Capstone Project

Play store App Review Analysis

By,
Yamini.

Steps followed in analysis of Play store data

1) Understanding problem statement
2) Viewing the data
3) Data understanding and summary
4) Missing value analysis
5) Data Cleaning and transformation
6) Feature engineering
7) Outliers detection
8) Correlation plots
9) Feature important variables
10) Data Visualization
11) Some case studies

Problem statement

- The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market.
- Each app (row) has values for category, rating, size, and more. Another dataset contains customer reviews of the android apps.
- **Explore and analyze the data to discover key factors responsible for app engagement and success.**

Dataset1 – Play store data(First 5 rows using head())

	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
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

Dataset2 – User reviews data

	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
3	10 Best Foods for You	Works great especially going grocery store	Positive	0.40	0.875000
4	10 Best Foods for You	Best idea us	Positive	1.00	0.300000

Data understanding and summary

- There are two datasets given for analysis in which play store data tells about the apps and its related success factors with their updated details whereas user reviews dataset only tells about the users sentiment with their reviews on different apps.
- There are 10841 rows,13 columns in Play store data and there are 64295 rows, 5 columns in the User reviews data.
- When I check for datatypes in play store data it says 11 object variables and 1 float variable which is not how it has to be because there are some variables which are categorized as object instead it has to be in float and vice-versa.
- In User reviews dataset there are 3 object and 2 float variables.

Summary of datasets

`Store_df.describe()`

	Rating
count	9367.000000
mean	4.193338
std	0.537431
min	1.000000
25%	4.000000
50%	4.300000
75%	4.500000
max	19.000000

`Users_df.describe()`

	Sentiment_Polarity	Sentiment_Subjectivity
count	37427.000000	37427.000000
mean	0.182171	0.492770
std	0.351318	0.259904
min	-1.000000	0.000000
25%	0.000000	0.357143
50%	0.150000	0.514286
75%	0.400000	0.650000
max	1.000000	1.000000

Missing value analysis

- There are two ways of dealing with missing values. They are either to drop them entirely or to impute them using different techniques like statistical techniques(mean,median,mode), KNN imputer, based on domain knowledge, random imputation ..etc
- In Play store data there are more missing values in rating column than any other columns which have only few missing values. For this dataset I have used Systematic random sampling imputation method which preserves the variance and distribution of the variable.
- In User reviews dataset there are many missing values and the dataset is also huge compared to play store data. So I have dropped the missing values because its not possible to impute the data for review column which gives sentiment scores for that review. So, here dropping is the best option.

Data Cleaning and transformation

- There is a lot of data cleaning and pre-processing work to be done before making any analysis on play store data as the raw data is not so good for analysis.
- Firstly, when I checked for datatypes they do not seem to be the correct datatypes for the variables. So I have changed them after converted the values in them to proper format.
- For example, the size and price columns cannot be object variables. So converted them to numeric by making required changes so that the data is not distorted. In price column there is dollar sign attached to the values so I have removed punctuations(,) and currency symbol or notation. I have renamed the column as Price($) and size contains ‘MB’,’KB’ so I have created new column with units(MB,KB) and size values as another column.

Feature engineering

- New features like Size(MB,KB), Size values(actual given),Size(all in KB unit),Last updated month, last updated year, last updated quarter have been introduced for better analysis of Play store data.
- App category data values are capitalized for better representation and Installs column is also renamed as Installs(+) after removing the punctuations from it.
- In Size column there are some rows which says that the app size varies with device so that rows have replaced with 11.5M which is the average size of the app in the android device as play store is only contained in android devices.
- There is not much changes done in User reviews dataset as it contains the reviews and sentiment scores. I have not applied nlp techniques here because as it only analysis we can read the reviews this way better than systems way of removing stop words and punctuations. So User reviews data is ready to use for analysis and finding insights from them.

Numeric data after data cleaning in play store dataset

```
Store_df.describe()
```

	Rating	Reviews	Size	Installs(+)	Price($)	Size(Values as given)	Last_updated_day
count	10841.000000	1.084100e+04	10841.000000	1.084100e+04	10841.000000	10841.000000	10841.000000
mean	4.206485	4.441136e+05	20426.899908	1.546291e+07	1.027273	33.039627	15.609353
std	0.480321	2.927628e+06	21569.790208	8.502557e+07	15.948971	91.283080	9.561235
min	1.000000	0.000000e+00	1.000000	0.000000e+00	0.000000	1.000000	1.000000
25%	4.100000	3.800000e+01	6041.600000	1.000000e+03	0.000000	6.800000	6.000000
50%	4.300000	2.094000e+03	11776.000000	1.000000e+05	0.000000	11.500000	16.000000
75%	4.500000	5.476800e+04	26624.000000	5.000000e+06	0.000000	28.000000	24.000000
max	5.000000	7.815831e+07	102400.000000	1.000000e+09	400.000000	1020.000000	31.000000

Categorical columns in Play store data – Last 3 rows

```
Store_col_df.tail(3)
```

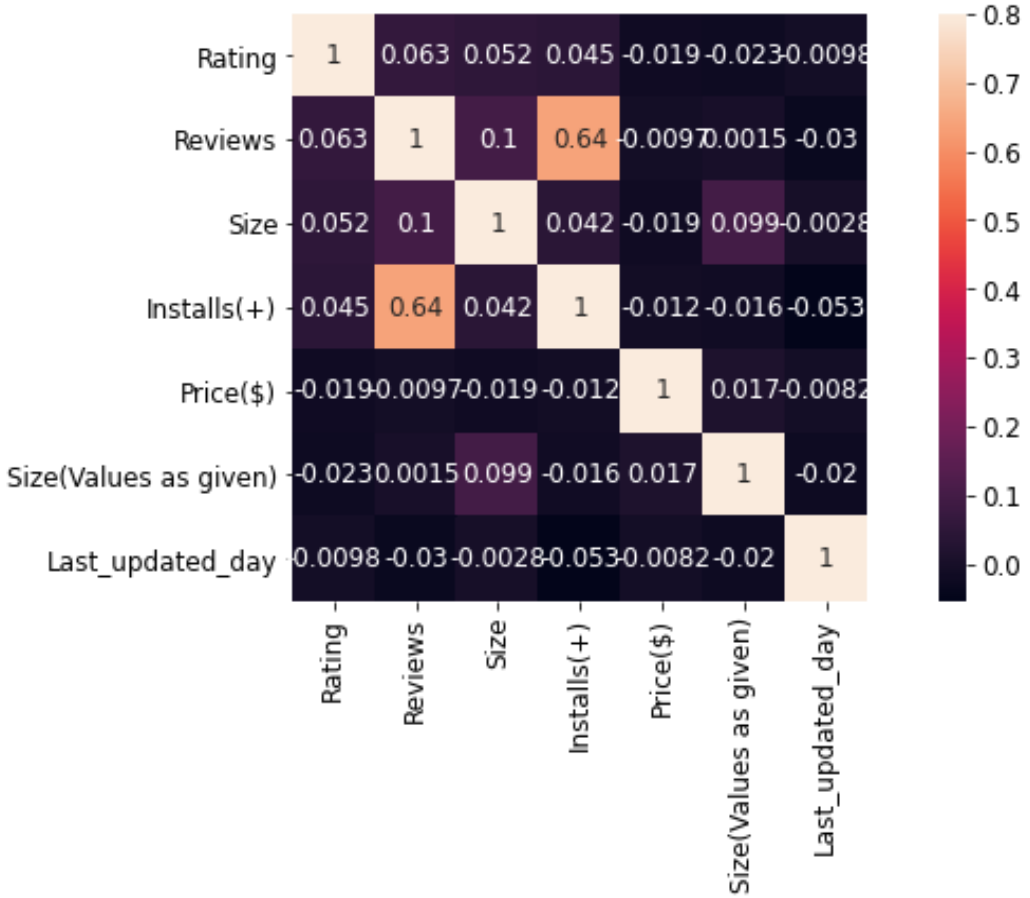
	App	Category	Type	Content Rating	Genres	Current Ver	Android Ver	Size(MB,KB)	Last_updated_month	Last_updated_Year	Last_updated_quarter
0838	Parkinson Exercices FR	Medical	Free	Everyone	Medical	1.0	2.2 and up	MB	Jan	2017	1
0839	The SCP Foundation DB fr nn5n	Books_and_reference	Free	Mature 17+	Books & Reference	Varies with device	Varies with device	Varies with device	Jan	2015	1
0840	iHoroscope - 2018 Daily Horoscope & Astrology	Lifestyle	Free	Everyone	Lifestyle	Varies with device	Varies with device	MB	July	2018	3

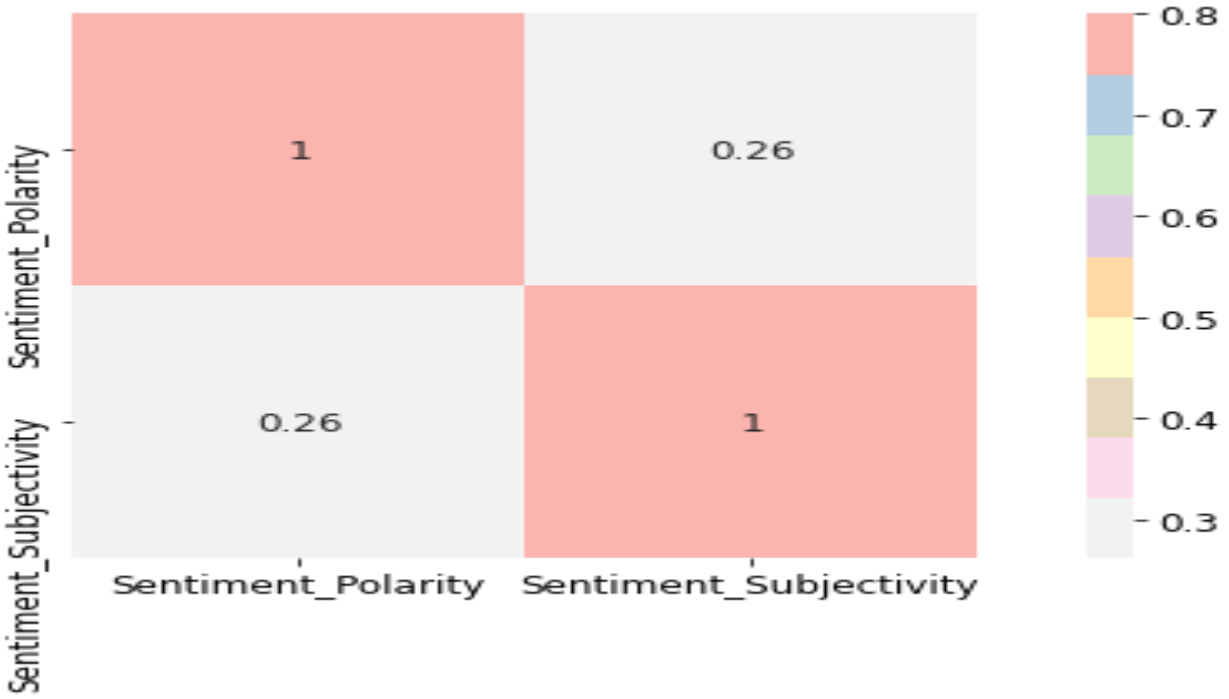
User reviews dataset

- There are 4 features which gives the information about user engagement with the andriod apps in the play store.
- Here there is data about the translated reviews given by the users, Sentiment polarity, Sentiment subjectivity and Sentiment for a given app.
- **Polarity** is float which lies in the range of [-1,1] where value is near to 1 means positive statement and values near to -1 means a negative statement.
- **Subjectivity** generally refers to personal opinion, emotion or judgment whereas objective refers to factual information. Subjectivity is also a float which lies in the range of [0,1].
- For example, if subjectivity value is near to 1 means its a public opinion and if its near to 0 means its a factual data.
- Sentiment finally gives whether it is positive, negative or neutral statement/reviews for a given app.

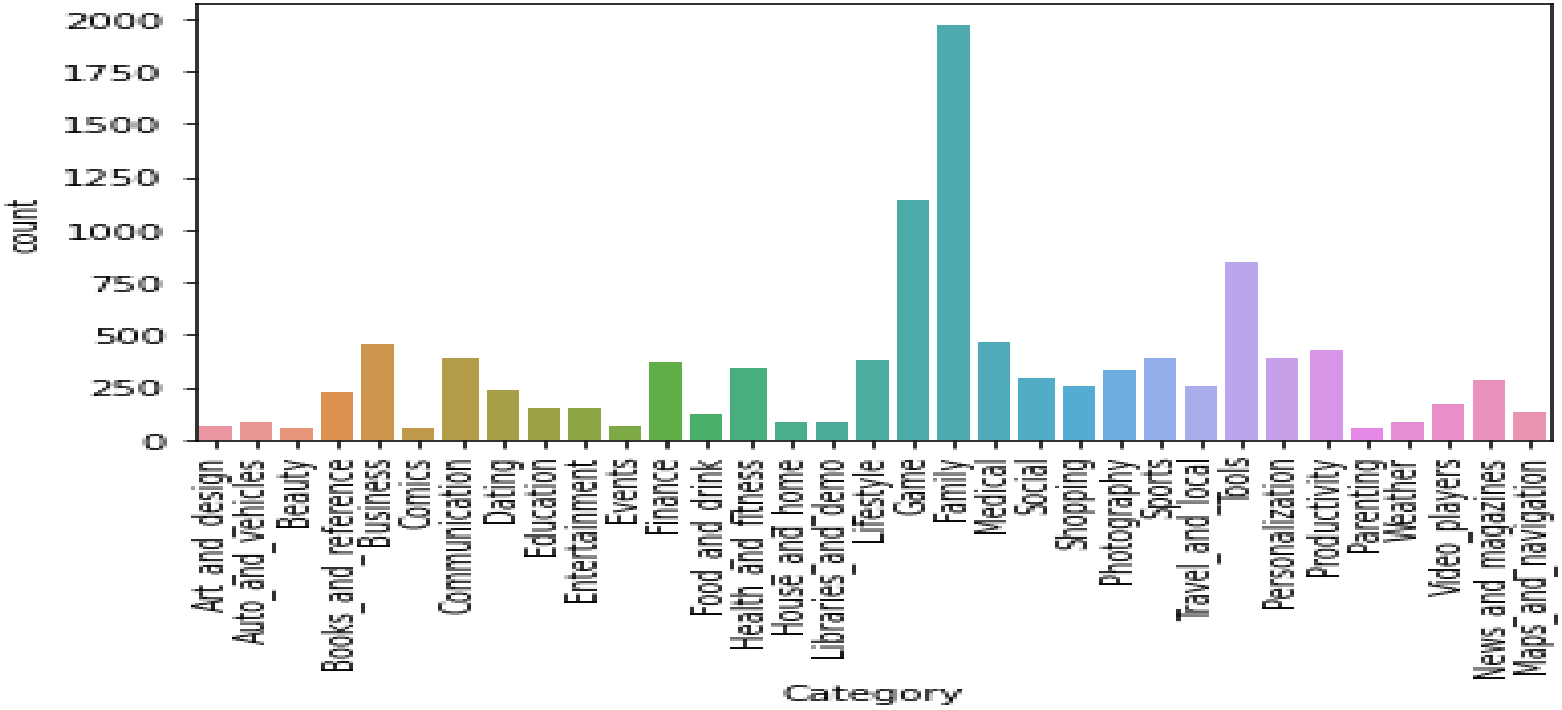
Data Visualization and analysis – Correlation plot

- The correlation between Reviews and Installs is higher and both are strongl y correlated. Hence multi-collinearity exists between these two features.
- Size and Reviews are next correlated variables with correlation of 0.1.
- All the other variables are not much correlated with each other.
- Based on the requirements or predict or variable only uncorrelated variable ('s)are chosen and whichever is have good correlation with target variable are preferred.

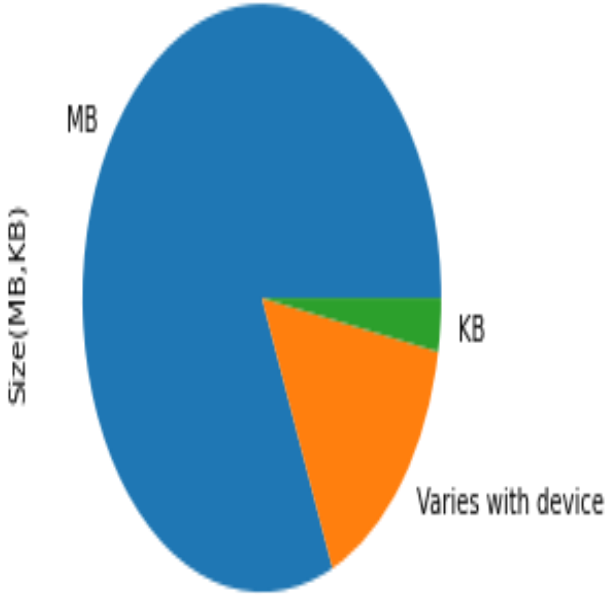


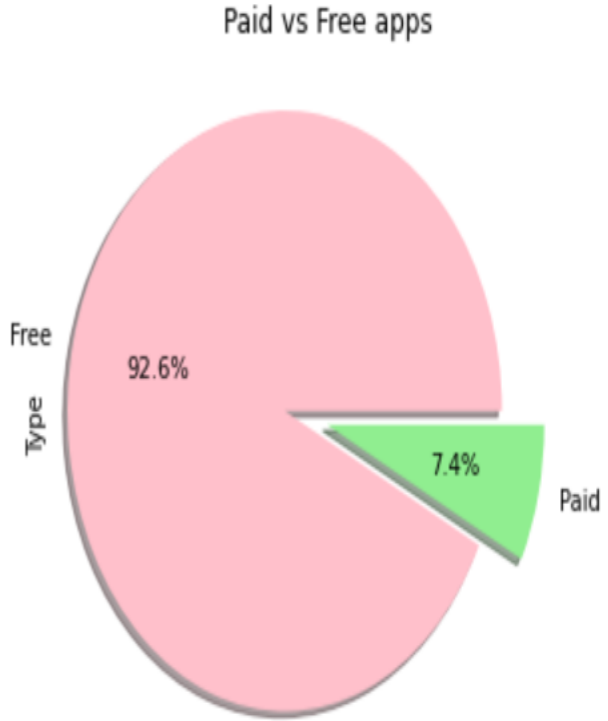


User Reviews dataset correlation plot

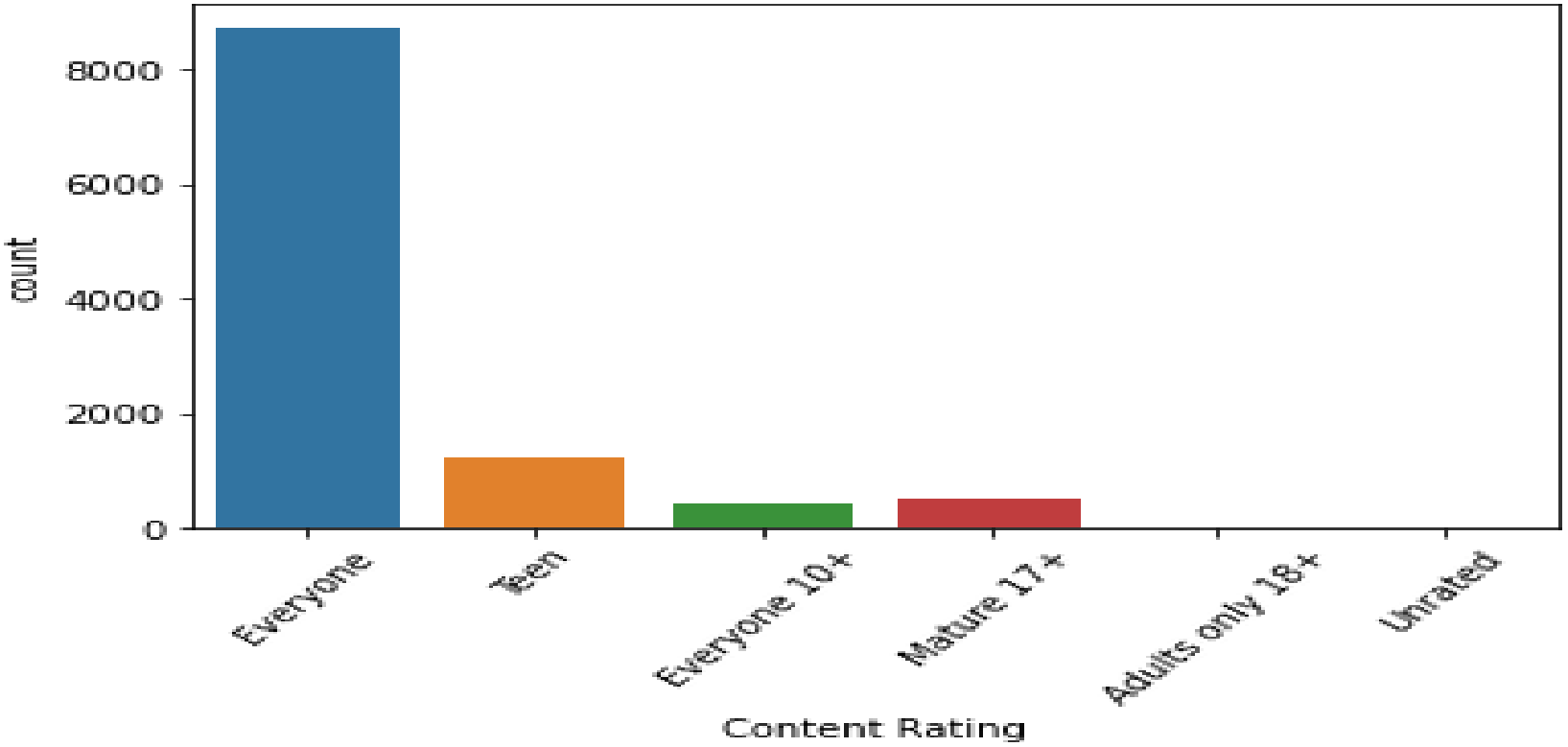


Different category apps in play store data

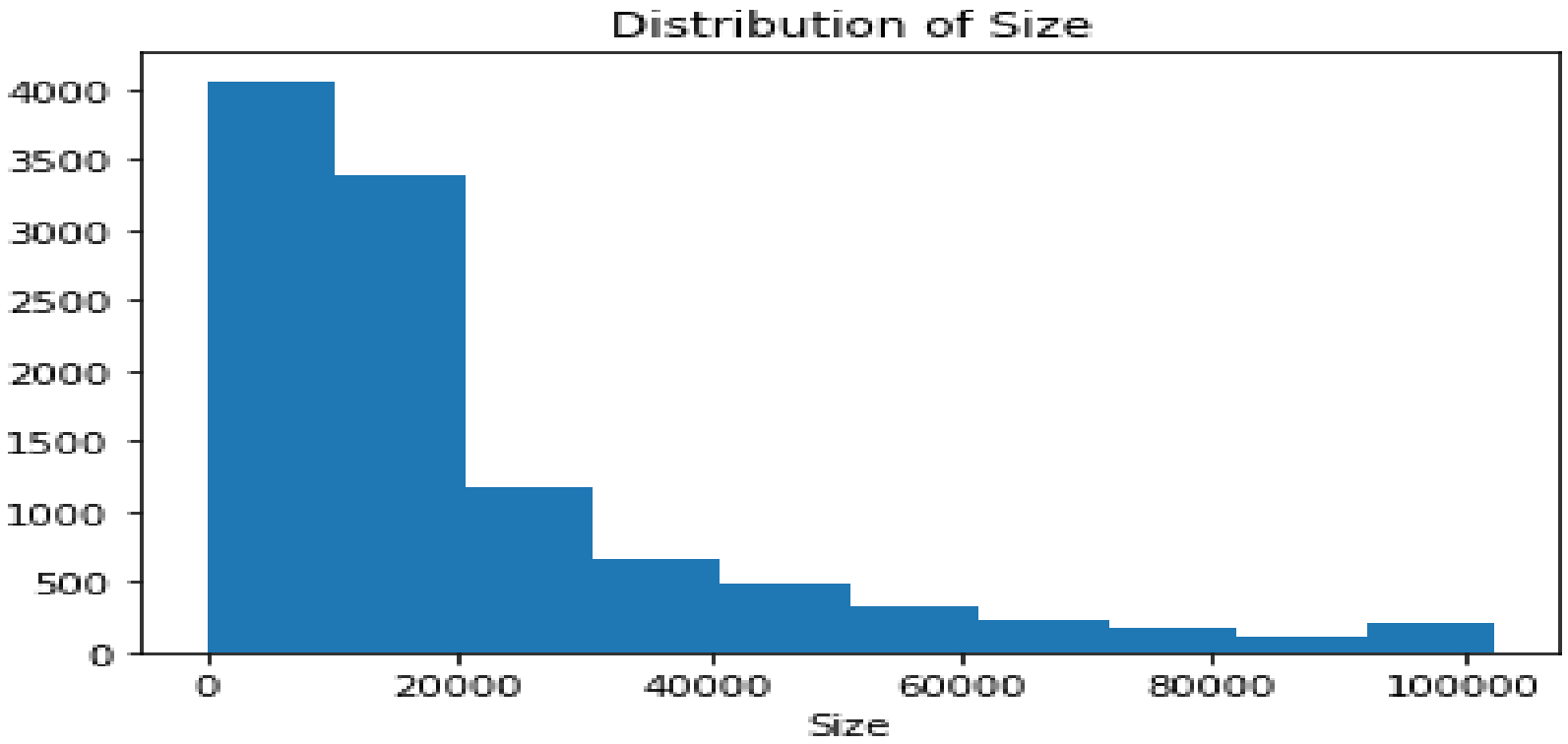




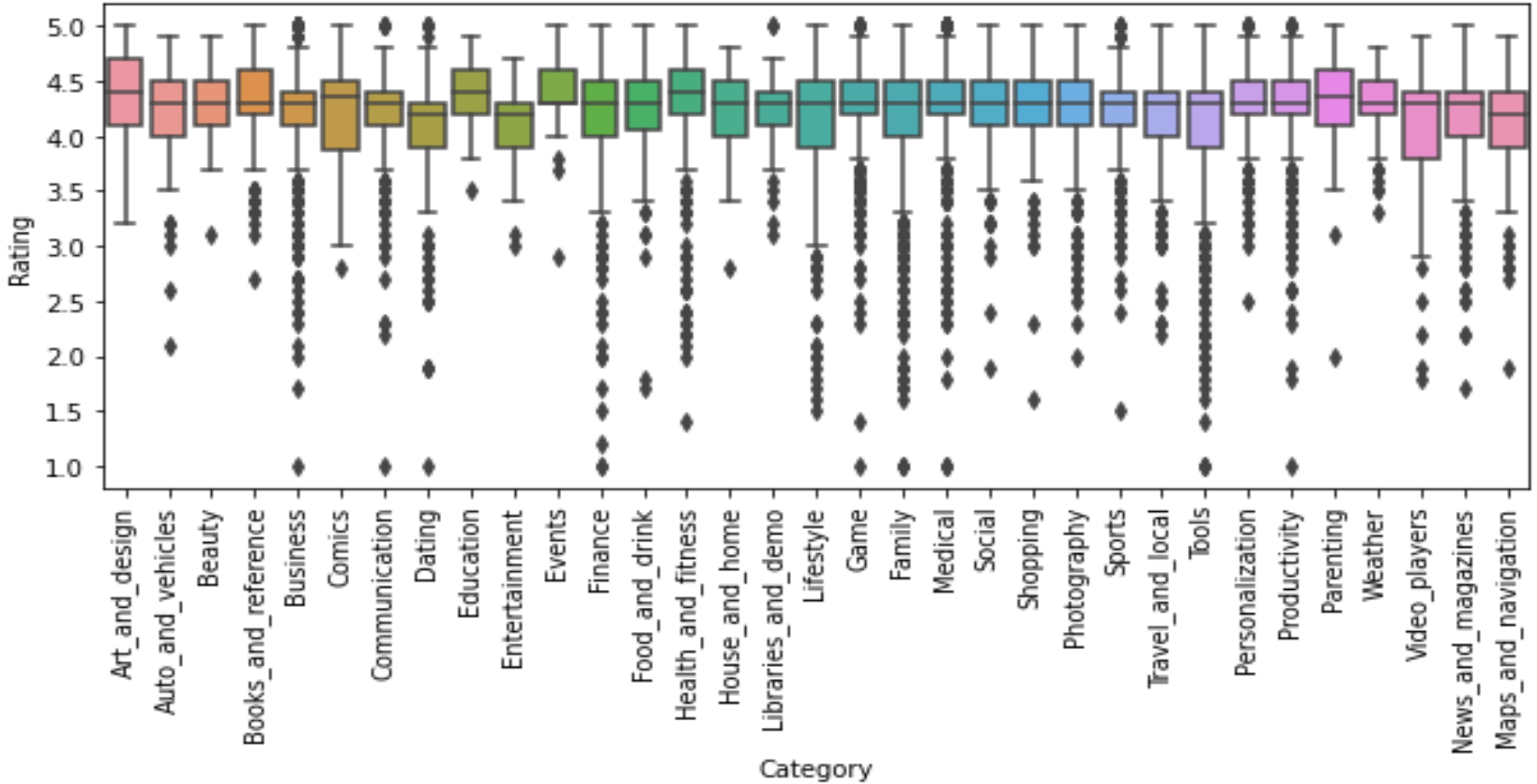
Pie chart representing the Size and Type of Apps distribution



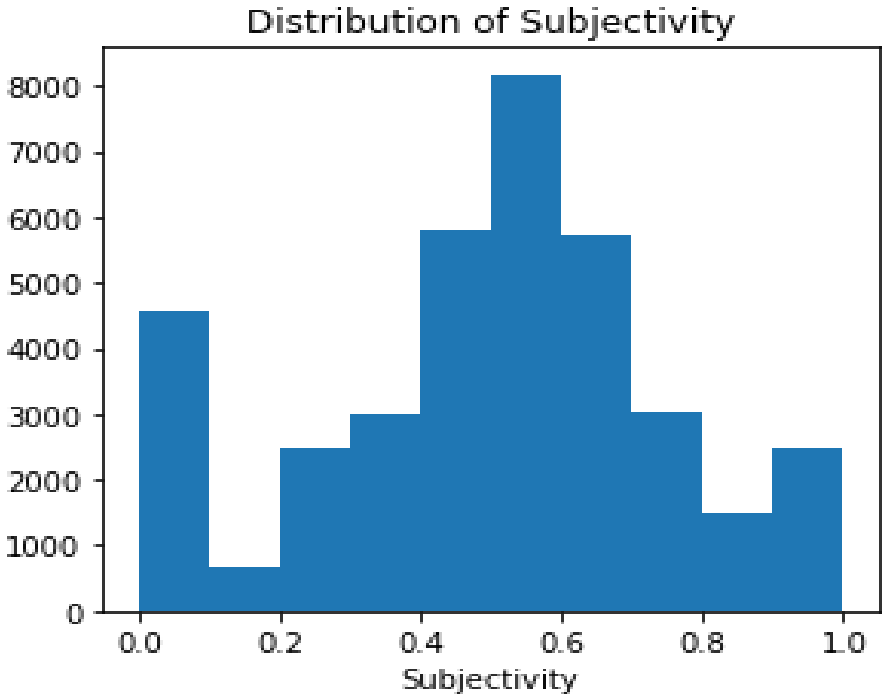
Content Rating of Apps – There are more apps which can be used by everyone

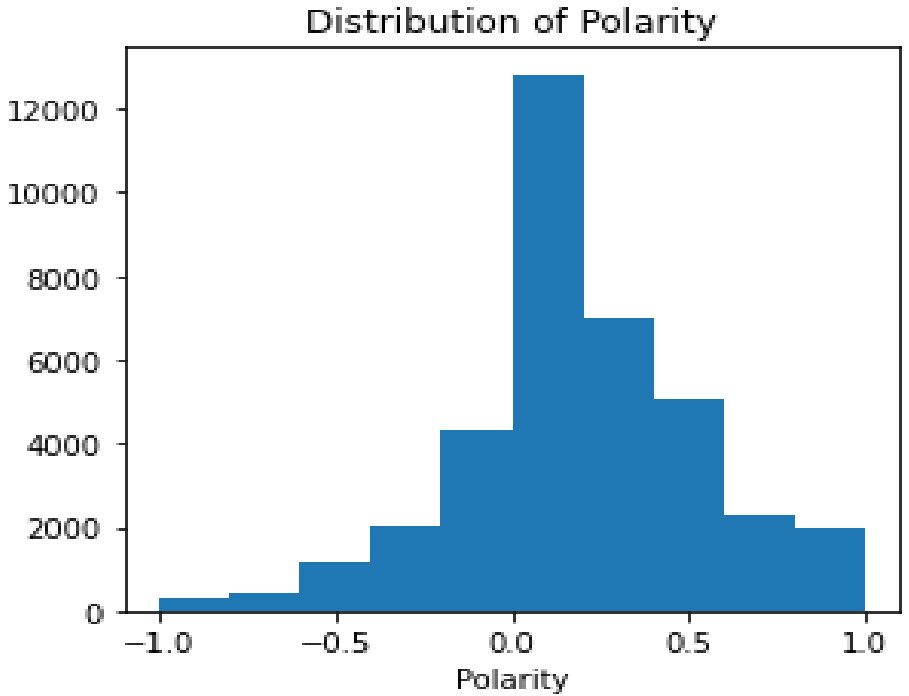


Size in KB – There are few apps with huge size(outliers)



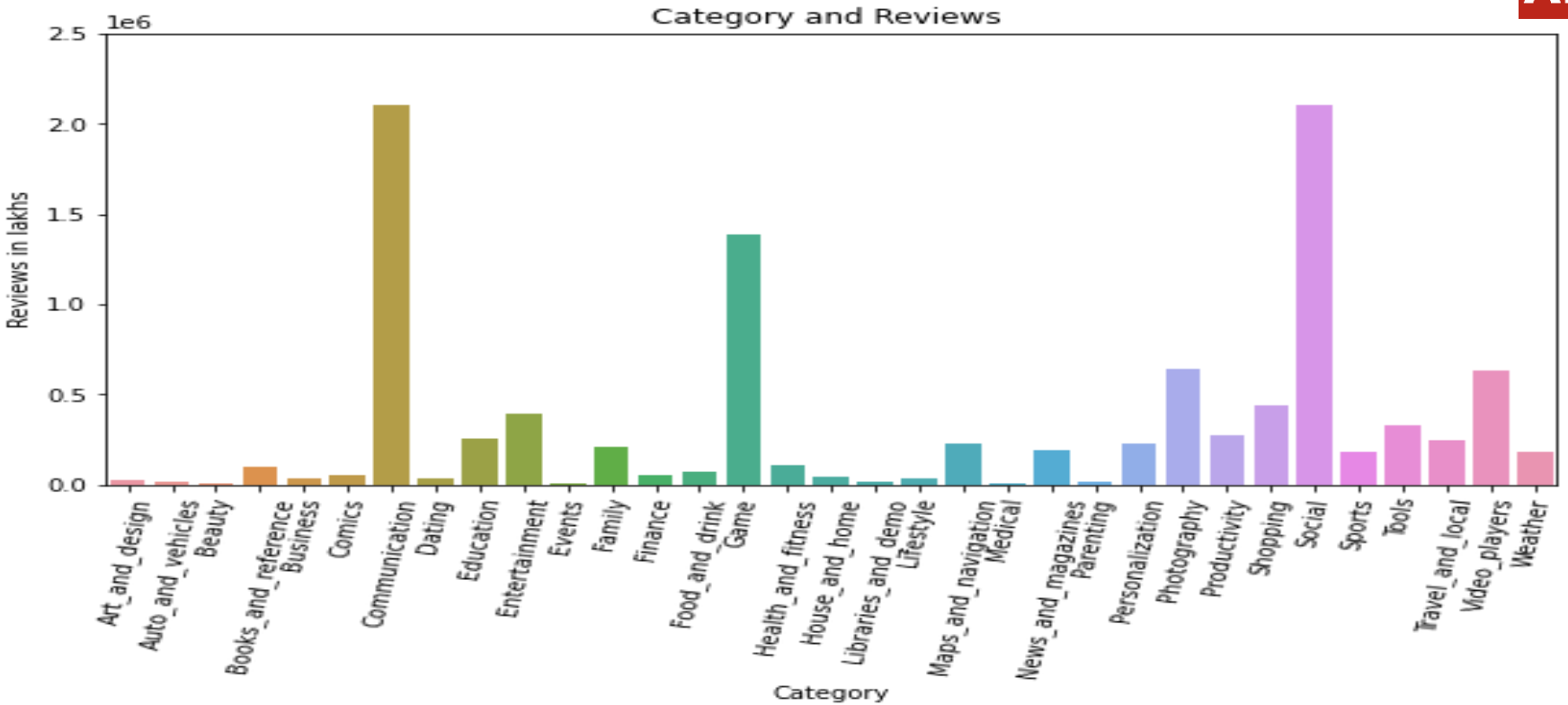
Category and rating of the apps – There are outliers where the rating is below 3.0 for most of the categories



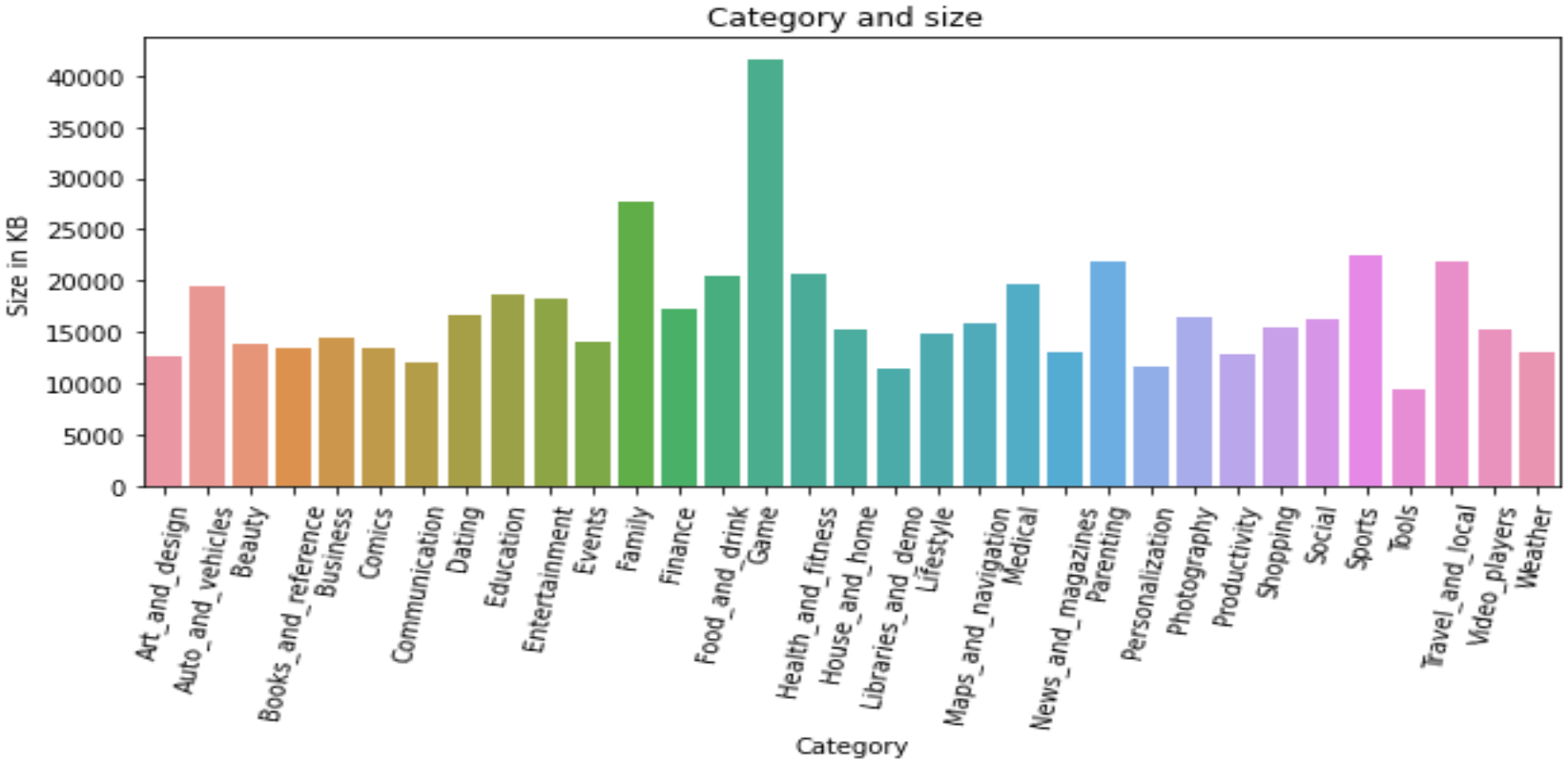


Subjectivity and Polarity in User reviews data – Polarity lies in between [-1,1] whereas subjectivity from [0,1] which says about sentiment(negative,positive) and opinion of the users respectively.

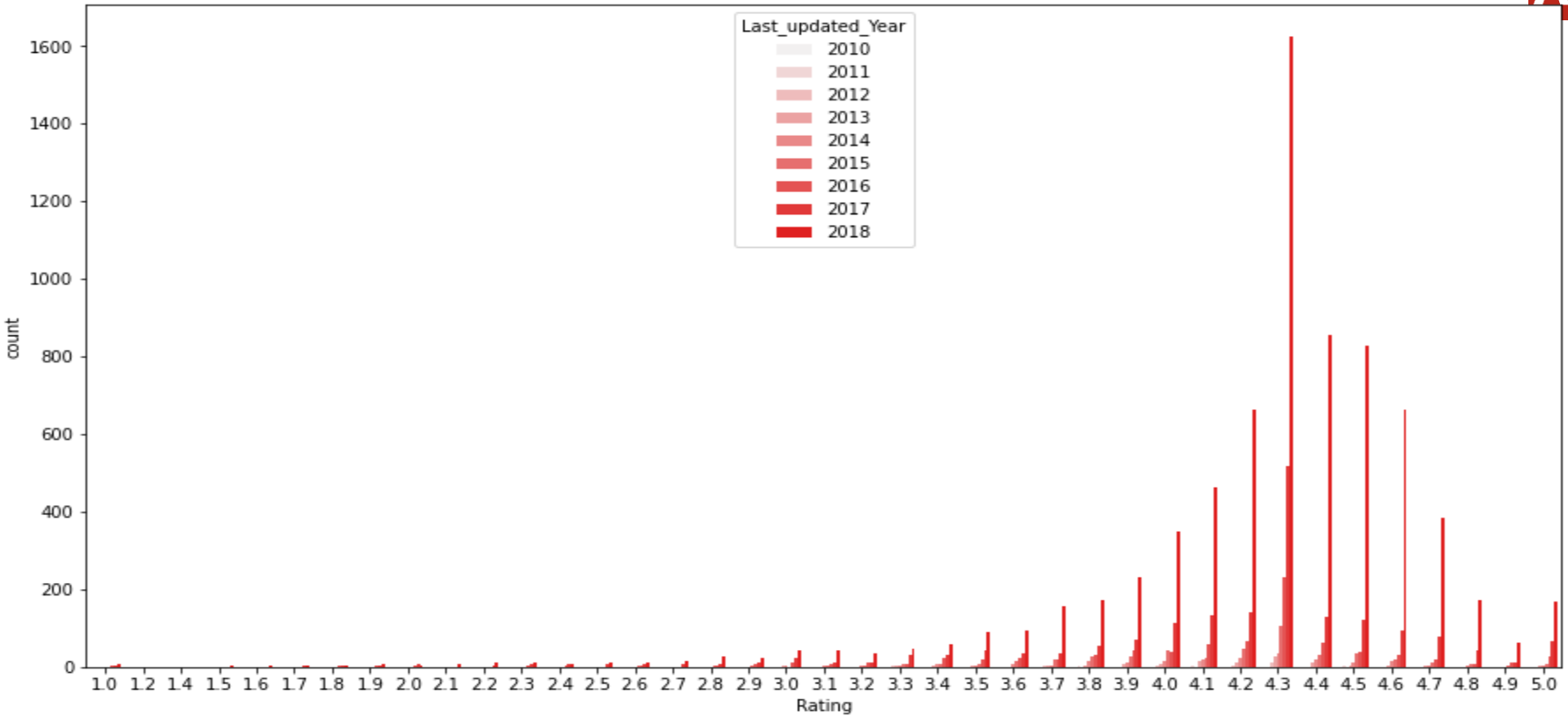
Most used words from the translated reviews feature in User reviews dataset



Category vs reviews for the play store apps – Bivariate analysis



Category vs Size of the app – Game category apps have high size than other apps



Ratings acquired in different year from 2010 to 2018. Ratings have increased in 2018 indicating grows and success of the apps

Conclusion

- The features in Play store data that mostly helps in **predicting the success rate of an app are the Rating, reviews, Installs and type of an app.**
- The features in Users review dataset that would help in the success rate or by which it can be said **that the app is successful or not. They are firstly Sentiment is useful then Sentiment polarity and subjectivity.**
- From the analysis it is seen that there are good number of apps with positive reviews than negative and neutral reviews.
- There are many categories of apps present in the play store and **the apps that are high in particular category is Communication and Social apps. It says these apps are more successful and have high app engagement.**
- Also **there is an increasing trend that can be observed from 2010 to 2018, the users who have rated the apps have increased. So the success rate of the apps has also grown.**
- Finally it can be said that app business is growing and all the category apps have shown significant improvement from initial years to present days. There are also many new improved versions and features in the apps which has influenced the app business to grow and succeed.

Thank you