Capstone Project – 1

Exploratory Data Analysis

Team **Champions** : Play Store App Review Analysis

Team Members

Varsha Rani

Vivek Chandrakant Pawar

Rabista Parween

Tushar Gaikwad

POINTS FOR DISCUSSION

- Problem Statement & Objective
- Data Summary
- Load the Data and Create Data Frame
- Univariate Analysis
- Skewness of Data
- Bivariate Analysis
- Correlation of Data
- Most installed category
- Content Rating vs Installs
- Review vs Rating
- App Size Distribution
- Sentiment Subjectivity Distribution
- Sentiment Distribution
- Sentiment Polarity Distribution
- Category vs Review vs Sentiment
- Conclusion

Problem Statement & Objective

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.

Here we are exploring and analyzing the data to discover key factors responsible for app engagement and success.

We have used the following Libraries :

- Pandas - manipulation of tabular data in Dataframes
- Numpy - mathematical operations on arrays
- Matplotlib - visualization
- Seaborn - visualization
- Plotly Express - visualization

GET IT ON Google Play

The objective of this project is to deliver insights to understand customer demands better and thus help developers to popularize the product.

Data Summary

df

It contains information about applications, having column like app name, size, installs,etc.

User_review_df

It contains user reviews for application containing information like Sentiment Distribution, Sentiment Polarity Distribution and Sentiment Subjectivity Distribution.

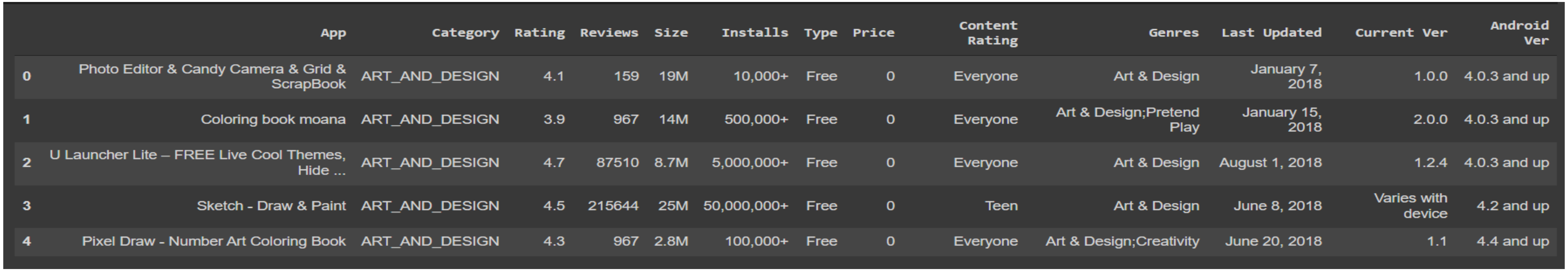
Merge_df

It contains combined information of all applications and user reviews sentiments.

Load the Data and Create Data Frame

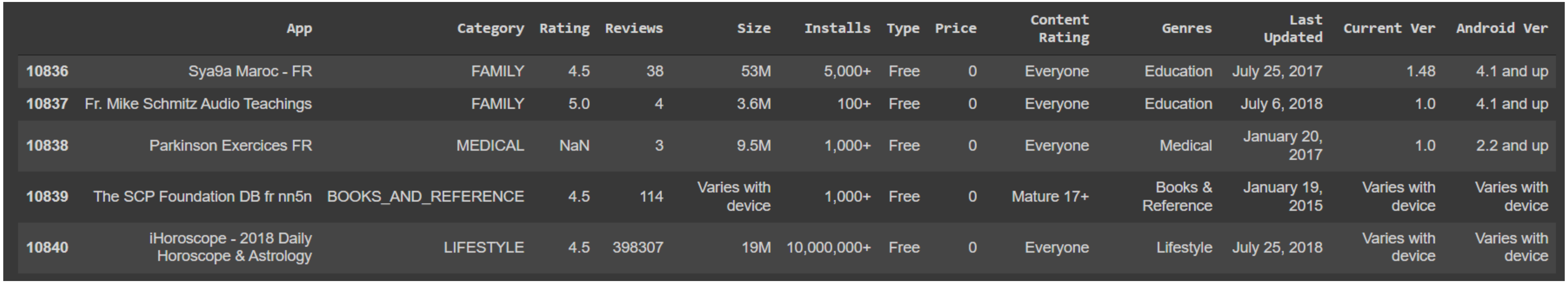
Data Frame creation using pandas

app_df.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

app_df.tail()



	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	0	Everyone	Education	July 25, 2017	1.48	4.1 and up
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3.6M	100+	Free	0	Everyone	Education	July 6, 2018	1.0	4.1 and up
10838	Parkinson Exercices FR	MEDICAL	NaN	3	9.5M	1,000+	Free	0	Everyone	Medical	January 20, 2017	1.0	2.2 and up
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	Varies with device	1,000+	Free	0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Varies with device
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19M	10,000,000+	Free	0	Everyone	Lifestyle	July 25, 2018	Varies with device	Varies with device

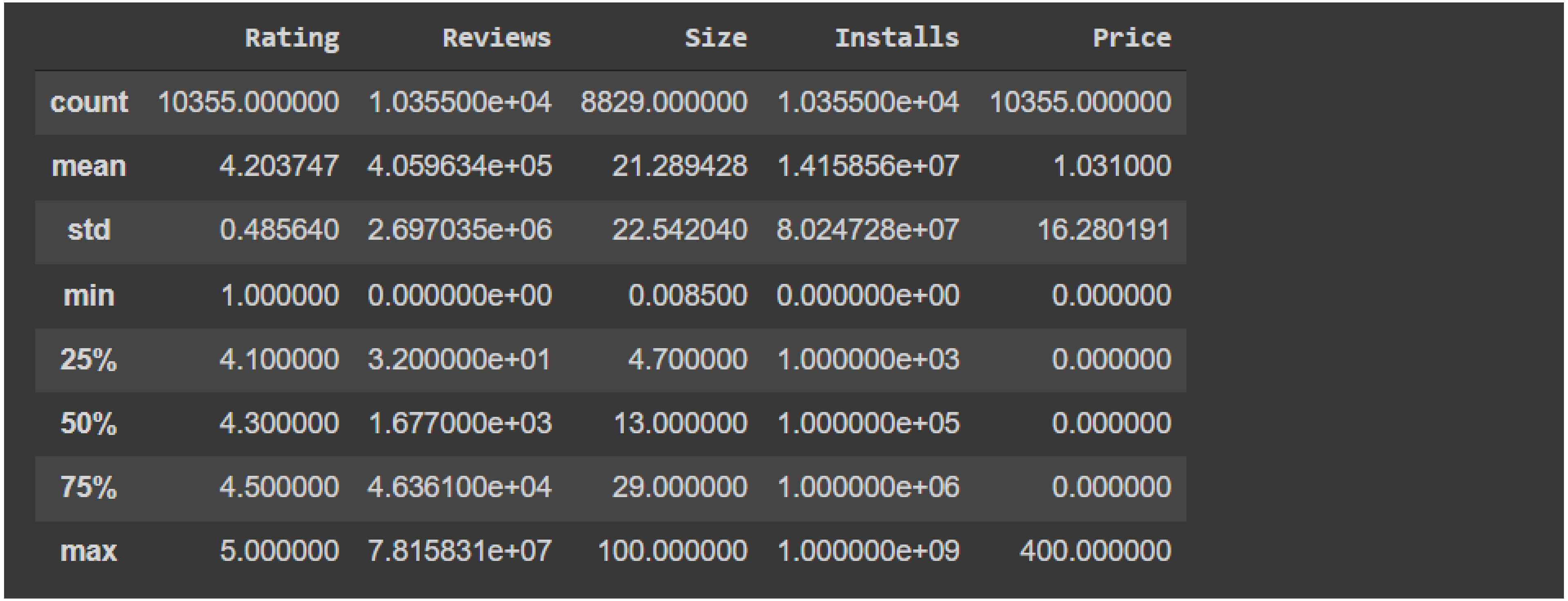
Cleaning the DataFrame

- Dropping Duplicates
- Checking the null values
- Filling the null values
- Dropping improper values within columns
- Converting the columns datatype

Univariate Analysis

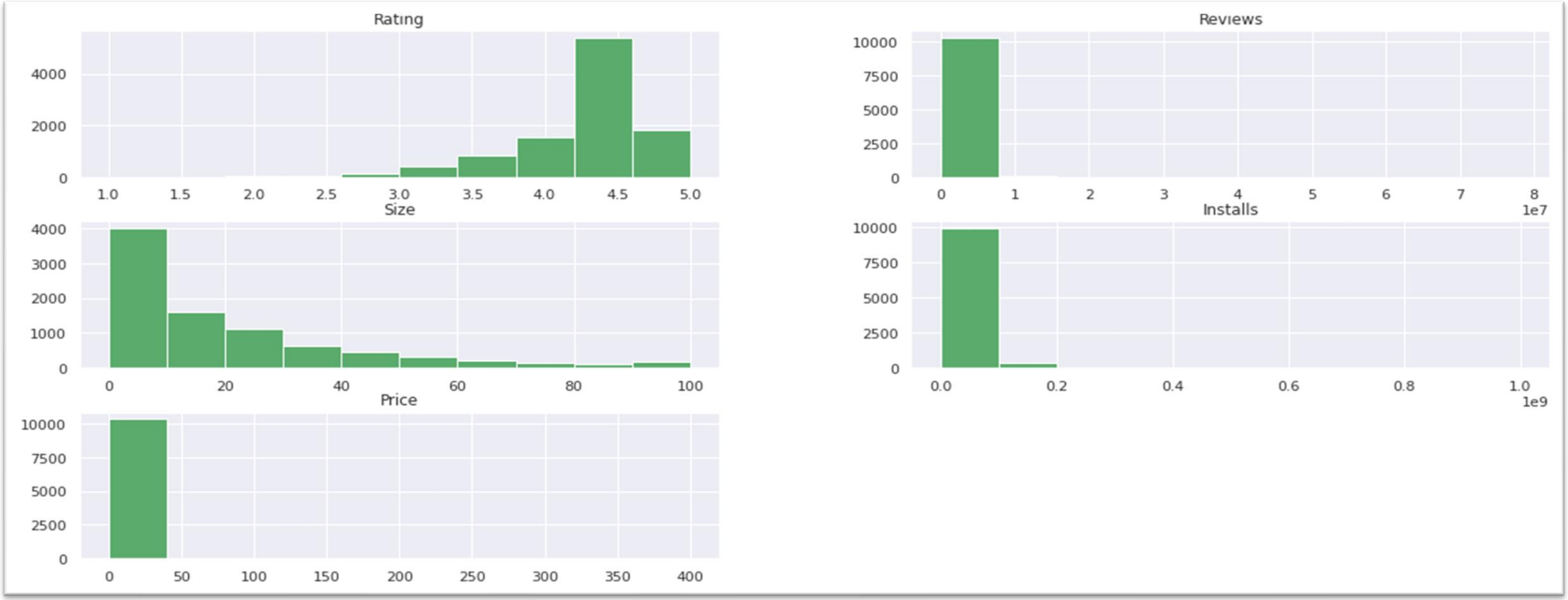
Univariate analysis is perhaps the simplest form of statistical analysis. Like other forms of statistics, it can be inferential or descriptive. The key fact is that only one variable is involved.

Statistical Data



	Rating	Reviews	Size	Installs	Price
count	10355.000000	1.035500e+04	8829.000000	1.035500e+04	10355.000000
mean	4.203747	4.059634e+05	21.289428	1.415856e+07	1.031000
std	0.485640	2.697035e+06	22.542040	8.024728e+07	16.280191
min	1.000000	0.000000e+00	0.008500	0.000000e+00	0.000000
25%	4.100000	3.200000e+01	4.700000	1.000000e+03	0.000000
50%	4.300000	1.677000e+03	13.000000	1.000000e+05	0.000000
75%	4.500000	4.636100e+04	29.000000	1.000000e+06	0.000000
max	5.000000	7.815831e+07	100.000000	1.000000e+09	400.000000

Skewness of Data

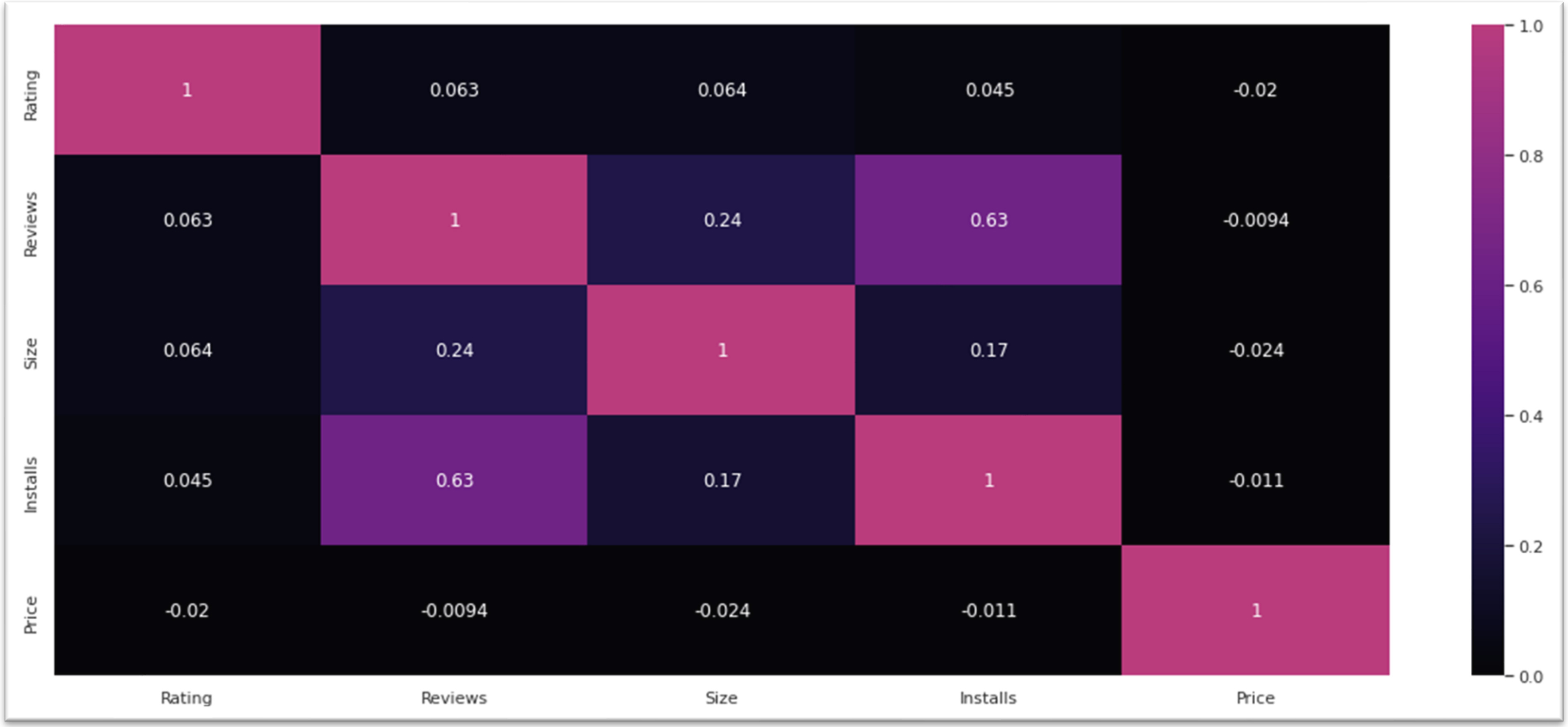


Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative, or undefined.

Bivariate Analysis

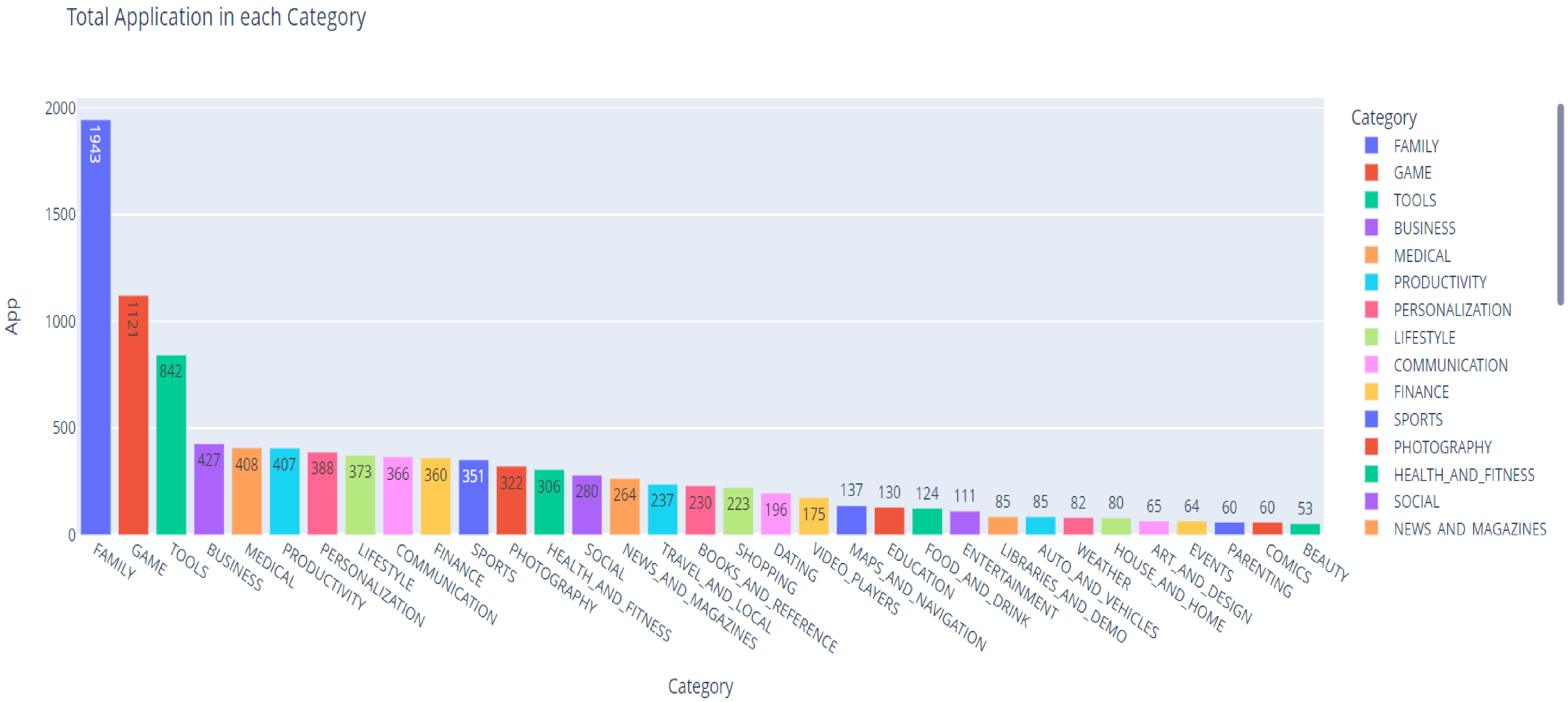
Bivariate analysis is one of the simplest forms of quantitative analysis. It involves the analysis of two variables, for the purpose of determining the empirical relationship between them.

Correlation of Data



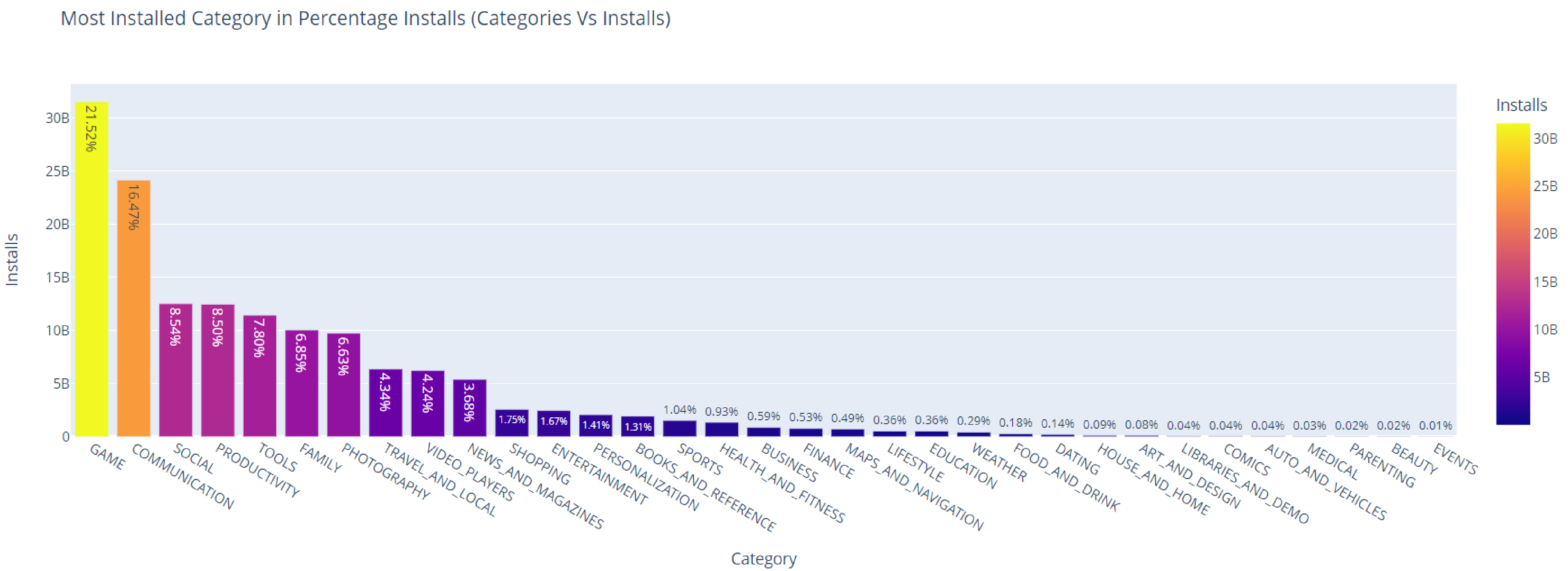
Correlation is used to find the relationship between two variables which is important in real life because we can predict the value of one variable with the help of other variables, who is being correlated with it.

Total Number of Application in each category



From the above graph, we can see the total number of applications in each category.

Categories Vs Installs

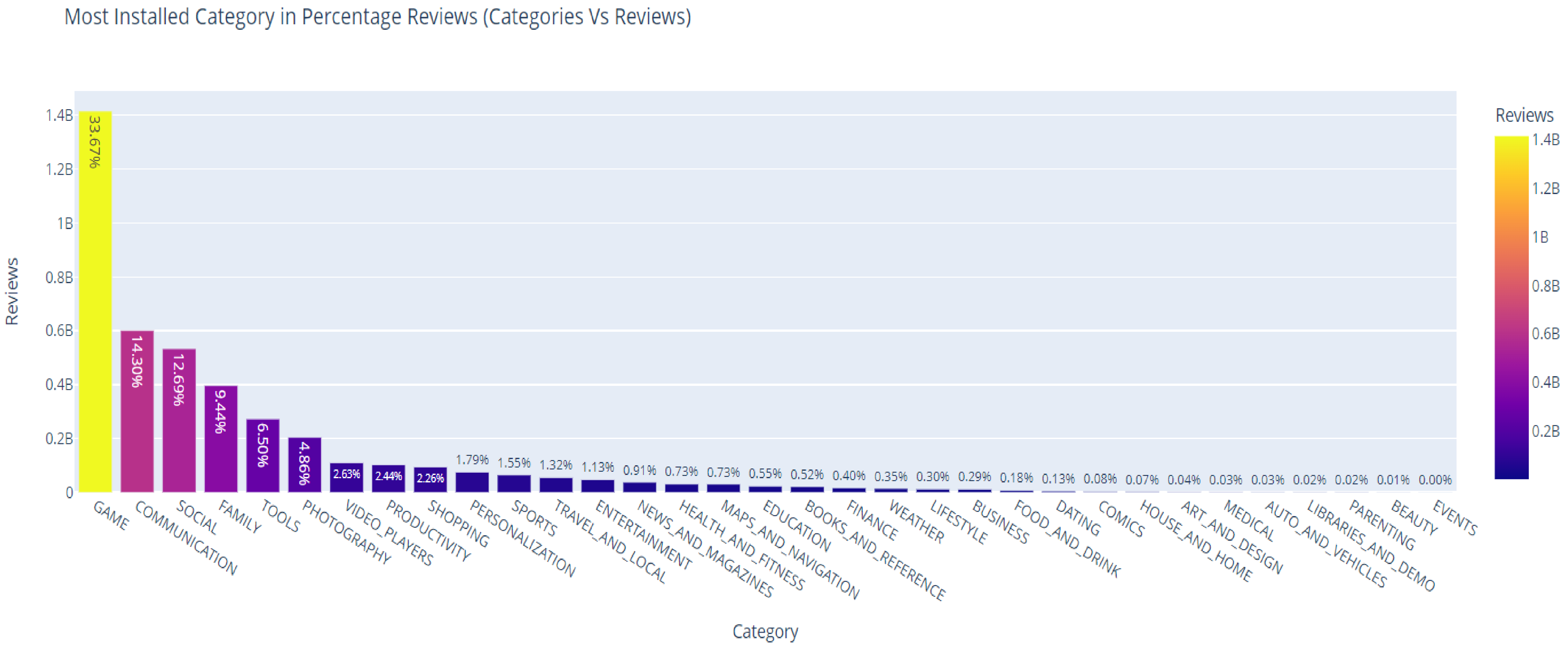


Most Installed Category in Percentage Installs (Categories Vs Installs)

Maximum number of applications are being installed from the category game followed by communication category. So, this graph is the indication of interest hold by the customer, so you must bend your technologies, accordingly.

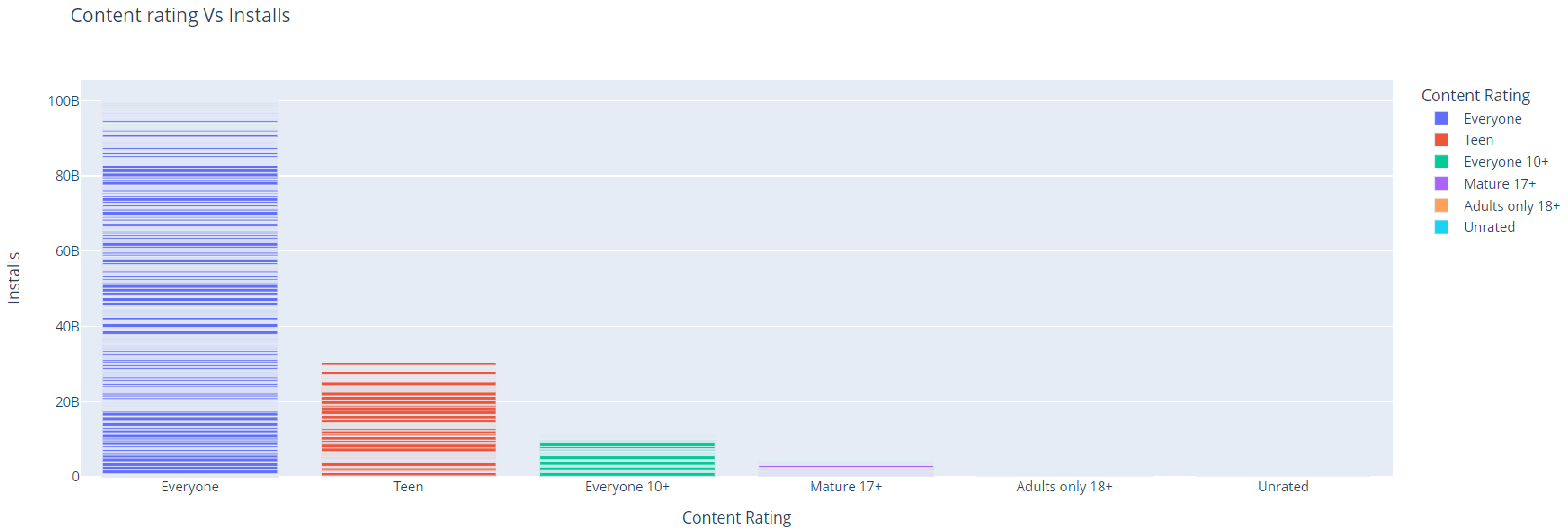
As we have seen correlation map that reviews, and installs are highly correlated with each other so here is the evidence that most positively reviewed category have the highest installation rate.

Categories Vs Reviews



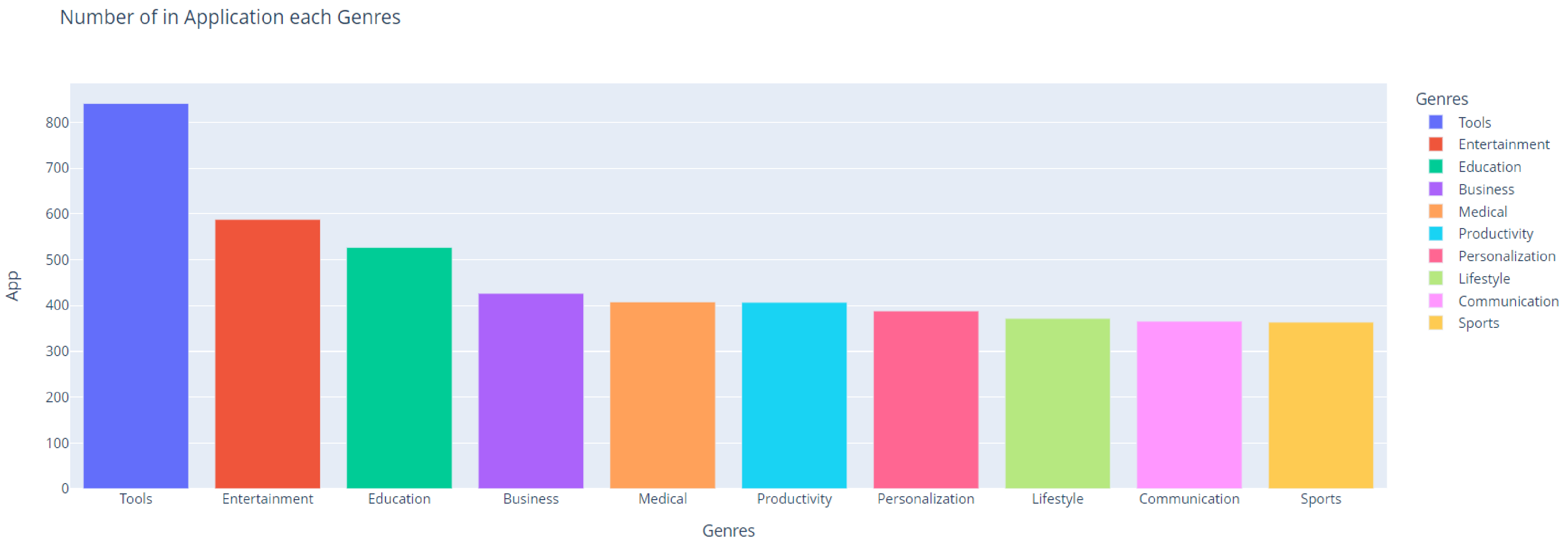
Reviews have the power to gain customer trust, and they encourage people to interact with the company. Customer interaction ultimately leads to improved profits for businesses. Gaming and Communication these two categories has highest percentage of reviews 33.67% and 14.30% respectively.

Content Rating Vs Installs



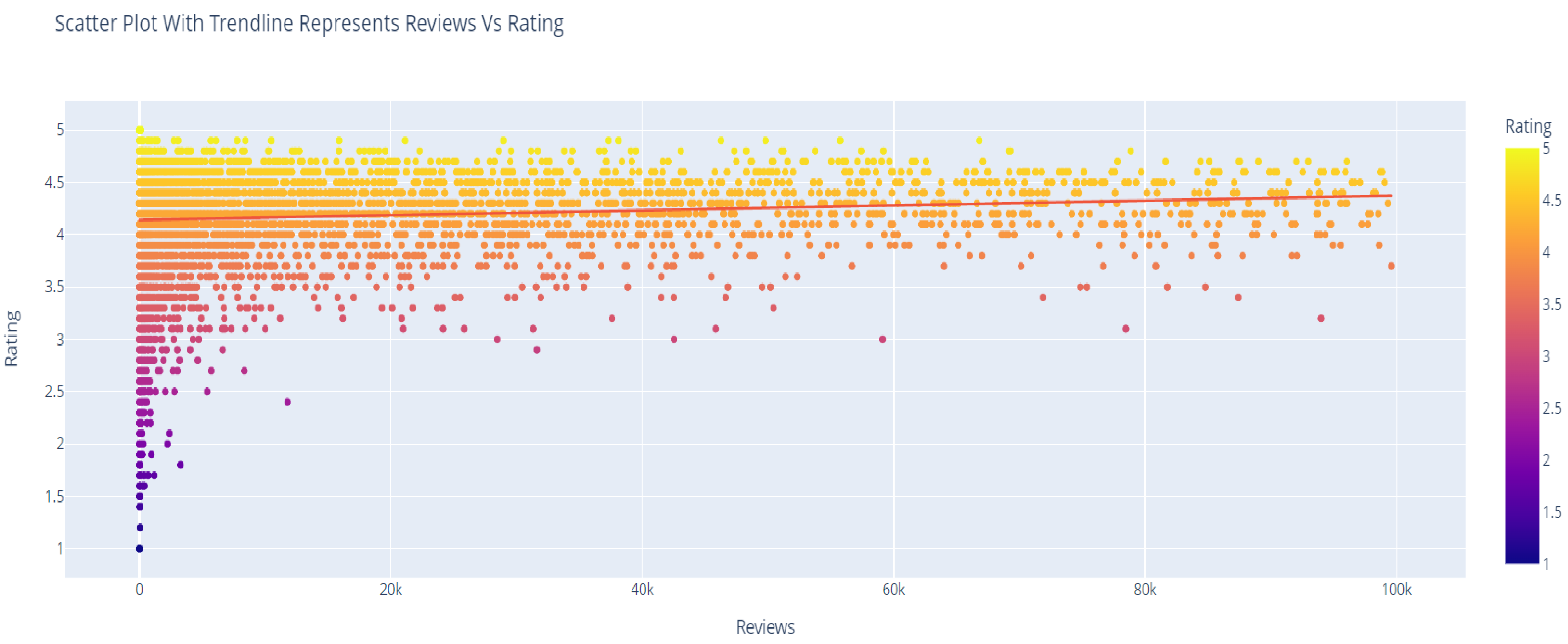
The applications which have a content rating for everyone are being installed most than other. Play Store place a greater level of importance on ratings and reviews than ever before. Apps with higher ratings and reviews rank high in search. If an app ranks high then there's a better chance of it being found and downloaded.

Number of Applications in each Genres



There are 10 different genres – Tools, Entertainment, Education, Business, Medical, Productivity, Personalization, Lifestyle, Communication and Sports. The above graph shows the number of applications in each genres.

Reviews Vs Rating



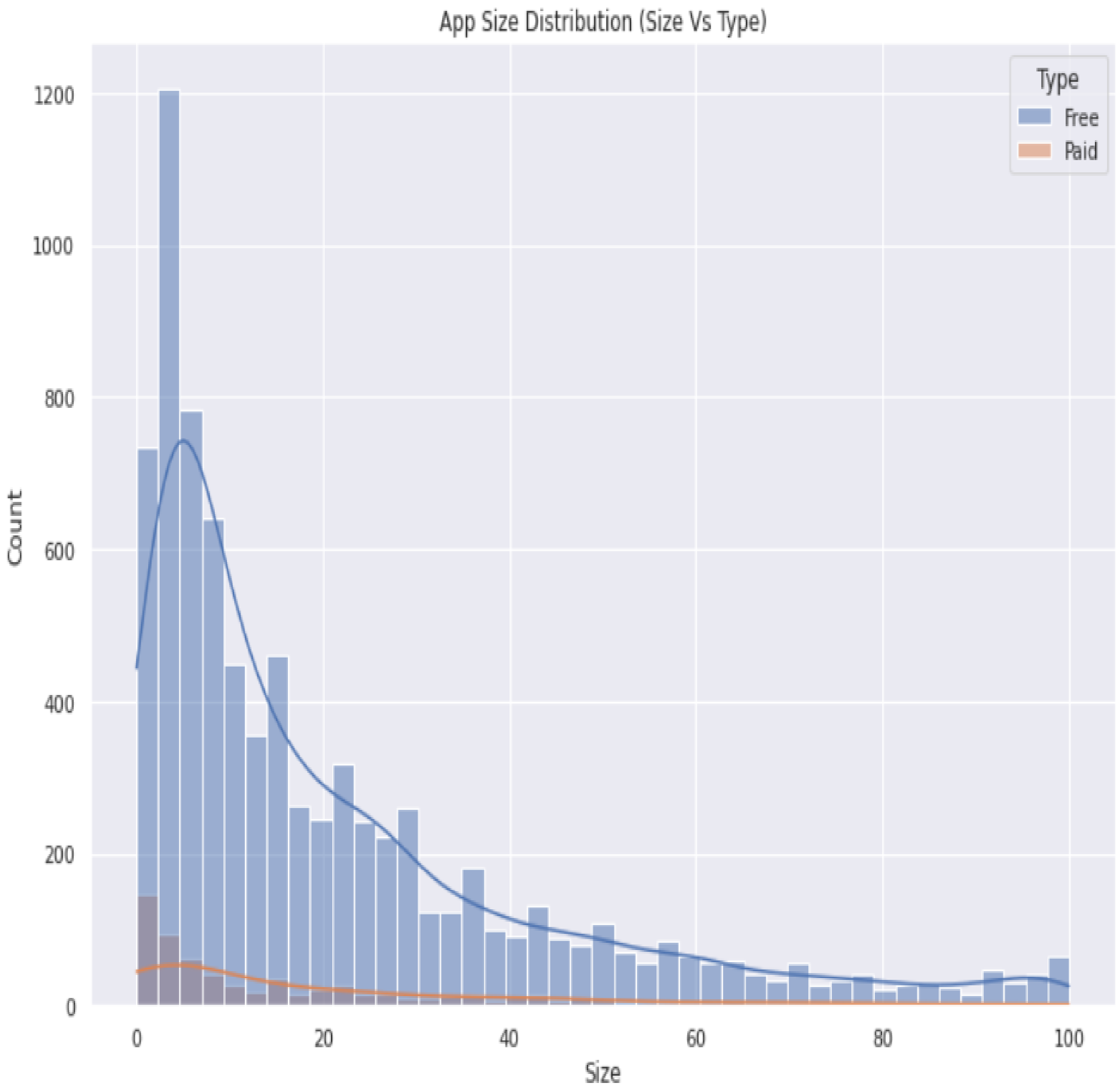
By looking at above scatter plot with trendline we are able toconclude that lesser the reviews on applications lesser the rating as well.

App size distribution

This histogram of apps size distribution tell us about the optimum size range most liked by the user.

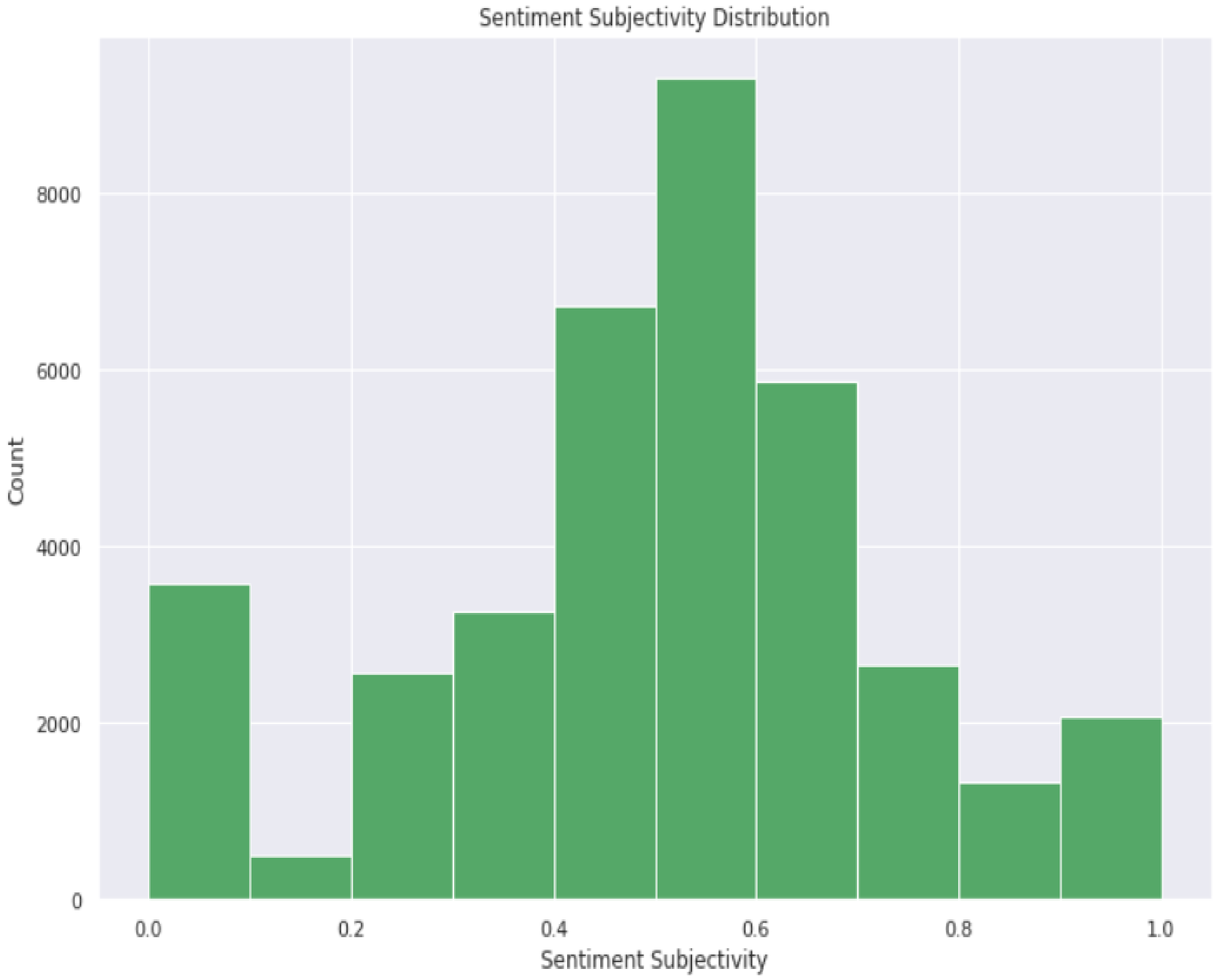
The size of your application has an impact on how fast your app loads, how much memory it uses, and how much power it consumes.

We can draw a conclusion that there are maximum number of applications whose range of size is between 0 to 25 or 30 MB.



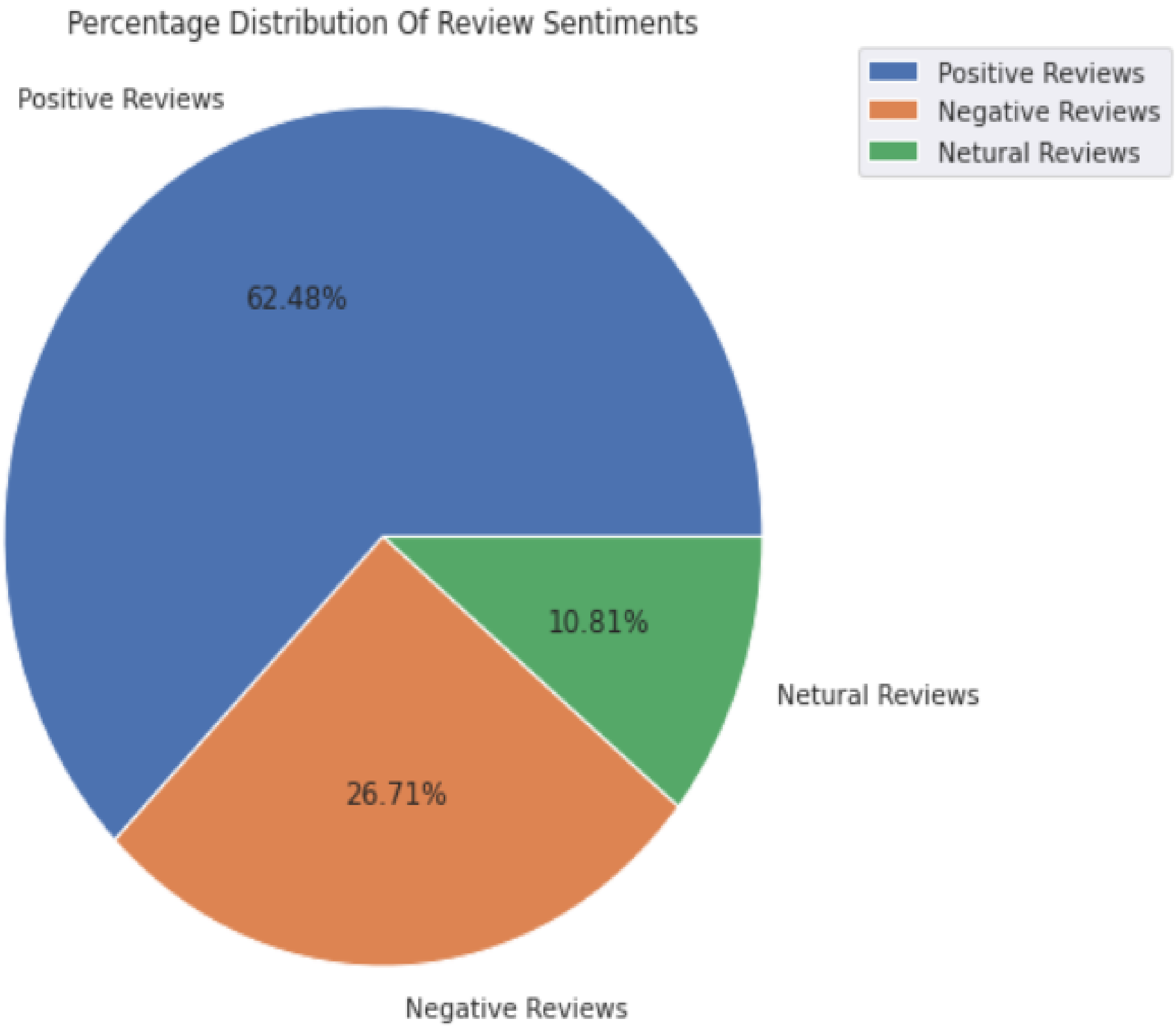
Sentiment Subjectivity Distribution

- Sentiment is the emotion, feeling, opinion, or views held or expressed by users.
- Sentiment subjectivity is float number value whose range lies in between 0 to 1, where 0 is very objective and 1 is very subjective.
- Sentiment subjectivity determines the judgement of review writer's how happy, disappointed, frustrated they are with the service of the application.



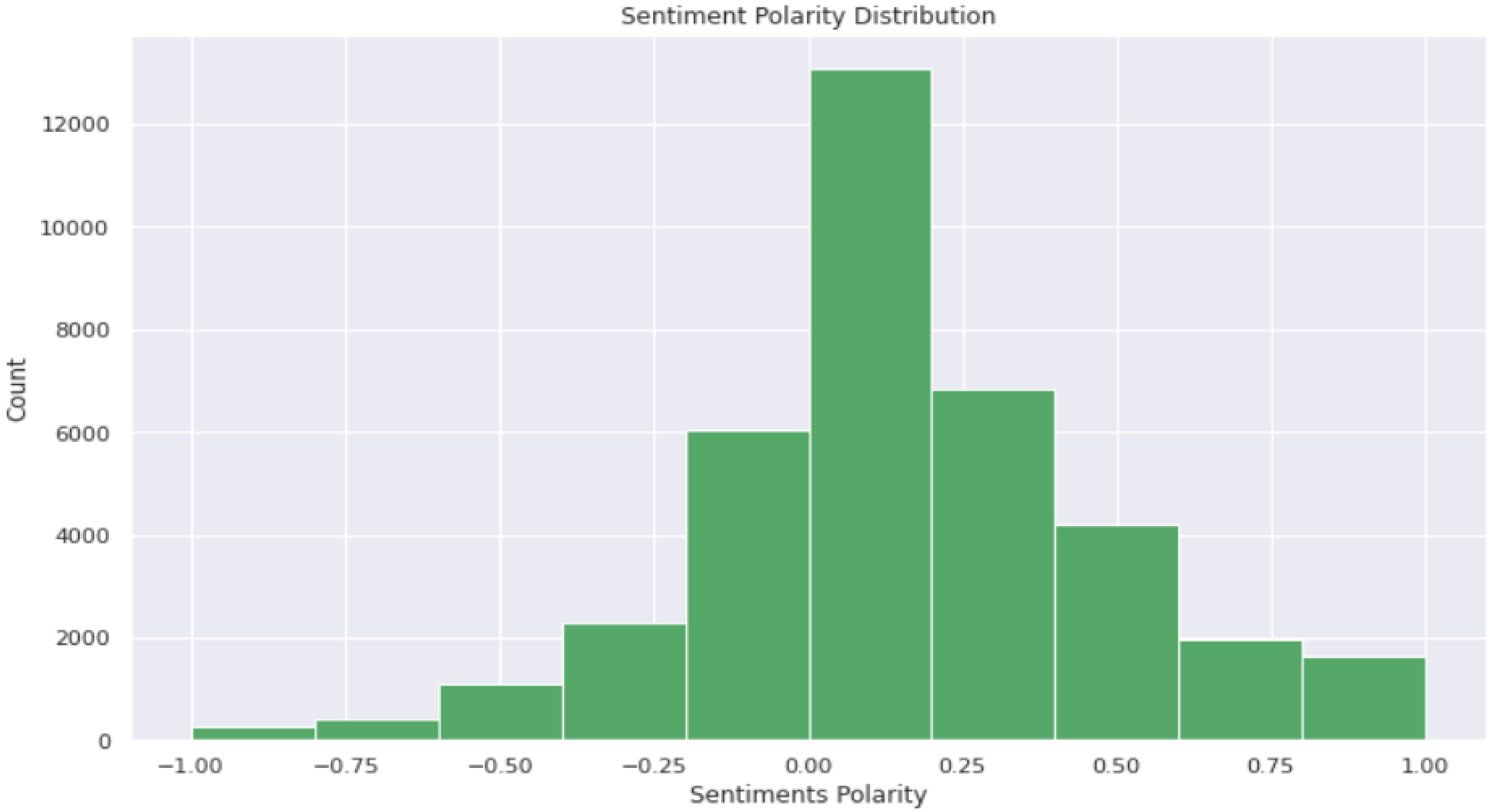
- For given google play store data, sentiment subjectivity range lies between 0.5 to 0.7 that's positive one.

Sentiment Distribution



From the above pie chart, it can easily be understood that there is around 62 of user reviews sentiment is positive, around 27% of reviews sentiment is negative and the remaining around 11% of reviews sentiment is neutral. If some apps have a higher percentage of positive Reviews sentiments, then it is sure that the app is performing its intended work, and people are enjoying it, they may share the app with somebody thus increases the number of installations. So, need to keep an eye-tracking on the review sentiment it is what decides whether the app is going to feature on google play store. By featuring I mean visibility of apps when someone searches for a category. If the app is not visible in the top 10 or 12 apps range then there are fewer chances of the app being installed.

Sentiment Polarity Distribution

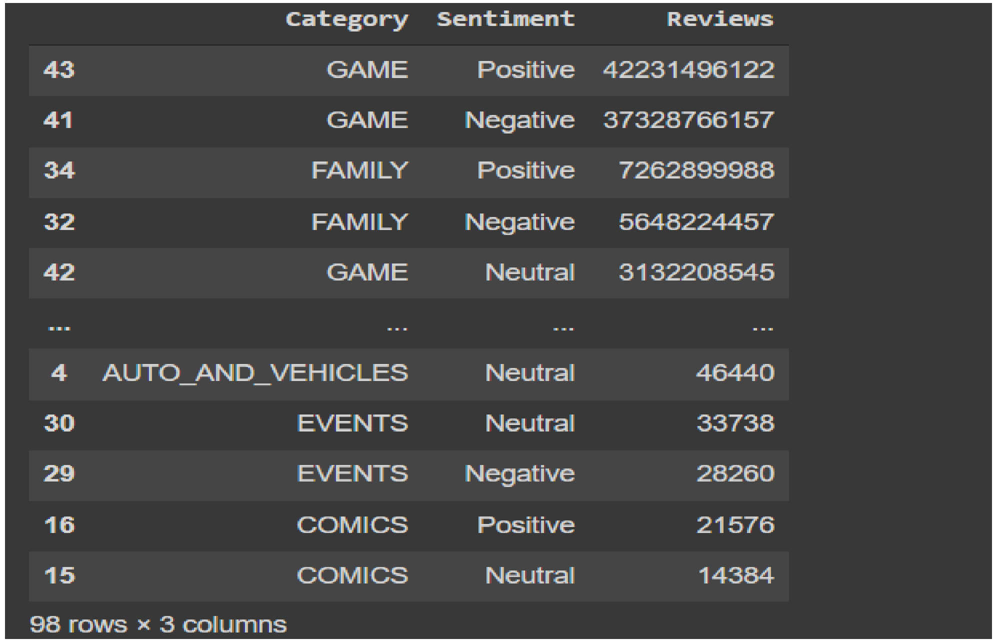


Sentiment polarity is a float value ranging from negative one to positive one. i.e., range (-1, 1), (dtype = float) where -1 means negative statement 1 means positive statement.

Sentiments and Count of sentiment polarity

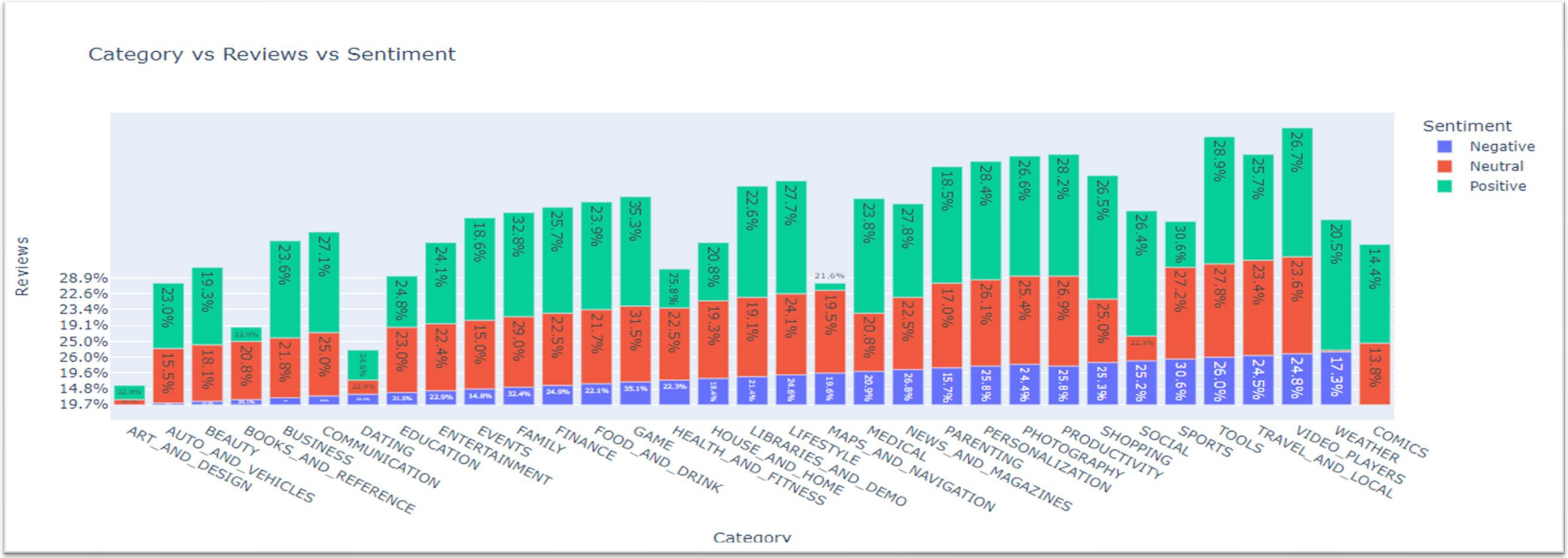
Sentiment	*Sentiment Polarity*
Negative	10108
Neutral	4089
Positive	23642

Reviews and Sentiment of users in each category



```
            Category  Sentiment      Reviews
43              GAME   Positive  42231496122
41              GAME   Negative  37328766157
34            FAMILY   Positive   7262899988
32            FAMILY   Negative   5648224457
42              GAME    Neutral   3132208545
...              ...        ...          ...
 4  AUTO_AND_VEHICLES   Neutral        46440
30            EVENTS    Neutral        33738
29            EVENTS   Negative        28260
16            COMICS   Positive        21576
15            COMICS    Neutral        14384
98 rows × 3 columns
```

Category Vs Review Vs Sentiment



We always treat neutral reviews as insignificant, but it is not the case always, Neutral reviews are never neutral. A neutral review affects both negative and positive reviews and can also reinforce both. Statistically, neutral reviews are a good thing because without them, positive reviews will be overestimated, and negative ones will be underestimated. Neutrals give consumers a better understanding of a brand or a business.

Conclusion

As per our EDA, an ideal application on the google play store should obey the following properties/characteristics:

1. **Reviews vs install**: We have experienced from the seaborn heatmap that reviews on the google play store are highly correlated with the rate of installation. Reviews are given by users as per their experience with the application. So reviews on the application should be examined properly to get to know the performance of the application, whether it is catering to the need of users, From review, we will get an idea on which aspect to work on.
2. Family Category has the most number of applications available on the google play store, and very few apps are available for the category beauty and parenting. We can see the people does not relay on apps for parenting and beauty. And for the comics the category is very saturated.
3. Game category has the highest percentage of installs that is 20.93% around 21% and second highest is communication have 19.48% percentage installs. So people's are more likely to connect with the people by using apps. And for entertainment purpose they are relaying on game as well.
4. Most space consuming category bar plot gives us the idea that which category has the most variety of applications available, and which has low. The category which consumes high space inside the google play store, it means it has more number of applications than other categories. So we have to take a signficant decision to decide category for our future applications.

Thank You!!

🙂