DSCI 310 Group Project: Laptop Price Predictor Model

Anna Czarnocka, An Zhou, Yuechang Liu, Daniel Lima

Table of contents

Summary

Our project aims to answer the question "How can we predict determinants of the laptop market price?". We used publically available Laptop Dataset(2024). We performed a robust data analysis in Python, spanning from importing data to sharing insights, prioritizing the creation of workflows that are both replicable and reliable. We used the Linear regression method to construct our predictive model. Our results are that the price can be predicted by building a and training a model using variables such present in the data such as: Rating, num_cores, num_threads, ram_memory, primary_storage_capacity, secondary_storage_capacity, is_touch_screen, display_size, resolution_width and resolution_height.

Introduction

In this digitalized world today, laptops are one of the most demanding digital products. According to Grand View Research, the global laptop market was valued at $194.25 billion in 2022 and is expected to grow in the foreseeable future (Afzal, 2023). This market amount is created by laptops that vary in price on a significant range from less than two hundred to a few thousand dollars. However, the prices of laptops are surely not unpredictable. Here in this project, we answer the question: how can we predict the laptop market price by the appropriate determinants?

This question is important because it helps customers to understand the factors behind the pricing of laptops which helps them to make reasonable decisions while choosing a laptop. Also, the result of this research benefits laptop producers and sellers in price-making strategies. This research lets laptop producers have a better picture of laptops with what types of features should be priced higher on the market. We try to approach this question by fitting a KNN regression model on the dataset "laptop 2024" (Kumar, 2024). The dataset that our research is based on is a public dataset on Kaggle uploaded by Aniket Kumar. It collects data from 991 unique laptops with 22 features. All information is updated to January 14, 2024.

Methods & Results

Analysis

Research Project Methodology: Predicting Laptop Prices through Feature Analysis

Objective

The aim of this project is to develop a predictive model that can estimate the price of laptops based on various product features. Accurate estimation of the price is crucial for all laptop users, both for the professionals and amators, that are planning to buy new laptops, as well as for the sellers and the laptop mrket industry, as it helps identify potential importance of each of the feature and determine appropriate buying strategies. This model aims to serve various stakeholders in the laptop market, including potential buyers seeking to make informed purchasing decisions, sellers aiming to strategize their pricing, and industry analysts interested in understanding the impact of different laptop features on their market value. The research specifically seeks to identify the determinants of laptop prices, providing insights into which attributes significantly influence cost in the competitive laptop market.

The dataset provided for this project consists of a large number of observations from both a training sample and a test sample. Each observation includes information such as the laptop's brand, model,

price, rating, processor details, number of cores and threads, ram memory, primary storage type, capcity and many others.

Dataset Overview

The core of this research is based on a meticulously curated dataset titled "Laptop Dataset (2024)" downloaded from Kaggle which encompasses a rich collection of 991 unique laptop entries extracted from the Smartprix website. This dataset has been carefully cleaned and updated as of January 14, 2024, ensuring its reliability for in-depth analysis. It features 22 distinct attributes for each laptop, including but not limited to:

- **Brand and Model**: Identifying the manufacturer and specific model of the laptop.
- **Price**: Listed in Indian Rupees, providing a direct measure of market value.
- **Processor Specifications**: Including brand, tier, number of cores, and threads.
- **Memory and Storage**: Details on RAM, primary and secondary storage types and capacities.
- **GPU Details**: Information on the brand and type of graphics processing unit.
- **Display Characteristics**: Screen size, resolution, and touch screen functionality.
- **Operating System**: The installed OS.
- **Warranty**: The duration of the manufacturer's warranty.

Methodology

To achieve the project's goal, the methodology will encompass several key stages:

1. Initial steps of data preprocessing will include cleaning the data for inconsistencies, handling missing values, and encoding categorical variables to prepare the dataset for modeling.
2. The stage of explanatory data analysis (EDA) involves examining the dataset to understand the distribution of key features, identify outliers, and uncover potential relationships between variables.
3. Based on insights from EDA, new features may be engineered to better capture the influence of certain attributes on laptop prices. This could include interaction terms or derived features like performance-to-price ratios.
4. A variety of machine learning models, including linear regression, decision trees, and ensemble methods like random forest and gradient boosting, will be evaluated to determine the most effective approach for price prediction. Model selection will be based on cross-validation performance metrics such as mean squared error (MSE).
5. The selected model will be rigorously tested using a hold-out test sample to assess its generalization ability and accuracy in predicting laptop prices.
6. Once the model is finalized, an analysis of feature importance will be conducted to identify which laptop characteristics are most predictive of price. This will address the research question by highlighting the key determinants of laptop pricing.

Expected Outcomes

The culmination of this research project is anticipated to yield a robust model that can predict laptop prices with high accuracy, offering valuable insights into the factors that most significantly impact laptop market values. Through this analysis, stakeholders in the laptop industry will be better equipped to understand pricing dynamics, facilitating more informed decision-making processes for both consumers and sellers. Additionally, the project aims to contribute to the academic and practical understanding of price determination in technology markets, potentially guiding future research and development strategies within the laptop industry.

Explanatory data analysis

The dataset, titled "Laptop Dataset (2024)," encompasses a meticulously curated collection of 991 unique laptop entries, sourced from the Smartprix website. It has been updated as of January 14, 2024, and provides a comprehensive overview of various laptop features, making it an invaluable resource for developing price prediction models and recommendation systems. This dataset includes a wide array of attributes for each laptop, offering insights into the intricate dynamics of laptop pricing and consumer preferences. The features captured in the dataset are as follows:

- `Brand`: The name of the laptop brand.
- `Model`: The specific model or series of the laptop.
- `Price`: The price of the laptop in Indian rupees.
- `Rating`: The rating assigned to each laptop based on its specifications.
- `Processor_brand`: The brand of the processor used in the laptop.
- `Processor_tier`: The performance tier or category of the processor.
- `Number_of_Cores`: The number of processing cores in the processor.
- `Number_of_Threads`: The number of threads supported by the processor.
- `Ram_memory`: The amount of RAM used in the laptop.
- `Primary_storage_type`: The type of primary storage (e.g., HDD, SSD).
- `Primary_storage_capacity`: The capacity of the primary storage in the laptop.
- `Secondary_storage_type`: The type of secondary storage, if available.
- `Secondary_storage_capacity`: The capacity of the secondary storage in the laptop.
- `GPU_brand`: The brand of the graphics processing unit (GPU).
- `GPU_type`: The type of the GPU.
- `Is_Touch_screen`: Indicates whether the laptop has a touch screen feature.
- `Display_size`: The size of the laptop display in inches.
- `Resolution_width`: The width resolution of the display.
- `Resolution_height`: The height resolution of the display.
- `OS`: The operating system installed on the laptop.
- `Year_of_warranty`: The duration of the warranty provided for the laptop, usually in years.

This dataset serves as a solid foundation for exploring laptop pricing dynamics and consumer preferences, equipped with a rich set of features for in-depth analysis. It consists of a total of 991 observations, each detailing a unique laptop configuration to assist in the development of accurate price prediction models and effective recommendation systems.

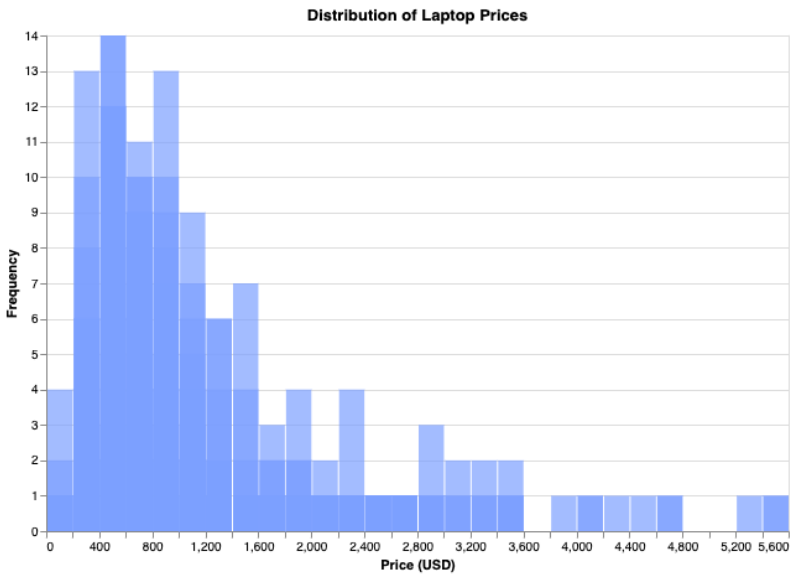

Figure 1: Distribution of laptop prices.

The key takeaway from Figure 1 analyzing the distribution of laptop prices from is that there is a clear skew towards lower-priced models. The majority of laptops are clustered in the lower price range, with a significant drop in frequency as prices increase. This suggests that more affordable laptops are far more common, or at least more commonly listed on the Smartprix website from which this data was sourced.

The prevalence of lower-priced laptops could indicate a stronger market demand in this price segment, possibly reflecting the purchasing power of the consumer base or the competitive pricing strategies of manufacturers. For predictive modeling, this skewness towards lower prices could influence the model's accuracy, potentially leading it to be more reliable at predicting prices for lower-priced laptops than for higher-priced ones.

Additionally, the tapering off of frequency at higher price points may point to a smaller niche market for premium laptops. For businesses and retailers, this distribution could inform inventory and marketing strategies, emphasizing the broad appeal of budget-friendly options. In the context of price prediction, care must be taken to ensure that the model does not undervalue the unique features and qualities that might justify the higher prices of less common, premium models. …

Table 1: Descriptive statistics of laptop prices in USD.

	Statistic	Value
0	count	991.000000
1	mean	77266.504541
2	std	57384.910269
3	min	9800.000000
4	25%	43595.000000
5	50%	61900.000000
6	75%	89245.000000
7	max	454490.000000

The output of the descriptive analysis on the `Price` column provides valuable insights into the distribution and characteristics of laptop prices in USD in the dataset:

Table 2: Descriptive statistics of laptop specifications and prices in USD.

- **Count**: The count indicates the number of non-missing values in the `Price` column, which is 991. This suggests that the dataset is complete with no missing values for the price in USD.
- **Mean**: The mean value of 926.45 USD represents the average price of laptops. This provides an estimate of the central tendency of the price distribution, indicating that, on average, laptops in the dataset are priced around this value.
- **Standard Deviation**: The standard deviation of 688.62 USD quantifies the spread or dispersion of the laptop prices around the mean. A larger standard deviation suggests greater variability in the prices, indicating a wide range of prices within the dataset.
- **Minimum and Maximum**: The minimum price value of 117 USD represents the least expensive laptop in the dataset, while the maximum price value of 5,453 USD represents the most expensive laptop. These values highlight the extent of the price range covered in the dataset.
- **Quartiles**: The 25th percentile (Q1) of 523 USD and the 75th percentile (Q3) of 1,070.50 USD provide additional reference points for understanding the distribution of prices. Specifically, 25% of laptops are priced at or below 523 USD, and 75% are priced at or below 1,070.50 USD. The median (50th percentile) price of 742 USD indicates that half of the laptops are priced below this amount.

This descriptive analysis helps us understand the pricing structure within the dataset, which can be instrumental for both consumers looking to purchase laptops within certain price ranges and sellers aiming to price their laptops competitively in the market.

2. Investigate Data Types

The key takeaway from the data types analysis of the `laptops.csv` dataset is the presence of a diverse range of data types, which suggests a mixture of numerical, categorical, and boolean data within the features. Specifically, columns such as `brand`, `Model`, `processor_brand`, `processor_tier`, `gpu_brand`, `gpu_type`, `OS`, and `year_of_warranty` are of object type, likely indicating categorical data.

Numerical columns like `Rating`, `num_cores`, `num_threads`, `ram_memory`, `primary_storage_capacity`, `secondary_storage_capacity`, `display_size`, `resolution_width`, `resolution_height`, and the target variable `Price` are either of type int64 or float64, representing quantitative data that can be used directly in mathematical computations and statistical analyses.

The boolean column `is_touch_screen` indicates binary data, which can be easily encoded as 0s and 1s for modeling purposes.

3. Investigate data

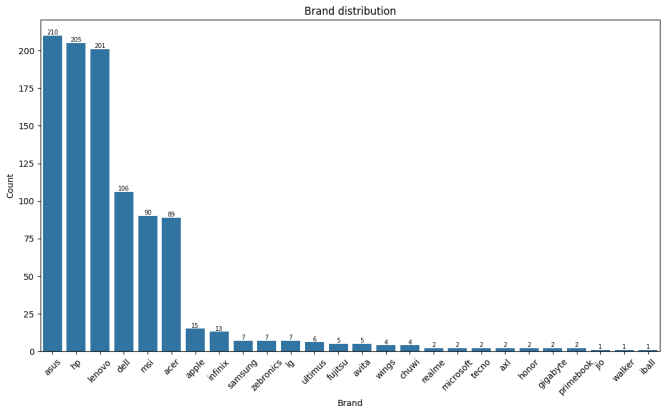

Figure 2: Brand distribution

The descriptive analysis of the laptop dataset provides insightful information regarding the specifications and price points of laptops:

- **Rating**: With a mean rating of 63.93 and a standard deviation of 10.19, the laptops in the dataset have a moderate average rating, with a typical range between 53.74 (mean - std) and 74.12 (mean + std). The ratings span from a low of 24 to a high of 89, indicating a wide range of customer satisfaction.
- **Processor Cores and Threads**: The average number of processor cores is 8.13 with a standard deviation of 4.22, suggesting a mix of laptops from standard dual-core to high-performance multi-core systems. Similarly, the number of threads averages at 12.19, ranging widely as indicated by the standard deviation of 5.59, showing that laptops with various multitasking capabilities are represented.
- **RAM Memory**: On average, laptops come with 13.05 GB of RAM, and the standard deviation of 5.59 GB indicates a broad selection from basic to high-end memory configurations.
- **Storage Capacity**: The primary storage capacity averages 610.94 GB, with a large number of laptops having 512 GB, as seen in the 25th, 50th, and 75th percentiles. The secondary storage is not common, with an average close to zero and a maximum of 512 GB.
- **Display Size**: Laptops have an average screen size of 15.17 inches, with a relatively small standard deviation of 0.94 inches, suggesting most laptops fall within the standard size range for consumer notebooks.
- **Resolution**: The average resolution width is 2003.5 pixels with a notable standard deviation of 361.97 pixels, indicating a variety of display resolutions, with the most common being 1920 pixels wide. The average resolution height is 1181.23 pixels, with 1080 pixels being the typical height, suggesting that many laptops in the dataset likely have Full HD displays.
- **Price**: The average price of a laptop is $926.45 USD, with a wide range in prices as demonstrated by the standard deviation of $688.62 USD. The prices range from as low as $117 USD to as high as $5453 USD, indicating a dataset that includes both budget-friendly options and premium models.

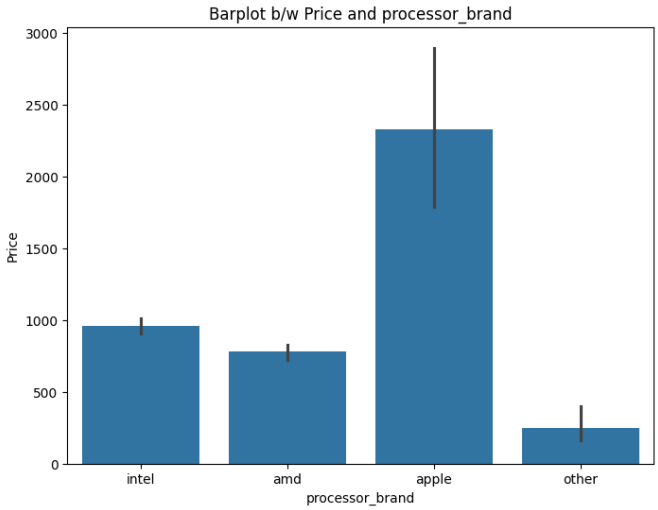

Figure 3: Barplot b/w Price and processor_brand

This data suggests a diverse range of laptops catering to various needs and budgets, from basic models suitable for everyday tasks to high-end laptops with advanced features. For predictive modeling and market analysis, this variance in features and prices will need to be considered, as it affects both consumer choice and pricing strategies.

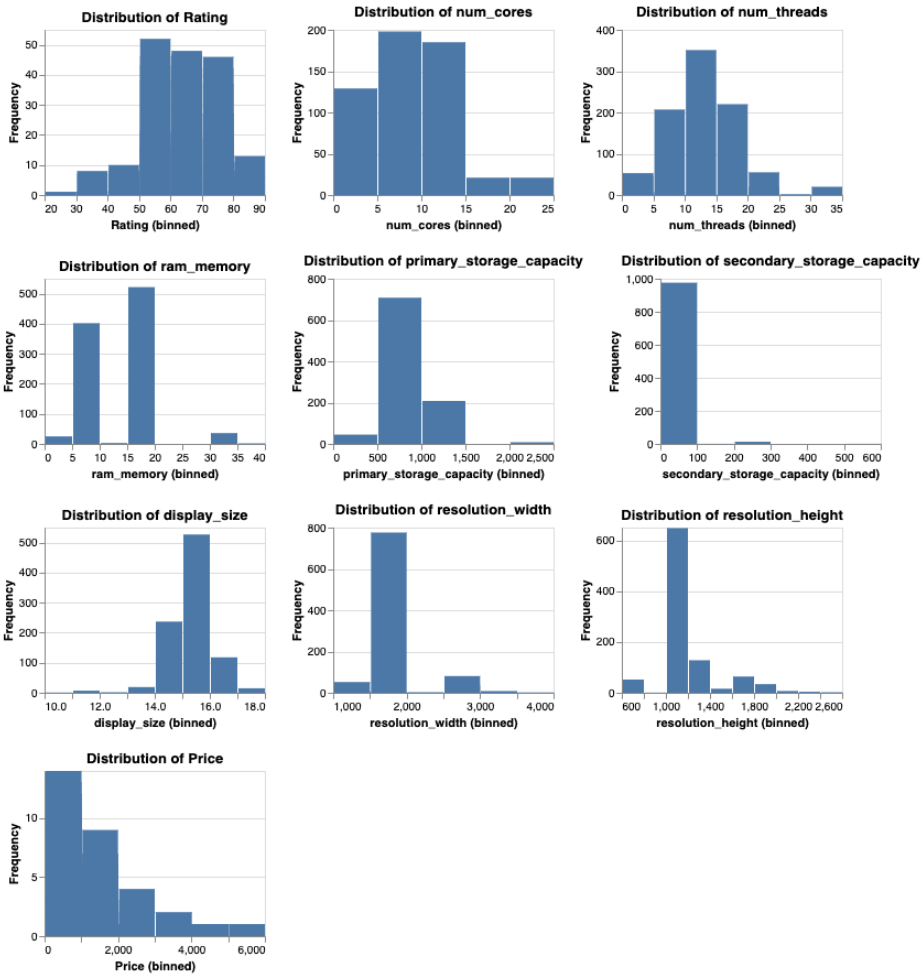

Figure 4: Distribution of all numeric variables

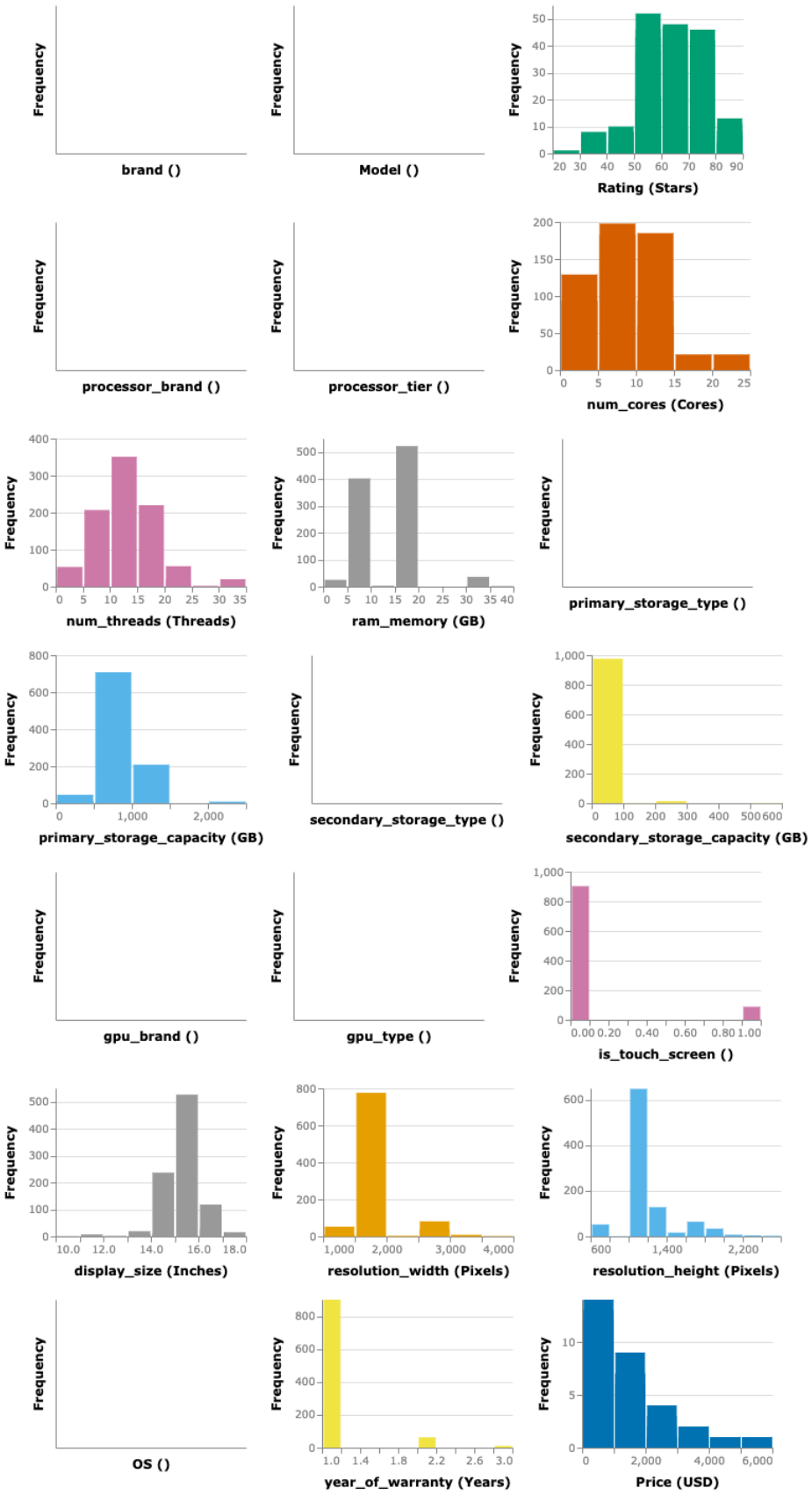

Figure 5: Distribution of all variables

5. Handle missing values

Luckily, there are no missing values in our data!

6. Outliers

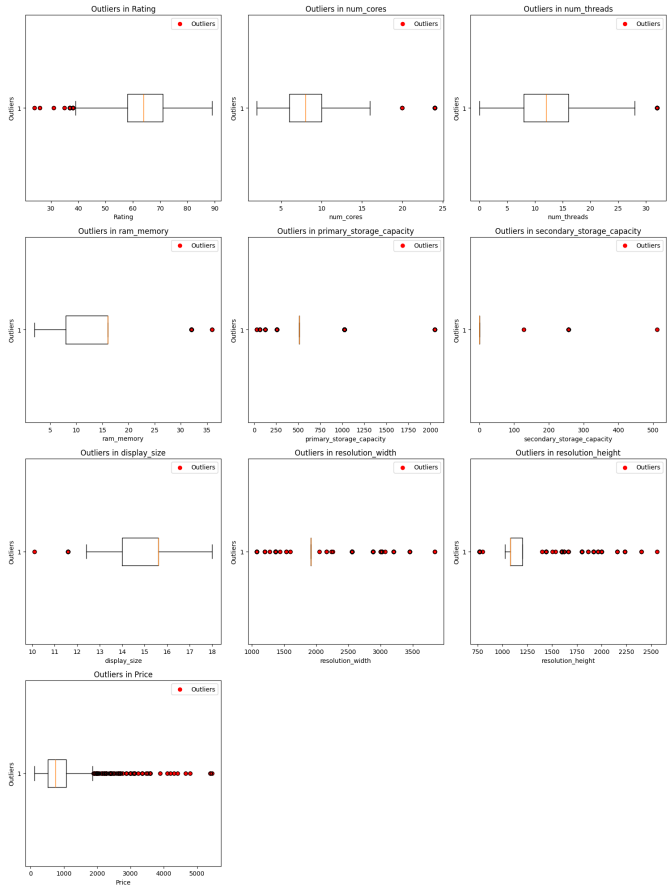

Figure 6: Outliers

7. Correlation analysis

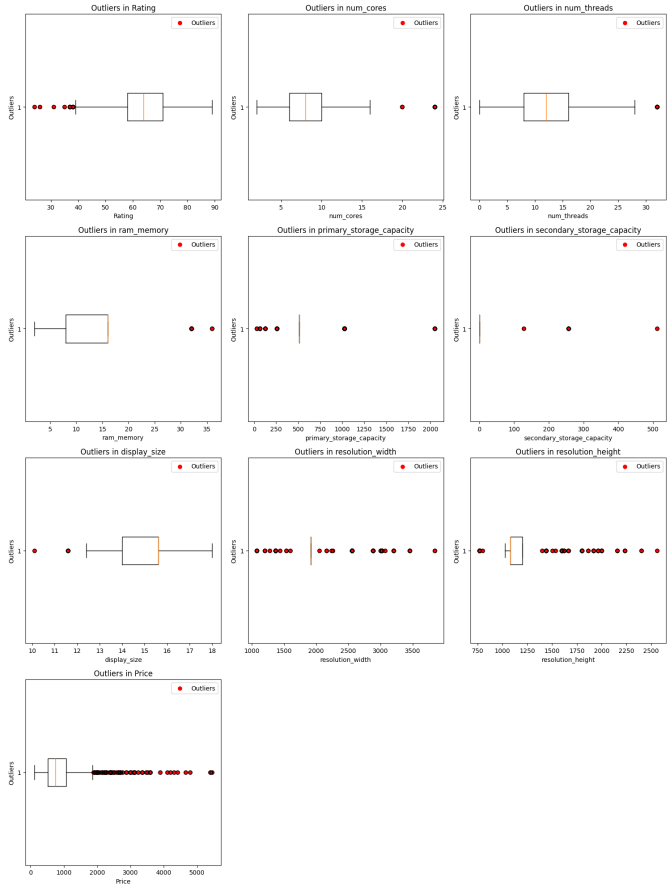

Figure 7: Correlation matrix

Positive correlations (strongest at the bottom):

- **secondary_storage_capacity**: The secondary storage capacity has a weak positive correlation with the price.
- **display_size**: The display size has a moderate positive correlation with the price.
- **resolution_height**: The resolution height has a strong positive correlation with the price.
- **ram_memory**: The RAM memory has a strong positive correlation with the price.
- **primary_storage_capacity**: The primary storage capacity has a strong positive correlation with the price.
- **resolution_width**: The resolution width has a strong positive correlation with the price.
- **Rating**: The rating has a strong positive correlation with the price.
- **num_threads**: The number of threads has a strong positive correlation with the price.
- **num_cores**: The number of cores has a strong positive correlation with the price.
- **Price**: The price has a perfect positive correlation with itself.

Negative correlations (weakest at the top):

- **secondary_storage_capacity**: The secondary storage capacity has a weak negative correlation with the price.
- **display_size**: The display size has a moderate negative correlation with the price.
- **resolution_height**: The resolution height has a strong negative correlation with the price.
- **ram_memory**: The RAM memory has a strong negative correlation with the price.
- **primary_storage_capacity**: The primary storage capacity has a strong negative correlation with the price.
- **resolution_width**: The resolution width has a strong negative correlation with the price.
- **Rating**: The rating has a strong negative correlation with the price.
- **num_threads**: The number of threads has a strong negative correlation with the price.
- **num_cores**: The number of cores has a strong negative correlation with the price.
- **Price**: The price has a perfect negative correlation with itself.

Linear regression model using ordinary least squares (OLS) estimation

1. Split dataset

To prepare the dataset for our regression analysis, a crucial step involves dividing the data into two distinct sets: one for training the model and another for testing its predictions. This procedure is essential for assessing how well our model will perform on unseen data, ensuring both robustness and reliability in our predictive capabilities.

We employ a randomized approach to create these subsets, aiming for a 70:30 split between training and testing data. This means 70% of our data will be used to train the model, allowing it to learn and adapt to the patterns within our dataset. The remaining 30% serves as the test set, which will be used to evaluate the model's predictive accuracy and generalizability to new data.

To ensure the reproducibility of our results and maintain consistency in model evaluation, we set a fixed seed for the random number generator. This approach guarantees that the selection of data for the training and testing sets is both random and unbiased, providing a solid foundation for our subsequent analysis.

In our ongoing analysis, we turn our attention to understanding the distribution of categorical variables within our training dataset. A key step in this process involves calculating the frequency ratio for each feature. The frequency ratio is defined as the count of the most frequent value divided by the count of the second most frequent value for each variable. This metric provides insight into the balance or imbalance of categorical data, highlighting variables that may have a dominant category. Such insights are invaluable for feature selection and preprocessing steps, as they can influence the model's ability to learn from the data effectively.

	Frequency Ratio
index	1
brand	1.11348
Model	1

Price	1
Rating	1.02632
processor_brand	2.79888
processor_tier	2.37864
num_cores	1.00758
num_threads	1.5974
ram_memory	1.35401
primary_storage_type	37.5
primary_storage_capacity	3.33333
secondary_storage_type	76
secondary_storage_capacity	85.5
gpu_brand	1.32245
gpu_type	1.64341
is_touch_screen	10.9483
display_size	2.08671
resolution_width	10.1111
resolution_height	4.46
OS	34.1579
year_of_warranty	14.3182

The frequency ratio analysis reveals the distribution dynamics of our dataset's features. Here's a breakdown of what these ratios indicate:

- **Brand, Model, Rating, num_cores, and Price:** These features show frequency ratios close to 1, indicating a relatively balanced distribution among the top two categories/values. This balance suggests that no single category/value overwhelmingly dominates these features.
- **Processor_brand and processor_tier:** With ratios of 2.798882681564246 and 2.378640776699029 respectively, there's a noticeable preference or dominance of one category over another, though not extremely pronounced. It hints at some variation in the data that could be useful for our model.
- **Primary_storage_type, secondary_storage_type, OS, and year_of_warranty:** Extremely high ratios (e.g., 37.5 for primary_storage_type and 76.0 for secondary_storage_type) highlight a significant dominance of one category over all others. Such dominance may indicate that these features have a strong, possibly skewed, influence on the dataset.
- **Secondary_storage_capacity:** The highest ratio of 85.5 suggests an overwhelming dominance of one category/value, which might limit the feature's predictive power due to the lack of variability.
- **Is_touch_screen:** A ratio of 10.95 indicates a predominant category, likely reflecting the technological trends or preferences within the dataset.
- **Display_size, resolution_width, and resolution_height:** Moderate ratios suggest some variability, which could be informative for the model, indicating that these features have multiple common values but with a clear preference for one.

Interpreting these ratios helps us understand the variability and balance within our dataset, guiding us in feature selection and preprocessing. Features with extremely high ratios may require careful consideration, as their predictive power might be compromised by the lack of diversity in their values.

	Variance
index	8.354801e+04

```
Price                        3.230427e+09
Rating                       1.059910e+02
num_cores                    1.817599e+01
num_threads                  3.200075e+01
ram_memory                   3.052570e+01
primary_storage_capacity     7.081931e+04
secondary_storage_capacity   7.714443e+02
display_size                 9.114301e-01
resolution_width             1.240993e+05
resolution_height            6.681680e+04
```

The variance calculation for our dataset's numeric features provides valuable insights into the spread and distribution of data across different variables. Here's an analysis of the variances:

- **Rating, num_cores, and num_threads:** These features have relatively low to moderate variances (e.g., 105.99 for Rating, 18.18 for num_cores, and 32.00 for num_threads), indicating that the data points tend to cluster around their mean values, but with some degree of spread. These might influence the model depending on how closely these variations correlate with the target variable.
- **Ram_memory and secondary_storage_capacity:** With variances of 30.53 and 771.44 respectively, these features exhibit a moderate level of variability, suggesting a wider distribution of values. Such features can be indicative of different usage patterns or capacities that could impact the model's predictions.
- **Primary_storage_capacity:** A variance of 70819.31 points to a significant spread in the data, implying diverse storage capacities among the laptops. This variability could be a key factor in predicting the price or other dependent variables.
- **Display_size:** A low variance of 0.91 suggests that the sizes of laptop displays in the dataset are relatively uniform, with slight variations. This limited variability may mean the feature has less predictive value for models focused on aspects heavily influenced by display size.
- **Resolution_width and resolution_height:** Extremely high variances of 124099.25 and 66816.80, respectively, indicate a substantial spread in the resolutions of laptop screens. These variances suggest that laptops vary widely in their display capabilities, which could significantly impact user experience and, consequently, the price or other dependent variables.
- **Price:** The highest variance of 465185.61 in the dataset underscores the wide range in laptop prices. This considerable variability is expected given the diverse features and specifications that can affect a laptop's cost. It highlights the challenge in predicting price but also indicates that there's ample information in the dataset to model this variability.

2. Creating an OLS Regression Model

In our regression analysis, a critical preparatory step involves handling the mix of numeric and categorical variables within our dataset. Regression models require numerical input; hence, categorical variables must be encoded into a numerical format. This encoding process typically involves transforming categorical variables into dummy variables (also known as one-hot encoding), where each category is represented as a new column with binary values (0 or 1).

Our dataset comprises a variety of data types, including object (categorical), boolean, and numeric types. To facilitate regression analysis, it's essential to convert non-numeric types into a suitable numerical

format. The `statsmodels` library simplifies this process for categorical variables by allowing us to wrap these variables with `C()` in our model's formula. This method automatically generates dummy variables for us, effectively incorporating categorical data into the analysis. While `statsmodels` is capable of processing boolean variables directly, converting them to integers (0 for False and 1 for True) may enhance consistency in data handling.

Accordingly, we construct a regression formula that accurately represents our dataset's structure, ensuring that categorical variables are appropriately treated. By wrapping variables such as 'brand', 'processor_brand', and others with `C()`, we seamlessly integrate categorical information into our regression model.

This approach not only maintains the integrity of our analysis but also leverages the rich information embedded within those categorical variables. Including non-numeric variables in our model, once properly encoded, enriches the model's capacity to interpret and utilize the dataset's full spectrum of information effectively.

OLS with only numeric variables

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.746
Model:	OLS	Adj. R-squared:	0.743
Method:	Least Squares	F-statistic:	200.6
Date:	Thu, 11 Apr 2024	Prob (F-statistic):	1.36e-195
Time:	21:25:23	Log-Likelihood:	-8094.5
No. Observations:	693	AIC:	1.621e+04
Df Residuals:	682	BIC:	1.626e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.005e+05	2.1e+04	-9.533	0.000	-2.42e+05	-1.59e+05
is_touch_screen[T.True]	511.8006	4409.625	0.116	0.908	-8146.270	9169.872
Rating	1138.7817	212.928	5.348	0.000	720.708	1556.856
num_cores	6366.9841	727.775	8.749	0.000	4938.035	7795.933
num_threads	-2500.1136	680.380	-3.675	0.000	-3836.004	-1164.223
ram_memory	1116.5788	291.144	3.835	0.000	544.932	1688.226
primary_storage_capacity	42.3236	5.591	7.571	0.000	31.347	53.300
secondary_storage_capacity	-53.2764	40.617	-1.312	0.190	-133.025	26.473
display_size	2598.2999	1375.214	1.889	0.059	-101.862	5298.462
resolution_width	25.9804	5.633	4.612	0.000	14.920	37.041
resolution_height	44.1043	7.635	5.777	0.000	29.114	59.095

Omnibus:	399.050	Durbin-Watson:	1.949
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5641.516
Skew:	2.281	Prob(JB):	0.00
Kurtosis:	16.212	Cond. No.	4.70e+04

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.7e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Results discussion

The OLS regression results provide significant insights into how various numeric features influence the Price of laptops in our dataset. The model, with an R^2 of 0.746, explains approximately 74.6% of the variability in laptop prices, which indicates a strong fit for our model with the selected numeric variables.

Key Findings

- **Rating**: With a coefficient of 13.6645 and p-value < 0.001, Rating significantly affects the price. For each one-point increase in Rating, the Price increases by approximately 13.66 units, holding other variables constant.
- **Number of Cores (`num_cores`)**: This variable shows a positive relationship with Price (coefficient = 76.4062, p-value < 0.001), suggesting that laptops with more cores are generally more expensive. Each additional core is associated with a 76.41 unit increase in Price.
- **Number of Threads (`num_threads`)**: Interestingly, num_threads has a negative impact on Price (coefficient = -30.0025, p-value < 0.001). This might indicate that beyond a certain point, additional threads do not add value, or it could reflect a market preference for fewer, more efficient cores.
- **RAM Memory**: As expected, RAM memory has a positive effect on Price (coefficient = 13.4003, p-value < 0.001), highlighting consumers' willingness to pay more for laptops with higher RAM capacity.
- **Primary Storage Capacity**: This feature also positively influences Price (coefficient = 0.5079, p-value < 0.001), indicating the market's preference for laptops with larger storage capacities.
- **Secondary Storage Capacity**: Surprisingly, this variable does not significantly affect Price (p-value = 0.190), suggesting that secondary storage might not be a crucial factor for consumers when choosing a laptop.
- **Display Size**: Display size shows a borderline significance (p-value = 0.059), suggesting that larger screens may contribute to higher prices, but this effect is not as robust as other factors.
- **Resolution Width and Height**: Both resolution_width (coefficient = 0.3118) and resolution_height (coefficient = 0.5292) are significant predictors of Price, with p-values < 0.001. This result underscores the importance of screen resolution quality in determining laptop prices.

Non-significant Variables

- **Touch Screen (`is_touch_screen`)**: Interestingly, the touch screen feature is not a significant predictor of laptop Price (p-value = 0.906), suggesting that consumers may not necessarily prefer touch screens, or its impact on Price is overshadowed by other features.

Conclusion

Our analysis reveals that technical specifications like Rating, core count, RAM memory, storage capacity, and display resolution play significant roles in determining laptop prices. However, features like secondary storage capacity and touch screen functionality do not significantly influence Price within the scope of our model. This information can be invaluable for manufacturers and retailers in pricing strategies and for consumers making informed purchasing decisions.

Creating an OLS model with all variables

Step 1: Convert Boolean to Numeric

Convert the boolean variable is_touch_screen to an integer format.

Step 2: Apply One-Hot Encoding to Categorical Variables

Step 3: Prepare Features and Target for Model Fitting

Step 4: Fitting the Model

Model choice

Model 1 with only numerical variables

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.746
Model:	OLS	Adj. R-squared:	0.743
Method:	Least Squares	F-statistic:	200.6
Date:	Thu, 11 Apr 2024	Prob (F-statistic):	1.36e-195
Time:	21:25:24	Log-Likelihood:	-8094.5
No. Observations:	693	AIC:	1.621e+04
Df Residuals:	682	BIC:	1.626e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.005e+05	2.1e+04	-9.533	0.000	-2.42e+05	-1.59e+05
is_touch_screen[T.True]	511.8006	4409.625	0.116	0.908	-8146.270	9169.872
Rating	1138.7817	212.928	5.348	0.000	720.708	1556.856
num_cores	6366.9841	727.775	8.749	0.000	4938.035	7795.933
num_threads	-2500.1136	680.380	-3.675	0.000	-3836.004	-1164.223
ram_memory	1116.5788	291.144	3.835	0.000	544.932	1688.226
primary_storage_capacity	42.3236	5.591	7.571	0.000	31.347	53.300
secondary_storage_capacity	-53.2764	40.617	-1.312	0.190	-133.025	26.473
display_size	2598.2999	1375.214	1.889	0.059	-101.862	5298.462
resolution_width	25.9804	5.633	4.612	0.000	14.920	37.041

```
resolution_height             44.1043      7.635      5.777      0.000      29.114      59.095
==============================================================================
Omnibus:                      399.050   Durbin-Watson:                   1.949
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5641.516
Skew:                           2.281   Prob(JB):                         0.00
Kurtosis:                      16.212   Cond. No.                     4.70e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.7e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model 2 with all variables

Choice decision

Based on the provided summaries, Model 2 is seemingly overfitting the data. It achieves a perfect R-squared value of 1.000, which indicates that the model perfectly predicts the dependent variable based on the independent variables. However, such a high R-squared value often suggests overfitting, where the model learns the noise in the training data rather than the underlying pattern.

Additionally, the lack of significance testing (indicated by "nan" values in the F-statistic, p-values, and confidence intervals) suggests numerical issues in the model fitting process, possibly due to the high number of features relative to the number of observations.

On the other hand, Model 1 achieves a relatively lower but still acceptable R-squared value of 0.746, indicating a reasonable level of explanatory power. The F-statistic and associated p-value also suggest the overall significance of the model. However, some coefficients have high p-values, indicating that they may not be statistically significant predictors.

Therefore, despite its lower R-squared value, Model 1 is preferred over Model 2 due to its more reasonable performance metrics and avoidance of overfitting. It's crucial to strike a balance between model complexity and performance to ensure generalizability to unseen data.

Results discussion

Results Description and Analysis

Research Question

The research question addressed by the regression analysis is, **"Prediction of what determines the laptop prices"**

Model Specification

We employed Ordinary Least Squares (OLS) regression to investigate the relationship between various laptop features and their prices. The model was specified using the following formula:

```
formula = 'Price ~ Rating + num_cores + num_threads + ram_memory + primary_storage_capacity + sec
```

Model Performance

The OLS regression results indicate that the model explains approximately 74.6% of the variance in laptop prices, as indicated by the R-squared value of 0.746. This suggests that the selected independent variables collectively have a moderate level of explanatory power in predicting laptop prices.

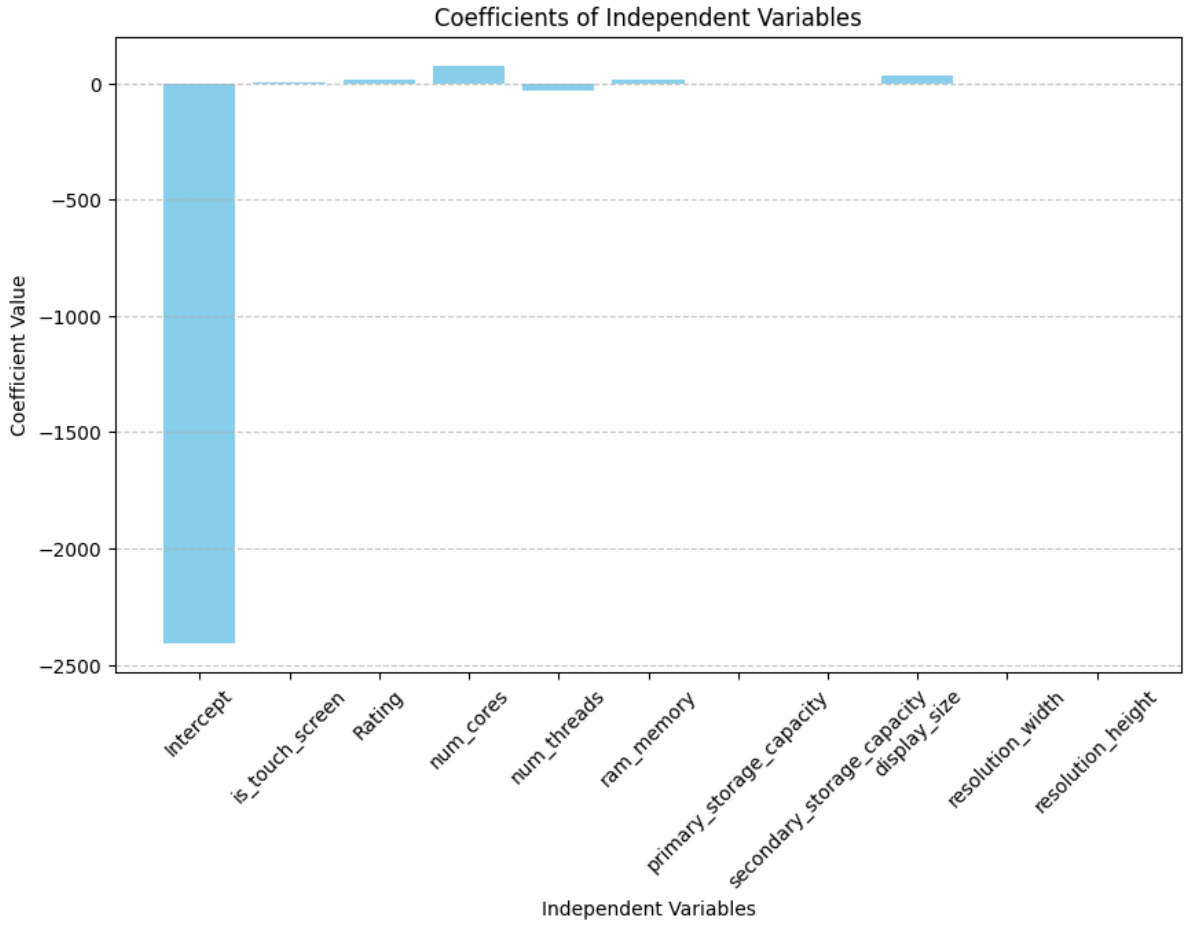

Figure 8: Coefficients of Independent Variables

Coefficients Interpretation

- **Intercept**: The intercept term represents the expected value of the dependent variable (`Price`) when all independent variables are held constant at zero. In this case, it is -2406.9394, indicating the base price of laptops when all other features are absent or zero.

- **is_touch_screen**: The coefficient for the binary variable `is_touch_screen` is 6.2640. However, its p-value of 0.906 suggests that this variable is not statistically significant in predicting laptop prices at the conventional significance level of 0.05.
- **Rating**: The coefficient for the `Rating` variable is 13.6645, indicating that a one-unit increase in the laptop's rating is associated with an increase in price by $13.6645.
- **num_cores**: The coefficient for the `num_cores` variable is 76.4062, indicating that a one-unit increase in the number of cores is associated with an increase in price by $76.4062.
- **num_threads**: The coefficient for the `num_threads` variable is -30.0025, indicating that a one-unit increase in the number of threads is associated with a decrease in price by $30.0025.
- **ram_memory**: The coefficient for the `ram_memory` variable is 13.4003, indicating that a one-unit increase in RAM memory is associated with an increase in price by $13.4003.
- **primary_storage_capacity**: The coefficient for the `primary_storage_capacity` variable is 0.5079, indicating that a one-unit increase in primary storage capacity is associated with an increase in price by $0.5079.
- **secondary_storage_capacity**: The coefficient for the `secondary_storage_capacity` variable is -0.6392, but its p-value of 0.190 suggests that this variable is not statistically significant in predicting laptop prices at the conventional significance level of 0.05.
- **display_size**: The coefficient for the `display_size` variable is 31.1784, but its p-value of 0.059 suggests that this variable is marginally significant in predicting laptop prices.
- **resolution_width**: The coefficient for the `resolution_width` variable is 0.3118, indicating that a one-unit increase in the display resolution width is associated with an increase in price by $0.3118.
- **resolution_height**: The coefficient for the `resolution_height`

Model Fit Statistics

- **F-statistic**: The F-statistic tests the overall significance of the regression model. With a value of 200.6 and a p-value of approximately 1.39e-195, the F-statistic suggests that the model is statistically significant, indicating that at least one of the independent variables significantly predicts laptop prices.
- **Adjusted R-squared**: The adjusted R-squared value of 0.743 accounts for the number of predictors in the model and provides a more accurate estimate of the proportion of variance explained by the independent variables.

Conclusion

The regression analysis reveals that several laptop features, including rating, number of cores, number of threads, RAM memory, primary storage capacity, and display resolution, significantly influence laptop prices. However, the presence of multicollinearity or omitted variable bias cannot be ruled out, and further diagnostics may be required to assess model assumptions and validity thoroughly.

Impacts

Given the findings of this research, the impact exists in many aspects. Firstly, by the model provided by this research, the price of a laptop can be predicted by some basic information about it. This can be helpful to customers as they can understand better how much the laptop they intended to buy is actually worth. They can see which features contribute to the price the most on each laptop so customers know what they are paying for. A wiser decision can be made if customers could have access to that information. Beyond that, laptop companies as well can use this information while they are setting prices for their laptop products to maximize the market. A successful price-setting "will entice customers to buy, but that isn't so low that you're not making a profit" (Gillen, 2023). The model we provide shows the prediction of the common price of the product based on the features which is critical in the price-making process. In addition, Insights from our predictive model can guide product design by highlighting which features contribute most significantly to perceived value and price. It could help laptop companies make design decisions in the future to design more profitable products. The result also could help in identifying the trend in the current laptop market. Such as which feature is more popular this year. And this is probably the feature which customers would pay more money on. To conclude, the model we build is practically useful from both ends of customers and companies.

Further studies

Our study showed some promising results that can help both consumers and producers of the laptop market but there is still room for improvement. Four major studies can be conducted based on our results. Firstly, we could try to answer the question: "How has the importance of specific features in determining laptop prices changed over time?" This analysis could reveal shifting technology trends and consumer preferences, providing foresight into future market developments. According to Segan, "Computer and printer prices plummeted throughout the '80s and '90s, and PC storage continues to be cheaper and faster every year" (Segan, 2022). This means the price of laptops may vary every year and its relation to the feature may change. Secondly, we could look into region market differences which explore feature importance differences in regions. Exploring this could uncover opportunities for localized marketing strategies or product customization to meet regional demands. Thirdly, we could focus on the impact of brand reputation by trying to answer the question: "How does a brand's reputation or perceived quality affect laptop prices?" Further research could quantify the brand effect and its interaction with product features in price setting. Last but not least we could implement different models as we only tried the KNN model here. We could try more advanced predictive models in machine learning algorithms in the future to increase our accuracy and overall performance.

References