

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	2
2.2	Investigating Missing Values	2
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	6
4	Prepraring the Dataset for Machine Learning	7
5	Machine Learning Models	7
5.1	Linear Model	7
6	Source	12

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

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.

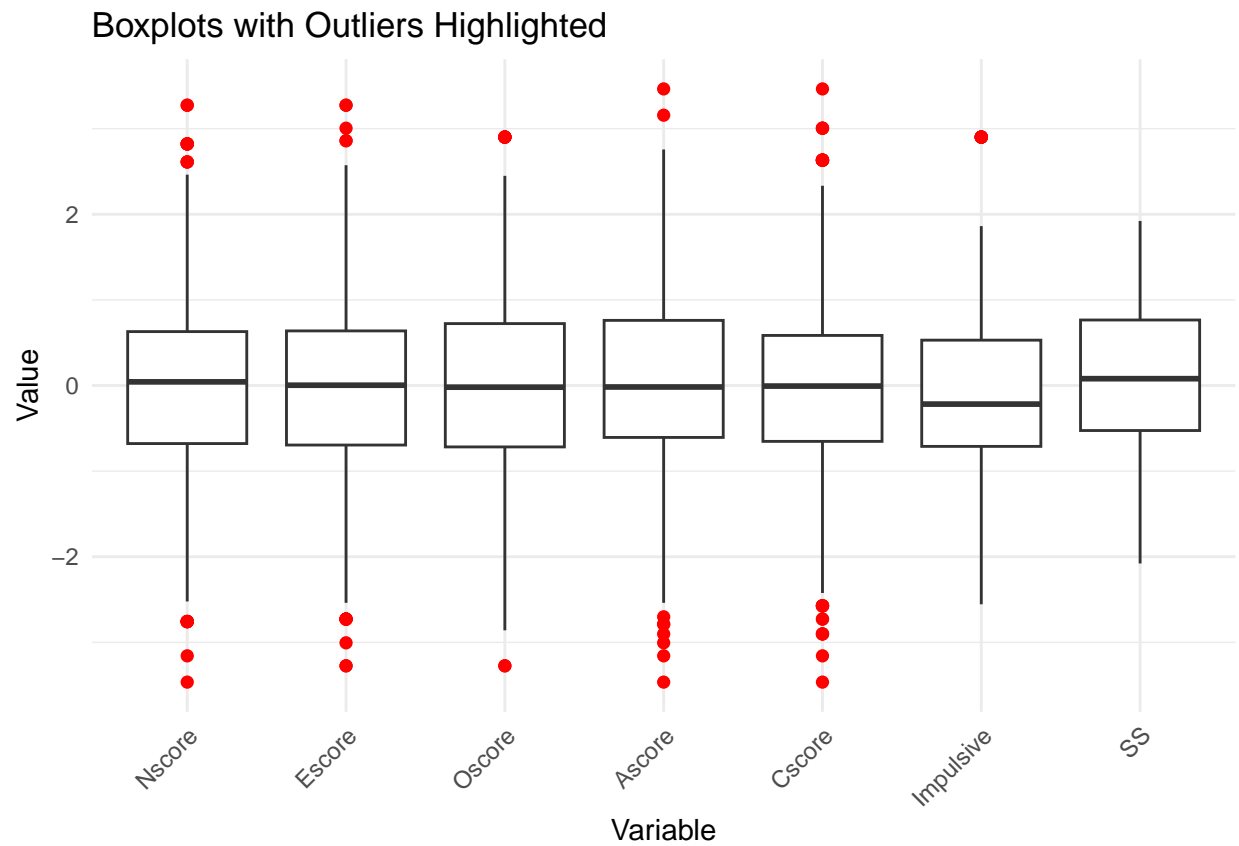
2.2 Investigating Missing Values

```
## 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.

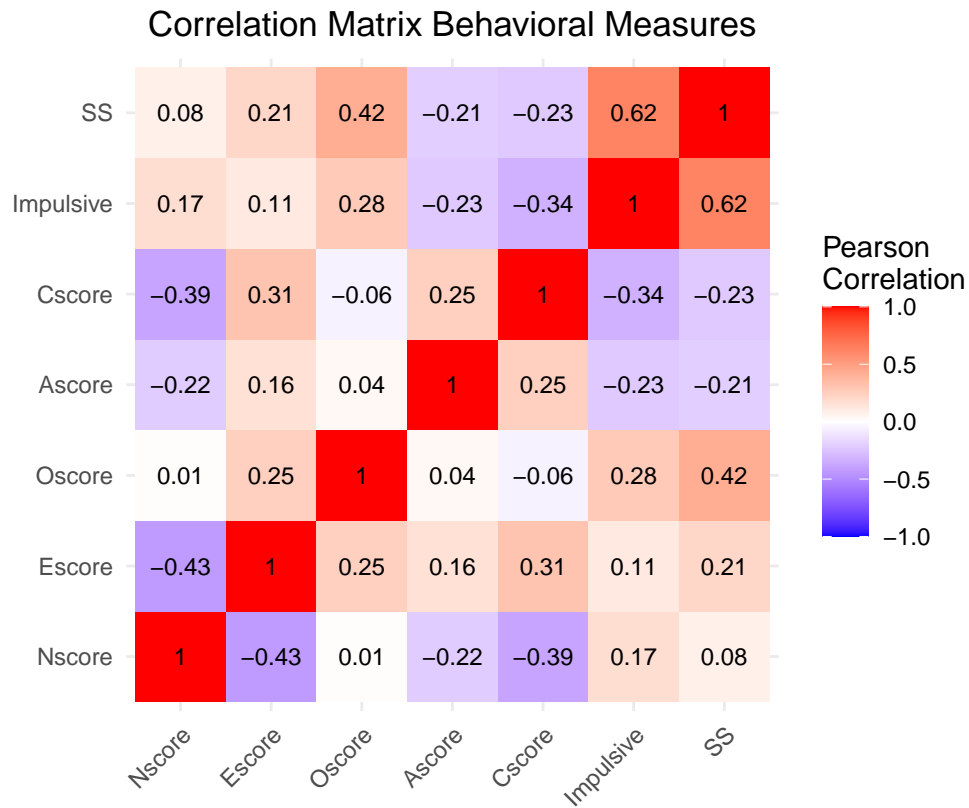
2.3 Investigating Outliers



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.

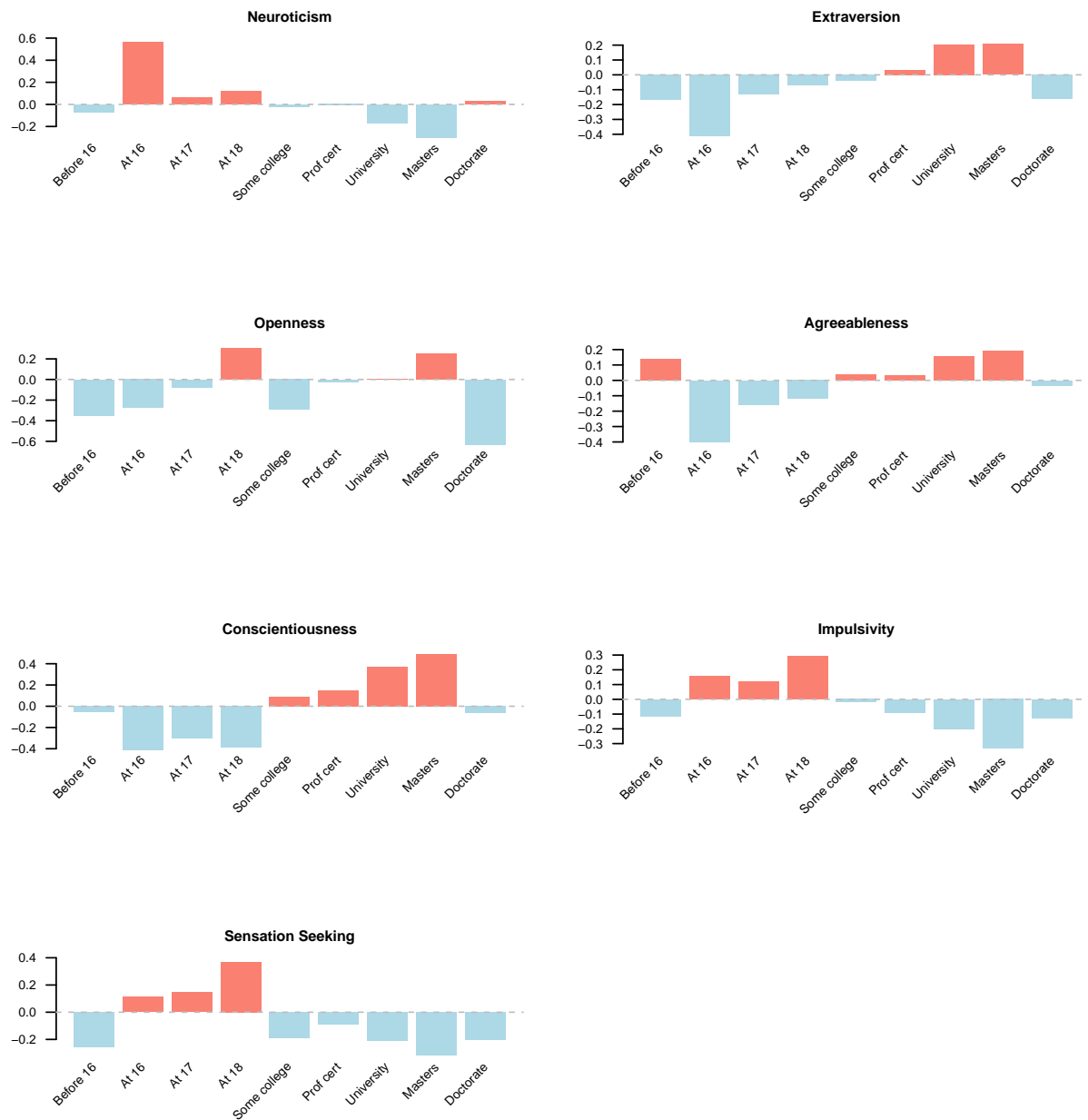
3.2 Comparing Behavioral Measure for Gender



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

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%

Semeron is a non existing drug that was introduced to the questionnaire. With only 0.42% of respondents reporting usage of Semeron. This would indicate that the survey data is likely of good quality, with most respondents providing attentive and truthful answers regarding their substance use.

4 Preparing the Dataset for Machine Learning

Since the main focus of the project is implementing machine learning models we decided to prepare our data for this purpose. Just like we converted our original dataset to be more human readable for data exploration we have changed our dataset to be more machine readable. The sex column was changed to binary data and for all the Drug columns, Education and Age we converted the data to ordinal data.

For the Ethnicity and Country columns we used a technique called One-Hot Encoding, where we transform a categorical variable with multiple possible values into multiple binary (0 or 1) columns. Each new column represents one possible category from the original variable, and for each observation, exactly one of these new columns will have the value 1 (hence “one-hot”) while all others will be 0.

It prevents the machine learning algorithm from assuming an arbitrary numerical relationship between categories. For example, if you simply encoded “USA”=1, “UK”=2, “Canada”=3, the algorithm might incorrectly assume that “Canada” is somehow “greater than” or “three times more important than” “USA”.

5 Machine Learning Models

5.1 Linear Model

As linear regression is not the ideal model for our dataset when making predictions we decided to use linear regression to better understand what factors influence drug use and focus on the better suited models on making predictions.

Table 2: Linear Regression Models for Drug Usage (Usage Level 0-6)

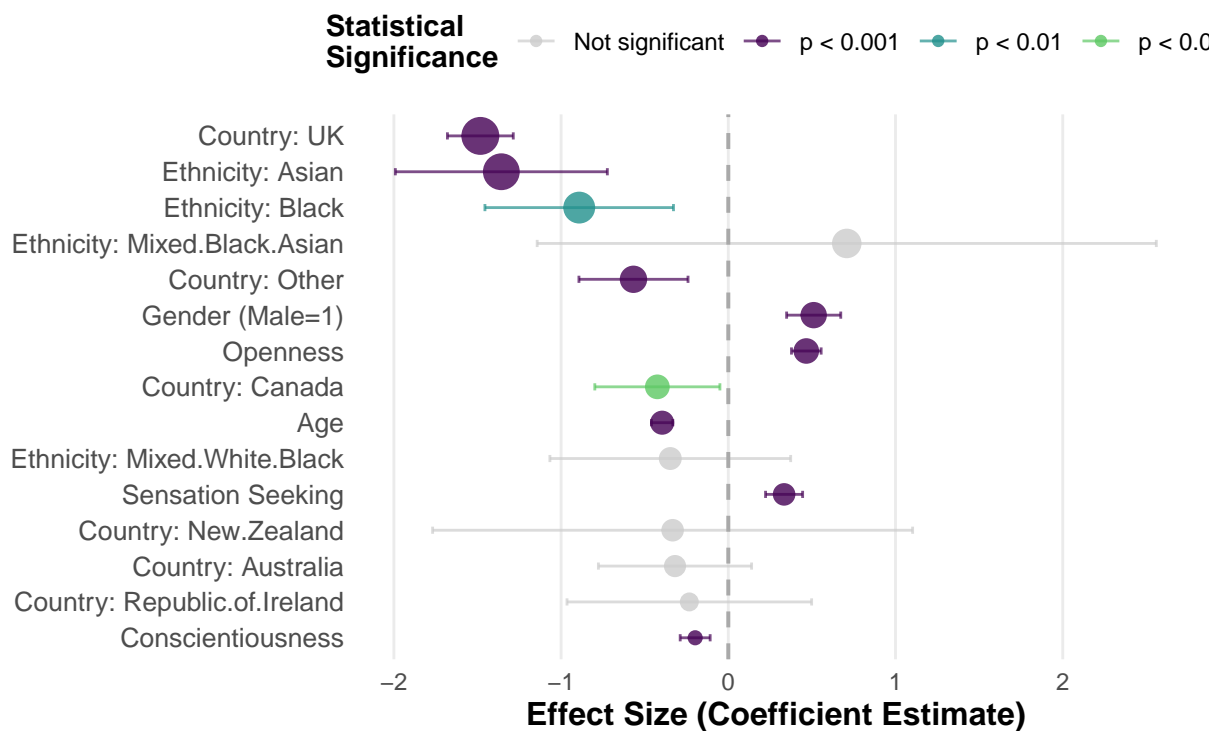
Variable	Drug Models				
	Cannabis	Alcohol	Nicotine	Coke	Ecstasy
Intercept	5.387***	3.929***	4.925***	1.588***	2.295***
Age	-0.396***	-0.031	-0.216***	-0.095***	-0.307***
Gender (Male=1)	0.511***	0.043	0.377***	0.216**	0.344***
Education Level	-0.116***	0.089***	-0.160***	-0.005	-0.026
Neuroticism	-0.112*	0.049	0.109	0.123**	-0.002
Extraversion	-0.098*	0.102**	0.009	0.113**	0.113**
Openness	0.467***	-0.040	0.158**	0.029	0.175***

Agreeableness	-0.037	-0.031	0.010	-0.144***	-0.026
Conscientiousness	-0.198***	-0.031	-0.198**	-0.095*	-0.169***
Impulsivity	0.017	-0.052	0.128	0.035	-0.003
Sensation Seeking	0.334***	0.204***	0.293***	0.272***	0.257***
N	1885	1885	1885	1885	1885
R ²	0.499	0.094	0.197	0.195	0.291
Adjusted R ²	0.494	0.083	0.188	0.186	0.283
F-statistic	88.484	9.151	21.715	21.454	36.412

Significance levels: * p<0.05; ** p<0.01; *** p<0.001

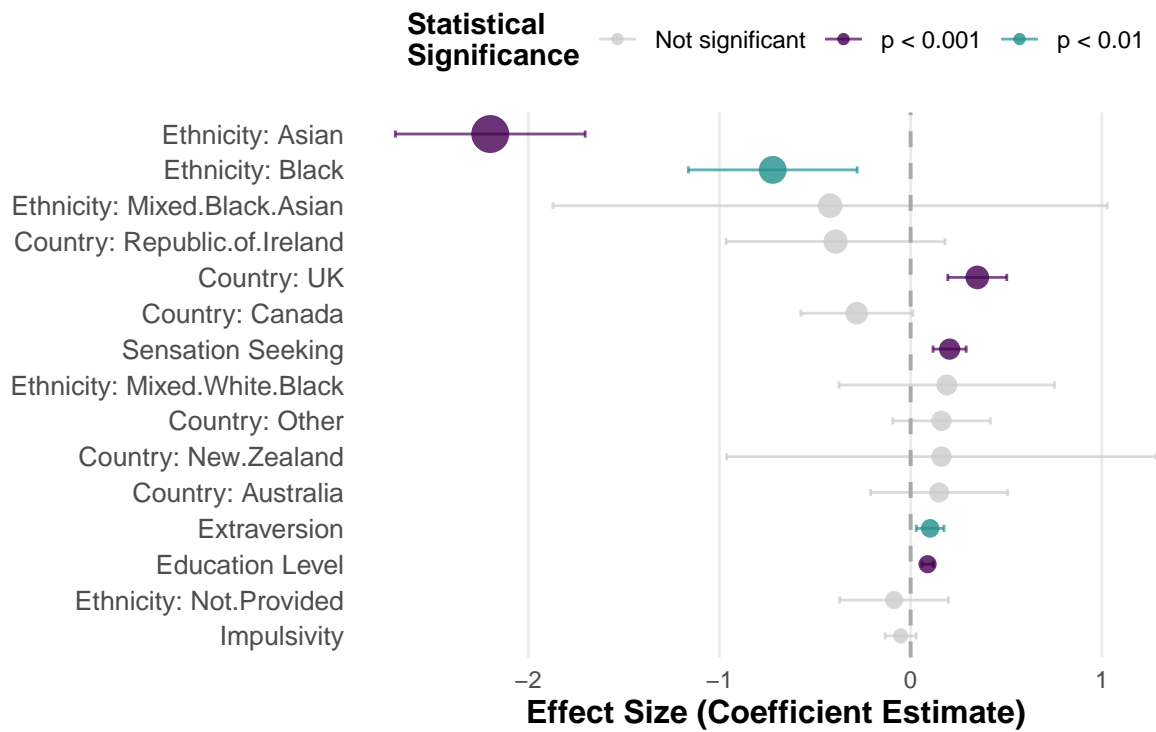
Predictors of Cannabis Usage

Estimated coefficients with 95% confidence intervals



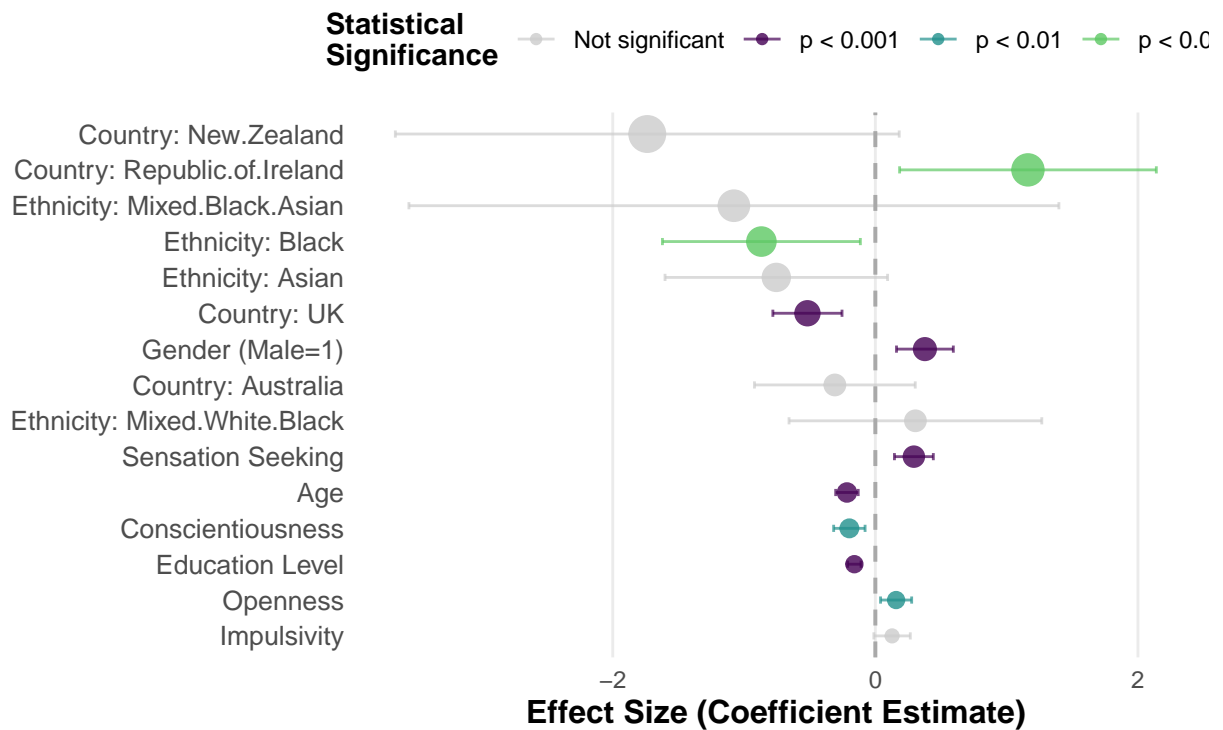
Predictors of Alcohol Usage

Estimated coefficients with 95% confidence intervals



Predictors of Nicotine Usage

Estimated coefficients with 95% confidence intervals



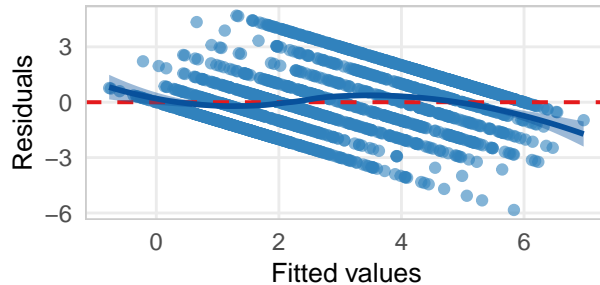
```
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
```

Cannabis Usage Model Diagnostics

Diagnostic Plots for Linear Regression Model

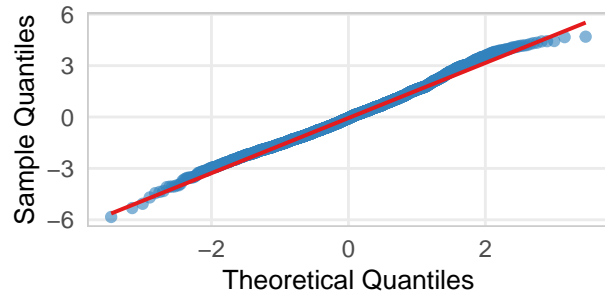
Residuals vs Fitted

Should show random scatter around the zero line



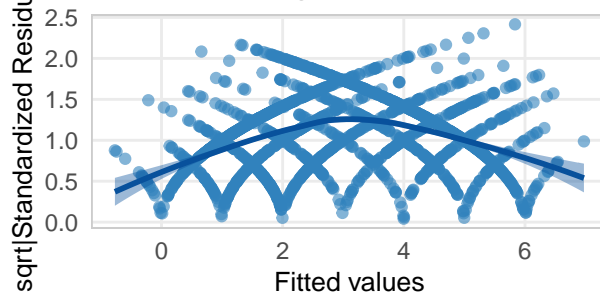
Normal Q-Q Plot

Points should follow the diagonal line



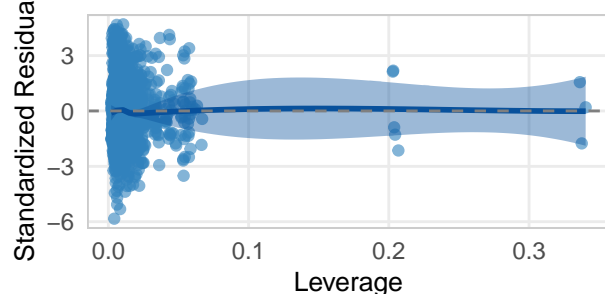
Scale-Location

Should show homogeneous variance



Residuals vs Leverage

Identifies influential cases



```
## TableGrob (2 x 1) "arrange": 2 grobs
##   z      cells   name      grob
## 1 1 (1-1,1-1) arrange gtable[arrange]
## 2 2 (2-2,1-1) arrange gtable[arrange]
```

```
##
## ===== SUMMARY OF DRUG CONSUMPTION ANALYSIS =====
```

```
## Linear regression models were fit for 5 drugs
```

```
## The best-fitting model was for Cannabis with Adjusted R2 = 0.494
```

```
##
```

```
## Key personality traits predicting drug use:
```

```
##
```

```
## Significant traits for Cannabis usage:
```

```
##   - Nscore: negative effect (-0.112) *
##   - Escore: negative effect (-0.098) *
##   - Oscore: positive effect (0.467) ***
##   - Cscore: negative effect (-0.198) ***
##   - SS: positive effect (0.334) ***
```

```
##
```

```
## Significant traits for Alcohol usage:
```

```
##   - Escore: positive effect (0.102) **
##   - SS: positive effect (0.204) ***
```

```
##
```

```
## Significant traits for Nicotine usage:
##   - Oscore: positive effect (0.158) **
##   - Cscore: negative effect (-0.198) **
##   - SS: positive effect (0.293) ***
##
## Significant traits for Coke usage:
##   - Nscore: positive effect (0.123) **
##   - Escore: positive effect (0.113) **
##   - Ascore: negative effect (-0.144) ***
##   - Cscore: negative effect (-0.095) *
##   - SS: positive effect (0.272) ***
##
## Significant traits for Ecstasy usage:
##   - Escore: positive effect (0.113) **
##   - Oscore: positive effect (0.175) ***
##   - Cscore: negative effect (-0.169) ***
##   - SS: positive effect (0.257) ***
```

6 Source

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