**Play store App Review Analysis**

**Mayank Ghai, Mrugesh Patel**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

The Project “Play Store App and Review Analysis” is based on discovering key understandings about the different categories of apps available on the Google Play Store, the installation per category, the count of reviews per app, and the sentimental of the reviews per app.

In this analysis, we’re provided with a dataset with some records. We did a basic inspection, and data cleaning to remove suspicious data and to avoid an error. We’ve used different plots to visualize our analysis in the easiest way. This can be used to know which type or category of app is installed the most, count of the applications available per category, distribution of the rating how much of these apps are paid or are available for free, to know the sentiment of the reviews, the polarity of the sentiment.

***Keywords: EDA, Google Play Store, Data visualization, and exploration.***

**1. Problem Statement**

Google Play Store is one of the most used applications on any android phone as it allows the user to get any app that he or she requires. It has many categories to choose the type of application that the user wants to install on his device. Doing a data analysis on the data of the google play store along with the user review dataset will help us to get to know the most dominant category among all the available ones. It also allows us to track the number of installs per category and the most popular application among people of different age groups, the polarity of the sentiment of the reviews given to each application. It can also help us to fetch various information about the different applications like the size, the rating, the version available, and the price of the applications if a paid version of them is available. All this information can help us to fetch meaningful insights from the dataset provided.

The data provided to us have two datasets one about the application information and the second one telling us the reviews on each application available. After combing the two datasets together the number of observations went over 7 lakhs which is distributed among 17 columns and it is a mix of categorical and numeric values. Explore and analyze the data to discover key understandings.

**2. Introduction**

We live in an era where data is produced and circulated in an enormous amount. Those data can be collected and allow us to infer meaningful results and make well-informed decisions. However, as the number of data increases, it becomes difficult to analyze the data, here we need to visualize the data to help us in conducting data analysis. By using visualization tools, we can deliver a message to our audience and inform them about our findings.

This project explores a dataset from a technology company, maps the result clearly through visualization tools, and gives new insight to the public and other relevant parties.

We will explore and visualize the dataset from the Google Play Store App and User Reviews dataset, using basic exploratory data analysis techniques. We will find out the distribution of every application listing based on their category, type of the application available:- free or paid version, including price range, size of the application, number of installations per genre, and other related factors.

This store provides the application that can be used to install on the device. It is one of the largest online stores that provide the services. As we are moving into an era, where technology is one of the one things that can help us to keep up with this fast-moving world.

Our goal here is to explore the data and find useful insights from the data and find out different relations between the columns.

This EDA project aims to discover patterns that lead to a successful application on the Google Play Store. This will be done by analyzing the historical data collected from the Google Play Store according to the dataset source.

Google Play-Store Dataset columns description:-

* **App:** Application name
* **Category:** Category the app belongs to
* **Rating:** Overall user rating of the app (as when scraped)
* **Reviews:** Number of user reviews for the app (as when scraped)
* **Size:** Size of the app (as when scraped)
* **Installs:** Number of user downloads/installs for the app (as when scraped)
* **Type:** Paid or Free
* **Price:** Price of the app (as when scraped)
* **Content:** Rating Age group the app is targeted at - Children / Mature 21+ / Adult
* **Genres:** An app can belong to multiple genres (apart from its main category). For eg, a musical family game will belong to Music, Game, or Family genres.
* **Last Updated:** Date when the app was last updated on Play Store (as when scraped)
* **Current Ver:** Current version of the app is available on the Play Store (as when scraped)
* **Android Ver:** Min required Android version (as when scraped)

User Review Dataset column description:-

* **App:** Application name
* **Translate Reviews:** Reviews on various applications given by the user
* **Sentiment:** Positive, Negative, or Neutral
* **Sentiment Polarity:** Tell us about the polarity of the sentiment of the translated reviews that we received.
* **Sentiment Subjectivity:** Tell us about the subjectivity of the various reviews according to the translated review.

|  |  |
| --- | --- |
| **Fig 1: Age Group of the user we are targeting** | **Fig 2 : Distribution of the Free App and Paid App** |
| **Fig 3: Rating in the application** | **Fig 4: The Outler** |

|  |
| --- |
| **Fig 5: The Sentimental Polarity and the Sentimental Subjectivity** |
| **Fig 6 : The Original Google Dataset** |
| **Fig 7: The Original Google Dataset** |

# 3. **Steps involved:**

1. **Acquire data**:- For this project, we are using Google colab notebook IDE with a python programming language to write our script. Google Play Store dataset is available at<http://kaggle.com/.> This data file includes all needed information to find out more about the type of application, the size of the application, the current version of the application, the android version required for the application, and necessary metrics to make predictions and draw conclusions.
2. **Exploratory Data Analysis**:- After acquiring and loading the dataset, we performed this method by comparing all the columns. This process helped us to figure out patterns, spot anomalies, detect outliers, and find fascinating relations among the variables. By conducting EDA, we can turn an almost usable or unusable dataset into a usable dataset. The major components of exploratory data analysis are Cleaning datasets and exploring and visualizing data.

* **Cleaning dataset:** The next step is cleaning up the data. Often the data we load has various faults, such as typos, missing values, and incomplete data. By cleaning up, the data quality will be better to use for further analysis. Cleaning the data is a crucial step in Exploratory Data Analysis so that it can help us to get the dataset that can be used to get meaningful insights from the data.
  + **Null and unfitting values Treatment:** Google Play Store dataset contains many null values, so we tried to preserve as many rows as possible by replacing null values with suitable values. For categorical data, we will replace the null values with the mode of the categorical columns and for numerical data, we will replace the null values with the median of the numerical value in order to make it more presentable.

The user review dataset also has some missing values same procedure is done for this dataset based on whether the column is categorical or numerical.

* + **Non-Graphical and Graphical Univariate Analysis:** We analyze just one variable to describe the data and find patterns that exist within it. For which we used different graphs like a bar, histogram, box, etc. By which we came to know about some outliers.
  + **Outlier treatment:** Outlier treatment is required in this dataset only in the rating columns as it is the only column that has a value that can actually qualify to be an outlier as its value is 19 while the rating is confined to 1 to 5.
  + **Dropping Duplicates:** Both datasets have many duplicates in them and we do have to take care of these duplicates as they will create a hindrance in the proper analysis of the dataset
* **Data Manipulation:** To get meaningful insight from the dataset, we do have to manipulate the dataset because there are some typos that will not help us to get what we want. There are many columns in the dataset that have to be manipulated like:-
  + - Removing a **$ sign** from the Price as it was creating a hindrance while doing the analysis of the data.
    - Removing a **+ sign** from the Installs as it was creating a hindrance while doing the analysis of the data.
    - As the size of the applications are in MB (megabytes) and KB (kilobytes) so in order to make the data even we just have converted all the data to MB (megabytes)
* **Merging of Dataset:** The merging of the two datasets is done as both have the same column and keeping it as common both the dataset is merged. The merging of the dataset is really important as it was difficult to fetch insights from two datasets. In the merged data frame, we have four new columns i.e. Translated Review, Sentiment, Sentiment Polarity, and Sentiment Subjectivity. Sentiment basically determines the attitude or the emotion of the writer, i.e., whether it is positive or negative, or neutral. Sentiment Polarity is a float that lies in the range of [-1,1] where 1 means a positive statement and -1 means a negative statement. Sentiment Subjectivity generally refers to personal opinion, emotion, or judgment, which lies in the range of [0,1].
* **Data Visualization:** This step is the last and the most crucial step in the Exploratory Data Analysis as the above-mentioned steps are used here to get eloquent insights that can be represented graphically to get a clear representation of the data and what it is going to tell the viewer. Data visualization actually helps to ask various questions and to get the required answer with the help of a graphical representation. As it is a very large dataset and looking at the Excel datasheet will not going to help us if we need the answer to the question that will arise while doing the analysis.

While doing the analysis few questions came to our mind and these questions and their respective pictorial representation are as follows:-

* + - Installs per Category
    - Pricing vs Category
    - Number of Reviews per Category
    - What are the top 20 apps present in the google play store per the Genres?
    - Count of applications for each Category
    - Number of Installed applications for each category
    - Distribution of Rating
    - Highest and Lowest Genres
    - Rating according to Genres
    - How many apps were installed according to their type?
    - Different Android Version
    - What is the percentage of paid to free apps?
    - What is the most dominant category?
    - What price should my paid app have?
    - What are the number of installs for apps priced more than $100?
    - What is the most successful category?
    - Distribution of app sizes
    - Rating per category and type
    - Distribution of content rating per categories
    - Distribution of subjectivity
    - Does Sentiment subjectivity proportional to sentiment polarity
    - Word-cloud
    - Percentage of Review Sentiments

|  |  |
| --- | --- |
| **Fig 8: Number of Install per Category.** | **Fig 9: Price per Category** |
| As we can see applications that have the highest number of install comes from the communication, photography, and social category. | As we can see applications that have the highest price come from the social, tools, medical and personalization category |
| **Fig 10: Number of Install per Category.** | **Fig 11: Top 20 Genres** |
| The Category that has the highest number of reviews are social, games, photography, family, and communication. | The maximum number of apps present in the google play store comes under Tools, Entertainment, and Education Genres but as per the installation and requirement in the market plot, the scenario is not the same. Maximum installed apps come under Communication, Photography, and Social Genres. |

|  |  |
| --- | --- |
| **Fig 12: Count of applications for each Category** | **Fig 13: Number of installed applications for each c per Category** |
| As we can see that the application under Family, game, and tools has the highest number of applications under their category | Games, Communication, and Productivity have the highest number of installed applications. |

|  |  |
| --- | --- |
| **Fig 14: Highest Rating** | **Fig 15: Lowest Rating** |
| It shows the highest rating among the different Genres Rating | It shows the lowest rating among the different Genres Rating |

|  |  |
| --- | --- |
| **Fig 16: Distribution of Ratings** | **Fig 17: Count of applications in each category differentiated by their type** |
| This graph shows that around 62.4% of the rating are distributed from 4.1 to 4.5 with 4.3 being the mean | Family, Games, and Tools are the categories that have the highest count of application differentiated by their type |

|  |  |
| --- | --- |
| **Fig 18: Different AndroiVersionson** | **Fig 19: Pie chart representing the Type of application distributed among the dataset** |
| Shows the different versions of android that are required by the application. | Around 92.6% of the applications available on the store are free |

|  |  |
| --- | --- |
| **Fig 20: Most Dominant category** | **Fig 21: Number of install type wise according to the category** |

|  |  |
| --- | --- |
| **Fig 22: Distribution of the app size** | **Fig 23: Most successful category** |
|  | **Finance and lifestyle are valued the most but it's interesting to find that there are apps with values of more than $100.**  **On average, an app price is $1** |

|  |  |  |
| --- | --- | --- |
| **Fig 24: Distribution of content rating per category** | | |
|  | | |
| **Fig 24: The most Successful category** | | |
| The most successful category among all of them is communiaction | | |
| **Fig 25: Number of reviews type-wise according to Categories** | **Fig 26: What are the number of installs for apps priced more than $100?** |
|  |  |

|  |  |
| --- | --- |
| **Fig 27: Percentage of Review Sentiments** | **Fig 28: Does sentiment subjectivity proportional to sentiment polarity** |
| So 64,65% of the reviews come under positive, 23.18% are negative while the rest comes under neutral | **From the above scatter plot it can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in a maximum number of cases, shows a proportional behavior, when variance is too high or low** |
|  |  |
| **Fig 29: Distribution of Subjectivity** | **Fig 27: Word-Cloud** |
| **It can be seen that the maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this, we can conclude that the maximum number of users give reviews to the applications, according to their experience.** | Most of the words are excessive, kids bad, and allowed that are typed in the reviews |

**4. Conclusion:**

That’s it! We reached the end of our exercise.

Starting with loading the data, So far we have done identification of variables and data types, Missing value treatment, Non-Graphical and Graphical Univariate Analysis, outlier treatment, and Correlation Analysis and learned that

* As we can see applications that have the highest number of install comes from the communication, photography, and social category.
* As we can see applications that have the highest price come from the social, tools, medical, and personalization categories.
* The Category that has the highest number of reviews are social, games, photography, family, and communication.
* As we can see from the above plots: The maximum number of apps present in the google play store comes under Tools, Entertainment, and Education Genres but as per the installation and requirement in the market plot, the scenario is not the same. Maximum installed apps come under Communication, Tools, and Productivity Genres.
* As we can see that the application under Family, game, and tools has the highest number of applications under their category
* As we can see from the above two plots: The maximum number of apps present in the google play store comes under Communication, Social, Productivity and Tools, Genres but as per the installation and requirement in the market plot, the scenario is not the same. Maximum installed apps come under Communication, Tools, and Productivity Genres.
* **It looks like certain app categories have more free apps available for download than others. In our dataset, the majority of apps in the Family, Games, and Tools categories were free to install. At the same time Family, Tools, and Medical categories had the biggest number of paid apps available for download.**
* It can be concluded that the number of free applications installed by the user is high when compared with the paid ones.
* The trend is games have a much larger download size than any other category. If you're developing anything other than a game, you need to optimize your download size. on Average, any app should have a size of 18.15 MB.
* This is an interesting finding, our first assumption for the most dominant category was family.
* Finance and lifestyle are valued the most but it's interesting to find that there are apps with values of more than $100.

On average, an app price is $1

* It can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in a maximum number of the case, shows a proportional behavior, when variance is too high or low
* Most successful category based on the number of installs is the communication category.
* So 64,65% of the reviews come under positive, 23.18% are negative while the rest comes under neutral
* It can be seen that a maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this, we can conclude that a maximum number of users give reviews to the applications, according to their experience.

**References-**

1. GeeksforGeeks
2. Stack Overflow
3. medium