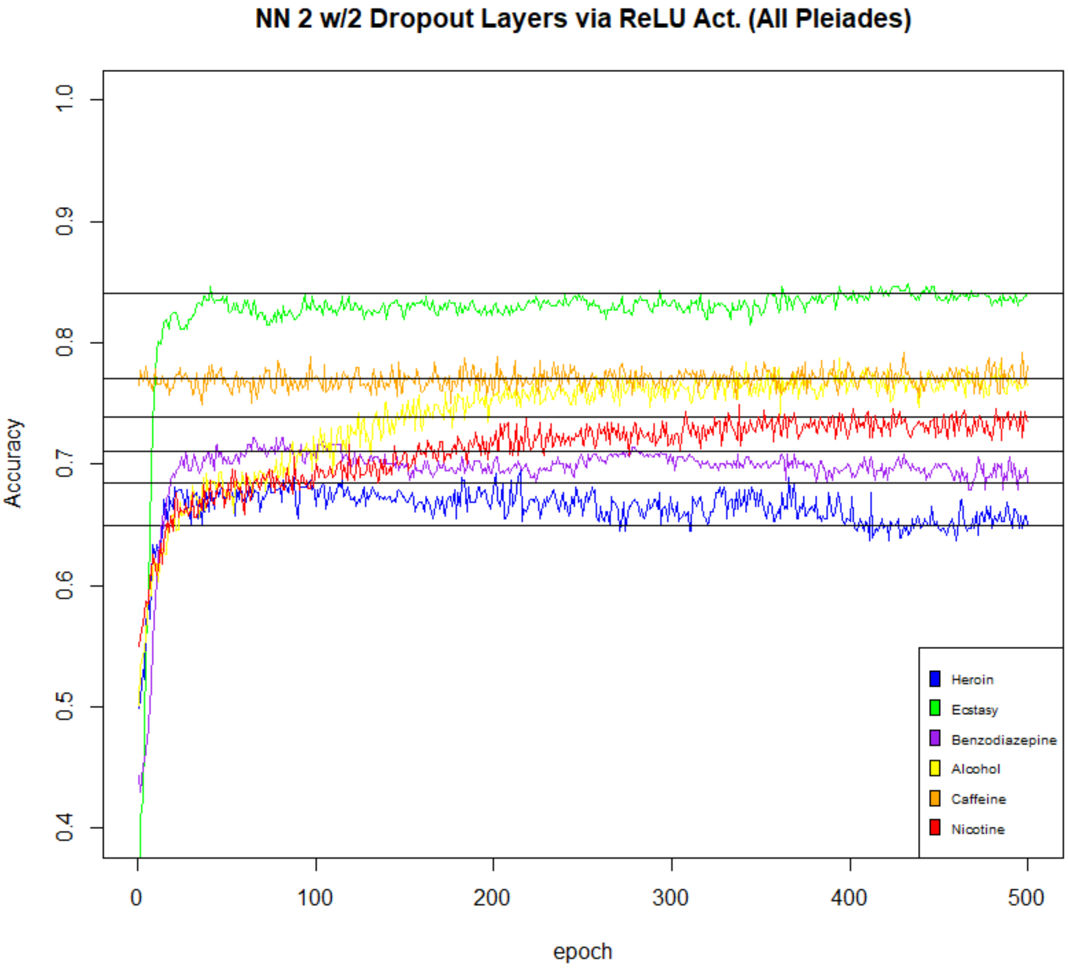

Simple Neural Net 1 via ReLU Activation
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
100
200
300
400
500
epoch
🟦 Heroin
🟩 Ecstasy
🟪 Benzodiazepine
🟨 Alcohol
🟧 Caffeine
🟥 Nicotine

Figure 23:


NN 2 w/2 Dropout Layers via ReLU Act. (All Pleiades)
Accuracy
0.4
0.5
0.6
0.7
0.8
0.9
1.0
epoch
0
100
200
300
400
500
🟦 Heroin
🟩 Ecstasy
🟪 Benzodiazepine
🟨 Alcohol
🟧 Caffeine
🟥 Nicotine

I would attempt a different cut-off point for the definition of a user that would allow for a more even split and therefore better analysis. This means that we should be skeptical of the value of accuracy for the alcohol and caffeine. I was hoping to discover something like this by including these commonly used drugs. It's a challenge to the data: Strong skewing of user groups. However, this challenge was only mild with Nicotine and was not an obstacle for the three pleiades. It seemed like the Neural Networks under-performed compared to Ridge Regression with best lambda for all drug groups except Nicotine, for which it yielded somewhat equal results. This level of accuracy suggests that the variables in our data set (without Country and Ethnicity) are solid predictors for drug use, given the proper machine learning technique.

The clusters that predict whether or not someone is a user consistently appeared in the EDA. The results from K-Means were especially illuminating. Cluster 1 seemed to be a unique identifier of the traits associated with users who mostly use a few drugs from the Ecstasy pleiade. The traits of this cluster are relatively normal. This was a surprising finding, but it suggests that there might be two classes of Ecstasy users. This cluster may also relate to the literature that suggests the reasoning behind using Ecstasy Pleiade drugs is not necessarily to numb pain but to have new experiences and that reasoning is very different compared to other motivations behind drug use. Cluster 2 reflected the traits of people who have only used Alcohol and Caffeine within the past decade. Characterized by less Openness, more Conscientiousness, less Sensation Seeking, and less Impulsiveness, this finding is consistent with the literature discussed previously and the correlations found in the EDA. Cluster 3 is characterized by heavy drug users who use more Ecstasy pleiade drugs than the first cluster. It is characterized by more Openness, less Conscientiousness, more Sensation Seeking, and more Impulsiveness. This confirms the correlations found previously with most drugs and it further reaffirms the literature review discussed previously.

6.0 Discussion

If I had more time, I would have liked to have worked in the Neural Network for better accuracy. I would have also liked to explore the role of age and education further. However, these two categorical variables had strong clustering with personality traits. If I could have any other extra variable, it would be the context in which people do drugs. I would expect that the association between character traits and the social environment of drug users would overlap. If I could have an additional variable, it would be income. Based on the clustering discussed earlier, I would expect that the Ecstasy Pleiade would have a split between users with not only certain character traits but also certain economic backgrounds. Overall, my analysis confirmed what previous literature has found in the relationship between psychometric data and drug use.

References

[1] Cooper, M. 1994. "Motivations for alcohol use among adolescents: Development and validation of a four-factor model." *Psychological Assessment* 6(2): 117–128. https://doi.org/10.1037/1040-3590.6.2.117.

[2] Dash, G., Slutske, W., Martin, N., Statham, D. , Agrawal, A., and M. Lynskey. 2019. "Big Five personality traits and alcohol, nicotine, cannabis, and gambling disorder comorbidity." *Psychology of Addictive Behaviors* 33(4): 420–429. https://doi.org/10.1037/adb0000468.

[3] Dash, G.. Martin, N., and W. Slutske. 2021. "Big Five personality traits and illicit drug use: Specificity in trait-drug associations." *Psychology of Addictive Behaviors*. Advance online publication. https://doi-org.lib-proxy.fullerton.edu/10.1037/adb0000793.

[4] Milivojevic, D., Milovanovic, S., Jovanovic, M., Svrakic, D., Svrakic, N., Svrakic, S., and C. Cloninger. 2012. "Temperament and Character Modify Risk of Drug Addiction and Influence Choice of Drugs." *American Journal on Addictions* 21 (5): 462-467.

[5] Fehrman, E., Muhammad, A., Mirkes, E., Egan, V. and A. Gorban. 2017. "The Five Factor Model of personality and evaluation of drug consumption risk." https://arxiv.org/abs/1506.06297v2.

[6] Fehrman, E., and V. Egan. 2016. "Drug consumption, collected online March 2011 to March 2012, English-speaking countries." *Inter-university Consortium for Political and Social Research* [distributor]. https://doi.org/10.3886/ICPSR36536.v1

[7] Kroencke, L., Kuper, N., Bleidorn, W., and J. Denissen. 2021. "How does substance use affect personality development? Disentangling between- and within- person effects." *Social Psychological and Personality Science* 12(4): 517–527. https://doi.org/10.1177/1948550620921702

[8] Merriam-Webster Medical Dictionary. 2021. "Etiopathogenesis." Accessed December 2021. https://www.merriam-webster.com/medical/etiopathogenesis.

[9] National Institute on Drug Addition. 2006. "Animal Experiments in Addiction Science" Last modified April 01. https://archives.drugabuse.gov/news-events/nida-notes/2006/04/animal-experiments-in-addiction-science.

[10] National Institute on Drug Addition; National Institutes of Health; U.S. Department of Health and Human Services. 2020. "MDMA (Ecstasy/Molly) Drug Facts." Last modified June 2020. https://www.drugabuse.gov/publications/drugfacts/mdma-ecstasymolly.

[11] Schneider Jr., R., Ottoni, G., de Carvalho, H., Elisabetsky, E., and D. Lara. 2015. "Temperament and character traits associated with the use of alcohol, cannabis, cocaine, benzodiazepines, and hallucinogens: evidence from a large Brazilian web survey." Brazilian Journal of Psychiatry 37(1): 31-39. https://doi.org/10.1590/1516-4446-2014-1352.

[12] Simons, J., Correia, C., Carey, K., and B. Borsari. 1998. "Validating a five-factor marijuana motives measure: Relations with use, problems, and alcohol motives." *Journal of Counseling Psychology* 45(3): 265–273. https://doi.org/10.1037/0022-0167.45.3.265.

[13] Slutske, W., Heath, A., Madden, P., Bucholz, K. Statham, D., and N. Martin. 2002. "Personality and the genetic risk for alcohol dependence." *Journal of Abnormal Psychology*, 111(1): 124-133. https://doi.org/10.1037/0021-843X.111.1.124.

[14] Soto, C. and J. Jackson. 2020. "Five-Factor Model of Personality," *Oxford Bibliographies in Psychology*. Oxford: Oxford University Press.

[15] Terracciano, A., Löckenhoff, C., Crum, R., Bienvenu, O., and P. Costa (Jr). 2008. "Five-Factor Model personality profiles of drug users." *BMC Psychiatry* 8(1), Article 22: https://doi.org/10.1186/1471-244X-8-22

[16] University of California, Irvine (UCI). 2016. "Drug consumption (quantified) Data Set." https://archive.ics.uci.edu/ml/machine-learning-databases/00373/.

[17] U.S. Department of Health and Human Services. 2016. "Sidebar: The Many Consequences of Alcohol and Drug Misuse." https://addiction.surgeongeneral.gov/sidebar-many-consequences-alcohol-and-drug-misuse.

[18] Vreeker, A., Brunt, T., Treur, J., Willemsen, G., Boomsma, D., Verweij, K., and J. Vink. 2020. "Comparing ecstasy users and non-users in a population-based and co-twin control design across multiple traits." *Addictive Behaviors* 108, Article 106421: https://doi.org/10.1016/j.addbeh.2020.106421.

******* INCLUDE R Code File as Appendix
Appendix: R Code