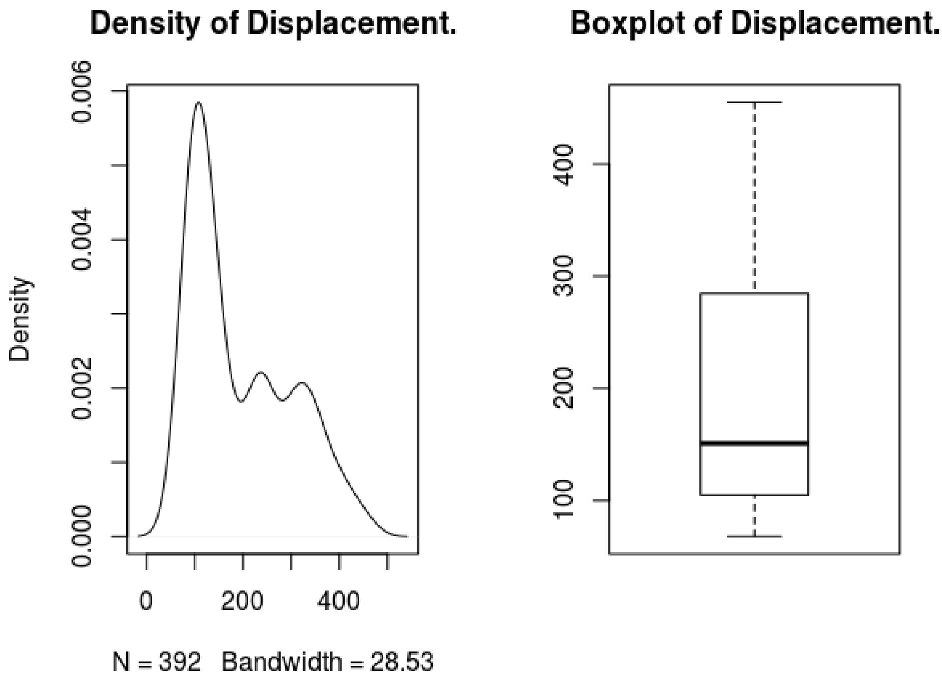
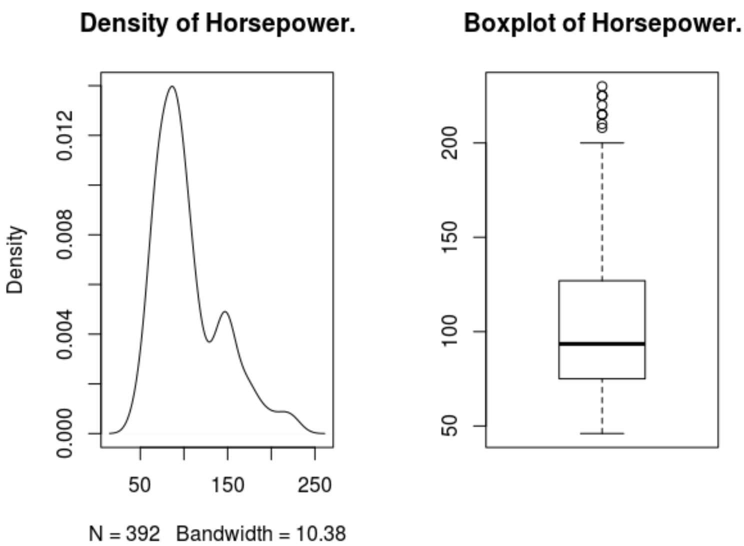
Sai Kiran Putta – 62.

Question – 1
The question mainly deals with the Exploratory Data Analysis.
Following are my observations, steps taken and inferences.

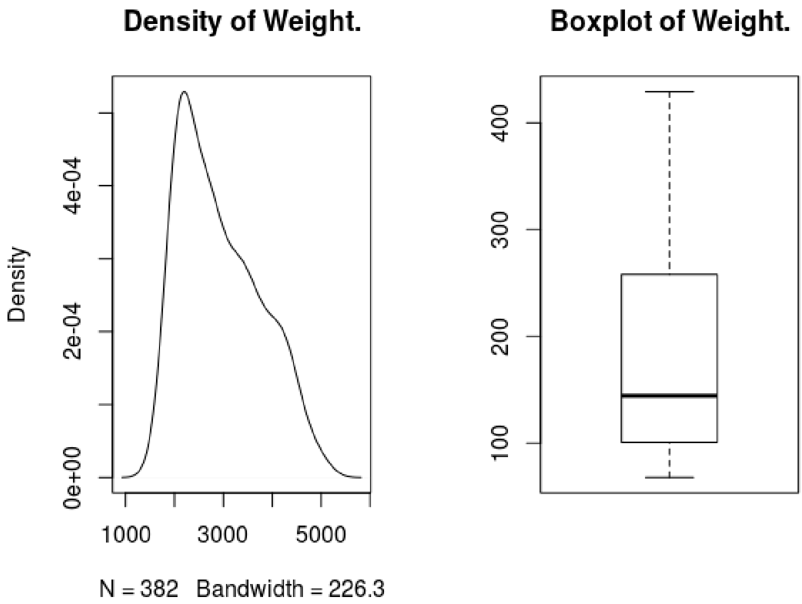
- There seems to be no NA values in this dataset.
- There are some features which are supposed to be categorical, but are numeric. Hence we need to convert them. The features that are being converted are, cylinders, year and origin.
- From the first glance, it looks like ‘name’ feature has a lot of values and it is actually categorical. So if we run regression on it we are going to have many features. Hence we need to delete it. Since we are going to delete it we can add a new feature of Company name from the name so that we don’t loose too much information from the deleted name feature.
- There seems to be many spelling error in the company names like, ‘chevroelt’ instead of ‘chevrolet’. And also abbrevations like ‘vw’ for ‘volkswagon’
- Following are some plots.



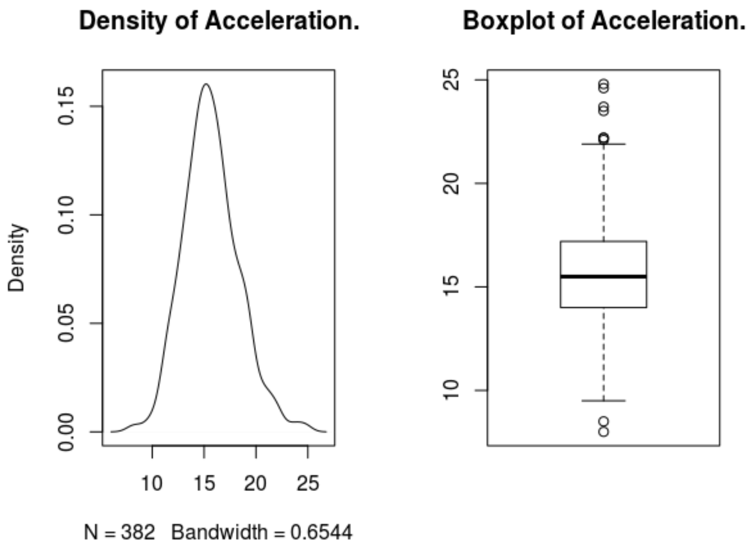
Looks like there are no outliers in the data. And data looks like a Tri-modal data.



The data looks like a Bi-modal data. And there are some outliers that should be removed.



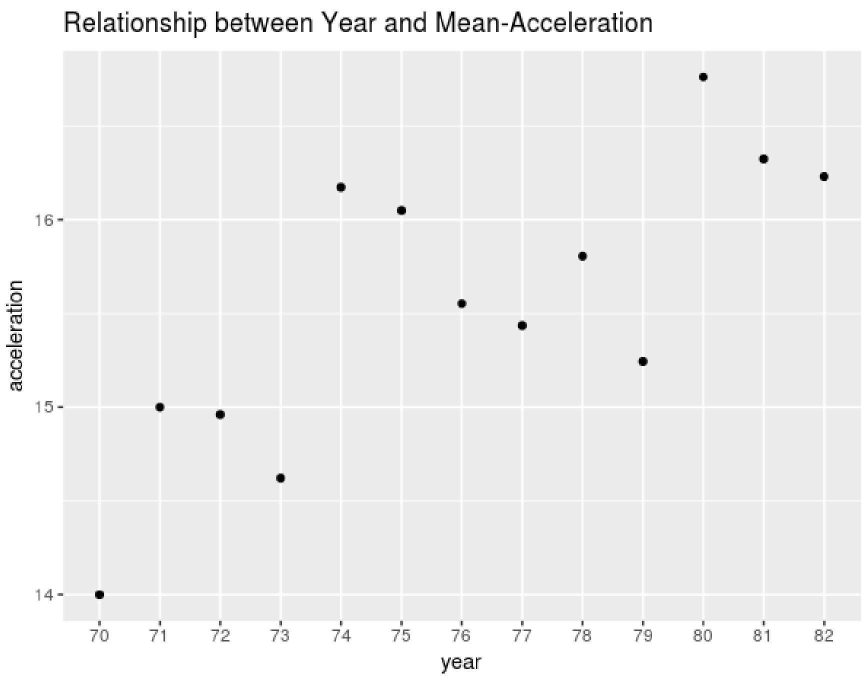
The data looks like a right tailed normal distribution without any outliers.



The data is a normal distributed data with some outliers that should be removed.

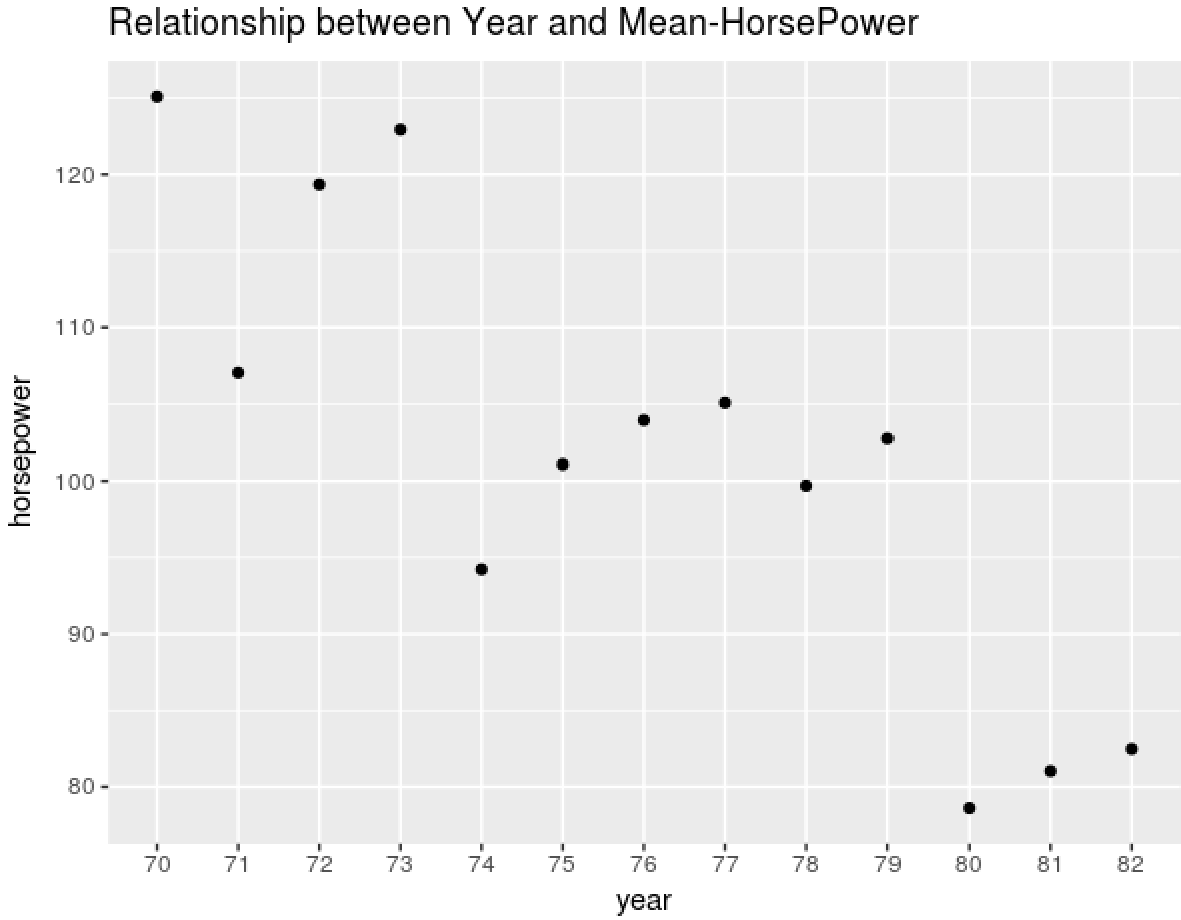
Let’s check how the pair plot looks for this dataset.

From the glance to this plot there seems to be some pattern for [horsepower, mpg] , [horsepower, displacement], [weight, mpg], [weight, displacement], [weight, horsepower].



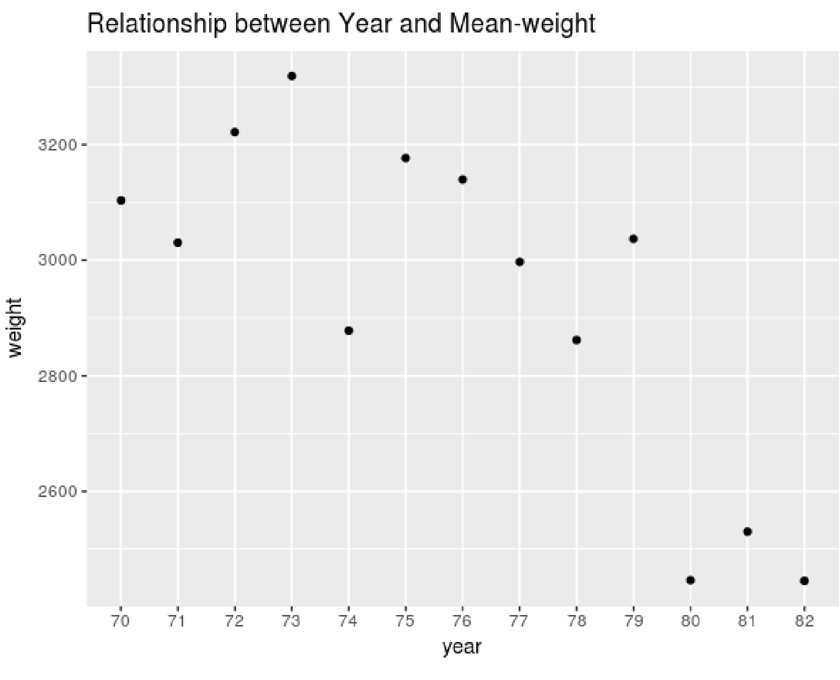
Looks like there is clear upward trend between the year and acceleration. Infact there is a 76% correlation between the year(converting into numeric) and acceleration.

Let’s check if there is any trend for year and horsepower.



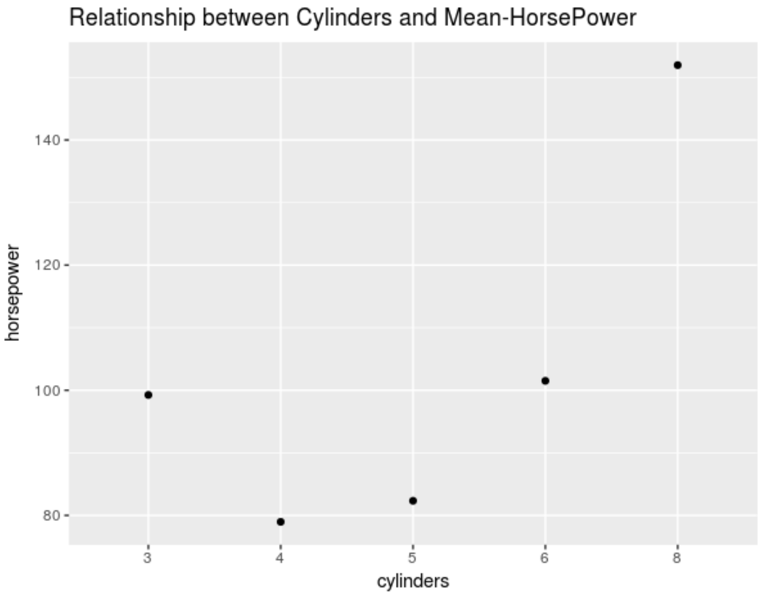
Clear downward trend can be seen with correlation -84%

Relationship between year and mean-weight.

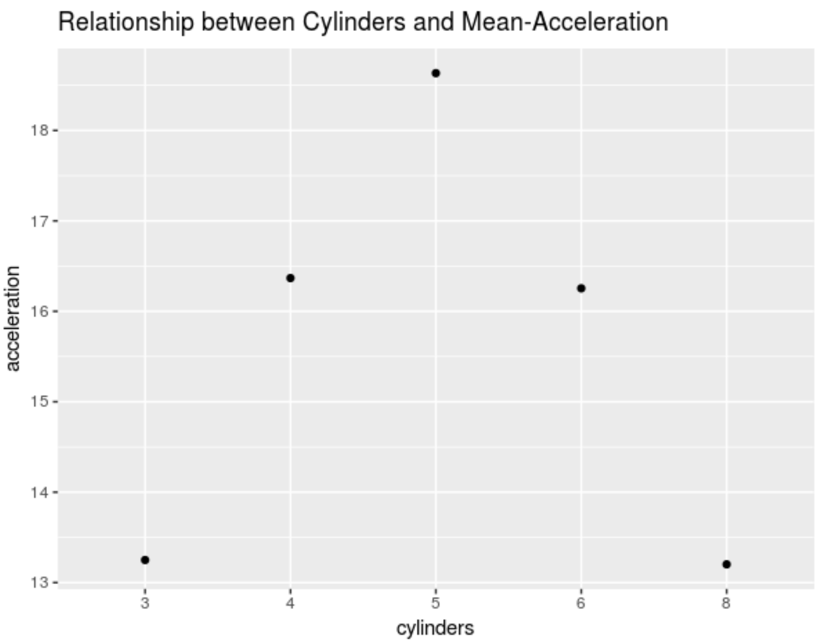


Clear downward trend can be seen with correlation -77%
It makes sense why the acceleration increased. It could probably have been due to decrease in weight over time!

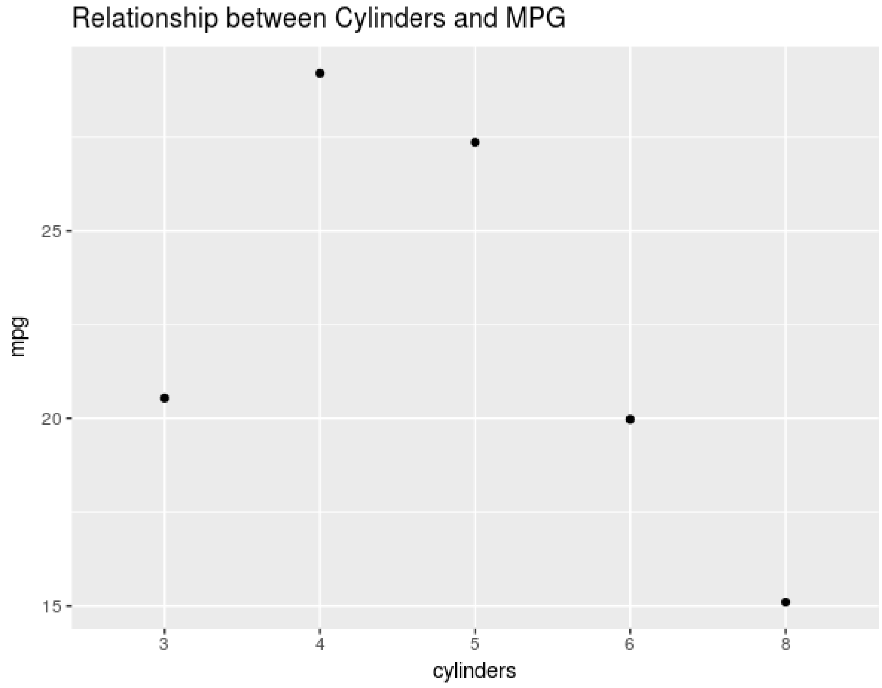
It makes intuitive sense that with more number of cylinders, more will be horsepower and thus accelleration. Let's check what the data says!



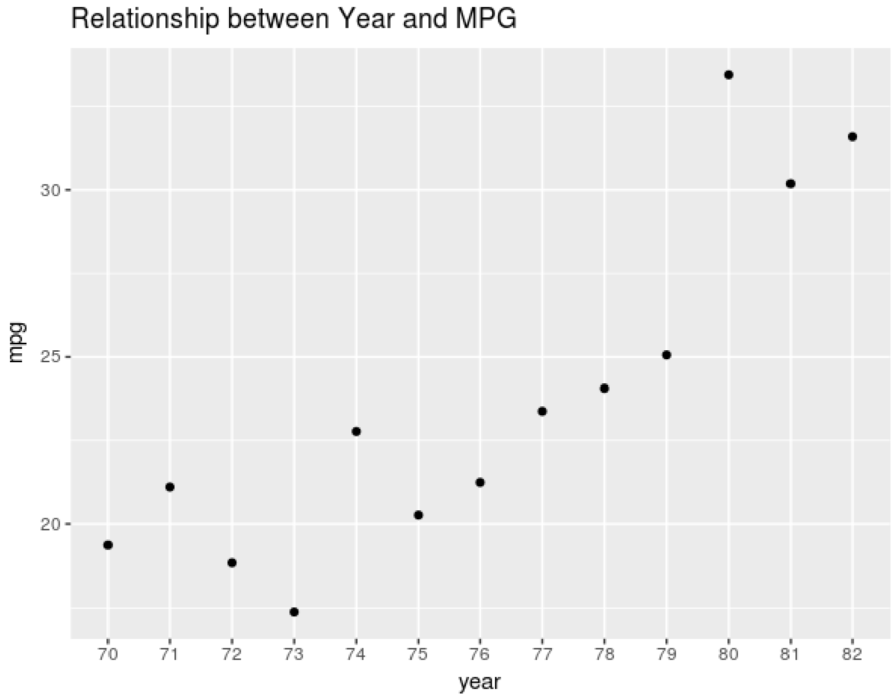
Looks like with respect to getting horsepower out of the engine, 4,5 design doesn't comply.
But the general trend is what we expected. That is horsepower increases with increase in cylinders.



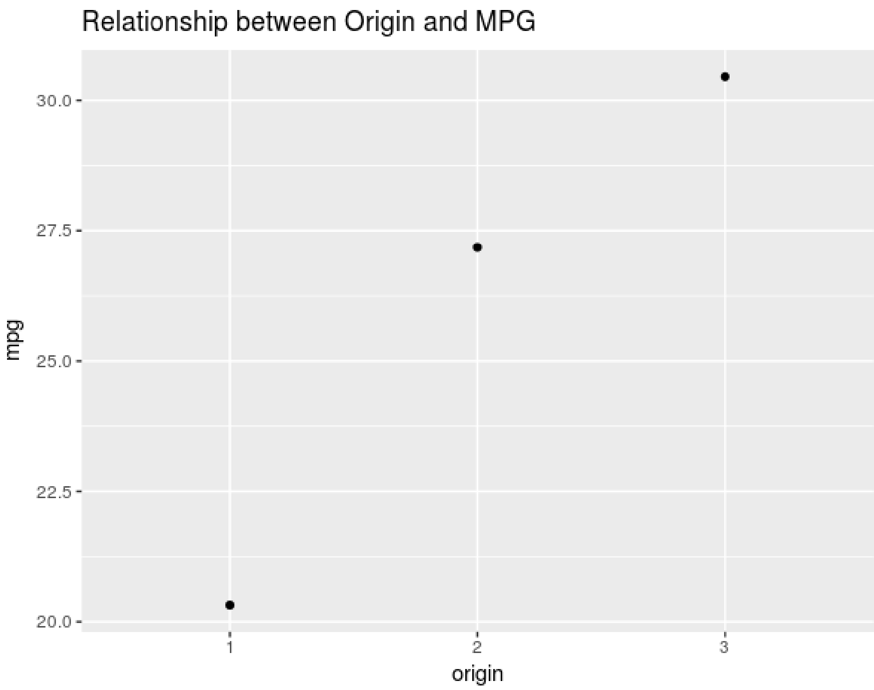
But interestingly 5 cylinder engine has great acceleration. Looks like it is the tradeoff engineers have to make with acceleration and horsepower while designing an engine for a car.



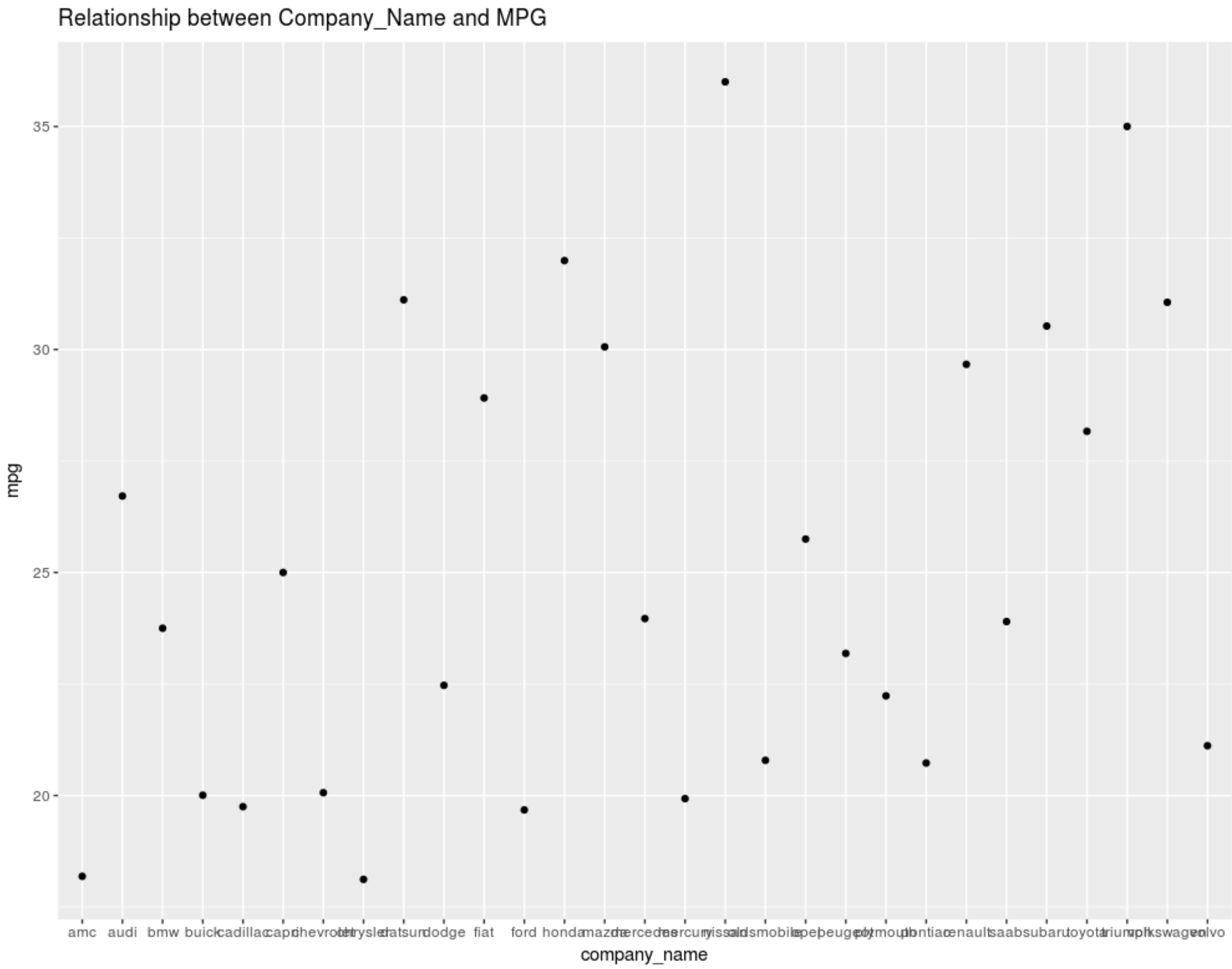
Like one can expect the mileage goes down with more cylinders (sports cars!)



Looks like technology has improved over the years!



clearly increasing!



Chrysler has the lowest Mileage!
Nissan has the highest mileage!

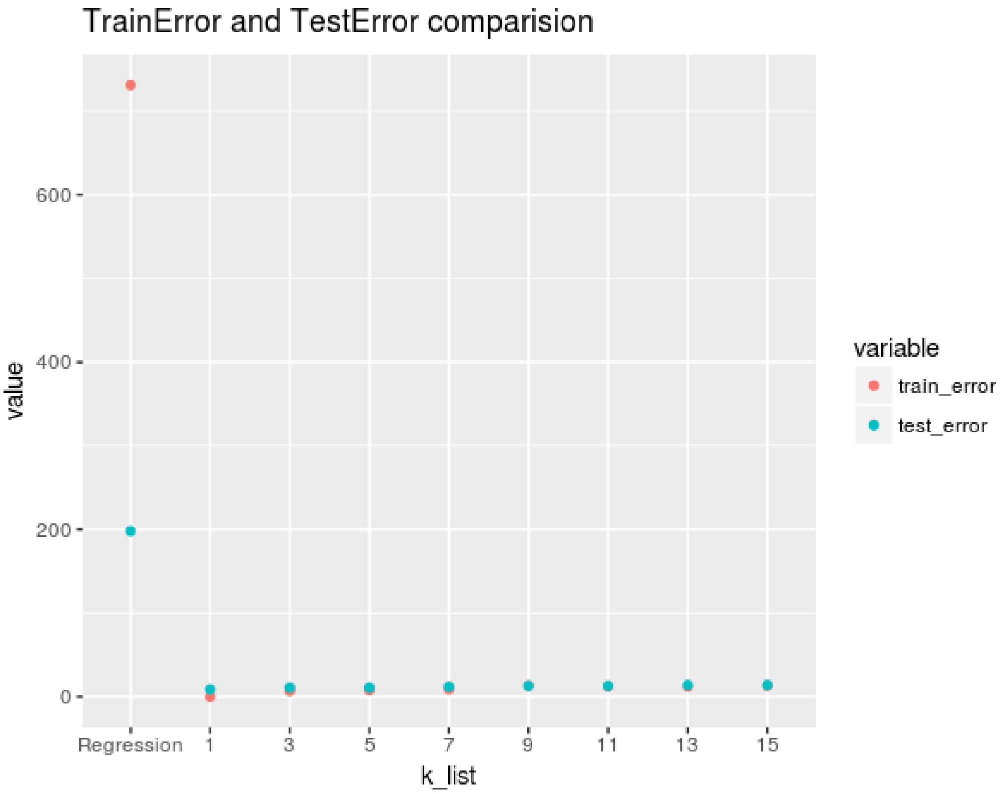
Question – 2
This question mainly deals with applying regression and it's observations.
a) Features, Cylinders4, Cylinders5, Cylinders6, Cylinders8, horsepower, weight, acceleration, year77, year78, year79, year80, year81, year82, company_namebuick, company_namecadillac,company_nameoldsmobile,
company_nameplymouth, company_namepontiac are important.

b) While year71-year76 seem not significant. Years 77-82 are important. While others have a very low coefficient
77-82 have high coefficient. For example for 82, the coefficient is 6.9. This means for unit increase (basically yes or no since it is a factor)
MPG increases by 6.9. Like we saw in graphs, as the years progressed MPG increased drastically.

c) *, : are alternate ways to do accomplish the same thing. That is two way interaction between the features.
Let's look only at the numeric two way interaction for simplicity.
In the two way interaction apart from the ones that are mentioned without interaction,
#[displacement,weight] and [horsepower,acceleration] are found to be important.

Question – 3

In this question, I'm first converting the 2, 3 to 0, 1 respectively.



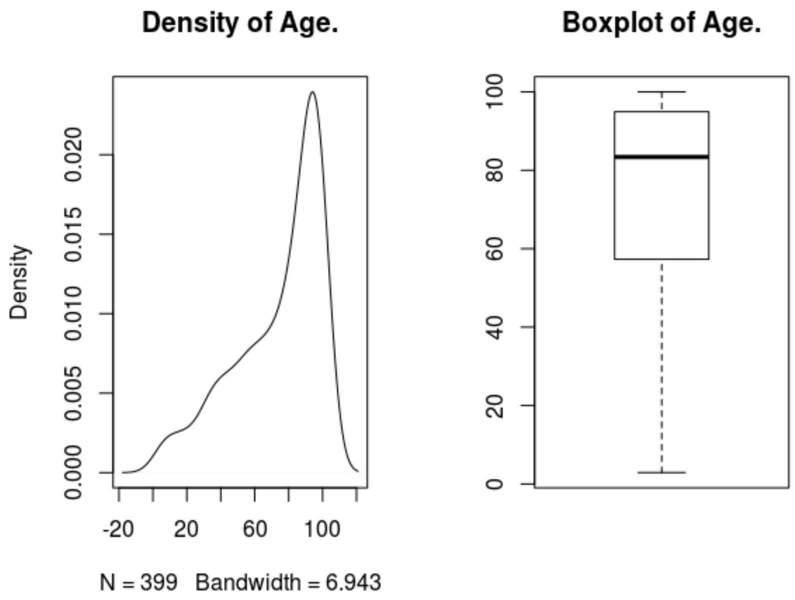
The error from regression so high that we are not able to see the intricacies of KNN clearly. First let's remove the regression error from this graph and then plot.

We can now compare the plots clearly. Like in theory, for lower values of k the error would be less but as the k value increases the error creaps up. That is the decision boundary smooths out and there is less chance of overfitting.

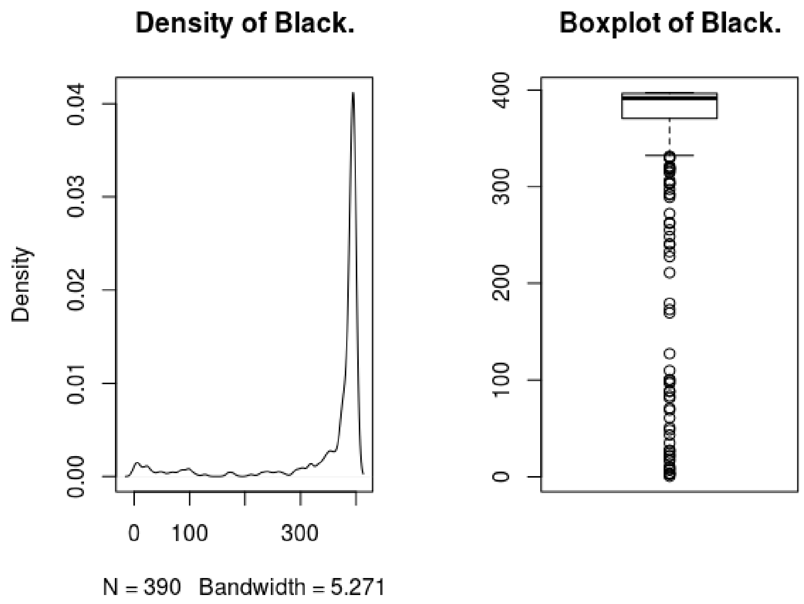
Question – 4

Let’s begin with checking if there are any NA values in the dataset. There are no outliers in this dataset.
It looks like there are some categorical variables which are labelled numeric. Let's convert them. The features that are being converted are chas and rad.

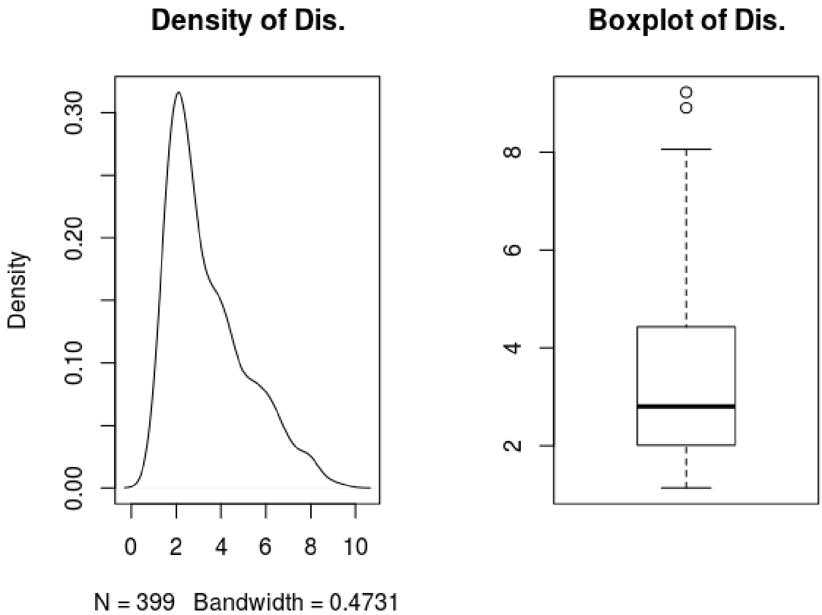
Let’s look at the distribution of the data.



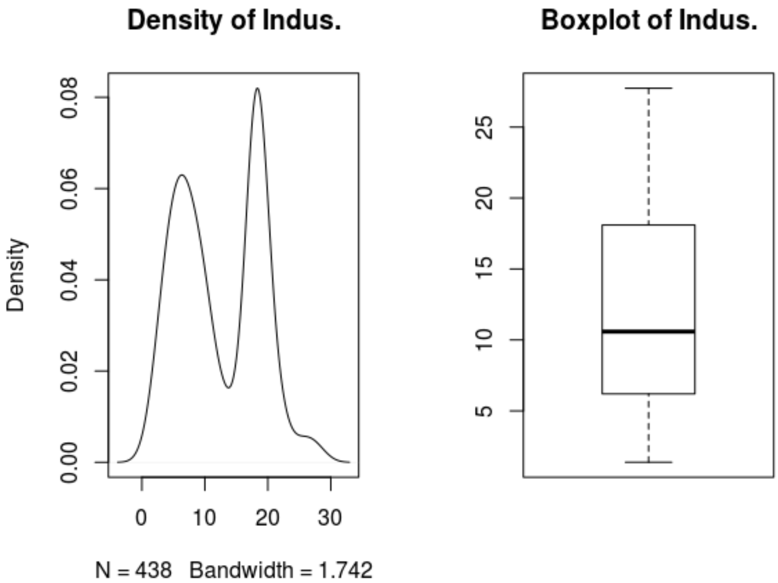
The data looks like left skewed without any outliers.



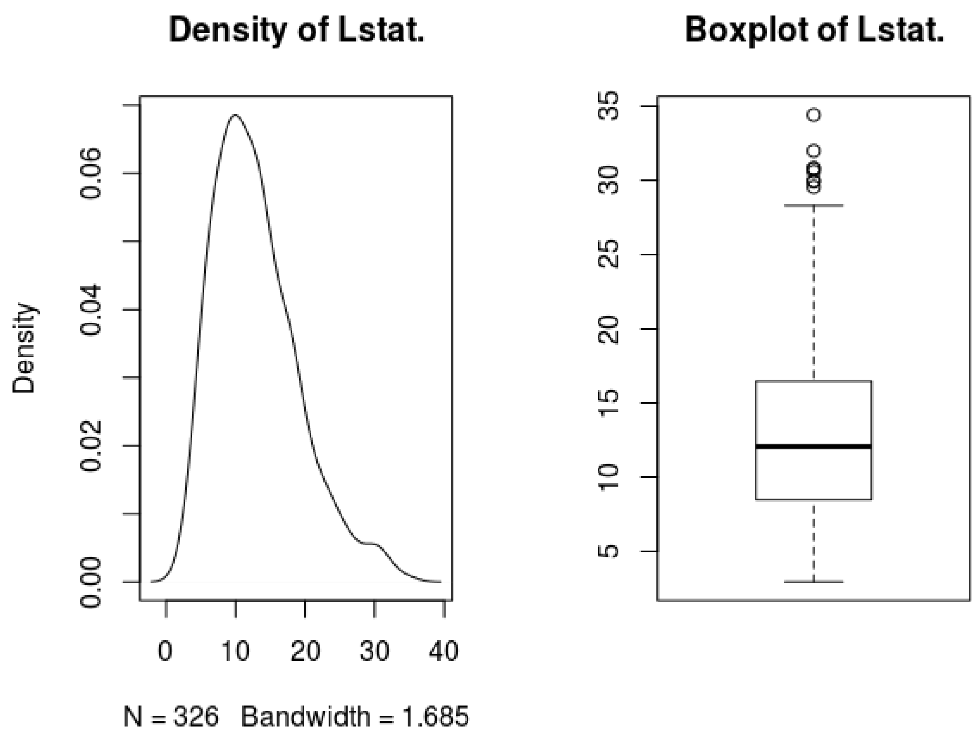
The data seems to be heavily skewed towards left with many outliers.



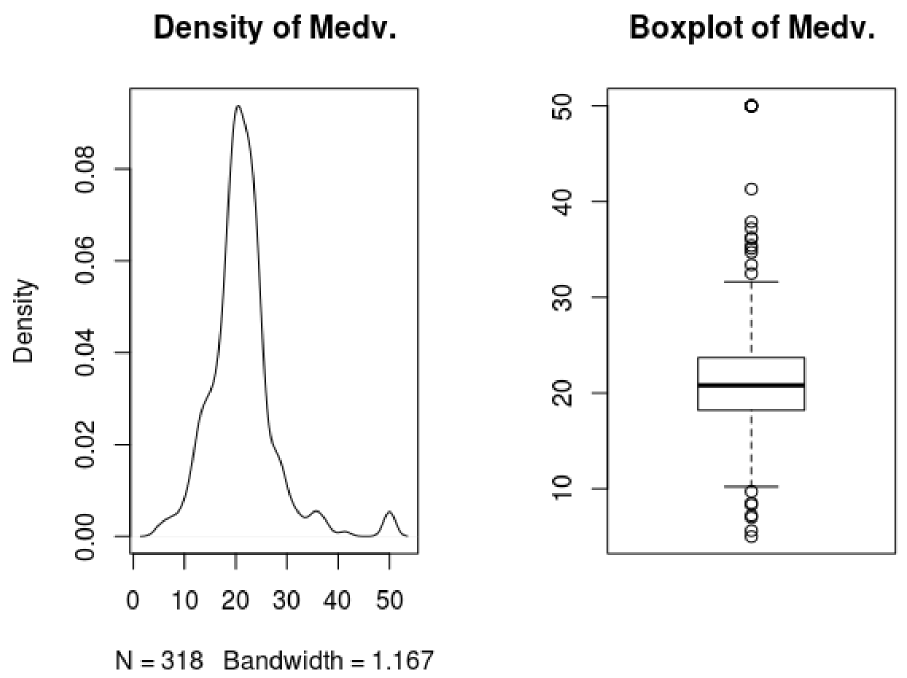
The data is right skewed with a couple of outliers.



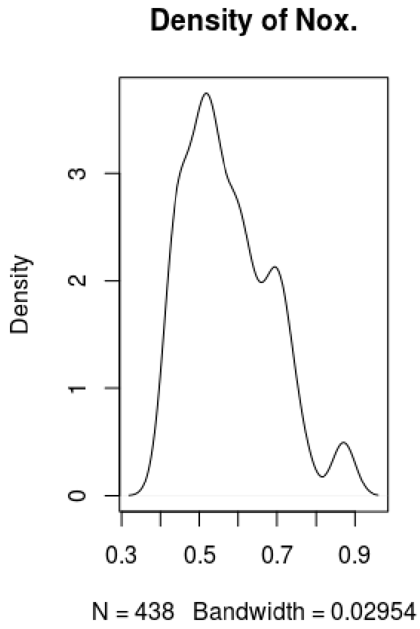
The data is a bimodal data with no outliers.



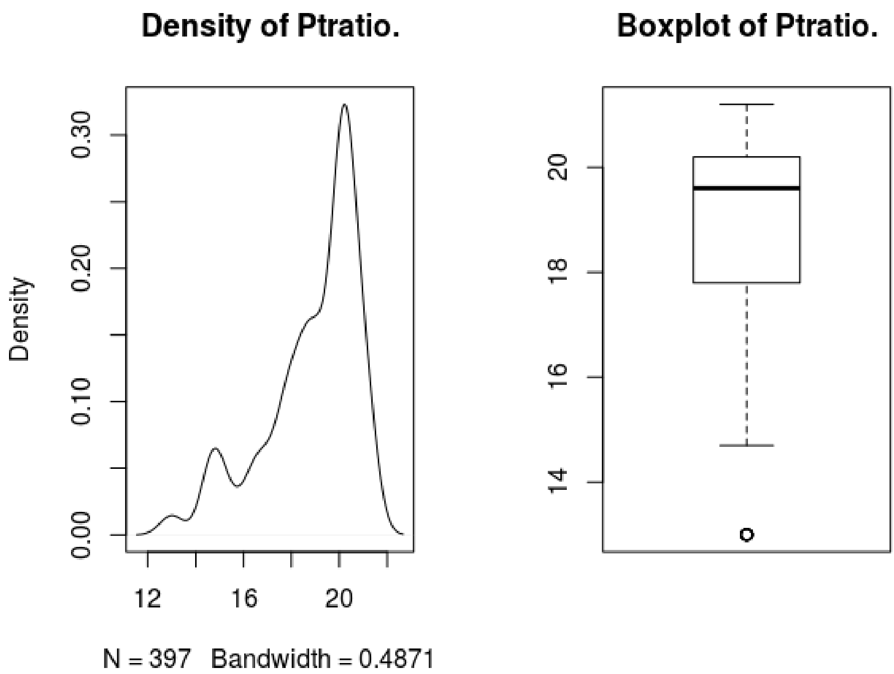
The data seem to be right skewed with some outliers.



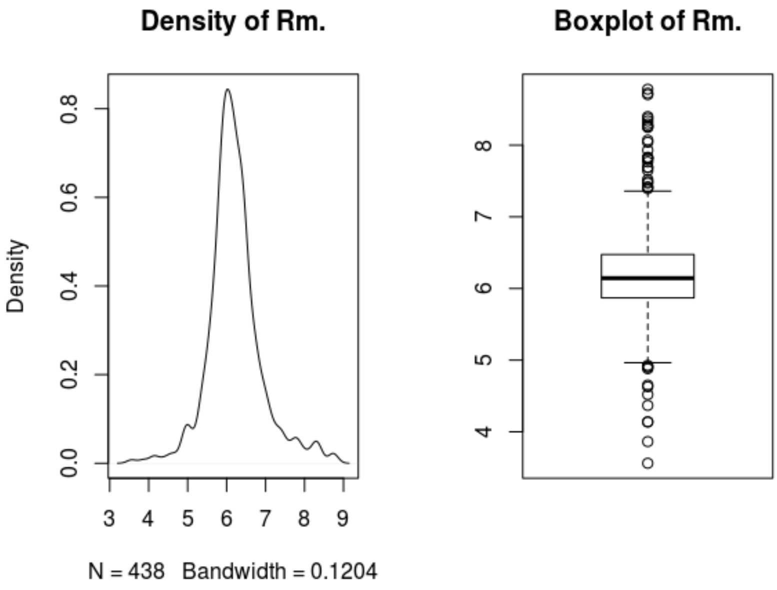
The data seems to be a right skewed with many outliers that are to be removed.



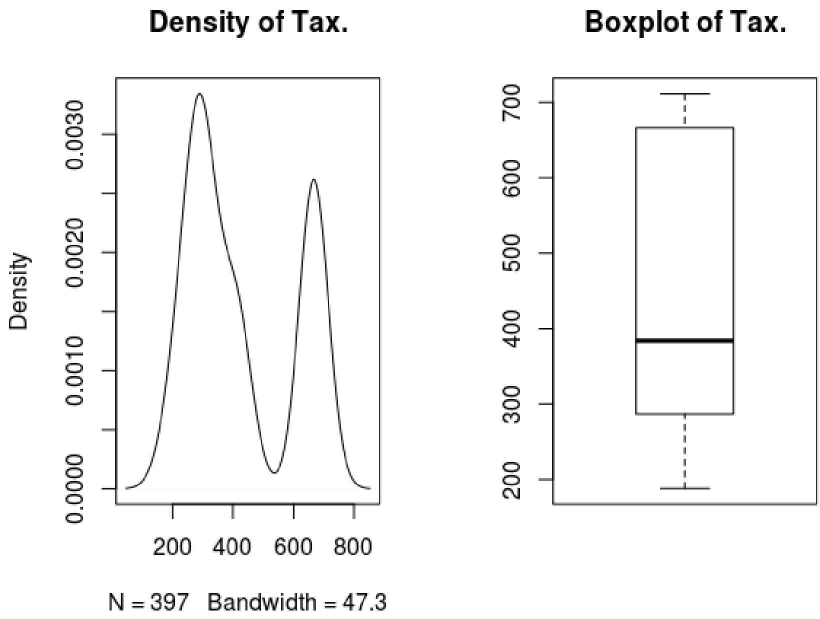
There seems to be no outliers in this dataset.



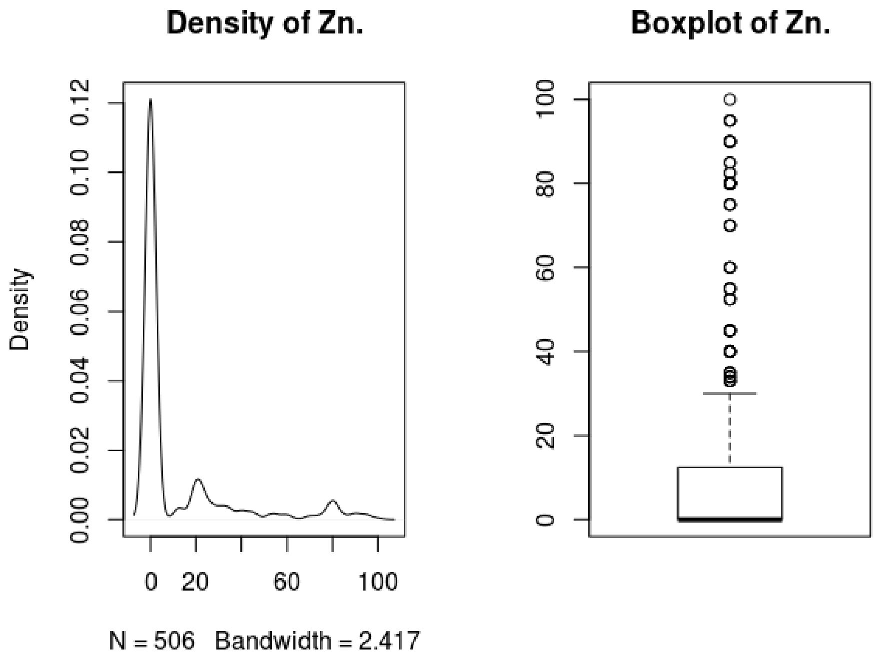
There looks like there are few outliers in this dataset that are to be removed. The data looks like left skewed.



The data looks to be in a normal distribution with left and right tails with many outliers that are to be removed.

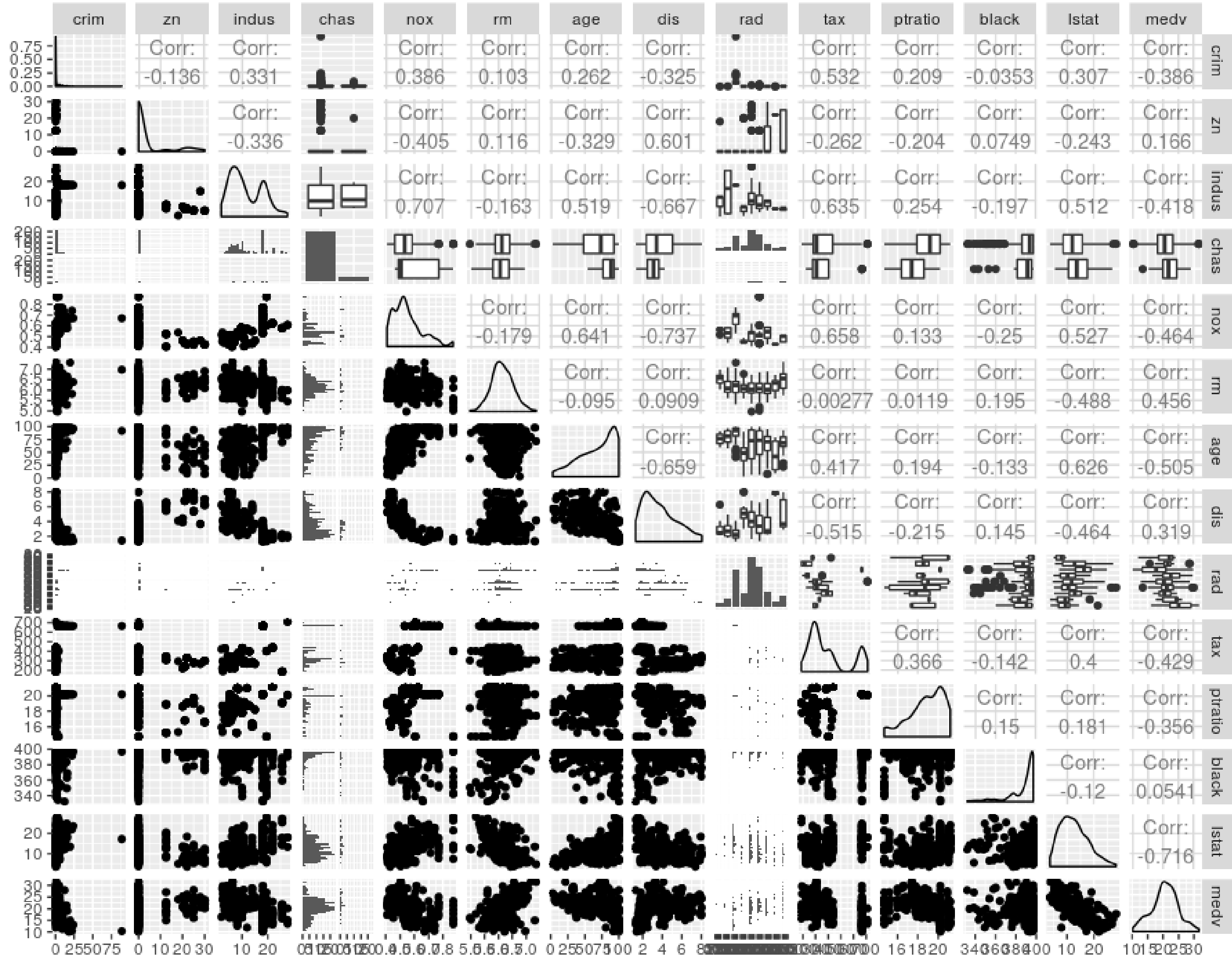


The feature seems to be a bi-modal data with no outliers.



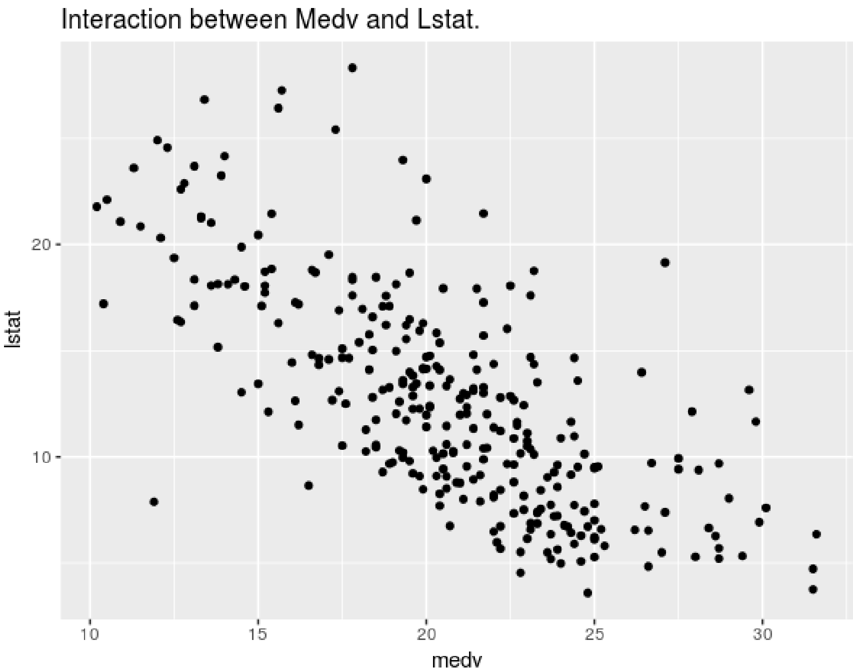
The feature is heavily skewed towards right with many outliers.

Let’s look at the pairplot for this dataset.

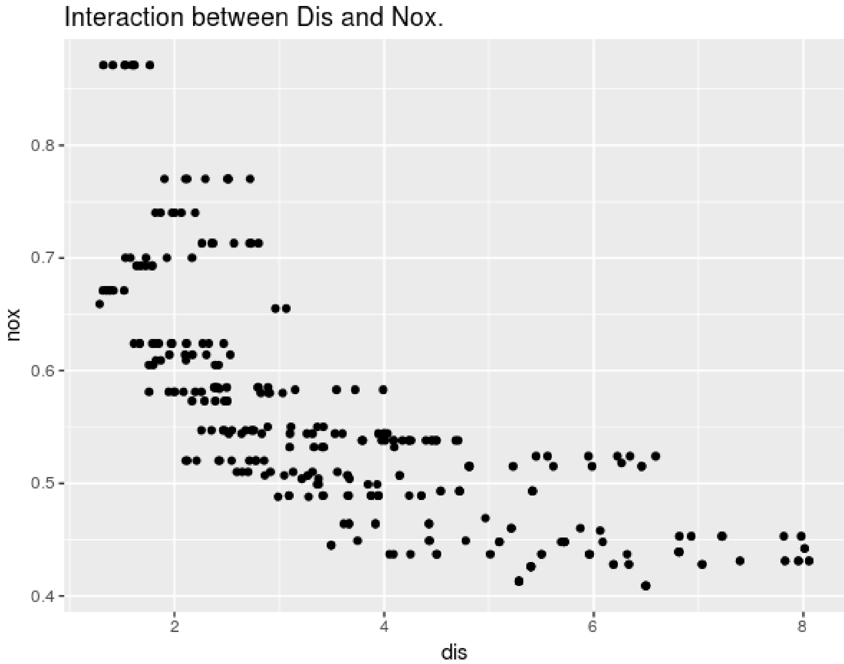


From the pairplots above. Most of the data plotted looks like blobs of data.
Except for medv,lstat and dis,nox, rm,lstat and rm,medv. Let's plot them seperately for a cleaner look.

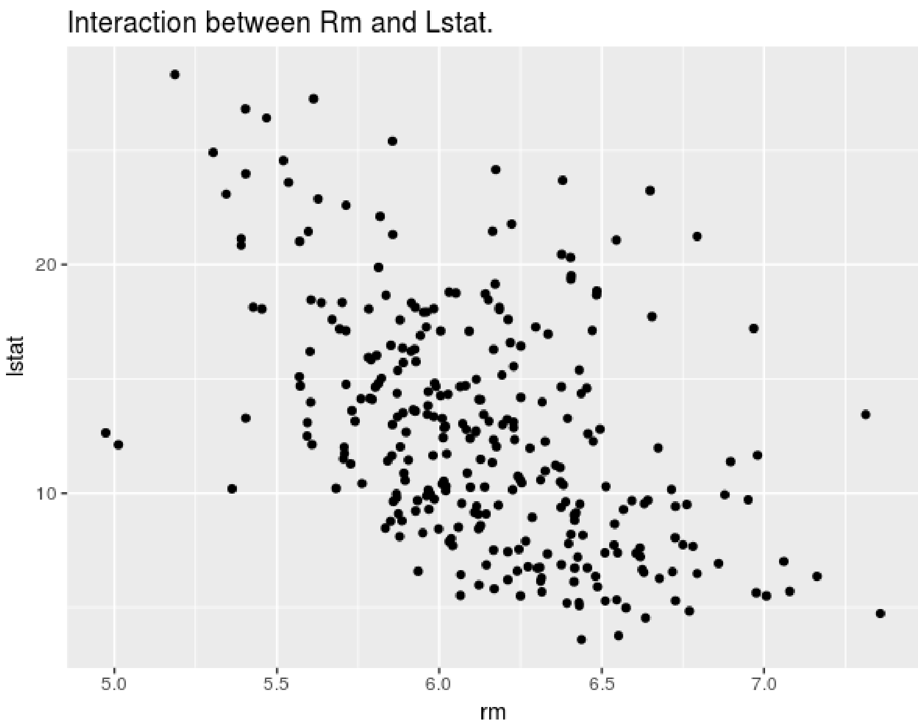
Let’s look at some explicit relationships.



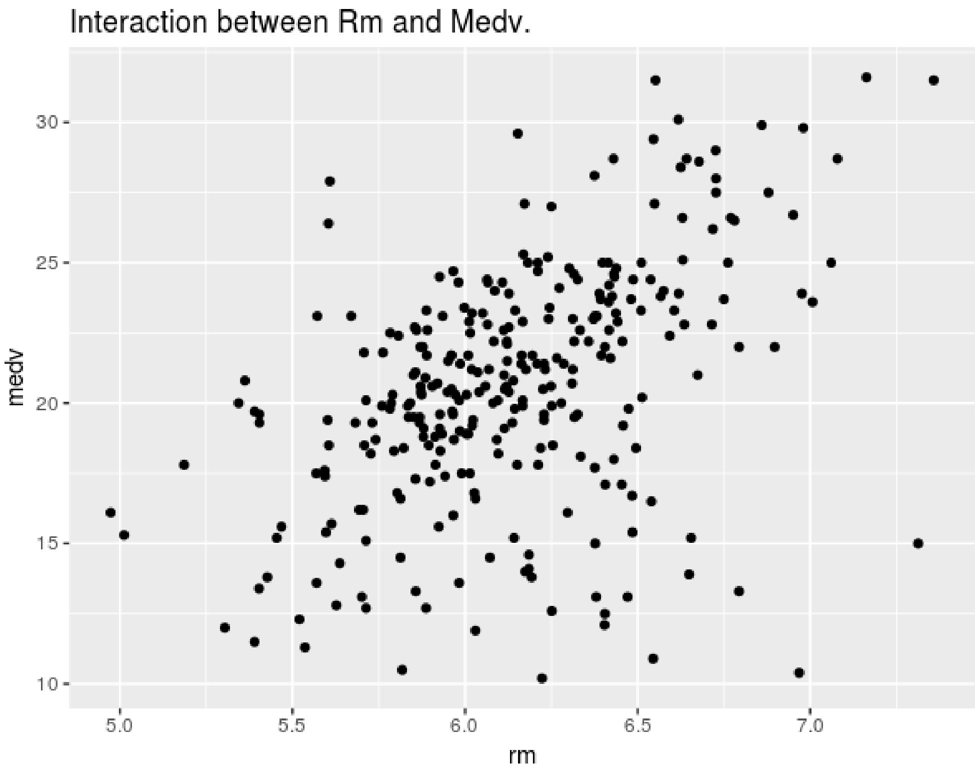
There seems to be a downward trend between price of homes and % of lower status population.



The relationship makes sense. The closer you are to the employment center, more is Nitric Oxide concentration (Pollution).



The relationship makese sense. As the % of lower population increases no of rooms decrease.



As the no of rooms increase the price of the house also increases.

b)
From the pairplots, there doesn't seem to be any relationship.
Let's do a regression to see the relationship between per capita crime rate and others.

Features rm, age, dis, rad24, rad7, ptration and medv seem to be important.

c)
Lets take the highest 20 observations where there is maximum crime. By taking that following are the obervations.
All zn = 0
All Indus = 18.1
All chas = 0
rad = 24 (that is more pollution)
ptratio = 20.2

For tax rates.
Lets take the highest 20 observations where there is maximum tax rates. By taking that following are the obervations.
Higher age(mean = 93),
All Zn = 0
Many chas = 0 (16 in total)
Many rad = 24 (17 times. More pollution)

High ptratio. Lets take the highest 20 observations where there is maximum ptratio. By taking that following are the obervations.
All zn = 0
All chas =0
All rad = 4
All tax = 666
All ptratio = 20.2.

d)
No of obervations where there are more than seven rooms per dwellings is 6.

There are no dwellings where there are more than eight dwellings. May be I might have deleted them in outliers.