

Drug Consumption

Nhat Bui, Johan Ferreira, Thilo Holstein

2025-03-06

Contents

1	Introduction	2
2	Cleaning and formatting the dataset	2
2.1	Fomattting the dataset	3
2.2	Investigating missing values	3
2.3	Investigating outliers	3
3	Exploratory Data Analysis	4
3.1	Correlation Between Behavioral Measures	4
3.2	Comparing Behavioral Measure For Gender	5
3.3	Comparing Education Level with Behavioral Measures	6
3.4	Analysis of Seremon Usage	7
4	Source	7

1 Introduction

Drug use is a significant risk behavior with serious health consequences for individuals and society. Multiple factors contribute to initial drug use, including psychological, social, individual, environmental, and economic elements, as well as personality traits. While legal substances like sugar, alcohol, and tobacco cause more premature deaths, illegal recreational drugs still create substantial social and personal problems.

In this data science project, we aim to identify factors and patterns potentially explaining drug use behaviors through machine learning techniques. By analyzing demographic, psychological, and social variables in our dataset, we'll aim to uncover potential predictors, use machine learning methods to understand the complex relationships surrounding drug consumption, demonstrating how machine learning can reveal insights into behavioral patterns. While our findings won't directly inform interventions, this project showcases how data-driven approaches can enhance our understanding of complex social phenomena and provide valuable practice in applying machine learning to real-world datasets.

The database contains records for 1,885 respondents with 12 attributes including personality measurements (NEO-FFI-R, BIS-11, ImpSS), demographics (education, age, gender, country, ethnicity), and self-reported usage of 18 drugs plus one fictitious drug (Semeron). Drug use is classified into seven categories ranging from "Never Used" to "Used in Last Day." All input attributes are quantified as real values, creating 18 distinct classification problems corresponding to each drug. A detailed description of the variables can be found in the Column Description text file.

2 Cleaning and formatting the dataset

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##   group_rows

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##   combine

##
## Attaching package: 'tidyr'

## The following object is masked from 'package:reshape2':
##
##   smiths
```

2.1 Fomattting the dataset

The original data set had all the values for most of the variables set to a random floating number representing a specific categorical value, we believe this was done in order to remove bias from the dataset. As the requirements of this project is different form the data set's original intention we had to replace these values with the original values in order to complete all the required steps for our project.

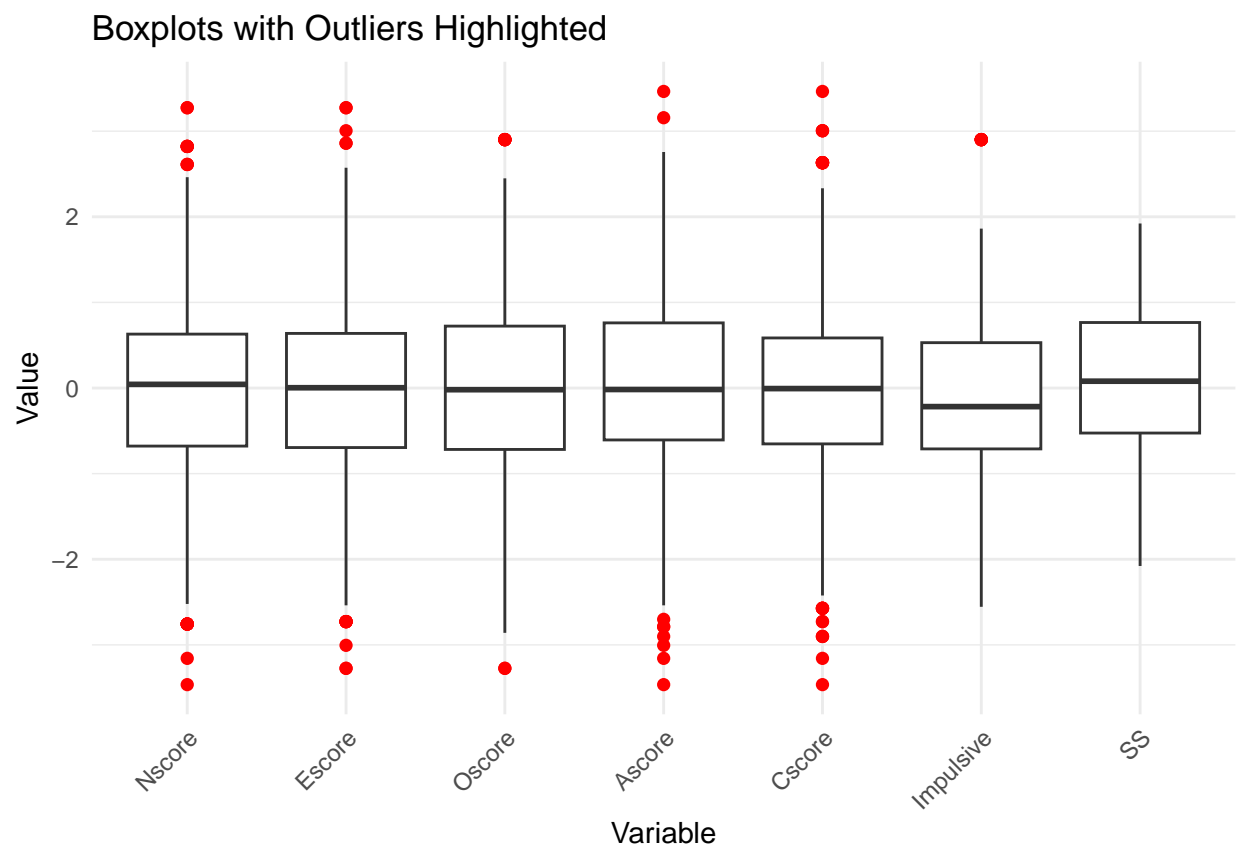
2.2 Investigating missing values

```
## NA values by column:
```

```
## Education Ethnicity
##          99          83
```

Only two columns contain missing values, affecting approximately 5% of the 1885 observations. Given the nature of these variables and the completeness of the rest of the data, we assume participants deliberately withheld this information. Therefore, we replaced the missing values with “Not Provided”, allowing us to treat these instances as a distinct category.

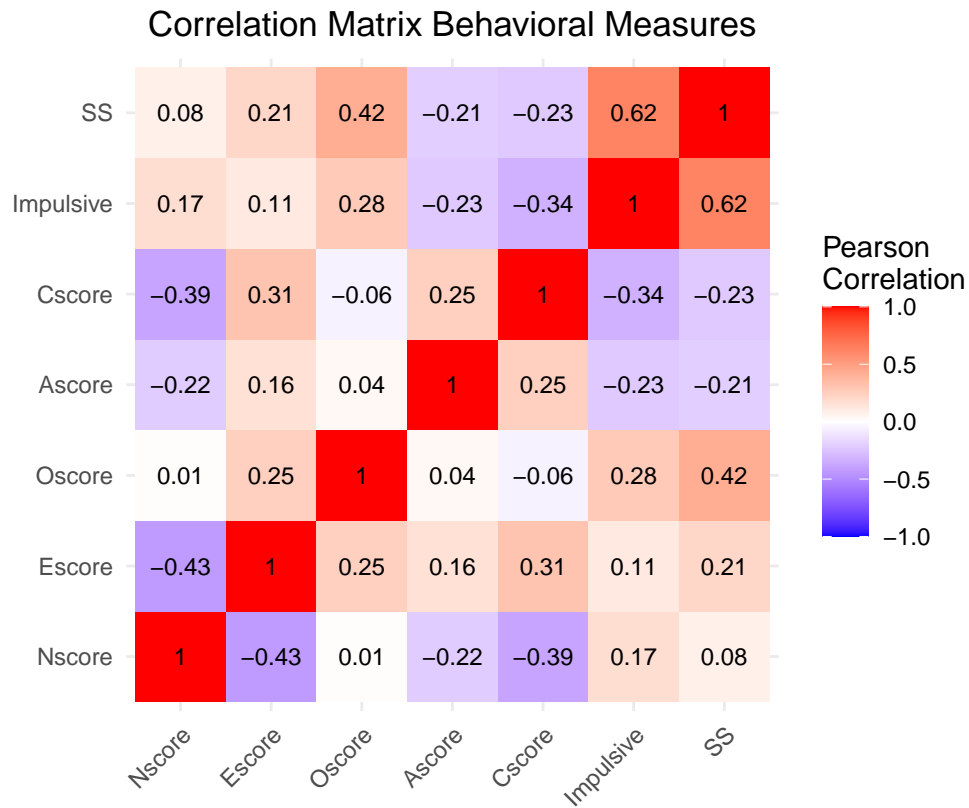
2.3 Investigating outliers



As can be seen from the box plots our data set has some values that are outside of the upper and lower bounds. All though these values are technically outliers they are not extreme, still fall inside of the range of our expected values and conforms to a normal distribution.

3 Exploratory Data Analysis

3.1 Correlation Between Behavioral Measures



The correlation matrix shows that certain personality traits tend to cluster together. For example, SS (Sensation Seeking) has a positive correlation with Escore (Extraversion), Oscore (Openness) and Impulsive while they in turn also have positive correlations with each other and a negative correlation to Cscore (Conscientiousness) and Ascore (Agreeableness) while they have a positive correlation with each other.

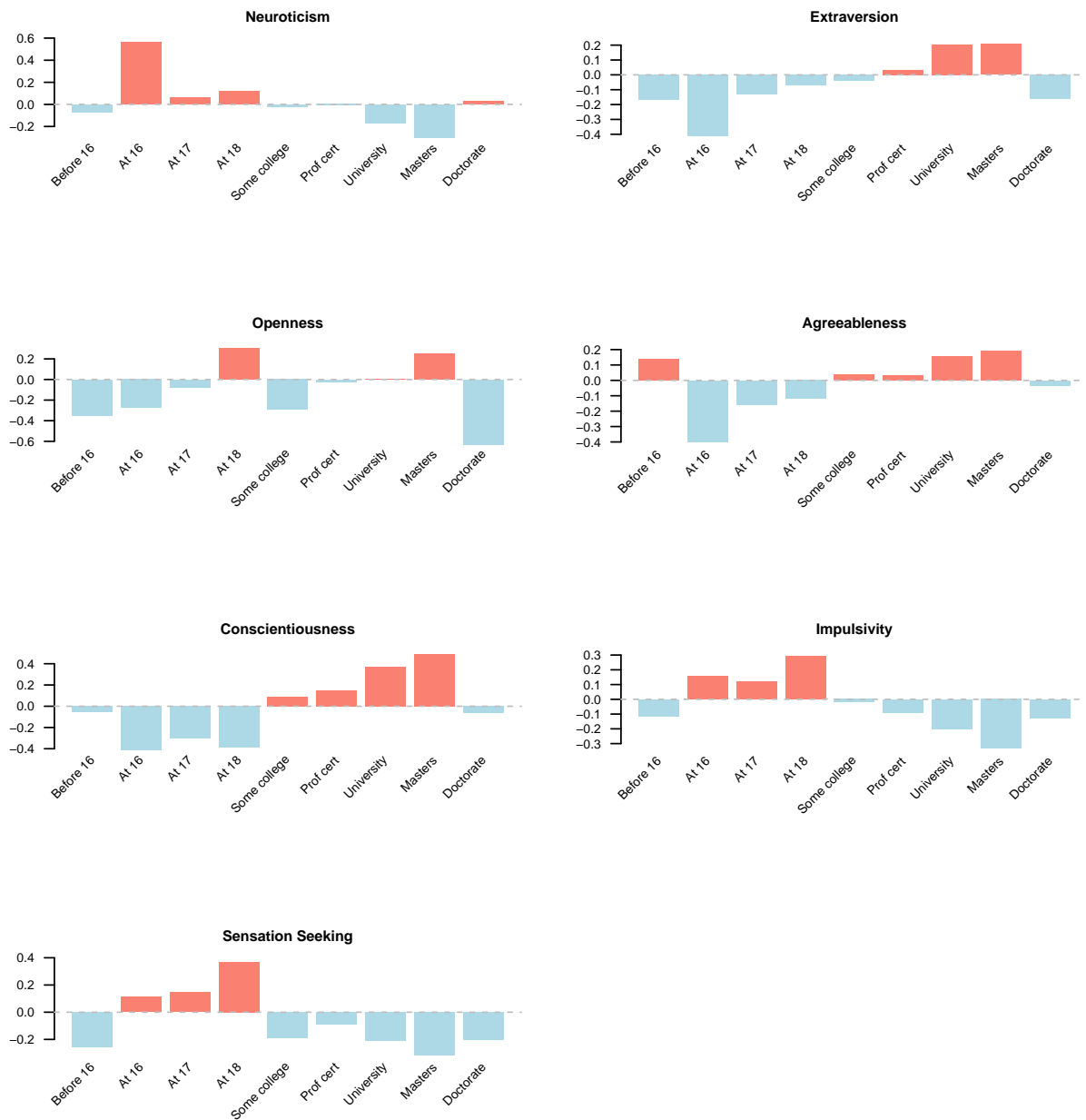
3.2 Comparing Behavioral Measure For Gender



The mean of all the Behavioral Measures is 0, the chart show the mean score broken down by gender for each Behavioral Measures. That chart shows that males tend to be more sensation seeking and impulsive but also more open, where females tend to be more impulsive but also more agreeable and conscientious.

3.3 Comparing Education Level with Behavioral Measures

Personality Traits by Education Level



It is not very clear at first glance but when you study the table closely it becomes clear that traits that can be perceived as bad like Neuroticism, Impulsivity and Sensation Seeking is more prevalent with lower education levels including Not Provided and steadily decrease as the level of education increases.

3.4 Analysis of Seremon Usage

```
# Create a nicely formatted table for the detailed counts
semeron_table <- semeron_counts %>%
  mutate(Percentage = paste0(round(Count / sum(Count) * 100, 2), "%")) %>%
  rename(`Usage Category` = Semer) %>%
  kable(caption = "Semeron Usage Categories",
        booktabs = TRUE,
        col.names = c("Usage Category", "Count", "Percentage"))

# For PDF output, apply specific styling
if(knitr::is_latex_output()) {
  semeron_table <- semeron_table %>%
    kable_styling(latex_options = c("striped", "hold_position"),
                  full_width = FALSE,
                  position = "center") %>%
    row_spec(0, bold = TRUE) %>%
    column_spec(1, width = "5cm") %>%
    column_spec(2, width = "2cm") %>%
    column_spec(3, width = "2.5cm")
}

# Display the table
semeron_table
```

Table 1: Semeron Usage Categories

Usage Category	Count	Percentage
Never Used	1877	99.58%
Used in Last Decade	3	0.16%
Used in Last Year	2	0.11%
Used over a Decade Ago	2	0.11%
Used in Last Month	1	0.05%

4 Source

<https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified>