Capstone Project – 1

EDA On Play Store App Reviews

By

SHAIK AHMAD BASHA

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 have two data sets. One is play store data which contains each app (row) . And the other dataset is user reviews.

Our main objective is to explore and analyze these datasets to discover key factors responsible for app engagement and success.

Work Flow:


Data understanding
Data Cleaning and Manipulation
Exploratory Data Analysis

Data Understanding

After collecting data its very important to understand the data. The App dataset has 10841 rows and 13 columns The Review dataset has 37427 rows and 5 columns:

Features of Play Store App dataset:

- **App :** Name of the App
- **Category :** Category of the app
- **Rating :** Rating given for the app
- **Reviews :** Total number of reviews for the app
- **Size :** Size of the app (MB)
- **Installs :** How many installs done for the app
- **Type :** Type of the app
- **Price :** Price of the app
- **Content Rating :** It is age appropriate or not
- **Genres :** Genre of the app
- **Last Updated :** Date of latest update
- **Current Ver :** Current version of the app
- **Android Ver :** Android version of the app

Features of User Reviews dataset:

App : Name of the App

Translated Review : Text of the review

Sentiment : The review is good or not

Sentiment Polarity : The polarity of sentiment measure how negative or positive the context is.

Sentiment Subjectivity : The subjectivity of sentiment.

Data Cleaning

- In Data Cleaning we have to check for null values and duplicated values
- The Play Store App dataset has 5 features with null values.
- In handling null values, for reviews feature, we have to replace null values with median value.
- The Play Store App dataset has 483 duplicated values. we can remove them.
- The User Review dataset has 4 features with null values. So we can remove them.
- The User Review dataset has 7735 duplicated values.
- So we should remove those duplicated values.

```
# Checking for Null values for app data
app.isnull().sum().sort_values(ascending = False)
```

```
Rating            1474
Current Ver          8
Android Ver          3
Type                 1
Content Rating       1
```

```
[13] # Checking null values for Review Data
     review.isnull().sum().sort_values(ascending = False)
```

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

```
# Checking for duplicated values
app.duplicated().value_counts()
```

```
False    10358
True       483
dtype: int64
```

```
# Checking duplicated values after removing null values
review.duplicated().value_counts()
```

```
False    29692
True      7735
dtype: int64
```

Data Manipulation

- Here, we can observe that Reviews, Installs, Price features has Object data type which is in string format. So we need to convert them into numerical values for our analysis.
- Also created a new column named ‘Revenue’ which is multiplication of price and number of installs for only paid apps.
- Also created a new feature named ‘year’ which contains the latest updated year of app.
- Also created a new column named ‘score’ which is based and positive and negative reviews for User Reviews dataset

```
# Inspecting Reviews column
new_data_app['Reviews'].dtype
```

```
dtype('O')
```

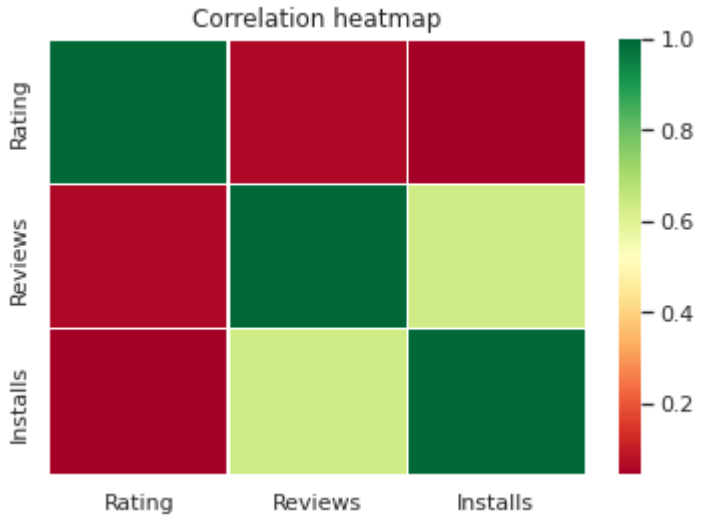
```
# Inspecting Installs Column
new_data_app['Installs'].dtype
```

```
dtype('O')
```

```
# Inspecting price column
new_data_app['Price'].dtype
```

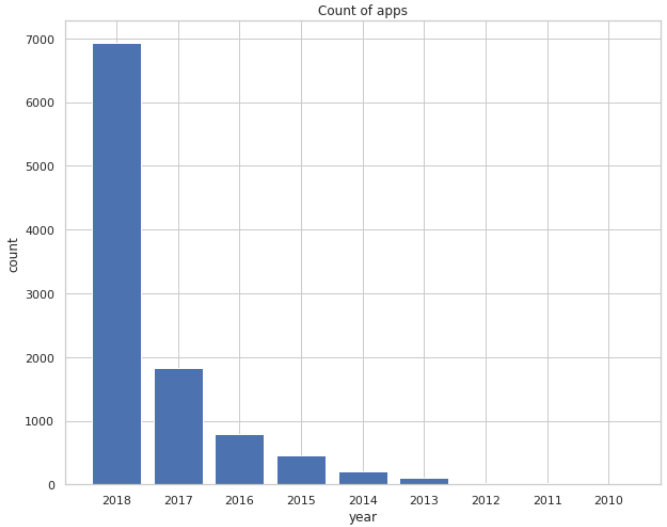
```
dtype('O')
```

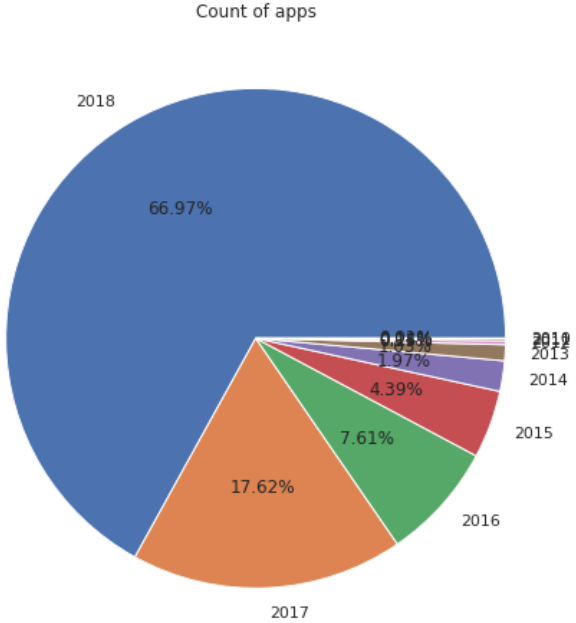
Exploratory Data Analysis :



From the above heatmap, we can say that there are no variables with multicollinearity in the data

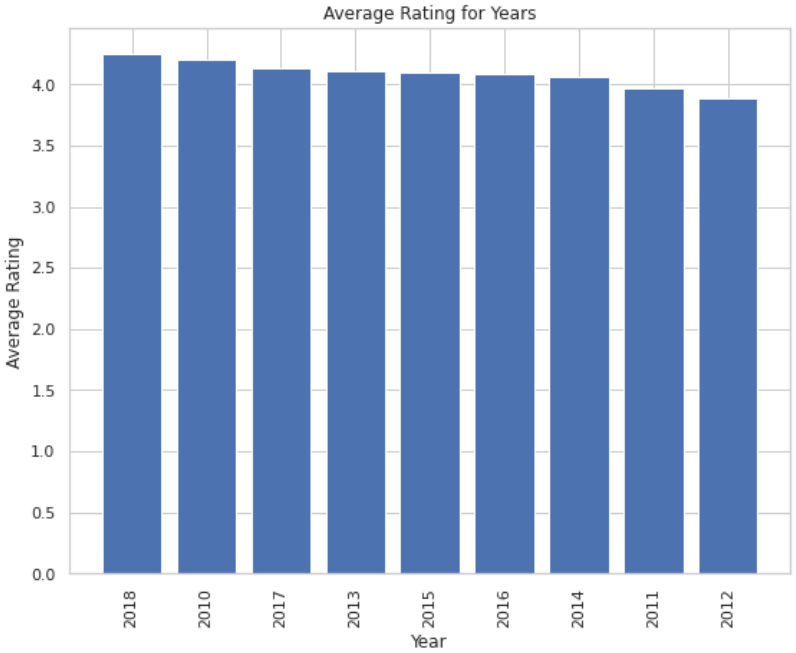
Year Analysis:

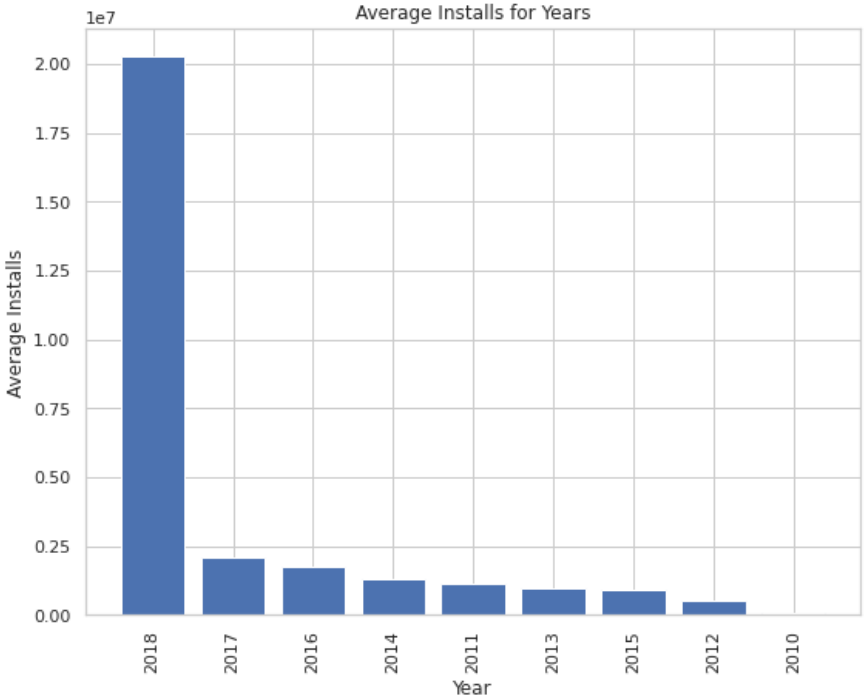




From the above charts, we can observe that most of the apps(66.97%) are updated in 2018.

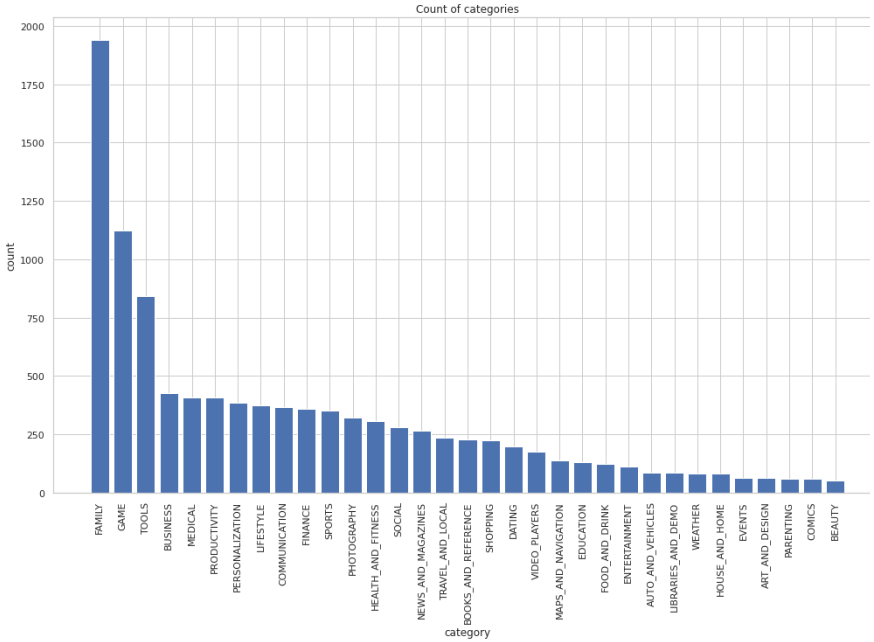
Year Analysis Based On Ratings And Installs:

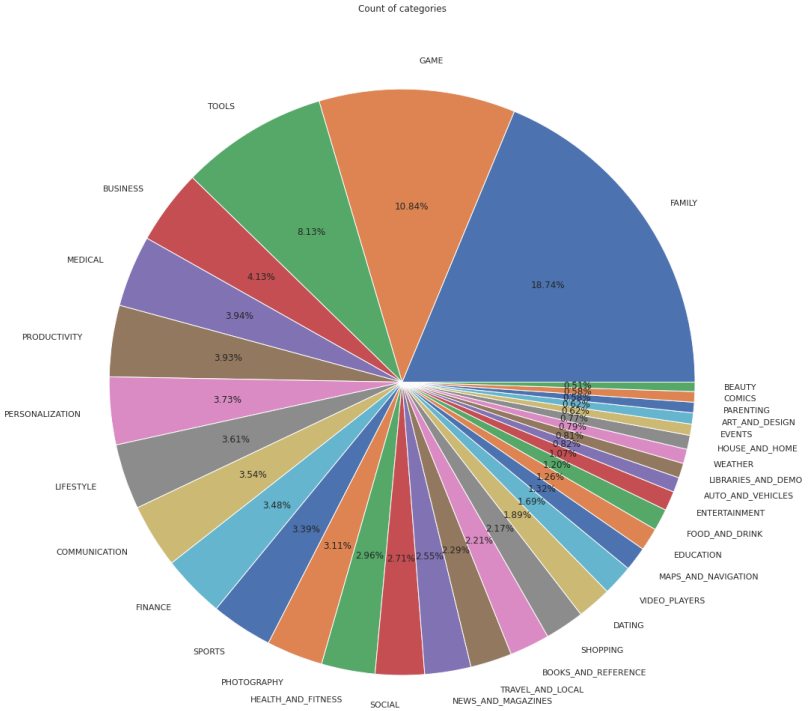




From the above charts, we can observe that 2018 apps have more number of installs and high ratings..

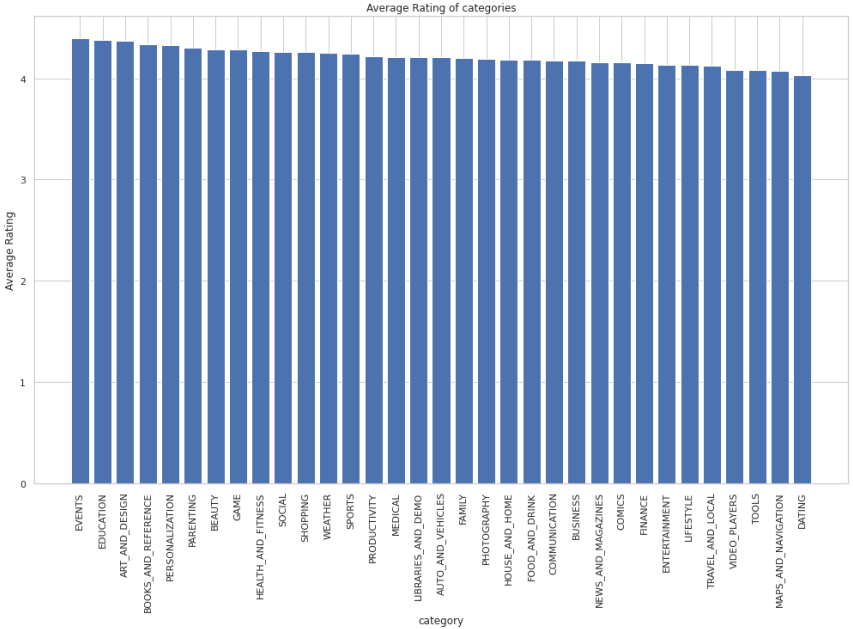
Category Analysis:

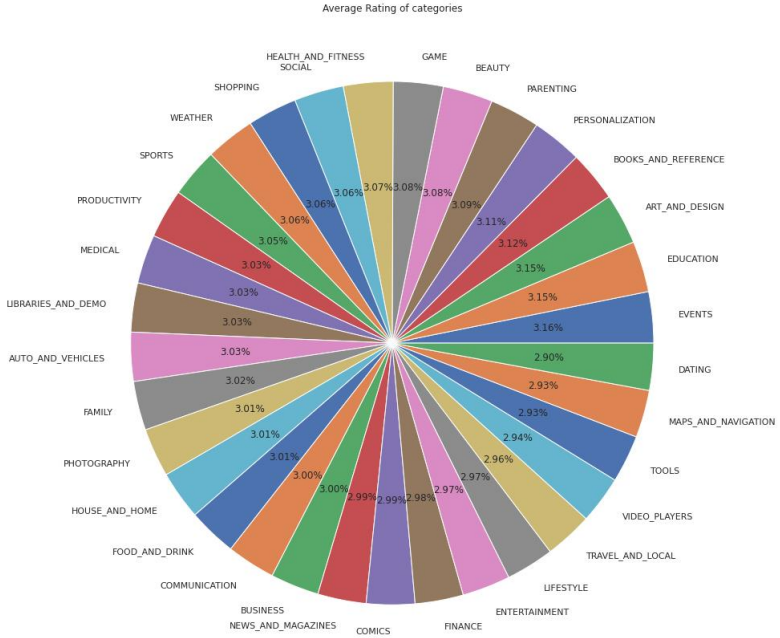




From the above charts, there are total of 33 categories and the most apps are from FAMILY category followed by GAME category and the least is BEAUTY category

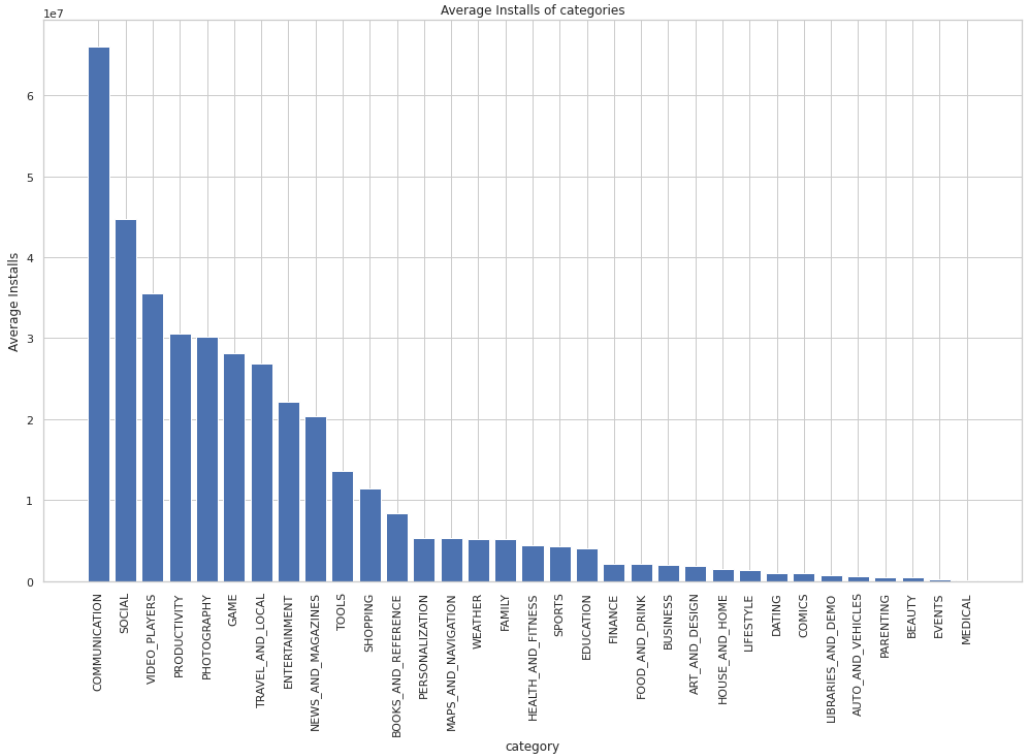
Category Analysis Based on Ratings :





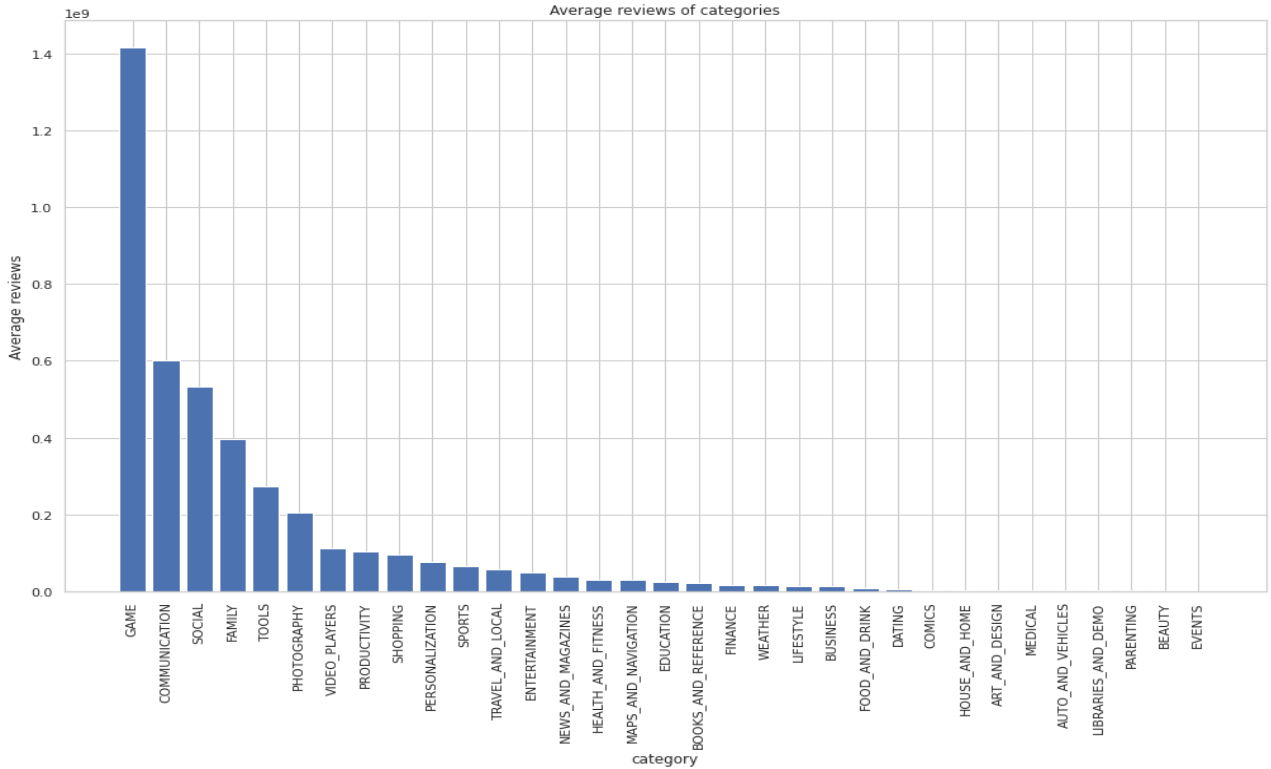
From the above charts, we can say that EVENTS, EDUCATION, ART_AND_DESIGN, BOOKS_AND_REFERENCE, PERSONALIZATION are the categories with highest average ratings

Category Analysis Based on Installs :



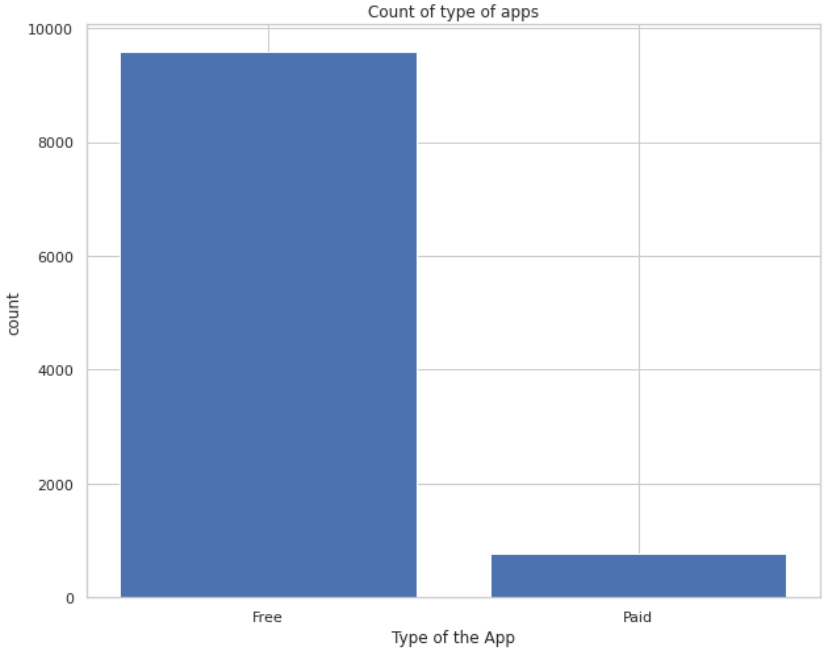
From the above plot we can observe that COMMUNICATION category has highest installs followed by SOCIAL and the least is MEDICAL

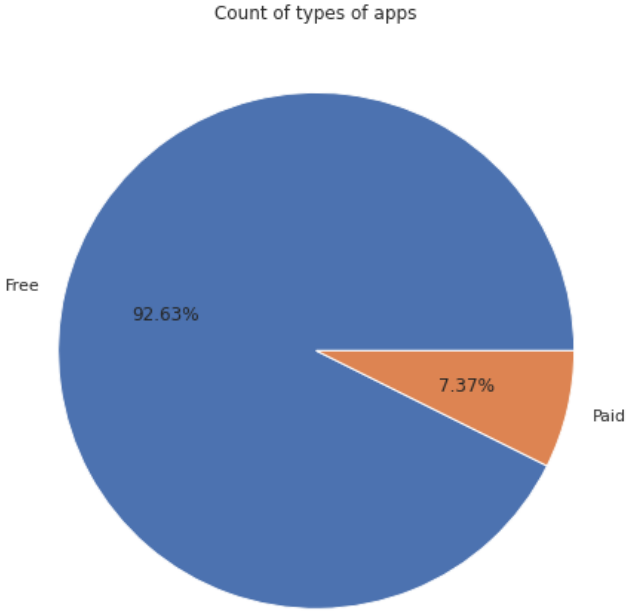
Category Analysis Based on Reviews :



From the above charts we can observe that GAME followed by COMMUNICATION category has high reviews and least is EVENTS.

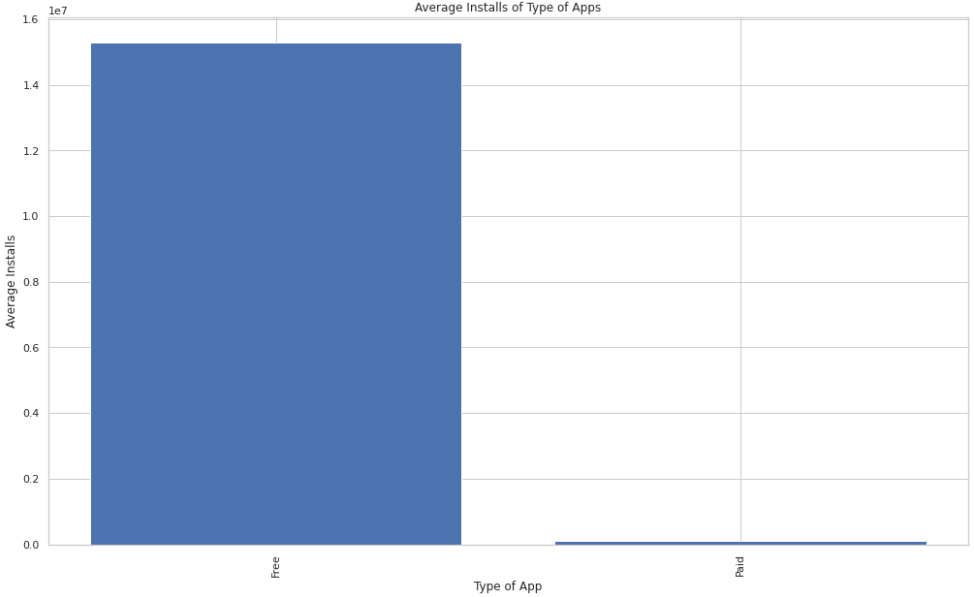
Type of Apps Analysis :

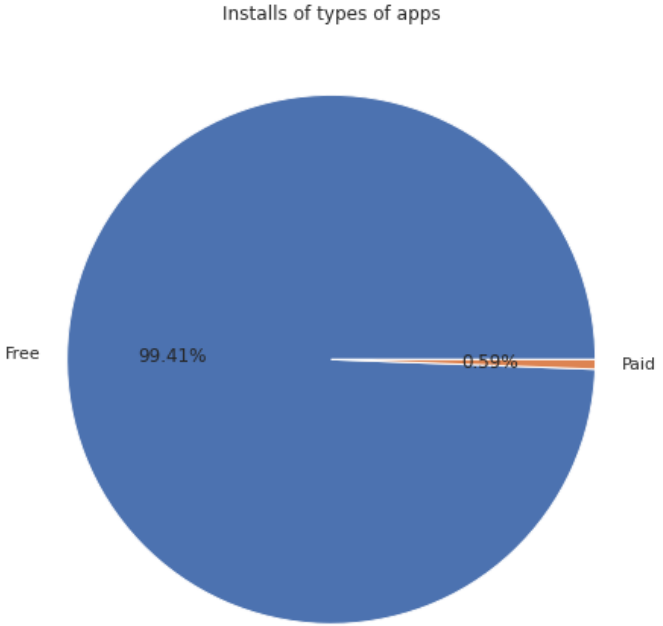




From the above chart we can observe that most of the apps are free (92.63%) and only 7.37% of apps are paid apps.

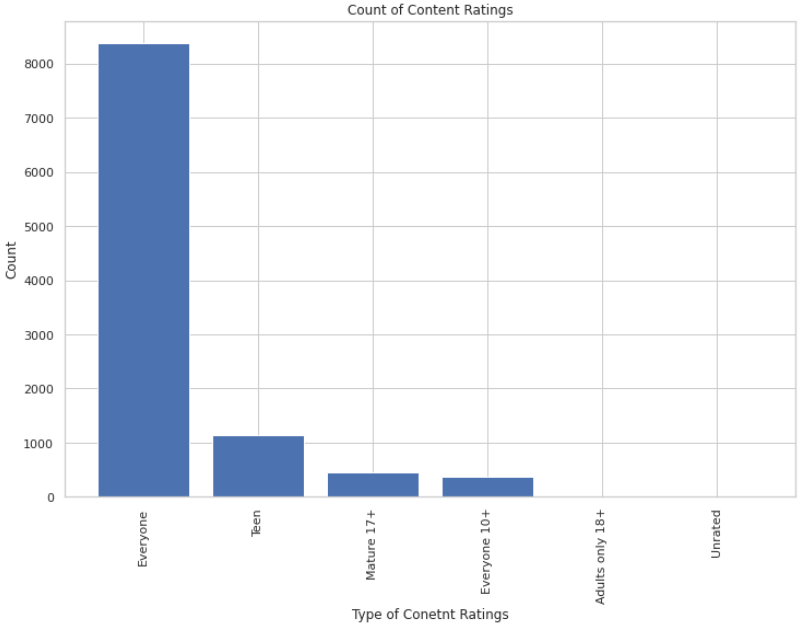
Type of Apps Analysis Based on Installs:

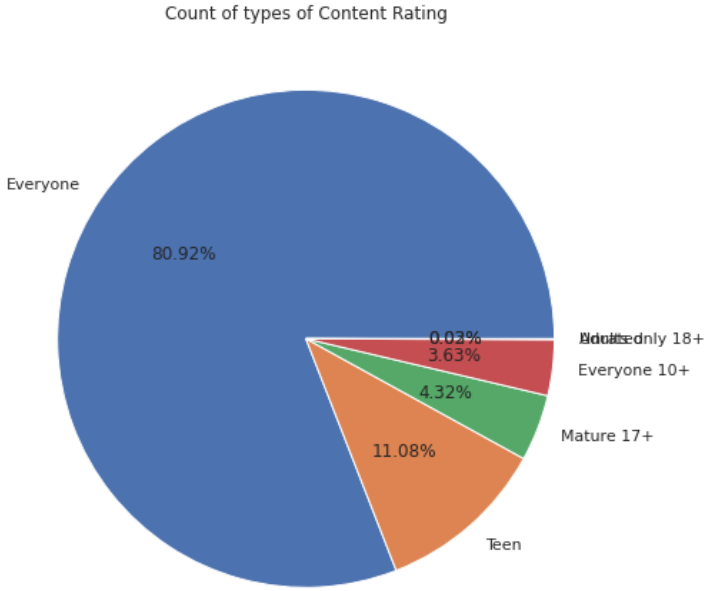




From the above visualizations, we can say that Free Apps has high installs (99.41%) than paid apps (0.59%)

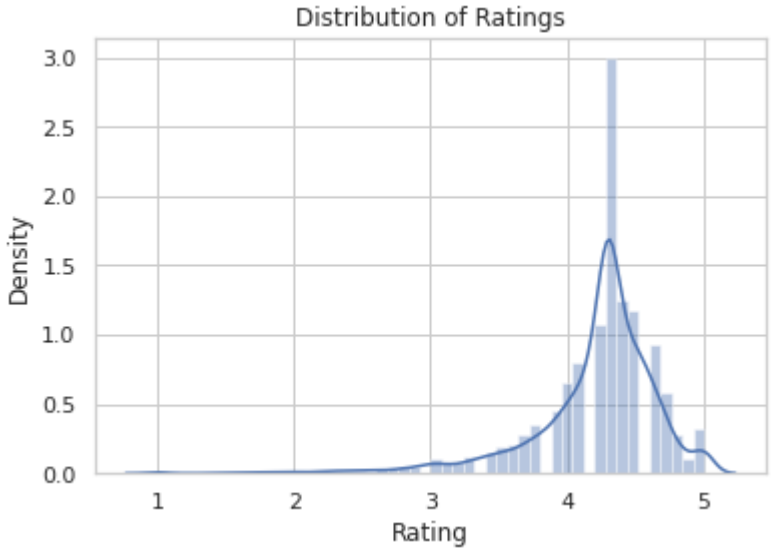
Type of Content Rating Analysis :





From the above chart we can say that Everyone content rating apps are more than others. That means 80.92% apps are for everyone.

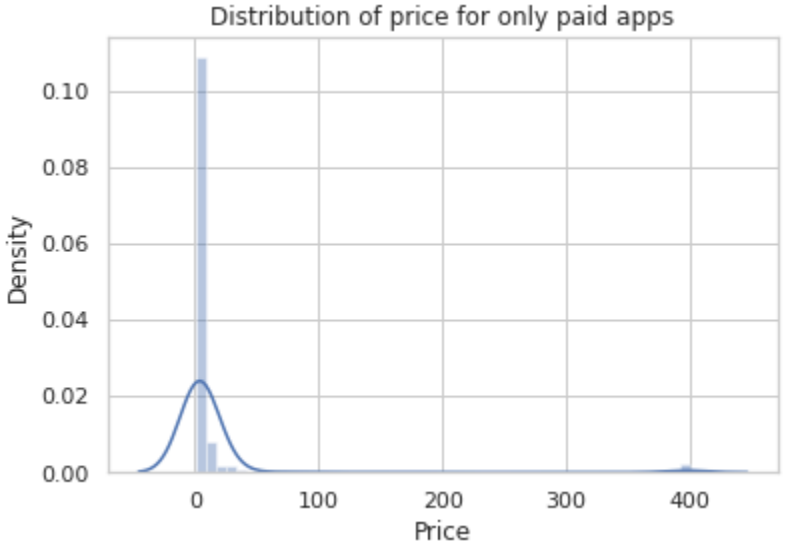
Rating Analysis :



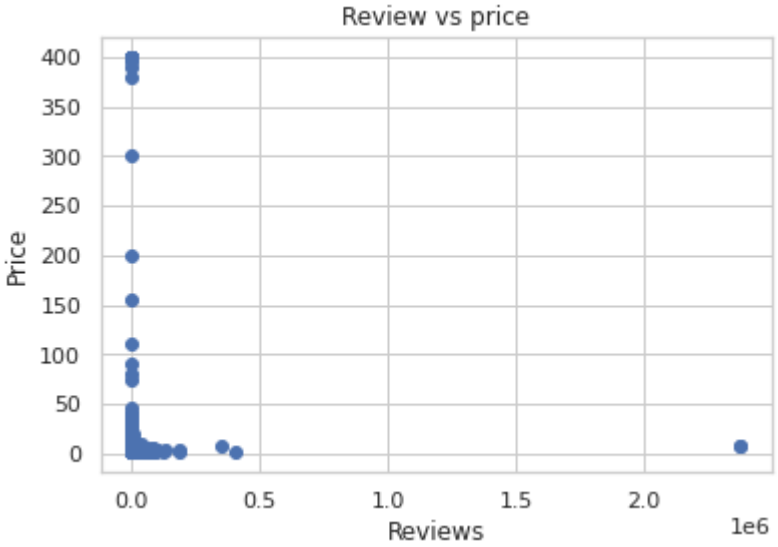
From the above chart we can say that :

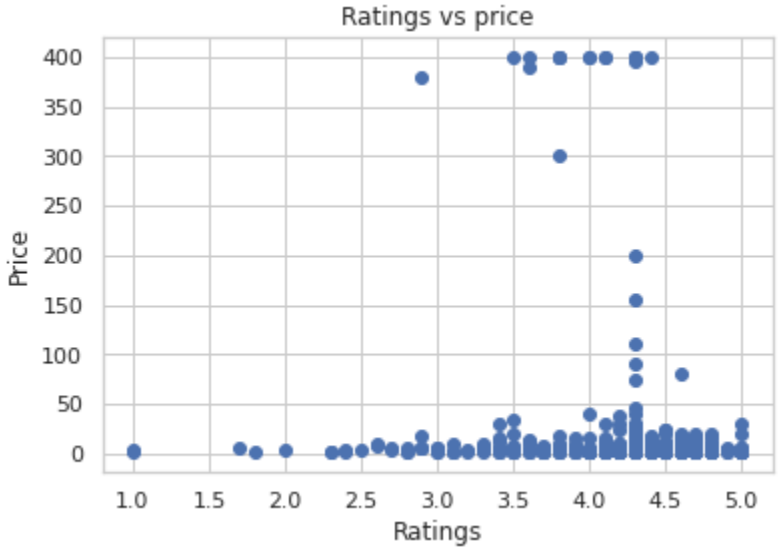
- **The above distribution plot is negatively skewed.**
- **The most number of ratings are in range of 4 to 5.**

Price Analysis :



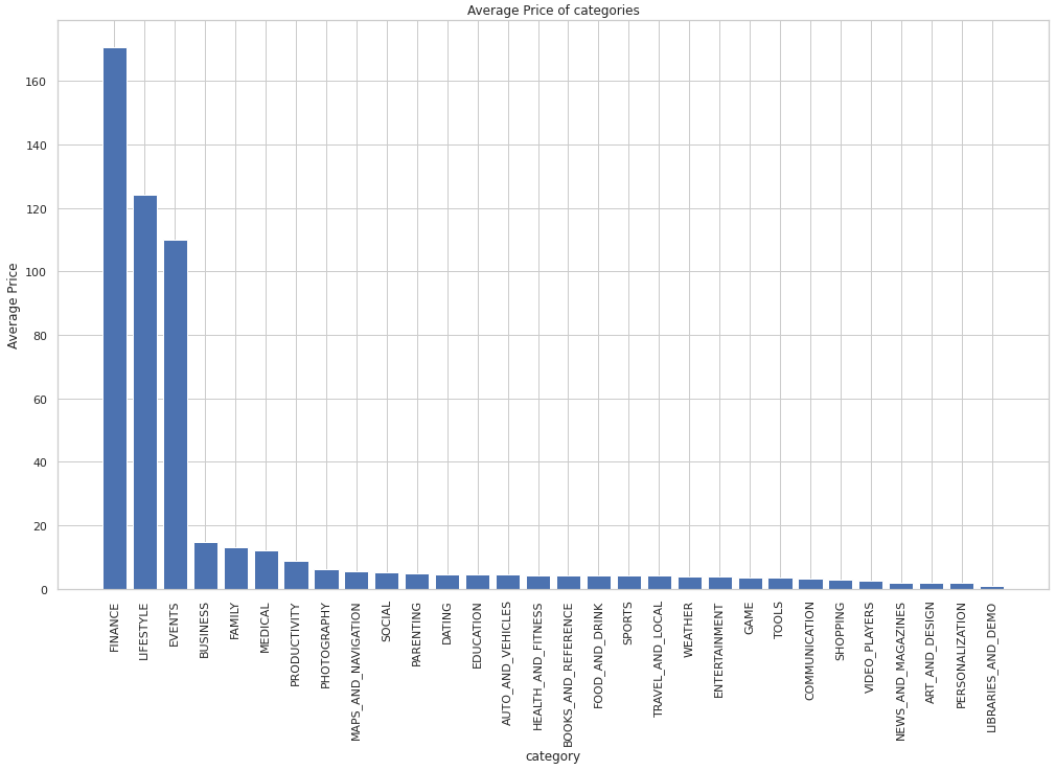
From the above distribution plot, we can say that most prices are in the range of 0 to 50 and a very few are in the range of 50 to 450





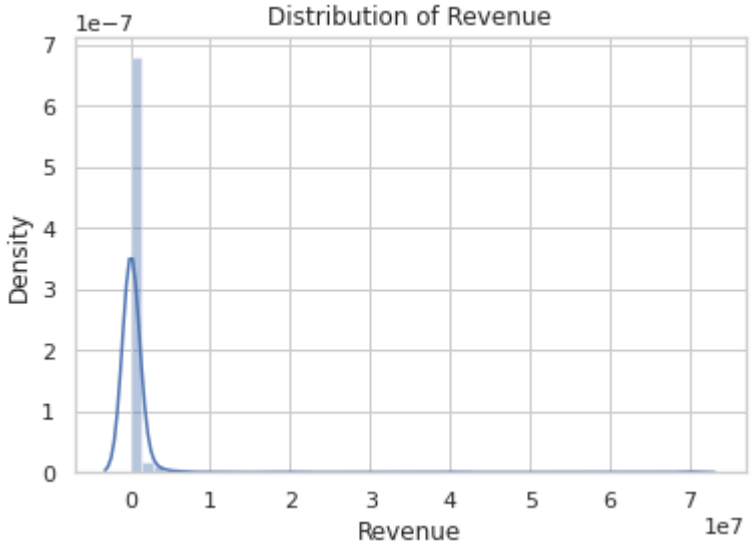
From the above scatter plots, we can observe that

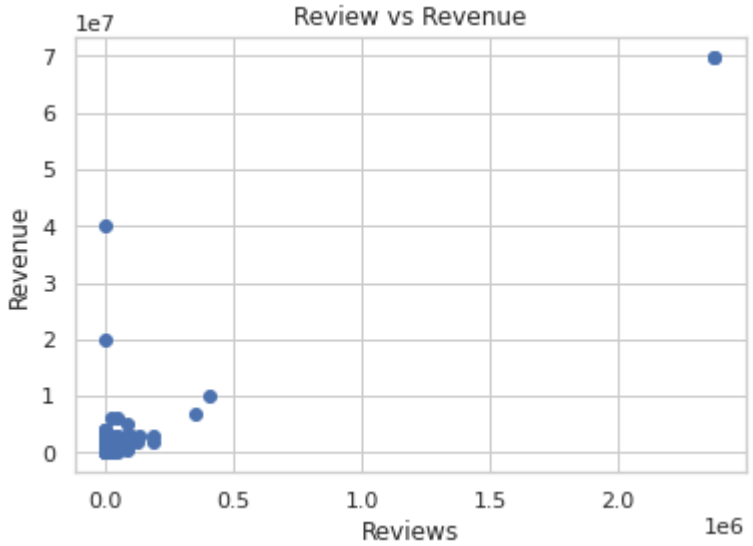
- **Most of the apps has prices of less than 50 USD**
- **And very few apps has high prices.**



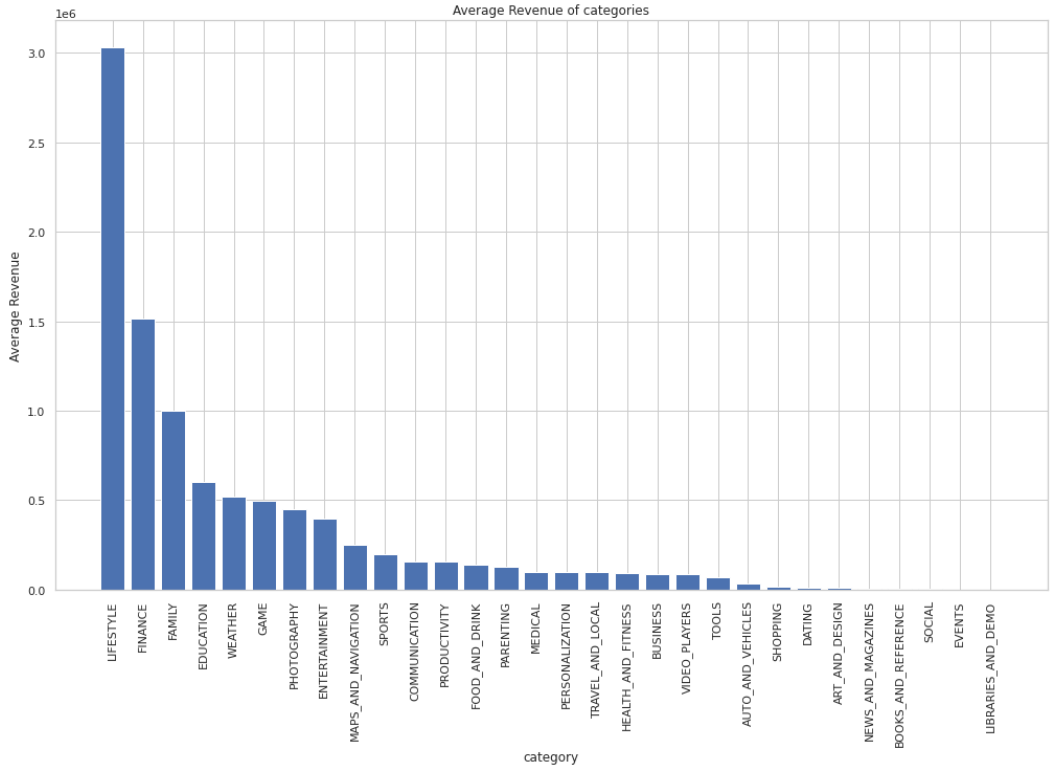
From the above visualization, FINANCE, LIFESTYLE category has high average prices and the least is SOCIAL.

Revenue Analysis Of Paid Apps:

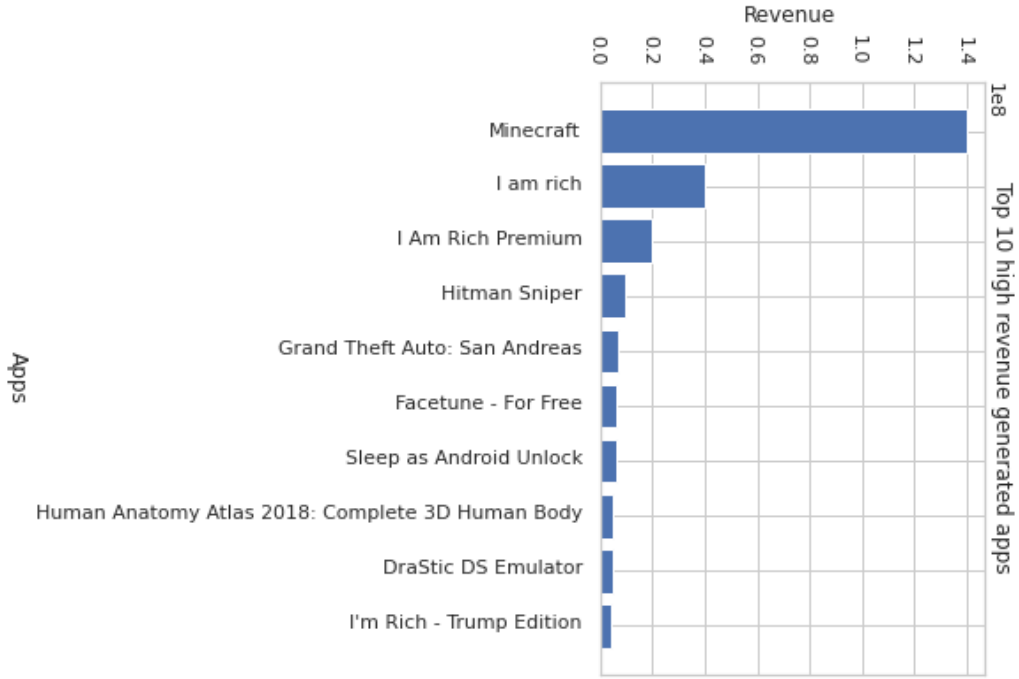




From the above plots we can observe that very few paid apps have highest revenue. Most of the paid apps have some decent amount of revenue.

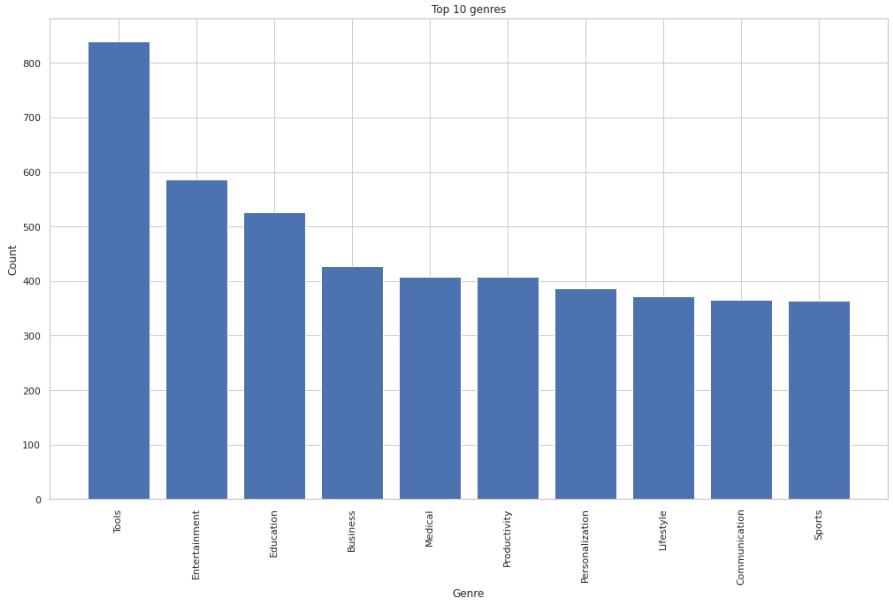


From the above visualizations, LIFESTYLE category has highest average revenue.



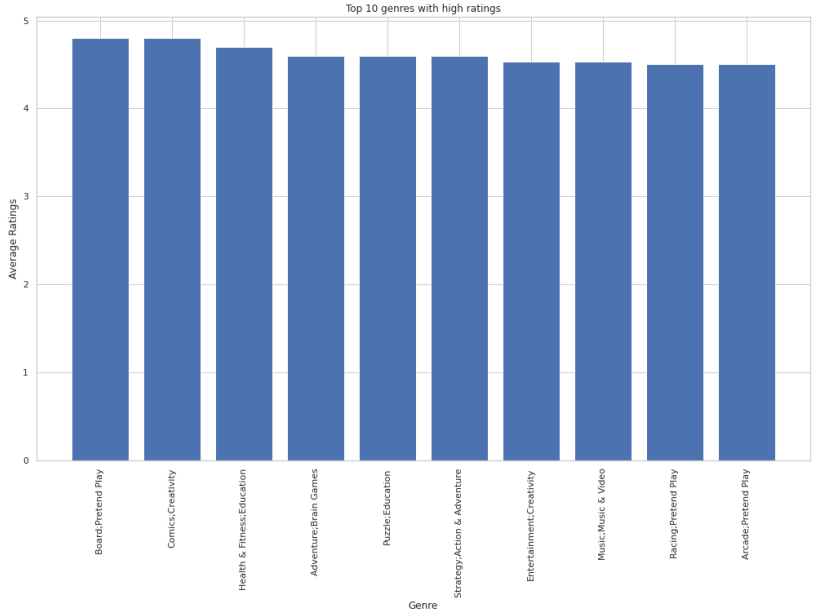
- **From the above visualization, The top 5 high revenue apps are Minecraft, I am rich, I Am Rich Premium, Hitman Sniper, Grand Theft Auto: San Andreas.**

Genre Analysis :



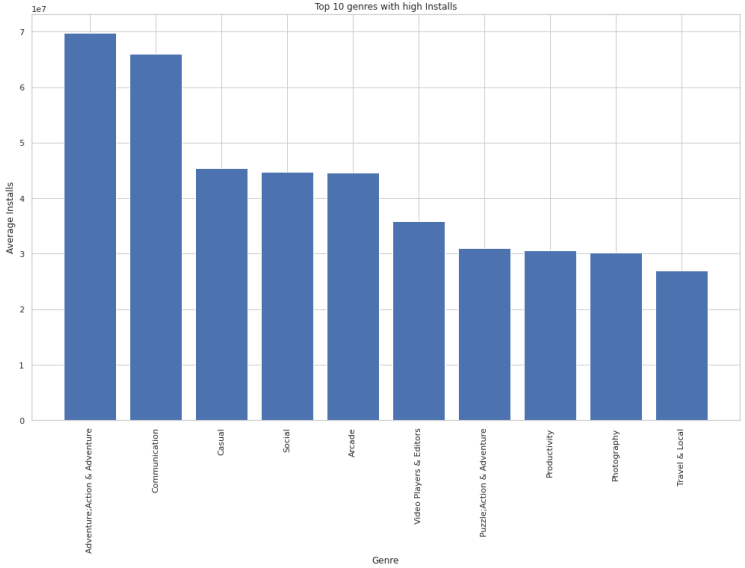
- There are total 115 genres exists in the data
- The above bar graph displays top 10 genres with most number of apps.
  And they are Tools, Entertainment, Education, Business, Medical, Productivity, Personalization, Lifestyle, Communication, Sports.

Genre Analysis Based On Ratings :



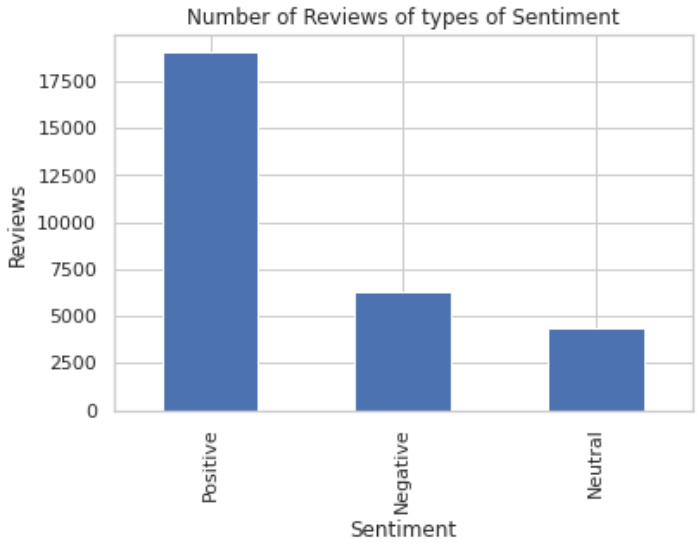
- From the above visualization, we can see the genres with highest ratings.
- They are Board;Pretend Play, Comics;Creativity, Health & Fitness;Education, Adventure;Brain Games, Puzzle;Education, Strategy;Action & Adventure, Entertainment;Creativity, Music;Music & Video, Racing;Pretend Play, Arcade;Pretend Play.

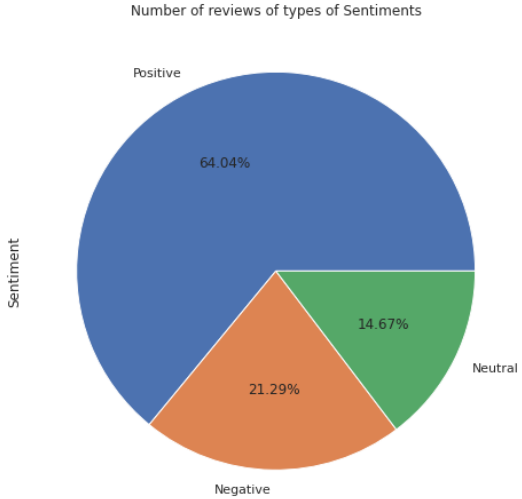
Genre Analysis Based On Installs :



- From the above visualization, we can see the genres with highest installs.
- The top 10 genres with high installs are Adventure;Action & Adventure, Communication, Casual' 'Social, Arcade, Video Players & Editors, Puzzle;Action & Adventure, Productivity, Photography, Travel & Local.

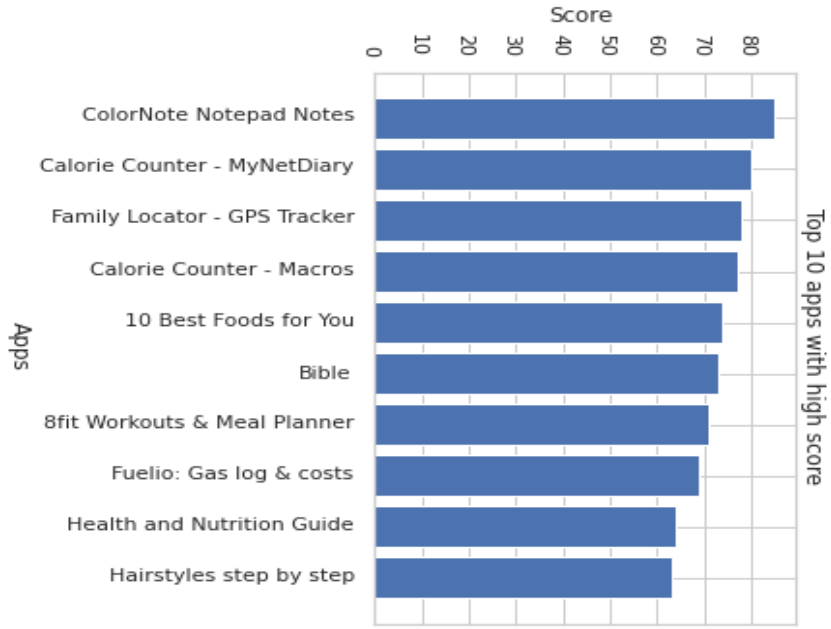
Sentiment of Reviews Analysis :





- **From the above visualizations, most reviews are positive i.e., 64.04%, 21.29% reviews are Negative, and 14.67% reviews are Neutral.**

Top 10 Popular Apps :



- From the above visualization, we can observe the apps with more positive reviews.
- ColorNote Notepad Notes App has high positive reviews.

Final Conclusions :

- Most apps are from FAMILY category followed by GAME category and the least is BEAUTY
- The top 5 Categories with high ratings are EVENTS, EDUCATION,ART_AND_DESIGN, BOOKS_AND_REFERENCE, PERSONALIZATION
- COMMUNICATION category has highest installs and the least is MEDICAL
- The top 5 Categories with high installs are COMMUNICATION, SOCIAL, VIDEO_PLAYERS, PRODUCTIVITY, PHOTOGRAPHY
- The top 5 Categories with high reviews are GAME, COMMUNICATION, SOCIAL, FAMILY, TOOLS.
- Most of the apps(66.97%) are updated in 2018.
- 2018 updated apps have higher rating followed by 2010 and least is 2012.

- 2018 updated apps have higher Installs followed by 2017 and least is 2010.
- 2018 updated apps have higher Reviews followed by 2017 and least is 2010.
- Most of the apps are free (92.63%) and only 7.37% of apps are paid apps.
- Free Apps has high installs (99.41%)than paid apps (0.59%)
- Most number of ratings are in range of 4 to 5 and most prices of paid apps are in the range of 0 to 50 USD *and a very few are in the range of* 50 to 450 USD.
- A very few paid apps have highest revenue. Most of the paid apps have some decent amount of revenue.
- LIFESTYLE category has highest average revenue.
- FINANCE, LIFESTYLE category has high average prices and the least is SOCIAL.
- The top 5 revenue apps are Minecraft, I am rich, I Am Rich Premium, Hitman Sniper, Grand Theft Auto: San Andreas.
- There are total 115 genres exists in the data. The top 10 genres with most number of apps. And they are Tools, Entertainment, Education, Business, Medical, Productivity, Personalization, Lifestyle, Communication, Sports.

- Comics;Creativity and Board; Pretend play genres have highest rating and the Parenting;Brain Games genre has least ratings.
- Adventure;Action & Adventure and Communication genres have high installs and Board; Pretend Play genre has very few installs
- Facebook app has higher reviews.
- Most of the reviews are positive i.e., 64.04%, 21.29% reviews are Negative, and 14.67% reviews are Neutral.
- The most popular apps (more positive reviews) are ColorNote Notepad Notes, Calorie Counter - MyNetDiary, Family Locator - GPS Tracker, Calorie Counter - Macros, 10 Best Foods for You, Bible, 8fit Workouts & Meal Planner, Fuelio: Gas log & costs, Health and Nutrition Guide, Hairstyles step by step

THANK YOU