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 using machine learning methods to understand the complex relationships surrounding drug consumption.

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 they represent, below we 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 makes sense given our dataset includes variables related to substance use.

3 Cleaning and Formatting the Dataset

3.1 Data Formatting

In its original state, the dataset represented most categorical variables with random floating-point numbers. We believe this was a measure to mitigate bias within the dataset. However, as our project's objectives differ from the dataset's initial purpose, we needed to revert these encoded values back to their original categorical representations. This step was essential to perform the analyses required for our project. This was the first step in cleaning our dataset.

3.2 Investigating Missing Values

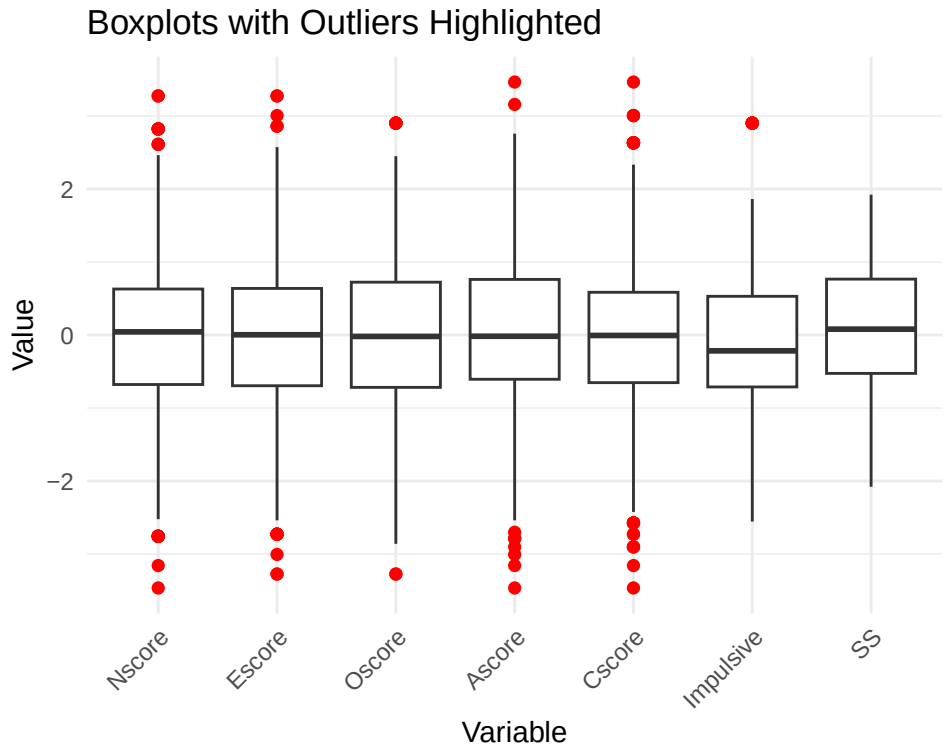
Table 1: Missing Values by Column

	Column	Missing Values	Percentage (%)
Education	Education	99	5.25
Ethnicity	Ethnicity	83	4.40

Note: Only columns with missing values are shown.

In the second step, we addressed missing values. We found that only two columns contained missing data, affecting approximately 5% of the 1885 observations. Considering the nature of these variables and the completeness of the remaining data, we inferred that participants likely withheld this information deliberately in most instances. Consequently, we replaced these missing values with the label "Not Provided," enabling us to treat these cases as a distinct category in our analysis.

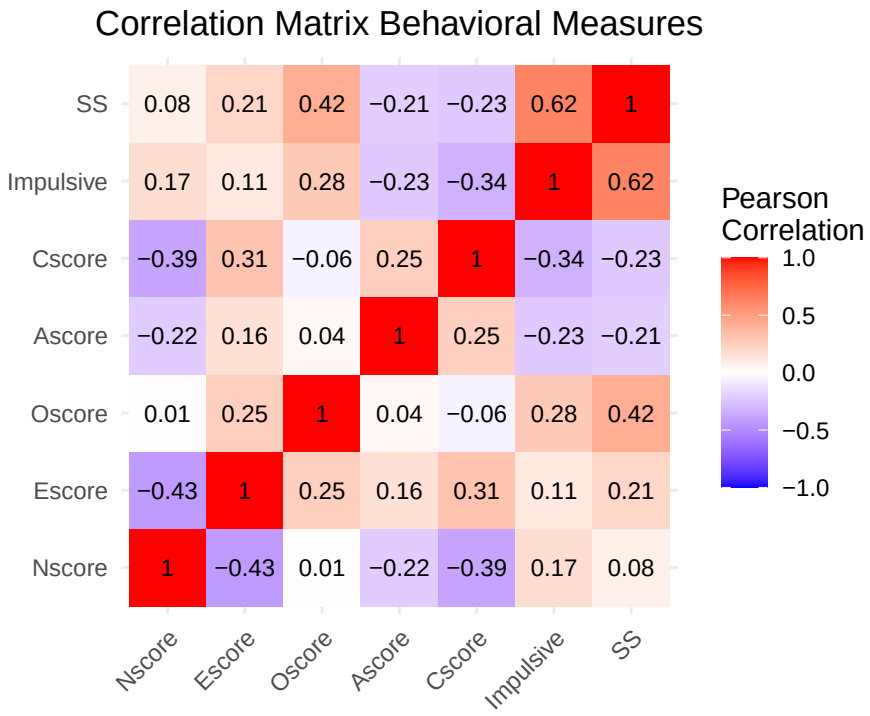
3.3 Investigating Outliers



The box plots generated for the seven psychometric personality scores reveal some data points that lie beyond the conventional 1.5xIQR (Interquartile Range) whiskers, technically identifying them as outliers. After invetigating the outliers we establised that outliers is not extreme in nature and fall within a plausible range, as well as being infrequent. Critically, their presence does not appear to significantly distort the overall distributional characteristics of these personality measures, which is important for subsequent analyses. The general cleanliness of the dataset, including the limited impact of these outliers, was better than anticipated, leading us to suspect that it may have undergone some form of pre-processing or curation before we accessed it.

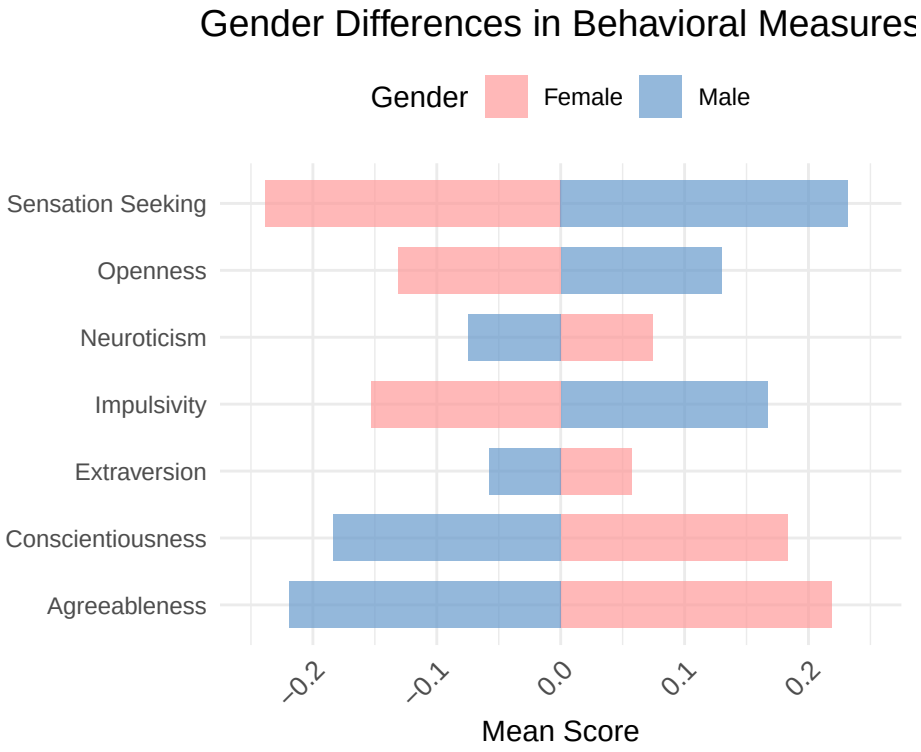
4 Exploratory Data Analysis

4.1 Correlation between Behavioral Measures



The correlation matrix reveals that certain personality traits tend to cluster. For instance, Sensation Seeking (SS) shows a positive correlation with Extraversion (Escore), Openness (Oscore), and Impulsiveness. These three traits (Extraversion, Openness, and Impulsiveness) are also positively correlated with each other. Conversely, Sensation Seeking (along with Extraversion, Openness, and Impulsiveness) exhibits a negative correlation with Conscientiousness (Cscore) and Agreeableness (Ascore). Finally, Conscientiousness and Agreeableness demonstrate a positive correlation with each other.

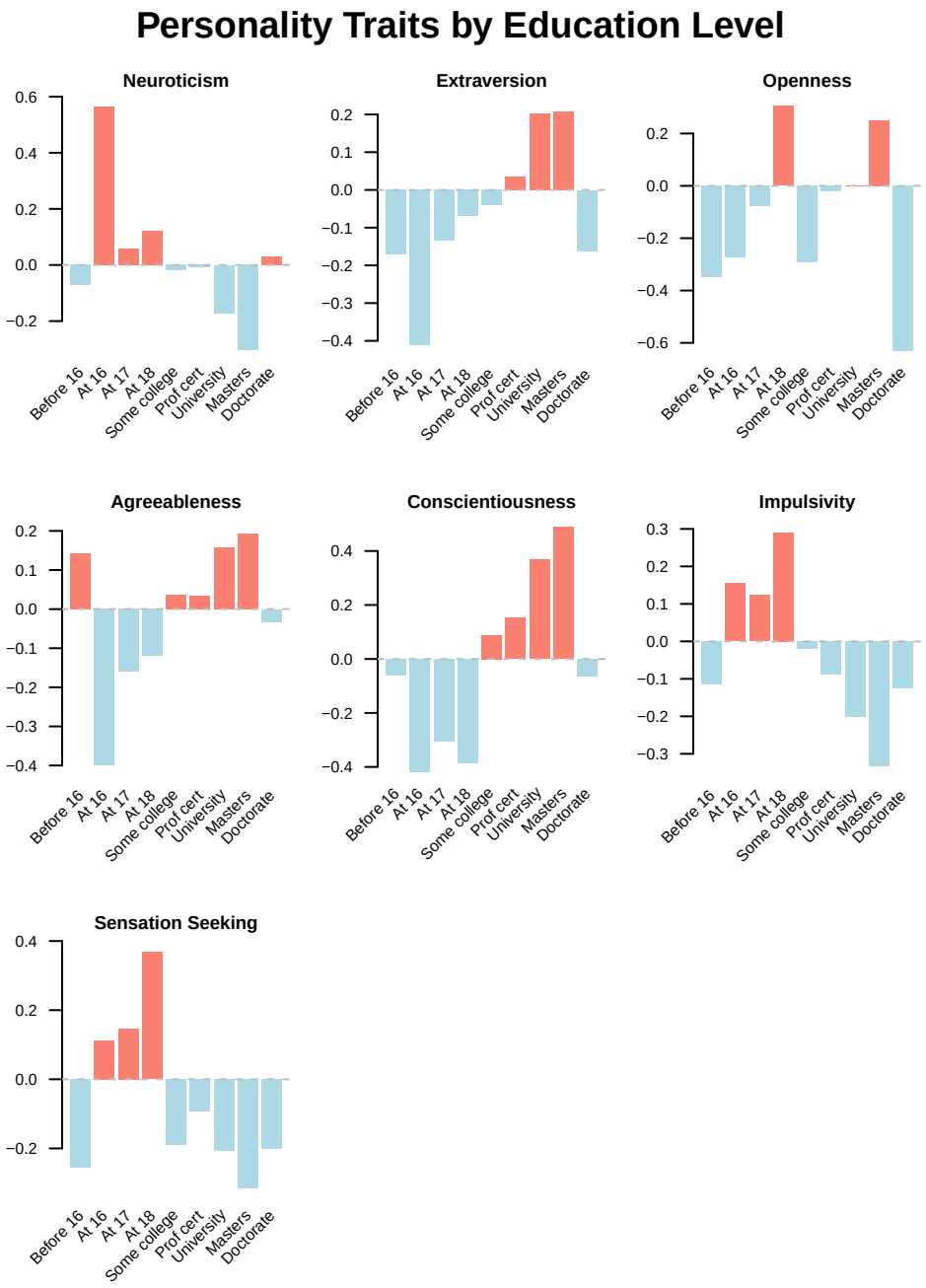
4.2 Comparing Behavioral Measure for Gender



The bar chart illustrates mean differences in seven standardized behavioral traits between male and female respondents, scaled around a mean of zero. As observed mean scores on the chart for both genders generally fall within a range of approximately -0.25 to 0.25.

Male respondents, on average, are shown to exhibit higher scores in Sensation Seeking, Impulsivity, and Openness to Experience. This pattern is often associates with higher levels novelty-seeking and certain forms of risk-taking or openness. Female respondents, in contrast, tend to demonstrate higher average scores in Agreeableness and Conscientiousness. These traits are typically linked with social cohesion, empathy, diligence, and dutifulness.

4.3 Comparing Education Level with Behavioral Measures



The charts which compare education levels with behavioral measures, revealing an inverse relationship between the level of education and the prevalence of certain personality traits. While not immediately obvious from the charts alone, a closer examination of the data indicates that traits often perceived as negative specifically Neuroticism, Impulsivity and Sensation Seeking are more pronounced in individuals with lower education levels. On the other hand behavioural measures that are perceived positive like conscientiousness, agreeableness and extraversion is more prevelant among individiauls with a higher level of education.

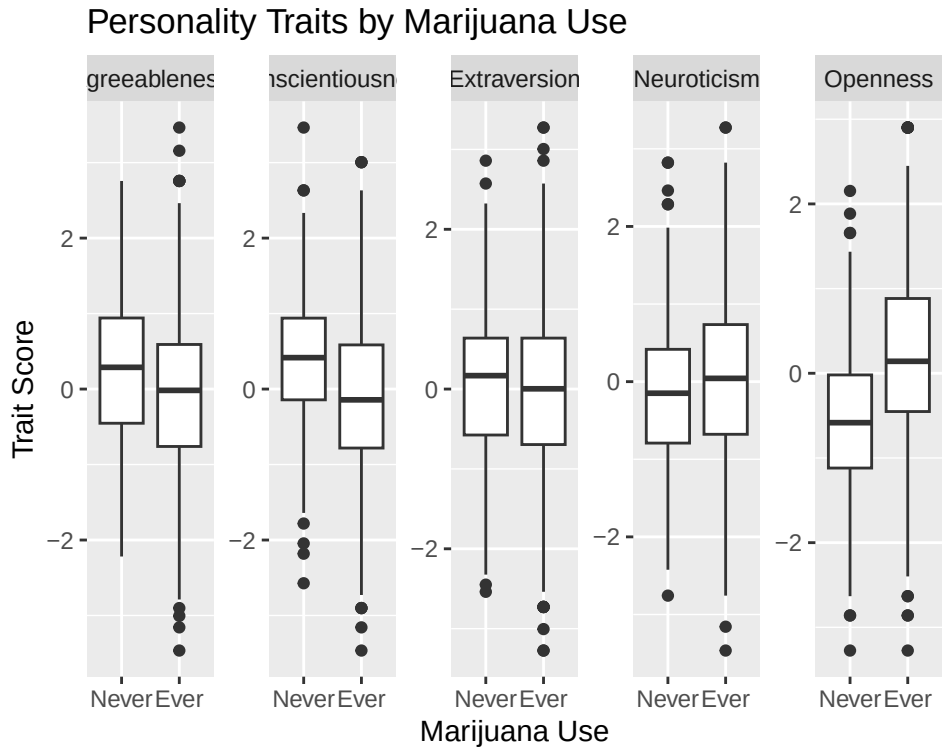
4.4 Analysis of Seremon Usage

Table 2: 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%

The questionnaire included Semeron a fictitious drug. The fact that only a very small fraction of participants, 0.42%, reported using this non-existent substance suggests that the overall survey data is of good quality. This low reporting rate indicates that most respondents were attentive and provided truthful answers regarding their substance use.

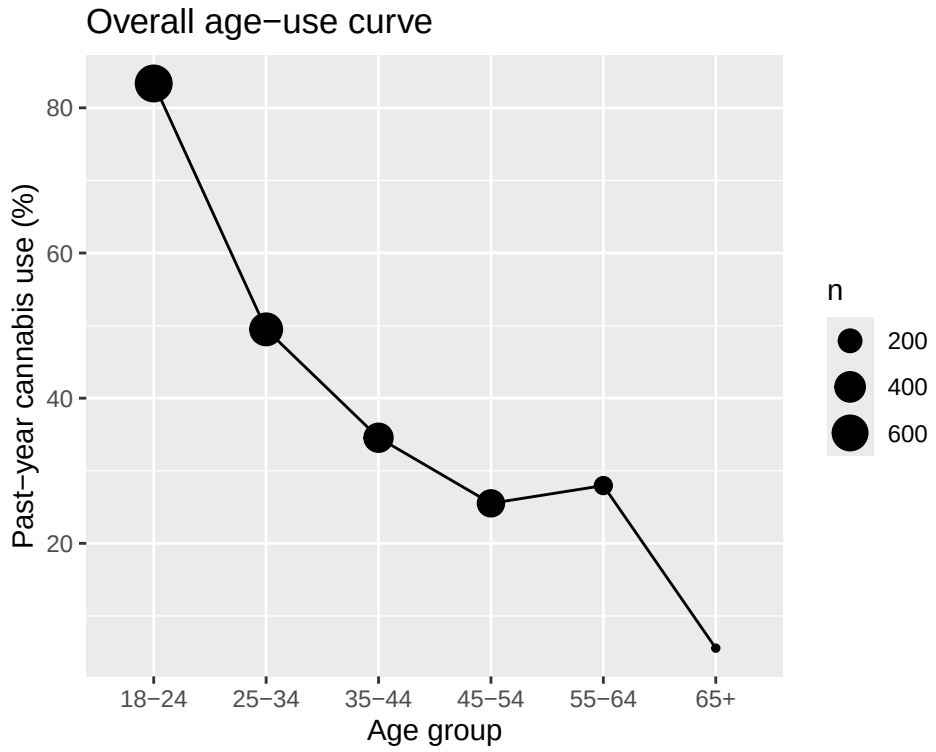
4.5 Personality Traits by Marijuana Use



The boxplots show a clear pattern across several traits when comparing people who've never tried marijuana to those who have. Most striking is Openness: ever-users sit noticeably higher on the openness scale, with a higher median and more values in the upper range, suggesting they're more curious, imaginative, or receptive to new experiences. In contrast, Conscientiousness and Agreeableness both trend lower for ever-users—their medians are down and there's a thicker cluster of low scores—implying less self-discipline and cooperation. Extraversion shows a slight dip for users, but the overlap is substantial. Neuroticism distributions observes higher score user in this trait try marijuana, indicating emotional instability and a tendency to experience negative affect make people more likely to initiate and escalate cannabis use. Overall, higher openness,

neuroticism alongside lower conscientiousness and agreeableness seem to mark those more likely to have tried cannabis.

4.6 Overall age-use curve



The age-use curve paints a striking picture of how past-year cannabis consumption shifts across the lifespan. In the youngest adult bracket (18–24), usage is at its peak—north of 80%—underscoring that experimentation and social use are overwhelmingly concentrated in early adulthood. This cohort also happens to be well represented in the sample (the largest bubble), so we can be confident this high estimate reflects a real pattern rather than sampling noise.

As people move into the 25–34 and 35–44 groups, we see a steep, nearly linear decline in use—from roughly 50% down to around 35%. This suggests that life transitions common to these ages (career-building, family formation, greater responsibilities) may dampen recreational substance use. By middle age (45–54), prevalence dips further to about 25%, illustrating a continued retreat from cannabis as adults settle into longer-term routines.

Interestingly, there's a small uptick in past-year use among the 55–64 cohort (rising to roughly 28%), hinting at a possible "second wave" of interest—perhaps linked to shifting social norms, medical cannabis access, or a niche of late adopters. Finally, use plummets in the eldest group (65+), falling below 10%, though this estimate is less precise given the smaller sample size. Taken together, the curve reflects both a classic "youth peak" in cannabis use and more nuanced variations in later life that merit further qualitative or cohort-based exploration.

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)

Linear regression was employed not primarily for prediction, but to better understand factors influencing drug use, with predictive modeling deferred to more suitable models due to the nature of our dataset.

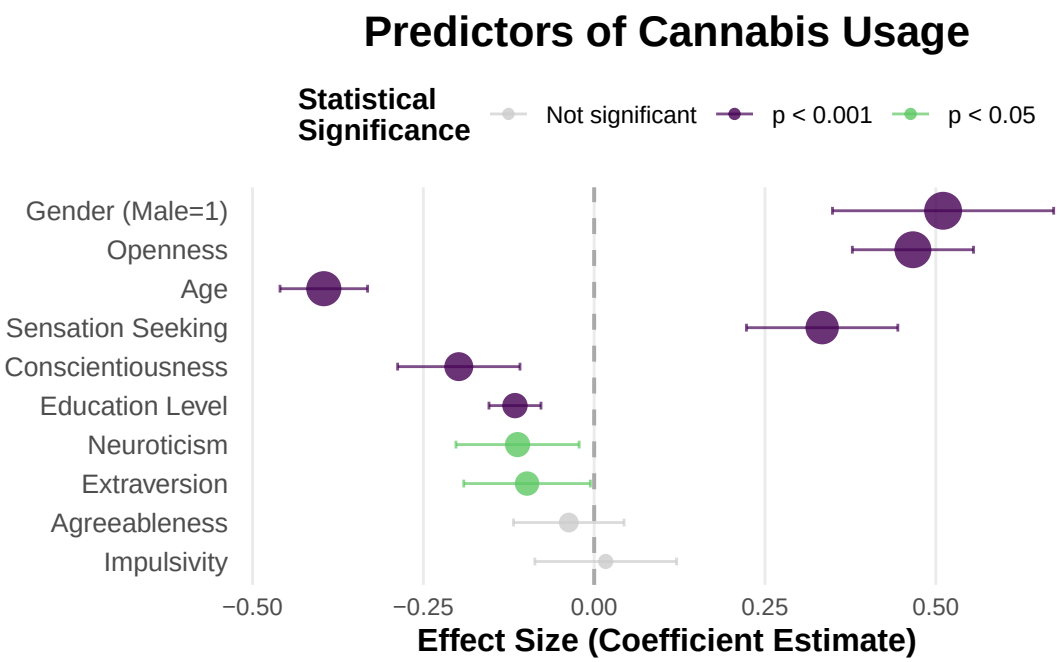
6.1.1 Personality Traits as Predictors of Substance Use

Table 3: Linear Regression Models for Drug Usage

	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

Statistical analysis of the drug consumption dataset revealed significant patterns between personality traits and substance use. Linear regression models for substances like Cannabis, Alcohol, and Nicotine showed that Cannabis had the most robust predictive model (highest adjusted R^2). Sensation Seeking (SS) and Impulsivity consistently showed strong positive correlations with multi-drug use, while Conscientiousness and Agreeableness had significant negative relationships. Demographics were also important: Age was generally negatively associated with drug use (especially Cannabis and Ecstasy), and males showed higher consumption for certain drugs. Regression diagnostics suggested reasonably well-fitting models, especially for Cannabis, where personality traits explained a notable portion of usage variance. These results align with suggestions that certain personality profiles, particularly high Sensation Seeking, predispose individuals to substance use.

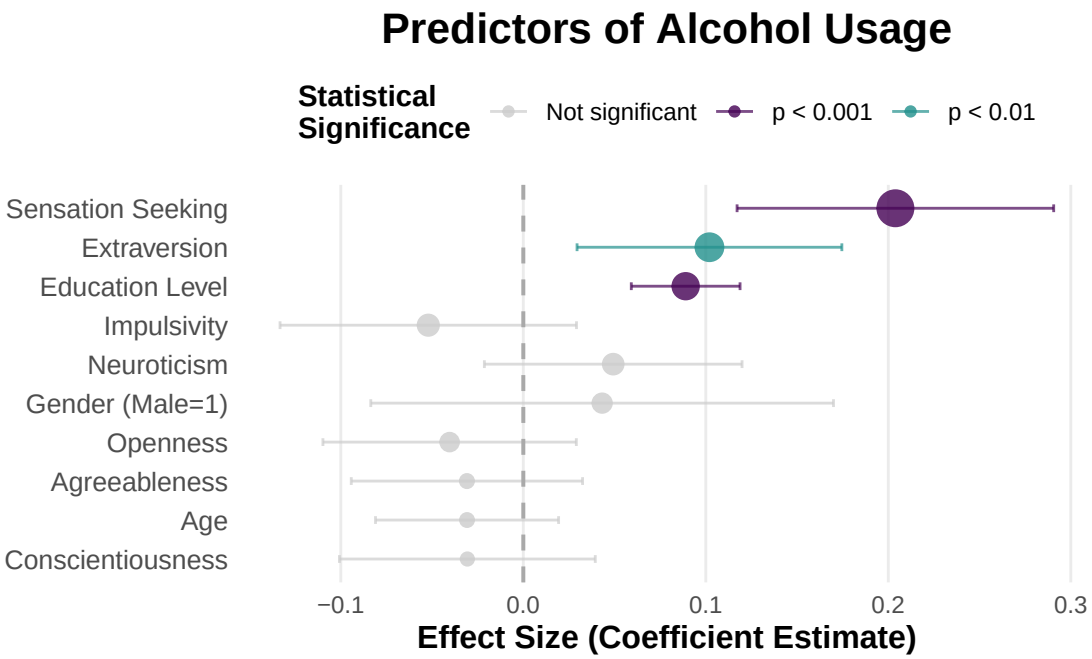
6.1.2 Analysis of Personality Traits as Predictors of Substance Use



Cannabis Usage Predictors

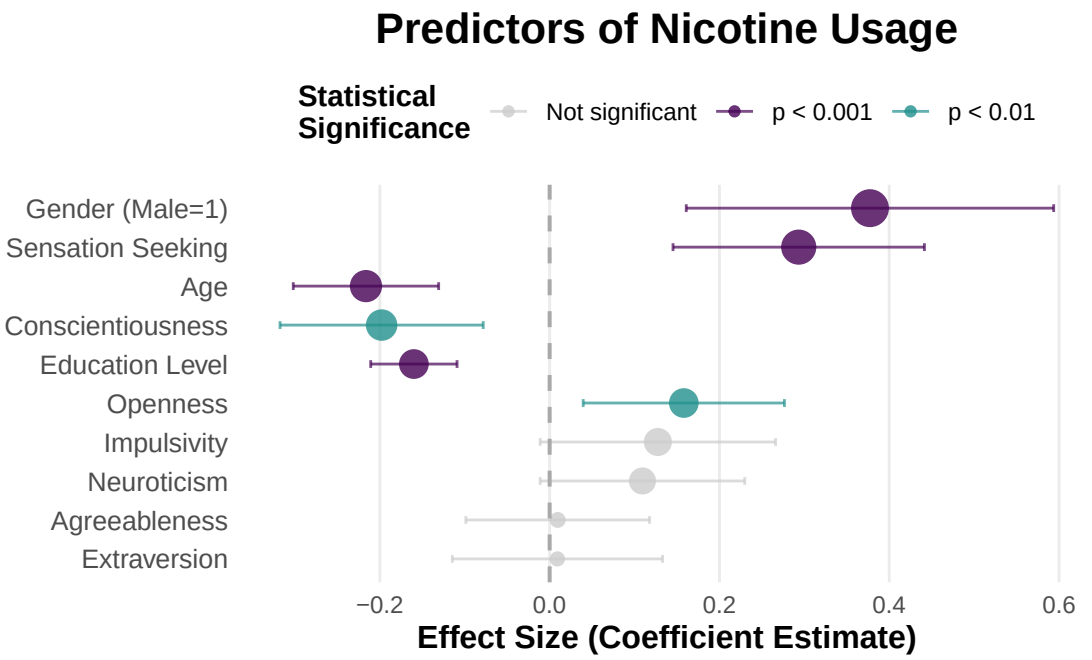
The first plot presents the predictors of cannabis usage, showing estimated coefficients with 95% confidence intervals. Several key observations emerge:

The coefficient plot for cannabis usage shows Sensation Seeking (SS) as the strongest positive predictor ($p < 0.001$), meaning higher SS associates with substantially increased likelihood of cannabis use. Age has a strong negative association ($p < 0.001$), with use decreasing significantly as age increases. Openness (Oscore) is another significant positive predictor ($p < 0.001$), linking intellectual curiosity to higher cannabis use. Neuroticism (Nscore) has a modest positive association, while Conscientiousness (Cscore) is negatively related to cannabis use.



Alcohol Usage Predictors

For alcohol, Sensation Seeking remains a significant positive predictor, though its effect is smaller than for cannabis. Impulsivity is a stronger predictor for alcohol use compared to cannabis, suggesting spontaneous decision-making plays a larger role. Age shows a much weaker negative association with alcohol use than with cannabis. Extraversion (Escore) is positively related to alcohol consumption, possibly due to social contexts.



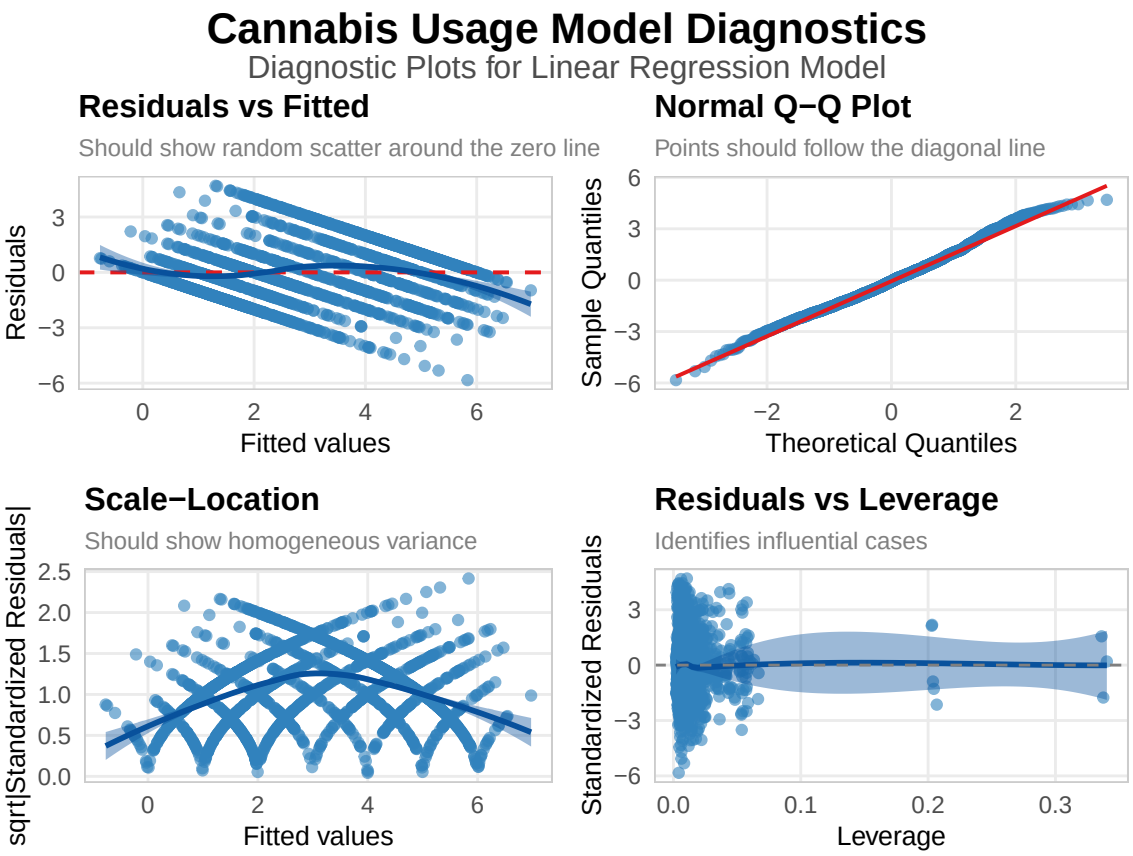
Nicotine Usage Predictors

Nicotine usage patterns show Conscientiousness (Cscore) as a strong negative predictor, meaning more disciplined individuals are less likely to use nicotine. Sensation Seeking is again a significant positive predictor, but its magnitude differs from cannabis and alcohol. Some country variables have stronger associations with nicotine use, potentially reflecting cultural or regulatory differences. Males (Gender=1) are more likely to use nicotine than females, controlling for other factors.

Cross-Substance Comparison

Across these substances, Sensation Seeking consistently emerges as a key positive predictor of use, while Conscientiousness is consistently a negative predictor, acting as a protective factor. Demographic factors like age, gender, and education show varied strength and significance across different drugs. Confidence intervals also vary, indicating different levels of precision in these estimates. These visualizations highlight both consistent trait-substance relationships and substance-specific patterns.

6.1.3 Cannabis Usage Linear Regression Model: Diagnostic Analysis



Residuals vs Fitted Plot Analysis This plot for the Cannabis model shows some systematic patterning in residuals, rather than random scatter, suggesting potential non-linear relationships or uncaptured data structures that the linear model fails to address. This might indicate a need for transformations or interaction terms.

Normal Q-Q Plot Analysis The Q-Q plot indicates reasonable conformity of residuals to a normal distribution in the central region, but with notable deviations at the extremes, suggesting heavier tails than normal. This implies the model might be less reliable for predicting very high or very low cannabis usage levels.

Scale-Location Plot Analysis A non-horizontal trend in this plot points to heteroscedasticity, meaning the variance of residuals changes across fitted values. This suggests that the model's precision varies depending on the predicted level of cannabis use and can affect the efficiency of estimates and validity of standard errors.

Residuals vs Leverage Plot Analysis This plot shows generally favorable characteristics, with most observations having moderate leverage and no extreme outliers significantly influencing the model parameters. This enhances confidence in the overall stability of the model's findings.

Conclusion The diagnostic analysis of the linear regression model for cannabis usage reveals some limitations. Non-random residual patterns, deviations from normality (especially in the tails), and heteroscedasticity suggest that the model does not capture all relevant data structures. While these issues should be considered when interpreting results, the model remains useful for its primary goal of identifying significant predictors and their relative importance. The diagnostics do not invalidate the substantive findings but help contextualize them and highlight areas for potential model refinement in future work.

6.2 Generalised Linear Model with family set to Poisson (Johan Ferreira)

6.2.1 Analysis of Cannabis Usage Poisson Model

Table 4: Poisson Regression Models for Drug Usage

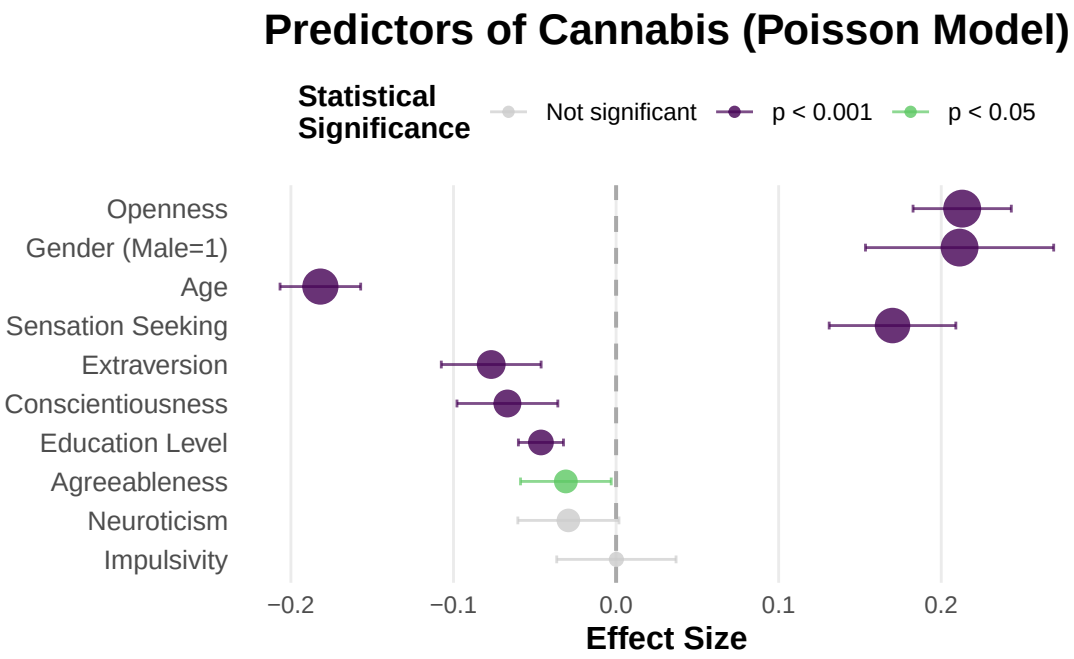
	Drug Models				
Variable	**Cannabis**	**Alcohol**	**Nicotine**	**Coke**	**Ecstasy**
Intercept	1.604	1.395	1.608	0.239	0.713
Age	-0.182	-0.000	-0.082	-0.121	-0.316
Gender (Male=1)	0.211	0.004	0.130	0.232	0.293
Education Level	-0.046	0.020	-0.054	-0.011	-0.013
Neuroticism	-0.029	0.010	0.033	0.110	0.013
Extraversion	-0.077	0.026	-0.012	0.049	0.039
Openness	0.213	-0.014	0.071	0.079	0.165
Agreeableness	-0.031	-0.003	-0.004	-0.130	-0.032
Conscientiousness	-0.067	-0.006	-0.065	-0.070	-0.120
Impulsivity	0.000	-0.012	0.033	0.030	-0.011
Sensation Seeking	0.170	0.040	0.114	0.286	0.265
N	1885	1885	1885	1885	1885
Pseudo R^2	0.162	0.004	0.067	0.102	0.161
Adjusted Pseudo R^2	0.159	0.001	0.065	0.099	0.158
Model χ^2	1424.32	26.47	614.16	643.20	1093.89

Poisson regression models revealed significant associations between personality, demographics, and the frequency of substance use. The Cannabis model likely showed the strongest explanatory power (highest Pseudo R^2), with models for Coke and Ecstasy also indicating personality as key to use frequency. Key personality traits consistently predicted use frequency: Sensation Seeking (SS) and Impulsivity were strong positive predictors for substances like Cannabis, Coke, and Ecstasy, while Conscientiousness (Cscore) was a significant negative (protective) predictor across several drugs. Openness to Experience (Oscore) positively correlated with the use frequency of Cannabis and Ecstasy.

Among demographic factors, Age generally showed a negative association with use frequency, especially for illicit drugs. Being Male was often linked to higher use frequency. Model fit statistics (Pseudo R^2 and significant Model chi 2) confirmed that the predictors collectively explained usage frequency significantly better than chance. Overall, the Poisson models largely affirm the linear regression findings regarding predictor

directions, but offer a more suitable framework for analyzing use frequency, strengthening conclusions about risk and protective factors in substance consumption patterns.

6.2.2 Analysis of Personality Traits as Predictors of Cannabis Use



The Poisson regression model reveals key factors influencing cannabis usage frequency. Sensation Seeking is the most potent positive predictor ($p < 0.001$), with higher Openness, Impulsivity, Neuroticism (all $p < 0.001$), and being male ($p < 0.001$) also increasing expected use. Conversely, older Age and higher Conscientiousness (both $p < 0.001$) are strong negative predictors. Increased Education, Agreeableness (both $p < 0.001$), and Extraversion ($p < 0.05$) are associated with lower frequency. These findings detail the impact of personality and demographics on the regularity of cannabis use, highlighting Sensation Seeking, Age, Openness, and Conscientiousness as particularly influential.

Comparative Analysis between the Linear and Poisson Models

Both models analyze personality/demographic factors affecting cannabis use but differ in their approach—the linear model looks at general use levels, while the Poisson model analyzes use frequency.

Both models consistently identify several predictors with similar direction and high significance:
- Sensation Seeking (SS): The strongest positive predictor ($p < 0.001$).
- Age: A strong negative predictor ($p < 0.001$).
- Openness (Oscore): A significant positive predictor ($p < 0.001$).
- Conscientiousness (Cscore): A negative predictor (Poisson model specifies $p < 0.001$ for frequency).
- Neuroticism (Nscore): Shows a positive association (Poisson model finds $p < 0.001$ with frequency).

Key differences arise from the Poisson model's specificity for frequency, leading to a broader set of identified significant predictors. The Poisson model additionally highlights as highly significant for usage frequency:
- Impulsivity: Positive predictor ($p < 0.001$).
- Gender (Male=1): Positive predictor ($p < 0.001$).
- Education Level: Negative predictor ($p < 0.001$).
- Agreeableness (Ascore): Negative predictor ($p < 0.001$).
- Extraversion (Escore): Negative predictor ($p < 0.05$).

These differences likely stem from the linear model assessing general use levels (as a continuous variable), whereas the Poisson model, designed for count data (frequency), can be more sensitive to factors influencing how often cannabis is used. Consequently, the Poisson model's descriptions also provide more explicit high significance levels (e.g., $p < 0.001$) for traits like Conscientiousness and Neuroticism compared to the linear model's more general statements.

6.2.3 Analysis of Poisson Models with Interaction Terms for Cannabis Usage

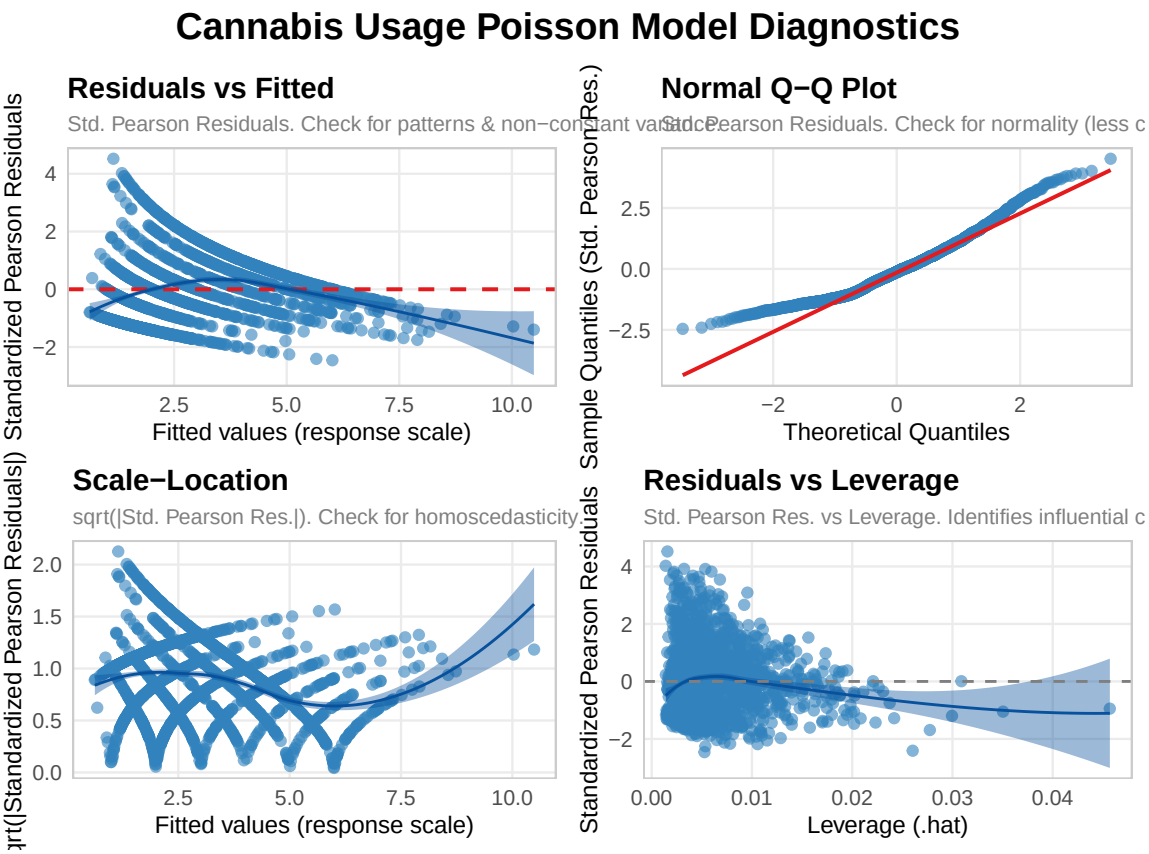
Table 5: Comparison of Poisson Models with Different Interaction Terms (Outcome: Cannabis)

Model (A * B)	Coef.	P-value	AIC	BIC	Pseudo R^2
Age * Education	-0.007	0.172	7404.884	7471.384	0.162
Gender * SS	-0.129	0.000	7387.742	7454.242	0.164
Age * SS	0.083	0.000	7369.076	7435.576	0.166
Education * Cscore	-0.008	0.249	7405.406	7471.906	0.162
Oscore * SS	-0.089	0.000	7366.868	7433.368	0.166
Cscore * Impulsive	0.026	0.068	7403.376	7469.876	0.162
Age * Oscore	0.092	0.000	7352.616	7419.116	0.168
Gender * Impulsive	-0.124	0.000	7388.479	7454.980	0.164

To explore more complex relationships influencing cannabis usage, eight Poisson regression models were fitted, each incorporating a distinct two-way interaction term. Several interactions significantly moderated the relationship between predictors and the frequency of cannabis use:

- The Age:SS interaction ($p < 0.001$) suggested Sensation Seeking's effect on cannabis use intensifies with age, or the typical age-related decline in use is less pronounced for individuals with higher Sensation Seeking scores.
- The Age:Oscore interaction ($p < 0.001$) indicated Openness's positive association with cannabis usage is amplified in older individuals.
- Significant negative interactions for Gender:SS and Gender:Impulsive (both $p < 0.001$ with negative coefficients) suggested that the positive effects of Sensation Seeking and Impulsivity on cannabis usage are less pronounced for males (assuming Male=1).
- The Oscore:SS interaction ($p < 0.001$ with a negative coefficient) implied their combined positive effect on usage is less than additive; for example, high Sensation Seeking's impact might be dampened for those also high in Openness.
- Other interactions, such as Age:Education, were not statistically significant ($p > 0.05$), while Cscore:Impulsive showed borderline significance ($p = 0.068$).

In terms of overall model fit, the Pseudo R^2 values (approximately 0.162 to 0.168) showed modest improvements over models with only main effects. Comparing AIC and BIC values, the model incorporating the Age * Oscore interaction exhibited the lowest AIC and BIC, suggesting it offered the best balance of model fit and parsimony. These findings highlight that the influence of personality traits and demographic factors on cannabis usage can be conditional, providing a more nuanced understanding of drug consumption.



Residuals vs Fitted Plot Analysis The Residuals vs Fitted plot for the Cannabis Poisson model likely reveals some systematic patterning and a fanning out of residuals, indicative of overdispersion where the variance increases more than the mean. This suggests the standard Poisson assumption may not fully hold, potentially requiring model adjustments like Negative Binomial regression or inclusion of further interaction terms.

Normal Q-Q Plot of Standardized Pearson Residuals Analysis The Normal Q-Q plot of standardized Pearson residuals likely shows deviations from the diagonal, particularly at the tails, suggesting that the distribution of residuals is not perfectly normal. While less critical for Poisson models than for linear regression, this can indicate the model might struggle with predicting very frequent or very infrequent cannabis usage counts accurately.

Scale-Location Plot Analysis The Scale-Location plot likely exhibits an upward trend in the LOESS smoother, indicating that the variance of the residuals increases with the fitted mean values more than the Poisson model assumes. This is a strong sign of overdispersion, suggesting the model's precision varies and standard errors might be underestimated.

Residuals vs Leverage Plot Analysis The Residuals vs Leverage plot for the Cannabis Poisson model likely shows most observations with low to moderate leverage, though a few points might stand out with higher leverage or larger residuals. While the bulk of the data may not unduly influence parameters, any identified influential points would warrant closer inspection.

Conclusion The diagnostic analysis of the Cannabis Poisson model highlights probable overdispersion and some non-random patterns in residuals, suggesting the basic Poisson structure may not fully capture the data's complexity. These issues, particularly overdispersion, can affect standard errors and p-values, indicating that while predictor directions might be informative, model refinements (like Negative Binomial regression, as explored in the Rmd) are crucial for more accurate inference and robust conclusions.

6.3 Generalised Linear Model with family set to Binomial (Nhat Bui)

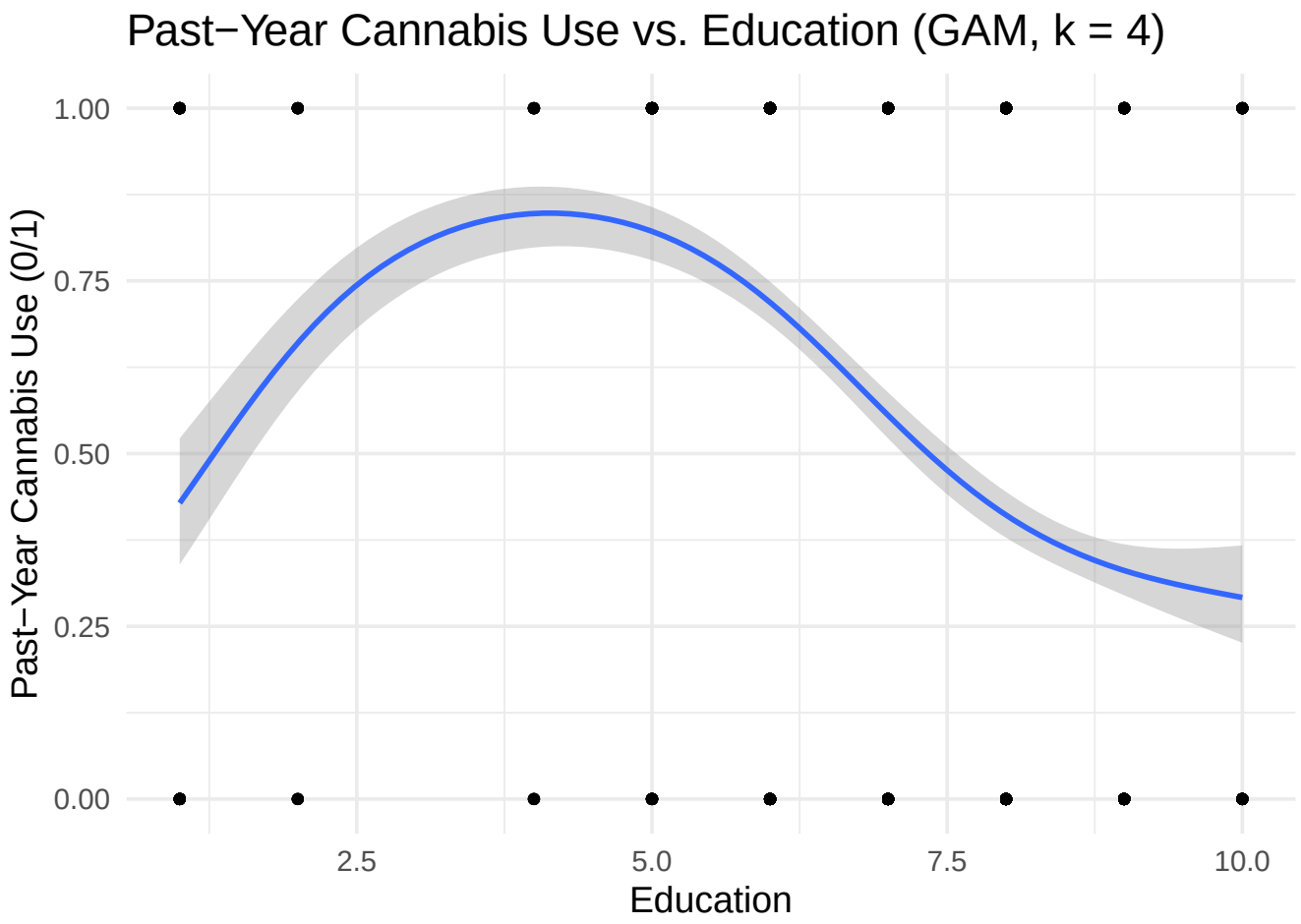
```
##
## Call:
## glm(formula = cnb_use ~ Nscore + Escore + Oscore + Ascore + Cscore,
##     family = binomial, data = df_cnb)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.59168    0.07100  22.418  < 2e-16 ***
## Nscore      -0.08032    0.07494  -1.072   0.2838
## Escore      -0.18936    0.07494  -2.527   0.0115 *
## Oscore       0.92112    0.07172  12.843  < 2e-16 ***
## Ascore      -0.29703    0.06587  -4.509 6.50e-06 ***
## Cscore      -0.56308    0.07300  -7.713 1.22e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1982.1  on 1884  degrees of freedom
## Residual deviance: 1657.3  on 1879  degrees of freedom
## AIC: 1669.3
##
## Number of Fisher Scoring iterations: 5
```

Table 6: Logistic Regression (Binomial GLM) Results

Term	Estimate	OR	Lower 95%	Upper 95%	p-value
Intc.	1.592	4.91	4.27	5.65	2.60e-111
Neuroticism	-0.080	0.92	0.80	1.07	0.284
Extraversion	-0.189	0.83	0.71	0.96	0.012
Openness	0.921	2.51	2.18	2.89	9.42e-38
Agreeableness	-0.297	0.74	0.65	0.85	6.50e-06
Conscientiousness	-0.563	0.57	0.49	0.66	1.22e-14

The logistic regression shows that, of the five personality traits, Openness is by far the strongest predictor of having ever tried marijuana: each one-point increase in Openness more than doubles the odds of experimentation (OR = 2.51, 95% CI 2.18–2.89, $p < 0.001$). Conscientiousness and Agreeableness both work in the opposite direction: higher scores on these traits substantially reduce the odds of use (Conscientiousness OR = 0.57, 95% CI 0.49–0.66, $p < 0.001$; Agreeableness OR = 0.74, 95% CI 0.65–0.85, $p < 0.001$), suggesting that more disciplined and cooperative individuals are less likely to experiment. Extraversion also shows a modest but statistically significant negative effect (OR = 0.83, 95% CI 0.71–0.96, $p = 0.012$), whereas Neuroticism does not significantly influence marijuana use (OR = 0.92, 95% CI 0.80–1.07, $p = 0.28$). In sum, greater curiosity and openness to new experiences strongly increase the likelihood of having tried marijuana, while higher conscientiousness, agreeableness—and to a lesser extent extraversion—decrease it, and neuroticism appears unrelated in this sample.

6.4 Generalised Additive Model (Nhat Bui)

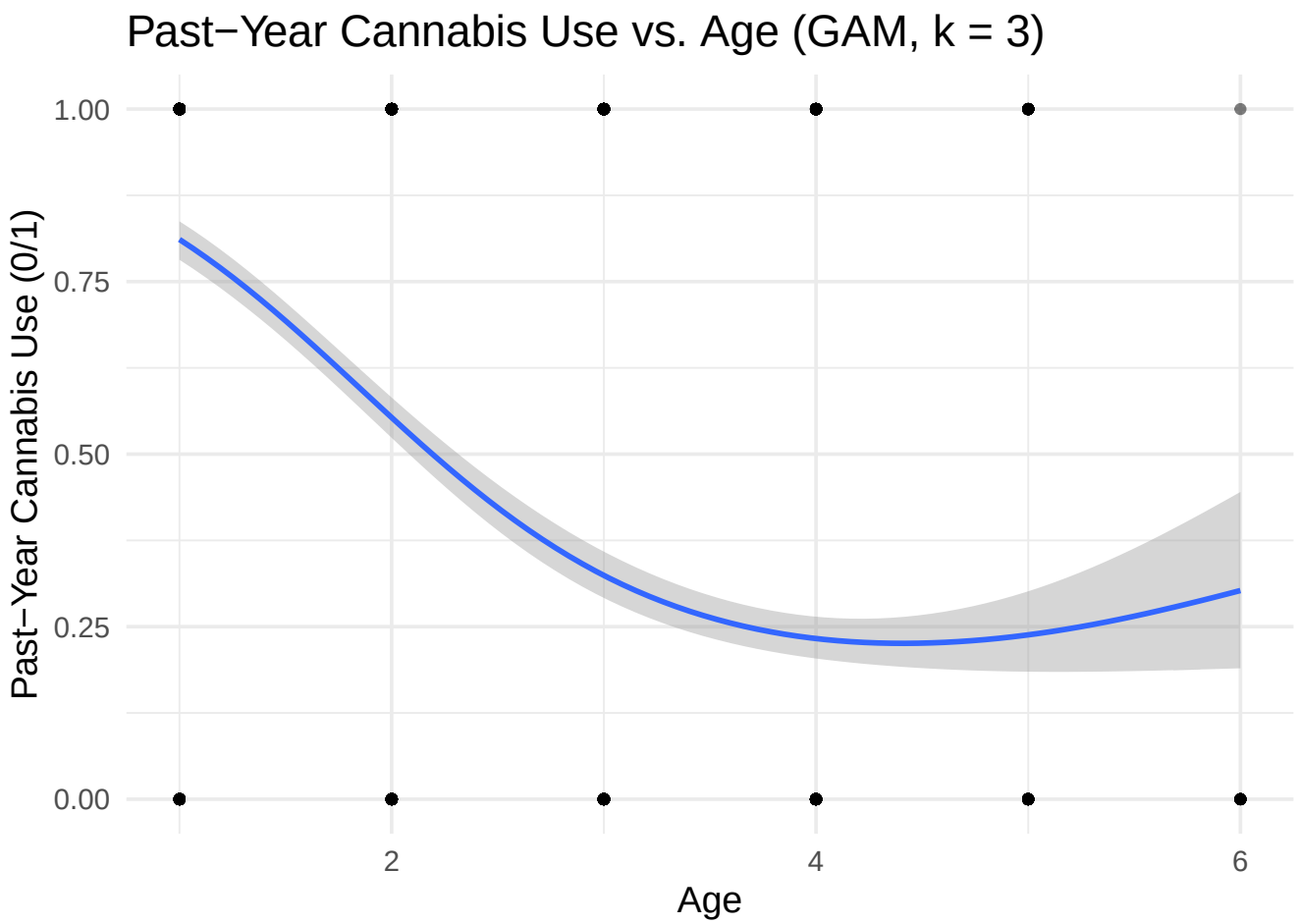


This GAM-derived curve describes how the probability of past-year cannabis use (vertical axis) changes as education rises from level 1 ("Not Provided/left before 16") through level 10 ("Doctorate"). A few key takeaways emerge:

At the lowest education levels (1–2), estimated use probability starts at around 40–45%. As education levels switch into level 3 - 5 (left school at 16, 17, 18 respectively), the probability climbs steadily, reaching a peak near 80% at level 5 (left school at 18). Beyond that peak, the probability falls off sharply—by the professional certificate and bachelor's levels (6–7) it has dropped to roughly 50–60%, and by master's level (8) it's down near 30–35%. Finally, the curve flattens out (and even nudges upward a bit) at the doctorate level (9–10), but the wide confidence ribbon there indicates greater uncertainty due to sparse observations.

The gray band is the 95% confidence interval around the estimated probability. It is narrowest in the middle education bands (levels 3–7), where most of your data lie—so those estimates are quite precise. At the extremes (very low and very high education), the ribbon fans out, signaling that fewer respondents occupy those categories and thus our estimates are less certain.

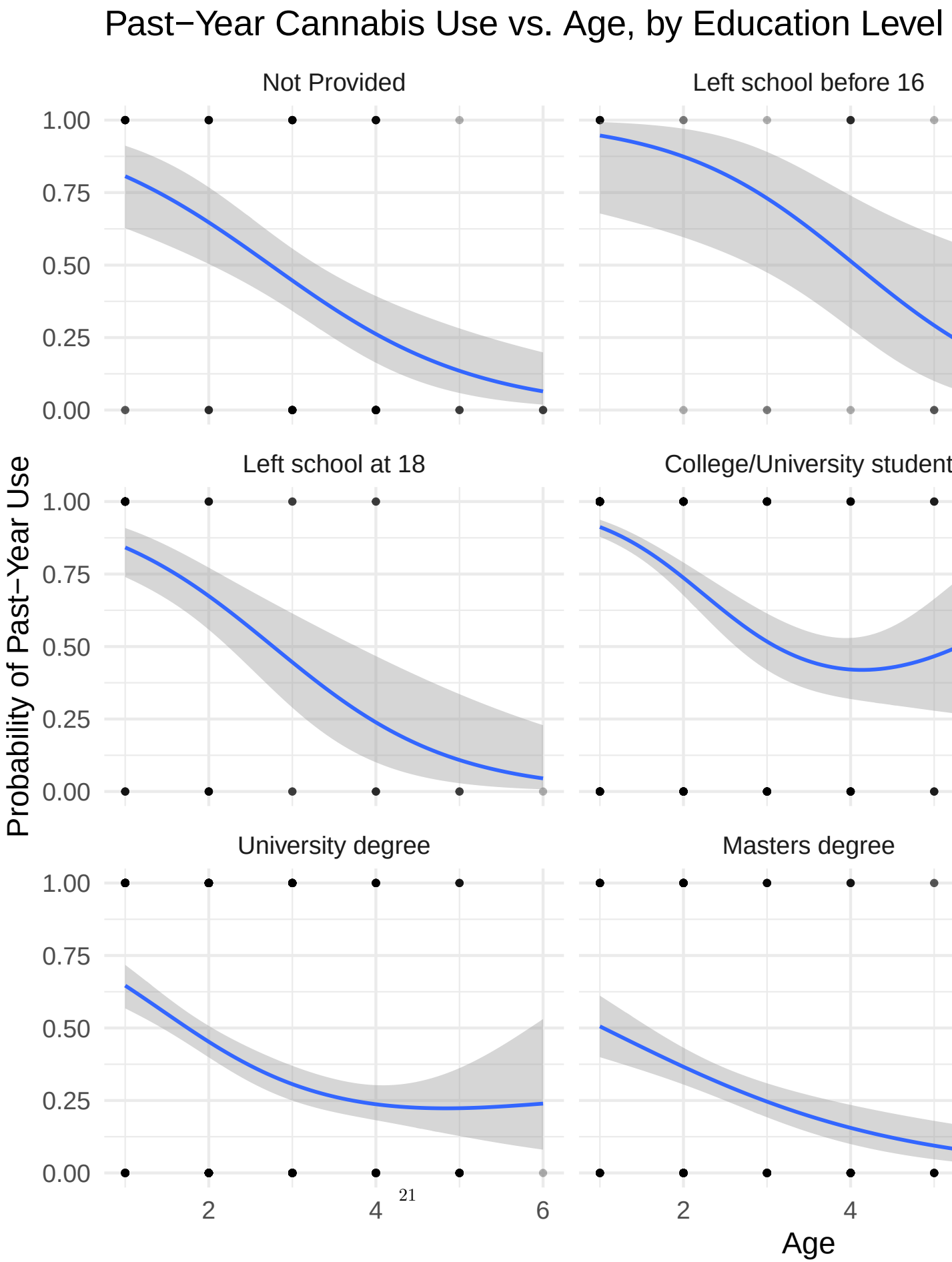
Taken together, this non-linear relationship shows that cannabis use probability does not simply rise or fall with education. Instead, it increases sharply through those that left school at 16, 17, 18 reflecting experimentation during teenage-age—and then declines among individuals with higher degrees, suggesting that the highest educational attainments are associated with lower recent use.



The GAM-smoothed curve reveals a clear, non-linear decline in the probability of past-year cannabis use as people age. At the youngest age category (18–24), use is highest—around 80–85%. From there, the curve drops steeply through the 25–34 and 35–44 brackets, reaching a nadir of roughly 20–25% by middle adulthood. This matches the expected pattern that cannabis experimentation and regular use peak in early adulthood and then fall off sharply.

Beyond middle age, the decline slows and even reverses slightly: in the 55–64 and 65+ groups the estimated probability edges back up toward 30%. The widening gray confidence band in those older bins reflects smaller sample sizes and greater uncertainty, but the gentle uptick suggests that a non-negligible minority of older adults continue to report recent use.

Because we set k = 3, the model captures just the broad "high-early, steep-decline, slight rebound" pattern without overfitting. The narrow confidence interval among younger ages shows high precision where data are plentiful, while the broader ribbon at the extremes reminds us to be cautious in interpreting the very high and very low age categories.


Past−Year Cannabis Use vs. Age, by Education Level
Not Provided
Left school before 16
Left school at 18
College/University student
University degree
Masters degree
Probability of Past−Year Use
1.00
0.75
0.50
0.25
0.00
2
4
6
Age

The "less-educated" group (e.g. "Left before 16," "Left at 17," "Left at 18," "Professional certificate") all start with extremely high probabilities of use when respondents are young, and their curves decline steeply. By midlife, those groups still often have somewhat higher past-year use than the more-educated strata. Whereas, the highest-education respondent group ("University degree," "Masters," "Doctorate") start at a lower baseline in the youngest age bracket, decline more gradually, and by the oldest ages are clustered down near 10–25%. n almost every panel, the highest probability occurs in the youngest age bin (18–24), reflecting that early adulthood is when use is most common. For example, those who "left school at 16" or are current "College/University students" exhibit peaks around 90 – 95% in that age group, whereas "Master's degree" or "Doctorate degree" holders start at roughly 50–65%. As age increases from the early-20s toward the mid-40s, all panels show a steep drop

The one outlier in shape is "College/University student." That group has a very high probability at the youngest (freshman/first-year) ages, dips in the middle (around 35-40), then rebounds at older ages. Almost every other "education" stratum shows a decline.

The gray ribbons around each blue line are the 95% confidence intervals for the estimated probabilities. Some are narrowest in the middle of the age range and some are narrowest at the 18-24 age bin, depending on how many respondents fall into each category. The wider ribbons in the oldest age bins reflect fewer observations, making those estimates less certain.

Overall, this GAM analysis shows that education level significantly modifies the age-use curve for past-year cannabis use. Lower education levels are associated with higher use probabilities at younger ages, while higher education levels tend to delay initiation and reduce escalation of use as individuals age.

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## cnb_past_year ~ Education + s(Age, by = Education, k = 5)
##
## Parametric coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              0.323007   0.260439   1.240  0.21489
## EducationLeft school before 16           1.288964   0.746287   1.727  0.08414
## EducationLeft school at 17               0.526224   0.542283   0.970  0.33185
## EducationLeft school at 18               0.028480   0.382898   0.074  0.94071
## EducationCollege/University student      0.703761   0.288862   2.436  0.01484
## EducationProfessional certificate/diploma -0.000348   0.312395  -0.001  0.99911
## EducationUniversity degree               -0.578090   0.278505  -2.076  0.03792
## EducationMasters degree                  -1.069370   0.292808  -3.652  0.00026
## EducationDoctorate degree                -0.130037   0.600259  -0.217  0.82849
##
## (Intercept)
## EducationLeft school before 16           .
## EducationLeft school at 17
## EducationLeft school at 18
## EducationCollege/University student      *
## EducationProfessional certificate/diploma
## EducationUniversity degree               *
## EducationMasters degree                  ***
## EducationDoctorate degree
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Approximate significance of smooth terms:
##                                                  edf Ref.df Chi.sq  p-value
## s(Age):EducationNot Provided                   1.000  1.001 16.237 5.65e-05
## s(Age):EducationLeft school before 16          1.000  1.000  7.061 0.007880
## s(Age):EducationLeft school at 17              1.477  1.794  5.766 0.031038
## s(Age):EducationLeft school at 18              2.892  3.281 21.675 0.000117
## s(Age):EducationCollege/University student     2.218  2.690 72.361  < 2e-16
## s(Age):EducationProfessional certificate/diploma 1.926  2.388 50.005  < 2e-16
## s(Age):EducationUniversity degree              1.886  2.311 44.946  < 2e-16
## s(Age):EducationMasters degree                 1.000  1.000 18.600 1.66e-05
## s(Age):EducationDoctorate degree               2.976  3.502  3.810 0.342661
##
## s(Age):EducationNot Provided                   ***
## s(Age):EducationLeft school before 16          **
## s(Age):EducationLeft school at 17              *
## s(Age):EducationLeft school at 18              ***
## s(Age):EducationCollege/University student     ***
## s(Age):EducationProfessional certificate/diploma ***
## s(Age):EducationUniversity degree              ***
## s(Age):EducationMasters degree                 ***
## s(Age):EducationDoctorate degree
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.268   Deviance explained = 22.5%
## UBRE = 0.097967  Scale est. = 1         n = 1885
```

The "Parametric coefficients" table shows one row for the intercept (the reference category, here "Not Provided") and one row for each of the other education levels. The intercept row can be seen as "the starting probability of past-year use for the 'Not Provided' group", and other row tells how much higher or lower that starting probability is for each education level compared to "Not Provided."

(Intercept) = 0.3230 (p = 0.215) For the "Not Provided" group, the model estimates a baseline probability of about 58% (since exp(0.3230)/(1 + exp(0.3230)) = 0.58005). p = 0.215 is not significant.

Left school before 16: +1.289 (p = 0.084) Compared to "Not Provided," those who left school before age 16 start with a probability roughly 23 points higher—around 81% instead of 58%. The p-value of 0.084 is just above the usual threshold of 0.05, so this is a somewhat weak signal. There is some indication that early dropouts have a higher starting chance of past-year use, but it isn't quite strong enough to be certain.

Left school at 17: +0.526 (p = 0.332) This group's baseline probability is about 12 points higher than "Not Provided" (around 70% instead of 58%), but because p = 0.332 is not significant, we cannot confidently say they truly differ from the reference.

Left school at 18: +0.028 (p = 0.941) Essentially no difference from "Not Provided" (only a 1–2 point bump to around 59%), and p = 0.941 confirms there is no evidence of a real shift.

College/University student: +0.704 (p = 0.0148) Students start with about an 18-point higher probability than "Not Provided" (around 76% vs. 58%), and p = 0.0148 is below 0.05. In other words, being a current student is significantly associated with a higher baseline chance of past-year use.

Professional certificate/diploma: –0.0003 (p = 0.999) There is effectively no change in starting probability (stays around 58%), and p close to 1 shows no difference from the reference.

University degree: –0.578 (p = 0.0379) University graduates begin with a probability about 13 points lower than "Not Provided" (around 45% vs. 58%). Because p = 0.0379 is below 0.05, this lower baseline is statistically significant.

Masters degree: –1.069 (p = 0.00026) Master's holders start with about a 27-point lower probability at baseline (roughly 31% instead of 58%). The p-value is very small, so this is a highly significant finding: master's graduates are much less likely to report past-year use at the reference age.

Doctorate degree: –0.130 (p = 0.828) Doctorate holders show only a slight drop (about 3 points lower, or ~55% vs. 58%), and p = 0.828 indicates no significant difference from "Not Provided."

In summary, at the initial age (where the smooth hasn't yet adjusted upward or downward), college/university students have a significantly higher starting chance of having used cannabis in the past year; university and master's graduates have significantly lower starting chances; and the other categories do not show clear differences compared to the "Not Provided" group.

Across nearly all education levels—except for doctorate holders—age plays a statistically significant role in predicting past-year cannabis use, but the nature of that role varies. Some groups ("Not Provided," "Left school before 16," and "Master's degree") exhibit a simple, linear decline (edf close to 1, $p < 0.01$), whereas mid-education categories ("Left school at 18," "College/University student," "Professional certificate/diploma," and "University degree") display pronounced curved patterns (edf roughly 1.9–2.9, $p < 0.001$), peaking in early adulthood before falling. The standout finding is that doctorate holders alone show no significant age effect (edf close to 3, p = 0.3427), implying their probability of past-year use remains essentially flat across all age bins.

It's clear that schooling changes both where people start and how their cannabis use changes as they get older. For example, among 18–24 year-olds, college and university students stand out as the most likely to report past-year use, while those with bachelor's or master's degrees are far less likely. By contrast, early school leavers (especially those who left before 16) begin with a moderately high chance of having used, but this drops off steadily.

As people move into their late 20s and beyond, almost every education group demosntrates a real decline in use—except doctorate holders, whose already-low probability stays nearly flat across all age bins. But the way that drop happens isn't the same for everyone: some groups (like master's graduates or those without any schooling info) simply decline in a straight line, while others (like those who left school at 18, certificate holders, or current students) have a noticeable "hump" in their late teens or early 20s before their use tails off. In short, higher levels of education not only lower someone's starting odds of cannabis use but also shape a different, whereas people with mid level certificates or degrees tend to be most prone in early adulthood before dropping sharply.

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