Hyperiondev

TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Introduction

Summary of the data set

The ‘Automobile.txt’ file is a text file that contains information about different makes of vehicles or automobiles. Each row in the file represents a single vehicle, and the columns of the file contain various pieces of information about the vehicle, such as the make, engine size and curb-weight.

There is a total of unique 26No. data columns and 205No. uncleansed rows.

DATA CLEANING & MISSING DATA

The following methods have been used in cleansing the dataset, where appropriate visualisations are included.

1) Head() – To show data rows and columns, gaining an appreciation of value data types.
2) Info() – To show a summary of the dataframe, including the number of rows and columns, the column names, and datatypes.
3) Missingno.matrix() visualisation – From observation, some of the rows have ‘?’ values which implies information is missing. 41No. ‘0’ values were changed to null values using replace().

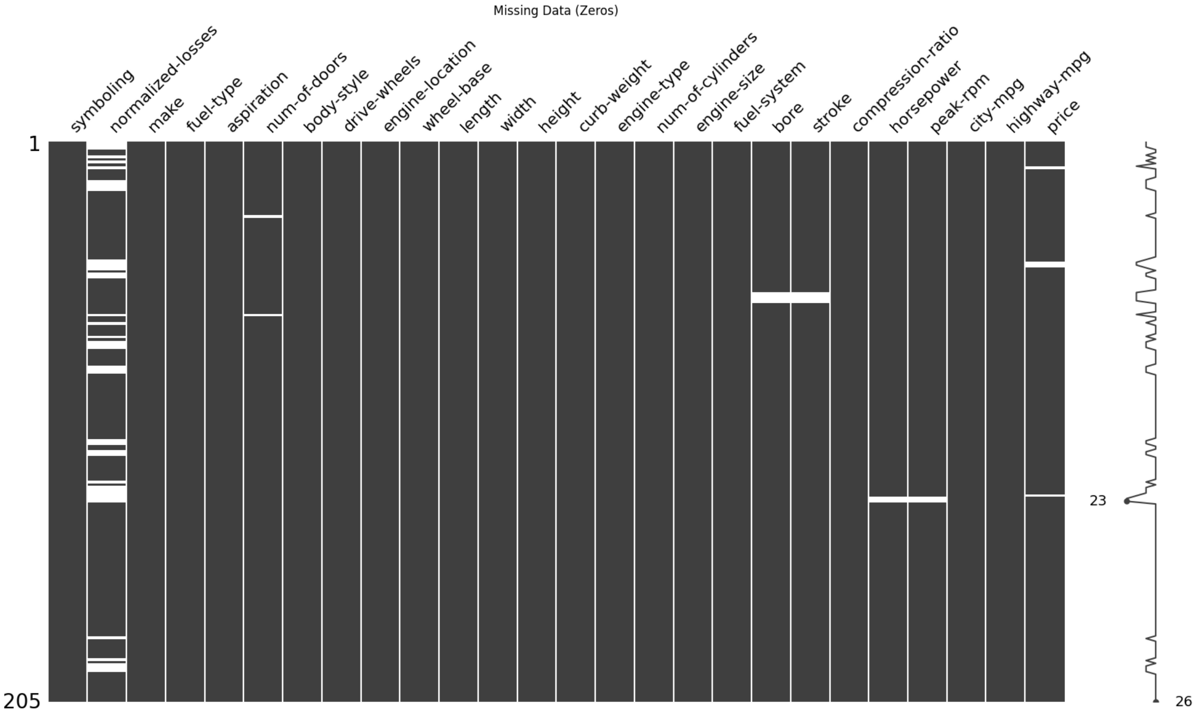

Figure 1: Missingno Matrix (null values)

4) Isnull() and Drop() – Knowing where any null values are and what our analysis will include allows for the removal of unnecessary data columns. Removing unnecessary data columns simplifies the dataframe. The following data 8No. columns were removed:

    a. Normalized-losses – dropped.

5) Drop_duplicates() – Given that each row is a separate vehicle, we do not want to duplicate data that may skew any assessment. Duplicate rows were removed. For this assessment, the dataset included no duplicates.
6) Astype() – To allow for full assessment, price, horsepower, peak-rpm, stroke, bore and engine-size data values cast from object class to integer.
7) New columns – “avg-mpg” , “body-volume” and “density” added as new columns to main dataframe. Body-volume and density not representative of individual vehicle but should allow for condensation of multiple data columns to one single variable to be used in the analysis.
8) Minmax scaling() – Due to differences in magnitude and distribution, continuous data columns have been scaled to compare data with different magnitudes and ranges so that one feature does not dominate the other feature. These scaled data columns are added to the main dataframe.

DATA STORIES AND VISUALISATIONS

Exploratory data analysis has been used to explore various aspects of the automotive industry such as what makes a car more fuel efficient, what increases the power of a car’s engine and what features influence a cars price. Different visualisations have been used to easier understand and interpret what the data is representing.

Assumptions

1) The given data set contains no unit information, therefore for assessment comments and visualisations presented will be unitless. A higher value is assumed as being good or bad concerning each data column title.

Patterns and Trends

1) Describe() & Data Distribution with Histograms – To show basic statistics of the numerical columns in the dataframe and review frequency distribution. From this we can review which numerical data is being included and look at the shape of the data included in the dataset.

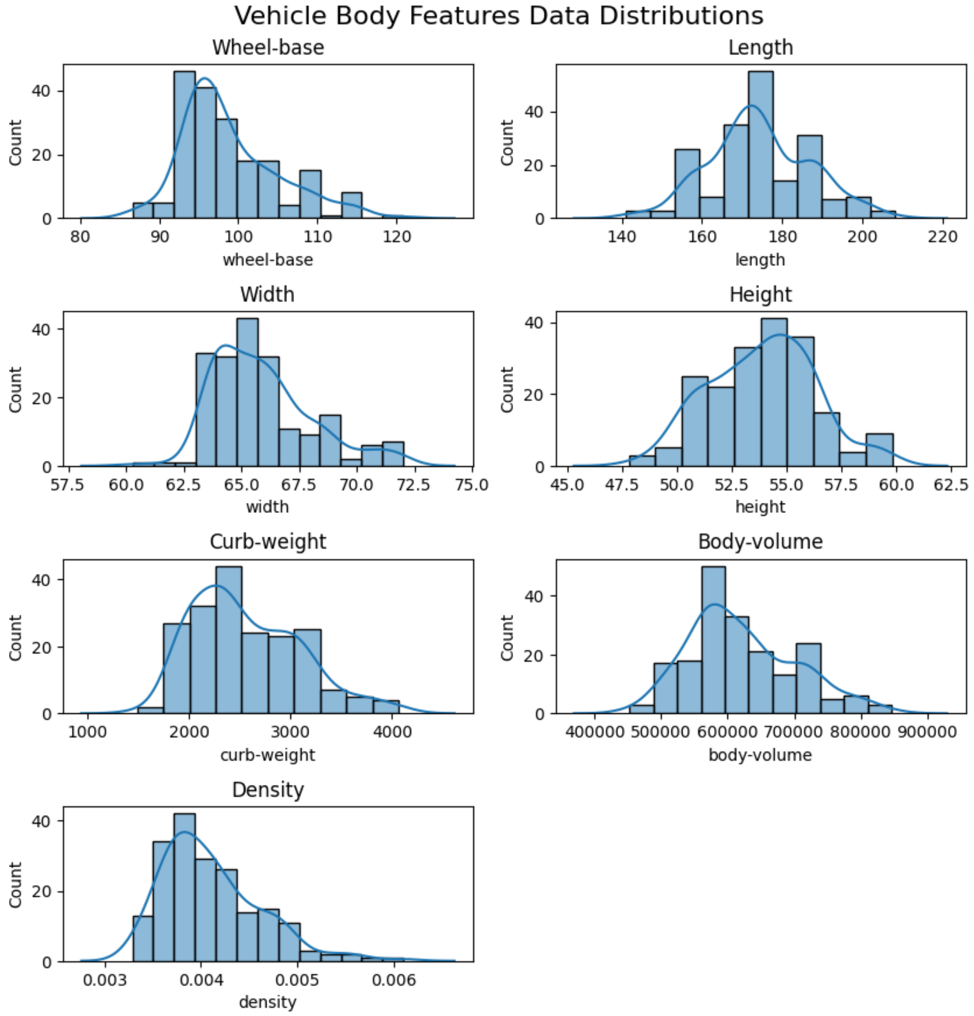

Figure 2: Seaborn Vehicle Features Histplots Subplots

- The positively skewed distribution of wheelbase, width, curb-weight, body-volume, and density suggests that most vehicles have lower values for these features, with fewer manufacturer's vehicles have higher values. This can be interpreted as most vehicles have a smaller of wheelbase, width, curb-weight, body-volume, and density, and it is less common or normal to have vehicles with higher feature values.
- Length and height appear normally distributed, with most data at the bells centre with fewer data points at the extreme left and right sides. This tells us that in terms of length and height , scoring is independent and less likely to be biased. This suggests that other factors set may set these decisions such as regulatory requires or engineering limitations with specific manufacturers/makes.

2) Scatterplot matrices were used to quickly review and visualise the relationships between general vehicle features in the dataset. The following observations were made comparing avg-mpg, horsepower, engine-size, curb-weight, fuel-type, and price.

- The data suggests a negative relationship between avg-mpg and horsepower, engine-size, and curb-weight. This means that an increase in these factors tends to reduce the avg-mpg that can be achieved. The relationship appears exponential tending to zero.
- A higher priced vehicle tends to also be less fuel-efficient.
- Diesel vehicles appear more fuel efficient compared to gas-fuelled vehicles.
- There is a positive correlation between horsepower, engine-size, and curb-weight, indicating that more horsepower typically means a larger engine-size and heavier vehicle.
- The data also shows that vehicle curb-weight is strongly influenced by the size of its engine.
- The plots of price against engine-size and curb-weight reinforces that horsepower is the largest influence on price, rather than avg-mpg.

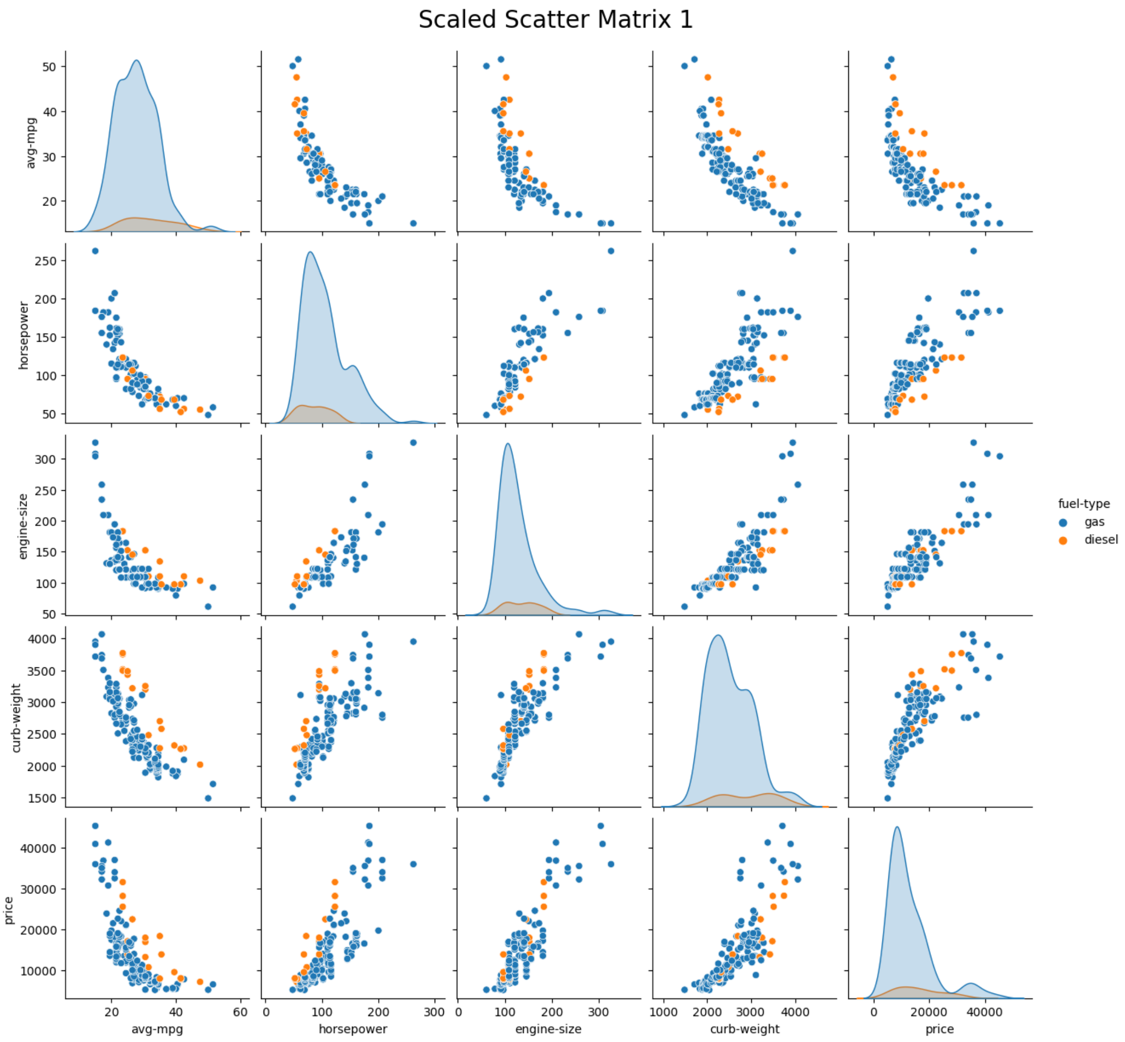

Figure 3: Scatterplot Matrix Scaled Features 1

3) Histplot subplots to present City vs. Highway MPG.

- City MPG and Highway MPG have similar distribution shapes, both are weakly positively skewed with most data points at the centre and fewer data points at the extreme ends. The longer right tail suggests that some vehicles are slightly more favourable of being fuel efficient in a city environment.
- However, when comparing the two, Highway MPG is more favourable with a mean and median of around 30MPG compared to 25MPG for City MPG. This suggests that other factors, such as driving style and sudden acceleration, play a role in determining a vehicle's MPG. Understanding that external factors may contribute to less fuel efficiency can explain the negative correlation between vehicle pricing and average-mpg.

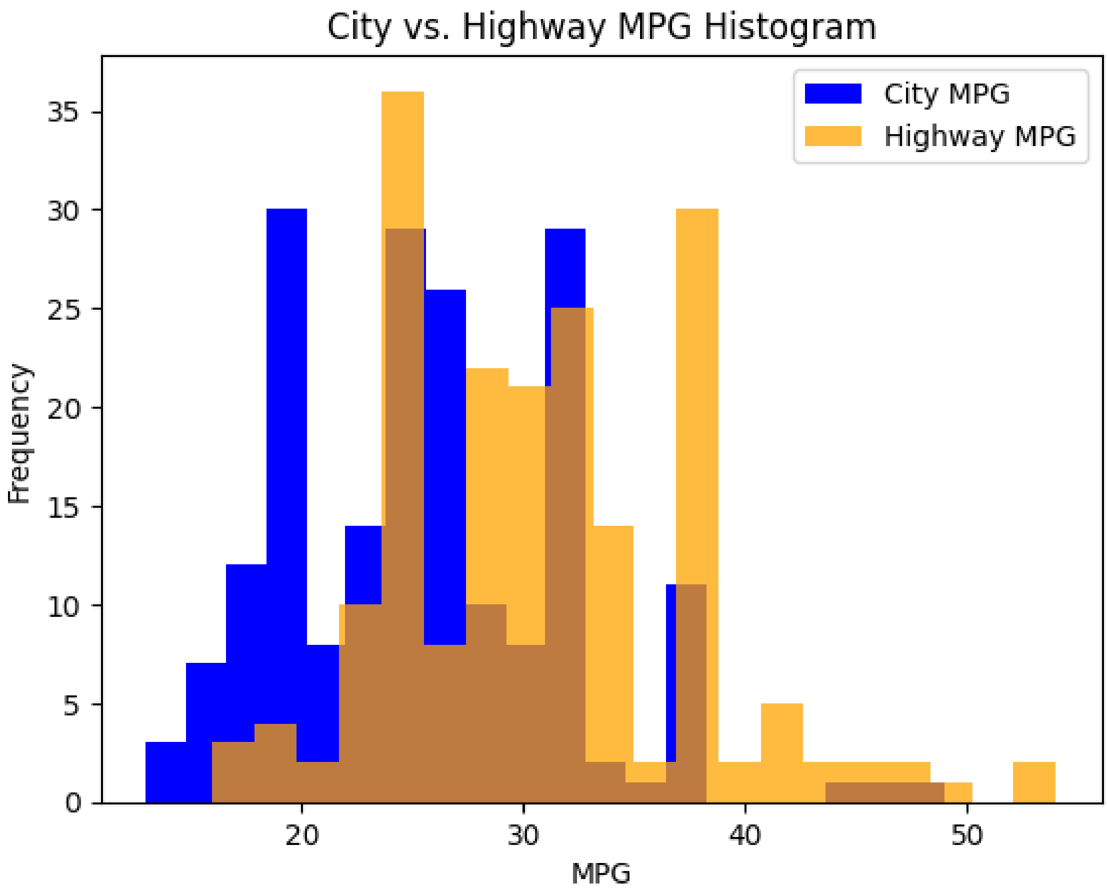

Figure 4: Histplot subplots Fuel Efficiencies

4) Scatterplot matrices were used to quickly review and visualise the relationships between vehicle engine features in the dataset. The following observations were made comparing avg-mpg, horsepower, engine-size, curb-weight, fuel-type, and price.

- The data shows a positive correlation between engine-size and cylinder bore, as both relate to engine volume.
- There is no correlation between engine-size and stroke.
- Higher compression ratios are found in diesel-fuelled cars.
- There is a weak correlation between engine-size and peak-rpm, with larger engine-size tending to have lower peak-rpm, however, there are not enough data points to confirm the strength of this trend. Diesel vehicles typically operate at a lower peak-rpm compared to gas.
- There appears to be no correlations between peak-rpm and bore, stroke, compression ratio, and horsepower.
- There is a weak inverse exponential correlation between horsepower and bore size.

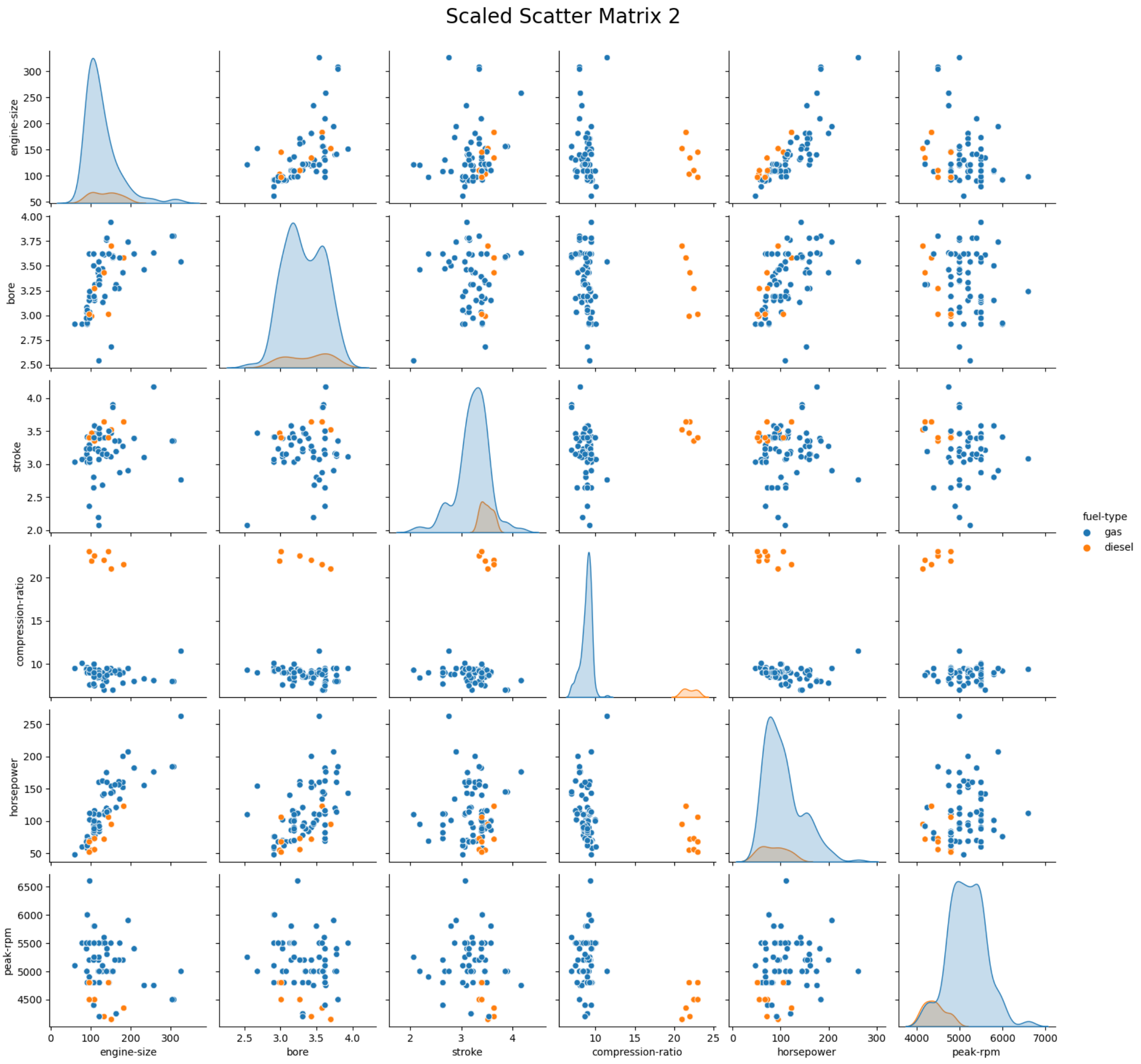

Figure 5: Scatterplot Matrix Scaled Features 2

5) A deeper look and analysis into engine features and the relationships between the number of engine cylinders, horsepower, engine-type and average mpg by aspirations by boxplot show the following.

- Fewer cylinders tend to result in higher average MPG but lower horsepower. Specifically, when comparing turbocharged engines to standard aspiration engines, the data shows that turbocharging significantly increases horsepower and decreases average MPG in 4-cylinder engines. However, in 6-cylinder engines, turbocharging improves average mpg, with only marginal losses in horsepower.
- Regardless of fuel-type turbocharging increased horsepower and lowered average MPG.
- The data suggests that OHC diesel engines are the most fuel efficient among all engine types but both OHC gas and diesel engines have the least horsepower. OHC engines with standard aspiration are more fuel efficient.
- OHCV gas engines are the least fuel efficient and have the most horsepower.
  The changes in fuel efficiency and horsepower observed in the data are due to a combination of factors. Further research is needed to understand the specific technology and design differences between the various engine types, as well as the fuel and air balance in the combustion process for each engine and fuel type, and the effect of aspiration on these factors.

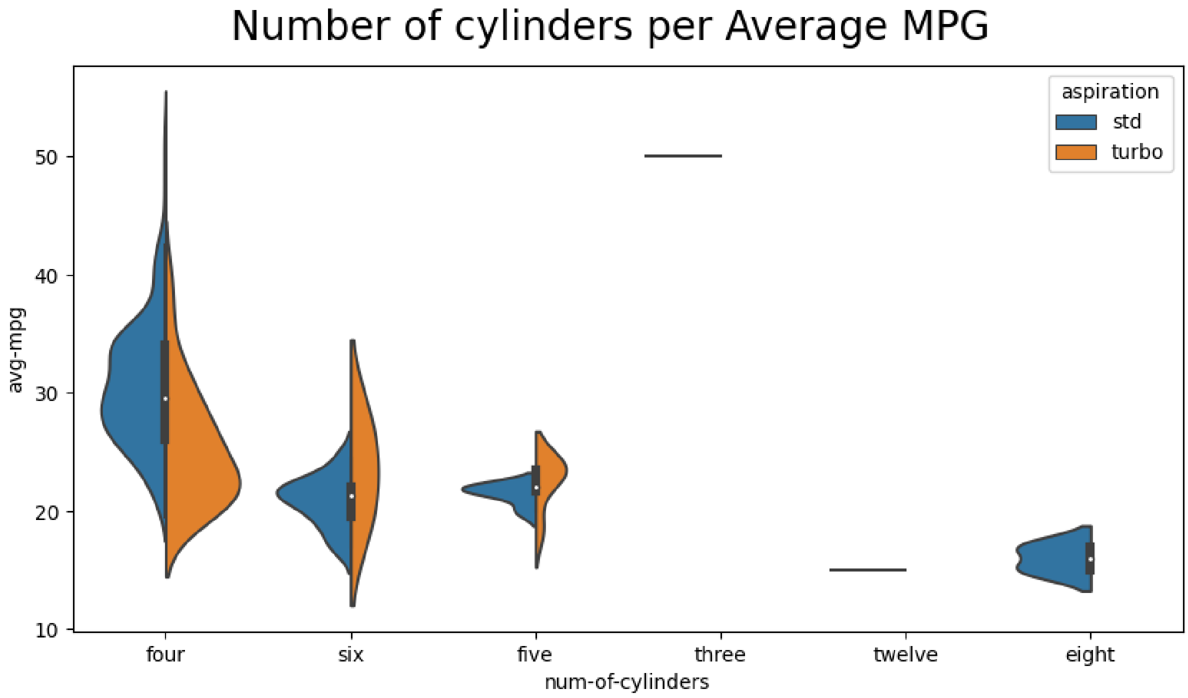

Figure 6: Boxplot Engine Fuel/Horsepower 1

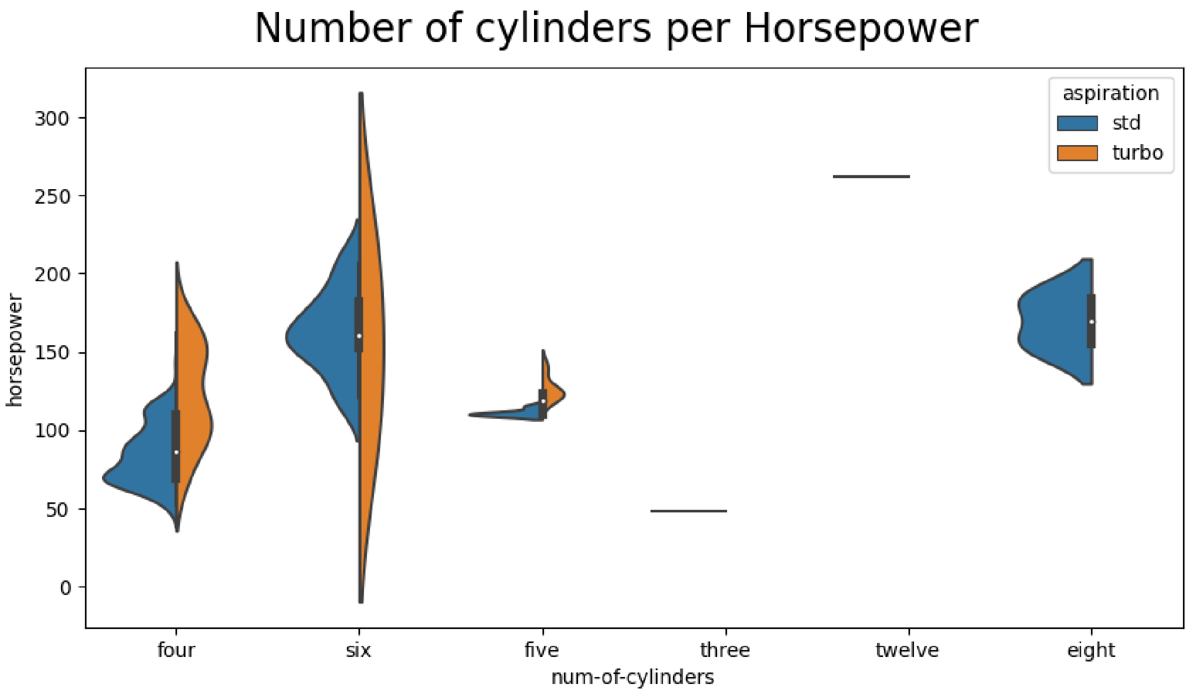

Figure 7: Boxplot Engine Fuel/Horsepower 2

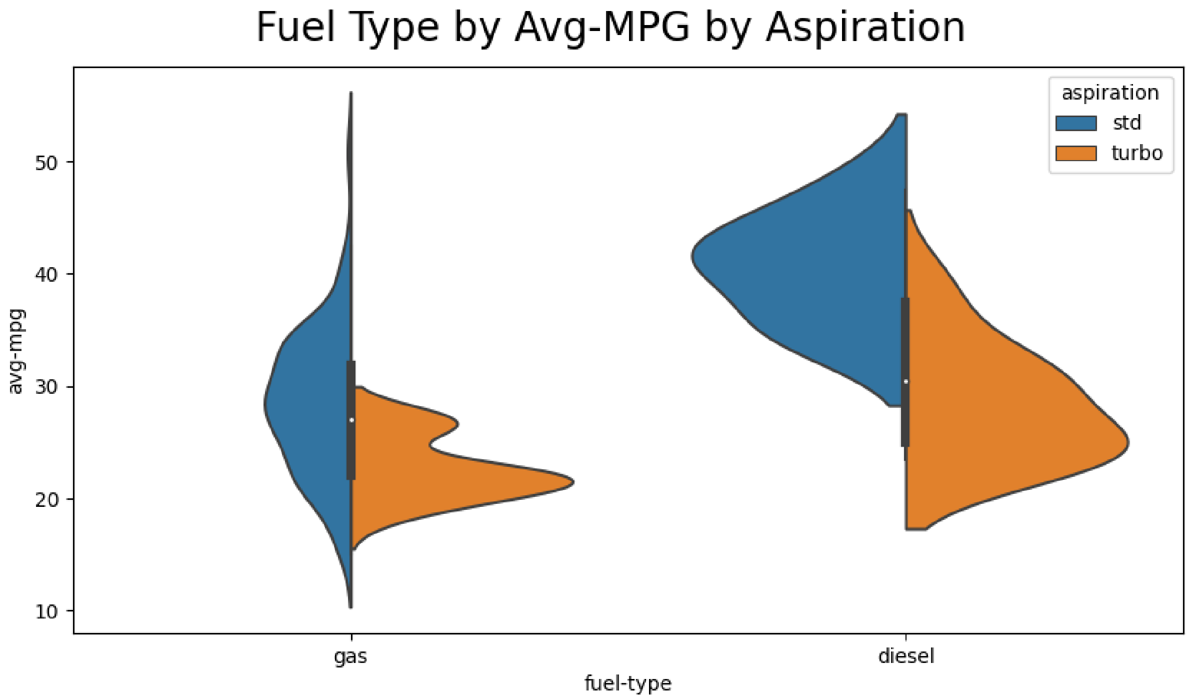

Figure 8: Boxplot Engine Fuel/Horsepower 3

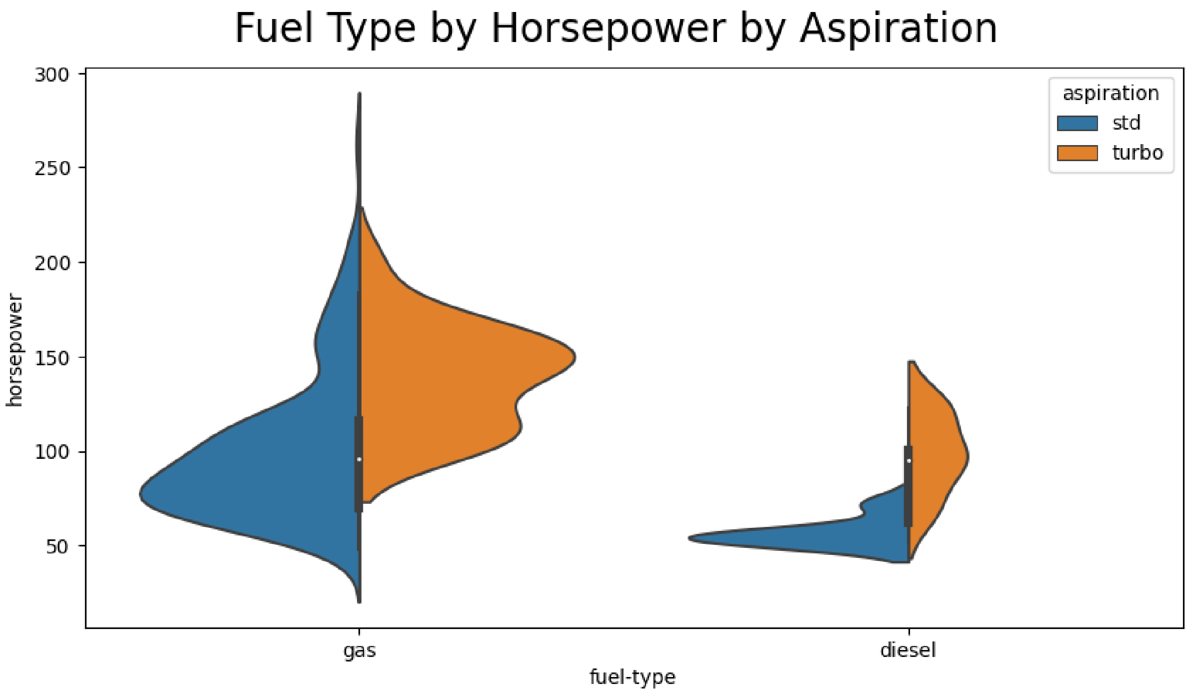

Figure 9: Boxplot Engine Fuel/Horsepower 4

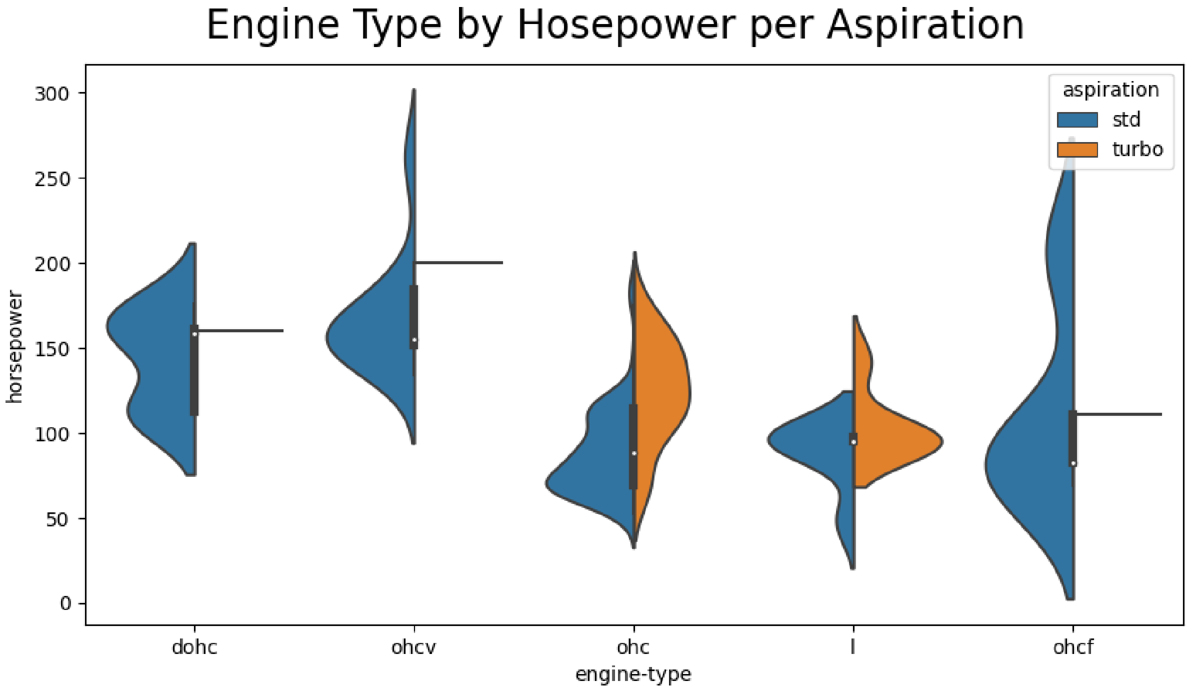

Figure 10: Boxplot Engine Fuel/Horsepower 5

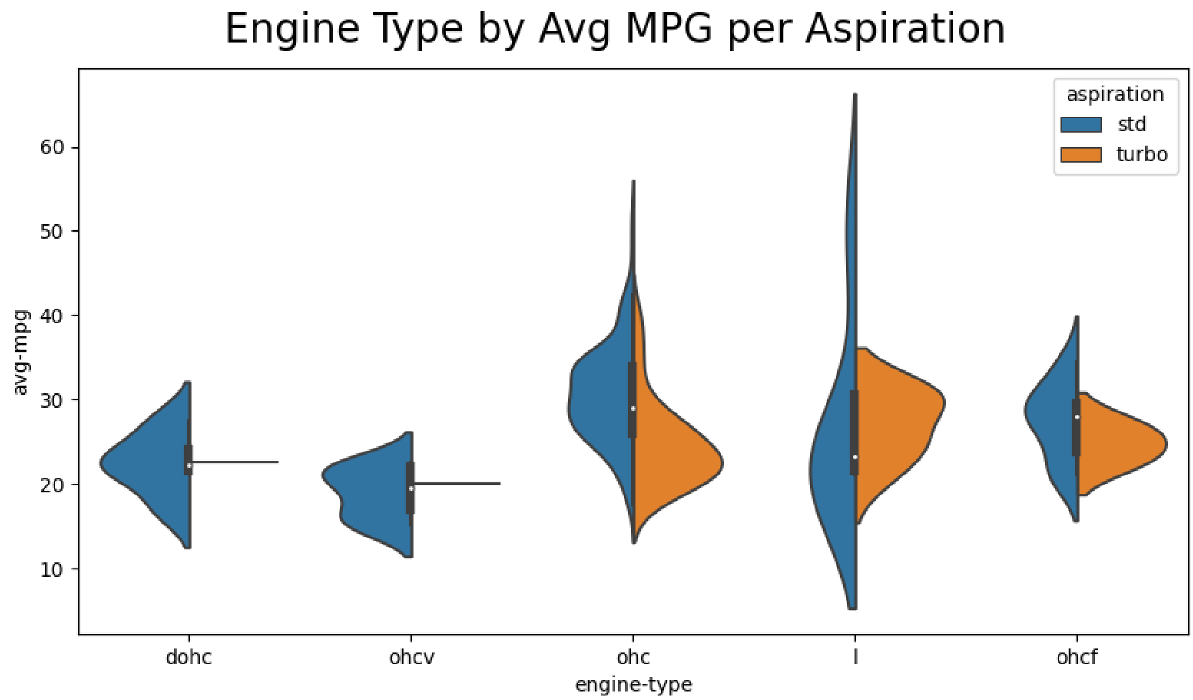

Figure 11: Boxplot Engine Fuel/Horsepower 6

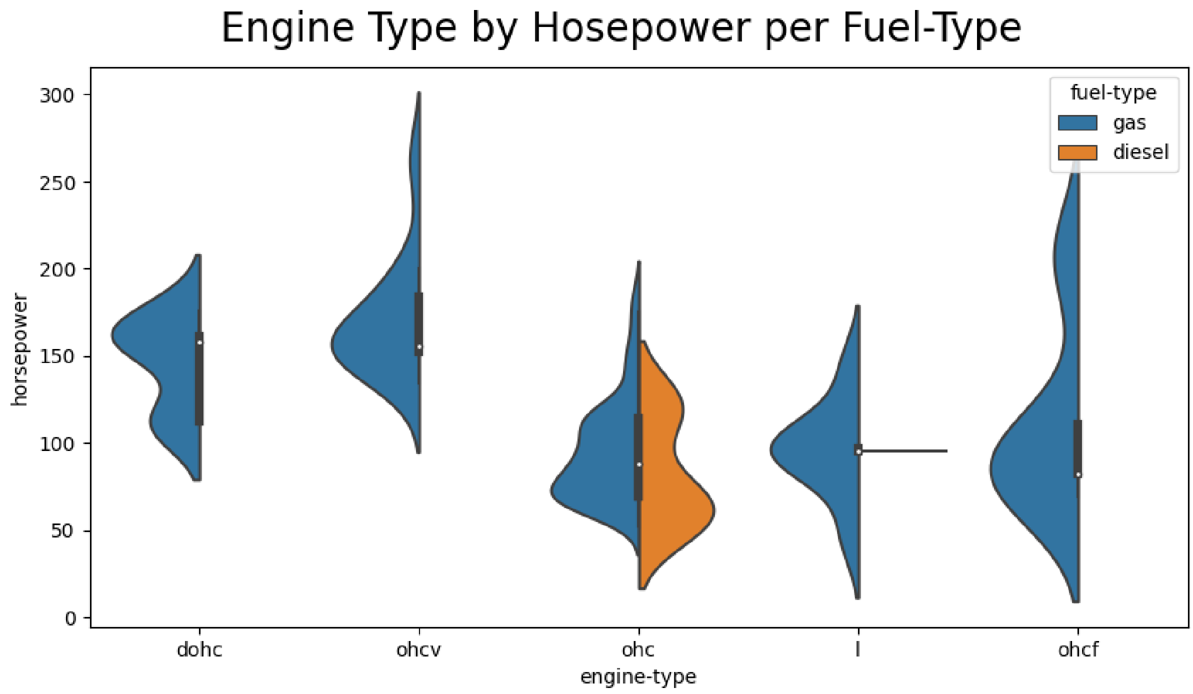

Figure 12: Boxplot Engine Fuel/Horsepower 7

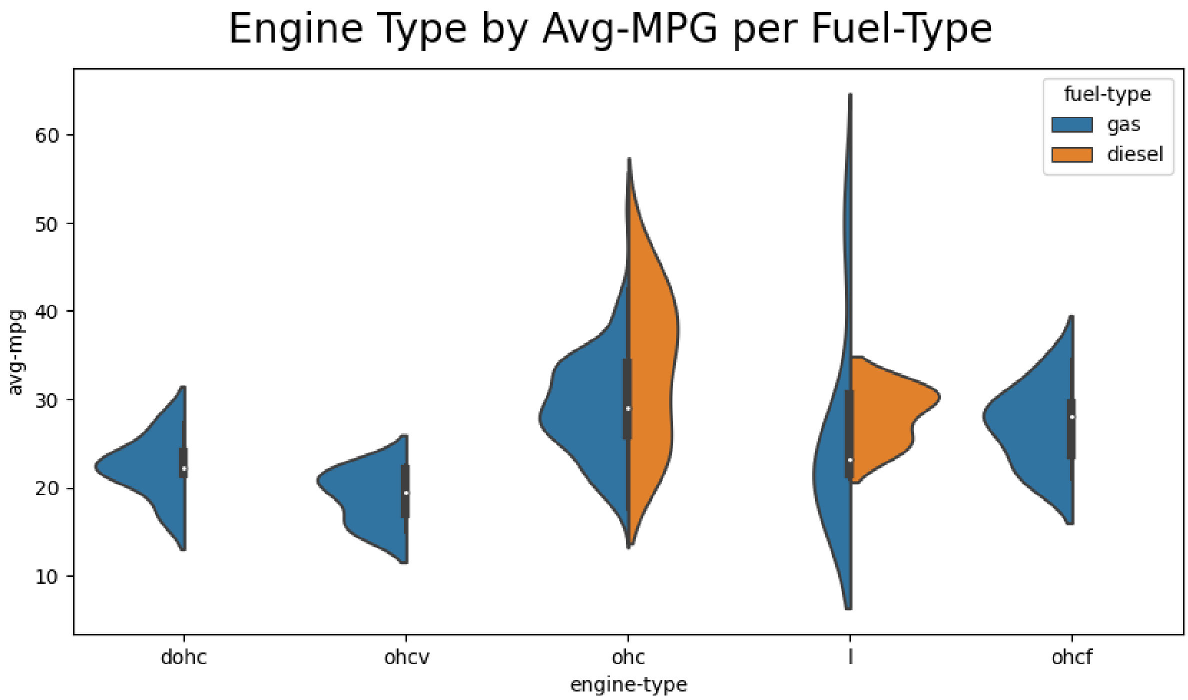

Figure 13: Boxplot Engine Fuel/Horsepower 8

6) The following plots look further into which features have the most impact on fuel efficiency.

- As seen from the scatterplot matrices, once scaled, it is clear which data features have the most impact on fuel efficiency. Figures 14 and 15 separate engine and vehicle body features.
- Engine features that have the worst impact on fuel efficiency are horsepower, engine-size, and bore (and by relationship, the number of cylinders) as seen in Figure 14. On the other hand, compression ratio is seen to improve fuel efficiency.
- Concerning vehicle body features, Figure 15 shows that length, width, and height (and by relationship, body volume) leads to poor fuel efficiency. Additionally, it is shown that curb-weight and density have the worst impact on fuel efficiency.

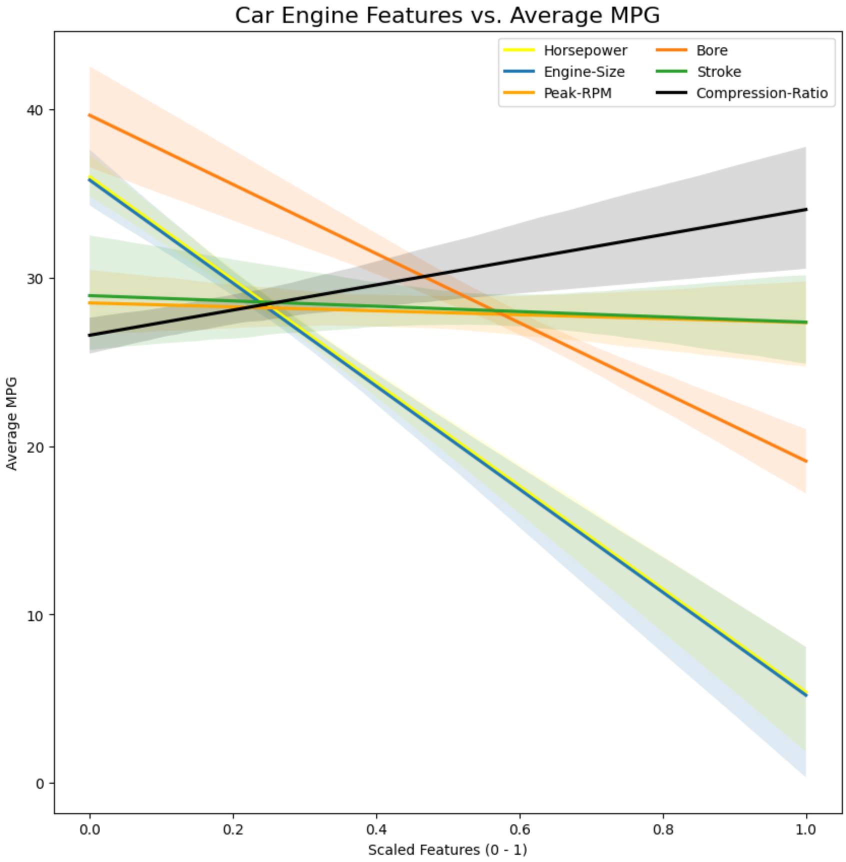

Figure 14: Regplot Engine Features by Average MPG

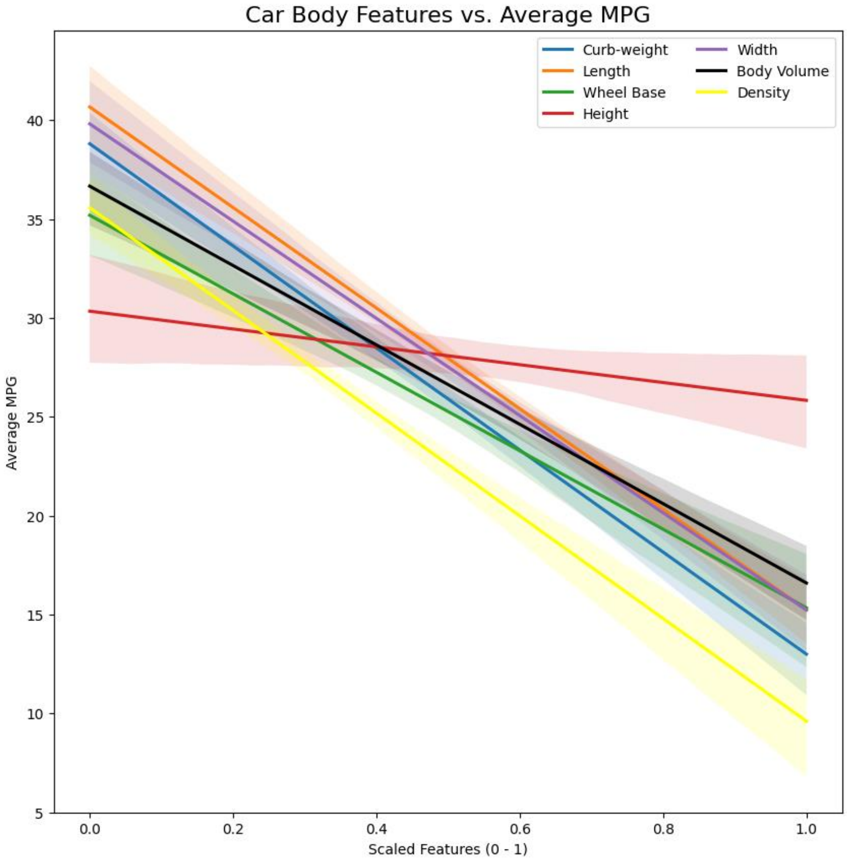

Figure 15: Regplot Vehicle Body Features by Average MPG

7) The following plots look further into how vehicles are priced and whether anything in addition to engine horsepower influences retail value.

- Figure 16 shows from this dataset that the most popular car is Toyota followed by Nissan then Mitsubishi. Mercury, Isuzu and jaguar are the least represented in the dataset.

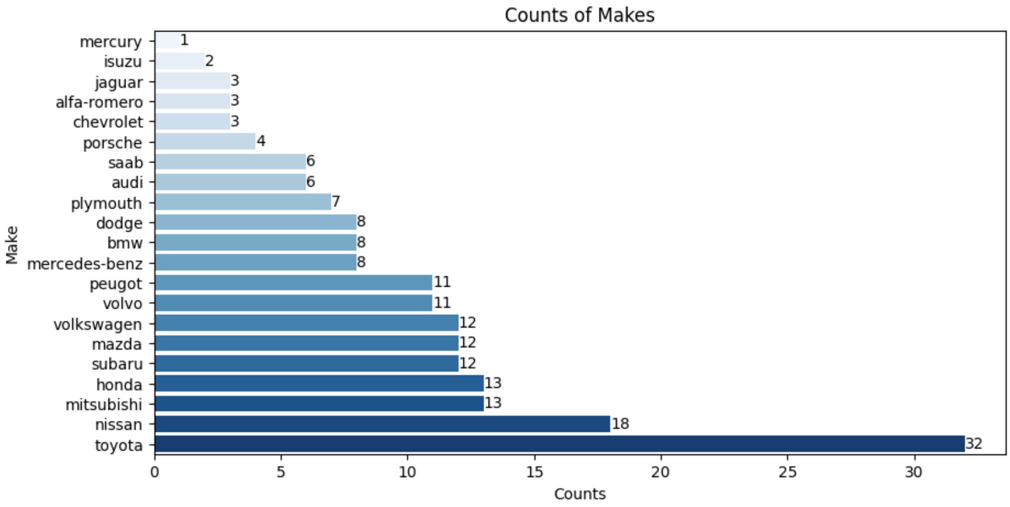

Figure 16: Manufacturers Vehicle Counts in Dataset

- Figure 17 shows from this dataset that the most expensive vehicles by manufacturer are Jaguar , Mercedes-Benz, and porches. Toyota, Nissan, and Mitsubishi are relatedly well priced by comparison.

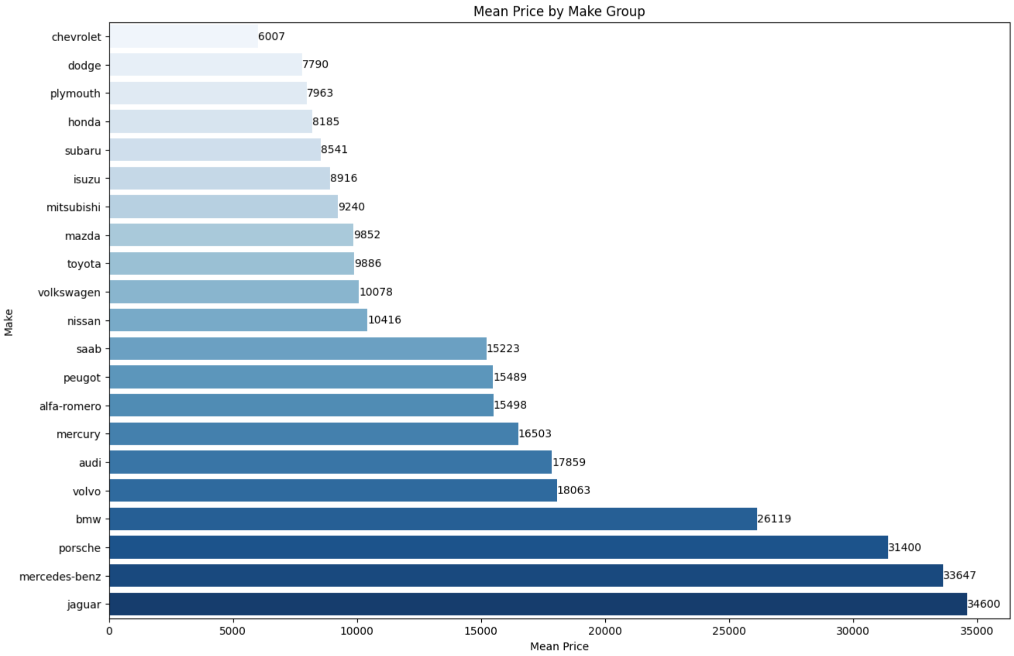

Figure 17: Manufacturers Vehicles Mean Price

- Figure 18 shows Jaguar and Merced-Benz having the highest mean curb-weight vehicles, possibly justifying their higher price point. Note of the top three most expensive makes, Porsche is not in the top three heaviest.

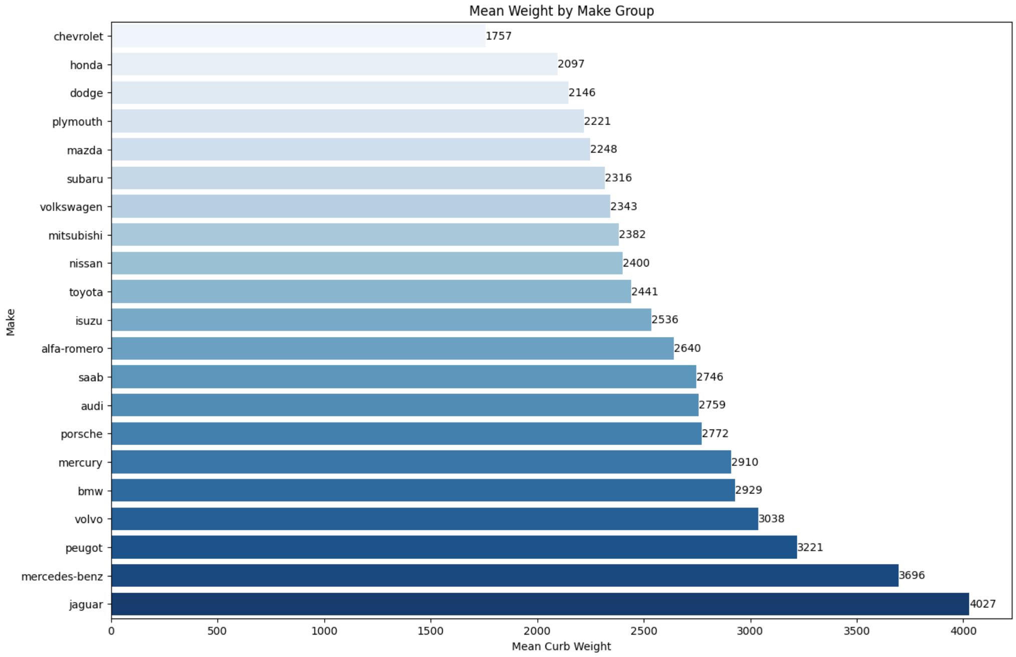

Figure 18: Manufacturers Vehicles Mean Curb Weight

- Figure 19 shows Jaguar and Porsche having the highest mean horsepower vehicles, again possibly justifying their higher price point.

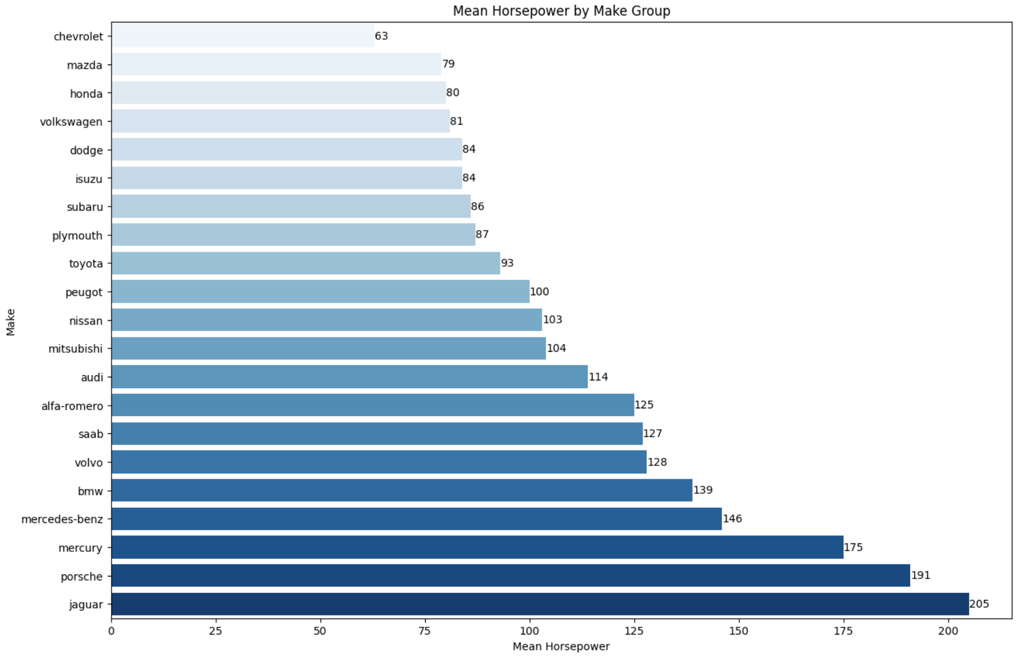

Figure 19: Manufacturers Vehicles Mean Horsepower

- Figure 20 shows Chevrolet, Hondas, Dodge and Volkswagen have the best average MPG. These also happen to be the cheapest, lightest, and least powerful vehicles in the dataset.

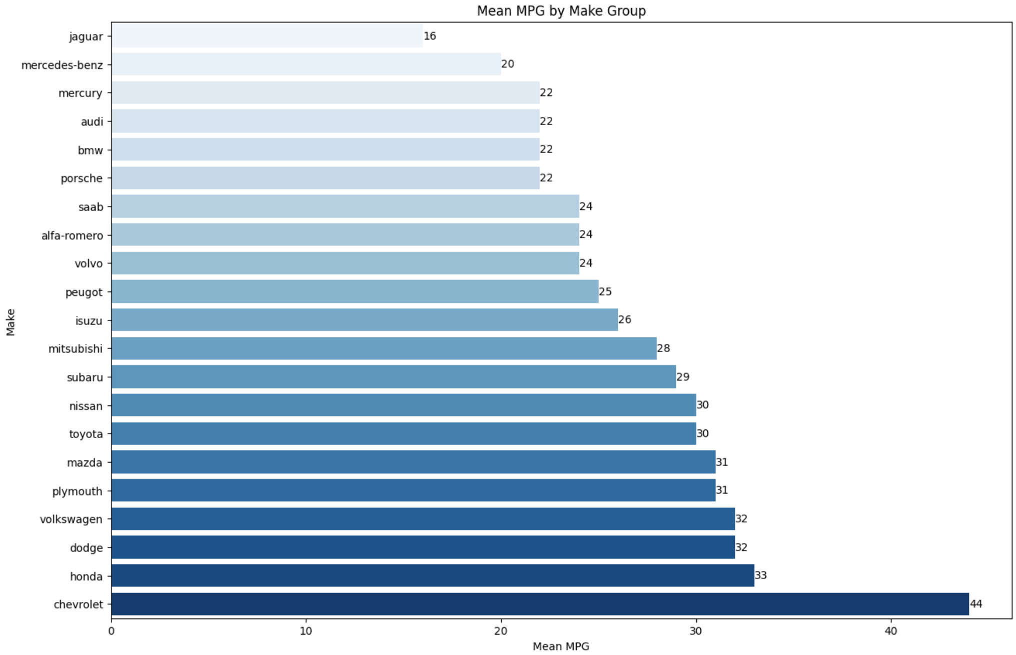

Figure 20: Manufacturers Vehicles Mean MPG

- Figure 21 shows most vehicles in this dataset are priced at the bottom of the collective range, being positively skewed.

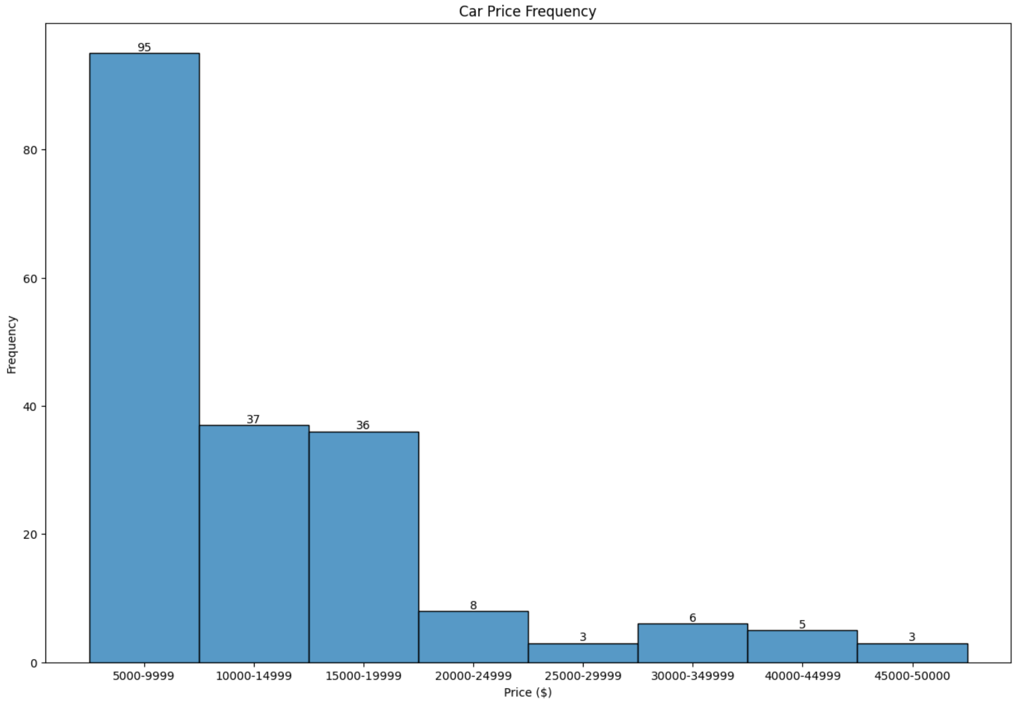

Figure 21: Vehicle Price Frequency ($5000 Bins)

- Figure 22 shows convertibles and hardtops have a higher mean price compared to hatchbacks which have the lowest mean price.

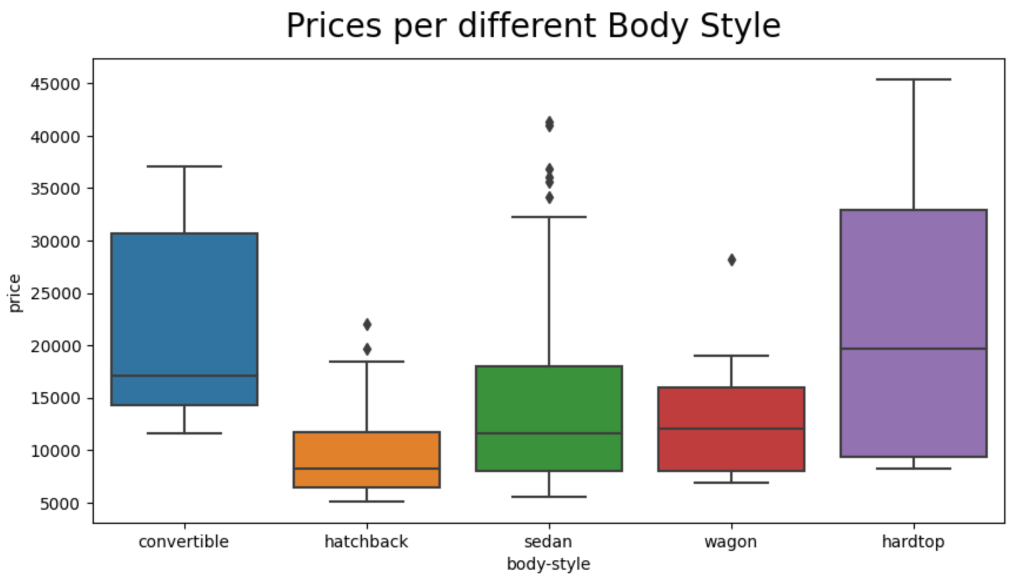

Figure 22: Boxplot Vehicle Price by Body Style

- Figure 23 shows convertibles and hardtops have similar horsepower ranges. Hardtops have a higher density of vehicles with higher horsepower. Diesels have less horsepower and belong to what would be classed as larger vehicles.

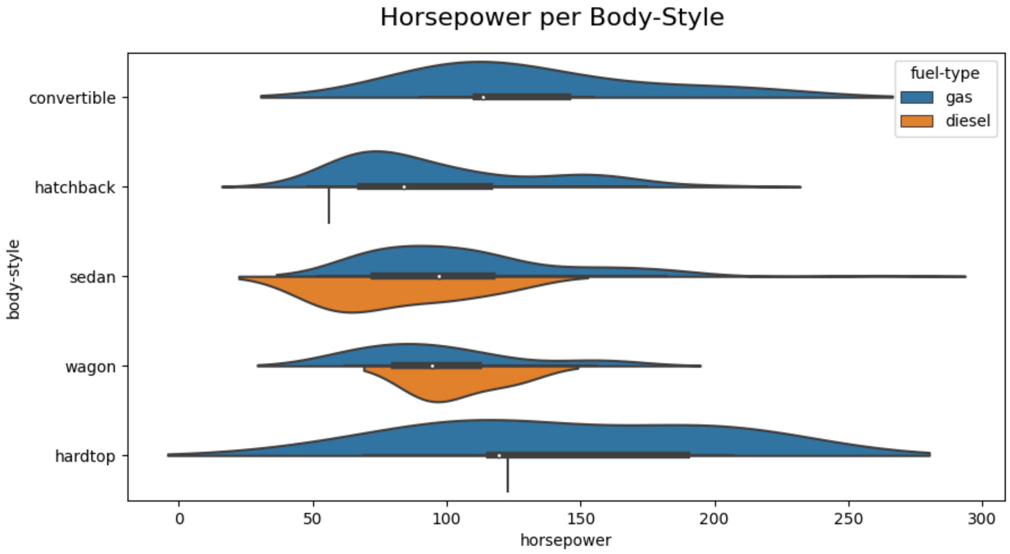

Figure 23: Violinplot Horsepower by Body-Style

- Figure 24 shows sedans and hatchbacks have better average MPG than convertibles, wagons, and hardtops.

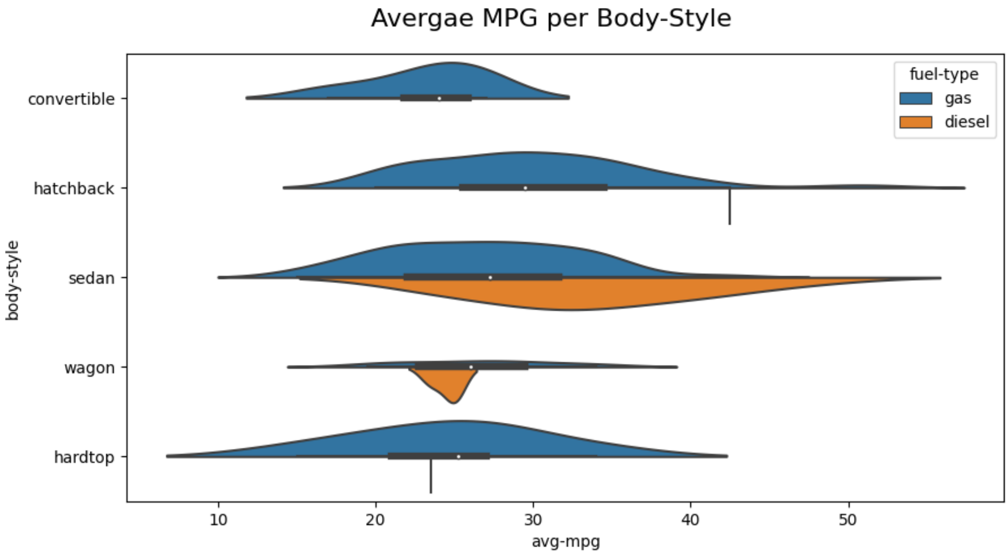

Figure 24: Violinplot Average MPG by Body-Style

- Figure 25 shows wagons and sedans having the high density of vehicles with larger curb-weight values. From Figure 23 and 24, despite Sedans have a heavier curb-weight they are typically more fuel efficient due to the lower horsepower.

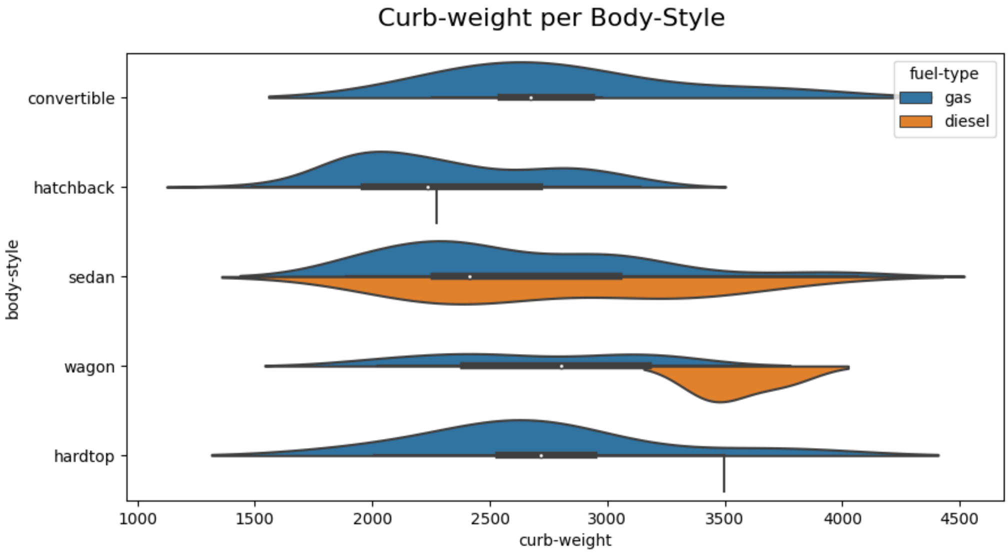

Figure 25: Violinplot Curb-Weight by Body-Style

- Figure 26 shows fwd and 4wd are of similar prices and distributions within the dataset. Whereas rwd has a much larger price range. It is also noticeable that vehicles with engines in the rear are more expensive than those at the front.

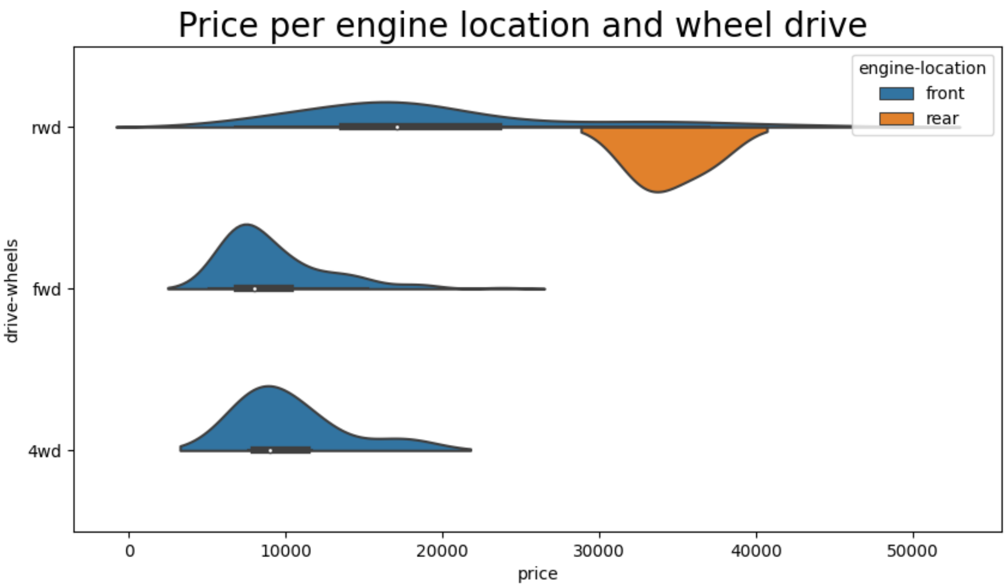

Figure 26: Violinplot Drive Wheels by Price by Engine Location

- Figure 27 shows that rwd are priced higher than fwd or 4wd vehicles as in figure 23, and diesels are similarly prices.
-

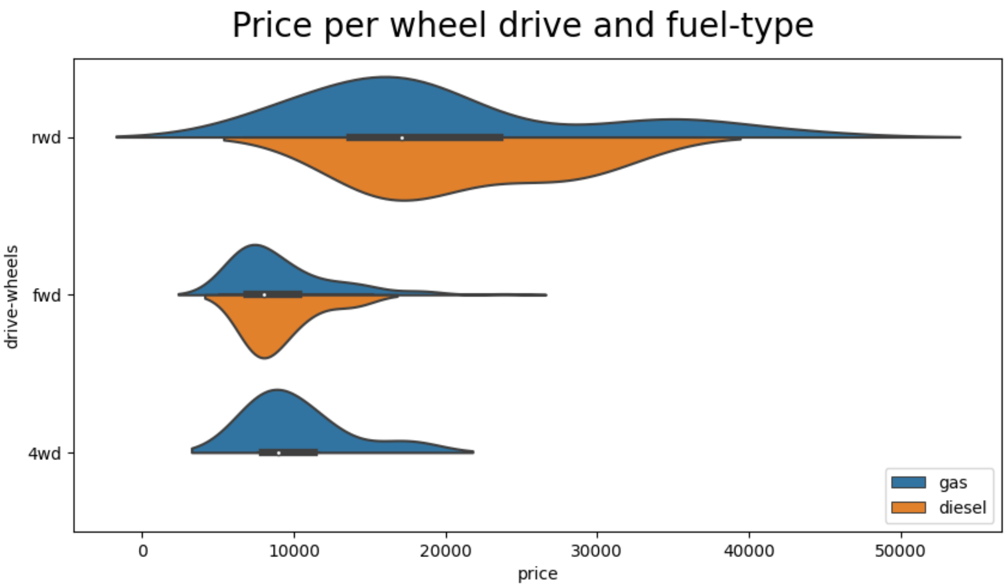

Figure 27: Violinplot Drive Wheels by Price by Aspiration

- Figure 28 siss the distribution densities of both gas and diesel vehicles with standard aspiration are similar though gas fuelled vehicle prices extend much further.

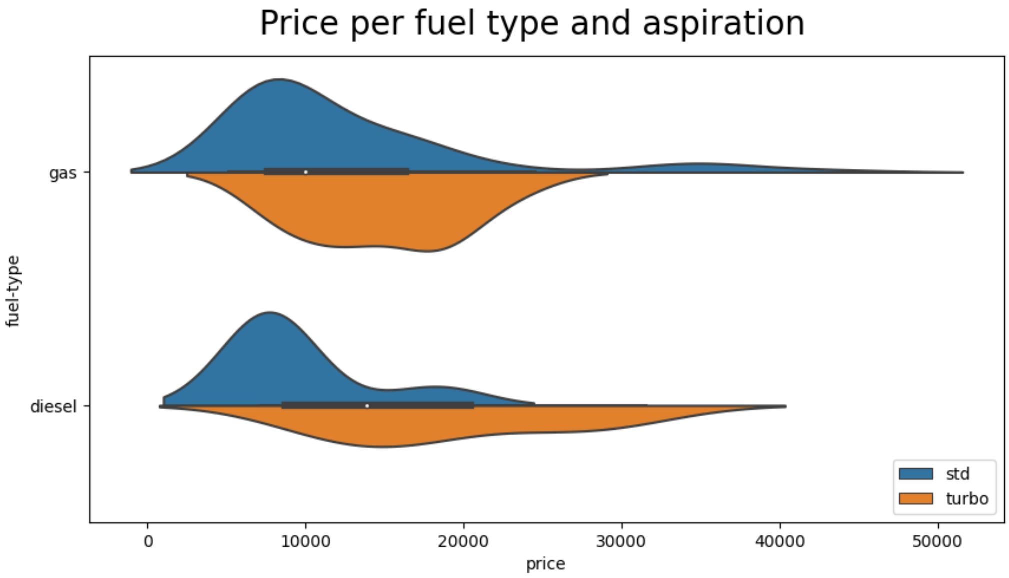

Figure 28: Violinplot Fuel Type by Price by Aspiration

Main Findings from Movies Dataset

The analysis of the given data suggests that the fuel efficiency and horsepower of vehicles are affected by a combination of factors, including the number of cylinders, engine type, aspiration, horsepower, engine size, curb weight, and compression ratio.

The data shows that fewer cylinders tend to result in higher average MPG but lower horsepower, specifically when comparing turbocharged engines to standard aspiration engines. It also shows that turbocharging increases horsepower and decreases average MPG regardless of fuel type.

OHC diesel engines are the most fuel efficient among all engine types, but both OHC gas and diesel engines have the least horsepower. OHC engines with standard aspiration are more fuel efficient. OHCV gas engines are the least fuel efficient and have the most horsepower.

Factors such as driving style and external factors may also play a role in determining a vehicle's MPG. Understanding these factors can help to improve the fuel efficiency and performance of vehicles in the future.

In terms of pricing, vehicles with higher MPG like those from Chevrolet, Honda and Dodge tend to be cheaper due to their reduced weight and horsepower. This is particularly true for hatchbacks wagons, and sedans. Conversely, manufacturers like Jaguar, Mercedes-Benz and Porsche tend to produce heavier and more powerful vehicles, which command a higher average price. Convertibles also tend to be more expensive due to their heavier weight and more powerful engines.

THIS REPORT WAS WRITTEN BY: Karl Gibson