**Play Store App Review Analysis**

**Gaurav Makode**

**Data science trainee,**

**AlmaBetter, Bangalore**

**Email**: [gauravmmakode002@gmail.com](mailto:gauravmmakode002@gmail.com)

**Github**: [[GitHub - gauravmakode/Play-Store-App-Review-Analysis-](https://github.com/gauravmakode/Play-Store-App-Review-Analysis-%0c)](https://github.com/gauravmakode/Play-Store-App-Review-Analysis-)

**Abstract:**

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. In this way, an application's prosperity is normally dictated by the quantity of installation of the application and the client appraisals that it has gotten over its lifetime instead of the income is created. Application (App) ratings are feedback provided voluntarily by users and function important evaluation criteria for apps. However, these ratings can often be biased due to insufficient or missing votes. Additionally, significant differences are observed between numeric ratings and user reviews. This Study aims to predict the ratings of Google Play Store apps

using machine learning Algorithms. I have tried to perform Data Analysis and prediction into the Google Play store application dataset that I have collected from Kaggle. Using Machine Learning Algorithms, I have tried to discover the relationships among various attributes present in my dataset such as which application is free or paid, about the user reviews, rating of the application.

**1.Problem Statement**

Data is taken from the Google play store dataset. Every row contains various entries regarding a certain app. We will be doing Exploratory data analysis on this data set, which is a very important step in data science cycle, as it not only helps in taking very initial business decisions but also in preparing the data for further modelling for use in machine learning algorithms. Our objective will be to structure the data, clean it and present certain trends that we observe that can help us draw very preliminary conclusions about the probability of success of a newly launched app.

# INTRODUCTION

Machine learning approaches are essential for us to take care of numerous issues. In this paper, we present machine learning models and structures in detail. Machine learning has numerous applications in numerous perspectives and has incredible advancement potential.

In future, it is predictable that machine learning could set up ideal speculations to clarify its exhibitions. In the meantime, its capacities of unsupervised learning will be improved since there is much information on the planet however it isn't relevant to add names to every one of them. It is additionally anticipated that neural system structures will turn out to be increasingly unpredictable with the goal that they can separate all the more semantically important highlights. In addition, profound learning will consolidate with support adapting better and we can utilize these points of interest to achieve more assignments.

# Google Play store and User Review Analysis

In today’s scenario we can see that mobile apps playing an important role in any individual’s life. It has been seen that the development of the mobile application advertise has an incredible effect on advanced innovation. Having said that, with the consistently developing versatile application showcase there is additionally an eminent ascent of portable application designers inevitably bringing about high as can be income by the worldwide portable application industry.

With enormous challenge from everywhere throughout the globe, it is basic for a designer to realize that he is continuing in the right heading. To hold this income and their place in the market the application designers may need to figure out how to stick into their present position. The Google Play Store is observed to be the biggest application platform. It has been seen that although it creates more than two-fold the downloads than the Apple App Store yet makes just a large portion of the cash contrasted with the App Store. In this way, I scratched information from the Play Store to direct our examination on it.

With the fast development of advanced cells, portable applications (Mobile Apps) have turned out to be basic pieces of our lives. Be that as it may, it is troublesome for us to follow along the fact and to understand everything about the apps as new applications are entering market each day. It is accounted for that Android market achieved a large portion of a million applications in September 2011. Starting at now, 0.675 million Android applications are accessible on Google Play App Store. Such a lot of applications are by all accounts an extraordinary open door for clients to purchase from a wide determination extend. We trust versatile application clients consider online application surveys as a noteworthy impact for paid applications. It is trying for a potential client to peruse all the literary remarks and rating to settle on a choice. Additionally, application engineers experience issues in discovering how to improve the application execution dependent on generally speaking evaluations alone and would profit by understanding a huge number of printed remarks.

We develop Android apps & release on Play Store. As an Developer or say Business Perspective it’s very important to know whether users are enjoying the app or facing any issues. To know this Play Store has a Ratings & reviews section for each app released on play store. Users can submit the ratings and has a freedom to write a review for a particular app. This approach is quite a lengthy to rate & review app i.e. navigate to Play store to submit feedback or redirect leaving a current app workflow to open Play Store App link using URI. We never wanted our customers to leave our application, but with this flow, we are forced to redirect the control to Play store app.

# Google Play store Dataset

Mobile apps are everywhere. They are easy to create and can be money making. Because of these two factors, more  
and more apps are being developed. In this notebook, we will do a comprehensive analysis of the Android app  
market by comparing over ten thousand apps in Google Play across different categories. We'll look for insights in the  
data to devise strategies to drive growth and retention.

**Following two files in the dataset which consists different of features:**

**1.1 datasets/play\_store\_data.csv:**This file contains all the details of the apps on Google Play. 13 features describe a given app.  
• **App:** Name of the app  
• **Category:** Category of the app. Some examples are ART\_AND\_DESIGN, FINANCE, COMICS, BEAUTY etc.  
• **Rating:** The current average rating (out of 5) of the app on Google Play  
• **Reviews:** Number of user reviews given on the app  
• **Size:** Size of the app in MB (megabytes)  
• **Installs:** Number of times the app was downloaded from Google Play  
• **Type:** Whether the app is paid or free  
• **Price:** Price of the app in US$  
• **Content Rating:** A content rating (also known as maturity rating) rates the suitability of TV broadcasts,  
movies, comic books, or video games to its audience. To show which age group is suitable to view media and  
entertainment.  
• **Genres:** A category of artistic, musical, or literary composition characterized by a particular style, form, or  
content  
• **Last Updated**: Date on which the app was last updated on Google Play  
• **Current Ver**: Current Version means a version of the software that is currently being supported by its  
publisher.  
• **Android Ver:** Android versions (codenames) are used to describe the various updates for the open-source  
Android mobile operating system.

**1.2 datasets/user\_reviews.csv:**This file contains a random sample of 100 most helpful first user reviews for each app. The distribution of positive  
and negative reviews in each category has been pre-processed and passed through a sentiment analyzer.

• **App:** Name of the app on which the user review was provided. Matches the App column of the play\_store\_data.csv file  
• **Translated Review:** The pre-processed user review text.  
• **Sentiment:** Sentiment category of the user review - Positive, Negative or Neutral.  
• **Sentiment Polarity**: Sentiment score of the user review. It lies between [-1,1]. A higher score denotes a  
more positive sentiment.

# Data Cleaning and Preparation

Preprocessing is important into transitioning raw data into a more desirable format. Undergoing the preprocessing process can help with completeness and compellability. For instance, you'll see if certain values were recorded or not. Also, you'll see how trustable the info is. It could also help with finding how consistent the values are. We need preprocessing because most real-world data are dirty. Data can be noisy i.e. the data can contain outliers or simply errors generally. Data can also be incomplete i.e. there can be some missing values.

The available data is raw and unusable for Exploratory data analysis, so before we do anything with the data we will have to explore and clean it to prepare it for data analysis.

* **Step1**: We write a function play store info (), that will display 5 attributes about all the columns: Data type, Count of non-null values, Count of null values, number of unique values in that column and percentage of null value in that columns in the play store dataset.
* **Step2**: we start off with the column ‘Type’ we can see that it has one null value. We checked this row and found out from the play store that it is a free app. We use filllna() function of the pandas library to fill this value.
* **Step 3**: We drop the columns ‘Current Ver’, ‘Android Ver’ and ‘last updated’ from our dataset using the drop() function of the pandas library.
* **Step 4**: We can see that the ‘Rating’ column has 1474 null values. Due to low variations in the rating values and a lot of repeated values the ‘median’ would be a suitable statistical indicator to replace the null values with. We calculate the mode of the column using the median () aggregate method, and fill this value in place of null values using the fillna() function.
* **Step 5:** We can see that the ‘Reviews’ column despite being a numerical indicator is of the ‘object’ data type, we will convert this to ‘int’ data type using the as type(int) function.
* **Step 6:**  We can see that the size column, which should be numeric, is of the data type ‘object’, it also has characters ‘k’ and ‘M’ in the values which stand for kilobytes and Megabytes, we will replace the ‘k’ with 1000 and ‘M’ with 1000000. Some values also have ‘+’ sign in them, which will be removed. Next, we will convert this column into ‘int’ datatype.
* **Step 7:** The ‘Installs’ column values contain the characters ‘+’ and ‘,’ which are going to prevent us from converting this column into a numeric datatype. We will get rid of these using the strip() and replace() functions.
* **Step 8:** The values in the column ‘Price’ might have the ‘$’ sign in some values and the column is of the datatype ‘object’. We will first remove the ‘$’ sign using the **strip()** function and then convert the column into ‘int’ datatype.
* **Step 9:** Handling the duplicates in the App column we drop the no of duplicate rows that are present in the App columns.
* **Step 10:** We write a function Ur info(), that will display 5 attributes about all the columns: Data type, Count of non-null values, Count of null values ,number of unique values in that column and percentage of null value in that columns in the User review dataset.
* **Step11:** In the User review dataset the columns are App, Translated Review, Sentiment, Sentiment Polarity, Sentiment Subjectivity in this total 26863 NaN value are present so we drop them using dropna() function.
* 3. **Exploratory Data Analysis.**

Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.

EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. In this article, we will understand EDA with the help of an example dataset. We will use **Python** language (**Pandas** library) for this purpose.

## Free vs Paid

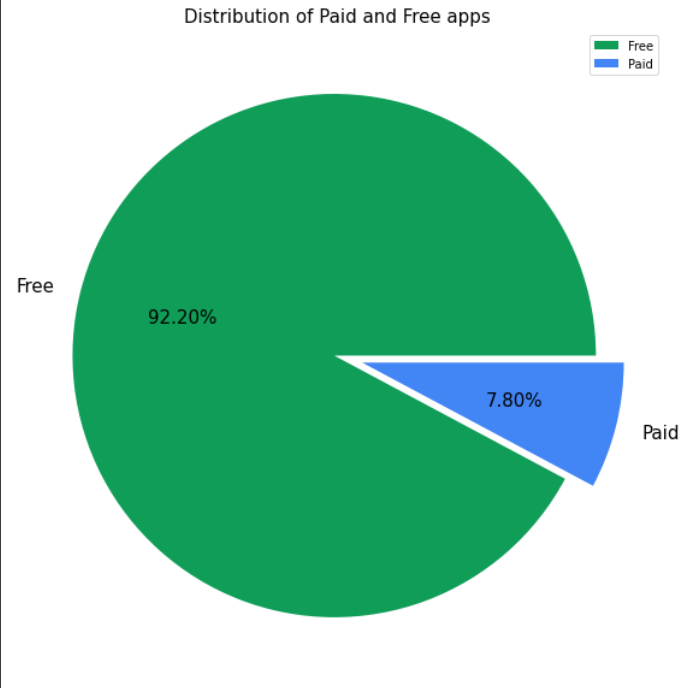
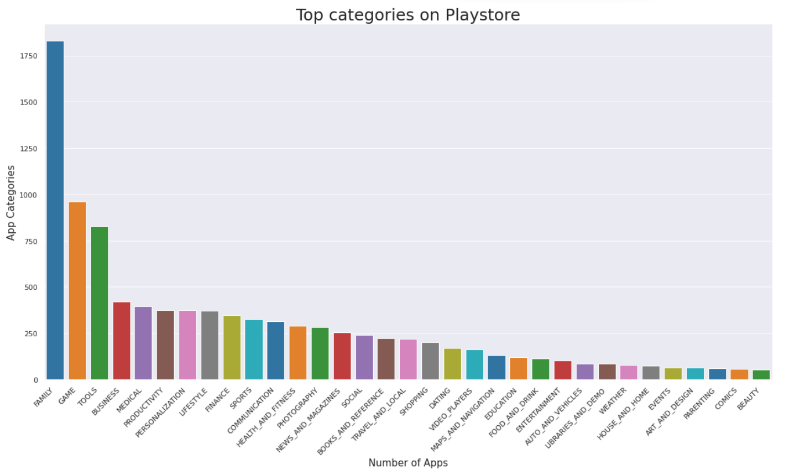


Fig -1:Free vs Paid

Here we can see that 92.20% apps are free, and 7.80% apps are paid on Google Play Store, so we can say that Most of the apps are free on Google Play Store.

## 3.2Top Category of Play store

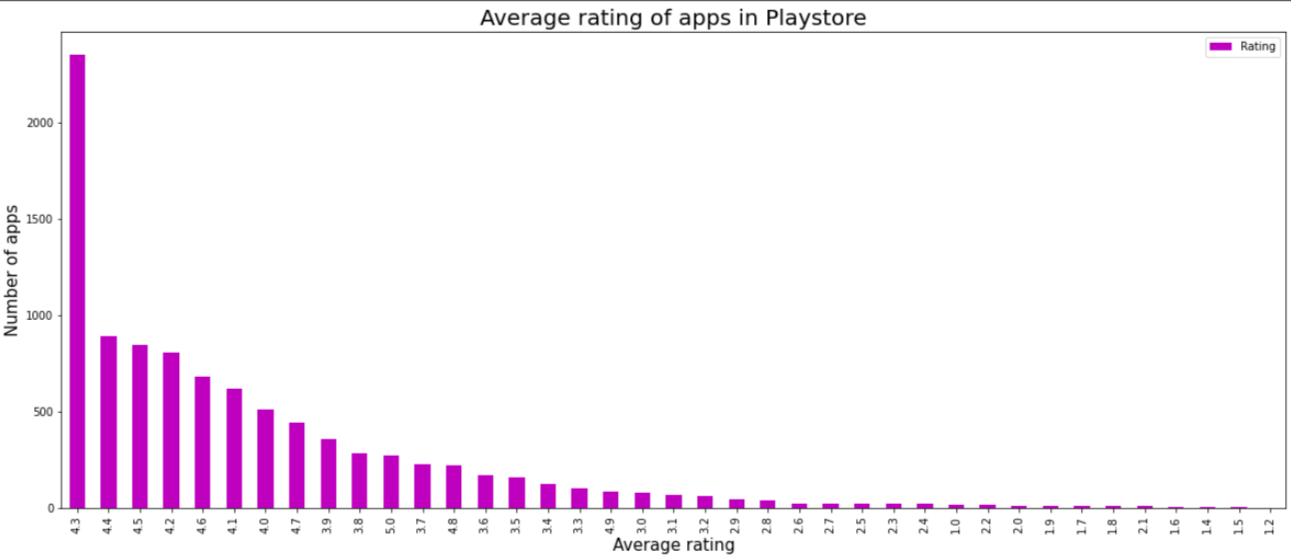
There are lot of category wise apps are available on playstore so the below curve show hoe the apps are distributed.



**Fig -2: Top Categories on Playstore**

So, there are all total 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.

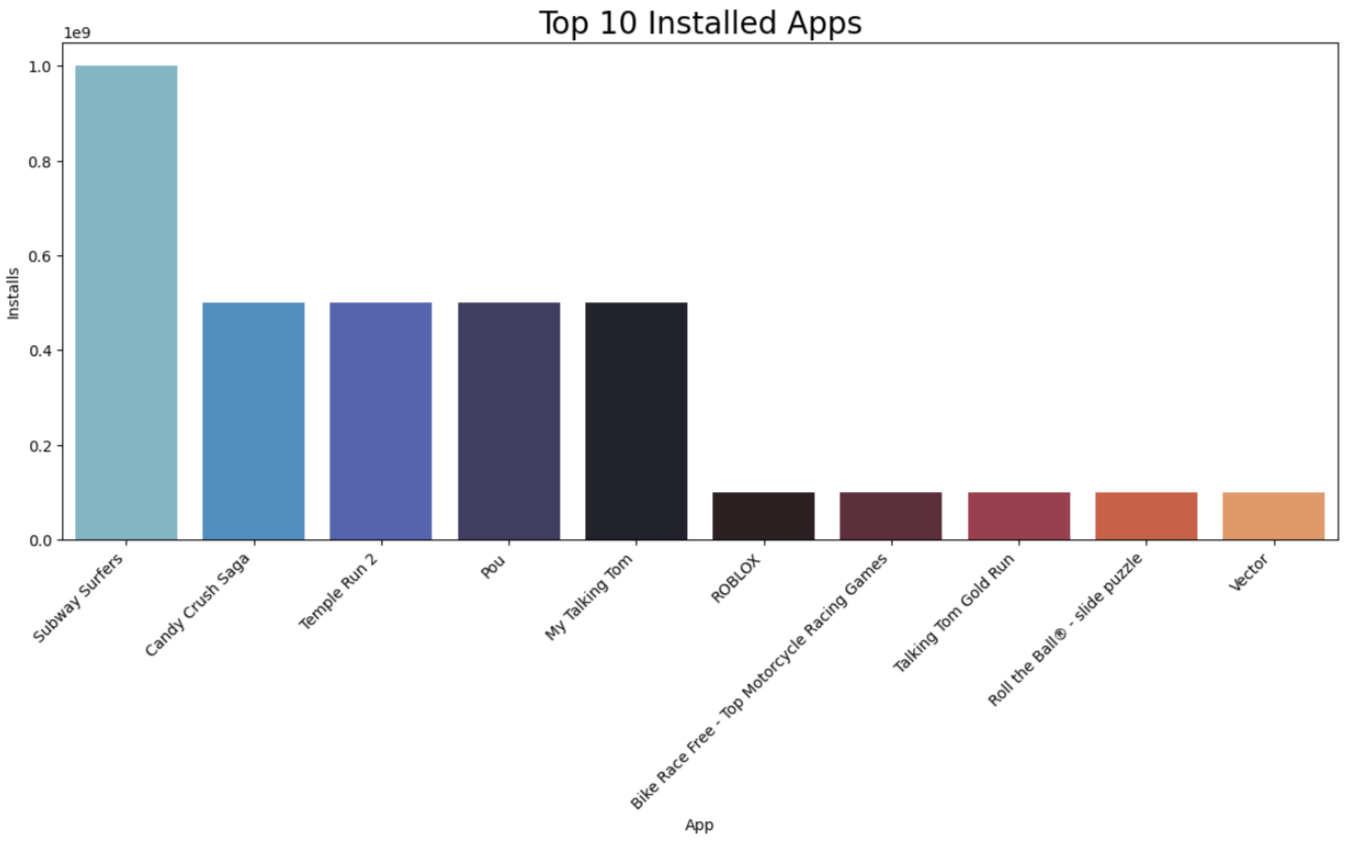
**3.3 Average rating of the apps**



**Fig -3: Average rating of the apps**

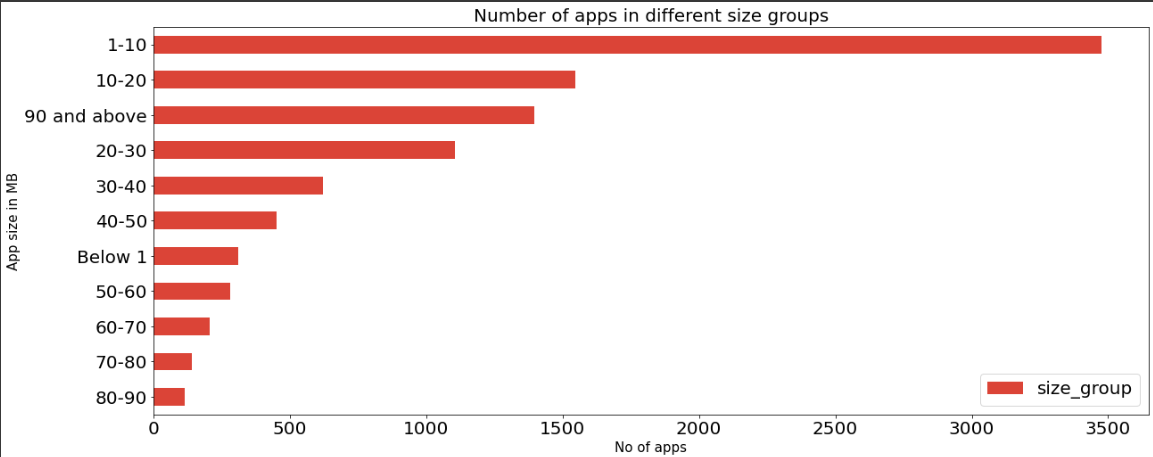
There are more No. of app rated 4.3 on play store .

**3.4 Top installed apps**

****

Subway Surfer has the highest no. of install, And next top 4 app has almost equal no. of install **.**

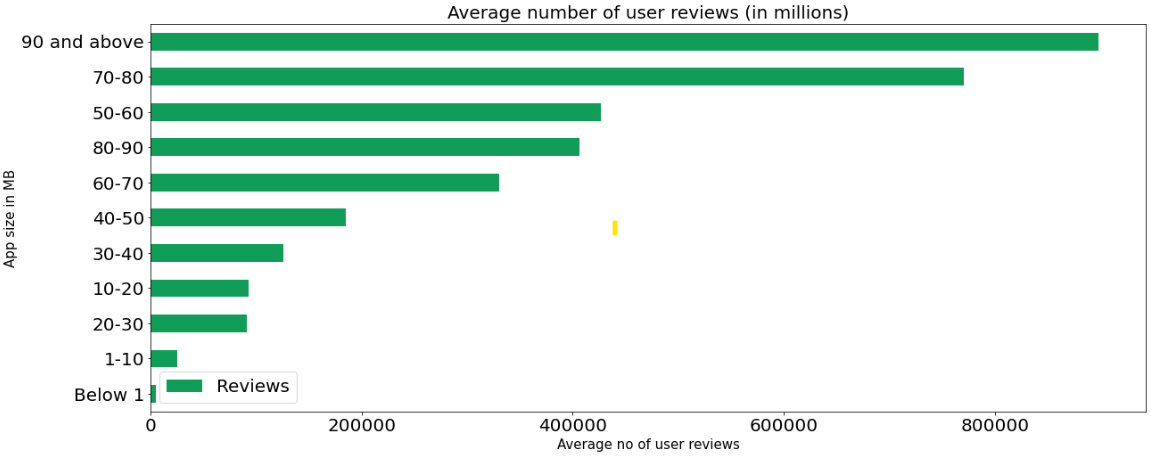
**3.5** **Distribution of apps based on size.**

****

**Fig -5: Distribution of apps based on size.**

There are more no. of app installed which size is between 1 to 10 mb.

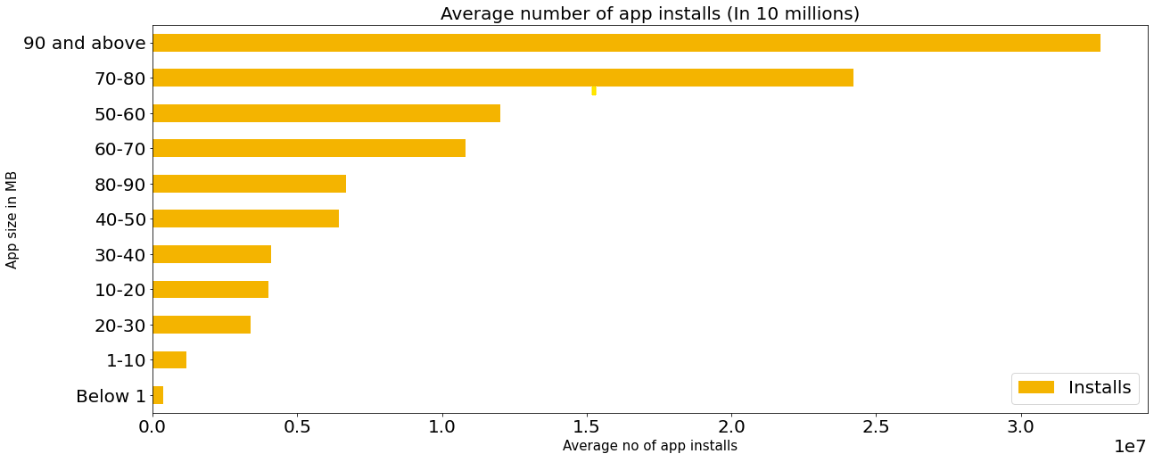
3.4 **Average no. of user reviews.**



**Fig -6: Average no. of user reviews.**

Apps of size 90 mb and above has more no. of user reviews **.**

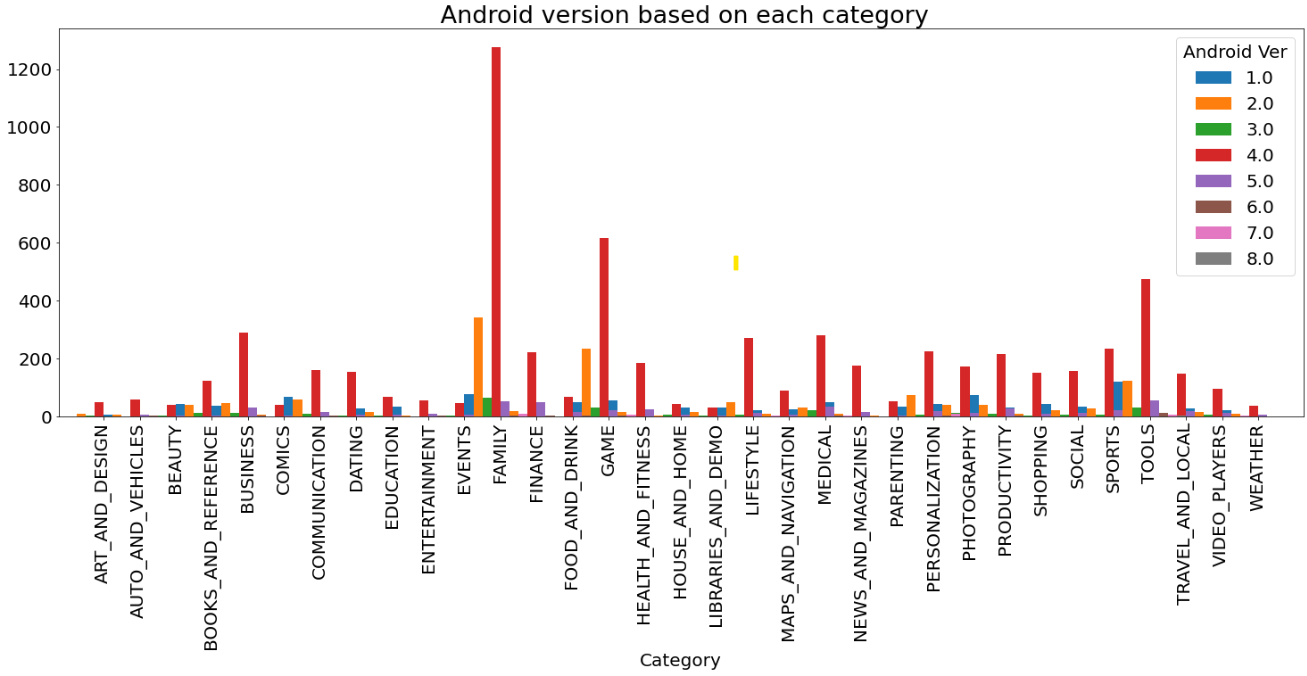
**3.5 Average no. of App installs.**

****

**Fig -7: Average no. of App installs.**

There are more no. of apps install whose size is 90mb or greater than 90mb .

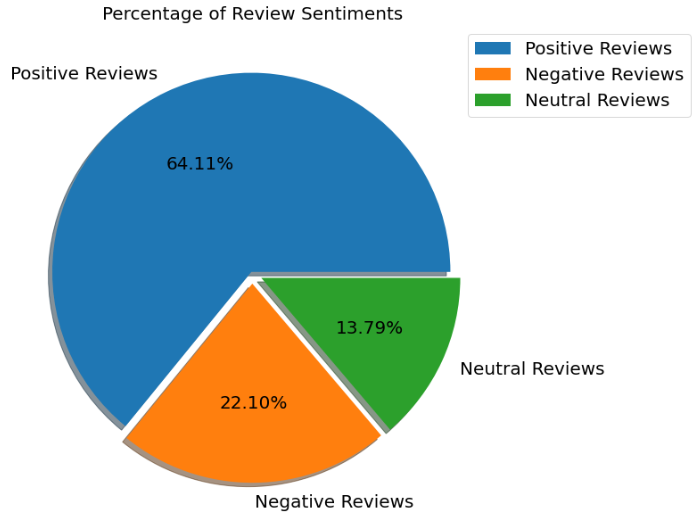
**3.6 Android version based on each category**



**Fig -8: Android version based on each category.**

In every category largest no. of apps are made on Android version 4.0

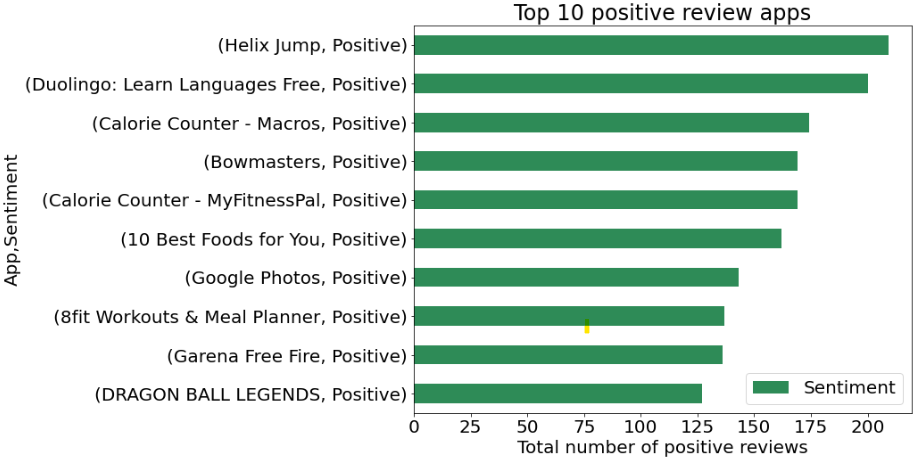
3.7 **Review sentiments.**



**Fig -9 Review sentiments.**

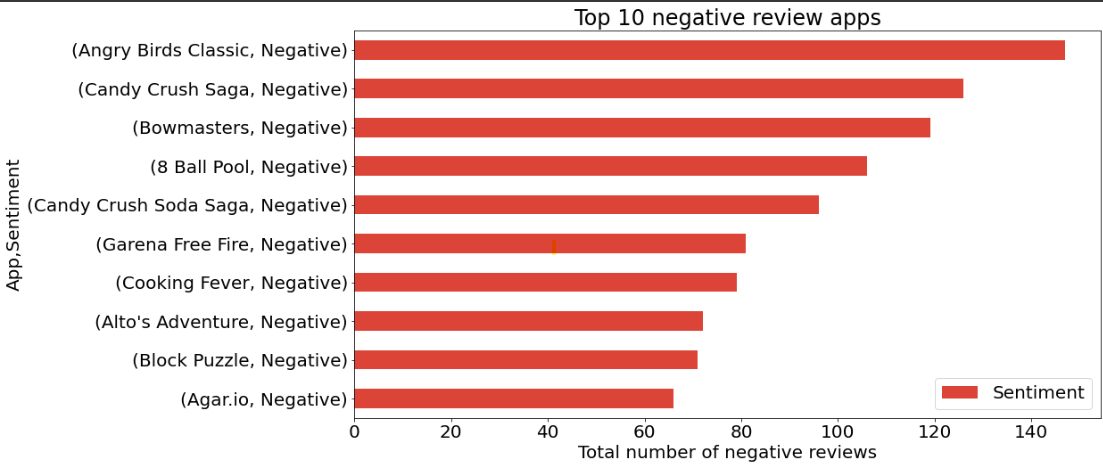
There are more no. of percentage of positive reviews which is 64.11% . There are 13.79% of neutral reviews . There are 22.10% of negative reviews **.**

**3.8 Top 10 positive review apps.**



**Fig -10: Top 10 positive review apps.**

3.9 **Top 10 nagative review apps**.



**Fig -11: Top 10 nagative review apps.**

## Relationship between sentiment subjectivity proportional to sentiment polarity

Chart, scatter chart

Description automatically generated

**Fig -12: Google play store Reviews Sentiment Analysis**

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

**Conclusion~**

* Average rating of (active) apps on Google Play Store is 4.17.
* Thus, a paid app that is higher in size may not perform well in the market.
* Users prefer to pay for apps that are light-weighted
* Most of the top rated apps are optimally sized between ~2MB to ~40MB - neither too light nor too heavy.
* Most of the top rated apps are optimally priced between ~1to 30 - neither too cheap nor too expensive.
* Users tend to download a given app more if it has been reviewed by a large number of people.
* Health and Fitness apps receive more than 85% positive reviews.
* Game and Social apps receive mixed feedback 50% positive and 50% negative

**References-**

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya