Play Store App Review Analysis

Abhilasha Rani Goel, Biswajeet Sethi
PVN Malleswara Rao, Sanjay P Malviya
Data science trainees,
AlmaBetter, Bangalore

Abstract:

A few thousands of new applications are regularly uploaded on Google play store. A huge number of designers working freely on designing the apps and making them successful. With the enormous challenge from everywhere throughout the globe, it is important for a developer to know whether he/she is continuing the correct way or not.

Since most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, in-application adverts and memberships add to the achievement of an application. In this way, an application's prosperity is normally dictated by the quantity of installation of the application and the client appraisals that it has gotten over its lifetime instead of the income is created.

The objective of this experiment is to deliver insights to understand customer demands better and thus help developers to popularize the product. We have tried to discover the relationships among various attributes such as which application is free or paid, what are the user reviews, rating of the application.

Keywords: Google Play Store, user sentiments, rating distribution

1. Introduction

In today’s scenario we can see that mobile apps playing an important role in any individual’s life. With enormous challenge from everywhere throughout the globe, it is important for a designer to realize that he/she is continuing in the right way or not. 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 dataset with 10k Play Store applications is available to analyze the market of android. It can be examined to analysis the different category such as family, communication, entertainment, tools, music, camera etc.

In this project we examine the different attributes present in the data set that affect the popularity of the application. We focused on to answer the questions like, what makes an app popular, what should be the price and size of the app, is there some trends in user sentiments.

2. Understanding the Data

In our data set we have two csv files for data analysis:

1. Play Store data
2. User Reviews

At first, we analysis the play store data and in the play store data we have 10841 rows and 13 columns & in the user review data we have 64295 rows and 5 columns of data. We have to take the maximum outcomes from the data which help us to analysis the which type of app is most preferable and comparisons between different insights. Our goal is to filter and make plots accordingly for a better EDA with respect to the final data.

Let's go through the 13 columns present in the play store data set:

- Apps: this column contains information about some different apps present in google play store app
- Category: it contains categories of different apps
- Rating: rating of apps by users
- Reviews: responses of app users
- Size: it contains the size of each app
- Installs: no. of installed each app
- Type: in the type we can able to know is our app is free or paid
- Price: in this column the price of each app is present
- Content rating: from this column we can able to know the app belongs to which age group people
- Genres: this consist of genres for each app
- Last updated: in the column we can see when app was last updated
- Current version: from this column we can find the app's current version
- Android version: from this we can find the android version used for the app

3. Analytical Problems

In this experiment, we examine the data set to give the answer of the following questions:

- Is there any app which has the rating greater than 5?
- Is there any correlation between the columns?
- Are the reviews more than installs as only those who installed can review the app?
- Is there any Null values, Duplicate files and outliers?
- What is the distribution of Reviews?
- What are the top Content Rating values? Are there any values with very few records?
- Which category have a greater number of apps and which category have least number of apps?
- Find the skew, mean and median of rating.

- What is the range of rating given to maximum apps?
- Does the installation number affect all other columns?
- Which type of apps are installed mostly, either free or paid?
- What is the maximum price of paid apps on play store and which app/s is/are expensive?
- Which app under the paid category earn most money?
- The apps under which age group are installed mostly?
- Which user sentiment has the maximum frequency?
- Which apps have the most positive sentiment and which apps have the most negative sentiments?
- The apps under which category are installed most and apps under which category have least installation number?

4. Steps Involved

After loading the dataset, we can start the exploration but before that, we need to check and see that the dataset is ready for performing several exploration operations or not, so let's first have a look at the structure and the manner in which the data is organized.

- **Data Cleaning**

  Our data set contains a large number of null values in the rating column, so we drop them. Some of the columns have a smaller number of null values, so we replace the null values in these columns with the mode value of that particular column.

  Our data set also contain the duplicate rows for a single application. We also drop the duplicate rows because the rows contain the identical data.

  Also drop the rows, which have rating greater than 5.

- **Data Transforming**

  From the information of data frame, we can see that all the columns except rating have the object data type but some of the columns like, reviews, size, installs and price have the numerical value. So, we have to transform them in proper data type and also remove the unwanted values from the numerical columns like '+' and ',' from installs and '$' from price. In the size column we have some values in KB and some values in MB, so we transform all the values in MB.

- **Exploratory Data Analysis**

  After establishing a good sense of each feature, we proceeded with plotting a pairwise plot between all the quantitative variables to look for any evident patterns or relationships between the features. There is a high variance in the number of installs and in number of reviews. To overcome this problem, we add two new columns to the data frame named:

log_installs and log_review, which contain the logarithmic values of installs and review columns, respectively.

5. Single Variate Analysis

After that we analysis all the columns one by one to examine whether the particular column contain some useful information or not:

Category

We breakdown the apps by category and observe that family and game categories have the maximum number of apps in the play store. Weather, house and home, comics, events, beauty, and parenting are the categories which have a few numbers of apps.

Rating Distribution

All the apps in play store have the rating between 0.5 to 5. Maximum apps have the rating between 3.8 to 4.5. Our analysis also observes the following information:

The skew of distribution of ratings -1.733457613883763
The mean of distribution of ratings 4.171309267241382
The median of distribution of ratings 4.3

Installs

We analysis the install column to observe the effect of size, price, rating, content rating, android version on app installation number. We can analysis that for each and every category number of app installation does not depend on the size. The free apps installed mostly. The apps which can be used by everyone is more installed than the apps which can be used by a particular age group. Rating of mostly installed apps is between 4 and 5.

Price

We feel that the apps which is free are more popular. That is confirmed by the data as free apps dominate the store at 92.6%, making it over 12.5 times greater number from paid apps. It also observes that the apps which have less price are more popular. The maximum paid apps have the price less than 50$. Some of the apps also have the price greater than 250$. The names of these apps were typically something to the effect of “I am rich”. We also examine the top 15 most earning apps, and surprisingly “I am rich” is the top most earning app, which belongs to lifestyle category and also installed by 1,00,000 users.

We also observe some interesting facts that the apps which has less size are installed mostly. The size of the app should between 2MB – 50MB. The apps which is suitable for every age group is mostly downloadable. The last year is 2018, in which the most apps are

updated. The apps under gaming category have the maximum installation number. The apps under parenting, comics, beauty and events category are installed by a very few users.

6. User Review Analysis – Users Sentiments

The user review data set also have a large number of null values, so we decide to drop all the null values. Then, we check the frequency of the user sentiments. The positive sentiment is present for the maximum apps. We also examine the top 15 apps having positive review and top 15 apps having negative review.

7. Conclusion

For the apps to be popular and mostly downloadable, a developer should focus on:

- The app should be free. For no ads application, the price of the app should be less than 10$.
- Paid app should be designed in small size and to meet the user expectation.
- The size of the app should be as small as possible, preferably between 2MB – 50MB.
- There is a positive correlation between installs and review.
- The Game category have a good potential for developing an app, because this is the demanding category.
- Apps which are available for everyone are most installed apps.
- Users prefer the apps which are compatible with android version 4.1 and above.