**Automobile Data Set**

**Exploratory Analysis**

Kathleen Zopff, [kzopff@bellarmine.edu](mailto:kzopff@bellarmine.edu)

1. **INTRODUCTION**

The data being explored in this analysis is the Automobile Data Set, found on the UCI Machine Learning Repository website. This data set has information about different types of cars and their specifications, as well as their price and their risk rating for insurance. This data set was chosen to explore due to its variety of variables as well as its applicable information.

1. **DATA SET DESCRIPTION**

This data set contains 205 samples with 26 columns with various data types. A complete listing of the variables and their data types is shown in **Table 1**.

**Table 1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| Symboling | Ordinal/int64 | 0% |
| Normalized Losses | Ratio/object | 20% |
| Make | Nominal/object | 0% |
| Fuel Type | Nominal/object | 0% |
| Aspiration | Nominal/object | 0% |
| Number of Doors | Nominal/object | 0.98% |
| Body Style | Nominal/object | 0% |
| Drive Wheels | Nominal/object | 0% |
| Engine Location | Nominal/object | 0% |
| Wheel Base | Ratio/float64 | 0% |
| Length | Ratio/float64 | 0% |
| Width | Ratio/float64 | 0% |
| Height | Ratio/float64 | 0% |
| Curb Weight | Ratio/int64 | 0% |
| Engine Type | Nominal/object | 0% |
| Number of Cylinders | Nominal/object | 0% |
| Engine Size | Ratio/int64 | 0% |
| Fuel System | Nominal/object | 0% |
| Bore | Ratio/object | 1.95% |
| Stroke | Ratio/object | 1.95% |
| Compression Ratio | Ratio/float64 | 0% |
| Horsepower | Ordinal/object | 0.98% |
| Peak RPM | Ratio/object | 0.98% |
| City MPG | Ratio/int64 | 0% |
| Highway MPG | Ratio/int64 | 0% |
| Price | Ratio/object | 1.95% |

1. **Data Set Summary Statistics**

In this section, the data set is explored by finding the summary statistics (Table 2), the proportions of each value of categorical data (Tables 3-12), and the correlations between the quantitative data (Table 13). A heatmap, shown in Figure 1, is also used to better visualize the correlations between the data in the set.

**Table 2: Summary Statistics for Automobile Data Set**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Variable Name* | *Count* | *Mean* | *Standard Deviation* | *Min* | *25th* | *50th* | *75th* | *Max* |
| Symboling | 205 | 0.8341 | 1.2453 | -2 | 0 | 1 | 2 | 3 |
| Wheel Base | 205 | 98.7566 | 6.0218 | 86.6 | 94.5 | 97 | 102.4 | 102.9 |
| Length | 205 | 174.0493 | 12.3373 | 141.1 | 166.3 | 173.2 | 183.1 | 208.1 |
| Width | 205 | 65.9078 | 2.1452 | 60.3 | 64.1 | 65.5 | 66.9 | 72.3 |
| Height | 205 | 53.7249 | 2.4435 | 47.8 | 52 | 54.1 | 55.5 | 59.8 |
| Curb Weight | 205 | 2555.5659 | 520.6802 | 1488 | 2145 | 2414 | 2935 | 4066 |
| Engine Size | 205 | 126.9073 | 41.6427 | 61 | 97 | 120 | 141 | 326 |
| Compression Ratio | 205 | 10.1425 | 3.9720 | 7 | 8.6 | 9 | 9.4 | 23 |
| City MPG | 205 | 25.2195 | 6.5421 | 13 | 19 | 24 | 30 | 49 |
| Highway MPG | 205 | 30.7512 | 6.8864 | 16 | 25 | 30 | 34 | 54 |

Table 3: Proportions for Make (n=3)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Toyota | 32 | 15.61% |
| Nissan | 18 | 8.78% |
| Mazda | 17 | 8.29% |
| Honda | 13 | 6.34% |
| Mistsubishi | 13 | 6.34% |
| Subaru | 12 | 5.85% |
| Volkswagen | 12 | 5.85% |
| Peugot | 11 | 5.37% |
| Volvo | 11 | 5.37% |
| Dodge | 9 | 4.39% |
| BMW | 8 | 3.9% |
| Mercedes-Benz | 8 | 3.9% |
| Plymouth | 7 | 3.41% |
| Audi | 7 | 3.41% |
| Saab | 6 | 2.93% |
| Porsche | 5 | 2.44% |
| Isuzu | 4 | 1.95% |
| Jaguar | 3 | 1.46% |
| Chevrolet | 3 | 1.46% |
| Alfa-Romero | 3 | 1.46% |
| Renault | 2 | 0.98% |
| Mercury | 1 | 0.49% |

Table 4: Proportions for Fuel Type (n=4)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Gas | 185 | 90.24% |
| Diesel | 20 | 9.76% |

Table 5: Proportions for Aspiration (n=5)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| STD | 168 | 81.95% |
| Turbo | 37 | 18.05% |

Table 6: Proportions for Number of Doors (n=6)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Four | 114 | 55.61% |
| Two | 89 | 43.41% |
| (Missing Data) | 2 | 0.98% |

Table 7: Proportions for Body Style (n=7)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Sedan | 96 | 46.83% |
| Hatchback | 70 | 34.15% |
| Wagon | 25 | 12.2% |
| Hardtop | 8 | 3.9% |
| Convertible | 6 | 2.93% |

Table 8: Proportions for Drive Wheels (n=8)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| FWD | 120 | 58.54% |
| RWD | 76 | 37.07% |
| 4WD | 9 | 4.39% |

Table 9: Proportions for Engine Location (n=9)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Front | 202 | 98.54% |
| Rear | 3 | 1.46% |

Table 10: Proportions for Engine Type (n=15)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| OHC | 148 | 72.2% |
| OHCF | 15 | 7.32% |
| OHCV | 13 | 6.34% |
| DOHC | 12 | 5.85% |
| L | 12 | 5.85% |
| Rotor | 4 | 1.95% |
| DOHCV | 1 | 0.49% |

Table 11: Proportions for Number of Cylinders (n=16)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Four | 159 | 77.56% |
| Six | 24 | 11.7% |
| Five | 11 | 5.37% |
| Eight | 5 | 2.44% |
| Two | 4 | 1.95% |
| Twelve | 1 | 0.49% |
| Three | 1 | 0.49% |

Table 12: Proportions for Fuel System (n=18)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| MPFI | 94 | 45.85% |
| 2BBL | 66 | 32.2% |
| IDI | 20 | 9.76% |
| 1BBL | 11 | 5.37% |
| SPDI | 9 | 4.39% |
| 4BBL | 3 | 1.46% |
| MFI | 1 | 0.49% |
| SPFI | 1 | 0.49% |

Table 13: Correlation Table/Tables

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Symboling* | *Wheel Base* | *Length* | *Width* | *Height* | *Curb Weight* | *Engine Size* | *Compression Ratio* | *City MPG* | *Highway MPG* |
| *Symboling* | 1 | -0.53 | -0.36 | -0.23 | -0.54 | -0.23 | -0.11 | -0.18 | -0.04 | 0.03 |
| *Wheel Base* | -0.53 | 1 | 0.87 | 0.8 | 0.59 | 0.78 | 0.57 | 0.25 | -0.47 | -0.54 |
| *Length* | -0.36 | 0.87 | 1 | 0.84 | 0.49 | 0.88 | 0.68 | 0.16 | -0.67 | -0.7 |
| *Width* | -0.23 | 0.8 | 0.84 | 1 | 0.28 | 0.87 | 0.74 | 0.18 | -0.64 | -0.68 |
| *Height* | -0.54 | 0.59 | 0.49 | 0.28 | 1 | 0.3 | 0.07 | 0.26 | -0.05 | -0.11 |
| *Curb Weight* | -0.23 | 0.78 | 0.88 | 0.87 | 0.3 | 1 | 0.85 | 0.15 | -0.76 | -0.8 |
| *Engine Size* | -0.11 | 0.57 | 0.68 | 0.74 | 0.07 | 0.85 | 1 | 0.03 | -0.65 | -0.68 |
| *Compression Ratio* | -0.18 | 0.25 | 0.16 | 0.18 | 0.26 | 0.15 | 0.03 | 1 | 0.32 | 0.27 |
| *City MPG* | -0.04 | -0.47 | -0.67 | -0.64 | -0.05 | -0.76 | -0.65 | 0.32 | 1 | 0.97 |
| *Highway MPG* | 0.03 | -0.54 | -0.7 | -0.68 | -0.11 | -0.8 | -0.68 | 0.27 | 0.97 | 1 |

A picture containing table

Description automatically generated

**Figure 1: Correlation Heatmap from Automobile Data Set**

1. **DATA SET GRAPHICAL EXPLORATION**

Here, the data is explored visually through the use of different types of graphs. The distribution of the quantitative data is shown using histograms, as shown in **Figure 2**. To compare all of the quantitative data, a pair plot of scatter plots was used, as shown in **Figure 3.** The frequency of all values for each categorical variable is displayed using horizontal bar charts, as shown in **Figure 4.** To further explore the relations between the quantitative data, an example of a regression model of two of the data set’s variables is shown in **Figure 5.**

*Chart, waterfall chart

Description automatically generated*

**Figure 2: Collection of Histograms for all quantitative variables in the Automobile Data Set**

Chart

Description automatically generated

**Figure 3: Pairwise scatter plots of the numerical data in the Automobile Data Set organized by their Body Style.**

Graphical user interface, application

Description automatically generated

**Figure 4: Collection of horizontal bar charts for the categorical data in the Automobile Data Set**

*Chart, scatter chart

Description automatically generated*

**Figure 5: Comparison of City MPG to Engine Size in a scatter plot with a regression line**

1. **SUMMARY OF FINDINGS**

Much of the data in this data set is as expected. The relationships between dimensions of the cars have significantly high correlation compared to other variables in the data set, which makes sense as the individual dimensions of the car are somewhat dependent on each other as cars are of a usual proportion. Some interesting relationships to note are the relationships between the City and Highway MPGs with the dimensional variables of the cars’ sizes. These relationships are consistently negative—some stronger than others. An example of one of these relationships is shown in **Figure 5.** The negative nature of their relationship suggests that as cars get smaller, their gas mileage improves. This is an important takeaway from the data as it gives information on the efficiency of different cars which is a great concern when looking to buy one.