Google Play Store Analysis

Trung Quan Lai

# Abstract

Extracting information from data has become a remarkable demand in modern society. Affordable devices, as well as technology development, increasingly helps to collect data from anywhere more accessible than ever. However, the use of information and how to turn it into a valuable form is still at the exploring stage, which combines the number of knowledge and appropriate processes. In this paper, Google Play Store dataset is used to gain useful information which supports making the right decisions in business and also figure out the market trend. Different analysis techniques such as univariate and bivariate analysis are applied based on particular data from the dataset. Finally, coming up with results as well as advantages and limitation which need to improve for further experiments.

# Introduction

## Research background

Data analytics demand is on the rise, and soon, it would become an integral part of every other organization. People are living in the age of information and internet has made it easy for anyone to gather whatever information. Even learning and building a plane or a car by using online materials that could be afforded by higher education in the past. Nowadays, it is available for everyone who could get an internet connection and how people are going to use this tremendous amount of data to gain benefits. Data analytics is also a combination of processes to extract information from datasets along with domain knowledge, programming, mathematics and analytical skills to arrive at a decision-making process or reach a conclusion with the help of data.

## Data analysis topic

In this paper, I am going to focus on Google Play Store Apps dataset for in-depth analytic of the Android application market which is an inspiration to app-making business decisions. Based on the analytic, it could help to indicate which option has a better chance of being trending overtime upward as well as predicting market movements and learning new tactics. Besides that, the data can also be valuable information to identify new topic ideas to develop a product which is the kind of mass consciousness of customers.

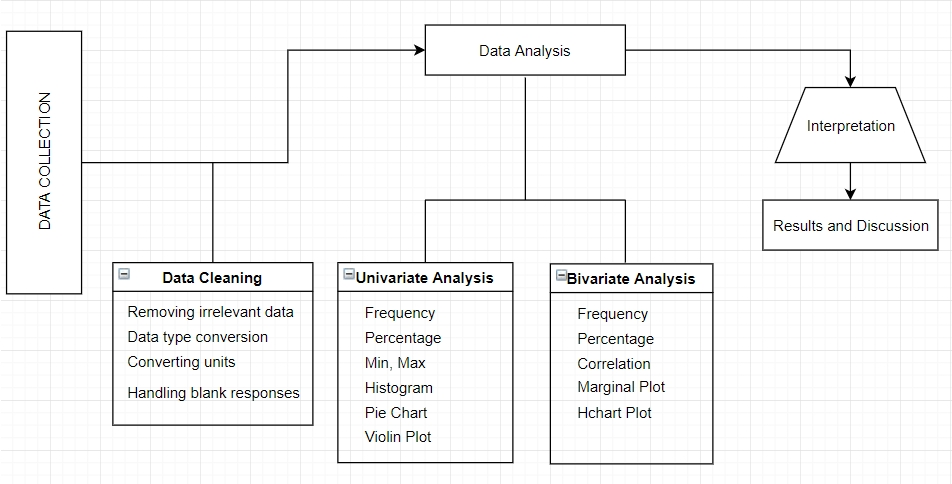
## Literature review on related data analysis methods

Univariate analysis: it is the first type of analysis that can do on single variable as the name indicates and also known as descriptive statistics which focus on a single attribute at a time. There are plenty of ways to perform univariate analysis such as quantifying univariate analysis following by Min/Max, Average, Median, Mode, Varian, Standard Deviation and so on and it is usually represented by a bar char or pie chart.

Bivariate analysis: it is an analysis of the relationship between two variables which usually represent as X and Y. Generally, in bivariate or two variables analysis, it tries to find out how to do changes in X effect some Y variables such as using X variable to predict the Y. The X also has plenty of different names such as independent variable, explanatory, predictor or the covariate and the Y is the dependent variable, the outcome or response.

# Methodology

## Flowchart of the data analysis process



## Explanation of each module in data analysis

Data collection: it is a very first step in data analysis which requires reliable sources so that the final result could be used for other experiments such as Machine Learning. This dataset was collected from the Google Play Store itself by using Python script to trigger a searching algorithm and end up with more than 10,000 applications.

Data cleaning: the raw dataset contains messy data on it which does not follow any data type. It needs to be formatted and created categorical variables before stating analyzing. Some techniques can be used in this step such as data type conversion, handling blank or nan values, removing irrelevant data and so on.

Data analysis: In this stage, different variables will be analyzed in order to gain information as much as possible by comparing two or more relevant points from the dataset. Furthermore, visualizing them on a graph which depends on the data itself and purpose of the information such as frequency or percentage, after that using the appropriate chart to visualize them.

Interpretation, results and discussion: throughout the whole process, it should come up with a conclusion that supports the objective of the experiment. It can be a prediction for further investments or show trending on the market. Based on these results, it can analyst the quality of the dataset.

## Difficulties and Solutions

The dataset contains more than 10 classes with lot of information which is difficult to decide which one is better for analysis. To come up with the solution, it needs to switch and test on different variables as well as try to visualize them on the graph. Besides that, data quality is also a big challenge due to the wide range of different sources so that Kaggle is one of the most famous dataset websites for experiments and researches. Furthermore, using different plotting library to visualize the result in order to point out data meaning totally.

# Experiments and Results

## Datasets introduction

The Google Play Store is one of the biggest Android markets which contains a number of applications with a huge range of categories. Analysis and exploration information from the dataset would be enormous potential to predict market trend as well as able to draw a general picture of the Android market in the future to drive business success. The dataset "Google Play Store Apps" is obtained from the Kaggle website [1] which contains more than 9000 applications and 13 features. Each column in the raw dataset contains different valuable information for further analysis, however, it needs to perform some cleaning techniques before it is ready to use.

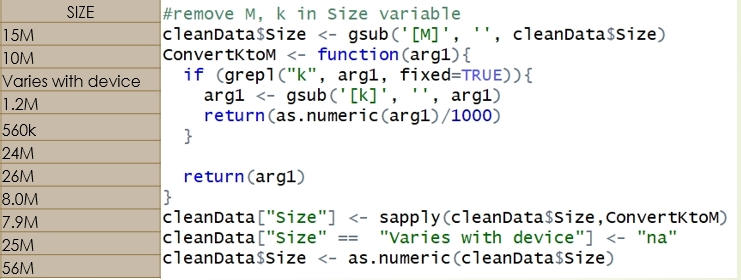
The variables in the dataset:

|  |  |  |
| --- | --- | --- |
| No | Variables | Meaning |
| 1 | App | Application name |
| 2 | Category | The category which the application belongs to |
| 3 | Rating | User rating on the app |
| 4 | Reviews | Number of user reviews on the app |
| 5 | Size | Installation size |
| 6 | Installs | Number of people has installed/downloaded the app |
| 7 | Type | Type Free or Paid |
| 8 | Price | Price of the app |
| 9 | Content Rating | Targeting group of the app such as children, adult, mature etc. |
| 10 | Genres | Subcategory of the app from the main category |
| 11 | Last Updated | Last time the app was updated |
| 12 | Current Ver | Current version of the app on Google Play Store |
| 13 | Android Ver | Minimum requirement of Android version |

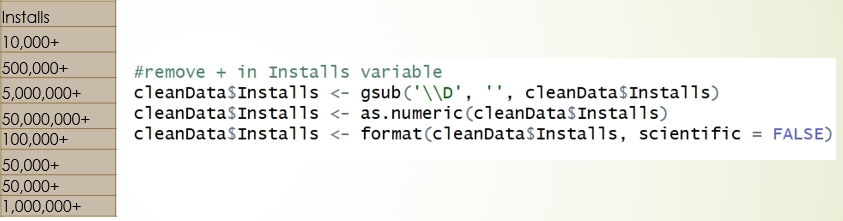
## Experiment setup

Data cleaning has been the most important part of data analysis because this is the very first step after data collection. It needs to be done right at the first time before going to do experiment because it decides the quality of the final result. In this dataset, some issues need to be addressed:

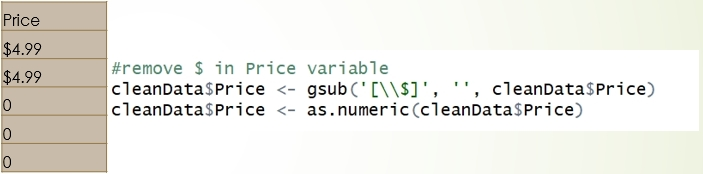
* Size: having a quick look at column size, it contains number mixing with M and K which means Millions and thousands. These mixing numbers would be converted to numeric by removing M and K at the end before turning the number to a numeric type. The M number would be kept as it is and K number would be divided to 1000. Besides that, some value has “Varies with device” string on it that I decide to convert to nan for further analysis.



* Installs: there are 2 different types in this column “1,000+”, “Free”. With the first type, the “+” sign would be removed and turn it to numeric and “Free” string would be converted to nan.



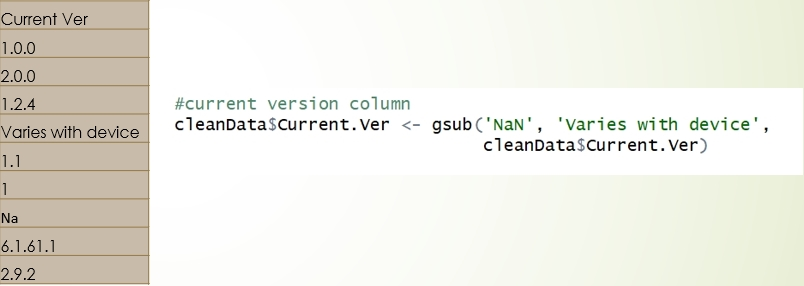
* Price: data value is a $ sign with a number “$3.5”, with this kind of data the $ sign will be removed and convert a value to numeric.

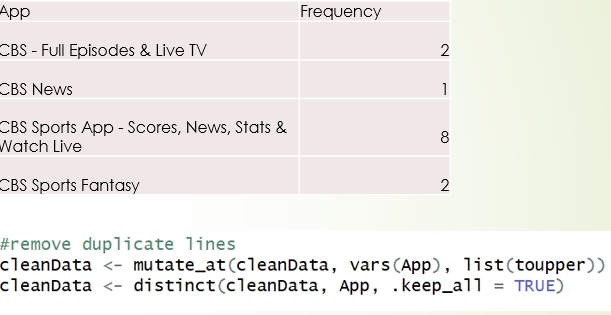


* Last update: changing the format to data time format.



* Current Version: having 3 types of data, they are Number, “Varies with device” string and “nan”. The number would be converted into numeric type and “nan” would be converted to “Varies with device” due to its low frequency, just 8 values.

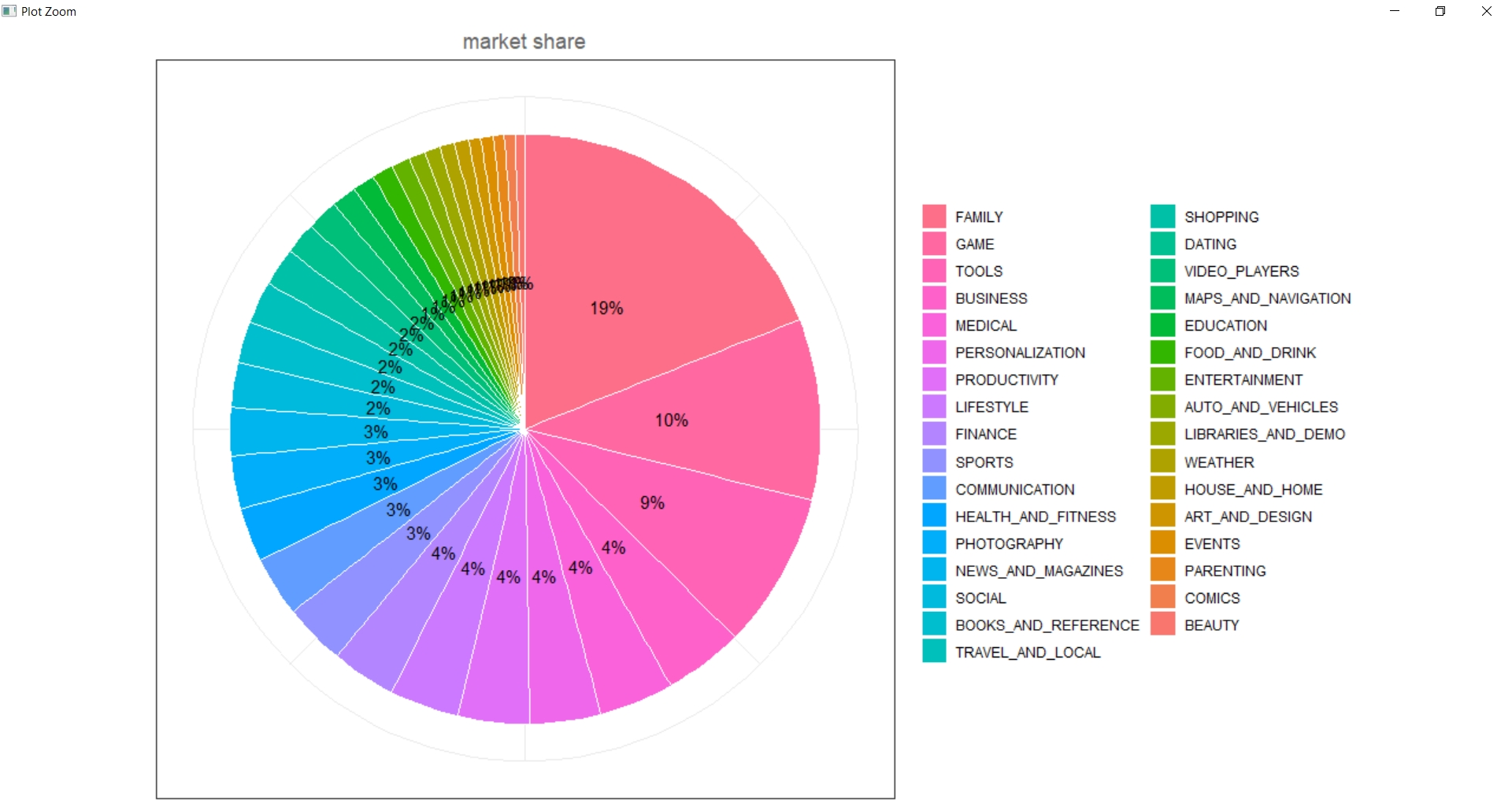


* Duplicated records: some records are duplicated in different forms such as Uppercase, Lowercase and Uppercase first character in the word. Solution for this issue, all records are going to change to Uppercase then doing distinct function.
* 

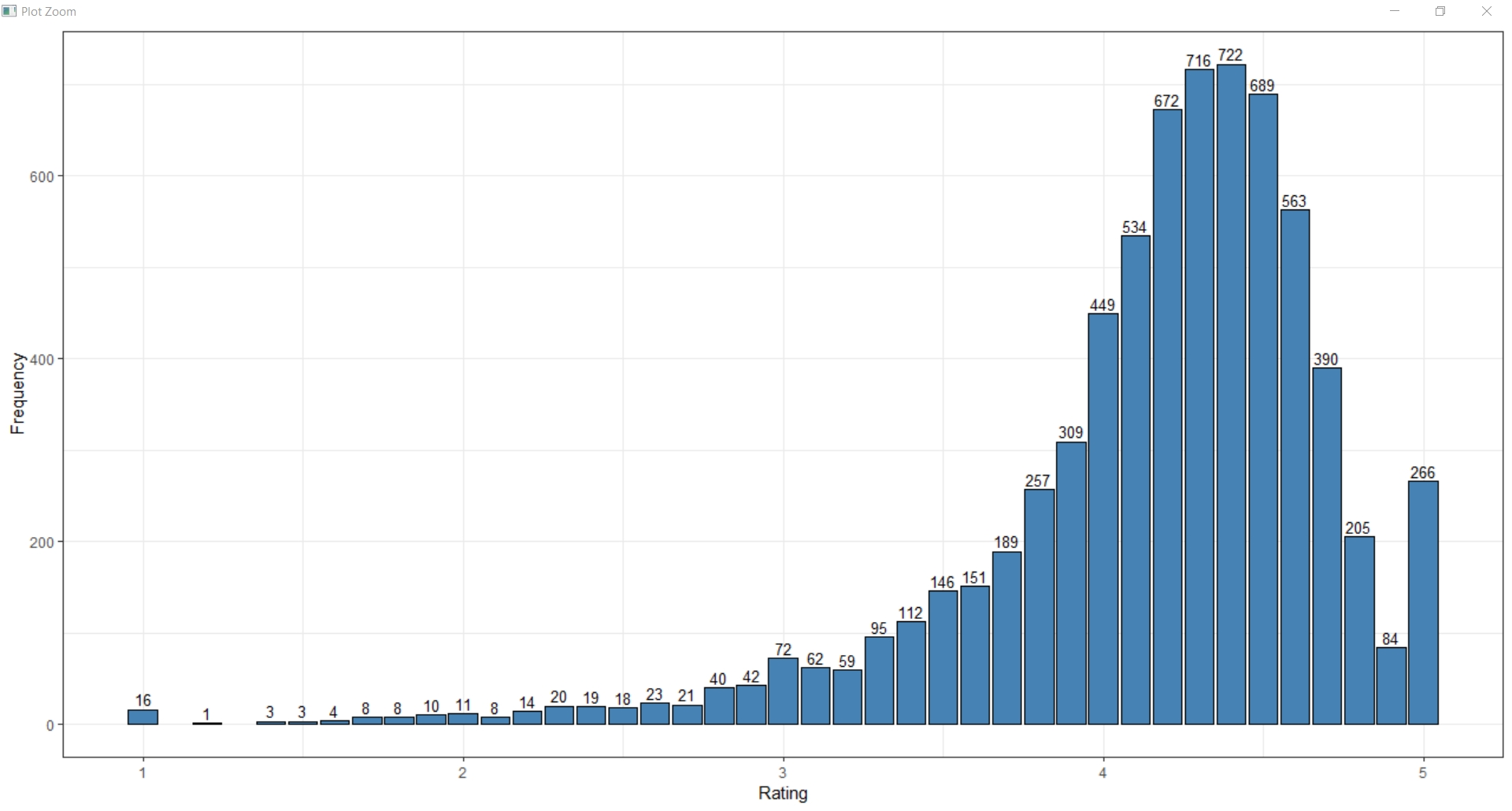
## The procedure of experimentation

In this section, I am going to extract information from the dataset based on Univariate and Bivariate analysis. Some important information would be focused on the experiment.

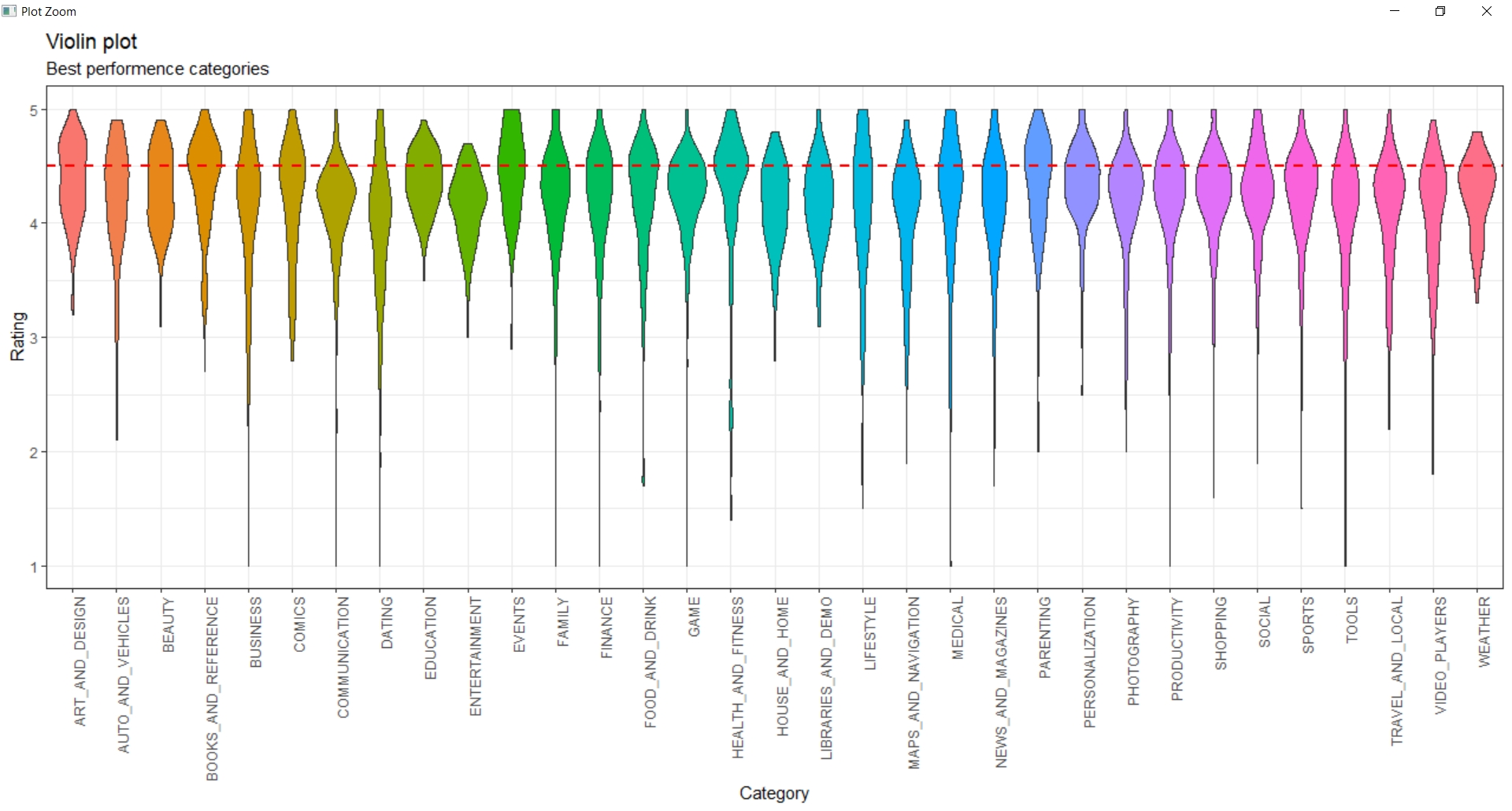
1. Breaking down the Android market by Category: there is three most significant type of the market are FAMILY, GAME and TOOLS which are 19%, 10% and 9% respectively. Other parts share the percentage from 4% to lower than 1%.



1. Rating overall applications: With the rating average is 4.1732 score, this figure shows a good performance across the market. Let plot the performance by categories to see which one has a good rating.

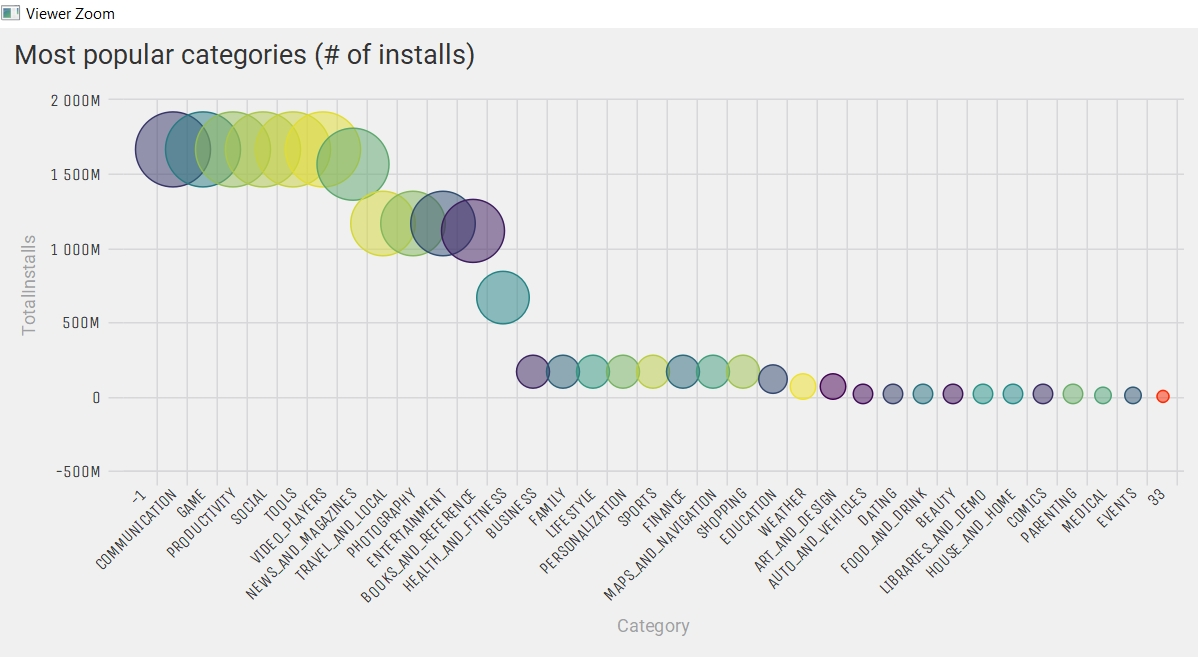


1. Best performance by Categories: Based on the Violin Plot, BOOK\_AND\_REFERENCE and HEALTH\_AND\_FITNESS have the highest performance which more than 50 % has greater than 4.5 ratings. In the other side, COMMUNICATION, DATING and ENTERTAINMENT have more than 50% lower than the average.

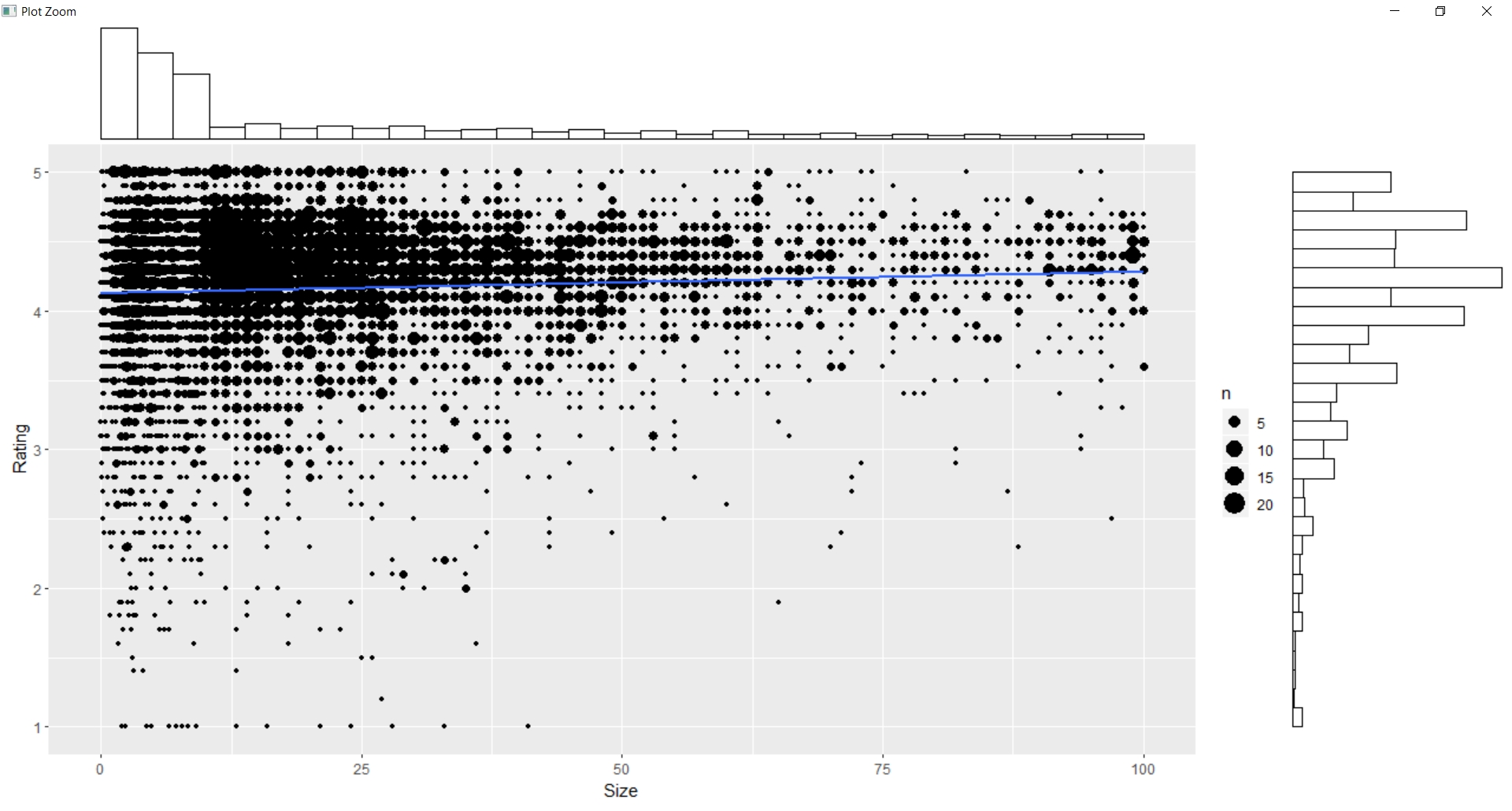


1. Most popular categories by a number of installations: The most significant number of installations are belonging to COMMUNICATION, TOOLS, GAMES, PRODUCTIVITY and SOCIAL which are more than 1,500 million downloaded. On the other hand, the lowest number belongs to

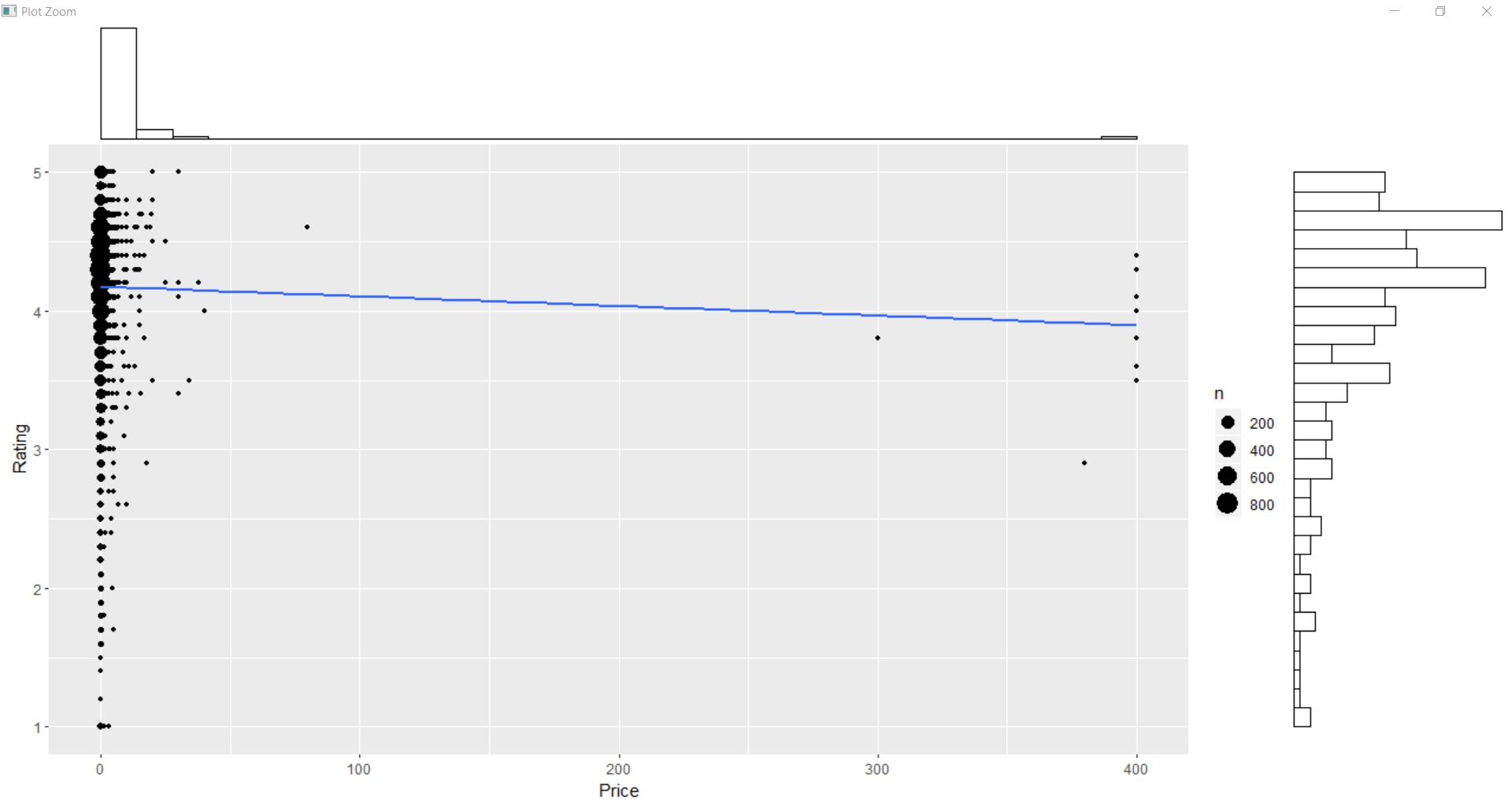
COMICS, EVENT and PARENTING.



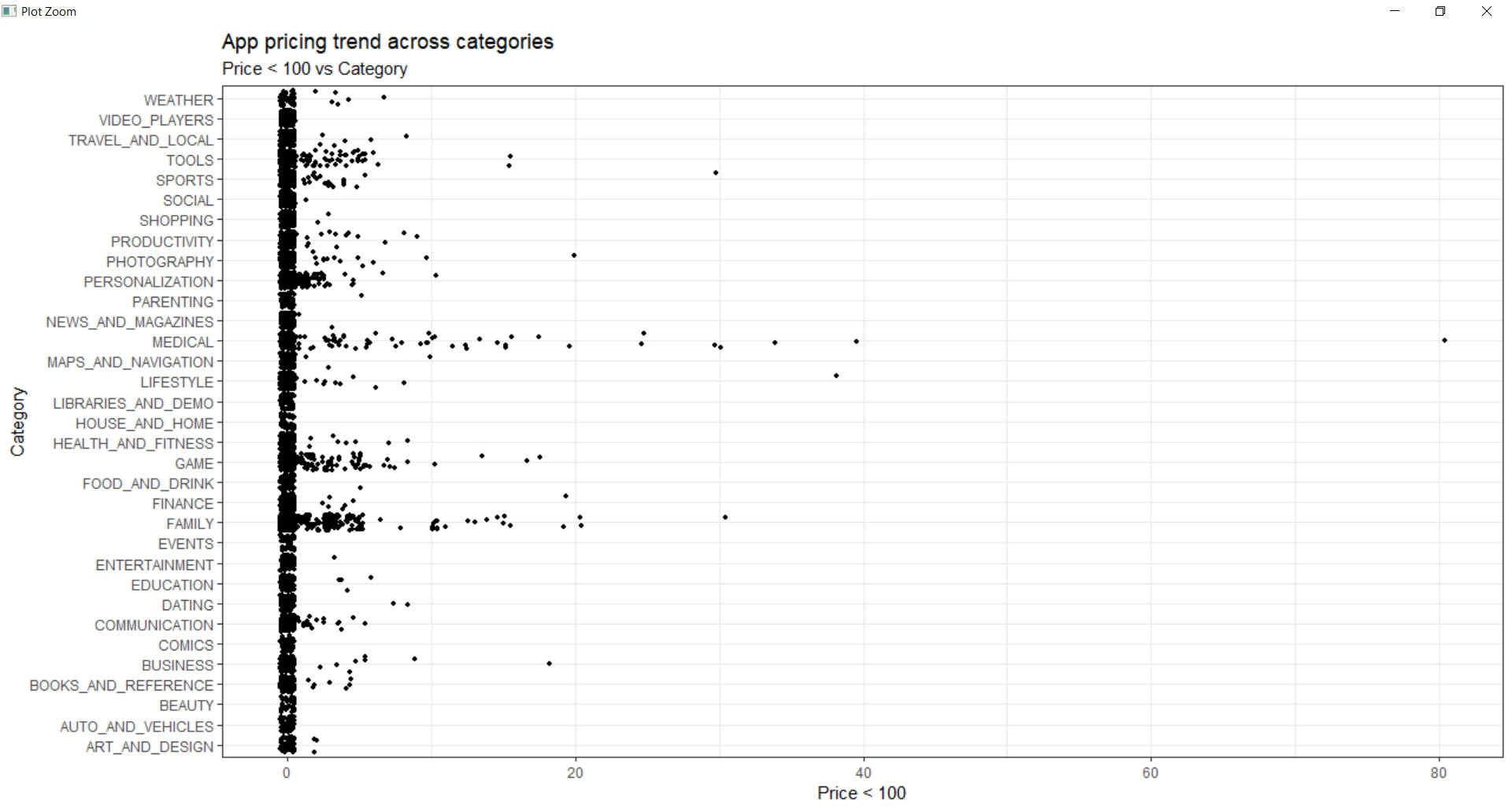
1. Correlation between Size and Rating: The size of application effects the rating apparently. Almost 4.5 rating applications has the size between 2MB to 40MB, the rating drops rapidly while the size increases.



1. Correlation between Rating and Price: The price is between 2 dollars to 25 dollars having good ratings which is more than 4 stars. Some applications are sold more than 300 dollars which is incredible and also have low performance.



1. App pricing trend across categories: apparently, the most expensive applications are MEDICAL and FAMILY while others have below 25 dollars which are acceptable.



1. Percentage between Free and Paid app over the market: clearly, the proportion of free applications fell into HOUSE\_HOME, LIBRARIES\_AND\_DEMO, EVENTS, COMICS, BEAUTY and AUTO\_AND\_VEHICLES categories which is 100 % of free applications. Others share between free and paid, MEDICAL and PERSONALIZATION has the highest percentage with 20% and 23% respectively.



## Result interpretation

Rating average is 4.17, it shows that customers have a great experience when using the application, especially BOOK\_AND\_REFERENCE and HEALTH\_AND\_FITNESS categories. They also prefer the applications with lightweight in size between 2MB and 40MB. Besides that, Price is another factor for choosing an application which makes customers happy with free of charge or at least in the acceptable range less than 25 dollars. The most expensive application fell into the category related to MEDICAL and FAMILY which prove that customers are willing to pay higher than the standard cost for their health and loved ones.

# Conclusion

## Advantages and limitations

Google App Store is one of the most popular markets for Android mobile operating system. With more than 10,000 application across the store, it contributes brief information about the current stage as well as supporting decision making to business values. There are plenty of variables and features which are not limited in this project for doing experiments. They can work well together in order to provide valuable information on any purposes. However, nan value needs to be addressed in the future work at the first stage of the collection process, it decides the quality of dataset and also reduces the loss of information which might lead to an inappropriate conclusion.

## Direction for future improvement

Machine Learning is an interesting field which can use this dataset to predict future work such as rating or application performance. Combination of some features in the dataset and using Linear Regression model to fit the data. It can come up with a nice fit or even show up a general trend which helps people to prevent mistakes in the design stage. By doing this, more features and examples are two essential things which need to keep improving and contributing to the dataset.