Car Price Predictions

Ismail AlKamal
Oct 29, 2021

Introduction

Business Case:

A Chinese car manufacturer wants to enter the US market. We are required to model the price of cars. This will help management understand how exactly the price varies across different features. They can then accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels.

Dataset:

[1] Large dataset of different type of cars across the US market.

[2] 205 entries and 26 features

[3] Target variable is the Price

[4] Source: https://www.kaggle.com/hellbuoy/car-price-prediction

Features

DATA DICTONARY

1	Car_ID	Unique id of each observation (Numeric)
2	Symboling	Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.(Categorical)
3	CarName	Name of car company (Categorical)
4	fueltype	Car fuel type i.e gas or diesel (Categorical)
5	aspiration	Aspiration used in a car (Categorical)
6	doornumber	Number of doors in a car (Categorical)
7	carbody	body of car (Categorical)
8	drivewheel	type of drive wheel (Categorical)
9	enginelocation	Location of car engine (Categorical)
10	wheelbase	Weelbase of car (Numeric)
11	carlength	Length of car (Numeric)
12	carwidth	Width of car (Numeric)
13	carheight	height of car (Numeric)
14	curbweight	The weight of a car without occupants or baggage. (Numeric)
15	enginetype	Type of engine. (Categorical)
16	cylindernumber	cylinder placed in the car (Categorical)
17	enginesize	Size of car (Numeric)
18	fuelsystem	Fuel system of car (Categorical)
19	boreratio	Boreratio of car (Numeric)
20	stroke	Stroke or volume inside the engine (Numeric)
21	compressionratio	compression ratio of car (Numeric)
22	horsepower	Horsepower (Numeric)
23	peakrpm	car peak rpm (Numeric)
24	citympg	Mileage in city (Numeric)
25	highwaympg	Mileage on highway (Numeric)
26	price	Price of car (Numeric)

Data Cleaning

- Data has no missing values

- Data has no duplicates

- **Car_Name** feature can be engineered into **Car_Model** and **Car_Manufacturer**.

Example:
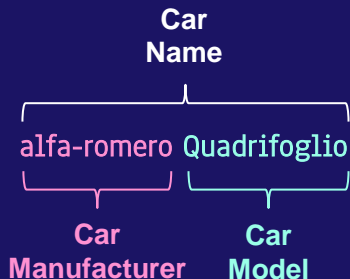
```
# Splitting Car_Name into Car_Manufacturer and Car_Model

car_manufacturer = pd.Series(ds1["Car_Name"].map(lambda x:x.split(" ")[0]),name="Car_Manufacturer")

car_model = pd.Series(ds1["Car_Name"].map(lambda x:" ".join(x.split(" ")[1:])),name="Car_Model")

df = pd.concat([car_manufacturer,car_model,ds1],axis=1)

df.drop(columns="Car_Name",inplace=True)
```

Car Name

alfa-romero Quadrifoglio

Car Manufacturer   **Car Model**

- Fixed a number of **Car_Manufacturer** misspellings such as "maxda" , "vokswagen", "toyouta", etc..

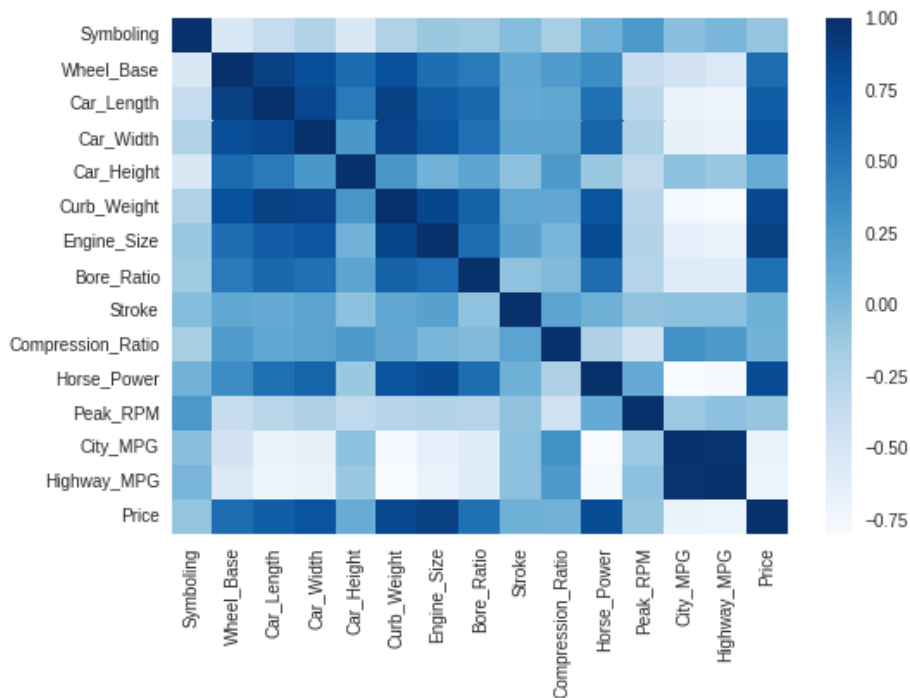- Fixed Audi car model "100 ls" and "100ls" by stripping the whitespace.

Exploratory Data Anaylsis
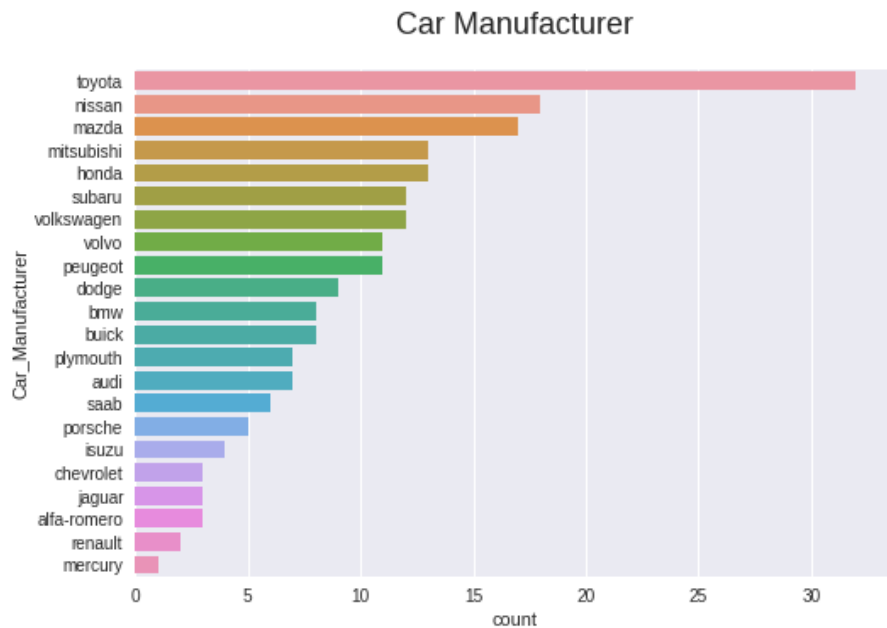
Car Manufacturer

Japanese car manufacturers make about half the samples in the dataset. Toyota being the most frequent.
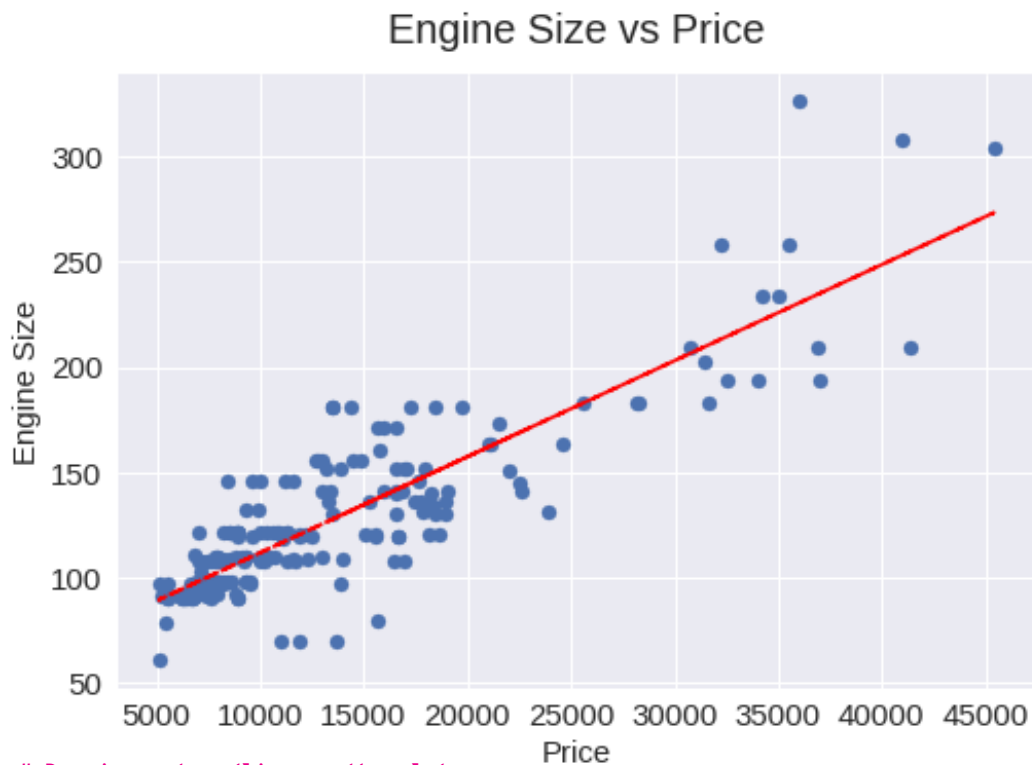
Exploratory Data Anaylsis

Correlation Heatmap (Numerical Only)

Top Factors affecting price:

- Engine Size (+)
- Curb Weight (+)
- Horse Power (+)
- Car Width (+)
- Four Cylinder Type (-)
- Highway MPG (-)
- City MPG (-)
- Car Length (+)
- Rear Wheel Drive (-)
- Forward Wheel Drive (+)

Exploratory Data Anaylsis

Engine Size vs Price

Engine Size is the most important feature in determining car price with a correlation of 87.4%
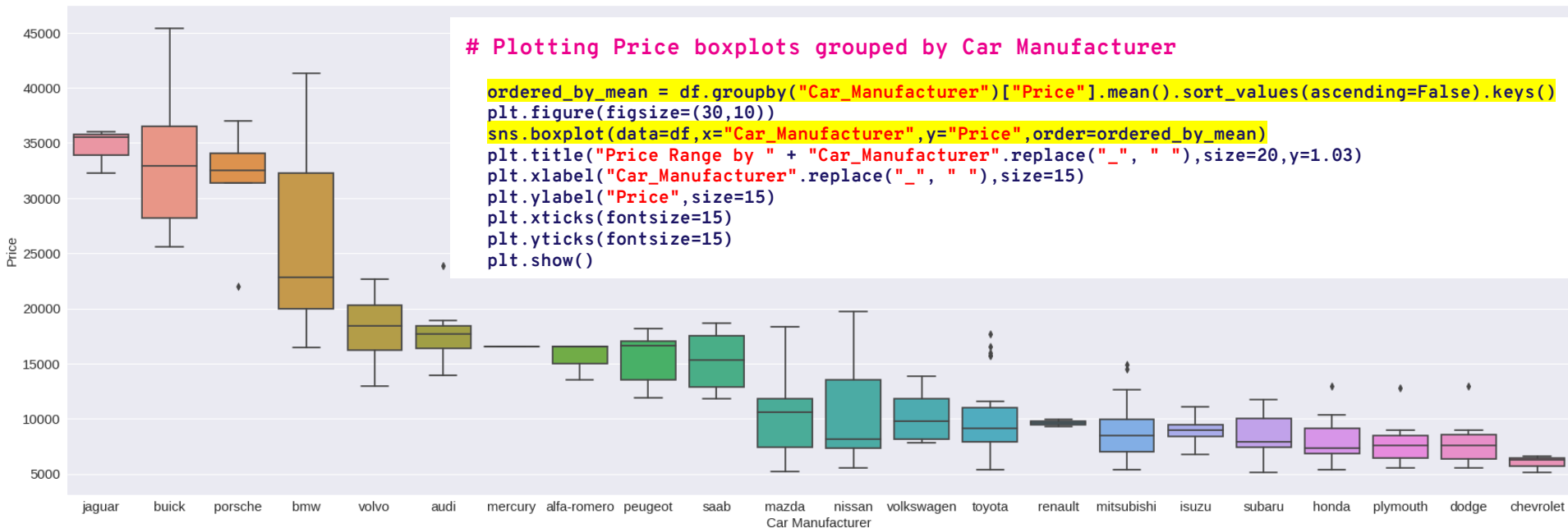
```
# Drawing a trendline on the plot.
z = np.polyfit(df["Price"], df["Engine_Size"], 1)
p = np.poly1d(z)
plt.plot(df["Price"], p(df["Price"]), "r--")
```

Exploratory Data Anaylsis

Price Range by Car Manufacturer

```
# Plotting Price boxplots grouped by Car Manufacturer

ordered_by_mean = df.groupby("Car_Manufacturer")["Price"].mean().sort_values(ascending=False).keys()
plt.figure(figsize=(30,10))
sns.boxplot(data=df,x="Car_Manufacturer",y="Price",order=ordered_by_mean)
plt.title("Price Range by " + "Car_Manufacturer".replace("_", " "),size=20,y=1.03)
plt.xlabel("Car_Manufacturer".replace("_", " "),size=15)
plt.ylabel("Price",size=15)
plt.xticks(fontsize=15)
plt.yticks(fontsize=15)
plt.show()
```

- Jaguars, Buicks, Porsches and BMWs are the most expensive.
- Hondas, Plymouths, Dodges and Chevrolets are the least expensive

THANKS!

Do you have any questions?