Capstone Project

Play Store App Reviews Analysis

by

Umesh Rathod

Pranav Dhabale

Rohit Rajput

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
- Inference

Data Summary

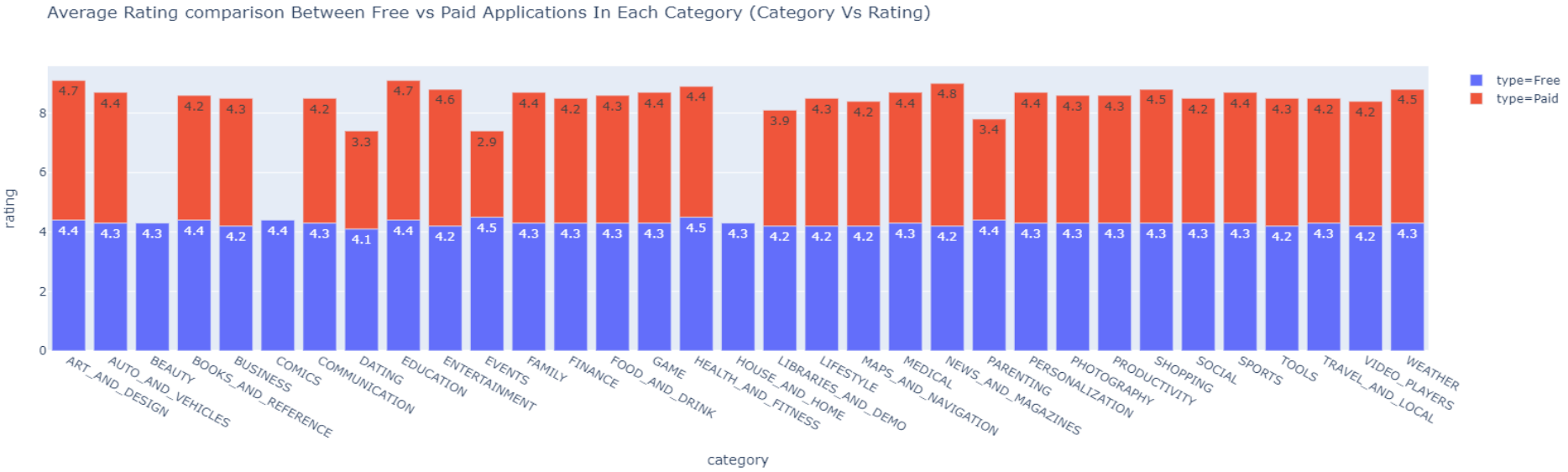
Problem statement: The Google Play Store is the largest app market in the world. It generates more than double the downloads of the Apple App Store but makes only half the money as the Apple Store. Explore and analyze the data to discover the key factor responsible for app engagement and success.

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 version, 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

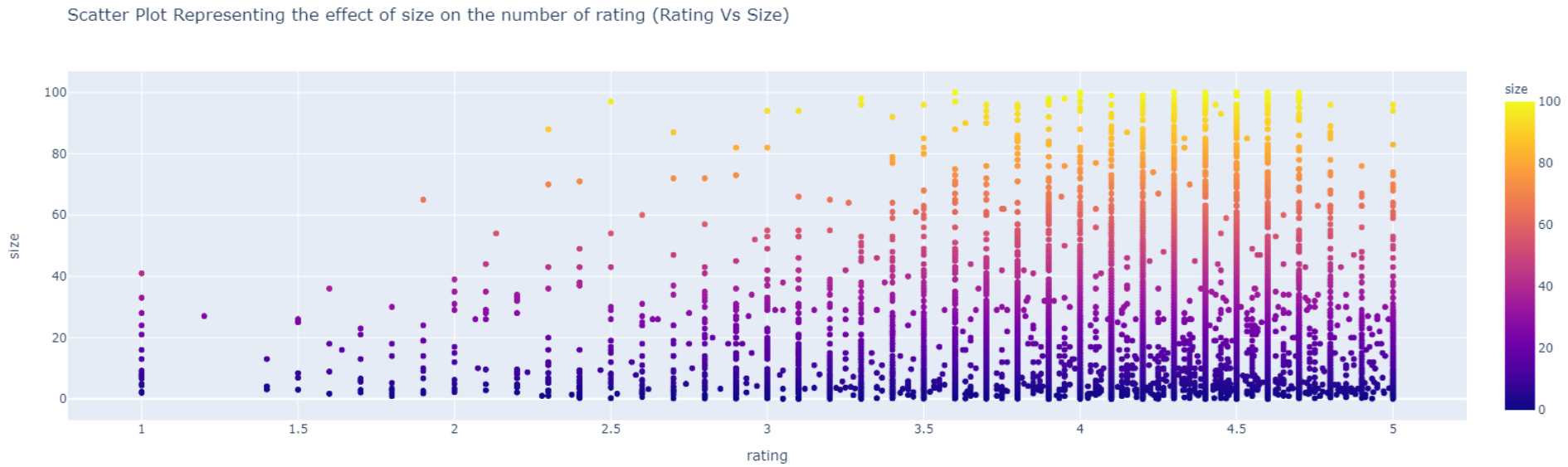
Average Rating Distribution per Categories



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.

According to a survey by Apptentive, 59 % of people usually or always check ratings before downloading a new app.

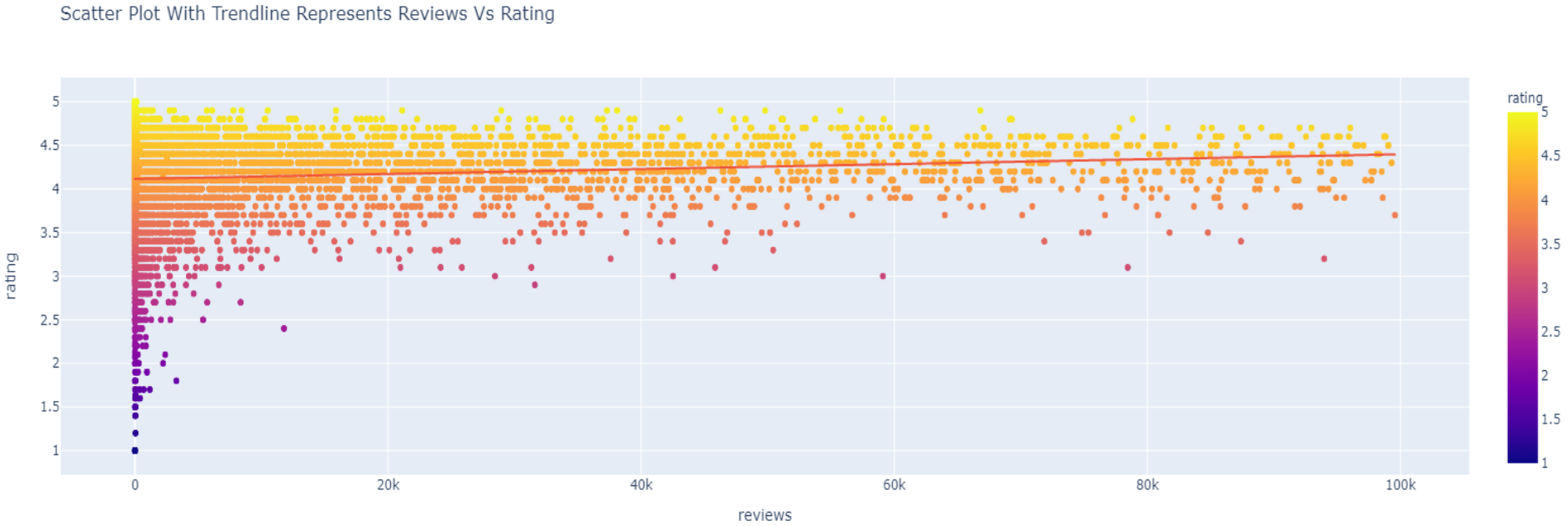
Rating vs Size



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

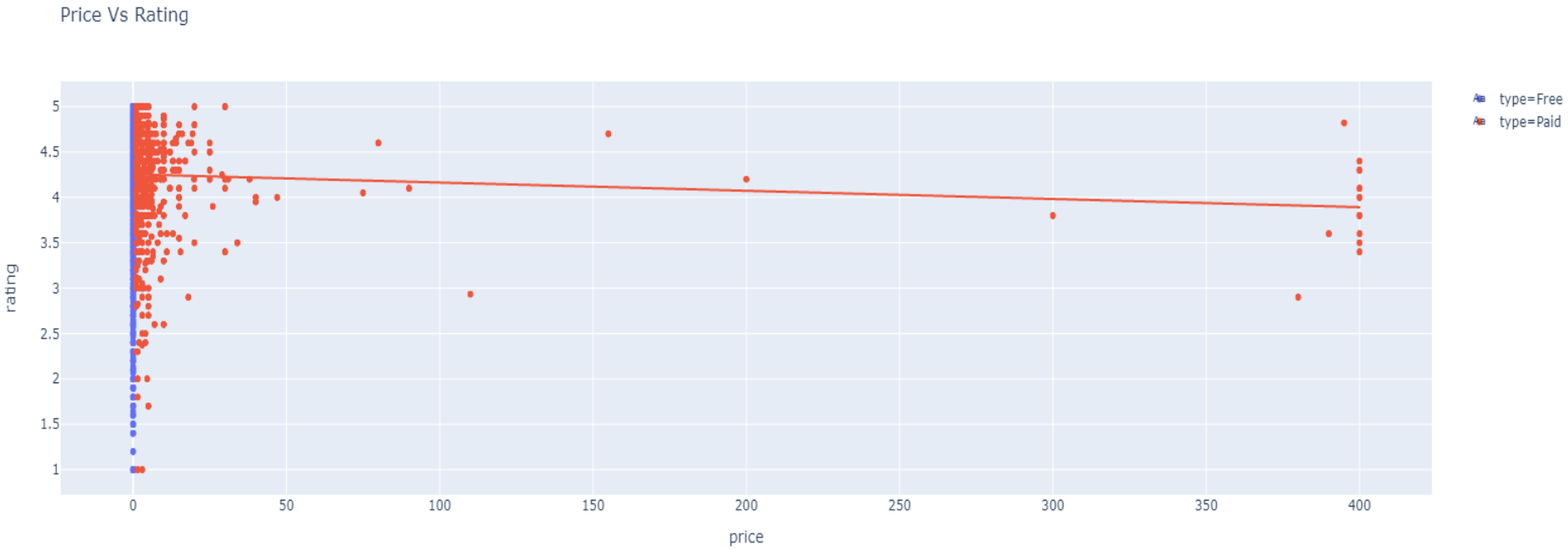
Smaller apps download faster and have higher install success rates, so it's important to monitor and optimize your app size for best results.

Reviews Vs Rating



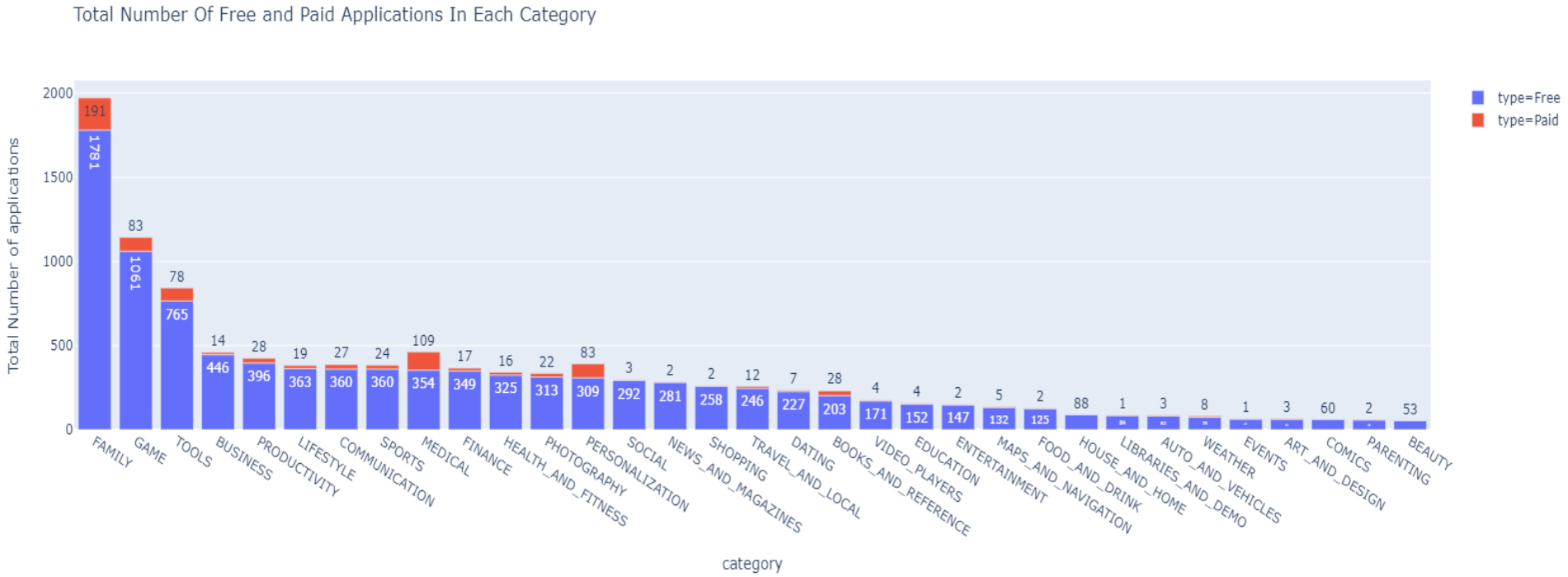
Obviously by looking at above scatter plot with trendline we are able to conclude that lesser the reviews on applications lesser the rating as well.

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.

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.

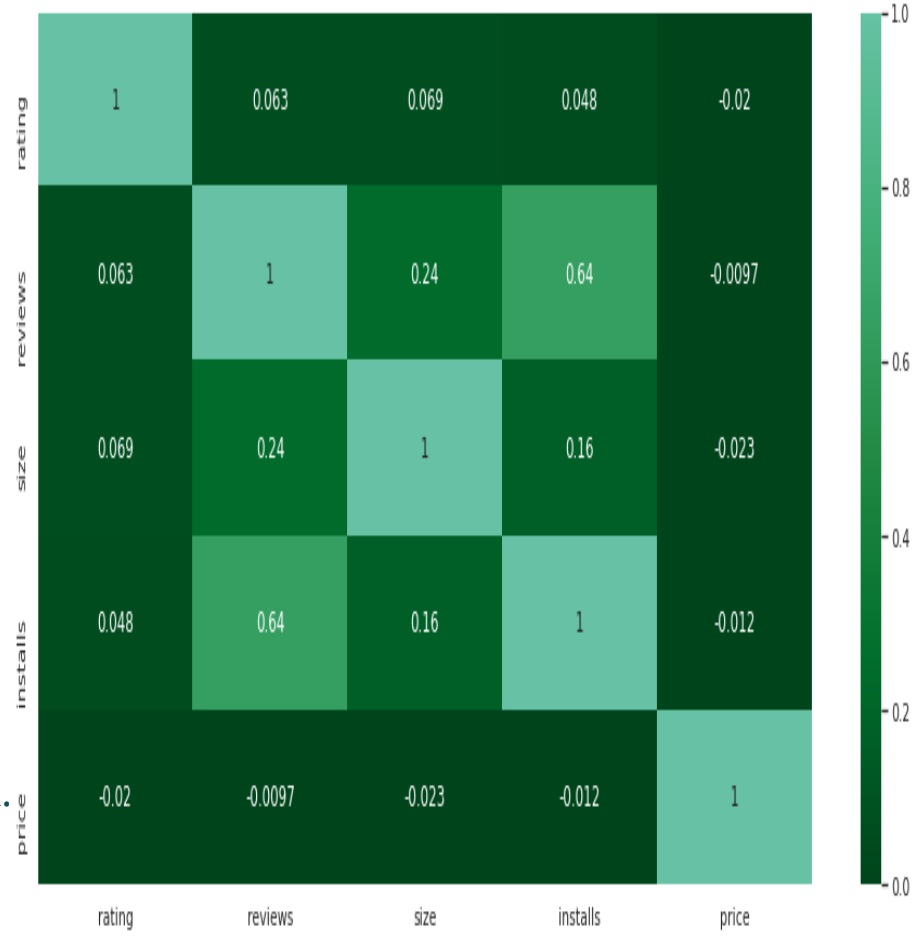
Heatmap Correlation Between Features

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.

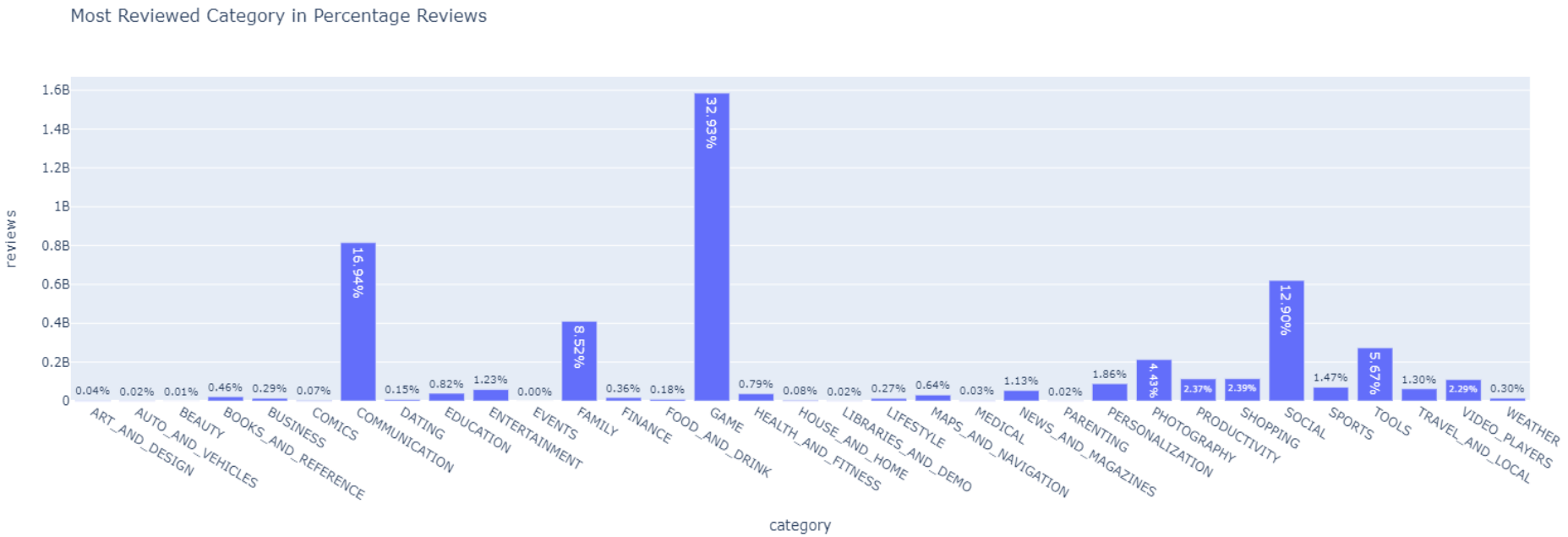
It means there is 64% chance of installing an app by the user if reviews are good. The average installation rate is positively correlated with user reviews, rating.

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.



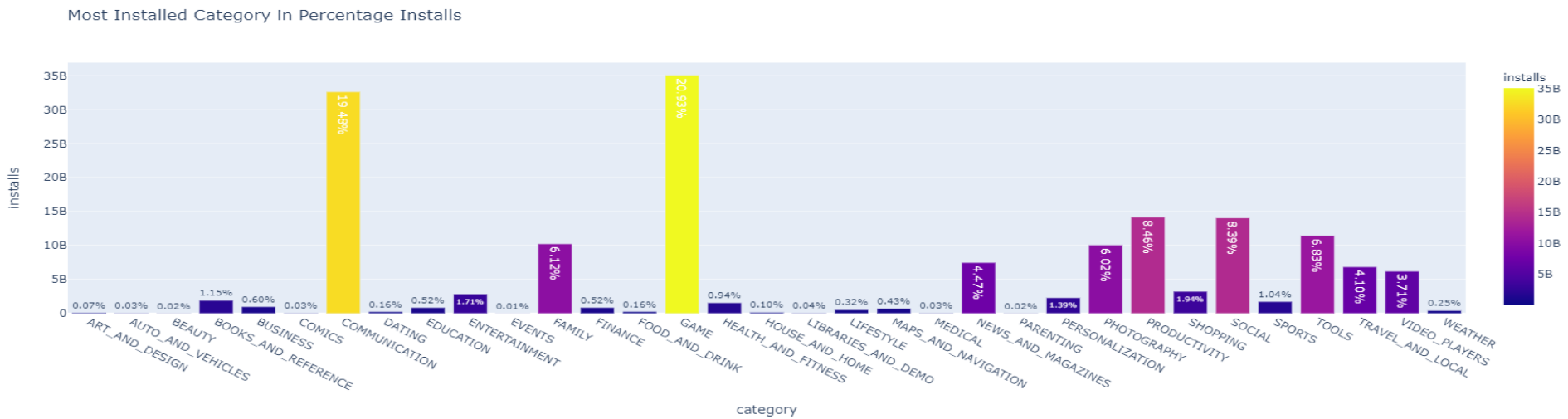
Most reviewed category



Reviews not only have the power to influence consumer decisions but can strengthen a company's credibility. 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 32.93% and 16.94% respectively.

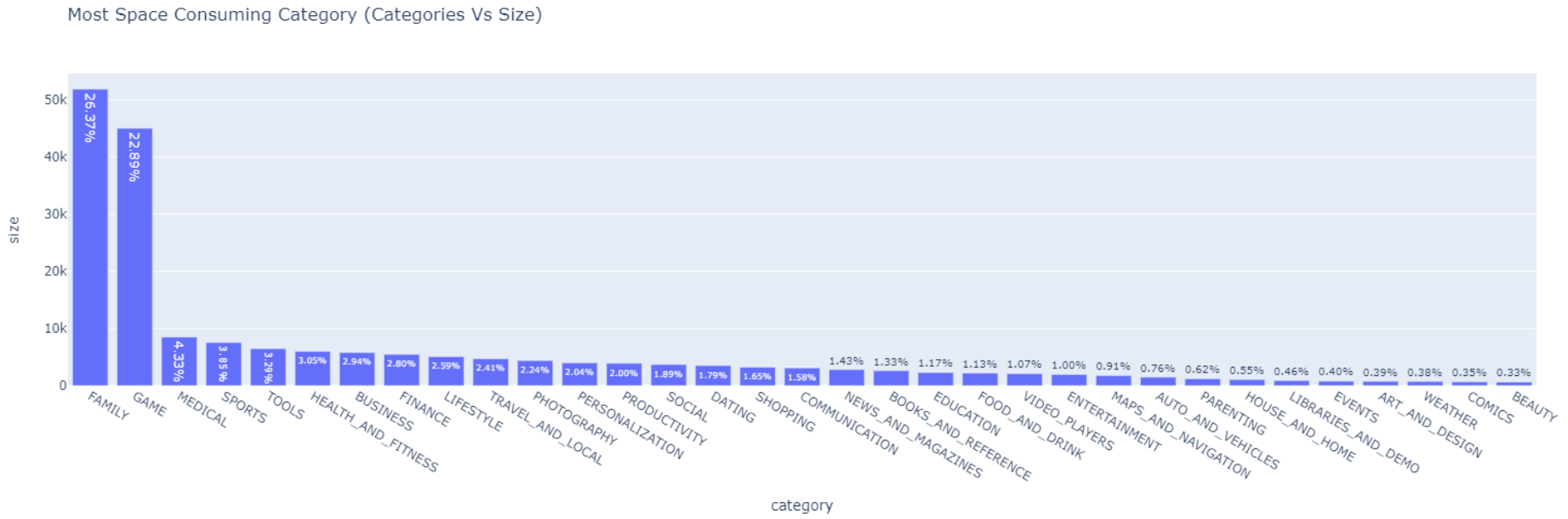
Most Installed Category



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.

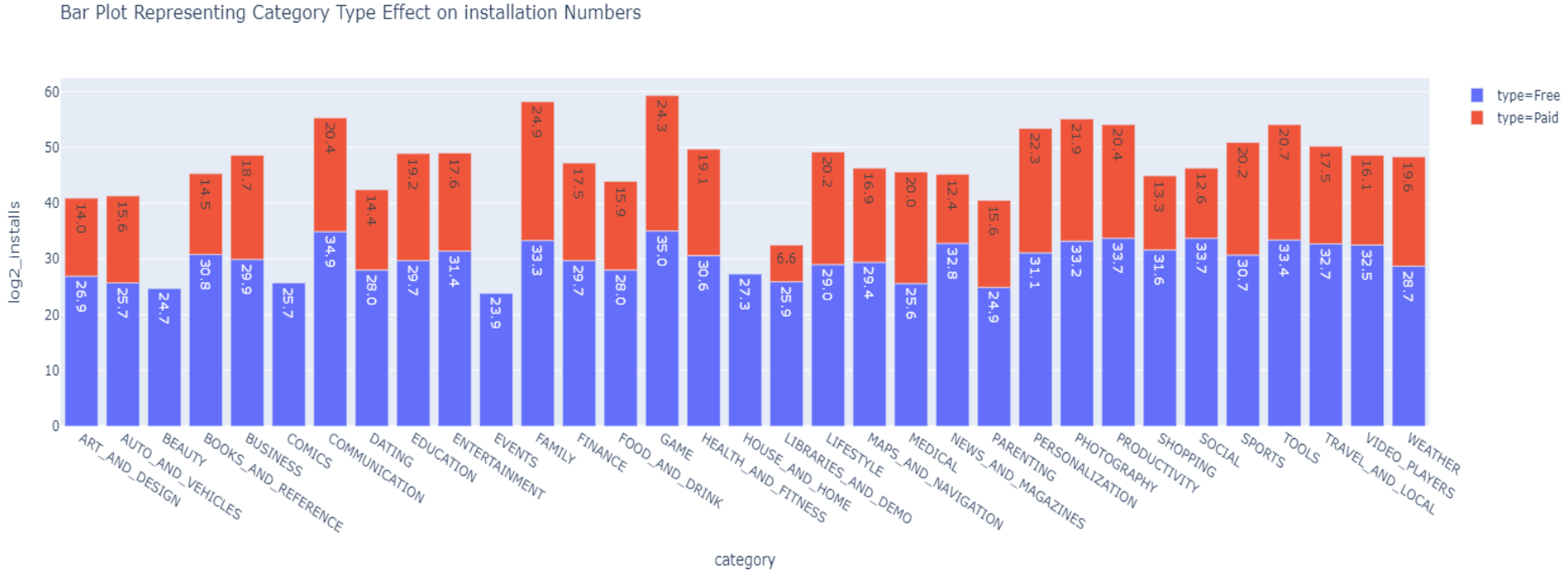
As we have seen heatmap correlation map that reviews, and installs are highly correlated with each other so here is the evidence that most positively reviewed category have the highest installation rate.

Most Space Consuming Category



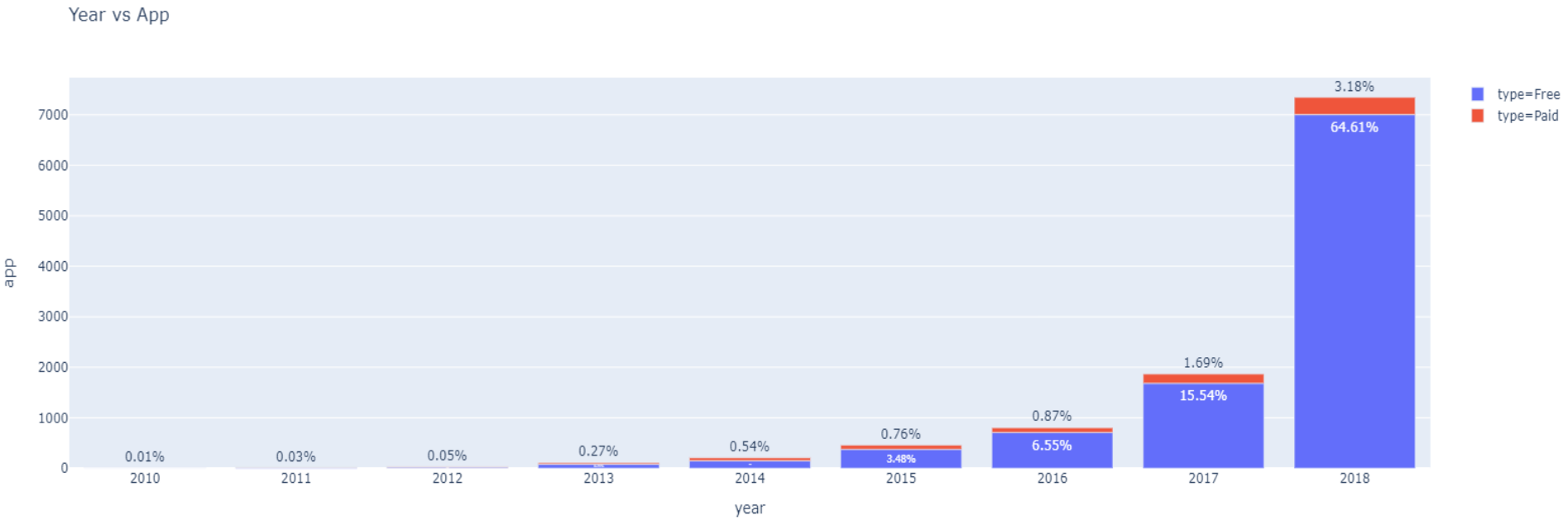
This bar plot represents the most space-consuming category on google play store. As we can see family and game are two categories that consume 26.37% and 22.89% space in the google play store, respectively. It means you will get more varieties of applications in these two categories; it also means there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

Category type effect



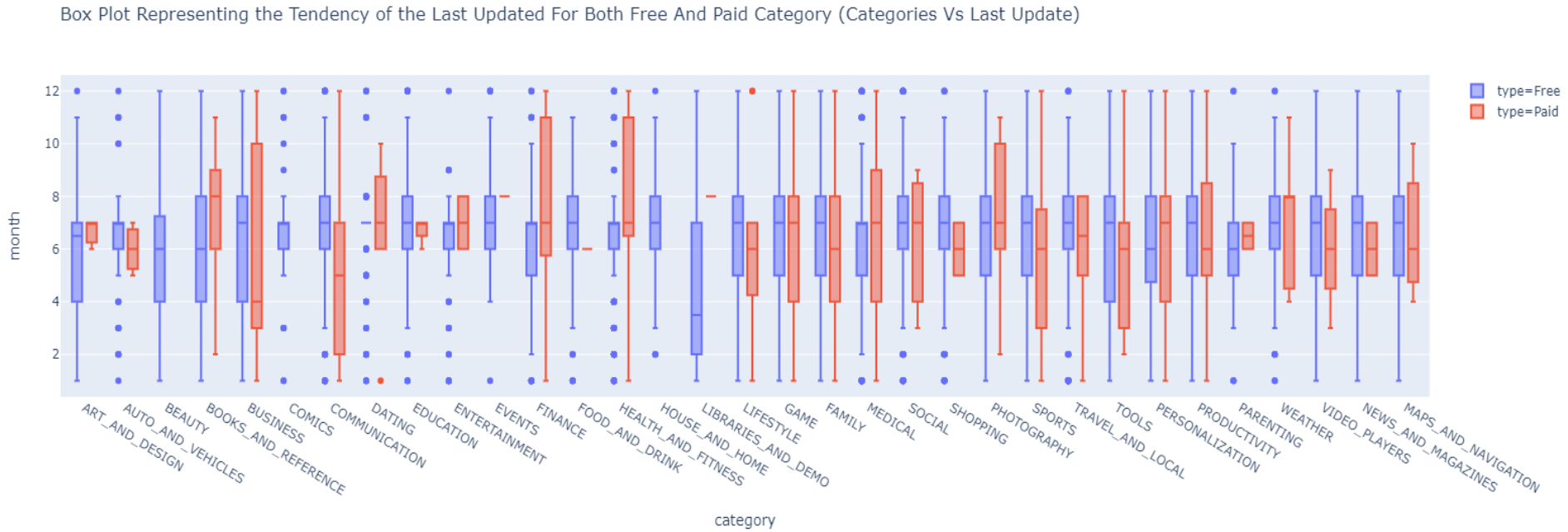
This bar plot indicates that the installation for the unpaid or free category application are much higher than that of paid one. The number of installs are converted into np.log2(installs) for the purpose of appropriate visibility of free and paid type installation.

Year vs App



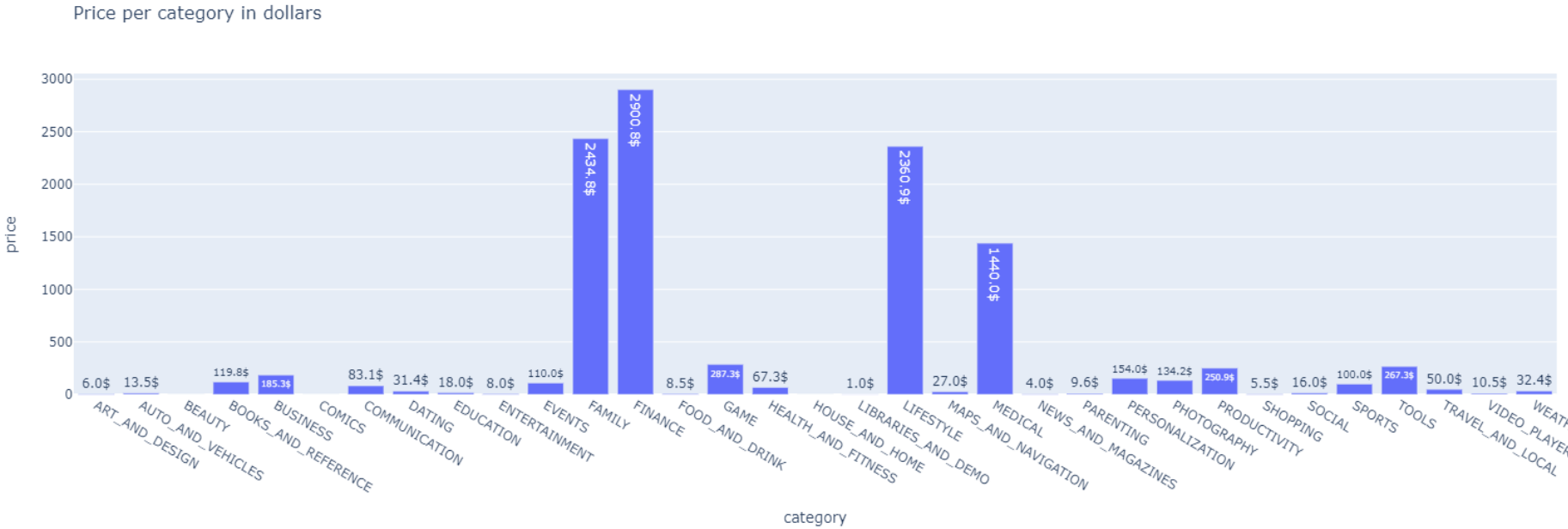
we can conclude that before 2012 there were no paid apps, but with the years passing free apps has been added more in comparison to paid apps, By comparing the apps updated or added in the year 2011 and 2018 free apps had increased from 0.01% to 64.61% and paid apps had gone from 0.27 to 3.18%. So, we can conclude that most people are after free apps.

Category vs Last Update



The above boxplot provides the number of application has been updated/added in each month for both free and paid applications. We can conclude from the above box plot that free categories people usually update more often. we can take inference that most applications are added or updated in the month of July, it is true for both kinds of applications free and paid.

Price Per Category



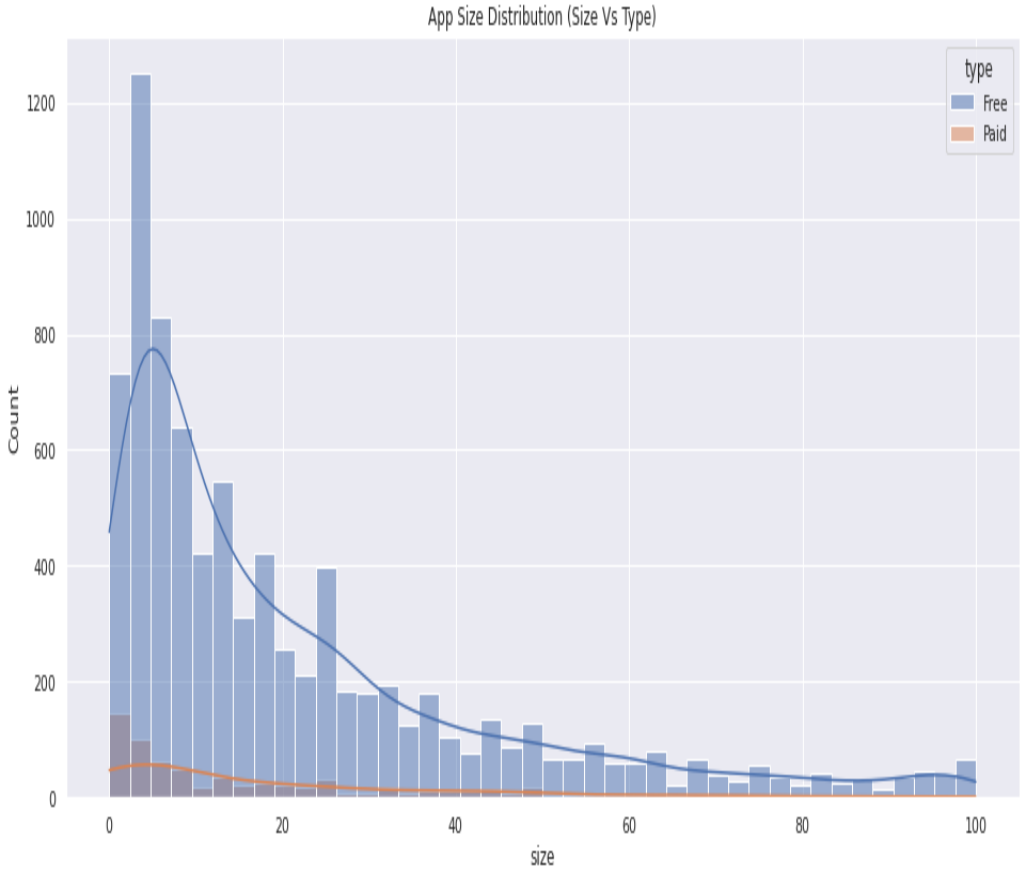
One can clearly understand that applications from the category finance have high price than any others, family category applications are second high priced applications, at third we have lifestyle category applications and at last, we have medical category applications, these four categories application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



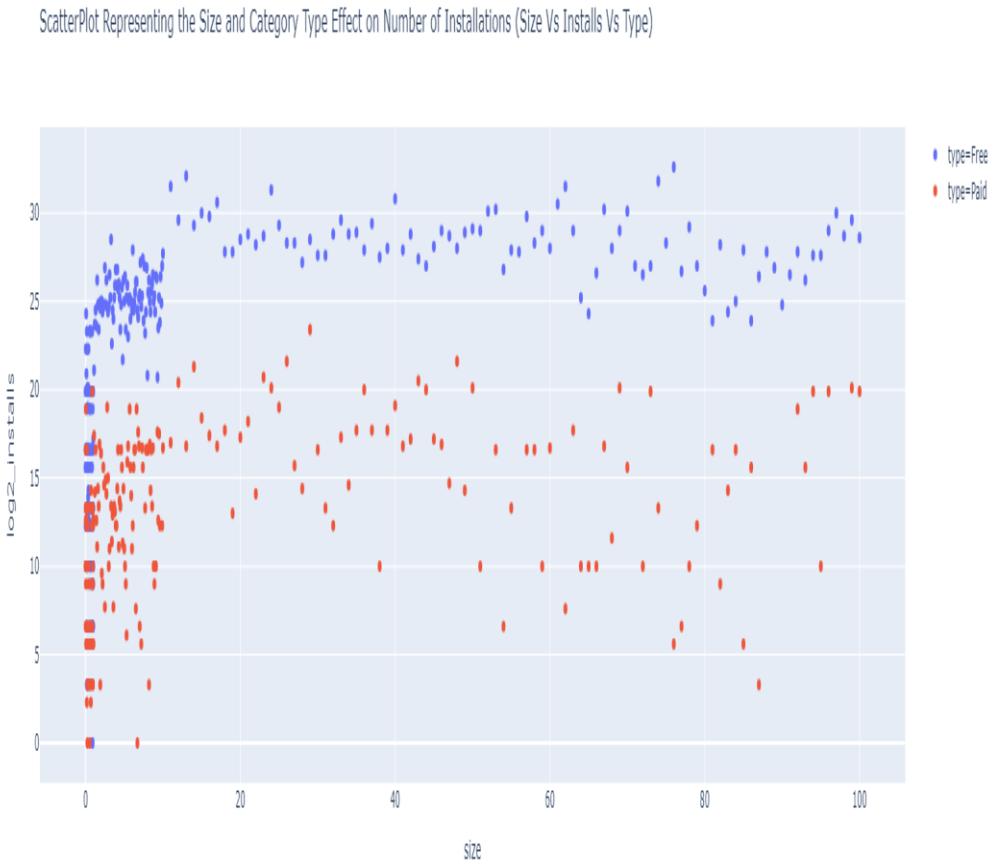
Combined effect of type, size on number of installs

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.

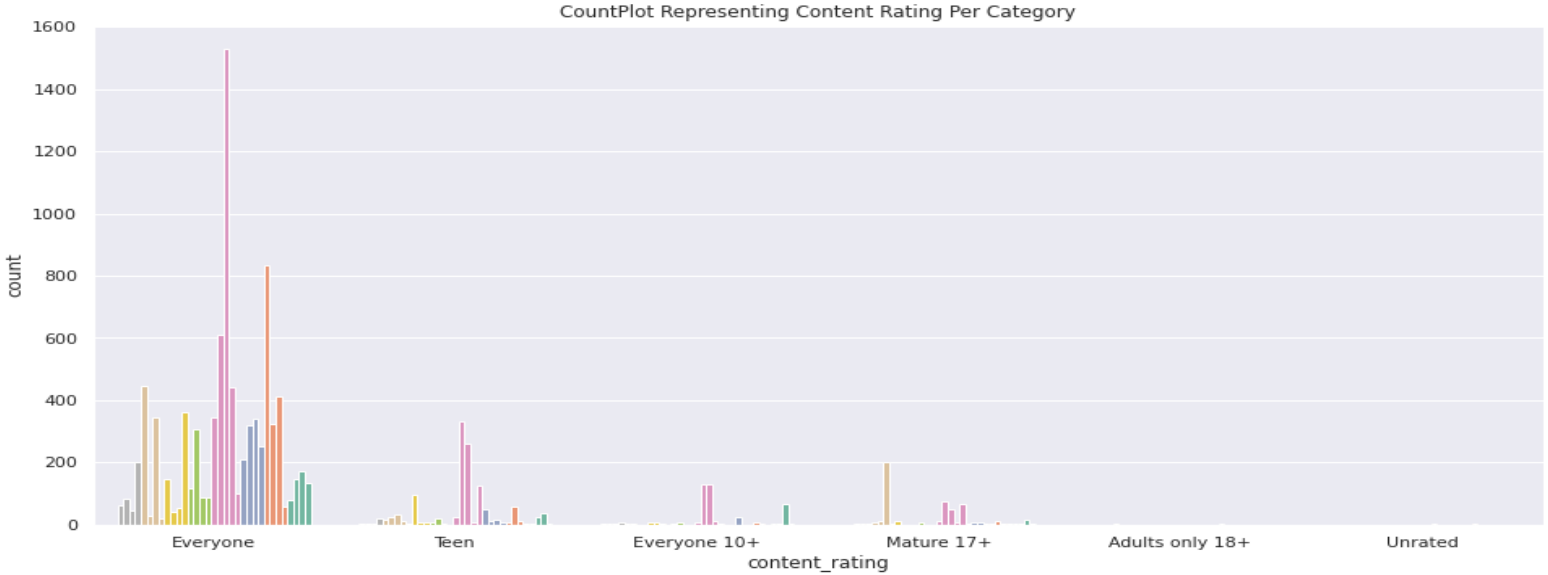
The scale of installs is converted into np.log2 scale, it is implemented to have explicit visualization.

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.

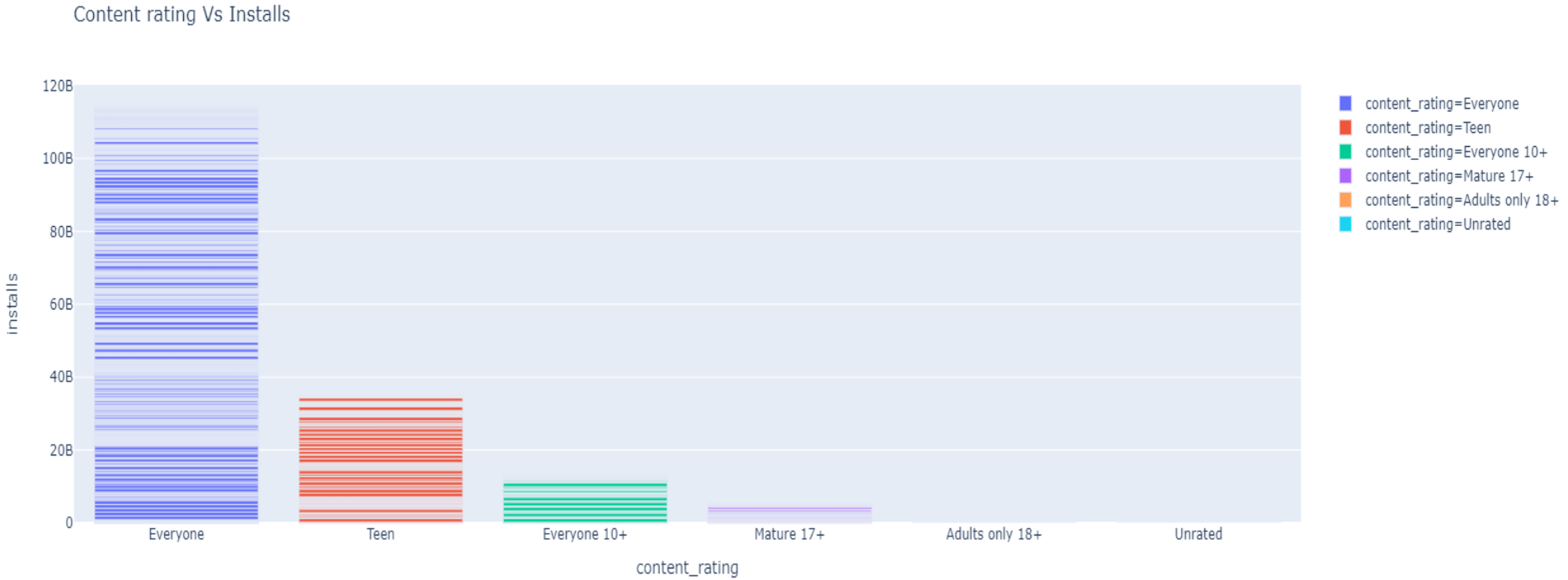


Content Rating For Each Category



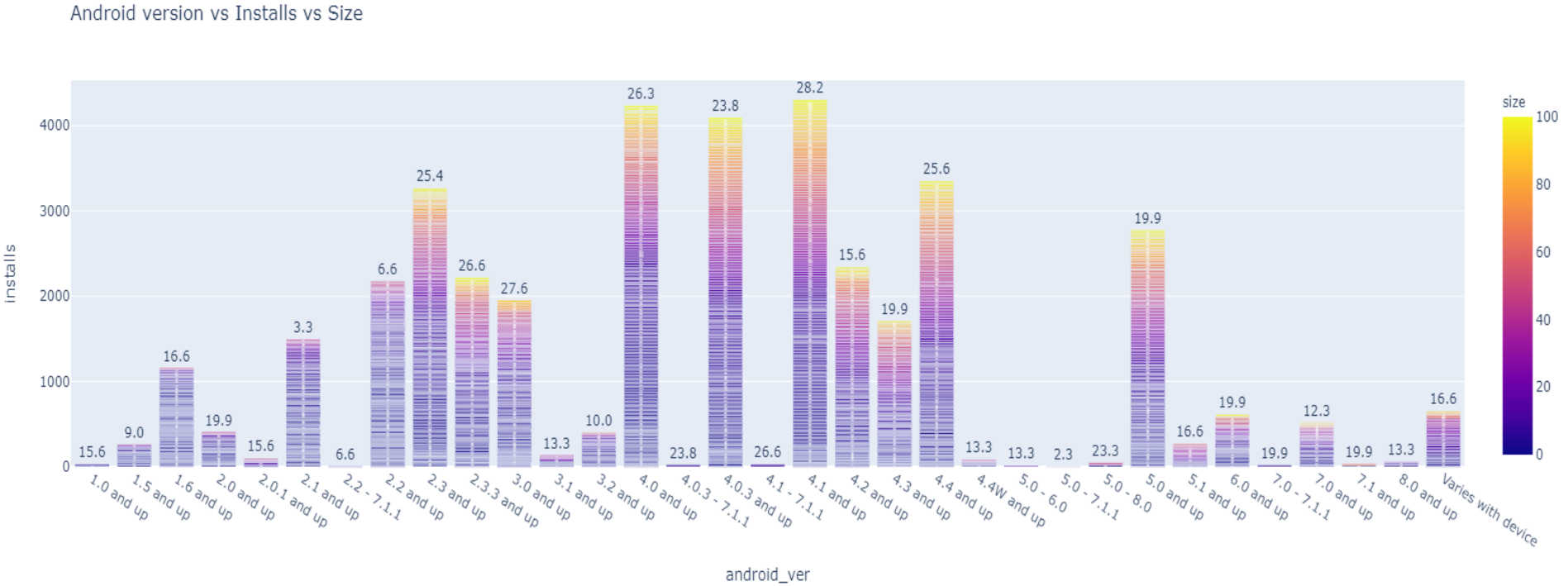
Content ratings are used to describe the minimum maturity level of content in apps. However, content ratings don't tell you whether an app is designed for users of a specific age. From above count plot we can conclude that Most of categories has content for everyone, excluding dating category which is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

Android Version vs Size vs Installs



From the above bar plot, we can see the range of application size and installs in each android version. 4.1 and up android version has the highest number of installs and the size of the application is mostly in the range of 0 to 40 Mb.

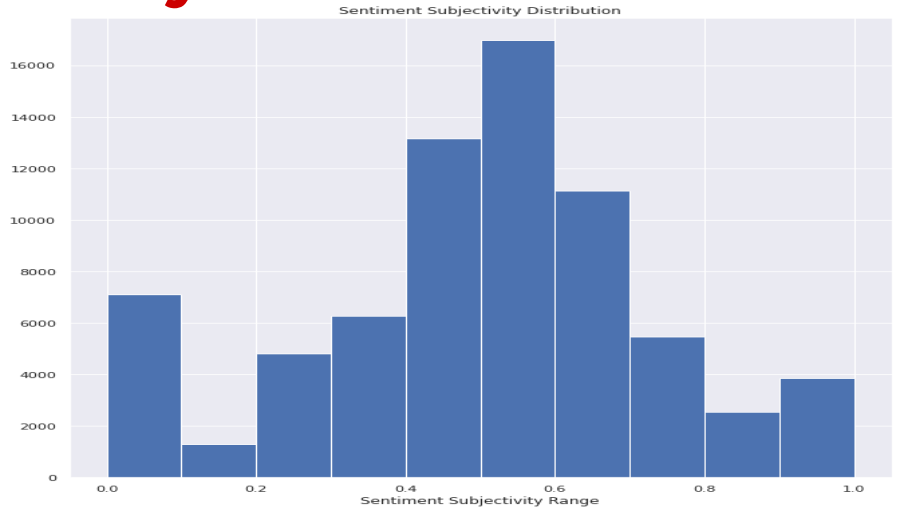
Sentiment Subjectivity and Polarity distribution

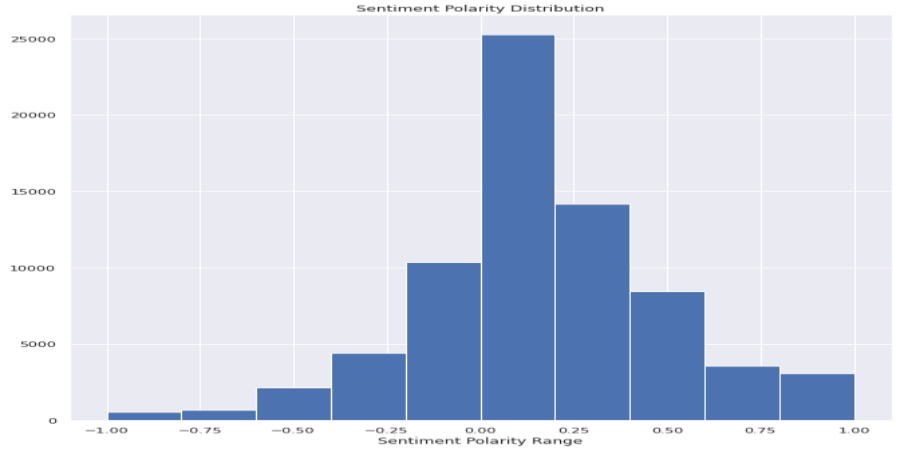
Sentiment is the emotion, feeling, opinion, or views held or expressed by users, sentiment subjectivity is float number value whose range lies in between 0 to 1. where one is very objective and 1 is very subjective.

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

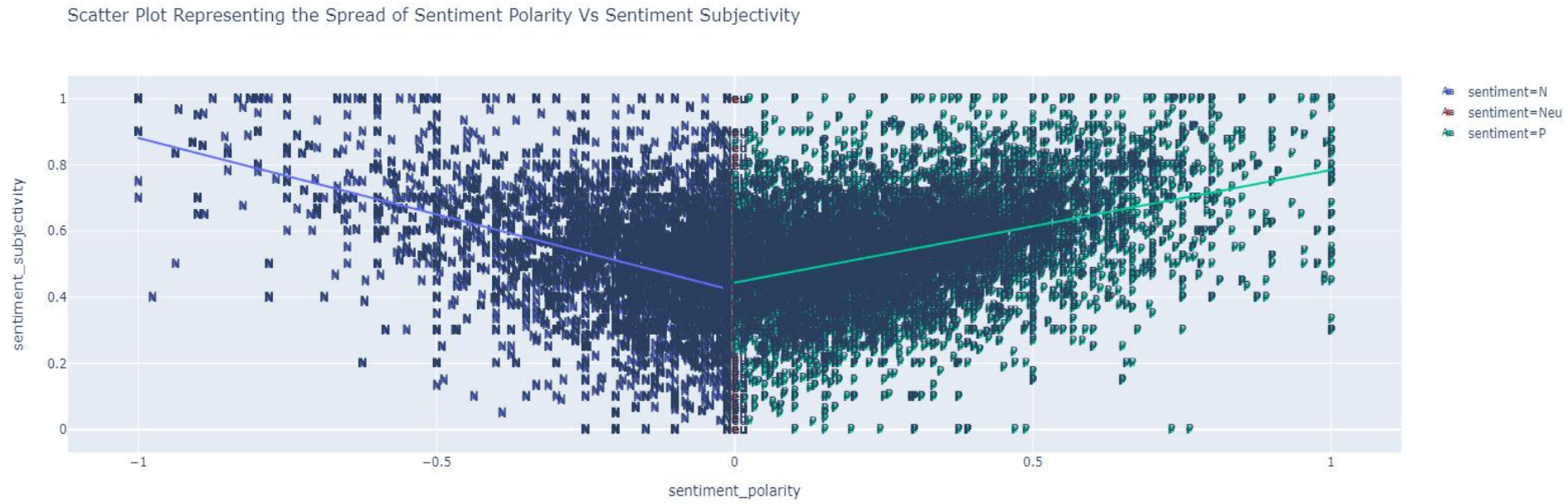
For given google play store data, sentiment subjectivity range lies between 0.5 to 0.7 that's positive one.

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



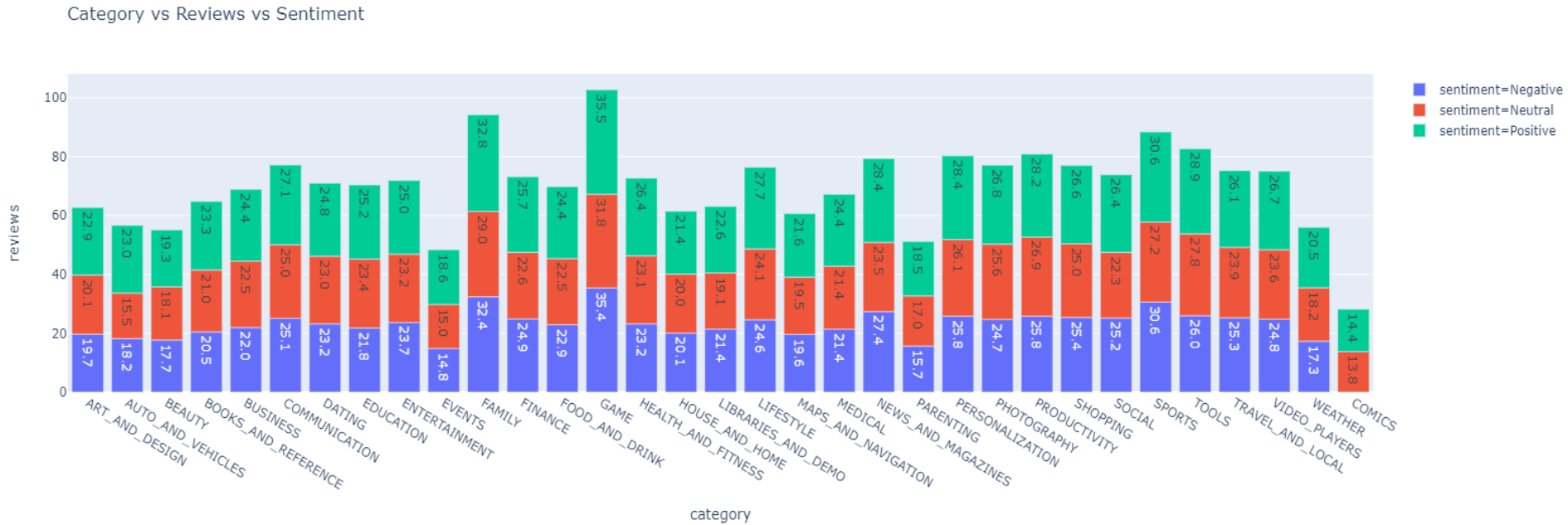


Sentiment Subjectivity Spread Analysis



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.

Category vs Reviews vs Sentiment

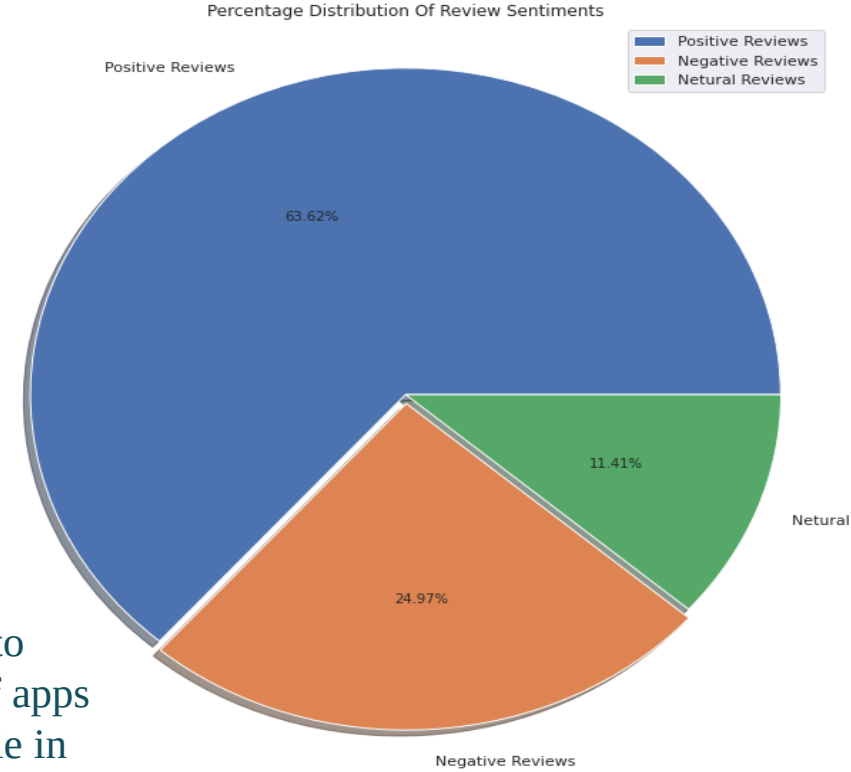


We always treat neutral reviews as insignificant, but it is not the case always, Neutral reviews are never neutral. A neutral review affects both negative and positive reviews and can also reinforce both. Statistically, neutral reviews are a good thing because without them, positive reviews will be overestimated,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

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.

Conclusion

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

1. Category Type: Before 2012 there were no paid apps, but with the years passing free apps has been added more in comparison to paid apps, By comparing the apps updated or added in the year 2011 and 2018 free apps are increases from 0.01% to 64.61% and paid apps are gone from 0.27 to 3.18%. So, we can conclude that most people are after free apps.

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 know the performance of the applications

4. The most installed category: As we have explored applications belong to the category gaming and followed by communication are being installed the most, this is the clue to choose the category for application.

5. Price vs Rating vs Review vs Size: Reviews correlate about 0.063 to rating, as price increases there are minimum downloads. The inclusion of a feature that lets users notify developers if they are unsatisfied with an update could go a long into not plummeting an app with a low rating and installs to the ground. You may encourage users to rate your app and write an App Store review.

Thank You!