

Capstone Project - 1

Play Store App Review Analysis Exploratory Data Analysis

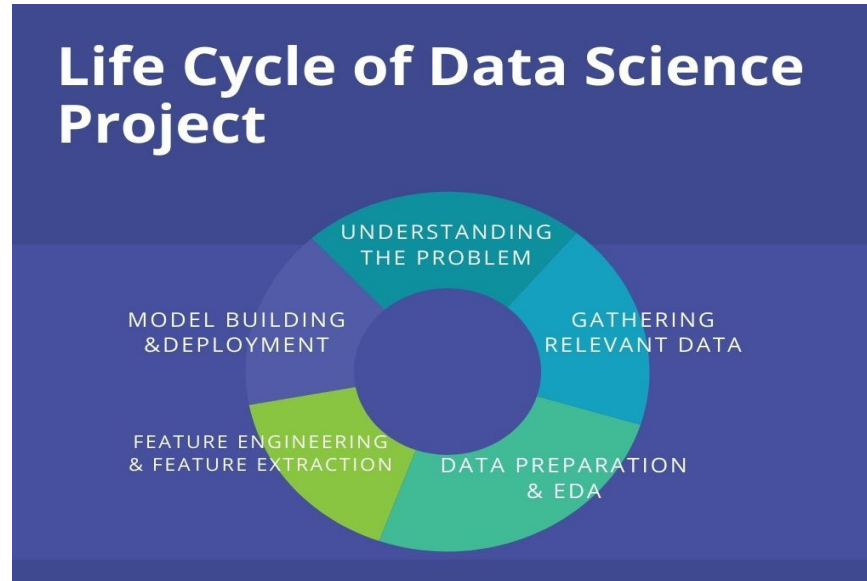
Individual Project:
Avisikta Majumdar

Table Of Contents

- **Understanding EDA**
- **Understanding the dataset(play-store)**
 - Descriptive statistics
 - Data cleaning
 - Data Visualization
- **Understanding the dataset(user_reviews)**
 - Descriptive statistics
 - Data cleaning
 - Data Visualization
- **DataFrame Merging(play-store & review)**

● *What does Exploratory Data Analysis (EDA) means ??*

Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset. EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better.



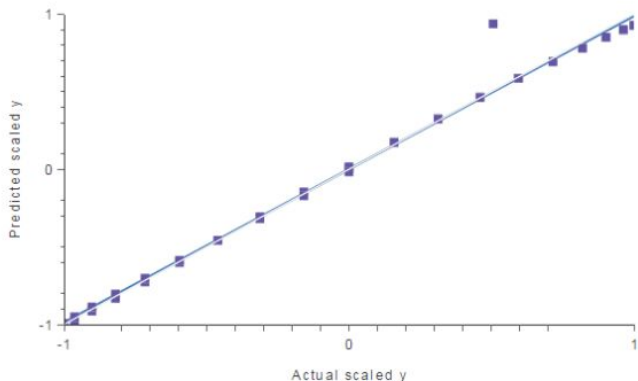
Problem Statement



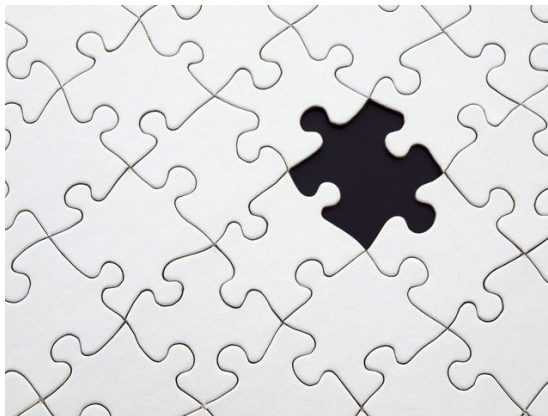
The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market. Explore and analyze the data to discover key factors responsible for app engagement and success.

Processes under EDA

- *Identify categorical & numerical column and change the datatype*
(if any numerical data is in string format then we have to convert it)
- *Handling missing values*
 - Using dropna method
 - Using mean , median , mode
- *Handling outliers*
 - Using IQR
 - Boxplot
 - Multivariate Method



Multivariate Method



Missing Values



Outliers

Libraries Used

- **NumPy** (Numerical Python)
- **Pandas** (data reading & data cleaning)
- **Matplotlib**(data visualization)
- **Seaborn** (data visualization)
- **Word Cloud** (graphical representation of words)

Data Summary

(Play Store Data)

- **App:-** Name of the application.
- **Category :-** This column will tell us in which type of application is this.
- **Rating :-** Rating of that application
- **Reviews :-** No of reviews present for this application
- **Size :-** Size of this app (Kb & Mb both present)
- **Installs :-** NO of times this app got installed
- **Type :-** Type of the application
- **Price :-** Price of that application
- **Content Rating :-** Rating of the content
- **Genres :-** Tells us which genres this application is belongs to
- **Last Updated :-** when this application got update last time
- **Current Ver :-** Current version of that application.
- **Android Ver :-** Android version of that application.

Basic Data Exploration

(Play Store Data)

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
#   Column          Non-Null Count  Dtype
---  -
0   App              10841 non-null  object
1   Category         10841 non-null  object
2   Rating           9367 non-null   float64
3   Reviews          10841 non-null  object
4   Size             10841 non-null  object
5   Installs         10841 non-null  object
6   Type             10840 non-null  object
7   Price            10841 non-null  object
8   Content Rating   10840 non-null  object
9   Genres           10841 non-null  object
10  Last Updated     10841 non-null  object
11  Current Ver      10833 non-null  object
12  Android Ver      10838 non-null  object
dtypes: float64(1), object(12)
memory usage: 8.3 MB
```

- Shape :- (10841, 13)

Columns

Numerical

- Rating
- Reviews
- Price
- Size
- Installs
- Curr ver
- Android Ver

Categorical

- App
- Category
- Type
- Content Rating
- Genres
- Last Updated

Data Cleaning

(Play Store Data)

Checking NaN values

	No Of Total Values	No of NaN values	%age of NaN values
Rating	10841	1474	13.60
Current Ver	10841	8	0.07
Android Ver	10841	3	0.03
Type	10841	1	0.01
Content Rating	10841	1	0.01
App	10841	0	0.00
Category	10841	0	0.00
Reviews	10841	0	0.00
Size	10841	0	0.00
Installs	10841	0	0.00
Price	10841	0	0.00
Genres	10841	0	0.00
Last Updated	10841	0	0.00




	No Of Total Values	No of NaN values	%age of NaN values
App	8886	0	0.0
Category	8886	0	0.0
Rating	8886	0	0.0
Reviews	8886	0	0.0
Size	8886	0	0.0
Installs	8886	0	0.0
Type	8886	0	0.0
Price	8886	0	0.0
Content_Rating	8886	0	0.0
Genres	8886	0	0.0
Last_Updated	8886	0	0.0
Current_Ver	8886	0	0.0
Android_Ver	8886	0	0.0
Log_installs	8886	0	0.0
Gaming Category App	8886	0	0.0

Data Cleaning

- Removing dollar(\$) sign from **Price** column


Price	
233	0
234	\$4.99
235	\$4.99



Price	
233	0.00
234	4.99
235	4.99

- Removing plus(+) sign from **Installs** columns


Installs	
0	10,000+
1	500,000+
2	5,000,000+



Installs	
0	10000
1	500000
2	5000000

- Converting **Installs** value by using *Log transformation*

```
Min. install value :-0
Max. installs value:- 1000000000
```



```
Min. install value :- -inf
Max. installs value:- 30.0
```

Data Cleaning

	Size
Varies with device	1695
11M	198
12M	196
14M	194
13M	191
...	...
1,000+	1
892k	1
387k	1
458k	1
67k	1

462 rows × 1 columns



As you can see Size in in MB & KB both format we have to convert it into MB & Will also replace "Varies with devices" with NaN value

	Size
11.000	198
12.000	196
14.000	194
13.000	191
15.000	184
...	...
0.027	1
0.414	1
0.647	1
0.039	1
0.942	1

460 rows × 1 columns

TOP 5 Genres & Categories

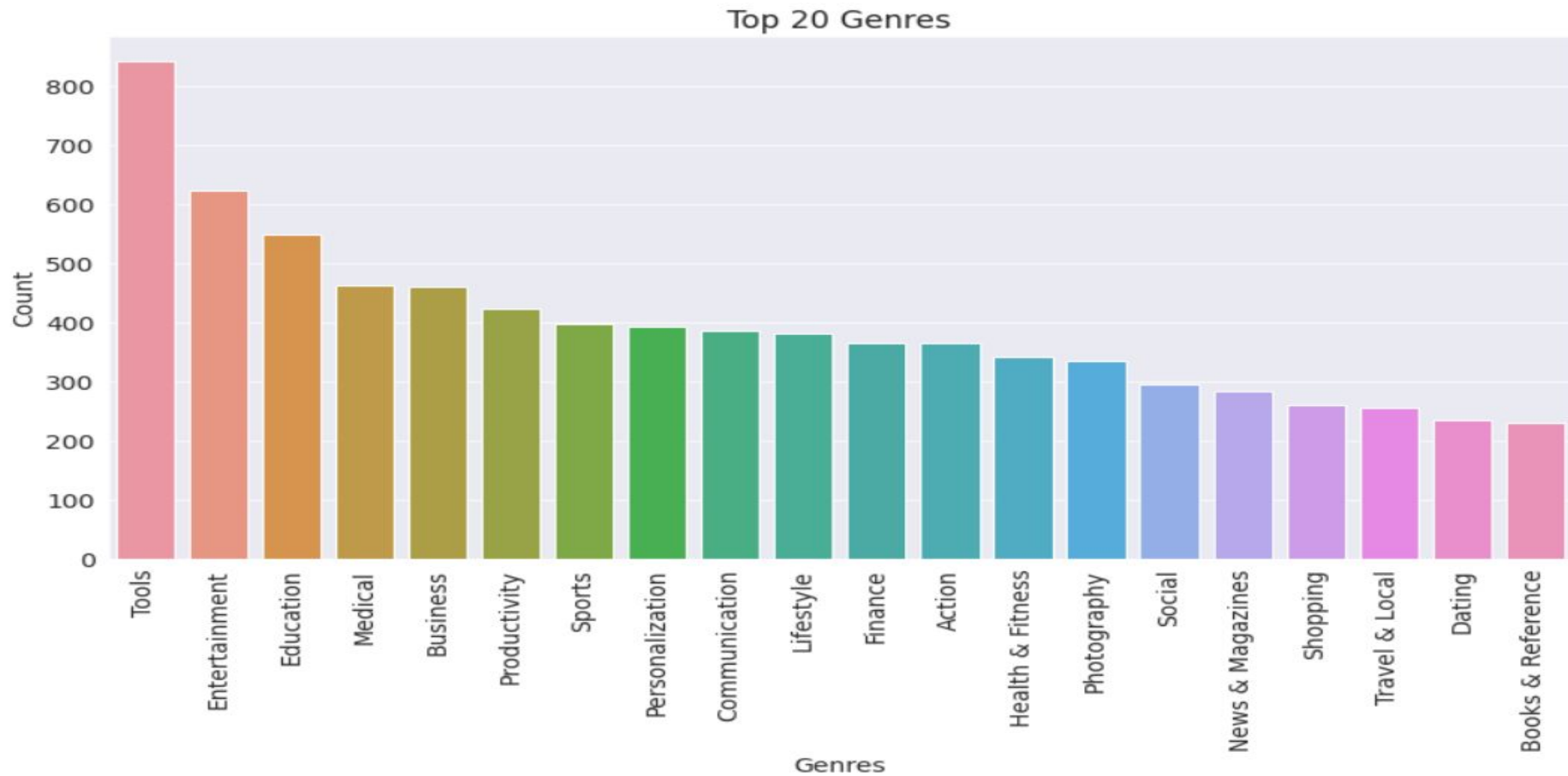
TOP 5 Genres

	Genres	Count
0	Tools	842
1	Entertainment	623
2	Education	549
3	Medical	463
4	Business	460

Top 5 category

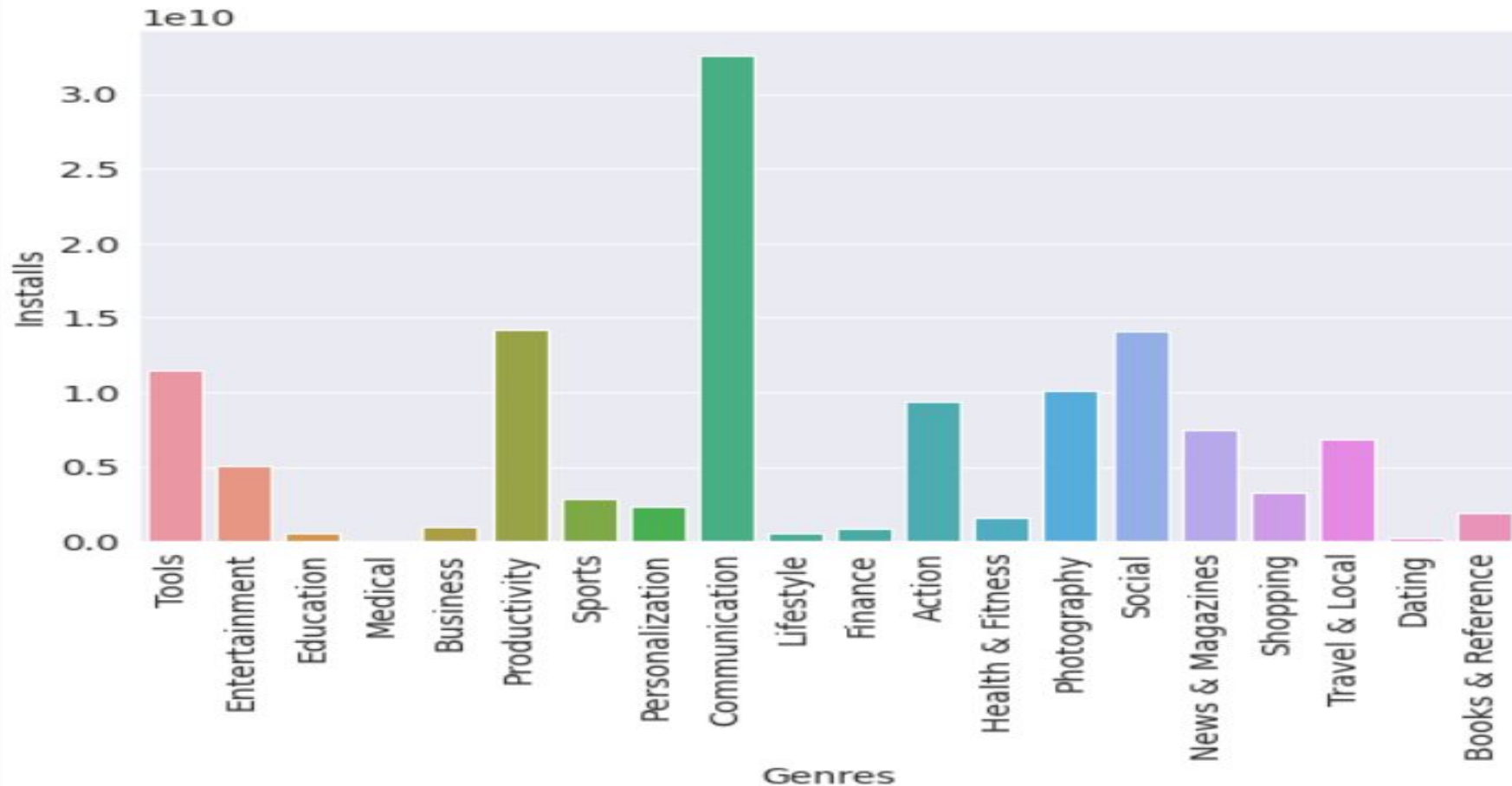
	Category	Count
0	FAMILY	1972
1	GAME	1144
2	TOOLS	843
3	MEDICAL	463
4	BUSINESS	460

Genres

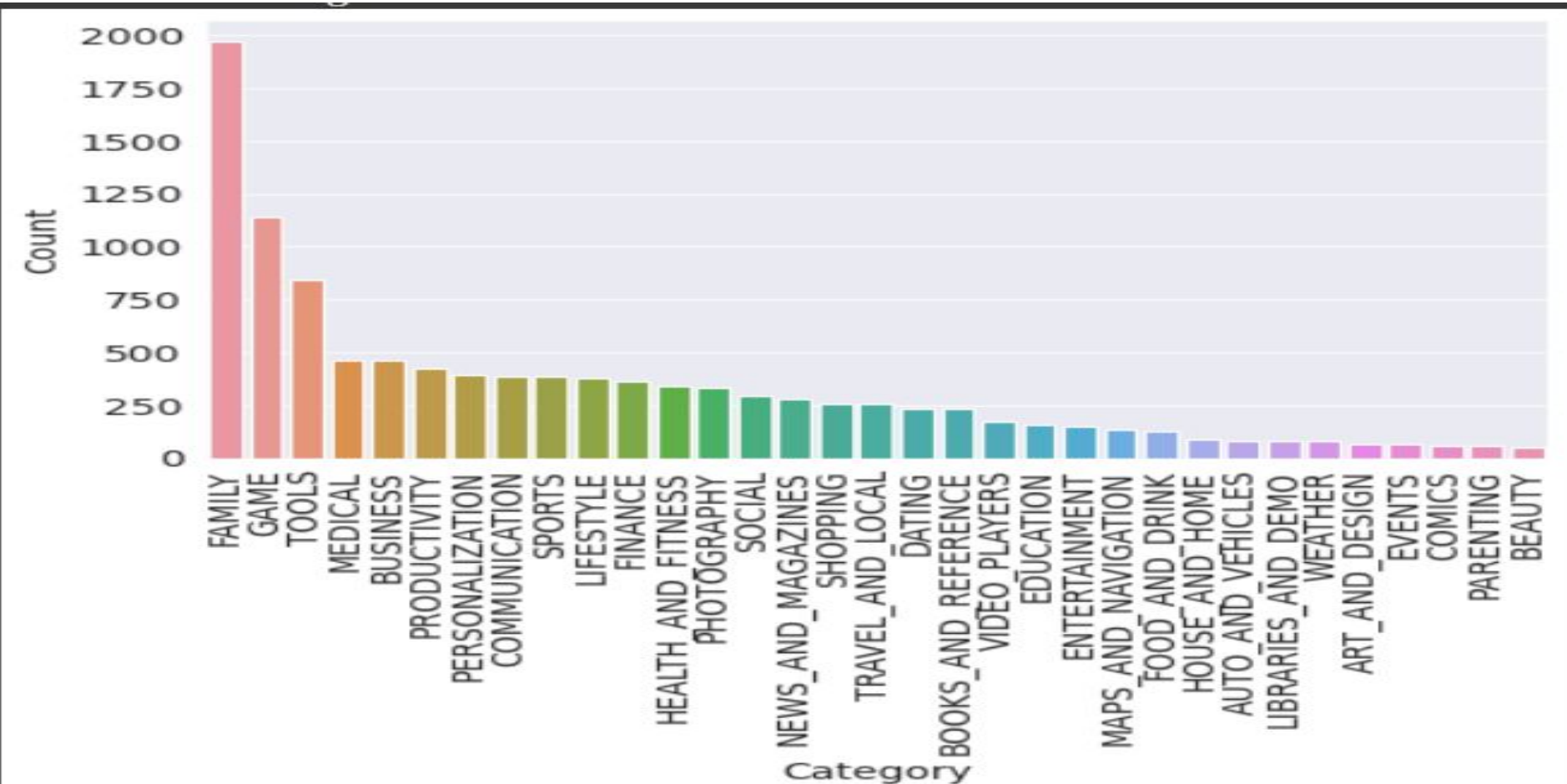


Installs according to Genres

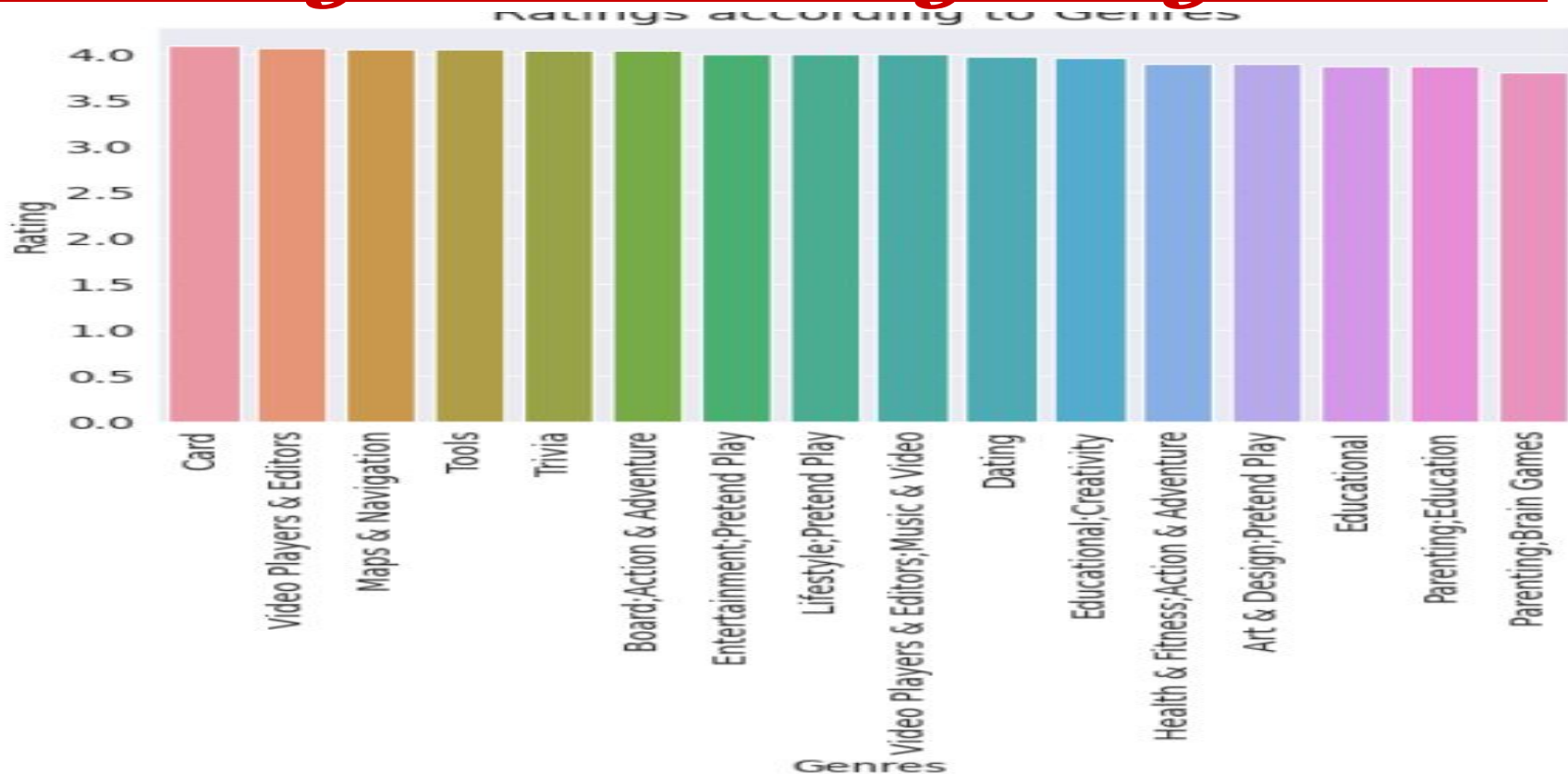
AI



Count Of Applications For Each Category AI

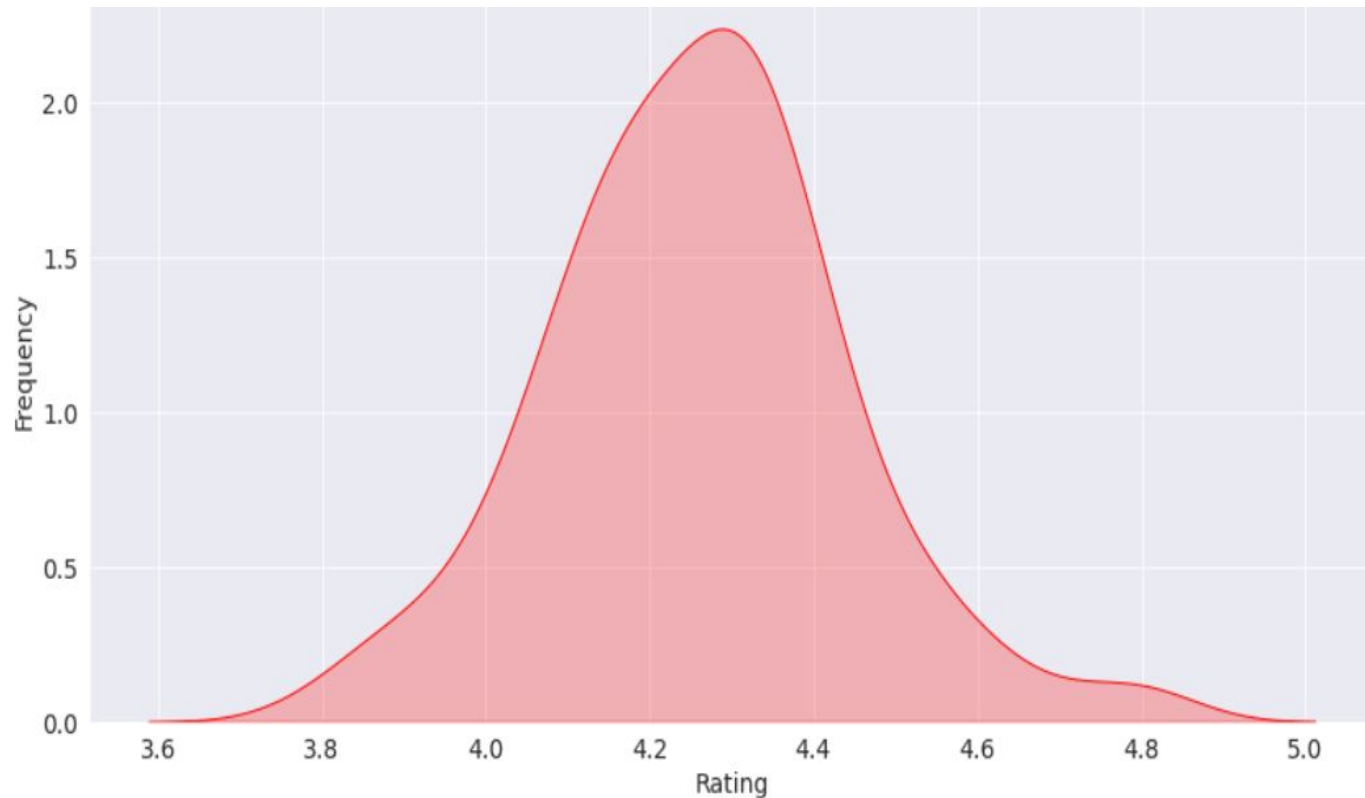


Rating according to genres



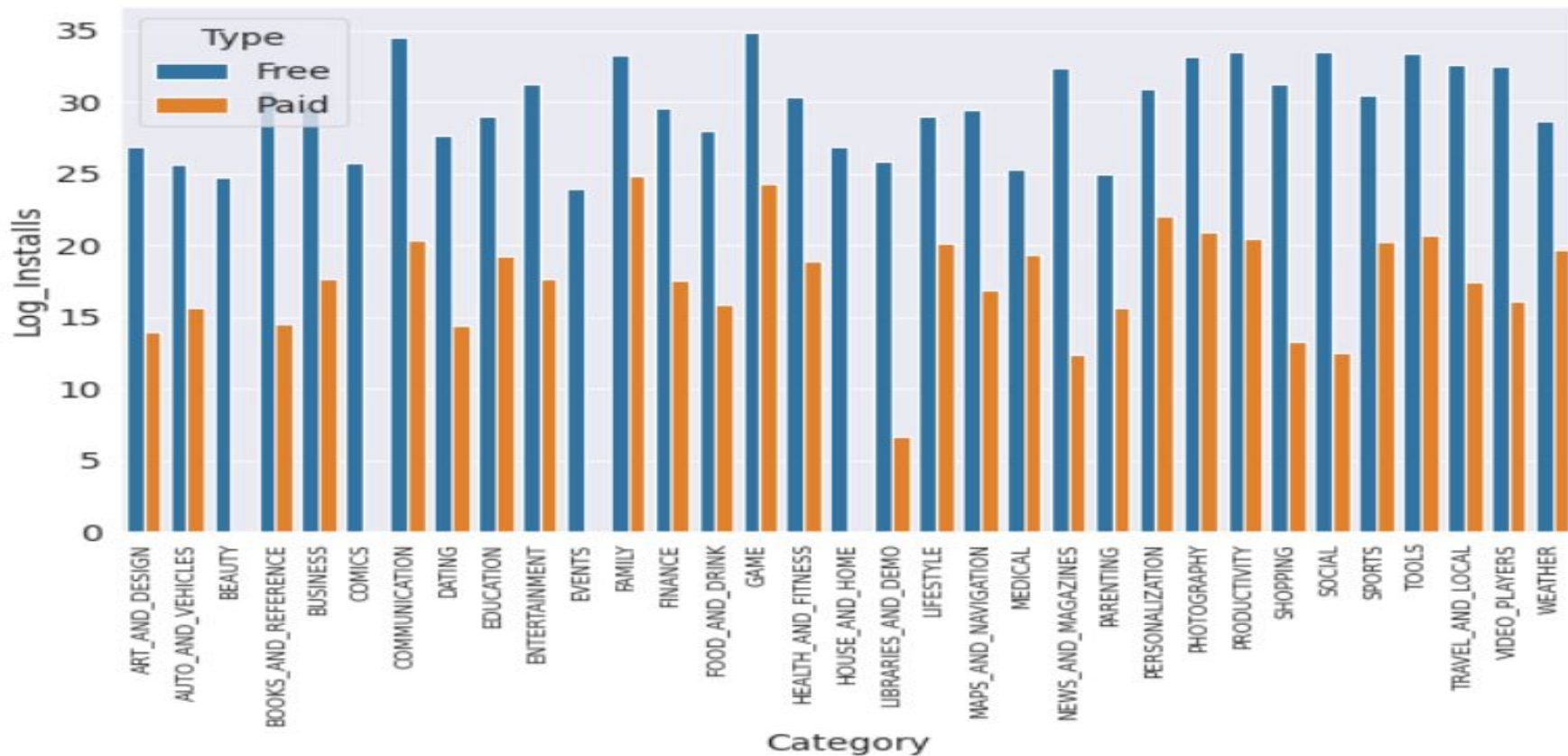
Category wise **Card** got max review then **video players & editors** and so on...

Distribution Of Rating



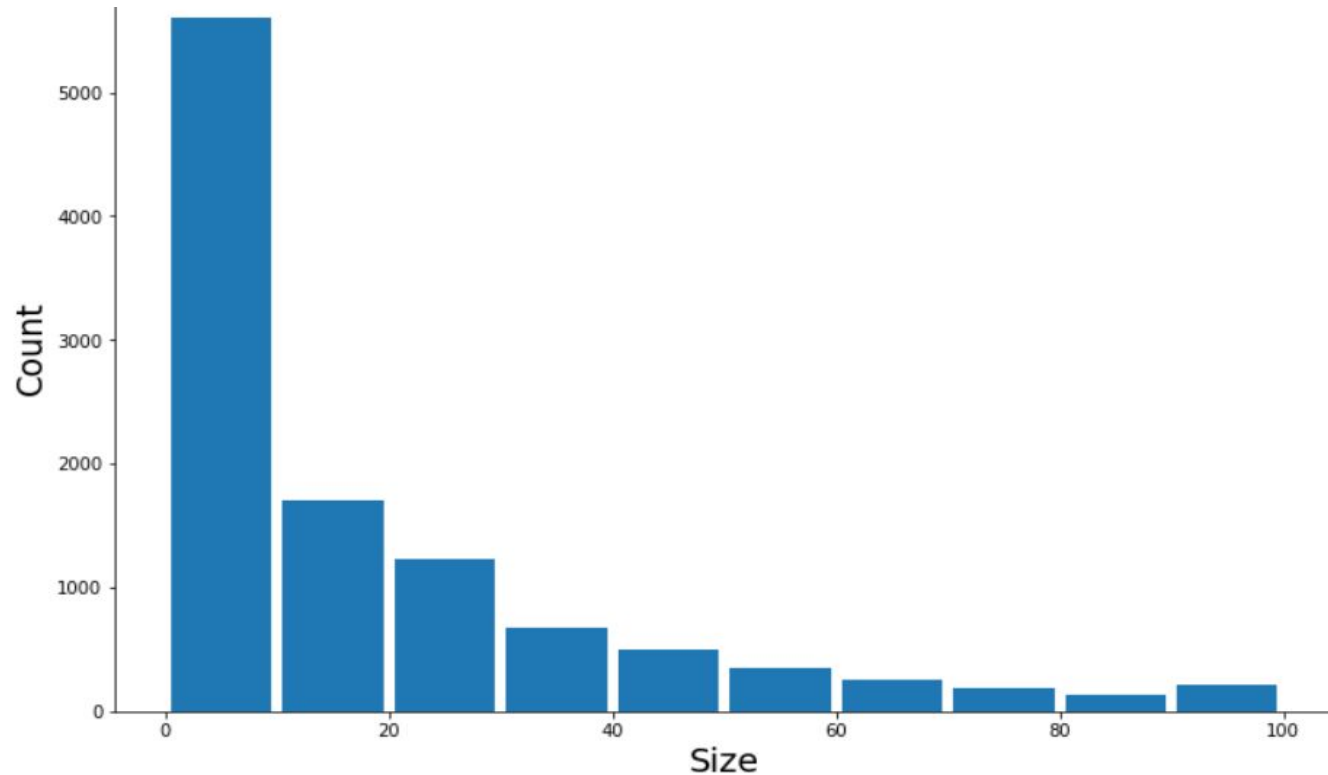
- Most of the application got rating between **3.9 to 4.5**
- There are very few app. Which got rating **5.0**
-

Number of installs type wise according to Category



It can be conclude that the number of free applications installed by the user are high when compared with the paid ones.

Distribution of Size



We can conclude that maximum number of applications present in the dataset are of small size.

Data Summary

(User Reviews Dataset)

- **App:** - Name of the application.
- **Translated_Review:** - This column is containing all the reviews in english.
- **Sentiment:** - sentiment basically determines the attitude or emotion of the user
- **Sentiment_Polarity:-** This value lies between -1 to 1 where
 - -1 means neg.
 - 0 means neutral
 - 1 means positive
- **Sentiment_Subjectivity :-** This column generally refer to personal opinion, emotion or judgement, which lies in the range[0 , 1]

Data Exploration & Data Cleaning

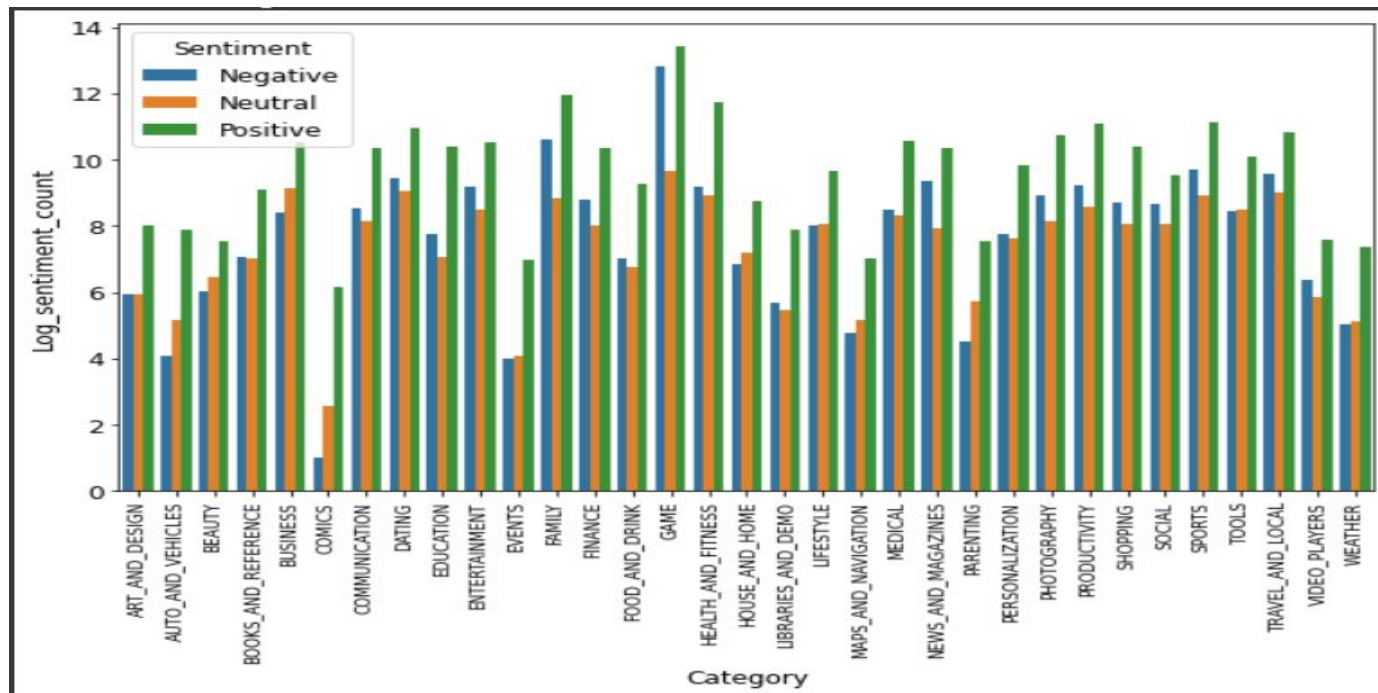
- All the features' are categorical
- *Shape* :- (64295, 5)
- **NaN values** :- Except **App** column all the columns is containing NaN values(41% of total values are NaN)

	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
0	10 Best Foods for You	I like eat delicious food. That's I'm cooking ...	Positive	1.00	0.533333
1	10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.288462
2	10 Best Foods for You	NaN	NaN	NaN	NaN



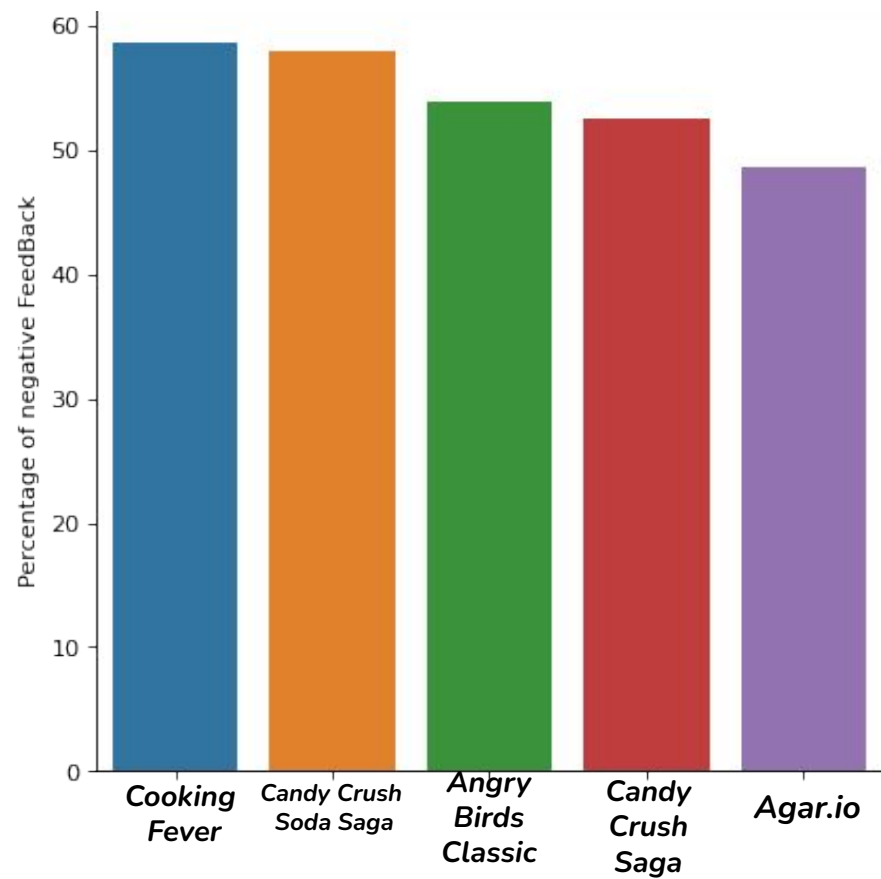
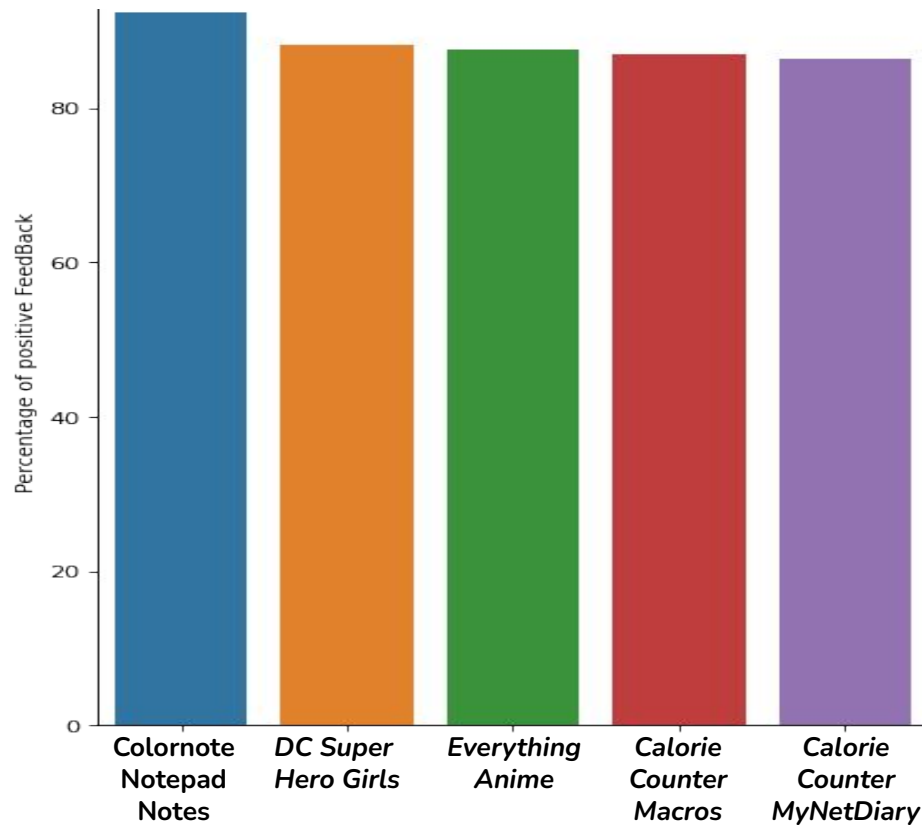
	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
0	10 Best Foods for You	I like eat delicious food. That's I'm cooking ...	Positive	1.00	0.533333
1	10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.288462
3	10 Best Foods for You	Works great especially going grocery store	Positive	0.40	0.875000

Number of installs type wise according to Genres



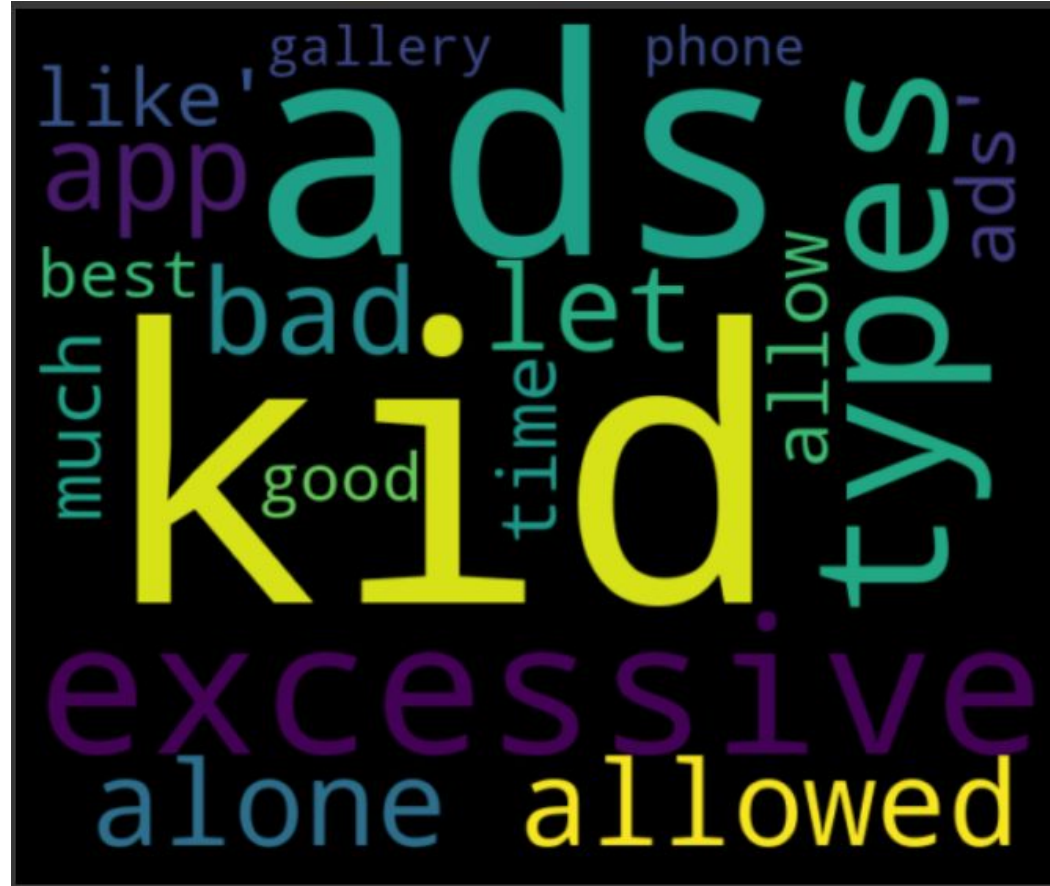
We can say that **Game** is having highest positive review as well as negative review

Top 5 apps with most negative and positive feedback

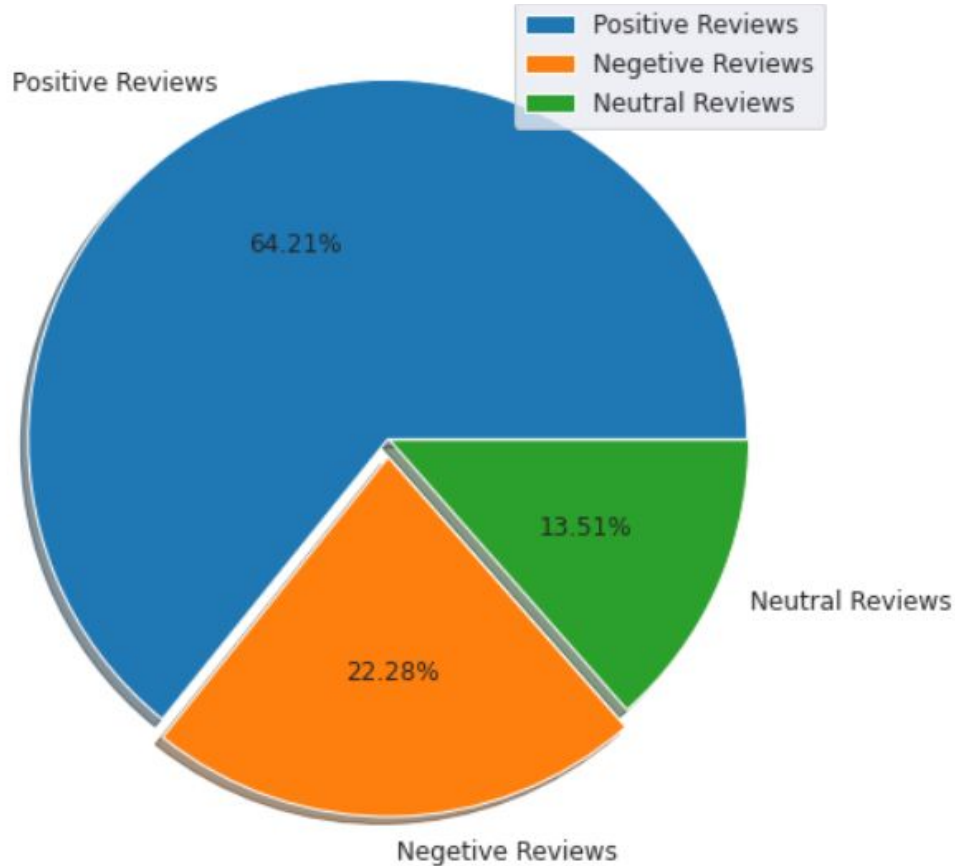


Word clouds or tag clouds are *graphical representations* of word frequency that give greater prominence to words that appear more frequently in a source text. The larger the word in the visual the more common the word was in the document(s).

- Kid
- ads
- types
- excessive
- alone
- allowed



Proportions of Apps on sentiment



64% customers are satisfied



13.5% customers are not reaching us with pos. Or neg. feedback



22% customers are not satisfied

Conclusion

- *Most of the rating is in between 4.0 to 4.5*
- *The number of free applications installed by the user are high as compared with the paid ones.*
- *Bulky applications are less installed by the user.*
- *Category wise GAME got highest no of pos. Review as well as neg. Review.*
- *Max. no of sentiment subjectivity lies between 0.4 to 0.7. From this we can conclude that maximum number of users give reviews to the applications, according to their experience.*
- *The most occurred word according to WordCloud*
 - *Kid*
 - *ads*
 - *types*
 - *excessive*
 - *alone*
 - *allowed*
- *65% customers are satisfied*
- *25% are not satisfied with the application hence they leave a neg review*

