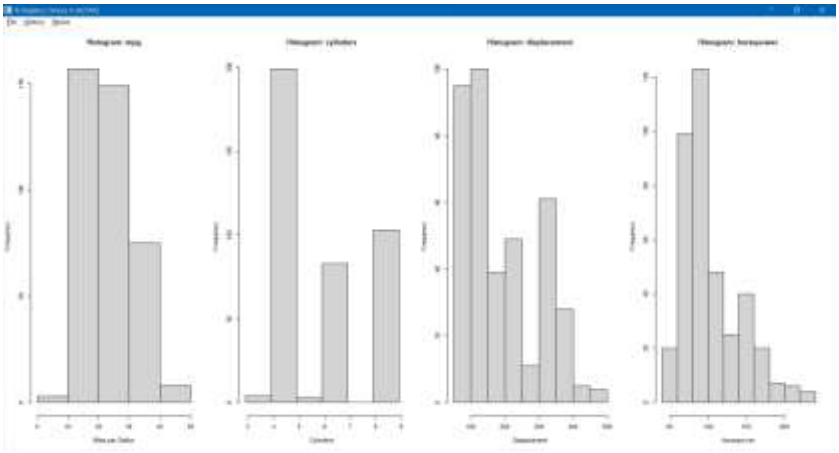
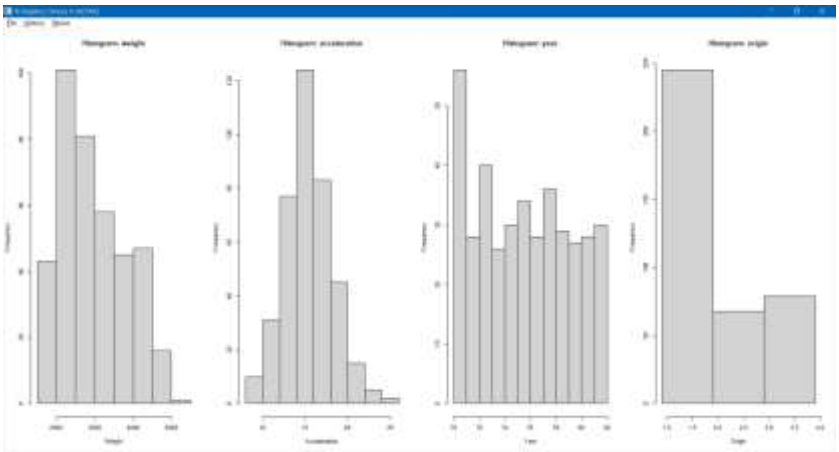
Kunal Mehta
50352960
kunalkam

Homework 1

Problem 1

a. Exploratory Data Analysis – As instructed in the question, I went about examining the dataset superficially at first, displaying the dimensions, the names of variables and a short snippet of the dataset. Of course, at the very beginning, I cleared my environment and loaded the lattice, MASS and ISLR libraries. On first glance at the values itself, I determined that the name variable was merely informative and would not lend any influence when it came to developing a predictive model.
b. The next step was to plot histograms of each of the variables with customized breaks when necessary. The screenshots of the same are as displayed:

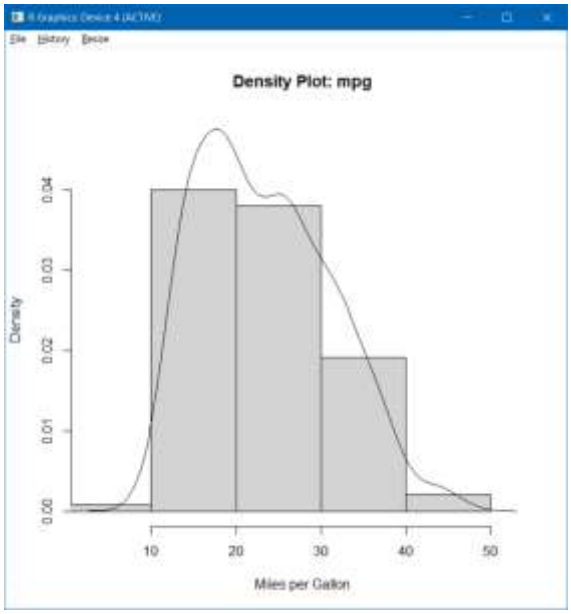




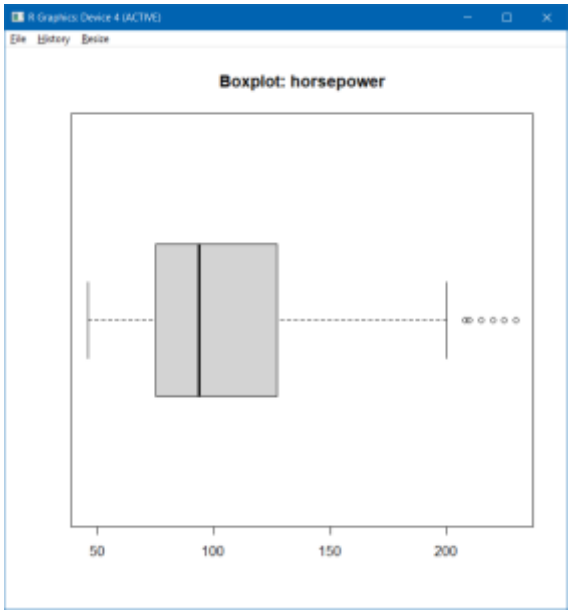
Owing to the values that cylinders variable took, I made 6 breaks of width 1 starting at 2.9. Similarly for origin, I made 3 breaks starting from 0.9 with width 1. The minimum and maximum values of year were 70 and 82, thus I made sure the histogram had exactly 12 breaks.

Kunal Mehta
50352960
kunalkam

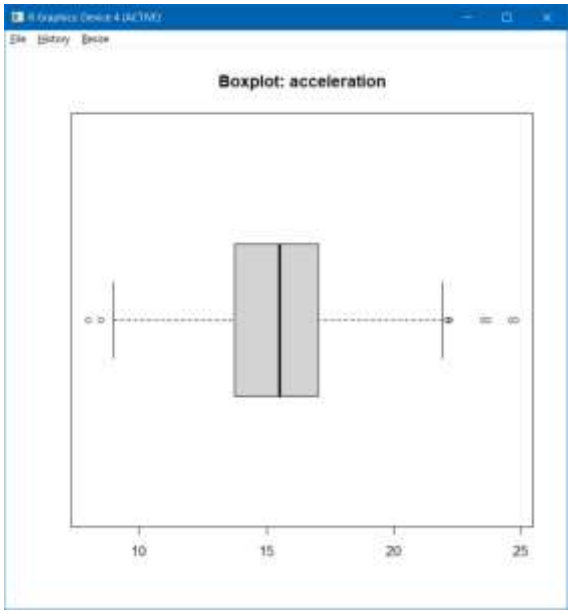
c. The crudeness of histograms as spoken about in the lab was immediately apparent, and I did not believe density plots would provide me much better insights, therefore I created one for the mpg variable.



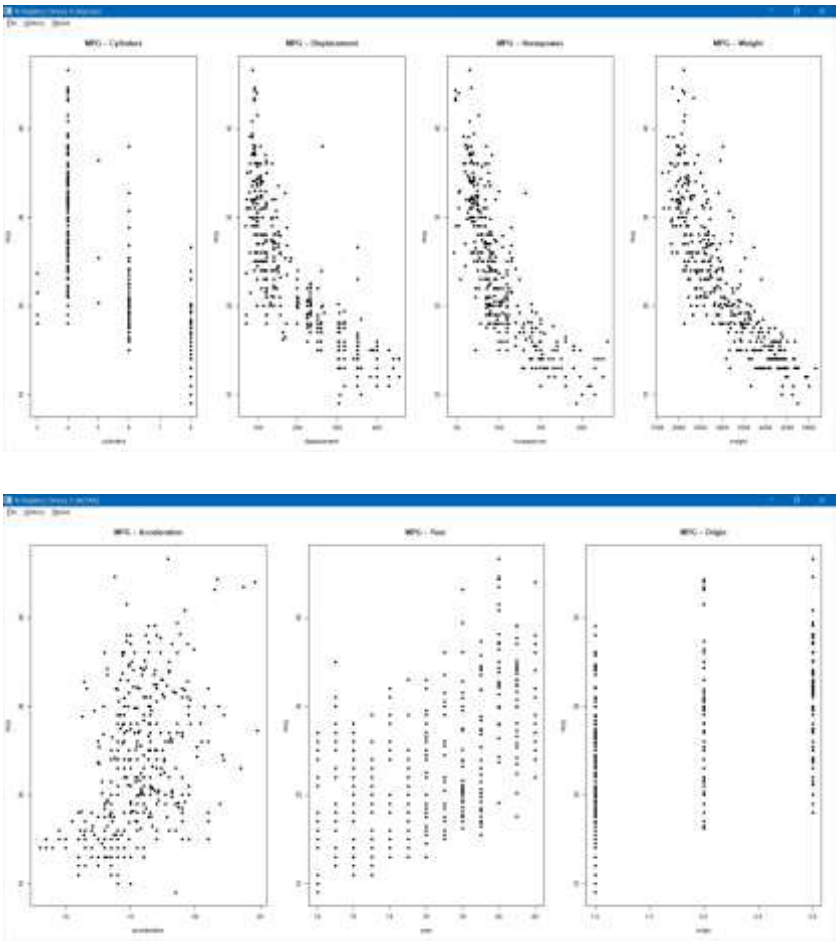
d. I was interested in what boxplots for each of the variables would show me and I was not disappointed. The boxplots showed me outliers for the horsepower and acceleration variables. The rest of the variables were pretty much what I expected.



Kunal Mehta
50352960
kunalkam

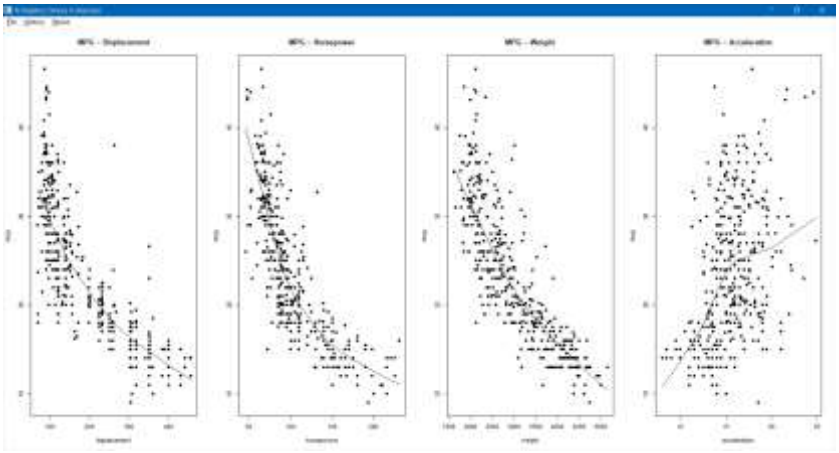


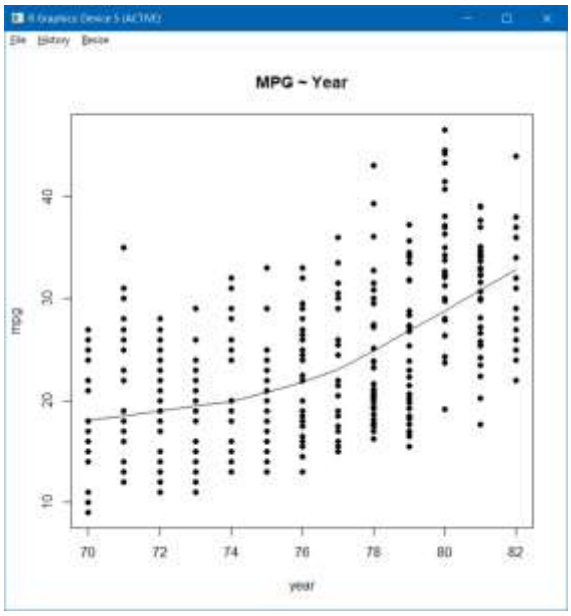
e. The next step was to create scatter plots to show the effect of each variable on mpg. This is as shown below.



Year and Origin were the only two which showed any clear positive effect on mpg. A greater number of cylinders, higher displacement, horsepower and weight, all showed

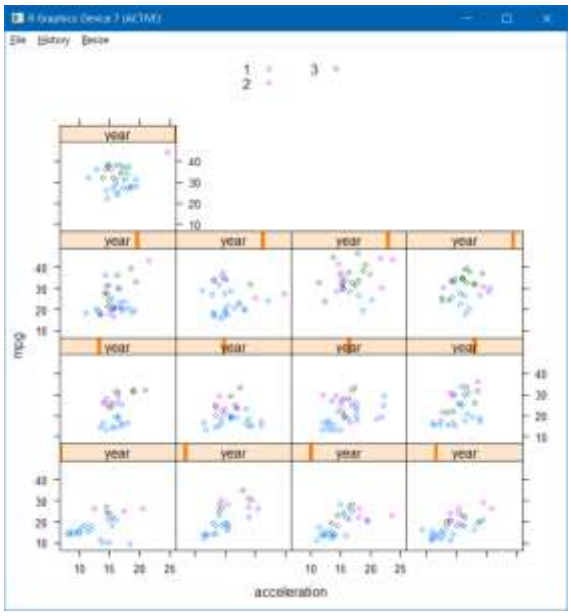
a decrease in mpg. Acceleration was a variable that didn't seem clearly trending in any way. At either rate, I fit smooth curves to confirm those which were even remotely ambiguous.

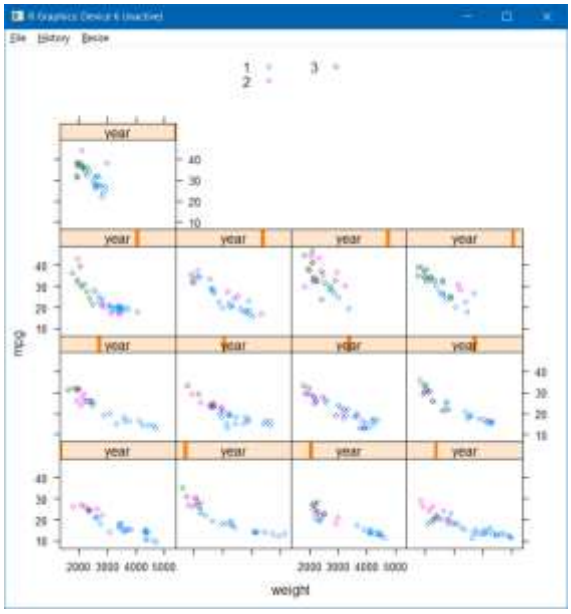




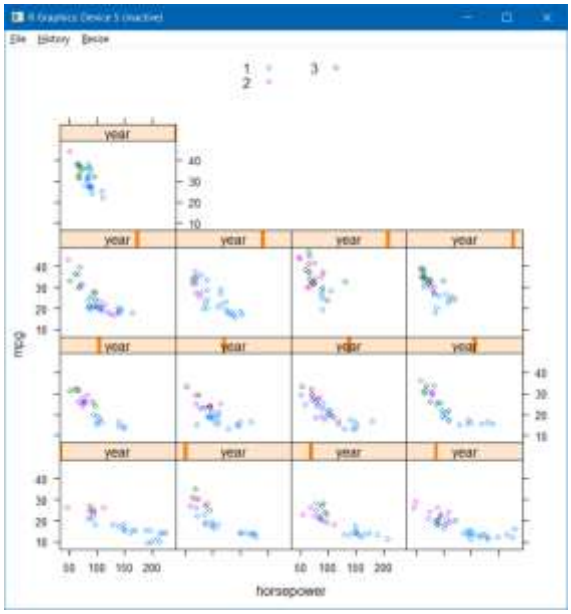
Kunal Mehta
50352960
kunalkam

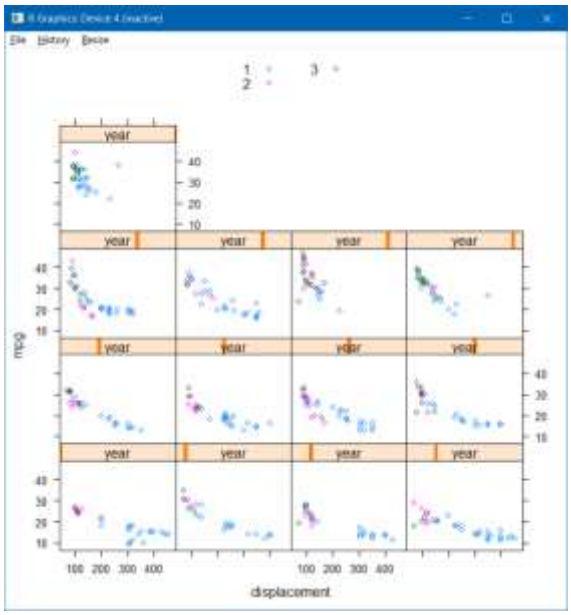
f. Since we had to make a predictive model with mpg as response variable, I set out to examine scatterplots broken down by multiple factors. Keeping the origin variable as the grouping argument, I plotted scatter plots of mpg with displacement, horsepower, weight and acceleration keeping the year as the factor.





Kunal Mehta
50352960
kunalkam





These did not reveal anything more than what I already knew. Acceleration was fuzzy, while the rest show a clear downward trend.

g. The next step was to eliminate the name variable from the data. I was reasonably sure that any more exploration was not going to reveal any new insights about the data to me. I still wanted to examine the outliers the boxplots revealed and therefore I subset them and examined each record to see if they were merely values that happened to be outliers but genuine, or if they were spurious or negative. On examining these outliers, few as they were, it seemed that they were simply extreme in nature but not ingenuine. Therefore I let them stay in the data and saved the cleaned data to a file called CleanAuto.RData

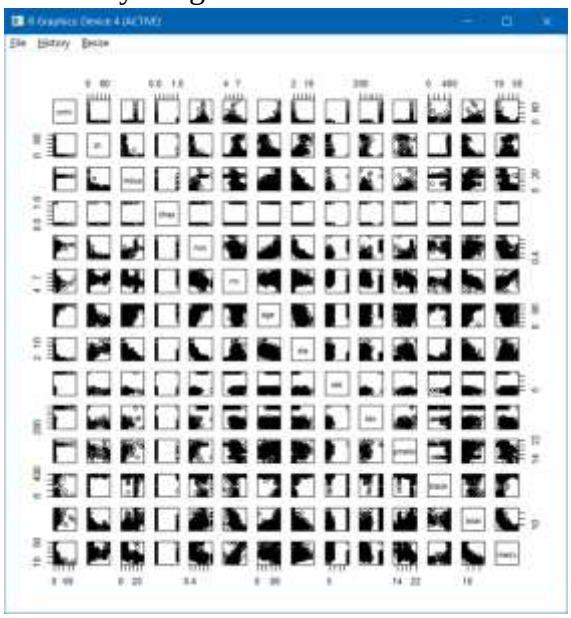
Kunal Mehta
50352960
kunalkam

Problem 2

a. Using the lm() function I performed multiple regression on the data I saved in CleanAuto.RData. Upon applying linear regression on mpg with all the variables (implying that terms are cylinders + displacement + horsepower…) the adj.r.squared value is approximately 0.818.
b. To check which predictors have a significant response to the response, I performed linear regression with all of them and got similar answers in precise numerical terms. The acceleration variable is not a significant predictor, while displacement, horsepower and weight are significant predictors.
c. The coefficients variable for year suggests slight variance from mpg.

Problem 3

a. Pairwise scatterplots – not very insightful.



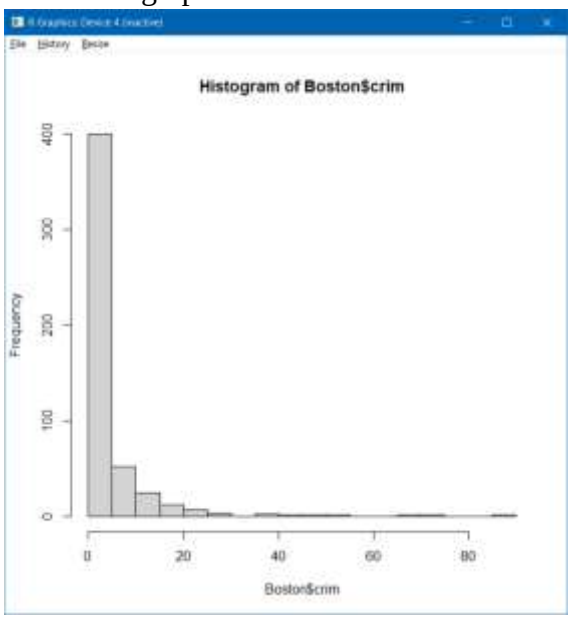
b. The predictors associated with crime per capita

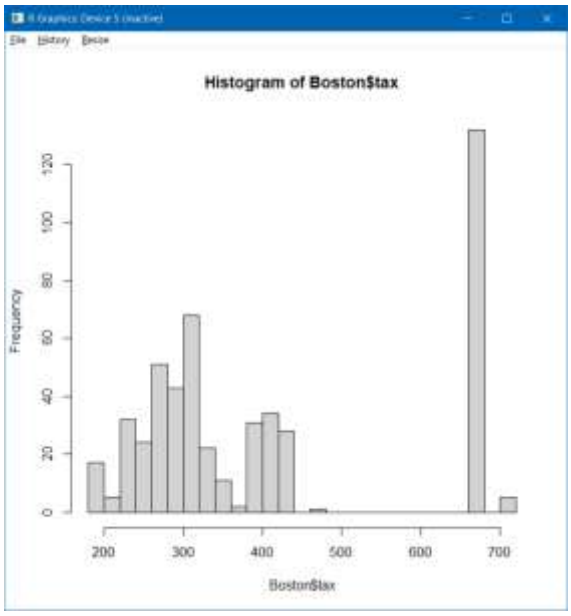
```
                [,1]
zn      -0.20046922
indus    0.40658341
chas    -0.05589158
nox      0.42097171
rm      -0.21924670
age      0.35273425
dis     -0.37967009
rad      0.62550515
tax      0.58276431
ptratio  0.28994558
black   -0.38506394
lstat    0.45562148
medv    -0.38830461
>
```

There definitely is correlation between crime rate and per capita crime.

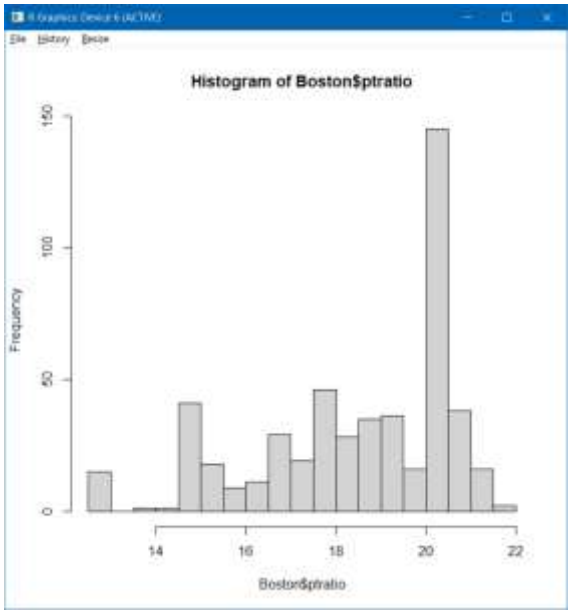
Kunal Mehta
50352960
kunalkam

c. There are definitely suburbs with high crime, tax and pupil to teacher ratios in the dataset. These are shown in the graphs below.





Kunal Mehta
50352960
kunalkam



Based on the mean, median and max values, I separated the records that I found as extreme and stored them in the highcrime, hightax, hightax2 and highptr variables.

d. There are 64 suburbs with 7 rooms and 13 suburbs with 8 rooms as their rooms per dwelling averages. The suburbs with more than 8 rooms per dwelling seem to be typical upper class suburbs with high tax rates, low crime rates, an elderly population.