**Capstone Project-1 Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| 1. Name: - Balkrishna Tiwari   Email:- [balkrishnatiwari942@gmail.com](mailto:balkrishnatiwari942@gmail.com)  Contribution:   1. Understood the business requirement and prepared path to follow for completion of the project. 2. Analyzed raw data, and standardized content for further analysis using tools of Pandas and python iterations. 3. Generated visualization based on standardized data using Matplotlib and Seaborn libraries. 4. Prepared technical documentation of this project ensuring massage is clearly delivered and work is properly documented. 5. Prepared power presentation on project, ensuring all the analytical details are covered. 6. Presented whole project in a video massage. |
| **Please paste the GitHub Repo link.** |
| Balkrishna Github Link: - https://github.com/BalKrishna-Tiwari/Playstore-App-Analysis.git |
| **Summery of Capstone Project**  **Play Store App and Review analysis** |
| Exploratory data analysis (EDA) is an approach developed by Python programmers to analyze the raw data being generated at various sources, and explore hidden trends using visual techniques. Data Science or EDA within it can be summarized into five steps that is, capture data, clean, analyze, summarize and communicate or visualize. With the available details we were expected to run EDA and Data analysis techniques on google play store data sets and review data sets, in order to come up with occlusive observations regarding app business with historical trends. These observations are expected to remain inline with formulating insight such that details are helpful to new app makers in deciding parameters for new app launch, in order to achieve increased probability of its success within audience.  We have received two data sets namely App data set that contains all the aggregated details regarding apps that are offered on play store, and the second data set is Review data, that contains records of each written reviews made by users on particular app.  At first, I went on to check condition of data set and found lots of deformity in it. The play store data structure contains 10841 rows and 13 columns. The data contains list of app and values of variables associated to it like current rating, number of reviews, category of app, android version required, current version of app etc. Most of the values mentioned in columns were non-numerical type all though its nature appears numerical. I have applied Uom standardization using customary defined function to such cases. I have also introduced some logical mutations so that string values can be converted into either numerical data (like converted smallest android version to numbers) or into purely unique value (converted multiple genres associated into primary single genres). There were around 13% of rows with duplicate app name in this data set. The Review data set contains 64295 rows and 5columns, most of the data values in this data frame appeared sorted but it has 26868 rows with null comment.  Initially I have tried to use merged data of play store data and review data, but later realized that clubbing two will concentrate our visualization towards app that has received most number of reviews and hence dropped this idea and went on to analyses app data separately and merged data separately for sentiment analysis only.  With two cleaned data sets in hand I have went on to perform various kinds of graphical visualization to reveal any trend and correlation between variables. I have also performed statistical analysis on rating distribution, price distribution to check for there nature. With all these efforts I have generated a number of conclusive results that will certainly helpful in the development and promotion of new app.  From EDA of both data sets it was revealed that most apps have received rating of more then 3.5. Rating reviews of apps are highly correlated to installs made to app. 92.3% of the total app present on play store are free. 81% of the app can be assessed by users belonging to any age group. Most of the app on play store have size 5MB to 30MB, or we can say most apps are of lower sizes. Android version 4.0 is the suitable to run most of the apps available. Most of the paid apps charges installation cost between 0 to 1$. Minecraft is the app that has generated highest revenue from installations only.  By going through EDA of play store and its reviews data I can say that people prefer to use free app more. In order to maintain apps relevant to users it is necessary that developer should add updates at regular interval and remain open to incorporate changes based on customer reviews. Game appears to be the most installed category at the same time we have noticed highest percent of negative comment in this category. It signifies existing market gap and possibility of new app in this segment. Dating is one of the underrated categories that again has a room for new players. Game and Family are the top app categories which have received written reviews and hence holds good ground of interests from user point.  More detailed report could have been generated if more number of review data would have available with us, and instead of install values were segmented, it would have been great if actual install numbers were available in data set.  From the process that we have followed and result there upon which we have received, we can conclude that the objective of producing valuable insight from play store data has reached. And the critical observations made to resolve our problem statement will be helpful to developers in order to release new app on play store. |