EDA Capstone Project

Play Store App Review Analysis

By- Ranajay Biswas

Points of Discussion :

The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market.
In this exploratory data analysis project, we going to focus on these following topics -

- Loading the data in our work environment.
- Summarizing dataset.
- Cleaning & Imputation.
- Visualization.
- Discover key factors responsible for app engagement and success.
- Conclusion.

Step 1 : Reading the datasets

We have been given two datasets. The first dataset being the - Play Store dataset and the second one is - User Reviews dataset. For keeping it simple, we will first deal with our Play Store data. Complete the whole process of reading, understanding, feature engineering, visualizing and drawing all the possible insights based on the information present.

We have the data as a csv file format. We will use Python Pandas **read_csv**() function to read the dataset and convert it to tabular format. This tabular data is also called Pandas dataframe.

Now we can perform EDA on our data.

Step 2 : Summarizing Dataset

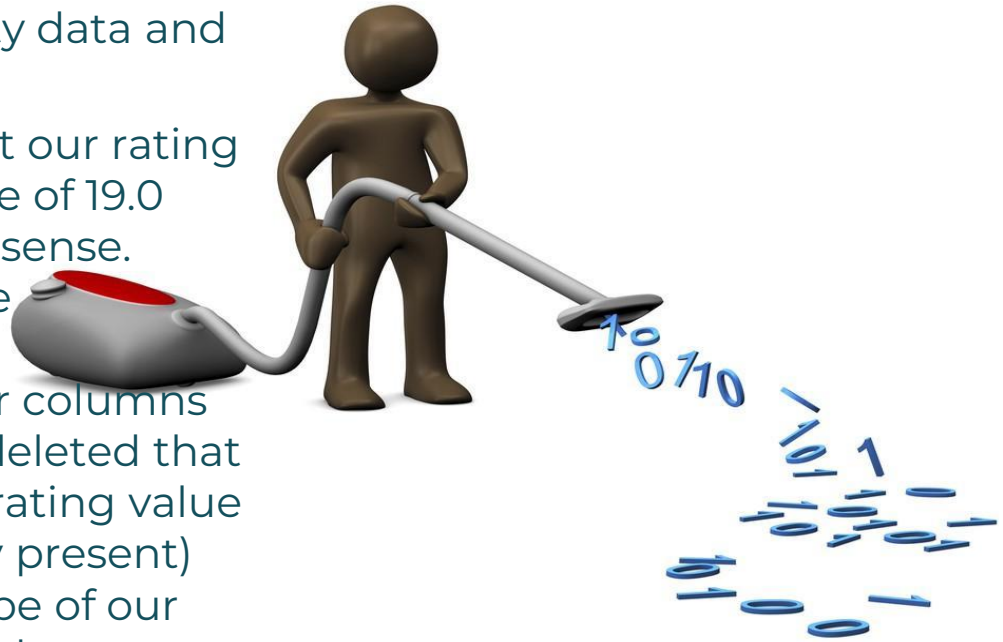
In this Step, we try to understand the dataset in hand. We followed the processes mentioned below –

1. *Shape of our dataset* : The findings tell us the our dataset comprises of 10841 rows and 13 columns.
2. *Identifying Columns* : Performing the pandas column function on our dataframe, we get the column names. Which in this case are - 'App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type', 'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver', 'Android Ver'.
3. *Checking Data Types* : Initially our dataset only has 'Rating' column as float64 type data. All the other columns have categorical values.
4. *Null Values Detection* : We tried to find which column has how many null values. So that we can start cleaning or filling up those values accordingly. Rating column have 1474 null values, Current version and Android version column had 8 and 3 null values respectively. Type and Content Rating columns each has 1 null value. We will need to take care of them.
5. *Identifying the Outlier* : We plot a boxplot on our dataset and found out that our rating column has one outlier.

Step 3 : Cleaning and Filling

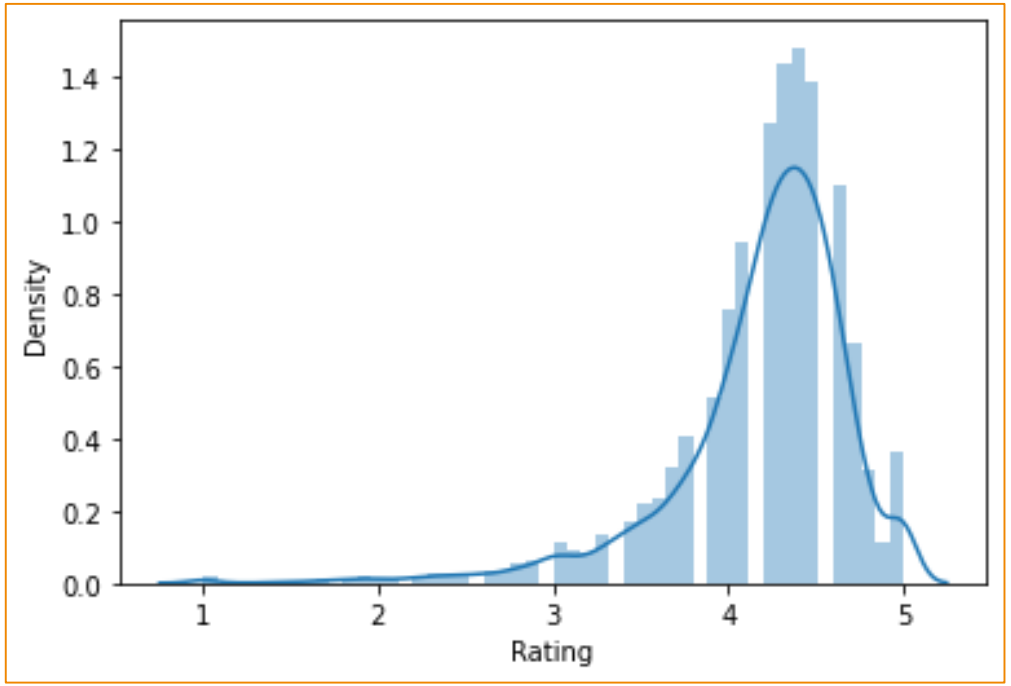
This step is crucial for removing faulty data and filling in missing values.

- Plotting a boxplot, we could tell that our rating column has an outlier with the value of 19.0
  In this case, it does not make any sense. Because the rating values should be between 1.0 and 5.0
- We don't know if the values in other columns for that row are true or not. So, we deleted that row and any such row which has a rating value not between 1.0 to 5.0 (If there's any present)
- After doing so, we checked the shape of our dataframe. It only deleted one row, the row which had the outlier present in it.



Step 3 Continues ...

- Now that we have got rid of the outlier, we can fill in the null values. We plotted a distribution plot to see the skewness of the data.



Filling the null values in Rating col :

Since our data looks to be negatively skewed, we are going to take the median value instead of the mean to fill the null values our 'Rating' column.

All the null values have been indeed filled with median value.

Step 3 Continues ...

Now that ‘Rating’ column has no null values, we shift our focus to other columns.

- Columns like ‘Type’, ‘Current Ver’ and ‘Android Ver’ still have null values. As they are not numerical values, we chose to take the mode values to replace the null values in those columns and finally we check our dataset again to confirm we don’t have any other null values in our dataset.
- The next challenge that stands before us now is that some of the columns have data types which is Object type. But we need them as numerical for visualizations and to draw meaningful insights.

Data Processing :

- We are choosing ‘Price’, ‘Reviews’, ‘Size’ and ‘Installs’ columns to convert their values numeric. We started by finding out the character(s) that we can remove to make the values numerical and meaningful for us to evaluate.
- After converting, we now have five numerical columns- ‘Rating’, ‘Price’, ‘Reviews’ , ‘Size’ and ‘Installs’

Step 4 : Feature Selection

Currently, we are doing only EDA, not working on any ML models. But we need to keep in mind that our feature selection should always be most relevant with our goals, it should be easier to debug and understand.

For this project, we need to understand the app market.
So, the goal is to figure out the following –

- Ratings for various apps categories.
- Which kinds of apps have had the most engagements.
- Which apps people preferred over other categories.
- The size distribution of the applications.
- How prices varied in different categories & the price factor influence on user engagement.
- What kinds of apps made the most money etc.

So in order to answer those, what we did was that we picked certain columns like 'Category', 'Rating', 'Reviews', 'Installs', 'Size', 'Content Rating', 'Price' etc.

We did group-by on 'Category' , then we try to visualize the correlations between category and these other columns and get some insights.

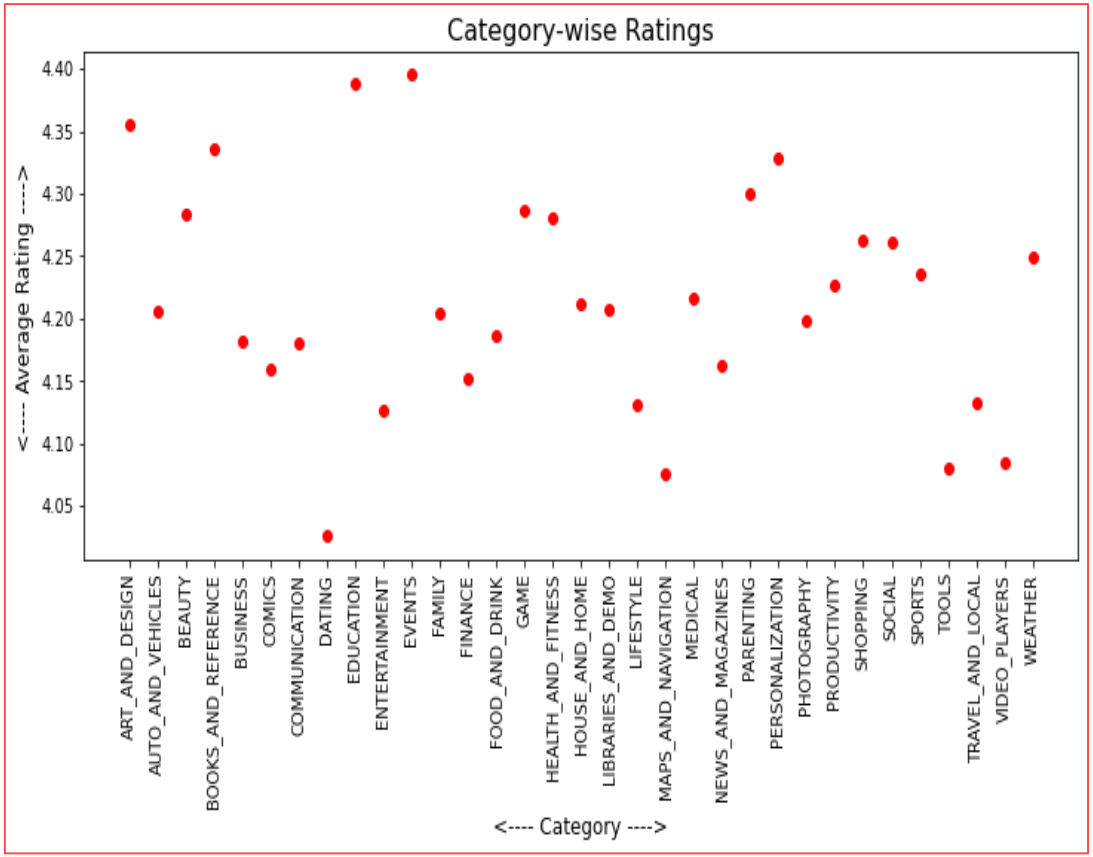
Step 5 : Visualization

This is the part where we are plotted different diagrams and charts to visualize our data.

Let’s check for some trends and patterns.

For this we used python libraries like Matplotlib and Seaborn.

Step 5 : Visualization



According to Ratings :

Plotting the scatterplot for Rating vs Category, we can see that the average ratings lie between 4.1 to 4.3 for majority of the app categories.

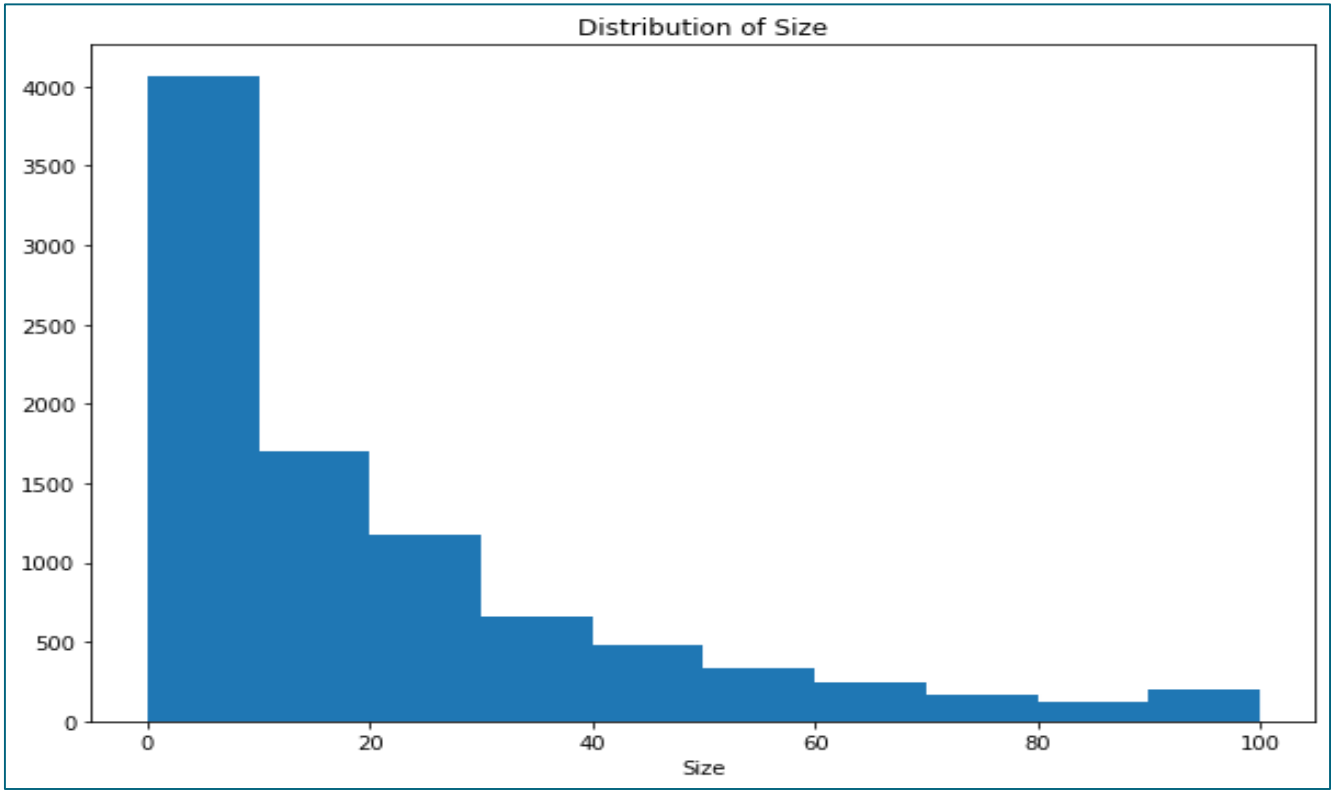
- The lowest rated app category is Dating apps. We can say that User Experience with these apps have not been great.
- The top 3 app categories according to user ratings are –

1. Events
2. Education
3. Art and design

Step 5 : Visualization

Distribution of App-sizes :

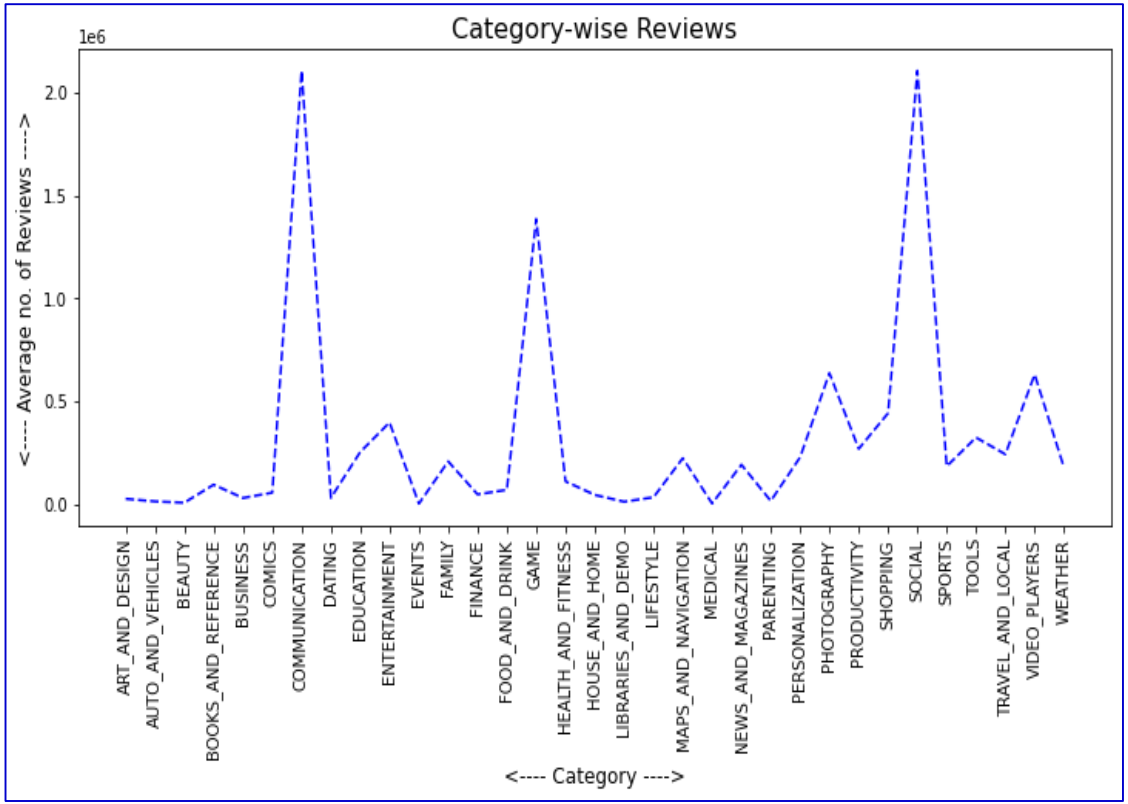
From the histogram, we can say that most of the apps in our datase t have smaller size somewhere between 0-30



Step 5 : Visualization

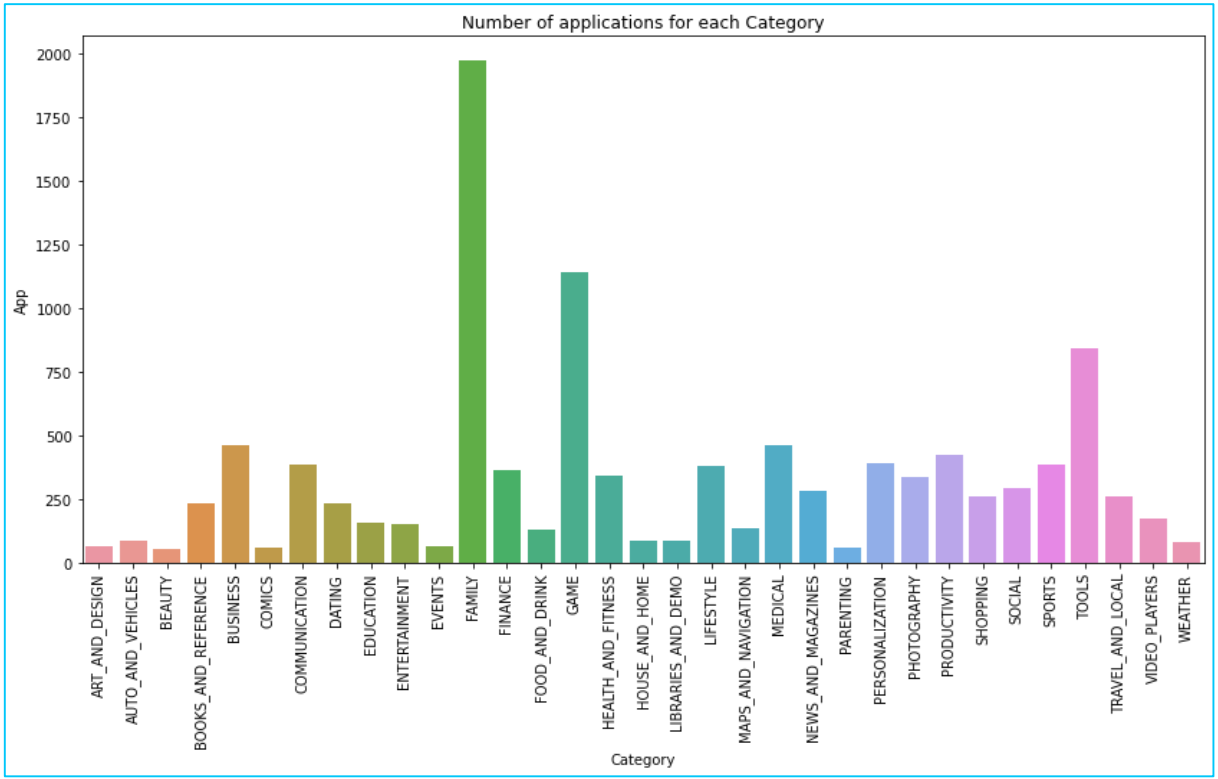
According to no. of Reviews :

- App categories belonging to Communication and Social got the most number of reviews.
- The 3rd most reviewed app category is Gaming
- The rating for these app categories were somewhere between 4.15 to 4.30



Step 5 : Visualization

Number of Apps in Each Category :



Conclusion :

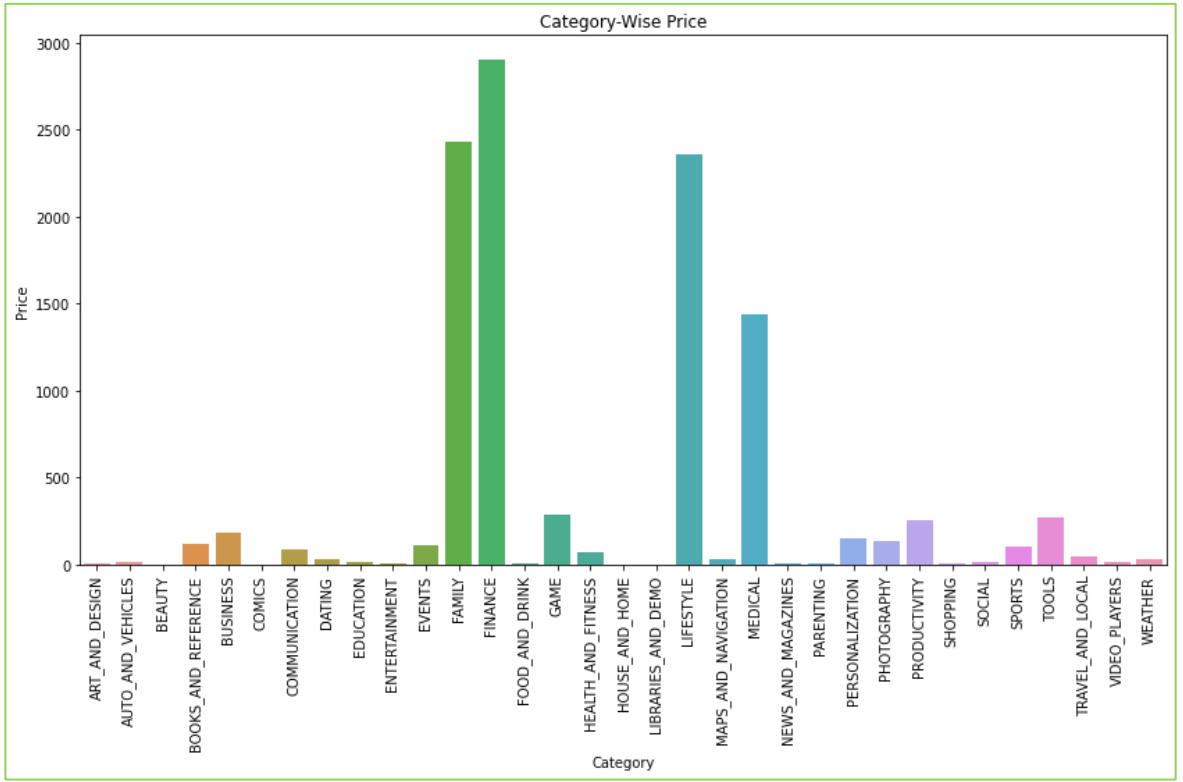
Family category has the most number of apps present in our dataset, Games & Tools being 2nd and 3rd respectively.

Step 5 : Visualization

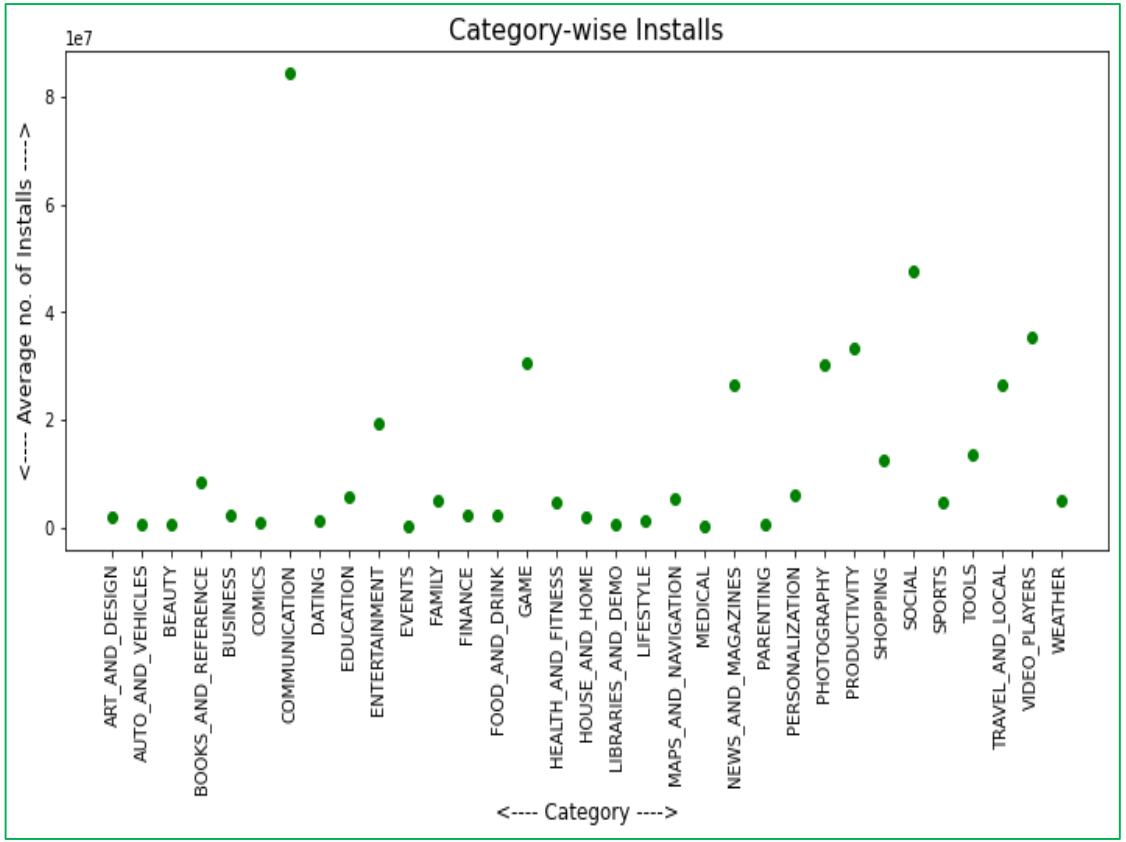
According to Prices :

- Finance, Family, Lifestyle and Medical apps have charged the most amount of money among all the app categories.
- Number of reviews that these apps received are low, with ratings somewhere between 4.1 to 4.25

We will later see how app prices have influenced installations.



Step 5 : Visualization

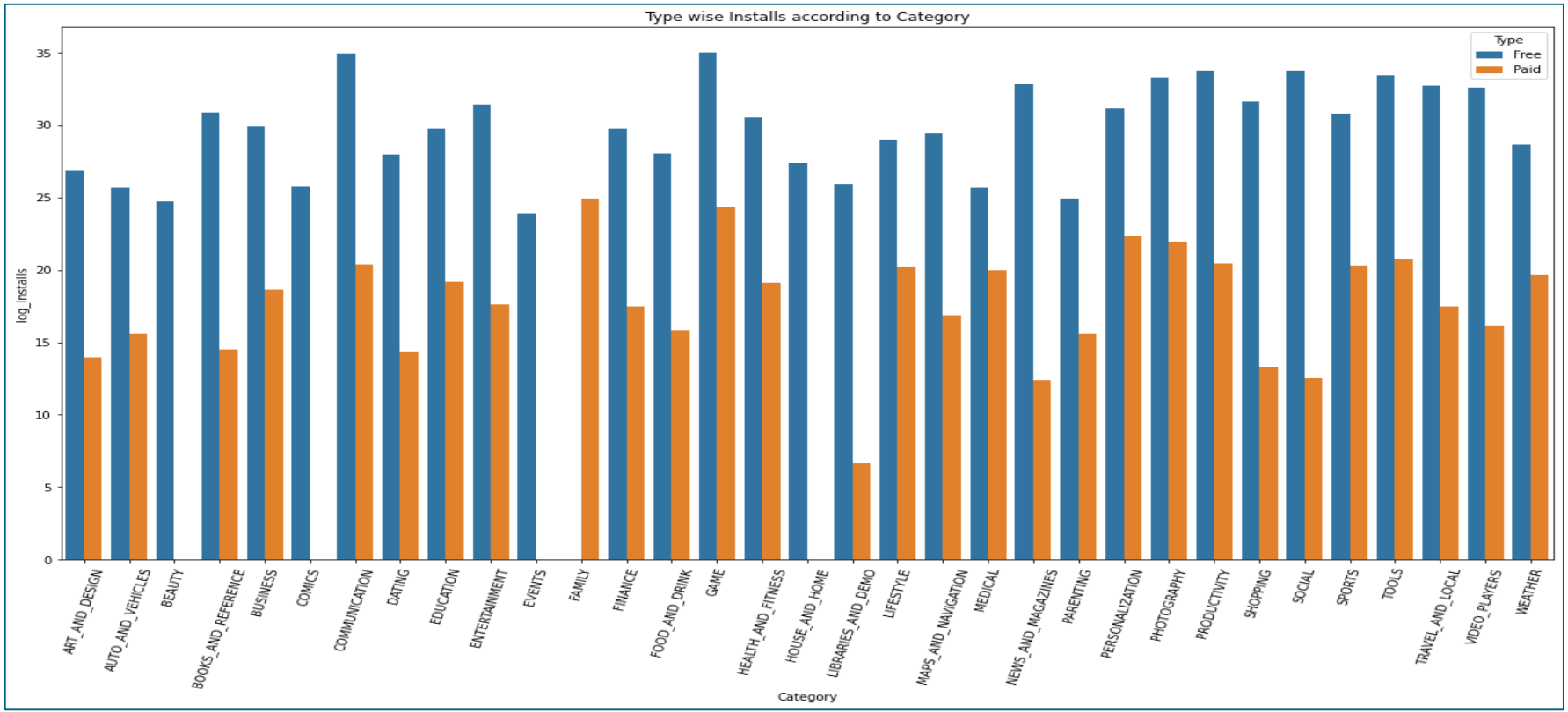


Installations in each category :

Communication apps have the highest number of Installs. A lot more than any other category. Followed by, Social apps.

Next, we are going to see how app price might have influenced Installations.

Step 5 : Visualization

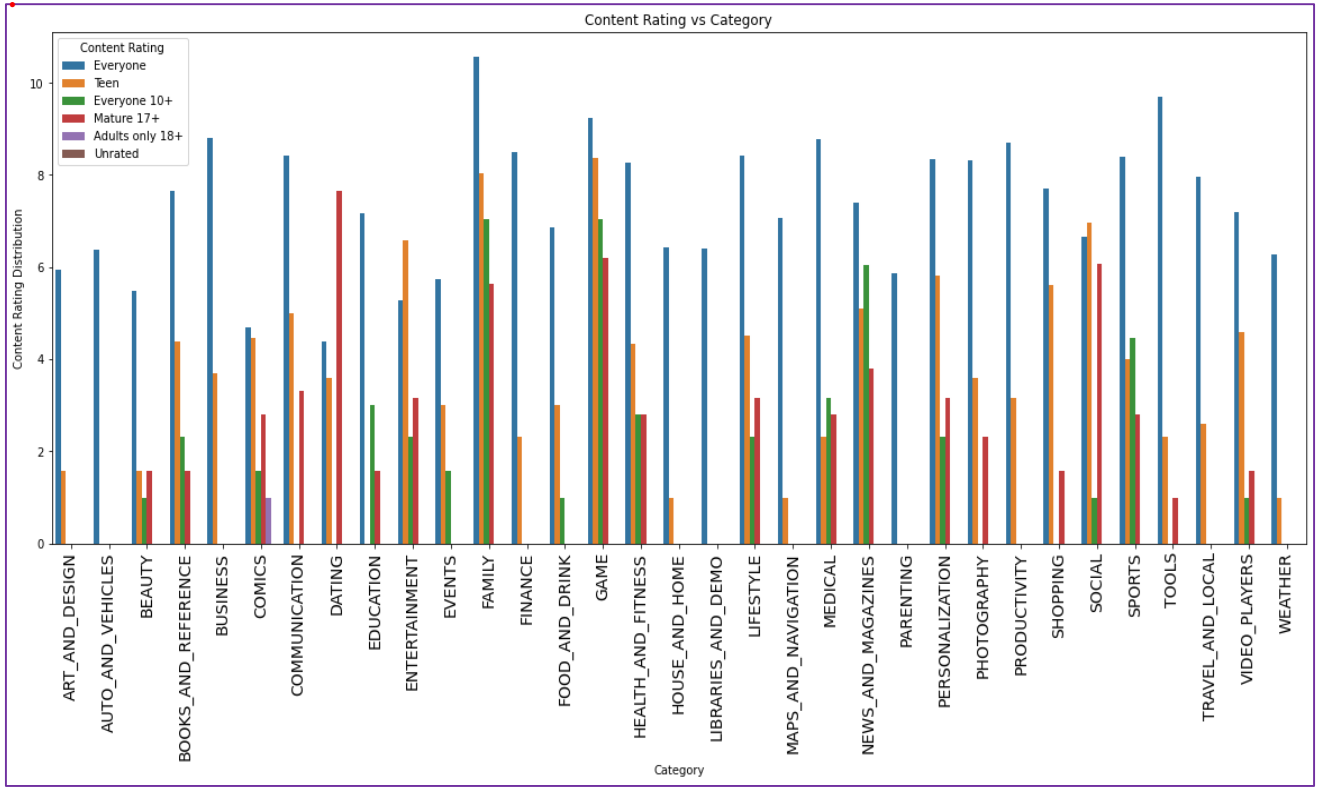


Conclusion : Not much surprise here... Free apps were installed by a lot more users for almost every category, compared to paid apps.

Step 5 : Visualization

Content Rating VS Category

.



- For every category in our data, there were applications with Content Rating - 'Everyone'..
- 'Comics' is the only content rating type which have noticeable amount of applications available for Adults.

With that, we were done with the analysis on the first dataset. Let's move on to the second one...

Part 2 : EDA on User Reviews Dataset

Our second dataset is another csv file named 'User Reviews'. We will first try to take a brief look at our dataset to figure out which necessary steps we need to follow to make best use of this dataset.

- *Reading Dataset and Understanding Features* : Using Pandas, we read the dataset in our python notebook and make a data-frame out of it.

- *Data Summary* : Checking the shape, we find this dataset has 64295 rows and 5 columns. Out of these 5 columns, only two of them are numerical and other three are object type.

The names of the columns are – App, Translated Review, Sentiment, Sentiment Polarity and Sentiment Subjectivity.

- *Null Values* : Checking the null values, we find the data has 26868 null values in each column except for App column. App column has no null values.

Digging a little deeper, we find out that where there is one null value in a row, the other 3 columns also have null values in that same row.

- ***Cleaning, Filtering and Feature Selection*** :

This process involves chopping off unnecessary and nonsensical data. As we found out in our previous step that the rows where only app name column values were present out of all the columns, those rows don't make any sense. And we cannot fill this null values, as we have no way of determining what values should we be putting there. So, the only option left was that we drop all those rows.

Applying this process left us with 37427 rows with no null values. PERFECT !

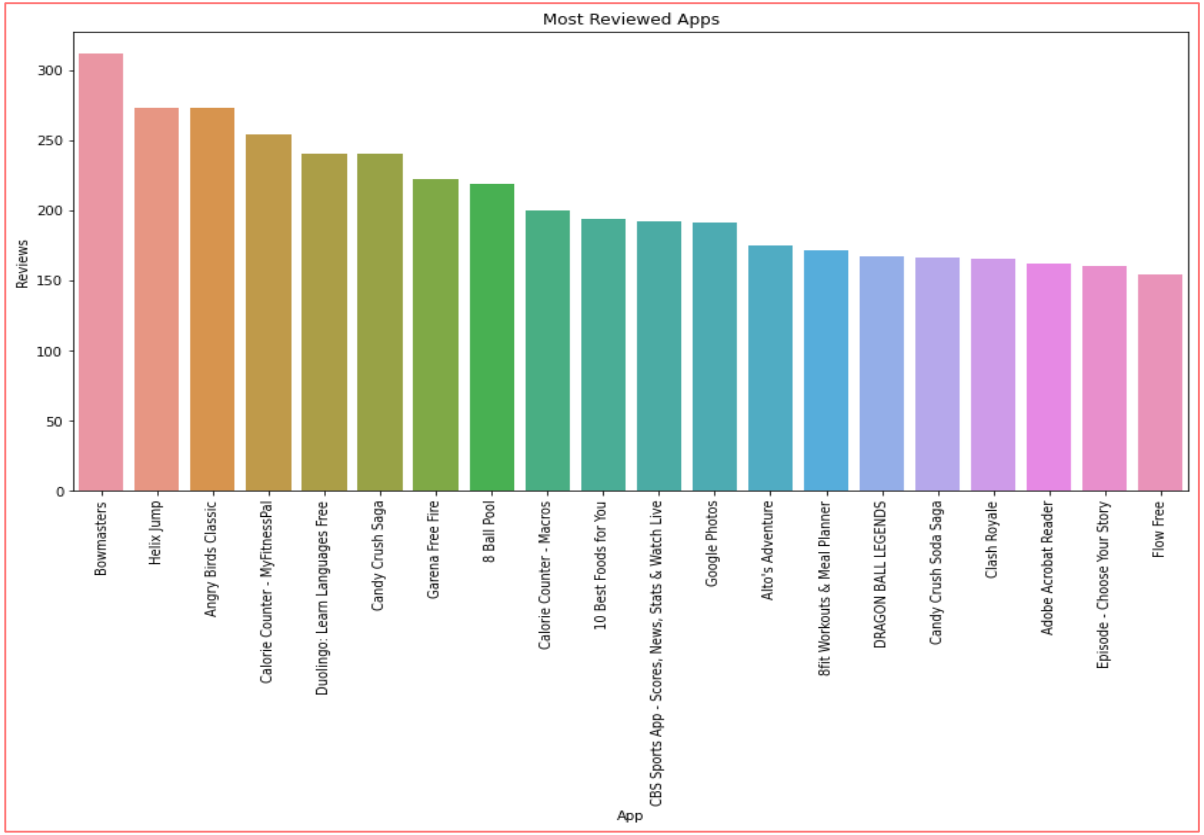
- ***Visualization*** **:**

We have multiple reviews for the same app coming from different users. What we are going to do is that we are going to group by our dataset on App name to see which apps had what results.

We can figure out which apps have been the most popular in terms of user reviews. And among the given reviews which apps have most positive or negative reviews etc.

- *Visualization* :

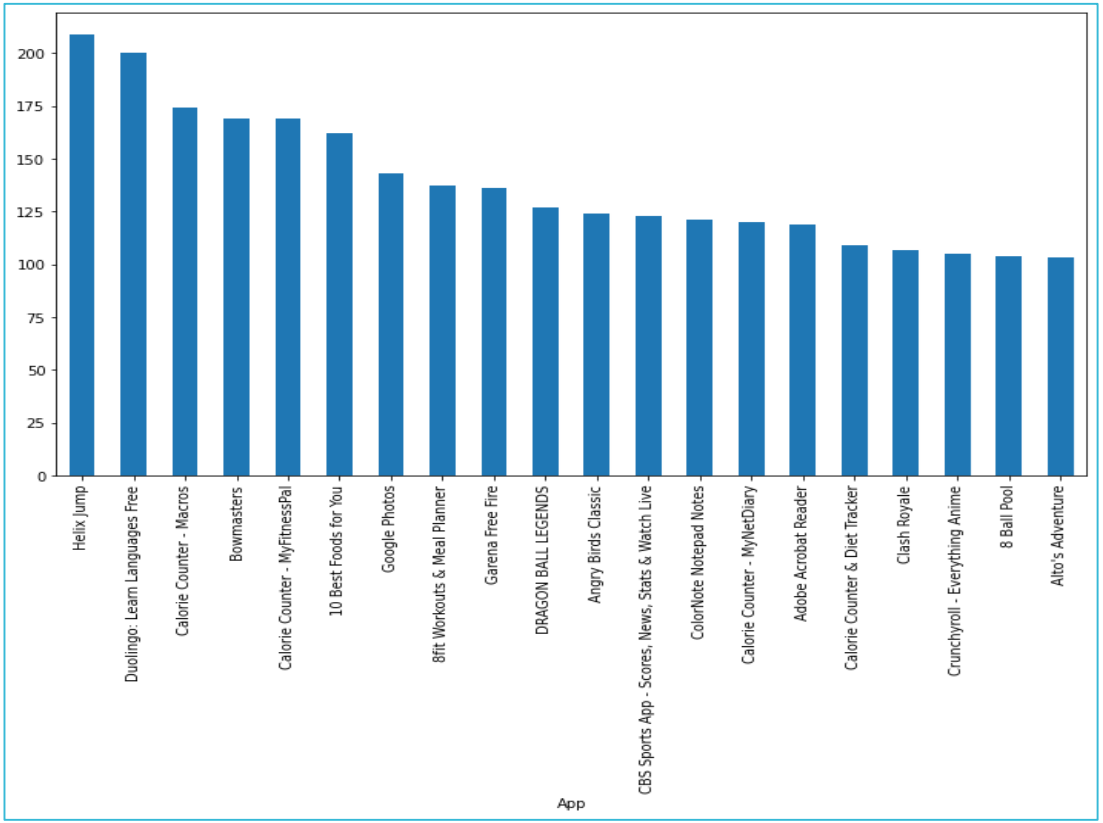
- There were numerous apps in the data. Let's see the Top 20 Apps with the most number of user reviews in our dataset are -



The top 3 most reviewed apps are –

1. Bowmasters
2. Angry Birds Classic
3. Helix Jump

- ***Visualization*** :
  - Top 20 apps with the most positive sentiment or positive reviews in our dataset are -



Top 3 apps with most positive reviews were –

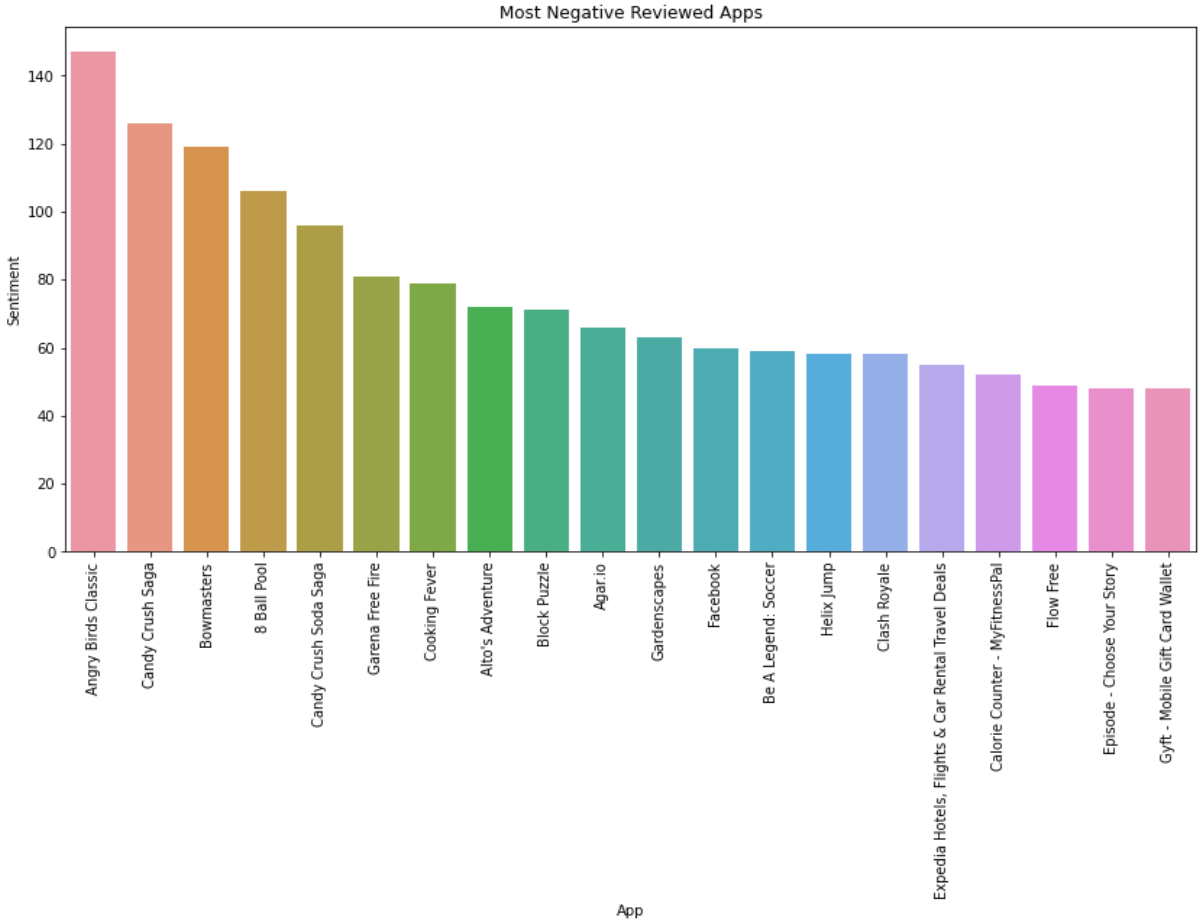
1. Helix Jump
2. Duolingo: Learn Languages Free
3. Calorie Counter - Macros

- *Visualization* :

Apps that received the most negative reviews were –

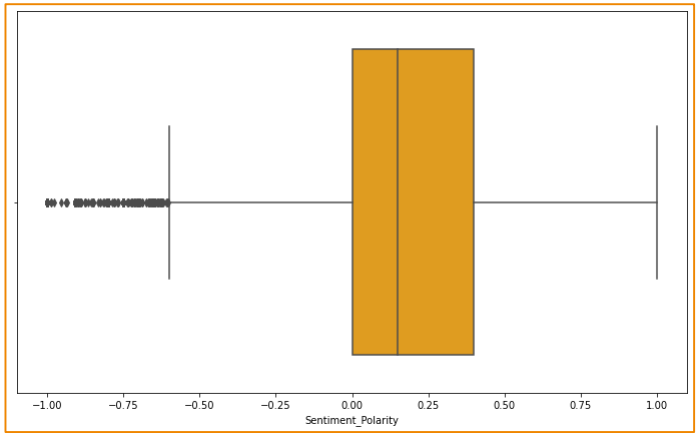
1. Angry Birds Classic
2. Candy Crush Saga
3. Bowmasters

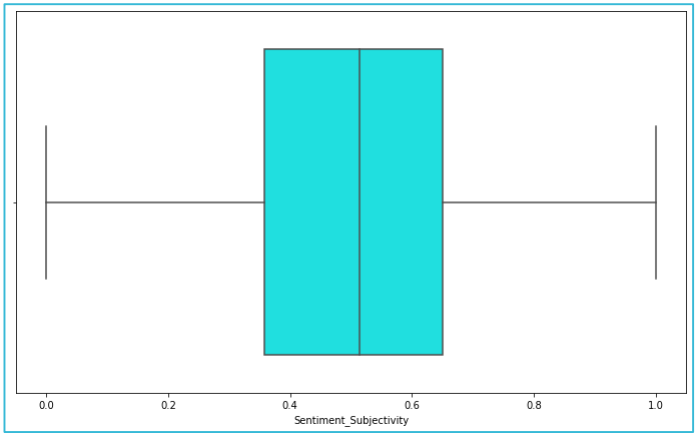
These apps did not satisfy the users much.



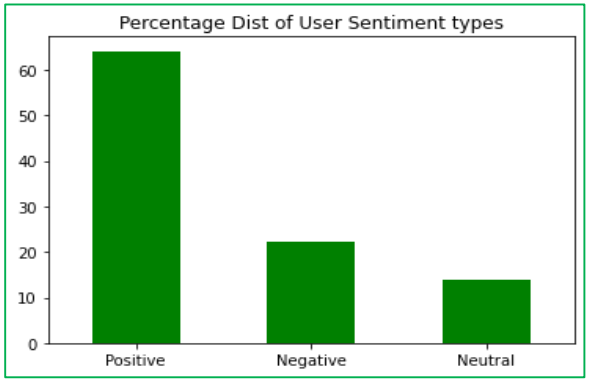
- **Visualization**

Sentiment Polarity & Sentiment Subjectivity :





Sentiments of users in percentage :



Model Implementation

- *Our next goal was to see if we can build a basic model that will take user reviews as an input and spit out the sentiment of that user regarding the product (in this case, App).*

Model Selection : *So, as per requirement, we went for the Supervised ML model that is the Sentiment Analysis model.*

Feature selection : *Translated Review column was chosen as the feature and the Sentiment column as dependent variable.*

Processing : *Data cleaning, text pre-processing was done. We cleansed the feature by removing punctuations and stop words. Next, data splitting was done for training and testing, then we finally fit the model.*

Results : *The model gave us approx. 90% accuracy. Testing with some random values, the results were satisfactory enough for a very basic model.*

Steps & Discoveries :

Exploratory Data Analysis is the most important step that needs to be performed for every dataset that we gather, regardless of the source of the data.

Depending on the data, this can be a lengthy process consisting of multiple steps.

It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand, before diving in deep with models and automations. Through EDA, we perform initial investigations to find patterns, spot anomalies and to check assumptions with the help of summary statistics and graphical representations.

For this EDA project, we used Python Pandas and Numpy libraries for computation and manipulation of the datasets. Matplotlib and Seaborn libraries were used for visualizations.

As we had two datasets with different kinds of features, we needed to have different approaches and follow different steps to draw conclusions from each data. We went through the datasets one-by-one to summarize the characteristics of the data and draw meaningful insights.

Steps & Discoveries :

- *EDA on Play Store dataset* : For this dataset, we followed the following process of –

- *Summarizing the data - to get the information about columns, data types, missing values and outliers etc.*

- *Treating the Outlier – There was only one outlier. So in this case, we removed it.*

- *Filling in the null values – This dataset having different kinds of data types for each column we had to choose the right central tendency very carefully for filling in the values.*

- *Selecting Features - We chose 'Category', 'Rating', 'Reviews', 'Installs', 'Size', 'Content Rating' and 'Price' columns for summary statistics and data visualization.*

- *Appropriating the data type – We changed the data type from string to numeric for these following columns - 'Rating', 'Reviews', 'Installs', 'Size' & 'Price'. This step was crucial for performing any computational and visualization process on these columns.*

Steps & Discoveries :

All these process left us with nice and tidy dataset consisting meaningful values with desired data types.

Then performing various computation, statistics method and visualization process, we tried to come up with answers of the following questions –

➢ *What was the average rating for each of our app categories?*

- *Average ratings for majority of the app categories, lie between 4.1 to 4.3*

➢ *From the ratings, what can we tell about which apps did well & which apps didn't?*

- *Events, Education, Art and design apps in our play store data had the highest average ratings. Whereas Dating apps category was the lowest in terms of ratings, which tells us user experience for these apps have not been great.*

➢ *What is the size distribution for apps in our data?*

- *Our data has high density for size between 0-30 MB. There were some apps who were close to 100 MBs or so.*

➢ *Which Categories have the most number of apps in our data?*

- *Most number of apps were from Family category , followed by Games & Tools respectively.*

Steps & Discoveries :

- *Which categories have the most number of reviews?*
- *App categories belonging to Communication and Social got the most number of reviews. The 3rd most reviewed app category is Gaming. Rating for these columns are somewhere between 4.15 to 4.30*

- *Most popular apps?*
- *Communication and Social apps have been installed a lot more than other app categories. Probably due to quarantine and work from home situation.*

- *Which apps charged most prices ?*
- *Finance, Family, Lifestyle and Medical are the top 4 categories in our dataset that stand out.*

- *How app prices might have affected installations?*
- *As expected, free applications were installed and used by more users than paid apps.*

- *What was the Content Ratings distribution for app categories in the dataset?*
- *All the app categories in the data had apps which were labelled - 'Everyone'*
- *'Teen' apps are another popular Content Rating type.*
- *Comics is the only category which have noticeable amount of applications available for Adults.*

Steps & Discoveries :

With that, we have managed to understand –

- ❑ What kind of apps are more popular among users.
- ❑ For which applications more prices can be charged.
- ❑ Different areas of content rating that could use more focus.
- ❑ Standards set by other applications in a certain category etc.

Now moving on to the 2nd dataset ---

- *EDA on Apps Review dataset* : For this dataset, we followed the following steps and came to following conclusions –

- *Data Summary : Loading the dataset in our dataframe, we checked the summary to realize the shape, data types, variables, column values etc.*
- *Data Cleaning & Filling in null values : After checking for null values, we found that more than 25000 Rows in this data were only having one column value and missing the rest of all the values.*

Cases like this, we can't just put central tendency values. Because the data being related to user sentiment and reviews, we can not determine what the value should be for such large no. of rows . And most probably these are just corrupted data-points. So, we and got rid of them.

Steps & Discoveries :

Next, we did some visualizations to see –

➢ *Which apps were the most reviewed in the dataset?*

- *We made visualization for top 20 apps with the most reviews. The top 3 of them are –*

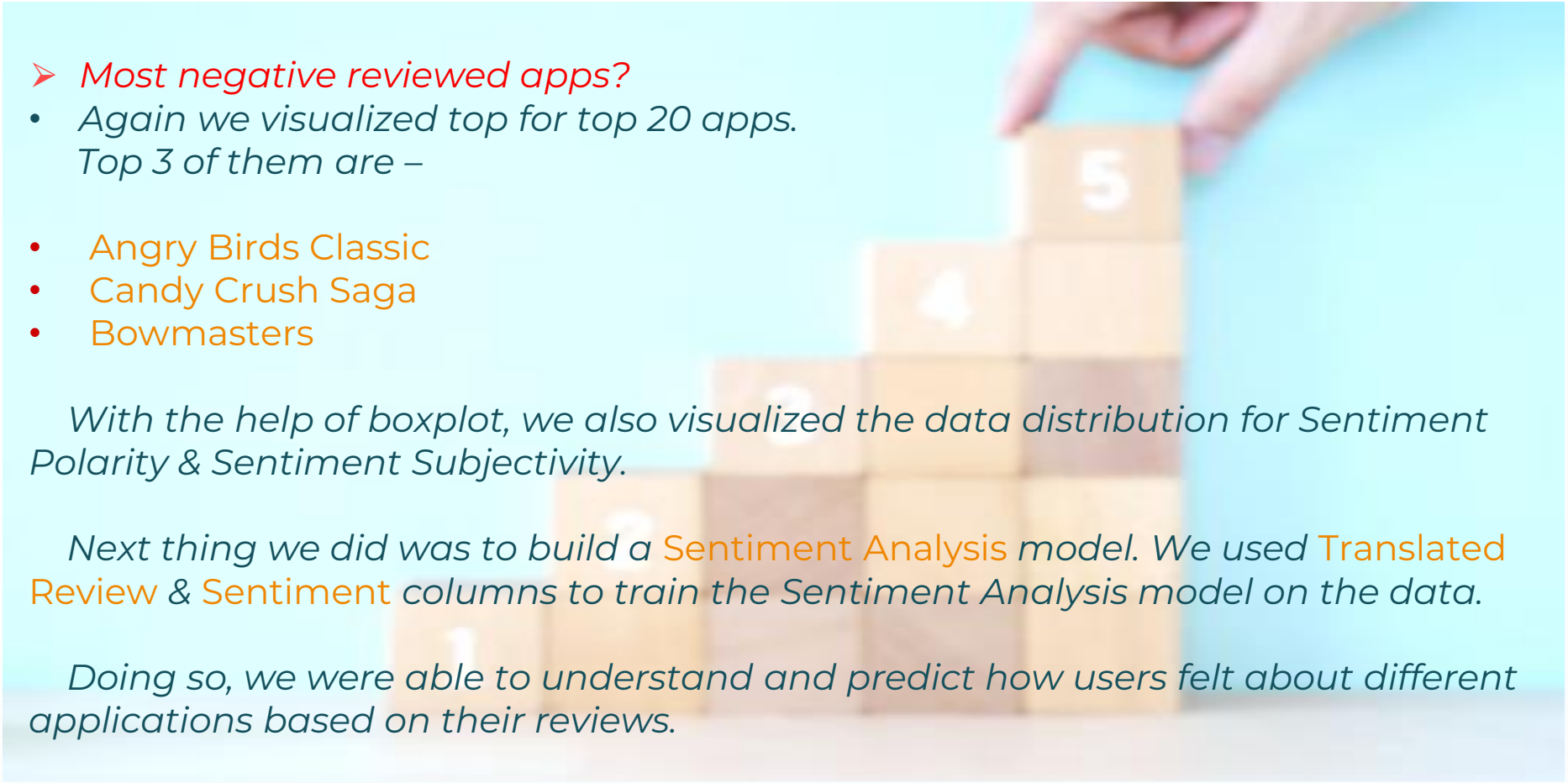
- Bowmasters
- Angry Birds Classic
- Helix Jump

➢ *Which apps received the most positive reviews?*

- *Again we visualized top for top 20 apps. Top 3 of them are –*

- Helix Jump
- Duolingo: Learn Languages Free
- Calorie Counter - Macros

Steps & Discoveries :



➢ *Most negative reviewed apps?*

- *Again we visualized top for top 20 apps. Top 3 of them are –*

- Angry Birds Classic
- Candy Crush Saga
- Bowmasters

With the help of boxplot, we also visualized the data distribution for Sentiment Polarity & Sentiment Subjectivity.

Next thing we did was to build a Sentiment Analysis model. We used Translated Review & Sentiment columns to train the Sentiment Analysis model on the data.

Doing so, we were able to understand and predict how users felt about different applications based on their reviews.

Conclusion :

In conclusion, we can say that these datasets have the potential of delivering with insights to better understand and process customer demands and their sentiments regarding the android apps, which can help in making better decisions in future for developers.

Analysing the actual ratings or popularity and comparing them with predicted results can tell us if the apps are performing as expected, better or worse compared to other apps.

Also, the dataset gives a clear indication of which apps are in demand, so that developers can popularize the app store with similar products.

The Sentiment Analysis model helps us to figure out users experiences, not only for given observations, but also for future observations.

Hopefully this exploratory data analysis will help to ease these processes.

Thank You...