Capstone Project

Exploratory Data Analysis On Google Play Store Apps Reviews.

Google play

PRESENTED BY

Ashish Mali
Data Science Trainee,
AlmaBetter

INDEX

INTRODUCTION

- As Internet hits the Indian market in 1995, has done an major impact on Indian economy in such a short period.
- In 2020, the Internet economy had grown to 16% of which 8% was driven by apps, In percent of India's GDP.
- The Internet economy contributed up to $537.4 billion to India's GDP in 2020, of which a minimum of $270.9 billion was contributed by apps. Apps were contributing 70% to the mobile traffic.
- India Internet is expected to cross triple digits GMV for the first time in 2021 and eventually become $250 billion scale and 10% of private consumption in 2025.
- India has 504 million active Internet users in 2020.
- Nearly 70% of the active Internet population in India are daily users.
- As of 2020, Android held a share of 95.23 percent of the mobile operating system market in India.
- Google Play Store is the largest app distribution platform owning 97 per cent market share in India.

NEED OF PLAY STORE EDA

- About 80% of the mobile users uses the mobile devices and apps to accomplish their day to day needs
- India has been rated as the world's third largest mobile app user.
- According to Statista, India's app market revenue is expected a compounded annual growth rate (CAGR) of 9.2%.
- The Progressive Policy Institute (PPI) expects India to overtake the US as the country with the largest developer population by 2024.

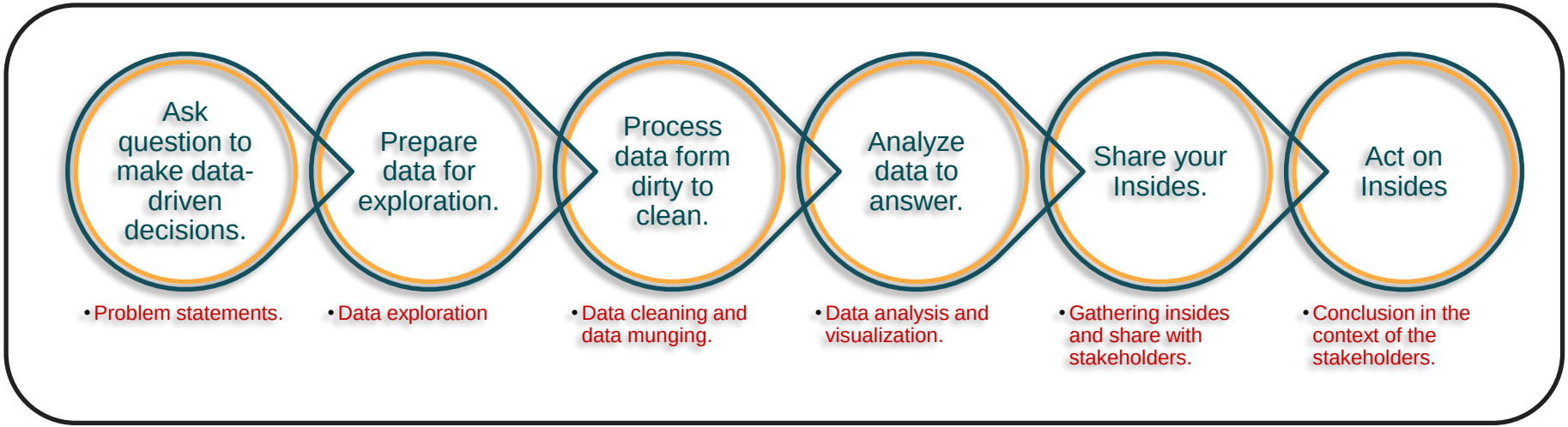
With a 14% yearly market growth since 2016, India is the fastest growing smartphone market globally.

In 2021, Android held a share of 95.84 percent of the mobile operating system market in India.

India had almost **560 million internet users**. The download almost 12.3 billion apps and spend 17 hours on the internet. (survey conducted in 2018)

In the first half of 2021, Google Play Store generated approximately $23.4 billion from in-app purchases, subscriptions, and premium apps and games.

➢ FRAMEWORK FOR ANALYSIS



- We analyze using this framework. As we progress through the phases, we get closer to the desired inside.
- We are now moving forward as the framework's steps progress.

Before we Begin…

- The Data file we have…
  - play store data.csv
  - user review data.csv
- Given Problem statement
  - We have to perform EDA On the both the datasets and find Out the factors responsible for successful application.
- The Context we have
  - We have to perform analysis taking the context of Developer.

Let's first know the things which Are in our hand and then move Forward towards the analysis.

➢ ASK QUESTION TO MAKE DATA DRIVEN DECISIONS

Q.1) Find out the number of application per category Also, find out the category which having highest number of installs.?

Q.2) Find out the top categories on play store based on the rating. Also analyze the Rating attribute.?

Q.3) Find out the application which having highest number of user engagement in the category which having highest number of application. Also fetch top 10 apps in any category based on the user engagement.?

Q.4) Among all the apps fetch maximum and minimum number of installed apps, query all it's attributes?

Q.5) Find out the percentage of free apps over a paid apps? Also check the distribution of “Content_Rating”.?

Q.6) Explore the Paid attribute and find the app which generated the highest revenue.?

Q. 7) Find out the percentage of paid apps available in each category, Also find out the category which generate the maximum revenue.?

Q.8) Find out the effect of application rating over it's user engagement.?

Q.9) Explore the Bivariate relationship between the attributes of the play store dataset.?

Q.10) Explore the relationship between Size and Rating attribute with respect to various category's. Also find if there is any relationship between Size and Installs.?

Q.11) Analyze the Distribution of Free and Paid apps updated over the Month.?

Q.12) Is frequent update has affect the rating. Also analyze the distribution of apps update over the years.?

Q.13) Explore the correlation between the different attribute of the play store data. Also check is there any correlation between the play store dataset and user reviews dataset.?

Q.14) Determine which type of sentiment dominates the most while reviewing the application.?

Q.15) Find the apps which has the maximum number of positive reviews.?

Q.16) Find the apps which has the maximum number of negative reviews.?

Q.17) Find out which reviews are the most, fact based or opinion based.?

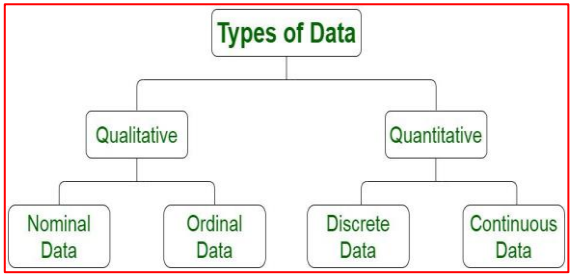
Q.18) Is Sentiment_Subjectivity proportional to Sentiment_Polarity.?

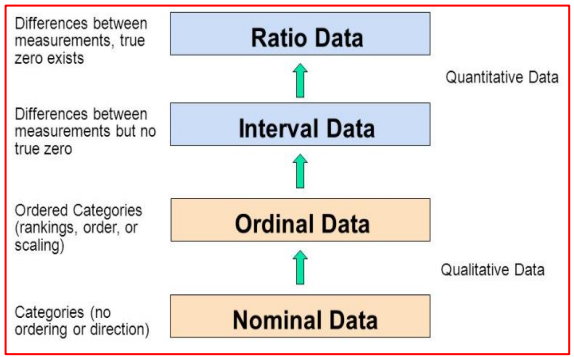
➢ PREPARE DATA FOR EXPLORATION

☛ The Flavors Of Data And The Scales Of Data Measurement.

- ⮡ Data which is the form of strict row/columns called **Structured Data.**
- ⮡ Data which is present in audio, images etc are **Unstructured Data.**

- ⮡ Data which in the form of text or string called **Qualitative Data.**
- ⮡ Data which is in the form of numeric values called **Quantitative Data.**





- ⮡ **The nominal level** :- Names Only.
- ⮡ **The ordinal level** :- Has an Order.
- ⮡ **The interval level** :- Also has a meaningful distance.
- ⮡ **The ratio level** :- Extension of interval level and has meaningful zero.

➢ PREPARE DATA FOR EXPLORATION

- In this phase we explore the data in detail and prepare it for the exploration.
- Let’s find out what type of data that each attribute holds in play store dataset.
  - **App -** represent the name of the application - **Qualitative data, & Scale is Nominal.**
  - **Category -** broad section of useability of the application - **Qualitative data, & Scale is Nominal.**
  - **Rating -** overall likeness of the app using numbers - **Quantitative data, & Scale is Ordinal.**
  - **Reviews -** number of people revived that particular application - **Quantitative data, & Scale is Interval.**
  - **Size** - size of application in megabits (unit of measurement) - **Quantitative data, & Scale is Interval.**
  - **Installs -** number which indicate how many time the application is downloaded on any device - **Quantitative data, & Scale is Interval.**
  - **Type -** represent whether it's paid or free - **Quantitative data, & Scale is Nominal.**
  - **Price -** purchase price of app if it's paid - **Quantitative data, & Scale is Ratio level.**

➢ PREPARE DATA FOR EXPLORATION

- **Content_Rating -** minimum age recommended to use application - **Qualitative data, & Scale is Nominal.**
- **Genres** A sub category for the app - **Qualitative data, & scale is Nominal.**
- **Last_Updated –** holds the date-time value when app got update - **Quantitative data, & Scale is Ratio level**

Let's do the same with the user review dataset

- **App -** name of the application - **Qualitative data, & Scale Nominal.**
- **Translated_Review -** string for text as review - **Qualitative data, & Nominal.**
- **Sentiment** typically contain three values - **Qualitative data, & Scale is Ordinal.**
- **Sentiment_Polarity -** A fixed range of integer values - **Quantitative data, & Scale is Ratio level.**
- **Sentiment_Subjectivity -** Also, a fixed range of integer values - **Quantitative data, & Scale is Ratio level.**

➢ PROCESS DATA FORM DIRTY TO CLEAN

- **Handling Nan/Missing values**
- “Category”, “Rating”, “Type”, “Genres”, “Current_ver” and “Android_ver” contains null values.
- Maximum null values are present in “Rating” attribute which account to 13.6% of total record.
- Minimum null values are present in “Category”, “Type” and “Genres” these all three contains 1 null values each.

“Category” and “Genres” attribute

- Both contains 1 null value each.
- Percentage impact is 0.09% only
- We drop the records that corresponds to null values.

“Type” attribute

- Contains only one null value
- Corresponding “Paid” attribute shows zero means it’s a free app.
- Replace the NaN value with “Free”

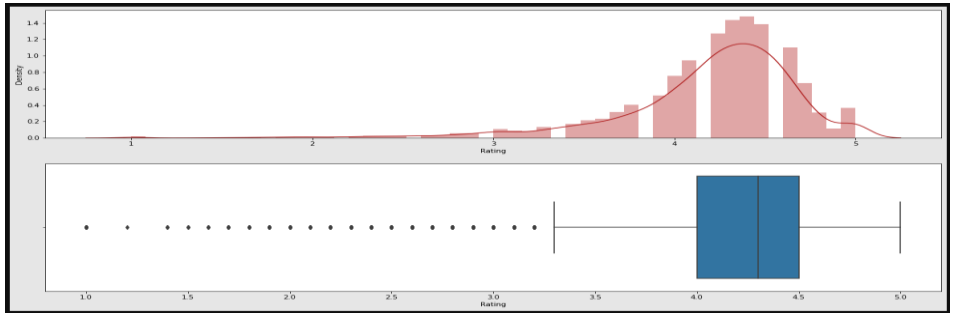
“Current_Ver” and “Android_Ver” attribute

- No particular value from which we can replace the NaN value.
- The percentage impact is just 0.074% and 0.018%.
- We dropped the NaN values for these attributes as well.

➢ PROCESS DATA FORM DIRTY TO CLEAN

“Rating” attribute

- Attribute contains 13.6% of NaN values, so dropping the records is costly operation.
- Using distplot we see the distribution of rating and box plot for the outliers.



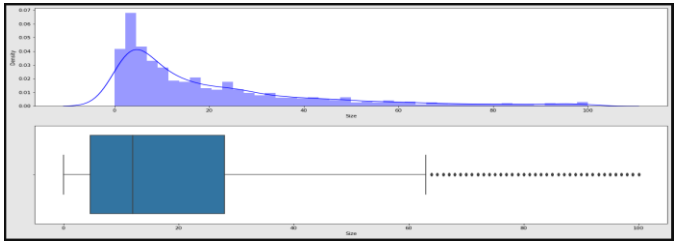
.

- From the distplot visualizations, it is clear that the ratings are left skewed.
- We know that if the variable is skewed, the mean is biased by the values at the far end of the distribution. Therefore, the median is a better representation of the majority of the values in the variable.
- The median of the entries (excluding the NaN values) in the 'Rating' column comes to be 4.3
- Hence we will impute the NaN values in the Rating column with its **median.**

➢ PROCESS DATA FORM DIRTY TO CLEAN

↪ Handling Duplicate, Data type, anomaly in the Dataset

- The “App” attribute records contains 483 duplicated values with accounts to 5% of total records.
- Except “Rating” attribute which has a data type of “float64”, All other attribute has a data type of “object” (String).
- For “Size” attribute we have the symbol like “MB” or “KB” to show the size of apps. We remove them and converted to single unit of measurement called megabits.
- In the “Size” attribute we also have 12.7% of records which carries “Varies with device” value. So we have to handle them.



- It is clear from the above two diagrams that the Size attribute values are skewed towards right.
- replacing those values with any of the central tendency, may cause bias analysis. Therefor, it is better idea to drop the these records.

➢ PROCESS DATA FORM DIRTY TO CLEAN

↪ **Handling Duplicate, Data type, anomaly in the Dataset**

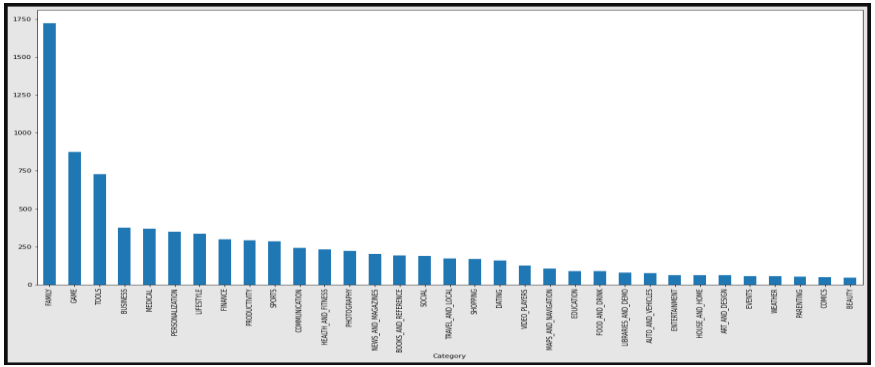
- The “Installs” attribute contains the values with special symbol like “+” & “,”. Also correct it’s data type
- The “Price” attribute also contains the special symbol like “$” & “,”. Also correct it’s data type.
- The “Content_Rating” attribute also contain the special symbol like “+”
- The “Last_Update” attribute data type converted to date-time.
- For this analysis we don’t required the “Current_Ver” attribute, so we dropped it.
- The final number of records present after cleaning the dataset are 8422, and and having 12 features.

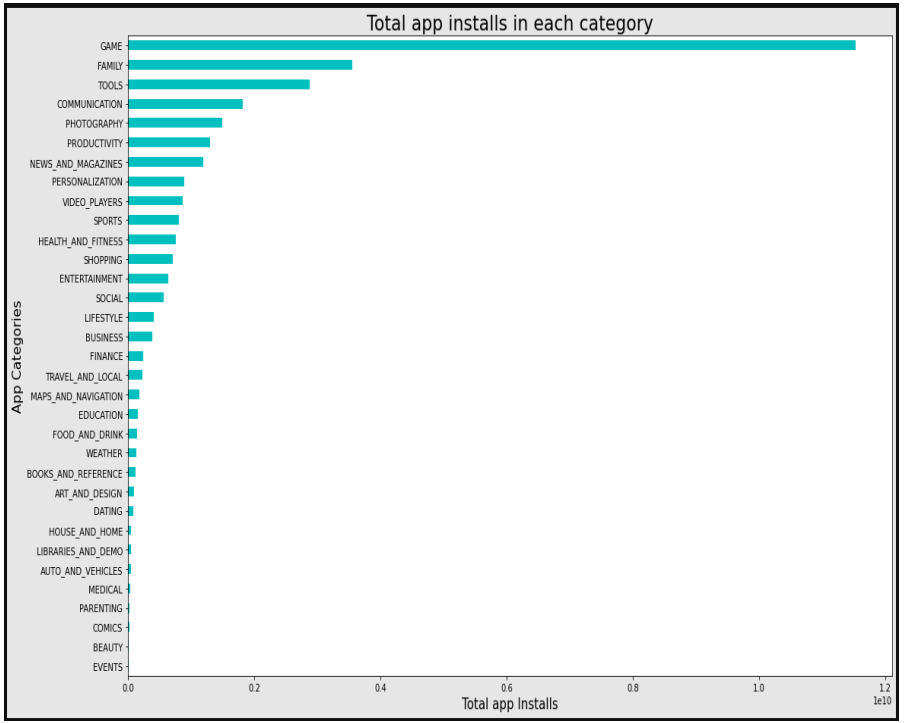
The user review dataset

- There are a total of 26868 rows containing NaN values in the Translated_Review attribute. (41.7% of total)
- We can say that the apps which do not have a review (NaN value insted) tend to have NaN values in the columns Sentiment, Sentiment_Polarity, and Sentiment_Subjectivity in the majority of the cases.
- There are few exceptions for our assumption. But these records are accounted as an error, because for the "Sentiment", "Sentiment_Polarity" and "Sentiment_Subjectivity" there is a mandatory condition that the values corresponding to the "Translated_Review" **must be non-null**. As there remaining three are the dependent variable on the "Translated_Review" attribute.
- It's better to drop all the "NaN" values corresponding to the dataset.

➢ ANALYSE DATA TO ANSWER

1. Top category with most apps & No of Installs per category.

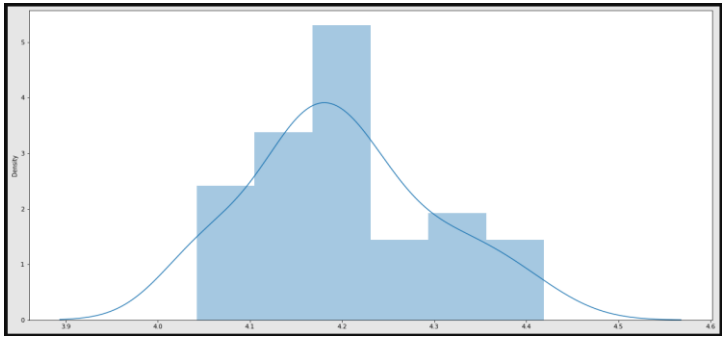


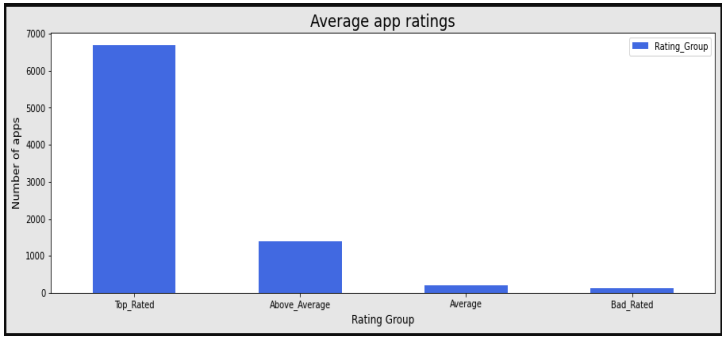


- There are all total **33 categories** in the dataset from the analysis we can come to a conclusion that in play store most of the apps are under **FAMILY & GAME** category and least are of **EVENTS & BEAUTY** Category.
- The **Game, Family and Tools** categories has the highest number of installs compared to other categories of apps.
- The category which having the maximum number application present is FAMILY. Whereas the GAME category having the highest number of installed application.

ANALYSE DATA TO ANSWER

2. Distribution of rating and average app rating.

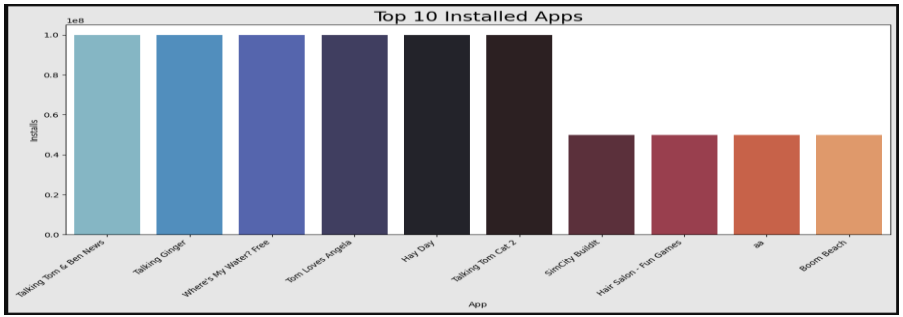


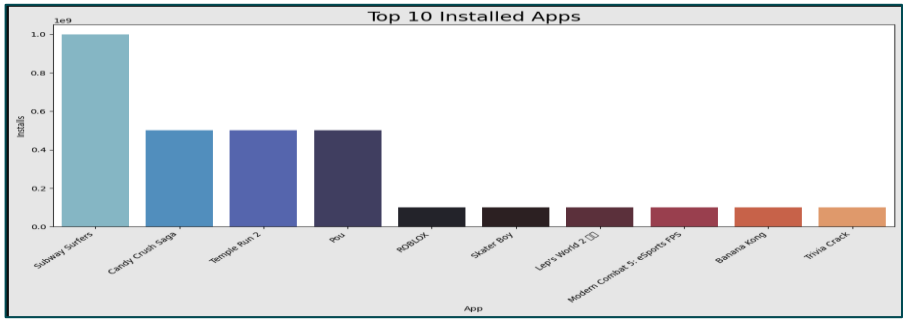


- The mean of the average ratings (excluding the NaN values) comes to be 4.2.
- The median of the entries (excluding the NaN values) in the 'Rating' column comes to be 4.3. From this we can say that 50% of the apps have an average rating of above 4.3, and the rest below 4.3.
- From the visualizations, it is clear that the ratings are left skewed.
- We know that if the variable is skewed, the mean is biased by the values at the far end of the distribution. Therefore, the median is a better representation of the majority of the values in the variable
- 4-5: Top rated
- 3-4: Above average
- 2-3: Average
- 1-2: Below rated

ANALYSE DATA TO ANSWER

3. Top app in top category and top ten apps in any category.





- We know that top category is “FAMILY”.so fetching top 10 apps in that category.
- The application which are highest in terms of user engagement are "Free" and having the content rating to everyone means no age limit.
- one the most important factor which highlighted here is that the gamification of the application draws more user compared to rest ones.
- Top 10 installed apps in “GAME” category.

1) Subway surfers
2) Candy crush saga
3) Temple run 2
4) Pou
5) Roblox
6) Skater boy
7) Lep’s world 2
8) Modern combat 5: esports FPS
9) Banna kong
10) Trivia crack

➢ ANALYSE DATA TO ANSWER

4. Compare max and min app installs and query all its attribute.

❑ Maximum no of times installed application.

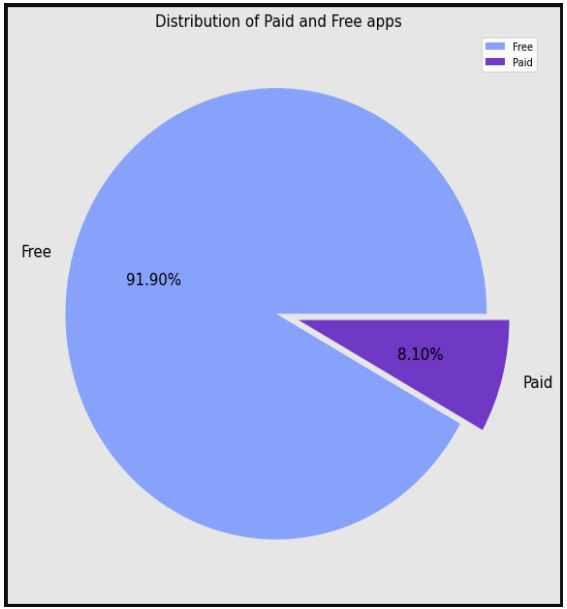
	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content_Rating	Genres	Last_Updated	Android_Ver	Rating_Group
1654	Subway Surfers	GAME	4.5	27722264	76.0	1000000000	Free	0.0	Everyone 10	Arcade	2018-07-12	4.1 and up	Top_Rated
3736	Google News	NEWS_AND_MAGAZINES	3.9	877635	13.0	1000000000	Free	0.0	Teen	News & Magazines	2018-08-01	4.4 and up	Above_Average

❑ Minimum no of times installed application.

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content_Rating	Genres	Last_Updated	Android_Ver	Rating_Group
4465	Popsicle Launcher for Android P 9.0 launcher	PERSONALIZATION	4.3	0	5.5	0	Paid	1.49	Everyone	Personalization	2018-07-11	4.2 and up	Top_Rated
5307	Ak Parti Yardım Toplama	SOCIAL	4.3	0	8.7	0	Paid	13.99	Teen	Social	2017-07-28	4.1 and up	Top_Rated
5486	AP Series Solution Pro	FAMILY	4.3	0	7.4	0	Paid	1.99	Everyone	Education	2017-07-30	4.0 and up	Top_Rated
5945	Ain Arabic Kids Alif Ba ta	FAMILY	4.3	0	33.0	0	Paid	2.99	Everyone	Education	2016-04-15	3.0 and up	Top_Rated
6692	cronometra-br	PRODUCTIVITY	4.3	0	5.4	0	Paid	154.99	Everyone	Productivity	2017-11-24	4.1 and up	Top_Rated
7434	Pekalongan CJ	SOCIAL	4.3	0	5.9	0	Free	0.00	Teen	Social	2018-07-21	4.4 and up	Top_Rated
8081	CX Network	BUSINESS	4.3	0	10.0	0	Free	0.00	Everyone	Business	2018-08-06	4.1 and up	Top_Rated
8614	Sweden Newspapers	NEWS_AND_MAGAZINES	4.3	0	2.1	0	Free	0.00	Everyone	News & Magazines	2018-07-07	4.4 and up	Top_Rated
8871	Test Application DT 02	ART_AND_DESIGN	4.3	0	1.2	0	Free	0.00	Everyone	Art & Design	2017-03-14	4.2 and up	Top_Rated
9337	EG | Explore Folegandros	TRAVEL_AND_LOCAL	4.3	0	56.0	0	Paid	3.99	Everyone	Travel & Local	2017-01-22	4.1 and up	Top_Rated
9719	EP Cook Book	MEDICAL	4.3	0	3.2	0	Paid	200.00	Everyone	Medical	2015-07-26	3.0 and up	Top_Rated
9905	Eu sou Rico	FINANCE	4.3	0	2.6	0	Paid	30.99	Everyone	Finance	2018-01-09	4.0 and up	Top_Rated
9917	Eu Sou Rico	FINANCE	4.3	0	1.4	0	Paid	394.99	Everyone	Finance	2018-07-11	4.0.3 and up	Top_Rated
9934	I'm Rich/Eu sou Rico/أنا غني/我很有錢	LIFESTYLE	4.3	0	40.0	0	Paid	399.99	Everyone	Lifestyle	2017-12-01	4.1 and up	Top_Rated

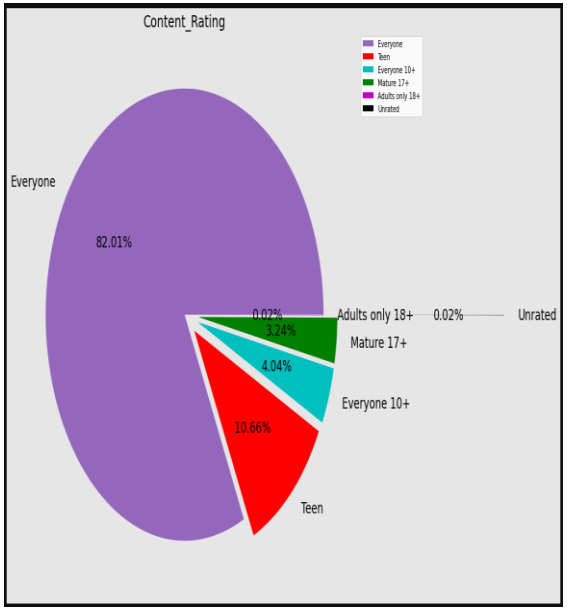
➢ ANALYSE DATA TO ANSWER

5. Distribution of free vs paid & content rating.



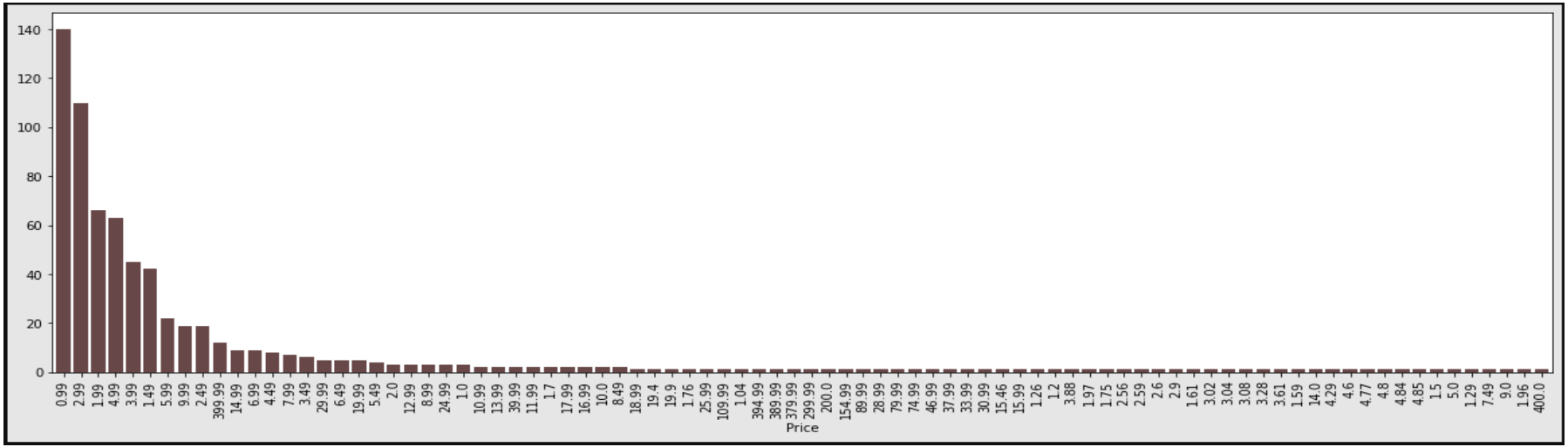
- Around 92% apps where free and the 8% apps where paid.

- Around 82% apps where rating of "Everyone". which accounts highest in terms of percentage.



➢ ANALYSE DATA TO ANSWER

6. Exploring the paid apps attribute.

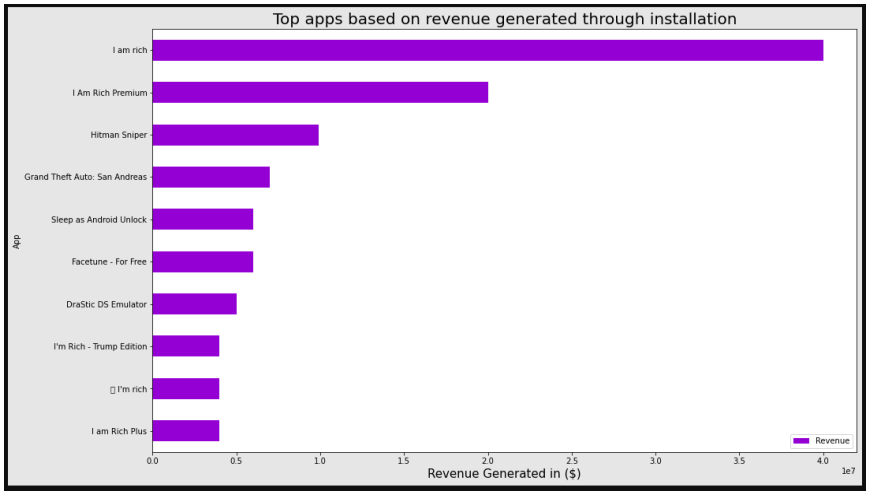


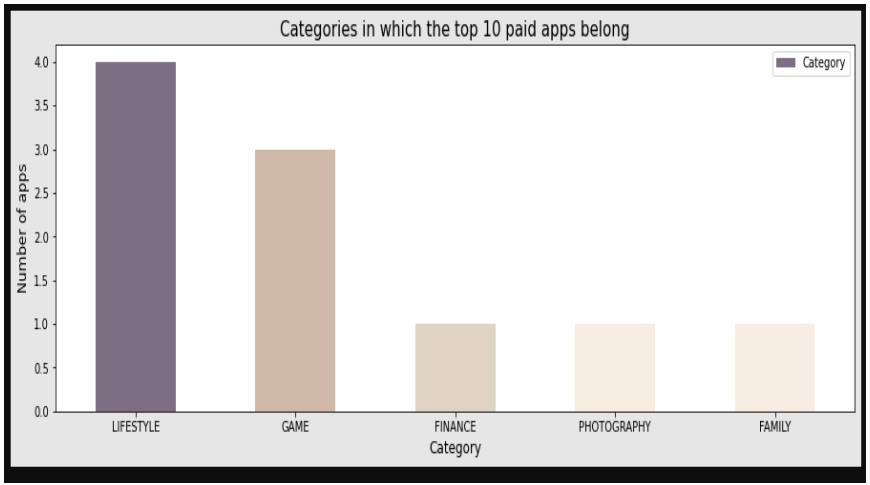
- ❑ Price vs number of applications.
- ❑ Price of the application varies form 0.99 – 400 USD.

➢ ANALYSE DATA TO ANSWER

6. Exploring the paid apps attribute.

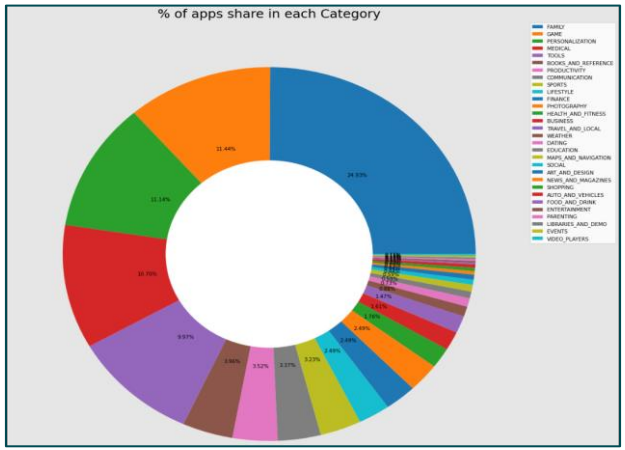
- A better way to determine the top apps in the paid category is by finding the revenue it generated through app installs here we also assume that **number of installations = paid app user.**
- **Revenue generated through installs = (Number of installs) x (Price to install the app)**





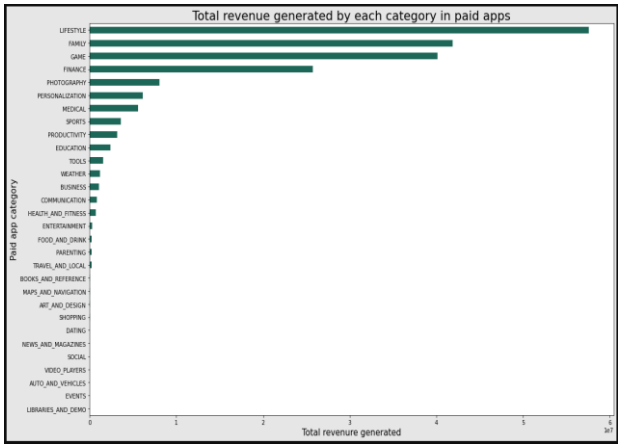
ANALYSE DATA TO ANSWER

7. Finding out the category wise paid app distribution and revenue generated by each category.



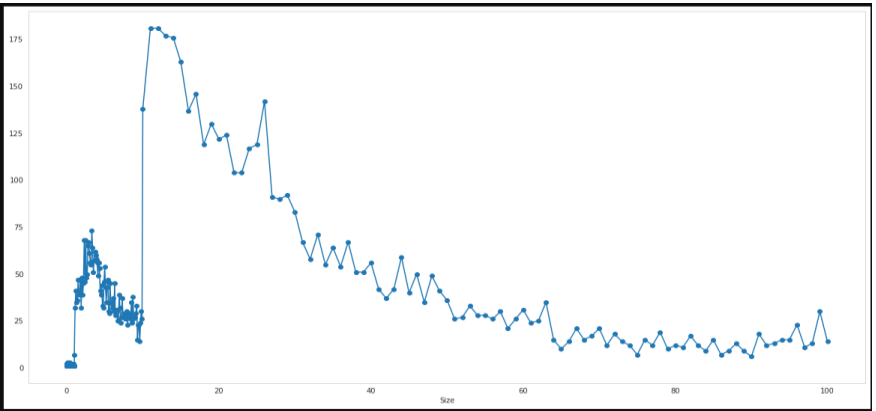
- FAMILY and GAME having the greatest number of paid applications
- Lest are present in EVENTS and VIDEO PLAYERS.

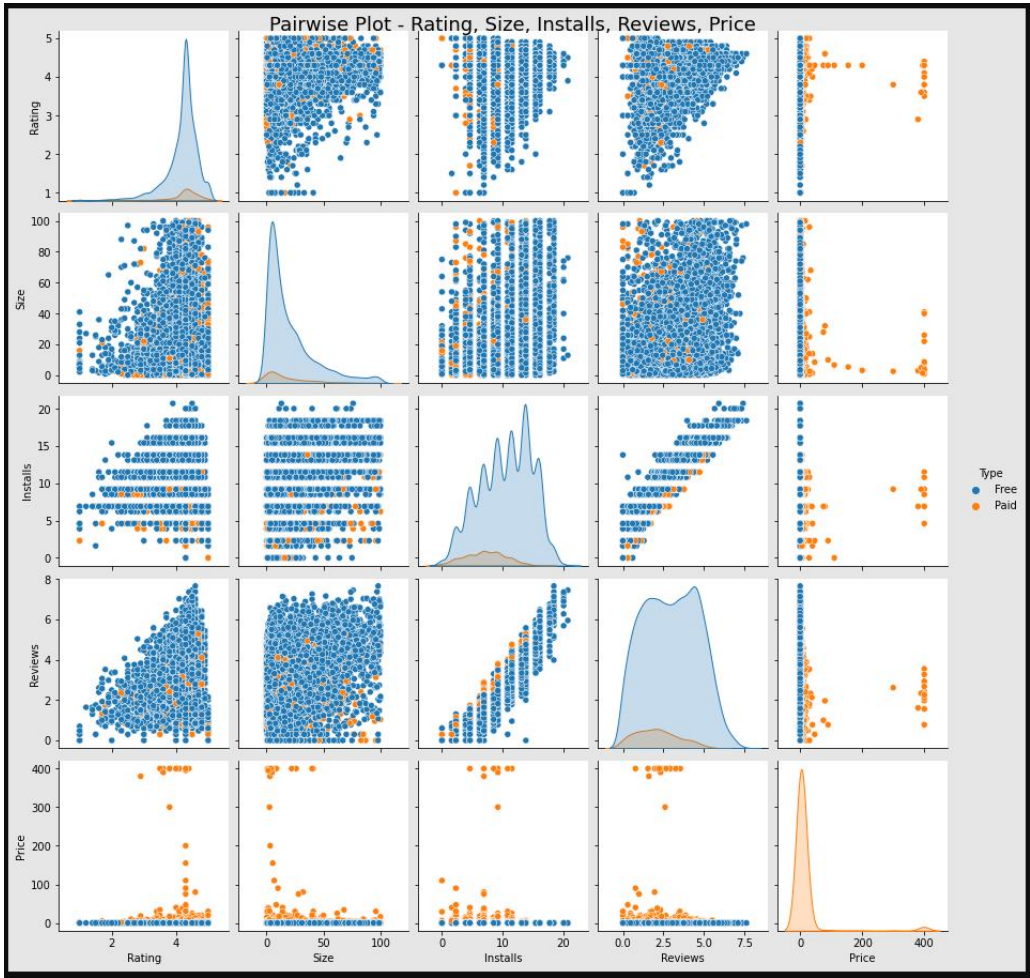
- The "LIFESTYLE" category dominated with respect to revenue collected which is by the way only have 2.5% of the total applications. This is may be because the case that the application under it charges the more in terms of money.



8. Bivariate relationship between play store attribute

- As we know there is no direct linear-relationship between install and rating but, we can say that as number of installs increase their increase in the rating of the application.
- Size of the application can affect the rating as less the size is more the rating of the application.
- Most of the apps are light-weight.
- We also see that the greater the user engagement the greater the rating. so, there is direct relationship between them.
- The inside we can drown form size vs installs comparison is that as size increase the number of installs decreases. and the best ideal size for application is under 20MB.





9. Exploring the relationship between size and installs attribute

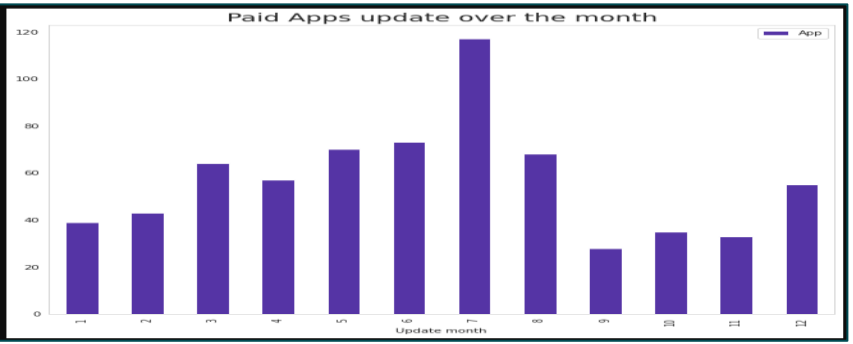
- Upon observing the pair plots, following categories of application show negative trend. Which means that the rating of the application will decreases as the size increases.
    1. NEWS_AND_MAGAZINES
    2. TOOLS
    3. TRAVEL_AND_LOCAL
    4. SHOPPING
    5. LIFESTYLE
    6. HOUSE_AND_HOME
    7. FOOD_AND_DRINK

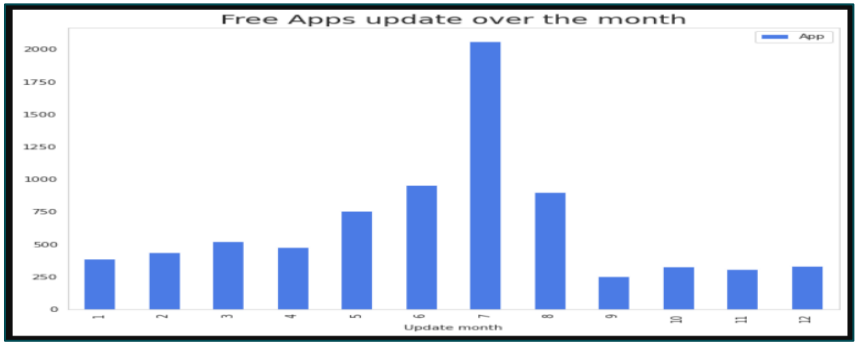
- The FAMILY category which holds maximum number of applications tends to have a neutral with respect to size.

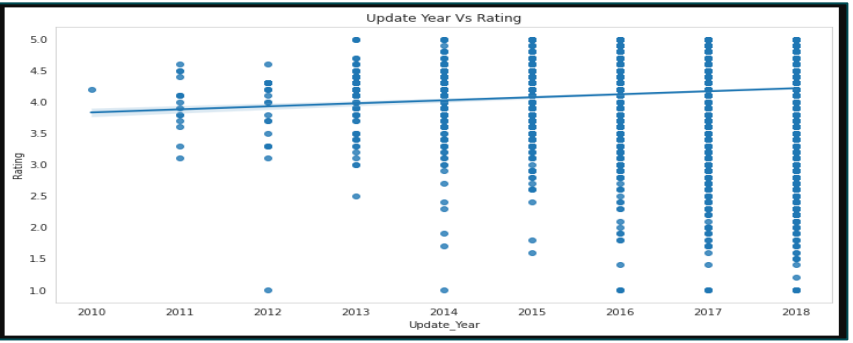
- Interesting fact about a COMIC category is that as size increase the rating of the application increase

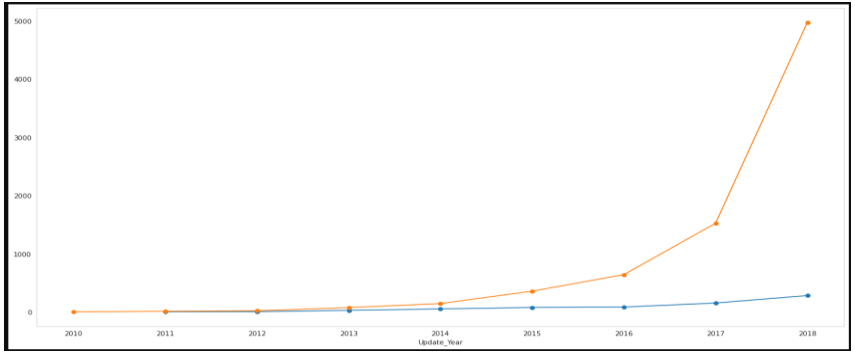
ANALYSE DATA TO ANSWER

10. Analysis of app update over rating in the span month & one year



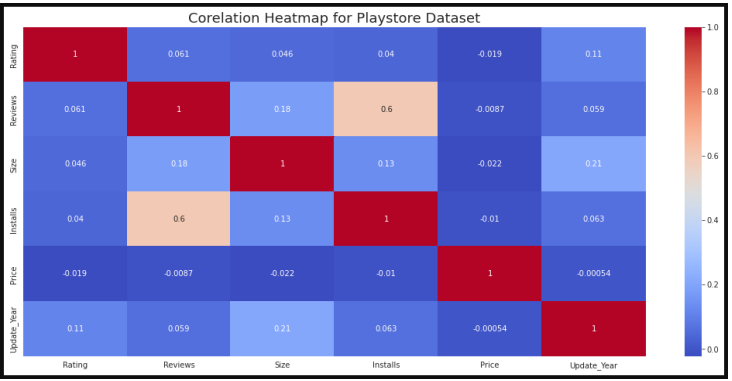


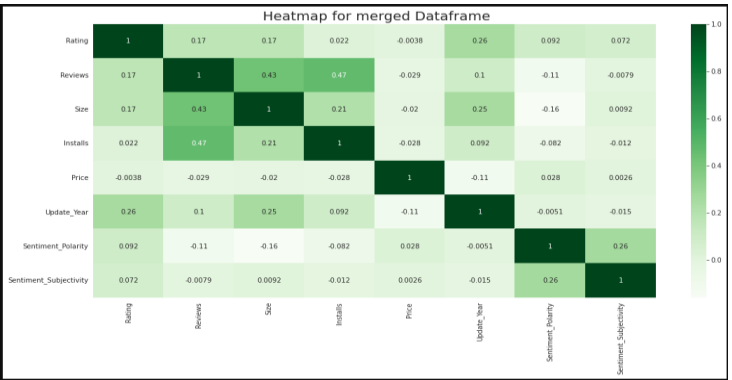




ANALYSE DATA TO ANSWER

11. Correlation heat maps for play store and user review dataset

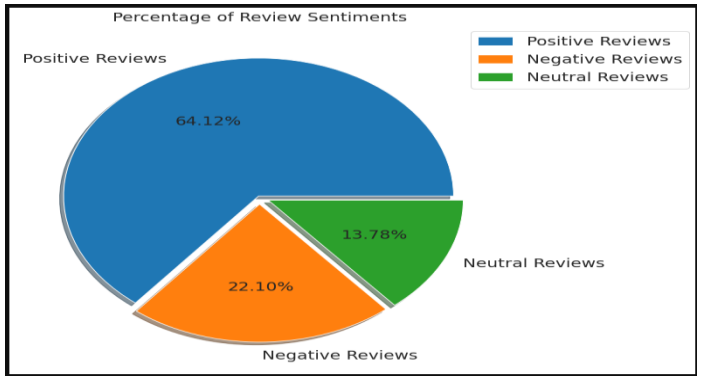


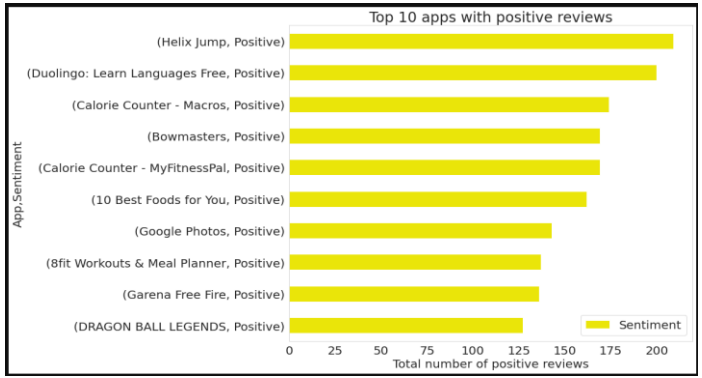


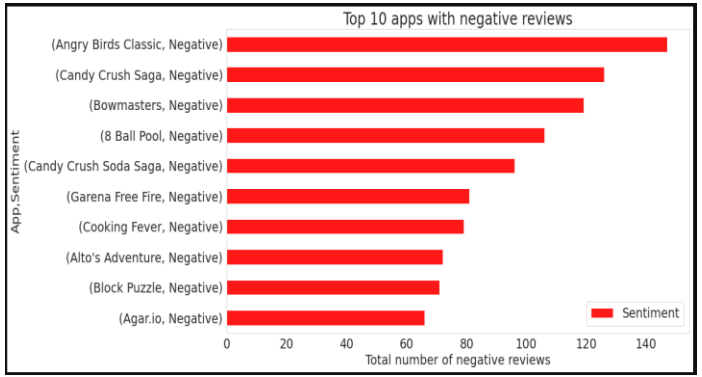
- There is a strong positive correlation between the Reviews and Installs column. This is pretty much obvious. Higher the number of installs, higher is the user base, and higher are the total number of reviews dropped by the users.
- The Price is slightly negatively correlated with the Rating, Reviews, and Installs. This means that as the prices of the app increases, the average rating, total number of reviews and installs fall slightly.
- The Rating is slightly positively correlated with the Installs and Reviews column. This indicates that as the average user rating increases, the app installs, and number of reviews also increase.

ANALYSE DATA TO ANSWER

12. Sentiment Analysis of user review dataset.



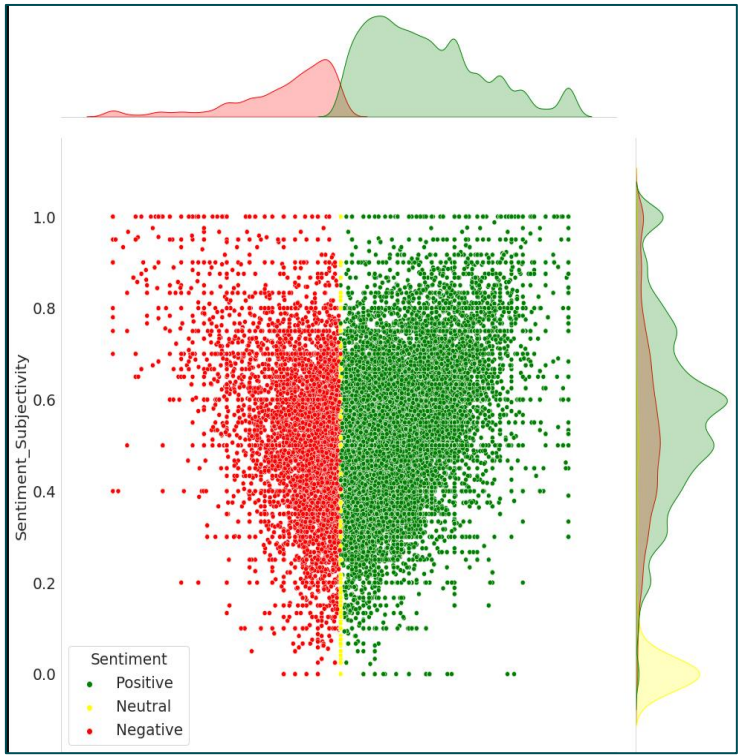


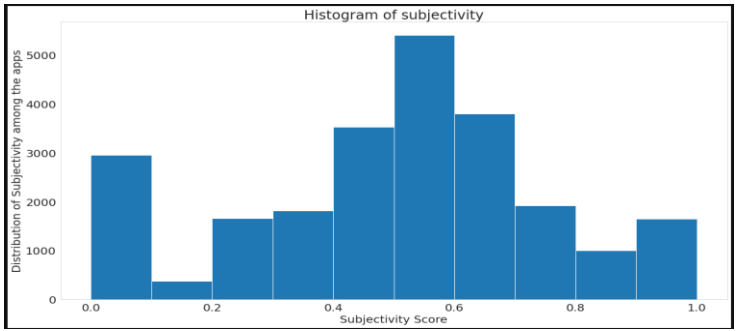


- Helix Jump has highest no of positive reviews
- Angry Birds has highest no of negative review
- Positive review has the highest share 64.12%

ANALYSE DATA TO ANSWER

13. Is sentiment subjectivity is proportional to sentiment polarity?





- The maximum number of reviews from the users are form their own experience (0.4-0.6)
- Sentiment subjectivity is not always proportional to sentiment polarity but in maximum number of cases, shows a proportional behavior, when variance is too high or low.

➢ SHARE YOUR FINDING

- The app's name should accurately describe its value propositions. because majority of successful application have this quality.
- Launching the apps in the category which having more and easy user reach such as "FAMILY", "GAME" etc.
- As per our analysis most of the apps are free it's around 92%, so if possible, try to launch the app with "Free" type. as it's increasing the user engagement.
- We also seen that the number of installs is correlated with the rating of the application as number of installs increases so as the application rating.
- One important inside we get with respect to the size of the app is that as the size of the application increases the installation of the app decrease. So, if we release a new apps in market make sure it's under 20MB.
- Content rating also affect the user engagement as more restricted your content rating is the more restricted your user engagement. So, try to keep user rating as "Everyone".
- Make sure the app will get the update at regular interval, as it's an important factor for user engagement and performance of the application. In our analysis we seen that most of the apps will get their app update at July month.
- As we seen form subjectivity most of the reviews as the objective, so for successful apps it's more important to keep eye on the user reviews and early resolution of the problems.
- We also seen our analysis that there is strong relationship between Install and reviews. for the new apps reviews are the important tool for increase the user engagement.
- For the apps it's more important have the android version compatibility above version 4. As we won't explore this attribute much but by using simple sorting, we can confirm this.

➢ Challenges & Future Work

- For this analysis we face most of the challenges in Ask and Process phase.
- In Ask phase of analysis process we wrap our head around the dataset to ask a question that really draws some insides. for that we have to do lot's of iterations.
- In Process phase of the analysis we handled null and duplicated values. the maximum number of NaN values are present in Rating attribute which approx. 14% (rounding-off) of the total records. Also there are some duplicated records as well which accounts to 5% of the total records.
- There were only 816 common apps in the merged data frame of both the play store and user reviews. Which is accounts only 10% of the total. We could have provided more valuable analysis if we had at least 70% - 80% of the data in the merged data frames.
- We might have filled the Reviews column's 14% NaN values by using the User Reviews' 42% NaN values to better grasp the attitudes within each category.
- There is so much more that can be discovered. For example, we have a Current_Version and an Android_Version that can be explored in depth, and we can provide more analysis to show how these factors affect and must be considered when developing an app for users.
- We can investigate the relationship between app size and Android_Version on the number of installs.
- By creating models that can aid in even better interpretation, machine learning can assist us in deploying more insights. Since this is something we can work on, we have left it for future work.


THANK
YOU !