# Google Play Store Analytics with R

Final Report

Sumadhuri Damerla,Shaoor Jan,Ashjan Khan

## Contents



# 1 Introduction to the dataset

## 1.1 Data source

The data source used for this analysis is the *2018 google play store*(https://www.kaggle.com/lava18/google-play-store-apps) collected from Kaggle.

## 1.2 Description of the dataset

The dataset is a collection of web-scraped data of 10,000 apps from Google Play Store. Google Play Store originally referred as the Android Market, is Google's official store and portal for Android apps, games and other content for Android-powered phone, tablet or Android TV device. As of May 2017, it has over two billion monthly active users, the largest installed base of any operating system, and as of January 2020, the Google Play Store features over 2.9 million apps[13].

The variables of the dataset are as follows:

1) App (Name) – Name/Title of the application
2) Category (App)- Category/Domain to which the app belongs to
3) Rating (App)- Overall user rating of the app
4) Reviews (User)- Number of user reviews for the app
5) Size (App)- Space or memory that the app takes up
6) Installs (App)- Number of user downloads/installs
7) Type (Free/Paid)-Apps may be free or paid depending on the developer's choice
8) Price (App)-Price of the app if not free
9) Content Rating - Age group the app is based off at - Children / Mature 21+ / Adult

10) Genres (Detailed Category)- An app can belong to multiple genres, For eg, a musical family game will belong to Music, Game, Family genres.
11) Last Updated (App)- Date when the app was last updated on Play Store
12) Current Version (App)- Current version of the app available on Play Store

13) Android Version (Support) – minimum version of android it takes to have the app on the device

# 2 Purpose of the project

- The aim of our project is to find out if we can predict ratings of an app based on different variables and we intend to summarise the different factors that influence the success of an app.These analysis might also help the developer community to build more successful apps by taking accurate data-based decisions, and focusing on those aspects of applications that matters most.
- Also,since this is the first time we are doing data analysis using R, it is a fun way to learn and to strengthen the concepts learned during the course by taking a hands-on approach.

# 3 Intended audience of the project

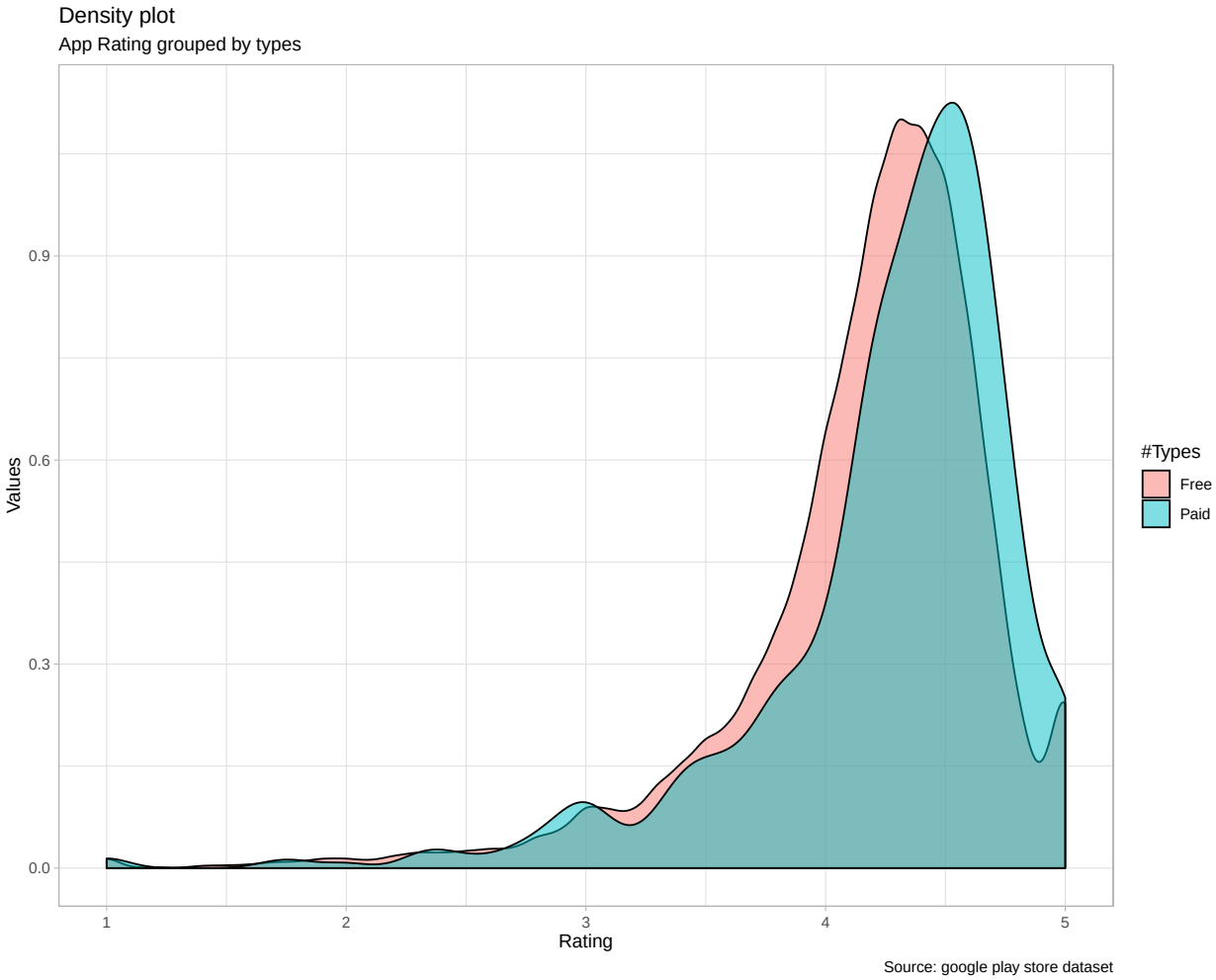
- We believe there's a diverse set of audience who might be interested in our project. As of Febraury 2020, 73.3% of the mobile operating system market share belongs to Android devices[11]. This large community consists of the general public who use android devices and appstore,the developer community and anyone who wants to understand how the app market works.
- This project is primarily intended for the growing developer community. It will help them make data backed decisions before launching their application. Besides developers, it is also helpful for tech journalists, Google Play Store users or any other interested party.

# 4 Exploratory Data Analysis to find factors for success of an app

Generally, the most successful apps have high ratings and high installs.To look at which app makes it to the top, we consider ratings and installs, so we explore these to find any relationship or trends

## 4.1 Rating column in depth

### 4.1.1 Distribution of rating

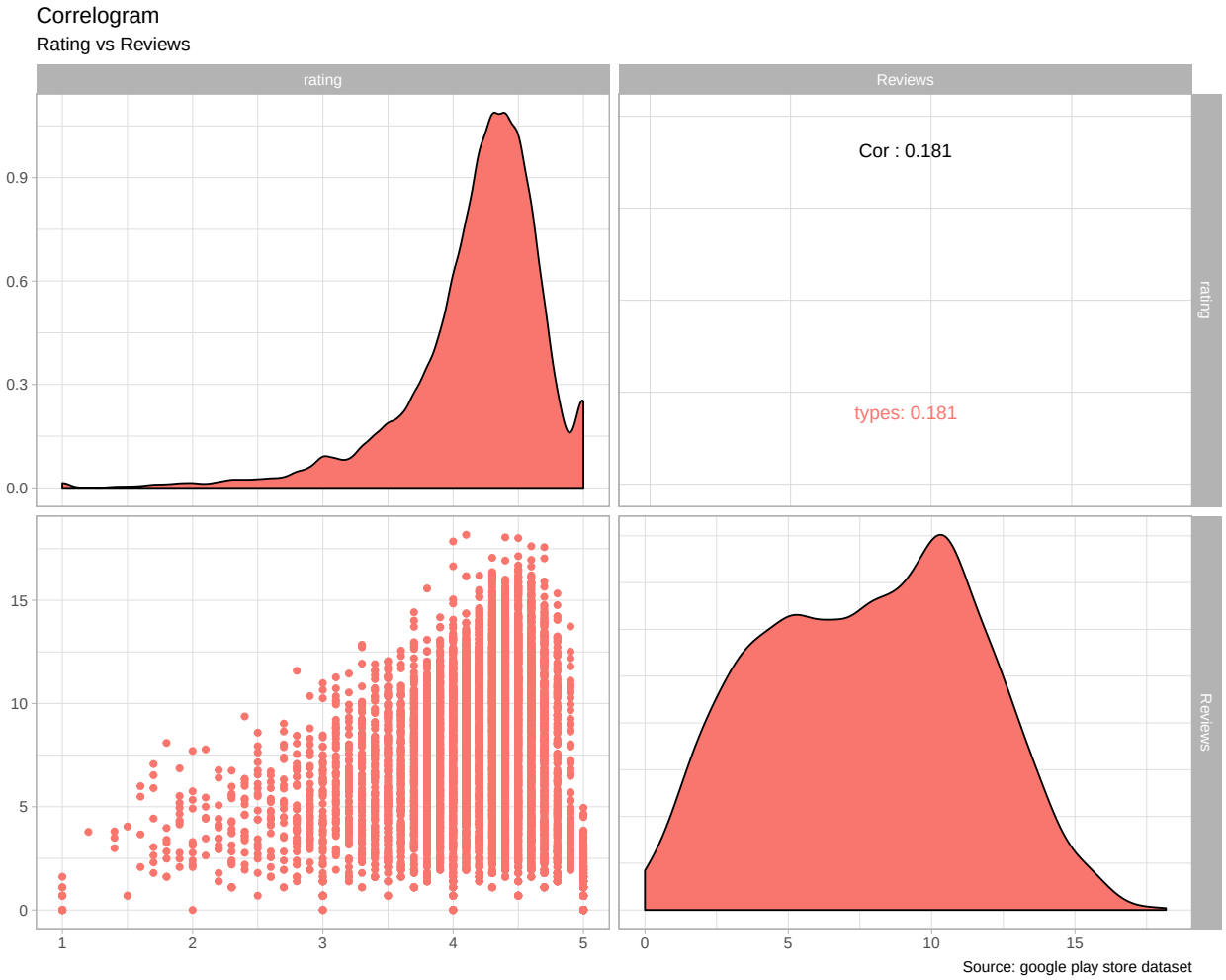



- We can observe that the number of apps with low ratings are less in number, and most apps have a ratings between 3-5.

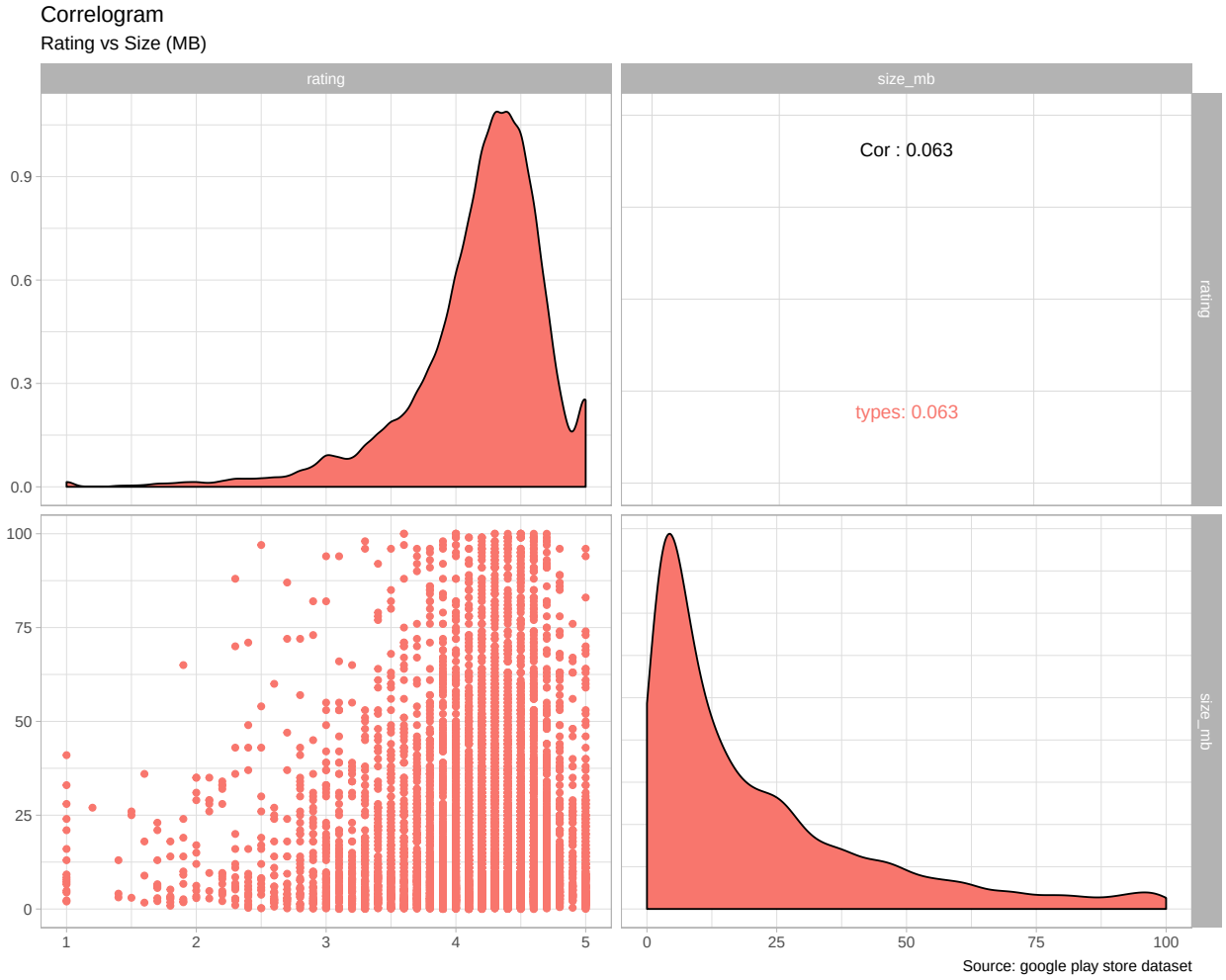
First, we plot correlograms of ratings versus different columns to find any relationship. Correlograms are useful to understand the relationship between different numerical variables. If the correlogram index is 1, it means that the variables are directly proportional to each other.

## 4.2 Correlogram plots

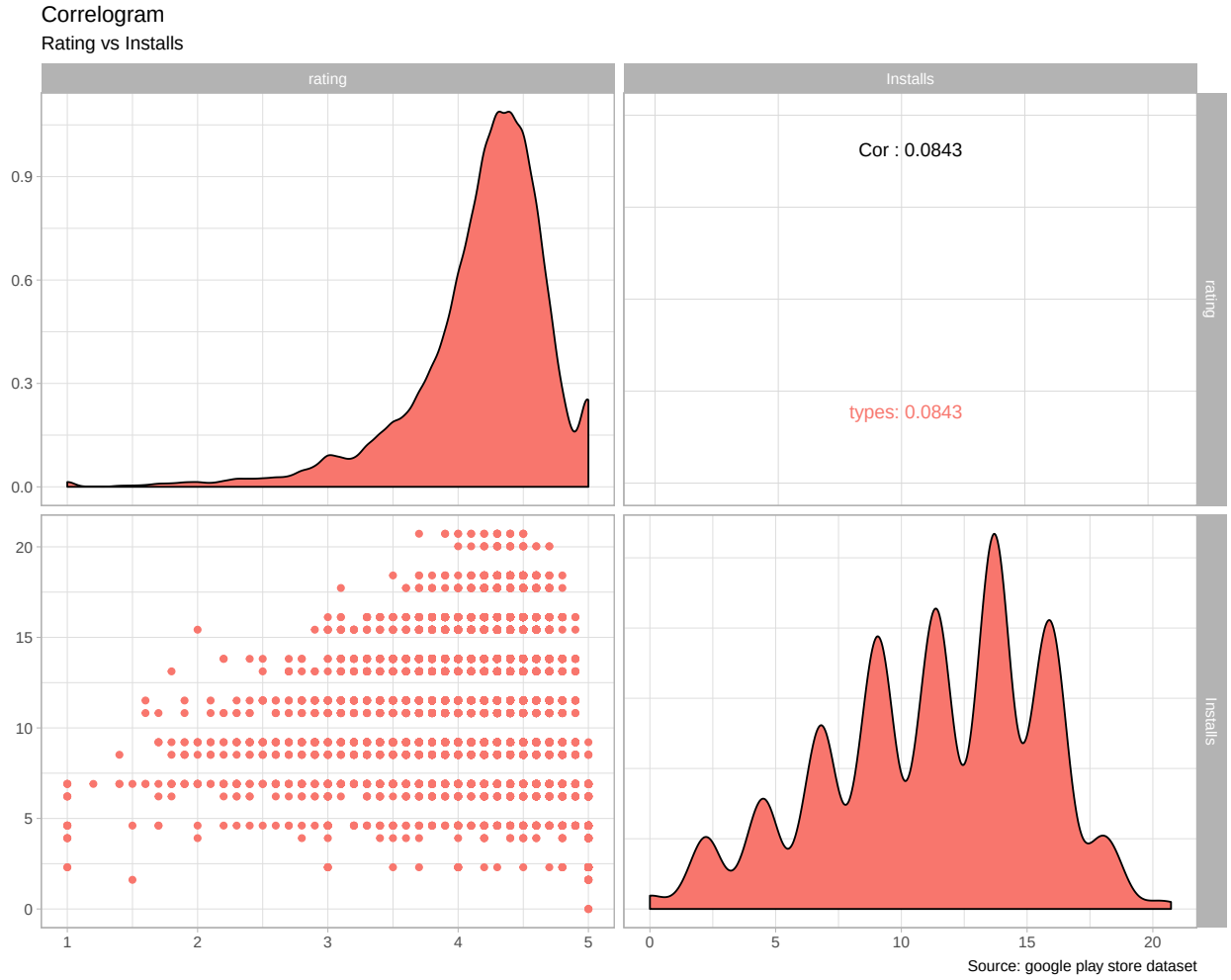
### 4.2.1 Rating vs Reviews



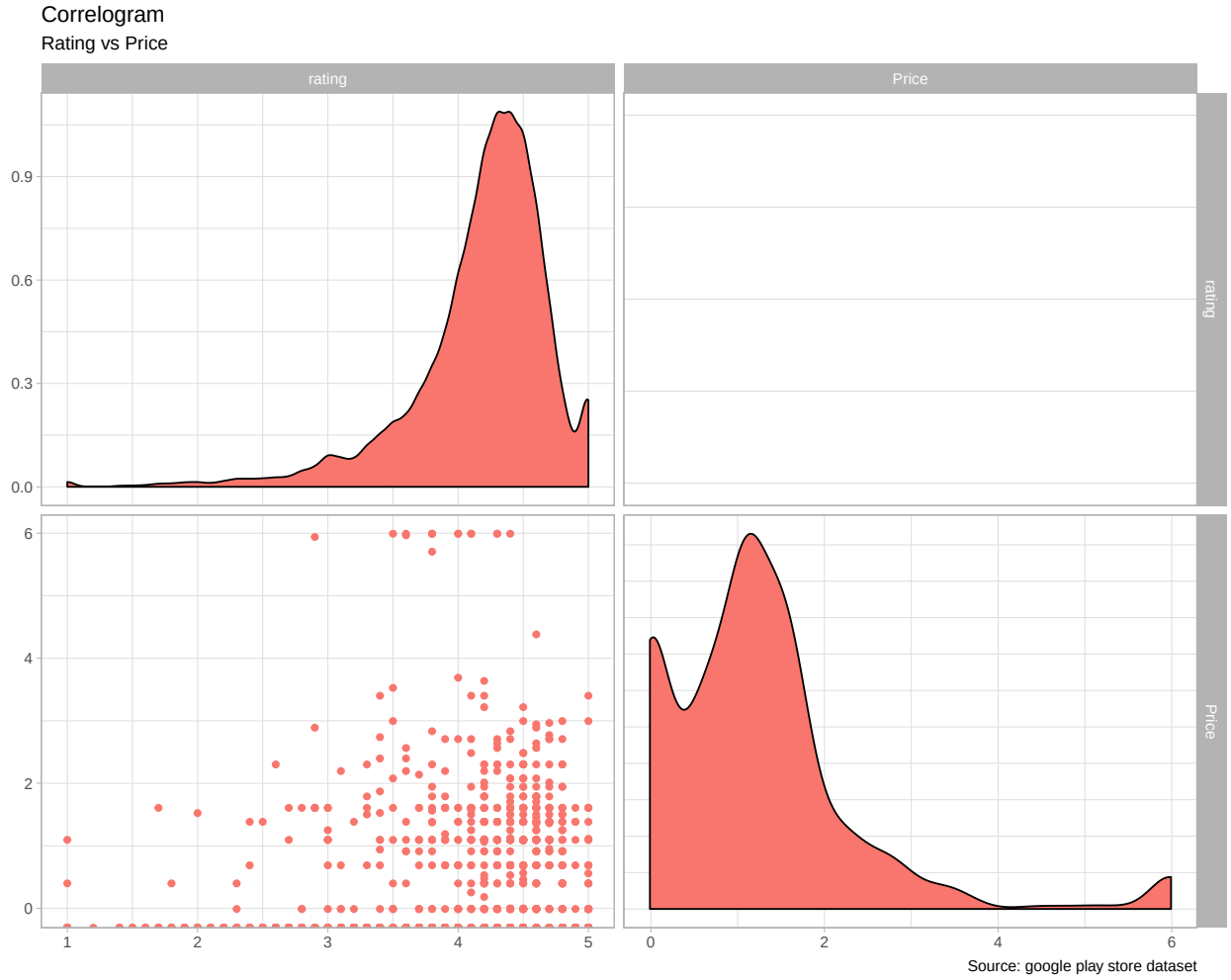
### 4.2.2 Rating vs Size



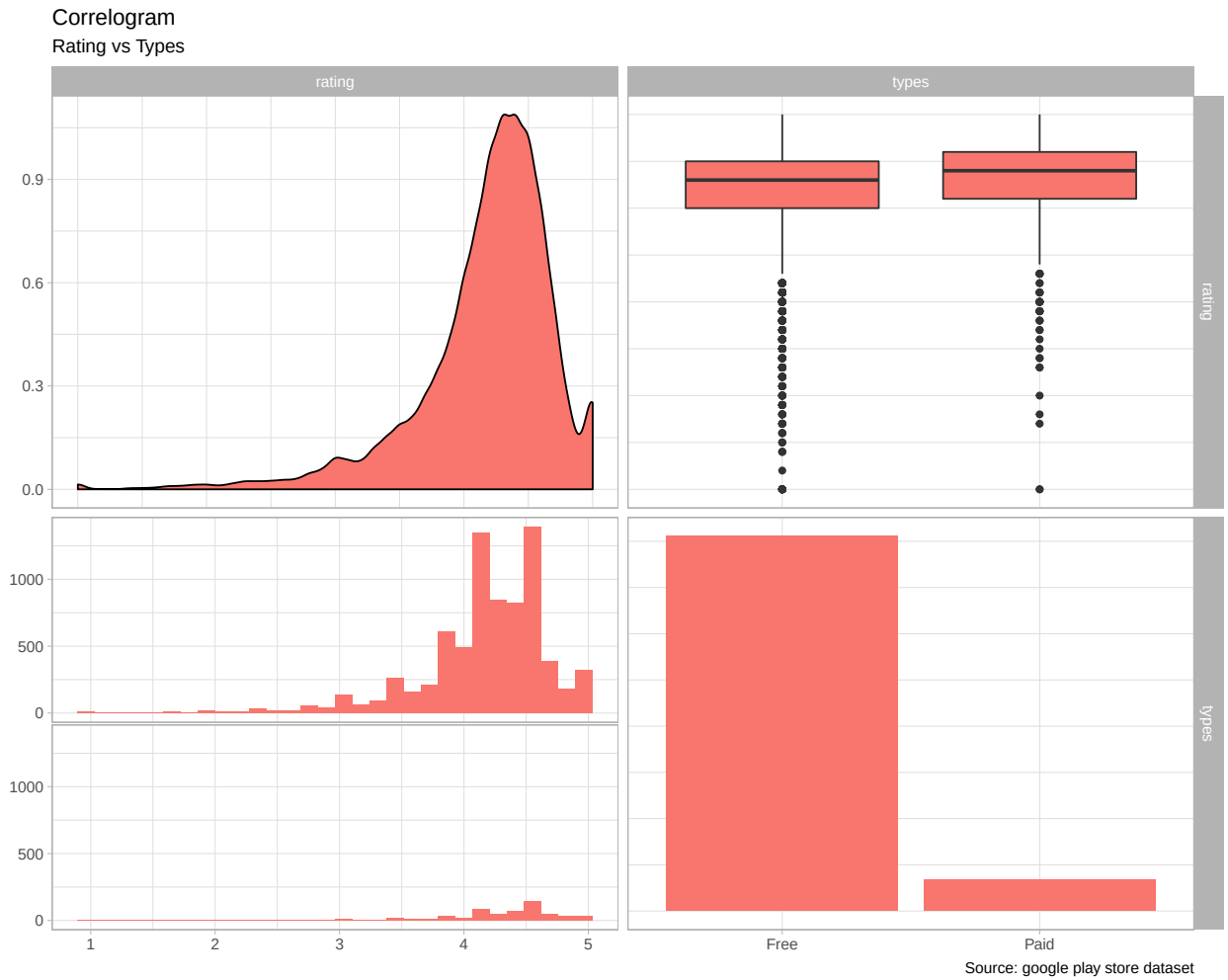
### 4.2.3 Rating vs Installs



### 4.2.4 Rating vs Price

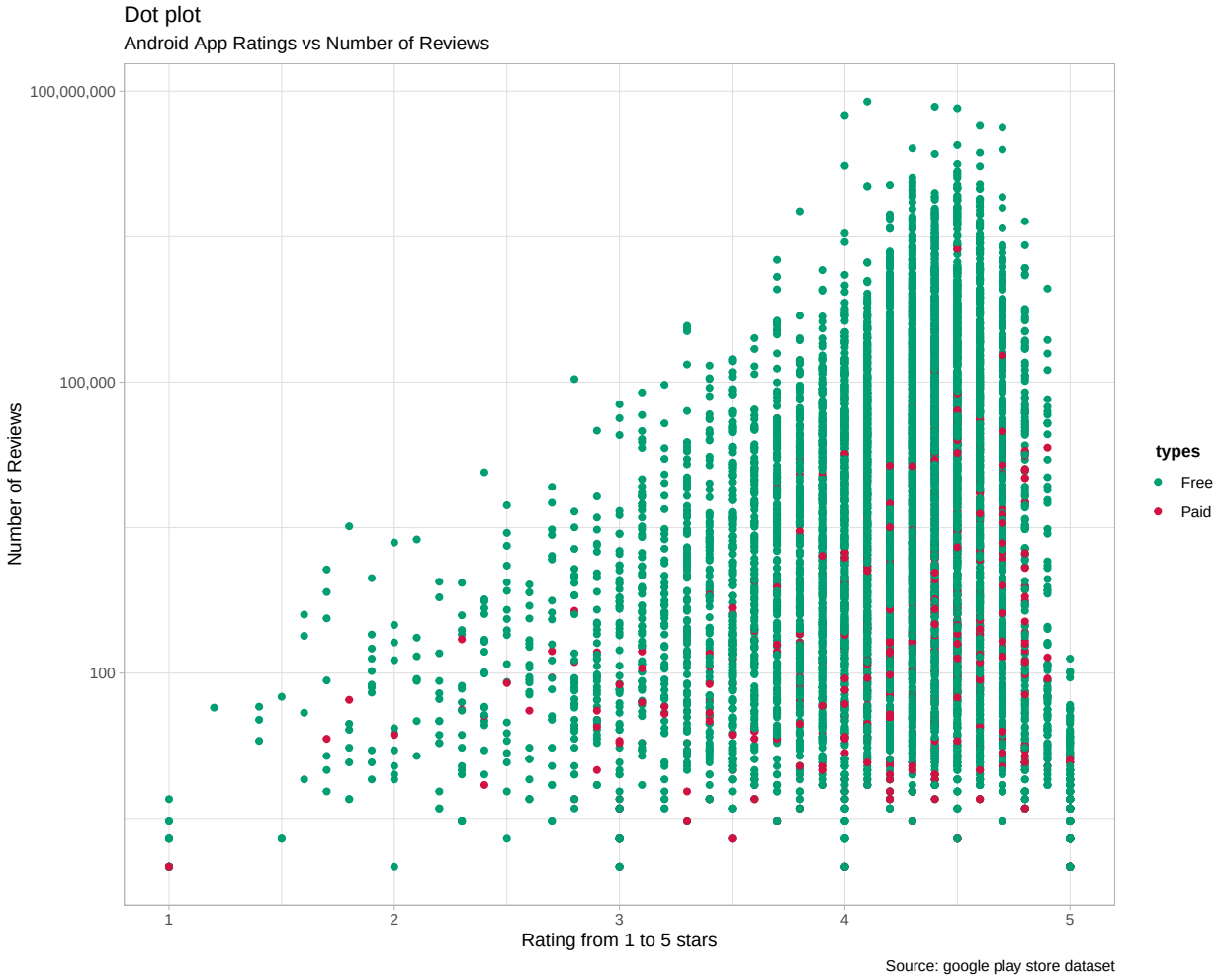


### 4.2.5 Rating vs Types




**Finding:**
+ Each correlation index talks about the relationship between plotted columns. If the index is 1, it means they have linear relationship + Each plot on the diagonal refers to the density plot of the respective column + We can observe that there is no significant linear relationship between rating and the plotted numerical variables.

### 4.2.6 Plot of reviews vs app ratings

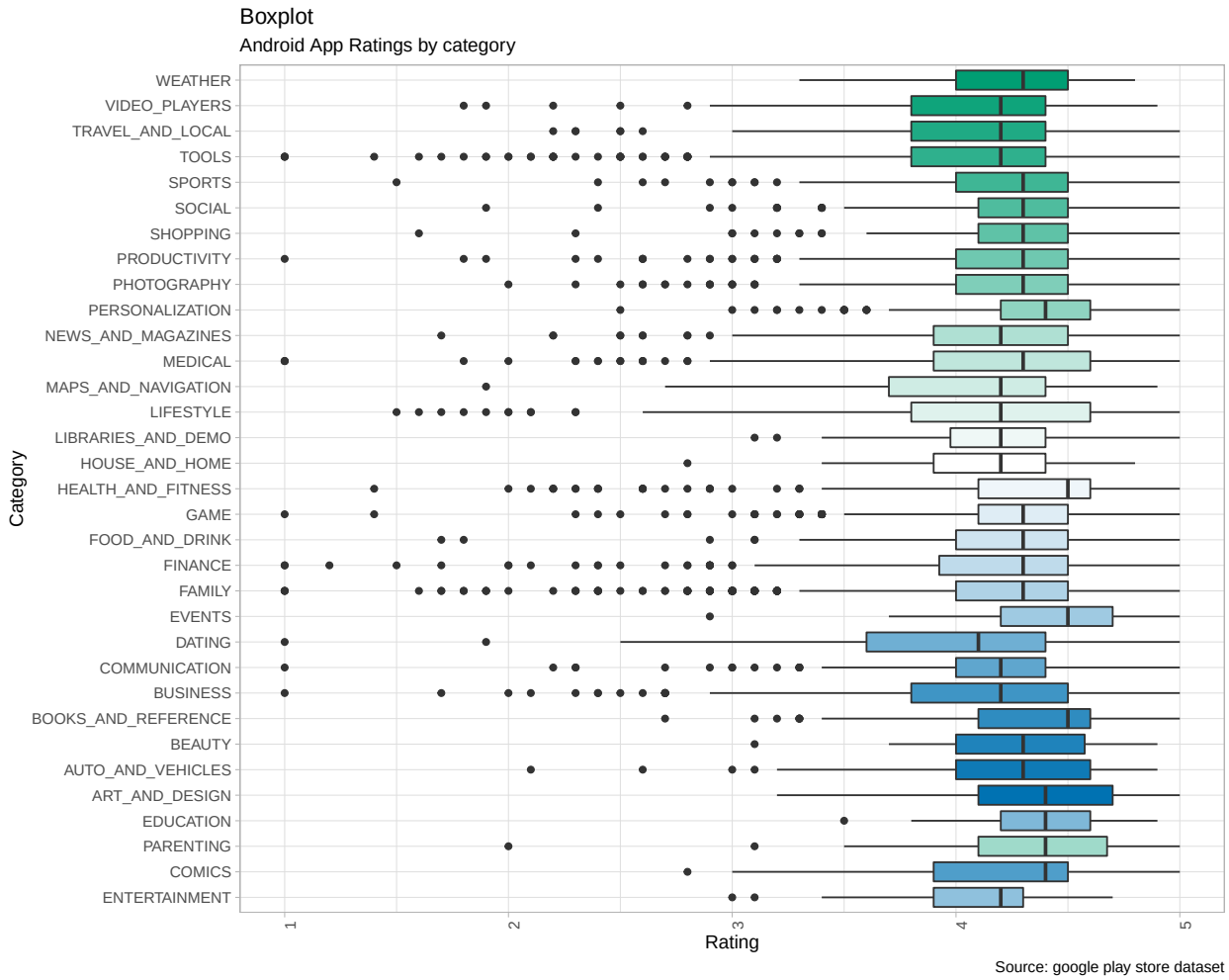



**Finding:** We can observe that the number of reviews influence the ratings. Generally, as the number of reviews increase, the rating is higher.

We now explore other factors that might potentially influence rating

### 4.2.7 App rating vs category

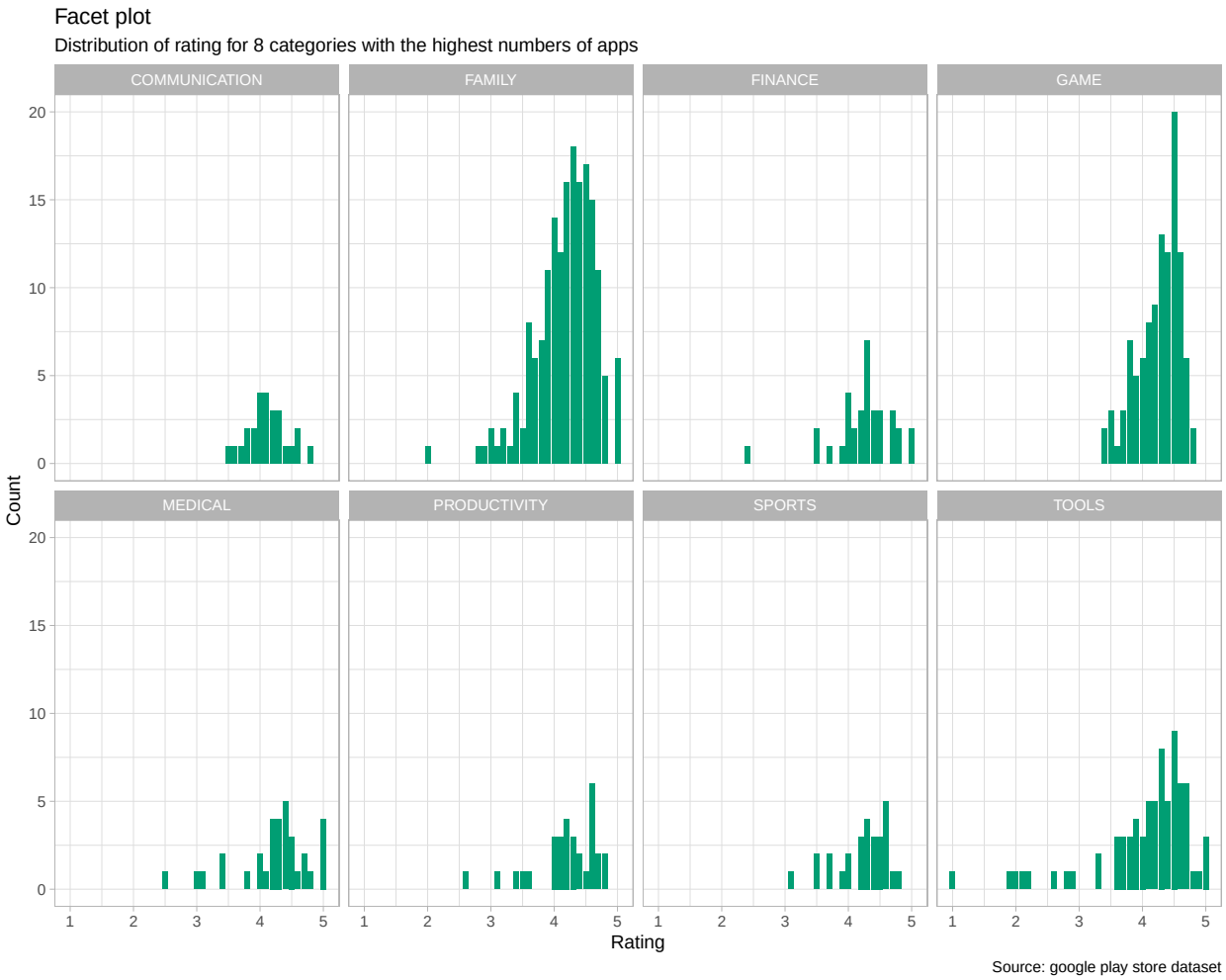
To check the relationship between category and rating, we plotted a box plot with rating on y-axis and category on x-axis.




**Finding:** This graph shows that for some categories like TOOLS, FAMILY, FINANCE and LIFESTYLE a great majority of applications fall below first quartile. Thus, even though median rating is high, deviation from median is significant.

### 4.2.8 Distribution of rating for 8 categories with the largest numbers of apps

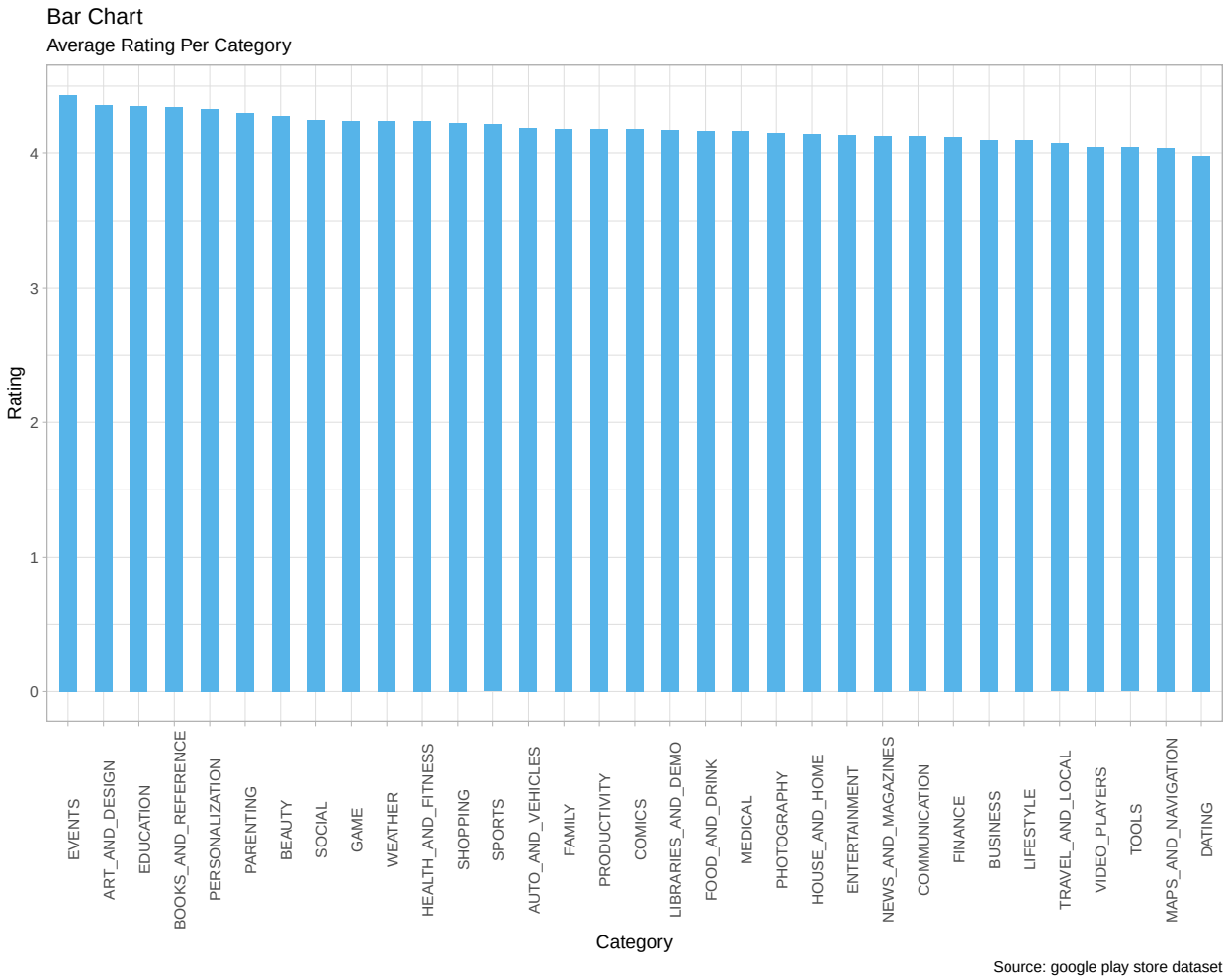
Here we look at the distribution of rating across different categories. We chose 8 categories with the largest number of applications.




**Finding:** The distribution of rating varies significantly across each category.

### 4.2.9 Average rating per category

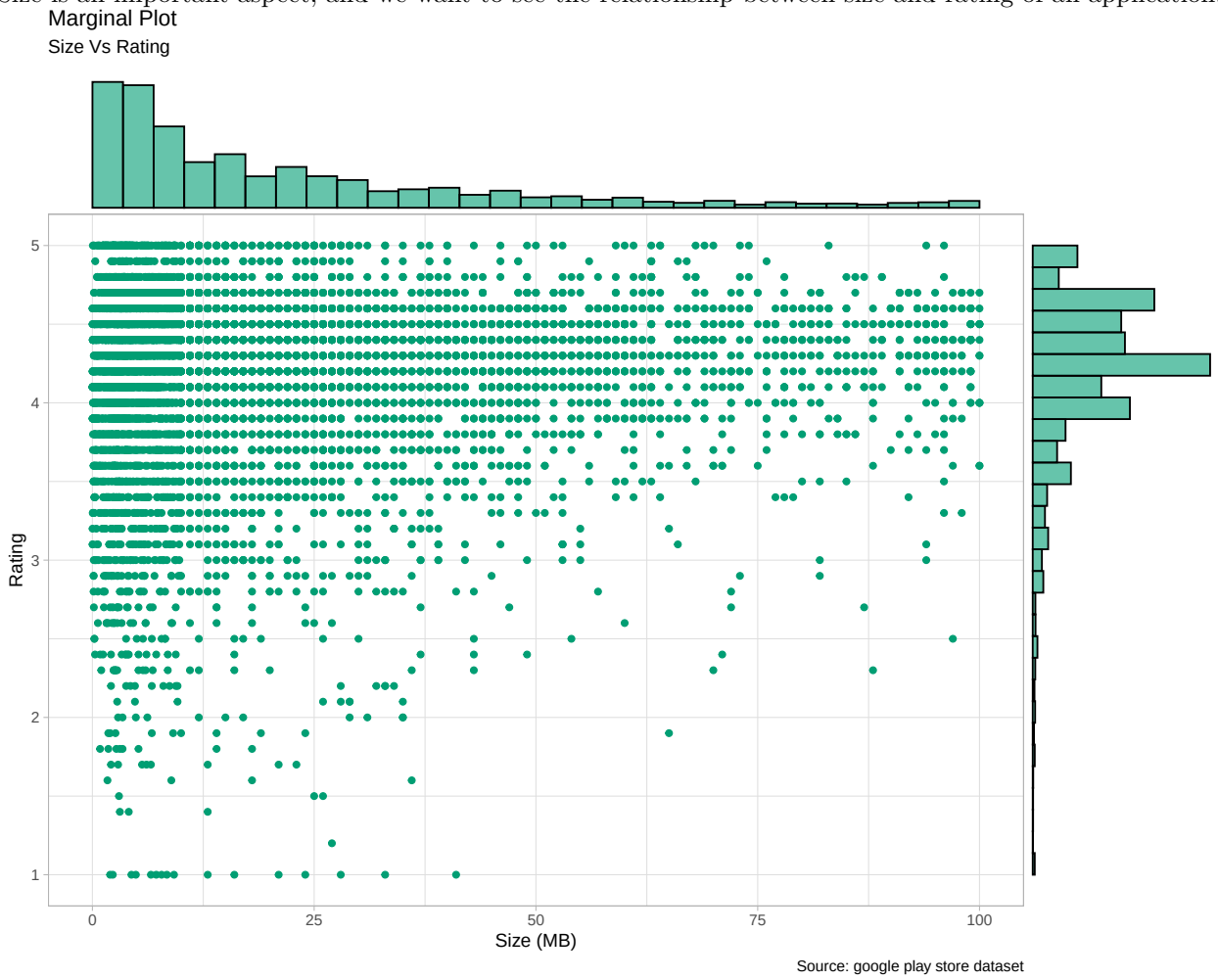
In previous graph we observed that distribution of rating as per category varies significantly. Now we want to find out what the average rating per category is.




**Finding:** This graph shows that the average rating per category is not very different. Still the "EVENTS" category has the highest average rating, and "DATING" category has the least average rating.

### 4.2.10 Size and rating

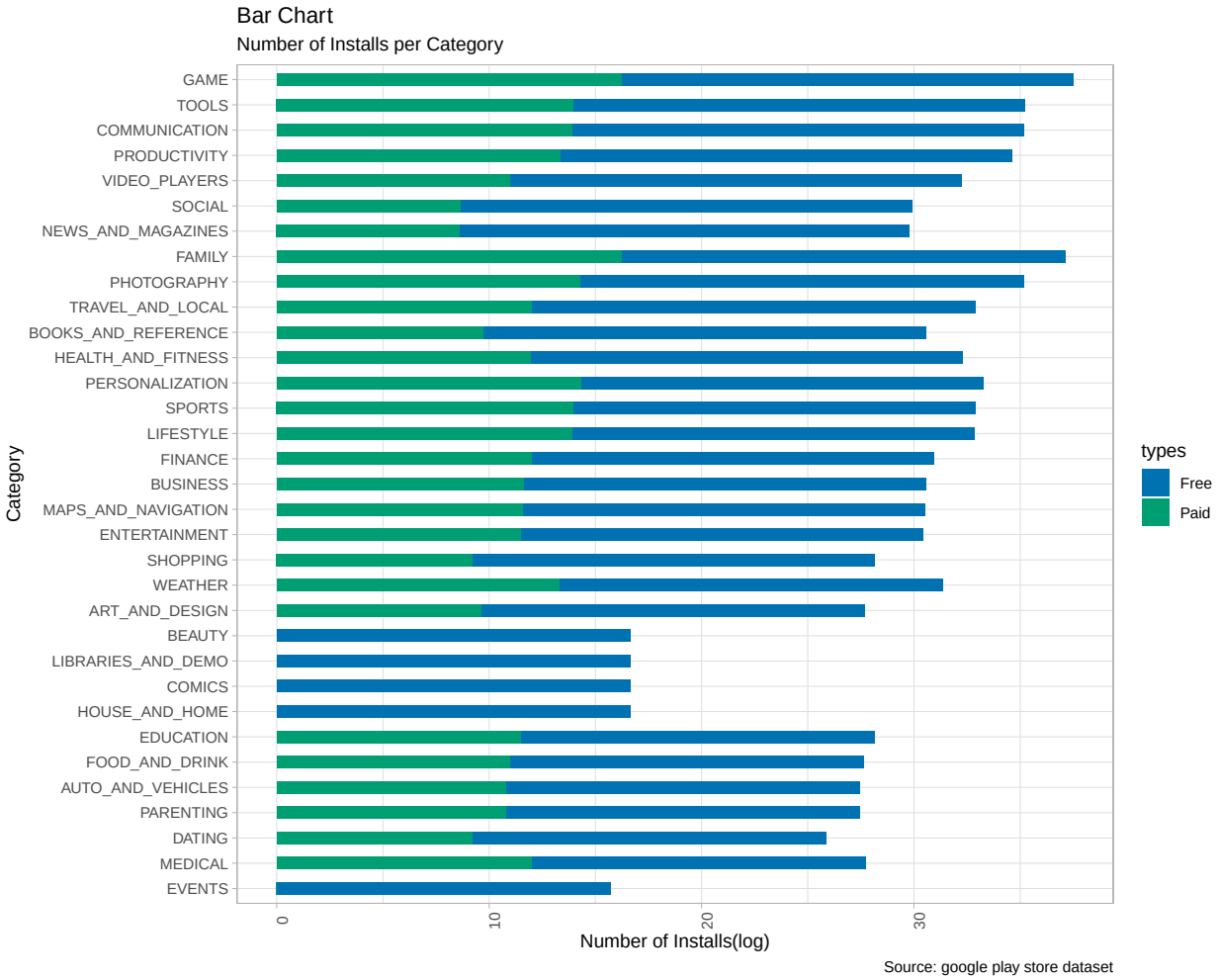
Size is an important aspect, and we want to see the relationship between size and rating of an application.




**Finding:** We can see that majority of applications with their sizes under 25 MB, have a good rating(4).

## 4.3 Installs column in depth

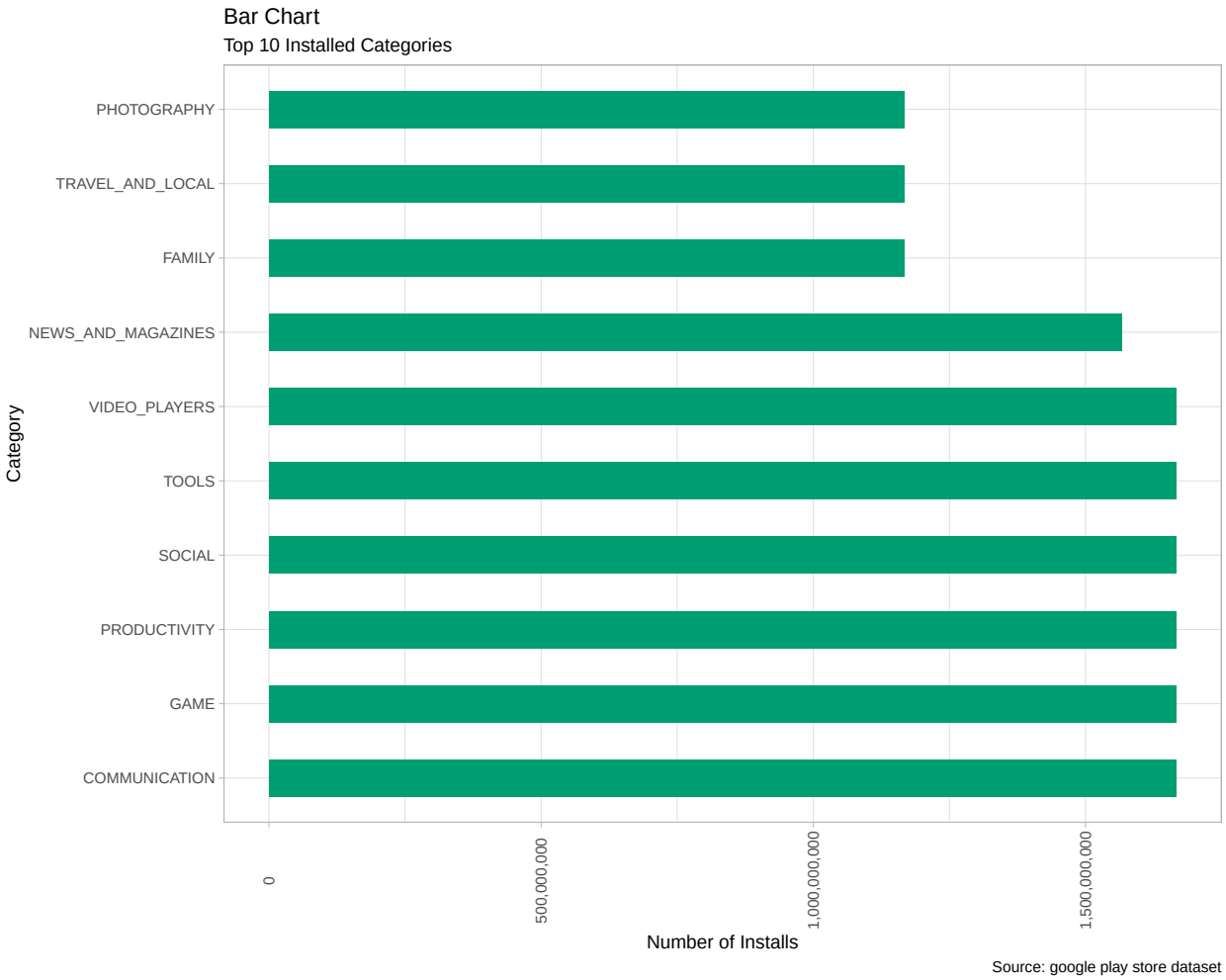
### 4.3.1 Number of installs per category




**Finding:** The graph shows the log of number of installs (the values of installs varied from 0 to 1 billion) vs CATEGORY. FAMILY and GAME has the highest number of installs. EVENTS, HOUSE_AND_HOME, COMICS, LIBRARIES_AND_DESIGN and BEAUTY have the least number of installs.

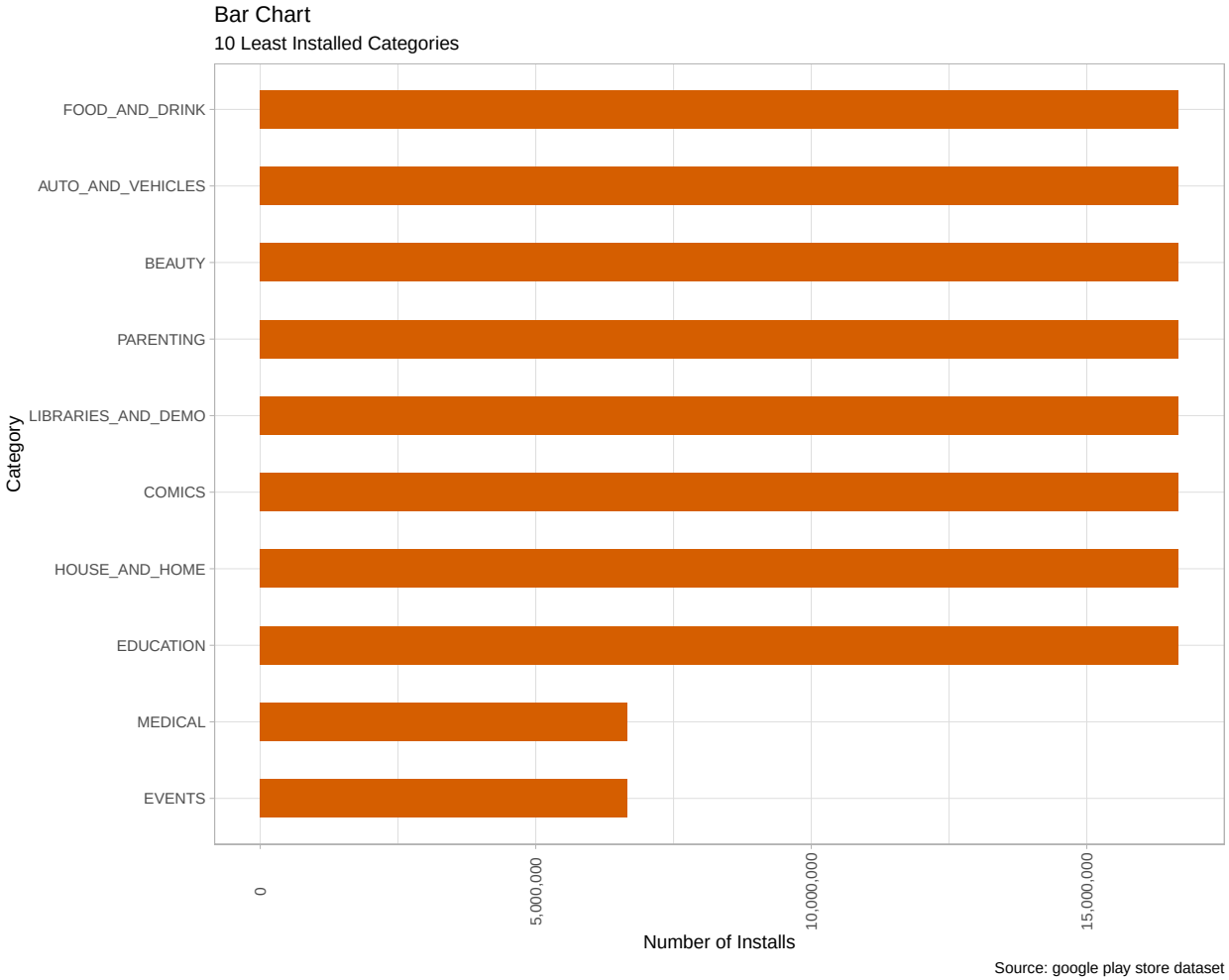
### 4.3.2 Top 10 installed categories

Top 10 categories with greatest number of installs.




**Finding:** COMMUNICATION has the highest number of installs.

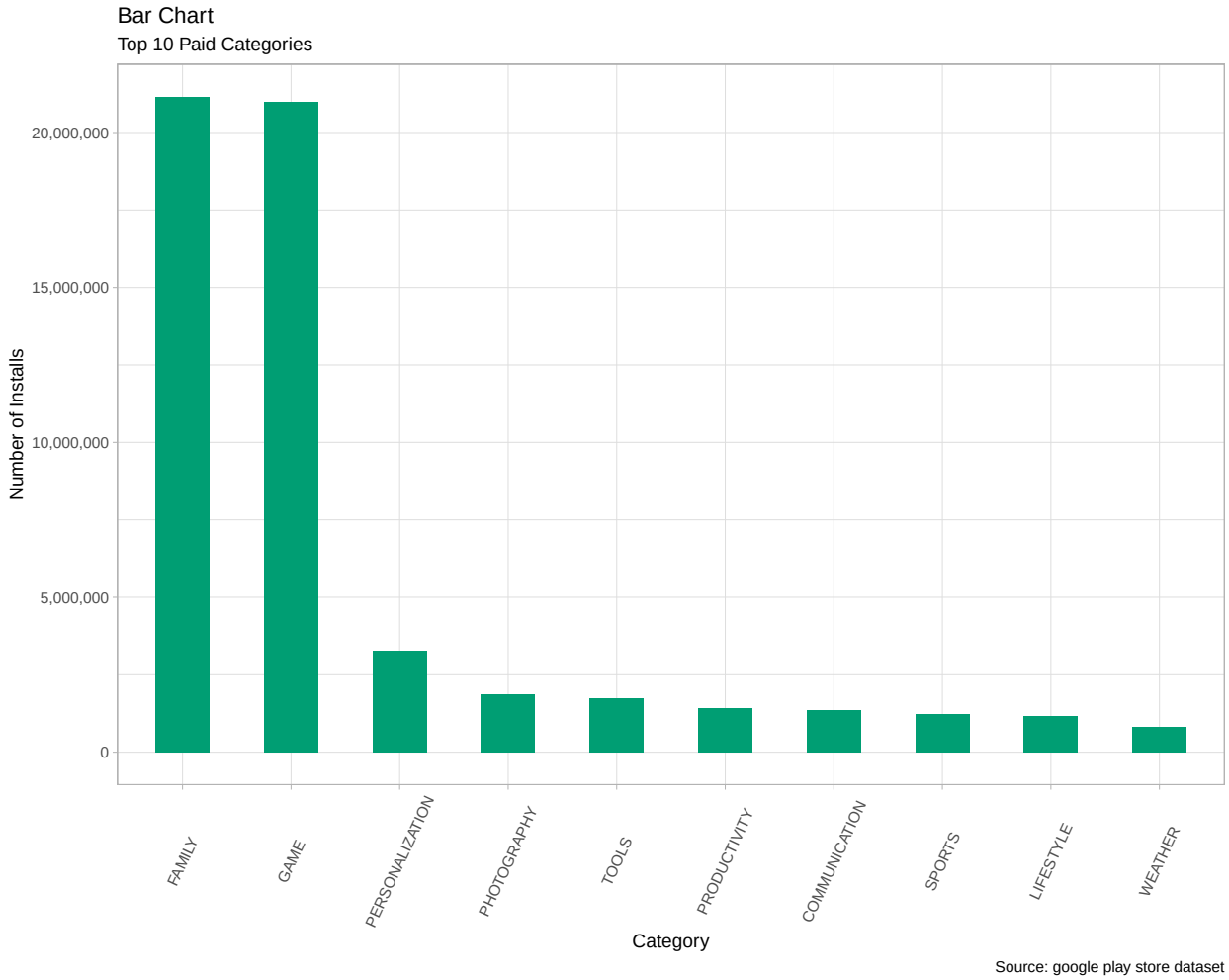
### 4.3.3 10 least installed categories




**Finding:** Events has the least number of installed applications.

### 4.3.4 Top 10 paid Categories

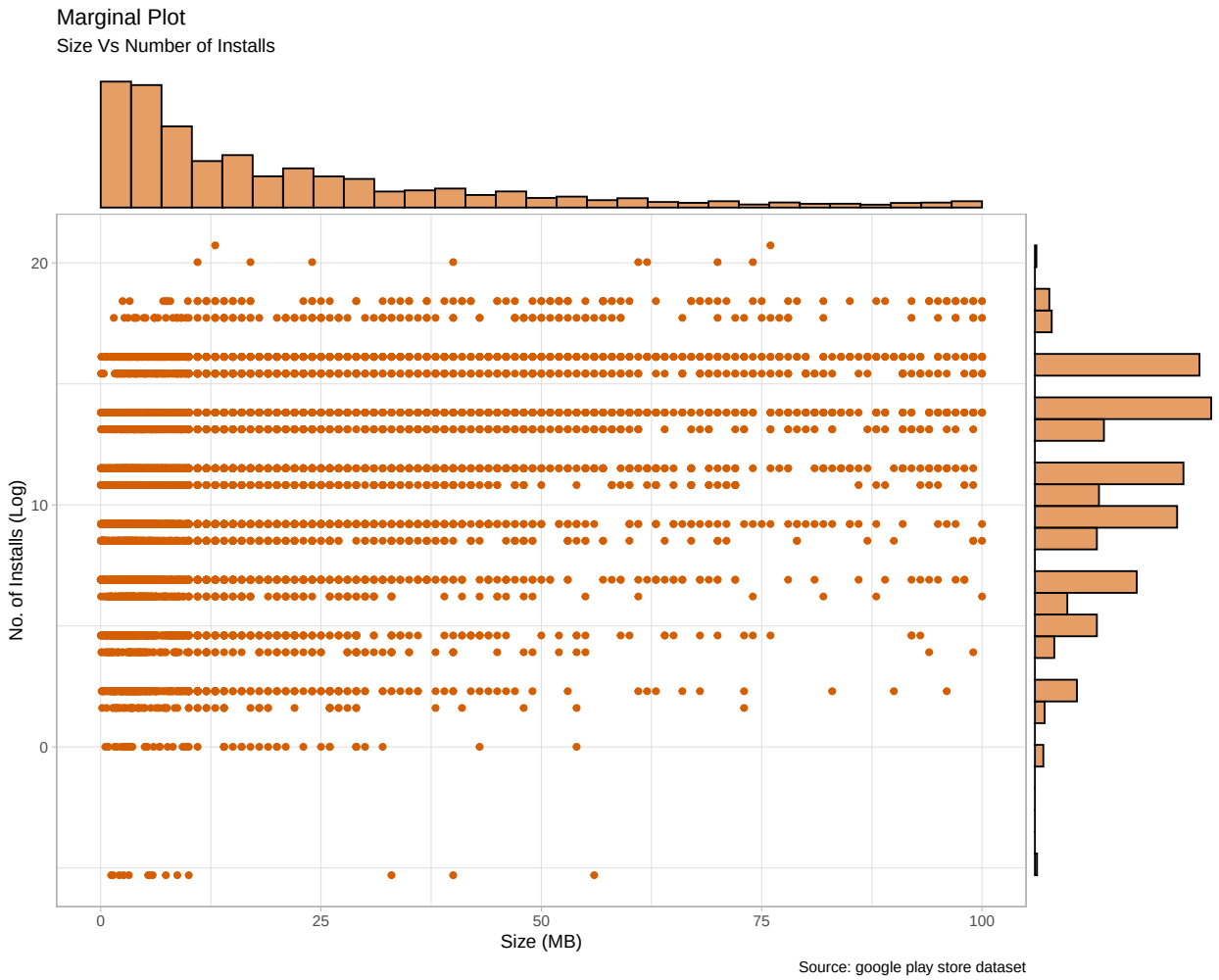
Top 10 categories with the highest number of installs for paid applications.




**Finding:** FAMILY and GAME has the greatest number of installs.

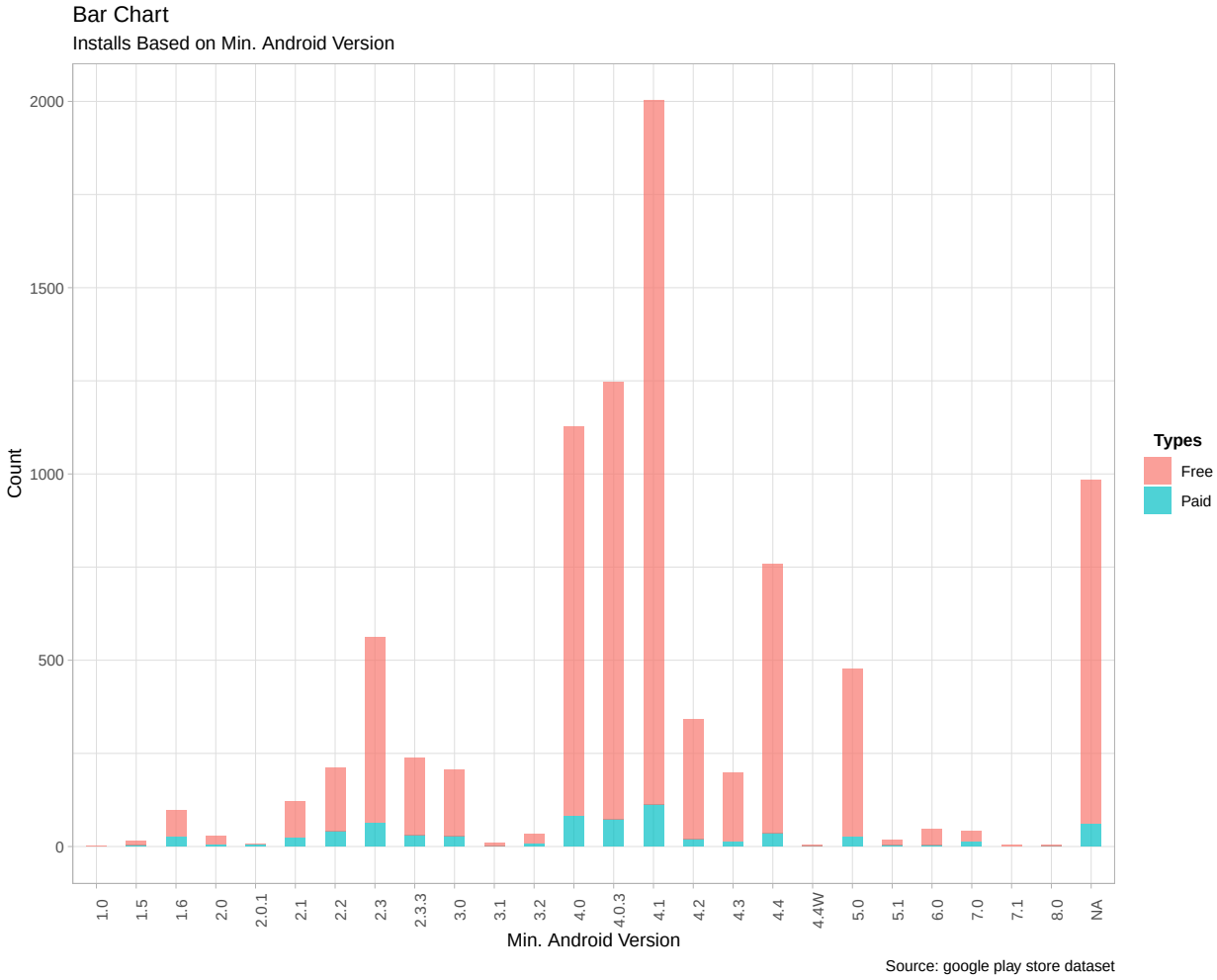
### 4.3.5 Size Vs. number of installs

Size is an important characteristic of an application. Large applications might reduce the number of installs, as it reduces to targeted audience. We can test this by plotting size against installs.




**Finding:** Optimally sized Applications, with sizes between 5MB and 30MB, gets the greatest number of installs.

### 4.3.6 Number of installs based on support by minimum android version




**Finding:** 4.1 android version has the maximum number of installs.

#Models

#Conclusion

# 5 References

[1] Dataset;
[2] R markdown ;
[3] Stackoverflow ;
[4] GGally;
[5] Custom color pallete;
[6] Tidying data;
[7] Colors;
[8] TidyVerse;
[9] ggplot2;
[10] Google playstore Kernel by Danilodiogo;
[11] market share ;
[12] Global game jam ;
[13] Play Store Statistics ;