# *Capstone Project*

# *Data Analytics*

**WEB SCRAPING & EDA**

**(www.datafolkz.co.in)**

by

***SRINIVASA REDDY JAYAVARAPU***

## Objective:

The project is to scrap any one product from one of the E-Commerce website given from the following list and visualize the insights of the Data.

1. www.flipkart.com
2. www.snapdeal.com
3. www.ajio.com

I have chosen mobiles as the product from the website www.flipkart.com as target for my project.

## Approach for Scraping:

Search the product (in my case mobiles) in www.flipkart.com and the site gives set of 24 product in a paged manner and maximum of 40 pages will return as results.

In this project two level approach is used to achieve the desired objective.

First level get the 24 products url links from each page in loop and second level open each product page from url link received, once landed on the specific product web page, identify the required fields to be scrapped and get the respective html tags with associated class Ids.

This way the basic details like the Mobile Phone Brand name, Model Name, Model Number, Actual Price, offered Price (Discounted Price) Discount and Product ratings information.

The Technical details of each product are available in Specifications section. The Specification section is broadly divided into 10 Tables.

The Required technical details are scrapped from Table tags in a looped manner.

As all Technical details are not mentioned for each product it is not possible to get the required field with respective unique Html tag, hence a loop is run through entire 10 tables for specification name and if found then extract the respective specification value.

After getting all the information formatted the data into a Pandas Dataframe and then saved into a csv file with name 'Flipkart_Mobiles.csv'.

## Approach for Data Cleaning:

The Data from 'Flipkart_Mobiles.csv' is loaded into Pandas Dataframe and following cleaning activities are performed.

1. Renaming the Column Names:
    a. First and foremost, this activity can be avoided while saving the scrapped data into csv file by giving the appropriate name.
    b. But intentionally skipped embedding the appropriate symbols into columns name as it may not be the case that always we do get the data in the desired format.
    c. To enhance my learning skills I took this approach, if in case someone else is scrapped in this way and given for cleaning. How to handle the data like, price information missing the appropriate currency symbol, and Discount in % , rather than the amount, Battery units (capacity) taken into column name (as mAh) etc.

2. Identify Null (missing) values and fill with appropriate:

3. Splitting the column:
    a. As the 'Rating_Count' column contains both Number of Ratings (Rating_Count) given and Number of Reviews (Review_Count).
4. Convert the Identified columns (for visualization) into Numeric data type (wherever required).
5. Drop the Duplicate rows, if any exists.

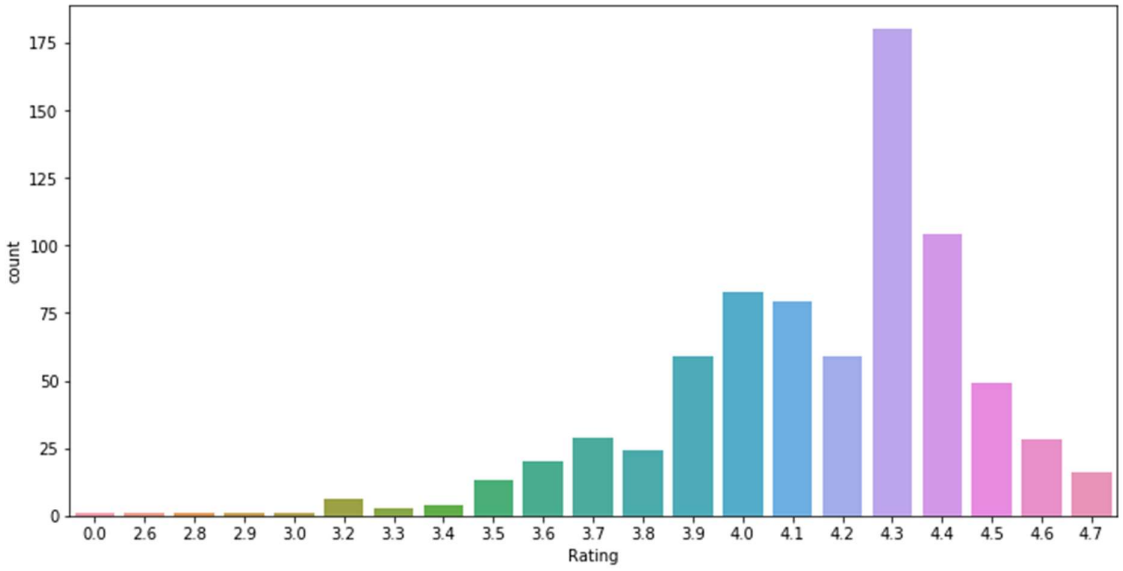
Now the Data is ready for Visualization.

## Approach for Visualization:

The list of columns identified for visualization are

1. Brand
2. Offered Price
3. Discount
4. Rating
5. Number of Ratings (Rating_Count)
6. Number of Reviews (Review_Count)
7. Primary Camera available
8. Secondary Camera available
9. Battery

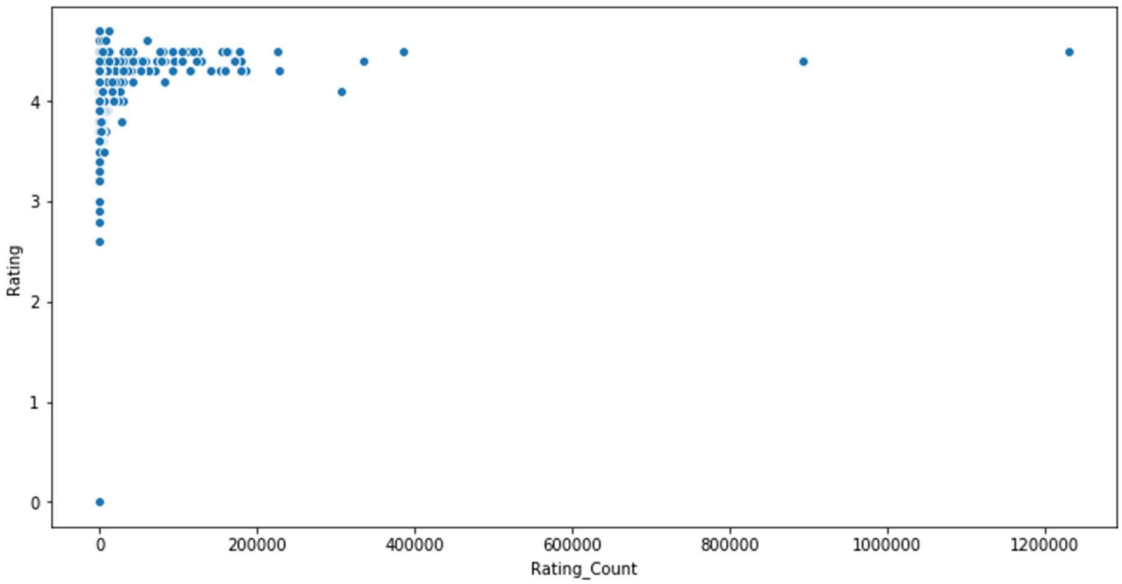
### 1. Count plot of Ratings




Observation:

- The satisfied customers showed interest give ratings and given majority of them above 3.5
- 4.3 is given for a greater number of mobile phones.

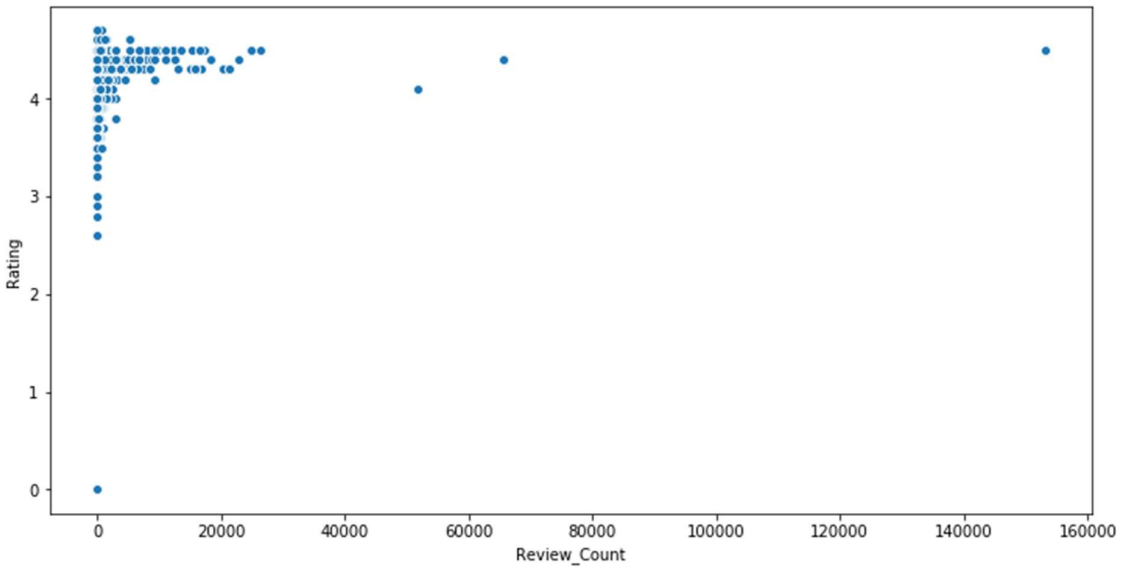
2. **Scatter plot - Rating_Count vs Rating (0 to 5)**




Observation:

- Most of the Mobiles phones are rated above 4.0
- The number of customers given rating below 4 are considerably less in number.
- A very small number phones got rating count more than 4 Lakhs.
- There are few mobiles phones which are not rated or Zero rating.
- No one rated below 2.5.

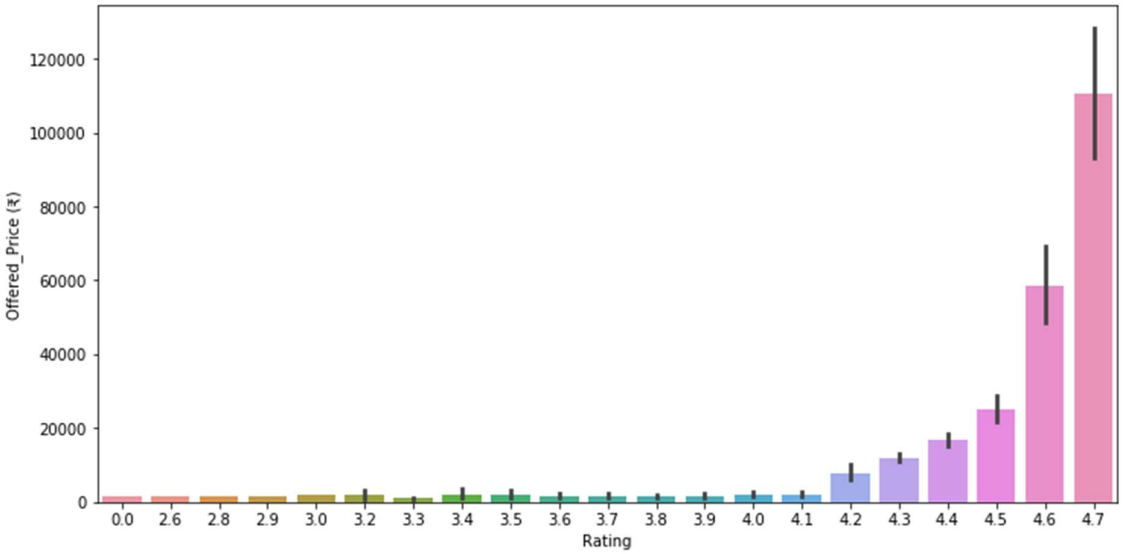
3. **Scatter plot Review_Count Vs Rating**




Observation:

- Mostly the people given rating are also reviewed the product.
- We can see similarity between Rating_Count Vs Rating and Review_Count Vs Rating.

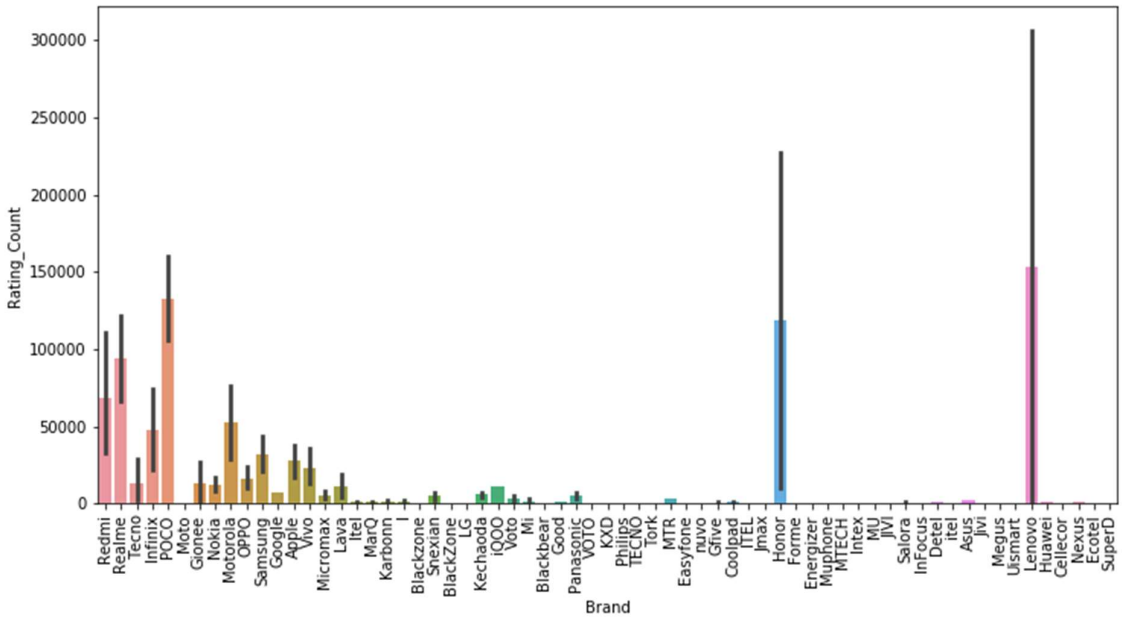
**4. Bar plot – Rating Vs Offered Price (Discounted Price)**




Observation:

- The phones with more than 20,000 Price are highly rated.
- As the cost of the phone increases, the satisfaction index of the customers increased.

**5. Bar plot – Rating Count Vs Brand**

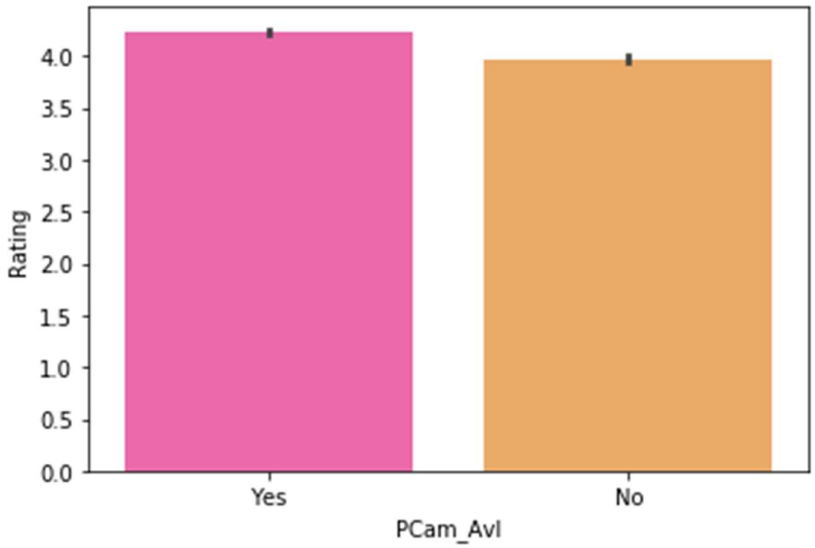



Observation:

- Though the Lenovo brand rated by many people the satisfaction index is distributed across all models. i.e few models received very few ratings and few models received very high number of ratings.

- The POCO brand is rated by many people in a consistent manner w.r.to number of ratings across all models.

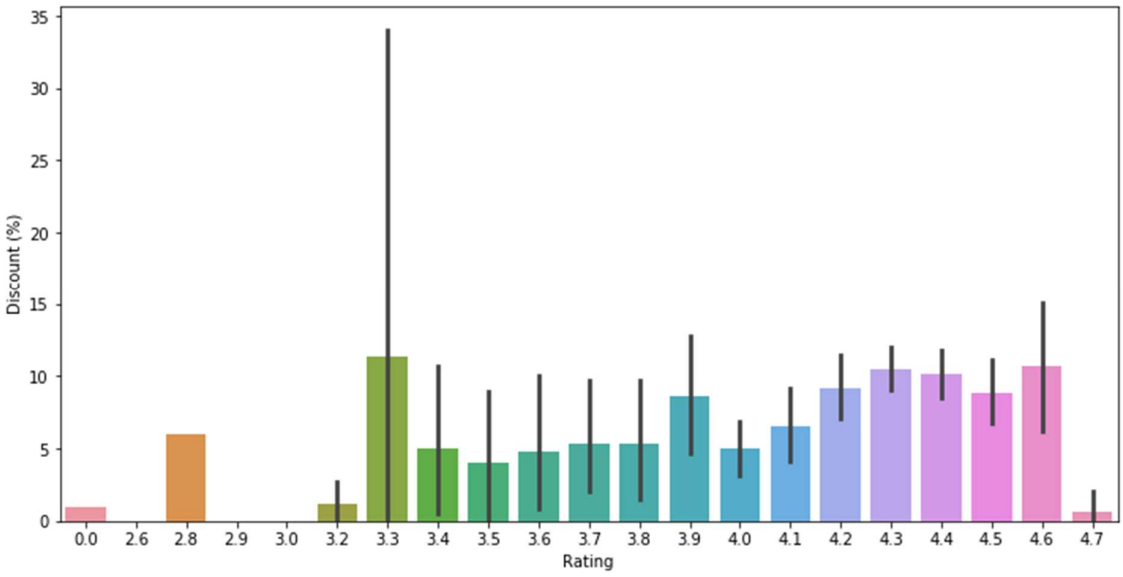
**6. Bar Plot – Primary Camera Availability (Yes / No)**




Observation:

- The rating is not really dependent the Primary camera availability.
- Almost Phones with Primary Camera and without Camera got same rating.

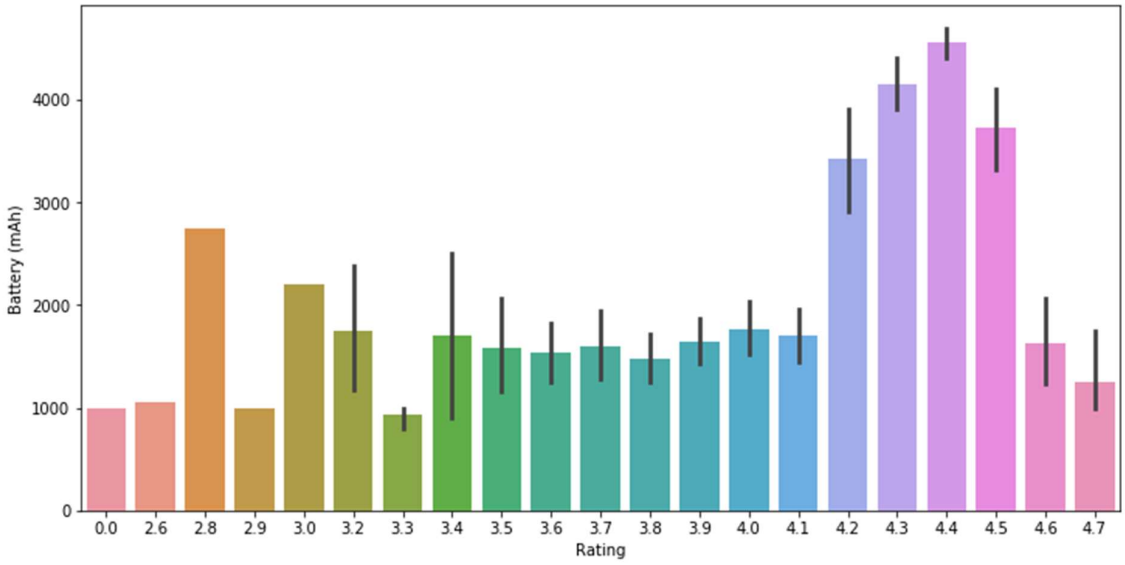
**7. Bar plot – Rating Vs Discount (%)**




Observation:

- The discount with 5% to 7% have ratings between 3.4 to 4.1.
- Discount with more than 10% phones are rated above 4.1
- Along with other features discount played vital role in rating.

**8. Bar Plot – Rating Vs Battery**




Observation:

- The phones with high Battery Capacity are rated high.
- The Phones with above 300 mAh battery capacity are rated above 4.1.
- The phones with battery capacity from 1000 mAh to 2000 mAh are rated between 3.4 to 4.1