Exploratory Data Analysis Of laptops dataset

Usama Ali

2023-09-30

Objective

- The objective of this project was to perform an exploratory data analysis of a laptop dataset using R and ggplot2 for data visualization. The analysis aimed to extract valuable insights from the dataset and provide clear visualizations to enhance understanding.

Dataset used:

- About: This dataset provides a comprehensive collection of information on various laptops, enabling a detailed analysis of their specifications and pricing. It encompasses a wide range of laptops, encompassing diverse brands, models, and configurations, making it a valuable resource for researchers, data analysts, and machine learning enthusiasts interested in the laptop industry.
- Source: Kaggle.com
- License: CCO Public Domain
- Uploaded By: Juan Merino

```
library(tidyverse)
library(skimr)
library(markdown)
```

Importing the libraries

```
laptops <- read.csv('laptops.csv')
```

Importing Dataset

Checking for duplicates & removing

```
#viewing the first rows and structure
glimpse(laptops)
```

Taking a look and cleaning

```
## Rows: 2,160
## Columns: 12
## $ Laptop       <chr> "ASUS ExpertBook B1 B1502CBA-EJ0436X Intel Core i5-1235U/~
## $ Status       <chr> "New", "New", "New", "New", "New", "New", "New", "New", "~
## $ Brand        <chr> "Asus", "Alurin", "Asus", "MSI", "HP", "MSI", "Lenovo", "~
## $ Model        <chr> "ExpertBook", "Go", "ExpertBook", "Katana", "15S", "Cross~
## $ CPU          <chr> "Intel Core i5", "Intel Celeron", "Intel Core i3", "Intel~
## $ RAM          <int> 8, 8, 8, 16, 16, 32, 8, 8, 8, 16, 8, 16, 16, 16, 8, 8, 16~
## $ Storage      <int> 512, 256, 256, 1000, 512, 1000, 256, 512, 256, 512, 256, ~
## $ Storage.type <chr> "SSD", "SSD", "SSD", "SSD", "SSD", "SSD", "SSD", "SSD", "~
## $ GPU          <chr> "", "", "", "RTX 3050", "", "RTX 4060", "", "", "", "RTX ~
## $ Screen       <dbl> 15.6, 15.6, 15.6, 15.6, 15.6, 17.3, 14.0, 15.6, 15.6, 16.~
## $ Touch        <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No~
## $ Final.Price  <dbl> 1009.00, 299.00, 789.00, 1199.00, 669.01, 1699.00, 909.00~
```

```
#basic descriptive stats for our dataset & Na values
skim_without_charts(laptops)
```

Table 1: Data summary

Name	laptops
Number of rows	2160
Number of columns	12
Column type frequency:	
character	8
numeric	4
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Laptop	0	1	36	129	0	2160	0
Status	0	1	3	11	0	2	0
Brand	0	1	2	16	0	27	0
Model	0	1	2	14	0	121	0
CPU	0	1	8	21	0	28	0
Storage.type	0	1	0	4	42	3	0
GPU	0	1	0	18	1371	45	0
Touch	0	1	2	3	0	2	0

Variable type: numeric

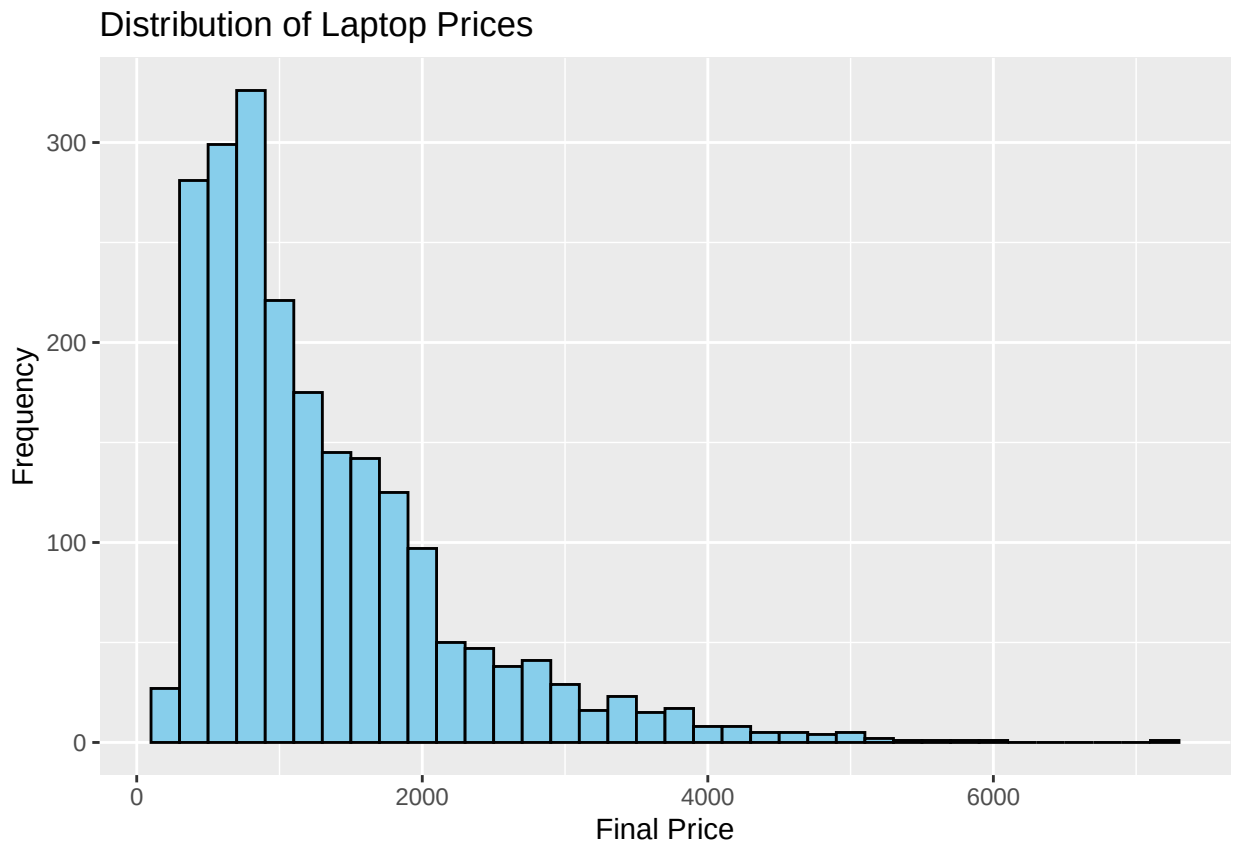
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
RAM	0	1	15.41	9.87	4.00	8.00	16.00	16.00	128.00
Storage	0	1	596.29	361.22	0.00	256.00	512.00	1000.00	4000.00
Screen	4	1	15.17	1.20	10.10	14.00	15.60	15.60	18.00
Final.Price	0	1	1312.64	911.48	201.05	661.08	1031.95	1708.97	7150.47

```
#removing missing values
laptops <- na.omit(laptops)
```

Plot Analysis

Price Distribution:

```
#histogram for price distribution
ggplot(laptops, aes(x = Final.Price)) +
  geom_histogram(binwidth = 200, fill="skyblue", color = 'black')+
  labs(title = 'Distribution of Laptop Prices',
       x = 'Final Price', y = 'Frequency')
```

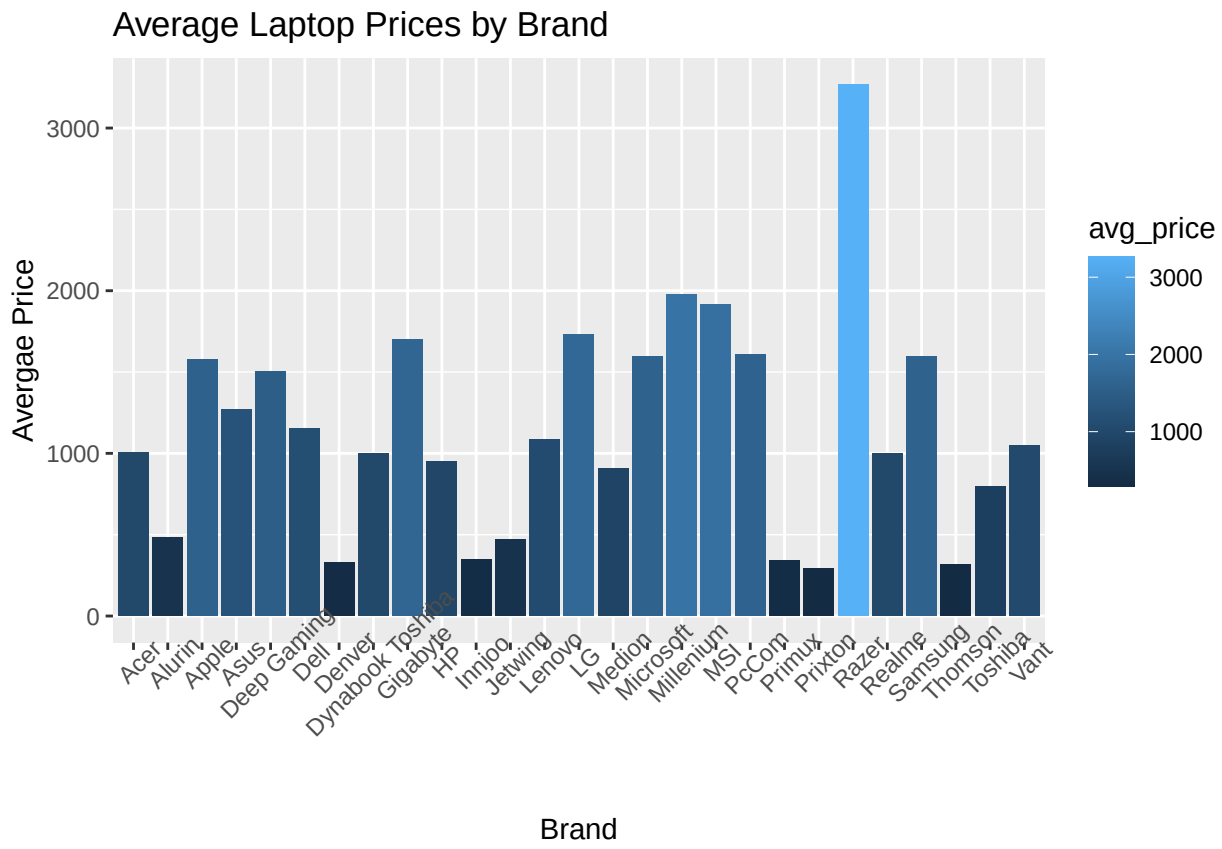


- The histogram displays a right-skewed distribution. Three prominent peaks are observed, with frequencies exceeding 350 for one peak.

- As prices increase beyond $300, the frequency gradually declines until the $5000 price range. Between $5000 and $6000, the distribution becomes less visible. Laptops priced above $6000 are exceedingly rare in the dataset.

Brand Analysis - Average Price:

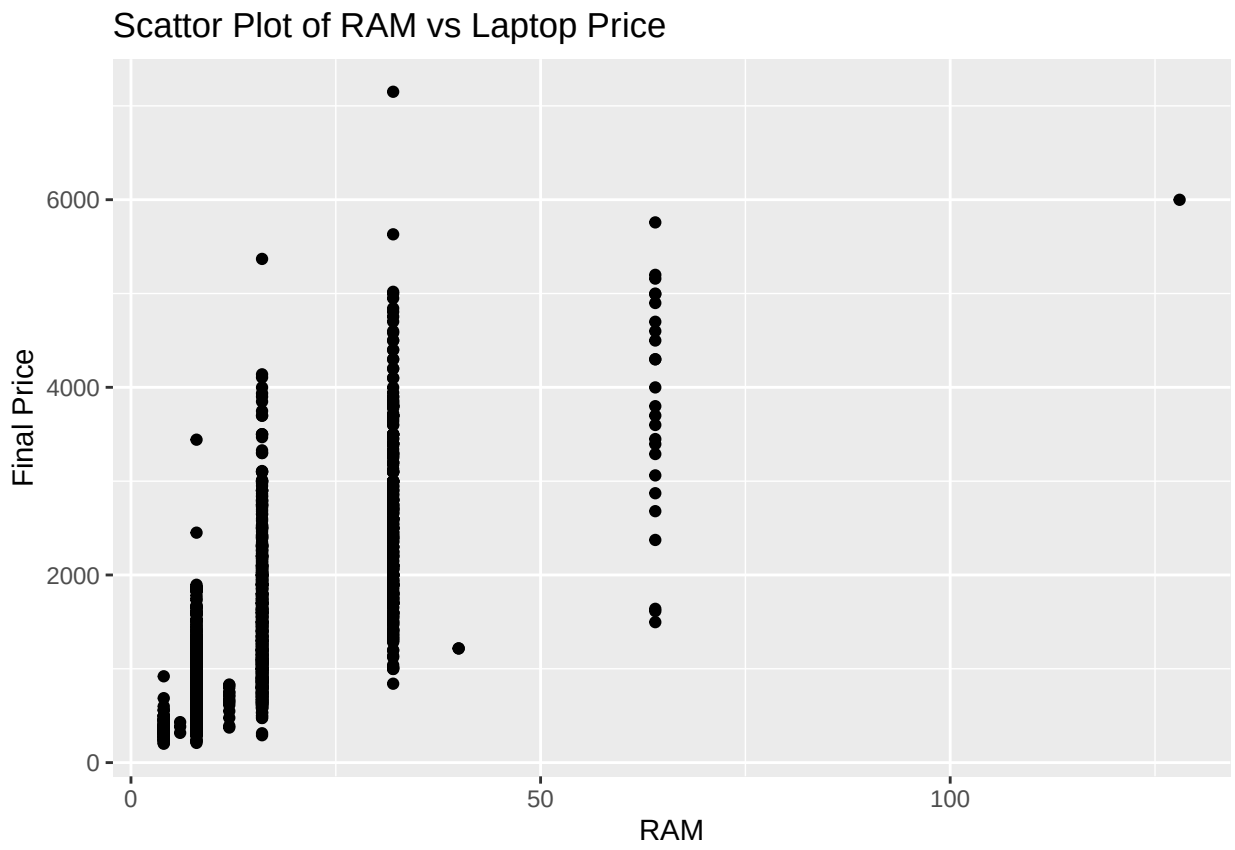
```
#brand analysis, avg_price for each brand bar chart
brand_avg_prices <- laptops %>%
  group_by(Brand) %>%
  summarise(avg_price = mean(Final.Price))
ggplot(brand_avg_prices, aes(x = Brand, y = avg_price,
                             fill = avg_price))+
  geom_bar(stat = 'identity') +
  labs(x = 'Brand', y = 'Avergae Price',
       title = 'Average Laptop Prices by Brand') +
  theme(axis.text.x = element_text(angle = 45))
```



- The bar chart presents the average prices of laptops by brand.
- Razer stands out as the highest-priced brand, with laptops peaking at around $3500 in average price.
- Millenium and MSI are the second and third highest-priced brands, with laptops sitting near the $2000 price mark on average.
- Popular brands like Apple, Microsoft, and Gigabyte have laptops priced at nearly $1500 on average.

RAM & Price Relationship:

```
#RAM & Price relationship via scatter plot
ggplot(laptops, aes(x = RAM, y = Final.Price))+
  geom_point()+
  labs(title = 'Scattor Plot of RAM vs Laptop Price',
       x = "RAM", y = 'Final Price')
```

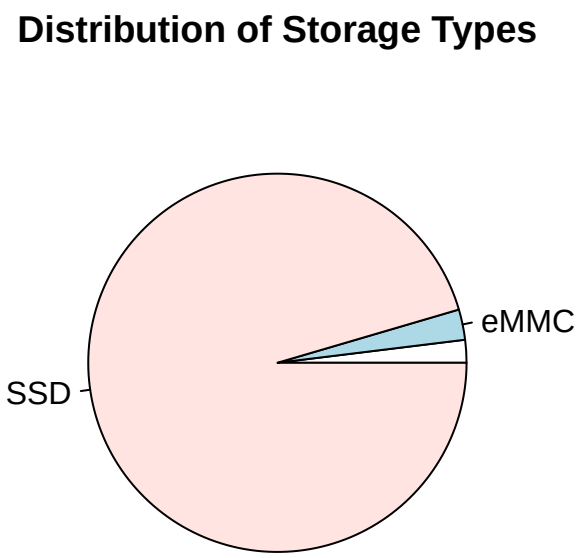


- The scatter plot reveals the relationship between laptop RAM (in GB) and prices.
- For laptops with 4GB RAM, a cluster is observed in the $200-$800 price range.Laptops with 8GB RAM cluster between $200 and $2000, with fewer above $2000. The 16GB RAM category is dense, spanning $200-$4000, with fewer observations beyond $3000.

-64GB RAM laptops are sparsely represented. An isolated 120GB RAM laptop is priced around $6000, suggesting uniqueness.

Storage Type Distribution:

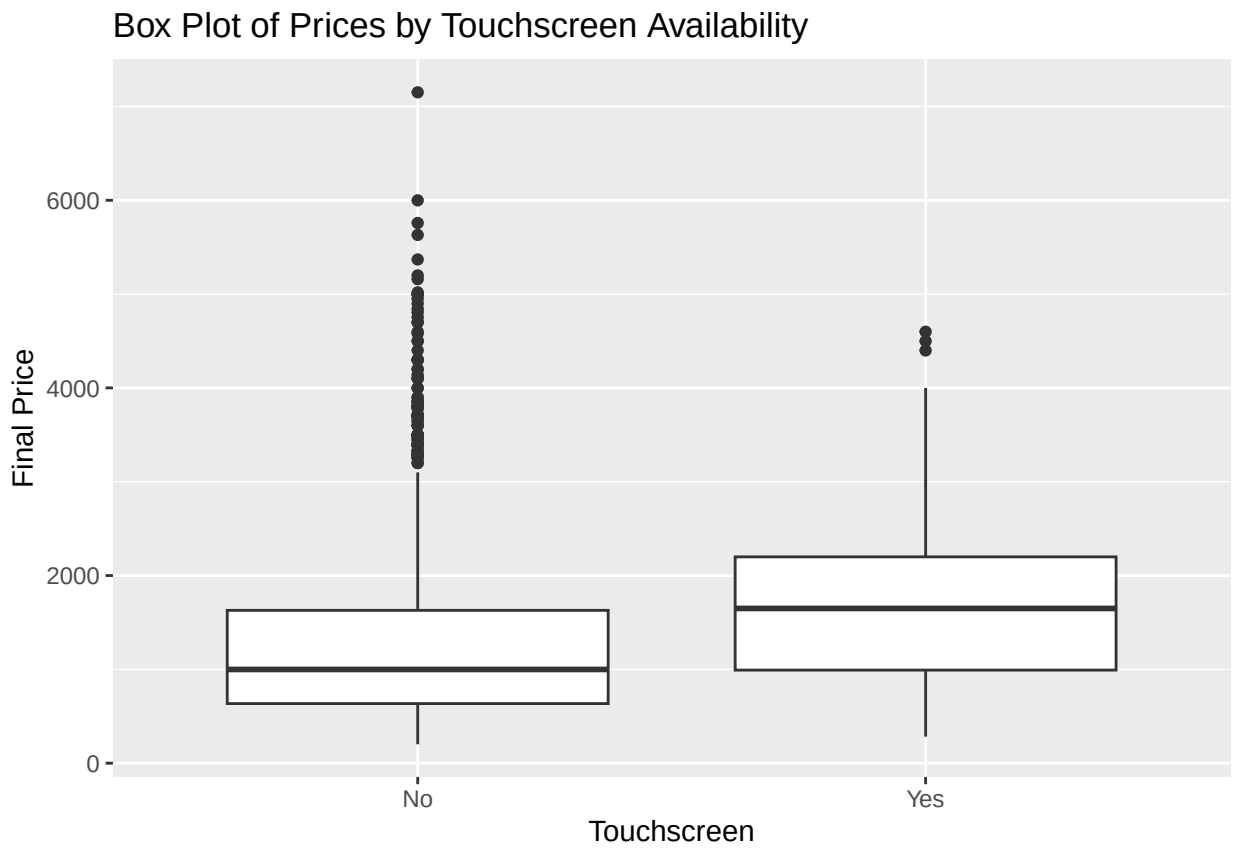
```
#distribution of storage types using pie chart
storage_distribution <- table(laptops$Storage.type)
pie(storage_distribution, labels = names(storage_distribution),
    main = "Distribution of Storage Types")
```



- The pie chart illustrates the predominance of SSD storage types among laptops.
- SSD storage types represent the majority of laptops in the dataset, while eMMC storage types constitute a very small proportion.

Touchscreen vs. Non-Touchscreen Prices:

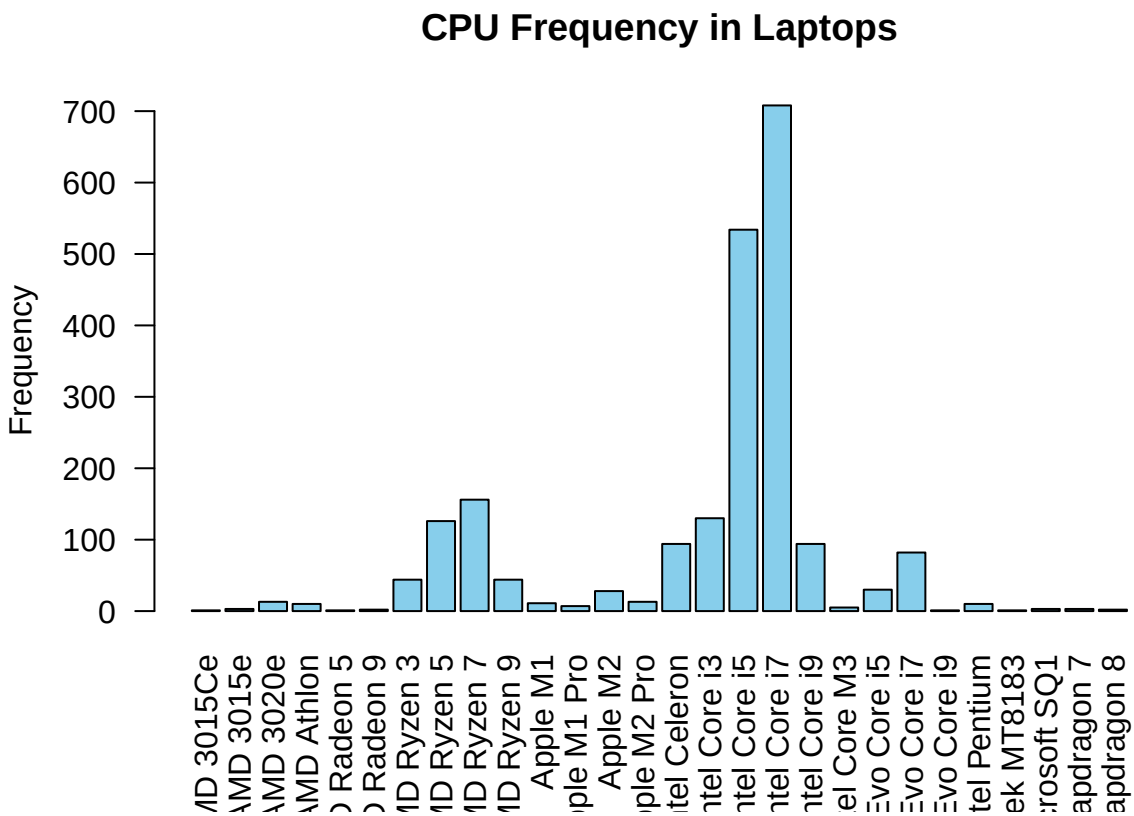
```
#5.box-plot for avg prices of touchscreen vs non-touchscreen
ggplot(laptops, aes(x = Touch, y = Final.Price))+
  geom_boxplot()+
  labs(x = 'Touchscreen', y = 'Final Price',
       title = 'Box Plot of Prices by Touchscreen Availability')
```



- The box plot compares laptop prices between two categories: "Yes" (touchscreen) and "No" (non-touchscreen).
- The median for the "No" category sits at approximately $1000, while for the "Yes" category, it's slightly above $1500.
- In the "No" category (non-touchscreen), the majority of data points are clustered above the 3rd quartile, ranging from approximately $3000 to $6000. There is a notable outlier priced above $7000.
- The median is positioned between the 1st and 2nd quartiles but is also closer to the center.
- In the "Yes" category (touchscreen), the pattern is similar, with most data points concentrated above the 3rd quartile, at around $4500.
- The median leans toward the 3rd quartile rather than the 1st quartile, indicating differences in price distribution between the two categories.

CPU Analysis - Most Common CPU

```
#CPU analysis, most common CPU barplot
cpu_freq <- table(laptops$CPU)
common_cpu <- names(which.max(cpu_freq))
barplot(cpu_freq,
        main = 'CPU Frequency in Laptops',
        ylab = "Frequency", col = 'skyblue',
        las= 2)
```



- The bar chart displays the frequency of different CPU types among laptops.
- Intel Core i7 CPUs are the most prevalent, with a frequency of approximately 700. Intel Core i5 CPUs follow as the second most common, with a frequency of around 500.
- AMD Ryzen 7 CPUs are the third highest in frequency, at approximately 150.
- Ryzen 5 CPUs are the fourth highest in frequency.

Recommendations:

- Consider exploring the unique features and specifications of laptops from brands like Razer, Millenium, and MSI, as they offer higher-priced options with potential performance benefits.
- For consumers seeking laptops within the $200 to $800 price range, focus on laptops with 4GB and 8GB RAM, which are prevalent and offer good value.
- Manufacturers should take note of the popularity of Intel Core i7 and Core i5 CPUs and potentially offer a wider range of laptops featuring these processors.