## 📋 **Abstract**

A few thousands of new applications are regularly uploaded on Google play store. A huge number of designers working freely on designing the apps and making them successful. With the enormous challenge from everywhere throughout the globe, it is important for a developer to know whether he/she is continuing the correct way or not. Since most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, in-application adverts and memberships add to the achievement of an application. In this way, an application's prosperity is normally dictated by the quantity of installation of the application and the client appraisals that it has gotten over its lifetime instead of the income is created. The objective of this experiment is to deliver insights to understand customer demands better and thus help developers to popularize the product. We have tried to discover the relationships among various attributes such as which application is free or paid, what are the user reviews, rating of the application. Inserting image...

Inserting image...

## **📖 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 analyze 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 analyze the data to discover key factors responsible for app engagement and success.

Inserting image...

### **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.

Inserting image...

## **📋Problem Statements**

1. What are the top categories on Play Store?
2. Are majority of the apps Paid or Free?
3. How importance is the rating of the application?
4. Which categories from the audience should the app be based on?
5. Which category has the most no. of installations?
6. How does the count of apps vary by Genres?
7. How does the last update have an effect on the rating?
8. How are ratings affected when the app is a paid one?
9. How are reviews and ratings co-related?
10. Let's us discuss the sentiment subjectivity.
11. Is subjectivity and polarity proportional to each other?
12. What is the percentage of review sentiments?
13. How is sentiment polarity varying for paid and free apps?
14. How Content Rating affect over the App?
15. Does Last Update date have an effect on rating?
16. Distribution of App update over the Year.
17. Distribution of Paid and Free app updated over the Month.

## **📔 What is Exploratory Data Analysis?**

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets for patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset and summarize their main characteristics, often employing data visualization methods. It is an important step in any Data Analysis or Data Science project. It helps determine how best to manipulate data sources to get the answers you need.

EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better and make it more attractive and appealing.

The following are the various steps involved in the EDA process:

1. **Problem Statement** - We shall brainstorm and understand the given data set. We shall study the attributes present in it and try to do a philosophical analysis about their meaning and importance for this problem.
2. **Hypothesis** - Upon studying the attributes present in the data base, we shall develop some basic hypothesis on which we can work and play with the data to look for the varied results which we can get out of it.
3. **Univariate Analysis** - It is the simplest form of analyzing the data. In this we would initially pick up a single attribute and study it in and out. It doesn't deal with any sort of co-relation and its major purpose is to describe. It takes data, summarizes that data and finds patterns in the data.
4. **Bivariate Analysis** - This analysis is related to cause and the relationship between the two attributes. We will try to understand the dependency of attributes on each other.
5. **Multivariate Analysis** - This is done when more than two variables have to be analyzed simultaneously.
6. **Data Cleaning** - We shall clean the dataset and handle the missing data, outliers and categorical variables.
7. **Testing Hypothesis** - We shall check if our data meets the assumptions required by most of the multivariate techniques.

Inserting image...

## **📖 Steps Involved**

After loading the dataset, we can start the exploration but before that, we need to check and see that the dataset is ready for performing several exploration operations or not, so let’s first have a look at the structure and the manner in which the data is organized.

### Data Cleaning

Our data set contains a large number of null values in the rating column, so we drop them. Some of the columns have a smaller number of null values, so we replace the null values in these columns with the mode value of that particular column. Our data set also contain the duplicate rows for a single application. We also drop the duplicate rows because the rows contain the identical data. Also drop the rows, which have rating greater than 5.

### Data Transforming

From the information of data frame, we can see that all the columns except rating have the object data type but some of the columns like, reviews, size, installs and price have the numerical value. So, we have to transform them in proper data type and also remove the unwanted values from the numerical columns like ‘+’ and ‘,’ from installs and ‘$’ from price. In the size column we have some values in KB and some values in MB, so we transform all the values in MB.

### Exploratory Data Analysis

After establishing a good sense of each feature, we proceeded with plotting a pairwise plot between all the quantitative variables to look for any evident patterns or relationships between the features. There is a high variance in the number of installs and in number of reviews. To overcome this problem, we add two new columns to the data frame named: log\_installs and log review, which contain the logarithmic values of installs and review columns, respectively.

### Single Variate Analysis

After that we analysis all the columns one by one to examine whether the particular column contain some useful information or not:

### Category

We breakdown the apps by category and observe that family and game categories have the maximum number of apps in the play store. Weather, house and home, comics, events, beauty, and parenting are the categories which have a few numbers of apps.

### Data wrangling

Apart from this, two new columns were added to the main data frame, namely, “Rating Group”, and “Revenue”. This is done to improve simplify the analysis and come up with different meaningful visualizations

* **Rating Group:** This column groups the apps based on the average user rating. (4-5: Top rated, 3-4: Above average, 2-3: Average, 1-2: Below average).
* **Revenue:** This column gives the revenue generated by the app through app installs alone. By doing these operations on the original dataset, we are ready with the data pipeline, and data visualizations can be done on it. All the apps in play store have the rating between 0.5 to 5. Maximum apps have the rating between 3.8 to 4.5.

### Installs

We analysis the install column to observe the effect of size, price, rating, content rating, android version on app installation number. We can analysis that for each and every category number of app installation does not depend on the size. The free apps installed mostly. The apps which can be used by everyone is more installed than the apps which can be used by a particular age group. Rating of mostly installed apps is between 4 and 5.

## **🛠 Challenges Faced**:

* Reading the dataset and comprehending the problem statement. Our major challenge was data cleaning.
* Handling the error, duplicate and NaN values in the dataset.
* 13.60% of reviews were NaN values, and even after merging both the dataframes, we could not infer much in order to fill them. Thus, we had to drop them.
* The merged data frame of both play store and user reviews, had only 816 common apps. This is just 10% of the cleaned data, we could have given more valuable analysis, if we had at least 70% - 80% of the data available in the merged dataframes.
* User Reviews had 42% of NaN values, which could have been used for developing an understanding of the category wise sentiments, which would help us to fill 13.60% NaN values of the Reviews column.
* There is so much more which can be explored. Like we have current version, android version available which can be explored in detail and we can come out with more analysis where we can tell how does these things effect and needs to be kept in mind while developing app for the users.
* We can explore the correlation between the size of the app and the version of Android on the number of installs.
* Machine learning can help us to deploy more insights by developing models which can help us interpret even more better. We have left this as future work as this is something where we can work on.
* Designing multiple visualizations to summarize the information in the dataset and successfully communicate the results and trends to the reader.

## **📋 Conclusion:**

Most of the apps are free so developers should focus on creating free apps to have a huge customer base. If developing paid apps then apps size should not be greater than 40mb. More Apps should be in the category like Events,Beauty,Parenting as they have not been explored much but still quite popular with huge installations. In order to retain the customer base apps should be updated regularly Developers should develop apps such that their content is available for everyone. Bulky apps should be developed in the category like Game, Family. If developing paid apps then its price should not be high and size should be less than 20mb. Apps belonging to Game and Family Category have high negative reviews therefore they should be developed carefully. Like this there can be a lot of conclusions but we have tried to cover the most important ones. 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.
* Helix Jump has the highest number of positive reviews and Angry Birds Classic has the highest number of negative reviews.

-----------------------------------------------------