Laptop Price Prediction Model

(by Harsh Singh)

Introduction

Objective

- Develop a Machine Learning web application to predict laptop prices based on various features such as brand, specifications accurately for SmartTech Co. The model aims to provide a reliable tool for setting competitive prices, understanding market dynamics, and making data-driven decisions to enhance pricing strategies and overall market performance.

Scope

- Data exploration and preprocessing.
- Feature engineering.
- Model development and evaluation.
- Real-time prediction
- Interpretability and Insights

Data Set

The dataset contains numerous columns with a lot of noise and redundant information. Despite these challenges, through effective ,data cleaning and feature engineering, we can extract relevant features and achieve better results. Although the dataset is relatively small, we are confident that our approach will yield good accuracy.

	Unnamed: 0.1	Unnamed: 0	Company	TypeName	Inches	ScreenResolution	Cpu	Ram	Memory	Gpu	OpSys	Weight	Price
0	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	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	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	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	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

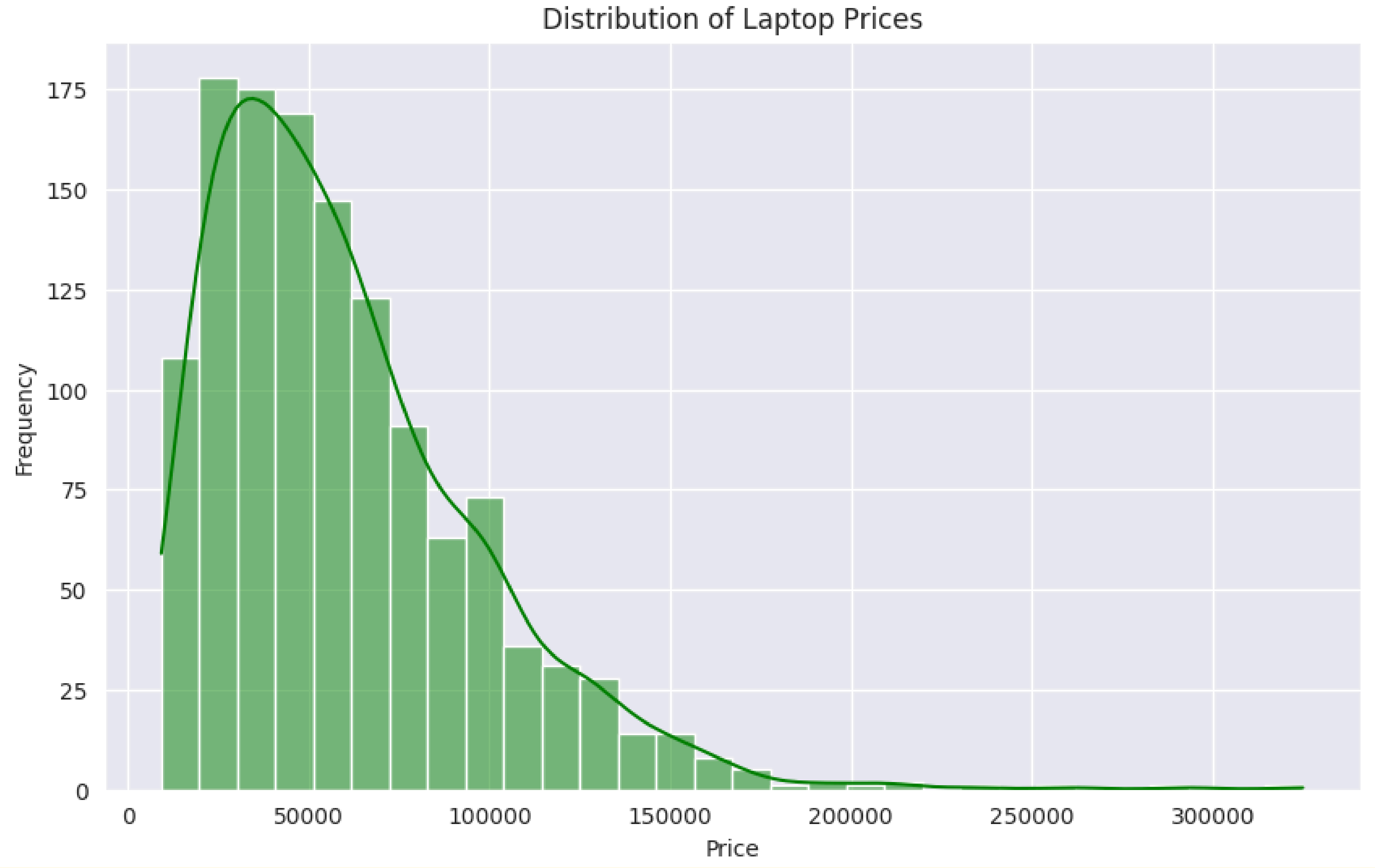
Data Preprocessing

- Import necessary libraries like pandas, numpy, matplotlib,seaborn and scikit-learn
- Load the Dataset and take a basic understanding of dataset like shape,info etc.
- Total Rows - 1303 and Total Columns - 11
- Null values - 30 in each Column

- Dropped unwanted columns ('Unnamed: 0.1', 'Unnamed: 0')
- Handled missing values by dropping rows with null values
- Replaced invalid entries (e.g., replacing '?' with mode values in columns like Weight, Inches, Memory)
- Converted columns to appropriate data types (e.g., Weight to float, Ram to int)
- Remove str values kg, GB from Weight and Ram columns respectively.

EDA (Exploratory Data Analysis) and Feature Engineering

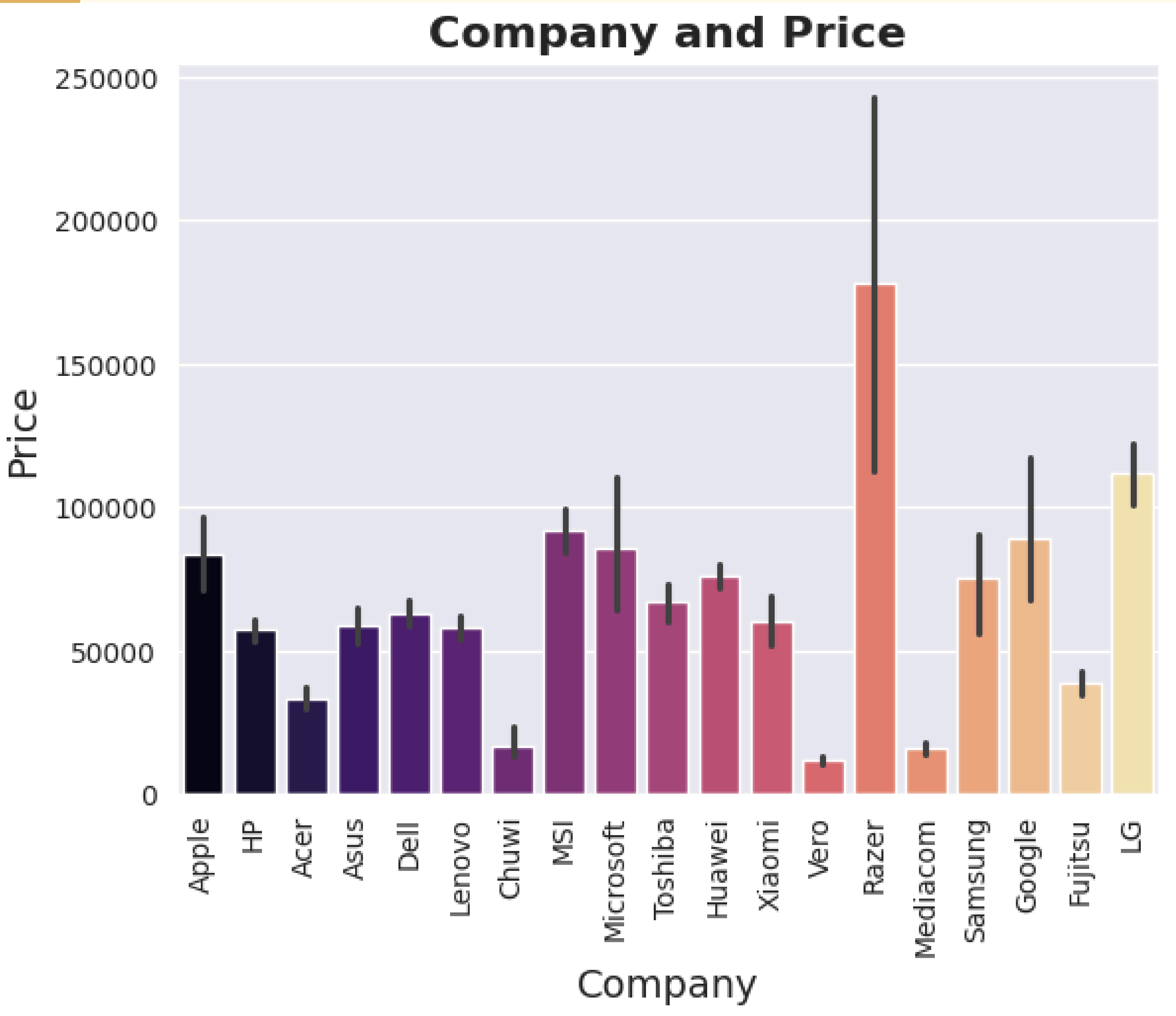
EDA helps in understanding the impact of each feature on the target variable(Price). We will explore each column to assess its influence, performing necessary preprocessing and feature engineering along the way. The goal is to prepare and clean the data for better machine learning modeling, aiming for high performance and generalized models. Let’s start analyzing and preparing the dataset for predictions.

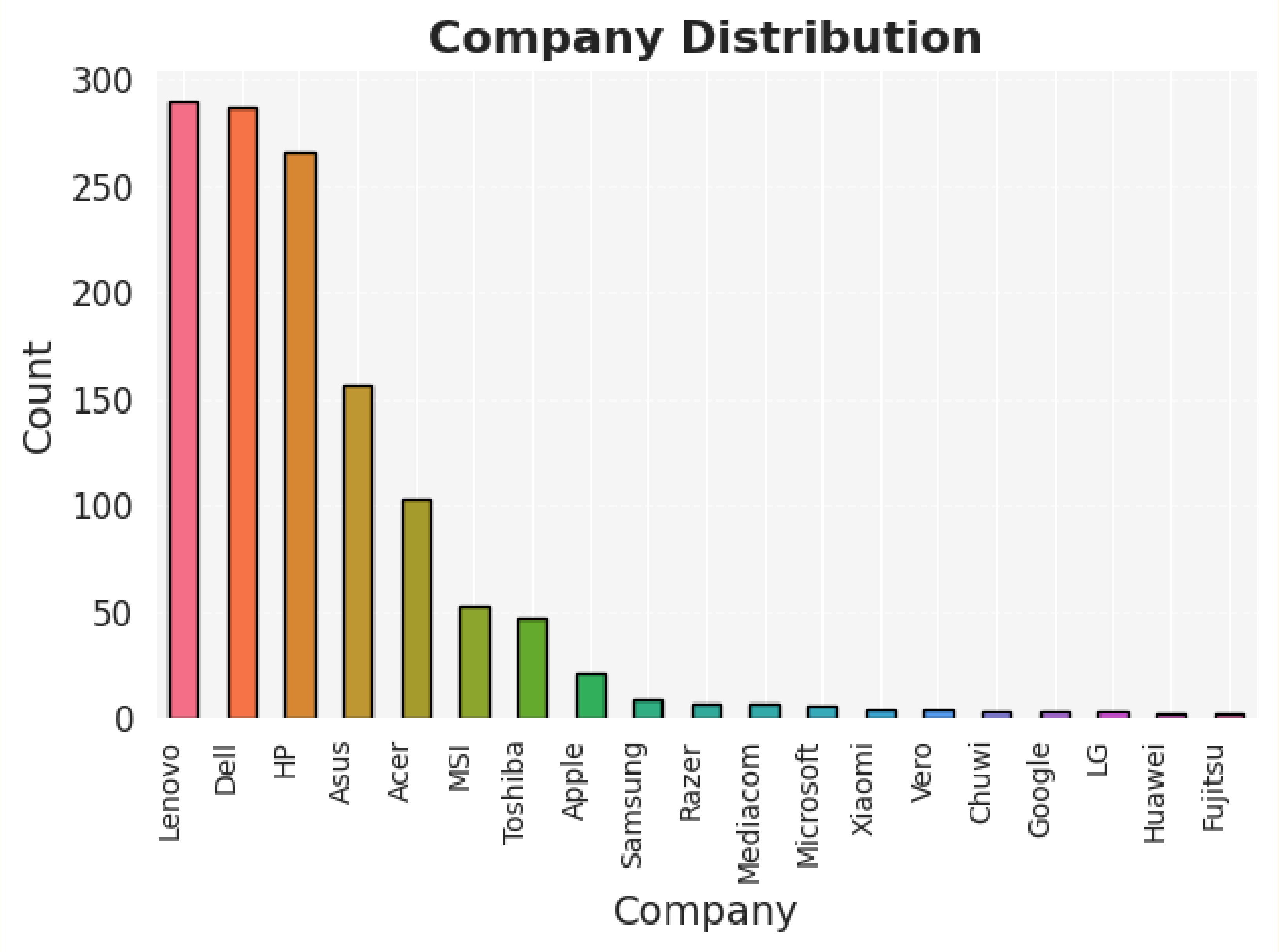


1. Price column(target) distribution

When working with a regression model, it's crucial to understand how the target variable is distributed. In our case, a right-skewed distribution of the target variable (laptop prices) indicates that lower-priced laptops (budget laptops) are more common than higher-priced ones (premium laptops).

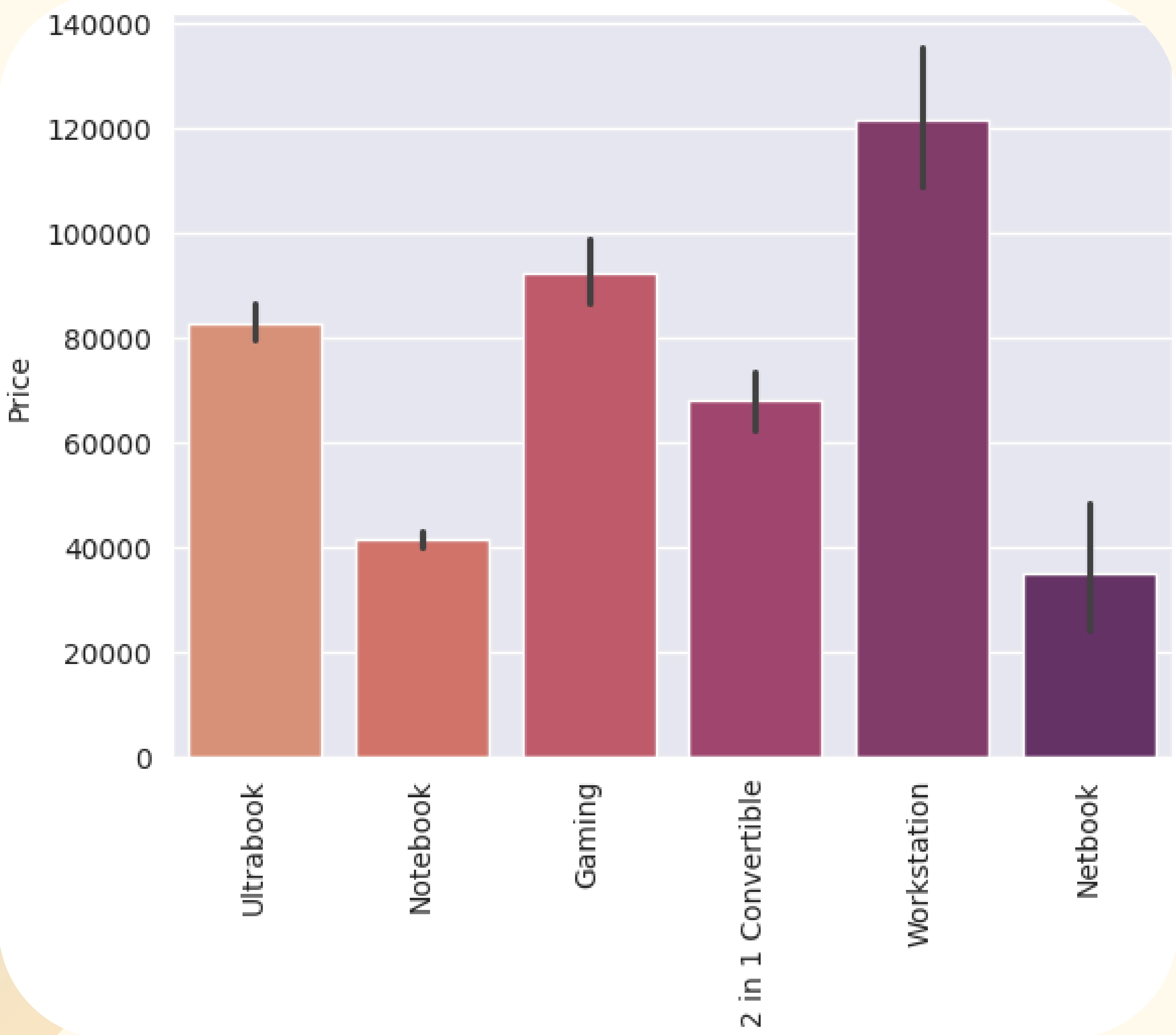
2. Compnay Column

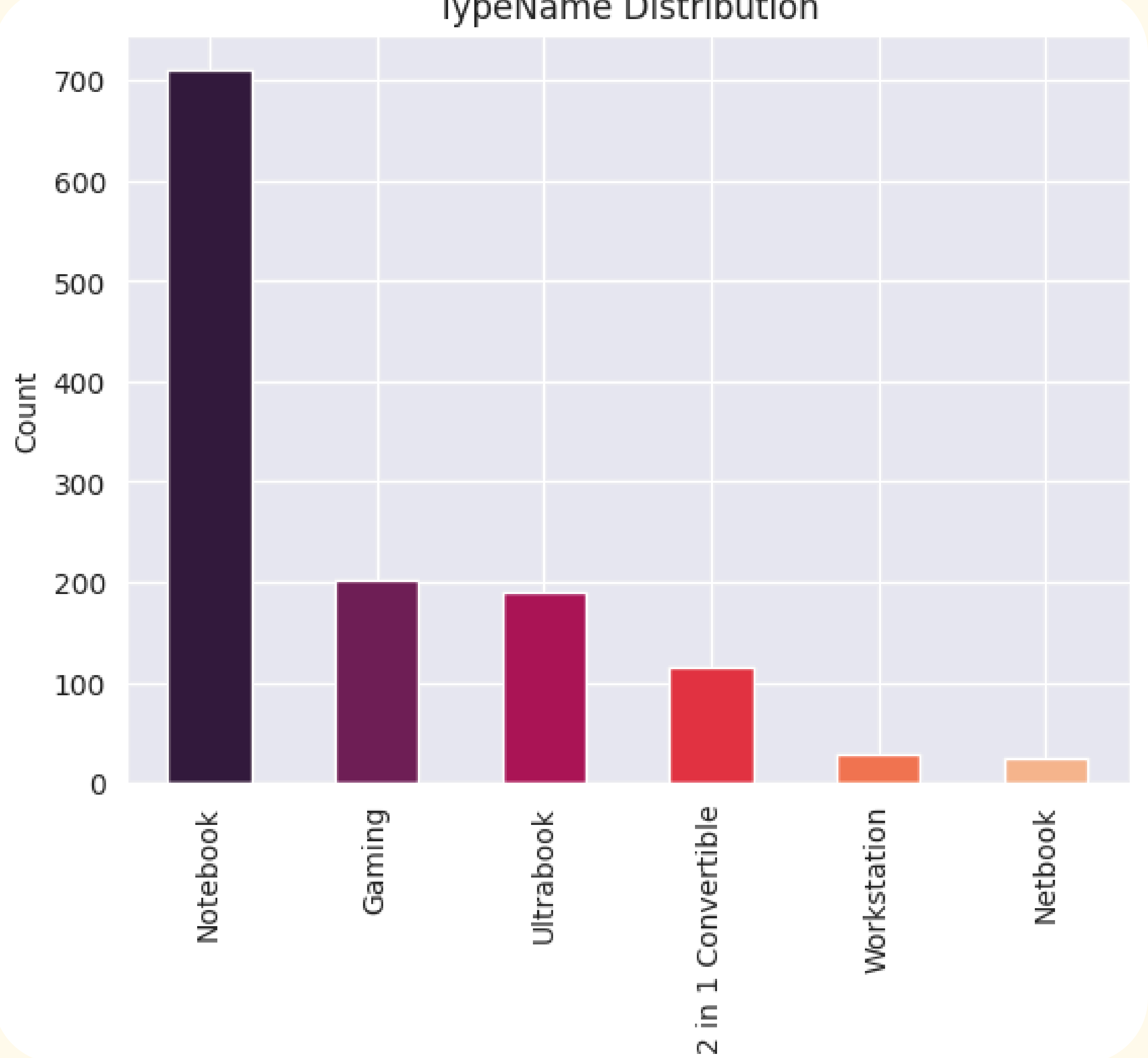




As we Observe from the graph we see that the prices of laptop varies with different brands. Razer, Apple, LG, MSI, Google and Microsoft laptops are expensive and others are in budget range.

3. Laptop Type

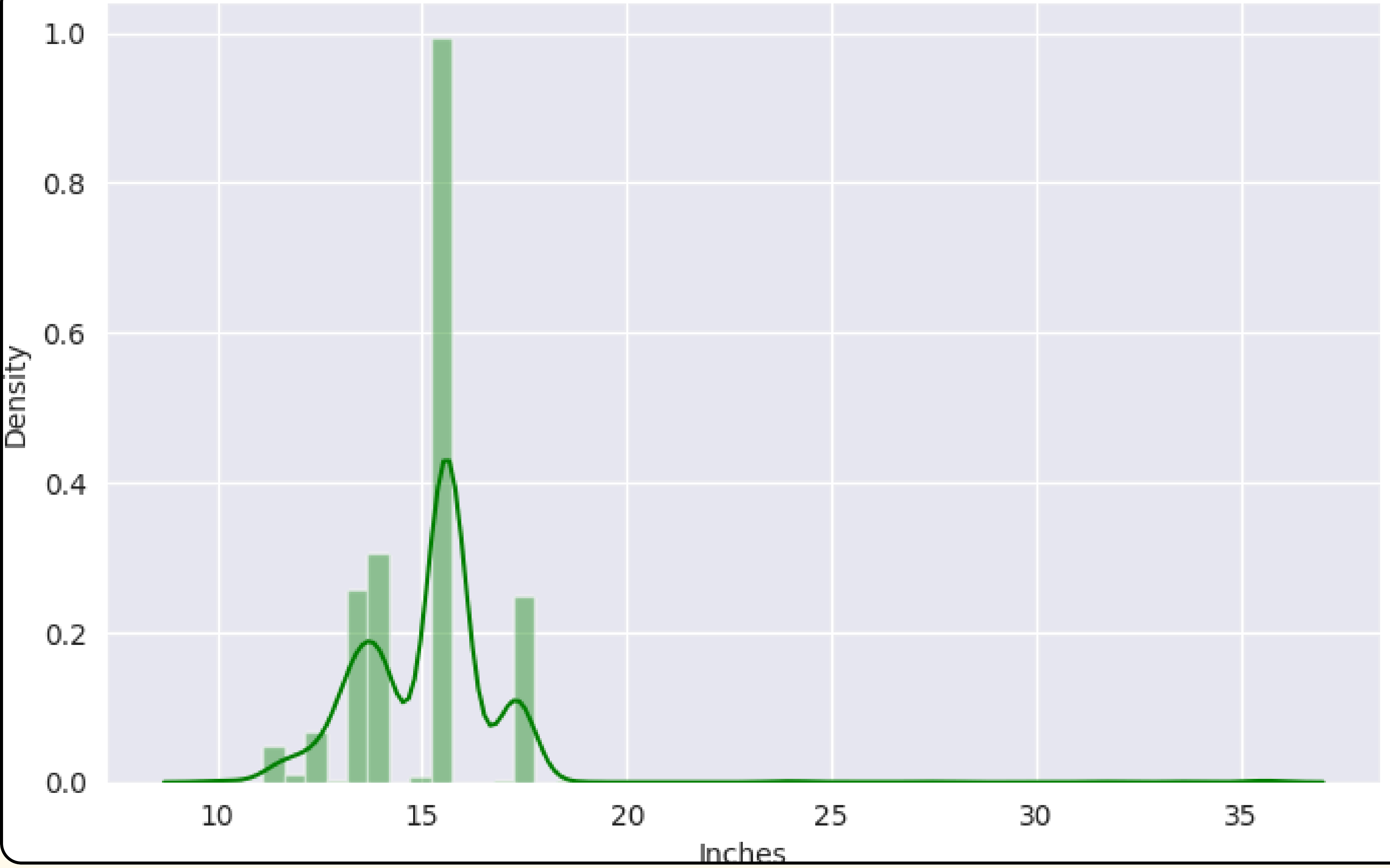


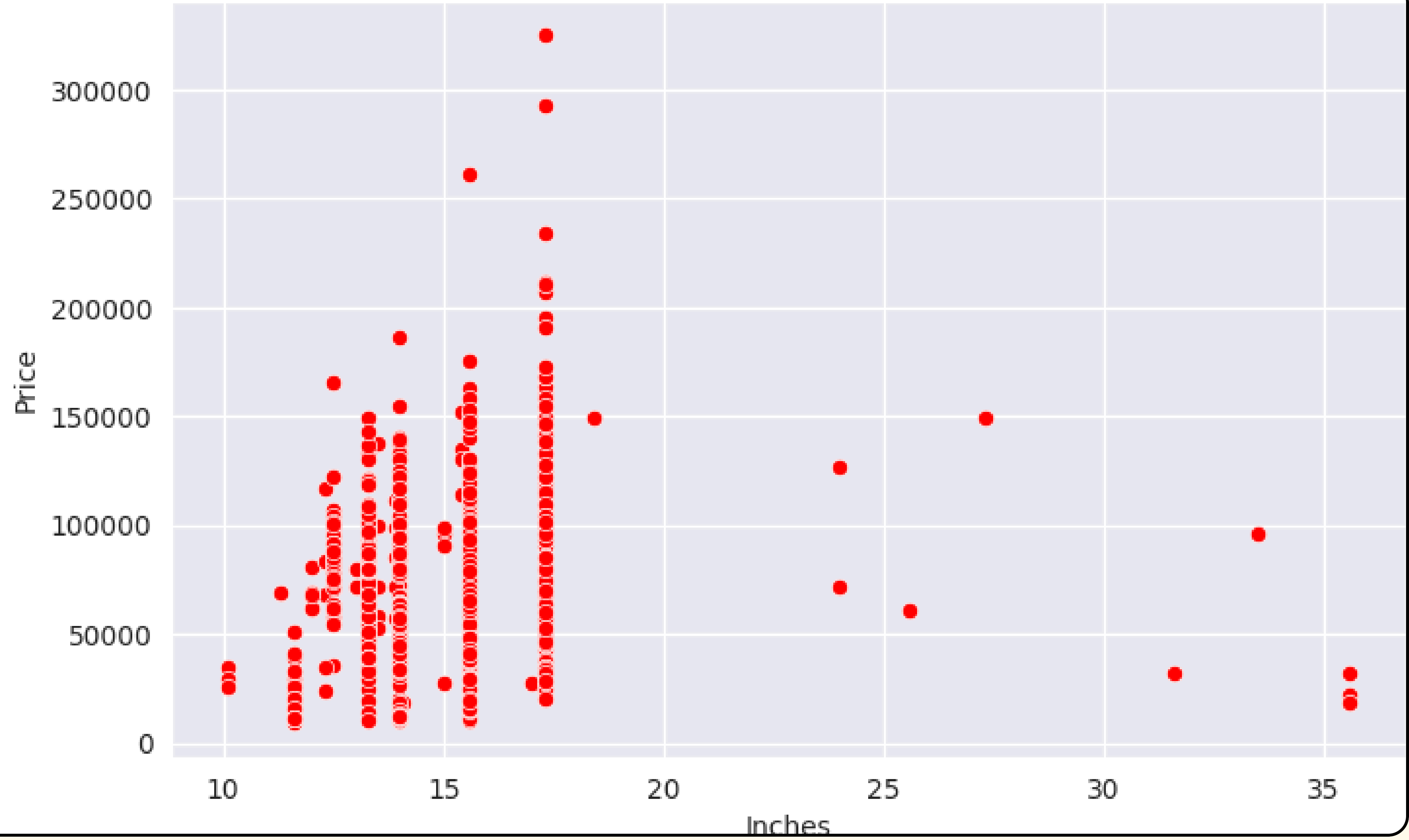


Most of the people prefer noteboook,because it is under budget and the same we can conclude from our data. We also conclude that price of laptop varies with laptop type.

4.Size of laptop

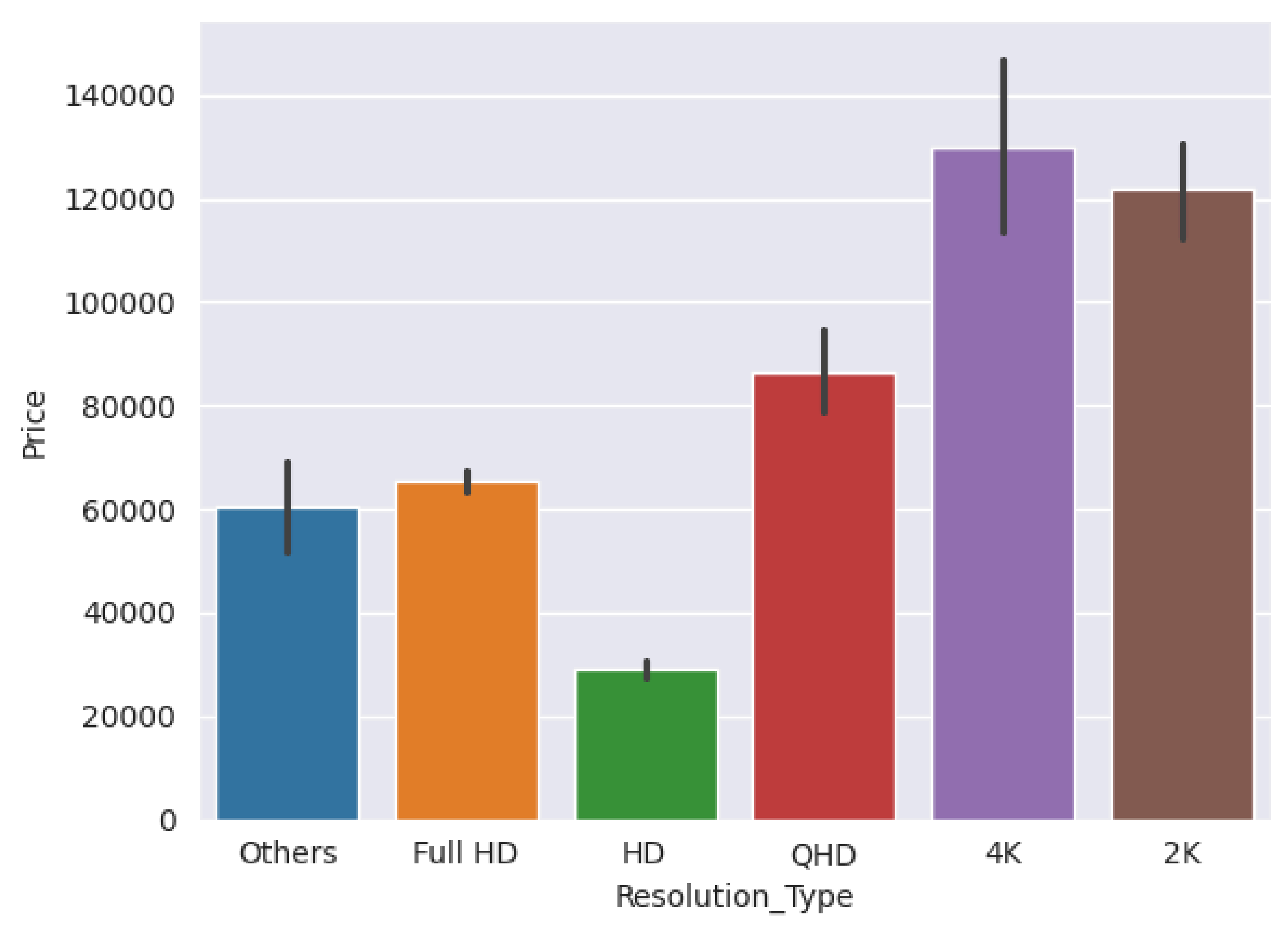
The correlation coefficient between size and price is 0.04, indicating a very weak relationship between the size and price of a laptop.



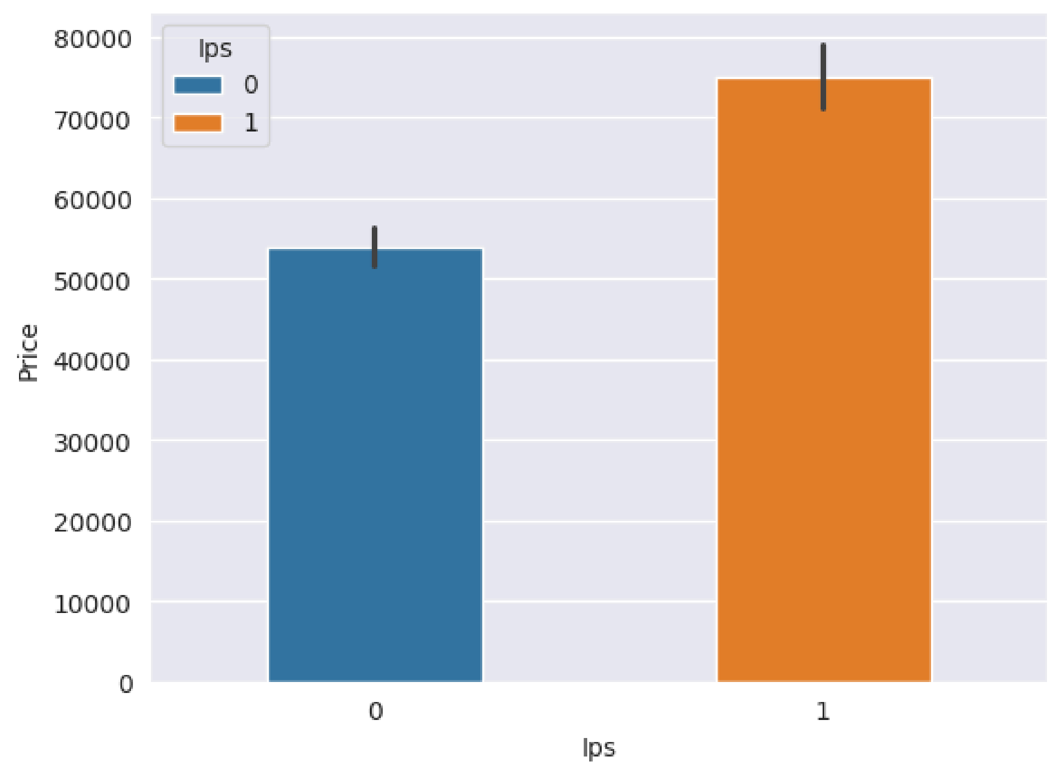


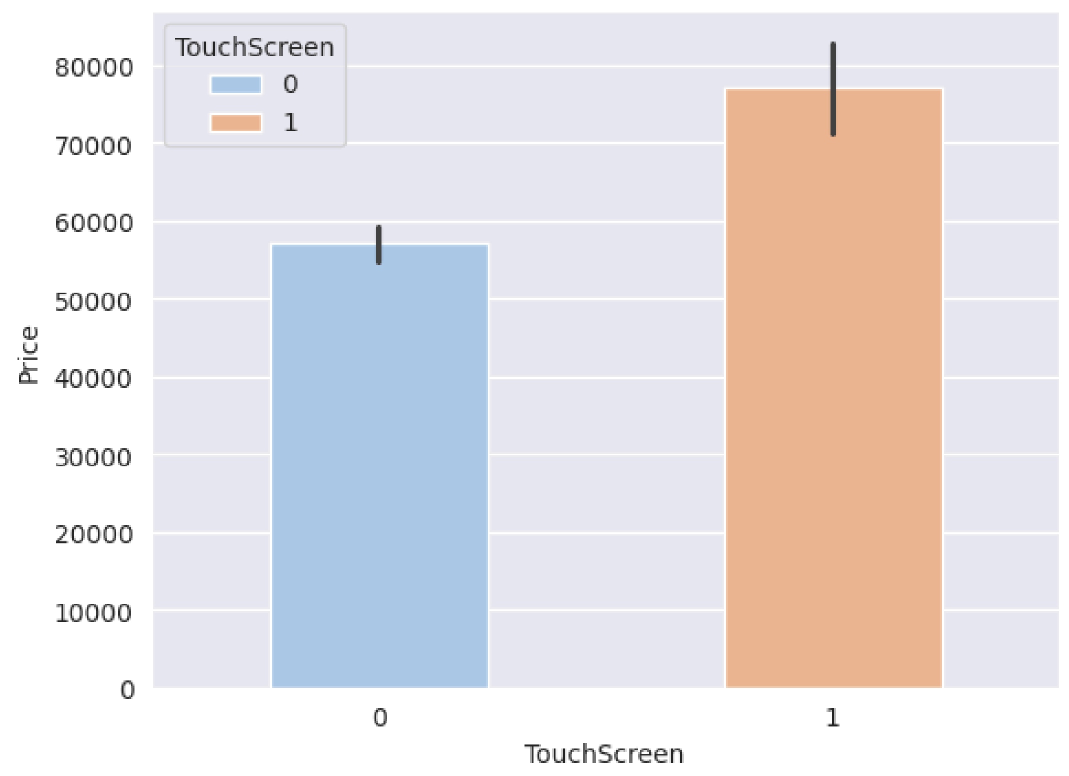
5. Screen Resolution

Screen resolution contains a lot of information and is quite noisy. Before any analysis, we need to perform feature engineering on it. By examining the unique values, we can extract three new columns: IPS panel presence, touchscreen capability, and resolution (X and Y axis).



- Similarly, the IPS column is binary: 1 for IPS panels and 0 for non-IPS. Although IPS laptops are less common in our dataset, they tend to be higher priced.

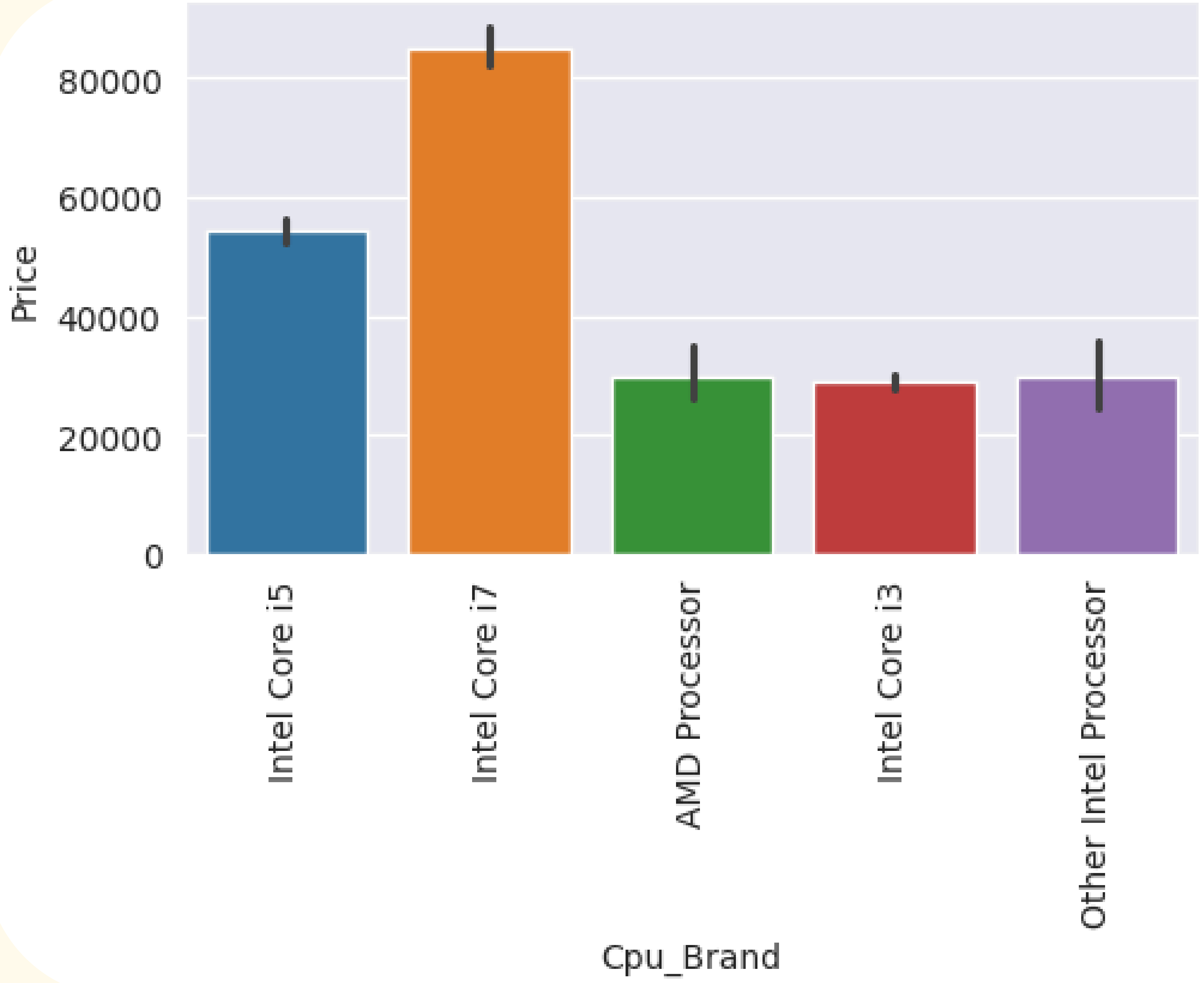


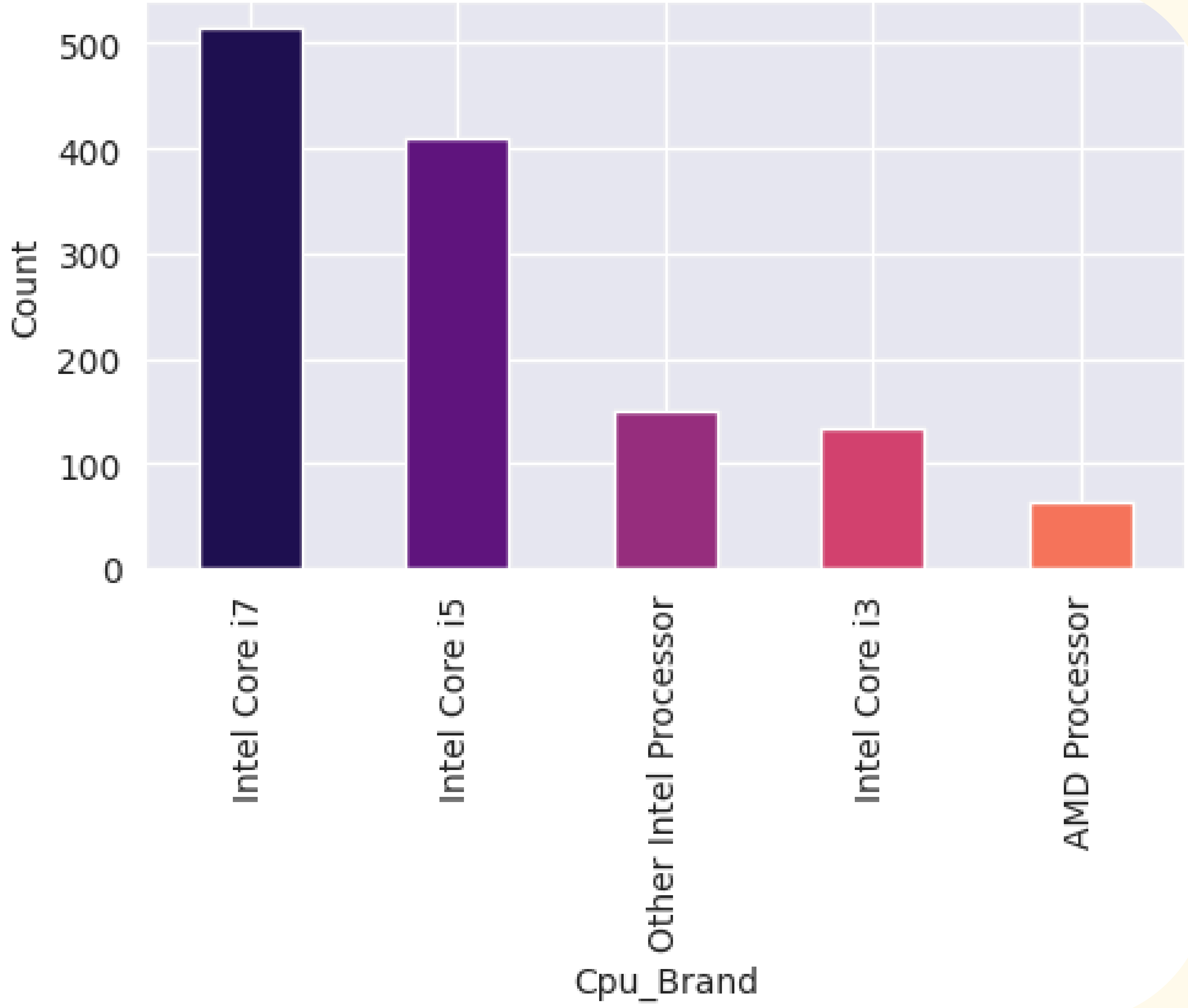


- We encoded the touchscreen column as a binary variable: 1 for touchscreen and 0 for non-touchscreen. The graph shows that touchscreen laptops are generally more expensive, which aligns with real-life observations.
- Using the corr method, we found that laptop size (inches) has a weak correlation with price, while X and Y axis resolution have a strong correlation. To leverage this, we combined these into a single column, pixels per inch (PPI), which has a correlation of 0.47 with price. We then dropped the unnecessary columns: screen resolution, inches, and X and Y resolutions.

6. CPU

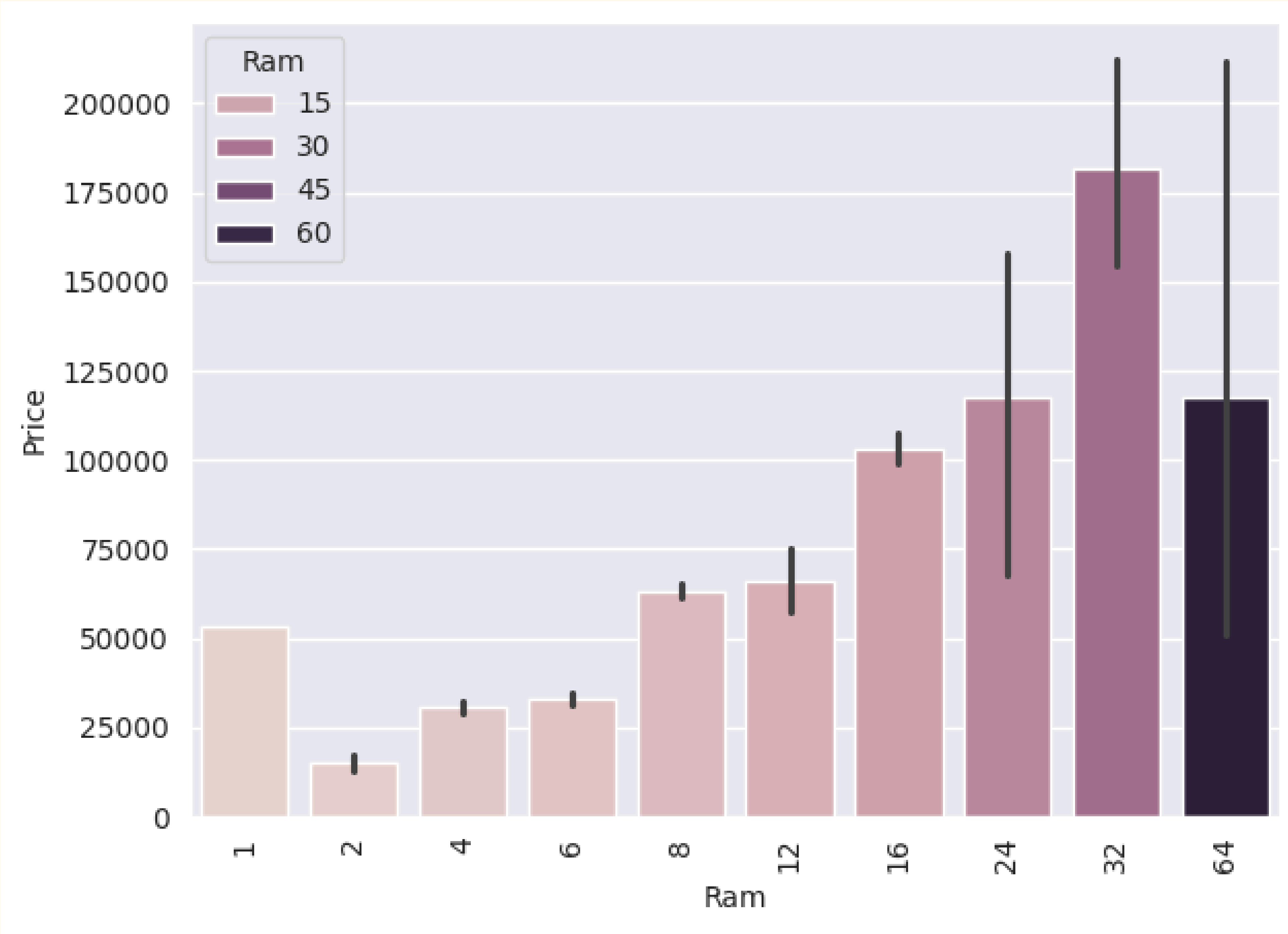
The CPU column has a lot of information, with 118 different categories. To simplify, we extracted the first three words to identify the preprocessor type, focusing on five categories: i3, i5, i7, other Intel processors, and AMD processors. We found that laptops with i7 processors are the most expensive, while i5, i3, AMD, and other processors fall into a similar price range. So, the type of preprocessor significantly influences the laptop's price.

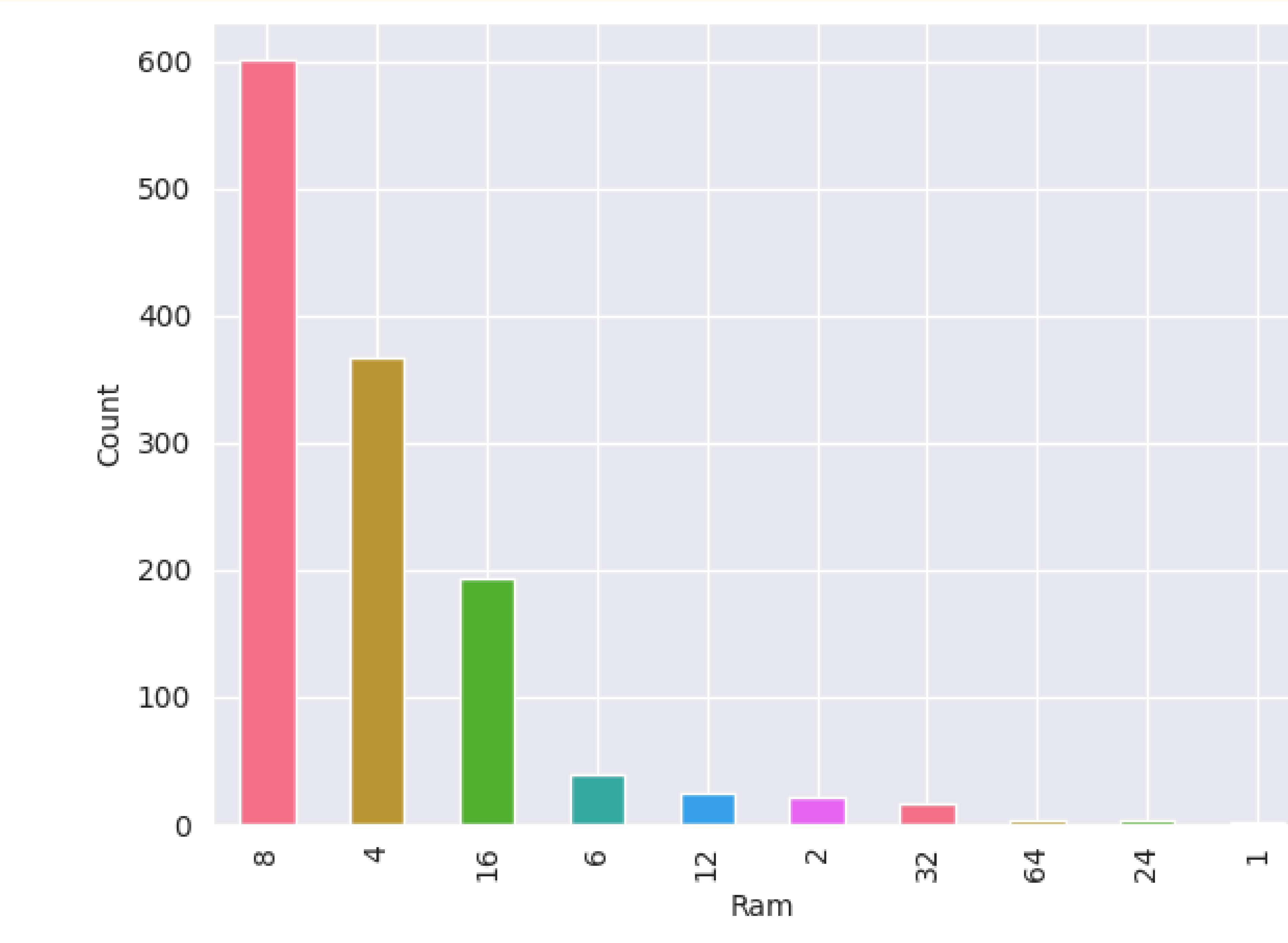




7. RAM

From the graph, we can see that RAM affects the price: as RAM increases, so does the price. The correlation between price and RAM is quite high at 0.68. 8 GB RAM is the most popular choice among buyers.





8. Memory

The Memory column is quite noisy and includes various types such as HDD, SSD, Flash storage, and Hybrid storage. Many laptops come with both HDD and SSD, or have an external slot for upgrades, making this column potentially disruptive for our analysis if not handled properly.

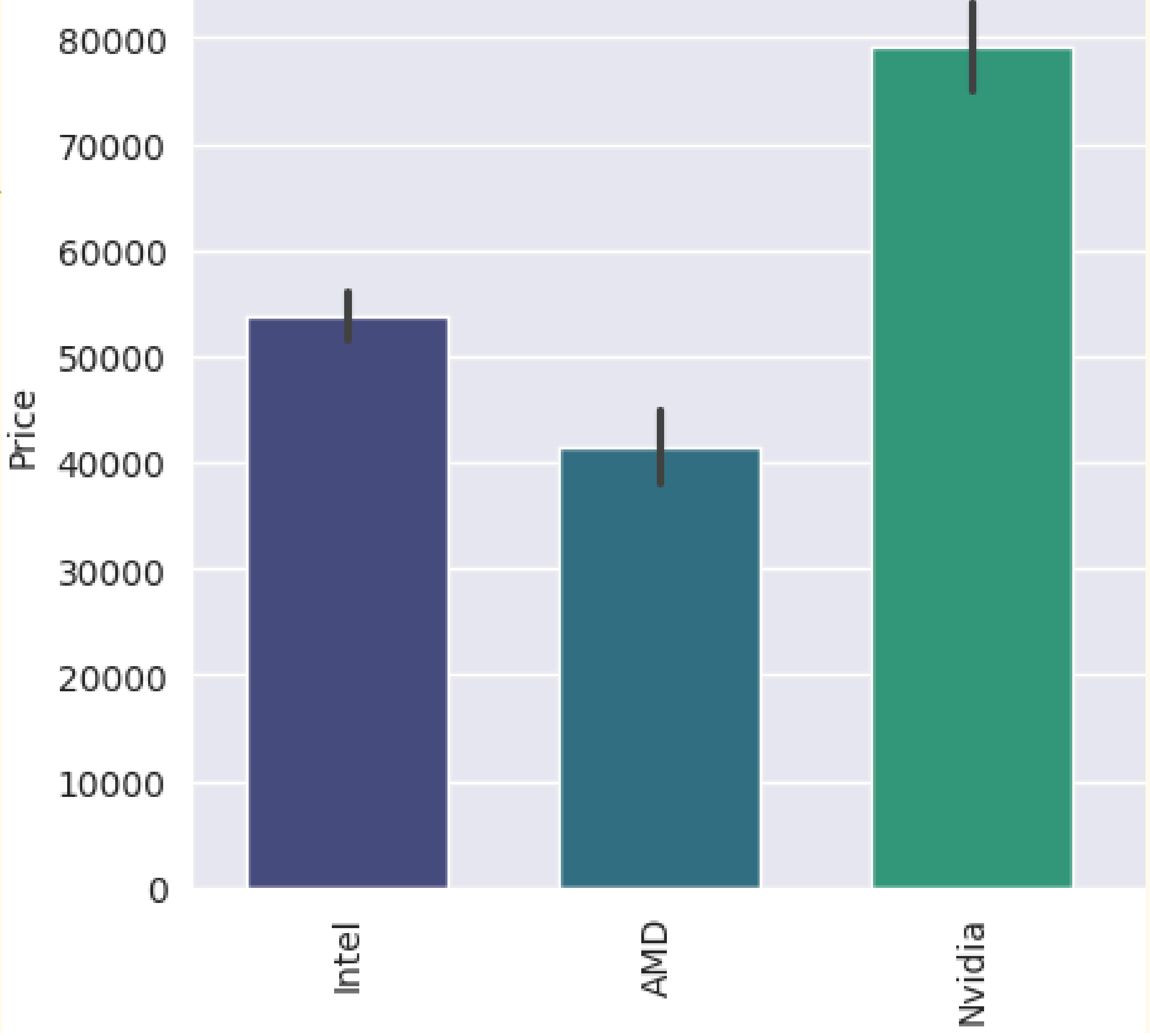
By examining the value counts, we identified four categories: HDD, SSD, Flash storage, and Hybrid. Looking at the correlations with price:

- **SSD: 0.67**
- **HDD: -0.1**
- **Hybrid: 0.01**
- **Flash storage: -0.04**

Since Hybrid and Flash storage show very little to no correlation with price, we decided to drop this column, along with the Cpu and memory columns, which are no longer required.

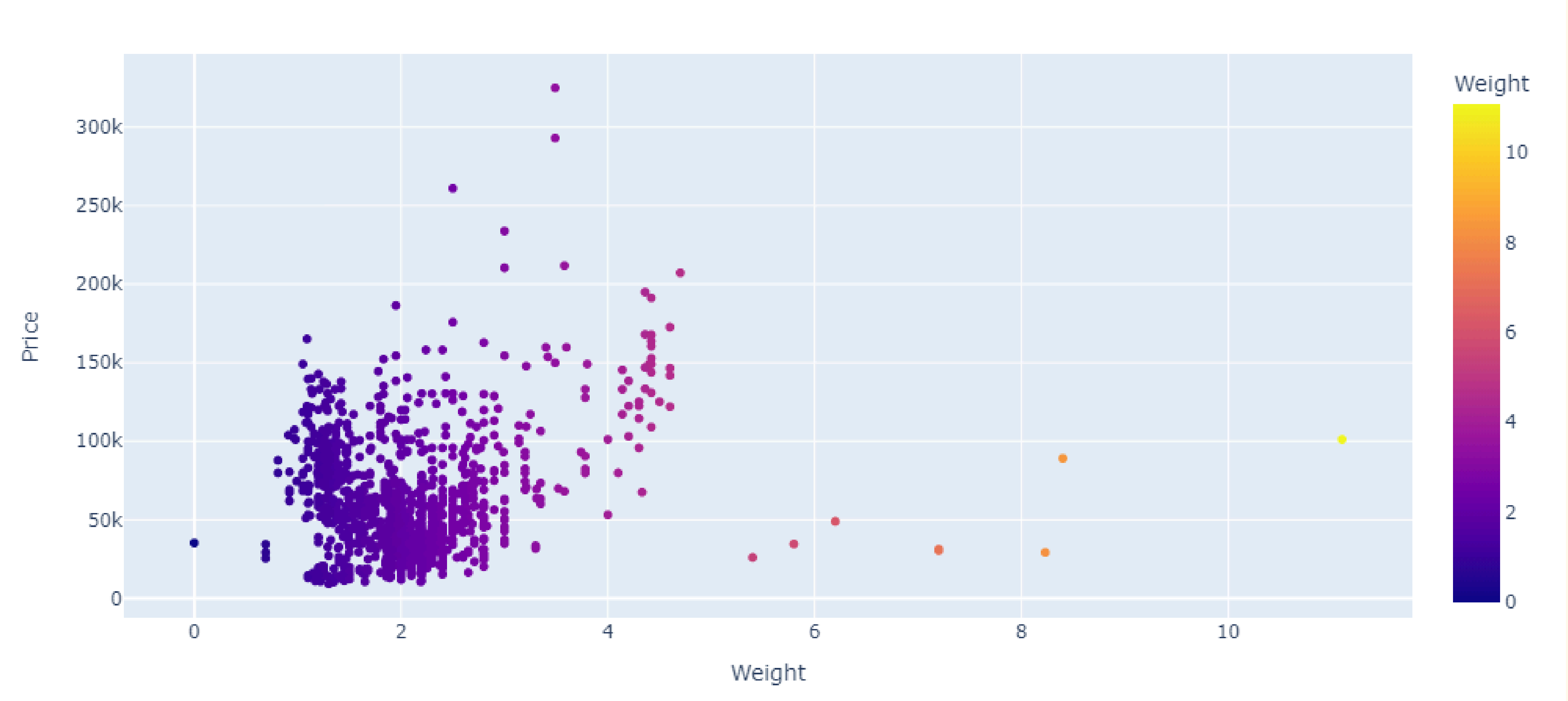
9. GPU

We have categorized GPU(Graphical Processing Unit)s based on their brands: Intel, AMD, and Nvidia. To achieve this, we created a new column called Gpu_brand by extracting brand names from the existing GPU column. We excluded the row that contains ARM GPU, as it's the only exception. Once this transformation completed, we removed the original GPU column from the dataset.

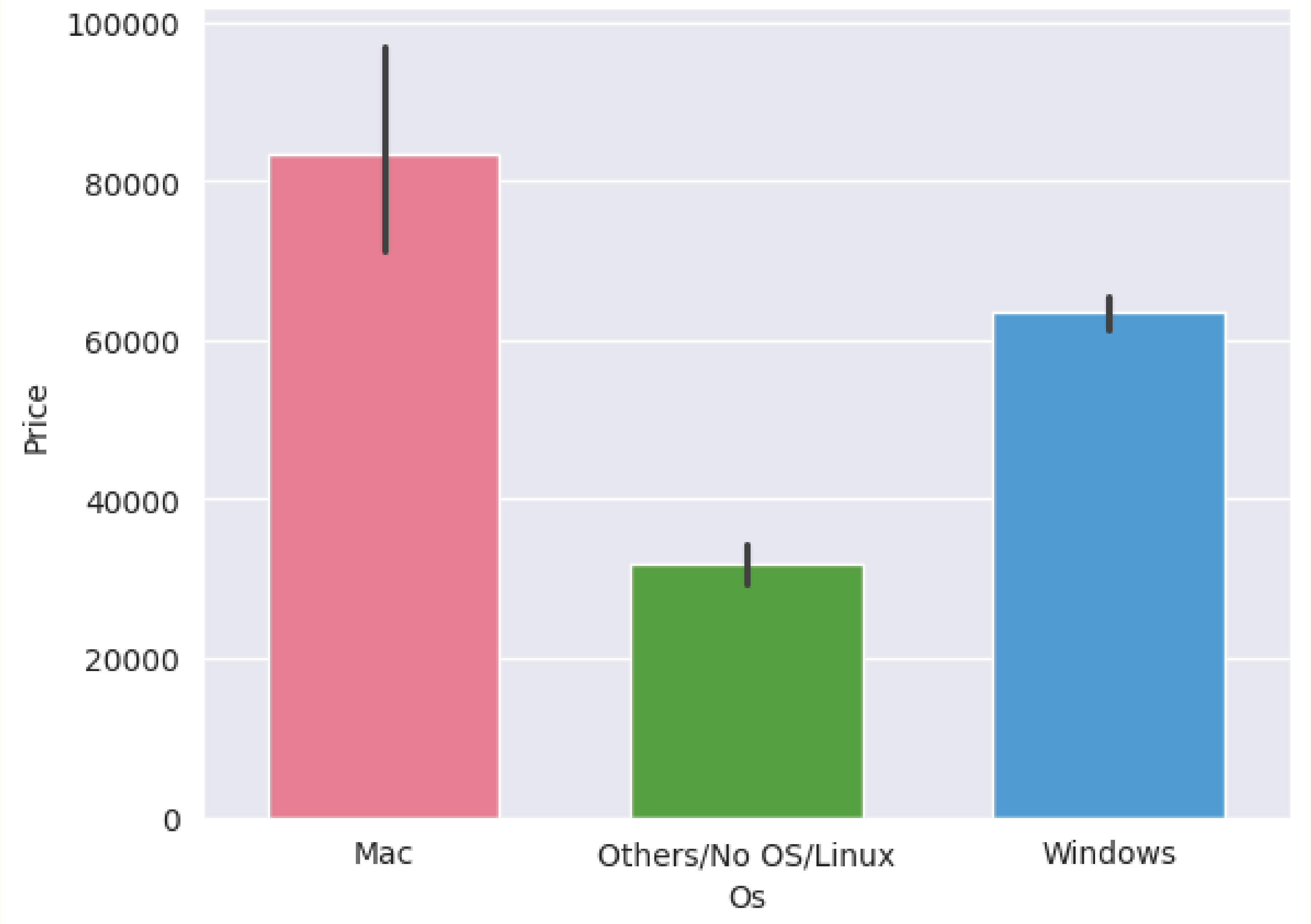


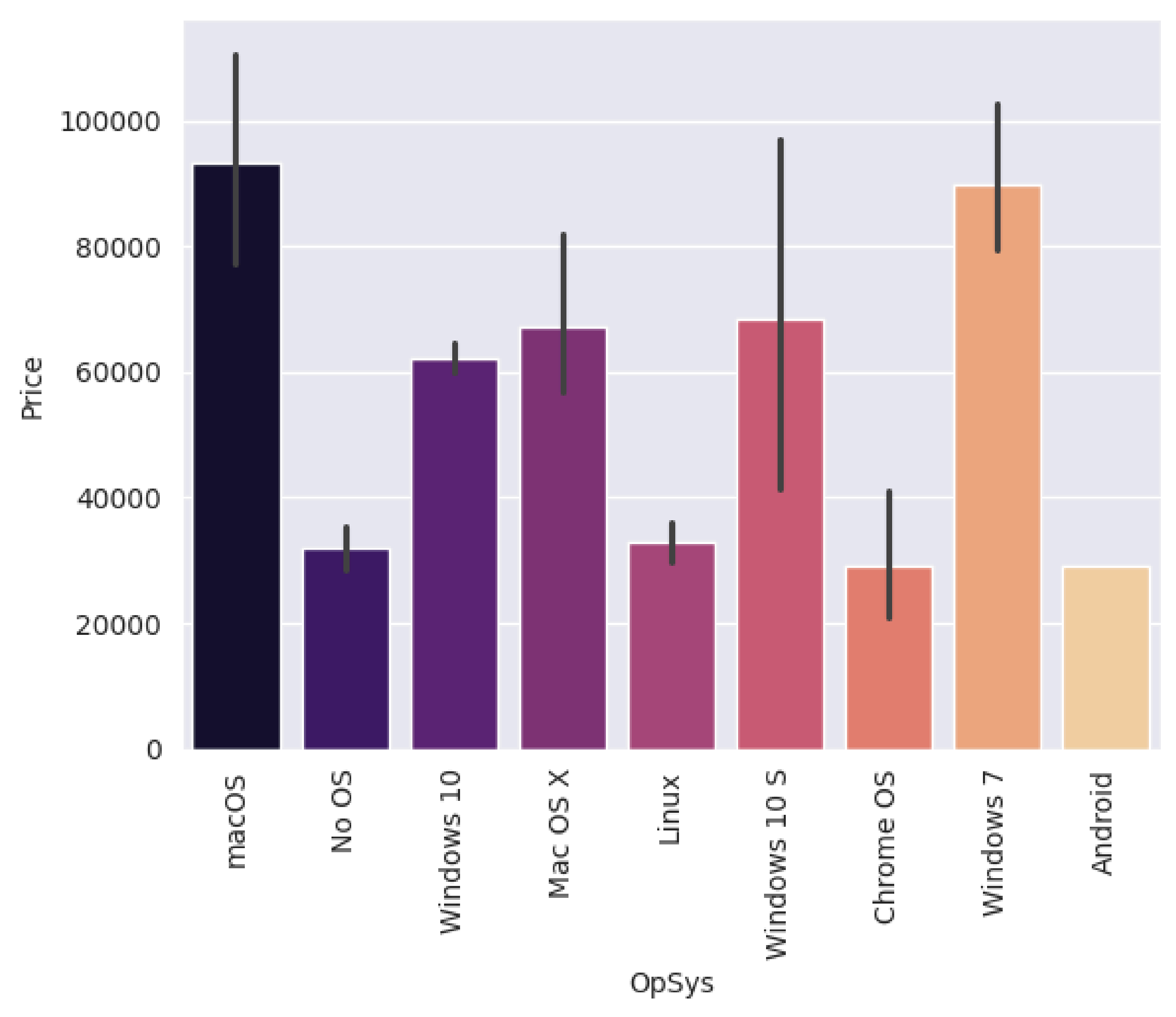
10. Weight

The weight column shows a very weak correlation of 0.18 with the price, indicating a minor influence on pricing of laptop.

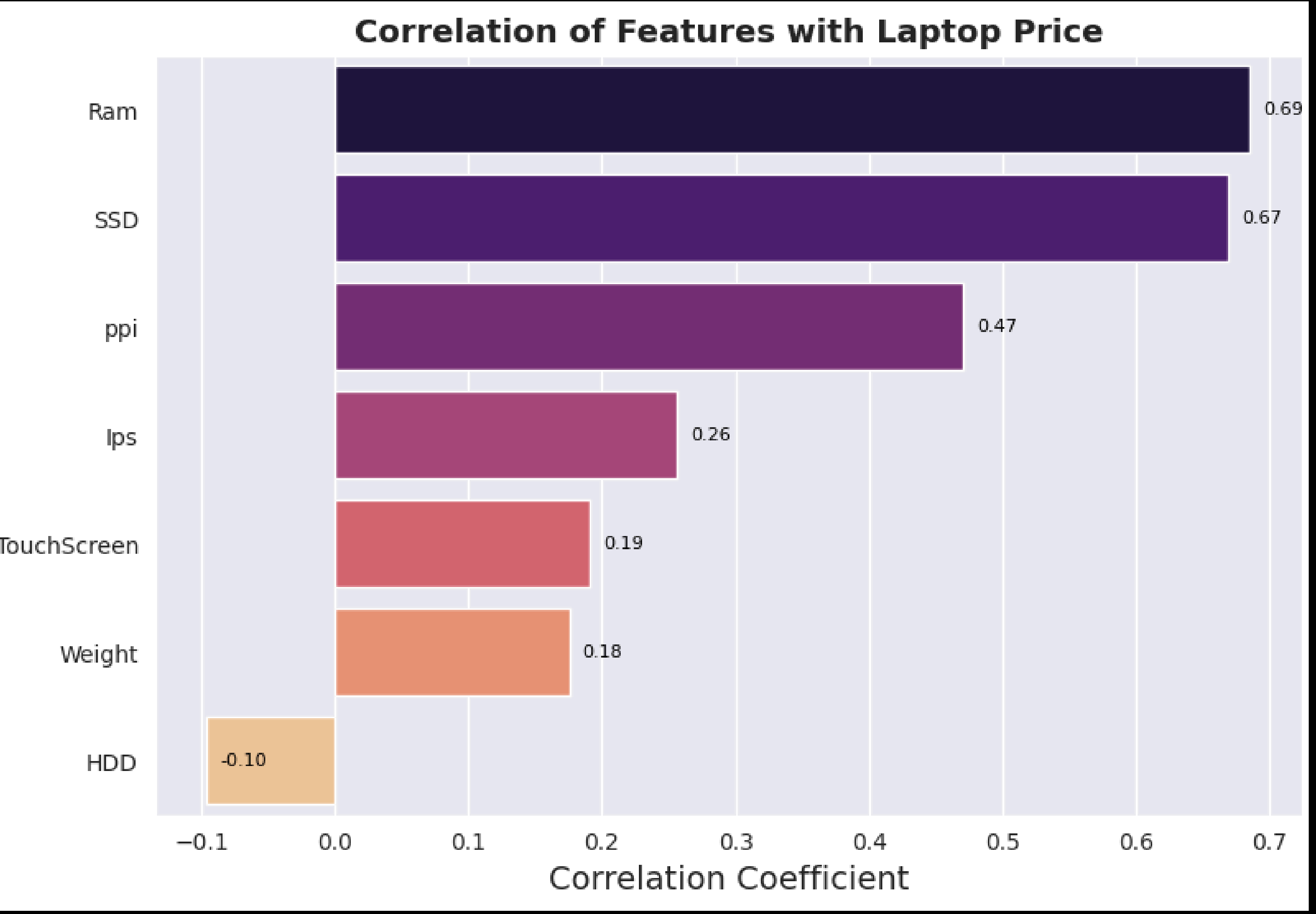


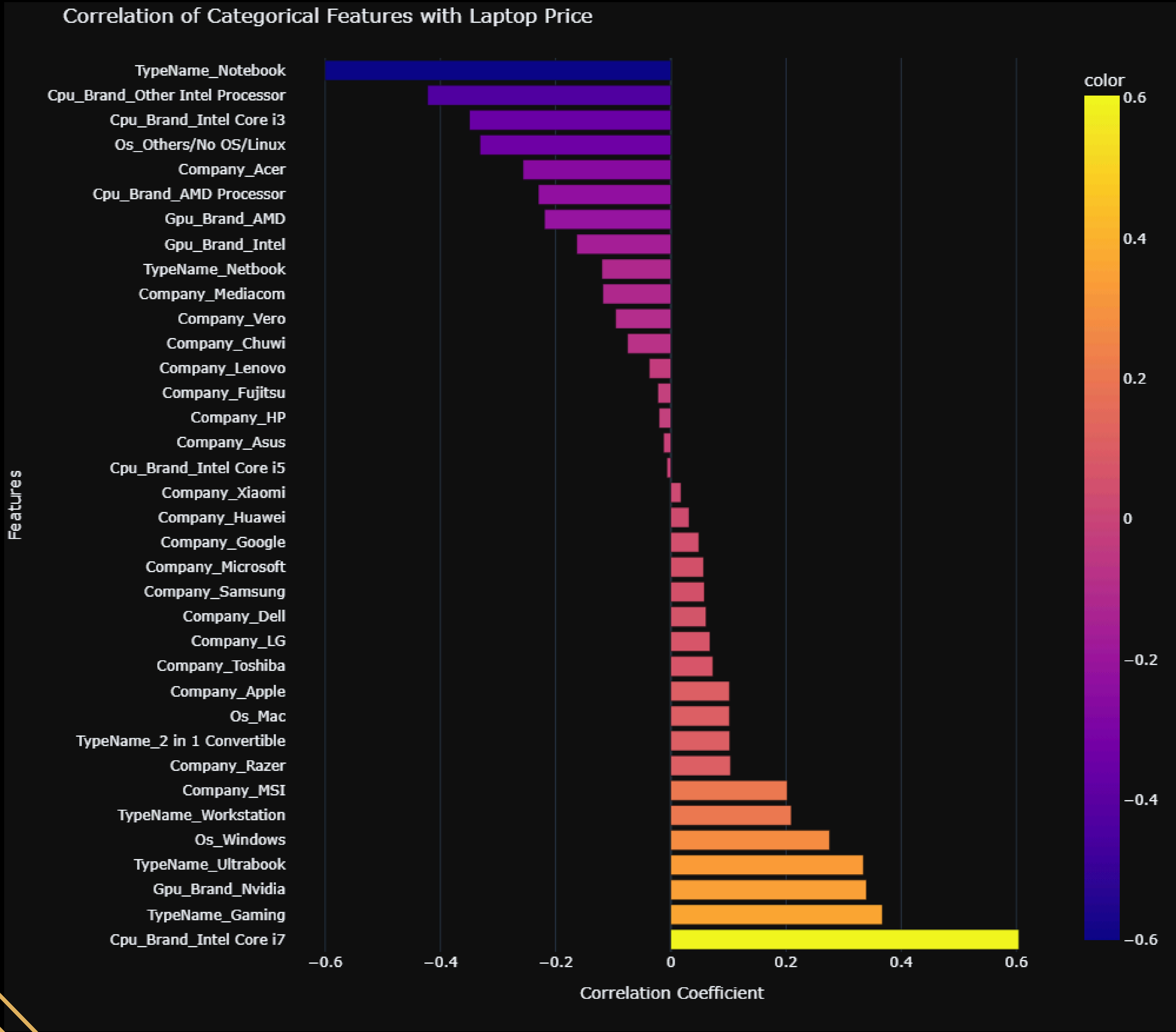
11. Operating System





There are various categories of operating systems. We grouped all Windows categories together, grouped Mac into another category, and classified the remaining operating systems under "Others." According to the graph, Mac remains consistently the most expensive, which is a commonly observed trend.



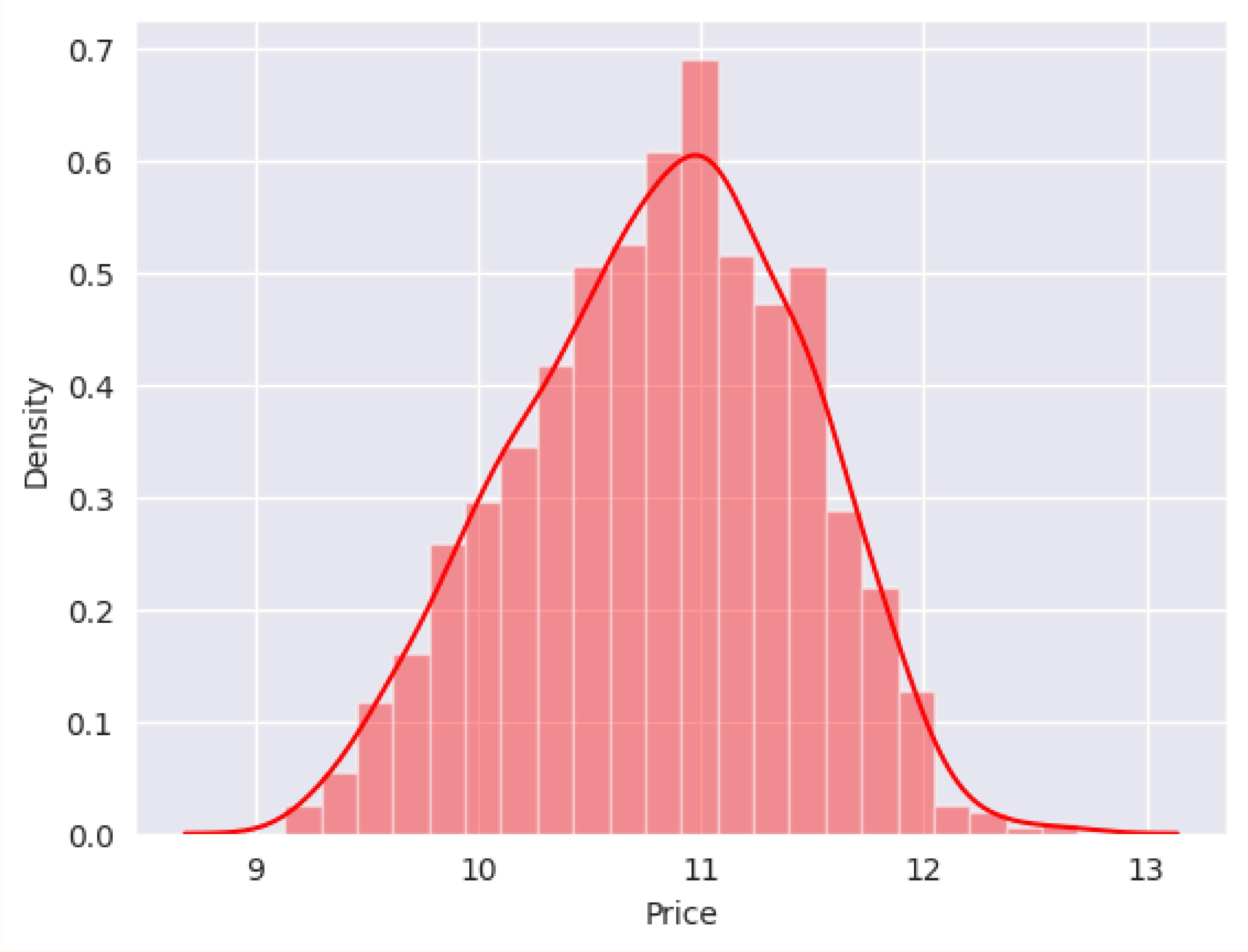


Key Features Identified:

- Company
- Ram
- Cpu Brand
- Gpu Brand
- SSD
- TypeName and ppi

Log-normal distribution

We observed that the target variable's distribution was right-skewed. Transforming it to a normal distribution will improve the algorithm's performance. We apply the log transformation to the price, which normalizes the distribution as shown in the graph. During separation of dependent and independent variables, we take the log of the price and apply the exponent when displaying the results.



Machine Learning Modeling

- We imported the necessary libraries from scikit-learn, a powerful library for machine learning in Python. The primary libraries we used were train_test_split for splitting the data and other scikit-learn modules for building and evaluating our models.

```
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import OneHotEncoder
from sklearn.linear_model import LinearRegression, Ridge
from sklearn.neighbors import KNeighborsRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor, AdaBoostRegressor
from sklearn.svm import SVR
from xgboost import XGBRegressor
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error, mean_absolute_percent
```

- We implemented a pipeline to streamline the training and testing process. First, we used a column transformer to encode categorical variables, which is the first step. After that, we created an object of our algorithm and passed both the steps to the pipeline. Using the pipeline object, we predict the price on new data.

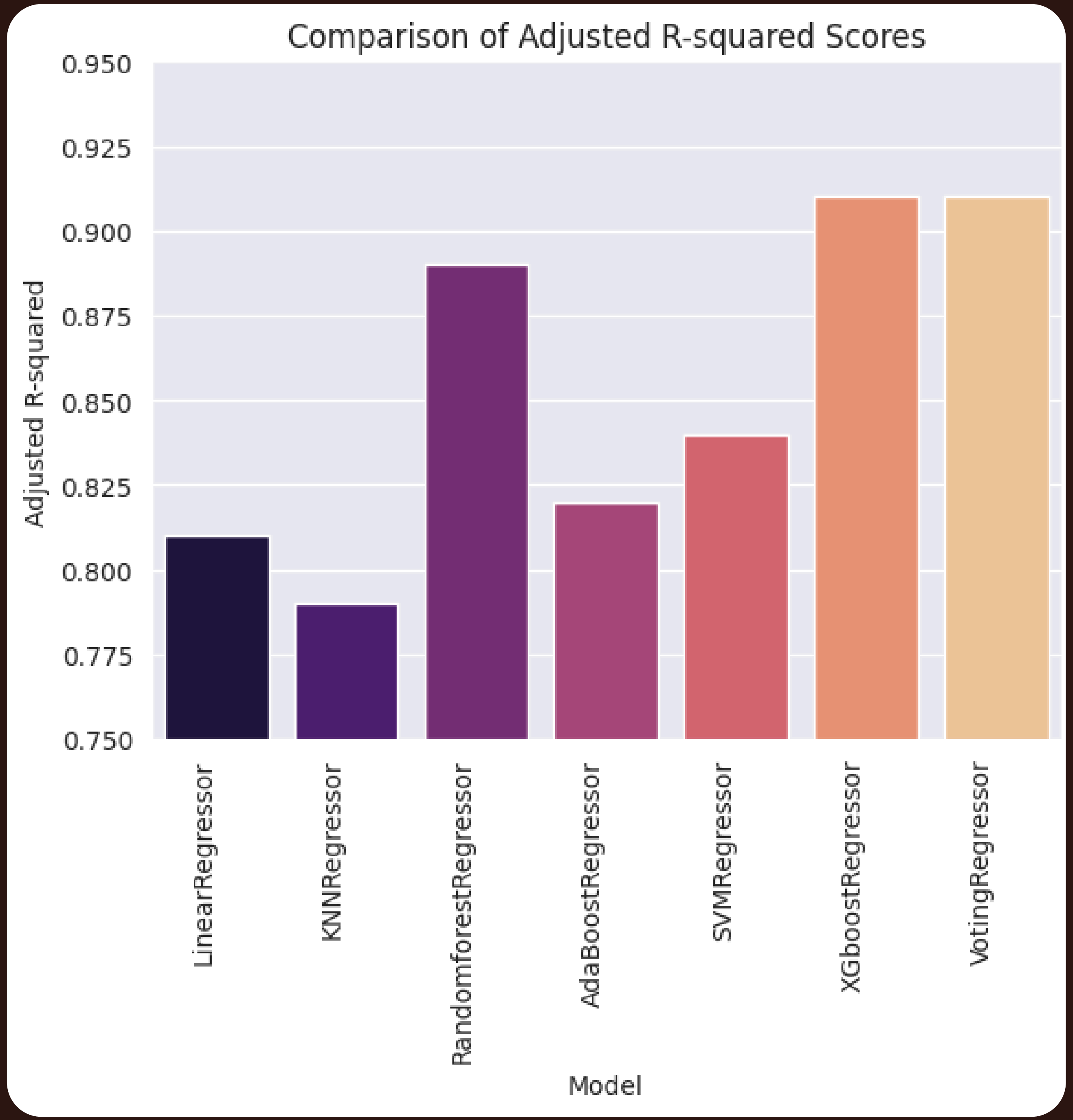
	Company	TypeName	Ram	Weight	TouchScreen	Ips	ppi	Cpu_Brand	HDD	SSD	Gpu_Brand	Os
0	Apple	Ultrabook	8	1.37	0	1	226.983005	Intel Core i5	0	128	Intel	Mac
1	Apple	Ultrabook	8	1.34	0	0	127.677940	Intel Core i5	0	0	Intel	Mac
2	HP	Notebook	8	1.86	0	0	141.211998	Intel Core i5	0	256	Intel	Others/No OS/Linux
3	Apple	Ultrabook	16	1.83	0	1	220.534624	Intel Core i7	0	512	AMD	Mac
4	Apple	Ultrabook	8	1.37	0	1	226.983005	Intel Core i5	0	256	Intel	Mac
...	...	...	...	...	...	...	...	...	...	...	...	...

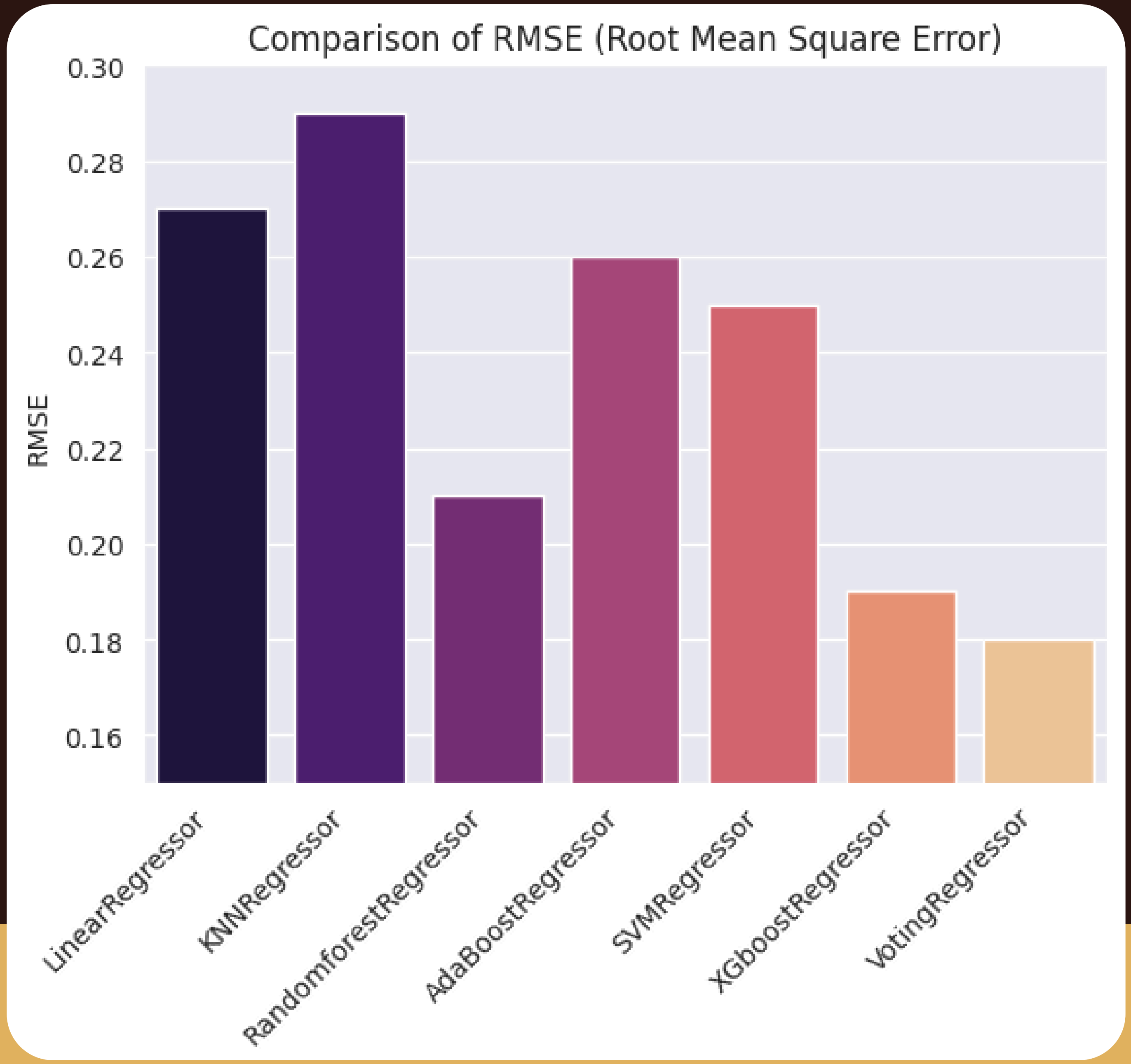
Model Selection

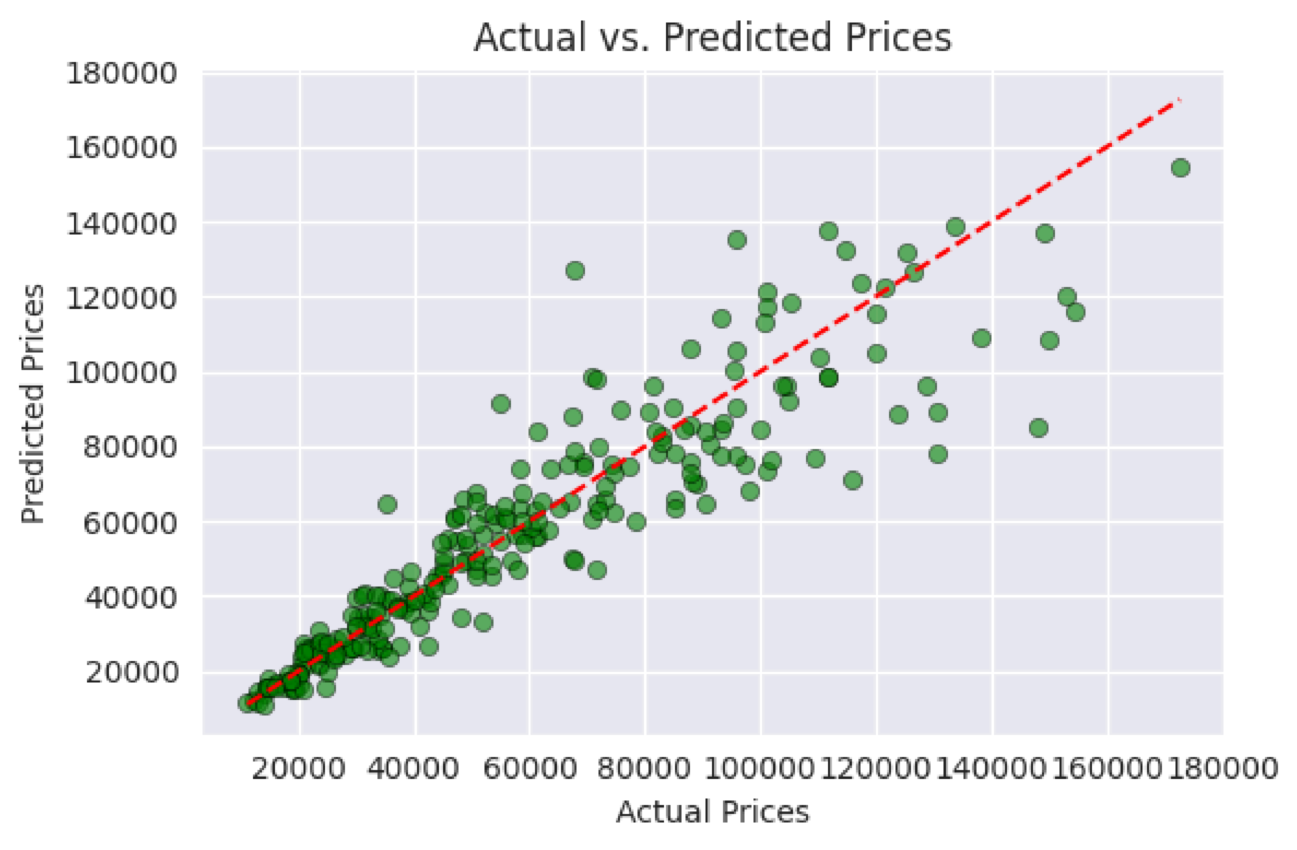
We used several models for our regression analysis, including Linear Regression, KNN Regressor, Random Forest Regressor, SVM Regressor, XGBoost Regressor, and Voting Regressor. We evaluated these models using metrics such as Adjusted R^2 score and RMSE (Root Mean Squared Error).

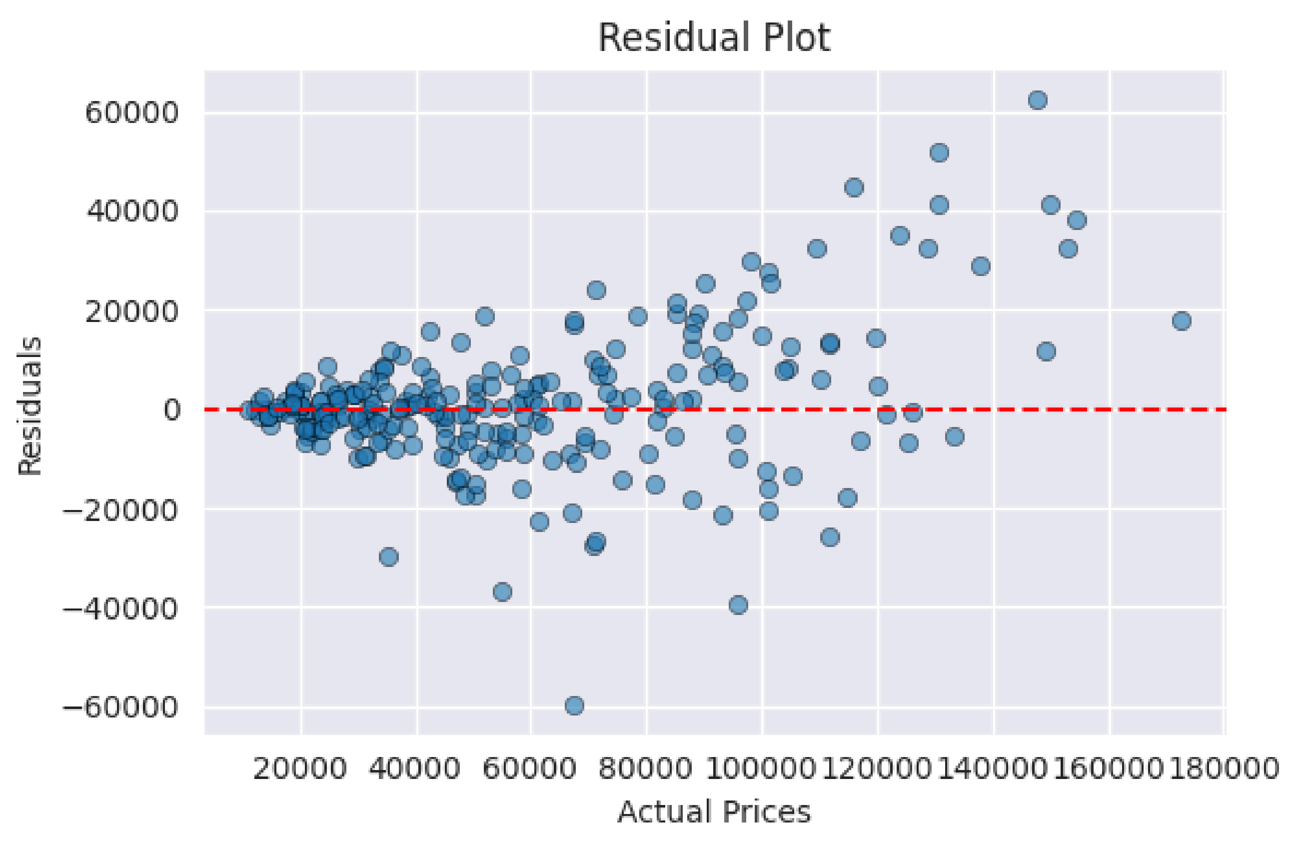
By looking at the graph, we concluded that the Voting Regressor and XGBoost Regressor are the best fit for the model. We chose the Voting Regressor for our final model due to its slightly better performance.

	Model	AdjustedR2_score	RMSE
0	LinearRegressor	0.81	0.27
1	KNNRegressor	0.79	0.29
2	RandomforestRegressor	0.89	0.21
3	AdaBoostRegressor	0.82	0.26
4	SVMRegressor	0.84	0.25
5	XGboostRegressor	0.91	0.19
6	VotingRegressor	0.91	0.18

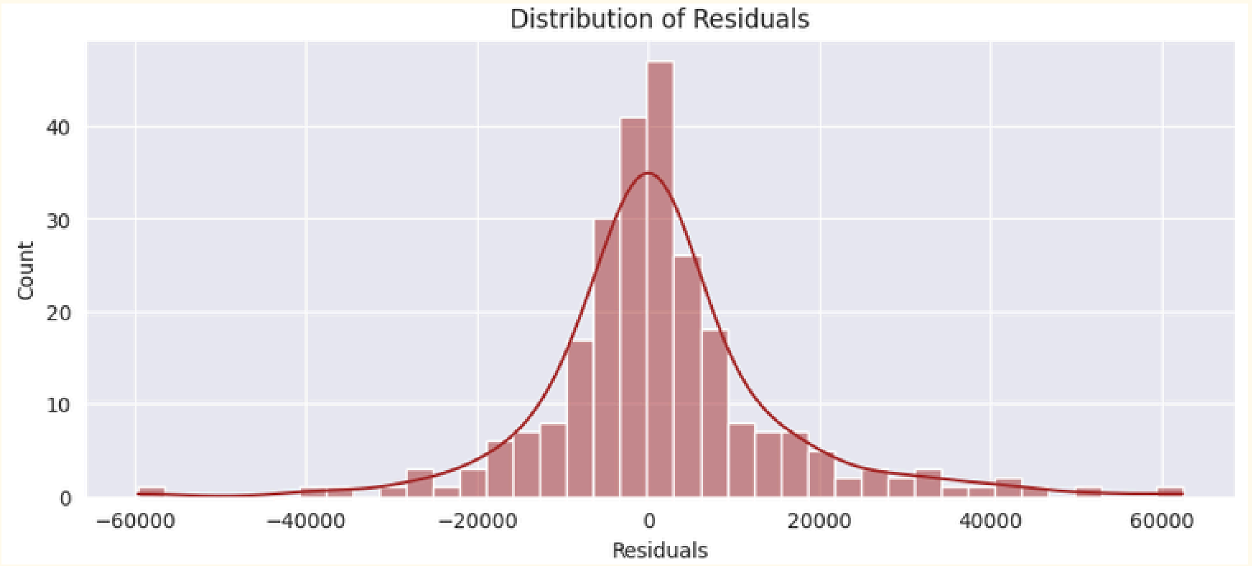








- Strong Performance: The model shows a strong linear relationship between actual and predicted prices.
- Price Deviation: Predictions are less accurate for more expensive laptops. This is because of less amount of data present.
- Normal Residuals: Errors are roughly normally distributed around zero, but with some variability.

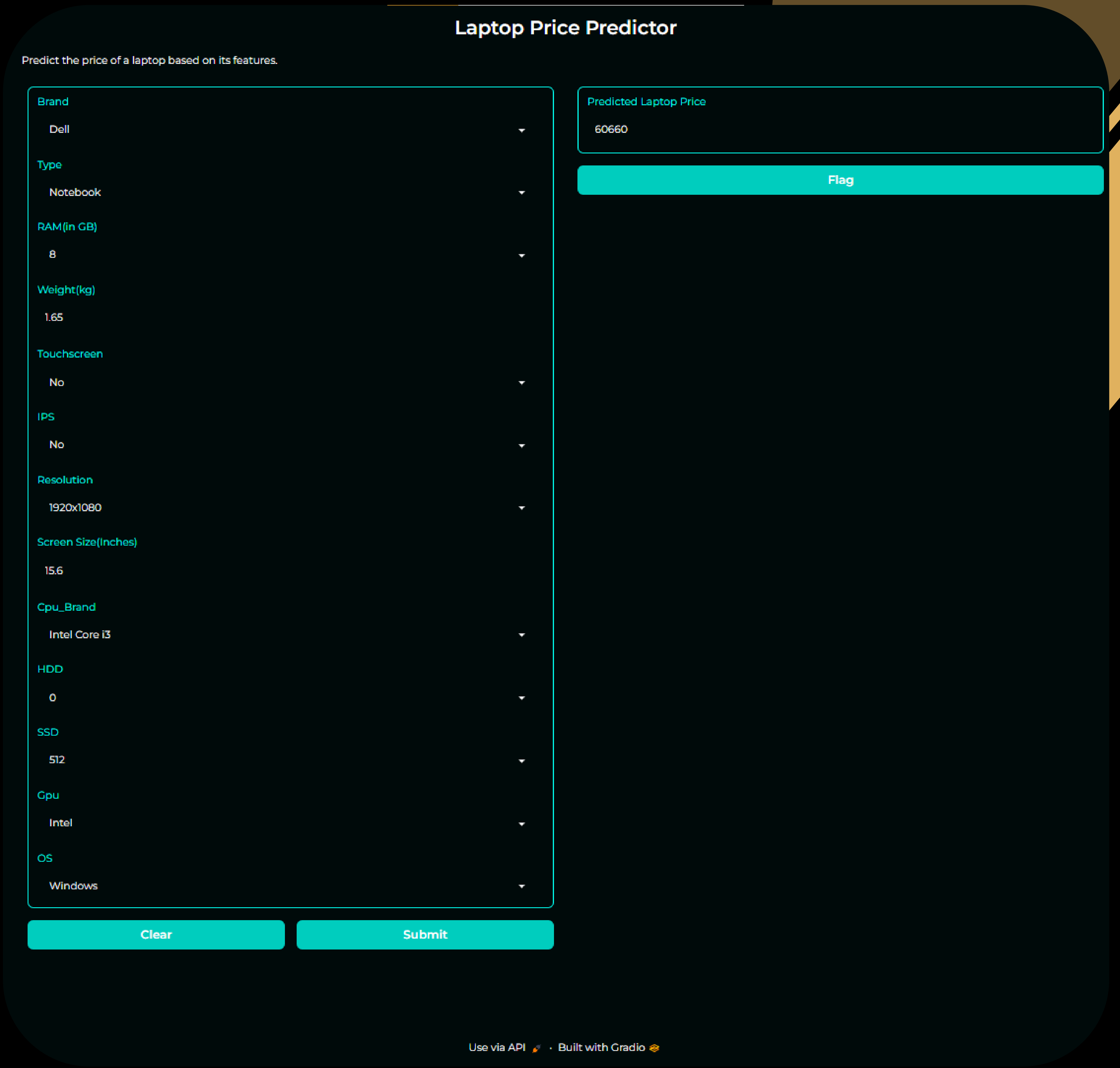


- The predicted price distribution closely matches the actual price distribution, with only minor deviations, suggesting a generally good fit.

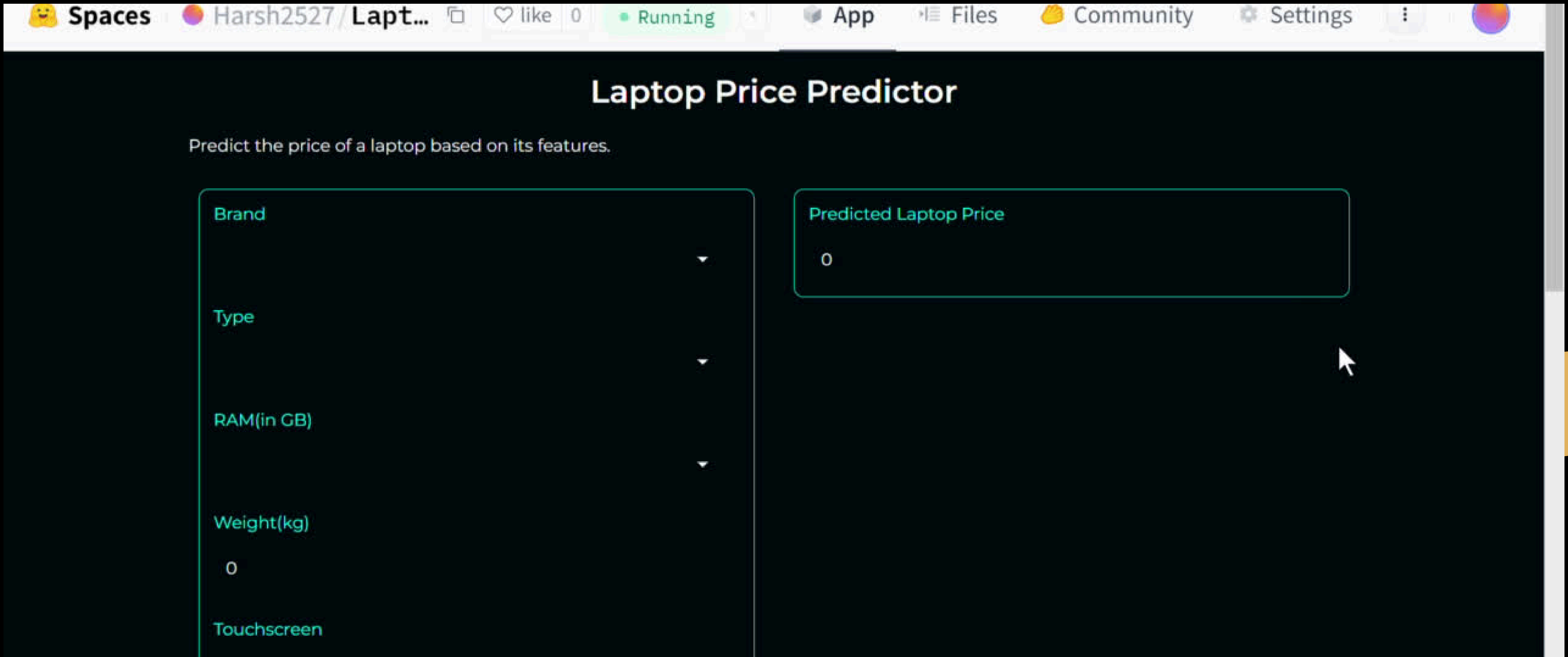
Model Deployment

- We used Gradio to develop a web app for predicting laptop prices. The app will feature a form that collects user inputs corresponding to the dataset features. Using the saved model, we will predict the output and display it to the user.
- Gradio, an open-source Python package, enables rapid development of demos and web applications for machine learning models
- Deployed web app on Hugging face.

[LaptopPricePredictor Link](https://huggingface.co/spaces/Harsh2527/LaptopPricePredictoModel)



Real Time Prediction



Key Insights

Brand Influence

- Certain brands (e.g., Apple, Razer, LG, Google, Microsoft, MSI) have higher average prices
- Mid-range brands (e.g., HP, Lenovo, Dell, Asus, Acer) are competitively priced
- The brand significantly influences the price

Features Impact

- RAM size significantly affects the price(Critical for performance)
- PPI(pixels per inch) - Larger, higher-resolution screens display drive up prices.
- Storage type (SSD) affects the price (Critical for performance)
- Processor brand and type influence price
- Graphical Processing Unit(GPU) affects the price of laptop

Market Positioning

- High-End Market: Premium brands like Apple , Razer and MSI for consumers prioritizing performance and design.
- Mid-Range Market: Brands like HP, Dell, Acer and Lenovo, offering good performance at a reasonable price.
- Budget Market: Lesser-known and budget models, targeting students and cost-sensitive users.
- Niche Segments: Specialized laptops for gaming and professional use, for users seeking specific features.

Conclusion

Project Outcome:

- Developed a reliable model for laptop price prediction
- Gained insights into factors influencing laptop prices

Recommendations for (Smart Tech Co.) :

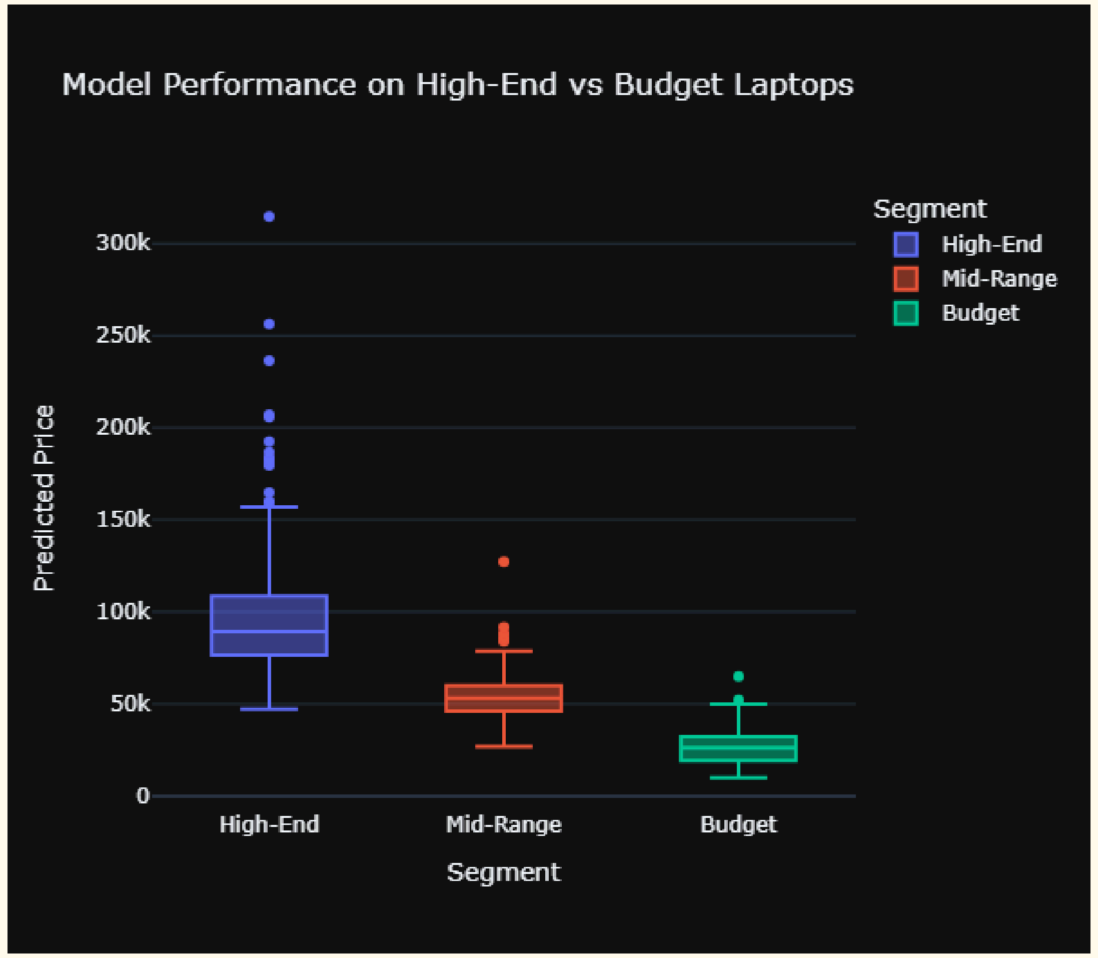
- Utilize the model for pricing new laptops products
- Consider brand positioning and key features of laptops in marketing strategies
- Regularly update the model with new data for improved accuracy

Questions & Answers

- **Can the model accurately predict the prices of laptops from lesser-known brands?**
  - Generally accurate for well-known brands
  - Lesser-known brands may have less consistent predictions due to fewer data points
  - Model performance improves with more data from lesser-known brands

- **Does the brand of the laptop significantly influence its price?**
  - The brand significantly influences the price
  - High-end brands (e.g., Apple,Razer,Google,Microsoft,&MSI) command higher prices
  - Lesser-known brands often have lower prices but competitive specifications

- **How well does the model perform on laptops with high-end specifications compared to budget laptops?**
  - The model demonstrates good performance across both segments, with higher variability and potential outliers in high-end laptop predictions and consistent, reliable performance in budget laptop predictions.



Questions & Answers

How does the model perform when predicting the prices of newly released laptops not present in the training dataset?

- Predictions for newly released laptops can be less accurate
- Requires regular updates with new data to maintain accuracy
- Incorporating recent trends and technology changes is crucial for model relevance

What are the limitations and challenges in predicting laptop prices accurately?

Challenges:

- Limited data from lesser-known brands affects prediction accuracy
- Rapid changes in technology and market trends require frequent model updates
- Price anomalies and outliers can skew predictions

Limitations:

- The model might not account for sudden market shifts
- Performance on newly released laptops may vary