Available online at : http://www.advancedscientificjournal.com
http://www.krishmapublication.com
IJMASRI, Vol. 3, issue 5, pp. 962- 972, May -2023
https://doi.org/10.53633/ijmasri

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY ADVANCED SCIENTIFIC RESEARCH AND INNOVATION (IJMASRI)

ISSN: 2582-9130 **IBI IMPACT FACTOR 1.5** **DOI: 10.53633/IJMASRI**

RESEARCH ARTICLE

EXPLORATORY DATA ANALYSIS AND SUCCESS PREDICTION OF GOOGLE PLAY STORE APPS

Samarth Gaba[1] and Sachin Garg[2]
[1, 2] Department of Information Technology, Maharaja Agrasen Institute of Technology, Rohini, Delhi
Email: samarthgaba27@gmail.com, sachingarg@mait.ac.in

Abstract

In a fiercely competitive worldwide market, the Google Play Store is constantly being updated with a large number of new apps developed either alone or by teams. Despite the fact that the majority of applications available on the Play Store are free, it is still unclear how in-app purchases, adverts, and subscriptions are paid for. As a consequence, user feedback and installation count rather than cash produced are often used to gauge an application's performance. App ratings, which are voluntarily submitted user input, are a key criterion for app assessment. However, owing to inadequate or absent votes, these ratings often exhibit bias. Additionally, the differences between user reviews and numerical ratings are evident. The purpose of this study is to use machine learning techniques to forecast app ratings on the Google Play Store. The dataset utilised in this research was gathered from Kaggle and includes details on a number of characteristics, such as an app's price, user feedback, and app ratings, as well as whether it is a free or paid download. We seek to identify links between these traits via data analysis and prediction utilising machine learning methods.

Keywords: Google Play Store Apps, Ratings Prediction, Exploratory Data Analysis, Machine Learning.

Problem statement

The Google Play Store served as the source of the dataset for this study. A unique app is represented by each row in the dataset, which also includes numerous items pertaining to that app. Our study focuses on doing this dataset's exploratory data analysis (EDA), a critical phase of the data science cycle. Making early business choices and preparing the data for later modelling using machine learning algorithms are two of the many functions of EDA. Our study' main goal is to organise and clean the data while spotting patterns that might provide early indications of the possibility that freshly released applications will be successful. By using EDA, we hope to get some early insights that will help us determine the likelihood that an app would be successful

Introduction

This research study focuses on providing comprehensive information on machine learning models and frameworks because it recognises the critical role that machine learning plays in solving a variety of issues. Machine learning has a wide range of uses in several fields and has a great deal of room for development.

It is projected that machine learning will develop the best theories to account for its performance in the future. Additionally, given the enormous quantity of data that is accessible on a worldwide scale but may not always be labelled, unsupervised learning skills are anticipated to advance. Additionally, it is anticipated that neural.

Analysis of google play store and user reviews

In the present environment, smartphone applications are very important to people's life. The growth of the market for mobile applications has had a big influence on digital technology. As a result, the worldwide mobile app business has seen significant revenue growth as a result of the constantly growing mobile app market. Because of the fierce competition throughout the globe, app developers must make sure they are headed in the correct route. Developers need to identify solutions to protect their existing position in order to keep their income and market share. The biggest application marketplace is widely acknowledged to be the Google Play Store. It's interesting to note that despite having more than twice as many downloads as the Apple App Store, the Play Store only brings in half as much money. As a result, information was scraped from the Play Store to carry out our investigation.

Mobile applications are becoming an essential part of our lives because to the explosive proliferation of smartphones. However, the rapid flood of new apps makes it difficult to keep up with the ever-expanding app market and remain educated about every app. There are presently over 0.675 million Android applications accessible on the Google Play App Store, up from half a million in the Android market as of September 2011. Users have a broad selection of alternatives to pick from as a result of this enormous number of applications. We think that, especially for commercial apps, consumers of mobile apps largely depend on online app evaluations. It is challenging for prospective consumers to go through all the written reviews and ratings in order to make a wise choice. Similar difficulties arise when app performance is improved merely based on overall ratings, and app developers would significantly benefit from comprehending the voluminous written feedback.

We publish our Android applications on the Play Store as Android app developers. From a commercial standpoint, it is critical for us to understand if customers are satisfied with our app or having any problems. Users may submit ratings and reviews for individual apps in the Ratings & Reviews area of the Play Store. However, the procedure for rating and evaluating an app might be laborious, forcing users to either launch the Play Store app via a URI link or travel to the Play Store to provide comments without interfering with their current app activity. Because the present flow pushes users to move control to the Play Store app, we want to make sure that users may submit feedback without leaving our application.

Playstore data from google

Almabetter, the world's biggest community for data scientists to explore, analyse, and exchange data, provided the dataset utilised in this study. This dataset focuses on data that was web scraped from 10,000 Google Play Store apps, enabling a thorough examination of the Android market. The information includes a number of app use categories, including music, camera, and more. Users can forecast if a certain application will get a better or lower rating level thanks to it. Additionally, this information might be a useful tool for upcoming app ideas. Because internet data is often changed, using an offline dataset provides accurate measurements. The researcher intends to use this information to use Hive to analyse numerous aspects, including ratings, the cost of an app, and more. Additionally, forecasts will be generated about attributes like user ratings and reviews.

The data set contains the following columns:

- **App:** This Column contains the name of the app

- **Rating:** The average rating given to the app by users on the Play Store. Ratings range from 0 to 5.
- **Installs:** The approximate number of times the app has been downloaded from the Play Store.
- **Category:** The category to which the app belongs. There are 33 unique values in this column
- **Reviews:** The number of user reviews received for the app.
- **Size:** The size of the app in terms of memory space it occupies on the device after installation.
- **Type:** Indicates whether the app is free or paid.
- **Price:** For paid apps, this column contains the price of the app; for free apps, it contains the value 0.
- **Genre:** Specifies the genre or sub-division to which the app belongs.
- **Last updated:** The date of the last update for the app on the Play Store.
- **Content Rating:** Indicates the targeted audience and age group for the app.
- **Android version:** Indicates the minimum version of the Android OS required to install the app
- **Current version:** Information about the current version of the app available on the Play Store.

User review and data set

The user reviews data frame consists of 64,295 rows and 5 columns, each identified as follows:

- **App:** Contains the name of the app, accompanied by an optional short description.
- **Sentiment:** Indicates the attitude or emotion expressed by the reviewer. It can be categorized as 'Positive', 'Negative', or 'Neutral'.
- **Translated Review:** Provides the English translation of the review left by the app user.
- **Sentiment Polarity:** Represents the polarity of the review, ranging from -1 to 1. A value of 1 indicates a 'Positive statement', while -1 represents a 'Negative statement'.
- **Sentiment Subjectivity:** Measures the degree of subjectivity in the reviewer's opinion, ranging from 0 to 1. A higher subjectivity value suggests that the reviewer's opinion aligns closely with the general public's opinion, while a lower subjectivity value indicates that the review is more factual and less influenced by personal opinion.

Python

Data scientists like Python as a programming language because of its large selection of built-in library functions and its active community. Python provides a large selection of tools and resources for data analysis and machine learning applications, with over 70,000 libraries readily accessible. Python is a popular option for data scientists because of how simple it is to learn compared to other computer languages. Data scientists often use Python because of its extensive ecosystem of libraries. One of them, Pandas, is a popular open-source library created with data manipulation and analysis in mind. It offers strong capabilities for working with structured data, which makes it crucial for data scientists. Furthermore, the built-in libraries in Python come in very handy for visualising data, whether it be in the form of scatterplots, heat maps, graphs, or three-dimensional data. For data scientists, these libraries make it easier to visualise data, facilitating data exploration and analysis.

Data preparation and cleaning

Preprocessing is essential for converting raw data into a format that is more appealing. It fulfils crucial functions including assuring data dependability and completeness as well as evaluating value consistency. Preprocessing is essential because imperfect real-world data are often present. Incomplete data, which contains missing numbers, and noise, which describes outliers or mistakes in the data, might be included in this. The quality of the dataset is increased by the use of preprocessing methods, such as controlling outliers, resolving missing values, and maintaining data consistency, resulting in more dependable and accurate analysis.

It is essential to preprocess and clean the raw data to make it appropriate for analysis before beginning the Exploratory Data Analysis (EDA)

process. The first dataset could show irregularities or have missing numbers, which call for cautious management. We may successfully prepare the data for further analysis by extensively investigating and cleaning it, so assuring the correctness and dependability of our results throughout the EDA step.

- **Step1**: To learn more about the columns in the play store dataset, we construct a method called play_store_info(). It shows details such as each attribute's data type, the number of unique values, the number of non-null values, and the proportion of null values in each column.

- **Step2**: Next, we go on to the 'Type' column where we see that it only contains one null value. We determine after further study that it relates to a no-cost app. We use the pandas library's fillna method to handle this missing value and replace it with the right value.
- **Step 3**: We choose to remove the "Current Ver," "Android Ver," and "Last Updated" variables from our dataset as part of data pretreatment. Utilising the drop function offered by the pandas package, this is accomplished.
- **Step 4**: There are 1474 null entries in the 'Rating' column. We choose to utilise the "median" as an acceptable statistical indicator to impute the missing data because of the low variability in rating values and the prevalence of repeated values. We replace the null values with the estimated median value by computing the median using the median technique and then using the fillna function.
- **Step 5:** Examining the "Reviews" column, we see that, while being a numerical indication, it is now given the "object" data type. We will use the astype(int) method to change the data type to 'int' in order to fix issue.
- **Step 6:** The 'Size' column is now categorised as a 'object' data type, despite the fact that it should ideally include a number property. It includes symbols like "k" and "M," which stand for kilobytes and megabytes, respectively. The '+' symbol is also present in certain values. We will delete the "+" symbol, change "k" to "1000," and "M" to "1000000." The column will then be changed to the "int" data type.
- **Step 7:** In the 'Installs' column, the values include characters such as '+' and ',' which hinder our ability to convert the column into a numeric data type. We will remove these characters using the strip and replace functions.
- **Step 8:** The data type is presently "object," and entries in the "Price" column may include the dollar sign. To fix this, we'll first use the strip method to get rid of the dollar symbol ($), then change the column's data type to "int."

- **Step 9:** We use the drop_duplicates method to eliminate rows with duplicate values in the 'App' column in order to deal with duplicates in the 'App' column.
- **Step 10:** We develop the 'Ur_info' function, which offers details on the User review dataset. Data type, count of non-null values, count of null values, number of unique values, and the percentage of null values in each column are the five properties that this function provides for each column.
- **Step11:** We note that the columns in the user review dataset are "App," "Translated Review," "Sentiment," "Sentiment Polarity," and "Sentiment Subjectivity." There is a total of 26863 NaN values spread throughout these columns. We tackle this by removing rows that contain NaN values using the dropna method.

Analysing exploratory data

Any project involving data analysis or data science must include exploratory data analysis (EDA) since it enables us to learn from the information and derive relevant conclusions. EDA entails carefully going through the information to

look for patterns, anomalies (outliers), and to create hypotheses based on what we know about the data.

During EDA, we compute summary statistics for the dataset's numerical variables and provide visualisations to better comprehend the data. In this article, we will use an example dataset, the Python programming language, and the Pandas package to demonstrate the EDA process.

Paid vs Free

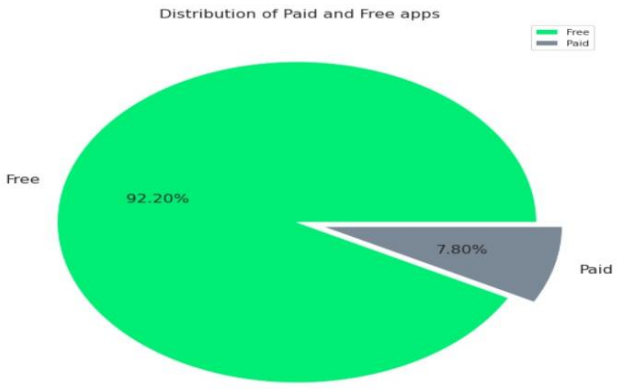

Fig -1: Paid Vs Free

We can see from the study of the Google Play Store dataset that a large percentage of the apps—roughly 92.2%—are offered without charge. However, 7.8% of the applications demand money in order to be downloaded and installed. This shows that the majority of applications available on the Google Play Store are free for consumers to download, while a lesser percentage of apps have a fee.

Rating

We noted the following findings in the supplied figure, which shows the distribution of app ratings:

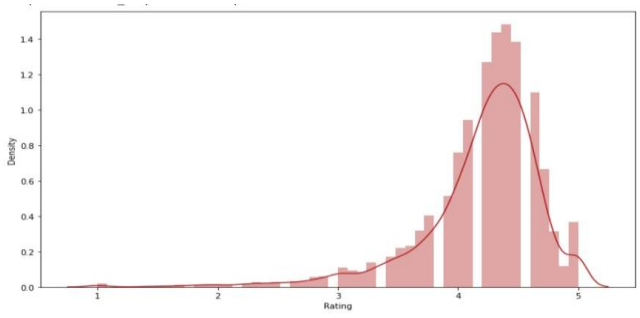

Fig -2: Distribution of App ratings

- Without include the NaN values, the mean of the average ratings is computed to be 4.2
- Without including NaN values, the median value of the ratings in the 'Rating' column is determined to be 4.3. According to this, 50% of the applications have an average rating higher than 4.3, while the other 50% have a rating lower than 4.3.
- We can see from the distplot visualisation that there are more applications with higher ratings since the distribution of ratings is left-skewed.
- It's crucial to remember that when a variable is skewed, the extreme values near the distribution's tails might have an impact on the mean. Therefore, for the majority of the rating values, the median offers a more indicative measurement of the centre tendency.
- These findings help us comprehend the general distribution and central tendency of the dataset's app ratings.

Size distribution of apps

The following findings may be drawn from the supplied plot, which depicts the distribution of app sizes on the Google Play store:

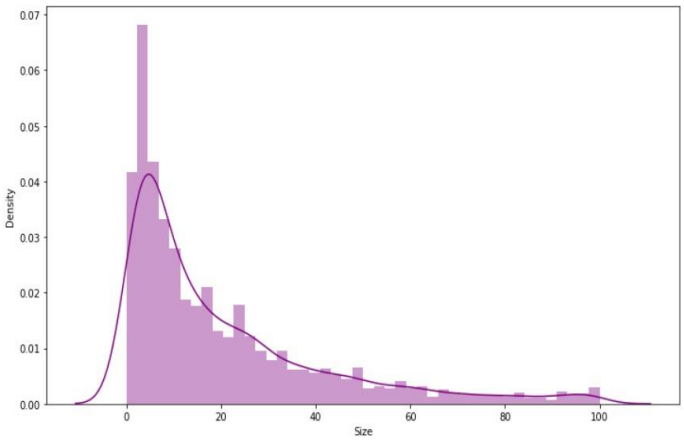

Fig -5: Distribution of App Size

- There are more applications with smaller sizes than bigger ones, according to the data in the Size column, which is biassed to the right.
- It's crucial to keep in mind that a substantial number of the items in this column contain the value "Varies with device." The statistics and

visualisations that come from replacing these numbers with a measure of central tendency, such as the mean or median, may be inaccurate. As a result, the dataset does not modify these values.

- We can better understand the distribution and variability of app sizes on the Google Play store by taking into account the skewness of the data and the existence of "Varies with device" entries.

Recent paid apps

A whopping 82% of the apps available on the Google Play store are made to be appropriate for all users, regardless of age. The remaining applications, which target to certain demographics, have explicit age limitations. target audiences.

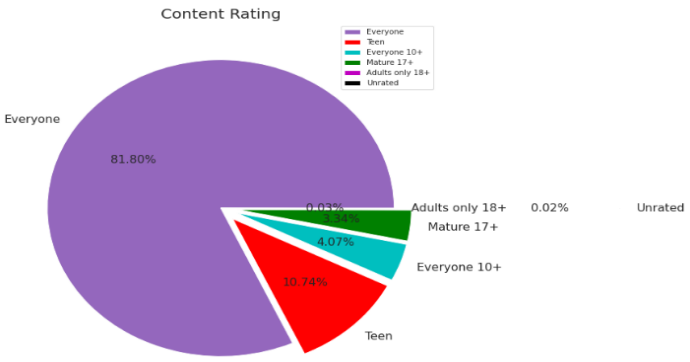

Fig -6: Content rating

Top playstore category

Figure 7 shows how applications are distributed across different categories on the Play Store. There are 33 distinct categories in the dataset.

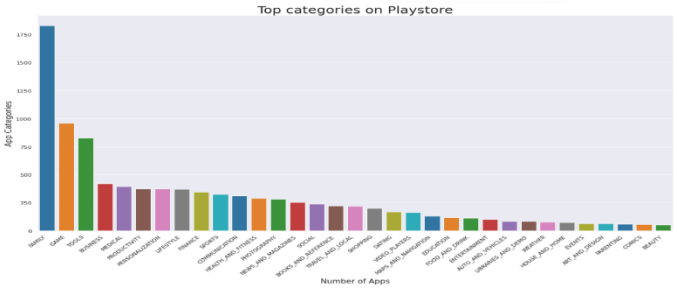

Fig -7: Top Playstore Categories

The visualisation makes it clear that the FAMILY and GAME categories have the most applications, while the EVENTS and BEAUTY categories contain the fewest.

No. of installs per category

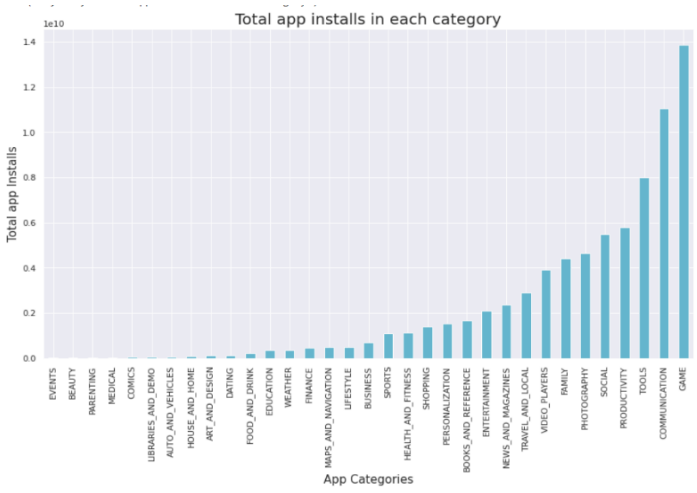

Fig -8: No. of Installs Per Category

By examining the data, it is possible to identify the app category with the greatest number of installations. When compared to other app categories, it has been shown that the Game, Communication, and Tools categories have the most installations. This demonstrates that certain categories are well-liked by consumers and draw plenty of installs.

Average App ratings

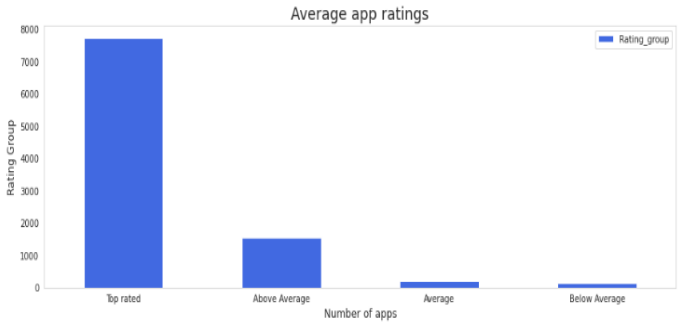

Fig -9: Average App Ratings

We may divide the ratings into intervals to portray them in a more understandable manner. We may group the ratings into the following categories using the provided standards:

- 4-5: Top rated
- 3-4: Above average
- 2-3: Average
- 1-2: Below average

We can offer a deeper picture of the distribution of ratings and categorise them based on their degree of quality or pleasure by dividing the ratings into these intervals.

Premium app with category

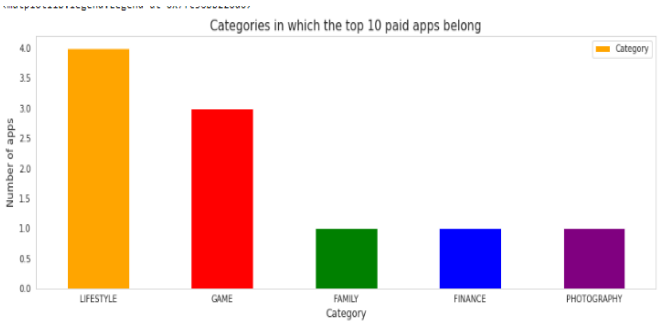

Fig -10: The best paid apps per category

The data reveals that the gaming and leisure categories have the greatest amount of premium applications. This indicates that creators of applications in these categories are more likely to recoup their investment by requesting a charge to acquire or use their software. It suggests that customers who are interested in lifestyle and gaming apps could have to pay for them rather than go for free alternatives.

Sentiment index of user reviews

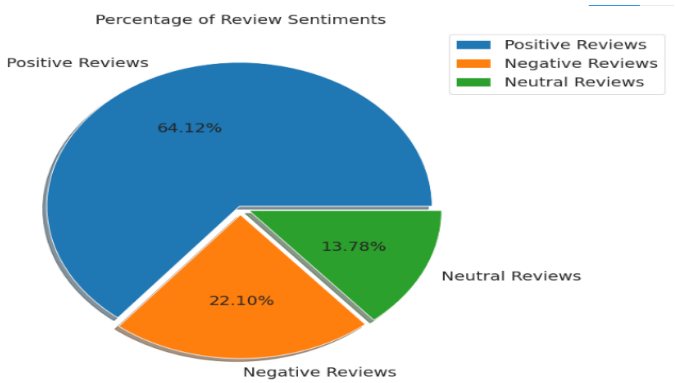

Fig -11: User Review Sentiments as Percentage

The pie chart makes it clear that a sizeable percentage of the applications offered on the Play Store have received positive reviews from customers. On the other hand, several of the applications have gotten unfavourable reviews.

Top 10 apps with good ratings

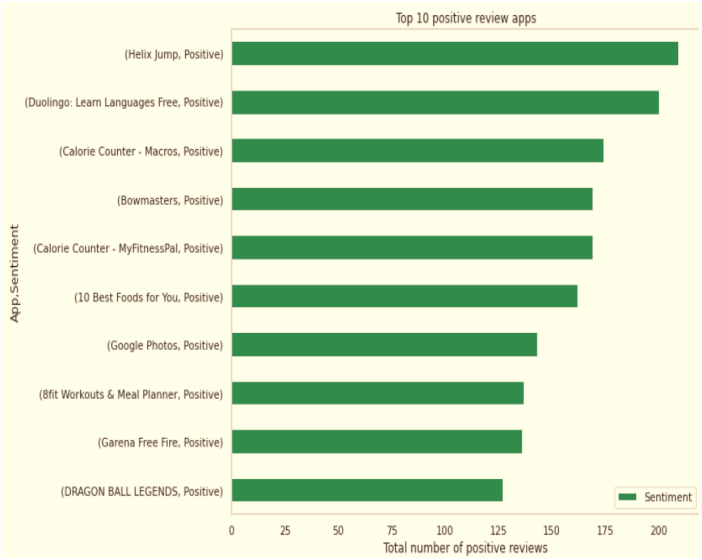

Fig -12: Top 10 Apps with Good Reviews

Top 10 apps for negative reviews

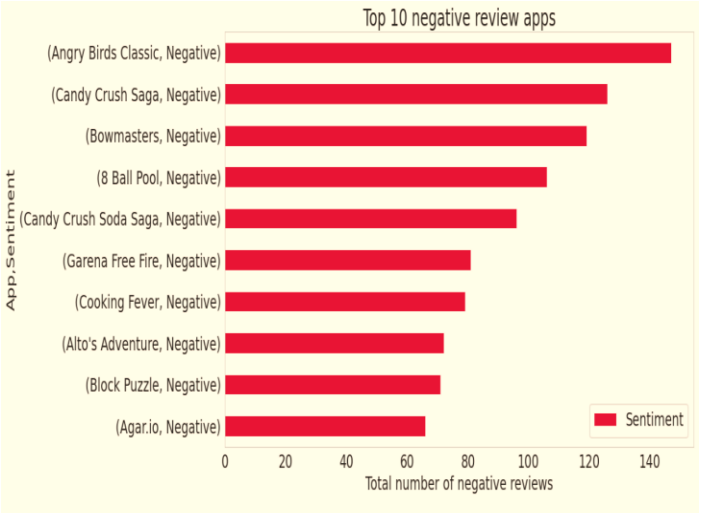

Fig -13: Top 10 Apps with Negative Reviews

Top 10 apps with negative reviews

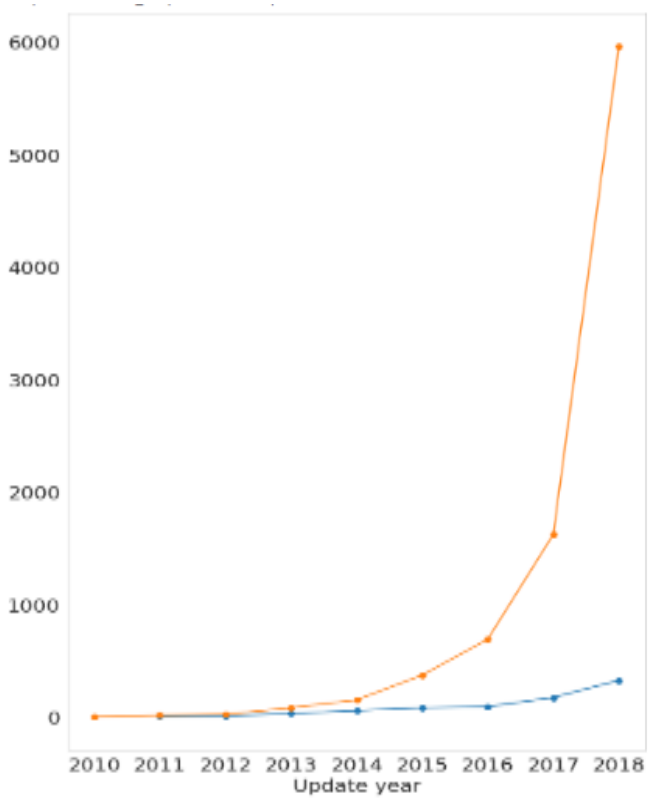

Fig -14: Top 10 Apps with Negative Reviews

Availability of apps updates

Several inferences may be made by examining the plotted data on applications that have been updated or added over time for both the free and premium categories. First off, the Play Store didn't provide any paid applications until 2011. But as time went on, more and more free applications were uploaded, outpacing the amount of commercial apps. It is clear from a comparison of the applications updated or introduced between 2011 and 2018 that the proportion of free apps rose dramatically from 80% to 96%, while the proportion of commercial apps shrank from 20% to 4%. Since free applications account for the vast majority of app updates and additions, this pattern reveals a significant preference among users for them.

Monthly distribution of app updates

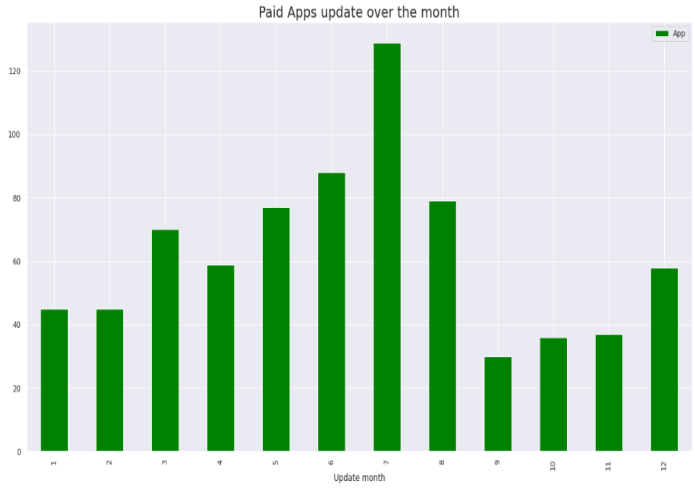

Fig -16: Monthly update for free apps

According to the statistics, around half of the applications in the sample had updates or additions in July. While 25% of the applications received upgrades or additions in August, the other 25% were spread out among the other months. It's interesting to note that the majority of premium applications got upgrades in July as well, reflecting the trend seen for free apps.

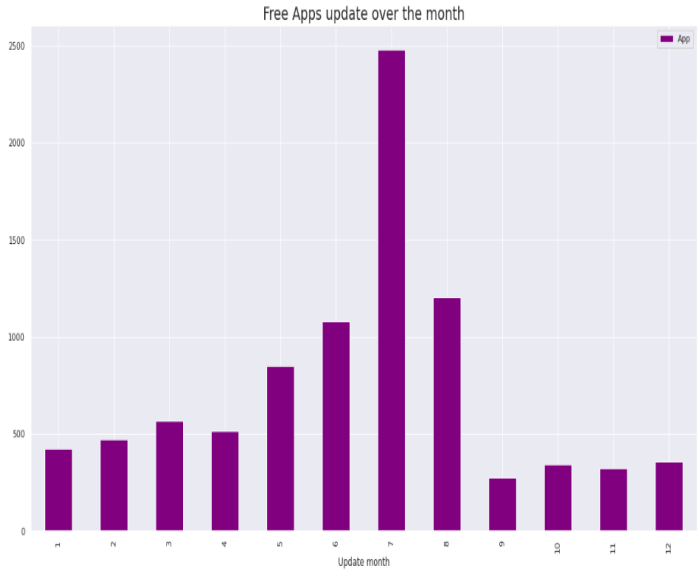

Fig -15: Paid Apps update over the month

Relationship between sentiment Polarity and sentiment subjectivity

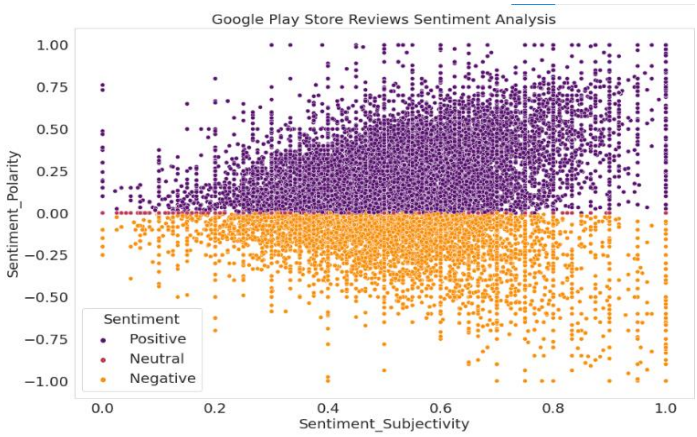

Fig -17: Analysis of the sentiment in Google Play Store Reviews

The scatter plot study suggests that sentiment subjectivity and sentiment polarity may not always have a straight proportional connection. Though sentiment subjectivity and sentiment polarity often display proportionate behaviour, this is not always the case. When the variation in the data is neither too great nor too low, this trend is valid.

Distribution of Subjectivity

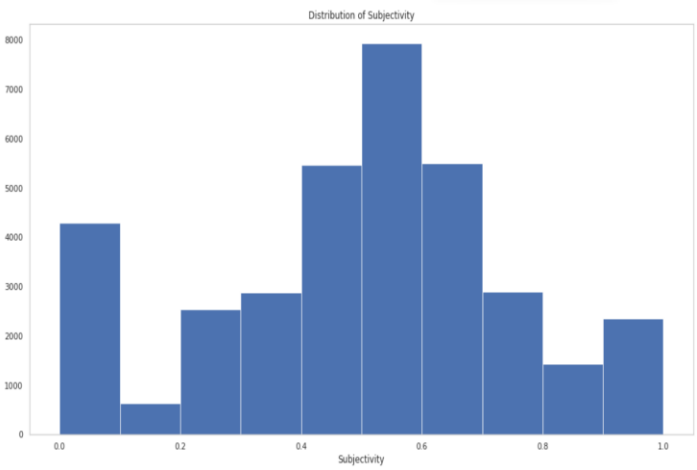

Fig -18: Subjectivity Distribution

0 - objective(fact), 1 - subjective(opinion)

The bulk of the evaluations lie between 0.4 and 0.7, according to the measurement of sentiment subjectivity. This shows that rather than just expressing the facts as they are, a significant portion of people provide ratings for programmes based on their own experiences and opinions.

Relationship between the dataset's many characteristics

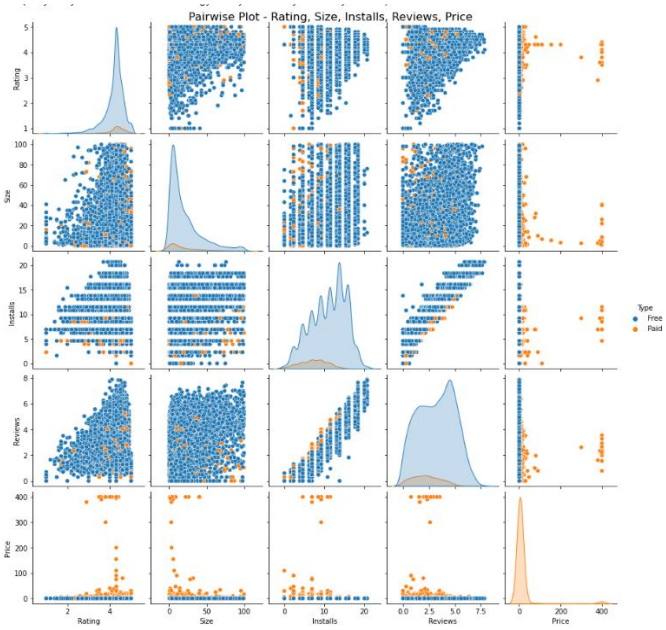

Fig -19: Shows a Pair wise plot

- The fact that most of the applications on the platform are free suggests that people favour free apps
- Ratings for paid applications are often higher, with a focus around 4-rating threshold.
- An app's number of installs and number of reviews are positively correlated, indicating that well-liked applications get more user feedback.
- The dataset has a lot of lightweight applications, which suggests that consumers prefer apps that take up less space on their smartphones.

Heatmap of correlation

A correlation matrix offers a thorough breakdown of the correlation rates between various dataset variables. By tabulating the correlations between all feasible pairings of variables, it aids in summarising and spotting trends in the data

On the other hand, a correlation heatmap uses a color-coded design to graphically depict the correlation matrix. It enables a more logical and simple interpretation of the correlation patterns. connection coefficients have values between -1 and 1, with a value of 1 denoting a strong positive connection and a value of 0 indicating no correlation between the variables.

It's crucial to understand that correlation does not always suggest a causal connection between variables. High correlation does not prove causality, but it does point to a probable relationship between variables. The main purpose of correlation analysis is to find statistical dependencies and correlations between variables in a dataset. Play store Correlation Heatmap.

Play store Correlation Heatmap

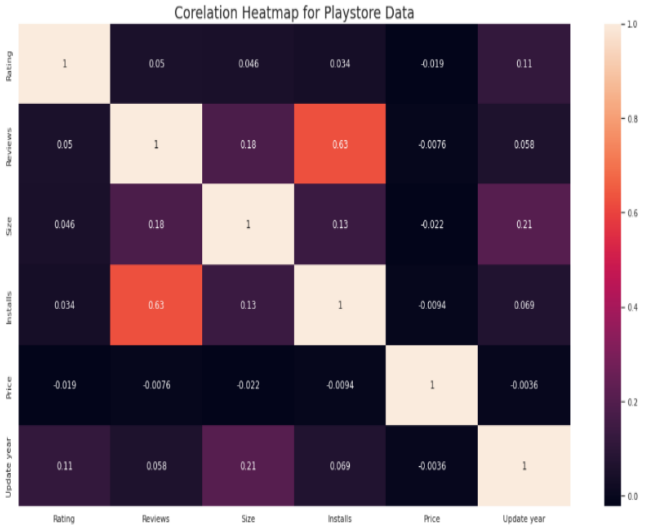

Fig -20: Heatmap of Correlation

- The Reviews and Installs columns have a significant positive link, according to the correlation study. This conclusion makes sense given that increased app installs often result in a bigger user base and more user reviews.
- Furthermore, In addition, the Price column and the Rating, Reviews, and Installs columns have a tiny inverse relationship. This implies that the average rating, total number of reviews, and number of installs all have a tendency to somewhat decline when app costs rise.
- Finally, there is a modest positive link between the Installs and Reviews columns and the Rating column. This shows that there is a trend for both app installs and the number of reviews to rise when the average user rating rises.

These findings are based on the statistical link between the variables in the dataset, but it's crucial to remember that correlation does not indicate causality

Merged Data frame Heatmap

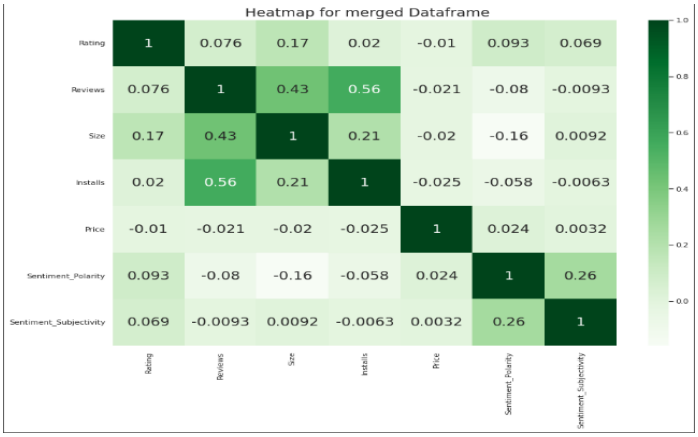

Fig -21: Heatmap for Merged Data Frame

Conclusion

Several patterns and hypotheses have been identified via exploratory data analysis, revealing potential contributors to the popularity of an app among Play Store users. These results provide marketers and app developers useful information. It is crucial to keep in mind that these are based on observational analysis, and further investigation or testing is required to prove causal correlations. Nevertheless, these insights provide a place to start when analysing user preferences and behaviour inside the Play Store ecosystem, which helps in strategic app development and marketing strategy decision-making.

- 92% or so of the applications in the sample are no cost.
- The age limits on 82% of the applications are absent.
- When it comes to the number of apps, "Family" is the most competitive category.

- "Family" (1906 applications), "Game" (926 apps), and "Tools" (829 apps) are the top three categories with the most apps.
- Among the most popular categories are "Tools," "Entertainment," "Education," "Business," and "Medical."
- There are both free and premium applications, totaling 7749 with ratings above 4.0 and 8783 less than 50 MB in size.
- The category with the most average app installations is "Game."
- Nearly 80% of the applications have high ratings.
- Over a billion people have downloaded 20 free applications.
- The only paid programme with more than 10 million downloads is Minecraft, which also makes the most money from the installation charge.
- "Finance" is the category with the highest average installation cost for premium applications.
- The average app size in the Google Play store is 12 MB.
- The most apps are installed on average when their size varies depending on the device.
- The most average user evaluations are seen for apps larger than 90 MB, a sign of their popularity.
- The majority of reviews for "Helix Jump" are good, whereas the majority of those for "Angry Birds Classic" are unfavourable.
- According to the combined dataset's total sentiment count, 64% of the feelings are positive, 22% are negative, and 13% are neutral.
- Sentiment polarity and sentiment subjectivity do not significantly correlate.

Reference

1 Kaggle.com. (2018). Google Play Store Apps.[online]https://www.kaggle.com/lava18/google-play-store- apps [Accessed 3 Mar. 2020].

2 "Mining and Analysis of Apps in Google Play," Proceedings of the 9th International Conference on Web Information Systems and Technologies, 2013.Google playstore: number of apps 2018(2018).[online] https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/ [Accessed 3 Mar. 2020].

3 Amit Chile., Gundalwar. P.R. (2019). Analysis of Google Play Store Application. [online]http://ijraset.com/fileserve.php?FID=24134 [Accessed3 Mar. 2020].

4 Denoeux, T and Skarstein-Bjanger. M. (2000). Induction of decision trees from partially classified data, in: Proceedings of the 2000 IEEE International Conference on Systems, Man and Cybernetics (SMC'00), IEEE, Nashville, TN, 2000, pp. 2923–2928.

5 Harman, M., Jia, Y. and Zhang, Y. (2012). App store mining and analysis: Msr for app stores. In 2012 9th IEEE Working Conference on Mining Software Repos- itories (MSR), pages 108–111.

6 Rajeswari, R.P., Juliet, K and Aradhana. (2017). "Text classification for student data set using naive bayes classifier and KNN classifier," Int. J. Com-put. Trends technol., Vol. 43, no. 1 pp. 8-12, 2017.

7 Jong, J. (2011). Predicting rating with sentiment analysis.http://cs229.stanford. edu/proj2011/Jong-PredictingRatingwithSentimentAnalysis.pdf.

*****