Fundamental of Data Science Mini Project

Car Price Prediction (Linear Regression)

Tzu-Yao Lin

Business Understanding

- ***Target user:***
  Automobile company

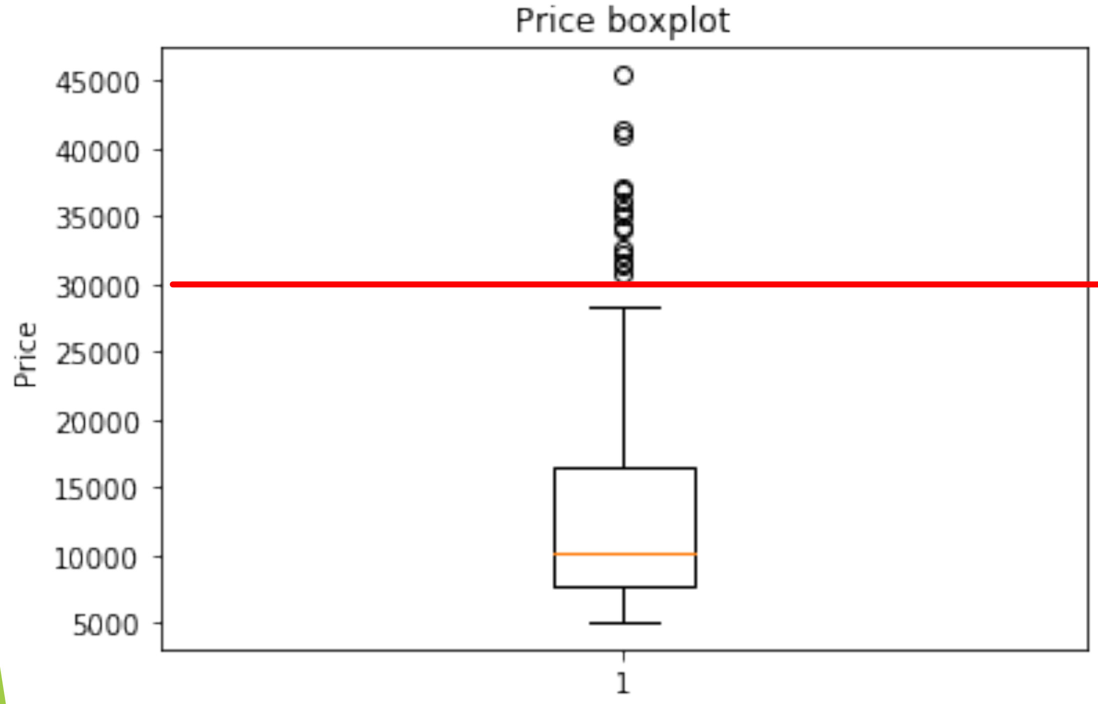
- ***Define Problem:***
  Which variables are significant in predicting the price of a car

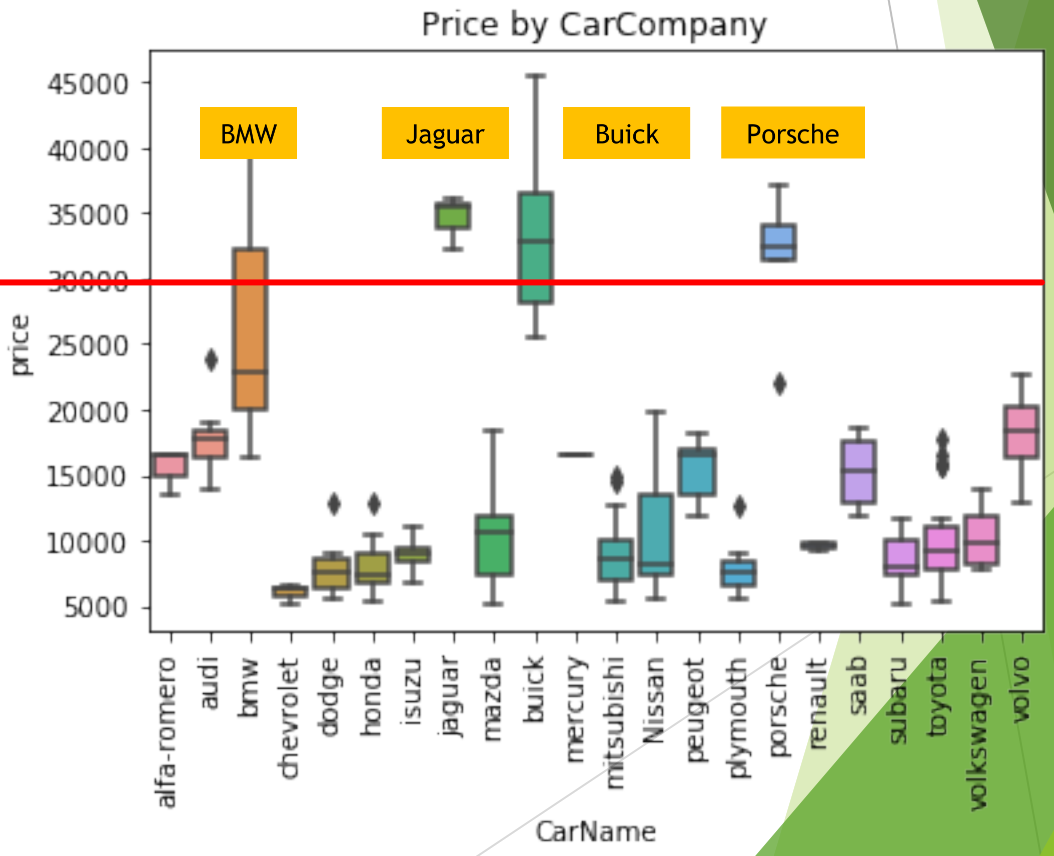
- ***Business Goal:***
  Company can design their cars and meet certain price based on the model

DATA DICTONARY		
1	**Car_ID**	Unique id of each observation (Interger)
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	**carCompany**	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(Dependent variable**	Price of car (Numeric)

Data Understanding

- **Only 4 car company's car price over 30000 (BMW, Jaguar, Buick and Porsche)**
- **It is reasonable to keep these outliers**





Data cleaning & Preprocessing

Detecting missing value

car_ID	0
symboling	0
CarName	0
fueltype	0
aspiration	0
doornumber	0
carbody	0
drivewheel	0
enginelocation	0
wheelbase	0
carlength	0
carwidth	0
carheight	0
curbweight	0
enginetype	0
cylindernumber	0
enginesize	0
fuelsystem	0
boreratio	0
stroke	0
compressionratio	0
horsepower	0
peakrpm	0
citympg	0
highwaympg	0
price	0

Correct misspelling company name

maxda → mazda

porcshce → porsche

toyouta → toyota

vokswagen → volkswagen

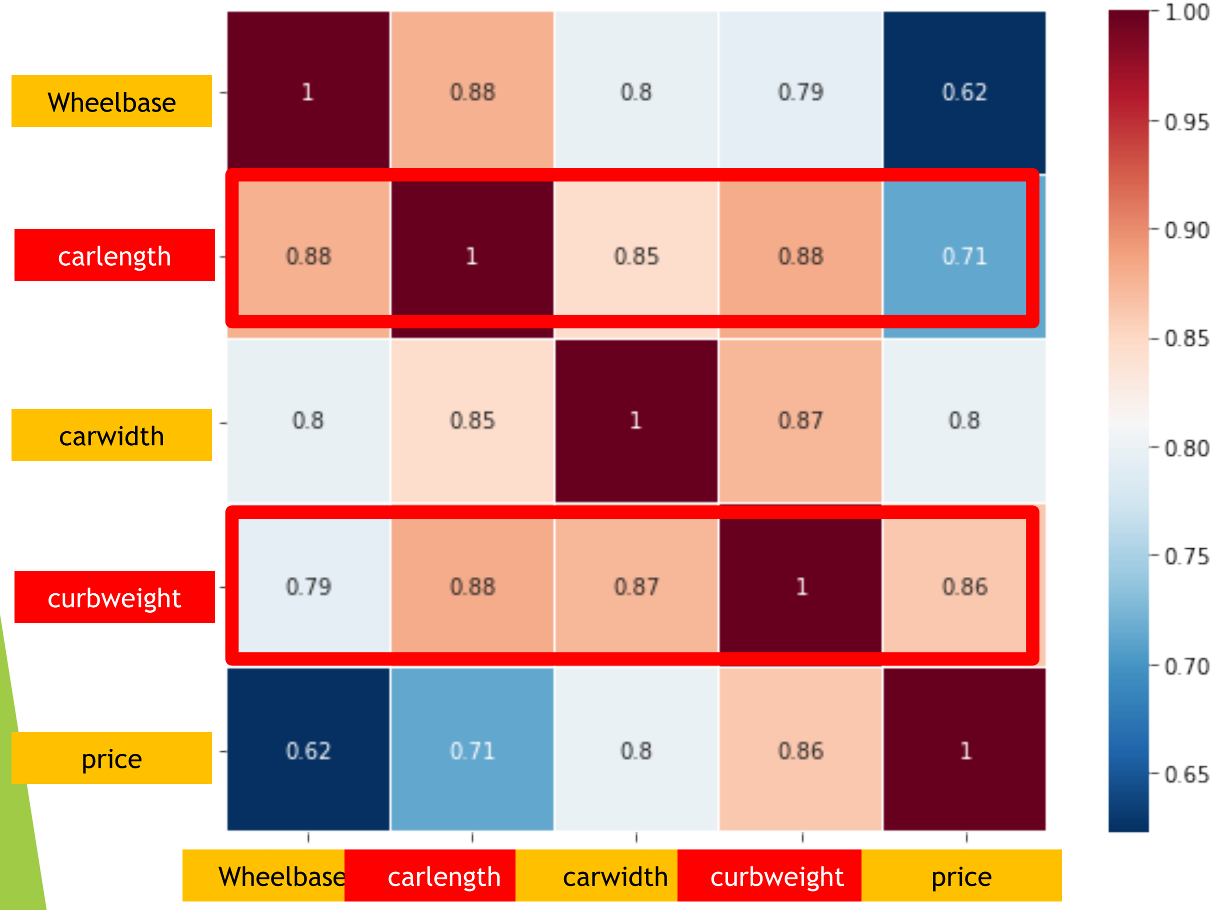
nissan → Nissan

Splitting data

70% Training	30% Test

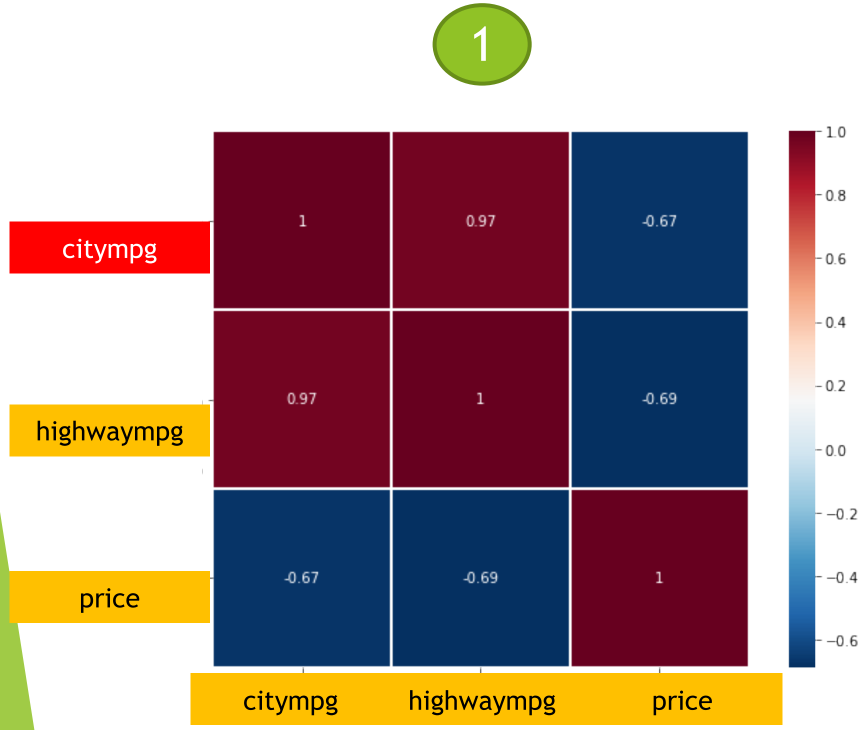
Feature Selection(Continuous)

- ***Check correlation between each features***
- ***According to the MSE result, drop carlength and curbweight***



Feature Selection(Continuous)

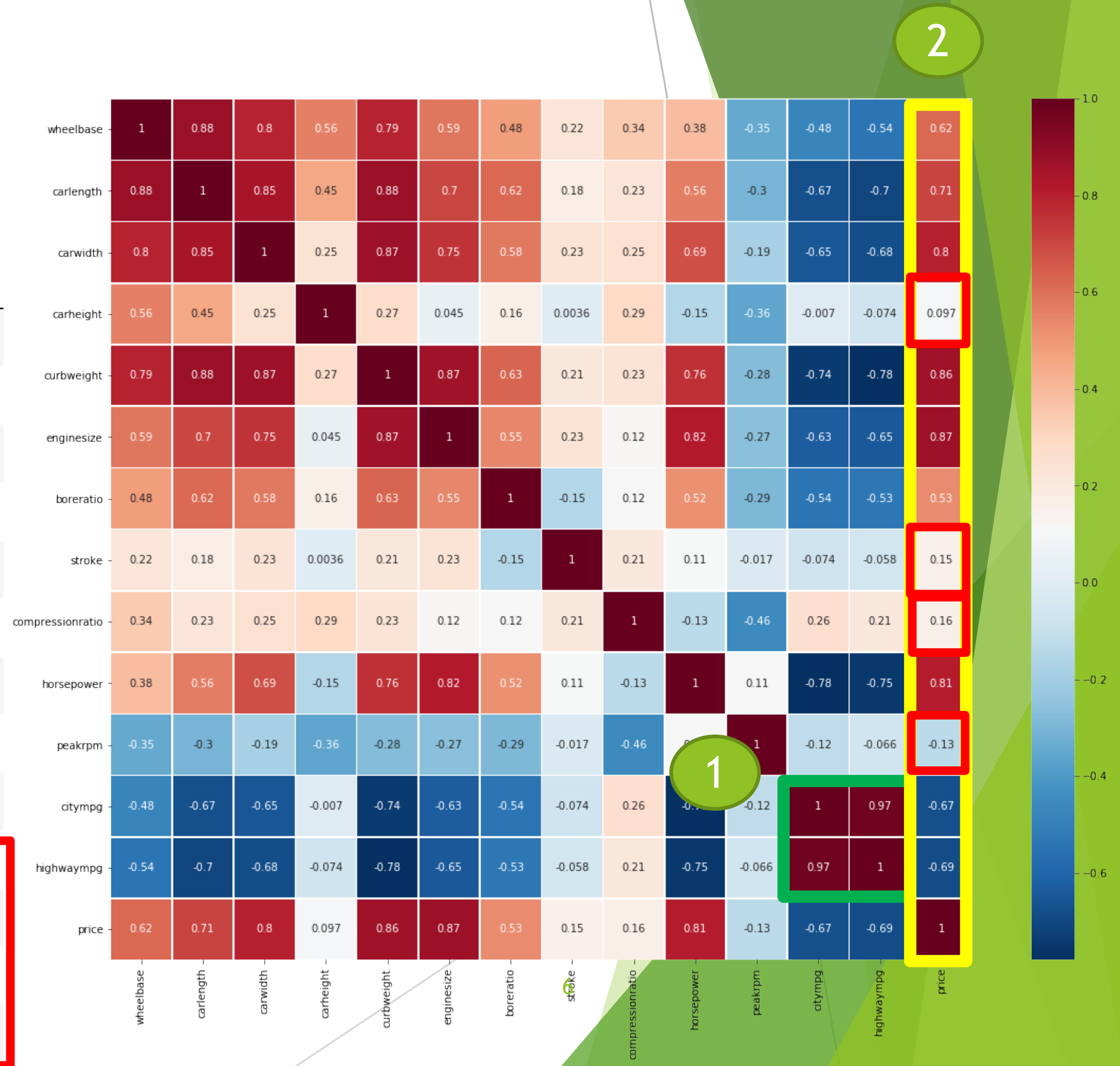
- *Check correlation between each features*
- *Check F-score*
- *Drop the features with four lowest F-score*



2

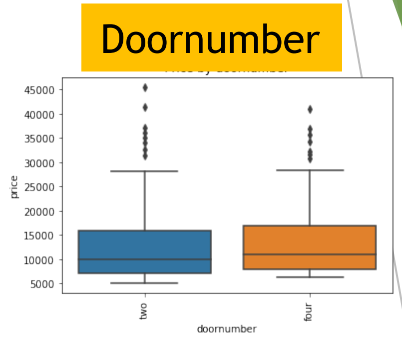
Feature	F-Score
enginesize	430.488900
curbweight	407.216278
horsepower	261.777750
carwidth	249.591029
carlength	146.424930
highwaympg	126.999719
citympg	117.557542
wheelbase	89.248846
boreratio	56.125469
compressionratio	3.744818
stroke	3.371650
peakrpm	2.327430
carheight	1.329005

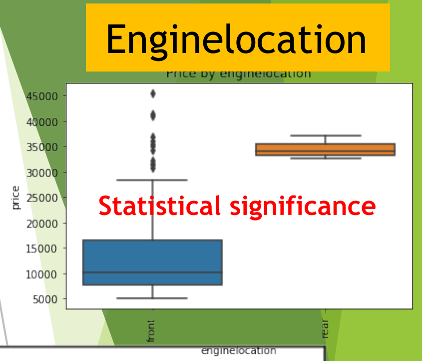
Drop



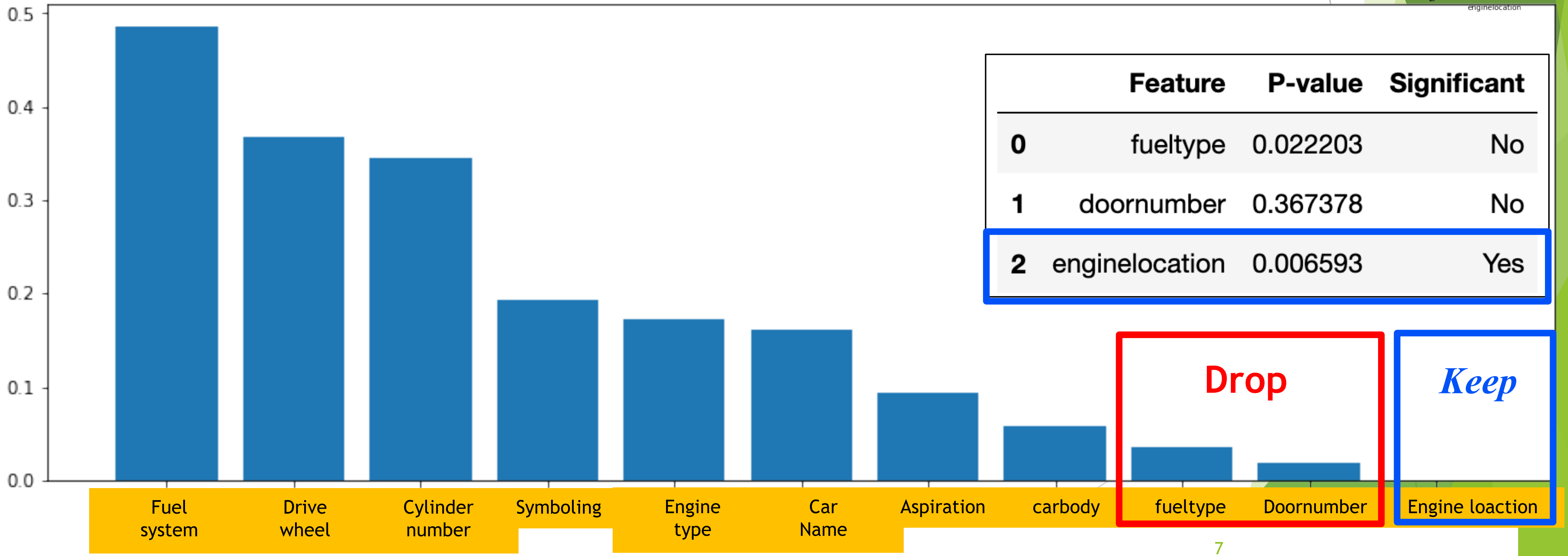
Feature Selection(Categorial)

- ***Run mutual_info_test and P-value test***
- ***Plot the feature versus Price***
- ***Drop fueltype and doornumber and keep engineloaction***



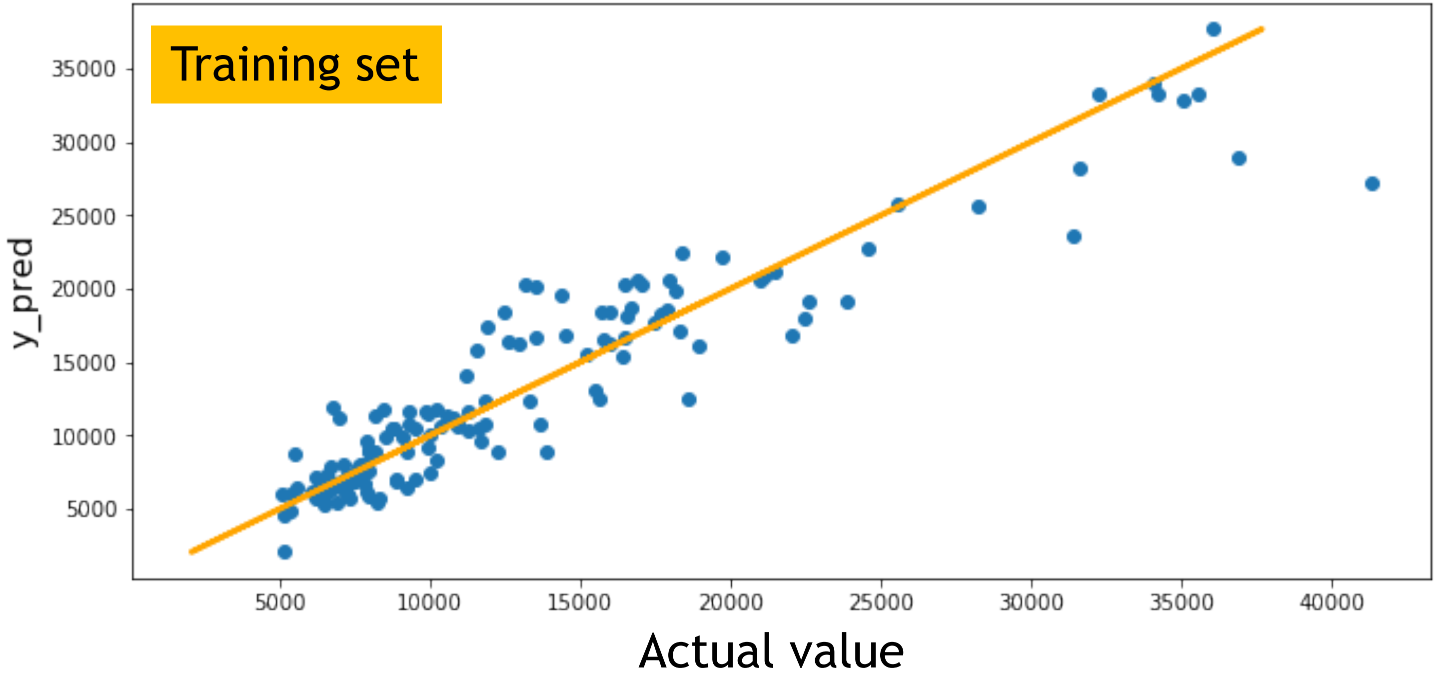


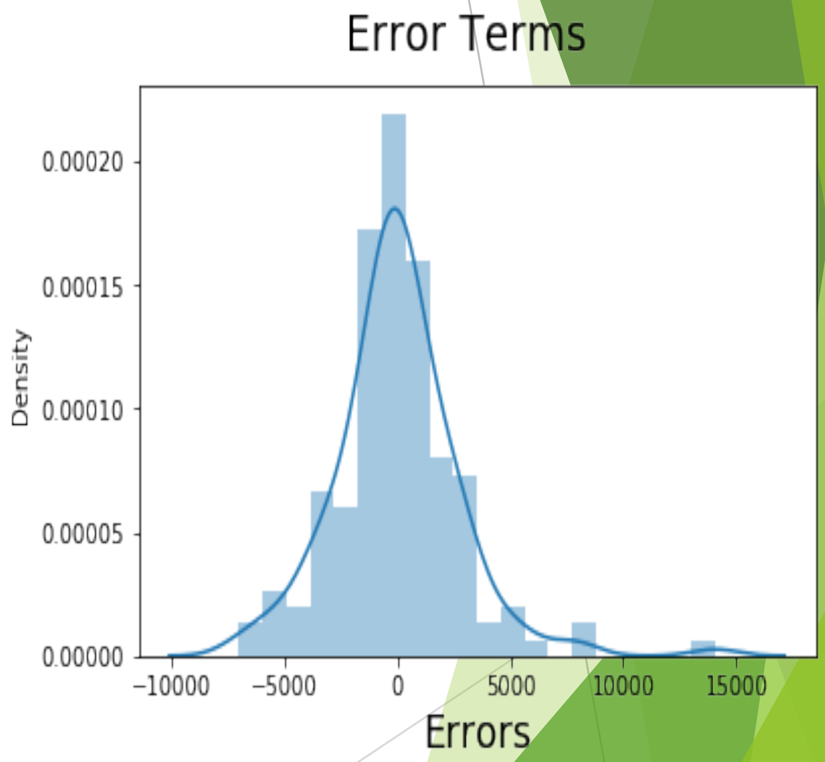
	Feature	P-value	Significant
0	fueltype	0.022203	No
1	doornumber	0.367378	No
2	enginelocation	0.006593	Yes



Result (training data)

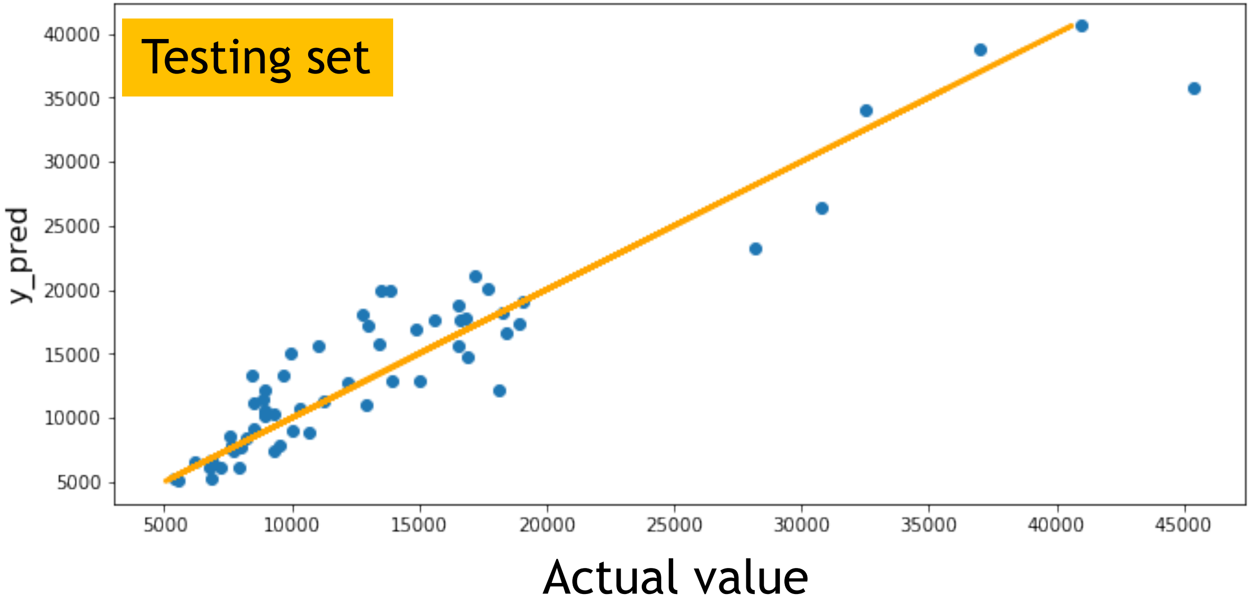
- *Training set: R-squared : 0.89 , Mean Squared Error : 6939060.69*
- *Residual : mean = ~0 , Standard deviation: 2643.46*
- *90% confidence interval:* ± 4348.49

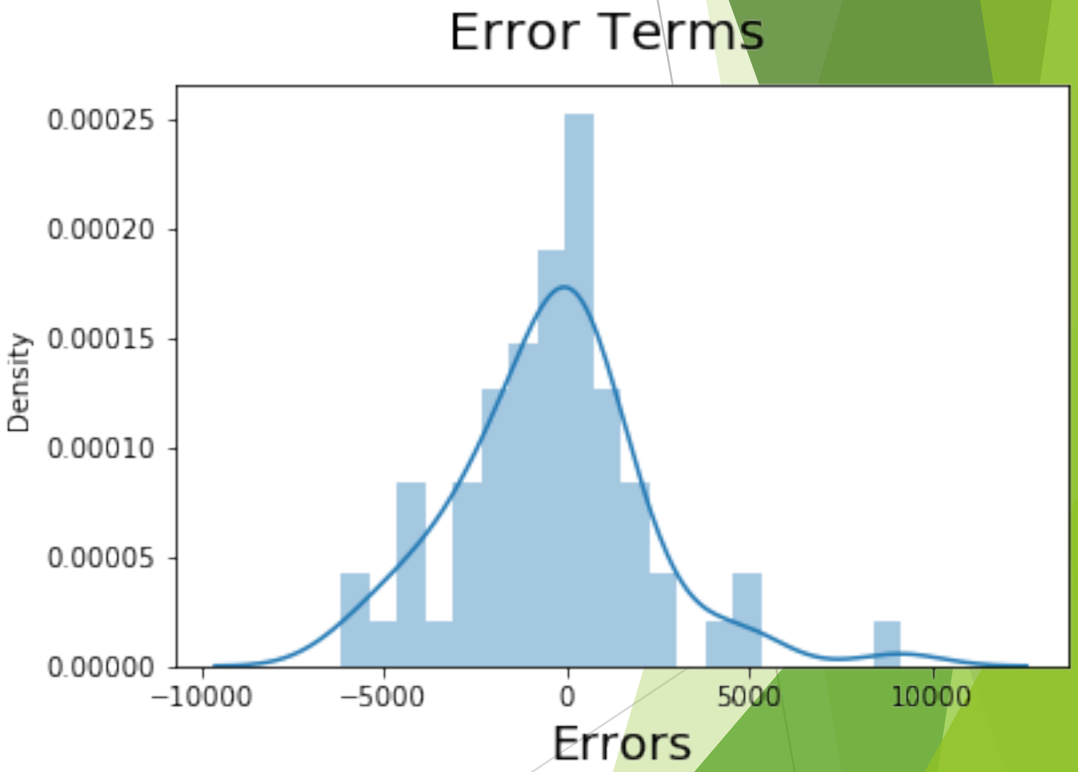




Result (testing data)

- *Training set: R-squared : 0.89 , Mean Squared Error : 7839580.17*
- *Residual : mean = -449.764, Standard deviation: 2799.92*
- *90% confidence interval:* ± 4605.87





Future work

- The model shows high variance in the medium price range. Since, we get the same R-squared values on training set and test set, we might try put more features to train the model.
- The model can't predict accurately the high-priced cars. I should collect more data to figure the issue should be high variance or high bias.
- When performing feature selection, use at least two methods to make the final decision.

Reference

Kaggle : https://www.kaggle.com/goyalshalini93/car-data