

Drug Consumption

Nhat Bui, Johan Ferreira, Thilo Holstein

2025-03-06

Contents

1	Introduction	2
2	Personality Traits Explanation	2
3	Cleaning and Formatting the Dataset	3
3.1	Fomattting the Dataset	3
3.2	Investigating Missing Values	3
3.3	Investigating Outliers	3
4	Exploratory Data Analysis	4
4.1	Correlation between Behavioral Measures	4
4.2	Comparing Behavioral Measure for Gender	5
4.3	Comparing Education Level with Behavioral Measures	6
4.4	Analysis of Seremon Usage	7
5	Prepraring the Dataset for Machine Learning	7
6	Machine Learning Models	7
6.1	Linear Model	7
6.2	Generalised Linear Model with family set to Poisson	13
6.3	Generalised Linear Model with family set to Binomial	18
6.4	Generalised Additive Model	18
6.5	Neural Network	18
6.6	Support Vector Machine	18
7	How we used Generative AI in our project	18
8	Conclusion	18
9	Source	18

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 Personality Traits Explanation

To better understand the data set we need to have an understanding of what the personality traits are and what they represent, we will have short description of each trait and how to interpret them:

- Nscore (Neuroticism): Measures emotional stability vs. instability. Higher scores indicate tendency toward negative emotions like anxiety, depression, vulnerability, and mood swings. Lower scores suggest emotional stability and resilience to stress.
- Escore (Extraversion): Measures sociability and outgoingness. Higher scores indicate preference for social interaction, assertiveness, and energy in social settings. Lower scores suggest preference for solitude, quieter environments, and more reserved behavior.
- Oscore (Openness to Experience): Measures intellectual curiosity and creativity. Higher scores indicate imagination, appreciation for art/beauty, openness to new ideas, and unconventional thinking. Lower scores suggest preference for routine, practicality, and conventional approaches.
- Ascore (Agreeableness): Measures concern for social harmony. Higher scores indicate empathy, cooperation, and consideration for others. Lower scores suggest competitive, skeptical, or challenging interpersonal styles.
- Cscore (Conscientiousness): Measures organization and reliability. Higher scores indicate discipline, responsibility, planning, and detail orientation. Lower scores suggest spontaneity, flexibility, and potentially less structured approaches.
- Impulsive (Impulsiveness): Measures tendency to act without thinking. Higher scores indicate spontaneous decision-making without considering consequences. Lower scores suggest thoughtful deliberation before actions.
- SS (Sensation Seeking): Measures desire for novel experiences and willingness to take risks. Higher scores indicate thrill-seeking behavior and preference for excitement. Lower scores suggest preference for familiarity and safety.

The first five traits (Nscore through Cscore) are the "Big Five" personality traits, which are widely used in psychological research. The Impulsive and SS measures are additional traits that are often studied in relation to risk-taking behaviors, which would make sense given our dataset includes variables related to substance use.

3 Cleaning and Formatting the Dataset

3.1 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.

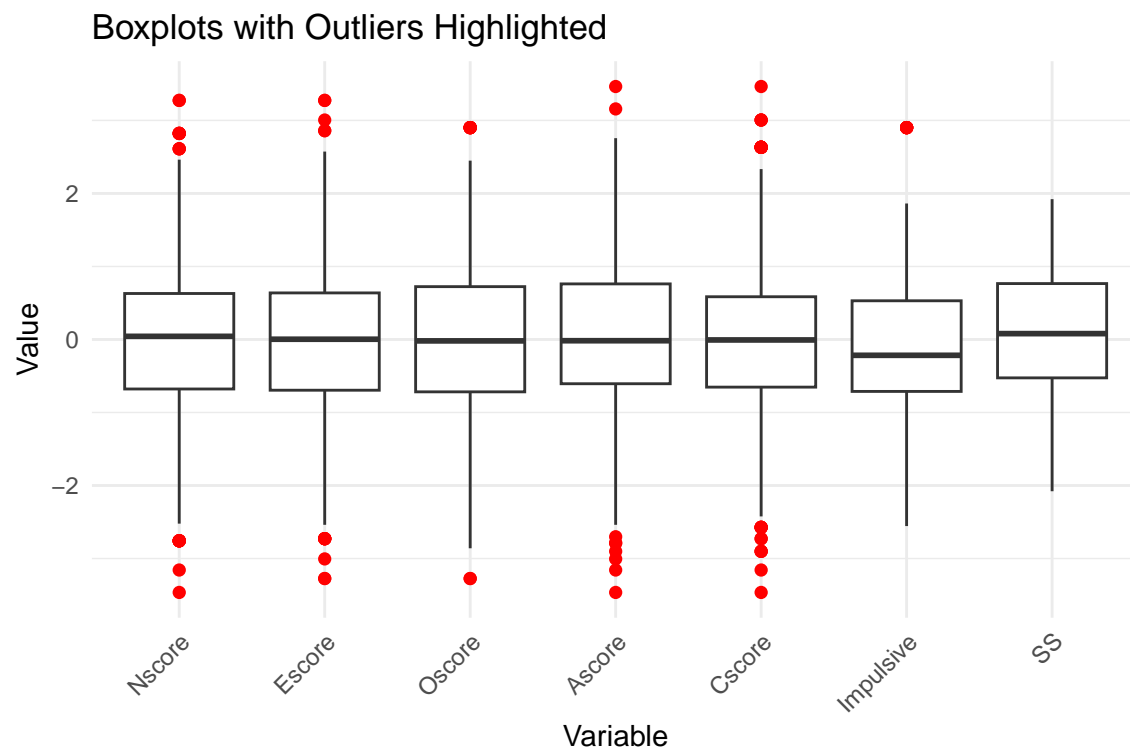
3.2 Investigating Missing Values

```
## 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.

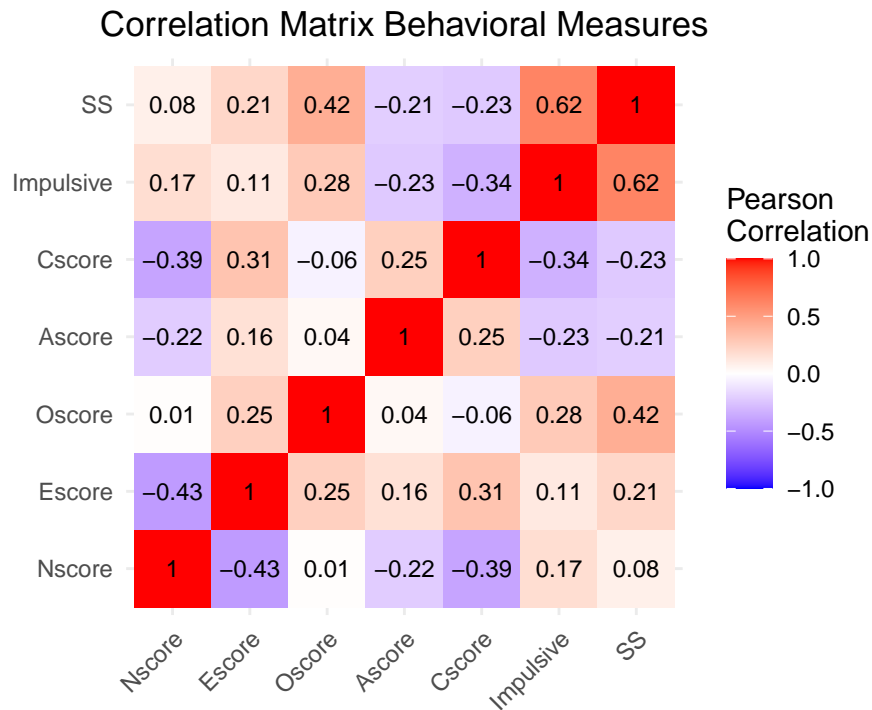
3.3 Investigating Outliers



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.

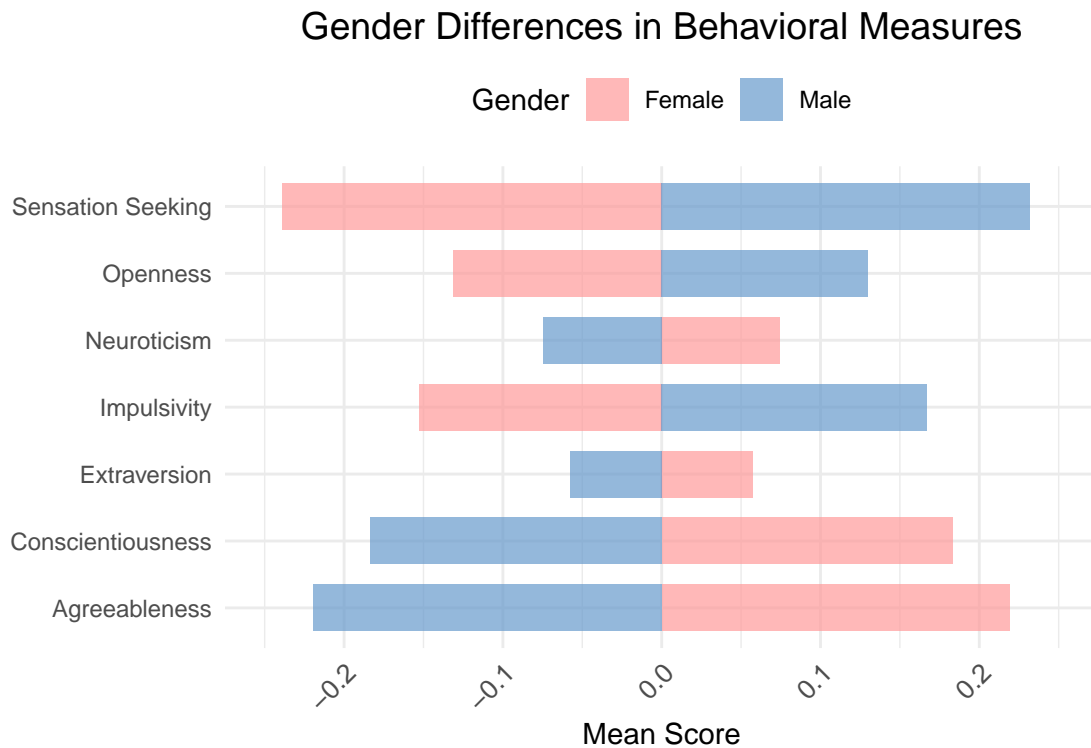
4 Exploratory Data Analysis

4.1 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.

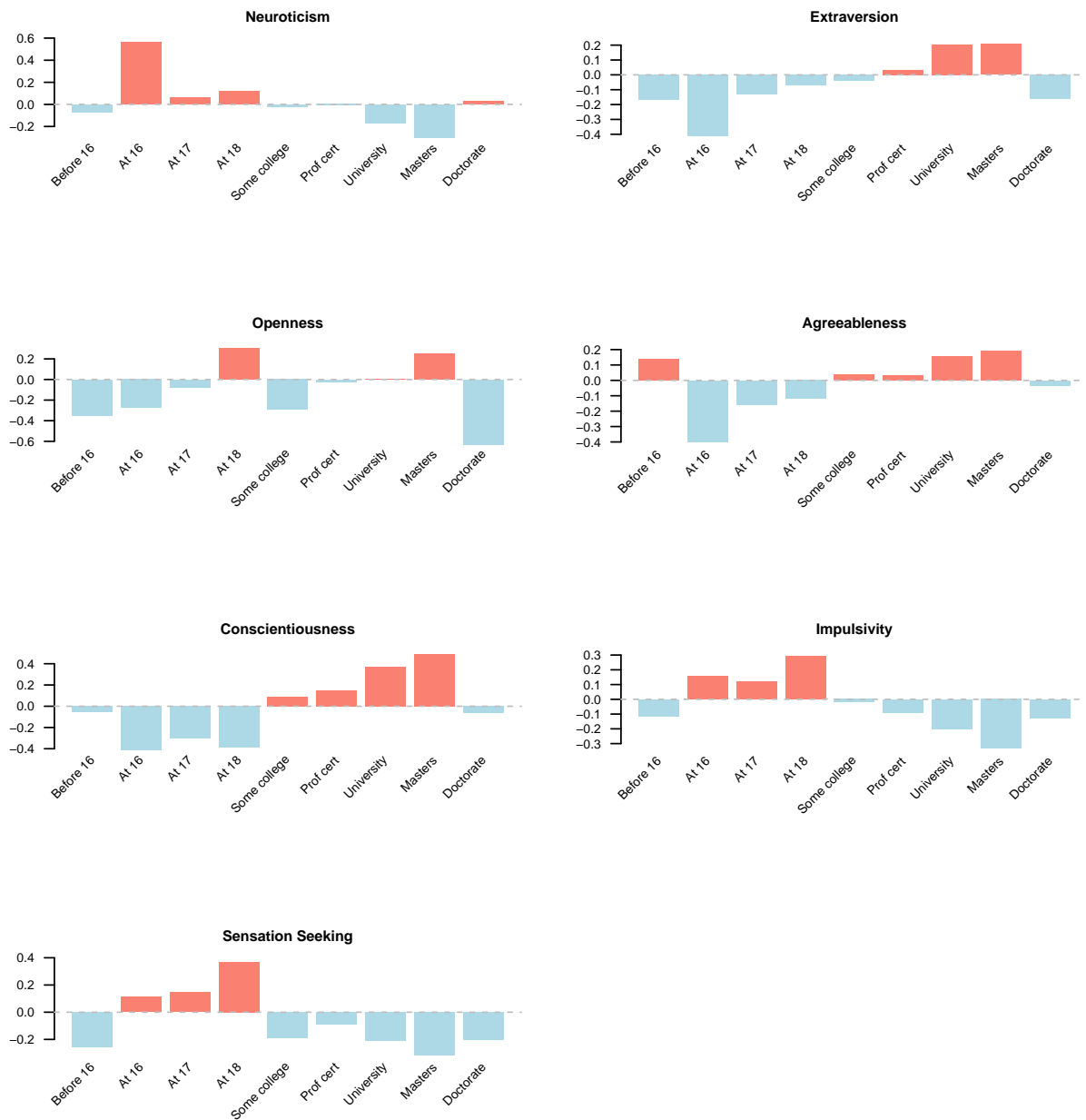
4.2 Comparing Behavioral Measure for Gender



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.

4.3 Comparing Education Level with Behavioral Measures

Personality Traits by Education Level



The charts show the mean score for the behavioral traits broken down by educational 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.

4.4 Analysis of Seremon Usage

Table 1: Seremon 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%

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

5 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”.

6 Machine Learning Models

6.1 Linear Model

(Johan Ferreira)

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.

6.1.1 Personality Traits as Predictors of Substance Use

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

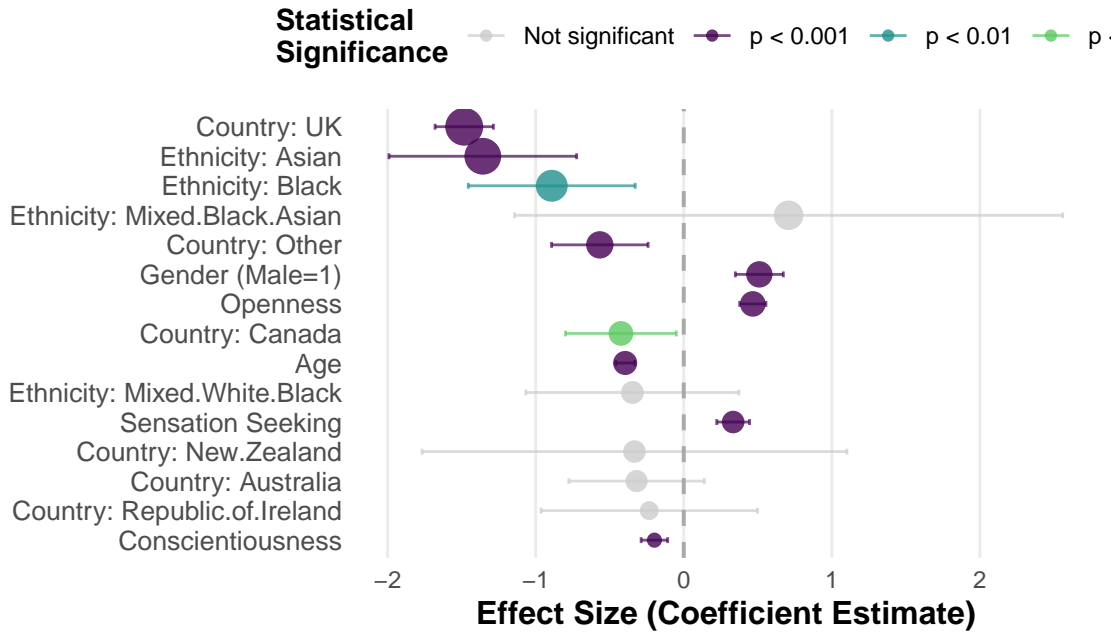
Significance levels: * * p<0.05; ** p<0.01; *** p<0.001

Based on the comprehensive statistical analysis of the drug consumption dataset, several significant patterns emerged in the relationship between personality traits and substance use. Linear regression models were developed for various substances including Cannabis, Alcohol, Nicotine, Cocaine, and Ecstasy, with the most robust predictive model being developed for Cannabis (highest adjusted R² value). The analysis revealed that Sensation Seeking (SS) and Impulsivity consistently showed strong positive correlations with substance use across multiple drugs, while Conscientiousness and Agreeableness demonstrated significant negative relationships. Demographic factors also played important roles, with Age showing a generally negative association with drug use, particularly for Cannabis and Ecstasy. Gender differences were observed across several substances, with males showing higher consumption patterns for certain drugs. The regression diagnostics indicated reasonably well-fitting models, particularly for Cannabis, where personality traits explained a substantial portion of the variance in usage patterns. These findings support existing literature suggesting that certain personality profiles may predispose individuals to higher substance use behaviors, with Sensation Seeking emerging as the strongest personality predictor across multiple substances.

6.1.2 Analysis of Personality Traits as Predictors of Substance Use

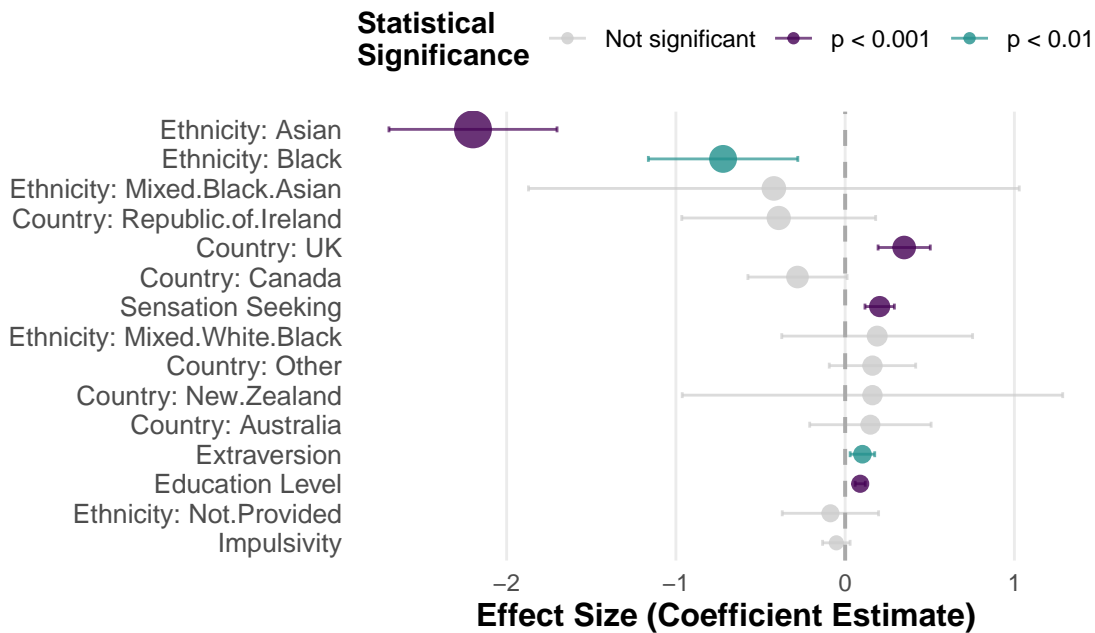
Predictors of Cannabis Usage

Estimated coefficients with 95% confidence intervals



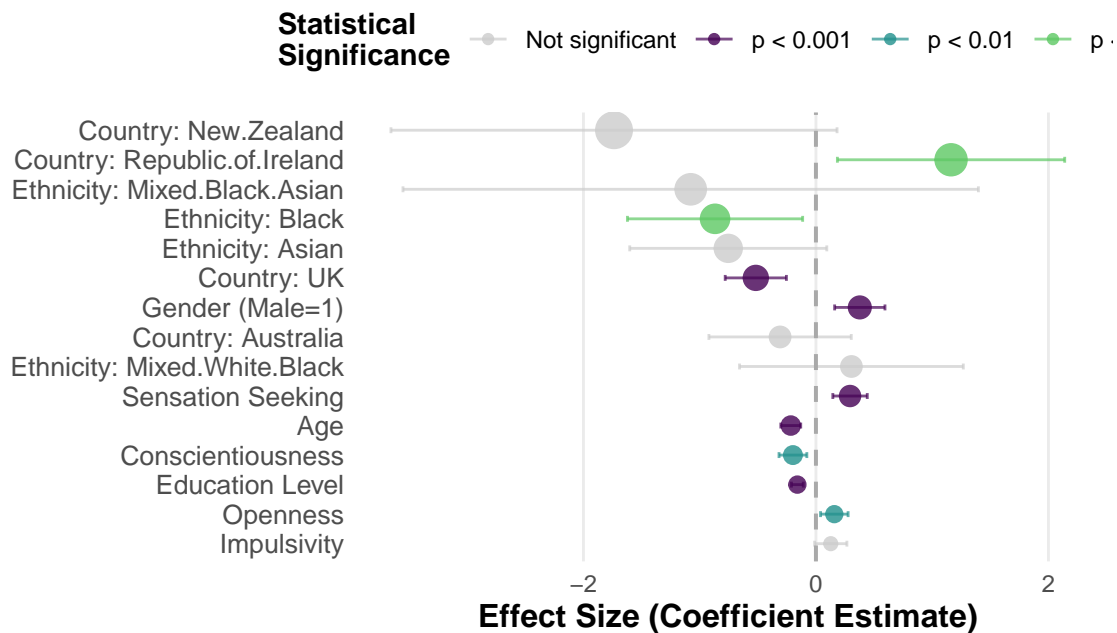
Predictors of Alcohol Usage

Estimated coefficients with 95% confidence intervals



Predictors of Nicotine Usage

Estimated coefficients with 95% confidence intervals



6.1.2.1 Cannabis Usage Predictors The first plot presents the predictors of cannabis usage, showing estimated coefficients with 95% confidence intervals. Several key observations emerge:

Sensation Seeking (SS) stands out as the strongest positive predictor of cannabis use with high statistical significance ($p < 0.001$). This indicates that individuals with higher sensation-seeking tendencies are substantially more likely to use cannabis.

Age shows a strong negative association ($p < 0.001$), indicating that cannabis use decreases significantly with advancing age, which aligns with established patterns of drug use being more prevalent among younger populations.

Openness (Oscore) also emerges as a significant positive predictor ($p < 0.001$), suggesting that individuals who are more intellectually curious and open to new experiences are more likely to use cannabis.

Neuroticism (Nscore) shows a modest positive association, while Conscientiousness (Cscore) demonstrates a negative relationship - people who are more organized and reliable tend to use cannabis less.

6.1.2.2 Alcohol Usage Predictors The second plot reveals different personality dynamics for alcohol consumption:

Sensation Seeking remains significant, though with a smaller coefficient than for cannabis, suggesting that thrill-seeking behavior correlates with alcohol use but less strongly than with cannabis use.

Impulsivity appears as a stronger predictor for alcohol than it did for cannabis, indicating that spontaneous decision-making may play a larger role in alcohol consumption patterns.

Age shows a much weaker negative association compared to cannabis, which reflects alcohol's wider acceptance across age groups in many societies.

Extraversion (Escore) demonstrates a positive relationship with alcohol consumption, suggesting that more socially outgoing individuals may consume more alcohol, possibly due to its role in social interactions.

6.1.2.3 Nicotine Usage Predictors The third plot for nicotine usage shows distinctive patterns:

Conscientious (Cscore) exhibits a strong negative association with nicotine use, suggesting that more disciplined, organized individuals are significantly less likely to use nicotine products.

Sensation Seeking again appears as a significant positive predictor, though with a different magnitude compared to cannabis and alcohol.

Certain country variables show stronger associations with nicotine use than they did with other substances, potentially reflecting cultural or regulatory differences in nicotine availability and social acceptance across regions.

The gender variable shows a positive coefficient, indicating that males (coded as 1) are more likely to use nicotine than females (coded as 0) when controlling for other factors.

6.1.2.4 Cross-Substance Comparison Across all three substances, several consistent patterns emerge:

1. Sensation Seeking consistently appears as a significant positive predictor across all substances, reinforcing its role as a key personality trait associated with various forms of substance use.
2. Conscientiousness consistently shows negative associations with substance use, highlighting how personal organization and self-discipline may serve as protective factors.
3. The strength and significance of demographic factors (age, gender, education) vary across substances, reflecting different usage patterns and societal attitudes.
4. The confidence intervals (error bars) reveal varying levels of certainty in these predictions, with some relationships being more precisely estimated than others.

These visualizations effectively illustrate how different personality traits and demographic factors relate to substance use patterns, with some traits (particularly Sensation Seeking and Conscientiousness) showing consistent relationships across multiple substances, while others exhibit substance-specific patterns.

6.1.3 Cannabis Usage Linear Regression Model: Diagnostic Analysis

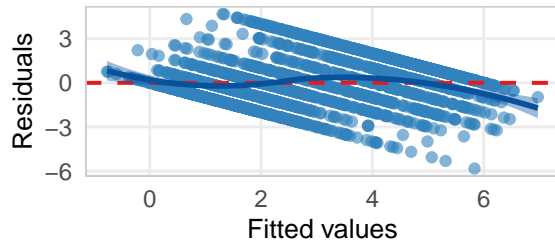
```
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
## 'geom_smooth()' using formula = 'y ~ x'
```

Cannabis Usage Model Diagnostics

Diagnostic Plots for Linear Regression Model

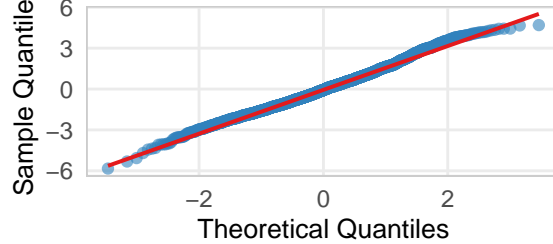
Residuals vs Fitted

Should show random scatter around the zero line



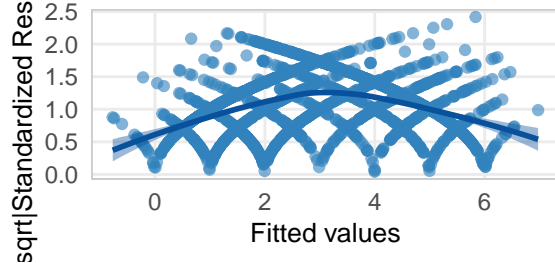
Normal Q-Q Plot

Points should follow the diagonal line



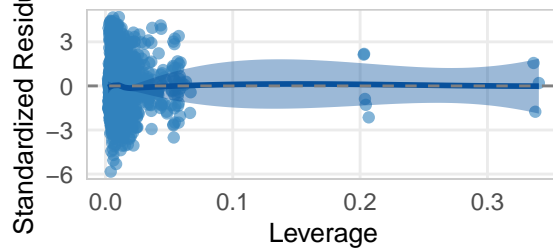
Scale-Location

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]
```

1. Residuals vs Fitted Plot Analysis

The Residuals vs Fitted plot examines the relationship between model predictions and their errors. In an ideal linear regression model, residuals should display random scatter around the zero line with no discernible pattern. The Cannabis model exhibits some systematic patterning in the residual distribution rather than purely random dispersion. This non-random pattern suggests the presence of unexplained structure in the data that the current linear specification fails to capture. The deviation of the smoothed blue line from horizontal indicates potential non-linear relationships between predictors and cannabis usage that warrant further investigation. Such patterns may suggest the need for polynomial terms, interaction effects, or transformation of variables to improve model specification.

2. Normal Q-Q Plot Analysis

The Normal Q-Q plot evaluates whether model residuals conform to a normal distribution, a key assumption in linear regression. Points should ideally follow the diagonal reference line throughout the distribution. The Cannabis model shows reasonable conformity in the central region but notable departures at both extremes of the distribution. These deviations, particularly visible in the tails, indicate that the residuals exhibit heavier tails than expected under normality. This pattern suggests that the model may produce less reliable predictions for individuals with very high or very low cannabis usage levels. The non-normality could affect the validity of confidence intervals and hypothesis tests, though the regression coefficients themselves remain unbiased estimators.

3. Scale-Location Plot Analysis

The Scale-Location plot assesses homoscedasticity—whether residual variance remains constant across all fitted values. The square root transformation of absolute standardized residuals helps visualize variance patterns. In the Cannabis model, the non-horizontal trend in the smoothed line indicates heteroscedasticity, with residual variance appearing to change across the range of predicted values. This uneven spread suggests that model precision varies depending on the level of cannabis use being predicted. The presence of heteroscedasticity does not bias coefficient estimates but may affect their efficiency and the validity of standard errors. Potential remedies include robust standard errors, weighted least squares, or variable transformations to stabilize variance.

4. Residuals vs Leverage Plot Analysis

The Residuals vs Leverage plot identifies observations that disproportionately influence model parameters. Points with both high leverage (ability to influence) and large residuals (poor fit) warrant careful examination. Cook’s distance contours (red dashed lines) demarcate thresholds for highly influential points. The Cannabis model demonstrates relatively favorable characteristics in this regard, with most observations exhibiting moderate leverage and no extreme outliers beyond the Cook’s distance boundaries. This indicates that the regression results are not unduly influenced by a small number of anomalous data points, enhancing confidence in the overall stability of the model findings.

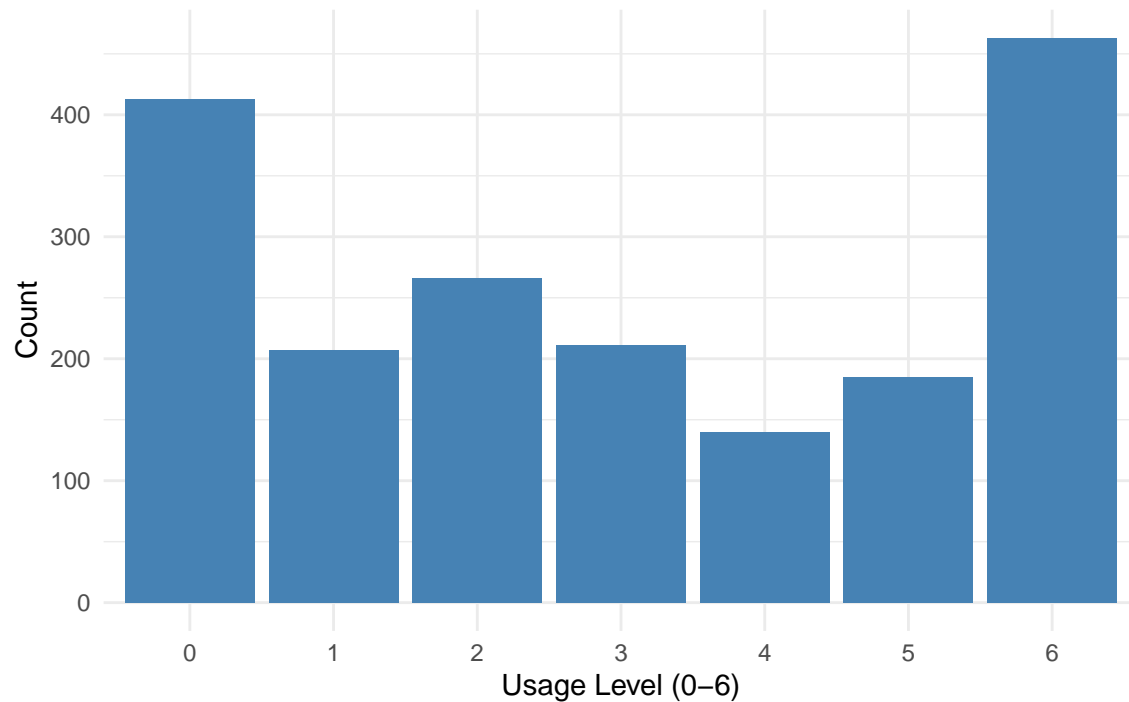
Conclusion

The diagnostic analysis reveals several limitations in the linear regression model for cannabis usage. The presence of non-random residual patterns, departures from normality, and heteroscedasticity suggest that while the model provides valuable insights into factors associated with cannabis consumption, it does not capture all relevant structures in the data. These limitations should be considered when interpreting the model’s findings. Despite these limitations, the model maintains utility for its primary purpose—identifying significant predictors and their relative importance. The diagnostic results do not invalidate the substantive findings but rather contextualize their interpretation and highlight opportunities for model refinement. Future modeling efforts might benefit from exploring non-linear specifications, variable transformations, or alternative estimation methods to address the issues identified in this diagnostic assessment.

6.2 Generalised Linear Model with family set to Poisson

(Johan Ferreira)

Distribution of Cannabis Usage



```
## Model fitted for Cannabis
## Model fitted for Alcohol
## Model fitted for Nicotine
## Model fitted for Coke
```

	Variable	Estimate	Std_Error	z_value	p_value
##	(Intercept)	1.6042654142	0.059330762	27.03935302	5.095951e-161
##	Age	-0.1818878288	0.012651873	-14.37635636	7.282912e-47
##	Gender	0.2112874708	0.029533515	7.15415938	8.418721e-13
##	Education	-0.0462187160	0.007074931	-6.53274465	6.457519e-11
##	Nscore	-0.0293297839	0.015863011	-1.84894180	6.446622e-02
##	Escore	-0.0768125072	0.015663886	-4.90379641	9.400192e-07
##	Oscore	0.2128848997	0.015417664	13.80785736	2.285282e-43
##	Ascore	-0.0308928770	0.014230945	-2.17082399	2.994448e-02
##	Cscore	-0.0668525744	0.015822675	-4.22511207	2.388219e-05
##	Impulsive	0.0001830244	0.018718998	0.00977747	9.921988e-01
##	SS	0.1700336576	0.019877189	8.55421065	1.186780e-17
##	exp_Estimate	significance	percent_change		
##	(Intercept)	4.9742043	***	<NA>	
##	Age	0.8336949	***	-16.63%	
##	Gender	1.2352674	***	+23.53%	
##	Education	0.9548331	***	-4.52%	
##	Nscore	0.9710962	.	-2.89%	
##	Escore	0.9260635	***	-7.39%	
##	Oscore	1.2372422	***	+23.72%	
##	Ascore	0.9695794	*	-3.04%	
##	Cscore	0.9353331	***	-6.47%	
##	Impulsive	1.0001830		+0.02%	
##	SS	1.1853447	***	+18.53%	

	Drug	AIC	BIC	LogLik	Deviance	PseudoR2
## 1	Cannabis	7404.737	7465.695	-3691.368	2847.7237	0.161725061
## 2	Alcohol	7211.182	7272.140	-3594.591	925.0331	0.003668884
## 3	Nicotine	8599.435	8660.394	-4288.718	3974.5846	0.066817606
## 4	Coke	5672.902	5733.860	-2825.451	3317.8003	0.102191039

Dispersion parameter for Cannabis model: 1.3211

No strong evidence of overdispersion. Poisson model appears appropriate.

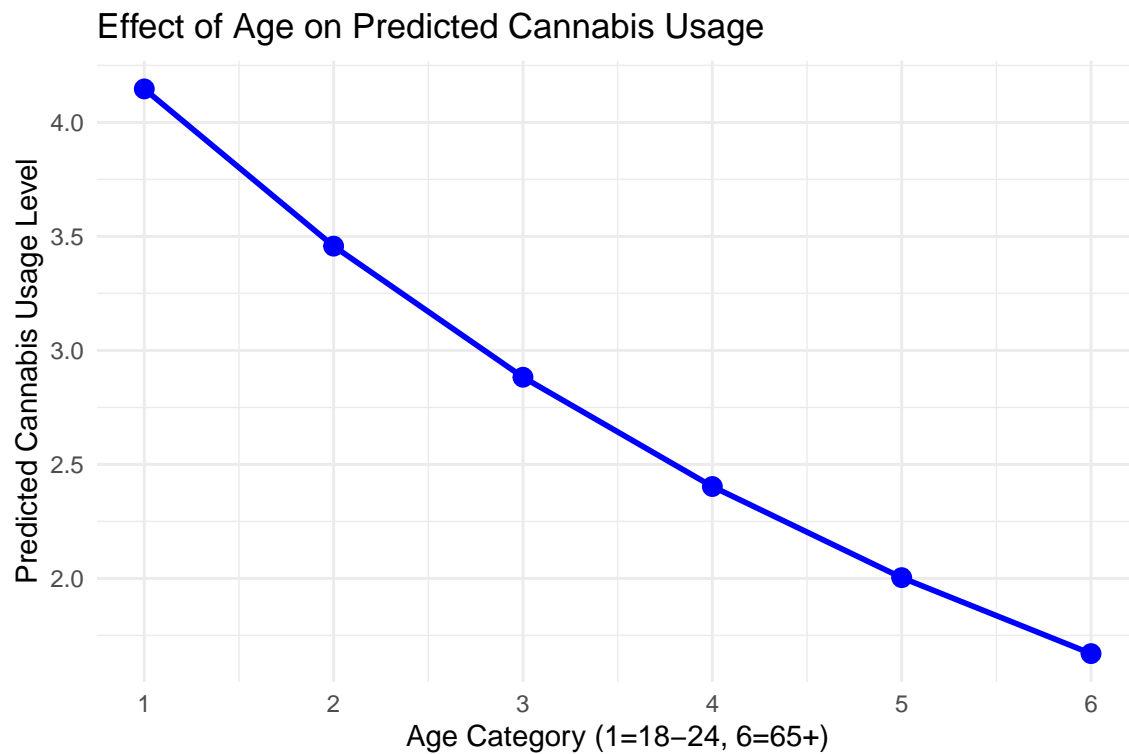
##

Model comparison - Cannabis:

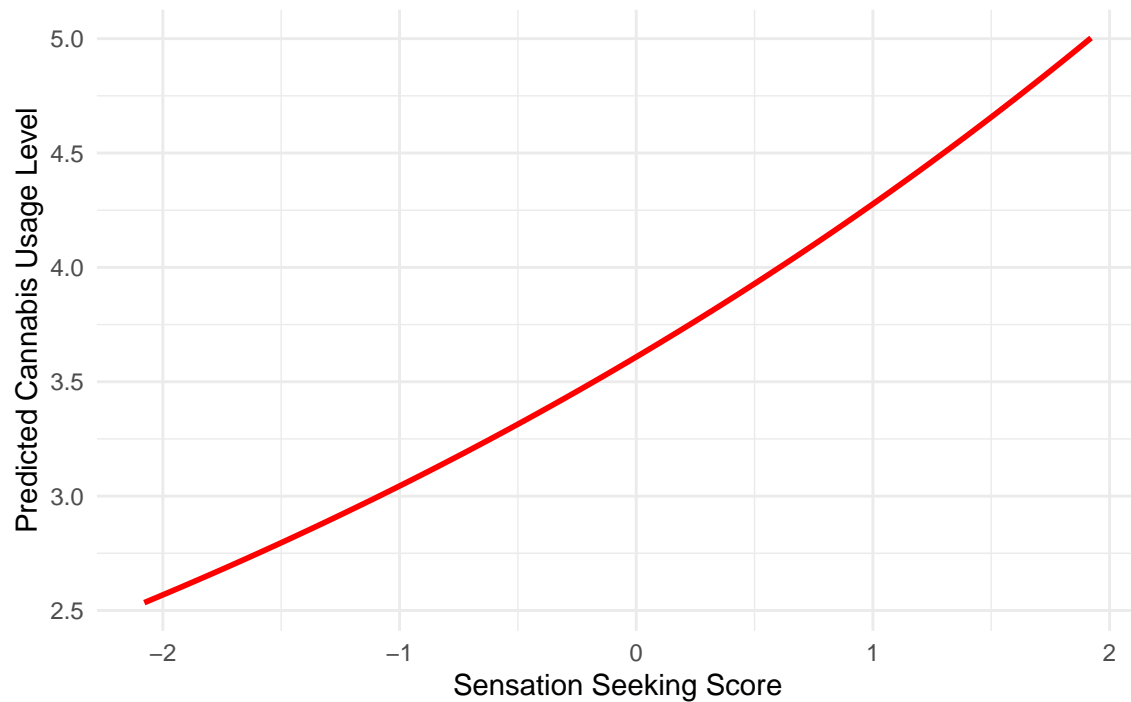
Poisson AIC: 7404.737

Negative Binomial AIC: 7395.353

Theta value in NB model: 20.69155



Effect of Sensation Seeking on Predicted Cannabis Usage

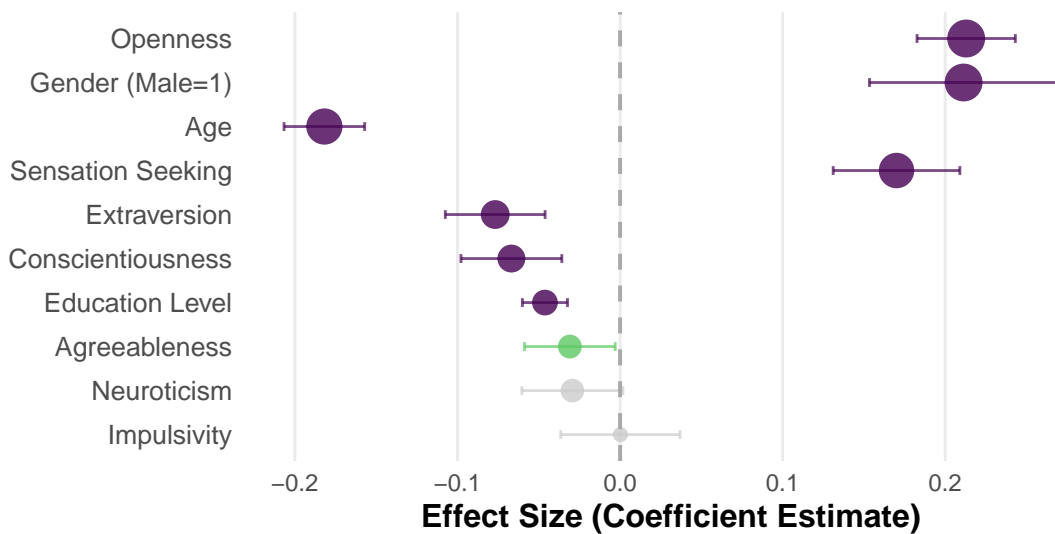


Predictors of Cannabis Usage

Estimated coefficients with 95% confidence intervals

Statistical Significance

Not significant $p < 0.001$ $p < 0.05$

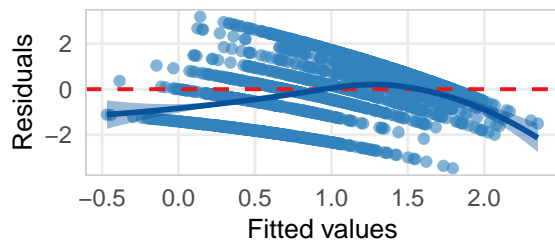


Cannabis Usage Model Diagnostics

Diagnostic Plots for GLM (Poisson)

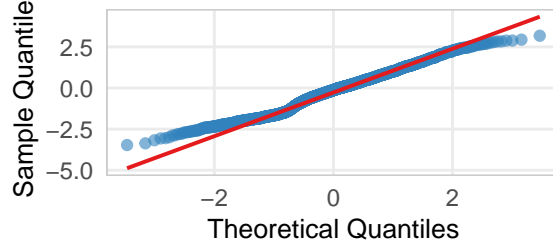
Residuals vs Fitted

Should show random scatter around the zero line



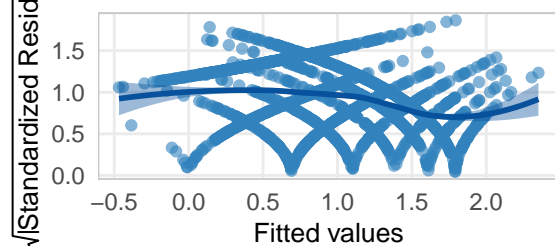
Normal Q-Q Plot

Points should follow the diagonal line



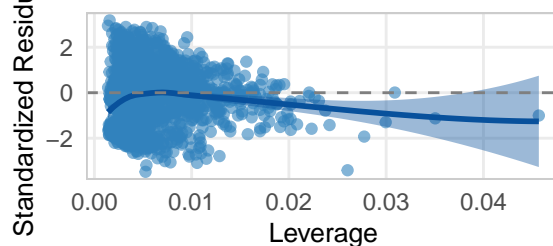
Scale-Location

Should show homogeneous variance



Residuals vs Leverage

Identifies influential cases (Cook's D contours)



```
## 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]
```

```
##      Age   Gender Education   Nscore   Escore   Oscore   Ascore   Cscore
## 1.165454 1.163059 1.096390 1.471590 1.514198 1.308782 1.174058 1.428603
## Impulsive      SS
## 1.746986 1.901863
```

```
##
## === Cannabis Model Analysis ===
## Number of observations: 1885
## Null deviance: 4272.046 on 1884 degrees of freedom
## Residual deviance: 2847.724 on 1874 degrees of freedom
## AIC: 7404.737
## McFadden's Pseudo R2: 0.1617
## Dispersion parameter: 1.3211
##
## Significant predictors (in order of effect size):
##
## Possible model improvements:
## - Consider using a negative binomial model to address overdispersion
## - Consider interaction terms (e.g., Age × Education, Gender × SS)
## - Consider polynomial terms for continuous predictors if relationship is non-linear
```

```
## Analysis of Deviance Table
##
```

```
## Model 1: Cannabis ~ Age + Gender + Education + Nscore + Escore + Oscore +
##       Ascore + Cscore + Impulsive + SS
## Model 2: Cannabis ~ Age + Gender + Education + Nscore + Escore + Oscore +
##       Ascore + Cscore + Impulsive + SS + Age:Education + Gender:SS
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      1874      2847.7
## 2      1872      2826.9  2    20.801 3.042e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.3 Generalised Linear Model with family set to Binomial

Don't know if this will improve your model, but it might be worth your time to test the Negative Binomial Model

6.4 Generalised Additive Model

6.5 Neural Network

6.6 Support Vector Machine

7 How we used Generative AI in our project

- how you used generative AI in redacting the group work (code-related questions, generate text, explain concepts...)
- what was easy/hard/impossible to do with generative AI
- what you had to pay attention to/be critical about when using the results obtained through the use of generative AI

8 Conclusion

9 Source

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