An Exploratory Data Analysis of

Google Play Store App Reviews

Sanjeev Hegde
Data science Trainee, AlmaBetter

Abstract: Android is a mobile operating system loved by millions across the globe. Android devices are popular due to their user-friendly features and millions of apps hosted on Google Play store. Play Store is the official platform wherein a user can download, rate and review the downloaded application(s). An analysis performed based on user reviews for the applications on the play store can provide countless insights in development of applications. In this regard, an Exploratory Data Analysis has been performed on two sets of given data with various parameters using Python. From the analysis, category of apps containing the most number of apps, category of apps downloaded the most, category wise rating for apps on play store, how reviews effect user preference of one app over other, how subjective and polarised user reviews are, how paid apps are reviewed and several other questions have been answered.

Keywords: Play store, Exploratory Data Analysis, Python, Data cleaning, Data visualization, Sentiment analysis

Problem statement: To perform Exploratory Data Analysis on given data sets containing information about apps on Google Play store thereby providing valuable insights based on various parameters

Introduction: Play store which is the official app store for android devices has been developed and operated by Google LLC since the year 2008. Play store hosts numerous services apart from applications such as music, books, movies and other television content. Applications on Play store are either free, charged or can contain in-app purchases through google play balance. By the year 2017, over 3 million apps were hosted on the play store. Key features of play store include features such as secure applications, application rating for specific users such as adults, kids etc, User ratings and reviews wherein user can rate and review downloaded applications based on his experience, Application size, description and data policy of developers enabling transparency and beta programs wherein a user can register for trying out apps in developmental stages. It is noteworthy that with above features, play store contains humongous data for apps as experienced by the end users which can serve as torch bearer for application developers post proper analysis. As the objective of the project, we hereby provide meaningful insights into the given two sets of data such as categorical implications on applications, user review and ratings impacting installs, sentiment analysis of apps etc are discussed in the further details of this technical paper with the help of Python programming language.

1. ***The data sets:*** The objective required analysis of two data sets with the following details.:

1.1 Play store Data: Among the two files analysed, Play store Data file contained numerical data along with categorical details of the apps. The data set had 10,841 entries against 13 parameters. The list of parameters along with their details are as below:

- **App:** The name of the application to which data was provided in the data set
- **Category:** The category of to which given application belonged to. Includes Family, Games, Communication etc.
- **Rating:** Average user rating for the application
- **Reviews:** Total number of reviews the app has gained since its launch
- **Size:** Size of the application
- **Installs:** total number of installs for the app
- **Type:** Entries include Free and Priced accordingly for the given app
- **Price:** Price of the application
- **Content Rating:** The type of users whom the application is meant for. Examples include teen, adult, kids, everyone etc.
- **Genres:** The genre which application belongs to
- **Last Updated:** The last **date** of updating the app by the developer
- **Current Ver:** Present Version of the application
- **Android Ver:** Minimum android version required to run the application

- ***1.2 User Reviews:*** 2nd data set analysed contained partially processed data from user reviews. This data set consisted 64,295 entries against 5 parameters on which sentiment analysis was performed. The parameters include:
- **App:** The name of the application to which data is provided
- **Translated Review:** This column contains NLP processed version of the user reviews for given application by each user.
- **Sentiment:** Sentiment of the review is either positive or negative with the app summarizing his experience with the app.
- **Sentiment Polarity:** A quantifying measure of the Sentiment in each user's review. Value varies from -1 to +1 wherein negative number closer to -1 indicates highly negative review and vice-versa
- **Sentiment Subjectivity:** A quantitative measure of the subjectivity of a user's review. Value varies from 0 to 1 wherein number closer to zero indicate an objective review type of review which is more of a fact-based review and numbers closer to 1 indicate subjective review.

2 ***Data Cleaning and Preparation:*** Given data sets were cleaned and prepared for performing Exploratory Data Analysis using Python programming language. The process included:

2.1 Modification of the Play store review data set which consisted of a skewed row. The row containing such data was appropriately aligned using methods in Pandas library in Python.

2.2 Null values in numerical entries were replaced with Median values to avoid them being affected by extreme values

2.3 Null values in non-numerical data were replaced by the mode values for the data set thereby ensuring no effect of any extreme or abnormal values.

2.4 Unit conversion was achieved wherever necessary such is size in columns wherein size of each app was converted into Kilobytes (Kb).

2.5 Non-essential characters in numerical data such as commas, "+" and "$" were eliminated.

2.6 Every data was assigned appropriate data type to ensure precise results

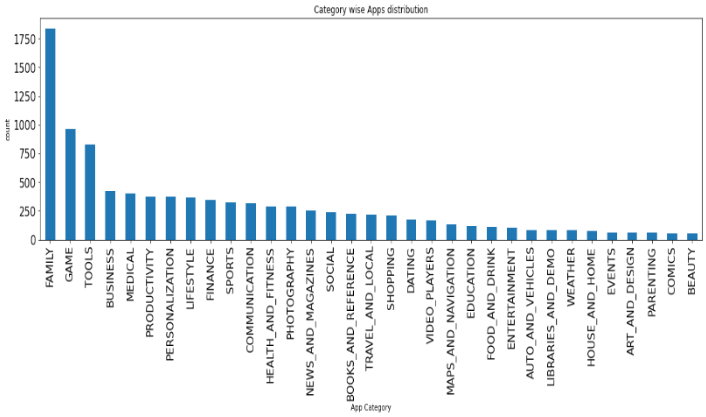
2.7 Non-essential parameters such as Translated review which consisted of processed words of user's review were dropped from the analysed data set to facilitate better memory usage

2.8 Duplicate entries in the data were removed based on the app name and Last updated date leaving the data set with 9701 unique entries.

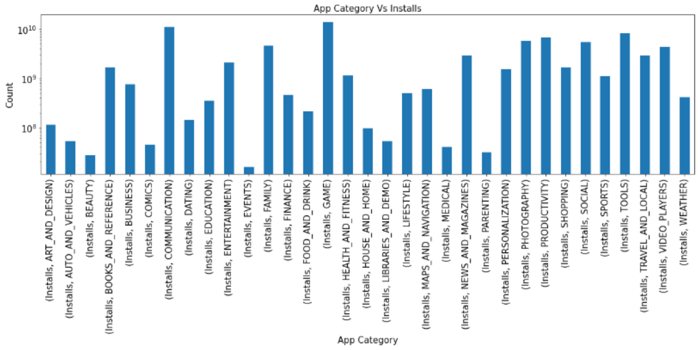
3. <u>Exploratory Data Analysis:</u> The next step in data analysis is to graphically or visually represent the data from the 2 processed data sets. This is mainly done with the help of "matplotlib" and "seaborn" libraries in Python. The findings from our analysis are summarized below:

3.1 Statistical Overview of the data sets: There are total of 9,701 unique apps in the data set against which various parameters are provided. Average ratings for apps in data sets is about 4.2 stars while the average size is about 16MB. Largest app is 100MB in Size while smallest app is 1Kb. Costliest app is priced at $400. Standard deviation for ratings from mean is about 0.49 stars.

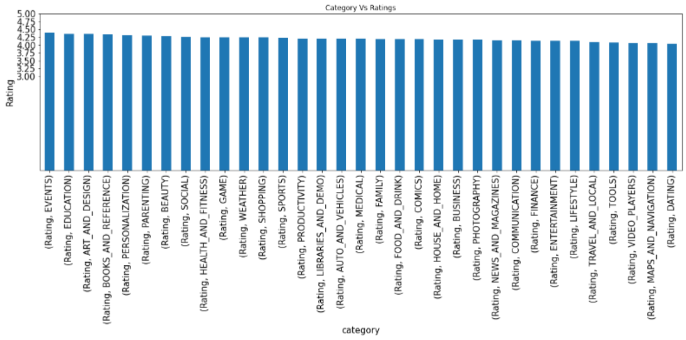
3.2 Category wise app distribution: Among the apps in the given data sets, highest number of them belong to Family category followed by Games and Tools.



3.3 Category and Installs: Among the apps given in data sets, it is found that maximum number of users have installed apps belonging to Games category followed by apps belonging to Communication category.



3.4 Category and Rating: Category with highest rating is Event category with 4.4 stars while it is to be noted that all category of apps have average rating above 4 stars.

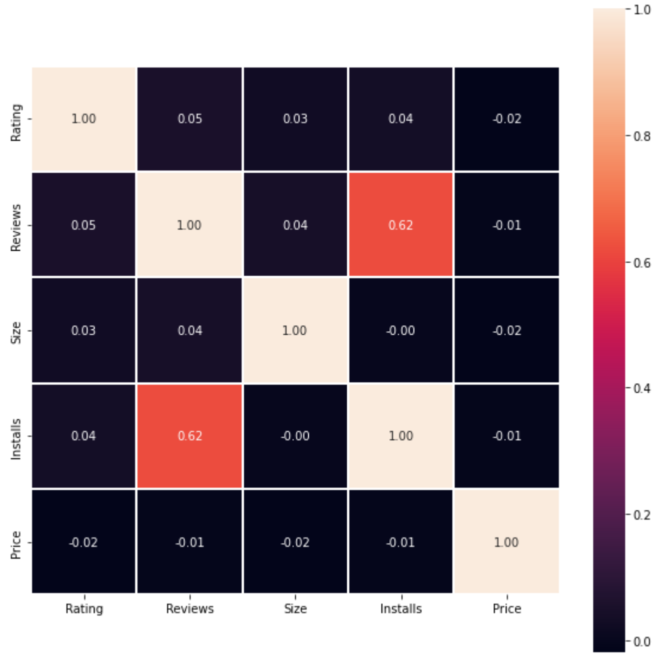


3.5 Apps with Highest Rating: There are 27 apps with 5 stars rating among the given apps and top app with 5 stars rating is "Ek Bander Ne Kholi Dukan" belonging to Family category and Entertainment Genre with 10,000 plus installs.

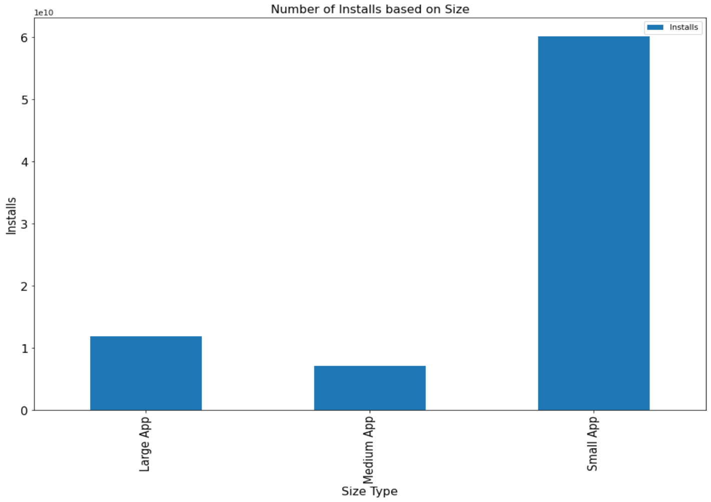
3.6 Top 5 apps based on installs: Top Installed apps have been downloaded more than 10^9 times which includes apps such as Facebook, WhatsApp, Instagram, Subway Surfers etc.

3.7 Installs by Pricing: It is clear from the analysis that free apps were mostly downloaded as compared to paid apps wherein free apps had 79,16,18,73,646 installs in total compared to paid apps which were installed only 5,73,63,881 times. Also, costliest app is "I'm Rich - Trump Edition" with price of 400$ and 3.6 stars rating with more than 10,000 installs.

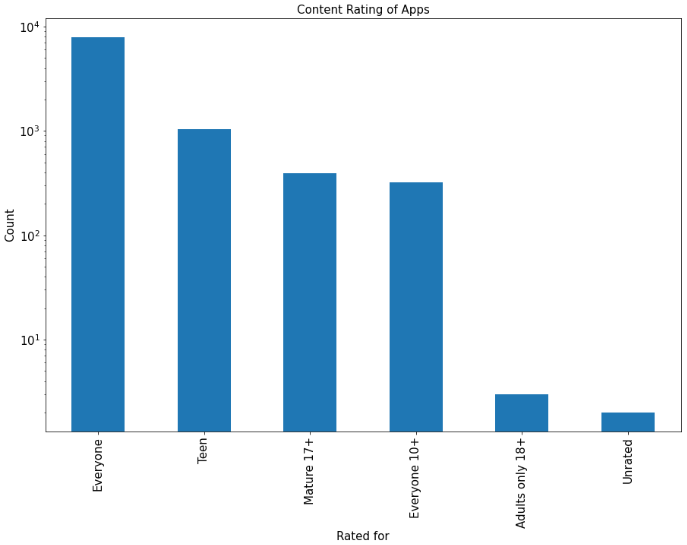
3.8 Number of reviews and installs: Correlation heatmap plotted using "seaborn" library reveals an important relation between number of reviews and installs wherein correlation coefficient of 0.62 exists between the two parameters. This implies that users have a tendency to install apps with more installs compared to other apps. The heatmap is represented below:



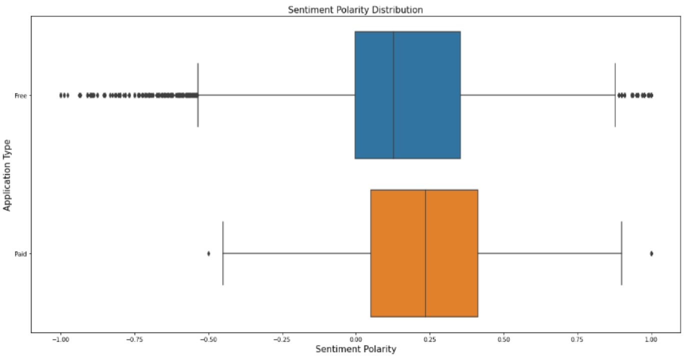
3.9 App size and Installs: Size of the apps were categories into 3 types based on the range of their size wherein apps with size less than 25MB were considered to be small apps, apps with size between 25MB and less than 50MB were considered to be medium apps while apps with size more than or equal to 50MB were considered to be large apps. From the analysis it is found that Small apps have highest number of installs of 60,18,44,04,390 while Large apps were second mostly installed apps with 11,90,39,84,833 installs and medium apps were installed 7,13,08,48,304 times.



4.0 Content Rating and Apps: In the given data sets, it is found that 7937 apps are rated for Everyone. While Teen rated apps are 1042 in number.

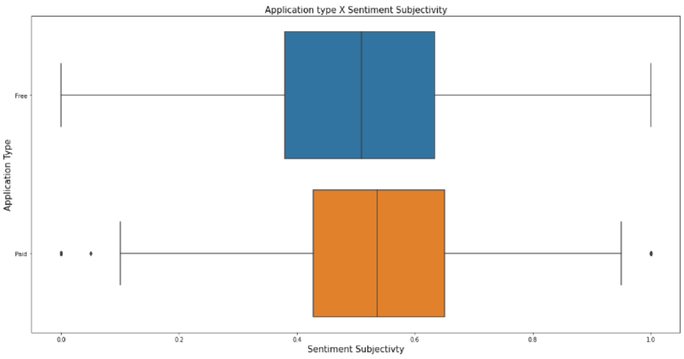


4.1 Analysis of User Sentiment: User Sentiment Analysis plays a key role in understanding the shortcomings in a particular segment wherein an application is developed. In this regard, a box plot is plotted for the sentiment polarity to understand how majority of the apps are performing in the given data set.



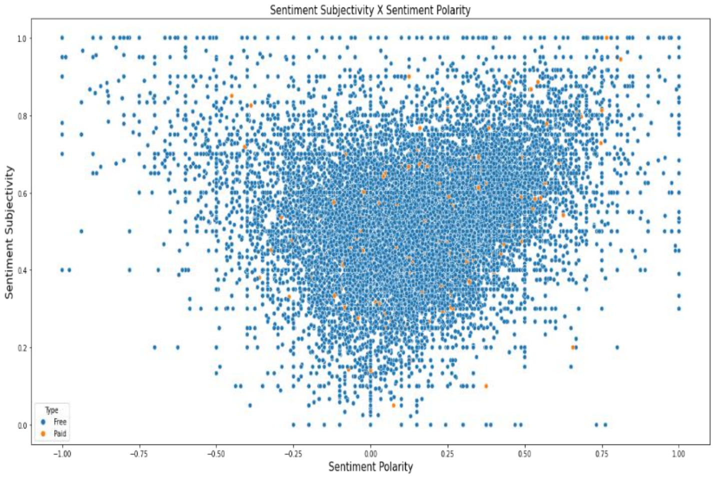
Thus, from the above plot, it can be inferred that user sentiment in free apps is slightly positive with the median around 0.15. Similarly paid app users tend to have reviewed apps in even more positive way with median being around 0.25.

4.2 Sentiment subjectivity: Sentiment subjectivity is a measure of the subjectiveness of a review in the given data set. A boxplot is plotted for visualizing the distribution of given app type with respect to sentiment subjectivity.



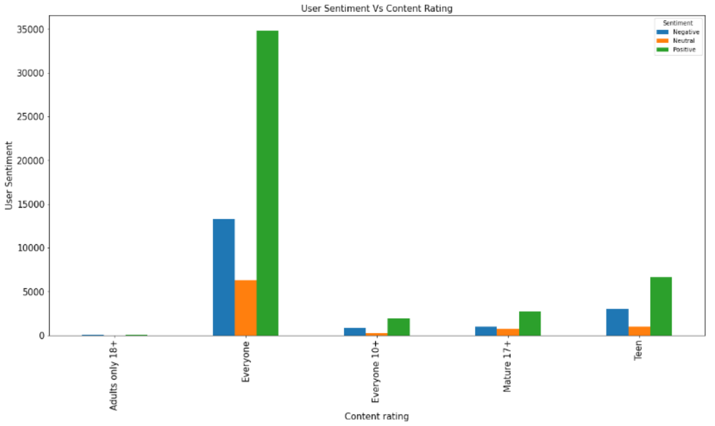
Hence, it can be inferred that free app reviews are relatively objective in comparison with paid apps whose reviews are slightly more subjective. Also, given user review data set contains few outsider values implying extremely objective or subjective reviews.

4.3 App distribution: The following plot gives us the distribution of applications with respect to sentiment polarity and sentiment subjectivity.



Since most apps lie in the quadrant with subjectivity>0.5 and polarity>0.5, it can be generalised that app reviews are subjective and positive in nature in comparison with other subjectivity and polarity parameters considered together.

4.4 Content rating and Type of sentiment: With type of sentiment experienced by the age group an app is built for, an appropriate product market fit can be established. The following graph represents such relationship between the two afore mentioned parameters.



Hence, it can be inferred that apps developed for everyone have highest positive, negative and neutral reviews in comparison to apps designed for other users. Also, it can be concluded that most apps have gained positive user reviews in every segment.

Conclusion: After analysis of the two data sets, it can be concluded that most apps have positive user experience while mostly preferring apps with higher number of reviews, ratings and smaller size. Further, It can be said that developing application rated for everyone is likely to be downloaded more in comparison with other niche audience.

References:

- *https://github.com/*
- *https://stackoverflow.com/*
- *https://pandas.pydata.org/docs/*
- *https://www.almabetter.com/*
- *https://www.kaggle.com/*