Exploratory Data Analysis on the Automobiles Dataset

Introduction

The following EDA will be focused on the Automobiles dataset. The dataset consists of 193 observations and 24 variables, after the data has been cleaned. Each observation provides information concerning a specific automobile, according to the variables outlined. This exploratory data analysis will explore the inter-variable relationships as well as the influence that different variables had on the price of automobiles.

The variables include:

Variable	Definition	Description	Data Types
make	The manufacturer or brand of automobile.	Alpha-Romero, Audi, BMW, Chevrolet, Dodge, Honda, Isuzu, Jaguar, Mazda, Mercedes-Benz, Mercury, Mitsubishi, Nissan, Peugot, Plymouth, Porsche, Saab, Subaru, Toyota, Volkswagen, Volvo	Nominal
fuel-type	The type of fuel used by the automobile.	Gas or Diesel	Nominal
aspiration	The way in which air enters the cylinders of the engine.	std = Standard turbo = turbocharged	Nominal
num-of-doors	The number of doors.	2 doors or 4 doors	Ordinal
body-style	The body-type of the automobile.	Convertible, Hatchback, Sedan, Wagon, Hardtop	Nominal

drive-wheels	The drive type.	fwd = front-wheel drive, rwd = rear-wheel drive, 4wd = four-wheel drive	Nominal
engine-location	Position of the engine	Front or rear	Nominal
wheel-base	The distance between the front axels and the rear-axels.	-	Continuous.
length	The length of the automobile.	-	Continuous
width	The width of the automobile.	-	Continuous
height	The height of the automobile.	-	Continuous
curb-weight	The weight of the automobile, excluding any load.	-	Continuous
engine-type	The type of engine.	dohc = Double overhead camshaft ohcv = Overhead valve ohc = Overhead Camshaft l – Inline Engine ohcf = Overhead Camshaft, Front	Nominal
num-of-cylinders	Number of engine cylinders.	3, 4, 5, 6, 8, 12	Ordinal
engine-size	The size of the engine.	-	Continuous

fuel-system	Fuel injection type	mpfi = Multi-point fuel injection 2bbl = Two-barrel carburettor mfi = Multi-point fuel injection 1bbl = Single barrel carburettor spfi = Single-point fuel injection idi = Indirect injection diesel spdi = Single-point direct injection	Nominal
bore	The diameter of the cylinder bore.	-	Continuous
stroke	Piston stroke length	-	Continuous
compression-ratio	Engine compression ratio	The ratio of the cylinder volume.	Continuous
horsepower	The engine output measured in horsepower.	-	Continuous
peak-rpm	Peak engines revolutions per minute.	-	Continuous
city-mpg	Fuel efficiency in the city.	Miles per gallon	Continuous
highway-mpg	Fuel efficiency on the highway.	Miles per gallon	Continuous
price	The price of the car.	-	Continuous

Data cleaning

The Automobiles dataset was loaded using pandas and the structure of the dataset was viewed.

The size of the dataset was inspected and found to consist of 205 observations and 26 columns. The 'symboling' and 'normalized-losses' columns were removed from the dataset, leaving the dataset with 24 columns.

The dataset was inspected for duplicated observations and none were found.

The data types in each column were checked and there were inconsistencies found. While the following variables: bore, stroke, horsepower, peak-rpm and price, are integers they were all saved under the object datatype.

Column Name	Data Type
make	object
fuel-type	object
aspiration	object
num-of-doors	object
body-style	object
drive-wheels	object
engine-location	object
wheel-base	float64
length	float64
width	float64
height	float64
curb-weight	int64
engine-type	object
num-of-cylinders	object
engine-size	int64
fuel-system	object
bore	object
stroke	object
compression-ratio	float64
horsepower	object
peak-rpm	object
city-mpg	int64
highway-mpg	int64
price	object

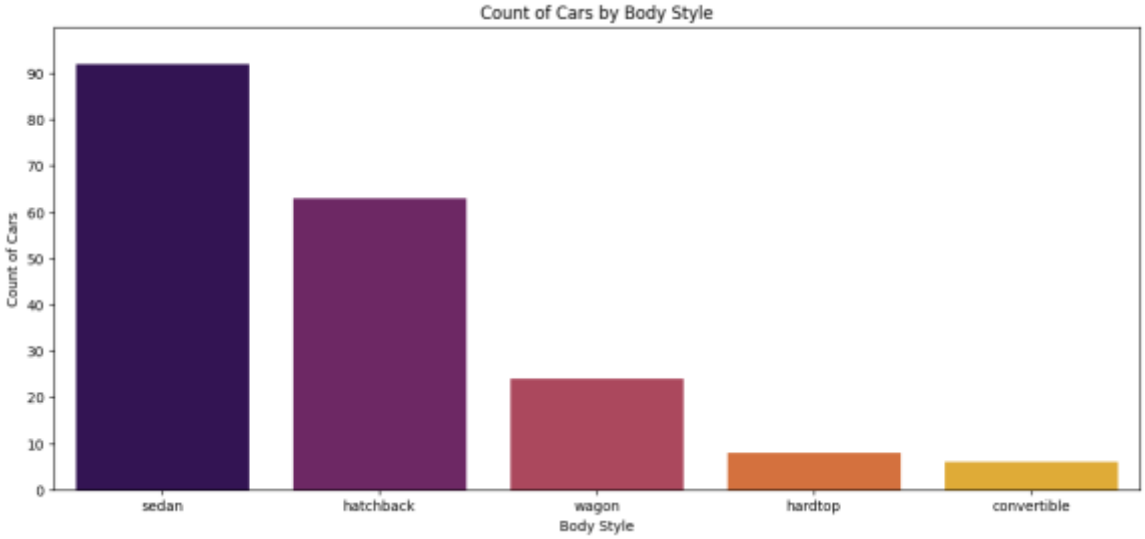
The datatypes of these variables were corrected as follows:

The bore and stroke were changed to a float datatype in order to preserve their decimal points and keep the data accurate. While the horsepower, peak-rpm and price were changed to integers.

Column Name	Data Type
make	object
fuel-type	object
aspiration	object
num-of-doors	object
body-style	object
drive-wheels	object
engine-location	object
wheel-base	float64
length	float64
width	float64
height	float64
curb-weight	int64
engine-type	object
num-of-cylinders	object
engine-size	int64
fuel-system	object
bore	float64
stroke	float64
compression-ratio	float64
horsepower	int64
peak-rpm	int64
city-mpg	int64
highway-mpg	int64
price	int64

The ‘body-style’ column was inspected and the Sedan body-type was found to be the most common.

Body-Style	**Number of Observations**
Hatchback	63
Convertible	6
Sedan	92
Wagon	24
Hardtop	8



The ‘price’ column was inspected for uniqueness as well as the maximum and minimum prices of the vehicles. It was determined that the maximum price was 45400, while the minimum price was 5118.

The ‘make’ column was then inspected for uniqueness. There were 2 spelling errors corrected, ‘Alfa-Romero’ was changed to ‘Alfa-Romeo’ and ‘Peugot’ was changes to ‘Peugeot’. The following 21 brands were identified:

- Alfa-Romeo
- Audi
- BMW
- Chevrolet
- Dodge
- Honda
- Isuzu
- Jaguar
- Mazda
- Mercedes-Benz
- Mercury
- Mitsubishi
- Nissan
- Peugeot
- Plymouth
- Porsche
- Saab
- Subaru
- Toyota
- Volkswagen
- Volvo

The uniqueness of the following engine related variables were inspected:

‘num-of-cylinders’ column - Cars ranged from having 3 cylinders to 8 cylinders.

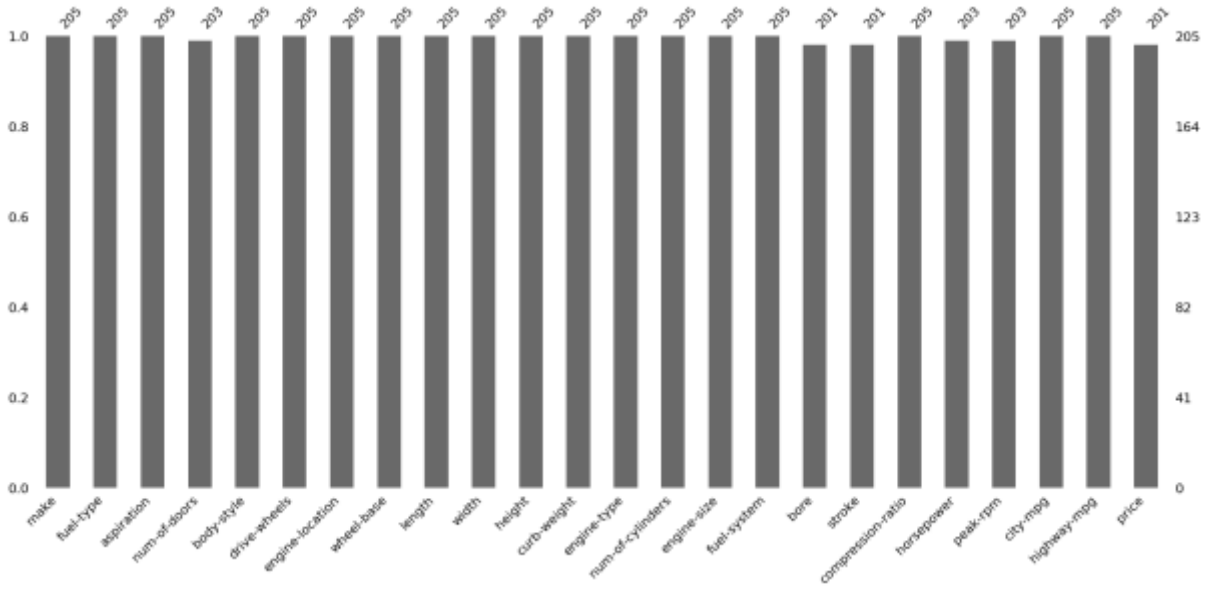
‘aspiration’ column - It was found that cars vary based on the engine aspiration type. The engine aspirations are either standard or turbo-charged.

Missing data

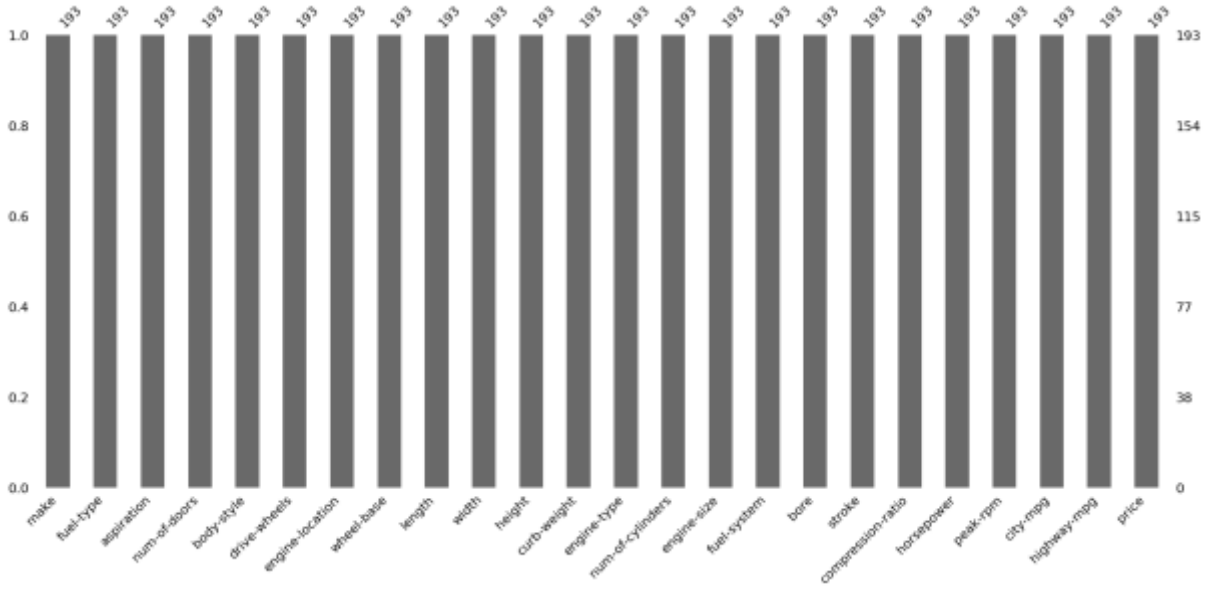
Missing values were captured as question marks (?) in the dataset. Missing values were found in the following columns:

- Num-of-doors – 2 missing values
- Bore – 4 missing values
- Stroke – 4 missing values
- Horsepower – 2 missing values
- Peak-rpm – 2 missing values
- Price – 4 missing values

The question marks in the above columns were converted into NA values and then visualised. The following bar graph depicts the rows with missing data.

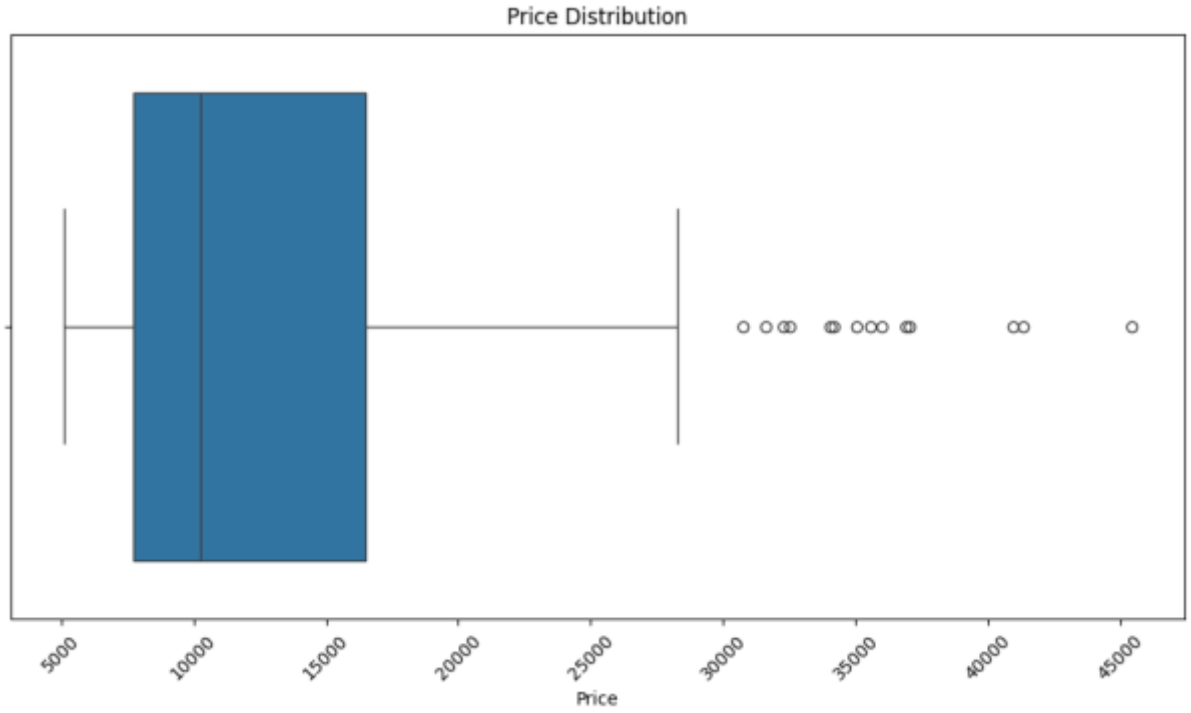


After being converted to NA values, the missing data was dropped from the table, resulting with the 193 remaining observations.



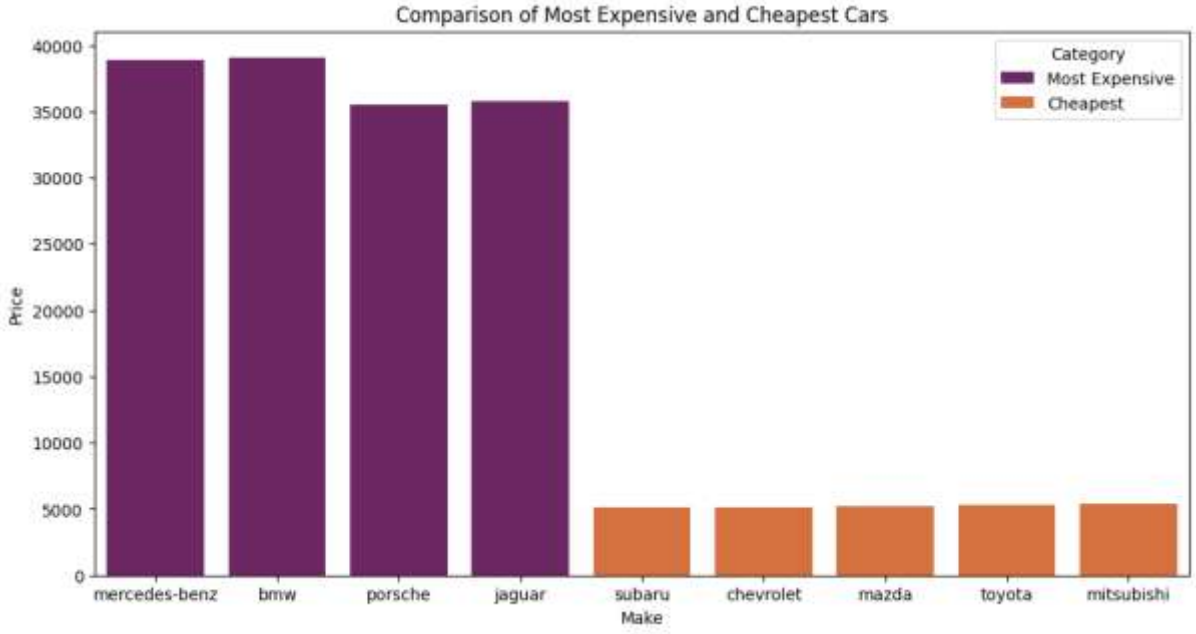
Data Stories and Visualisations

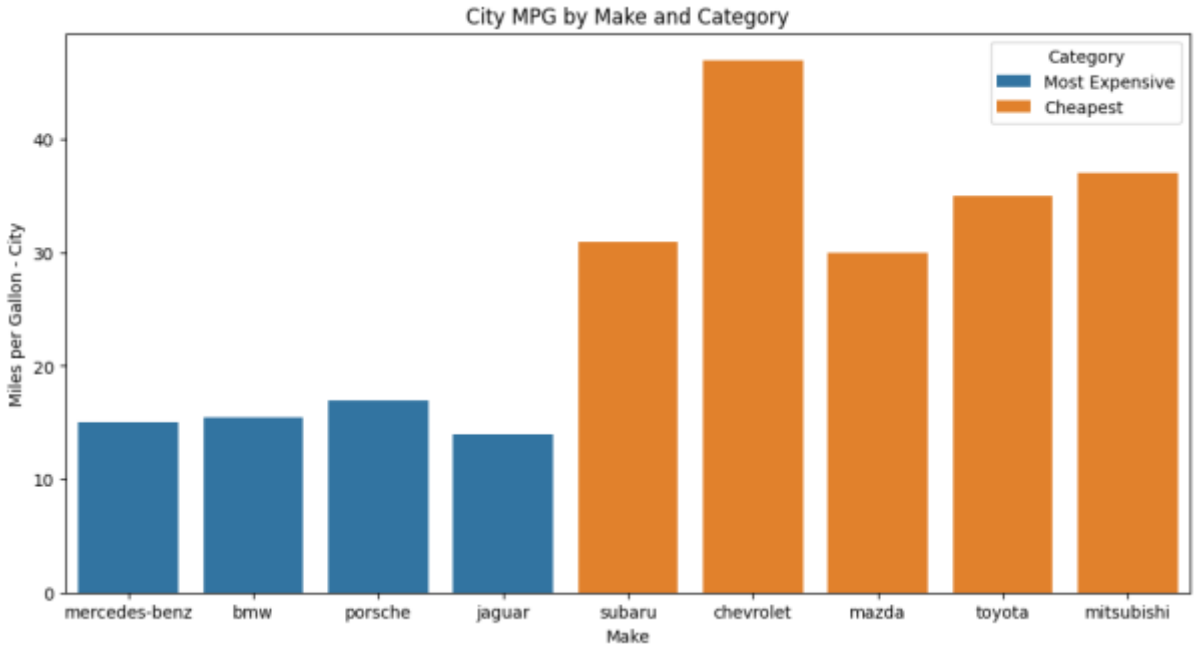
Price Distribution



The boxplot shows the spread of the prices of cars in the dataset. From the boxplot we see that the lowest price is around 5000, the median is roughly around 10 000, and the highest price is an outlier around 45 000. The data is skewed to the right and most of the automobiles are priced in the low to mid-range. The outliers may reflect high-end or luxury automobiles.

Price Distribution in Relation to Fuel Consumption

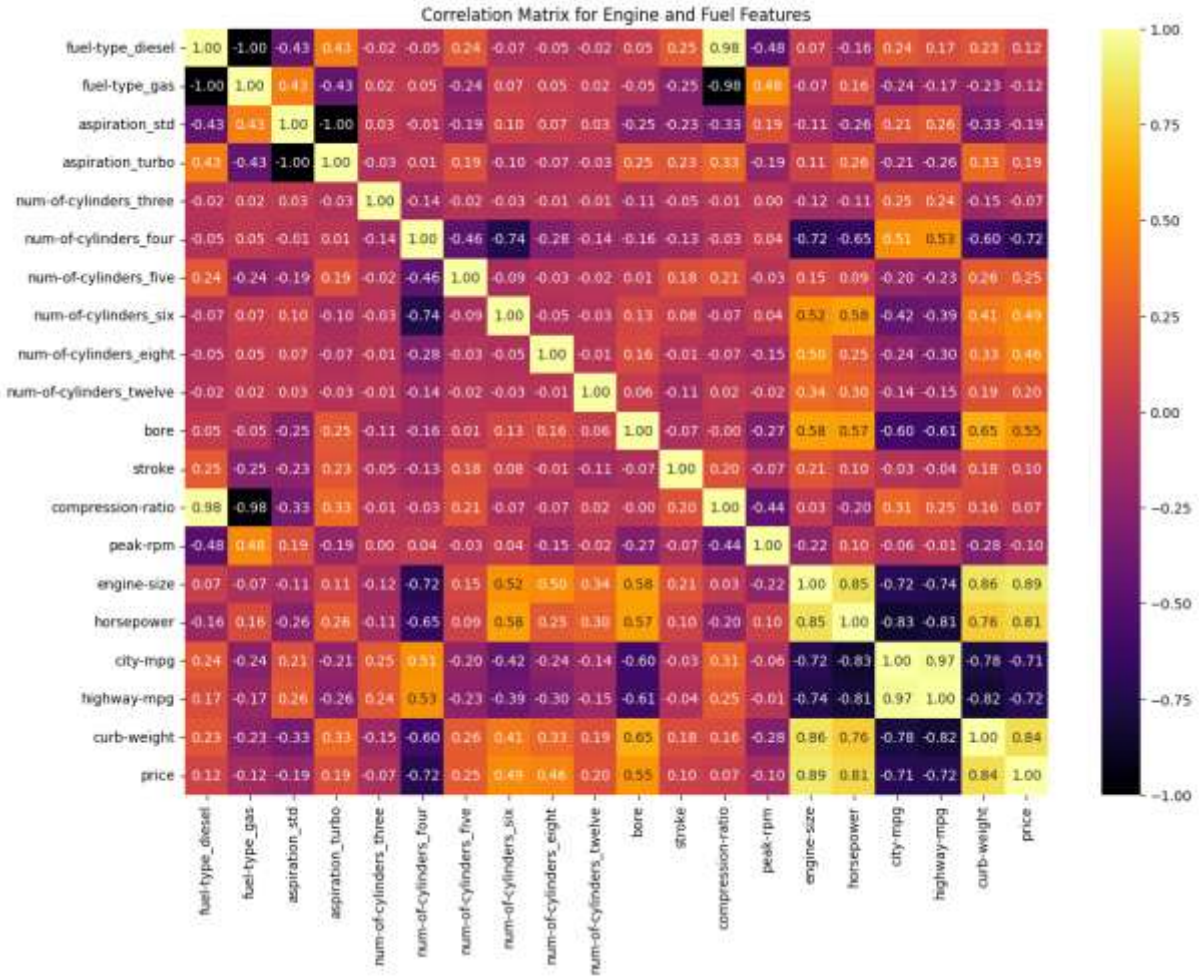




Mercedes-Benz, Porsche, BMW and Jaguar are the more expensive, higher end cars. While, Subaru, Chevrolet, Mazda, Toyota and Mitsubishi are the more affordable cars, in the lower price range.

The cheaper cars are more fuel efficient than the expensive cars. Mercedes-Benz is not only the most expensive make, but it also has the highest fuel consumption of around 10 miles per gallon in the city. While Chevrolet, a cheaper car, has the lowest fuel consumption of around 45 miles per gallon in the city. This makes the cheaper cars an affordable option to buy but also an affordable option to maintain.

However, the increased fuel consumption in more expensive cars can be attributed to bigger engine sizes and power.



From the correlation matrix, engine-size and horsepower are the two main performance related factors that influence the price. The engine-size and price have a positive correlation of 0.89, therefore cars with larger engines are values at higher prices. The horsepower and price have a positive correlation of 0.81, showing that faster, more powerful cars are valued at higher prices. This is further proven by the engine-size and horsepower also have a strong positive correlation of 0.85. As the engine-size increases, so does the horsepower, and thereby the price.

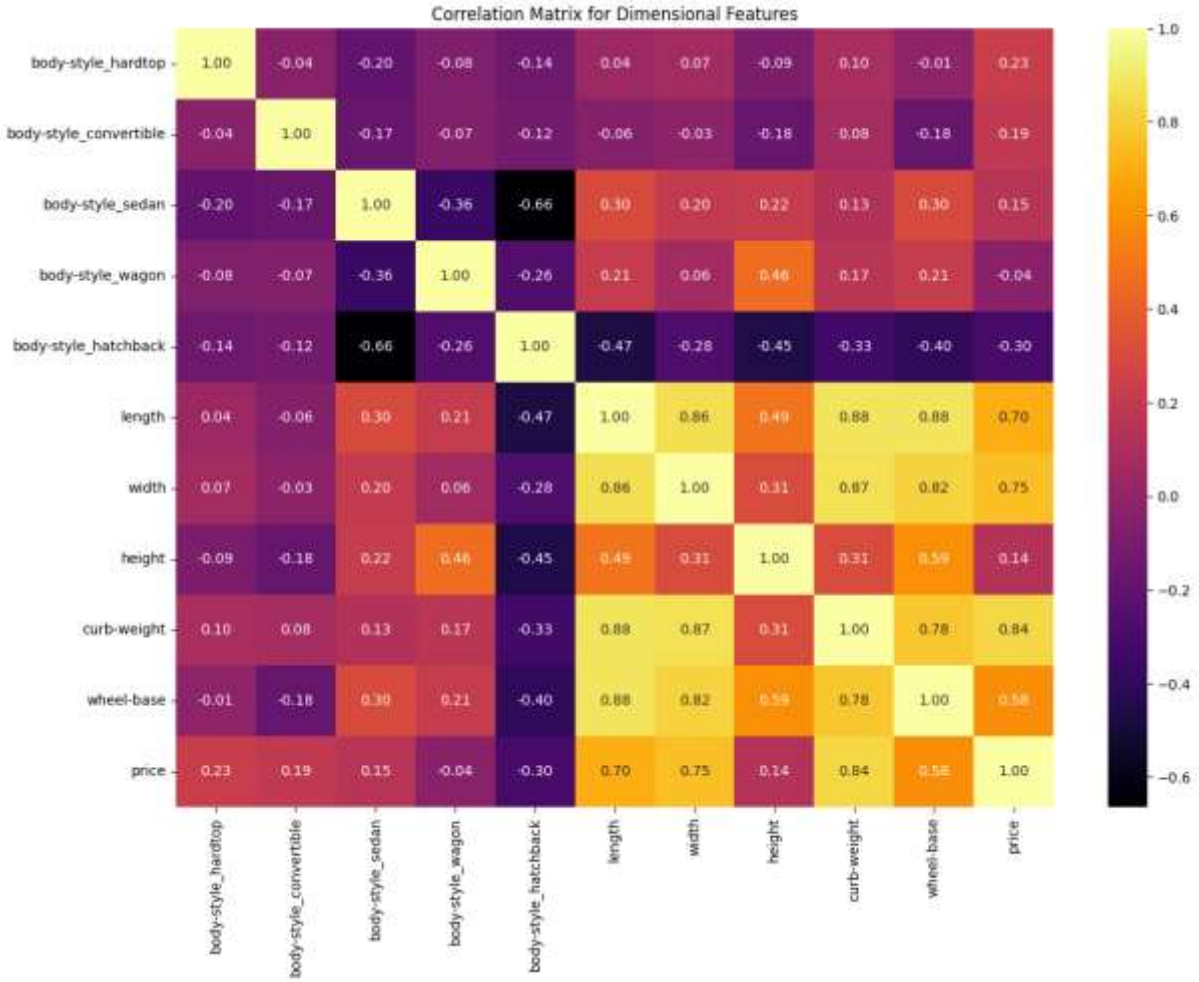
Additionally, the fuel efficiency for a car in the city (city-mpg) has a strong positive correlation with the fuel efficiency on a highway (highway-mpg), 0.97 correlation. From this we can conclude that fuel efficiency is consistent, if cars are fuel efficient in the city then they are more likely to be more fuel efficient on the highway and vice versa.

However, fuel efficiency has an inverse relationship to engine size and horsepower. This negative relationship is proven by the following negative correlations:

- Engine-size and city-mpg: -0.72
- Engine-size and highway-mpg: -074
- Horsepower and city-mpg: -0.83
- Horsepower and highway-mpg: -0.81

From the findings above it can be inferred that as that as the engine-capacity and power of the car increases, the car becomes less fuel efficient, as a larger input is required for a stronger output.

Price Distribution in Relation to Vehicle Dimensions



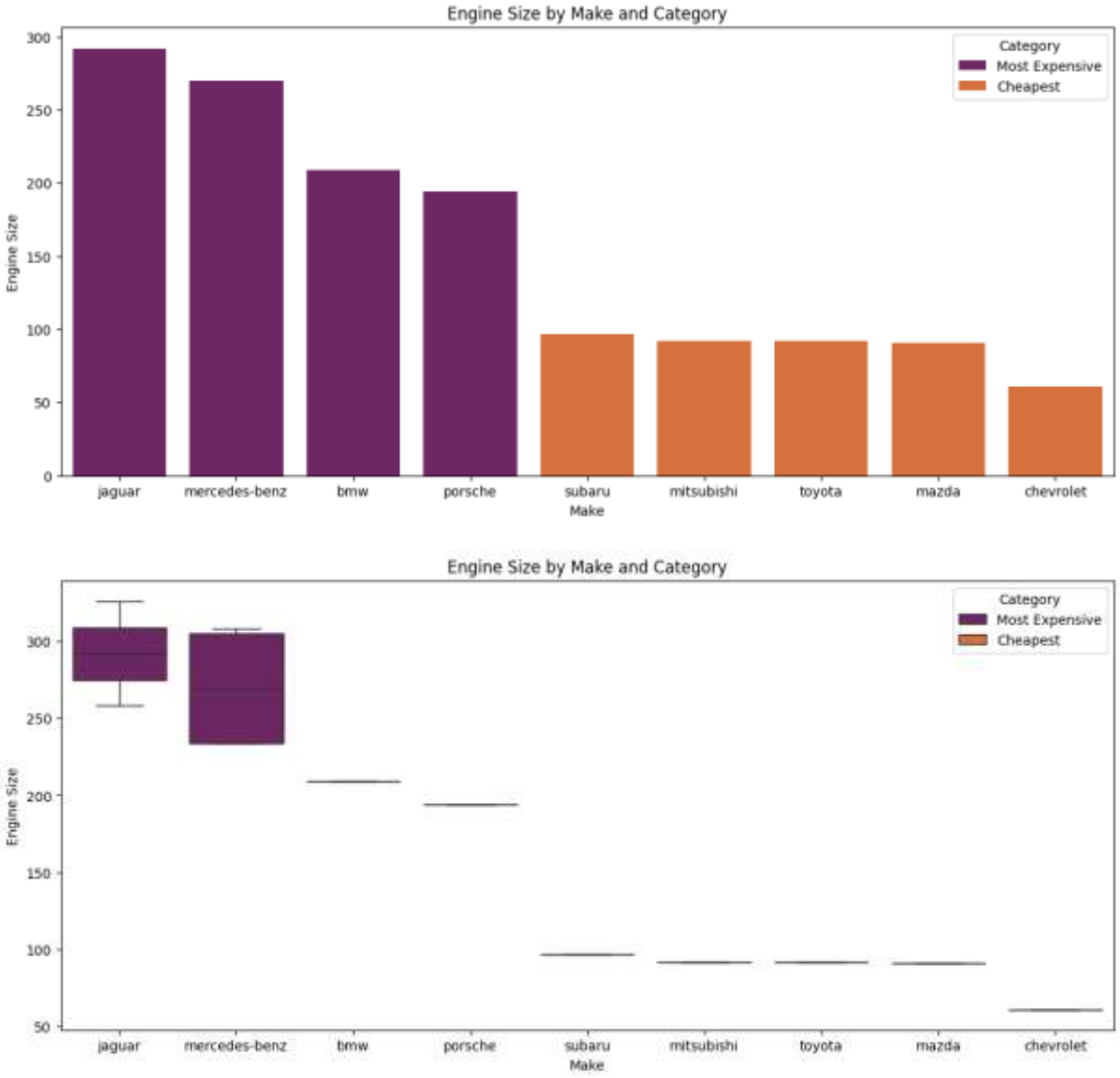
From the correlation matrix above, the strongest positive correlations that influence the price are:

- Curb-weight and price: 0.84
- Width and price: 0.75
- Length and price: 0.70

A general relationship can be derived from the data above. As a vehicle's weight and size increases, so does its market value or price. This relationship may be tied to consumer trends as larger vehicles may be thought of as more luxurious or better quality as they present as studier vehicles. The relationship may also be tied to larger cars being more complex to design and engineer.

Breakdown of Key Features

Engine-size Distribution



From the graphs above, the bar plot depicts the engine size of the different vehicle makes. The boxplot provides insight into the spread of the engine-size.

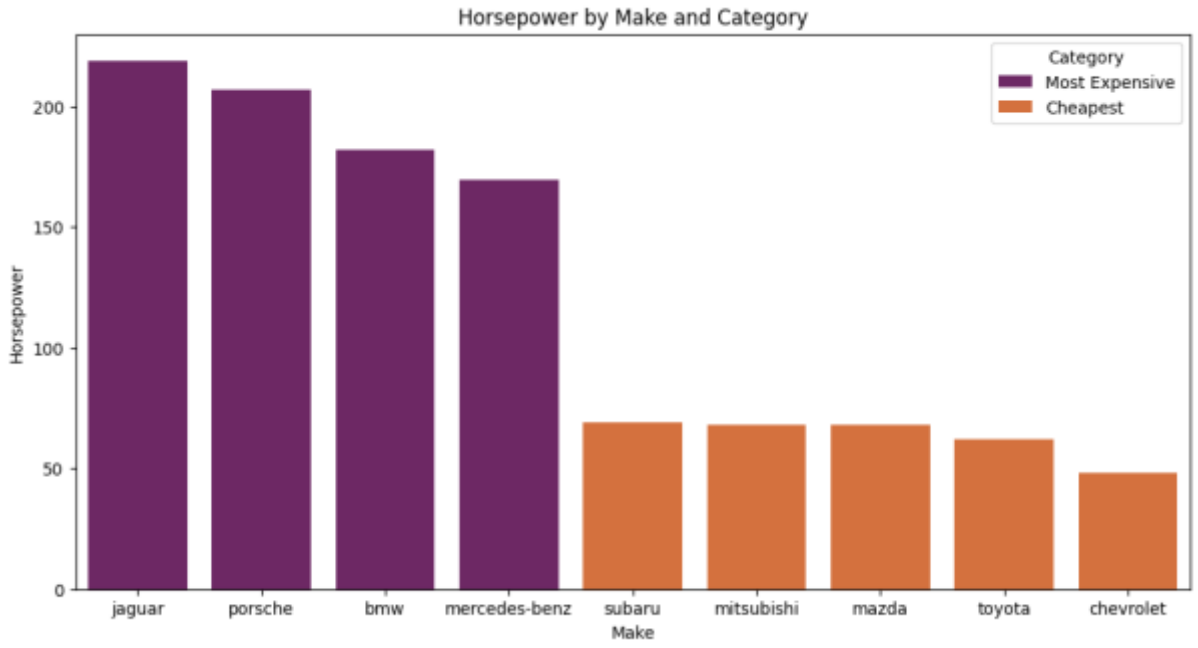
There is a clear distinction between high-end manufacturers and low-end manufacturers. The expensive cars (Porsche, BMW, Mercedes-Benz and Jaguar) feature larger engine sizes ranging from around 200 to greater than 300. While the cost effective manufacturers (Chevrolet, Mazda, Mitsubishi, Toyota and Subaru) feature smaller engine sizes ranging from around 60 to 100. This further proves that the engine size of a car impacts the price, with higher end cars having larger engine capacities.

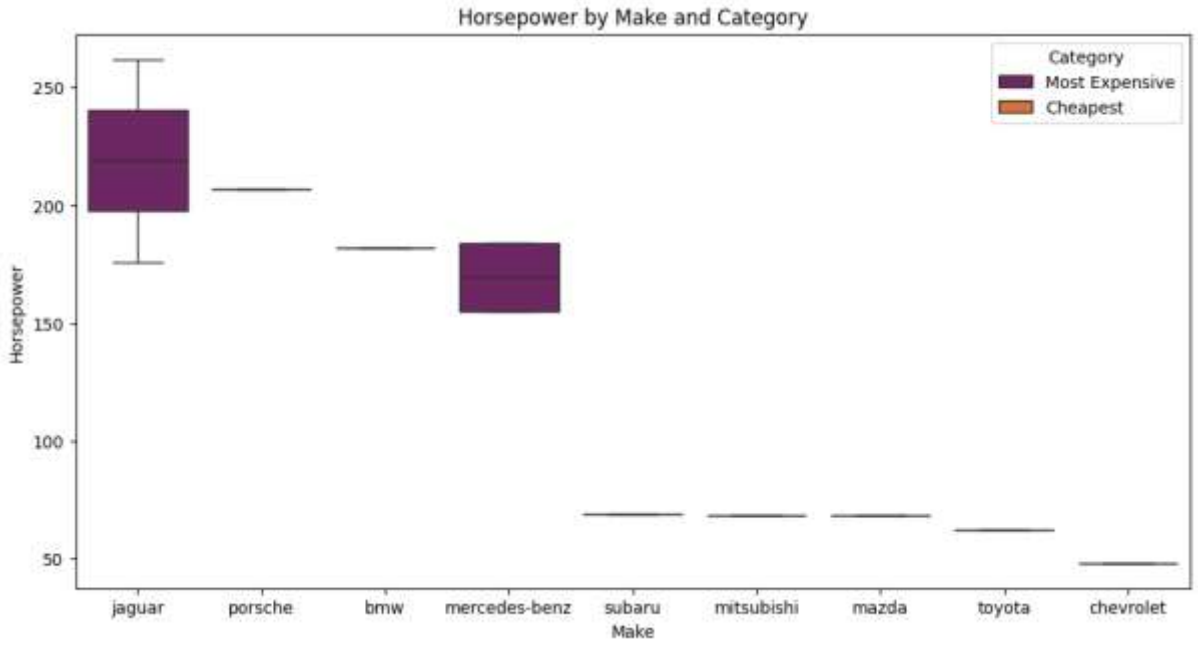
There is also a degree of variation across the different manufacturers. Mercedes-Benz and Jaguar offer the most variation, indicated by the spread of their boxplots. This presents the consumer with more choices when selecting a vehicle, based on engine size. There is little to no variation in the engine size of the following makes: BMW, Porsche, Subaru, Mitsubishi, Toyota, Mazda, and

Chevrolet, indicated by their narrow boxplots. While these brands have little to no variation, it can be inferred that they prioritise consistency across their brands.

These findings provide further prove that engine-size influences car prices, as larger engine-sizes are associated with higher-end cars. The clear positive relationship between engine-size and fuel consumption can also be seen by, Chevrolet having the smallest engine-size and the lowest fuel consumption and Jaguar having the largest engine-size and largest fuel consumption.

Horsepower Distribution





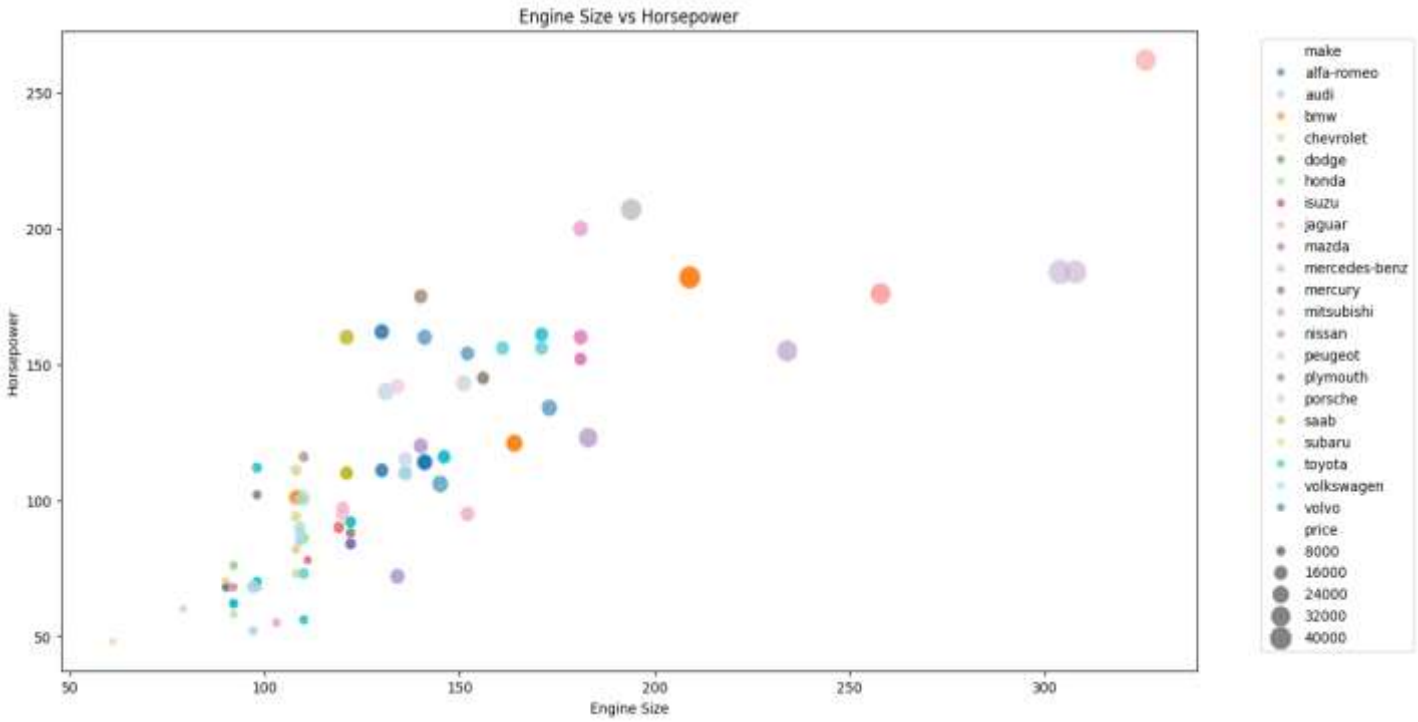
From the graphs above, the bar plot depicts the spread of the horsepower across higher-end and lower-end vehicles.

The expensive cars (Porsche, BMW, Mercedes-Benz and Jaguar) have a larger horsepower ranging from around 150 to greater than 250. Therefore showcasing their stronger, more performance related engines. Whereas the cost effective cars (Chevrolet, Mazda, Mitsubishi, Toyota and Subaru) have a horsepower ranging from around 50 to 75, therefore having weaker engines.

There is little to no variation in the horsepower of the following cars: BMW, Porsche, Subaru, Mitsubishi, Toyota, Mazda, and Chevrolet. This is consistent as the data shows that these makes have little to no variation in their engine size.

While the engine size of the Mercedes-Benz is larger and has more variation which will result in a range of horsepower levels. BMW and Porsche have a higher more consistent level of horsepower in their vehicles.

Engine-size vs Horsepower

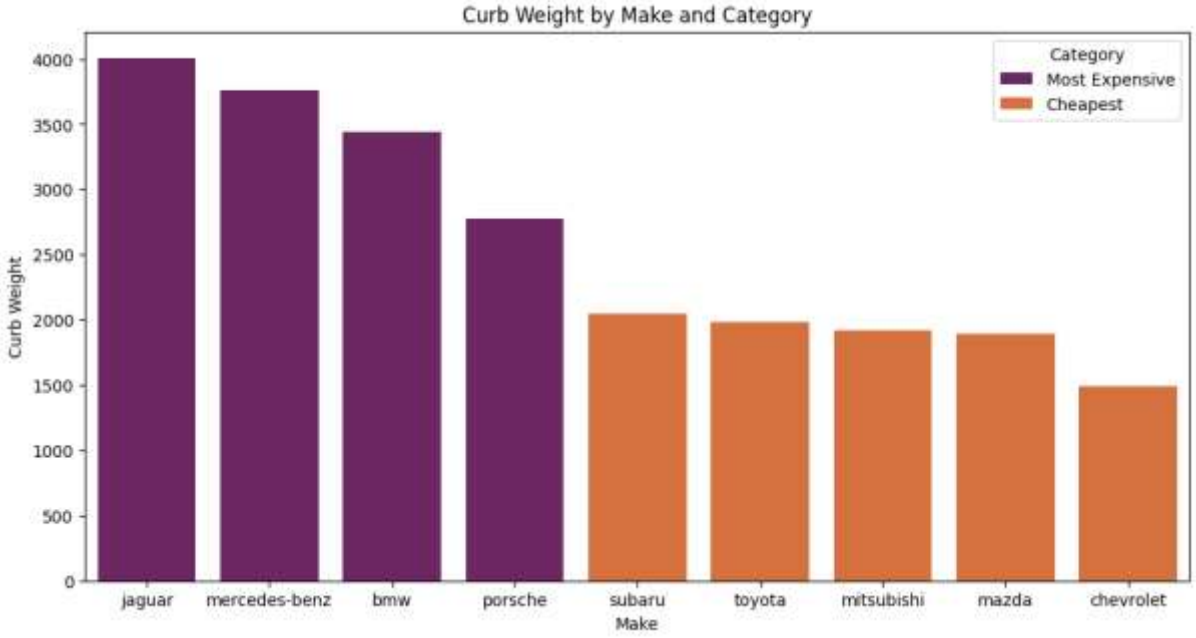


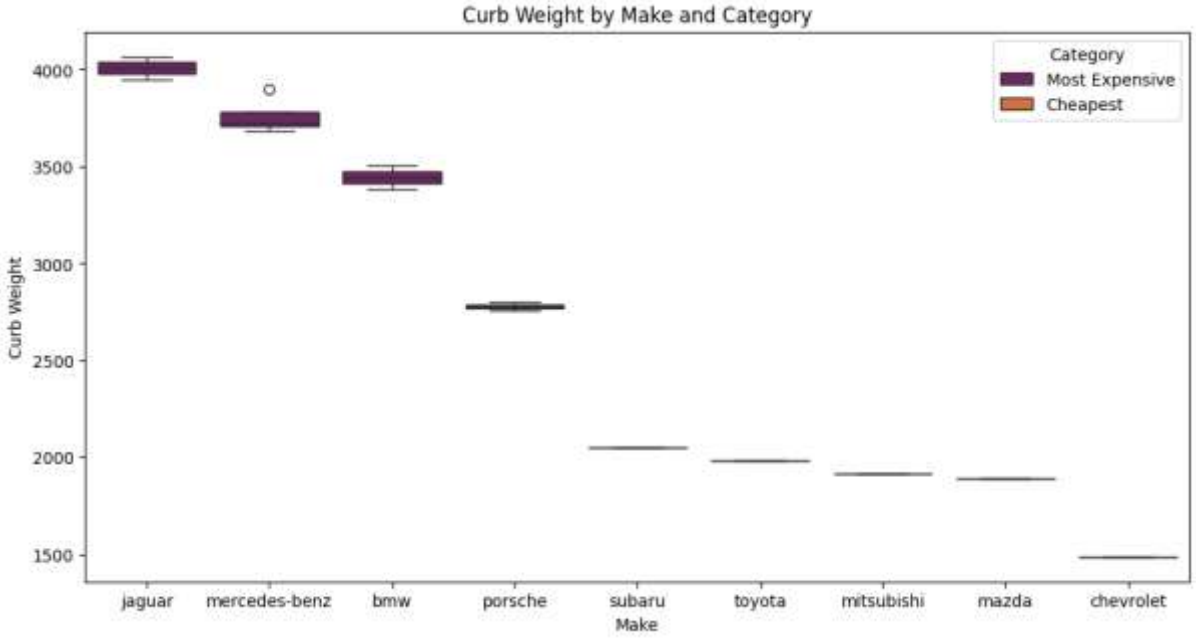
The scatterplot above illustrates the general positive relationship between engine size and horsepower. As the engine size increases, the horsepower increases as well. The size of the bubbles represents the price of the vehicles, showing that vehicles that have larger engines and higher horsepower tend to be more expensive.

The upper right section of the graph is populated by the larger, more highly priced cars, of the following makes: Jaguar and Mercedes, with Porsche being in the mid-range. These manufacturers have larger engine sizes and greater horsepower. This provides more evidence for the positive relationship between engine-size, horsepower and price. The greater the engine size and horsepower, the greater the price. This proves that higher end cars prioritise performance.

The more economical brands, occupy the lower left section of the graph. Some of these brands include: Chevrolet, Toyota, Peugeot and Mazda. While these, vehicles have smaller engine sizes and lower horsepower, they meet the consumer demand for vehicles that are affordable and fuel efficient.

Curb Weight Distribution





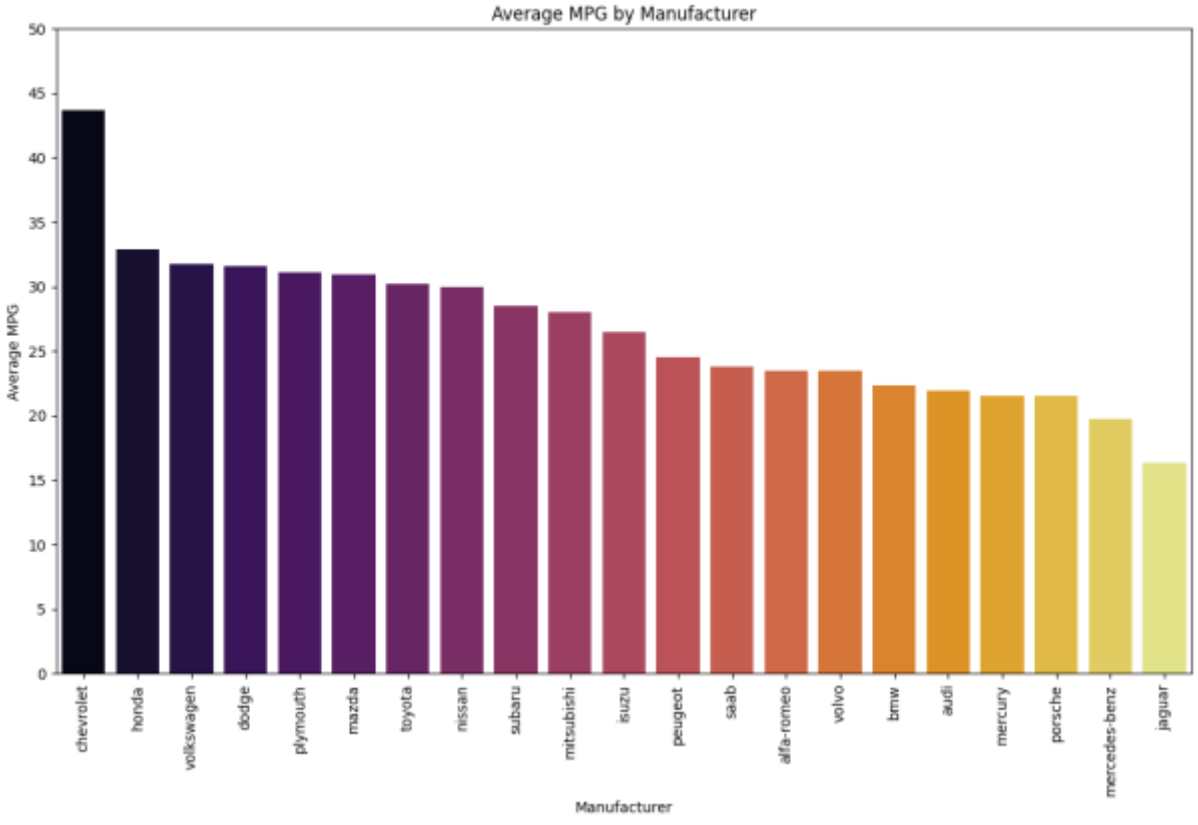
The curb weight refers to the total weight of the vehicle excluding load (passengers and cargo).

The higher-end vehicles (Jaguar, Mercedes-Benz, BMW and Porsche) have larger curb weights than the lower end vehicles (Subaru, Toyota, Mitsubishi, Mazda and Chevrolet). This finding is supported by the earlier correlation matrix – “Correlation Matrix for Engine and Fuel Features”. The correlation between engine-size and curb-weight is 0.86. Therefore, larger curb-weights can be linked to larger

engines and consequently greater horsepower. This relationship can be further explained as follows: Larger vehicles will require a stronger output, therefore a larger engine and greater horsepower, this will in turn raise the price of the car.

Jaguar, Mercedes Benz and BMW are not only the largest but also reflect variation in their data spread, compared to the narrow boxes of Subaru, Toyota, Mitsubishi, Mazda and Chevrolet. From this, it can be inferred that the higher end cars reflect a range of diverse of body-types compared to the lower end cars. It can also be inferred that the lower end cars stick to a standard lower weight across all models to maintain their fuel efficiency and cost effectiveness.

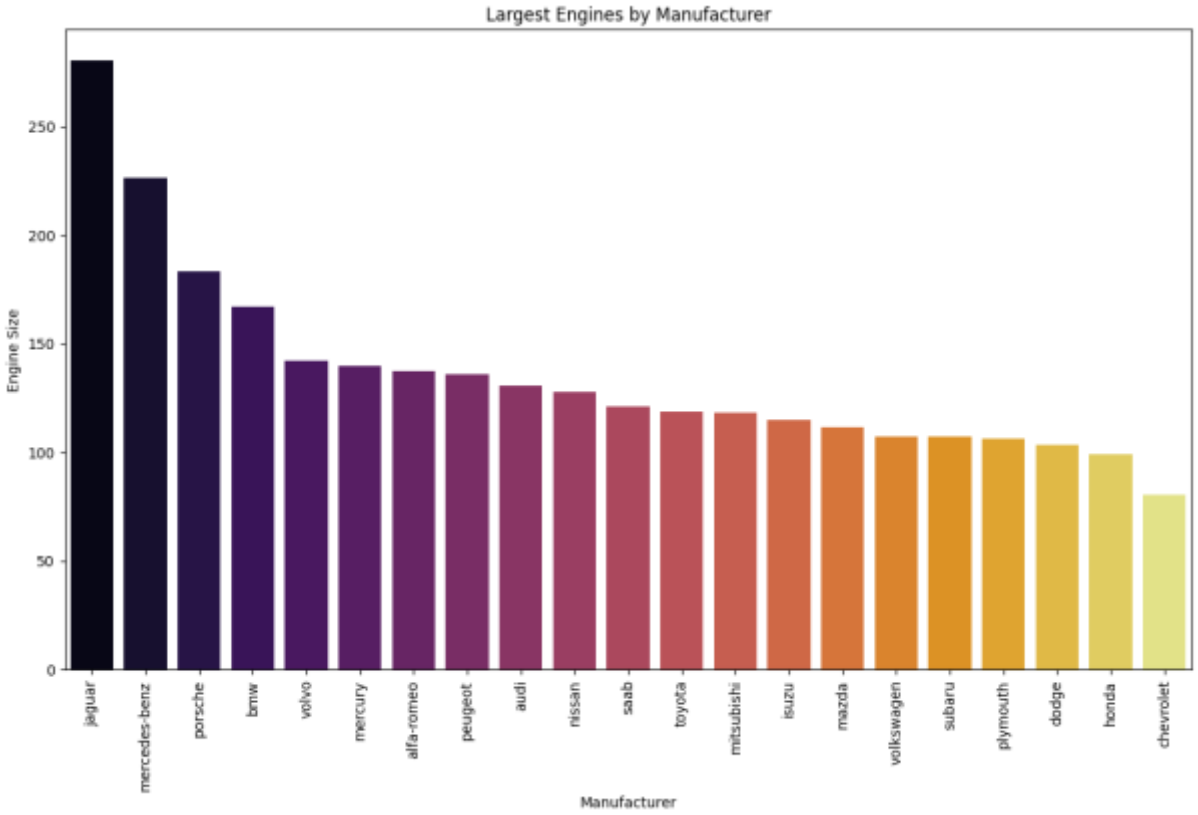
Comparison of MPG



From the graph above, Chevrolet has the highest average miles per gallon, between 40 and 45. This is consistent with previous findings as Chevrolet had the smallest engine size, lowest horsepower and therefore the highest average miles per gallon, making it the most fuel efficient manufacturer.

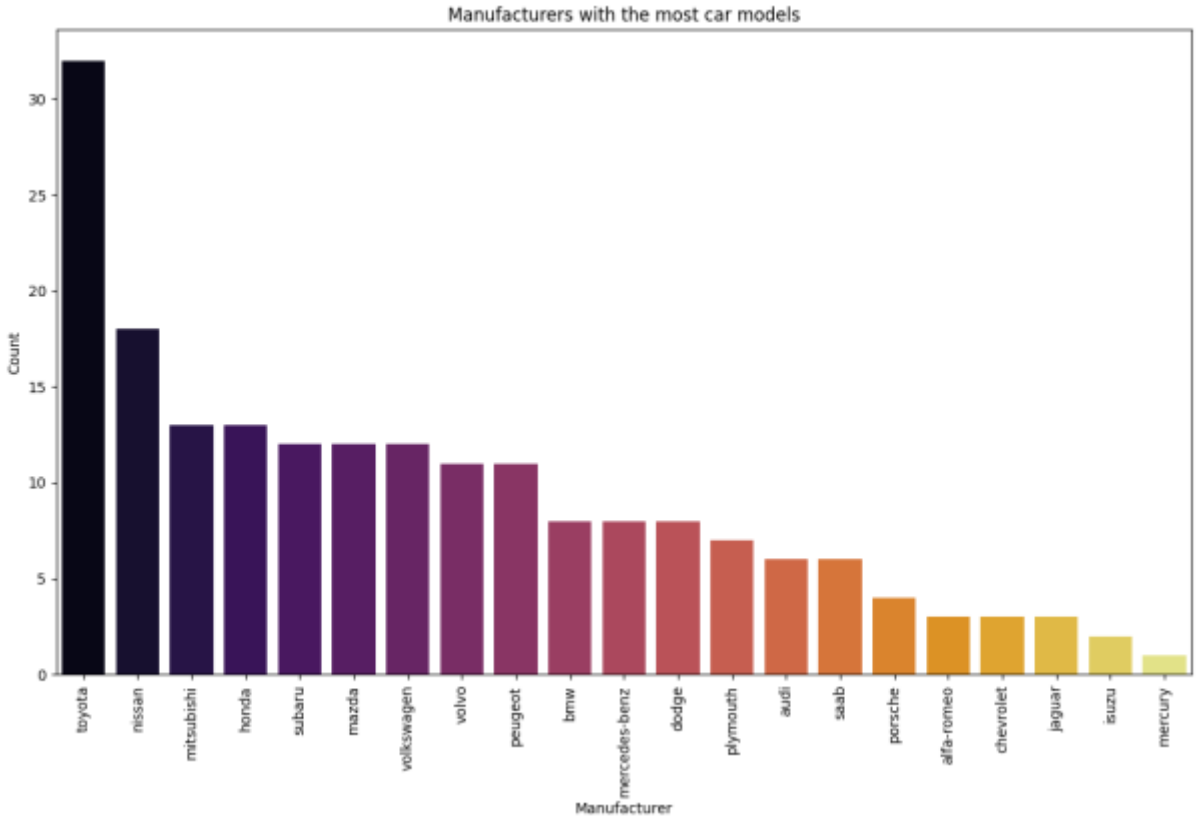
This further reinforces the notion that smaller cars, less powerful cars have an increased fuel efficiency.

Vehicles with the Largest Engine



From the graph above, Jaguar has the largest engine size, greater than 250. This is consistent with previous findings as Jaguar was presented the largest horsepower, the highest fuel consumption and it was priced among the more expensive cars. It can be concluded that Jaguar, like other luxury brands, prioritises high performance vehicles.

Most Common Manufacturer



The manufacturer that appeared the most frequently in the dataset was Toyota. This suggests that Toyota has the widest range of car models therefore the brand prioritises consumer accessibility and variation.

Conclusion

Upon analysis of the data, it can be concluded that affordability linked to practicality and efficiency, while luxury vehicles trade fuel efficiency for size and horsepower. Manufacturers are aligned themselves with their target markets. Brands that are cost effective are focused on efficiency and accessibility, while luxury brands are focused on performance and exclusivity. There is not clear "superior" option when choosing a car. The decision depends on the features and price point an individual is looking for.

This report was written by: Leah Govender