**Capstone Project**

On

Play Store APP Review Analysis



Submitted by

Himanshu Rathod

Email id- [himanshurathod755@gmail.](mailto:sabitendumahapatra@gmail.com)com

GitHub link- <https://github.com/Mr-Himanshu123/EDA_Project.git>

Project Summary

Data science can be summarized into five steps: capture, maintain, process, analyze, and communicate. Mobile app distribution platforms such as Google Play get flooded with several thousand new apps every day, with many more thousands of developers working independently or in a team to make them successful. With immense competition from all over the globe, it is imperative for a developer to know whether he is proceeding in the right direction. The app’s success is usually determined more by the number of installs and the user ratings that it has received over its lifetime than by the revenue it generates. To gain a deeper understanding of the Google Play Store data, we conducted exploratory data analysis in order to find relationships with specific features, such as how the number of words in an app's name affects installs, so that we can use these to determine which apps will have the highest chances of succeeding.

In the initial phase, we focused on the data cleaning and preparation in order to ensure that we gave them the best possible outcome of our analysis. Our major challenge was data cleaning. In data cleaning, we have performed a few steps to ensure the quality of the data, such as removing null values. During the data cleaning step, we found that 13.60% of ratings were null values, and even after merging both the data frames, we could not infer much in order to fill them. Thus, we had to drop them.

User Reviews had 42% of null values, which could have been used for developing an understanding of the category-wise sentiments, which would help us fill 15.60% of the null values in the Rating column.

The merged data frame of both the Play Store and user reviews had only 816 common apps. This is just 10% of the cleaned data; we could have given more valuable analysis if we had at least 70%–80% of the data available in the merged data frames.

With the cleaned data, we have performed exploratory data analysis to understand our dataset, like the number of installations for each category. We observed the distribution of all numerical data and the frequencies of categorical data. We have also found out the correlation of several data points, like category vs. install, rating vs. review, sentiment subjectivity vs. sentiment polarity, rating, and so on.

After completion of the analysis, we concluded that users prefer more free apps. Most of the apps present in the Play Store are more or less the same size, so size doesn’t affect their decisions much. The majority of the apps in the dataset are small in size. It was found that most of the apps that are present on the Google Play Store have ratings between 4 and 5. The dominant content rating of the app is everyone, and communication is the most frequent genre.

We found the most popular apps based on two factors: the number of installs and the number of reviews. Facebook is the most popular app with this logic, and WhatsApp has followed.

Most of the reviews have positive sentiment, while negative and neutral reviews have a low number of reviews. Most of the reviews fall on the [-0.50, 0.75] polarity scale, implying that the extremely negative or positive sentiments are significantly low.

Most of the reviews show a mid-range of negative and positive sentiments. Sentiment subjectivity is not always proportional to sentiment polarity, but in the maximum number of cases, it shows linear behaviour when variance is too high or too low.

The scope of the problem is not limited to what is being considered for this project. Considering the results and processes we have executed, we can conclude that we have met the objective of the group project, which was to analyse the apps on the Google Play Store and determine trends for the Google Play Store.