82

Analysis of Drug Use by Age

Final Project Report

1 Introduction

Drug use has been an important topic of interest for many years. Knowing the frequency of drug use within a particular age group can help rehabilitation companies and other businesses better market to the deterrence of these particular drugs. Without any knowledge of the percentage each drug is used, a company may be wasting its resources. On top of that, less people will be inclined to receive help. The goal of this project is to analyze the percentage of drug use within age groups and evaluate the meaning behind the results. Specifically, this project will look into the analysis of drug-use for baby boomers and dive in further to drug-use between the ages of 12 and 21 by predicting if there is a spike in drug use percentage at a certain age and what age has the most users.

2 The Data Set

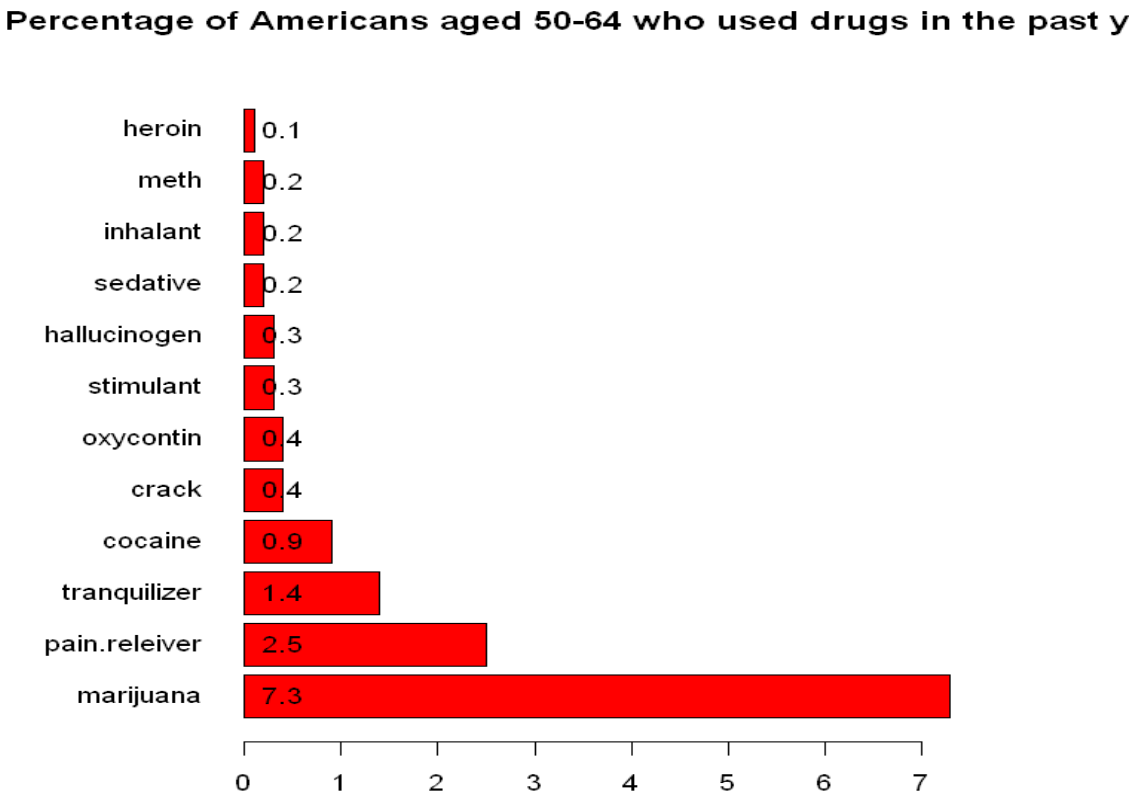
The data set used in the project can be found on the *FiveThirtyEight* website and is entitled "Drug-Use by Age". The source is a survey done in 2012 asking people to indicate which drug they have done in the past year, and how many times they have done each drug during that period. The survey includes alcohol, marijuana, and many types of illicit drugs. Here is the data in a table:

age	n	alcohol.use	alcohol.frequency	marijuana.use	marijuana.frequency	cocaine.use	cocaine.frequency	crack.use	crack.frequency	...	oxycontin.use
12	2798	3.9	3	1.1	4	0.1	5.0	0.0	-	...	0.1
13	2757	8.5	6	3.4	15	0.1	1.0	0.0	3.0	...	0.1
14	2792	18.1	5	8.7	24	0.1	5.5	0.0	-	...	0.4
15	2956	29.2	6	14.5	25	0.5	4.0	0.1	9.5	...	0.8
16	3058	40.1	10	22.5	30	1.0	7.0	0.0	1.0	...	1.1
17	3038	49.3	13	28.0	36	2.0	5.0	0.1	21.0	...	1.4
18	2469	58.7	24	33.7	52	3.2	5.0	0.4	10.0	...	1.7
19	2223	64.6	36	33.4	60	4.1	5.5	0.5	2.0	...	1.5
20	2271	69.7	48	34.0	60	4.9	8.0	0.6	5.0	...	1.7
21	2354	83.2	52	33.0	52	4.8	5.0	0.5	17.0	...	1.3
22-23	4707	84.2	52	28.4	52	4.5	5.0	0.5	5.0	...	1.7
24-25	4591	83.1	52	24.9	60	4.0	6.0	0.5	6.0	...	1.3
26-29	2628	80.7	52	20.8	52	3.2	5.0	0.4	6.0	...	1.2
30-34	2864	77.5	52	16.4	72	2.1	8.0	0.5	15.0	...	0.9
35-49	7391	75.0	52	10.4	48	1.5	15.0	0.5	48.0	...	0.3
50-64	3923	67.2	52	7.3	52	0.9	36.0	0.4	62.0	...	0.4
65+	2448	49.3	52	1.2	36	0.0	-	0.0	-	...	0.0

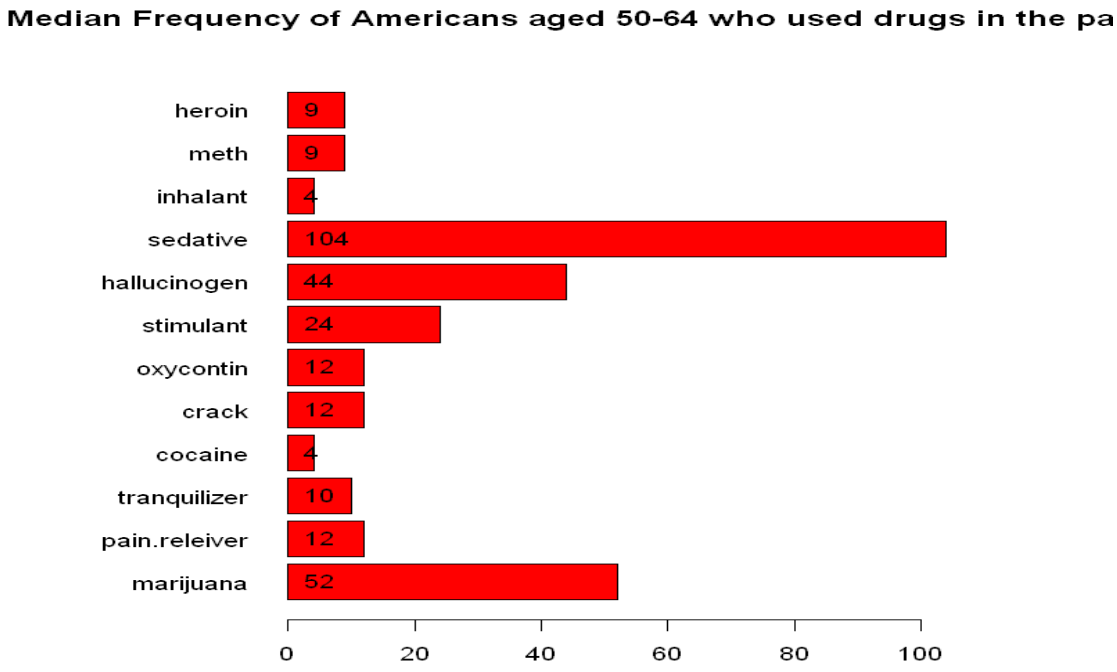
As shown, the data is more generalized and mainly age-focused; However, this dataset can be useful in answering many age-based questions on the frequency of drug use.

3 Related Work

An existing piece of work examines the percentage of each drug-use for baby boomers using this set of data. The study was created since there were many reports of higher rates of illicit drug use among the generation of baby boomers. Looking at the analysis of the data below it is clear that the most common drug use other than alcohol is marijuana.



The article also mentions the importance of the frequency of use for each drug. Trying or doing a drug once or twice is not the same significance as frequently using the drug. By comparing the frequency of each drug, it can be seen that the drugs with a higher percentage of use, do not always have a higher frequency.



Even though there was a low percentage of members within this age group doing sedatives, the median number of times it was used is 104. That is about once every three days, which is a cause for concern.

4 Methods for Further Analysis

Perhaps a more interesting and important question is to evaluate drug use between the ages 12 and 21. There has been substantial evidence over the years that drug use before a person's brain is fully developed could cause permanent brain damage. So, a good question to evaluate is when does the percentage of drug use peak between these ages and is there a spike in usage percent at any age period. The focus will be on marijuana usage.

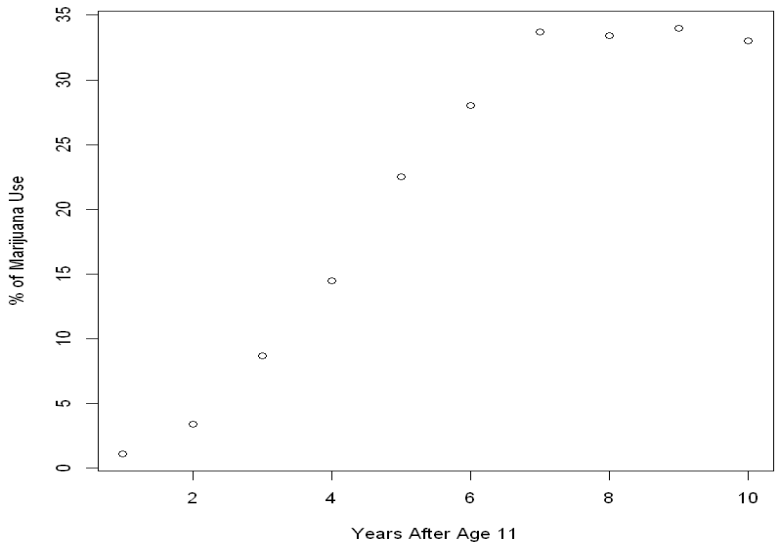
One way to visualize this question is to use a polynomial regression. This replaces the standard linear model with the following polynomial function:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i$$

Where epsilon is the error term and xi's are the predictors. In R, the data will be fitted to this model predicting marijuana usage to age. The which.max() function will give the highest percentage value and can be indexed to give the corresponding age that is predicted to have the highest percentage of users.

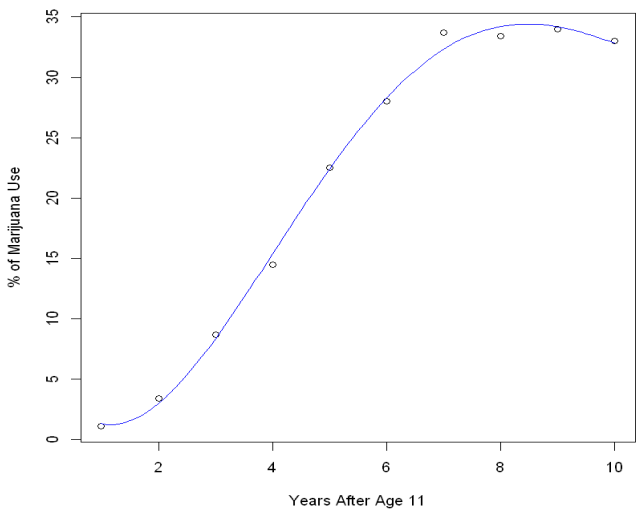
5 Experiment and Discussion

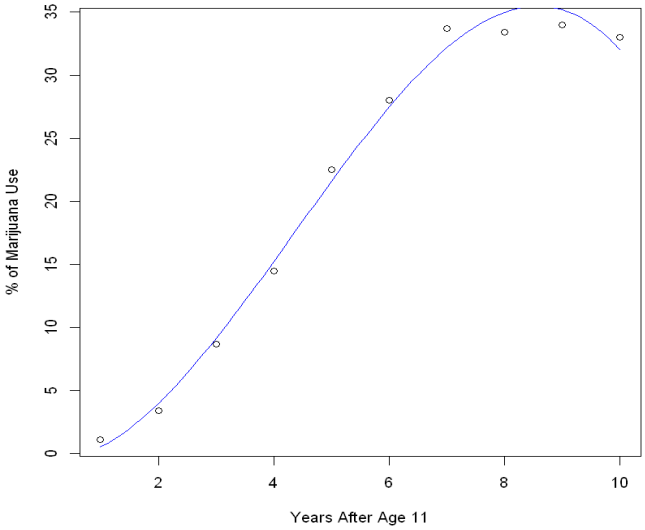
The following is a scatter plot of the data points for marijuana usage between the age of 12 and 21:



By looking at this graph, it can be reasonably seen that the growth rate is not linear. It was best to fit a polynomial model between degree 2 and 4.

Polynomial Regression was used to evaluate the question at hand. The two polynomials fitted were degree 3 and degree 4. The best degree polynomial to fit the data was a polynomial of degree 4. The p-values and adjusted r-squared were looked at to determine if it was a good fit. The outcome of the fit is featured below with degree 4(left) and degree 3(right):





There was no significant jump in percentage of marijuana use between these ages. However, the highest jump in usage occurs from age 15 to 16. It can also be seen that the percentage levels out around seven years after the age of 11 (age 18). This could mean two different scenarios: people tend to not try marijuana after the age of 18 is they have not already, or certain users quit while other begin to try the drug after this age. Since the data set does not look at the same people over the years, it is hard to determine which possibility is accurate.

Based on the prediction the maximum point on the function is 19.5. This means that the highest percentage of users occur at the age of 19 - almost 20 - years old. The great thing about a fitted polynomial function is that the beta coefficients found from lm() can be used to predict drug use percentage in future years. For example, the 3 degree regression polynomial had the equation: y = -0.57 - 0.3576*x + 1.5549*x^2 - 0.1193*x^3, which would actually predict a decrease in drug usage looking five years into the future for the same generation.

6 Conclusion

Overall, it can be concluded that for ages 12 to 21, the largest spike in percentage of drug use is age 15 to 16. Also, between ages 19-21 there is less of a growth in usage. It could even be predicted that this same generation would show a decrease in drug usage after the age of 21.

This evaluation can be crucial when it comes to deterring individuals from trying marijuana. Businesses/companies like *Truth*, a company that advises against drugs, can use this data to market to the right audience age. In this case, they could advise in a certain way to people around the age of 15 the costs of pursuing marijuana and maybe there will be less of a jump in usage.

7 References

[1] https://fivethirtyeight.com/features/how-baby-boomers-get-high/

[2] https://github.com/fivethirtyeight/data/tree/master/drug-use-by-age

[3]https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6062a083acbfe82c7195b27d/1617076404560/ISLR%2BSeventh%2BPrinting.pdf