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
Raman Kumar	ramank445522@gmail.com	**Data Visualization & Conclusion**
Rajat Chaudhary	rajat.25.chaudhary@gmail.com	**Data Visualization**
Anukriti Shakyawar	shakyawaranukriti@gmail.com	**Data Filtering**
Deepmala Srivastava	srvdeepmala@gmail.com	**Data Cleaning: Identifying and removing duplicate entries. Removing visual impurities like "+", "$" sign.**

Github Link:- https://github.com/ramank123/Play-Store-App-Review-Analysis-EDA-

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.csv". One containing13 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'.

We started by performing some fundamental research on our dataset. When we did this, we found the basic information regarding our data set such as columns, data types and we also found out that we have missing values, data duplication and a few other issues as well. 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 and we remove the outliner. In 'Rating' therewere 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 observedone 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.
We also plotted 'Category' vs 'Size' strip plot, '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.