Capstone Project 1

Exploratory Data Analysis (EDA) on Play Store App Review Analysis

Team Members

Ajay Pedhekar

Narendra Ghodse

Uday Firake

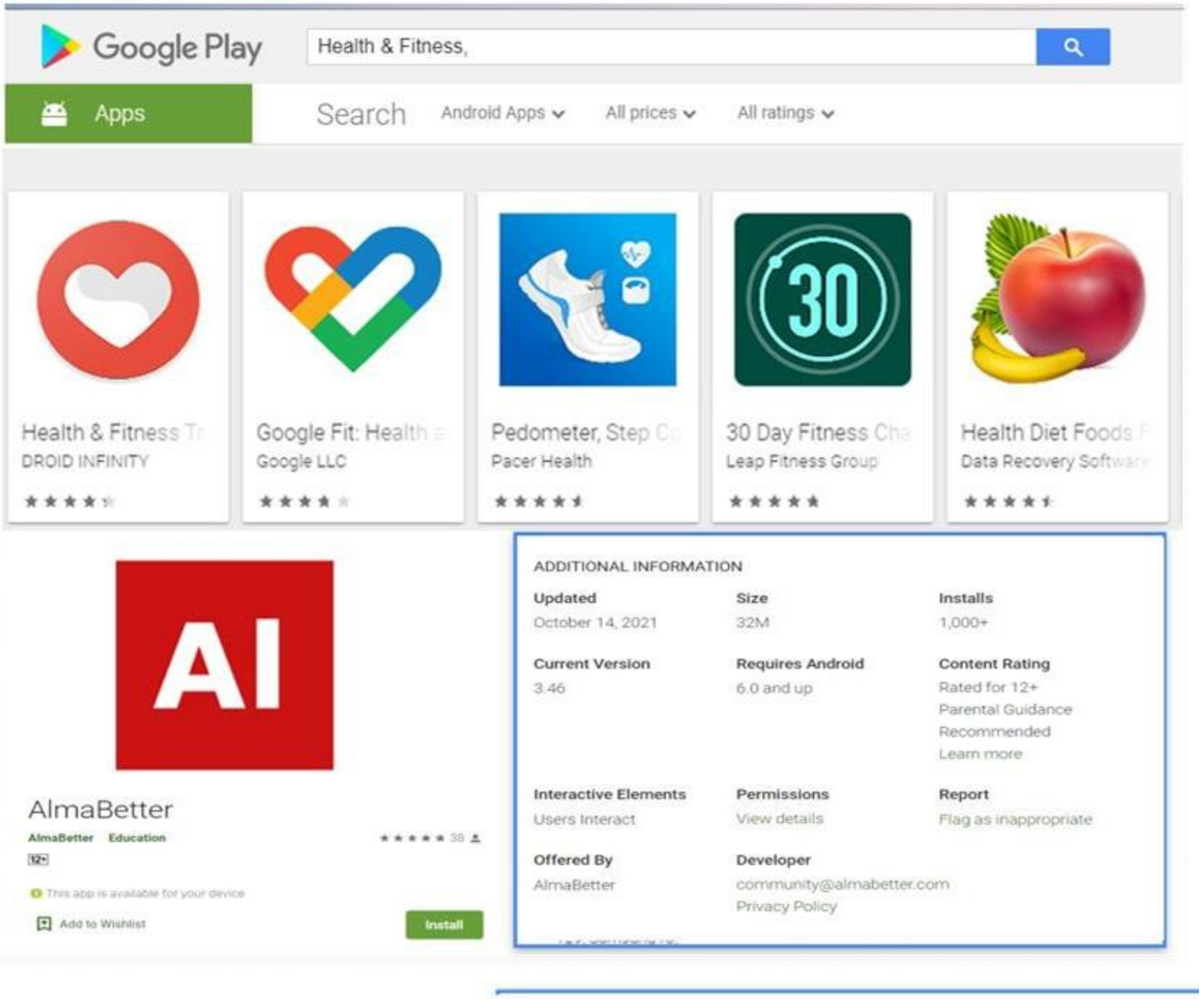
Saurabh Kumar

Sandhya Sah

Contents

Introduction

- Google Play Store is a formerly Android Market, developed by Google. It serves as the official app store for certified devices running on the Android operating system, It is allowing users to browse and download applications developed with the Android software development kit (SDK) and published through Google.
- Google Play also serves as a digital media store, offering music, books, movies, and television programs.
- It is the largest and most popular Android app store with around three million different apps available for the users to download.



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.

Each app (row) has values for category, rating, size, and more. Another dataset contains customer reviews of the android apps.

Explore and analyze the data to discover key factors responsible for app engagement and success.

Data Summary

A. Play store Dataset

This file contains data of the Play store applications. It contains 10,841 rows of data with different application names and following columns:

App: It defines names of different applications to be reviewed.

Category: Category of the application it belongs to such as family, game, beauty, business, entertainment, education...etc.

Rating: Overall user rating of the application. The users have rated the app out of 5, with 1 being the lowest rating and 5 being the highest.

Reviews: The number of user reviews each app has received.

Size: The memory size needed to install the application.

Installs: The number of times each application has been installed by users.

Data Summary (Cont'd)

Type: Tells about the Free or paid version of the app.

Price: Notifies the price of the app.

Content Rating: This column specifies the intended audience of the app such as teens, mature 21+, or everyone.

Genres: The sub-category for each app. Example: for the Education category, this could be Education: Pretend Play, for example.

Last Updated: Release date of the most recent update for the app.

Current Ver: The app's current version.

Android Ver: The oldest version of Android OS supported by the app.

Data Summary (Cont'd)

B. **User Reviews Dataset**

This dataset contains the result of the sentiment analysis. It has 64,295 rows of data with the following additional attributes:

App: Name of the applications.

Translated_Review: Either the original review in English, or a translated version if the original review is in another language. Moreover, the text in each review was pre-processed and attributed with three new features.

Sentiment: The result of the sentiment analysis conducted on a review dataset. The value is either Positive, Neutral, or Negative.

Sentiment_Polarity: Sentiment polarity is a value indicating the positivity or negativity of the sentiment. It ranges from -1 (most negative) to 1 (most positive).

Sentiment_Subjectivity: Sentiment Subjectivity is a value ranging from 0 to 1, which indicates the subjectivity of the review. Here, lower values indicate the review which is based on factual information, and higher values indicate the review that is based on personal opinions, public opinions or judgements.

Data Exploration

A. Let's Explore play store dataset

```
[11] #Looking at the shape of data.
     play_store_data_df.shape
```

```
(10841, 13)
```

```
# Looking at the columns of data.
play_store_data_df.columns
```

```
Index(['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type',
       'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver',
       'Android Ver'],
      dtype='object')
```

```
[13] # Information about data.
     play_store_data_df.info()
```

```
<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: 1.1+ MB
```

```
# To get first 10 rows of data.
play_store_data_df.head(10)
```

	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
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
5	Paper flowers instructions	ART_AND_DESIGN	4.4	167	5.6M	50,000+	Free	0	Everyone	Art & Design	March 26, 2017	1	2.3 and up
6	Smoke Effect Photo Maker - Smoke Editor	ART_AND_DESIGN	3.8	178	19M	50,000+	Free	0	Everyone	Art & Design	April 26, 2018	1.1	4.0.3 and up
7	Infinite Painter	ART_AND_DESIGN	4.1	36815	29M	1,000,000+	Free	0	Everyone	Art & Design	June 14, 2018	6.1.61.1	4.2 and up
8	Garden Coloring Book	ART_AND_DESIGN	4.4	13791	33M	1,000,000+	Free	0	Everyone	Art & Design	September 20, 2017	2.9.2	3.0 and up
9	Kids Paint Free - Drawing Fun	ART_AND_DESIGN	4.7	121	3.1M	10,000+	Free	0	Everyone	Art & Design;Creativity	July 3, 2018	2.8	4.0.3 and up

```
[15] # To get last 10 rows of data.
     play_store_data_df.tail(10)
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
10831	payermonstationnement.fr	MAPS_AND_NAVIGATION	NaN	38	9.8M	5,000+	Free	0	Everyone	Maps & Navigation	June 13, 2018	2.0.148.0	4.0 and up
10832	FR Tides	WEATHER	3.8	1195	582k	100,000+	Free	0	Everyone	Weather	February 16, 2014	6	2.1 and up
10833	Chemin (fr)	BOOKS_AND_REFERENCE	4.8	44	619k	1,000+	Free	0	Everyone	Books & Reference	March 23, 2014	0.8	2.2 and up
10834	FR Calculator	FAMILY	4.0	7	2.6M	500+	Free	0	Everyone	Education	June 18, 2017	1.0.0	4.1 and up
10835	FR Forms	BUSINESS	NaN	0	9.6M	10+	Free	0	Everyone	Business	September 29, 2016	1.1.5	4.0 and up
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	0	Everyone	Education	July 25, 2017	1.48	4.1 and up
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3.6M	100+	Free	0	Everyone	Education	July 6, 2018	1	4.1 and up
10838	Parkinson Exercices FR	MEDICAL	NaN	3	9.5M	1,000+	Free	0	Everyone	Medical	January 20, 2017	1	2.2 and up
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	Varies with device	1,000+	Free	0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Varies with device

✓ 0s completed at 10:21 PM

Data Exploration (Cont'd)

B. Let's Expore user reviews dataset

```
[18] # Looking at the shape of data.
     user_reviews_df.shape
```

```
(64295, 5)
```

```
# Looking at the columns of data.
user_reviews_df.columns
```

```
Index(['App', 'Translated_Review', 'Sentiment', 'Sentiment_Polarity',
       'Sentiment_Subjectivity'],
      dtype='object')
```

```
[20] # Information about data.
     user_reviews_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64295 entries, 0 to 64294
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   App                     64295 non-null  object
 1   Translated_Review       37427 non-null  object
 2   Sentiment               37432 non-null  object
 3   Sentiment_Polarity      37432 non-null  float64
 4   Sentiment_Subjectivity  37432 non-null  float64
dtypes: float64(2), object(3)
memory usage: 2.5+ MB
```

```
[21] # To get first 10 rows of data.
     user_reviews_df.head(10)
```

```
# To get first 10 rows of data.
user_reviews_df.head(10)
```

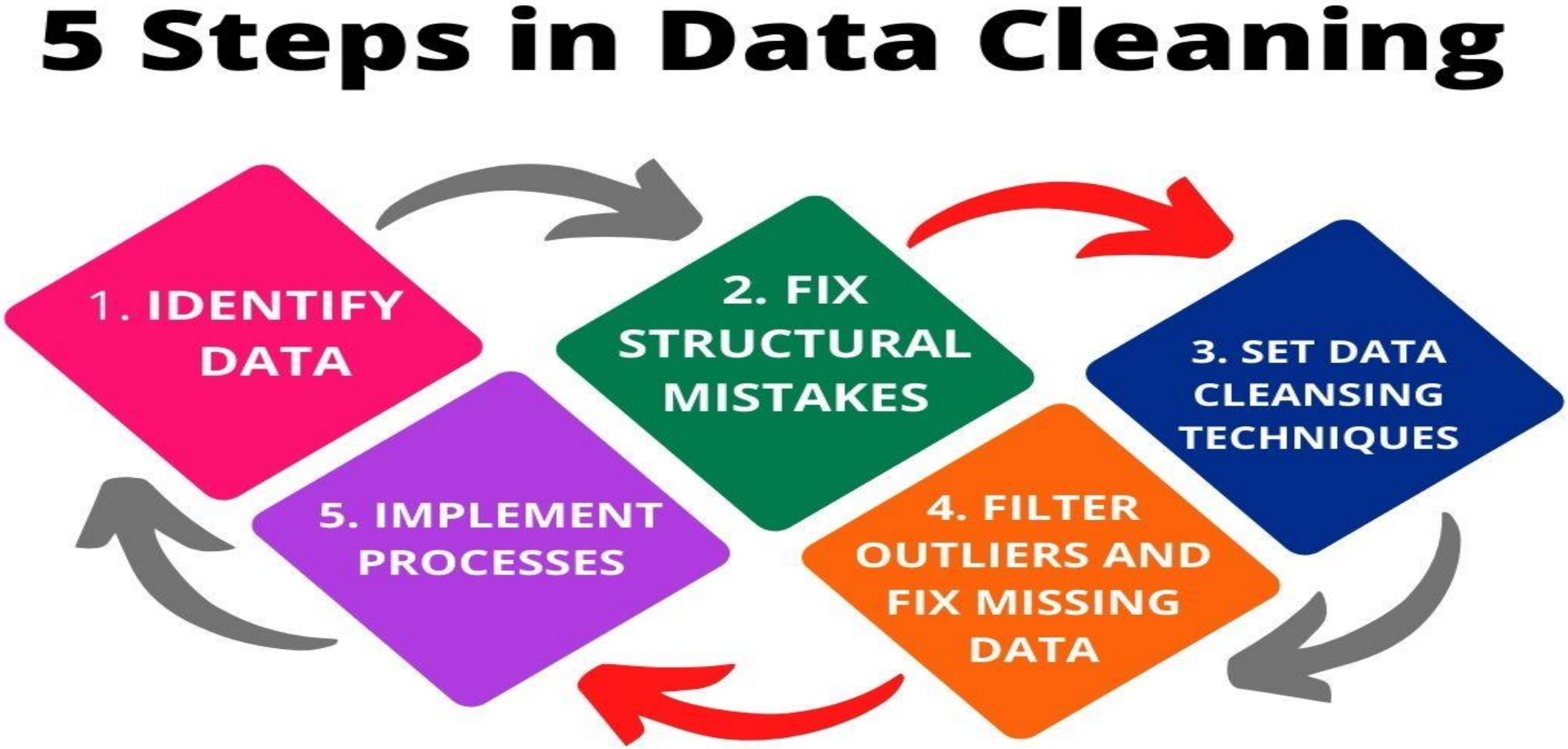
	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
0	10 Best Foods for You	I like eat delicious food. That's I'm cooking ...	Positive	1.00	0.533333
1	10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.288462
2	10 Best Foods for You	NaN	NaN	NaN	NaN
3	10 Best Foods for You	Works great especially going grocery store	Positive	0.40	0.875000
4	10 Best Foods for You	Best idea us	Positive	1.00	0.300000
5	10 Best Foods for You	Best way	Positive	1.00	0.300000
6	10 Best Foods for You	Amazing	Positive	0.60	0.900000
7	10 Best Foods for You	NaN	NaN	NaN	NaN
8	10 Best Foods for You	Looking forward app,	Neutral	0.00	0.000000
9	10 Best Foods for You	It helpful site ! It help foods get !	Neutral	0.00	0.000000

```
[22] # To get last 10 rows of data.
     user_reviews_df.tail(10)
```

	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
64285	Houzz Interior Design Ideas	NaN	NaN	NaN	NaN
64286	Houzz Interior Design Ideas	NaN	NaN	NaN	NaN
64287	Houzz Interior Design Ideas	NaN	NaN	NaN	NaN
64288	Houzz Interior Design Ideas	NaN	NaN	NaN	NaN
64289	Houzz Interior Design Ideas	NaN	NaN	NaN	NaN
64290	Houzz Interior Design Ideas	NaN	NaN	NaN	NaN
64291	Houzz Interior Design Ideas	NaN	NaN	NaN	NaN
64292	Houzz Interior Design Ideas	NaN	NaN	NaN	NaN
64293	Houzz Interior Design Ideas	NaN	NaN	NaN	NaN
64294	Houzz Interior Design Ideas	NaN	NaN	NaN	NaN

Data Cleaning

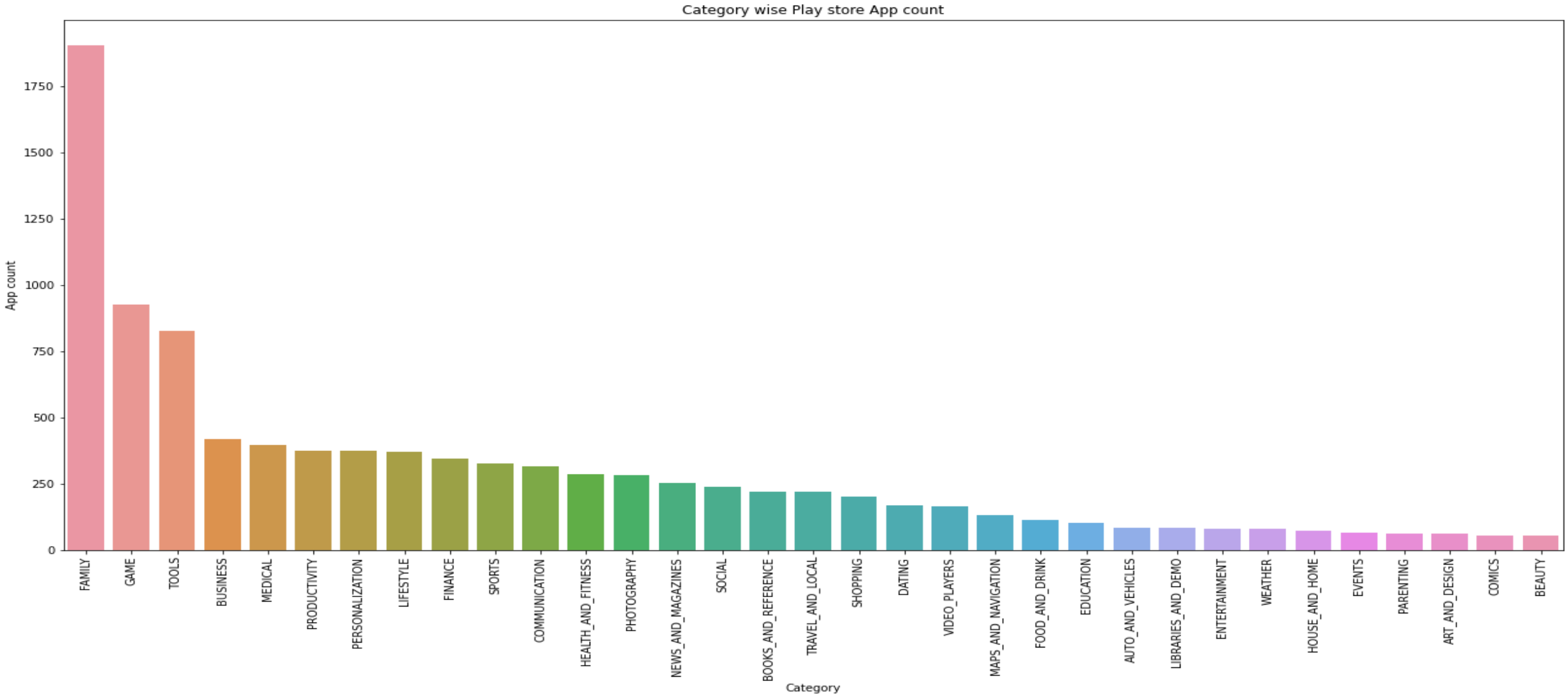
Data Cleaning (Cont'd)



Data Visualization

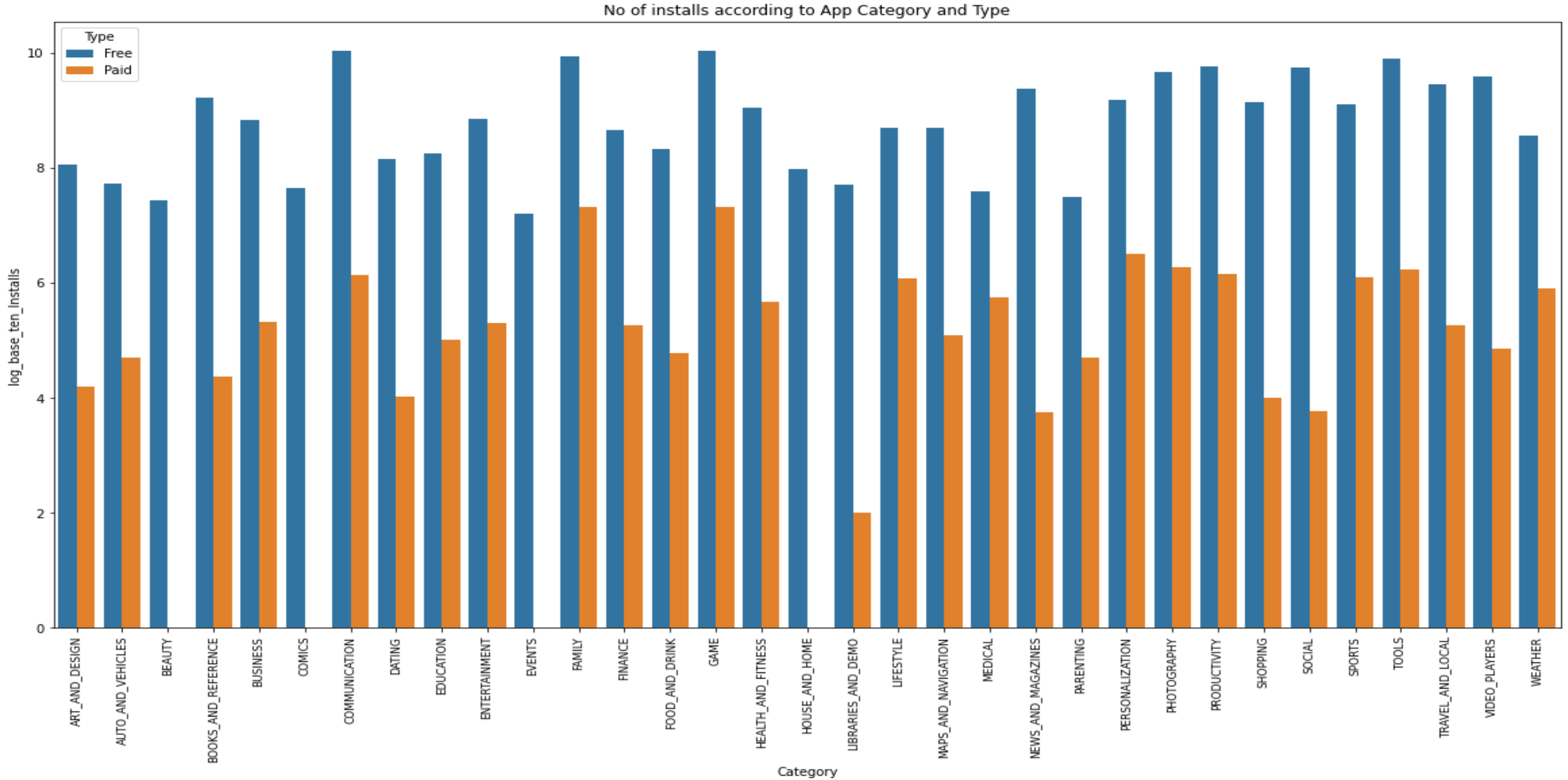
The number of apps in each category.

- As we can see the plot maximum no. of apps present in the play store are comes under Family, Games and Tools Category.
- And minimum no. of apps are present in Comics, Art and design, Beauty category.



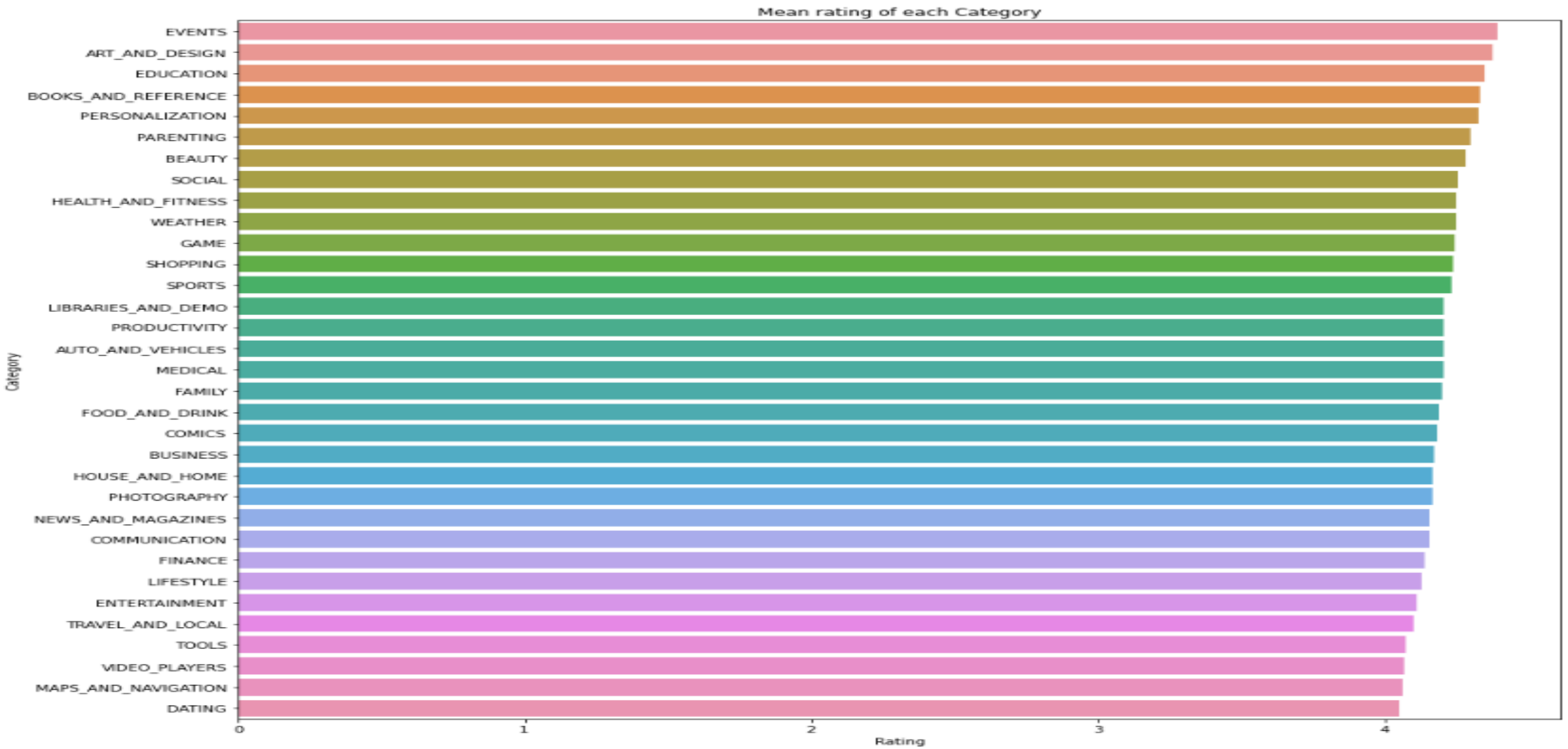
Comparing Category And App Type (Free, Paid) by count of installs.

- We can deduce that The Game, Communication, Tools and Family category has the highest number of Installation.



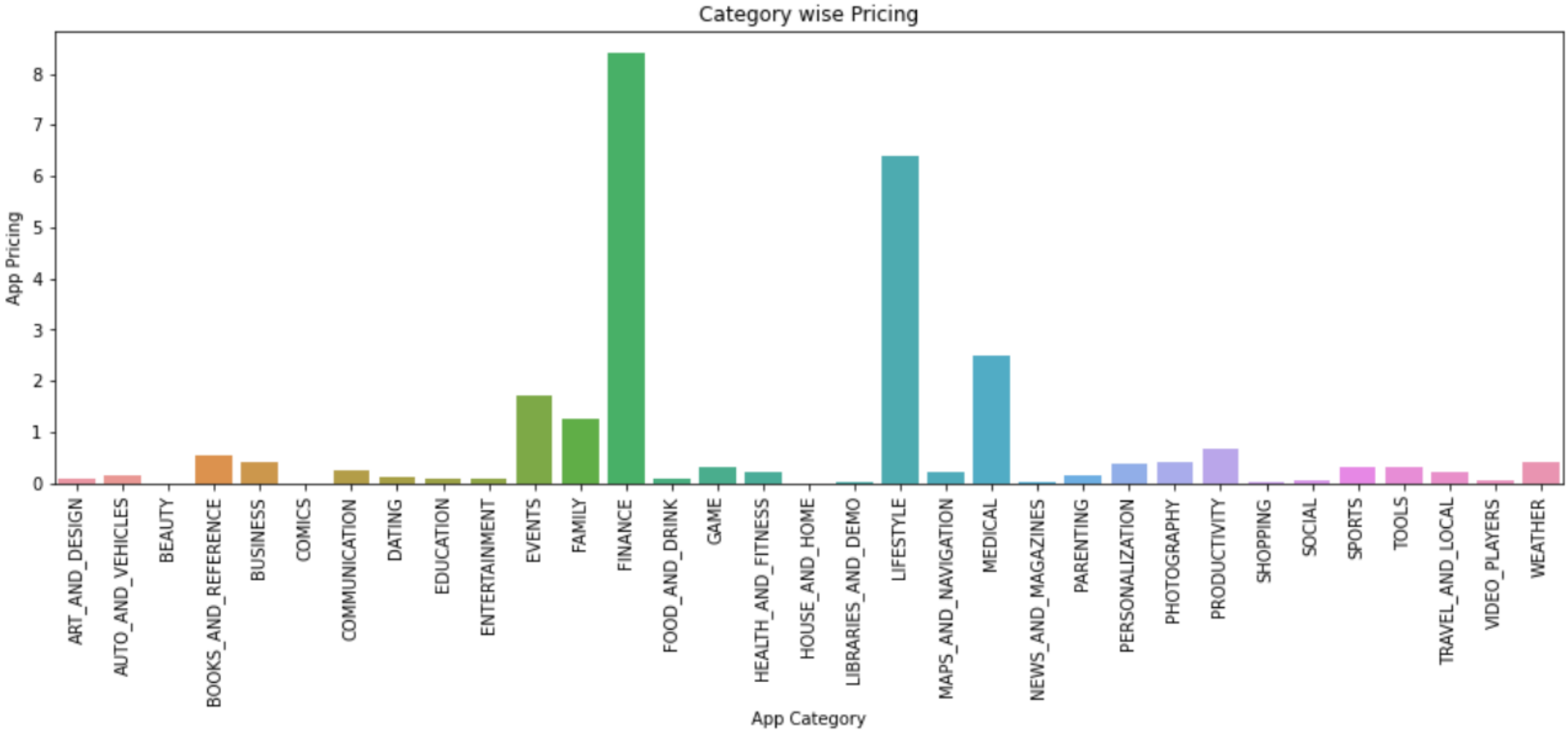
The average rating of each category.

- In the given plot we can say that maximum average rating is high for Event, Art and design, Education and Books category apps.
- The minimum average rating for Dating, Maps and navigation, Video players and Tools category apps.



Category Wise Pricing Visualization.

- The majority of apps except Finance, Lifestyle, Events & Medical, Family categories were free to install.
- Lifestyle, Finance, and Medical categories had the biggest number of paid apps available for download.



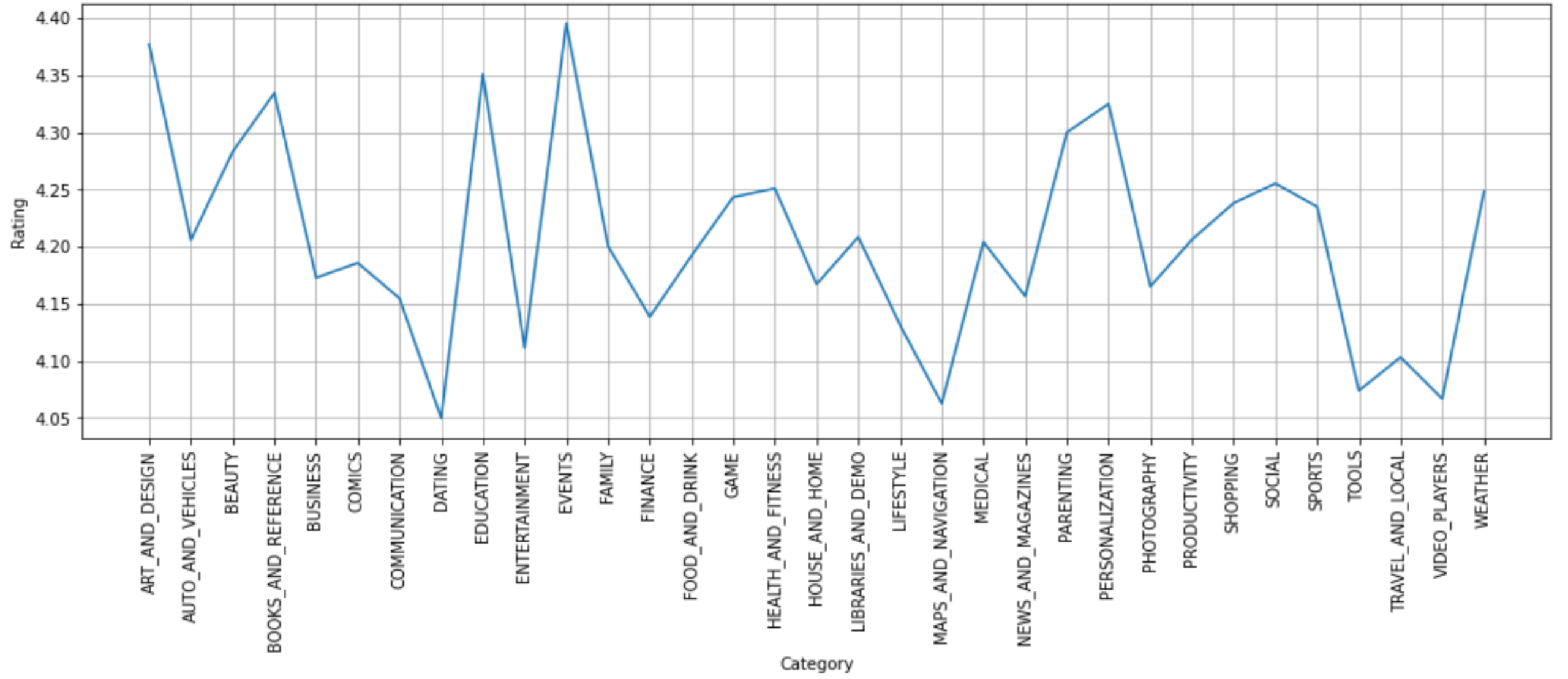
Rating visualization of Apps for given Category.

A. Highest Ratings apps :

Events, Art and Design, Education, Books and reference categories has got 4.3 and above ratings.

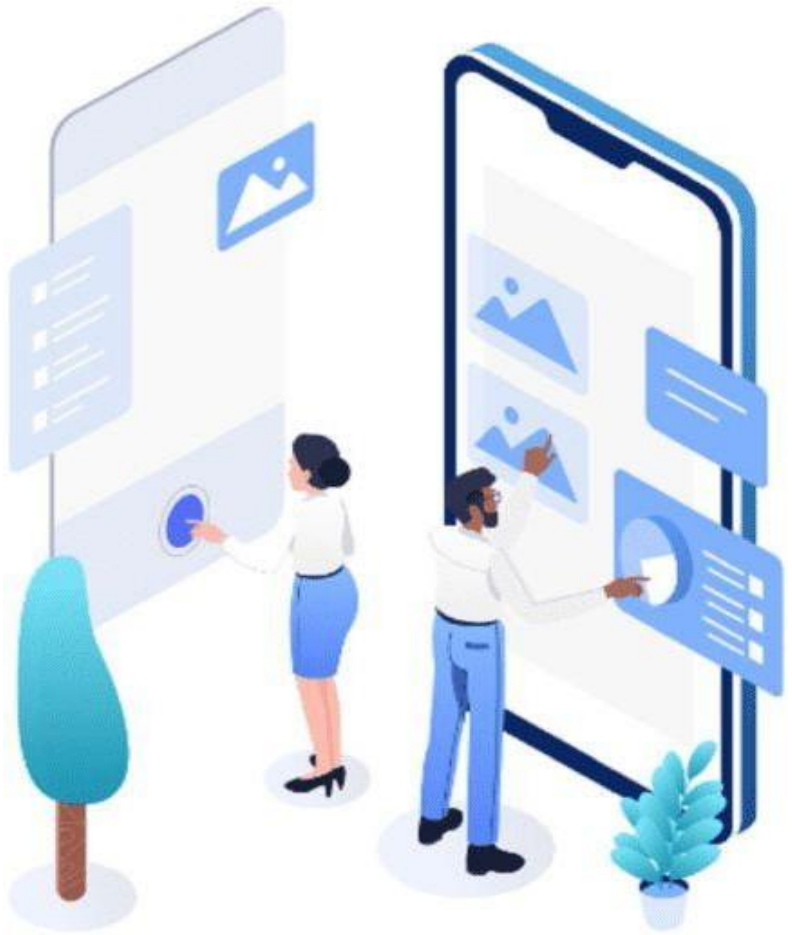
B. Lowest Ratings apps :

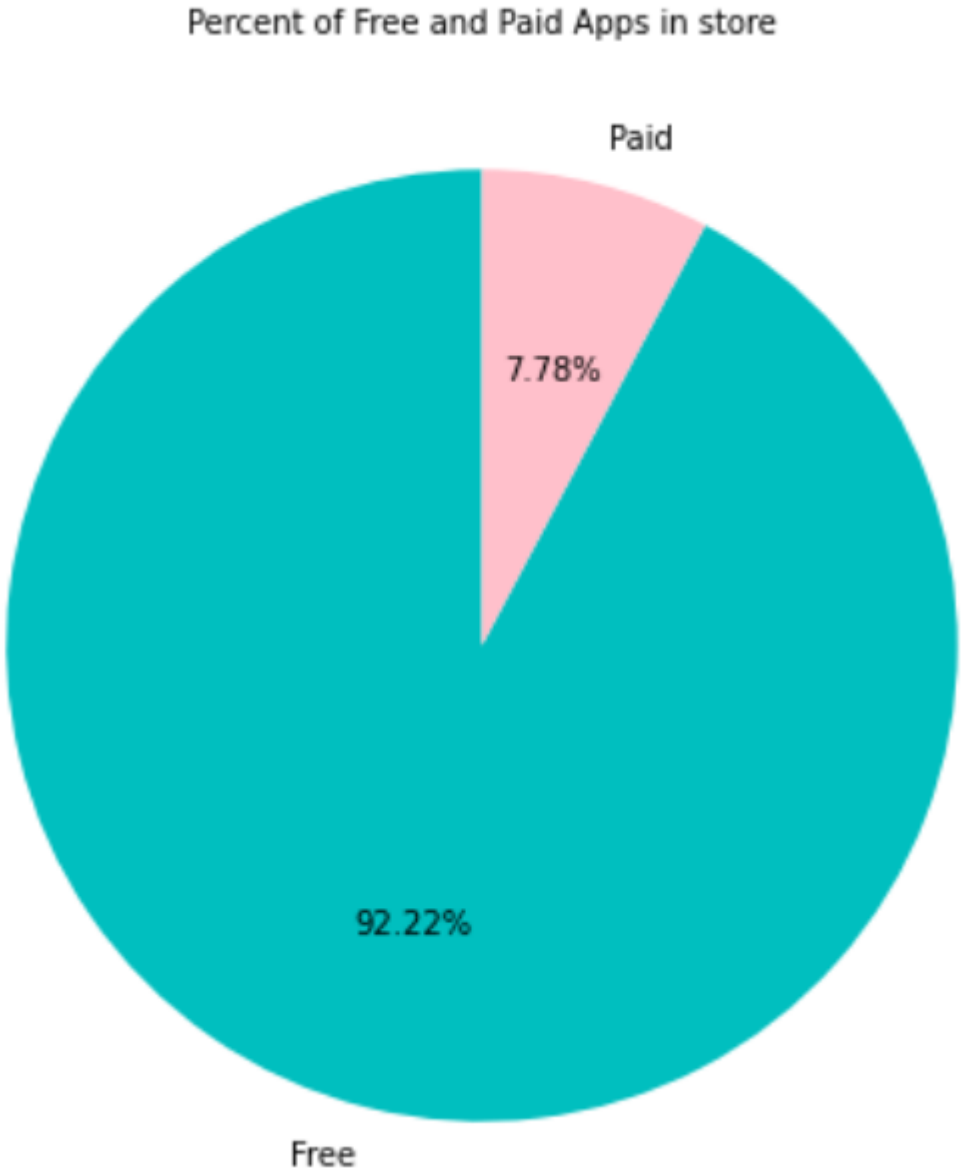
Dating, Maps and navigation, Video players, Tools category has got lowest ratings.



Percentage of free vs paid apps in play store

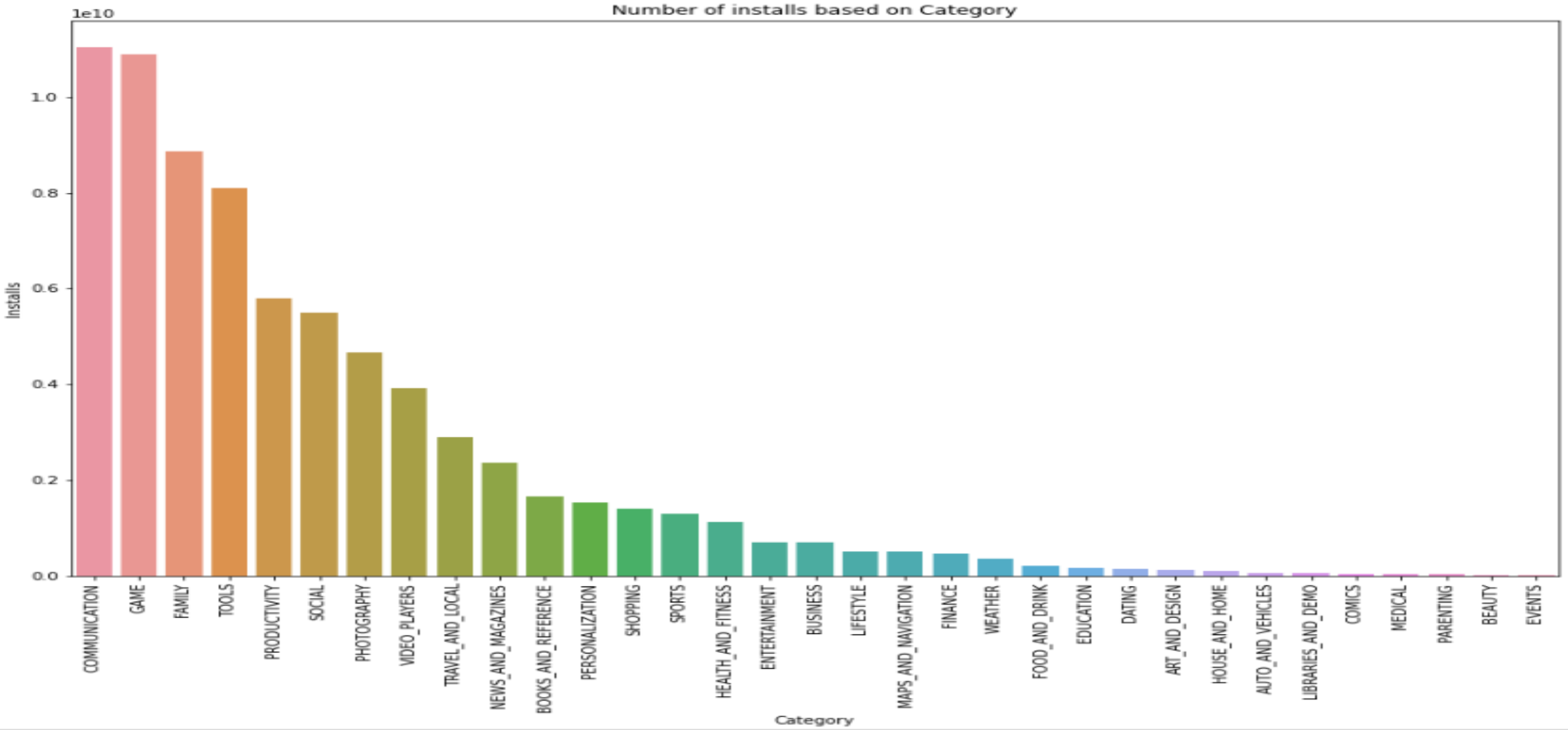
- Here we can see that there are 92.22% of apps are free and only 7.78% of Apps are paid on Play store.





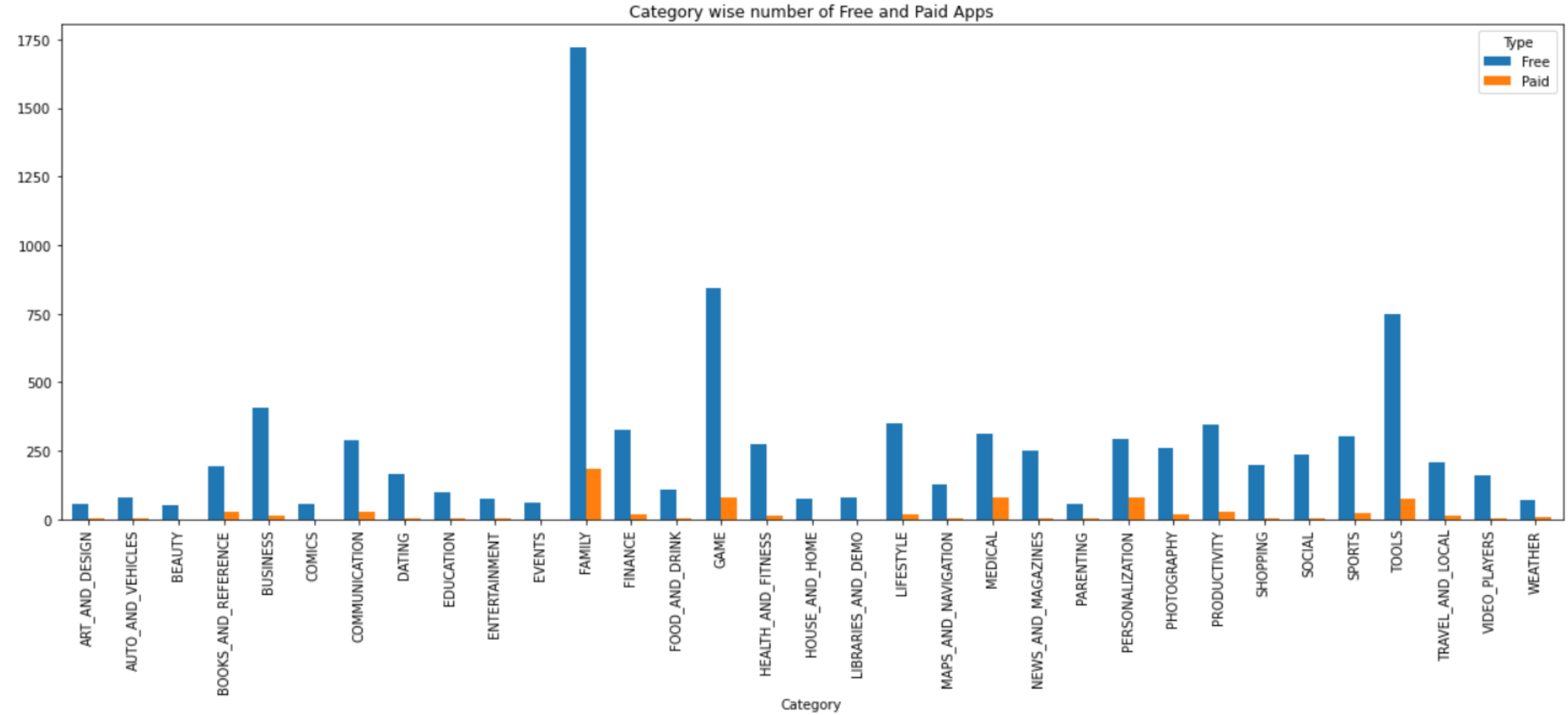
The Number of Apps Installs based on Category.

- As we can see, the Maximum number of apps present in google play store comes under Communication, Family & Games Category.
- But as per the installation and requirement in the market plot, scenario is not the same. Here maximum installed apps comes under Games, Communication, Family and Tools.



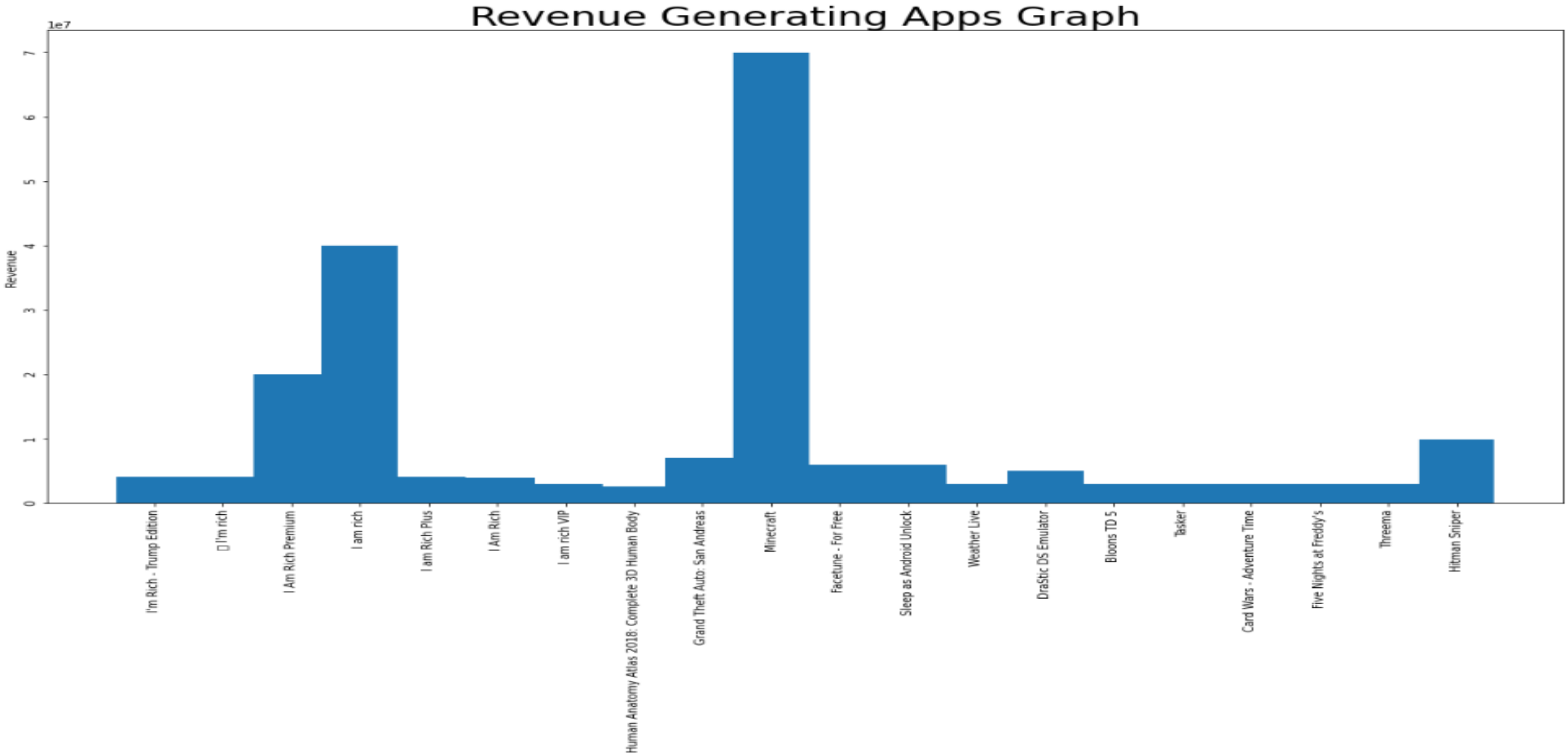
The number of free and paid apps category wise.

- In the given plot we can say that maximum free apps are Family, Tools, business and Games category apps.
- And the higher paid apps are also Family, Games, Medical and Tools category apps.



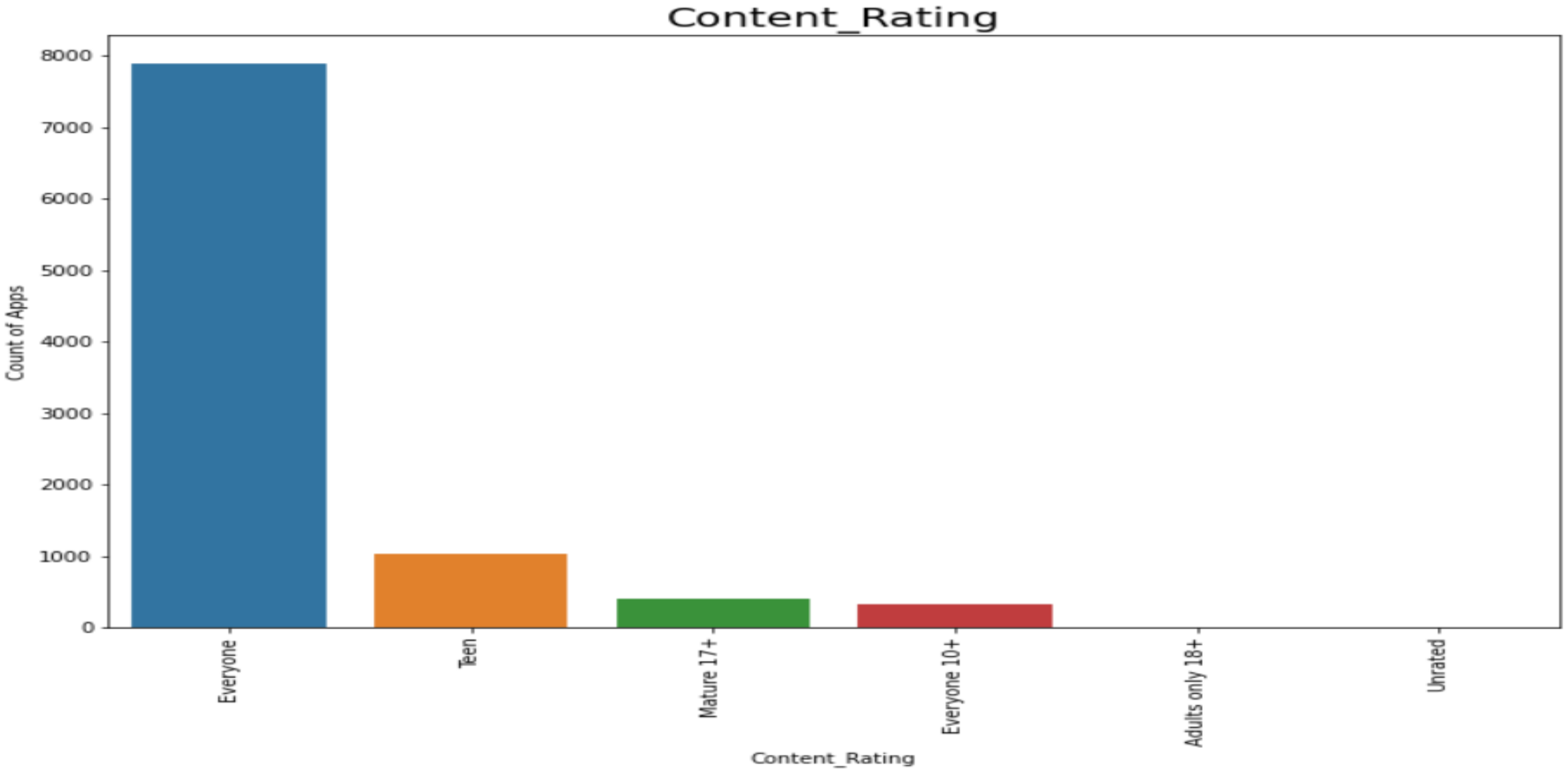
Graph of Revenue for given Apps.

- The majority of apps, which generate revenue is between 2 to 8 million .
- Minecraft is generating highest revenue among all, which is about 7 million.



Comparing Number of Apps Age Group(Content rating) wise.

- we can see from the graph that most of the apps are for “everyone” (Approx. 7900) so that it could capture the most of the customer-base.
- As we can see the plot minimum no. of apps are in google play store comes under Unrated, Adults only 18+ content rating.

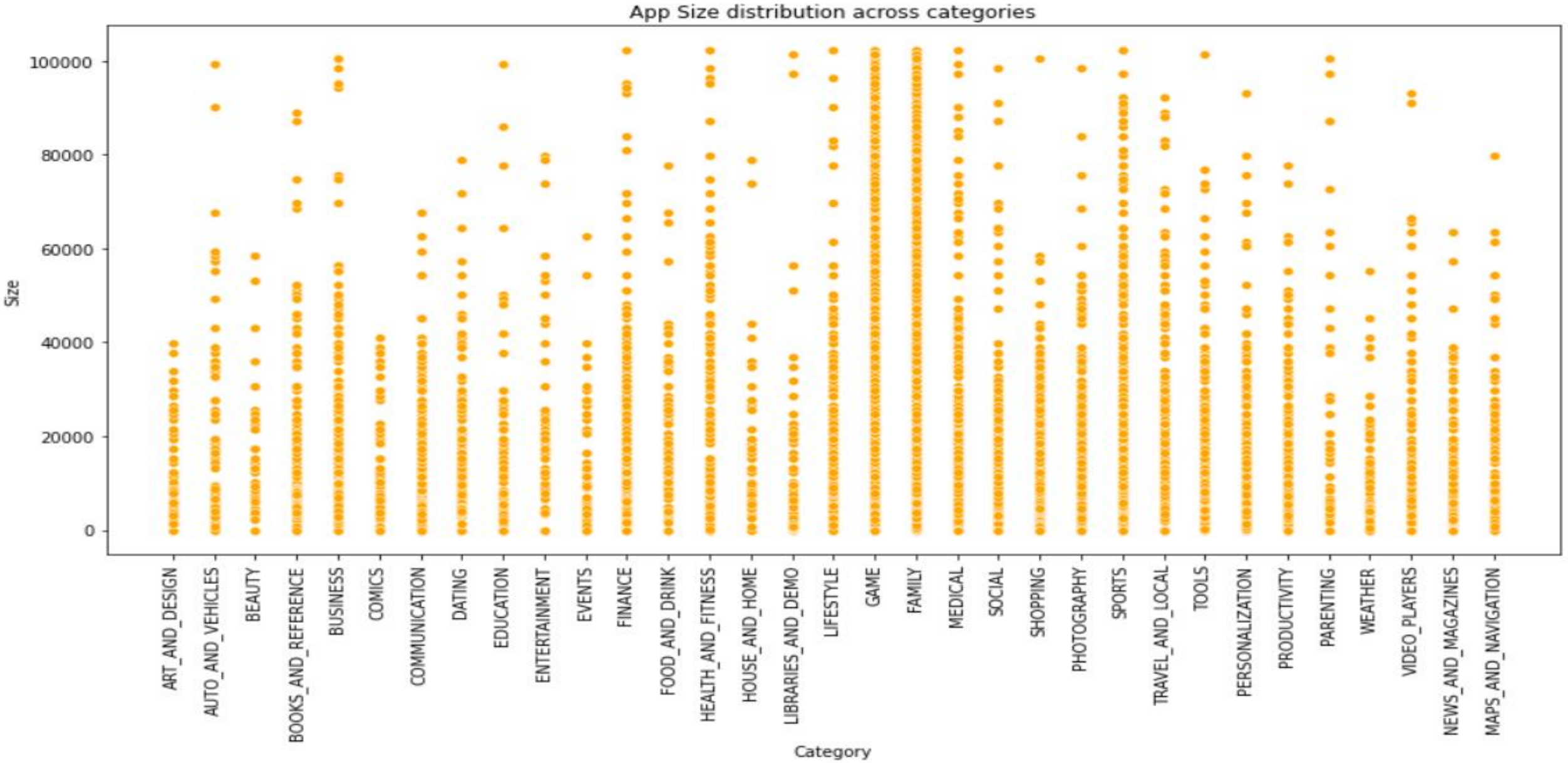


Correlation Between Different Parameters

- Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together.
- In simple terms, it tells us how much does one variable changes for a slight change in another variable. It may take positive, negative and zero values depending on the direction of the change.
- We can see that installs and reviews have the strongest correlation.

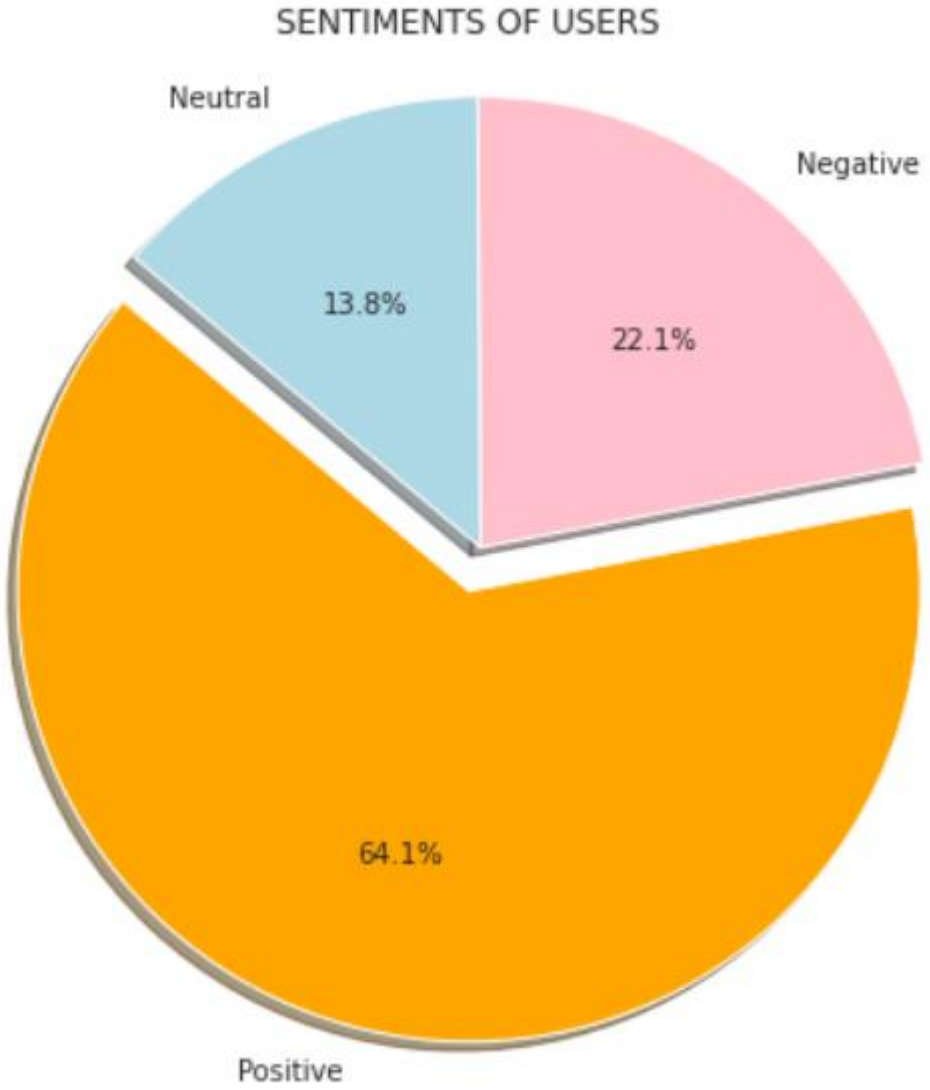
The distribution of Size across different Category.

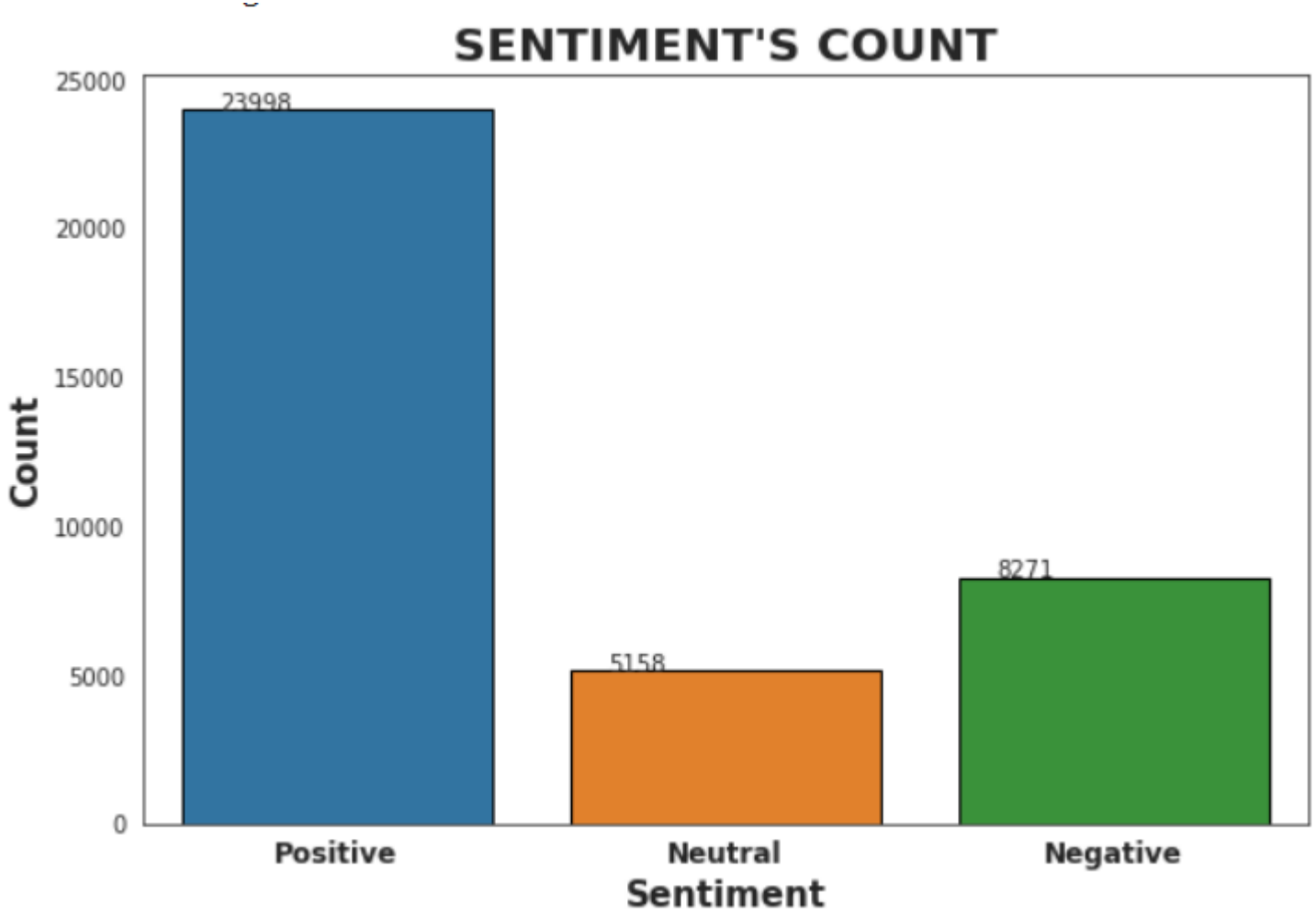
- As we see the scatter plot we can say that maximum no. of apps size distribution on Family, Games and Sports.
- Most of users preferred less M B size apps.



Sentiments Of Users & Sentiments Counts

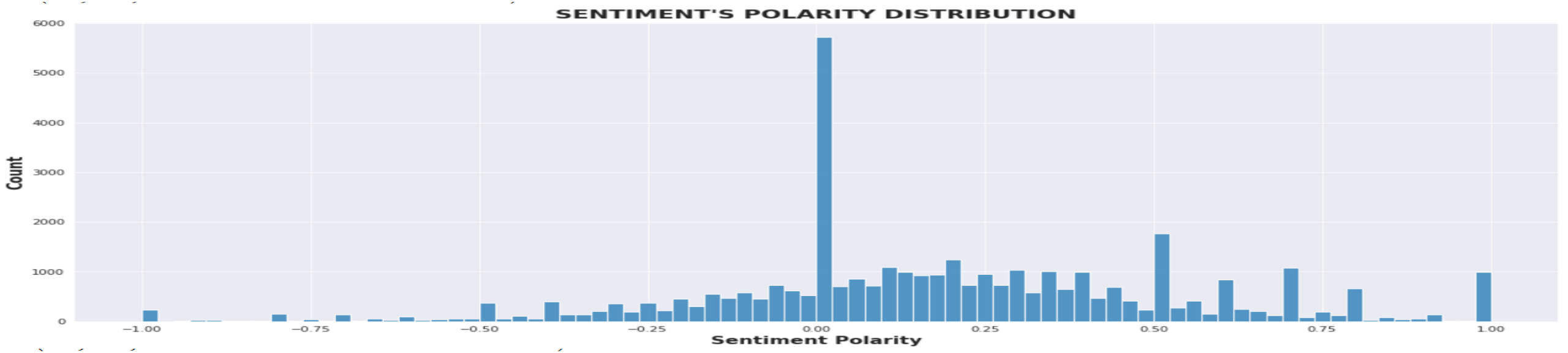
- Most of the sentiments users are positive and having review percent of 64.1
- Also the high counts of sentiments are positive of 23998.

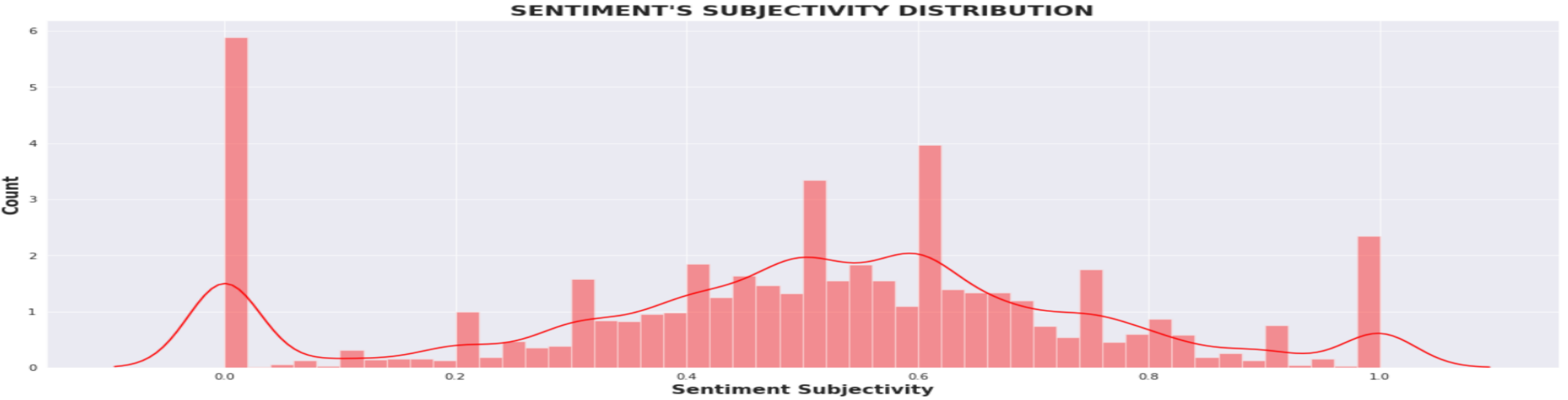




Sentiment Polarity Distribution & Sentiment Subjectivity Distribution

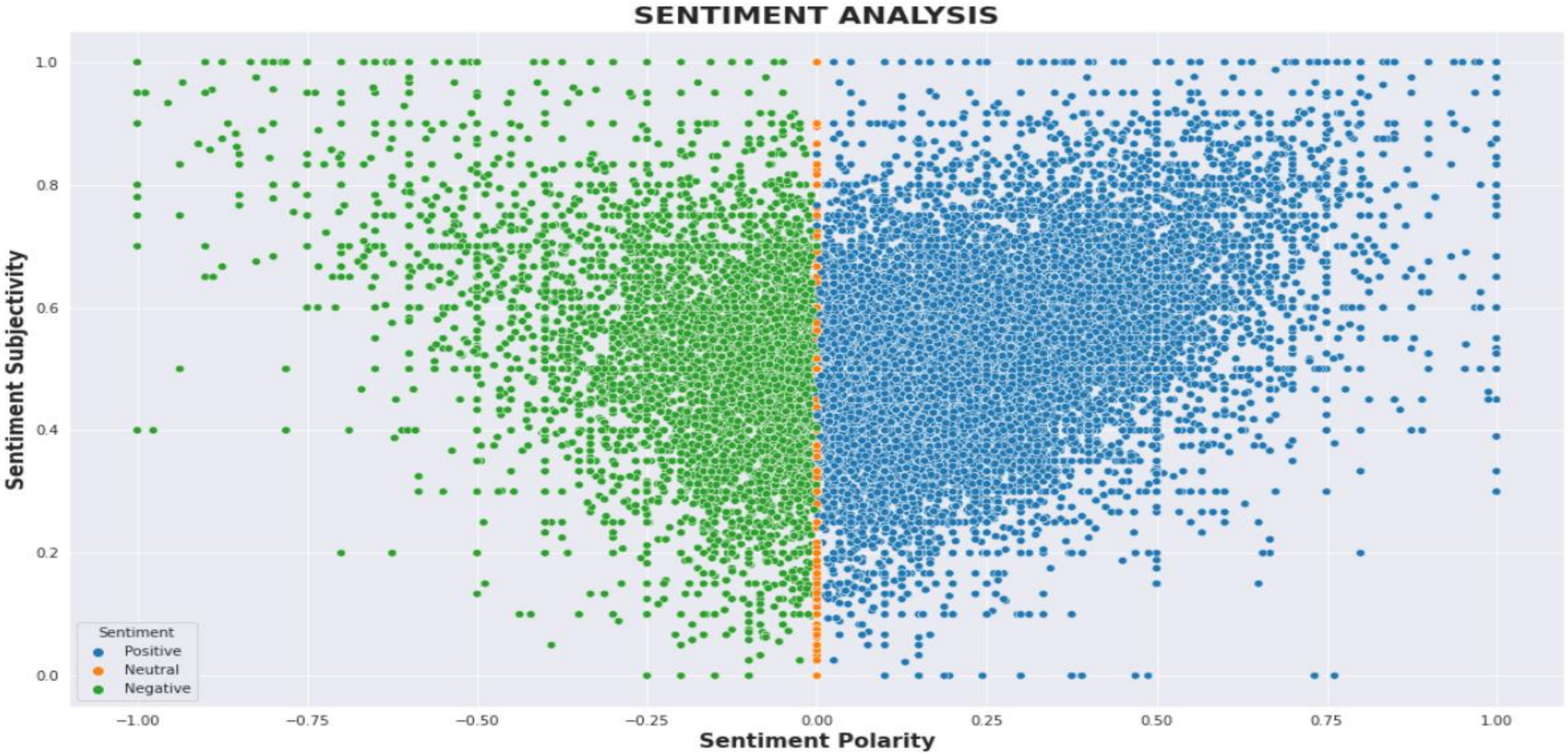
- Sentiment Polarity is a float which lies in the range of [-1,1] In this plot the maximum no. of positive reviews lies between 0 to 0.5.
- Sentiment Subjectivity generally lies in the range of [0,1] In that maximum no. of sentiment subjectivity lies between 0.4 to 0.8.





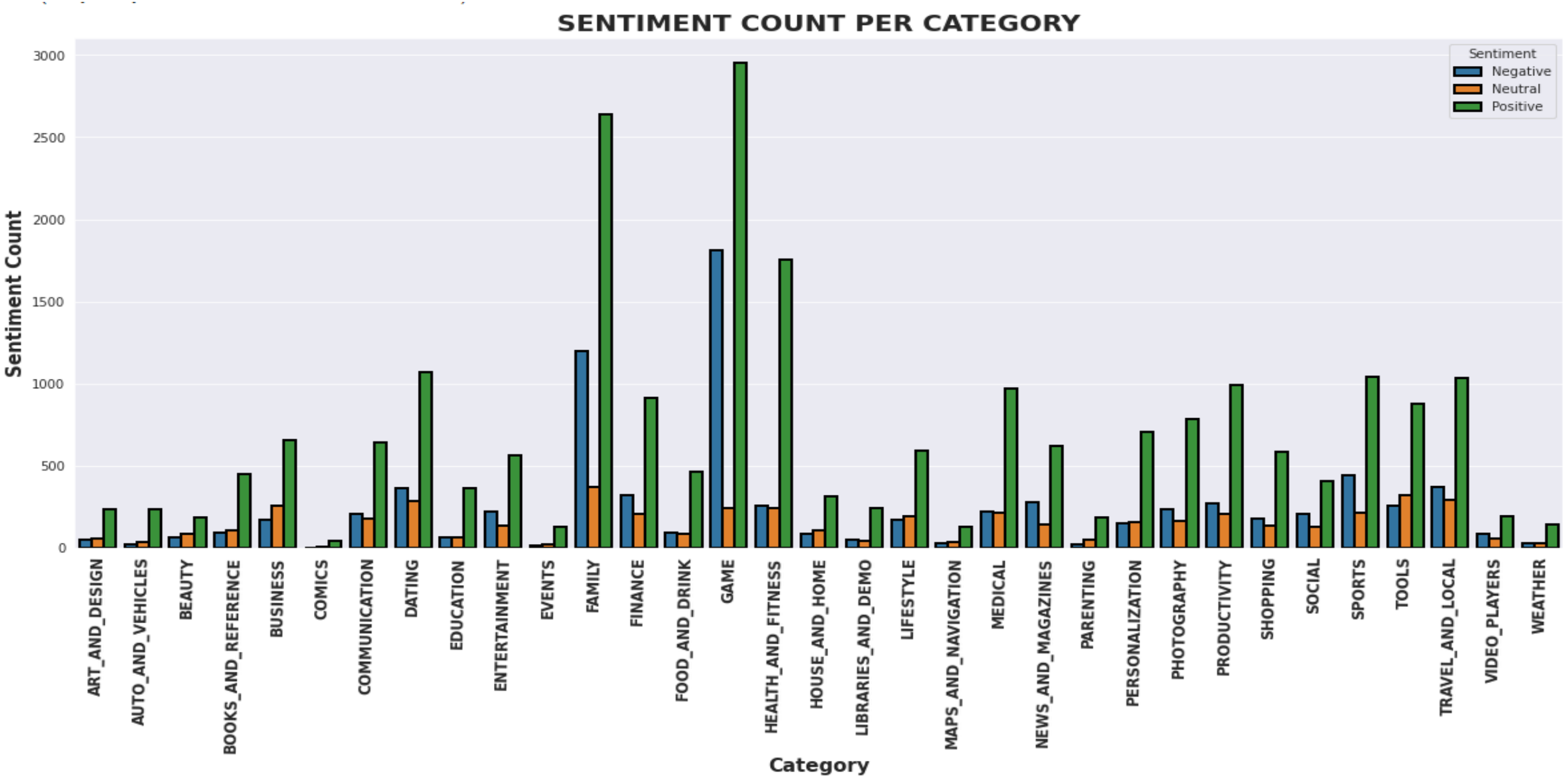
Sentiment Analysis on Polarity & Subjectivity

- From the given scatter plot it can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in some number of case Sentiment Subjectivity proportional to Sentiment Polarity.



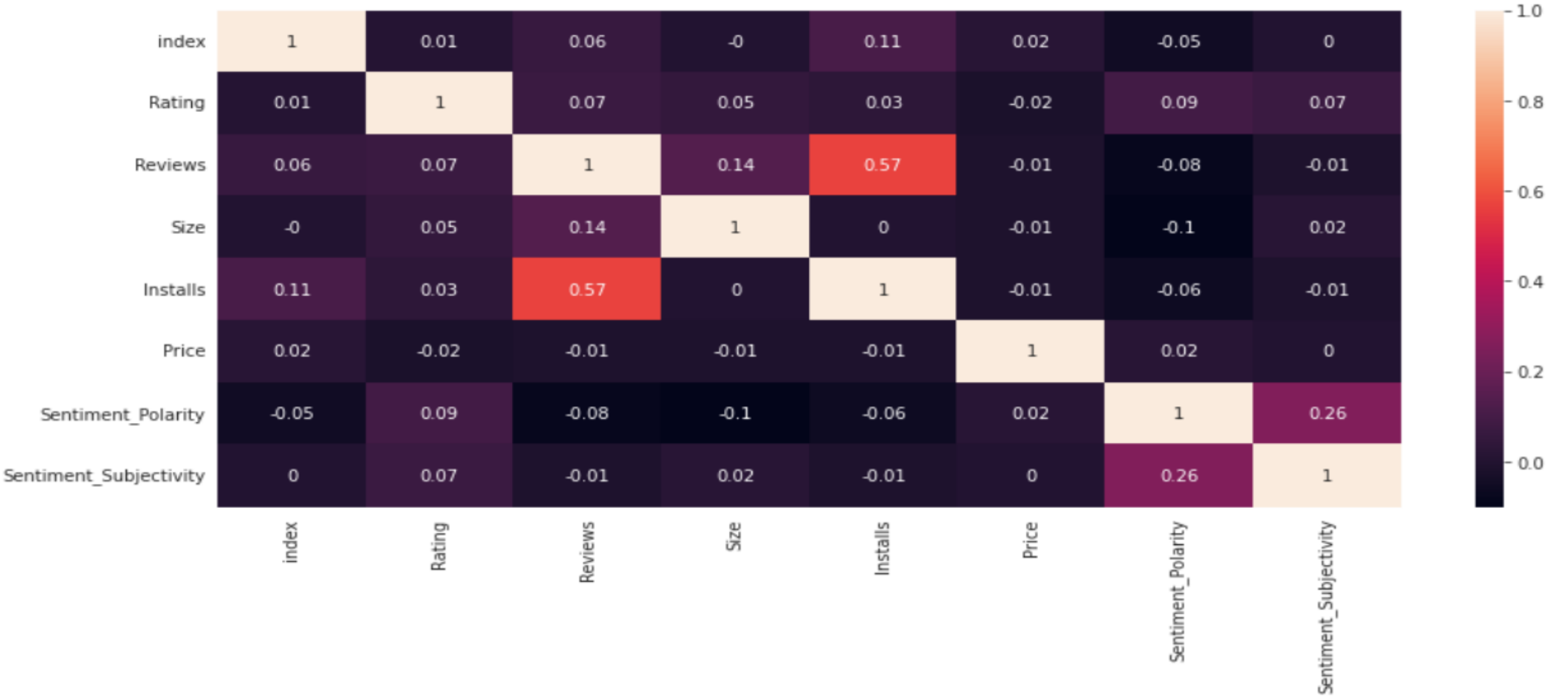
Sentiment Analysis on different Category basis

- As we can see in bar chart maximum number of positive category are Games, Family & Health and fitness.
- And we can say the maximum number of negative category are Dating, Games, Family & Travel and Local.



Correlation of Merged Dataset

- The Correlation tells us how much does one variable changes for a slight change in another variable. It may take positive, negative and zero values depending on the direction of the change.
- We can see that installs and reviews have the strongest correlation.



Conclusion

- The majority of users prefer free apps. The size of the app does not directly affect their decision to use paid or free apps.
- The highest number of apps have been published in the Family, Games, and Tools categories on the Play store, but the highest number of installations have come from the Games category.
- The apps which have the ratings above 4, are actually targeting all the people and not a certain age group.
- Dating, Maps, Travel and Video Player are the categories which have received the lowest ratings from the user and we know that these apps have high human interaction.
- On average the application size is in between 10 MB to 30 MB.
- Medical, Family, game, Tools apps are the expensive apps.
- This helps us to understand the application market and tells us as to which would be the right category to invest money to make good profits.
- The sentiment does not influence the final rating of the apps because of the proportional behavior.
- Popular categories have a more positive sentiment.
- Positive reviews are higher than negative and neutral but the sentiment subjectivity is not always. proportional to sentiment polarity.
- The pie chart of review sentiments and observe that the percentage of positive sentiments is near about 64%.
- Graph of sentiment subjectivity and observe the maximum number of sentiment subjectivity lies between 0.4 to 0.8. From this we can conclude that the maximum number of users give reviews to the applications, according to their experience.

THANK YOU