Multivariate Analysis of Drug Consumption Patterns

An Applied Statistics Project in Data Science

Spring 2025

Course: AD (Multivariate Data Analysis Part)
Degree: Data Science and Engineering

Kaggle/UCI Dataset: Drug Consumption

Authors:
Matías Mora
Jordi Ferré
Roger Velilla

Universitat Politècnica de Catalunya
June 1, 2025

Contents

1 Introduction

The dataset used in this project was obtained from the UCI Machine Learning Repository and contains self-reported information on personality traits, sociodemographic factors, and drug consumption. The goal of this analysis is to explore the dataset using multivariate statistical techniques in order to identify meaningful patterns and relationships among the variables. Our motivation lies in understanding the factors associated with drug consumption—one of the most alarming and socially impactful phenomena worldwide. Substance use not only affects individual health and behavior but also has far-reaching consequences on public health, safety, and socioeconomic structures. Detecting underlying profiles is not only useful for identifying individuals at risk, but also essential for gaining a deeper understanding of the complex mechanisms behind this widespread issue.

2 Exploratory Data Analysis

2.1 Dataset Description

This dataset contains self-reported information from $n = 1885$ individuals regarding their sociodemographic profile, psychological traits, and drug consumption behavior. Each row corresponds to a different respondent. After an initial cleaning process, no missing values were found.

The variables can be classified as follows:

- **Sociodemographic variables:** All categorical variables, including: Age (6 intervals), Gender (Male/Female), Education level (ranging from no formal education to doctoral degree), Country of residence (e.g., UK, USA, Canada), and Ethnicity (e.g., White, Black, Asian, Mixed, Other).
- **Psychological tests:** Nscore (Neuroticism), Escore (Extraversion), Oscore (Openness to experience), AScore (Agreeableness), and Cscore (Conscientiousness), all derived from the NEO-FFI-R model. Additionally, Impulsive measures impulsiveness according to the BIS-11 scale, and SS refers to sensation seeking as measured by the ImpSS scale. Each of them are handled as numerical variables.
- **Substance use variables:** Participants reported their frequency of use for 18 substances, including common legal drugs such as alcohol, caffeine, and nicotine, as well as less commonly used or illicit substances such as heroin, cocaine, LSD, and amphetamines. The original scale ranged from CL0 (Never Used) to CL6 (Used in Last Day). A fictitious drug, Semer, was included to detect unreliable answers; all respondents who reported use of Semer were removed, and the variable was discarded.

A brief correlation analysis among the psychological traits shows that the strongest relationship appears between Impulsive and SS, suggesting that individuals who tend to act impulsively also exhibit a higher propensity for seeking novel and intense experiences. Additionally, we observe moderate correlations between AScore and Cscore, indicating that more agreeable individuals also tend to be more disciplined and organized. Negative

correlations between Nscore and both Escore and Cscore suggest that individuals with higher levels of neuroticism are generally less sociable and less conscientious—patterns that align with established findings in personality psychology.

It is also worth noting that most participants fall into the 18–24 age category, the vast majority of participants are White, and the most represented country of residence is the United Kingdom, followed by the United States. Gender, on the other hand, is almost perfectly balanced across the dataset.

In terms of drug use, we also find a vast imbalance. The class CL0 (Never Used) is the most frequent response in many of the drug-related variables, especially for less commonly used substances such as heroin, LSD, or amphetamines. In contrast, more socially accepted and widely consumed substances such as alcohol, nicotine, and caffeine exhibit more balanced usage distributions across categories, as expected. This overall imbalance in category frequencies—will be now managed.

2.2 Dataset Preprocessing

Several preprocessing steps were applied to prepare the dataset for multivariate analysis:

- The Education variable, originally composed of 9 detailed levels, was grouped into three categories: No studies, Compulsory studies, and Advanced studies. This way we managed the large imbalance and favored the interpretation.
- Drug consumption variables were recoded from the original 7-level scale into 3 interpretable (and balanced) categories :
  - For frequent-use substances (alcohol, caffeine, nicotine): Last day, Last month, More than a month ago.
  - For the rest: Never, Recent, More than a month ago.
- Outliers were identified using standardized z-scores on the continuous psychological traits. Individuals with any absolute z-score greater than 3 were removed from the dataset ($n = 30$).
- To assess normality, Shapiro–Wilk tests and histograms were evaluated. Most variables followed distributions close to Gaussian. However, Impulsive and SS, being discretized variables, deviated more clearly. A Box–Cox transformation was applied but did not significantly improve normality. Further details are provided in Appendix A.1.

Depending on the method used in each stage of the analysis, only a selected subset of the substance use variables will be considered—typically those with greater interpretability or prevalence in the sample.

We will as well consider the creation of other variables throughout the multivariate analysis, to perform a deeper exploration of potential interactions and gain a more comprehensive understanding of the underlying structure in the data.

3 Multivariate Analysis

3.1 Principal Component Analysis

To reduce dimensionality while preserving the psychological structure of the data, we performed a Principal Component Analysis (PCA) on the seven standardized continuous variables. All of them are going to be considered since PCA is not especially sensitive to little deviations from normality. It must be said that we are only considering the psychological profile to group individuals in this PCA, as these are the only numerical variables available. Therefore, we do not expect to find a separation as clear as in MDS. However, if the psychological profile is indeed relevant enough to create a visible separation, that would provide us with valuable insight on how drug consumption can affect psychological traits (and viceversa).

In order to interpret the results more clearly, we focus only on a subset of the categorical variables — those related to the most socially relevant or frequently analyzed substances: **Alcohol, Nicotine, Cannabis, Cocaine, Ecstasy, Heroin, Amphetamines,** and **Benzodiazepines**. These, together with the sociodemographic variables such as **Gender, Education, Age, Country,** and **Ethnicity**, are included as *supplementary categorical variables*.

According to the scree plot (see B, the first three components explain approximately **72%** of the total variance, with large dominance of the first two. The interpretation (which is endorsed by Figure 1) is the following:.

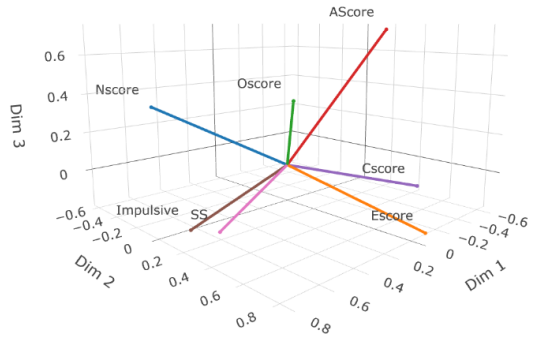

Figure 1: 3D plot of the dimension contributions of numerical variables

- **Dimension 1** shows strong positive loadings for Impulsive (0.79), SS (0.75), Nscore (0.48), and Oscore (0.43), and strong negative loadings for Cscore (-0.64) and AScore (-0.48). This dimension contrasts individuals who are impulsive, sensation-seeking, emotionally reactive, and open to experience against those who are conscientious, organized, and prosocial.

- **Dimension 2** has relevant positive loading from Escore (0.69) and moderate contributions from Oscore (0.49) and Cscore (0.41), and a moderate negative loading from Nscore (-0.53). This dimension may represent a balance between sociability and emotional stability, contrasting extroverted, open, and organized individuals with those who are emotionally unstable and introverted.

- **Dimension 3** (13.5% of the variance) is mainly defined by AScore (0.72), followed

by Oscore (0.45) and Nscore (0.32). It reflects a contrast between individuals who are cooperative, empathetic, and sensitive to others' needs, and those who tend to be more distant, emotionally reserved, or interpersonally rigid.

To interpret the principal component space in relation to group characteristics, we projected the supplementary categorical variables, in order to observe possible separations. Among all these relationships, we selected two that showed the most significant and interpretable patterns. We will onyl use the first two dimensions, in which we observe the following interpretable patterns:

- First, **Dimension 1** effectively separates individuals with recent consumption of harder drugs, such as cocaine, heroin, or amphetamines, as seen in figure 2 as an example. Respondents who have used these substances in the last year or month are concentrated towards the *positive* side of Dimension 1, suggesting they tend to score higher in impulsivity, sensation seeking, neuroticism, and openness, and lower in agreeableness and conscientiousness. This aligns with expectations, as recent or frequent users of such substances are typically associated with higher disinhibition and emotional vulnerability. It must be noted that "non-recent" consumers appear more dispersed; this may be due to the fact that each plot focuses on a single substance. As a result, individuals who consume one hard drug but not another may appear as non-recent users and misalign the interpretation, even though they naturally have a similar psychological profile than others.

- This effect does **not appear** for legal or more commonly used substances (e.g., alcohol, nicotine, or cannabis), which show no clear differentiation along any principal axis. These drugs are widely used across different personality profiles, and therefore PCA does not distinguish a typical user profile for them.

- Secondly, we observe (see figure 3) that **individuals from the United States** tend to project clearly on the positive side of Dimension 1. This might indicate a tendency towards more impulsive or novelty-seeking psychological profiles in the U.S. subgroup of respondents in comparison with the rest of the world, or reflect cultural or reporting biases.

- On the other hand, **Dimension 2** does not separate any categorical variable meaningfully. There is no observable trend linking this axis with age, gender, ethnicity, or any drug use category. This suggests that traits like extraversion, openness, or neuroticism, captured in this dimension, do not have a direct statistical relationship with either sociodemographic traits or drug consumption.

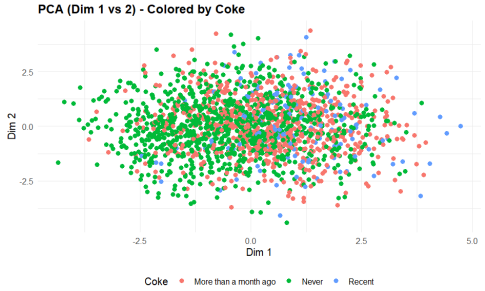

Figure 2: PCA (Dim 1 vs 2) colored by Cocaine use

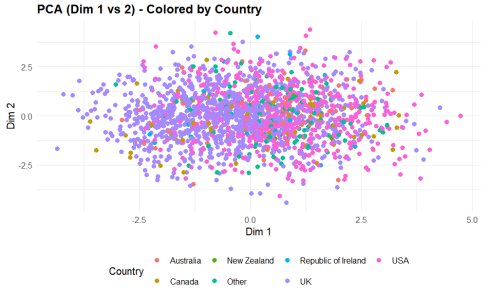

Figure 3: PCA (Dim 1 vs 2) colored by Country

3.2 Multidimensional Scaling Analysis

To capture relationships between individuals across both numerical and categorical variables, we applied **Multidimensional Scaling (MDS)** using the **Gower distance**, which is suitable for mixed data types. We used the same data subset as in the PCA analysis, including psychological traits, sociodemographic variables, and use of selected drugs, but we will focus on categories we have not already drawn conclusions. Our study will consist on, just like in PCA, discerning how the categorical variables are grouped in the now 2 dimensional distribution of our data.

We applied classical metric MDS via `cmdscale`, as well as nonlinear techniques like `isoMDS` and `sammon`. However, the plots yielded by all three methods were visually similar, and the Shepard diagram was not clearly linear in non-metric alternatives. Thus, we will focus on the metric configuration due to its interpretability and stable structure. Indeed, we observed through the Scree Plot that he first two MDS dimensions provide a faithful low-dimensional representation of the data, justifying a 2D analysis.

A key advantage of this method over PCA is the incorporation of categorical variables in distance computation. This results in a much clearer separation of individuals with similar drug use profiles, rather than just psychological similarities. Firstly, we can clearly see a dense cloud on the left hand side of the configuration, which actually corresponds a **well-defined cluster of non-consumers of hard drugs**, since they all appear to have the category "never" for every hard drug labeling of the map, as seen in figures 4 and 5 for instance.

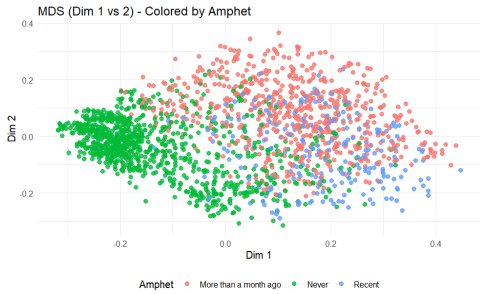

Figure 4: MDS (Dim 1 vs 2) colored by Amphetamine use

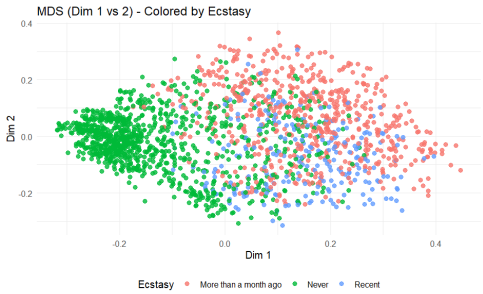

Figure 5: MDS (Dim 1 vs 2) colored by Ecstasy use

On the other end of the configuration, we find recent or frequent users of any hard drugs like heroin, amphetamines, and ecstasy, grouped more to the right.This affirms our previous PCA hypothesis, since MDS clearly supports the that strong drug users tend to share common psychological and demographic traits, such as drug historial. However, legal and commonly consumed substances like alcohol or nicotine are more evenly distributed across the map, suggesting that their consumption is not strongly tied to a specific general profile, as it may seen in figure 6.

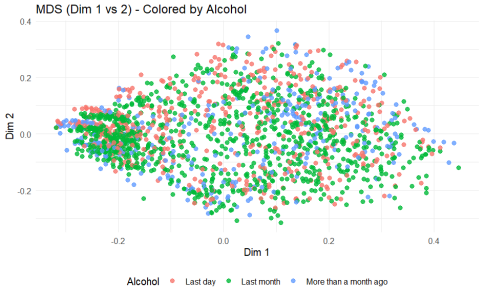

Figure 6: MDS (Dim 1 vs 2) colored by Alcohol use

Regarding sociodemographic variables (apart from country, which we have already considered), we observe distinct patterns that suggest meaningful associations with psychological traits and drug use. **Gender differences** are particularly evident (see figure 8): **males tend to cluster on the right-hand side of the MDS space**, an area associated with higher drug involvement, while females are more frequently located on the left cluster — suggesting that men are more likely to engage in the use of harder substances.

In addition, **Dimension 2** appears to reflect an **age-related gradient**, as figure 7 shows. Younger individuals, especially those in the 18–24 group, are concentrated in the lower part of the map, whereas older participants — particularly those aged 45–54 and 55–64 — occupy the upper region. No other relation has been seen in this axis, so the separation is clear.

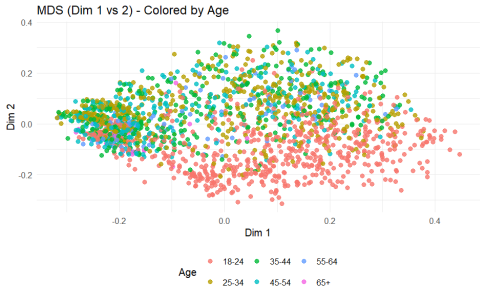

Figure 7: MDS (Dim 1 vs 2) colored by Age group

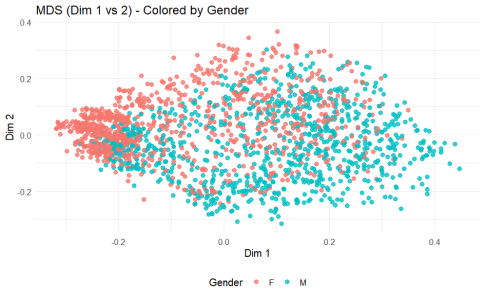

Figure 8: MDS (Dim 1 vs 2) colored by Gender

Lastly, we wanted to analyze how individuals differ in terms of their underlying psychological traits along the MDS dimensions. When coloring the configuration by variables such as Sensation Seeking (SS) and Agreeableness (AScore), it becomes evident that the axis which separates recent users of hard drugs from non-users (X axis) also corresponds to a contrast in psychological profile. Specifically, drug users tend to score higher in SS and lower in AScore, suggesting a personality pattern characterized by greater impulsivity, thrill-seeking behavior, and reduced prosocial tendencies. This observation reinforces the findings from our PCA analysis, where these same traits were associated with positive values in the primary dimension, aligning with a profile of emotional reactivity, disinhibition, and reduced interpersonal sensitivity. Relations can be seen in figures 10, 9.

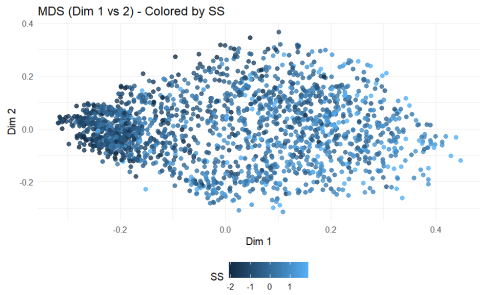

Figure 9: MDS (Dim 1 vs 2) colored by Sensation Seeking (SS)

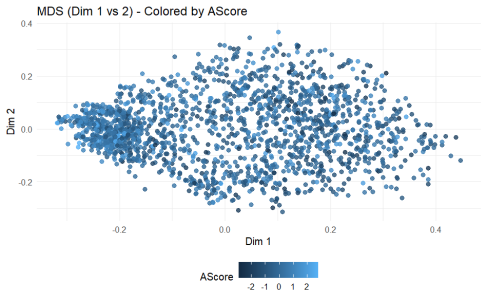

Figure 10: MDS (Dim 1 vs 2) colored by Agreeableness (AScore)

3.3 Correspondence Analysis

In this section, we explore the relationship between education level and patterns of hard drug use. Unlike previous sections, we will include every hard drug variable — excluding only cannabis, alcohol, caffeine, and chocolate — to construct a comprehensive behavioral categorization. Given the number and imbalance of these variables, analyzing them individually would obscure general patterns. Instead, we define a unified factor based on the most recent consumption of *any* hard drug, assigning individuals to levels such as *Used in Last Day*, *Used Last Month*, etc. This variable condenses drug involvement into an interpretable form and, when analyzed in conjunction with education level, may reveal structural associations between educational attainment and the severity or recency of drug use.

The CA map reveals a clearly dominant first dimension, accounting for **93.96%** of the total inertia, suggesting that most of the variation in the table can be interpreted along a single axis.

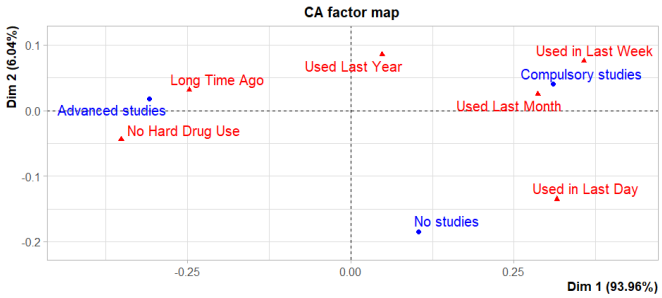

Figure 11: Correspondence Analysis of Education vs. Drug Use Status

Individuals with *no studies* are positioned closest to the most recent usage categories—*Used in Last Day* and *Used Last Week*—indicating a pronounced tendency toward ongoing or frequent hard drug consumption. Those with *compulsory studies* appear aligned with intermediate levels of use, such as *Used Last Month*, suggesting more sporadic but still relatively recent engagement. In contrast, individuals with *advanced studies* cluster near the categories *Long Time Ago* and *No Hard Drug Use*, implying either discontinued or absent drug use histories.

This spatial arrangement supports a robust inverse relationship: as education level increases, the likelihood and frequency of hard drug use diminishes notably.

3.4 Multiple Correspondence Analysis

We will apply Multiple Correspondence Analysis to investigate how categorical patterns of drug use relate to key sociodemographic variables. The analysis proceeds in two stages: the first focusing on *hard drugs* (e.g., heroin, crack, LSD), and the second on *legal or commonly used substances* (alcohol, cannabis, and nicotine). To assess the associations of drug use with personal characteristics, we include **gender**, **age group**, and **country** as *supplementary variables*. These variables are projected onto the MCA space to aid interpretation but do not influence the construction of the principal dimensions, ensuring that the observed structure is driven solely by patterns of drug consumption.

Hard drugs. The first analysis uses only the consumption variables related to hard drugs: `Amphet, Coke, Crack, Ecstasy, Heroin, Ketamine, LSD, Meth`. As shown in figure 22, the first two dimensions capture a substantial share of the total inertia and provide a clear and interpretable configuration.

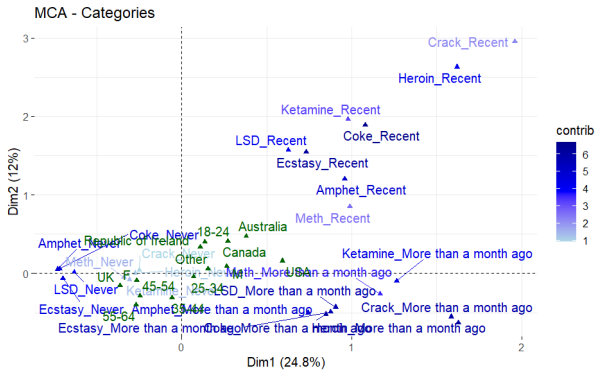

Figure 12: MCA - Categories for Hard Drug Consumption and sociodemographics

The plot, together with a dimension contribution analysis, reveals several key structures:

- **Dimension 1** (23.8%) distinguishes **never-users** from all others. Categories like `LSD_Never`, `Ecstasy_Never`, and `Amphet_Never` cluster on the left, while past and recent users project to the right.
- **Dimension 2** (18.9%) further separates **recent users**, who occupy the upper-right region (e.g., `Crack_Recent`, `Heroin_Recent`), from those who used drugs in the past (e.g., `Coke_More than a month ago`), who appear lower.

Sociodemographic variables project meaningfully onto the MCA space. While most groups tend to cluster near the center — suggesting no extreme associations — we observe that **younger individuals (18–24)** and respondents from **non-European countries**, particularly **the USA**, are displaced toward the upper-right region of the map. This area aligns with categories representing *recent* or *past* use of hard drugs. In contrast, **older age groups** (45–64) and **European countries** such as the UK or Ireland remain closer to the origin, indicating weaker associations with strong drug use. Notably, the USA

appears especially close to categories denoting **past consumption** (e.g., "More than a month ago"), which may reflect a pattern of experimentation or discontinued use rather than ongoing abuse. We will use the help of indivudal plots.

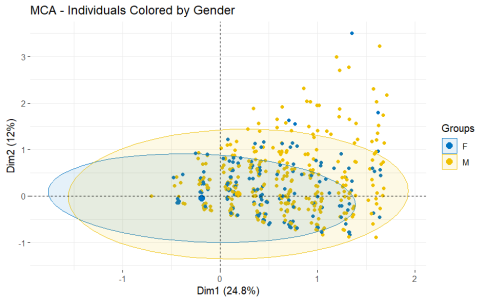

Figure 13: MCA - Individuals colored by Gender

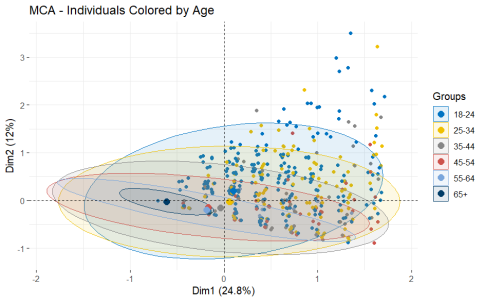

Figure 14: MCA - Individuals colored by Age group

These individual-level plots confirm what the category map suggests: men are more represented in zones of higher drug use, even though the centroid is concentric with the female respondents, and younger age groups align with those same patterns, since their ellipse show a higher dispersion towards the upper right corner.

Legal drugs. We next applied MCA to the variables related to *alcohol, cannabis,* and *nicotine*. As shown in Figure 21, the first two dimensions account for a quite vast majority of the explained inertia, and — most importantly — provide the clearest and most interpretable structure. Higher-order dimensions do not reveal additional meaningful separation between categories, suggesting that the essential relationships are already captured in the 2D map.

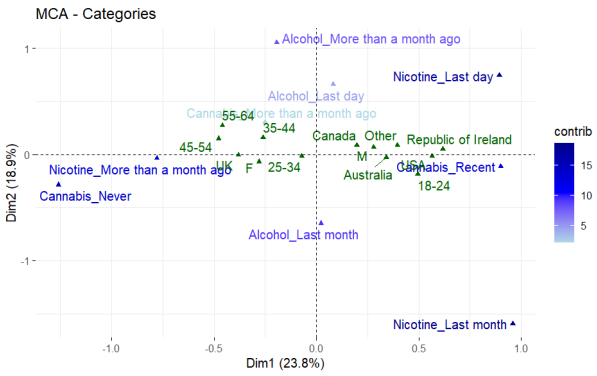

Figure 15: MCA - Categories for Alcohol, Cannabis, Nicotine and sociodemographics

- **Dimension 1** clearly contrasts **non-users** (e.g., `Cannabis_Never`, `Nicotine_More than a month ago`) against **recent users** (e.g., `Cannabis_Recent`, `Nicotine_Last day`).
- **Dimension 2** separates **heavier users** of alcohol and nicotine (e.g., `Nicotine_Last day`, from **moderate users** (e.g., `Nicotine_Last month`, `Alcohol_Last month`). `Alcohol_More than a month ago` may disalign, but its contribution is low.

Again, younger participants appear in zones associated with recent cannabis use, as well as USA, — which is consistent with more permissive legal frameworks and cultural acceptance. **Males** also cluster closer to high-consumption categories, whereas **females** tend to be located nearer non-use areas. It is also safe to say that older generations tend to cluster around categories indicating no cannabis use and appear to be former users of nicotine. Additionally, individuals from European countries show a markedly lower association with legal or commonly used substances, suggesting overall lower consumption rates in these regions.

3.5 Clustering Analysis

3.5.1 Methodology and Dimensionality Reduction

In this section, we investigate the presence of distinct psychological profiles in the dataset using clustering techniques. The analysis is based on the seven standardized continuous variables.

Since these dimensions tend to be intercorrelated, we applied Principal Component Analysis (PCA) to reduce dimensionality and mitigate collinearity. We retained the first **three principal components**, which together explained over 70% of the variance in the data.

To validate the robustness of our findings, we also conducted a parallel clustering analysis on the raw standardized personality traits. Both approaches led to highly consistent results in terms of cluster composition and associated patterns of substance use. To avoid redundancy and simplify interpretation, we present only the PCA-based clustering results in the main report. The full code and additional comparisons with the raw-data clustering are provided in the accompanying script.

Hierarchical clustering using Ward's method on the PCA scores suggested an optimal partition into three clusters, as supported by:

- the **elbow method** on total within-cluster sum of squares (see Figure 23),
- the **maximum average silhouette width** at $k = 3$
- and a visal inspection of the **dendrogram**, shown below.

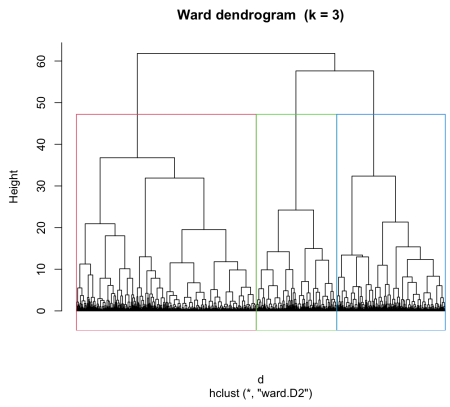

Figure 16: Ward dendrogram from hierarchical clustering on PCA scores.

3.5.2 Cluster Profiling

A k-means algorithm with $k = 3$ was run on the PCA scores, using the centroids obtained from Ward's method as initial seeds. The resulting silhouette average was 0.26, indicating moderate but meaningful separation.

Figure 17 displays the result of the k-means clustering ($k = 3$) projected onto the first two principal components. Each point corresponds to an individual, and the shape and color indicate the assigned cluster. The ellipses help visualize the compactness and separation of the clusters in the reduced PCA space.

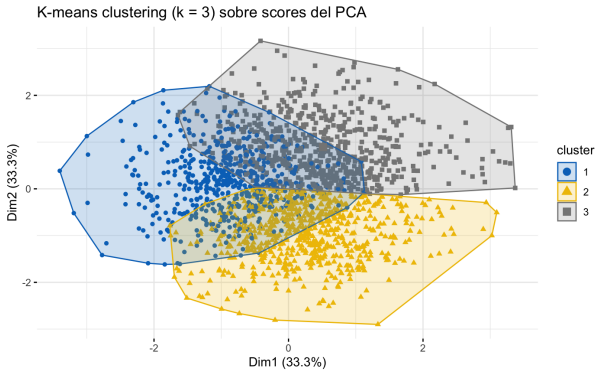

Figure 17: K-means clustering with $k = 3$ projected on the first two principal components.

Table 1 summarizes the average value of each personality trait within the three clusters.

float

Table 1: Mean of personality traits per cluster

Cluster	Nscore	Escore	Oscore	AScore	Cscore	Impulsive	SS
1	-0.59	0.85	0.60	0.25	0.34	0.34	0.51
2	-0.22	-0.19	-0.65	0.28	0.45	-0.80	-0.86
3	0.75	-0.53	0.14	-0.51	-0.77	0.53	0.43

Each cluster corresponds to a distinct psychological profile. Below is a summary of the key characteristics of each group:

- **Cluster 1: Sociable, open, emotionally stable.**
  This group is characterized by high extraversion and openness, moderate agreeableness and conscientiousness, low neuroticism, and moderately high impulsivity and sensation seeking. These individuals likely enjoy socializing and new experiences while maintaining emotional balance and moderate self-control.

- **Cluster 2: Conventional, structured, low-impulsive.**
  Members of this cluster score very low on openness, impulsivity, and sensation seeking, while maintaining high levels of conscientiousness and agreeableness. They appear disciplined, conformist, and risk-averse. Their low extraversion suggests a more introverted lifestyle, and the low neuroticism indicates good emotional control.

- **Cluster 3: Emotionally unstable and impulsive.**
  This profile stands out for its high neuroticism, low agreeableness and conscientiousness, and high impulsivity and sensation seeking. These individuals are more prone to emotional instability, poor impulse control, and may seek stimulation as a coping strategy or lifestyle preference.

3.5.3 Relationship Between Psychological Profiles and Substance Use

To examine how personality profiles relate to substance use, we analyzed the distribution of alcohol, nicotine, cannabis, cocaine, and ecstasy consumption across the three clusters. The results are presented in a series of stacked bar charts (see Annex for selected figures).

The most relevant findings are summarized below:

- **Cluster 1** (sociable and open) shows high levels of alcohol and cannabis use, likely reflecting socially driven or recreational consumption patterns. Ecstasy and cocaine use are also present but less prevalent than in Cluster 3.
- **Cluster 2** (structured and reserved) reports the lowest levels of use across all substances, especially for stimulants such as cocaine and ecstasy. This is consistent with their low sensation-seeking and impulsivity scores.
- **Cluster 3** (impulsive and neurotic) exhibits the highest levels of recent use for most substances, particularly nicotine, cannabis, cocaine, and ecstasy. These patterns align with the psychological profile of high impulsivity, emotional instability, and low conscientiousness.

3.6 Discriminant Analysis

In this section we will use discriminant analysis to check if personality traits can be used to predict whether or not a person is a drug user. Firstly, let us define what we mean by drug user. Amongst all the substances listed, there are four which we label as *frequent use drugs*, which are chocolate, alcohol, nicotine and caffeine. Now, all the other drugs will be labeled *hard drugs*. We will say somebody falls into the category of **DrugUser** if he/she has taken some hard drug recently.

Performing Shapiro-Wilk test into our variables gives that only Nscore follows normality. Still, by looking at the histograms and QQ-plots of the other variables, they seem normal, excepts maybe for Impulsive and SS. Still, we will be taking them into account, as we will see they are good predictors.

QDA

Performing Box's M-test shows, with high confidence, that the covariance matrices of our variables are not equal. Therefore, the assumption of homogeneity of covariance matrices is violated, and we cannot apply LDA. Thus, we will be applying QDA, which gives us the following results:

	Predicted Non-DrugUser	Predicted DrugUser
Non-DrugUser	649	239
DrugUser	221	767

Table 2: Prediction results of the QDA.

The model has a 74.52% correctness rate, and checking at the Q-statistic we see that it is greater than 3.84, indicating that it performs significantly better than the random model. Also, we check that it is not biased, as it correctly labels 73.75% of Non-DrugUsers, and $76,24\%$ of DrugUsers. One can visualize this classification projected onto the first two principal components on Figure 29.

Regarding the coefficients of the model, we showcase them in Table 3. We see how DrugUsers are caracterized by having significantly higher SS, Impulsive and Oscore values than Non-DrugUsers, while having lower Cscore, and maybe Ascore (not very significant difference for this last one). See how this is consistent with the results we obtained in the Clustering Analysis section, where Cluster 2 had the lowest SS, Impulsive and Oscore values by far, and ranked the lowest in drug consumption.

	Nscore	Escore	Oscore	AScore	Cscore	Impulsive	SS
Non-DrugUser	-0.142	0.057	-0.386	0.181	0.301	-0.314	-0.445
DrugUser	0.126	-0.054	0.341	-0.163	-0.272	0.292	0.292

Table 3: Mean of personality traits of DrugUsers against Non-DrugUsers

We now apply the stepwise function to see which variables are the most significant. If we run it backwards, SS is the last one to be deleted, and Impulsive the second last one. If we run it forwards, it only takes SS. This indicates that SS may be the only statistically significant variable. If we perform the model only taking variable SS, we achieve a 70.20% correctness rate, and the results are shown in Table 5.

Observe that it makes sense that SS (recall, stands for sensation-seeking personality trait) predicts if someone is a DrugUser decently. This justifies our choice of taking SS and Impulsive variables into account even if they didn't look normal.

Naive Bayes

Applying the Chi-Squared test we obtain that all our variables satisfy the independence assumptions. Now, this means we can apply Naive Bayes, which leads to the following results:

	Predicted Non-DrugUser	Predicted DrugUser
Non-DrugUser	653	238
DrugUser	235	750

Table 4: Prediction results of the Naive Bayes model.

This prediction has a 74.78% correctness rate, and similarly to the QDA one, does not

look biased. We also check that the Q-statistic is greater than 3.84, and thus this model looks significantly better than the random model.

3.7 Hotelling T^2

The objective of this test is to check if two groups are significantly different. We will be studying the same groups as in the Discriminant Analysis section, that is, Group A: Drug Users; Group B: Non-DrugUsers. That is, we are doing the test $\begin{cases} H_0 & : \mu_1 = \mu_2 \\ H_1 & : \mu_1 \neq \mu_2 \end{cases}$, where μ_1 and μ_2 are the averages of groups A and B, respectively. Recall that there are 7 variables, so μ_i is a vector of the average of the 7 variables, corresponding to an average psychological profile.

In order to apply this test, all the variables need to satisfy normality, and also have equal covariance matrices. Normality is not a problem, but we've already mentioned that we don't have homogeneity. Thus, we apply an aproximation of the Hotelling T^2 test using a chi-squared distribution. After computing the Q-statistic we check that it is greater than the critical value, that is, we conlcude, with high confidence, that there is a significant difference between groups, which aligns with the separability found in the previous section.

3.8 MANOVA

Lastly, we will perform a MANOVA analysis. This test checks if more than two groups have the same mean or not. We will be grouping people into 4 categories, depending on which hard drug they have taken recently, particularly, Group 1 **Stimulants**: Amphet, Coke, Meth and Ecstasy; Group 2 **Hallucinogenous**: LSD, Mushrooms, Ketamine and Legalh; Group 3 **Depressants**: Benzos, Heroin and Amyls; Group 4 **Cannabis**.

We recall that we are doing the test $\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1 : \mu_i \neq \mu_j \text{ for some } i, j \end{cases}$. Performing the test tells us, with a lot of confidence, that we should reject H_0, that is, that there is a significative difference between some of the groups.

More interestingly, we also conducted a Tukey post-hoc test on all the personality trait variables. The trait that showed the greatest variation was Nscore, which reflects levels of anxiety, sadness, and anger. This score was highest among individuals using Depressant drugs and lowest among Cannabis users, suggesting a strong link between emotional instability and depressant use. Notably, Sensation Seeking (SS) was the second most variable trait, peaking among Stimulant users and reaching its lowest levels among Depressant users. This pattern reinforces the idea that different substance types are associated with distinct psychological profiles, particularly regarding emotional reactivity and impulsive behavior.

4 Conclusions

This project aimed to uncover and interpret the complex relationships between psychological traits, sociodemographic characteristics, and patterns of drug consumption through

a broad set of multivariate analysis techniques.

Principal Component Analysis (PCA) revealed a distinct psychological profile associated with recent users of hard drugs. Individuals scoring high in impulsivity, sensation seeking, and neuroticism—and low in agreeableness and conscientiousness—were clearly separated along the first principal component, suggesting that personality plays a key role in distinguishing between consumers and non-consumers.

Multidimensional Scaling (MDS) further supported this distinction by incorporating both numerical and categorical variables via Gower distance. The resulting configuration showed a clear spatial separation between non-users and recent or frequent users of hard substances. It also revealed noticeable differences in consumption patterns across age and gender groups, enriching the interpretation provided by PCA.

Correspondence Analysis (CA) showed a strong inverse relationship between education level and drug use severity. Individuals with lower levels of education were more frequently associated with recent consumption of hard drugs, while those with advanced studies tended to cluster near non-user profiles.

Multiple Correspondence Analysis (MCA) provided additional insights by aligning categorical consumption patterns with sociodemographic traits. Recent or past use was more prevalent among young males—especially from the United States—whereas older individuals and those from European countries were more often linked to non-consumption categories. This analysis highlighted clear behavioral tendencies by age, gender, and region.

Cluster analysis uncovered three meaningful psychological profiles. One group, marked by high emotional instability and impulsivity, exhibited the highest levels of drug use. Another, characterized by more conventional and structured personality traits, showed minimal consumption. These clusters revealed how psychological dimensions can segment the population into interpretable and behaviorally relevant subgroups.

Discriminant analysis using Quadratic Discriminant Analysis (QDA) and Naive Bayes classifiers achieved classification accuracies above 74%. Sensation seeking emerged as the most important predictor, demonstrating that psychological traits alone can be effective in anticipating substance use status.

Hotelling's T^2 test confirmed a statistically significant difference between the psychological profiles of users and non-users, reinforcing the idea that these two groups differ not only in behavior but also in underlying psychological structure.

MANOVA, followed by Tukey post-hoc tests, revealed significant differences in the psychological profiles of users across different drug categories. In particular, users of Depressant drugs stood out, exhibiting higher levels of anxiety, sadness, and anger, along with the lowest levels of sensation seeking.

In conclusion, this project illustrates the power of multivariate techniques to extract meaningful and interpretable patterns from complex behavioral data. The findings offer valuable insights into the psychological and sociodemographic factors influencing drug consumption, with potential implications for research, prevention, and public policy.

References

[1] Obey Khadija. *Drug Consumptions (UCI Dataset)*. Kaggle, 2022. Disponible en: https://www.kaggle.com/datasets/obeykhadija/drug-consumptions-uci

[2] Fehrman, Craig; Muhammad, Omar; Egan, Vincent. *Drug Consumption (Quantified)*. UCI Machine Learning Repository, 2017. Disponible en: https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified

[3] Costa, Paul T. y McCrae, Robert R. *Revised NEO Personality Inventory (NEO PI-R) y NEO Five-Factor Inventory (NEO-FFI)*. Psychological Assessment Resources, 1992.

[4] Patton, J.H.; Stanford, M.S.; Barratt, E.S. Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology*, 1995, 51(6), 768–774.

[5] Zuckerman, M.; Kuhlman, D.M.; Joireman, J.; Teta, P.; Kraft, M. A comparison of three structural models for personality. *Journal of Personality and Social Psychology*, 1993, 65(4), 757–768.

A Data Preprocess

A.1 Numerical Variables Histograms

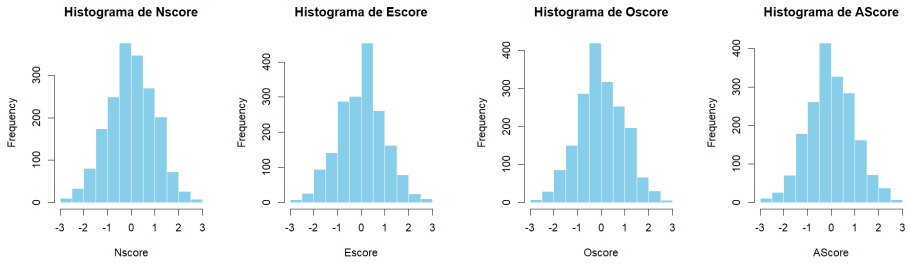

Figure 18: Histograms of Nscore, Escore, Oscore, Ascore.

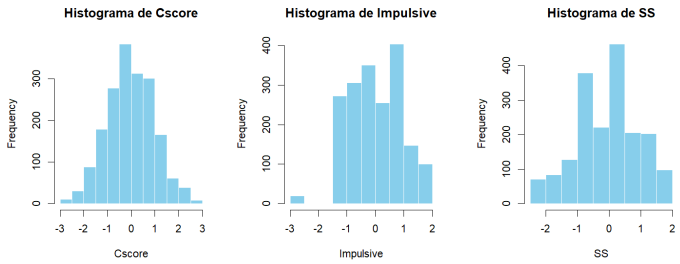

Figure 19: Histograms of Cscore, Impulsive, SS.

B Principal Component Analysis

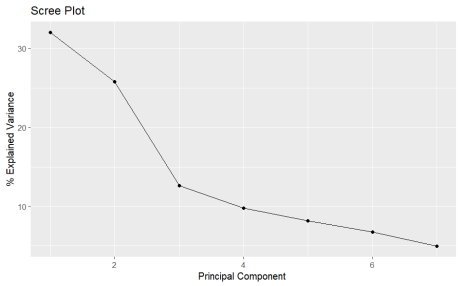

Figure 20: Scree Plot of PCA

C Multiple Correspondence Analysis

C.1 Dimension Inertia Contributions

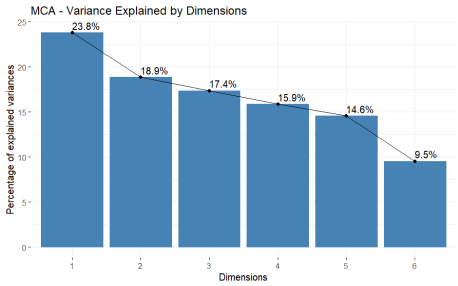

Figure 21: Contributions to the first two dimensions – Legal drugs MCA

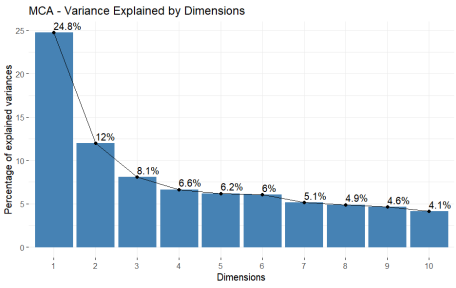

Figure 22: Contributions to the first two dimensions – Hard drugs MCA

D Clustering Analysis

D.1 Methodology and Dimensionality Reduction

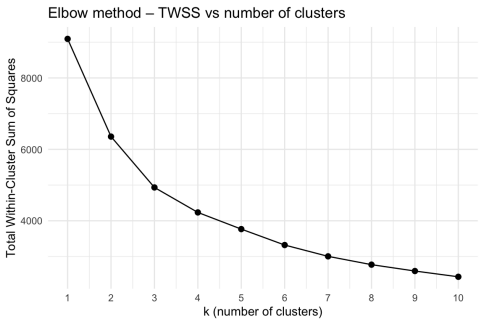

Figure 23: Total Within-Cluster Sum of Squares for different values of k.

D.2 Relationship Between Psychological Profiles and Substance Use

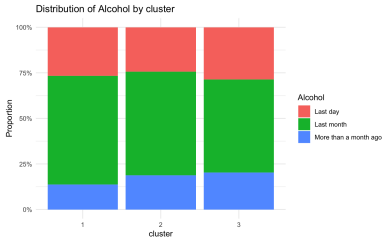

Figure 24: Distribution of alcohol consumption across clusters.

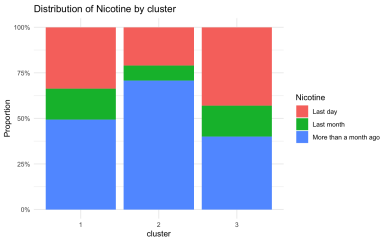

Figure 25: Distribution of nicotine consumption across clusters.

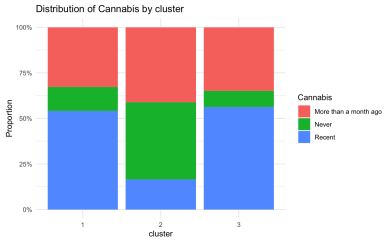

Figure 26: Distribution of cannabis consumption across clusters.

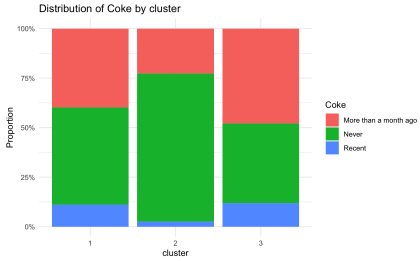

Figure 27: Distribution of coke consumption across clusters.

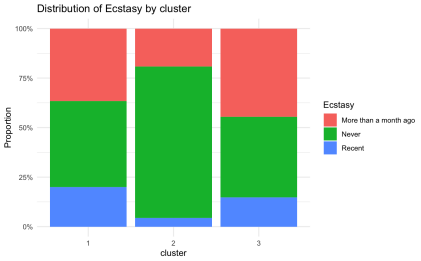

Figure 28: Distribution of ecstasy consumption across clusters.

E Discriminant analysis

E.1 QDA classification projected onto the first two principal components

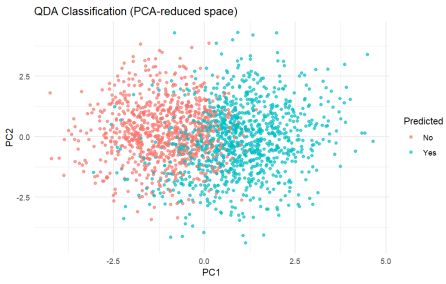

Figure 29: QDA classification projected onto the first two principal components

E.2 Prediction results of QDA only taking variable SS

	Predicted Non-DrugUser	Predicted DrugUser
Non-DrugUser	610	278
DrugUser	281	707

Table 5: Prediction results of the QDA using only the variable SS.