CAPSTONE PROJECT SUMMARY

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email, Contribution and Github links :-

Name	Email	Contribution
Md Sazil Sharif	mdsazilsharif@gmail.com	Data Analysis, Power point presentation, Summary
Asadullakhan Pathan	asadullaapathan@gmail.com	Powerpoint presentation, Technical Documentation, Data filtering
Adil Imam	adil.imam12@gmail.com	EDA , Technical Documentation, Final touch to the project
Sushil kumar singh	sushilsinghrajput.333@gmail.com	Technical Documentation, Data Cleaning, Data filtering
Madhulika Kumari	mpradhan1990@yahoo.com	EDA, Data analysis , Data Cleaning

Github Repo Links :- **https://github.com/sushilksing808/EDA_1st-capstone-project**

Please write a short summary of your Capstone project and its components.

Describe the problem statement, your approaches and your conclusions. (200-400 words)

- Data science can be summarized into five steps: **capture, maintain process, analyse, and communicate**. Play store is an application for android users which allows the users to download millions of applications for entertainment purposes like gaming, watching movies, fitness, reading books, doing businesses etc.
- In this capstone project we have compared thousands of applications across various categories. We have analysed the data to discover key factors responsible for app engagement and success helping the developers to work and capture the android market.
- We have been provided with 2 Dataset files – 'Play store csv' and 'User Reviews". One containing 13 databases namely **'App', 'Category', 'Ratings', 'Reviews', 'Types', 'Size' , 'Installs' , 'Genres' , 'Price' , 'Content Rating' , 'Last Updates' , 'Current Version' and 'Android Version' and another file containing databases namely 'App' , 'Translated Review' , 'Sentiment' , 'Sentiment Polarity' and 'Sentiment Subjectivity'**.
- First we have performed Data Wrangling over the raw data. We then analyzed the data, database by database. We then checked for any duplicate data present to be removed. Then we checked for any errors or null values present. Then we filtered it one by one.
- 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.
- 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.

- 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.
- 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.

- 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.
- Most of the reviews are of Positive Sentiment, while Negative and Neutral have low number of reviews, 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.
- 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.
- 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. Percentage of free apps is 92%, Percentage of apps with no age restrictions is 82%, and Percentage of apps that are top rated is 80%.
- There are 20 free apps that have been installed over a billion times. Minecraft is the only app in the paid category with over 10M installs. This app has also produced the most revenue only from the installation fee.
- The apps whose size is greater than 90 MB has the highest number of average user reviews, the median size of all apps in the play store is 12 MB. The apps whose size varies with device has the highest number average app installs. Helix Jump has the highest number of positive reviews and Angry Birds Classic has the highest number of negative reviews.
- We also plotted graph for sentiments and noted that 64% are positive while 22% are negative and rest 13% are neutral. We also plotted ‘Category vs density’, ‘Category vs rating’, ‘Category vs review’, ‘Category vs install’, ‘Category vs paid/Free’, ‘App vs rating’, ‘Sentiments vs review’ etc. graphs.
- Our motive in whole project was to analyse the data and find out main components that affect user’s decision to download app. After completion of analysis I concluded that user prefer more of free apps.

 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.