**EDA on Play Store App Reviews**

**Praveen Kumar Gutti, Usha Rani Bolimera**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

Google play store is the official app store for all devices operating on the Android OS. It allows the users to browse and download the apps that are developed with the android software development kit (SDK). Apart from offering android applications and games, it also serves as a digital media store offering music, books, movies, and television programs. User ratings and reviews can significantly increase the number of app downloads; hence it is important to analyse the parameters which lead to users giving positive feedback and higher rating. Though this exploratory data analysis, we can understand and discover the key factors responsible for app engagement and success.

***Keywords*: *Correlation heat map, EDA, Outliers, Data Cleaning.***

**Problem Statement:**

Two datasets are provided, one with basic information and the other with user reviews for the respective app. We must examine and evaluate the data in both datasets in order to identify the important characteristics that influence app engagement and success.

## Data Summary:

## We are provided with two datasets:

* **Play\_store\_data**: It contains the basic details of the app like number of user reviews, ratings, etc.
* **User reviews**: It contains the user reviews and its sentiment score for the respective app.

We need to explore and analyse the data to discover key factors responsible for app engagement and success.

**The contents of play\_store\_data are:**

### **App**: It contains the name of the app with a short description (optional).

* **Category**: This section gives the category to which an app belongs. In this dataset, the apps are divided among 33 categories.
* **Size**: The disk space required to install the respective app.
* **Rating**: The average rating given by the users for the respective app. It can be in between 1 and 5.
* **Reviews**: The number of users that have dropped a review for the respective app.
* **Installs**: The approximate number of times the respective app was installed.
* **Type**: It states whether an app is free to use or paid.
* **Price**: It gives the price payable to install the app. For free type apps, the price is zero.
* **Content rating**: It states which age group is suitable to consume the content of the respective app.
* **Genres**: It gives the genre(s) to which the respective app belongs.
* **Last updated**: It gives the day in which the latest update for the respective app was released.
* **Current Ver**: It gives the current version of the respective app.
* **Android Ver**: It gives the android version of the respective app.

### **The contents of User Reviews are:**

* **App**: It 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.

**Data Cleaning:**

* **Android Ver**: The 3 rows containing NaN values were dropped from the dataset.
* **Current Ver**: The 8 rows containing NaN values were dropped from the dataset.
* **Type:** One row containing NaN value was replaced with a relevant entry.
* **Rating:** The 1470 NaN values were imputed with its median value.
* **App**: The duplicate values in the dataset were dropped.

Apart from this the data present in different columns was manipulated to make them easier to analyse. Also, the datatype of the entries was changed in some cases to make the data relevant.

The resultant number of rows post cleaning the data: 9649

**Data Visualization:**

It deals with the graphical representation of data, from which we can draw conclusions and take different business decisions.

**Correlation heatmap**

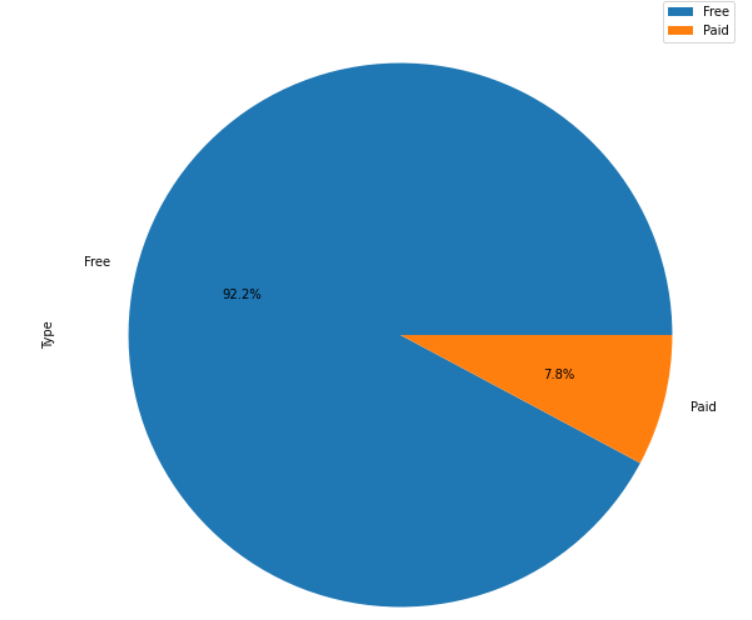
A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables.

* The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution.
* The value of correlation coefficient lies between -1 and 1.
* If the correlation coefficient is above zero (positive), we can say that the two variables are positively correlated. Here, the increase in the value of one of the variables will result in the increase of the second variable.
* If the correlation coefficient is below zero (negative), we can say that the two variables are negatively correlated. Here, the increase in the value of one of the variables will result in the decrease of the second variable.
* If the correlation coefficient is zero, we can say that there is no relation between these two variables. They are independent.
* Plotting the correlation heatmap for the play\_store\_data, we get:



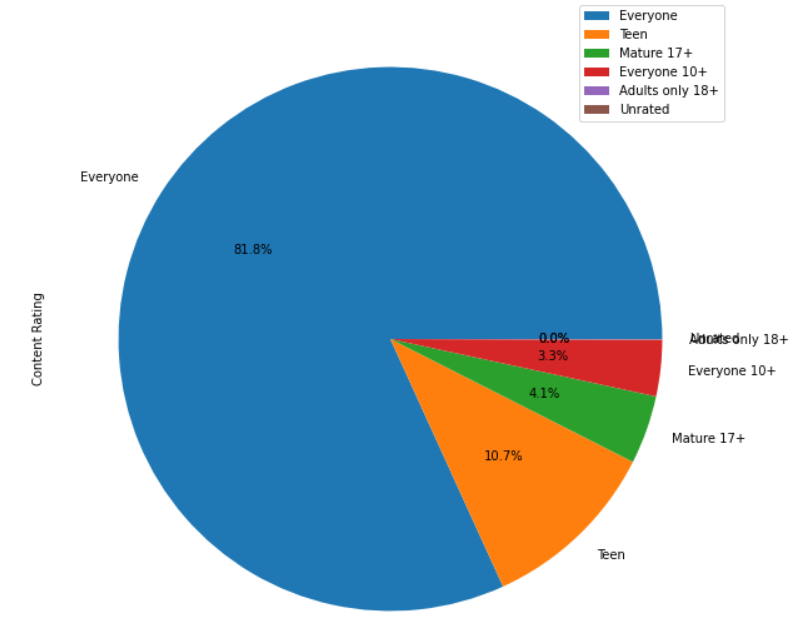
* There is a strong positive correlation between the Reviews and Installs column. This is pretty much obvious. Higher the number of installs, higher is the user base, and higher are the total number of reviews dropped by the users.
* The Price is slightly negatively correlated with the Rating, Reviews, 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.

**Free apps in the database:**



* Approximately 92% of the apps in the play store are free to install.

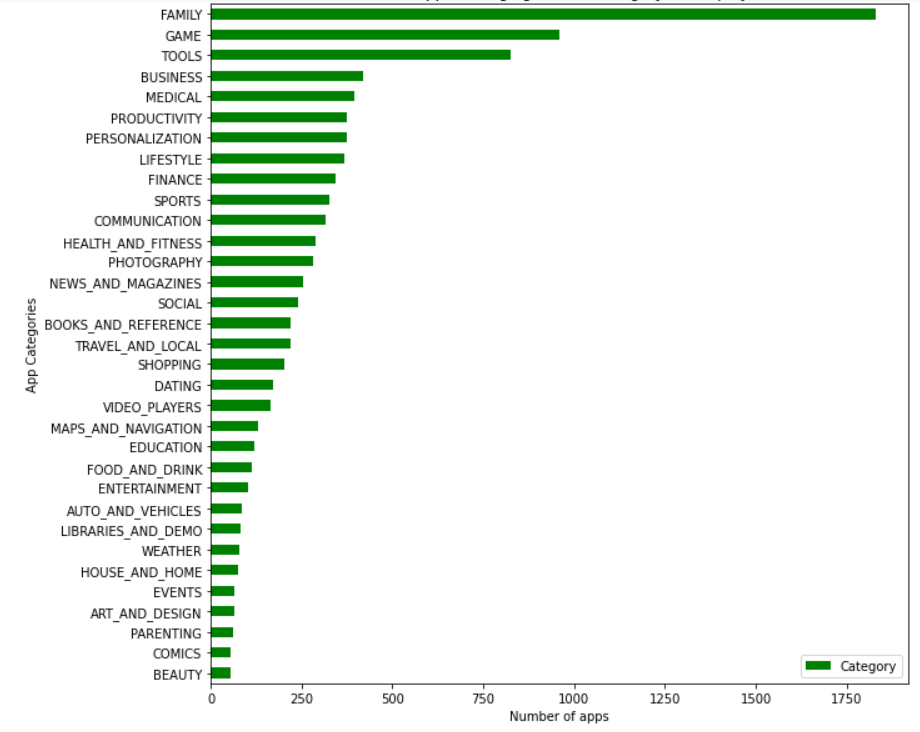
**Content rating:**



* Approximately 82% of the app in the play store have to age restrictions to install and use the app.
* The rest of the apps have certain age restrictions on it.
* Around 11% of the apps are rated as “Teen”, which means that the user must be at least a teenager (13 years old) to install and use the respective app.
* Around 4% of the apps are rated as “Mature 17+”, and around 3% of the apps as “Everyone 10+”

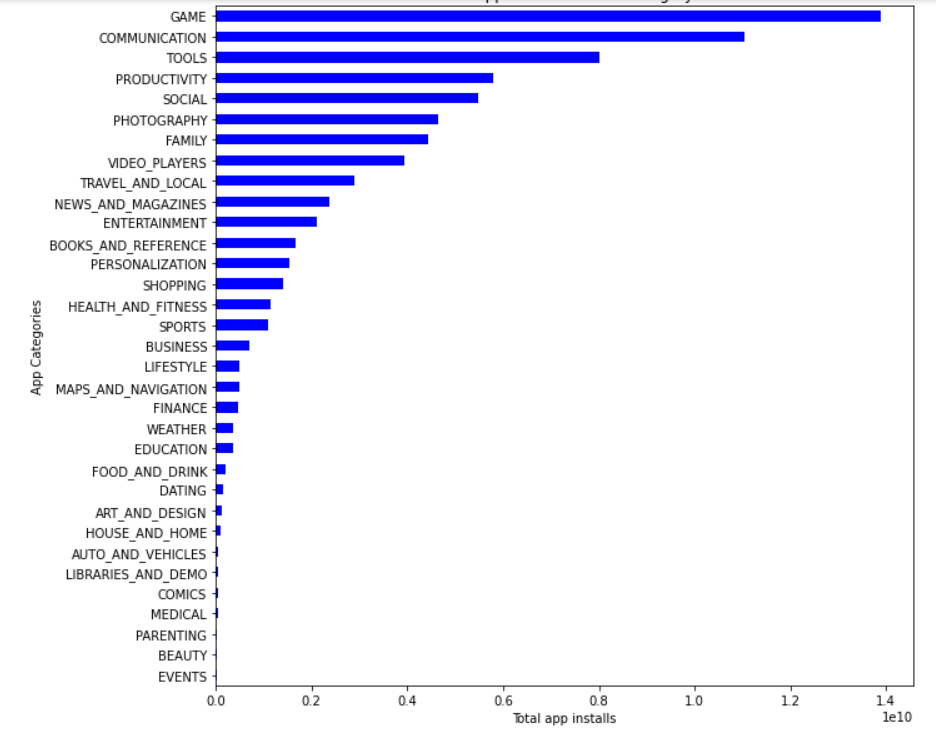
**Number of apps in each category:**

* The apps in the dataset are divided among various categories based on its applications and use-cases.
* In this dataset, the apps are divided into a total of 33 categories.
* The higher the number of apps in a category, the more competitive it is to launch an app in the said category.
* From the bar graph below, we can say that the “Family” category has the highest number of apps, followed by the “Game” and “Tools” category. From this we can say that these categories are the most competitive to get in to.



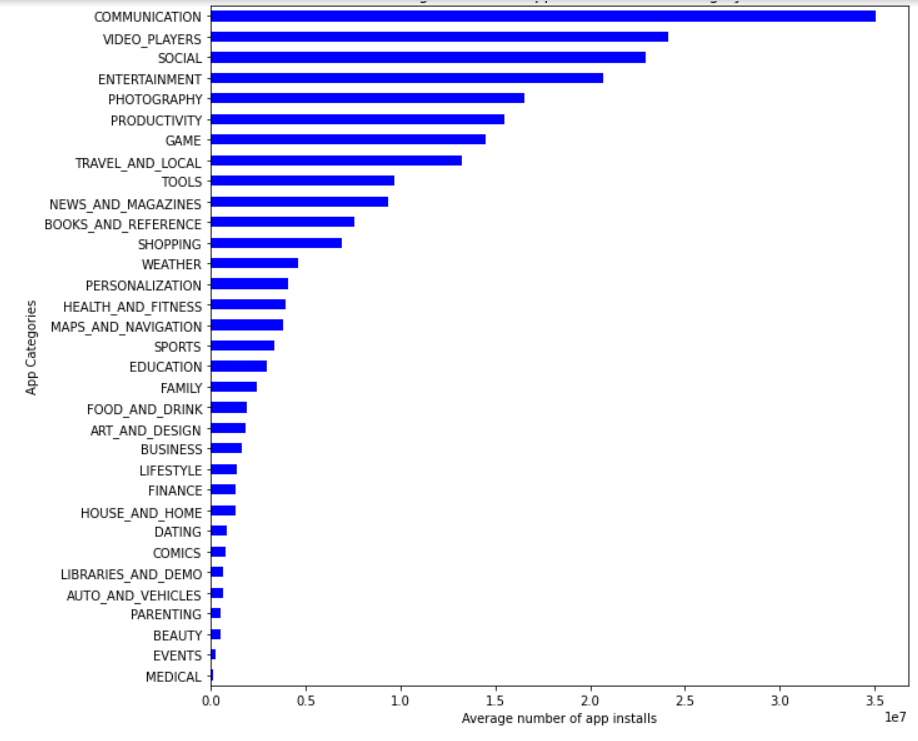
**Total app installs in each category**

* We can say that the total number of installs and reviews for an app shows its popularity among the users.
* The below bar plot gives the distribution of the total app installs in each app category.
* This measure is useful in determining the popularity of apps categorically.
* Hence, the “Game”, “Communication”, and “Tools” are the most popular categories compared to the rest.



**Average app installs in each category**

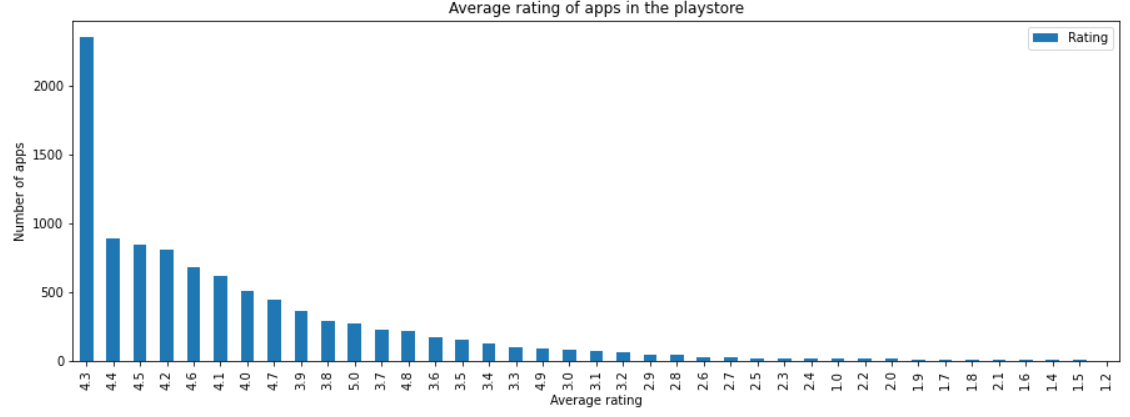
* The below plot gives the average number of app installs categorically.
* From this representation, we can get a fair idea on how popular an average app is in each category.
* Hence, we can say that an average app in the “Communication”, “Video Players”, and “Social” category are more popular than an average app in other categories.



**Average app rating**

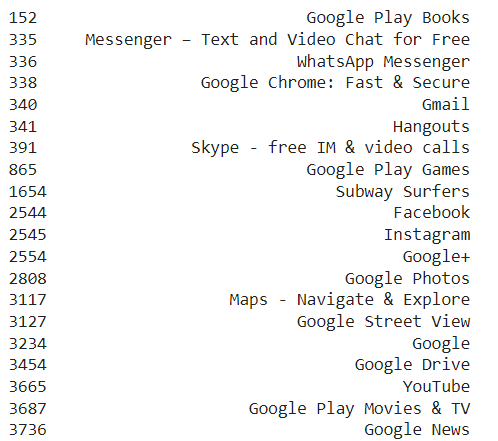
* The median user rating for all the apps in the play store is 4.3.

This metric can be used by a developer to find and study the apps which are not popular among the users and see what mistakes they are doing. Also, this metric can be used to find and study the apps which are popular among the users and implement some strategies in their app.



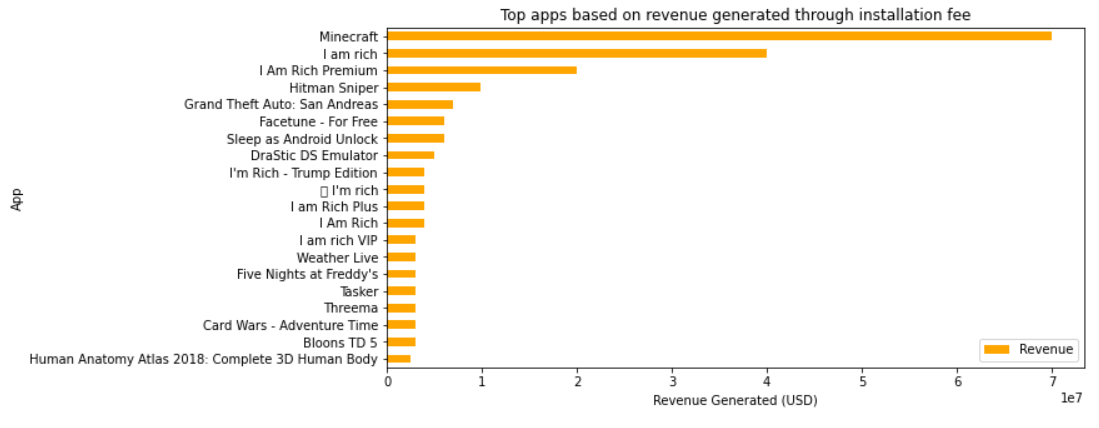
**Top apps that are of free type**

* There is a total of 20 free apps with over one billion installs.
* The top categories in which these apps fall are Communication (6), Social (3), Video Players (2), Travel and Local (2)



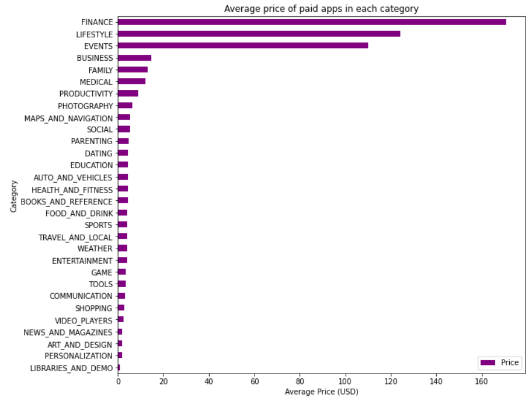
**Top apps that are of paid type:**

* The paid apps charge the users a certain amount to download and install the app. This amount varies from one app to another.
* 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 USD 0.99 to USD 400.
* In order to select the top paid apps, it won't be fair to look just into the number of installs. This is because the apps that charge a lower installation fee will be installed by a greater number of people in general.
* Here a better way to determine the top apps in the paid category is by finding the revenue it generated through app installs.
* This is given by: Revenue generated through installs = (Number of installs) x (Price to install the app)



* The above bar plot gives the top paid apps considering the revenue generated through the installation fee.
* Minecraft is the most installed paid app with over 10 M app installs.
* The top categories in which these apps fall are Lifestyle (5), Family (5), and Game (4).

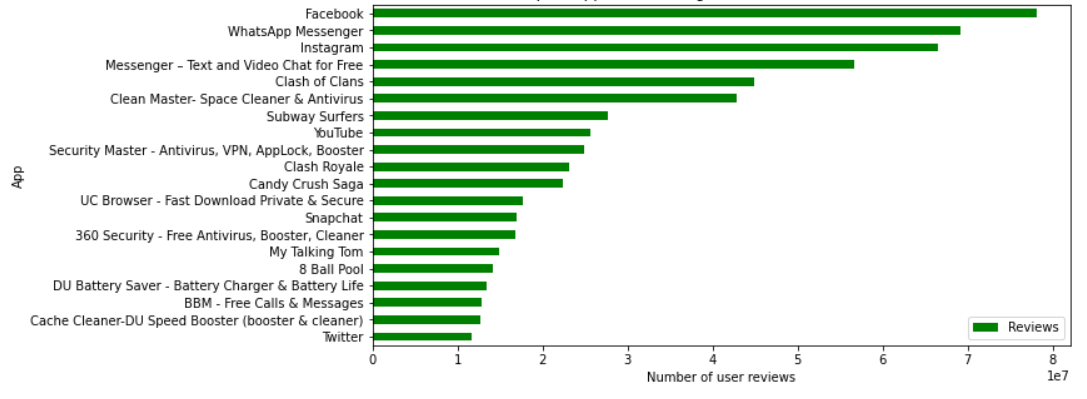
**Average price of paid apps in each category:**



* This visualization gives a fair idea on the amount of installation fee charged by the paid apps categorically.
* The paid apps in the “Finance”, “Lifestyle”, and “Events” category are on average significantly more expensive than the paid apps in other categories.
* If the developer wants to develop an app in these categories, s/he has the liberty to charge a higher fee compared to that of other categories.

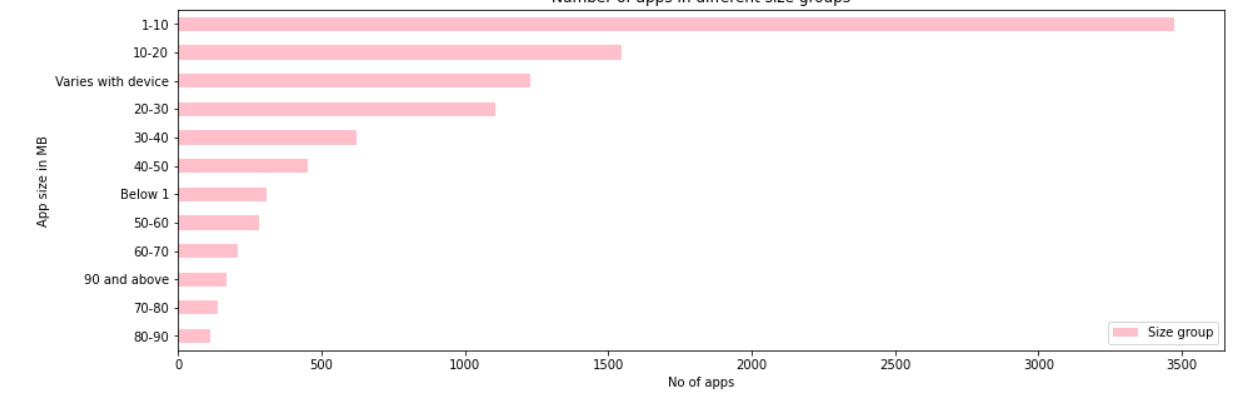
**Apps with the highest number of reviews:**

* We can state that the apps with more reviews, whether positive, negative, or neutral, are more popular than the others.
* This is because the number of user reviews indicates that these individuals have engaged with the app's content and have written their opinions on it.
* From the visualization below, we can say that “Facebook” is the most popular app in the Play Store based on the number of reviews.

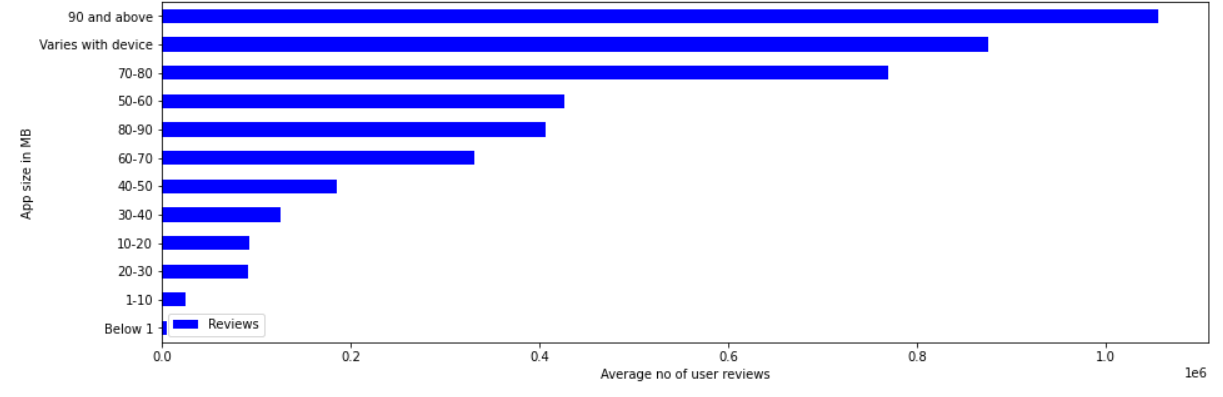


**Distribution of apps based on its size:**

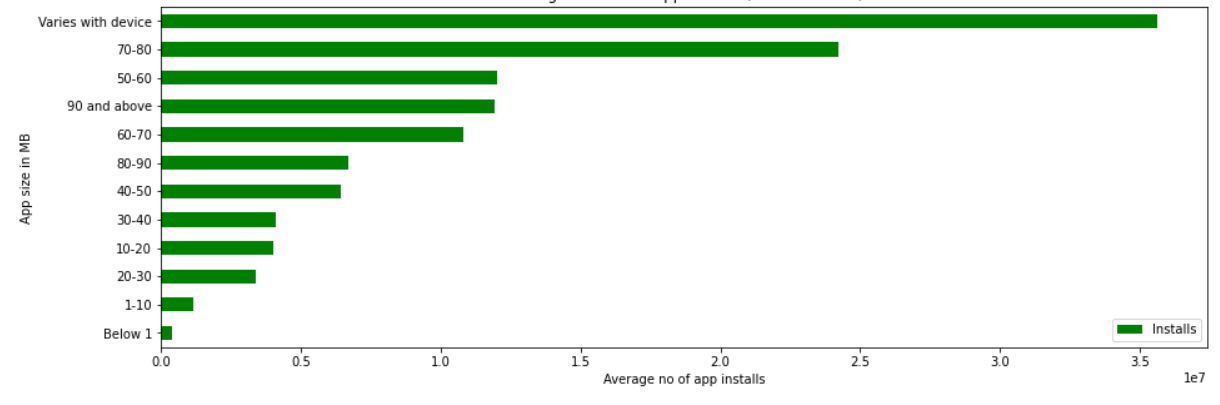
* The size of an app in our database varies from 100 MB to 0.0083 MB. We can analyse the size of the apps if we can group them into certain intervals.
* Here, we’ll group the data in the size column as follows into intervals of 10 each: (< 1 MB, 1-10, 10-20, 20-30, ..., 90-100, 'Varies with device').
* The visualization below gives the number of apps present in each size group. The higher this number, the higher is the competition.
* The groups “1-10” and “10-20” have the highest number of apps compared to the remaining size groups.
* If the developer wants to launch an app within this size range, the app will face tight competition from the apps that are already present.



* The plot below gives the average number of user reviews per app in each size group. We can see that the apps whose size is above 90 MB tops this list.



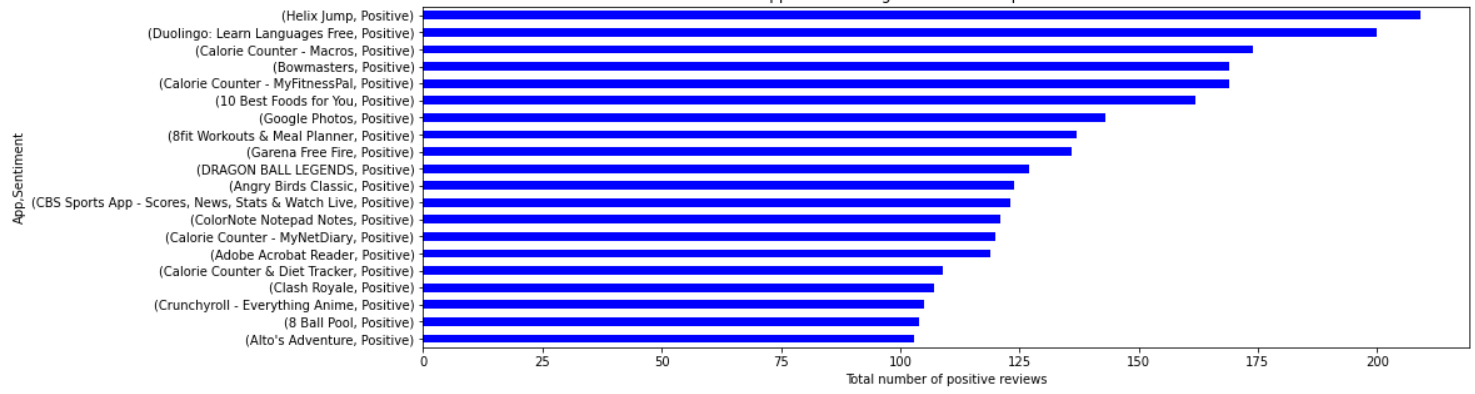
* The plot below gives the average number of app installs in each size group. We can see that the apps whose size varies with device tops this list.

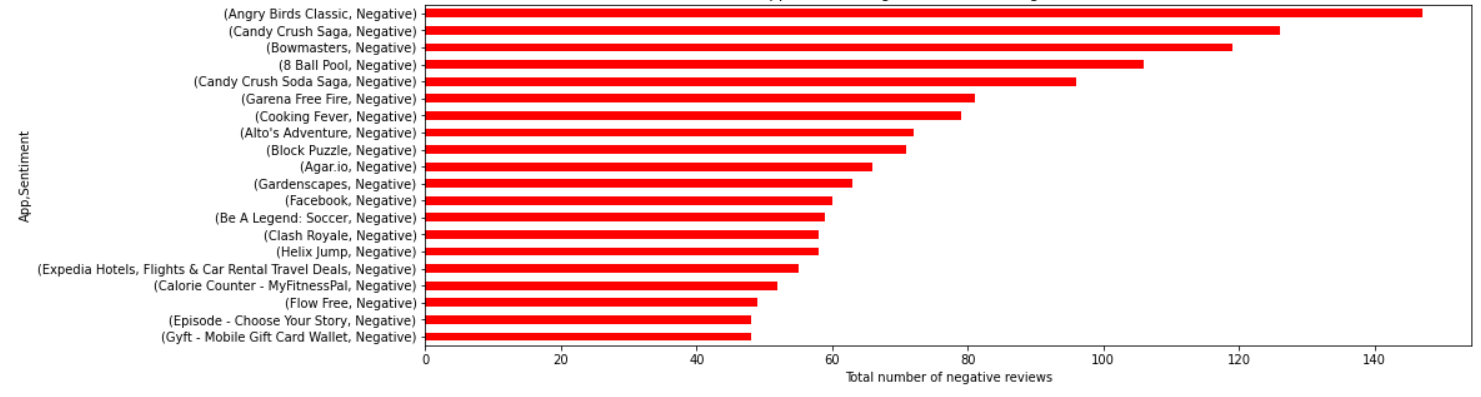


* As we have established earlier that the number of user reviews and installs gives the popularity of an app, we can say the same about this as well.
* The majority of the apps in the play store are in the size range of 1-20 MB, but when it comes to popularity, the apps which are bulky are more popular than the former.

**Apps with the highest number of positive and negative reviews:**

* A good number of positive reviews for an app indicates that a lot of users are happy with the services offered by that app.
* From the user\_reviews dataset, “Helix Jump” has the highest number of positive reviews.
* If an app has a lot of negative reviews, it indicates that a lot of users are dissatisfied with the services offered by that app.
* From the user\_reviews dataset, “Angry Birds Classic” has the highest number of negative reviews.





**Conclusion:**

These are some of the aspects that the developer should research before proceeding with the app development. By conducting a simple exploratory data analysis (EDA) on the play store dataset, we not only eliminate avoidable risks of failure, but we may also be able to provide better ideas for building the app.

* Percentage of free apps = ~92%
* Percentage of apps with no age restrictions = ~82%
* Most competitive category: Family
* Category with the highest number of installs: Game
* Category with the highest average app installs: Communication
* Percentage of apps that are top rated = ~80%
* There are 20 free apps that have been installed over a billion times
* Minecraft is the only app in the paid category with over 10M installs. This app has also produced the most revenue only from the installation fee.
* Category in which the paid apps have the highest average installation fee: Finance
* Most popular app in the Play Store based on the number of reviews: Facebook
* The median size of all apps in the play store is 12 MB.
* The apps whose size varies with device has the highest number average app installs.
* The apps whose size is greater than 90 MB has the highest number of average user reviews, i.e., they are more popular than the rest.

**References:**

* GeeksforGeeks
* Analytics Vidhya
* Smartlook
* Stackoverflow
* Towards data science
* Python libraries documentation
* Data camp