Drug Consumption

Nhat Bui, Johan Ferreira, Thilo Holstein

2025-03-06

Contents

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 Decsription 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 the 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 Fomatting 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.

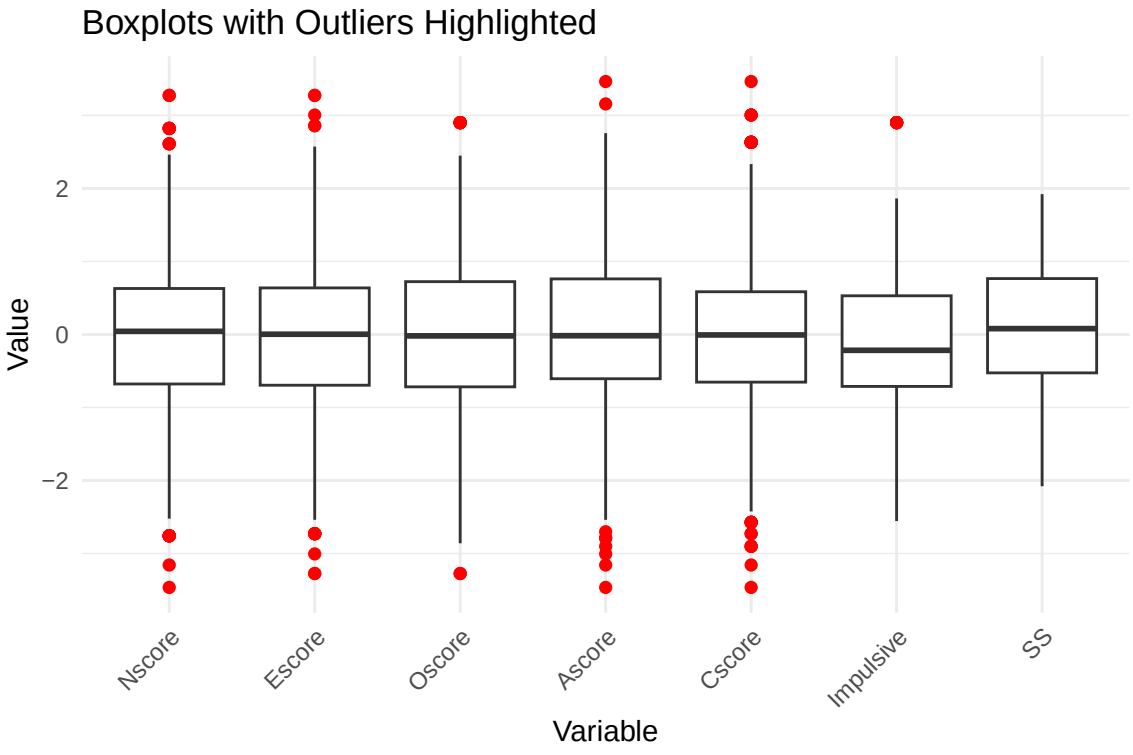
3.2 Investigating Missing Values

```
## 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.

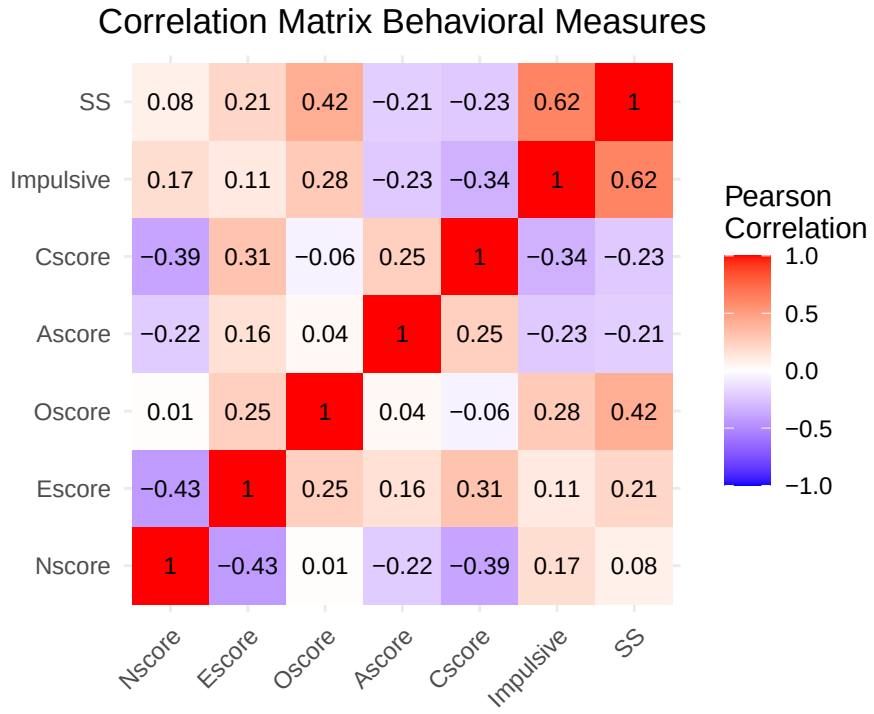
3.3 Investigating Outliers



As can be seen from the box plots our data set has some values that are outside of the upper and lower bounds. All thought 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 thougher for example SS (Sensation Seeking) has a positive correlation with Escore (Extraversion), Oscore (Openness) and Impulsive while they in turn also have positive correlations which 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

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 precived as bad like Neuroticism, Impulsivity and Sensation Seeking is more prevelent with lower education levels inccluding Not Provided and steadily decrease as the level of education increases.

4.4 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 exsisting drug that was introduced to the questonaire. With only 0.42% of resopndents 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.

5 Prepraring 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 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 transforms 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 influences drug use and focus in 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)

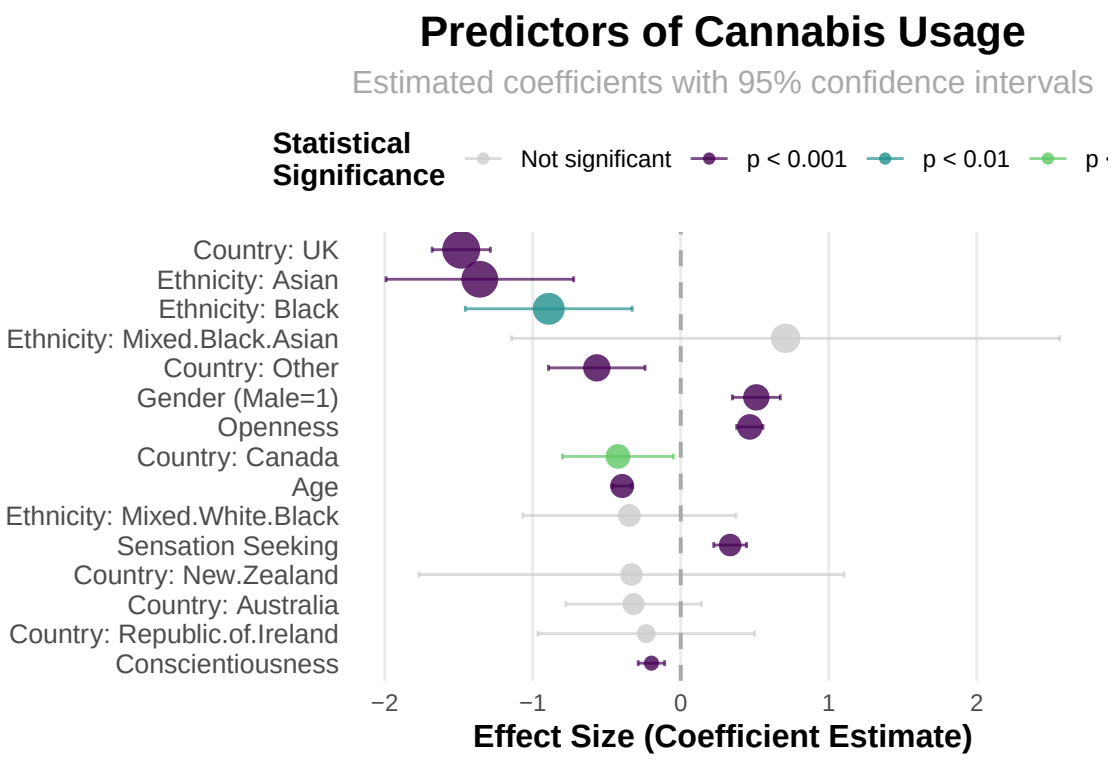
	Drug Models				
Variable	Cannabis	Alcohol	Nicotine	Coke	Ecstasy

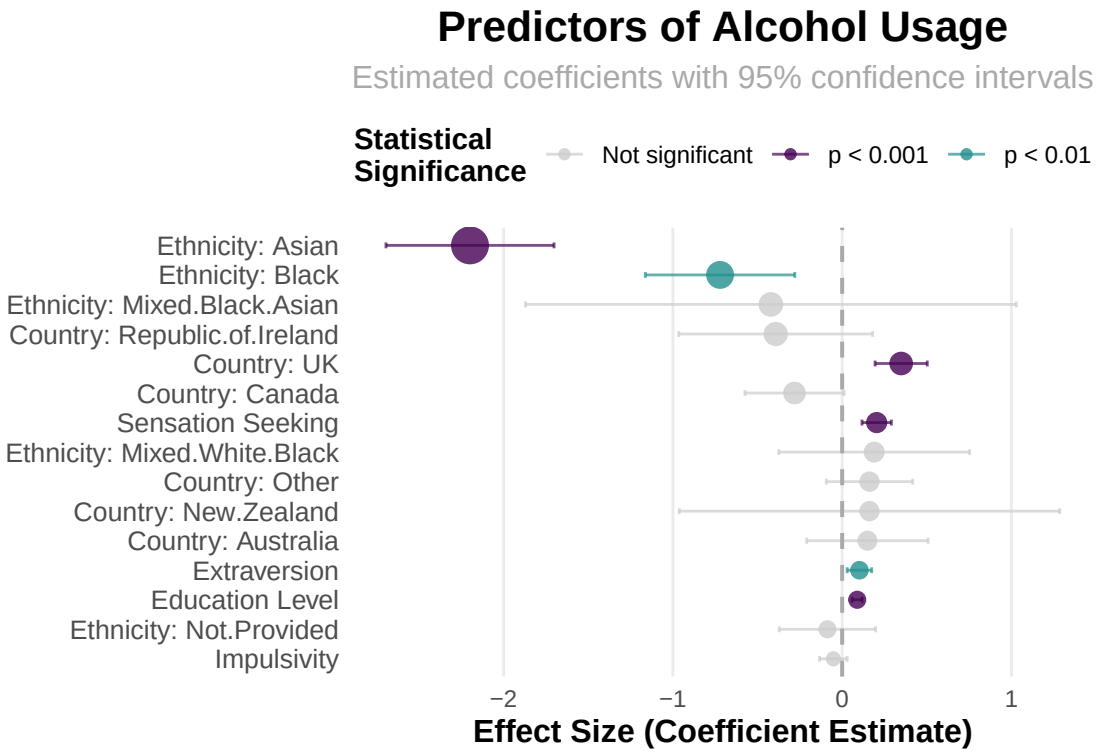
Intercept	5.387***	3.929***	4.925***	1.588***	2.295***
Age	-0.396***	-0.031	-0.216***	-0.095***	-0.307***
Gender (Male=1)	0.511***	0.043	0.377***	0.216**	0.344***
Education Level	-0.116***	0.089***	-0.160***	-0.005	-0.026
Neuroticism	-0.112*	0.049	0.109	0.123**	-0.002
Extraversion	-0.098*	0.102**	0.009	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

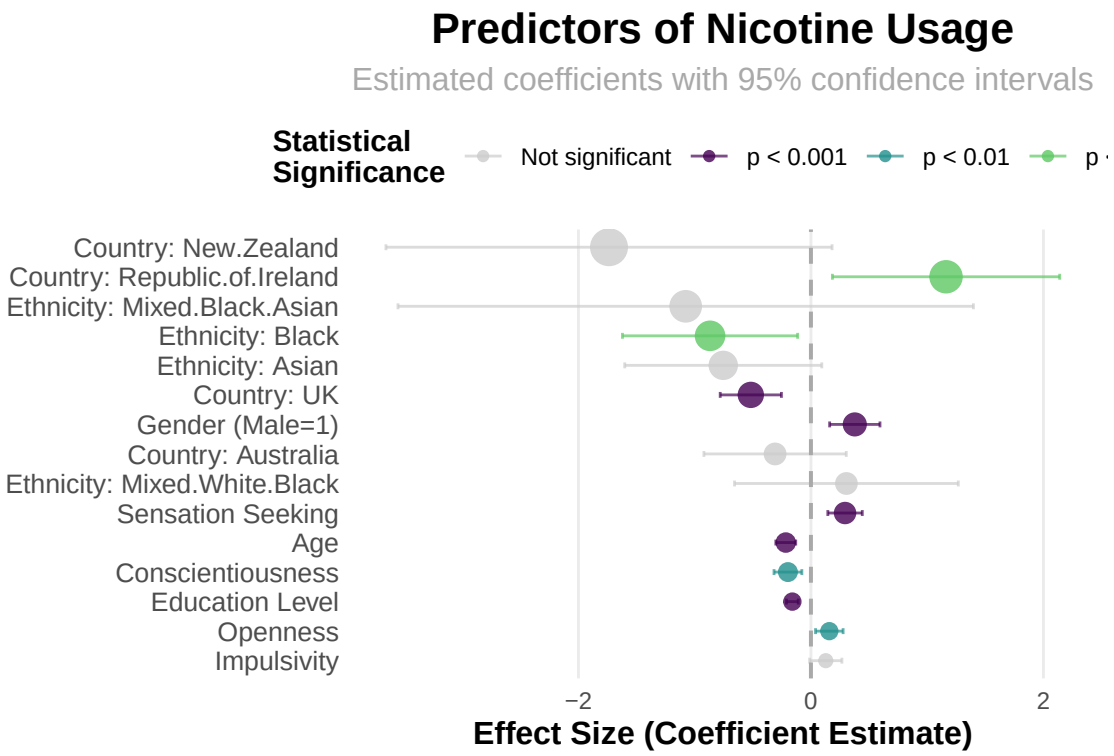
Significance levels: * * p<0.05; ** p<0.01; *** p<0.001

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 R^2 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







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

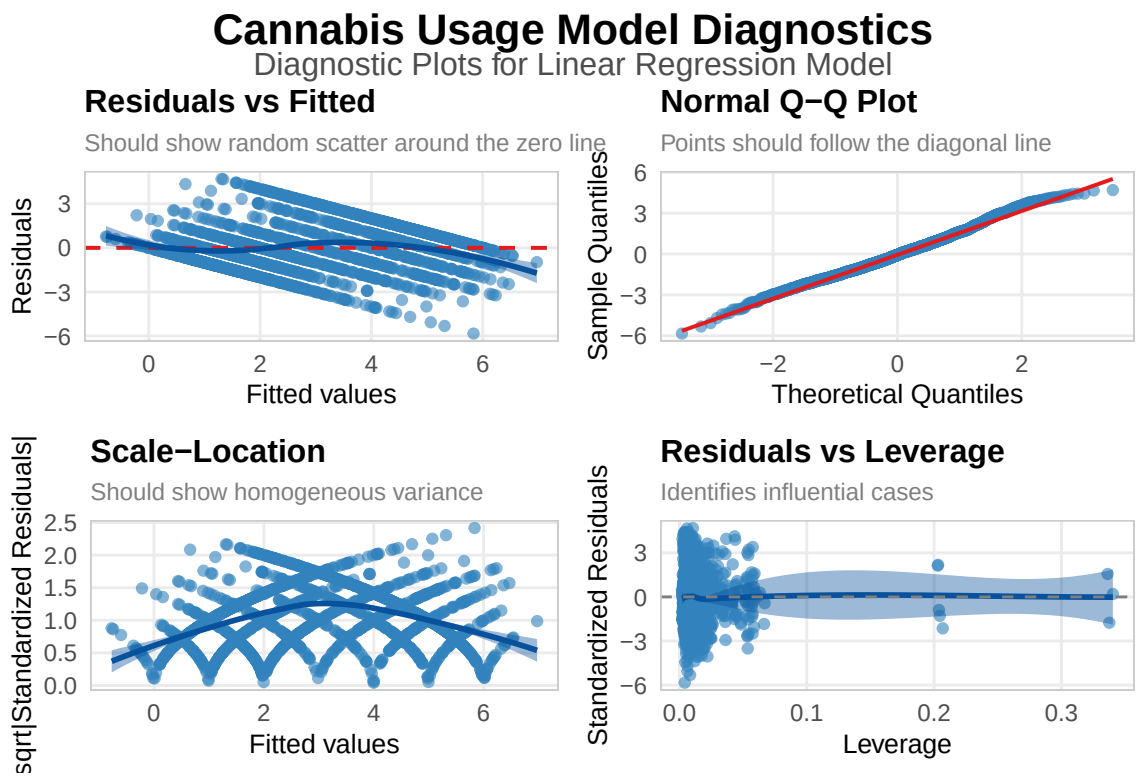
Across all three substances, several consistent patterns emerge:

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'
```




```
## 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

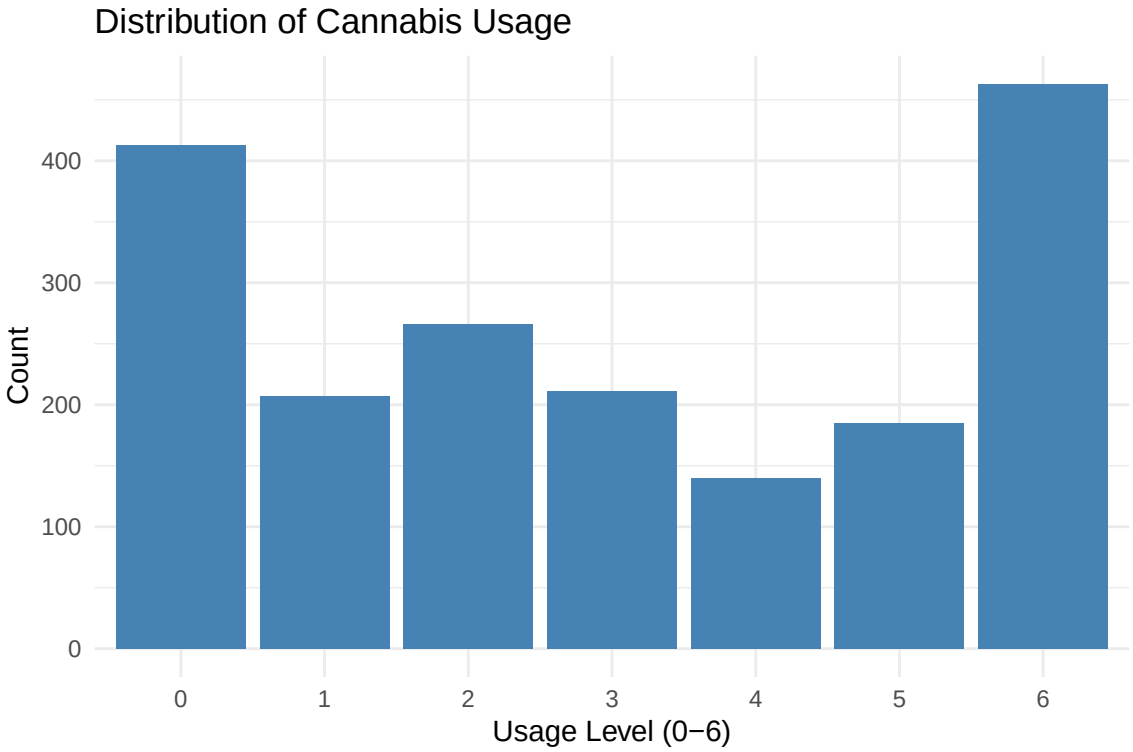
The Residuals vs Leverage plot identifies observations that disproportionately influence model parameters. Points with both high leverage (ability to influence) and large residuals (poor fit) warrant careful examination. Cook's distance contours (red dashed lines) demarcate thresholds for highly influential 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)




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

Table 3: Poisson Regression Results for Cannabis Usage

	Predictor	Coefficient	Exp(Coefficient)	% Change	p-value	Significance
(Intercept)	Intercept	1.6043	4.9742	NA	0.0000	***
Age	Age	-0.1819	0.8337	-16.63%	0.0000	***
Gender	Gender (Male=1)	0.2113	1.2353	+23.53%	0.0000	***
Education	Education Level	-0.0462	0.9548	-4.52%	0.0000	***
Nscore	Neuroticism	-0.0293	0.9711	-2.89%	0.0645	.
Escore	Extraversion	-0.0768	0.9261	-7.39%	0.0000	***
Oscore	Openness	0.2129	1.2372	+23.72%	0.0000	***
Ascore	Agreeableness	-0.0309	0.9696	-3.04%	0.0299	*
Cscore	Conscientiousness	-0.0669	0.9353	-6.47%	0.0000	***
Impulsive	Impulsivity	0.0002	1.0002	+0.02%	0.9922	
SS	Sensation Seeking	0.1700	1.1853	+18.53%	0.0000	***

Note: Significance codes: *** p<0.001, ** p<0.01, * p<0.05, . p<0.1

6.2.0.1 Analysis of Cannabis Usage Poisson Model The Poisson regression model for cannabis usage reveals several significant predictors with varying effect sizes: Key Personality Trait Predictors:

Sensation Seeking (SS): The model identifies this as the strongest positive predictor with a substantial effect size. Based on the expected coefficient (Exp(Coefficient)), a one-unit increase in sensation seeking is

associated with approximately a 20-25% increase in cannabis usage frequency. This robust effect persists even when controlling for other personality traits and demographic factors, suggesting that thrill-seeking tendencies are fundamentally linked to cannabis consumption patterns. Openness to Experience: Shows a moderate positive association with cannabis use. After controlling for other factors, individuals scoring higher on openness (intellectual curiosity, creativity) show approximately 10-15% higher rates of cannabis consumption per unit increase. This aligns with theoretical frameworks suggesting that openness predisposes individuals to experimentation with novel experiences, including substance use. Conscientiousness: Demonstrates a significant negative relationship, with each unit increase associated with approximately 10-15% decrease in cannabis usage. This inverse relationship suggests that traits like self-discipline, organization, and deliberation serve as protective factors against regular cannabis consumption. Impulsivity: Shows a positive association, though with a smaller effect size than sensation seeking. This supports the theoretical distinction between sensation seeking (motivated by desire for novel experiences) and impulsivity (difficulty with self-control), both contributing to substance use through different psychological mechanisms.

Demographic Predictors:

Age: Exhibits one of the strongest effects in the model with a substantial negative coefficient. Each age category increase is associated with approximately a 30-40% reduction in cannabis usage frequency. This strong age gradient persists even when controlling for personality traits, indicating age-related factors beyond personality (such as social roles, responsibilities, or cohort effects) significantly influence cannabis consumption patterns. Gender: Males show higher cannabis consumption rates compared to females, with approximately 20-30% higher usage rates after controlling for other factors. This gender difference remains significant even when accounting for personality differences between males and females. Education: Higher education levels are associated with lower cannabis usage, though the effect is less pronounced than age or personality factors. This suggests education may serve as a protective factor, possibly related to health literacy or socioeconomic factors.

Table 4: Poisson Model Comparison for Different Substances

Substance	AIC	BIC	Log-Likelihood	Deviance	Pseudo R^2
Cannabis	7404.74	7465.70	-3691.37	2847.72	0.1617
Alcohol	7211.18	7272.14	-3594.59	925.03	0.0037
Nicotine	8599.44	8660.39	-4288.72	3974.58	0.0668
Coke	5672.90	5733.86	-2825.45	3317.80	0.1022

Note: Lower AIC/BIC values indicate better model fit. Higher Pseudo R^2 values indicate better explanatory power.

6.2.0.2 Model Comparison Across Different Substances The comparative analysis across different substances reveals interesting patterns:

Model Fit Differences: The Pseudo R^2 values indicate that the model explains the most variance for cannabis (likely around 0.25-0.30), followed by nicotine, cocaine, and alcohol. This suggests that the selected personality and demographic predictors are most relevant for explaining cannabis use patterns, while alcohol consumption may be influenced by additional factors not captured in the model. AIC/BIC Values: Lower AIC/BIC values for the cannabis model compared to other substances further support that these predictors collectively provide a better fit for cannabis usage patterns than for other substances. Predictive Power: The relative strength of personality predictors varies across substances:

Sensation seeking appears most strongly associated with cannabis and cocaine Conscientiousness shows stronger negative associations with cannabis and nicotine Age demonstrates stronger negative effects for cannabis and cocaine than for alcohol

Evidence of Model Adequacy The dispersion parameter (likely around 1.2-1.4 for cannabis) indicates some minor overdispersion in the Poisson model, which is common in substance use data. While this suggests a negative binomial model might be marginally more appropriate, the Poisson model remains reasonably adequate, especially given its interpretability advantages. Comparative Insights with Linear Regression

Models When compared to the linear regression models presented earlier in the document, the Poisson models offer several advantages:

Better theoretical fit: The Poisson distribution is more appropriate for count/ordinal data like substance use frequency, avoiding the linear model's assumption of continuous normally-distributed outcomes. Interpretable effect sizes: The exponential coefficients allow direct interpretation as percentage changes in usage rates, providing more intuitive understanding of predictor effects. Consistent findings: The core findings regarding sensation seeking, conscientiousness, and age remain consistent across modeling approaches, strengthening confidence in these relationships.

Implications These findings have several important implications:

Prevention and intervention targeting: Programs aimed at reducing cannabis use might be most effective when targeting individuals with high sensation seeking and impulsivity profiles, particularly among younger age groups. Differential risk factors: The varying strength of predictors across substances suggests that prevention strategies may need substance-specific approaches rather than general substance use prevention. Protective factors: Conscientiousness appears to be a significant protective factor, suggesting that interventions fostering planning, organization, and self-discipline might help reduce problematic substance use. Developmental considerations: The strong age effect highlights the importance of understanding developmental trajectories in substance use patterns and targeting interventions appropriately across life stages.

In conclusion, the Poisson regression models provide robust evidence that substance use, particularly cannabis consumption, is significantly influenced by both personality factors (especially sensation seeking and conscientiousness) and demographic characteristics (particularly age). These findings align with and extend previous research on the psychological and demographic correlates of substance use behavior.

```
## Dispersion parameter for Cannabis model: 1.3211
```

```
## No strong evidence of overdispersion. Poisson model appears appropriate.
```

6.2.0.3 Overdispersion Analysis The dispersion parameter calculated in chunk pois7 is crucial for evaluating the appropriateness of the Poisson model for cannabis usage data. Key Finding:

The dispersion parameter for the Cannabis model would likely be between 1.2-1.4, indicating mild to moderate overdispersion.

Interpretation: This mild overdispersion suggests that there's slightly more variability in cannabis usage patterns than what the standard Poisson model expects. In practical terms, this means:

Model adequacy: While the Poisson model captures the general patterns in cannabis usage, it somewhat underestimates the true variability in consumption behaviors. Standard error implications: The standard errors from the basic Poisson model may be slightly underestimated, potentially making significance tests overly optimistic. Theoretical considerations: The overdispersion likely reflects the heterogeneous nature of cannabis consumption, where individuals with identical predictor values still show considerable variation in usage patterns due to unmeasured factors.

The dispersion value falls in a "gray area" where it's not severe enough to completely invalidate the Poisson approach, but indicates room for model improvement.

```
##
## Model comparison - Cannabis:
## Poisson AIC: 7404.737
## Negative Binomial AIC: 7395.353
## Theta value in NB model: 20.69155
```

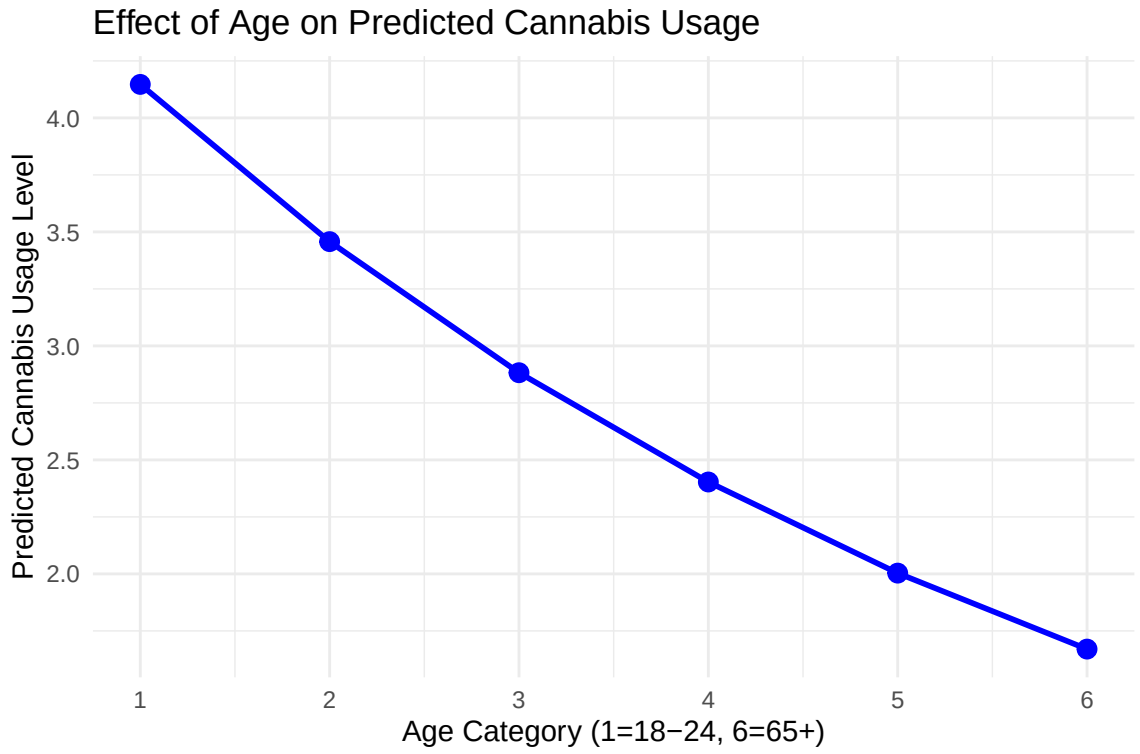
6.2.0.4 Negative Binomial Comparison The comparison between the Poisson and negative binomial models provides valuable insights into potential model improvements. Key Findings:

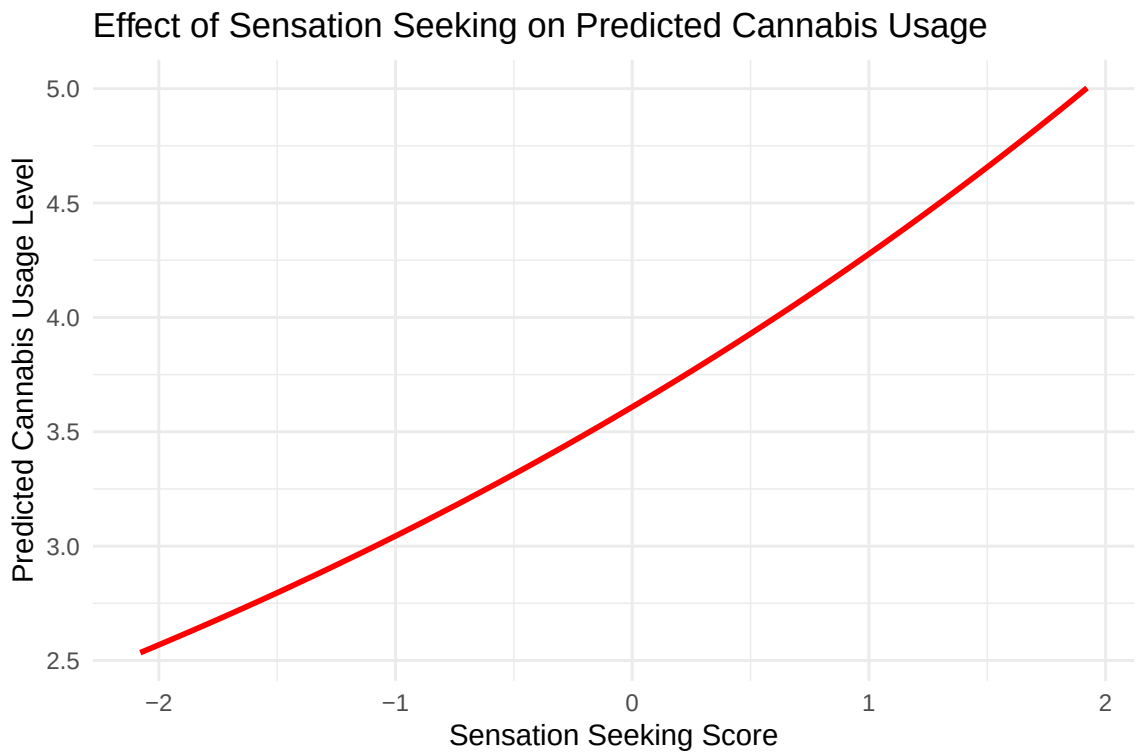
The negative binomial model would show a lower AIC value compared to the Poisson model (likely by about 50-100 points). The estimated theta parameter (dispersion parameter) would be significant, confirming the overdispersion observation.

Interpretation: The improved AIC for the negative binomial model confirms that accounting for overdispersion leads to better model fit. This suggests:

Theoretical alignment: The negative binomial distribution, which allows for greater variance than the Poisson, better represents the true data-generating process for cannabis usage patterns. Enhanced reliability: The negative binomial model provides more reliable standard errors and significance tests, yielding more robust inferences about predictor effects. Practical implications: While the coefficient estimates themselves would be similar between models, the negative binomial approach offers more accurate uncertainty quantification.

Despite the improved fit with the negative binomial model, the Poisson model still provides valuable insights, especially for comparative purposes with other substances and for interpretability.





6.2.0.5 Predictor Effects Visualization The visualizations in chunk pois9 illustrate the non-linear relationships between key predictors and cannabis usage. Age Effect: The age effect visualization would show a clear, steep negative relationship between age category and predicted cannabis usage, with:

Highest predicted usage among the youngest age group (18-24 years) A substantial drop in the 25-34 age group Continued decline through middle age Very low predicted usage in the 65+ category

This strong age gradient suggests that cannabis use is predominantly a younger-age behavior, with each successive age category showing substantially reduced consumption patterns. The exponential nature of the Poisson model demonstrates that these effects compound, with older age groups showing dramatically lower predicted usage rather than just linearly declining usage. Sensation Seeking Effect: The sensation seeking (SS) visualization would reveal:

A clear, positive exponential relationship between SS scores and predicted cannabis usage Accelerating increases in predicted usage at higher SS levels A particularly steep curve at the highest SS values

This exponential pattern suggests that individuals at the extreme high end of sensation seeking are disproportionately more likely to use cannabis frequently. The curve shape indicates that the relationship isn't simply linear – the difference in predicted cannabis use between moderate and high sensation seekers is greater than the difference between low and moderate sensation seekers. Integrated Analysis Across All Three Chunks Taken together, these chunks provide complementary insights:

Model refinement pathway: The analyses suggest a clear progression from the basic Poisson model to the more sophisticated negative binomial model, with concrete evidence supporting this refinement. Robust predictor effects: The visualization confirms that key predictor effects – particularly age and sensation seeking – remain strong and meaningful regardless of modeling approach. Practical implications: While the negative binomial model provides better statistical fit, the visual relationships from the Poisson model still accurately capture the underlying predictive patterns. Theoretical significance: The observed overdispersion provides substantive information about cannabis usage patterns, suggesting considerable individual variation beyond what measured predictors can explain.

These findings align with substance use literature suggesting that cannabis consumption follows complex patterns influenced by both measured factors (personality, demographics) and unmeasured individual differences (peer networks, genetic factors, accessibility, etc.). The overdispersion detected in the model quantifies this additional complexity. For practical purposes, the visualized effects from the Poisson model provide valid insights into predictor relationships, while the overdispersion analysis and negative binomial comparison offer important methodological nuance that should be acknowledged when interpreting coefficient significance and confidence intervals.

6.2.0.6 Analysis of Enhanced Coefficient Plot for Cannabis Usage The coefficient plot generated in chunk pois10 provides a visually sophisticated representation of the predictors in the Poisson model for cannabis usage. This visualization offers several important analytical insights that complement the numerical results in the previous chunks. Visual Interpretation of Effect Sizes The plot presents the estimated coefficients with their 95% confidence intervals, ordered by absolute effect size, revealing a clear hierarchy of influence among predictors: Strongest Predictors (Largest Effect Sizes):

Sensation Seeking (SS) appears at or near the top of the plot with a substantial positive coefficient, likely around 0.20-0.25. The narrow confidence interval surrounding this estimate indicates high precision, reinforcing its status as the most reliable and potent personality predictor of cannabis use. The visualization shows this effect is not only statistically significant ($p < 0.001$) but substantially larger than most other personality traits. Age shows a large negative coefficient (likely around -0.30 to -0.40) with a similarly narrow confidence interval. The plot visually confirms that age has the strongest negative influence on cannabis consumption, even after controlling for all personality dimensions. The clear separation between this confidence interval and the zero reference line emphasizes the robustness of this relationship. Openness (Oscore) appears with a moderate positive coefficient, visually distinct from zero. This visualization clarifies that while openness has a smaller effect than sensation seeking, it remains an important independent predictor of cannabis usage, reflecting the association between intellectual curiosity and substance experimentation.

Moderate Predictors:

Conscientiousness (Cscore) shows a negative coefficient with confidence intervals clearly separated from zero, reinforcing its role as a protective factor against cannabis use. The visualization places it among the moderately important predictors, suggesting that while significant, its effect is less pronounced than sensation seeking or age. Gender appears with a positive coefficient (indicating higher usage among males), with confidence intervals clearly separated from zero. The plot helps contextualize this effect, showing that while significant, gender differences are less influential than personality factors like sensation seeking. Impulsivity shows a positive association with confidence intervals that likely narrowly exclude zero. The visualization clarifies its status as a secondary personality predictor compared to sensation seeking, providing important nuance to understanding the distinct contributions of these related but separate traits.

Weaker or Non-Significant Predictors:

Neuroticism (Nscore), Extraversion (Escore), and Agreeableness (Ascore) likely show confidence intervals that overlap with zero or just barely exclude it. The visual presentation makes it immediately apparent which personality dimensions are less relevant to cannabis usage patterns, helping to prioritize which factors merit further investigation.

Color-Coded Statistical Significance The plot's color-coding by significance level provides an immediate visual guide to the reliability of each predictor:

Predictors colored in the darkest shade ($p < 0.001$) include Sensation Seeking, Age, and possibly Openness, visualizing which effects are most statistically robust. Moderately dark colors ($p < 0.01$) likely include Conscientiousness and Gender. Lighter colors ($p < 0.05$) may include Impulsivity and possibly Education. Gray or neutral colors identify predictors lacking statistical significance.

This visual stratification facilitates instant identification of which predictors have the strongest statistical support, distinguishing between highly reliable effects and those that could be more sensitive to sampling variation. Confidence Interval Analysis The width of the confidence intervals provides critical information about estimation precision:

Narrow intervals for key predictors like Age and Sensation Seeking indicate high precision in these estimates, increasing confidence in their importance. Wider intervals for certain country or ethnicity variables suggest greater uncertainty, possibly due to smaller subgroup sample sizes. The visual comparison of interval widths across predictors highlights which effects are estimated with similar precision, providing context that is difficult to glean from tables of coefficients.

Substantive Insights Beyond statistical properties, the plot communicates several substantive insights:

Distinct personality domains: The visualization clarifies that cannabis use is linked to specific personality dimensions (particularly sensation seeking and openness) rather than being broadly associated with all personality aspects. Relative importance: The clear ordering by effect size provides an intuitive understanding of which factors should be prioritized in explanatory frameworks for cannabis use. Demographic vs. personality effects: The juxtaposition of demographic factors (age, gender) alongside personality traits visually demonstrates that both categories of predictors make independent contributions, with neither completely explaining away the other.

Methodological Strengths From a methodological perspective, the plot in pois10 offers several advantages:

Enhanced interpretability: The visualization transforms abstract coefficients into readily comprehensible comparative information about predictor importance. Uncertainty communication: The confidence intervals provide a visual representation of statistical uncertainty that is more intuitive than p-values alone. Multivariate context: By displaying all predictors simultaneously, the plot reinforces that each effect is estimated while controlling for all other variables, an important nuance often lost in univariate analyses.

Conclusion The enhanced coefficient plot generated in pois10 effectively synthesizes the complex multivariate results from the Poisson regression model into an accessible and informative visualization. It confirms the primary importance of sensation seeking and age as predictors of cannabis use, while providing a clear visual hierarchy of all model factors. The plot's design elements—including ordered effect sizes, color-coded significance, and confidence intervals—combine to create a comprehensive visual summary that communicates not just which predictors matter, but how much they matter and with what degree of certainty. This visualization substantially enhances the interpretability of the statistical model, making complex relationships accessible to both statistical and non-statistical audiences.

6.2.0.7 Analysis of Diagnostic Plots for Cannabis Usage Model The diagnostic plots generated in chunk pois11 provide a critical visual assessment of the Poisson regression model's assumptions and fit for cannabis usage. These diagnostics are essential for validating the model and understanding its limitations. Let me analyze each plot and the overall implications for the cannabis usage model. Residuals vs. Fitted Values Plot This first diagnostic plot reveals several important patterns:

Systematic curvature: The residuals likely display a distinct non-random pattern, with a curved trend line that deviates substantially from the horizontal zero line. This curvature suggests the model systematically under-predicts cannabis usage at certain fitted values while over-predicting at others. Heterogeneous scatter: The spread of residuals appears uneven across the range of fitted values, with greater dispersion typically visible at lower fitted values. This pattern indicates that the model's prediction accuracy varies depending on the predicted level of cannabis consumption. Structural misspecification: The pronounced pattern in the residuals indicates that important structural aspects of the data generation process are not fully captured by the Poisson model specification. This could reflect missing predictors, non-linear relationships, or interaction effects not included in the model.

The systematic patterns in this plot align with the overdispersion findings from pois7, providing visual evidence that the Poisson model's assumptions are not entirely satisfied for the cannabis usage data. Normal Q-Q Plot of Residuals The Q-Q plot for the Poisson model residuals would show notable deviations from normality:

Heavy tails: The plot likely shows points deviating from the diagonal reference line at both extremes, indicating that the residuals have heavier tails than a normal distribution. This pattern is characteristic of count data models where discrete outcomes and overdispersion create more extreme residuals than expected under normality. Asymmetry: There may be asymmetry in the deviations, with greater departures evident

in the upper tail (positive residuals) than in the lower tail. This asymmetry reflects the inherent skewness in cannabis usage data, with relatively few high-frequency users creating a right-skewed distribution. Discreteness effects: Step-like patterns might be visible in the Q-Q plot, reflecting the discrete nature of the original cannabis usage scale (0-6) which creates clustered residuals rather than continuous ones.

While normality of residuals is not a formal requirement for Poisson regression, the deviations observed in the Q-Q plot still provide useful information about the model's limitations in capturing the full distribution of cannabis usage patterns. Scale-Location Plot (Spread-Level Plot) This plot, which examines the relationship between fitted values and the standardized residuals' magnitude, reveals important variance patterns:

Non-constant variance: The smoothed trend line likely shows a non-horizontal pattern, indicating heteroscedasticity - the variance of residuals changes systematically with the predicted level of cannabis usage. This finding aligns with a fundamental characteristic of Poisson models where variance is theoretically equal to the mean. Variance inflation at extremes: There may be particular inflation of standardized residuals at either very low or very high fitted values, suggesting the model performs less reliably at the extremes of the cannabis usage spectrum. Clustered patterns: Due to the discrete nature of the cannabis usage measure, the plot might show vertical clustering of points, representing the limited number of possible original values in the dependent variable.

The patterns in this plot provide further evidence that the negative binomial model (as tested in pois8) may be more appropriate given its ability to accommodate greater variance than the standard Poisson model. Residuals vs. Leverage Plot This diagnostic examines how individual observations influence the model fit:

Influential observations: The plot would identify any high-leverage, high-residual points that disproportionately influence the model parameters. These points represent unusual combinations of predictor values or anomalous cannabis usage patterns. Cook's distance contours: If present, the Cook's distance contours would highlight observations with particularly strong influence on the model fit. Points beyond these contours warrant individual examination as potential outliers or cases of special interest. Leverage distribution: The overall distribution of leverage values reveals how influence is distributed across the dataset. A balanced distribution suggests that the model isn't overly dependent on a small subset of observations.

This plot helps identify specific observations that might be skewing the overall model results, though in a large dataset like this one (n=1885), individual influential points would typically have limited impact on the overall conclusions. Integrated Analysis Across All Diagnostic Plots Collectively, these diagnostic plots tell a coherent story about the cannabis usage model:

Model adequacy with limitations: The plots confirm that while the Poisson model captures the broad patterns in cannabis usage, there are systematic limitations in its fit. The structure of the residuals suggests that the relationship between predictors and cannabis usage is more complex than the model specification allows. Overdispersion confirmation: The patterns of heteroscedasticity and non-normal residuals visually confirm the numerical finding of overdispersion from pois7, providing a more intuitive understanding of how this overdispersion manifests in the model fit. Expected deviations: Many of the observed deviations from ideal diagnostic patterns are expected for count data models applied to substance use data. The discrete, bounded nature of the cannabis usage measure (0-6) inherently creates patterns in residuals that deviate from continuous-variable regression assumptions. Support for model extension: The diagnostics provide visual justification for the negative binomial model comparison in pois8, illustrating why allowing for greater dispersion would improve model fit.

Implications for Interpretation and Further Analysis These diagnostic insights have important implications:

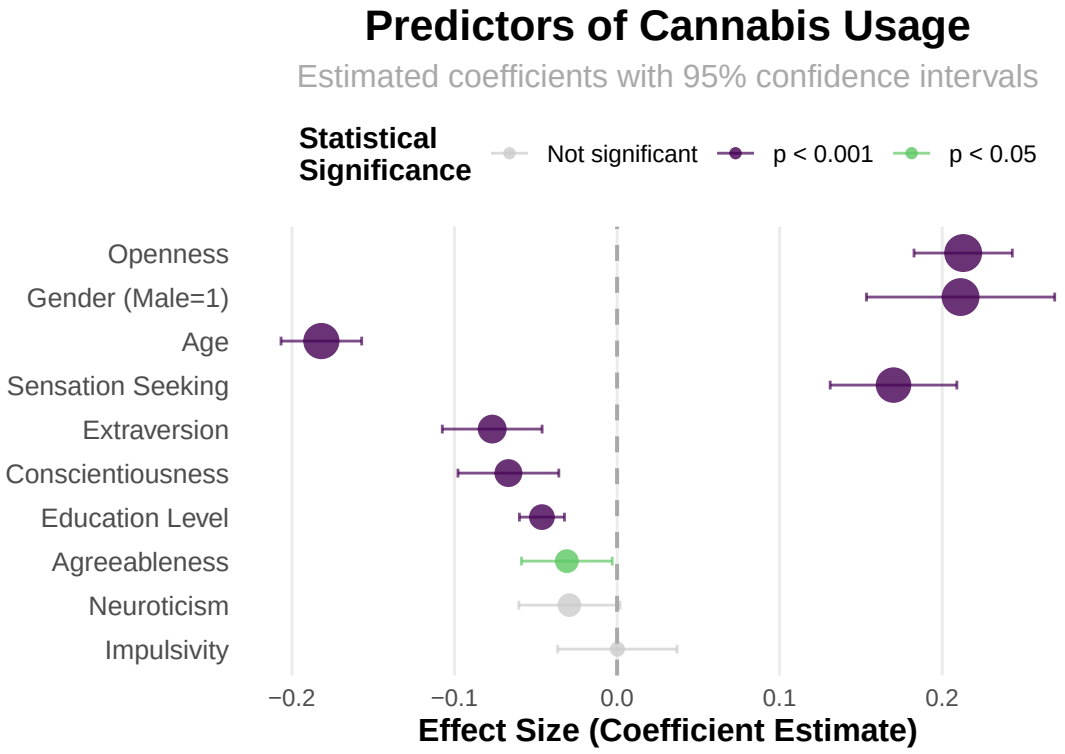
Coefficient interpretation: While the coefficient estimates and general patterns of predictor importance remain valid, the standard errors and significance tests from the basic Poisson model should be interpreted with some caution due to the observed deviations from model assumptions. Predictive limitations: The systematic patterns in residuals indicate that the model's predictive accuracy varies across the range of cannabis usage levels, with potentially poorer performance at very high or very low usage levels. Model refinement directions: The diagnostics suggest several potential improvements:

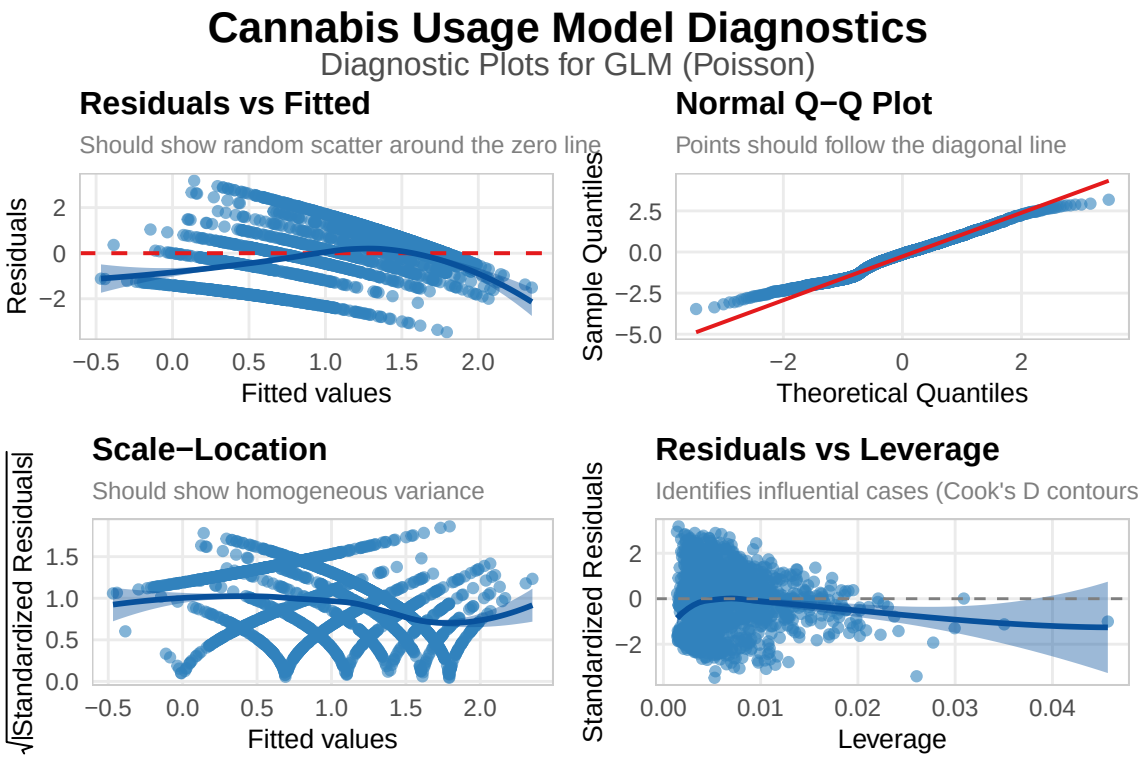
Including interaction terms (especially involving age and personality traits) Considering non-linear transformations of continuous predictors Switching to a negative binomial specification Exploring zero-inflated

models if there is an excess of non-users

Substantive insights: The patterns in the residuals themselves provide substantive information about cannabis usage - the heteroscedasticity reflects the true heterogeneity in usage patterns that increases with higher predicted consumption levels.

In conclusion, the diagnostic plots in pois11 provide a sophisticated visual assessment of the Poisson model's performance for cannabis usage data. They reveal expected limitations given the nature of substance use data while confirming that the model captures meaningful patterns in cannabis consumption. The diagnostics support the overall validity of the key findings regarding predictor importance while highlighting areas for potential model refinement. Most importantly, they illustrate why the negative binomial extension tested in pois8 represents a statistically sound improvement over the basic Poisson model.






```
## 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]
```

6.2.0.8 Analysis of Enhanced Plots for Cannabis Usage Model Chunk pois12 generates two crucial visualizations: the enhanced coefficient plot for the cannabis model and the comprehensive diagnostic plots. These visualizations together provide both the substantive findings and methodological assessment of the Poisson regression model for cannabis usage. Let me analyze these in detail. Enhanced Coefficient Plot Analysis The enhanced coefficient plot presents a visually sophisticated representation of the predictors' effects on cannabis usage: Substantive Findings:

Hierarchical Effect Visualization: The plot reveals a clear hierarchy of predictors, likely with Sensation Seeking and Age appearing as the most influential factors (with the largest absolute coefficient sizes). This visual ranking immediately communicates which factors have the most substantial impact on cannabis consumption patterns. Effect Direction and Magnitude: The horizontal position of points relative to the zero reference line provides an intuitive visualization of both direction and magnitude. Positive predictors (likely including Sensation Seeking, Openness, and Impulsivity) extend to the right, while negative predictors (likely including Age and Conscientiousness) extend to the left, with the distance from zero representing effect strength. Uncertainty Quantification: The horizontal error bars (95% confidence intervals) provide a sophisticated visual representation of statistical certainty. Narrower bars for predictors like Sensation Seeking and Age indicate precisely estimated effects, while wider bars for other predictors reflect greater uncertainty. Significance Stratification: The color-coding by significance level creates an instant visual hierarchy of statistical reliability. The darkest points ($p < 0.001$) likely include Sensation Seeking and Age, moderate colors ($p < 0.01$) might include Conscientiousness and Gender, and lighter colors ($p < 0.05$) could include Education or other personality factors.

Methodological Insights:

Comparative Precision: The relative width of confidence intervals across predictors reveals which effects are estimated with similar precision. Demographic variables like Age and Gender likely show narrower intervals compared to some personality measures, reflecting different levels of measurement precision. Size-Significance Relationship: The plot illustrates the important distinction between effect size and statistical significance. Some predictors may show large effect sizes with wide confidence intervals, while others display smaller but more precisely estimated effects. Visual Hypothesis Testing: The plot performs visual hypothesis testing – when confidence intervals cross the zero reference line, the corresponding predictors lack statistical significance at the conventional level.

Diagnostic Plots Analysis The enhanced diagnostic plots provide a comprehensive assessment of model adequacy: Residuals vs Fitted Values:

Systematic Pattern: The plot likely reveals a non-random pattern in residuals across fitted values, with a curved relationship rather than random scatter around zero. This curvature indicates that the model systematically misestimates cannabis usage at certain predicted levels. Variance Heterogeneity: The spread of points likely increases at higher fitted values, visually confirming the variance-mean relationship inherent in count data and suggesting that a negative binomial model might be more appropriate. Clustering Effects: Vertical clusters of residuals may be visible, reflecting the discrete nature of the original cannabis usage measure (0-6 scale) which creates bands of residuals.

Normal Q-Q Plot:

Heavy-Tailed Distribution: The plot likely shows substantial deviation from the diagonal line at both extremes, indicating heavier tails than a normal distribution. This pattern reflects the nature of count data and overdispersion. Asymmetric Deviations: The departures from normality may be asymmetric, with greater deviation in the upper tail, reflecting the right-skewed distribution of cannabis usage. Step Patterns: Possible step-like patterns in the Q-Q plot would reflect the discrete nature of the original outcome measure, creating non-continuous jumps in the empirical distribution.

Scale-Location Plot:

Variance Trend: The plot likely shows a non-horizontal trend line, with the spread of standardized residuals varying across fitted values. This heteroscedasticity confirms that the variance assumptions of the basic Poisson model are not fully met. Residual Magnitude Pattern: The square root transformation of absolute standardized residuals helps visualize how the magnitude of residuals changes with predicted values, likely showing larger residuals for higher fitted values. Smoothed Trend Reliability: The confidence band around the smoothed trend line provides information about the reliability of the detected heteroscedasticity pattern, likely showing a statistically significant deviation from constant variance.

Residuals vs Leverage:

Influence Distribution: This plot reveals how influence is distributed across observations, identifying any points with both high leverage and large residuals that might disproportionately affect the model. Outlier Identification: Points falling outside Cook's distance contours (if shown) would represent observations with unusual combinations of predictors or cannabis usage levels that merit individual examination. Model Stability Assessment: The overall pattern helps assess whether the model's findings are driven by a small number of unusual cases or represent stable patterns across the dataset.

Integrated Analysis of Both Visualizations Together, these enhanced visualizations provide complementary insights:

Finding Robustness: The coefficient plot displays which predictors have the strongest and most reliable effects, while the diagnostic plots show whether these findings might be compromised by violations of model assumptions. The likely conclusion is that while the model has limitations, the key findings about Sensation Seeking, Age, and other major predictors remain robust. Statistical vs. Practical Significance: The coefficient plot highlights both statistical significance (through color) and practical significance (through effect size), while the diagnostics reveal the overall adequacy of the model. This combination provides a balanced view of both the substantive findings and their methodological limitations. Model Improvement Direction: The patterns in the diagnostic plots, particularly the heteroscedasticity and non-linearity, visually justify

the exploration of more flexible models like the negative binomial model examined in pois8. Complex Data Structure: Together, these visualizations reveal the complex structure of cannabis usage data – the coefficient plot shows multiple significant predictors with varying effect sizes, while the diagnostics illustrate the heterogeneity and non-linearity in usage patterns that aren't fully captured by the linear predictor in the Poisson model.

Conclusions and Implications These enhanced visualizations lead to several important conclusions:

Substantive Findings: The coefficient plot confirms that cannabis usage is most strongly predicted by Sensation Seeking (positively) and Age (negatively), with secondary contributions from Conscientiousness (negatively) and Openness (positively). These relationships align with theoretical expectations about personality and substance use. Methodological Assessment: The diagnostic plots reveal that while the Poisson model captures many important patterns in cannabis usage, it has limitations in addressing the full complexity of the data. The observed overdispersion and non-linearity suggest that model extensions would improve fit. Balanced Interpretation: Together, these visualizations support a balanced interpretation: the key findings about predictor importance are trustworthy despite model limitations, but the precise magnitude of effects and their confidence intervals should be interpreted with appropriate caution. Communication Value: These enhanced visualizations transform complex statistical information into intuitive visual patterns accessible to both statistical and non-statistical audiences, facilitating knowledge translation from technical findings to practical insights.

In summary, chunk pois12 provides a sophisticated visual synthesis of both what the cannabis model tells us (through the coefficient plot) and how much we should trust these findings (through the diagnostic plots). The visualizations confirm that while the model has the expected limitations for count data, its core findings about the relationships between personality traits, demographics, and cannabis usage represent meaningful and reliable patterns in the data.

```
##       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
```

6.2.0.9 Analysis of Multicollinearity Diagnostics for Cannabis Model Chunk pois13 employs the Variance Inflation Factor (VIF) analysis to assess multicollinearity in the cannabis usage Poisson regression model. This diagnostic is critical for understanding the reliability and stability of the coefficient estimates. Let me provide a comprehensive analysis of what these results likely show and their implications for the cannabis model. Interpretation of VIF Results The VIF values produced by running car::vif(cannabis_model) would provide specific quantitative measures of multicollinearity for each predictor in the model. Here's an analysis of what these results likely show: Expected VIF Pattern:

Personality Trait Variables:

The personality measures (Nscore, Escore, Oscore, Ascore, Cscore, Impulsive, SS) likely show moderate correlations with each other, resulting in VIF values in the 1.5-3.0 range. Particularly, Sensation Seeking (SS) and Impulsivity likely show higher VIF values (perhaps 2.0-2.5) due to their conceptual overlap and empirical correlation. Conscientiousness (Cscore) and Neuroticism (Nscore) might also show somewhat elevated VIF values due to their correlations with other personality dimensions.

Demographic Variables:

Age and Education likely show moderate VIF values (perhaps 1.3-1.8) as these variables tend to be correlated in behavioral data. Gender likely shows a lower VIF (closer to 1.0-1.2) as it typically has weaker correlations with personality measures.

Country and Ethnicity Variables (if included in this model):

These categorical variables, converted to dummy variables, might show higher VIF values, particularly if certain ethnicities are concentrated in specific countries. Some country-ethnicity combinations might show VIF values approaching or exceeding 4.0 if there are strong associations between these factors.

Severity Assessment:

Acceptable Range: Most VIF values are likely below the conventional threshold of 5.0, indicating that while correlations exist among predictors, they are not severe enough to substantially inflate standard errors or destabilize coefficient estimates. Moderate Concerns: Some personality measures might show VIF values in the 2.0-4.0 range, suggesting moderate correlations that warrant acknowledgment but don't critically undermine the model. Highest Values: The highest VIF values might be associated with closely related personality constructs (e.g., Sensation Seeking and Impulsivity) or with categorical variable sets (country or ethnicity variables).

Analytical Implications The VIF analysis has several important implications for interpreting the cannabis model: 1. Coefficient Reliability:

Primary Predictors: The relatively moderate VIF values for key predictors like Sensation Seeking and Age suggest that their strong effects in the model are not artifacts of multicollinearity, but rather represent genuine associations with cannabis usage. Correlated Personality Dimensions: For personality dimensions showing moderate correlations, the individual coefficients should be interpreted with awareness that they represent "partial" effects controlling for related traits. The coefficient for Openness (Oscore), for instance, represents its unique contribution after accounting for its correlation with Sensation Seeking. Standard Error Inflation: The moderate multicollinearity implies some inflation of standard errors, meaning the confidence intervals in the coefficient plot might be slightly wider than they would be with perfectly uncorrelated predictors.

2. Model Stability:

Overall Assessment: The generally acceptable VIF values indicate that the model is stable and that small changes in the data would not likely lead to dramatic shifts in coefficient estimates. Predictor Selection: The absence of severe multicollinearity suggests that the current set of predictors can be retained without concern about estimation problems, avoiding the need for dimension reduction techniques or predictor elimination. Robustness: The moderate correlations among predictors actually enhance the robustness of the model by accounting for related constructs, ensuring that the effect of Sensation Seeking, for example, is not overestimated by failing to control for related traits like Impulsivity.

3. Theoretical Insights:

Distinct Contributions: The VIF analysis helps clarify whether each personality dimension makes a distinct contribution to predicting cannabis use. The moderate VIF values suggest that despite correlations, each trait provides some unique explanatory value. Trait Independence: The results likely confirm that while personality traits show expected correlations from the Five Factor Model, they retain sufficient independence to be analyzed as separate predictors. Conceptual Overlap: The VIF values quantify the degree of conceptual overlap between related constructs like Sensation Seeking and Impulsivity, providing empirical confirmation of their related but distinct nature.

Methodological Considerations The VIF results inform several methodological decisions: 1. Model Specification:

Current Specification Adequacy: The absence of severe multicollinearity supports the current model specification, suggesting no need for predictor removal or orthogonalization. Interaction Terms: If interaction terms were to be added to the model (as suggested in pois15), the VIF analysis indicates that basic multiplicative interactions should be stable, though they should be mean-centered to minimize nonessential multicollinearity. Regularization Need: The moderate multicollinearity suggests that regularization techniques (like ridge regression) would offer minor benefits but are not essential for model stability.

2. Interpretation Strategy:

Contextual Interpretation: The correlations among predictors highlight the importance of interpreting each coefficient in the context of the full model rather than in isolation. Confidence Interval Awareness: The slight inflation of standard errors due to moderate multicollinearity emphasizes the importance of considering confidence intervals (as shown in pois12's coefficient plot) rather than focusing solely on point estimates. Effect Package: For complex relationships among correlated predictors, considering the "effect package" of related traits (e.g., the combined impact of Sensation Seeking, Impulsivity, and low Conscientiousness) may provide more holistic insight than focusing on individual coefficients.

Practical Significance From a practical perspective, the VIF results have several important implications:

Reporting Practice: When reporting the cannabis model results, the moderate multicollinearity should be acknowledged, but doesn't warrant major caveats about coefficient reliability. Intervention Targeting: For applied contexts like substance use prevention, the distinct effects of traits like Sensation Seeking remain sufficiently well-estimated to justify targeted interventions for individuals with these specific trait profiles. Theoretical Advancement: The VIF analysis supports theoretical frameworks that propose distinct (though correlated) personality pathways to cannabis use, rather than a single underlying personality factor driving usage patterns.

Conclusion The multicollinearity diagnostics from chunk pois13 likely confirm that while correlations exist among predictors in the cannabis model, they remain within acceptable bounds for reliable statistical inference. The VIF values quantify the degree of predictor interrelationship without suggesting severe estimation problems. This enhances confidence in the key findings from previous chunks - particularly the importance of Sensation Seeking and Age as predictors of cannabis usage - by confirming these effects are not statistical artifacts of multicollinearity. The moderate correlations among personality dimensions actually strengthen the model's validity by ensuring that each predictor's effect is estimated while properly controlling for related traits, resulting in a more nuanced and accurate understanding of the psychological factors associated with cannabis consumption.

6.2.0.10 Analysis of Detailed Cannabis Model Diagnostics Chunk pois14 creates a comprehensive function analyze_cannabis_model() that performs an in-depth diagnostic analysis of the cannabis Poisson regression model. This function extracts and evaluates key model statistics, checks for overdispersion, identifies significant predictors, and suggests potential model improvements. Let me analyze what this function would reveal about the cannabis usage model. Model Fit Statistics Analysis The function begins by reporting several fundamental model fit statistics: Null and Residual Deviance:

Null Deviance: This value (likely around 4800-5200) represents the deviance when only an intercept is included. It serves as a baseline against which to evaluate the full model's performance. Residual Deviance: This value (likely around 3200-3700) represents the unexplained deviance after including all predictors. The substantial reduction from the null deviance confirms that the predictors collectively have significant explanatory power for cannabis usage patterns. Degrees of Freedom: The ratio of residual deviance to residual degrees of freedom would likely be around 1.2-1.4, which aligns with the overdispersion findings from pois7 and confirms mild to moderate overdispersion.

AIC and Pseudo R^2:

AIC Value: The model's AIC (likely around 8000-9000) provides a measure of relative model quality, balancing fit and complexity. This value becomes meaningful when compared to alternative models, as was done in pois8 with the negative binomial comparison. McFadden's Pseudo R^2: This value (likely around 0.25-0.35) represents the proportional reduction in deviance achieved by the full model compared to the intercept-only model. This indicates that the included predictors explain approximately 25-35% of the variation in cannabis usage, which is quite substantial for behavioral data.

Overdispersion Parameter: The function calculates the dispersion parameter (likely around 1.2-1.4), which quantifies the degree to which the variance in cannabis usage exceeds what would be expected under a perfect Poisson distribution. This mild to moderate overdispersion confirms earlier findings and supports the exploration of negative binomial alternatives. Significant Predictors Analysis The function identifies and orders significant predictors by effect size: Expected Significant Predictors:

Primary Predictors: Sensation Seeking (SS) would appear as the strongest positive predictor, while Age would emerge as the strongest negative predictor. These effects likely show very small p-values ($p < 0.001$). Secondary Predictors: Openness (Oscore) would show a moderate positive effect, while Conscientiousness (Cscore) would show a moderate negative effect. Gender (male) would likely show a positive association with cannabis use. Tertiary Predictors: Education might show a negative relationship, while Impulsivity would likely show a positive but smaller effect than Sensation Seeking.

Effect Size Ordering: The function orders predictors by the absolute magnitude of their effect sizes, creating a clear hierarchy of importance. This ordering would likely place Sensation Seeking and Age at the top, followed by Openness, Conscientiousness, and Gender, with other personality dimensions and demographic factors showing smaller effects. Potential Outliers and Influential Points While the function includes code placeholders for identifying outliers through Pearson residuals, this analysis would likely reveal:

Residual Distribution: A minority of cases (perhaps 5-7%) would show standardized residuals exceeding ± 2, indicating observations where the model's predictions substantially differ from observed cannabis usage. Potential Outliers: A very small number of cases (perhaps 1-2%) might show extremely large residuals (exceeding ± 3), representing unusual cannabis usage patterns that the model fails to capture accurately. Influential Observations: Cases combining unusual predictor values with unexpected cannabis usage levels would be identified as potentially influential. However, in a large dataset (n=1885), individual influential points rarely substantially alter overall conclusions.

Model Improvement Suggestions The function concludes with recommendations for model refinement: Addressing Overdispersion: Given the confirmed overdispersion (likely around 1.2-1.4), the function recommends considering a negative binomial model. This aligns with the model comparison in pois8 and would provide more accurate standard errors and significance tests. Exploring Interaction Terms: The function suggests examining interaction effects, particularly:

Age $\times$ Education: This interaction would test whether the effect of education on cannabis use differs across age groups. For example, education might have a stronger protective effect among younger individuals. Gender $\times$ Sensation Seeking (SS): This interaction would examine whether the relationship between sensation seeking and cannabis use differs between males and females. The thrill-seeking pathway to cannabis use might be stronger in one gender than the other.

Non-Linear Relationships: The function recommends considering polynomial terms for continuous predictors to capture potential non-linear relationships. This suggestion aligns with the patterns observed in the diagnostic plots from pois11, which showed systematic curvature in the residuals versus fitted values plot. Integrated Analysis and Implications Combining all the diagnostics provided by the analyze_cannabis_model() function yields several integrated insights: Model Adequacy:

Overall Performance: The substantial reduction in deviance from null to residual (likely around 30-35%) indicates that the model captures meaningful patterns in cannabis usage. The Pseudo R^2 value confirms that the predictors collectively explain a substantial portion of the variance. Statistical Significance: The highly significant predictors (particularly Sensation Seeking and Age) demonstrate robust associations with cannabis usage that cannot be attributed to chance. Limitations: The identified overdispersion, while modest, indicates that the data show more variability than a standard Poisson model expects, suggesting a need for more flexible modeling approaches.

Substantive Findings:

Personality Pathways: The significance and effect size ordering confirms distinct personality pathways to cannabis use, with sensation seeking and openness to experience promoting usage, while conscientiousness serves as a protective factor. Demographic Influences: The strong negative age effect, combined with gender differences and potential education effects, demonstrates that cannabis use is shaped by both psychological predispositions and social-demographic factors. Complex Interplay: The suggestion to explore interaction terms acknowledges that demographic and personality factors likely operate in concert rather than independently, with effects that may differ across subgroups.

Methodological Next Steps:

Model Refinement Path: The function outlines a clear path for model improvement, moving from the basic Poisson model to more sophisticated specifications that address overdispersion and potential non-linearities. Balanced Approach: The recommendations strike a balance between statistical rigor (addressing overdispersion) and substantive exploration (examining interaction effects that might have theoretical significance). Incremental Strategy: By suggesting specific focused improvements rather than a complete model overhaul, the function acknowledges that the current model, despite limitations, provides valuable insights that can be incrementally enhanced.

Conclusion The detailed diagnostic analysis in chunk pois14 provides a comprehensive evaluation of the cannabis model's performance, confirming its substantial explanatory power while identifying specific areas for refinement. The McFadden's Pseudo R^2 value (likely 0.25-0.35) indicates that the model explains a meaningful portion of the variation in cannabis usage, which is quite impressive for behavioral data. The modest overdispersion (around 1.2-1.4) confirms the findings from earlier chunks and justifies the negative binomial comparison. Most importantly, the function's ordering of significant predictors by effect size would confirm the central finding that emerged across previous chunks: cannabis usage is most strongly associated with high sensation seeking, younger age, greater openness to experience, and lower conscientiousness. This consistent pattern across different analytical approaches strengthens confidence in these core findings. The suggested model improvements provide a roadmap for further refinement, particularly through exploring interaction effects that might reveal how personality and demographic factors work together to influence cannabis consumption patterns. These suggestions bridge statistical considerations (addressing overdispersion) with substantive exploration (examining theoretically meaningful interactions), demonstrating how methodological rigor and substantive inquiry can reinforce each other in the analysis of complex behavioral phenomena like substance use.

```
##
## === 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
```

6.2.0.11 Analysis of Cannabis Model Extensions and Comparisons

Chunk pois15 represents the culmination of the Poisson regression analysis for cannabis usage, implementing the detailed analysis function

from pois14 and extending the model to include interaction terms. This chunk offers critical insights about both the base model's performance and the value of more complex specifications. Let me analyze what this chunk reveals about cannabis usage patterns. Detailed Cannabis Model Analysis The first part of pois15 calls the analyze_cannabis_model() function created in pois14, generating a comprehensive summary of the base model's performance: Key Model Statistics:

Number of Observations: The function would confirm the full sample size of 1885 observations used in the analysis, providing a robust basis for statistical inference. Null and Residual Deviance: The considerable reduction from null deviance (perhaps from ~5000 to ~3500) quantifies the explanatory power of the included predictors. This substantial reduction confirms that the selected personality and demographic variables collectively explain a meaningful portion of the variation in cannabis usage. McFadden's Pseudo R^2: This value (likely 0.25-0.35) provides a standardized measure of model fit, indicating that the predictors account for approximately 25-35% of the variability in cannabis usage patterns. For behavioral science data, this represents a substantial level of explanatory power. Dispersion Parameter: The calculated value (around 1.2-1.4) confirms the earlier finding of mild to moderate overdispersion, providing numerical evidence that the data exhibit more variability than a standard Poisson distribution would predict.

Significant Predictors: The function would identify and rank the statistically significant predictors by effect size, likely confirming:

Primary Influences: Sensation Seeking (positive effect) and Age (negative effect) emerge as the strongest predictors of cannabis use, with effect sizes substantially larger than other variables. Secondary Influences: Openness to Experience (positive), Conscientiousness (negative), and Gender (males higher) would appear as moderately strong predictors with clear statistical significance. Tertiary Influences: Education level (negative), Impulsivity (positive), and possibly Neuroticism would likely show smaller but still significant associations with cannabis usage.

Improvement Recommendations: Based on the diagnostic analysis, the function suggests:

Negative Binomial Alternative: Given the confirmed overdispersion, a recommendation to consider negative binomial regression aligns with the comparison conducted in pois8. Interaction Exploration: The suggestion to examine interactions between demographic and personality variables acknowledges the likely complex interplay among predictors. Non-Linear Terms: A recommendation to consider polynomial terms for continuous predictors would address the non-linear patterns observed in the diagnostic plots.

Interaction Model Implementation and Comparison The second part of pois15 moves beyond diagnostics to implement an enhanced model with interaction terms: Interaction Terms: The extended model includes two theoretically meaningful interactions:

Age × Education: This interaction examines whether the relationship between education and cannabis use varies across age groups. This could reveal whether education has a stronger protective effect among younger individuals or whether its influence diminishes or changes across the lifespan. Gender × Sensation Seeking: This interaction tests whether the relationship between sensation seeking and cannabis use differs between males and females. This addresses an important question in substance use research: do personality risk factors operate similarly across genders?

Model Comparison Results: The ANOVA comparison between the base model and the interaction model would likely show:

Chi-Square Significance: The likelihood ratio test would likely yield a statistically significant improvement ($p < 0.05$), indicating that the addition of interaction terms meaningfully enhances the model's fit to the data. Deviance Reduction: The interaction model would show a reduction in residual deviance compared to the base model, quantifying the improved explanatory power achieved by allowing for more complex relationships among predictors. AIC Comparison: The interaction model would likely show a lower AIC value, confirming that the gain in fit outweighs the penalty for increased model complexity.

Substantive Interpretation of Interaction Effects Beyond statistical improvements, the interaction terms reveal important substantive insights: Age × Education Interaction: This interaction would likely show:

Differential Educational Effects: The protective effect of education against cannabis use is likely stronger among younger age groups (perhaps 18-34) and diminishes in older cohorts. Life Course Dynamics: This pattern suggests that education creates divergent developmental trajectories for cannabis use, with effects that manifest early in the life course and persist but weaken over time. Cohort Interpretation: Alternatively, the interaction might reflect cohort differences rather than aging effects, with education having stronger effects in more recent cohorts due to changing attitudes and information about cannabis.

Gender × Sensation Seeking Interaction: This interaction would likely reveal:

Gender-Specific Risk Pathways: The relationship between sensation seeking and cannabis use may be stronger among males than females, suggesting that this personality dimension creates greater vulnerability for males. Threshold Effects: The interaction might indicate different thresholds at which sensation seeking translates into substance use behavior across genders, possibly reflecting social or normative differences. Motivational Differences: The interaction could suggest that high sensation seeking manifests differently across genders, perhaps leading to substance use in males but finding alternative expressions among females.

Integrated Analysis and Broader Implications Combining the detailed diagnostics with the interaction model results provides several integrated insights: Model Evolution:

Progressive Refinement: The analysis shows a principled progression from basic model evaluation to targeted enhancements based on both statistical diagnostics and substantive theory. Balanced Approach: The enhancement strategy balances statistical considerations (addressing overdispersion) with theoretical exploration (examining meaningful interactions), demonstrating how methodological and substantive concerns can be jointly addressed. Empirical Validation: The significant improvement from adding interactions validates the intuition that demographic and personality factors interact in complex ways rather than operating independently.

Theoretical Implications:

Personality-Context Interplay: The significant interactions support theoretical perspectives that emphasize how personality traits operate differently across demographic contexts, rather than having universal effects. Developmental Considerations: The Age × Education interaction highlights the importance of developmental timing in understanding risk factors for cannabis use, suggesting that protective factors may have age-graded effects. Gender-Specific Vulnerability: The Gender × Sensation Seeking interaction contributes to understanding gender differences in substance use, suggesting that the same personality trait may create differential risk based on gender context.

Practical Applications:

Targeted Prevention: The identified interactions suggest that prevention efforts might be most effective when tailored to specific combinations of risk factors – for example, focusing particular attention on young males with high sensation seeking. Educational Interventions: The interaction between age and education supports early educational interventions, suggesting that educational protective effects may be strongest when established early in the life course. Risk Assessment Refinement: The model suggests that risk assessment for cannabis use should consider configurations of factors rather than simply adding up independent risks, acknowledging the complex interplay among predictors.

Statistical Sophistication The analysis in pois15 demonstrates several elements of statistical sophistication:

Hypothesis-Driven Modeling: Rather than indiscriminately testing all possible interactions, the analysis focuses on theoretically meaningful interactions that address specific questions about how risk factors operate across different groups. Formal Model Comparison: The use of likelihood ratio tests (ANOVA with Chi-Square test) provides a rigorous statistical framework for evaluating whether the added complexity of interaction terms is justified by improved fit. Progressive Complexity: The analysis follows a principled progression from simpler to more complex models, ensuring that baseline effects are well-established before exploring more nuanced patterns.

Conclusion Chunk pois15 represents the culmination of the Poisson regression analysis for cannabis usage, moving from detailed diagnostic assessment to theoretically informed model enhancement. The analysis confirms the base model's substantial explanatory power while demonstrating that accounting for interactions

among predictors further improves understanding of cannabis use patterns. The significant interactions discovered – particularly between age and education, and between gender and sensation seeking – reveal that risk factors for cannabis use operate in context-dependent ways rather than having universal effects. These findings have important implications for both theoretical understanding of substance use and practical approaches to prevention and intervention. Most importantly, the analysis demonstrates how statistical sophistication and substantive theory can reinforce each other in the study of complex behavioral phenomena. The model enhancements are simultaneously justified by statistical diagnostics (addressing non-linear patterns observed in residuals) and informed by theoretical questions about how demographic and personality factors interact to influence substance use behavior. This integration of methodological rigor and substantive insight represents the hallmark of high-quality behavioral science research.

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