**Capstone Project-1 Submission**

**Play Store App Review Analysis**

**Vinayaka R Kulkarni**

**Prajwal Singh**

**Jayanthi Challagunda**

**Samarth Kushwaha**

**-- Data science trainees,**

**AlmaBetter, Bangalore**

----------------------------------------------------------------------------------------------------------------------------------------------

**TEAM Members:**

**Vinayaka R Kulkarni -** [vinayakark94@gmail.com](mailto:vinayakark94@gmail.com)

**Prajwal Singh -** [singhprajwal048@gmail.com](mailto:singhprajwal048@gmail.com)

**Jayanthi Challagunda –** [jayanthi0c@gmail.com](mailto:jayanthi0c@gmail.com)

**Samarth Kushwaha –** [kushwahasa@rknec.edu](mailto:kushwahasa@rknec.edu)

**GitHub Link~~**

**Prajwal Singh-**  <https://github.com/PrajwalSingh048/Playstore-App-Review-Analysis>

**Abstract -** A few thousand new applications are added to Google Play Store each day, with a progressively enormous number of designers working independently or in groups to make them successful, with the enormous challenge coming from everywhere. Since most Play Store applications are free, the income model is obscure and inaccessible regarding how in-app purchases, advertisements, and memberships contribute to an application's success. As a result, the success of an application is usually determined not by the amount of revenue it brings in, but by the number of installations and assessments it receives. Users rate applications (apps) voluntarily and serve as important criteria for evaluating them. However, these ratings can often be biased due to insufficient or missing votes. Additionally, significant differences are observed between numeric ratings and user reviews. We 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.

# Problem Statement

The data comes from the Google Play store dataset. There are various entries regarding each app in each row. This data set will be used for exploratory data analysis, which is an important step in the data science process because it prepares data for further modelling used in machine learning algorithms as well as securing very initial business decisions. To help us draw some preliminary conclusions about the chances of a newly launched app's success, we will structure the data, clean it, and present certain trends we observe.

# Google Play store and User Review Analysis

In today’s abstract we will see that traveling apps perform a vital perform in some individual’s life. it's happened that the expansion of the traveling application promote has an out of this world impact on leading modification. Having aforesaid that, related to the sometimes developing adjustable application reveal expert is to boot a far-famed ascension of moveable use designers inevitably inflicting success extreme as will be profit for one worldwide light-weight application producing.

The quick development of advanced cells, transferable applications (Mobile Apps) have clothed to be basic things of our lives. All constant, it's onerous for the USA to follow at the side of the facts and to understand everything relating to the apps as new applications ar coming back into the market on a daily basis. it's accounted for that automaton market achieved a substantial portion of one,000,000 applications in solar calendar month 2011. starting presently, 0.675 million automaton applications ar accessible on Google Play App Store. Such loads of applications ar by all accounts a rare open door for purchasers to urge from a decent determination extend. we've got a bent to trust versatile application purchasers ponder on-line application surveys as an interesting impact for paid applications. it's creating a trial for a attainable shopper to examine all the literary remarks and rating to settle on a variety. additionally, application engineers experience issues in discovering some way to enhance the appliance execution hooked in to evaluations alone and would profit by understanding a large vary of written remarks.

We develop automaton apps & unharness on Play Store. As AN Developer or say Business Perspective it’s important to understand whether or not users area unit enjoying the app or facing any problems. to understand this Play Store incorporates a Ratings & reviews section for every app free on play store. Users will submit the ratings and incorporates a freedom to write down a review for a specific app. This approach is kind of a long to rate & review app i.e. navigate to Play store to submit feedback or direct effort a current app work flow to open Play Store App link mistreatment URI. we have a tendency to ne'er wished our customers to go away our application, however with this flow, we tend to area unit forced to direct the management to Play store app.

# Google Play store Dataset

The dataset consists of Google play store application and is taken from Almabetter, which is the world’s largest community for data scientists to explore, analyze and share data.

This dataset is for Web scratched information of 10k Play Store applications to analyze the market of android. Here it is a downloaded dataset which a user can use to examine the Android market of different use of classifications music, camera etc. With the assistance of this, client can predict see whether any given application will get lower or higher rating level. This dataset can be moreover used for future references for the proposal of any application. Additionally, the disconnected dataset is picked so as to choose the estimate exactly as online data gets revived all around a great part of the time. With the assistance of this dataset, We will examine various qualities like rating, free or paid and so forth utilizing.

### The Data Set contains the following columns:

* **App:** This Column contains the name of the app
* **Category:** This contains the category to which the app belongs. The category column contains 33 unique values.
* **Rating:** This column contains the average value of the individual rating the app has received on the play store. Individual rating values can vary between 0 to 5.
* **Reviews:** This column contains the number of people that have given their feedback for the app.
* **Size:** This column contains the size of the app i.e. The memory space that the app occupies on the device after installation.
* **Installs:** This column indicates the number of time that the app has been downloaded from the play store, these are approximate values and not absolute values.
* **Type:** This column contains only two values- free and paid. They indicate whether the user must pay money to install the app on their device or not.
* **Price:** For paid apps this column contains the price of the app, for free apps it contains the value 0.
* **Content Rating:** It indicates the targeted audience of the app and their age group.
* **Genre:** This column contains to which genre the app belongs to, genre can be considered as a sub division of Category.
* **Last updated:** This column contains the info about the date on which the last update for the app was launched.
* **Current version:** Contains information about the current version of the app available on the play store.
* **Android version:** Contains information about the version of the android OS on which the app can be installed.

# User Review Dataset

* User reviews data frame has 64295 rows and 5 columns. The 5 columns are identified as follows:
* **App:** Contains the name of the app with a short description (optional).
* **Translated Review:** It contains the English translation of the review dropped by the user of the app.
* **Sentiment:** It gives the attitude/emotion of the writer. It can be ‘Positive’, ‘Negative’, or ‘Neutral’.
* **Sentiment Polarity:**It gives the polarity of the review. Its range is [-1,1], where 1 means ‘Positive statement’ and -1 means a ‘Negative statement’.
* **Sentiment Subjectivity:** This value gives how close a reviewer’s opinion is to the opinion of the general public. Its range is [0,1]. Higher the subjectivity, closer is the reviewer’s opinion to the opinion of the general public, and lower subjectivity indicates the review is more of a factual information.

# Python

Most of the info scientist use python due to the good built-in library functions and therefore the decent community. Python now has 70,000 libraries. Python is simplest programing language to select up compared to other language. That is the most reason data scientists use python more often, for machine learning and data processing data analyst want to use some language which is straightforward to use. That is one among the most reasons to use python. Specifically, for data scientist the foremost popular data inbuilt open-source library is named panda. As we have seen earlier in our previous assignment once we got to plot scatterplot, heat maps, graphs, 3-dimensional data python built-in library comes very helpful.

# Data Cleaning and Preparation

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.

1. We write a function play store info (), that will display 5 attributes about all the columns: Data type, Count of non-null values,
2. We start of by Finding the percentage of null values in each column by using isnull().sum() function
3. We drop the null values from columns ‘Current Ver’, ‘Android Ver’ and ‘Type’ from our dataset using the .notna() function of the pandas library.
4. We head on to find mean and median in the Rating column excluding the NaN values using the median() and mean()
5. Moving ahead we plot an sns.distplot of the “Rating” Column. We replaced all the “NAN” vales with the median value of Rating column using .fillna() function of the pandas library.
6. Handling the duplicates in the App column we drop the no of duplicate rows that are present in the App columns.
7. 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 'Lambda’ function.
8. 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 using the 'Lambda’ function.
9. Next, we change the datatype of Last updated to date\_time for easy processing.
10. We now Bring the size column to same units that is "MB” using the lambda function.
11. Converting the reviews section to type “INT”.

# 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

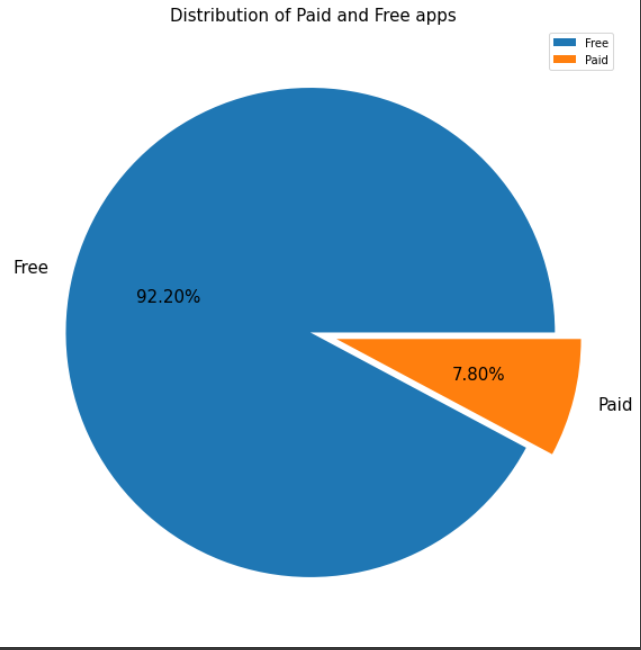


Fig -1: Free vs Paid

Here we can see that 92.2% 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.

## Rating:

In the below plot, we plotted the apps Rating

Chart, histogram

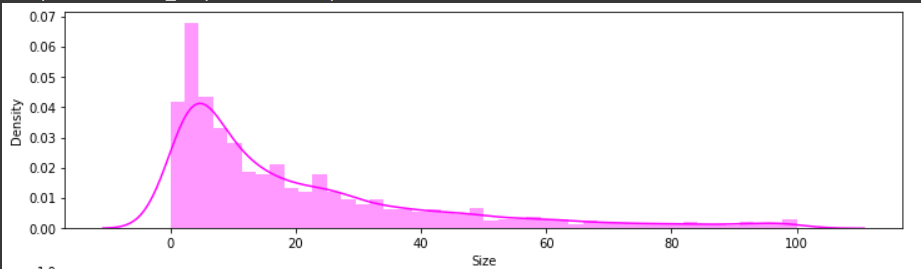
Description automatically generated

**Fig -2**: Distribution of App rating

* 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.

## Distribution of App Size

The below curve represents the variation of the size of apps available on Google Play store

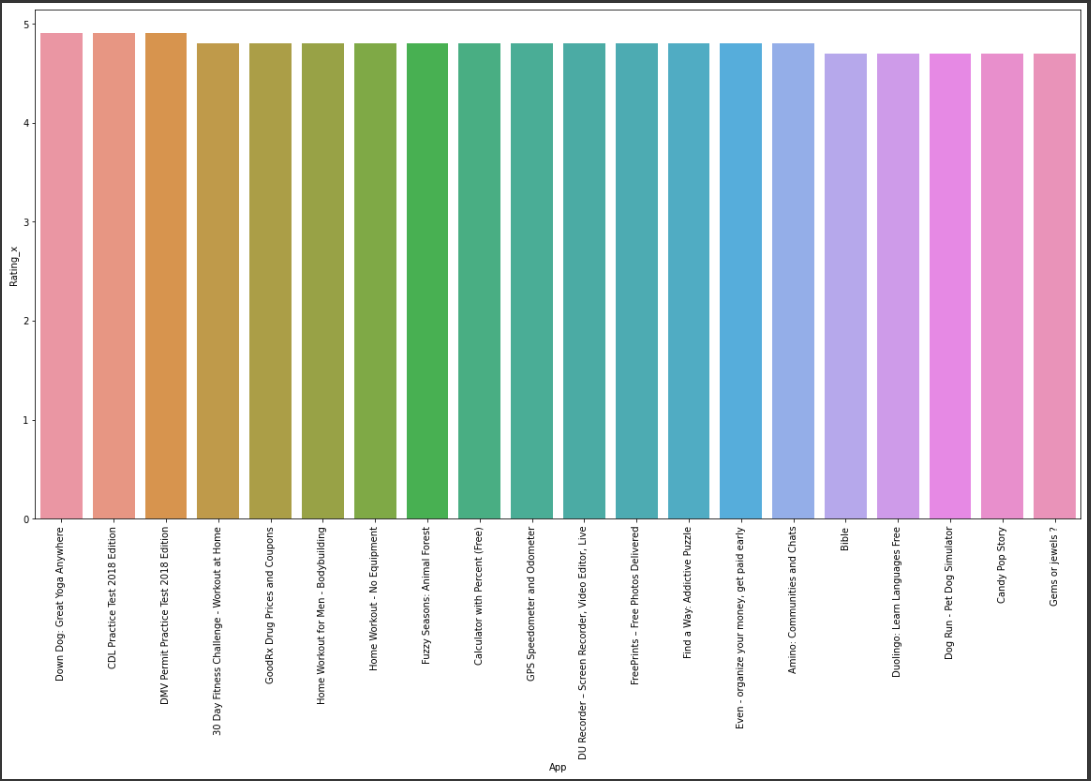


**Fig -5**: Distribution of App Size

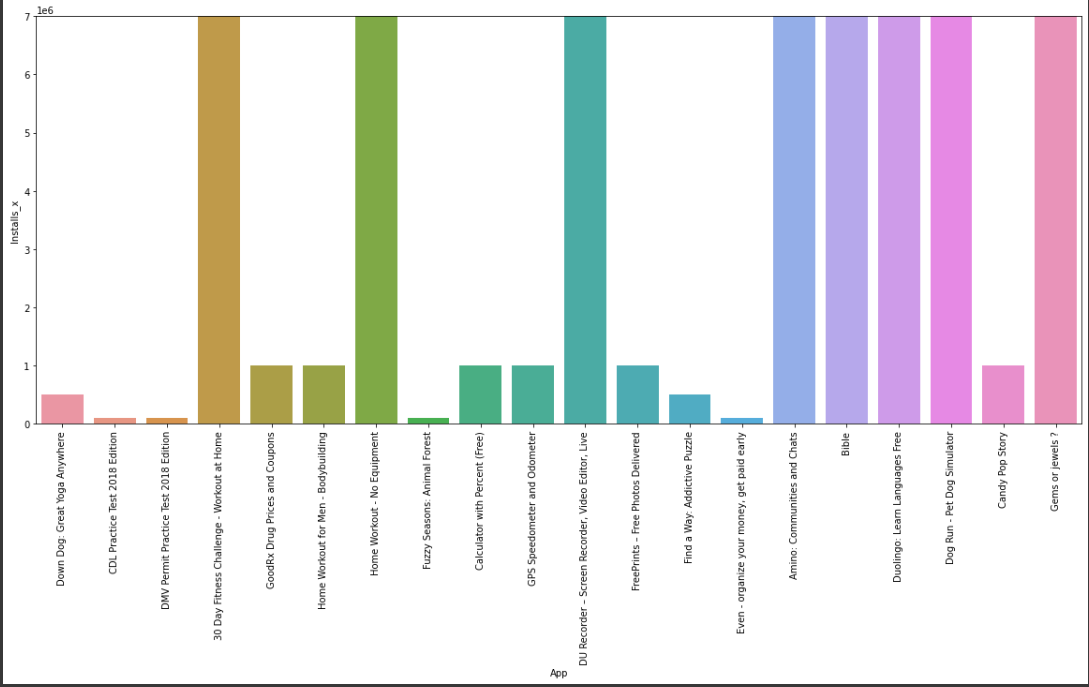
* IIt is clear from the visualizations that the data in the **Size** column is skewed towards the right.
* Also, we see that a vast majority of the entries in this column are of the value **Varies with device**, replacing this with any central tendency value (mean or median) may give incorrect visualizations and results. Hence these values are left as it is.

## Most Popular Apps [20]

The below



**Fig –6 Top 20 Apps with highest ratings**

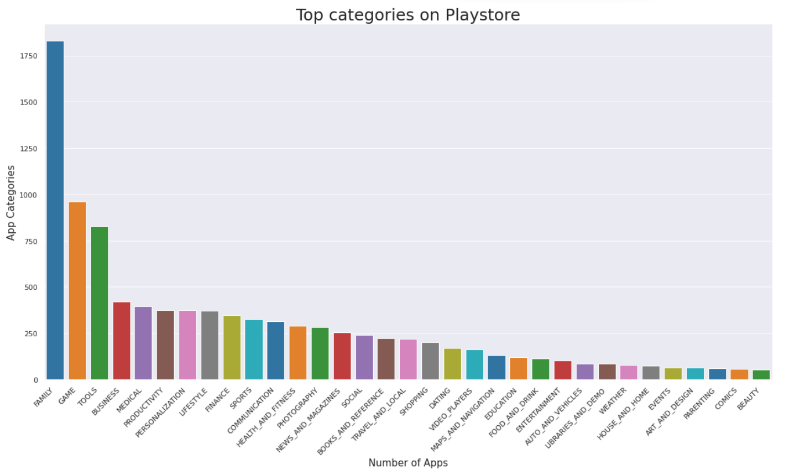


**Fig –7 Number of installations**

* 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 Category of Play store

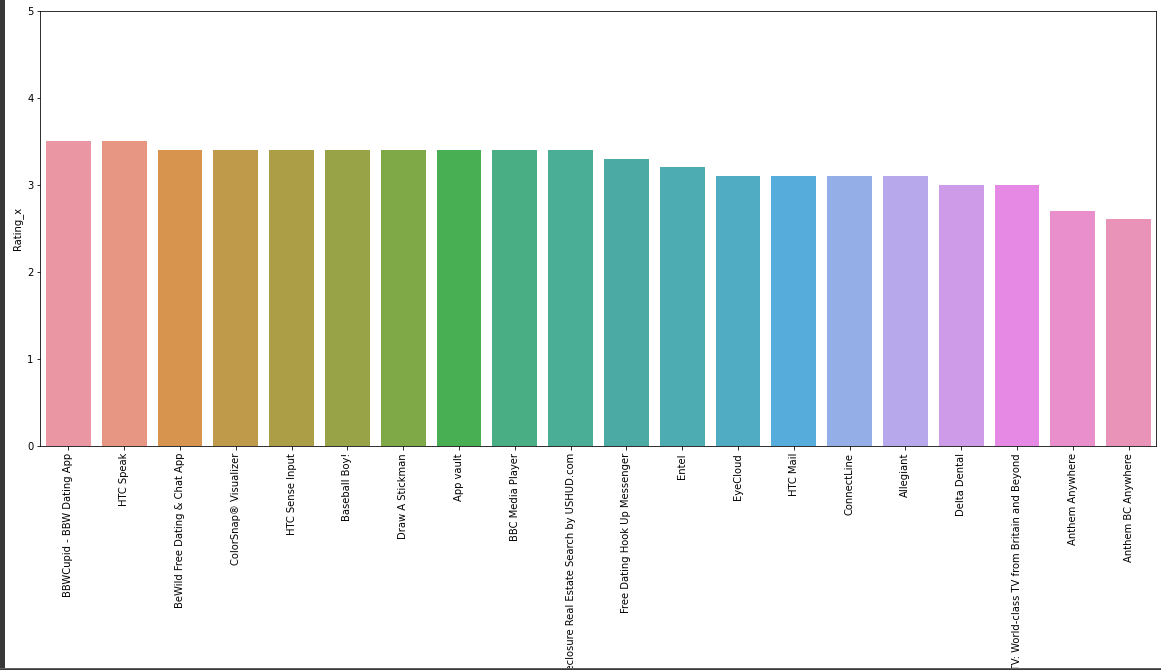
There are lot of category wise apps available on play store so the below curve shows how the apps are distributed.



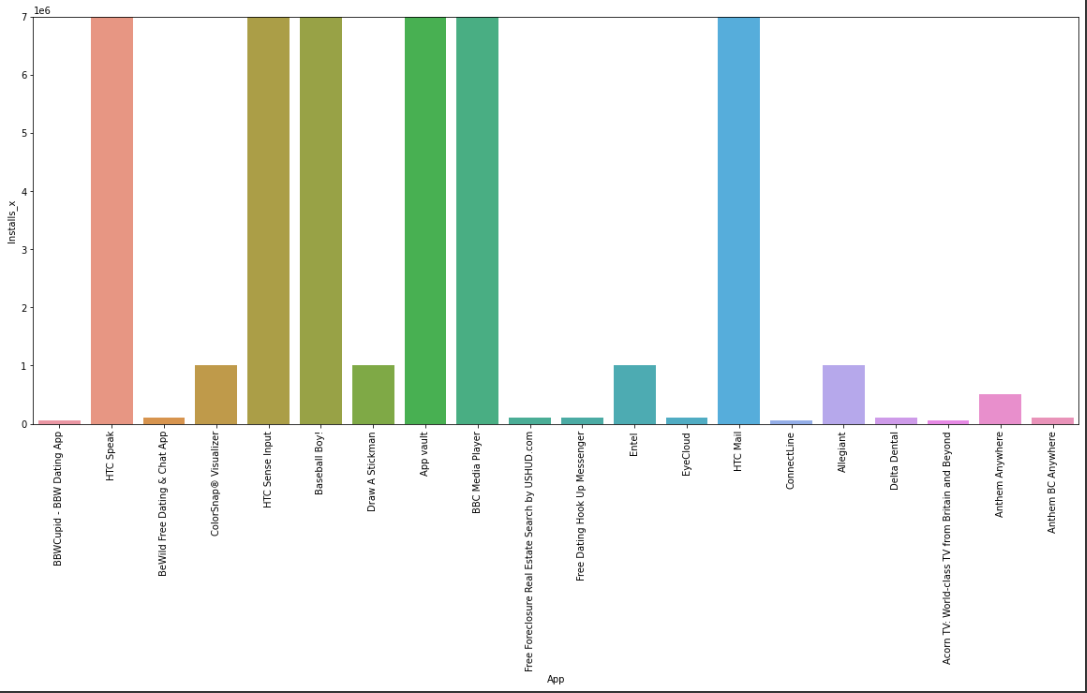
**Fig -8: Top Categories on Playstore**

So, 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.

## Bottom 20 in terms of Popularity



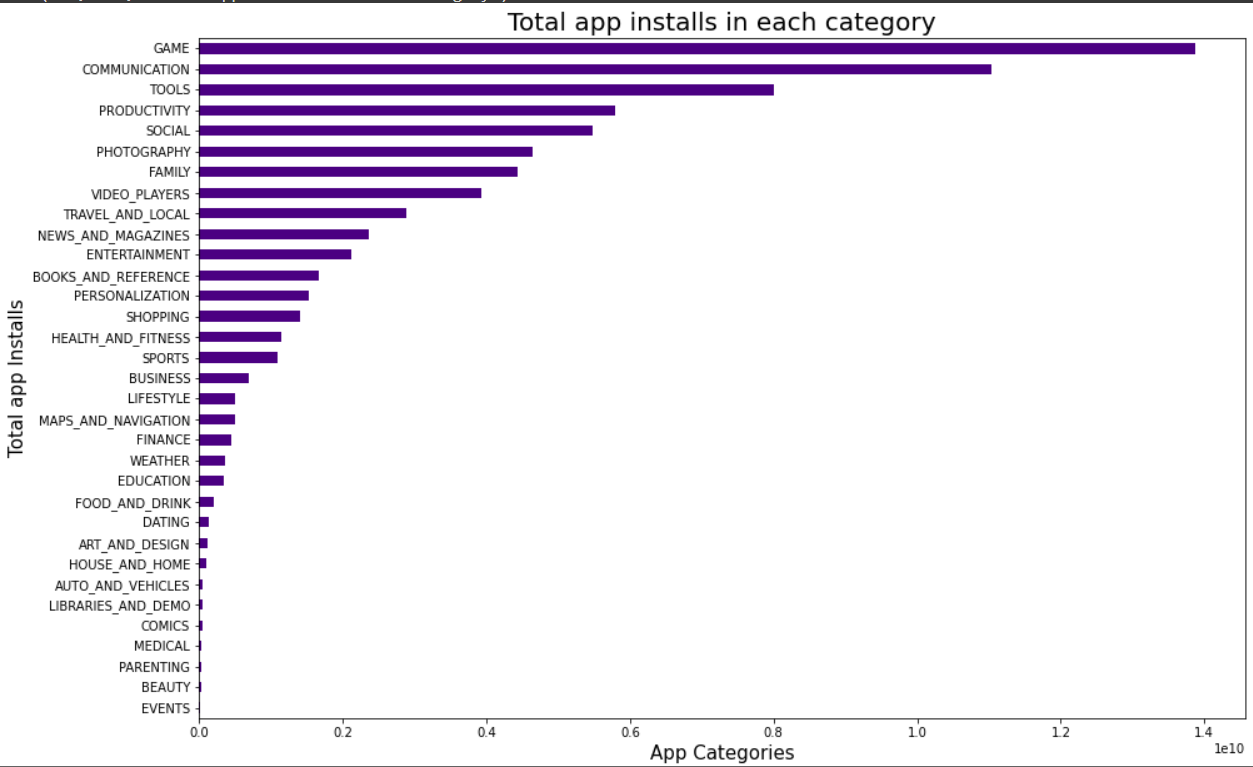
**Fig -9: Least Popular apps**



**Fig –10 No. of installs**

* This picture conveys that least popular apps have low ratings on playstore.

## Which Category apps are most installed



**Fig -11: Most Popular Category**

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

**Conclusion:**

Through exploratory data analysis we have observed some trends and have made some assumptions that might lead to app success among the users in the play store.

* Percentage of free apps = ~92%
* Percentage of paid apps= 7.80%
* Which category apps are most installed? -- Game
* Which is the topmost category (in terms of number of apps) on play store – Family
* Top 3 Apps in terms of ratings are- Down Dog: Great Yoga Anywhere, CDL practice test 2018 Edition, DMV Permit practice.
* Lowest 3 apps in terms of rating are -- Anthem Anywhere, Anthem BC Anywhere.

**References~**

* GeeksforGeeks
* Analytics Vidhya
* Stackoverflow
* Towards data science
* Python libraries documentation.

Thank You