EDA on Play Store App Reviews

Bhaskar Purimitla ,Tejaswi Padarthi - Data Analysts

Glucksort Private Limited

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.

- **Installs**: The approximate number of times the respective app was installed.
- **Type**: It states whether an app is free to use or paid.
- **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.

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.
- **Android Ver:** The 3 rows containing NaN values were dropped from the dataset.

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:

- **App**: The duplicate values in the dataset were dropped based on the this 'App' column.
- **Rating:** The 1463 NaN values were imputed with its corresponding category median value.
- **Size:** Converted all the values into a single units like(all values in MB's).
- **Type:** One row containing NaN value was replaced with a mode of the column.
- **Price :** The '$' symbol was removed , and converted into numeric datatype.
- **Current Ver**: The 8 rows containing NaN values were dropped from the dataset.
- **Android Ver**: The 3 rows containing NaN values were dropped from the dataset.

Added Columns:

- **Revenue Column**: Added to the Playstore dataset by multiplying Price and Installs columns to estimate app revenue.
- **Size_group**: Added size_group column for size intervals in playstore data.
- **Sentiment Numeric Column:** Added to the Reviews dataset, mapping sentiments to numeric values:

  0: Positive

  1: Negative

  2: Neutral

- Simplifies sentiment analysis.

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

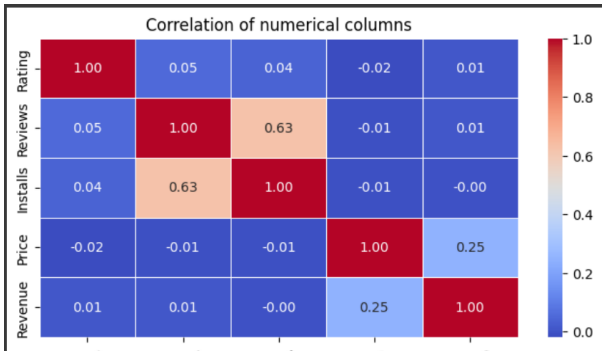
Data Visualization:

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.

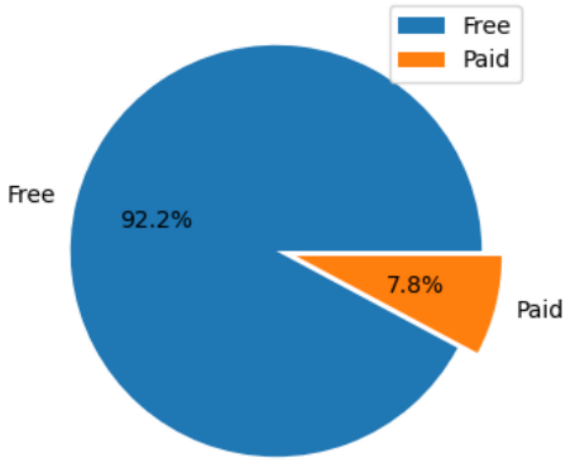
- The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution.

- 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:



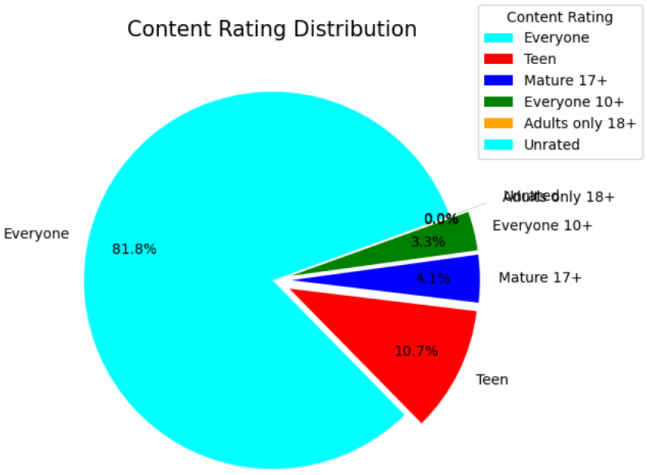
- There is a strong positive correlation between the Reviews and Installs column. This is pretty much obvious. Higher the number of installs, higher is the user base, and higher are the total number of reviews dropped by the users.
- The Price is slightly negatively correlated with the Rating, Reviews, and Installs. This means that as the prices of the app increases, the average rating, total number of reviewers and installs fall slightly.
- The value of correlation coefficient lies between -1 and 1.
- 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:

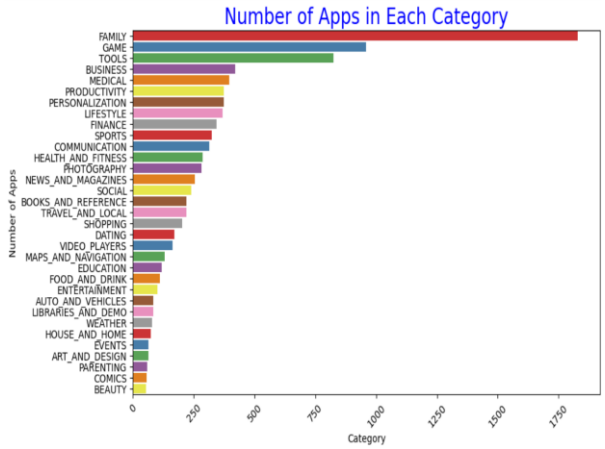


- Approximately 82% of the app in the play store have No age restrictions to install and use the app.

- 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.

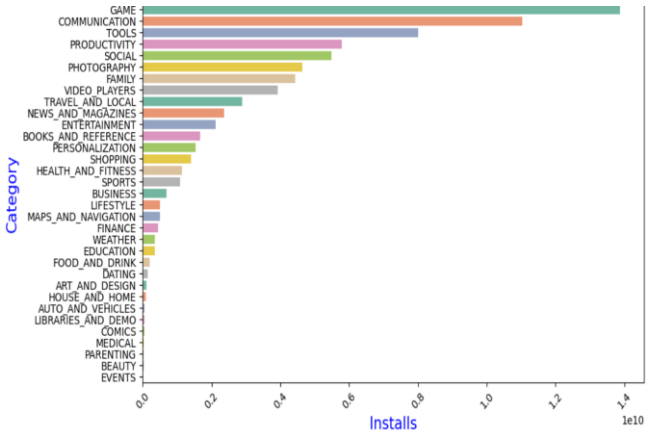


Total app installs in each category:

- We can say that the total number of installs and reviews for each category 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.

- 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

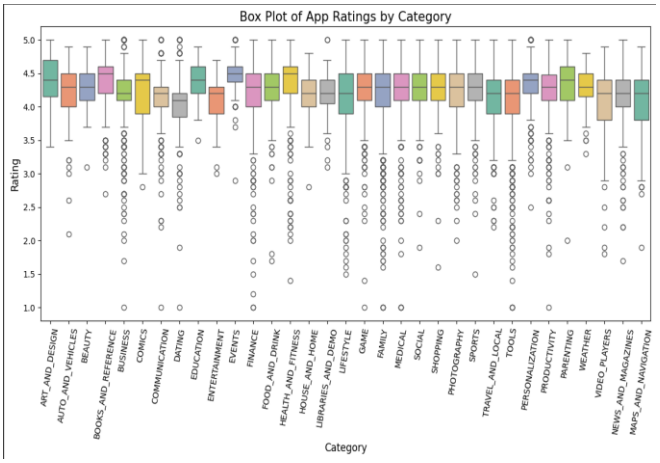
- Hence, the **Game**, **Communication**, and “**Tools**” are the most popular categories compared to the rest.



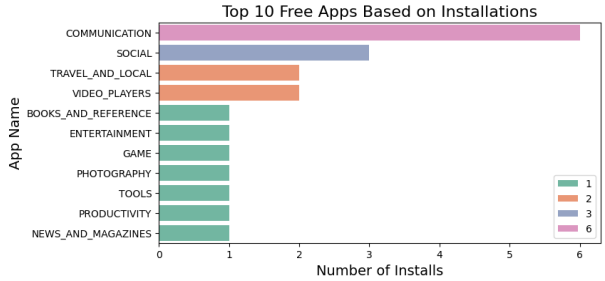
Rating distribution across different Categories:

- The Distribution of user rating for all the categories as shown below.

This metric can be used by a developer to find and study the categories 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 categories which are popular among the users based on Rating and implement some strategies in their app.



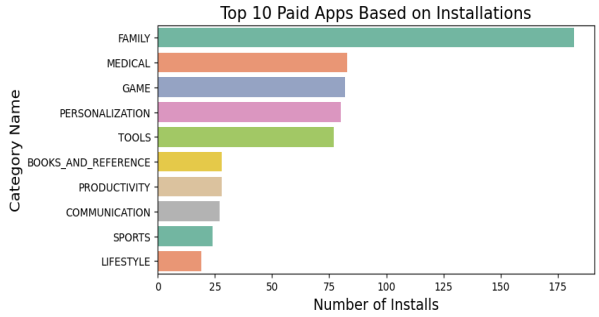
Top 10 apps that are of Free type :



- 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).

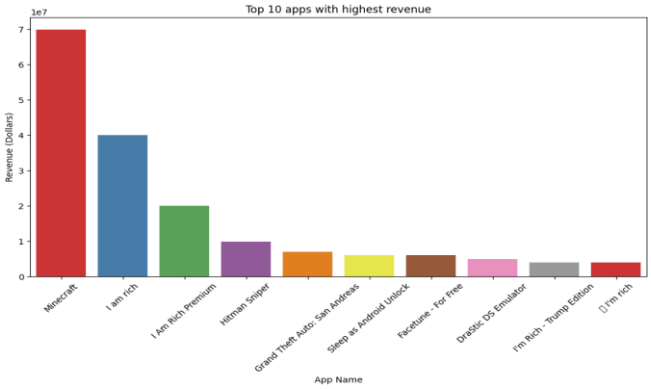
Top apps that are of Paid type:

- The paid apps charge the users a certain amount to download and install the app. This amount varies from one app to another.

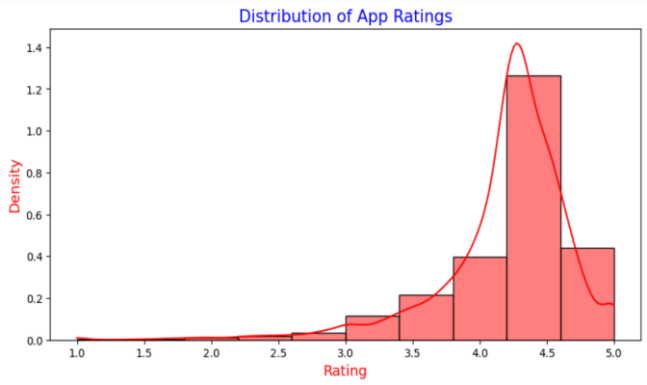


- In order to select the top paid apps, it won't be fair to look just into the number of installs. This is because the apps that charge a lower installation fee will be installed by a greater number of people in general.
- There are a lot of apps that charge a small amount whereas some apps charge a larger amount.
- 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).

Top Apps By Revenue:

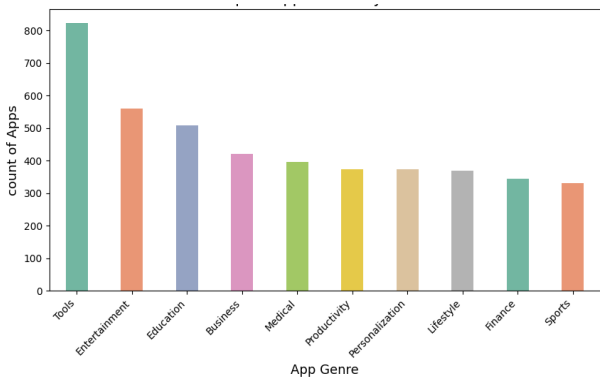


Distribution of App Ratings:



- The majority of ratings are concentrated **between 4 and 4.5** stars, indicating strong user satisfaction.
- Positive ratings highlight the app's success in meeting user expectations.
- This trend can be leveraged to attract new users and build credibility.

Top 10 Genres Based on App Count:



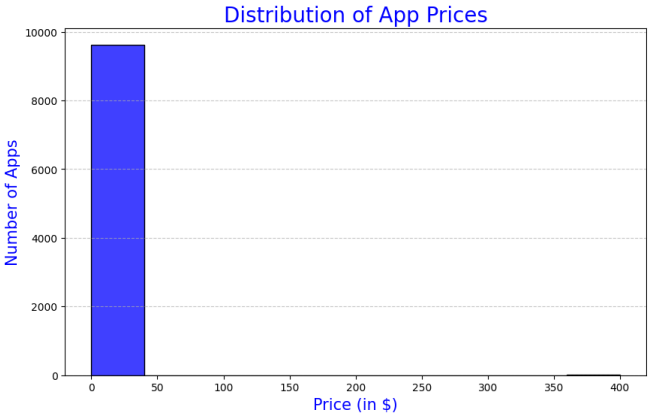
- The graph highlights the **top 10 genres** with the highest number of apps on the Play Store, showcasing their popularity and dominance.
- It is evident that most apps belong to these genres, indicating user interest and developer focus in these categories.
- This trend suggests that developers have extensively explored these genres, leading to a competitive landscape within them.
- To stand out and address untapped opportunities, developers should consider focusing on less saturated genres to create innovative applications.

Price Distribution of Apps:

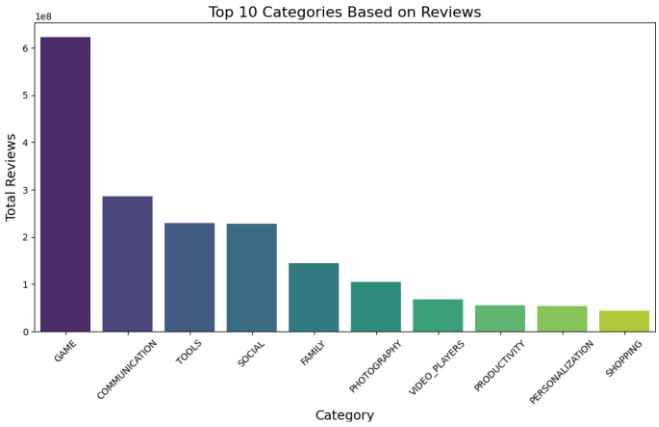
- The histogram reveals that the majority of app prices are concentrated between **$0** and **$50**, indicating affordability for most users.
- This price distribution suggests that developers target a broad audience by keeping app prices accessible.
- The concentration in this range highlights a potential opportunity to explore premium pricing for niche or high-value apps.

Rating Given by different Age group:

- Analysing lower ratings can identify improvement areas to enhance user experience.

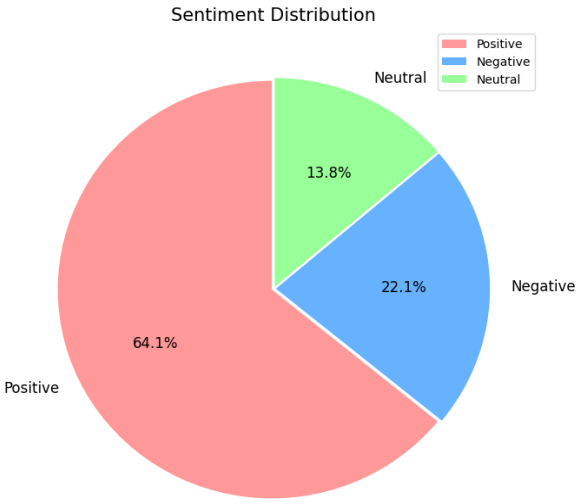


Top 10 Categories with Highest no of Reviews:



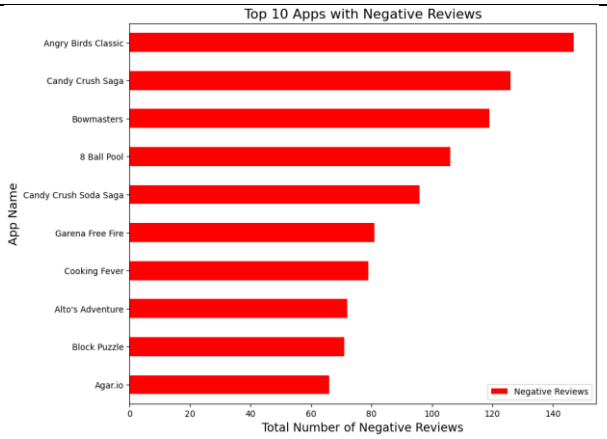
- The graph highlights the top 10 categories based on the number of reviews, providing a clear view of user engagement across categories.
- Categories such as **Games** and **Social** receive the highest number of reviews, reflecting their immense popularity and widespread user interaction.
- This indicates that apps in these categories are highly effective in capturing user attention and encouraging feedback.
- Developers aiming for higher user engagement and visibility can benefit from

Sentiment Distribution in reviews Dataset:



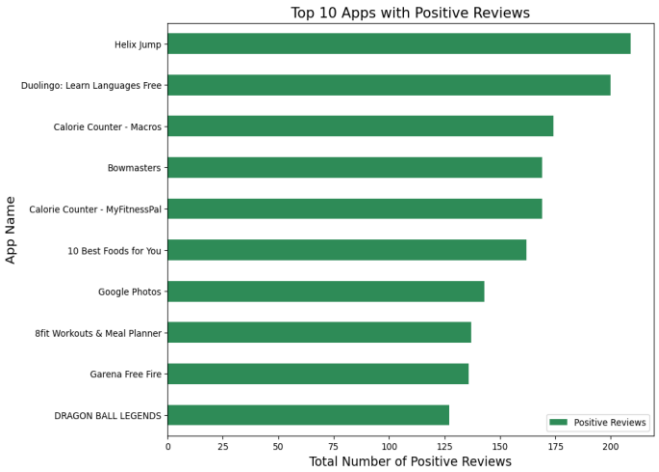
- The graph shows that **64.1%** of reviews **express positive sentiment**, highlighting strong user satisfaction with most apps.
- **Negative sentiment, at 22.1%,** points to areas where apps can be improved to better meet user expectations.
- **Neutral sentiment** makes up **13.8%,** indicating an opportunity to engage these users and convert them into positive reviewers.
- These insights emphasize the importance of sustaining positive feedback while addressing concerns to improve the overall user experience.
- These findings can help developers understand user preferences and identify areas of improvement for apps with lower positive feedback.

Top Apps with most Negative Reviews:



- The graph for negative reviews reveals the top 10 apps with the highest
- **"Angry Birds Classic"** and **"Candy Crush Saga"** top the list, indicating that these apps have attracted significant negative sentiment from users.
- **"Agar.io"** ranks last in the top 10, with the least negative reviews among the group, suggesting relatively better user perception compared to others.
- These insights can help developers focus on improving the user experience for apps with higher negative reviews while analysing the reasons behind dissatisfaction.
- percentage of negative feedback from users.

Top Apps with most positive reviews:

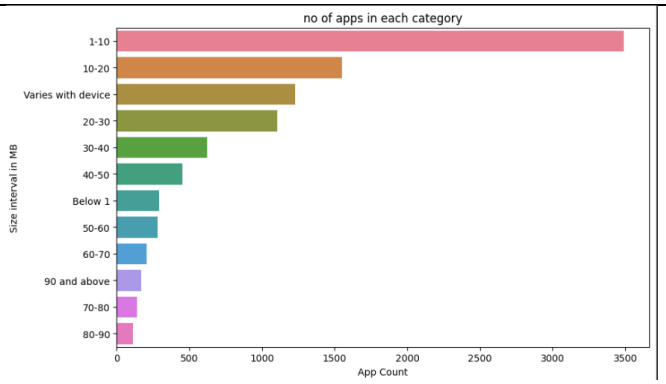


- The graph highlights the top 10 apps with the highest number of positive reviews,

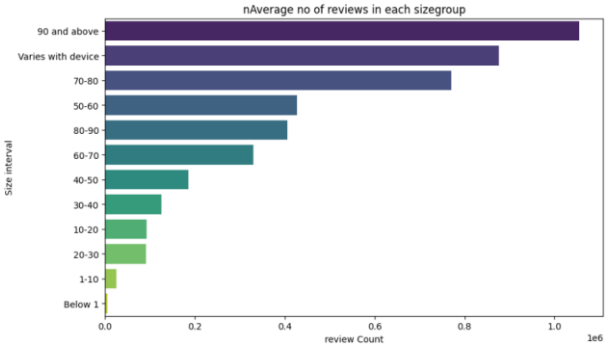
- providing insights into user satisfaction for each app.
- **Helix Jump** app leads the list with the highest number of positive reviews, followed by **Duolingo: Learn Languages Free**, indicating strong user approval for these apps.
- On the other hand, **Dragon Ball Legends** ranks last among the top 10 apps in terms of positive reviews, suggesting relatively
- lower user satisfaction compared to the others.

Distribution of apps based on its size:

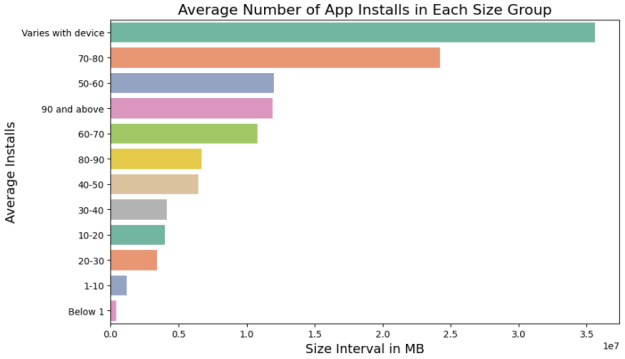
- The size of an app in our database varies from 100 MB to 0.0083 MB.
- We can analyse the size of the apps if we can group them into certain intervals.
- Here, we'll group the data in the size column as follows into intervals of 10 each: (< 1 MB, 1-10, 10-20, 20-30, ..., 90-100, 'Varies with device').
- 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.

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.

- Most of the apps are **Free** with **92.2%.**
- The category with highest average rating is **EVENTS.**
- The category with lowest average rating is **DATING.**
- Most competitive category: **Family**
- Almost all apps are targeting **everyone** with **81.8%.**
- All age groups has almost same Average rating.
- The Genre **TOOLS** has highest no of apps.
- Most Competitive category based on installations : **Games**
- The app with highest revenue : **Minecraft**
- Most of the apps contains positive reviews rather than negative reviews.
- **Helix Jump** has highest positive reviews.
- **Angry** Birds Classic has highest negative reviews.
- Majority of the Prices are between **0** to **50 dollars**.
- Most of the ratings fall between **3.0** to **4.5**
- The app with highest no of installations without any price (Free) is **Hitman Sniper.**
- 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.

Resources:

1.Complete EDA using python: https://medium.com/@ugursavci/complete-exploratory-data-analysis-using-python-9f685d67d1e4
2.W3schools for python Libraries.
3.Greeksforgreeks
4.chatGPT.