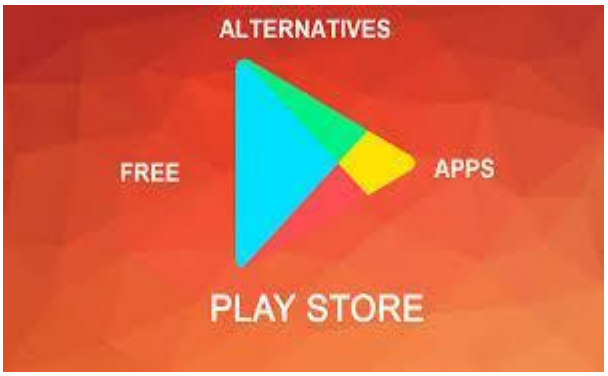
CAPSTONE PROJECT:

Play Store App Review Analysis



Group Name: data_battalion

Group Members: 01.Ishan Sharma (group Leader)

02. Mohd Ashif Khan

03. Nitesh Pawar

04. Virender Chib

05. Kismat Choudhary

1 ABSTRACT:

Software application is vital because specific software is required in almost every industry, in every business ,and for each function. It becomes more important as time goes on. Google play store is with a few thousands of new applications regularly with a progressively huge number of designers working freely or on the other hand in a group to make it successful, with the enormous challenge from everywhere throughout the globe. . A huge number of designers work freely on designing the apps and making them successful. Since most Play store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, adverts and memberships add to the achievement of an application. In this way, an application's prosperity is normally dictated by the quantity of installation of the application and the client appraisals that it has gotten over its lifetime instead of the income created. The objective of this experiment is to deliver insights to understand customer demands better and thus help developers to popularize the product. We have tried to discover the relationships among various attributes such as which application is free or paid, what are the user reviews, rating of the application. The objective of this experiment is to deliver insights to understand customer demands better and thus help developers to popularize the product. We have tried to discover the relationships among various attributes such as which application is free or paid, what are the user reviews, rating of the application.

The study aims to predict the rating of google play store apps and their analysis based on different criterias with the help of Python datasets. We have tried to perform data analysis, data cleaning and data visualization.

***Key Words*:** Google Play Store Apps, Ratings Prediction, Exploratory Data Analysis, Data Cleaning, Data visualization

1. PROBLEM STATEMENT

Data is taken from the Google play store dataset. There are two datasets as Play store dataset & User Reviews. Each app has values for category,rating,size and more.another dataset contains customer reviews of the android apps. We will be doing Exploratory data analysis on this data set, which is a very important step in the data science cycle, as it not only helps in taking very initial business decisions. Our objective will be to structure the data, clean it and present certain trends that we observe that can help us draw very preliminary conclusions about the probability of success of a newly launched app.

2. INTRODUCTION

In today's world we can see that mobile is a part of life.Mobile phones have many more applications and these apps are available on the platform of google playstore for android versions mobile phone. It is flooded with millions of applications and it provides a wide collection of data on features like ratings, price and number of downloads and app description. Many apps are being developed as apps are easy to create and it's lucrative. But it's important for developers to know which apps are loved by customers and are trending in the market so that they develop only those apps and also there is a high competition between app providers producing similar applications. Analyzing customer needs is one of the most difficult tasks in the business world today. Hence proposing to analyze data to the developer that what customer is likely to download, which category got the maximum downloads this all plays a crucial role in app development. With enormous challenges from everywhere throughout the globe, it is important for a designer to realize that he/she is continuing in the right way or not. To hold this income and their place in the market the application designers may need to figure out how to stick into their present position. The dataset with approximately 10000 Play Store applications is available to analyze the market of android. It can be examined to analyze the different categories such as family, communication, entertainment, tools, music, camera etc.

2.1 GOOGLE PLAY STORE AND USER REVIEW ANALYSIS

In this project we are going to analyze the different attributes present in the data set that affect the popularity of the application. We focused on the questions like what makes an app popular, what should be the price and size of the app,the different genres, age group and reviews by customers. In our project we are dealing with the two datasets csv files for data analysis: - PLAY STORE DATA & USER REVIEWS. At first, we analyzed the play store data and in the play store data we have 10841 rows and 13 columns. In the user review dataset we have 64295 rows and 5 columns. We have to take the maximum outcomes from the data which help us to analyze which type of app is most preferable and comparisons between different insights. Our goal is to filter and make plots accordingly for a better EDA with respect to the final data. We need to explore and analyze the data to discover key factors responsible for app engagement and success.

2.2 GOOGLE PLAY STORE DATASET

This dataset has 10841 rows and 13 columns. The 13 columns are identified as follows:

- **App:** This Column contains the name of the app
- **Category:** This contains the category to which the app belongs. The category column contains 33 unique values
- .**Rating:** This column contains the average value of the individual rating the app has received on the play store. Individual rating values can vary between 0 to 5.
- **Reviews:** This column contains the number of people that have given their feedback for the app.
- **Size:** This column contains the size of the app i.e. The memory space that the app occupies on the device after installation.
- **Installs:** This column indicates the number of times that the app has been downloaded from the play store, these are approximate values and not absolute values.
- **Type:** This column contains only two values- free and paid. They indicate whether the user must pay money to install the app on their device or not.
- **Price:** For paid apps this column contains the price of the app, for free apps it contains the value 0.
- **Content Rating:** It indicates the targeted audience of the app and their age group.
- **Genre:** This column contains to which genre the app belongs to, genre can be considered as a sub division of Category.
- **Last updated:** This column contains the info about the date on which the last update for the app was launched.
- **Current version:** Contains information about the current version of the app available on the play store.
- **Android version:** Contains information about the version of the android OS on which the app can be installed.

2.3 USER REVIEWS DATASET

This user reviews dataset has 64295 rows and 5 columns. These 5 columns are identified as follows:

- **App:** Contains the name of the app with a short description.
- **Translated Review:** It contains the English translation of the review dropped by the user of the app.
- **Sentiment:** It gives the attitude/emotion of the writer. It can be ‘Positive’, ‘Negative’, or ‘Neutral’.
- **Sentiment Polarity:** It gives the polarity of the review. Its range is [-1,1], where 1 means ‘Positive statement’ and -1 means a ‘Negative statement’.
- **Sentiment Subjectivity:** This value gives how close a reviewer’s opinion is to the opinion of the general public. Its range is [0,1]. Higher the subjectivity, closer is the reviewer’s opinion to the

opinion of the general public, and lower subjectivity indicates the review is more of factual information.

2.4 Python for Data Science

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured, object-oriented and functional programming. Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed. Most of the info scientists use python due to the good built-in library functions and therefore the decent community. Python now has 70,000 libraries. Python is the simplest programming language to select compared to other languages. That is the most reason data scientists use python more often,

2.5 DATA CLEANING AND PREPARATION

Preprocessing is important in transitioning raw data into a more desirable format. Undergoing the preprocessing process can help with completeness and compellability. For instance, you'll see if certain values were recorded or not. Also, you'll see how trustable the info is. It could also help with finding how consistent the values are. We need preprocessing because most real-world data is dirty. Data can be noisy i.e. the data can contain outliers or simply errors generally. Data can also be incomplete i.e. there can be some missing values. Preprocessing is important in transitioning raw data into a more desirable format. Undergoing the preprocessing process can help with completeness and compellability. For instance, you'll see if certain values were recorded or not. Also, you'll see how trustable the info is. It could also help with finding how consistent the values are. We need preprocessing because most real-world data is dirty. Data can be noisy i.e. the data can contain outliers or simply errors generally. Data can also be incomplete i.e. there can be some missing values.The available data is raw and unusable for Exploratory data analysis, so before we do anything with the data we will have to explore and clean it to prepare it for data analysis.

- **Step 1:-** In the first step, to filter the user reviews dataset, we remove the null values in the rows of data sets.and with help of describe() method we find out the mean, mode, max, min values. The duplicates i.e. repeating entry we had removed here.
- **Step 2 :-** In the second step, we examine every column of data sets whether there is any missing value or repeated values. If there is, then we remove it and take only unique values by applying the unique() method.
- **Step 3:-** Now we have created the separate numerical data and categorical data for two different datasets as “num_dataf” and “cat_dataf”. This is the final step of data cleaning where we get all clean data available for further data visualization.
- **Step 4 :-** This step include the different problem statements as:
  Problem 1:- How many Android Version Supported Apps across the Whole Database?
  Problem 2:- What are the top most competing categories in the play store?
  Problem 3:- What is the Paid and Free apps ratio from all apps?
  Problem 4:- What are the update details of apps by year?
  Problem 5:- How is the App pricing trend across popular categories?
  Problem 6:- What are the Sentiment Data Across the All Reviews?

Problem 7:- Is there any correlation between Sentiment polarity/Sentiment subjectivity with Installs/Rating/Reviews/Size/Price/Last Updated Day/ Last Updated Month/Last Updated Year?

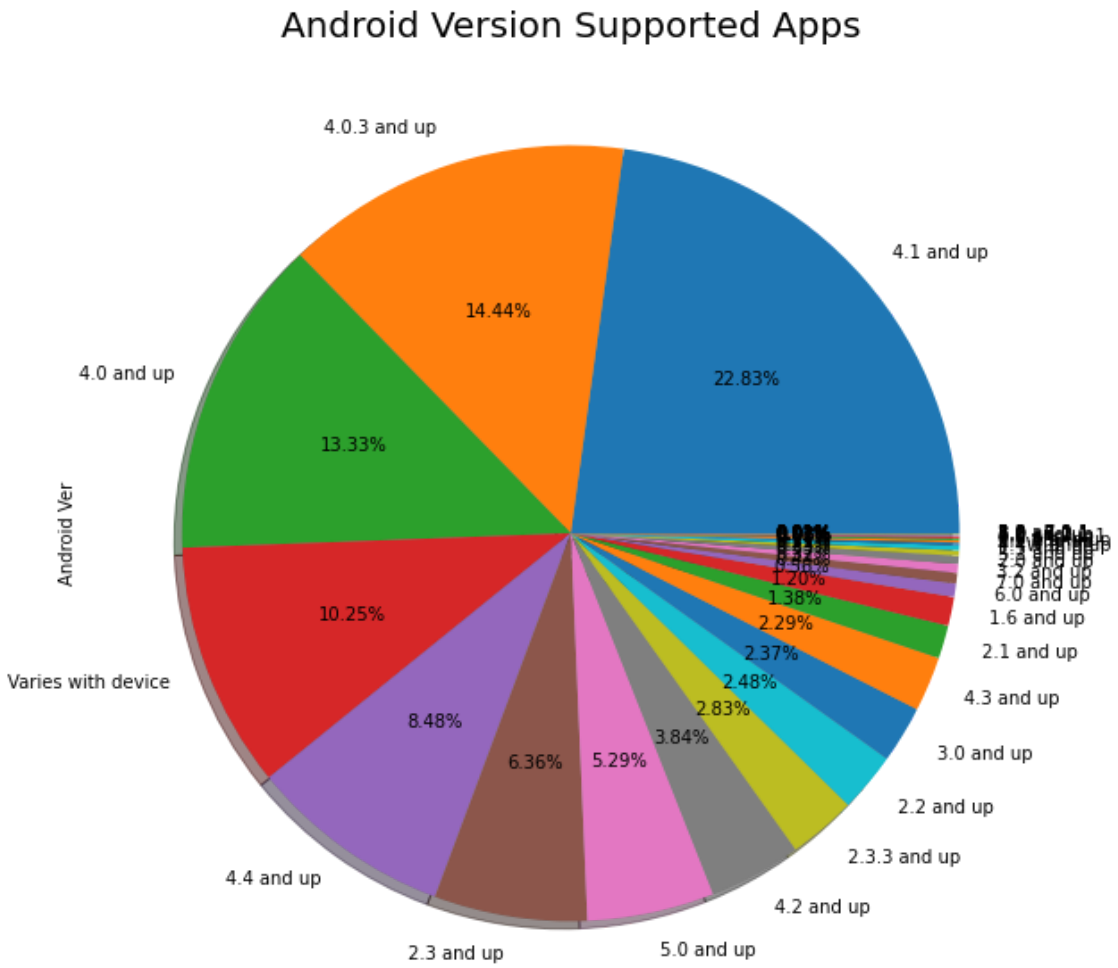
- **Step 5 :-** In this step solved all the 7 problems with the help of different graphs.
- **Step 6 :-** This is the last step in which we conclude the project details and the data which we clean and visualize.

3. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.

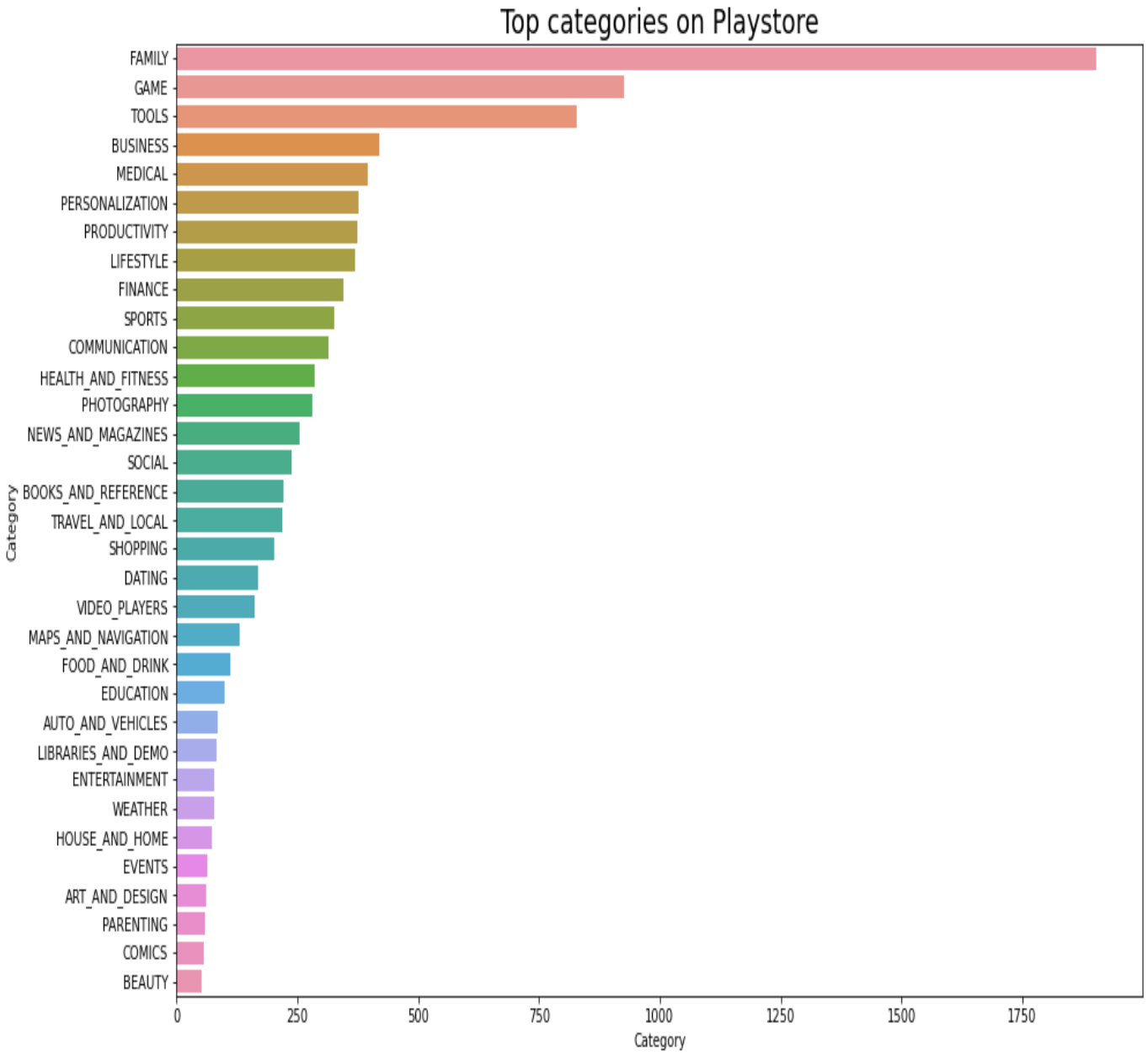
EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. In this project, we will understand EDA with the help of an example dataset. We will use **Python** language (different **Pandas** library) for this purpose.

3.1 Android Version Supported Apps



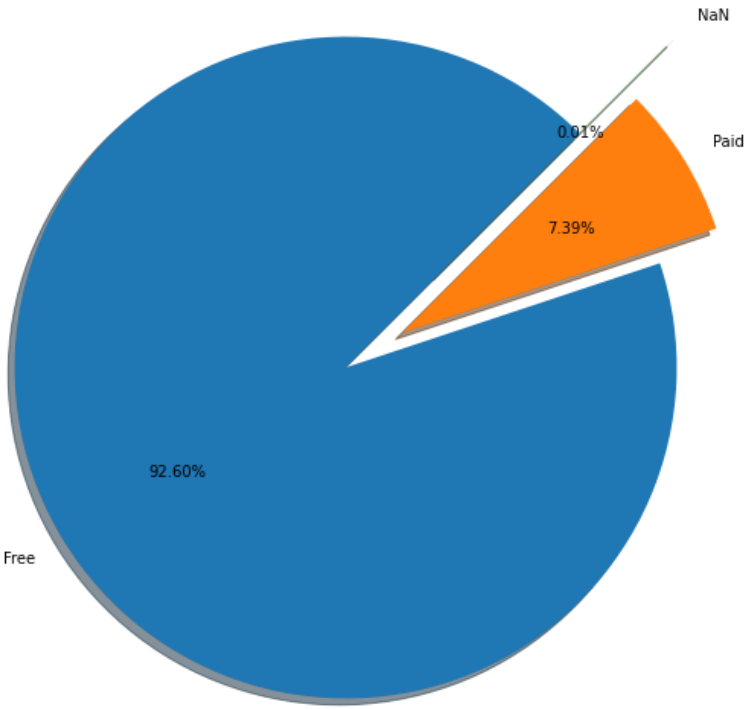
After identifying the total distribution percentage on data, given details of more app supported Android OS versions. Basically android 4.0 and above version supported app ratio is very high and more than 60% app's support only on android 4.0 and above version.

3.2 Categories of app in play store



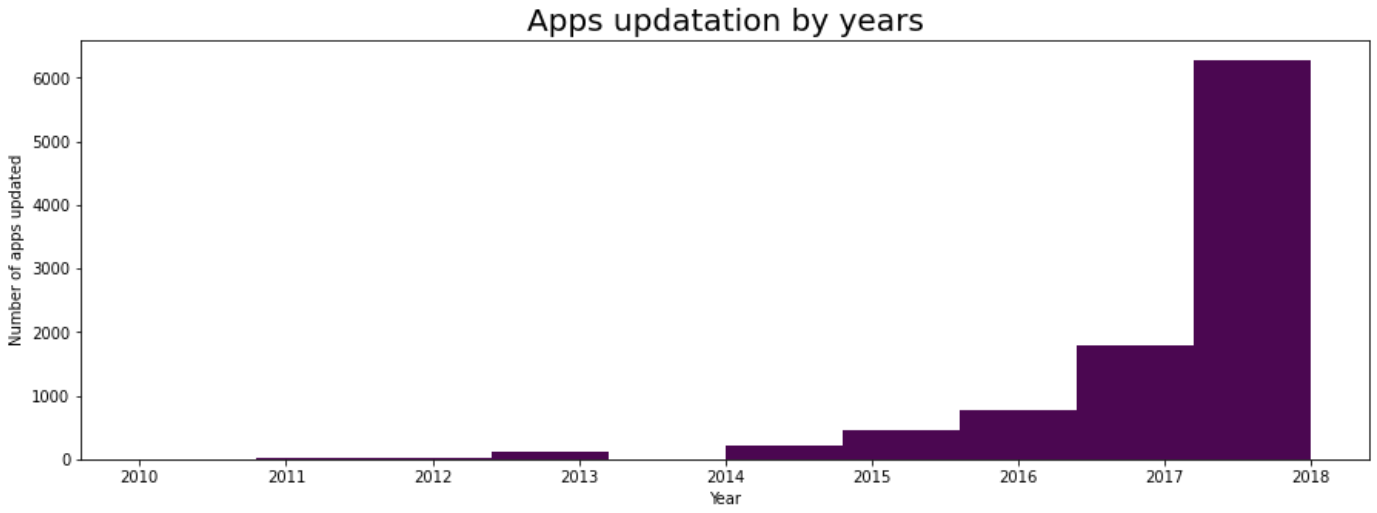
There are 33 categories of apps on play store datasets. We can clearly see that the top category of apps that customers like most are **family apps** and **gaming apps**. Approximately 1750 apps for the family are available on play store. Nearly 1000 gaming categories are there on the Google play store platform.

3.3 PAID v/s FREE APPS



There are 92.60% apps that are free available on the playstore which goes up to 8275. And 7.39% apps are those which are available at some cost i.e. they are paid apps. Total paid apps are 611.

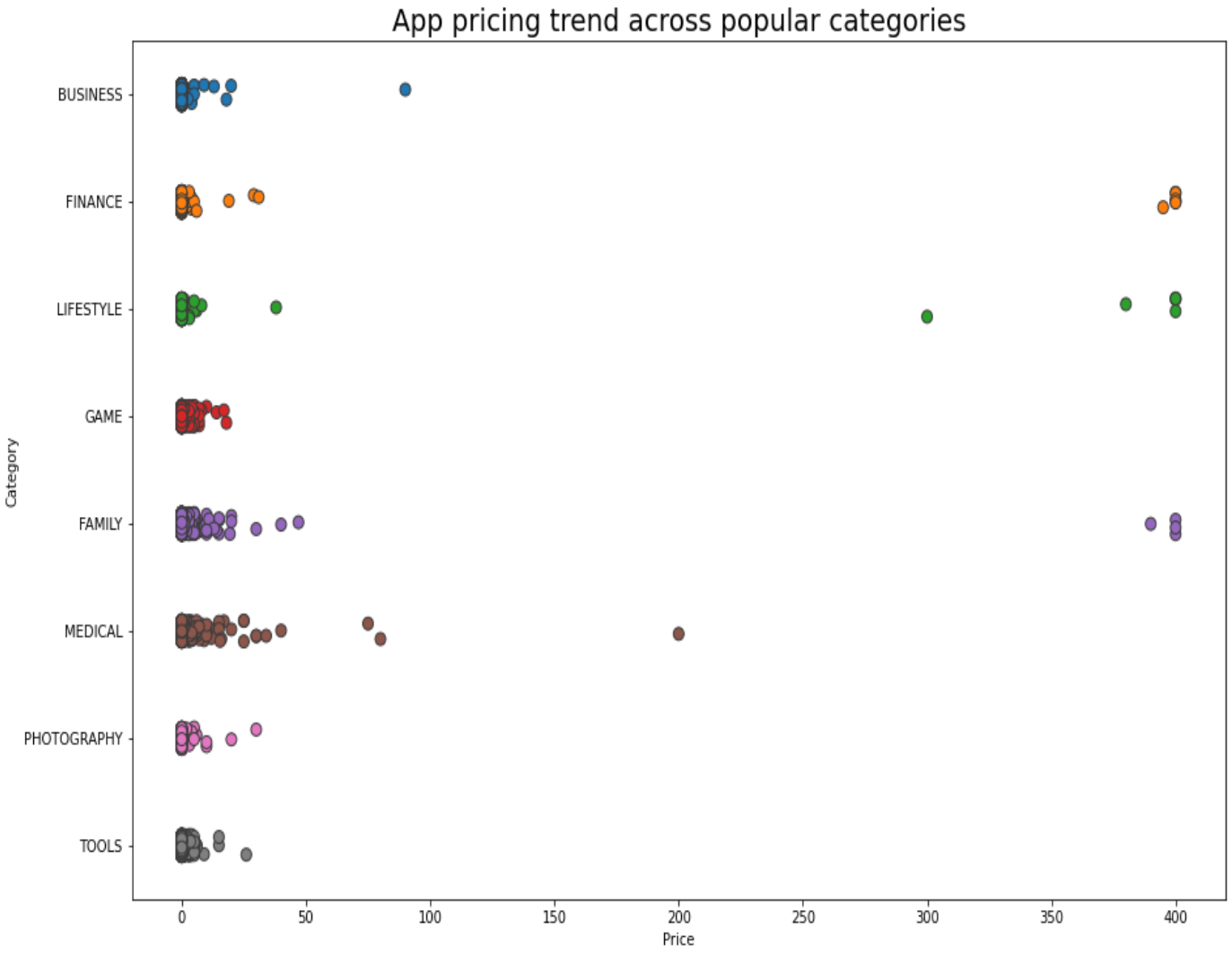
3.4 UPDATION DETAILS OF APPS BY YEAR



This is the graph of no of apps updated v/s app updation by years. This graph shows that there are 6000+ apps that have the latest updated version in 2018 which is the maximum value among all the previous year. Only a few

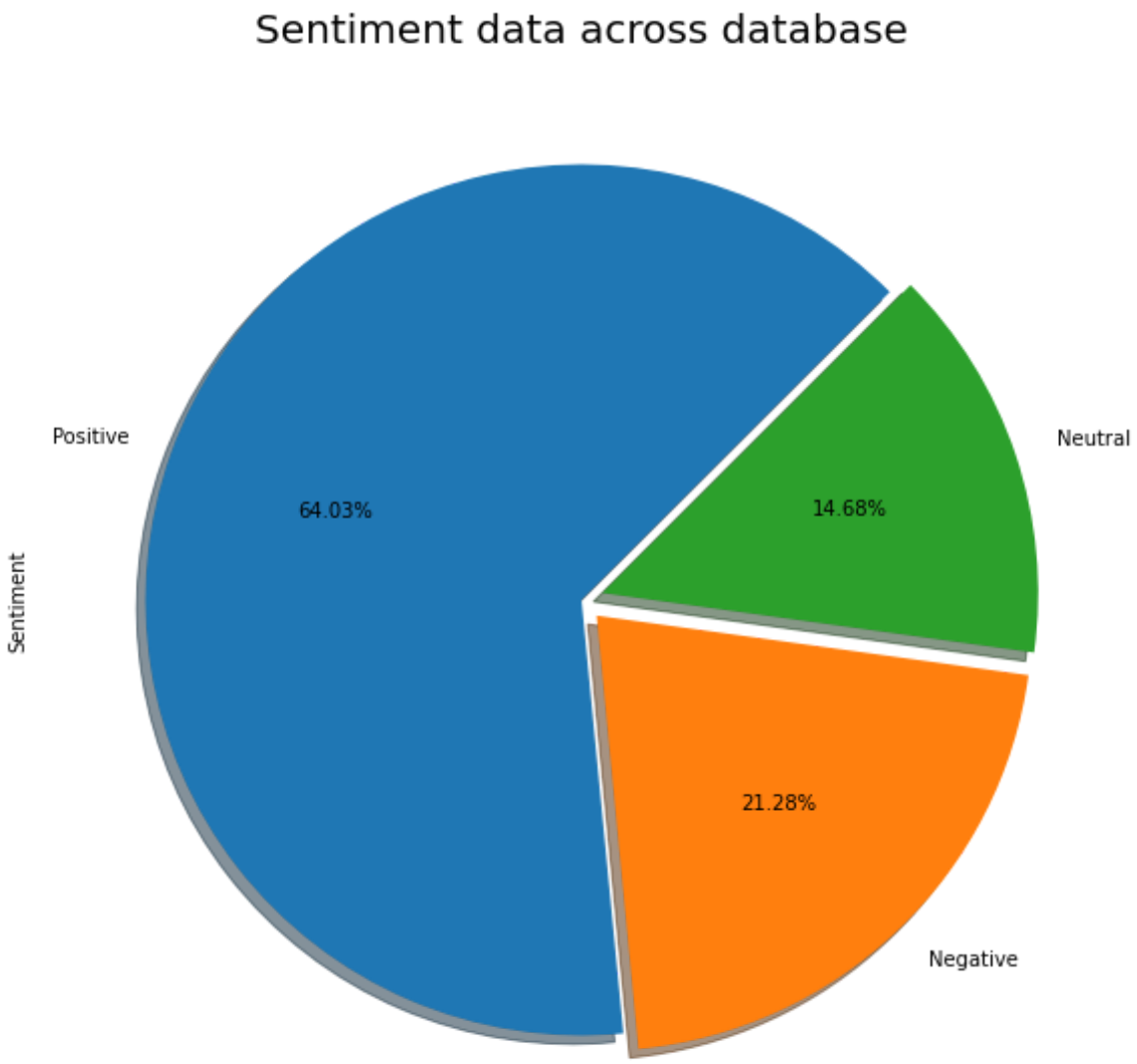
hundred apps were updated in 2011 which was very low. But as the graph shows, the app's updating ratio increases year by year.

3.5 App pricing in popular categories



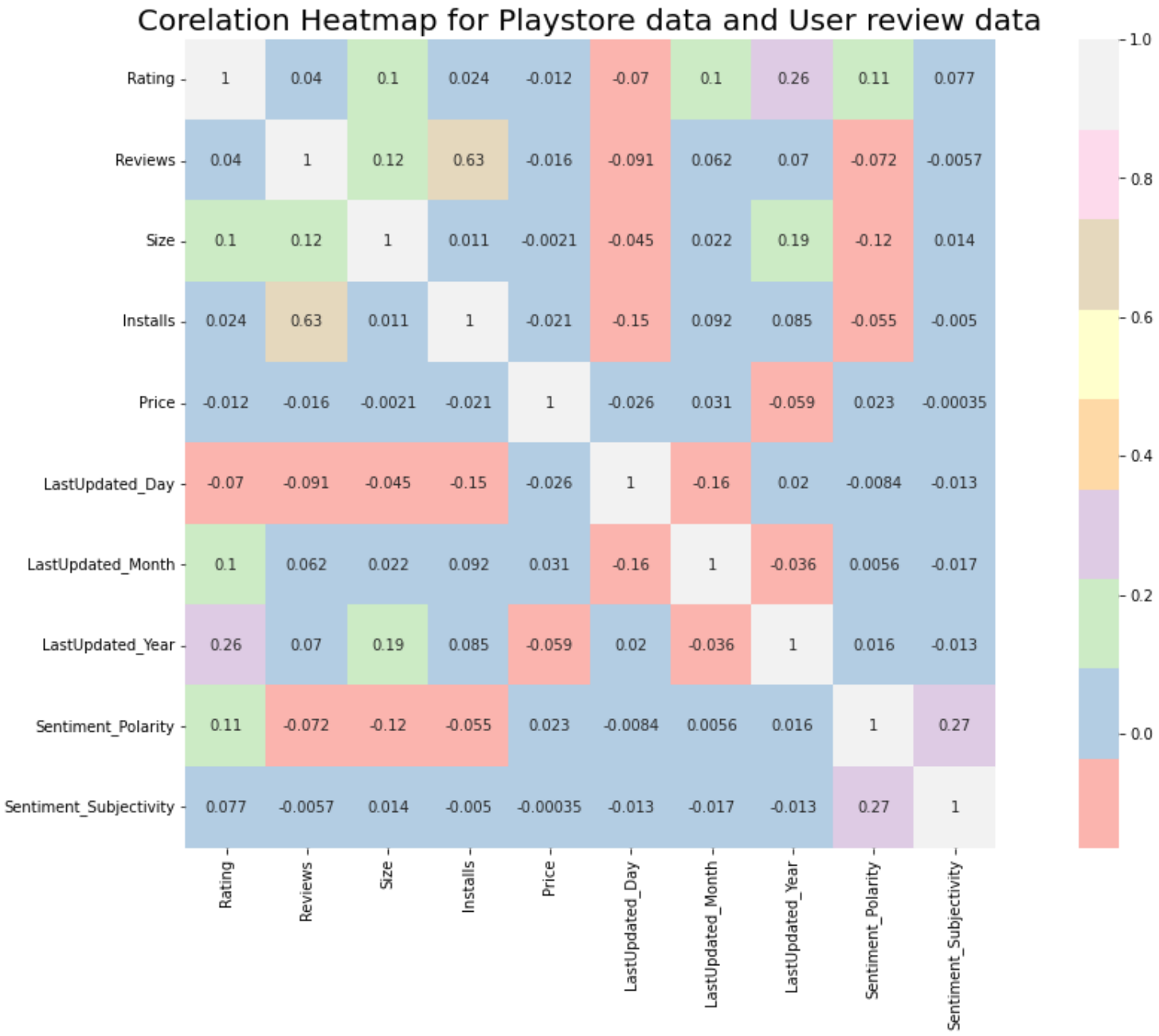
This scatter plot graph shows the relation between the category of apps and the price of different apps which are available on the play store. The maximum price of the apps goes up to 400 dollars and the minimum price may be 0.1 dollar. But we can see that most of the apps have low prices and only few are high priced. Family and medical apps are mostly priced apps as we can see in the graph.

3.6 SENTIMENTS ACROSS ALL REVIEWS



There are three sentiments in the app reviews. They are positive, negative, and neutral. To analyze the sentiments of the customers we made a pie chart for it. We can see 64.03% of customers are satisfied with the apps and they give positive feedback. But there are some customers who give negative feedback. The negative feedback customers are 21.28%.along with positive and negative feedback 14.68% customers stay neutral i.e. they don't give any feedback neither positive nor negative.

3.7 CORRELATIONS OF PLAY STORE DATA AND USER REVIEW DATA BASED ON DIFFERENT FACTORS



We add Both the data frames that are related to each other. In these correlation heat map some values are negative and some are positive, for that purpose we merge the both data frames to obtain the most appropriate results and yes, we observed that :

1. Size and sentiment polarity are negatively correlated (-0.12): There are lots of reason of disliking those apps which have large size. First of all, it consumes more storage; takes more RAM and needs a high speed connection for its execution.

2. There is a positive correlation between reviews and number of installs (0.63) because as the reviews increase, people start noticing the app and install them.

3. There is a slightly positive correlation (0.27) between sentiment polarity and sentiment subjectivity that means if the user's shares positive reviews then there is chance that users are sharing their personal opinion and not genuine information.

4. CONCLUSION

Through exploratory data analysis we have observed some trends and have made some assumptions that might lead to app success among the users in the play store.

- Android Version Supported Apps: - 60% app's support only on android 4.0 and above version.
- Categories of app in play store :- most like or preferred by customers family apps and gaming apps
- Percentage of free apps :- 92.60% apps are free available on the play store
- Percentage of Paid apps: - 7.39% apps are paid apps.
- Updation details of apps by year: - Latest 6000+ apps updated in 2018 which is the maximum count.
- App pricing in popular categories: - we see prices up to 400 dollar & mostly medical and family apps are paid.
- Sentiments of reviews: - 64.03% of customers gave positive feedback while the negative feedback is 21.28%.
- correlations of play store data and user review data based on different factors

5. REFERENCES

✔ https://www.geeksforgeeks.org/
✔ https://stackoverflow.com/
✔ https://docs.python.org/3/library/
✔ https://www.academia.edu/
✔ https://www.w3schools.com/python/

THANK YOU!!