CAPSTONE PROJECT

Play Store App Review

(Exploratory Data Analysis)

Individual Project By

Sourabh Deshmukh

ROADMAP FOR EDA

01	Understanding the problem?	Analysing the problem is very important because output from the data is mainly depends upon which questions are asked to the data.
02	Cleaning of Data	Analyzing the given dataset, checking its sanity and then proceeding with further analysis.Checking for Null Values,Duplicate values.
03	Data Visualization	A picture is worth 1000 words, taking it seriously and tried to explain relation among variables by different types of visualization techniques.
04	Conclusion	In last step summing up results and came up with the inferences and recommendations.

PROBLEM STATEMENT ?

Explore and analyze the data to discover key factors responsible for app engagement and success.

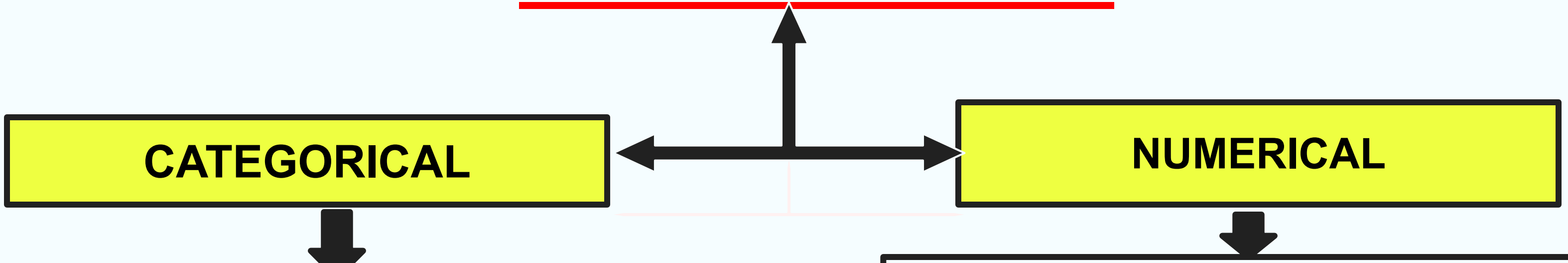
WHAT IS APP’S SUCCESS AND ENGAGEMENT ?

- App Engagement is the ways in which users interact with an app. App engagement is defined by a set of metrics that measure user interaction with an app.
- Following are the few user engagement metrics:

1. Number of Downloads.	4. Session interval
2. Total Number of User's (New & Old).	5. Session length.
3. Active app users.	6. Retention

- A successful app combines three aspects in a smart way:
- Market, User and the Product itself.
- All these factors need to work together to give users a unique value, great usability, and good performance. Accessibility is the last but not least key feature of a successful mobile app.

DATA SUMMARY



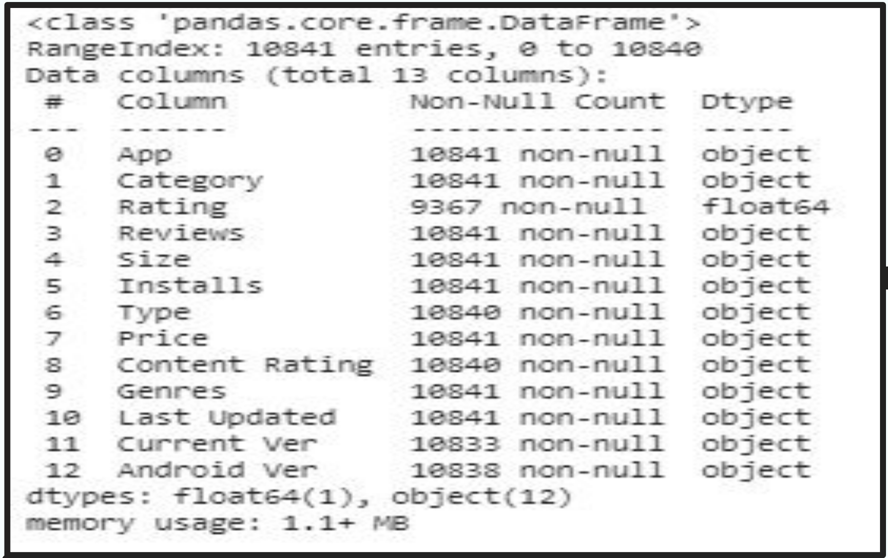
CATEGORICAL

NUMERICAL

1. **App (Object)** : Application name
2. **Category (Object)** : Category the app belongs to
3. **Type (Object)** : Paid or Free
4. **Content Rating(Object)** : Age group the group is targeted at -Everyone /Teen/Everyone10+/Matured 17/Adults only 18+/Unrated.
5. **Genres (Object)** : Genres the app belongs to
6. **Last Updated(Object)** : Date when the app was last updated on Play Store.

1. **Rating (Float) :** Overall user rating of the app
2. **Reviews (Object) :** Number of user reviews for the app
3. **Size(Object) :** Size of the app
4. **Installs(Object)** : Number of user downloads/installs for the app
5. **Price(Object)** : Price of the app in dollar
6. **Current Ver(Object):**Current Version of the app available on Play Store
7. **Android Ver(Object) :** Min required Android version

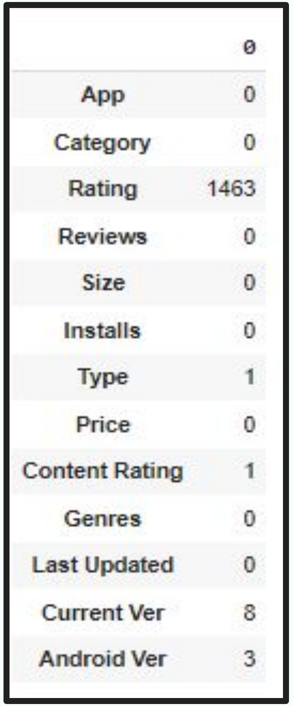
DATA CLEANING



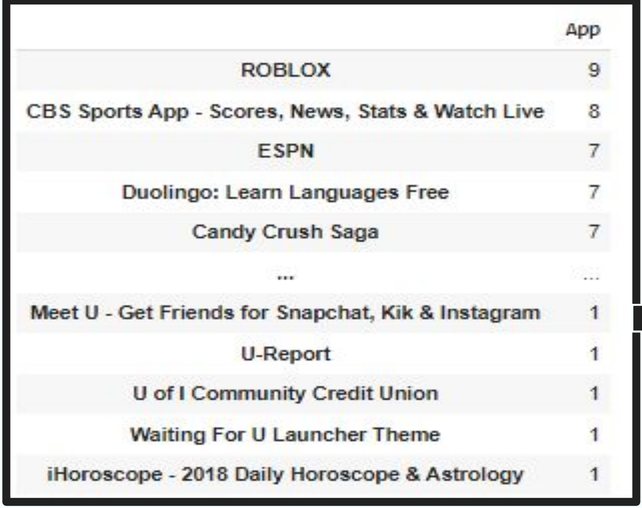
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             10841 non-null  object 
 1   Category        10841 non-null  object 
 2   Rating          9367 non-null   float64
 3   Reviews         10841 non-null  object 
 4   Size            10841 non-null  object 
 5   Installs        10841 non-null  object 
 6   Type            10840 non-null  object 
 7   Price           10841 non-null  object 
 8   Content Rating  10840 non-null  object 
 9   Genres          10841 non-null  object 
 10  Last Updated    10841 non-null  object 
 11  Current Ver     10833 non-null  object 
 12  Android Ver     10838 non-null  object 
dtypes: float64(1), object(12)
memory usage: 1.1+ MB
```

DATA SHAPE (10841,13)

TOTAL NULL VALUES

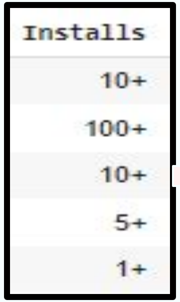


	0
App	0
Category	0
Rating	1463
Reviews	0
Size	0
Installs	0
Type	1
Price	0
Content Rating	1
Genres	0
Last Updated	0
Current Ver	8
Android Ver	3

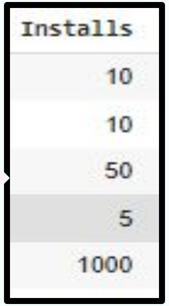


	App
ROBLOX	9
CBS Sports App - Scores, News, Stats & Watch Live	8
ESPN	7
Duolingo: Learn Languages Free	7
Candy Crush Saga	7
...	...
Meet U - Get Friends for Snapchat, Kik & Instagram	1
U-Report	1
U of I Community Credit Union	1
Waiting For U Launcher Theme	1
iHoroscope - 2018 Daily Horoscope & Astrology	1

TOTAL DUPLICATE VALUES : 483



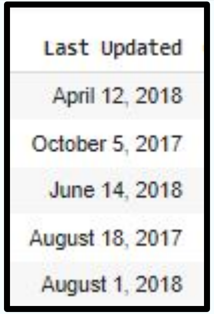
Installs
10+
100+
10+
5+
1+



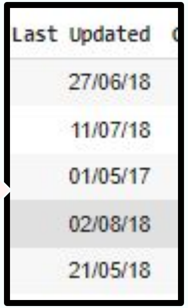
Installs
10
10
50
5
1000

REMOVING “ + “ FROM THE INSTALLS

CHANGING TO DATE TIME-TIME FORMAT

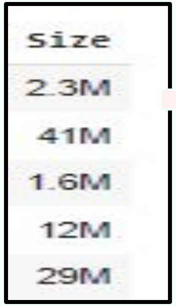


Last Updated
April 12, 2018
October 5, 2017
June 14, 2018
August 18, 2017
August 1, 2018

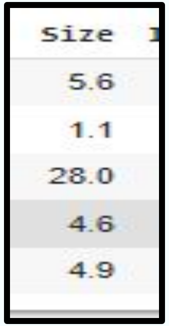


Last Updated
27/06/18
11/07/18
01/05/17
02/08/18
21/05/18

REMOVING “ M” FROM THE SIZE

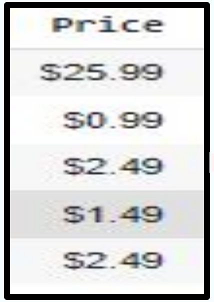


Size
2.3M
41M
1.6M
12M
29M

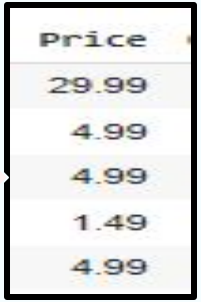


Size
5.6
1.1
28.0
4.6
4.9

REMOVING “ $ “ FROM THE PRICE



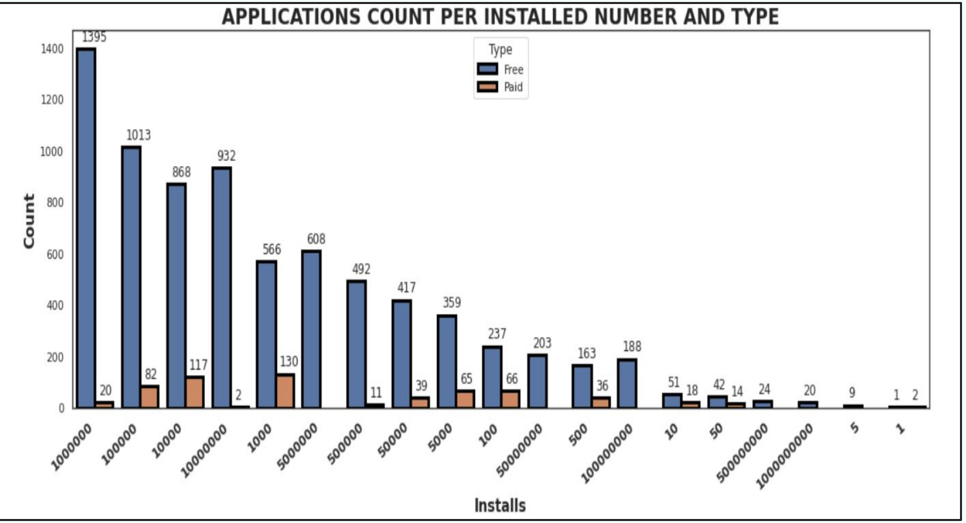
Price
$25.99
$0.99
$2.49
$1.49
$2.49

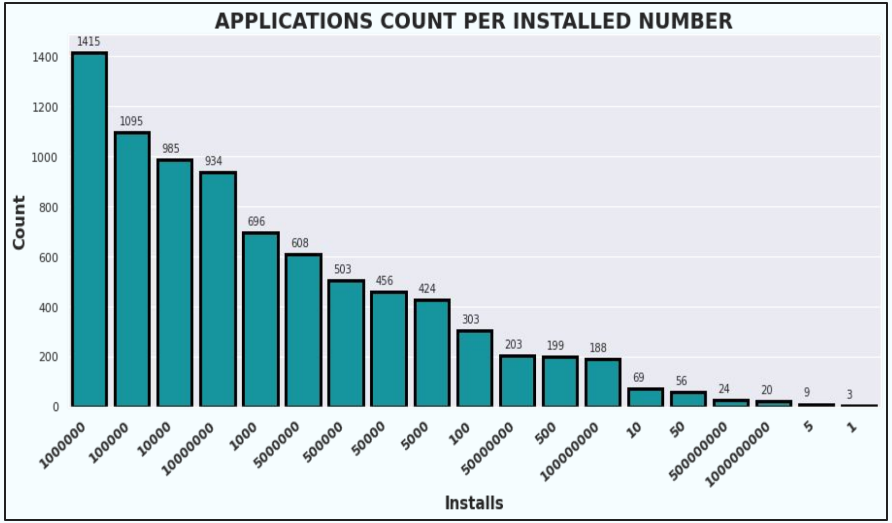


Price
29.99
4.99
4.99
1.49
4.99

DATA SET SHAPE AFTER CLEANING : (8190,13)

INSTALLS AND TYPE

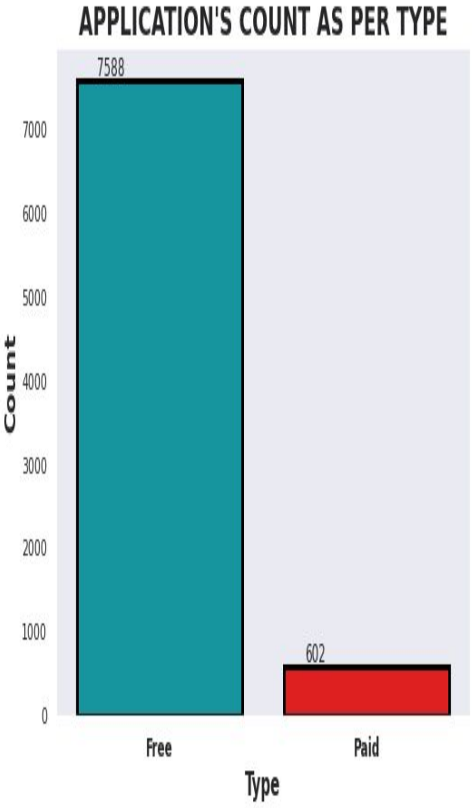

APPLICATIONS COUNT PER INSTALLED NUMBER AND TYPE

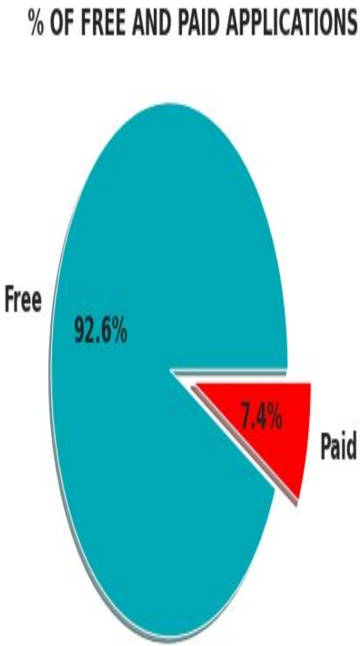

APPLICATIONS COUNT PER INSTALLED NUMBER

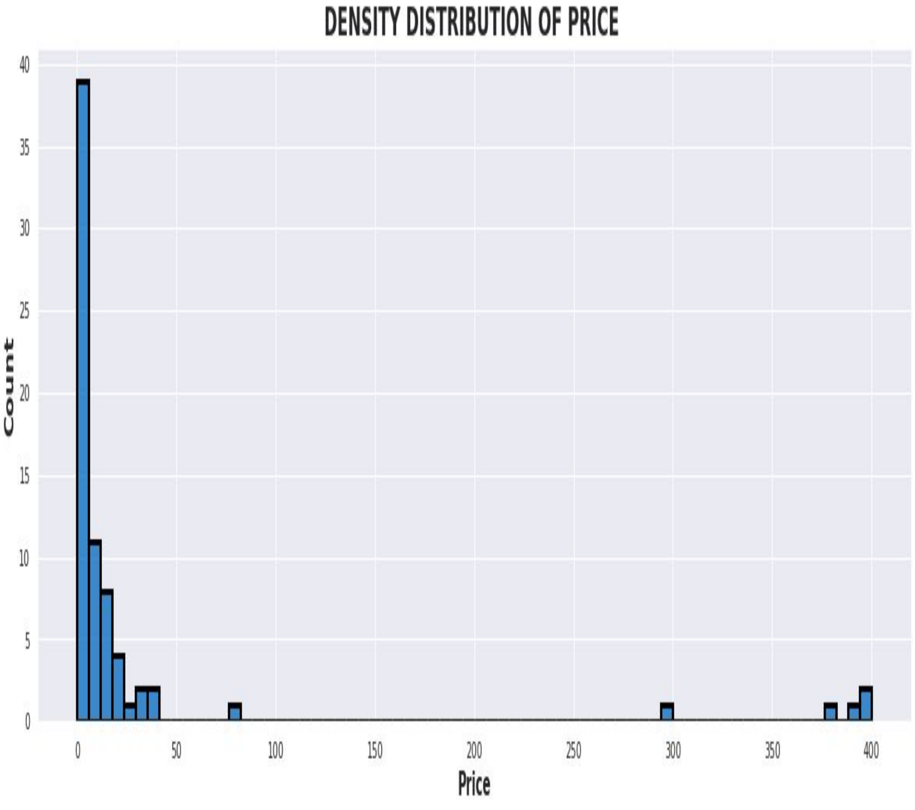
- **The above figure gives us the impression that the applications have been downloaded at least once and at maximum more than one billion times.**
- **The most number of applications are with over 1 million and less than 5 million installs.**
- **There are only 20 applications in Big Billion Club**

PRICE

- It's clearly visible that there are maximum number of Free applications in the data set ,which is about 93.7 % of total applications.







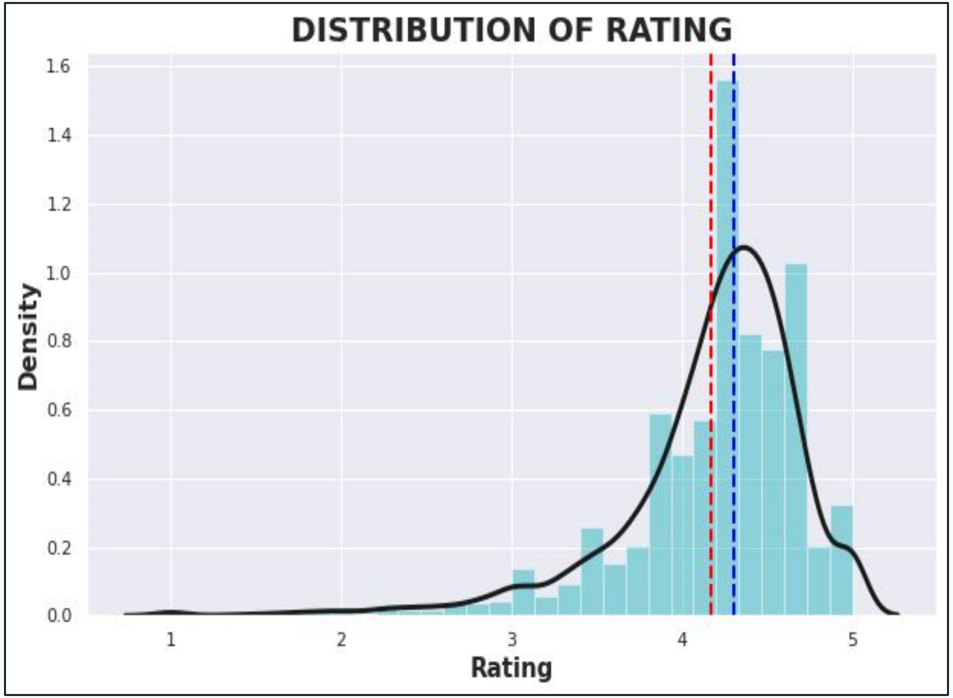
RATING AND SIZE

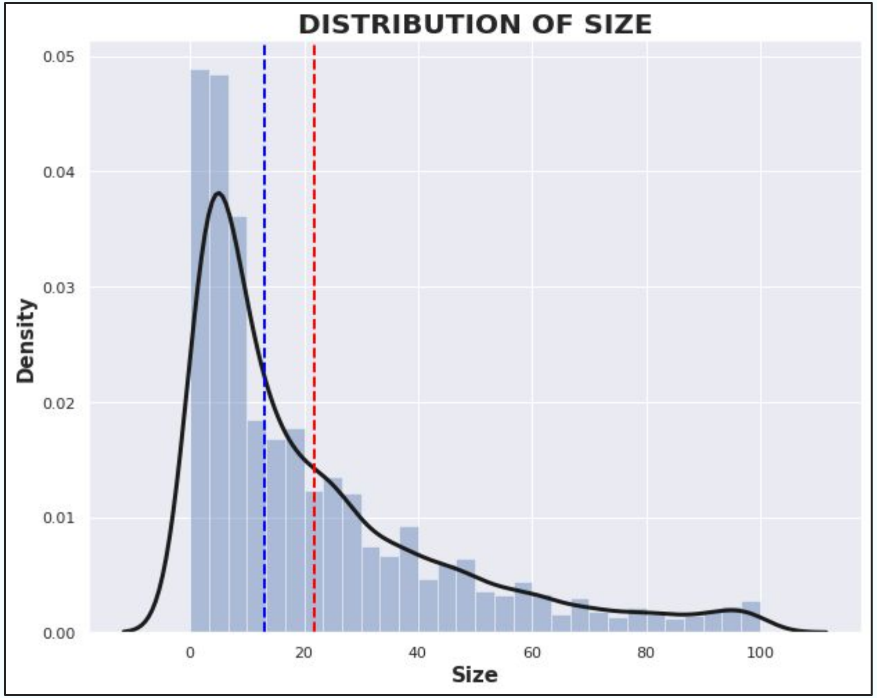
The average rating for all the application is 4.17 and the 50% of applications get rating between 4 to 4.5.

Due to the fact that rating is not mandatory, most users choose to skip it.

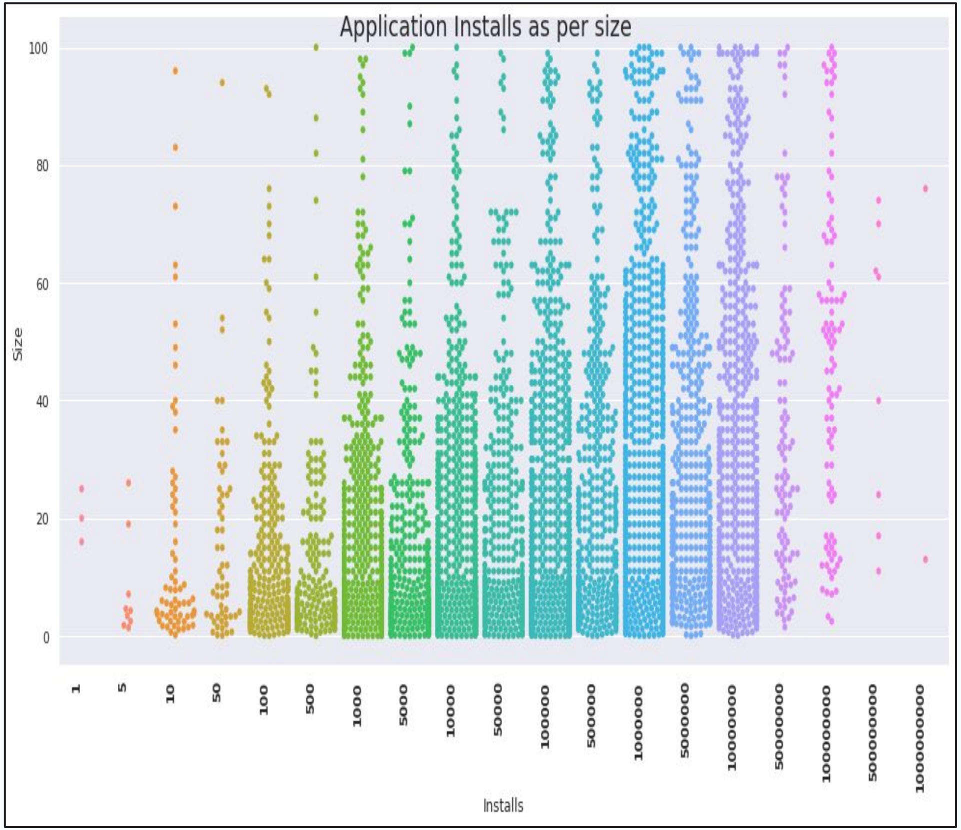
The application size lies between the "0 MB" and the "100 MB".

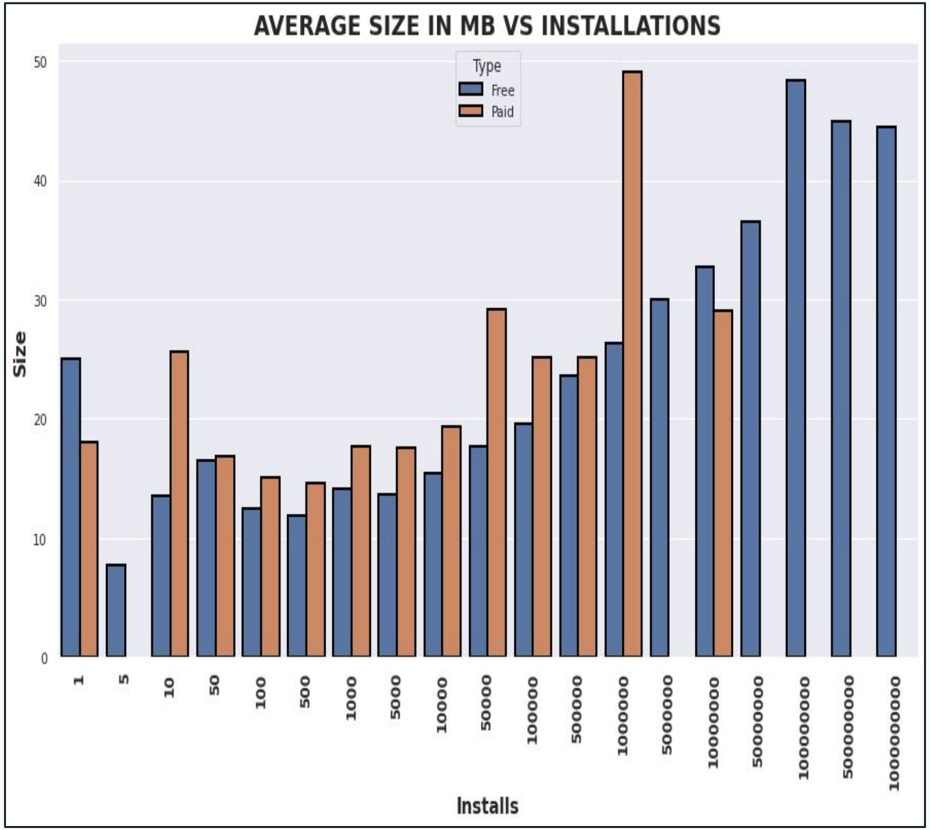
75% of application sizes lies between the range 0 MB to 33 MB.





INSTALL VS SIZE VS TYPE



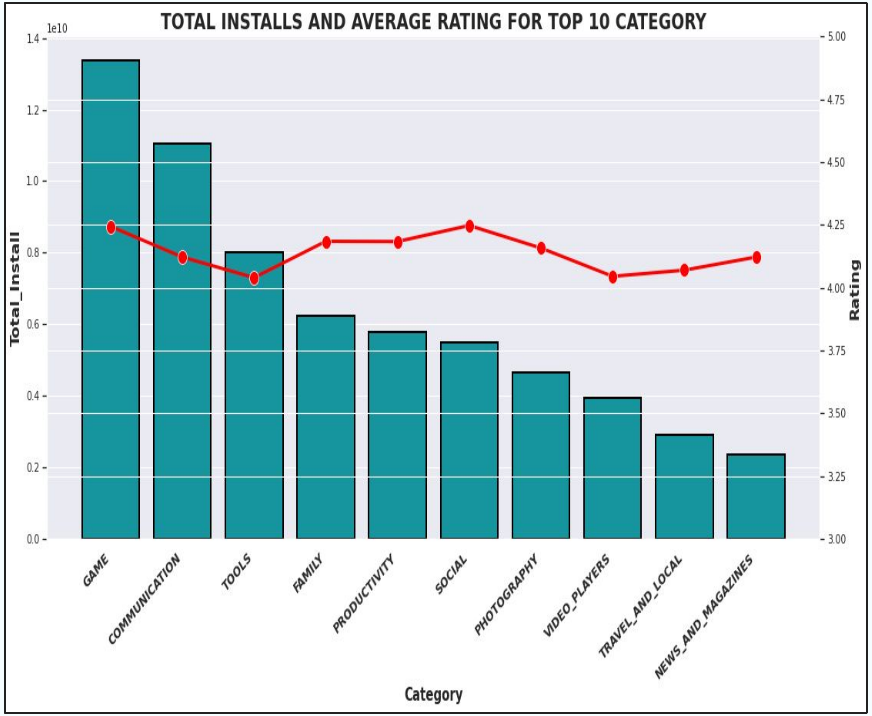


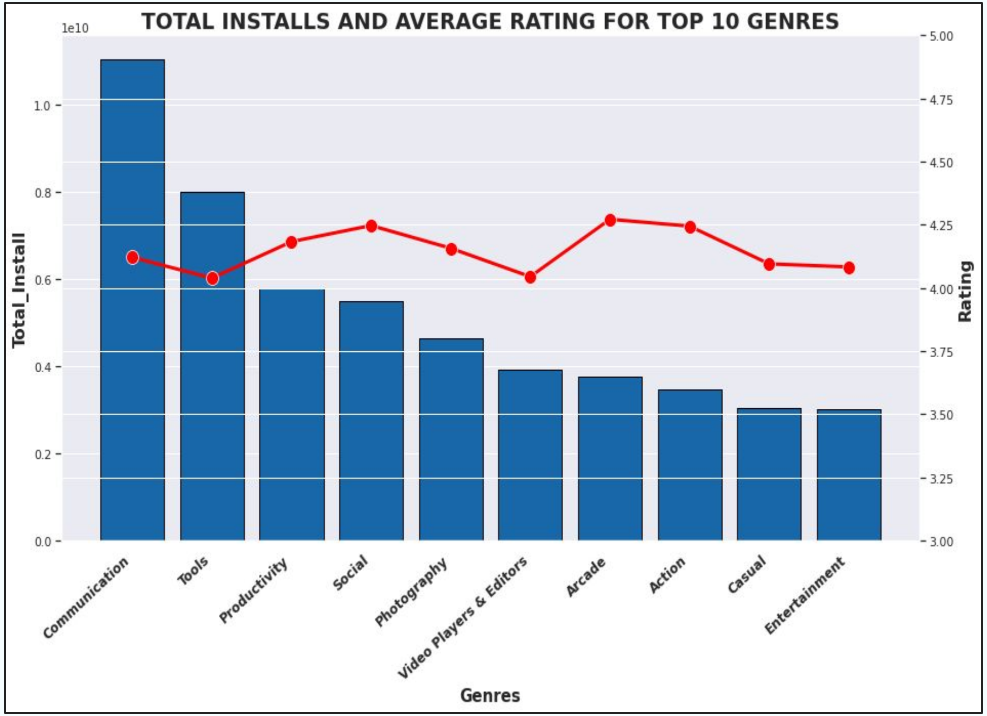
CATEGORY AND GENRES

Gaming category is clearly at the top of the chart with most installations and an average rating of 4.25.

Communication just falls short of Game in terms of installations and is ranked second in the most installed category.

Genres are the subcategories of the feature category so there few categories are divided into the genres and some are not.



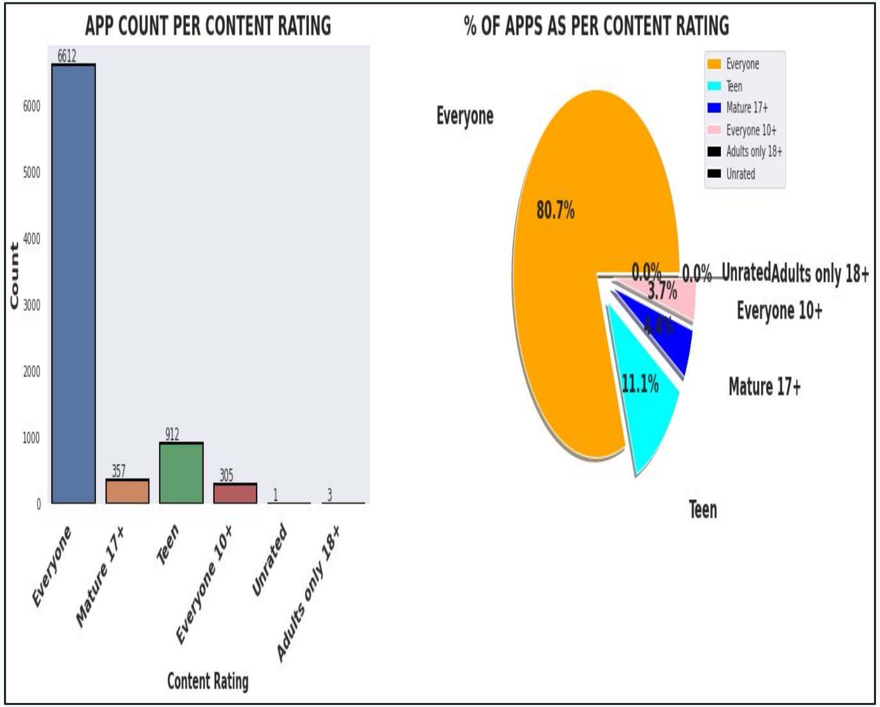


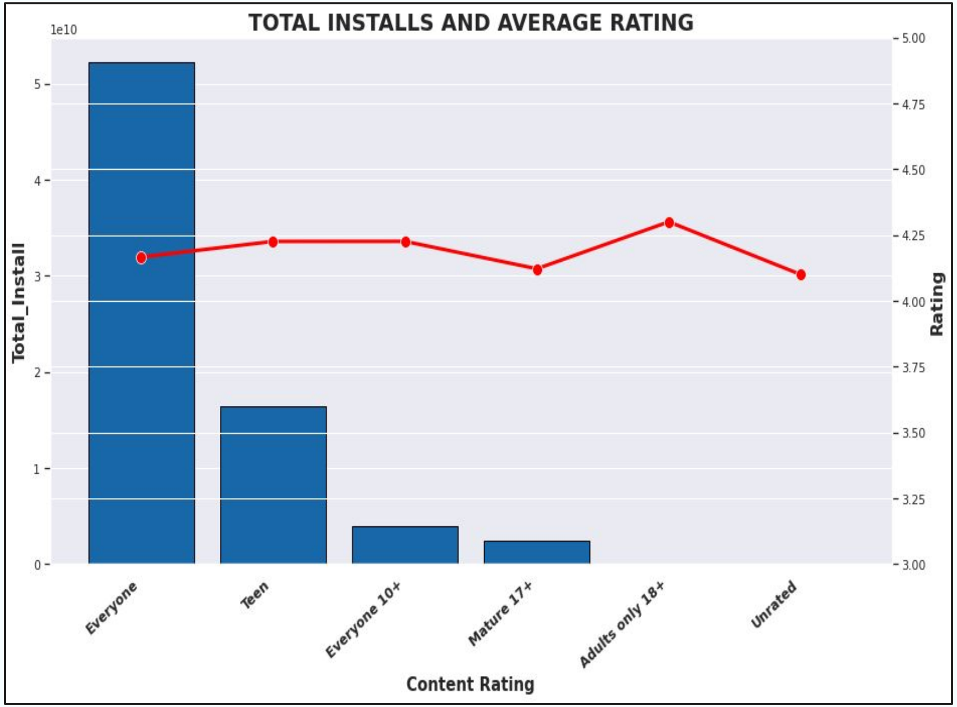
CONTENT RATING

It's clearly visible that about 80% apps don't have any restriction,which falls under "Everyone" category.

Only one app falls under "Unrated" and three falls under "Adults only" category.

We observed significant relationships between the Install and Contain Rating in the above plots.Applications with content rating type "Everyone", "Everyone 10+" and "Teen" have the most installations, with an average rating exceeding 4.2.

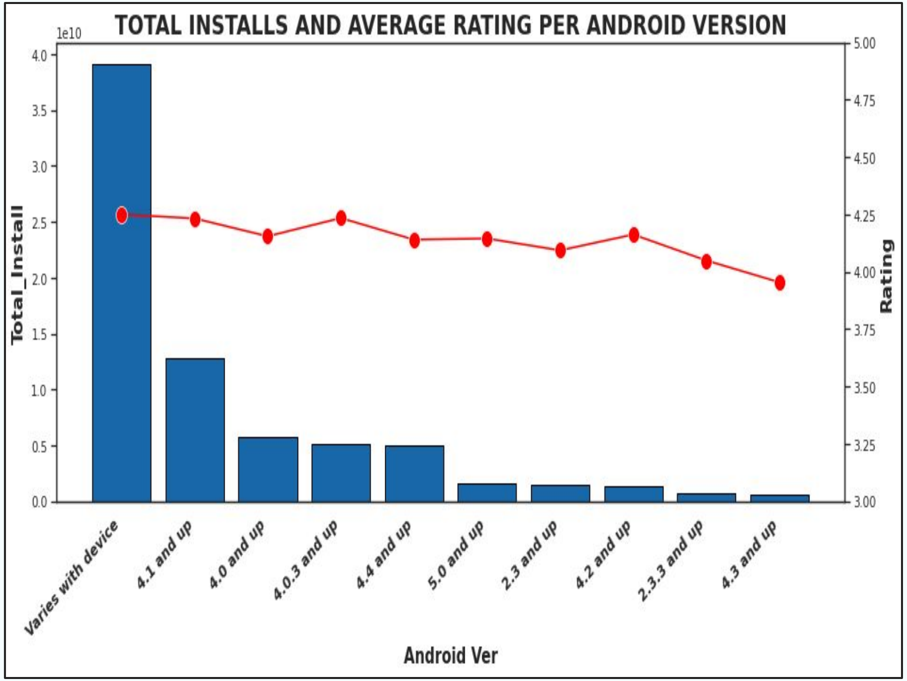


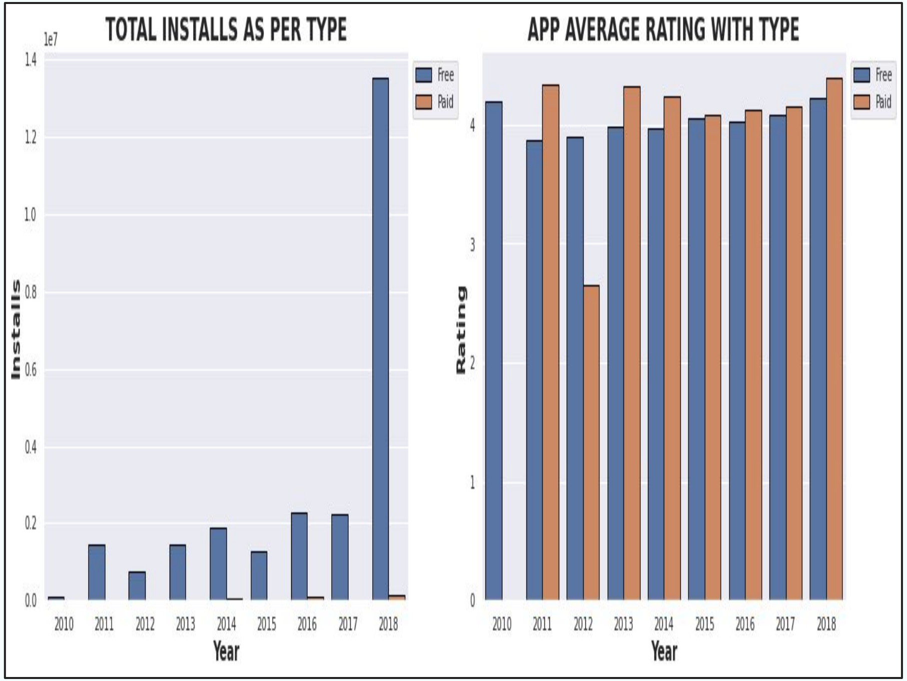


ANDROID VERSION AND LAST UPDATED

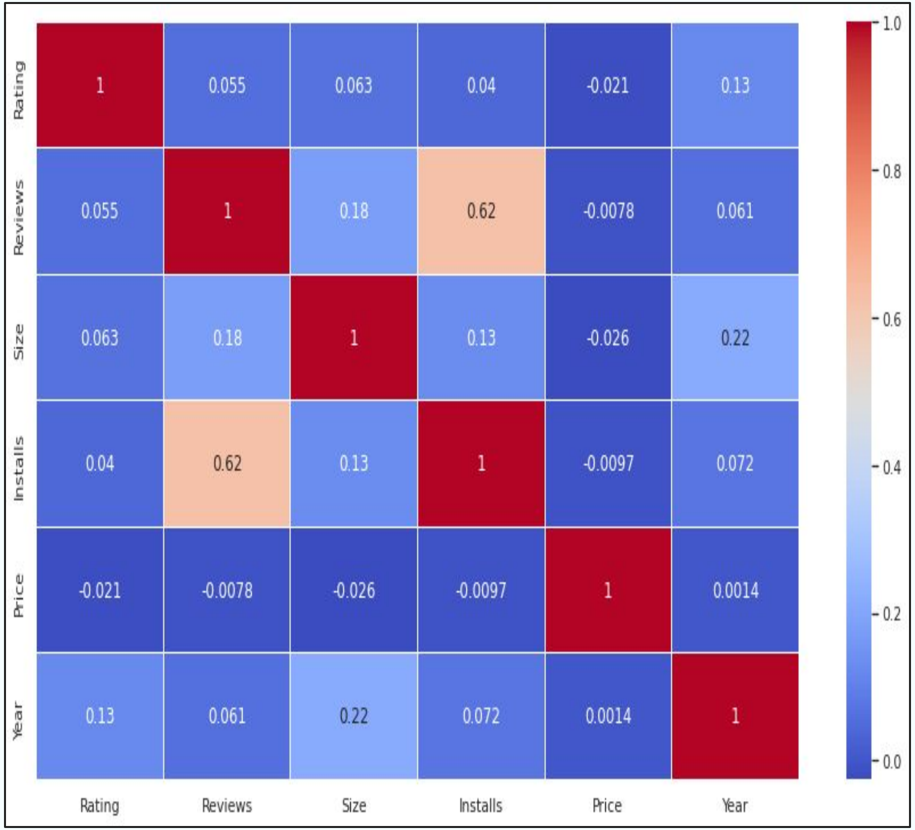
Applications which permits short range of applications version , applications which permits only latest android versions have less installs.

It's clearly visible that the applications which are updated after 2017 the amount of installations has increased drastically about 5 times of previous years which shows us that newer app have more installations.





CORRELATION



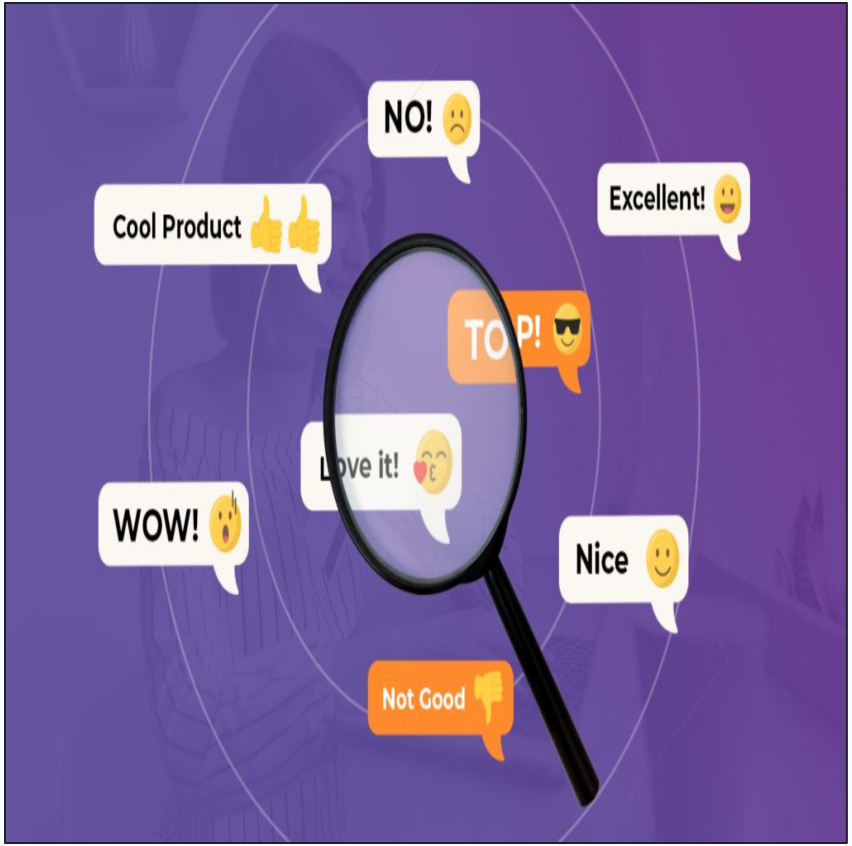
	Rating	Reviews	Size	Installs	Price	Year
Rating	1	0.055	0.063	0.04	-0.021	0.13
Reviews	0.055	1	0.18	0.62	-0.0078	0.061
Size	0.063	0.18	1	0.13	-0.026	0.22
Installs	0.04	0.62	0.13	1	-0.0097	0.072
Price	-0.021	-0.0078	-0.026	-0.0097	1	0.0014
Year	0.13	0.061	0.22	0.072	0.0014	1

1. Reviews have highest Positive Correlation with Installs with correlation value of +0.62.

2. Price is only variable which have Negative Correlation between with the Installs with correlation value of -0.0097.

3.Rating also Negatively correlated with the Price with correlation value of -0.021.

SENTIMENT ANALYSIS

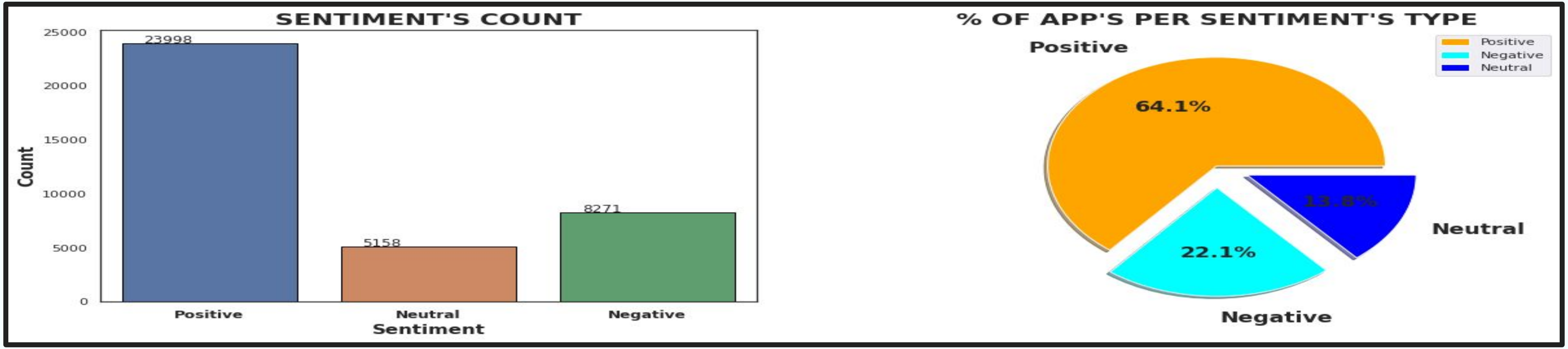


What is sentiment analysis?

Mining user review data to determine how people feel about your App or service can be done using a technique called sentiment analysis.User reviews for apps can be analyzed to identify if the mood is positive, negative or neutral about that app.

DATA SUMMARY

1. **App(Object) :** Application name
2. **Translated_Review (Object) :** User review (Preprocessed and translated to English)
3. **Sentiment (Object) :** Positive/Negative/Neutral (Preprocessed)
4. **Sentiment_Polarity (Object) :** Sentiment polarity score(>0 - Positive , < 0 - Negative)
5. **Sentiment_subjectivity(Object) :**Sentiment subjectivity score(>0.5-public opinion,<0.5-factual information)

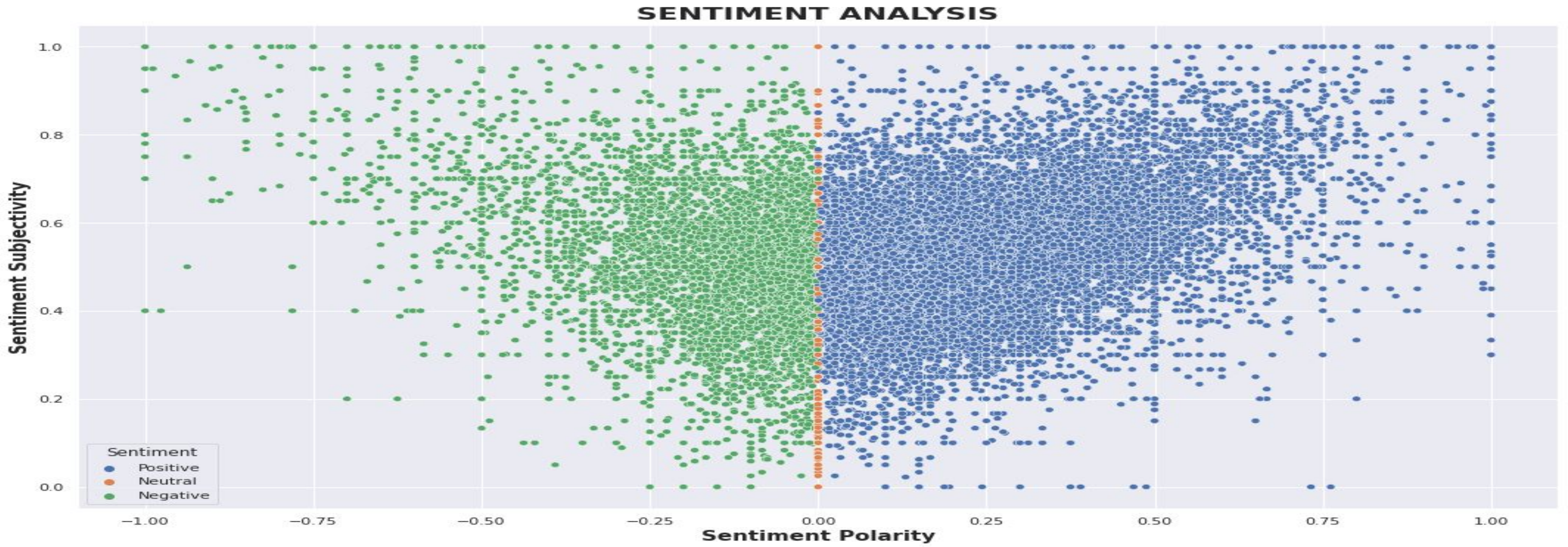


SENTIMENT POLARITY

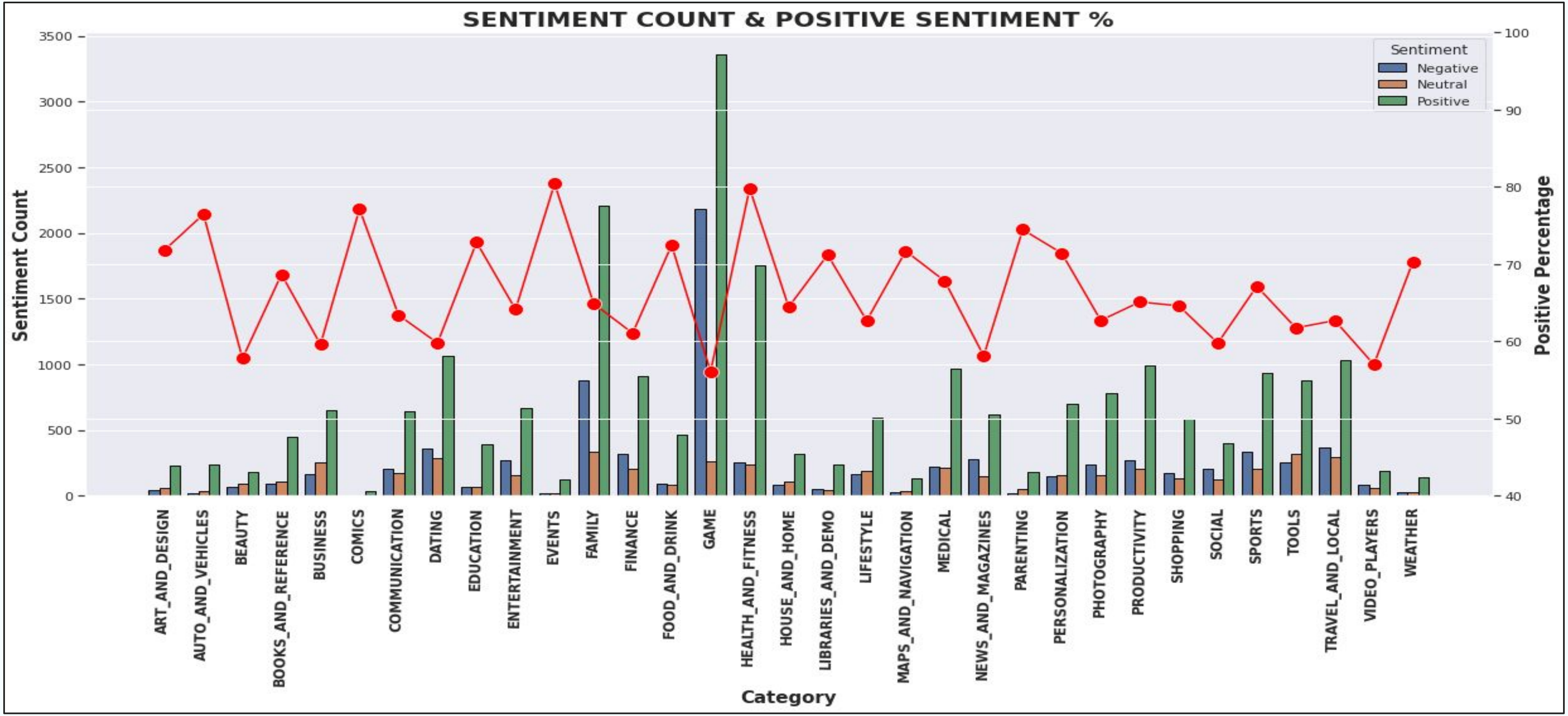
The mean of subjectivity is 0.50 and 50 % of subjectivity values lies between 0.35 to 0.65.

According to standard practices, 0.50 is a good subjectivity score, and we will not discard the values with higher subjectivity, since the mean subjectivity score for data is 0.492770..

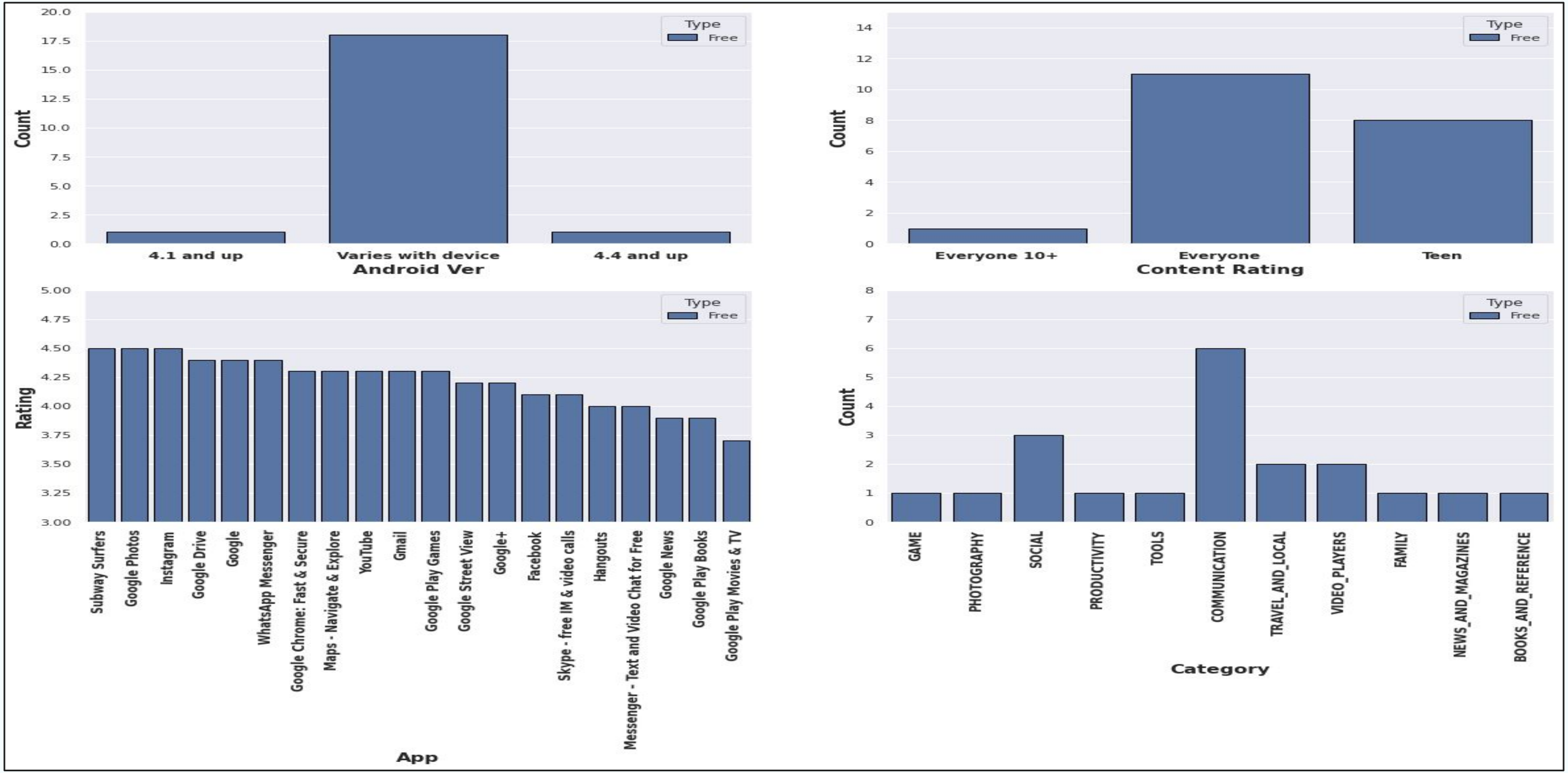
The mean value of sentiment polarity is 0.182171 and 50% of the values lie between 0 and 0.40, for this reason 64.00 % of sentiments are positive.



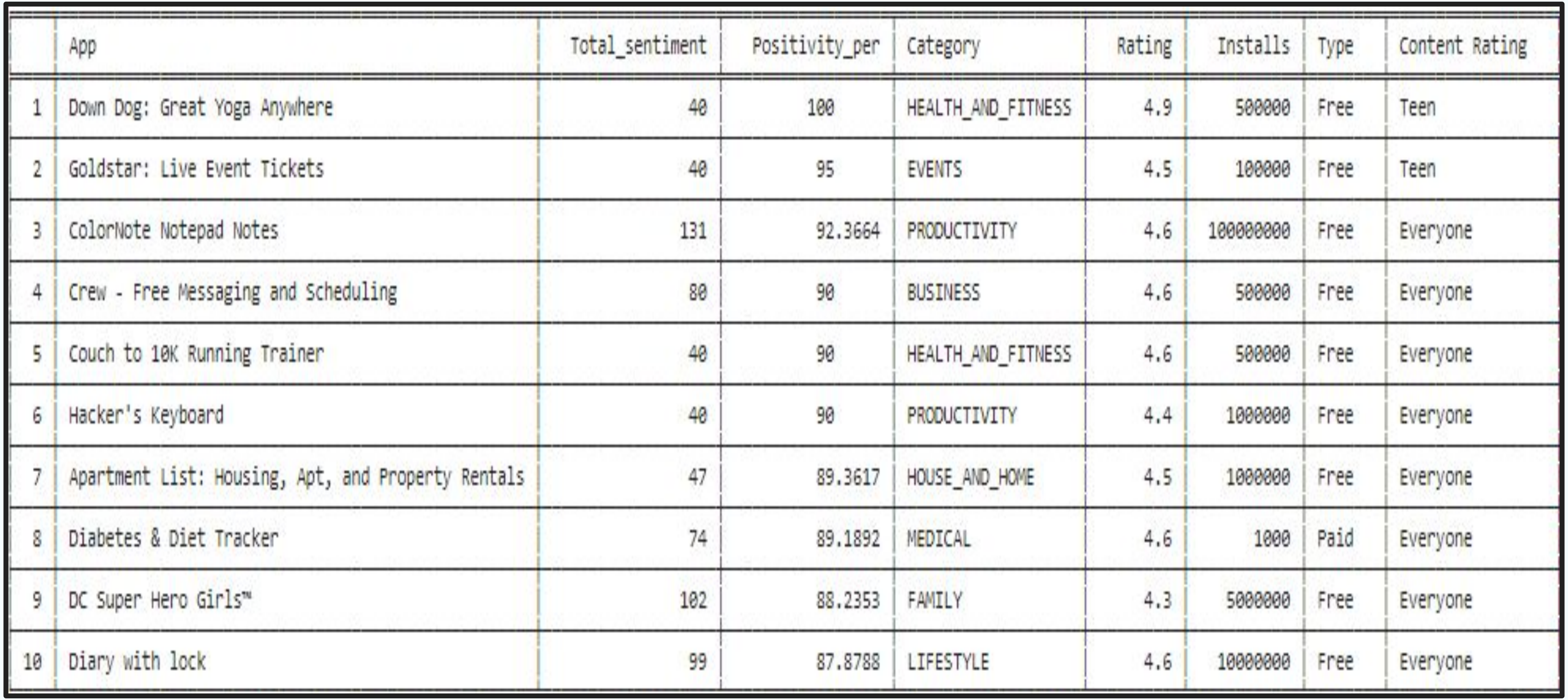
SENTIMENT ANALYSIS BY CATEGORY



APP’S WITH ONE BILLION INSTALLS



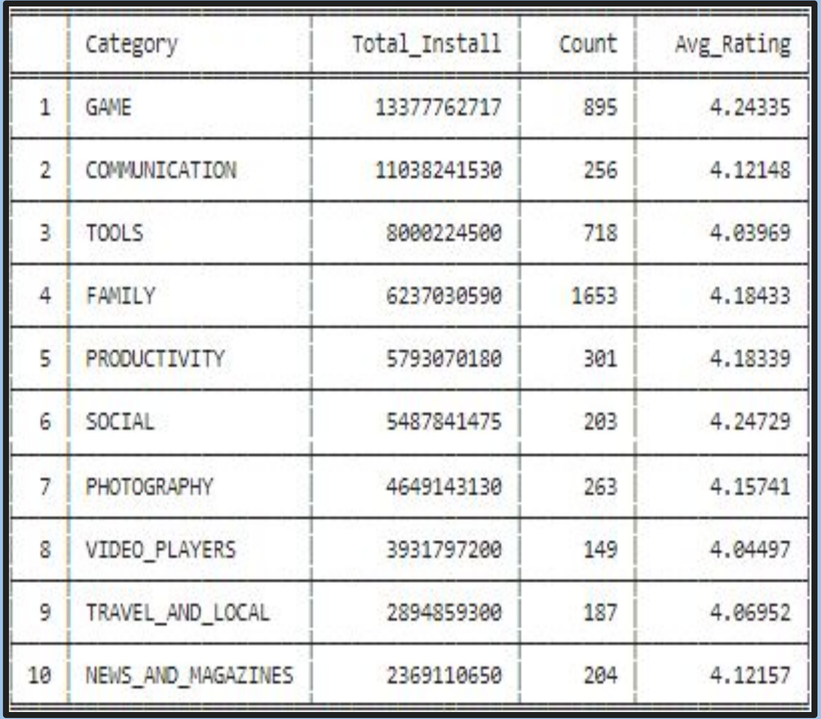
TOP 10 APPS AS PER POSITIVE SENTIMENT



	App	Total_sentiment	Positivity_per	Category	Rating	Installs	Type	Content Rating
1	Down Dog: Great Yoga Anywhere	40	100	HEALTH_AND_FITNESS	4.9	500000	Free	Teen
2	Goldstar: Live Event Tickets	40	95	EVENTS	4.5	100000	Free	Teen
3	ColorNote Notepad Notes	131	92.3664	PRODUCTIVITY	4.6	100000000	Free	Everyone
4	Crew - Free Messaging and Scheduling	80	90	BUSINESS	4.6	500000	Free	Everyone
5	Couch to 10K Running Trainer	40	90	HEALTH_AND_FITNESS	4.6	500000	Free	Everyone
6	Hacker's Keyboard	40	90	PRODUCTIVITY	4.4	1000000	Free	Everyone
7	Apartment List: Housing, Apt, and Property Rentals	47	89.3617	HOUSE_AND_HOME	4.5	1000000	Free	Everyone
8	Diabetes & Diet Tracker	74	89.1892	MEDICAL	4.6	1000	Paid	Everyone
9	DC Super Hero Girls™	102	88.2353	FAMILY	4.3	5000000	Free	Everyone
10	Diary with lock	99	87.8788	LIFESTYLE	4.6	10000000	Free	Everyone

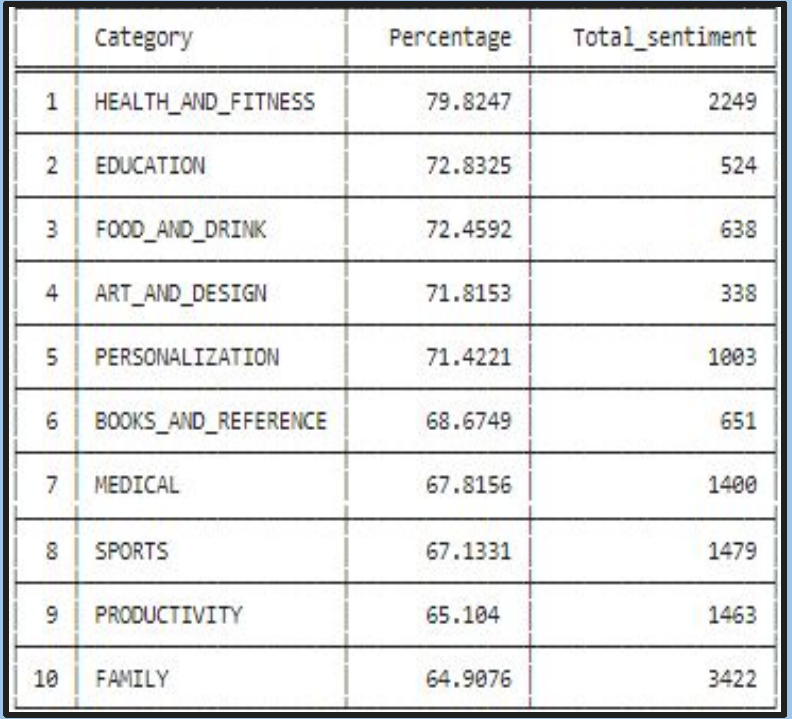
TOP 10 CATEGORY

AS PER INSTALLS



	Category	Total_Install	Count	Avg_Rating
1	GAME	13377762717	895	4.24335
2	COMMUNICATION	11038241530	256	4.12148
3	TOOLS	8000224500	718	4.03969
4	FAMILY	6237030590	1653	4.18433
5	PRODUCTIVITY	5793070180	301	4.18339
6	SOCIAL	5487841475	203	4.24729
7	PHOTOGRAPHY	4649143130	263	4.15741
8	VIDEO_PLAYERS	3931797200	149	4.04497
9	TRAVEL_AND_LOCAL	2894859300	187	4.06952
10	NEWS_AND_MAGAZINES	2369110650	204	4.12157

AS PER POSITIVE SENTIMENT



	Category	Percentage	Total_sentiment
1	HEALTH_AND_FITNESS	79.8247	2249
2	EDUCATION	72.8325	524
3	FOOD_AND_DRINK	72.4592	638
4	ART_AND_DESIGN	71.8153	338
5	PERSONALIZATION	71.4221	1003
6	BOOKS_AND_REFERENCE	68.6749	651
7	MEDICAL	67.8156	1400
8	SPORTS	67.1331	1479
9	PRODUCTIVITY	65.104	1463
10	FAMILY	64.9076	3422

CONCLUSION

1. It is hard to conclude an app's success solely from its Rating since the average Rating is 4.17 and the majority of apps get ratings between 3.8 and 4.5.

2. The number of installations of app is very effective measure for app's success and engagement. Only 20 app's have more than Billion installations.

3. From the data it's clearly visible that User's like to install Free applications. Approximately 93% of the apps in the Google Play Store are free to download, and the price is negatively correlated with both installs and ratings.

4. Small sized applications are installed more, but for higher installed applications amounts, big sized apps are also more prevalent.

5. Even though category GAME has the highest total installs, category COMMUNICATION is the most successful category due to the highest average installs, higher percentage of positive sentiment, and six apps with more than billion installs compared to just one app from category GAME.

6. Apparently, users like freedom, as most of the Top apps according to Installs and Positive Sentiment percentage fall under content rating EVERYONE and Android version VARIES BY DEVICE.

7. App’s updated at latest have the highest installs and good rating, which indicates user's tends to download apps which are latest.

8. Positive sentiments for app's under the category HEALTH_AND_FITNESS are the highest.

9. About 63% of sentiments are positive and the Top 10 categories also have a positive sentiments percentage around 60%, except GAME which has a 56% positivity rate.

THANK

YOU !