Project EDA

Somesh Bagadiya

2023-11-02

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
library(tinytex)
```

```
## Warning: package 'tinytex' was built under R version 4.3.2
```

INTRODUCTION:

I have a keen interest in the field of technology and gadgets, specifically laptops in this case. Working with this data will help me gain valuable knowledge regarding how laptops are categorized based on their specifications, how they are priced, which laptop is suitable for a particular user, which brand is more expensive, which brand is more affordable, and so on. The data comprises a combination of specifications and categorical information.

```
data <- read.csv("./Laptop_Data.csv")
head(data, 5)
```

```
##    brand   model processor_brand    processor_name processor_gnrtn ram_gb
## 1 Lenovo A6-9225             AMD A6-9225 Processor            10th      4
## 2 Lenovo Ideapad             AMD          APU Dual            10th      4
## 3  Avita    PURA             AMD          APU Dual            10th      4
## 4  Avita    PURA             AMD          APU Dual            10th      4
## 5  Avita    PURA             AMD          APU Dual            10th      4
##   ram_type ssd  hdd      os os_bit graphic_card_gb     weight display_size
## 1     DDR4   0 1024 Windows     64               0 ThinNlight         15.6
## 2     DDR4   0  512 Windows     64               0     Casual         15.6
## 3     DDR4 128    0 Windows     64               0 ThinNlight         15.6
## 4     DDR4 128    0 Windows     64               0 ThinNlight         15.6
## 5     DDR4 256    0 Windows     64               0 ThinNlight         15.6
##   warranty Touchscreen msoffice latest_price star_rating ratings reviews
## 1        0          No       No        24990         3.7      63      12
## 2        0          No       No        19590         3.6    1894     256
## 3        0          No       No        19990         3.7    1153     159
## 4        0          No       No        21490         3.7    1153     159
## 5        0          No       No        24990         3.7    1657     234
```

DATA:

Data Source:- This data is available on Kaggle, below is the link to the dataset https://www.kaggle.com/datasets/kuchhbhi/latest-laptop-price-list (https://www.kaggle.com/datasets/kuchhbhi/latest-laptop-price-list)

Data Collection:- The author scrapped the data from flipkart.com, they used an automated chrome web extension tool called Instant Data Scrapper to gather the data. It is a observational study.

Units of observation:- Each row represents a laptop along with its specifications, current price, reviews, and ratings.

No data cleanup required.

Exploratory Data Analysis:

```
summary(data)
```

```
##     brand              model           processor_brand    processor_name    
##  Length:896         Length:896         Length:896         Length:896        
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##  processor_gnrtn        ram_gb         ram_type              ssd        
##  Length:896         Min.   : 4.000   Length:896         Min.   :   0.0  
##  Class :character   1st Qu.: 4.000   Class :character   1st Qu.: 256.0  
##  Mode  :character   Median : 8.000   Mode  :character   Median : 512.0  
##                     Mean   : 8.531                      Mean   : 432.3  
##                     3rd Qu.: 8.000                      3rd Qu.: 512.0  
##                     Max.   :32.000                      Max.   :3072.0  
##       hdd              os                os_bit      graphic_card_gb
##  Min.   :   0.0   Length:896         Min.   :32.00   Min.   :0.000  
##  1st Qu.:   0.0   Class :character   1st Qu.:64.00   1st Qu.:0.000  
##  Median :   0.0   Mode  :character   Median :64.00   Median :0.000  
##  Mean   : 226.9                      Mean   :59.18   Mean   :1.199  
##  3rd Qu.: 512.0                      3rd Qu.:64.00   3rd Qu.:2.000  
##  Max.   :2048.0                      Max.   :64.00   Max.   :8.000  
##     weight           display_size     warranty     Touchscreen       
##  Length:896         Min.   :12.2   Min.   :0.000   Length:896        
##  Class :character   1st Qu.:15.6   1st Qu.:0.000   Class :character  
##  Mode  :character   Median :15.6   Median :1.000   Mode  :character  
##                     Mean   :15.3   Mean   :0.692                     
##                     3rd Qu.:15.6   3rd Qu.:1.000                     
##                     Max.   :17.3   Max.   :3.000                     
##    msoffice          latest_price     star_rating      ratings       
##  Length:896         Min.   : 13990   Min.   :0.00   Min.   :    0.0  
##  Class :character   1st Qu.: 45490   1st Qu.:0.00   1st Qu.:    0.0  
##  Mode  :character   Median : 63494   Median :4.10   Median :   19.0  
##                     Mean   : 76310   Mean   :2.98   Mean   :  367.4  
##                     3rd Qu.: 89090   3rd Qu.:4.40   3rd Qu.:  179.5  
##                     Max.   :441990   Max.   :5.00   Max.   :15279.0  
##     reviews       
##  Min.   :   0.00  
##  1st Qu.:   0.00  
##  Median :   3.00  
##  Mean   :  46.15  
##  3rd Qu.:  23.25  
##  Max.   :1947.00  
```

```
nrow(data)
```

```
## [1] 896
```

```
ncol(data)
```

```
## [1] 21
```

```
numeric_columns <- sapply(data, is.numeric)

# Print the results
print(numeric_columns)
```

```
##           brand           model processor_brand  processor_name processor_gnrtn
##           FALSE           FALSE           FALSE           FALSE           FALSE
##          ram_gb        ram_type             ssd             hdd              os
##            TRUE           FALSE            TRUE            TRUE           FALSE
##          os_bit graphic_card_gb          weight    display_size        warranty
##            TRUE            TRUE           FALSE            TRUE            TRUE
##     Touchscreen        msoffice    latest_price     star_rating         ratings
##           FALSE           FALSE            TRUE            TRUE            TRUE
##         reviews
##            TRUE
```

```
print(sum(numeric_columns))
```

```
## [1] 11
```

```
print(sum(is.na(data)))
```

```
## [1] 0
```

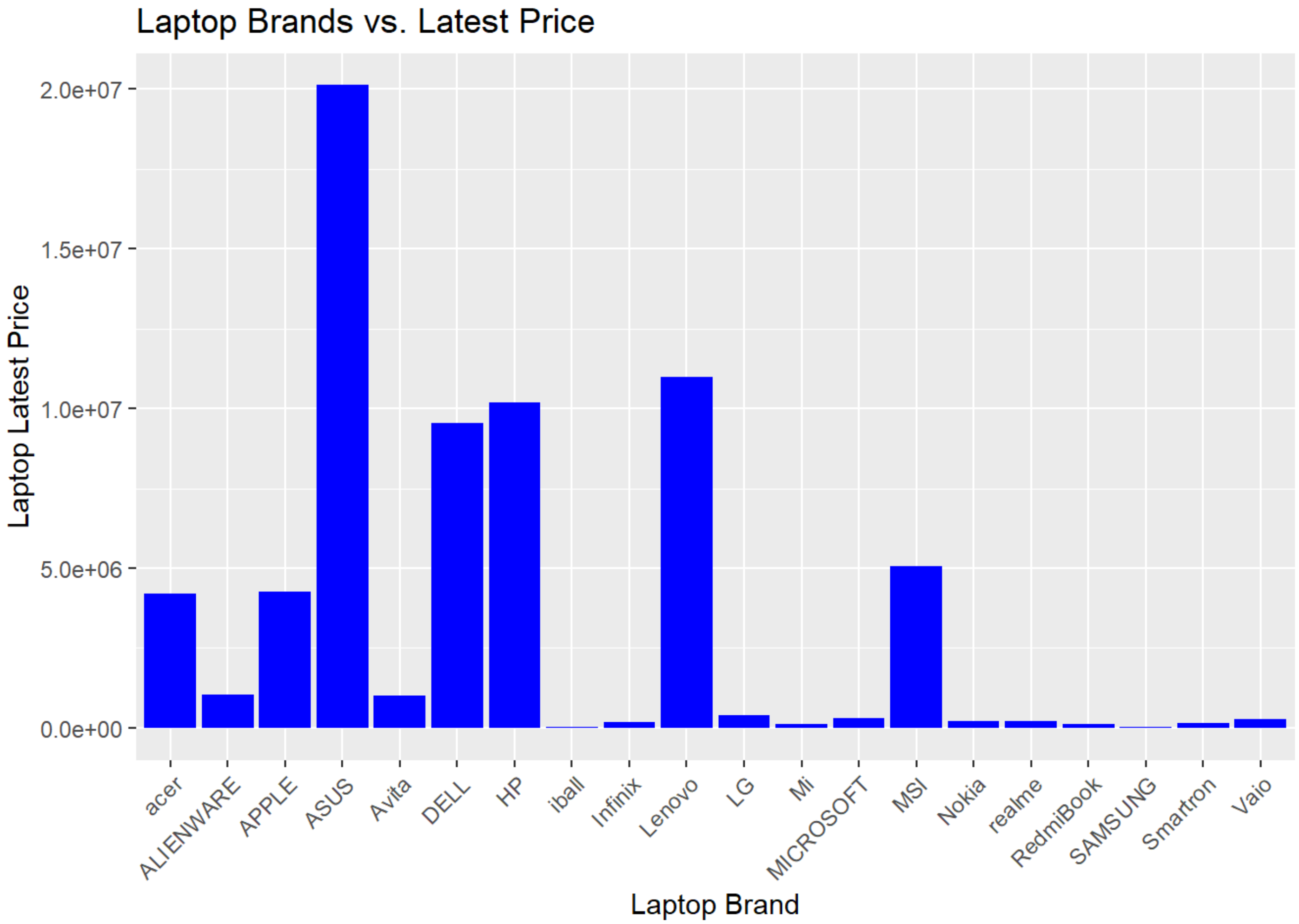
```
print(sum(duplicated(data)))
```

```
## [1] 21
```

Above is the summary of the data. There are a total of 896 rows. There are a total of 26 columns. There are 11 numeric columns out of 26. There are no null values in the data. There are 21 duplicate rows in the data.

```
p <- ggplot(data, aes(x = `brand`, y = `latest_price`)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(x = "Laptop Brand", y = "Laptop Latest Price") +
  ggtitle("Laptop Brands vs. Latest Price") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Display the plot
print(p)
```

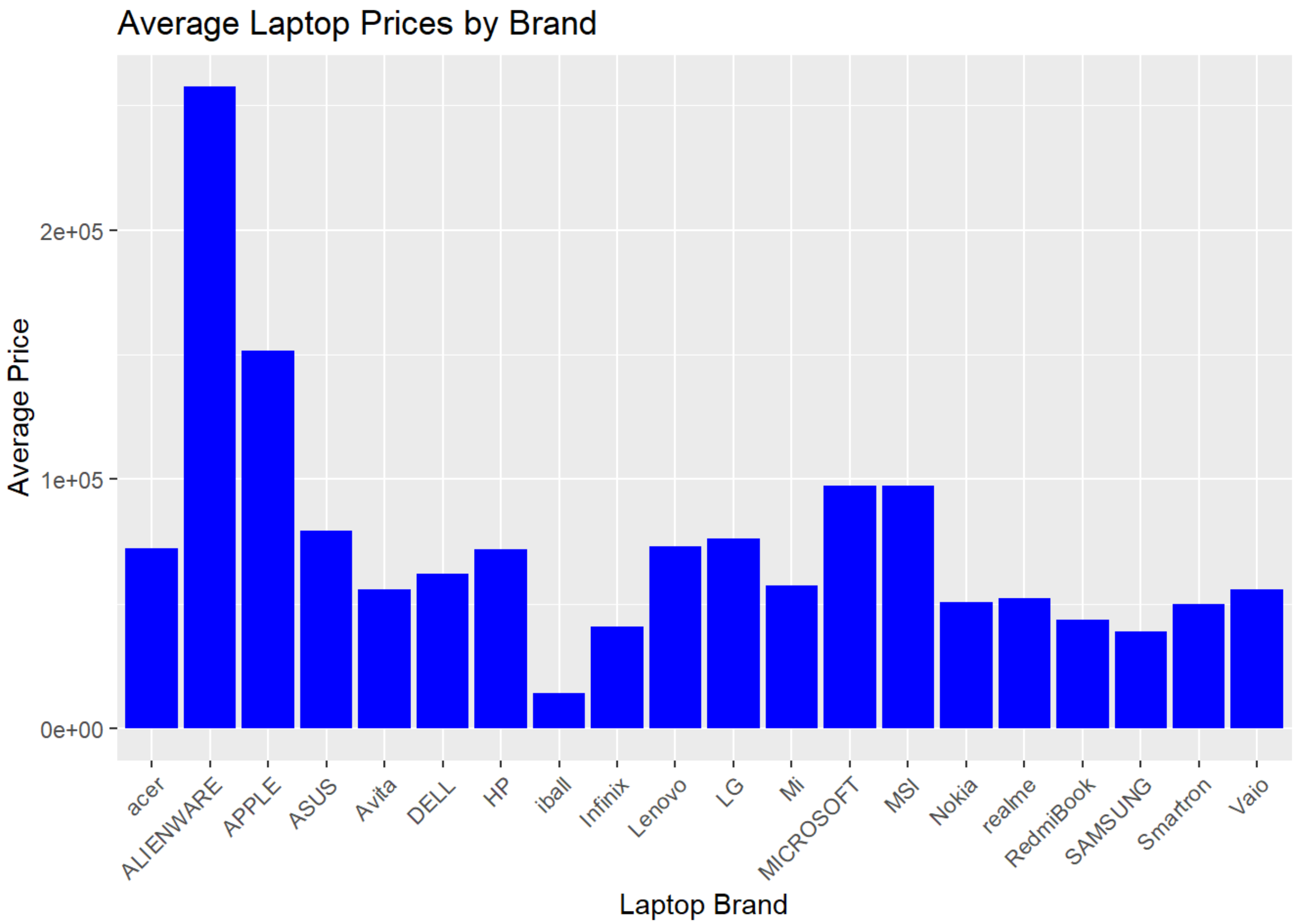


Above graph will show us brands with highest laptop price in rupees.

```r
avg_prices <- aggregate(latest_price ~ brand, data = data, FUN = mean)
colnames(avg_prices) <- c("brand", "Average Price")

p <- ggplot(avg_prices, aes(x = `brand`, y = `Average Price`)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(x = "Laptop Brand", y = "Average Price") +
  ggtitle("Average Laptop Prices by Brand") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Display the plot
print(p)
```



Above graph will show us brands with there average price in rupees.

Questions for next stage:

1. Is the average price of the products equal to the reference value Rs. 73k?
2. Is the average star rating of the products equal to the reference value 3.2?