Flipkart Laptops Dataset Analysis

Yanbo Tong

Rutgers University - School of Communication and Information

17:610:560 Foundations of Data Science

PhD. Haein Kong

May 7, 2025

Topic Statement

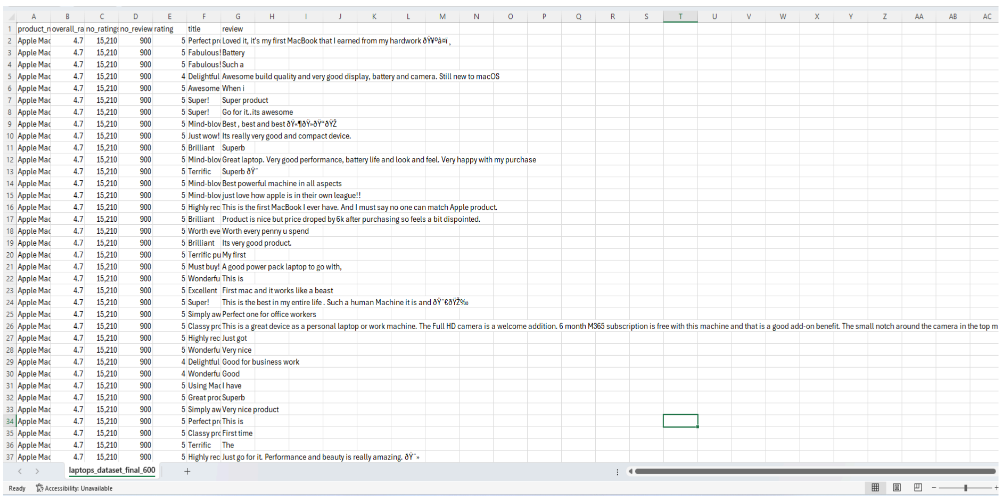
- Topic: Analysis of Customer Reactions (Numeric Variables) in Flipkart Laptop Reviews
- Motivation and Purpose

Understanding the numeric variables, i.e. overall rating, number of ratings, etc., number of reviews, etc., in product reviews is crucial for e-commerce platforms and manufacturers to enhance product offerings and customer satisfaction. By analyzing Flipkart laptop reviews via machine learning regression models, we aim to discover the factors affecting customer satisfaction. In addition, using the data visualization techniques and creating a scatter plot and bar charts, this analysis can represent the distributions between the independent variables and dependent variables, and provide actionable insights for improving product features and marketing strategies.

Data

- Source: Data Source: Source: The dataset is from Kaggle, titled "Laptop reviews dataset (Flipkart)" by Gitaditya Maddali.
  https://www.kaggle.com/datasets/gitadityamaddali/flipkart-laptop-reviews
- Preview of Data: See the screenshot of the original, uncleaned dataset (laptops_dataset_final_600.csv) below
- Basic Information (The four bold variables are the ones that will be discussed further in this project)
  - Number of Rows: 24,113
  - Number of Columns: 7
  - Column Names and Meanings:

- product_name: Name of the laptop model reviewed
- **overall_rating**: Average rating of the product
- **no_ratings**: Total number of ratings given to the product
- **no_reviews**: Total number of reviews for the product
- **rating**: Rating given by an individual reviewer (out of 5)
- title: Short summary of the review.
- review: Full text of the customer review



	A	B	C	D	E	F	G
1	product_n	overall_ra	no_ratings	no_review	rating	title	review
2	Apple Mac	4.7	15,210	900	5	Perfect pr	Loved it, it's my first MacBook that I earned from my hardwork ðŸ¥ºâ¤ï¸
3	Apple Mac	4.7	15,210	900	5	Fabulous!	Battery
4	Apple Mac	4.7	15,210	900	5	Fabulous!	Such a
5	Apple Mac	4.7	15,210	900	4	Delightful	Awesome build quality and very good display, battery and camera. Still new to macOS
6	Apple Mac	4.7	15,210	900	5	Awesome	When i
7	Apple Mac	4.7	15,210	900	5	Super!	Super product
8	Apple Mac	4.7	15,210	900	5	Super!	Go for it..its awesome
9	Apple Mac	4.7	15,210	900	5	Mind-blov	Best , best and best ðŸ«¶ðŸ»ðŸ"ðŸŽ
10	Apple Mac	4.7	15,210	900	5	Just wow!	Its really very good and compact device.
11	Apple Mac	4.7	15,210	900	5	Brilliant	Superb
12	Apple Mac	4.7	15,210	900	5	Mind-blov	Great laptop. Very good performance, battery life and look and feel. Very happy with my purchase
13	Apple Mac	4.7	15,210	900	5	Terrific	Superb ðŸ˜
14	Apple Mac	4.7	15,210	900	5	Mind-blov	Best powerful machine in all aspects
15	Apple Mac	4.7	15,210	900	5	Mind-blov	just love how apple is in their own league!!
16	Apple Mac	4.7	15,210	900	5	Highly rec	This is the first MacBook I ever have. And I must say no one can match Apple product.
17	Apple Mac	4.7	15,210	900	5	Brilliant	Product is nice but price droped by 6k after purchasing so feels a bit dispointed.
18	Apple Mac	4.7	15,210	900	5	Worth eve	Worth every penny u spend
19	Apple Mac	4.7	15,210	900	5	Brilliant	Its very good product.
20	Apple Mac	4.7	15,210	900	5	Terrific pu	My first
21	Apple Mac	4.7	15,210	900	5	Must buy!	A good power pack laptop to go with,
22	Apple Mac	4.7	15,210	900	5	Wonderfu	This is
23	Apple Mac	4.7	15,210	900	5	Excellent	First mac and it works like a beast
24	Apple Mac	4.7	15,210	900	5	Super!	This is the best in my entire life . Such a human Machine it is and ðŸ˜€ðŸŽ‰
25	Apple Mac	4.7	15,210	900	5	Simply aw	Perfect one for office workers
26	Apple Mac	4.7	15,210	900	5	Classy prc	This is a great device as a personal laptop or work machine. The Full HD camera is a welcome addition. 6 month M365 subscription is free with this machine and that is a good add-on benefit. The small notch around the camera in the top m
27	Apple Mac	4.7	15,210	900	5	Highly rec	Just got
28	Apple Mac	4.7	15,210	900	5	Wonderfu	Very nice
29	Apple Mac	4.7	15,210	900	4	Delightful	Good for business work
30	Apple Mac	4.7	15,210	900	4	Wonderfu	Good
31	Apple Mac	4.7	15,210	900	5	Using Mac	I have
32	Apple Mac	4.7	15,210	900	5	Great proc	Superb
33	Apple Mac	4.7	15,210	900	5	Simply aw	Very nice product
34	Apple Mac	4.7	15,210	900	5	Perfect pr	This is
35	Apple Mac	4.7	15,210	900	5	Classy prc	First time
36	Apple Mac	4.7	15,210	900	5	Terrific	The
37	Apple Mac	4.7	15,210	900	5	Highly rec	Just go for it. Performance and beauty is really amazing. ðŸ˜»

Methods

- Data Manipulation
  - Data Cleaning

    Before visualizing and analyzing the dataset, I first cleaned the dataset for more diverse and accurate results. To achieve this goal, data cleaning can be divided into three approaches: handling missing values, identifying and removing duplicates, and standardizing ratings. To be more specific, I removed the missing values of “review” and “ratings,” identified and removed duplicate reviews, and standardized the ratings to the scale of 1 to 5. These three steps ensure data integrity so that the analysis’s result can be more accurate. Before data cleaning, the original dataset had nearly 24.000 records in total, which is difficult and less convenient for the data analysis and visualization, but it was reduced to half of the original dataset (11,463 rows) after finishing these three steps.

```
{r}
# Data Cleaning Process
# Check if there are missing values of "review" and "rating"
laptop_cleaned <- laptop[!is.na(laptop$review) & !is.na
(laptop$rating), ]
laptop_cleaned
```

```
{r}
# Identify and remove duplicate records
laptop_after_clean <- laptop_cleaned[!duplicated(laptop_cleaned[c
("review", "rating")]), ]
laptop_after_clean
cat("Before removing duplicates:", nrow(laptop_cleaned), "\n")
cat("After removing duplicates:", nrow(laptop_after_clean), "\n")
```

  - Data Prepossessing

In the data preprocessing stage, I created four new variables: three variables based on the product's operating systems and one based on the "rating". To create the new variable "sentiment, " I took advantage of the dplyr library and the mutate function, which the code is shown in the RStudio in the screenshot below. These two steps aim to discover the total number sold of the three operating systems and summarize customers' ratings in the three words: "positive" for 4 and 5, "neutral" for 3, and "negative" for 1 and 2.

```
# Data Preprocessing
# Creating new features: Two new features based on the computer's
product name: ios_system and windows_system
laptop_standardized$ios_system <- grepl("Macbook",
laptop_standardized$product_name, ignore.case = TRUE)

# Create windows_system: TRUE if product_name contains Windows/Win
(case-insensitive)
laptop_standardized$windows_system <- grepl
("HP|DELL|Lenovo|SAMSUNG|ASUS|Acer|Primebook|CHUWI|Ultimus|MSI",
laptop_standardized$product_name, ignore.case = TRUE)

# Create chromebook_system
laptop_standardized$chromebook_system <- grepl("Chromebook",
laptop_standardized$product_name, ignore.case = TRUE)

# Save and view the updated csv file
write.csv(laptop_standardized, "laptop_updated.csv", row.names =
FALSE)
updated <- read.csv("laptop_updated.csv")

# View the first few rows in the console
head(updated)
```

```
{r}
# Calculate the number of laptops with ios system, windows system,
and chromebook
sum(updated$ios_system == TRUE)
sum(updated$windows_system == TRUE)
sum(updated$chromebook_system == TRUE)
```

```
{r}
# Create a new column 'sentiment' based on 'rating' (e.g., 4-5 as
'positive,' 3 as 'neutral,' 1-2 as 'negative')
library(dplyr)

updated <- updated |>
  mutate(sentiment = case_when(
    rating >= 4 ~ "positive",
    rating == 3 ~ "neutral",
    rating <= 2 ~ "negative"
  ))
head(updated[c("rating", "sentiment")])
```

- Data Visualization

In order to show the relationship between three numeric variables: overall_ratings, no_reviews, and no_rating, I used the ggplot2 library to create the scatter plot and three bar charts. For the bar charts among the three variables, they showed basically the same distribution trends: left-skewed distribution between each two of them.

```
{r}
# Create the bar chart to represent the distribution of average
rating of the product (overall_rating) and number of reviews
(no_reviews)

library(ggplot2)

# Assuming df_view has one row per product with columns:
overall_rating and no_reviews
ggplot(updated, aes(x = factor(overall_rating), y = no_reviews)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Number of Reviews by Overall Rating",
       x = "Overall Rating",
       y = "Number of Reviews") +
  theme_minimal()
```

```
{r}
# Create the bar chart to represent the distribution of average
rating of the product (overall_rating) and number of received
ratings (no_ratings)
library(ggplot2)

# Assuming df_view has one row per product with columns:
overall_rating and no_ratings
ggplot(updated, aes(x = factor(overall_rating), y = no_ratings)) +
  geom_bar(stat = "identity", fill = "navyblue") +
  labs(title = "Number of Rating by Overall Rating",
       x = "Overall Rating",
       y = "Number of Ratings") +
  theme_minimal()
```

```
{r}
# Create the bar chart to represent the distribution of number of
received ratings (no_ratings) and the number of received reviews
(no_reviews)
library(ggplot2)

# Assuming df_view has one row per product with columns:
overall_rating and no_ratings
ggplot(updated, aes(x = factor(no_ratings), y = no_reviews)) +
  geom_bar(stat = "identity", fill = "purple") +
  labs(title = "Number of Reviews by Number of Ratings",
       x = "Number of Ratings",
       y = "Number of Reviews") +
  theme_minimal()
```

```
{r}
# Data visualization
# Create the Scatter Plot to show the relationship between overall
rating and the number of reviews
# Count number of reviews per overall_rating
overall_rating_counts <- table(updated$overall_rating)
rating_df <- as.data.frame(overall_rating_counts)
colnames(rating_df) <- c("overall_rating", "no_reviews")

# Use ggplot2 for the scatter plot
library(ggplot2)

scatter_plot <- ggplot(rating_df, aes(x = overall_rating, y =
no_reviews)) +
  geom_point(color = "skyblue", size = 3) +
  labs(title = "Overall Rating vs Number of Reviews",
       x = "Overall Rating",
       y = "Number of Reviews") +
  theme_minimal()
scatter_plot
```

- Data Analysis

  To better predict the “overall_rating”, I created two regression models using mlbench and caret libraries: the first one only has one independent variable, “no_rating,” and the other one has two independent variables: “no_ratings” and “no_reviews.” Both of the two regression models use the 80:20 as the Train: Test Split. Finally, calculate the RMSE (Root Mean Squared Error) to determine which regression model is the better one to predict the “overall_rating.”

```r
{r}
# Data Analysis
# Create the regression model to predict "overall_rating" using
"no_ratings"
library(mlbench)
library(caret)

df <- select_if(updated, is.numeric)
??updated

# Train/test split (80/20)
set.seed(123)
split_idx <- createDataPartition(df$overall_rating, p = 0.8, list
= FALSE)
train <- df[split_idx, ]
test <- df[-split_idx, ]
dim(train)
dim(test)
# Fit linear model
model <- lm(overall_rating ~ no_ratings, data = train) # prev
model <- train(overall_rating ~ no_ratings,
               data = train,
               method = "lm")
# Predict on test set
pred <- predict(model, newdata = test)

# Evaluate performance
actual <- test$overall_rating
rmse <- RMSE(pred, actual)

# Results
summary(model)
rmse
```

```r
{r}
# Create the regression model to predict "overall_rating" using
"no_ratings" and "no_reviews"
library(mlbench)
library(caret)

df <- select_if(updated, is.numeric)
??updated

# Train/test split (80/20)
set.seed(123)
split_idx <- createDataPartition(df$overall_rating, p = 0.8, list
= FALSE)
train <- df[split_idx, ]
test <- df[-split_idx, ]
dim(train)
dim(test)
# Fit linear model
model2 <- train(overall_rating ~ no_ratings + no_reviews,
               data = train,
               method = "lm")
# Predict on test set
pred2 <- predict(model2, newdata = test)

# Evaluate performance
actual <- test$overall_rating
rmse2 <- RMSE(pred2, actual)

# Results
summary(model2)
rmse2
```

Findings

- Regression Models
    - The second regression model is better than the first one because it has a smaller RMSE (rmse2 = 0.21791).

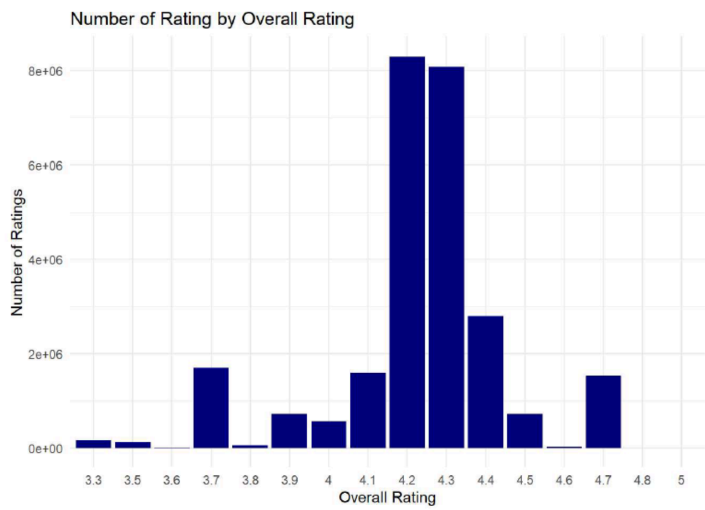
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86066 -0.06593  0.04203  0.13921  0.65162
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  4.148e+00  2.971e-03 1396.074  < 2e-16 ***
## no_ratings   2.620e-05  1.858e-06   14.105  < 2e-16 ***
## no_reviews  -1.433e-04  1.980e-05   -7.239 4.89e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2201 on 9169 degrees of freedom
## Multiple R-squared:  0.03763,    Adjusted R-squared:  0.03742
## F-statistic: 179.3 on 2 and 9169 DF,  p-value: < 2.2e-16
```

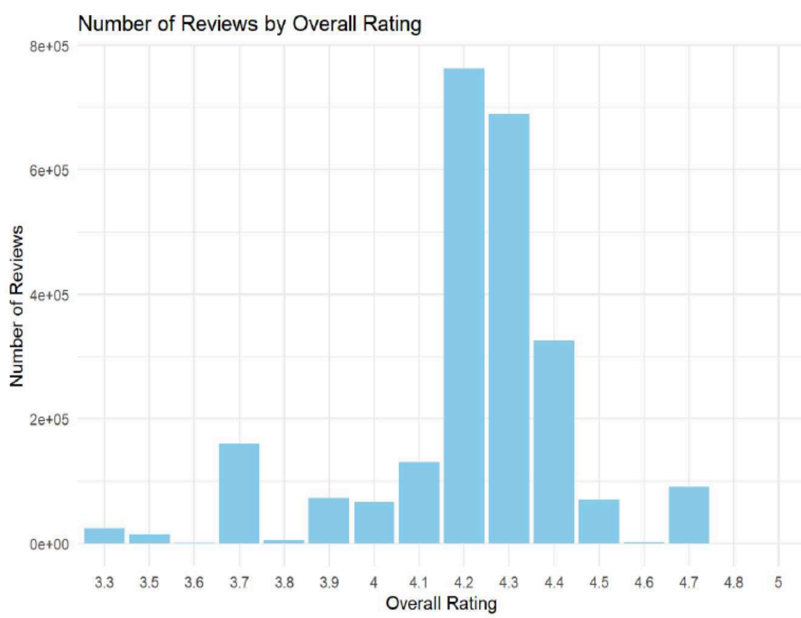
```
rmse2
```

```
## [1] 0.2179098
```

```
# The second regression model with three variables("overall_rating", "no_ratings",and "no_review
s") is better than the first one because it has lower RMSE.
```

- Key Data Visualizations and Explanations
    - Left-skewed distribution between the number of reviews and the overall rating
        - Same distribution between the number of ratings and the overall rating





Conclusion

- Summary and what I learned from the analysis

  Besides the distribution between the three numerical variables, I found some other conclusions. Customers are more willing to buy laptops with the Windows system rather than the iOS system. The most frequent overall ratings with the most reviews are **4.2 and**

4.3 out of 5.0. And the product's overall_rating is closely related to **both** the two independent variables: no_ratings and no_reviews

- Connection with Topic Statement and Motivation

  Although Flipkart sells a large number of laptops, based on the bar charts above, it doesn't have a very high rating out of 5. It means that the electronic store needs to fix its marketing strategies to improve its service quality, instead of only selling laptops.

Limitations and Suggestions

Without a doubt, as an open-source dataset, the project still has some limitations that come from the dataset. For example, the dataset only includes the primary laptop brands; some brands and products are not included in it. Also, it lacks some essential features that can influence the customer's ratings, including the price, warranty, and return rates, etc. To refine this situation, the publishers should regularly update the features in the dataset so that further research can consult directly with the updated dataset. Besides, due to the numeric values and different comments in the dataset, it is difficult to analyze all the variables. So some word variables like the "reviews" and "title" are not included in the data analysis section. Further researchers need to leverage the text-mining approaches to deal with these two variables.

References

- "Laptop reviews dataset (Flipkart)" by Gitaditya Maddali.

  https://www.kaggle.com/datasets/gitadityamaddali/flipkart-laptop-reviews