Course Number: Data 620 *Instructor: Osei Tweneboah*

Personality Traits, Demographics, and Drug Usage

Vedika Shaily[1,*]

1. Ramapo College; vshaily@ramapo.edu
★ Correspondence: e-mail@ramapo.edu

Abstract: This study used the Drug Consumption (Quantified) dataset obtained from the UC Irvine Machine Learning Repository dataset. This dataset was taken from a 2017 study that gathered online responses regarding personality traits and drug usage. The study hypothesized that traits such as neuroticism and sensation seeking would be associated with more drug use, alongside low agreeableness. Both unsupervised(hierarchical clustering) and supervised(logistic regression with lasso) methods were used. In the end, the project found that people fell into archetypes that determined the types of drugs they used, rather than certain personality traits having a black and white impact, it was the way these traits interacted with each other that determined the drugs used. Additionally, the supervised learning techniques found that education had opposite effects on alcohol and cannabis use, with higher education indicating more recent alcohol use and less cannabis use. Limitations of this study include a lack of diverse demographic features, and an inability to evaluate one of the strongest predictors of cannabis use, the country code. Further research should seek to address these limitations.

Keywords: drug usage; substance abuse; hierarchical clustering, lasso

---

1. Introduction

Drug usage is arguably one of the most significant public health topics throughout history–with interventions primarily focused on prevention and safe treatment[1]. Outside of the way that it interferes with users lives and their futures, it often can even lead to death; over 105,000 individuals died in the U.S. from drug-involved overdoses in the U.S. in 2023. This included both illicit and prescription drugs[2].

This highlights another issue in the medical field, where overprescription of addictive substances often leads to people developing addictions that they may not have had they not been prescribed those drugs, or even just prescribed such a large amount of the drug. Understanding the different risk factors for individuals in relation to substance abuse can help both public health and medical professionals take action to prevent usage. Studies have already shown that prevention is an effective intervention, especially in adolescents.Therefore, it is important for scientists and other individuals in the field to understand the potential demographics and individuals who may be more inclined towards engaging in substance usage.

The NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking) personality traits are all measurable traits well-established to be significant in studies[3]. For example, the “Big Five” (NEO-FFI-R) was tested to see if it could apply to larger populations while maintaining its Cronbach’s α, keeping it at 0.67 – 0.81, further solidifying its efficacy while allowing researchers to observe different patterns previously established reemerging—for example, women exhibit more

agreeableness and neuroticism, while younger individuals are more extraverted and less agreeable[4]. Therefore, using them as a basis for examining which individuals may need interventions could be fruitful. This study specifically used unsupervised clustering, namely hierarchical, to observe what clusters naturally form in the personality types, and then what drug usage these clusters had associated with them[5]. Then, a logistic regression used with lasso was used to evaluate what things are more closely related with alcohol and cannabis usage, as they are both commonly used recreationally here in the US[6][7][8].

2. Materials and Methods

The dataset chosen for this project is the Drug Consumption (Quantified) dataset obtained from the UC Irvine Machine Learning Repository. This dataset includes 1885 instances, in which each row represents an individual, their demographic information(such as their age, ethnicity, nationality, gender, and level of education), their NEO-FFI-R personality traits (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), and their drug usage throughout their lives for 18 drugs, both legal and illegal. The drugs observed were: alcohol, amphet, amyl, benzos, caffeine, cannabis, chocolate, coke, crack, ecstasy, heroin, ketamine, legalh, LSD, meth, mushrooms, nicotine, semer(a fictitious drug), and VSA(volatile substance abuse)[9]. The data was collected through a study—"The Five Factor Model of personality and evaluation of drug consumption risk"—performed by members of various Universities and Hospitals in the United Kingdom and in Iraq. All 1885 instances were obtained via online surveys, and they were entered into the data accordingly[10].

The data for the demographic and personality traits were represented in their z-scored forms—this made it prepared for analysis. The drug usage entries were classified as CL0, CL1, CL2, CL3, CL4, CL5, and CL6. Respectively, these represent never used, used over a decade ago, used in the last decade, used in the last year, used in the last month, used in the last week, and used in the last day. In the dataset, there were no missing values, and all of the columns were going to be used, so the cleaning was minimal to none. The next step of the analysis was to begin the hierarchical clustering. The SciPy library was used for this, with Matplotlib being used for subsequent visualization. When the data was loaded in, it was loaded in a X(features) and y(targets) format—both of these being dataframes. The first step involved checking which kind of linkage to use; single, average, or complete. Initially of the three average was deemed the most ideal, however further analysis revealed that the clusters were extremely small, with some being only one or two instances. Therefore, Ward linkage was used instead, as Ward linkage minimizes the variance between clusters[11]. Then, the amount of clusters in the dendrogram depending on the height was checked, and at a height of 30, six clusters were chosen. This amount of clusters was chosen as it gave a good amount of unique clusters while still being small enough to be comprehensible for the audience. For example, initially five clusters were used, but the sixth cluster(which split the largest of the five clusters into two) was found to contain two groups that were distinct between one another, with one having very high neuroticism individuals who had low conscientiousness and extraversion, and the second group having high neuroticism individuals who were also high sensation seeking. These

groups also ended up having different drug habits, which will be discussed further in the discussion section.

After the clusters were established, a new column was created in the dataframe to assign each instance to whatever cluster it was assigned to. This enabled further analysis of the results. Things that were looked at were cluster sizes, personality traits found in each cluster, and then the top three most used drugs in each cluster. A heatmap was created in order to view the concentration of each personality trait looked at throughout the data for each cluster. It should be noted that when quantifying drug usage, only "recent" drug usage was looked at—for the purposes of this study, this included within the last year, month, week, and day. Subsequently, the top three was done a second time excluding drugs that are not explicitly illegal in any nation, specifically caffeine and chocolate[12]. This was so that more "controversial" drugs could be observed. Finally, a heatmap was done via Seaborn in order to observe the proportion of recent users for each drug.

Next, for the supervised stage, the data was modeled via a L^1 penalized logistic regression with lasso. This was done for both alcohol consumption and cannabis use, separately, but identically. Within this, there were a few things specified. First, penalty="l1" specified that the L^1 penalty was being used, ,solver="saga" specified that saga was being used over liblinear as it is faster and better for larger datasets. Class imbalance was handled by class_weight="balanced"(chosen due to high prevalence of certain habits), and the iteration cap was raised to 20 000 to eliminate convergence warnings that were appearing[13][14]. Next, the predictors were the personality traits and the demographic information(age band, gender, education level, country, ethnicity). The data was split into a training and test set, scaled, and then scikit-learn's logistic regression CV was performed on them. This was chosen because it will yield a shorter list of risk factors. After the model was fitted, an ROC curve was plotted, and then the coefficients were printed out. These coefficients are what was used to see what traits were significant(or not equal to 0).

3. Results

3.1 Figures

3.1.1 Hierarchical Clustering

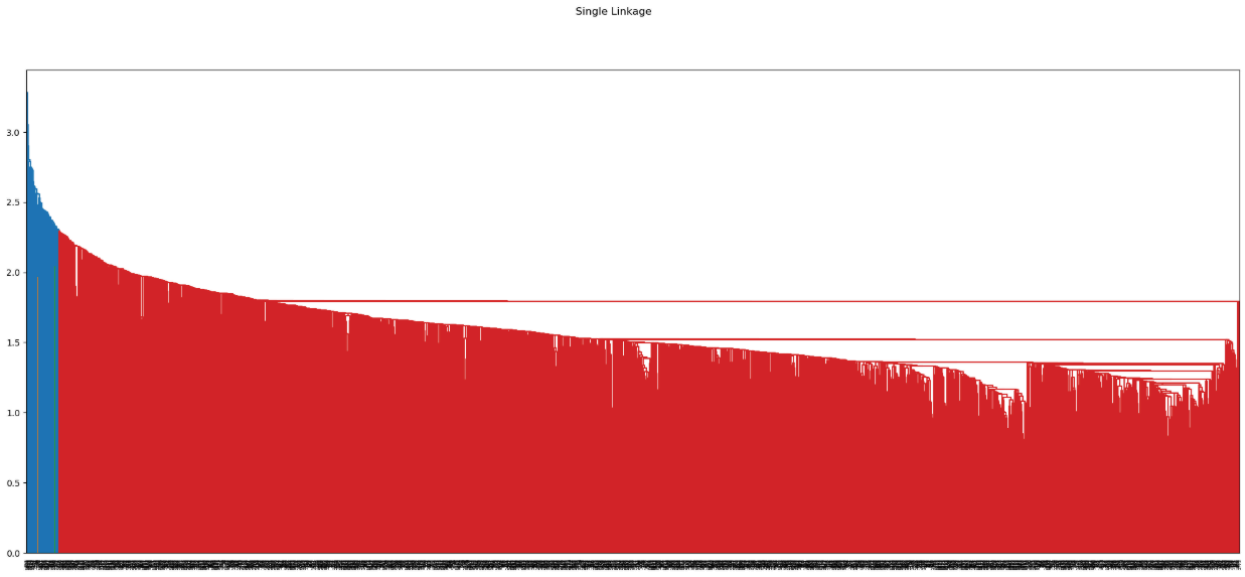

Figure 1: This figure shows the dendrogram produced by the Single Linkage.

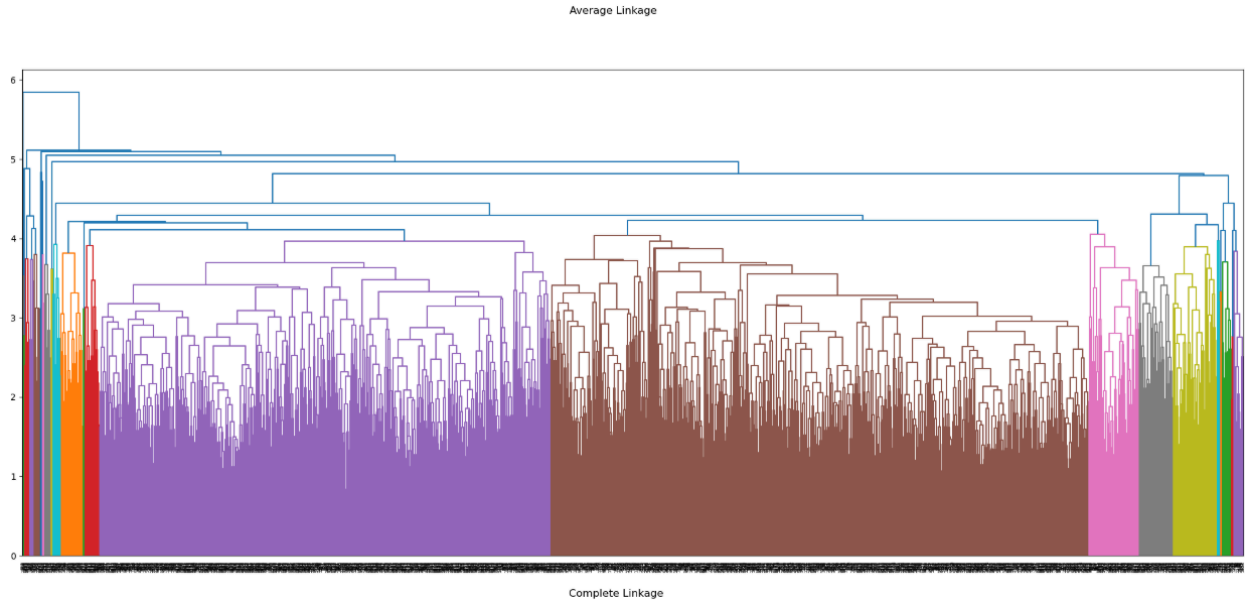

Figure 2: This figure shows the dendrogram produced by the Average Linkage.

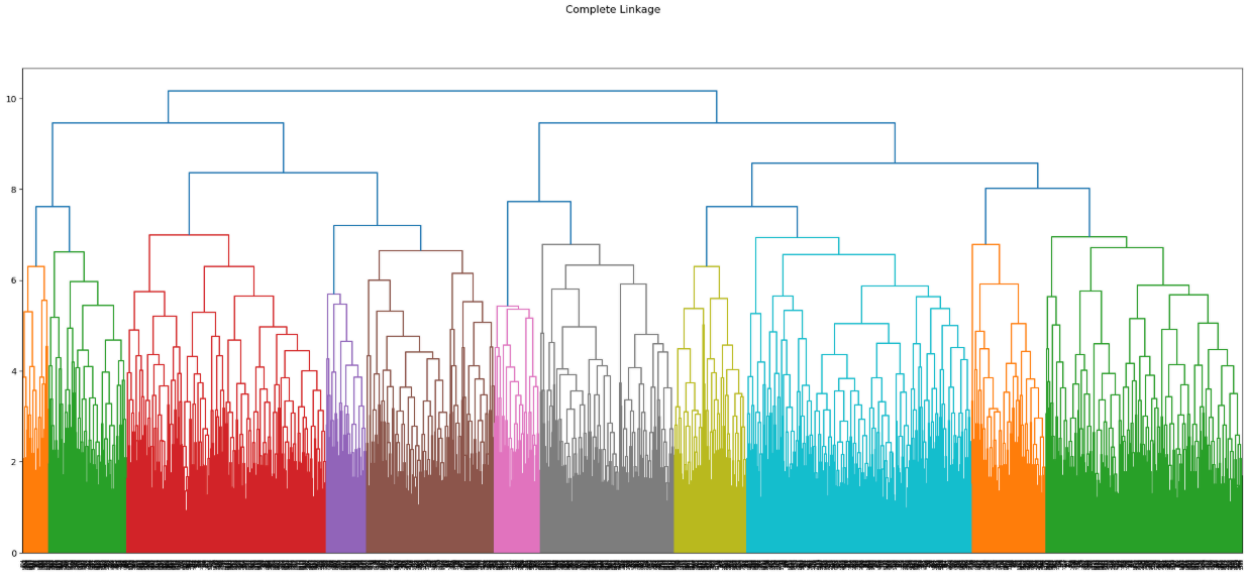

Figure 3: This figure shows the dendrogram produced by the Complete Linkage.

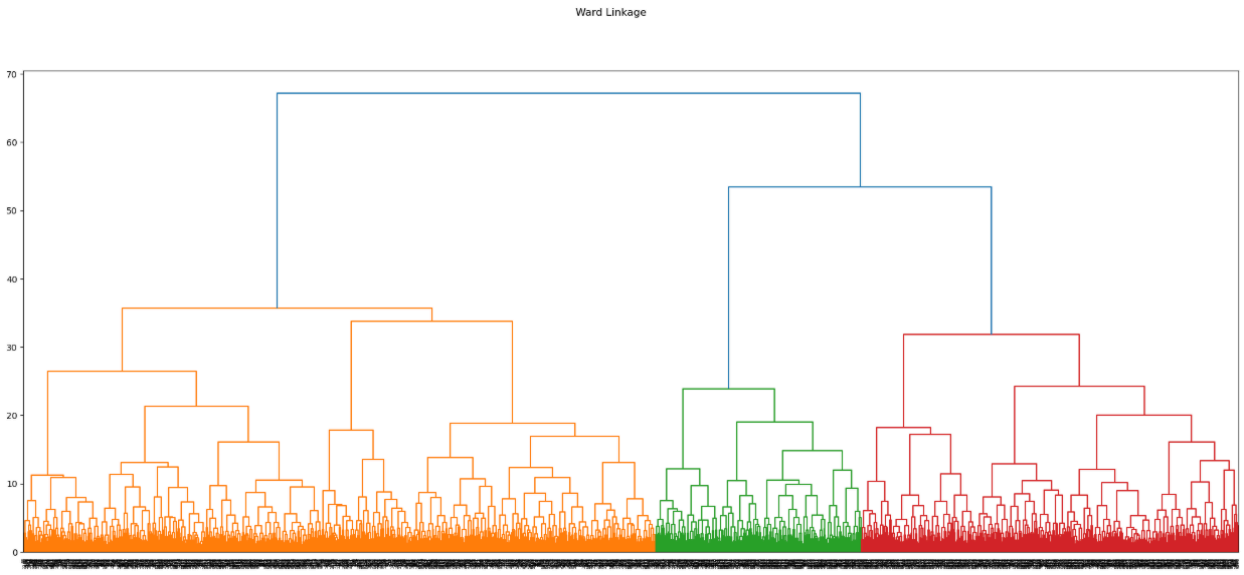

Figure 4: This figure shows the dendrogram produced by the Ward Linkage.

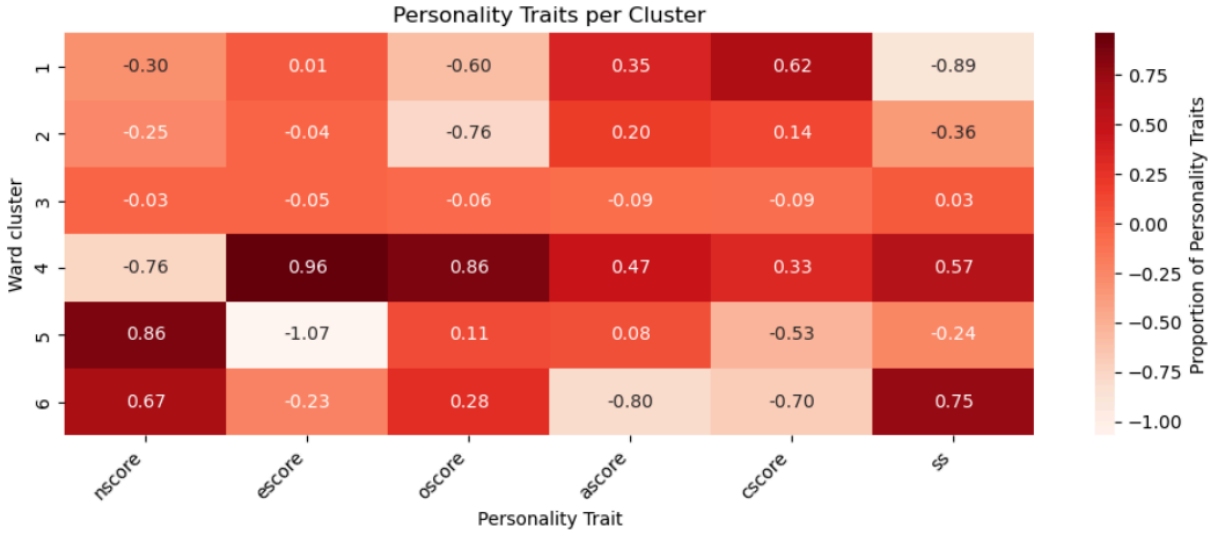

Figure 5: This shows a heatmap for the personality traits per cluster. The darker the color is, the more concentrated or the higher proportion of people in that cluster possess that personality trait. It is on a -1 to 1 scale as the initial entries were all the standardized z-scores.

```
Top-3 most-used (excluding caffeine & chocolate):
Cluster 1: ['alcohol', 'nicotine', 'cannabis']
Cluster 2: ['alcohol', 'nicotine', 'cannabis']
Cluster 3: ['alcohol', 'nicotine', 'cannabis']
Cluster 4: ['alcohol', 'cannabis', 'nicotine']
Cluster 5: ['alcohol', 'cannabis', 'nicotine']
Cluster 6: ['alcohol', 'cannabis', 'nicotine']
```

Table 1: This table shows the top three most used “recent” drugs per each cluster, excluding caffeine and chocolate.

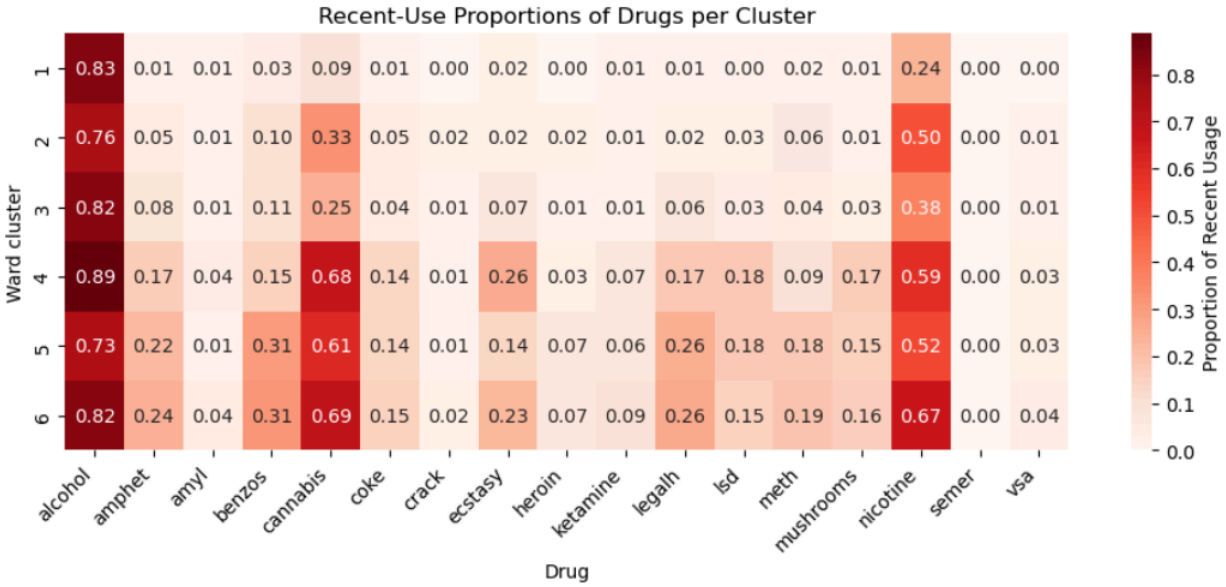

Figure 6: This heat map shows the recent-use proportions of drugs per cluster. The deeper the shade of red, the higher the recent usage percentage is.

3.1.2 LogisticRegressionCV and Lasso

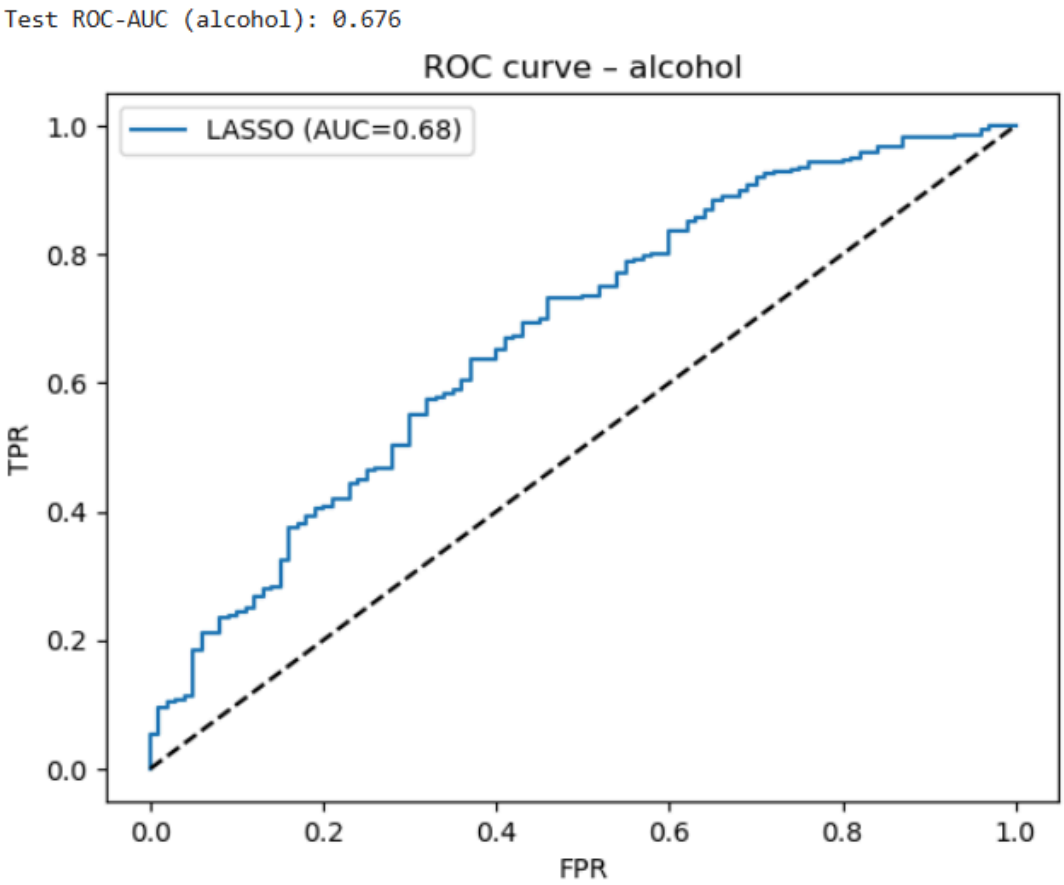

Figure 7: This graph shows the ROC curve for the alcohol lasso model.

education	0.218807
escore	0.204590
country	0.132630
ss	0.108223
age	-0.102779
ethnicity	0.016506
oscore	0.007509

Table 2: This table shows the coefficients obtained from the model, which can then be used to determine the impact of these features on alcohol usage.

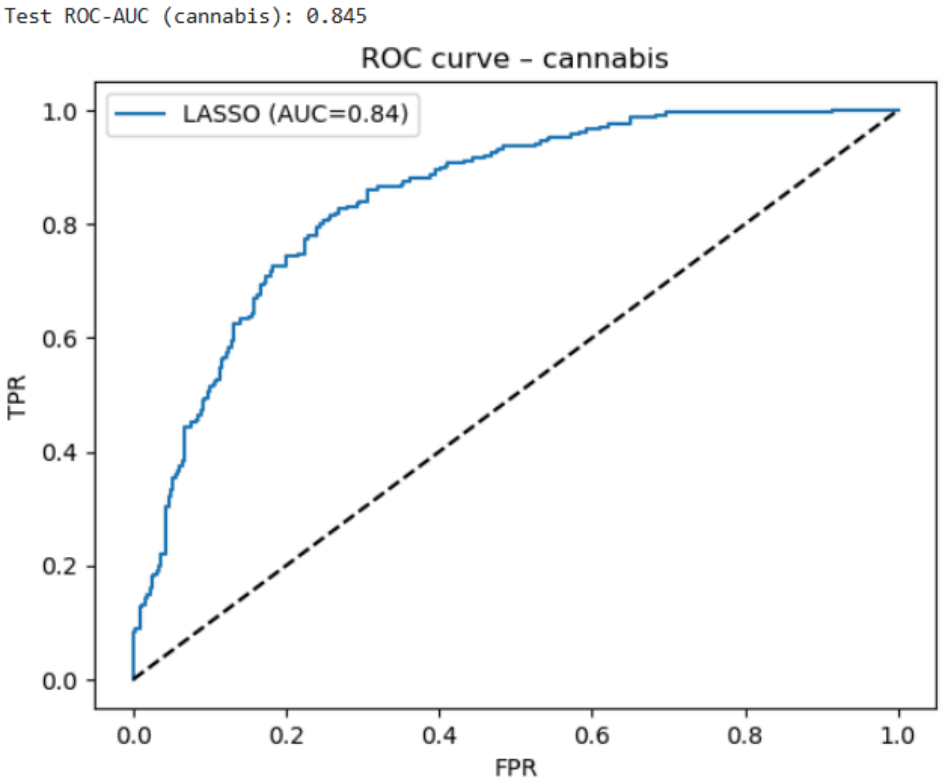

Figure 8: This graph shows the ROC curve for the cannabis lasso model.

country	-0.783876
age	-0.605497
oscore	0.589124
education	-0.515506
ss	0.370974
gender	-0.238483
cscore	-0.210063
nscore	-0.173264
ethnicity	0.101572
ascore	0.049302
impuslive	0.000426

Table 3: This table shows the coefficients obtained from the model, which can then be used to determine the impact of these features on alcohol usage.

4. Discussion

This study hypothesized that as traits like neuroticism and sensation seeking increased and agreeableness decreased, that the usage of illicit drugs would be more present. In addition to this, the study hypothesized that older individuals would be less likely to be recent users, and that traits like neuroticism would be associated with cannabis and alcohol use. The first steps began with the unsupervised clustering of the data—specifically of the traits and demographic characteristics.

Figures 1-4 show the dendrograms for Single, Average, Complete, and Ward linkage. Figure 1 shows very nonoptimal branching, to the point that you can't even make out the development of the tree well. In contrast, average and complete linkage have clearer branching, however, as mentioned earlier, the clusters produced by them were extremely disproportionate. As a result, Ward linkage was used in the end.

Figure 5 is a heatmap that displays the concentrations of the personality traits in each cluster. Figure 6 is a heatmap that displays the concentrations of recent drug usage per cluster. Combining the information gleaned from both of these charts yields interesting possibilities. In cluster 1, we see high conscientiousness, and high openness. This is interesting because the most common drug used here was alcohol, with all the other concentrations for drugs being incredibly low. As the personality traits indicate, this is a type of person who is careful of how they conduct themselves and what they are doing(conscientiousness) and who is polite and open to experiences(openness)—combined with only alcohol being high, this suggests an individual who may only be a social drinker. Cluster two has no standout personality traits, all of them being slight concentrations. This suggests that this cluster can serve as the "average" person. Here, the concentration of alcohol drinkers is once again high(0.76) however, there is also a notable concentration of nicotine users(0.50) and a slight cannabis presence(0.33). This immediately accounts for what are commonly seen as the most present/common drugs used by the average person. The next cluster, cluster 3, shows individuals who are even closer to the 0 value for traits than the previous cluster—once again, these individuals have high alcohol usage rates(0.82), however cannabis and nicotine use are both lower—0.25 and 0.38 respectively. This confirms that groups closer to the center will have lower usage rates for all drugs except for alcohol, cannabis, and nicotine. Cluster 4 shows individuals who have extremely high extraversion(0.96), high openness(0.86), and high sensation seeking(0.57). As predicted, this results in a group who has higher drug usage rates with illicit substances—immediately alcohol, cannabis, and nicotine climb to their highest values yet(0.89, 0.67, and 0.58), and substances like ecstasy(0.26), amphetamines, benzos, LSD, and other illicit substances see a spike from their previous scores. This makes sense because the person with these traits is the kind of person who is a thrill seeker who is always willing to try new things(extreme extravert with openness to new experiences who is specifically seeking sensation). It makes sense that they would have tried more drugs across the board than the average person. Next, cluster 5 shows an extreme introvert(-1.07 extraversion) who also has high neuroticism(0.86), with relatively low conscientiousness and sensation seeking. These kinds of traits indicate someone who is anxious and potentially more prone to self medication—the usage rates line up, with alcohol being high(0.73), cannabis—which is commonly used as an anti-anxiety drug—being high(0.61), and legal prescription drugs and benzos also seeing a spike compared to previous rounds, but no stimulants standing out. Lastly, cluster 6 shows individuals with high sensation seeking(0.75), high neuroticism(0.67), low agreeableness(-0.70) and low conscientiousness(-0.80). The kind of person this can be interpreted as is someone who is anti-social, lacks regard for others, and seeks out stimulation without much regard for its impacts. This profile, fittingly, has drug spikes across the board, with both stimulants and more common drugs like alcohol, cannabis, and nicotine. This fits with the personality described.

The next analysis was the supervised logistic regression and lasso that was performed to see what traits had impacts on recent alcohol and cannabis use. Figure 7 shows the ROC curve for the lasso alcohol model—the 0.676 ROC-AUC indicates that there was a good class separation. Similarly, Figure 8, which displays the ROC curve for recent cannabis use, also has a high ROC-AUC at 0.845, which is actually even better. Table 2 shows the coefficients provided by the alcohol model. Here, we see that a one-SD increase in the respondent's education level raises the log-odds of having drunk alcohol recently by 0.219. Additionally, more extraverted respondents are more likely to have drank alcohol recently, which ties into the social drinkers point that was brought up earlier. The other coefficients present have relatively low indications. Table 3 displays the same concept but for the cannabis model. Here, we see that country code is a strong indicator, specifically that a higher country code means that your recent likelihood of drinking is higher—this is difficult to interpret in the country code context. Other coefficients that show up are older age meaning that someone is less likely to have used cannabis recently, that people with higher openness scores are more likely to have used cannabis recently, people with less education are more likely to have used cannabis recently, females are more likely to have used it recently, and then that higher sensation seeking also raises the likelihood of recent cannabis use. The further coefficients all begin to crawl closer to 0, indicating less of an impact on the likelihood of recent cannabis consumption.

5. Conclusion

In this study, the relationship personality traits and some demographic traits and their relationship to drug usage habits. There were some interesting conclusions. There appears to be evidence that a correlation towards "neutral" or less extreme manifestations of the personality traits tended towards conservative drug usage—namely, only alcohol appeared at a high proportion of recent usage. Conversely, individuals with more extreme scores tended towards using more drugs at higher proportions. The study hypothesized that as neuroticism and sensation seeking increased, and agreeableness decreased, that illicit drug usage would increase. This study found that it was more nuanced than that, there were more combinations and archetypes of people that had to be considered depending on the cluster formed, and through that a greater understanding of the different correlations and relationships that existed in the clusters was gleaned.

The supervised learning yielded results that were interesting as well, for example, that higher education levels indicated more alcohol drinking but less cannabis usage, and vice versa. Something that could be explored is the mapping of country codes and the relationship with them and cannabis use, to potentially see any climate or geological links considering how high the indication was. Further analysis could also take a look at incorporating additional demographic features, such as things like criminal history, kids, marital status, etc. These things could help address some of the limitations of the study, which would allow for further research and interventions to be developed for vulnerable populations.

Author Contributions: Conceptualization, V.S.; methodology, V.S..; software, Python; validation, V.S.; formal analysis, V.S..; investigation, V.S; resources, V.S.; data curation, V.S..; writing—original draft preparation, V.S.; writing—review and

editing, V.S..; visualization, V.S.; supervision, O.T.; project administration, V.S.; All authors have read and agreed to the published version of the manuscript."

Data Availability Statement: While in the reference section, the links to the datasets used are provided here.
https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified

References

1. Lo, T. W., Yeung, J. W. K., & Tam, C. H. L. (2020). Substance Abuse and Public Health: A Multilevel Perspective and Multiple Responses. *International journal of environmental research and public health*, *17*(7), 2610. https://doi.org/10.3390/ijerph17072610

2. Coi, A et. al.(2025).Updated Overdose Statistics 2025: Trends in Drug-Related Deaths. *Addiction Group.* https://www.addictiongroup.org/resources/overdose-statistics/

3. Carducci, B. J. (2009). *The psychology of personality: Viewpoints, research, and applications*. John Wiley & Sons.

4. Körner, A., Czajkowska, Z., Albani, C., Drapeau, M., Geyer, M., & Braehler, E. (2015). Efficient and valid assessment of personality traits: population norms of a brief version of the NEO Five-Factor Inventory (NEO-FFI). *Archives of Psychiatry & Psychotherapy*, *17*(1).

5. Tweneboah, O. (2025). Cluster_Analysis. Ramapo College of New Jersey.

6. Tweneboah, O. (2025). Logistic_Regression. Ramapo College of New Jersey.

7. Tweneboah, O. (2025). Variable_Selection_Regulaarization. Ramapo College of New Jersey.

8. Patrick M. E. (2025). Daily or near-daily cannabis and alcohol use by adults in the United States: A comparison across age groups. *Addiction (Abingdon, England)*, *120*(4), 779–782. https://doi.org/10.1111/add.16748

9. Fehrman, E., Egan, V., & Mirkes, E. (2015). Drug Consumption (Quantified) [Dataset]. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5TC7S.

10. Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V., & Gorban, A. N. (2017). The five factor model of personality and evaluation of drug consumption risk. In *Data science: innovative developments in data analysis and clustering* (pp. 231-242). Springer International Publishing.

11. W3Schools. (n.d.) Machine Learning - Hierarchical Clustering. *Refsnes Data*. https://www.w3schools.com/python/python_ml_hierarchial_clustering.asp

12. Jackson M. T. (n.d.) Is caffeine illegal anywhere in the world? *ChefsResource.* https://www.chefsresource.com/is-caffeine-illegal-anywhere/

13. Mila. (2025). Fine-Tuning Logistic Regression Models Using Lasso and Ridge Regularisation. *MyCollegeAI*. Fine-Tuning Logistic Regression Models Using Lasso and Ridge Regularisation

14. scikit-learn developers. (n.d.) LogisticRegressionCV. *scikit-learn.* https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegressionCV.html