CAR PRICES

Prediction of Car prices in US Market

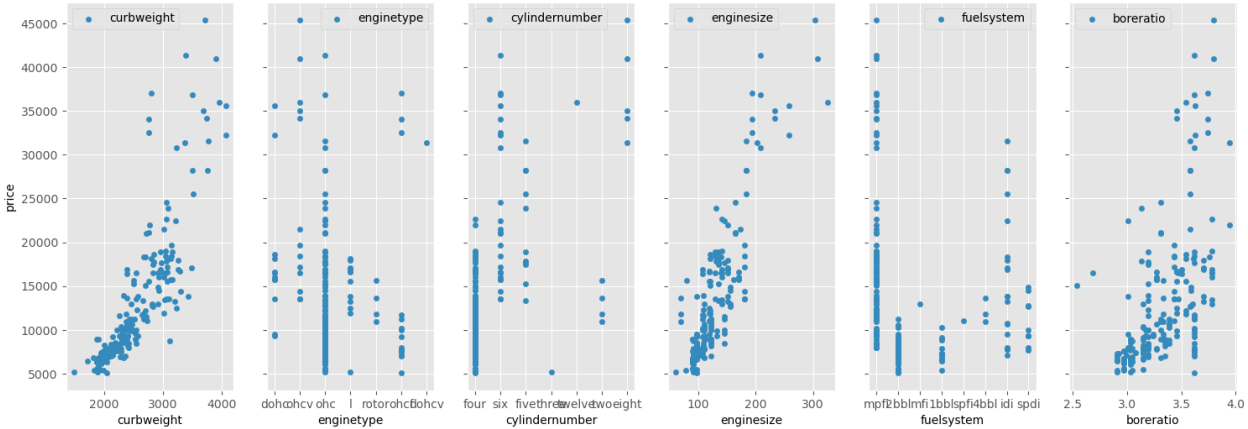
OUTLINE

- Business Problem
- Data
- Modelling
- Evaluation & Results

BUSINESS PROBLEM

To provide analysis on car prices in US market to predict **what features drives the car prices** in US market which will be used by automobile consulting companies to advice potential car Manufacturers.

DATA



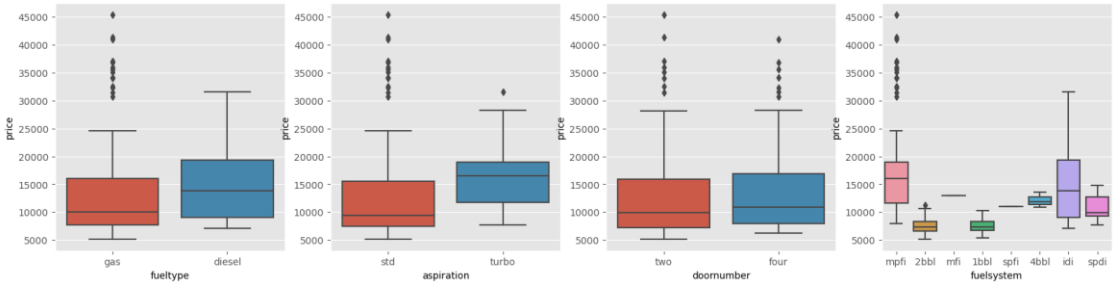
DATA UNDERSTANDING

- The analysis is based on a data set of approximately **300 rows and 26 columns on car prices**. The data includes many different types of information about each cars. They were categorized in to two different types of data by visualizing using scatter plots

  - Continuous Data- Wheel base, Car length, curb weight, Engine size, Bore ratio, High mpg etc.

  - Categorical Data – Car Company, Fuel type, Car body, Drive wheel, Engine location, Engine type, Cylinder number, Fuel system etc.

DATA



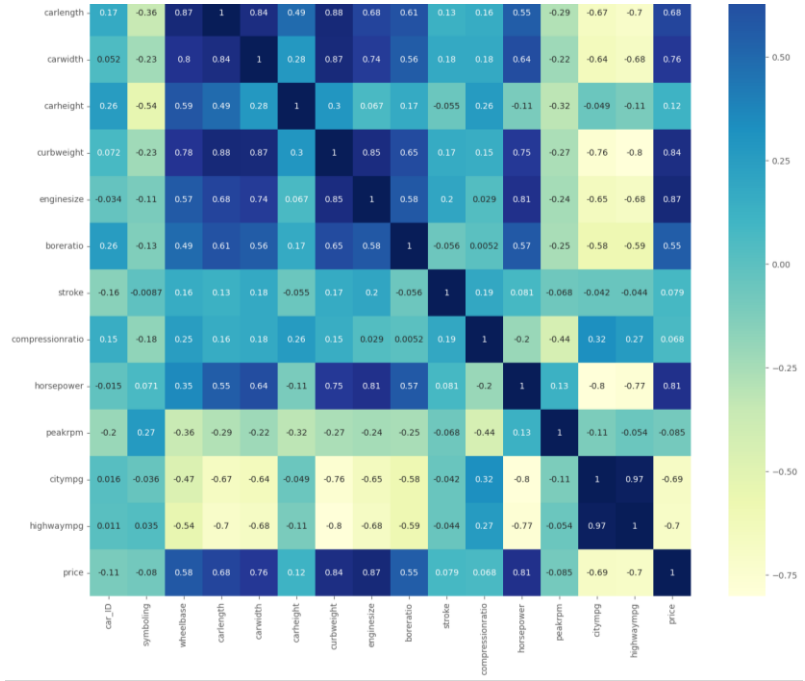
DATA CLEANING

Categorical Variables

1. Visualization of Categorical data using Box plots
2. Dropping insignificant variables on "Price" of the car
3. Deriving new variable for "Car names"
4. Creating Dummies

Continuous Variables

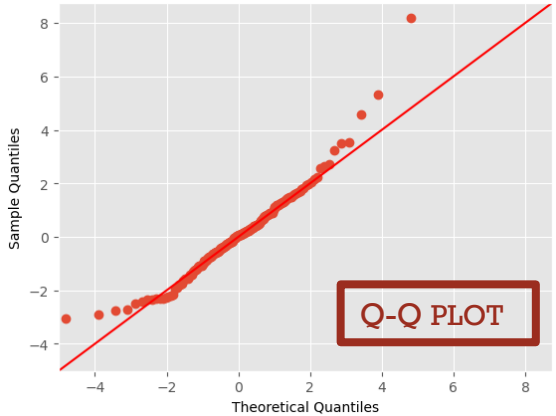
1. Visualization of Continuous data and their relation to Target variable using Heat maps
2. Dropping insignificant variables on "Price" of the car
3. Checking the distribution of Continuous variables using Scatter matrix and doing Log Transformation



MODELLING

Iteration 1

- Checking for Multi Collinearity and dropped variables which are strongly correlated
- Adjusted R square **0.942** but there are many insignificant variables based on the P values
- Model Assumptions were checked using model fit distribution and Regression plots



OLS Regression Results

Dep. Variable:	price	R-squared:	0.951
Model:	OLS	Adj. R-squared:	0.942
Method:	Least Squares	F-statistic:	104.3
Date:	Sat, 22 Jul 2023	Prob (F-statistic):	3.35e-96

	coef	std err	t	P>|t|	[0.025	0.975]
const	5959.2355	2034.478	-2.929	0.004	-9974.995	-1943.476
wheelbase	143.4767	61.695	2.326	0.021	21.700	265.253
curbweight	4.9075	1.152	4.261	0.000	2.634	7.181
boreratio	1650.7734	1130.657	-1.460	0.146	-3882.524	580.977
horsepower	79.5440	13.441	5.918	0.000	53.013	106.075
highwaympg	87.7562	56.105	1.564	0.120	-22.987	198.499
CarCompany_highend	8494.9656	818.117	10.384	0.000	6880.124	1.01e+04
CarCompany_med	1971.1379	434.839	4.533	0.000	1112.830	2829.446
fueltype_diesel	1846.2703	873.806	-2.113	0.036	-3571.035	-121.506
fueltype_gas	4112.9652	1240.157	-3.316	0.001	-6560.851	-1665.079
fuelsystem_idi	-1846.2703	873.806	-2.113	0.036	-3571.035	-121.506
fuelsystem_mfi	-1708.0866	1824.608	-0.936	0.351	-5309.594	1893.421
fuelsystem_mpfi	-491.0412	634.541	-0.774	0.440	-1743.532	761.449
fuelsystem_spdi	-1655.0740	838.889	-1.973	0.050	-3310.917	0.769
fuelsystem_spfi	446.1957	1794.484	0.249	0.804	-3095.849	3988.241

Omnibus:	66.245	Durbin-Watson:	1.407
ob(Omnibus):	0.000	Jarque-Bera (JB):	288.388
Skew:	1.193	Prob(JB):	2.38e-63

MODEL SUMMARY

MODELLING

Iteration 2

- Data was split into Train and test data first to avoid any transformation on test data
- Min MAX Scaling on continuous variable and same scaling on test data
- Recursive feature elimination to select the

```
selected_columns = Xtrain.columns[selector.support_ ]
selected_columns
```

```
Index(['wheelbase', 'curbweight', 'horsepower', 'CarCompany_highend',
       'carbody_convertible', 'enginelocation_rear', 'enginetype_dohcv'],
      dtype='object')
```

- Adjusted R square **0.917** and all the variables seems to be significant based on the P values except one(engine type)
- Dropped the insignificant variable and the and run regression fit
- Adjusted R square **0.916** and all the variables seems to be significant

OLS Regression Results

Dep. Variable:	price	R-squared:	0.921
Model:	OLS	Adj. R-squared:	0.917
Method:	Least Squares	F-statistic:	225.7
Date:	Sat, 22 Jul 2023	Prob (F-statistic):	2.99e-71
Time:	10:15:26	Log-Likelihood:	198.70
No. Observations:	143	AIC:	-381.4
Df Residuals:	135	BIC:	-357.7
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1140	0.068	1.677	0.096	-0.020	0.248
wheelbase	0.2011	0.057	3.523	0.001	0.088	0.314
curbweight	0.2392	0.079	3.026	0.003	0.083	0.396
horsepower	0.4363	0.069	6.311	0.000	0.300	0.573
CarCompany_highend	0.2579	0.021	12.549	0.000	0.217	0.299
carbody_convertible	0.1378	0.036	3.855	0.000	0.067	0.209
enginelocation_front	-0.2092	0.069	-3.050	0.003	-0.345	-0.074
enginetype_dohcv	-0.1290	0.073	-1.772	0.079	-0.273	0.015

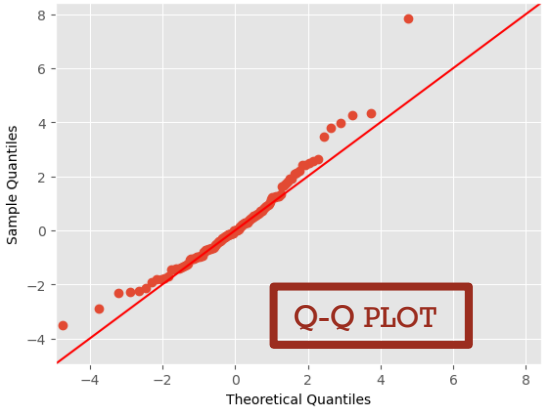
Omnibus:	47.368	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	140.577
Skew:	1.263	Prob(JB):	2.98e-31
Kurtosis:	7.149	Cond. No.	34.6

Dep. Variable:	price	R-squared:	0.919
Model:	OLS	Adj. R-squared:	0.916
Method:	Least Squares	F-statistic:	258.8
Date:	Sat, 22 Jul 2023	Prob (F-statistic):	8.35e-72
Time:	10:36:52	Log-Likelihood:	197.05
No. Observations:	143	AIC:	-380.1
Df Residuals:	136	BIC:	-359.4
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0946	0.013	-7.032	0.000	-0.121	-0.068
wheelbase	0.1926	0.057	3.360	0.001	0.079	0.306
curbweight	0.2750	0.077	3.570	0.000	0.123	0.427
horsepower	0.3827	0.063	6.109	0.000	0.259	0.507
CarCompany_highend	0.2531	0.021	12.327	0.000	0.212	0.294
carbody_convertible	0.1378	0.036	3.823	0.000	0.067	0.209
enginelocation_rear	0.2321	0.068	3.421	0.001	0.098	0.366

Omnibus:	50.974	Durbin-Watson:	1.939
Prob(Omnibus):	0.000	Jarque-Bera (JB):	166.640
Skew:	1.326	Prob(JB):	6.52e-37
Kurtosis:	7.576	Cond. No.	24.4

MODEL SUMMARY



EVALUATION

Model Evaluation

- Applied same feature elimination and dropped insignificant variables on the Test data same as train data
- Check the Mean Squared Error of both Train and Test data

```
Train Mean Squared Error: 0.00372061063925577 42
Test Mean Squared Error:  0.0045711971018560 8
```

There does not seem to be a big difference between the train and test MSE!
Test MSE is slightly higher than the training MSE which indicates that the model is performing well on unseen data and is not overfitting to the training data.

RESULTS

- The main features which Contributes heavily on the prices of cars are:
  1.Wheelbase
  2.Curb weight
  3.Horsepower
  4.CarCompany_highend brand
  5.Enginelocation_rear
  6.Carbody_convertible

- There is positive relation ship between Price and and these variables.

	coef	std err	t	P>|t|
const	-0.0946	0.013	-7.032	0.000
wheelbase	0.1926	0.057	3.360	0.001
curbweight	0.2750	0.077	3.570	0.000
horsepower	0.3827	0.063	6.109	0.000
CarCompany_highend	0.2531	0.021	12.327	0.000
carbody_convertible	0.1378	0.036	3.823	0.000
enginelocation_rear	0.2321	0.068	3.421	0.001

THANK YOU!

EMAIL: e.gajanani9@gmail.com
GITHUB: @Gajas9
LINKEDIN: LINKEDIN.COM/IN/GAJA-SANCHAYAN/