Play Store App Review Analysis

Chetan Chaudhari
Data science trainee,
AlmaBetter, Bangalore

Abstract:

Android is an operating system for mobiles. Play Store is the official app store of google for android. Google Play store is an App store containing thousands of applications for Android mobiles which is useful in day-to-day life.

Our analysis can help us understand the play store application trends and features liked by the wide variety of users.

1.Problem Statement

The data provided by google playstore contains reviews, rating, sentiment, installation like number of installs and price of the paid application.

The main objective is to do an Exploratory Data Analysis (EDA), I will be doing Exploratory data analysis on this data set, which is a very important step in data science cycle, as it not only helps in taking very initial business decisions but also in preparing the data for further modelling for use in machine learning algorithms. Our objective will be to structure the data, clean it and present certain trends that I observe that can help us draw very preliminary conclusions about the probability of success of a newly launched app.

2. Introduction

The Google Play Store is a digital store for various types of apps like Game, Communication, Family etc.

People most commonly use the app to download gaming apps. However, the Play Store also sells e-books, TV shows, and movies. Each category also has its own standalone app (except apps and games), so you can browse that way as well.

The Google play store initially release on October 22, 2008 with android operating system. It received many UI updates over the years, along with additional content and functionality. It changed its name to the Play Store in March 2012 and it’s been that way ever since. While Google Play is strongly associated with Android, it is not a part of the stock Android experience. It’s actually an extra piece of software for Google’s specific Android experience. Thus, just because a device runs Android doesn’t mean it automatically has Google Play Store support. OEMs must adhere to a specific set of rules to get Google apps and the Play Store is part of that package. There are alternate app stores available for android as well.

2.1 <u>Google Play store Dataset</u>

The dataset consists of Google play store application reviews there rating pricing etc. information and is taken from Almabetter, which is the world’s largest community for data scientists to explore, analyze and share data.

The dataset consists of information about 10K application. In this data set I will examine various qualities like rating, free or paid and so forth utilizing Hive and after that I will likewise do forecast of various traits like client surveys, rating etc.

The data set contains the following columns:

- **App:** This Column contains the name of the app
- **Category:** This contains the category to which the app belongs. The category column contains 33 unique values.
- **Rating:** This column contains the average value of the individual rating the app has received on the play store. Individual rating values can vary between 0 to 5.
- **Reviews:** This column contains the number of people that have given their feedback for the app.
- **Size:** This column contains the size of the app i.e. The memory space that the app occupies on the device after installation.
- **Installs:** This column indicates the number of time that the app has been downloaded from the play store, these are approximate values and not absolute values.
- **Type:** This column contains only two values- free and paid. They indicate whether the user must pay money to install the app on their device or not.
- **Price:** For paid apps this column contains the price of the app, for free apps it contains the value 0.
- **Content Rating:** It indicates the targeted audience of the app and their age group.
- **Genre:** This column contains to which genre the app belongs to, genre can be considered as a sub division of Category.
- **Last updated:** This column contains the info about the date on which the last update for the app was launched.
- **Current version:** Contains information about the current version of the app available on the play store.
- **Android version:** Contains information about the version of the android OS on which the app can be installed.

2.2 User Review Dataset

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. How it works:

In the play store, I can find millions of applications on the single go. I choose an application and get it installed.
If the application is free, you can download it normally, but if the application is paid, you have to make a payment before getting it installed on your device.
Playstore
provides options for rating and reviews, for users to rate the applications for better engagement and recommendations for next time. Recommendations plays a major role in user engagement in playstore.

4. Types of Applications:

There are many types of applications in playstore, from free to paid application. Playstore is also categorized in many categories. They also have the option for age group restrictions on applications. Playstore has thousands of applications divided into a ton of categories.

5. Steps involved:

● Data Exploration

Firstly, I did data exploration by checking all the data frames and values associated with the rows. I encountered around 1500 null values in order to get more frequent results. It also consists of outlier in rating.

● Data Cleaning

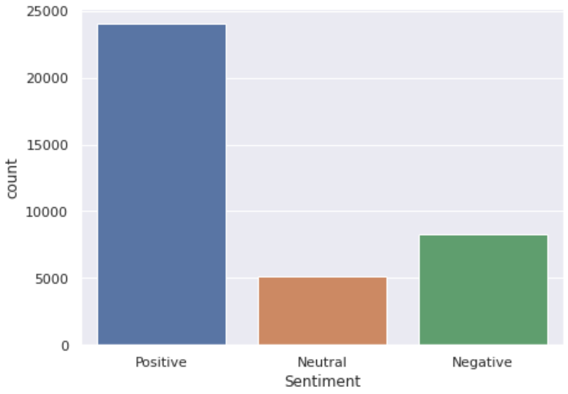
The available data is raw and unusable for Exploratory data analysis, so before I do anything with the data I will have to explore and clean it to prepare it for data analysis. So, I use fillna () function along with median value of rating to fill the null values (NaN) with median of rating. I also use >5 to remove the outlier from rating.
The remaining null values I replace with mode of their particular column.
I also remove the special character to make the clean.

● Exploratory Data Analysis

I started doing exploratory data analysis.
I found out about the apps which are free and paid. I also found out most installed category of the apps.
I tried to give conclusive analysis to developers so that this data could be far more useful to them. The overall

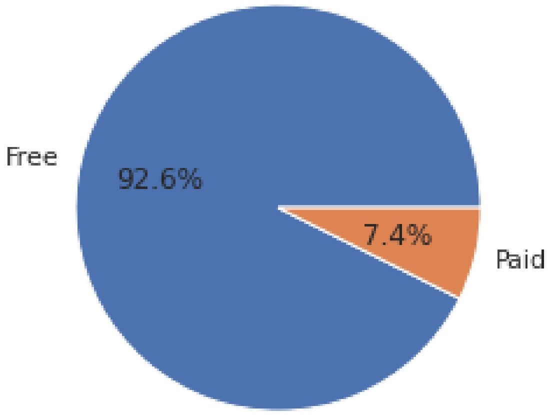
analysis revolves around comparisons of types in apps.

- **Count vs sentiments Graph for user review.**



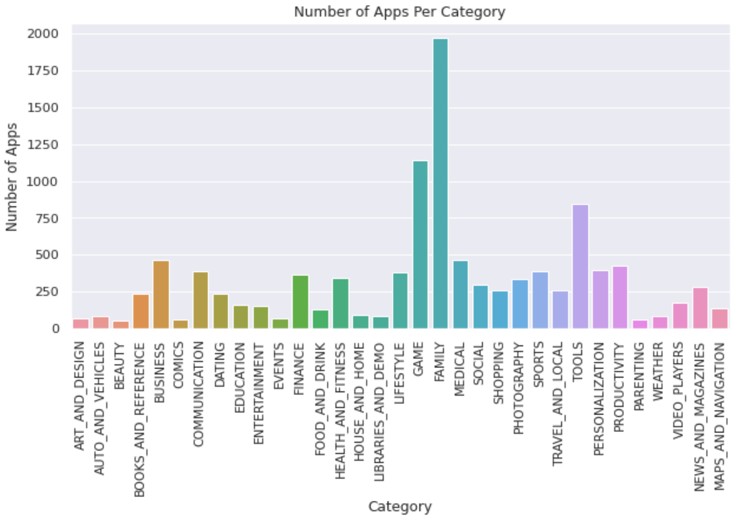
As I can see from the graph people with positive sentiments tend to give more feedback as compared to neutral and negative sentiments. Positive sentiments indicating user satisfaction with the product provided by the developer or the administrator.

- **Free vs Paid**



As I can see from the upper pie chart is that free applications consist of a major portion of Google Play store. Paid applications are significantly less as compared to free applications. I can also conclude that developers are most dependent on post-business instead of pre-business taking money up front.

- **Apps per Category**



From the above plotting I know that most of the apps in the play store are from the categories of 'Family', 'Game' and also 'Tools.

I also did analysis for Rating distribution, Number of installs for each category, Distribution of apps in terms of there rating, size and type.

6. Libraries:

1. Pandas:

It is a Python library for data analysis. Started by Wes McKinney in 2008 out of a need for a powerful and flexible quantitative analysis tool, pandas has grown into one of the most popular Python libraries. It has an extremely active community of contributors. Pandas is built on top of two core Python libraries—matplotlib for data visualization and NumPy for mathematical operations. Pandas' acts as a wrapper over these libraries, allowing you to access many of matplotlib's and NumPy's methods with less code. For instance, pandas. plot() combines multiple matplotlib methods into a single method, enabling you to plot a chart in a few lines. Before pandas, most analysts used Python for data munging and preparation, and then switched to a more domain specific language like R for the rest of them workflow. Pandas introduced two new types of objects for storing data that make analytical tasks easier and eliminate the need to switch tools: Series, which have a list-like structure, and Data Frames, which have a tabular structure.

2. Matplotlib:

It is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

3. Seaborn:

It is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas.

Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that I can switch between different visual representations for same variables for better understanding of dataset.

7. Conclusion:

The Google play store is the collection of apps which we need in day-to-day life it is flooded with the app. I provide you some useful insights regarding the trending of the apps in the play store. As per the graphs visualizations shown above, most of the trending apps (in terms of users' installs) are from the categories like GAME, COMMUNICATION, and TOOL even though the number of available apps from these categories are twice as much lesser than the category FAMILY. The trending of these apps is most probably due to their nature of being able to entertain or assist the user. So, I say that GAMING app are most popular app in google play store which attracted the audience most.

As per the above charts most of the apps having good ratings of above 4.0 are mostly confirmed to have high number of reviews and user installs. Most of the apps are having high number of reviews are from the categories of SOCIAL, COMMUNICATION and GAME like Facebook, WhatsApp Messenger, Instagram, Messenger – Text and Video Chat for Free, Clash of Clans etc.

FINANCE and LIFESTYLE category apps are the expensive apps in google play store. As per the above charts I observe that people with positive sentiments tend to give more feedback as compared to neutral and negative sentiments. We also see that most of the apps having rating 4.3

So, I learn that the current trend in the Android market is mostly from these categories which either entertaining, communicating or assisting apps.

References -

- GeeksforGeeks
- Analytics Vidhya
- Stackoverflow
- Towards data science
- Python libraries documentation