Exploratory Data Analysis in Automobile Data Set

Introduction

The EDA for this Capstone Project was performed on the automible data set using the automible.txt file. The data has values concerning the specifics of automobiles such as the make, fuel system and other variables of interest.

Data Cleaning

In order to turn the data meaningful information, it must be cleaned. The function info() is used to get the data types of each column and so can be converted into data type that would be easier to manipulate if necessary. The drop() and drop\_duplicates() functions remove the redundant and duplicate columns respectively in the DataFrame. Some columns' data had to be cleaned individually to come up with comprehensive quantitative analysis, namely the 'price', 'horsepower’, and 'peak-rpm' columns which were variables used in compiling this report. All the mentioned columns had non-numeric values ('?'). In the 'price' and 'horsepower' data, the '?' values were then replaced with the means of the respective means and then the datatypes were changed from object to int. The non-numeric data in the 'peak-rpm' data was first converted to be null and then the datatype was changed to int. Outliers must be taken into consideration as they incorrectly skew data. An outlier in the 'horsepower' data was identified and removed.

Missing Data

There were no null fields in the data set and this determined using the isnull().sum() function. The dropna() function to drop rows with missing values, if there were any.

Data Stories and Visualisations

The quantitative data was taken into consideration as plots of the numerical data were made. Univariate histograms (Figure 1) were plotted and the results are as follows:  
- The 'engine-size' data is positively skewed and within range of 60-190 units.  
- The 'peak-rpm' data relatively normal with most values within range of 4500-5800 units.  
- The 'curb-weight' data is mostly in the range of 1900-3200 units.  
- The 'horsepower' data is postively skewed with range of 50-125 units.  
- The 'price' data is mostly in the range of 5000-18000 units.

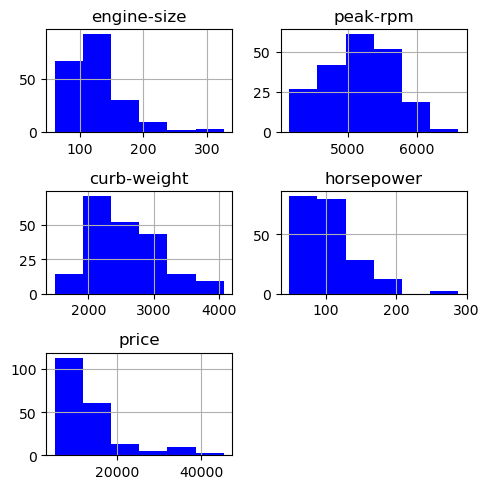


Figure 1

Bivariate analysis demonstrates the relationships between and two given variables. Pairplots (Figure 2) were made using the above data and the different pairwise combinations demonstrate positive linear relationships, and this corroborates with the figures in the correlation matrix (Figure 3). Particularly, the correlation coefficient between price and engine size is 0.86 as the engine size strongly influences the price of a car as bigger engines are more costly to manufacture. The visualisations which were outputted showing the relationships between the engine size and curb weight, suggesting a positive relationship strong correlation of 0.85 coinciding with the fact that the amount of power, determined by engine size, that an automobile has is relative to it curb weight.

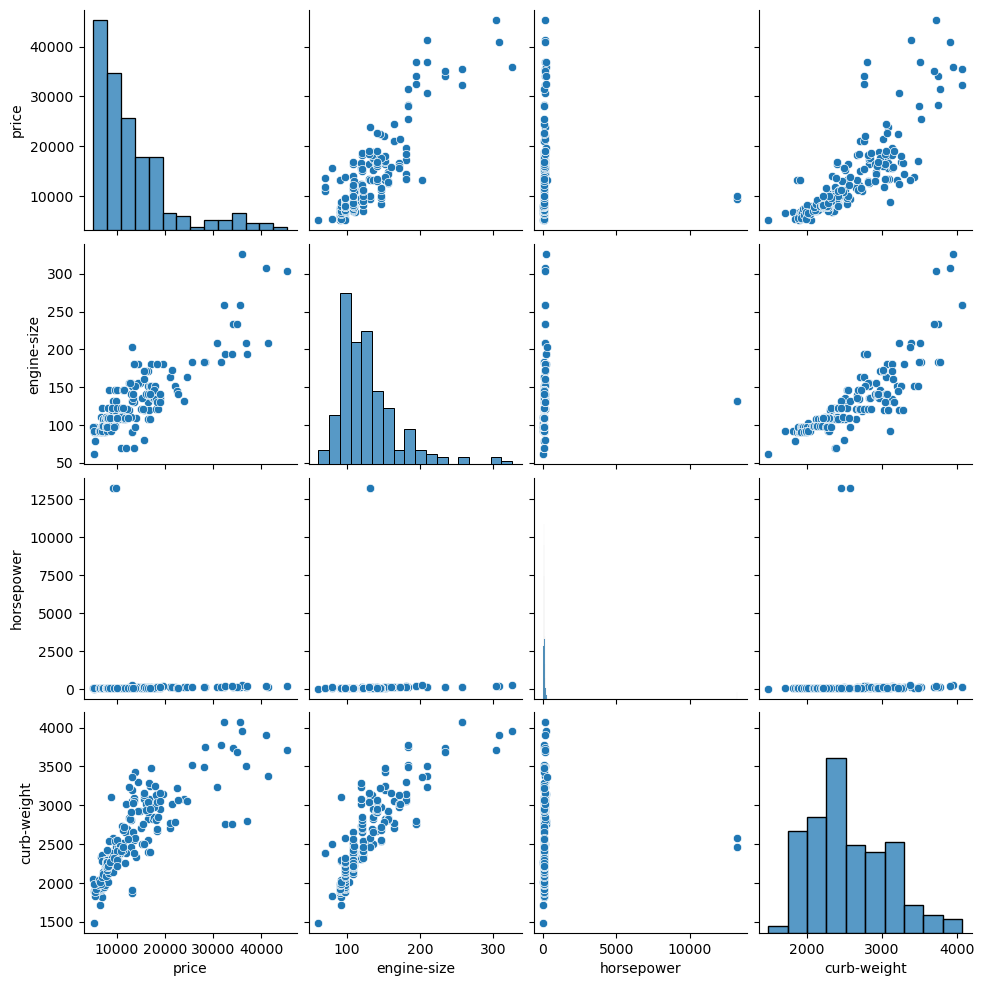


Figure 2

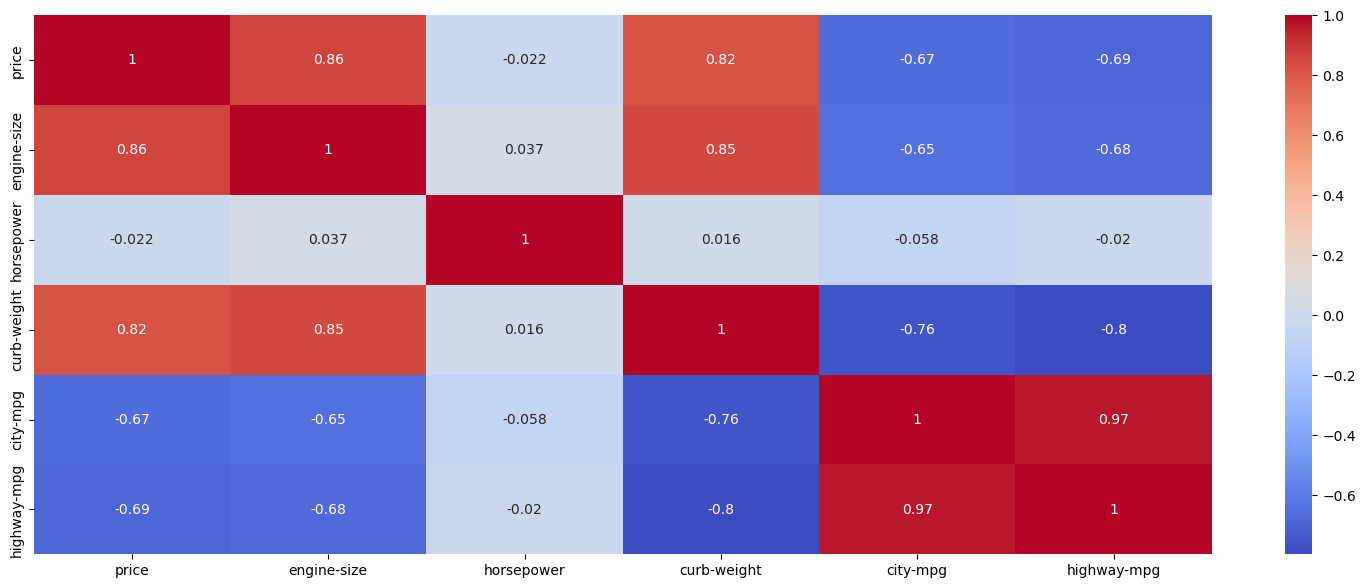


Figure 3

The qualitative data was also analysed. Univariate bar graphs (Figure 4) were plotted of the columns: 'engine-type', 'body-style' and 'fuel-type' which contain categorical data. Most of the automobiles (70% <) have the OHC engine type, which is most common due to better fuel efficiency and power putput. The sedan was the most popular body style with approximately 50%. Gas (~90%) of vehicles used gas instead of diesel as fuel. This may be because the average price of gas cars (12922.69) is lower than that of diesel cars (15838.15). The comparative box plots (Figure 5) show that diesel cars are overall costlier.

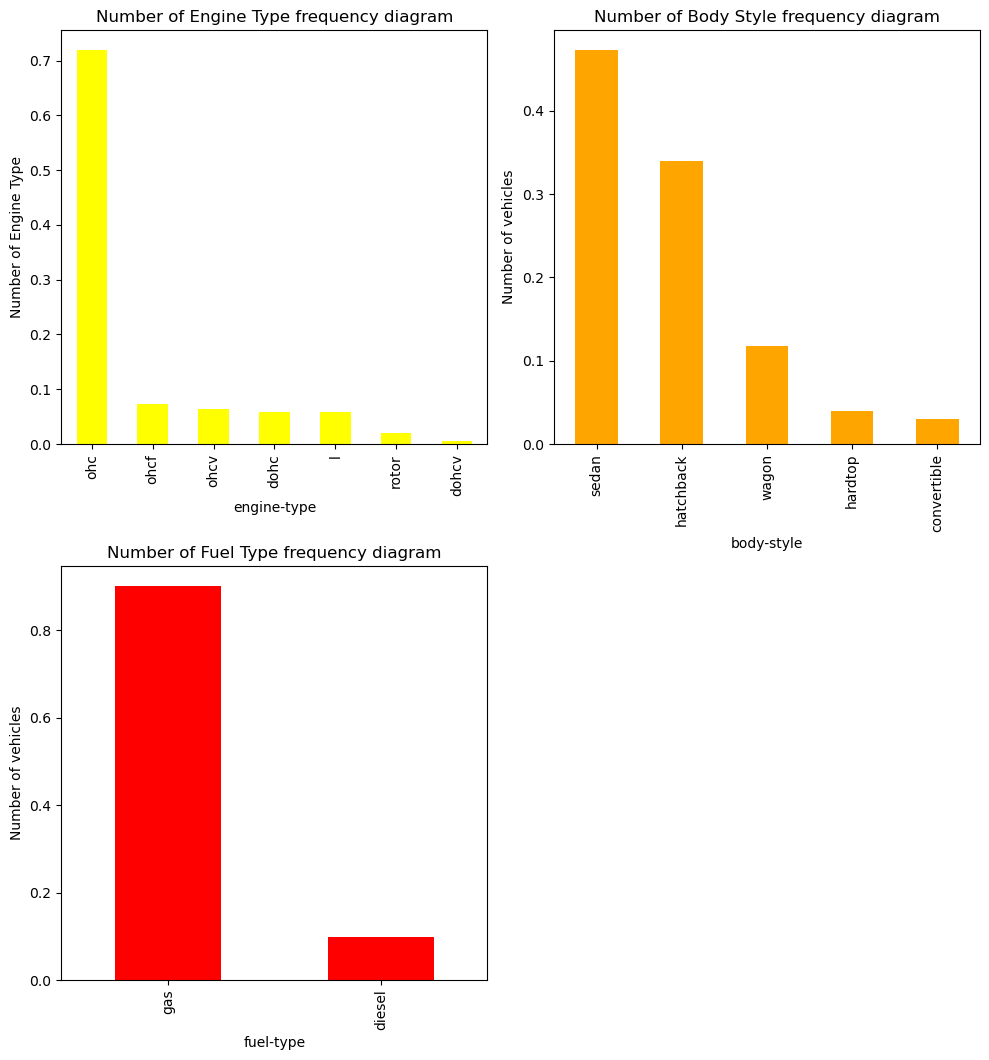


Figure 4

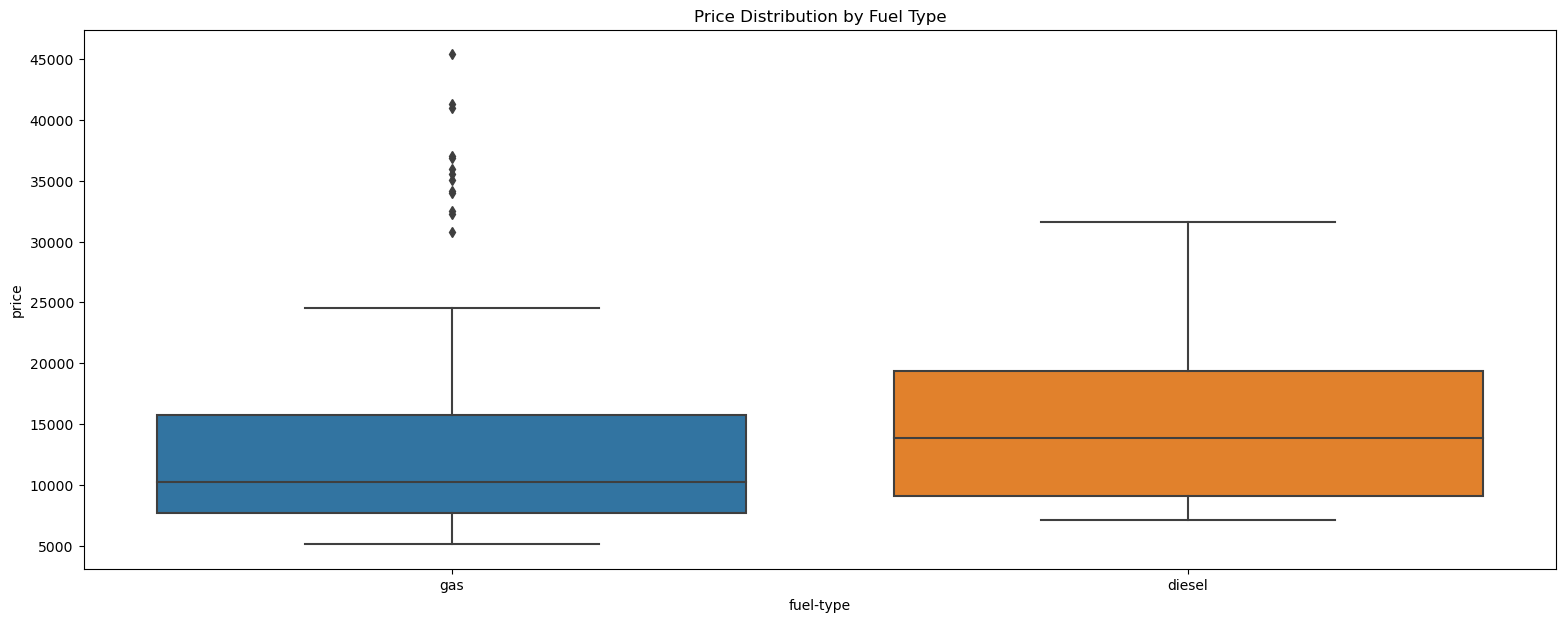
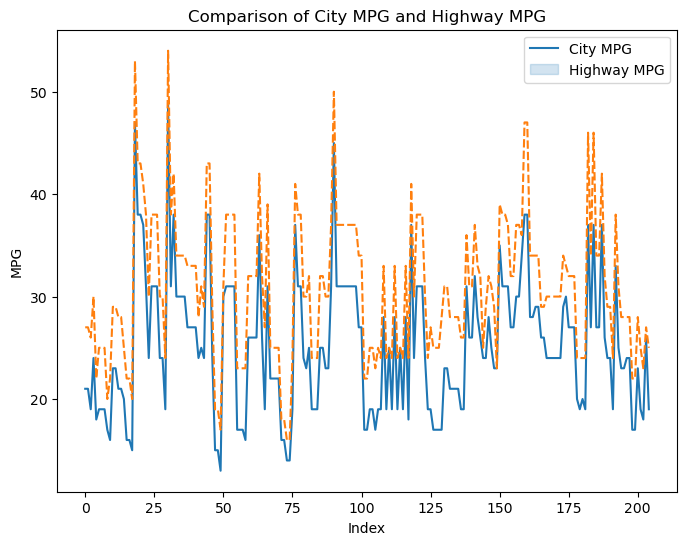


Figure 5

Both a line plot and box plot (Figure 6) were created to compare the city and highway MPGs in order to compare fuel usage discrepancies that may arise due to travelling in these two settings. It is clear from both visualisations that travelling in a city causes more fuel consumption and this may result from constant stopping and starting and being stationary at traffic lights/stop signs and due to traffic. There is a strong, positive, linear relationship between city MGP and highway MPG with a correlation efficient of 0.95.



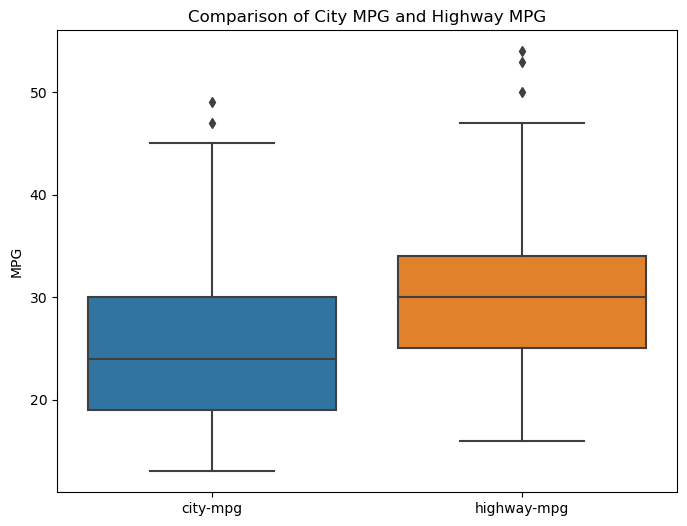


Figure 6

A bar plot (Figure 7) of the frequency of the top 10 car makes was plotted with Toyota being the most popular brand and this is seconded by the fact that Toyota dominated the global car market in 2022, 2021 and 2020. This may be price related because Toyota produces some of the cheapest cars shown by comparative box plot (Figure 8) diagram. Mercedes Benz produces the most expensive cars and does not appear as one of the popular brands.

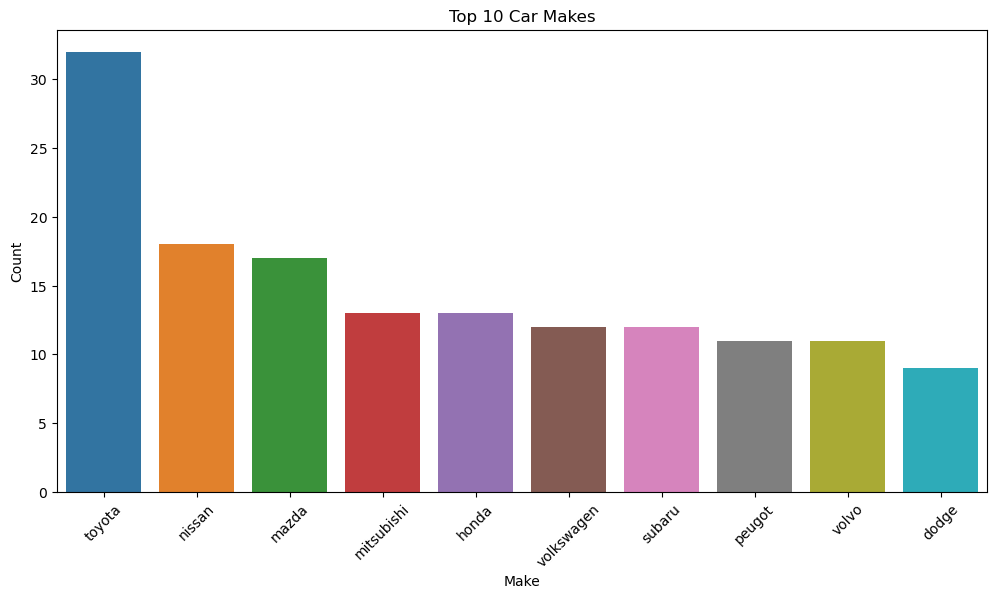


Figure 7

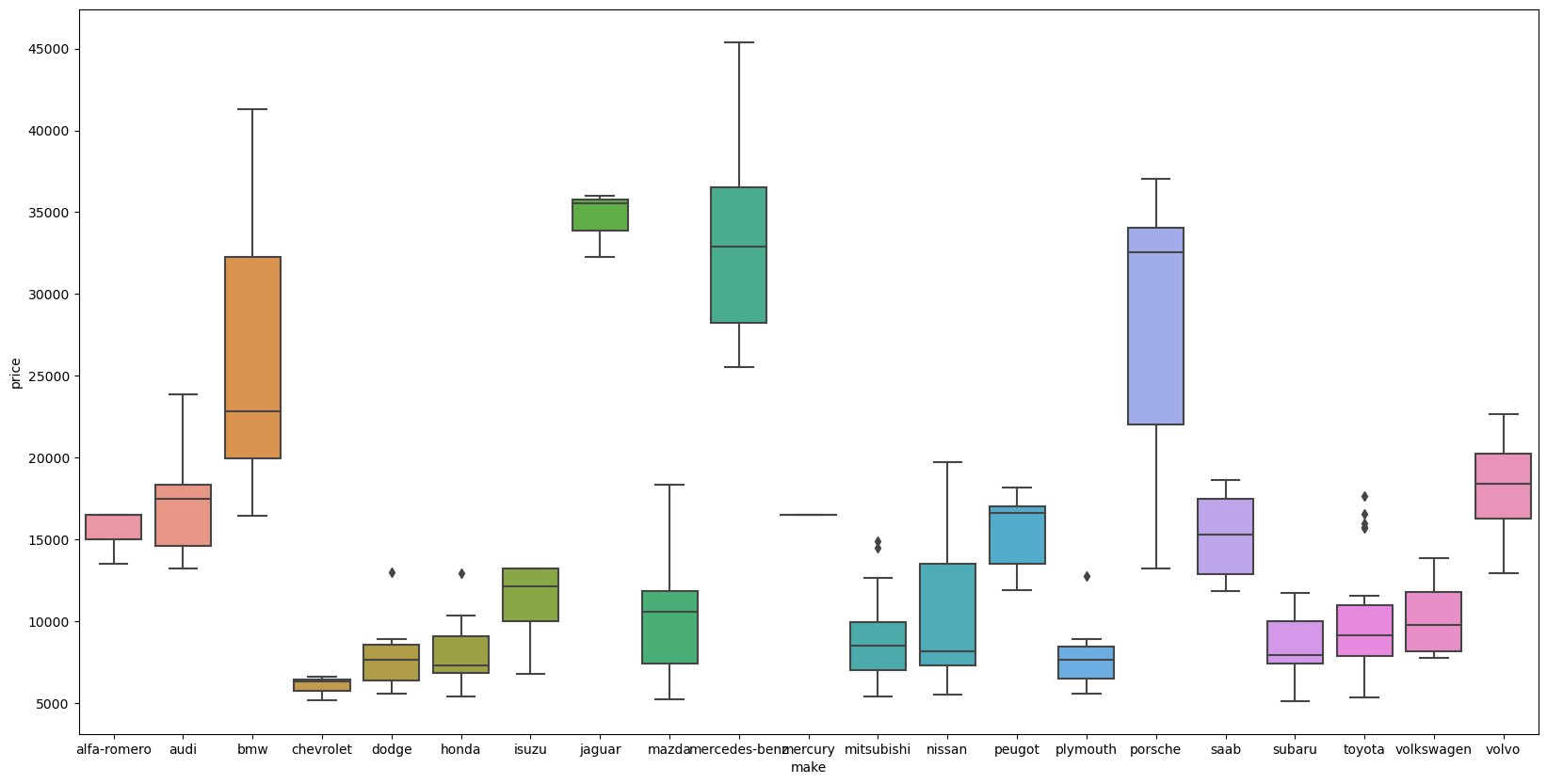


Figure 8