**Exploratory Data Analysis on Google Play Store Apps Reviews.**

Ashish Bhaidas Mali,

Data science trainee,

AlmaBetter, Bangalore.

# Abstract

With the rapid development of Internet technology and the global penetration of the internet, there is a surge in demand for customized applications for every need. As a result, several application distribution platforms are available to meet this demand. The Google Play Store is the most popular of these. It is the world's largest app store, hosting around 3 million apps in 2020, with 2.5 billion active Android users. Google App Store is a brand in and of itself, with over 2 billion active monthly users. The fact that the Google Android app store has received over 115 billion downloads in the last year demonstrates its growing popularity. With the growing popularity of the Android platform around the world, there has been a growing demand for Android apps to be recognized separately. Furthermore, mobile developers desired to upload Android apps and provide ongoing support. So, for that app, it is critical to understand which factors are responsible for the application's success in the Play Store. The goal of this project is to provide insights that will help developers better understand customer demands and thus popularize the product. We attempted to discover relationships between various attributes such as free or paid, reviews, application's rating, sentiment analysis so on and so forth. And find out the insides which then utilized by developer to build and launch the application on google play store

***Keywords: google play store, developer, app, attribute, data analysis, insides, sentiment analysis.***

# 1.Problem Statement

The Data is collected form the google play store dataset. Contains two files as “play store data.csv” and “user review data.csv” in the first file per application data is reflected and in the second file we have a user review and its variables.

Growing demand of the application and rapid growth in internet, there is surge in demand of the application for every need. Right form education to entertainment. So as per the rule of supply and demand we see the surge of application in the google play store. But the problem here is very few applications make it in a list of successful application. So, what’s drive the success of the application in the play store and what factors should developer consider to develop and launch the application and make that application on the top of the list. We analyze the successful application and draw some insides that help make better application.

# 2. Introduction

In today’s scenario we can see that mobile apps playing an important role in any individual’s life. With enormous challenge from everywhere throughout the globe, it is important for a designer to realize that he/she is continuing in the right way or not. 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 dataset with 10k Play Store applications is available to analyse the market of android. It can be examined to analysis the different category such as family, communication, entertainment, tools, music, camera etc. In this project we examine the different attributes present in the data set that affect the popularity of the application. We focused on to answer the questions like, what makes an app popular, what should be the price and size of the app, is there some trends in user sentiments. In our data set we have two csv files for data analysis: Play Store data User Reviews At first, we analysis the play store data and in the play store data we have 10841 rows and 13 columns & in the user review data we have 64295 rows and 5 columns of data. We have to take the maximum outcomes from the data which help us to analysis the which type of app is most preferable and comparisons between different insights. Our goal is to filter and make plots accordingly for a better EDA with respect to the final data. We need to explore and analyse the data to discover key factors responsible for app engagement and success.

# 3. Overview of Google Play Store and User Review Datasets.

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 a 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.

## 3.1. 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, I will examine various qualities like rating, free or paid and so forth utilizing Hive and after that I will likewise do forecast of various traits like client surveys, rating etc.

**The data set contains the following attributes:**

* **App:** This Column contains the name of the app which clearly defines its purpose.
* **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.

## 3.2. User Review Dataset

The user review dataset contains the five attributes, this dataset mainly deals with the user reviews of a apps and score them. We define these attribute 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.

# 4. Framework for Analysis

While doing EDA it’s good idea to follow certain process or framework that keep the analysis process organized. We divide our analysis process into six phases, by complecting each phase one by one we are getting close to our insides and conclusions. These six phases are;

* **Ask**
* **Prepare**
* **Process**
* **Analyze**
* **Share**
* **Act**

Let’s describe them in bit details.

1. **Ask**: Include the problem statement for analysis
2. **Prepare**: preparing the data for exploration.
3. **Process**: In this part we are going to clean the data, so that it can be ready for analysis.
4. **Analyze**: answering the question by analyzing the data using various tools.
5. **Share**: The insides we are gather form above process are shared among the stakeholders.
6. **Act**; Put the data into the context and apply its finding for the data-driven decision making.

In the above framework we used certain tool for the analysis get the insides form the data. These tools and their respective version are defined below.

* **Python – Version 3.9 and above**
* **NumPy – Version 1.21.5 and above**
* **Pandas – Version 1.4.2 and above**
* **Matplotlib – Version 3.5.1 and above**
* **Seaborn – Version 0.11.2 and above**

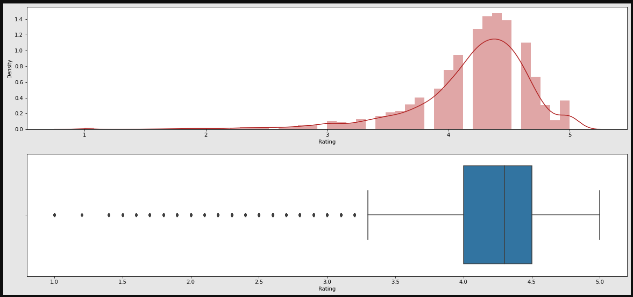
The above tool and their corresponding version are recommended It’s because the fact that python does not support backward compatibility.

# 5. 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.

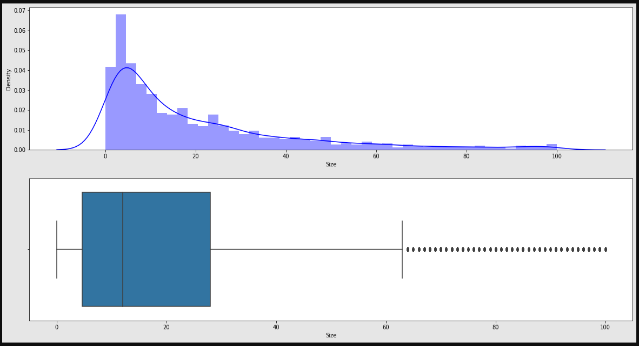
The available data is raw and unusable for Exploratory data analysis, so before we do anything with the data we will have to explore and clean it to prepare it for data analysis.

* **Step1**: We write a function metainfo(), 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 column 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 panda’s library to fill this value.
* **Step 3**: We drop the columns ‘Current Ver’ and ‘Android Ver’ from our dataset using the drop() function of the panda’s 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 call the metainfo() function, 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.

# 6. 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.

### 6.1 Top category with most apps & No of Installs per category.

There are all total 33 categories in the dataset from the analysis 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.

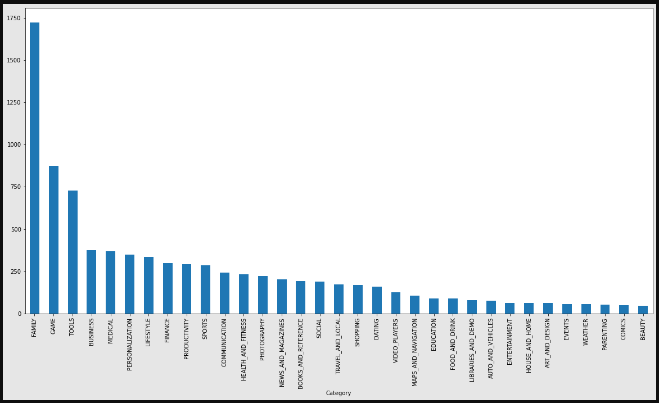
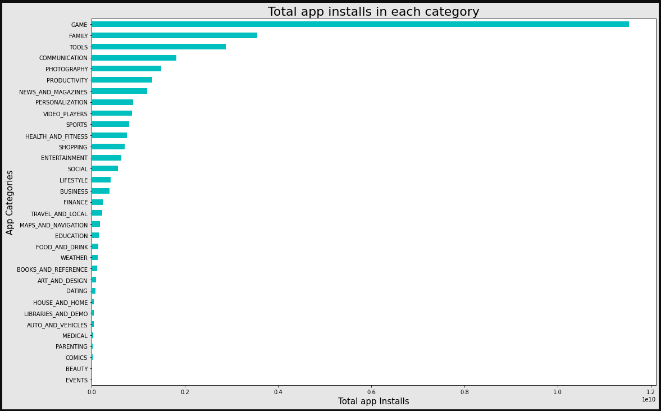


Fig -1: No. of apps per category.

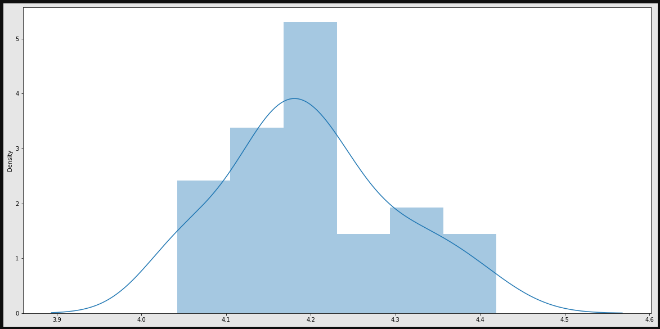
Categories of apps that has the maximum number of installs. The Game, Communication and Tools categories has the highest number of installs compared to other categories of apps.

Fig -2: Total apps installed in each category

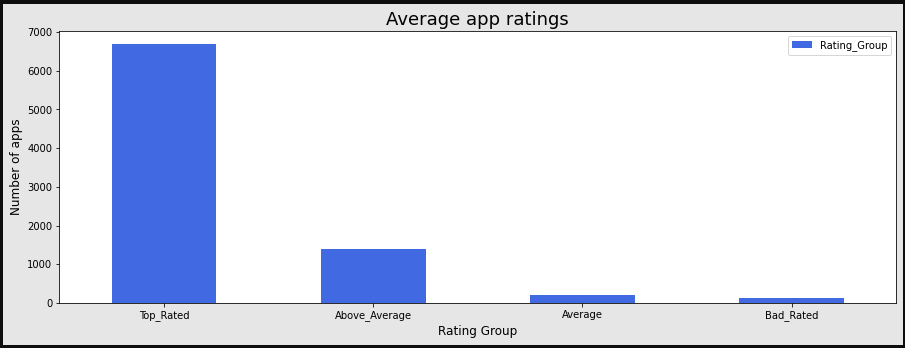
The category which having the maximum number application present is FAMILY. Whereas the GAME category having the highest number of installed application.

### 6.2 Top category based on rating, distribution of rating and average app ratings.

In the below plot, we plotted the apps Rating

Fig -3: Distribution of apps 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.

Fig -4: Average app rating

The rating available in the dataset is distributed so we can represent the ratings in a better way if we group the ratings between certain intervals. Here, we can group the rating as follows:

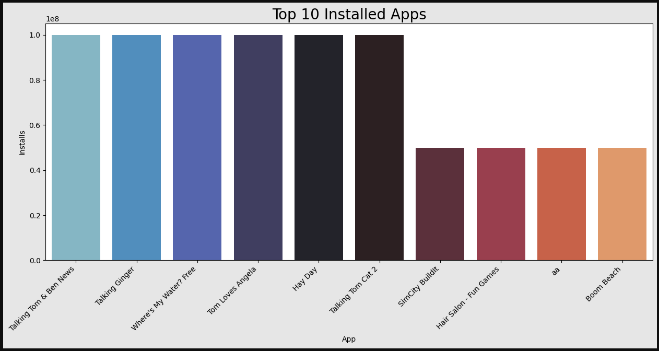
* 4-5: Top rated
* 3-4: Above average
* 2-3: Average
* 1-2: Below rated

### 6.3 Top app in top category and Top ten apps in any category.

We know that the highest number of applications are present in the "FAMILY" category. let's fetch the maximum number of installs for each app in this category.

Before we move forward it's important to understand what the term **User Engagement** means. So, for our case we define user engagement **as the number of installs of the application.**

Let’s fetch the top 10 apps form any category based on installation.

Fig -5: Top 10 apps in FAMILY category

1. The application which are highest in terms of user engagement are "Free" and having the content rating to everyone means no age limit.
2. one the most important factor which highlighted here is that the **gamification of the application draws more user compared to rest ones.**
3. Second, the more restricted the content rating of the application the more niche the customer base of application is

.

### 6.4 Compare max and min app installs and query all its attribute

These apps are;

Max installs:

1. Subway Surfers
2. Google News

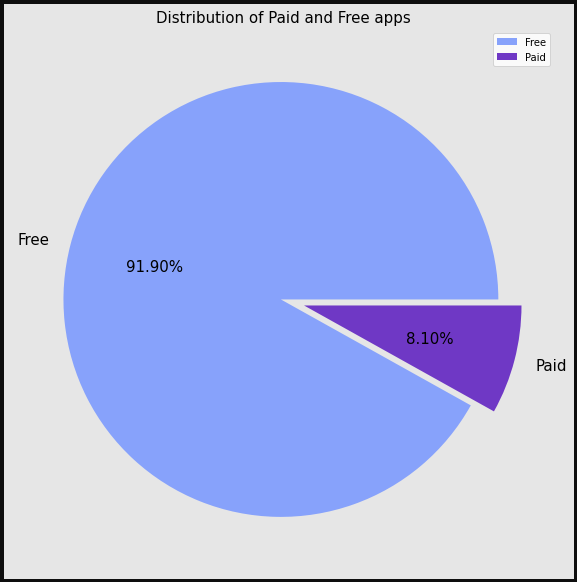
Min installs:

1. Popsicle Launcher for Android P 9.0 launcher
2. Ak Parti Yardım Toplama

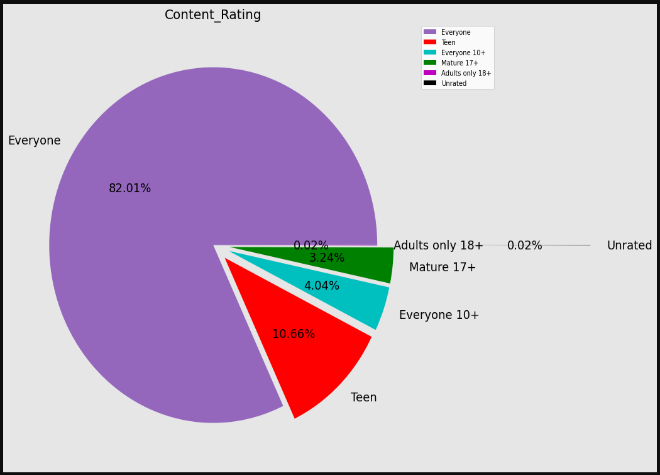
Highest number of installs belong to free app as expected. Among the lowest number of installs which is zero, most of app belong to paid version of app. We also seen that the app installs zero number of times, having a rating value in rating attribute, who this will be a case?

This is because, app may be downloaded from 3rd party source and rated in google play store, we don't roll-out is possibility.

### 6.5 Distribution of Free vs Paid & Content Rating

Fig -6: Paid vs Free application

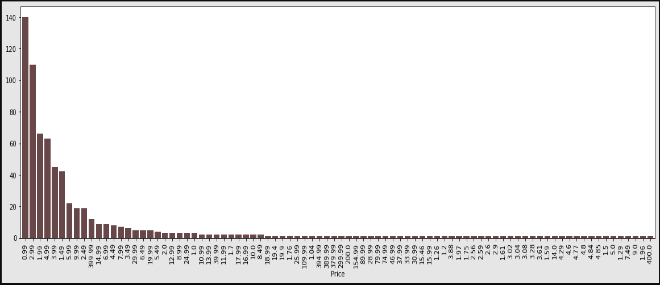
Around 92% apps where free and the 8% apps where paid.

Fig -7: Content Rating Distribution of apps

Around 82% apps where rating of "Everyone". which accounts highest in terms of percentage.

### 6.6 Exploring the Paid apps Attribute

As some of the apps in the play store are paid so we have to analyze this attribute deeply.

Fig -8: Price vs No. of apps

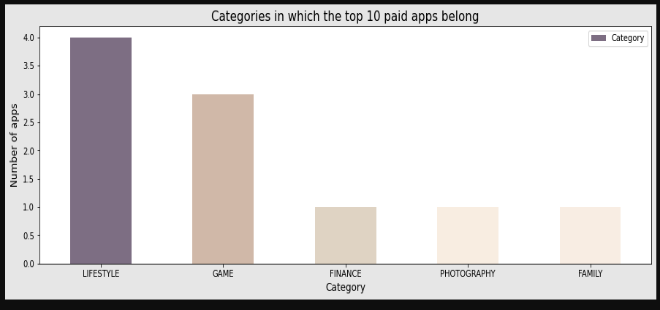
1. The paid apps charge the users a certain amount to avail their full services through app. This amount varies from one app to another.
2. There are a lot of apps that charge a small amount whereas some apps charge a larger amount. In this case the price to download an app varies from 0.99 to 400 USD.
3. In order to select the top paid apps, it won't be fair to look just into the number of installs. This is because for most of the application the installation is free.
4. Here a better way to determine the top apps in the paid category is by finding the revenue it generated through app installs Here we also assume that number of installations = paid app user.

This is given by:

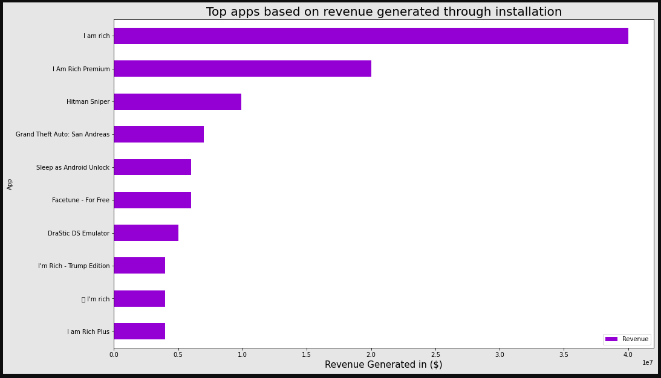
Revenue generated through installs = (Number of installs) x (Price to install the app)

* **I am rich**
* **I Am Rich Premium**
* **Hitman Sniper**
* **Grand Theft Auto: San Andreas**
* **Facetune - For Free**
* **Sleep as Android Unlock**
* **DraStic DS Emulator**
* **I'm Rich - Trump Edition**
* **💎 I'm rich**
* **I am Rich Plus**

Let’s find out the categories where top 10 apps belongs

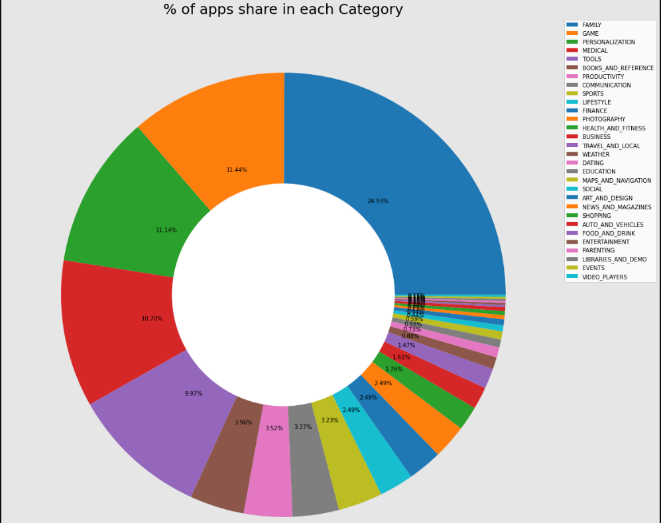
Fig -9: Top 10 Paid apps categories

Let’s Find out the apps which generated the maximum revenue

Fig -10: Top app with highest revenue

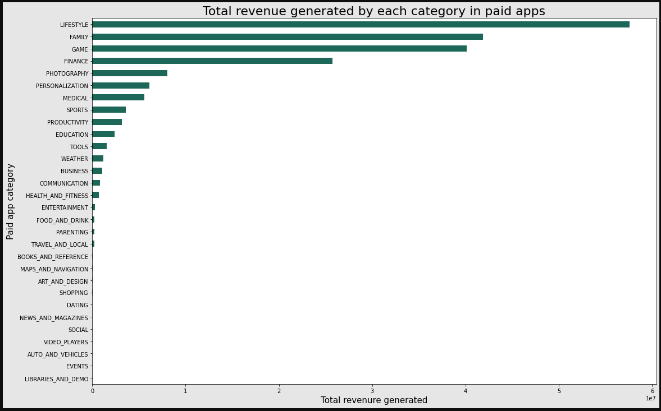
1. LIFESTYLE, GAME, FINANCE are the top three categories which top apps belong too.
2. I AM RICH apps having the five different versions for paid app

### 6.7 Finding out the category wise Paid app distribution and revenue generated by each category.

Fig -11: Paid app % in each category

1. FAMILY and GAME having the greatest number of paid applications
2. Lest are present in EVENTS and VIDEO PLAYERS.

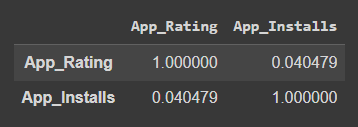
Let’s find out the revenue generated by each category.

Fig -12: Revenue by each category

The "LIFESTYLE" category dominated with respect to revenue collected which is by the way only have 2.5% of the total applications. this is may be because the case that the application under it charges the more in terms of money.

### 6.8 Application rating vs user engagement

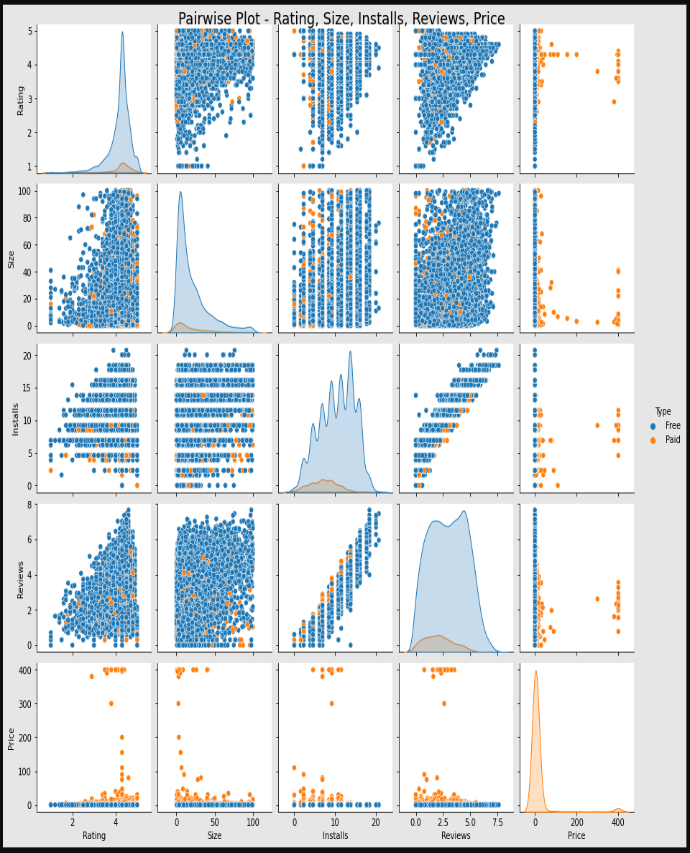
For this problem we have to find the Correlation Coefficient. As we know, the correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no linear relationship between the movement of the two variables.

Fig -13: Correlation coefficient

The linear correlation coefficient of approximately 0.04 **suggests that there is no appreciable linear correlation.**

### 6.9 Bivariate relationship between play store attribute

Univariate & Bivariate Analysis Pair plot is used to understand the best set of features to explain a relationship between two variables or to form the most separated clusters. It also helps to form some simple classification models by drawing some simple lines or make linear separation in our data-set.  
Plot a pairwise plot between all the quantitative variables to look for any evident patterns or relationships between the features

Fig -14: Bivariate relationship

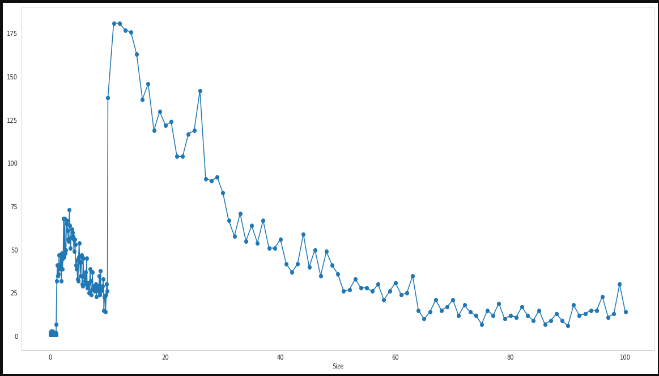
1. As we know there is no direct linear-relationship between install and rating but, we can say that as number of installs increase their increase in the rating of the application.
2. Size of the application can affect the rating as less the size is more the rating of the application.
3. Most of the apps are light-weight.
4. We also see that the greater the user engagement the greater the rating. so, there is direct relationship between them.

### 6.10 Exploring the relationship between size and installs attribute

Now, we have to deeply analyze the size and install attribute.

Fig -15: size and rating wrt category

Let’s explore the impact of size on installs.

Fig -16: size vs no of installs

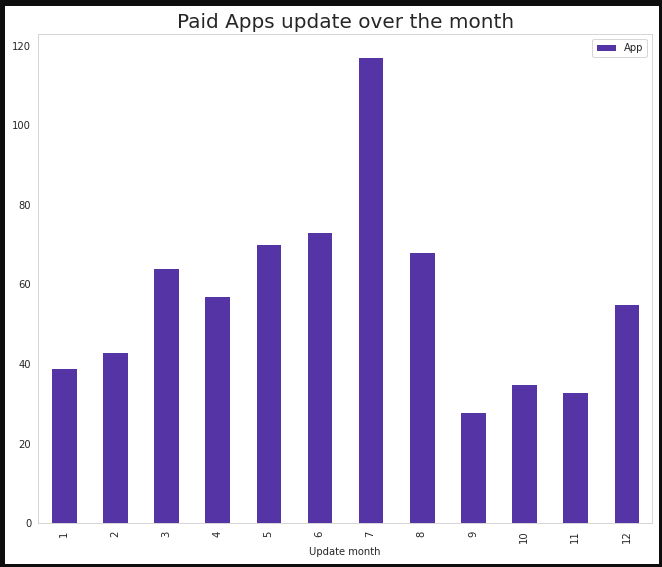
Upon observing the pair plots, following categories of application show negative trend. Which means that the rating of the application will decreases as the size increases.

1. NEWS\_AND\_MAGAZINES
2. TOOLS
3. TRAVEL\_AND\_LOCAL
4. SHOPPING
5. LIFESTYLE
6. HOUSE\_AND\_HOME
7. FOOD\_AND\_DRINK

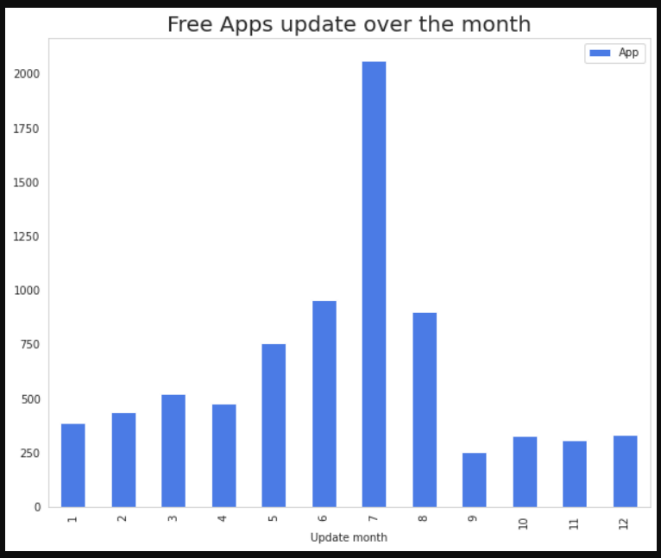
The FAMILY category which holds maximum number of applications tends to have a neutral with respect to size.  
Interesting fact about a COMIC category is that as size increase the rating of the application increase, but consider the fact that all the application in this category is less then 50MB size. The inside we can drown form size vs installs comparison is that as size increase the number of installs decreases. and the best ideal size for application is under 20MB.

### 6.11 Distribution of Free apps vs Paid apps updates over the Timeframe

Application update has a significant impact on the apps so let’s explore this.

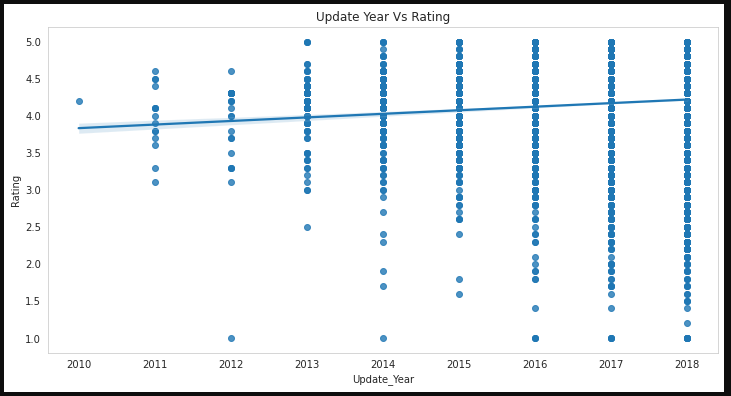
Fig -17: Paid app update over month

Let’s do it for the Free apps also,

Fig -18: Free app update over month

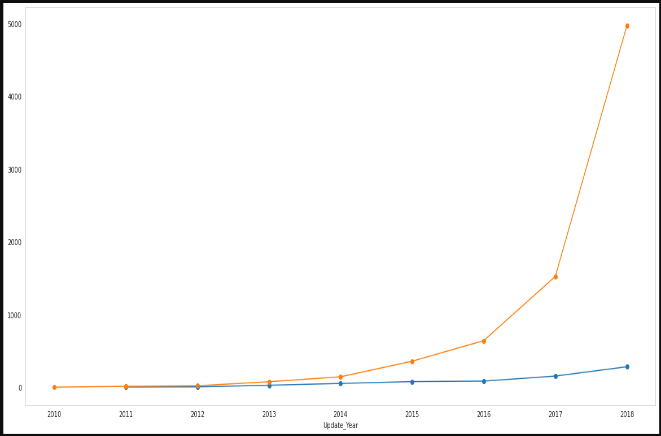
1. Both the categories of apps got maximum number of updates in the months of July.
2. In this data almost 50% apps are added or updated on the month of July, 25% of apps are updated or added on the month of August and rest of 25% remaining months.

### 6.12 Analysis of app update over rating in the span of one year

Fig -19: App update vs rating

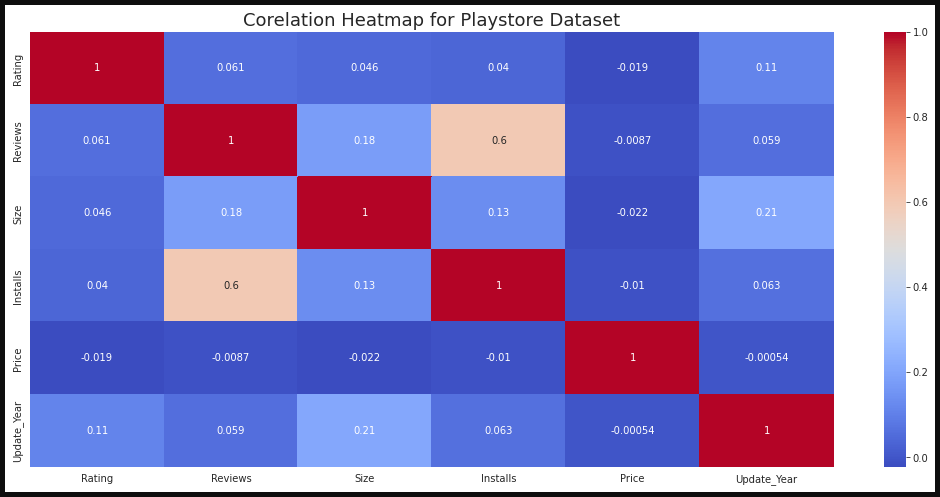
As we seen in the regplot that as apps frequently got update, so as the user reviews will also increases.

Let’s explore this in the context of the free vs paid.

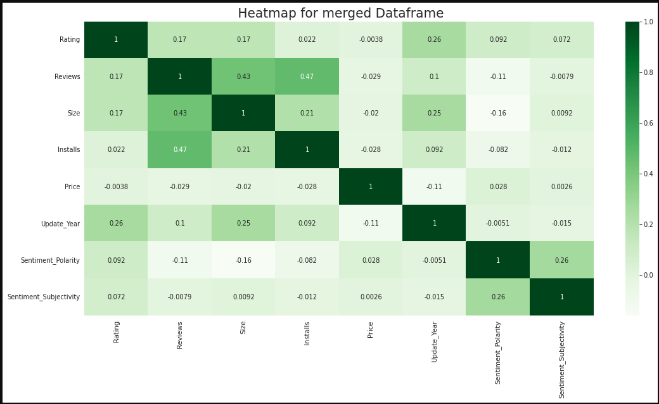
Fig -20: Free vs Paid app update over year

In the above plot, we plotted the apps updated or added over the years comparing Free vs. Paid, by observing this plot we can conclude that before 2011 there were no paid apps, but with the years passing free apps has been added more in comparison to paid apps, by comparing the apps updated or added in the year 2011 and 2018 free apps are increases from 80% to 96% and paid apps are goes from 20% to 4%. So, we can conclude that most of the people are after free apps.

### 6.13 Correlation Heat maps for play store and user review dataset

Fig -21: Correlation between attributes

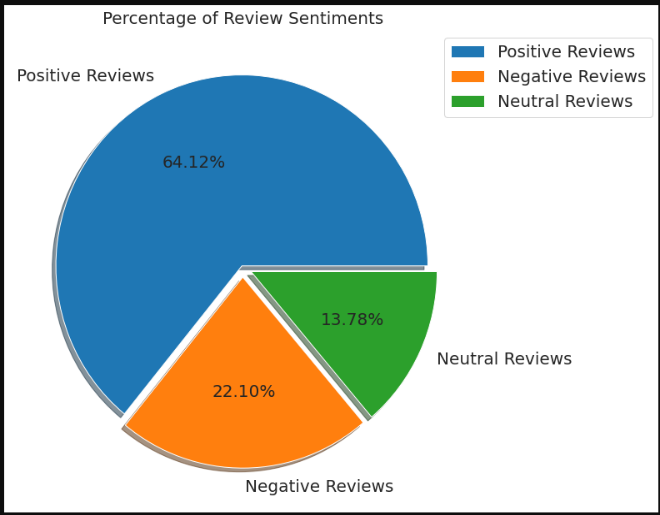
* There is a strong positive correlation between the Reviews and Installs column. This is pretty much obvious. Higher the number of installs, higher is the user base, and higher are the total number of reviews dropped by the users.
* The Price is slightly negatively correlated with the Rating, Reviews, and Installs. This means that as the prices of the app increases, the average rating, total number of reviews and installs fall slightly.
* The Rating is slightly positively correlated with the Installs and Reviews column. This indicates that as the average user rating increases, the app installs, and number of reviews also increase.

Fig -22: Merged data frame heat map

### 6.14 Sentiment Analysis of user review dataset

**Sentiment analysis (or opinion mining)** is a natural language processing (NLP) technique used to determine whether data is positive, negative or neutral. The Subjectivity detection and polarity detection are sub-tasks under sentiment analysis. Subjectivity detection aims to remove 'factual' or 'neutral' content, i.e., objective text that does not contain any opinion. Polarity detection aims to differentiate the opinion into 'positive' and 'negative'. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs. For this analysis we have given the extracted data in terms of numerical values, we just have to analyze these values. and draw the conclusion.

#### 1. Analyze which sentiment dominates the most while reviewing the application

Fig -23: Sentiment Percentage

1. Among all the reviews the positive reviews dominate the most with share of **64.12%**.
2. Positive reviews followed by negative reviews with share of **22.10%**.
3. The lest number of reviews are the neutral reviews with share of **13.78%**.

#### 2. Top 10 Apps with positive reviews

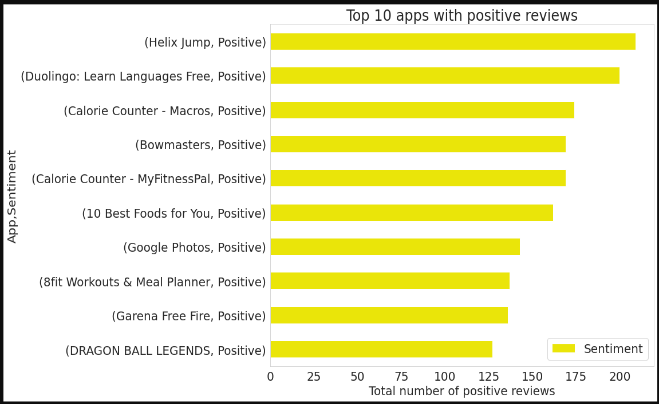


Fig -24: Top 10 positively reviewed apps

For above chart following are the application that has most positive reviews

1. Helix Jump
2. Duolingo
3. Calorie Counter
4. Bowmaster
5. Calorie Counter

#### 3. Top 10 Apps with negative reviews

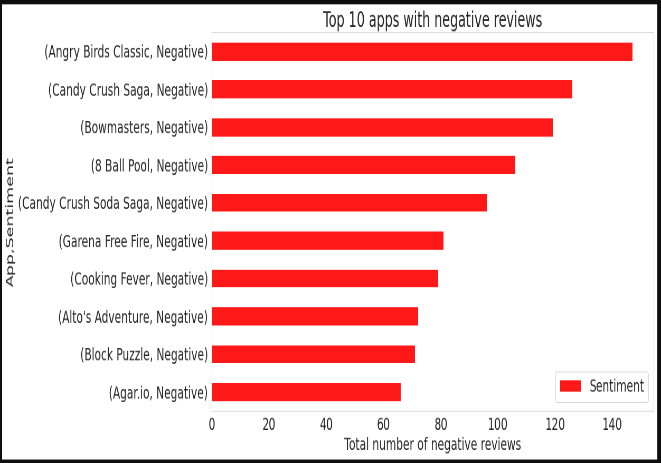


Fig -25: Top 10 negatively reviewed apps

For above chart following are the application that has most negative reviews

1. Angry Birds
2. Candy Crush
3. Bowmasters
4. 8 Ball Pool
5. Candy Crush Soda Saga

#### 4. Fact based or Opinion based review analysis

**For this question we have taken the "Sentiment\_Subjectivity" attribute into the account. As this help us to find out whether reviews are fact or opinion.**

As we all know that the Sentiment\_Subjectivity score are between 0.0 to 1.0 (float64 dtype) in which,  
**0.0 means objective (fact based)**  
&  
**1.0 means subjective (opinion based)**

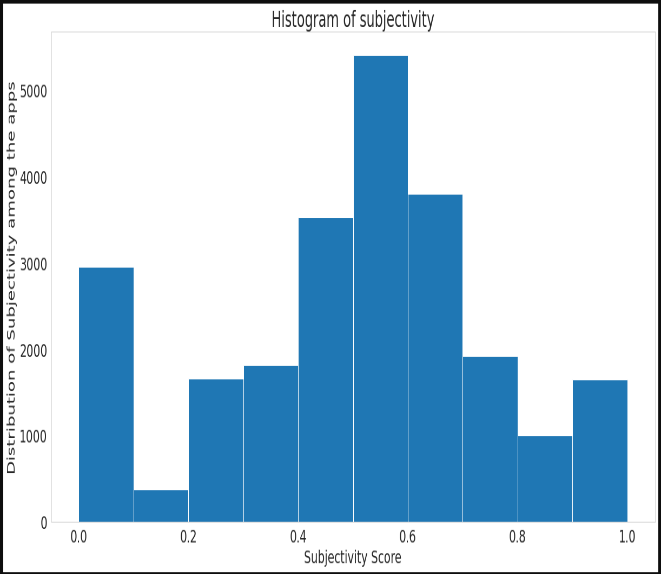


Fig -26: Histogram of Subjectivity

From the above visualization we seen that the highest number of sentiment subjectivity is in the range of 0.4 to 0.6. Which means that the maximum number of reviews from the users are form their own experience.

#### 5. Is Sentiment Subjectivity is proportional to Sentiment Polarity?



Fig -27: Polarity vs Subjectivity

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, shows a proportional behavior, when variance is too high or low.

# 7. Summery of Insides per attribute

1. **The App Attribute**

While analysing we see that the name of the application truly defines the app and what their purpose is for majority of successful applications. And clearly distinguish them from rest of the apps.  
For example, the app name **File commander** is the most installed app and we clearly state that is a **Office Utility** app, same goes for the **Subway Surfers** which belong to **Entertainment and Games.**

1. **The Category Attribute**

In the play store we have a category’s which divides all the application into their function and usability. For example, you can't find the **Gaming** application in **Business** category. We also find out that the **FAMILY** and **GAMES** category has maximum number of applications were launched or available.

1. **The Rating Attribute**

One of the top most matrix to determine the app performance. The most of the apps have a **rating above 4 stars.** The **average rating** for each application **is 4.3**  
We also see that the linear correlation coefficient between rating and review is approximately 0.04 which suggests that there is **no appreciable linear correlation.**

1. **The Reviews Attribute**

While analysis we seen that as **reviews of the application are increases so as the rating of the application also impacted**.  
It is one of the most important matrices for the improvement of the application. if there is, and it's always.

1. **The Size Attribute**

The **size** of most of the application **is small and below 20MB.**  
We also find out that the **size has not significant proportional relationship with rating.** But, in some cases we see that size do matter.

1. **The Installs Attribute**

This matrix is useful for measuring the user engagement of the application.  
Most of the user engagement is done in **FAMILY, GAME, and TOOLS** categories.  
We also seen that the max app are available in the **FAMILY** category but, of we look at the max installed apps in the category it comes out to be **GAME**.

1. **The Type Attribute**

It determines whether the application is **Free or Paid**  
While analysing we seen that the most of the application are the of type **Free** which account to 92% of the total apps.  
Most number of **Paid** apps are available under the **FAMILY** which is obvious as this very category has a maximum number of applications. The app **I'm Rich** is the most expensive app which cost around 400 USD. for it's installation and use.

1. **The Price Attribute**

This attribute defines the price per application in USD.  
If the app is **Free** then **0** otherwise cost of the application. max price of the application is around 400 USD.

1. **The Content\_Rating Attribute**

Direct the recommended age to use the application. The rating **Everyone** has highest number of user engagement. while **Adults**, **Only 18+** and **Unrated** has lowest number of user engagement.

1. **The Last\_Updated Attribute**

The attribute gives the latest update got by the app. we find out that the most number of apps got their update in the month of **July**.  
We also explored that, there where no **Paid** apps **before 2011** and **most of the people are after the Free apps**

**Let's summarize the finding from the User\_Review dataset.**

**The Sentiment Attribute**

This attribute structured whole argument into the three segments as **Positive**, **Negative** and **Neutral**. We find out that maximum number of reviews are **Positive** followed by **Negative**, and the lest among these are **Neutral** ones.

**The Sentiment\_Polarity & Sentiment\_Subjectivity Attributes**

Sentiment subjectivity is not always proportional to sentiment polarity but in maximum number of cases, shows a proportional behaviour, when variance is too high or low.  
Sentiment Polarity is not highly correlated with Sentiment Subjectivity.  
The polarity scale show that the sentiments which are extremes are the low in numbers, most of the reviews are lies between the mid-range of positive and negative.

# 8. Conclusion

As of now we know our findings and we have information in our hand. So, it's time for us to put this valuable information in the **context** and gain some knowledge.  
For this project we have a **context of developer** rather than an application which to be launched. so, form our analysis we should conform these following factors in order to launch the successful application.

1. The app's name should accurately describe its value propositions. because majority of successful application have this quality.
2. Launching the apps in the category which having more and easy user reach such as **"FAMILY"**, **"GAME"**etc.
3. As per our analysis most of the apps are free it's around **92%**, so if possible, try to **launch the app with "Free" type**. as it's **increasing the user engagement.**
4. We also seen that the number of installs is correlated with the rating of the application as number of installs increases so as the application rating.
5. One important inside we get with respect to the size of the app is that as the size of the application increases the installation of the app decrease. So, if we release a new apps in market **make sure it's under 20MB.**
6. Content rating also affect the user engagement as more restricted your content rating is the more restricted your user engagement. So, try to keep user rating as **"Everyone"**.
7. Make sure the app will get the update at regular interval, as it's an important factor for user engagement and performance of the application. In our analysis we seen that most of the apps will get their app update at **July month**.
8. As we seen form subjectivity most of the reviews as the objective, so for successful apps it's more important to keep eye on the user reviews and early resolution of the problems.
9. We also seen our analysis that there is strong relationship between Install and reviews. for the new apps reviews are the important tool for increase the user engagement.
10. For the apps it's more important have the android version compatibility **above version 4**. As we won't explore this attribute much but by using simple sorting, we can confirm this.

# 9. References

* Stackoverflow
* Python documentation
* Panda’s documentation
* Seaborn documentation
* Towards data science
* Statica
* wikipedia
* Google data analytics professional.