Capstone Project

Play Store App Review Analysis

By- Raushan Kumar

Objective and Problem statement

- In this **data analysis project** we have asked to analyze technical aspect how about building a good android application and discover **key factors responsible for app engagement and success**. We also see trend and insights of play store and **tried to answer following questions**:

1. Which category of App is most popular on Play store?
2. Which category of App has more installs?
3. Which are the top 10 genres of apps?
4. What is the Distribution of the ratings of the apps?
5. What is the average count of the number of reviews for each category of apps?
6. What is the percentage of Free and Paid apps in the dataset?
7. Which are the top 10 expensive Apps in play store?
8. What is the count of different age category of target audience?
9. What are the Top 10 installed apps in any category?
10. Android version based on each category.
11. Does the last update date has an effect on rating?
12. What is the distribution of rating per number of installs and type (paid or free)?
13. Is there any co-relation between rating, reviews and price columns together?.
14. Is there any co-relation between price of apps and ratings? does price responsible for reviews and installs of apps?
15. Histogram of Subjectivity
16. Is sentiment subjectivity proportional to sentiment polarity?
17. Percentage of Review Sentiments.

DATASET CONTENT- Play Store

The Play Store dataset contain following features:

- **App**- Name of the apps.
- **Category**- Category under which the app falls.
- **Rating**- Applications rating in play store.
- **Reviews**- Number of reviews given to apps.
- **Size**- Size of the apps.
- **Installs**- Number of installs of the apps.
- **Type**- If the app is free or paid.
- **Price**- Price of the apps.
- **Content Rating**- Appropriate target audience of the apps.
- **Genres**- Genres under which the app fall.
- **Last updated**- Date when the app last updated.
- **Current Ver**- Current version of the apps.
- **Android Ver**- Minimum android version required for the apps to run.

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content_Rating	Genres	Last_Updated	Current_Ver	Android_Ver
count	10841	10841	9367.000000	10841	10841	10841	10840	10841	10840	10841	10841	10833	10838
unique	9660	34	NaN	6002	462	22	3	93	6	120	1378	2832	33
top	ROBLOX	FAMILY	NaN	0	Varies with device	1,000,000+	Free	0	Everyone	Tools	August 3, 2018	Varies with device	4.1 and up
freq	9	1972	NaN	596	1695	1579	10039	10040	8714	842	326	1459	2451
mean	NaN	NaN	4.193338	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	NaN	NaN	0.537431	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	NaN	NaN	1.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	NaN	NaN	4.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	NaN	NaN	4.300000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	NaN	NaN	4.500000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	NaN	NaN	19.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content_Rating	Genres	Last_Updated	Current_Ver	Android_Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up

Findings- Play Store

	Datatypes	Count of non-null values	NaN values	% NaN Values	Unique_count
App	object	10841	0	0.00	9660
Category	object	10841	0	0.00	34
Rating	float64	9367	1474	13.60	40
Reviews	object	10841	0	0.00	6002
Size	object	10841	0	0.00	462
Installs	object	10841	0	0.00	22
Type	object	10840	1	0.01	3
Price	object	10841	0	0.00	93
Content_Rating	object	10840	1	0.01	6
Genres	object	10841	0	0.00	120
Last_Updated	object	10841	0	0.00	1378
Current_Ver	object	10833	8	0.07	2832
Android_Ver	object	10838	3	0.03	33

- From the above, we understand that **except for Rating columns**, we are having a good dataset.
- The number of null values are:
- **Rating** has 1474 null values which contributes **13.60%** of the data.
- We know that there is only **one feature with numeric type** i.e. Rating.
- Unique counts of **Type** and **Content ratings** are very small i.e. **3** and **6** respectively

DATASET CONTENT- User Review

- **App** - The name of the application.
- **Translated Review** - what the users feedback is about the application.
- **Sentiment** - tells us about a view or opinion of the user w.r.t. the application.
- **Sentiment Polarity** - Sentiment polarity for an element defines the orientation of the expressed sentiment, i.e., it determines if the text expresses the positive, negative or neutral sentiment of the user about the application.
- **Sentiment Subjectivity** - It refers to the text that contains text which is usually expressed by a human having typical moods, emotions, and feelings. Mostly it is a public opinion and not a factual information.

	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
count	64295	37427	37432	37432.000000	37432.000000
unique	1074	27994	3	NaN	NaN
top	Angry Birds Classic	Good	Positive	NaN	NaN
freq	320	247	23998	NaN	NaN
mean	NaN	NaN	NaN	0.182146	0.492704
std	NaN	NaN	NaN	0.351301	0.259949
min	NaN	NaN	NaN	-1.000000	0.000000
25%	NaN	NaN	NaN	0.000000	0.357143
50%	NaN	NaN	NaN	0.150000	0.514286
75%	NaN	NaN	NaN	0.400000	0.650000
max	NaN	NaN	NaN	1.000000	1.000000

	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
0	10 Best Foods for You	I like eat delicious food. That's I'm cooking ...	Positive	1.00	0.533333
1	10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.288462
2	10 Best Foods for You	NaN	NaN	NaN	NaN

Findings- User Review

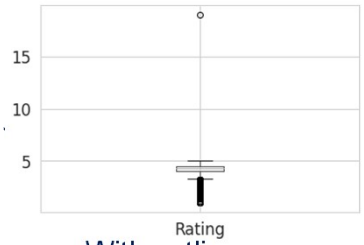
	Datatypes	Count of non-null values	NaN values	% NaN Values	Unique_count
App	object	64295	0	0.00	1074
Translated_Review	object	37427	26868	41.79	27994
Sentiment	object	37432	26863	41.78	3
Sentiment_Polarity	float64	37432	26863	41.78	6195
Sentiment_Subjectivity	float64	37432	26863	41.78	4530

- From the above, we understand that **only for App column**, we are having a good dataset.
- There so many of null values in dataset, we need to take care of that part.
- **Translated_Review** has 26868 null values which contributes **41.79%** of the data.
- **Sentiment** has 26863 null values which contributes **41.78%** of the data.
- **Sentiment_Polarity** has 26863 null values which contributes **41.78%** of the data.
- **Sentiment_Subjectivity** has 26863 null values which contributes **41.78%** of the data.
- We know that there is **two feature with numeric type** i.e. **Sentiment_Polarity** and **Sentiment_Subjectivity**.
- Unique count of **Sentiment** are very small i.e. **3** .
- Most of the **Translated_review** and **Sentiment** are **Good** and **Positive**.

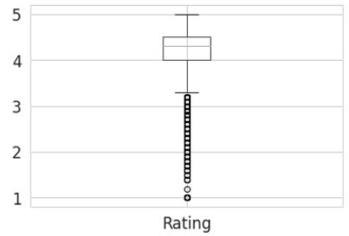
Data Cleaning - Univariate & Bivariate Analysis

Steps followed in data cleaning:

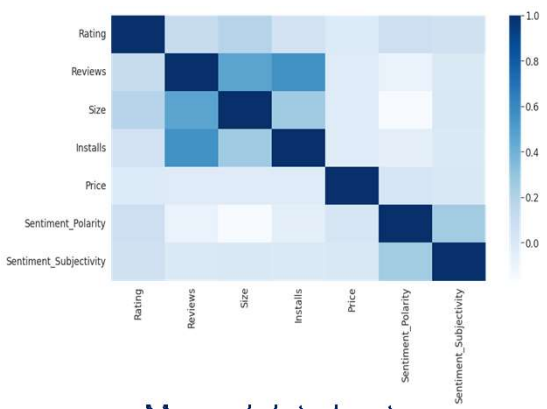
1. Handled missing values by dropping them in both dataset.
2. Merged two dataset in one.
3. Checked for noise or irrelevant data in columns/rows and removed them.
4. Removed outliners in data.
5. Removed duplicates from data.
6. Checked for correlated features.
7. Remove mostly or non-correlated data based on heat map.



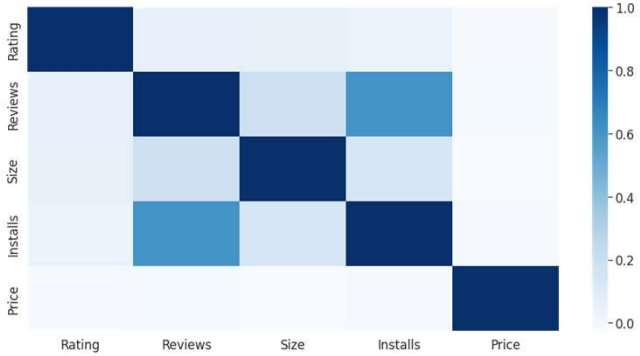
With outliners



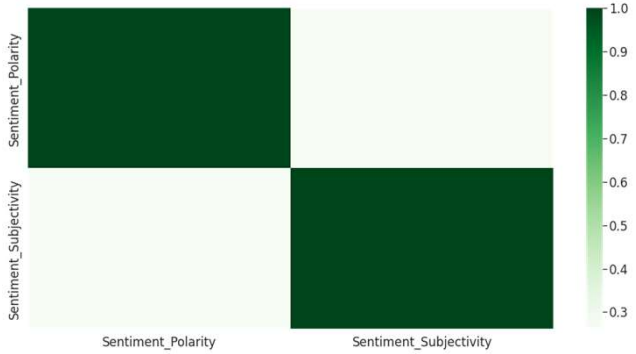
Without outliners



Merged data heatmap

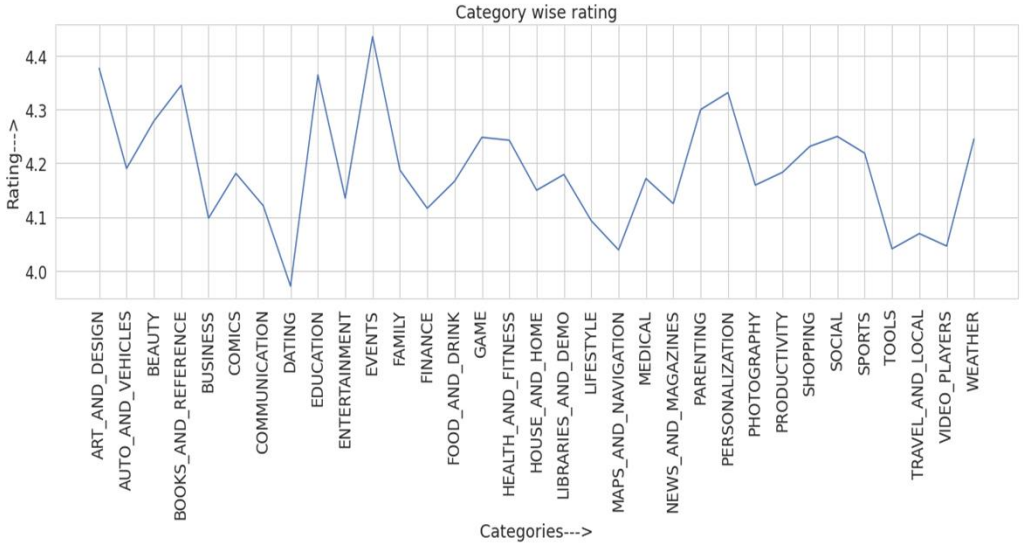


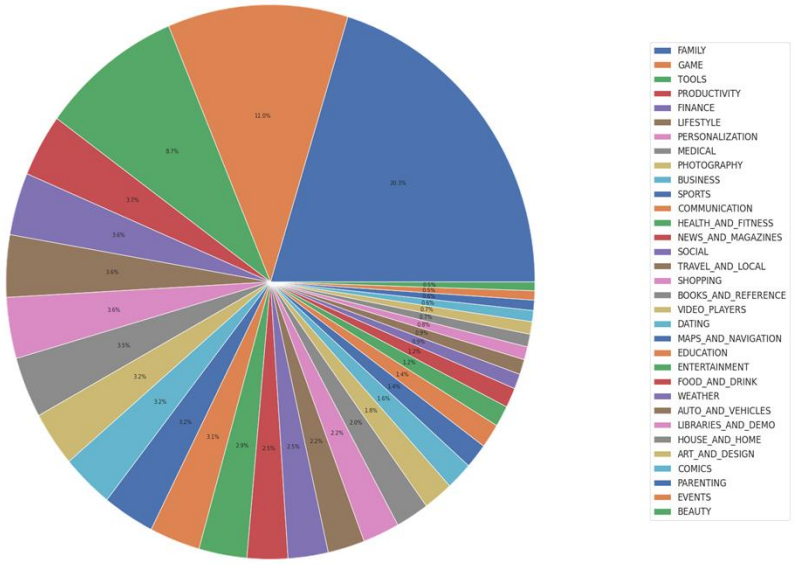
Play store heatmap



User review heatmap

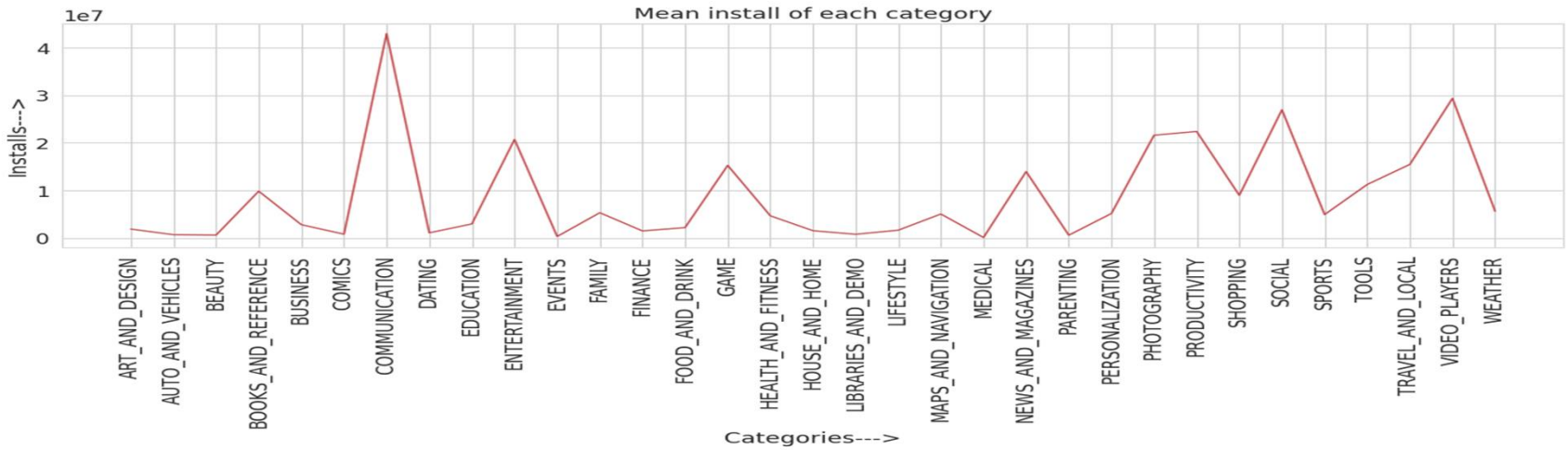
1. Which category of App is most popular on Play store?





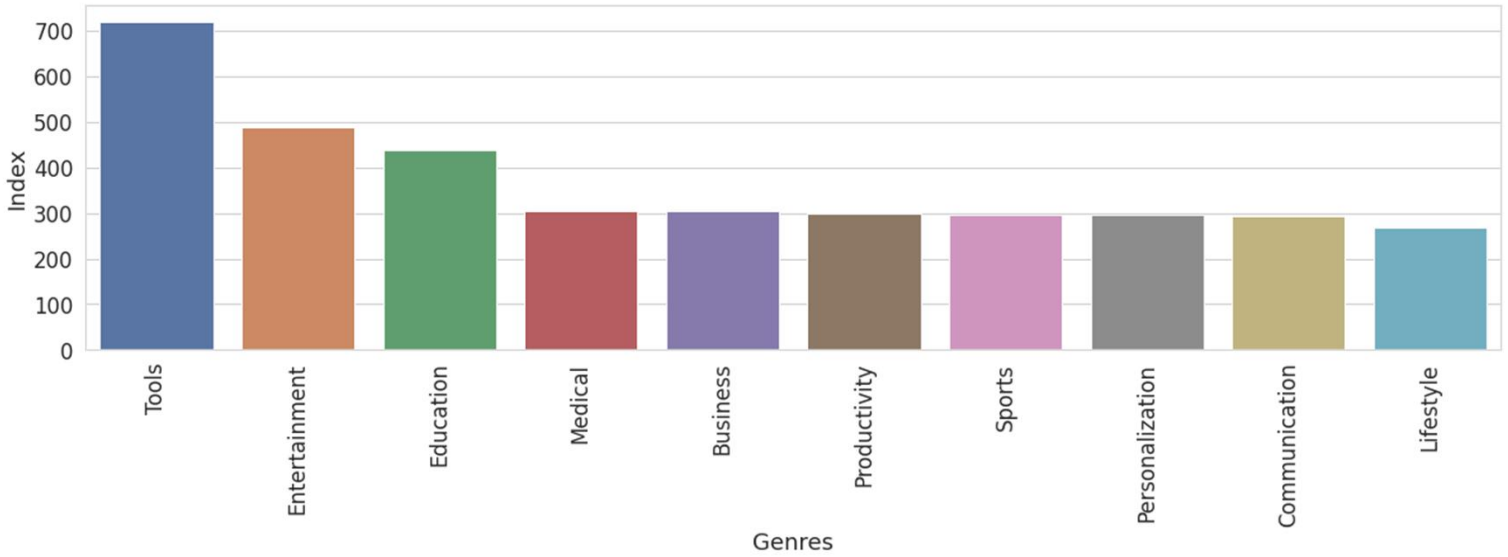
- Rating is highest for **family** and **game** category apps.
- Rating is low for **events** and **beauty** category apps

2. Which category of App has more installs?



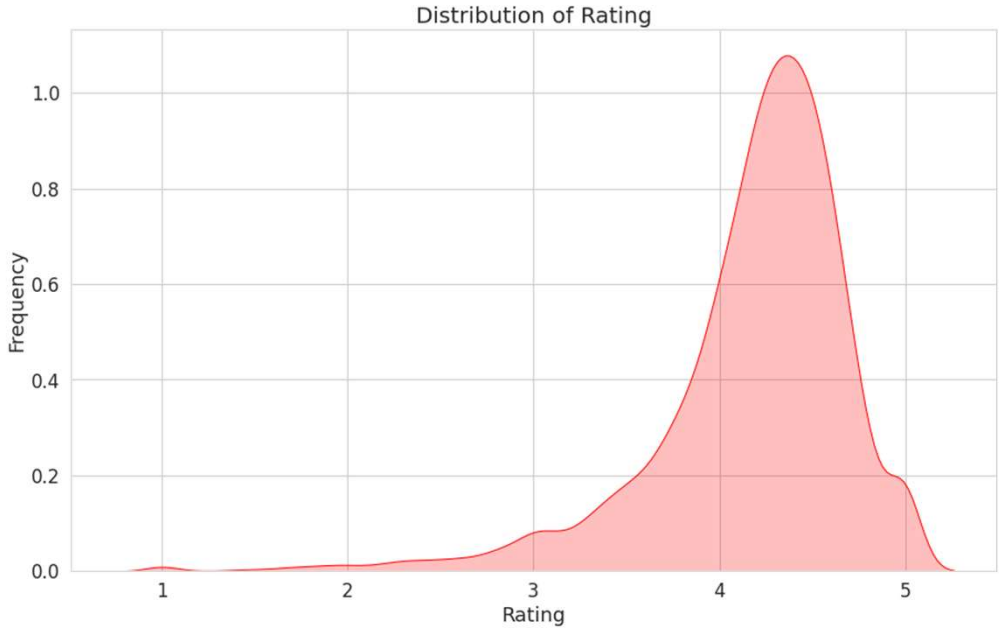
- Average installs is **higher for communication category** apps followed by **social category** apps.
- Average installation is **low** for categories such as **beauty, comics, dating, events, medical and parenting**.
- Even though the **average rating is quite high for event category** but the **mean install is quite low**.

3. Which are the top 10 genres of apps?



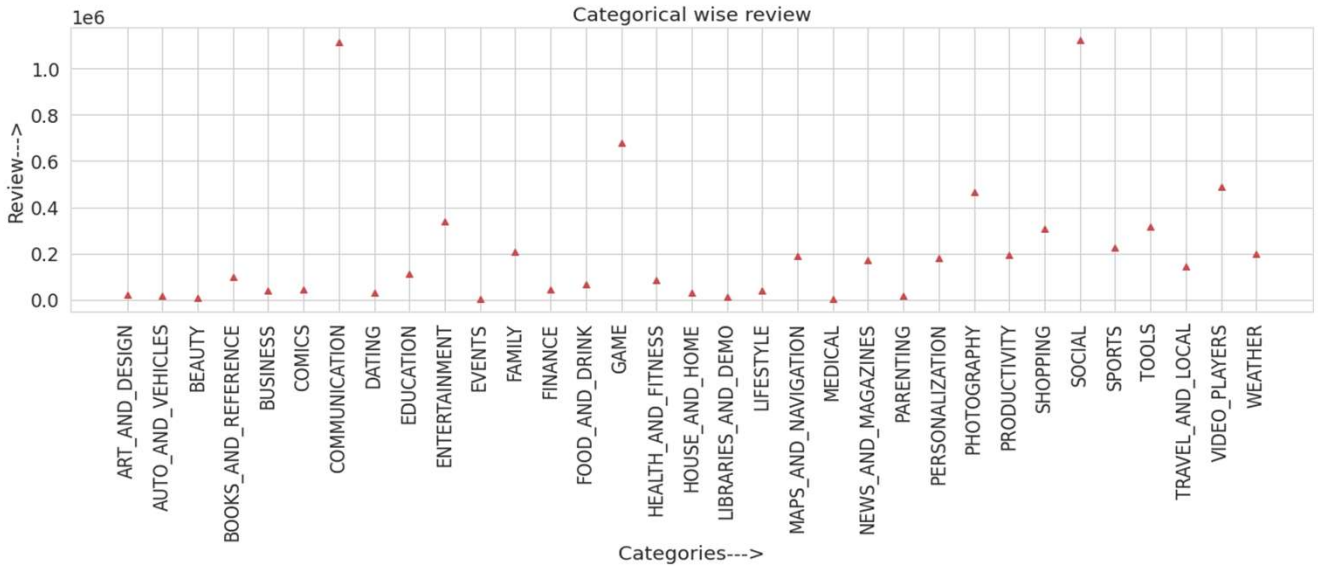
- **Tools genre** have the **highest** count followed by **Entertainment**.
- **Lifestyle genre** have the **lowest** count followed by **Communication**.

4. What is the Distribution of the ratings of the apps?



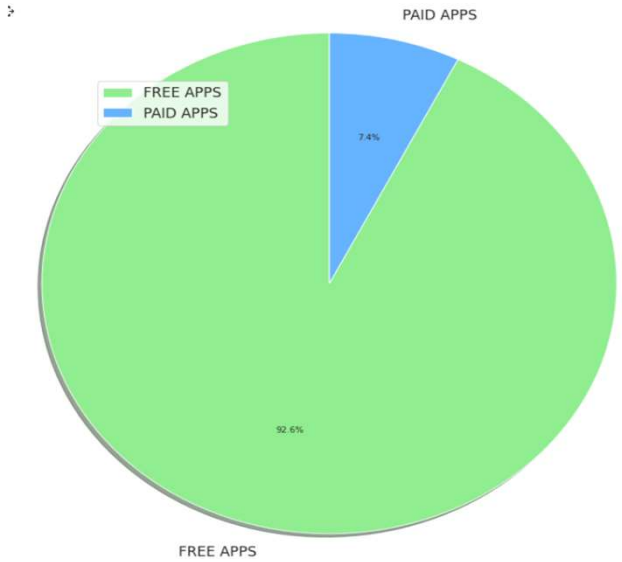
- From the above graph we can come to a conclusion that most of the apps in google playstore are rated in between **3.5 to 4.8**

5. What is the average count of the number of reviews for each category of apps?



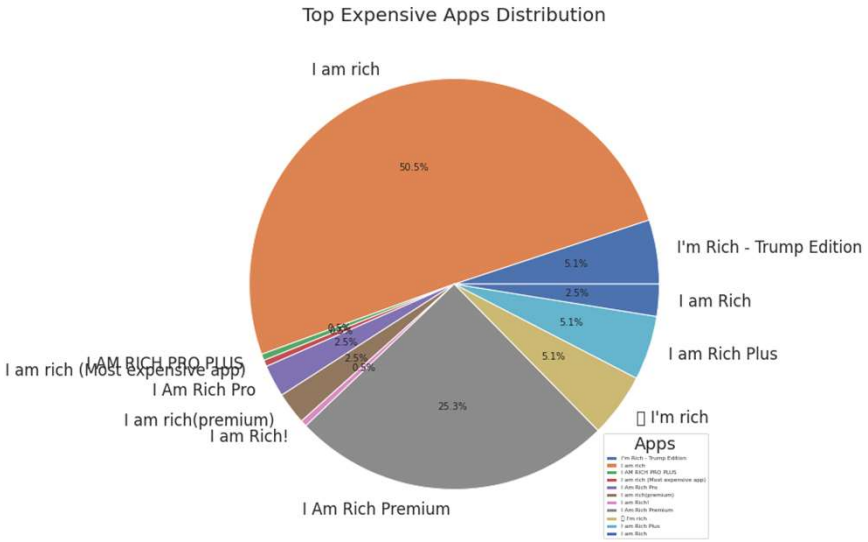
- Most user reviewed **communication** and **social category** apps.

6. What is the percentage of Free and Paid apps in the dataset



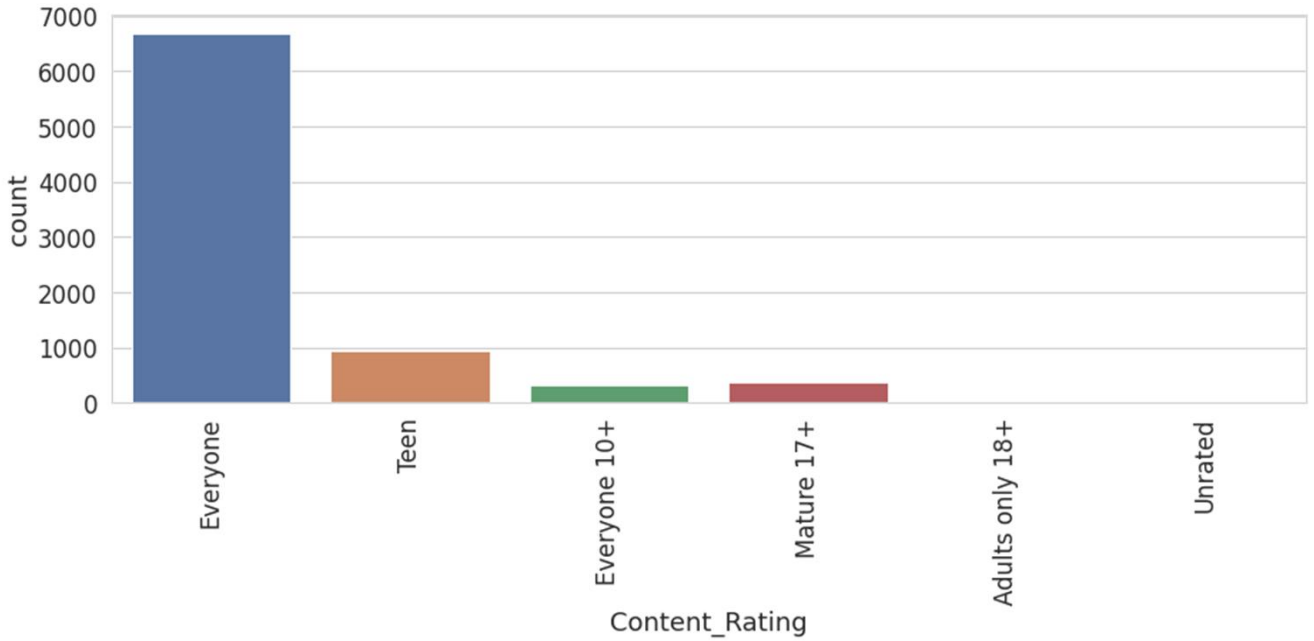
- Most of the apps in the dataset are **free type** approx **93%**.

7. Which are the top 10 expensive Apps in play store?



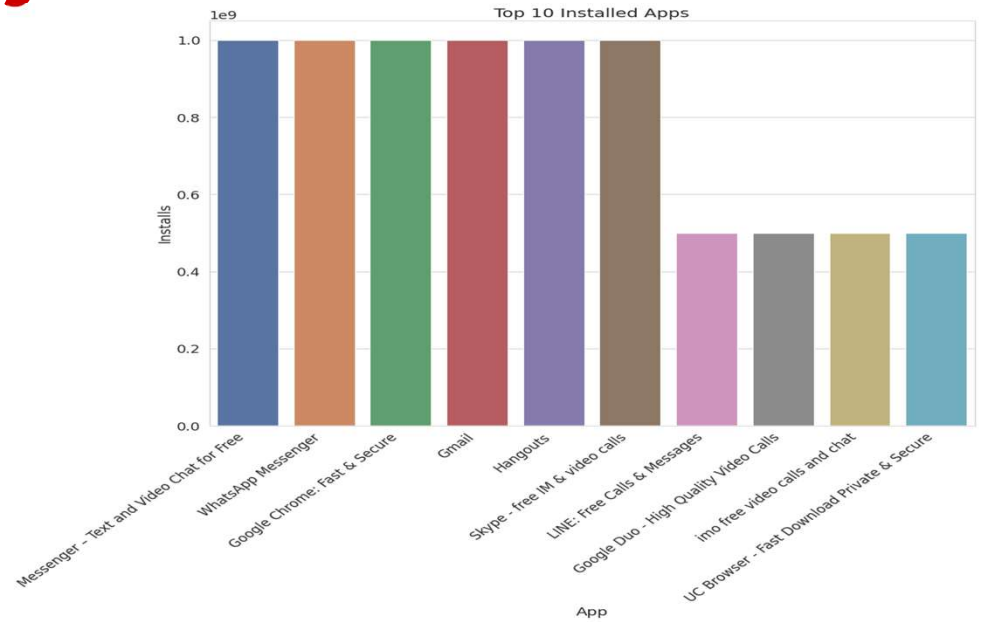
- From the above graph we can interpret that the **I am Rich** app is the **most expensive app** in the playstore. But this seems to be like a junk app. We need to further analyze if it is a junk app or not by deploying machine learning models in it.

8. What is the count of different age category of target audience?



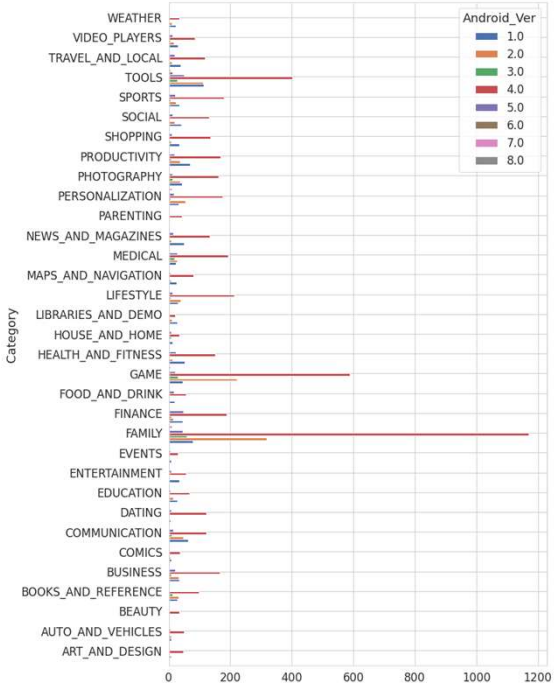
- **Everyone** age group has highest the target audience.

9. What are the Top 10 installed apps in any category?



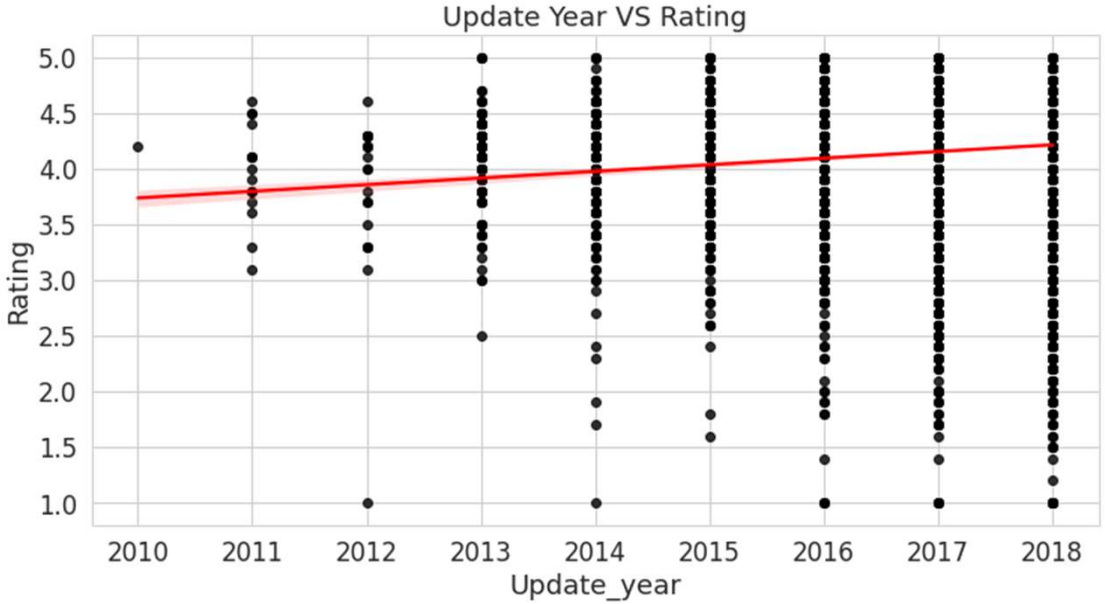
- From the above graph we can see that in the **Communication category Messenger-Text and Video Chat for Free, WhatsApp Messenger and Gmail has the highest installs.** In the same way we by passing different category names to the function, we can get the top 10 installed apps.

10. Android version based on each category



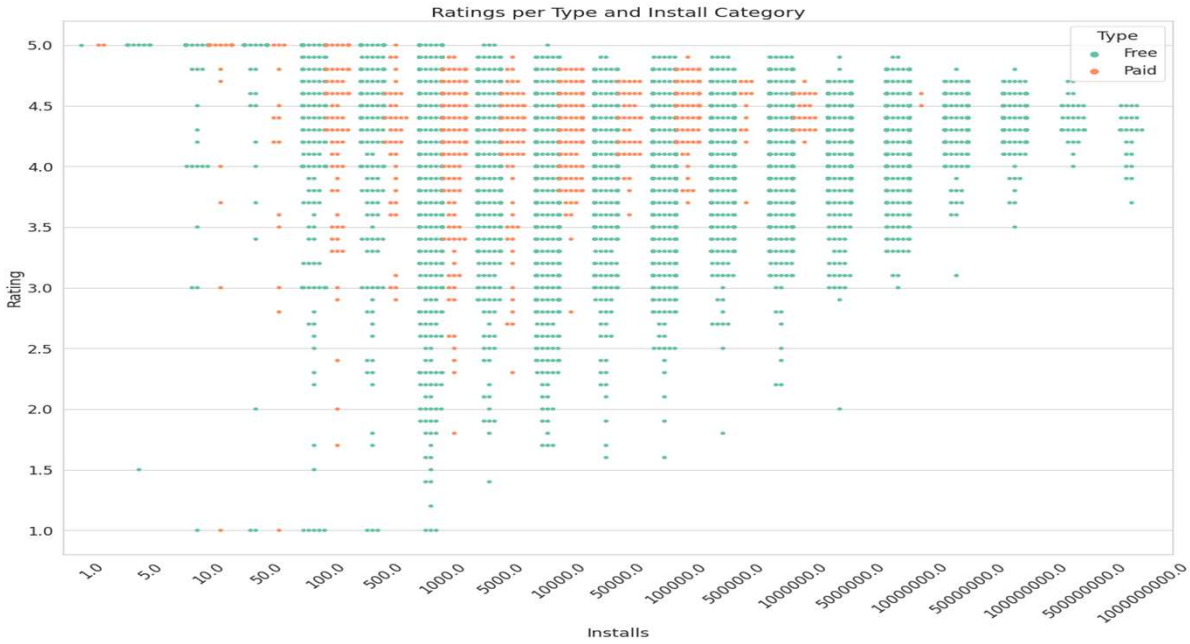
- It is clearly evident from the above plot that majority of the apps are working on **Android_Ver 4.0 and up**.

11. Does the last update date has an effect on rating?



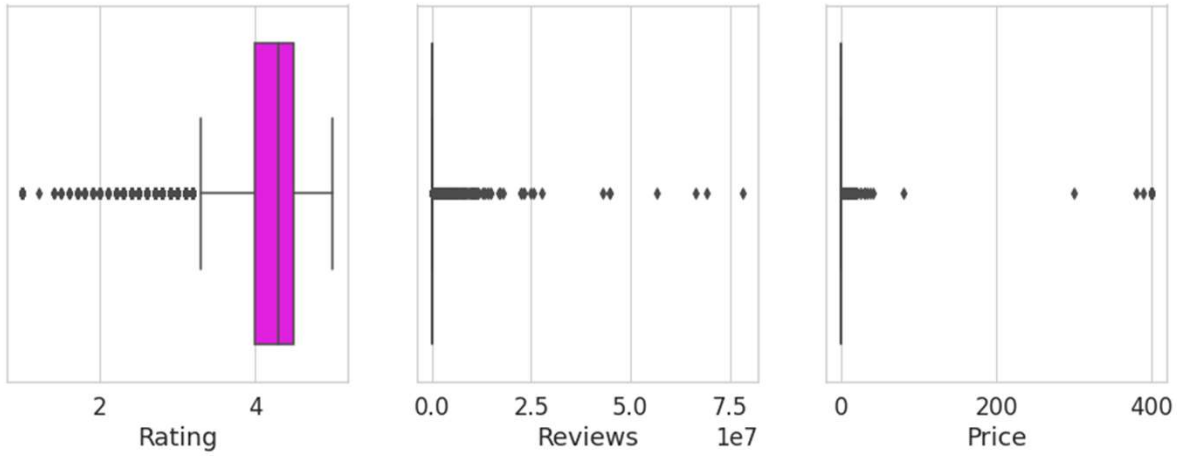
- From above graph, we can conclude, the app gets **more recent updates chances of getting a higher rating increases**.

12. What is the distribution of rating per number of installs and type (paid or free) ?



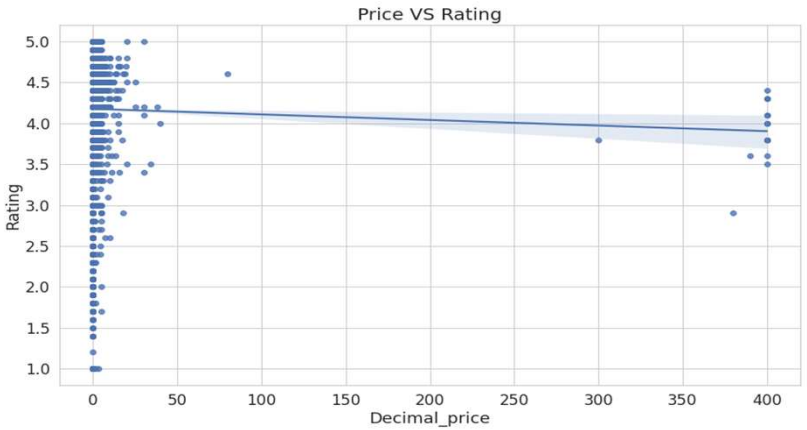
- Looks like rating is **distributed around 4.5** when its categorized per install category.
- Google play store have very **few paid apps**.

13. Is there any co-relation between rating, reviews and price columns together?.



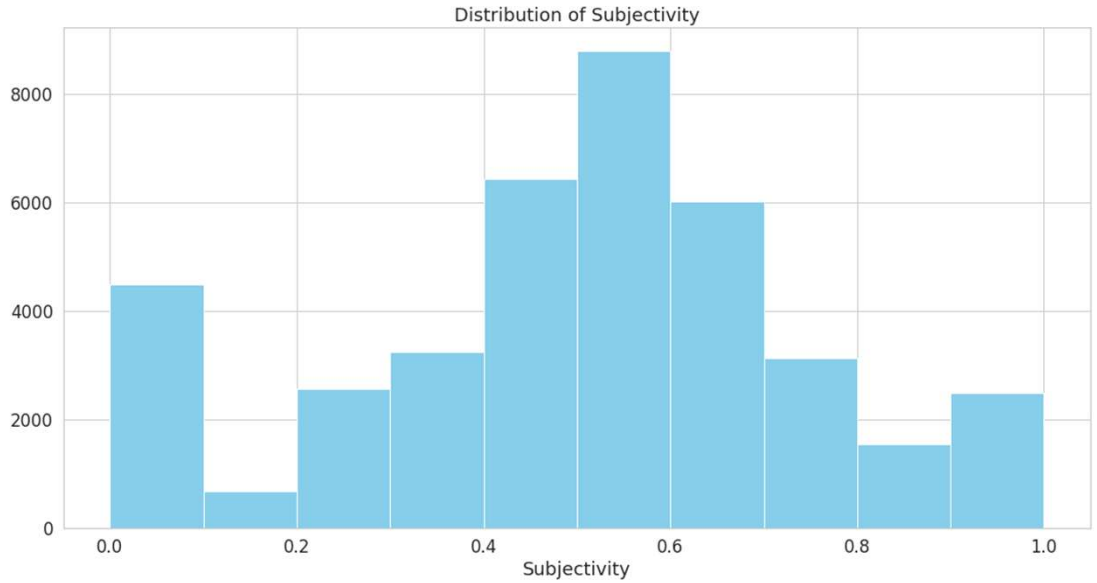
- We can see that most of the Ratings are between 4 and around 4.5 and 5.
- As far as Reviews are concerned, for **most of the Apps Reviews are not given**.
- Also for Price, most of the Apps are Free.

14. Is there any co-relation between price of apps and ratings? does price responsible for reviews and installs of apps?



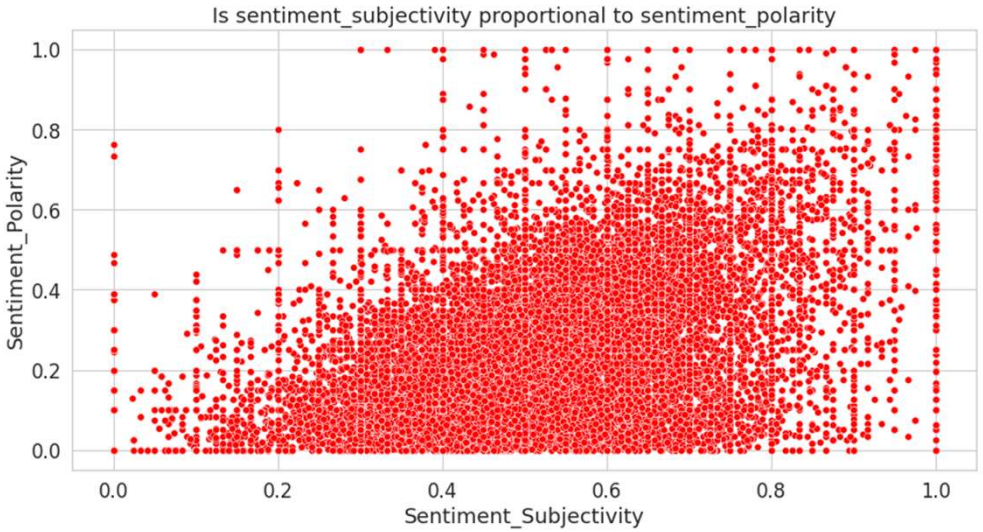
- From above we can clearly see as the price of app increases ratings received seems to decrease.
- From Price distribution graph we conclude that most of apps used is free.

15. Histogram of Subjectivity



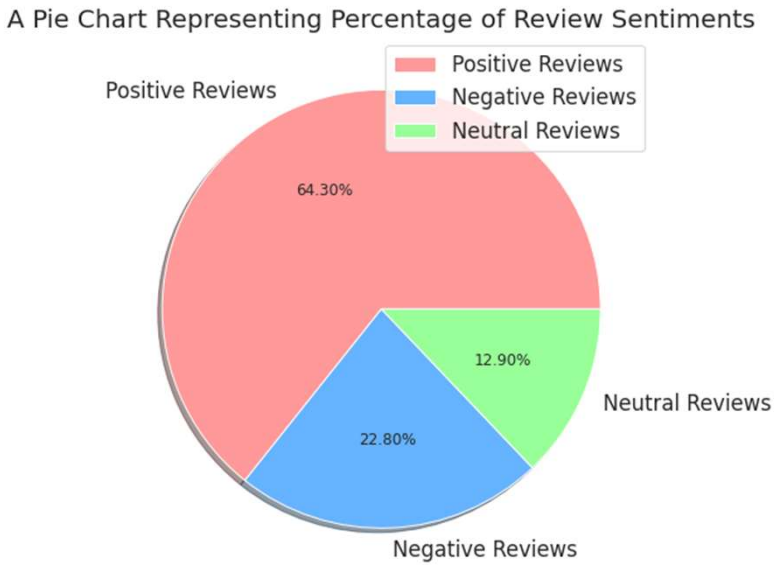
- **0 - objective(fact)**
- **1 - subjective(opinion)**
- It can be seen that maximum number of **sentiment subjectivity lies between 0.4 to 0.7**.
- From this we can conclude that **maximum number of users give reviews to the applications, according to their experience.**

16. Is sentiment subjectivity proportional to sentiment polarity?



- From the above scatter plot it can be concluded that **sentiment subjectivity is not always proportional to sentiment polarity but in maximum number of cases, it shows a proportional behavior, when variance is too high or low.**

17. Percentage of Review Sentiments



- Positive reviews are **64.30%**
- Negative reviews are **22.80%**
- Neutral reviews are **12.90%**

Conclusion

1. Rating is very important factor for installation of apps as **user mostly like to watch rating before using app**. so developer should also work on updating their content as per the ratings.

2. From all above we analyze rating and installation are related, so **owners should encourage to write review of their app**.

3. Number of **installation of free app is more compare to Paid app**, so developer can also consider this point for high reach.

4. They need to focus on **updating their apps regularly** as it attract more people.

5. More installation of app **in game app ,followed by communication ,productivity** so if one can thinking to develop an app will go for these.

Contd...

6. We can also see social media app is almost free, so **keeping social media app free** may be a great advantages over others.

7. App category like **events and beauty have not much reach**, so one can also keep this consideration.

8. Most of the apps are **downloaded by teens**, so users of other age category, must also be encouraged to install the apps.

9. From our above analysis we conclude that **sentiments of the user keep varying**, so owner need to **keep updating their app regularly basis on user feedback**.

Thanks