**ABSTRACT**

|  |  |
| --- | --- |
| App | Name of the App |
| Category | Category under which it falls |
| Rating | Application’s rating on play store |
| Reviews | Number of reviews of the app |
| Size | Size of the app |
| Installs | Number of Installation of the app |
| Type | Whether the app is free or paid |
| Price | Price of the app if it’s a paid app (0 if it’s a free app) |
| Content Rating | Appropriate target audience of the app |
| Genres | Genres under which the app falls |
| Last Updated | Date when the App was last updated |
| Current Version | Current version of the App |
| Android Version | Minimum android version required to support the App. |

Google play store is simply entertainment at our fingertips. It’s an official app store and a digital media store having enormous things to offer. Applications are either free or paid. Our team has worked on play store data which have two datasets, first contains basic information of apps and other is user reviews data. This information

can be used for predicting key factors responsible for app engagement & success story.

## INTRODUCTION

Play store is not just an app store, it’s a platform offering various digital content to its consumers.

The Google Play Store is home to android applications, music, movies, books, games and television programs. 81% of the apps are free of cost which has led to immense popularity of this

platform. As per google survey report 3000+ apps are being added every other day. The Google Play Store contains applications for the Android system only. This document reveals the dynamics of the Play Store app and gives actionable insights for the developers to work on and rule the Android market

## INTEGRAL METHODOLOGY:

First, we investigated some basic information of our dataset. On doing so we found out that our data needed some cleaning, some values were missing, and some datatypes were incorrect. We started with data cleaning and correcting the data types, followed by data visualization. We removed some unnecessary features and made it ready for analysis using different plots.

## DATASET DESCRIPTION:

1st dataset has 13 features and 10841 observations. Which are as:

* 2nd Dataset contains 5 features and 64295 observations. Which are as:

|  |  |
| --- | --- |
| App | The name of the application. |
| Translated Review | Review texts in English |
| Sentiment | View or opinion of users which can be Positive, Negative or Neutral. |
| Sentiment Polarity | Sentiment in numerical form ranging from -1 to 1. |
| Sentiment Subjectivity | Measure of the expressions of opinions, evaluations, feelings and speculations. |

## BREAKDOWN OF DATASETS:

Before proceeding to data visualization, we need to perform the following steps:

1. Importing required packages for future analysis.
2. Mounting drive and reading data files from Google drive.
3. Viewing all data information.
4. Dropping duplicates.
5. Finding and removing the outliers.
6. Removing special characters
7. Checking unique values, null count and datatypes of each column.
8. Segregation of numerical and categorical data.

## EXAMINING NULL / MISSING VALUES

Some values in our dataset are null or missing. These values affect the accuracy and performance of the models that predict the outcome, so these need to be handled. While analyzing our dataset the first thing we will do is to examine the null or missing values in our dataset. This makes our result accurate. Missing values are more in Size & Rating columns as can be seen by plotting graphs. Hence several methods are used to remove these values.

## DATA CLEANING:

Data cleaning is the foremost step in any data science project. Cleaner the data, better are the results. As the proverb goes by saying “More Data beats clever algorithm, but better data beats more Data” – Peter Norvig. To begin with our data cleaning, first we remove the duplicate values. Then we remove unnecessary characters in our dataset.

After doing so we find the unique values of each column and make the necessary changes in each column like converting datatypes, removing the null and ‘nan’ values.

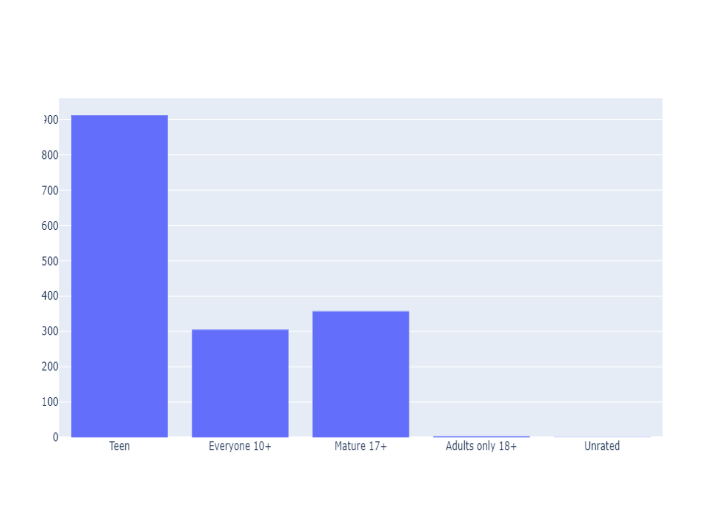
Lastly, we have done exploratory data analysis of our dataset.

**DATA VISUALIZATION:**

# Observation 1:

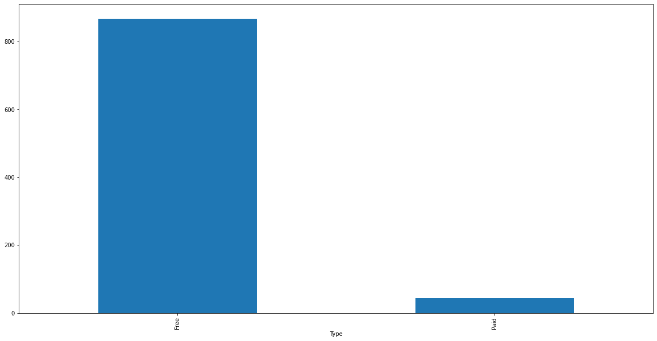
We plotted a graph all the content rating category

And we observed that the teen category is getting the most numbers compare to all other category in the graph.



# Observation 2:

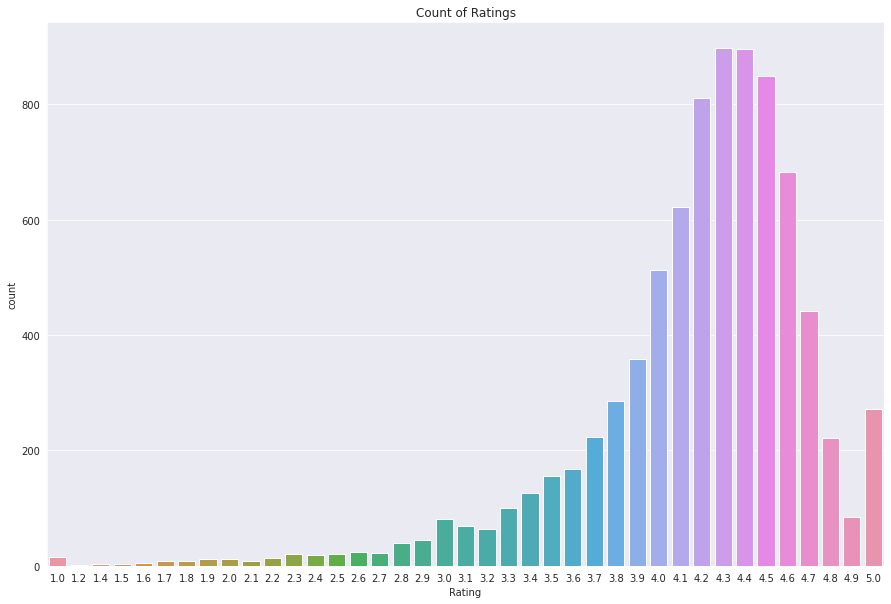
We plotted a graph of teen category to compare with the paid and free apps.



We got to know that number of free apps are very large as compare to paid in the teen of content rating.

# Observation 3:

We plotted a graph of number of rating on play store.



Most of the apps ratings are between 4.2 and 4.6. Maximum numbers of apps rated is 4.3 followed by 4.4.

**Observation 4:**

We plotted a graph of app rating across major categories.

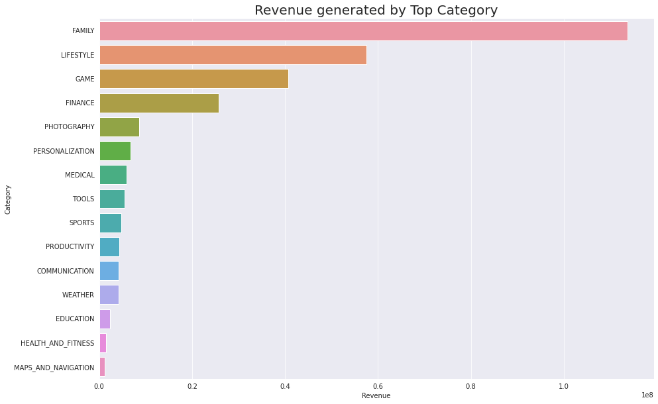
* As seen The Health and Fitness and Books and Reference is having the most rating greater than 4.5
* Half of the Dating apps have a rating lower than the average.

# Observation 5:

We have plotted a chart to find the revenue.

Formula to find the revenue:

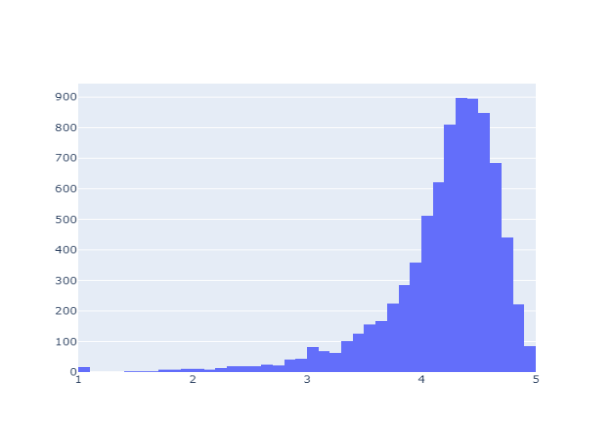
{Revenue = Install\*Price}



* Family is the highest revenue generated category
* Maps are the lowest

**Observation 6:**

We plot this graph to check the average rating given to apps on play store. As observed average rating of apps on Google Play Store is {4.17}



**Observation 7:**

We plotted a graph to the percentage of positive and negative comments of play store app



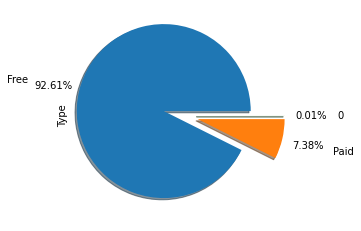
We observed that more than 50% of the sentiments are positive.

**Observation 8:**

We plotted a graph to compare age with sentiments.

**Observation 10:**

By plotting this graph we can check the relation between the paid apps and free apps on play store.

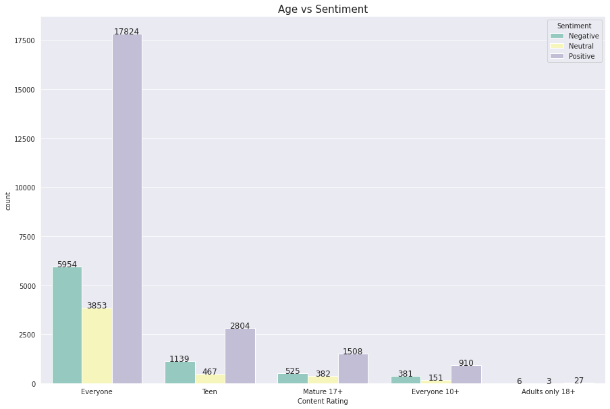


We observed that more than 90% of that apps on play store are free apps and only 7.38% of the apps are paid.

**Observation 11:**

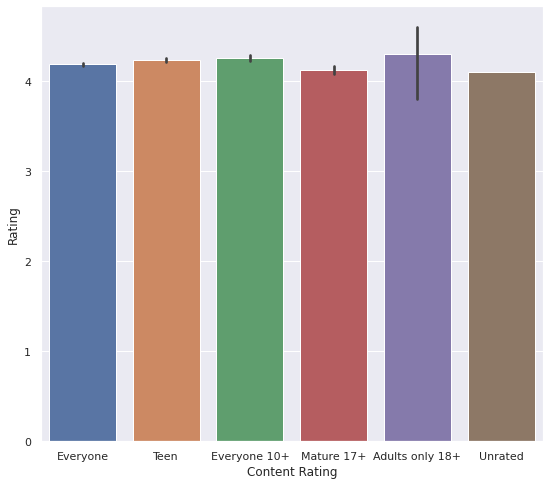
We plotted a correlation heatmap between both the dataset given ( play store and user review analysis) . We got to know that there are negative values coming whi

As seen the most positive sentiments (20028)



**Observation 9:**

We plotted this graph to check the rating given to different category of content rating.



We observed from the above map that only adult 18+ is been rated the most among all other categories.

# 

# 

# ADVANTAGES OF VISUALISATION:

Visualized data is processed faster and easier.

Better insights of the data are drawn which may be missed in traditional reports.

Helps us visualize trends which improve performance.

**CONCLUSION AND FUTURE WORK:**

The app developers can predict the outcome of the developed apps. Better insights are drawn from this visualization. Apps which need to be improved can be worked upon by the developers. The dataset contains immense possibilities to improve business values and have a positive impact.

We could add a system that would create application on its own by using the dataset and creating the best user interface by highly rated apps.

**ACKNOWLEDGEMENT**

This project is presented by-Amar Singh Chouhan  
 Ashwin Babu  
 Hitesh Malvi  
 Ninad Rathore  
 Vikas Chandra

Our sincere efforts have made us to accomplish the task of completing this project. We are extremely grateful to our instructors and mentors who have helped us to grow in this field. We would like to express our sincere gratitude to the celebrated authors whose phenomenal work has been consulted and referred in our project work. We also wish to convey our appreciation to our peers who provided encouragement and timely support in the hour of need. This project helps us to improve our skills and enhance our knowledge.

## REFERENCES

[https://seaborn.pydata.org/tutorial/color\_palett](https://seaborn.pydata.org/tutorial/color_palettes.html) [es.html](https://seaborn.pydata.org/tutorial/color_palettes.html)

[https://www.python-graph-gallery.com/196-](https://www.python-graph-gallery.com/196-select-one-color-with-matplotlib) [select-one-color-with-matplotlib](https://www.python-graph-gallery.com/196-select-one-color-with-matplotlib)https://www.analyticsvidhya.com/blog/2021/05/10-colab-tips-and-hacks-for-efficient-use-of-it/