Capstone Project-1

Play Store App Review Analysis

By
Vinayaka R Kulkarni
Prajwal Singh
Jyanathi Challagunda
Samarth Kushwaha

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?

Introduction

- Android is the most popular operating system in the world, with over 2.5 billion active users spanning over 190 countries.
- 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 **Google play store data set** and the other with **user reviews data set** 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.

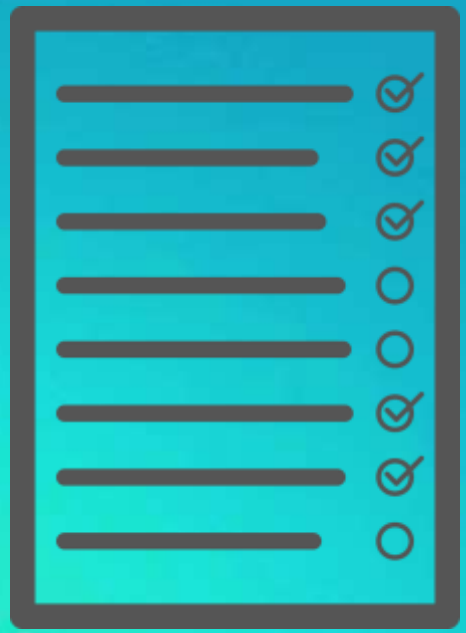
So, what factors influence an app's success?

An app is said to be successful if it has:

- When it has most number of installs and ratings.
- Which category apps are most installed.
- What is the percentage of installing free apps versus paid apps
- What are all the top categories apps.

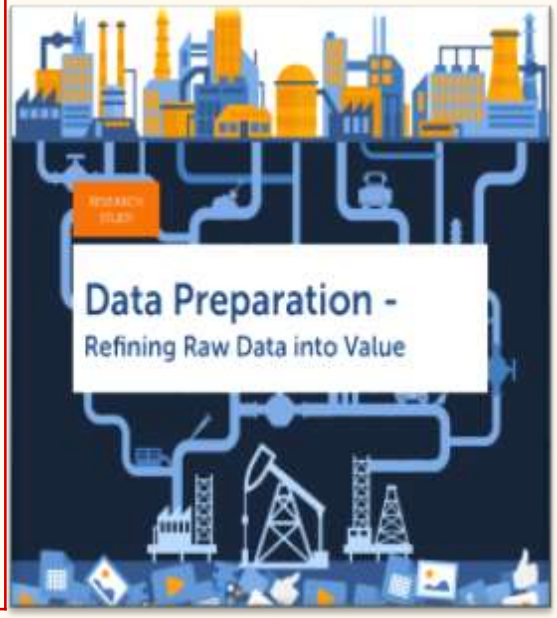
Agenda

- Introduction
- Category wise play store apps installs
- Category wise most popular apps
- Top 20 apps in play store considering all the parameters
- Average installs, category wise
- Most installed apps in communication category
- Average sizes of apps in each category
- Category wise percentage of paid apps
- Category wise top installed paid apps
- Category wise installed apps with content rating
- Percentage reviews sentiment distribution



Dataset Preparation

- **Loading the data sets:** Two datasets, First Play store app dataset and User Reviews dataset.
- **Import Libraries:** NumPy, Pandas, Seaborn and Matplotlib
- **Data cleaning:** Null values, Finding and removing Outliers, Removing duplicate data.
- **Data Imputation:** 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.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.
9.Last Updated: Contains the date on which the latest update of the app was released.
10.Current Version: Contains information on the current version of the app available on the play store.
11.Android Version: Contains information about the android versions on which the app is supported.

Attributes in User reviews

1. **App-** Application name
2. **Translated Review-** User review
3. **Sentiment-** Positive/Negative/Neutral
4. **Sentiment Polarity-** Sentiment polarity score
5. **Sentiment Subjectivity-** Sentiment subjectivity score

OVERVIEW OF ANALYSIS

Data Cleaning

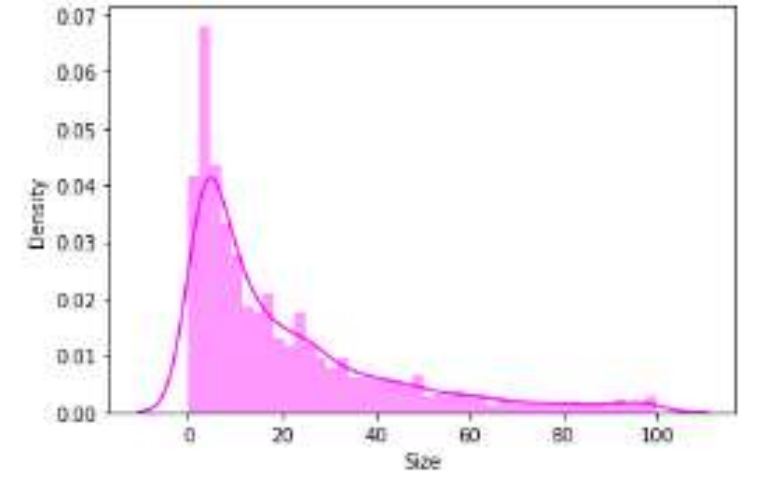
Understand the structure of the dataset and clean data before analysis

Data Exploration

Uncover initial patterns, characteristics, and points of interest using visual exploration

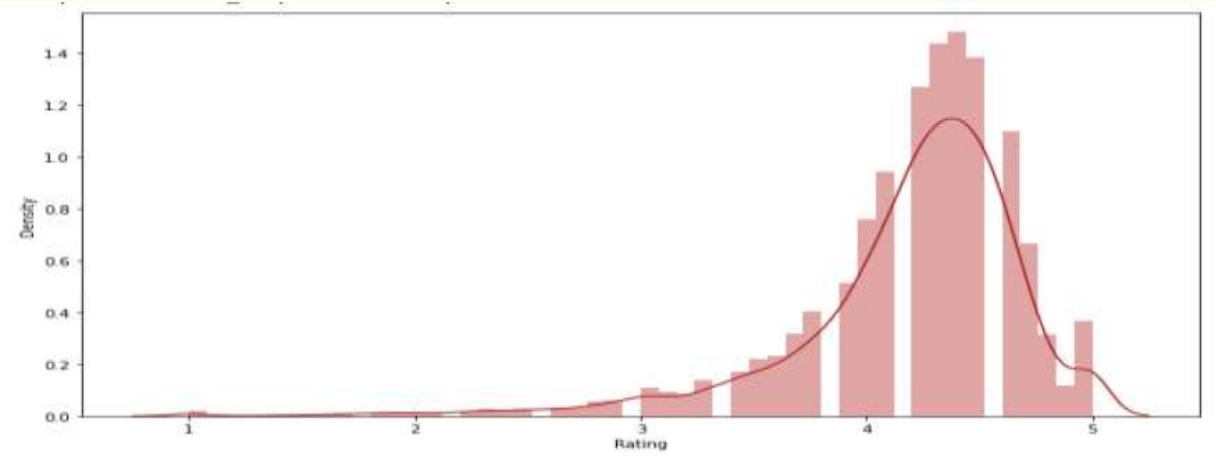
Distribution of App size

- The below curve represents the variation of the size of apps available on Google Play store.
- It is clear from the visualizations that the data in the Size column is skewed towards the right.
- Also, we see that a vast majority of the entries in this column are of the value Varies with device, replacing this with any central tendency value (mean or median) may give incorrect visualizations and results. Hence these values are left as it is.



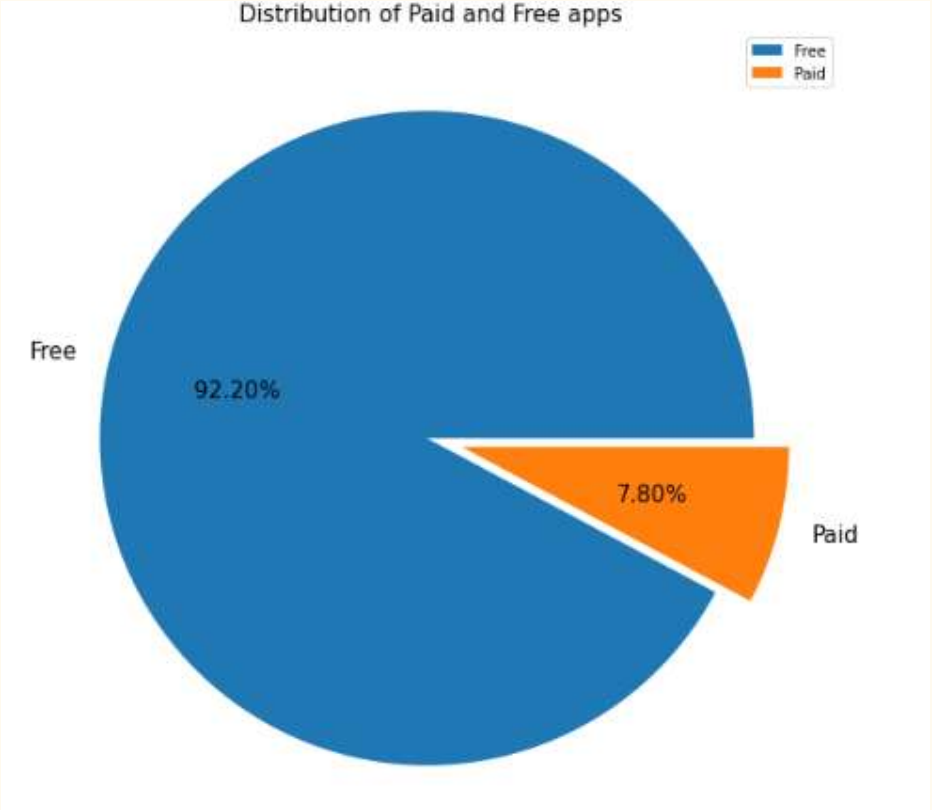
Distribution on Ratings

- The mean of the average ratings (excluding the NaN values) comes to be 4.2.
- The median of the entries (excluding the NaN values) in the 'Rating' column comes to be 4.3. From this we can say that 50% of the apps have an average rating of above 4.3, and the rest below 4.3.
- From the distplot visualizations, it is clear that the ratings are left skewed.
- We know that if the variable is skewed, the mean is biased by the values at the far end of the distribution. Therefore, the median is a better representation of the majority of the values in the variable.

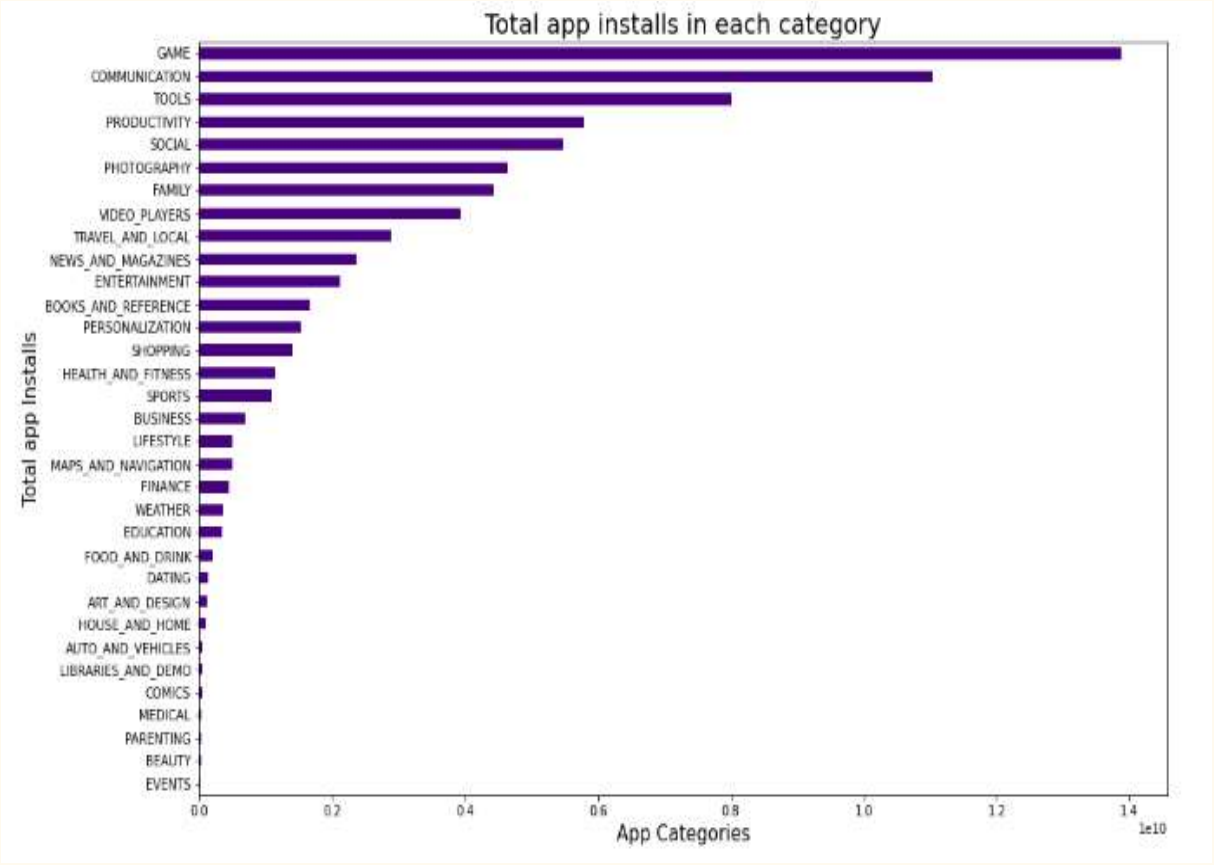


Percentage of Paid apps v/s Free apps

We can Observed from this Pie-chart is that **92.20% of Apps are free** and only **7.80% of Apps are paid** in Play store.

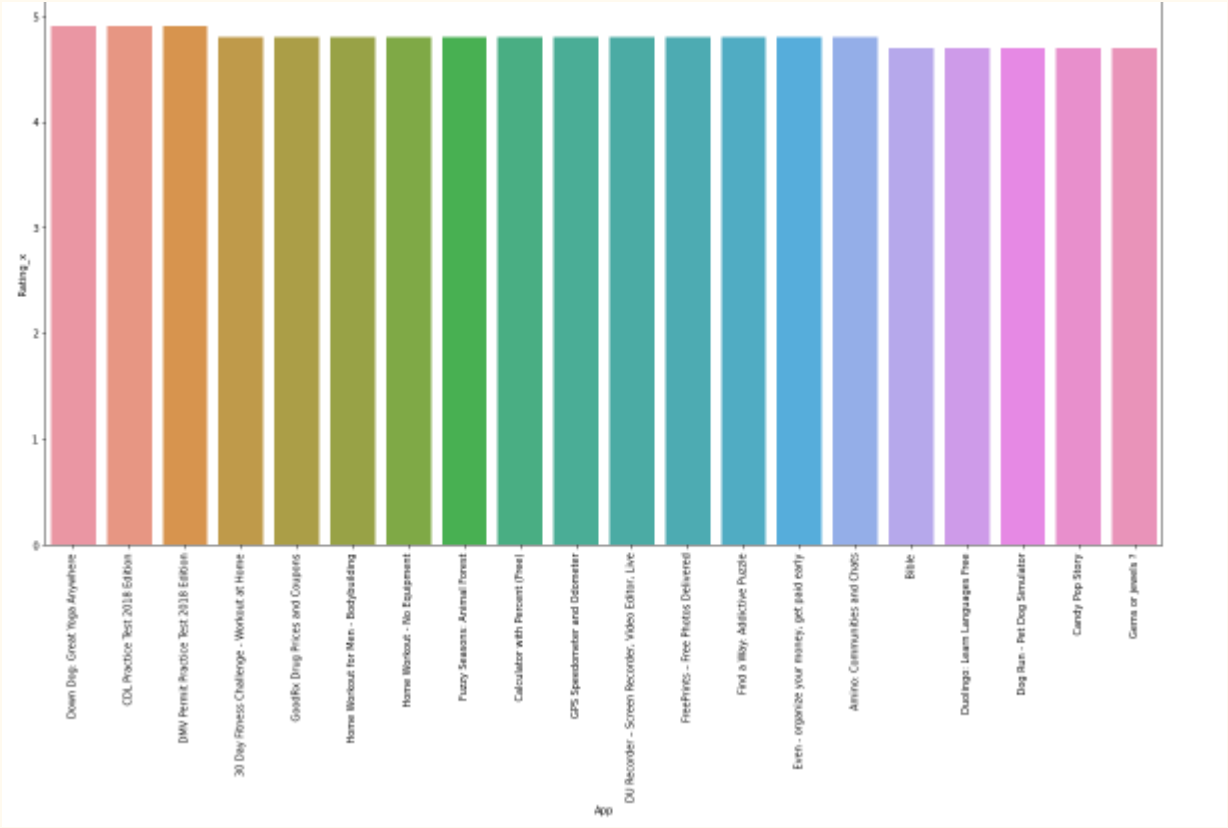


Category App's have most and least number of installs



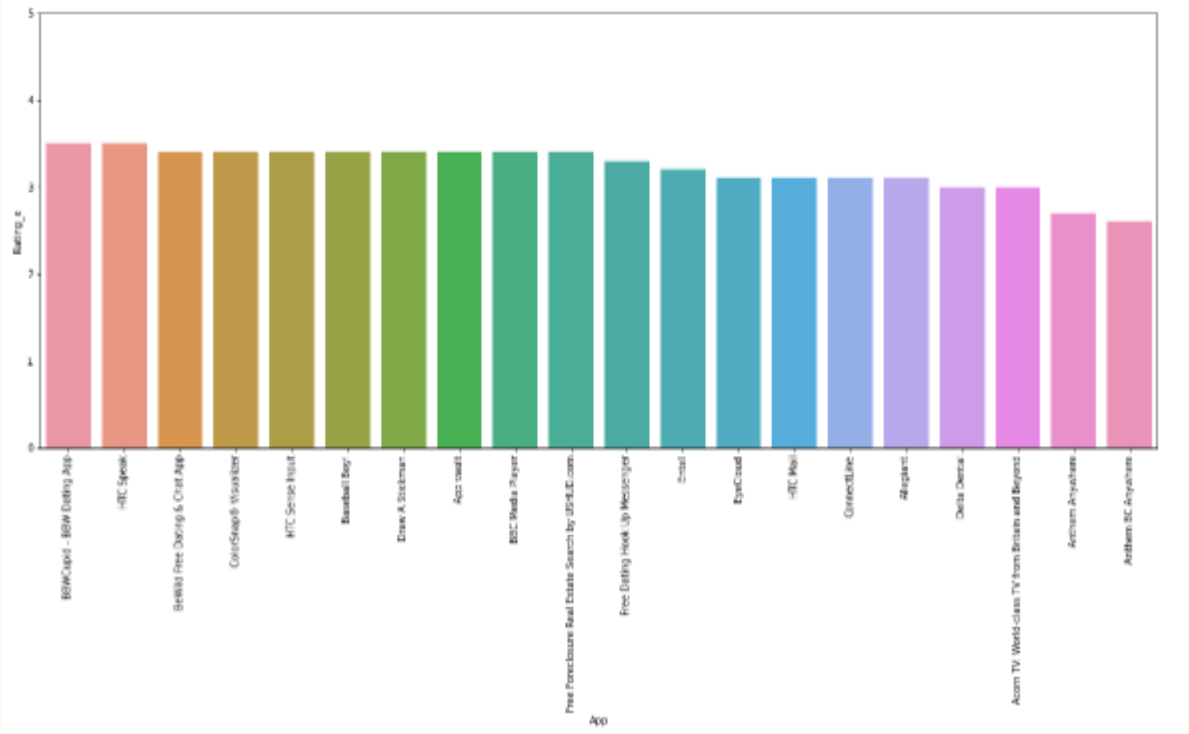
According to the Graph, we can say that **Game** is the category which has the most number of Downloads and installations. And the least installed category is **Events.**

Top 20 most popular apps based on Rating



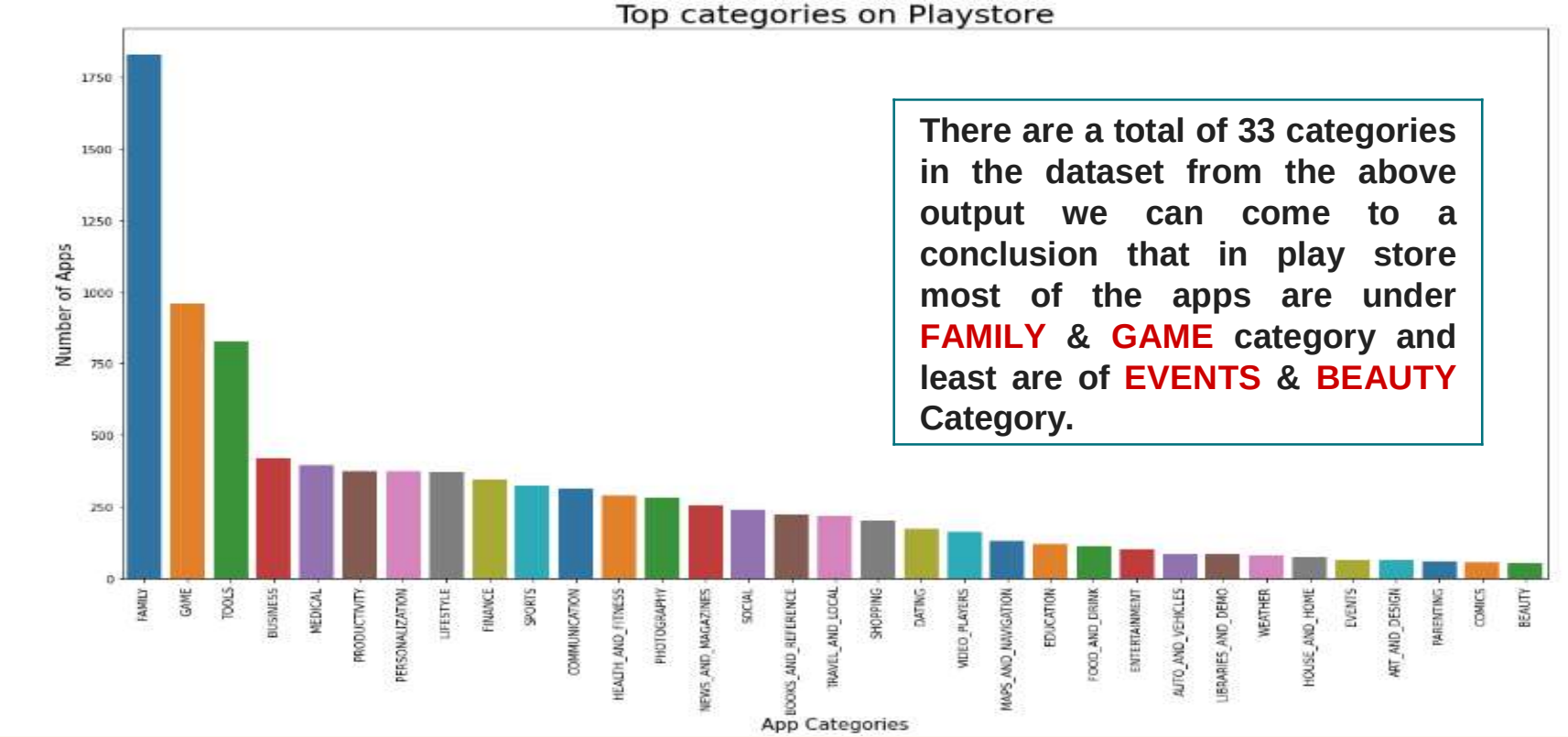
- We can see that all of the top 20 apps are rated close to 5 stars. We can say the **popularity** of an app can be determined by higher ratings.
- Top 3 Apps in terms of ratings are- **Down Dog: Great Yoga Anywhere, CDL practice test 2018 Edition, DMV Permit practice.**

Top 20 least popular apps based on Ratings



- **Lowest 3 apps in terms of rating are -- Anthem Anywhere, Anthem BC Anywhere.**

Top Categories on Play store



There are a total of 33 categories in the dataset from the above output we can come to a conclusion that in play store most of the apps are under **FAMILY** & **GAME** category and least are of **EVENTS** & **BEAUTY** Category.

Challenges Faced

- Reading the dataset and comprehending the problem statement.
- Examining the business KPIs for app development and devising a solution to the problem.
- Handling the error, duplicate and NaN values in the dataset.
- Designing multiple visualizations to summarize the information in the dataset and successfully communicate the results and trends to the reader.

Conclusion's

- **Family, Game and Tools are top three** categories having most number of Downloads.
- Most competitive category: **Family.**
- **92.19%** apps are **Free** and 7.81% apps are paid in type.
- **81.80%** apps have **Everyone** content rating.
- Category with the highest number of installs: **Game**
- Most of the apps (around 50%) have an average rating of **4.3;** The rest of the apps are below **4.3.**
- Most of the apps are under the size of **50 MB.**
- The median size of the apps in the play store is **12 MB**


THANK
YOU