

Google Play Store Apps

BAN 693-T Transitional Capstone Project under the guidance of Dr. Chonqui Wu





# Team –

Komal Jaiswal (KQ2587)

Sindhuja Neravetla (QF7287)

# Table of Contents

|  |  |
| --- | --- |
| Introduction….............................................................................. | 3 |
| Data……………………………………………………………... | 4 |
| **Part I** |  |
| Step by Step Process……………………………………………... | 4 |
| Summary Statistics………………………………………………. | 5 |
| Correlation Matrix ………………………………………………. | 8 |
| Model Performance……………………………………………… | 8 |
| Benchmark………………………………………………………. | 8 |
| Model Evaluation and Validation………………………………... | 9 |
| Conclusion………………………………………………………... | 9 |
| **Part II** |  |
| Step by Step Process……………………………………………… | 10 |
| Summary Statistics………………………………………………... | 11 |
| Regression Model to Predict Installs……………………………… | 15 |
| Conclusion…………………………………………………………. | 19 |
| Future Scope………………………………………………………... | 19 |
| References…………………………………………………………... | 20 |

# Introduction

Google Play is a digital distribution service developed by Google, an app store for an android operating system. Applications are available through Google Play, either free of charge or at a cost.  A user can download these apps directly on the android device. Google Play contains all kinds of applications like games, maps, To-do tasks, food, and so on. Almost 190 countries use the google play store, and the google play store generates more than double downloads of the Apple app store. They are easy to create and can be lucrative. Because of these two factors, more and more apps are being developed. In this project, 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.

This project lets us apply the skills from Data Analysis and Machine learning

What would be more useful than being able to predict the ratings of an app? Or what type of android app should be built for it to become successful? Or which month should the app be released to attract more users? The Google 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.

This project looks at the Android mobile apps from the perspective of an analyst for a company which builds both the free and paid mobile apps. Our job is to enable the team of developers to make data-driven decisions with respect to the kind of apps they build. To this end, we analyze free apps and paid apps by number of users to determine which kinds of apps are likely to attract more users. That way, the company's developers have data to inform what kind of apps they build.

Predictive analytics use prediction models that predict ratings of an app by assessing their number of downloads, the category it belongs, number of reviews on play store, and app size. We also tried to use prediction model that predicts the success, that is predicting the number of installations.

**We divided our analysis into two parts -**

**Part I:**

* Predict if an app would have a high rating or low rating knowing its number of downloads, the category it belongs, number of reviews on play store, and app size using classification methods like KNN and Support Vector machines.

**Part II:**

* To be able to recommend the developers what kind of free or paid application should they build for it to become successful.
* Which Month the app needs to be released for it to become successful.
* Predict Installations using Regression Model.

# Data

The Data set is collected from kaggle.com. While each row represents an app and each column contains that app’s attributes.



The raw data contains 10841 rows (apps) and 13 columns (features). At a quick glance, the columns that might be useful for the purpose of our analysis are ‘App’, ‘Category’, ‘Rating’, ‘Reviews’, ‘Size’, ’Price’, ’Installs’, ‘Content Rating’, ‘Genres’ and, ‘Last Updated’. In the first part of analysis our primary focus was on ‘Ratings’, where as in the second part we focused on categories, price and installation.

# Part I

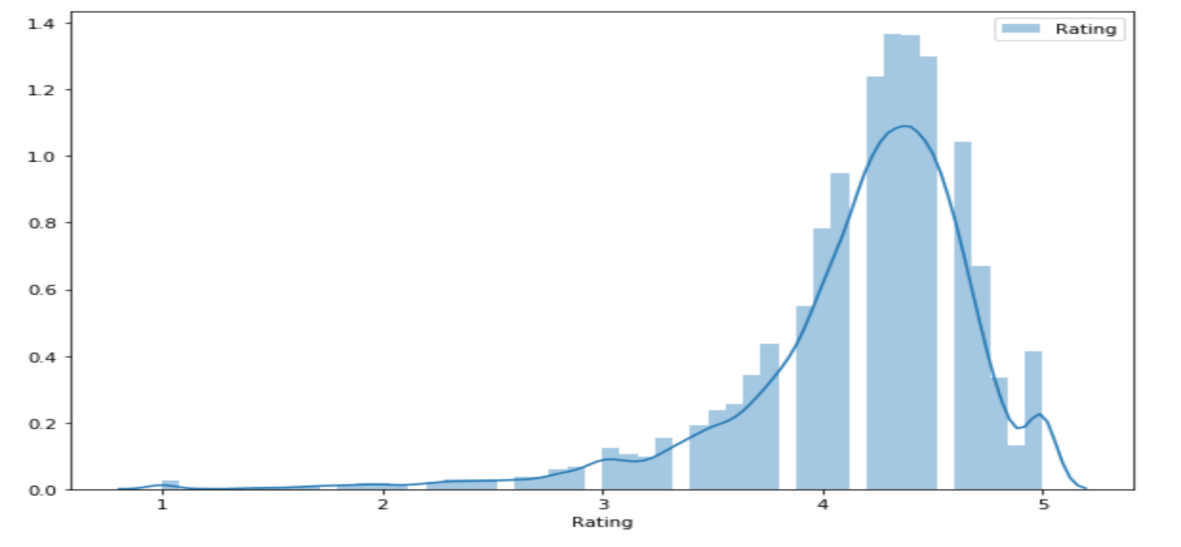
**Objective:** Predict Google Play Store App Ratings High or Low?

# Step by Step Process –

* Introducing and exploring data
* Data Cleaning
* Feature Engineering
* Correlation matrix
* K-Nearest Neighbor Model
* Random Forest Model
* Comparing Models based on metrics

# Summary Statistics –

## Rating:

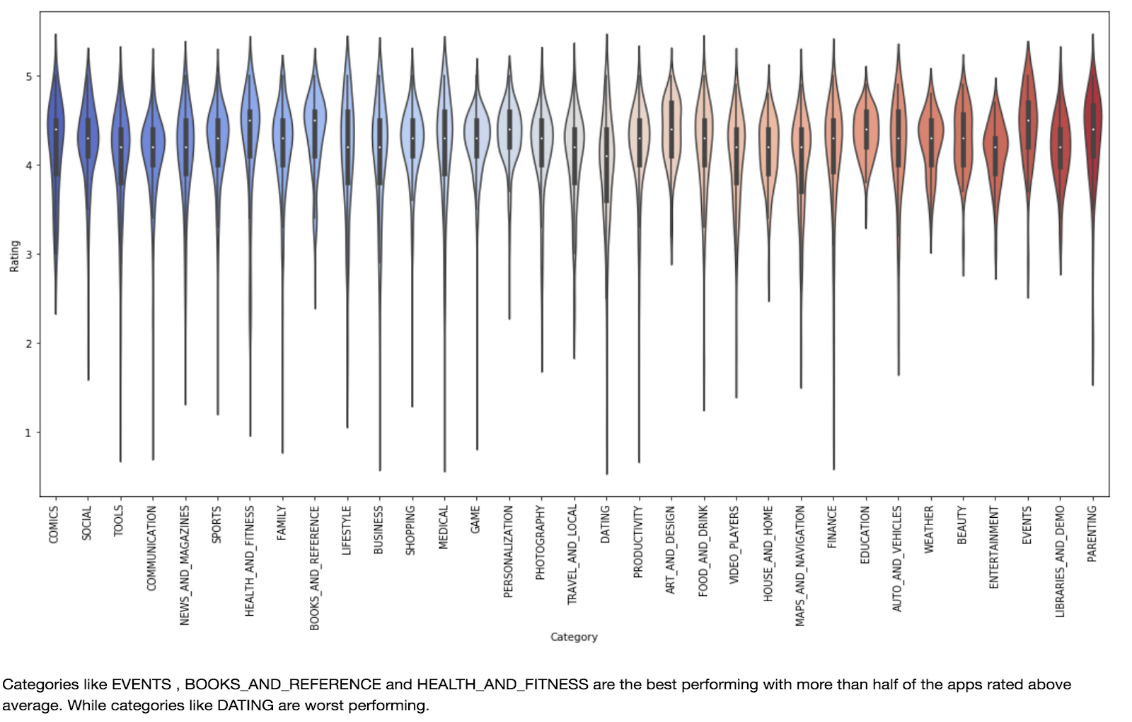


The average rating in the google play store is 4.173.

## Ratings Vs Categories

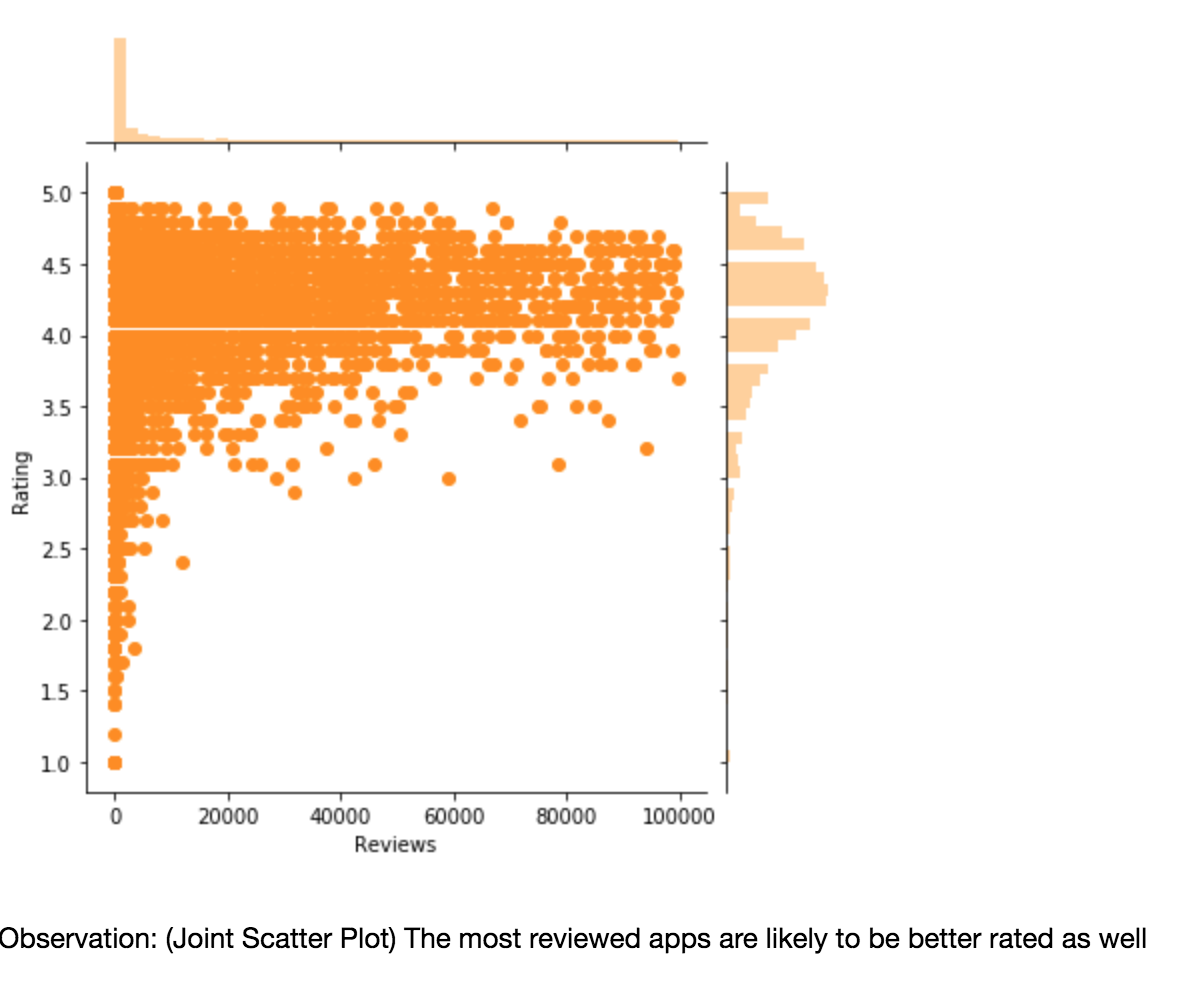
## https://lh5.googleusercontent.com/OcbgnBMgmFKCfCdasXU0VzoEd0kNXwo7jlGfcMkAtyBUubGDSuHuHT3IiQs_sF-0flELcVAlNNLeL1ccX35FYJUhkvNgFAli64xpKChBEHfz7MwYEB2KMIQJ0LhHguO_I_o_VPLv

Here, we can observe the rating is varied by the categories.



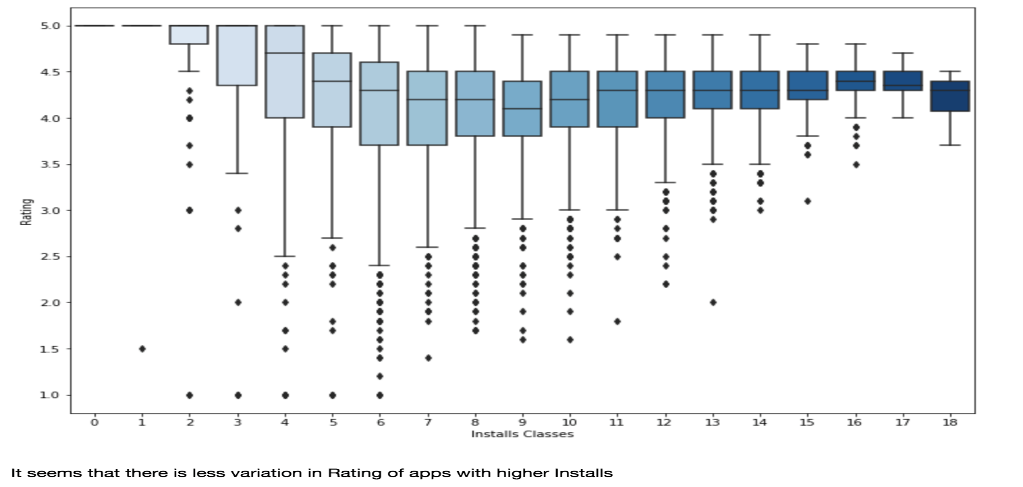
Categories like EVENT, BOOKS\_AND\_REFERENCE and Health&Fitness are the best performing with more than half of the apps rated above average. While Categories like DATING were worst performers.

## Rating Vs Reviews



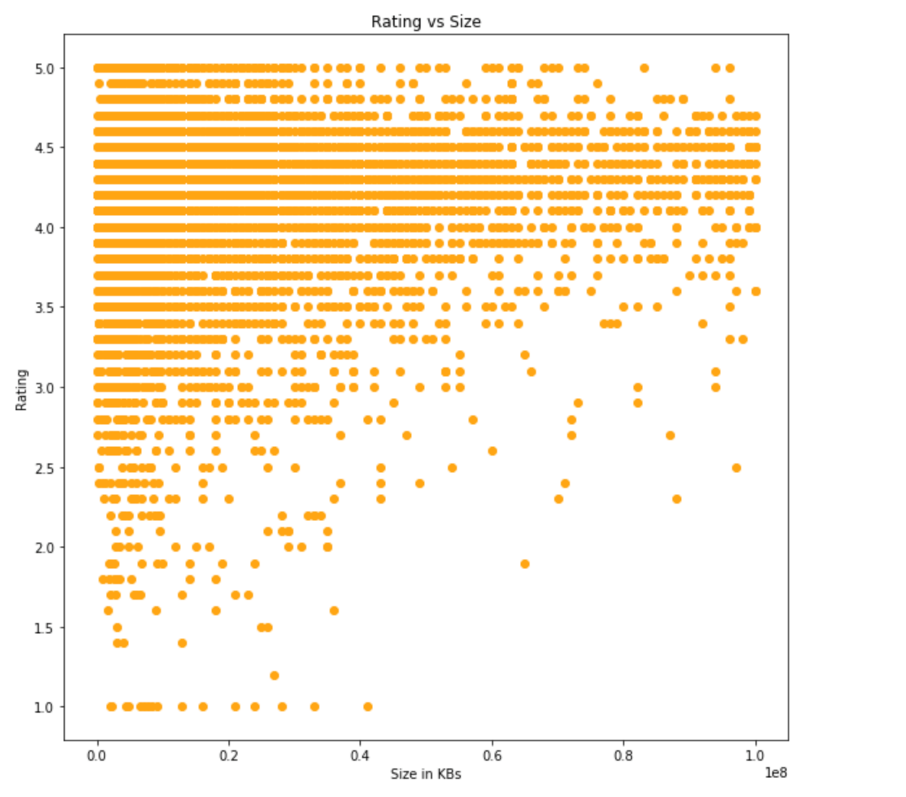
The most reviewed apps are most likely to be better rated as well.

## Rating vs installs:



It seems that there is less variation in Rating of apps with higher installs

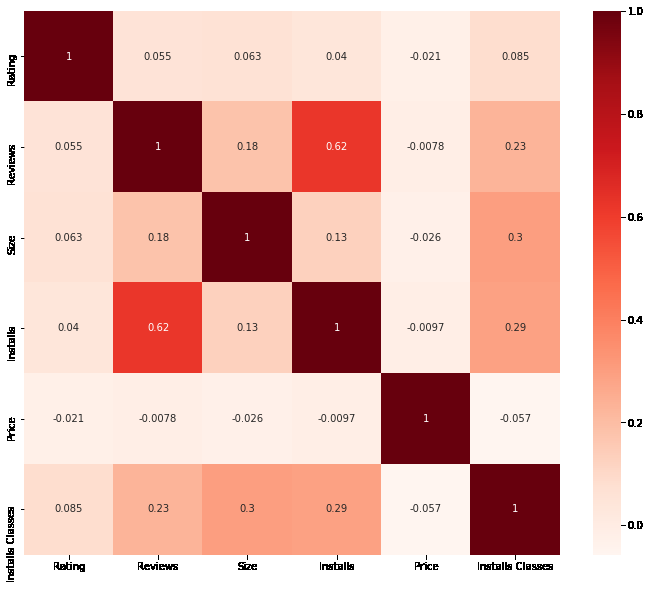
## Rating Vs Size



It is evident from the plot that the mid-sized apps tend to perform mostly better.

# Correlation Matrix

A correlation matrix is a table showing correlation coefficients between variables. Here, we used the Pearson Correlation Coefficients to determine the correlation. The values lie between -1 and 1. From the correlation matrix below, we can see that there is a strong positive linear relationship between Reviews and Installs. It indicates, there will be an increase in reviews with rising in the number of installations. We can also see that there is a negative linear relationship between price and remaining variables. With an increase in the price, there will be a decrease in installations, ratings, reviews, and others. It could be because a lot of people prefer to use free apps.



# Model Performance

* The metric used to measure performance of our models is Accuracy.
* It is calculated by correctly classified points divided by the total number of points.
* Compare the two models K-Nearest Neighbors Model and Random Forest Model to check which model has the highest accuracy.

# Benchmark

After Reviewing few of models defined in the Kaggle Kernels, we defined our Benchmark as 85% Accuracy for the classification models. We used two different models like K-Nearest Neighbors Model and Random Forest Model to check the accuracy.

## K-Nearest Neighbors Model

The k-nearest neighbors’ algorithm is based around the simple idea of predicting unknown values by matching them with the most similar known values. Building the model consists only of storing the training dataset. To make a prediction for a new data point, the algorithm finds the closest data points in the training dataset — its "nearest neighbors".

## Random Forest Model

The RandomForestRegressor class of the sklearn.ensemble library is used to solve regression problems via random forest. The most important parameter of the RandomForestRegressor class is the n\_estimators parameter. This parameter defines the number of trees in the random forest.

## Model Evaluation and Validation

Table comparing different model performances:

|  |  |  |
| --- | --- | --- |
| **Models** | **Accuracy of Training Set** | **Accuracy of Test Set** |
| K-Nearest Neighbors’ Model | Train = 0.904 (90.4%) | Test = 0.907 (90.7%) |
| Random Forest Model | Train=0.907 (90.7%) | Test = 0.905 (90.5%) |

As we can see that both the algorithms KNN and Random forest gives good accuracy with approximately 90%, therefore these algorithms can be used for Predicting Ratings

We can apply any of these models to a sample and be correct 90% of the time.

## Final Results compared to the benchmark

As we have got 90% Accuracy, we have crossed our benchmark of 85%.Therefore the final model and results are significant enough to solve the problem adequately.

## Conclusion

* One of the exciting parts of the project was performing Exploratory Data Analysis
* After performing exploratory data analysis on the data-set we came out with the below analysis:
* The rating is on a scale of 1 - 5 with 1 being the lowest and 5 being highest
* The mean rating is 4.17 while the median rating is 4.3. This implies that average rating is greater than 4.1
* The minimum App size is 8.5 kb with maximum size being 100 Mb
* The FAMILY, EVENTS and TOOLS are the most dominating applications in the play store
* Categories like EVENTS , BOOKS\_AND\_REFERENCE and HEALTH\_AND\_FITNESS are the best performing with more than half of the apps rated above average.
* Categories like DATING are worst performing.
* The most reviewed apps are likely to be better rated as well
* It seems that there is less variation in Rating of apps with higher Installs
* Higher installed apps have less variation in their ratings.
* Mid-sized apps tend to perform mostly better.
* Only 7.3% of the Apps in the store are paid apps
* We can observe that even in the paid apps most of the apps are cheap!
* Even the most expensive apps do not have exceptionally better rating and even less rated than many cheaper apps.
* Pricing of the Apps is found to be mostly above Average(Normal)
* Too Expensive apps have average Rating less than others.
* The Adult or Unrated apps are almost nil
* There is not much effect of the Content Rating on the Rating, the better rating trend in adult and unrated apps is due to very few observations.
* The Dating genre is the least rated on average, but there is not a very vast difference to the highest averagely rated Events genre

**Predicting ratings of app: Whether it’s a low rating or high rating:**

* By observing the performance of the two models, we concluded that both the model's K-Nearest Neighbors and Random Forest  gives the highest accuracy.
* Therefore, these two models can be used to predict high and low ratings of google play store apps.

# **Part II**

## Objective

1. To be able to recommend the developers what kind of free or paid application should they build for it to become successful.
2. Which Month the app needs to be released in order for it to become successful.
3. Predict Installations using Regression Model.

## Step by Step Process

1. **Google Play Store App success**
   1. Exploring the Data
   2. Cleaning the Data
   3. Analyzing the Data (Free App):

Most Common Free Apps by Genre

Most Popular Free Apps by Genre

Most Popular Free Game Apps on Google Play Store

* 1. Analyzing the Data (Paid App):
  2. Most Common Paid Apps by Genre

Most Popular Paid Apps by Genre

Most Popular Paid Game Apps on Google Play Store

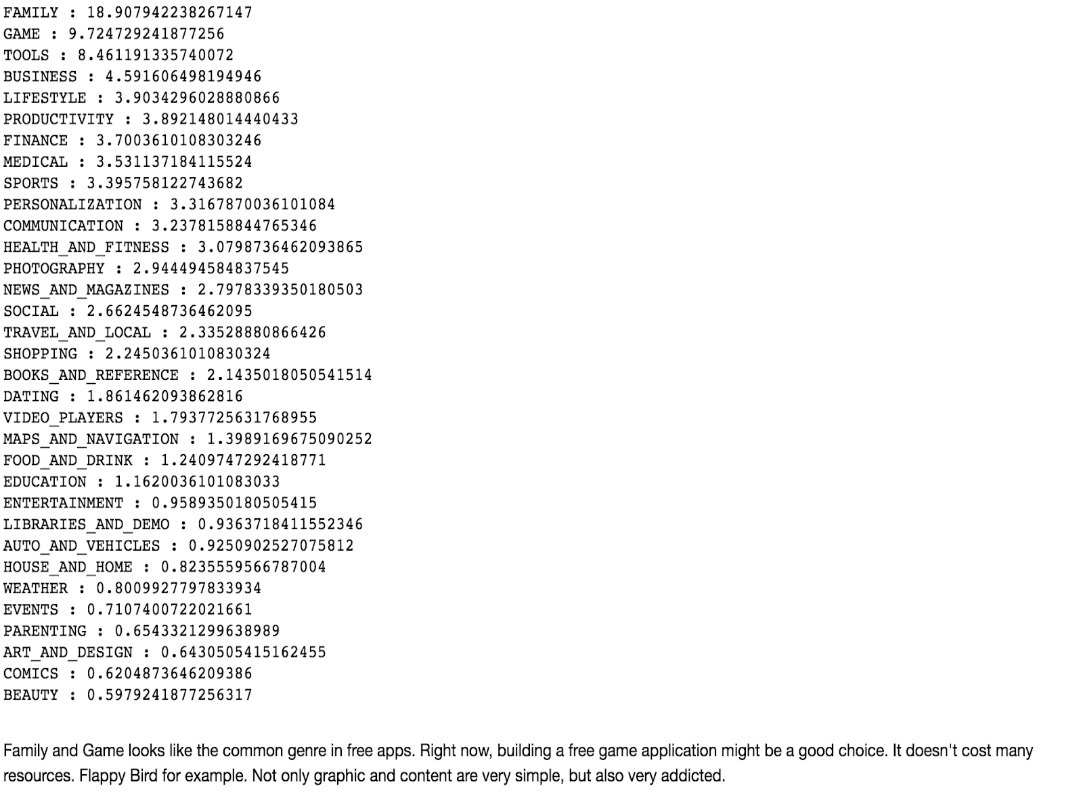
Most Popular Paid Game Apps on Google Play Store

1. **Google Play Store App -Which is the best month to release?**
2. Rating Versus Reviews Plot
3. Box plot for Rating Versus Month
4. Plot Reviews Versus Month to see which month has higher reviews
5. **Regression to Predict Installations**
6. Univariate analysis of Installs
7. Bivariate analysis of Installs
8. Content Rating Vs Installs
9. Category Vs Installs
10. Distribution of Installs Vs Rating
11. Cleaning the Data
12. Perform Regression

## Summary Statistics

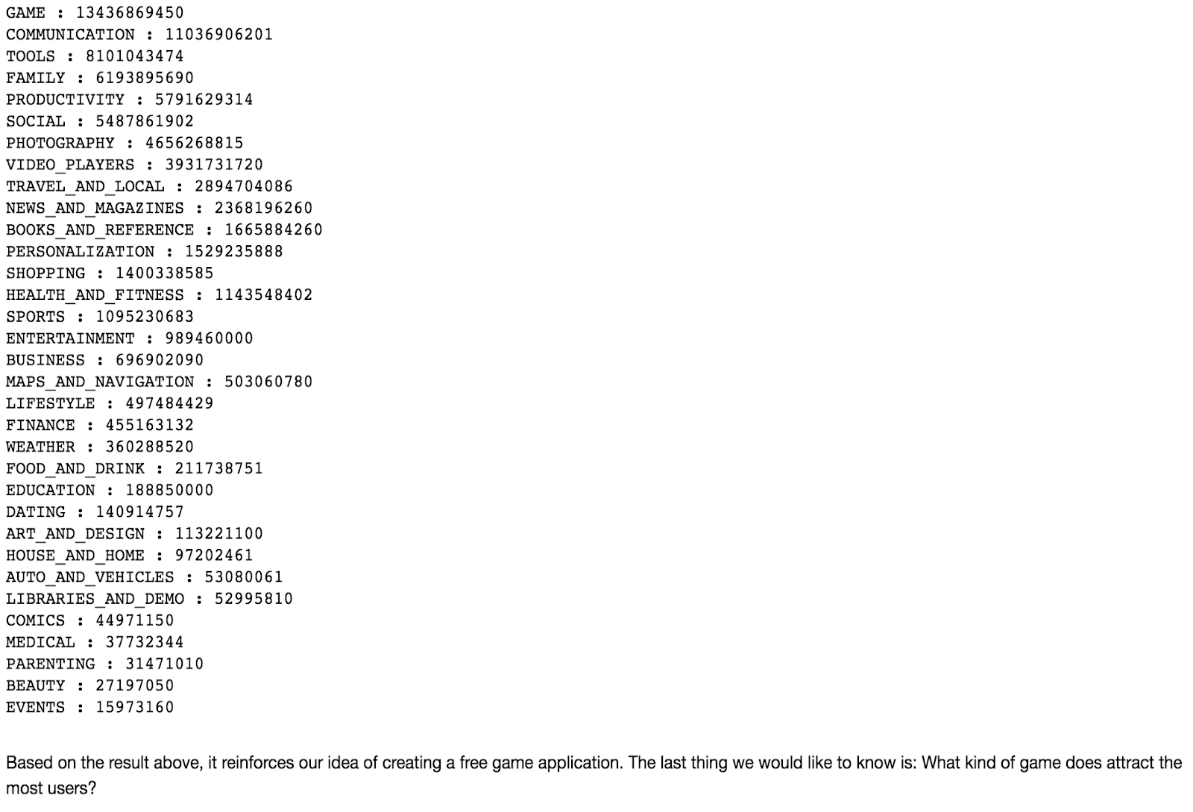
## Google Play Store App success

Most Common Free Apps by Genre -



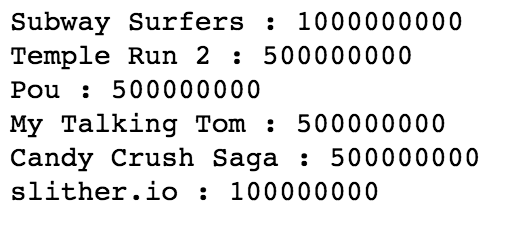
Family and Games looks like the most common genre in free apps. Right now, building a free game application might be a good choice. It doesn't cost many resources. Flappy Bird for example. Not only graphic and content are very simple, but also very addicted.

Most Popular Free Apps by Genre –

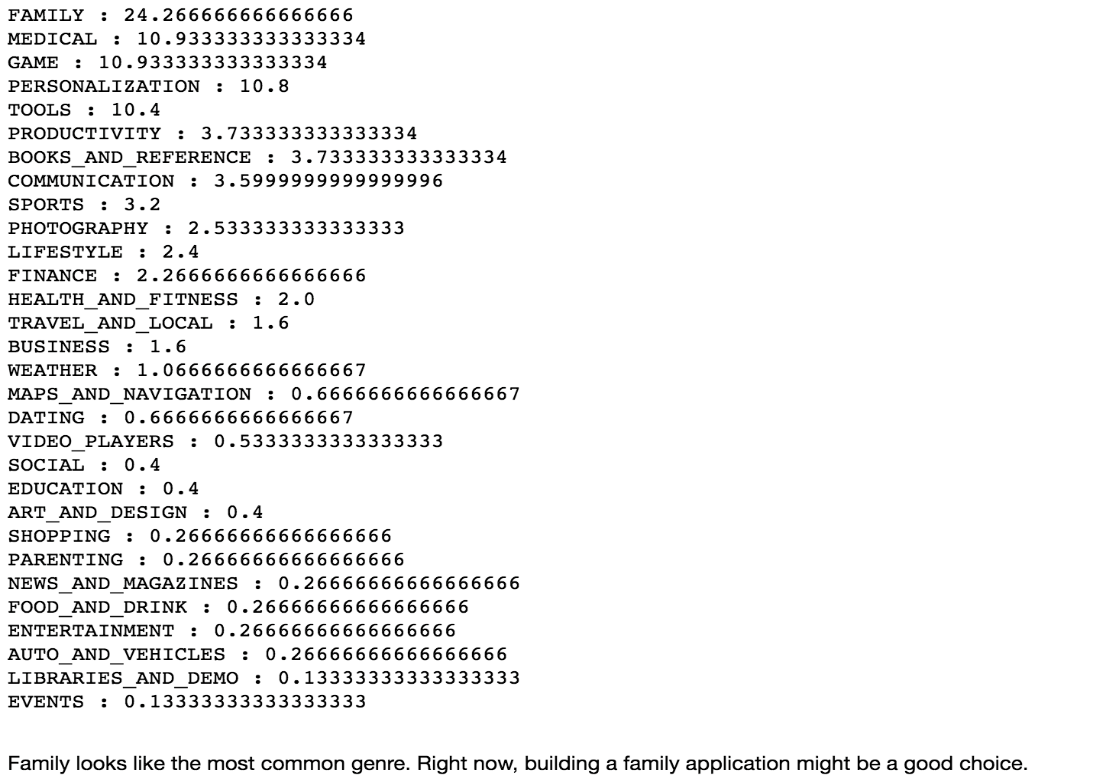


Based on the result above, it reinforces our idea of creating a free game application. The last thing we would like to know is: What kind of game does attract the most users?

Most Popular Free Game Apps in Google Play Store –

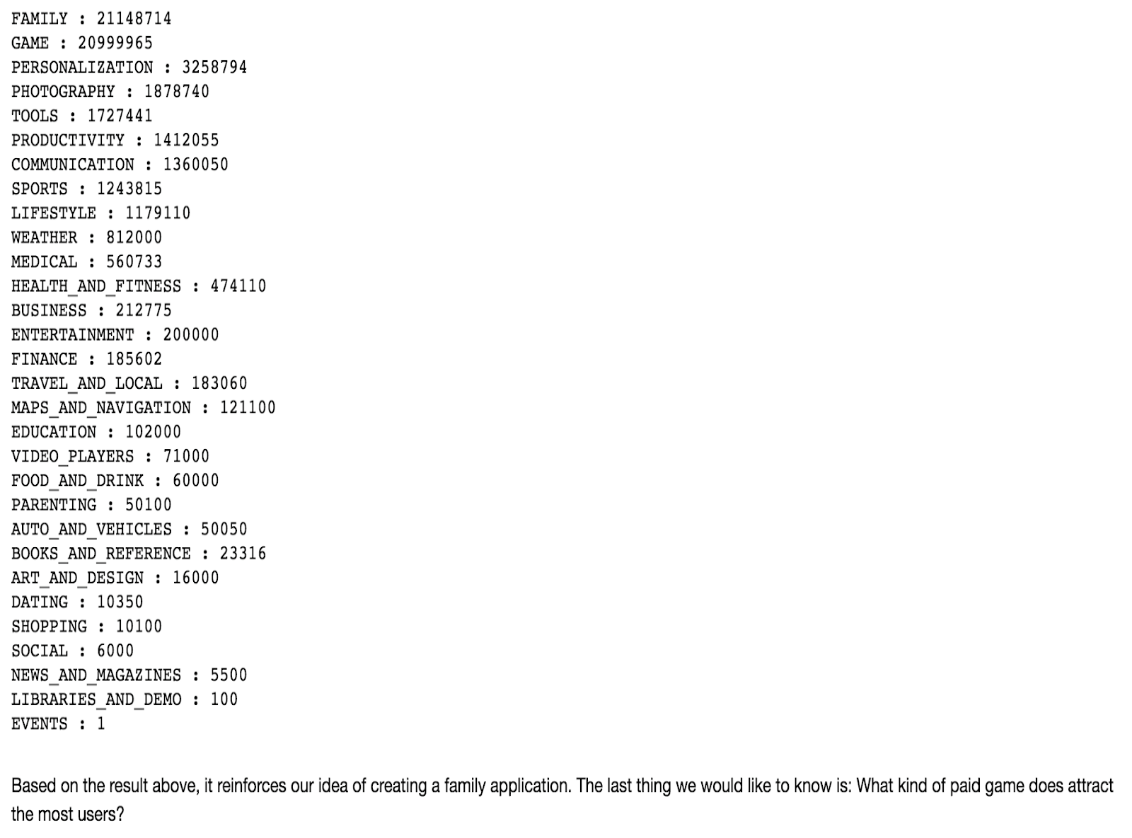


Most Popular Paid Apps by Genre –



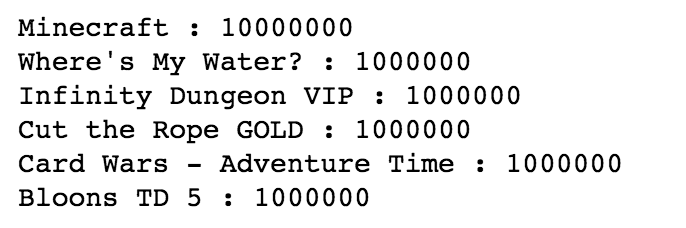
Family looks like the most common genre. Right now, building a family application might be a good choice.

Most Popular Paid Game Apps in Google Play Store –



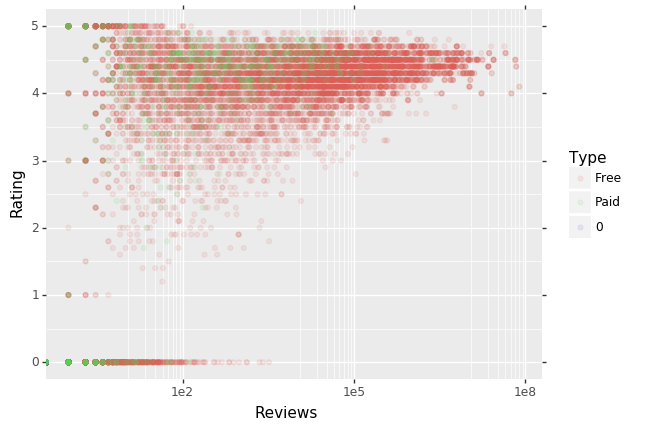
Based on the result above, it reinforces our idea of creating a family application. The last thing we would like to know is: What kind of paid game does attract the most users?

Most Popular Paid Game Apps in Google Play Store –



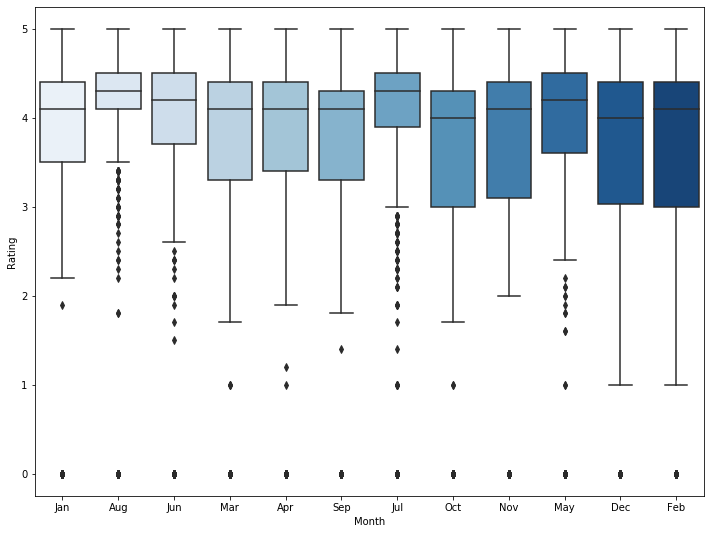
## Google Play Store App -Which is the best month to release?

Rating Versus Reviews –

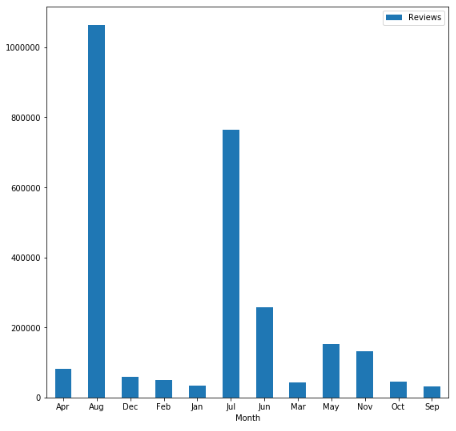


From this plot it is clear that apps with greater number of reviews always have higher rating.

Box Plot of Ratings by Month –



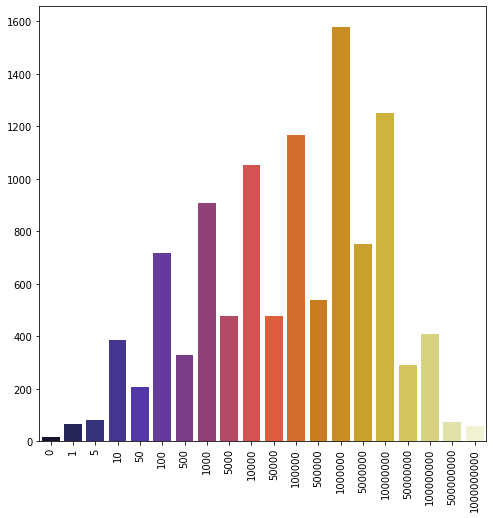
Plot for Reviews by Month –



The apps those were released or updated in the months of August and July have a huge number of reviews. Apps that are released during January and September did not receive great number of reviews.

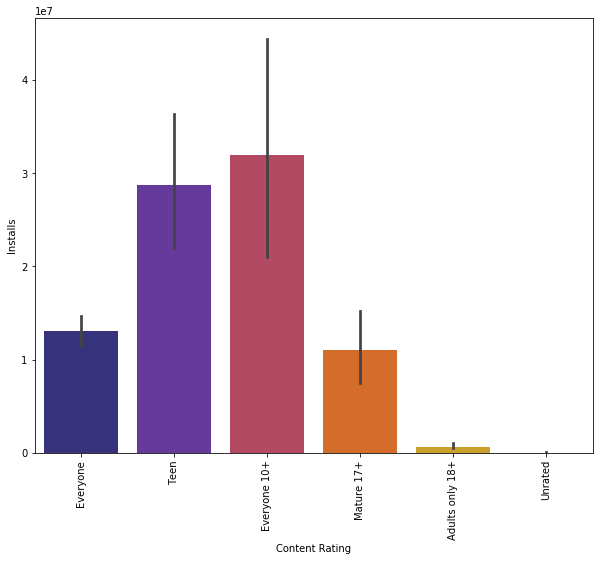
## Regression to Predict Installations

Univariate Analysis of Installs



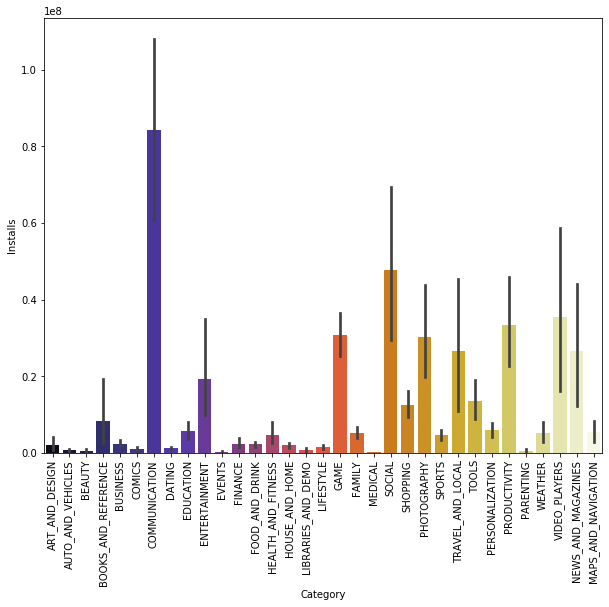
This Analysis shows the count of each Installation value appearing in the dataset. The dataset contains more than 1500 apps with 1M installations. Almost zero apps with no downloads, it means every app is at least downloaded by one user.

Bivariate Analysis of Installs Vs Ratings



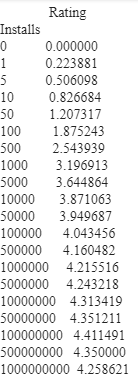
From the above Bivariate analysis of Installs and content rating it is clear that apps with content rating Everyone 10+ has a greater number of installs.

Bivariate Analysis of Installs Vs Category –



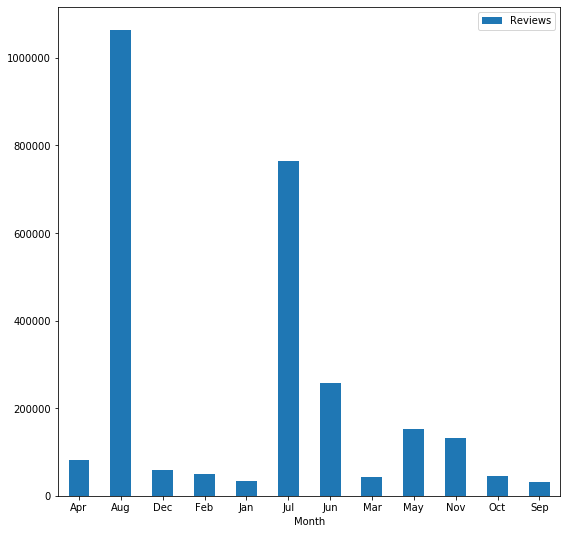
It is clear that communication, Social and Game applications are the most downloaded apps. The least installed are Parenting, Events and, Beauty apps.

Distribution of Installs Vs Rating -



There are no apps with zero rating. There are 1Billion apps with 4.26 Rating and 100 Million apps with 4.41 ratings.

Reviews Vs Month -

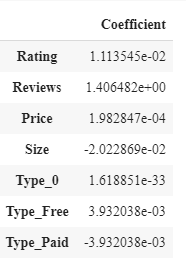


From these plots, we can see that more reviews often means high ratings. And developers should consider releasing apps and updates in the summer months, preferably August.

# C. Regression Model to Predict Installs

Multi Linear Regression – Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

Here, we used multiple regression to predict number of installs based on rating, reviews, size, content rating, price and type(Free/Paid). This model is used to see how these variables influence a user to download an application.



From the coefficients above we can form an equation –

Installs = 0.01 (Rating) + 1.41 (Reviews) + 0.02 (Price) – 0.20 (Size) + 0.003 (Type Free) – 0.33 (Type Paid)

The above equation shows us the relationship between Installs and these variables. These measures tell us the amount of increase/decrease in installs that would be predicted by a 1-unit increase/ decrease in variables. There will be 14 % increase in the installs with increase in the number of reviews, similarly there will be 20% decrease in the number if installs with 1 unit increase in size.

To evaluate the performance of regression model we used three measures - Mean Absolute Error, Mean Squared Error and, Root Mean Squared Error.

Mean Absolute Error - MAE is the sum of absolute differences between our target and predicted variables.

Mean Squared Error – It tells how close a regression line is to a set of points.

Root Mean Squared Error - It is the standard deviation of predicted errors.



# Conclusion:

1. **Google Play Store App success**

**Free app –**

Arcade and Casual are the most popular game genre on Android. Candy Crush is a dominator in its category.

Arcade game would be the best choice. Subway Surfers, Temple Run are endless runner mobile games. They are easy to

play, nice graphics and users may pay attention during gameplay which is a good thing because they could see our ads.

**Finally, what kind of free app developers should build -**

--Arcade game

--Nice graphic

--Easy and fun to play like Subway Surfers, Temple Run, Candy Crush.

**Paid app -**

Family apps are the most popular paid apps on Android.

Family Apps similar to Minecraft, Where's My Water?

Infinity Dungeon VIP, Cut the Rope GOLD, Card Wars - Adventure Time and Blooms TD 5 can be built

1. **Google Play Store App -Which is the best month to release?**

It is clear from the analysis that, more reviews often mean higher ratings. And it is also seen that there is higher number of reviews during the summer season i.e., during the months of July and August. From this it is clear that the apps receive greater number of reviews and rating during summer which is the holiday season. That will be reasonable for the developers to release their app during these months to have a huge success.

1. **Regression to Predict Installations**

It can be seen from our analysis that, apps with higher ratings and reviews has greater downloads compared to others. It could be because the user who wants to download an app will always check for reviews and ratings to see how good it is and also how far it can help them. So, maintaining a greater number of reviews and ratings for the app will be ideal. The communication and social apps are the apps with more downloads.

# Future Scope:

* Find Profitable app profiles for Google Play Store
* Predict Reviews
* Predict app success more accurately with data like Developer information.
* We can make more accurate prediction of app usage by using more variables like android version, Current version of the app because these two variables also paly a key role while installing an application.

# References:

<https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222>

<https://www.kaggle.com/>

<https://www.kaggle.com/lava18/google-play-store-apps>