Play Store App Review Analysis

Sumit Kumar Dhibar
Data science trainee,
AlmaBetter, Bangalore

Abstract:

Play Store serves as the official app store for the Android operating system, allowing users to browse and download applications developed with the Android SDK and published through Google.

This project aims to use the knowledge gained in Data Wrangling class on Goggle Play Store database.

1. Problem Statement

The Play Store apps data has enormous potential to drive making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market.
Each app (row) has values for category, rating, size and more. Another dataset contains customer reviews of the android apps.
Explore and analyse the data to discover key factors responsible for app engagement and success.

Data Description

The dataset contains details of Android applications present on Google Play. There are 13 includes that depict each application and an aggregate of 10841 applications. Following variables were initially included:

- App – Name of the App
- Category – Category of the app
- Rating – Overall user rating of the app out of 5 on the Play Store
- Reviews – Number of user reviews for the app
- Size - Size of app
- Installs - Number of user downloads/installs for the app
- Type - Paid or Free
- Price - Cost of the App
- Genres - An app can belong to multiple genres (apart from its main category)
- Last updated - Date when the app was last updated on Play Store
- Current Ver - Current version of the app available on Play Store
- Android Ver - Minimum required Android Version

2. Introduction

Nowadays each individual comes up with lots of interesting ideas related to creating a new app. Using different tools and applications helping in the creation of these mobile apps. It is even easy to implement those ideas. But the situation is they are not sure if their app would be successful or what kind of app they should develop to make it successful and reach more no people. Seems this data are of apps that are currently available in Play Store. Let's see if it can help us in detecting the target factors responsible for app engagement and success. It would prove useful for our developers if we can make some conclusions from it.

Android is expanding as an operating system. It has captured around 74% of the total market which is a true indicator of the huge amount of population using android. Our goal is to help android developers to know what is the motivating factor for people to download an app. It will also help to find out the factors that affect someone's decision to download an app. I would like to analyse category, reviews, price, ratings and installs for this purpose and find out how they are inter related.

3. Objective of our Project-

* Installation of application by users according to the categories.
* Mostly demanded applications in playstore.
* Factors that affect the installation of application by the user.

4. Steps involved:

- loading the data into data frame
- cleaning the data
- extracting statistics from the dataset
- exploratory analysis and visualizations like Checking the installation according to rating, types of application, top 20 installs according to genre by using visualization.
- conclusion

We can move to first step of data analysis by cleaning the data that will make the results more accurate.

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.

Cleanse and validate data

This step is for removing faulty data. Some tasks are here:

- Removing extraneous data
- Conforming data to a standardized pattern.

Exploratory Analysis and Visualization

Exploratory data visualizations (EDVs) are the type of visualizations we assemble when we do not have a clue about what information lies within our dataset.

Distribution of App Rating

Average rating of application in store is around 4.3, which is very high. This plot can be used to look whether the original ratings of the app matches the predicted rating to know whether the app is performing better or worse compared to other apps on the Play Store.

Number of Installed applications for each category

From the EDA we can conclude that, maximum number of apps present in google play store comes under Family, Games and Tools Category but as per the installations and requirements in the market place, this is not the case. Maximum installed apps comes under Games, Communication and Tools.

Distribution of App Size

It can be seen that maximum application's size lies between 0–10 MB.

Distribution of Subjectivity

It can be seen that maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this we can conclude that maximum number of users give reviews to the applications according to their experience.

Count of applications in each category differentiated by their type

It looks like certain app categories have more free apps available for download than others. In our dataset, the majority of apps in Family, Games and Tools, as well as Social categories were free to install. At the same time Family, Personalization and Medical categories had the biggest number of paid apps available for download.

Apps installed according to its type

It can be concluded that the number of free applications installed by the user are very high when compared with the paid ones. As we have converted number of installs to it's log, that is why the difference in the plot between free and paid apps seems to be low.

Size impact the number of installs of any application

It is clear that size may impact the number of installations. Bulky applications are less installed by the user.

Sentiment Polarity and Sentiment Subjectivity

In the another dataframe, we have three new columns i.e.. Sentiment, Sentiment Polarity and Sentiment Subjectivity. Sentiment basically determines the attitude or the emotion of the reviewer, i.e., whether it is positive or negative or neutral. Sentiment Polarity is a float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement. Sentiment Subjectivity generally refer to personal opinion, emotion or judgment, which lies in the range of [0,1].

Sentiment is divided for different type of reviews

It can be seen that the number of positive reviews are way higher than negative and neutral ones.

Conclusion

I started from scratch where the dataset we took was totally raw. I did a lot of cleaning on the data provided to bring it in a cleaner, representable form. Missing values were also removed in this process.

After completion of this project I got some conclusions -

- Users prefer to install free applications more.
- Communication category apps are in high demand for all type of users.
- A part of our population (18-30 yr age people) use Gaming category applications a lot and give reviews according to their

Sentiments. So developers have to keep proper attention while developing those apps or making change.

- Users installs the application depends according to previous rating and reviews

References

1. Analytics Vidhya
2. Kaggle
3. Medium