**Automobile Data Set**

**Exploratory Analysis**

Brandon Thomas, [bthomas@bellarmine.edu](mailto:bthomas@bellarmine.edu)

Richard Osborn, [rosborn@bellarmine.edu](mailto:rosborn@bellarmine.edu)

1. **INTRODUCTION**

This data set is a list of cars from 1985 Ward’s Automotive Yearbook that contain a variety of variables pertaining the certain types of cars. This can be found at <https://archive.ics.uci.edu/ml/datasets/Automobile>. This data set was chosen for its mix of continuous and categorical variables. Also, cars provide various methods of comparison that make for a good dataset to perform an exploratory analysis on.

1. **DATA SET DESCRIPTION**

This data set contains 205 samples with 26 columns with various types of data. There are 15 continuous variables and 11 discrete variables. Rows with missing data were removed for simplicity of data analysis but are included in the complete listing shown in Table 1.

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

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| *Symboling* | *interval/float64* | *0%* |
| *Normalized Losses* | *ratio/object* | *20%* |
| *Make* | *nominal/object* | *0%* |
| *Fuel Type* | *nominal/object* | *0%* |
| *Aspiration* | *nominal/object* | *0%* |
| *Number of Doors* | *nominal/object* | *1%* |
| *Body Style* | *nominal/object* | *0%* |
| *Drive Wheels* | *nominal/object* | *0%* |
| *Engine Location* | *nominal/object* | *0%* |
| *Wheelbase* | *ratio/float64* | *0%* |
| *Length* | *ratio/float64* | *0%* |
| *Width* | *ratio/float64* | *0%* |
| *Height* | *ratio/float64* | *0%* |
| *Curb Weight* | *ratio/object* | *0%* |
| *Engine Type* | *nominal/object* | *0%* |
| *Number of Cylinders* | *nominal/object* | *0%* |
| *Engine Size* | *ratio/float64* | *0%* |
| *Fuel System* | *nominal/object* | *0%* |
| *Bore* | *ratio/object* | *2%* |
| *Stroke* | *ratio/object* | *2%* |
| *Compression Ratio* | *ratio/float64* | *0%* |
| *Horsepower* | *ratio/object* | *1%* |
| *Peak Rpm* | *ratio/object* | *1%* |
| *City Mpg* | *ratio/float64* | *0%* |
| *Highway Mpg* | *ratio/float64* | *0%* |
| *Price* | *ratio/object* | *2%* |

**Explanation of Variables**

Symboling - Cars are initially assigned a risk factor symbol associated with its price. Then, if it is riskier (or less), this symbol is adjusted by moving it up (or down) the scale. A value of +3 indicates that the auto is risky, -3 that it is probably safe.

Normalized Losses – The relative average loss payment per insured vehicle year. This value is then normalized for vehicles in certain size classifications to get the normalized loss value of insurance payments.

Make – The manufacturer of the vehicle.

Fuel Type – The kind of fuel required for the vehicle to run properly.

Aspiration – How air is taken into the engine to be used for combustion.

Number of Doors – Amount of doors the vehicle has.

Body Style – Category given to a vehicle based on the shape and number of doors it has.

Drive Wheels – The number and which wheels power the vehicle to move.

Engine Location – Shows where the engine is located in reference to the front or back of the vehicle.

Wheelbase – The distance between the axles of the front and back wheels of a vehicle.

Length – The total length of the car.

Width – The width of the car.

Height – The height of the car.

Curb Weight – The weight of a car with no occupants or baggage.

Engine Type – The type of engine, often classified by the number of cylinders or displacement mechanism.

Number of Cylinders – The number of cylinders in the engine where combustion takes place.

Engine Size – Refers to the space an engine’s pistons operate in, often also referred to as engine displacement.

Fuel System – Contains the fuel tank, filter, pump, injectors/carburetor and named for the different parts that make it up.

Bore – The diameter of the cylinder in each car’s engine.

Stroke – The distance within the cylinder the piston travels.

Compression Ratio – The ratio of volume in a cylinder when a piston is at the beginning of a stroke compared to when a piston is at the end of its stroke.

Horsepower – The power an engine produces, specifically the power needed to move 550 pounds one foot per second.

Peak RPM – The maximum revolutions per minute of the crankshaft of a car’s engine.

City MPG – The average miles per gallon of a gas for a car in the city.

Highway MPG – The average miles per gallon of gas for a car on the highway.

Price – The market price the vehicle was purchased for on day of purchase.

1. **Data Set Summary Statistics**

These tables contains statistical information pertaining to different attributes of the car. This includes the count, mean, standard deviation, min and maxis, and each quarter percentile, respectively.

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Symbolling** | **Normalized\_losses** | **Wheel\_base** | **Length** | **Width** | **Height** | **Curb\_weight** | **Engine\_size** | **Bore** | **Stroke** | **Compression\_ratio** | **Horsepower** | **Peak\_rpm** | **City\_mpg** | **Highway\_mpg** | **Price** |
| **Count** | **159** | **159** | **159** | **159** | **159** | **159** | **159** | **159** | **159** | **159** | **159** | **159** | **159** | **159** | **159** | **159** |
| **Mean** | **0.7358** | **121.132** | **98.2641** | **172.4128** | **65.707** | **53.8993** | **2461.1345** | **119.2264** | **3.3** | **3.2365** | **10.1611** | **95.83** | **5113.8364** | **26.522** | **32.0817** | **11445.729** |
| **Std** | **1.193** | **35.6512** | **5.1674** | **11.5231** | **1.947** | **2.2687** | **481.9412** | **30.4607** | **0.2673** | **0.2948** | **3.889** | **30.7185** | **465.75** | **6.097** | **6.4591** | **5877.8561** |
| **Min** | **-2** | **65** | **86.6** | **141.1** | **60.3** | **49.4** | **1488** | **61** | **2.54** | **2.07** | **7** | **48** | **4150** | **15** | **18** | **5118** |
| **25%** | **0** | **94** | **94.5** | **165.65** | **64** | **52.25** | **2065.5** | **97** | **3.05** | **3.105** | **8.7** | **69** | **4800** | **23** | **28** | **7372** |
| **50%** | **1** | **113** | **96.9** | **172.4** | **65.4** | **54.1** | **2340** | **110** | **3.27** | **3.27** | **9** | **88** | **5200** | **26** | **32** | **9233** |
| **75%** | **2** | **148** | **100.8** | **117.8** | **66.5** | **55.5** | **2809.5** | **135** | **3.56** | **3.41** | **9.4** | **114** | **5500** | **31** | **37** | **14719.5** |
| **max** | **3** | **256** | **115.600** | **202** | **71.7** | **59.8** | **4066** | **258** | **3.94** | **4.17** | **23** | **200** | **6600** | **49** | **54** | **35056** |

A picture containing graphical user interface

Description automatically generatedIt is interesting that the average horsepower is only 95.83HP. You would expect this to be higher since most cars in the data set seem to be commuter cars that have would typically have between 150-200 horsepower. This could imply there are outliers that are bringing the horsepower average down. With a minimum of 48 and a maximum of 200, you would expect the mean to be closer to 200 than 48, so a median may be a more accurate method of comparison.

Table

Description automatically generated

When comparing city\_mpg and highway\_mpg, there is a noticeable difference in the means, since highway driving is done with minimal stopping resulting in better gas mileage. Highway miles per gallon is higher than expected but also has a higher standard deviation. Based on the empirical rule, 95% of the data should fall between 19 and 45 highway\_mpg. This shows that there is a lot of variability in highway mpg. This leaves more uncertainty on how to truly reflect the mean when the interval containing most of the data (95%) has such a large spread. A better measure of central tendency may be to use the median.

Table 3: Proportions for Automobile Data Set (Shown as Barcharts in Figure 7)

|  |  |  |
| --- | --- | --- |
| *Symboling* | *Frequency* | *Proportion (%)* |
| *-3* | *0* | *0%* |
| *-2* | *3* | *1.9%* |
| *-1* | *20* | *12.58%* |
| *0* | *48* | *30.19%* |
| *1* | *46* | *28.93%* |
| *2* | *29* | *18.24%* |
| *3* | *13* | *8.18%* |

With most of the frequency being at 0 or 1, this shows that most cars taken in this sample were considered risky or somewhat risky.

|  |  |  |
| --- | --- | --- |
| *Make* | *Frequency* | *Proportion(%)* |
| *Toyota* | *31* | *19.50%* |
| *Nissan* | *18* | *11.32%* |
| *Honda* | *13* | *8.18%* |
| *Subaru* | *12* | *7.55%* |
| *Mazda* | *11* | *6.92%* |
| *Volvo* | *11* | *6.92%* |
| *Mitsubishi* | *10* | *6.29%* |
| *Dodge* | *8* | *5.03%* |
| *Volkswagen* | *8* | *5.03%* |
| *Peugot* | *7* | *4.40%* |
| *Plymouth* | *6* | *3.77%* |
| *Saab* | *6* | *3.77%* |
| *Mercedes-Benz* | *5* | *3.14%* |
| *BMW* | *4* | *2.52%* |
| *Audi* | *4* | *2.52%* |
| *Chevrolet* | *3* | *1.89%* |
| *Jaguar* | *1* | *0.63%* |
| *Porsche* | *1* | *0.63* |

Overall, the highest proportion of cars in the sample were Toyota’s and Nissan’s. This would make sense since they are typically cheaper cars affordable to the general population when compared to a BMW or Jaguar.

|  |  |  |
| --- | --- | --- |
| *Fuel Type* | *Frequency* | *Proportion(%)* |
| *Gas* | *144* | *90.57%* |
| *Diesel* | *15* | *9.43%* |

|  |  |  |
| --- | --- | --- |
| *Aspiration* | *Frequency* | *Proportion(%)* |
| *Standard* | *132* | *83.02%* |
| *Turbo* | *27* | *16.98%* |

|  |  |  |
| --- | --- | --- |
| *Number of Doors* | *Frequency* | *Proportion(%)* |
| *Four* | *95* | *59.75%* |
| *Two* | *64* | *40.25%* |

The majority of cars are gas vehicles, which is the more widely used fuel type.

The majority of cars also have standard aspiration, with turbo usually being more present in faster, more expensive, higher-end cars.

Four doors were more frequent in the sample but there were more two door cars than expected.

|  |  |  |
| --- | --- | --- |
| *Body Style* | *Frequency* | *Proportion(%)* |
| *Sedan* | *79* | *49.69%* |
| *Hatchback* | *56* | *35.22%* |
| *Wagon* | *17* | *10.69%* |
| *Hardtop* | *5* | *3.14%* |
| *Convertible* | *2* | *1.26%* |

Sedans and hatchbacks are the most popular car styles, with them making up a large proportion of the data set. Conversely, hardtops and convertibles make up a very small proportion as they are much less common.

|  |  |  |
| --- | --- | --- |
| *Drive Wheels* | *Frequency* | *Proportion(%)* |
| *FWD* | *105* | *66.04%* |
| *RWD* | *46* | *28.93%* |
| *4WD* | *8* | *5.03%* |

Over 90% of the cars had two-wheel drive (front and rear wheel drive combined), which can be expected as many vehicle owners do not have a need for 4 wheel drive in their daily activities.

|  |  |  |
| --- | --- | --- |
| *Engine Location* | *Frequency* | *Proportion(%)* |
| *Front* | *159* | *100%* |

All cars had a front-engine location, with engines in the back still being uncommon today except in cars designed for maximal speed and horsepower.

|  |  |  |
| --- | --- | --- |
| *Engine Type* | *Frequency* | *Proportion(%)* |
| *(OHC)* | *123* | *77.36%* |
| *(OHFC)* | *12* | *7.55%* |
| *(DOHC)* | *8* | *5.03%* |
| *(L)* | *8* | *5.03%* |
| *(OHCV)* | *8* | *5.03%* |

The overhead camshaft had over 77% of the proportion, the standard engine style. The other types are much more specialized and are not as customary.

|  |  |  |
| --- | --- | --- |
| *Number of Cylinders* | *Frequency* | *Proportion(%)* |
| *Four* | *136* | *85.54%* |
| *six* | *14* | *8.81%* |
| *five* | *7* | *4.4%* |
| *eight* | *1* | *0.6%* |
| *three* | *1* | *0.6%* |

Four-cylinder and six-cylinder cars are the most common in the data set, because of their creation of a balanced engine and their frequency of use in commuter cars. Volkswagen has historically used five-cylinder engines, but three-cylinder and eight cylinders are uncommon.

|  |  |  |
| --- | --- | --- |
| *Fuel System* | *Frequency* | *Proportion(%)* |
| *(MPFI)* | *64* | *40.25%* |
| *(2BBL)* | *63* | *39.62%* |
| *(IDI)* | *15* | *10.79%* |
| *(1BBL)* | *11* | *6.92%* |
| *(SPDI)* | *5* | *3.14%* |
| *(MFI)* | *1* | *0.6%* |

The multi-point fuel injection and the two-barrel carburetor have the highest proportion, with the latter typically being used on smaller cars.

Table 4: Correlation Table/Tables for Automobile Data Set

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Normalized\_losses | Wheel\_base | Length | Width | Height | Curb\_weight | Engine\_size | Bore | Stroke | Compression\_ratio | Horsepower | Peak\_rpm | City\_mpg | Highway\_mpg | price |
| Normalized\_losses | 1 | -0.060086 | 0.035541 | 0.109726 | -0.413702 | 0.125858 | 0.207820 | -0.031558 | 0.063330 | -0.127259 | 0.290511 | 0.237697 | -0.235523 | -0.188564 | 0.202761 |
| Wheel\_base | -0.060086 | 1 | 0.871534 | 0.814991 | 0.555767 | 0.810181 | 0.649206 | 0.578159 | 0.167449 | 0.291431 | 0.516948 | -0.289234 | -0.580657 | -0.611750 | 0.734419 |
| Length | 0.035541 | 0.871534 | 1 | 0.838338 | 0.499251 | 0.871291 | 0.725953 | 0.646318 | 0.121073 | 0.184814 | 0.672063 | -0.234074 | -0.724544 | -0.724599 | 0.760952 |
| Width | 0.109726 | 0.814991 | 0.838338 | 1 | 0.292706 | 0.870595 | 0.779253 | 0.572254 | 0.196619 | 0.258752 | 0.681872 | -0.232216 | -0.666684 | -0.693339 | 0.843371 |
| Height | -0.413702 | 0555767 | 0.499251 | 0.292706 | 1 | 0.367052 | 0.111083 | 0.254836 | -0.091313 | 0.233308 | 0.034317 | -0.245864 | -0.199737 | -0.226136 | 0.244836 |
| Curb\_weight | 0.125858 | 0.810181 | 0.871291 | 0.870595 | 0.367052 | 1 | 0.888626 | 0.645792 | 0.173844 | 0.224724 | 0.790095 | -0.259988 | -0.762155 | -0.789338 | 0.893639 |
| Engine\_size | 0.207820 | 0.649206 | 0.725953 | 0.779253 | 0.111083 | 0.888626 | 1 | 0.595737 | 0.299683 | 0.141097 | 0.812073 | -0.284686 | -0.699139 | -0.714095 | 0.841496 |
| Bore | -0.031558 | 0.578159 | 0.646318 | 0.572554 | 0.254836 | 0.645792 | 0.595737 | 1 | -0.102581 | 0.015119 | 0.560239 | -0.312269 | -0.590440 | -0.590850 | 0.533890 |
| Stroke | 0.063330 | 0.167449 | 0.121073 | 0.196619 | -0.091313 | 0.173844 | 0.299683 | -0.102581 | 1 | 0.243587 | 0.148804 | -0.011312 | -0.020055 | -0.012934 | 0.160664 |
| Compression\_ratioo | -0.127259 | 0.291431 | 0.184814 | 0.258752 | 0.233308 | 0.224724 | 0.141097 | 0.015119 | 0.243587 | 1 | -0.162305 | -0.416769 | 0.278332 | 0.221483 | 0.209361 |
| Horsepower | 0.290511 | 0.516948 | 0.672063 | 0.681872 | 0.034317 | 0.790095 | 0.812073 | 0.560239 | 0.148804 | -0.162305 | 1 | 0.074057 | -0.837214 | -0.827941 | 0.759874 |
| Peak\_rpm | 0.237697 | -0.289234 | -0.234074 | -0.232216 | -0.245864 | -0.259988 | -0.284686 | -0.312269 | -0.011312 | -0.416769 | 0.074057 | 1 | -0.052929 | -0.032777 | -0.171916 |
| City\_mpg | -0.235523 | -0.580657 | -0.724544 | -0.666684 | -0.199737 | -0.762155 | -0.699139 | -0.590440 | -0.020055 | 0.278332 | -0.837214 | -0.052929 | 1 | 0.971999 | -0.692273 |
| Highway\_mpg | -0.188564 | -0.611750 | -0.724599 | -0.693339 | -0.226136 | -0.789338 | -0.714095 | -0.5980850 | -0.012934 | 0.221483 | -0.827941 | -0.032777 | 0.971999 | 1 | -0.720090 |
| Price | 0.202761 | 0.734419 | 0.760952 | 0843371 | 0.244836 | 0.893639 | 0.841496 | 0.533890 | 0.160664 | 0.209361 | 0.759874 | -0.171916 | -0.692273 | -0.720090 | 1 |

Table

Description automatically generatedWhen looking at the correlation table, much of the data has a positive correlation. For example, the length and width have a large, positive correlation that is intuitive. Engine size and horsepower also have a large, positive correlation as a larger engine often results in more horsepower.

Table

Description automatically generated

**Figure 5 – Heatmap of Automobile Dataset Correlations**

Chart

Description automatically generated

1. **DATA SET GRAPHICAL EXPLORATION**

**Figure 6 - Pairwise Plot/ Scatter Plots of Automobile Data Set’s Continuous Variables**

Chart, histogram

Description automatically generatedOne of the most interesting distributions is the compression ratio. Most of the data is under 10 while there are a few outliers that are over 20 which affect the graph. Interestingly, the stroke, bore engine size, height, width, and length all have a line that follows the shape of a normal distribution. The normalized losses and wheelbase tend to look more like a chi square distribution however, potentially showing the impact of larger sample size.

Chart, histogram

Description automatically generated

These graphs are normally distributed with few significant outliers. Besides price, most of the data has little to no skew. There is a higher frequency of cars that are under $10,000, which makes sense given the relatively low mean car price.

Chart, scatter chart

Description automatically generatedThroughout all of the data, there is a clear distinction between the different values of the compression ratio. It is either under 10 or over 20, with no points falling in the middle in any of the pair plots. This suggests a common ratio that is optimal for most vehicles and suggests that more expensive cars often have a higher compression ratio.

Chart

Description automatically generated

Chart, scatter chart

Description automatically generatedThe city MPG and the highway MPG are positively correlated with each other, taking a linear shape. This shows that as your city MPG increases, your highway MPG also increases. Another interesting feature is that curb weight has a negative correlation with city MPG and highway MPG. This would be expected because the heavier your car is, the less likely you are to get a good MPG. This is demonstrated with the positive correlation to engine size and curb weight since you need a larger engine to power a bigger vehicle.

**Figure 7 – Distribution Barcharts of Automobile Data Set’s Categorical Variables**

**Chart, bar chart

Description automatically generated**

**Chart, bar chart

Description automatically generated**

**Chart, waterfall chart

Description automatically generated**

**Chart, histogram

Description automatically generated**

Actual Proportions and descriptions of data are shown in Table 3.

**Figure 8 – Boxplot of Make and Price**

**Chart, box and whisker chart

Description automatically generated**

The higher proportion of cars were cheaper with most of the data occupying the lower realm of the figure than the few expensive ones. For the Jaguar and Porsche, there was not enough data to make a reliable boxplot graph as there was only one instance of each.

**Figure 9 – Boxplot of Make and Engine Size**

Chart, waterfall chart

Description automatically generatedNissan typically had the highest standard deviation on engine size, with the sample containing multiple Nissans with different engine types. Audi and Plymouth had one of the tightest deviations, with them using similar engines in all their cars. For cars with a straight line, this meant that all of the cars in the sample had the same size engine or there was only one instance of each car type.

**Figure 10 – Boxplot of Make and Normalized Losses**

Chart, box and whisker chart

Description automatically generatedThe higher value vehicles typically have higher normalized loss values, since it costs more to insure the vehicle each year. Toyota has the largest spread meaning that to insure their vehicles is not consistent and they may have a wide range of vehicle quality.

**Figure 11 – Boxplot of Make and Wheel Base**

Chart, box and whisker chart

Description automatically generatedA large amount of these brands kept the same size wheelbase for all of the cars that were included in the sample. This reduces variability and allows multiple different vehicles to have the same wheelbases.

**Figure 12 – Boxplot of Make and Bore Size**

Chart, waterfall chart

Description automatically generated**A**ll of the makes have a bore between 2 and 4, with Porsche and Volvo having the highest average bore size.

**Figure 13 – Boxplot of Make and Stroke**

**Chart, waterfall chart

Description automatically generated**

Most of the cars fall within the same range for stroke. Higher-powered cars such as Jaguar and Mercedes-Benz are slight outliers as they have a larger stroke to produce more power output for their more expensive, powerful vehicles.

**Figure 14 – Boxplot of Make and Compression Ratio**

Chart, histogram

Description automatically generatedThis graph is interesting since most of the cars used the same general compression ratio for their cars. However, Volkswagen and Peugot had other makes of cars that had high compression ratios, skewing the distribution.

**Figure 15 – Boxplot of Make and Horsepower**

Chart, box and whisker chart

Description automatically generated**W**ith Nissan having the largest deviation for engine size, it makes sense that they would have a higher deviation of horsepower. The different size engines produce different horsepower. Most companies produce similar cars, so the other makes have smaller, tighter spreads.

**Figure 16 – Boxplot of Make and Peak RPM**

**Chart, waterfall chart

Description automatically generated**

Most makes have a peak rpm that is set by the company itself. This is to protect the engine from working too hard and overheating which can cause engine failure or damage. BMW has the biggest deviation of peak rpm as they produce a variety of luxury cars. They set a specific rpm limit for each distinct style of car they produce.

**Figure 17 – Boxplot of Make and City MPG**

**Chart, box and whisker chart

Description automatically generated**

Chevrolet has the highest average city mpg. Most other cars fall between 15 and 30 miles per gallon. For example, Honda and Audi have higher city mpg deviation, but lower highway mpg deviation. This shows they are more consistent when driving on highways and will get the best gas mileage.

**Figure 18 – Boxplot of Make and Highway MPG**

Chart, box and whisker chart

Description automatically generatedThis is very similar to the city mpg, with the deviation typically being smaller with highway mpg. The averages are mostly all higher than the city mpg but retain much of the same shape and spread as the boxplot of city mpg

1. **SUMMARY OF FINDINGS**

Overall, this dataset gave insight into the proportion of cars that were surveyed. While there was some missing data throughout the set, it was mostly complete, and the missing data was dealt with so as not to affect the end results. Much of the data was either a nominal or ratio data type, which was to be expected since the categories were mainly continuous variables. When looking at the proportions of each of the categories, there are results that confirmed general life observations. For example, you would expect most of the cars to have a front engine and that most of the cars were sedans and hatchbacks as could be noticed by a quick survey while on the road as the general population more often buys cars of these types.

The correlations that were explored in this analysis could be useful for new car buyers. For example, normalized losses have the highest positive correlation with horsepower and price. This implies that they are more often expensive to own and can lead to more price loss due to insurance coverage. So as a new buyer, it may be better to settle for a car with lower horsepower and price to avoid high insurance fees. This data set provided useful insight pertaining to cars in 1985, much of which still holds true today even though it was over 35 years ago.