Scientific Article

Predicting Drug Consumption Behavior from Personality Traits: A Quantitative Analysis Using the UCI Drug Consumption Dataset

Robben Wijanathan

Abstract

This project analyzes drug consumption patterns using a multivariate dataset containing demographic, personality, and behavioural attributes. Through preprocessing, exploratory data analysis, visualization, and predictive modelling in R, the study aims to identify key factors associated with drug usage and evaluate model performance. The results highlight meaningful behavioural trends across demographic groups and reveal significant correlations between personality traits and substance use.

Keywords: *Drug Consumption, Personality Traits, Machine Learning, Classification, Behavioral Analysis, Social Science*

1. Introduction

1.1. Background

Understanding the behavioural and demographic factors that influence drug usage is essential for designing effective prevention and intervention strategies. The dataset used in this study includes self-reported information on nineteen drug types, along with personality trait scores and demographic attributes such as age, gender, education, ethnicity, and country. These variables offer a broad view of the social and psychological context surrounding substance use.

Through structured data preprocessing, exploratory analysis, and visualisation, meaningful patterns are identified across user groups. Machine learning models are then developed to predict drug consumption and assess the contribution of each factor. This work aligns with student outcomes involving the development of analytical models, evaluation of model performance, and drawing contextual conclusions from real-world data.

1.2. Objectives

- To preprocess and explore the dataset, identifying trends in demographic and personality factors.
- To visualize drug usage distributions and behavioural relationships using appropriate R techniques.
- To build and evaluate predictive models for drug consumption using statistical and machine learning methods.
- To interpret model results and communicate insights through clear data.

2. Methods

2.1. The Dataset

The dataset used in this study is the *Drug Consumption (Quantified)* dataset obtained from the UCI Machine Learning Repository. It contains 1,885 instances and 32 attributes. The first attribute is an ID column, followed by demographic variables, personality trait scores, and 19 drug usage indicators.

```r
url <- "https://raw.githubusercontent.com/
    RobbenWijanathan/drug-consumption-regression
    /main/drug_consumption.csv"
data <- read_csv(url)

head(data)
```

Dataset Overview

```
> head(data)
# A tibble: 6 × 32
     ID   Age    Gender Education      Country Ethnicity Nscore
      Escore  Oscore AScore Cscore Impulsive
```

```
  <dbl> <chr> <chr>  <chr>          <chr>   <chr>       <dbl>
       <dbl>   <dbl>  <dbl>  <dbl>      <dbl>
1      2 25–34 M       Doctorate ...de UK       White
      −0.678   1.94    1.44     0.761  −0.143     −0.711
2      3 35–44 M       ...Professional UK       White
      −0.467   0.805  −0.847   −1.62   −1.01      −1.38
3      4 18–24 F       Masters ...degr UK       White
      −0.149  −0.806  −0.0193   0.590   0.585     −1.38
4      5 35–44 F       Doctorate ...de UK       White
      0.735  −1.63   −0.452   −0.302   1.31      −0.217
5      6 65+   F       Left school ... Canada   White
      −0.678  −0.300  −1.56     2.04    1.63      −1.38
6      7 45–54 M       Masters ...degr USA      White
      −0.467  −1.09   −0.452   −0.302   0.939     −0.217
# i 20 more variables: SS <dbl>, Alcohol <chr>, Amphet <chr
    >, Amyl <chr>, Benzos <chr>,
#   Caff <chr>, Cannabis <chr>, Choc <chr>, Coke <chr>,
    Crack <chr>, Ecstasy <chr>, Heroin <chr>,
#   Ketamine <chr>, Legalh <chr>, LSD <chr>, Meth <chr>,
    Mushrooms <chr>, Nicotine <chr>,
#   Semer <chr>, VSA <chr>
```

Output

2.1.1. Demographic Attributes

Five demographic attributes are included:

- **Age**: Categorical with levels {18–24, 25–34, 35–44, 45–54, 55–64, 65+}.
- **Gender**: {Male, Female}.
- **Education**: {Left school before 16, Left at 16, Left at 17, Left at 18, Some college/no certificate, Professional certificate/diploma, University degree, Masters degree, Doctorate degree}.
- **Country**: {Australia, Canada, New Zealand, Other, Republic of Ireland, UK, USA}.
- **Ethnicity**: {Asian, Black, Mixed-Black/Asian, Mixed-White/Asian, Mixed-White/Black, Other, White}.

2.1.2. Personality Trait Attributes

Seven continuous variables represent personality scores based on standardized psychological assessments:

- **Nscore**: Neuroticism
- **Escore**: Extraversion
- **Oscore**: Openness to Experience
- **Ascore**: Agreeableness
- **Cscore**: Conscientiousness
- **Impulsive**: Impulsiveness
- **SS**: Sensation Seeking

Each score is a continuous numerical value derived from the original psychometric questionnaire.

2.1.3. Drug Usage Attributes

The dataset contains 19 drug-related variables representing consumption patterns across a wide range of substances. Each variable is encoded on an ordinal scale (CL0–CL6), indicating recency of use:

Code	Meaning
CL0	Never used
CL1	Used over a decade ago
CL2	Used in the last decade
CL3	Used in the last year
CL4	Used in the last month
CL5	Used in the last week
CL6	Used in the last day

These ordinal values quantify drug-use recency, enabling both classification and regression analyses.

List of Drug Attributes

- **Alcohol** — Alcohol consumption class.
- **Amphet** — Amphetamines consumption class.
- **Amyl** — Amyl nitrite consumption class.
- **Benzos** — Benzodiazepine consumption class.
- **Caff** — Caffeine consumption class.
- **Cannabis** — Cannabis (marijuana) consumption class.
- **Choc** — Chocolate consumption class.
- **Coke** — Cocaine consumption class.
- **Crack** — Crack cocaine consumption class.
- **Ecstasy** — MDMA (ecstasy) consumption class.
- **Heroin** — Heroin consumption class.
- **Ketamine** — Ketamine consumption class.
- **Legalh** — "Legal highs" (synthetic recreational drugs) consumption class.
- **LSD** — Lysergic acid diethylamide consumption class.
- **Meth** — Methadone consumption class.
- **Mushrooms** — Psilocybin (magic mushrooms) consumption class.
- **Nicotine** — Nicotine/tobacco consumption class.
- **Semer** — Semeron, a *fictional drug* included to detect unreliable or invalid survey responses.
- **VSA** — Volatile substance abuse (inhalants) consumption class.

2.2. Data Visualization

Data visualization was performed in **R** using libraries such as `tidyverse`, `ggplot2`, and `fmsb`. Different visualization techniques were selected based on the type and characteristics of the variables.

For the **demographic attributes**, radar plots were used to compare average drug usage across categories such as Age, Gender, Education, Country, and Ethnicity. Radar plots are well suited for this purpose because they allow multivariate, category-based comparisons on a unified scale, making it easy to observe differences in usage patterns across multiple drug types simultaneously.

For the **personality attributes**, Spearman correlation heatmaps were employed. Personality traits are continuous variables, while drug usage levels are ordinal; therefore, Spearman's rank correlation is appropriate for capturing monotonic relationships without assuming linearity.

2.2.1. Spearman's Rank Correlation Coefficient

Let n denote the number of respondents. For each respondent $i \in \{1, \dots, n\}$ we observe a value of a personality trait P which we denote as x_i and a value of a drug usage attribute D, denoted y_i.

For a trait P, consider the values $x_1, \dots, x_n$. We then sort their values from smallest to largest and assign to each x_i the rank $R(x_i)$. The rank corresponds to the position of a value in this sorted list, so the smallest would be have rank 1, and the largest has rank n. Same values would receive the averaged rank.

The same is applied to the drug-use values $y_1, \dots, y_n$, to obtain $R(y_i)$ for each respondent. Now that we have $R(x_i)$ and $R(y_i)$, we can define the rank difference

$$d_i = R(x_i) - R(y_i)$$

Now we can compute the spearman's rank correlation coefficient between a trait P and drug-use attribute D, denoted $\rho_s(P, D)$ as such

$$\rho_s(P, D) = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}.$$

This yields $\rho_s(P, D) \in [-1, 1]$. Values closer to $+1$ indicate a strong positive monotonic association between P and D, values close to -1 indicate a strong negative monotonic association, and values near 0 indicate little or no monotonic association.

2.2.2. Correlation Matrix and the Heatmap

Let personality traits be indexed by $j = 1, \dots, J$ and drugs by $k = 1, \dots, K$, where $J = 7$ for 7 personality traits, and $K = 18$ for 18 drug usage attributes (ignoring the fictional drug, semer). For each pair (j, k), we compute

$$\rho_{j,k} = \rho_s(P_j, D_k),$$

and collect these into a correlation matrix

$$C = \begin{bmatrix} \rho_{1,1} & \cdots & \rho_{1,K} \\ \vdots & \ddots & \vdots \\ \rho_{J,1} & \cdots & \rho_{J,K} \end{bmatrix}$$

This matrix C will then be visualized as a heatmap, with rows corresponding to personality traits and columns to drugs, and each cell (j, k) colored according to the value of $\rho_{j,k}$. The heatmap enables clear visual identification of positive, negative, and neutral associations between personality dimensions and drug-use frequency across all substances at once.

Together, these visualization techniques provide a comprehensive overview of how demographic and psychological factors relate to drug consumption patterns in the dataset.

2.3. Predictive Modeling

We then construct a model to predict how demographic and personality variables jointly determine or at least relate to drug-use behavior, based on the dataset used. Each participant possesses a set of characteristics but potentially consumes multiple substances. Since the 18 drug-use variables might not be independent phenomena, it is only appropriate to model them simultaneously rather than in isolation. Therefore, we adopt a multivariate (multiple-output) regression model, which generalizes ordinary least squares to vector valued outputs. This approach accounts for shared factors across substances, avoids model fragmentation by not artificially treating each drug as a separate, unrelated problem.

2.3.1. Encoding Output Values

Each drug-use variable is recorded as categorical ordinal levels (CL0-CL6), representing the recency of consumption. To preserve their ordered structure while enabling regression, we map them to integer scores:

$$\text{CL}k \mapsto k, \quad k \in \{0, \dots, 6\}$$

2.3.2. Encoding Categorical Predictors

The explanatory variables, which we will use as predictors, consist of both demographic attributes and personality traits. The personality traits are continuous values obtained from standardized instruments. However, the demographic attributes are categorical, which are not immediately useble in a regression analysis, that's why we have to encode the values, specifically, using the one-hot encoding with a reference.

For each categorical variable C, where C has L levels:

$$C \in \{c_1, \dots, c_L\}.$$

We fix one level, say c_L as the **reference category**, which we omit from the final vector to avoid collinearity. For the remaining $L-1$ levels, we define dummy variables:

$$d^{(1)} = 1\{C = c_1\}, \ldots, d^{(L-1)} = 1\{C = c_{L-1}\}$$

where

$$1\{A\} = \begin{cases} 1 & A, \\ 0 & \neg A \end{cases}$$

In our case, it means for a dummy variable $d^{(i)}$ corresponding to a possible value c_i of C, the value of $d^{(i)}$ is 1 if for the category C, a respondent has the value c_i, and 0 if otherwise. That means, if a respondent has the value 0 for all $d^{(i)}$ it corresponds to them having the reference value c_L.

We then collect these into a single vector. For a respondent i, the dummy vector is

$$d_i = \begin{bmatrix} d_{i1} & \ldots & d_{ir} \end{bmatrix}^\top \in \mathbb{R}^r$$

where r is the total number of dummy variables. In the dataset, each category C has L_C levels:

- $L_{\text{Age}} = 6$
- $L_{\text{Gender}} = 2$
- $L_{\text{Education}} = 9$
- $L_{\text{Country}} = 8$
- $L_{\text{Ethnicity}} = 7$

Since we removed one value as the reference from each, $r = 27$.

2.3.3. Combined Predictors

For each respondent i, the continuous-valued predictors, viz. the personality traits, can directly form a vector as such:

$$z_i = \begin{bmatrix} z_1 & \ldots & z_7 \end{bmatrix}^\top \in \mathbb{R}^7$$

Each z_i corresponds to one personality trait. Then concatinate the vector z_i with the dummy vector d_i of the encoded categorical predictors, viz. the demographical attributes into one large vector:

$$x_i = \begin{bmatrix} z_i \\ d_i \end{bmatrix} \in \mathbb{R}^p$$

Where $p = 34$. We then combine the predictor vectors of all n respondents into a single matrix:

$$X = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix} \in \mathbb{R}^{n \times p}$$

2.3.4. Multivariate Regression Framework

Let $m = 18$ denote the number of drugs (ignoring the fictional drug, semer), and let

$$y_i = \begin{bmatrix} y_{i1} & \ldots & y_{im} \end{bmatrix}^\top \in \mathbb{R}^m$$

represent participant i's drug use data. We form another matrix combining that of all n respondents:

$$Y = \begin{bmatrix} y_1 & \ldots & y_n \end{bmatrix} \in \mathbb{R}^{n \times m}.$$

We are now able to form the multivariate model which is of the form:

$$Y = XB + E,$$

where $B \in \mathbb{R}^{p \times m}$ contains regression coefficients and $E \in \mathbb{R}^{n \times m}$ is the unknown residual matrix. The j^{th} column of B corresponds to the regression parameters for predicting the use of drug j. Componentwise, for participant i and drug j,

$$y_{ij} = \beta_{0j} + \sum_{k=1}^{p} \beta_{kj} x_{ik} + \varepsilon_{ij}$$

with β_{0j} the intercept specific to substance j.

2.3.5. Parameter Estimation

The regression coefficients are then obtained via least squares by minimizing the total squared deviation accross all participants and substances:

$$\hat{B} = \arg\min_B ||Y - XB||_F^2 = \arg\min_B \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - (XB)_{ij})^2,$$

where $||\cdot||_F$ denotes the Frobenius norm. When $X^\top X$ is invertible, this problem admits the closed-form solution

$$\hat{B} = (X^\top X)^{-1} X^\top Y.$$

2.3.6. Prediction and Evaluation

For a new individual with predictors x_{new}, the model yields a vector of predicted drug use scores:

$$\hat{y}_{\text{new}} = x_{\text{new}}^\top \hat{B}.$$

We can assess the overall performance as well as that for each drug using the root-mean-square error (RMSE):

$$\text{RMSE}_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{ij} - \hat{y}_{ij})^2}, \quad \text{RMSE}_{\text{avg}} = \frac{1}{m} \sum_{j=1}^{m} \text{RMSE}_j$$

3. Result & Discussion

3.1. Data Cleansing

Before conducting any analysis, the dataset was examined for missing or invalid values. Since missing data may bias statistical results, we performed a completeness check by calculating the total number and percentage of `NA` values in each column. The following R code was used:

```r
colSums(is.na(data))
```

Checking for Missing Values

The output below shows that **all columns have 0 missing values**, indicating that the dataset is complete and requires no imputation or removal of incomplete entries.

```
> colSums(is.na(data))
      ID      Age   Gender Education  Country Ethnicity
       0        0        0         0        0         0
  Nscore   Escore   Oscore
       0        0        0
  AScore   Cscore Impulsive       SS  Alcohol   Amphet
       0        0        0         0        0         0
   Amyl   Benzos    Caff
       0        0        0
Cannabis    Choc    Coke   Crack  Ecstasy   Heroin
       0        0        0         0        0         0
Ketamine  Legalh      LSD
       0        0        0
    Meth Mushrooms Nicotine   Semer      VSA
       0        0        0         0        0
```

Output (NA Summary)

Since `Semer` represents a fictional drug, only the class `CL0` is valid. All rows containing any other `Semer` value must be removed. The code below converts drug class labels from "CL0–CL6" to numeric levels, filters out invalid `Semer` entries, and finally removes the `Semer` column.

```r
drug_cols <- c("Alcohol","Amphet","Amyl","Benzos","Caff","Cannabis",
               "Choc","Coke","Crack","Ecstasy","Heroin","Ketamine",
               "Legalh","LSD","Meth","Mushrooms","Nicotine","VSA")

```

```r
all_to_convert <- drug_cols
if ("Semer" %in% names(data)) all_to_convert <-
    c(all_to_convert, "Semer")

df <- data %>%
  mutate(across(all_of(intersect(all_to_convert,
     names(.))),
                ~ as.numeric(gsub("^CL", "", as.
    character(.)))))

if ("Semer" %in% names(df)) {
  df <- df %>%
    filter(!is.na(Semer) & Semer == 0) %>%  #
    keep only CL0
    select(-Semer)                          #
    remove Semer column
}
```

Dataset Cleaning for Semer

3.2. Data Visualisation

To explore behavioural patterns in the dataset, several visualisation techniques were used. Each plot was chosen to suit the variables analysed, enabling meaningful comparisons across demographic groups and examining how personality traits relate to drug usage.

3.2.1. Demographic-Based Drug Usage Patterns

Radar plots were used to compare average drug consumption across demographic groups. Because all substances share the same ordinal scale (CL0–CL6), radar charts effectively highlight multivariate patterns across the nineteen drug types.

Figures 1–5 show usage profiles for Age, Country, Education, Ethnicity, and Gender. Overall consumption levels appear broadly similar across groups, though minor differences emerge for several substances, particularly Cannabis, Coke, and Ecstasy.

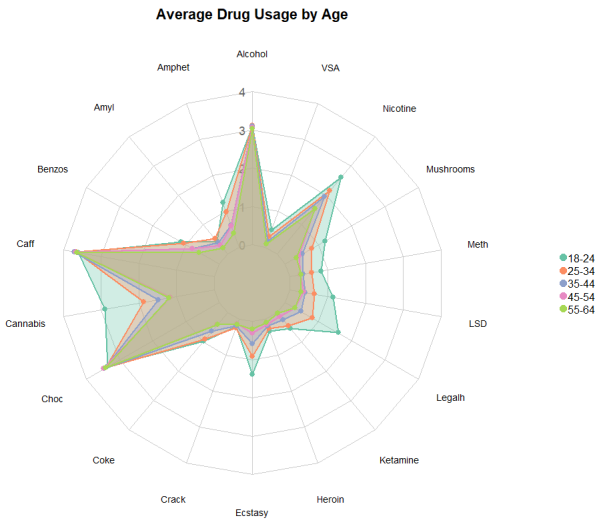

Figure 1. Average Drug Usage by Age

Figure 1 shows that the **18–24** age group exhibits the largest overall radar shape, indicating generally higher usage across most substances. The most notable difference between age groups appears in **Cannabis**, where younger respondents show a clear increase compared to older groups.

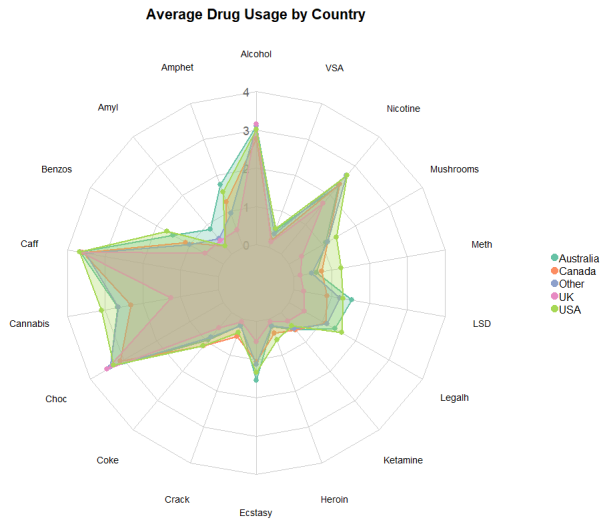

Figure 2. Average Drug Usage by Country

Figure 2 indicates that the **USA** exhibits the largest overall radar shape, suggesting higher average usage across most substances. However, **Australia** surpasses the USA in several specific drugs, notably **LSD**, **Amyl**, and **Amphet**. In contrast, the **UK** shows the smallest radar area, reflecting consistently lower usage levels relative to the other countries.

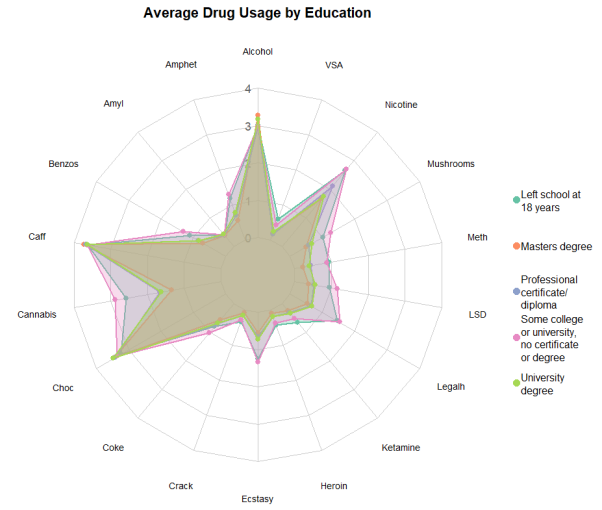

Figure 3. Average Drug Usage by Education

Figure 3 shows that individuals with **some college or university education** display the highest overall drug usage, followed closely by those who **left school at 18 years**. This pattern aligns with the age-based results, where younger groups exhibit higher consumption. In contrast, **Master's degree** holders show the lowest overall usage, suggesting that higher education levels may correlate with reduced drug involvement. However, this trend does not apply to widely used legal substances such as **Caffeine**, **Chocolate**, and **Alcohol**, where differences remain minimal across education groups.

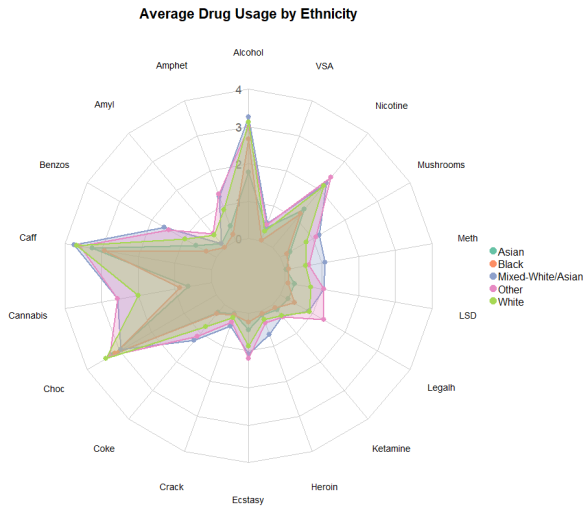

Figure 4. Average Drug Usage by Ethnicity

Figure 4 shows that average drug usage is broadly similar across ethnic groups, with only small variations in most substances. The **Mixed-White/Black** and **White** groups display slightly larger overall radar shapes, indicating marginally higher usage, while the **Asian** group tends to show the lowest levels. The most noticeable differences appear in substances such as **Cannabis**, **Benzos**, and **Ecstasy**, where certain groups exhibit modestly elevated consumption compared to others.

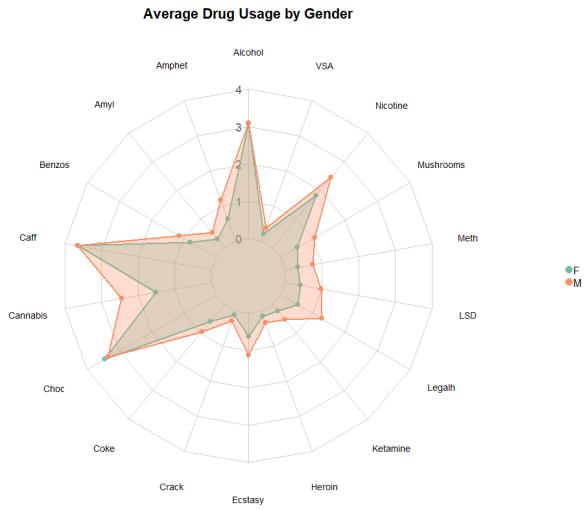

Figure 5. Average Drug Usage by Gender

Figure 5 shows that **males** exhibit slightly higher overall drug usage than **females**, resulting in a marginally larger radar shape. The biggest differences appear in **Cannabis**, **Nicotine**, and **Ecstasy**, where male usage is noticeably higher. For common legal substances such as **Caffeine**, **Chocolate**, and **Alcohol**, however, both genders show nearly identical consumption patterns.

Across all demographic, the radar charts reveal a consistent pattern in overall drug consumption. The substances with the highest average usage levels are:

- **Caffeine (Caff)**
- **Alcohol**
- **Chocolate (Choc)**
- **Nicotine**

These drugs appear prominently across all plots because they are *legal, widely accessible, and socially normalized*. Their consumption is therefore more frequent, leading to higher recency-of-use scores. Additionally, the **Legalh** (legal highs) attribute shows moderate usage, suggesting that certain synthetic or emerging recreational substances are somewhat familiar or accessible within the population.

In contrast, substances such as **Heroin**, **Crack**, **Methadone (Meth)**, **Ketamine**, and **LSD** consistently exhibit low average usage across all demographic categories. This aligns with their restricted legality, higher perceived risks, and limited availability.

These radar charts indicate that legal or socially accepted substances dominate usage patterns, while illegal, stigmatized, or high-risk drugs remain infrequently used, regardless of demographic segmentation.

3.2.2. Personality–Drug Associations

To assess the relationship between personality traits and drug usage, a Spearman correlation heatmap was constructed (Figure 6). Spearman's rank correlation is appropriate for this analysis because drug-use levels are ordinal, and personality measures are continuous. This approach captures monotonic relationships without assuming linearity.

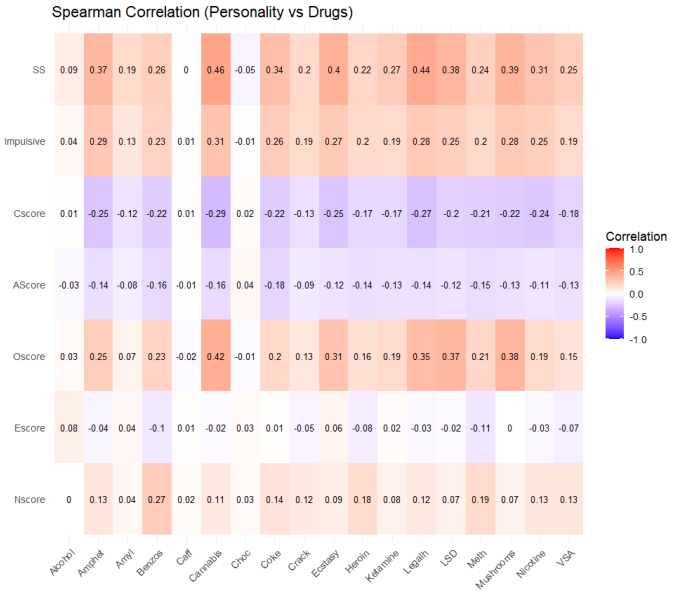

Figure 6. Spearman Correlation Between Personality Traits and Drug Usage

Figure 6 shows clear links between personality traits and drug usage. **SS** (Sensation Seeking), **Oscore** (Openness), and **Impulsiveness** have the strongest positive correlations, indicating that individuals with high impulsivity, novelty-seeking, or sensation motivation are more likely to use substances such as **Cannabis**, **Mushrooms**, **Ecstasy**, and **Amphetamines**.

In contrast, **Cscore** (Conscientiousness) and **Ascore** (Agreeableness) consistently show negative associations, suggesting that more organised and cooperative individuals tend to avoid drugs. **Escore** (Extraversion) and **Nscore** (Neuroticism) show only weak effects, except for a moderate positive link between Neuroticism and **Benzodiazepine** use.

Legal substances such as **Alcohol**, **Caffeine**, and **Chocolate** display minimal correlations, reflecting their normalisation. Overall, drug use is most strongly associated with high sensation seeking and openness.

3.2.3. Drug Bar Plot Distribution

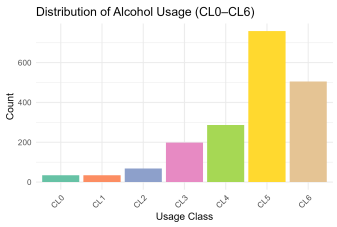

Figure 7. Distribution of Alcohol Usage

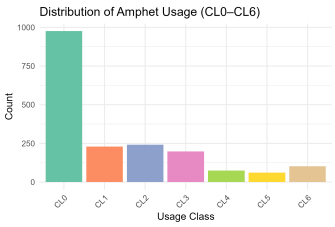

Figure 8. Distribution of Amphet Usage

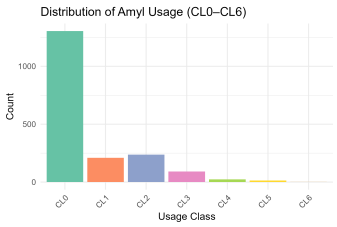

Figure 9. Distribution of Amyl Usage

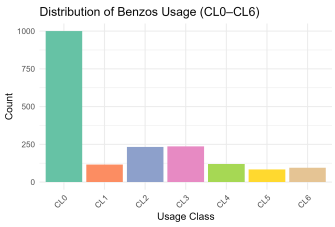

Figure 10. Distribution of Benzos Usage

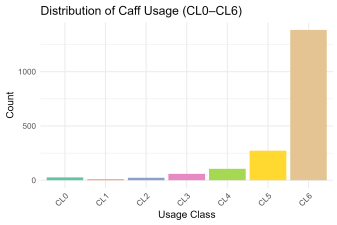

Figure 11. Distribution of Caffeine Usage

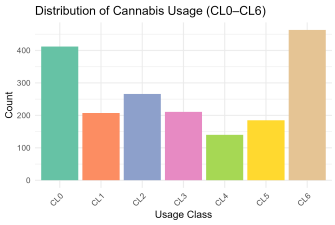

Figure 12. Distribution of Cannabis Usage

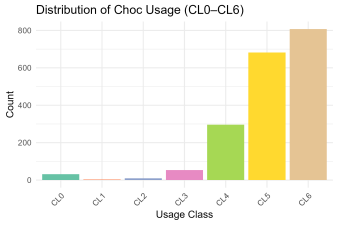

Figure 13. Distribution of Chocolate Usage

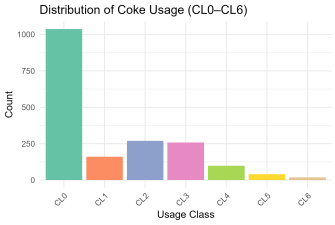

Figure 14. Distribution of Cocaine Usage

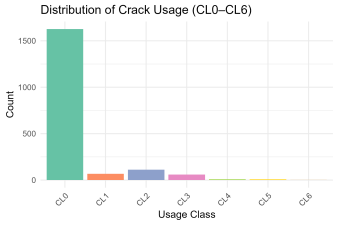

Figure 15. Distribution of Crack Usage

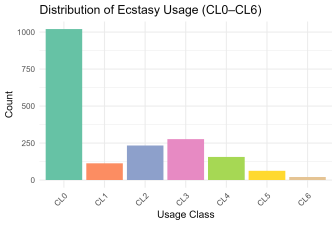

Figure 16. Distribution of Ecstasy Usage

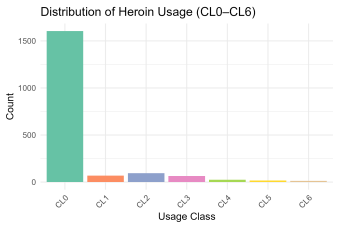

Figure 17. Distribution of Heroin Usage

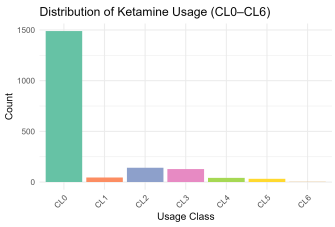

Figure 18. Distribution of Ketamine Usage

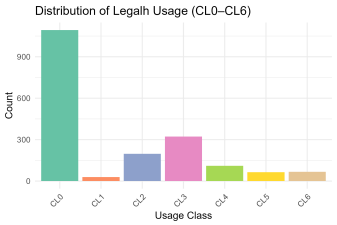

Figure 19. Distribution of Legal High Usage

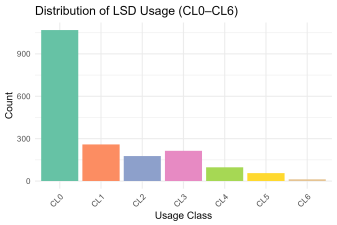

Figure 20. Distribution of LSD Usage

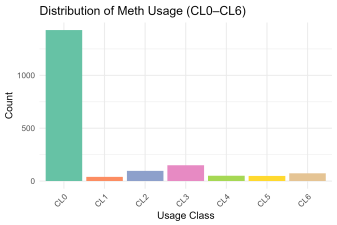

Figure 21. Distribution of Meth Usage

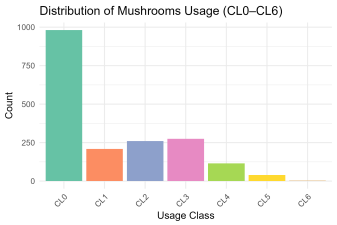

Figure 22. Distribution of Mushrooms Usage

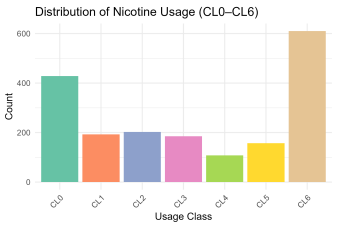

Figure 23. Distribution of Nicotine Usage

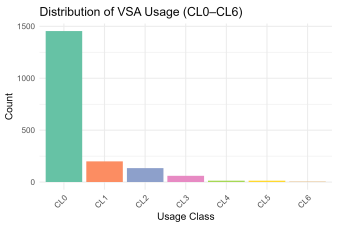

Figure 24. Distribution of VSA Usage

Across most substances, the usage distributions exhibit a strong right skew, with **CL0 (never used)** consistently dominating. This pattern reflects the illegality, rarity, and social stigma associated with many drugs, resulting in the majority of individuals reporting no lifetime usage.

- **Chocolate (Choc).** Chocolate displays the opposite trend: a pronounced left skew, with **CL6 (used yesterday)** being the highest class. This indicates that chocolate is consumed routinely and is deeply normalised in daily life.
- **Caffeine (Caff).** Caffeine shows an even stronger left-skewed distribution, overwhelmingly concentrated in **CL6**. This aligns with caffeine's status as a near-daily stimulant for most individuals.
- **Alcohol.** Alcohol usage peaks at **CL5 (used last week)**, with CL6 also relatively high. This reflects alcohol's legality and cultural integration, where consumption is frequent but often episodic (e.g., weekends or social settings).
- **Nicotine.** Nicotine exhibits a **bimodal distribution**, with large peaks at CL0 and CL6. Many individuals abstain entirely, while those who do use nicotine tend to do so daily, consistent with its addictive and legal nature.
- **Cannabis.** Cannabis shares a similar two-peak pattern (high CL0 and high CL6). This likely reflects differences in regional legality and accessibility: some participants abstain due to restrictions, while others consume frequently in areas where it is legal or socially accepted.
- **Other illicit drugs (e.g., Coke, Ecstasy, Mushrooms).** These substances follow the typical right-skewed pattern, with CL0 dominating and usage sharply decreasing in higher classes. This is consistent with their illegality, limited availability, and higher perceived risk.

3.3. Predictive Modeling

Using the analysis above, we can create a predictive model using multivariate regression to predict drug usage based on a person's personality score and demographic information. We may start by applying preprocessing methods within the code.

```
library(readr)
library(dplyr)

url_quantified <- "https://raw.githubusercontent.com/RobbenWijanathan/drug-consumption-regression/main/drug_consumption_quantified.csv"
df <- read_csv(url_quantified)

drug_cols <- c("Alcohol","Amphet","Amyl","Benzos","Caff","Cannabis",
               "Choc","Coke","Crack","Ecstasy","Heroin","Ketamine",
               "Legalh","LSD","Meth","Mushrooms","Nicotine","VSA")

df[drug_cols] <- lapply(df[drug_cols], function(x) {
  as.numeric(gsub("CL", "", x))
})

df <- df[df$Semer == "CL0", ]

df$Age <- as.factor(df$Age)
df$Gender <- as.factor(df$Gender)
df$Education <- as.factor(df$Education)
df$Country <- as.factor(df$Country)
df$Ethnicity <- as.factor(df$Ethnicity)

df_cleaned <- na.omit(df)
df_cleaned <- df_cleaned %>% select(-Semer, -ID)

target_drugs <- c("Alcohol","Amphet","Amyl","Benzos","Caff","Cannabis","Choc","Coke","Crack","Ecstasy","Heroin","Ketamine",
                  "Legalh","LSD","Meth","Mushrooms","Nicotine","VSA")

predictors <- df_cleaned %>%
  select(-all_of(target_drugs))

X<- model.matrix(~ ., predictors)
X <- as.data.frame(X[, -1])
df_numeric <- cbind(X, df_cleaned[, target_drugs])

predictor_cols <- setdiff(names(df_numeric), target_drugs)
predictor_cols <- predictor_cols[complete.cases(predictor_cols)]
predictor_cols
corr <- cor(df_numeric[, target_drugs], df_numeric[, predictor_cols])
best_predictors_clean <- gsub("[0-9.-]+$", "", predictor_cols)  # Remove trailing numbers
best_predictors_clean <- unique(best_predictors_clean)
best_predictors_clean

formula_multi <- as.formula(
  paste("cbind(", paste(target_drugs, collapse=","), ") ~ ",
        paste(best_predictors_clean, collapse = " + "))
)
formula_multi

set.seed(123)

n <- nrow(df_numeric)
train_index <- sample(1:n, size = 0.7*n)

train_data <- df_cleaned[train_index, ]
test_data  <- df_cleaned[-train_index, ]
model_multi <- lm(formula_multi, data = train_data)

summary(model_multi)

predictions <- predict(model_multi, newdata = test_data)
head(predictions)

actual <- test_data[, drug_cols]
rmse <- sqrt(colMeans((predictions - actual)^2))
rmse
```

Predictive Modelling

3.4. Predictive Modelling Results and Interpretation

In this section we present and interpret the results of the multivariate regression model introduced in Section 2.3.4, with particular attention to the research objectives formulated in Chapter 1. Recall that the project aims (i) to build and evaluate predictive models for drug consumption using statistical and machine-learning methods, and (ii) to interpret model results and communicate insights in a clear, data-driven manner. Here we address these objectives for the multivariate linear regression model that links demographic and personality features to the recency of use for eighteen different substances.

3.4.1. Research Objectives Revisited

For convenience, we restate the two objectives most directly related to the predictive modelling component:

- **Objective 3:** To build and evaluate predictive models for drug consumption using statistical and machine-learning methods.
- **Objective 4:** To interpret model results and communicate insights through clear data.

The analysis in the remainder of this section directly targets these two objectives. First, we describe how the multivariate regression model was trained and evaluated. Second, we summarise its performance in terms of explained variance and prediction error for each drug. Finally, we interpret the patterns in these results in the context of demographic and personality factors.

3.4.2. Model Specification and Data Partitioning

Let n denote the number of respondents after preprocessing. As described in Chapter 2, the original dataset contains 1,885 instances and 32 attributes. After dropping rows with missing values, we obtain $n = 1{,}876$ complete cases. For each participant $i \in \{1, \dots, n\}$, we collect $p = 34$ predictors in a feature vector

$$x_i \in \mathbb{R}^p,$$

which includes demographic variables (age group, gender, education level, country and ethnicity), personality trait scores (Nscore, Escore, Oscore, AScore, Cscore), and additional behavioural traits (Impulsive, SS). These predictors are assembled into a design matrix

$$X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in \mathbb{R}^{n \times p}.$$

The outcome variables correspond to recency of use for $m = 18$ non-fictional drugs (Alcohol, Amphet, Amyl, Benzos, Caff, Cannabis, Choc, Coke, Crack, Ecstasy, Heroin, Ketamine, Legalh, LSD, Meth, Mushrooms, Nicotine, and VSA). For each participant i, we denote by

$$y_i = (y_{i1}, \dots, y_{im})^\top \in \mathbb{R}^m$$

the vector of recency codes (CL0–CL6, encoded as integers $0, \dots, 6$).

Stacking these into a response matrix

$$Y = \begin{bmatrix} y_1^\top \\ \vdots \\ y_n^\top \end{bmatrix} \in \mathbb{R}^{n \times m},$$

the multivariate regression model is

$$Y = XB + E,$$

where $B \in \mathbb{R}^{p \times m}$ collects the regression coefficients for all drugs and $E \in \mathbb{R}^{n \times m}$ is the residual matrix. Column j of B corresponds to the parameters for predicting drug j.

For model training and evaluation, the cleaned dataset is randomly split into a 70% training set and a 30% test set. Concretely,

$$n_{\text{train}} = \lfloor 0.7 \times 1876 \rfloor = 1313, \qquad n_{\text{test}} = 1876 - 1313 = 563.$$

The multivariate regression is estimated on the training subset, and then used to generate out-of-sample predictions on the test subset.

Although the model is estimated jointly in matrix form, it is equivalent to fitting $m = 18$ separate linear regression models

$$Y_{\cdot j} = X\beta_j + \varepsilon_j, \qquad j = 1, \dots, m,$$

where $Y_{\cdot j}$ denotes the jth column of Y, $\beta_j \in \mathbb{R}^p$ is the coefficient vector for drug j, and ε_j is the residual vector. The software output for each response confirms that all eighteen regressions are jointly highly significant: for example, for Alcohol we obtain a residual standard error of 1.283 on 1,279 degrees of freedom, a multiple R^2 of 0.1135, and an F-statistic of 4.961 on 33 and 1,279 degrees of freedom with $p < 2.2 \times 10^{-16}$. Similar omnibus F-tests for the other drugs also have p-values below 2.2×10^{-16}, indicating that the predictors have statistically detectable associations with recency of use across all substances.

3.4.3. Explained Variance and Predictive Accuracy

To evaluate how well the model explains and predicts drug-use recency for each substance, we consider two complementary metrics:

- The *training-set* coefficient of determination R^2 for each univariate regression, which measures the proportion of variance in the recency code explained by the predictors.
- The *test-set* root mean squared error (RMSE) for each drug, defined as

$$\text{RMSE}_j = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i \in \text{test}} \left(\hat{y}_{ij} - y_{ij}\right)^2},$$

  where $\hat{y}_{ij}$ is the model prediction for participant i and drug j. Because the outcome is encoded on the CL0–CL6 scale, the RMSE has a straightforward interpretation as the typical prediction error measured in units of "recency categories".

Table 1 reports the training-set R^2 and test-set RMSE for each of the eighteen substances. The R^2 values are shown with four decimal places, while RMSE values are rounded to two decimal places.

Several points are worth highlighting:

- The R^2 values exhibit substantial heterogeneity across substances, ranging from approximately 0.06 (Caff, Choc) to almost 0.50 (Cannabis). Substances with comparatively higher R^2 include Cannabis ($R^2 \approx 0.50$), LSD ($R^2 \approx 0.38$), Legal highs ($R^2 \approx 0.38$), Mushrooms ($R^2 \approx 0.37$), and Amphetamines ($R^2 \approx 0.29$). For these drugs, the model captures a non-trivial portion of the variation in recency of use.
- For more commonly consumed substances like Alcohol and Nicotine, the R^2 values are more modest (≈ 0.11 and ≈ 0.21, respectively), indicating that a substantial share of individual-level variation in recency of use remains unexplained by the available predictors.

Table 1. Per-drug training-set R^2 and test-set root mean squared error (RMSE) for the multivariate regression model. The outcome for each drug is encoded on the CL0–CL6 scale, so the RMSE can be interpreted as the typical error measured in "recency classes".

Drug	R^2 (train)	RMSE (test)
Alcohol	0.1135	1.27
Amphet	0.2854	1.61
Amyl	0.1657	1.03
Benzos	0.2712	1.69
Caff	0.0594	1.17
Cannabis	0.4965	1.61
Choc	0.0572	1.03
Coke	0.2306	1.40
Crack	0.1350	0.74
Ecstasy	0.3004	1.38
Heroin	0.1731	0.94
Ketamine	0.1519	1.23
Legalh	0.3837	1.43
LSD	0.3838	1.21
Meth	0.2442	1.53
Mushrooms	0.3686	1.19
Nicotine	0.2104	2.21
VSA	0.1496	0.91

- Across all eighteen drugs, the test-set RMSE ranges from about 0.74 recency classes (Crack) to about 2.21 recency classes (Nicotine), with an average RMSE of roughly 1.31. In practical terms, this means that the model is typically off by about one category on the CL0–CL6 scale for most substances (e.g. predicting "used in the last year" when the truth is "used in the last month").
- Prediction errors tend to be larger for some widely used substances: for example, the RMSE is 1.27 for Alcohol, 1.61 for Amphetamines, 1.61 for Cannabis, and as high as 2.21 for Nicotine. This suggests that, even when the model detects significant relationships between predictors and outcomes, individual-level recency of use for these substances is difficult to forecast precisely.

From the perspective of **Objective 3** (building and evaluating predictive models), these results show that the multivariate regression framework is able to learn statistically significant patterns and achieve non-trivial predictive accuracy, particularly for certain illicit drugs such as Cannabis, LSD, Legal highs and Mushrooms. However, the relatively low to moderate R^2 values for many substances, combined with RMSEs around one recency class, indicate that the model has clear limits in its ability to predict individual drug-use behaviour.

3.4.4. Interpretation in Terms of Demographic and Personality Factors

Beyond aggregate measures of fit, the regression coefficients provide insight into how demographic and personality factors relate to substance use, addressing **Objective 4** (interpreting model results and communicating insights).

At a high level, the estimated coefficients for the demographic variables (age, country and ethnicity) display systematic patterns:

- For several illicit drugs (e.g. Amphet, Coke, Ecstasy, Mushrooms), older age groups tend to have lower recency scores compared to younger participants, holding other variables fixed. This is consistent with a life-course pattern in which experimentation and intensive use are more concentrated in younger adulthood.
- Country and ethnicity indicators sometimes show significant effects, reflecting cross-national and cultural differences in both the prevalence and social acceptability of particular substances. For example, certain illicit drugs show lower predicted use in some countries relative to others, even after controlling for personality traits.

Turning to personality traits, the Big Five scores (Nscore, Escore, Oscore, AScore, Cscore) generally exhibit weaker and less consistent associations with recency of use:

- Across most substances, Big Five coefficients are small and only occasionally significant, indicating that broad traits such as Extraversion or Conscientiousness are weak standalone predictors of drug-use recency.
- Some exceptions appear: higher Conscientiousness (Cscore) is sometimes linked to lower recency of use, while Neuroticism (Nscore) and Extraversion (Escore) show sporadic positive or negative effects, though these patterns are less consistent than those of the behavioural traits below.

In contrast, the behavioural traits Impulsivity and Sensation Seeking (SS) exhibit clearer and more robust associations:

- Sensation Seeking (SS) shows positive and often significant coefficients for many substances (e.g., Alcohol, Amphetamines, Amyl, Benzos), suggesting that high-SS individuals are more likely to report recent or frequent use.
- Impulsivity also tends to have positive coefficients for multiple drugs, with particularly strong effects for some (e.g., Amphetamines), consistent with its link to unplanned or compulsive substance use.

Broad personality traits offer some explanatory power, but more specific behavioural tendencies, such as sensation seeking and impulsivity, are more directly associated with recent substance use. Demographic factors (age, country, ethnicity) also shape these relationships, contributing to the heterogeneous predictive performance observed across drugs.

3.4.5. Low R^2 Values and Limits of Predictability

Although the multivariate model is statistically significant for all eighteen substances, the R^2 values in Table 1 remain modest. This indicates that a large share of individual variation in drug-use recency is not captured by our predictors, and this result requires careful interpretation.

First, the modest R^2 values may partly reflect limitations of the modelling approach itself. The analysis relies on a linear specification with a fixed set of demographic and personality covariates. Non-linear relationships, interactions between traits, or additional contextual variables (such as peer networks, life events, mental health conditions, or economic factors) are not included. Incorporating such information, or using more flexible modelling techniques (e.g. tree-based methods or neural networks), could potentially improve predictive performance.

However, a more fundamental perspective from recent computational social science also applies here. Several authors argue that many complex social behaviours are inherently difficult to predict at the individual level, even with rich data and carefully tuned models. Outcomes such as educational achievement, mental health, criminal justice involvement, or economic hardship often show substantial randomness, path dependence, and sensitivity to unobserved factors. In these settings, there may be a hard upper bound on achievable R^2, beyond which additional model complexity yields only marginal improvement (Hofman, Sharma & Watts, 2017).

Empirical evidence for this view comes from the Fragile Families and Child Wellbeing Study. In that project, more than four thousand cases and nearly thirteen thousand predictors were provided to over a hundred modelling teams using modern machine-learning methods. The task was to predict six adolescent outcomes (including GPA, grit, eviction, and material hardship). Despite the richness of the data and the sophistication of the models, the best out-of-sample R^2 values were about 0.20 for GPA and material hardship, and only around 0.05 for the remaining outcomes (Salganik et al., 2020). This shows that even optimally tuned non-linear models with extensive feature sets can explain only a modest share of variance in many social-behavioural outcomes.

Taken together, these points suggest that the low to moderate R^2 values in our drug-consumption models should not be viewed solely as a failure of the specific methodology used here. Instead, they likely reflect both:

1. The limits of a relatively simple linear model with a restricted set of predictors (demographics and personality scores), and
2. Intrinsic limits on how precisely individual drug-use behaviour can be predicted from such variables alone, given the complex and context-dependent nature of substance use.

From the standpoint of **Objective 3**, these findings emphasise that predictive models in social and behavioural domains should be assessed not only by raw R^2 values but also by realistic expectations about the difficulty of the prediction task. From the standpoint of **Objective 4**, the results highlight the importance of using predictive models primarily to understand broad patterns and risk factors, such as the roles of age, country, sensation seeking, and impulsivity, rather than as precise tools for forecasting individual behaviour.

4. Conclusion

In summary, the multivariate regression model presented in this report:

- Identifies statistically significant associations between demographic factors, personality traits, and the recency of use for eighteen different substances.
- Exhibits moderate explanatory power for several drugs particularly Cannabis, LSD, Legal highs, and Mushrooms, while providing weaker predictive performance for others, with typical errors of roughly one recency class.
- Demonstrates that behavioural traits such as Sensation Seeking and Impulsivity, together with demographic context, contribute more meaningfully to prediction accuracy than the broader Big Five personality dimensions.
- Highlights both methodological and fundamental limitations on the predictability of individual drug consumption behaviour using only demographic and personality-based features.

Overall, these findings directly address the predictive modelling and interpretive objectives outlined in Chapter 1. They offer a clear and quantitatively grounded evaluation of the extent to which drug consumption patterns can, and cannot, be inferred from the available demographic and psychological variables.

References

[1] Han, J., Pei, J., & Tong, H. (2022). *Data mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.

[2] Pearson, R. K. (2018). *Exploratory data analysis using R*. CRC Press.

[3] Healy, K. (2019). *Data visualization: A practical introduction*. Princeton University Press.

[4] Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science (New York, N.Y.), 355*(6324), 486–488. https://doi.org/10.1126/science.aal3856

[5] Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., et al. (2020). *Measuring the predictability of life outcomes with a scientific mass collaboration*. Proceedings of the National Academy of Sciences, *117*(15), 8398–8403. https://doi.org/10.1073/pnas.1915006117