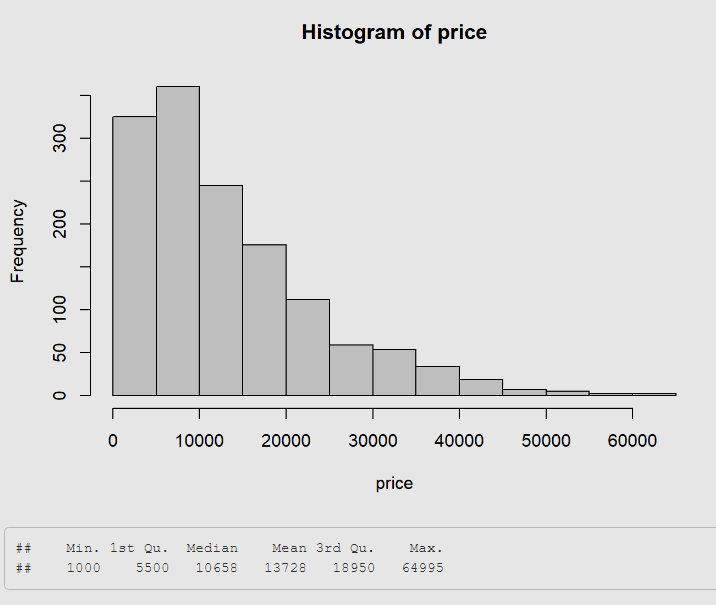
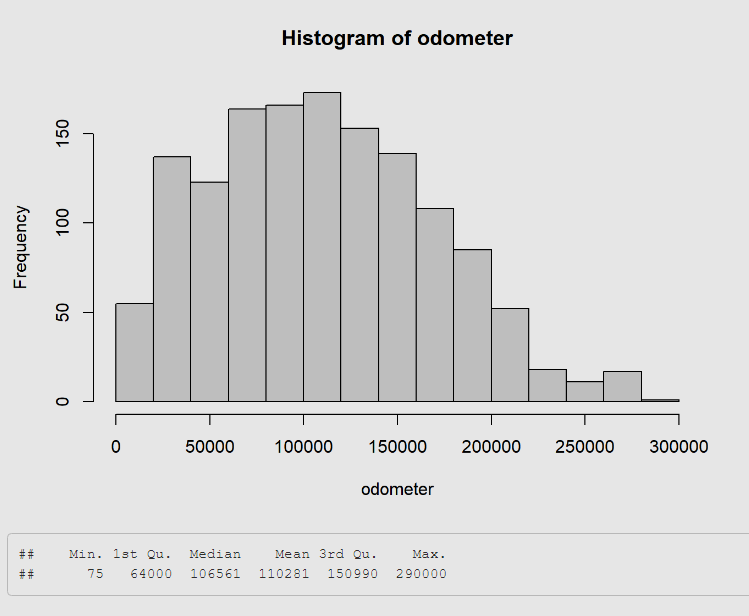
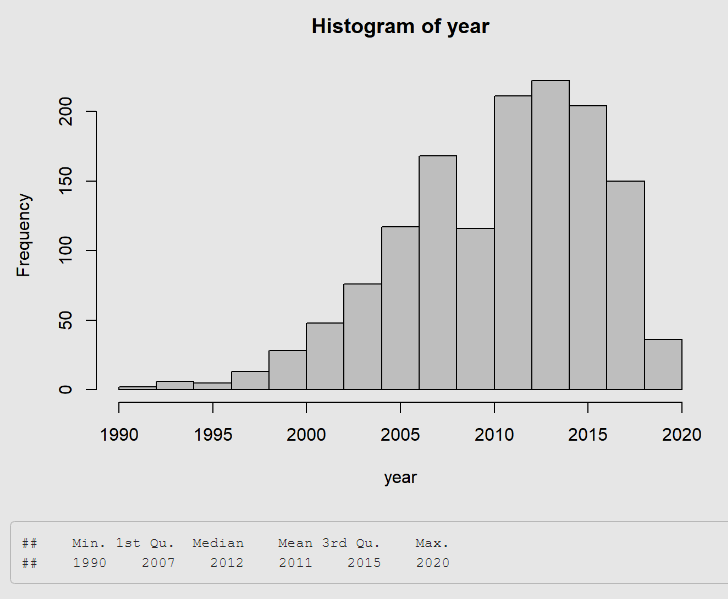
Exploratory Analysis

For our Exploratory analysis stage, we started off by creating histograms for all variables in the dataset, we then followed that by narrowing it down to the variables that we planned on using in our model, which we agreed upon price, odometer, and year. The reasoning behind selecting these variables is as follows, price gives us what a value as to what the car was sold for, this will be our response variables. While the odometer and year variable give us something to use as explanatory variables. We started our explanatory analysis by creating histograms for all the selected variables.



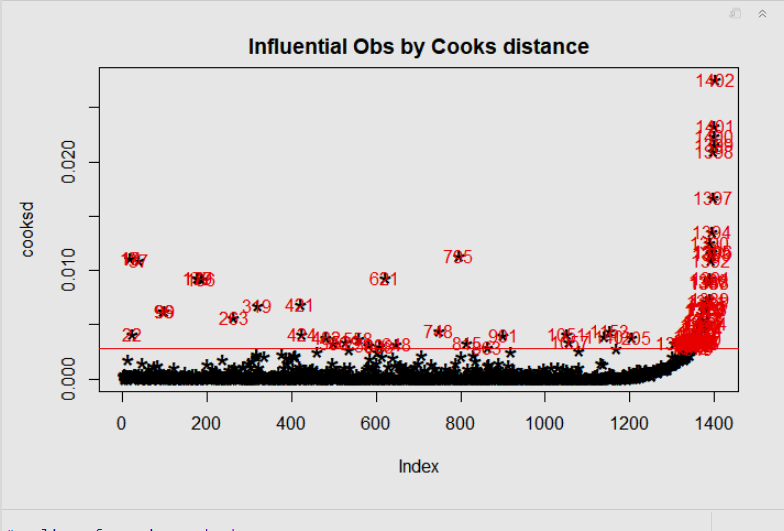




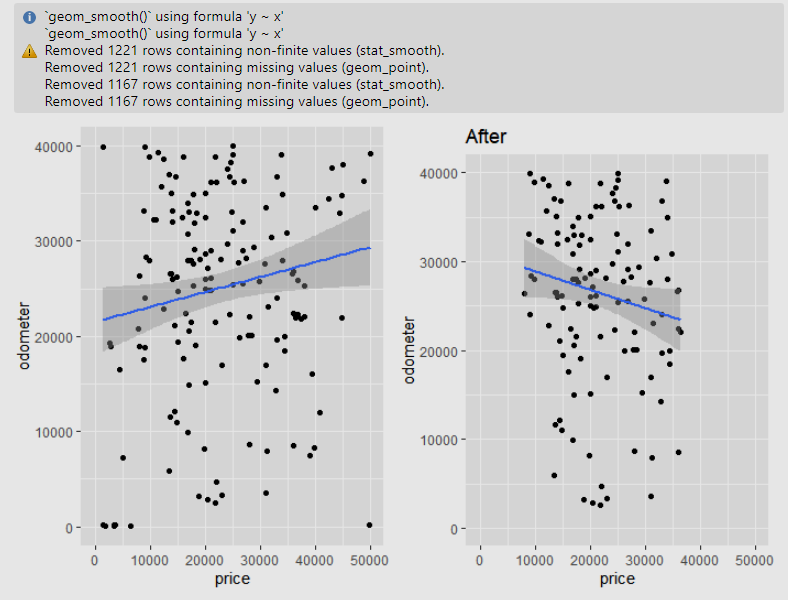
Upon looking at the histograms we noticed that the histogram for price seems to be skewed to the right. The histogram for odometer is also skewing to the right but it is not as much as price, contrary to price and odometer the histogram for year is skewing to left instead. So, these histograms told us that our data does not a bit of modifying to be able to use them for our model. To get a better understanding of the skewness we started off by creating summaries for each of the selected variables. Which showed us the values of price dramatically increases after 3rd Quadrat, from 18000 to 64000. This explains the skewness in the graph, since there aren’t many values after the 3rd quadrant, but it is causing the skewness. We did the same for odometer and year, which showed us that odometer has the same issue after 3rd quadrant where the values dramatically increase from 150000 to 290000. This directs us to think that odometer and price have a positive correlation number. We followed our analysis by creating summary for year, in which we found the values dramatically increase from Min to 1st quadrant, this directs us to realize that price and year has a negative correlation value.

After histogram and summaries, we decided to take care of the outliers and influential points, we used cooks’ distance to take care of the outliers. To clearly see the outliers marked by R, we created a graph and added a cutoff line, which then marked all the indexes of the outliers and removed the outliers. We first started with price and odometer, the cooksd graph looked like there are a lot of influential points, especially after the 1300 index. This can be also be seen in the outlier’s part, where about 1221 outliers were removed by R. The graph for outliers turned a lot different, especially the midline for the outliers switched from facing in the upward direction to a downward direction.

Price and Odometer:



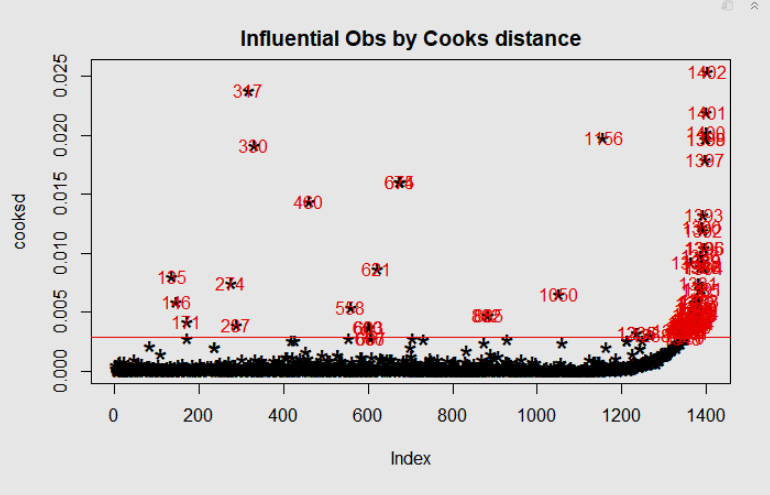
Influential point:



Outliers:

We did the same cooksd influential points and outliers for price and year. The cooksd graph looks like the influential points drastically increase after the 1300 index, this leads us to believe our dataset has some issues after index 1300. Also, there are a lot more outliers with higher than 0.010 cooksd in price and year. We then started working on outliers, and upon running the code the results were quite different from price and odometer, since the code only removed around 3 variables. This makes us to believe the cooksd and influential points were a lot like price and odometer, since the number for removed variables was a significantly less than before. After taking care of the outliers and influential points we started working on the data mining technique and analysis of results.

Influential points:



Outliers:

