Capstone Project

Play store app review analysis

Mohammed Saudh

Exploratory data analysis on the Google Play store Data Set

1. Defining the purpose of the exercise
2. Exploring the Dataset
3. Preapring and cleaning the data
4. Drawing useful insights using visualisations

Play Store Data analysis-Why??

Play store has a vast amount of data that can be analysed to understand trends among users and draw valuable insights which in turn will help app making companies and developers make robust decisions regarding new projects.

Data Summary

Columns present in the Data Set :

1.App : This column Contains the name of the app for each observation.

2.Category : This column Contains Category to which the app belongs.

3.Rating : This column contains the average rating for the app.

4.Reviews : This column contains the number of reviews that the app has received on the play store.

5.Size : This column contains the amount of memory the app occupies on the device

6.Installs : This column contains the number of times that the app has been downloaded and installed from the play store.

7.Type : This column contains the information whether the app is free or paid.

8.Price: If the app is a paid app, this column contains the data about its price.

9. Content Rating: This column contains the maturity rating of the app i.e. the age group of the audience for which it is suitable.

Data Summary

10.Genres : This column contains the data about to which genre the app belongs. Genres can be considered as a further division of the group of Category.

11.Last Updated : Contains the date on which the latest update of the app was released.

12.Current Version : Contains information on the current version of the app available on the play store.

13.Android Version : Contains information about the android versions on which the app is supported.

Data Pipeline:

Data Processing 1: In this step the null values in the ratings column were replaced with the column mode1.

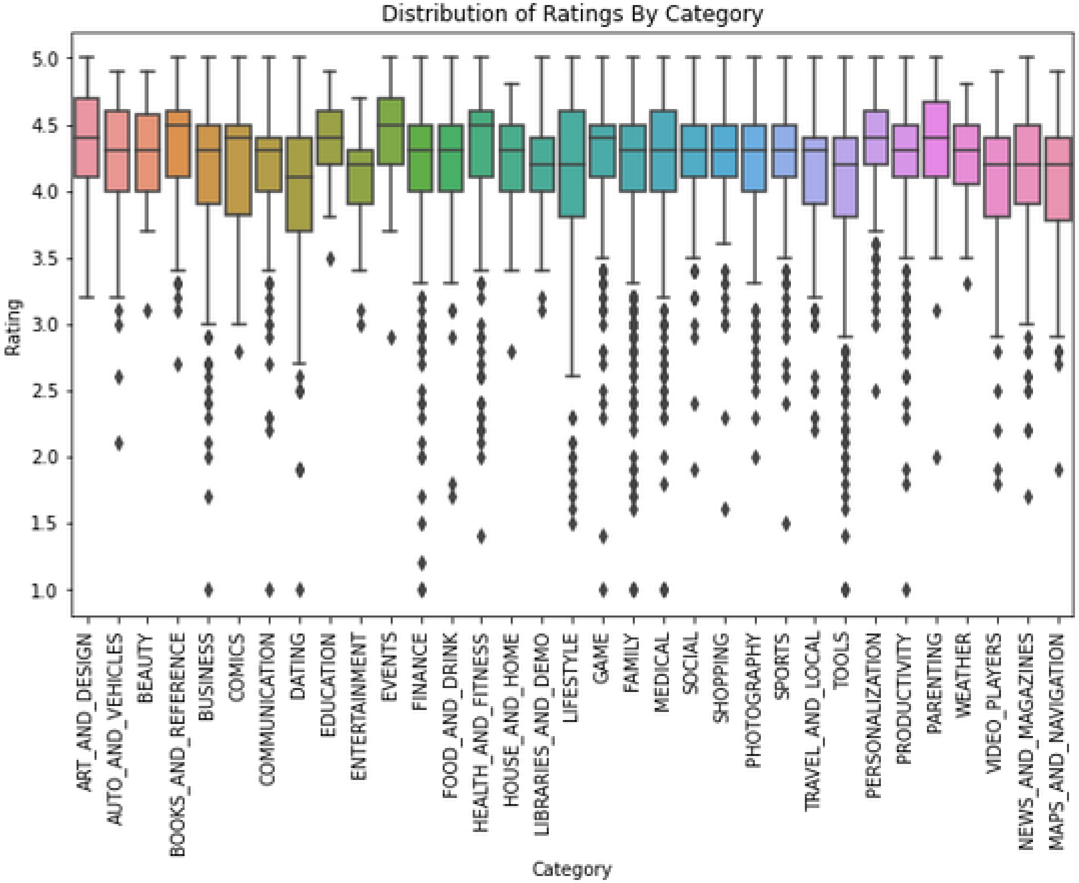
Data Processing 2: In this step, some of the columns which were present as object data type were converted to numeric data type.

Data Processing 3 : In this step, certain unnecessary columns were dropped.

EDA : In this step we do some exploratory data analysis in order to see the trends and draw some valuable insights.

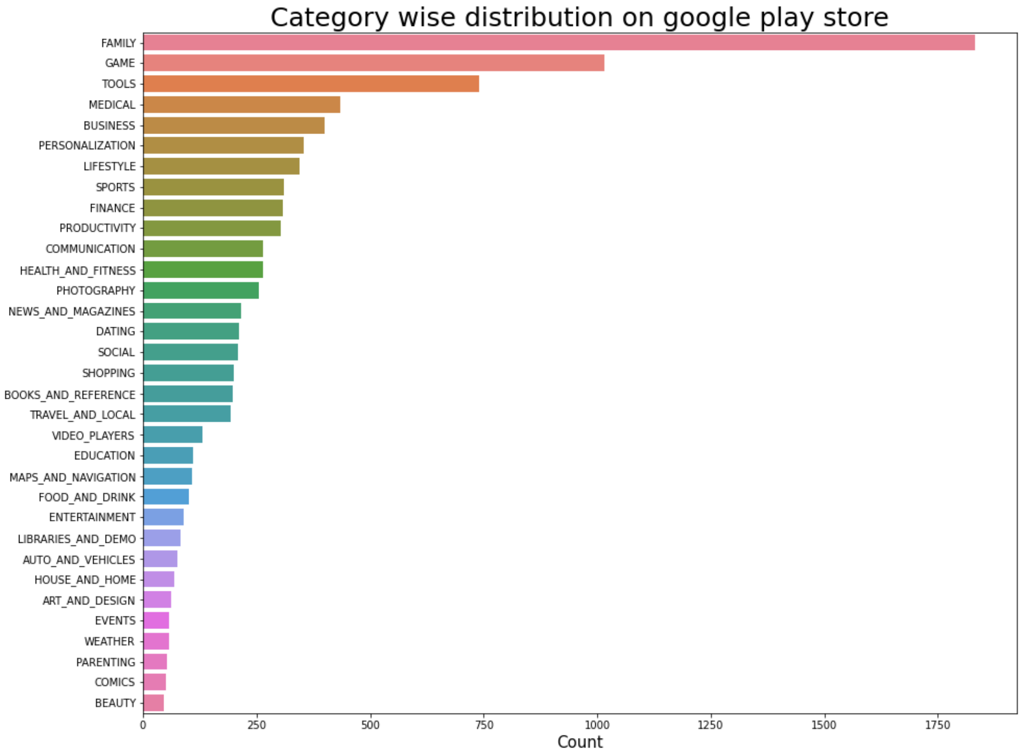
EDA

The first visualisation done was the distribution of the ratings category wise on a box plot. From this we can see that most ratings for the apps over all the categories are similar and lie between the values 3.8 and 4.8



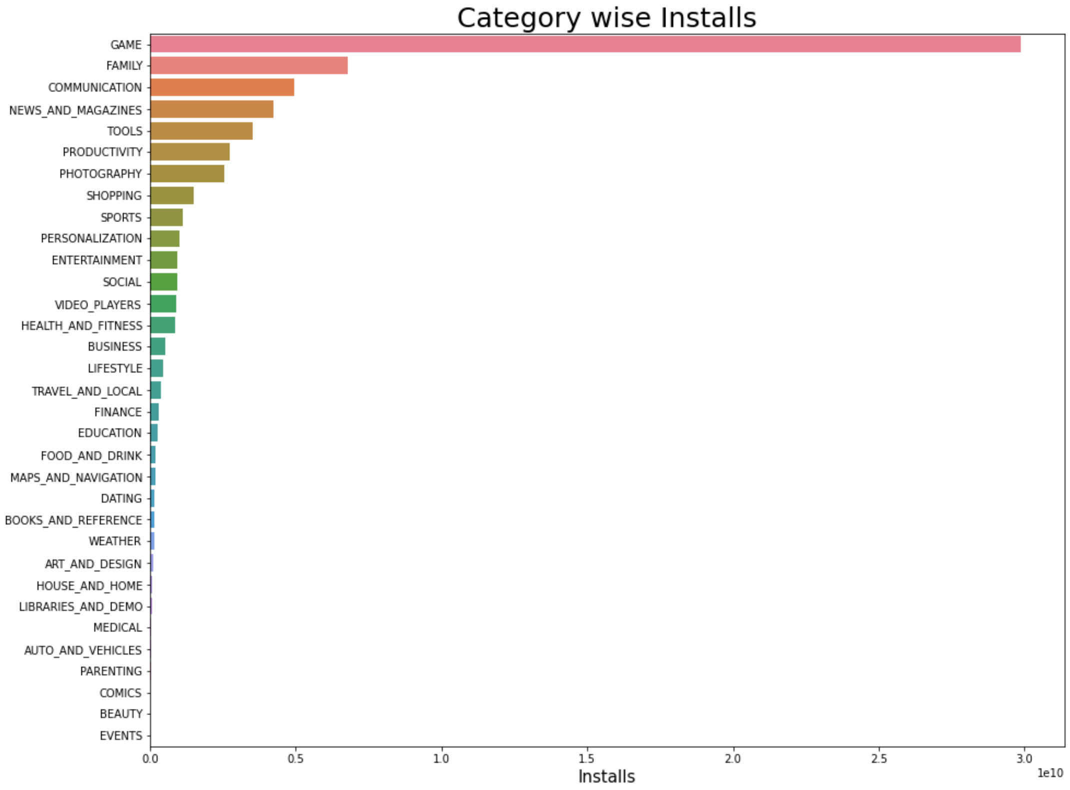
EDA(cont.)

From the plot of Category wise distributions on play store, we can see that the highest number of apps available on the play store belong the category 'Family'.



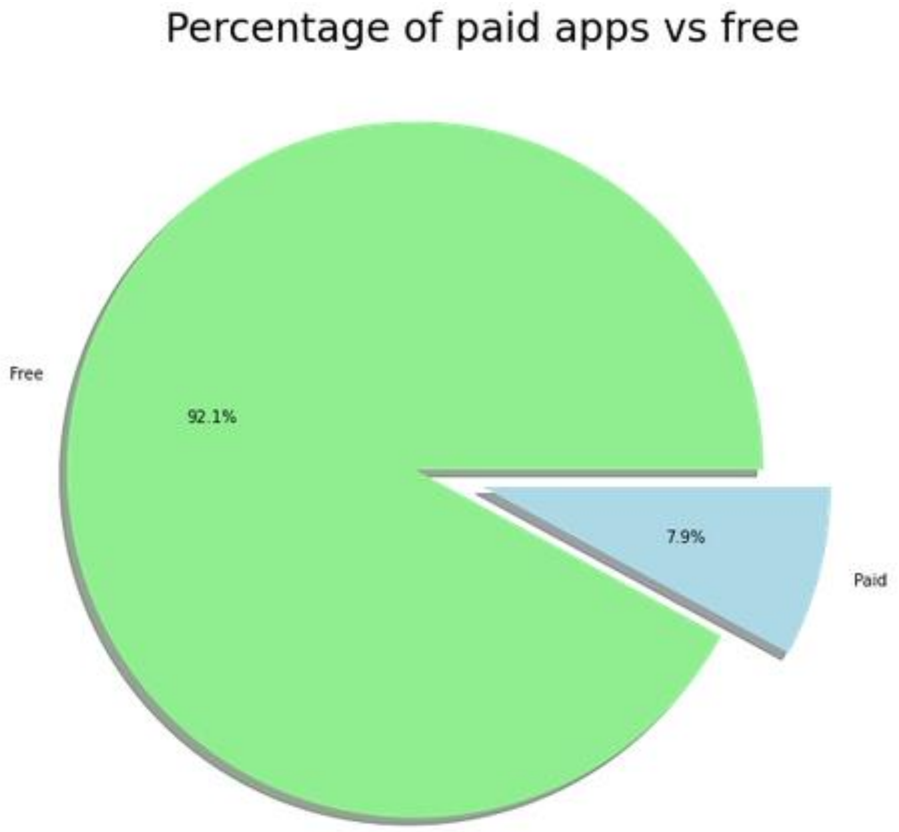
EDA(cont.)

From the plot of Category wise Installs on play store, we can see that the highest number of installs on the play store are from the category ‘Games’, despite the fact that ‘Family’ Category has the highest number of apps. Thus we can safely conclude that ‘Games’ is the most popular category among users on play store.



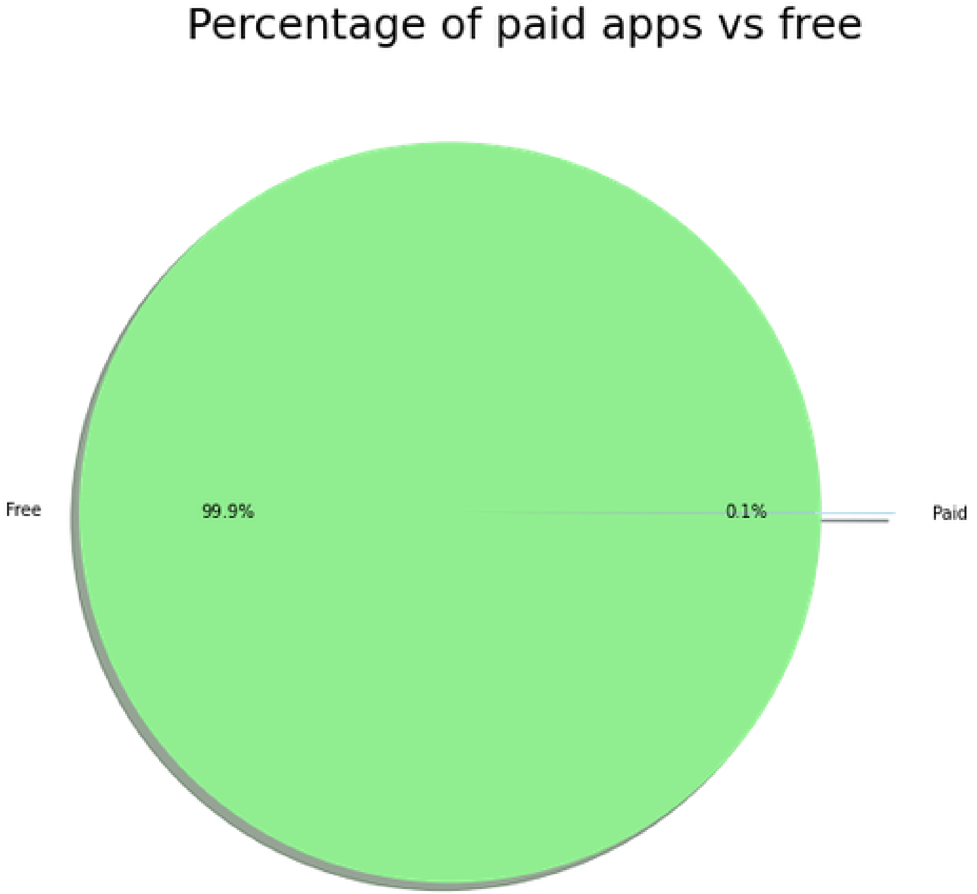
EDA(cont.)

The following pie chart shows the distribution of free and paid apps on the Play store. From the graph we can note that 92% of the apps on Google play store are Free and 7.9% are paid.



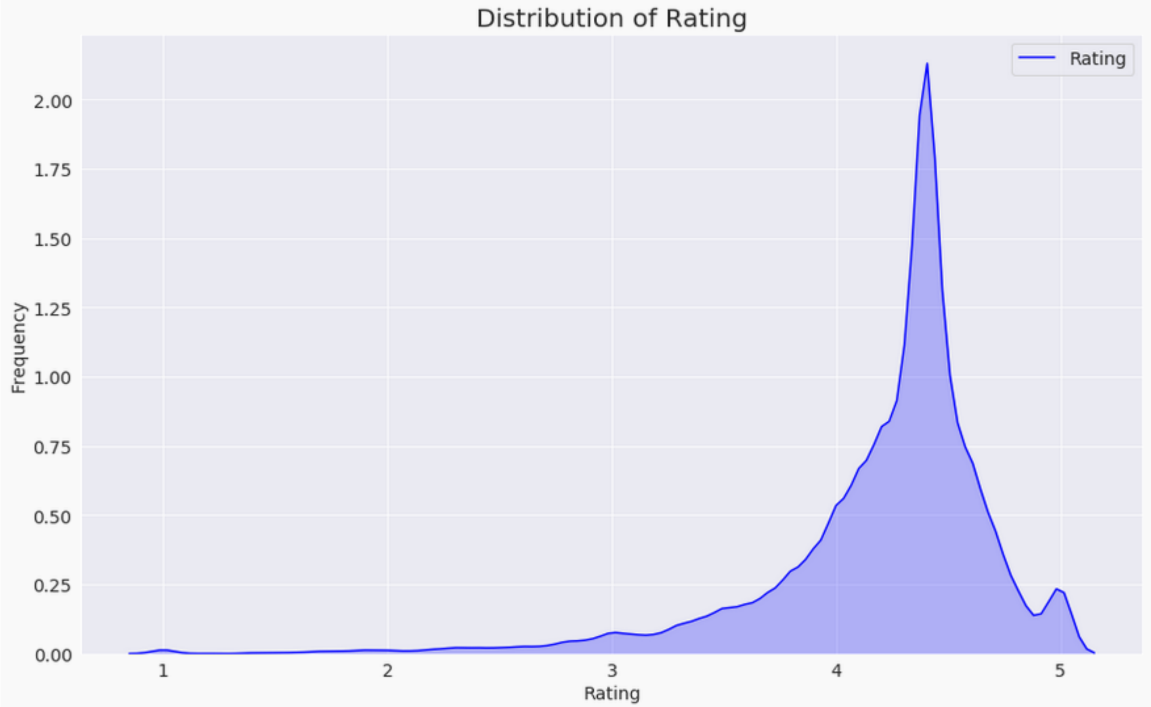
EDA(cont.)

The following pie chart shows the number of installs of free and paid apps on the Play store. The trend is very different compared to the distribution of apps, the number of paid apps on the play store make up 7.9% whereas the installs are a mere 0.1%. This shows reluctance amongst the play store users in installing paid apps on their devices.



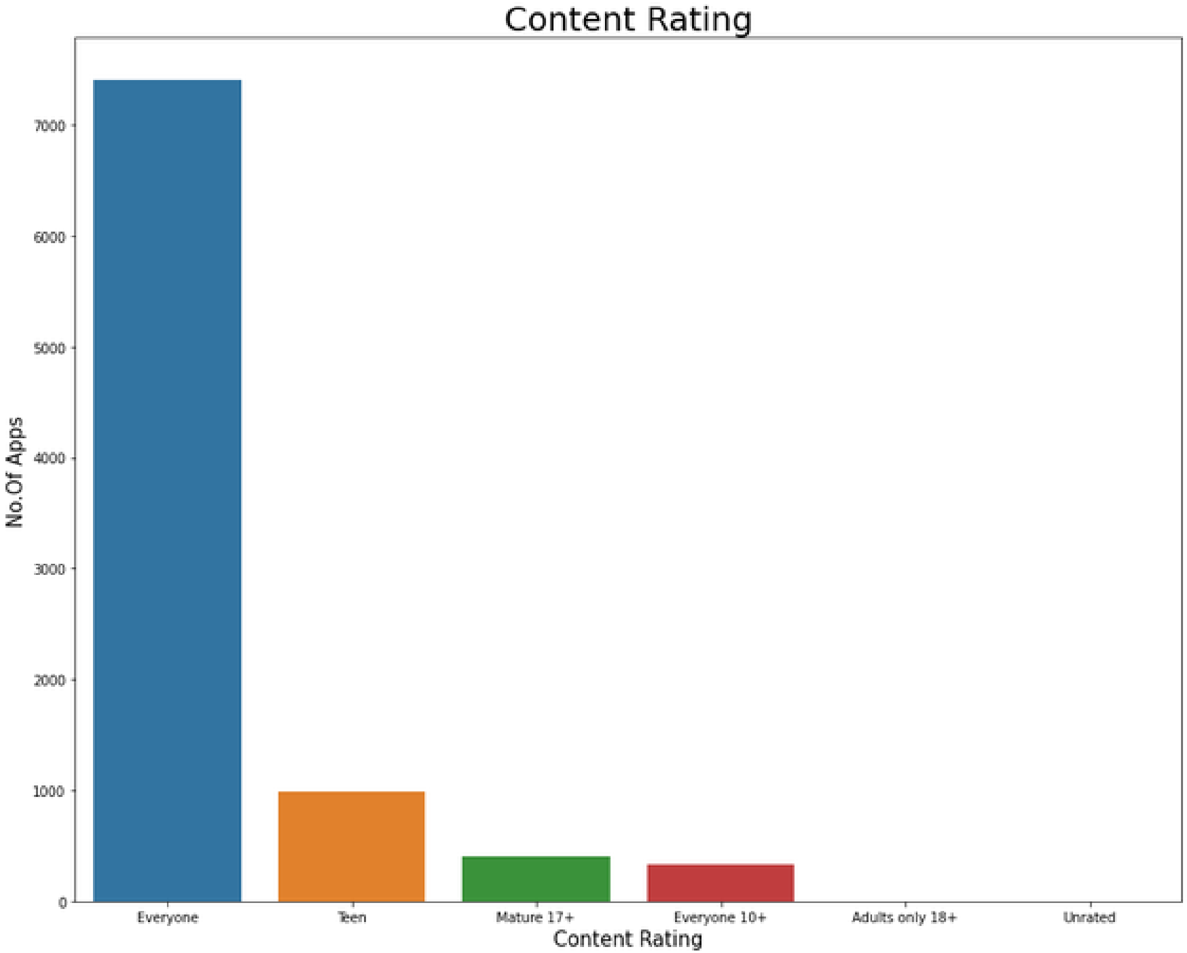
EDA(cont.)

The following distribution curve shows the distribution of ratings for all the apps combined. It can be seen that most of the apps in the play store are rated between 3.5 and 4.8. If an app has a rating that is below 3.5 then it shows that the app is in need of improvements.



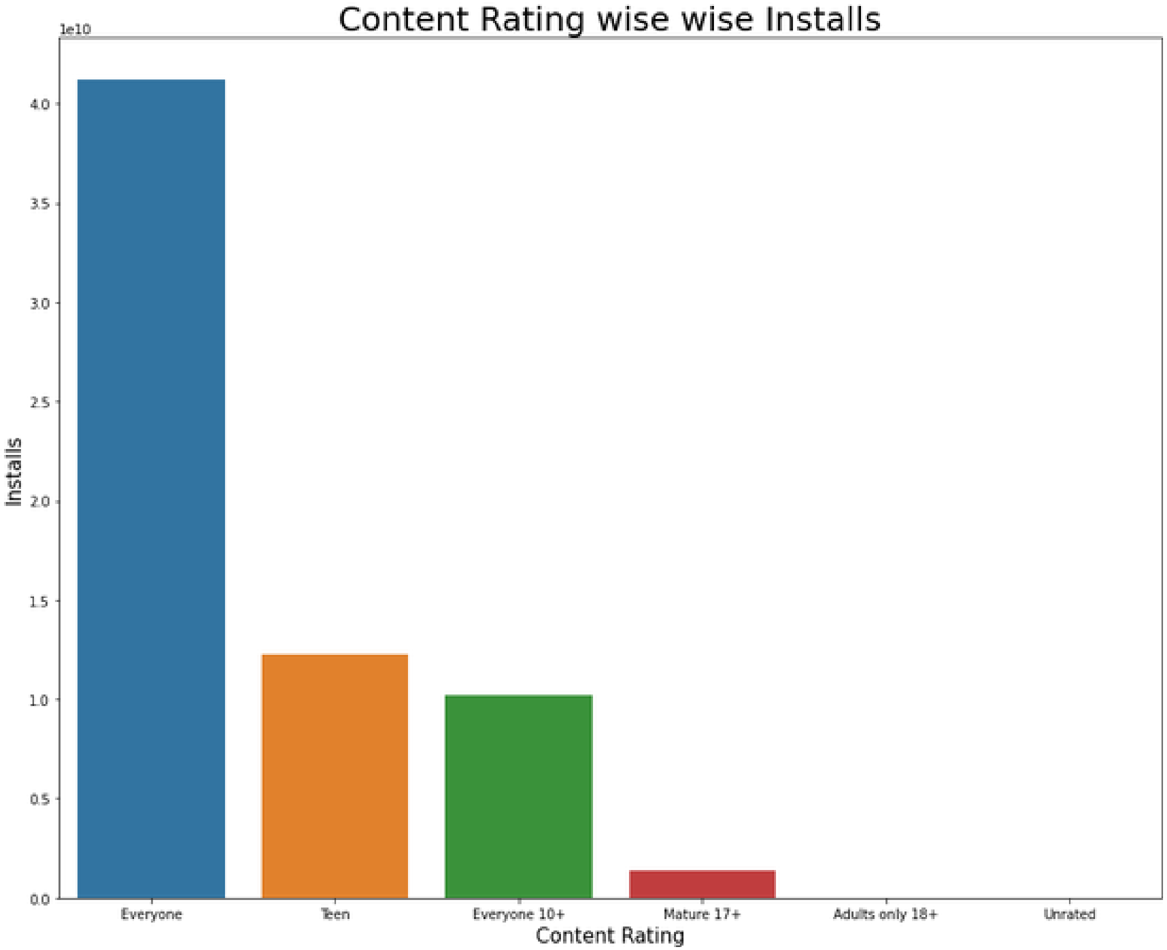
EDA(cont.)

The following bar plot shows the distribution of apps by content rating. From the graph we can se that most of the apps present on the play store(almost 70%), belong to the content rating 'Everyone'.



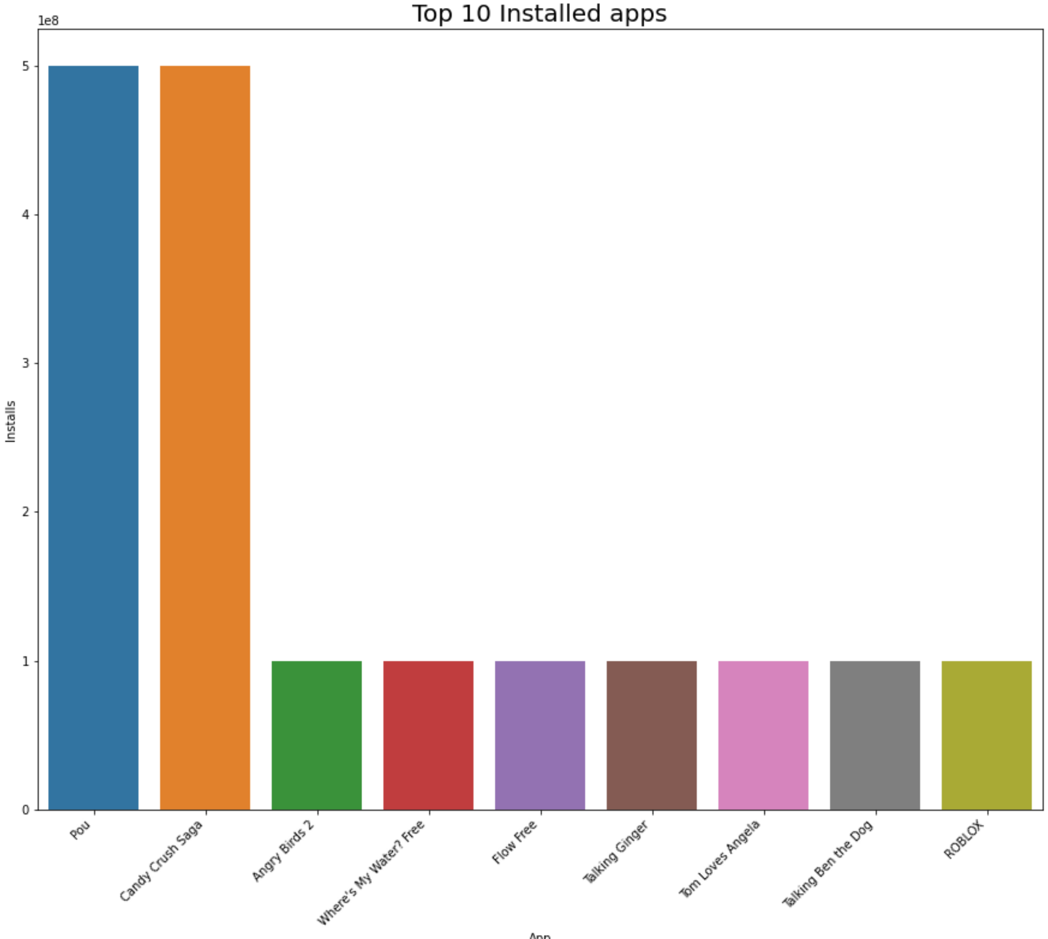
EDA(cont.)

The following bar plot shows the installs across content rating. The trend here is similar to the distribution of apps across content ratings as the content rating ‘Everyone’ occupies a majority of the installs followed by ‘Teen’.



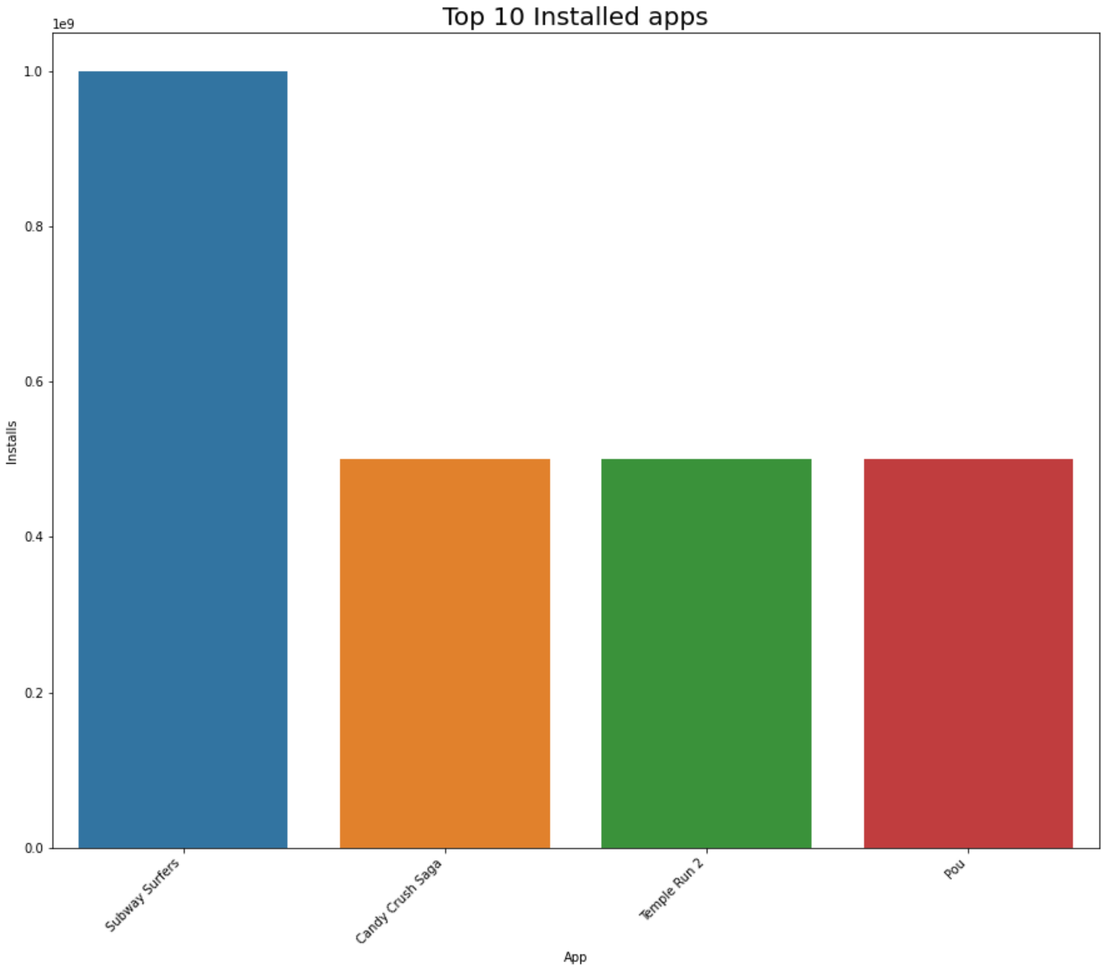
EDA(cont.)

The following bar plot shows the top installed apps in the category ‘family’, further looking into the play store reveals that most of these apps are either virtual pets or light, casual games.



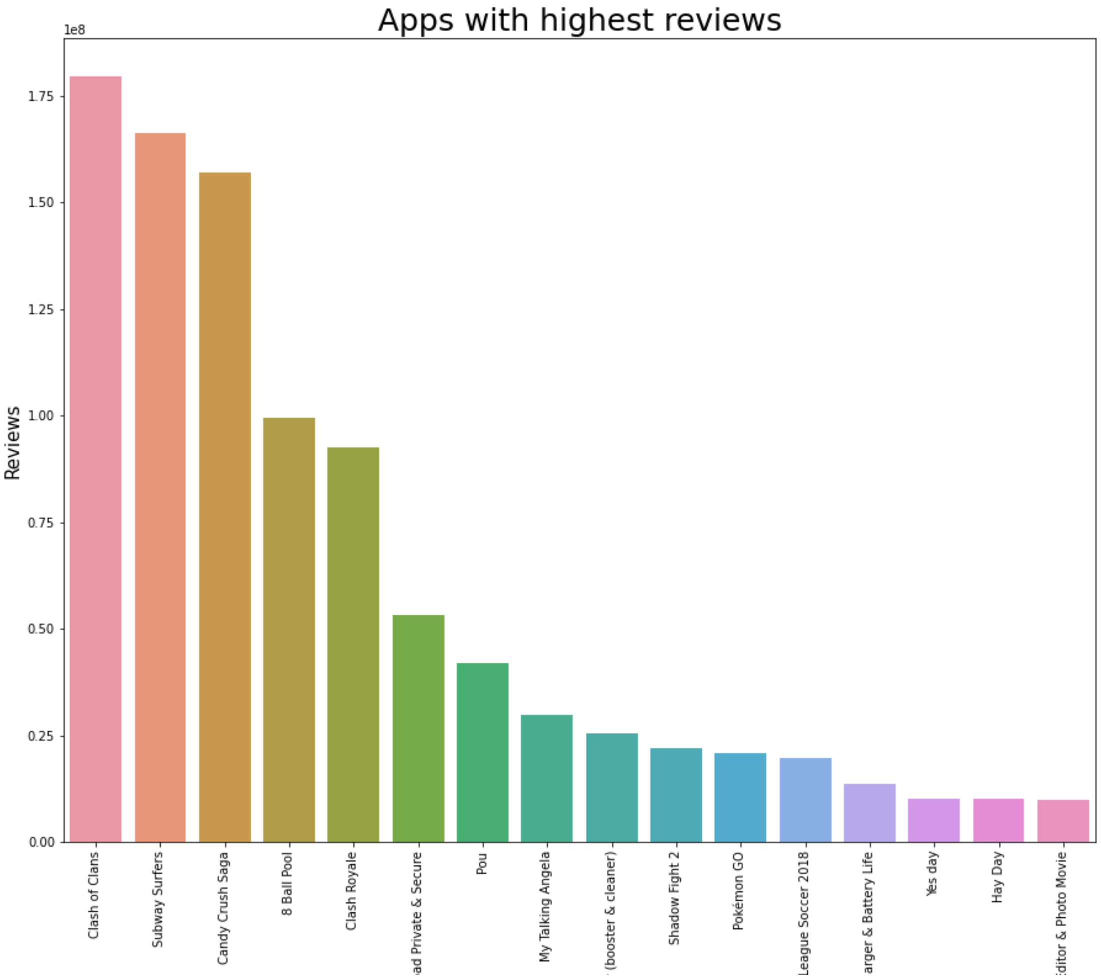
EDA(cont.)

The following bar plot shows the top installed apps in the ‘Games’ category. Further looking into the play store reveals that these apps are light, casual, single player games.



EDA(cont.)

The following bar plot shows the apps with the highest reviews. Most apps appearing in this list are the same apps that appear in the top apps installed in the ‘family’ and ‘game’ category.



Conclusion :

Depending on the data analysis of the play store data, if a suggestion was to be made, it would be recommended to create an app that resembles the virtual pet apps or light casual games that all members of the family can play and similarly enjoy. It would be highly recommended not to create an app that is paid as the data clearly shows reluctance among the play store users to install paid apps.