****

**Play store App Review Analysis**

Awake Chaudhary,

Data science trainees,

AlmaBetter, Bangalore.

Email: [awakechaudhary@gmail.com](mailto:awakechaudhary@gmail.com)

LinkedIn: <https://www.linkedin.com/in/awake-chaudhary-741006146/>

Github: <https://github.com/Awakechaudhary>

Github repo. : <https://github.com/Awakechaudhary/Play-store-app-review-analysis>

**Abstract:**

Google play store is a digital distribution service operated and developed by Google.

It contains android apps divided among different categories, some of them are free and some of them are paid.

Our analysis on the database can help us to find various key factors which are responsible for play store popularity and help to make it more better platform for android users.

**Keywords: Exploratory Data Analysis (EDA), Reviews, Ratings, Category, Installs.**

1. **Problem Statement**

Data provided by Play store, which is operated and developed by Google. Play store has a million numbers of similar or different kinds of apps which are tagged by different Genres. These apps are available for both free and paid. They are also providing different features for the users to choose and download the best apps from it.

We will be doing Exploratory data analysis on this data set, which is a very important ,as it not only helps in taking very initial business decisions but also in preparing the data for further modeling for use in machine learning algorithms.

They are also providing different features for the users to choose and download the best apps from it.

1. **Introduction**

Play store is an Android Market serves as the official app store for certified devices running on the Android Operating system. Developed and Operated by Google, launched on 6th March, 2012. Approximately 3.48 million apps are in the Play store. Play store apps have their own features such as Ratings, Reviews, Size and more. From the problem statement given, we should analyze the given database and should come up with the key factors that increased the number of users, long term usage etc., the objective of this project is to deliver insights to understand customer demands better and thus help developers to popularize the product.

* **Size**: Size is a factor which is not the same for all the devices, and it fully depends on the developers.
* **Reviews**: Given by the customers on the basis of their experiences which helps in betterment of apps
* **Installs**: Installs gives us the data of total installed users count.
* **Type**: Type defines whether the app is free or paid.
* **Price**: Here price is mentioned for the paid apps, for free apps 0 will be given.
* **App name:** This is the column which contains the name of the app.
* **Category**: The Category is a column by which the apps got separated based on the application and purpose.
* **Rating**: Rating plays a huge role in finding the correct apps. It was manually given by the users.
* **Genres**: Genres are like tags, an app can be coming under more than one Genre, based on the usage.
* **Android Version**: This feature gives us the supported device versions of android.

The project is done by using Python language and its library such as Numpy, Pandas, Matplotlib, seaborn.

1. **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

1. **Exploring the database**

**Two databases:**

1. **Play store:**

* Shape of this database is (10841, 13).
* It contains 13 columns, which are as follows:

'App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type', 'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Version', 'Android Version'.

* Out of these 13 columns, “Rating” column is of numeric data type and it also have some none values.

1. **User review:**

* Shape of this database is (64295, 5)
* It contains 5 columns, which are as follows:

App, Translated\_Review, Sentiment, Sentiment\_polarity, Sentiment\_Subjectivity.

* Out of 5 columns Sentiment\_polarity and Sentiment\_Subjectivity have numeric datatype.

1. **Exploratory Data Analysis(EDA):**

**Step1: Data cleaning:**

Data cleaning on raw data is the very first step, it helps us to do EDA in more effective manner and makes data more efficient.

**Play store database:**

* In Category column ,”1.9” is the invalid category name which need to be dropped.
* I dropped the row which contains “1.9” as category.
* Reviews column datatype=object, which should be of numeric datatype so, I changed its datatype to numeric.
* The data set contains many strings like “M”, “k”, “+”, “$”, “NaN”, etc.

So, I replaced the string by making a new function “string\_replace”.

* This data base also contains a row which show paid app price=0.

So, I dropped that row too from the data base.

* In this database the dtype of column “Last\_updated” which contains date of app updation is not of datetime type.

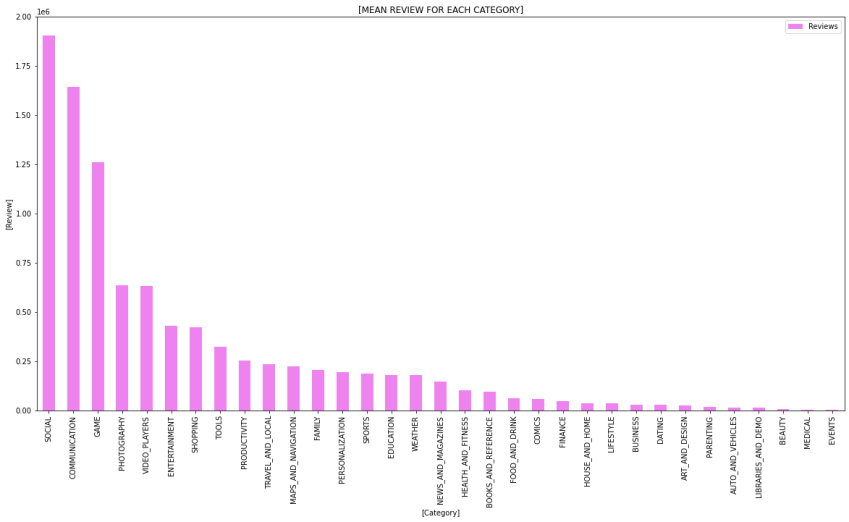
So,I changed its dtype to datetime.

**User Reviews database:**

* In this Data type I found many null values so I make a new database which contains only non-none values.

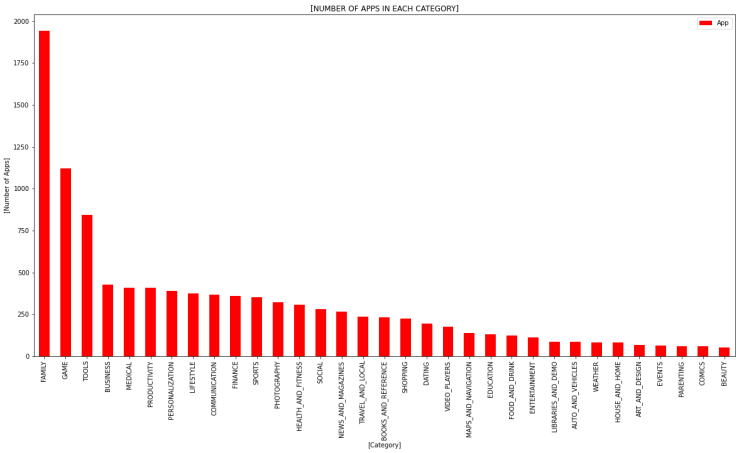
**Step 2 : Draw insights:**

* **Reviews:**
* Reviews is something by which people share their experiences, views with any app on play store by writing about it .
* Here by analyzing we can say that people has given most reviews on apps which belongs to **Social** category on top followed by **Communication ,Games, photography,** etc.
* And least reviews on **Events, Medical & Beauty** category apps.
* If this is the scenario then why is that like this, so many questions arises let’s see number of apps in each category.

****

* **Number of apps in each Category:**
* Here by analyzing we can say that **Family, game, tools, business, medical, productivity** are the some top categories which contains maximum number of apps.

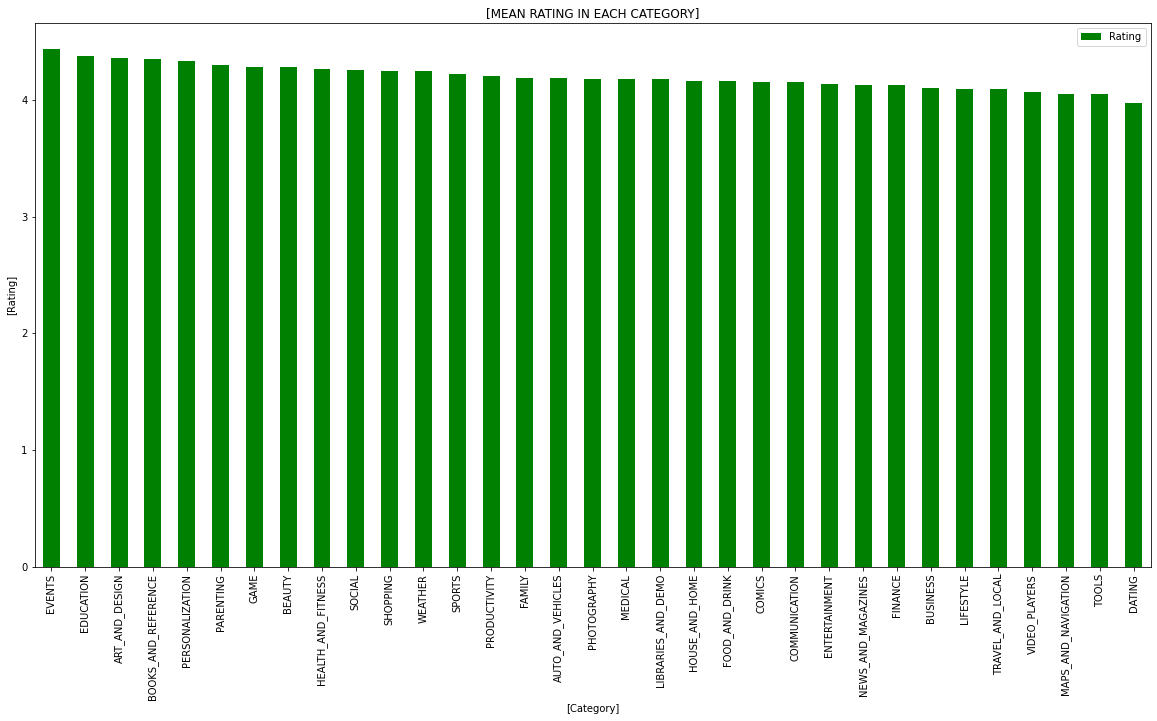
And **events, parenting comics, beauty** are the some categories which contain least number of apps.

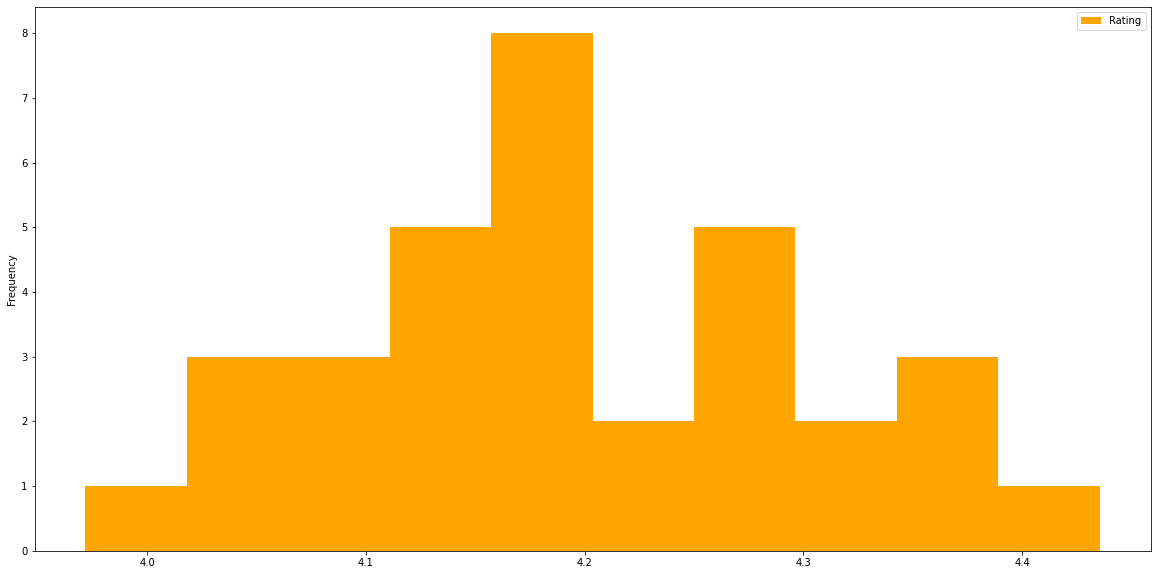
****

* **Rating :**
* Rating is something which shows the people love toward it.

So, rating analysis is very important.

* By analyzing we can say that maximum rating is of Event category followed by education.
* The maximum rating in category is around **4.2** which is considered good.
* Moderate rating lies between 2.5 to 4.
* And below average or poor apps are in the range of 0 to 2.5 which are less in count.

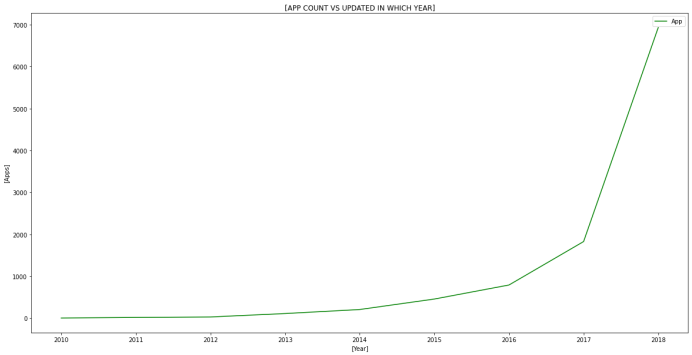
****

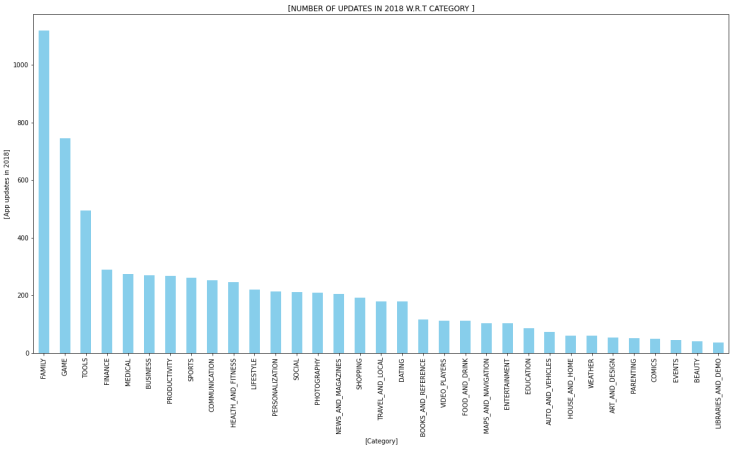
****

* **App Updates:**
* App developer ratify the app performance, reliability, etc. by updating app.
* Usually update add more features in app and makes them more user friendly.

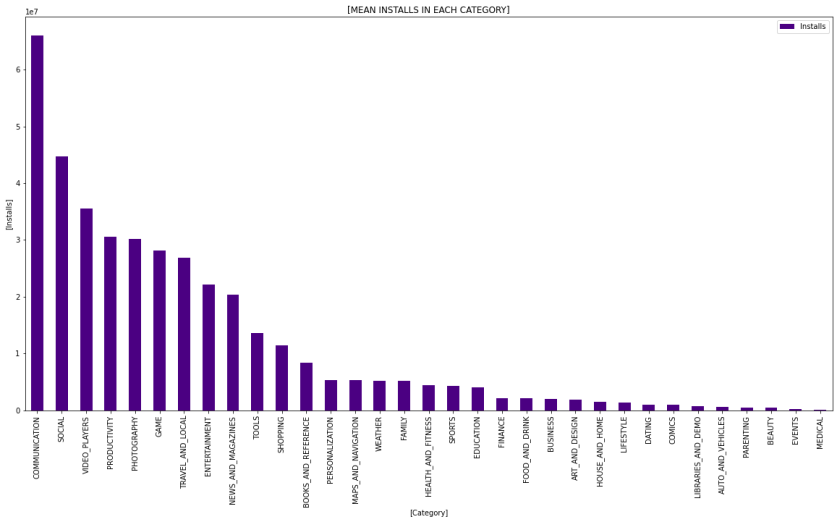
So, apps updates play an important role.

* By analyzing it seems that more apps updates come in year **2018**, which are continuously increasing.
* **Family Apps** Category got the most Updates in year 2018.

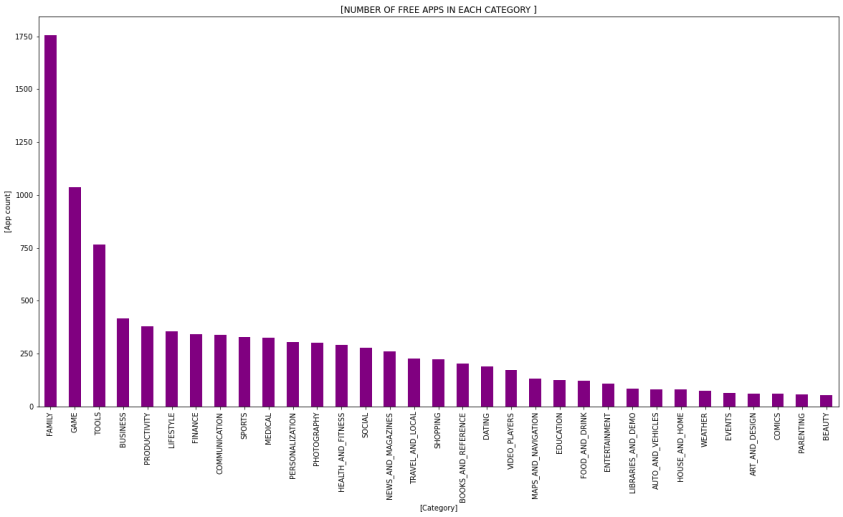
****

****

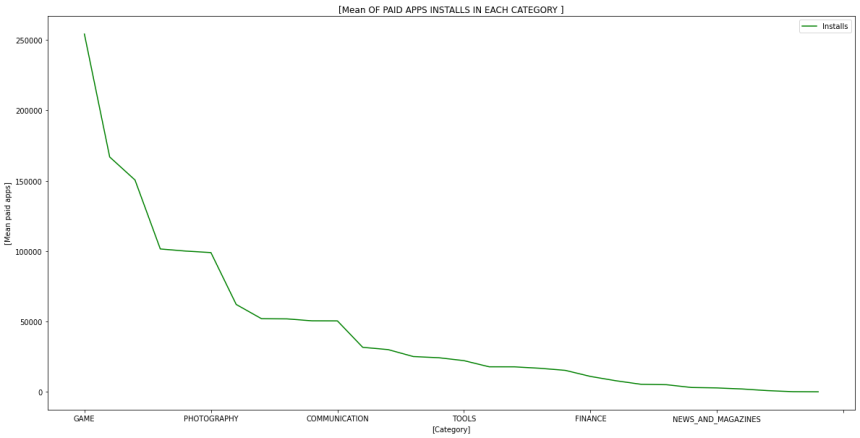
* **Installs:**
* Installs, Review, Rating are related to each other because without installing an app from play store we can’t rate and review that app.
* **Communication** is top category in terms of installation.
* **Medical, Events, Beauty** are among the least categories in terms of installation.

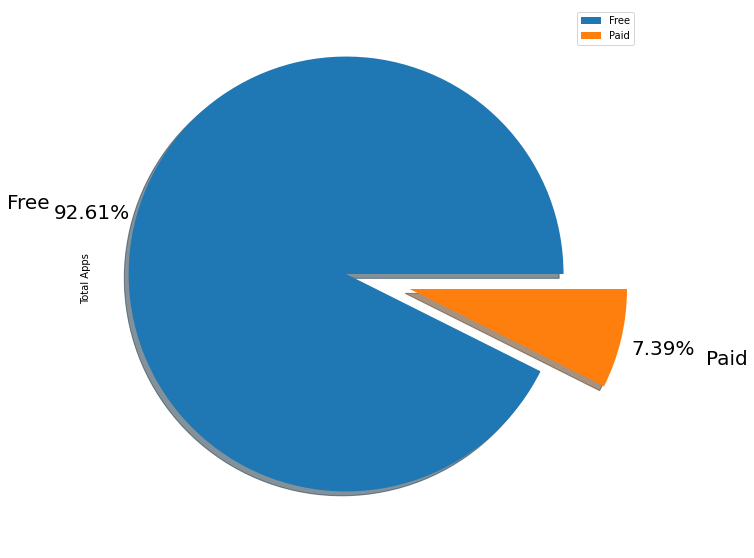


* **Genres:**
* The analyze is done between the top Genres which was having more number of downloads
* As we saw here, the Play store having more number of applications in the genres like **Tools, Entertainment, Education** and etc.
* The developers are mostly focusing on these genres because of the people's daily basis requirements.
* **Free Apps**
* There are **9591** free apps i.e.92.61%
* When comparing, people are showing more interest on free apps like **Family, Game, Tools** Category**.**

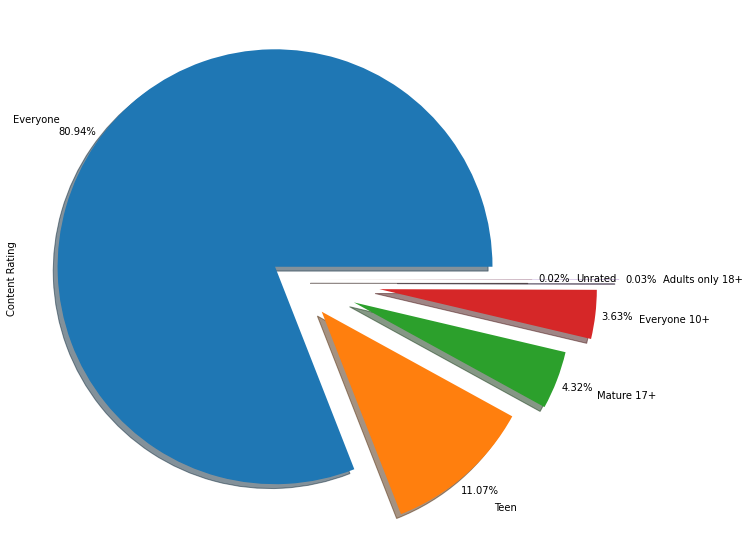
****

* **Paid Apps**
* There are only **765** paid apps i.e**. 7.39%.**
* When it’s coming, commercial people are preferring apps like Games, Photography, and Personalization.
* People preferring less on Educational, Event, and Art & Design.
* Most Paid apps are installs from the **Game** category.

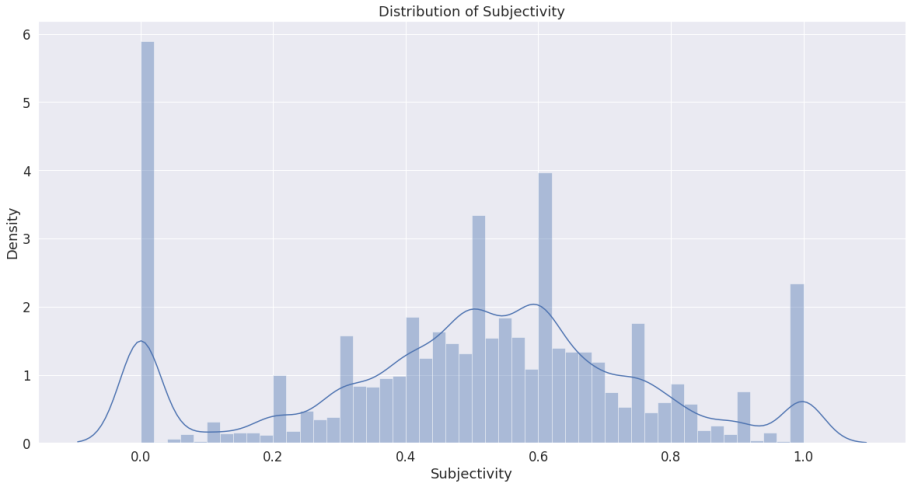




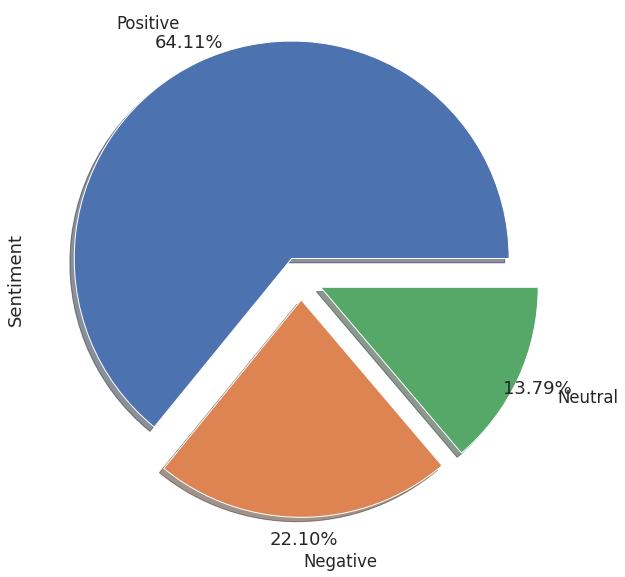
* **Impact of Rating:**
* The content rating shows the results for general contents as high.
* The content rating type **‘Everyone’** has the most percentage value of 80.39%.
* **‘Teen’** contents are second in the order with the percentage of 11.14%.
* Adult’s only and unrated contents are least in this plot, 0.03% and 0.02% respectively.
* So it can be concluded that most of the contents are generic.



* **User sentiment analysis: Distribution of Subjectivity**
* Subjectivity lies mostly between 0.5 and 0.65.
* It shows that the average content and apps reviews subjectivity are mostly relevant.
* Subjectivity of 100% has slightly occurred frequently.
* The nearly 0 subjectivity has a considerable amount of frequency.

****

* **Sentiment percentage**
* The sentiment plot shows the results for positive reviews as high.
* Sentiment type **‘Positive’** has the most percentage value of **64.11%.**
* ‘Negative’ reviews are with the percentage of **22.10%.**
* ‘Neutral’ percentage is **13.79%.**

****

* **Play store data correlation**
* The correlation heat map shows that the **rating, review, size, install** are positively correlated to each other whereas **negatively** correlated with **price**.

****

* **User review data correlation**
* The correlation heat map shows that the sentiment subjectivity and polarity are relatively **less correlated** to each other.
* It’s only 0.26 on scale of 0 to 1

****

**Conclusion:**

Through exploratory data analysis on the basis of some trends many assumptions can be drawn that might lead to app success among the users in the play store , Some of them are**:**

* Percentage of free apps = **~92%**
* People are showing more interest on free apps like **Family, Game, and Tools** Category**.**
* Percentage of Paid apps = **~8%**
* Most Paid apps are installs from the **Game** category.
* People have given most reviews on apps which belong to Social category on top followed by **Communication, Games, photography,** etc.
* **Family, game, tools, business, medical, productivity** is the some top categories which contains maximum number of apps.

And **events, parenting comics, beauty** are the some categories which contain least number of apps.

* The maximum rating in category is around **4.2** which are considered good.
* Most apps updates come in year **2018.**
* **Family Apps** Category got the most Updates in year 2018.
* **Communication** is top category in terms of installation.
* **Medical, Events, Beauty** are among the least categories in terms of installation.
* Play store having more number of applications in the genres like **Tools, Entertainment, Education** and etc.
* The content rating type ‘**Everyone’** has the most percentage value of **80.39%.** ‘**Teen**’ contents are second in the order with the percentage of **11.14%.** Adult’s only and unrated contents are least in this plot, 0.03% and 0.02% respectively. So it can be concluded that most of the contents are generic.
* Subjectivity lies mostly between **0.5 and 0.65**. It shows that the average content and apps reviews subjectivity are mostly relevant. Subjectivity of 100% has slightly occurred frequently.
* The sentiment plot shows the results for positive reviews as high. Sentiment type ‘Positive’ has the most percentage value of 64.11%.‘Negative’ reviews are with the percentage of 22.10%. ‘Neutral’ percentage is 13.79%.

As we know that the Junk data is eliminated, so more assumptions can be drawn from data sets

**References:**

1. Stackoverflow 2. GeeksforGeeks

3. matplotlib.org 4.seaborn.pydata.org

.