Web Scraping Project

Submitted By : Sonali Samparna

Outline:
Scrape any of the following website(s) and scrape details for any of the products like mobile phones, TV's, Laptops, personal health equipment or anything related to your domain.
Choices of websites:
1. Flipkart.com
2. Ajio.com
3. Snapdeal.com
You can choose the website and products from the above.

What you need to do?
For example, for a product called mobile phones, capture the data, clean the data and create visualizations which help in understanding the budget phones, and high end phones. Please make sure that data has to be visualized as per the features, camera pixels, no of sim slots, and the data / plots should be displayed in such a way that the user should be able to understand which phone to buy if he has a set of features.
For example: If I have 25000 INR as my budget, and if I need a single sim with some company like Samsung in mind, I should be able to select the phones from a group of phones.

Web Scraping - Approach

Webpage – www.filpkart.com
Product – Mobile-Phones

Step1: Access the webpage www.flipkart.com and search the product (mobiles). There will be 24 products per page and at bottom you can have total numbers pages of total results for your product (mobiles).

The link will Open as:
https://www.flipkart.com/search?q=mobiles&otracker=search&otracker1=search&marketplace=flipkart&as- show=on&as=off
So to traverse through each page logic is written to loop through till 10 (have scraped data for 10 pages) and pass the iteration value +1 to the base link like this

https://www.flipkart.com/search?q=mobiles&otracker=search&otracker1=search&marketplace=flipkart&asshow=on&as=off&page=1

Step 2: Identify the fields to be scrapped and look for associated class Ids like Mobile-Phone name, Stars, Ratings & Reviews, Actual Price, Discount and Discounted Price etc..

Step 3: All details of a mobile are present under a specific div in this case "_3ply-c row" lopping through this div based on specific tags the data was processed into a dictionary flipkart_temp the following data was fetched – **model name, mobile name (first value from model name), stars for that mobile, ratings, reviews, discount price, actual price, discount provided** were scraped.
Remove unnecessary symbols like '₹', % off, gb , mp which would help in cleansing of data.
After the details of a mobile are scrapped then dictionary df_flipkart flipkart_temp is appended to a list flipkart_data

after the all the pages have been iterated and the data has been appended to the list then the list is put into a dataframe "df" and this dataframe is transferred to a .csv file flipkart_data.csv which will be used for Exploratory Data Analysis (EDA)

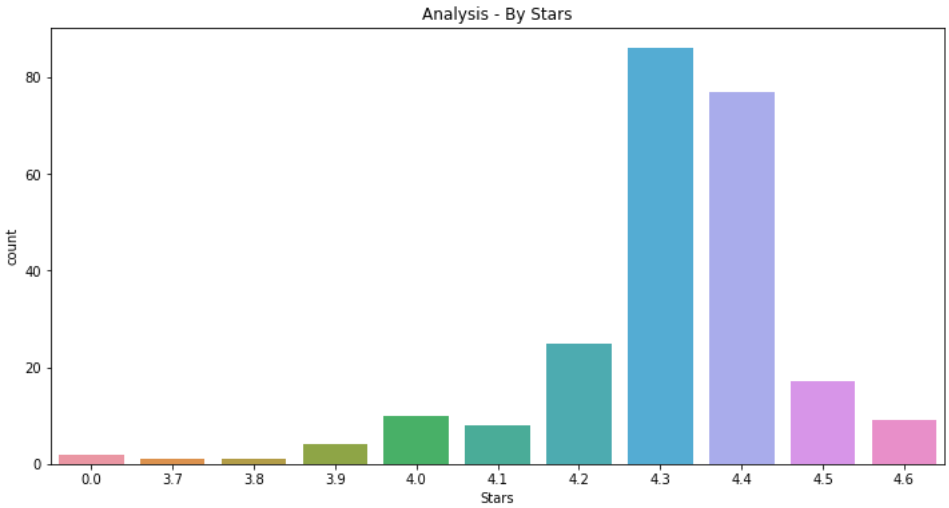
Step 4: Data Cleansing, Removals of NULLS, Impute columns

- Read the csv file 'flipkart_data.csv'.
- Ensure the dimensions of dataset Rows, columns
- Verify the dataset (Top 5 rows or so)
- Identify the data types of each columns
- Apply transformation if required(Here, Discounted Price, Rating and Actual Price showing as object convert into int)

- Check statistical summary of DataFrame (Min, Max, Count....)
- Check Null values
- Perform Data Analysis by Mobile Name, Company name, Stars, Ratings and etc.

Step 5: Data Analysis:

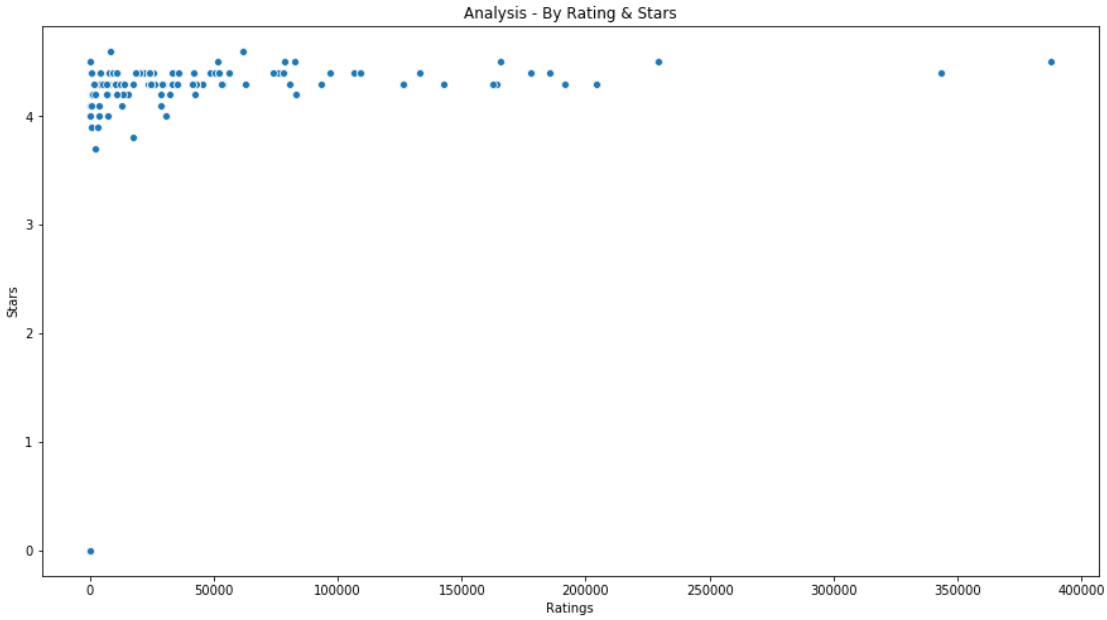
1. Analysis By Stars(Review)



Observation:

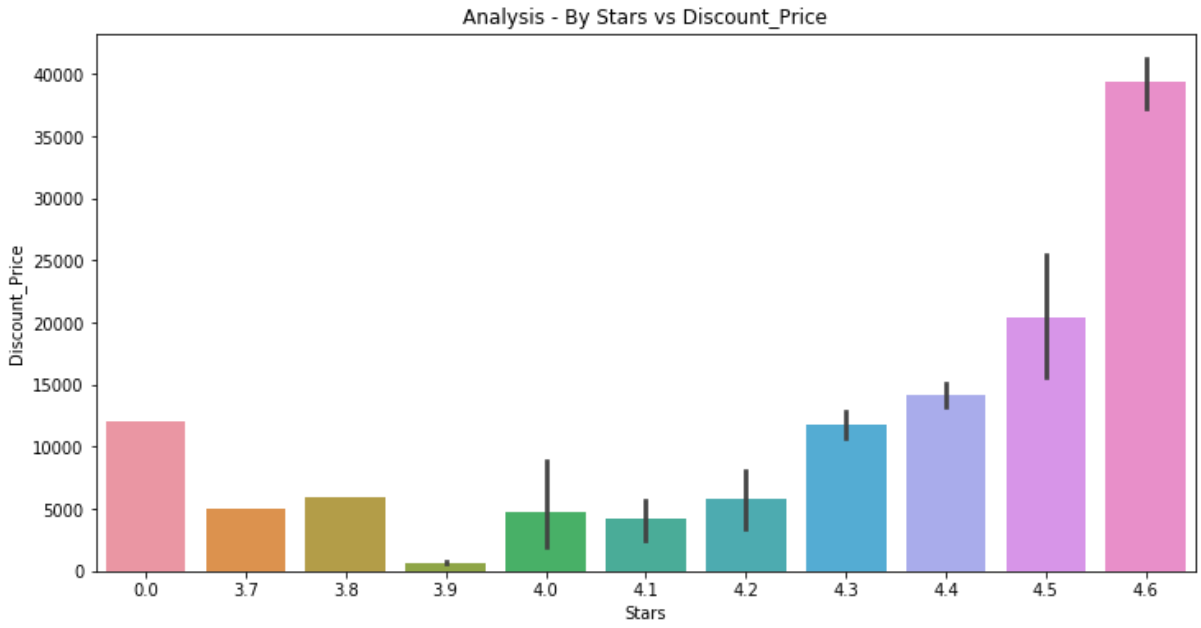
- The satisfied customers showed interest give ratings and given majority of them above 4
- 4.3 is given for a greater number of mobile phones

2. Analysis by Ratings and Stars



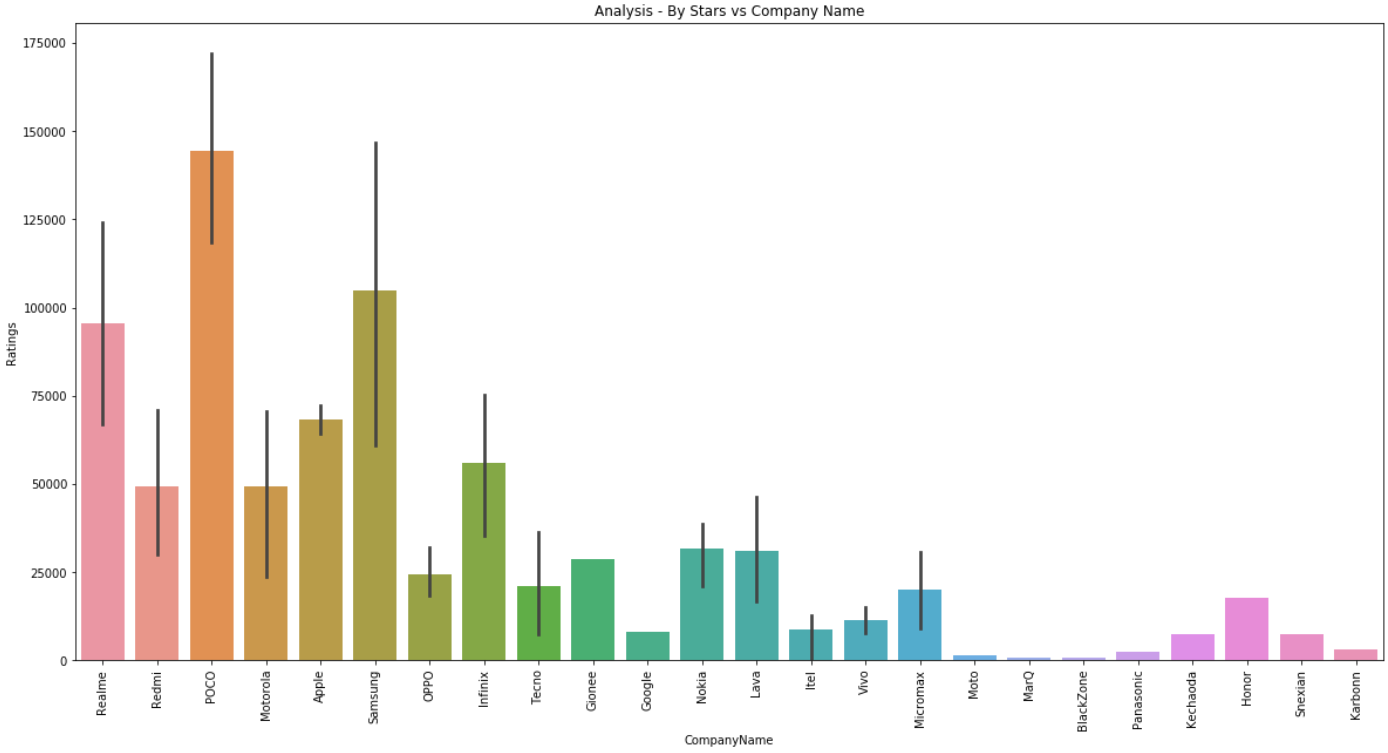
Observation:
- Most of the Mobiles phones are rated between 4 and 4.5¶
- Only few customers rated below 4 shows customer satisfaction

3. Analysis - By Stars vs Discount Price



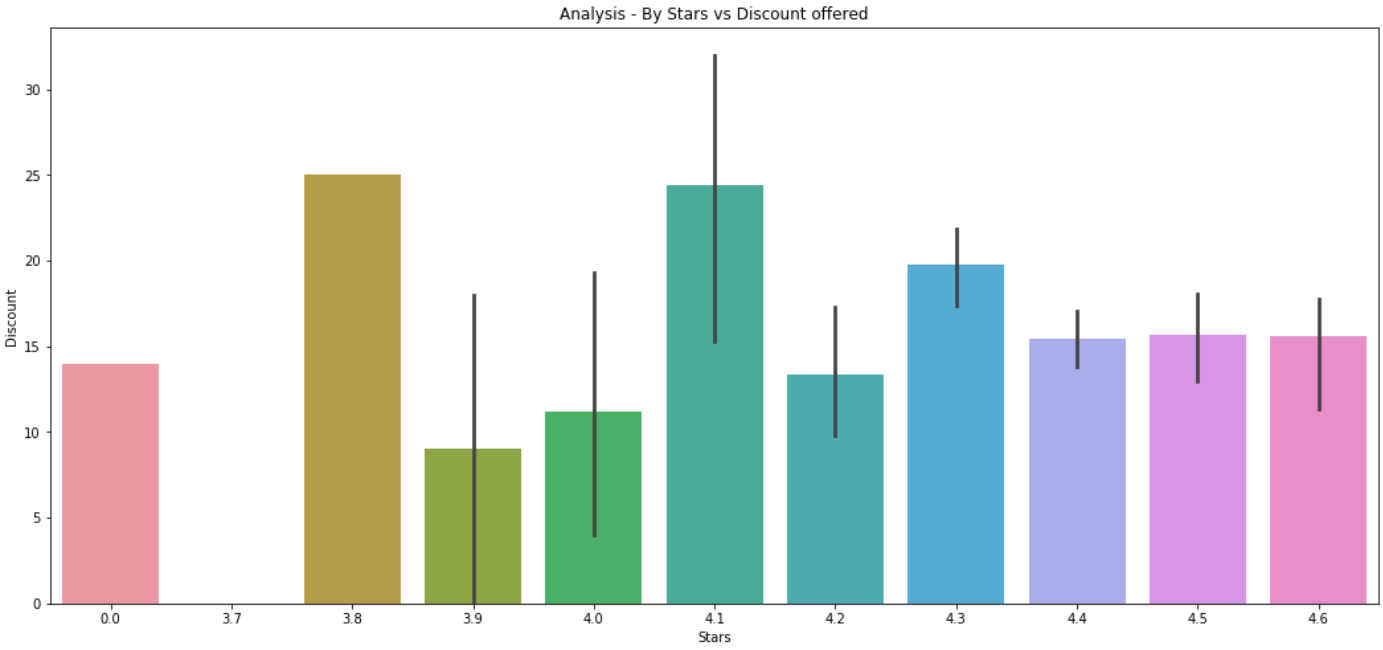
Observation:
- Customers are more satisfied with Costlier mobile-phones.
- Costlier the product(Mobile),Increase in the star rating.

4. Analysis - By Stars vs Company Name



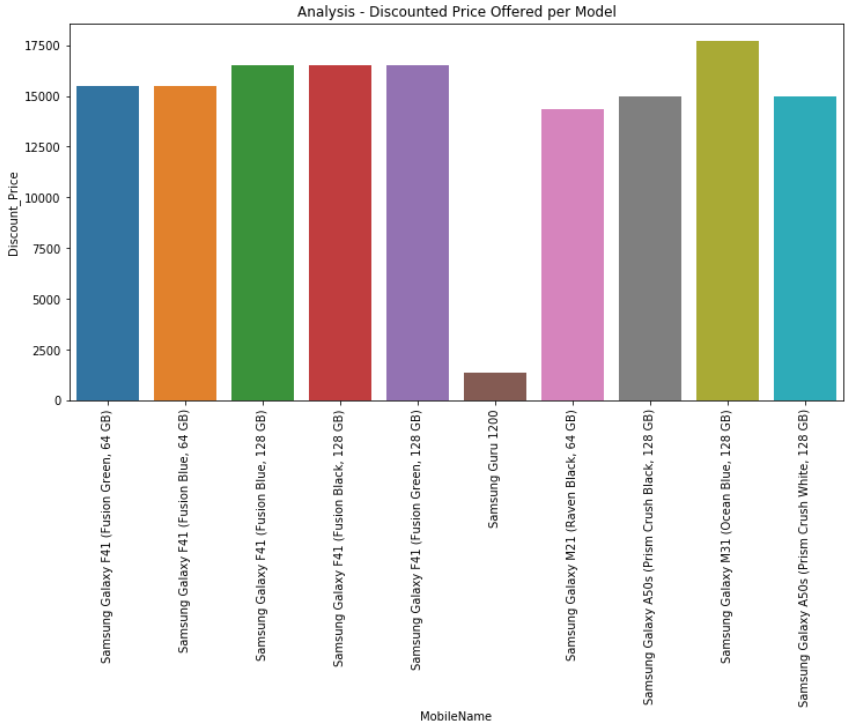
Observation:
- Samsung, Poco & Realme has received high number of Ratings.
- POCO brand is rated by most customers.

5. Analysis - By Stars vs Discount offered



Observation:
-Discount is an important factor impacting Ratings.
-Discount is directly proportional to Ratings.

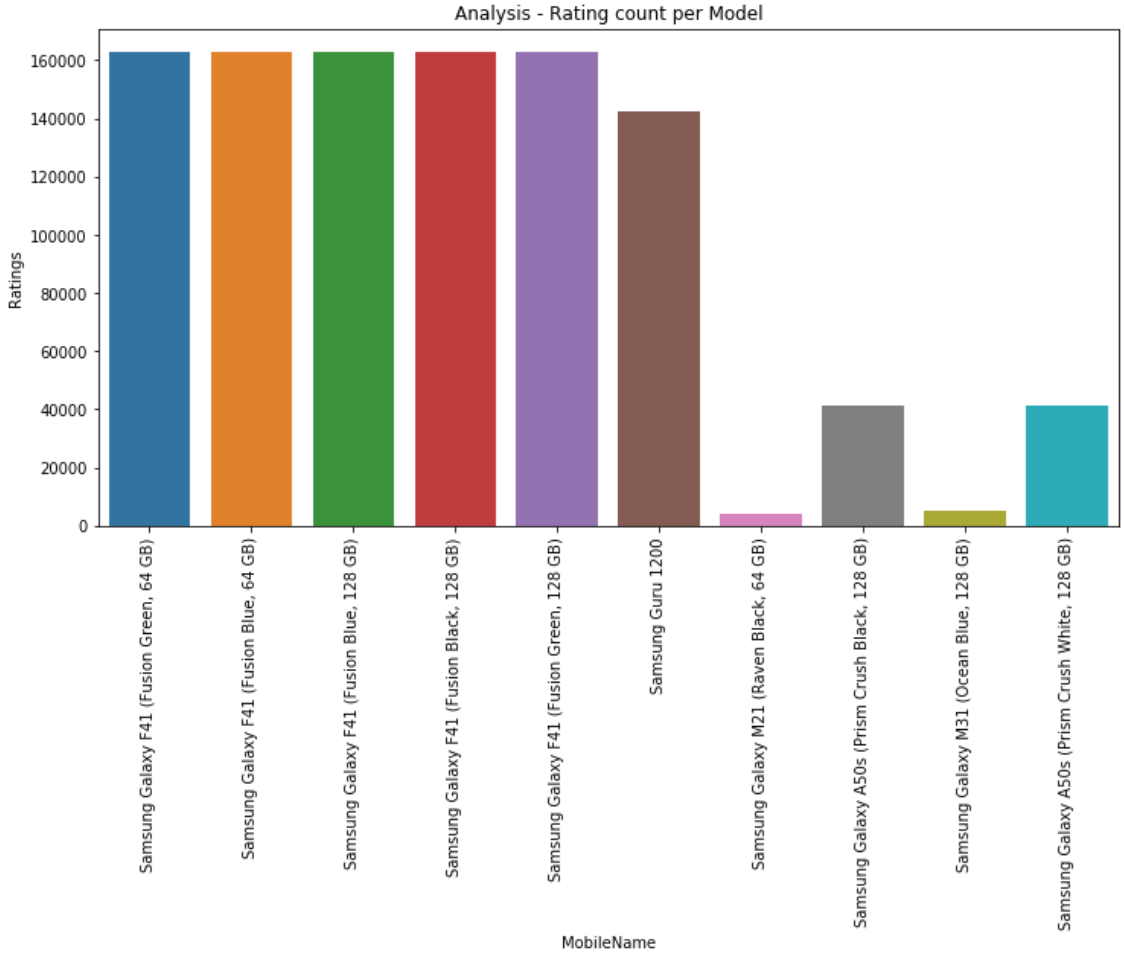
6. If I have 25000 INR as my budget and company like Samsung from the group of phones



Observation

- Based On discounted price "Samsung F41" is the best choice.

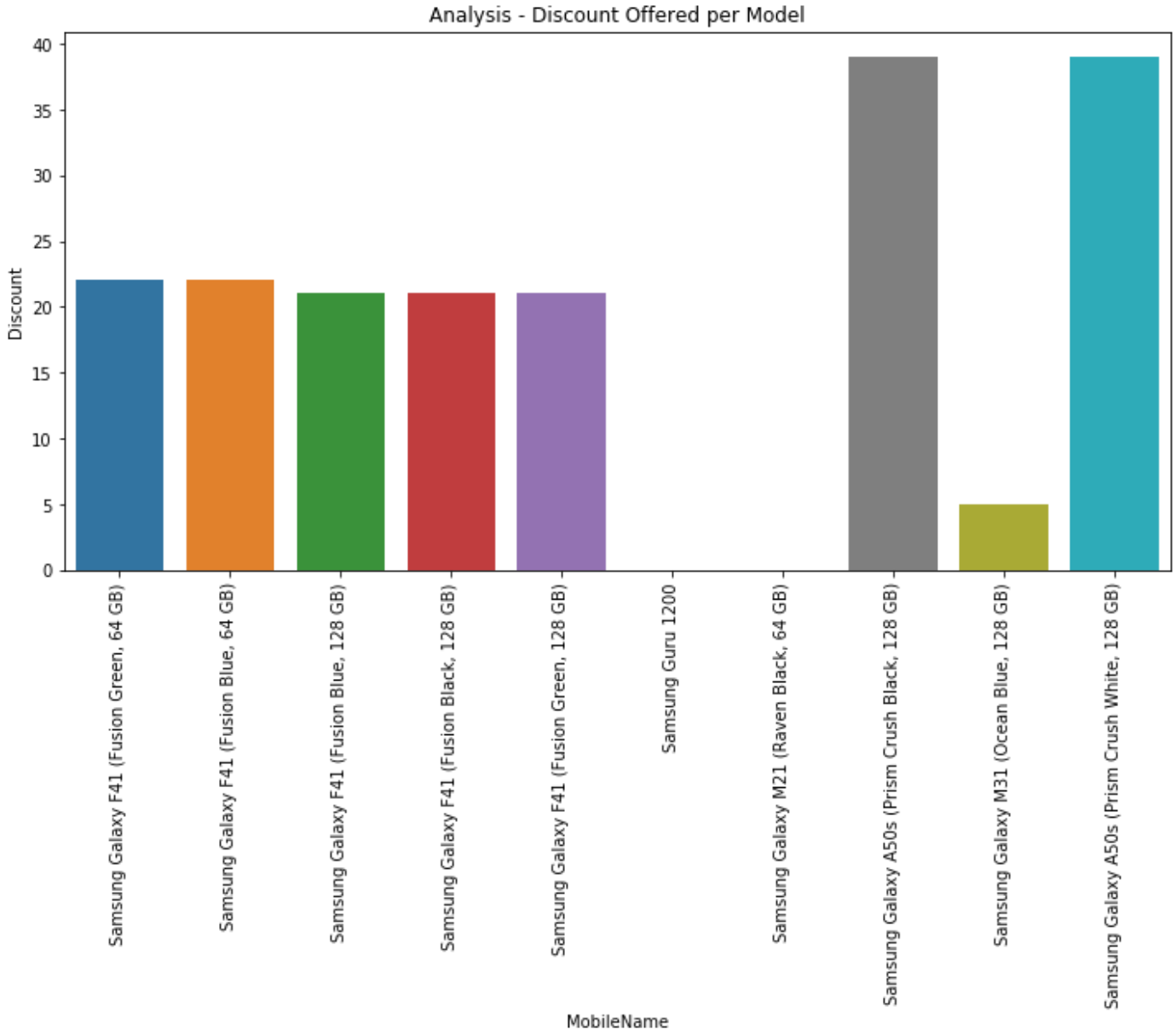
- Analysis -Rating count per model



Observations:

- Rating of "Samsung F41" is more.(All colours)

- Analysis -Discount Offered per Model



<u>Observation</u>:

- Based on Customer Price, Rating and discount offered, **'Samsung Galaxy F41'** would be best option available.