EDA CAPSTONE PROJECT

Play Store App Reviews Analysis

By-Ghanshyam Menariya
and Brijesh Patel
(Cohort Berlin)

POINT FOR DISCUSSION

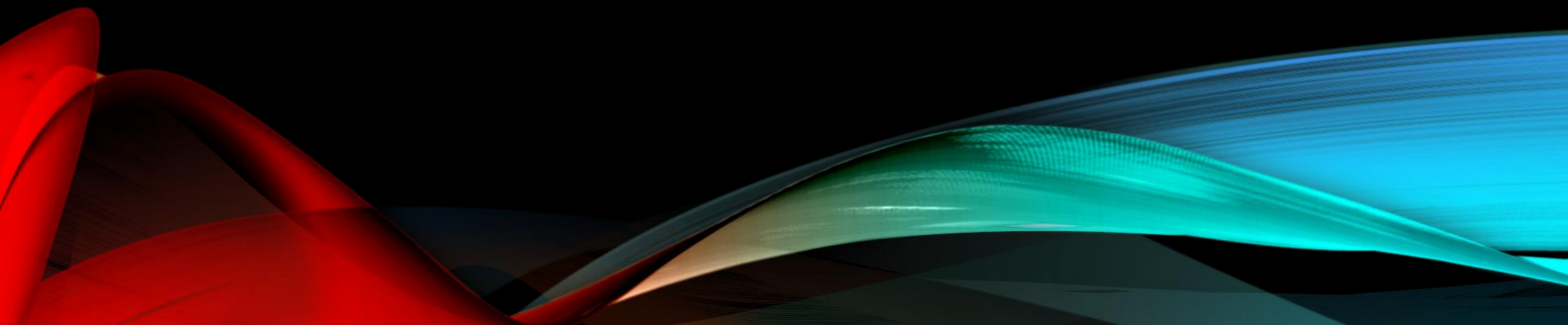
- Data Summary
- Count of app. in each category differentiated by their type
- Count which Category app. Has most download
- Distribution of Rating
- Top 20 Genres
- Top Category App which have most Price
- installed applications for each Category
- Content type according to there Rating and Type
- count the category have most reviews
- Sentiment type wise according to Genres
- Representing Percentage of Review Sentiments
- Conclusion

DATA SUMMARY

- **df** : it contains the all information about the dataframe
- **df_app_count** : it contain the information about Category , Type , App
- **top_20_genres_installs** : it Contains top 20 Genres
- **cate_count** : it contain the count of all category values
- **top_20_category_installs** : it contain top 20 installed Category
- **cr_group** : it have contain the information of Category and Reviews
- **user_reviews_df** : it contain the new users review csv
- **merged_df** : its have merged dataframe of user_reviews_df
- **Counts** : list of merged sentiment data

Cleanse and Validate Data

This step is crucial for removing faulty data and filling in gaps. Important tasks here includes:
Removing extraneous data
Filling in missing values.
Conforming data to a standardized pattern



DATASET MAY CONTAIN DUPLICATE VALUES FOR PARTICULAR APPLICATION:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Androi
53	ROBLOX	GAME	4.5	4447388	67M	100,000,000+	Free	0	Everyone 10+	Adventure;Action & Adventure	July 31, 2018	2.347.225742	4.1 a
01	ROBLOX	GAME	4.5	4447346	67M	100,000,000+	Free	0	Everyone 10+	Adventure;Action & Adventure	July 31, 2018	2.347.225742	4.1 a
48	ROBLOX	GAME	4.5	4448791	67M	100,000,000+	Free	0	Everyone 10+	Adventure;Action & Adventure	July 31, 2018	2.347.225742	4.1 a
41	ROBLOX	GAME	4.5	4449882	67M	100,000,000+	Free	0	Everyone 10+	Adventure;Action & Adventure	July 31, 2018	2.347.225742	4.1 a
70	ROBLOX	GAME	4.5	4449910	67M	100,000,000+	Free	0	Everyone 10+	Adventure;Action & Adventure	July 31, 2018	2.347.225742	4.1 a
16	ROBLOX	FAMILY	4.5	4449910	67M	100,000,000+	Free	0	Everyone 10+	Adventure;Action & Adventure	July 31, 2018	2.347.225742	4.1 a
38	ROBLOX	FAMILY	4.5	4450855	67M	100,000,000+	Free	0	Everyone 10+	Adventure;Action & Adventure	July 31, 2018	2.347.225742	4.1 a
06	ROBLOX	FAMILY	4.5	4450890	67M	100,000,000+	Free	0	Everyone 10+	Adventure;Action & Adventure	July 31, 2018	2.347.225742	4.1 a
27	ROBLOX	FAMILY	4.5	4443407	67M	100,000,000+	Free	0	Everyone 10+	Adventure;Action & Adventure	July 31, 2018	2.347.225742	4.1 a

ROBLOX app is having identical rows with difference in number of reviews. It may have happened that for the same app, the data has been scraped in different points of time. So we have kept row of an app with maximum number of reviews, assuming it to be the latest one.

```
# Converting KB to MB
df['Size'] = df['Size'].apply(lambda x: str(x).replace('Varies with device', 'NaN') if 'Varies with device' in str(x) else x)
df['Size'] = df['Size'].apply(lambda x: str(x).replace('M', '') if 'M' in str(x) else x)
df['Size'] = df['Size'].apply(lambda x: str(x).replace(',', '') if 'M' in str(x) else x)
df['Size'] = df['Size'].apply(lambda x: float(str(x).replace('k', '')) / 1000 if 'k' in str(x) else x)
df['Size'] = df['Size'].apply(lambda x: float(x))
```

CONVERT KB TO MB

Since, size of the applications present in the datset are in MB and KB. Therefore, for ease in data processing, entire size column is converted to MB.

```
df['Installs'] = df['Installs'] .map(lambda x: x.rstrip('+'))
df['Installs'] = df['Installs'].map(lambda x: x.replace(',',''))
```

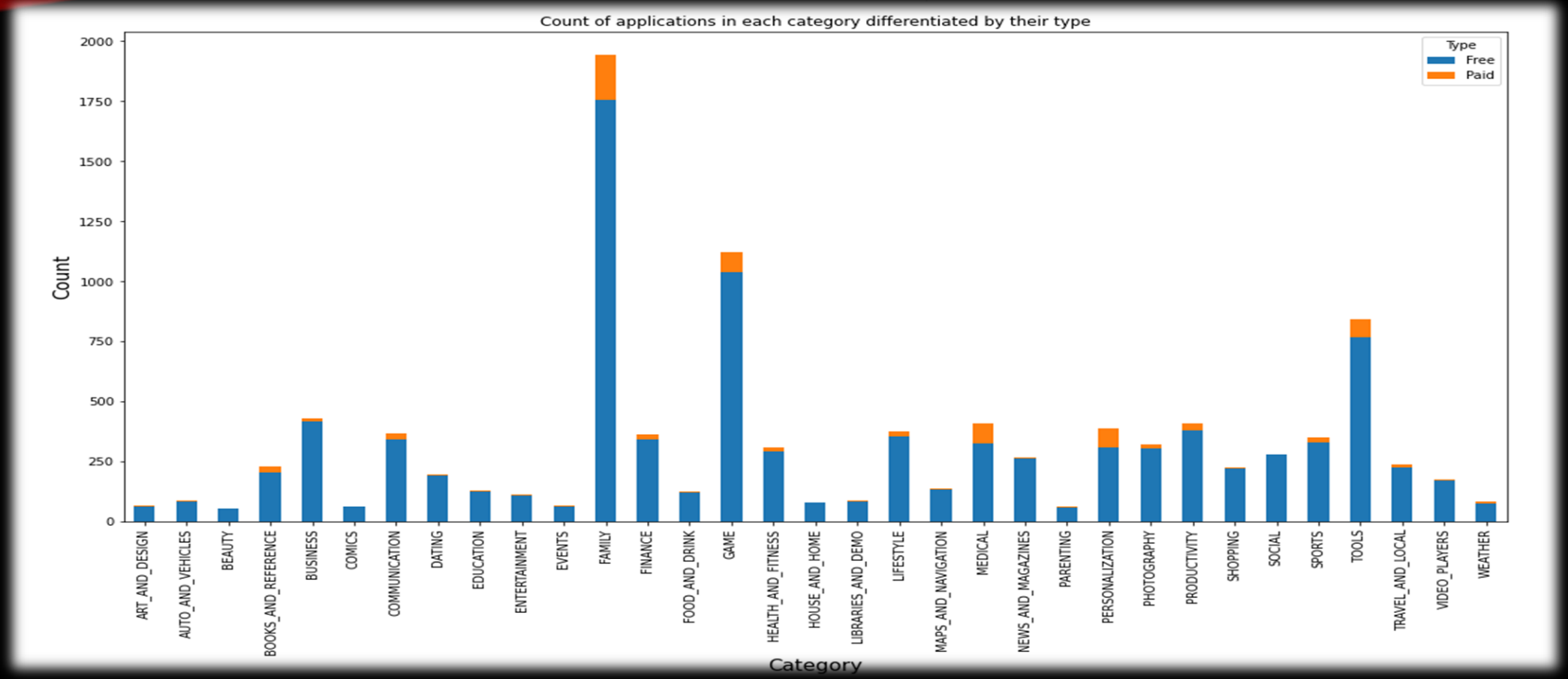
CLEAN DATA COLUMNS

No we clean data columns of installs, Remove + in installs and remove coma

EXPLORATORY ANALYSIS AND VISUALIZATION

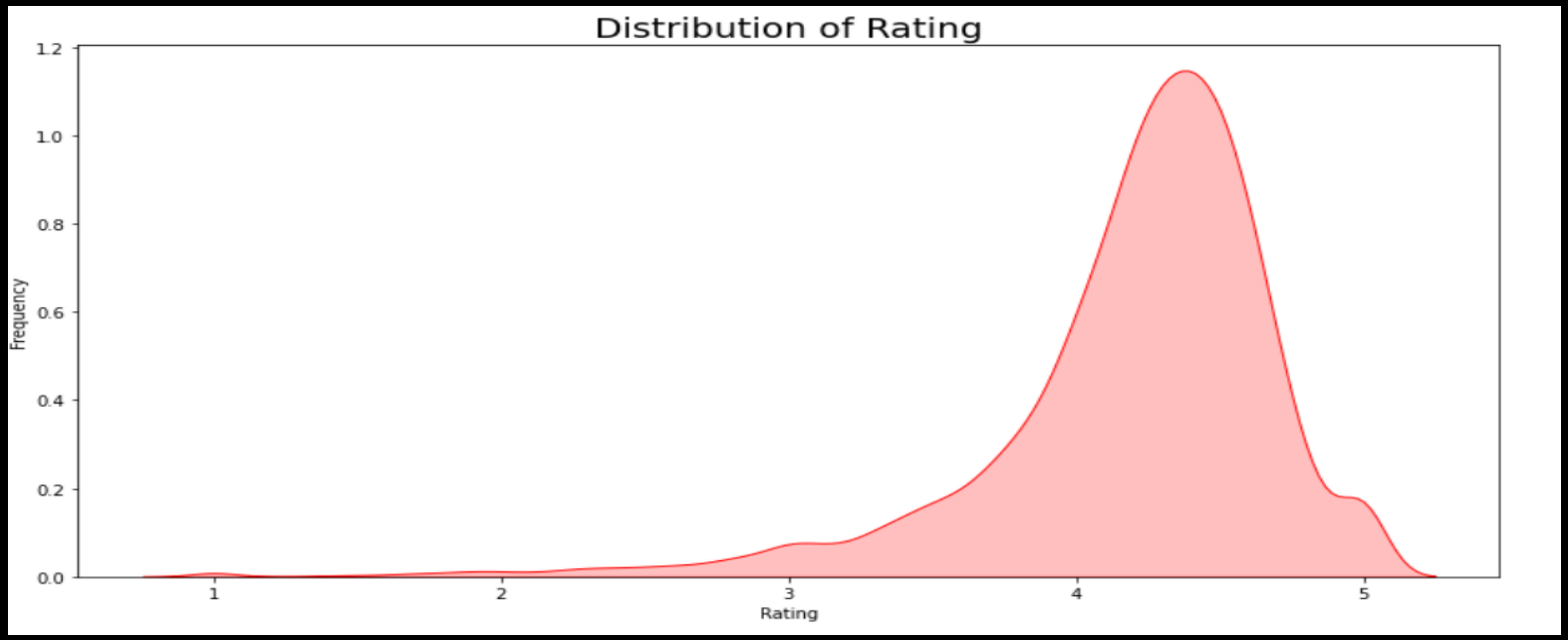
EXPLORATORY DATA VISUALIZATIONS ARE THE TYPE OF VISUALIZATIONS WE ASSEMBLE WHEN WE DO NOT HAVE A CLUE ABOUT WHAT INFORMATION LIES WITHIN OUR DATASET.

Count of Applications in each category differentiated by their type



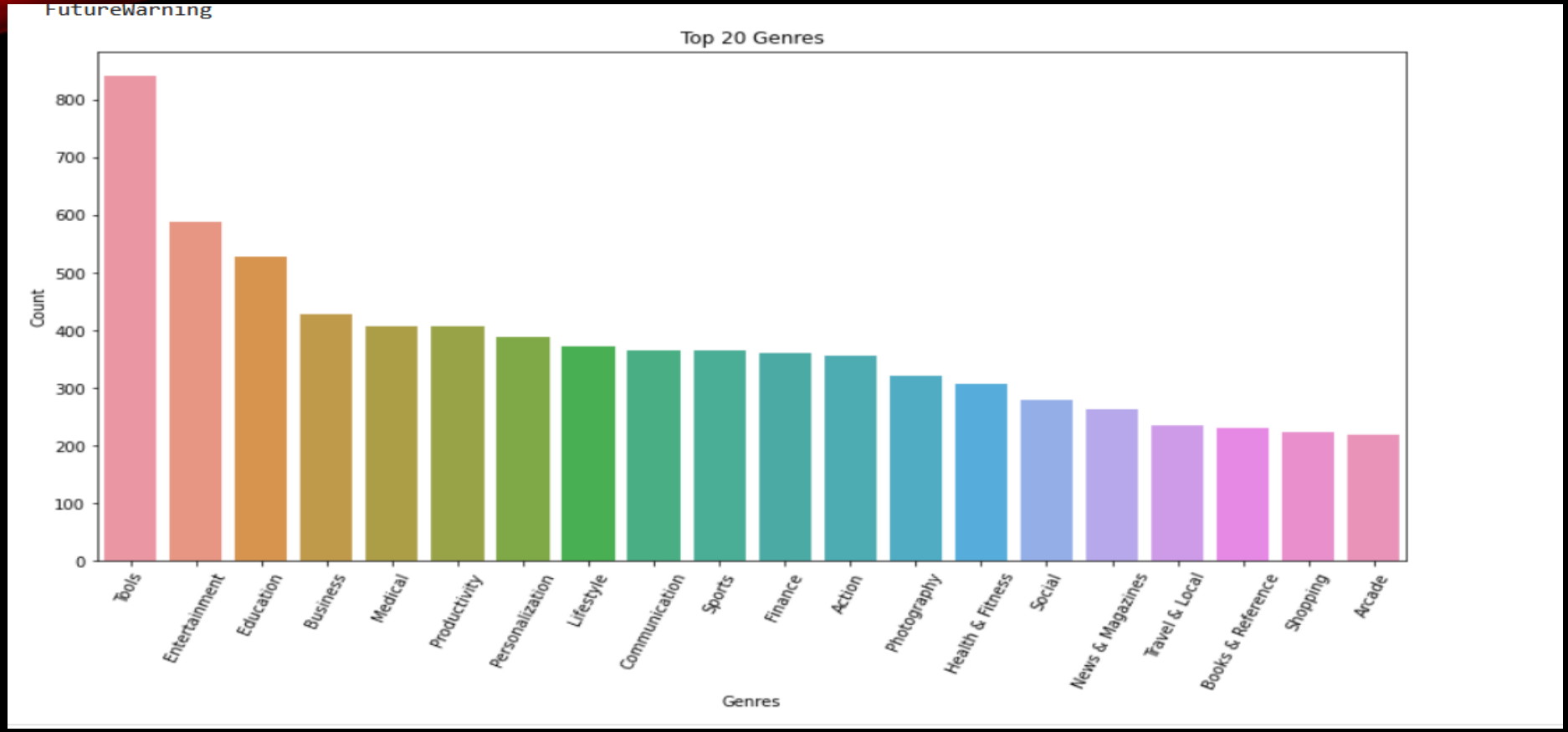
It looks like certain app categories have more free apps available for download than others. In our dataset, the majority of apps in Family, Games and Tools, as well as Social categories were free to install. At the same time Family, Personalization and Medical categories had the biggest number of paid apps available for download.

DISTRIBUTION OF APP RATING



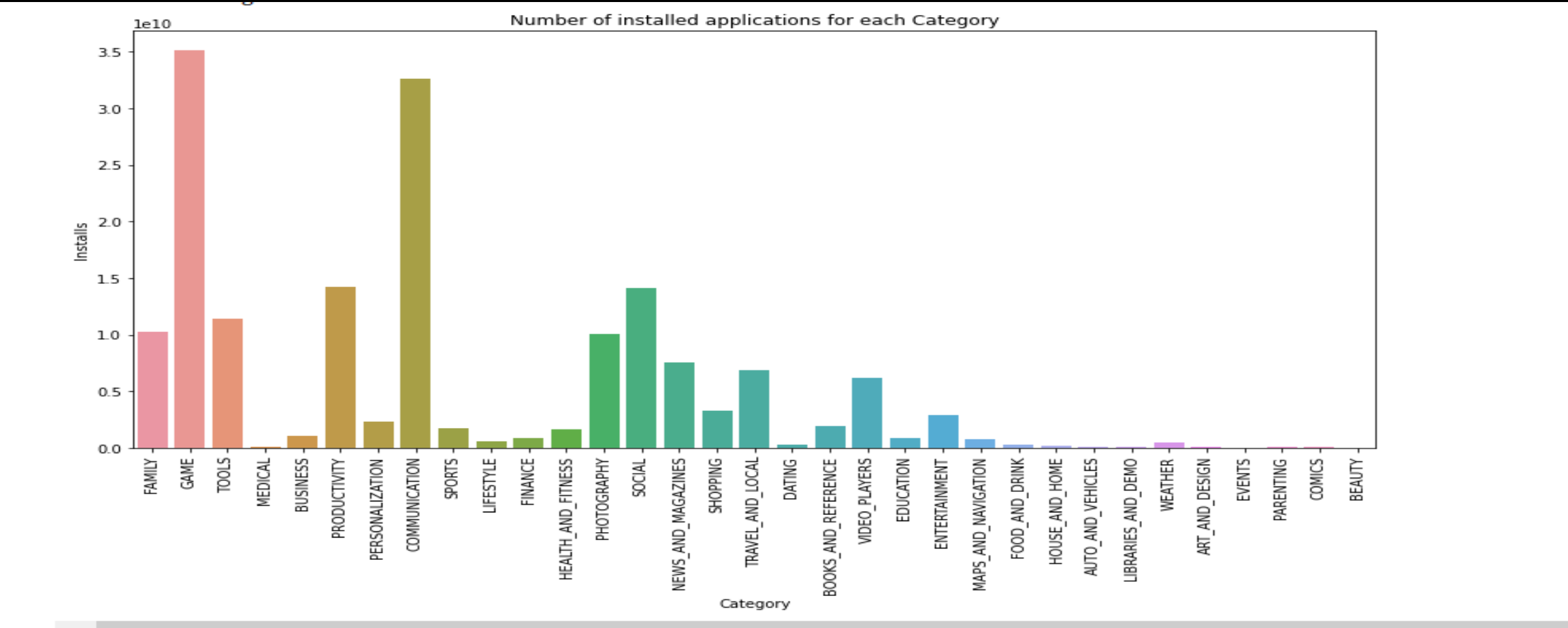
Average rating of application in store is around 4.3, which is very high. This plot can be used to look whether the original ratings of the app matches the predicted rating to know whether the app is performing better or worse compared to other apps on the Play Store.

TOP 20 GENRES

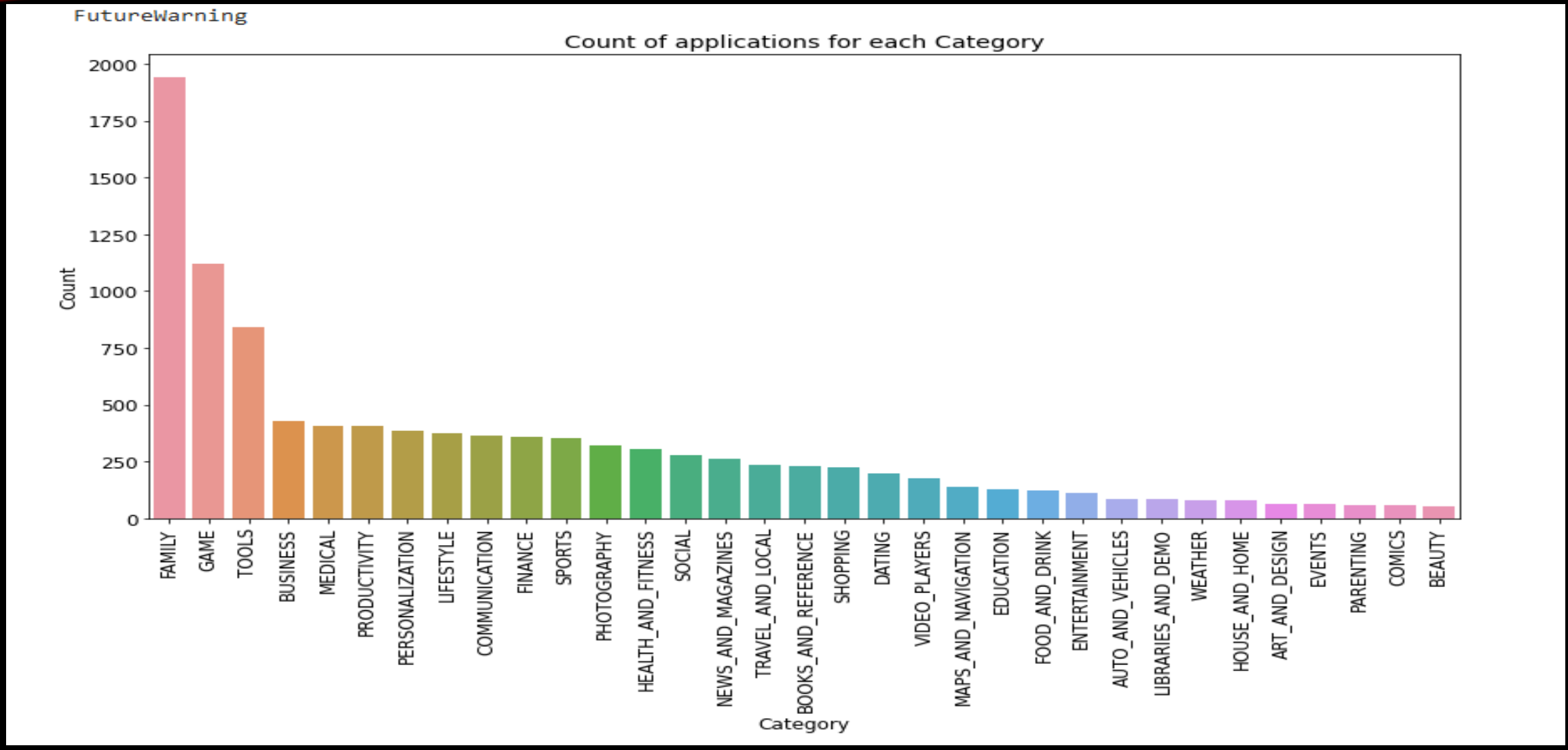


As well as we can sea the tools and entertainment genres have most of count.

NUMBER OF INSTALLED APPLICATIONS FOR EACH CATEGORY



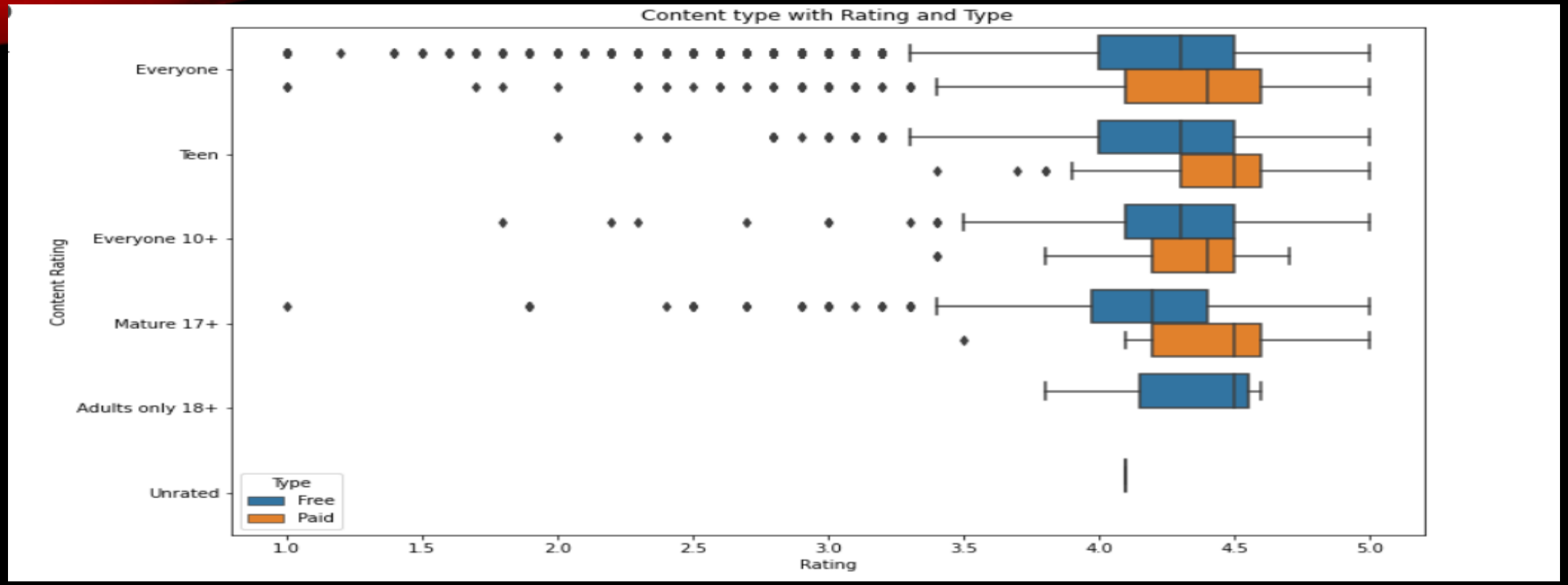
COUNT OF APPLICATIONS FOR EACH CATEGORY



NUMBER OF INSTALLED APPLICATIONS FOR EACH CATEGORY

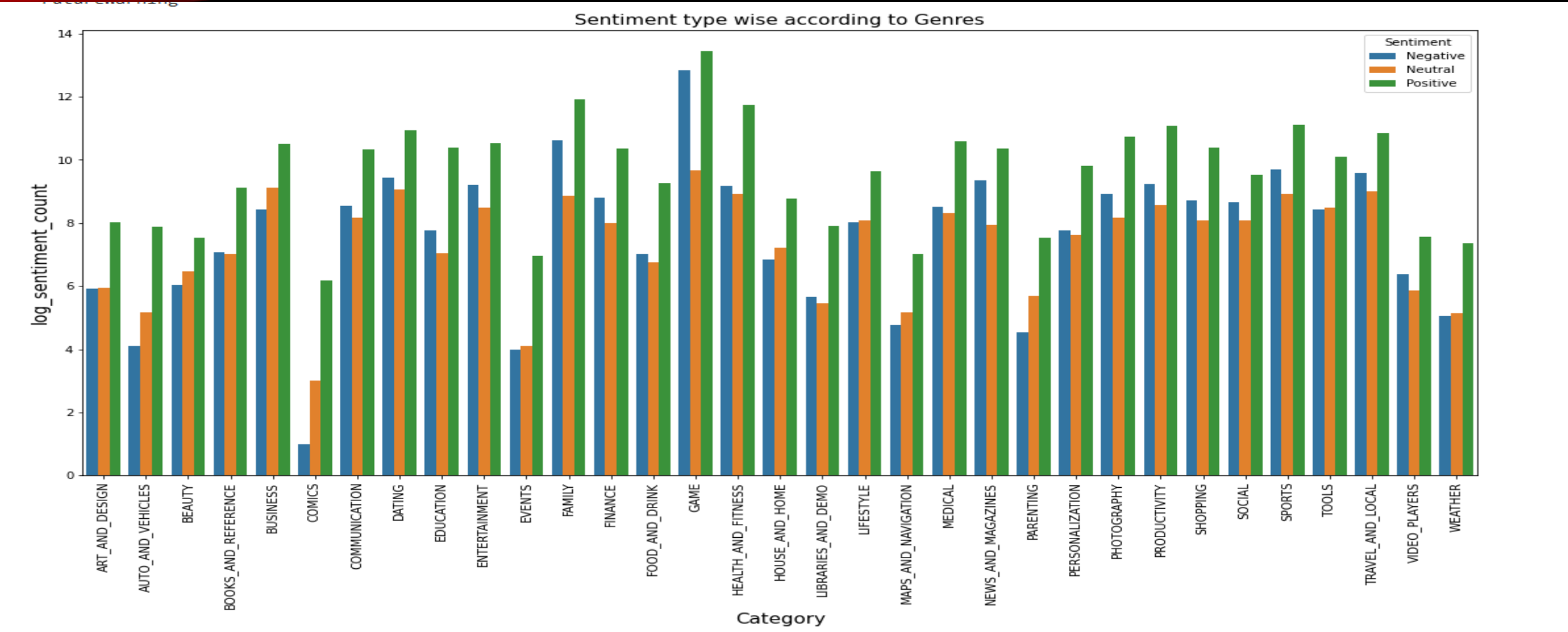
From the above two plots we can conclude that, maximum number of apps present in google play store comes under Family, Games and Tools Category but as per the installations and requirements in the market place, this is not the case. Maximum installed apps comes under Games, Communication and Tools.

CONTENT TYPE HAVE MOST RETING AND TYPE



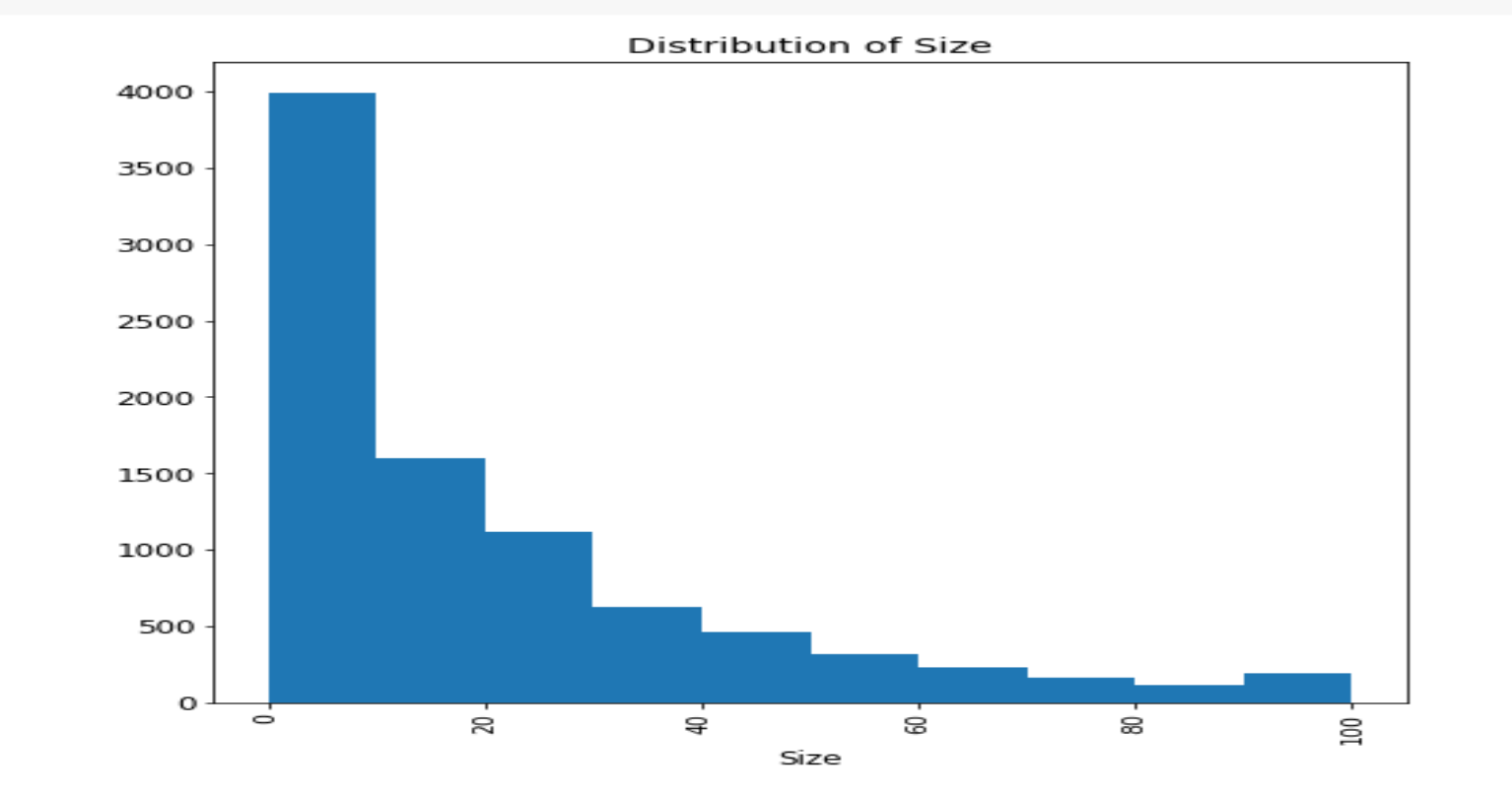
Now we can see the everyone content which have free type that's have most of rating .

DISTRIBUTION OF TYPE OF REVIEWS, CATEGORY WISE IN THE DATASET



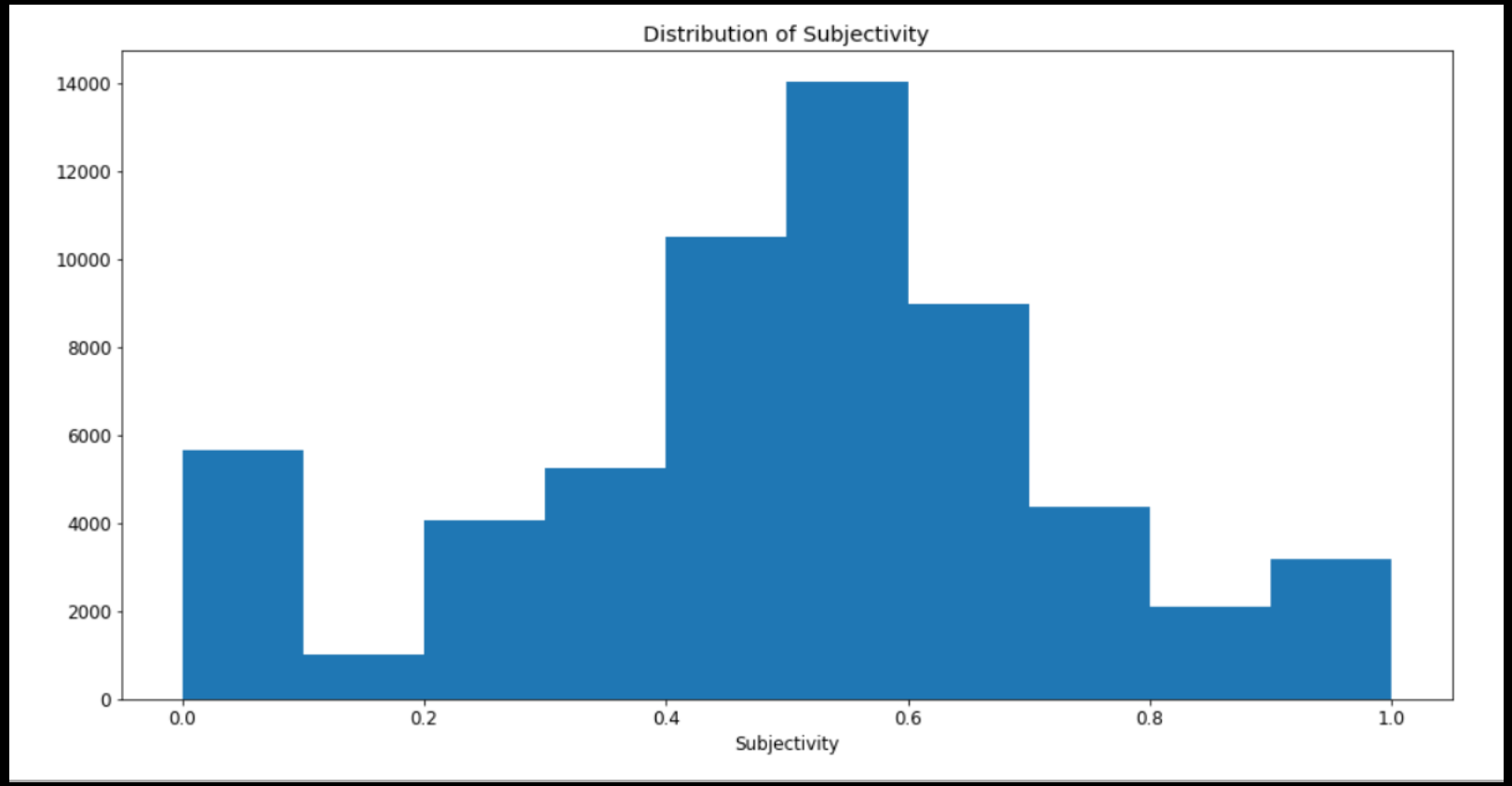
It can be seen from the above plot that the number of positive reviews are way higher than negetive and neutral ones.seen.

DISTRIBUTION OF APP SIZE



It can be seen from the above plot that maximum application's size lies between 0–10 MB.

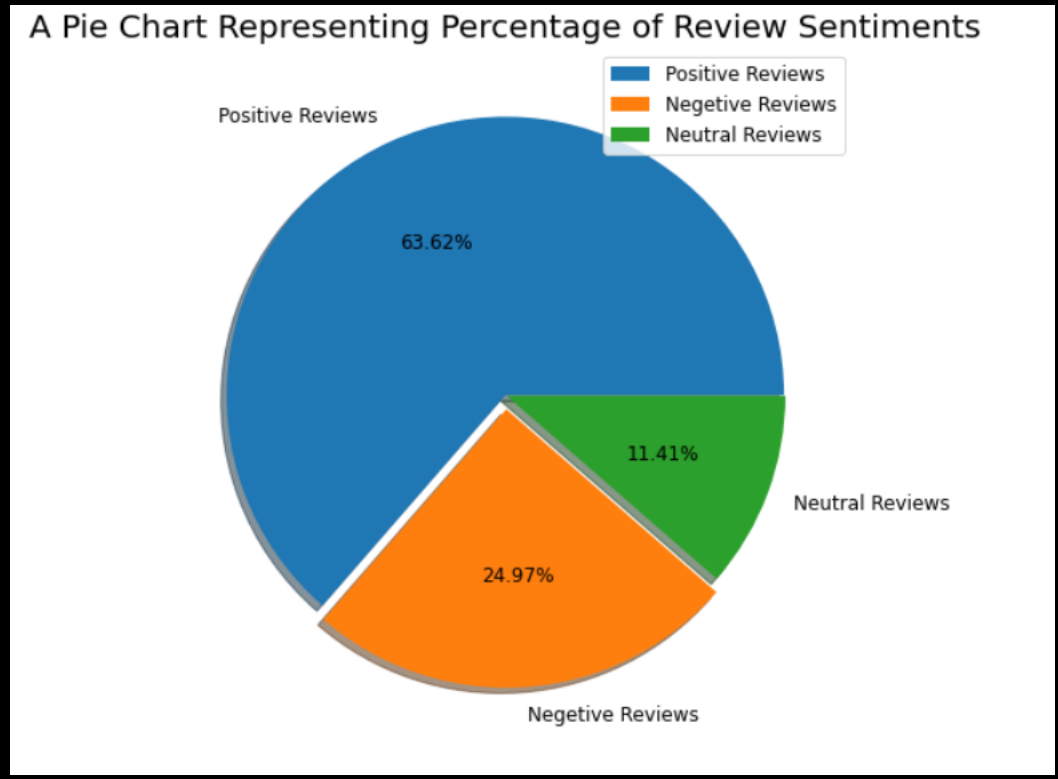
DISTRIBUTION OF SUBJECTIVITY



It can be seen that maximum number 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.

Representing Percentage of Review Sentiments

It can be seen from the above plot that the number of positive reviews are way higher than negative and neutral ones.



CONCLUSION

- The dataset contains possibilities to deliver insights to understand customer demands better and thus help developers to popularize the product. Dataset can also be used to look whether the original ratings of the app matches the predicted rating to know whether the app is performing better or worse compared to other apps on the Play Store.

THANK YOU