

EDA Capstone Project

Play Store App Review Analysis

Team: Data Defenders

Shubham Verma

Saraswat Mukherjee

Lokesh Tokas

Shubham Sartape

PROBLEM

- 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 analyse the data to discover key factors responsible for app engagement and success.

OBJECTIVE

- The objective of this project is to deliver insights to understand customer demands better and thus help developers to popularize the product.
- The Dataset consist of 10k Play Store apps for analysing the Android market.
- It consists of in total of 10841 rows and 13 columns.

FLOW

Discussion of Play Store dataset will involve various steps such as:

- Loading the Data into Data Frame and importing useful Libraries for Analysis
- Cleaning the Data / Data Wrangling
- EDA and Visualizations
- Conclusion

LOADING THE DATA AND IMPORTING LIBRARIES

- Here we will import the Data from the path

```
[1] from google.colab import drive
    drive.mount('/content/drive')

Mounted at /content/drive

[4] working_directory = '/content/drive/MyDrive/AlmaBetter/Python/'
    data = pd.read_csv(working_directory + 'Play Store Data.csv')
    reviews = pd.read_csv(working_directory + 'User Reviews.csv')
```

- Similarly we will import all the necessary Libraries which we will be using for the Analysis and Visualization.

```
import pandas as pd
import numpy as np

import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots

import matplotlib.pyplot as plt
import seaborn as sns

from ast import literal_eval as le
```

UNDERSTANDING AND CLEANING THE DATA

- The data consists of 10,841 rows and 13 Columns.
- Columns consists of details regarding the Apps such as Category, Rating, Reviews, Size, etc...
- The Columns required for the Analysis can be cleaned by dropping the null values or simply replacing it to make the analysis easier.

```
[5] data.shape

(10841, 13)

[6] data.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
```

UNDERSTANDING AND CLEANING THE DATA

Like here we have cleaned the Column 'Install' by replacing the '+' Sign, removing 'commas' between the numbers and converting to 'int' Format.

```
[15] data['Installs'].value_counts()
```

```
1,000,000+    1579
10,000,000+   1252
100,000+      1169
10,000+       1054
1,000+        907
5,000,000+    752
100+          719
500,000+      539
50,000+       479
5,000+        477
100,000,000+  409
10+           386
500+          330
50,000,000+  289
50+           205
5+            82
500,000,000+ 72
1+            67
1,000,000,000+ 58
0+            14
0              1
Name: Installs, dtype: int64
```

```
data['Installs'] = data['Installs'].apply(lambda x : str(x).replace('+',''))
data['Installs'] = data['Installs'].apply(lambda x : str(x).replace(',',''))
data['Installs'] = pd.to_numeric(data['Installs'])
data['Installs'].dtype

dtype('int64')
```

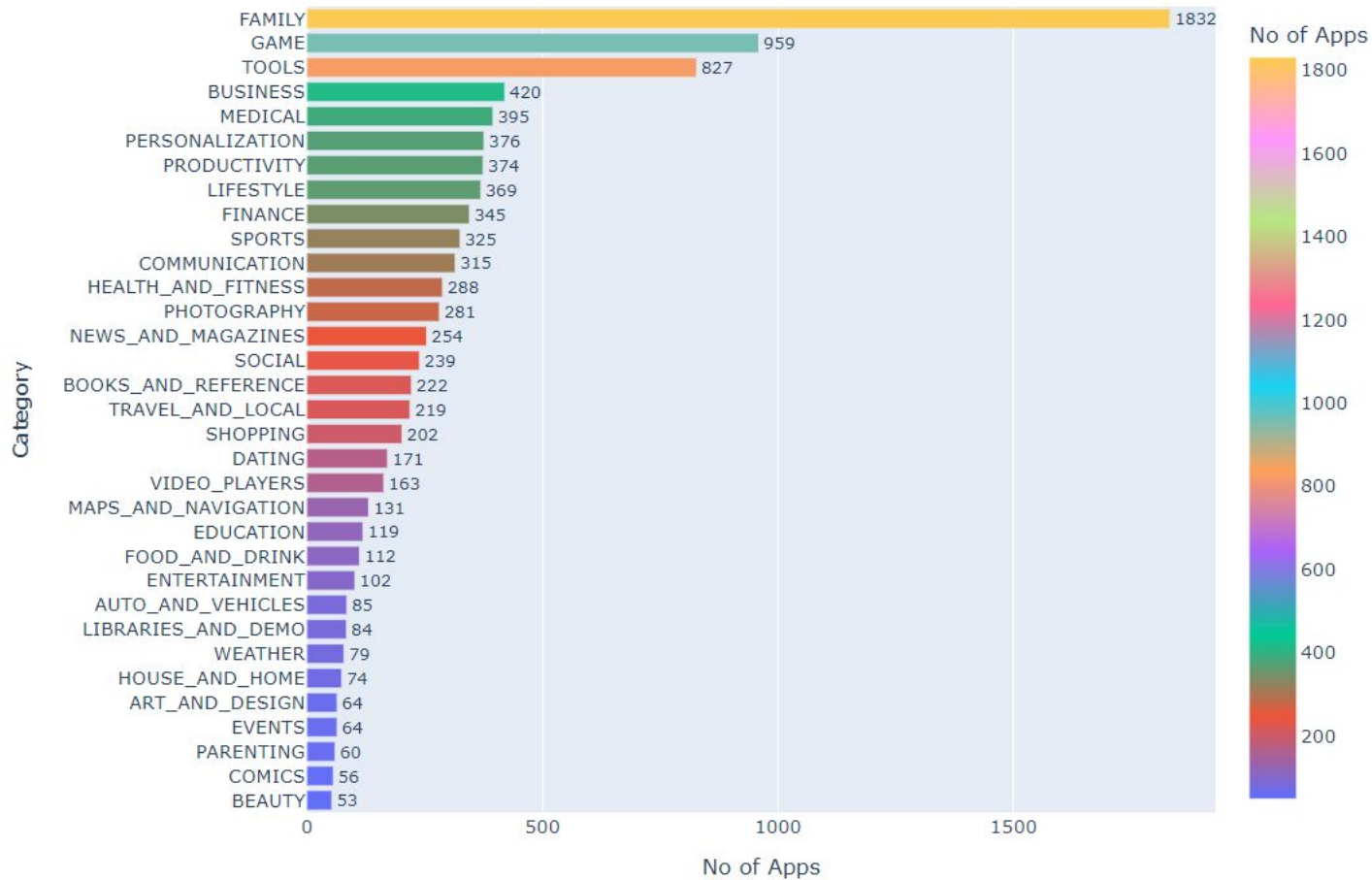


```
data['Installs'].value_counts()
```

```
1000000    1579
100000000  1252
100000     1169
10000      1054
1000       907
5000000    752
100        719
500000     539
50000      479
5000       477
100000000  409
10         386
500        330
50000000   289
50         205
5          82
500000000  72
1          67
100000000  58
0          15
Name: Installs, dtype: int64
```

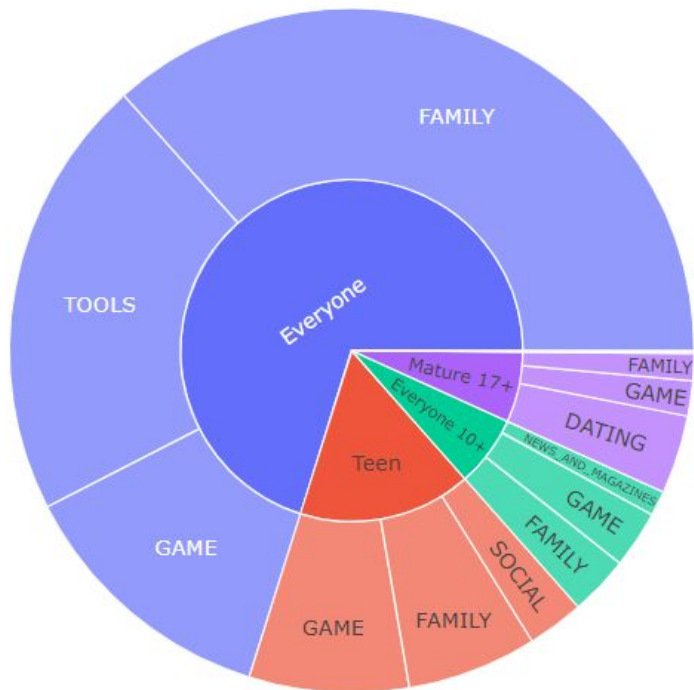
EDA and Data Visualization

Top Category in play store.

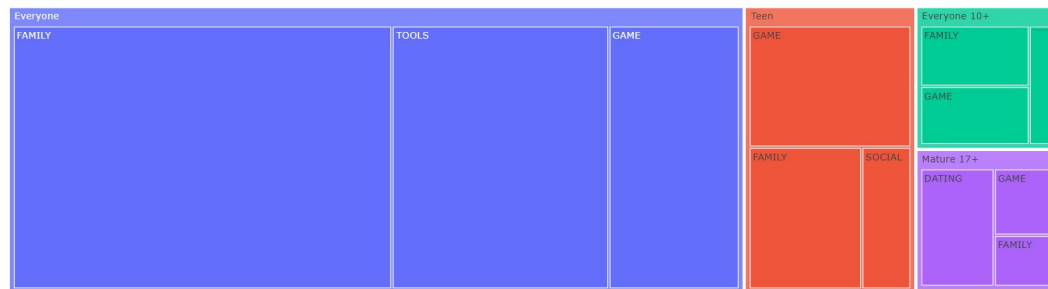


Top 3 Categories in Play Store by Content Rating

Pie Chart

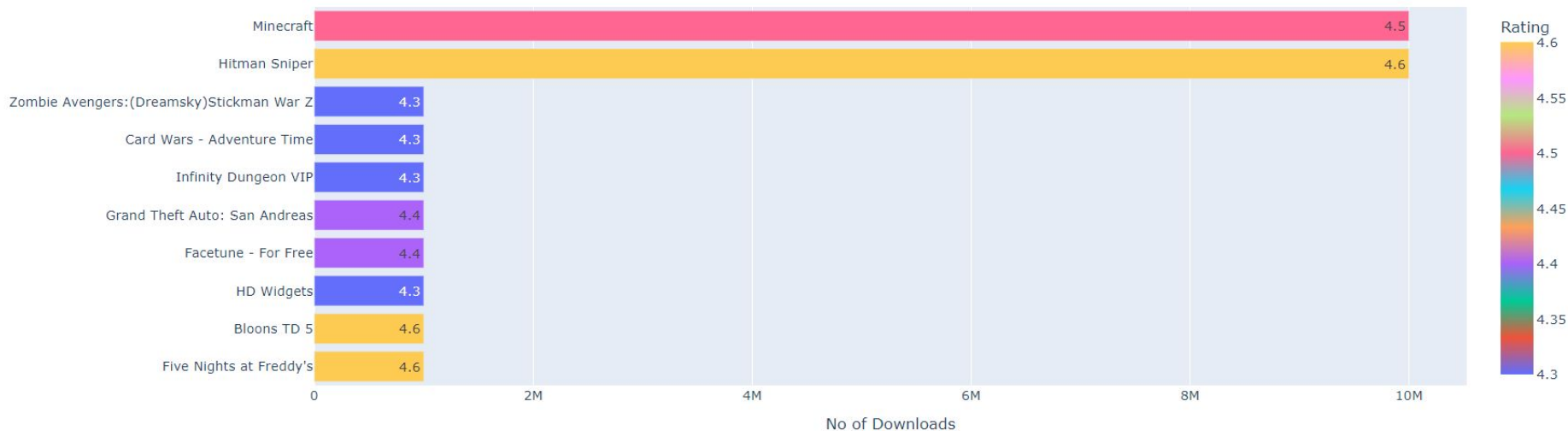


Tree Map

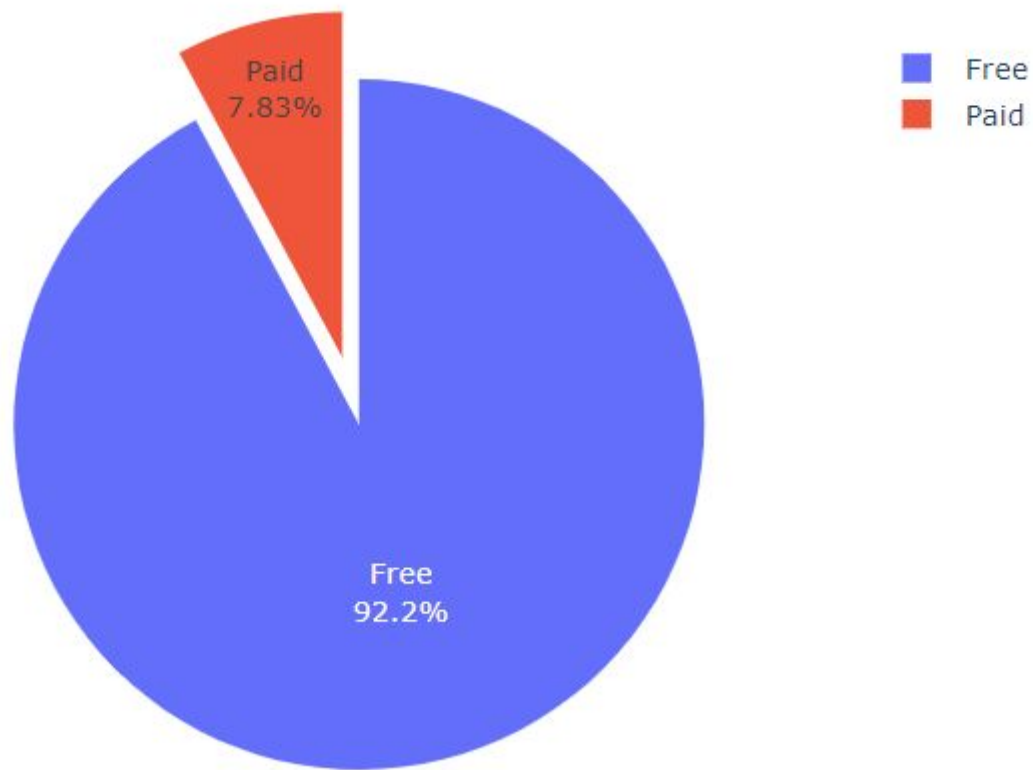


Top 10 most installed paid apps and their ratings.

Top 10 Paid Apps on Play Store

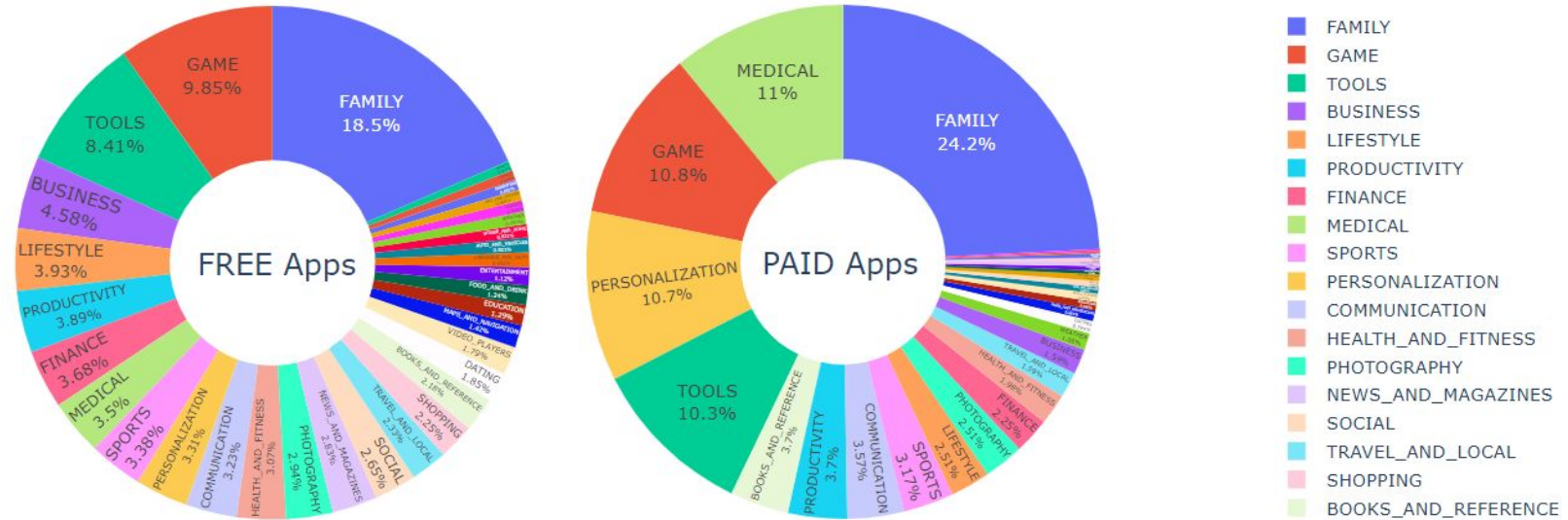


Paid vs Free app in Play Store



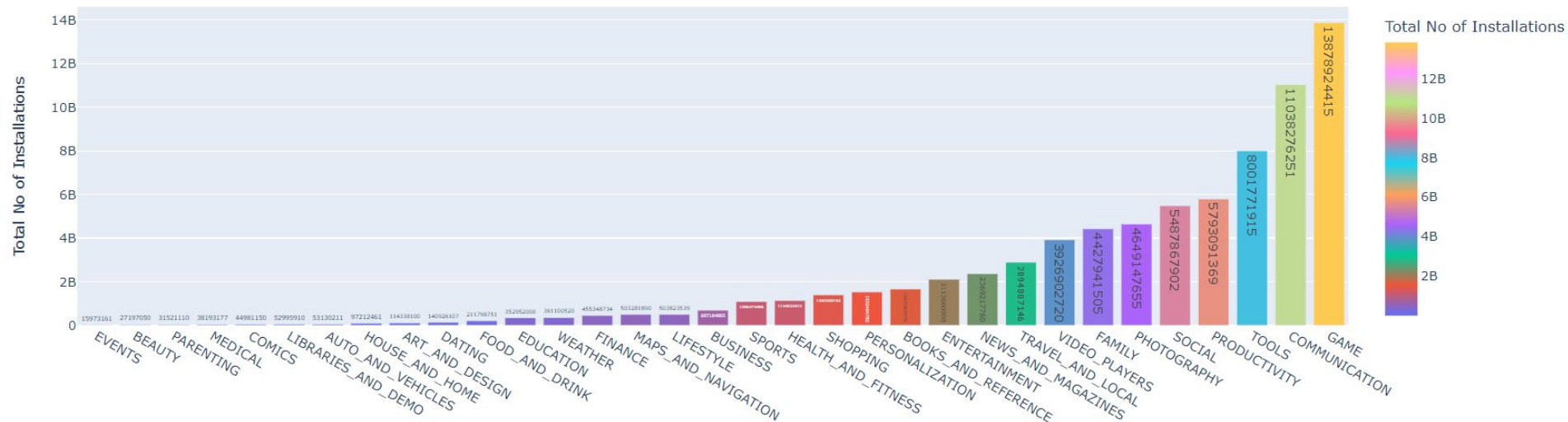
Paid Apps and free Apps by category in play store

Free vs Paid Top Categories in Play Store



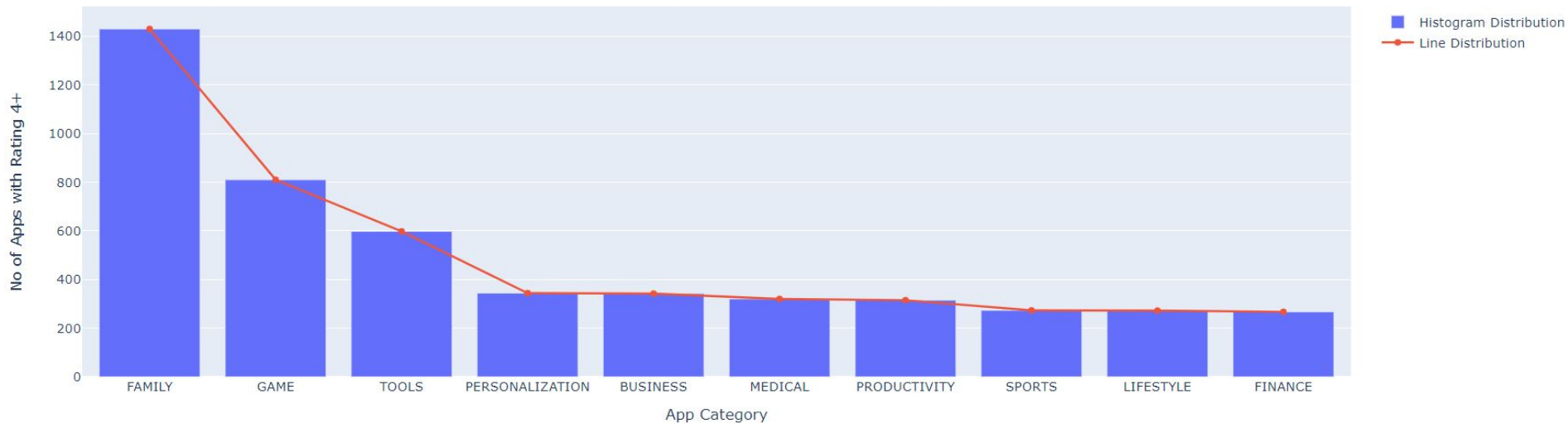
App Installs in each Category

Distribution frequency by app installation no and category

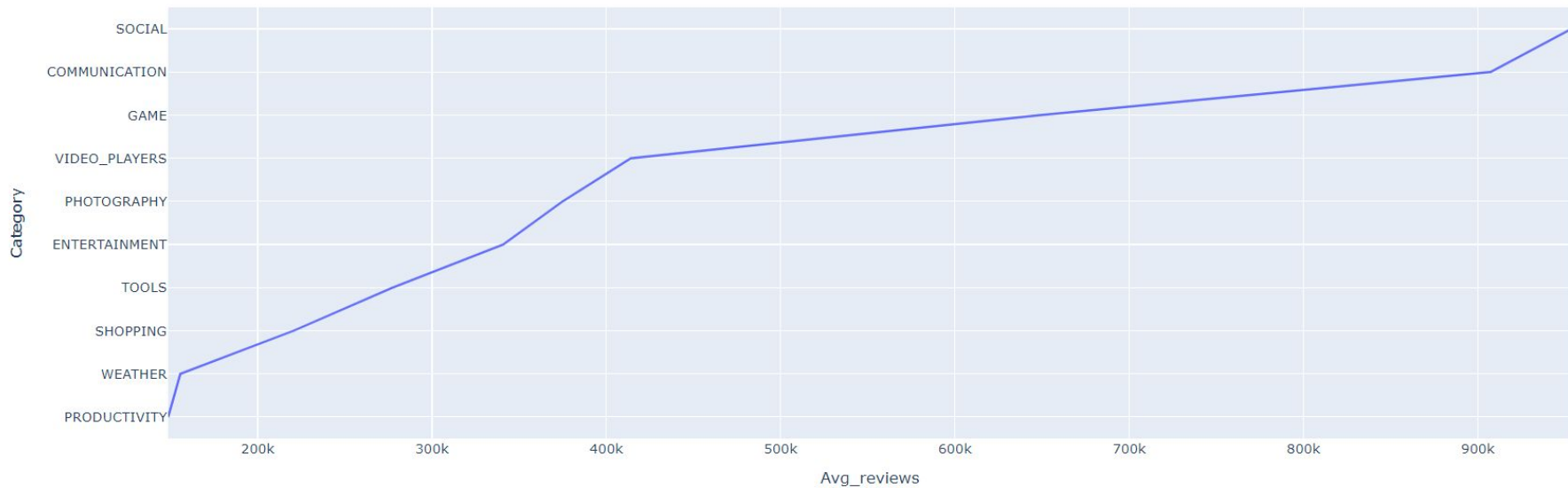


Top 10 category with Rating 4.0 and above.

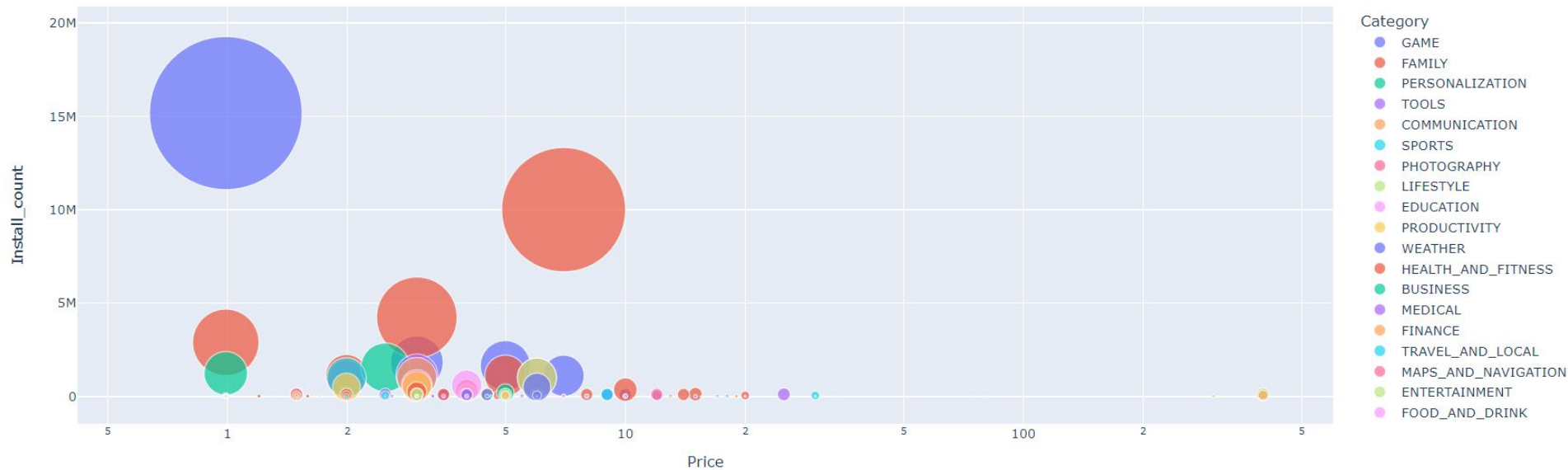
Top 10 category with Rating 4.0 and above



Top 10 category with highest avg no of reviews.

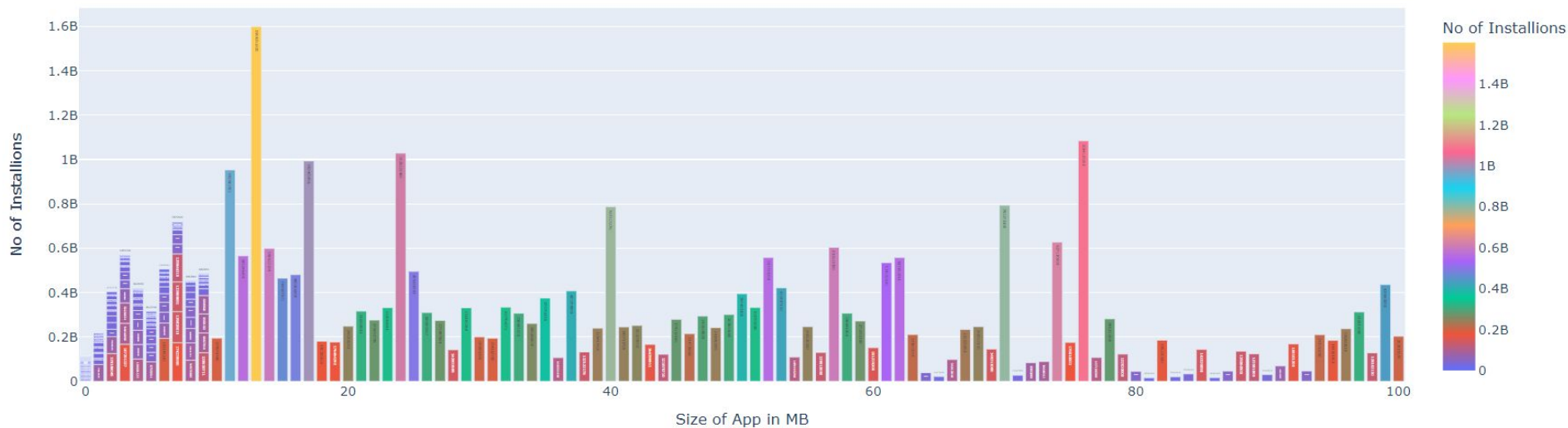


Paid app in each category by Price and no of installation.



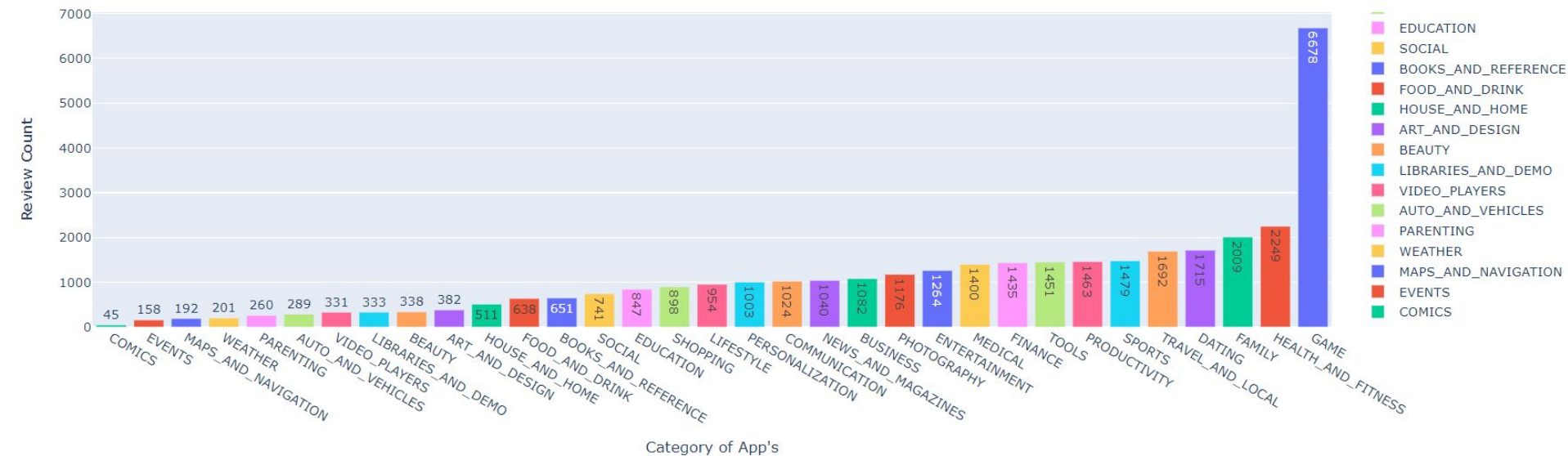
App installation by Size.

Distribution frequency by app installation no and category



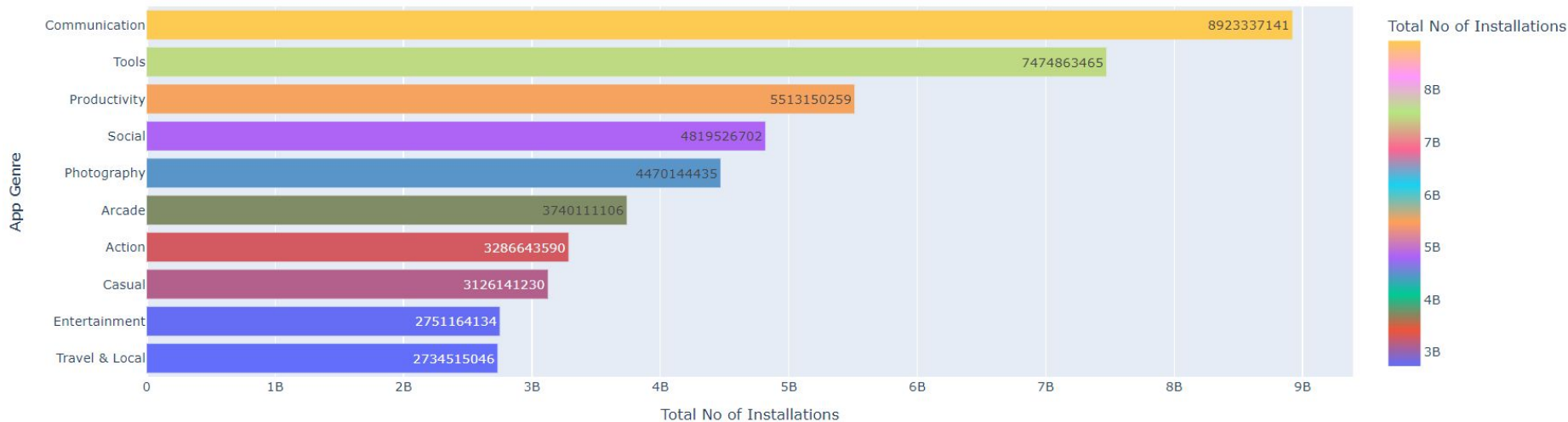
Top App category with most reviews.

Top App category with most reviews.



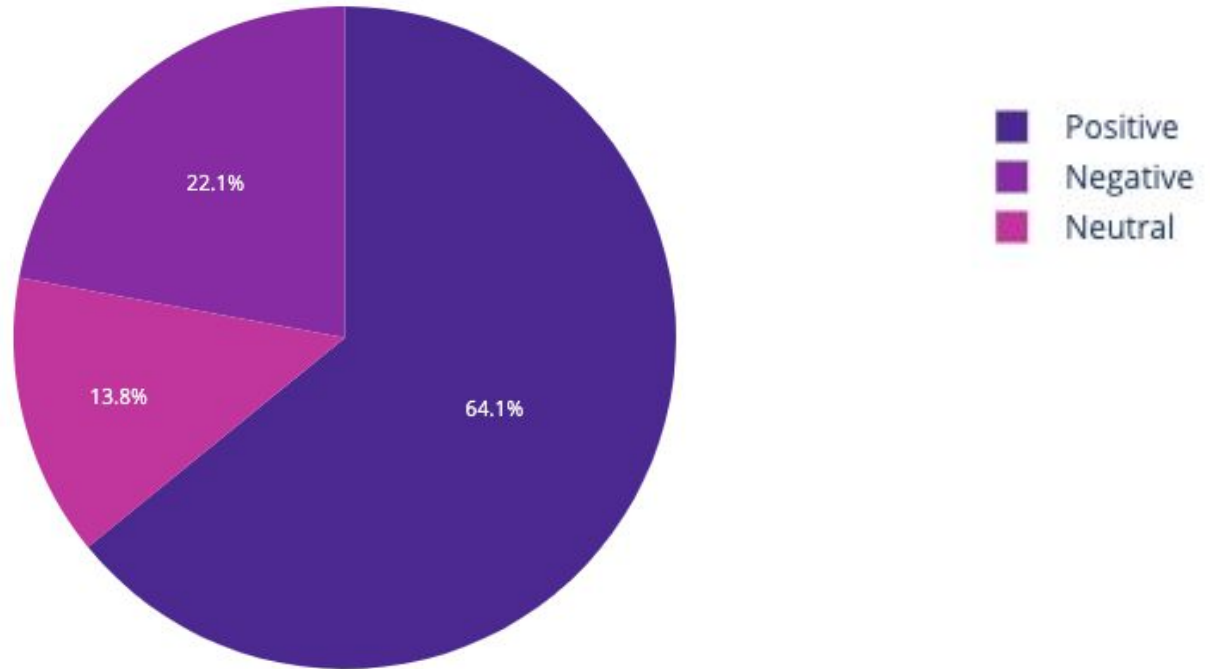
Total Installed Apps by top 10 Genres having rating above 4.0

Total Installed Apps by top 10 Genres having rating above 4.0



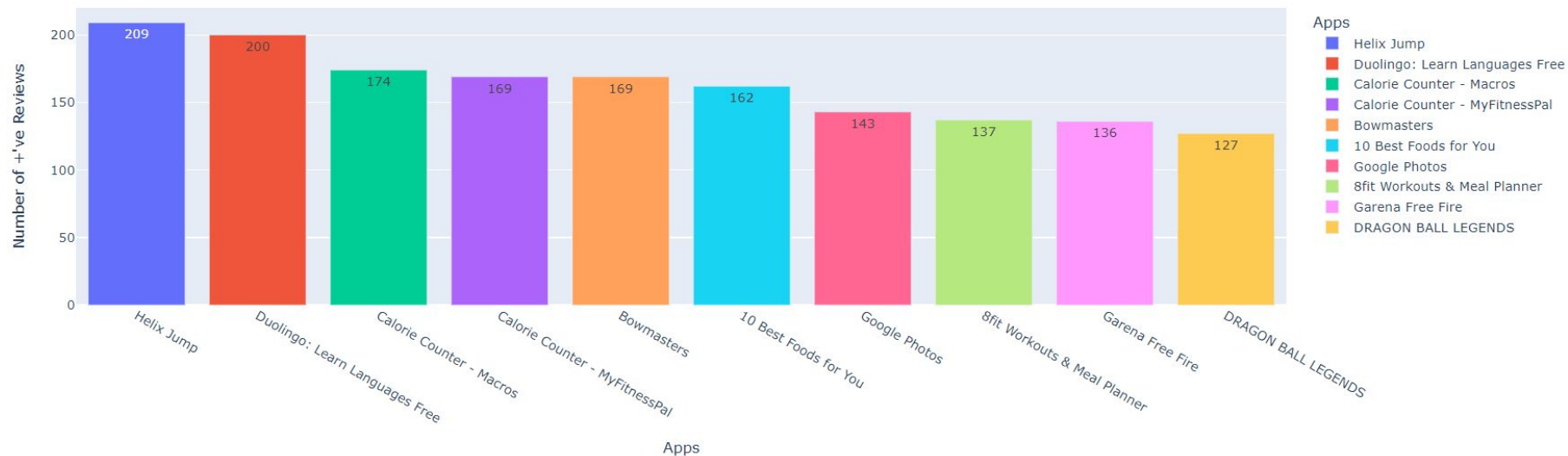
Percentage of Review Sentiments

Percentage of Review Sentiments

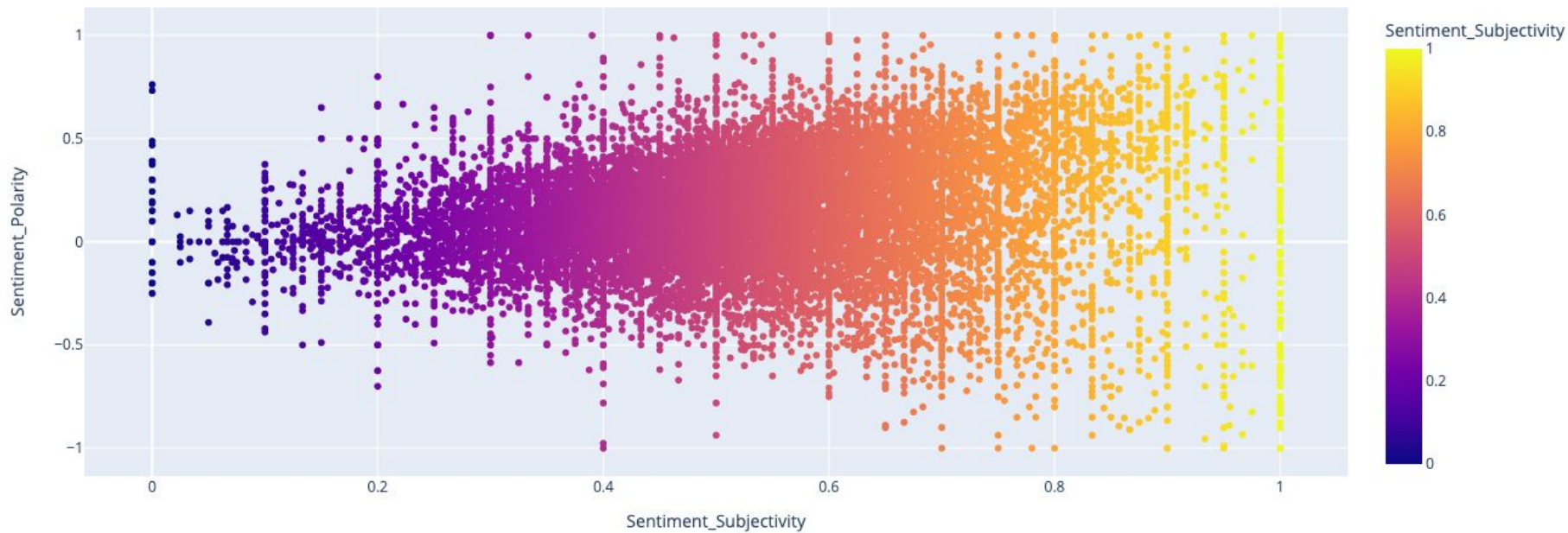


Top 10 Apps with Maximum Positive Reviews

Top 10 Apps with Maximum Positive Reviews



Is sentiment subjectivity proportional to sentiment polarity?



CONCLUSION

- Category – Apps in **Family, Tools & Game** category are more likely to have higher success rate.
- Content Rating – Apps with content rating of **Everyone** has dominating market share.
- Paid apps – Top installed apps in **Paid** type are mostly games with average rating of 4+ with installation numbers above 1M.
- Market Share – **Free** type apps dominates the play store market share at **92.2%** while **Paid** types are only at **7.83%**
- Installation – Apps in the Game category have an overall highest number of installations.
- Rating - Apps in the Family, **Game & Tools** category are more likely to have higher user ratings.
- User Engagement – Apps in the Social & **Communication** category have a much higher user engagement ratio as seen by comparing Review data.
- Size – Apps with **size** between **7-24MB** have higher installation numbers.
- User Comments – Users are more compelled to give feedback for apps in the Game category, leading to an improved app experience.
- Most reviewed apps - 'Helix Jump', 'Duolingo: Learn Languages Free', 'Calorie Counter – Macros', These Apps have the highest number of Positive reviews.

CHALLENGES

- Data Cleaning

Challenges faced in finding the unwanted values/null values and replacing the same with suitable values for Data Analysis.

- EDA and Visualization

The research for finding a compatible library for visualization with lots of trial & errors and finally settling for one (plotly) with vibrant and precise plotting which improved the visualization analysis to some extent.

- Learning

Lots of new ways of EDA and visualization were learnt in the process.

Thank You