GOOGLE PLAY STORE APP REVIEW ANALYSIS

RAM MANOHAR THAKUR,
SAYAN BISWAS,
SNEHAL RAGIT,
SHRADHA LAKHADIVE
DATA SCIENCE TRAINEE
AlmaBetter, Bangalore

1. ABSTRACT:

Google Play Store is a digital distribution service operated by Google. Google Play Store mainly works for Android devices, from where Android users can download various apps as per their needs on their phones That is, it acts as an application store. Mainly here the apps are free to download on their phones. As various developers are publishing their apps on Google Play Store on one hand and users are using them on the other hand, From which a financial side of the developers is also getting prosperous. Google arranges app rating for users which basically allows users to tell how they feel about an app after using it but it is completely personal.

Keywords- Google Play Store Apps, Ratings Prediction, Exploratory Data Analysis, Feature engineering, Data Mining

2. INTRODUCTION:

Currently there are at least 25 lakh apps on play store. Today we will show how we analyzed the apps obtained from the dataset through the Python library on this project. The data set has 13 columns and 10841 rows. After cleaning the data, we analyze which apps are most popular and why? Which apps are the least popular and why? What is the problem with the apps that are least popular and what should be done to bring them back to popularity.

PROBLEM STATEMENT :

1. How many apps are free or paid?
2. Which apps are at the top of their popularity?
3. Which apps category have been downloaded the most?
4. The most popular category among the categories that appear on the Play Store dataset?
5. What percentage of apps are free?
6. The name of the most downloaded apps among the paid apps?
7. How many free apps have been downloaded over a million?
8. Which apps received the most positive and the least negative reviews?

We got the above data from Alma Better. This data contains a description of different types of apps in each row. This method is considered to be one of the most important methods of data science. Through this method we can perform other tasks besides making basic business decisions. But our primary objective should be – Cleaning up these data so that we can present them in front of the presenter is a visualization of the complete data.

3. GOOGLE PLAY STORE APP REVIEW ANALYSIS :

In today's era mobile applications are seen playing an important role in every people's life as per their needs. People are directly and indirectly involved with mobile applications. Developers are earning from this while users are using it to complete their needs.

Google Play Store is the world's largest app store in the present era. A designer from all over the world with huge challenges from everywhere realizes that he is moving in the right direction. Designers are realizing that this is a very large market and to survive here they have to create new applications day after day and they also have to provide proper maintenance. There are currently 2.7 million apps on the Play Store which are closely related to people's lives.

as an example: As a developer if I create an app and publishing on the Play Store doesn't stop my work. Google Play Store has a separate section of ratings for different apps to see if users can use that app properly or have problems. In rating section users can express their opinion after using that app.

4. DESCRIPTION OF DATASET :

4.1 PLAY STORE DATASET:

The google play store app review analysis data set is provided to us by Alma Better. From this data set a user can experiment about various topics including android market. Through testing, on one hand, you will get an idea about various apps, on the other hand, you will get an idea about which apps will be more popular in the future. It had 10841 rows and 13 columns . so let's analyze it .

This dataset holding following columns:

- **App:** This section contains the name of apps.
- **Category:** The category indicates which type the app is. It is seen here that there are 33 unique values
- **Rating:** This column shows the average calculations of users by taking feedback from them. The value of opinion is between zero and five.
- **Review:** This column carries out how many users voted on that particular app.
- **Size :** How much Android device space a user needs to install the app is the work of the size column.
- **Installs :** It keeps track of how many people have downloaded the apps. However, this is not exact data.
- **Type :** This column tells you whether you have to pay to download this app or it can be used for free.
- **Price :** The prize column indicates how much money to pay to download this app if the type column tells you it's not free If free, its value is zero.
- **Content Rating :** This section indicates that which audience or age group's people this app targeted.
- **Genres :** The column of the genre indicates which sub-category the app belongs to
- **Last Updated :** This column indicates when was last updated for users to use it better after it was published on a PlayStore.
- **Current Ver :** Here is a description of which version this app can currently run on.
- **Android Ver :** This column determines how much the user's Android version needs to be to use the app.

4.2 USER REVIEW DATASET:

This dataset contains 64295 rows & 5 columns. let's describe 5 different columns.

- **App:** It contains the name of the apps.
- **Translated_Review :** This section contains an English translation of the opinions expressed by specific users of a particular app after using the app.
- **Sentiment :** The sentiment column carries exactly what their opinion is i.e. positive or negative after users use particular app .
- **Sentiment_Polarity :** This column gives the polarity of the review whose limitation is (1,-1) i.e. -1 is completely negative and 1 is completely positive.
- **Sentiment_Subjectivity :** Subjective sentences generally refer to **personal opinion, emotion or judgment** whereas objective refers to factual information. Subjectivity is

also a float which lies in the range of [0,1]

5. DATA CLEANING & PREPARATION :

The data must be cleaned up before it can be prepared. Because it is seen that almost all the data in the real world is unclean and unusable. There is a lot of data that has no value, it is not possible to work with all this data. So first we will clean them up and then start the work of data preparation.

- STAGE 1 - We have run df.head() & df .tail() for getting data how it look is. We also run df.collumn() & df.shape() to get full columns data & get how many row & collumns present here.

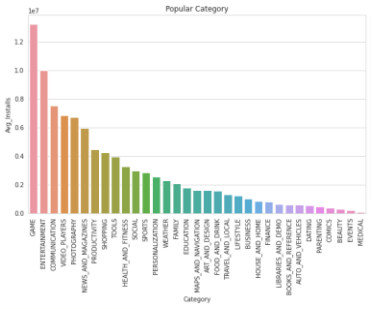
We have pass a code df.info() [df=play store dataset] which shows us full information about columns.

We also check duplicate data & found it,so we have drop it from dataset & check the shape of dataset.we also trying to find out missing value in column , fill it 'FREE' to use it as usable & drop unusable columns(android ver & cureent ver). Then we have showing plot distribution.

- STAGE 2- We have Columns like Reviews, Size, Installs, Price are of object type so we can change their data type to integer/float.We can see the size of column, which should be numeric , is of the data type 'object', it also has characters 'k' (kilobyte) and 'M' (megabyte) in the values.So we have
removing M and replacing K with e-3. Some values also found '+' , ',' '$' sign in them, which will be removed.after we looked at the 'last updated' column it contains the date on which the app is updated/launched last time. It is of object type so we have to convert date in the date-time format. Now we have got final dataset to explore it.

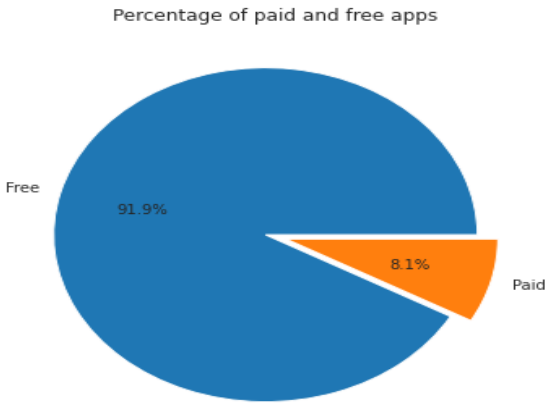
6. Which category app present most & which category app install most:

We found Family category most number of apps & least number of apps found in Comics ,Beauty and a lot of category. We also find out 'Games' category's apps install most of times & 'Medical' category least of times.



7. FREE APP v PAID APP

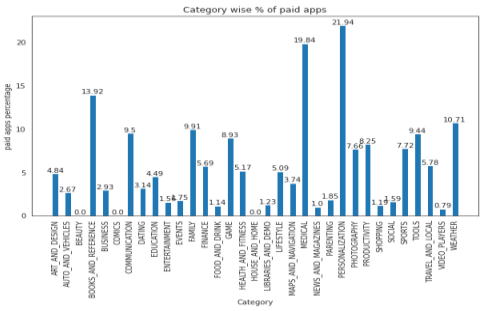
We have create a pie chart to find out Free and Paid users interface.From that we have got 91.9% apps are free & 8.1% apps are paid.so we can say that maximum apps of playstore are free & user use it.



8. PAID CATEGORY SECTION

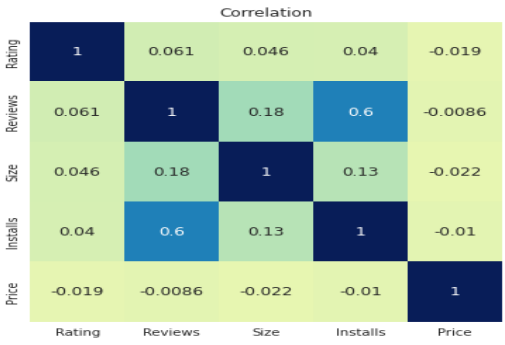
We look into paid app section & got

'Personalization' category got high user & 'Beauty', 'Comics' got zero user.so we can say least category should developed.



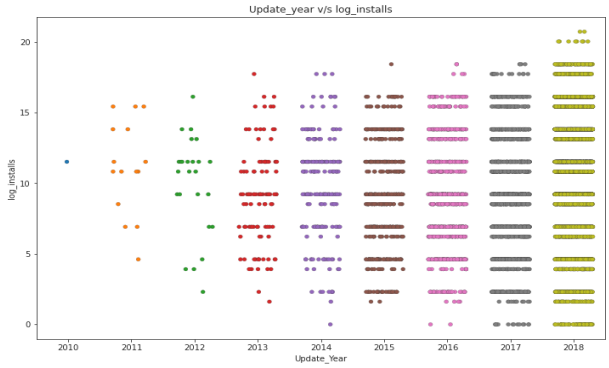
9. CORRELATION HEATMAP:

We have created correlation heatmap to find out more clearly about this dataset. It is a powerful tool for summarizing a huge dataset for visualizing and identifying patterns in given data.We say about pricing category and clearly said that with pricing every category gives negative impact so we have to improve it .



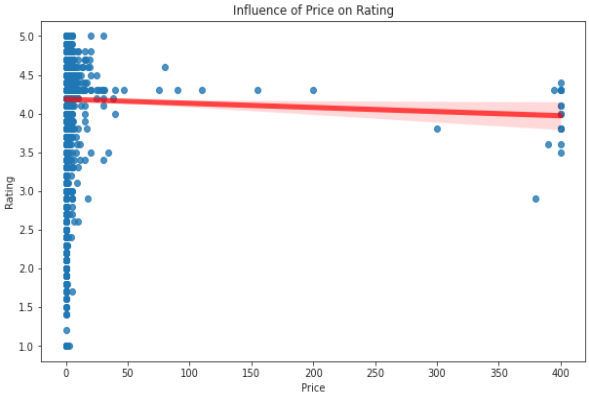
10. UPDATE YEAR & DETAILS

We have say from this data that if app isn't update or more attractive to user it would be go down.



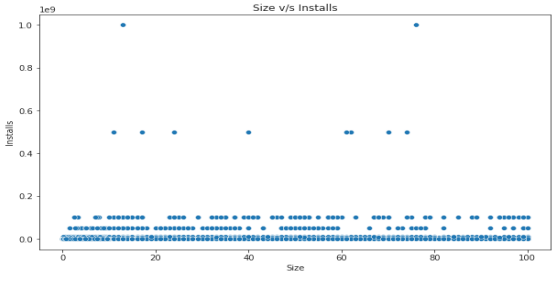
11. INFLUENCE OF PRICE ON RATING

We have create a scatterplot & found that maximum apps cost are below $100 . There is negative relation between price & rating. Rating decreases with incresing price.



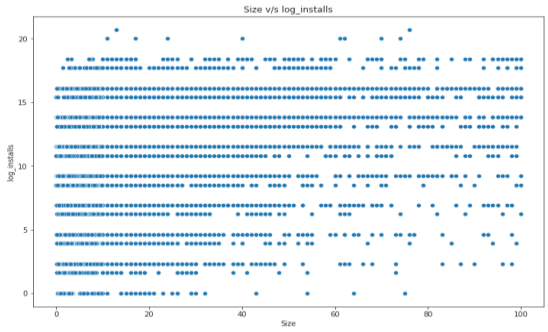
12. SIZE v INSTALLS

When we created another scatterplot to find out app installs according to size we got that more than 80mb apps install less. So we can say people like less size apps.



13. LOG INSTALLS & SCATTER PLOT SIZE

To install log we have got more clear image for installation & size .



14. DISPERSION SYSTEM

If we analyze the data set too fine then the representation of the average data may not be accurate because it will change if there are large differences between the data.

15. CORRELATION AMONG VARIABLES

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables. A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.

16. GRAPHICAL REPRESENTATION

It is very important to represent this data set through graph map diagram. Because of how essential the results obtained from the data set are to business stockholders, this is the main goal of our project. Most of the graphical analysis we have done includes bar charts, tables charts, scatterplot, histograms, etc.

- **CONCLUSION**- Through the analysis of exploratory data, we have tried to observe that which apps appear on the Play Store have attracted people more and which apps have attracted the crowd the most. The project also shows the reasons for the apps that could not attract the crowd. We can explore more:

a) First of all we can say that 91.9% apps are free & 8.1% apps are paid.

b) Most apps are available in Family category.

c) Most favorite category of users is Game.

d) The most disliked category by users is medical.

e) Personalization is the most popular category among paid apps.

f) Free apps have an average rating of 4.18

g) Paid apps have an average rating of 4.26

h) The number of free apps downloaded more than 1 billion is 20.

i) People prefer small size apps more.

j) Most of the paid apps cost less than $100

k) User mostly liked updated apps.

l) Paid apps have lower negative sentiment polarity than free apps.

REFERENCES:

- Kaggle
- Towards Data Science (TDS)
- Pew Research Center
- GreeksforGreeks
- Python Libraries
- GitHub