SCHOOL OF COMPUTING AND INFORMATION TECHNOLOGY
BACHELOR OF COMPUTER SCIENCE

5th Semester

# Statistical Inference and Modeling (ITS66904)

## Group Assignment (30% )

**Group Name: The Bandits**

| Student ID | Student Name |
|---|---|
| 355572 | Kushal Pandey |
| 355612 | Manoj Paudel |
| 355491 | Kabin Shrestha |
| 355454 | Rojan Lal Shrestha |
| 355637 | Ritin Karanjit |

# Table of Contents

## Table of Contents

# Abstract

This research paper presents a Laptop Price Prediction Model, which involves cleaning and organizing data, exploring data patterns, and building predictive models using statistical methods and machine learning. The main goal is to create a prediction model that accurately predicts the laptop prices when based on the features and specifications given by the users, benefiting both buyers and sellers.

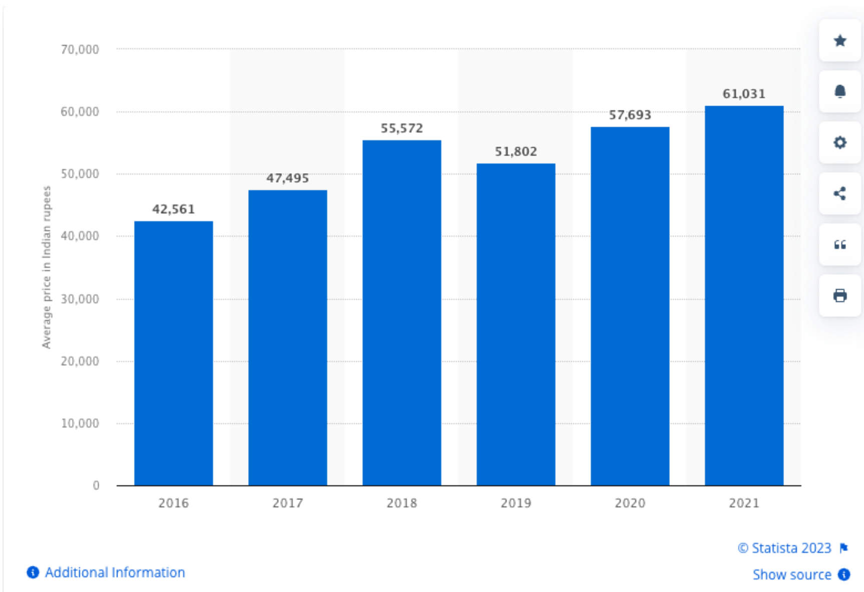
The prediction system relies on supervised machine learning and the models used are Linear Regression, Support Vector Regressor and Random Forrest. To develop an efficient prediction model, data cleaning and preprocessing is done to ensure the data quality while exploratory data analysis has helped finding data patterns, and insights to what factors influence the price the most. Proper feature engineering and transforming makes the data suitable for the machine learning algorithms to build the model.

After the model is built, highest R2 score of 99% is achieved by the Random Forrest, with which we can deduce that the model can predict the price of laptop with minimum accuracy of 90%. Once this project is implemented properly, it will help buyers budget properly for their laptops based on their specific needs. It will prevent them for overpaying for the laptops in the market.

# Introduction

In today’s age of science and technology, we all revolve around different types of tech products for our regular daily life. The most major gadget that is the most important among all for all students/ workers/ business people etc. is a *Personal Laptop*. And the majority of the consumers of the product are the students. It is a must have equipment for them for making their learning activity more effective and easier. Since most of the students are unemployed and they lack in self owned money, it is very difficult for them to be able to afford it. And when presented with a variety of specifications and prices, it is really difficult for them to decide whether they are getting the proper rate or not.

While a laptop only seems to be a simple tech product, there are a lot of factors that affect the price of the product. The processor that it is using, the amount of ram it consists of, the amount of graphical content it can handle with its graphic card, the operating system that it is using and also the display quality it is using and so on. Since most of the product can have variations in prices with its constituent parts, it's really difficult to assume the prices of the products.




*Source: Average price of laptops in India 2016-2021*
*Published by Shangliao Sun, Dec 22, 2022*

Most of the distributors/sellers are willing to make as much profit as they can while selling the product while the consumers are sometimes not aware of the product current running price and the marked price by the sellers. So most consumers are overpaying due to their lack of awareness of the proper valid price. As the price goes higher and higher, it will be more and more difficult for people with a low budget to afford it hence will create a technological gap for people who can't afford it. The most important sector is in education, the unavailability will create a huge gap in the study progress.The students with low resources will be capped to their learning impractically too. A decline in laptop sales caused by price increases may cause computer businesses to spend less on study and development. Future laptop models could have less cutting-edge features and see a slower pace of technological development as a result.

Our prediction model uses the data that consists of numerous possible parts of the laptop and predicts its price with proper accuracy. The data sees the previous sold price which is accurate and is able to predict the price with different  parts.

With our proper analysis and prediction model, when deployed the users will be able to see the predicted price and be able to buy the product at the best rate possible.

# Related works

“Laptop Price Prediction with Machine Learning Using Regression Algorithm”,
A. D. Siburian, “Laptop Price Prediction with Machine Learning Using Regression Algorithm”, *JUSIKOM PRIMA*, vol. 6, no. 1, pp. 87-91, Sep. 2022.

This research paper presents a Machine Learning model from data acquisition, Data Cleaning, and Feature Engineering for the Pre-Processing, Exploratory Data Analysis stages to modeling based on regression algorithms. The proposed method is to predict laptop prices using Random Forrest Regressor, Gradient Boosting Regressor and XGBoost Regressor algorithms.

The proposed method of predicting laptop prices using RF Regressor, GB Regressor and XGB Regressor algorithms was successful, with the XGBoost algorithm having the highest accuracy result of 92.77%. The model created can predict laptop prices with a minimum accuracy above 80% produced by the RF Regressor. The XGBoost model that has been created can also be used to make predictions in real-time using a web-based Machine Learning application. The study concludes that with a model that can predict the price of a laptop, employees can more easily determine the laptop that suits their needs.

# Data

For our project, we are using a Kaggle dataset with 1303 entries on Laptop Price Prediction. This dataset is open to the public. It includes a wide range of attributes that provide a comprehensive picture of different laptops from various manufactures and its listing price with other attributes. The dataset has 12 columns,

| **Features** | |
|---|---|
| Company | The different manufacturing company names. |
| TypeName | The laptop type such as,<br>(Notebooks, Ultrabook, Gaming laptops etc.) |
| Inches | The screen size of the laptop. |
| ScreenResolution | The screen resolution of the laptop, display quality. |
| Cpu | The processor types with speed. |
| Ram | The RAM capacity of the laptop |
| Memory | The Hard Disk, SSD storage capacity. |
| Gpu | The different GPU configuration. |
| OpSys | The different operating systems. |
| Weight | The weight of the laptop. |
| Price | Price of different laptops in INR. (Target variable of our model) |

The original source of the data comes from Amazon with the help of data scraping from the website.

**Statistical Analysis:**

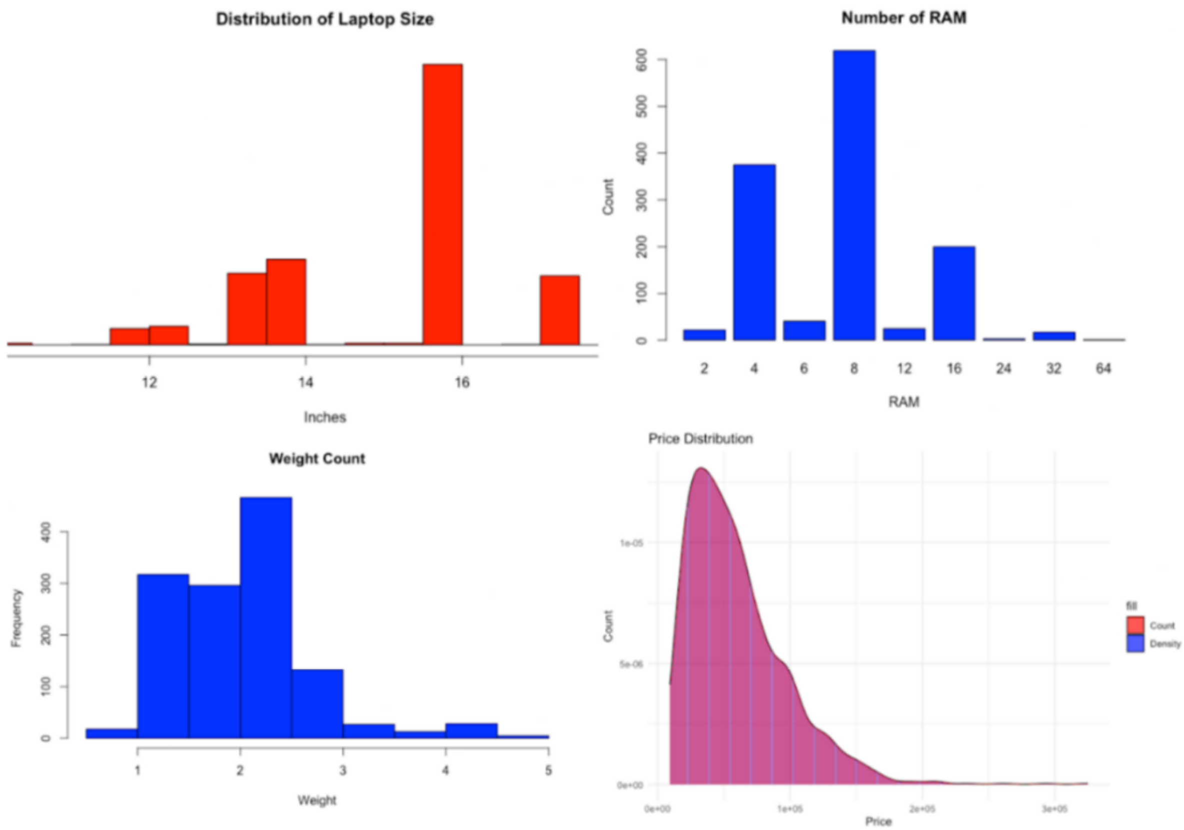
We now extract and reorganize our data to better understand the underlying factors that contribute to the price of laptops.The columns for screen resolution, memory, and CPU contain a wide range of information, including details about touchscreens, Ips panels, i3, i5, and i7 processors, SSDs, and HDDs and so on. In the Method section below, we'll go over data cleaning for this column in more detail.

If we look at the summary of statistics for numerical variable in dataset such as Inches, Ram, Weight, Price:

| **Inches** | **Ram** | **Weight** | **Price** |
|---|---|---|---|

| | | | |
|---|---|---|---|
| **Min :** 10.10 | **Min :** 2.000 | **Min :** 0.690 | **Min :** 9271 |
| **1st Qu.:** 14.00 | **1st Qu.:** 4.000 | **1st Qu.:** 1.500 | **1st Qu.:** 31915 |
| **Median :** 15.60 | **Median :** 8.000 | **Median :** 2.040 | **Median :** 52055 |
| **Mean :** 15.02 | **Mean :** 8.382 | **Mean :** 2.039 | **Mean :** 59870 |
| **3rd Qu.:** 15.60 | **3rd Qu.:** 8.000 | **3rd Qu.:** 2.300 | **3rd Qu.:** 79274 |
| **Max. :** 18.40 | **Max. :** 64.000 | **Max. :** 4700 | **Max. :** 324955 |

The distribution of these numerical variables is visualized in the following histograms:




The histogram plots shown above show that the majority of the laptops in our dataset are between 15 and 16 inches in size, have 8 GB of RAM, and weigh more than 2 kg. If we examine the Price plot, we may see that it is right-skewed; however, for better model prediction in the upcoming EDA section, we will convert it to normal using log transformation.

# Methodology

## Exploratory Data Analysis (EDA)

In the initial stages of our laptop price prediction project, we conducted Exploratory Data Analysis (EDA) to better understand our dataset and lay the foundation for our modeling efforts. The following steps were taken during the EDA phase:

1. **Data Cleaning**: We began by cleaning the dataset to ensure its quality. This involved addressing missing values, handling outliers, and resolving any inconsistencies.

| Column | After Data Cleaning |
|---|---|
| RAM | Removed ‘GB’ from the data and made numeric |
| Weight | Removed ‘kg’ form the data and made numeric |
| X | Removed the column |
| Screen Resolution | Derived other columns like Touchscreen, Resolution, ppi, |
| CPU | Transformed the column into CPU brand name |
| GPU | |
| Processor | Removed unnecessary metrics from the column. |

2. **Summary Statistics**: We calculated essential summary statistics such as mean, median, standard deviation, and quartiles for numerical features. This helped us understand central tendencies and the spread of these features.

3. **Correlation Analysis**: We assessed the relationships between features and the target variable, which is laptop prices. Correlation matrices aided in understanding which features exhibited strong or weak correlations with the target.

4. **Feature Engineering**: Based on the insights gathered during EDA, we performed feature engineering. New features were created, like the ‘ppi’ metric was created from the ‘Screen Resolution’ data. Existing ones were transformed or dropped based on their relevancy to enhance the predictive power of our model.

5. **Data Visualization –** We used GGPlot2 library for the visualizations of data and the correlations between the different variables of our dataset. We also used R’s built in methods for a few graphs.

## Statistical Methods Used

In our laptop price prediction project, we employed several statistical methods to develop and evaluate predictive models. These methods included:

1. **Linear Regression**: Linear regression was used to model the relationship between laptop prices and various predictor variables. This method assumes a linear relationship and provides interpretable coefficients, allowing us to understand the impact of each feature on laptop prices.
2. **Support Vector Regressor (SVR)**: SVR is a powerful regression technique that can capture non-linear relationships in the data. It uses a kernel function to map data into a higher-dimensional space and find the optimal hyperplane for regression.

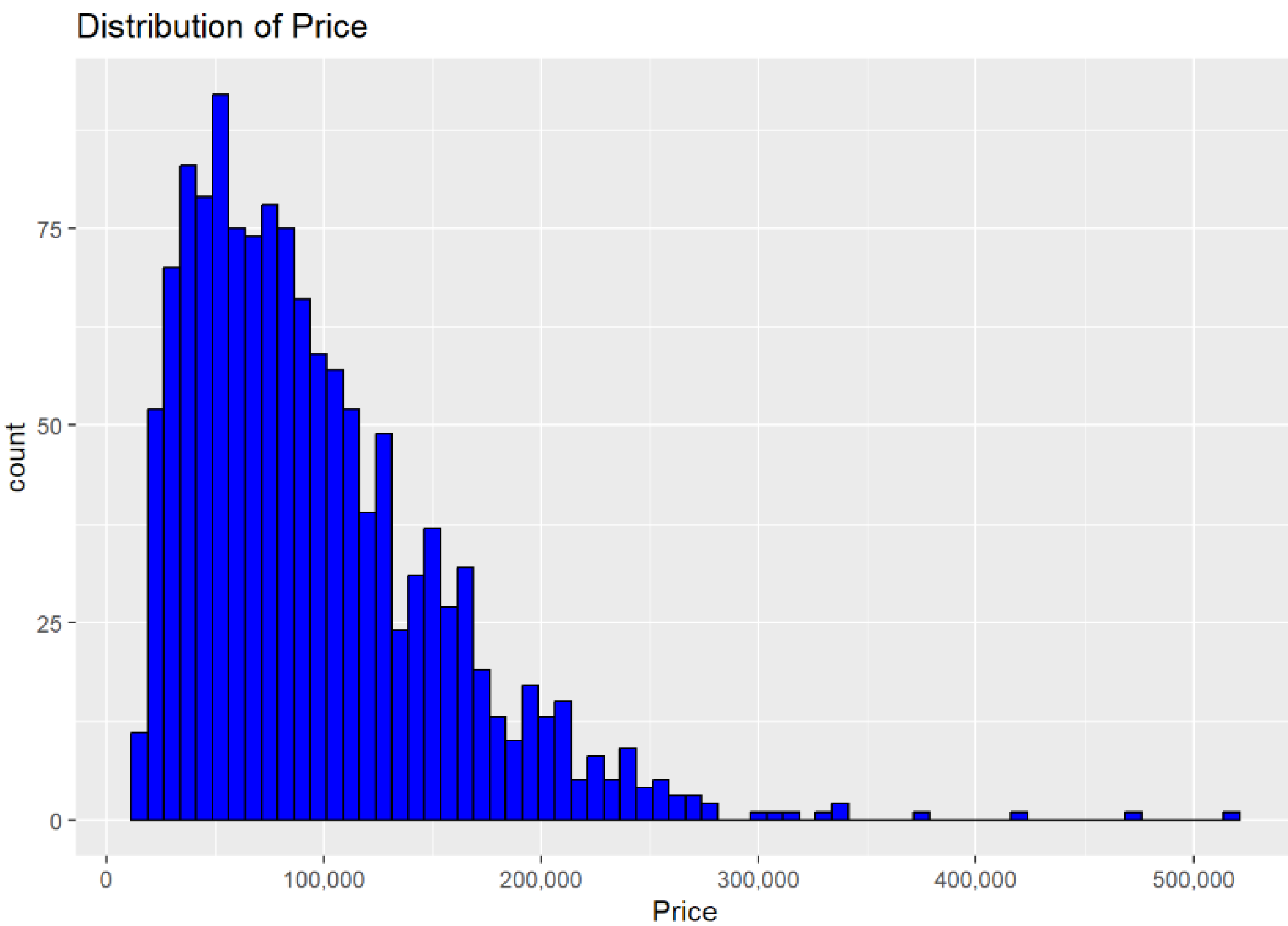
3. **Random Forest**: We also utilized Random Forest, which is an ensemble learning method. It combines multiple decision trees to create a robust and accurate predictive model. Random Forest is particularly effective when dealing with complex data relationships.

By combining EDA with these statistical methods, we aimed to develop a reliable and accurate laptop price prediction model that could provide valuable insights for pricing strategies and decision-making.

# RESULTS

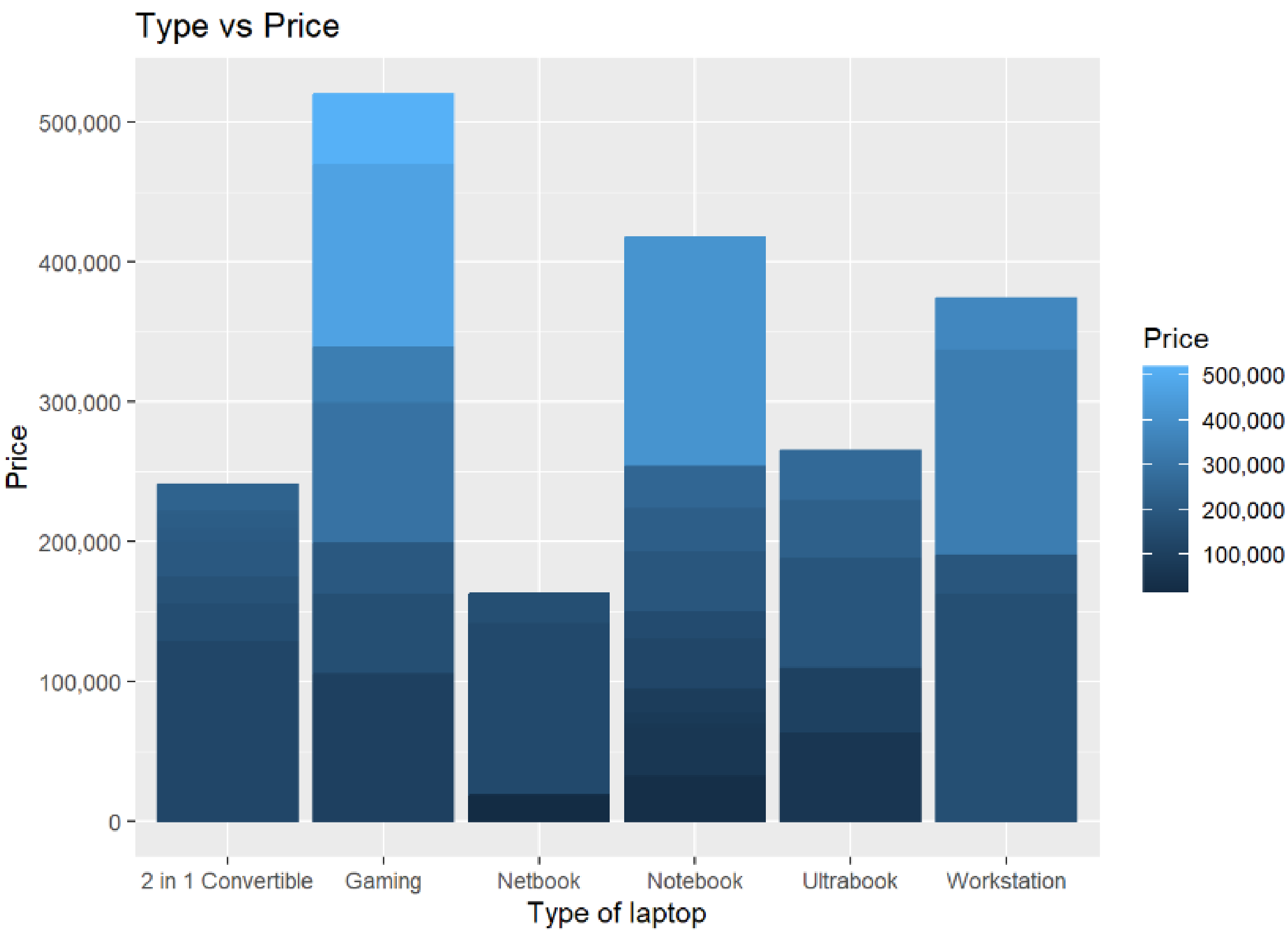
## Result of EDA

With the help of EDA, we found useful insights and data pattern that helped us to be on the right track for the prediction model.

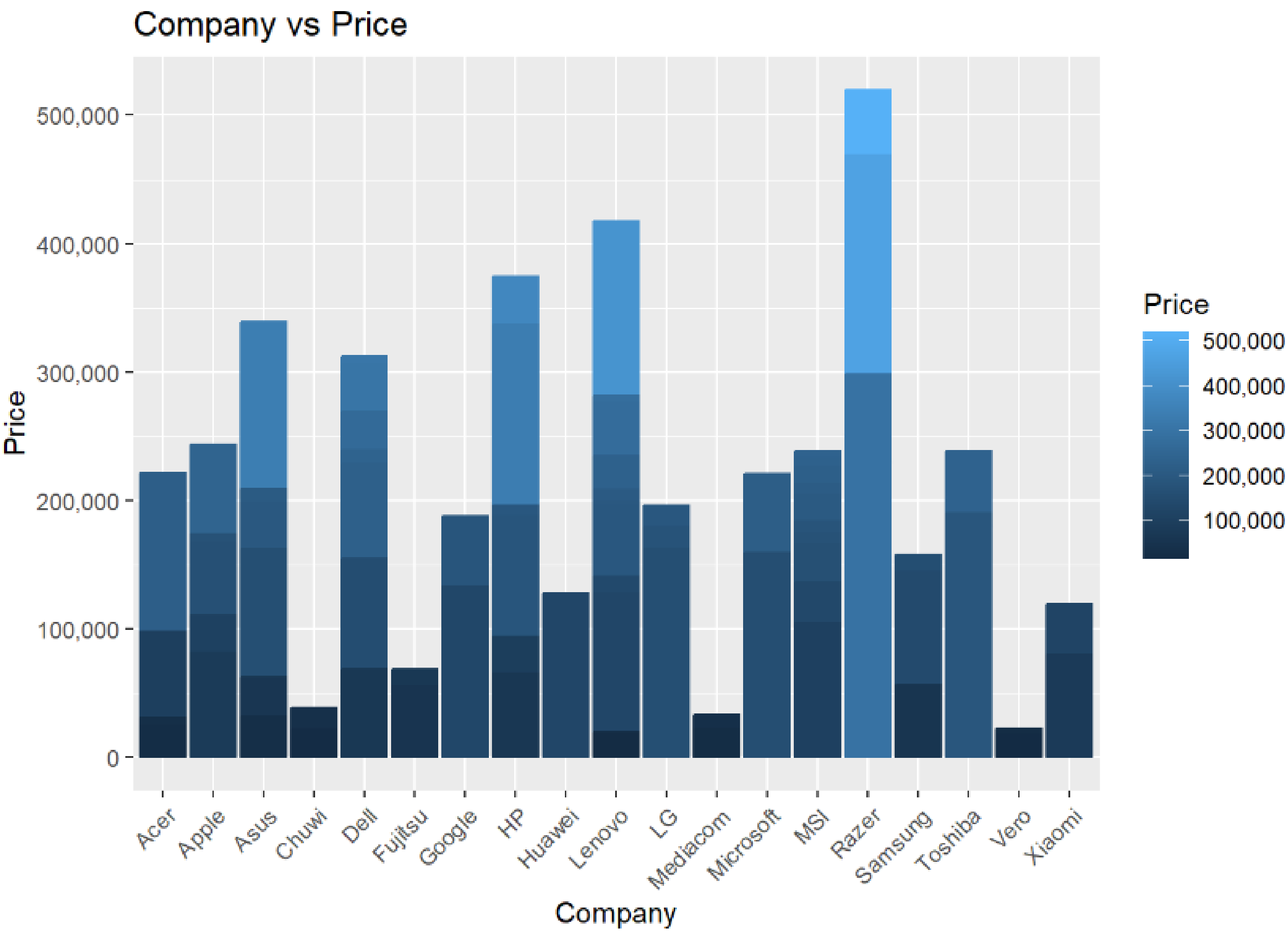



Most of the laptops are priced less than Rs. 100000, while a few fall under the expensive zone of more than Rs. 300000.

Lets see what type of laptops does our dataset contain.



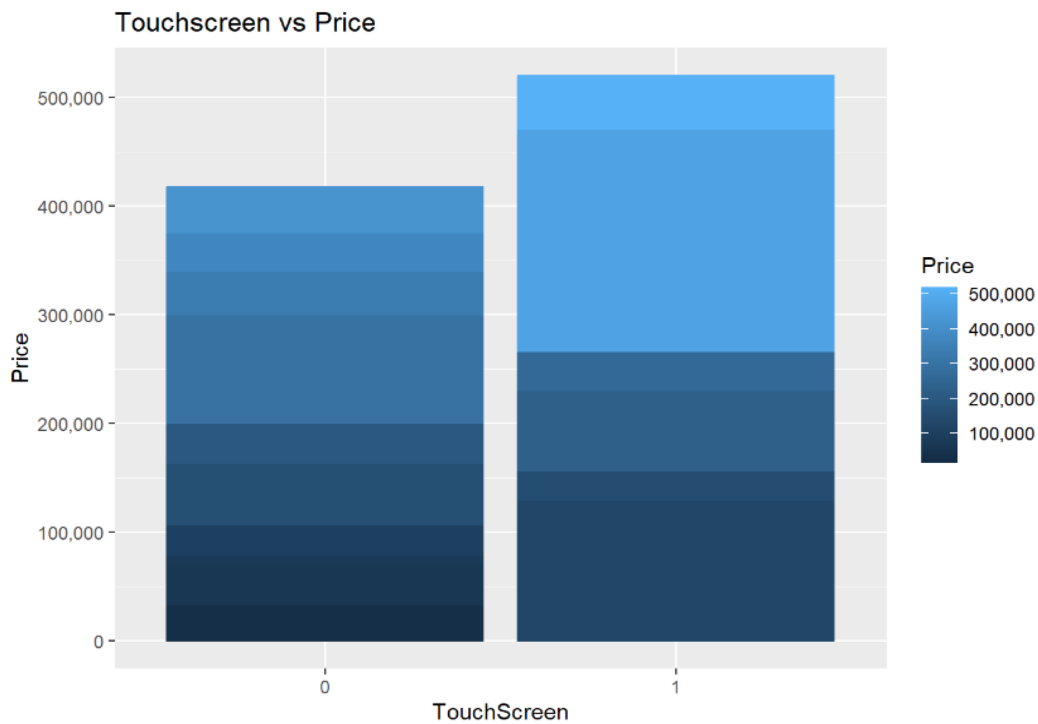

We can clearly see that Gaming laptops are the most expensive and the Netbook are the inexpensive ones. This makes sense because the gaming laptops are loaded with high-end specs whereas netbooks are designed only to surf the internet primarily.

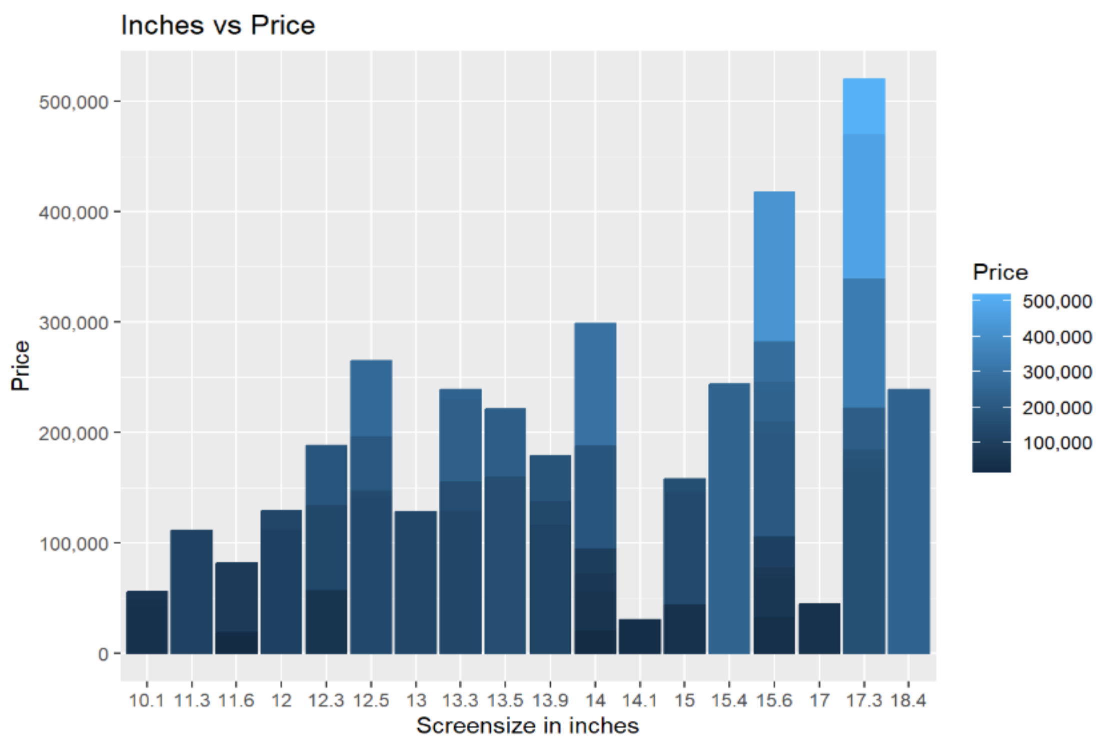



There are a lot of companies that manufacture and sell laptops on the market. The price of a laptop may be influenced by the brand value and the trust of the consumers towards the brand.

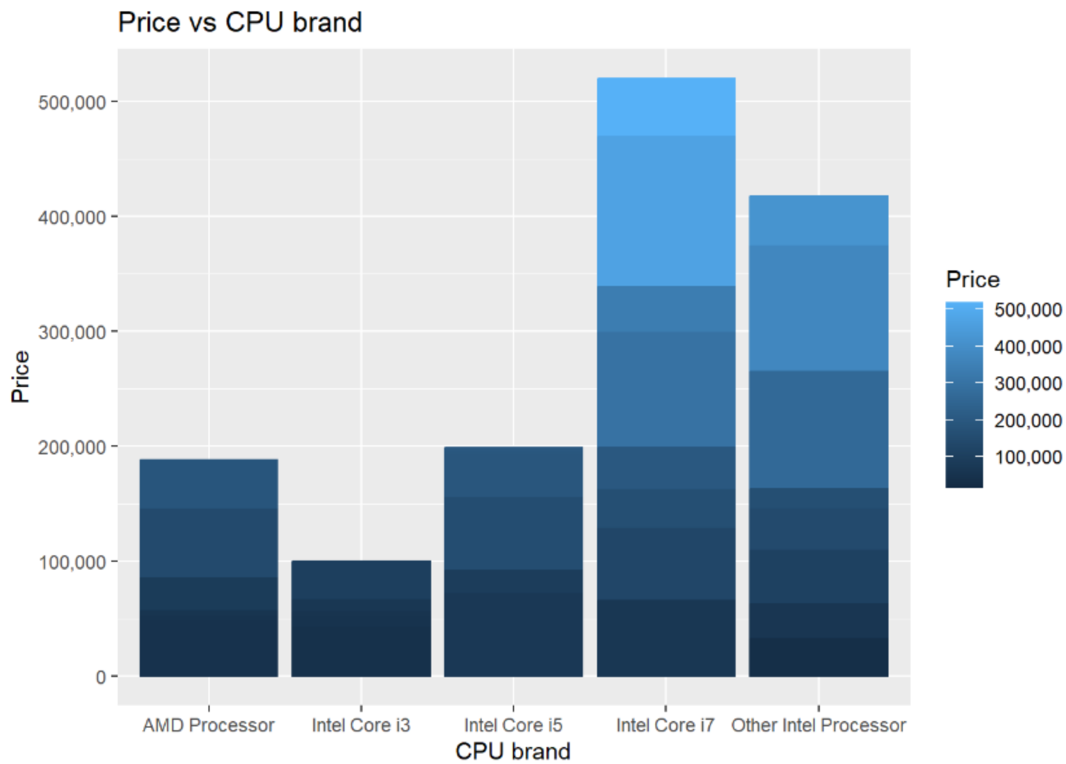
The bar chart helps us know which company has the most expensive laptops and in our case it happens to be the brand ‘Razor’, due to its core market in Gaming Laptops.

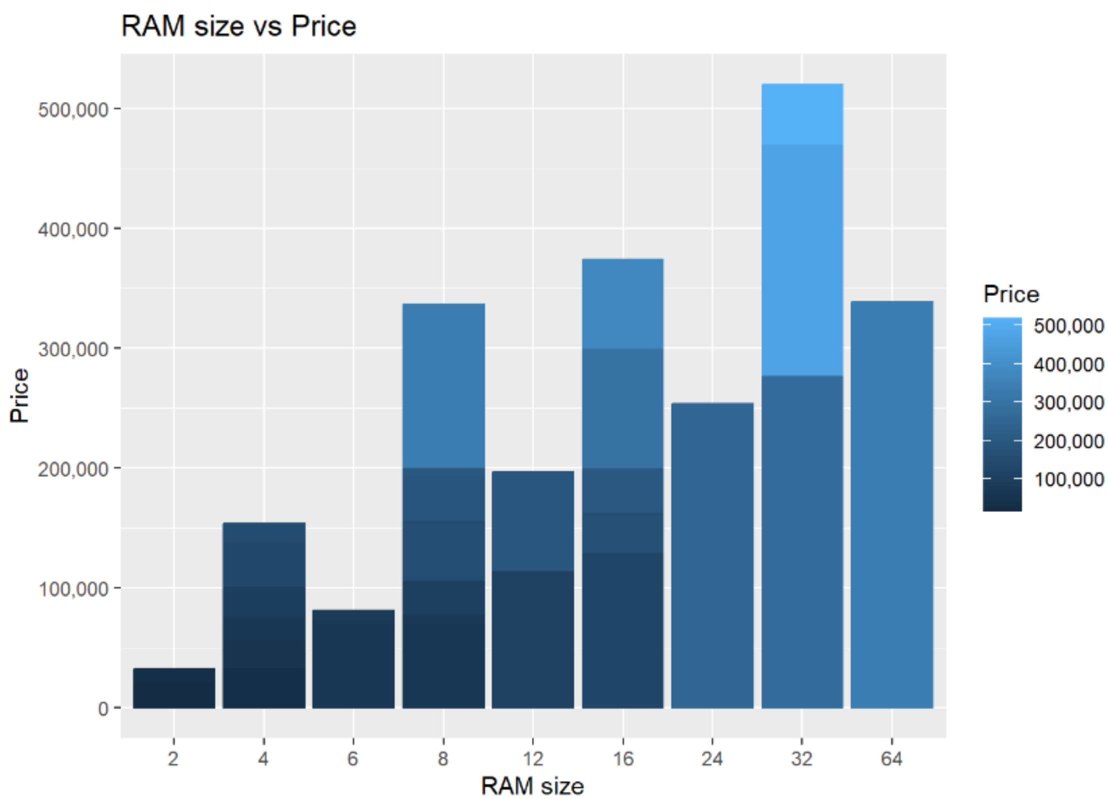
After transforming the ‘Screen Resolution’ column to other features, we were able to look at the correlation between the features and the price of the laptops.



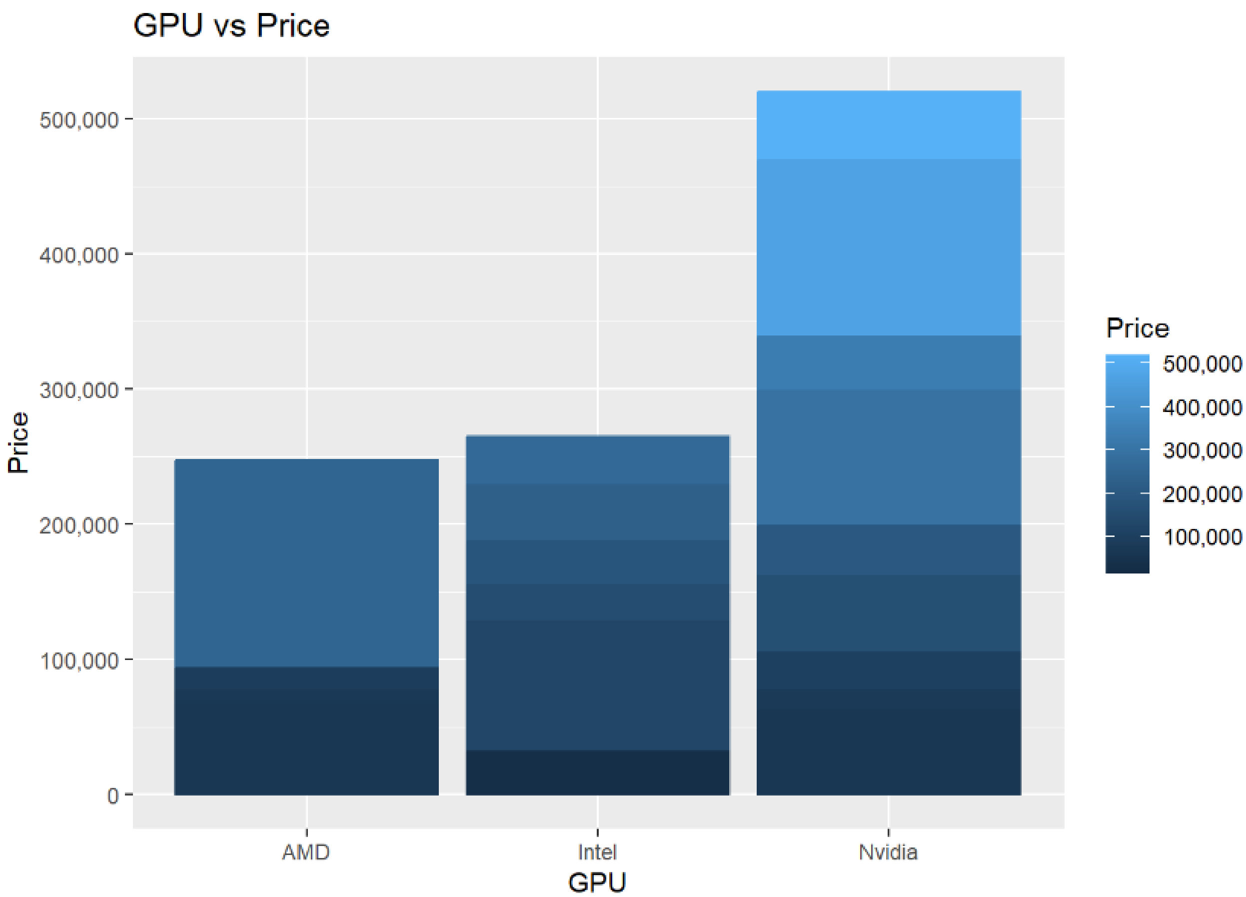

This helped us know that touchscreen laptops are slightly more expensive than non-touchscreen laptops and as the laptop size increases, price also goes up.




RAM and CPU plays a great role in determining the price of the laptop.

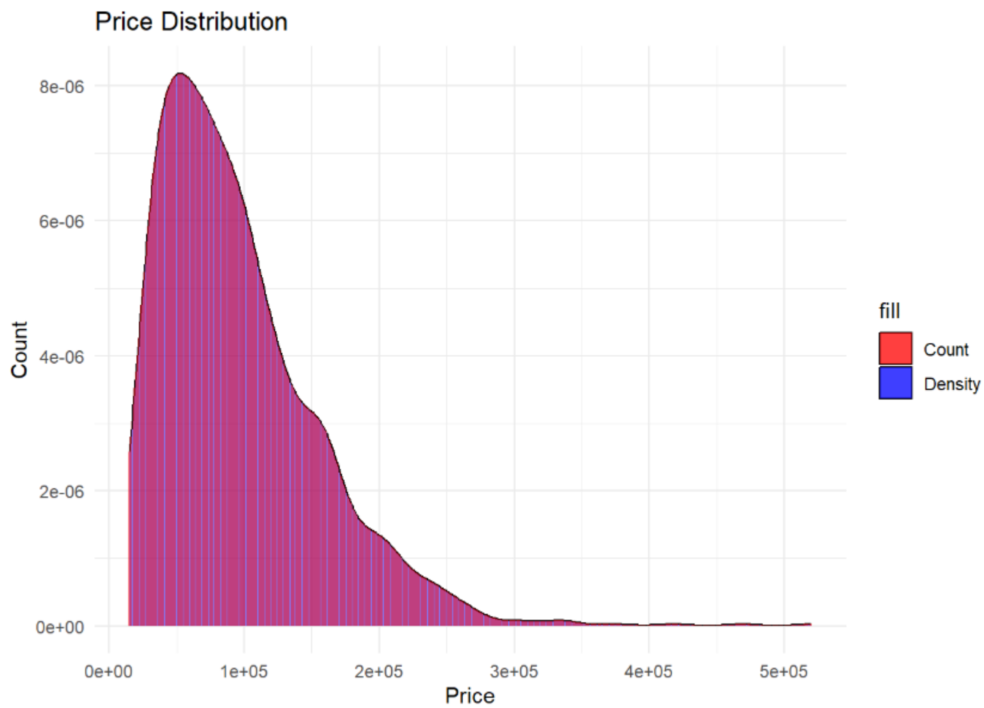
Lets visualize how our laptop price changes based on the GPU specification of the laptop classified by the brand name.

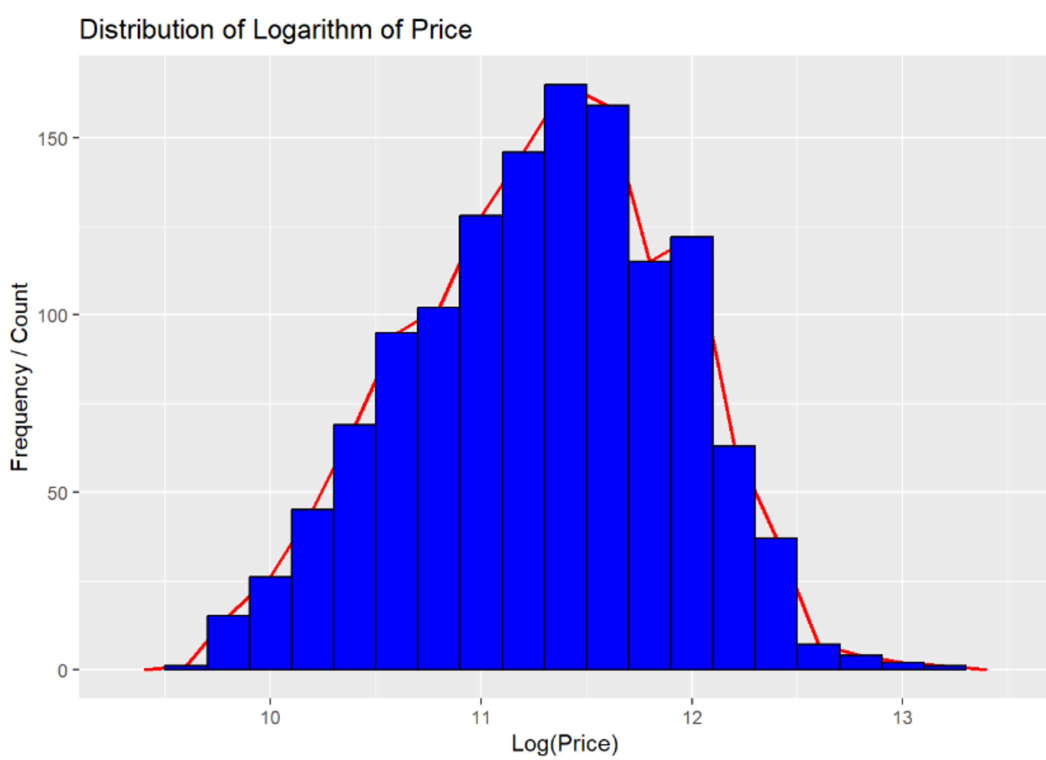



Laptops having Nvidia graphics cards tend to be highly expensive in the market. These graphics cards are commonly found on gaming laptops, which was deduced to be the most expensive among other types of laptops. Intel and AMD integrated graphics are most found on notebooks and ultrabooks, which are less expensive in the market.

# Normalization

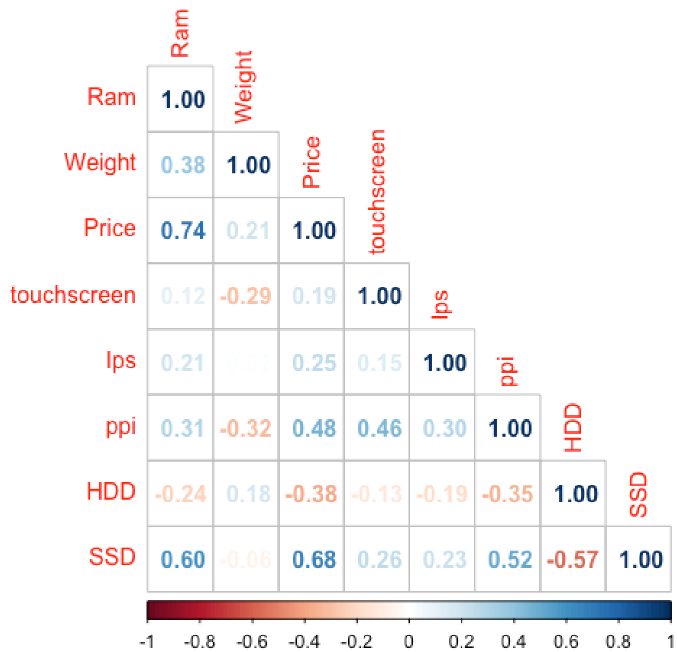
Log transformation helps in a way to normalize our data so that the model training is effective.



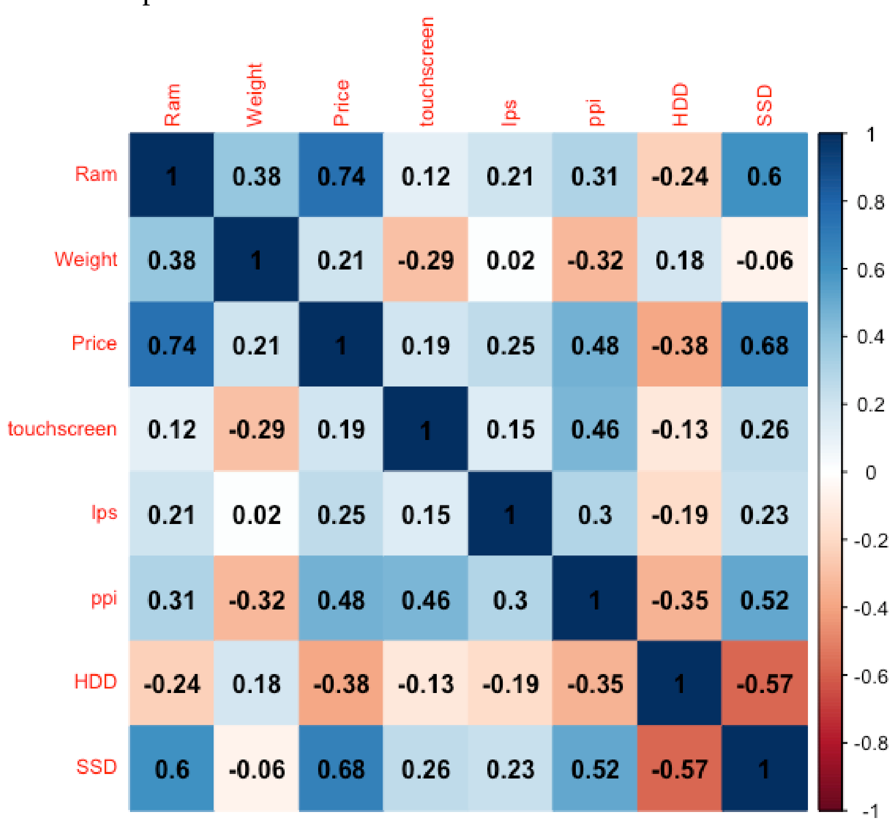



# Correlation Analysis

Correlation Matrix




Correlation HeatMap



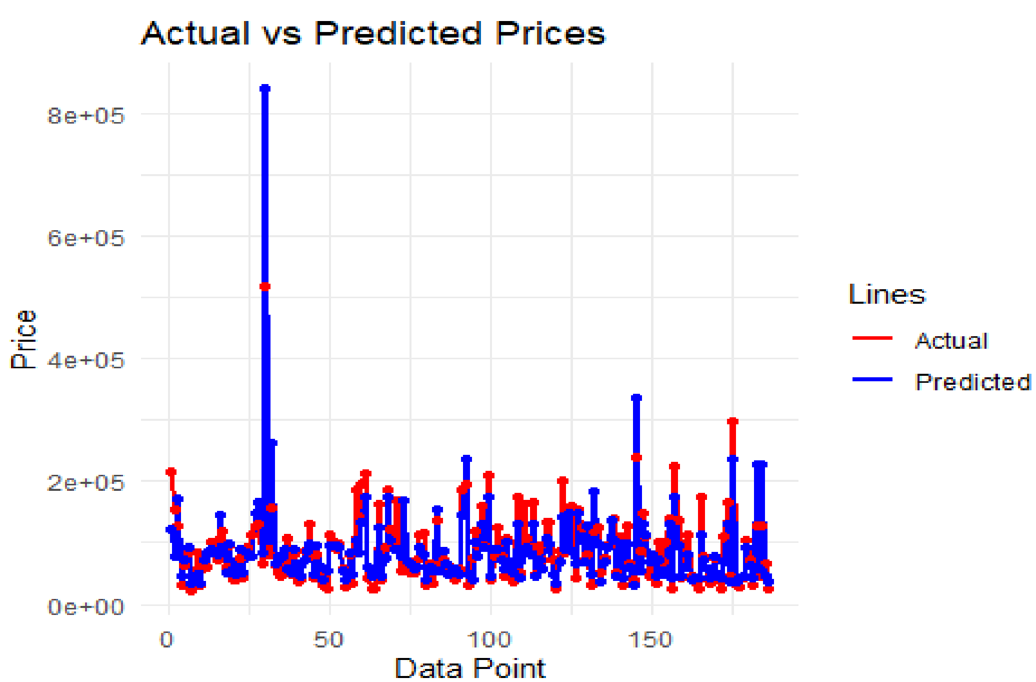
## Model Performance

### Linear Regression

```
Residual standard error: 0.3168 on 1039 degrees of freedom
Multiple R-squared:  0.7027, Adjusted R-squared:  0.6993
F-statistic: 204.7 on 12 and 1039 DF,  p-value: < 2.2e-16
```

Result of applying the model to the dataset

```
##       predicted       True
## 4     122327.03 216312.54
## 5      78473.61 153753.29
## 8     172447.70 127445.76
## 11     45891.78  29409.71
## 16     72465.59  63509.76
## 20     92636.32  85162.75
## 21     35188.39  21993.98
## 24     43286.19  35688.22
## 31     47189.73  85077.50
## 32     34368.07  31286.02
## 50     72303.86  65453.41
## 59     84367.65  58821.12
## 87     88353.35 101871.36
## 88     85125.72  89425.15
```



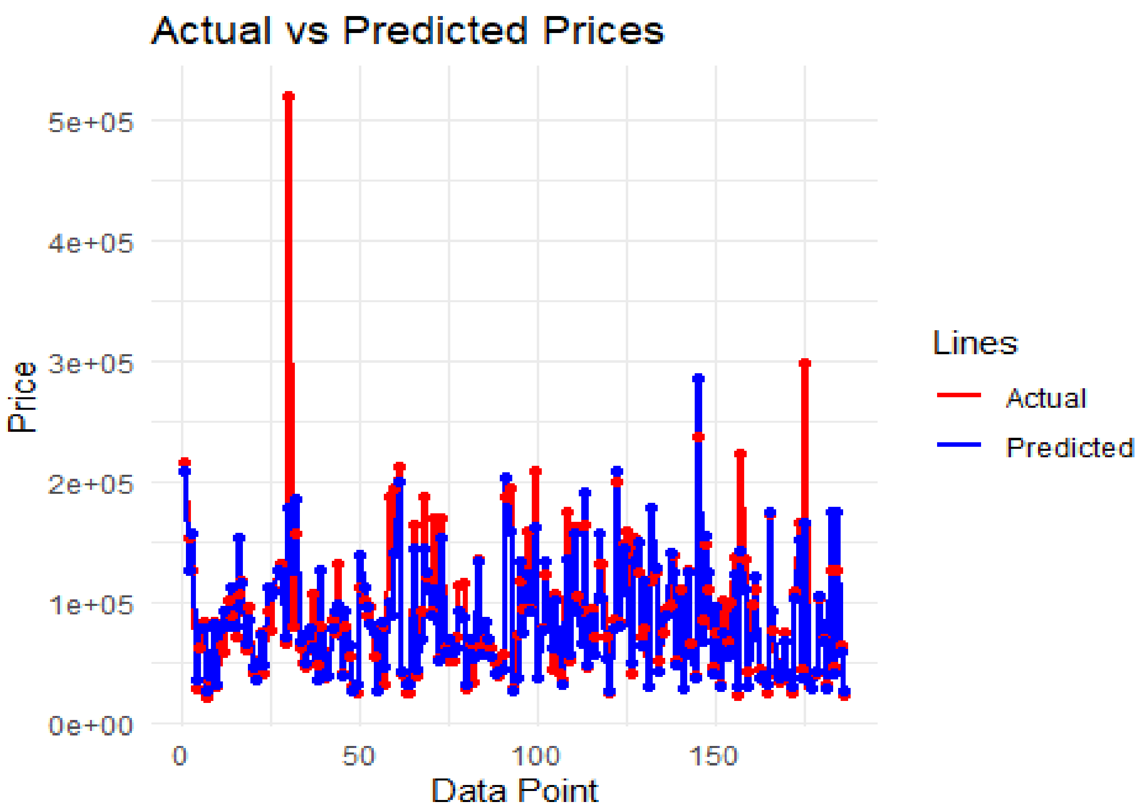
## Support Vector Regressor

```
## [1] "R2 score: 0.86142574859494"
```

```
## [1] "MAE: 0.172006651734998"
```

Applying the model to 10 random records of our dataset

```
##      predicted      True
## 4    209937.19 216312.54
## 5    128139.30 153753.29
## 8    158456.41 127445.76
## 11    35985.46  29409.71
## 16    79593.33  63509.76
## 20    78590.49  85162.75
## 21    28069.53  21993.98
## 24    39803.77  35688.22
## 31    82688.90  85077.50
## 32    33185.27  31286.02
```



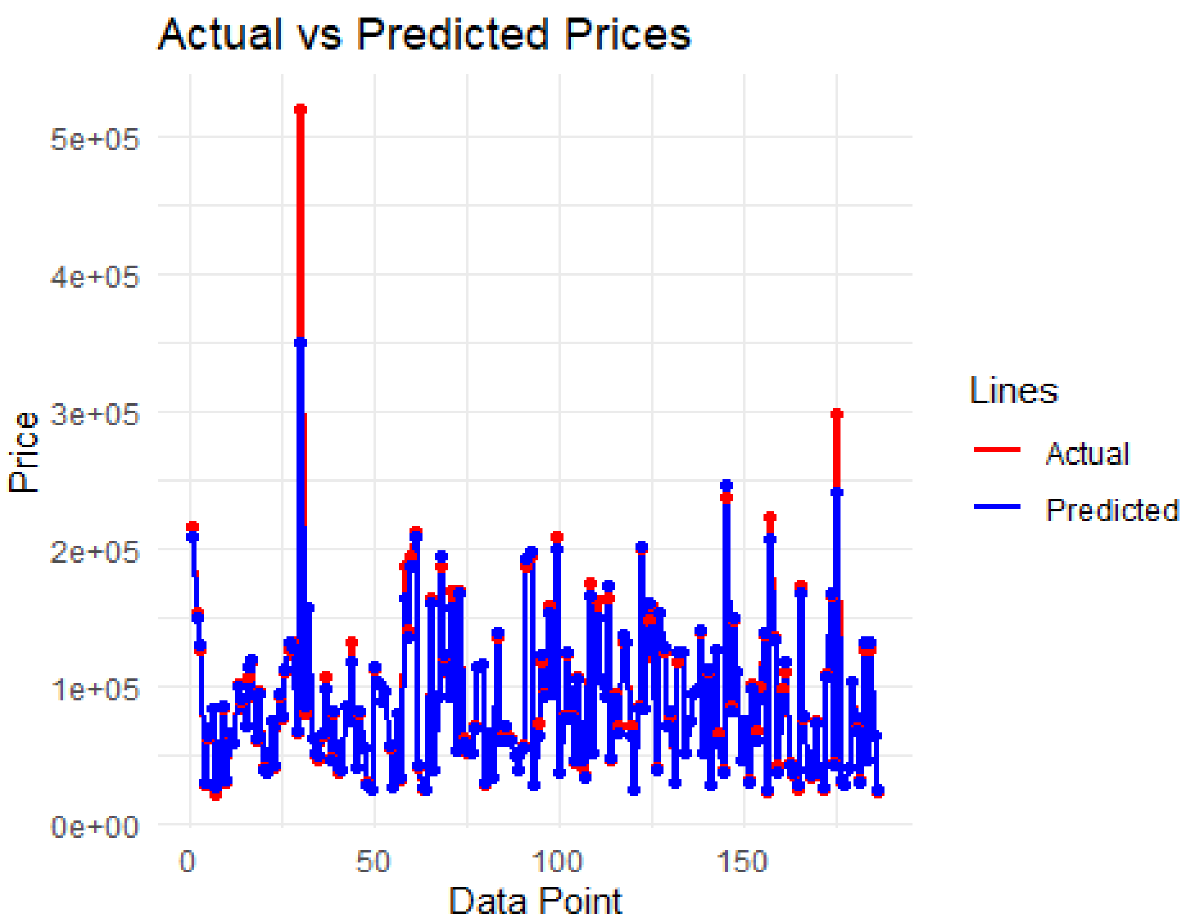
## Random Forrest

```
## [1] "R2 score: 0.992549931979812"
```

```
## [1] "MAE Score: 0.0325191100620161"
```

Applying the model to 10 random datapoints in our dataset

| | Predicted | True |
|---|---|---|
| ## 190 | 85475.23 | 81752.83 |
| ## 193 | 157804.19 | 157282.56 |
| ## 195 | 62837.94 | 63254.02 |
| ## 202 | 52141.16 | 50893.06 |
| ## 206 | 50010.37 | 46801.15 |
| ## 220 | 66479.50 | 65555.71 |
| ## 222 | 98263.32 | 108179.71 |
| ## 230 | 47746.56 | 49017.60 |
| ## 248 | 83094.32 | 80900.35 |
| ## 249 | 41678.73 | 38012.08 |



## Model Evaluation

| Models | R2 Score |
|---|---|
| Linear Regression | 70% |
| Support Vector Regressor | 86% |
| Random Forrest | 99% |

The results of this page is an upgrade to the results of the paper

# Limitations and Future study

Predicting laptop prices can be a challenging task due to various factors that influence the pricing of laptops.The huge amount of consumer-customizable setups offered poses a serious difficulty. Models find it challenging to predict costs for personalized computers with any degree of accuracy since these customizations might result in distinctive pricing patterns. Additionally, as older models try to keep up with these quick changes, the laptop market's fast-paced nature characterized by regular product introductions and price changes can render them out of date. Further study should investigate improved data sources, real-time data integration, advanced modeling techniques, and the inclusion of unstructured data, such as user reviews and preferences, to improve the accuracy as well as durability of laptop cost forecasting models.

Several scenarios can lead to poor performance of the prediction models:

1. **Rapid technological advancements**:
   The rapid rate of development in the technology sector is well-known. Models trained on historical data may find it challenging to anticipate pricing of laptops with these new characteristics due to the constant introduction of new components and functions. This is due to the model's inability to effectively predict how these additional factors would affect the pricing because it was not trained on datasets that included them.

2. **Insufficient Data**:
   To train on and create reliable predictions, machine learning models need a lot of data. The model may not be able produce accurate forecasts if there is not enough data, especially for specific laptop kinds or feature combinations. When there is a paucity of data,

overfitting may occur, where the predicted result performs well on training data but badly on new, unexpected information.

3. **<u>High Price Variability</u>:**
It can be challenging for models to produce reliable forecasts if laptop prices vary widely. This could be the result of a number of things, such as different pricing in different geographic areas, prices that change because of offers or discounts, and price variations across stores. Inaccurate forecasts may result from models' inability to include these variances.

4. **<u>Lack of Domain Knowledge</u>:**
A number of technological preferences, like the kind of CPU, the quantity of memory, and the size and quality of the display, affect laptop pricing. Without a knowledge of these technical details and how they affect the cost, models could have trouble providing accurate predictions.

5. **<u>External variables can significantly affect laptop costs</u>:**
Examples include changes in the economy, trade regulations, or supply chain problems. These variables are frequently challenging to forecast and include  models, which can lead to incorrect price predictions.

These limitations offer opportunities for future development as well as unresolved issues that must be solved with::

One strategy for addressing short-term price swings is to include analysis of time series in the prediction model. This might aid the model's capacity to learn and forecast rapid changes in pricing.Engineering features is  Improved feature engineering could handle complicated features and high volatility. For instance, Principal Component Analysis (PCA) dimensionality reduction methods could be used to streamline high-variance features. More advanced encoding strategies for complex features could be investigated. Data enrichment may be a solution if the data set lacks pertinent features. This can entail collecting extra information with the missing features. For instance, data that other sources that provide Processor information could be incorporated into the dataset if CPU information is lacking.

# Conclusion

In conclusion, predicting laptop prices is a difficult  effort that requires examining and understanding a wide range of variables, from technical details to market patterns. Despite the difficulties, it's a useful tool for both customers and companies. While corporations can use it for competitive pricing and strategic planning, it can also assist consumers in making educated purchasing decisions.

The shortcomings of current prediction models include their inability to account for seasonal trends, large volatility in features, complicated features, and short-term price fluctuations. However, these difficulties also offer chances for additional study and development. These restrictions can be overcome using methods including time series analysis, enhanced engineering of features, enriching the data, and seasonal decomposition.

Additionally, the investigation of machine learning algorithms for laptop estimation of prices is an exciting field of study for the future. These models could perhaps increase the precision of predictions by identifying intricate links in the data.

The techniques for forecasting laptop costs will ultimately change as the technological market does. It's a field of research that will only continue to become more significant and advanced.

# References


https://www.statista.com/statistics/1235280/india-average-price-of-laptops/

A. D. Siburian, “Laptop Price Prediction with Machine Learning Using Regression Algorithm”, *JUSIKOM PRIMA*, vol. 6, no. 1, pp. 87-91, Sep. 2022.

Sun, S. (2022) *India: Average price of laptops 2021*, *Statista*. Available at: https://www.statista.com/statistics/1235280/india-average-price-of-laptops/ (Accessed: 04 October 2023).

*Laptop Prediction & Comparison Using Machine Learning - ijirt*. Available at: https://ijirt.org/master/publishedpaper/IJIRT158273_PAPER.pdf (Accessed: 28 September 2023).

Ijraset *IJRASET Journal for Research in Applied Science and Engineering Technology*, *Laptop Performance Prediction*. Available at: https://www.ijraset.com/research-paper/laptop-performance-prediction (Accessed: 01 October 2023).

Bhattacharya, P.S. (2021) *Laptop price prediction by machine learning*, *Medium*. Available at: https://medium.com/analytics-vidhya/laptop-price-prediction-by-machine-learning-7e1211bb96d1 (Accessed: 04 October 2023).

*Laptop price prediction system* (no date) *PPT*. Available at: https://www.slideshare.net/MDRIAZHASAN/laptop-price-prediction-system (Accessed: 04 October 2023).