

**TASK**

**Exploratory Data Analysis on the Automobile Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

Summary of the data set

The automotive data set contains 205 observations of vehicles from 22 car makers. This data set contains automotives ranging from luxury vehicles (Mercedes-Benz, BMW) to budget friendly options (Dodge, Chevrolet). It is a good representation of the overall automotive market.

The data set to be analysed contains vehicle technical specifications that include engine internal properties, exterior physical properties, insurance characteristics, and cost of vehicles. The set has been serialised into 26 columns. This data set seems to have been obtained in 1985. Hence, the low prices, and the engine types that seem popular during this period whereas these types are not popular in the current market dating a few decades back.

This data can be used by various sectors ranging from insurance to vehicle fleet management by private organisation and/or government institutions. The symbolling and normalised losses variables are of interest for insurance purposes to decide on the premiums to be charged on vehicle owners or users. Fleet management (car rentals, some government departments, etc) are likely to be interested in the price, body-style, number of doors, etc for the transportation of people and light baggage.

Purpose of EDA

The objective of this EDA is to gain insights about the automotive data set. It is to gain an understanding of the structure of the data set including the variables and their data types. The exercise will also identify parameters or other information that is missing in the data set which can be used to benefit decision makers. Data will also be contextualised to get a better understanding of its meaning and purpose.

The EDA will extract the following information:

* Errors in the data set
* Outliers and Anomalies or noise
* Variable types and Meaning

Various relationships between variables will be brought to light. We will also look at trends that can be used to predict other variables that are beneficial for business decisions such as price, insurance risk rating, etc. What factors affect the price of a vehicle? What increases the risk rating of a vehicle and what influences the loss payment per insured vehicle?

Engineering specifications of automobiles will also be looked at to gain insights that affect the technical specifications of automobiles. Does the curb-weight of a vehicle affect the engine size to be installed and the type of aspiration required for this engine? What are prevalent features of a vehicle and the reasons thereof?

Description of the Data Set Variables

The data is of mixed type comprising of both categorical and continuous variables.

- symboling: assigned insurance risk rating of the car [ORDINAL]

- normalized-losses: relative average loss payment per insured vehicle [RATIO]

- make: brand of the car [NOMINAL]

- fuel-type: type of fuel used by the car [DICHOTOMOUS]

- aspiration: type of aspiration system used by the car [DICHOTOMOUS]

- num-of-doors: number of doors on the car [ODRDINAL]

- body-style: body style of the car [NOMINAL]

- drive-wheels: type of drive wheels used by the car [ORDINAL]

- engine-location: location of the engine in the car [DICHOTOMOUS]

- wheel-base: distance between the centers of the front and rear wheels of the car [RATIO]

- length: length of the car [RATIO]

- width: width of the car [RATIO]

- height: height of the car [RATIO]

- curb-weight: weight of the car without occupants or baggage [RATIO]

- engine-type: type of engine used in the car [NOMINAL]

- num-of-cylinders: number of cylinders in the engine [ORDINAL]

- engine-size: size of the engine in cubic centimeters [RATIO]

- fuel-system: type of fuel system used by the car [NOMINAL]

- bore: diameter of the cylinders in the engine [RATIO]

- stroke: length of the piston stroke in the engine [RATIO]

- compression-ratio: ratio of the volume of the combustion chamber when the piston is at the bottom of its stroke to the volume of the combustion chamber when the piston is at the top of its stroke [RATIO]

- horsepower: maximum power output of the engine in horsepower units [RATIO]

- peak-rpm: revolutions per minute at which the engine produces its maximum power output [RATIO]

- city-mpg: miles per gallon (MPG) of the car in city [RATIO]

- highway-mpg: MPG of the car in highway [RATIO]

- price: manufacturer's retail price (MSRP) of the car [RATIO]

**DATA CLEANING**

Data cleaning was performed where missing duplications, redundancy and units formatting was checked and corrected as required.

Duplications

The data does not contain independent duplications. However, there are observations that are identical except for the price data value. It is appreciated that these observations could be a result of vehicle models of the same make being sold at different years such as a 1985 model and 1984 model, with former having been slightly improved and hence a higher price. Unfortunately, each observation does not have a year in which it was observed to ascertain the above assumption.

Redundancy

All categories or series of observations are unique without any ambiguity or similar titles or names given to some categories. As an example, there is not an observation which has a car make named ‘Mercedes Benz’ and just ‘Mercedes’ or just ‘Benz’. All car makes have unique names without repetitions.

Encoding or Text Pre-processing

There are categorical values that have been given in word format for the number of doors and the number of cylinders variables. These variables have been encoded into integer values using the ‘word2number’ python extension library. It is appreciable and visually easy to identify quantities when represented in digital format as opposed to word format. See the following: eight vs 8, twelve vs 12.

Units Formatting

The audience of this EDA is assumed to be in South Africa. The data is presented in imperial units mostly used in North America. For ease of reference and contextual appreciation, all imperial units to be presented in graphical format has been converted to metric units which understandable to the audience. The average volumetric consumption of fuel over travelled distance is better understood in litres per 100km than in miles per gallon. Dimensions measured in inches are converted to millimetres. Engine sizes presented in cubic inches have been converted to litres. It is easy to understand a statement that says: a 2-litre engine compared to a 122 cubic inch engine. Horsepower is easily understood in watts of power. In this part of the world, an engine’s power is usually shown in kilowatts than in horsepower.

Along with units formatting, the wording for imperial variables was amended to metric wording. Fuel consumption given as ‘city-mpg’ now reads ‘city-L/100km’. The units have been changed accordingly in labels for various plots presented below.

There are other units that are foreign but easy to understand such as the US dollar currency used to present car prices. Such units remain unchanged. Mass is also given in kilograms and not pounds. Unitless variables such as ratios remain unchanged as well.

**MISSING DATA**

Missing data can lead to poor assumptions if not checked and handled correctly. There are both categorical and continuous data missing in the automobile data set.

Price

The most important continuous variable with missing values is the price of each model. Missing continuous data constitutes 23% of the data set, hence it is unwise to drop or ignore this data set. Imputation was carried out on much of this missing portion of continuous data. Price is MNAR as it is related to a car's make, curb weight, horsepower, external dimensions, and other factors.

Price was interpolated using existing prices for given weight of cars for a particular make where the existing prices were observed. An intelligent weighted average on curb weight and horsepower could prove to be more accurate than interpolation. The accuracy of this method is improved with more observations for a car make in question. Few observations such as those of an Isuzu where there is a total of 4 observations with 2 of the 4 missing prices, are likely to have inaccurate imputed values which can lead to a biased decision.

Normalized Losses

The normalized losses are missing completely at random (MCAR). These values, like symbolling risk factors are derived from factors not inherently related to the specifications of an automobile. They are derived by insurance companies and have little correlation to the mechanical and physical properties of a car. So as not o loose data, missing values were converted to NumPy’s NaN values.

Bore and Stroke

The missing categorical data for bore and stroke are as a result of the nature of these variables for a rotary engine. A rotor type or rotary engine is a combustion engine that uses a rotor(s) instead of the popular reciprocating pistons to convert fuel into mechanical energy. In general, a rotary engine consists of a rotor that is reuleaux triangle shaped (a triangle with sides curved in) rotating in a stadium shaped chamber. Therefore, a rotor engine has a rotor and a stadium shaped chamber instead of a piston and a cylinder. Hence, bore and stroke variables are irrelevant for this type of an engine. Bore and Stroke are MNAR. The missing values will be replaced by NaN.

Peak-rpm and Horsepower

The missing values for peak-rpm and horsepower are related to one another. Estimating them using the available information such as engine-size, compression ratio, is possible but difficult. These variables are also a function of other unavailable data in the data set. Peak-rpm and horsepower are MAR. These values were replaced by NaN values.

Number of Doors

Imputation was to replace the missing number of doors using the mode of number of doors for the same car length and body style. This can be described as MNAR missingness. The number of doors can be closely approximated using the modal value of number of doors for a particular body style where the value is missing and for all car lengths of that model that are equal to the car length with the missing number of doors.

**DATA STORIES AND VISUALISATIONS**

The following visualisations explore the relationships that exist between the data set’s variables.

Price vs Horsepower

A car with high power is pricier than those with low horsepower. Diesel engine vehicles are costly than their gas counterparts that produce the same amount of horsepower.

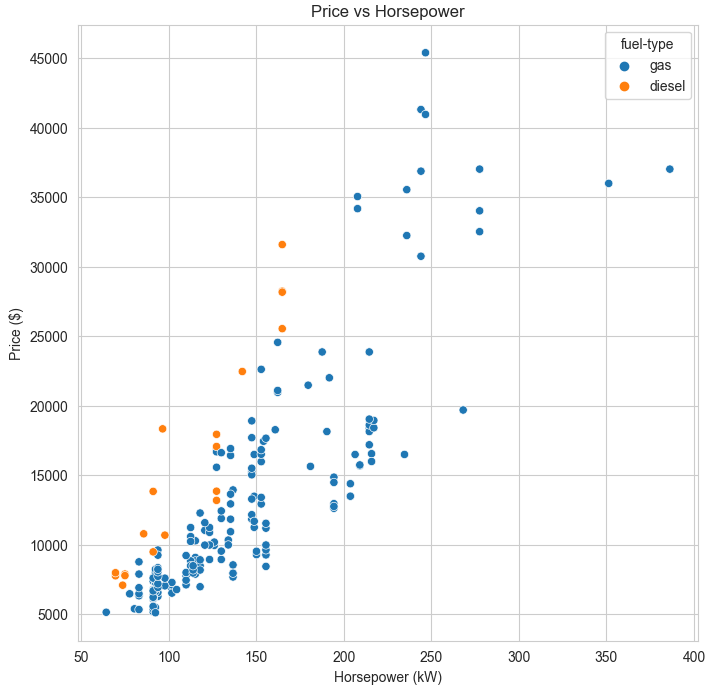


Figure : Scatter plot of Price vs Horsepower

Price vs (Wheelbase, Length, Width)

An automobile's external 3D dimensions play a role in the pricing of a vehicle. The greater the dimensions, the pricier a vehicle. Although this means that it the increase is due to an increase in usable space in the car, it is not always true. Evidence of this is shown by the 2 door vehicles which have the same price as their 4 door counterparts despite having a much shorter wheelbase of about half a meter. The price of 2 door vehicles is less dependent on wheelbase.

There are only 2 vehicles out of 205 with engines located at the rear end, and the prices of such vehicles are higher compared to cars of the same width with engines located at the front. The size of an engine is also positively correlated to a car's dimensions. Looking at cars with at least 6 cylinders, these engines are usually of the V-type and are large. It makes good sense for the width or other dimensions to increase to accommodate these mammoth engines.

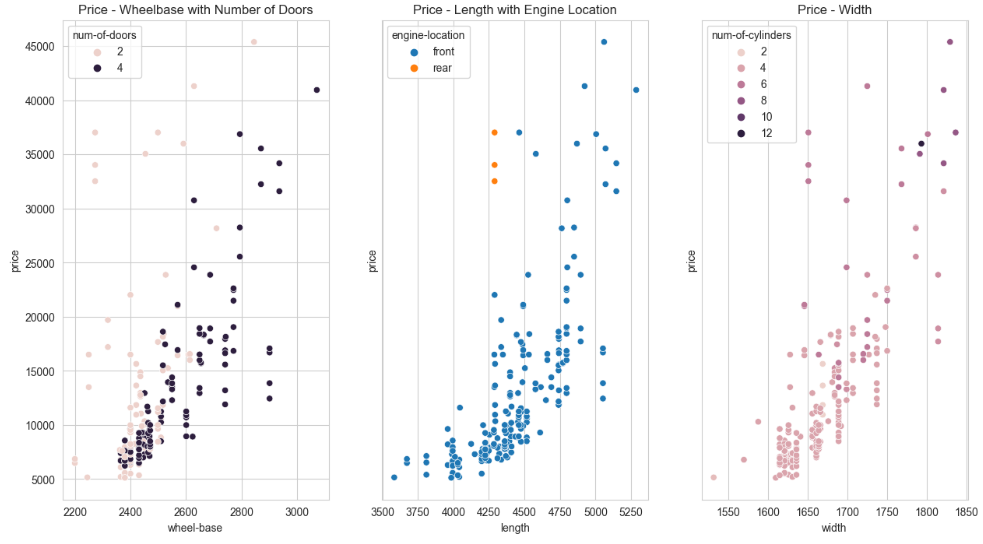


Figure : Scatterplots of Price vs Wheelbase, Length, Width

Curb Weight vs Horsepower, Engine Size, and Type of Aspiration

The size of an engine affects its weight which in turn affects the weight of the vehicle. We notice that, without any increase in engine size, forced induction (turbo charged) engines result in in a heavier vehicle compared to standard aspiration engines with a similar volumetric size. A turbo charged engine produces more power compared to a naturally aspirated engine of the same size. With more power, a vehicle needs the right curb weight for its safety.

A picture containing text, screenshot, diagram, number

Description automatically generated

Figure : Scatterplot of Curb-weight and Engines Size

Price vs Engine Size and Drive Wheels

Engine size has a significant role in increasing the price of a vehicle. The wheel drive system also has huge impact on the cost of a vehicle.

The affordable cars to drive are those with a forward wheel drive system. Their popularity is also high at about 60%. For a 37% of rear-wheel-drive cars, the price varies greatly by almost 3 times as much as the forward wheel and 4-wheel drive cars. Rear wheel drive cars also expensive to own than other wheel drive systems.

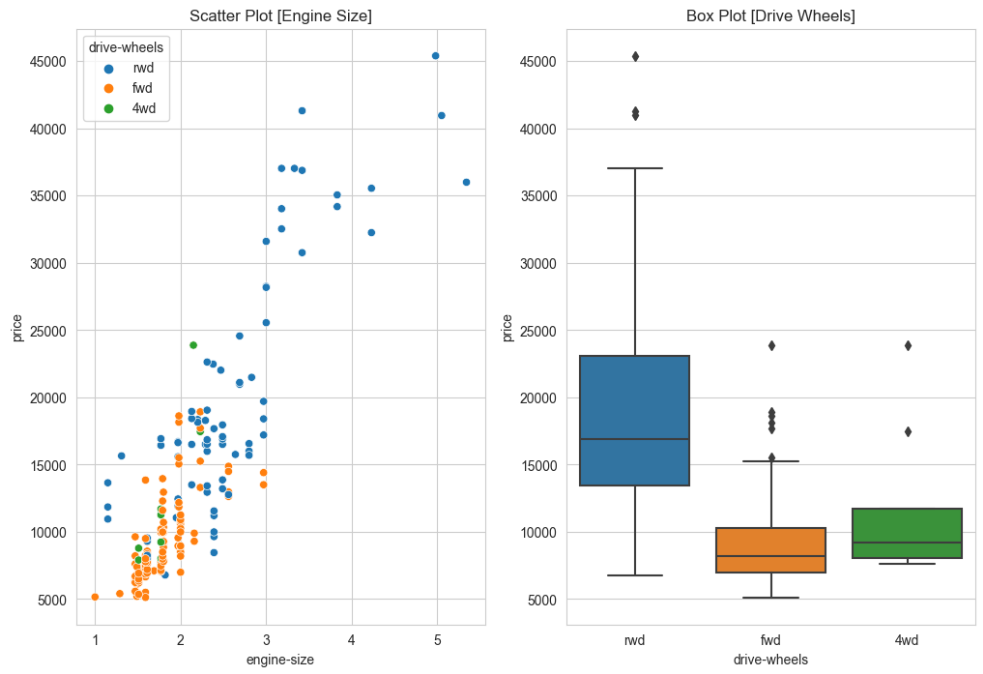


Figure : Scatter plot (Price and Engine size) and a Whisker plot (Price and Drive Wheels)

Distribution of Drive Wheels of Cars per Make

Forward wheel drive cars are the most popular at almost 60% of cars in the market. Few car makers (Audi, Subaru, and Toyota) have 4-wheel drive vehicles.

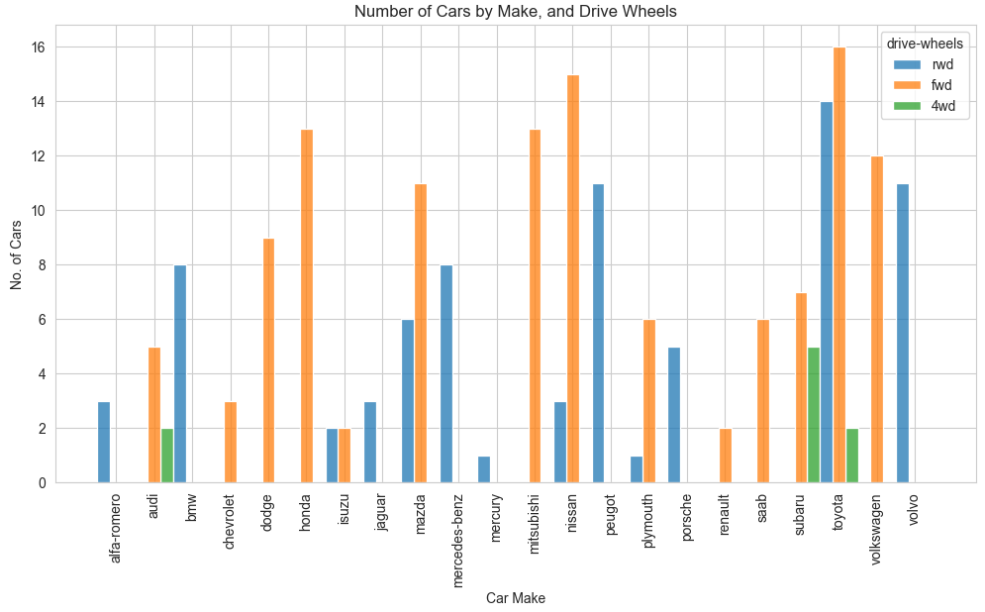


Figure : A bar graph of the number of cars with a particular wheel drive per car make.

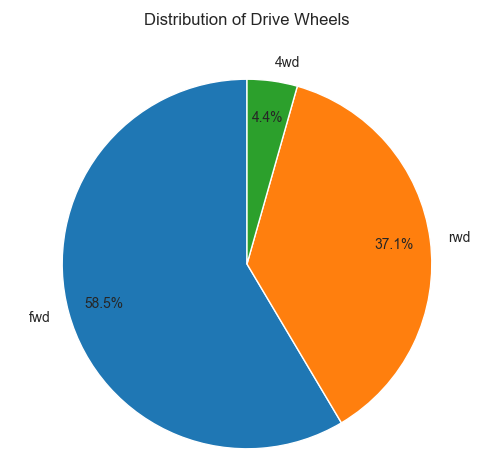


Figure : A Pie Chart for Distribution of Drive Wheels

Average Price of Each Wheel Drive Type per Make

The expensive high-end cars are Jaguar, Mercedes-Benz, Porsche and BMW all with average prices above $25 000. All the car makers drive their vehicles using the rear wheels. Looking at all other car makes; rear wheel drive cars are expensive averaging $15 000 compared to front wheel drives at average of $10 000 for each car make. Notably, Isuzu cars seem to cost the same regardless of the drive wheel arrangement and that 50 % of price values for this make were imputed.

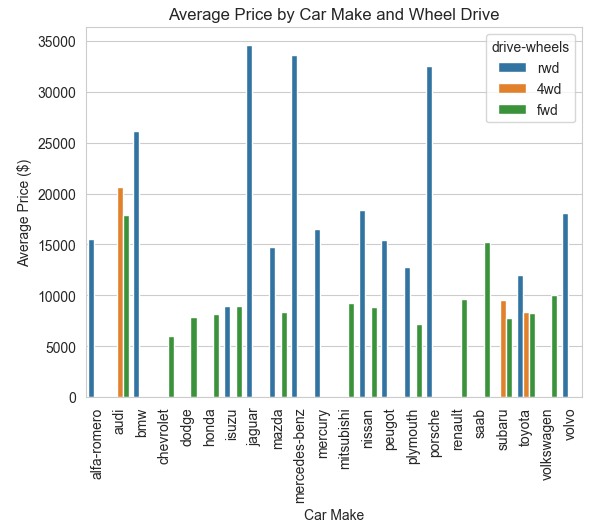


Figure : A histogram of Average Price of each Wheel Drive Type per Car Make

Average Price of Each Wheel Drive Type per Make

From the whisker plots below, hatchbacks seem to have a low variance seen by the shortest IQR and low average price. Hatchbacks are generally affordable vehicles and have prices that are reliably close to the average price. Convertibles and hardtops vary greatly in price and have the fewer cars in the market.

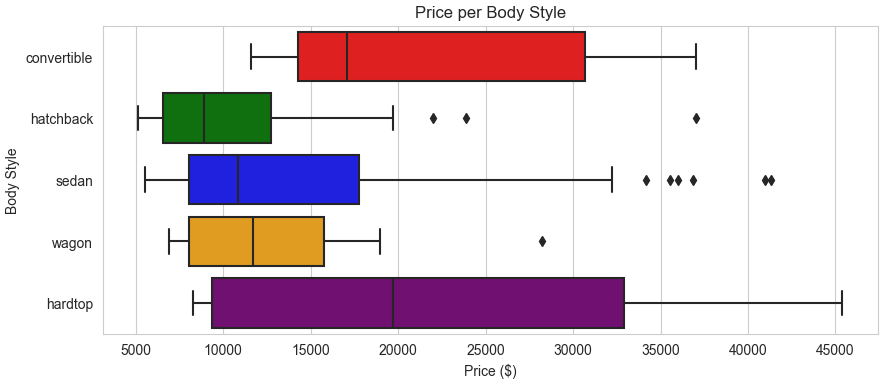


Figure : Whisker Plots of Average price per Body Style

The market has over 80% of cars as sedans and hatchbacks. Notice that hatchbacks are also the cheapest cars in the market.

A picture containing text, diagram, screenshot, circle

Description automatically generated

Figure : A Pie Chart for Distribution of Body Styles

Horsepower vs Engine Size for Engine Types

The greater the size of an engine, the higher the power output. However, the power output also largely depends on the type of an engine being used. Notably, a rotor engine is one of the small sized engines with a higher power to size ratio despite having a few data points (4 rotor engines) for this data and from only 1 car maker. A double overhead cam with values (dohcv) seems to be producing a higher power output, however this cannot be ascertained with only one engine of this type given. Notice that power output is also dependent on the aspiration of the engine which is not indicated in this plot.

In general, ohcv engines are large compared to most engine types.

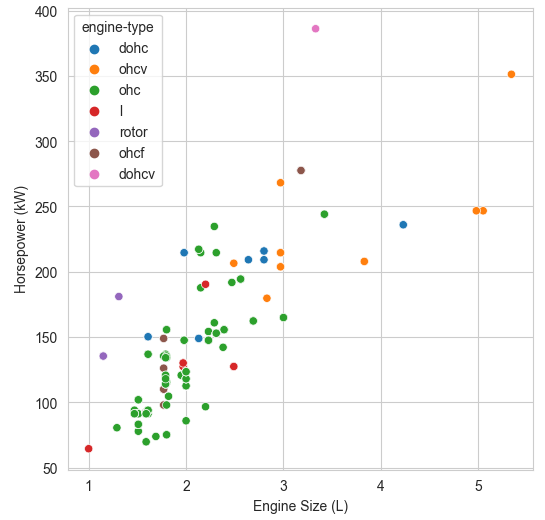


Figure : A Scatterplot of Horsepower vs Engine Size

Distribution of Engine Types

Which engine type is the most popular and why? As a reminder, this data set represents cars manufactured in the 1980s. These conclusions are for this period and will vary with present trends. Double overhead camshaft (dohc) engines are known to be popular today.

Over 7 out of 10 cars have an OHC or Over Head Camshaft engine. An OHC engine is the most popular because of its simplicity and better performance.

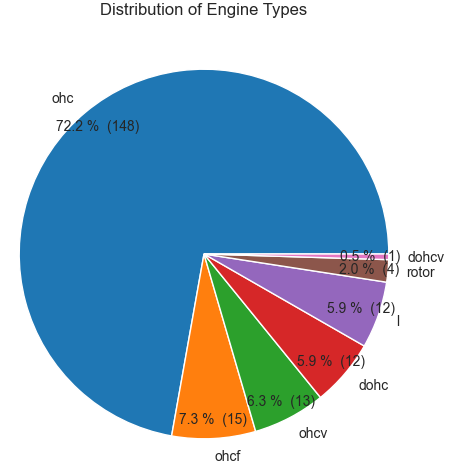


Figure : A Pie Chart for Distribution of Engine Types

An ohc engine is the most affordable in the market hence the popularity in the market amongst other factors.

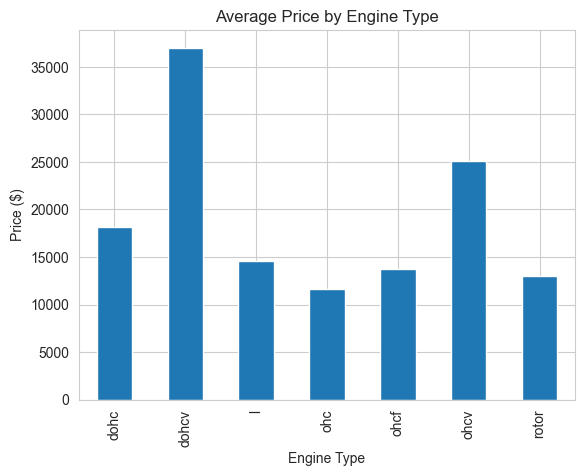


Figure : A Bar Graph of Average Price for each Engine Type

Distribution of Cars by Risk (Symbolling)

1 in every 10 cars has excellent safety features and is the safest with a risk or symbolling rating of +3. Most cars at 32% of the market have average safety features and are considered moderate risk at a rating of 0, while 26% percent of cars are just above average at a rating of +1. Thankfully, there are no cars that are very risky, however, there are vehicles that are risky but have minor features to mitigate risk with a rating of -2 at 1.5% of the cars (3 out of 205 cars). 10% of vehicles have safety features that are below average.

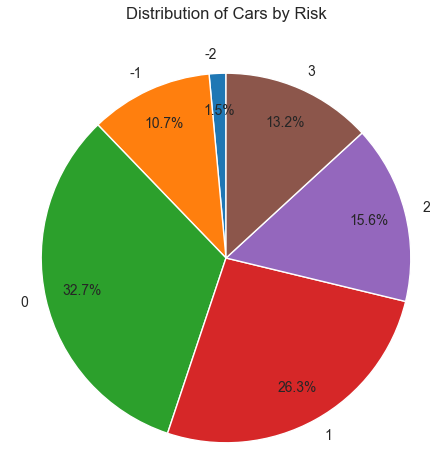


Figure : A Pie Chart for Distribution of Risk amongst Cars

Average Price of Cars vs Risk Rating

One would expect otherwise, however, cars that have above average to good safety features are some of the cheapest cars in the market on average. Expectedly, cars with excellent safety features are pricier on average.

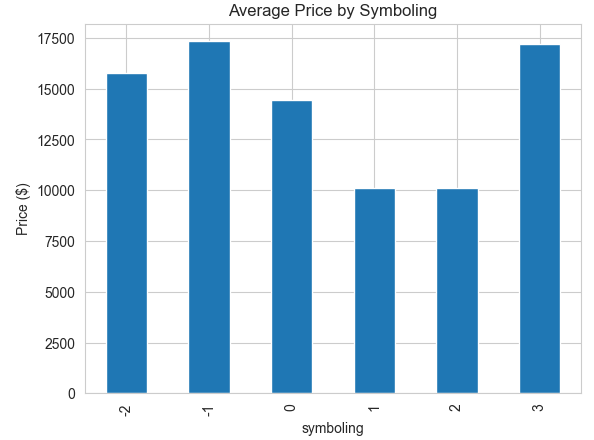


Figure : A Bar Graph of Average Price of Cars per Risk Rating

Normalized Losses vs Car Height

Insurance premiums are partially based on the height of a car. Lower cars are more likely to get seriously damaged in accidents than high cars. High cars usually have big or larger diameter wheels, making them less likely to get damaged in collision with smaller cars. There is little correlation between a car's make and the insurances losses associated with it.

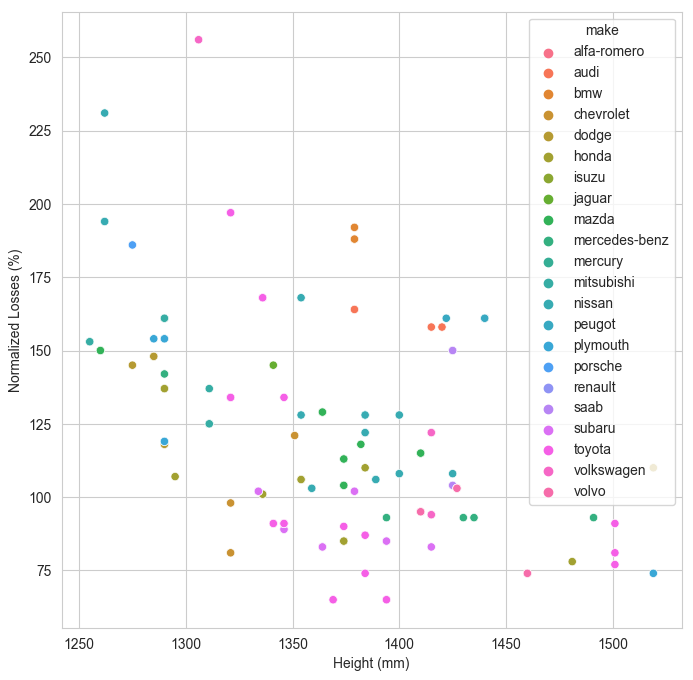


Figure : A Scatterplot of Normalized Losses for Car Heights

Average Fuel Consumption vs Engine Size

The bigger an engine is, the more fuel it uses per kilometre for a given speed in both city and highway driving conditions. The difference in fuel consumption between city and highway driving conditions can vary by as much as 5 litres per 100 km. This difference is higher than the total city and highway fuel average consumptions for small engines (less than 1.5 litres).

The spikes or high jumps in both city and highway fuel consumptions seen for small engines are from rotary engines. Rotary engines are light in weight and produce more horsepower than their counterparts of the same size. However, this high power is produced from significantly high volumes of fuel required to drive the engine. Hence these spikes are not an anomaly but a reality with rotary engines.

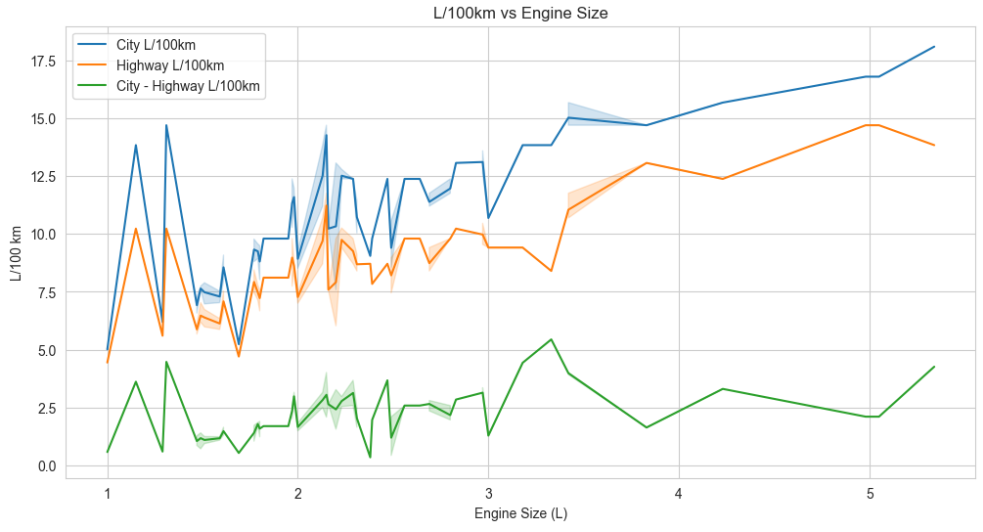


Figure : Line Graphs of Average Fuel Consumption in Cities and Highways and their Differences for Engine Sizes

**THIS REPORT WAS WRITTEN BY: FORTUNE NCUBE**

