Capstone Project

Exploratory Data Analysis

Project Title:

Play Store App Review Analysis

By: Ajay Tiwari

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.
  Each app (row) has values for category, rating, size, and more.
  Another dataset contains customer reviews of the android apps.

- We have to explore and analyze the data to discover key factors responsible for app engagement and success.

Objective of our Project:

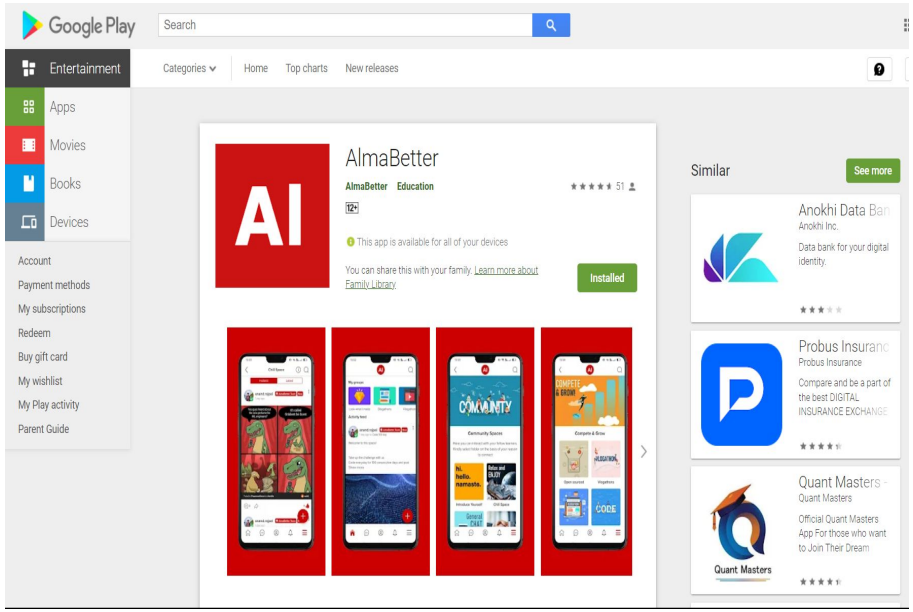
- Through data wrangling the process of cleaning and unifying messy and complex data sets for easy access and analysis.
- So the main objective of our project is to detect the key factors responsible for customer base analysis with apps through some exploratory data analysis and find out the selected features to draw conclusions.

Index

Introduction

- The **Google Play Store** is the official pre-installed app on android devices which provides access to the Google Play store.

- It allows users to browse and download music, books, magazines, movies, television programs, and applications from Google Play. The Devices segment of Google Play is not accessible through the Play Store.



Data pipeline

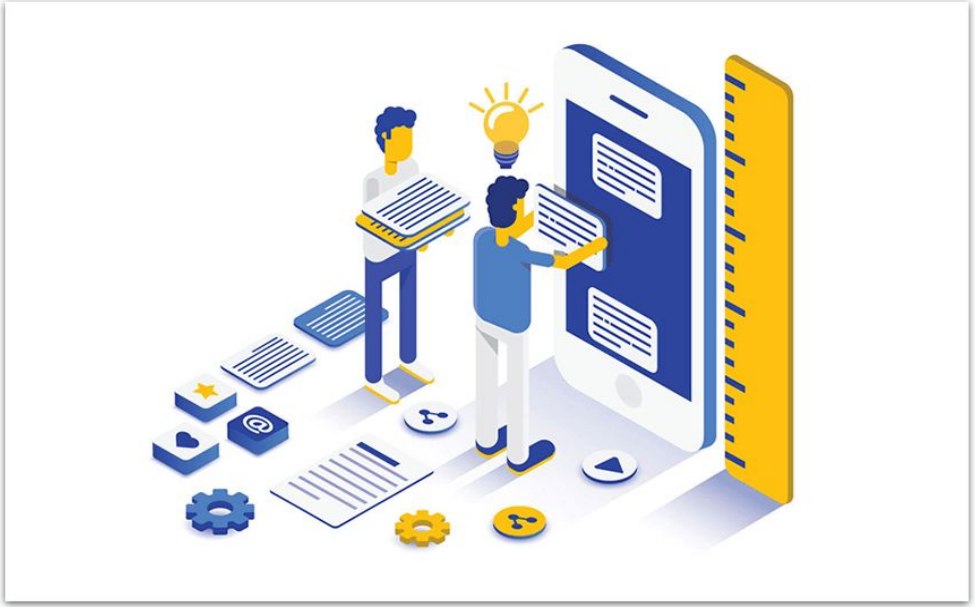
- **Data Wrangling-** The process of cleaning and unifying messy and complex data sets for easy to access and analysis, which includes treatment of null values and removing duplicates.

- **Data Processing-** In this process, we manually examined each feature and processed the null values, by replacing them with the mean or median values, wherever required. for easy interpretation.

- **EDA -** In this part, we did some exploratory data analysis(EDA) on the selected features to draw conclusions.

Data Summary

- Collected descriptive information on over 9637 apps across 30 different categories in the Google Play Store.
- Categories included Business, Family, Game, Tool, Education, Social, Finance, Medical etc.
- Total number of rows 9637 and variable columns are 13



Variables Description In Play Store & Reviews

- 'App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type', 'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver', 'Android Ver'

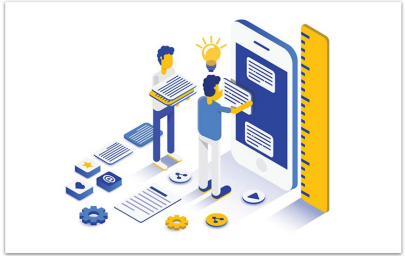
- 'App',
- 'Translated Review',
- 'Sentiment',
- 'Sentiment Polarity',
- 'Sentiment Subjectivity'

Variable Breakdown In play store:

- **Apps** - Application name
- **Categories** - Category of given application.
- **Rating** - Rating is a reflection of how users respond to the apps. Rating of app is between 0 and 5.
- **Reviews** - Total number of reviews.
- **Size** - The amount of space required to install the app. This feature shows the size of the application.
- **Installs** - This feature represents the number of installations on devices.
- **Type** - This feature tells us whether an app is free
- **Price** - This column shows the price of certain apps.
- **Content Rating** - Apps use a separate content rating system. Titles rated ALL, have content that may be suitable for all ages.

- **Genres** - App genres help us determine the type of apps that are being built. It is a detailed description
- **Last Updated** - This feature represents the date when the application was last updated. Updating the apps gives us access to the latest features and it also improves security and stability of the applications.
- **Current Version** - A positive integer used as an internal version number. This number is used only to determine whether one version is more recent than the another, with higher numbers indicating more recent versions.
- **Android Version** - Android version indicates the version of Android platform. The bigger the version number, the newer the Android is.

Variable Breakdown In User Reviews:

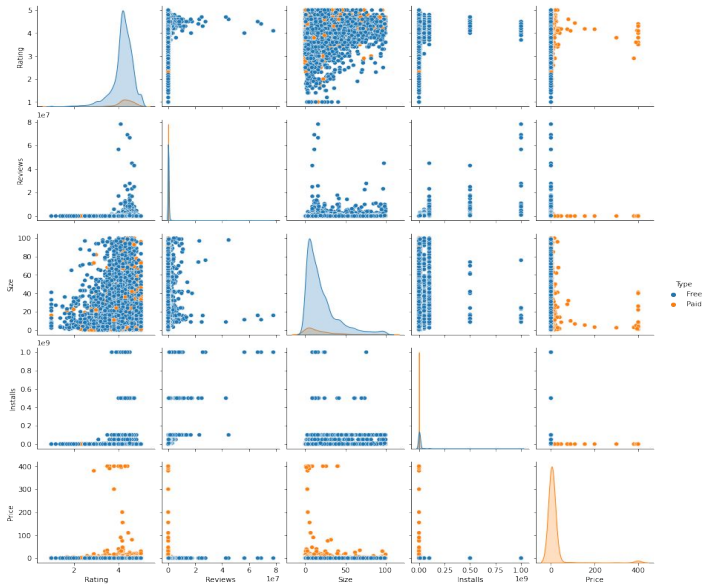


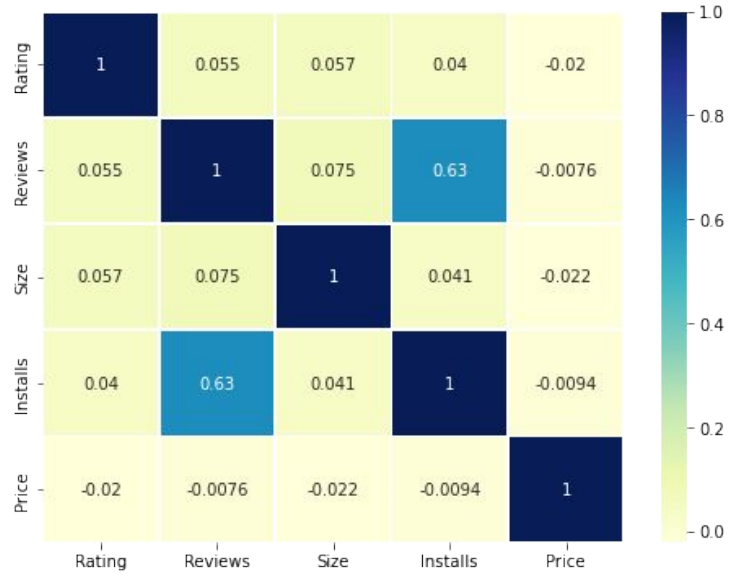
- **Apps –** Application name.
- **Translated Review** – Review for giving app.
- **Sentiment** - Boolean values of user translated reviews; 'positive', 'negative and 'neutral‘.
- **Sentiment Polarity** - Sentiment polarity for an app defines the orientation of the expressed sentiment. It is calibrated value of translated reviews that varies from -1 to 1.
- **Sentiment Subjectivity** - Sentiment Subjectivity is basically pitch of translated reviews that varies from 0

Exploratory Data Analysis

Relationship Analysis

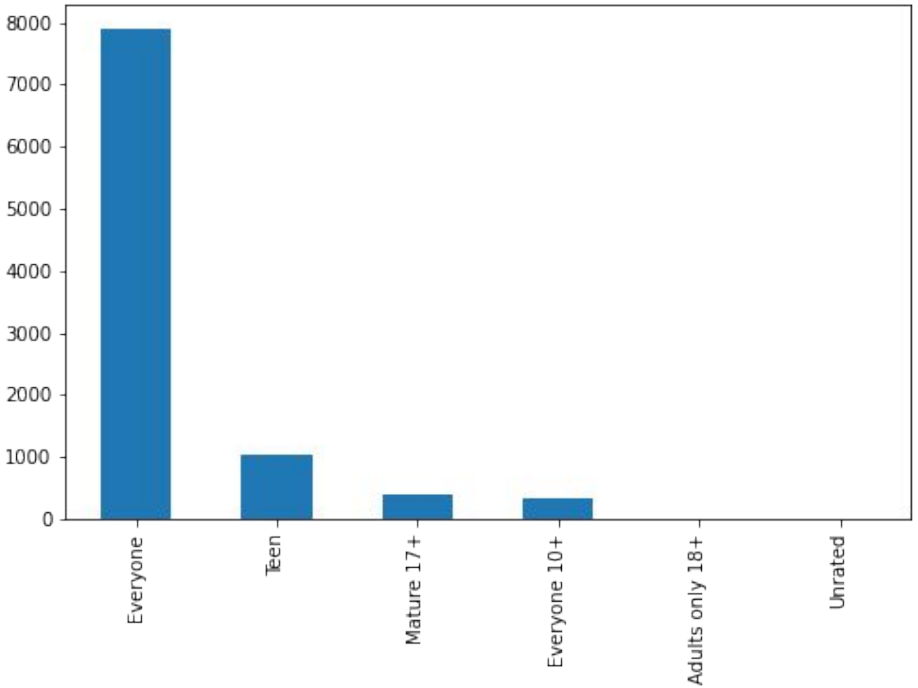
- We can see that 'Rating' and 'Reviews' are highly positive correlated.
- 'Installs' and 'Size' are also correlated positive.
- Most of the columns are negative correlated with 'Price' column





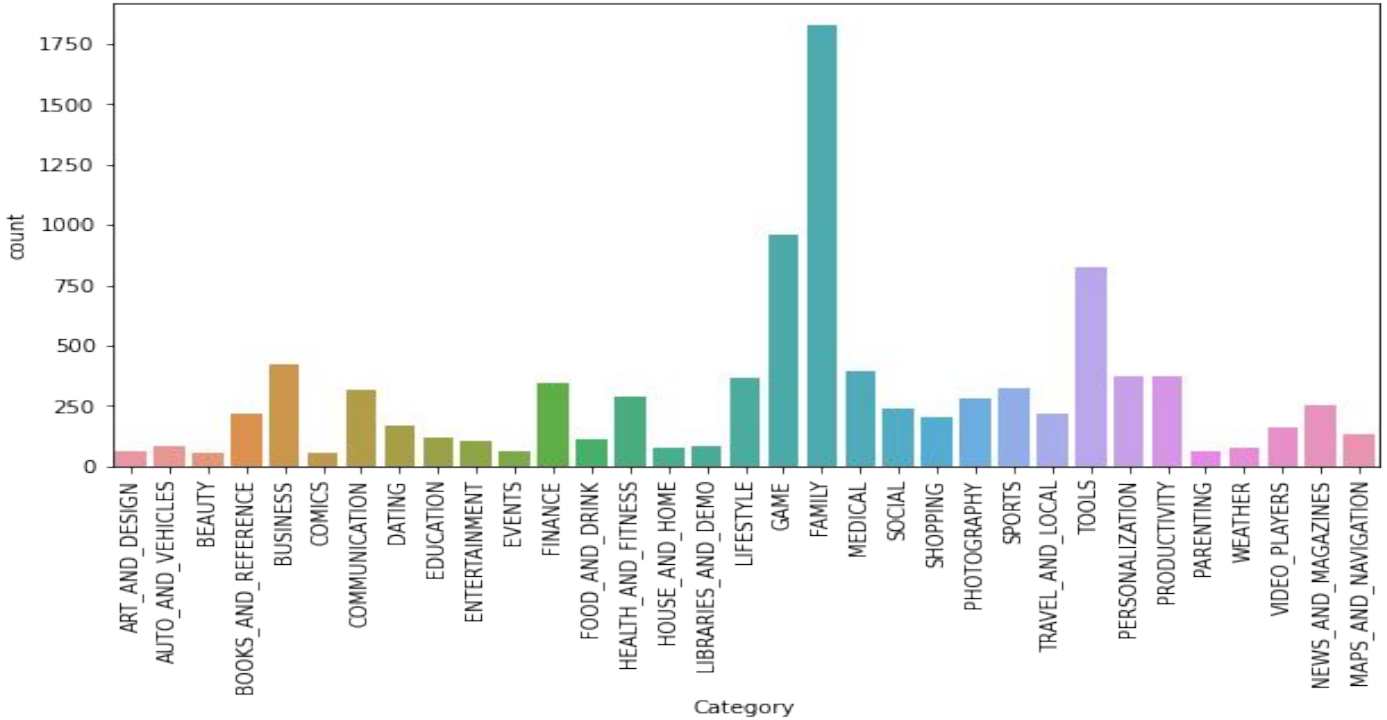
Content Rating

- we can see from the graph that most of the apps are for “everyone” (Approx 8000) so that it could capture the most of the customer-base.
  Only fews are rated for "Adults only 18+" and "Unrated".



Category wise app popularity

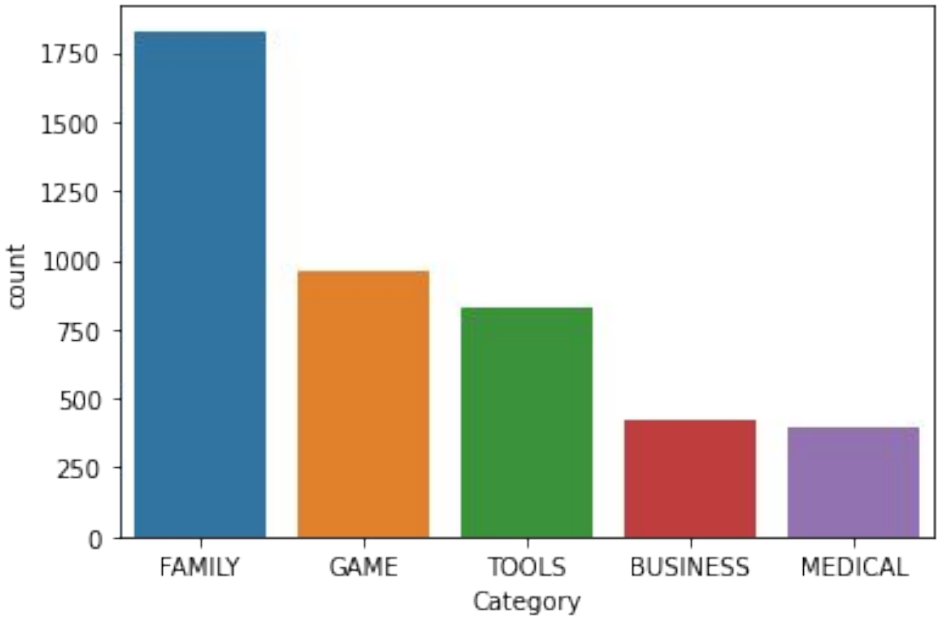
- Family, Games & Tools were the most common categories of the total number of apps in our dataset, respectively.



Top 5 category

Top 5 Categories are:

- 1. Family
- 2. Game
- 3. Tools
- 4. Business
- 5. Medical

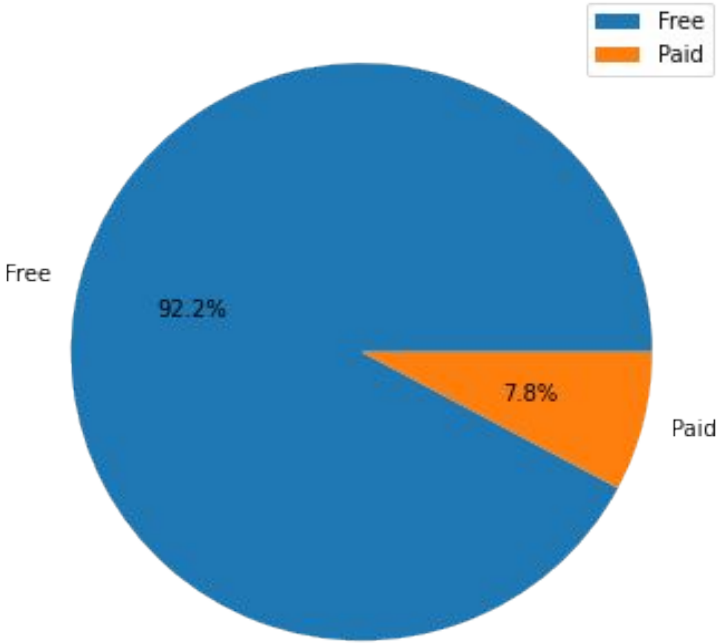


Free Vs. Paid Apps

- Here we can see that there are

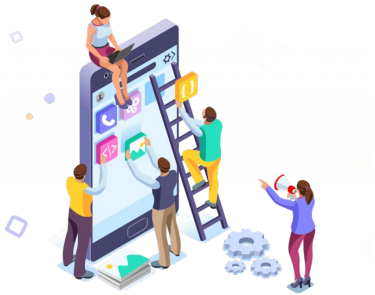
  92.2% of apps are free and only

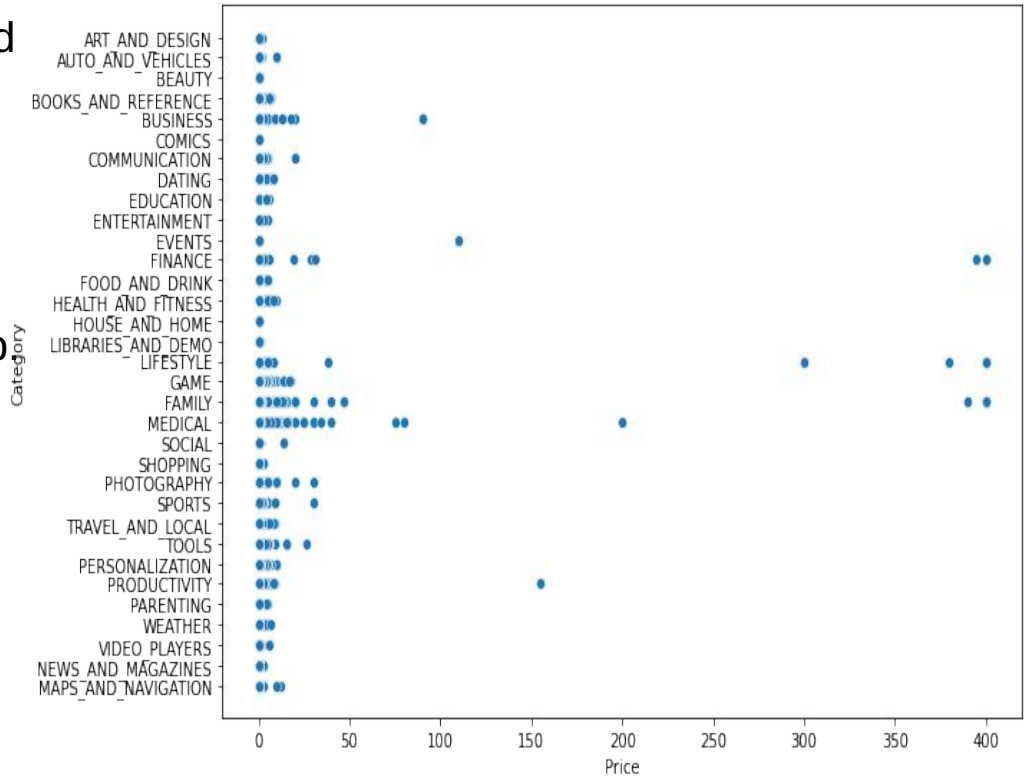
  7.8% of Apps are paid on Playstore.



Category Wise Pricing Visualization

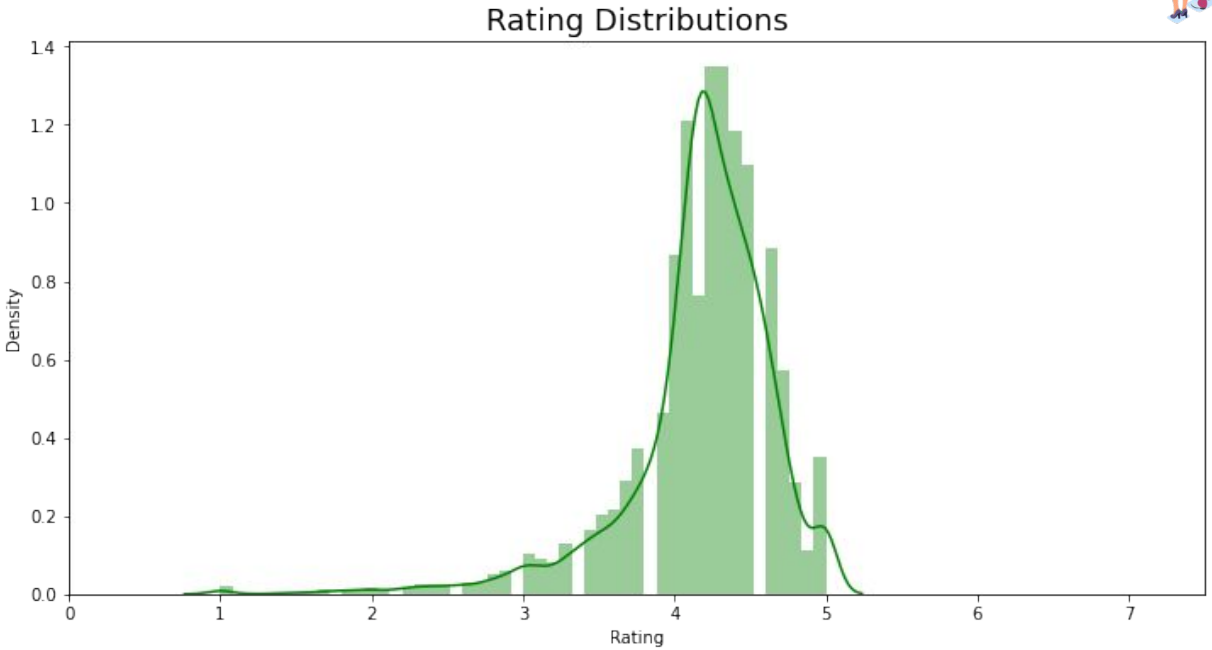
- Family and Medical categories had the biggest number of paid apps available for download.
- The 'I'm Rich - Trump Edition' app is found to be the highest paid app





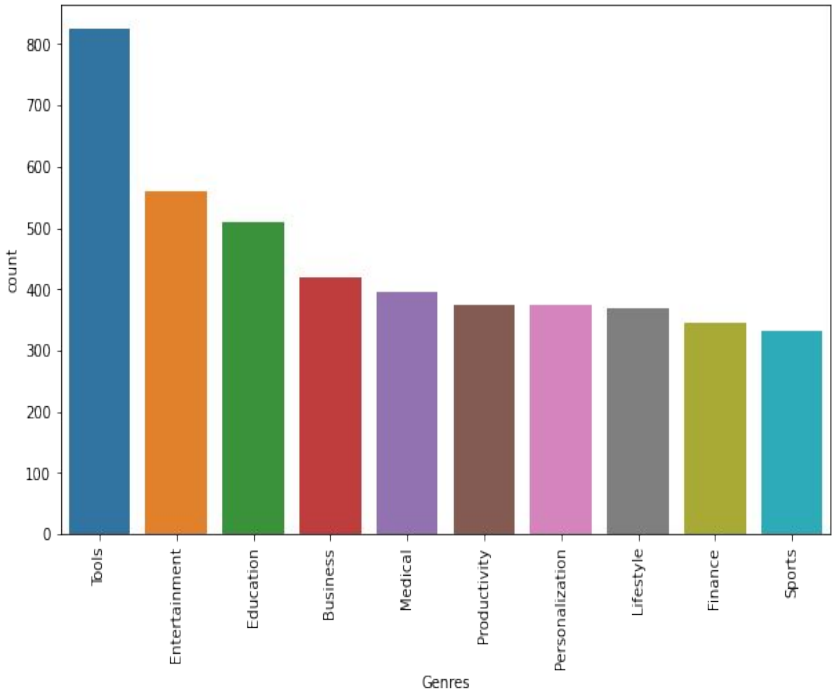
Distribution of User Ratings

- Maximum of the apps are rated around 4 to 4.5



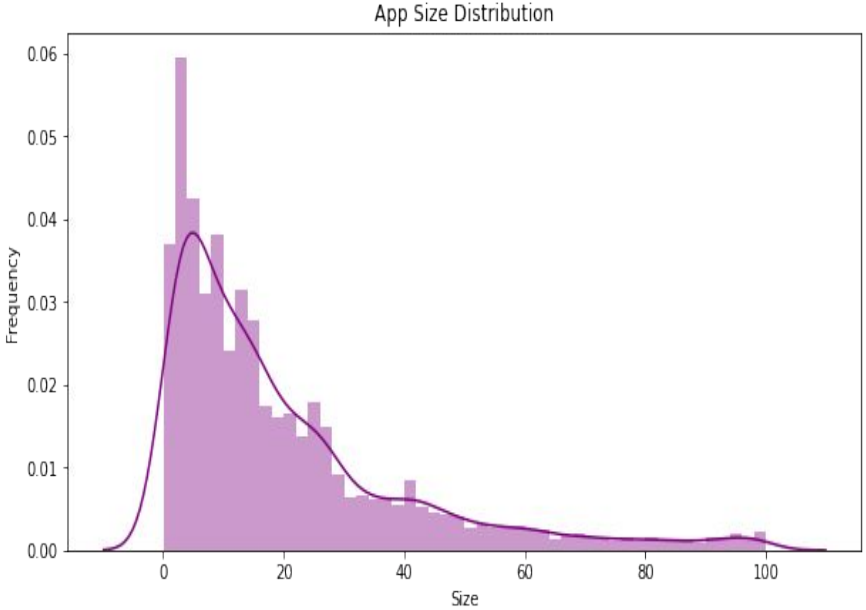
Genre wise rating analysis

- Most frequent genres are Tools,Entertainment,Education.
- There are 118 unique Genres.
- Comics;Creativity is the highest rated genre followed by Health & Fitness;Education and Books & Reference.



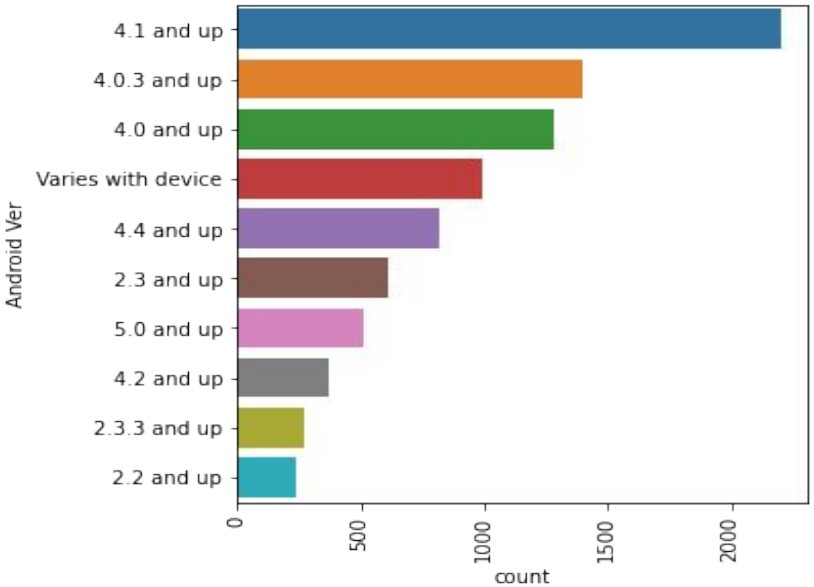
App Size Distribution

- Maximum apps are in the size category of 0 KB to 20000 KB i.e. 0MB to 20MB.Almost 65% of apps from total dataset size lies between 0-20 M.



Android version

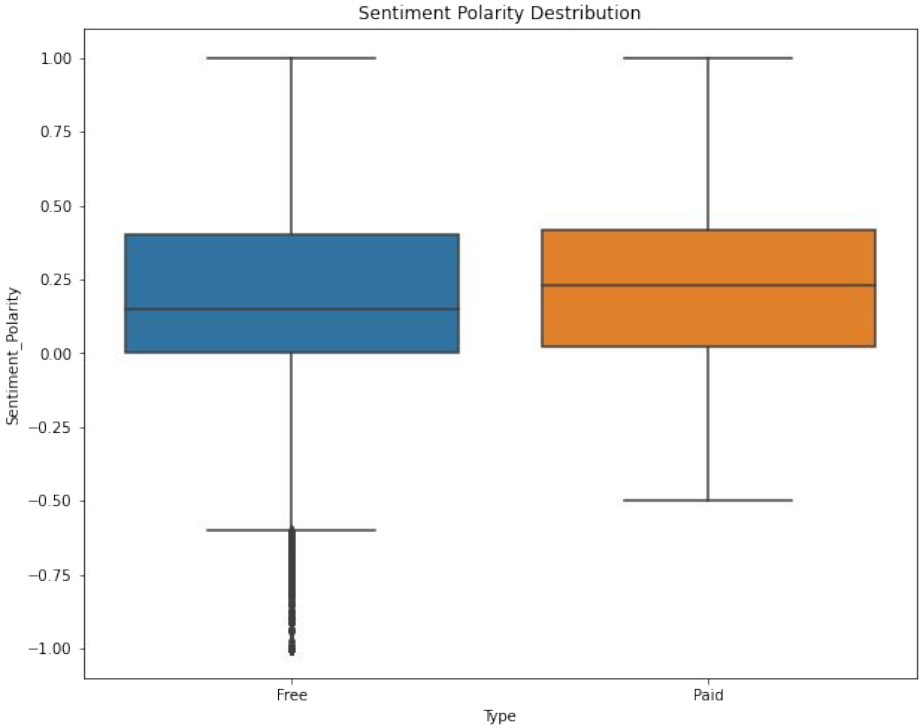
- Most of the apps are made compatible with Android version 4(Jellybean) and above.



Distribution Sentiment Types

Sentiment Polarity Between Free and Paid Apps

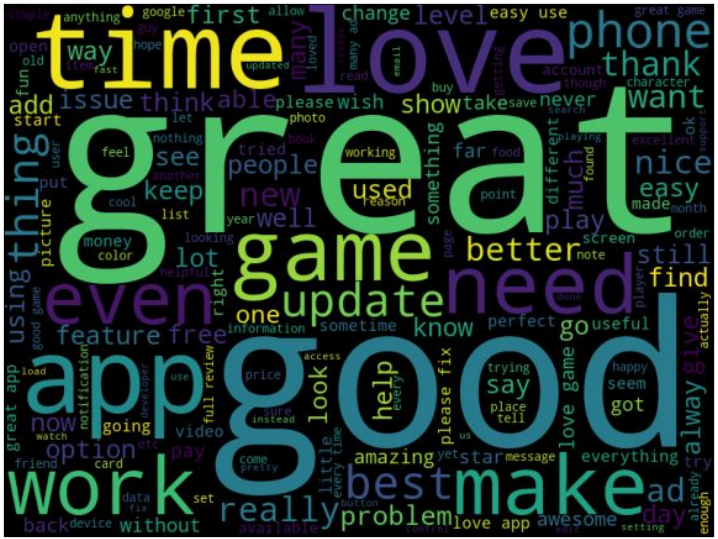
- Free apps relatively receive more negative comments (outliers on the negative Y-axis) than paid apps.



Sentiment Of Users

- We generate wordcloud image for negative and positive reviews.
- We can see that most positive sentiment reviews are from most popular categories. Positive reviews are higher than negative and neutral sentiment reviews.

Positive



Negative

Conclusion

The objective of this Play store Data analysis is to know how different parameters can affect the app ratings and reviews and understand how these can impact the Play store app industry as a whole. So we can conclude that:

- Users mostly prefer the free apps. App size does not affect the decision of using the paid or free apps much.
- The apps which have the higher rating above 4 are targeting all the people and not a certain age group.
- On average the application size is in between 0 MB to 20 MB.
- Family & Medical category Apps have the highest earning.
- Also at the same time Medical, Family, Tools, Game category apps are the most expensive apps which is clearly visible because these categories of the application market have the most invested money and indeed these are the right categories to make best profits.
- Customer ratings actually affect the category to release an app in that category.
- We can see that most positive sentiment reviews are from most popular categories.
- Positive reviews are higher than negative and neutral sentiment reviews.

THANK YOU