Data Preparation

The preparation of the automobile dataset involved a series of checks and transformations in order to turn the dataset into a suitable error-free format. Python package pandas was used to load the dataset and perform these checks and transformations. The first step was to load the dataset using pandas and manually check if the loaded dataset is identical to the contents to the file the dataset was loaded from. If identical, the data then structured into a human-readable table.

Physically Impossible Values

A function was created to determine any impossible values exists in the data. The function works by checking for values that fall outside its acceptable range for the corresponding feature. Values that fall outside the acceptable range are manually overwritten with a value closest to the acceptable range. This approach was justified due to being easy to perform and not losing information by omitting its value.

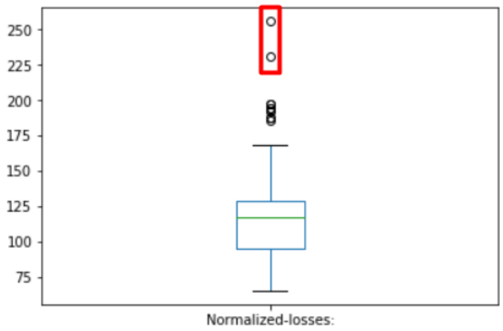
Missing Values

Handling missing values depends on the datatype of the feature. Missing values in features with numerical datatypes had their values replaced with the column mean. The justification for this method is because it eliminates guessing errors and its generally better to retain information rather than discarding it.

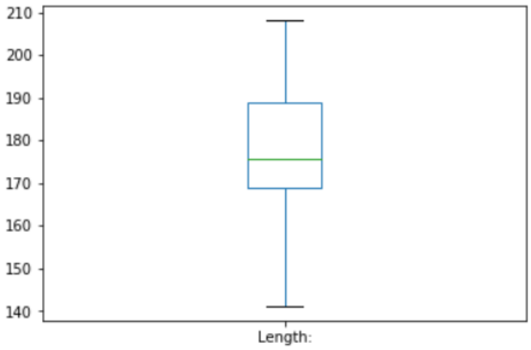
For cases involving strings such as the door-handle feature; these missing values are omitted because no other value is acceptable without requiring guessing errors or time-consuming investigation.

Outliers

Outliers are determined by plotting a boxplot for each feature that has a numerical datatype. In figure 1.0, outliers (highlighted in red) in normalized-losses are seen by the deviated circles relative to the others. A feature without any outliers can be seen in the length boxplot found in figure 1.1.



(Figure 1.0: Normalized-losses Feature Box Graph)



(Figure 1.1: Length Feature Box Graph)

Dealing with outliers is dependent on the nature of the outlier itself. Outliers require investigation and figuring out whether the outlier is a legitimate value or a human data entry error. Unfortunately, due to the complexing of outliers, the best course of action in this instance is to omit the value.

Typos

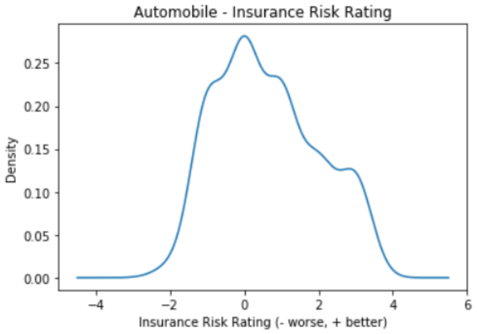
Typos were found using a function called value_counts(). The output of this function counts the total for each unique value. By scanning through the output of the function typos can easily be spotted. Each typo is manually validated within the context of the feature and fixed by overwriting the string with an appropriate value. The justification for this method allows for the information to be portrayed as intended rather than incorrectly.

Whitespaces & Inconsistencies

Whitespaces and other inconsistencies in the data source were handled using python functions. Whitespaces were removed using the strip() function and capitalization was lowercased using the lower() function. Like how typos were found, the function value_counts() serves the same purpose for finding inconsistencies and determining if they are legitimate. This approach required is easy to implement and has little or, in this case, no drawbacks.

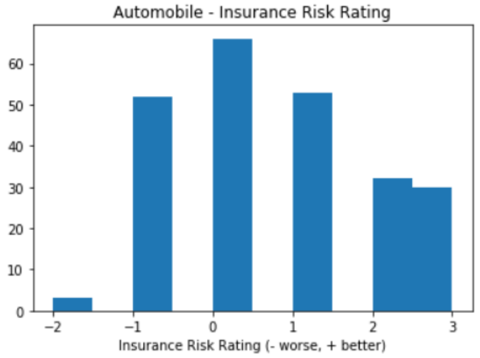
Data Exploration

Subsection 1.

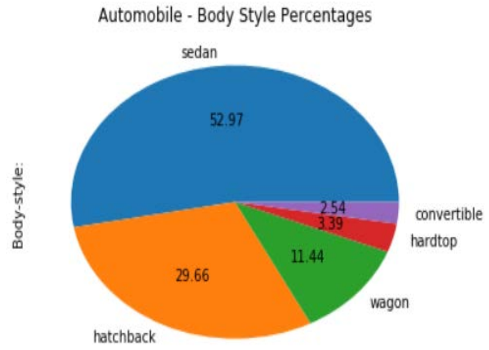


(Figure 2.0: Insurance Risk Rating Density Graph)

The insurance risk rating (symbolling) feature set was plotted using a density graph as seen in figure 2.0 above. A density graph was chosen for this feature set because of its smooth probability distribution over discrete values, and for determining where the largest concentration of values congregate. Because of the smooth continuous line, other graphs such as histograms detract from the purpose of visualizing data due to poor readability (unequal bin widths and vertical bar positioning) as seen in figure 2.1 below. As such, density is the ideal graph for visualizing the insurance risk rating.

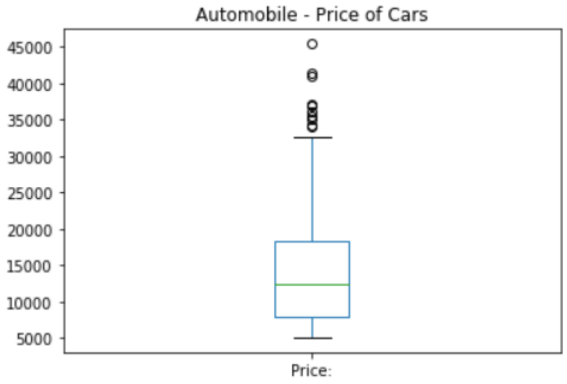


(Figure 2.1: Insurance Risk Rating Histogram Graph)



(Figure 2.2: Body Style Percentage Pie Graph)

A pie chart (shown in figure 2.2) was chosen for the body style feature set because of its categorical visualization properties. This graph is appropriate for this feature set because each category is divided into sectors within the circle. Because each sector in this graph represents a category of a body style, we can quickly determine what percentage of the out of the entire proportion a body style represents.

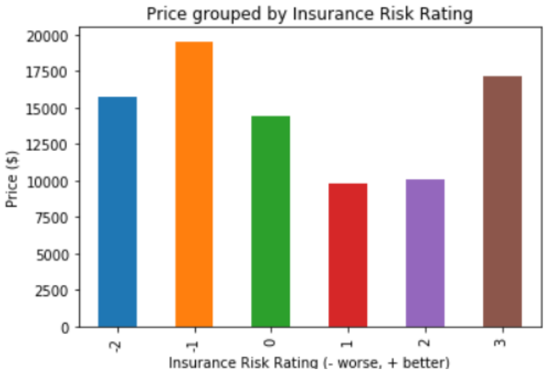


(Figure 2.3: Price Box-and-Whiskers Graph)

A box and whiskers graph is appropriate for the price feature set because it allows us to easily find the maximum and minimum value, quartiles, and the median price of automobiles. At a glance, we can observe how the price values are distribution. Other observations can include positive or negative skewness, outliers and extreme values.

Subsection 2.

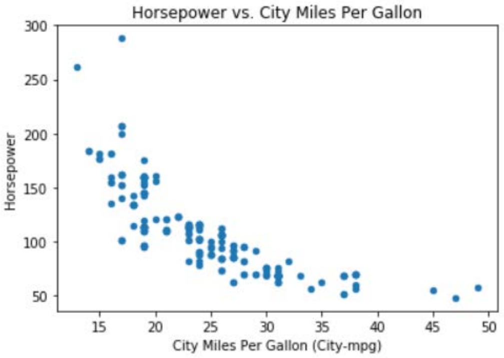
Hypothesis: - Higher prices negatively affects the insurance risk rating



(Figure 3.1: Price vs Insurance Risk Rating Bar Graph)

In figure 3.1, the average price is displayed for each insurance risk rating interval. At the worst interval (-2), the average price is approximately $15500 and at the positive end (3) of the insurance risk rating, the average price is approximately the same. By observing the price fluctuations at each interval, we can conclude that there no immediate relationship between these two features. Therefore, the hypothesis is rejected.

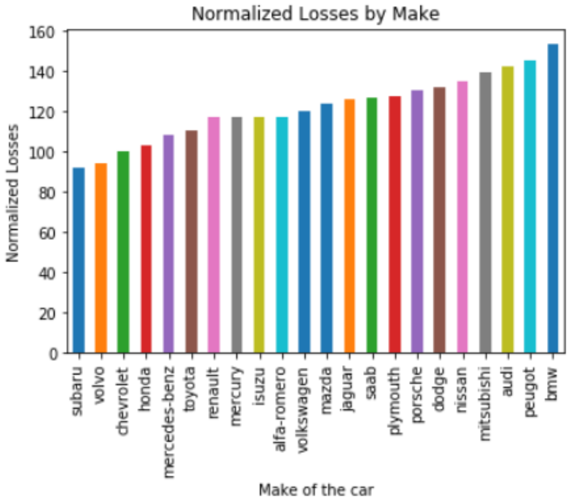
Hypothesis: - Higher horsepower decreases the city miles per gallon



(Figure 3.2: Horsepower vs City-mpg Scatterplot)

In figure 3.2 above, we can see a downhill trend between horsepower and city-mpg. As you observe from left to right, we can see horsepower (y-value) decreasing as city-mpg (x-value) increases. This observation indicates there is a negative relationship between the x and y values. Therefore, the hypothesis is accepted.

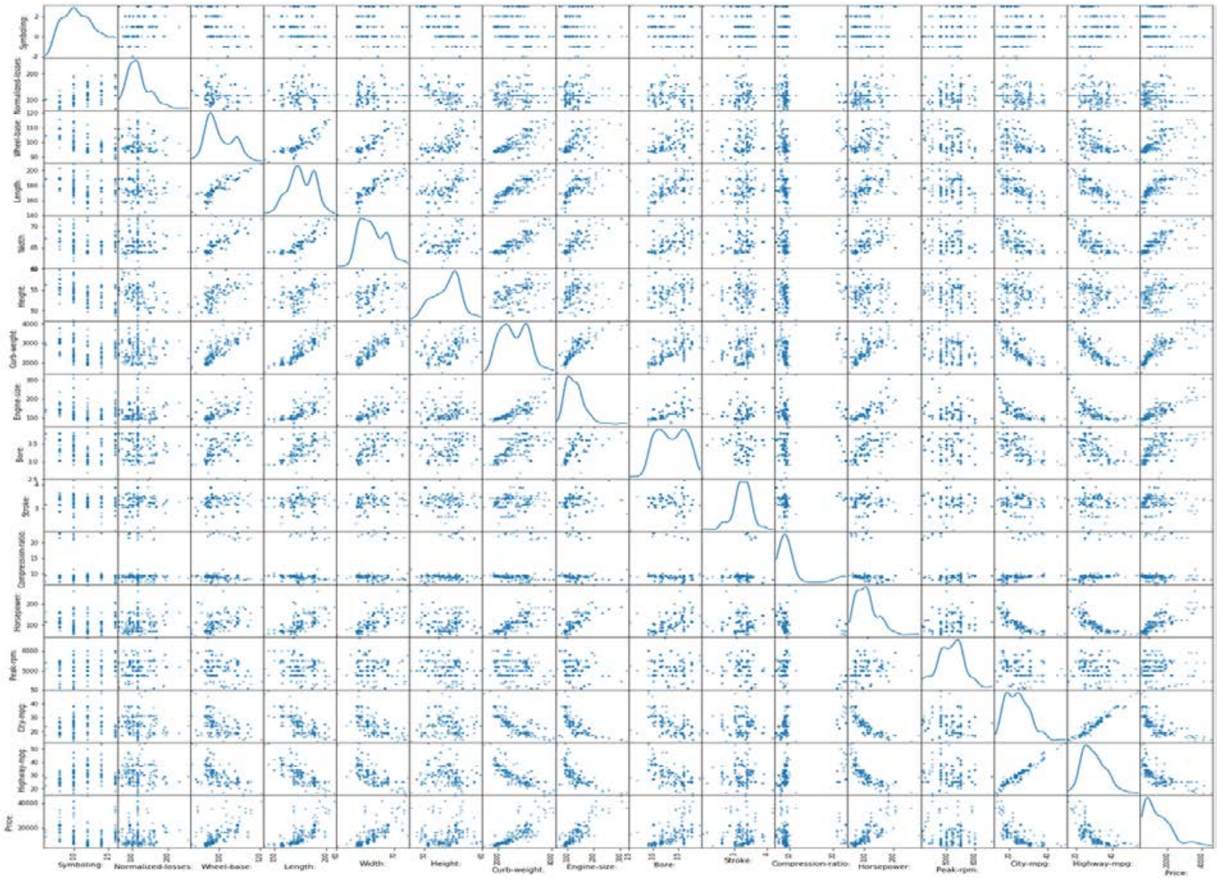
Hypothesis: - The make of an automobile affects normalized losses



(Figure 3.3: Normalized-losses vs Make Bar Graph)

In figure 3.3, the average normalized-loss for each make is plotted using a bar graph and sorted in ascending order by normalized-loss. From this graph, we can observe that the lowest normalized-loss is from Subaru and the highest loss is from BMW. We can also determine that the make of an automobile does affect the normalized-losses. Therefore, the hypothesis is accepted.

Subsection 3.



(Figure 4.0: Automobile Scatter Matrix)

Column Name	**Positive Correlation**	**Negative Correlation**
Symboling	No relations	No relations
Normalized-Losses	No relations	No relations
Wheel-Base	Length, width, curb-weight, price	No relations
Length	Wheel-base, width, curb-weight, engine-size	City-mpg, highway-mpg, price
Width	Wheel-base, length, curb-weight, engine-size, price	Highway-mpg
Height	No relations	No relations
Curb-Weight	Wheel-base, length, width, engine-size, price	City-mpg, highway-mpg
Engine-Size	Length, width, curb-weight, horsepower, price	City-mpg, highway-mpg
Bore	No relations	No relations
Stroke	No relations	No relations
Compression-Ratio	No relations	No relations
Horsepower	No relations	City-mpg, highway-mpg
Peak-RPM	No relations	No relations
City-MPG	Highway-mpg	Curb-weight, engine-size, horsepower, price
Highway-MPG	City-mpg	Length, width, curb-weight, engine-size, price
Price	Wheel-base, length, width, curb-weight, engine-size, horsepower,	City-mpg, highway-mpg