Julie Flament
#53203964
DATA 407
Professor Xioping Shi
April 8th, 2025

Data Analysis Project

Introduction

Drug consumption has become an increasingly prevalent and dangerous issue worldwide. The ease of access to both legal and illegal substances has made it essential to understand the trends behind drug use, particularly across different demographics. By identifying trends in drug consumption, we can discover more effective ways to combat this problem. This UCI Machine Learning Repository dataset includes information on drug consumption across multiple demographics. This analysis uses data from the UCI Machine Learning Repository on drug consumption across various demographics. The dataset includes information on the usage of substances such as alcohol, cannabis, nicotine, cocaine, heroin, and others, measured across seven frequency classes: “Never Used,” “Used Over a Decade Ago,” “Used in Last Decade,” “Used in Last Year,” “Used in Last Month,” “Used in Last Week,” and “Used in Last Day.” The drug “semeron” is a control substance used in this experiment to monitor respondents who overreport their drug consumption. All of these drugs are also assessed across different characteristics such as gender, education, country, ethnicity, and personality measurements, which include NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking). While the dataset tracks a wide range of variables, this analysis specifically focuses on the effects of gender and education level on drug consumption patterns. The goal is to uncover trends that can inform more effective strategies for addressing drug use in various population groups.

Methods

In this analysis, I chose to use stratified sampling to ensure that each subgroup is adequately represented, particularly given the varying accessibility of different drugs. For example, substances like chocolate and caffeine are legal and readily available to almost anyone, regardless of demographic. In contrast, illegal drugs like cocaine and ketamine are harder to access, leading to much lower consumption rates. Despite being illegal in some places, cannabis remains highly accessible and is often consumed at higher rates than other illegal drugs. Stratified sampling is also important due to the diversity of variables in the dataset, such as personality types, gender, education, and ethnicity. Some demographics are overrepresented, such as 91.25% White respondents and 55.38% from the UK, which could skew the results if not properly accounted for. Stratified sampling ensures that these groups are fairly represented,

leading to more accurate and generalizable trends across all demographics. Before starting the analysis, the dataset required cleaning. The is.na() function was used to check for missing values, but none were found, so no data entries were removed. Next, responses indicating consumption of the control drug Semeron (a fake substance designed to identify overclaimers) were removed. Out of the initial 1885 responses, only 1876 remained after this step. The data was then divided into different demographic groups, with 20% of each group selected for the analysis using stratified sampling. To facilitate a more efficient analysis, the drug consumption frequency classes in the dataset, originally labeled CL0 to CL6, were renamed with numeric values. Values like education level and gender were also stored as numeric variables.

Frequency Classes:

0	Never Used
1	Used Over a Decade Ago
2	Used in Last Decade
3	Used in Last Year
4	Used in Last Month
5	Used in Last week
6	Used in Last Day

Gender:

Female	0.48246
Male	-0.48246

Education Level:

Left school before 16 years	-2.43591
Left school at 16 years	-1.73790
Left school at 17 years	-1.43719
Left school at 18 years	-1.22751
Some college or university	-0.61113
No certificate or degree	-0.05921
Prefessionalcertificate/diploma	-0.05921
University Degree	0.45468
Masters Degree	1.16365
Doctorate Degree	1.16365

Figure 1

Analysis

To begin this analysis, it was important to determine the size of each subgroup to understand the distribution of the dataset (Figure 1). To continue obtaining an initial basic analysis of the data, the average consumption for each group, disregarding gender and education level, was found (Figure 2). In this plot, it is evident that the most commonly used drugs are caffeine and chocolate.

	gender_label	education	responses_count
	<chr>	<dbl>	<int>
1	Female	-2.44	2
2	Female	-1.74	8
3	Female	-1.44	3
4	Female	-1.23	7
5	Female	-0.611	34
6	Female	-0.0592	27
7	Female	0.455	58
8	Female	1.16	36
9	Female	1.98	11
10	Male	-2.44	3
11	Male	-1.74	11
12	Male	-1.44	3
13	Male	-1.23	12
14	Male	-0.611	67
15	Male	-0.0592	27
16	Male	0.455	37
17	Male	1.16	21
18	Male	1.98	6

>

However, due to these drugs being completely legal and easily obtainable for any demographic, they were omitted from the rest of the analysis.

The most frequently used drugs, after caffeine and chocolate, are alcohol, nicotine, and cannabis. Similarly, the least frequently used drugs are VSA, heroin, and crack. Further in this analysis, these six drugs will be compared.

However, to get a better initial insight on the trends and patterns in this data, the average overall drug use for each gender and education group was calculated.

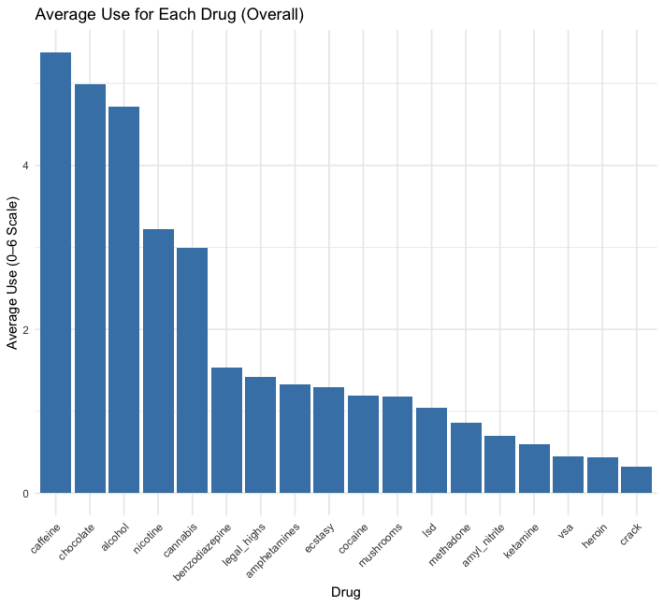

Figure 2

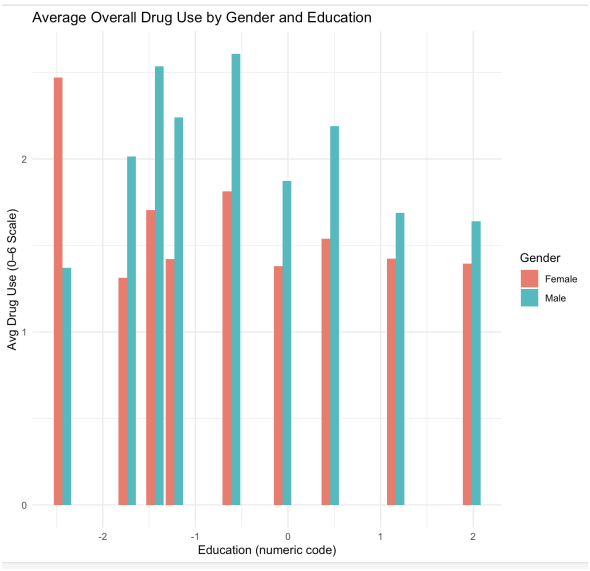

Figure 3

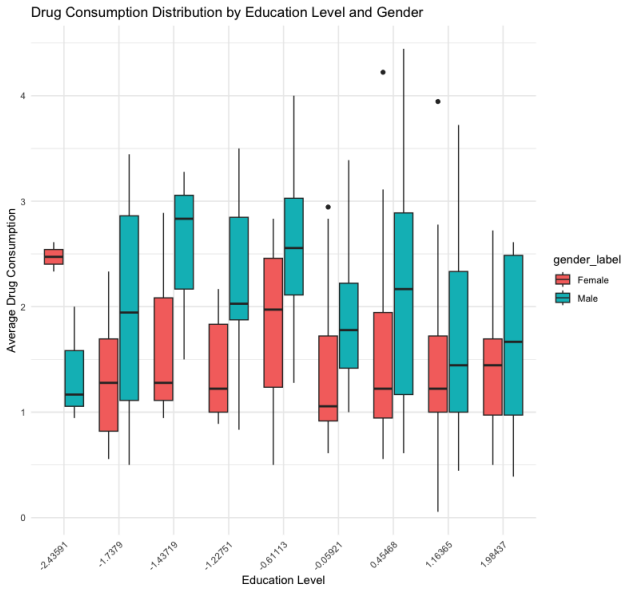

Figure 4

From these plots (Figure 3 and Figure 4), it is evident that males of almost any education level consume drugs more frequently than females. The only exception is among females who left school before the age of 16, where their drug consumption is higher than that of males. However,

this observation is based on only 2 responses from this demographic, making the analysis for this group less reliable. Additionally, it can be inferred that higher education levels generally correspond to lower drug consumption across both genders. With these overall trends in mind, I focused on a detailed comparison of the top three most and least frequently consumed drugs. I then compared these drugs across all gender and education level groups.

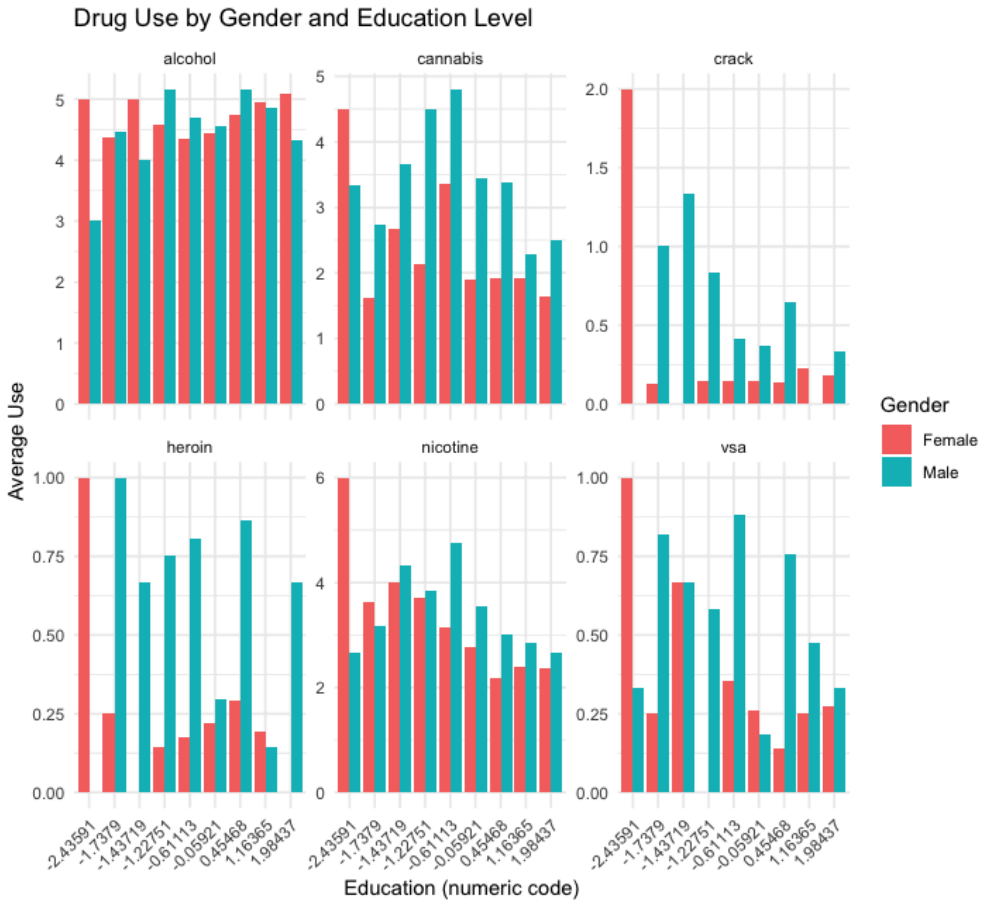

Figure 5

In Figure 5, it is evident that alcohol is the most commonly used drug across both genders, with very similar consumption rates for males and females across all education levels. In contrast, the other five drugs, cannabis, crack, heroin, nicotine, and VSA, show a more significant discrepancy between genders, with males consuming these drugs more frequently than females. Another notable trend in the plot is the impact of education level. As education level increases, drug consumption decreases for both genders. This trend is particularly strong for cannabis, crack, and VSA, while nicotine also shows a decline, although it is less pronounced than the others. To explore this further, we will examine the average drug consumption across different education levels.

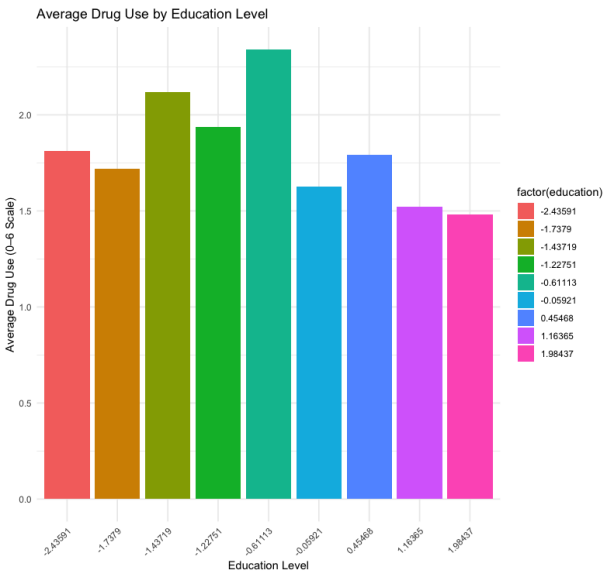

Figure 6

In this plot (Figure 6), we can infer that individuals with lower education levels (indicated by the leftmost bars in the plot) show the highest drug consumption, reflecting the trend that individuals with less education tend to consume more drugs. The education level consuming drugs most frequently is individuals in college or university. This could be due to several social and lifestyle factors, such as the college environment, which may expose individuals to substances like alcohol, nicotine, and cannabis, especially in social settings or parties. While the plot does not differentiate by gender, the consistent trend across education levels suggests that both males and females follows the same general pattern: individuals with higher education tend to consume drugs less frequently. To further analyze these two trends, I conducted linear regression for each of the six drugs listed above (alcohol, nicotine, cannabis, vsa, heroin, and crack).

```
[[1]]

Call:
lm(formula = formula, data = strat_sample)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1241 -0.6816  0.2835  1.0158  1.6573

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.71240    0.06384  73.814   <2e-16 ***
gender      -0.15210    0.13511  -1.126   0.2610
education    0.17049    0.06933   2.459   0.0144 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.233 on 370 degrees of freedom
Multiple R-squared:  0.01716,   Adjusted R-squared:  0.01184
F-statistic: 3.229 on 2 and 370 DF,  p-value: 0.0407


[[2]]

Call:
lm(formula = formula, data = strat_sample)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8870 -2.2020  0.0277  2.0277  4.2094

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2256     0.1209  26.671  < 2e-16 ***
gender       -0.9127     0.2560  -3.566 0.000410 ***
education    -0.5012     0.1313  -3.816 0.000159 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.336 on 370 degrees of freedom
Multiple R-squared:  0.08456,   Adjusted R-squared:  0.07961
F-statistic: 17.09 on 2 and 370 DF,  p-value: 7.971e-08
```

Figure 7

In this regression analysis (Figure 7), the first drug analyzed was alcohol. For alcohol, it was found that education level has a significant impact on consumption. This conclusion is based on the p-value of the education level variable, which is 0.0144, below the set significance level of 0.05. On the other hand, the p-value for gender is 0.2610, which is much higher than the significance threshold, indicating that gender does not have a statistically significant effect on alcohol consumption. The second drug analyzed was nicotine. In contrast to alcohol, both gender and education level were found to significantly influence nicotine consumption. The p-value for gender is 0.000410, well below the significance level, making gender a highly significant factor in nicotine consumption. Similarly, the p-value for education level is 0.000159, also far below the threshold, confirming that education level has a significant effect on nicotine use as well. Unlike alcohol, both gender

and education level strongly impact the consumption of nicotine. Based on the previous trends found above, we can infer that males with a lower education level are the most likely to consume nicotine frequently.

```
[[3]]

Call:
lm(formula = formula, data = strat_sample)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8258 -1.7453 -0.0729  2.0176  4.2547

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9915     0.1089  27.468  < 2e-16 ***
gender       -1.4684     0.2305  -6.371 5.59e-10 ***
education    -0.4622     0.1183  -3.908 0.000111 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.103 on 370 degrees of freedom
Multiple R-squared:  0.1567,    Adjusted R-squared:  0.1521
F-statistic: 34.36 on 2 and 370 DF,  p-value: 2.047e-14


[[4]]

Call:
lm(formula = formula, data = strat_sample)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8161 -0.6126 -0.2597 -0.1159  5.4373

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.45011    0.05055   8.904  < 2e-16 ***
gender      -0.40321    0.10698  -3.769 0.000191 ***
education   -0.07040    0.05489  -1.282 0.200488
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9763 on 370 degrees of freedom
Multiple R-squared:  0.04777,   Adjusted R-squared:  0.04262
F-statistic: 9.281 on 2 and 370 DF,  p-value: 0.0001167
```

Figure 8

The third drug analyzed in this regression was cannabis (Figure 8). Similar to nicotine, both gender and education level were found to have a significant impact on cannabis consumption. The p-value for gender was $5.59 \times 10^{-10,}$ and the p-value for education level was 0.000111; both of these are significantly lower than the threshold of 0.05, making them significant predictors. Like nicotine, we can conclude that males with a lower education level are most likely to consume cannabis more frequently than others. These values indicate that lower education levels and being male are associated with higher cannabis consumption. The fourth drug analyzed was VSA (Volatile Substance Abuse). Unlike the other three previous drugs, the only significant variable was gender, with a p-value of 0.000191. In contrast, education level had a p-value of 0.200488, much higher than the threshold of 0.05, meaning it does not significantly influence VSA consumption. Given the trends observed with the other drugs, we can infer that males of all education levels tend to consume VSA more frequently than females.

Figure 9

```
[[5]]

Call:
lm(formula = formula, data = strat_sample)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8233 -0.6106 -0.2797 -0.1490  5.7988

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.43936    0.05571   7.886 3.53e-14 ***
gender      -0.42423    0.11791  -3.598 0.000364 ***
education   -0.07361    0.06050  -1.217 0.224499
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.076 on 370 degrees of freedom
Multiple R-squared:  0.04364,   Adjusted R-squared:  0.03847
F-statistic: 8.443 on 2 and 370 DF,  p-value: 0.0002598
```

The fifth drug analyzed was heroin (Figure 9). In this regression analysis, similar to VSA, only gender was found to be significant, with a p-value of 0.000364 and an estimated coefficient of -0.42423. The p-value for education level was 0.224499, which is higher than the significance threshold of 0.05, meaning that education level is not considered a significant factor in heroin consumption. As a result, similar to VSA, we can conclude that males, regardless of their

education level, consume more heroin than females.

```
[[6]]

Call:
lm(formula = formula, data = strat_sample)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6812 -0.4643 -0.2509 -0.0889  5.5826

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.32702    0.04625   7.071 7.71e-12 ***
gender      -0.27343    0.09787  -2.794  0.00548 **
education   -0.09125    0.05022  -1.817  0.07005 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8931 on 370 degrees of freedom
Multiple R-squared:  0.03575,   Adjusted R-squared:  0.03054
F-statistic: 6.859 on 2 and 370 DF,  p-value: 0.001189
```

The fifth drug analyzed was heroin (Figure 9). In this regression analysis, similar to VSA, only gender was found to be significant, with a p-value of 0.000364 and an estimated coefficient of -0.42423. The p-value for education level was 0.224499, which is higher than the significance threshold of 0.05, meaning that education level is not considered a significant factor in heroin consumption. As a result, similar to VSA, we can conclude that males, regardless of their education level, consume more heroin than females.

Figure 10

The final drug analyzed was crack (Figure 10). Similar to VSA and heroin, only gender was found to be significant for crack consumption, with a p-value of 0.00548. For the three drugs that were the least consumed in the initial plot, the trend remained the same: education level did not have a significant impact on drug consumption, but gender was a very significant factor. The trend for these drugs follows the same pattern as with VSA and heroin: males consume drugs more frequently than females.

Conclusion

In this analysis, stratified sampling was used to examine how gender and education level impact drug consumption. The drugs analyzed in this study range from legal substances, such as caffeine, to illegal drugs like heroin and ketamine. While caffeine and chocolate are considered drugs in this dataset, they were excluded from the analysis due to their distinct characteristics, as they are not directly comparable to other drugs in terms of their accessibility and use patterns. Additionally, individuals who reported consuming Semeron, a control drug used to identify overclaimers, were also excluded from the analysis. The first trend identified was that males, across most education levels, tend to consume drugs more frequently than females. The second trend showed that higher education levels generally correlate with lower drug consumption, though this pattern is primarily observed for more commonly consumed drugs such as nicotine and cannabis. In contrast, illegal drugs, which are less accessible, were not significantly influenced by education level, indicating that factors other than education may play a larger role in their consumption. In conclusion, both gender and education level have a significant impact on

drug consumption, with males at lower education levels being more likely to consume drugs more frequently than other groups.

References

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