PLAY STORE APP REVIEW ANALYSIS

ANKIT RAI
DATA SCIENCE TRAINEE
ALMABETTER

ABSTRACT:

Play Store is an "app store", which allows users to browse and download applications, game, and other media onto their Android device. Users can leave their reviews and rating for the downloaded app in the play store, based on which Apps gets ranked in the Play Store. User review and rating also helps other users in understanding about the given app and in making decisions for downloading.

We have got an interesting task to analyse the play store apps data and discover the key factors responsible for the app engagement and its success. For this task we have been provided with two datasets namely – ***Play store Data*** and ***User Review*** Data. Our objective is to analyse each dataset and to take useful insight from them.

PROBLEM STATEMENT

The success of an app depends on various factors. The one of the objective for this exercise is to study about the already performing top factors which contributes to app's success. Further, we are required to draw correlations between various features of a given app, so that while designing the App our designers can take them into consideration.

ABOUT THE DATA

We have been provided two datasets – Play store and User Review.

The first Play store dataset is much informative as it gives much more information about the app.Following are the labels of the columns and their meanings.

- **APP -** Simply talks about the App name.
- **CATEGORY –** Talks about the 'Category' to which the given app belongs. An app "category" makes it easier for users to discover new apps. For example, a user could click on the "Lifestyle" category in the App Store to discover new Lifestyle related apps.
- **RATING –** Rating gives the mean average rating of all the ratings given by the users in their reviews. The Rating varies from 1 to 5 (all integers). 1 being the lowest while 5 being the highest.
- **REVIEWS –** Number of *Reviews* given by all the users in play store for the given app. This is important as it gives the reliability to the *Rating* column values. Larger the *Reviews* numbers, more reliable are *Rating* values.
- **SIZE :** Size of the App

- **INSTALLS:** Number of installations done for the given app. This is an important information as it represents the success of a given app. Higher installation numbers in dicates higher acceptance of that app in the market.

- **TYPE:** Whether the App is free or paid

- **PRICE:** Tells the price for paid apps.

- **CONTENT RATING COLUMN:** These ratings talks about the age group of people the given app is for.

- **GENRES:** Genre to which a given app belong.

- **LAST UPDATED:** When was the last time App got its updates

- **CURRENT VER:** Current version of the app

- **ANDROID VER:** Android Ver which is compatible for the app.

The Second datatset ***User Review*** talks mostly about the comments given by the user for an app and **the sentiment associated** with that comment. This sentiment again is an important k ey factor for deciding the success of an app.

STEPS INVOLVED:

- **DATA EXPLORATION:**
  After loading the dataset, the first thing that we did was primary exploration of our data. We first checked about the top and bottom five rows to get an idea of our dataset. We looked for our dataset size and shape, various columns and a summarise information about datatypes present. While exploring, we found that there were data which would be required to change in numerical type for to analyse. For E.g., *Reviews* column had data in object type, while since those data represents numerical values, they were supposed to be in integer type for analysation. Also, many columns had null values in it. By the end of our primary data exploration, we have understood that we need to first get our data cleaned to analyse.

- **DATA CLEANING:**
  - **DUPLICATES:**
    After buckling our belt for data cleaning, the first thing we did was to search for any duplicate rows. Having two or more records for the same app can have a significant impact on the efficiency and performance of our analysing techniques. Hence, after searching for the duplicate values, we eliminated them from the dataset.

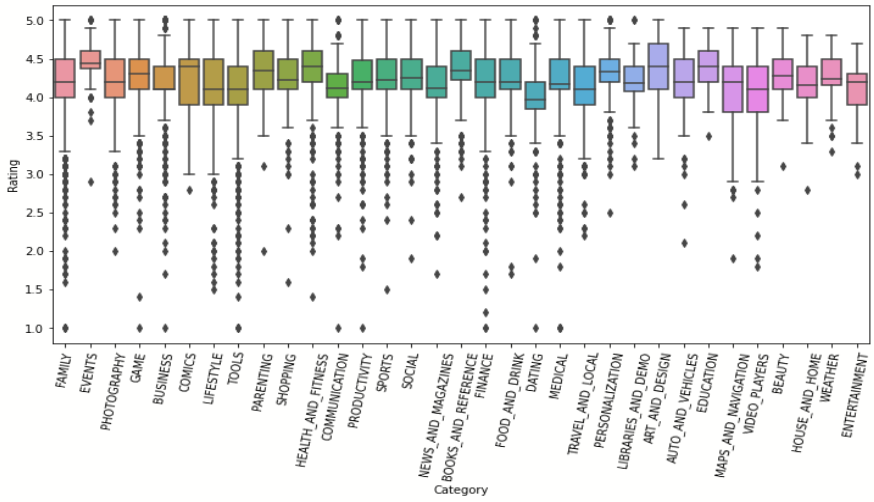
  - **NULL VALUES TREATMETNT:**
    Null values do not contribute any good in the analysation process and are used to indicate that you could have a value, but you don't know what that value should be yet. In our data summary we had found that there were five columns where null values were present -Rating, Type, Content Rating, Current Ver and Android Ver.

We had gone through column wise for the null values and replaced them with the median values of their respective categories.

- **CHANGING DATATYPES:**

Next challenge was to change object type datatypes to integer or float type depending on the column.
Sizes of the App were given in MB and kB and are of object data type. Our task was to replace the suffixes (MB & kB) and change the object data type to float type. For this we defined a function and applied it to the Size column. Similarly, we defined other functions to change other columns data.

- **HANDLING OUTLIERS**



Outliers are unusual values in your dataset, and they can distort statistical analyses and violate their assumptions. Outliers increase the variability in your data, which decreases statistical power. Consequently, excluding outliers can cause your results to become statistically significant.
To check for outliers, we plotted distribution graph for Install column and found that it had heavy Kurtosis (kurtosis measures extreme values in either tail.), which meant presence of lot of outliers. Similarly, by box plot also we could see that there were large number of outliers across various columns. For simplicity, we defined a function to detect an outlier. This function was able to detect any data which comes between 75percentile to 100 percentile range and between 0 percentile to 25 percentile range. After identifying these data, our function was designed to change the values to the nearest limit value (either 75 percentile or 25 percentile). This made the data statistically significant.

ANALYSIS OF DATA SET

For analysing top performing apps, categories, genres etc. we needed a strong basis on which these different variables can be evaluated. For example, if one has to find about top performing categories, then what parameters to be set to identify them. Going through the details provided we identified three important factors which indicates popularity or success for a given app. These three factors were:

- ***Number of Installs*** **-** This is one important objective for any app i.e. to get accessed through maximum number of users. But just getting numbers is never enough. Numbers with customer satisfaction is the goal. And there comes the role of ratings.

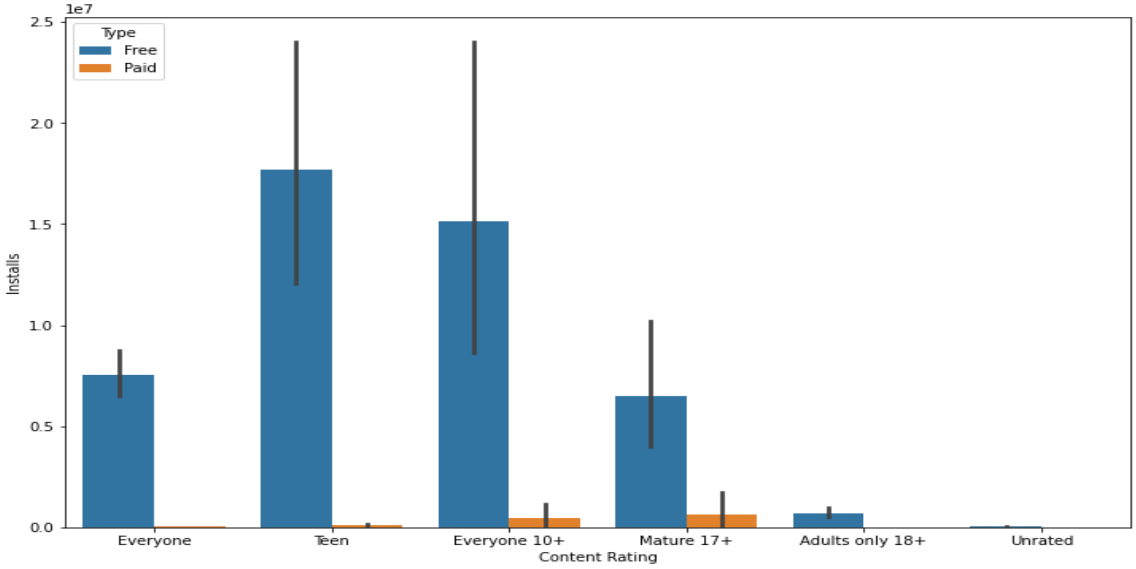
- ***Ratings*** **–** This represents the sentiment/ satisfaction level of the users for the app.

- ***Number of Reviews*** **–** Ratings are generally the average value of all the ratings given by the users. Hence, for lower numbers of feedback, ratings can be unreliable. Larger number of reviews represents the sentiments of large number of people which gives reliability to the ratings value.

That is why for every top performing variable, we considered above three factors for analysation. We looked features with highest number of installs and with highest ratings with respect to minimum certain number of reviews. Based on this method we found following top 3 results:

Merit Rank	Categories	Genres	Content Rating
1	Communication	Adventure; Education	Teen
2	Photography	Arcade	Mature 17+
3	Video_Players	Photography	Everyone 10+

We used bar charts to compare different values with their numbers in play store. Like when we wanted to compare different content-ratings based on their numbers in play store we plotted below bar chart.



To show percentage distribution of Free and Paid apps we used pie chart. These visualization tools helped to us give a clear vision for the further approach.

In User Review dataset, we analysed top apps based on their positive sentiment. More the positive sentiments number indicated higher engageability of the given app. But again, filtering apps just based on number of positive sentiments was not enough and inappropriate. This is because an app could have large number of comments with positive sentiments at the same time it may have large number of comments with negative sentiments as well. Hence, the requirement was to define a ***key factor*** which would indicate the extent of positive sentiment strongly. For this we defined a new variable called – Sentiment Ration. This was the ration of number of Positive sentiments to the number of negative sentiments. Higher the ratio, stronger the positive sentiments.

ANALYSIS OF MERGED DATA SET

Finally, we merged our two datasets, and analysed all important variables like- App, Categories, Genres e.t.c based on sentiment ratio. For the reliability of the sentiment ration, we ensured that the minimum reviews considered were 1000.

CONCLUSION

With the help of Analysation and Visualization tools finally we reached to the end of our EDA. We have gone through both datasets and made them ready for the analysation. Later we performed EDA by statistical measures and using visualizing tools. We have provided all important top performing features like - Category, Genres, Content Rating, Size, and type which is required for the development of App. We also discovered a new parameter called 'Sentiment Ratio' to extract apps which are having most comments of positive sentiment than negative sentiments.