EAST WEST UNIVERSITY

CSE303: Statistics for Data Science [Spring 2025]

Project Report

Course Code : CSE303

Course Title : Statistics for Data Science

Section : 3

Group Number: 21.

Topic Name : Laptop Price Prediction.

Submitted by:

Student Name	Student ID
1. Yesmin Akter	2021-3-60-072
2. Sharifuzzaman Mazumder	2022-1-60-003
3. Farhana Rahman Munira	2021-3-60-153

Submitted to:

Dr. Mohammad Manzurul Islam

Assistant Professor

Department of Computer Science & Engineering.

1. Introduction

The price of laptops varies extensively based on factors like brand, specifications, and functionality. The objective of this project is to predict the price of laptops from various features through machine learning models. The demand for laptops has numerous varieties, varying in price extensively based on factors like brand, functionality, and hardware specification. Most consumers lack the proper equilibrium of price versus performance. At the same time, the manufacturers have the focus to set the price fixing which would attract the buyers at the same time without getting harmed from it. The target of this project is to predict the laptop's price from their specifications and features. We selected a data of laptop prices because it has numerous useful data like processor type, RAM capacity, storage type, and screen size, all of which have an effect on the price. From these features, we hope to identify trends that can forecast laptop prices and give useful insights to buyers and firms. This project will not only affect consumers' decisions but also educate retailers and manufacturers regarding prices in a highly competitive market. Insights or outcome we want to reach are:

•Identifying which of the features have the greatest impact on laptop prices.
• Training machine learning models that can precisely predict laptop prices based on specifications available.
• Understanding pricing patterns across different brands, models, and hardware configurations.

2. Exploratory Data Analysis

The dataset reflects the cost of laptops based on features and hardware specifications. There are 23 columns and 1276 rows in this dataset. In data analysis, features (or variables) are normally classified as categorical or numerical based on the nature of information they present. There is both type available in this dataset.

There are 14 categorical key or columns in this dataset:

1. **Company**: Laptop Manufacturer.
2. **Product**: Brand and Model.
3. **TypeName**: Laptop Type (Notebook, Ultrabook, Gaming, …etc).
4. **OS**: Operating System installed.

5. **Screen**: screen definition (Standard, Full HD, 4K Ultra HD, Quad HD+).
6. **Touchscreen**: whether or not the laptop has a touchscreen.
7. **IPSpanel**: Whether or not the laptop has an IPSpanel.
8. **RetinaDisplay**: Whether or not the laptop has retina display.
9. **CPU_company**: The name of the CPU company.
10. **CPU_model**: The name of the model, series of CPU.
11. **PrimaryStorageType**: Primary storage type (HDD, SSD, Flash Storage, Hybrid).
12. **SecondaryStorageType**: Secondary storage type (HDD, SSD, Hybrid, None).
13. **GPU_company**: The name of the GPU company.
14. **GPU_model**: The name of the model, series of GPU.

There are 9 numeric key or columns in this dataset:

1. **Inches**: Screen Size.
2.**Ram**: Total amount of RAM in laptop (GBs).
3. **Weight**: Laptop Weight in kilograms.
4. **Price_euros**: Price of Laptop in Euros. (Target)
5. **ScreenW**: Screen width (pixels).
6. **ScreenH**: Screen height (pixels).
7. **CPU_freq**: Frequency of laptop CPU (Hz).
8. **PrimaryStorage**: Primary storage space (GB).
9. **SecondaryStorage**: Secondary storage space if any (GB).

3. Data Preprocessing

One notable observation from the analysis is that there were no missing or duplicate values in the dataset. When we start the analysis the whole dataset, we found that null values in three columns they are: Ram, ScreenW and Primary Storage. Then we try to find outliers in this three column and we get outliers in all the three column. The mean is the average of all values, so extreme values (outliers) can misshape it significantly higher or lower. The median is the middle value in a sorted list of numbers, which means it represents the central tendency without being affected by outliers. So, in this case we use median to fill-up all the null values. There we found that 14 columns are categorical where 3 columns (Product, CPU_model, GPU_model) we couldn't convert or mapping because of large number of unique values. Other 11 columns we covert categorical to

numerical and mapping them. Here, describe how we converted them to numerical values:

1. **Company**: Here we found 19 unique like values and mapping like, Apple: 0, 'HP': 1, 'Acer':2, 'Asus':3, 'Dell':4, 'Lenovo':5, 'Chuwi':6, 'MSI':7, 'Microsoft':8, 'Toshiba':9, 'Huawei':10, 'Xiomi':11, 'Vero':12, 'Razer':13, 'Mediacom':14, 'Samsung':15, 'Google':16, 'Fujitsu':17, 'LG':18.

2. **TypeName**: Here we found 6 unique values and mapping like that, Ultrabook': 0, 'Notebook': 1, 'Netbook':2, 'Gaming':3, '2 in 1 Convertible':4, 'Workstation':5

3. **OS**: Here we found 9 unique like values and mapping like, 'macOS': 0, 'No OS': 1, 'Windows 10':2, 'Mac OS X':3, 'Linux':4, 'Android':5, 'Windows 10 S':6, 'Chrome OS':7, 'Windows 7':8

4. **Screen**: Here we found 4 unique like values and mapping like, Standard': 0, 'Full H': 1, 'Quad HD+':2, '4K Ultra HD':3

5. **TouchScreen**: Here we found 2 unique like values and mapping like, 'No': 0, 'Yes': 1.

6. **IPSpanel**: Here we found 2 unique values and mapping like, 'Yes': 0, 'No': 1.

7. **RetinaDisplay**: Here we found 2 unique values and mapping like, 'Yes': 0, 'No': 1

8. **CPU_company**: Here we found 3 unique values and mapping like, 'Intel': 0, 'Amd': 1, 'Samsung':2.

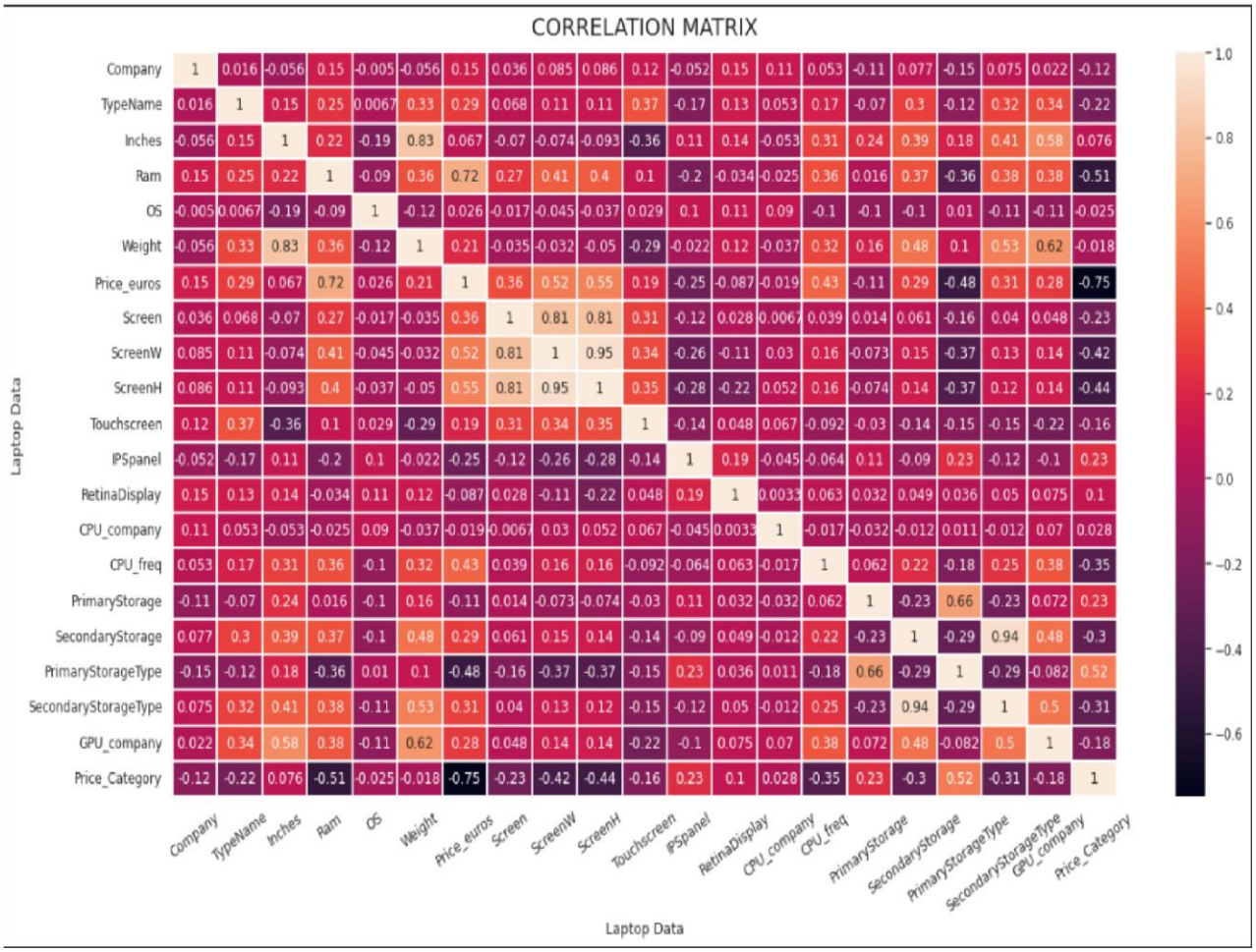
9. **PrimaryStorageType**: Here we found 4 unique values and mapping like, 'SSD': 0, 'Flash Storage': 1, 'HDD':2, 'Hybrid':3.

10. **SecondaryStorage** Type: Here we found 4 unique values and mapping like, 'No': 0, 'HDD': 1, 'SDD':2, 'Hybrid':3

11. **GPU_company**: Here we found 4 unique values and mapping like, 'Intel': 0, 'AMD': 1, 'Nvidia':2, 'ARM':3.

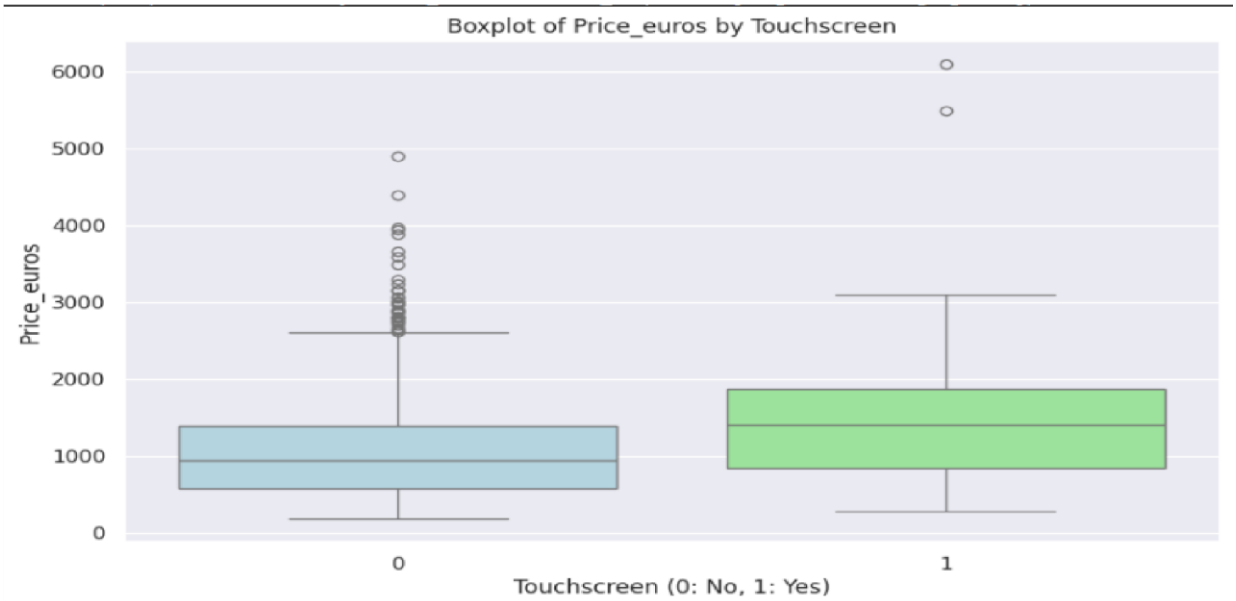
4. Exploratory Data Analysis (EDA):

Present the insights we gained from visualizing the data from the Laptop Prices Correlation Matrix. A correlation matrix is a statistical technique used to evaluate the relationship between two variables in a data set. The matrix is a table in which every cell contains a correlation coefficient, where 1 is considered a strong relationship between variables, 0 a neutral relationship and -1 a not strong relationship. Here given the table:

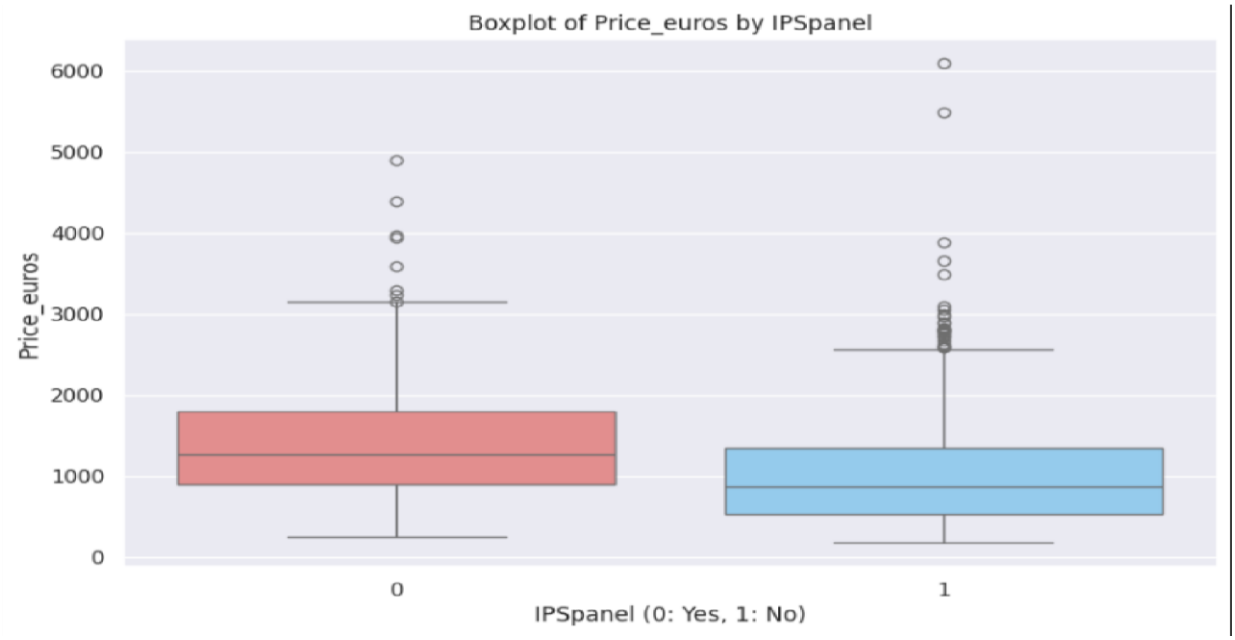


Here from the correlation heatmap matrix we found many interesting relationships between features or columns. There we found Price_euros and

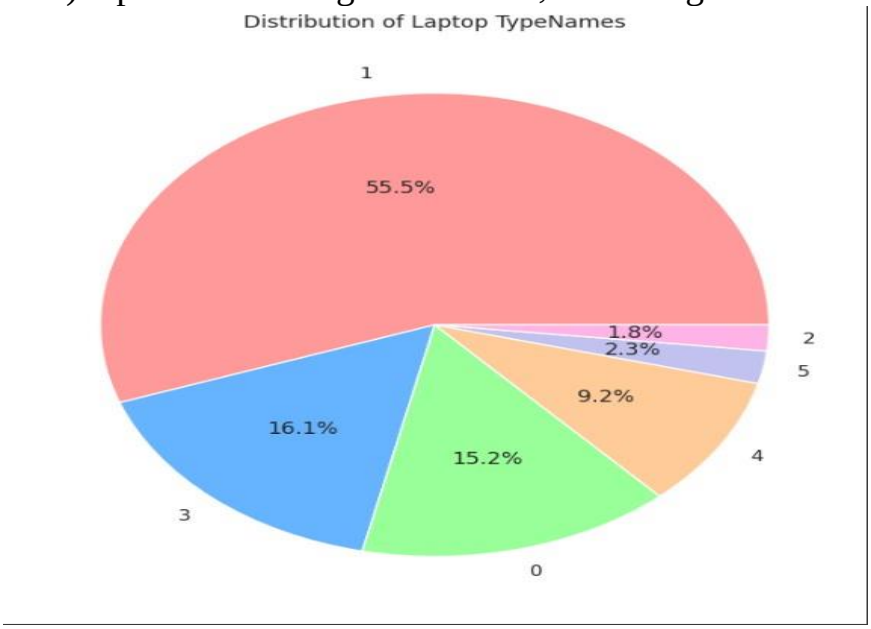
Touchscreen, Price_euros and IPSpanel, Primary Storage and Price Category, ScreenW and
ScreenH, Inches vs Weight, Screen vs ScreenH and Screen vs ScreenW etc. Now, here add the plots that we created (bar charts, pie charts, scatter plots, count plots, box plots and histogram etc.) and for each visualization explanation what information it reveals about the data.



The boxplot shows the distribution of laptop prices based on whether they have a touchscreen or not. The horizontal line inside the box represents the median price for each Touchscreen category. The box itself represents the interquartile range (IQR), which contains the middle 50% of the data. The IQR for laptops with touchscreens is slightly wider than the IQR for laptops without touchscreens. The lines extending from the box (whiskers) represent the range of the data, excluding outliers.

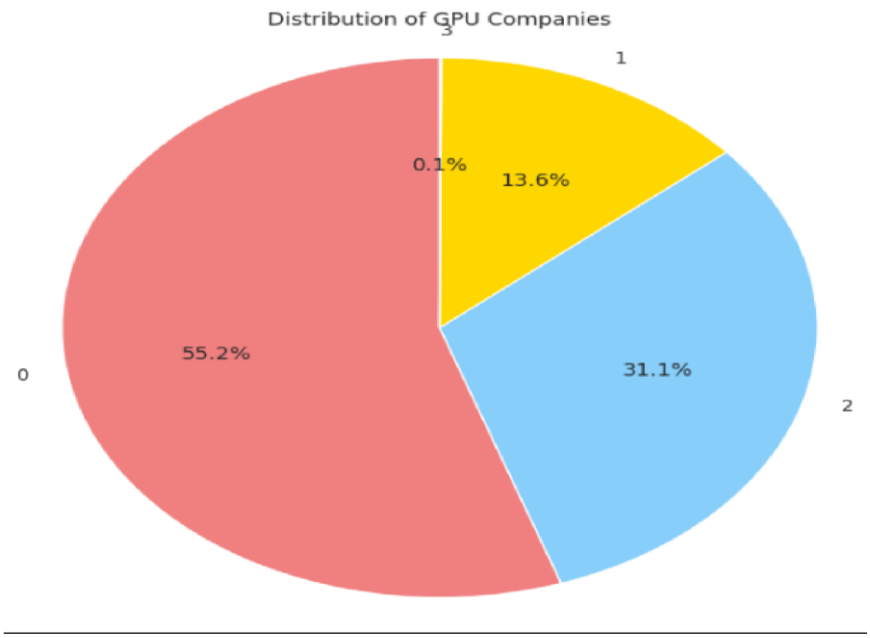


The boxplot shows the distribution of laptop prices based on whether they have an IPS panel or not. The horizontal line inside the box represents the median price for each IPSpanel category. The box itself represents the interquartile range (IQR), which contains the middle 50% of the data. The IQR for laptops with IPS panels is slightly wider than the IQR for laptops without IPS panels. The lines extending from the box (whiskers) represent the range of the data, excluding outliers.

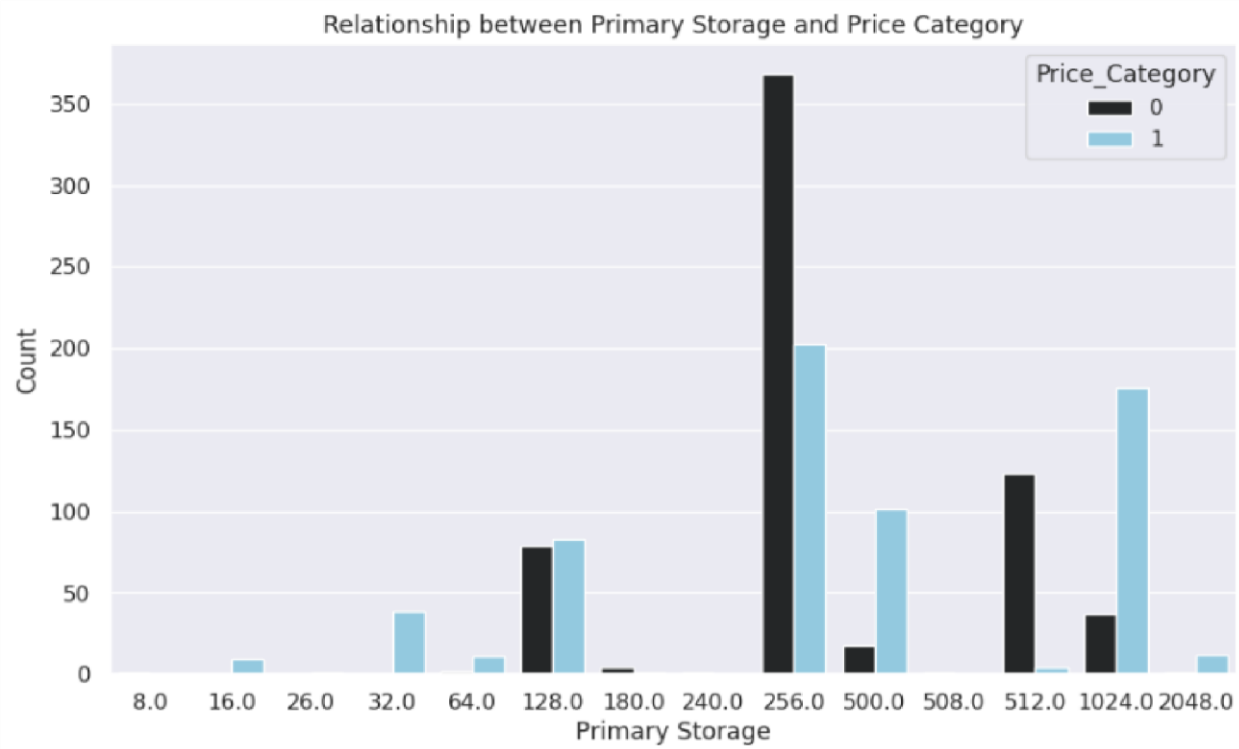


The pie chart visualizes the distribution of laptop TypeNames within the dataset where Ultrabook: 0, Notebook: 1 , Netbook:2, Gaming:3, 2 in 1 Convertible:4, Workstation:5. Each slice represents a different TypeName, and the size of the slice is proportional to the number of laptops belonging to that type. Notebooks are the

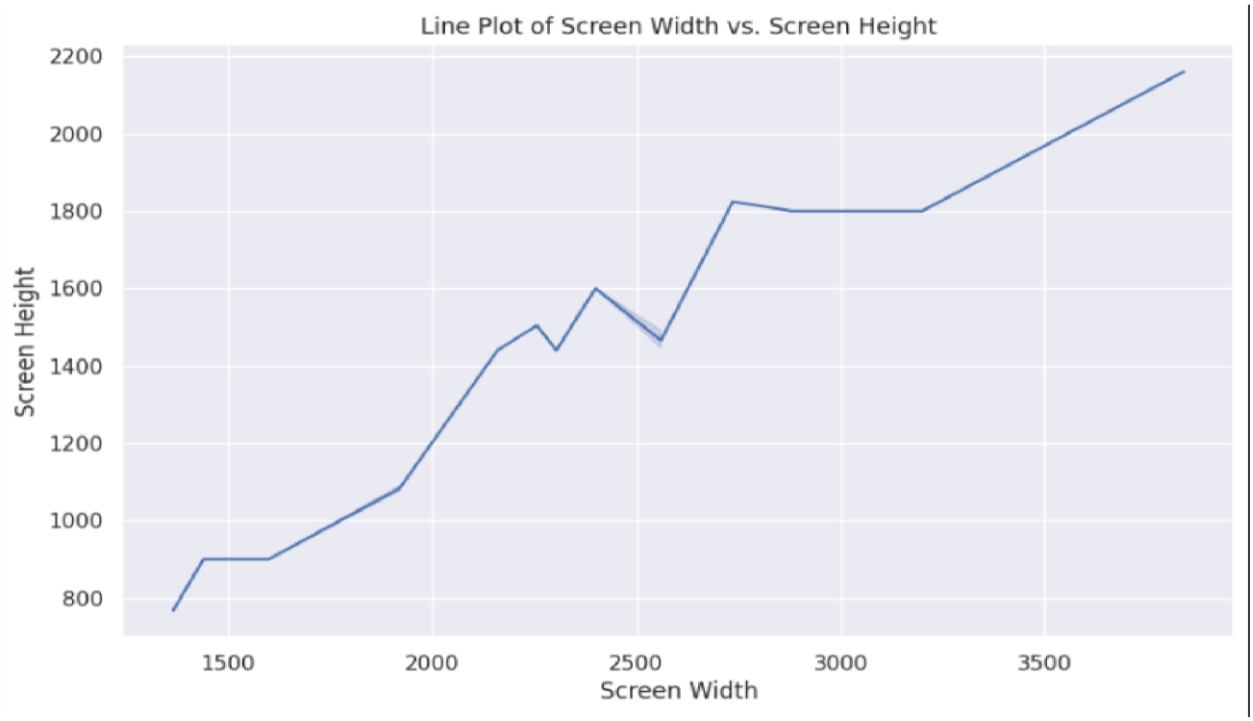
most prevalent type, comprising the largest portion of the dataset. Ultrabooks represent the second-largest category.



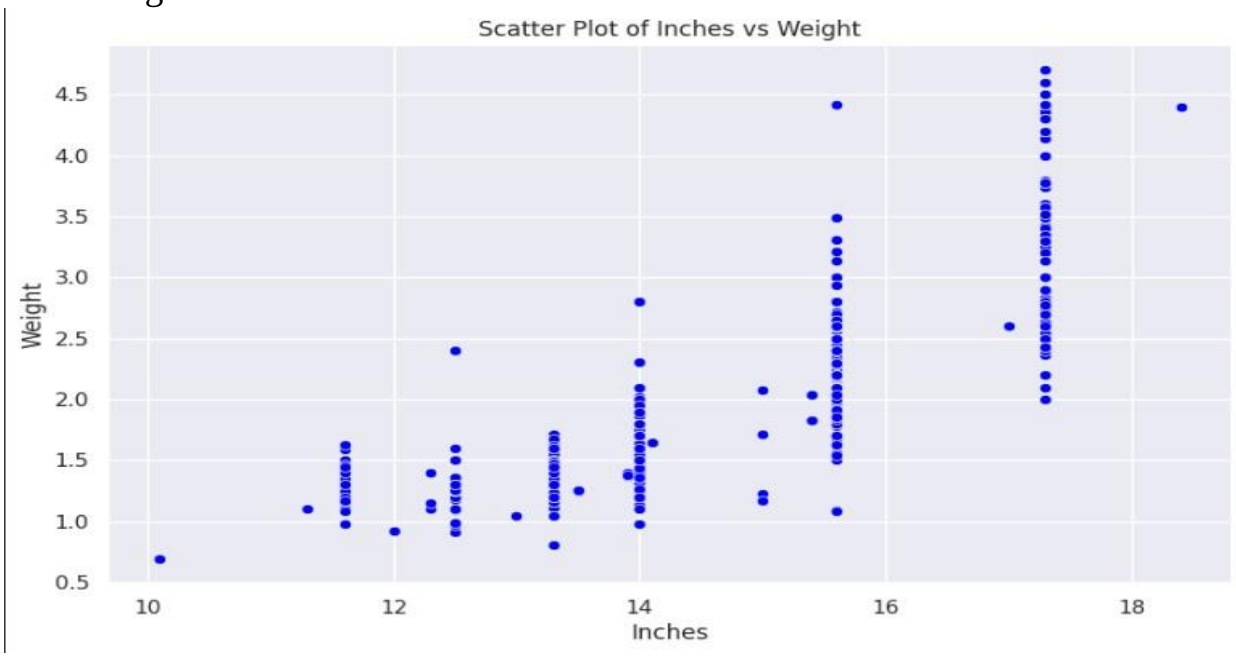
The pie chart displays the distribution of laptop GPU companies within the dataset Where Intel: 0, AMD: 1, Nvidia:2, ARM:3. Each slice of the pie represents a different GPU company, and its size is proportional to the number of laptops using GPUs from that company. Nvidia GPUs are the most common in the laptops in the dataset, constituting the largest portion of the pie. Intel GPUs are the second most frequent, making up a substantial portion of the laptops. AMD GPUs are present in a moderate number of laptops within the dataset. ARM GPUs are the least common in the laptops.



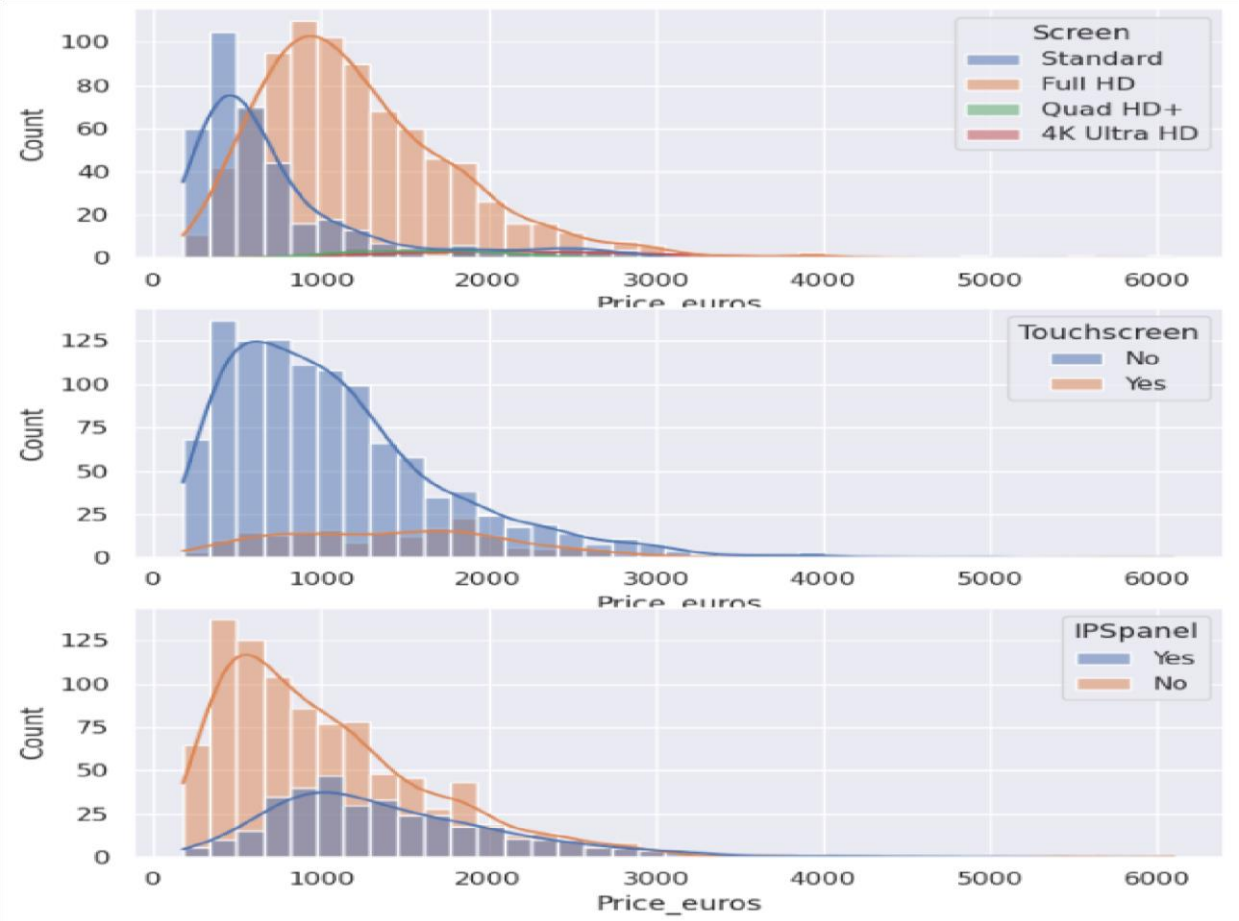
The count plot visualizes the distribution of the type of primary storage and the price category of the laptops. It shows the number of laptops with different types of primary storage (e.g., SSD, HDD). It also indicates the proportion of these laptops with either a high or low price.

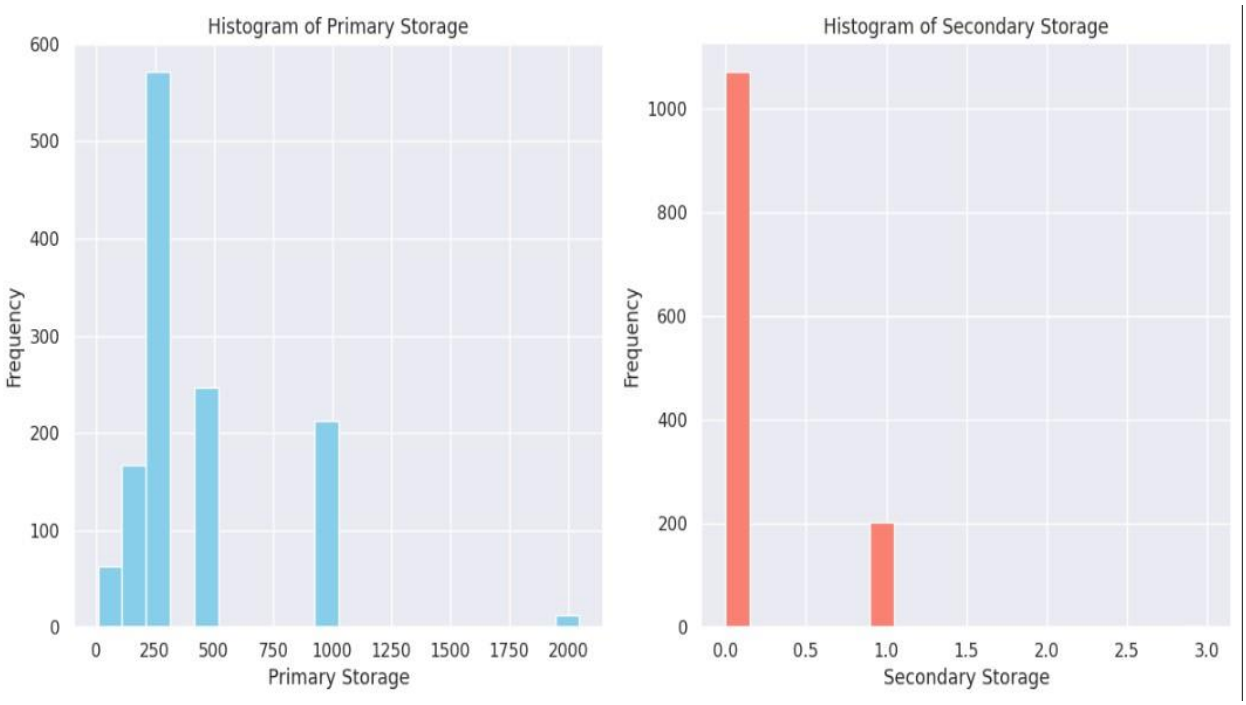


The line plot visualizes the relationship between screen width and screen height in the laptop dataset. It shows the general trend of how screen height changes with screen width. This indicates that laptops with wider screens tend to have greater screen heights.



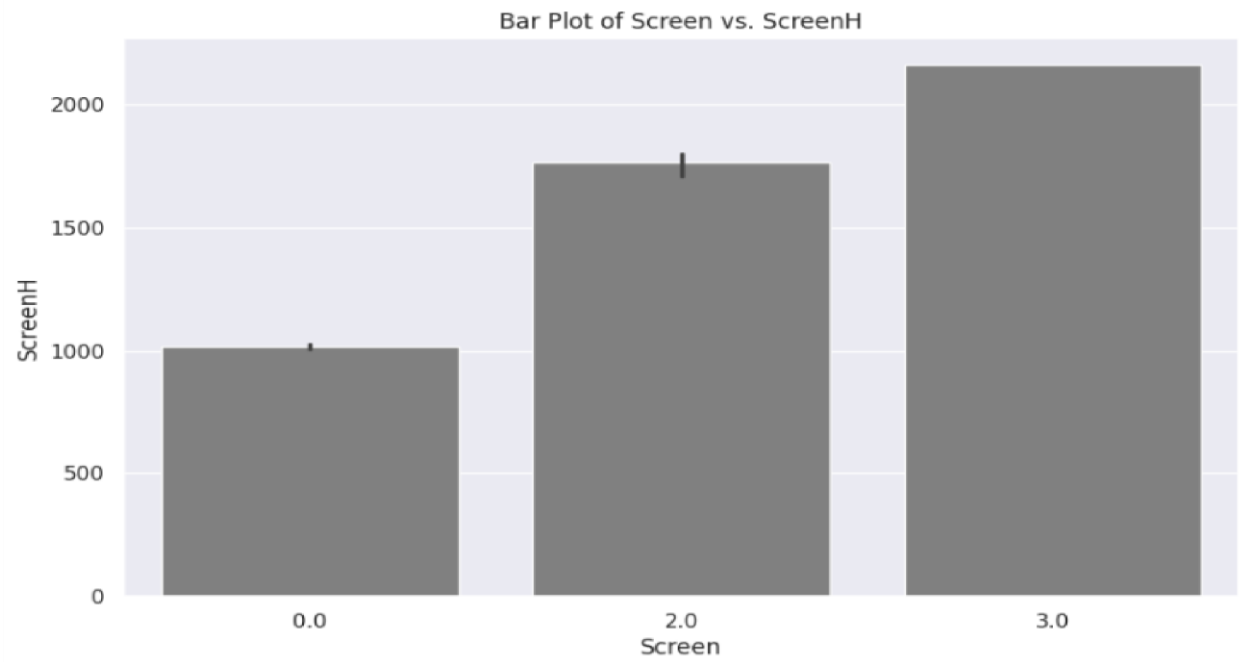
The scatter plot visualizes the relationship between screen size (Inches) and laptop weight. It is show that the screen size increases, the laptop weight also tends to increase. This indicates that larger laptops generally weight more.


Screen
Standard
Full HD
Quad HD+
4K Ultra HD
Count
Price_euros
Touchscreen
No
Yes
IPSpanel
Yes
No

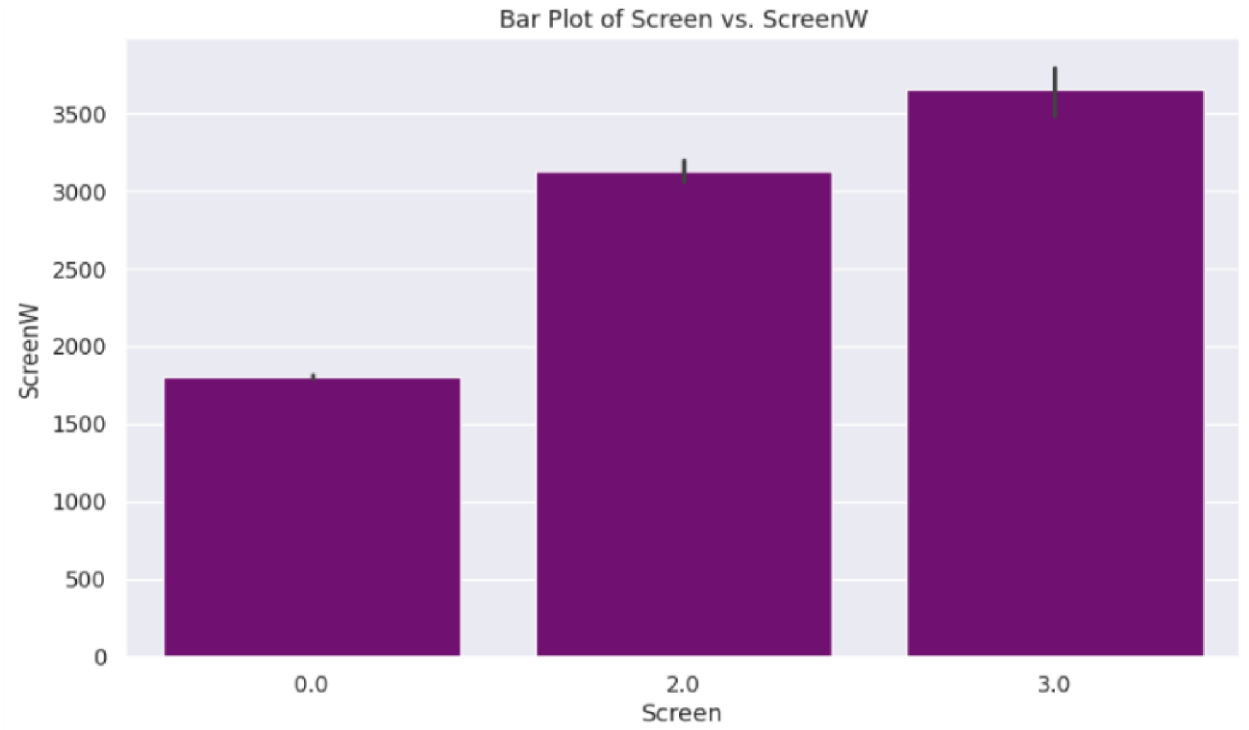


This histogram displays the distribution of primary storage capacities in the dataset. The x-axis represents the primary storage capacity in gigabytes. The y-axis represents the frequency or the number of laptops that have that particular storage capacity. The histogram shows that the majority of laptops in the dataset have a primary storage capacity in the lower ranges (presumably between 0-256 GB, if it's SSD or HDD). A smaller number of laptops have storage capacities in the higher ranges (for example above 512GB).

This histogram visualizes the distribution of secondary storage types or capacities in the dataset. The x-axis represents the secondary storage type (e.g., HDD, SDD, Hybrid). The y-axis represents the frequency or count of laptops with a specific secondary storage type. The majority of laptops have no secondary storage or less frequency. A small number of laptops have HDD as secondary storage.



The bar plot visualizes the average screen height (ScreenH) for different screen types. There screen generally has a higher or lower screen height. The x-axis represents different categories of screen displays (e.g., Standard, Full HD) while the y-axis represents the average screen height for each screen type.

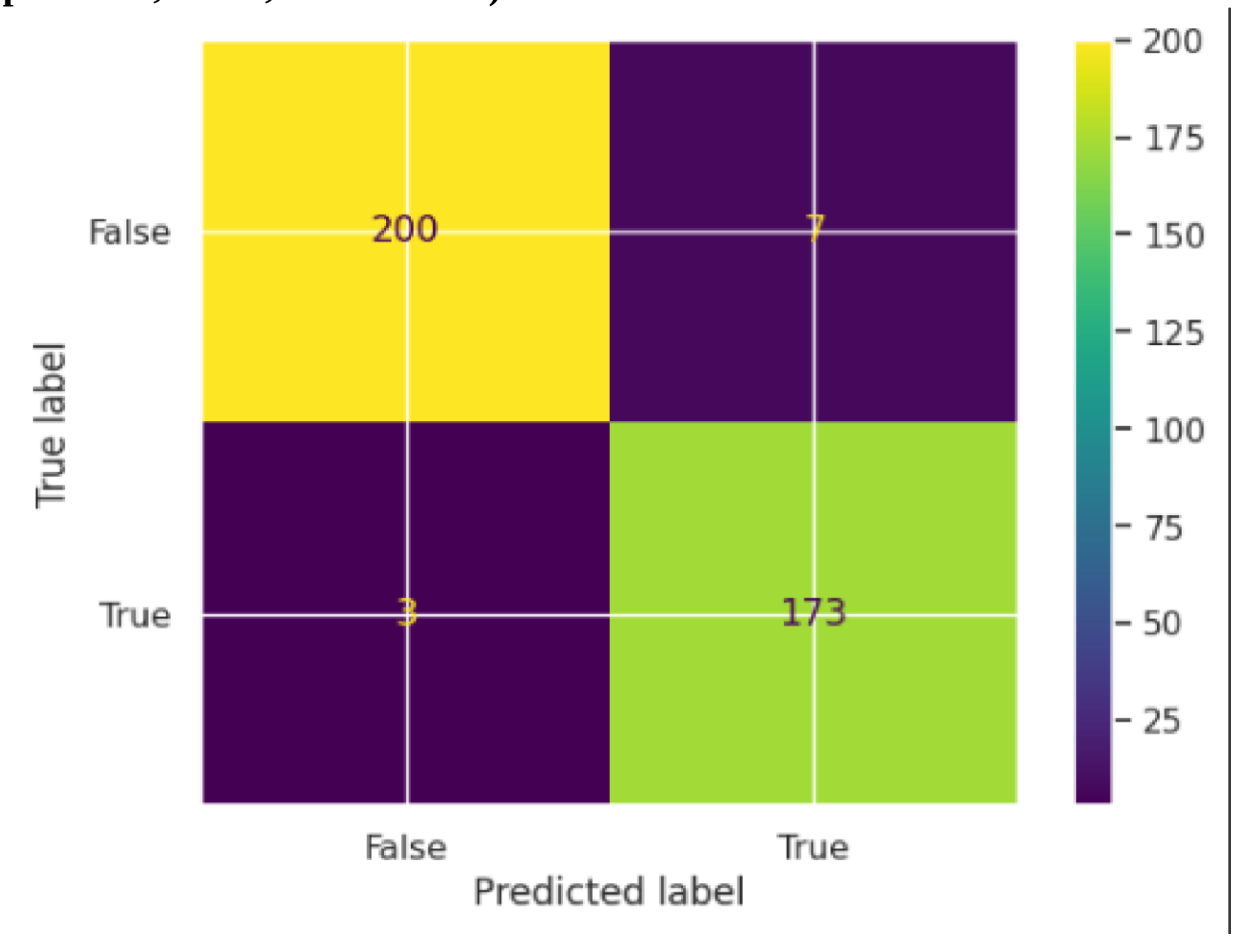


The bar plot visualizes the average screen width (ScreenW) for different screen types.There screen generally has a higher or lower screen width. The x-axis represents different categories of screen displays (e.g., Standard, Full HD) while the y-axis represents the average screen width for each screen type.

5. Machine Learning Models

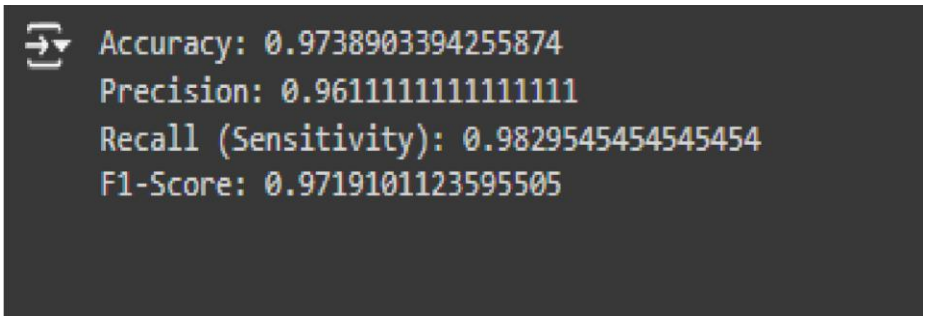
In a laptop price prediction project, using Logistic Regression is typically not the right choice because logistic regression is designed for classification problems, not for predicting continuous values like prices. Logistic regression is used when the target variable is categorical (e.g., yes/no, true/false), "low price" or "high price" categories. But, It is possible to convert continuous price values into categories (like "low" and "high") and then use logistic regression for classification and we do this. So, finally we run our project model into using Logistic Regression. Where we mapping price into low and high which include new column ‘Price_catagory’ and using median. Logistic regression will now be a classification problem where the target variable is the ‘Price_category’ (low,high) rather than a continuous price. So, we can convert price values into categories and run logistic regression to classify laptops into price categories like "low" or "high." However, we will lose some precision in price prediction but it’s not a big deal.

Here, Present the accuracy or other metrics (e.g., confusion matrix, precision, recall, F1 score etc.) into our models:



This is a confusion matrix, which is used to evaluate the performance of a classification model by comparing predicted labels with actual labels. Where,

True Positive (173): The model correctly predicted "True" when the actual label was "True."
True Negative (200): The model correctly predicted "False" when the actual label was "False."
False Positive (7): The model predicted "True" when the actual label was "False".
False Negative (3): The model predicted "False" when the actual label was "True".



Accuracy:
Overall correctness of the model's predictions. High accuracy means the model is correctly predicting a high proportion of laptop price categories.

Precision:
Out of all the laptops predicted as "High" price, how many were actually "High"? High precision means the model is making few false positive errors when predicting "High" price.

Recall (Sensitivity):
Out of all the laptops that are actually "High" price, how many did the model correctly identify as "High"? High recall means the model is identifying most of the actual "High" price laptops correctly.

F1-Score:
Harmonic mean of precision and recall. A balance between precision and recall. High F1-score means that the model is performing well in both correctly identifying "High" price laptops and avoiding false positive predictions for "High" price.

6. Conclusion

In this project, we aimed to predict laptop price categories (e.g., low, high) using Logistic Regression after converting continuous price values into categories. The logistic regression model did a good job of classifying laptops into price categories based on their features. While it worked well for this task, the model's performance could be improved by using more advanced techniques like decision

trees or random forests. These models are better at capturing complex patterns in the data. Still, the current model provides useful insights into how different laptop features, like RAM and processor type, relate to price categories. This makes it a helpful tool for both consumers and retailers to understand and classify laptops based on their specifications.

Finding:

- Processor and RAM: are the most influential features for determining laptop prices. Laptops with higher-end processors (e.g., Intel i7, AMD Ryzen 7) and more RAM (16GB or higher) tend to be classified into the "high price" category.

- Storage Type: also significantly affects price categories. Laptops with SSDs are more likely to be classified as "high price" compared to those with HDDs.

- Brand Influence: Premium brands like Apple are often associated with higher price categories, while brands like Dell, HP, and Lenovo are spread across both low and high categories, depending on other specifications.

- Accuracy: The confusion matrix shows that the logistic regression model performed well, with most laptops correctly classified into their respective price categories. The model correctly classified 200 "low price" laptops and 173 "high price" laptops, with only a few misclassifications (7 false positives and 3 false negatives).