LAPTOP PRICE PREDICTION FOR SMARTTECH CO.

CREATED BY ARCHITA SAHA

TABLE OF CONTENTS

PROJECT OVERVIEW AND SCOPE

Project Overview

The project overview outlines the collaboration between SmartTech Co. and the data science team to develop a machine learning model for predicting laptop prices accurately.

Project Scope

The scope of the project includes exploring the dataset, pre-processing the data, engineering meaningful features, developing and fine-tuning machine learning models, implementing a mechanism for real-time predictions, and presenting findings and insights to SmartTech Co. stakeholders.

BUSINESS PROBLEM

The business problem is to develop a robust machine learning model that accurately predicts laptop prices based on diverse specifications. This will enable SmartTech Co. to strategically position its laptops in the market and stay competitive.

BUSINESS OBJECTIVES

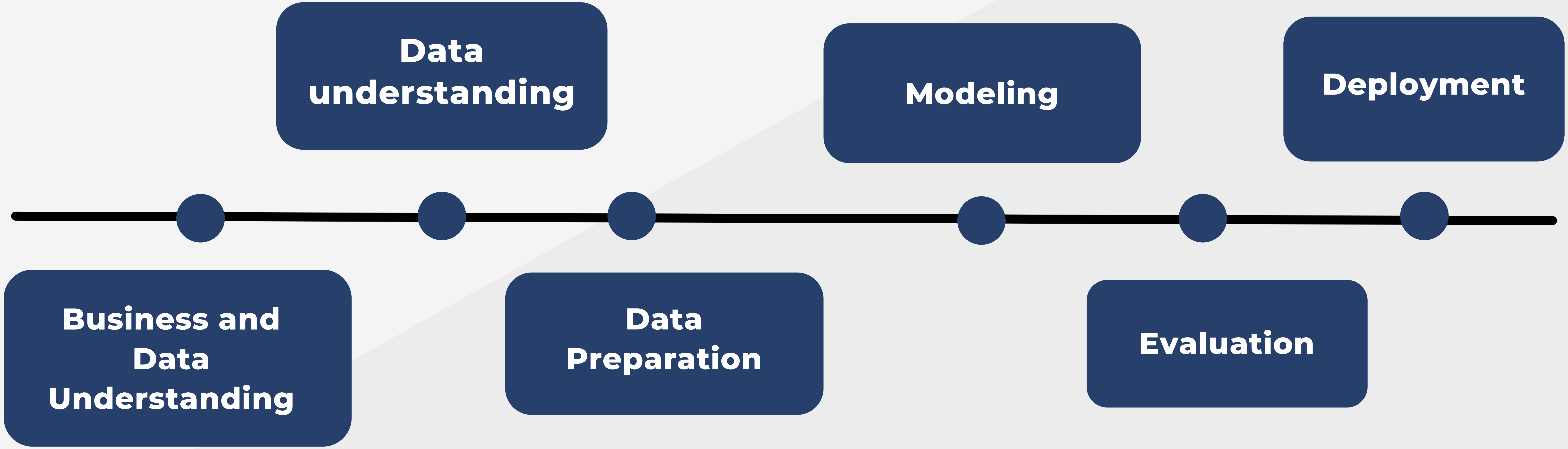
Develop a precise pricing model that accurately predicts laptop prices based on various features

Understand how different features contribute to pricing to strategically position SmartTech Co.'s laptops in the market

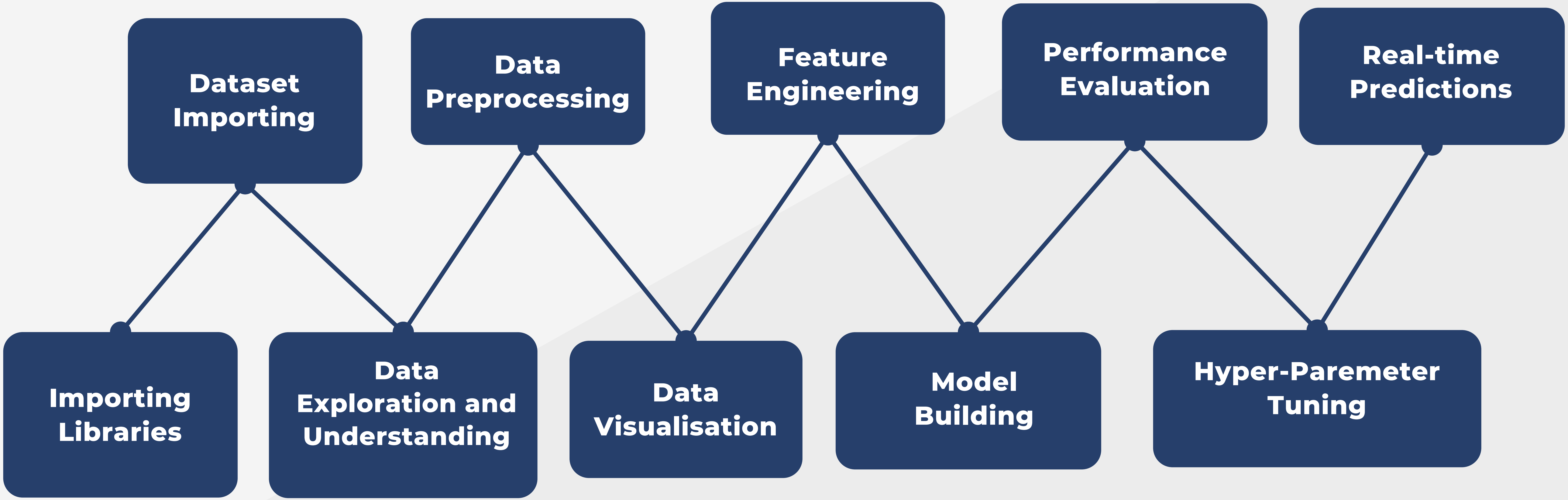
Assess the impact of brand reputation on pricing to enhance brand perception and market demand

CRISP-ML(Q) METHODOLOGY

To guide ML practitioners through the development life cycle, the Cross-Industry Standard Process for the development of Machine Learning applications with Quality assurance methodology (CRISP-ML(Q)) was recently proposed



PROJECT ARCHITECTURE



TECHNICAL STACKS

jupyter

pandas

matplotlib

plotly

seaborn

scikit learn

gradio

DATA COLLECTION AND UNDERSTANDING

The dataset consists of 1303 rows and 13 columns

Unnamed: 0.1	Unnamed: 0	Company	TypeName	Inches	ScreenResolution	Cpu	Ram	Memory	Gpu	OpSys	Weight	Price
0	0.0	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 2.3GHz	8GB	128GB SSD	Intel Iris Plus Graphics 640	macOS	1.37kg	71378.6832
1	1.0	Apple	Ultrabook	13.3	1440x900	Intel Core i5 1.8GHz	8GB	128GB Flash Storage	Intel HD Graphics 6000	macOS	1.34kg	47895.5232
2	2.0	HP	Notebook	15.6	Full HD 1920x1080	Intel Core i5 7200U 2.5GHz	8GB	256GB SSD	Intel HD Graphics 620	No OS	1.86kg	30636.0000
3	3.0	Apple	Ultrabook	15.4	IPS Panel Retina Display 2880x1800	Intel Core i7 2.7GHz	16GB	512GB SSD	AMD Radeon Pro 455	macOS	1.83kg	135195.3360
4	4.0	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 3.1GHz	8GB	256GB SSD	Intel Iris Plus Graphics 650	macOS	1.37kg	96095.8080

11 Columns left after 'Unnamed: 0.1' and 'Unnamed: 0' columns were dropped

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1273 entries, 0 to 1302
Data columns (total 11 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Company           1273 non-null   object
 1   TypeName          1273 non-null   object
 2   Inches            1273 non-null   object
 3   ScreenResolution  1273 non-null   object
 4   Cpu               1273 non-null   object
 5   Ram               1273 non-null   object
 6   Memory            1273 non-null   object
 7   Gpu               1273 non-null   object
 8   OpSys             1273 non-null   object
 9   Weight            1273 non-null   object
 10  Price             1273 non-null   float64
dtypes: float64(1), object(10)
memory usage: 119.3+ KB
```

DATA DICTIONARY

Columns/Features	Description
Company	Name of the laptop brand
TypeName	Type of the laptop used for various applications
Inches	Laptop Screen Size
ScreenResolution	Resolution of laptop screen in dimension pixels and type of screen
Cpu	Cpu brand and Type
Ram	RAM size in GB
Memory	Total Memory
Gpu	GPU brand and Type
OpSys	Operating system
Weight	Weight of laptop in kgs
Price	Price of laptop

DEALING WITH MISSING VALUES

Removed the null values as they were very less compared to shape

```
data.isnull().sum()
```

```
Unnamed: 0.1         0
Unnamed: 0          30
Company             30
TypeName            30
Inches              30
ScreenResolution    30
Cpu                 30
Ram                 30
Memory              30
Gpu                 30
OpSys               30
Weight              30
Price               30
dtype: int64
```

The columns “Weight” and “Inches” consists of garbage values might be due to data entry errors; those values were dropped

```
[12]: data.drop(data[data['Weight'] == '?'].index, inplace=True)
      data['Ram']=data['Ram'].astype('int32')
      data['Weight']=data['Weight'].astype('float64')
```

```
[13]: data.drop(data[data['Inches'] == '?'].index, inplace=True)
      data['Inches']=data['Inches'].astype('float64')
```

FEATURE ENGINEERING

The 'ScreenResolution' column has been divided into three seperate columns i.e 'Resolution_in_pixels', 'Touch_screen_laptop' and 'IPS_laptop' as most laptop screens are belong to these types

```
[21]: data['ScreenResolution'].value_counts()

[21]: ScreenResolution
      Full HD 1920x1080                                    493
      1366x768                                             274
      IPS Panel Full HD 1920x1080                          226
      IPS Panel Full HD / Touchscreen 1920x1080             52
      Full HD / Touchscreen 1920x1080                       45
      1600x900                                              23
      Touchscreen 1366x768                                  16
      Quad HD+ / Touchscreen 3200x1800                      14
      IPS Panel 4K Ultra HD 3840x2160                       12
      IPS Panel 4K Ultra HD / Touchscreen 3840x2160         11
      4K Ultra HD / Touchscreen 3840x2160                    9
      4K Ultra HD 3840x2160                                  7
      IPS Panel 1366x768                                     7
      IPS Panel Retina Display 2560x1600                     6
      IPS Panel Quad HD+ / Touchscreen 3200x1800             6
      Touchscreen 2560x1440                                  6
      IPS Panel Retina Display 2304x1440                     6
      Touchscreen 2256x1504                                  6
      IPS Panel Touchscreen 2560x1440                        5
      1440x900                                               4
      IPS Panel 2560x1440                                    4
```

Touch_screen_laptop	IPS_laptop	Resolution_in_pixels
0	1	2560x1600
0	0	1440x900
0	0	1920x1080
0	1	2880x1800
0	1	2560x1600

CONTD...

Various types of CPUs are available in the market, and so they were divided into the most famous one such as Intel and less famous ones are separated as others.

```
data['Cpu'].value_counts()
```

```
Cpu
Intel Core i5 7200U 2.5GHz     183
Intel Core i7 7700HQ 2.8GHz    141
Intel Core i7 7500U 2.7GHz     128
Intel Core i7 8550U 1.8GHz      71
Intel Core i5 8250U 1.6GHz      68
                              ...
AMD A9-Series 9420 2.9GHz        1
Intel Core i7 2.2GHz             1
AMD A6-Series 7310 2GHz          1
Intel Atom Z8350 1.92GHz         1
AMD E-Series 9000e 1.5GHz        1
Name: count, Length: 118, dtype: int64
```

```
[34]: def categorise_cpu(name):
    if name == 'Intel Core i7' or name == 'Intel Core i5' or name == 'Intel Core i3':
        return name
    else:
        if name.split()[0] == 'Intel':
            return 'Other Intel Processor'
        else:
            return 'AMD Processor'
```

CONTD...

The memory column has been transformed into four available types 'HDD', 'SDD', 'Hybrid' and 'Flash Storage'

```
data['Memory'].value_counts()
```

```
Memory
256GB SSD                  399
1TB HDD                    217
500GB HDD                  130
512GB SSD                  116
128GB SSD +  1TB HDD        92
128GB SSD                   74
256GB SSD +  1TB HDD        71
32GB Flash Storage          37
2TB HDD                     16
64GB Flash Storage          14
512GB SSD +  1TB HDD        14
1TB SSD                     13
256GB SSD +  2TB HDD        10
1.0TB Hybrid                 9
256GB Flash Storage          8
16GB Flash Storage           7
32GB SSD                     6
180GB SSD                    4
128GB Flash Storage          4
512GB SSD +  2TB HDD         3
16GB SSD                     3
512GB Flash Storage          2
1TB SSD +  1TB HDD           2
128GB SSD +  2TB HDD         2
256GB SSD +  500GB HDD       2
256GB SSD +  256GB SSD       2
```

HDD	SSD	Hybrid	Flash_Storage
0	128	0	0
0	0	0	128
0	256	0	0
0	512	0	0
0	256	0	0

But, the columns ‘Hybrid’ and ‘Flash Storage’ have most 0 values, so they are dropped

```
data['Hybrid'].value_counts()
```

```
Hybrid
0       1258
1000      11
508        1
Name: count, dtype: int64
```

```
data['Flash_Storage'].value_counts()
```

```
Flash_Storage
0      1197
32       37
64       15
256       8
16        7
128       4
512       2
Name: count, dtype: int64
```

CONTD...

Similarly, so many GPUs are there in market so only first first is kept as brand name to generalise better

```
[63]: data['Gpu'].value_counts()

[63]: Gpu
      Intel HD Graphics 620      271
      Intel HD Graphics 520      181
      Intel UHD Graphics 620      65
      Nvidia GeForce GTX 1050     64
      Nvidia GeForce GTX 1060     48
                                ...
      AMD Radeon R5 520            1
      AMD Radeon R7                1
      Intel HD Graphics 540        1
      AMD Radeon 540               1
      ARM Mali T860 MP4            1
      Name: count, Length: 110, dtype: int64
```

The Operating Systems are categorised under 'Windows, 'Mac', and 'Others'.

```
[67]: data['OpSys'].value_counts()

[67]: OpSys
      Windows 10      1044
      No OS             63
      Linux             61
      Windows 7         45
      Chrome OS         27
      macOS             13
      Mac OS X           8
      Windows 10 S       8
      Android            1
      Name: count, dtype: int64
```

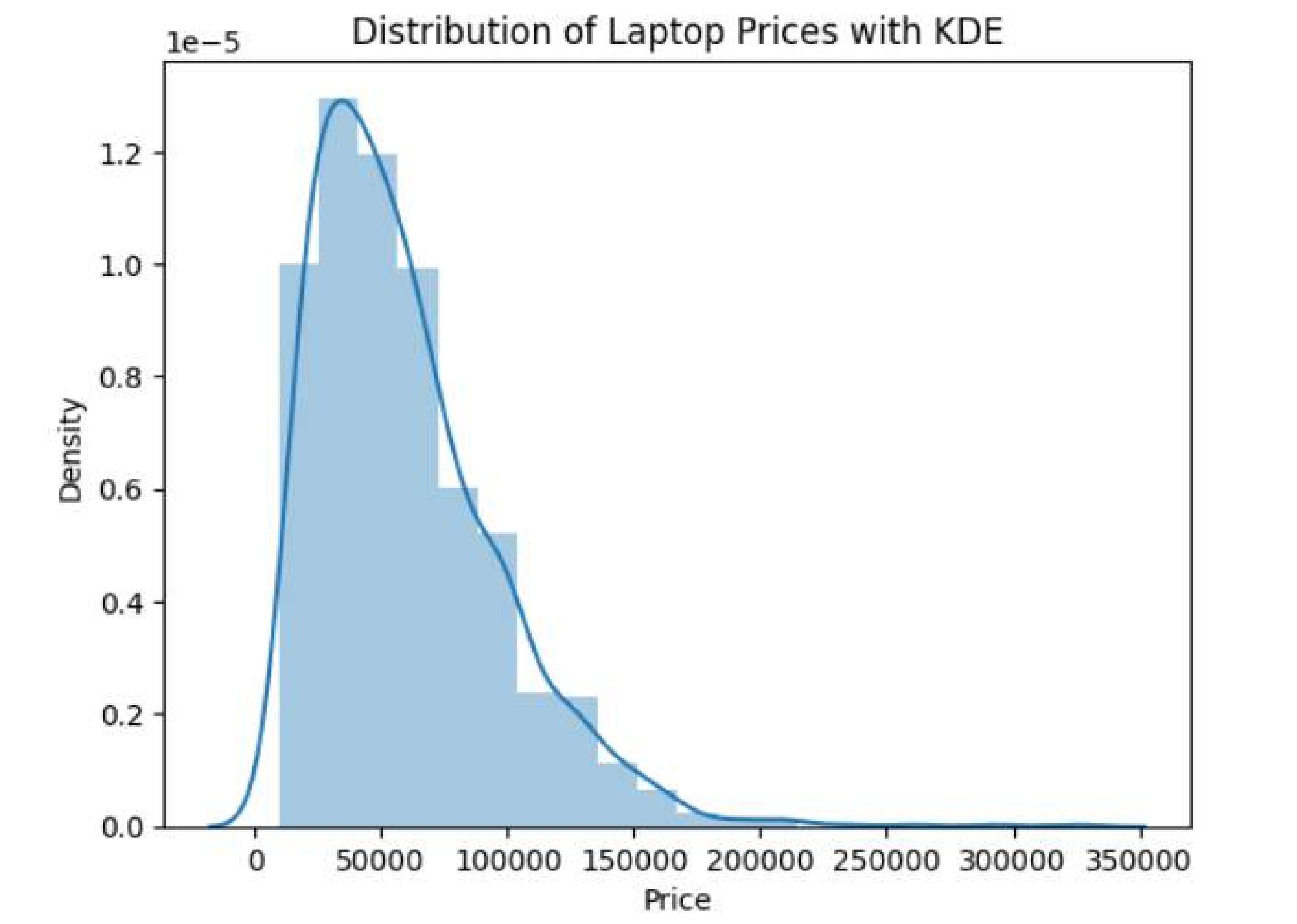
CONTD...

PPI (Pixels Per Inch) measures the pixel density of a display, indicating the number of pixels present per inch of screen space. It is calculated by dividing the diagonal resolution of a display by its diagonal size. PPI is important for determining the sharpness and clarity of images and text on screens. So, this feature has been added to predict laptop price more clearly.

$$PPI = \frac{\text{Total Pixels}}{\text{Screen size (inch)}}$$

```
[30]: data['PPI']=((width**2)+(height**2))**0.5/data['Inches']

[31]: data.info()
```

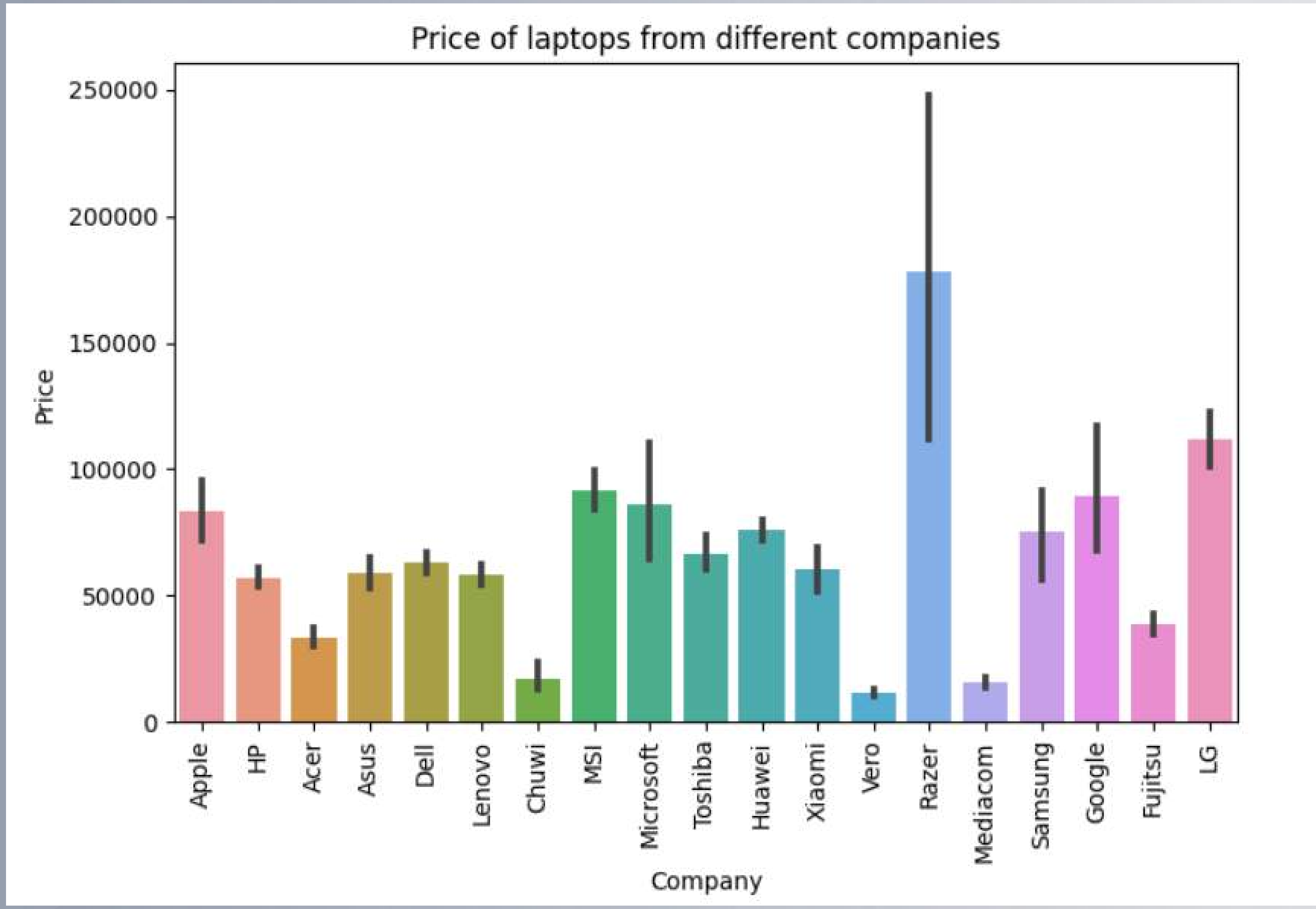
DATA VISUALISATIONS

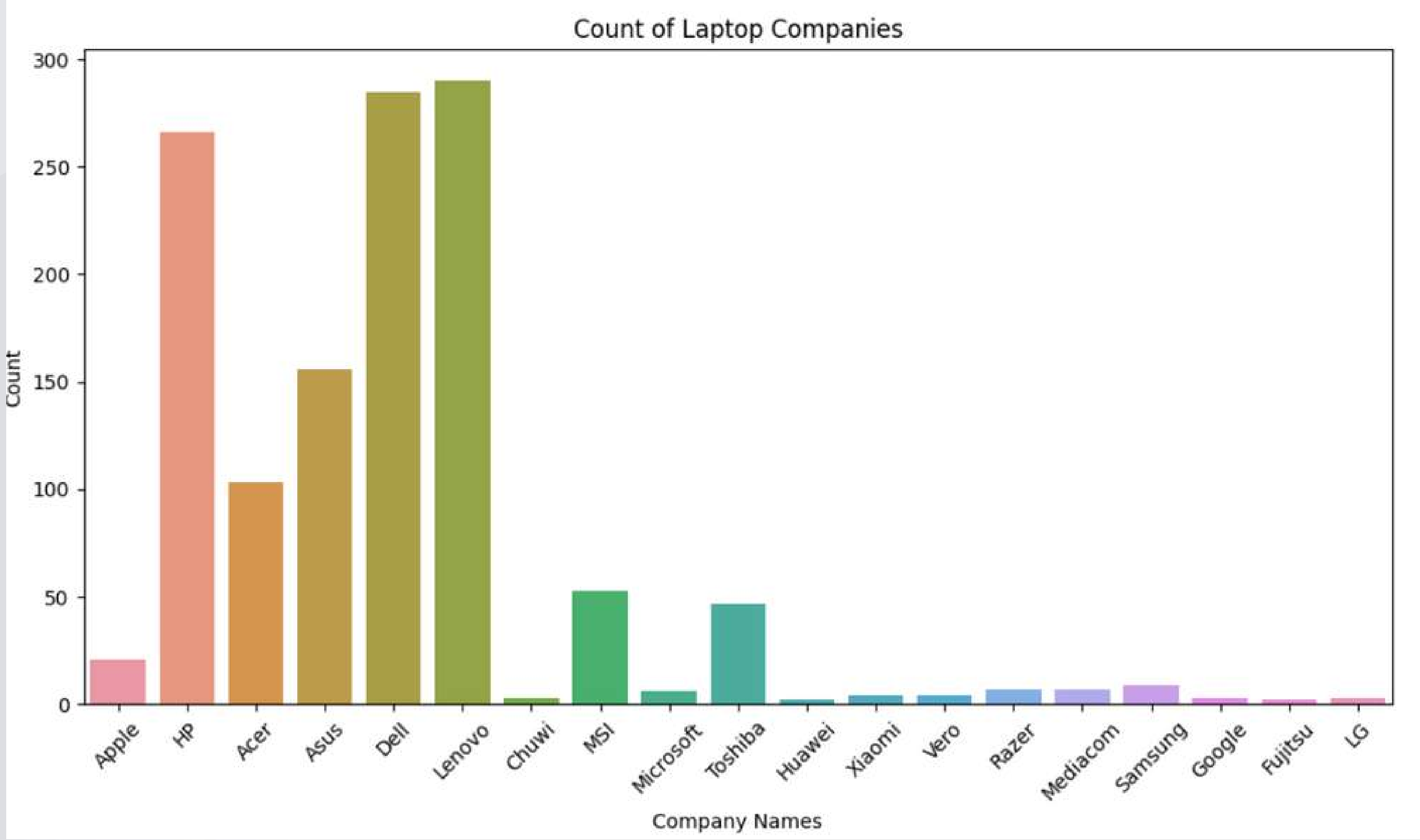


The laptop prices mostly range between 10,000 to 2,00,000. However, most laptops cost 50,000. This plot also shows that the distribution of laptop prices is right skewed or positive skewed.

CONTD...

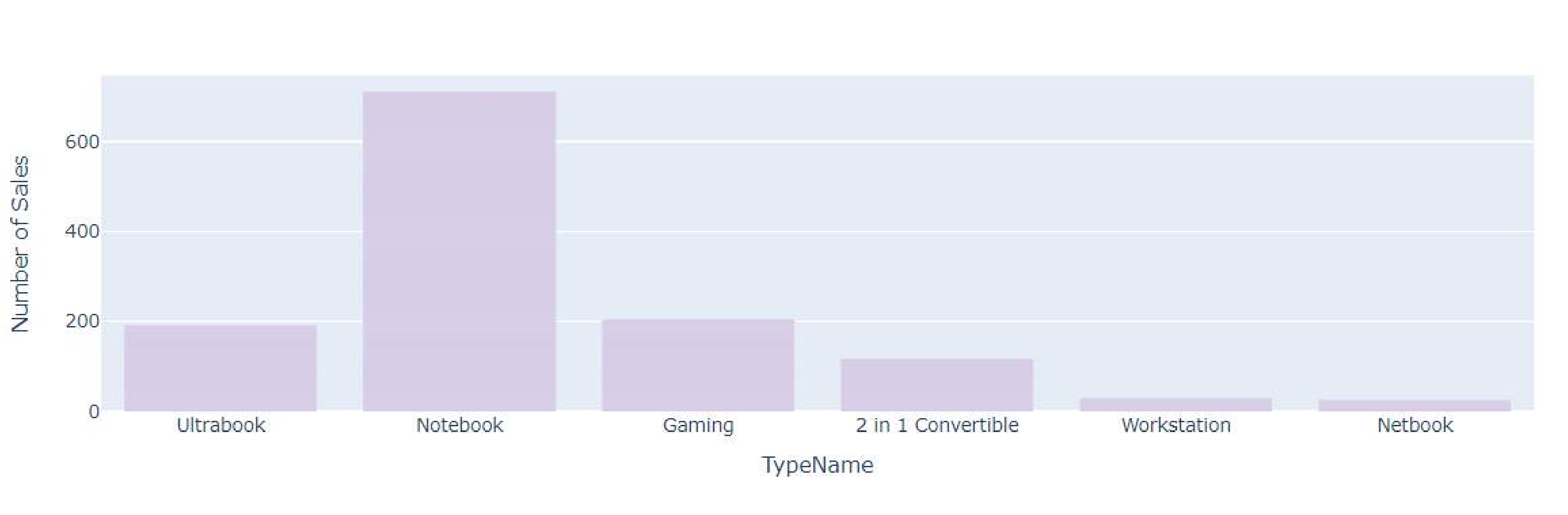
Razer laptops are most expensive. But, they are sold less. But, the mid-range laptops with affordable prices are sold most. Similarly, Lenovo, Dell, Asus, Acer, and HP fall in this category.

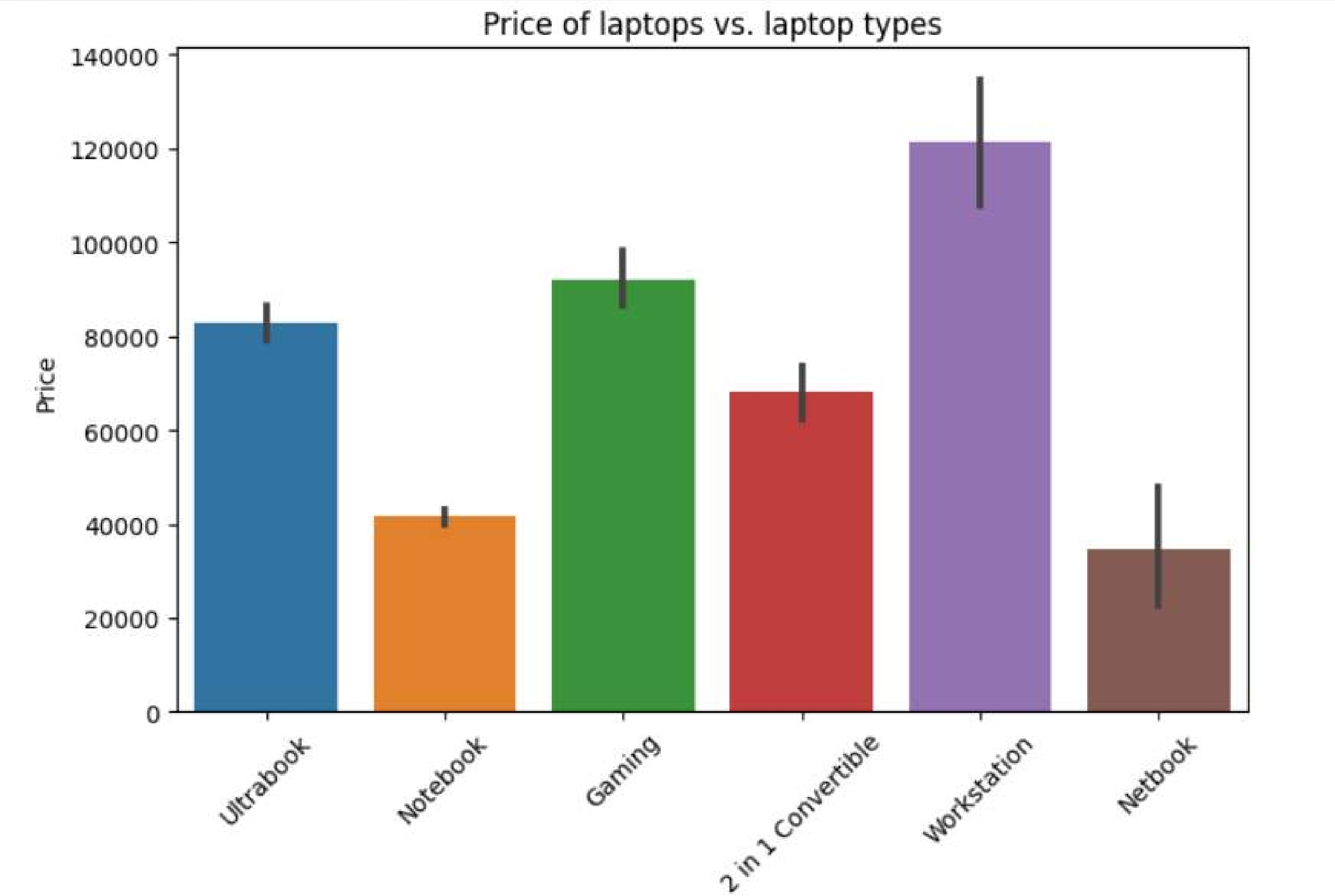




CONTD…

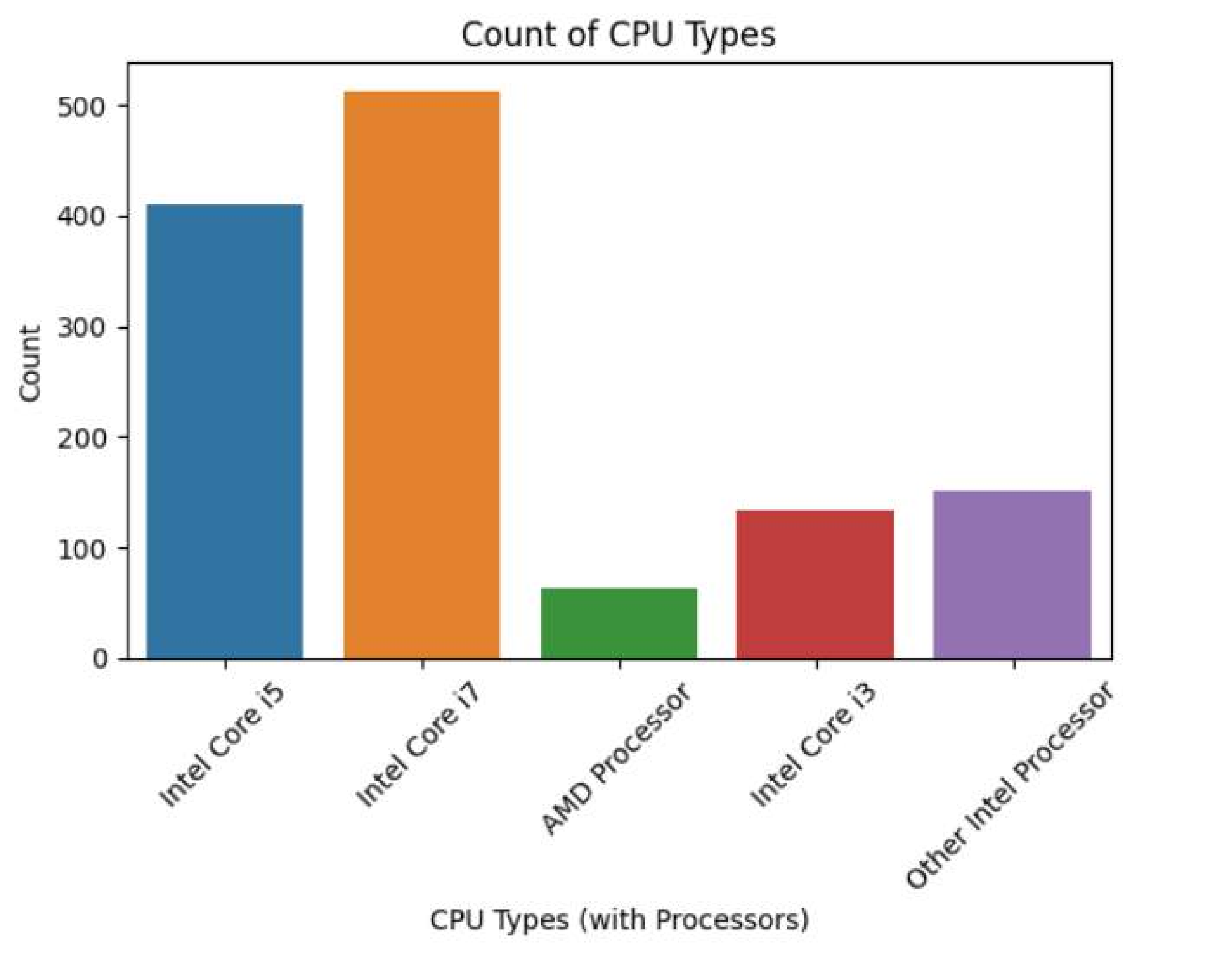
Notebook laptops are sold the most, followed by Gaming and Ultrabook. Here, Notebooks have affordable pricing of about 40,000 and hence they are sold most. Workstation, on the other hand, are the most expensive laptops that are sold less.

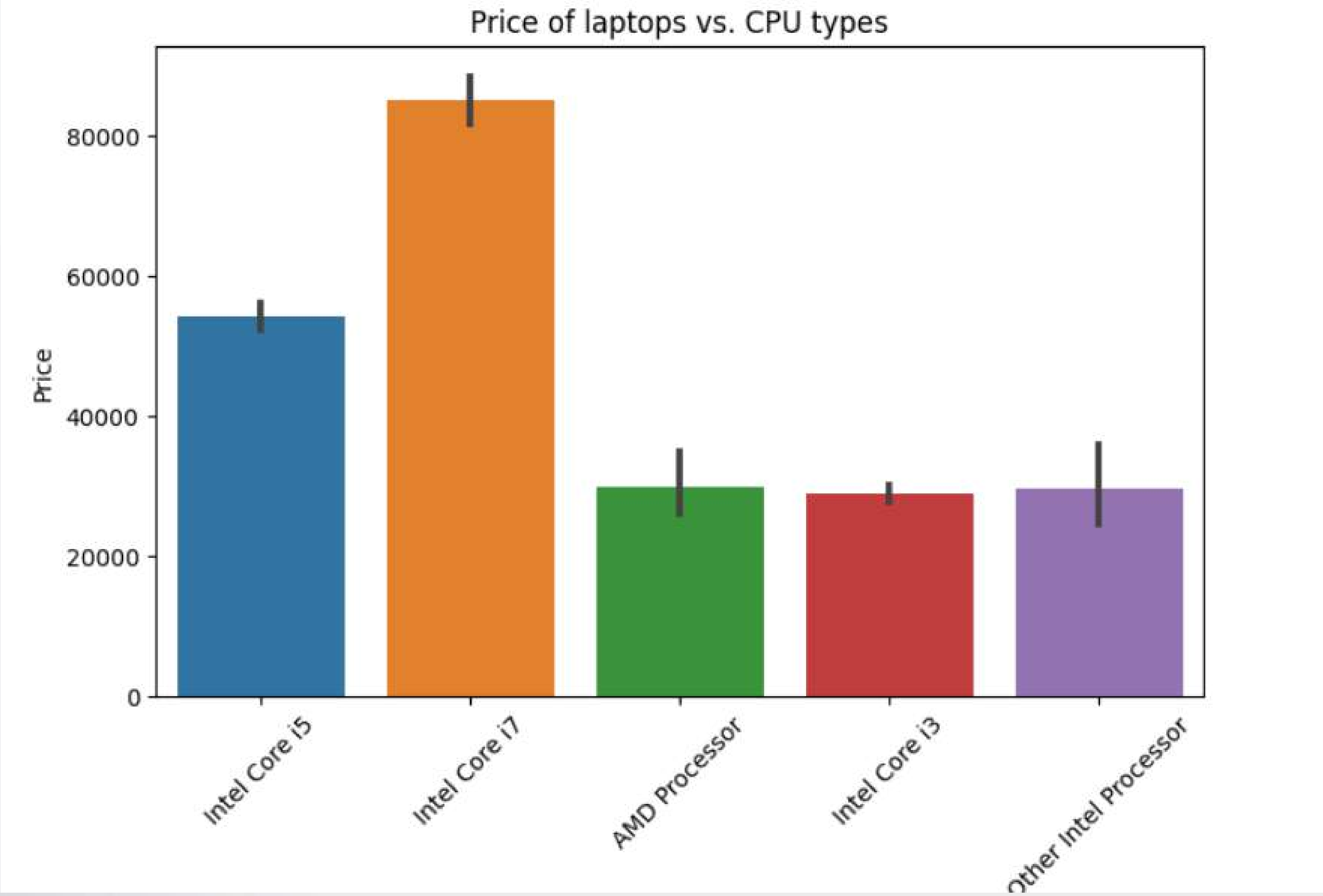




CONTD...

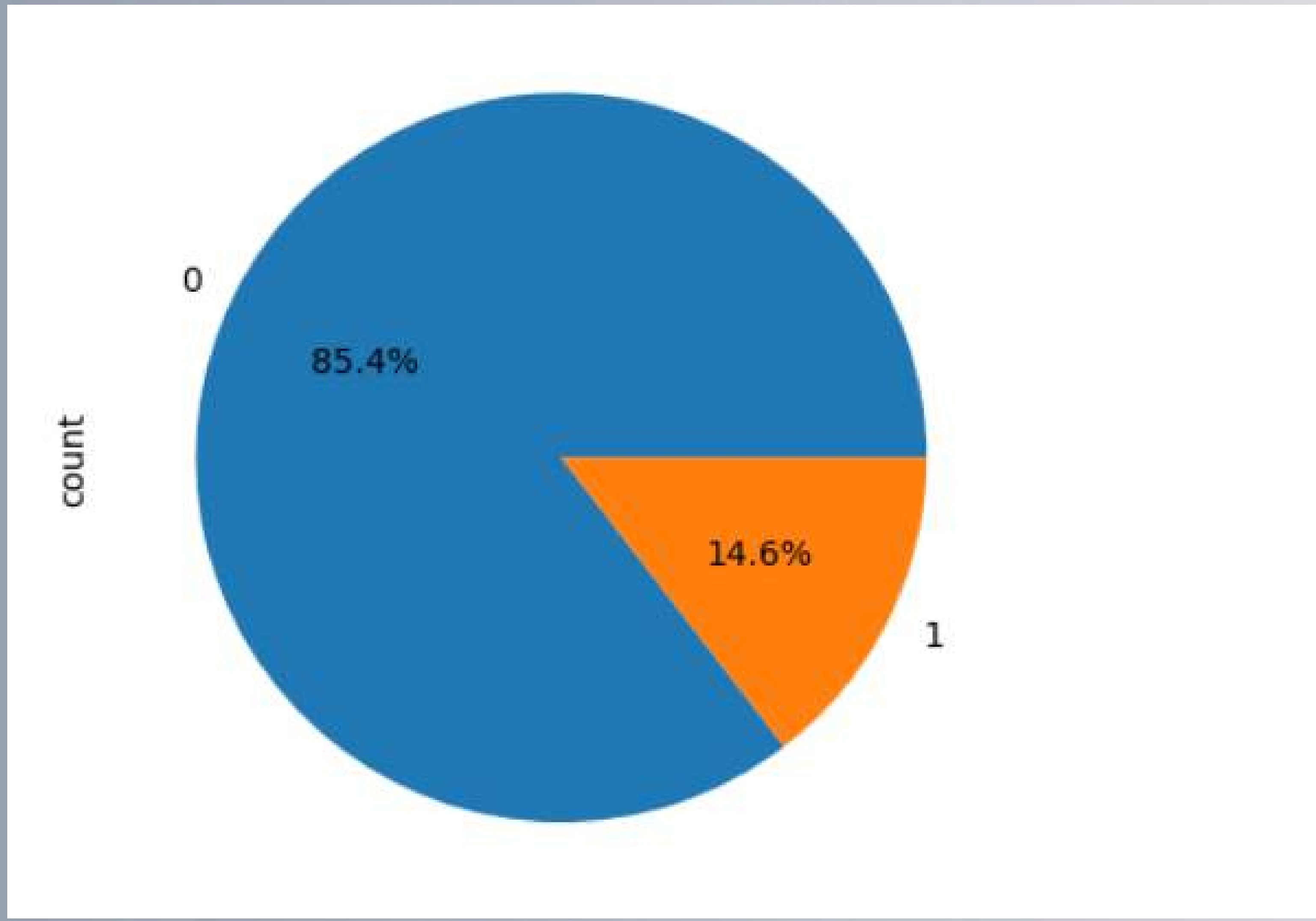
More number of intel core i7 processors are sold in the market compared to other different processors and brands. Similarly, it is also the most costly one, followed by intel core i5.

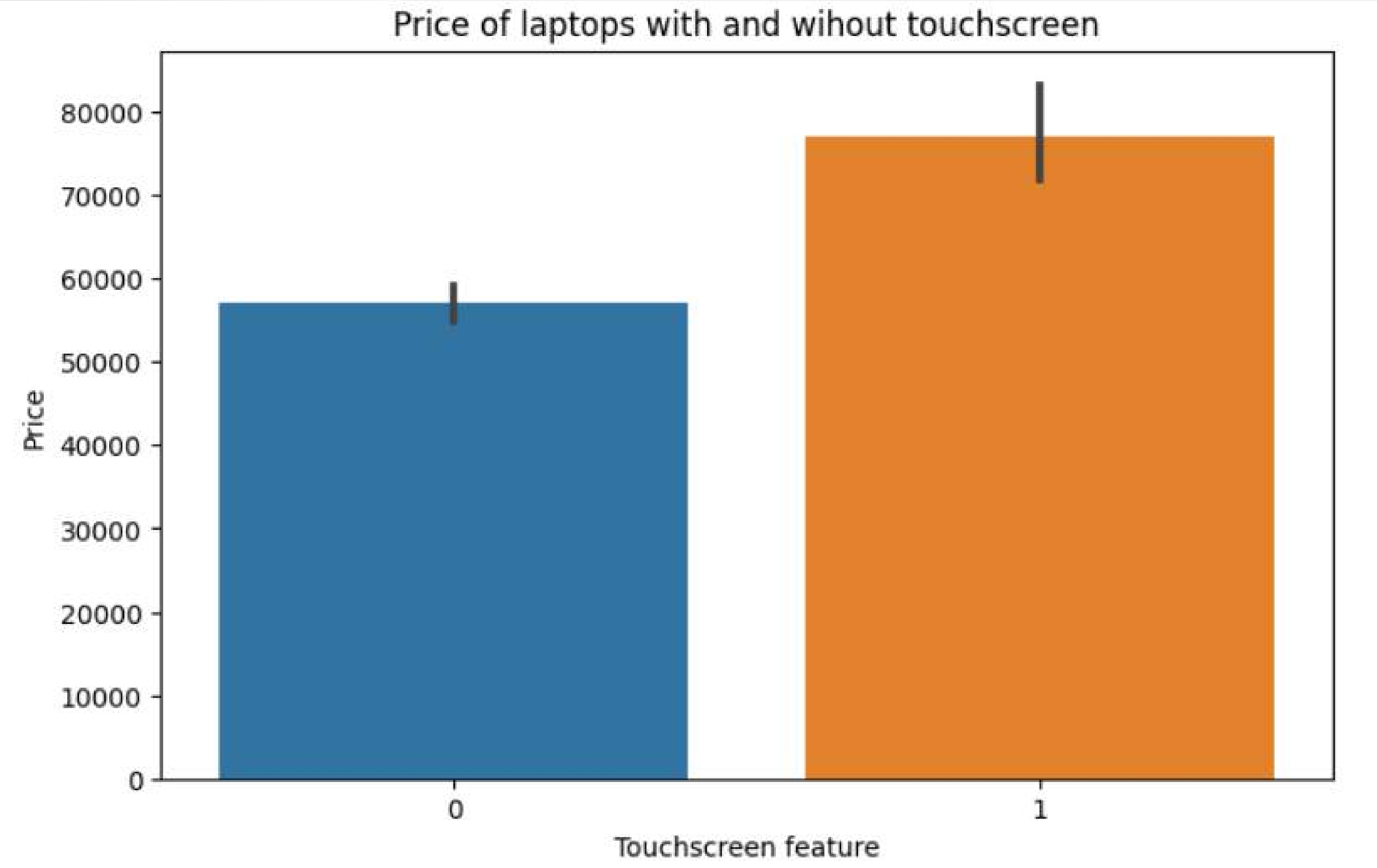




CONTD...

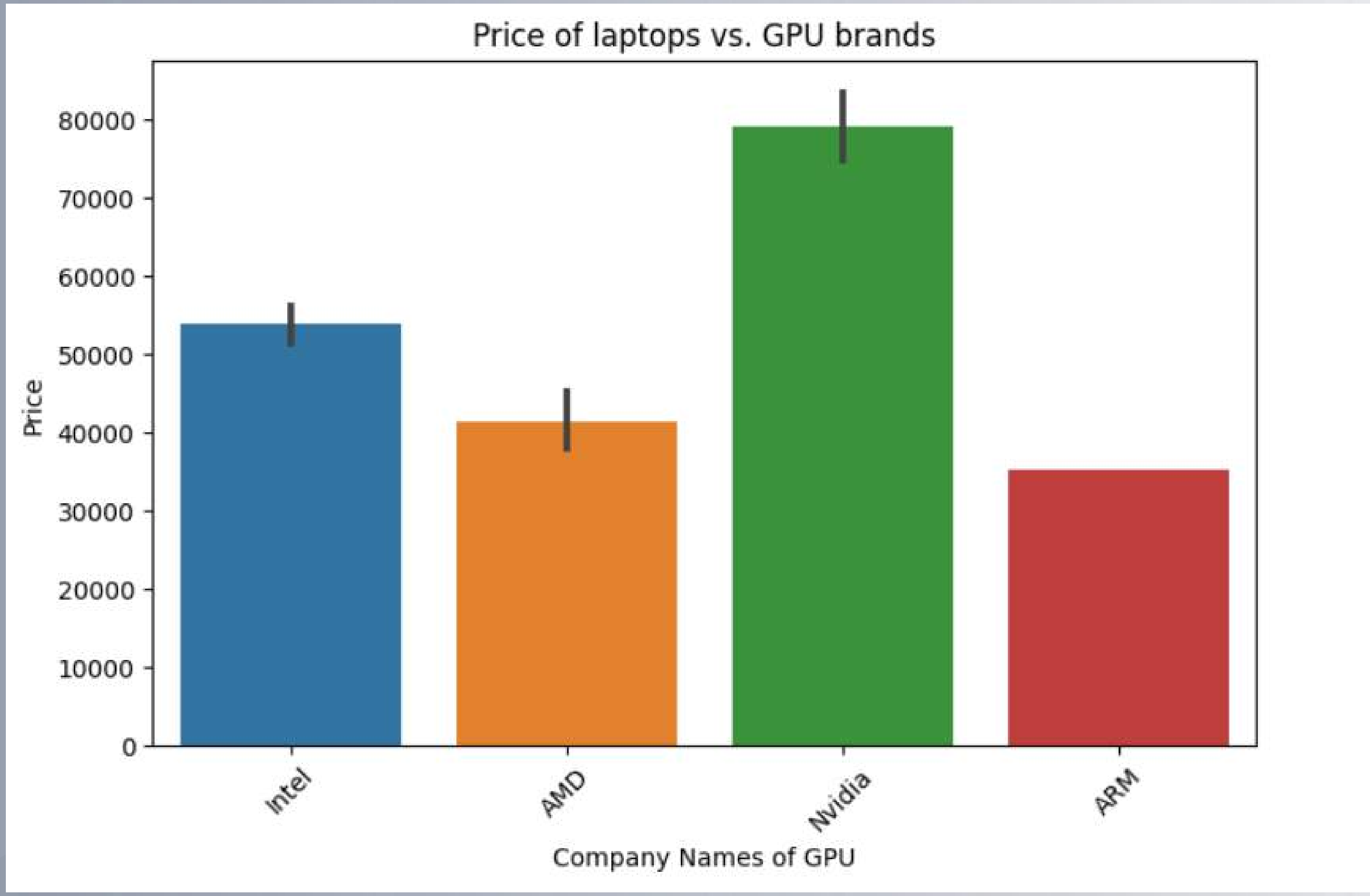
Most of the laptops sold are without touchscreen feature. Also, laptops with touchscreen feature are more expensive than without. Therefore, it might be possible that customers opt for cheaper option.

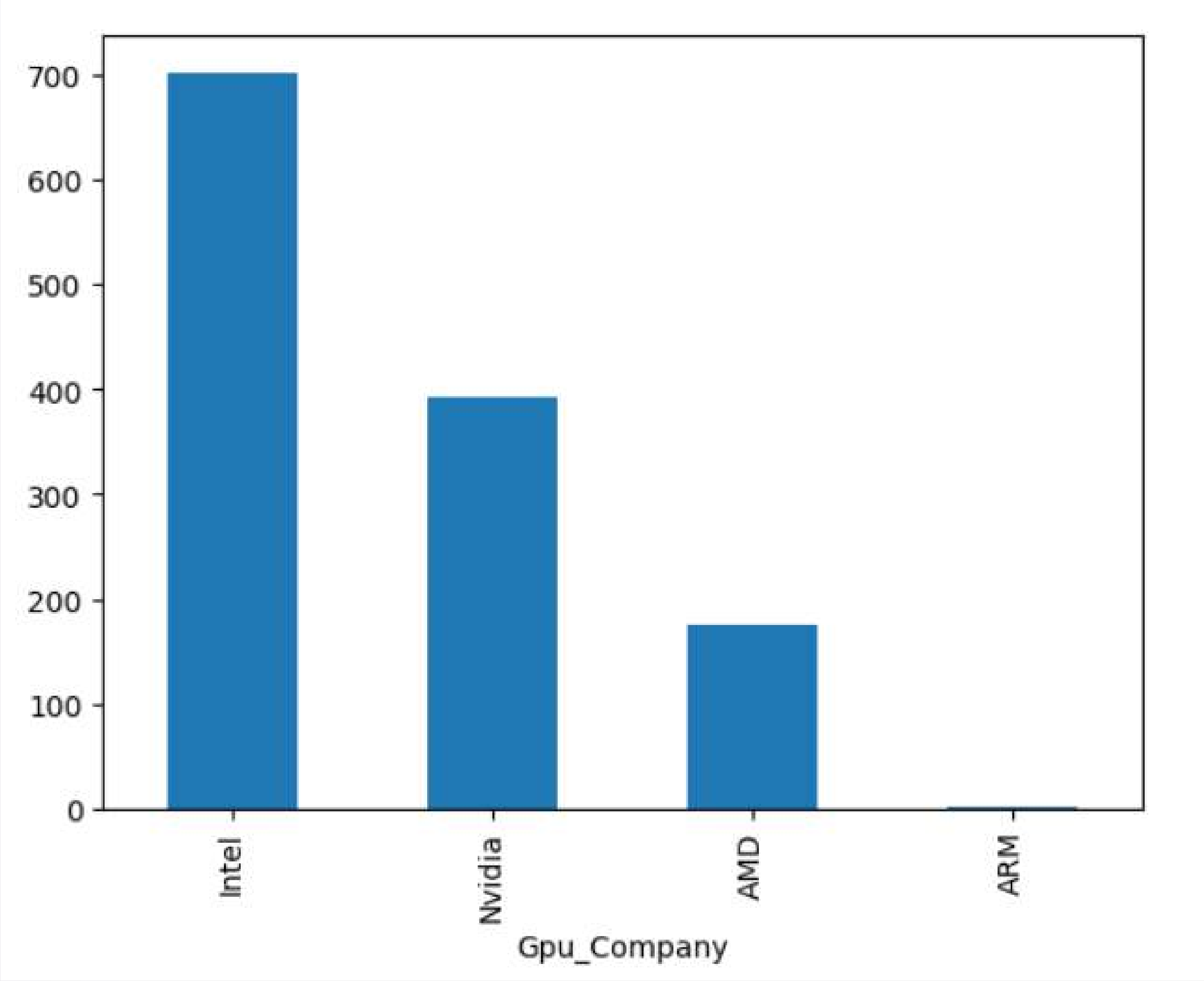




CONTD...

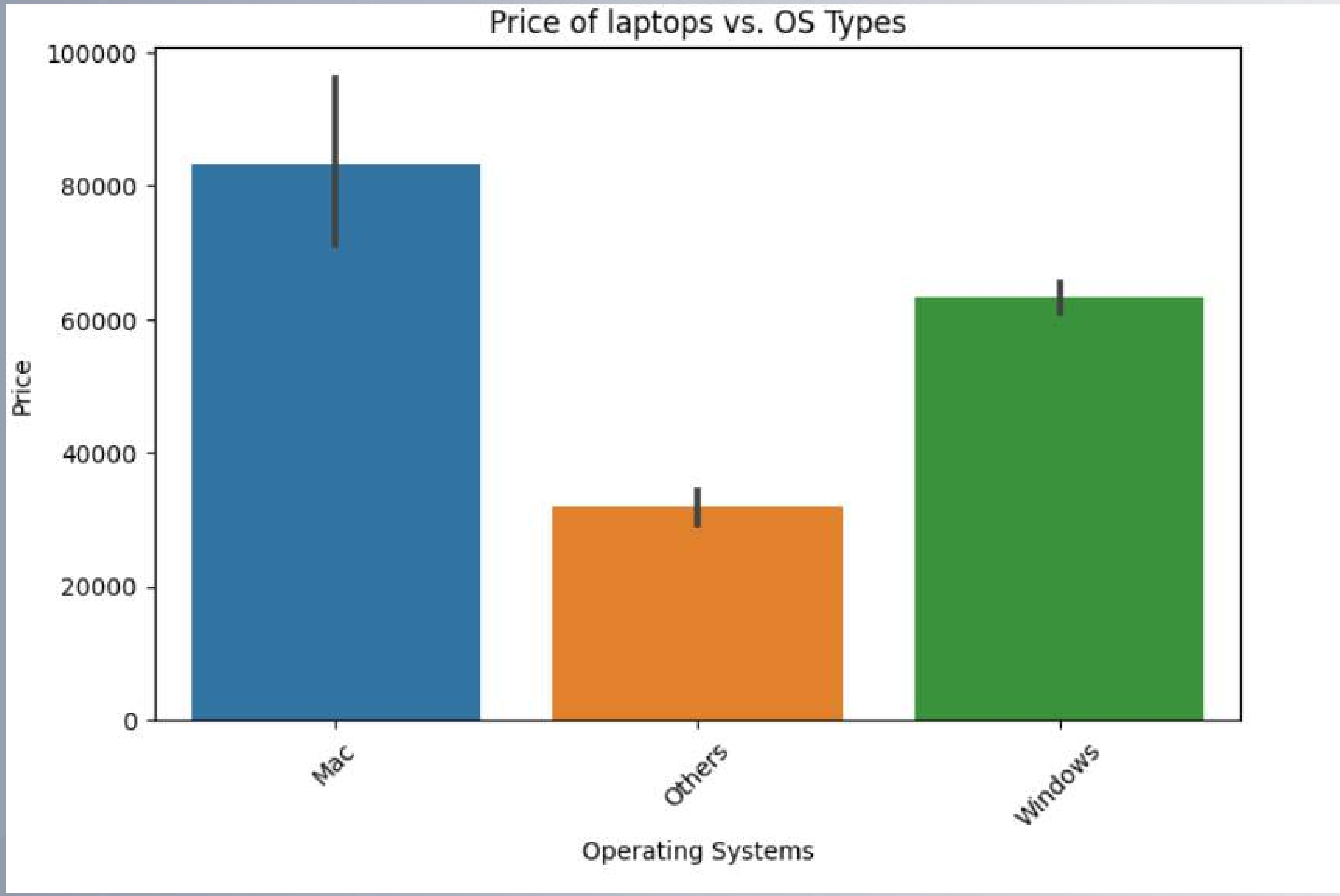
Nvidia Gpus are most expensive followed by Intel and AMD. On the other hand, Intel Gpus are most demanded by customers followed by Nvidia and AMD.

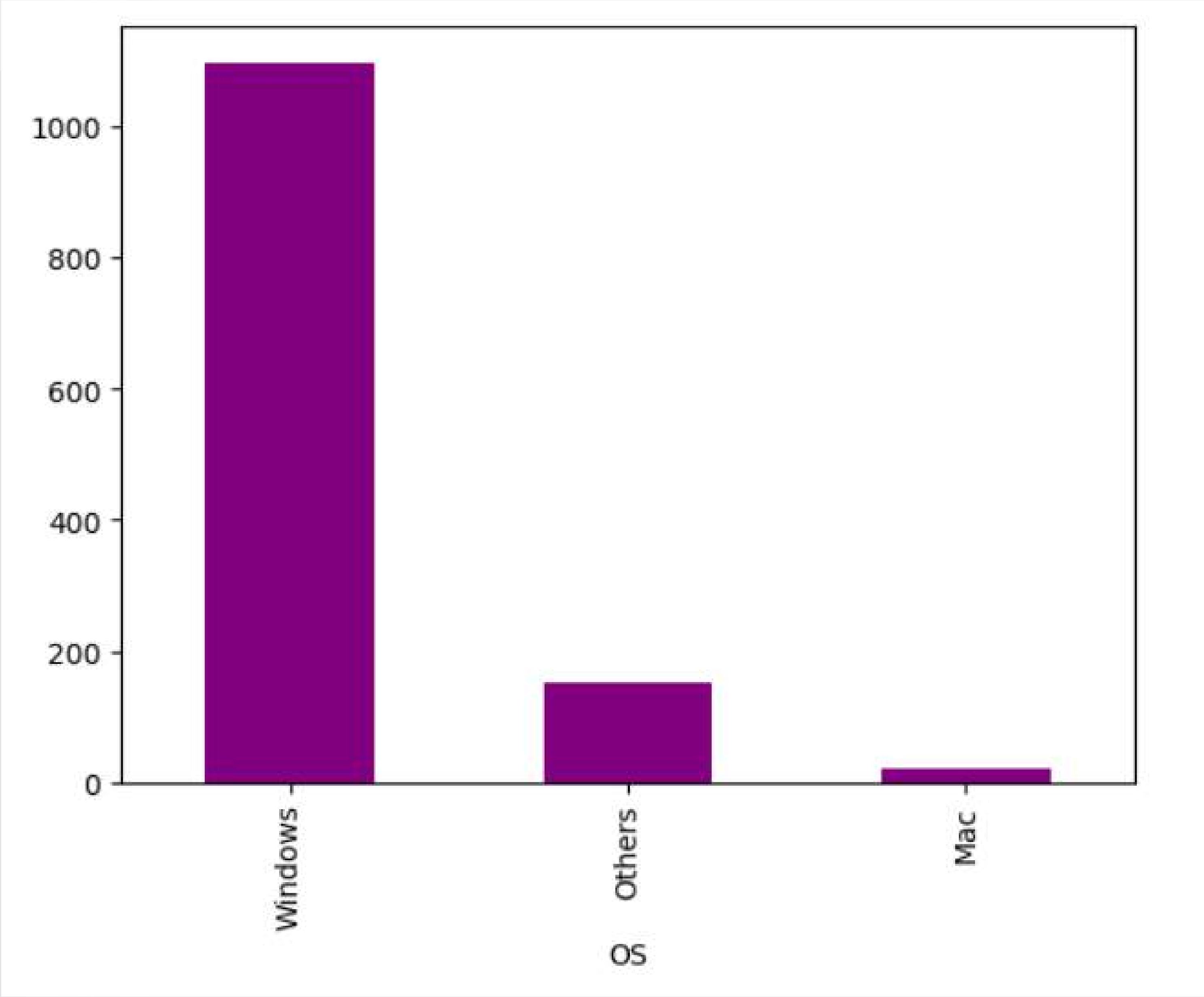




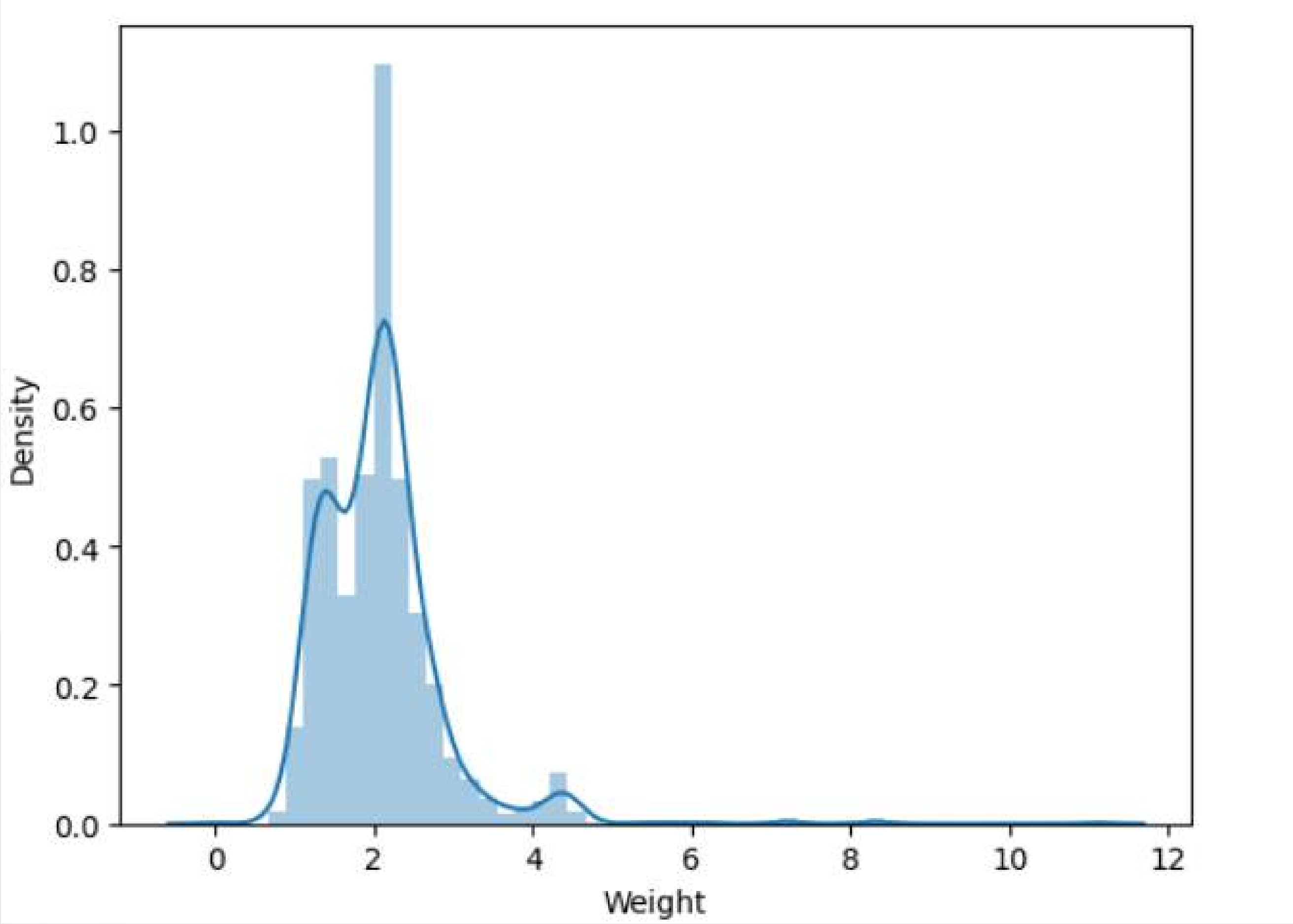
CONTD...

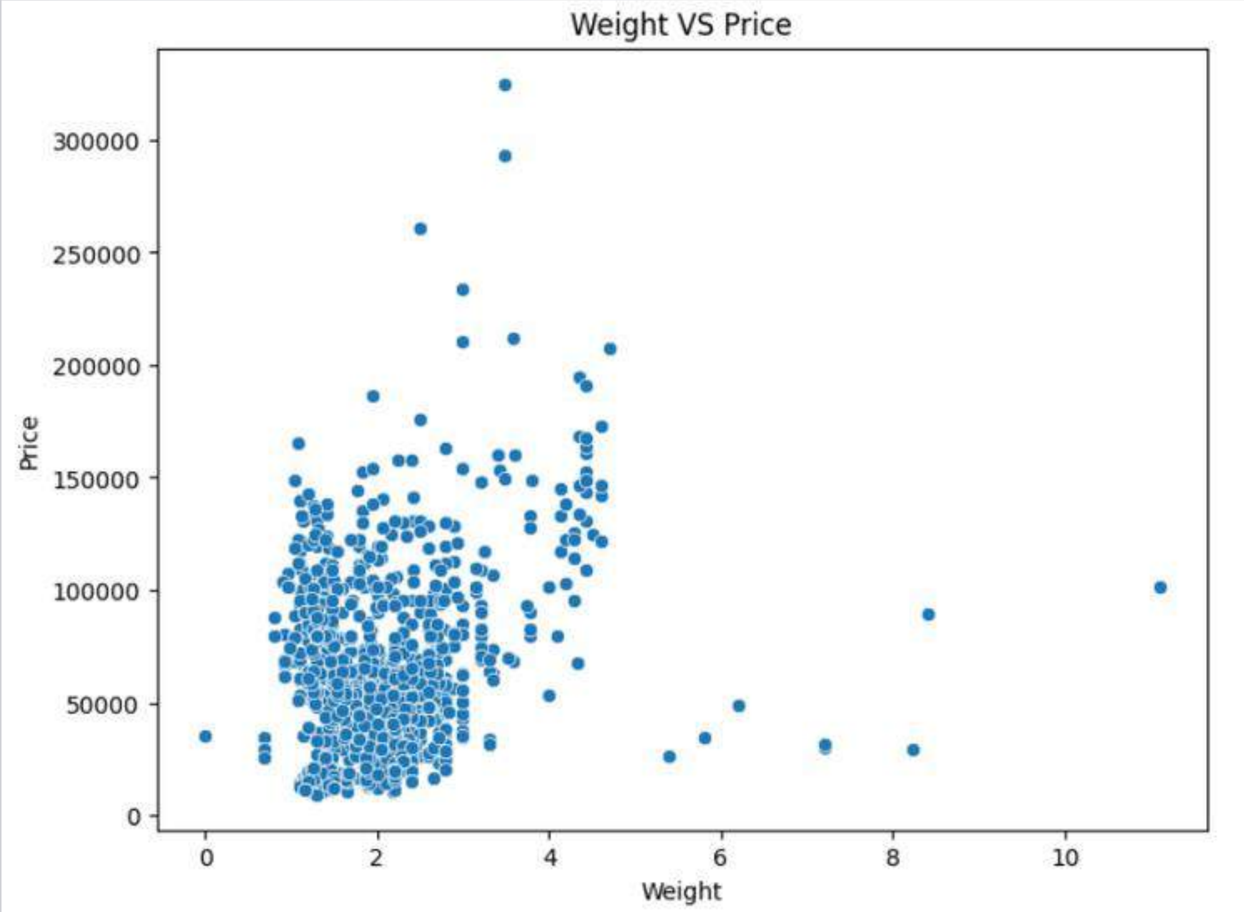
Mac OS are the most expensive, followed by Windows, and Others. Others mainly include No OS, Linux, Chrome, OS, etc. However, the most sold laptops have Windows OS. Mac OS can be afforded by few customers.





CONTD...

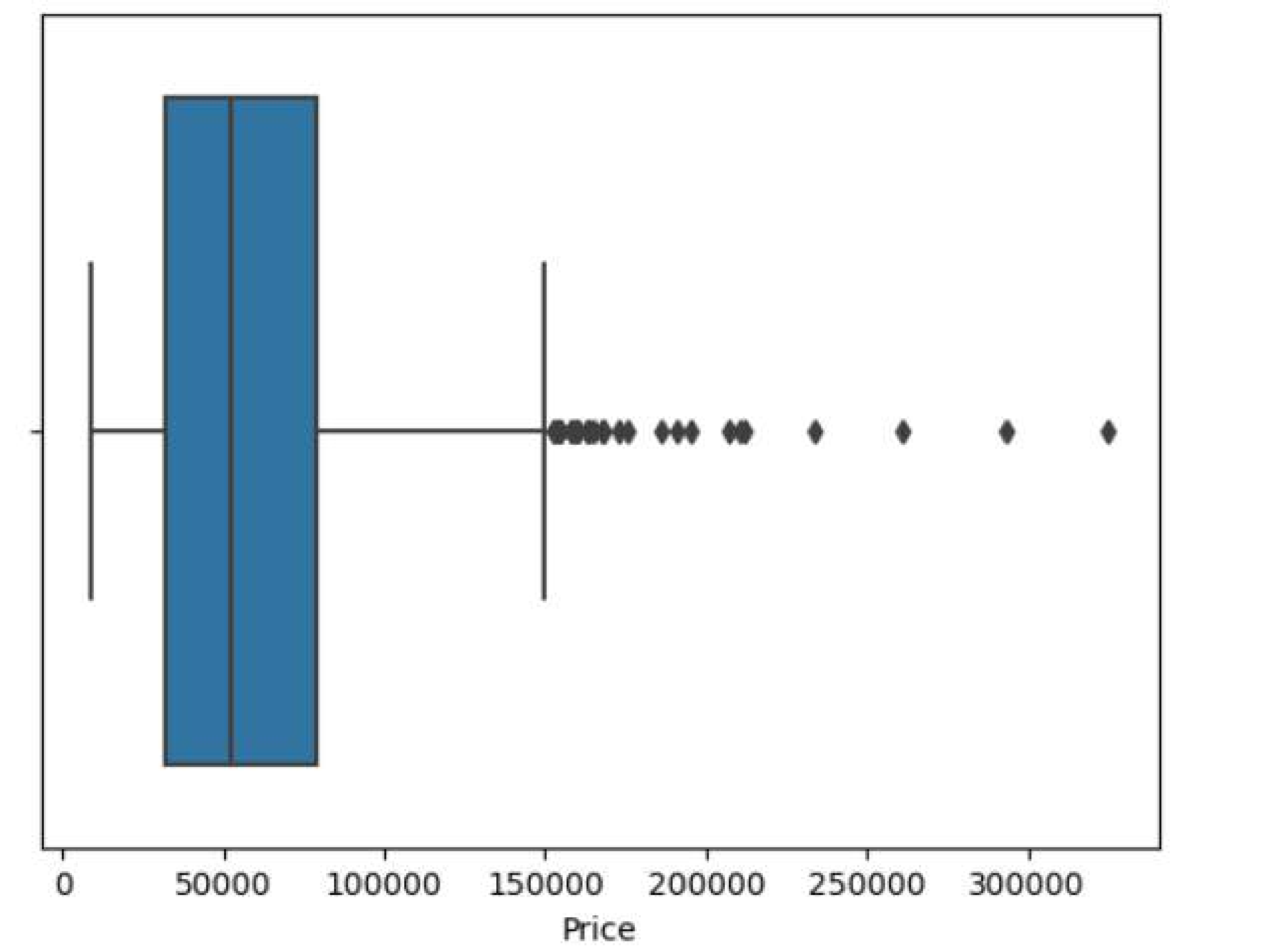


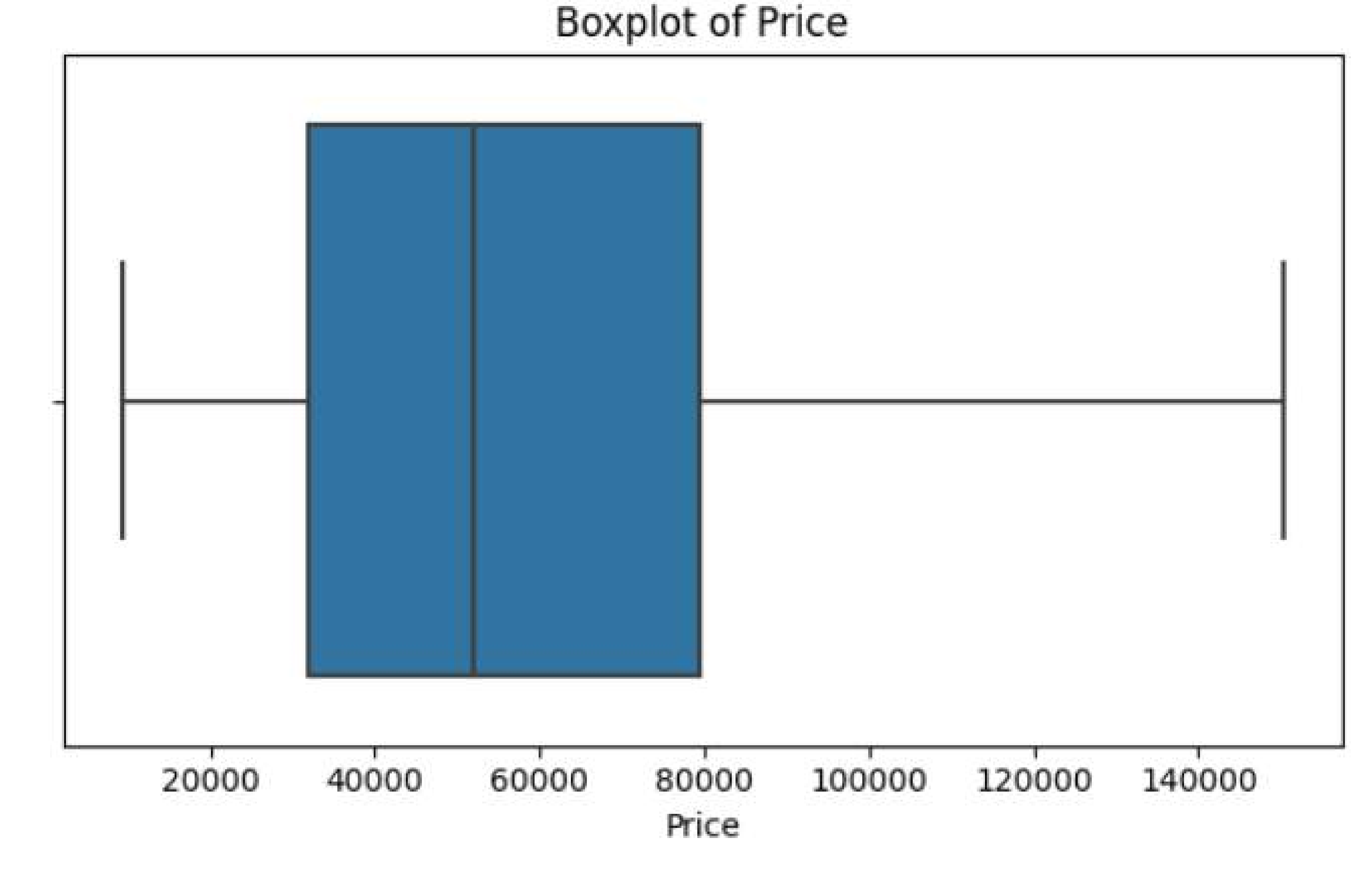


Most laptops weigh 2 kgs. The range is from 0 to 4 kg. However, no clear relationship can be observed between weight and price.

OUTLIER DETECTION AND HANDLING

Outliers are detected with boxplot. Then IQR method was applied with clipping to remove outliers.



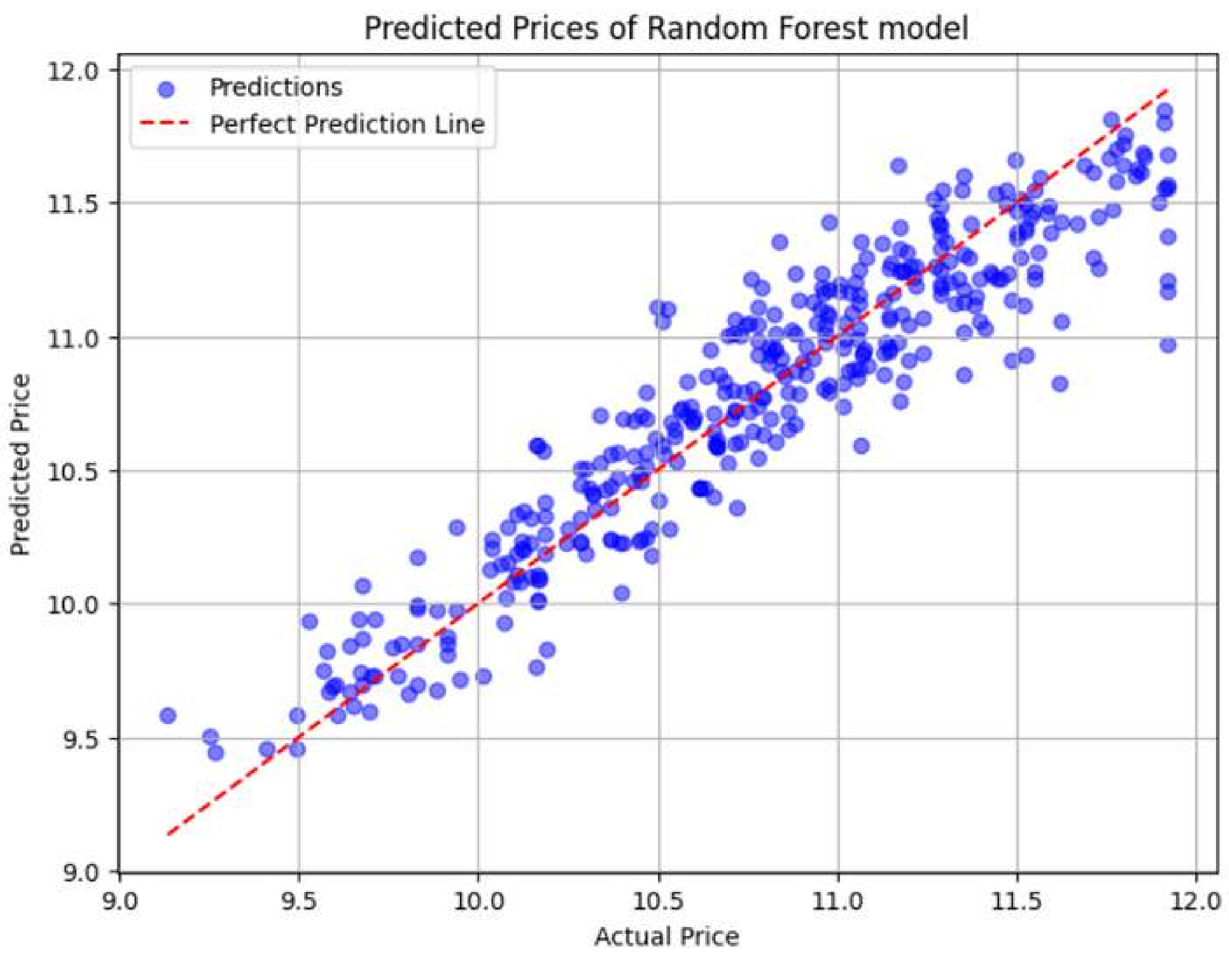


COMPARISON OF MODEL OUTCOMES

Model Name	R2 score	RMSE
Linear Regression	0.7959	0.2778
KNN Regressor	0.7983	0.2761
Decision Tree Regressor	0.8077	0.2696
SVR	0.8075	0.2697
Random Forest Regressor	0.8555	0.2336
Gradient Boosting Regressor	0.7875	0.2834

HYPERPARAMETER TUNING

RandomizedSearchCV is used for hyperparameter tuning because it efficiently explores a wide range of hyperparameter combinations by sampling from specified distributions, making it suitable for large parameter spaces. It can provide good results with fewer iterations compared to GridSearchCV, especially when computational resources are limited.

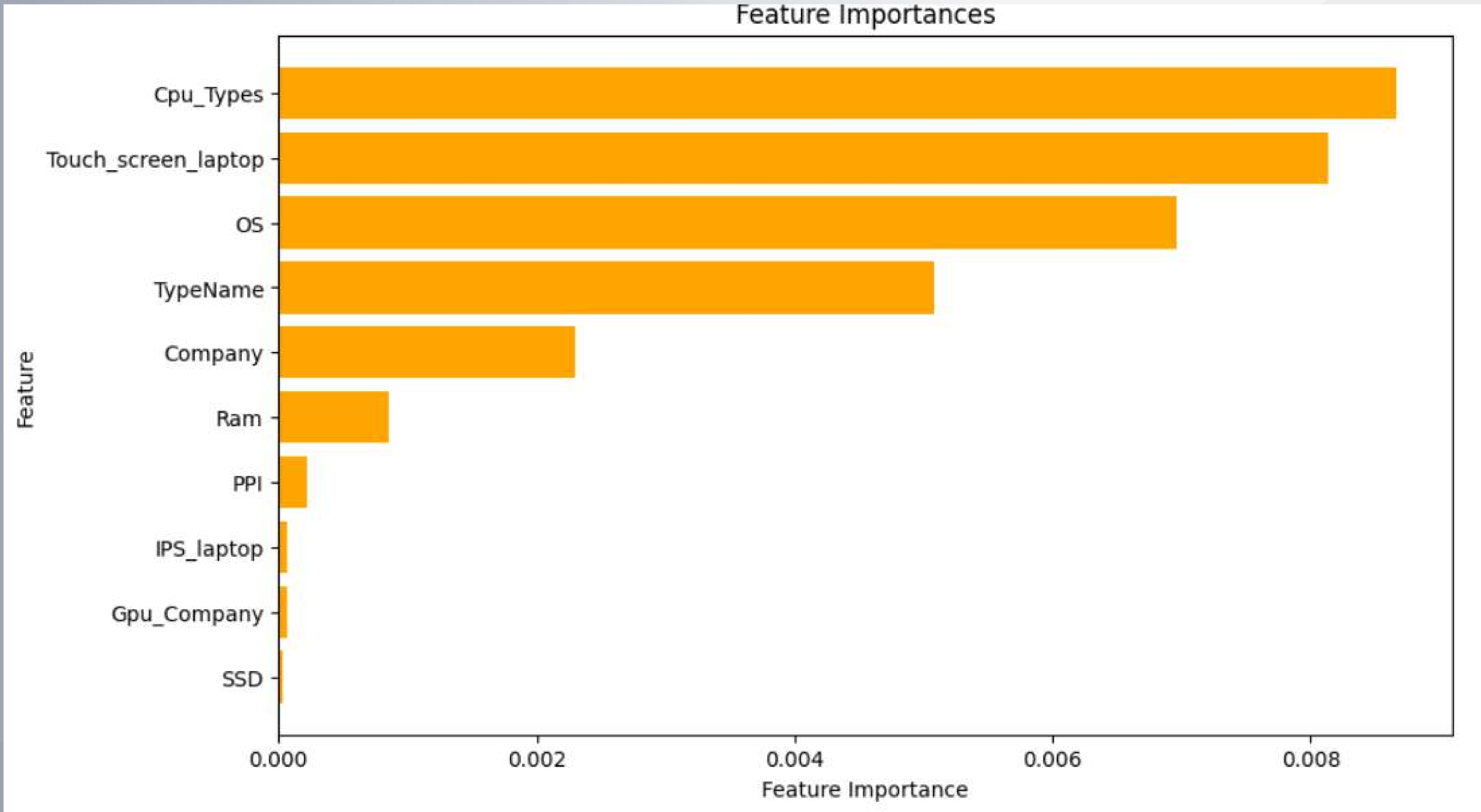


This figure displays a scatter plot comparing the actual prices of items to the predicted prices generated by the Random Forest Regressor model. The scatter points represent individual predictions, while the dashed red line illustrates perfect prediction alignment.

```python
# Define the pipeline
step1 = ColumnTransformer(transformers=[
    ('col_tnf', OneHotEncoder(sparse=False, drop='first'), [0,1,6,8,9])
], remainder='passthrough')

step2 = RandomForestRegressor()

pipe = Pipeline([
    ('step1', step1),
    ('random_search', RandomizedSearchCV(estimator=step2, param_distributions=param_grid,
                                         n_iter=100, cv=5, verbose=2, random_state=3, n_jobs=-1))
])
pipe.fit(X_train, y_train)

best_params = pipe.named_steps['random_search'].best_params_
best_score = pipe.named_steps['random_search'].best_score_

print("Best Parameters:", best_params)
print("Best Score:", best_score)

feature_importances = pipe.named_steps['random_search'].best_estimator_.feature_importances_

y_pred = pipe.predict(X_test)
```

```
Fitting 5 folds for each of 100 candidates, totalling 500 fits
Best Parameters: {'n_estimators': 100, 'min_samples_split': 6, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth':
92, 'bootstrap': False}
Best Score: 0.8670196422304134
R2 score: 0.8698643952682927
RMSE: 0.22185335243247434
```

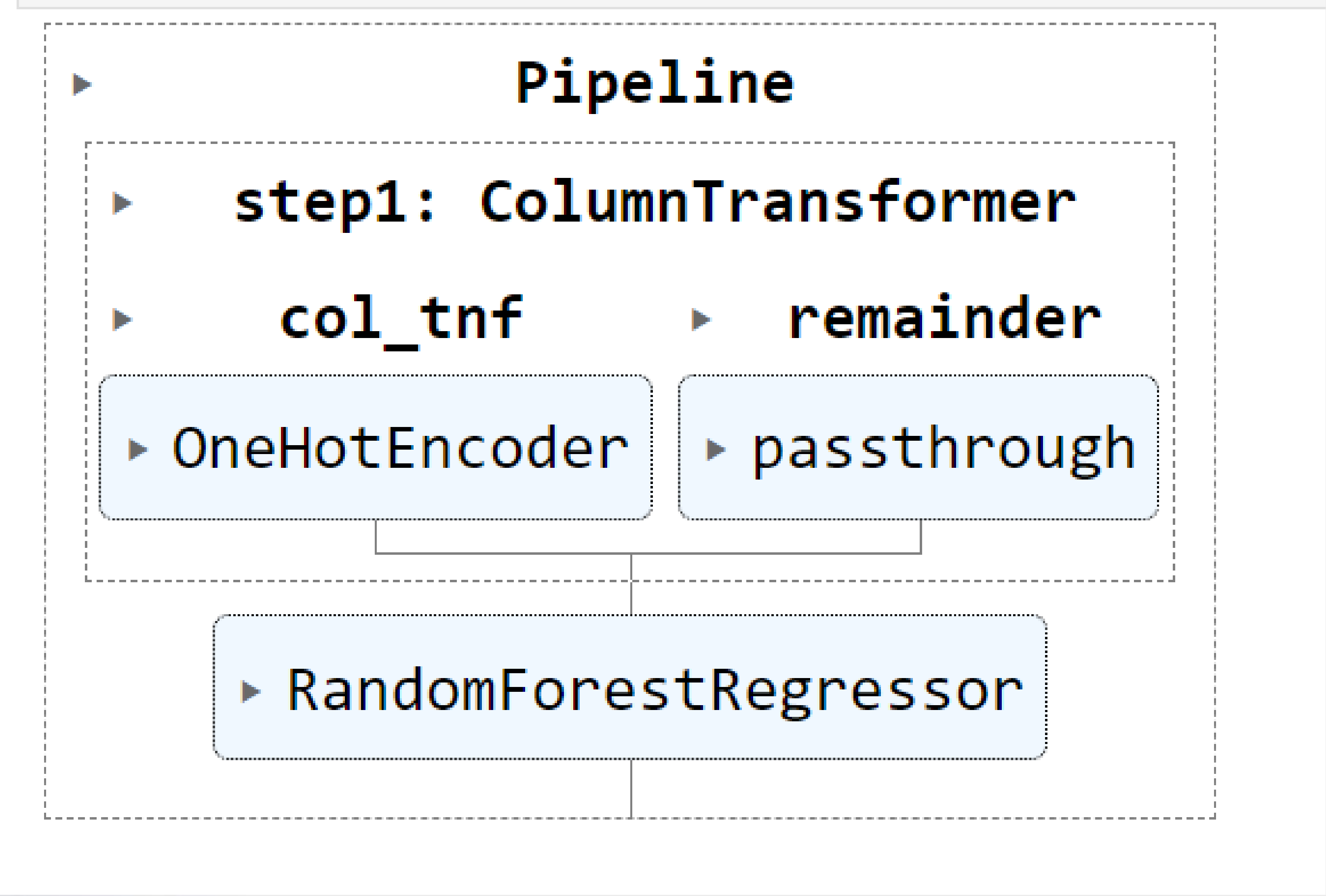
MOST IMPORTANT FEATURES



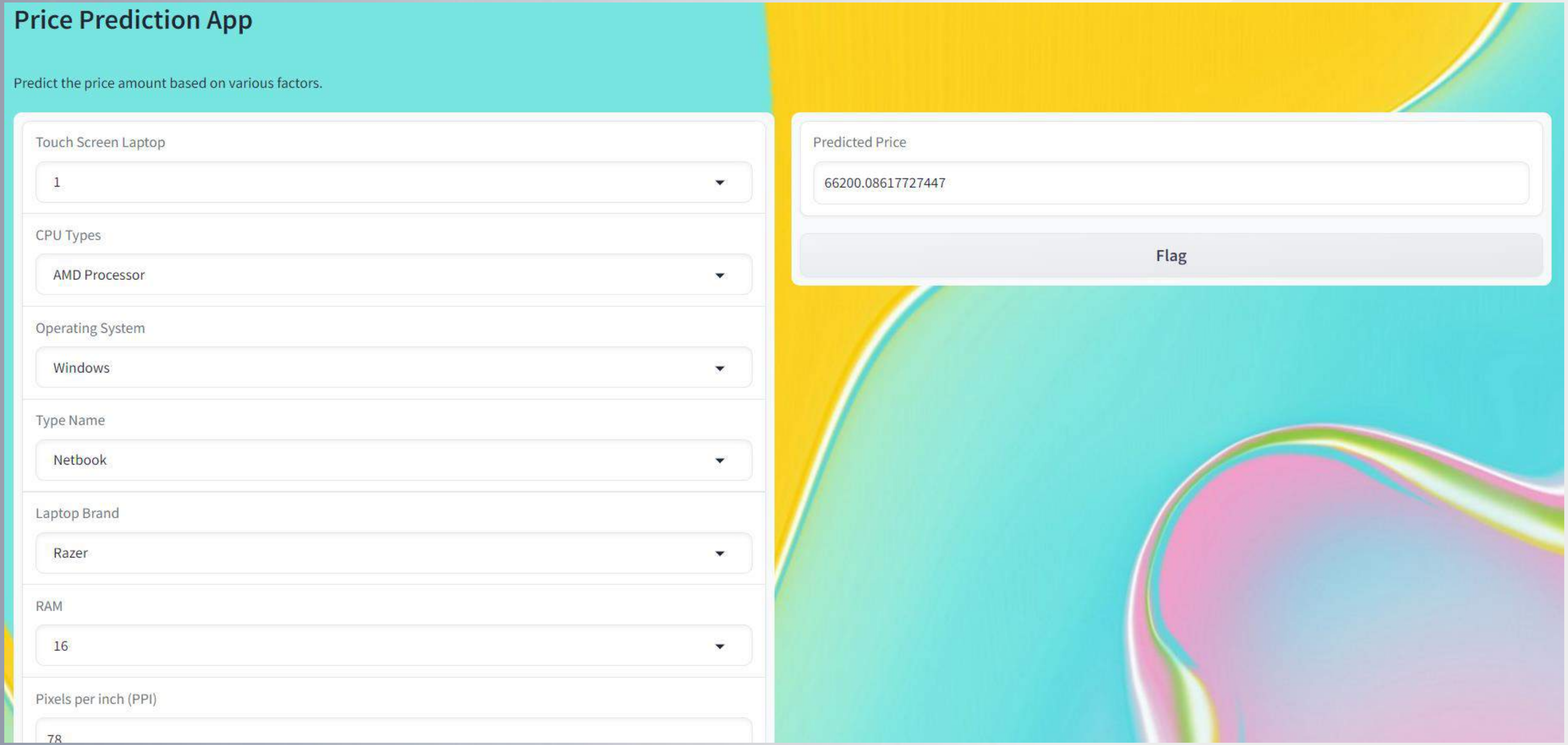
FEATURE SELECTION AND MODEL BUILDING

```
X_train_new
```

	Touch_screen_laptop	Cpu_Types	OS	TypeName	Company	Ram	PPI
945	0	AMD Processor	Windows	Gaming	Lenovo	16	141.211998
477	0	Intel Core i5	Windows	Notebook	Lenovo	8	157.350512
85	0	Intel Core i7	Windows	Gaming	Dell	16	141.211998
1034	0	Intel Core i7	Windows	Notebook	HP	8	141.211998
662	0	Intel Core i5	Windows	Notebook	Lenovo	4	141.211998
...	...	...	...	...	...	...	...
480	0	Intel Core i7	Windows	Notebook	Dell	8	141.211998
309	0	Intel Core i3	Windows	Notebook	HP	4	141.211998
507	0	Intel Core i5	Windows	Notebook	Acer	4	141.211998
541	0	Intel Core i3	Windows	Notebook	Dell	4	157.350512
1224	1	Intel Core i3	Windows	2 in 1 Convertible	Dell	4	146.860478



GRADIO APP FOR REAL-TIME PREDICTIONS



CHALLENGES FACED

- Ensuring project quality while iterating on data pre-processing, model training, and evaluation phases
- Dealing with missing values, categorical variables, and feature engineering which required careful handling and informed decision-making
- Choosing the most suitable algorithms among various options, considering their interpretability, and computational complexity

BUSINESS QUESTIONS

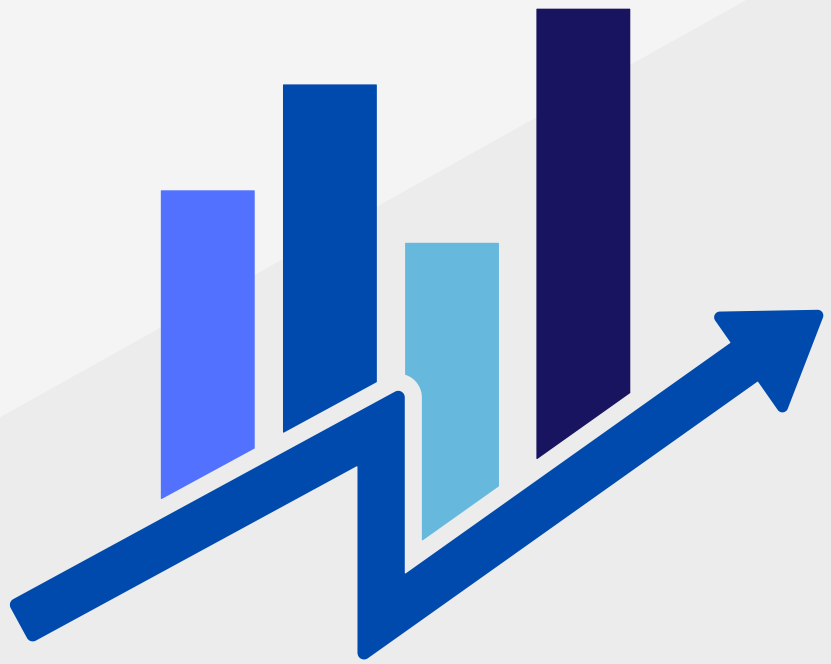
1. The features with the most significant impact on laptop prices, according to the model's insights, include whether the laptop has a touchscreen, the type of CPU, the operating system (OS), the laptop type, the brand, RAM capacity, and pixel density (PPI).
2. The model can predict the prices of laptops from lesser-known brands that are present in the dataset like Fujitsu and Vero.
3. The brand of the laptop significantly influences its price, with Razer laptops being the most expensive, followed by other mid-range brands like Lenovo, Dell, Asus, Acer, and HP. Lesser-known brands may offer more competitively priced options
4. High-end features include Gaming and Ultrabook categories, Intel Core i7 processors, laptops with touchscreen features, Nvidia GPUs, and Mac OS. These features are associated with higher pricing and are considered premium options compared to budget alternatives like Notebook laptops with Intel Core i5 processors and Windows OS. The model performs reasonably well on both categories.
5. Limitations and challenges in predicting laptop prices accurately can be changes in market trends or consumer preferences over time, and the complexity of factors influencing laptop prices beyond the features included in the model.
6. The model struggles when predicting the prices of newly released laptops not present in the training dataset as it lacks historical data on those specific models.

RECOMMENDATIONS

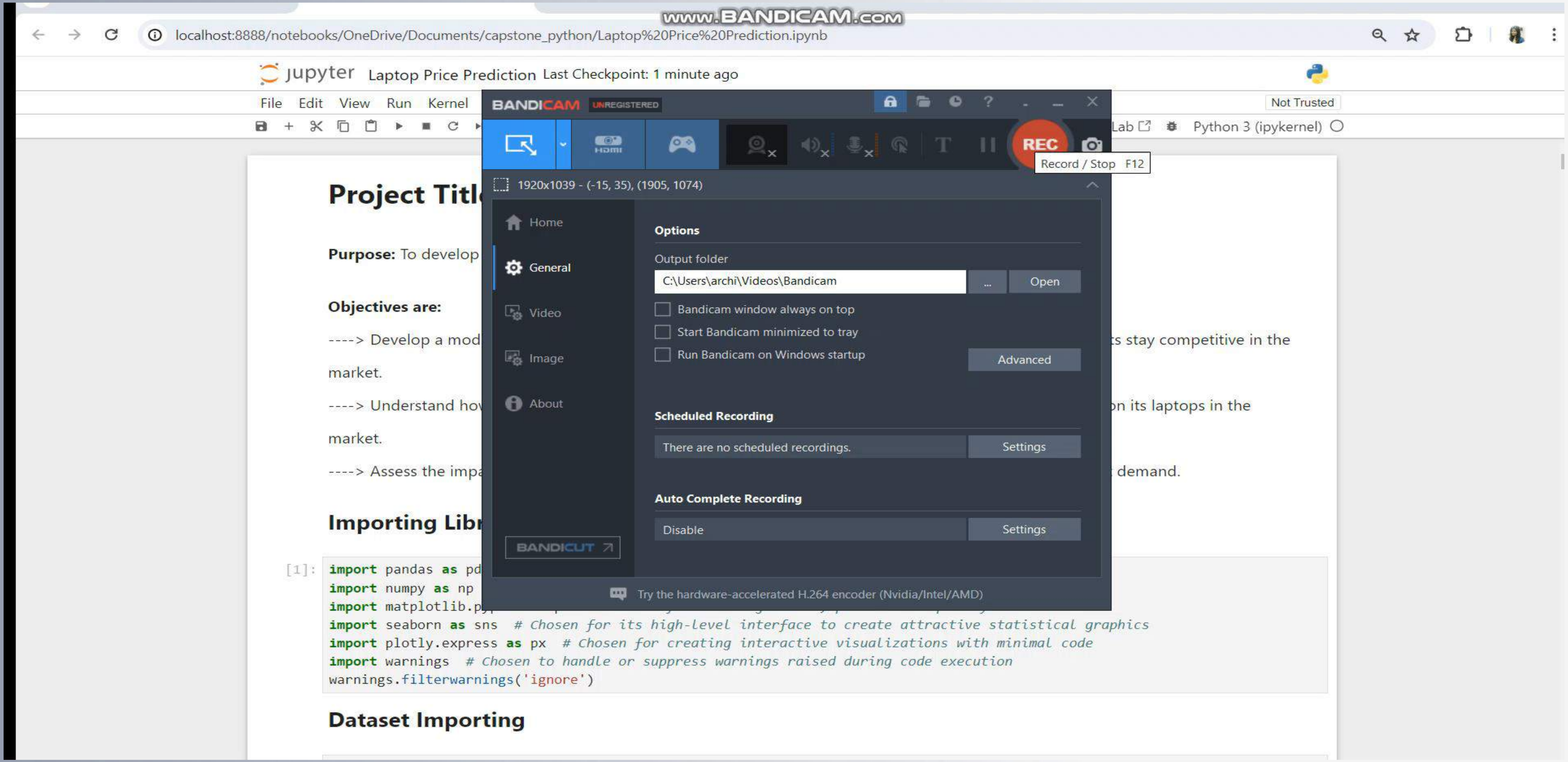
- Consider expanding the brand portfolio to include lesser-known brands like Fujitsu and Vero to offer more competitively priced options
- Invest in marketing efforts targeting high-end segments, emphasizing features like Gaming and Ultrabook categories, and Intel Core i7 processors, to capitalize on their association with premium pricing
- Stay updated on market trends and consumer preferences to adapt pricing strategies accordingly
- Educate customers on the value proposition of high-end features, highlighting their benefits in terms of performance, durability, and user experience to justify premium pricing and increase willingness to pay
- Continuously benchmark prices against competitors, particularly mid-range brands like Lenovo, Dell, Asus, Acer, and HP, to ensure competitiveness while maintaining profitability
- Consider promotional pricing strategies to stimulate demand and capture market share

FUTURE SCOPE

- Explore the implementation of other machine learning models such as XGBoost or Gradient Boosting Machines to enhance the accuracy and robustness
- Investigate deep learning architectures such as recurrent neural networks (RNNs) or long short-term memory (LSTM) networks to uncover complex relationships
- Refine the model to predict emerging features' impact on laptop prices



VIDEO RECORDING



THANK YOU

Hope you liked it.