Capstone Project

Google Play Store App Review Analysis

by

Shailendra Dubey

Points For Discussion

- Data summary
- Avg rating distribution per categories
- Heatmap correlation of features
- Most reviewed category
- Most space required category
- Most installed category
- Category type effect
- Size distribution
- Size effect
- Size vs installs vs type
- Sentiment subjectivity distribution
- Sentiment polarity distribution
- Percentage reviews sentiment distribution
- Conclusion

Data Summary

play_store_df : This data frame is having a shape (10841,13). It holds the 13 features which include apps, category, rating, reviews, size, installs, type, price, content rating, genres, last update, current update, android version. This 13 features and 10841-row labels contains all sort of information which can be analyzed, interpreted, and implemented in order to take better business discission.

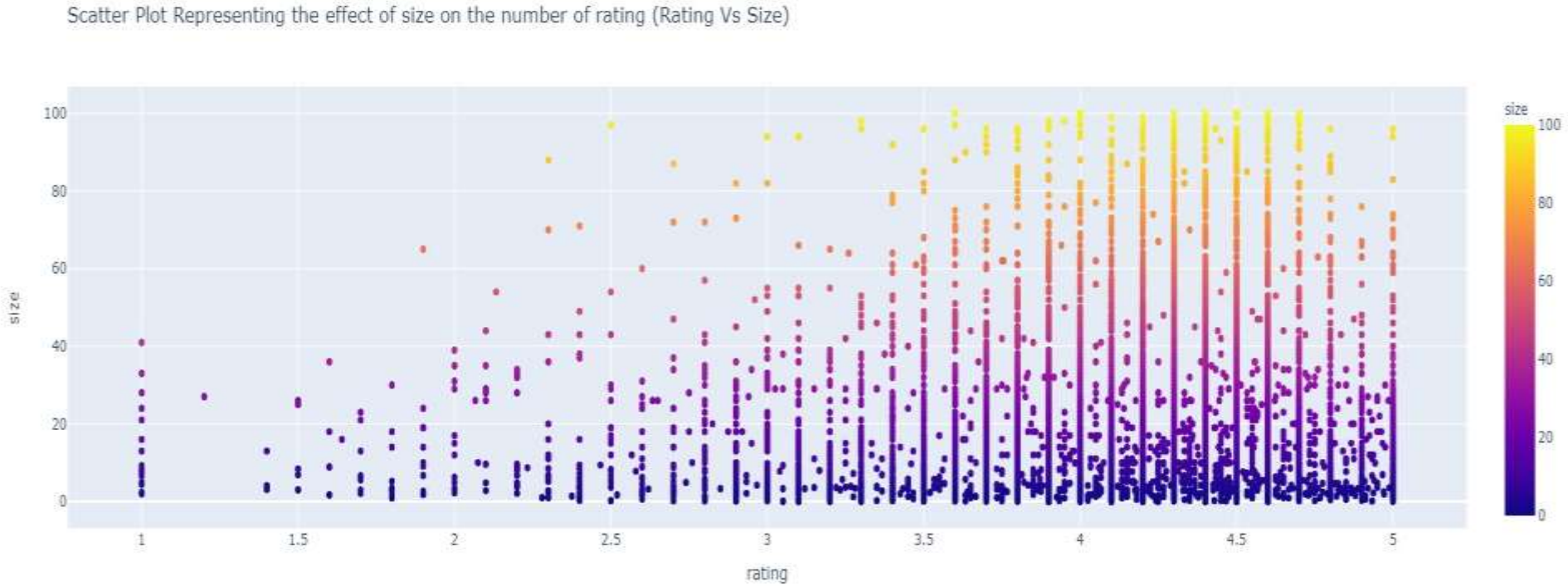
user_review_df : This data frame is having a shape (64295, 5). It consist of 5 features which involves app, translated view, sentiment, sentiment subjectivity, sentiment polarity.

merged_df1 : As the name suggest, this data frame is the result of merging above two data frame based on feature app. As app is the only feature that is common between these two data frame. merged_df encapsulate all the feature from play_store_df and user_review_df

Average Rating Distribution per Categories

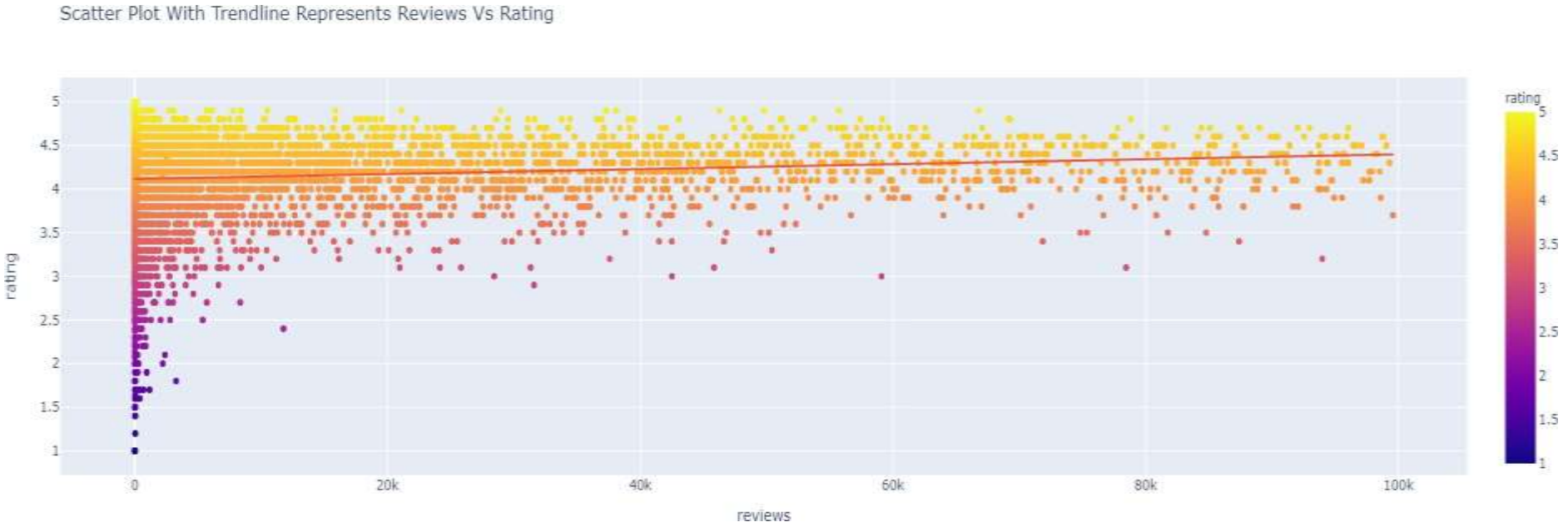
Your app's rating will affect its chances of being featured. Apps with 3 stars or lower will not be featured. The app rating is an important aspect of ASO (app store optimization). Negative mobile app reviews combined with a poor rating will hurt your app's rank, but great app reviews and high ratings will help increase your app's rank.

Rating vs Size



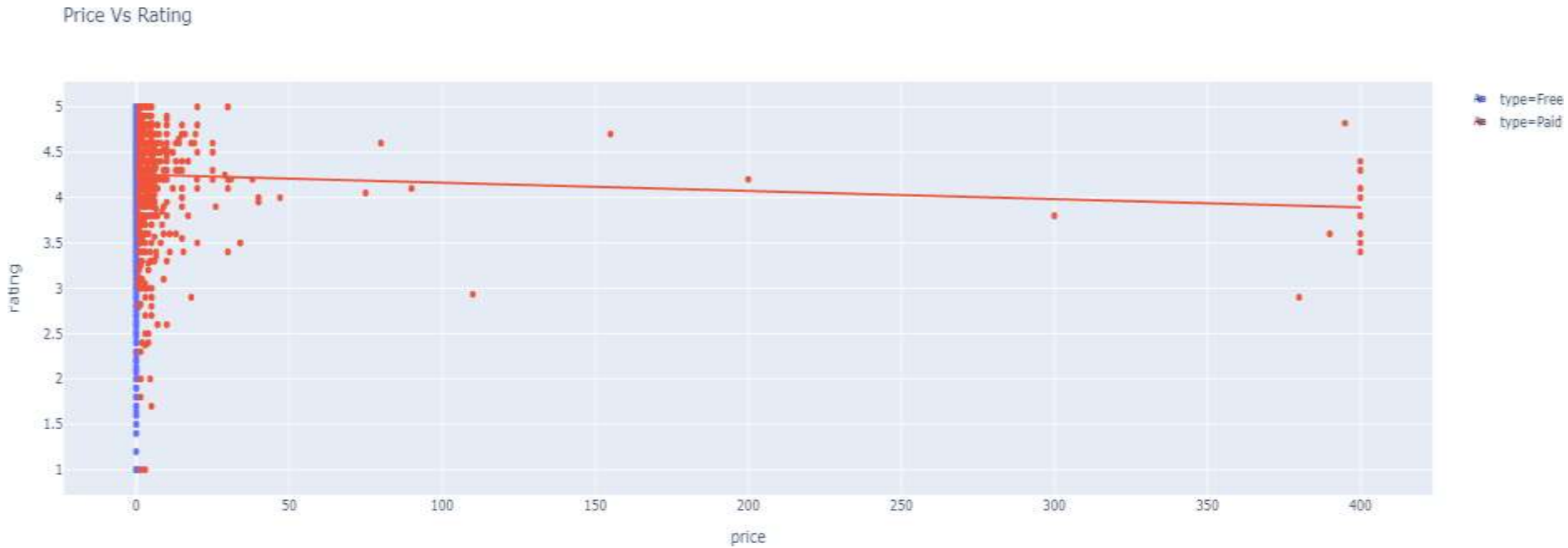
The above Scatter plot is evidence that there is more rating available on low-sized applications than that heavy-sized applications.

Reviews Vs Rating



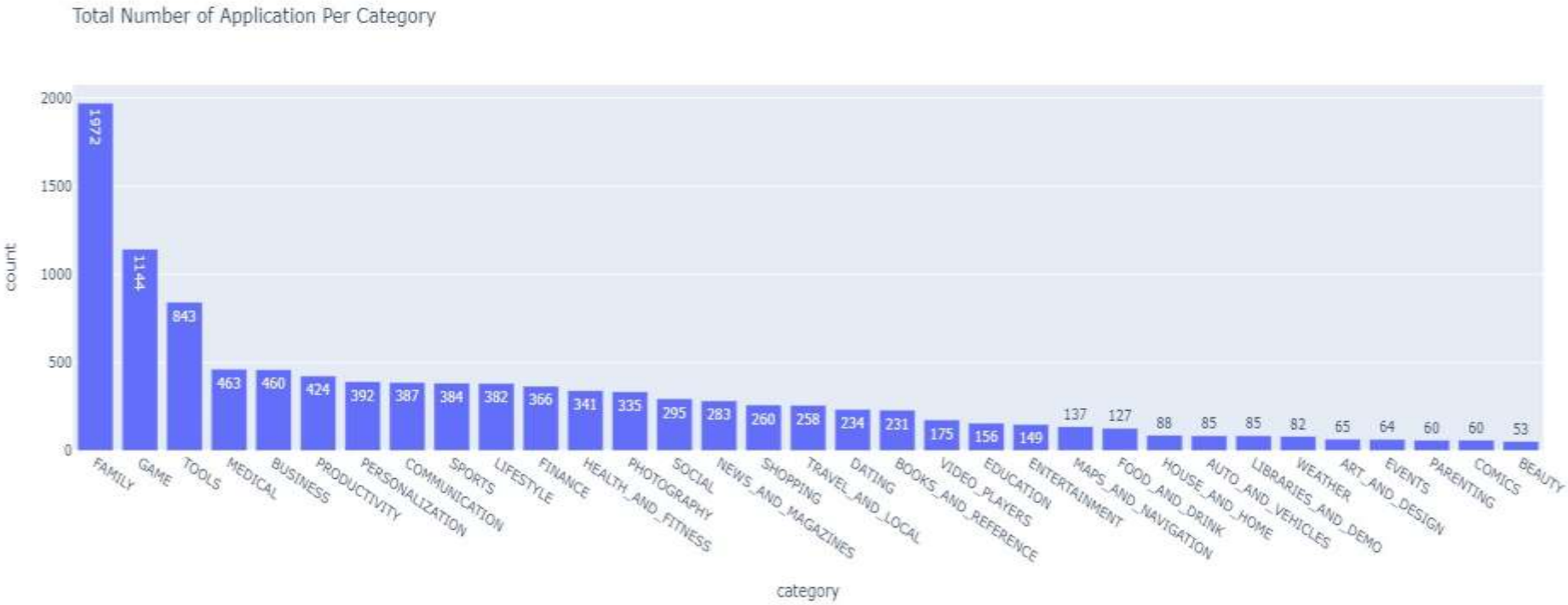
Obviously by looking at above scatter plot with trendline we are able to conclude that lesser the reviews on applications lesser the rating as well.

Price Vs Rating



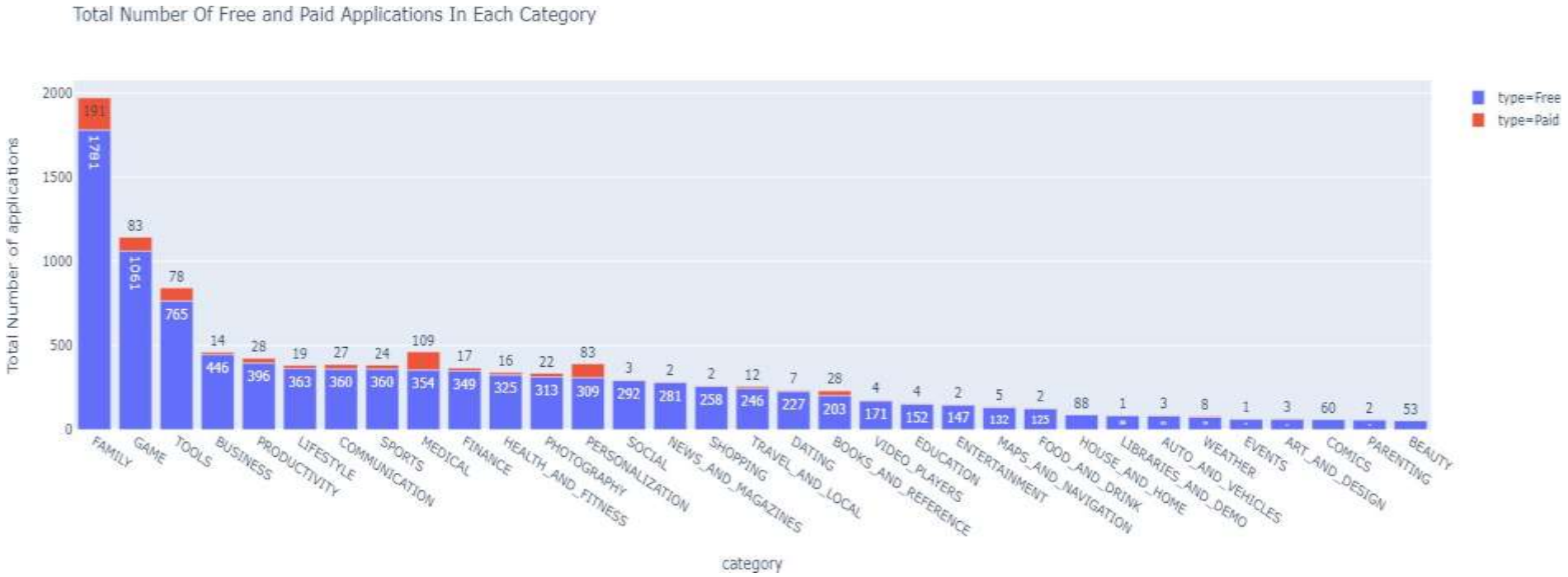
Inference - - Of course, as the price of the application increases, there are fewer downloads hence fewer reviews and ratings as we can visualize from the above scatter plot.

Number of Application Per Categories



Above bar plot represents total number of applications available per category. Family category is at first place with highest number of applications that is 1972, and game category settle down to second with 1144 applications, tool category is at third with 843 applications.

Total free and paid application in each category



From the above bar plot, we can see the total number of free and paid applications in each category. Number in red color denotes the paid applications whereas the number in sky-blue color denotes the free application in categories.

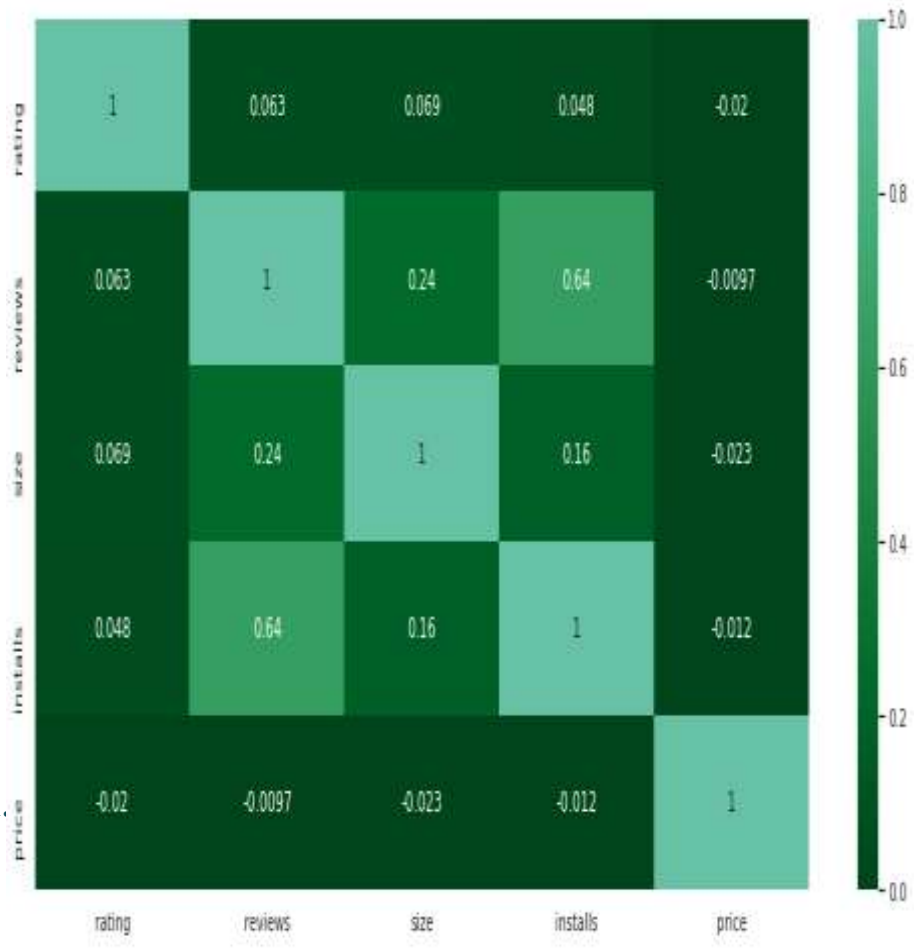
Heatmap Correlation Between Features

This is a heatmap chart that represents the relation between one feature and another one! It clearly denotes that customer reviews are highly correlated with the installation rate.

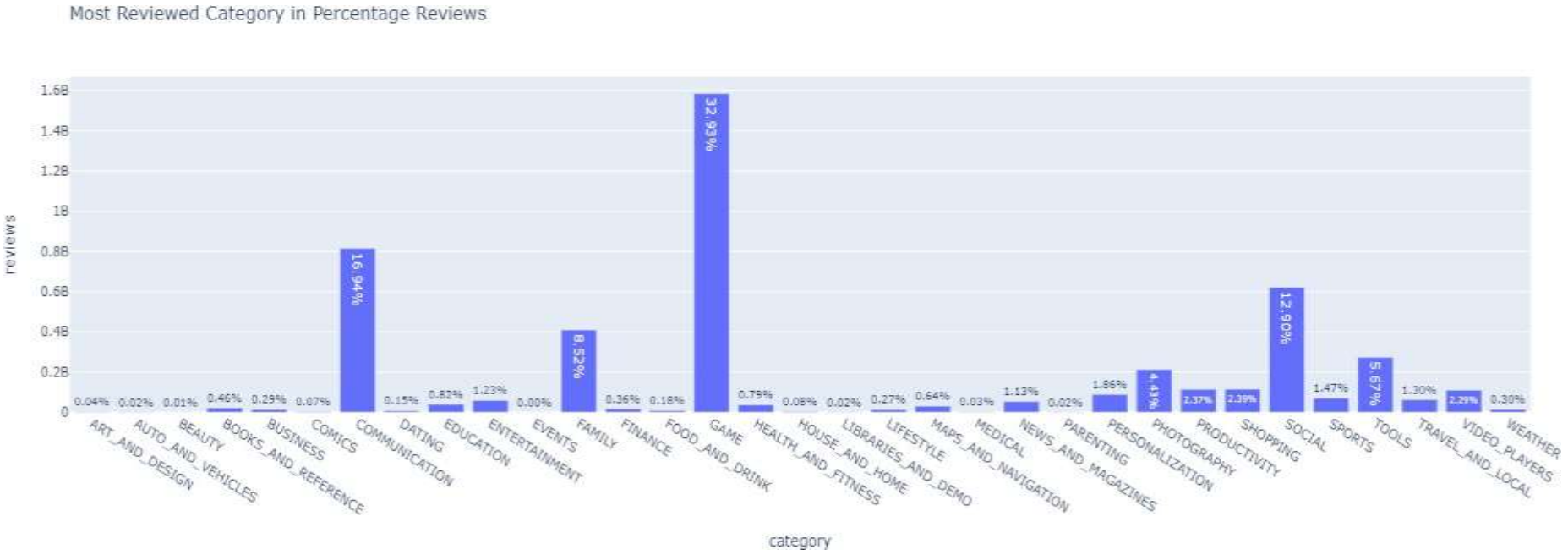
It means there is 64% chance of installing an app by the user if reviews are good. The average installation rate is positively correlated with user reviews, rating.

We can see the correlation between size and installation. But almost every customer tends to decline for installing the application that demands high space inside a device.

Price is in negative correlation with installs. It defines higher the price rate , lower will be the rate of installation.

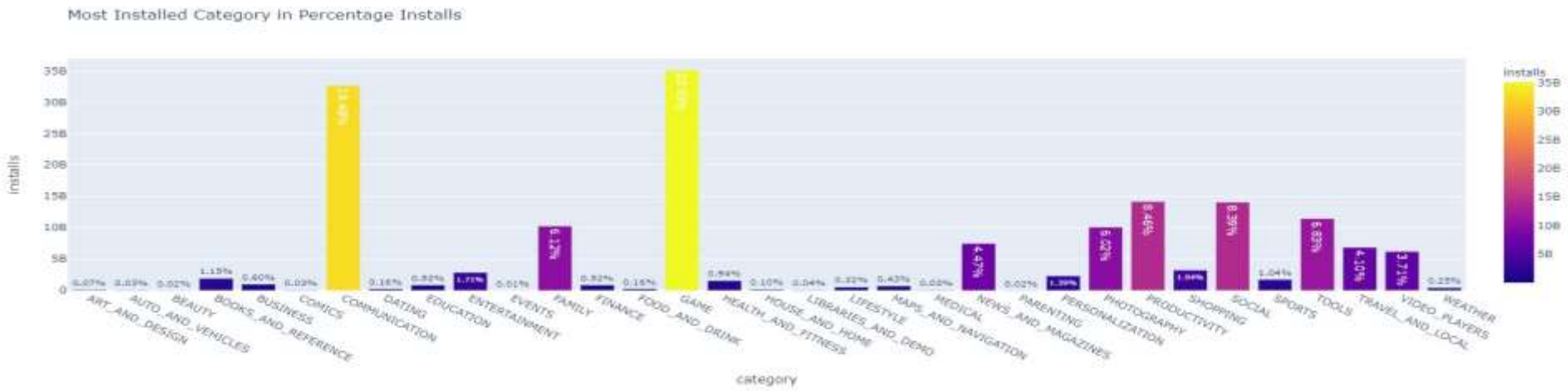


Most reviewed category



Gaming and Communication these two categories has highest percentage of reviews 32.93% and 16.94% respectively. There are around 1.6 billion reviews available on game category, and around 0.8 billion reviews on communication, least reviews can be observed in categories beauty, parenting, auto and vehicles, art and design.

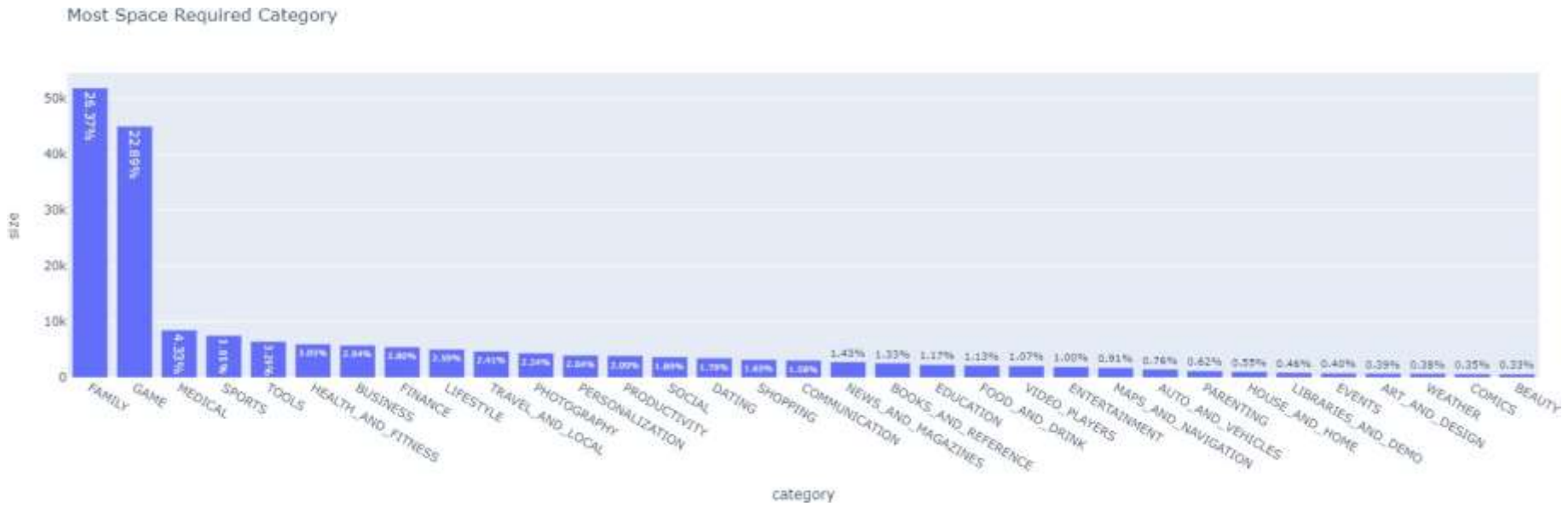
Most Installed Category



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

As we have seen heatmap correlation map that reviews, and installs are highly correlated with each so here is the evidence that most positively reviewed category have the highest installation rate. Gaming category and communication category have highest installation. Application from the category beauty are being installed least.

Most Space Required Category

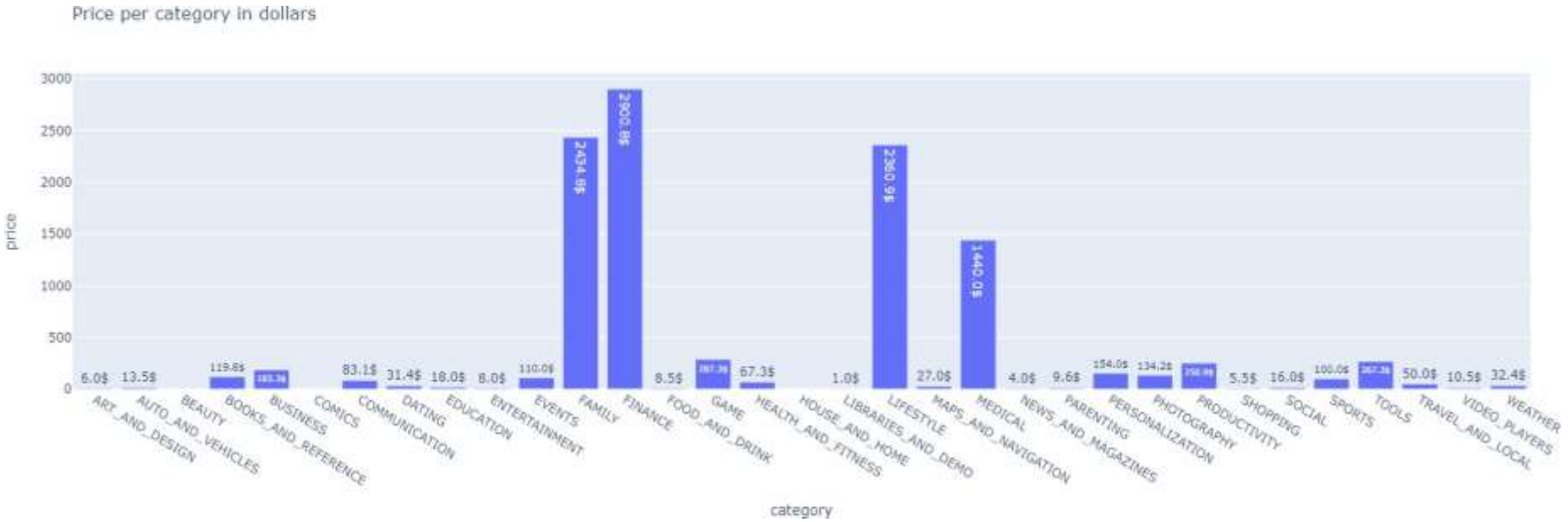


This bar plot represents the most space-consuming category on google play store. As we can see family and game are two categories that consume 26.37% and 22.89% space in the google play store, respectively.
It means you will get more varieties of applications in these two categories; it also means there are many applications available in these two categories when compared with other categories.

Category type effect

This bar plot indicates that the installation for the unpaid or free category application are much higher than that of paid one. The number of installs are converted into np.log2(installs) for the purpose of appropriate visibility of free and paid type installation.

Price Per Category



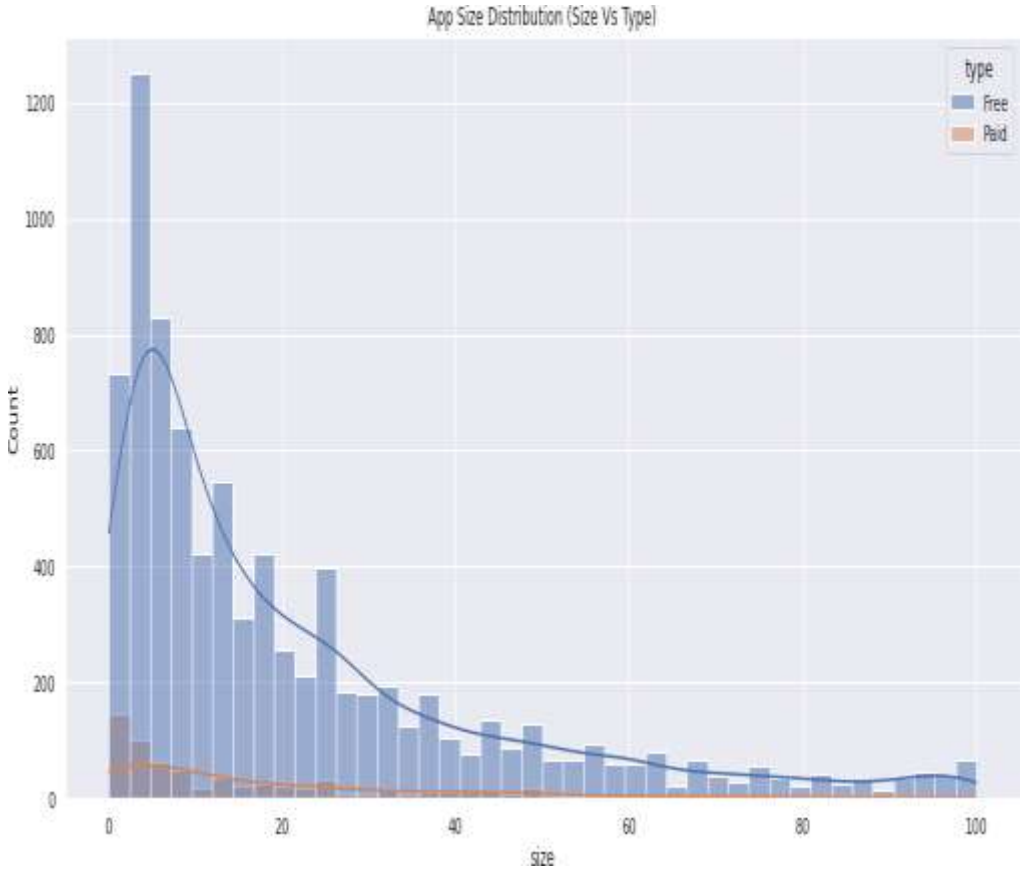
One can clearly understand that applications from the category finance have high price than any others, family category applications are second high priced applications, at third we have lifestyle category applications and at last, we have medical category applications, these four categories application charges high price when compared with others.

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.

By observing the above histogram with KDE (Kernel density estimation) KDE line for paid type applications represented in orange color whereas sky-blue colored KDE line for the free type applications we can draw a conclusion that there are maximum number of applications whose range of size is between 0 to 25 or 30 Mb.



Size Distribution For Free and Paid Applications

The above box plot shows that the median for the free type application is 13 MB whereas for the paid type application median is 9.5 MB. Most of the free type of applications has the size in the range of 0 to 30 MB, Whereas for the most paid applications is in the range of 0 to 28 MB.

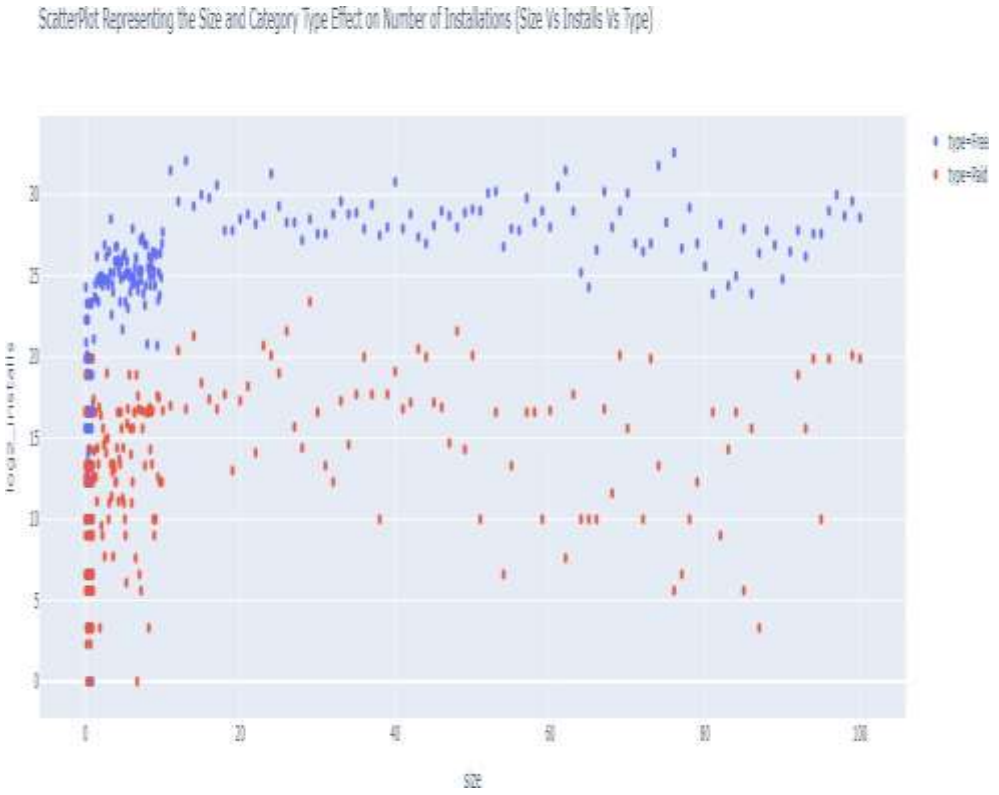
Combined effect of type, size on number of installs

We can conclude from the scatterplot representing the combined effect of type, and size on the number of installs. People more likely to install a free type app that requires less memory to function.

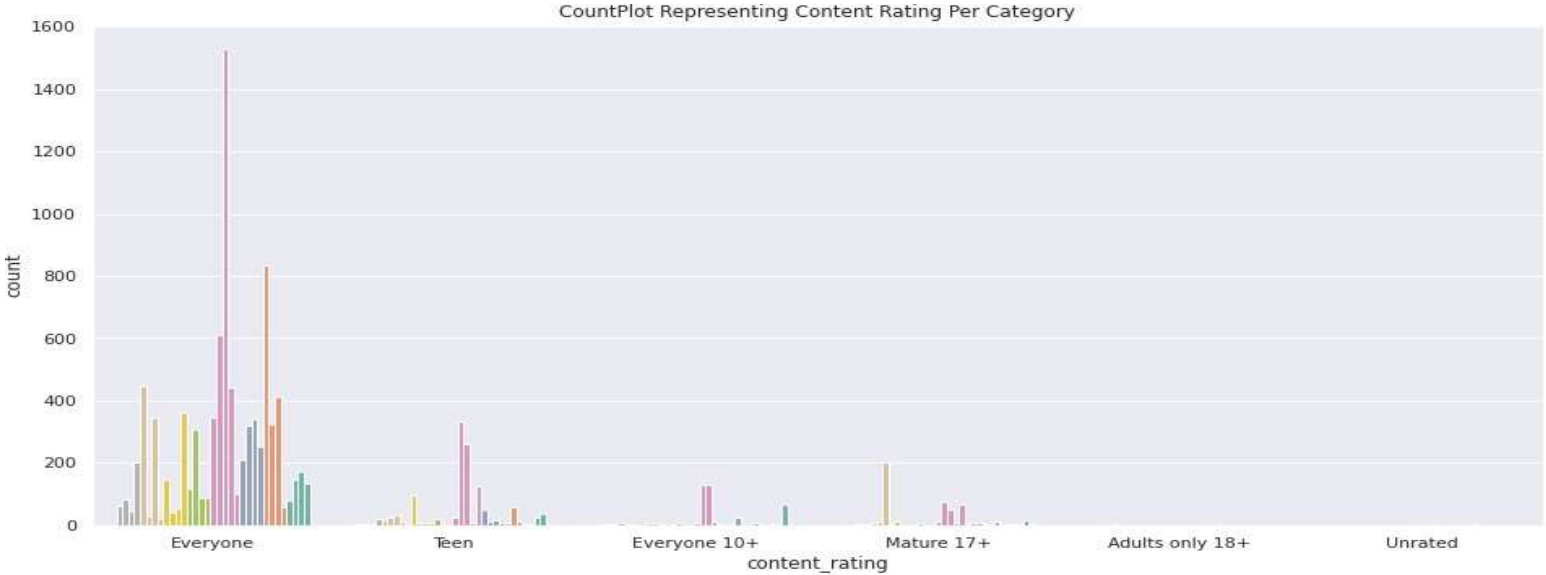
The scale of installs is converted into np.log2 scale, it is implemented to have explicit visualization.

Legends available on the upper left side of the scatterplot denotes the type of category the application coming from, it may be free or paid. The size is in the
MB unit.

So, we can easily draw a conclusion that the greatest number of installations are taking place within a smaller range of size.

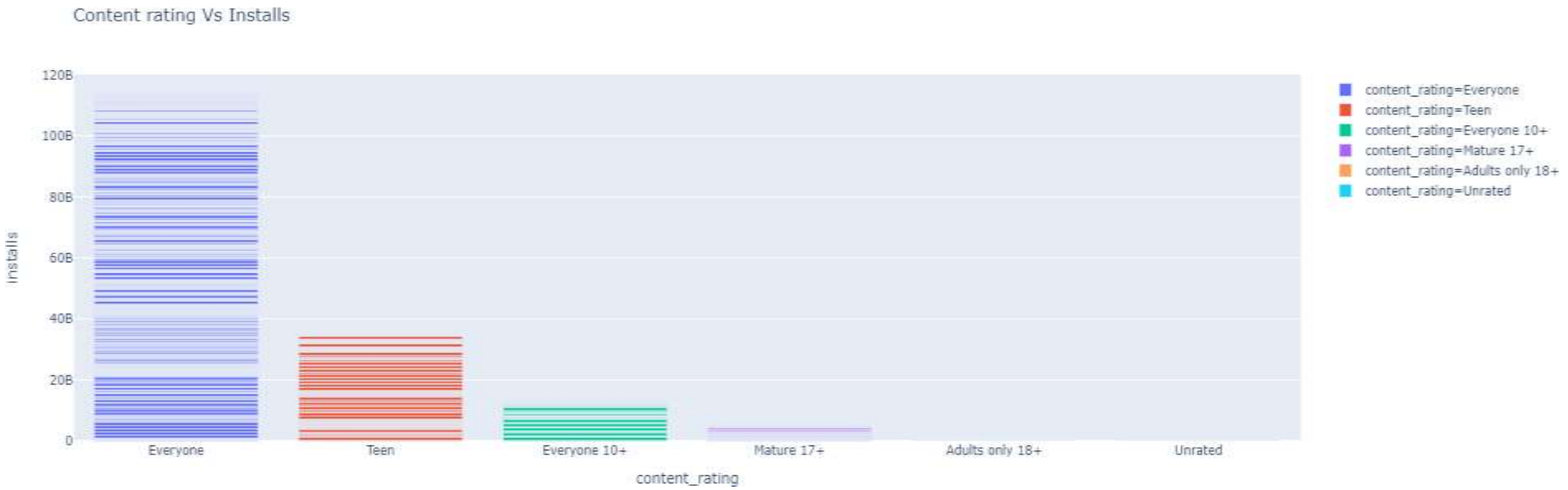


Content Rating For Each Category



Content ratings are used to describe the minimum maturity level of content in apps. However, content ratings don't tell you whether an app is designed for users of a specific age. From above count plot we can conclude that Most of categories has content for everyone, excluding dating category which is only for mature 17+. Some application from other categories like lifestyle, medical, social are for mature 17+.

Content Rating Vs Installs



The applications which have a content rating for everyone are being installed most than other.

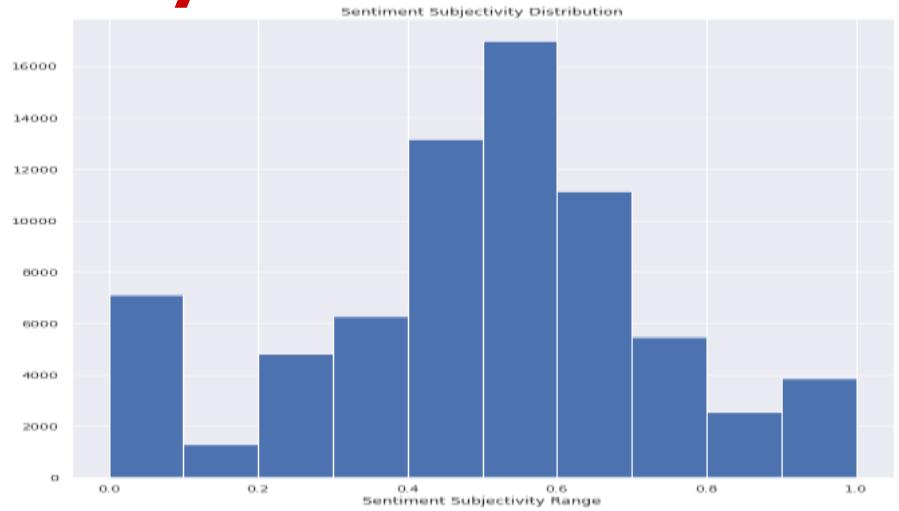
Sentiment Subjectivity and Polarity distribution

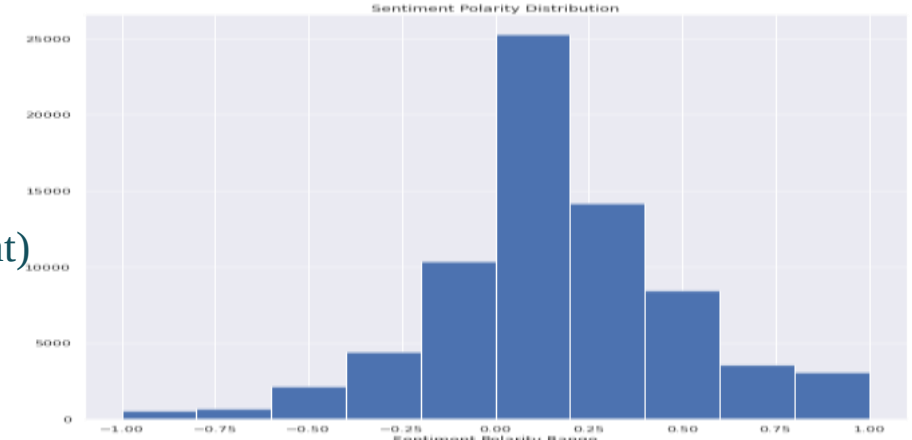
Sentiment is the emotion, feeling, opinion, or views held or expressed by users, sentiment subjectivity is float number value whose range lies in between 0 to 1. where one is very objective and 1 is very subjective.

Sentiment subjectivity determines the judgement of review writer's how happy, disappointed, frustrated they are with the service of the application.

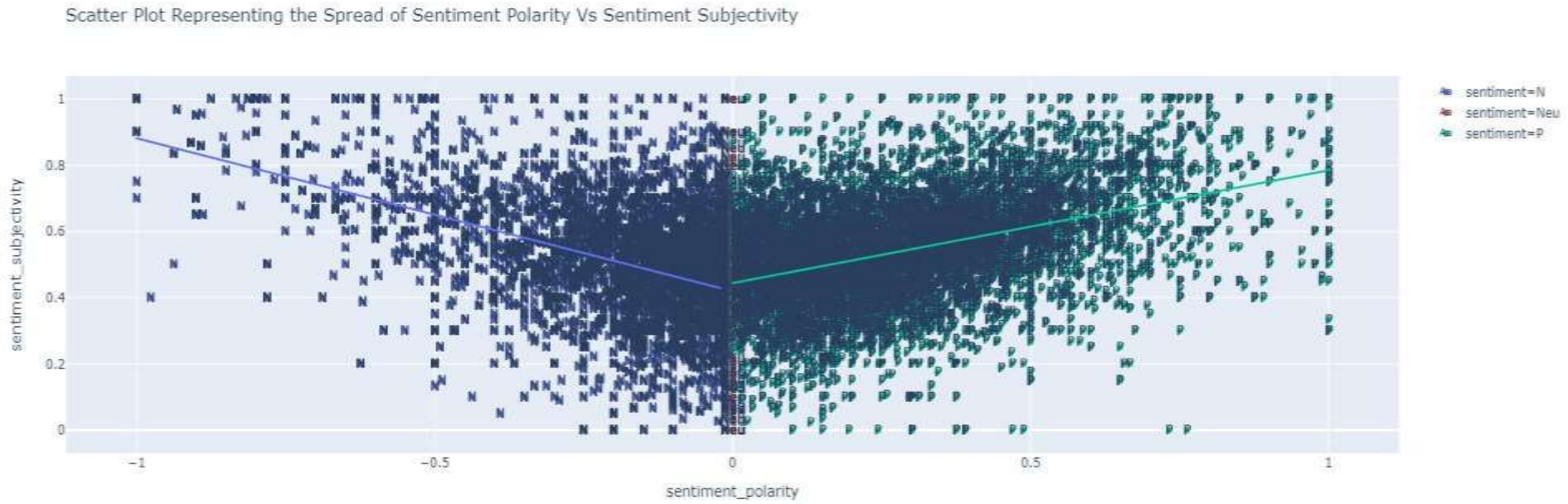
For given google play store data, sentiment subjectivity range lies between 0.5 to 0.7 that's positive one.

Sentiment 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





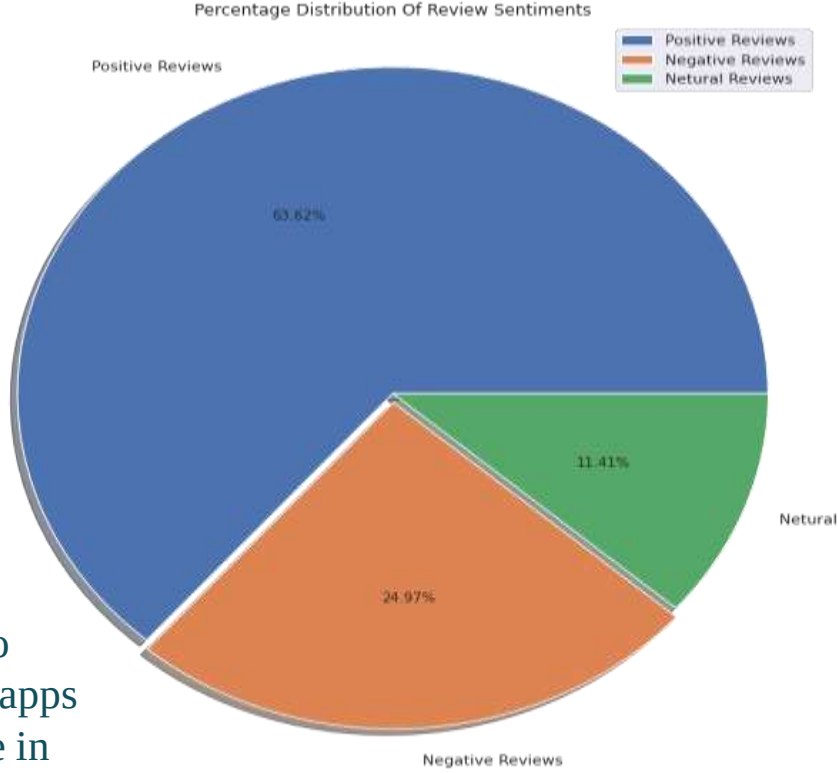
Sentiment Subjectivity Spread Analysis



The greenish portion of the scatterplot on the right-hand side represents the positive reviews spread you can see an alias p for positive, while the blueish portion on the left-hand side indicates the negative reviews and alias as N for negative spread finally the thin portion separating these two spreads is called neutral reviews and can be seen in red color in the above plot. Region above the trendline shows subjective statements and region below that is for objective statements.

Percentage reviews sentiment distribution

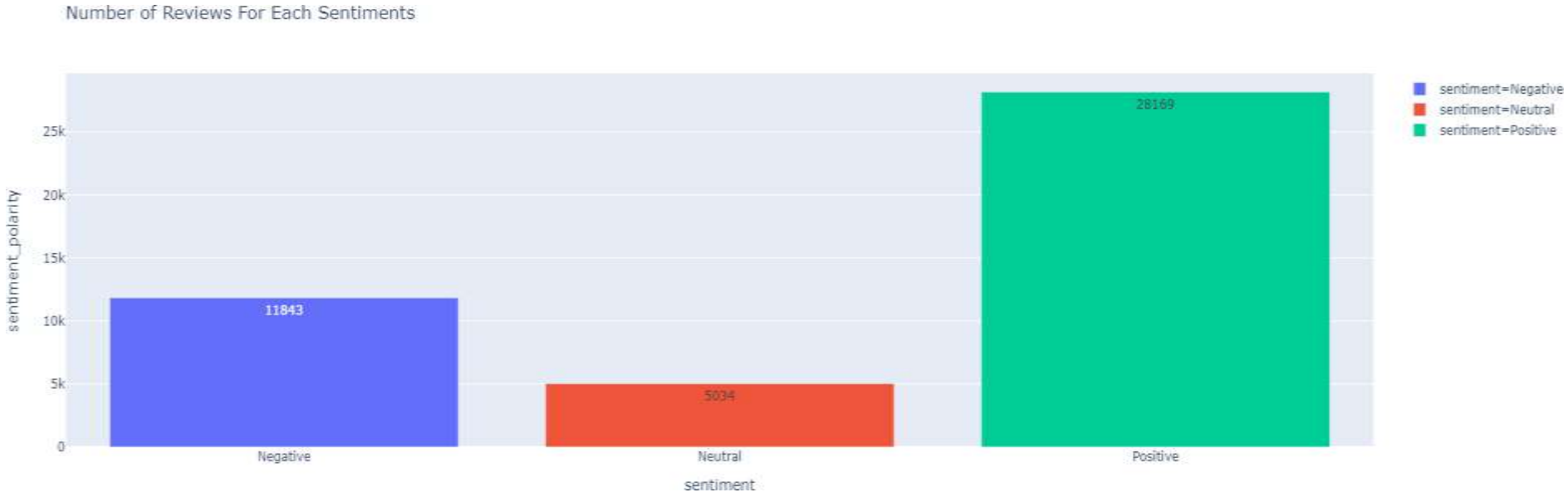
From the above pie chart, it can easily be understood that there is around 63 of user reviews sentiment is positive, around 25% of reviews sentiment is negative and the remaining around 11% of reviews sentiment is neutral.



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.

Count of Reviews Sentiment



The total number of positive reviews is 28169 which is way higher than 11843 negative reviews, and there are 5034 neutral reviews statements.
We have determined the sentiment of the review in terms of ratio which turns around 63:26:11 For positive reviews statement, negative reviews statement, and neutral reviews statements respectively

Conclusion

As per our EDA, an ideal application on the google play store should have the following properties.
1.Category Type: Almost every customer on the google play store expects that application should belong to the category free.

2.Size vs install vs type vs rating: As we have observed in the size vs installation vs type scatterplot, the ideal size of the application should be below 40 MB and max up to 50 MB. we have seen that peoples are less interested to install and use heavy-size applications even though the application is free of cost. There are more ratings on low-sized applications than that heavy-sized applications.

3.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.

4.The most installed category: As we have explored applications belongs to the category gaming and followed by communication are being installed the most, secondly, applications from the productivity category followed by the social media category is being installed the most, It gives us tips to choose domain as per the customer affection inclination.

Thank You!