Capstone Project 1

Play Store App Review Analysis

Presented By

- Divya P. Kedia
- Sakshi R. Ghugare

Content

- Introduction
- Objective
- Problem Statement
- Description of Data
- Cleaning the Data
- Data analysis and visualizations
- Conclusion

Problem statement

1. What are the top 10 apps available per category in Play store?
2. How many no of installs are there per category? What are the top two categories are observed?
3. What are the top 30 Genres for free installation?
4. What are the top 30 Genres for paid installation?
5. What are the top apps in play store based on review?
6. What are the mean rating for all categories?
7. What is the Rating distribution for apps?
8. What are the Types of Content rating for the apps?
9. Whether or not does the size of app matters?

Description of Dataset

There are two dataset :

1)Play store data 2) User review data

Play store data :-

App	Name of the application
Category	Category of the application
Rating	Rating given to the application
Review	Numbers of reviews given to the application
Size	Size of the application
Installs	Numbers of downloads of the application
Type	Free or Paid
Price	Price of the application if it is paid
Content rating	It is Age appropriate or not
Genres	Type of Genre the application belongs to
Last Updated	When the last time the application is updated
Current version	Current version of the application
Android version	Min android version required to run the application

Description of Dataset

2. User Review Data:

- **App:** An name of the app.

- **Translated Review :**Reviews being given by consumers.

- **Sentiment** : Sentiment given to an app by users (i.e. Positive, Neutral,Negative).

•**Sentiment Polarity:** The polarity of sentiment measures how negative or positive.

•**The context :** In the data we have, the polarity ranges from +1(Positive) to -1(Negative).

•**Sentiment Subjectivity:**
The subjectivity of a sentiment is how likely that sentiment is to be based on data or factual information, versus personal opinions or public opinions

Data Cleaning

When using data, most people agree that your insights and analysis are only as good as the data you are using. Essentially, garbage data in is garbage analysis out.

What is Data Cleaning?

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.
Data cleaning, also referred to as data cleansing and data scrubbing, is one of the most important steps for your organization if you want to create a culture around quality data decision-making.

Steps for data cleaning

Step 1: Remove duplicate or irrelevant observations
Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. Duplicate observations will happen most often during data collection

Step 2: Fix structural errors
Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization.

Step 3: Filter unwanted outliers
this is use for removing incorrect and unwanted outlier

Step 4: Handle missing data
Fixing mislabeled categories or classes, Types , Strange name conventions
Replacing missing values with mean, median or mode: Replacing missing values

Exploratory Analysis and Visualization

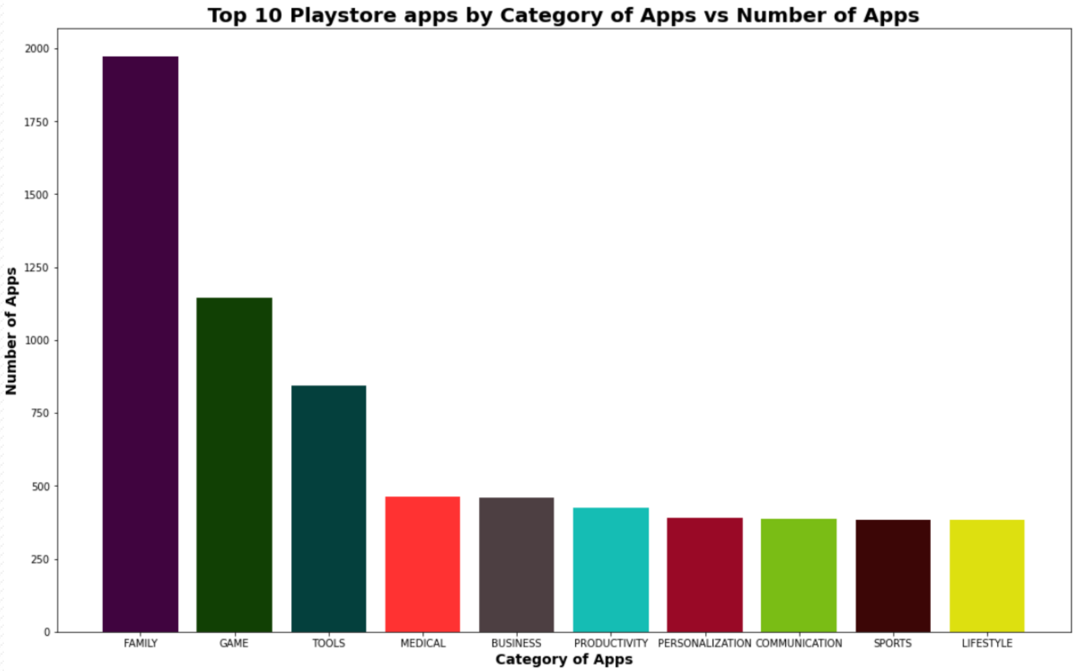
In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Data visualization is the graphic representation of data. It involves producing images that communicate relationships among the represented data to viewers of the images.

Top 10 play store apps by category of apps Vs number of apps

As we see here, the Play store having more number of applications in the genres like Tools, Entertainment, Education and etc.

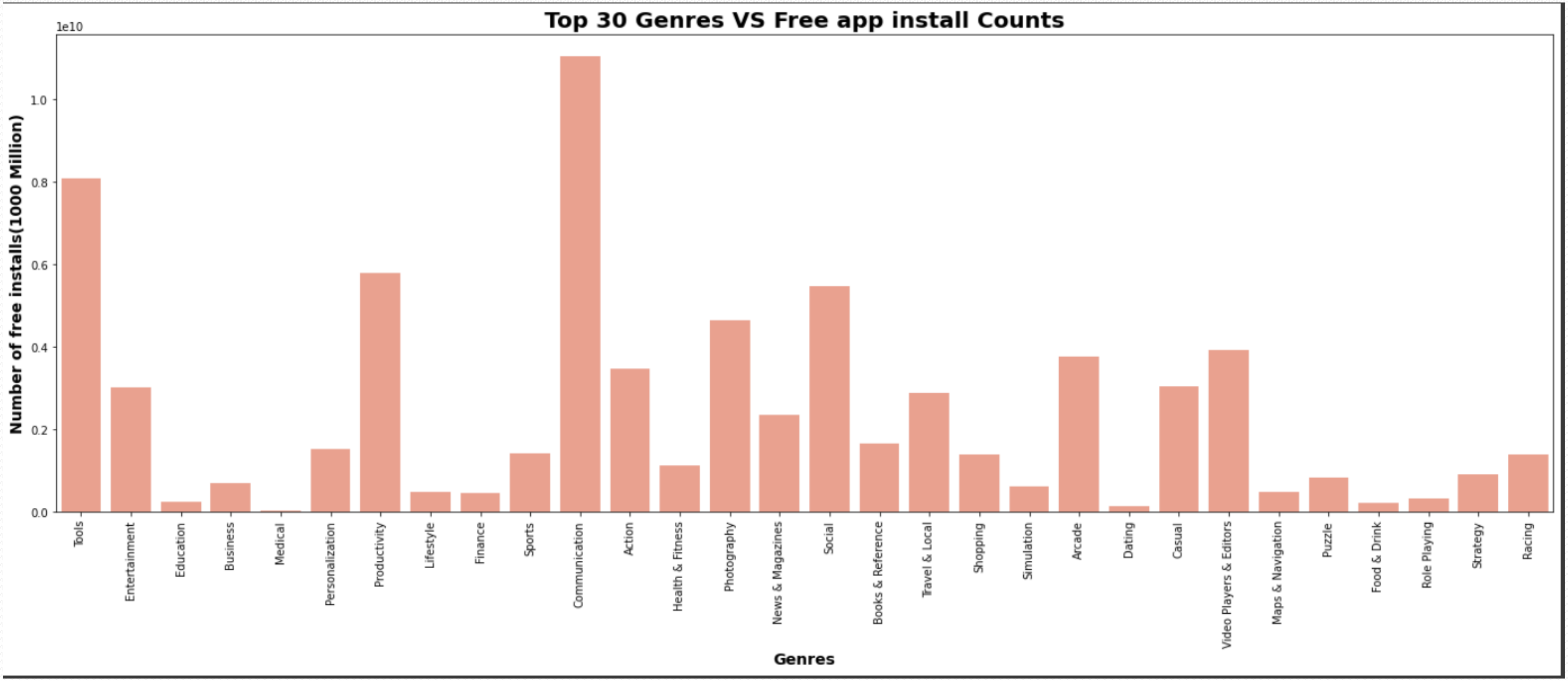
The developers are mostly focusing on these genres because of the people's daily basis requirements.

Genres like Educational, Parenting, Music are having comparatively less



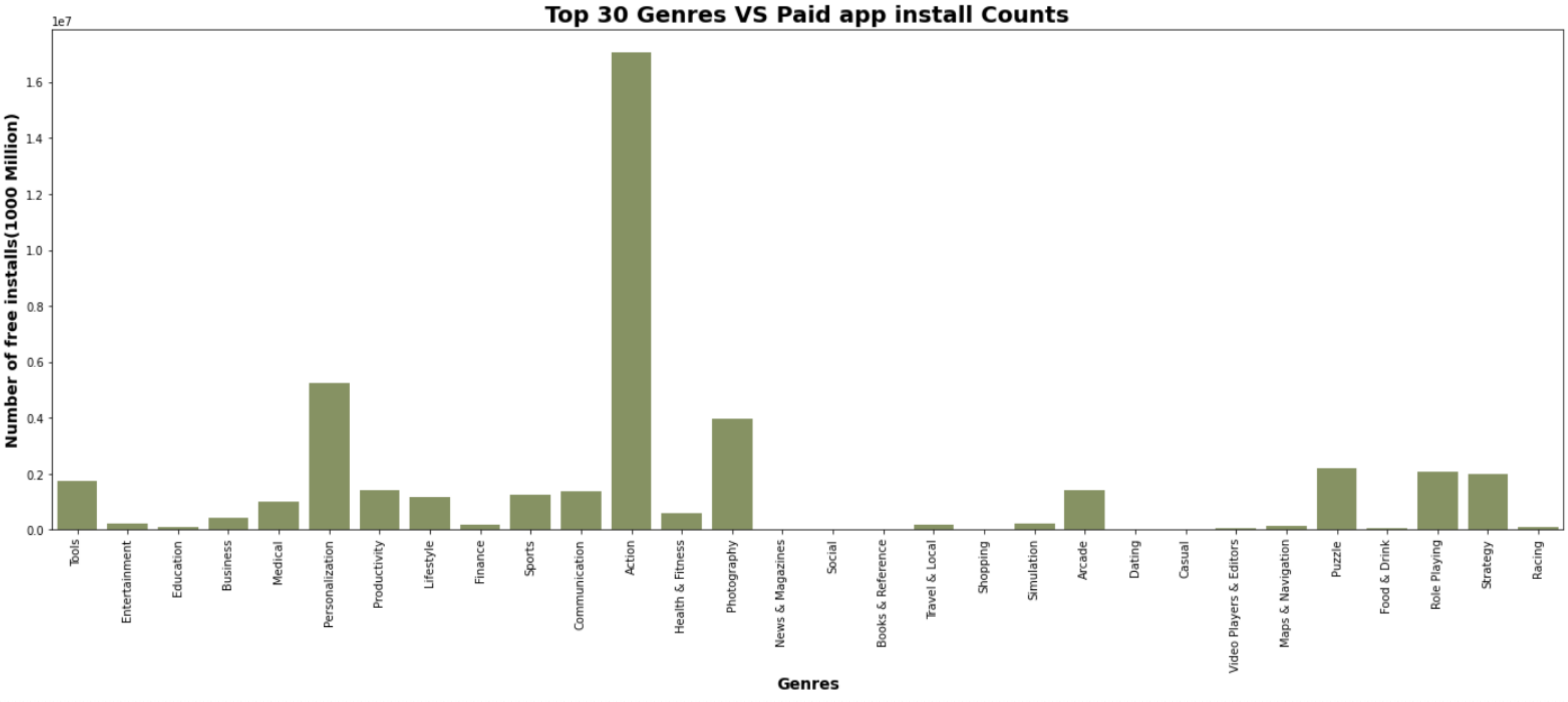
Top 30 genres Vs free app install counts

When comparing the both plots, people are showing more interest on free apps like communication, Tools etc
Again the Educational, Parenting and Music are the genres in the least Top free apps
When it’s coming, commercial people are preferring apps like Games, Photography, and Personalization.
People are preferring less on Educational, Event, and Art & Design.



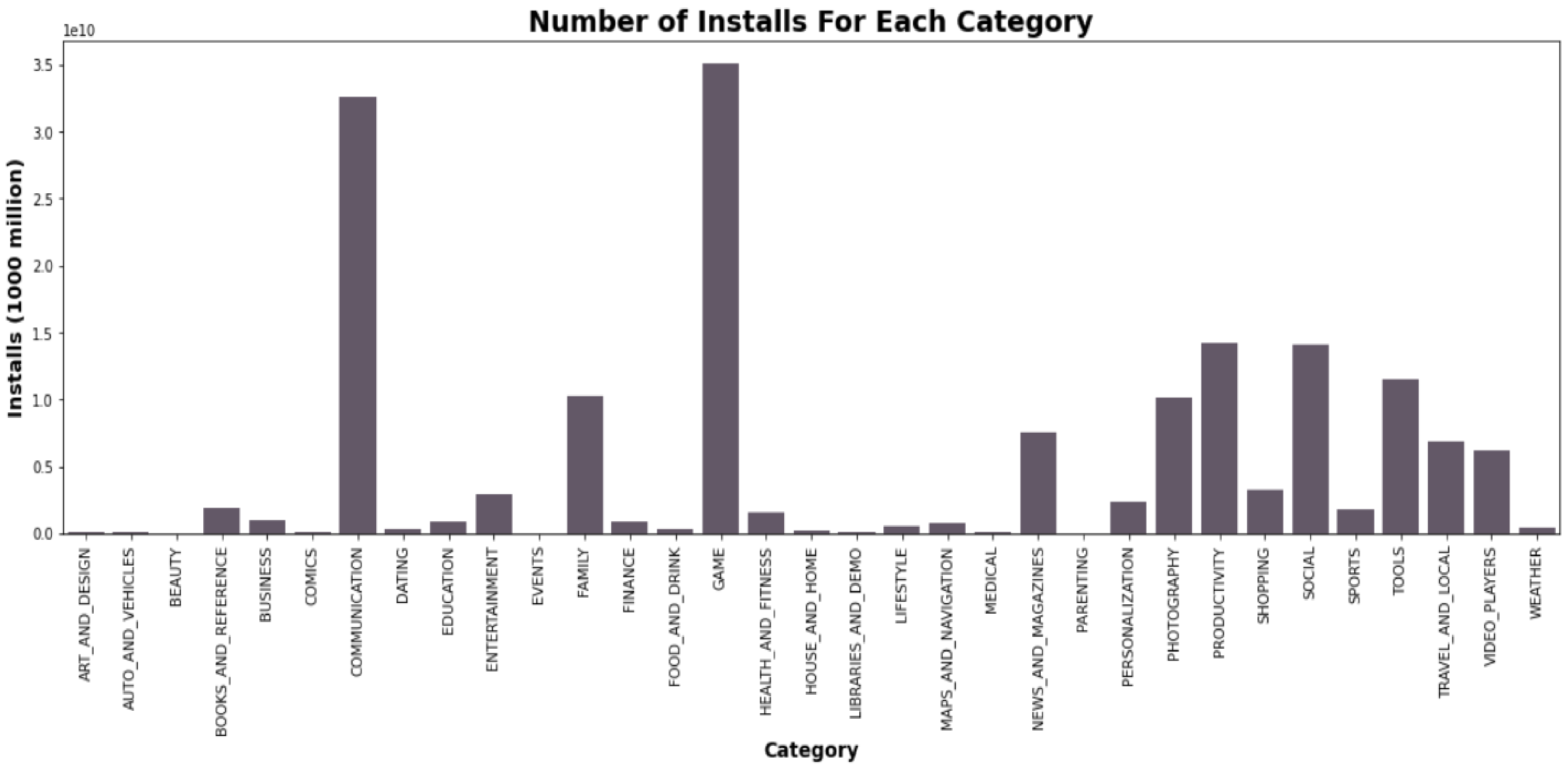
Top 30 genres Vs paid app install count

When it’s coming, commercial people are preferring apps like Games, Photography, and Personalization.
People are preferring less on Educational, Event, and Art & Design.



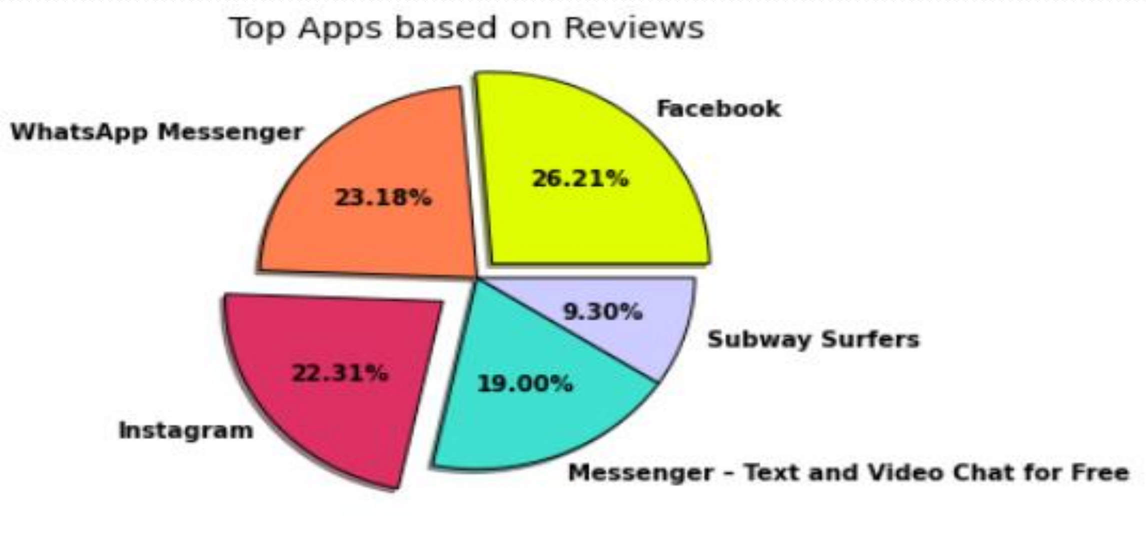
Number of installs for each category

From this Bar it is observed that "Game" and "Communication" Category has the most number of Installs.



Top apps based on reviews

As seen from the above Pie Chart "Facebook" and "Whatsapp“ have the most number of Reviews with 26.21% and 23.18% respectively



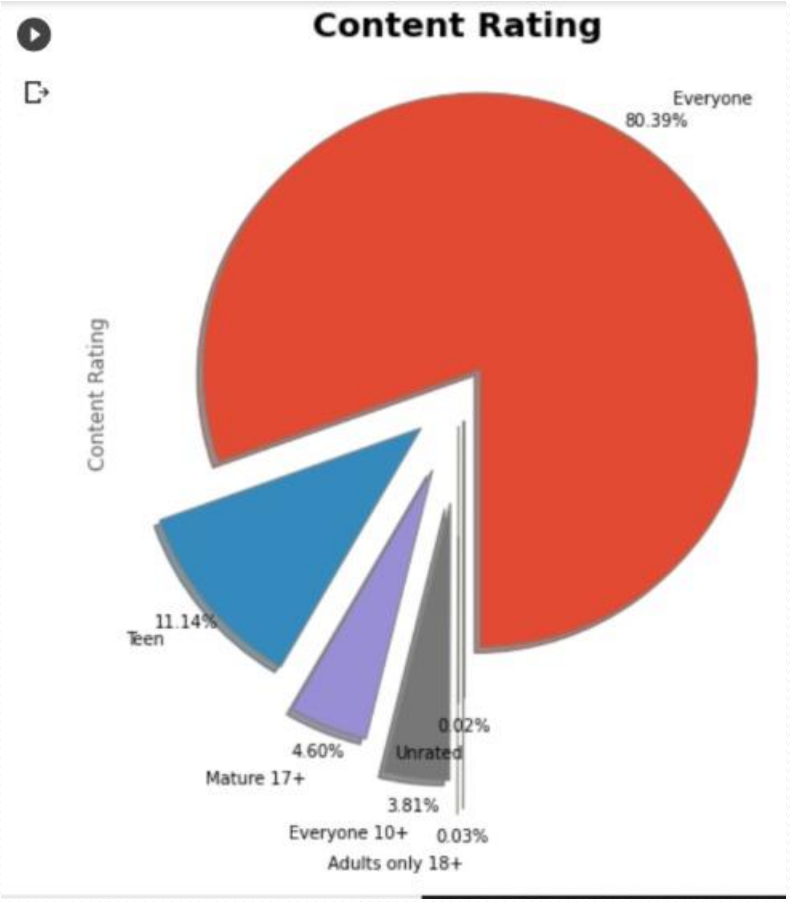
Content Rating

The content rating shows the results for general contents as high.

The content rating type ‘Everyone’ has the most percentage value of 80.39%.

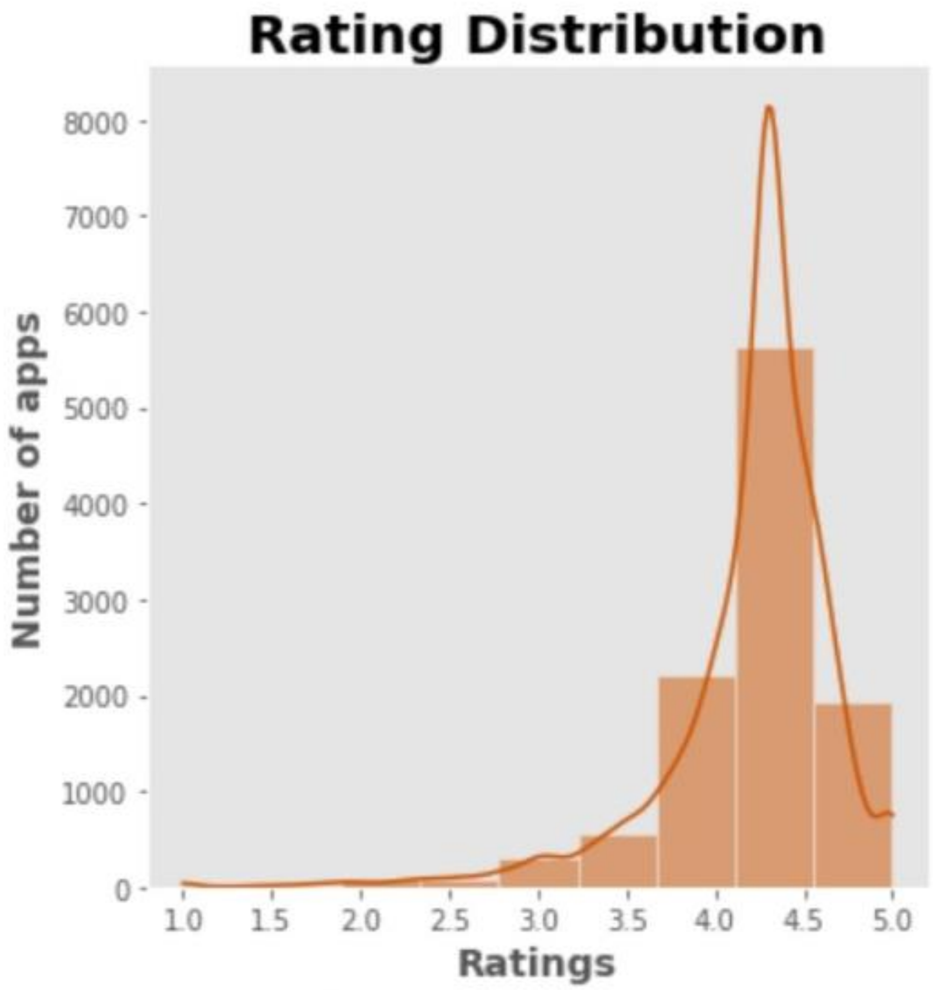
‘Teen’ contents are second in the order with the percentage of 11.14%.

Adult’s only and unrated contents are least in this plot, 0.03% and 0.02% respectively.



Rating distribution

From this distribution plotting, it is observed that most of the apps in the Play Store have the rating between 4 to 4.5 out of 5 which shows that most of the apps on play store are liked by the users.



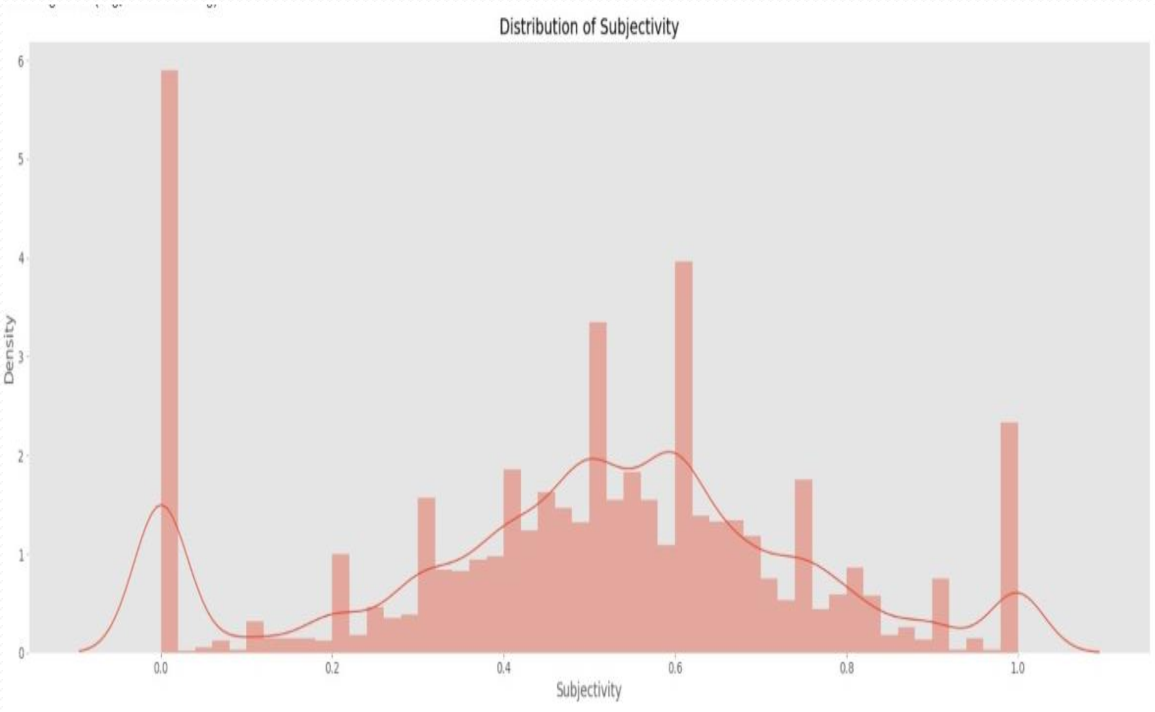
Distribution of Subjectivity

Subjectivity lies mostly between 0.5 and 0.65.

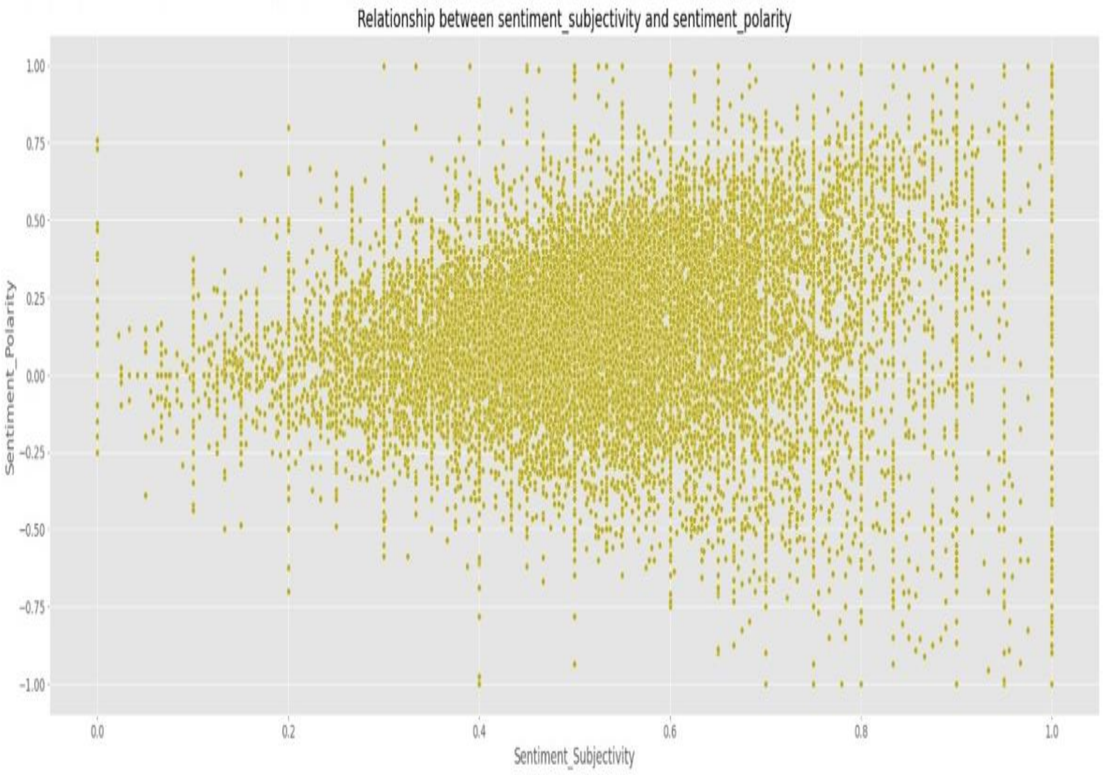
It shows that the average content and apps reviews subjectivity are mostly relevant.

Subjectivity of 100% has slightly occurred frequently.

The nearly 0 subjectivity has a considerable amount of frequency.



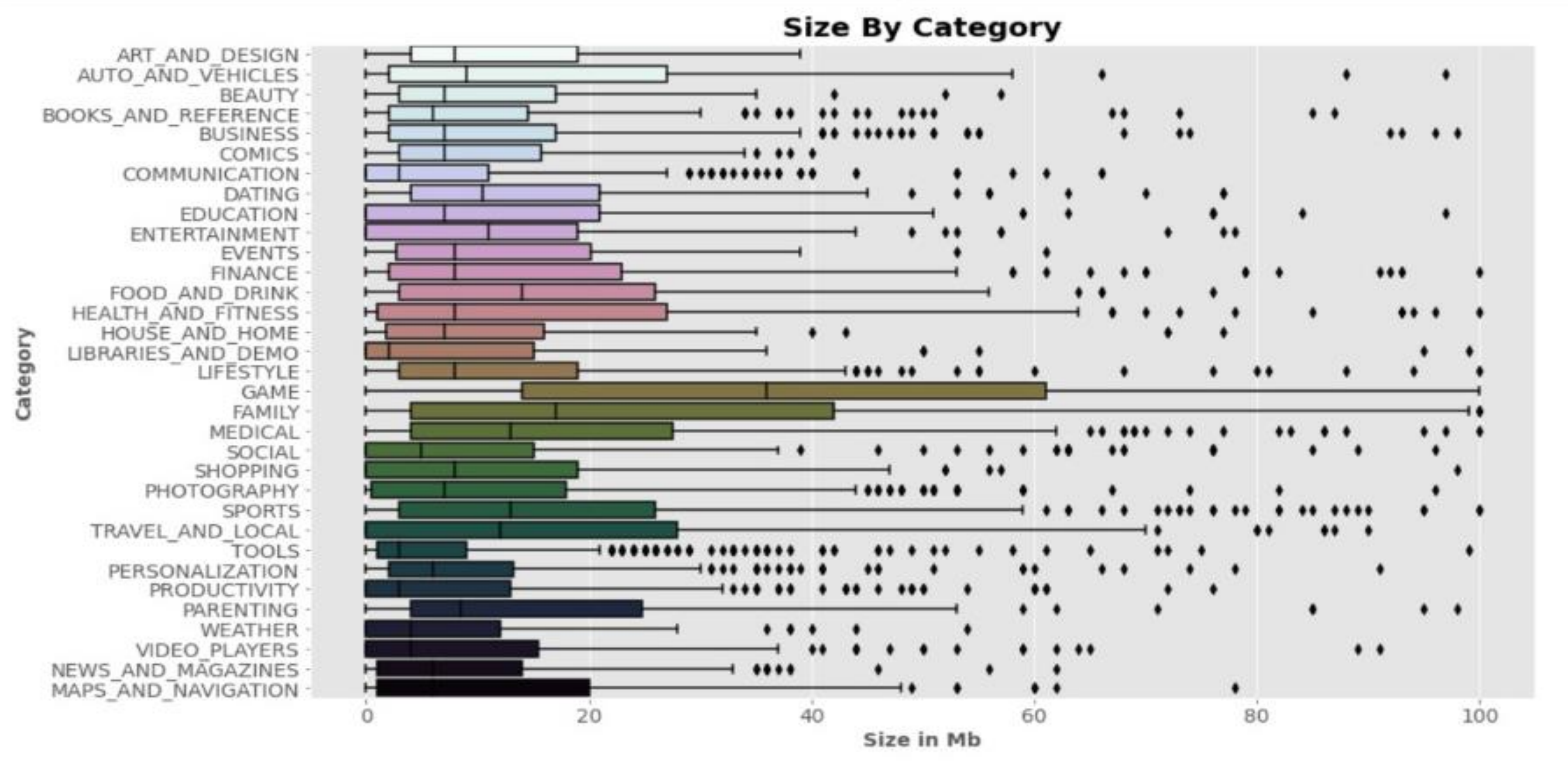
Relationship between sentiment subjectivity and sentiment polarity



From the above scatter plot it can be observed that sentiment subjectivity is not always proportional to sentiment polarity but in maximum number of case, shows a proportional behavior, when variance is too high or low

Size of the apps

Majority of the apps installed are no more than 100 megabytes in size. This shows that size does matter. Most consumers don't want to download apps that takes up too much space on their device storage.



CONCLUSION

Thus the app development companies could decide what application should be developed and they can also see the prediction of their developed application. In this they also get to see the categorized reviews of all the application in one interface which will help them decide which app is liked by the users and which apps need to be developed more. The dataset contains immense possibilities to improve business values and have a positive impact. It is not limited to the problem taken into consideration for this project**.**

Challenges

- Huge chunk of data was to be handled keeping in mind not to miss anything which is even of the little relevance.

- Computation time