# Week 2 Assignment – Simple Linear Regression

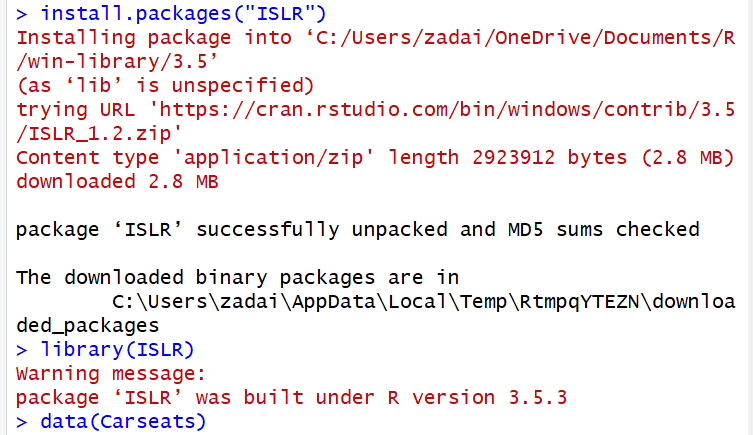
By Zach Adair

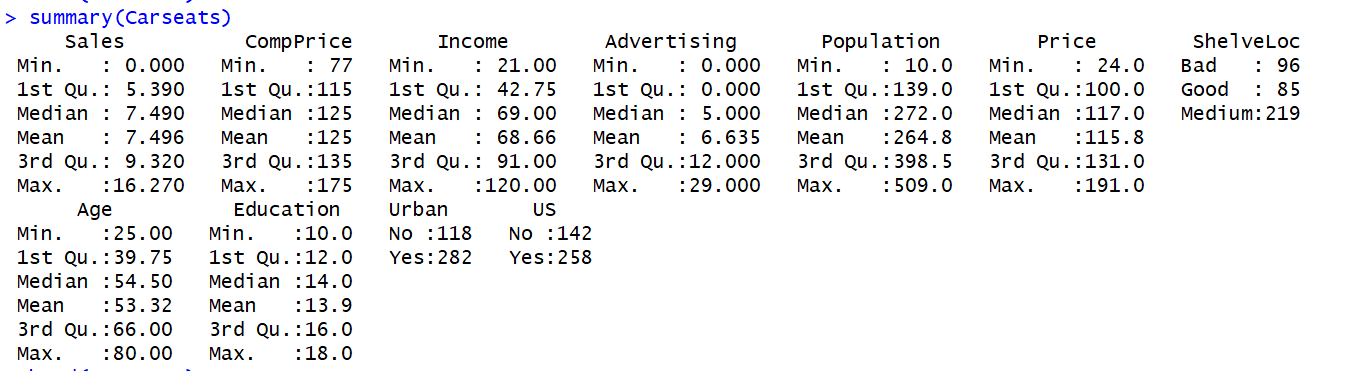
Regis University

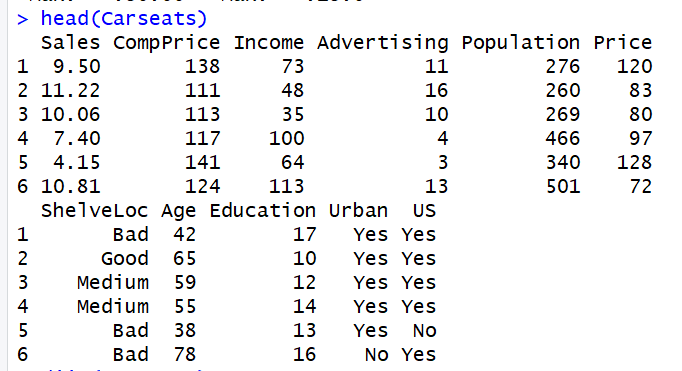
MSDS – 660 Statistical Methods and Experimental Design

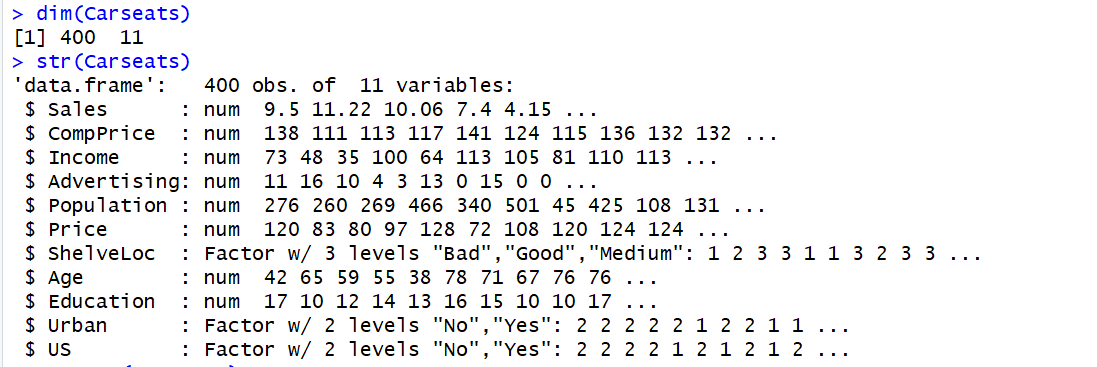
## Load and Briefly describe the Dataset

The dataset we are using is the Carseats dataset from the ISLR package which can be uploaded into RStudio.





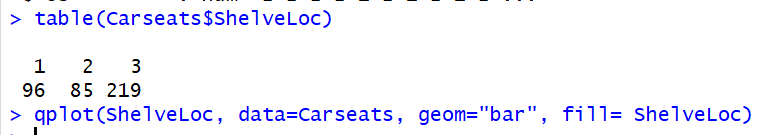


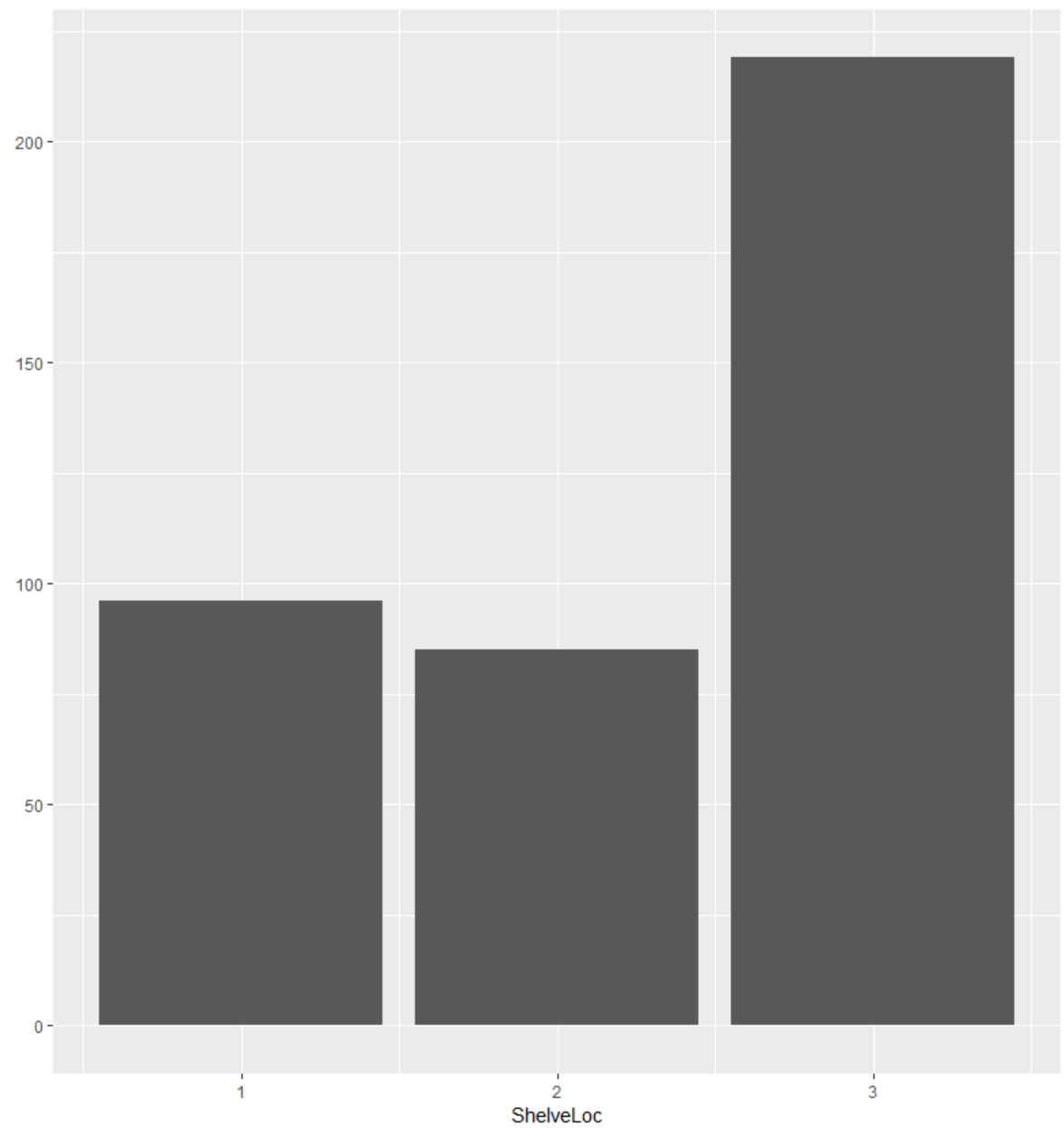


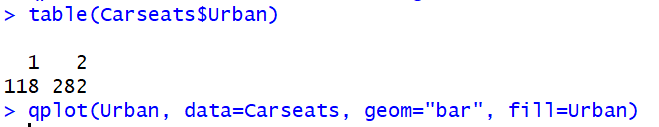
From the summary statistics collected below, there are 400 total observations and 11 total fields. It looks like a lot of the data is numerical with 8 columns being numerical while only 3 are different and they are Factors where the numbers mean something like Yes/No in the Urban or US column or Bad, Good, or Medium in the ShelveLoc field.

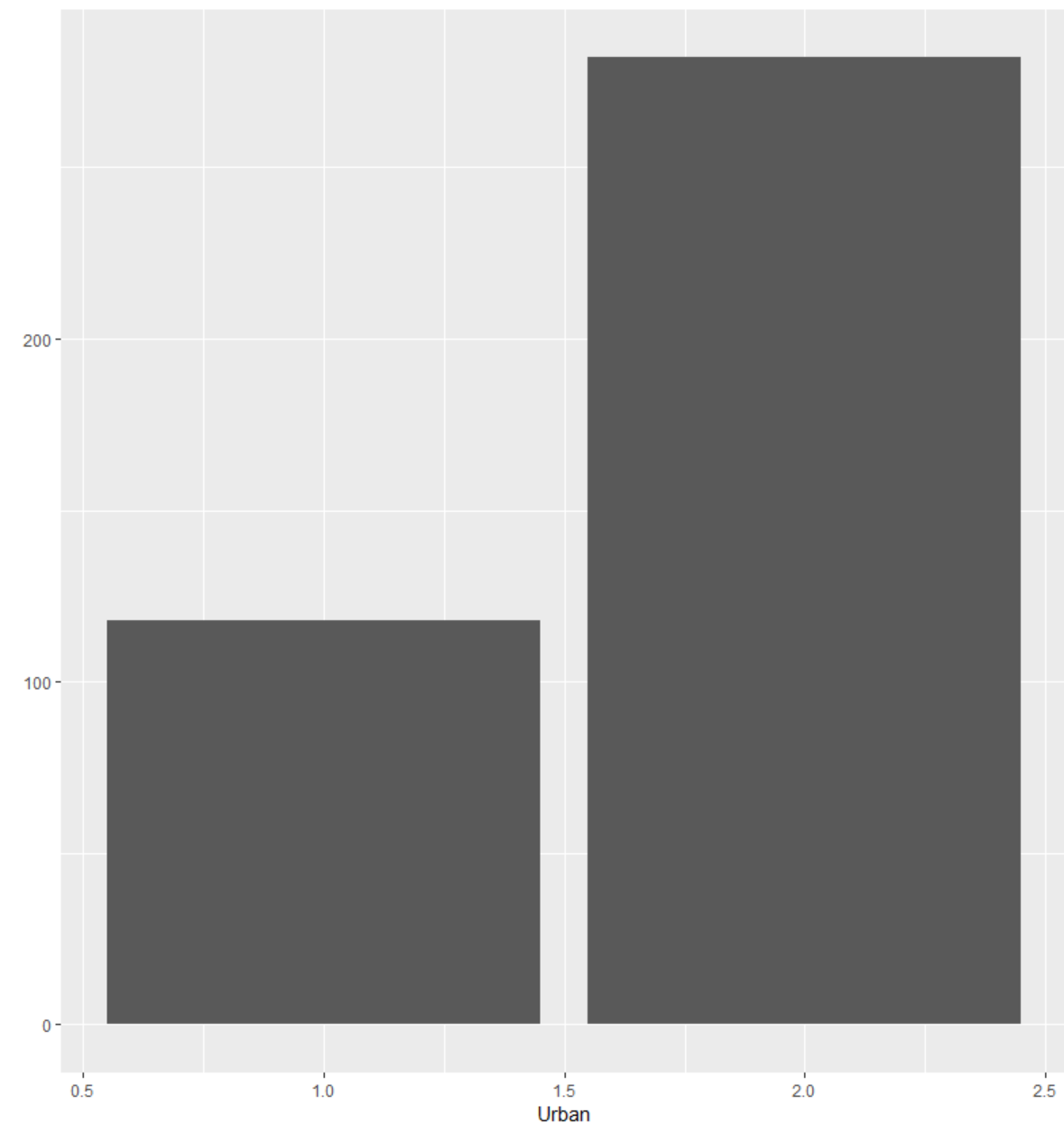
## Perform Data Exploration

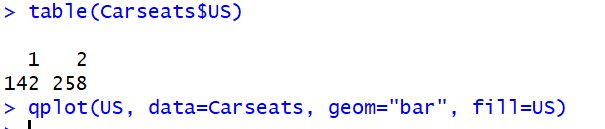
Start with some tables and plots of the data, where are we at with the distribution of some of the different variables in the dataset.

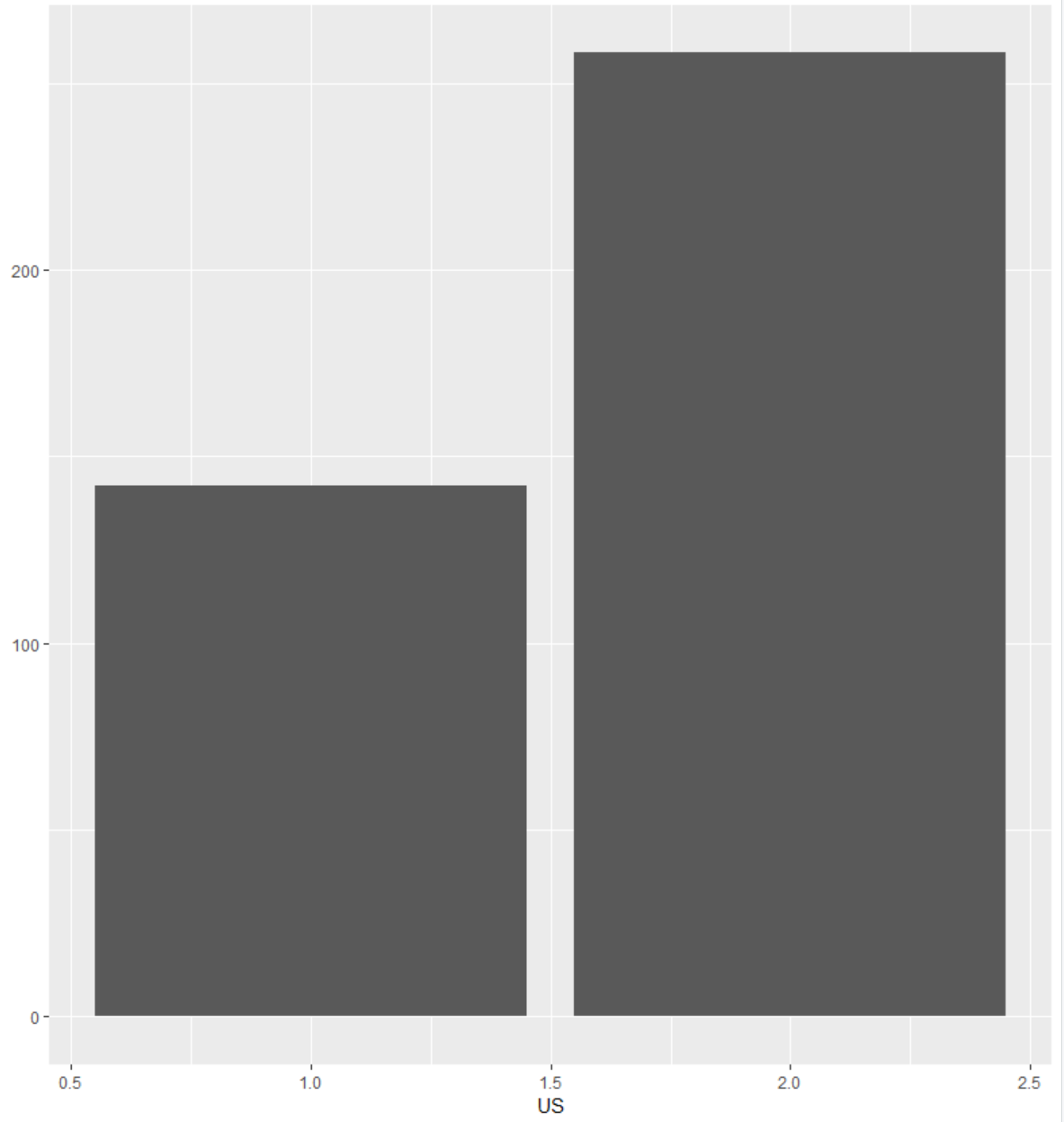






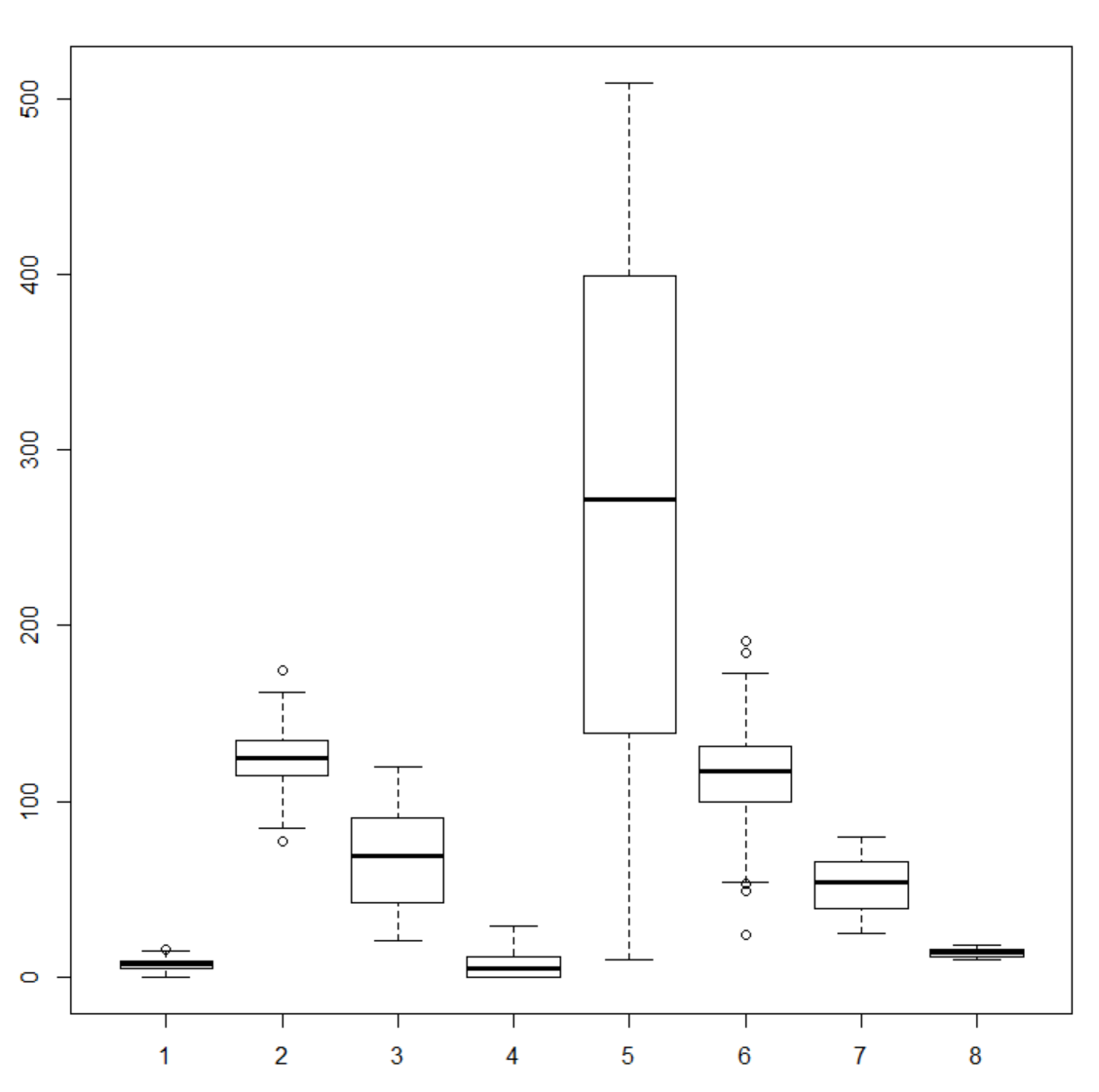






The three Bar Charts represent record counts for each of the individual fields that aren’t calculated by a number, it is easy to notice that the variables are not evenly distributed in any of the three fields I just tested.





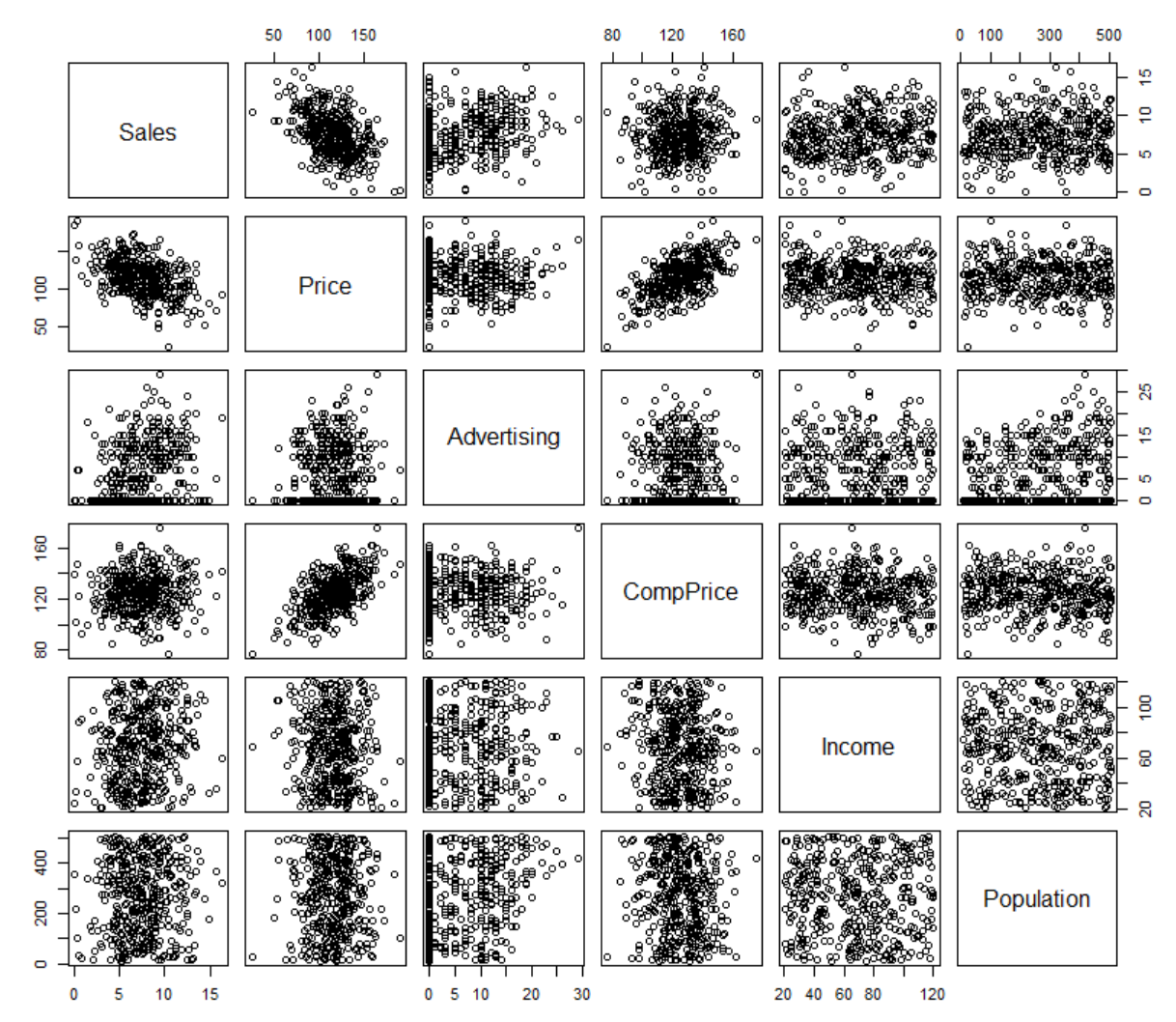
The different counting variables were spread out across the data in their own separate box plots. Not many outliers within the distributions with population having the most variance within it’s distribution by a wide margin, but when compared to the others on the same xy-plane that is something to expect.

From the data exploration done I can confidently state the distribution of the data is not evenly distributed.

## Perform Pairwise Scatterplots is there any Correlation?

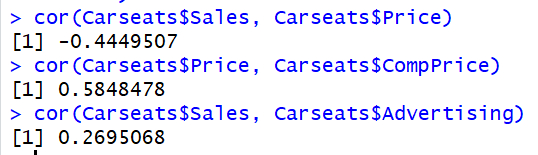
Scatterplots trying to find any correlation between variables:





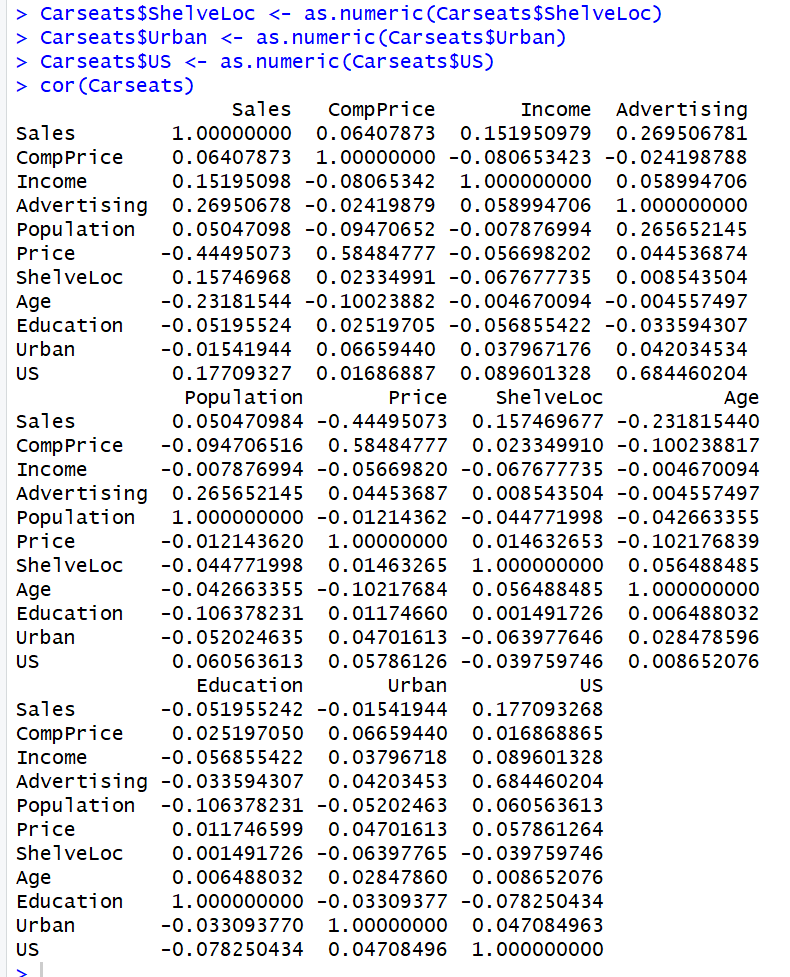
Looking at all different variables and comparing them against one another on the scatterplot matrix above there isn’t much of a correlation between two data points. The ones that look to might have a correlation is Sales vs. Price, Price vs. CompPrice and Sales vs Advertising. Unfortunately it doesn’t appear to be any correlation between several variables which I was surprised to not see at least one other for population and income.

Next, I want to test the correlation of the variables which appeared to have a bit of a correlation between one another. So there’s three sets of variables in which I would like to test the correlation of, those are Sales vs. Price, Price vs. CompPrice, and Sales vs. Advertising. I’m not expecting strong correlations but ones that would point to be pretty correlated from the way their numbers are displayed on the scatterplot matrix.

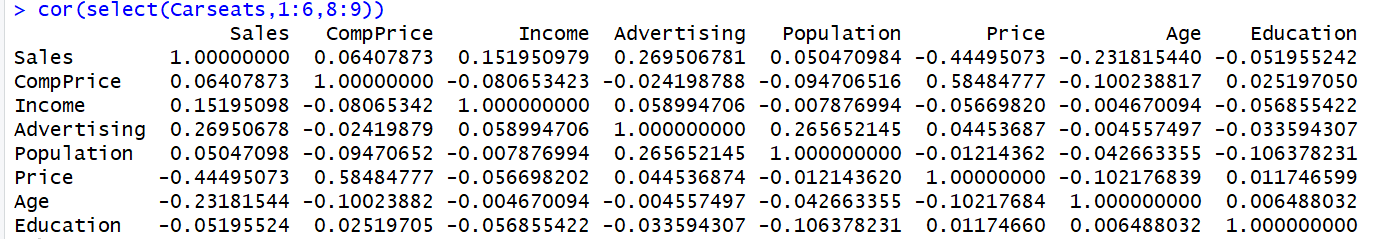


From the result, there isn’t much of a strong correlation between any of the variables I’ve isolated and tested. The strongest correlation appears to come between Price and CompPrice but that was about 0.58 which isn’t necessarily that strong. The correlation that was negative was Sales vs. Price which was at -0.44 which was to be expected as negative based on our scatterplot matrix.

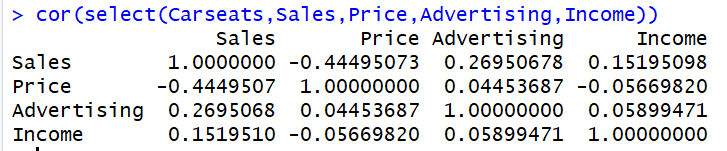
Instead of testing each variable individually for we can also take a comparison of each variable and find the correlation of each compared to a counterpart variable also in the dataset to test correlation. Doing this hopefully I’ll find a strong correlation that I didn’t catch from earlier.



From the correlations matrix we can confirm there just isn’t a huge correlation between any of the variables which I spotted earlier, which isn’t much of a surprise considering what I found didn’t have a strong correlation to start. Let’s take a little bit more of a refined look now and pull out the Shelve location and the Urban and US fields from the data and compare the correlations of all the other fields.



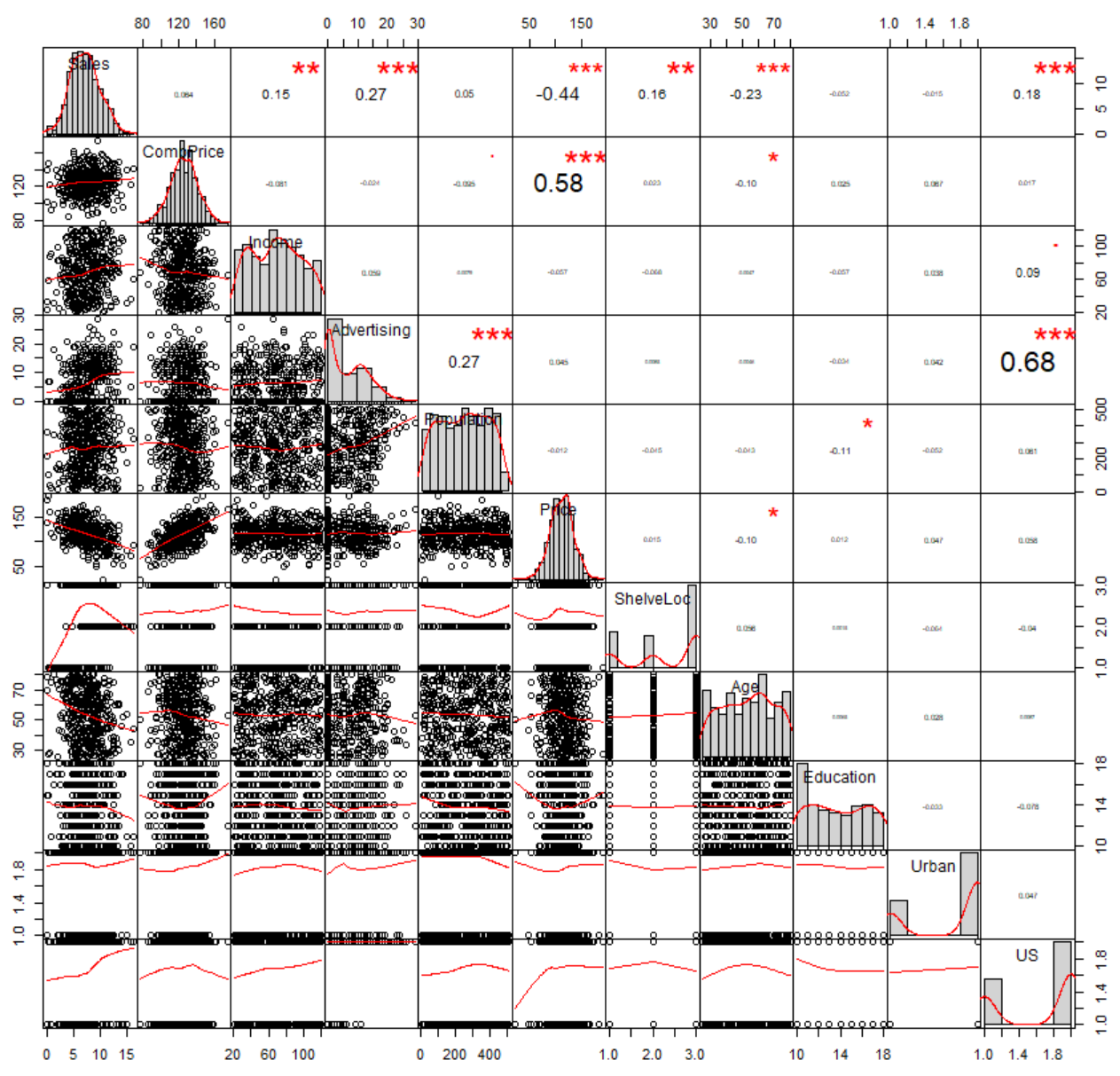
Refining the look doesn’t give me a better glance at what correlations could be really positive or negate, there are some small ones but nothing really eye catching. Now I’ll refine the fields again, this time just pulling out most of the ones I isolated at the beginning of measuring correlation: Sales, Price, Advertising and Income.



This is a much more compact look at all those different variables to view if there is any notable correlation, which besides the ones pointed out there doesn’t appear to really be any. From here I’m going to chart correlation upon the Carseats dataset and try to denote any analysis from that.

After bringing in the PerformanceAnalytics Package into RStudio.



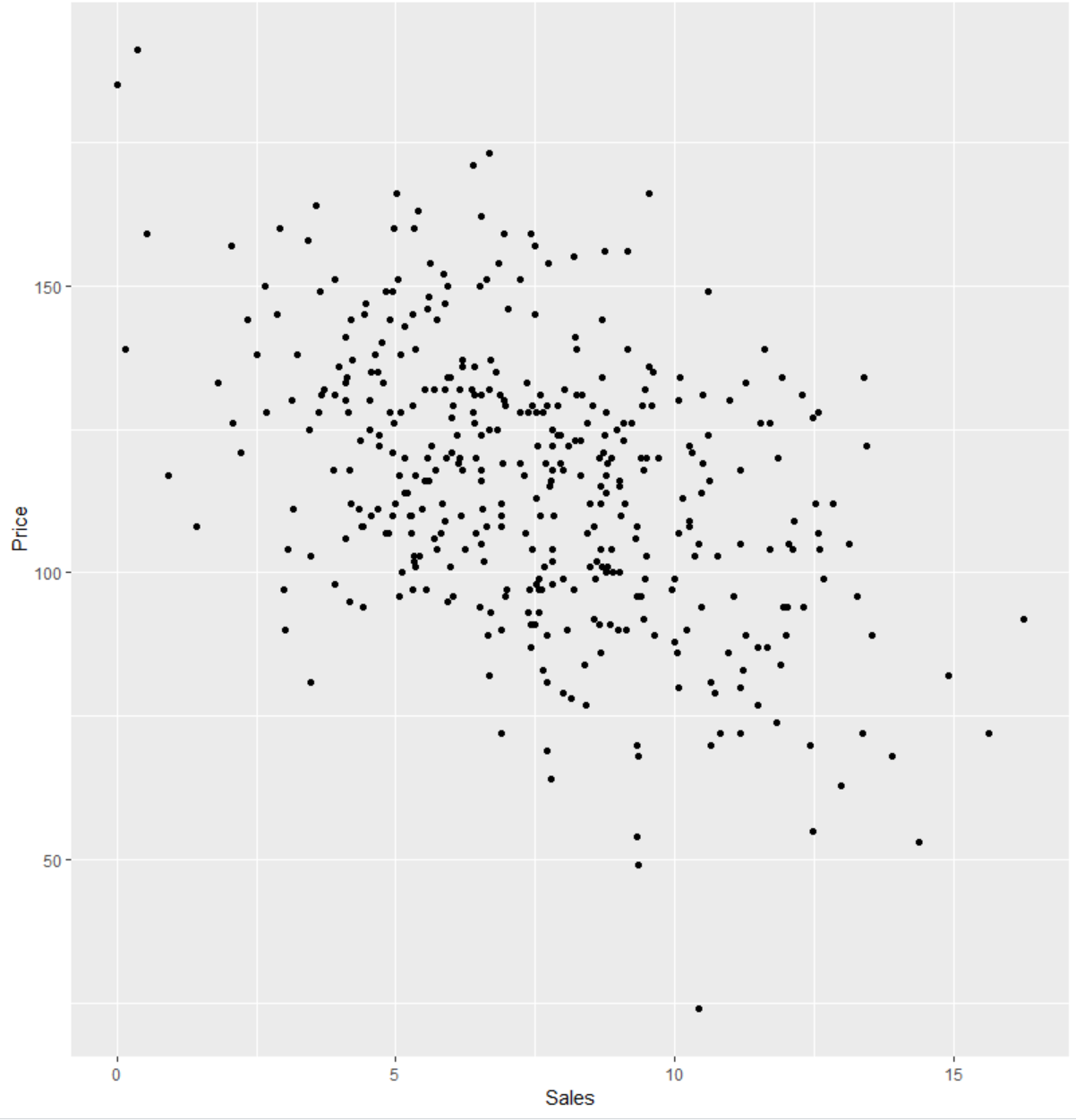


From our chart here were able to see much easier some bigger correlations I was unaware of, the distribution of data for each variable and then each relation is mapped out variable vs variable. The correlation relation that surprised me to see most is Advertising vs. US, which it really shouldn’t based on the distribution of US for advertising it seems the Carseats data didn’t really have a lot of foreign marketing. One that I wasn’t surprise to see on here but I didn’t highlight was CompPrice vs. Price, I just figured those two were a given and didn’t look very deeply into that relationship.

## Select an Independent and Dependent Variable and run a Scatterplot of the Variables

The variables I selected are x = Sales and y = Price. My goal is to predict the price based on the number of sales.





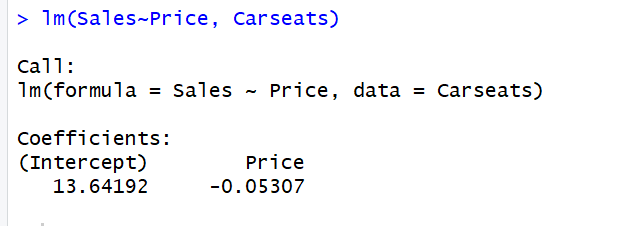
From the scatterplot it appears the data would say that as x increases, y decreases.

## Perform a Linear Regression on Variables Chosen using 5% Confidence Level

H0: As the Sales increase, the price decreases.

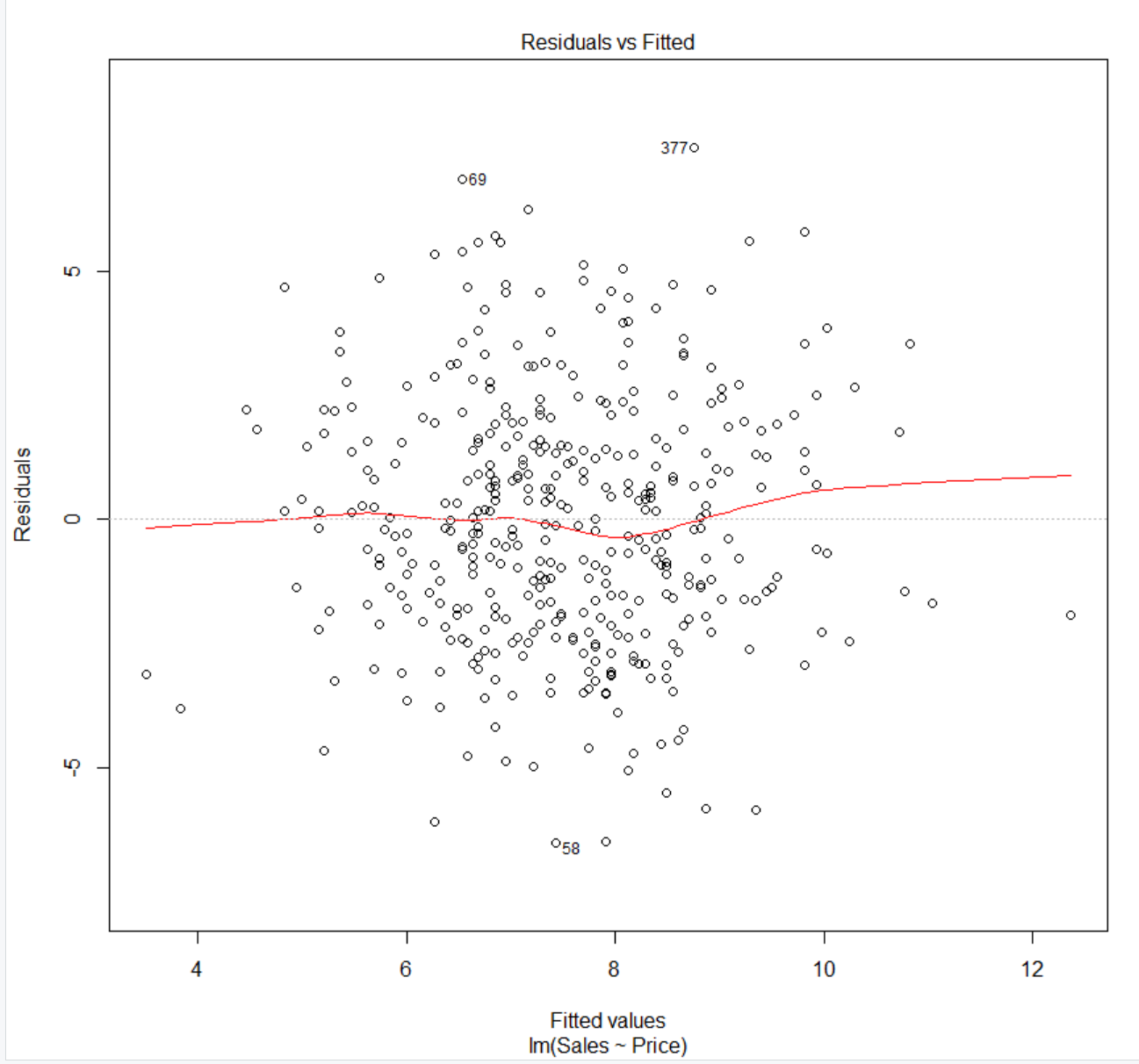
HA: As the Sales increase, the price stays the same or increases.

## Linear Regression

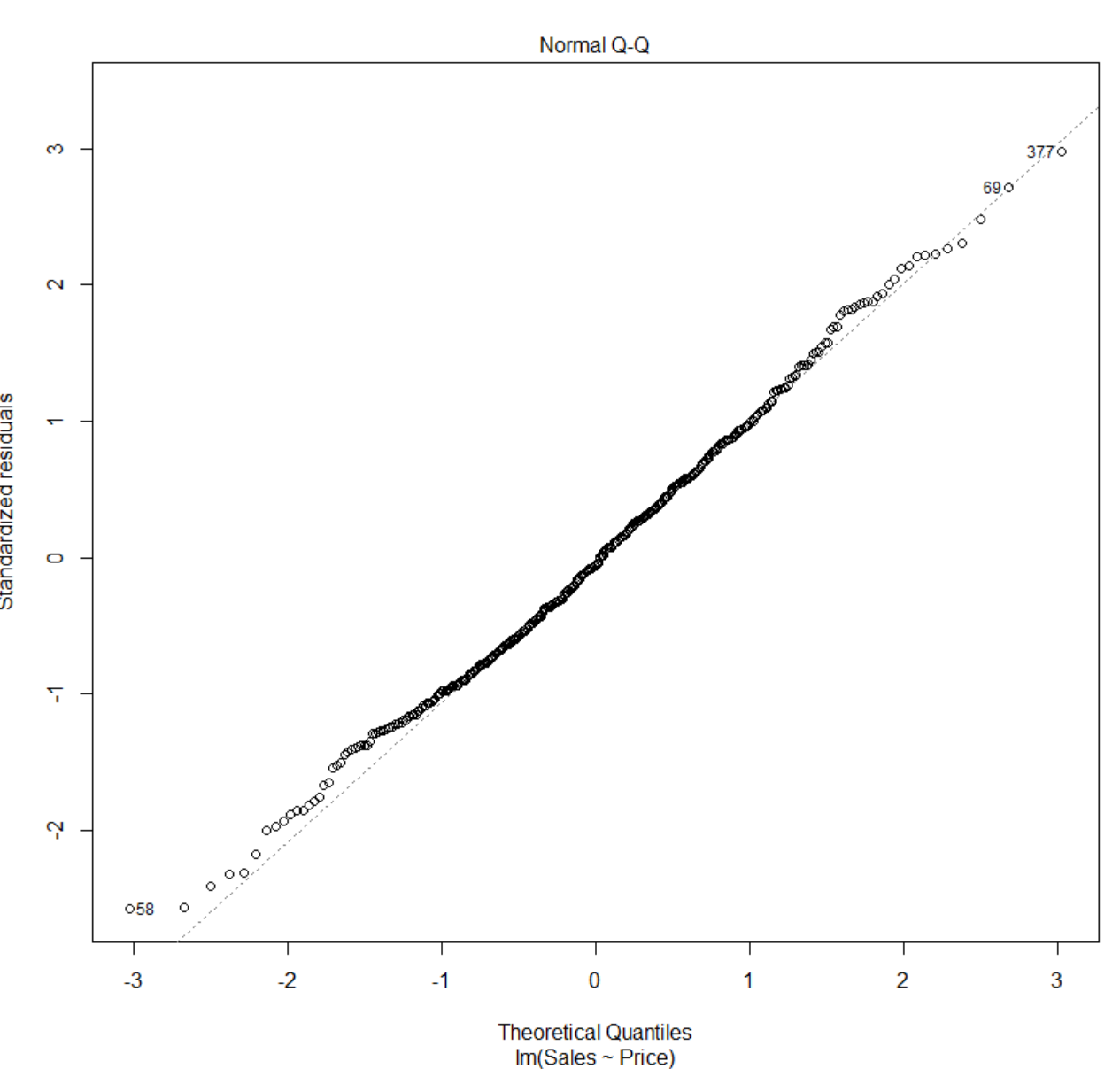


Then we plot the model that was just created:

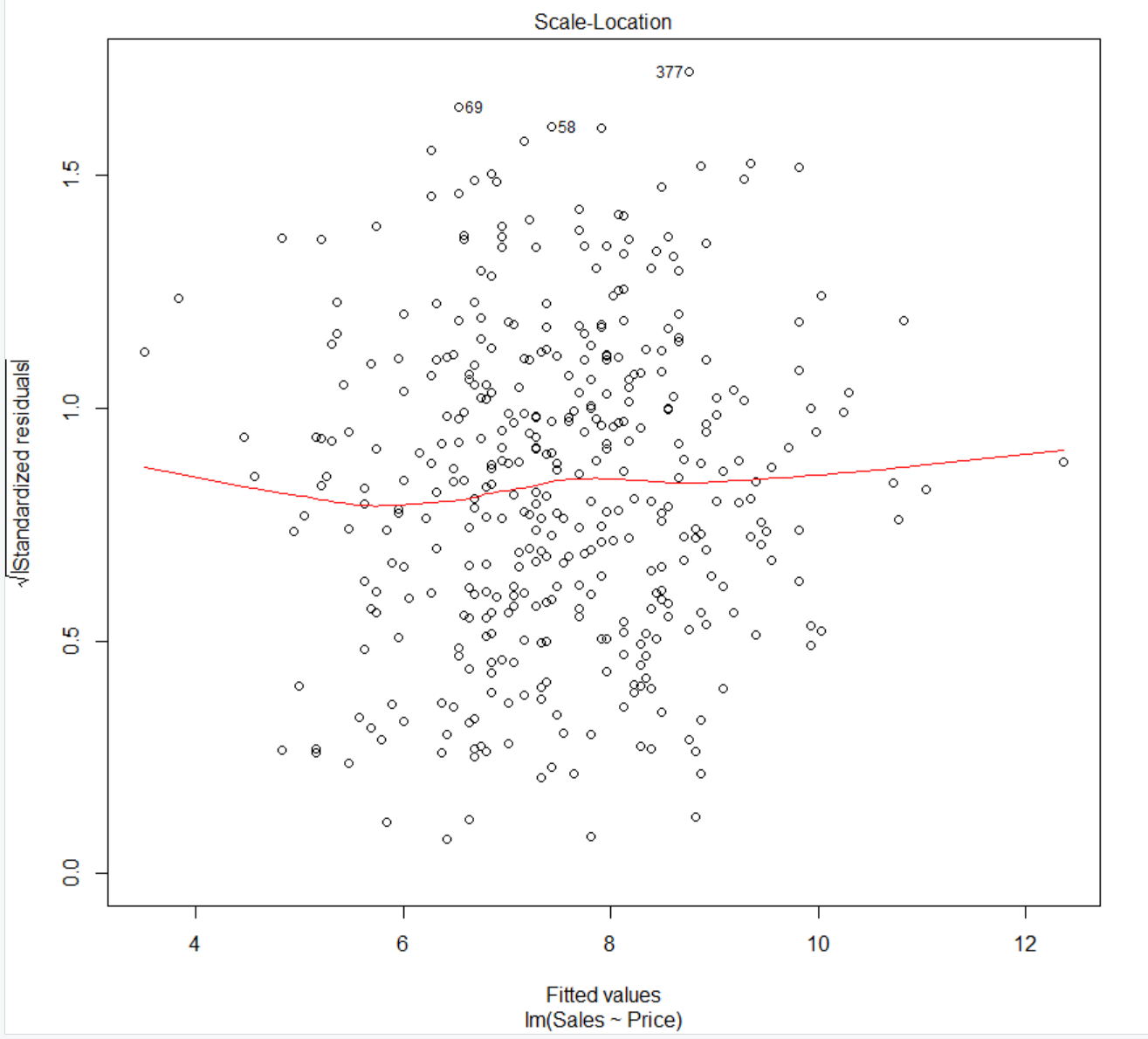




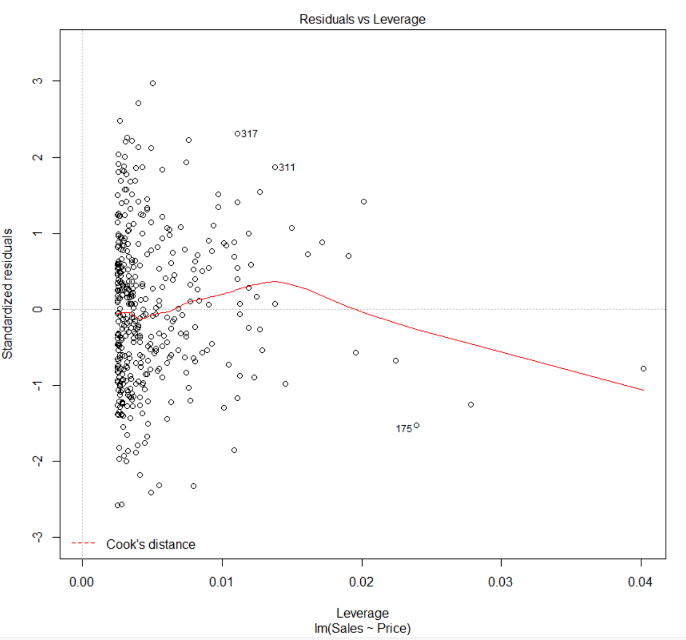
The residuals vs the fitted values appear to give a weak correlation between the two, but the data does appear to slightly rise as the fitted values rise.



The Normal Q-Q plot the data appears to move right along the (1,1) regression line with some tailing off on each of the ends but through a majority is in step with the line.



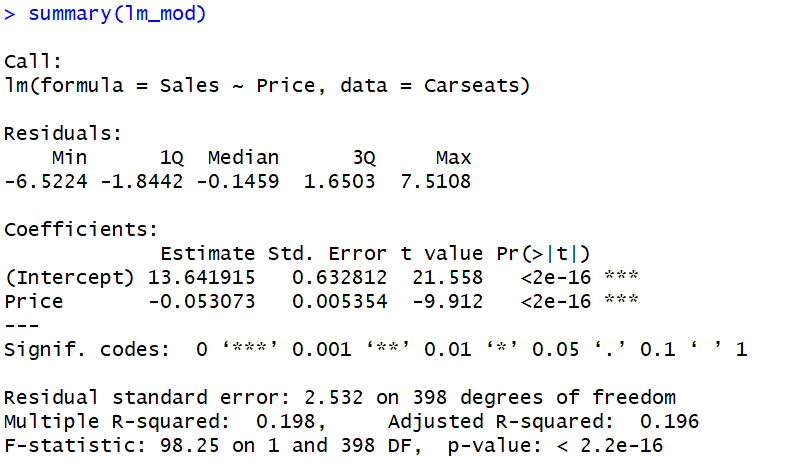
The Scale-Location plot is similar to the residual fitted line the only difference is were now find the square root of the residuals. This cause the data to be more spread out on the plot but the regression line appears to not change much between plot 1 and plot 3.



The last plot is measuring leverage and the standard residuals, A lot of the data appears really sporadic around low leverage point and the regression line begins by going up slightly but turns downward sharply.

From these linear model plots it is still unclear on if the effect of Sales on Price.

Next we will bring in some summary statistics which will give us the R-Squared values as well as the p-value.



## 7, 8, 9. Explain how to read the results, reject the null hypothesis or not? Is there a linear relationship between the predictor and response variables chosen? Verify model assumptions.

From our summary statistics for our model we can be for certain for how well the model performed based on the Multiple and Adjusted R-Squared numbers, though they are quite low, but the degrees of freedom being 398 is not good because that is really high and you would want your degrees of freedom to be at or as close to one as possible. From the p-value we would reject the null hypothesis I stated earlier. The p-value is a 2.2e-16 which is a very small number and whenever the p-value is less than or equal to our alpha level (in this case 0.05) we must reject the null hypothesis in favor of the alternative.

There is no evidence that there is a real linear relationship between the predictor and response variable. The reason why is based on the model, the plots came back inconclusively for there being an actual relationship between Sales and Price. Then in the summary statistics values we would deduce after seeing the R-squared values and a p-value that had us reject the null hypothesis that there is most likely no relationship between our predictor and response variables.

There is no way to verify any of the model assumptions because the model performed so poorly. From the plots we aren’t able to deduce any kind of linearity between the two variables, the data isn’t distributed normal and the variance appears to be sparse and populated mainly in the center where both sales and price are low but once each grows the data goes off into all directions.

## Summary

I found this exercise very interesting and I felt like I learned quite a bit about RStudio and more capabilities it possesses as well as learned some detailed information about Linear Regression. I was disappointed in my model I was hoping I could confidently find a relationship between sales and price, if I could choose over I would have chosen sales and profit because I believe those two variables have a better chance of having a relationship than sales and price.