Play Store Apps Review Analysis

Rajat Chaudhary, Anukriti Shakyawar,

Raman Kumar, Deepmala Srivastava

Team: Web Crawlers

Abstract:

Play Store is the first choice whenever we want to download an app on our android phone. It consists of a large amount of data that can be used to analyze and prepare a model that can be helpful to developers and companies also. In this data set we were provided with two data set in which first one is Play Store data set which consists of our basic information like the name of app, categories and so on and the second data set was User Reviews in which we have customers reviews for apps, both data sets are connected with a key column which is App which has the names of all the apps present in data set. The objective of this project is to deliver insights to understand customer demands better and thus help developers to popularize the product.

1. Introduction

Play Store is the biggest platform when it comes to download mobile applications for the android users, It was created by Google and comes under the name of Google Play Store previously known as Android Market. It provides a variety of applications from different categories like Games, Entertainment, Business, Tools and so on. Its applications are divided into two types Free and Paid in which 92.2% of apps are Free apps. This data set contains a tremendous amount of information which can be used by developers and companies and can be used to dominate the Android Applications market. With this analysis we can know the classifications through which apps are present in Play Store and the customer behavior to download these apps also, such type of information is very important when it comes to business.

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

There are some basic steps which we performed in our analyses. The first one is Integral Research, in this step we first uploaded the data and then perform basic operations such as info, shape and describe commands to know more about the data such as data type, columns, total number of entries and null values. This type of information gives us information to process to our next step. After that we performed data filtering, in this step we remove all the duplicate rows and the null values from our data set to provide more logical analyses and the third step was data cleaning in this step we cleared all the extraneous data (such as removing signs like $ and +), these steps are performed on the both the data sets. After the data filtering and cleaning we move to our next step which is data visualization for which we used a variety of packages such as Pandas, Numpy, Matplotlib, Seaborn, Word Cloud and Stopword.

3.1. Data Cleaning

In this step the first thing we need to do is find the null values in each data set which can be done by using info command. After that we remove all the null values from each column, when we are finished with this step we move on to our next step which is removing duplicate data. As you can see in many columns like Size, Content Rating we have different types of impurities like Null values in Size and data with incorrect information of data different from the required form is present in content rating column so, we have also removed such kinds of data also.

It is true that more data gives better analyses results but better data gives more accurate and clear results. So by doing these steps we can make our data more accurate.

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

3.2. Data Visualization

In this section we have performed various operations on our data to find meaningful answers or analysis and find a pattern in which users tend to download their apps.

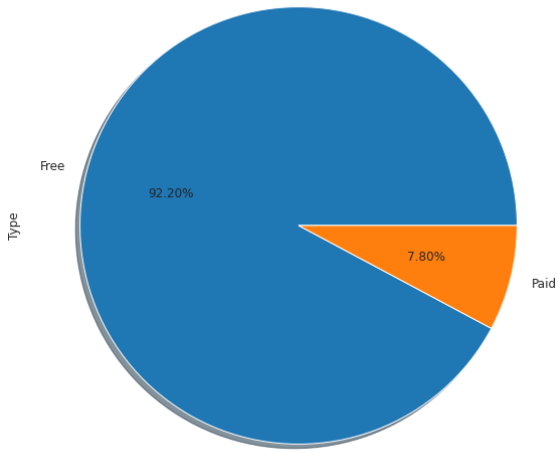

Fig.1: Distribution of Application Type

This graph shows the type of Applications present in the Play Store and in which percentage they are present. In this we can see that we have 2 types of applications. The first one is Free applications with 92.2% and the second is Paid applications with 7.8% of presence in the data set.

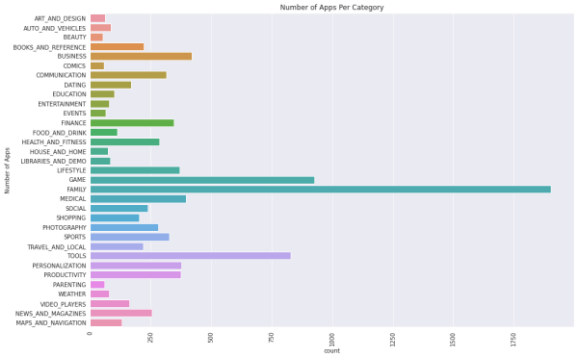

Fig.2: Number of Apps per Category

This graph shows the number of applications present in each Category. We can also see that Family, Games and Tools have the most number of applications as compared to other categories.

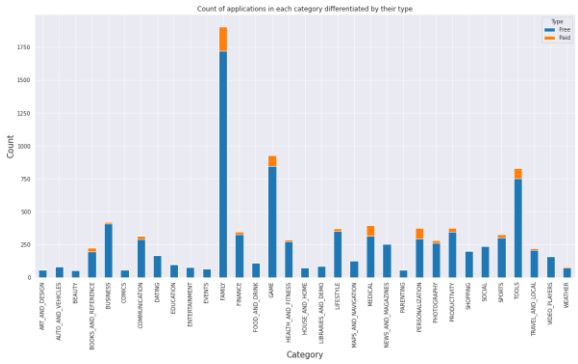

Fig.3: Number of apps per category differentiated by type

This graph shows the relation between the Category and the Type of apps, as discussed in previous we can also see here that Family, Tools and Games category have the highest number of apps. In which Medical and Personalization has the highest number of Paid apps as compared to other categories, on the other hand Business and Lifestyle has more number of Free apps as compared to Paid apps.

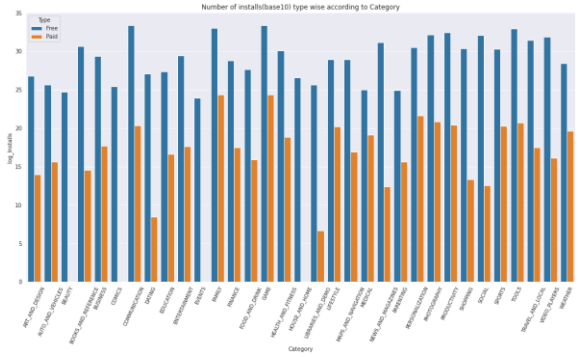

Fig.4: Number of installs type-wise according to categories

In this graph we can see the relationship between the number of installs and the type of apps in each category and we can see that the Free apps are more installed by the users as compared to Paid apps in each category. Majorly we can see that in all categories the difference between paid and free app

installs are high but in few cases we can see that this installs gap is very low.

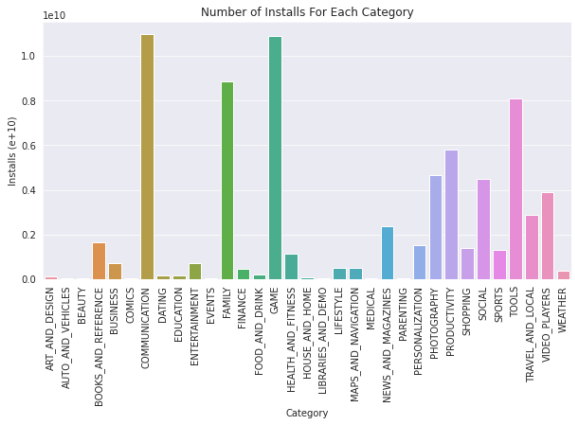

Fig.5: Number of installs for each category

In this graph we can see that the number of installation of apps in Communication, Game, Family and Tools are very high as compared to other categories, which means that the apps of these categories are the most downloaded as compared to other categories, hence concluding that the demand of these categories are higher in market.

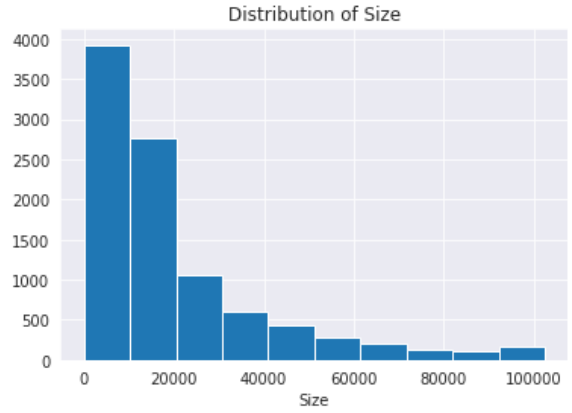

Fig.7: Distribution of Size of app

In this graph we can see that the number of light apps are more present in the Play Store as compared to bulky apps. According to the following trend we can see as the size of the apps increases their number are simultaneously decreasing.

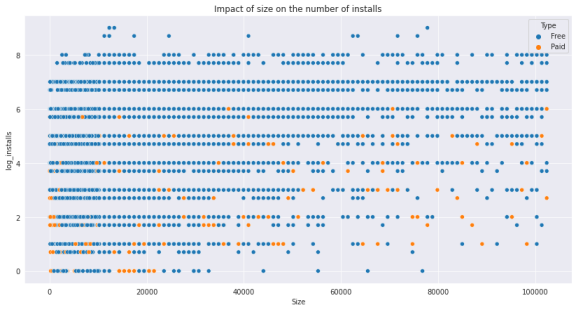

Fig 8. Impact of size on the number of installs

This graph shows the relation of Size and Installs of apps with their type. In this graph we can see that the less the size of the app the more it is downloaded, in the case of type also it shows that only a handful of paid apps with bulky size are installed at larger number but still the pattern shows that the app needs to be light in size to be more preferred by the users.

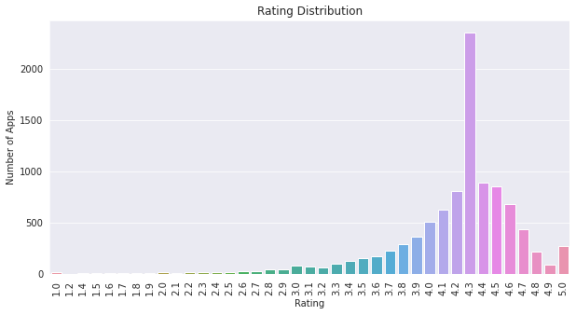

Fig 9. Distribution of App Rating

This graph shows the flow of ratings given by users and it is clear that the most high rating is 4.3. Most of the ratings given by users are in the range of 4 to 4.7.

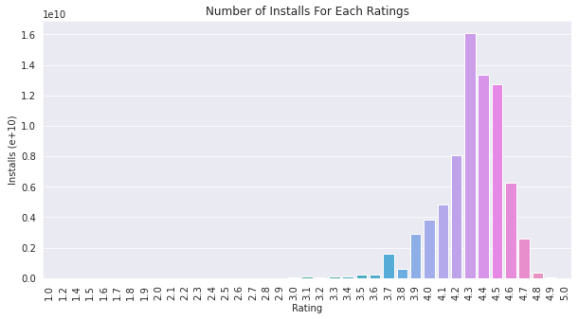

Fig 10. Install per Rating

As we can see in this plot that the Rating also has a major role in the preference by the users, apps with rating between 4.2 to 4.6 are majorly installed by the users.

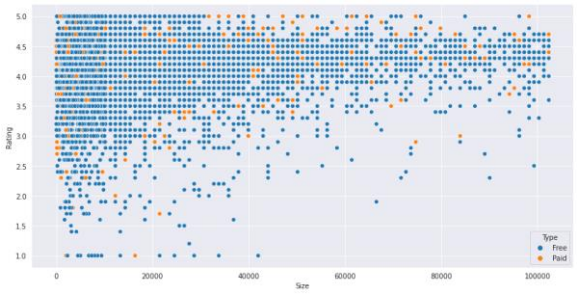

Fig 11. Rating based on Size

This graph shows the relation between the Rating and the Size along with the Type of the App. We can see that the Rating of light apps is higher than the rating of the apps which are bulky in size but the difference is the number of ratings these apps gain. On the other side we can see that apps which are bulky in size and has higher rating are generally paid apps. Free type bulky apps don't have higher ratings however this doesn't support in the case of light apps. Whether the app is paid or free if it is light in size it has a higher rating.

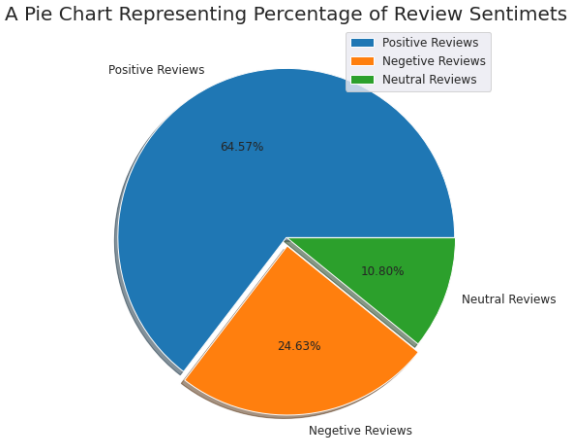

Fig 12. Reviews Sentiment

This pie chart shows the distribution of types of reviews. Positive reviews has the majority with 64.57% while Negative reviews and Neutral reviews are 24.63% and 10.8% respectively.

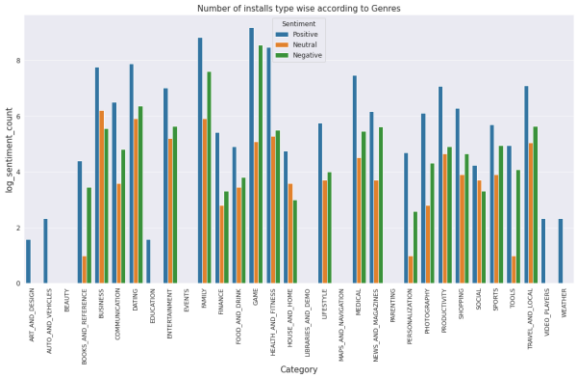

Fig 13. Number of Installs Type-wise according to Categories

Positive ratings are more prevalent in each area of Play Store data, while the difference between positive and negative reviews is noticeably smaller in categories like game and business. However, there are also categories, including "House," "Home," and "Business," where there are also a lot more positive evaluations than negative ones. Users' experiences in these circumstances vary; for some, they are poor, so they leave negative reviews; for others, they are average, so they leave

unbiased ratings; but, if such a thing emerges in the results, it is crucial for the app to enhance users' experiences.

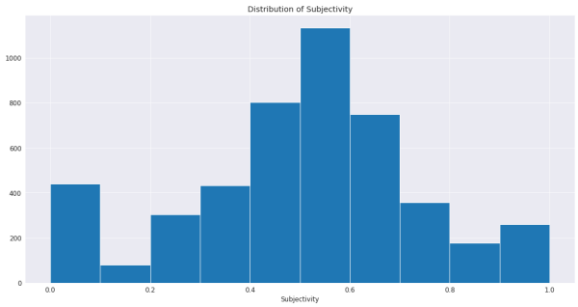

Fig 14. Distribution of Subjectivity

We might infer from this graph that there is a range of 0.3 to 0.7 where sentiment subjectivity is at its highest. It shows that the consumers have given most of the review on minimal to neutral usage of the apps.

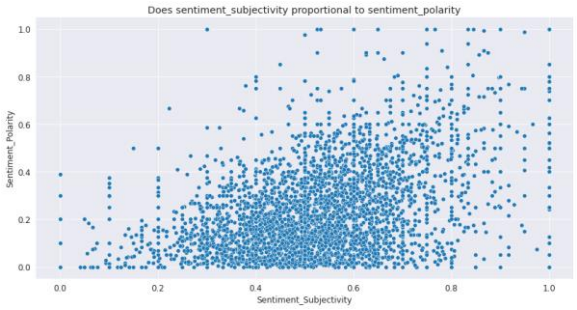

Fig 15. Sentiment subjectivity and Sentiment Polarity

Polarity is a float which lies between [-1,1] where -1 means a negative review and 1 means a positive review. Whereas, Subjectivity is calculated on emotions, personal experience or judgements, whether the consumer has given the review with right feelings in mind (it can be negative or positive) or has just given the review for the sake of writing something. Subjectivity is also measured between 0-1. From the aforementioned scatter plot, it can be inferred that while sentiment subjectivity does not always correlate with sentiment polarity, it does so more frequently when the variance is very large or low. It can be inferred from this graph that consumers have given reviews on an average usage of the apps as they are ranging between 0.3-0.7, and do not actually have established the relationship with it. It is also visible that the reviews mostly lie between 0-0.6, which means the reviews are neutral.

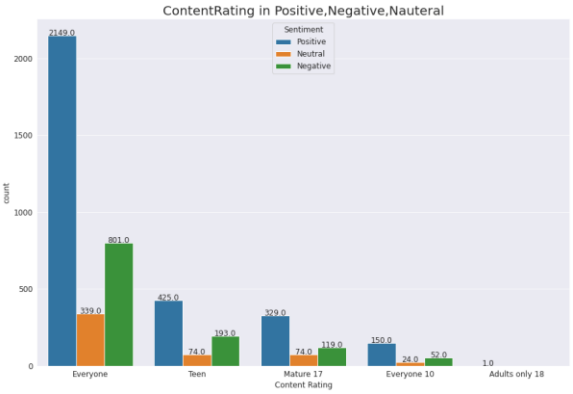

Fig 16. Content Rating based on Age

As we can see from the graph, the majority of the ratings originated from the Everyone category. As we moved forward with the age requirements for the applications, the ratings decreased and eventually reached 1 when it came to the Adults-Only category. When discussing the sentiment of the review, we can also notice that the majority of the positive ratings 2149 in everyone category applications and 801 negative ratings. Additionally, we find that the neutral ratings are weak compared to the positive and negative comments.

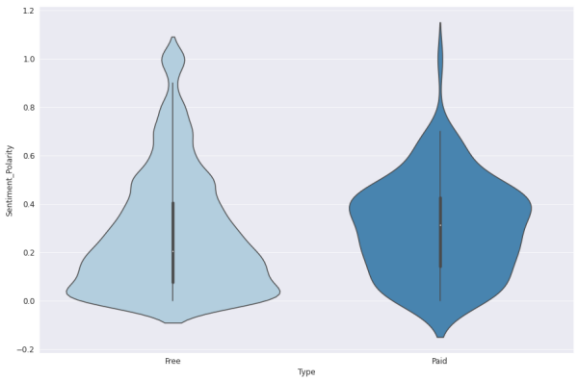

Fig 17. Sentiment polarity relation to the type of app

In this graph, we can see that the sentiment polarity for free apps is primarily at 0.1 and then declines, indicating that few users actually leave reviews after using the apps. In contrast, the sentiment polarity for paid apps is primarily between 0.1 and 0.4, with the highest value being 0.4, indicating that users actually use the apps first before leaving reviews.

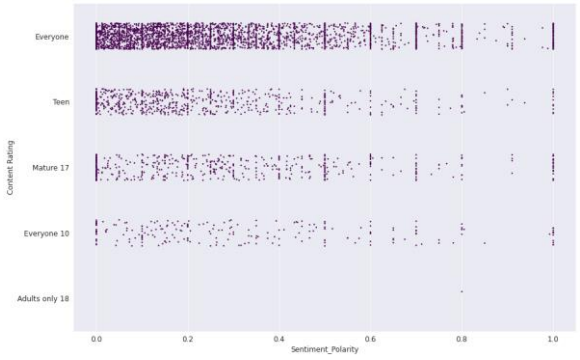

Fig 18. Content Rating relation with Sentiment Polarity

While sentiment polarity is low across the board, it is evenly distributed across the top 10 categories of content ratings. This suggests that users are initially using the app before leaving reviews, while it might also indicate that there are less reviews in that area. We have the most reviews in the Everyone category, but the majority of them fall between 0.1 and 0.4, and there aren't many that go higher. From this, we may infer that users did not really try the applications before leaving their ratings.

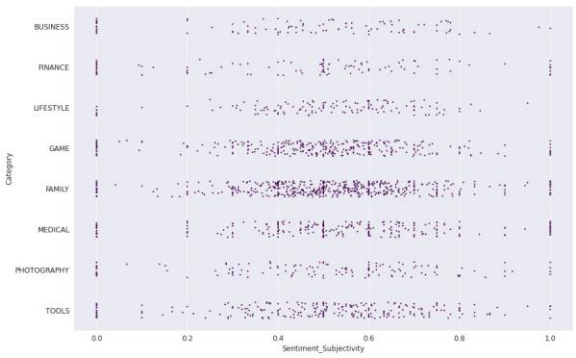

Fig 19. Categories Relation with Sentiment Subjectivity

The Family Category has the highest number of reviews, with Sentiment Subjectivity ranging from 0.4 to 0.6, indicating that users have rated the applications in this Category after using them. As can be seen in the graph, Sentiment Subjectivity ranges from 0.2 to 0.8 in all Categories. Which means that the reviews are not totally on non-usage or highly used experience, but on a neutral usage of the apps.

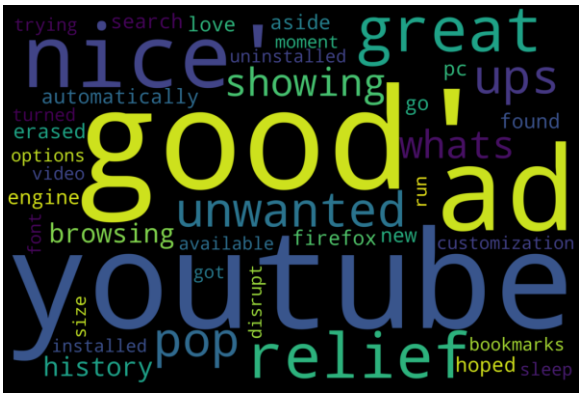

Fig 20. Word Cloud of Reviews

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

Conclusion

In this analysis of the Google Play Store Review Dataset, we found some useful insights about the apps and the user's behavior while installing those apps from Play Store. As per our graphs(Fig. 5) in Data Visualizations above most of the apps installed by users are from the Family, Games, and Tools categories. Even though the number of apps in this section is also high as compared to other categories , the user's attraction to apps in these sections is still very high. Other than that the type of app and the size of the app also contribute very highly in user app installs. In the above(Fig 7), we can see that users have installed free apps more as compared to paid apps. In the above graph (Fig 7) we can also see the relationship between installs, size, and type of the apps. In that graph we also analyzed that apps that are light in size are installed by the users as compared to the bulky apps, this can also be seen when we talk about the type of the app, in which the free apps are more installed as compared to the paid apps. But only a handful of free apps which are bulky in size are preferred by the users. This shows that the app needs to be light in size to be more preferable to the users. There is another major factor that defines the number of installs by the users and that is the rating of the app, in the graph (Fig 9) we can see that apps with high ratings are more downloaded. In the above graph (Fig 10) we show the relation of app rating, size, and the type of the app. In that we analyzed that the ratings of the light sized apps are more as compared to bulky apps, only a handful of bulky apps have good ratings. When we come to the type of the app we can see that it doesn't matter of the app is paid or free if it is bulky in size then it doesn't have good ratings. However, we can't control the ratings of the apps as it is given by the users but we only make the experience of the users better because when the first experience of the users is good then they will good reviews which we saw in sentiments subjectivity and polarity graph (Fig 14).

Basically any app made in the top 5 categories with a smaller size preferably free in type with higher number of review and so be in the positive polarity are more likely to be succeeded

Through this analysis, we can help many developers and businessmen who are in the business of app development and creation about which category to choose, what should be the size of their apps, and whether they should be paid for free because these details decide whether the app will be liked or preferred by the users or not.

Future Work

We may investigate the relationship between the app's size and Android version and the number of installations. We can also investigate user feedback and opinions in connection to the app's category.

A system that would automatically construct applications utilizing the data set and provide the best user interface by highly rated apps might be added to the software to improve it.

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