Capstone Project-1

: Play Store App Analysis

Team Member

Omkar Sargar

Utkarsh Shrivastava

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.
- We are provided with two datasets, namely: ‘Play Store Data’ and ‘User Reviews’.
- ‘Play Store Data’ has details about app category, rating, size, and more, while the ‘User Reviews’ table contains customer reviews of the android apps.
- Our goal was to explore and analyze the data to discover key factors responsible for app engagement and success.

Data Overview

The table below gives a general overview of both the datasets used.

Play_Store_Data

- **App**
- **Category**
- **Size**
- **Rating**
- **Reviews**
- **Installs**
- **Type**
- **Price**
- **Content Rating**
- Genres
- Last Updated
- Current Ver
- Android Ver
- **Rating Group**
- **Revenue**

User_reviews

- **App**
- **Translated Review**
- **Sentiment**
- Sentiment_Polarity
- Sentiment_Subjectivity

Data Summary

This data set contains Play Store App information for a wide variety of apps and its categories, includes information such as how many apps belong to a certain category ,app ratings out of five(5), how many downloads were made, User reviews for the apps in the dataset, the size of the app, type and Genres etc.

The Dataset for Play Store Apps includes data about Application Information as well as User Reviews for them which was to be analysed and proper insights should be taken out which can be useful to provider in future for making important decisions.

Outline

- Exploring and Cleaning the Dataset
- To establish relationship between various features of the Dataset.
- Present these relationships using various Data Visualization Techniques.
- Draw the useful insights from it.
- Conclusion.

Exploring and Cleaning the Dataset

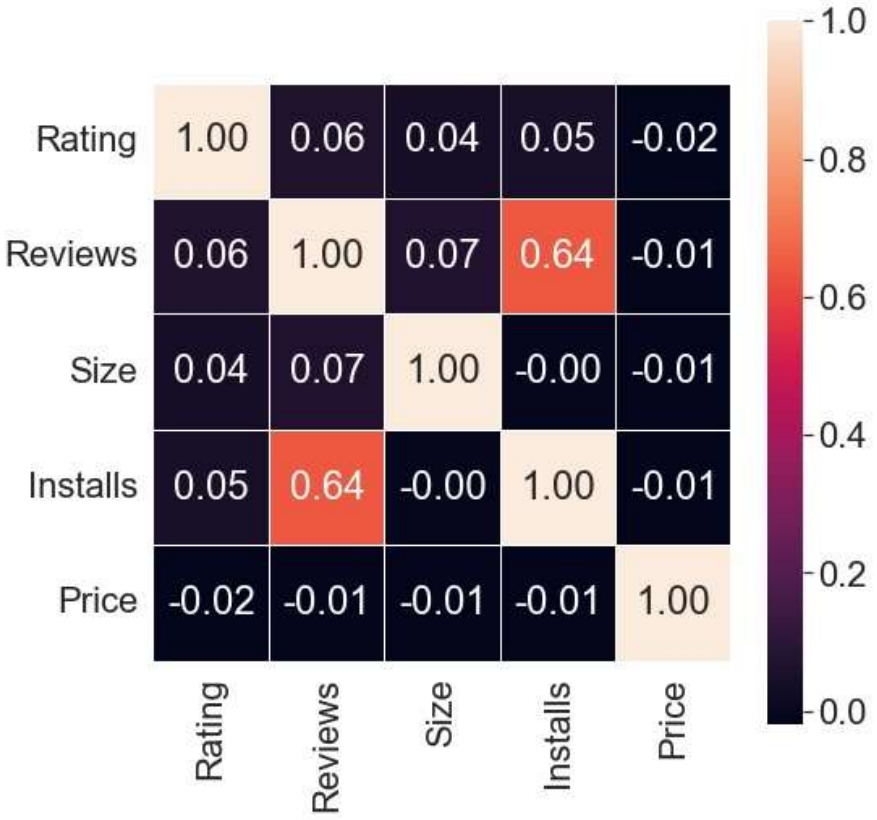
The first step is download dataset and check the missing value and null value.

- For Play Store Table, the numerical values were replaced with median values for columns.
- Similarly, for object type columns, missing points were replaced by mode.
- For User Reviews Table, there were missing values for over 26.8k reviews.
- Since empty values mean nothing for the table, the concerned rows were removed .

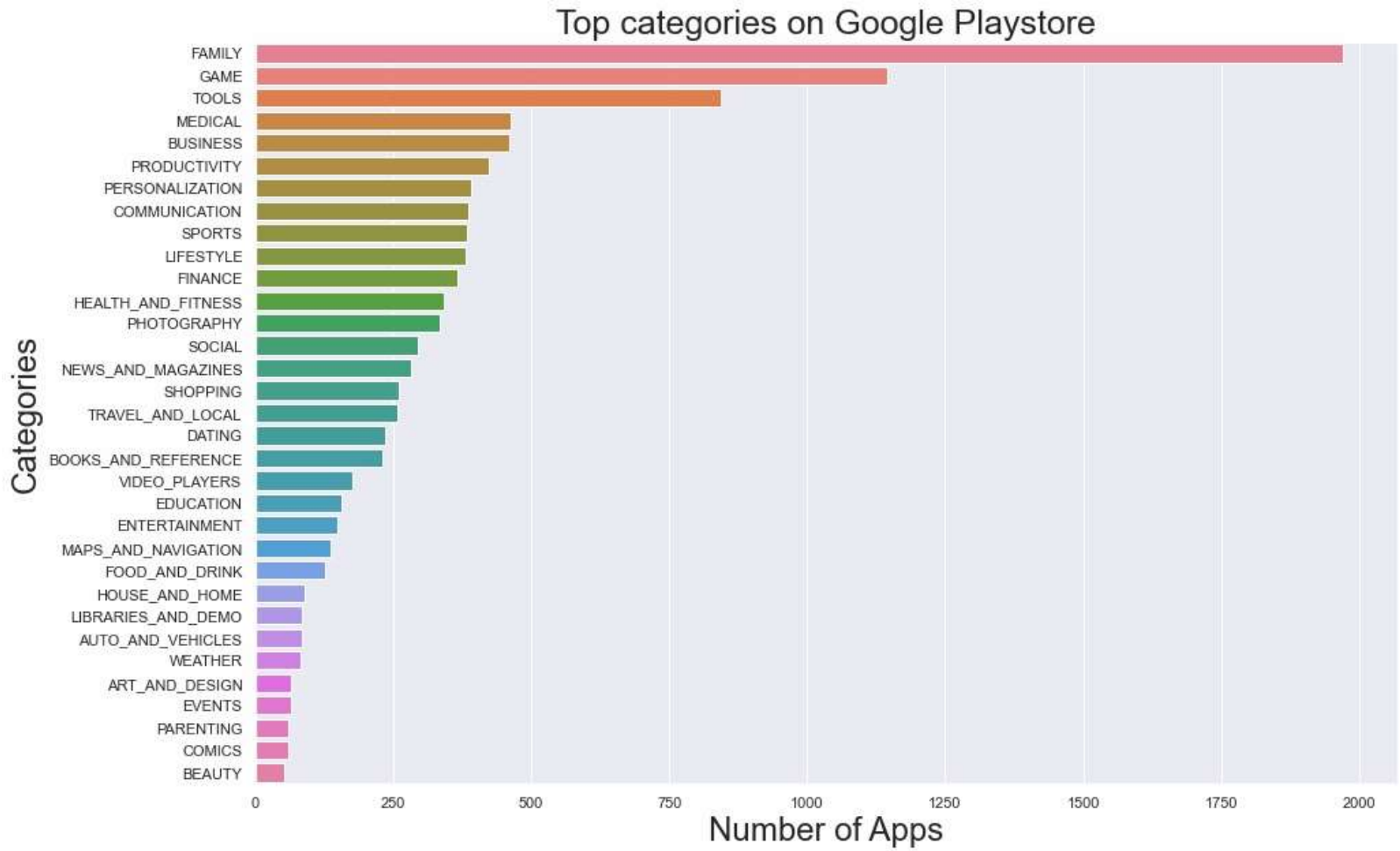
```
App                0
Category           0
Rating          1474
Reviews            0
Size               0
Installs           0
Type               1
Price              0
Content Rating     1
Genres             0
Last Updated       0
Current Ver        8
Android Ver        3
dtype: int64

App                        0
Translated_Review      26868
Sentiment              26863
Sentiment_Polarity     26863
Sentiment_Subjectivity 26863
dtype: int64
```

Correlation Matrix of the Data

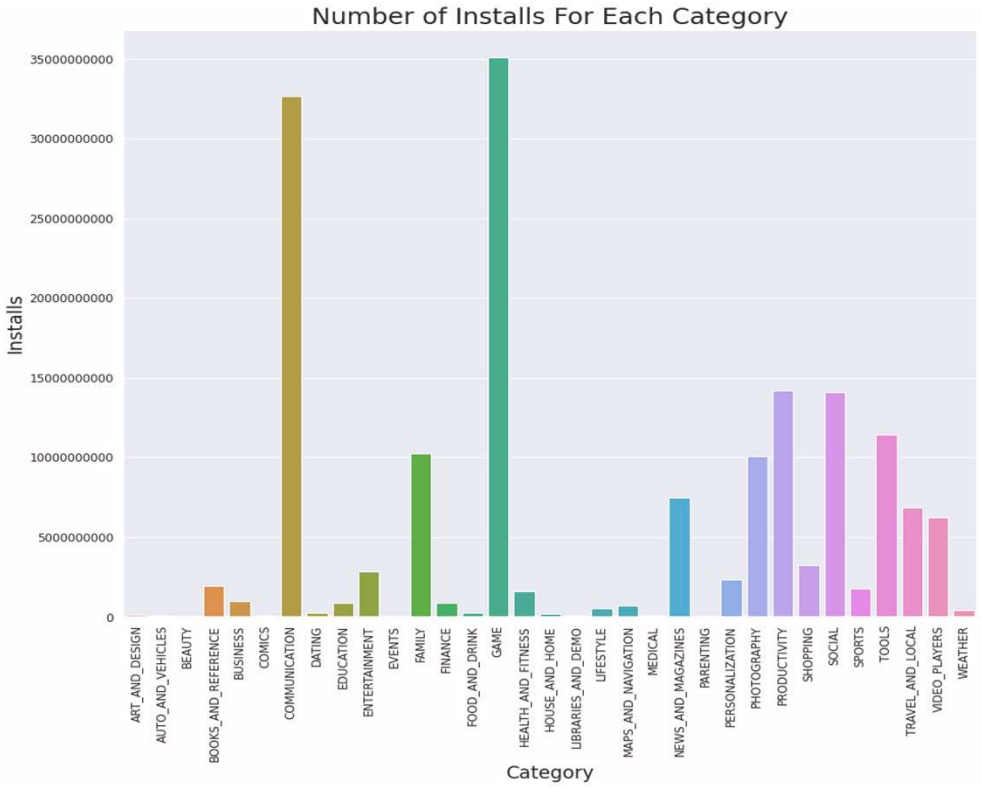
- Correlation matrix shows a correlation between columns if exists.
- There seem to be a correlation between the number of installs and number of reviews.
- There doesn't seem to be any significant correlation between any other variables/columns.

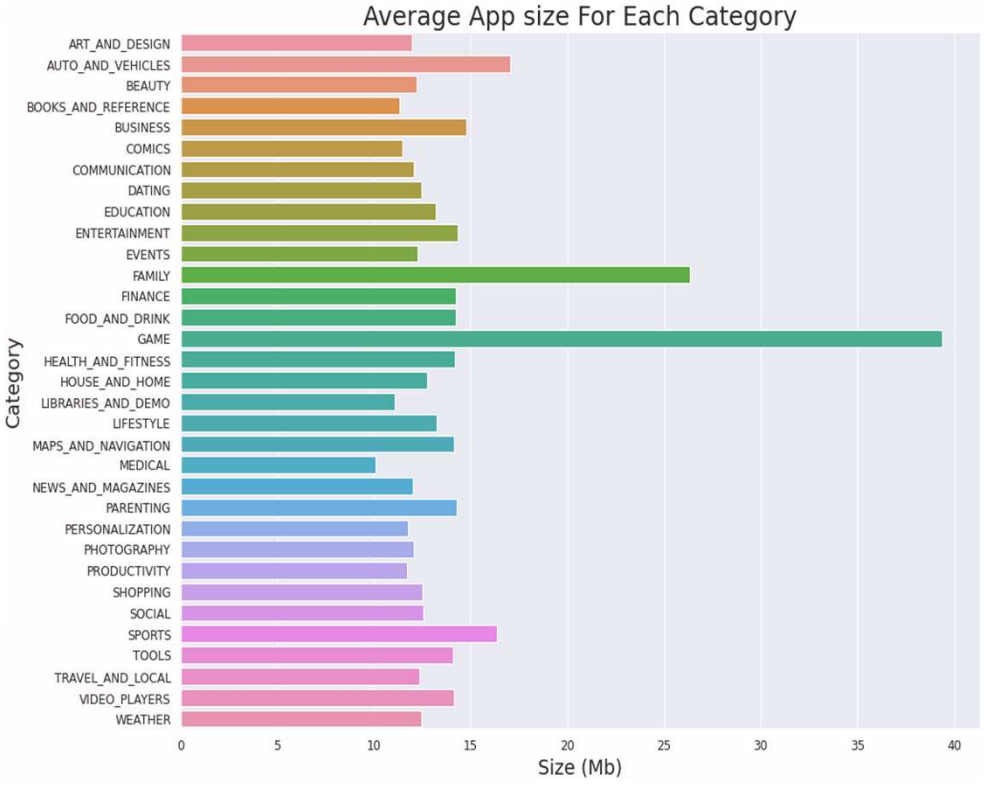


Categorical Distribution of Apps

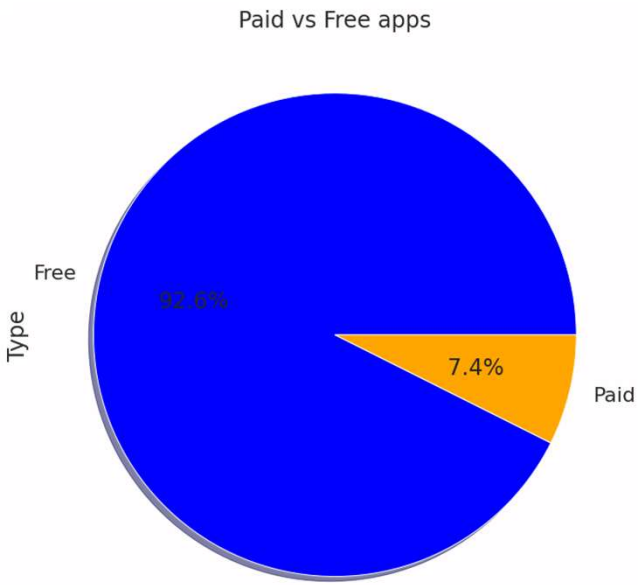


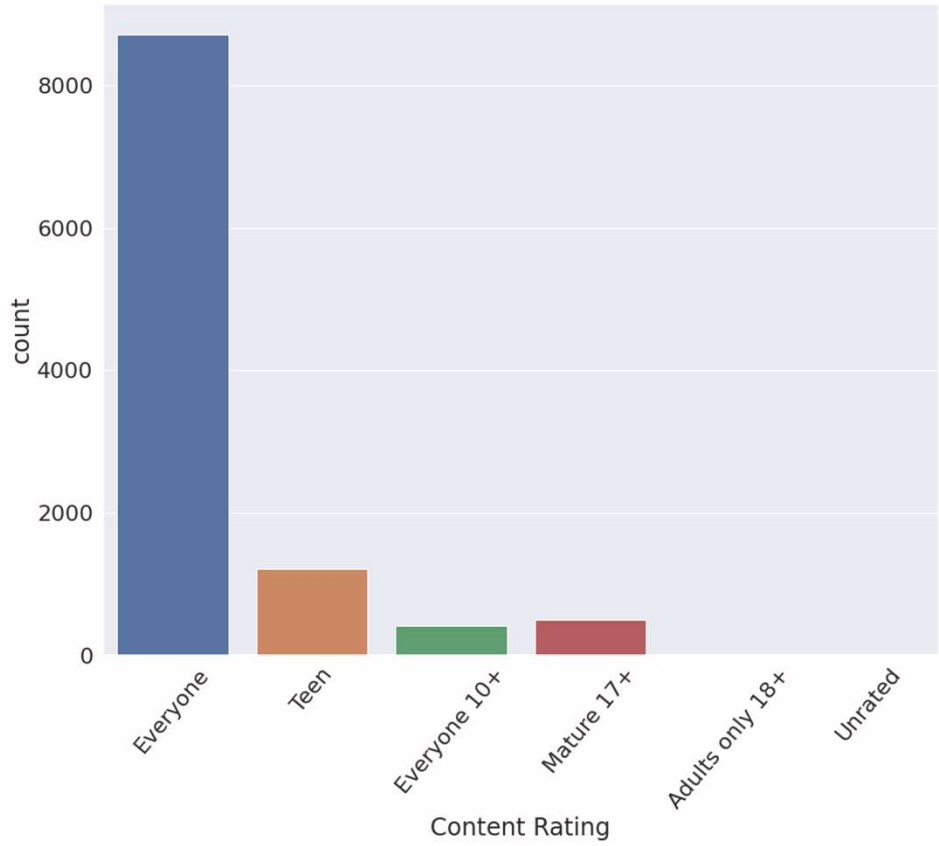
Analysis of Installation and Size data



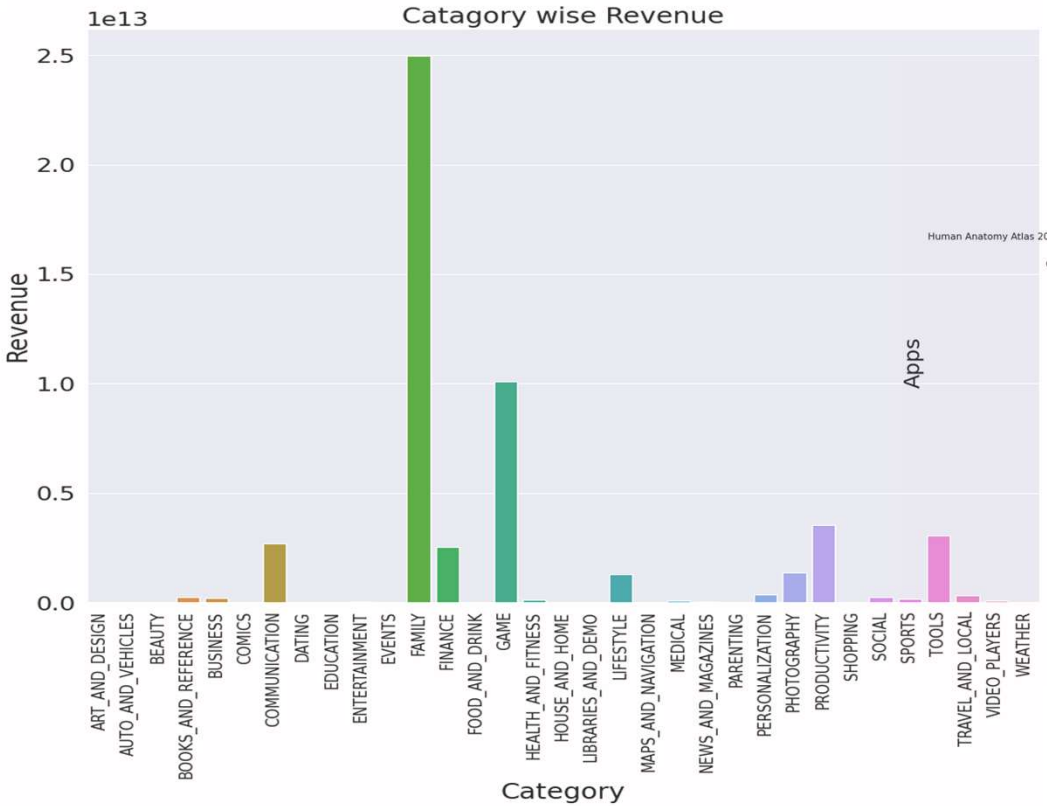


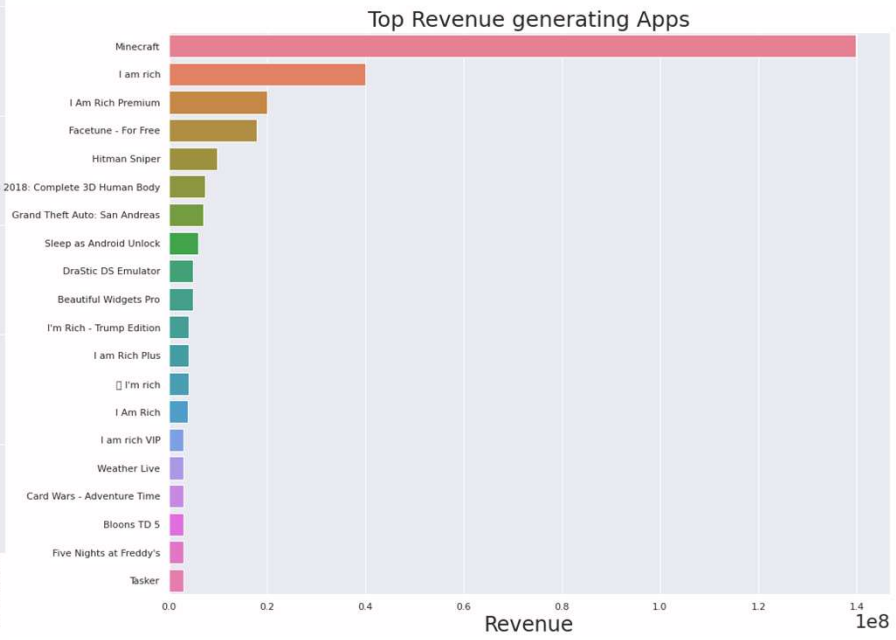
App Distributions





Revenue Analysis



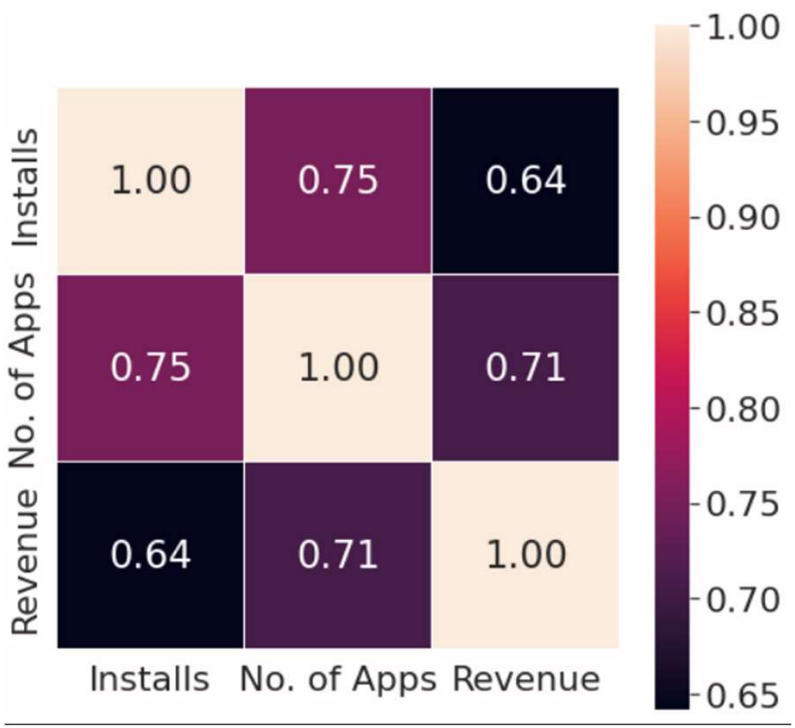


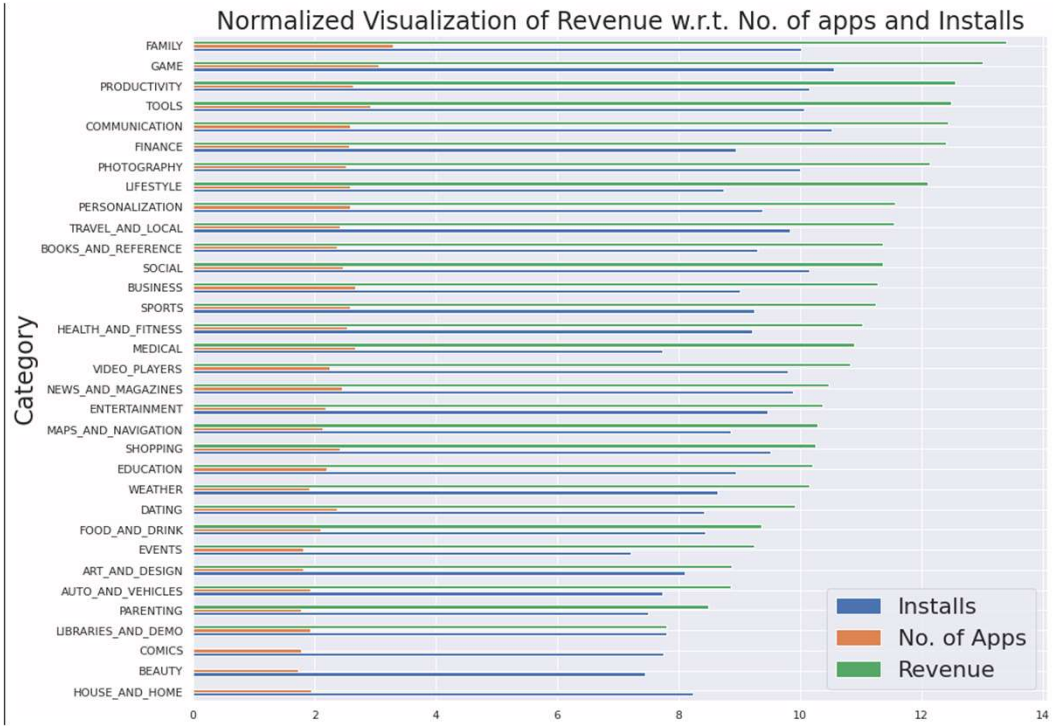
Most Expensive Apps

```
# top 5 expensive Apps
df = data.groupby(by =['App', 'Category']) [['Price']].sum().reset_index()
# we get the sum of price of every uique app but lets sort them Descending by price
df = df.sort_values(by = 'Price', ascending=False).head()
df
```

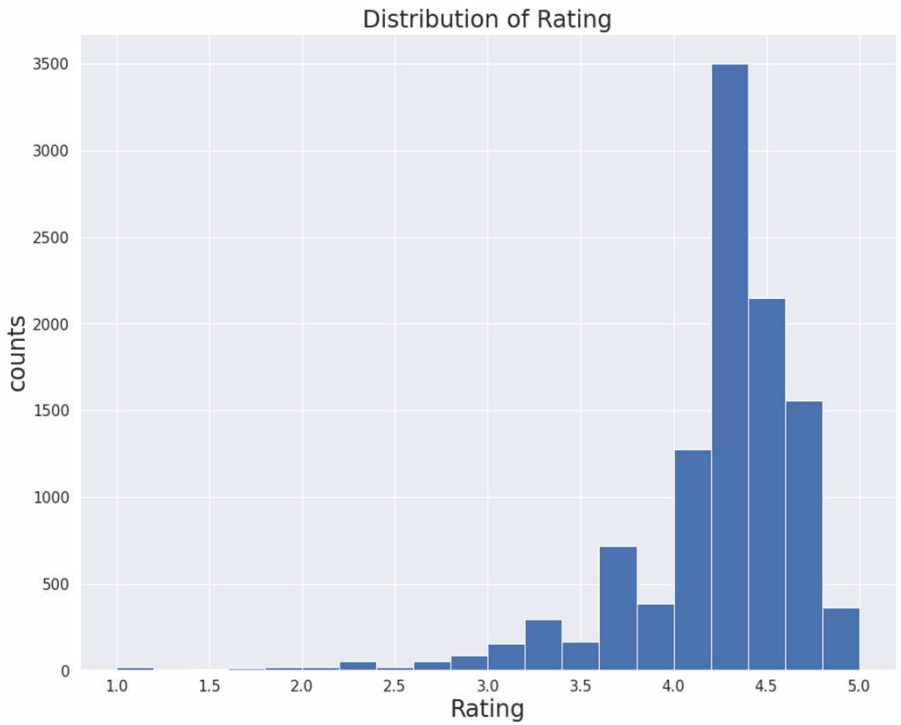
	App	Category	Price
5371	I'm Rich - Trump Edition	LIFESTYLE	400.00
9623	most expensive app (H)	FAMILY	399.99
5372	I'm Rich/Eu sou Rico/أنا غني/我很有錢	LIFESTYLE	399.99
5346	I Am Rich Pro	FAMILY	399.99
9739	💎 I'm rich	LIFESTYLE	399.99

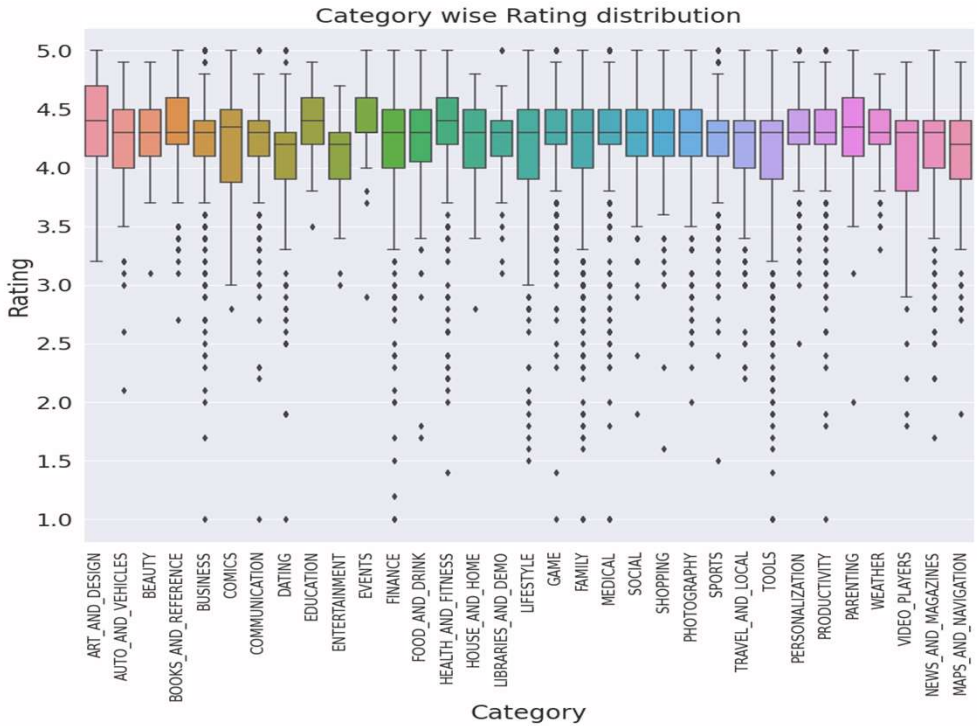
Revenue Correlation



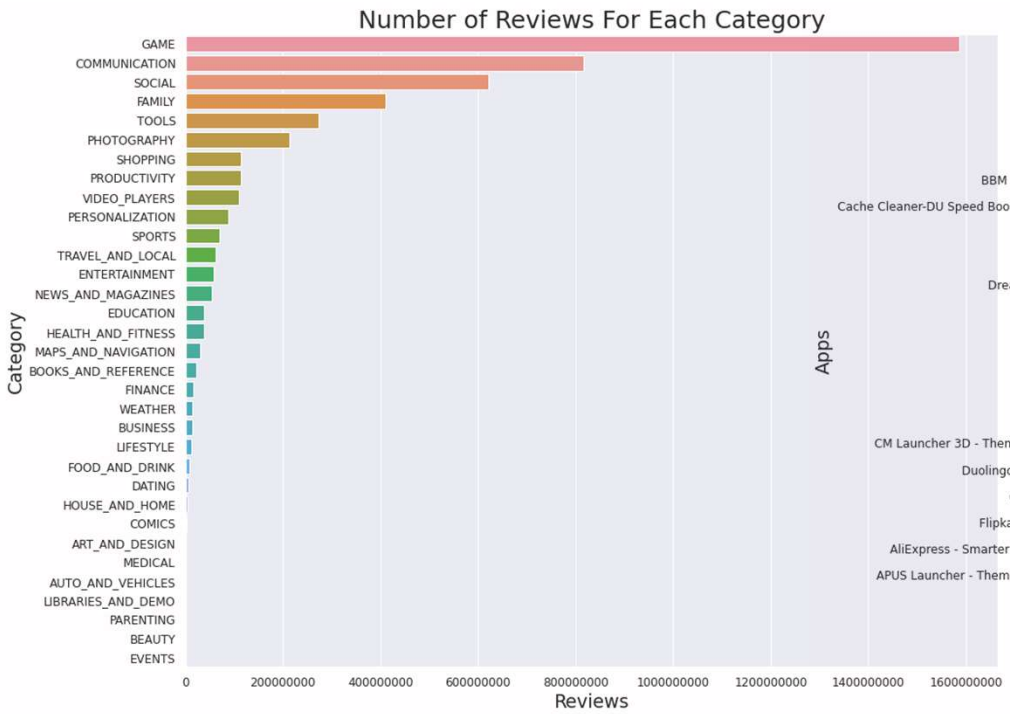


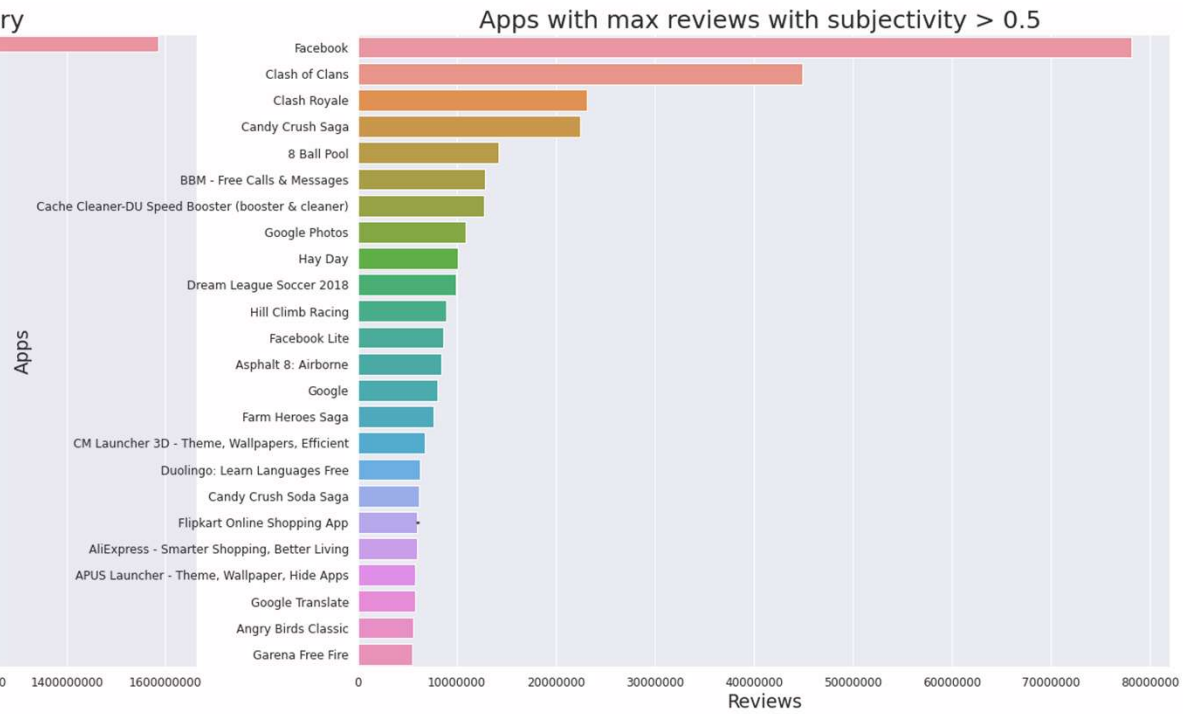
Rating Distribution



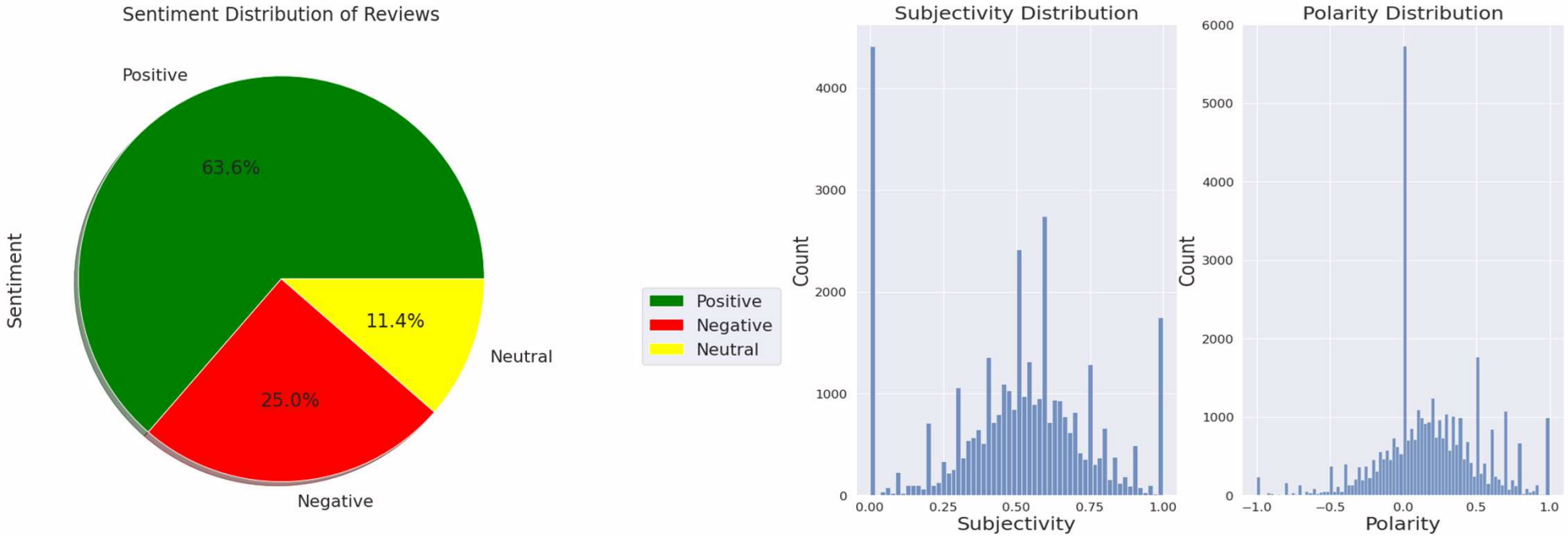


Review Analysis



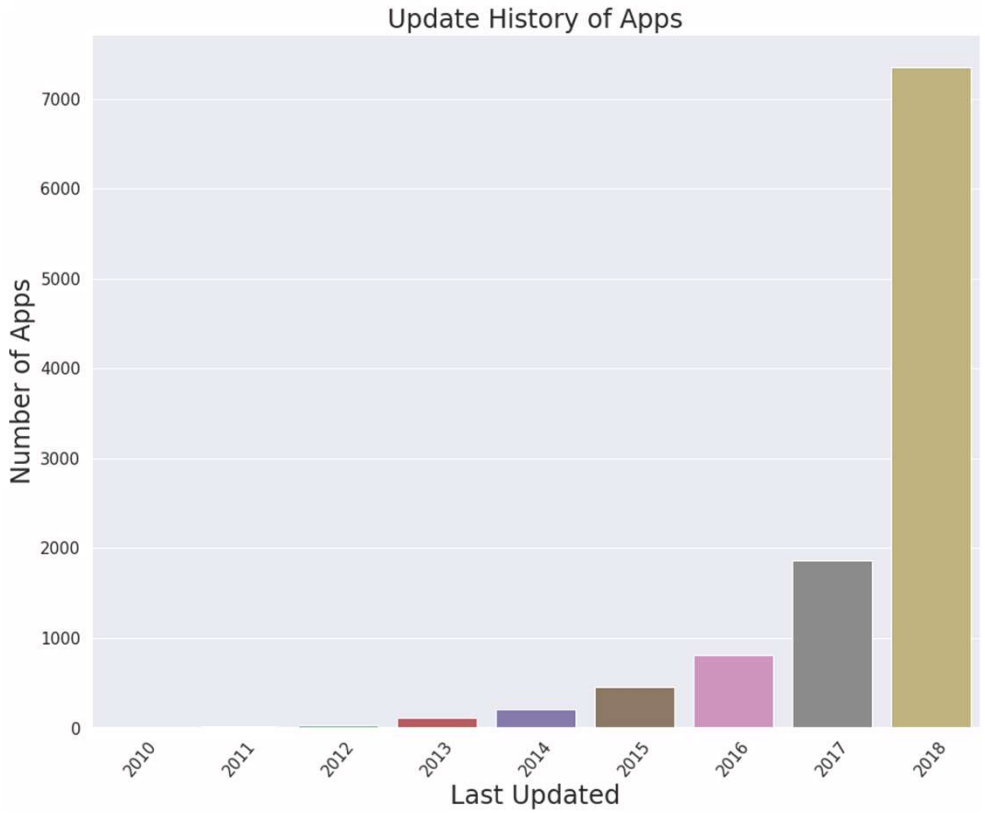


Sentiment Distribution of Apps



Active App Status

- There are 42 apps whose update hasn't been released since 2012.
- Over 820 apps haven't had an update since 2015. Possibly decommissioned or discontinued.
- There are 2761 apps with last updates between 2015 and 2017. Could be obsolete or discarded.
- There are 7348 apps with last update in 2018. Most likely Active.



Conclusions

- Most popular categories amongst developers include Family, Game and Tools; while the users prefer Games, Communication, Social and Productivity above others, followed by Tools and Family.
- Paid apps seem to occupy 7.4 % of the market share while 92.6% of the apps are free.
- Through top 10 most installed and most reviewed/popular apps, we can generate revenue by increasing advertisement in them. We can direct users towards our revenue generating apps. Also, paid versions of such free apps can also be introduced including some exclusive features.
- With least popular and least installed free apps, we probably don’t need to invest resources in them, so they can be pushed towards decommissioning. If the # installs is high with bad ratings, then those apps need to be improved.

Conclusions

- To cover a more broad user base, from the most profitable apps, we can conclude that more apps with low price and high profitability should be promoted.
- The Rating plot seems to be a skewed normal distribution left skewed around average 4.2 stars.
- The sentiment distribution shows the degree of polarity of the distributions. In general, we find 63.6% of reviews to be positive, 25% negative and 11.4% neutral.
- The boxplot of rating distribution can be used to check for differences between categories. # Ratings & Reviews signifies the level with which the user preferred the app. Categories with less outliers and a more balanced distribution and more installs can be a much safer investment than others.

Thank You !