**CMSE381 Final Project**

**1 Introduction**

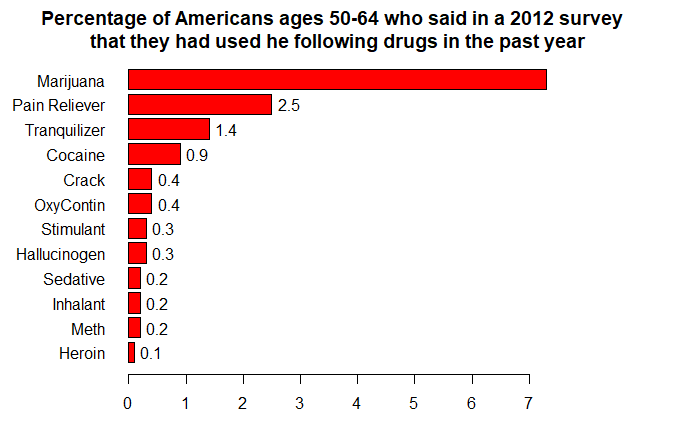
In this project I looked at the drug\_use data set from fivethirtyeight.com and looked at several of the variables in a sense to try and replicate the results of the attached website and also try to draw some conclusions of my own interests. After looking through the data set, I decided I thought it would be an interesting idea to see if there was any correlation between cocaine use and marijuana use in adults. I thought this would be an interesting thing to investigate since these seemed to be two of the most popular drug choices among most ages of those sampled in this survey. I also looked into predicting cocaine\_use from the drug\_use dataset.

**2 Dataset**

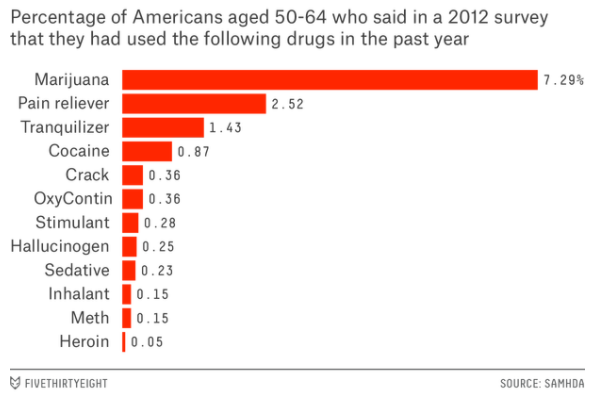
This dataset somes from a national survey on drug use and health from the substance abuse and mental health data archive. The data was collected in 2012 and taken from people of age 12 all the way to 65+. This survey seemed to focus more on the younger population rather than the older half of the population. The data set collected all sorts of drug use data from these individuals from more common drugs such as alcohol or marijuana, all the way to surveying them about their use of heroin and tranquilizers. In total the participants of the survey were asked to respond to their use of 13 different types of drugs. The data was made into a way that had columns of percentage of people that had used a specific drug in the last 12 months and had another type of column that had the median number of times a person used a drug within the past 12 months.

**3 Replication of Website Data**

For this project we were given an option to pick between four different topics to study, each coming with some prior knowledge that has already been plucked from the data. Here I am replicating the data that is from the article **“How Baby Boomers Get High”**. This article focuses on the older population that was surveyed in this dataset, which is a small selection and looks at what the most popular methods of drug use among this group. Here is my replication of their plot attached at the top of their article.



Here is the same plot that I was replicating from the website.



**4 Methods**

For this project I decided to use two different methods of machine learning models to find which method would work best. We compared these methods of classification against a baseline of linear least squares regression to see if this simple method could still be better than the more complicated methods, or hopefully that the other methods would find a more accurate result.

**4.1 Cross-Validation**

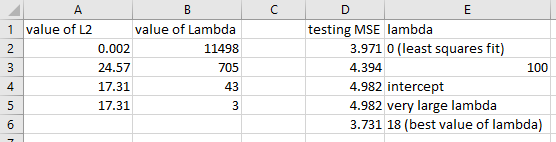
Here I looked to use the Validation Set Approach, Leave-One-Out Cross-Validation and k-fold Cross-Validation. I thought cross validation would be a great way to model this type of data since we would be dealing with a lower bias than we normally would using another technique. This technique also has multiple different instances we can use within the overlying title of cross-validation such as K-folds and Leave-One-Out. Through cross validation it enabled me to find the best optimized value of lambda for my further tests to compare the correlation between cocaine use to marijuana use.

**4.2 Ridge Regression and The Lasso**

Ridge regression I thought was a great choice for predicting things in this data set since I figured from the beginning that some of the variables that I would be looking at would be linearly related. Ridge Regression offers a way to reduce the overfitting that maybe a more complex model would do, but at the cost of introducing some bias, another reason why I paired this and cross-validation together. The ridge regression allows me to optimize the regularization of the coefficients made by the linear regression and push the estimated coefficients towards zero.

**5 Experiments and Discussion**

For the ridge regression evaluation of the dataset one of the things I used to test how accurate my predictions of my model were was mean squared error (MSE) and the L2 value. As we would expect, the L2 value increases to an extent as the value of lambda decreases. In my case we can see that this value seems to level off after lambda hits a value of around 100. As for the testing MSE, I tested many values of lambda ranging from 0 (which is the same as the line of least squares fit) to a very large number. It seems in this case that the testing MSE also levels off after the value of lambda hits a threshold in the same fashion the value of L2 does.



**6 Conclusion**

From these findings I have come to the conclusion that the correlation between cocaine use and marijuana use is definitely linear. I did have a hunch from the beginning that this relationship was linear, but I did feel that there was a chance there were some more factors at play than would be used in the simple least squares fit. I found by looking at this relationship there was no clear advantage of using any higher order polynomial like a quadratic or cubic fit using the validation set approach or the leave one out cross validation approach. But when we look at the k-fold cross validation approach where we look at higher order polynomials all the way up to 10th order, we find different results. Some of the higher order polynomials seem to have lower testing error but I see this as a result of over fitting of the data, especially since the data set im working with does not have a lot of observations to work with. Overall I found the relationship between cocaine use and marijuana use to be overall linear with the best machine learning modeling result coming from the ridge regression method while utilizing cross-validation.

**7 References**

<https://www.pluralsight.com/guides/linear-lasso-and-ridge-regression-with-r>

<https://www.statlearning.com/> (Class TextBook)

[https://www.geeksforgeeks.org/](https://www.geeksforgeeks.com/)

<https://www.stackoverflow.com/>

<https://fivethirtyeight.com/features/how-baby-boomers-get-high/>