

**TASK**

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

[](https://www.hyperiondev.com/)

**Introduction**

### Context

Dataset Source: https://www.kaggle.com/datasets/toramky/automobile-dataset

This dataset consists of data from 1985 Ward's Automotive Yearbook. Here are the sources

Sources:

1) 1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook.  
2) Personal Auto Manuals, Insurance Services Office, 160 Water Street, New York, NY 10038  
3) Insurance Collision Report, Insurance Institute for Highway Safety, Watergate 600, Washington, DC 20037

### Content

This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) it’s assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is riskier than its price indicates. 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. Actuarians call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.

The third factor is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc…), and represents the average loss per car per year.

Note: Several of the attributes in the database could be used as a "class" attribute.

**DATA CLEANING**

### Understanding Dataset

Uploaded the dataset into the Jupyter notebook as “car” dataset utilizing Pandas pd.read\_csv method. After uploading, utilized the below methods to preview and understand the dataset:

#Understanding the data set

car.info()

car.describe()

car.head(10)

A number of columns doesn’t seem necessary for the rest of data analysis, so dropping them!

car.drop(['symboling','normalized-losses','aspiration','engine-location','wheel-base','fuel-system','bore','stroke','compression-ratio'], axis = 1, inplace = True)

Utilized the below methods to identify and remove any duplicates:

#Check if any duplicate rows

car.duplicated().sum()

#Removing that duplicate row

car = car.drop\_duplicates()

**MISSING DATA**

Utilized isnull().sum(), nunique() and head() methods to understand the missing data

#Understanding missing values

print('\nMissing values    :',car.isnull().values.sum())

#looking at number of unique values per each field

car.nunique()

Reviewing the data by using the methods above, the below fields seem to have "?" character. The nature of the data in these fields based on the description of data led to the decision of replacing the missing data with mean values:

* horsepower
* peak-rpm
* normalized-losses
* bore
* stroke
* price

Further, the "num-of-doors" column is replaced with "four" doors as most of the cars seem to have four doors

#horsepower

df\_temp = car[car['horsepower']!='?']

normalised\_mean = df\_temp['horsepower'].astype(int).mean()

car['horsepower'] = car['horsepower'].replace('?',normalised\_mean).astype(int)

#stroke

df\_temp = car[car['stroke']!='?']

normalised\_mean = df\_temp['stroke'].astype(float).mean()

car['stroke'] = car['stroke'].replace('?',normalised\_mean).astype(float)

#price

df\_temp = car[car['price']!='?']

normalised\_mean = df\_temp['price'].astype(int).mean()

car['price'] = car['price'].replace('?',normalised\_mean).astype(int)

# peak-rpm

df\_temp = car[car['peak-rpm']!='?']

normalised\_mean = df\_temp['peak-rpm'].astype(int).mean()

car['peak-rpm'] = car['peak-rpm'].replace('?',normalised\_mean).astype(int)

# normalized losses

df\_temp = car[car['normalized-losses']!='?']

normalised\_mean = df\_temp['normalized-losses'].astype(int).mean()

car['normalized-losses'] = car['normalized-losses'].replace('?',normalised\_mean).astype(int)

# bore

df\_temp = car[car['bore']!='?']

normalised\_mean = df\_temp['bore'].astype(float).mean()

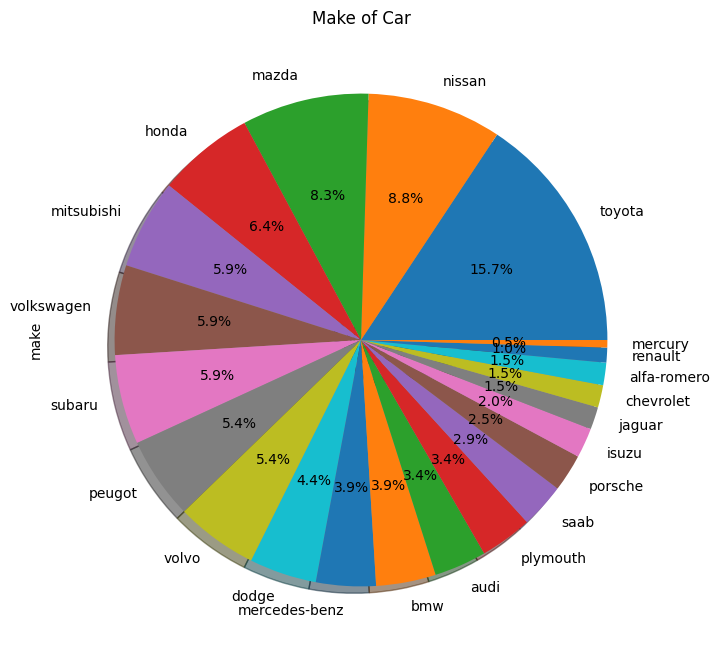
car['bore'] = car['bore'].replace('?',normalised\_mean).astype(float)

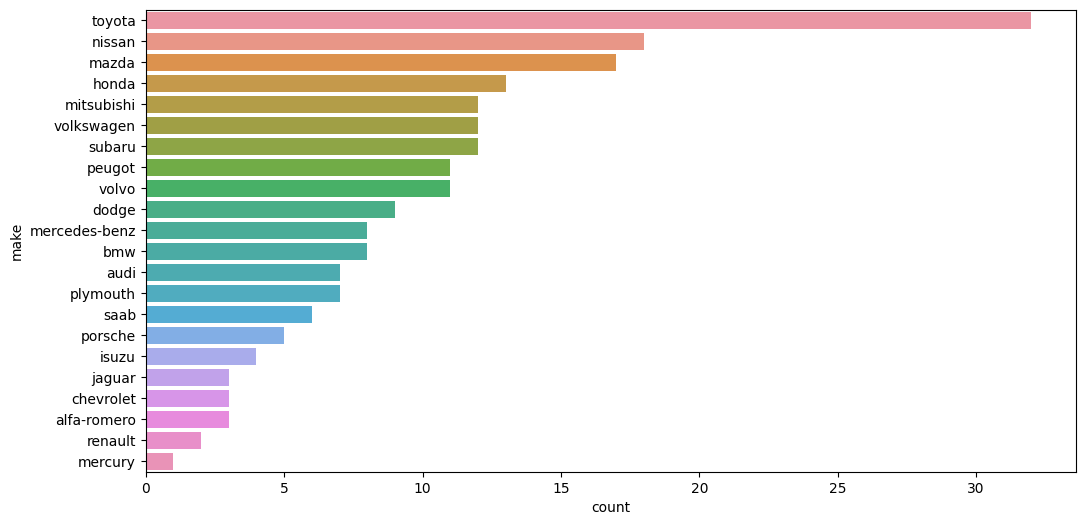
#num-of-doors

car['num-of-doors'] = car['num-of-doors'].replace('?','four')

**DATA STORIES AND VISUALISATIONS**

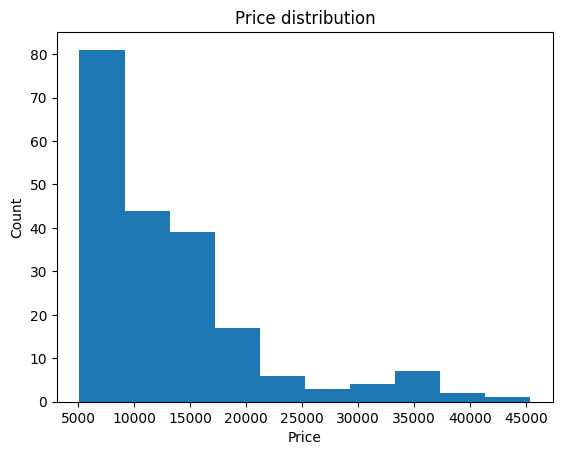
Visualization based on ‘Make’ of the Car:

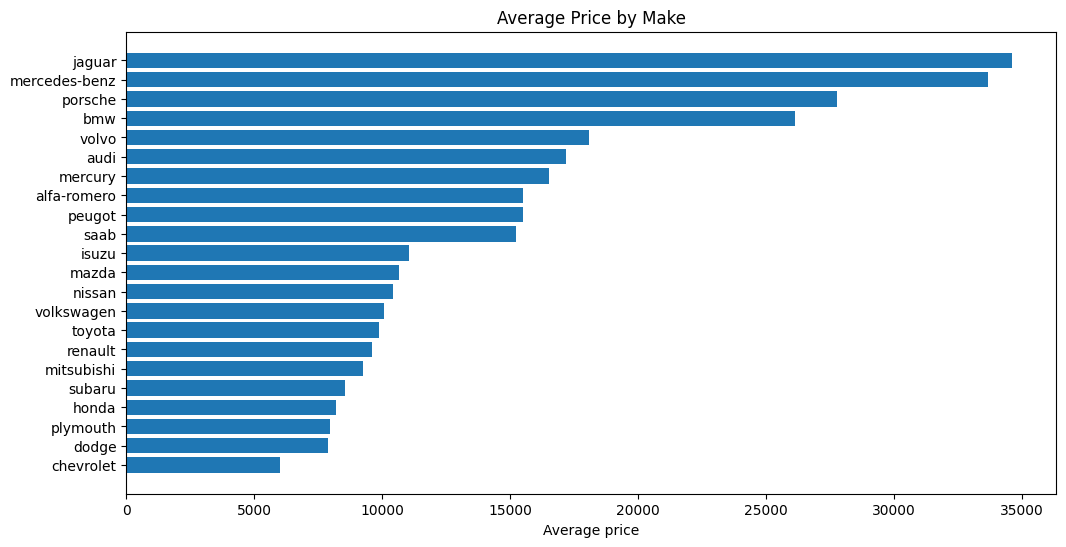




Observation: Toyota has the maximum number while Mercury has the least number of cars in the market

Visualization based on ‘Price’ of the Car:

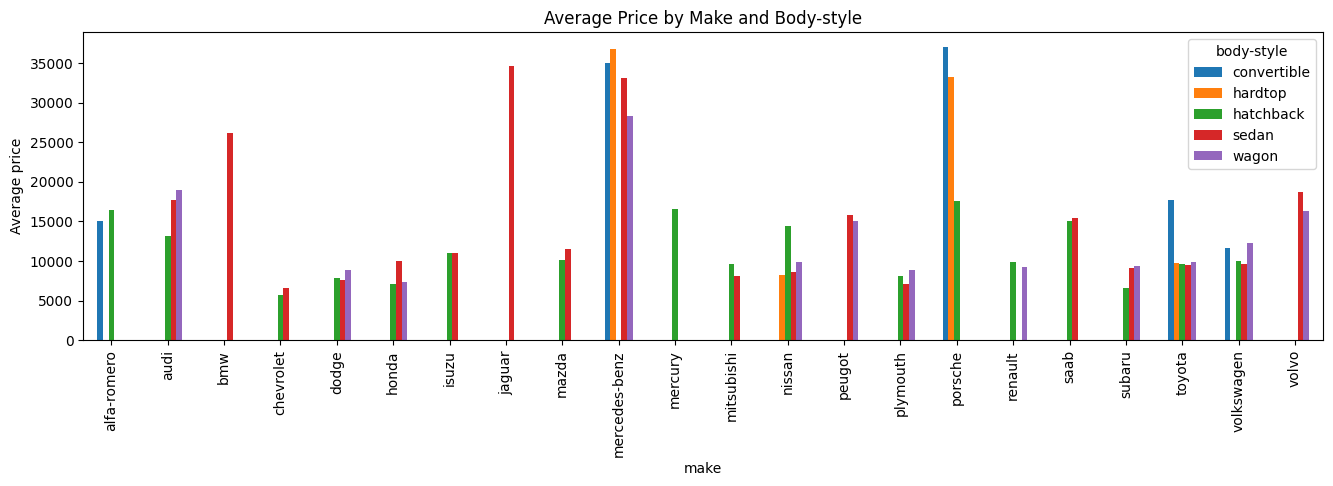




Observations: 1) Checking the distribution of cars by Price: Most of the cars are priced around 21k mark, while a few numbers of cars are in the higher price range from 21k to 45k

2) Jaguar seems to have a few cars in terms of count but highest average price

Visualization based on ‘Price, Make and Body style’ of the Car:

****

Observations: Based on the data and the graphs:

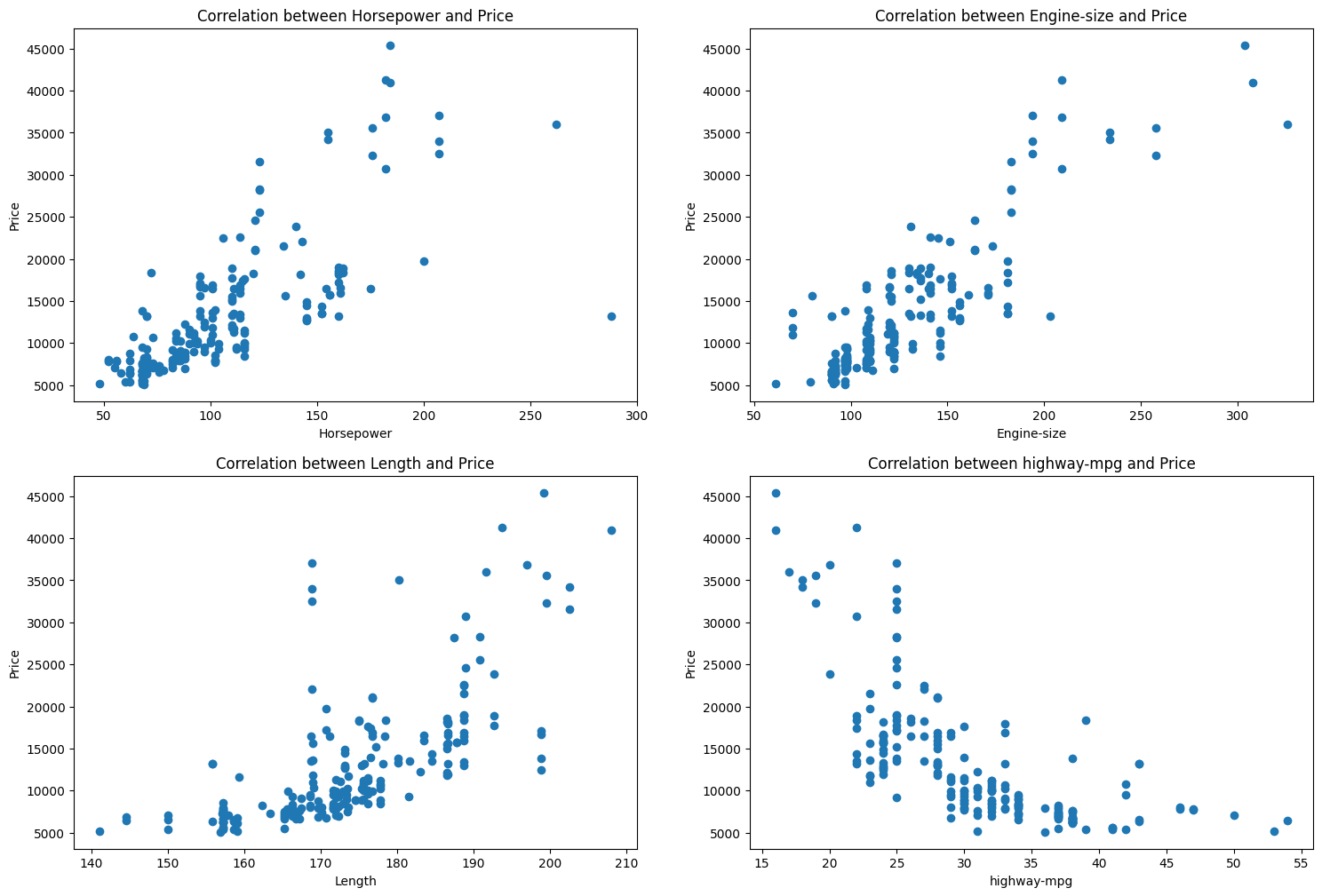
1) Merc, Porsche, BMW and Jaguar are most expensive across all body styles

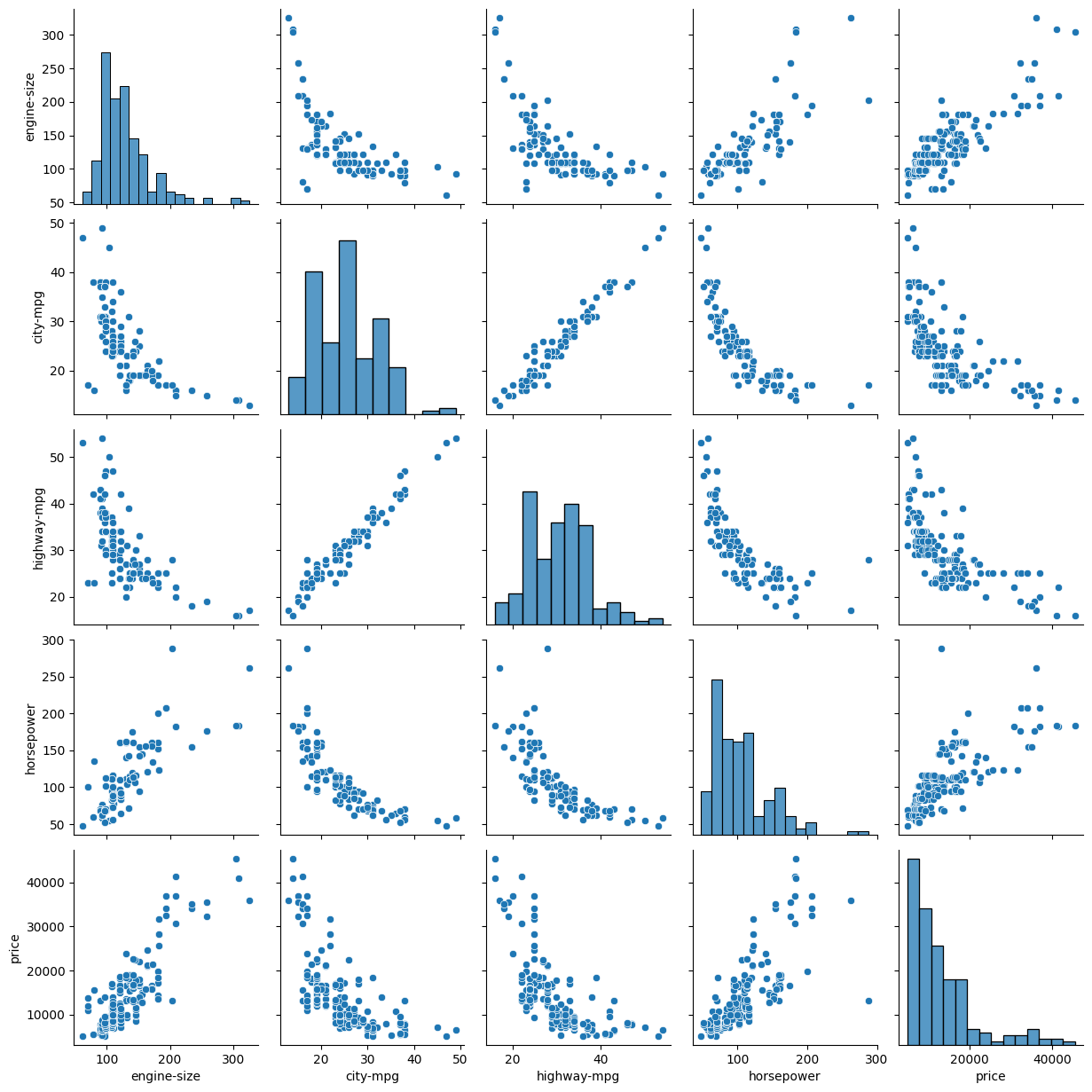
2) Generally, convertibles have a relatively higher price

3) Some manufacturers are specializing in certain body styles only to develop a "niche" - e.g., Jaguar and BMW on sedans (at relatively higher prices) and Mercury on hatchbacks

4) Toyota appears to be having cars across all body styles with relatively similar price range (barring convertibles) - indicates that their goal is to provide widest choice to customers at a relatively smaller price range

Visualization based on CORRELATION

****

****

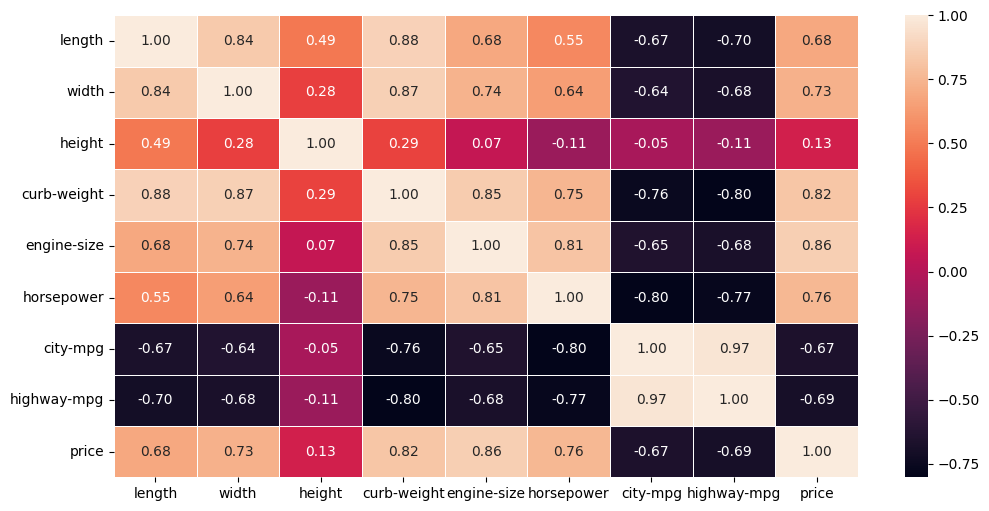
Observations: Based on the data and above graphs:

1) More expensive cars are not fuel efficient; however, the data could be skewed due to "supercars" i.e., extremely high-priced cars

2) Length and Price - have a positive correlation at 0.68 per heatmap below and per the scatterplot above.

3) Engine size and price - has a strong correlation at 0.86 per heatmap below and per the scatterplot above.

4) Horsepower and price - as a strong correlation at 0.76 per heatmap below and per the scatterplot above.

****

Observations from above heatmap:

1. Strong correlation: Car price is strongly correlated with length, width, curb-weight, engine-size, horsepower.
2. Strong correlation: Car price is strongly correlated with city-mpg, highway-mpg.

**THIS REPORT WAS WRITTEN BY: JENNINGS BALAVARI**

**DATE: 19TH JAN 2023**

