**Statistics 141A Final Project**

**Due: December 13, 2019**

**Professor Farne**

**Group #23**

**Collaborators:**

-Ryan Smith: Contribution : linear regression, basic data eploration, KNN classifier, LDA classifier, Logistic regression classifier, question 1, question 2, correlogram, multinomial LDA,
-Tony Linou:
-Dale Urbelis: Logistic regression models and plots, classification

Note: Our original data was deemed unusable by both the TA Amy Kim and Professor Farne after we had made our proposal, so a new data set and questions were permitted by Professor Farne, and a modified proposal more suited to the new data set is attached with this submission.

**Background:**

In our project, we use the automobile MPG dataset collected in 1983 by Ross Quinln and stored by Carnegie Mellon University, and also by University of California, Irvine, in their machine learning data set archives. The purpose of this data set is to provide data regarding mulitple vehicle factor's effects on the miles per gallon of each particular car. The data set contains 9 pieces of information on 398 vehicles produced between the years 1970 and 1983. The nine pieces of information on each of the 398 observed vehicle are: the fuel efficiency of the vehicle in the miles per gallon, the number of cylinders the engine of the vehicle has, the engine displacement(volume of the engine) in cubic centimeters, the horsepower of the vehicle, the weight in pounds of the vehicle, the acceleration of the car, quantified in terms of the time it takes the car to go from zero miles per hour to sixty miles per hour(0-60mph), the year the car was produced, the origin of the car, and the name of the car. Six of the observed vehicle characteristics in our data set are quantitative, mpg, number of cylinders, displacement, horsepower, weight, and acceleration. The remaining three are qualitative, and describe the make and model of the car, rather than being measured: the year the car was produced, the car origin, and name of the vehicle. Understanding the relationships between various characteristics of a vehicle that are controlled by the manufacture can give insight into how to exploit some of the features to improve the more desired characteristics, such as modifying the displacement of the engine in order to improve miles per gallon or accceleration.

**Statistical Questions:**

The initial question we seek to answer in this report is : 'What effect does the number of cylinders, and acceleration(0-60mph) of a vehicle have on its miles per gallon(MPG)?' After we answer this initial data exploration question, we will move onto our bigger goal. The primary goal of our project is to predict the class of miles per gallons(low, medium, high, etc) of vehicles based on their number of cylinders, engine displacement, weight, and acceleration. Our final question is which factors affect the miles per gallon of a vehicle the most?

**Methodology:**

To begin exploring the relationship between the explanatory and response variables, we will utilize several linear regression models, along numerous plots of the various variables. These plots include correlograms, scatterplot matrices,linear regression plots, and histograms. To answer the first question, we will make a linear model to predict the mpg versus acceleration + number of cylinders. We will analyze whether or not the normality assumption holds by plotting the QQ-plot of the residuals.We will then make a scatterplot matrix of the mpg, acceleration, and number of cylinders to find the basic relationship between the 3 variables. To

further analyze this relationship we will run numerous correlation tests, To answer the second question, which is regarding the predicted class assignments(good mpg, bad mpg, medium mpg) to the vehicles, we will explore various potential classification models. These models include performing logistic regression classification, linear discriminant analysis, and kth nearest neighbor classification. To answer the third question, regarding the most impactful factors on a vehicles' mpg, we will explore corellograms, in addition to several single factor linear model and multiple factor linear models. We will also analyze the correlation of each factor with MPG, to find which ones have the strongest positive or negative relation with MPG.

## Analysis

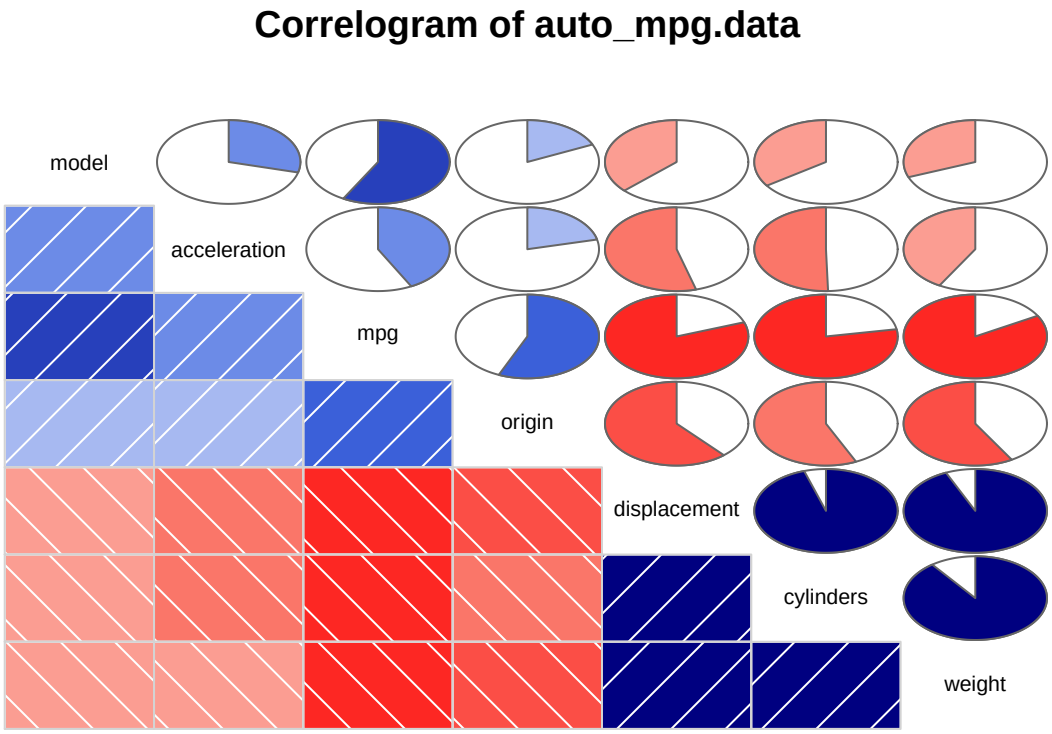
### Data Analysis

Our first goal was to get a general understanding of our data.

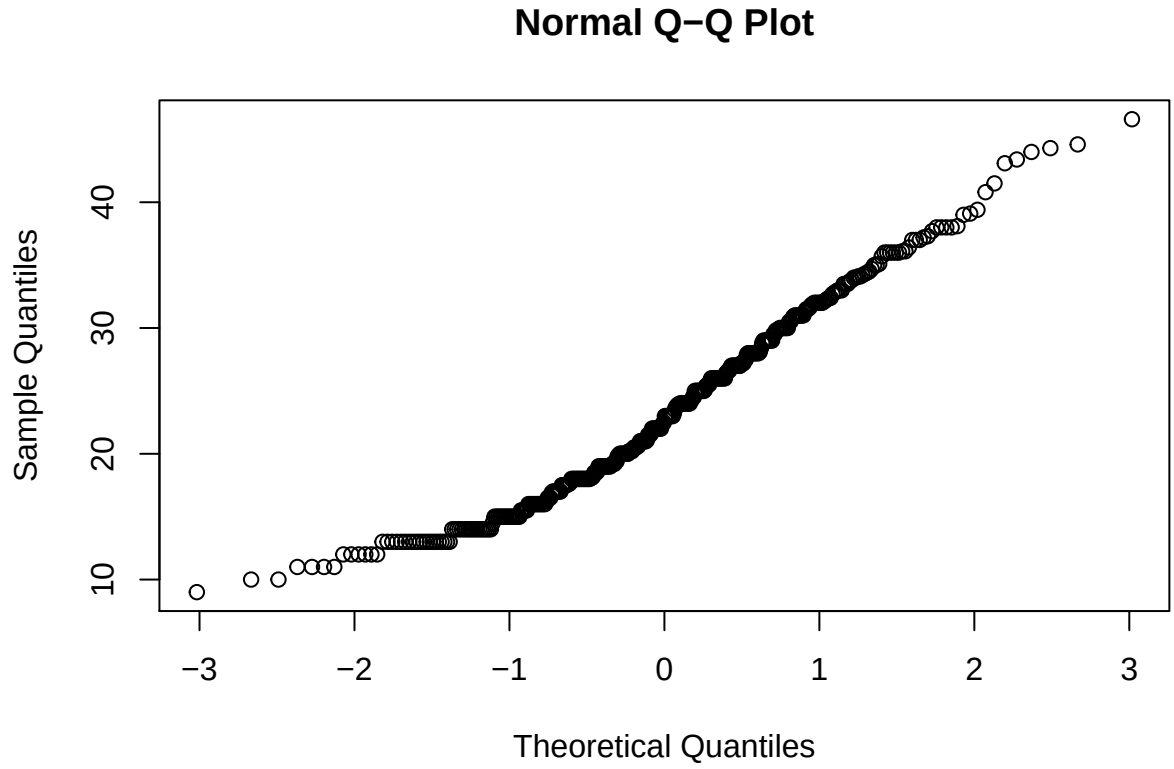
```
##   mpg cylinders displacement horsepower weight acceleration model origin
## 1  18         8          307      130.0   3504         12.0    70      1
## 2  15         8          350      165.0   3693         11.5    70      1
## 3  18         8          318      150.0   3436         11.0    70      1
## 4  16         8          304      150.0   3433         12.0    70      1
## 5  17         8          302      140.0   3449         10.5    70      1
## 6  15         8          429      198.0   4341         10.0    70      1
##                   car_name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

The data contains 9 different characteristics of 398 vehicles' mpg, cylinders, displacement, horsepower, weight, acceleration, model, origin, and name. After viewing the data and trying to manipulate it, it became apparent that there was missing data for horsepower, so in order to solve that issue we simply removed the data for the cars which had missing data. After this, there were 392 vehicles left.

A correlogram of the data set, showing the basic relationships between our variables is:




From this correlogram, we can see that mpg has a positive correlation with acceleration, and a negative correlation with displacement, number of cylinders, and weight. This means that a regression model will work well. The average mpg from the data set is 23.44. The QQ plot of mpg to see whether or not the mpg is distributed normally is shown below:




From the QQ plot we see that the mpg is relatively straight, so the normality assumption is not strongly

violated.

## Question 1: Mpg Vs Cylinders +(or) Acceleration

```
##
## Call:
## lm(formula = mpg ~ cylinders + acceleration)
##
## Coefficients:
##  (Intercept)     cylinders  acceleration
##      40.5704       -3.4624        0.1172
```

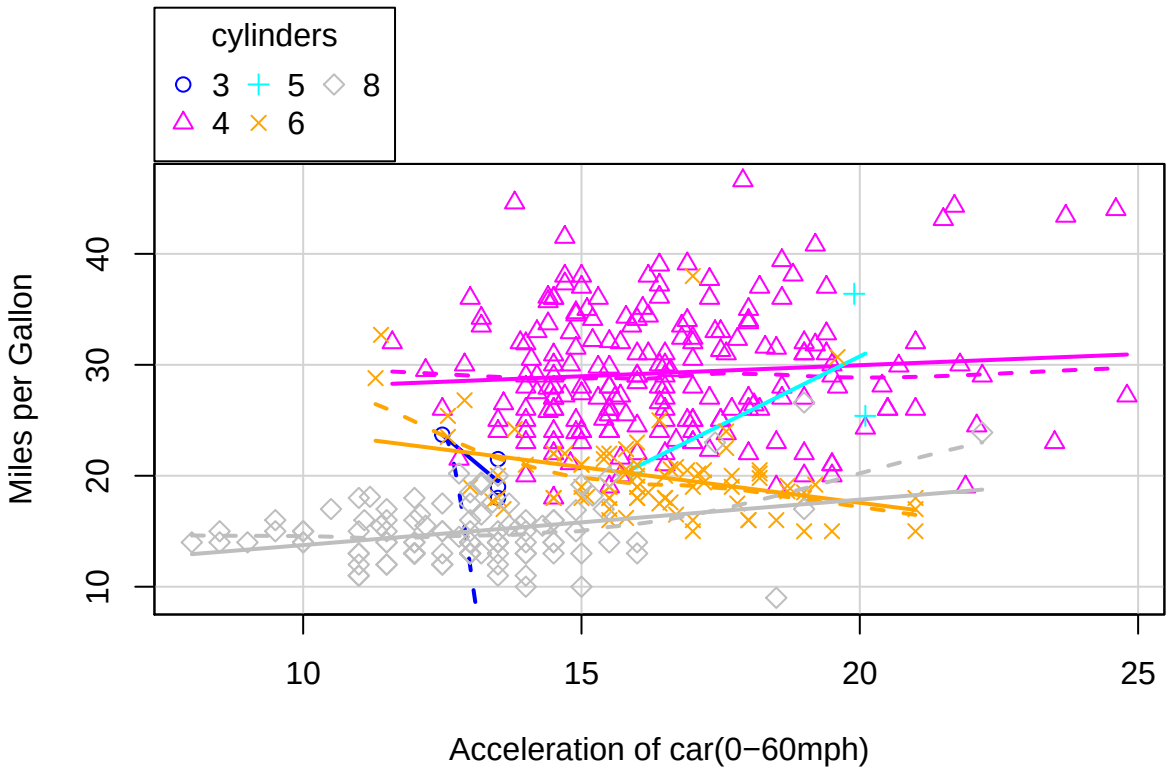
The QQ plot for the mpg vs acceleration+cylinder is straight for the vast majority of the data points, with the upper tail becoming pronounced, so its not perfect, and the data does contain several outliers. The normality assumption holds though, so the linear model is appropriate to predict the mpg, and the data we get from it is useful. From the linear model mpg = (acceleration + cylinders), we obtain the following regression coefficients: cylinders = 0.346, and acceleration = 0.1172, with an intercept of 40.57. This means that for every additional cylinder a car has, it's expected mpg goes down 3.462, and for every 1 second it's acceleration time(0-60mph) increases, it's expected mpg increases 0.1172. Next, we will find the correlation between these variables.

```
## [1] 0.4233285
```

```
## [1] -0.7776175
```

```
## [1] -0.0659962
```

We used the Pearson, Kendall, and Spearman formulas for the correlation coefficient estimates, but ultimately decided the Pearson correlation coefficient was sufficient for us to use. The Pearson correlation coefficient between mpg and acceleration is 0.432, between mpg and cylinders its -0.777, and between mpg and (cylinders+acceleration) is -0.06599. To explore why the (cylinders+acceleration) has a much lower correlation coefficient than mpg vs acceleration or cylinders alone, we will plot a conditional scatterplot matrix of mpg vs acceleration, given the number of cylinders.




This conditional scatterplot matrix makes it much more obvious that the number of cylinders dramatically effects the MPG of a vehicle, whereas the acceleration has little or no effect in general, having a slightly negative effect when the number of cylinders is 6.

## Question 2: Predicting the mpg class.

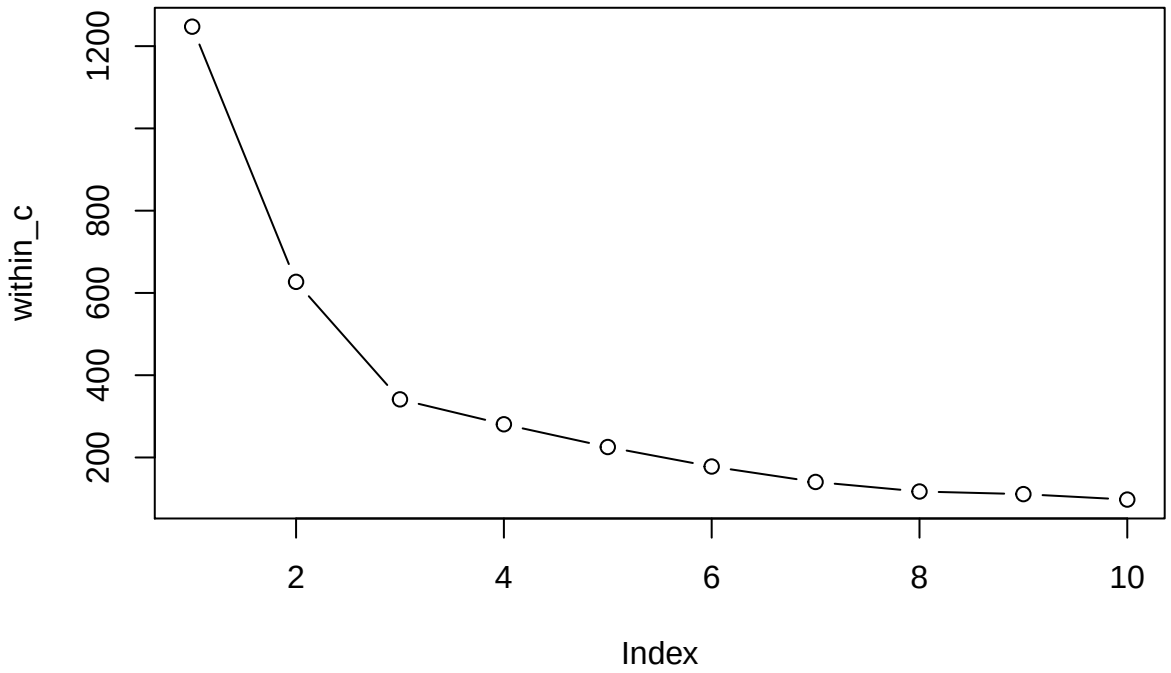
To begin the classification, we first sample 70% of the data and make it the training data set, with the remaining 30% of the data being set to be the test data.

For the prediction, we will use kth nearest neighbor, logistic regression, and linear discriminant analysis to predict the class of vehicles (good mpg, bad mpg). The variables we will use to predict the class are number of cylinders of the vehicle, it's engine displacement, weightm and acceleration. We will compare the error rate of the three different classification methods, and then we will calculate the optimal number of clusters for KNN, and re-apply the three methods using that quanitity. To begin, we apply linear discriminant analysis with two classes(good, bad) and predict which class each vehicle in the test data should be in. We then calulate the error rate of linear discriminant analysis, which is 0.1016949.

Next we apply logistic regression, training it with the training data set, and estimating the cluster membership of the test data. The error rate for logistic regression is: 0.1186441.

The final classification method we will use is Kth nearest neighbors, with k = 10, 20, and 30. The error rate for k = 10 is 0.1101695, the error rate for k =20 is 0.1186441, and the error rate for k = 50 is 0.1101695.

So, the classification method that minimized the prediction errors for 2 clusters was linear discriminant analysis, which incorrectly predicted the class of a vehicle in our test data set 10% of the time. Next, we will redo Kth nearest neighbor classification with different k values(number of neighbors considered), to find out what value of classes will allow us to best predict the classes. For each K, k =1,..,10 we will calculate the deviance within each cluster until we find the one that minimizes it.




From the graph, we can see that the within deviance is optimized when k = 3, based on the location of the elbow. This means that increasing K beyond 3 will provide insignificant improvements, which being much more computationally demanding. Now we will redo LDA, logistic regression, and knn classification methods with 3 classes(low mpg, medium mpg, and high mpg) to compare the accuracy to 2 classes. The 3 classes will be low mpg(0-16mpg), medium mpg(16-32mpg), and high mpg (>32 mpg).

Re-applying linear discriminant analysis to the data with 3 classes, our prediction error rate is 0.254.

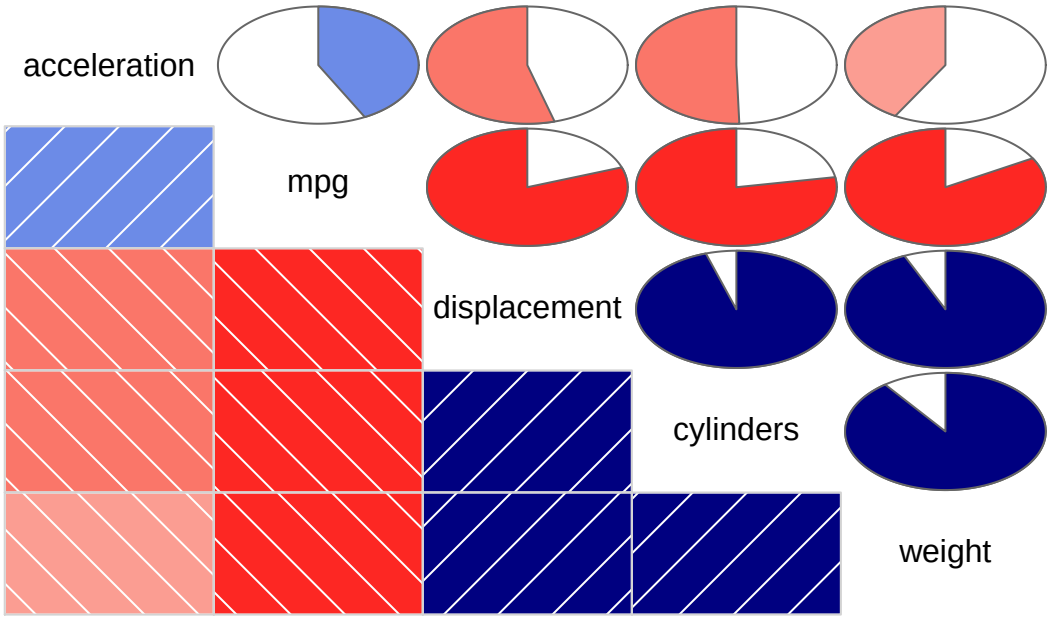
Re-applying logistic regression to the data with 3 classes, our prediction error rate is 0.2542373.

Re-applying K nearest neighbor to the data with 3 classes(low, medium, high), and k=10,20,30, we obtain prediction error rate for k = 3: 0.237, for k = 20: 0.2627119, k = 50: 0.2542373.

From this, we get the lowest prediction error rate for 3 clusters from the the KNN method, with 3 clusters and k = 3, which has a prediction error rate of 0.237. So, our final results are that when we are predicting 2 classes(good or bad mpg) is using the linear discriminant analysis method of classificiation. When we were predicting with 3 classes( low, medium, high MPG), the Kth nearest neighbor classification method was the most accurate when k=3.

## Question 3: Which factors affect mpg the most?

To answer the question of which factor(cylinder, displacement, weight, acceleration) affects the mpg of a vehicle the most, we first look at a correlogram of the variables of interest.




Next, we made several multiple variable linear regression models. The first has all 4 explanatory variables( cylinders, displacemenet,weight, acceleration), with an intercept of 41.61. When we remove the factor acceleration on the second linear model, the intercept increases by 2.5. When we remove acceleration and also weight from the third linear model however, the intercept drops dramatically by roughly 8. No other factor's presence affects the expectation by as much as weight, so the factor weight has the most effect on the mpg of the vehicle.

To confirm these findings, we will find the correlation coefficients between (mpg and cylinders), (mpg and displacement), (mpg and weight), and (mpg and acceleration). The correlation coefficient between mpg and number of cylinders is -0.777. The correlation coefficient between mpg and displacement is -0.805. The correlation coefficient between mpg and weight is -0.832. The correlation coefficient between mpg and acceleration is 0.432.

The correlation between weight and mpg being stronger than the correlation between mpg and any of the other variables and mpg confirms our previous findings that the weight of a vehicle affects its MPG more than it's number of cylinders, engine displacement, and it's acceleration.

## Discussion:

The results of our study indicate that several factors greatly affect the miles per gallon of a vehicle. Among these factors, we found that acceleration and the number of cylinders both affect the MPG, but the cylinders affects it much more significantly. We found a lot of the data and linear models we ran to be roughly normally distributed, which alloiwed us to use classificaition and other predictive models that require normality. One of our results was that of the factors that our data set contained, weight was the factor that affected the miles per gallon of a vehicle the most. We found that linear discriminant analysis is the best method to classify the data into two classes, but all of the methods, KNN,lda, and logistic regression provided similar prediction error rates. We believe that part of these results being similar is the sample size of the data. For several of the confusion matrices, there is less than 5 false predictions, which is too small of a quantity to be accurate when repeated. This is reaffirmed by changing the seed of the code, which changes the test and training data, altering our error rates significantly. When we found that k=3 is the ideal number of clusters for Kth nearest

neighbors, and then re-applied the classification methods with 3 classes, the predictive error rate for the kth nearest neighbors with 3 clusters became the best classification model. Another thing to consider is that our data is form 1983. Since then, many aspects of cars have evolved, such as the weight decreasing, the mpg increasing overall, and the acceleration getting faster. It would be an interesting followup to try and apply our predictive models to more current data, and see whether or not the relationships we found still exist.