Play Store Apps Review Analysis

Rajat Chaudhary, Anukriti Shakyawar,

Raman Kumar, Deepmala Srivastava

Team: Web Crawlers

Abstract:

One of the biggest and most well-known Android app shops is the Play store. It has a vast amount of data that may be utilized to create the best model. Two data sets were provided to us. The initial dataset was Play store data, which comprises 13 different features that can be used to forecast an app's success or failure based on various criteria. The second dataset included user reviews, which were evaluated based on five different criteria to determine which app was most popular with users. The question set for our project is to deliver insights to understand customer demands better and thus help developers to popularize the product.

1. Introduction

Plat Store is one of the marketplaces for downloadable software programs with the highest growth in mobile applications. Playstore is a platform that provides its users with a variety of digital content, not only an app store. Google created and runs the digital distribution service known as Google Play Store, formerly known as Android Market. It is a legitimate app store with a wide selection of media, including apps, books, magazines, music, movies, and television shows. Since 92.2% of the apps are free, this platform has experienced tremendous growth in popularity. The play store app's dynamics are revealed in this document, which also offers developers useful information they can use to dominate the android market. Given the explosive rise of Android-based gadgets and applications, it would be fascinating to do data analysis on the collected information to gain insightful knowledge from this data.

2. Data Description

DATASET 1: Playstore App data

In this dataset we had the data of apps which had 13 columns and 10841 entries, the following were:

1. App: *Name of the apps*

2. Category: *Category under which it falls.*

3. Rating: *Applications rating on PlayStore.*

4. Reviews: *Number of reviews of the app.*

5. Size: *Size of the app.*

6. Installs: *Number of Installations of the app.*

7. Type: *Whether the app is Free or Paid.*

8. Prize: *Price of the app if it's a Paid app, for Free apps, it's zero.*

9. Content Rating: *Appropriate target rating of the app.*

10. Genres: *Genres under which the app falls.*

11. Updated: *The date on which the app was last updated.*

12. Current Version: *Version of the app.*

13. Android Version: *Minimum android version required to support the app.*

DATASET 2: User Reviews data

In this data set we had the customer review who have experienced those apps, In this, we have 5 columns and 64295 entries. Here data set is classified by-

1. Apps

2. Translated Review

3. Sentiments

4. Sentiment Polarity

5. Sentiment Subjectivity

3. Analysis Methodology

Integral research, data cleaning and filtering, and data visualization make up our three-part analysis strategy. We started by performing some fundamental analysis on our dataset. When we did this, we found the basic information regarding our data set such as columns, and data types and we also found out that we have missing values, data duplication, and a few other issues as well, so we had to run a few steps to extract just the data we needed for exploration. Second, we cleaned up our data by eliminating duplicate entries, converting some variables into usable forms(like $ and + signs used in our data), and filtering out any data with null values. Third, we conducted data visualization. For this, we used a variety of tools, including Numpy, Pandas, Matplotlib, Seaborn, and WordCloud, to accomplish data visualization. We performed these steps on both of our DataSets, firstly on Playstore data and then on our User Reviews dataset.

3.1. Data Cleaning

While analyzing our dataset the first thing we will do is to examine the null or missing values in our dataset which is very important to remove because it might affect the accuracy and performance of our analysis and can also show false results at the end of our process. This makes our result accurate. There are many missing values in the Size & Rating columns which can be seen by plotting graphs. Hence several methods are used to remove these values.

The first step in any data science effort is cleansing the data. The outcomes are better when the data is cleaner. According to the proverbial adage by Peter Norvig, "More Data Beats Clever Algorithm, But Better Data Beats More Data." We initially remove duplicate values from the data before we start cleaning it up. Then we purge irrelevant characters from our dataset.

Following this, we identify the distinct values for each column and make the appropriate adjustments, such as changing the data types and getting rid of the null and "nan" values.

Lastly, we have done an exploratory data analysis of our dataset.

3.2. Data Visualization

In our Data Visualization, we performed many analyses to find relations in our data set. In which we first performed basic analysis such as Application Type, categories, and so on.

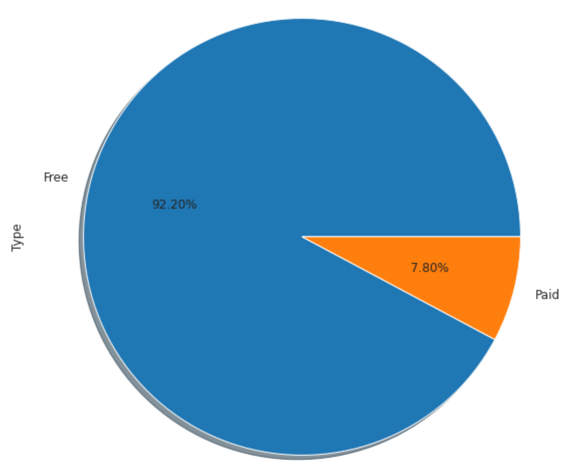

Fig.1: Distribution of Application Type

From the above graph, we can conclude that the majority of the apps in the Play Store are Free apps i.e. 92.2%, and only 7.80% of apps are paid apps.

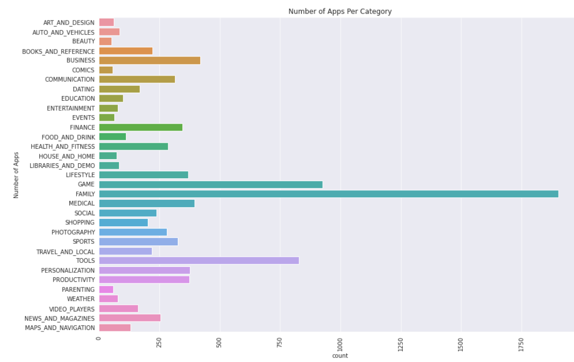

Fig.2: Number of Apps per Category

From this plotting we know that there are many categories in the Play Store and most of the apps are from the categories of 'Family', 'Game', and 'Tools' category.

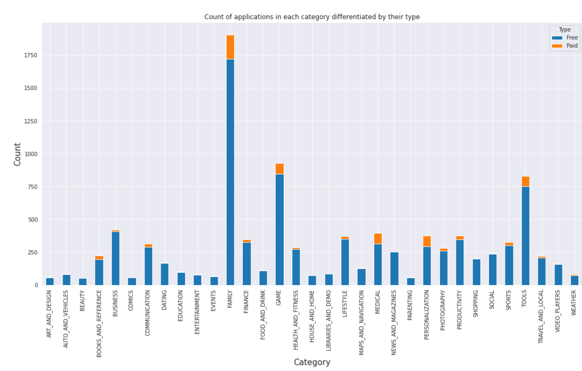

Fig.3: Number of apps per category differentiated by type

This graph shows that some app categories provide more free downloadable apps than others. Most of the apps under the Family, Games, and Tools, as well as Social categories in our dataset, were available for free download. At the same time, the most paid apps were available for download in the Family, Personalization, and Medical categories.

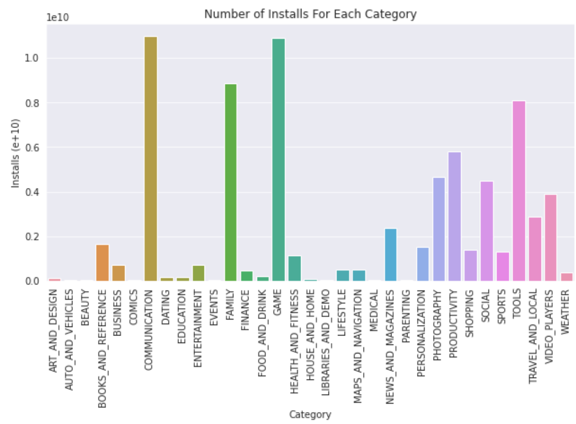

Fig.4: Number of installs for each category

The majority of apps available in the Play Store fall into the Family, Games, and Tools categories, although this is not true according to installations and market demands, as shown by the two plots shown above. Games, communication, and tools have the most installed apps.

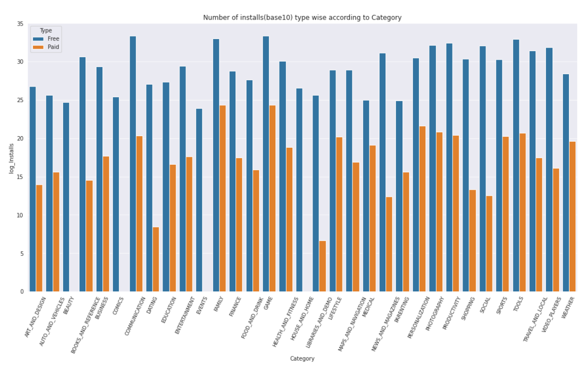

Fig.5: Number of installs type-wise according to categories

This graph shows the number of installations based on the Type and Category of apps. In this graph, we found out that app installations have a significantly higher proportion of free software than paid ones. The comparison comparing free and paid apps appears to show no variation because we converted the number of installs to its log.

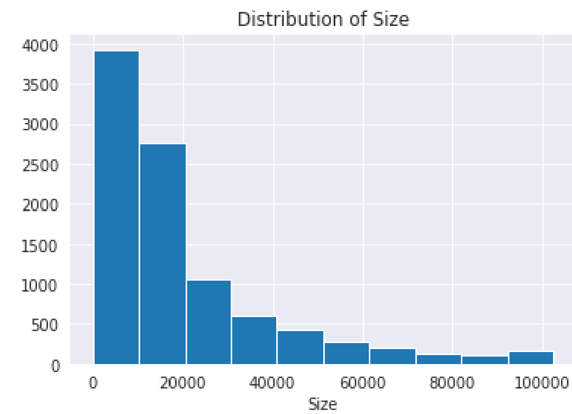

Fig.6: Distribution of Size of app

In this graph, we can see that most of the apps present in the Play Store are smaller in size and consume less memory and only a handful of apps are larger.

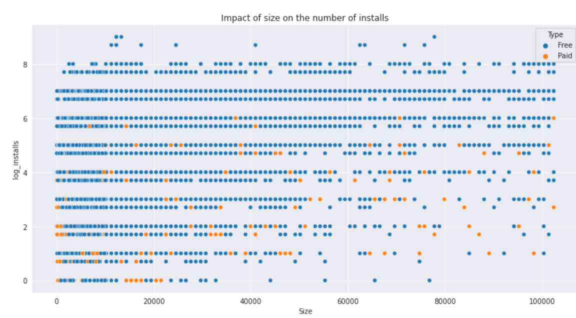

Fig 7. Impact of size on the number of installs

In this graph, we can see that the size of the application greatly impacts the number of installs by the user. On the other hand, we can see that the bulky applications are less downloaded. We can also derive from this graph that paid applications that are bulky in size are also less installed, which is the opposite in the free case, the bulky applications are still more downloaded than the paid bulky applications. Hence, we can say that size affects the installs of apps by the user.

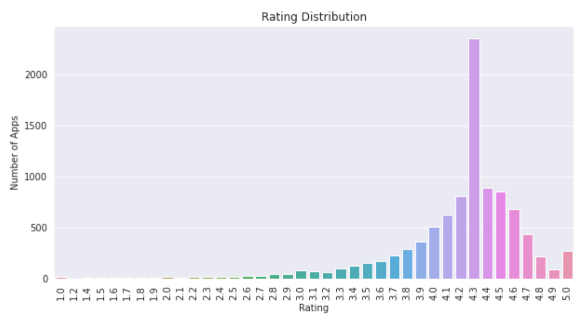

Fig 8. Distribution of App Rating

From this graph, we can say that most of the apps in the Play Store are having ratings higher than 4 or in the range of 4 to 4.7.

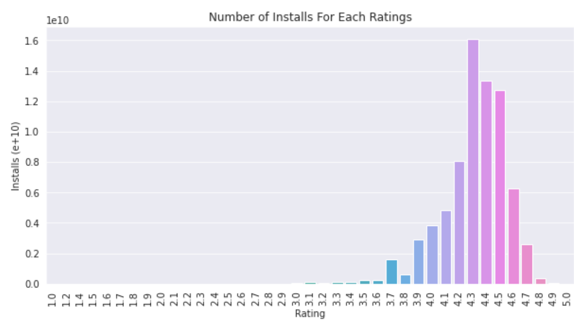

Fig 9. Install per Rating

In this graph, we can see that most of the apps downloaded by the customers are of higher ratings also which from 4.2 to 4.6 ratings.

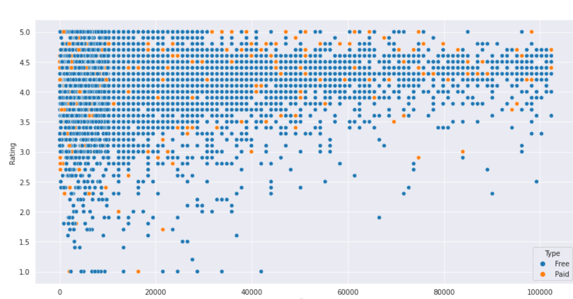

Fig 10. Rating based on Size

In this graph we can see that most of the apps with a higher rating are smaller apps in size, we also put paid and free apps in this also but the type of apps are evenly distributed so the type of the app doesn't affect the rating but the size of the app surely does.

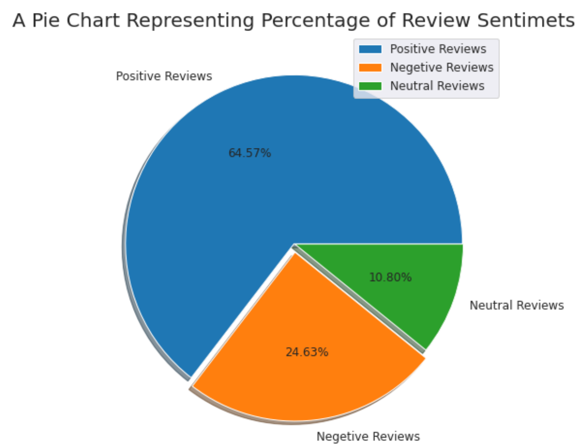

Fig 11. Reviews Sentiment

In this graph, we can see that most of the reviews are positive at 64.57% and negative reviews are only 24.63%. There are very less neutral reviews by the users with only 10.8%.

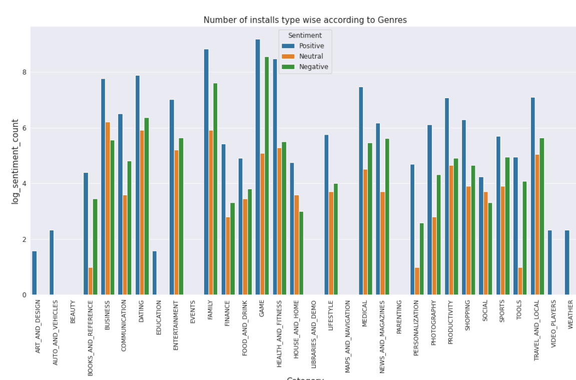

Fig 12. Sentiment count review-wise according to Category

In this, we can see that positive reviews are higher in each category of Play Store data, while in categories we can also see that there is very less difference in the positive and negative reviews such as Game and Business. But there also categories where neutral reviews are also very large as compared to negative reviews such as House and Home and Business. In such conditions the experience of the users also depends for some users it is bad so they provided negative reviews but for users, it is not that great so they provided unbiased reviews but if such a thing appears in results then it is important of the app to improve their user experience.

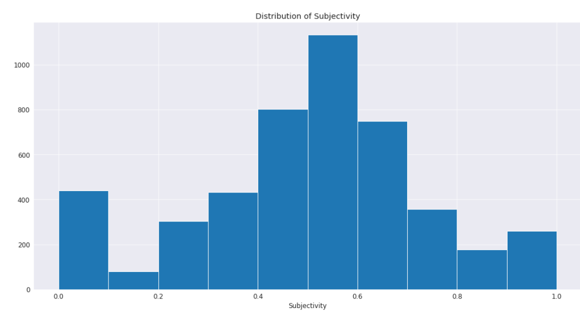

Fig 13. Distribution of Subjectivity

From this graph we can say that the maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this we can conclude that these reviews comes from the experience from the users while using these apps.

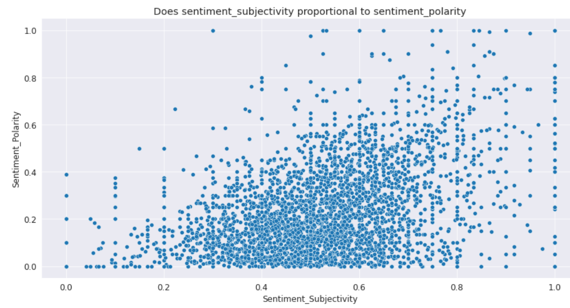

Fig 14. Sentiment subjectivity and Sentiment Polarity

Polarity is a basis that defines which types of reviews are given by the users, it varies in Positive, Neutral, and Negative reviews. When polarity goes low it shows that the review is either neutral or negative but when it goes upper it shows that the polarity is positive. Subjectivity, on the other hand, is based on emotions, personal experience, or judgments, and whether the customer gave the review with the correct feelings in mind (it might be critical or good) or simply for the sake of writing anything. Subjectivity is also graded from 0 to 1. According to the scatter plot, while sentiment subjectivity does not always correlate with sentiment polarity, it does so more frequently when the variance is very big or low. This graph indicates that users have submitted evaluations on an average usage of the applications, with ratings ranging from 0.3 to 0.7, but have not built a relationship with it. It is also clear that the majority of the reviews fall within 0-0.6, indicating that the reviews are impartial.

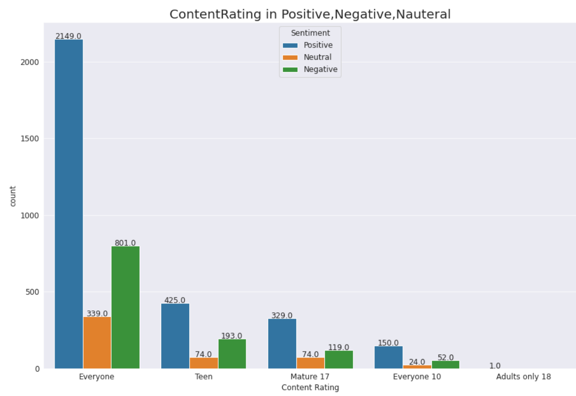

Fig 15. Content Rating based on Age

In this graph, we can see that most of the Ratings came from the Everyone category and the ratings went lesser when we move forward with the age criteria of the apps and went to even 1 when it comes to the adults-only category. While talking about the sentiment of the review we can also see that most of the positive ratings came from the everyone category apps with 2149 ratings and 801 negative ratings. In this also we can see that neutral rating are also low in comparison to positive and negative reviews

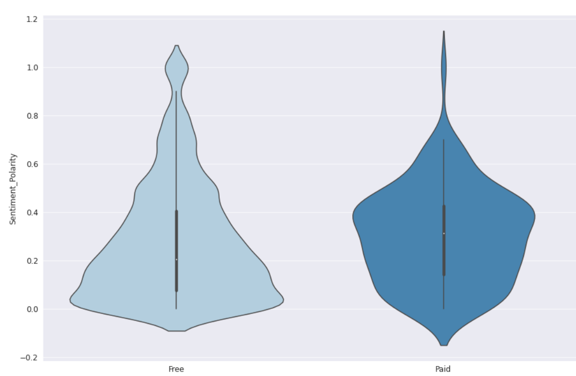

Fig 16. Sentiment polarity relation to the type of app

In this graph, we can see that in Free apps the sentiment polarity lies majorly at 0.1 and decreases after that which shows that only a handful of users give reviews after experiencing the apps, while on the other hand when the sentiment polarity of paid apps which falls majorly between 0.1 to 0.4 with highest in 0.4, in this we can say that customers first experience these apps and then gave their reviews about those apps.

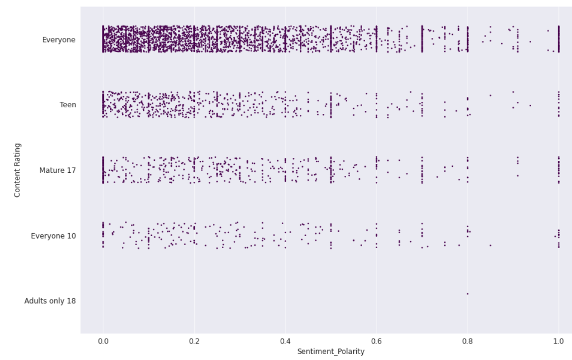

Fig 17. Content Rating relation with Sentiment Polarity

In this, we can see that sentiment polarity is low in all categories but in the Above 10 categories of Content Rating, we can see that sentiment polarity is evenly distributed, in this, we can say that users have first used the app and then they give their reviews but this can also be seen as less number of reviews in that category. In the Everyone category, we have the highest number of reviews but major part of those reviews lies between 0.1 to 0.4 and not more reviews are going higher than that, in this we can say that customers have not used those apps and gave their reviews beforehand.

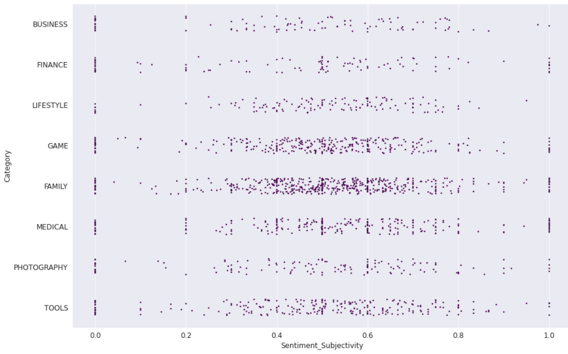

Fig 18. Categories Relation with Sentiment Subjectivity

In this graph we can see that Sentiment Subjectivity lies between 0.2 to 0.8 in all Categories, we can also see that the Family Category has the higher number of reviews in which Sentiment Subjectivity lies between 0.4 to 0.6 which shows that apps in this Category have been reviewed after using them.

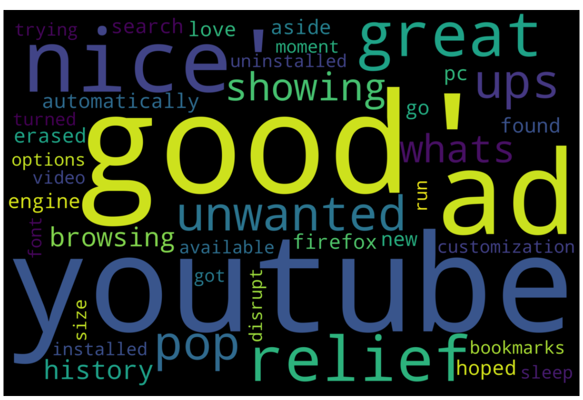

Fig 19. Word Cloud of Reviews

In this Word cloud, we can see what words are mostly used in the reviews.

Conclusion

We discovered some important insights regarding the applications and the user's behavior while installing such apps from the Google Play Store through this research of the Google Play Store Review Dataset. According to our graphs in Data Visualizations above (Fig. 5), the majority of apps installed by users are from the Family, Games, and Tools categories. Despite the fact that the quantity of applications in this category is similarly considerable in comparison to other categories, user attractiveness to apps in these categories is still quite strong. Aside from that, the type of app and the size of the app both have a significant role in user app installations. We can see from the above (Fig 7) that customers have installed more free apps than premium apps. The following graph (Fig 7) also shows the association between app installations, size, and kind. In that graph, we also discovered that people prefer apps that are small in size to those that are large. This can also be observed when we talk about the type of app, with free apps being more popular than premium apps. However, only a few free programs that are large in size are chosen by consumers. This demonstrates that the software should be small in size in order to be more appealing to consumers. Another important aspect that determines the number of app installations by users is the app's rating; as seen in the graph (Fig 9), apps with high ratings are more downloaded. The graph above (Fig 10) depicts the relationship between app rating, size, and app kind. We discovered that the ratings of light-sized apps are higher than those of bulky apps, with just a few huge apps receiving high marks. When it comes to app type, we can observe that whether the app is paid or free, if it is large in size, it does not have excellent reviews.

To know more about the reviews of the users we need to find pattern in which they gave their reviews. For this we first analysed the type of sentiments we have in reviews (Fig 11) in this section we saw that there are 3 parts in which reviews are divided- Positive, Neutral and Negative. Now we need to find out how these sentiments are divided and we found out that Game and Business have more positive reviews as compared to other categories.

Now, the main question is whether the app is used properly before the review or not for that we move on to our next graph(Fig 13) in which we found out that most the subjectivity lies between 0.4 to 0.7 and only a handful of them are 1. This shows that these apps are not used for a long time before giving their reviews.

However, we cannot influence the ratings of the applications as they are supplied by users; we can only improve the user experience since if the user's first experience is positive, they will offer positive ratings, as seen in the sentiments of subjectivity and polarity graph (Fig 14).

In another graph (Fig. 16) we can see that Paid apps are used before their reviews as compared to the Free type apps.

With this analysis, we can help many developers and business owners who are in the android app development and creation business by advising them on which category to choose, what size their apps should be, and whether they should be paid or free because these details determine whether the app will be liked or preferred by users.

Future Work

We can explore the correlation between the size of the app and the version of Android on the number of installs we can also explore reviews and sentiment of the users as per the category of the application.

To improve the program we could add a system that would create applications on its own by using the data set and creating the best user interface by highly rated apps.

References

1. https://www.kaggle.com/lava18/google-play-store-apps
2. https://jovian.ml/ritz1602-rs/course-project-google-play-store-dataset
3. https://jovian.ai/learn/data-analysis-with-python-zero-to-pandas
4. https://seaborn.pydata.org/examples/index.html
5. https://matplotlib.org/3.1.1/index.html