

Capstone Project-1 Submission

Name of Team Member, Email, and Contribution:

1. Name: - Ashwini R
2. Email ID:- Ashwini.rajendra58@gmail.com
 - Assisted in the notebook by contributing to Google Drive data connectivity, performing data cleaning and manipulation, and participating in Exploratory Data Analysis (EDA) visualization.
 - Contributed to the content of the PowerPoint presentation.
 - Played a role in the Technical Documentation by contributing to the content related to the problem statement, project goals, and the steps involved.
3. Name: - Dhanaraj S
4. Email ID: - dhanaraj.siddappa81@gmail.com
 - Played a part in the notebook by engaging in data cleaning, data manipulation, Exploratory Data Analysis (EDA) visualization, and finalizing the conclusion.
 - Played a role in the PowerPoint presentation by ensuring the inclusion of all the points to be covered.
 - Played a role in the Technical Documentation by contributing to the content related to the problem statement, project goals, and the steps involved.

Please paste the GitHub Repo link.

Sandeep Github Link: - <https://github.com/San13deep/Play-Store-App-Review-Analysis>

Ansh Bhatnagar Github Link: - <https://github.com/AnshRockstar/Playstore-Apps-Review-Analysis>

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

The field of data science can be distilled into five key steps: capturing, maintaining, processing, analyzing, and communicating data. Analyzing the Google Play Store applications has proven instrumental in developing more reliable and interactive apps. This insight is particularly valuable for app developers aiming to create applications tailored to specific categories identified in this analysis. The findings contribute to the construction of applications with well-defined and accurate objectives.

During the initial phase, our primary emphasis was on addressing the problem statements and conducting thorough data cleaning to deliver optimal results from our analysis. The most significant challenge we encountered was associated with data cleaning, wherein several steps were taken to enhance data quality, including the removal of NaN values. Through this process, we discovered that 13.60% of the reviews contained NaN values. Despite merging both data frames, we faced difficulties in deriving meaningful insights to fill these gaps, ultimately leading us to the decision to drop the NaN values.

Out of the cleaned data, the merged data frame containing both the Play Store and user reviews only included 816 common apps, constituting merely 10% of the dataset. A more comprehensive analysis could have been conducted if a larger portion, ideally between 70% and 80%, of the data had been available in the merged data frames.

The User Reviews column contained 42% NaN values, presenting an opportunity to gain insights into category-wise sentiments. This information could have been leveraged to fill the 13.60% NaN values in the Reviews column.

Utilizing the cleaned data, we conducted Exploratory Data Analysis to gain a deeper understanding of our dataset. This involved examining factors such as the number of installations for each category and exploring correlations between the app size, Android version, and the number of installs.

The overarching goal of our entire project was to scrutinize the data and identify the key factors influencing users' decisions to download an app. Upon completing the analysis, I arrived at the conclusion that users tend to prefer free apps. Additionally, since most apps in the Play Store are similar in size, it was observed that the size of the app doesn't significantly impact their decision-making process.

The analysis revealed that a majority of the apps available on the Google Play Store have ratings ranging from 4 to 5. Additionally, it was noted that the dataset contains a predominant number of applications characterized by small sizes.

Upon examination, we determined the most popular category of apps based on two criteria: the number of installs and the number of reviews.

Personalization emerged as the leader in the former category, while Sports claimed the top position in the latter.

The problem statement provides us with two datasets: the Play Store dataset and the User Review dataset. In the User Review dataset, it was noted that 42% of User Reviews were marked as NaN, presenting an opportunity to analyze category-wise sentiments. This information could have been employed to address the 13.60% NaN values in the Reviews column.

The majority of the reviews express a positive sentiment, with a comparatively lower number of reviews categorized as negative or neutral. Additionally, the sentiment polarity and subjectivity are considered in the analysis.

The compilation of reviews exhibits a broad spectrum of subjectivity, with the majority falling within the $[-0.50, 0.75]$ polarity scale. This suggests that sentiments tending toward extreme negativity or positivity are relatively infrequent. Instead, the reviews predominantly reflect a mid-range of both negative and positive sentiments.

Sentiment subjectivity doesn't always exhibit a direct proportionality to sentiment polarity. However, in the majority of cases, it demonstrates a proportional relationship, especially when the variance is either very high or very low.

Sentiment polarity does not display a strong correlation with sentiment subjectivity.

The dataset holds vast potential for enhancing business value and generating positive impacts. Its utility extends beyond the specific problem addressed in this project, opening up numerous other intriguing possibilities for exploration.

Based on the results and the implemented processes, we can affirm that we have successfully met the objective of this group project, which involves analyzing Google Play Store apps, identifying trends within the Google Play Store, and addressing both of our research questions.