**Capstone Project-2 Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| 1. Name: - Kartik Subhashrao Dhande. 2. Email ID:- [dhandekartik07@gmail.com](mailto:dhandekartik07@gmail.com)  * Contributed In notebook helped with Google diver data connectivity and data cleaning, data manipulation, and in EDA Visualization * Contributed for the contents of PPT. * Contributed in Technical Documentation in content of problem statement goal of project and steps involved.  1. Name: - Pranav Vilasrao Balpande. 2. Email ID: - [pranav.balpande@gmail.com](mailto:pranav.balpande@gmail.com)  * Contributed in notebook for data cleaning, data manipulation, and in EDA Visualization and finalizing the conclusion. * Contributed in PPT by making sure all the points to be covered. * Contributed in Technical Documentation in content of problem statement goal of project and steps involved      1. Name: - Sanket Rajendra Bhosale. 2. Email ID: - [sanketbhosale0023@gmail.com](mailto:sanketbhosale0023@gmail.com)  * Contributed in notebook for data cleaning, data manipulation, and in EDA Visualization and finalizing the conclusion. * Contributed in Technical Documentation in content of problem statement goal of project and steps involved  1. Name: - Kartik Anilrao Pisudde. 2. Email ID: - [pisuddekartik@gmail.com](mailto:pisuddekartik@gmail.com)  * Contributed In notebook helped with Google diver data connectivity and data cleaning, data manipulation, and in EDA Visualization * Contributed for the contents of PPT.  1. Name: - Puja Ashok Nehare. 2. Email ID: - [puja.nehare@gmail.com](mailto:puja.nehare@gmail.com)  * Contributed in notebook for data cleaning, data manipulation, and in EDA visualization and finalizing the conclusion. * Contributed in notebook helped with Google diver data connectivity and data cleaning, data manipulation, and in EDA visualization. |
| **Please paste the GitHub Repo link.** |
| Kartik Subhashrao Dhande GitHub Link: -  <https://github.com/KartikDhande007/play_store_app_review_analysis>  Pranav Vilasrao Balpande GitHub Link: -  https://github.com/pranav4536/Retail-Sales-Prediction  Sanket Rajendra Bhosale GitHub Link: -  <https://github.com/kartik-pisudde/capstone-project-playstore-review-analyze>  Kartik Anilrao Pisudde GitHub Link: -  <https://github.com/kartik-pisudde/capstone-project-playstore-review-analyze>  Puja Ashok Nehare GitHub Link: -  <https://github.com/pujanehare/capstone-project-2.git> |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| Sales forecasting refers to the process of estimating demand for or sales of a particular product over a specific period of time.  Businesses use sales forecasts to determine what revenue they will be generating in a particular timespan to empower themselves with powerful and strategic business plans. Important decisions such as budgets, hiring, incentives, goals, acquisitions and various other growth plans are affected by the revenue the company is going to make in the coming months and for these plans to be as effective as they are planned to be it is important for these forecasts to also be as good.  First step involved is understanding the data and getting answers to some basic questions like; What is the data about? How many rows or observations are there in it? How many features are there in it? What are the data types? Are there any missing values? And anything that could be relevant and useful to our investigation. Let’s just understand the dataset first and the terms involved before proceeding further.  Our dataset consists of two csv files, the first consists of historical data with 1017209 rows or observations and 9 columns with no null values. The second dataset was supplementary information about the stores with 1115 rows and 10 columns and a lot of missing values in a few columns.  The data types were of integer, float and object in nature.  missing values. Otherwise, it is better to replace them with appropriate values.  It is necessary to check and handle these values before feeding it to the models, so as to obtain good insights on what the data is trying to say and make great characterisation and predictions which will in turn help improve the business's growth.  The historical records dataset had no null values.  Handling missing values is an important skill in the data analysis process. If there are very few missing values compared to the size of the dataset, we may choose to drop rows that have   * The major challenge would be the computational time and RAM needed to work upon such a dataset in a cloud environment.   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 2 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. 8.Sentiment Polarity / Sentiment Subjectivity  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.  From the results and process we have implemented; we can conclude that we have achieved this group project objective which is analyzing the Google Play Store apps and determine trends of the Google Play Store and both of our research questions. |