Project no :- 2

Laptop price prediction

(Regression problem)

By

Ankit Gurudas Dhanore

Content

Problem Statement

- To help the customer in prediction of the price of the laptop based on specification, Hardware, Brand and all other things. so that that customer can save time on selection of model.

Objective

- The objective of creating a laptop price prediction model is to accurately predict the selling price of laptops based on various features such as brand, processor, memory, storage, and graphics card. The model will be trained on historical data of laptop prices and the corresponding features to identify the correlations and patterns between them. The end goal is to provide insights to businesses and consumers to make informed decisions regarding the purchase and pricing of laptops. Additionally, the model will also assist businesses in setting appropriate prices for their laptops, thereby helping them to stay competitive in the market.

Introduction

- Here we are trying to find out the price of the laptop ,as there are many laptop present in the market ,so on the basis of the specification we can predict the price.
- This will the help the customer to decide which laptop to select and which specifications are best suited for them.
- According to the data nearly 227 millions of laptops are sold every year, so there is very big market for the companies also.
- The price is continuous data so here we are going to use the regression modelling.

Importing Libraries

- Here we import all the libraries which are required for the project like sklearn for modelling and preprocessing , matplotlib and seaborn for data visualization etc.
- Then import the data set and see some basic information regarding the dataset like no of rows and columns and their data type.

Data summary

- The data set is from online website which tells the differents types of laptops according to company.
- the goal of this project is to guide the customer in selecting the laptop model by predicting the price of that laptop on the basis of specification by using regression modelling.
- This data contains 1303 rows and 12 columns
- **Company :-** This column shows the name of the company which manufacture the laptop.
- **Type name :-** This column shoes the type of the model whether is ultarbook, notebook, convertible etc.
- **Inches :-** As the name suggest this column tells us about the screen size of the laptop.
- **Screen resolution :-** This column tells us about the no of pixels and density of the pixels in x and y axis.
- **CPU :-** cpu stands for central processing unit or we can say that the main processor of the computer .
- **Ram** :- Ram stands for random access memory, it is a form of computer memory which can be read and changed in any order, typically used to store working data and machine code
- **Memory** :- This column shows the type of the storage and the capacity of the storage, like ssd is faster and lighter than hdd hence it is more expensive.
- **Gpu** :- Gpu stands for Graphics processing unit, a specialized processor originally designed to accelerate graphics rendering.
- **OpSys** :- This column tells us about the type of the operating system present in the laptop.
- **Weight** :- This column tells us about the overall weight of the laptop.
- **Price** :- This is our target column and this we have to find out on the basis of the specification.

Preprocessing

- First we checked for missing values but there are 0 null values present in the dataset.
- Then we checked for duplicate values but there also 0 duplicate values are present.
- Since there are no null or duplicate values are present in the dataset we jump to next step which is EDA.

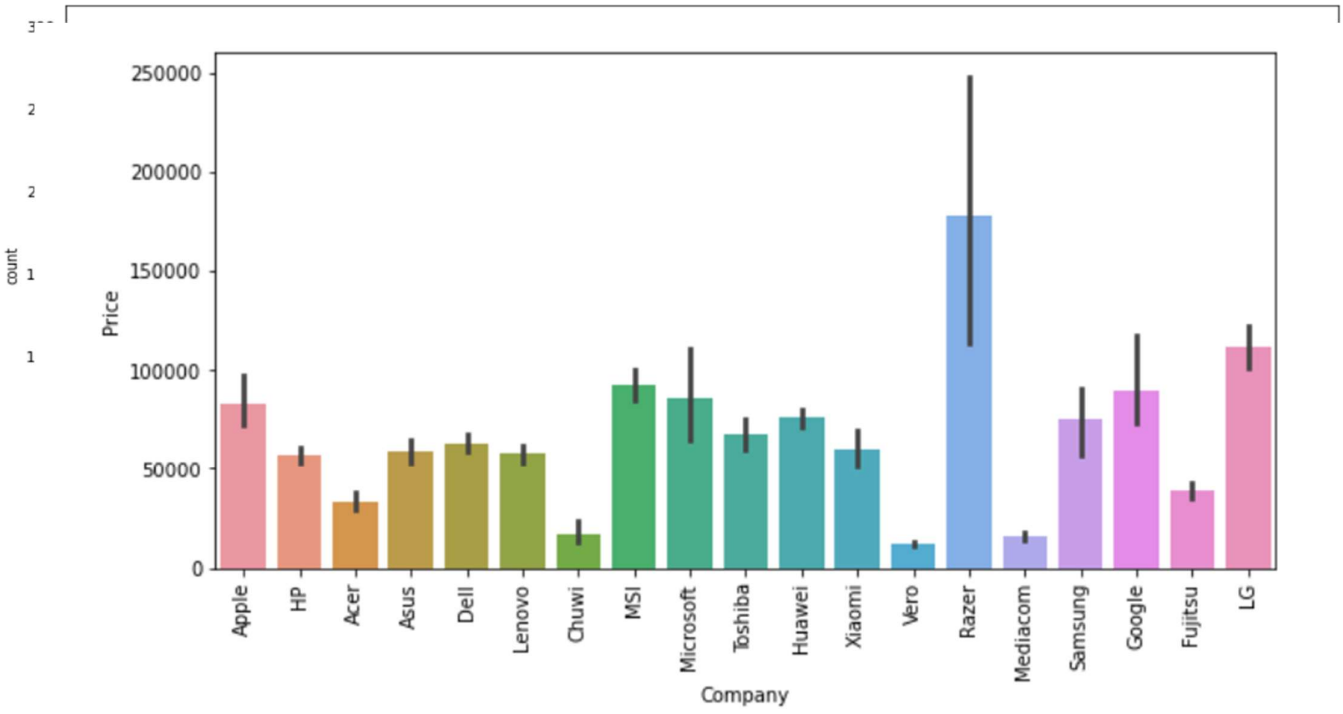
```
In [5]: df.duplicated().sum()  ## che
Out[5]: 0

In [6]: df.isna().sum()  # checking
Out[6]: Unnamed: 0          0
        Company             0
        TypeName            0
        Inches              0
        ScreenResolution    0
        Cpu                 0
        Ram                 0
        Memory              0
        Gpu                 0
        OpSys               0
        Weight              0
        Price               0
        dtype: int64
```

EDA (exploratory data analysis)

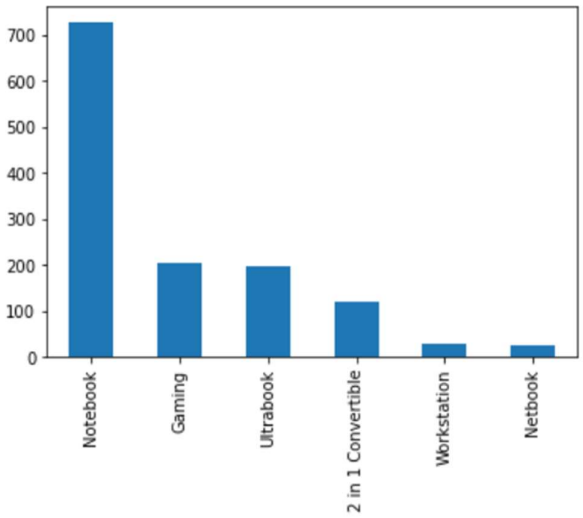
❖ For Company

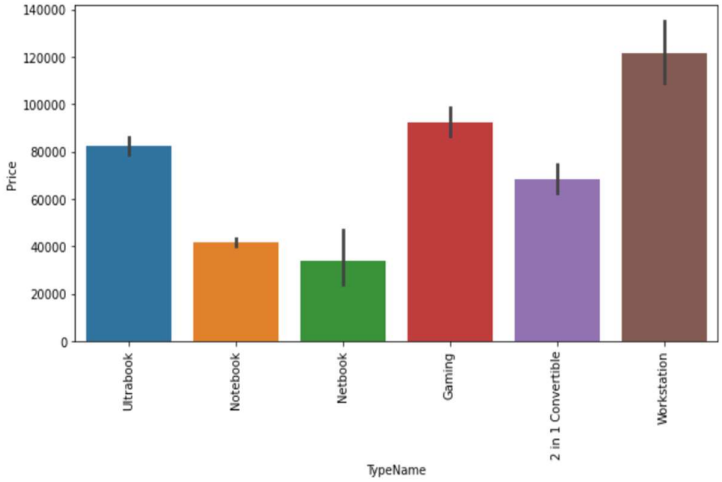
- Here by doing count plot we can see that the maximum no of laptops are of dell and lenovo followed by HP which aregreater than 250 models. and then asus and after that all other company



- And by doing bivariate analysis we can see that most expensive laptop are of Razer company which followed by apple, lg, msi, Samsung etc.

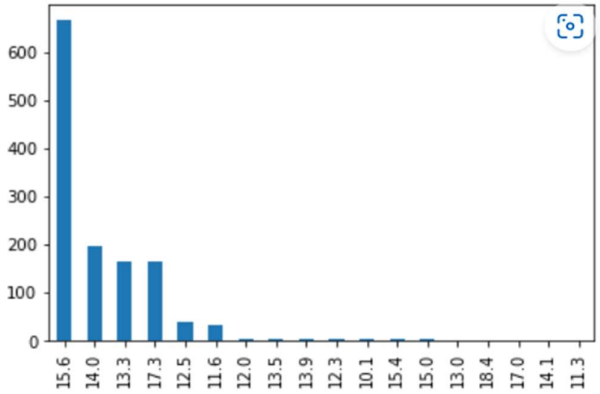
❖ For Laptop Type

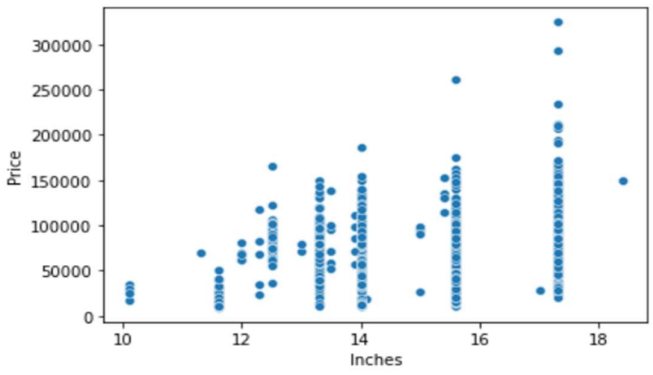




- From the above tables we can see that the maximum no of laptops are of notebook type then gaming and ultrabook.
- And if we see the price distribution of the them then we can see that workstation with avg of 1.2 lakhs are most expensive then ultrabook and gaming around 80k to 100k.

❖ For screen size



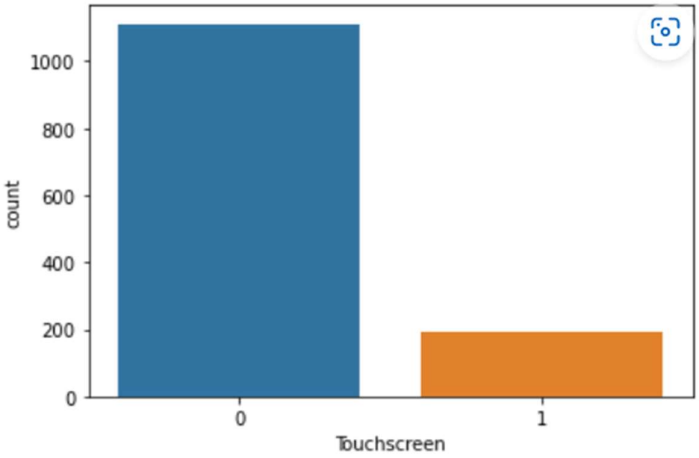


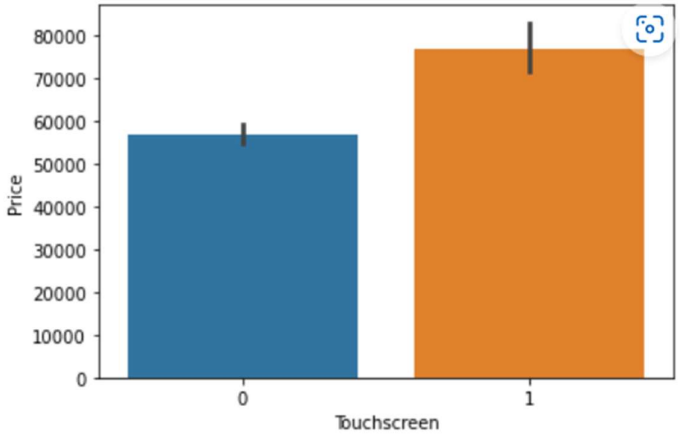
- Here we can clearly see that the the max no of laptops are of 15.6 inches.

- But we cannot see direct relation between the price and screen size.

❖ For screen resolution

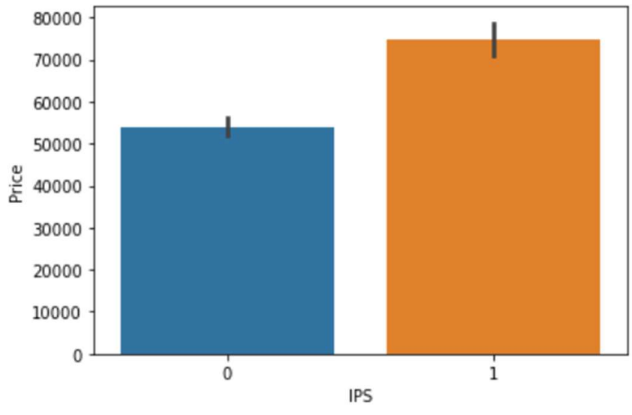
- Here we can see that the screen resolution column has so many complexities so we need to simplify it and turn it into usefull column.
- So from that column we have made one another column which is touch screen. Here 0 is non touch screen and 1 is touchscreen.

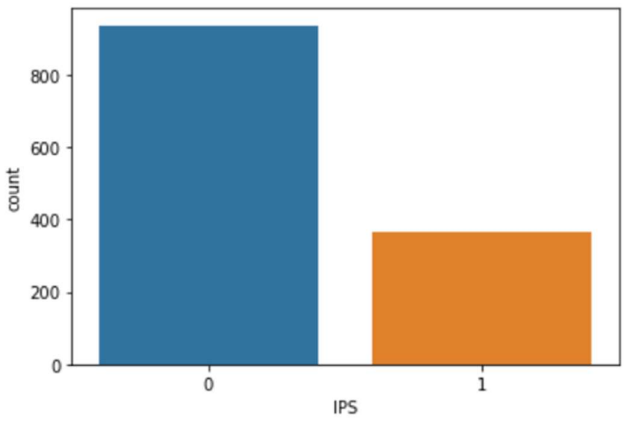




- From above table it is clear that laptop with touchscreen is more expensive.

- From that we have also made one more column which is ips which is type of screen type.

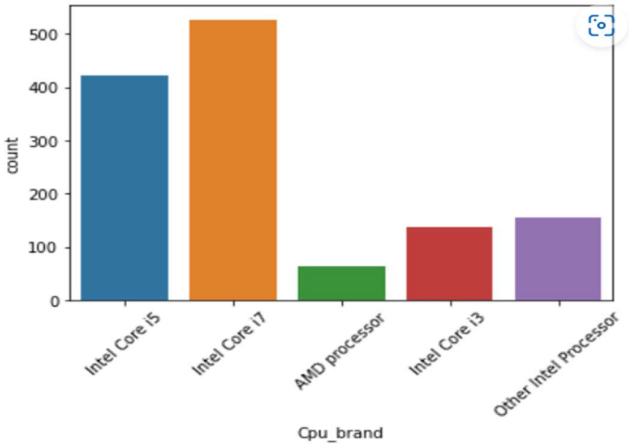


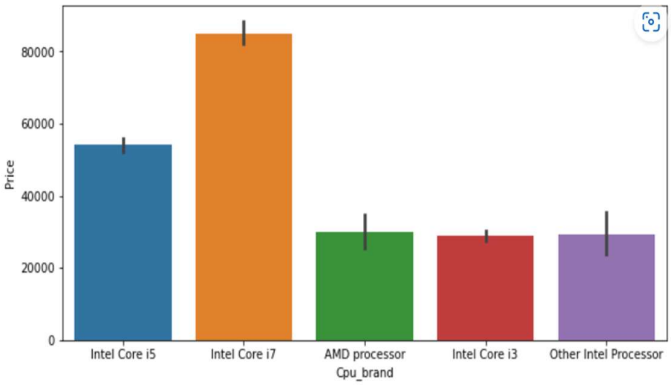


- The screen resolution is dependent on the pixel and size of the screen which is different for every screen size so it is converted to ppi (pixel per inches).

❖ For CPU

- It is clear that higher the cpu grade higher the price of the laptop will be.

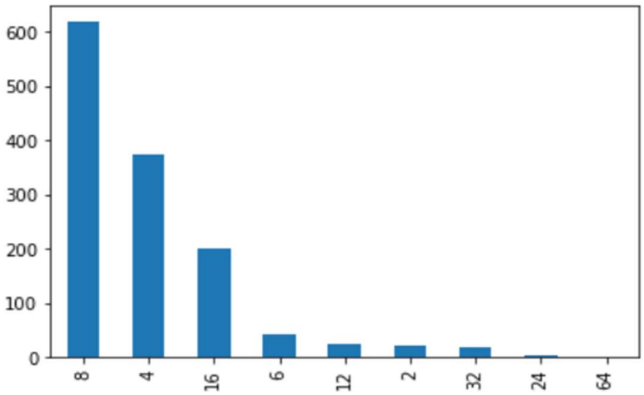


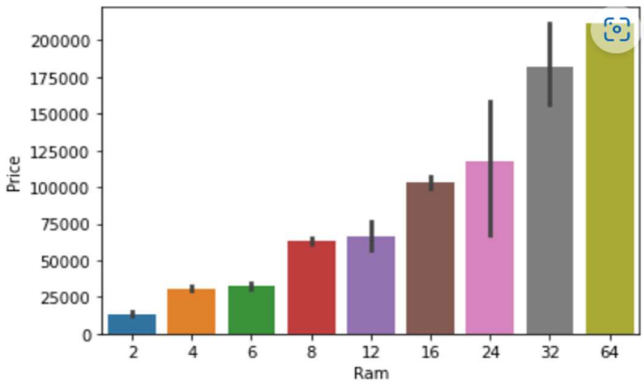


- Most expensive cpu in comparison to the type is intel core i7 then i5 follwed by amd and other brands.

❖ For RAM

- It is obvious that more the ram more higher the price will be





- So 64 gb model has highest price compare to all available.

❖ For memory

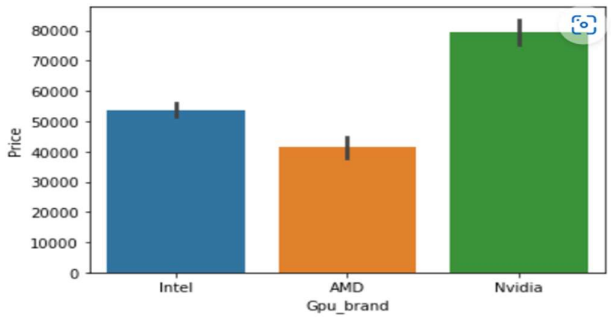
- It is shown from the data that ssd are more faster and reliable than the conventional hdd .

- But at the same time ssd are way more expensive than hdd for the same storage.

❖ For GPU brand

```
In [55]: df["Gpu_brand"].value_counts()

Out[55]: Intel     722
         Nvidia    400
         AMD       180
         Name: Gpu_brand, dtype: int64
```



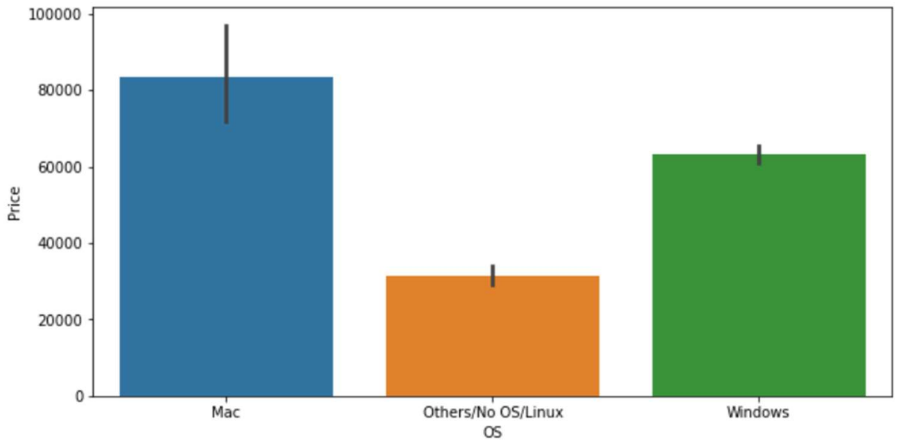
- There are 3 types of brand intel, nvidia, and amd.
- Out of this three Nvidia is most expensive.
- The price ranges between 40k upto 80k.

❖ For OpSys

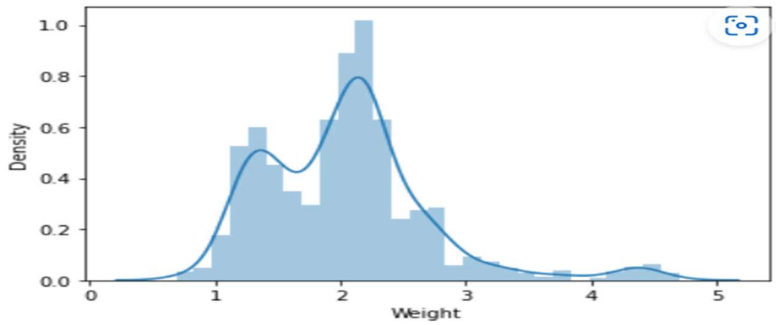
- There are many operating system in the market.
- But among all mac and windows are the most expensive which ranges between 80k to 100k.

```
|: ## Checking count of every OpSys
   df["OpSys"].value_counts()

|: Windows 10      1072
   No OS              66
   Linux              62
   Windows 7          45
   Chrome OS          26
   macOS              13
   Mac OS X            8
   Windows 10 S        8
   Android             2
   Name: OpSys, dtype: int64
```



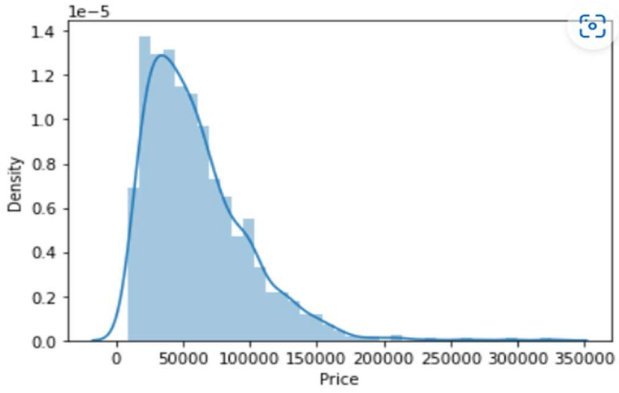
❖ For weight

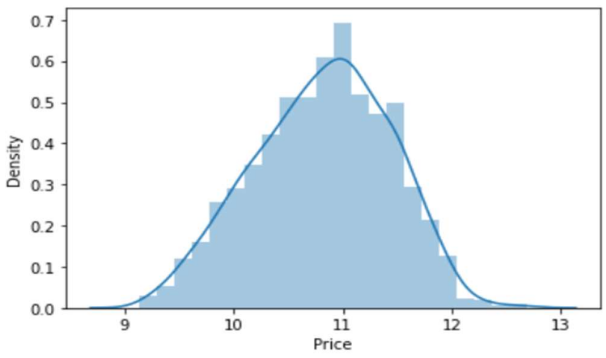


- It is normally distributed.
- It ranges between 1 to 5 kgs.

❖ For price

- The price is the target feature which is continuous.
- But it is skewed so we need to log transform to bring it to normal distribution.

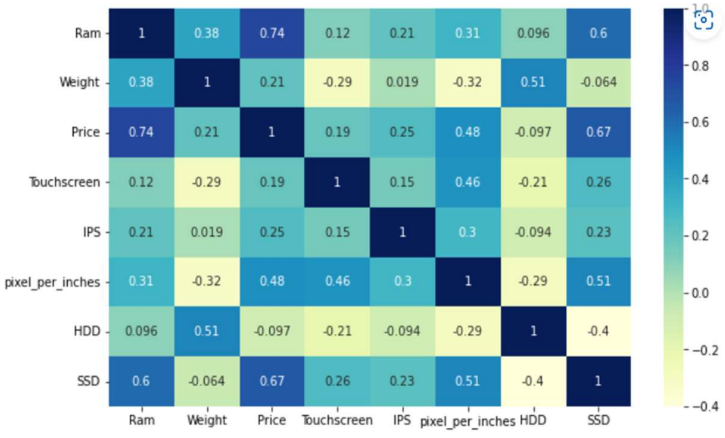




- This shows the graphs before and after log transformation.

❖ Multivariate analysis

- it shows no high multi colinearity between the features.
- All features are independent to each others.



Modelling

❖ Linear regression

- R2 score is :- 0.8368
- MAE score is :- 0.2038

❖ Rigde

- R2 score is :- 0.8366
- MAE score is :- 0.2052

❖ Lasso

- R2 score is :- 0.8362
- MAE score is :- 0.2038

❖ KNN

- R2 score is :- 0.8294
- MAE score is :- 0.1952

❖ Decision Tree

- R2 score is :- 0.8261
- MAE score is :- 0.2016

❖ SVM

- R2 score is :- 0.8086
- MAE score is :- 0.2013

❖ Random forest

- R2 score is :- 0.8864
- MAE score is :- 0.1651

❖ Adaboost

- R2 score is :- 0.8368
- MAE score is :- 0.2038

Model comparison

- It is clear that the Random forest is working best so we will go with Random forest algorithm.

Challenges faced

- Most difficult task was to do preprocessing as the data was very noisy and a bit tricky to simplify.
- Doing feature engineering was difficult

Conclusion

- With proper Preprocessing and feature engineering the random forest works best in this dataset.

Thank you