Capstone Project

EDA

Play store App Review Analysis

Team Members:

Mayank Ghai

Mrugesh Patel

Contents

1. Problem Statement
2. Data Reading and Exploration
3. Data Cleaning and Manipulation
4. Data Visualization
5. Conclusion from insights

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

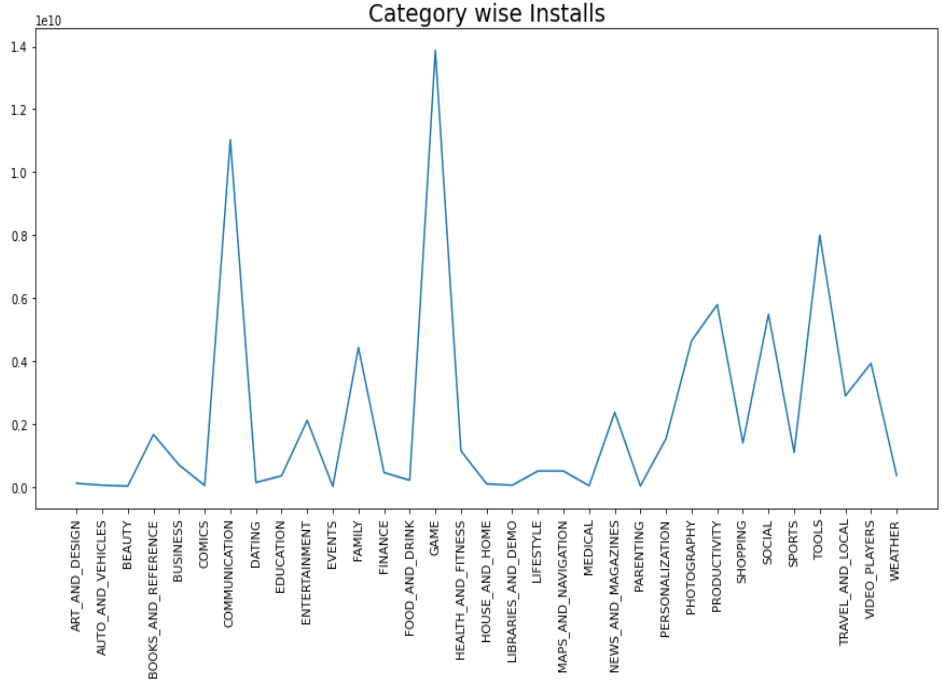
- First step in EDA to understand the dataset characteristics and data patterns
- Two Datasets, one with App meta-data and one with user reviews for various apps
- Play store app dataset has 12 different variables for each app entry
- User review dataset has reviews for 1074 unique apps and 4 features 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

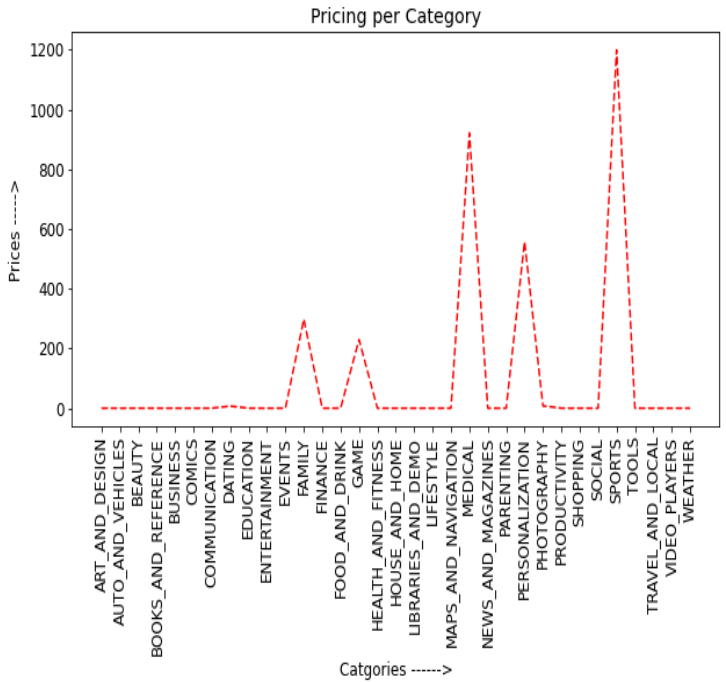
Flow Chart of EDA Pipeline

Data Scrubbing

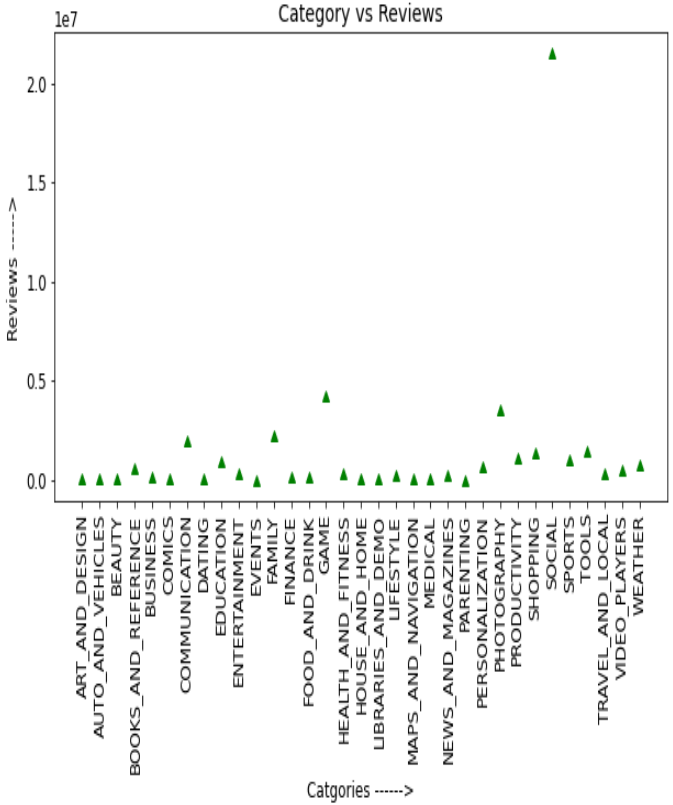
- **Next step in the EDA process to fix incorrect, incomplete, duplicate or otherwise erroneous data in a data set**
- **Two major processes:-**
  - **Data Cleaning :**
    - **Removing Null values, outliers, erroneous data, duplicates,**
    - **Removing Typographical or grammatical errors**
  - **Data manipulation**
    - **Replacing erroneous data with meaningful data**
    - **Converting variables to another data type for better processing**

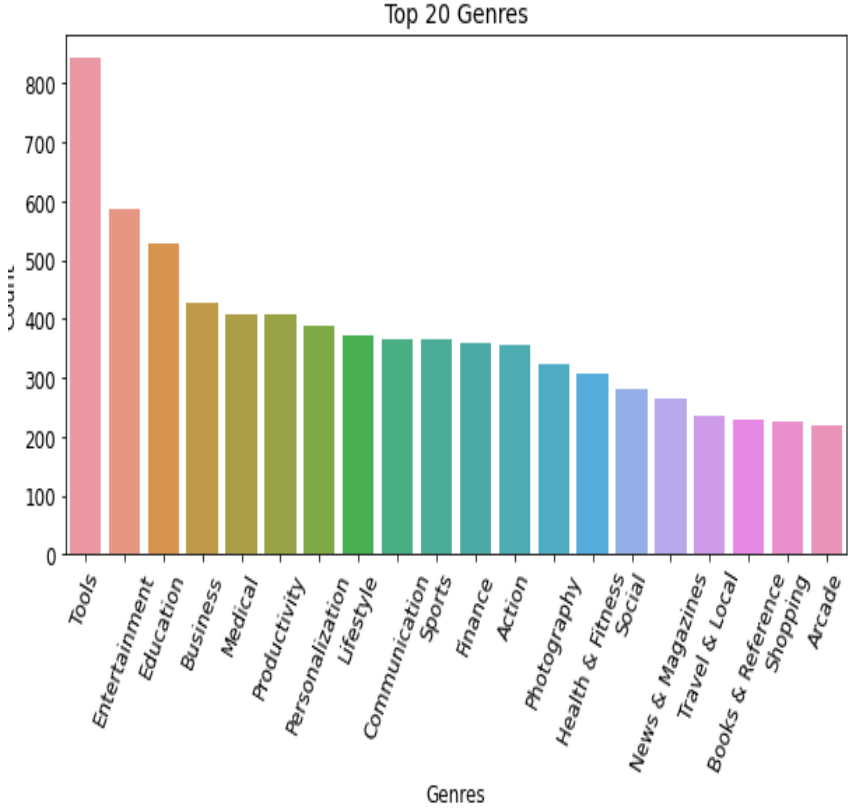
Data Visualisation

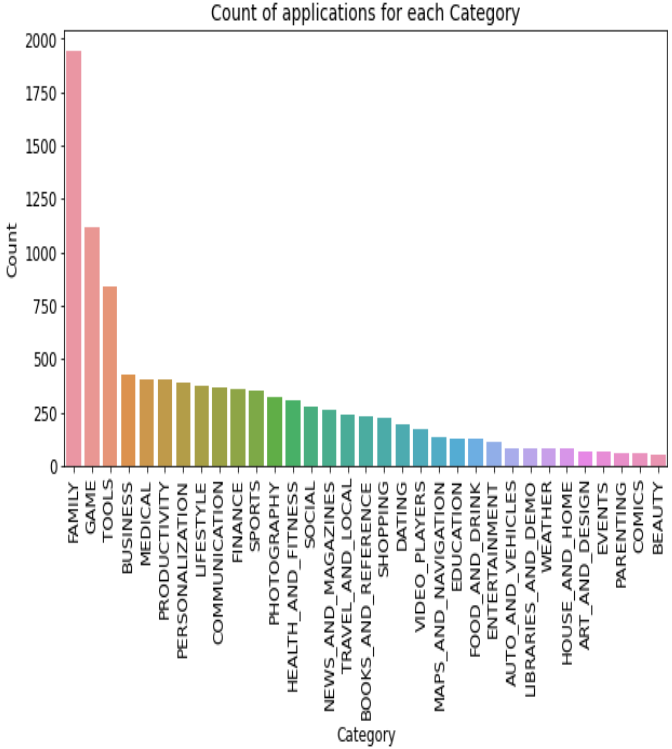


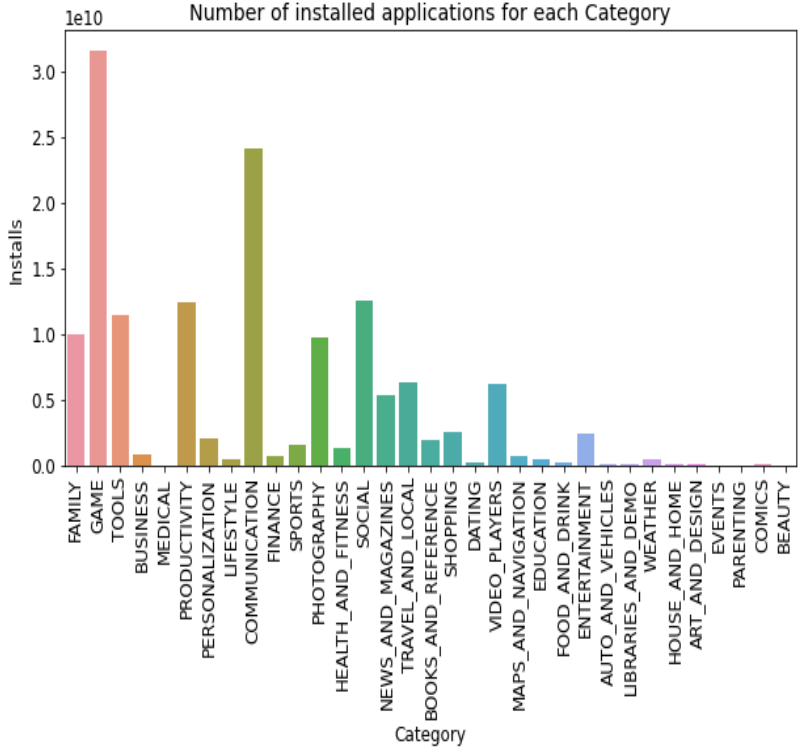


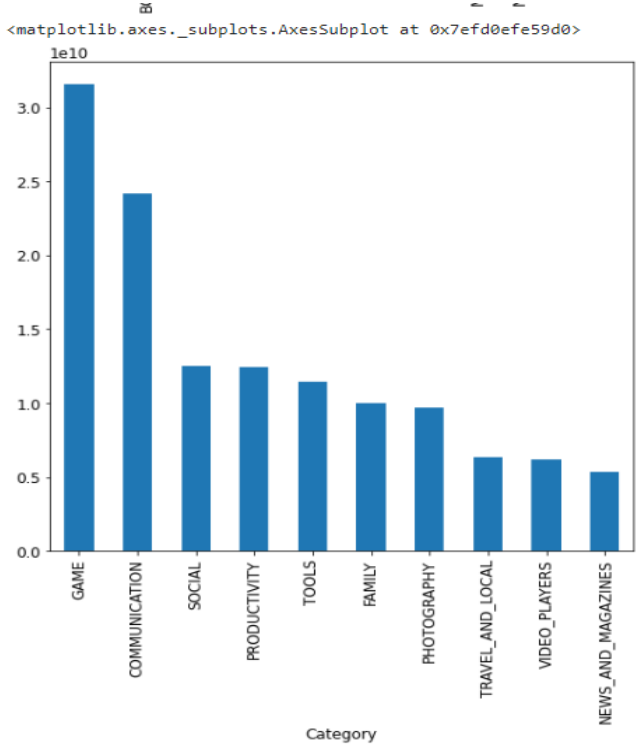
Data Visualisation



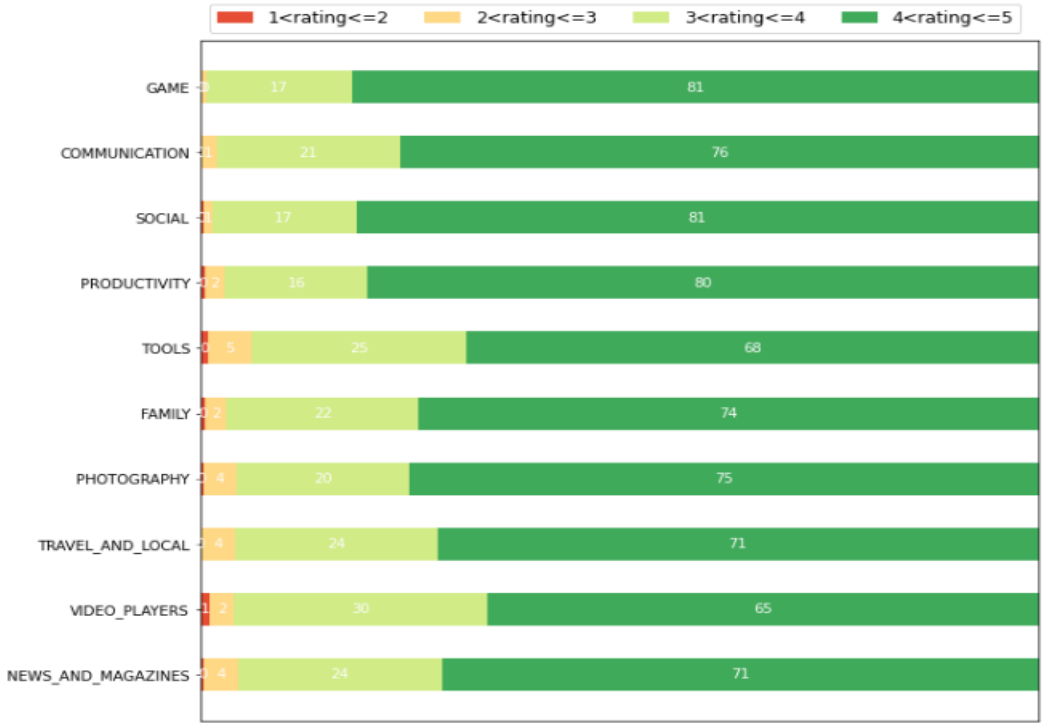



Count of applications for each Category
Count
Category
2000
1750
1500
1250
1000
750
500
250
0
FAMILY
GAME
TOOLS
BUSINESS
MEDICAL
PRODUCTIVITY
PERSONALIZATION
LIFESTYLE
COMMUNICATION
FINANCE
SPORTS
PHOTOGRAPHY
HEALTH_AND_FITNESS
SOCIAL
NEWS_AND_MAGAZINES
TRAVEL_AND_LOCAL
BOOKS_AND_REFERENCE
SHOPPING
DATING
VIDEO_PLAYERS
MAPS_AND_NAVIGATION
EDUCATION
FOOD_AND_DRINK
ENTERTAINMENT
AUTO_AND_VEHICLES
LIBRARIES_AND_DEMO
WEATHER
HOUSE_AND_HOME
ART_AND_DESIGN
EVENTS
PARENTING
COMICS
BEAUTY

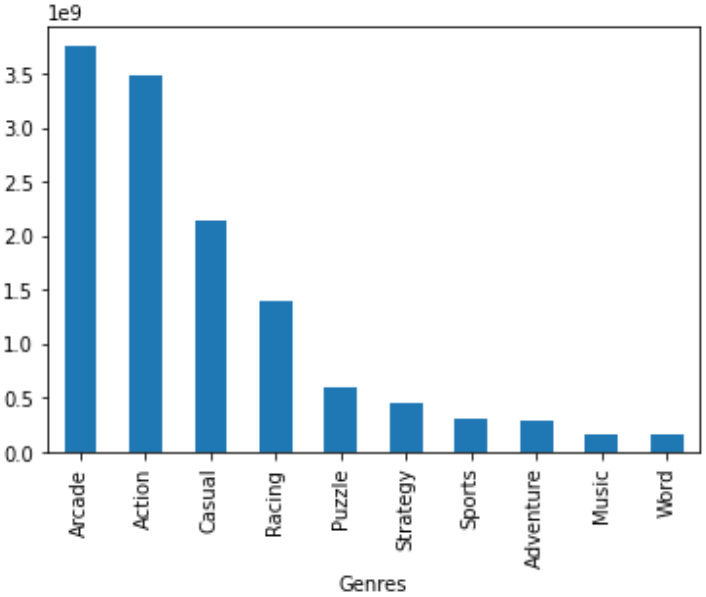

Number of installed applications for each Category
1e10
Installs
Category
3.0
2.5
2.0
1.5
1.0
0.5
0.0
FAMILY
GAME
TOOLS
BUSINESS
MEDICAL
PRODUCTIVITY
PERSONALIZATION
LIFESTYLE
COMMUNICATION
FINANCE
SPORTS
PHOTOGRAPHY
HEALTH_AND_FITNESS
SOCIAL
NEWS_AND_MAGAZINES
TRAVEL_AND_LOCAL
BOOKS_AND_REFERENCE
SHOPPING
DATING
VIDEO_PLAYERS
MAPS_AND_NAVIGATION
EDUCATION
FOOD_AND_DRINK
ENTERTAINMENT
AUTO_AND_VEHICLES
LIBRARIES_AND_DEMO
WEATHER
HOUSE_AND_HOME
ART_AND_DESIGN
EVENTS
PARENTING
COMICS
BEAUTY



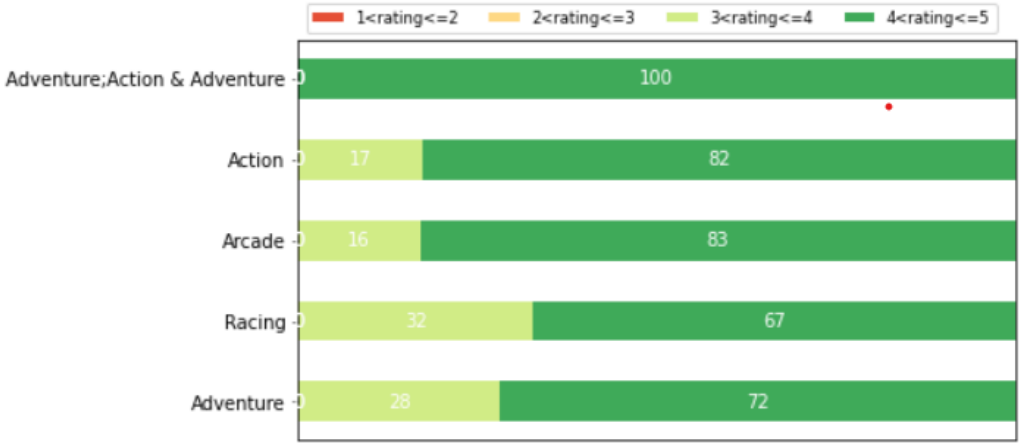
Number of Installs vs. category



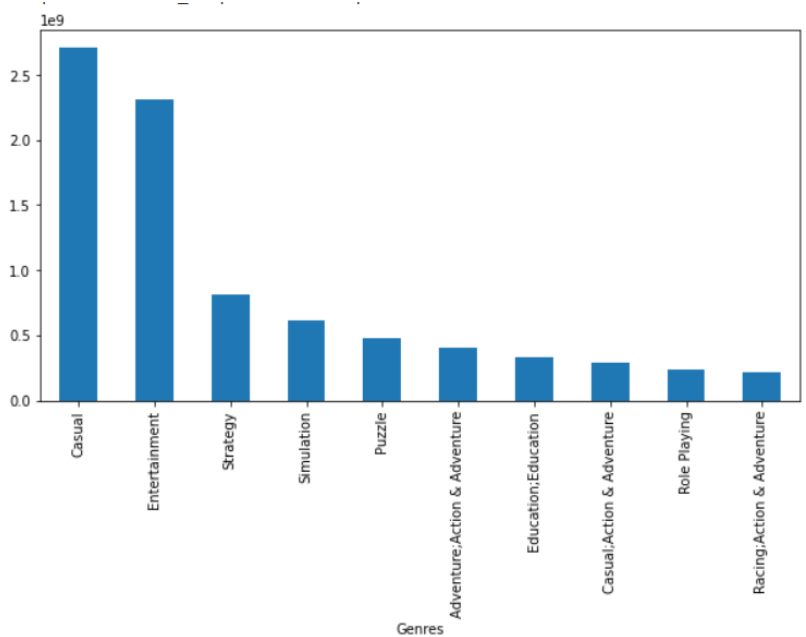
Stacked bar plot of rating bins for top categories according to installs



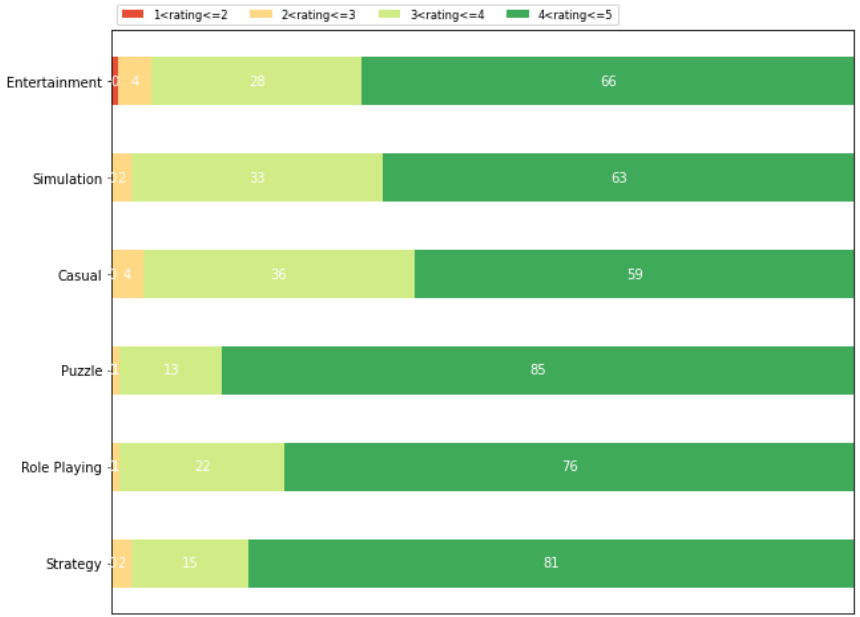
Number of Installs vs. category
(For Game Category)



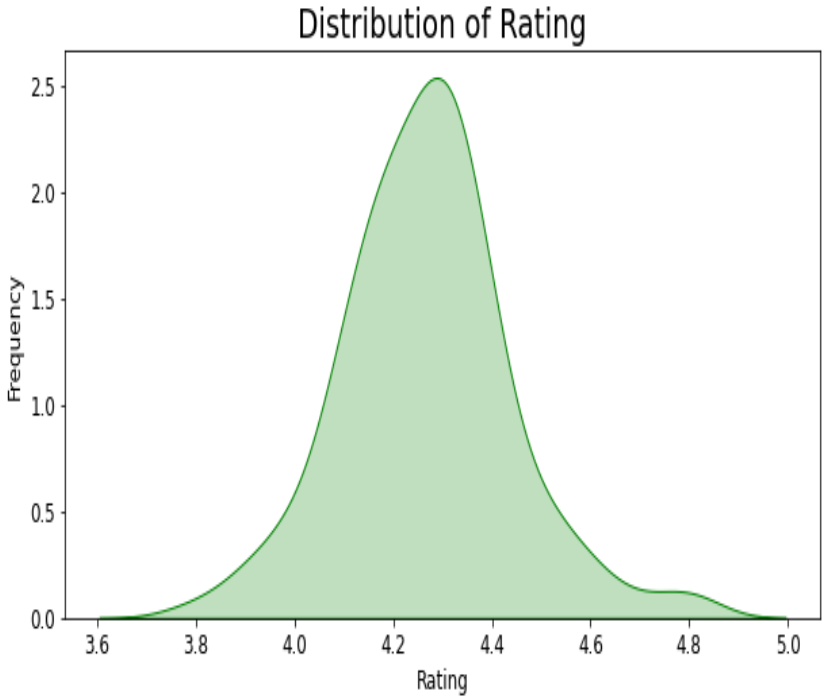
Stacked bar plot of rating bins for top genres according to installs
(For Game Category)

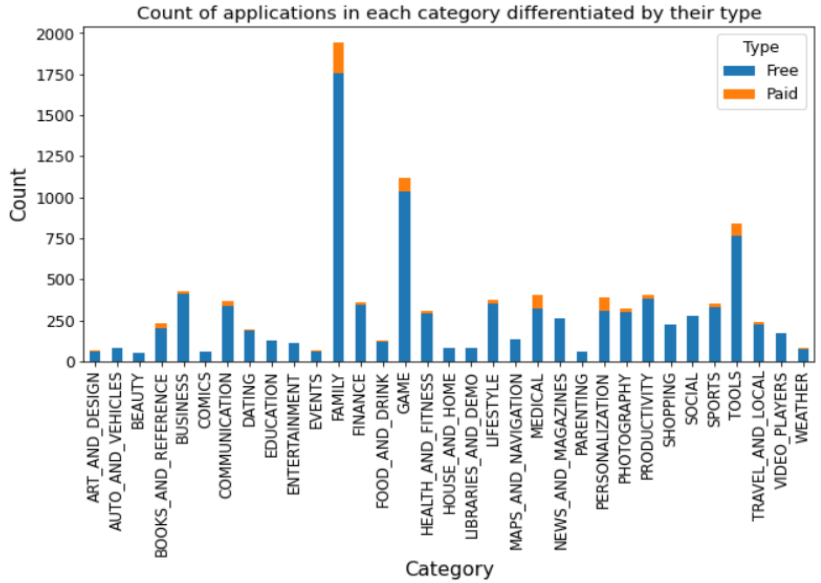


Number of Installs vs. category
(For Family Category)



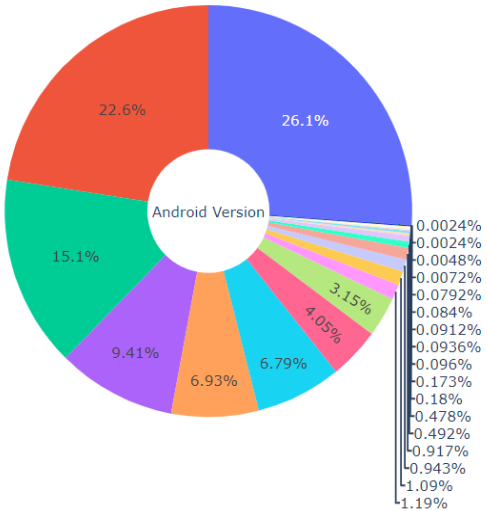
Stacked bar plot of rating bins for top genres according to installs
(For Family Category)

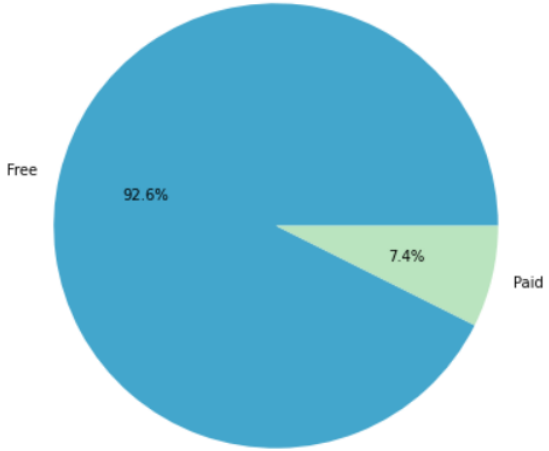

Distribution of Rating
Frequency
Rating
0.0
0.5
1.0
1.5
2.0
2.5
3.6
3.8
4.0
4.2
4.4
4.6
4.8
5.0


Count of applications in each category differentiated by their type
Type
Free
Paid
Count
Category
0
250
500
750
1000
1250
1500
1750
2000
ART_AND_DESIGN
AUTO_AND_VEHICLES
BEAUTY
BOOKS_AND_REFERENCE
BUSINESS
COMICS
COMMUNICATION
DATING
EDUCATION
ENTERTAINMENT
EVENTS
FAMILY
FINANCE
FOOD_AND_DRINK
GAME
HEALTH_AND_FITNESS
HOUSE_AND_HOME
LIBRARIES_AND_DEMO
LIFESTYLE
MAPS_AND_NAVIGATION
MEDICAL
NEWS_AND_MAGAZINES
PARENTING
PERSONALIZATION
PHOTOGRAPHY
PRODUCTIVITY
SHOPPING
SOCIAL
SPORTS
TOOLS
TRAVEL_AND_LOCAL
VIDEO_PLAYERS
WEATHER

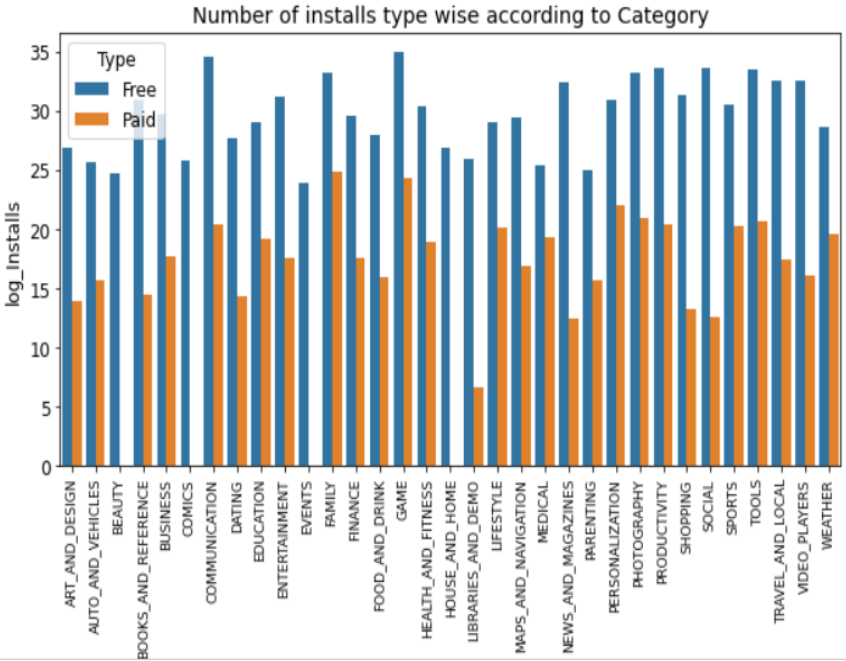
Data Visualisation

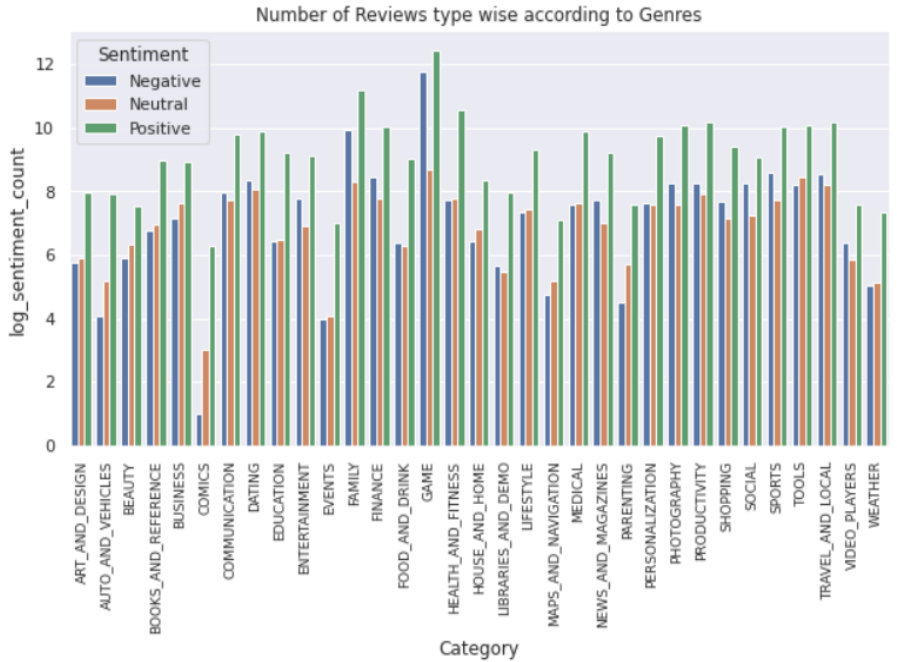
Different Android version

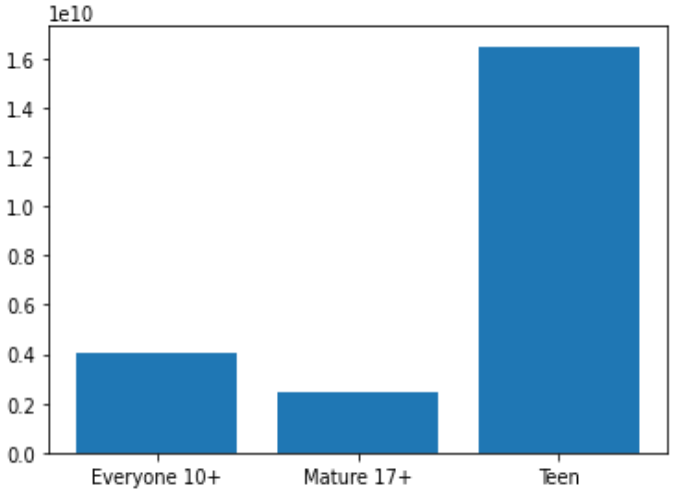


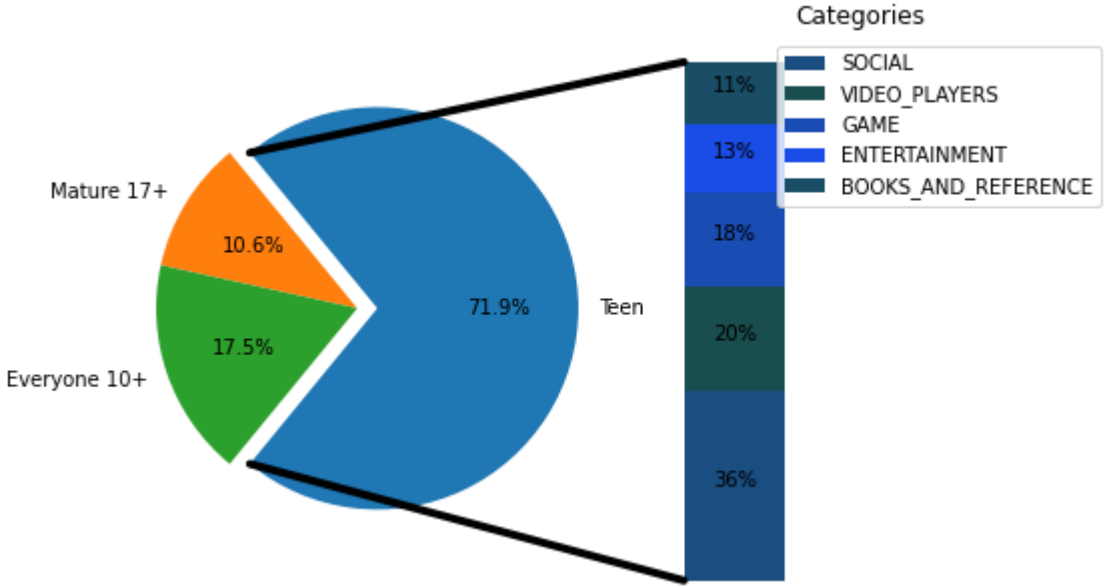


What is the percentage of paid to free apps?


Number of installs type wise according to Category
Type
Free
Paid
log_Installs
0
5
10
15
20
25
30
35
ART_AND_DESIGN
AUTO_AND_VEHICLES
BEAUTY
BOOKS_AND_REFERENCE
BUSINESS
COMICS
COMMUNICATION
DATING
EDUCATION
ENTERTAINMENT
EVENTS
FAMILY
FINANCE
FOOD_AND_DRINK
GAME
HEALTH_AND_FITNESS
HOUSE_AND_HOME
LIBRARIES_AND_DEMO
LIFESTYLE
MAPS_AND_NAVIGATION
MEDICAL
NEWS_AND_MAGAZINES
PARENTING
PERSONALIZATION
PHOTOGRAPHY
PRODUCTIVITY
SHOPPING
SOCIAL
SPORTS
TOOLS
TRAVEL_AND_LOCAL
VIDEO_PLAYERS
WEATHER

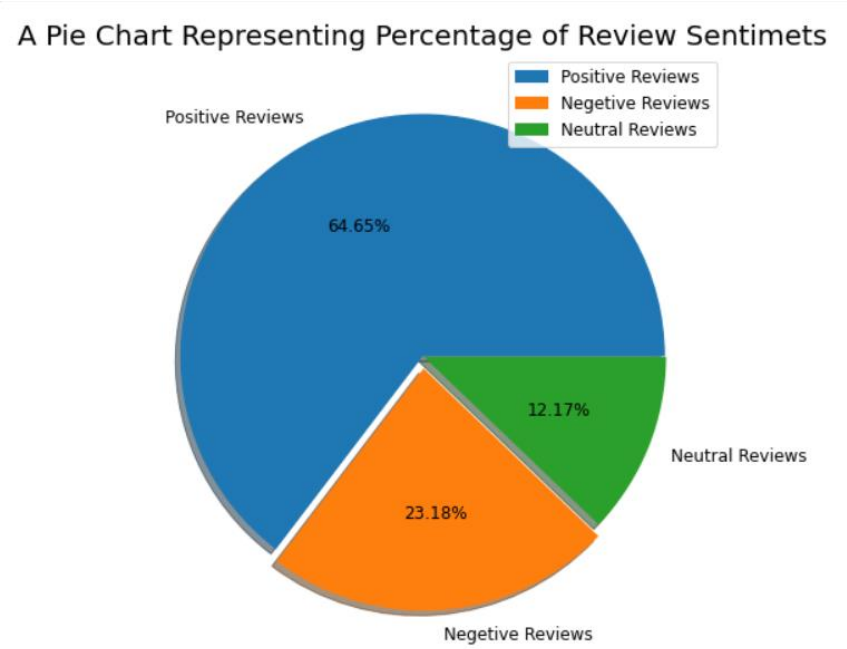

Number of Reviews type wise according to Genres
Sentiment
Negative
Neutral
Positive
log_sentiment_count
0
2
4
6
8
10
12
ART_AND_DESIGN
AUTO_AND_VEHICLES
BEAUTY
BOOKS_AND_REFERENCE
BUSINESS
COMICS
COMMUNICATION
DATING
EDUCATION
ENTERTAINMENT
EVENTS
FAMILY
FINANCE
FOOD_AND_DRINK
GAME
HEALTH_AND_FITNESS
HOUSE_AND_HOME
LIBRARIES_AND_DEMO
LIFESTYLE
MAPS_AND_NAVIGATION
MEDICAL
NEWS_AND_MAGAZINES
PARENTING
PERSONALIZATION
PHOTOGRAPHY
PRODUCTIVITY
SHOPPING
SOCIAL
SPORTS
TOOLS
TRAVEL_AND_LOCAL
VIDEO_PLAYERS
WEATHER
Category

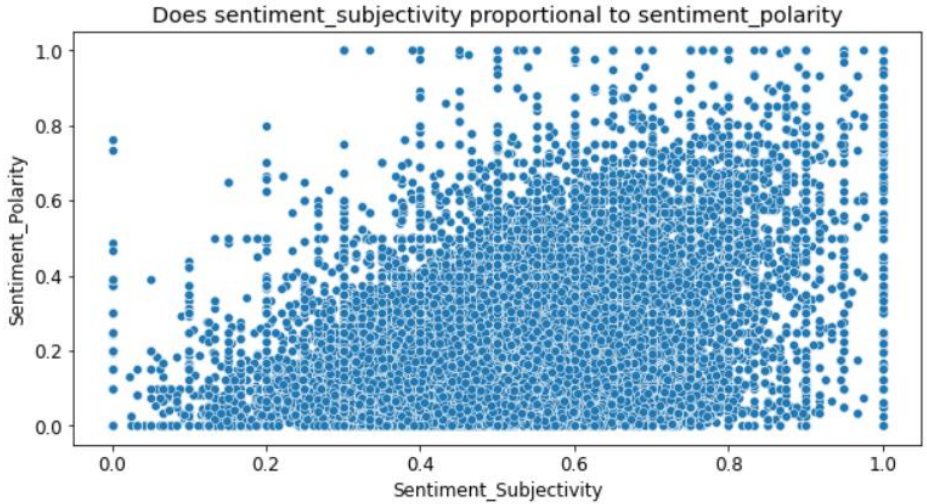


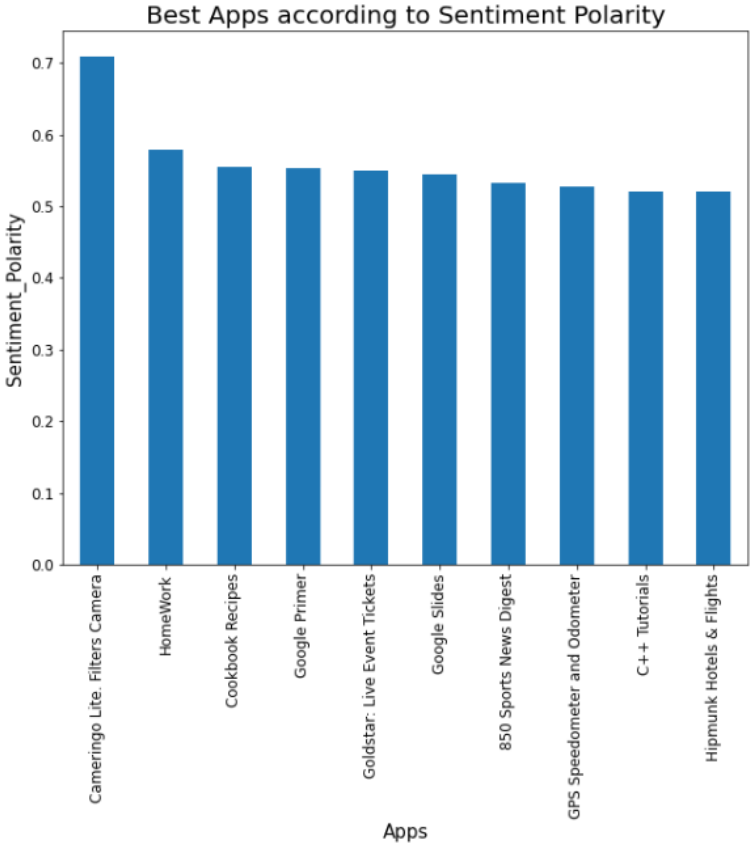


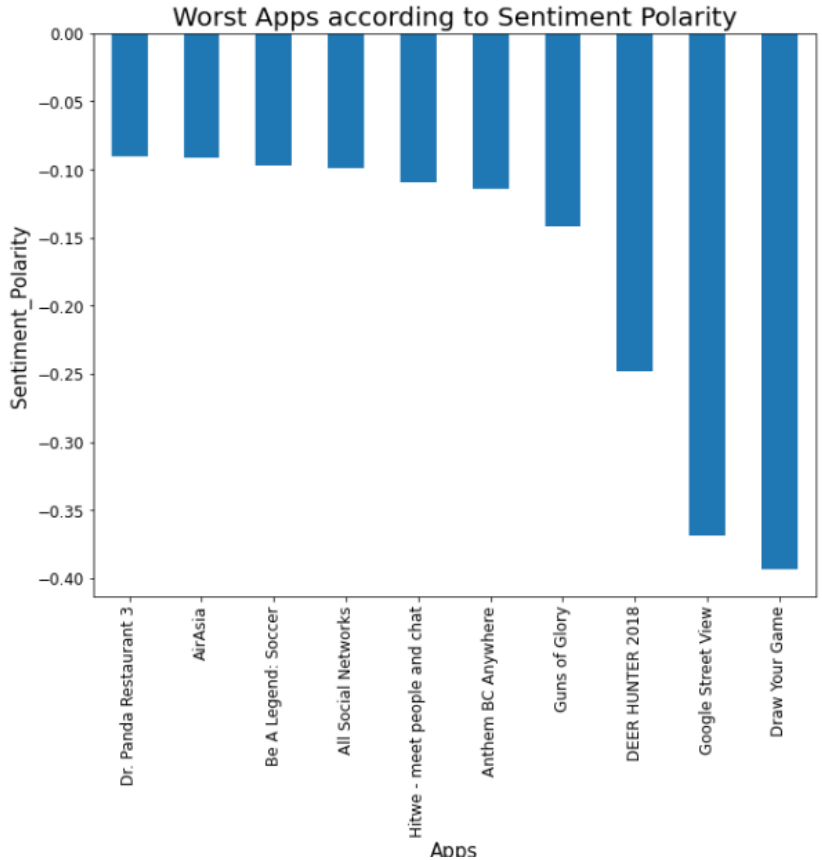
Pie chart of content rating type proportions with popular categories for teen type

Review Analysis






Best Apps according to Sentiment Polarity
Sentiment_Polarity
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Cameringo Lite. Filters Camera
HomeWork
Cookbook Recipes
Google Primer
Goldstar: Live Event Tickets
Google Slides
850 Sports News Digest
GPS Speedometer and Odometer
C++ Tutorials
Hipmunk Hotels & Flights
Apps


Worst Apps according to Sentiment Polarity
Sentiment_Polarity
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
−0.30
−0.35
−0.40
Dr. Panda Restaurant 3
AirAsia
Be A Legend: Soccer
All Social Networks
Hitwe - meet people and chat
Anthem BC Anywhere
Guns of Glory
DEER HUNTER 2018
Google Street View
Draw Your Game
Apps

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***

Challenges Faced

- Certain variables were having values that are in not ready to use the form, like size and Installs
- Limited data for paid-type apps
- Presence of Null values in the dataset, especially the Rating variable in the Google Play Store dataset and user review dataset

THANK YOU