Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Name	Email	Contribution
Sarthak Arora	Sarthak1611@gmail.com	Data Filtering, EDA, Technical Documentation, Summary
Jay Nandasana	nandasanajay@gmail.com	Data Filtering, Data Cleaning, EDA , Power point presentation
Madhavi Mali	madhavimali1996@gmail.com	EDA, Data cleaning, Data Analysis, Summary
Arshi Wani	arshiwani3@gmail.com	Data analysis, Data filtering, Technical Documentation
Pranjali Tete	pranjalitete@gmail.com	Power point presentation, Final touch to the project

Github Link:- https://github.com/madhavimali/PlayStore_DataDigger

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Play store is an application for android users which allows the users to download millions of applications for entertainment purposes like gaming, watching movies, downloading fitness applications, reading books, doing businesses etc.

In this capstone project we have compared thousands of applications across various categories. We have analyzed 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 began with ‘App’ database and removed all the duplicate rows present in it. Then we moved to ‘Category’ and we noticed that there is one outliner present. We observed that one row has been shifted, so we shifted it back to its original location and corrected the row. In ‘Rating’ there were some null values so we replaced it with the median of all the values present in that column. In ‘Installs’ we removed the ‘,’ and ‘+’ symbols and converted it to Int type. In ‘Type’ we observed one NaN value and converted it to ‘free’ to simplify the data. In ‘Size’ we converted all the entries to one single unit (from M to k). In price we removed the ‘$’ and converted it to float. Lastly, we converted ‘Last Updated’ to datetime datatype.

After this we performed EDA. We plotted pie chart for ‘Apps’ against ‘Android Version’. We observed that most of the apps required android version 4.0 and above.
Next we plotted bar graph for top categories, and found out that ‘Family’ , ‘Games’ and ‘Tools’ are the top three ones. For the ‘Genre’, the most popular genre is ‘Tools’ followed by ‘Entertainment’.
Next we plotted graph for most common ‘Rating’ that the app gets. We found out that the most common rating is around 4.3.
We also plotted graph between share of ‘paid’ vs ‘free’ app. We noted that there are approx. 93% free apps, while only 3% are paid.
Next, we plotted no. of apps that got updated in the following years. We then plotted pie chart for content rating and noted that around 81% are for everyone and 10.7% are for teens.
We also plotted graph for sentiments and noted that 64% are positive while 21% are negative and rest 14% are neutral. Heatmap for sentiment polarity and subjectivity is also shown.
We also plotted ‘Category’ vs ‘Size’ stripplot, ‘Content rating’ vs ‘Sentiment’ countplot , ‘Sentiment’ vs ‘Category’ and also ‘Category’ vs ‘Type’ graphs.

These observations can clearly help the developer to capture the android market.