**Used Vehicles Analysis**

DSC 341-601

Professor Hamilton

Group Members: Kavya Ramasharma, Jay Patel, Aamir Patel, Parth Rana

## **Data Description**

The dataset that we manipulated throughout this quarter contained information on cars sold in America. We accessed this data set of used cars sold on craigslist from Kaggle. After further analyzing the many variables that were present, we decided to contain the data significantly by only assessing the value of used cars sold in the state of Wisconsin. All variables that were insignificant to finding car values were removed to determine the price at which the car was sold, year, make, model, color, condition, engine type, fuel type, mileage, title status, transmission type, drivetrain, as well as the type of car sold.

Through modeling with R Studio, we were able to better analyze the dataset and determine which variables had significant effects on the price of a car. The initial step that we took prior to creating models, plots, and graphs was to prepare the data adequately. The data set was cleaned using Microsoft Excel to ensure that none of the required variables had any missing values. Variables that didn’t have as much of an effect on the price of cars were removed such as image links, and latitude and longitude for the car sales. Once the dataset was prepared, we were able to identify significant outliers and influential points that greatly affected the slope. Over 2,500 rows containing non finite values were identified and removed using scatterplots.

We created three histograms using the dataset to better understand the degree of variation and to show the distribution pattern. The histograms created were for price, odometer, and year. The histograms of price and odometer are skewed to the left, while the histogram of year is skewed to the right. Using R Studio, we were able to create a model that made it easier for us to calculate predictions for the value of a used car. Our first model made it possible to calculate the price of a car using variables such as odometer and year. The predictions that model 1 gave us using these variables were: $17216.014 for the first car, $18,382.075 for the second car, and $6,947.624 for the third car. For model 2, we used the same three predictions as model 1. However, model 2 only used the odometer variable to predict the value of a used car.

## **Exploratory Analysis of the Data**

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.

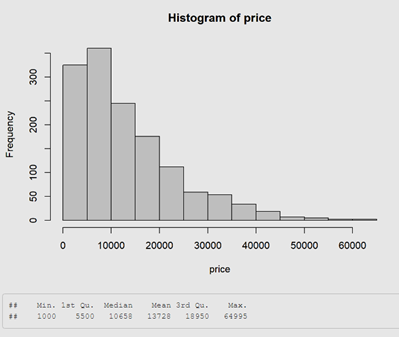


Figure 1

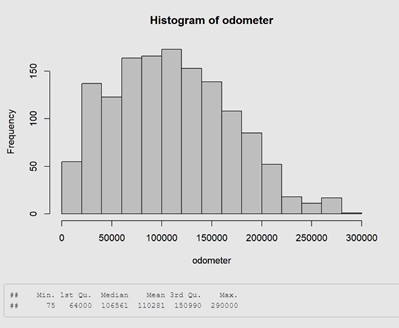


Figure 2

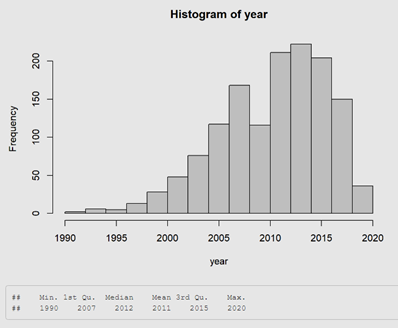


Figure 3

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 require 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 value of price dramatically increased after the 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 the 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 a summary for the year, in which we found the values dramatically increase from Min to 1st quadrant. This directs us to realize that price and year have 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 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:

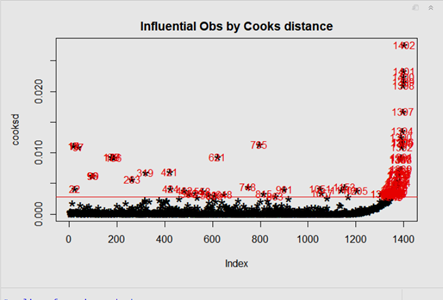
Influential point: 

Figure 4

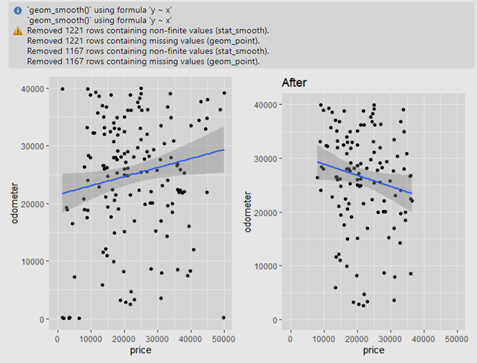
Outliers: 

Figure 5

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 believe the cooksd and influential points were a lot like price and odometer, since the number for removed variables was 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.

Price and Year:

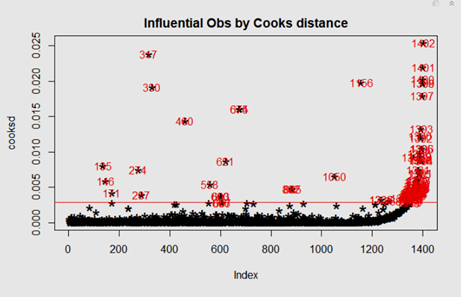
Influential points:

Figure 6

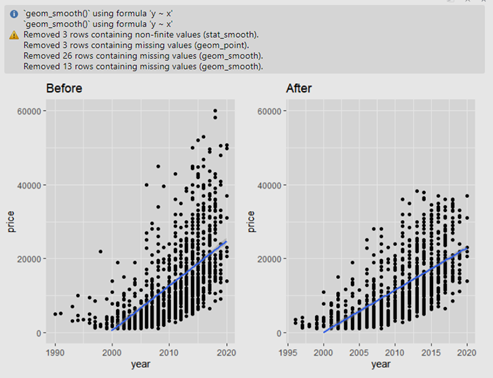
Outliers:

Figure 7

## **Data Mining Technique and Analysis of the Result**

After cleaning up our dataset by removing outliers and influential points. We used multiple different data mining techniques to interpret the data and put it into models which would make it easier for us to analyze the data. These models would also make it easier for us to make predictions about the value of used cars using variables like the vehicles model year, and odometer reading. We began the process by creating histograms for price, odometer, and year of the cars sold in Wisconsin. Our histogram for Price (Figure 1) concluded that the lowest price a car sold for was $1000 while the highest price was $64,995. The output also told us the median was $10,658 and the average price for a used car was $13,728. Our histogram for Odometer (Figure 2) told us the highest mileage on a used car was 290,000 miles, the lowest was 75, and the average was 110281 miles. Our histogram for year (Figure 3) told us the newest car sold was a 2020 model, the oldest was a 1990, and most of the used cars sold in Wisconsin were a 2011 model year. Using residual plots we were able to build a model for the price using the two variables year and odometer. Model 1- Price= -0.000001707 + 0.08573(year) - 0.04174(odometer). This model gave us an R-squared value of 0.4787. We can conclude from model 1, model year has a positive impact on the resale value of a car. Meaning the higher the year number the more the car is worth since we know cars that are newer are worth more we know this is accurate. Model 1 also told us odometer mileage has a negative impact on the resale value of a car. We can confirm this since we know that as a car obtains more miles on the odometer, it is worth much less. We decided to create another model using only the odometer variable. Model 2- Price= 0.0002175 - 0.08292(odometer). Model 2 gives us a R-squared value of 0.3146 and since the R-squared value for model 1 was greater than the model 2 value, we can conclude that model 1 is the better model when it comes to predicting the price of a used car. Using the models we created from residuals, we were able to make a couple predictions for the price of used cars with variables we made up.

We made a total of 3 predictions using both of the models we created. Our first prediction was a 2014 model year car with 20000 miles on it and was priced at $15,000. Our second prediction was a 2016 model year with 59000 miles and priced at $12,100. Our third prediction was a 2005 model year with 130000 miles and listed for $9,000.

## **Validation and Testing**

1. Detail Plan - It is important to lay out your roadmap for data validation. Setting your overall expectations for the number of iterations required in the roadmap will help you resolve the issues along the way. Our group created a project outline that entails data preparation, modeling, and deployment of our data. By creating a GANTT chart, our group was able to successfully meet our expectations.

2. Validating the database – Dealing with large datasets comes with great responsibility in determining the number of records, size of data, null values, and etc. Our group decided to focus on a particular set of data within the large dataset which is the state of Wisconsin. By performing this necessary step, we were able to load and execute the data in a friendly manner.

3. – Validate Data Formatting – The end-users should clearly understand data whether or not it is meeting our expectations or not.

4. Sampling – Since we only utilized data from Wisconsin we were able to test a small amount of data and see if it is meeting our requirements and expectations. Indeed it did meet our expectations and decreased the error rate for data and increased accuracy of the data. Then we proceeded further in our roadmap.

## **Discussion and Conclusion**

In conclusion, building these models helped us predict the price of used cars sold in Wisconsin. By using multiple variables, such as year and odometer, we were able to accurately predict the price that these cars would sell for. As emphasized above, price served as a response variable to year and odometer which were the explanatory variables that we had selected. Based on viewing the histograms created and analyzing whether they were skewed to the left or right, we were able to determine that odometer and price have a positive correlation value while price and year have a negative correlation value. As mentioned earlier, models and graphs were created and utilized to further understand and analyze the dataset that we chose to work with. Based on the two models that we created, we were able to determine direct correlations and better understand exactly how much the price of a used car is affected by factors like odometer and year. Model 1 confirmed that price decreases as a result of obtaining more miles on the odometer, as there was a clear negative impact on the resale value of a car. We were also able to determine that Model 1 was more accurate to use when predicting the price of a used car, compared to Model 2 since the R-squared value was greater than that of Model 2.