CAPSTONE PROJECT

FLIPKART PRODUCT DATA ANALYSIS

Vidhu Govind

DA/DS

November FN

Objective :

The main objective of this project to analyse product data from an e commerce website. The website chosen here is Flipkart and product chosen is laptops.

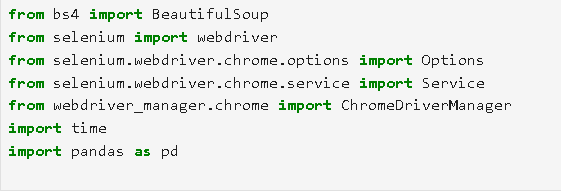
As part of analysis the main tasks are scrapping the data from website, performing EDA ,storing the data into a database in SQL , make unsupervised learning models and supervised learning models and do hyperparameter tunning.

Data Overview:

Data includes -Product name, product category, customer rating, total number of reviews, product price. further for analysis the product name was split into to further categories- Brand,Model,Ram,Rom,OS and Ram Type. Initially data from 10 pages were scrapped to get a data frame which contained about 258 rows, but after removing duplicates and nulls the final dataset contained 158 rows.

Data scrapping:

Flipkart was the website chosen and multipage web scrapping method was used to get the required data.

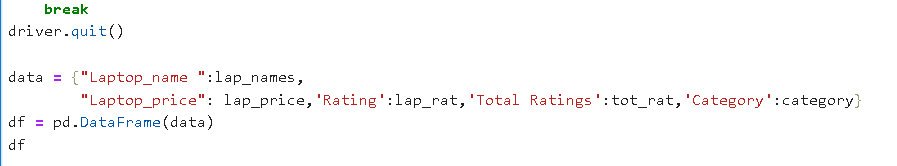


These are the libraries used for web scrapping.

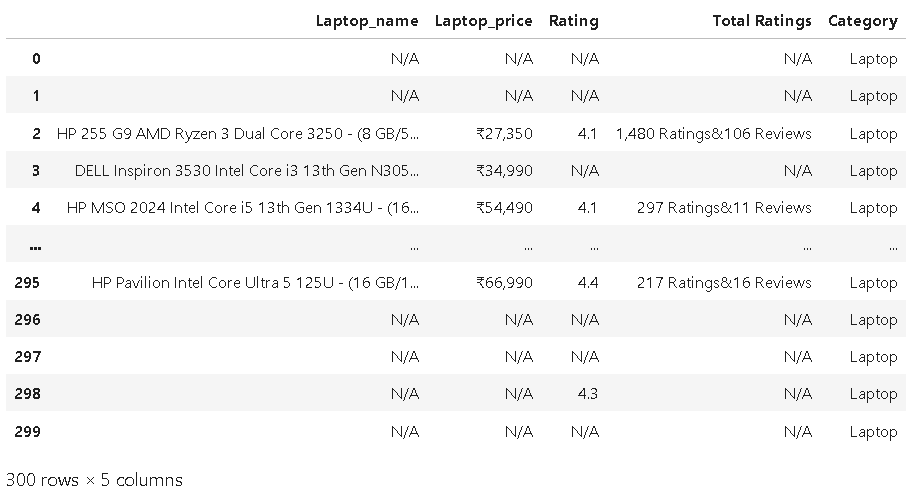
Code :







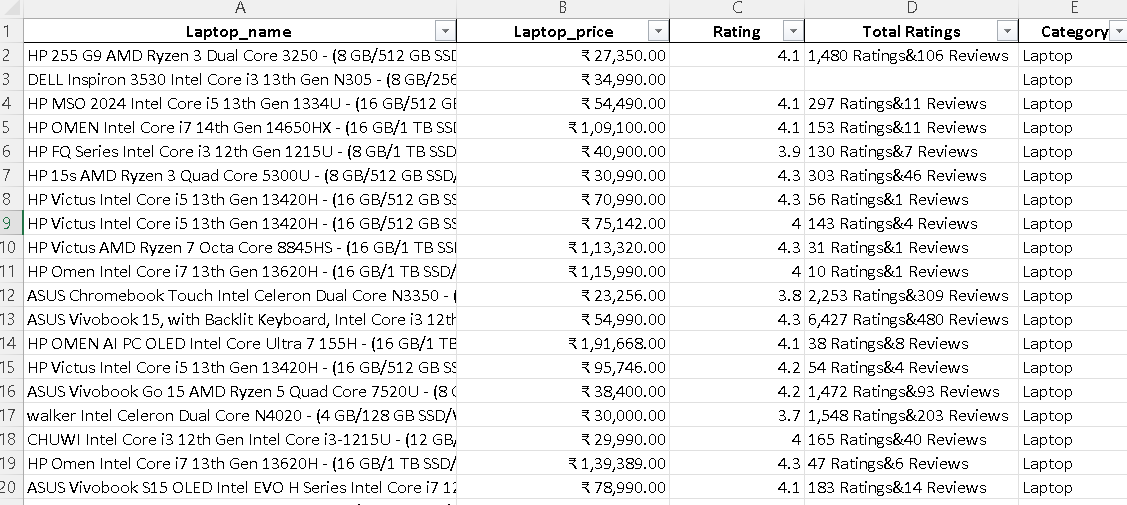
Data :



Data Storing:

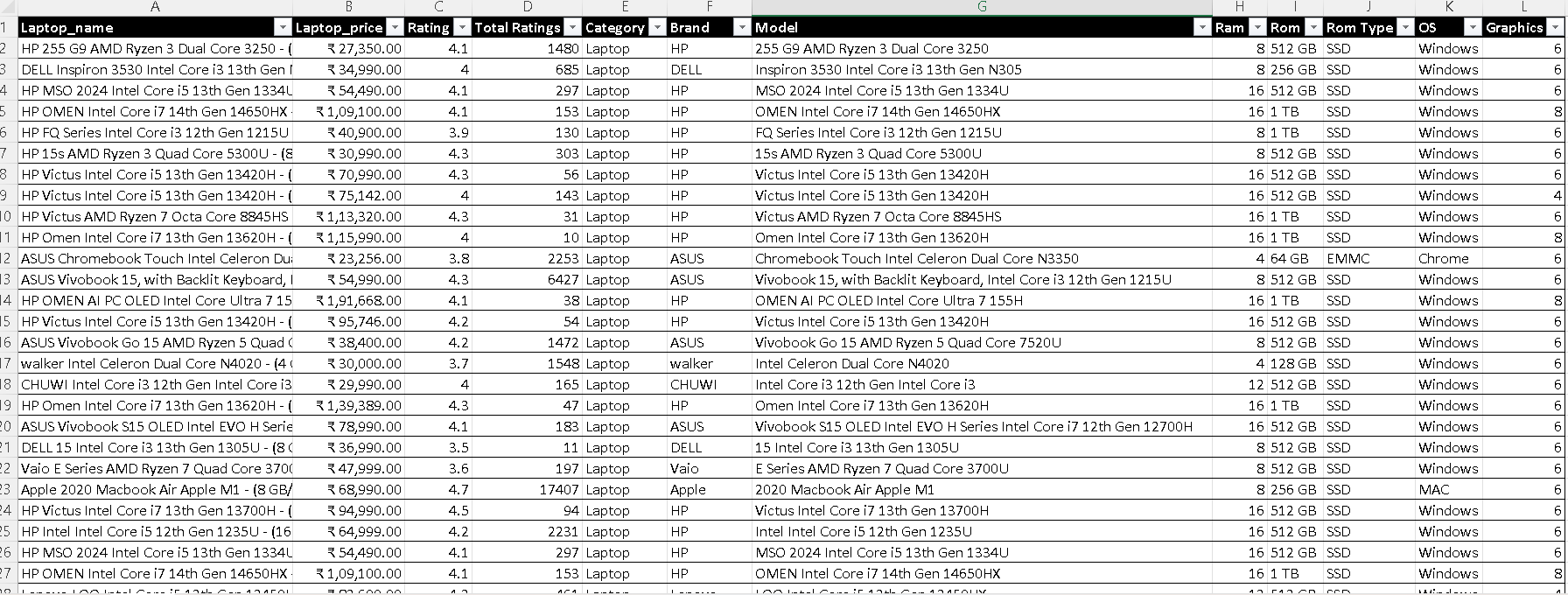


Data in excel:



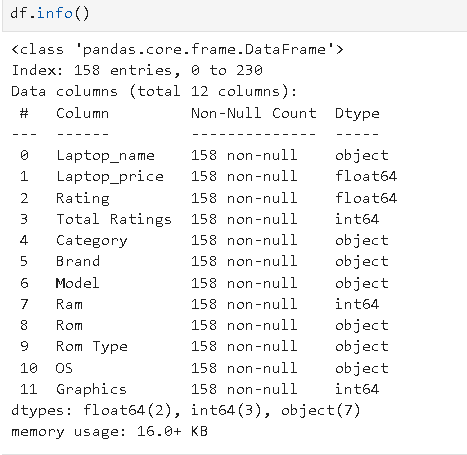
To save the data to a SQL Database, a database named ‘Capstone’ was created and data was imported using import data wizard after certain splitting and cleaning.

Final cleaned data:

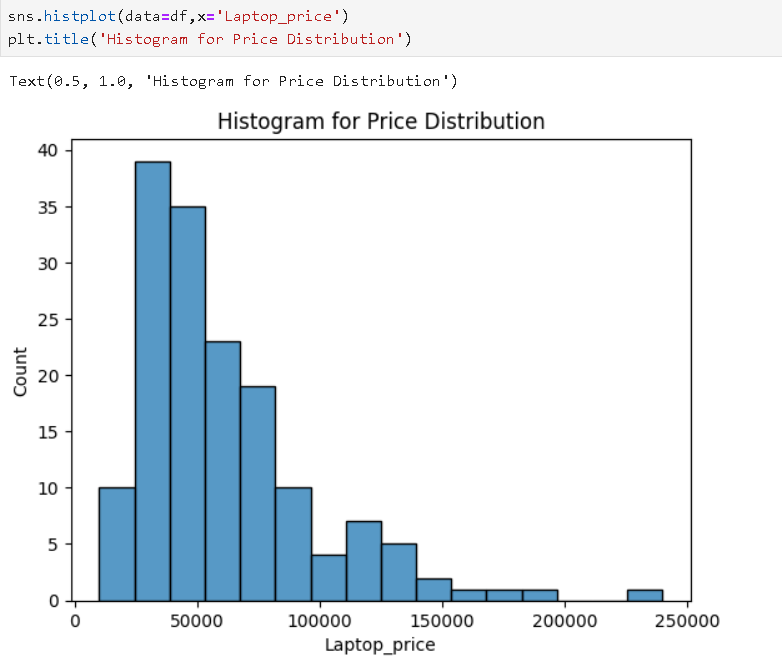


EDA:

Data information:

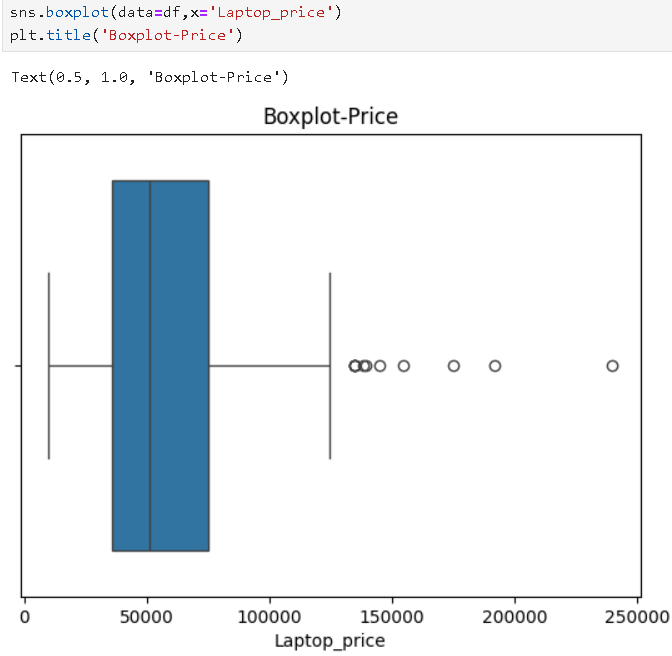


Distribution of Product Price:



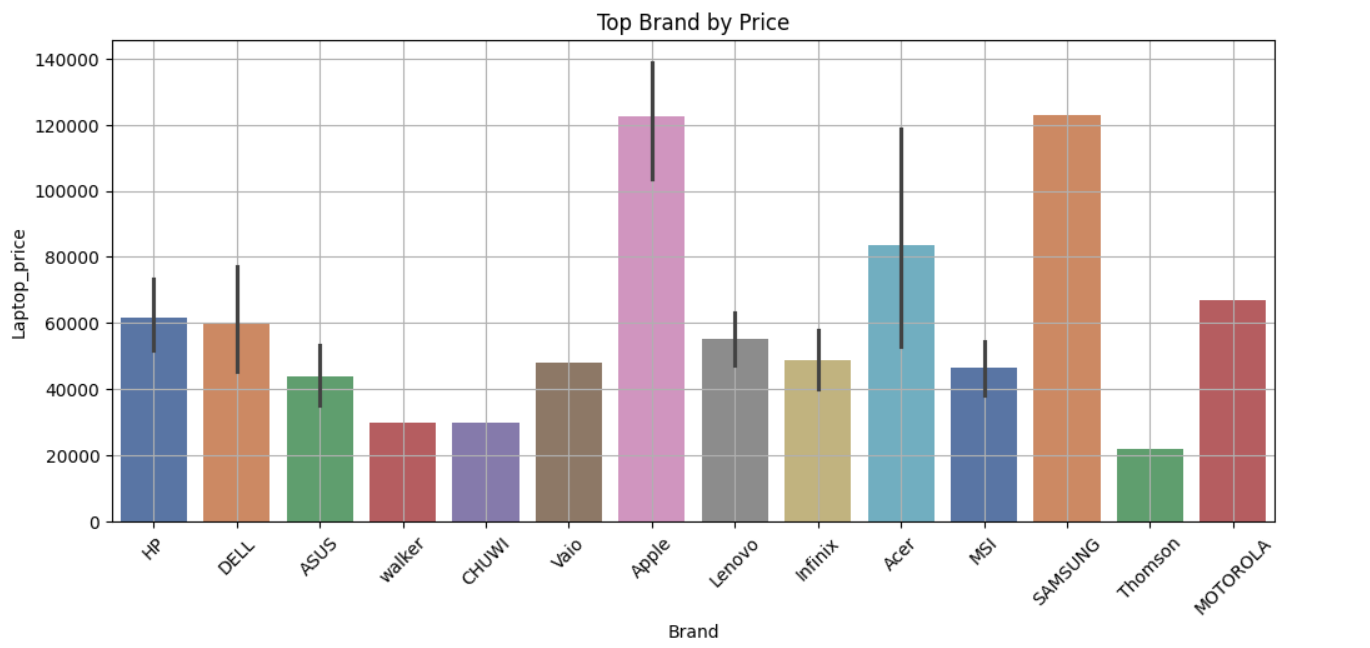
Only few outliers. Since its laptop price some high-end products can have higher pricing,

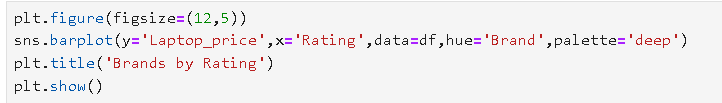
So, no need of handling them

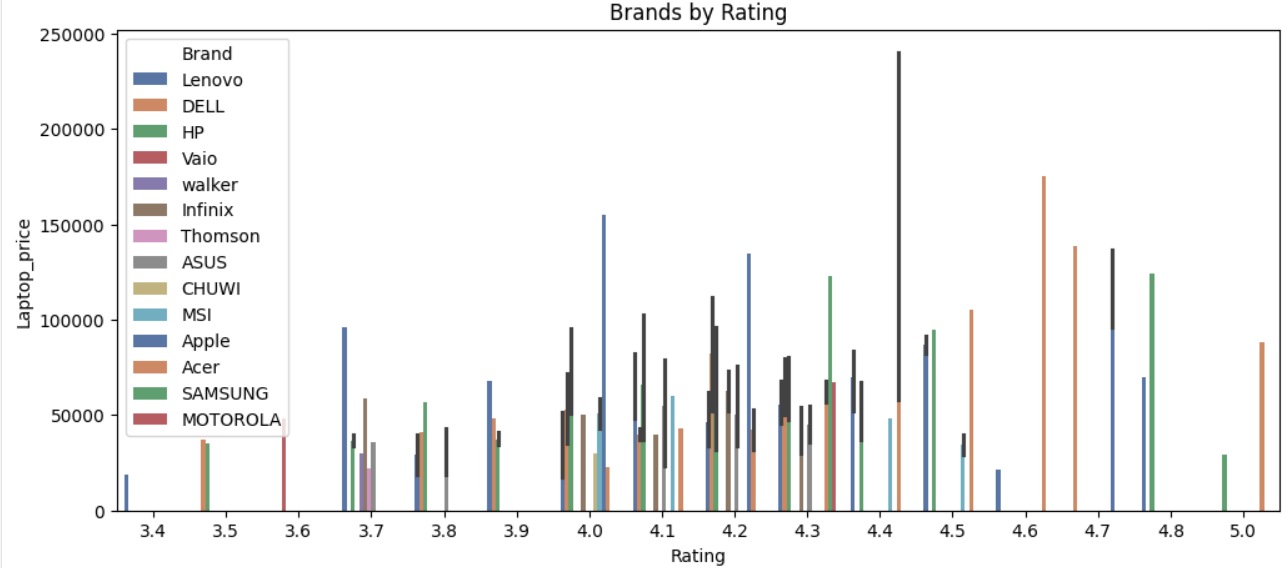


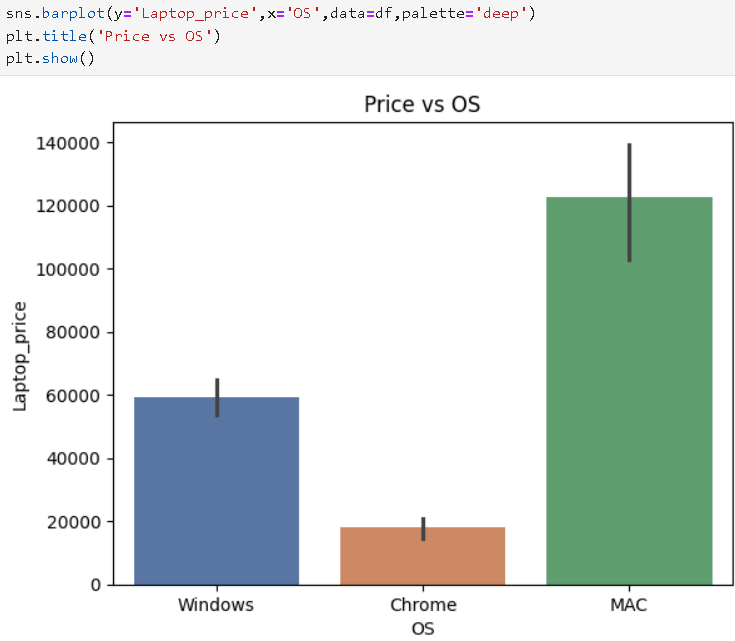
Analysis by Visualization:

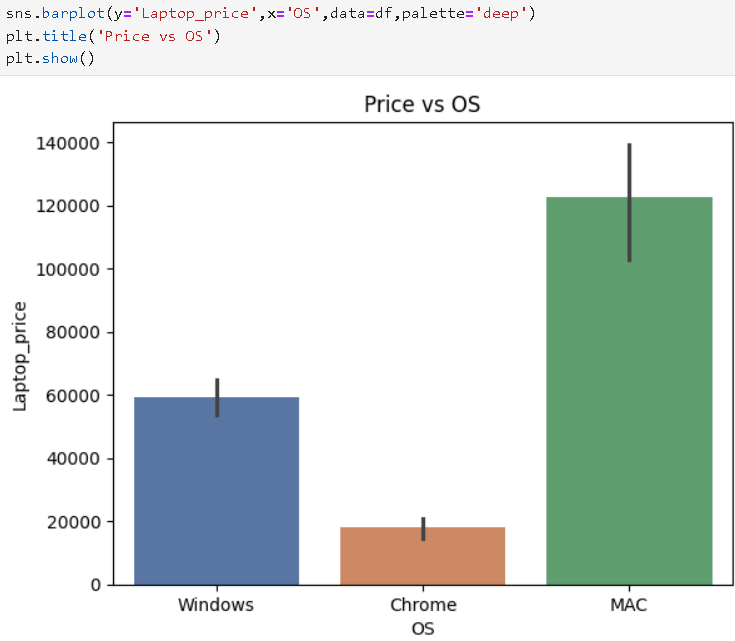


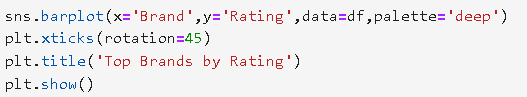


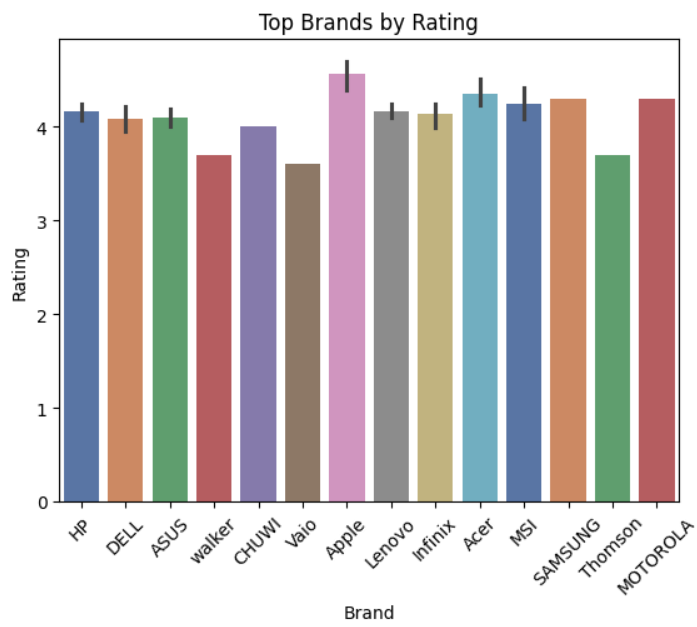


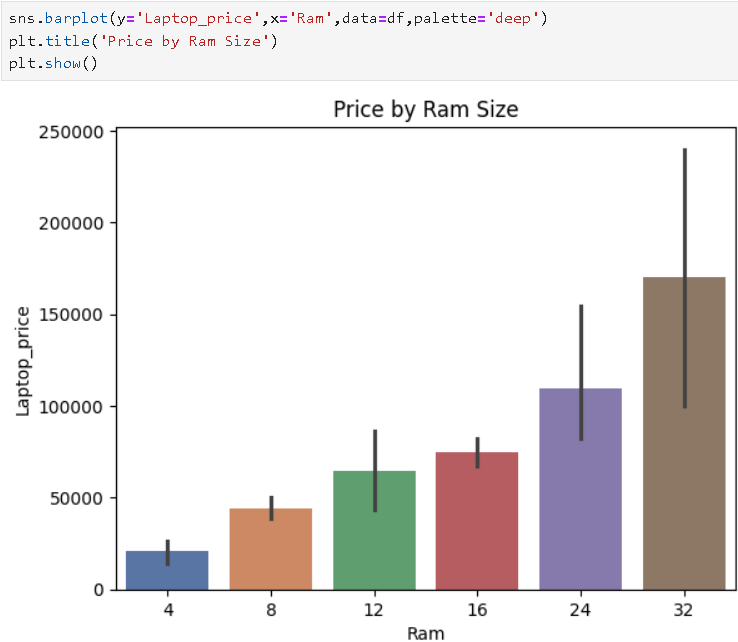


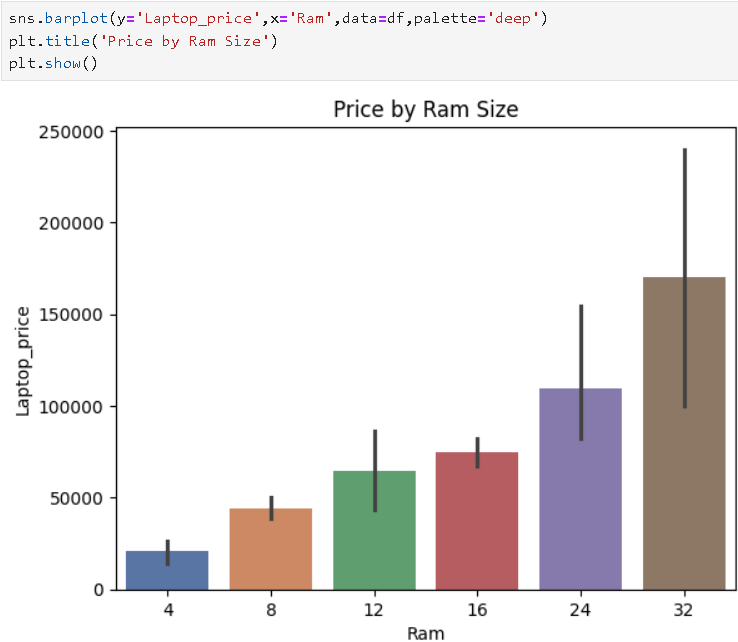


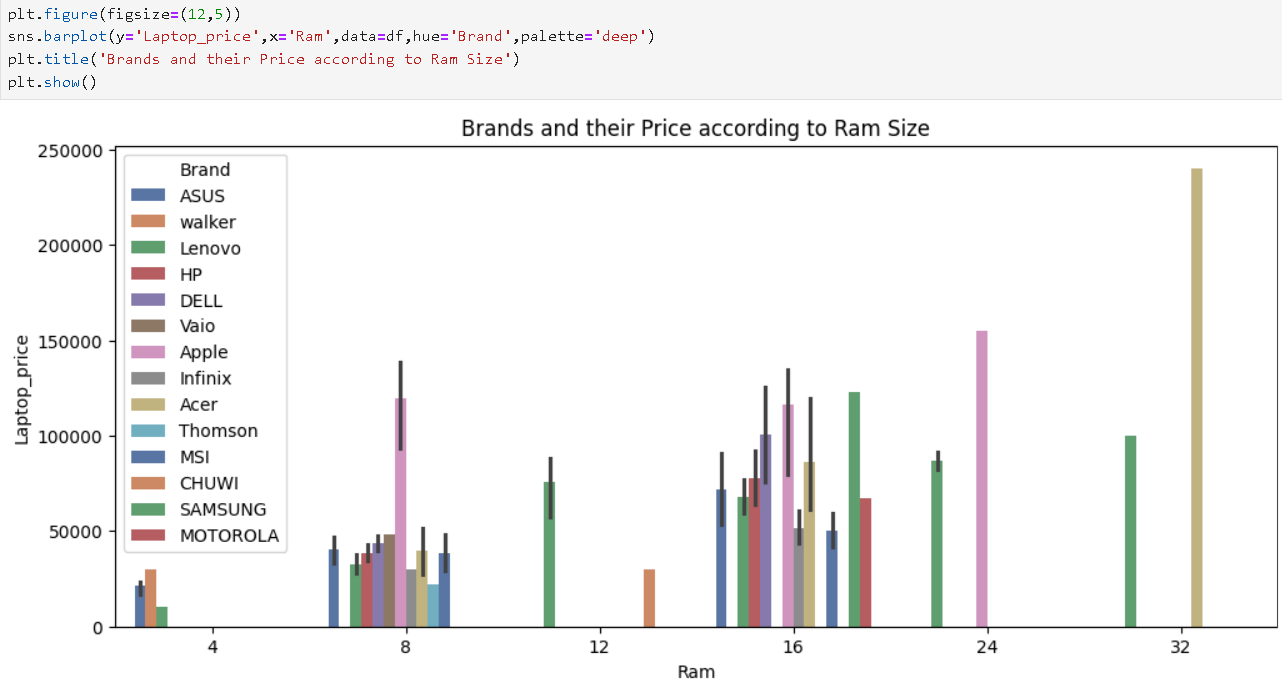


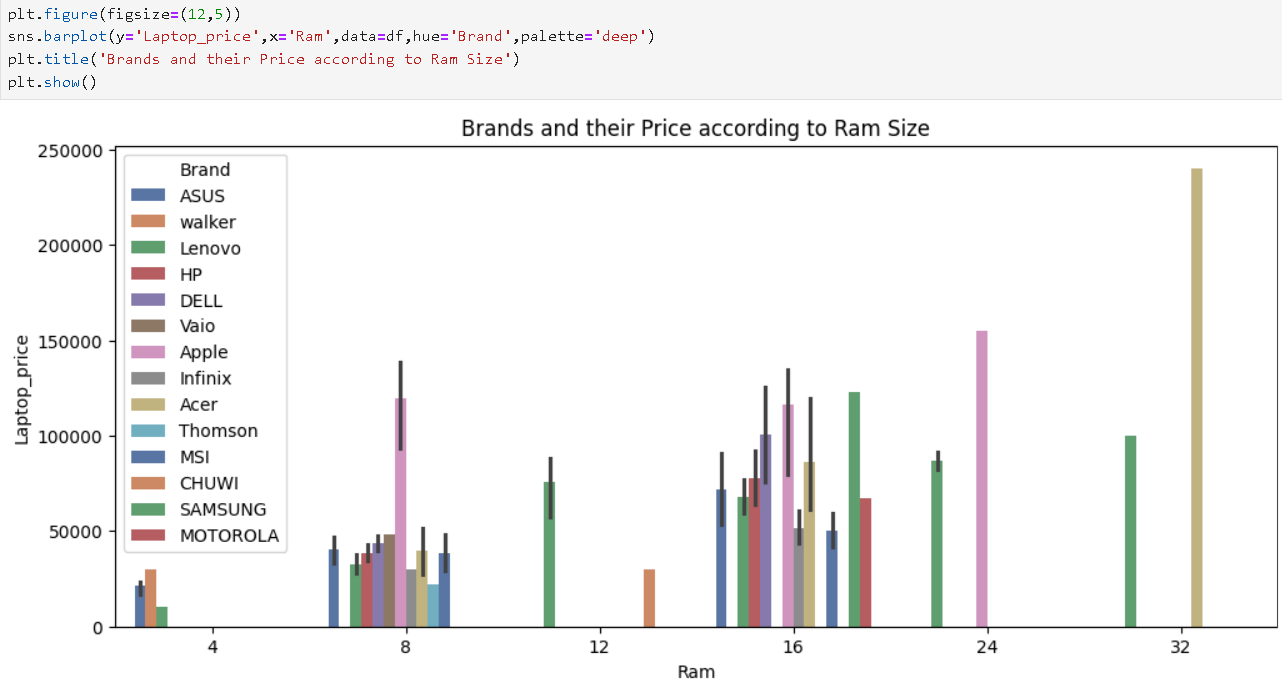


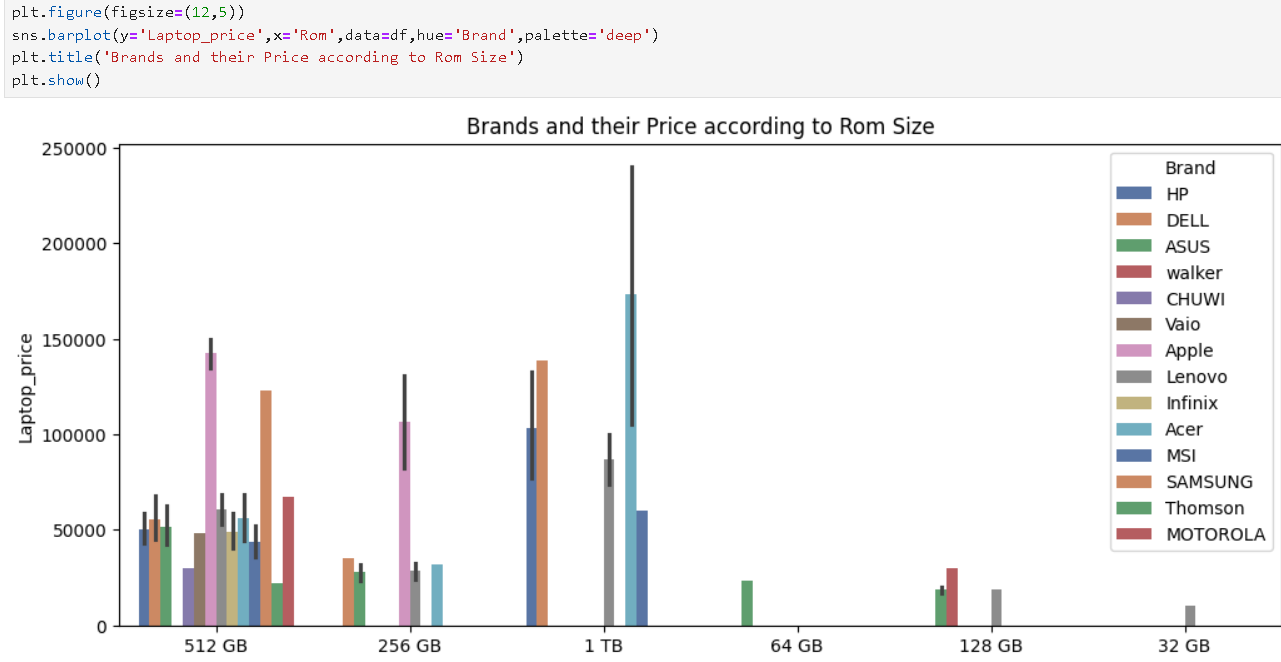


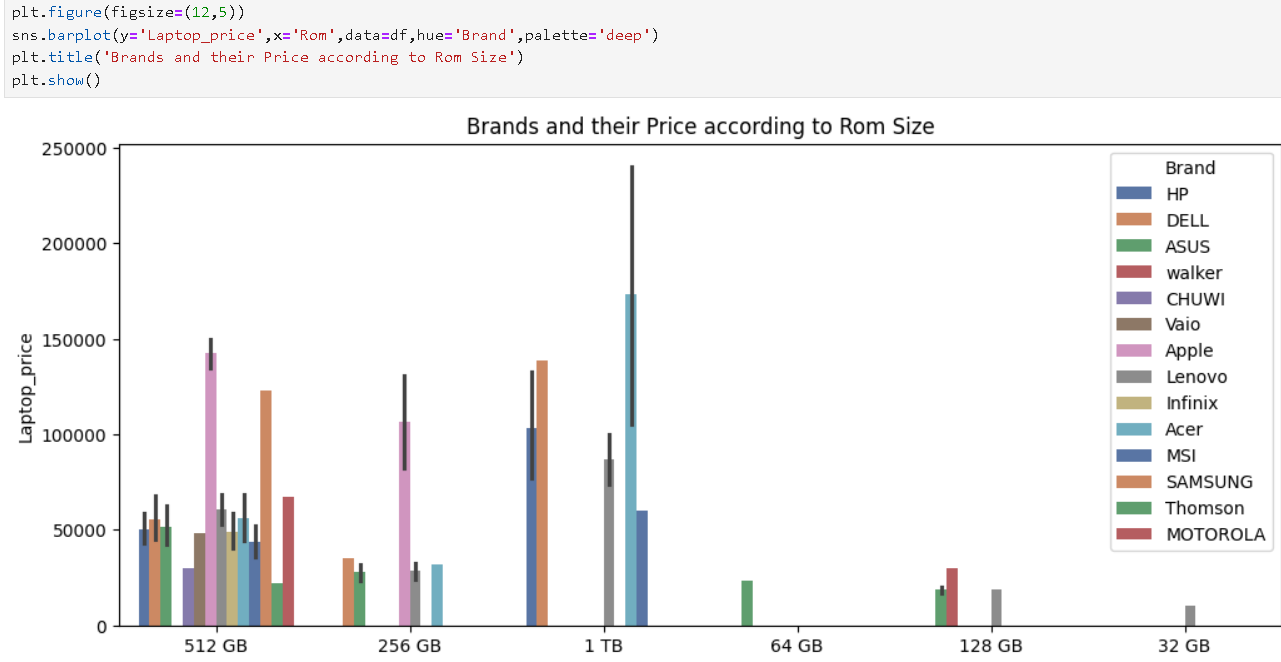




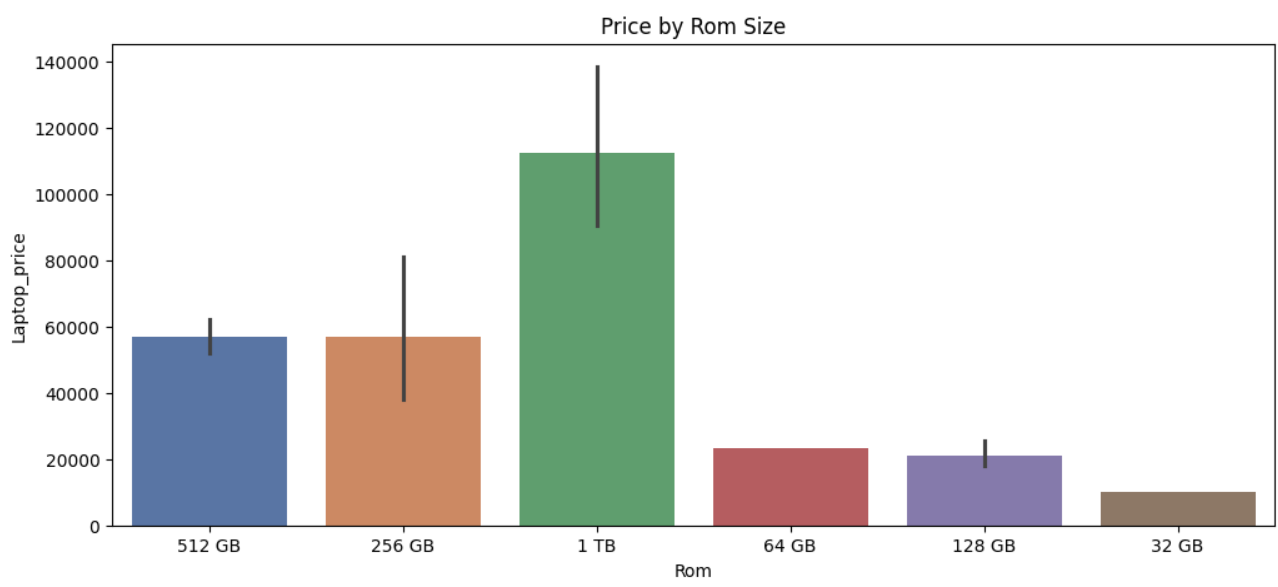


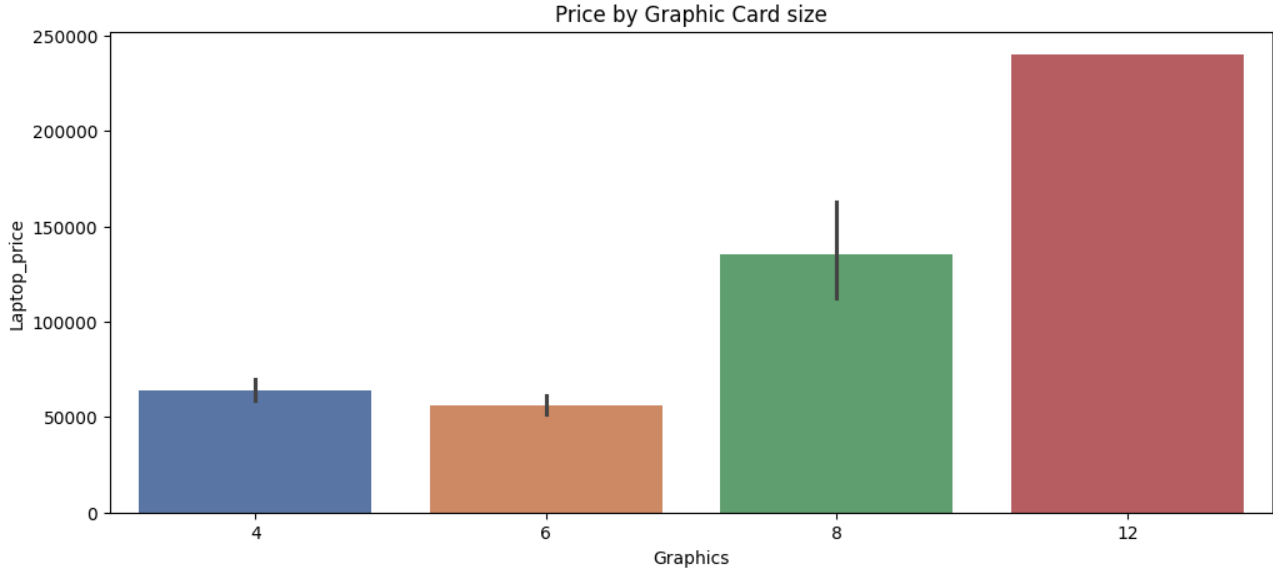


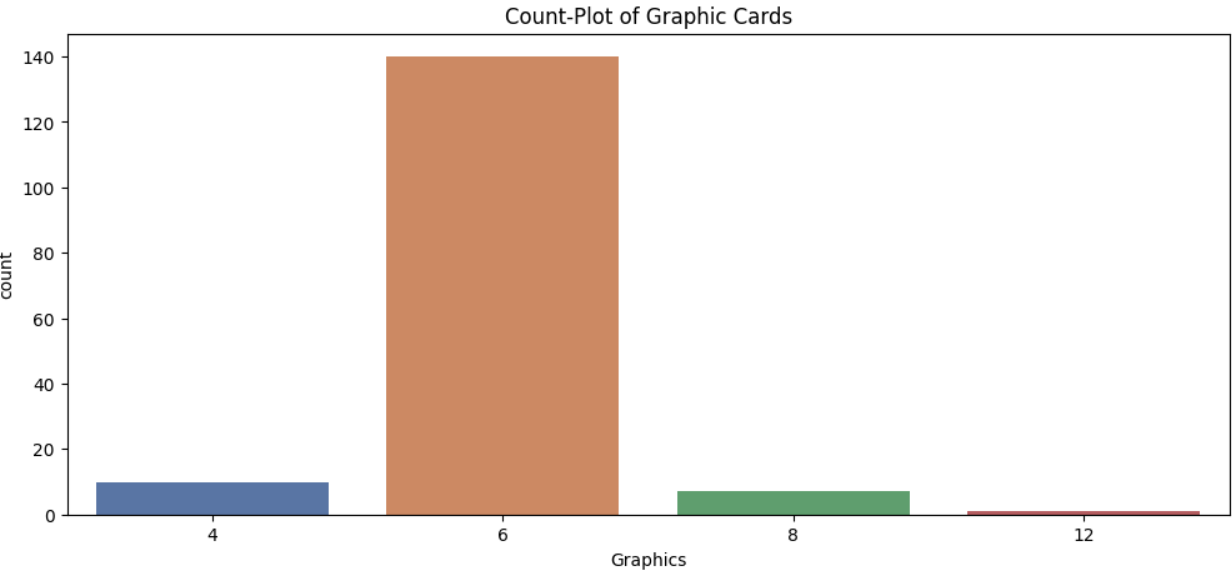


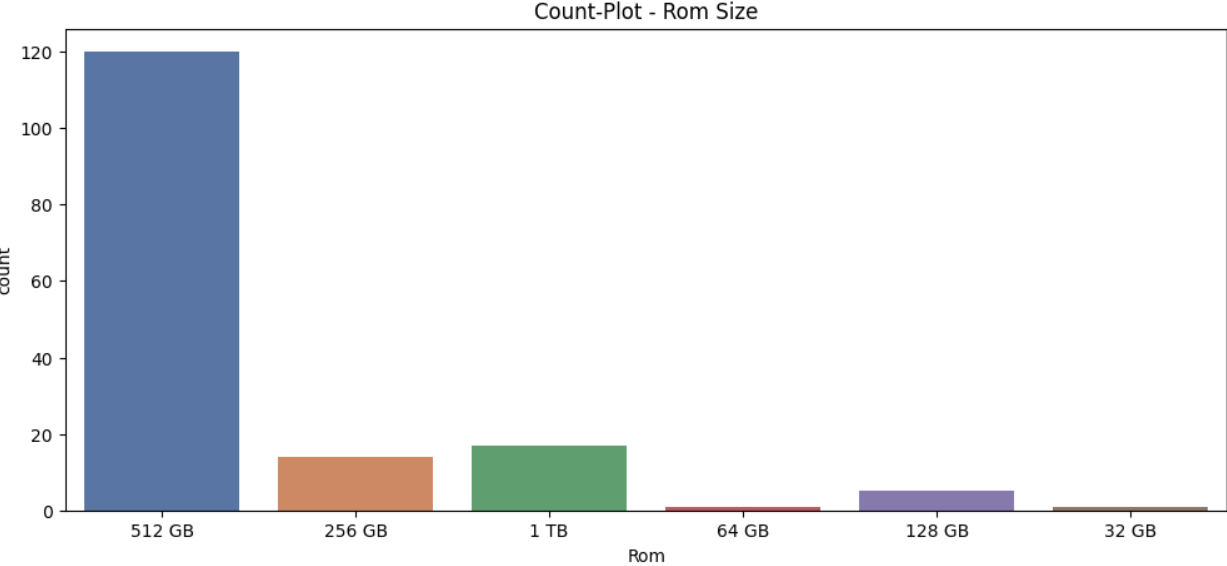


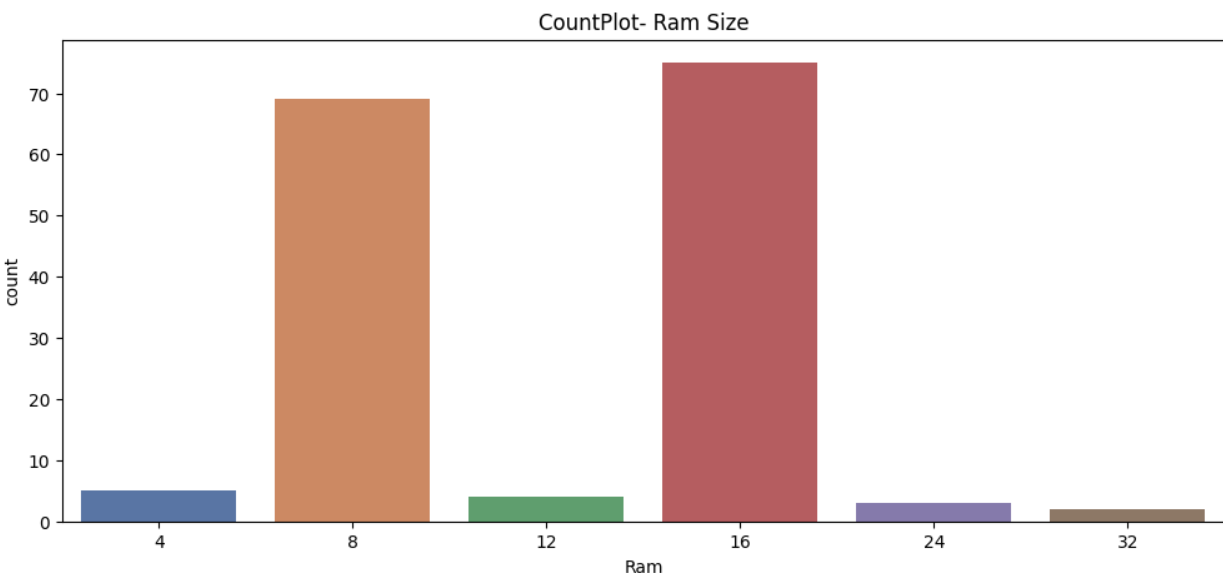
**More Plots:**

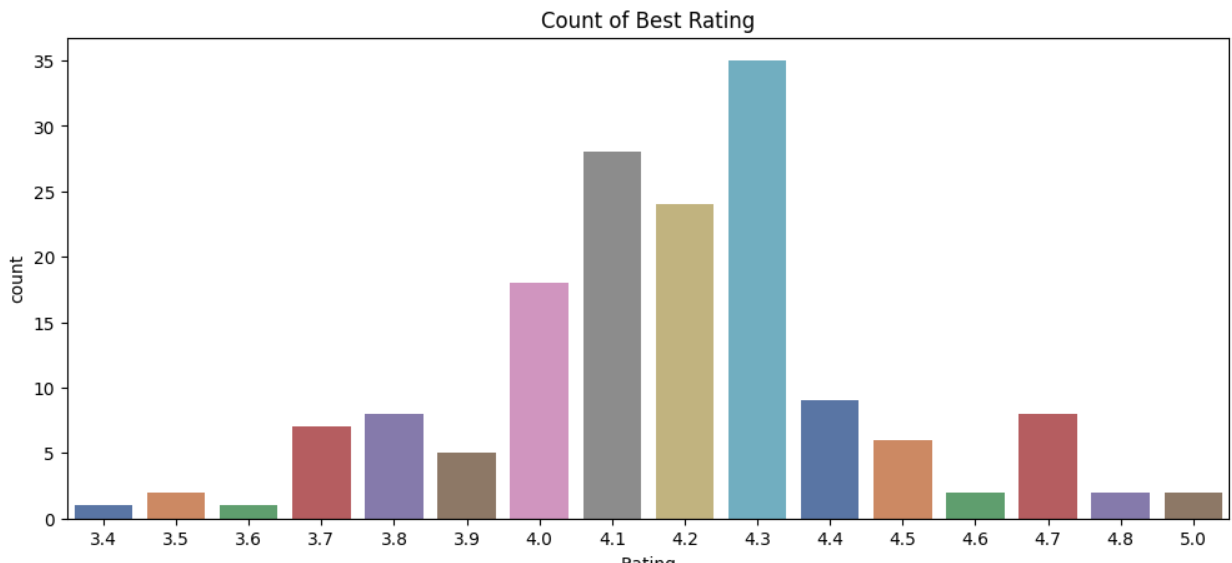


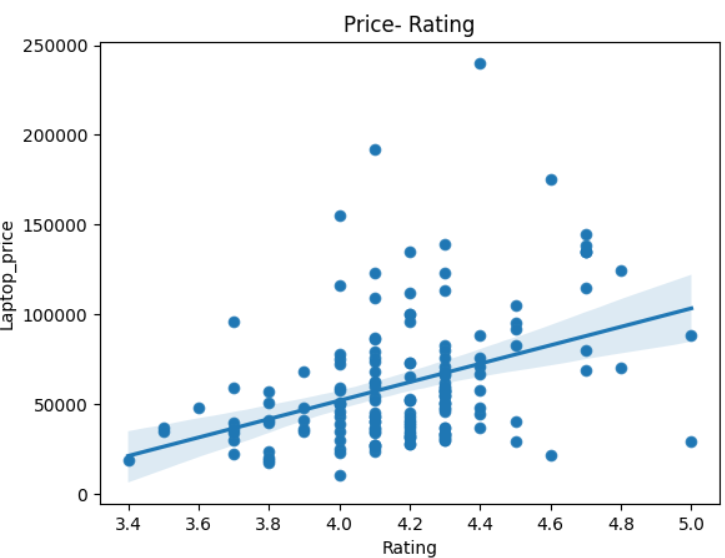




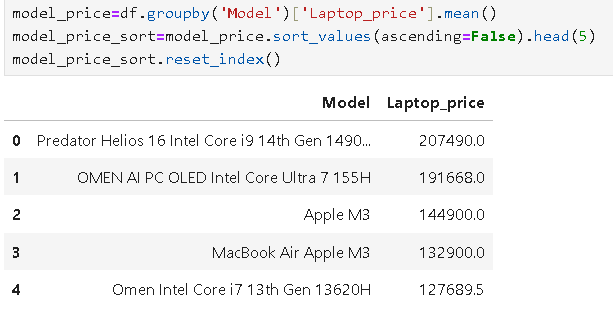




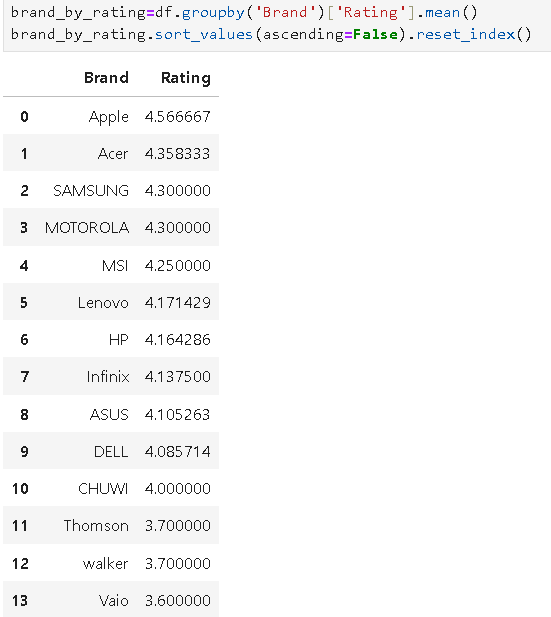




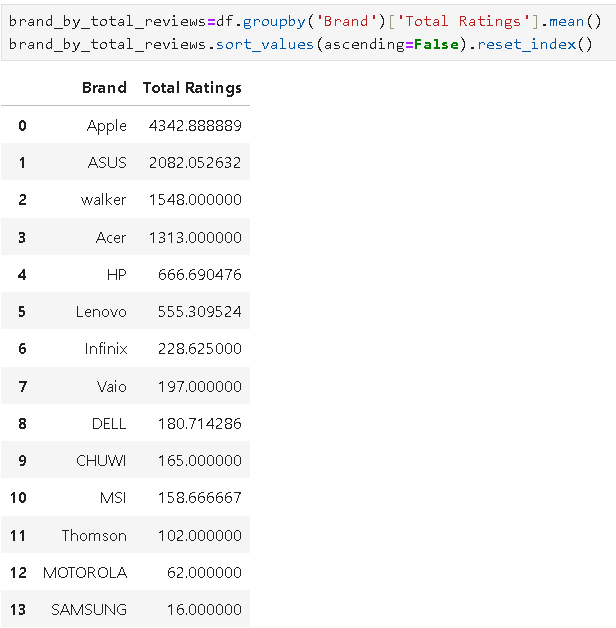
Highest priced laptops:



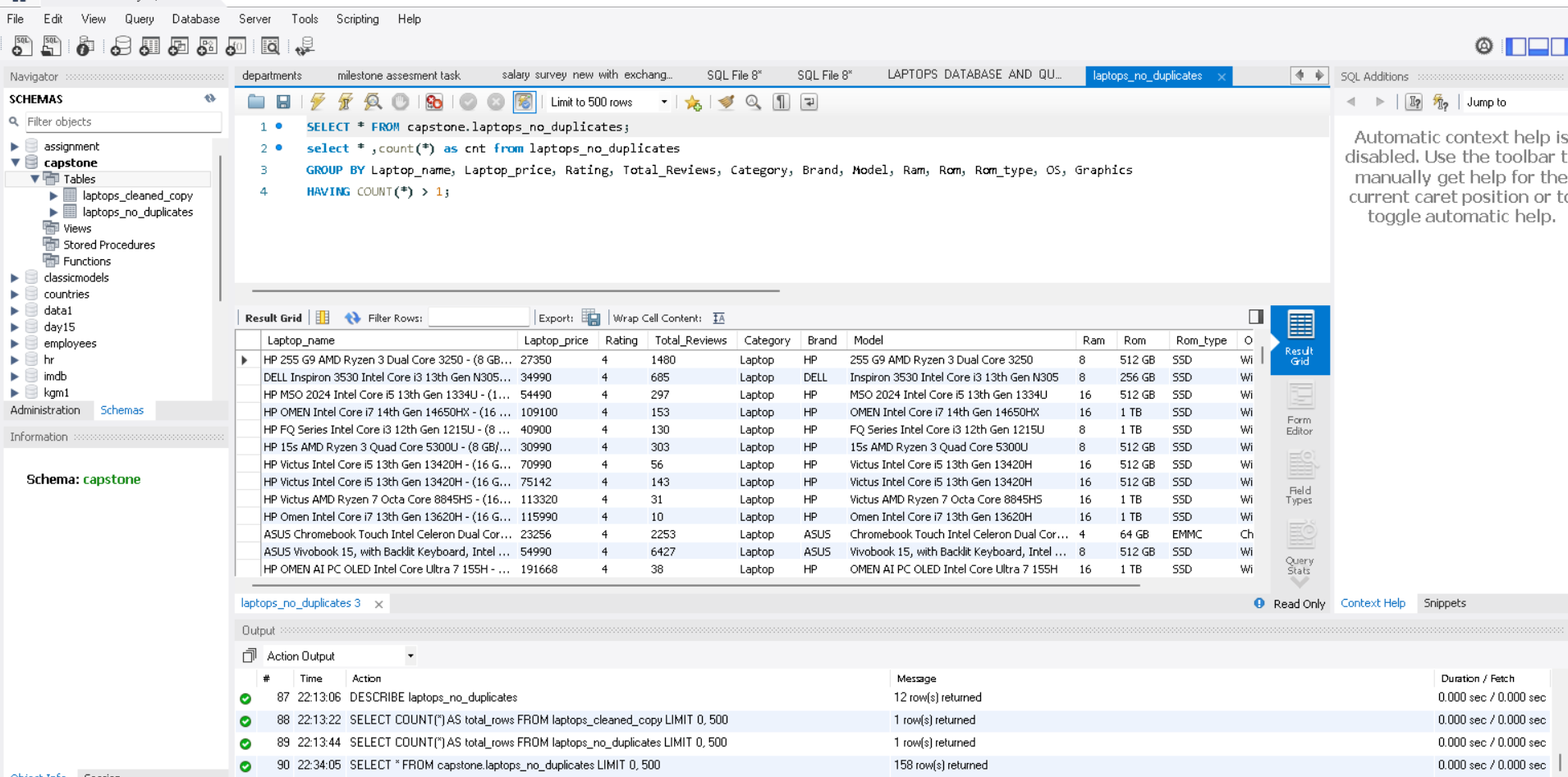
Best Brands by Customer Ratings:

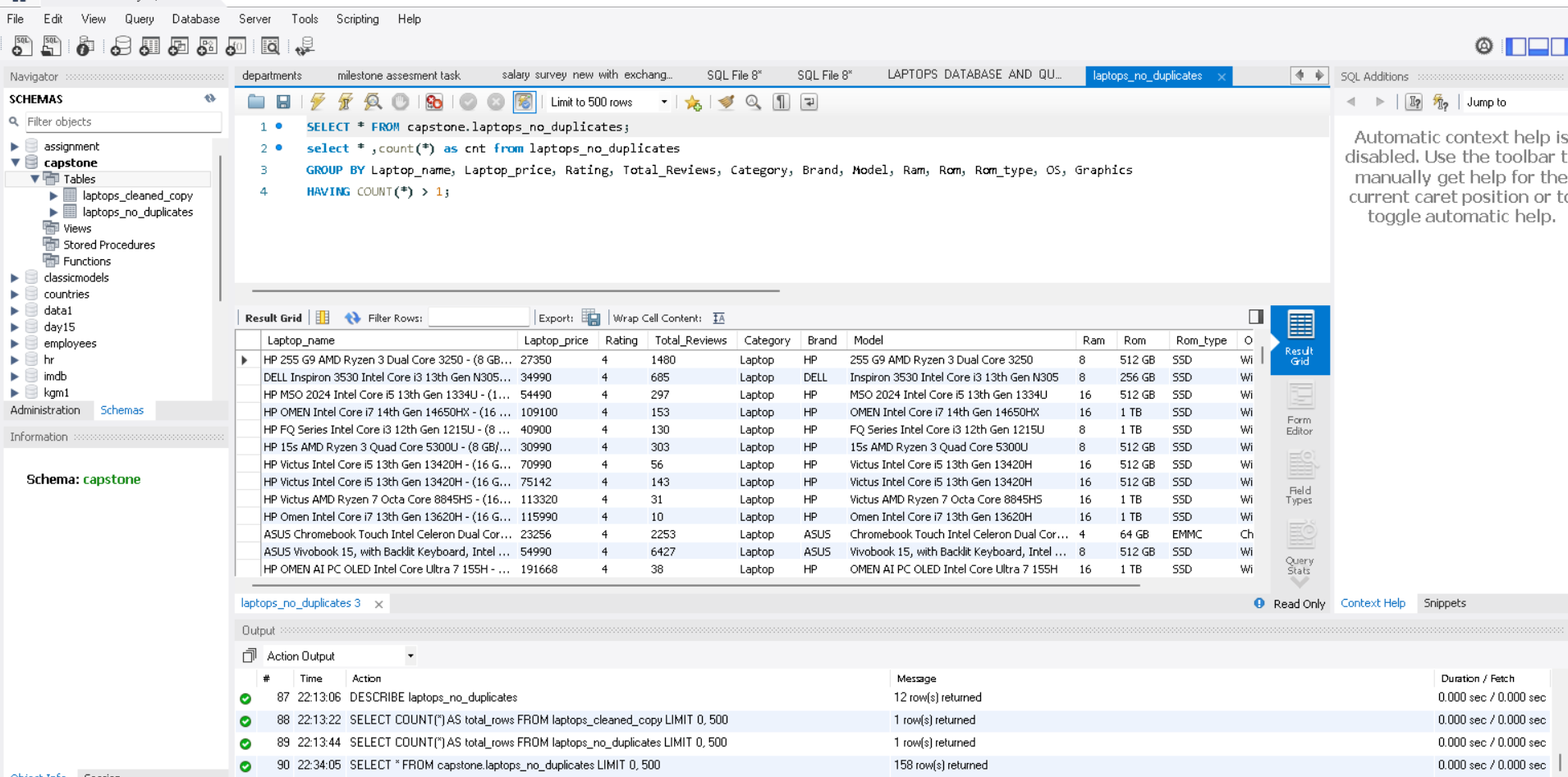


Most viewed or used Brands:



SQL DATABASE :



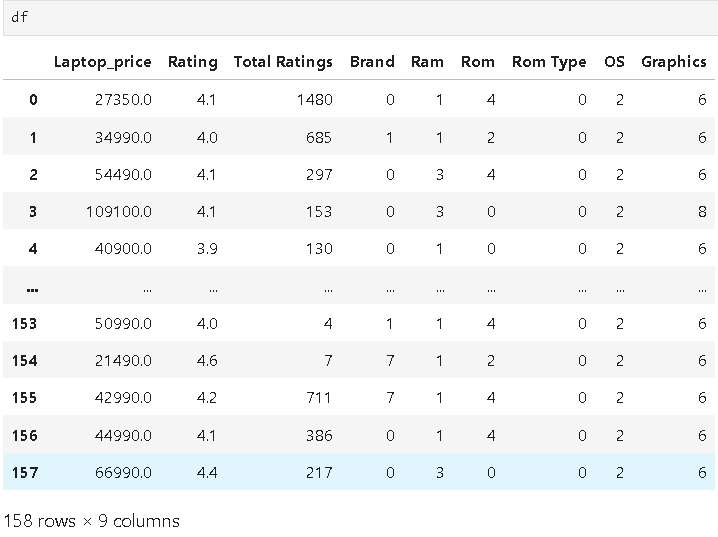


**Supervised and Unsupervised Learning:**

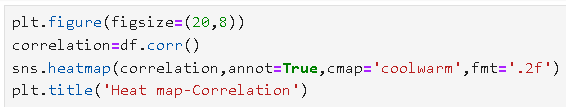
****

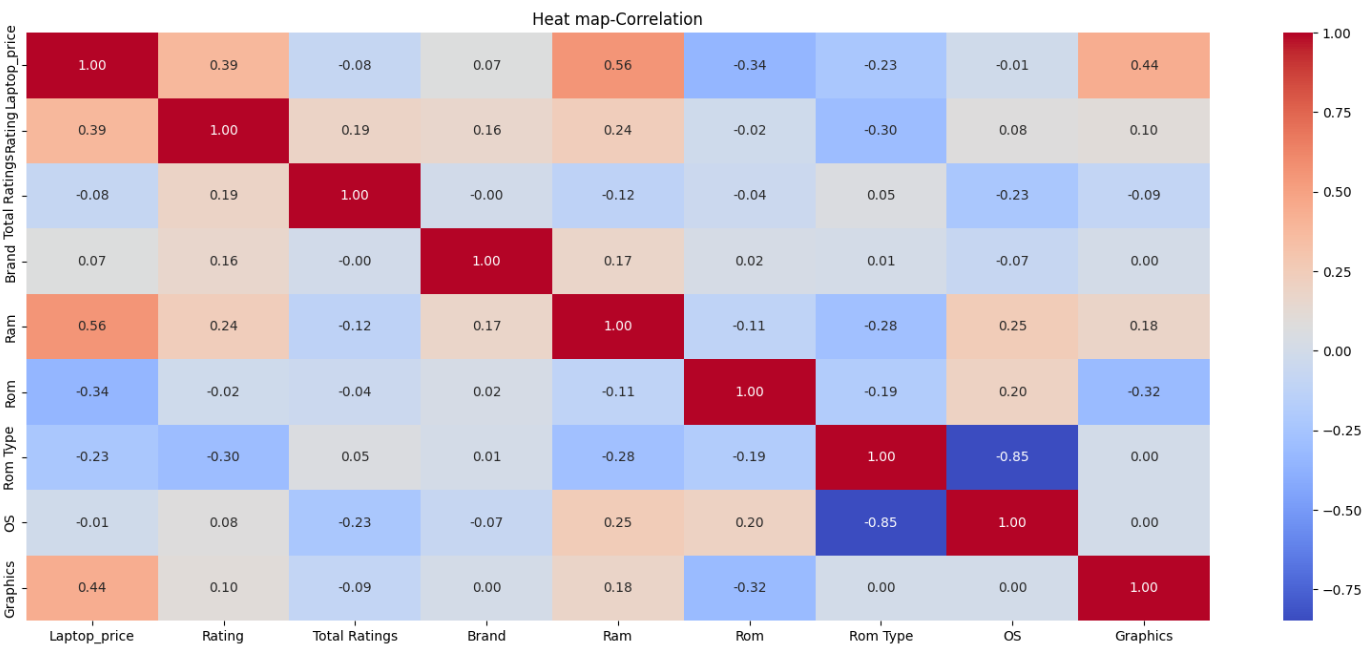
Label encoding was used for encoding Ram, Rom and OS, since they are ordinal. Brand names and Rom Type were mapped using ‘map’ function . Another encoding was also created where one hot encoding was used for brand and label encoding for rest. The results were similar for both models.

**Encoded DF:**

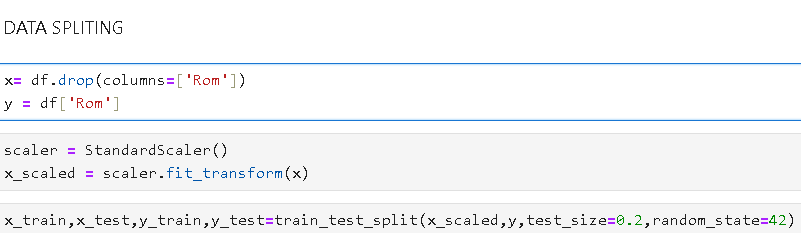


**Correlation Heatmap:**

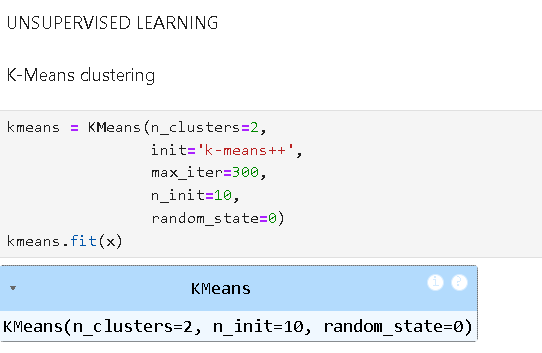




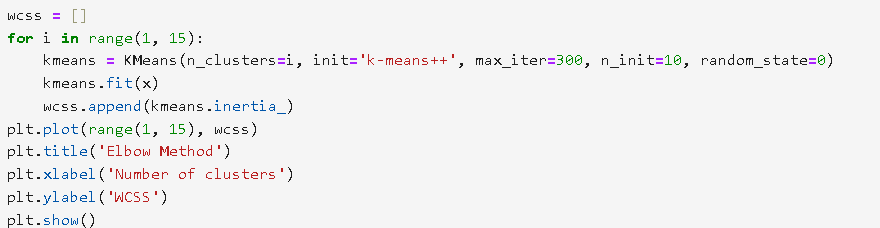
**Data splitting:**

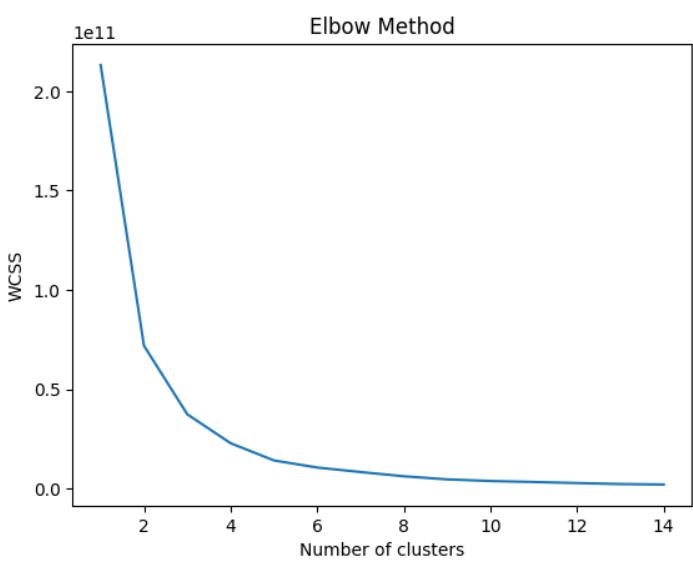


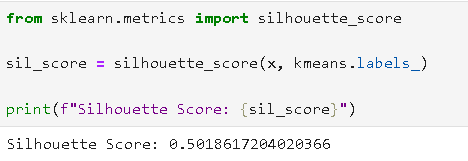
**Unsupervised Learning Model- K Means Clustering:**



**Finding number of clusters using elbow method and silhouette method:**



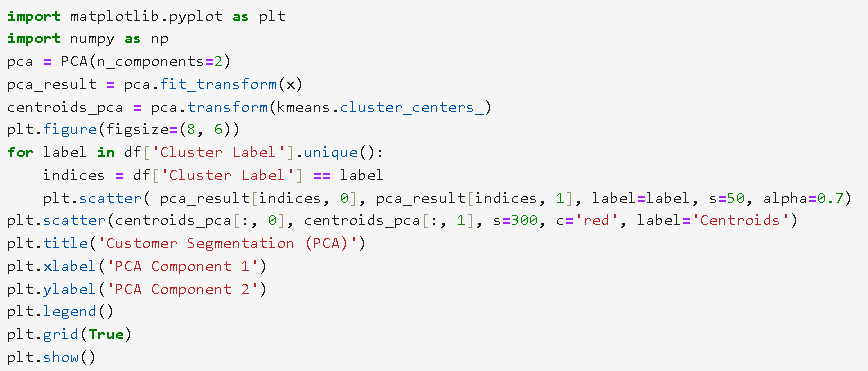




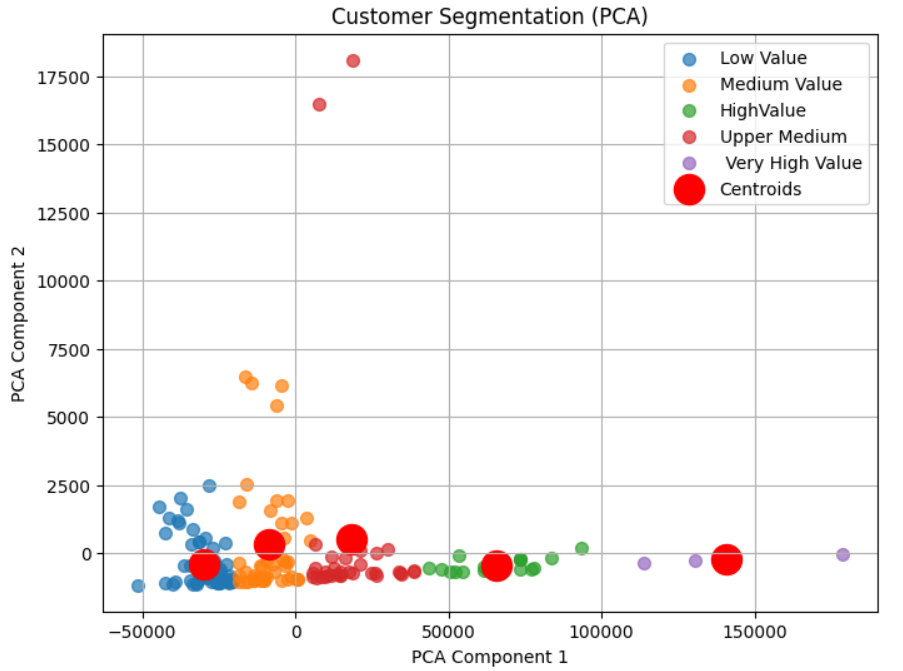
**So, the number of clusters is set to be 5.**

**Visualizing Clusters:**

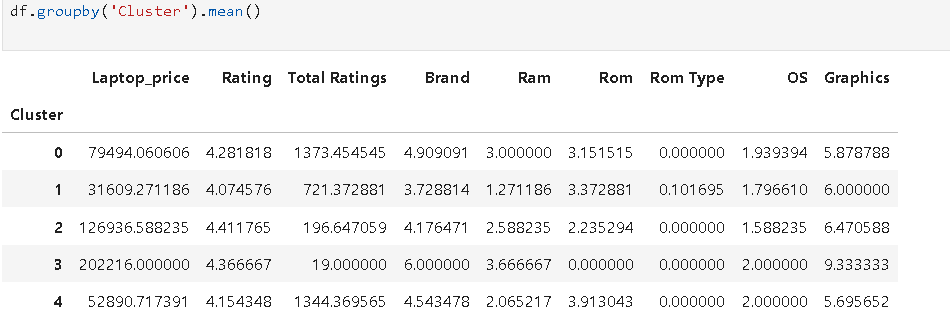
**Code ; PCA is used to handle multiclass.**

****

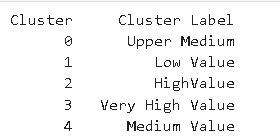
**Plot:**

****

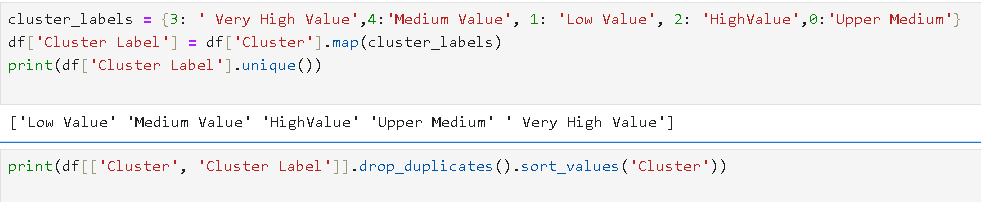
**Clusters:**

****

**Cluster Labels:**

****

**Code :**

****

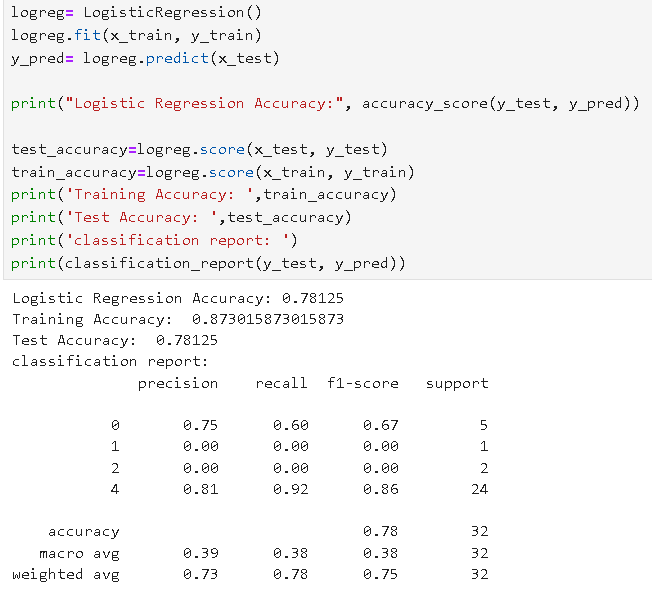
There are 5 centroids which are well clustered. Only two points lie outside which can related with outliers.

Clusters are not well-separated in PCA space. A few groups overlap, suggesting either:

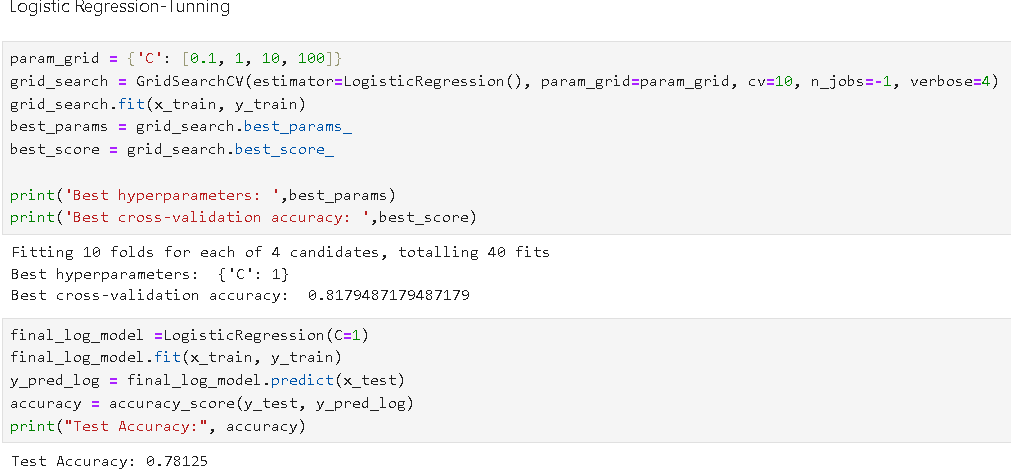
* The true separation lies in more dimensions than just 2
* Or some clusters are naturally closer (e.g., laptops with similar specs but different brands

**Supervised Learning Models:**

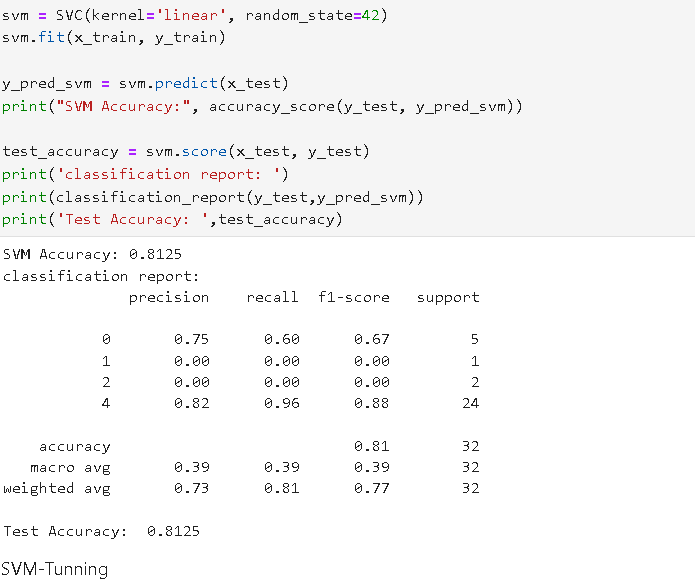
**Logistic Regression:**



* Hyperparameter Tunning:



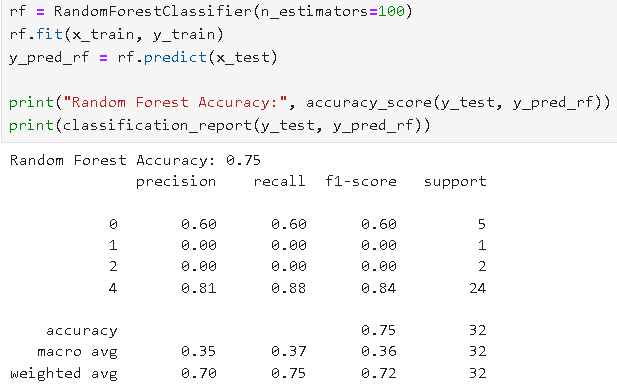
SVM Model:



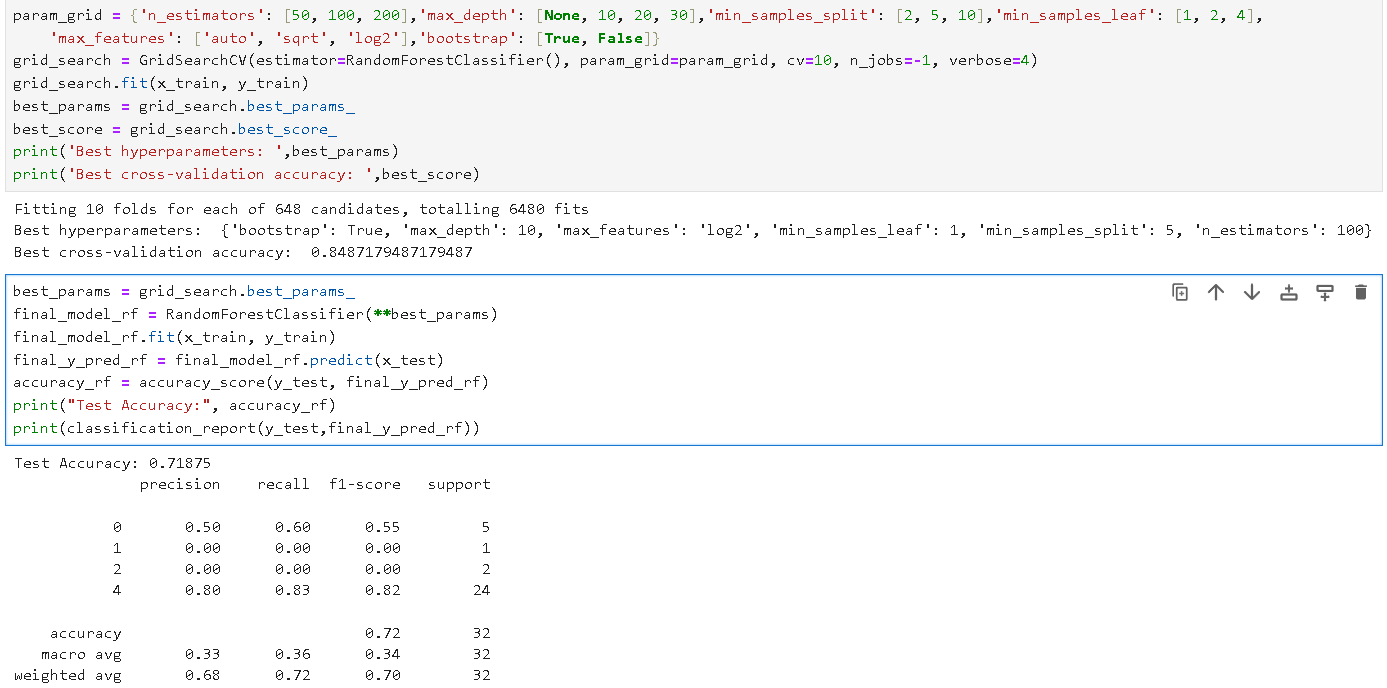
* Hyperparameter Tunning



Random Forest Classifier Model:

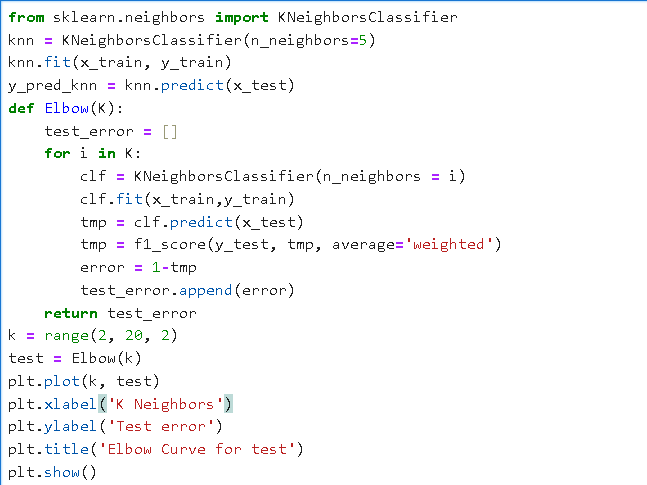


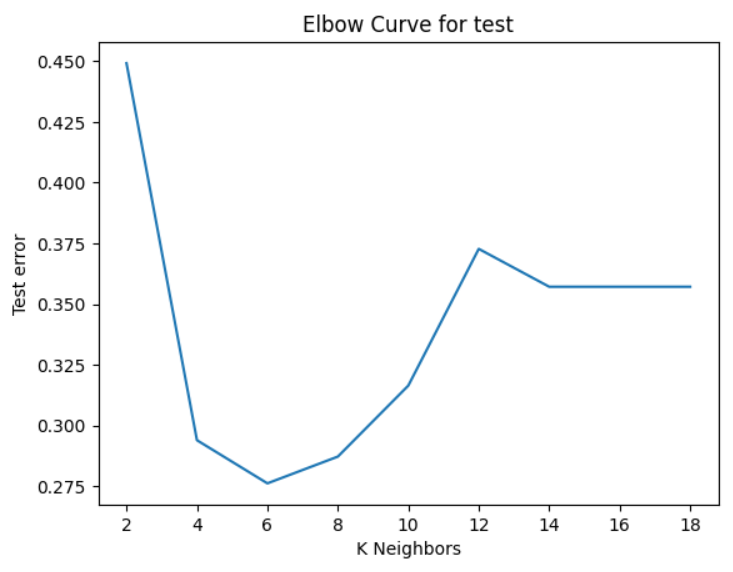
* Hyperparameter Tunning:

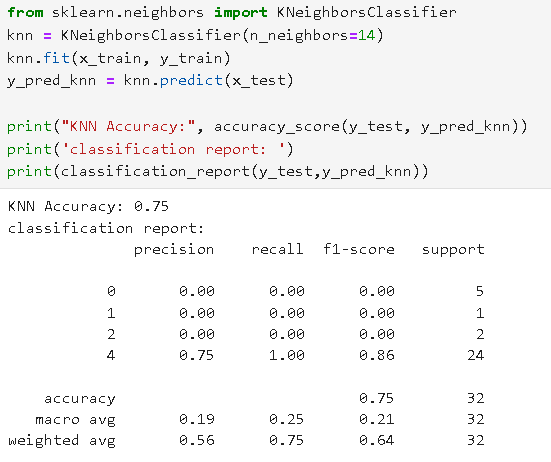


**KNN-Model:**

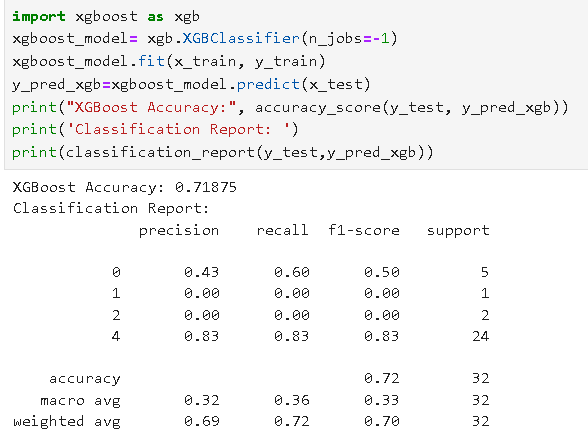
**Code:**



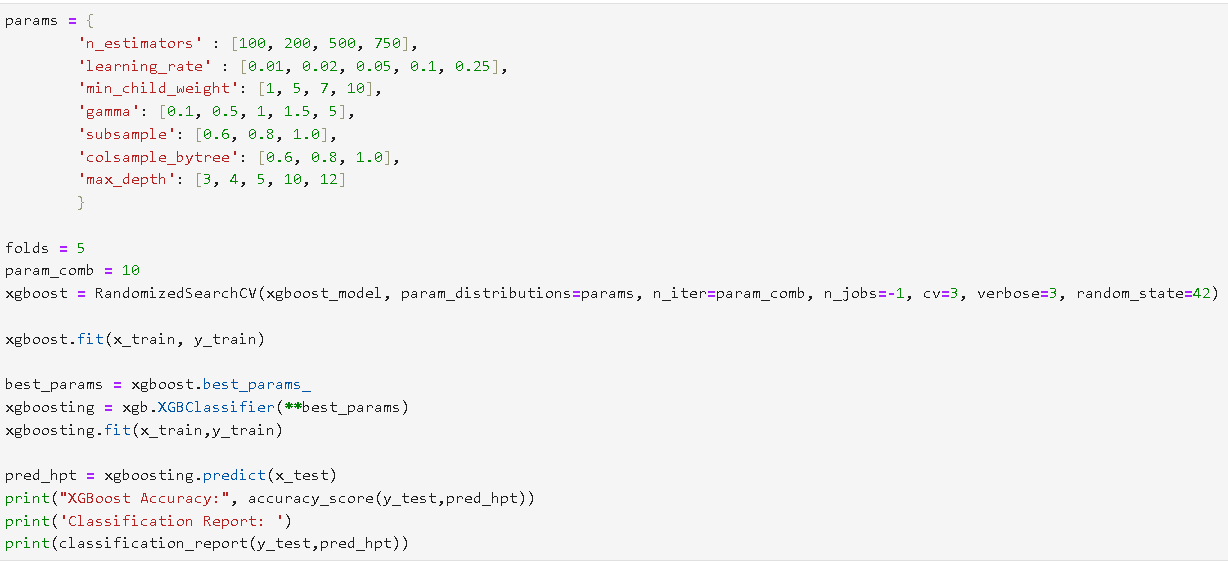


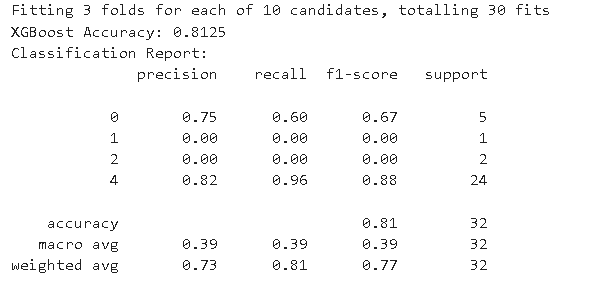


**XG Boost Classifier:**

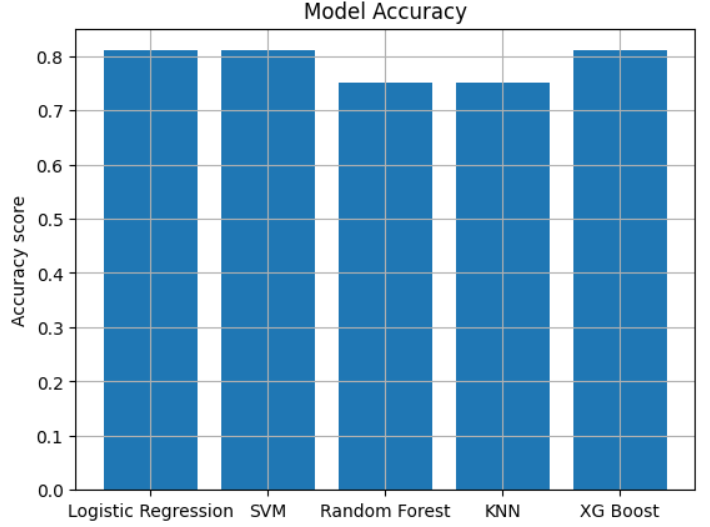
****

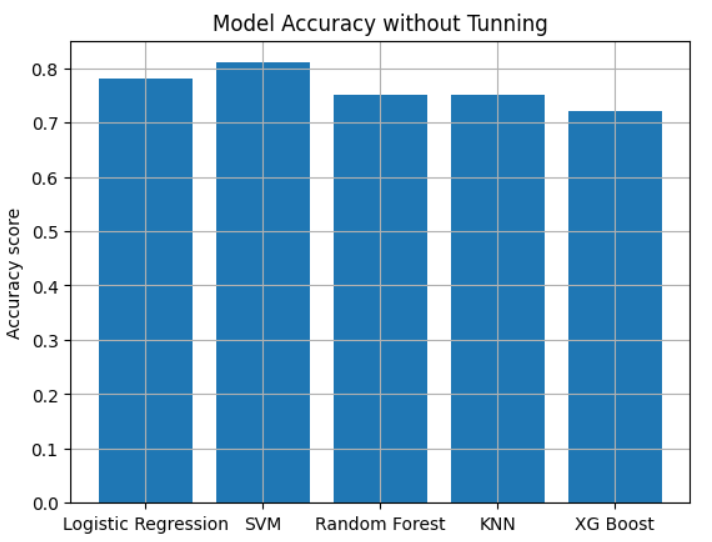
* **Hyperparameter Tunning:**

****

****

**MODEL ACCURACY:**

****

****

**EDA Highlights:**

* Price range mostly lies between mid-range (30k-60k), with few high-end outliers
* HP, DELL, and ASUS were the most frequent brands
* SSD is the dominant ROM type
* Windows OS dominates over Chrome and MAC OS
* Positive correlation seen between RAM, ROM and Price
* SSD is a key differentiator in pricing
* Brands like Apple and ASUS dominate high-end segments
* Most laptops cluster around Windows + SSD + 8/16 GB RAM configurations
* Apple , Samsung and Acer are top brands in the high-priced segment
* MSI,LENOVO AND ASUS are budget brands
* Mac OS is highest priced
* Apple , Acer ,Samsung, MSI and HP are the highest rated brands by customers
* Vaio, Walker and Thomson are least rated
* Apple has the highest price among most range of Ram and Rom
* Lenovo has budget price with larger Ram size ( also available in most ranges)
* 6GB and 4GB graphic cards have similar price
* 256 GB and 512 GB- 64GB and 128GB Rom’s have similar price
* So, opt for 6 GB and 512 GB when looking for budget laptops with high end features
* Highest priced model- Helios 16 intel core i9 14 the gen
* Most used/reviewed -Apple,Asus,Walker
* Avoid Walker ( Walker is most reviewed and rated very low)

**Unsupervised Learning: K-Means Clustering**

* Input: All encoded features including Laptop price; only Rom was excluded
* Preprocessing: Label encoding for ordinal, One-hot for nominal
* No scaling was used, since price is an important natural differentiator
* Silhouette Score:
  + With StandardScaler: -0.2 (bad clustering)
  + Without Scaling: 0.5 (good natural segmentation)
* Optimal Clusters Found: 5 (based on Elbow + Silhouette methods)
* Visualization: PCA projection in 2D with labelled centroids
* Insight: Clusters represent price + spec-based segments; slight overlap due to brand/spec similarities

**Results Summary:**

* Most models achieved comparable accuracy
* XGBoost significantly improved after tuning and matched SVM
* Logistic Regression and SVM showed strong baseline performance

**Hyperparameter Tuning**

* GridSearchCV and RandomizedSearchCV used
* Tuned parameters: n\_estimators, max\_depth, C, kernel, learning\_rate, k (neighbors)
* Improvement observed especially in SVM and Random Forest

**Conclusion**

This capstone project demonstrates the complete lifecycle of a data science project—from data acquisition and preprocessing to advanced machine learning modelling. By scraping real-world data from Flipkart, insightful patterns were uncovered about laptop brands, specifications, and pricing. Clustering helped segment the market into interpretable consumer groups, while supervised learning effectively predicted product categories. This project reinforces the value of applied machine learning in real-world business contexts, particularly for e-commerce product analysis and customer segmentation.

Few recommendations from analysis is to reduce visibility of low-rated brands or pair them with incentives (like discounts or customer education). Prioritize high-rated brands in search listings. Look for Lenovo/MSI models with 8GB+ RAM and SSDs for best performance-to-price ratio. Unless you're in the Apple ecosystem, Windows laptops offer more features per rupee.