Capstone Project Submission

Instructions:

i) Please fill in all the required information.
ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
1. Name : Divya Prakash Kedia **Email ID :** dpkedia2201@gmail.com **Contribution:** • Contributed in Google Colab Notebook with Data Cleaning, Data Manipulation, Data Wrangling and Data Visualization. • Preparation of the contents of the PowerPoint Presentation. • Preparation of the Technical Document by taking all the necessary contents into consideration. **2. Name :** Sakshi R. Ghugare **Email ID** : ghugaresakshi@gmail.com **Contribution:** • Contributed in Google Colab Notebook for Data Cleaning, Data Manipulation, and in EDA Visualization and finalizing the conclusion. • Prepared the power point presentation and to make sure all the content in the PPT are covered. • Contributed for preparation in Technical Documentation by considering all the necessary steps
Please paste the GitHub Repo link.
1. Divya Kedia Github Link: https://github.com/divyakedia/playstore-data-analysis **2. Sakshi Ghugare Github Link :** https://github.com/sakshighugare/EDA-Playstore-data-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 Google Play Store is home to not only apps and games, but some additional digital content as well. You can find movies, books, and music in the Play Store. Google basically launched the Play Store as a hub for its users to get all sorts of digital content. The main content would probably be apps and games, as the Play Store was mainly launched for Android smartphones.

The analysis of Google Play Store application aided to build most reliable and more interactive applications. This would be very useful for app developers to build an application focused on certain discussed category in this analysis. This analysis will help in building the application with precise and accurate objectives.

Discussion of Google play store dataset included various steps such as:

- Loading the data into data frame
- Cleaning the data
- Extracting statistics from the dataset
- Exploratory analysis and visualizations
- Questions that can be asked from the dataset
- Conclusion

In the Starting phase of the project , we have focused more on the understanding of the problem statements and data cleaning, in order to get best and reliable results out of our analysis. Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a recordset, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data

By Studying the DataFrame, we see noticed that,
The Shape of the DataFrame was (10841, 13), Column 'Reviews', 'Size', 'Installs' and 'Price' were the type of 'object', Column 'Size' had the strings representing size in 'M' as Megabytes, 'k' as kilobytes and also 'Varies with devices', Column 'Installs' had strings representing no. of installs with symbols such as ',' and '+', Column 'Price' had strings representing price with symbol '$'.
Thus, it was concluded that the data had to be cleaned for further process.

We also saw that in "Rating" there were 1473 null values. Hence we decided to replace them with the median of "Rating".

After the data cleaning the next step was data preparation. Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data, and the combining of data sets to enrich data. With the cleaned data, we performed Exploratory Data Analysis to understand our dataset like number of installations for each category We explored the correlation between the size of the app and the number of installs and so on.

After Data preparation next process was Exploratory Analysis and Visualization.
In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.
Data visualization is the graphic representation of data.
It involves producing images that communicate relationships among the represented data to viewers of the images.

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 than the paid one. Most of the apps present in play store are more or less of same size so size doesn't affect their decision much but most installed apps don't have size more than 100 Mb.

It was also found that Most of the apps that are present on the Google play store have rating between 4 and 4.5.
Content Rating of apps on Playstore we can see that most of the content on playstore are for "Everyone" followed by "Teen" , "Mature 17+" , "Everyone 10+", "Adults only 18+" and lastly "unrated" with the percentage of 80.39%, 11.14%, 4.60%, 3.18% , 0.03%, 0.02% respectively.

Sentiment Subjectivity lies mostly between 0.5 and 0.65. It shows that the average content and apps reviews subjectivity are mostly relevant. Subjectivity of 100% has slightly occurred frequently.
The nearly 0 subjectivity has a considerable amount of frequency.
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.

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.