**Play Store App Review Analysis**

**Sanjay Jaiswal**

**Sachin Dubey**

**Jayalaxmi Mekap**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

Google Play Store and formerly Android Market, developed by Google. It serves as the official app store for certified devices running on the Android operating system, It allows users to browse and download applications developed with the Android software development kit (SDK) and published through Google. Google Play also serves as a digital media store, offering music, books, movies, and television programs.

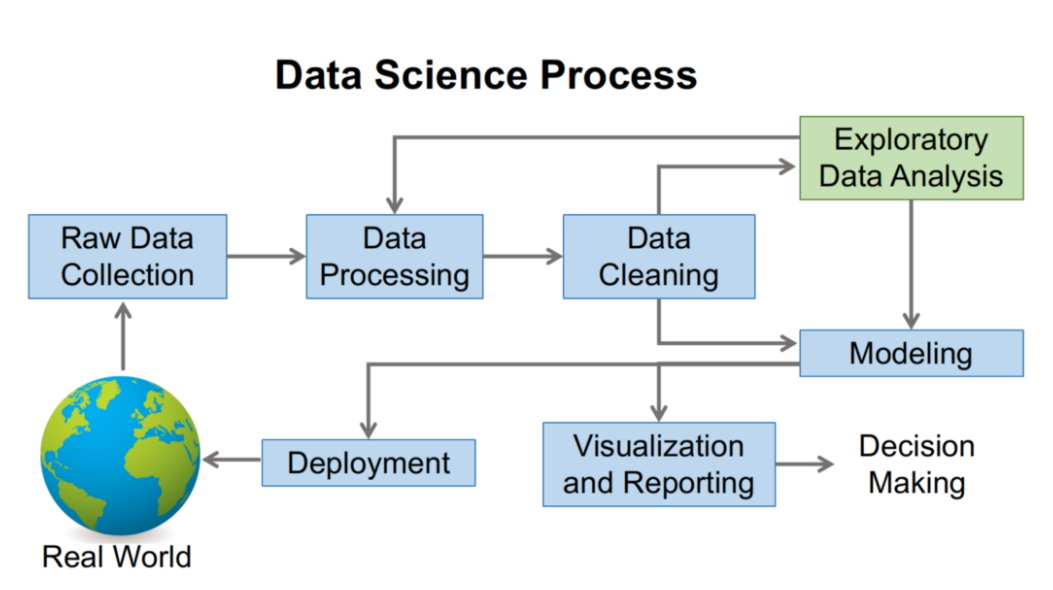
Google Play featured more than 3.5 million Android applications. Android Market was announced by Google on August 28, 2008, and was made available to users on October 22.

The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market

The objective of this project is to deliver insights to understand and analyze the customer demands better and which help to discover key factors responsible for app engagement and success.

Discussion of Google play store dataset will involve various steps such as:

* loading the data into data frame
* cleaning the data
* extracting statistics from the dataset
* exploratory analysis and visualizations
* questions that can be asked from the dataset
* conclusion



**1.Problem Statement**

The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market.

Each app (row) has values for category, rating, size, and more. Another dataset contains customer reviews of the android apps.

Explore and analyze the data to discover key factors responsible for app engagement and success.

**2. Introduction**

We were provided the in-CSV format and it is of 10k Play Store apps for analyzing the Android market and containing the following the 13 Columns -

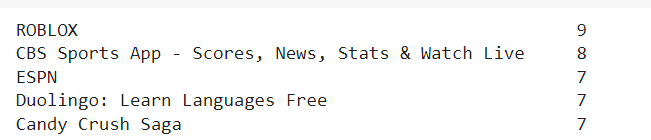
* **App**: This column hold name of App
* **Category**: The is defined for type of App Category and contain the also 30 type categories
* **Rating**: Rating is defined for user rating for respective app
* **Reviews**: This column contact the count of buyer review for the app
* **Size**: This is for App Size
* **Installs**: Number of installations
* **Type**: Whether App available free or Paid
* **Price**: Price of App if paid
* **Content Rating**: Given preference for user as per age wise or suitable for everyone
* **Genres**: It is containing the style of category
* **Last Updated**: App last update
* **Current Ver**: Currently App version
* **Android Ver:** Suitable for android version



**3. Data Cleanings and validations**

In this step removing faulty data and filling in gaps. The task to be crucial and important thus validating by following steps

* Removing extraneous data
* Handing in missing values.
* Data shifting in respective columns
* Conforming data to a standardized pattern.
* We found that Duplicate app’s values and It may have happened that for the same app, the data has been scraped at different points of time.

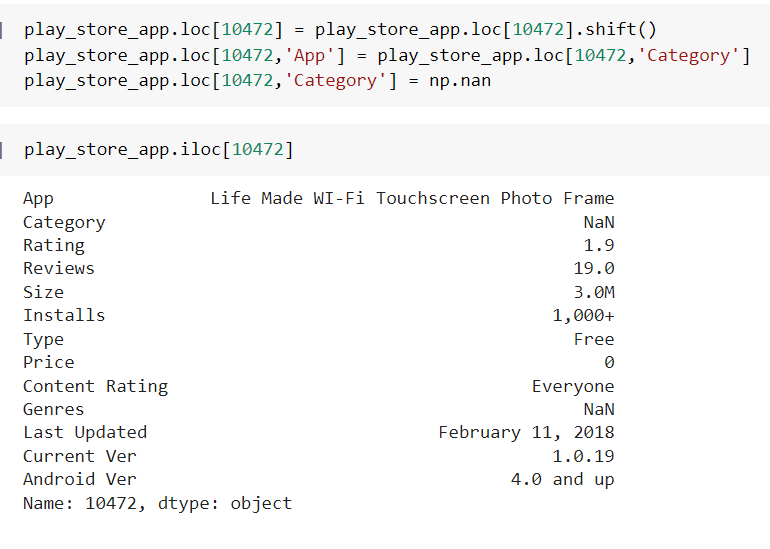




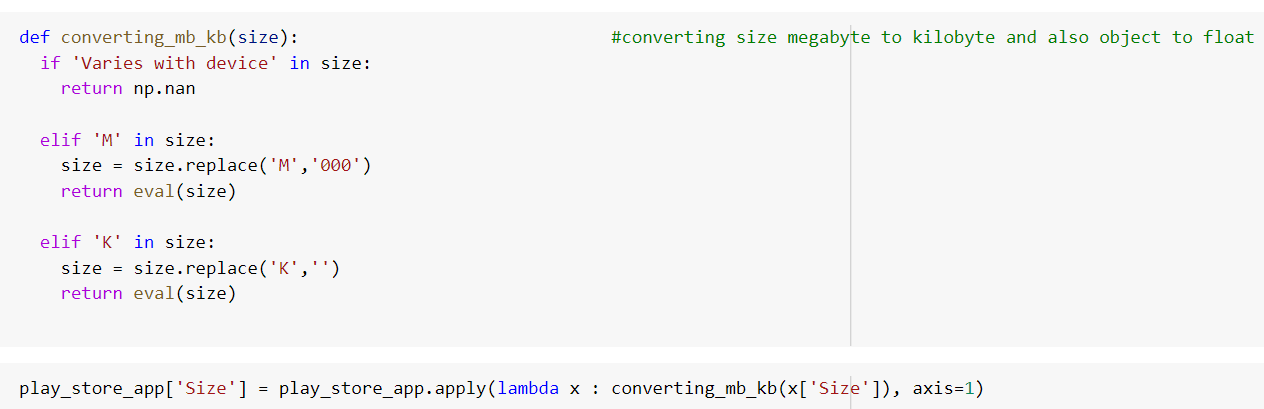
**4. Transform data**

Transforming data is the process of updating the format or value entries in order to reach a well-defined outcome, or to make the data more easily understood by a wider audience.

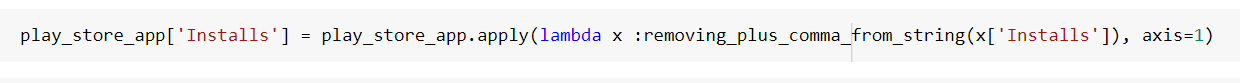
* Analyzed and found inappropriate value in the columns, so shifting in respective columns by using the shift method



* The application sizes available in the dataset are in MB and KB. For ease in data processing, converting the MB to KB then later removing KB from data value.



* Converting the number of installations in integer value for analyzing



* Removing ‘$’ symbol from price part



* Changing the date format for last update



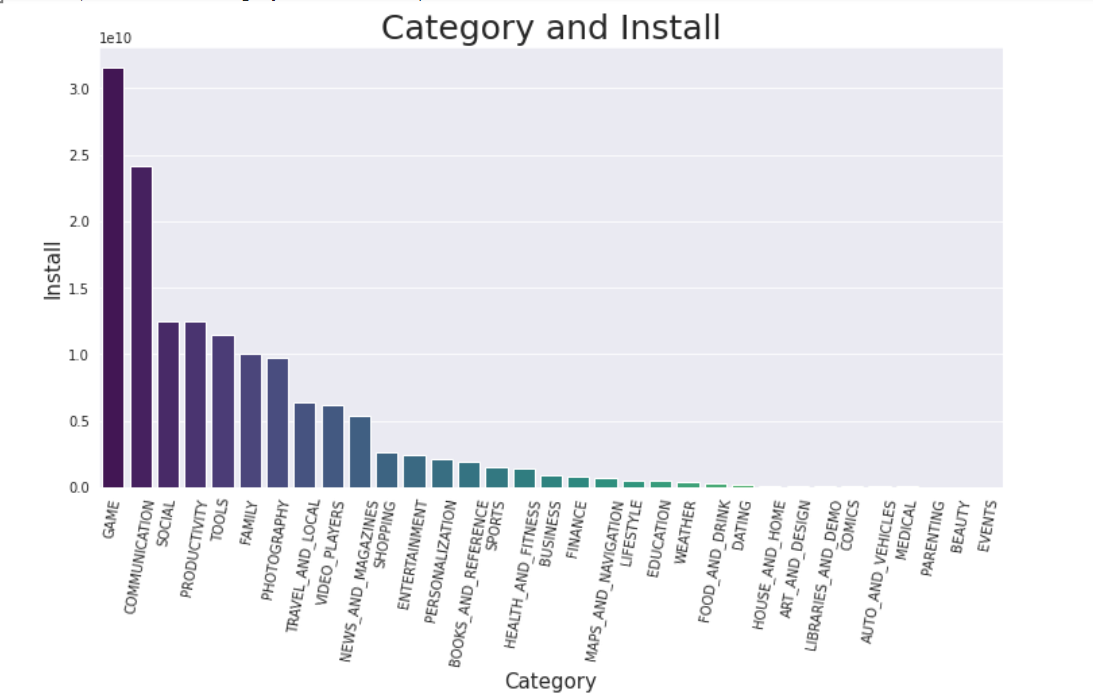
**5. Exploratory Analysis and Visualization**

Exploratory data visualizations are the type of visualizations we assemble when we do not have a clue about what information lies within our dataset.

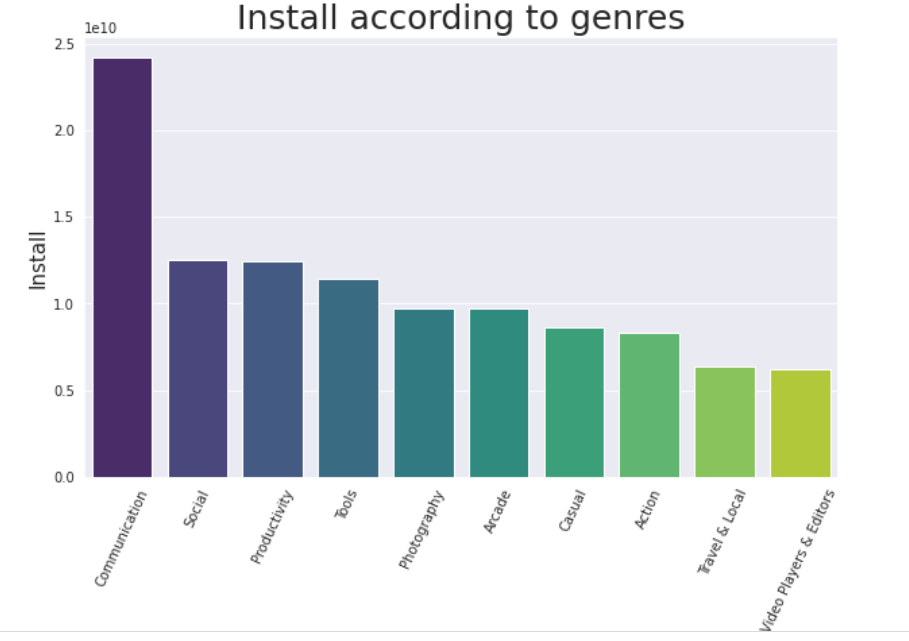
* Top app category in play store; the frequency of app installation indicates the engagement of top category, among the all Family, games and tools are top three category



* App installed according to category (Number of Installed applications for each category)

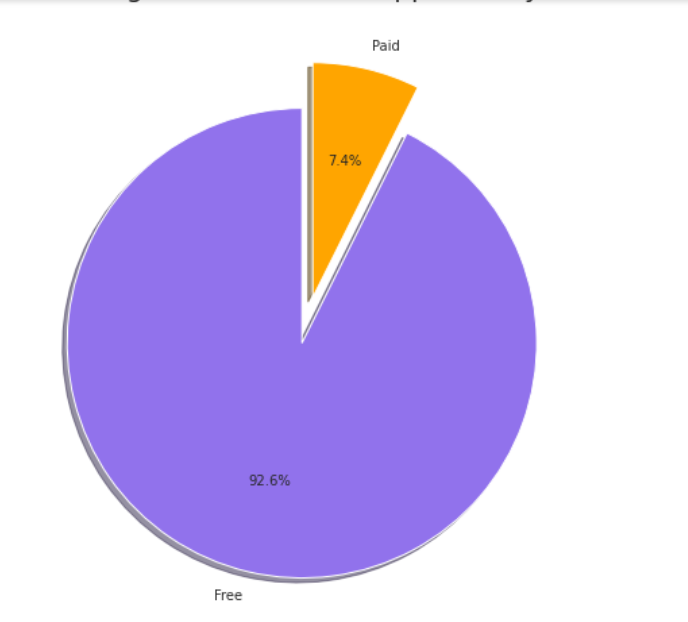


* App installed according to genres(Number of Installed applications for each genres)

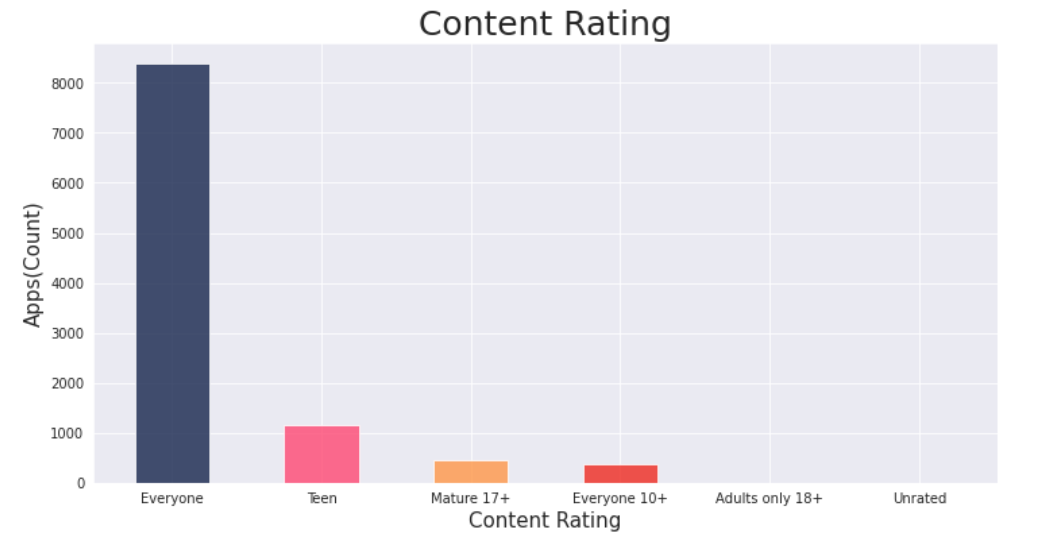


We can see from the above two plots: Maximum number of apps present in google play store comes under Tools, Entertainment and Education Genres but as per the installation and requirement in the market plot, scenario is not the same. Maximum installed apps come under Communication, Productivity and Social Genres

Percentage of Free Vs Paid Apps in Play store- It has been analyzing that free App has domination in app market



* Content Rating- Content ratings are used to describe the minimum maturity level of content in apps. However, content ratings don't tell you whether an app is designed for users of a specific age. Ratings are typically based on a number of factors, including sexual content, violence, drugs, gambling, and profane language.



**6. Contributions:**

1. Data wrangling
2. Treating Null values
3. Average reviews across each category
4. Top app category in play store(Count of apps in each category)
5. App installed according to category(Number of Installed applications for each category)
6. Top app genres in play store(Count of apps in each genres)
7. App installed according to genres(Number of Installed applications for each genres)
8. Percentage of Free Vs Paid Apps in Play store
9. Distribution of Apps by Rating, Size and price
10. Top Expensive and Earning App in Play Store
11. Average Installation of app across the year
12. What are the Top five installed apps in any category
13. Merging App review dataframes on apps
14. Distribution of Sentiment Subjectivity
15. Percentage of Review Sentiments
16. Distribution of Sentiment Polarity
17. Polarity Vs Subjectivity(Does sentiment\_subjectivity proportional to sentiment\_polarity)
18. Which are the Apps with highest number of reviews

**7. Programing Language:**

We have used python programing Language and used below library for EDA

Numpy

Pandas

seaborn

WordCloud

matplotlib

warnings