MICHIGAN STATE
UNIVERSITY

Drug Use Data Examination

CMSE 381

Section 001 SS21

Samantha Howard

80

Final Project

1. Introduction

Drug use is prevalent in all age demographics. This project evaluates the trend between the American population and use of cannabis. The populations are broken up into different age brackets for two data sets. One being from a related work, and one, more expansive set, pulled the other from a data archive. In this project the original results from the related work were recreated and then using the larger set to make a Support Vector Machine (SVM) model to see the correlation between age and use.

2. Related Work

In a related work, from a FiveThirtyEight article titled “How Baby Boomers Get High” shows a graph of the percent of the population aged 50 to 65 in the United States that were reported using various drugs in a 2012 survey. The graphs list the drug usage in decreasing order as a histogram plot that demonstrates the percent of the participants that used these drugs within the last twelve months. Below is a recreated graph of the table in the article.

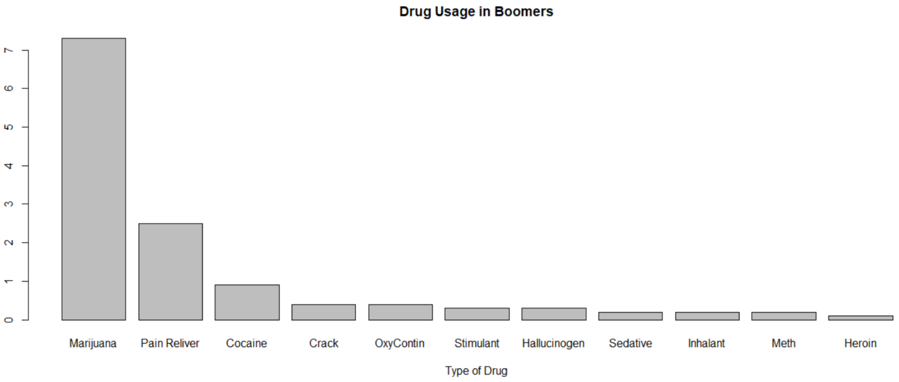


Figure 1: Drug Usage in Boomers Table

Notably alcohol was removed from the table due to its excessive use that the article states “dwarfed the differences among other drugs” (Barry-Jester). This is due to the consumption rate of alcohol among the demographic being 84.2 % and the next highest being marijuana at 7.3%. Included is that dwarfed result for comparison in figure 2 to demonstrate how its skews the perception of the data.

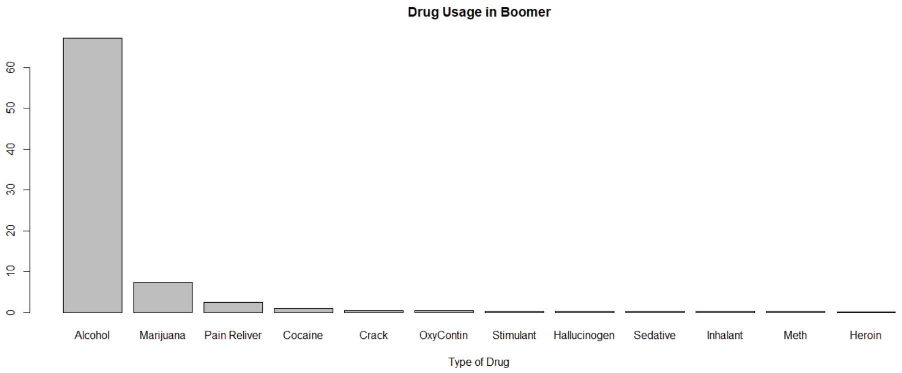


Figure 2: Drug Usage with Dwarfed Data

3. Dataset

Using two different data sets that were provided by FiveThirtyEight and another one provided by UC Irvine Machine Learning Repository both covering various drug usages across all ages. The dataset hosted by FiveThirtyEight is rather small, containing only the average population and frequency of drug use for 17 different age groups spanning across the range of 12 to 65+. The data set hosted by UCI expands on this concept by having individual entries, as opposed to an already calculated average that amounts to nearly 2000 different individuals surveyed.

4. Methods

In the model it was investigated as what was the correlation between age and users, with the assumption that perhaps younger adults would be among the highest users and that elderly would be among the lowest. Figure 3 is a graph of the use and frequency of use of cannabis over the last year, from the data supplied from FiveThirtyEight. With use being the percent of users reported using it and frequency being the average number of times it was used from those that use cannabis. As the graph indicates, the percent of users rises to peak at approximately young adults and then gradually have less users as they age. However, that prediction didn't cover the general trend of sharp rise in the frequency as the demographic ages. Notably, ages 30-40 had the highest average reported use per twelve months at over 80; this is significantly higher than any age range that classifies as young adult or teen.

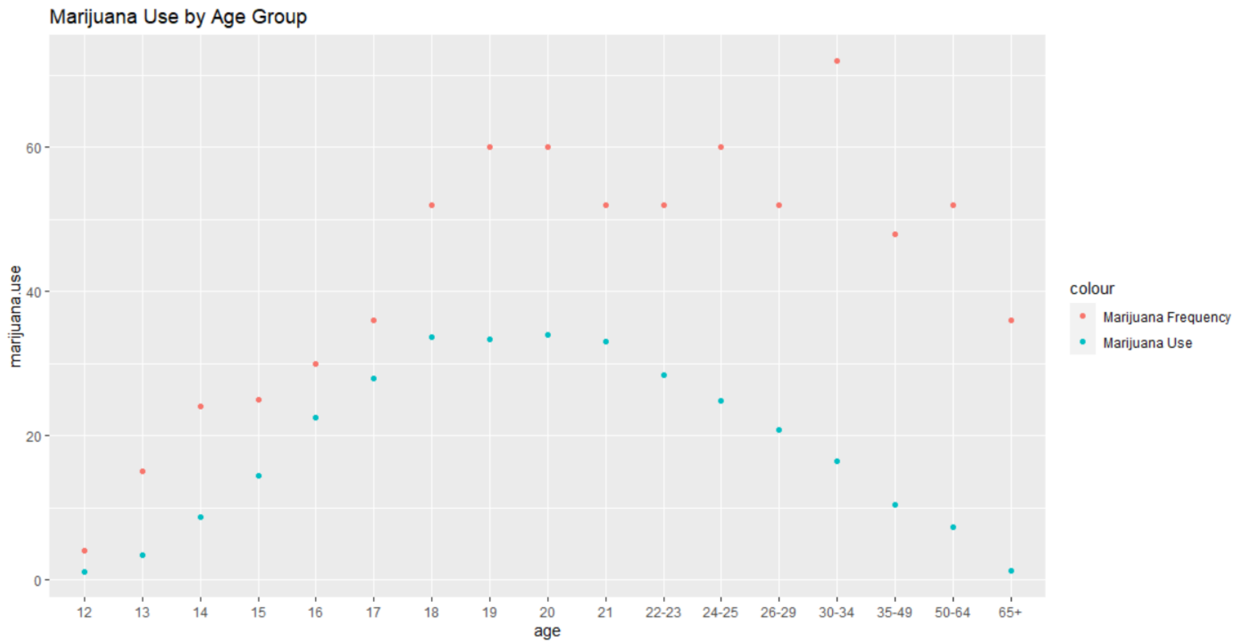
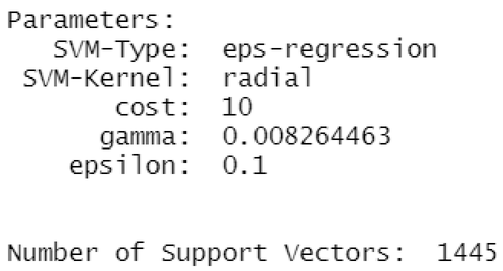


Figure 3: Plot of Marjuana Use by Age Group

Initially using a SVM model, the concept was to label cannabis users from non cannabis users from the UCI data. Non cannabis were classified as those that never used or or have used it once in the past. Cannabis users were the rest of that data, consisting of any given amount used within the given time period of the survey. Converting the seven categories of substance in the data into scalars that can be better modeled with. These values range from a low of 0, indicating the participants have never used cannabis to 6, meaning it was used as recently as yesterday. These then were identified as the y-values for the model, with the age being the x-values, with the model being constructed accordingly.

5. Results

From the two produced SVM models, both were shown to have very similar results. In the first model the results are shown in figure 4. The primary difference between the two models is this one specifying the kernel as linear was left out, so it defaulted to a kernel labeled as “radical”. The first model then yielded these results below.

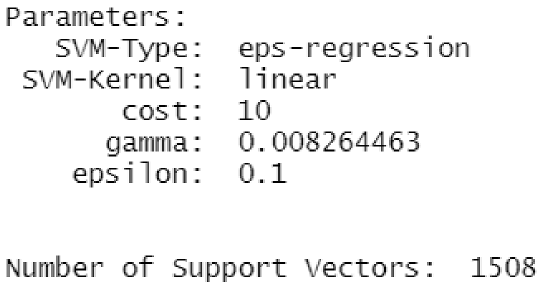


```
Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  10
      gamma:  0.008264463
    epsilon:  0.1


Number of Support Vectors:  1445
```

Figure 4: SVM model results

For the second model, featured in figure 5, the kernel was specified as linear and yielded similar results. Notably, the only change between these two were the count of support vectors identified. As both models have the identified gamma values as being identical in each case.



```
Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  linear
       cost:  10
      gamma:  0.008264463
    epsilon:  0.1


Number of Support Vectors:  1508
```

Figure 5: SVM linear results

6. Conclusions

Through various plots and SVM models the relationship between age and cannabis use was investigated. Two similar plots to the one in the related works from FiveThirtyEight was recreated, along with another plotting the general trends of use of cannabis with age and the frequency it is used. It was identified that young adults have a higher percent of their population that use, as compared to their elders in the age bracket of 50 to 65. However, those young adults are not the population with the highest frequency. The highest frequency is the 30 to 34 year old age group, with reporting an average of over 80 uses in the last twelve months.

7. References

Barry-Jester, A. M., &amp; Flowers, A. (2015, April 23). How baby boomers get high. Retrieved April 14, 2021, from https://fivethirtyeight.com/features/how-baby-boomers-get-high/

drug-use, (2018), GitHub repository, https://github.com/fivethirtyeight/data/ blob/master/drug-use-by-age/README.md

E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan and A. N. Gorban, "The Five Factor Model of personality and evaluation of drug consumption risk.," arXiv [Web Link], 2015