STEM Fellowship Big Data Challenge 2019

Countries' Control on Cannabis

Minh Nguyen, Ivy Liang, Alice Wu

University of Toronto

June 7, 2019

Abstract

This paper analyzes the relationship between a geographical location, its prevalence of cannabis use and its corresponding societal factors. Such factors include a country's GDP, health expenditure, life expectancy and democracy index. Despite many people talking about cannabis and its adverse effects in casual conversation, there is not a lot of statistical research on geographical spaces and their respective cannabis use. These correlations are particularly useful in the interest of legalization of marijuana but also general public curiosity. Datasets were obtained from organizations such as the UN Office of Drugs and Crime and The World Factbook - CIA. These reputable datasets were analyzed using software tools from R. Statistical analyses were constructed to interpret the correlation coefficients and general shape of the graph. Such analyses were conducted using large geographical regions as well as individual countries. The datasets indicated a correlation between an individual country's life expectancy, GDP and health expenditure. This analysis could also be used as a guideline for countries who have already legalized cannabis and want to potentially lessen harmful societal effects or countries wishing to legalize cannabis in the future. This could also be applicable to other drugs, whether illegal or recreational such as alcohol. Overall, this study indicates the correlation between a specific geographical location, its prevalence of cannabis use, effects on society and society's effects on cannabis use.

Keywords
cannabis, societal, impacts, GDP

1 Introduction

Cannabis is the hot topic of both casual conversation and scientific advancements in areas all over the world among various groups of people. Especially with the recent legalization of marijuana in Canada, individuals are quick to assume the prevalence of cannabis use in Canada. But how does the use of cannabis actually vary based on a country's societal factors? Some significant factors include a country's Gross Domestic Product (GDP), life expectancy, number of internet users and democratic status. Our analysis focuses on social science topic areas, primarily ways that certain factors influence commun. Since Altmetric data can be quantified, the results are much easier to display. Since the goal of

our analysis was to demonstrate the correlation between prevalence of cannabis use and various societal factors, Altmetric data was the appropriate fit for the conclusions we wanted to make.

There were a lot of data sources that fit our general objective, but we decided on two very reputable and organized data sets. We obtained data from the United Nations Office on Drug and Crime [1] about prevalence of drug use. The UN Office on Drug and Crime do a lot of research on drug use, abuse and its relation to crime. They collect a lot of data on a variety of drugs, including cannabis. This fit perfectly with our intention of analyzing drug use and societal factors. Additionally, we chose to use the World Factbook [2] for data on societal factors which includes a country's:

- Gross Domestic Product (GDP)
- Life expectancy
- Number of internet users
- Democratic status
- Health expenditure
- Education expenditure

As stated on their website, "The World Factbook provides information on the history, people, government, economy, geography, communications, transportation, military, and transnational issues for 267 world entities" [2]. This information is published under the Central Intelligence Agency (CIA), which is certainly very reliable.

A variety of statistical analyses were carried out using R with the above data. Since we were trying to determine the correlation between two quantitatively measured values, we decided to use linear regression. Our initial hypothesis was finding the relationship between geographical regions of the world and their respective cannabis use. But after some data wrangling, we discovered that there did not seem to be any significant correlation for each of the regions. This prompted us to do some additional statistical analysis on each country's prevalence of drug use and societal factors.

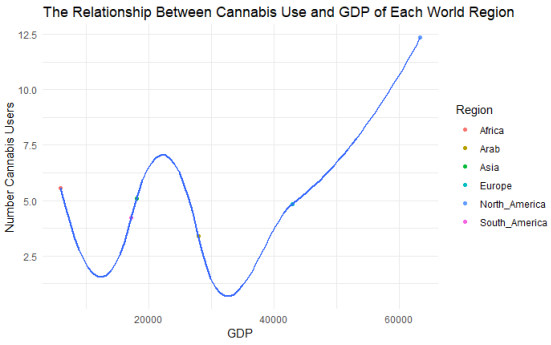
2 Materials & Methods

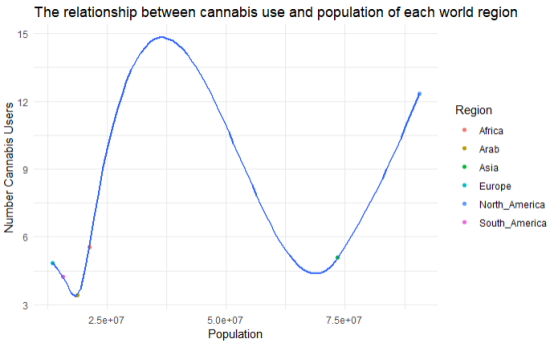
R was the statistical tool used to analyze and visualize the relationship between the prevalence of cannabis use and societal factors in both regions of the world and individual countries. Since we wanted to determine if there was a general global trend, we plotted each societal factor as a separate graph with the geographical regions. Throughout the analysis and visualization, we used the *tidyverse, dbplyr* and *ggplot2* libraries.
The data wrangling process started off by extracting all the data from their respective csv formatted files into data frames using the *read* function from the *tidyverse* library. We then proceeded with our initial hypothesis, and filtered the countries into 6 distinct world regions, including Europe, Africa, North America, Arab States, Asia Pacific and South/Latin America. Our end goal was to have an average of each societal factor for each region. So we then joined each societal factor table with all of the 6 regions' cannabis use and found the average in each region by dividing by the number of countries in that respective region. We did this for each region and each societal factor. For the data wrangling of our second hypothesis, we used a similar method of analysis. We joined each societal factor table with the cannabis use in that country along with using the *lm* function in R to create a standard linear regression model.

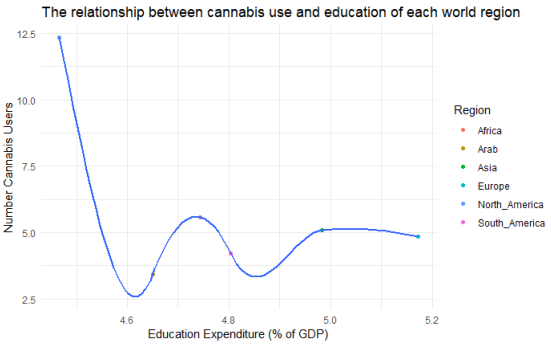
The data visualization process composed of using the *ggplot* function in the *ggplot2* library. We used a scatterplot with a line of best fit for all of the plots,

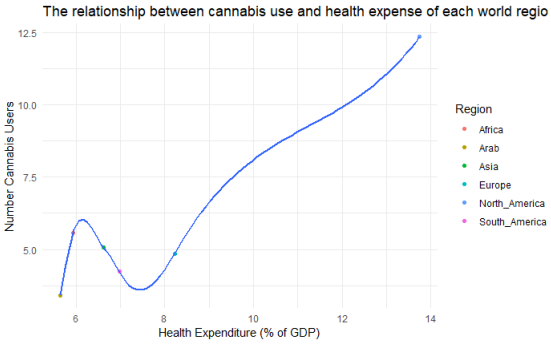
with *loess* as the smoothing function and the use of *se* to display a confidence interval around the fit line.

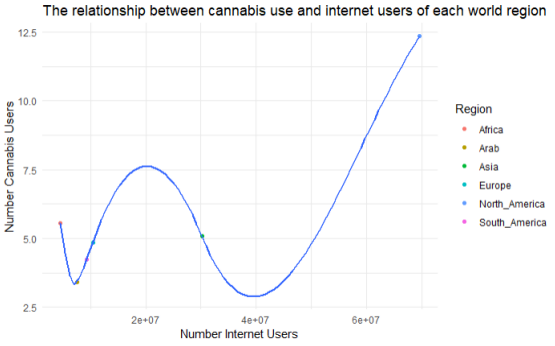
3 Results

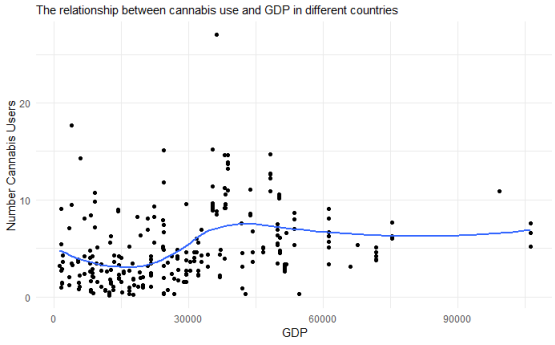


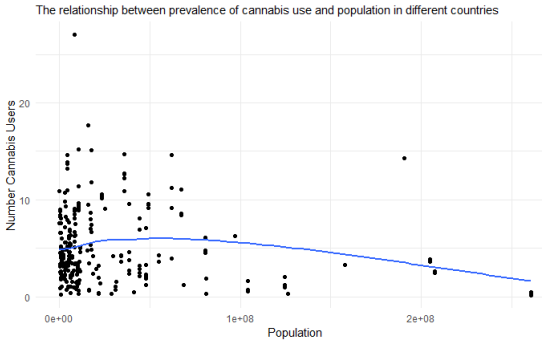


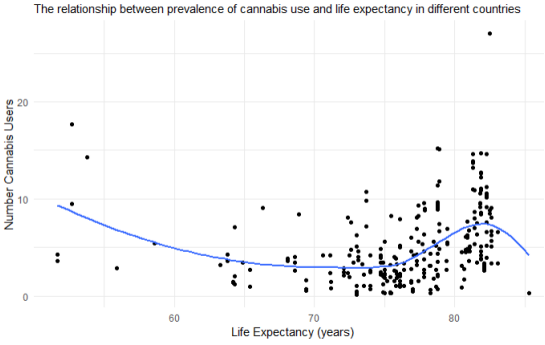


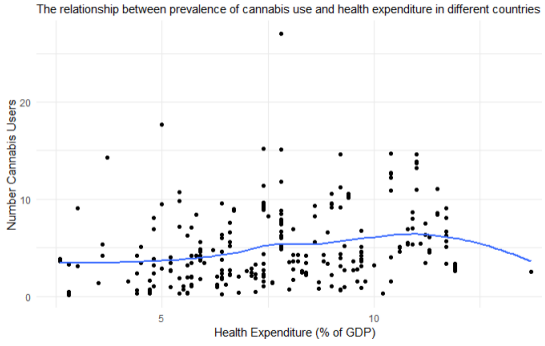












As shown in the graphs on the previous page, the does not seem to be any correlation between geographical regions, so-

cietal factors and prevalence of cannabis use. On the other hand, there does seem to be some correlation for individual countries, societal factors and prevalence of cannabis use. All four factors displayed above show a relatively direct correlation between the variables.

4 Discussion

Our initial hypothesis involved examining the relationship between geographical regions and their prevalence of cannabis use based on a variety of societal factors. But after analysis it is clear that these variables do not have a direct correlation with each other. The graph is oscillating between a large range of values, none of which particularly indicate a trend.

Our second hypothesis included examining the relationship between individual countries and their prevalence of cannabis use based on a variety of societal factors. This lead to a much more interesting result. The plot for the relationship between cannabis use and GDP in different countries showed that the lower a country's GDP, the less cannabis users they had. Though there are certainly exceptions, there is a clear cluster near the beginning of the data, which indicates a right skewed, unimodal distribution. This is an interesting result as we hypothesized that the higher a country's GDP the more cannabis users they would have as individuals have the resources to spend on recreational drug use. But our data says otherwise. Furthermore, for the relationship between prevalence of cannabis use and population in different countries, there is an even stronger cluster near the low end of the data, further described as a right skewed distribution. This indicates that the lower the population, the lower the cannabis prevalence. This is expected as smaller countries would purely have lower cannabis users than large countries, even if their entire country used cannabis. Additionally, the relationship between life expectancy and cannabis use is on the other side of the spectrum. The graph indicates that the data is left skewed, bimodal distribution, which indicates that the higher the life expectancy the more cannabis users in the country. This is interesting as cannabis is predominantly advertised as bad for one's health, but clearly countries who have a higher life expectancy also have more cannabis users. Finally, the relationship between cannabis use and health expenditure is an interesting result. The graph shows a relatively symmetric distribution of data.

There are many potential sources of error. As mentioned above, the population graph is extremely biased as even if country A had 10 inhabitants and all 10 were cannabis users, but country B had 100,000,000 inhabitants, but only 11 were cannabis users, the data would still indicate that country A had less cannabis users. Although this fact is true, this introduces a confounding variable into the data that is misleading. This fact also translates to why our initial hypothesis did not provide good results. The populations of each individual geographical region were so diverse that the data showed that variance. Moreover, our data may be skewed due to disclosure of information. Since cannabis is not legal everywhere around the world it is difficult to obtain data for countries in which cannabis is illegal. Whether the UN office collects data from emergency room patients or a self disclosed survey, the results are bound to be different. Very few people would voluntarily admit the fact that they used a recreational drug, as it might have severe consequences. Therefore, countries who appear to have a low cannabis prevalence rate, might in reality have a much higher one.

Conclusions

This analysis generates a lot of insight on the prevalence of cannabis use and its adverse effects. It is frequently discussed in a casual setting what the impacts of cannabis are, but now there is statistical evidence to back up people's assumptions. This is particularly useful for countries who are looking to perhaps legalize marijuana. Now they can examine how, for example, their GDP would change in terms of how many marijuana users they have. This could also be used by countries who have already legalized cannabis and want to potentially lessen harmful societal effects. This study guideline could also be applicable to other drugs, whether illegal or recreational such as alcohol. Perhaps an interesting topic would also be to investigate the relationship between firearms and various societal effects, as this seems to be a largely growing issue in today's society along with drug abuse.

References

[1] *Annual prevalence of drug use*. UNODC - United Nations Office on Drugs and Crime. Web. Accessed May 24, 2019. https://dataunodc.un.org/drugs/prevalence_table

[2] *The World Factbook*. Central Intelligence Agency. Web. Accessed May 24, 2019. https://www.cia.gov/library/publications/resources/the-world-factbook/