# Capstone Project

## Play Store App Review Analysis


**By**

**Vadlamani Shivani**

**Saiteja Ch**

**Data Science Trainee, AlmaBetter**

# WHY ANALYZE THE GOOGLE PLAY STORE?

Mobile App Market is set to grow 20% by 2023

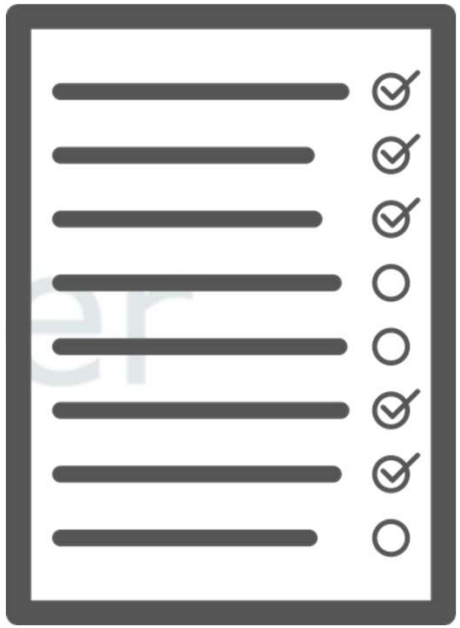
Android Apps comprise 90% of the Mobile App Market


What makes an App popular? Can we predict how popular it's going to be?

What are some interesting patterns in user behavior related to app usage & feedback

# Content

- Introduction
- Problem Statement
- Dataset Preparation
- Attributes in our dataset (Play store and user reviews)
- Overview of analysis
- Correlation Heat map
- Paid apps vs Free apps based on genre
- Content Rating as per category
- Count of applications in each genre
- Category of apps having most number of installs
- Finding top 10 apps based on ratings and reviews
- Percentage of review sentiments
- Finding Sentiment subjectivity proportional to sentiment polarity?
- Conclusion: Insights we found while analyzing the whole data



# Introduction

- Google play store is engulfed with a few thousands of new applications regularly with a progressively huge number of designers working freely or on the other hand in a group to make them successful, with the enormous challenge from everywhere throughout the globe. Since most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, adverts and memberships add to the achievement of an application.

- Google Play was launched on March 6, 2012, bringing together Android Market marking a shift in Google's digital distribution strategy .

- Android is the dominant mobile operating system today more than 85% of all mobile devices running Google's OS. The Google Play Store is the largest and most popular Android app store.
- There are more than 3.04 million apps found on Google Play Store.

- Actionable insights can be drawn for developers to work on and capture the Android market. The main goal of our project is-

1) The purpose of our project is to gather and analyze detailed information on apps in the Google Play Store in order to provide insights on app features and the current state of the Android app market.
2) The Objective of the project to Explore and analyze the data to discover key factors responsible for app engagement and success.

# Problem Statement

- Two datasets are provided, one with **basic information** and the other with **user reviews** for the respective app.
- We must examine and evaluate the data in both datasets in order to identify the important characteristics that influence app engagement and success.

## So, what factors influence an app's success?

An app is said to be successful if it has:

- A high average user rating
- A good number of positive reviews
- A good number of monthly average users
- High revenue per customer and so on.

# Dataset Preparation

- **Import Libraries:** NumPy, Pandas, Seaborn and Matplotlib
- **Loading the data sets:** Two datasets, First Play store app dataset and User Reviews dataset.
- **Data cleaning:** Null values, Finding and removing Outliers, Removing duplicate data.
- **Data preparation:** Filling the missing categorical values with mode and numerical values with median. Conversion of price, installs, reviews into numerical values.
- **Exploratory Data Analysis:** Analyzing the data sets to summarize their main characteristics using statistical graphics and data visualizations method.

# Attributes in Google Play store Data

1. **App :** This column Contains the name of the app for each observation.

2. **Category :** This column Contains Category to which the app belongs.

3. **Rating :** This column contains the average rating for the app.

4 .**Reviews :** This column contains the number of reviews that the app has received on the play store.

5. **Size :** This column contains the amount of memory the app occupies on the device.

6. **Installs :** This column contains the number of times that the app has been downloaded and installed from the play store.

7. **Type :** This column contains the information whether the app is free or paid.

8. **Price:** If the app is a paid app, this column contains the data about its price.

9. **Content Rating:** This column contains the maturity rating of the app i.e. the age group of the audience for which it is suitable.

**10.Genres:** This column contains the data about to which genre the app belongs. Genres can be considered as a further division of the group of Category.

**11.Last Updated:** Contains the date on which the latest update of the app was released.

**12.Current Version:** Contains information on the current version of the app available on the play store.

**13.Android Version:** Contains information about the android versions on which the app is supported

# Attributes in User reviews

**1. App :** It contains the name or identifier of the app.

**2. Translated Review :** It contains the text of the review

**3. Sentiment :** It contains the overall sentiment of the review (e.g. positive or negative)

**4. Sentiment Polarity :** It contains a measure of the positivity or negativity of the review

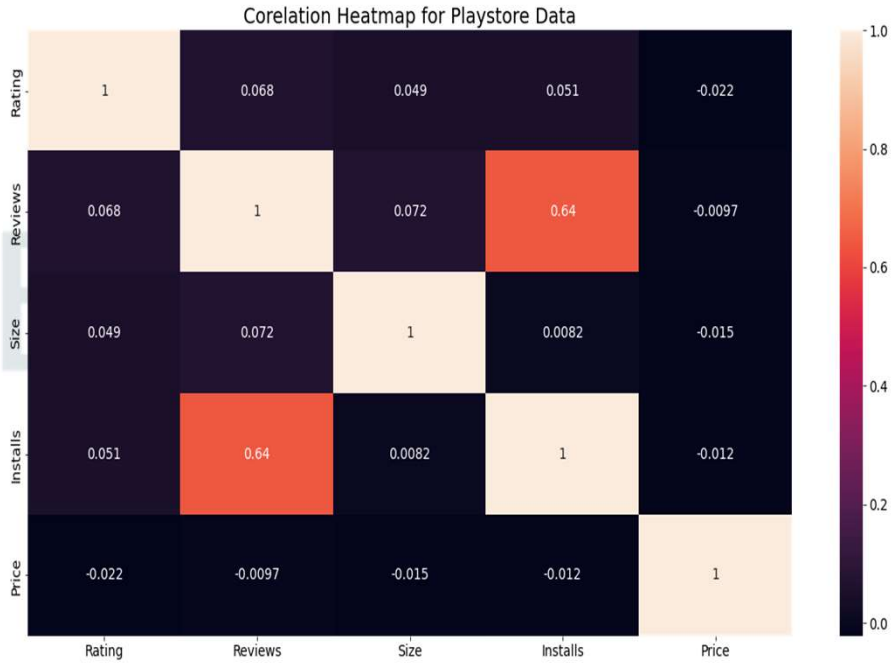
**5. Sentiment Subjectivity :** It contains a measure of the subjectivity or objectivity of the review. These columns could be useful for analyzing the sentiment of app reviews and understanding how users feel about different apps.

# OVERVIEW OF ANALYSIS

| Data Cleaning | Data Exploration | Predictive Modeling |
| --- | --- | --- |
|  |  |  |
| Understand the structure of the dataset and clean data before analysis | Uncover initial patterns, characteristics, and points of interest using visual exploration | Formulate a statistical model to forecast an outcome using relevant predictors |

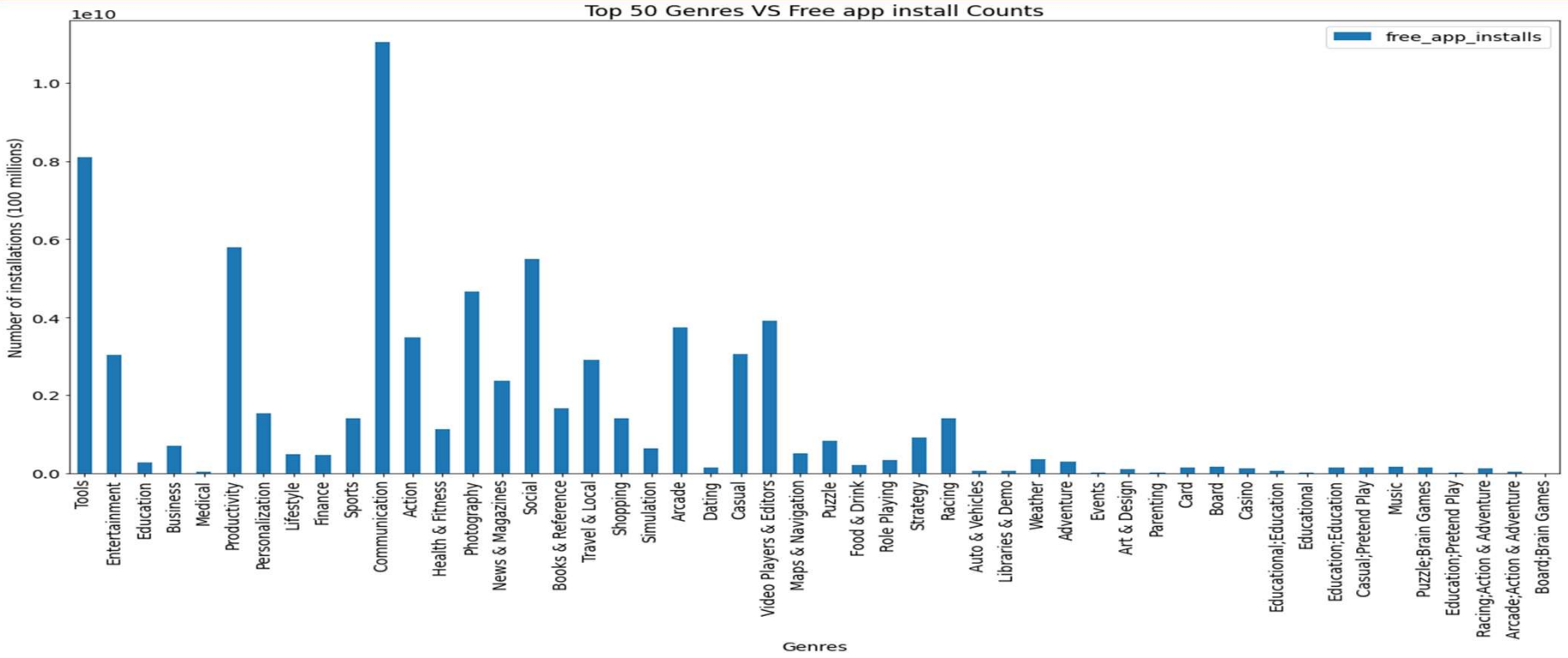
# Correlation Heatmap

- There is a majority **positive** correlation between the **Reviews** and **Installs**.
  - The Price is slightly **negatively** correlated with the **Rating**, **Reviews**, and **Installs**.
- The **Rating** is slightly **positively** correlated with the **Installs** and **Reviews**.

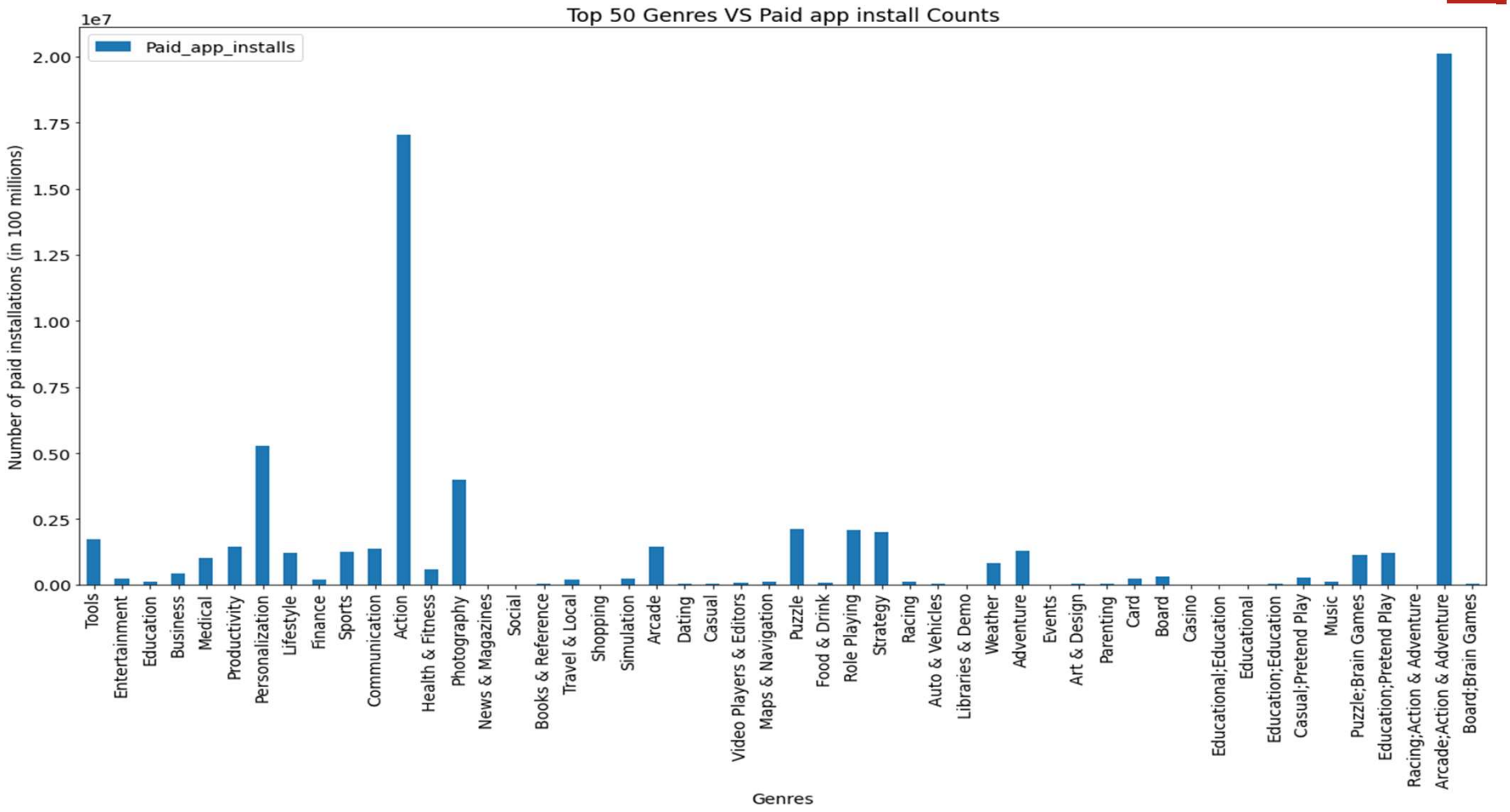


# Paid apps v/s Free apps based on Genre

We Observed that maximum number of apps installed and app count are  in Free apps.



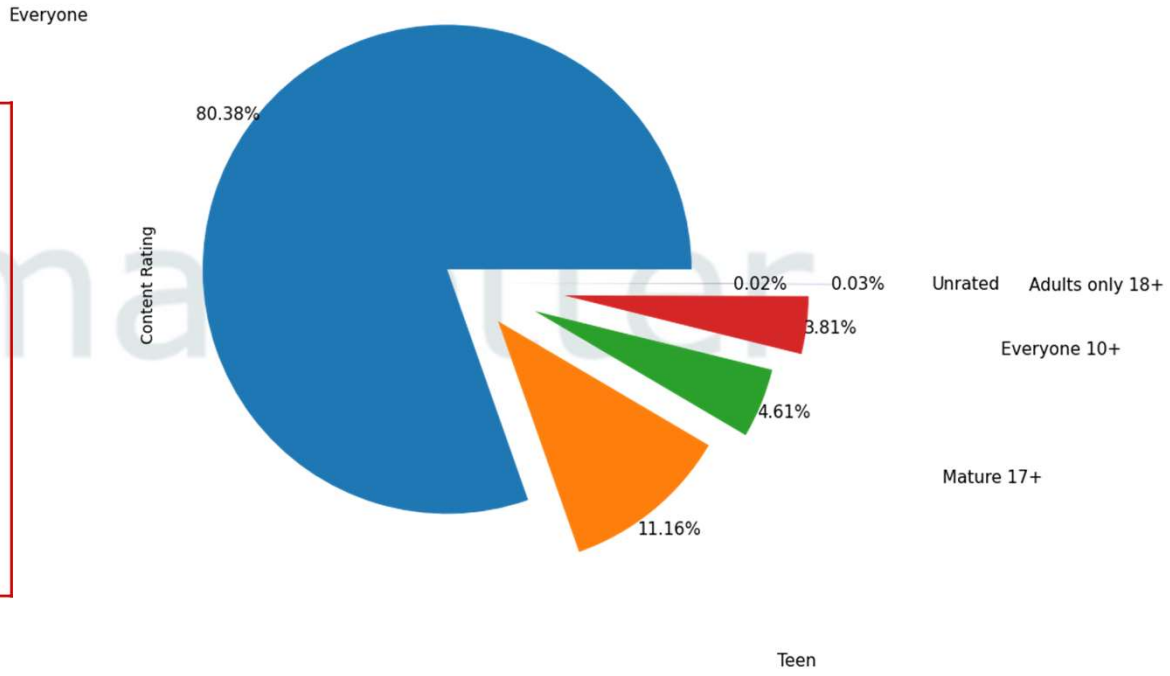
# Paid apps


Top 50 Genres VS Paid app install Counts
1e7
Paid_app_installs
Number of paid installations (in 100 millions)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Tools
Entertainment
Education
Business
Medical
Productivity
Personalization
Lifestyle
Finance
Sports
Communication
Action
Health & Fitness
Photography
News & Magazines
Social
Books & Reference
Travel & Local
Shopping
Simulation
Arcade
Dating
Casual
Video Players & Editors
Maps & Navigation
Puzzle
Food & Drink
Role Playing
Strategy
Racing
Auto & Vehicles
Libraries & Demo
Weather
Adventure
Events
Art & Design
Parenting
Card
Board
Casino
Educational;Education
Educational
Education;Education
Casual;Pretend Play
Music
Puzzle;Brain Games
Education;Pretend Play
Racing;Action & Adventure
Arcade;Action & Adventure
Board;Brain Games
Genres

# Content Rating

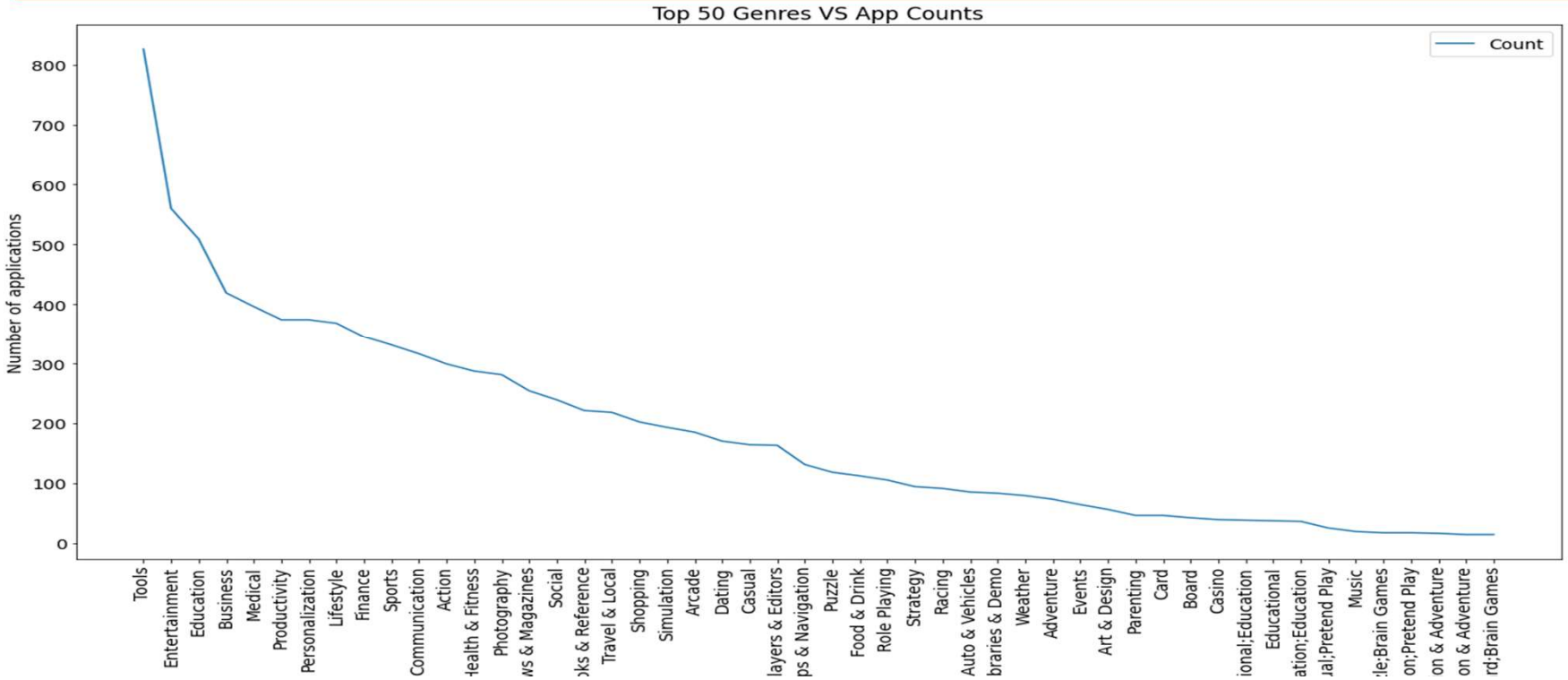
From the above plot we can see that Everyone category having majority of apps count.

A majority of the apps **(80.38%)** in the play store are can be used by everyone. The remaining apps have various age restrictions to use it.



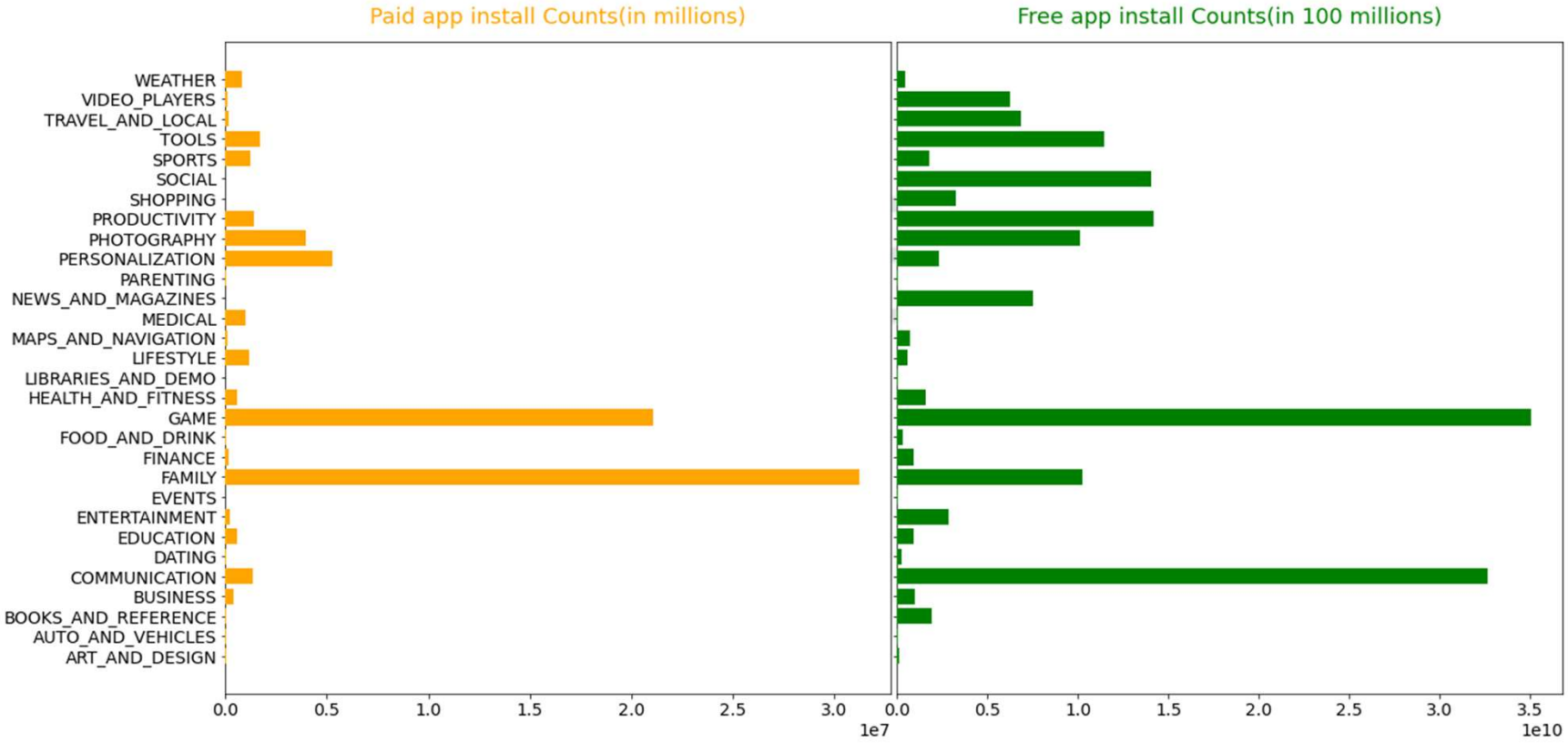
# Count of Applications in each Genre

Surprisingly **Tools, Business and Medical** apps are also at the Top Count of applications along with Entertainment**.**



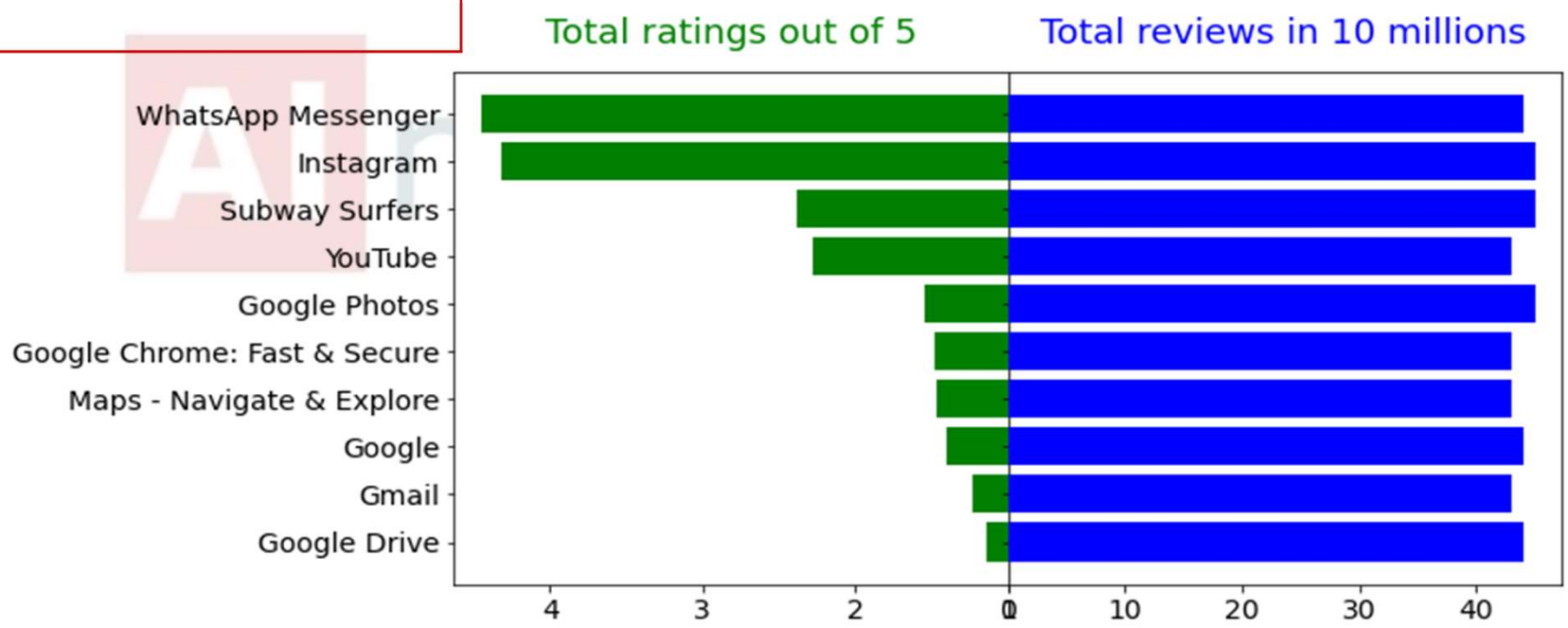
# Category App's have most number of installs

The Game, Communication and Tools
categories has the highest number of installs compared to other categories of apps.

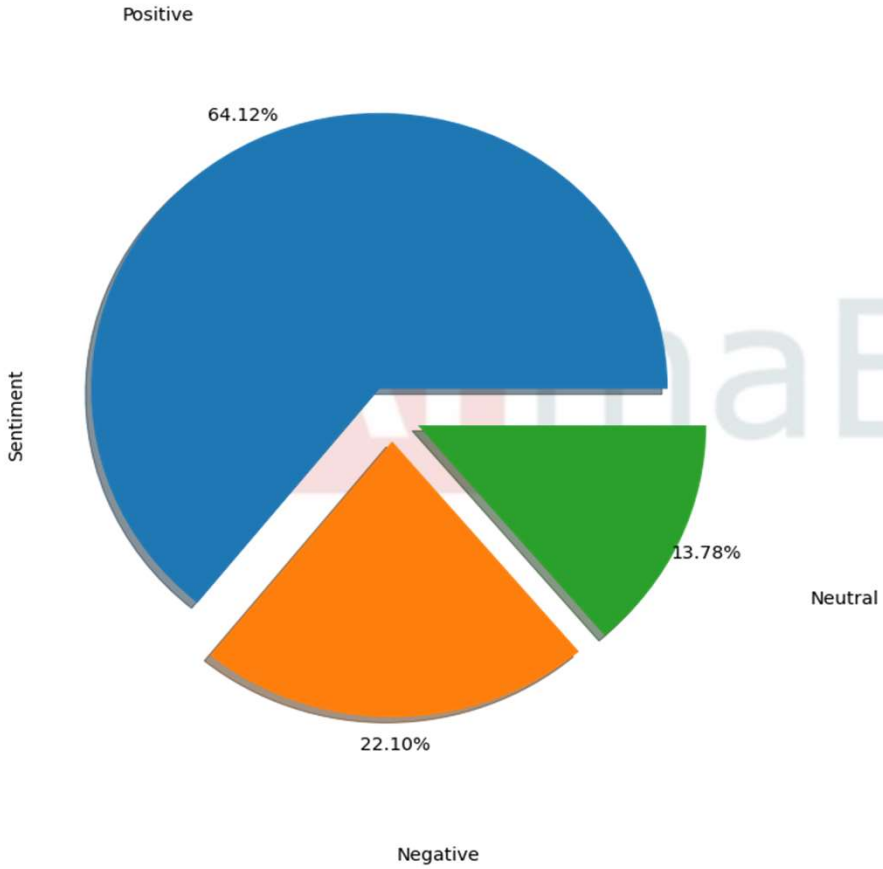


# Average rating of the apps based on some category

**Communication** , **Social** , **games** are the most rated among all categories and genres



# Percentage of Review Sentiments:

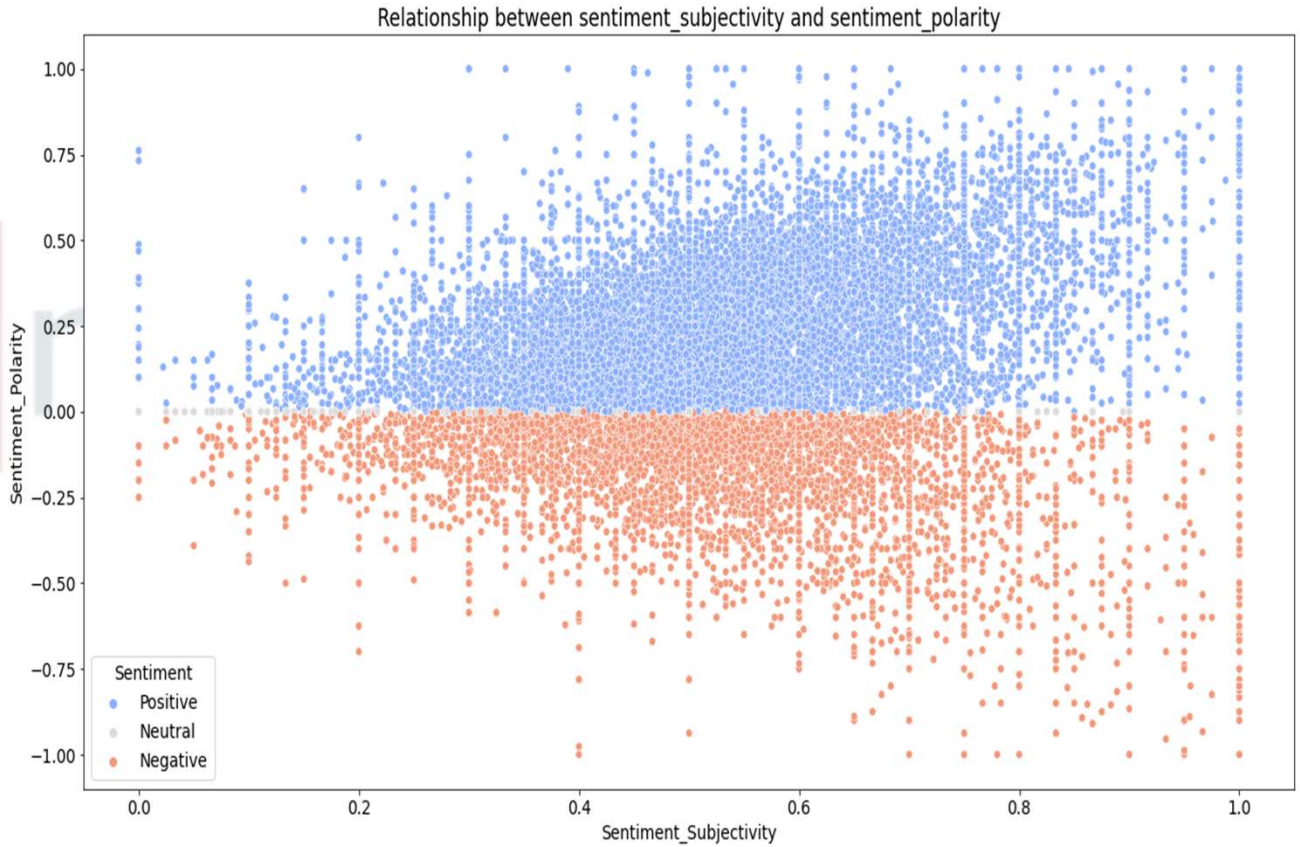



The number of **Unique** Apps from Play store and User reviews merged dataset are **816.**

From Sentiment column, **64%** are **Positive, 22%** are **Negative** and **14%** are **Neutral** values.

# Is sentiment_subjectivity proportional to sentiment_polarity?

From the above scatter plot it can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in maximum number of case, shows a proportional behavior, when variance is too high or low



# What have we Concluded so far ?

- **92.19%** apps are **Free** and 7.81% apps are paid in type.
- **81.80%** apps have **Everyone** content rating.
- **Events** category has a **highest mean rating of 4.39** and Dating category has lowest 4.05 rating.
- **Family, Game and Tools are top three** categories having 1906, 926 and 829 app count.
- Most competitive category: **Family**
- Category with the highest number of installs: **Game**
- Tools, Entertainment, Education, Business and Medical are top Genres.
- **Overall sentiment count** of merged dataset in which **Positive sentiment count is 64%, Negative 22% and Neutral 14%.**
- It's good to develop a **Free type** app and having a content rating for **Everyone.**
- Percentage of apps that are top rated = **81.80%**
- Price, Rating, Size **has no or very less correlation** with **Sentiment Polarity**.


Thank you!