**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| **Name : Gaurav Makode .**  **Email** : [gauravmmakode002@gmail.com](mailto:gauravmmakode002@gmail.com) |
| **GitHub Repo link.** |
| **Github Link** **:-**  [**https://github.com/gauravmakode/Play-Store-App- Review-Analysis-**](https://github.com/gauravmakode/Play-Store-App-%20%20Review-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)**  Google Play, also branded as the Google Play Store and formerly Android Market, is a digital distribution service operated and developed by Google. It serves as the official app store for certified devices running on the Android operating system and its derivatives as well as ChromeOS, allowing users to browse and download applications developed with the Android software development kit (SDK) and published through Google. Google Play has also served as a digital media store, offering games, music, books, movies, and television programs be. Content that has been purchased on Google Play Movies & TV and Google Play Books can be accessed on a web browser, and through the Android and iOS apps.  In the initial phase, we focused more on the problem statements and data cleaning, in order to ensure that we give them the best results out of our analysis. Our major challenge was data cleaning, In Data Cleaning, we have performed few steps to ensure the data quality such as removing NAN values. During the Data Cleaning step we found that 13.60% of reviews were NaN 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.  The merged data frame of both 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.  User Reviews had 42% of NaN values, which could have been used for developing an understanding of the category wise sentiments, which would help us to fill 13.60% NaN values of the Reviews column.  With the cleaned data, we have performed Exploratory Data Analysis to understand our dataset like number of installations for each category We explore the correlation between the size of the app and the version of Android on the number of installs and so on.  Our motive in whole project was to analyze the data and find out main components that affect users decision to download app. After completion of analysis I concluded that user prefer more of free apps. Most of the apps present in play store are more or less of same size so size doesn’t affect their decision much.  It was found that Most of the apps that are present on the google play store have rating in between 4 and 5.Also it was observed that Maximum number of applications present in the dataset are of small size.  We found most popular category of apps on two basis - Number of Installs and Number of reviews. Personalization wins in former criteria whereas Sports wins in later criteria.  In the problem statement we are given with Two datasets i.e. play store and User review data set in the user review dataset it was observed that User Reviews had 42% of NaN values, which could have been used for developing an understanding of the category wise sentiments, which would help us to fill 13.60% NaN values of the Reviews column.  Most of the reviews are of Positive Sentiment, while Negative and Neutral have low number of reviews.  Collection of reviews shows a wide range of subjectivity and most of the reviews fall in [-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 maximum number of case, shows a proportional behavior, when variance is too high or low.  Sentiment Polarity is not highly correlated with Sentiment Subjectivity.  The dataset contains immense possibilities to improve business values and have a positive impact. It is not limited to the problem taken into consideration for this project. Many other interesting possibilities can be explored using this dataset. |
|  |