Group 14: Claudia Tung, Will Everett, Alex Latendresse,

Shuhan Wang, Cody Greene, Manu Lopez

DSO 510

December 4th, 2022

How Education Level Influences Cannabis Consumptions

Abstract

Recreational cannabis consumption is slowly on the rise, so our group wanted to understand how education level influences whether someone has tried marijuana. Thus, we investigated whether not pursuing higher education has an effect on the likelihood that someone will consume cannabis within the next month. A data-driven approach is crucial for this topic, as data-based evidence can help eliminate biases and heuristics and better inform us about potential solutions to drug usage trends. We discuss our ideal experiment, which calls for random assignment, legalized cannabis usage and truthful responses, amongst other factors. We performed a regression analysis on a dataset from Kaggle and developed three models, where the dependent variable is level of cannabis consumption, and the independent variables are education level, gender, country and ethnicity. Our findings suggest that we can reject our null hypothesis that not pursuing higher education has no effect on the likelihood that someone will consume cannabis within the next month.

Introduction

Countries around the world are slowly beginning to decriminalize and legalize recreational marijuana. This gradual increase in cannabis consumption warrants more research on the substance, so our group decided to investigate how education level influences whether someone has tried marijuana recreationally. This topic is important because understanding how

education level affects cannabis consumption can guide policy makers and schools in creating a suitable approach to create drug usage policies, teach the side effects and responsible use of marijuana, develop proper drug prevention programs and improve rehabilitation facilities.

Our research hypotheses are as follows:

H_0: Not pursuing higher education has no effect on whether someone will consume cannabis within the next month.

H_a: Not pursuing higher education has an effect on whether someone will consume cannabis within the next month.

Our Approach

Sometimes the eye test can be misleading, and basing business decisions purely on intuition may not yield the most efficient results. The use of data analysis allows us to inform important business decisions, as the reasons for choosing potential strategies are backed up with statistical evidence. The use of data allows for organizations to create attainable and concrete goals, and then in turn optimizing performance with regards to these goals.

However, if these organizations are unable to understand the data generating process, their findings will not be useful. We must understand the limitations of our data, as systematic biases could potentially arise in our models. One example of this is eBay–sellers' reviews are not representative of their satisfaction ratings as buyers who were not pleased with their purchases instead sent negative emails. Understanding the possible omitted variables that may not be present is crucial. If the data is taken as is, the implications taken away are going to be affected, and conclusions and implications will be incorrect.

Once we are aware of our model's data generating process–in this case the relationship between education and cannabis consumption–after best attempting to eliminate these biases, we

are better informed to create these potential solutions. This data driven approach is important in our case; understanding our data generating process allows us to well-inform potential solutions to drug usage trends.

The Ideal Experiment

The ideal experiment would involve people randomly going to college or not so there wouldn't be a bias in any group. Also, if every country had legalized weed that would create uniformity within the data. The United States for example, some states legalized the use of marijana while others have outlawed it, this makes the likelihood of someone using drugs different across the country. Along the same vein, making sure that recreational use is legal compared to just medical purposes. Since the data was collected through surveys, everyone needed to be honest with their reporting.

Our Data

For our data, we have four independent variables. The independent variable we are most concerned about is **Education**, which is described by the highest education level obtained by the participant. **Education** here is a categorical variable, and takes the form of many different categories. These include the variables: *Left school before 16 years, Left school at 16 years, Left school at 17 years, Left school at 18 years, Some college or university, no certificate or degree, Professional certificate/ diploma, University degree, Masters degree,* and *Doctorate degree*. The other three independent variables present are: **Gender, Country,** and **Ethnicity**. All of these variables, like **Education**, are categorical variables. Our dependent variable here is **Cannabis Consumption**, determined in our dataset by the participant's most recent use of cannabis–categories here include *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*.

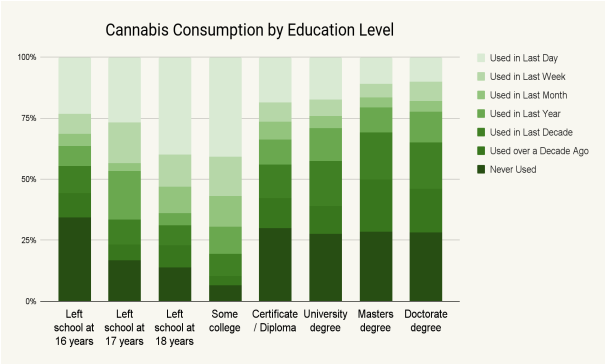
In order to create a more effective model, we underwent extensive data manipulation. This was heavily needed especially because our data here is purely categorical. We manipulated our categorical data through the use of dummy variables. For our exploratory variables, this was straightforward for **Country, Gender,** and **Ethnicity,** but because of the varying inputs of **Education**, we had to simplify our model and, grouping these inputs into just two different categories: Left School (containing people who did not finish high school, or people who finished high school but did not finish university) and Higher Education (those who finished university and have received some sort of degree or professional diploma). The variable, **Left School**, has a dummy score of 1 representing those who Left School, and a score of 0 indicating those with Higher Education.

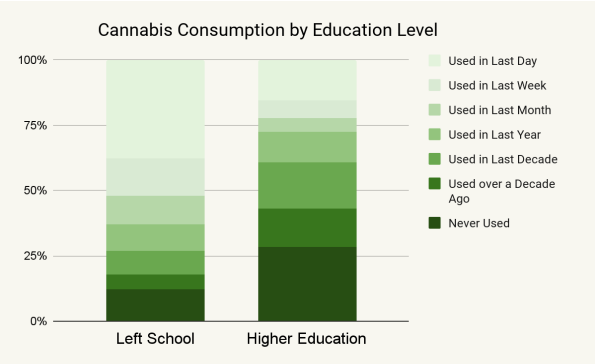
We also had to assign dummy variables for our response variable, **Cannabis Consumption,** and simplified it to just whether or not the participant has consumed cannabis in the last month. The variable, **Cannabis Consumption,** has a dummy score of 1 if the participant has consumed cannabis within the past month, and a score of 0 if otherwise.

Descriptive Statistics

Descriptive statistics helps our group to simplify the large amounts of research data on the relationship between cannabis consumption and the education level into a more sensible form. To better identify the pattern and basis features of the content of the dataset, we utilized a vertical stacked bar graph to explore the frequency of consuming cannabis on each education level and compare the pattern between each education level. Our finding shows that participants who have left school at 18 years exhibit a significantly higher response rate of approximately 38% of consuming cannabis in the last day, whereas only approximately 10% of participants who

have completed a doctorate degree used cannabis in the last day. From a broader perspective, it was also seen that participants with higher education consumes cannabis less regularly compared to participants who left schools on and before 18 years after grouping our education group data into two categories.





Main Findings and Analysis

In order to accurately determine whether there is a statistically significant relationship between education level and cannabis consumption, three regression models are needed to track the efficiency improvement of our analysis. In the first model, only the independent variable *Education* was included, which predicts that having a higher education background will increase the likelihood of consuming cannabis within the next month by 0.3543. However, the R-Squared value of 0.124 implies that only 12.4% variance of the dependent variable Cannabis Consumption is explained by the Education Level, which leads to our inclusion of other factors such as Gender, Ethnicities, and Countries into our second regression model. The second model reflects a higher R-Squared value of 0.325 and a decrease in the coefficient of Education from 0.3543 to 0.1683.

Through conducting further analysis, our group found that when filtering for educational level equal to “Left School”, there appears to be a stark difference in the cannabis consumption by gender. In other words, the impact that education level has on the frequency of consuming

cannabis is highly impacted by gender. Therefore, for our final regression model, we included an interaction term *Education*Gender* as one of the independent variables to capture this effect. The result yields that for participants who are male and have left schools before completing higher education, the likelihood of consuming cannabis within the next month will be 0.1143 + 0.0989 = 0.2132.

Conclusion

By looking at the model and its findings, it suggests that we can reject the null hypothesis that not *pursuing higher education has no effect on the likelihood that someone will consume cannabis within the next month*. In other words, education level affects the likelihood of consumption in cannabis. We were able to reach this decision by; using a data driven approach to eliminate potential biases, understand the data in more depth by visualizing, change the data more palatable for the objective by data cleaning and manipulation, run the regression analysis multiple times to compose a precise model. However, to conduct an ideal experiment, there were some limitations in our data. First is that it would have been better to have more diversity in countries and ethnicities. There are only seven different countries in the sample and the majority of it is data from the USA and UK. In terms of ethnicities, the data is 90% white and it would have been better to have had it balanced out by having more data of other ethnicities. Another limitation is the data of legality. We would be able to understand the data more precisely if we can know whether it's legal or not in each country or state where the data was collected. Finally, an overall bigger sample size would be better to have smaller standard deviations and conduct more accurate analysis.