**Data Analysis**

**Data Exploration: -**

For Data Exploration stage of our analysis our team utilized multiple techniques to find outliers, multicollinearity, and correlation between variables. All of our techniques helped us make the data more symmetric and remove the -obvious outliers. Starting from 5-num summary and histogram for both Price, and Odometer.

**Price**

**Histogram and 5 Num Summary**

we used 5 Num Summary and Histogram to get a general idea for the distribution of data, as shown in Figure 1 and Figure 2. In Figure 1, we can see that the distribution of data is very skewed to the right, we can also see that it is due to couple datapoints at around 69000. When get a more details about those data points on Figure 2, which tells us that there are 3 values that are causing the skewness. Upon removing those values, we can visually see in Figure 1.1 the skewness has decreased, we can also see that the skewness is still caused by 1 data point with value around 58000. So, we decided to do another around of data cleaning. Upon removing that data point we can see that we have minimized the skewness by a lot. However, we decided that it can be improved by making doing a log transformation. Upon doing the log transformation, we can see in Figure 1.3, that the skewness is eliminated and now it looks a lot like a normal bell curve. Our findings are reflected in the 5 num summary, we can see that the spread between each 5% increment is tighter in Figure 2.1 than Figure 2. After doing the log transformation the spread is even tighter.

**Scatter Plots**

To visualize and analyze the association between price and 8 variables, our team used scatter plots (Figure 3). Upon examining we determined that the Price and Odometer has a negative linear association and Price, and Year has a positive linear association. We could not determine the association between other 6 variables, due to them being dummy variables. We decided upon doing a linear regression model for our data, since the two variables we could determine association for were linear.

**Pearson Correlation Coefficient**

After scatterplots, our team analyzed the Pearson correlation coefficient values using Proc corr procedure(Figure3 & 4). This procedure allowed us to determine the correlation between price and other 8 variables. Starting from price and odometer the correlation is -0.42203, which is indicates a weak negative correlation. The correlation value between price and transmission is -0.01637, which also indicated a weak negative correlation. The correlation value between price and drive is 0.1515, which indicates a weak positive correlation between the two. The correlation value between price and type is 0.2384 which indicates a weak positive correlation. The correlation value between price and fuel is 0.1063, which indicates a weak positive correlation. The correlation value between price and condition is -0.03773, which indicates a weak negative correlation. The correlation between price and year is 0.2932 which indicates a weak positive correlation. The correlation between price and cylinders is 0.1504, which also indicates a weak positive correlation. After analyzing the correlation for price and other 8 variables, our team checked for the multicollinearity between other variables, and all the variables appeared to be independent, so we moved on to cleaning up odometer variable.

**Odometer**

Like Price, our team decided to follow similar steps for odometer, starting from data visualization, followed by cleanup, association, correlation, and finally multicollinearity.

**Histogram and 5 Num Summary: -**

We used Histogram and 5 Num Summary to visualize the data and get an idea on the spread of the data. Starting from Histogram (Figure 6), we can see that the skewness is caused by couple variables located at around 2000000, and 1000000. Once we saw the skewness, we looked at the 5num summary of the data (Figure 7), which confirmed our assumption that it is caused due to just two variables. The 5 Num Summary told us that the min of the dataset is at 0, Q1 is at 71000, Median is at 110000, Q3 is at 149000, and the max is at 425000. As a team we agreed to remove the variables that are acting as influential points, and upon removing those two variables, we can see in (Figure 8), that the skewness is minimized, but not eliminated, therefore our team decided to do another round of removing outliers. Upon removing that one outlier we can see that we have eliminated the skewness and now the graph looks like a normal bell shape (Figure 10). The findings are reflected on the 5 Num summary (Figure 11) as, our max went from being at 425000 to 293000, while other values like Q3, Median, Q1, and mean stayed the same. After visualizing and removing influential points, our team decided to find the association using scatterplots.

**Scatter Plots**

Like Price we used scatterplots, to visualize the association between variables (Figure 12), our findings were as expected with Odometer and Price having a negative linear association, followed by Odometer and year also having negative linear association. We could not determine the association of the other 6 variables due to them being dummy variables. Using the information, we got from the scatterplot our team moved on to finding out about the correlation between each variable.

**Pearson Correlation Coefficient**

We used the proc corr procedure to get the Pearson Correlation Coefficient between odometer and other 8 variables. This procedure also allowed to quickly see if there is any multicollinearity between any variables. Our findings are as follows, odometer and price have a collinearity of -0.4268, which indicates a weak negative correlation. Odometer and transmission have correlation of -0.05875, which indicates a weak negative correlation. Odometer and drive have correlation of -0.0587, which also indicates a weak negative correlation. Odometer and drive have a correlation of 0.1009, which indicates a weak positive correlation. Odometer and type have a correlation of -0.00304, which indicates a weak negative correlation. Odometer and fuel have a correlation of -0.0553, which indicates a weak negative correlation. Odometer and condition have a correlation of 0.05505 which indicates a weak positive correlation. Odometer and year have a correlation of -0.2182 which indicates a weak negative correlation. Lastly odometer and cylinders have a correlation of 0.0896, which indicates a weak positive correlation.

After looking at the correlation numbers for each variable against odometer, our team decided to check for correlation numbers for the other variables against each other, to track down the multicollinearity. Upon a detailed examination of the correlation values, we did not find any strong correlation of any variable between each other.

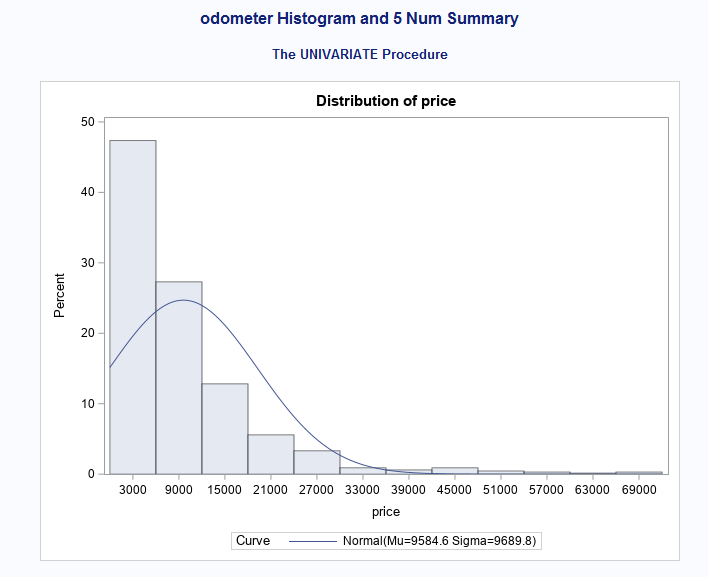


Figure 1

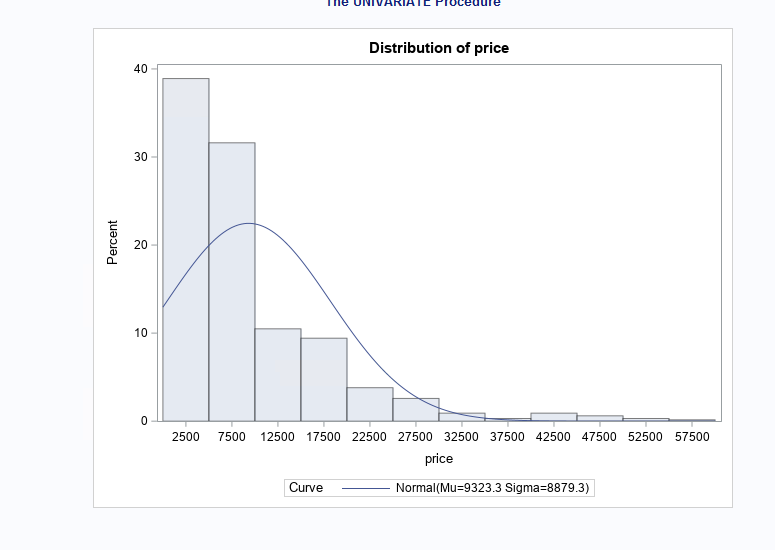


Figure 1.1

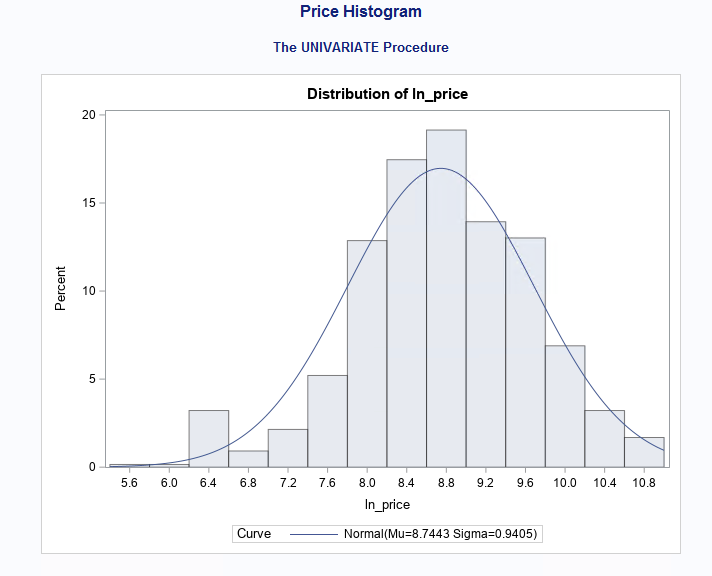


Figure 1.3

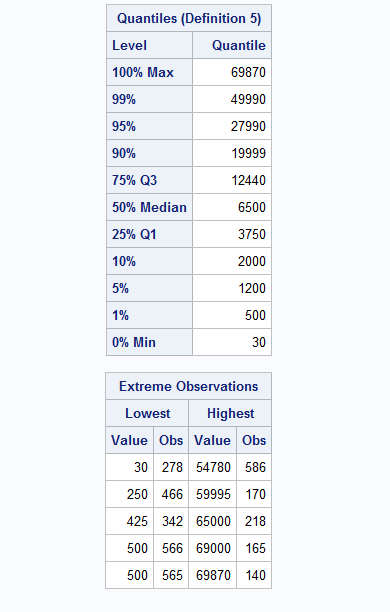


Figure 2

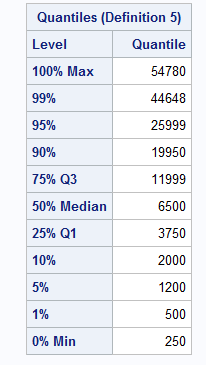


Figure 2.1

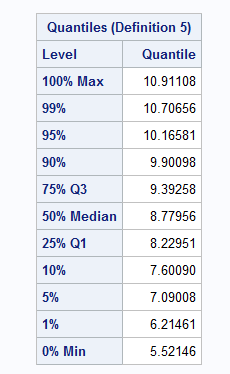


Figure 2.2

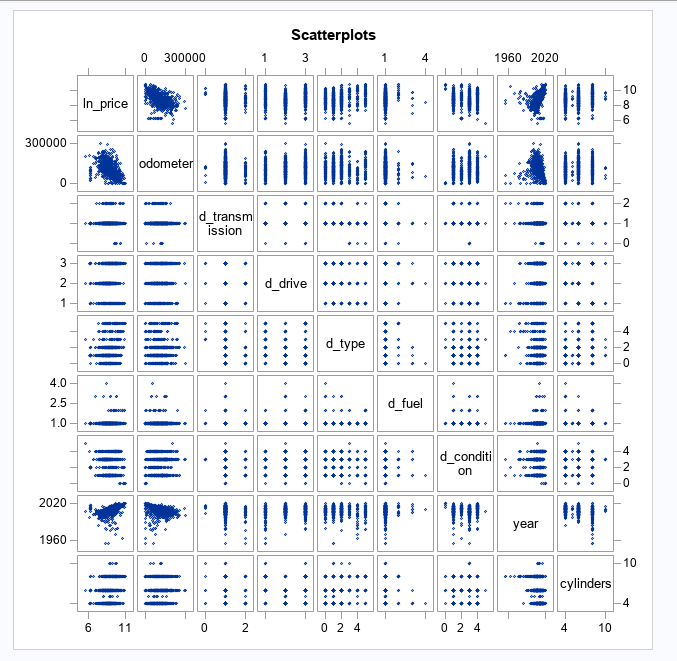


Figure 3

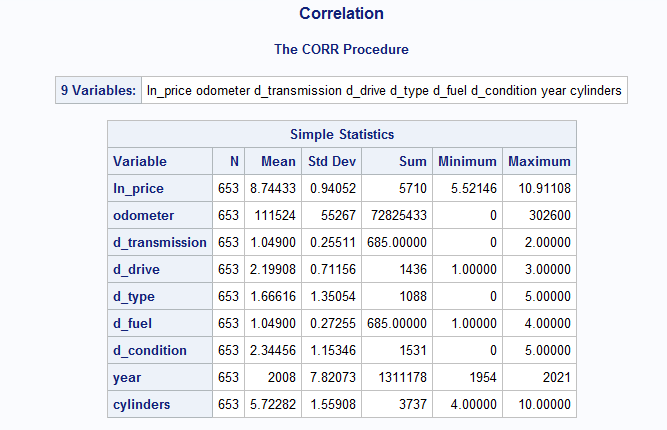


Figure 4

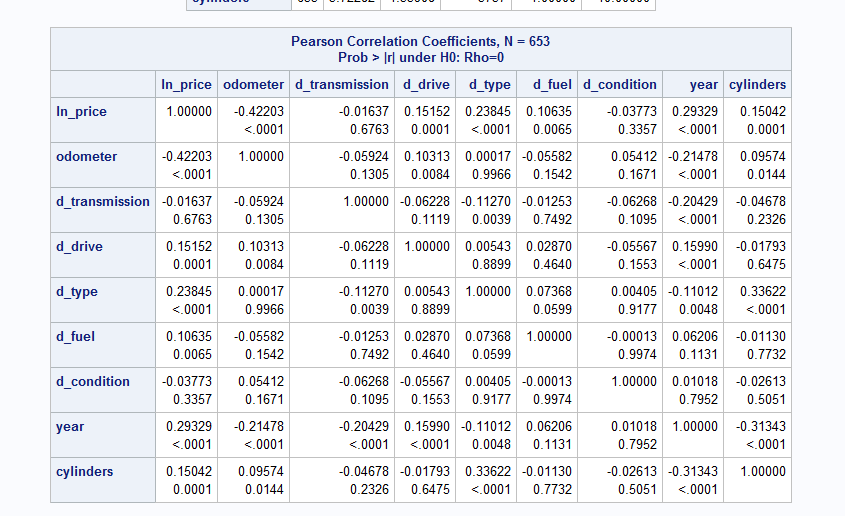


Figure 5

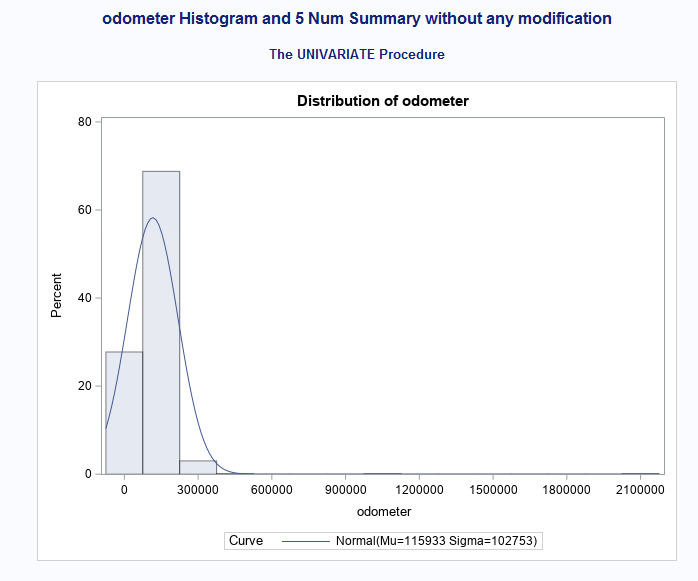


Figure 6



Figure 7

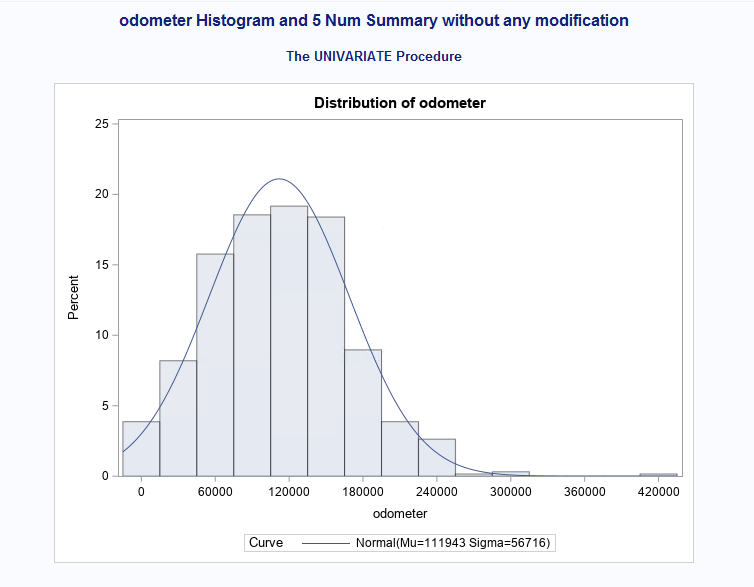


Figure 8



Figure 9

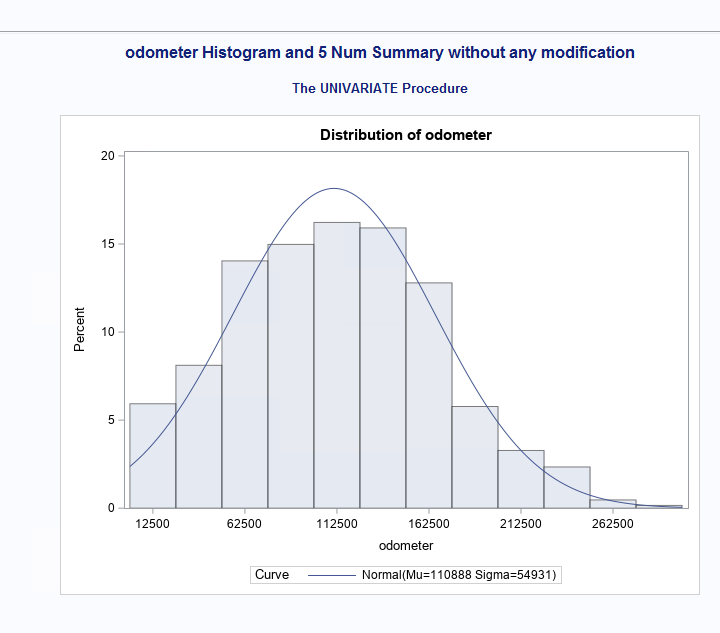


Figure 10

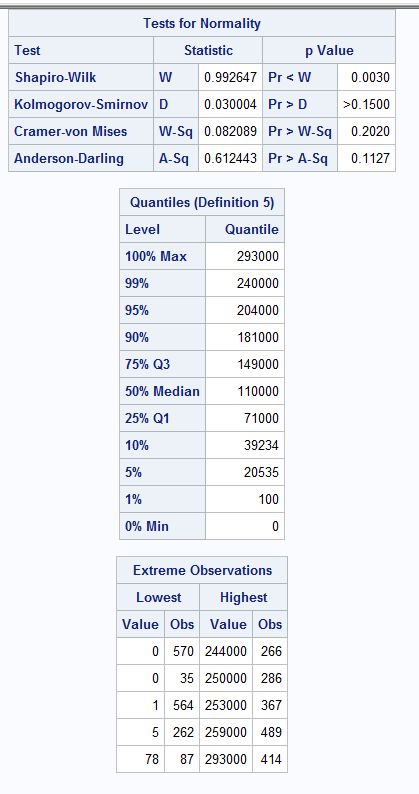
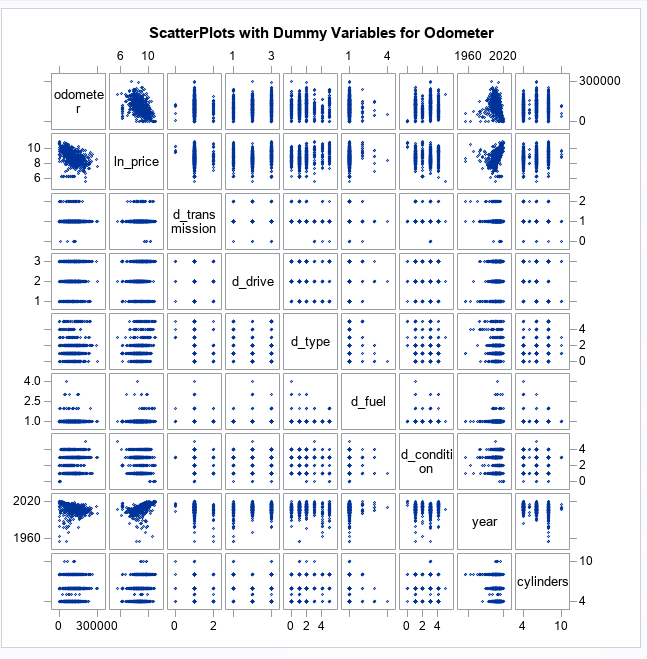


Figure 11

Figure 12