Capstone Project

EDA

Play store App Review Analysis

Team Members:

Mayank Ghai

Mrugesh Patel

Contents

Flow Chart of EDA Pipeline

Problem Statement

- One App installer Platform wants to know which apps to show & arrange on the homepage and which apps to recommend to customers
- They need to set up a homepage on their app which recommends apps that attracts more users.
- A search page where by default they can suggest top apps

1. Rank and sort App Categories based on their Installs and ratings
2. Determine the best genres and corresponding Apps for each category based on Installs and Ratings.
3. Which is the most distinct content rating type?
4. What are the most downloaded categories by most distinct content rating type?
5. Which are the 10 best and worst apps according to user Reviews?

Flow Chart of EDA Pipeline

Data Exploration and Reading

➢ First step in EDA to understand the dataset characteristics and data patterns

➢ Two Datasets, one with App meta-data and one with user reviews for various apps

➢ Play store app dataset has 12 different variables for each app entry

➢ User review dataset has reviews for 1074 unique apps and 4 features for each review entry

➢ There are several Null values in Rating column

➢ There are 33 distinct categories and more than 50 unique genres

➢ And there are 5 distinct content rating types

Flow Chart of EDA Pipeline
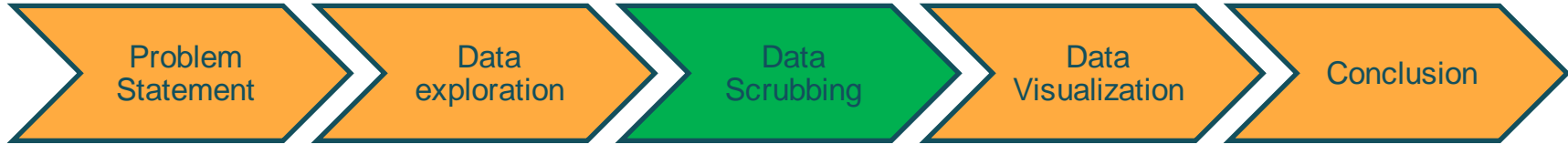
Data Scrubbing

- Next step in the EDA process to fix incorrect, incomplete, duplicate or otherwise erroneous data in a data set
- Two major processes:-

  - Data Cleaning :

    - Removing Null values, outliers, erroneous data, duplicates,

    - Removing Typographical or grammatical errors

  - Data manipulation

    - Replacing erroneous data with meaningful data

   - Converting variables to another data type for better processing

Data Scrubbing


```
[ ]  #checking the duplicated app entries in app column
     df.App.value_counts()

ROBLOX                                                    9
CBS Sports App - Scores, News, Stats & Watch Live        8
ESPN                                                      7
Duolingo: Learn Languages Free                           7
Candy Crush Saga                                         7
                                                         ..
Meet U - Get Friends for Snapchat, Kik & Instagram       1
U-Report                                                  1
U of I Community Credit Union                            1
Waiting For U Launcher Theme                             1
iHoroscope - 2018 Daily Horoscope & Astrology           1
Name: App, Length: 9659, dtype: int64
```

There are 9659 unique apps and several apps has multiple entries

• Duplicated entries removed from dataset

Distribution of Rating
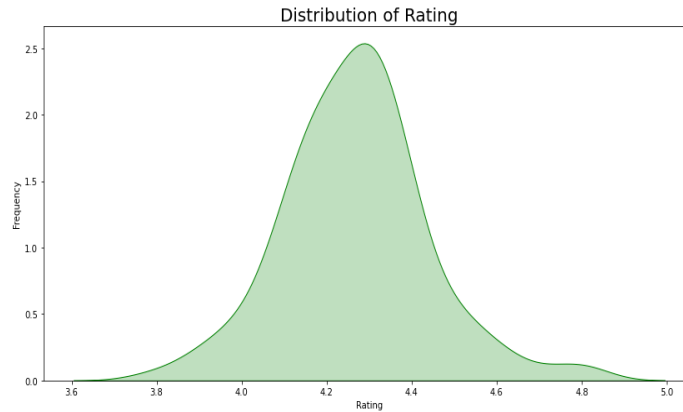
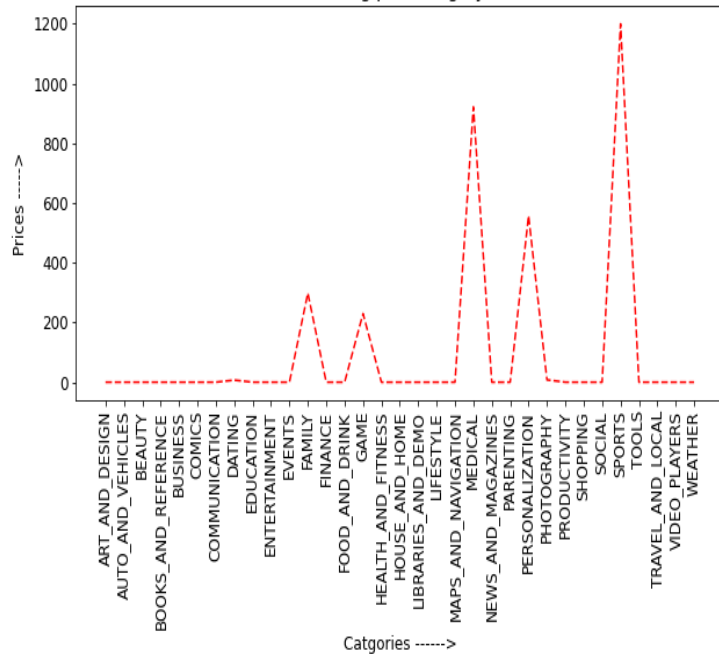• Null Values imputed with median of rating values

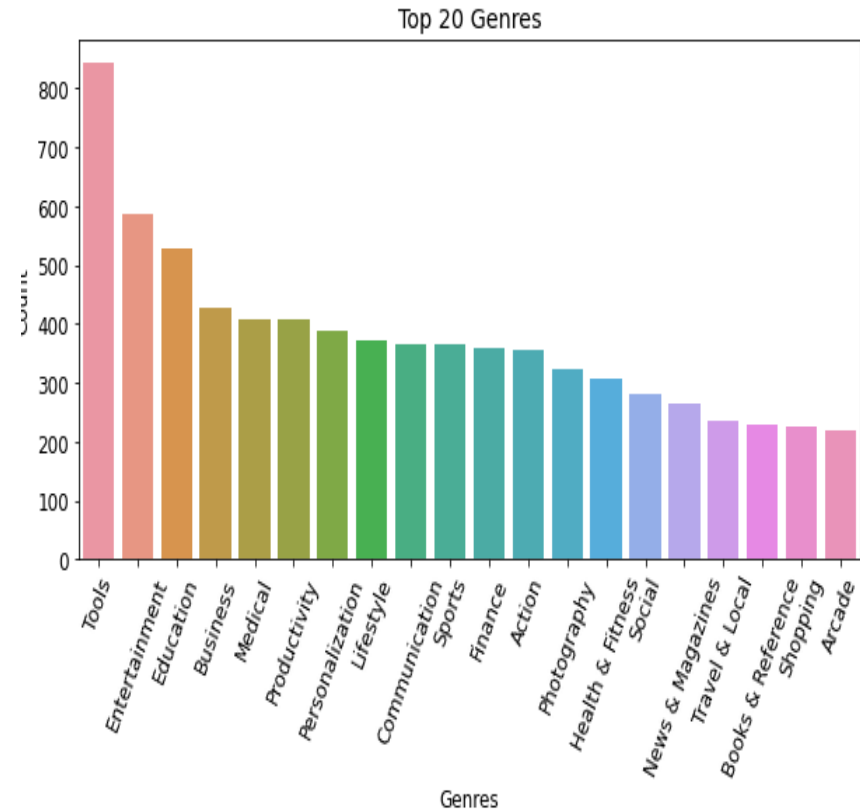• Values in size and Installs column converted for better usability
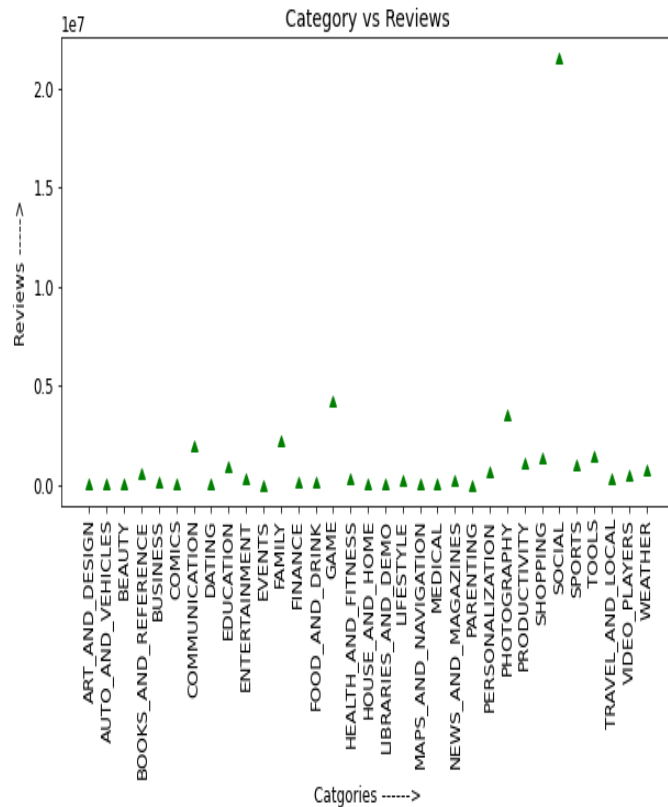
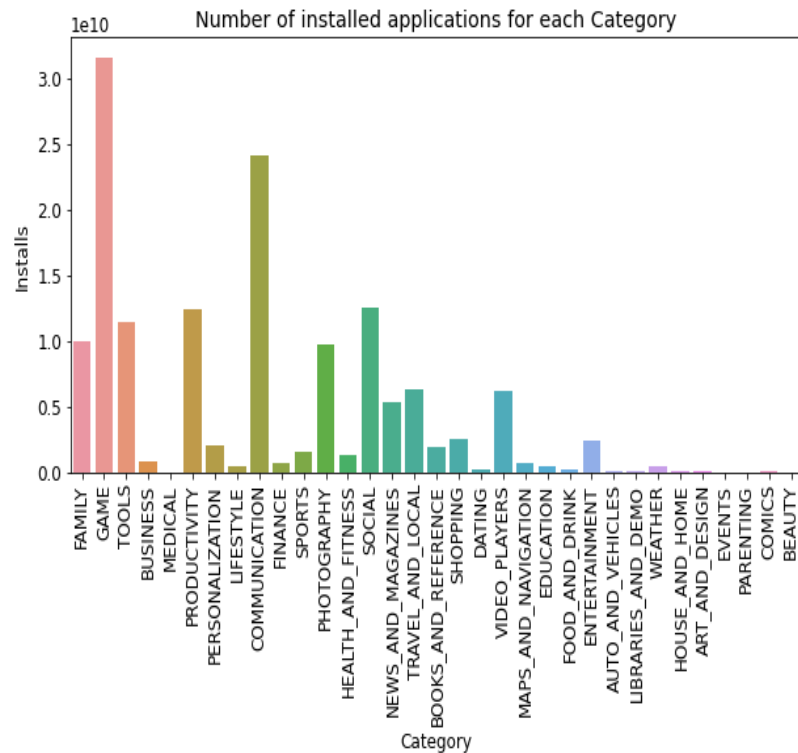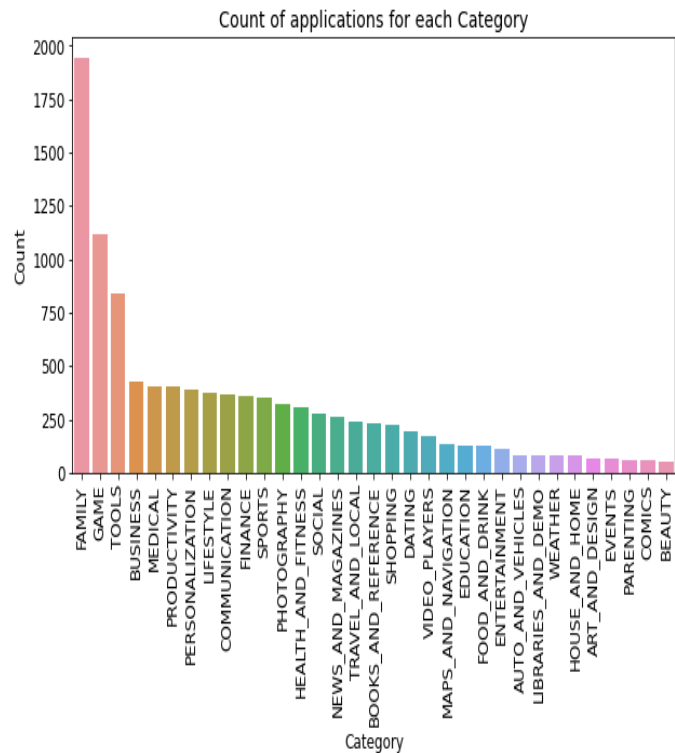# Data Visualisation

Category wise Installs
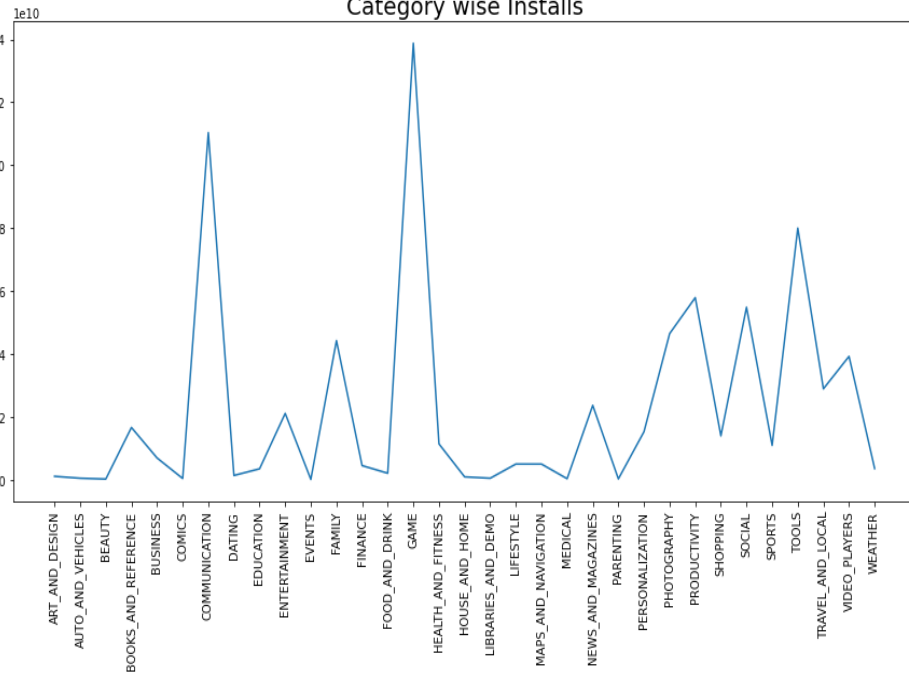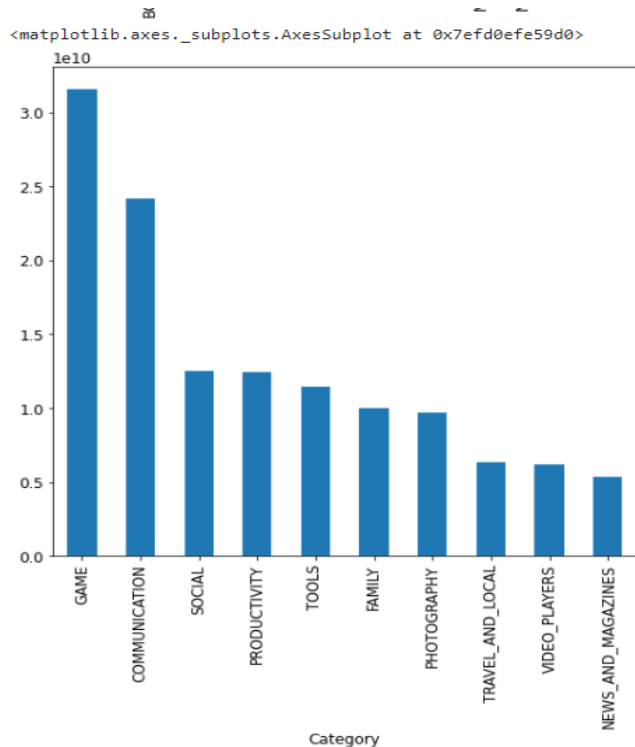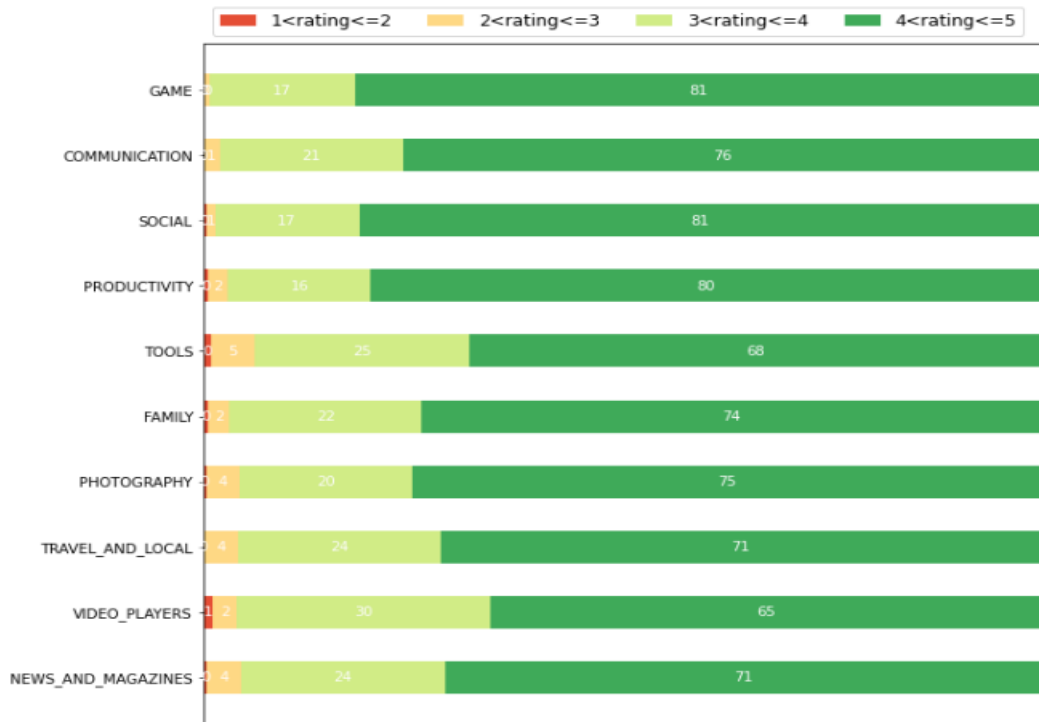

Pricing per Category

Data Visualisation

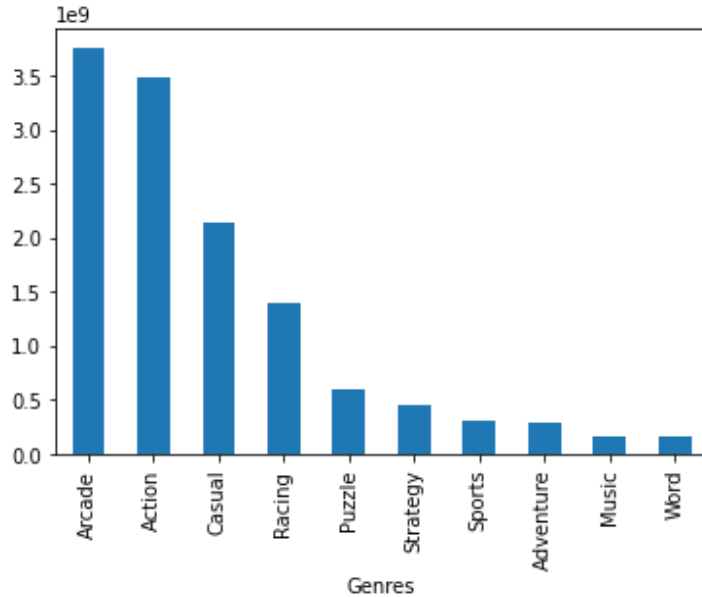# Data Visualisation

Data Visualisation

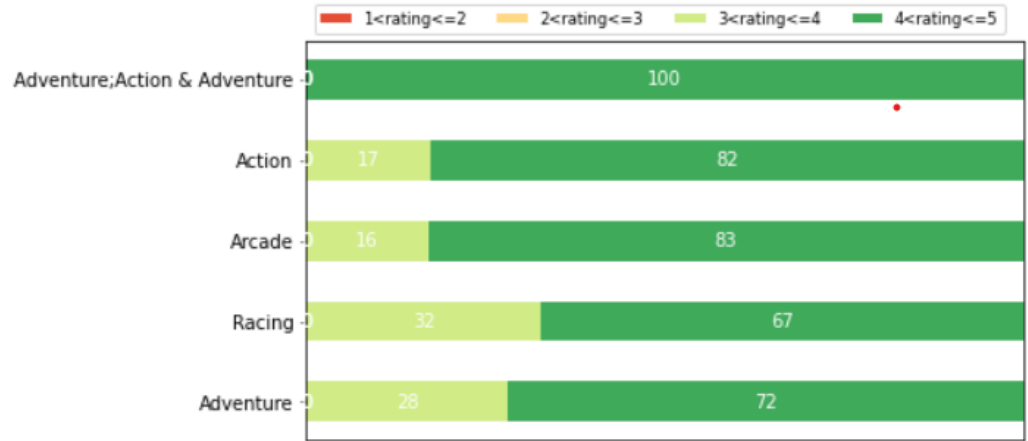*Number of Installs vs. category*

Stacked bar plot of rating bins for top categories according to installs
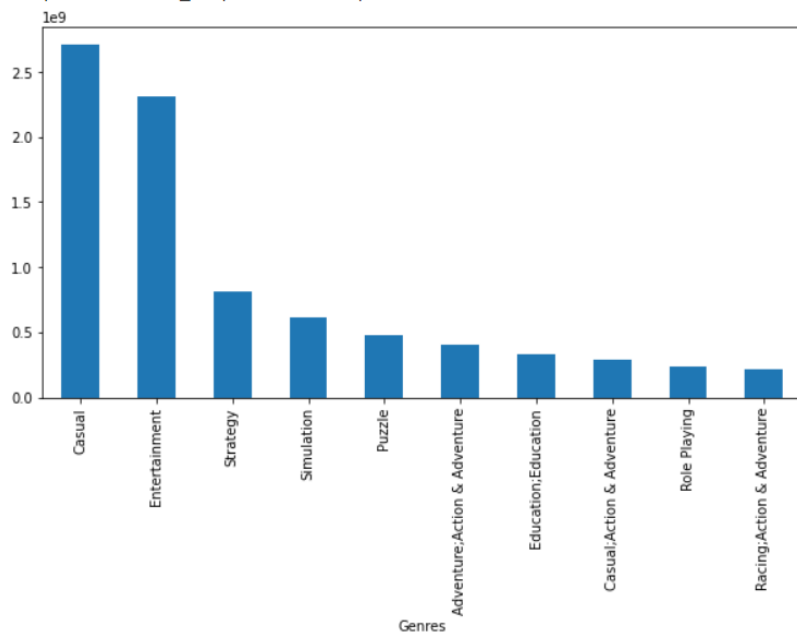
Data Visualisation

Number of Installs vs. category
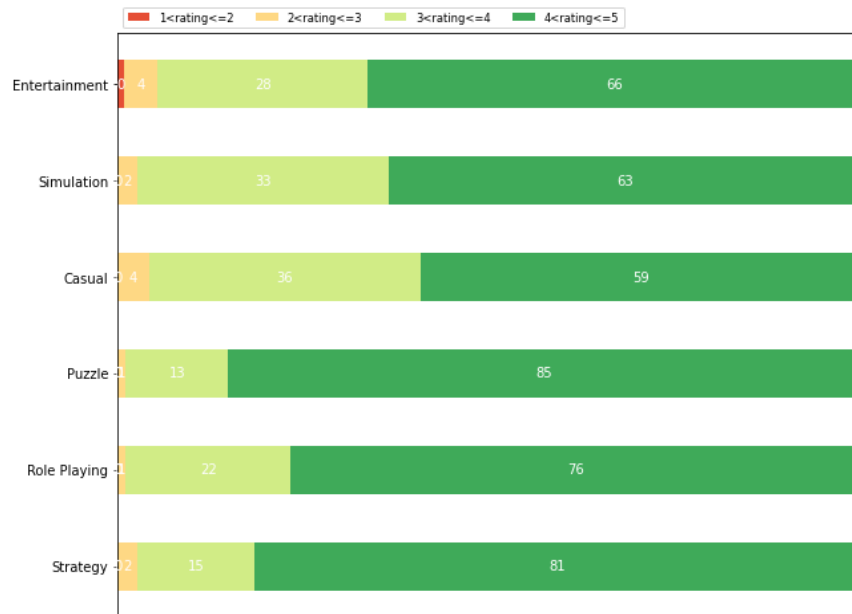(For Game Category)

Stacked bar plot of rating bins for top genres according to installs
(For Game Category)

Data Visualisation
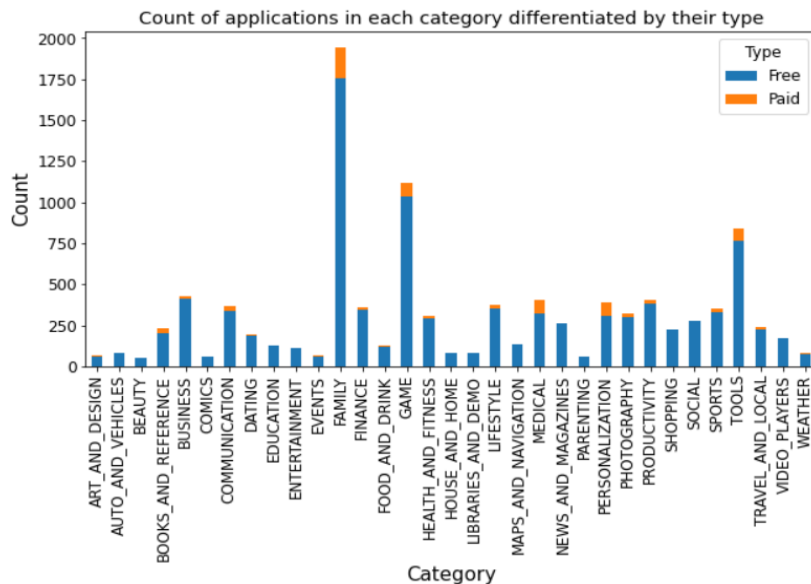

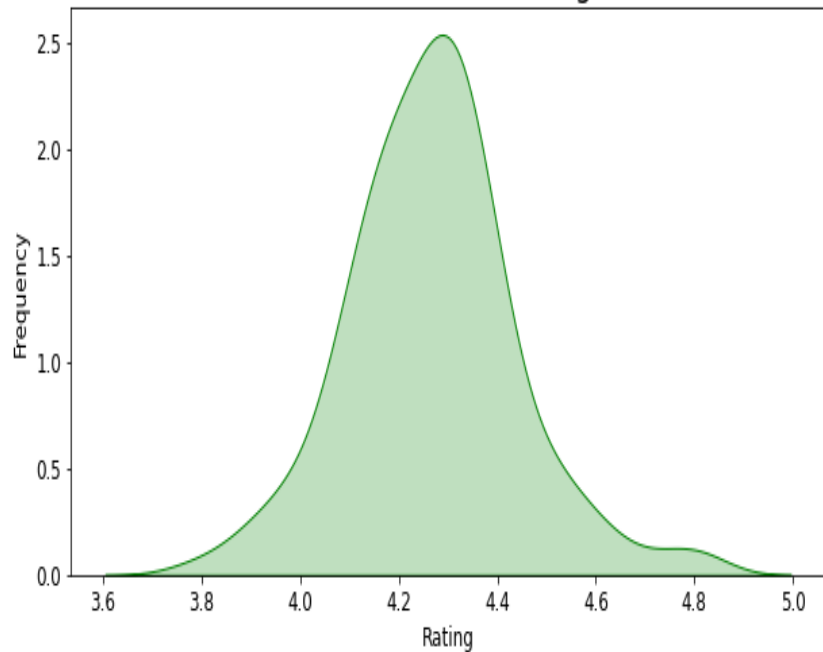*Number of Installs vs. category (For Family Category)*

Stacked bar plot of rating bins for top genres according to installs (For Family Category)
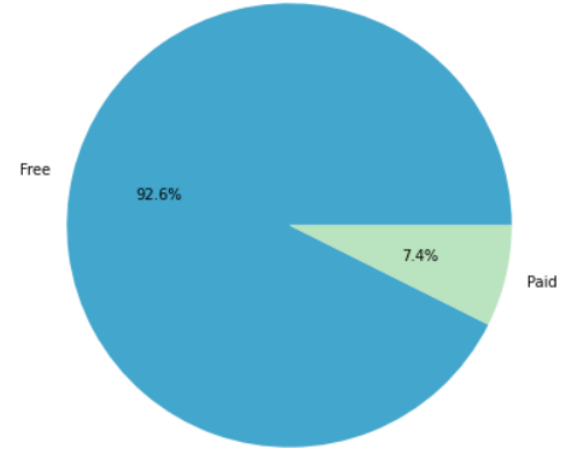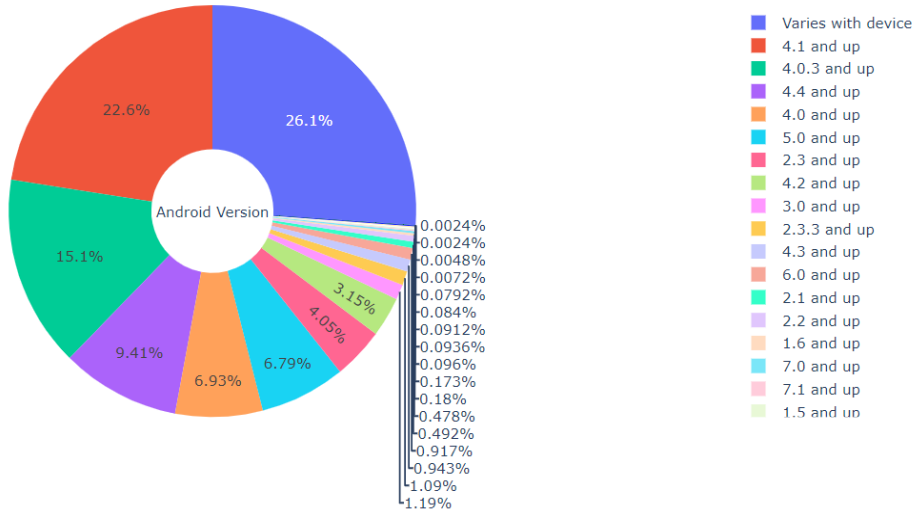
Data Visualisation

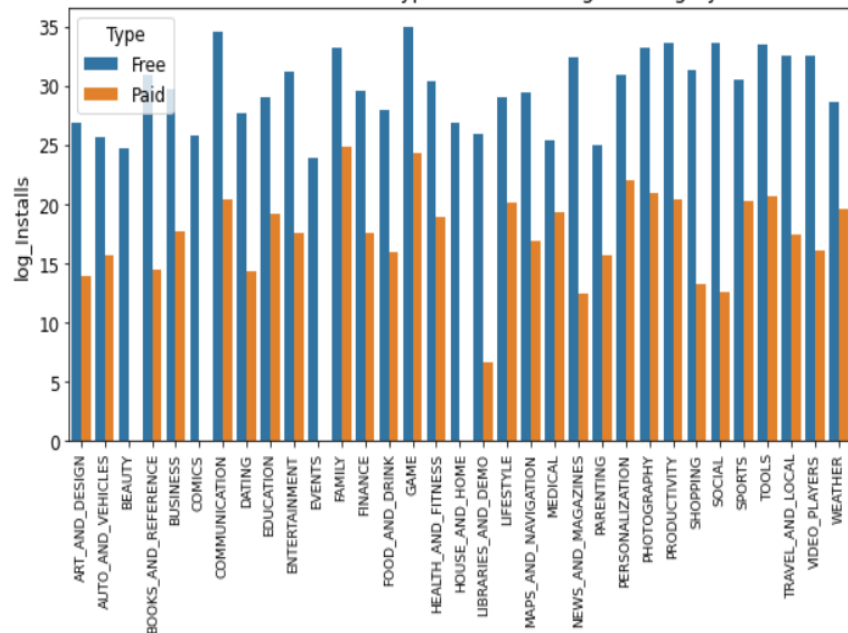# Data Visualisation

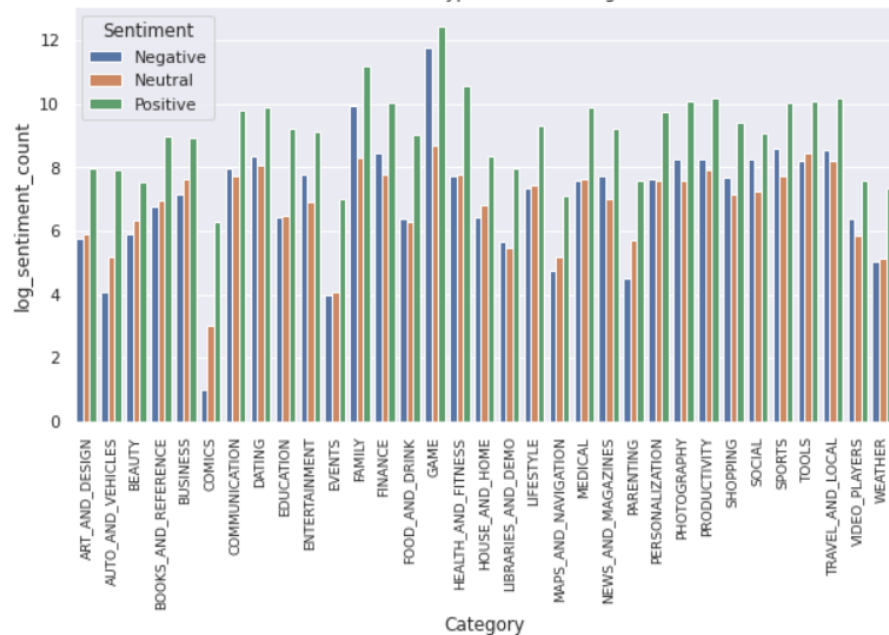**Different Android version**

What is the percentage of paid to free apps?

Data Visualisation

Number of installs type wise according to Category

Number of Reviews type wise according to Genres

Data Visualisation
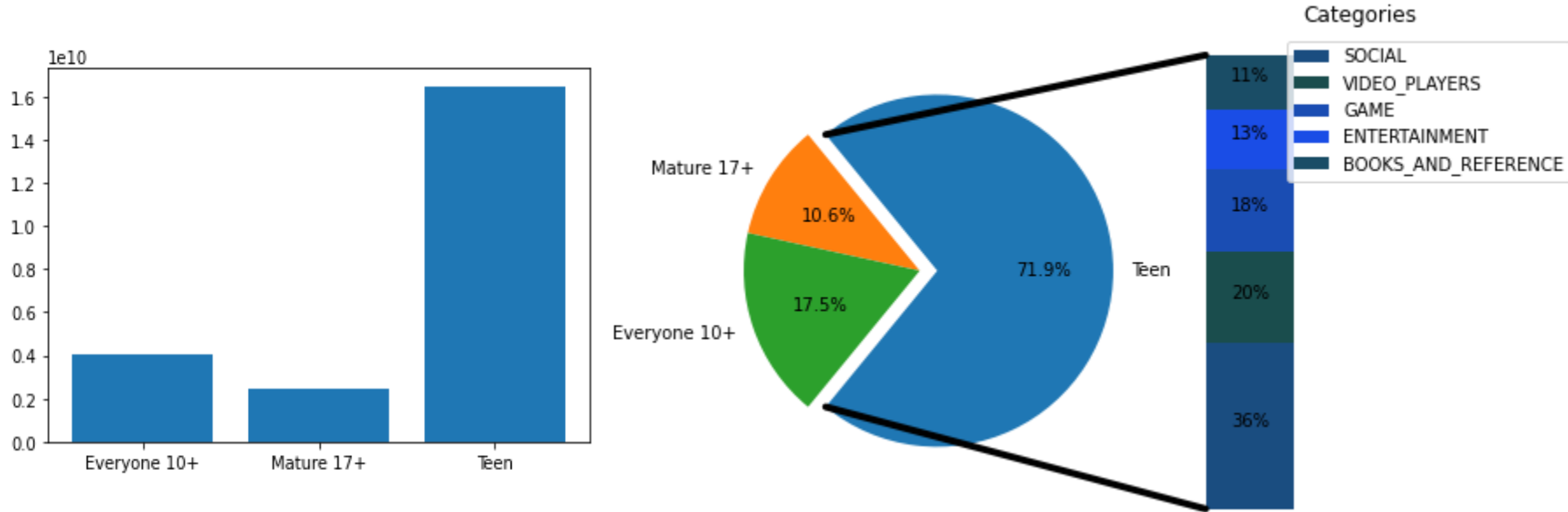
Pie chart of content rating type proportions with popular
categories for teen type

Review Analysis

A Pie Chart Representing Percentage of Review Sentimets

Legend:
- Positive Reviews
- Negetive Reviews
- Neutral Reviews

Positive Reviews: 64.65%
Neutral Reviews: 12.17%
Negetive Reviews: 23.18%

Does sentiment_subjectivity proportional to sentiment_polarity
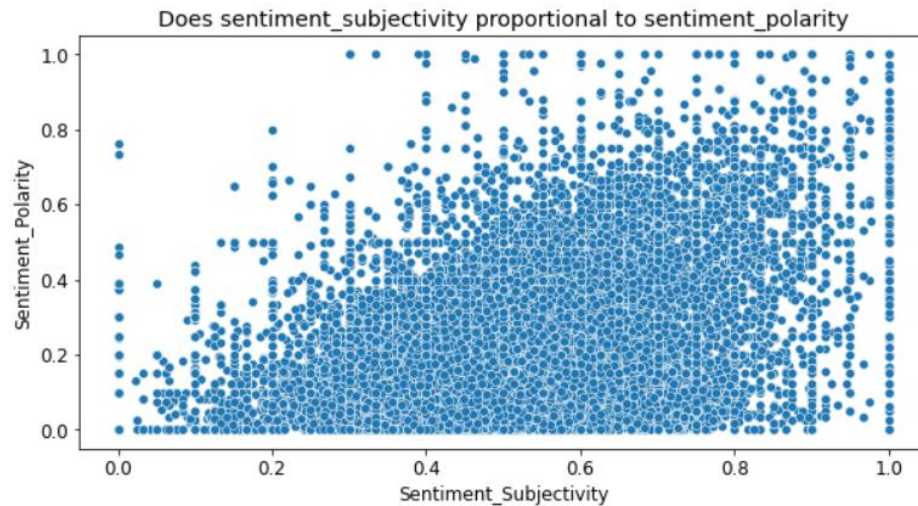
Data Visualisation

Flow Chart of EDA Pipeline

Conclusion

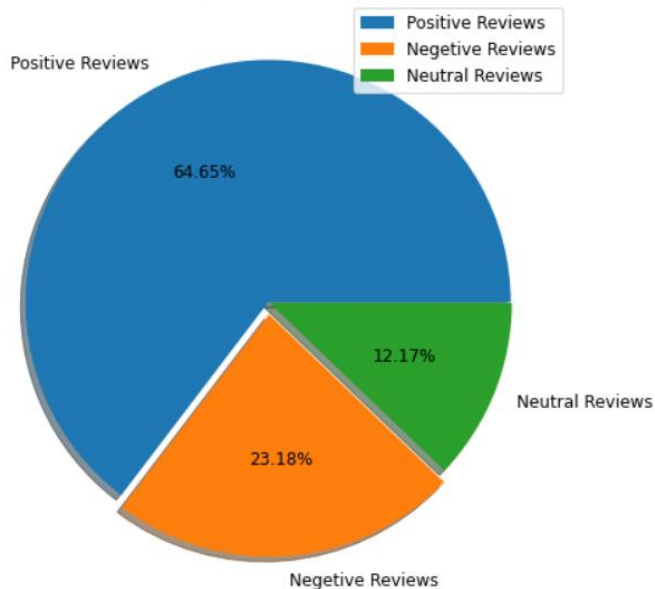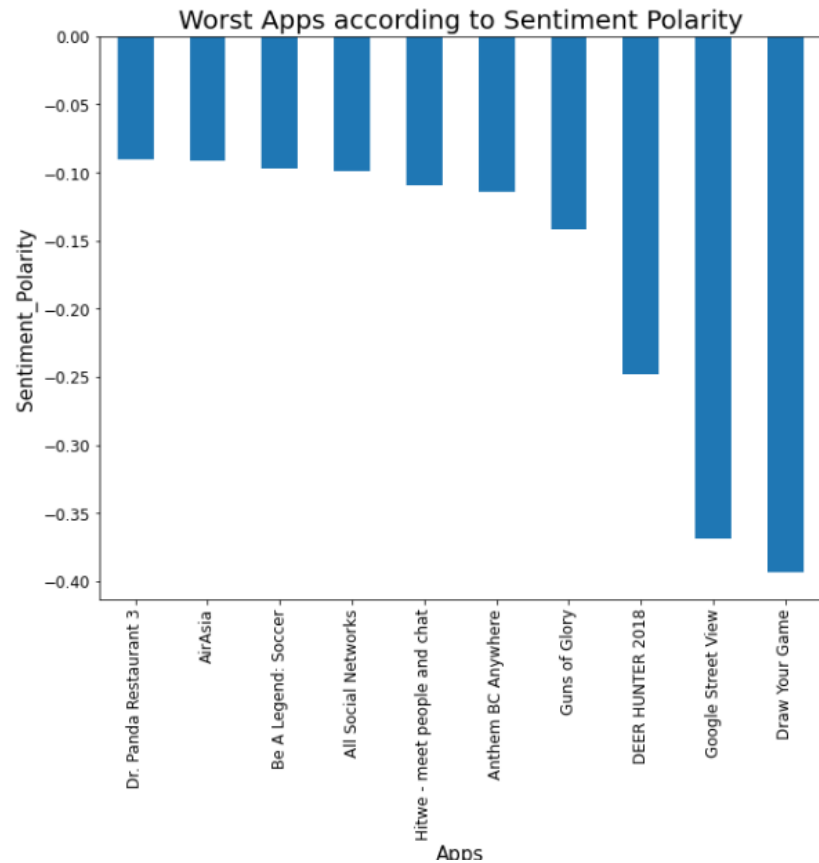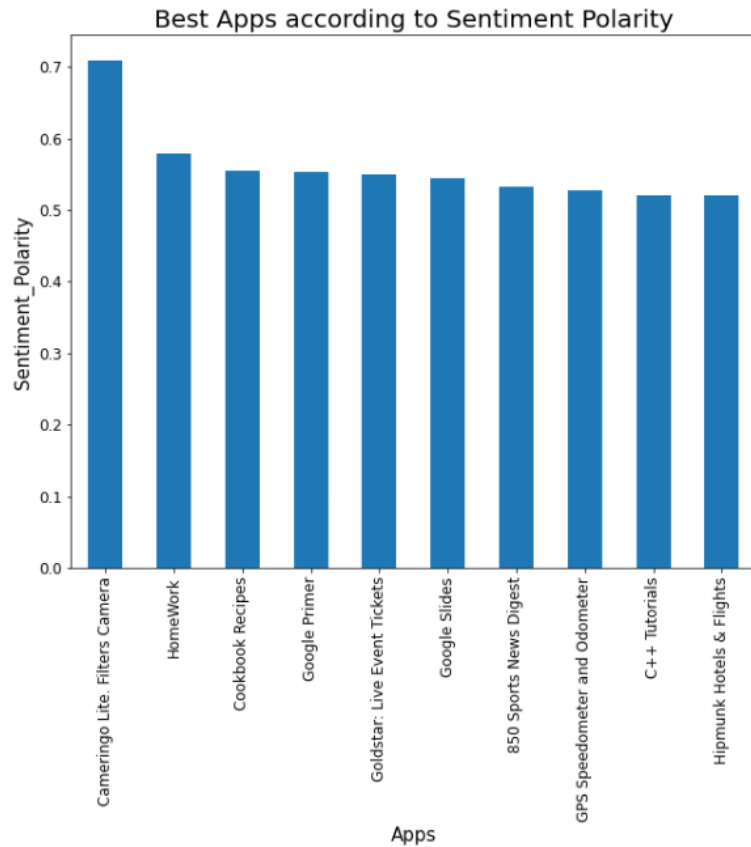- The top categories, *Game, Communication, Tools, Social, Productivity, and Photography.*

- The top categories with the highest pricing are *Sports, Tools, Social and medical*

- The Category that has the highest number of reviews are *Social, Games, Photography, Family, and Communication*.

- The maximum number of apps present in the google play store comes under *Tools, Entertainment, and Education* Genres but as per the installation and requirement in the market plot, the scenario is not the same. Maximum installed apps come under *Communication, Tools, and Productivity* Genres.

- As we can see that the application under *Family, Games and Tools* Genres has the highest number of applications under their category.

- For the Game Category, the Top three Genres are *Arcade, Action, and Casual.*

- Family Category, Top three Genres **Casual,** *Entertainment, and Strategy.*

- It can be concluded that the number of free applications installed by the user is high when compared with the paid ones.

- The most popular content rating type is *Teen*, which means that the most popular age demographic to look for is of age group 11-16.

- And the most popular categories for Teen type content rating are, *Social, Video_players, Game, Entertainment, and Books_and_reference* in that order.

Conclusion

- Among all the given reviews that are provided by the user, 64.65% of the reviews come under positive, 23.18% are negative while the rest comes under neutral.

- It can be seen that a maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this, we can conclude that a maximum number of users give reviews to the applications, according to their experience.

- The best applications according to Sentiment Polarity are *Cameringo Lite. Filters Camera, HomeWork, and Cookbook Recipes.*

- The worst application according to Sentiment Polarity is *Draw Your Game, Google Street View, and Deer Hunter 2018*

THANK YOU