**Google Playstore Exploratory Data analysis**

**Vartika Arora, Alma Better**

**Abstract:**

In today’s scenario we can see that mobile apps playing an important role in any individual’s life. It has been seen that the development of the mobile application advertise has an incredible effect on advanced innovation. Having said that, with the consistently developing versatile application showcase there is additionally an eminent ascent of portable application designers inevitably bringing about high as can be income by the worldwide portable application industry.

With enormous challenge from everywhere throughout the globe, it is basic for a designer to realize that he is continuing in the right heading. To hold this income and their place in the market the application designers may need to figure out how to stick into their present position. The Google Play Store is observed to be the biggest application platform. It has been seen that although it creates more than two-fold the downloads than the Apple App Store yet makes just a large portion of the cash contrasted with the App Store. In this way, I scratched information from the Play Store to direct our examination on it.

**1. Introduction**

**The Google play store** is one of the largest and most popular Android app stores. It has an

enormous amount of data that can be used to make an optimal model. The data set used in this project contains 13 different features that

can be used for predicting whether an app will be successful or not using different features.

This data set is scraped from the Google Play Store. This journal also gives detailed information on feature extraction and the complete Data visualization done on this data set.

**Exploratory Data Analysis (EDA)** is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. Applied early on in the data analytics process, EDA can help us learn a great deal about a dataset’s inherent attributes and properties.

**2. Problem Statement**

Objective is to analyze Google playstore data and do a comprehensive analysis of the Android app market by comparing apps in Google Play across different categories and genres. We'll look for insights in the data to devise strategies to drive growth and retention.

**3. Data Preparation**

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data.

**3.1 Gathering data**

This step is about getting to know the data and understanding what has to be done before the data becomes useful in a particular context. This can be done by reading the CSV file and doing initial statistical analysis.

Though the dataset may seem to have the correct data types for each column, we need to check it. Inconsistent data types will create issues while dealing with problems.

**The data consists of two files:**

**3.1.1 apps.csv:** This filecontains all the details of the applications on Google Play. There are 13 features that describe a given app.

Playstore data frame (apps) has 10841 rows and 13 columns. The 13 columns are identified as below:

* **App -** It tells us about the name of the application with a short description (optional).
* **Category –** It gives the category to the app.
* **Rating -** It contains the average rating the respective app received from its users.
* **Reviews -** It tells us about the total number of users who have given a review for the application.
* **Size -** It tells us about the size being occupied the application on the mobile phone.
* **Installs -** It tells us about the total number of installs/downloads for an application.
* **Type** - It states whether an app is free to use or paid.
* **Price -** It gives the price payable to install the app. For free type apps, the price is zero.
* **Content Rating -** It states whether or not an app is suitable for all age groups or not.
* **Genres -** It tells us about the various other categories to which an application can belong.
* **Last Updated -** It tells us about the when the application was updated.
* **Current Version -** It tells us about the current version of the application.
* **Android Version** - It tells us about the android version which can support the application on its platform.

**3.2 user\_reviews.csv:** This filecontains 100 reviews for each app. The ranking is based on the contribution to the sentiment analysis. The text in each review has been pre-processed and attributed with three new features:

Sentiment (Positive, Negative or Neutral), Sentiment Polarity and Sentiment Subjectivity.

* User reviews data frame has 64295 rows and 5 columns. The 5 columns are identified as follows:
* **App:** Contains the name of the app with a short description (optional).
* **Translated Review:** It contains the English translation of the review dropped by the user of the app.
* **Sentiment:** It gives the attitude/emotion of the writer. It can be ‘Positive’, ‘Negative’, or ‘Neutral’.
* **Sentiment Polarity:**It gives the polarity of the review. Its range is [-1,1], where 1 means ‘Positive statement’ and -1 means a ‘Negative statement’.
* **Sentiment Subjectivity:** This value gives how close a reviewer’s opinion is to the opinion of the general public. Its range is [0,1]. Higher the subjectivity, closer is the reviewer’s opinion to the opinion of the general public, and lower subjectivity indicates the review is more of a factual information.

**3.2 Data Cleaning**

This step is crucial for removing faulty data and filling in null values and treating outliers. Important task here includes:

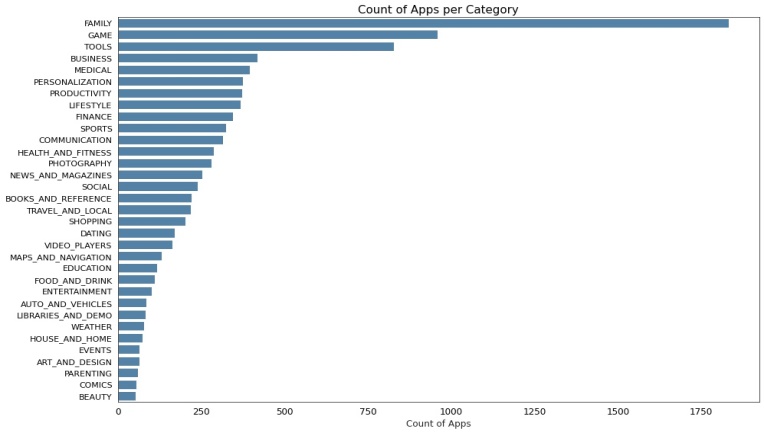
* Removing extraneous data
* Filling in missing/null values.
* Conforming data to a standardized pattern.
* Removing duplicate values
* Treating outliers

1. **We have only one column with numerical values, so we cleaned and changed the data type for Price, Installs and size column in order to analyze the data and draw some conclusions.**
2. Size column contains data with different units. 'M' stands for MB and 'k' stands for KB. To easily analyze this column, it is necessary to convert all the values to a single unit. In this case, we converted all the units to MB.
3. To convert all the values in the Installs column from string data type to integer data type, we dropped the '+' and ',' symbol from all the entries if present and then we can change its data type.
4. To convert the price column from string to float, we dropped the $ symbol from the all the values. Then we can assign float data type to those values.
5. App rating should be between 1, 5 any value above or below will be considered as an outlier and hence were removed from the main dataset.
6. Replaced null values with mean values.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**4. Data Analysis**

**4.1 Category wise distribution of apps**

**Fig 4.1**

The top 5 categories that majorly contribute to the playstore apps are Family, Games, Tools, Medical and Business category.

Whereas, Beauty, Comics, Parenting, Art and Design, and Event categories have the lowest share in the total playstore apps.

**4.2** **Top 100 Installed Apps**

Chart, bar chart, histogram

Description automatically generated

**Fig 4.2**

The top 100 most installed apps/popular apps are free apps majorly belongs to Communication, tools, games, social and family category.

**4.3 Percentage distribution of apps in different size groups**

Chart, pie chart

Description automatically generated

***Fig 4.3***

Around 44 % apps in playstore are very small or light apps, 32% small, 16.31 % medium and ~7% apps are heavy apps.

**4.4 Free vs. Paid Apps**

Chart, pie chart

Description automatically generated

***Fig 4.4***

There are 92.17% free and 7.83% paid apps in the playstore.

**4.4.1 Least expensive apps**

Chart, bar chart

Description automatically generated

***Fig 4.4.1***

In the paid category, A subset of 50 least expensive apps (mostly which were below $10.0 ) was observed . It is seen that top categories with least expensive paid apps are Tools, Family and Games.

And majority of these apps are very small in size.

**4.4.2 Most expensive apps**

Chart, bar chart

Description automatically generated

***Fig 4.4.2***

In the paid category, A subset of 50 most expensive apps (ranging b/w $30.0 to $400.0) was observed. It is seen that top categories with most expensive paid apps are Medical, Family and Finance.

Majority of most expensive apps are also very small in size.

**4.5 App size vs rating**

Chart, scatter chart

Description automatically generated

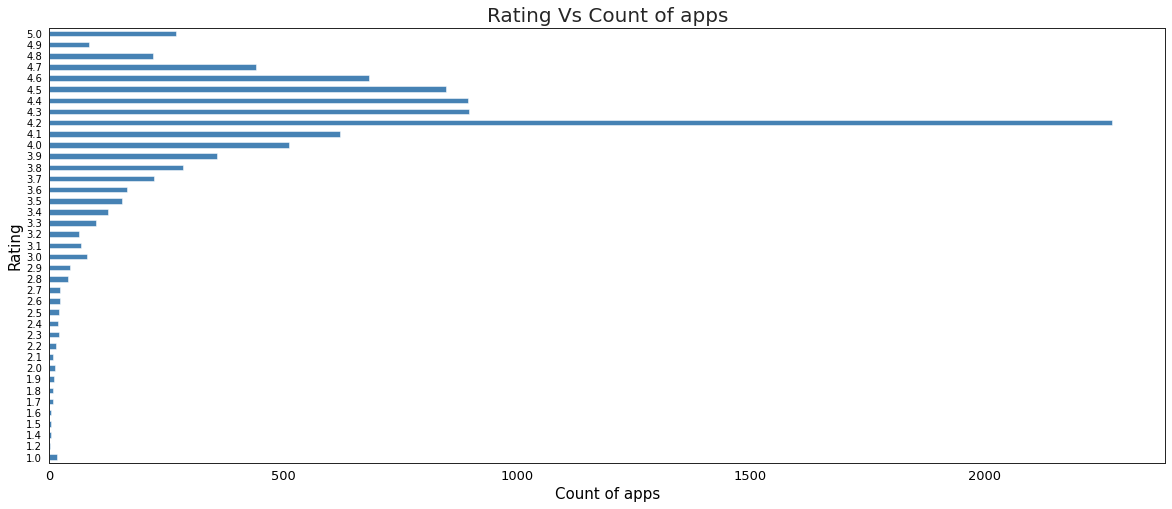
***Fig 4.5***

After observing the scatter plot, we see that majority of the apps are small in size and have good average rating.

On more close observation we can see apps with more data size mostly have good rating.

Hence, we can infer that the apps which have more data size have better performance, resulting in better user experience and Apps that are free also has positive impact on user's choice and preference.

**4.6 Rating vs count of apps**



Maximum apps have an average rating of 4.1 and majority apps have a good average rating (4.1-4.6).

80.23 % apps in playstore are top rated having average rating above and equal 4.0 and only 0.61% apps are lowest rated having average rating less than 2.0

**4.7 Percentage distribution of apps by content type**

Chart, pie chart

Description automatically generated

***Fig 4.7***

81.82 % apps in playstore is rated for everyone, 10.73% for teen , 4.07% for mature 17+, 3.33% for 10+ and 0.03 % rated only for adults 18+

A large portion of total apps available on playstore is rated for everyone and have a large and diverse user base.

**4**.**8 User reviews**

Chart, pie chart

Description automatically generated

***Fig 4.8***

From the above pie chart, we can say that most of the apps that are present on the play store has received positive review by the user while there are some apps which have negative reviews as well.

**4.8.1 Top 10 most reviewed apps**Chart, bar chart

Description automatically generated

**Fig 4.8.1**

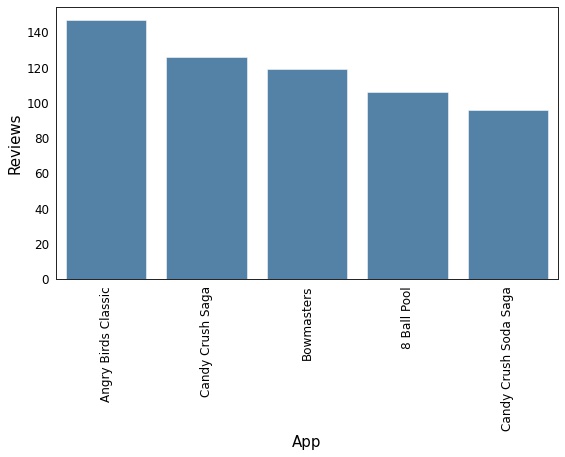
**4.8.2 Top 5 most positively reviewed apps**

Chart, bar chart

Description automatically generated

***Fig 4.8.2***

**4.8.3 Top 5 negatively reviewed apps**



***Fig 4.8.3***

* The top 5 most reviewed apps are Facebook, WhatsApp Messenger, Instagram, Messenger – Text and Video Chat for Free and Clash of Clans.
* Helix Jump has the highest number of positive reviews and Angry Birds Classic has the highest number of negative reviews.

1. **Conclusion**
2. All the top 100 apps installed are free apps and mostly belongs to communication, games and tools category. Hence, we can infer that if an app is free it will have larger and diverse user base.
3. We have observed that there is large user base of social, communication and games categories but there are not many apps available in these categories except for games. Hence, social and communication are the categories that can be tapped more.
4. If the app has content rating not for everyone, it will not have a large user base but if you make it more user centric targeting the selective audience the app will do well even if it is paid then.
5. Majority of the apps in playstore are small in size and have good/above average rating. While for paid apps, we have almost equal distribution in terms of size and rating.

On more close observation we can see apps with more data size are also given good rating.

Hence, we can infer that the apps which have more data size have better performance, resulting in better user experience and Apps that are free also has positive impact on user's choice and preference.

**6. Challenges**

* Examining and comprehending the problem statement, from business point of view.
* Handling the error, duplicate, null values and outliers in the dataset.
* To choose appropriate visualization for presenting data effectively. Making the charts more readable and comprehensive.
* Effectively communicating results.

**7. References**

* GeeksforGeeks
* <https://www.kaggle.com/>
* https://youtu.be/qR\_z4KlSKiA

**Github-**