**Google Play Store Data Analysis**

**Pankaj Verma**

**Data science trainees**

**Alma Better, Bangalore**

**Abstract:**

The use of applications is part of people daily lives for various activities. Applications are available through Google Play either free of charge or at a cost. Android is expanding as an operating system.There are more than 5 million Apps found on Google play store.

Our experiment can help to understand which category and area are more popular among the designers and developers on the play store. And also find out the factors that affect someone’s decision to download an app.

**Keywords:-** Mobile Apps, Reviews,Ratings, Sentiment Analysis, Colab Notebook,

**2. Introduction**

The use of applications is part of people daily lives for various activities. Applications are available through We focus on analyzing Google Play Store, the largest Android app store that provides a wide collection of data on features (ratings, reviews, type, install and number of downloads,) and descriptions related to application functionality. There are more than 5 million Apps found on Google play store.

Our experiment can help to understand which category and area are more popular among the designers and developers on google play

store. And also find out the factors that affect someone’s decision to download an app.

**1.Problem Statement**

Data provided from the play store Apps, 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. Explore and analyze the data to discover key factors responsible for app engagement and success.

It provides access to content on Google Play, including Apps, books, magazines, music, movies, and television programs

1. WHAT IS THE VALUE DISTIRBUTION OF RATING COLUMN?
2. HIGHEST NO. OF APPLICATION PER CATEGORY?
3. WHAT IS THE CORELATION  B/W CATEGORY AND TYPE?
4. HOW MANY INSTALLATIONS  PER CATEGORY?

## **3. DESCRIPTION OF DATA SET:**

1. **Play Store Data Set** (App, Category, Rating, Review, Size, Install, Type, current rating, genres, Last update, Current Var, Android Var)

**Shape:**(10841, 13)

1. **User Review Data Set** (App, Sentiment, Sentiment Polarity, Sentiment Subjectivity)

**Shape:**(64295, 5)

By diagnosing the data frame, we know that:

* There are 13 columns of properties with 10841 rows of data.
* Column 'Reviews', 'Size', 'Installs' and 'Price' are in the type of 'object'
* Values of column 'Size' are strings representing size in 'M' as Megabytes, 'k' as kilobytes and also 'Varies with devices'.
* Values of column 'Installs' are strings representing install amount with symbols such as ',' and '+'.
* Values of column 'Price' are strings representing price with symbol '$'.

**4. DATA CLEANING:**

* **Exploratory Data Analysis**

After loading the dataset we performed different methods to process data. By this we get the only that data which is free from all null

values, duplicity, spec. characters & other unwanted data.

**Null values Treatment**

* Our dataset contains a large number of null values which might tend to disturb our accuracy hence we dropped or replaced null values by the mean value of that particular columns, at the beginning of our project in order to get a better result.

**String & Spec. Characters Treatment**

* In size column, values are defined in ‘String’ of ‘M’, ‘N’ and also 'Varies with devices'. To perform different methods, we have converted them in ‘int’ by separating ‘M’, ‘N’ from the ‘int’ values and replacing ‘Varies with devices' with mean values.
* Column 'Installs' and are strings representing install amount with symbols such as ',' and '+'. So, to using the data from these column, just removed spec. characters and convert the values into ‘int’ or ‘float’.

**4.****Rating Distribution (EDA)**

From this distribution plotting, it implies that most of the apps in the Play Store are having rating higher than 4.7 and the range lies b/w 4.2 to 4.7.

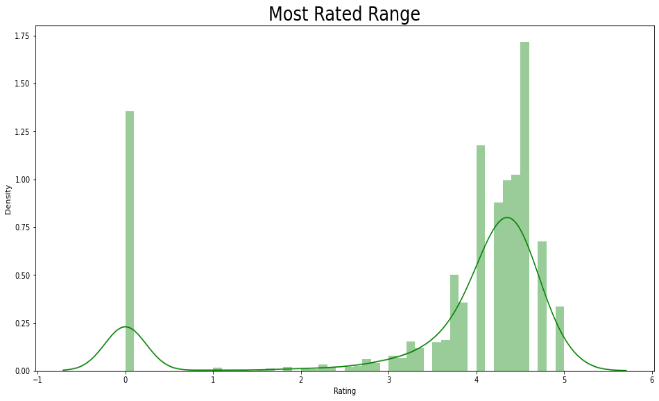


Fig.1 **Distribution of Rating**

**5. Highest Numbers of Application per Category**

The Bar plot shown below, representing the no of application per category, by which it’s easily visible the under **Family** category most numbers of application are available, then **Game and so on….**

Related data also given above the Fig.

FAMILY 1972

GAME 1144

TOOLS 843

MEDICAL 463

BUSINESS 460

PRODUCTIVITY 424

PERSONALIZATION 392

COMMUNICATION 387

SPORTS 384

LIFESTYLE 382

FINANCE 366

HEALTH\_AND\_FITNESS 341

PHOTOGRAPHY 335

SOCIAL 295

NEWS\_AND\_MAGAZINES 283

SHOPPING 260

TRAVEL\_AND\_LOCAL 258

DATING 234

BOOKS\_AND\_REFERENCE 231

VIDEO\_PLAYERS 175

EDUCATION 156

ENTERTAINMENT 149

MAPS\_AND\_NAVIGATION 137

FOOD\_AND\_DRINK 127

HOUSE\_AND\_HOME 88

AUTO\_AND\_VEHICLES 85

LIBRARIES\_AND\_DEMO 85

WEATHER 82

ART\_AND\_DESIGN 65

EVENTS 64

PARENTING 60

COMICS 60

BEAUTY 53

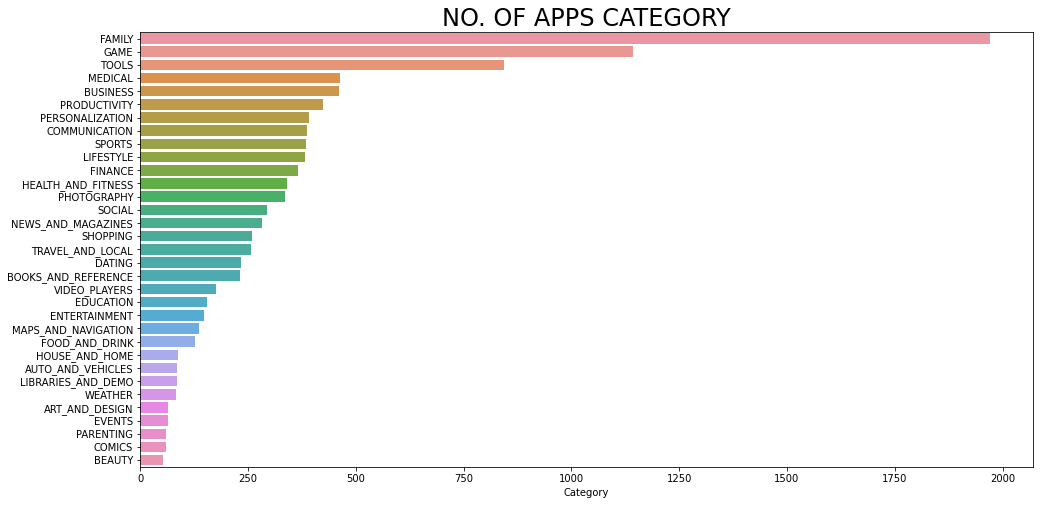
****

Fig.2**Application per Category**

## **6. Correlation Between Category and Type**

From the below bar plot showing the comparison of Paid and Free Apps. in each category.

Related data also given above the Fig.

Category Type

ART\_AND\_DESIGN Free 61

Paid 3

AUTO\_AND\_VEHICLES Free 82

Paid 3

BEAUTY Free 53

…………………………………………………………………………… ………………....

TRAVEL\_AND\_LOCAL Paid 12

VIDEO\_PLAYERS Free 171

Paid 4

WEATHER Free 74

Paid 8

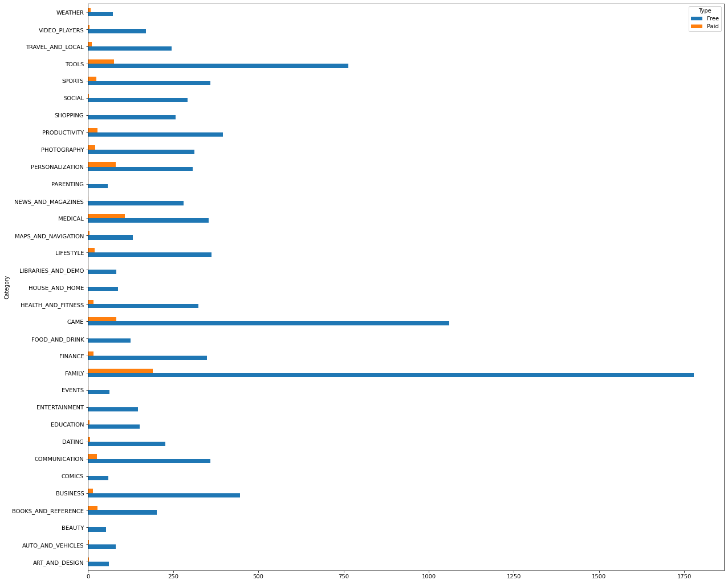
****

Fig.3 **Paid and free Apps. per Categories**

## **7. MANY INSTALLATIONS PER CATEGORY**

From the below bar plot **MEDICAL** have the highestnumber of installed applications followed by the **EVENT & BEAUTY.**

Related data also given above the Fig.

MEDICAL 1.150269e+05

EVENTS 2.495806e+05

BEAUTY 5.131519e+05

PARENTING 5.253518e+05

AUTO\_AND\_VEHICLES 6.250613e+05

LIBRARIES\_AND\_DEMO 7.411284e+05

COMICS 9.347692e+05

DATING 1.129533e+06

LIFESTYLE 1.407444e+06

ART\_AND\_DESIGN 1.912894e+06

HOUSE\_AND\_HOME 1.917187e+06

FOOD\_AND\_DRINK 2.156683e+06

BUSINESS 2.178076e+06

FINANCE 2.395215e+06

SPORTS 4.560350e+06

HEALTH\_AND\_FITNESS 4.642441e+06

WEATHER 5.196348e+06

FAMILY 5.201959e+06

MAPS\_AND\_NAVIGATION 5.286729e+06

EDUCATION 5.586231e+06

PERSONALIZATION 5.932385e+06

BOOKS\_AND\_REFERENCE 8.318050e+06

SHOPPING 1.249173e+07

TOOLS 1.358573e+07

ENTERTAINMENT 1.925611e+07

NEWS\_AND\_MAGAZINES 2.648876e+07

TRAVEL\_AND\_LOCAL 2.662359e+07

PHOTOGRAPHY 3.011417e+07

GAME 3.066960e+07

PRODUCTIVITY 3.343418e+07

VIDEO\_PLAYERS 3.555430e+07

SOCIAL 4.769447e+07

COMMUNICATION 8.435989e+07

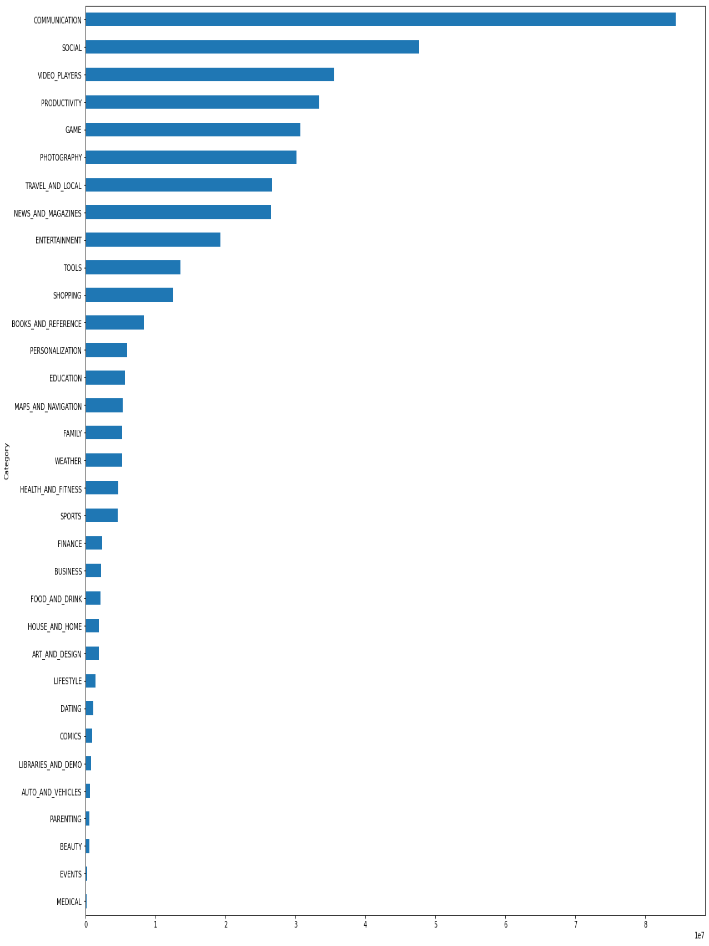
****

Fig.4 **Apps. Installed Per Category**

**8.Conclusion**

The Google Play Store Apps report provides some useful details regarding the trending of the apps in the play store. As per the graph’s visualizations shown above, most of the trending apps (in terms of users' installs) are from the categories like FAMILY, GAME, MEDICAL AND EVENTS AND BEAUTY.

As per some observations category FAMILY is leading in nos. of applications available to the end users and GAME on the second position on the same plot while BEAUTY, COMICS and EVENTS are at the bottom. And also the highest no of paid and free applications are available under the FAMILY category as it has max. nos. of application followed by GAME as above. But, MEDICAL category have hug lead in installations followed by EVENTS and BEAUTY which can be an opportunity for the designers and developers to work in that areas.

### Some important point: -

* **MOST NUMBERS OF APPLICATIONS HAVE RATING RANGE B/W APPROX 4.2 TO 4.7**
* **HIGHEST NO. OF APPLICATION COMES UNDER THE FAMILY FOLLOWED BY GAME....SO ON**.
* **BY COMPAIRING THE PAID AND FREE APPLICATION WE CAN CONCLUDE THAT MORE FREE APPLICATIONS ARE AVAILABLE UNDER ALL THE CATEGORIES.**
* **ALTHOUGH MORE NO. OF APPLICATIONS AVAILABLE IN FAMILY CATEGORY, BUT MOST NOS. OF APPLICATIONS INSTALLES ARE FROM MEDICAL,EVENTS AND BEAUTY.**

**References-**

1. Almabetter Tutorials
2. Stackoverflow
3. GeeksforGeeks
4. Youtubes