EDA Capstone Project

Play Store App Review

Team Data Avengers

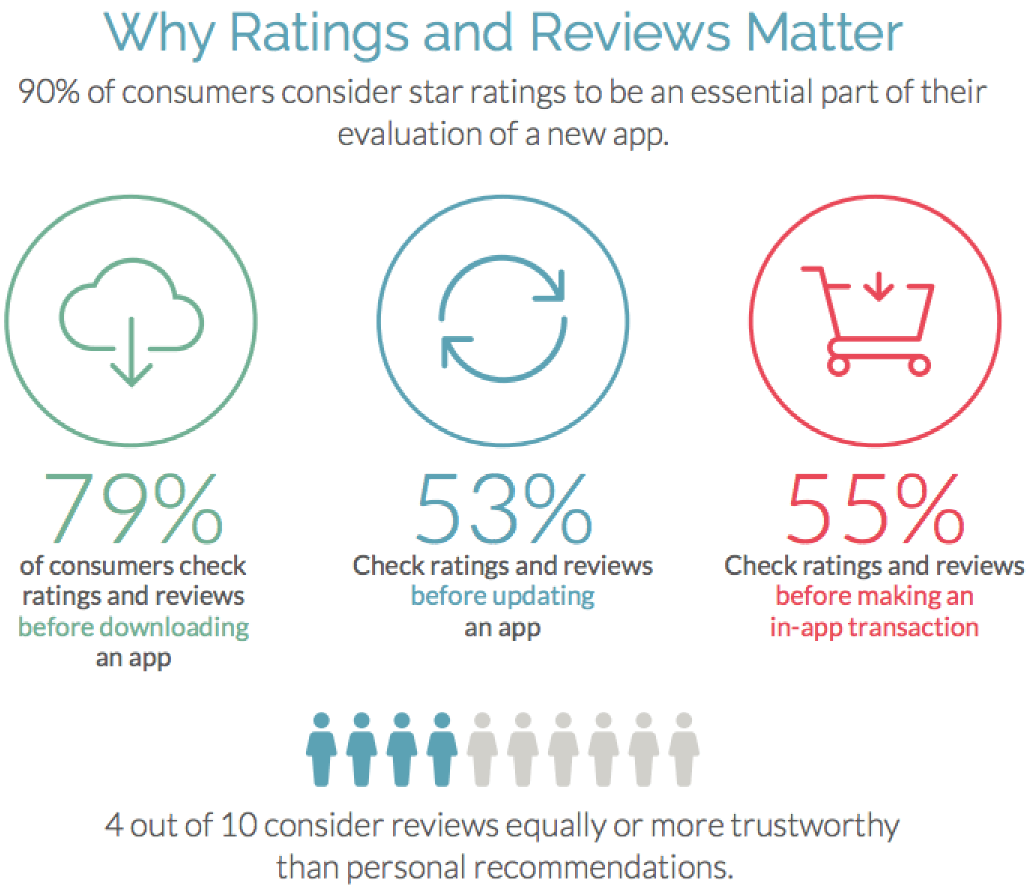
Ganeshkumar Patel, Akanksha Agarwal
Rahul Mishra, Shivam Singh
Dashrath Singh

Content

- Introduction
- Problem Comprehension
- Dataset Description
- Data Exploration
- Data Cleaning
- Extracting Statistics
- Exploratory & Visualization Analysis
- EDA on Review Dataset
- Conclusion

Introduction

- Mobile applications have enormous potential to drive one's business to reach great heights.
- Thus, it becomes essential to do comprehensive analysis of play store app data and understand the demand of users.



Why Ratings and Reviews Matter

90% of consumers consider star ratings to be an essential part of their evaluation of a new app.

79% of consumers check ratings and reviews before downloading an app

53% Check ratings and reviews before updating an app

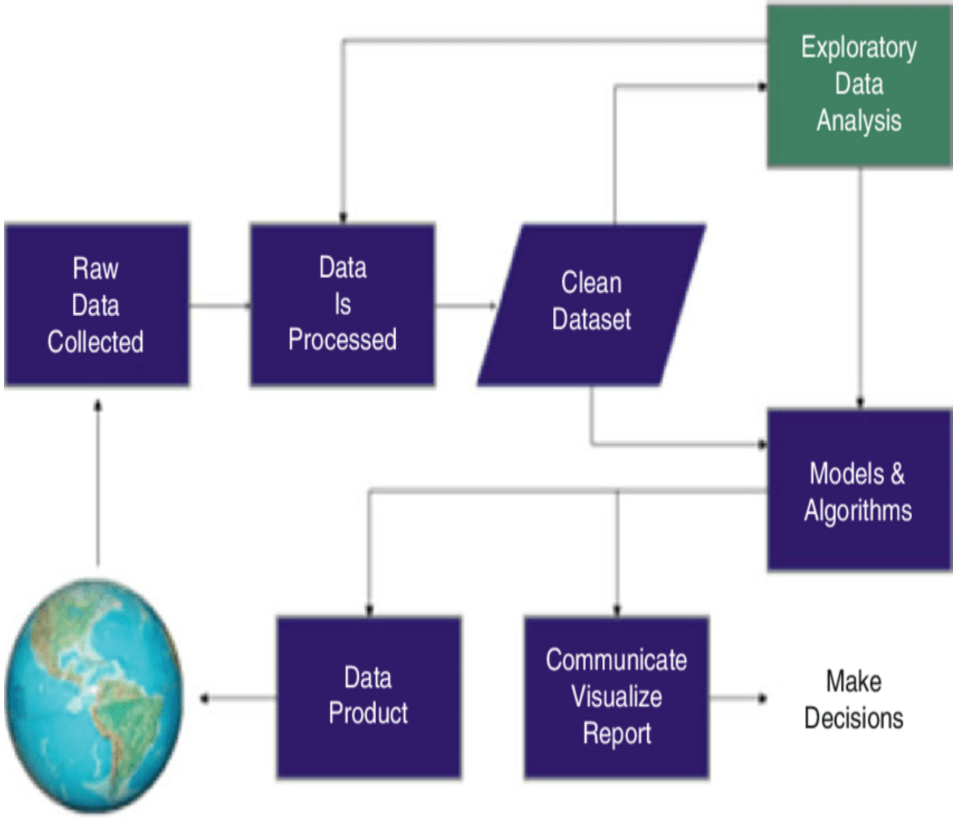
55% Check ratings and reviews before making an in-app transaction

4 out of 10 consider reviews equally or more trustworthy than personal recommendations.

Introduction

Discussion on Google play store dataset involves various steps such as:

- Problem Comprehension
- Loading the data into data frame
- Cleaning the data
- Extracting statistics from the dataset
- Exploratory analysis and visualizations



Problem Comprehension

The key performance indicators for app engagement and its success are:

- **Rating:** Overall user rating of the application. It lies between 1 to 5.
- **Reviews:** Total number of users reviews each app has received.
- **Size:** The memory size needed to install the application.
- **Installs:** Number of installs- An install takes place when a user has downloaded an app and successfully opens it for the first time.
- **Price:** The price of the app.

This analysis is performed to know, the categories, the genres which is most appealing to users and to see how does the current users react to certain parameters such as size of apps and their price.

Importing Libraries

Imported relevant libraries to perform the analysis on give datasets such as:

- Numpy
- Pandas
- Matplotlib
- Seaborn

Dataset reference - **AlmaBetter**

https://drive.google.com/drive/folders/1j6esDUtS0hmPddXMEerNjT4X2ol3XbX3?usp=sharing

Dataset Description

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.

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.

Dataset Description

B. Review 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 orignal 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. In this, the text in each review has been pre-processed and attributed with three new features.

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 Cleaning

It is a crucial step which includes:

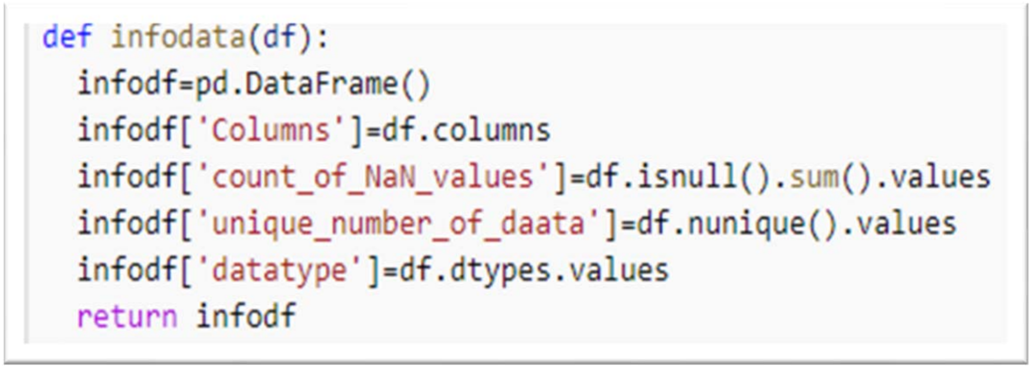
- Dropping Duplicates
- Finding NaN, Nulls and missing values
- Validating data to a standard pattern.

The three features that we will be working with most frequently henceforth are Installs, Size, and Price. A careful glance of the dataset reveals that some of these columns mandate data cleaning. Specifically, the presence of special characters (, $ +) and letters (M k) in the Installs, Size, and Price columns are making their conversion to a numerical data type difficult. Let's clean by removing these and converting each column to a numeric type

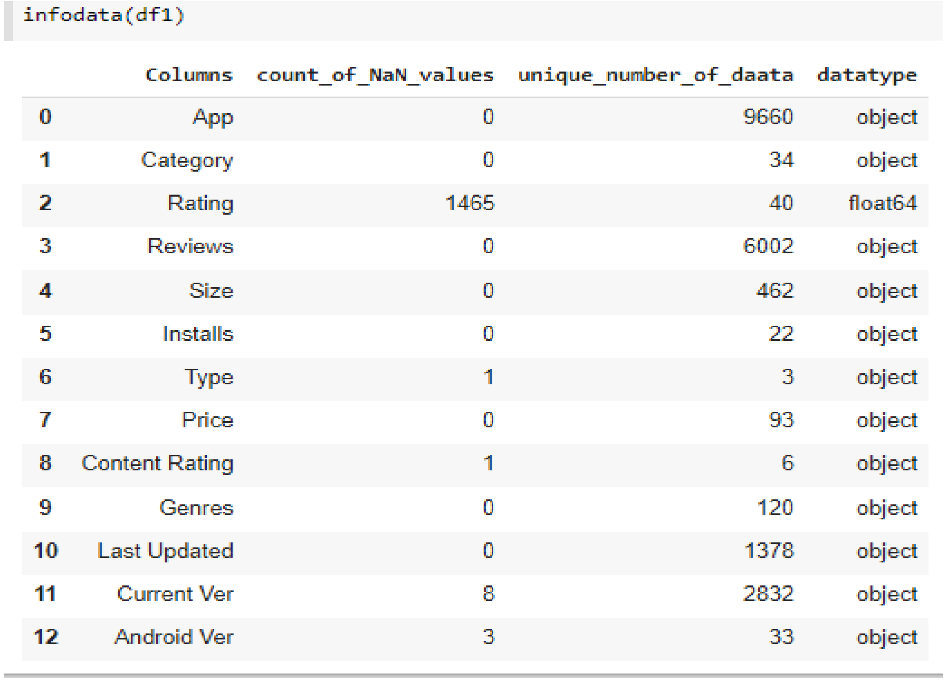
Data Cleaning

Steps which we have followed to make data Clean :

- **Dropping duplicate values.**
- **Define a function to get more info from any dataset at once.**



```
def infodata(df):
  infodf=pd.DataFrame()
  infodf['Columns']=df.columns
  infodf['count_of_NaN_values']=df.isnull().sum().values
  infodf['unique_number_of_daata']=df.nunique().values
  infodf['datatype']=df.dtypes.values
  return infodf
```

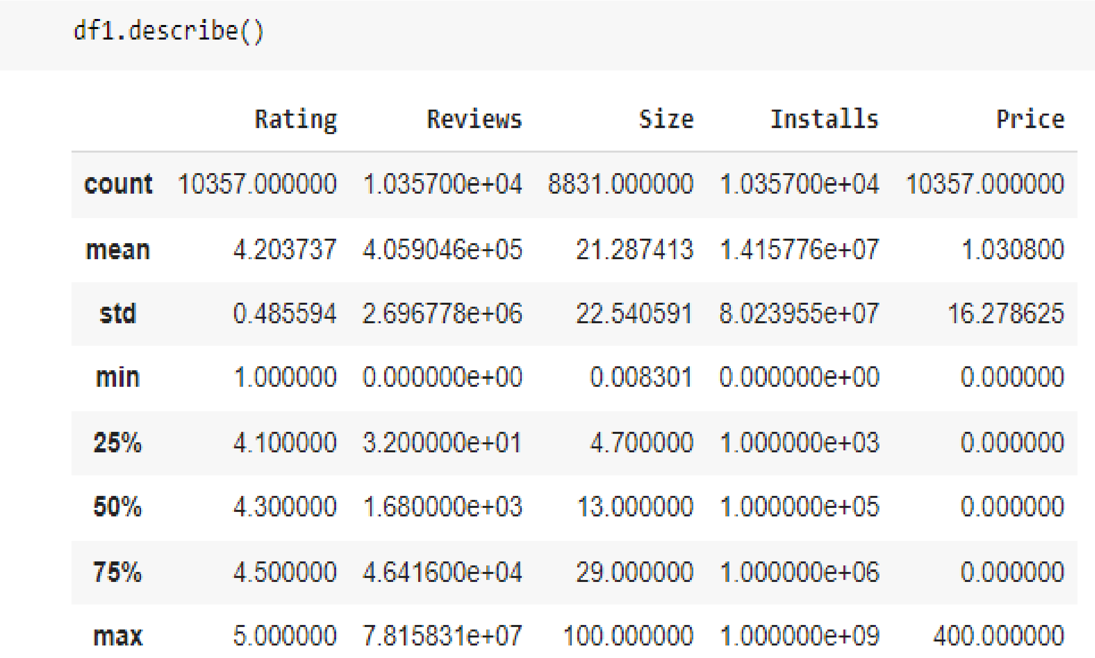


```
infodata(df1)
```

	Columns	count_of_NaN_values	unique_number_of_daata	datatype
0	App	0	9660	object
1	Category	0	34	object
2	Rating	1465	40	float64
3	Reviews	0	6002	object
4	Size	0	462	object
5	Installs	0	22	object
6	Type	1	3	object
7	Price	0	93	object
8	Content Rating	1	6	object
9	Genres	0	120	object
10	Last Updated	0	1378	object
11	Current Ver	8	2832	object
12	Android Ver	3	33	object

Statistical Analysis

Getting statistical values of numerical columns

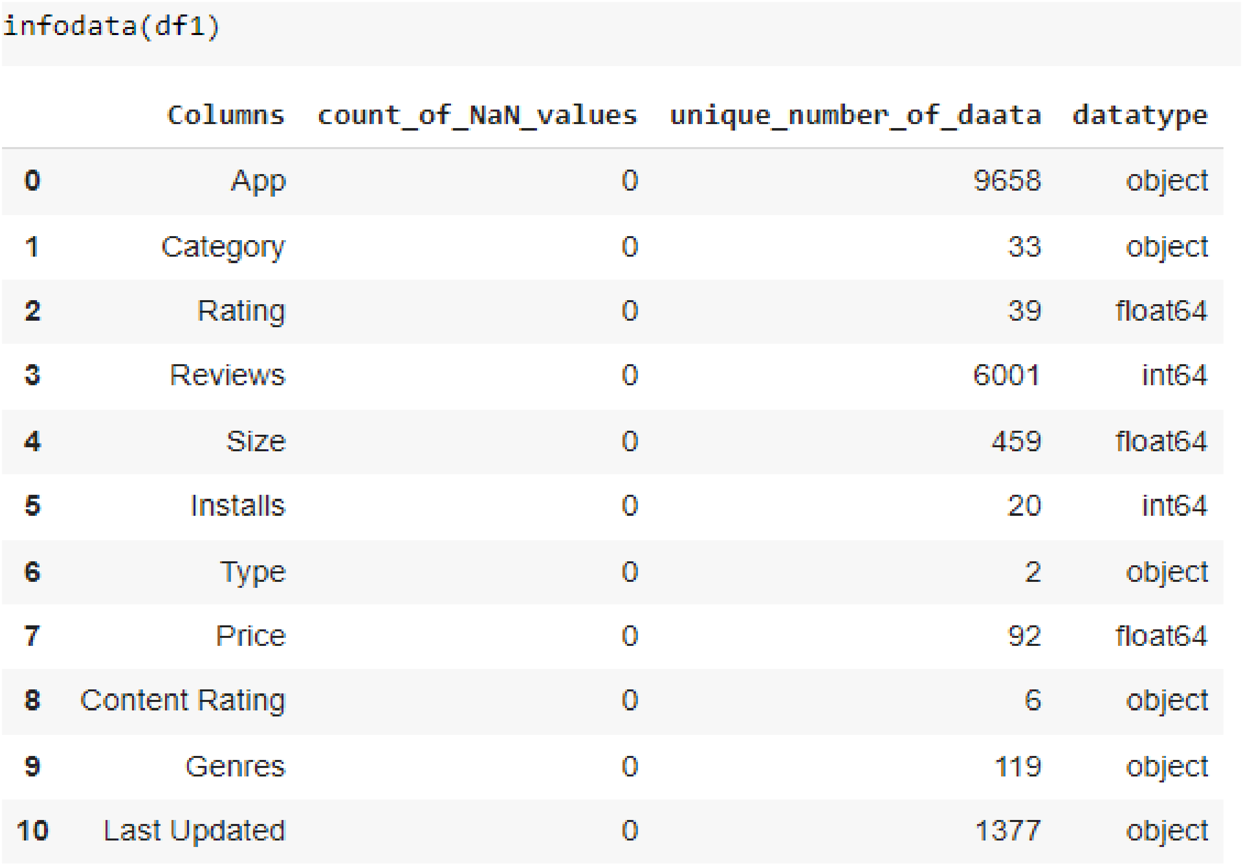


```
df1.describe()
```

	Rating	Reviews	Size	Installs	Price
count	10357.000000	1.035700e+04	8831.000000	1.035700e+04	10357.000000
mean	4.203737	4.059046e+05	21.287413	1.415776e+07	1.030800
std	0.485594	2.696778e+06	22.540591	8.023955e+07	16.278625
min	1.000000	0.000000e+00	0.008301	0.000000e+00	0.000000
25%	4.100000	3.200000e+01	4.700000	1.000000e+03	0.000000
50%	4.300000	1.680000e+03	13.000000	1.000000e+05	0.000000
75%	4.500000	4.641600e+04	29.000000	1.000000e+06	0.000000
max	5.000000	7.815831e+07	100.000000	1.000000e+09	400.000000

Now, we can see we have all five columns with numerical datatype.

First milestone reached !!



```
infodata(df1)
```

	Columns	count_of_NaN_values	unique_number_of_daata	datatype
0	App	0	9658	object
1	Category	0	33	object
2	Rating	0	39	float64
3	Reviews	0	6001	int64
4	Size	0	459	float64
5	Installs	0	20	int64
6	Type	0	2	object
7	Price	0	92	float64
8	Content Rating	0	6	object
9	Genres	0	119	object
10	Last Updated	0	1377	object

Data Cleaning Completed .

Exploratory Data Analysis

We will perform analysis through our five **KPIs (Key Performance Indicators)** i.e **'Rating**', **'Reviews**', **'Size**', **'Installs**' and **'Price**'. The analysis will be based on below method .

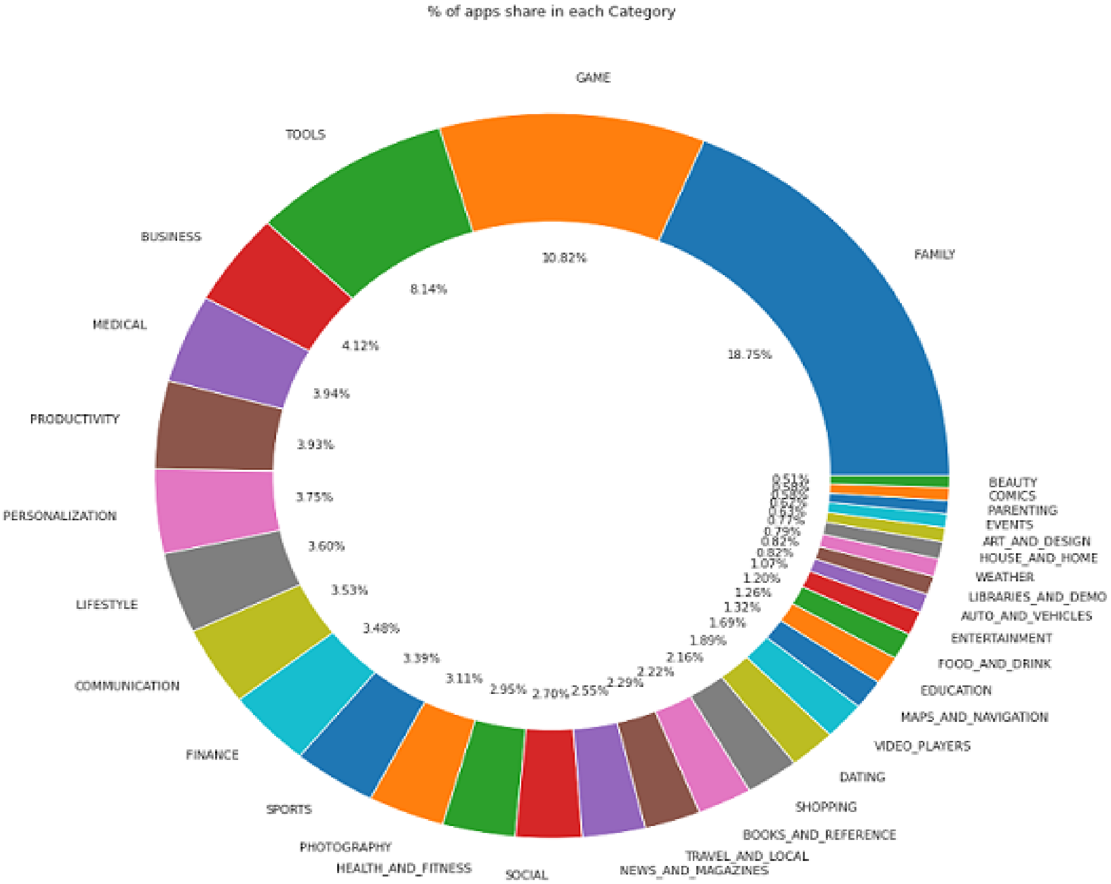
- **Univariate study-** Study of dependent variables such as 'Rating', 'Reviews', 'Size', 'Installs' and 'Price'
- **Multivariate study-** Study of multivariable and correlation between them.

Exploring 'Categories' of Apps

Which top 3 category shares more number of apps?

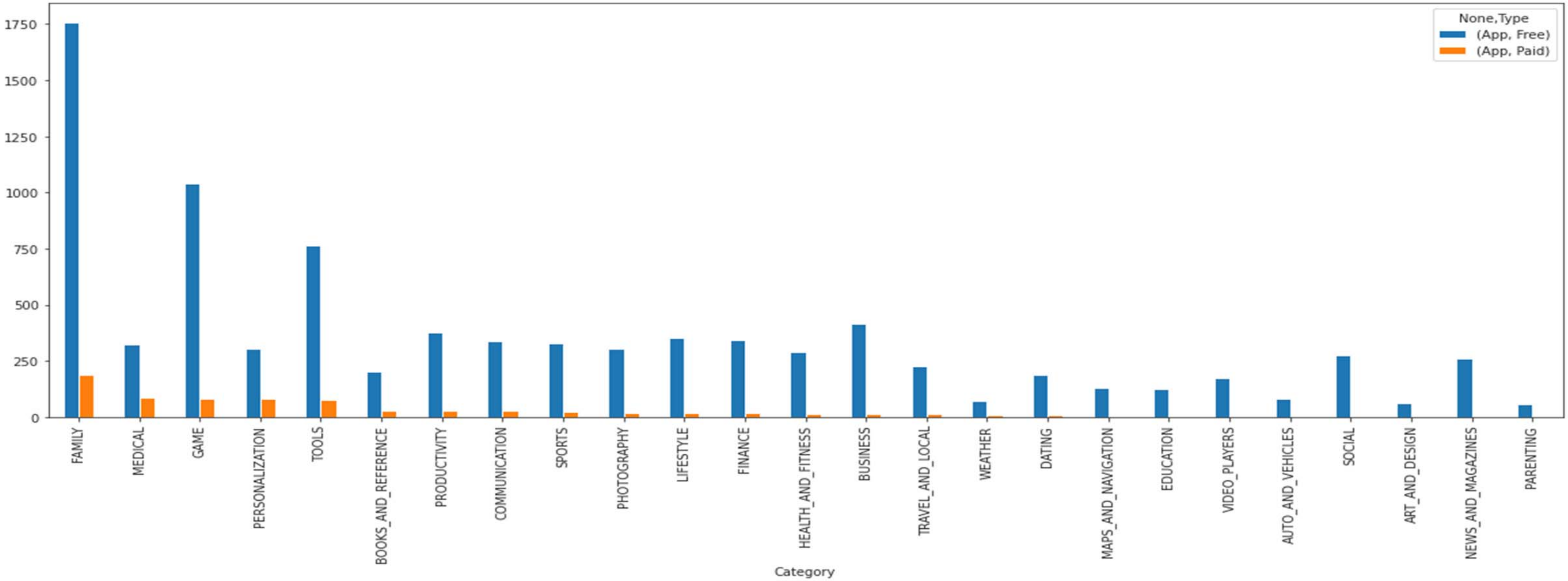
From the doughnut chart we can conclude that there are greater number of apps related to below categories:

- Family- 18.75%
- Game- 10.82%
- Tools-8.14%



Exploring 'Categories' of Apps

Which are the top 25 categories comprising maximum paid and free apps?

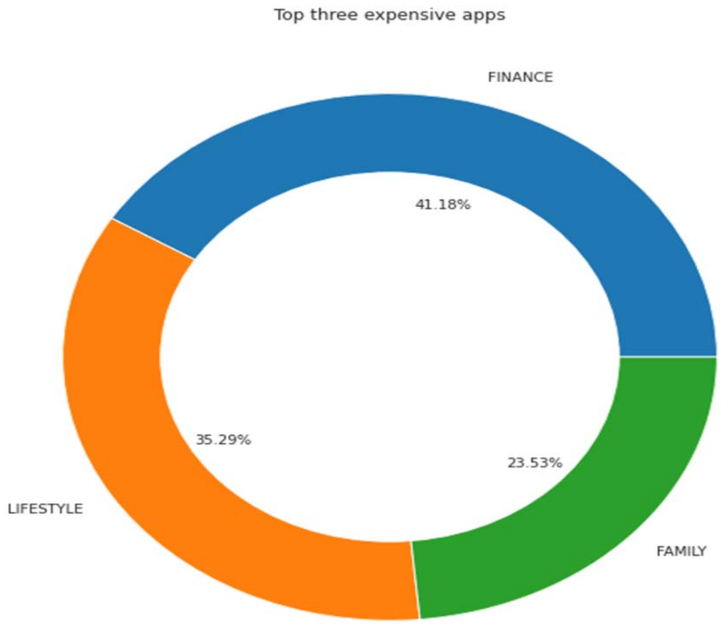


Which are the top three most expensive categories of apps?

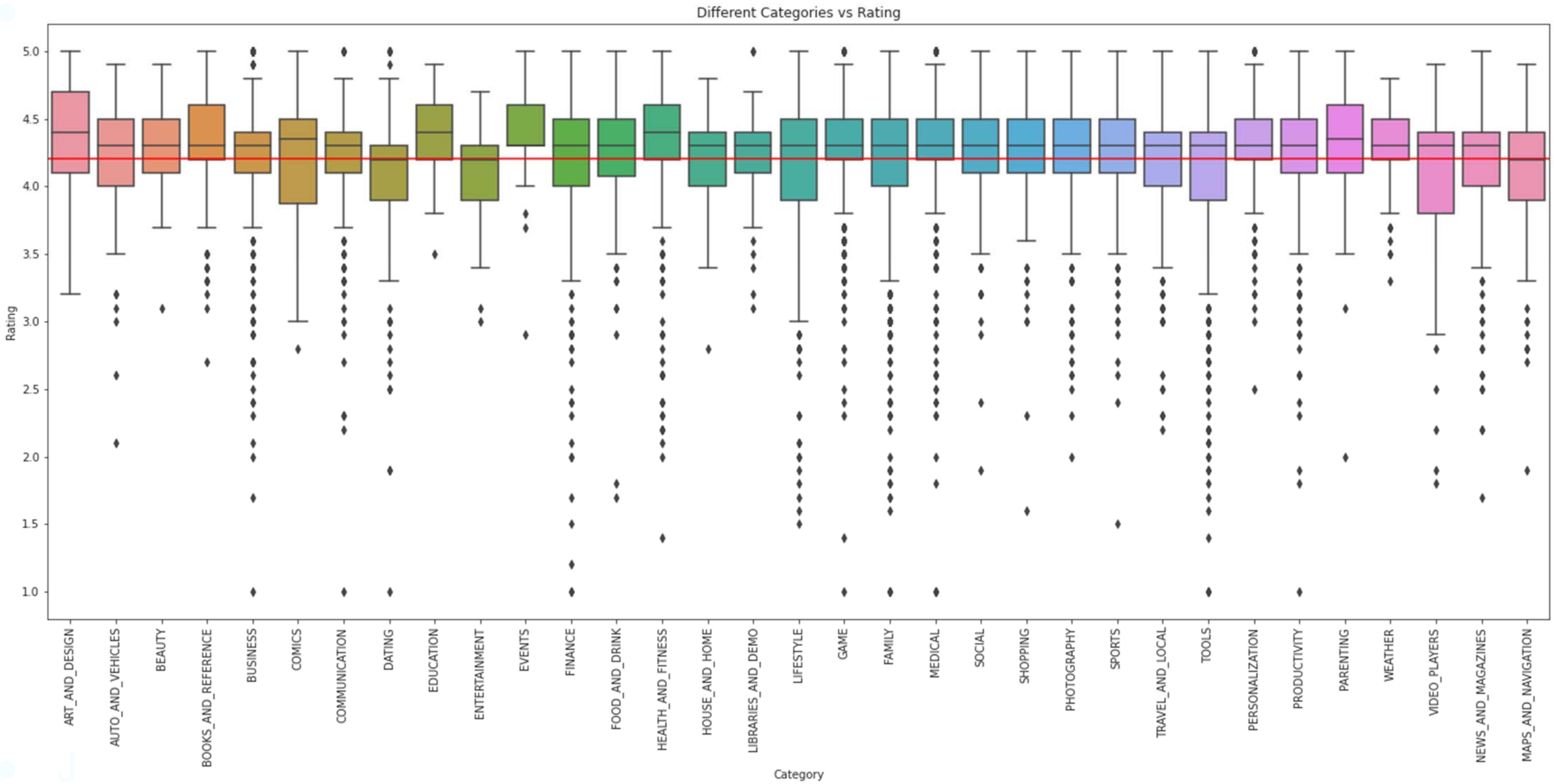
There are many factors to consider when selecting the right pricing strategy for your mobile app. It is important to consider the willingness of your customer to pay for your app. A wrong price could break the deal before the download even happens. Potential customers could be turned off by what they perceive to be a shocking cost, or they might delete an app they've downloaded after receiving too many ads or simply not getting their money's worth.

From the doughnut chart we are able to find top three most expensive categories :

- **Finance**
- **Lifestyle**
- **Family**

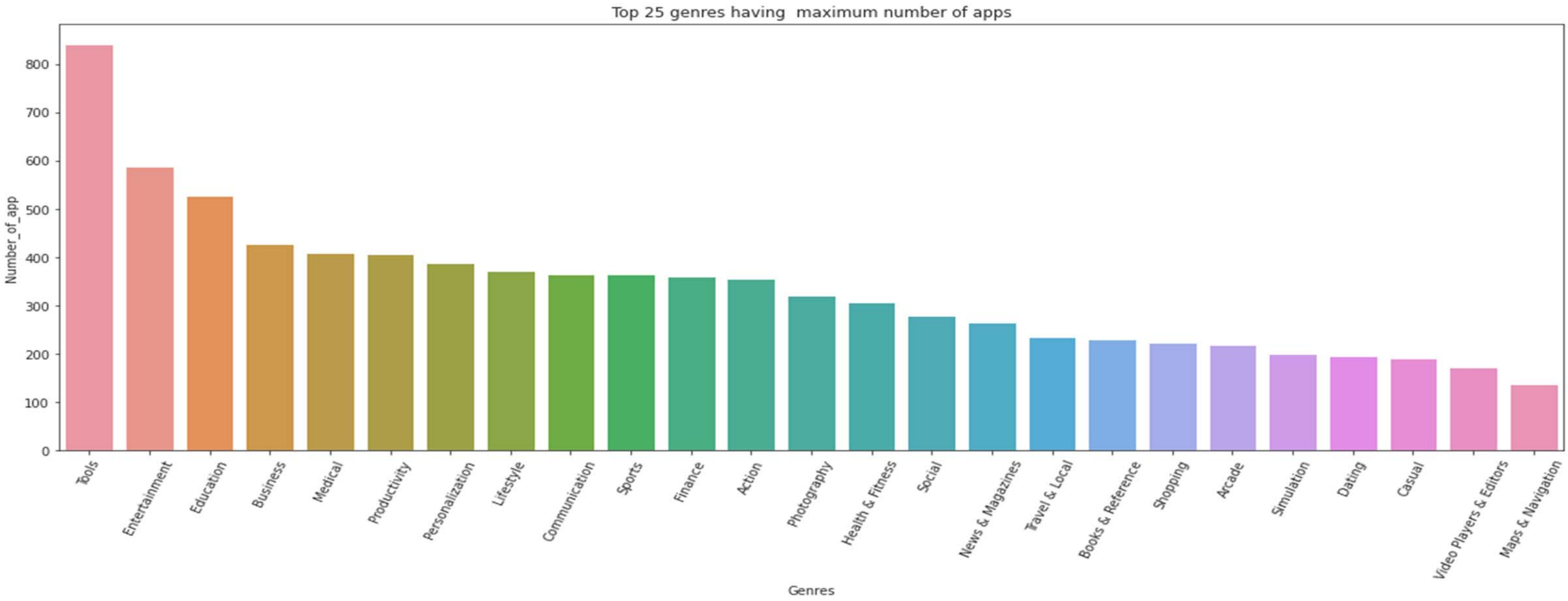


Which categories of apps are less popular?



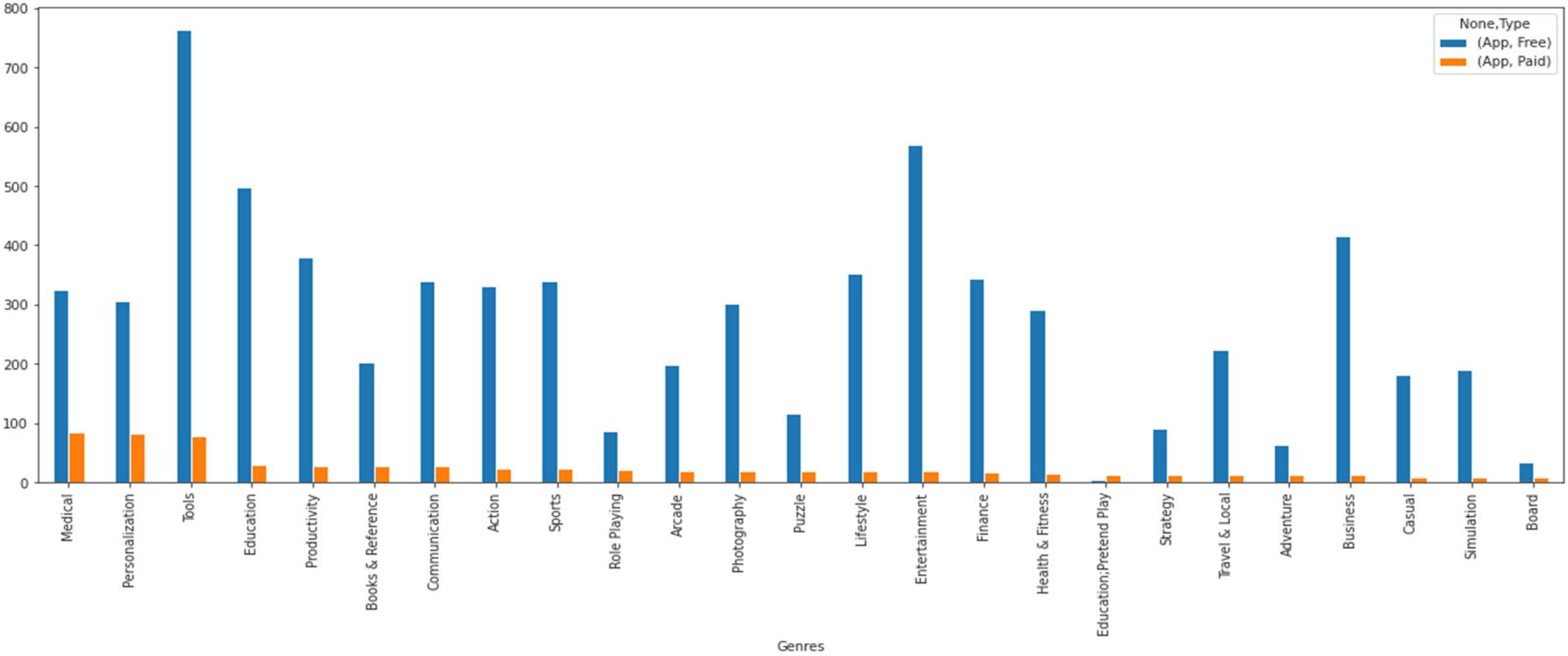
Diving into its Genres

Top 25 genres having maximum number of apps?

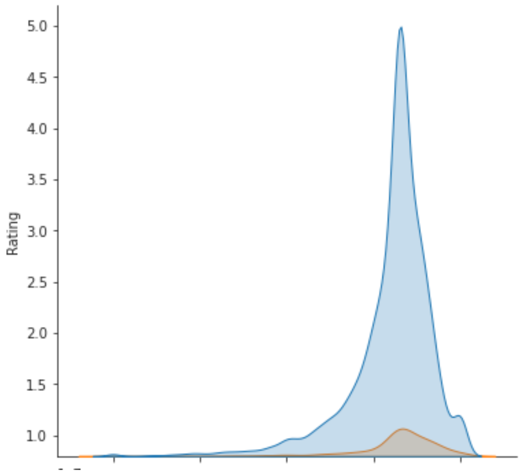


Diving into its Genres

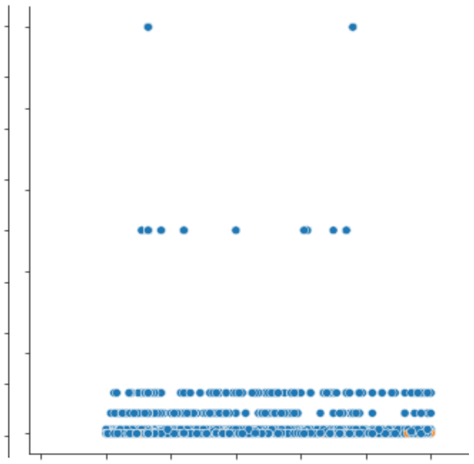
Top 25 Genres having maximum number of paid and free apps



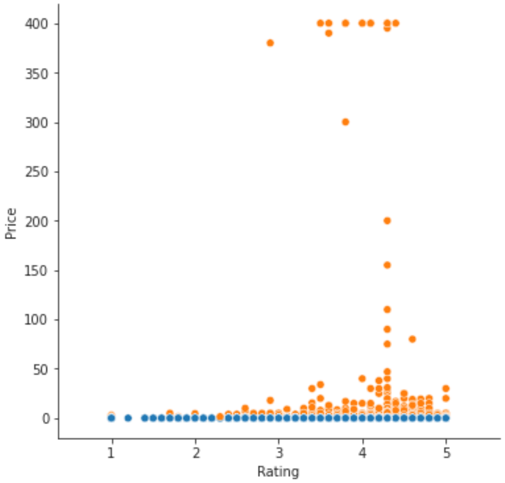
Study of KPIs through Multivariate plots



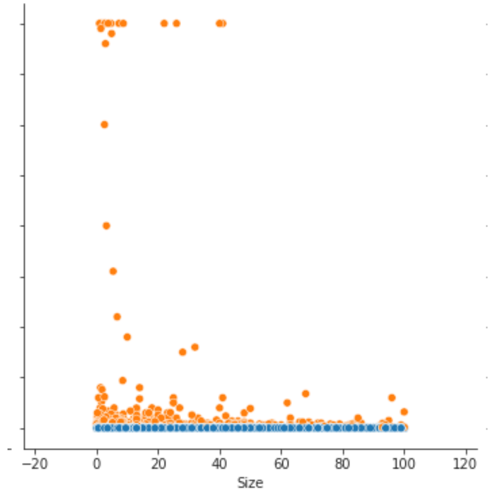
Distribution of Rating
Showing peak, near mode value of Ratings i.e at 4.3.



Size vs Install
The graph sows, that no. of installs does not depend on size



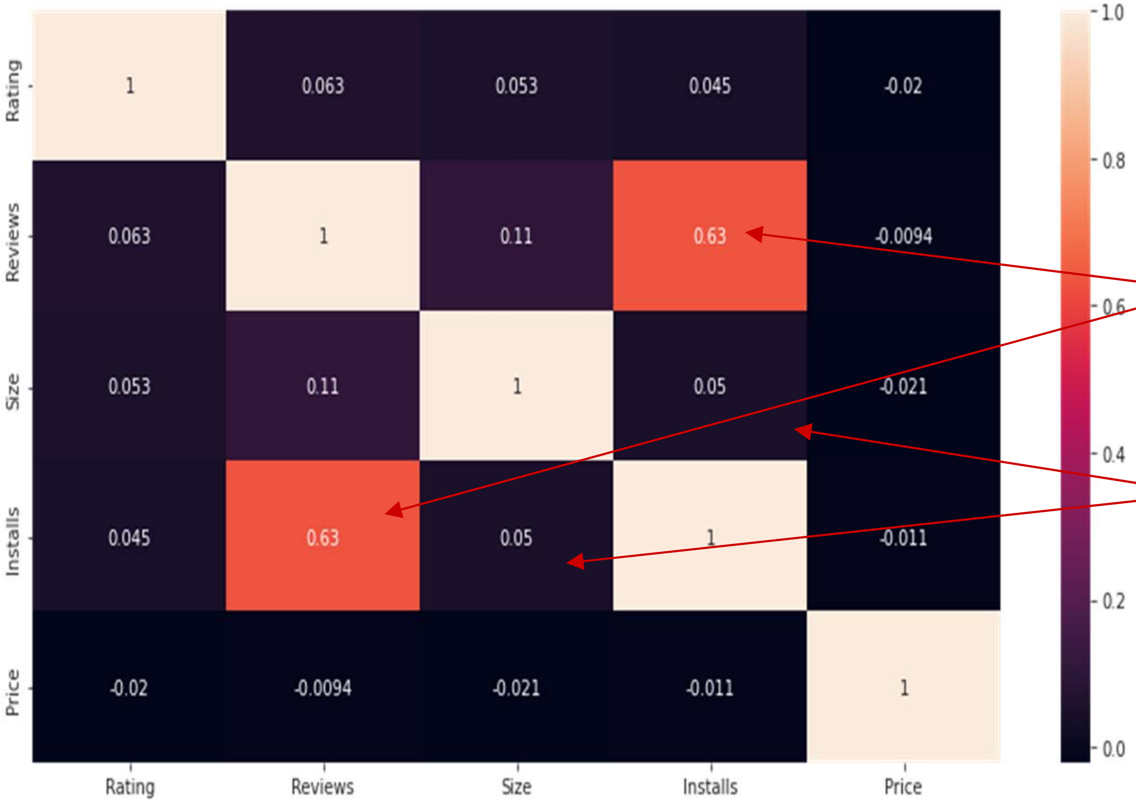
Price vs Rating
Showing that users have paid highly for good rated apps.



Size vs Price
The graph shows, that users have paid more for lite apps

Study of KPIs through Multivariate plots

Correlation between KPIs



	Rating	Reviews	Size	Installs	Price
Rating	1	0.063	0.053	0.045	-0.02
Reviews	0.063	1	0.11	0.63	-0.0094
Size	0.053	0.11	1	0.05	-0.021
Installs	0.045	0.63	0.05	1	-0.011
Price	-0.02	-0.0094	-0.021	-0.011	1

From the heatmap we can conclude:

- Users have installed more apps of the ones having high reviews i.e. installs and reviews are highly corelated.
- Installs of app doesn't depends on Size i.e. there is no correlation between them

EDA on Review Dataset

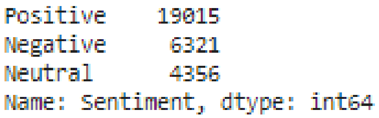
Which type of sentiments users generally prefer to give?

From the doughnut chart we are able to find top three most expensive categories :

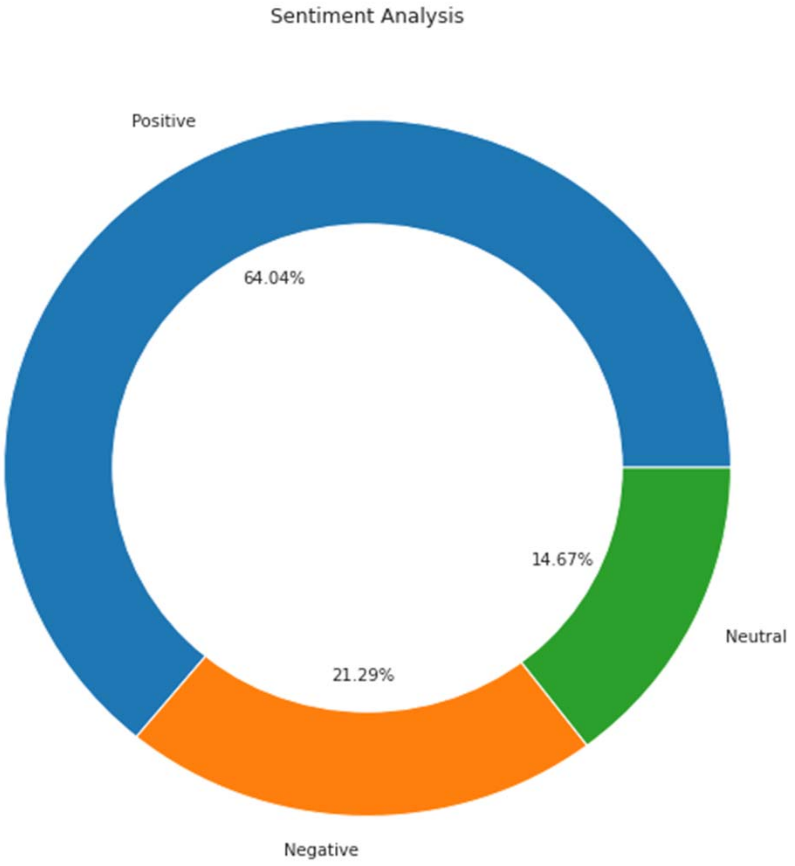
64.04% positive sentiments

21.29% Neutral sentiments

14.67% Negative sentiments

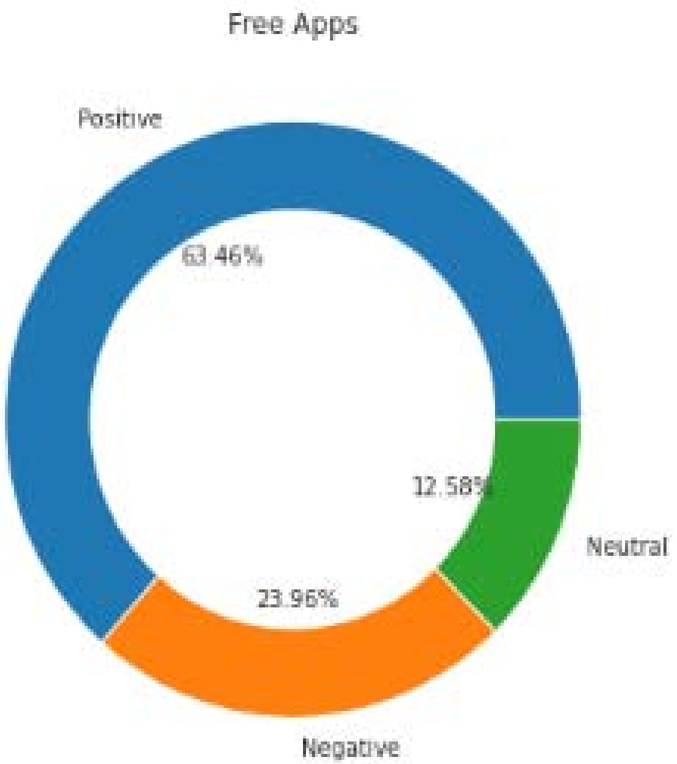


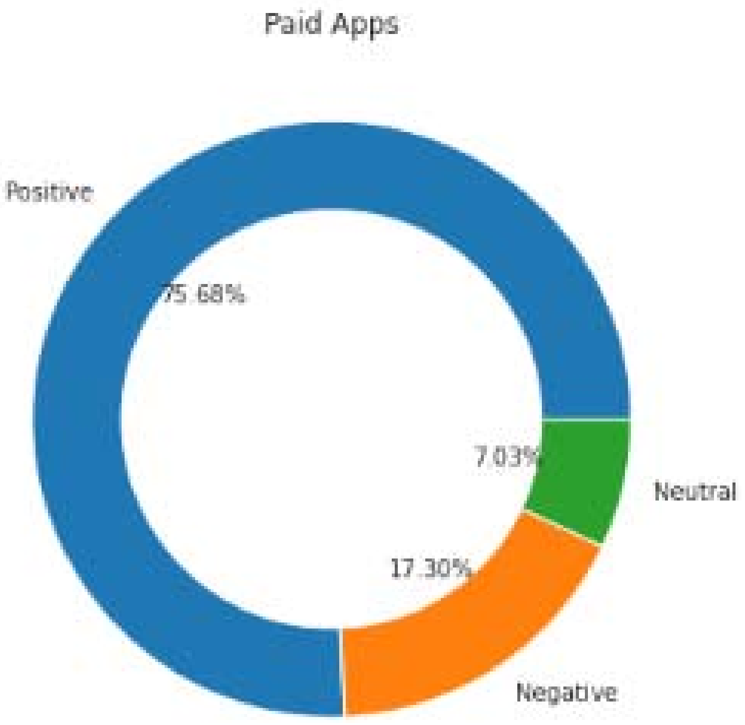
```
Positive    19015
Negative     6321
Neutral      4356
Name: Sentiment, dtype: int64
```



Exploratory Data Visualization

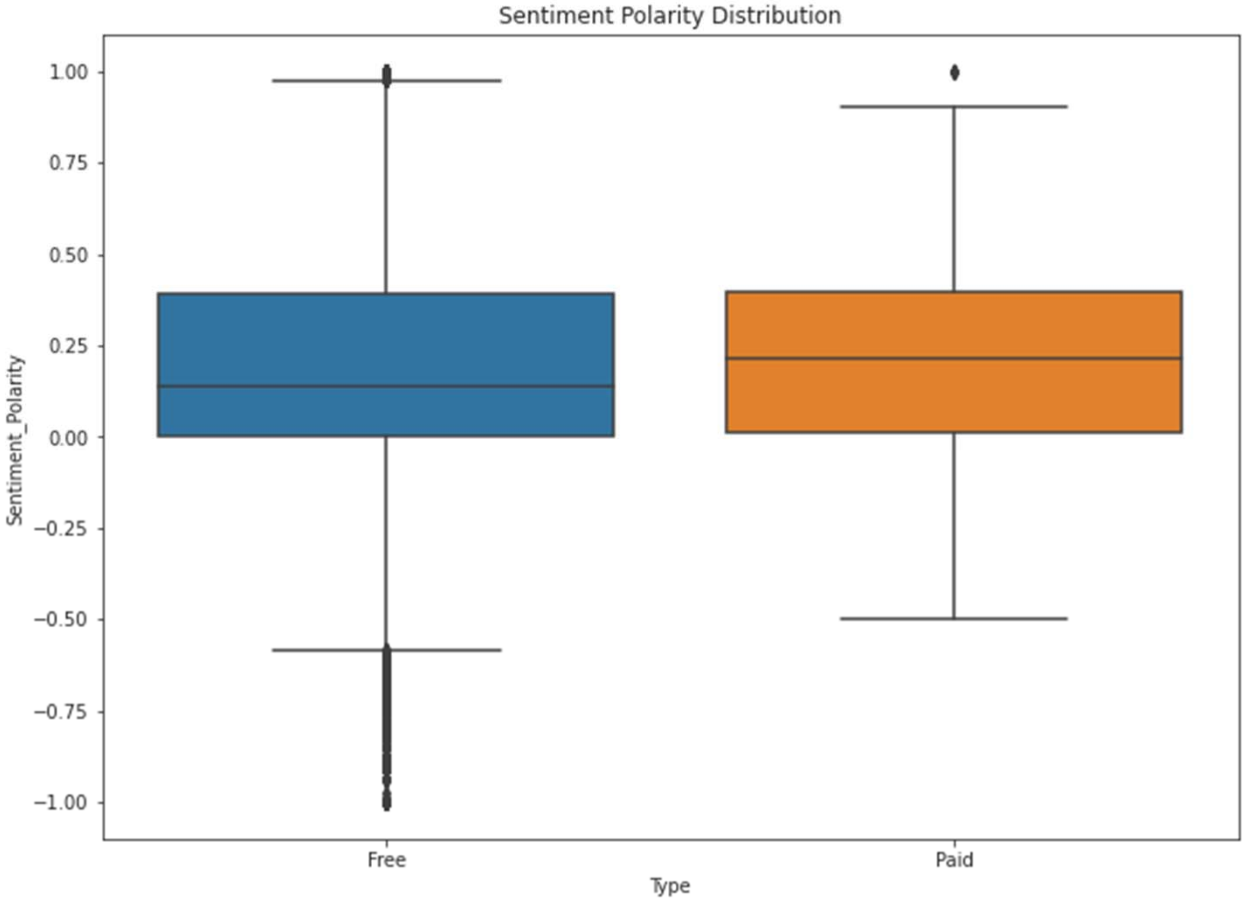
Does the sentiments differ for paid or free apps?



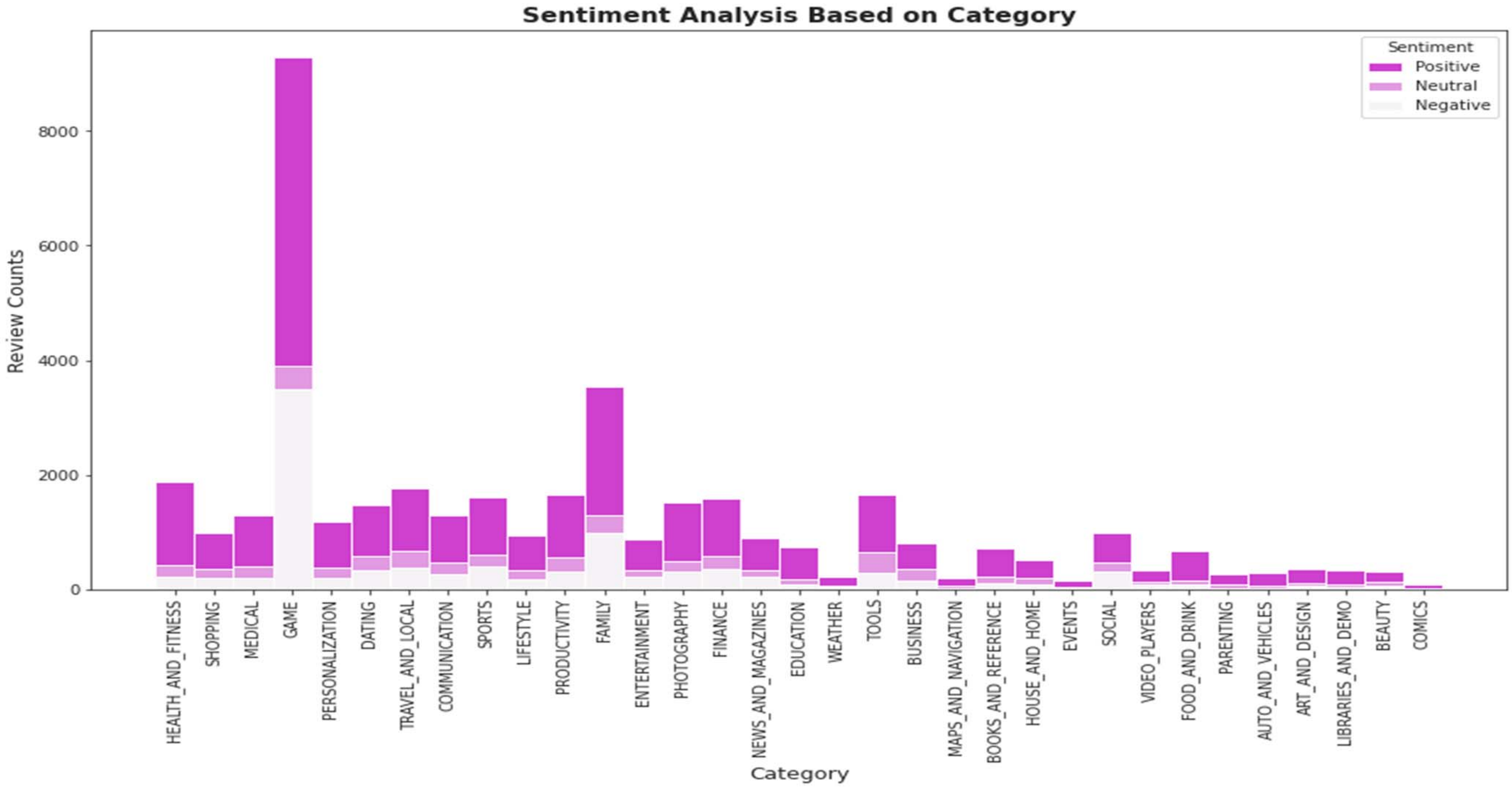


Exploratory Data Visualization

Sentiment polarity for an element defines the orientation of the expressed sentiment. Plotting sentiment polarity scores of user reviews for paid and free apps.

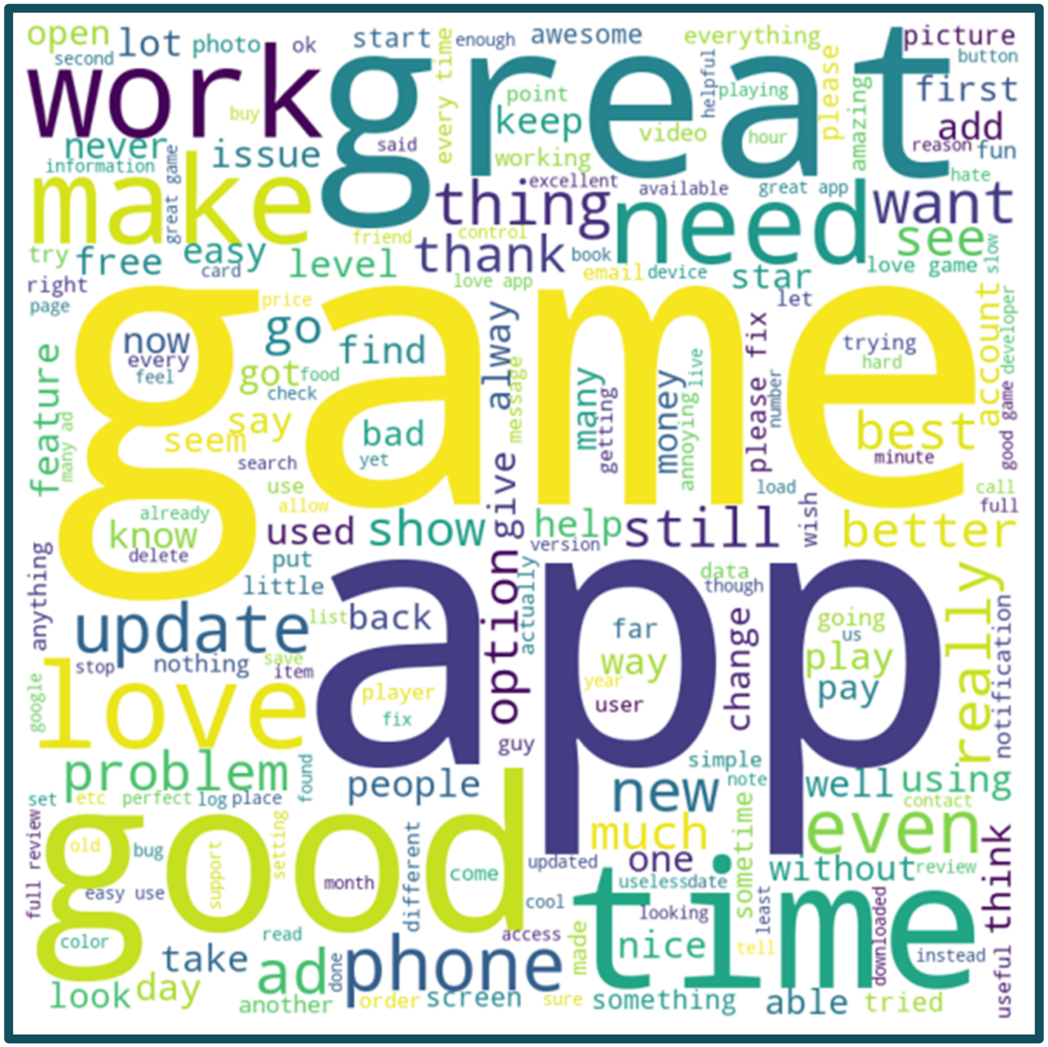


Which category of apps have got more number of reviews with positive sentiments?



Conclusion

- **Games and Medical have the highest market prevalence**
- **number of installs are higher for Communication, Social and productivity** genres
- **Tools, Entertainment, Education related genres have maximum number of apps among all 119 genres.**
- **Users have paid more for lite and high rating apps.**
- **Users have installed more apps of the ones having high reviews**
- **Installs of app doesn't depends on Size.**
- From the sentiment analysis we found **users showed more positive sentiments for paid apps and were harsh while showing sentiments for free apps.**
- Users gave **maximum reviews for game category** followed by family.



References

[1] Statista, Number of available applications in the Google Play store from December 2009 to March 2019, https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/,Online: accessed 22 May 2019.

[2] Statista, Number of mobile app downloads worldwide in 2017, 2018 and 2020 (in billions) https://www.statista.com/statistics/271644/worldwide-free-and-paid-mobile-app-store-downloads/, Online: accessed 22 May 2019.

[3] Fu, B., Lin J., Li, L., Faloutsos, C., Hong, J., Sadeh, N. (2013). Why People Hate Your App — Making Sense of User Feedback in a Mobile App Store, KDD '13, 1276-1284. Roma, P., & Ragaglia, D. (2016). Revenue models, in-app purchase, and the app performance: Evidence from Apple's App Store and Google Play, Electronic Commerce Research and Applications, 17, 173-190.

Thank
You