**Vietnamese National University HCMC**

**International University**

**School of Computer Science and Engineering**



PROJECT REPORT

**Topic: Google Play Store Apps**

**Assoc. Prof. Loan Nguyen, PhD**

**Course: Data Analysis**

**Group members:**

Trần Đình An ITDSIU20113

Nguyễn Sỹ Nguyên Ngọc ITDSIU20091

**Table of Contents**

**[INTRODUCTION 1](#INTRODUCTION)**

**[Motivation 1](#_Toc91764964)**

**[About the dataset 1](#_Toc91764965)**

**[Tools 2](#_Toc91764966)**

**[DATA PREPROCESSING 2](#_Toc91764967)**

**[Dataset’s overview 3](#_Toc91764968)**

**[Fill missing values 6](#_Toc91764969)**

**[Remove duplicates 7](#_Toc91764970)**

**[Convert problematic attributes 9](#_Toc91764971)**

**[DESCRIPTIVE ANALYSIS 11](#_Toc91764972)**

**[Distribution 11](#_Toc91764973)**

**[Features of the dataset 18](#_Toc91764974)**

**[DIAGNOSTIC ANALYSIS 27](#_Toc91764975)**

**[Relationship between installs and other attributes 28](#_Toc91764976)**

**[Differences in free and paid apps 29](#_Toc91764977)**

**[PREDICTIVE ANALYSIS 34](#_Toc91764978)**

**[Goal 34](#_Toc91764979)**

**[Implementation 34](#_Toc91764980)**

**[Result 35](#_Toc91764981)**

**[CONCLUSION 36](#_Toc91764982)**

**[REFERENCE 38](#_Toc91764983)**

# INTRODUCTION

As time goes by, the world moves forward into the technology era. Since the use of smartphones increases dramatically, the role of app stores has become more and more crucial for smartphone users can use them to their full potential.

Taking that into consideration, we decide to work on this project to give everyone an overview of the Google Play Store - one of the most popular app stores and what businesses/developers can do to improve their apps installation and sales.

## Motivation

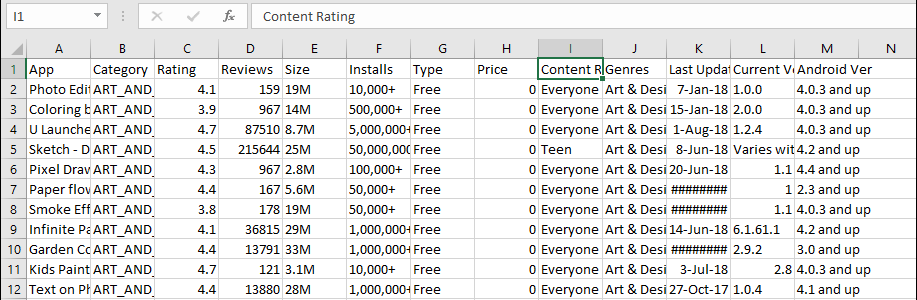
* We will preprocess and analyze the datasets to find features that are correlated to the installation to predict which kind of apps will be downloaded more over time.
* Our main objective is to analyze the sizes, genres, prices, … columns to see which apps in different categories customers prefer and tend to buy/download.
* Base on the characteristics of some top-installing apps, we can also help businesses/developers to improve their customer-approach solutions.

## About the dataset

The “Google Play Store Apps” dataset which is available on Kaggle.com, contains names of different apps, their category, their rating, etc.… (With 13 columns and 10,000+ rows)

**Figure 1**

*The raw dataset*



*Note.* By Lavanya, 2019

To analyze the dataset, we will use three different kinds of Analytics:

* Descriptive Analysis: we will find basic information about apps in different categories (Ex: In a category, which kind of age group prefer these apps the most? etc.…), we will also highlight potential relationships between categories (Ex: What do these top-installing apps in this category and these top-installing apps in that category have in common? etc.…)
* Diagnostic Analysis: we will translate the dataset into visualizations to uncover the reasons for some problems (Ex: Why do these apps in this category have a low installation? etc.…)
* Predictive Analysis: from the result of the descriptive analysis, we can make predictions about the future sale of apps in different categories; at the same time, we can also show businesses/developers what they can do to increase their sales and improve their customer’s satisfaction.

## Tools

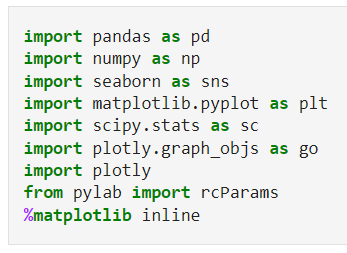
* Python: commonly used for data analysis, statistical computing, and organizing data.
* RapidMiner Studio: it is a software that allows data mining, text mining, and predictive analytics.

# DATA PREPROCESSING

In this part, we will use Python and the following libraries to clean and rearrange the dataset:

**Figure 2**

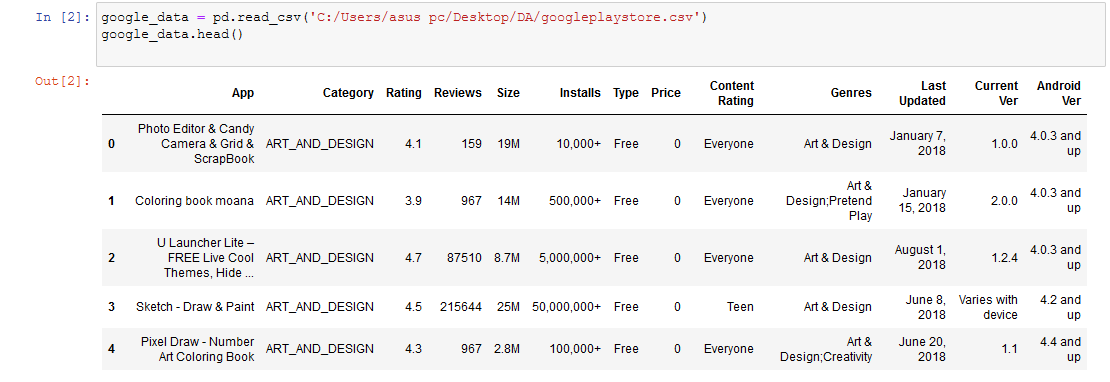
*Imported libraries*



After importing the libraries, the dataset will be imported and put inside a data frame, then we will use the function *“.head()”* to test if our data is the right type:

**Figure 3**

*The dataset*

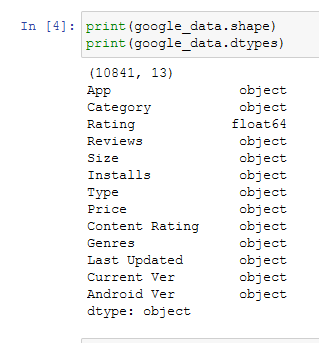


## Dataset’s overview

We use the *“.shape”* and *“.dtype”* function to see how many rows and columns are there in the dataset and type of each attribute, then print the result on the screen:

**Figure 4**

*Dataset’s overview*

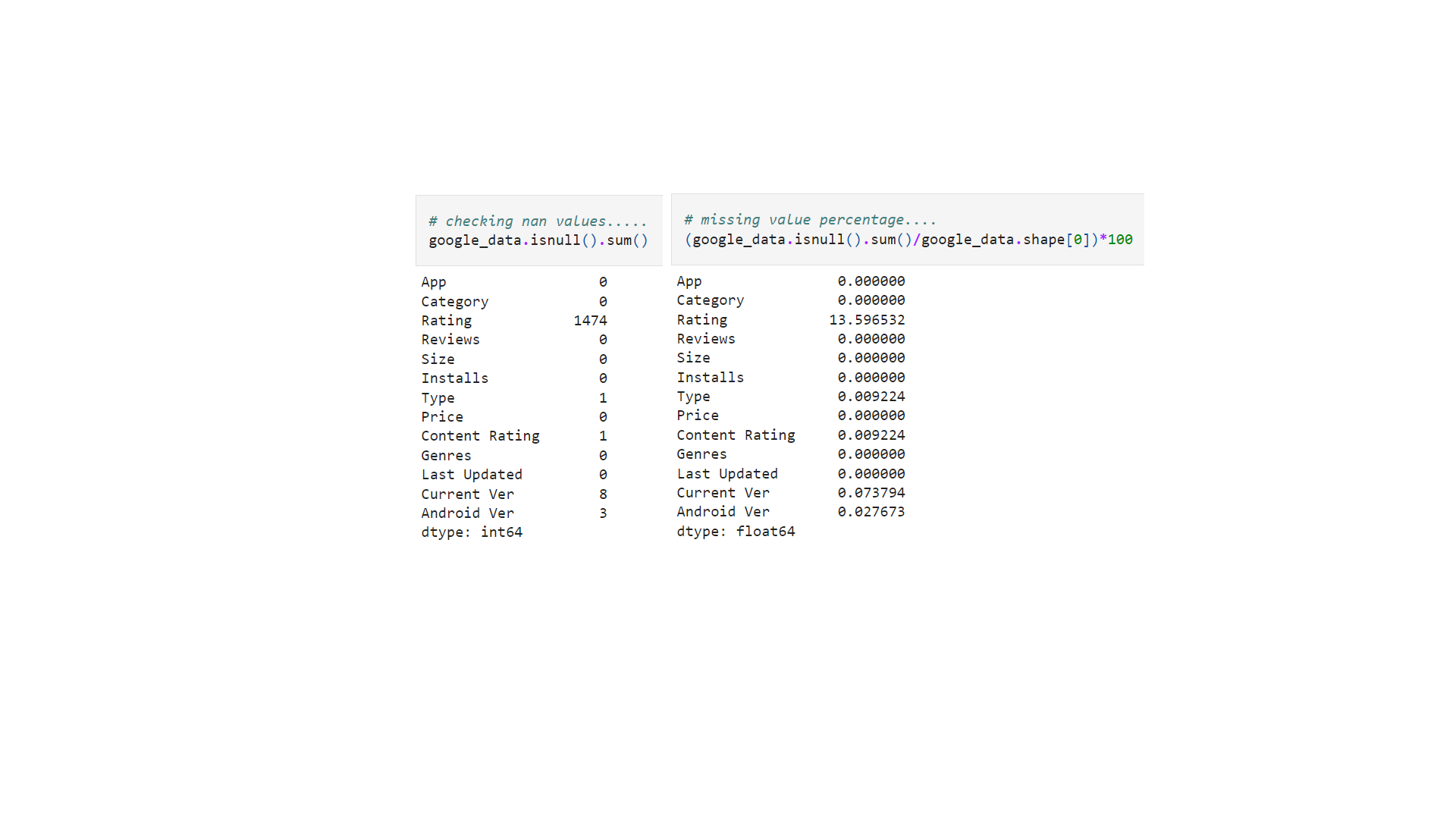


Only the *“Rating”* attribute has the *“float64”* data type, which means the values in this attribute are floating-point numbers. While all of the other's attribute has the *“object”* data type, implies that the remaining values are text or mixed numeric and non-numeric.

The first thing to deal with when cleaning a raw dataset is missing values, we can check the number of missing values and its percentage of the whole dataset with the *“****.****isnull()****.****sum()”* function:

**Figure 5**

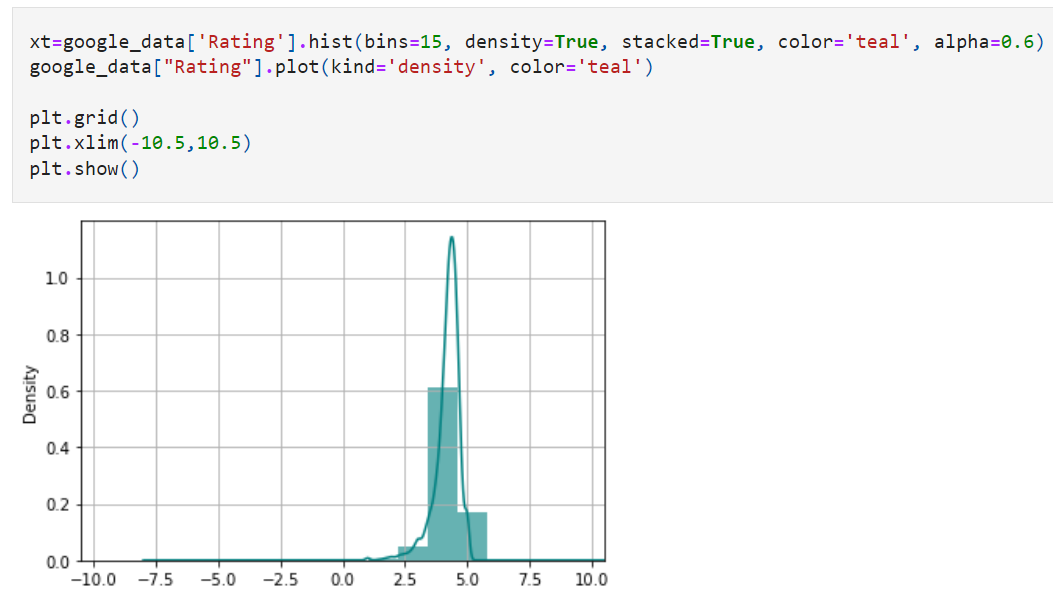
*The number and the percentage of missing value*



Most of the missing values fall into the *“Rating”* attribute, we will draw a histogram to find how the values in this attribute are distributed.

**Figure 6**

*Histogram of “Rating” column*

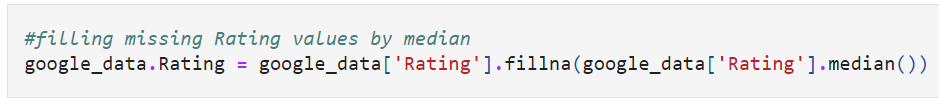


The histogram of the *“Rating”* attribute is left-skewed, hence we will replace the missing values in *“Rating”* with the median.

## Fill missing values

**Figure 7**

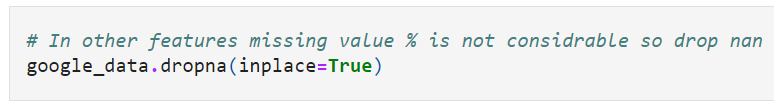
*Code for filling missing value*



Although the *“Rating”* column has a high percentage of missing values, the others are not. Therefore, we will remove them with the *“.dropna”* function:

**Figure 8**

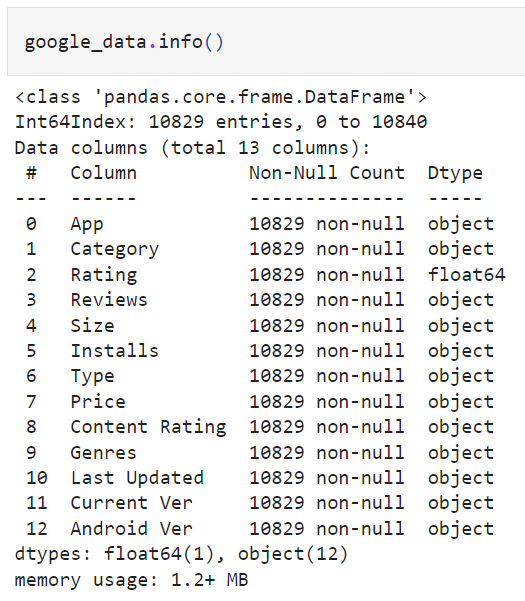
*Code for dropping missing value*



We will check if there are any more missing values with the *“.info()”* function:

**Figure 9**

*The dataset after being treated with missing values*

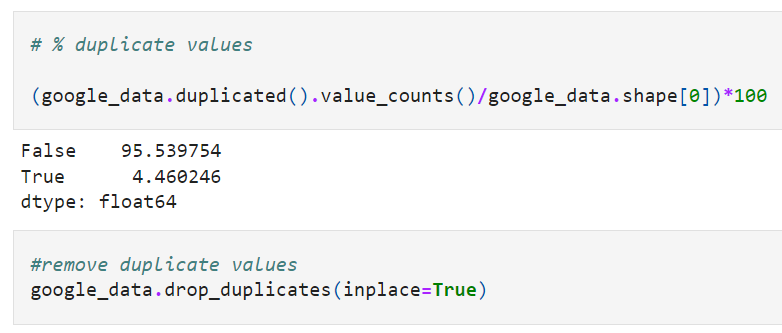


## Remove duplicates

After dealing with the missing values, we will find how many values are duplicated that can affect our analysis and remove them with the *“.drop\_duplicates”* function:

**Figure 10**

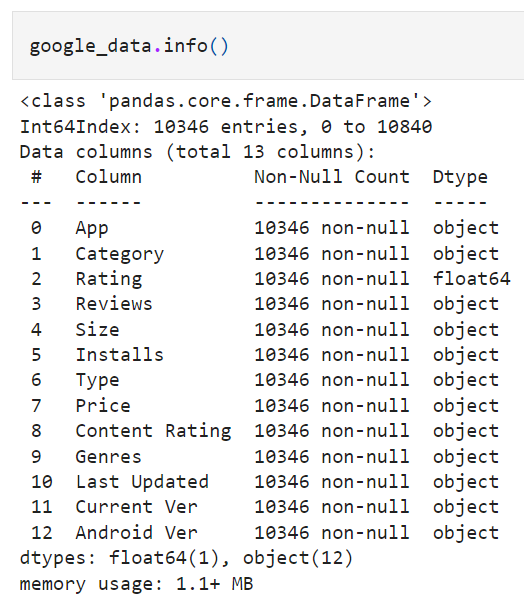
*The percentage of duplicated values*



Since *~4.4%* of the values are duplicated, the same amount will also be removed from the whole dataset:

**Figure 11**

*Dataset after dropping duplicated values*



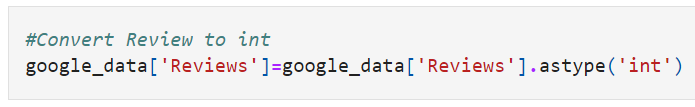
## Convert problematic attributes

Noticed that some attributes need to be changed its type to numeric, those attributes are *“Reviews”*, *“Price”*, *“Size”* and *“Installs”.* All of these attributes are being classified as *“object”* types at the moment.

Firstly, we will change the type of the *“Review”* attribute to integer with the *“.astype”* function:

**Figure 12**

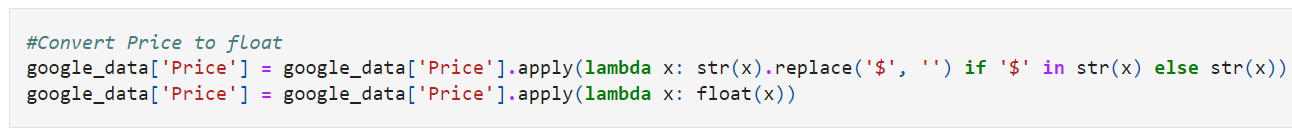
*Code for converting “Review” to int*



Secondly, for the *“Price”* attribute, at the beginning of each value greater than 0, there is a symbol *“$”.* We will remove this symbol and change the attribute type to *“float”*

**Figure 13**

*Code for converting “Price” to float*

**

Thirdly, the values in the “*Size”* attribute do not have the same unit. Therefore, we will change Mb and kb into byte and “Varies with device” into 0 with the *“.str.replace”* and *“.astype”* functions:

**Figure 14**

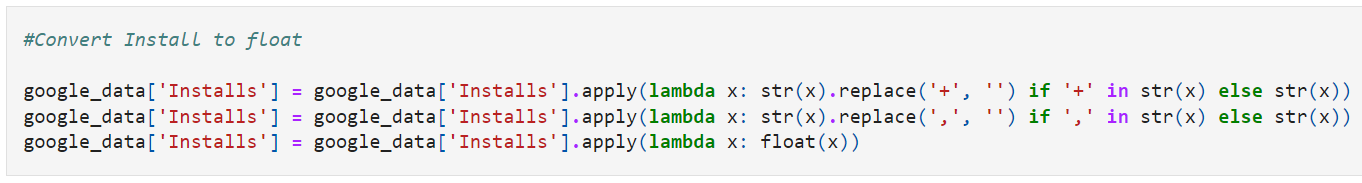
*Code for converting “Size”*



And lastly, the *“Installs”* attribute have commas between numeric character, *“+”* at the end of each value and the column type is *“object”.* We will remove those extra characters and change the attribute type to *“float”*:

**Figure 15**

*Code for converting “Install” to float*



Up to now, the problems in the dataset are solved and we are ready to move to the next part, which is *“Descriptive Analysis”*!

# DESCRIPTIVE ANALYSIS

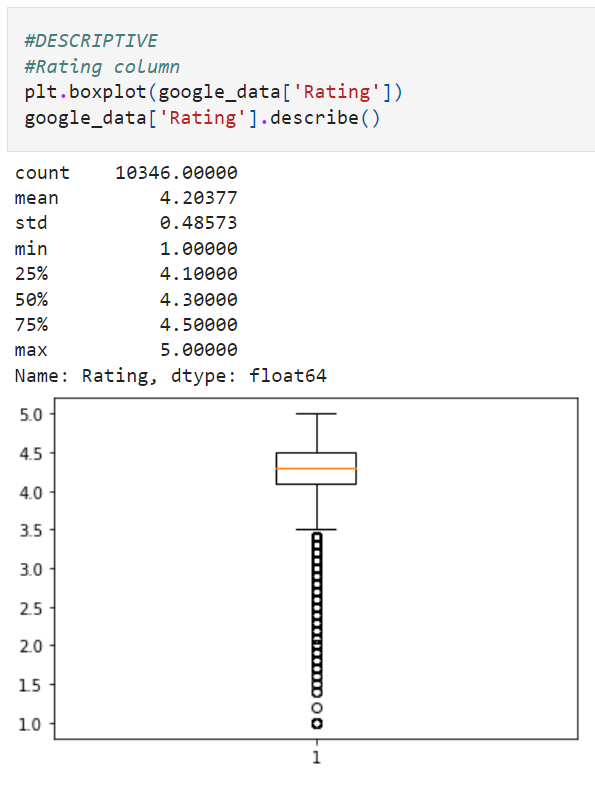
In this section, we will get to know our dataset more by finding the essential characteristics of it. The dataset will be dissected in order to determine what had happened in this dataset.

## Distribution

First, we will look at the distribution of the values in each column by drawing boxplot and KDE plot. For the *“Rating”* column, we will use boxplot:

**Figure 16**

*Distribution of “Rating” column*

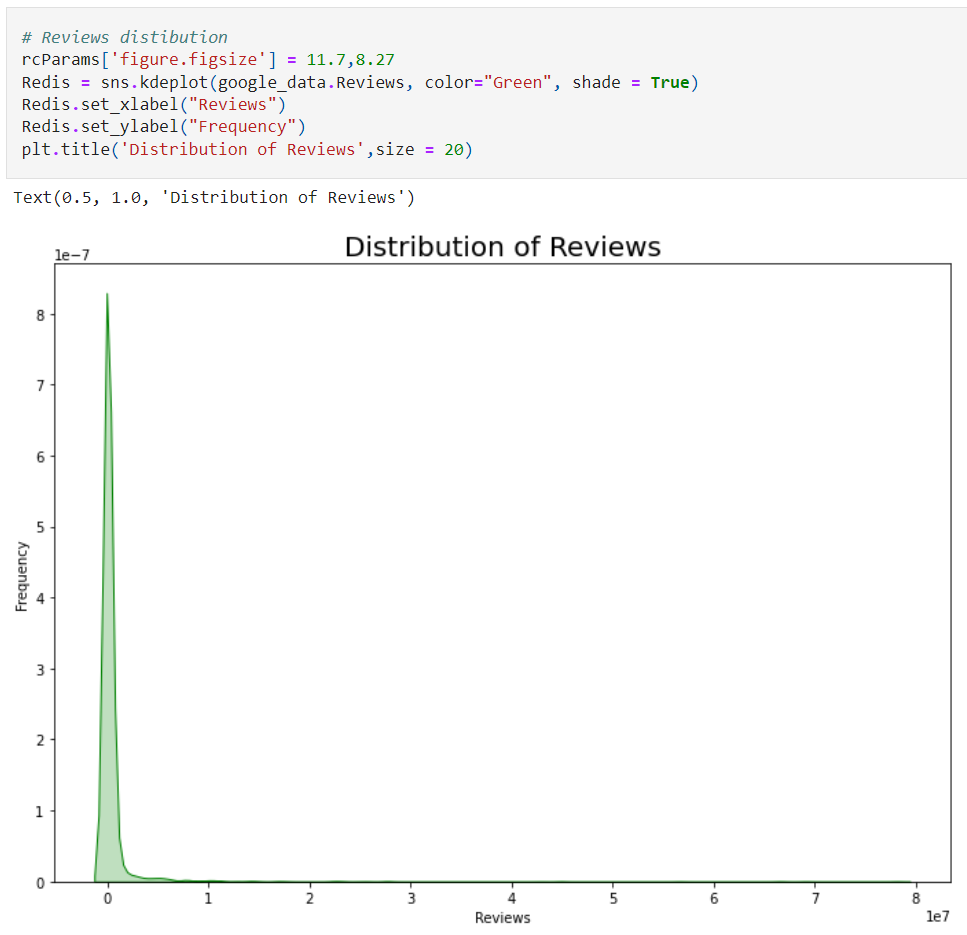


Despite the fact that the boxplot is pretty symmetric, there are a lot of outliers. The majority of the apps have their ratings at around 4 to 4.5, which is quite high.

To see the distribution of the *“Reviews”* column, we will use the KDE plot instead since the boxplot cannot represent this attribute the right way:

**Figure 17**

*Distribution of “Review” column*

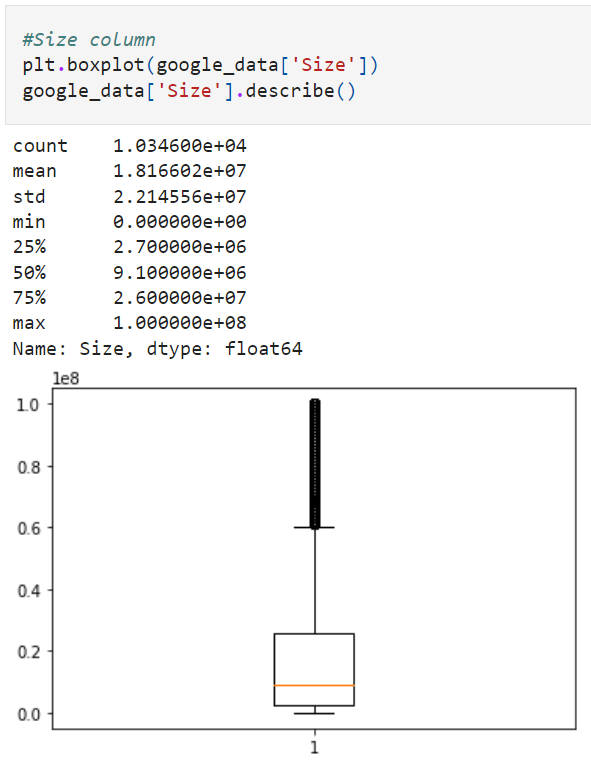


Most apps do not have any review at all and there appear to be a huge number of outliers in this plot as well. Hence, the values in the *“Reviews”* attribute do not spread out evenly.

Next up is the boxplot of the *“Size”* attribute:

**Figure 18**

*Distribution of “Size” column*

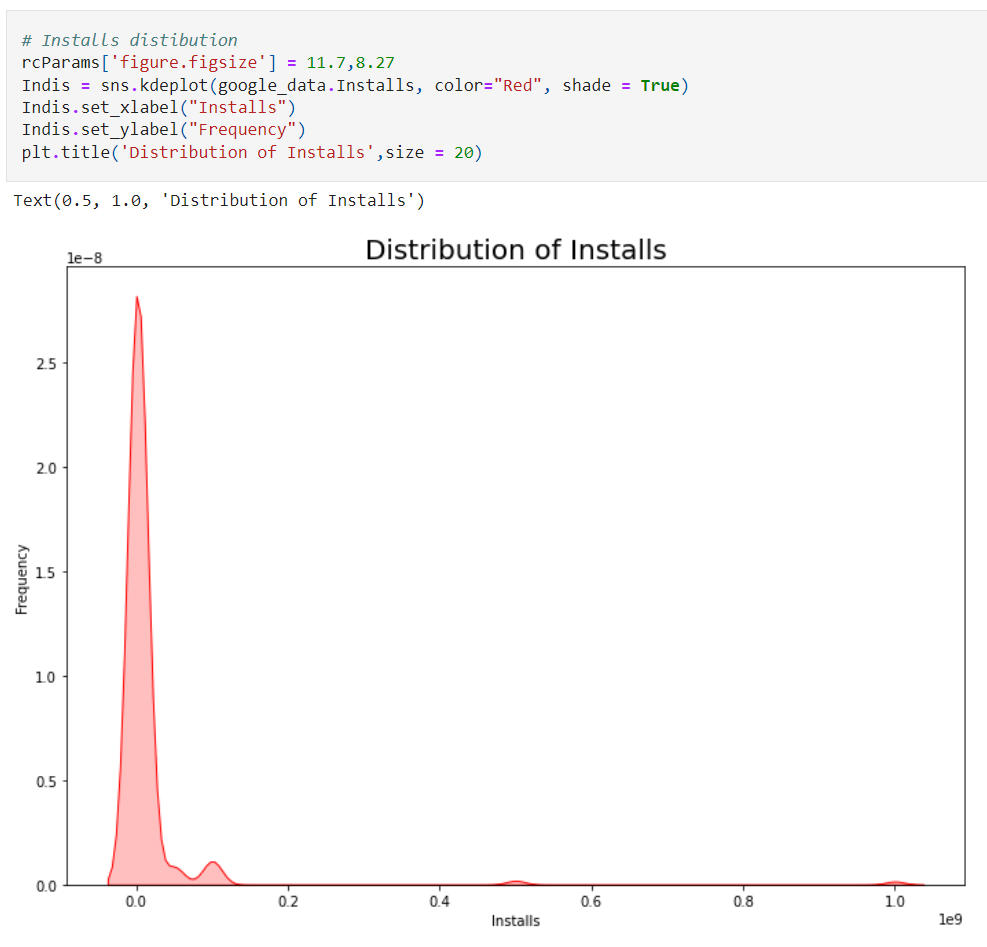


The boxplot of the *“Size”* attribute is right-skewed, that tells us that the number of lower values in this column is more than the number of higher one. Most apps are around 2.7-9.1MB in size.

The following KDE plot illustrates the distribution of *“Installs”* column.

**Figure 19**

*Distribution of “Install” column*

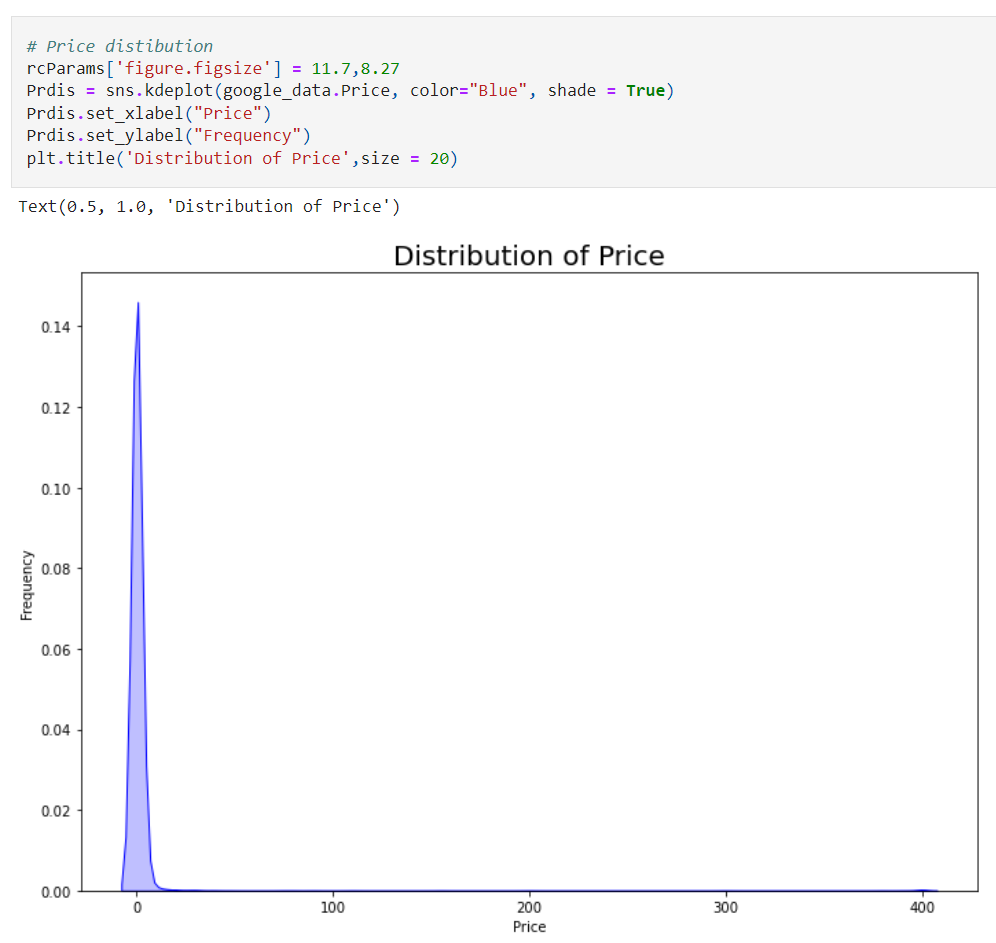


Similarly, to the *“Reviews”* attribute, most of the apps do not have anyone installed at all. The plot is also not evenly distributed as there are also a lot of outliers.

And the last column to be looked through is the *“Price”* column:

**Figure 20**

*Distribution of “Price” column*



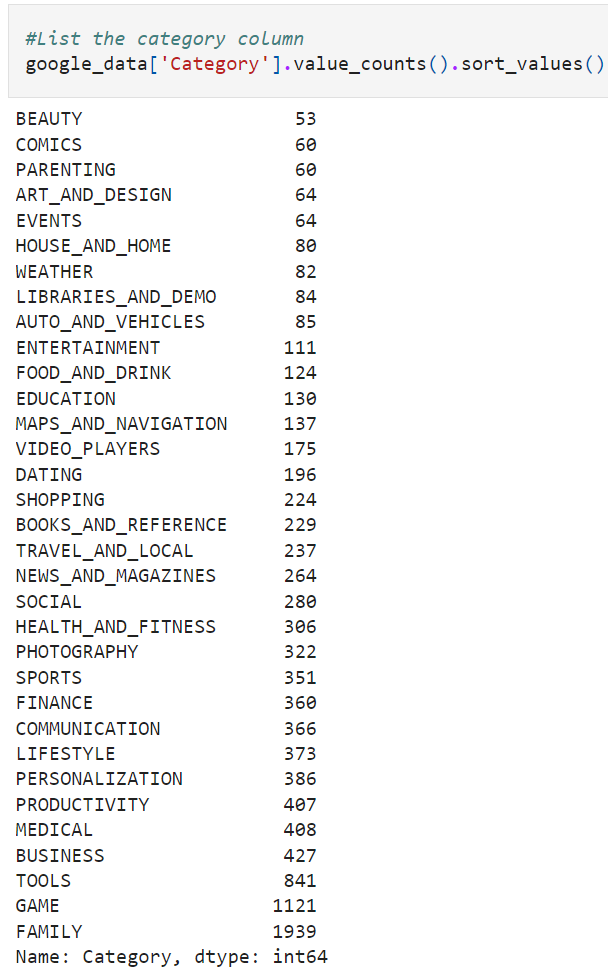
On the Google Play Store, we can find that a big number of apps are free, and the prices range from 0 to more than 400$.

## Features of the dataset

In the next section, we will find out how many categories available in the Appstore:

**Figure 21**

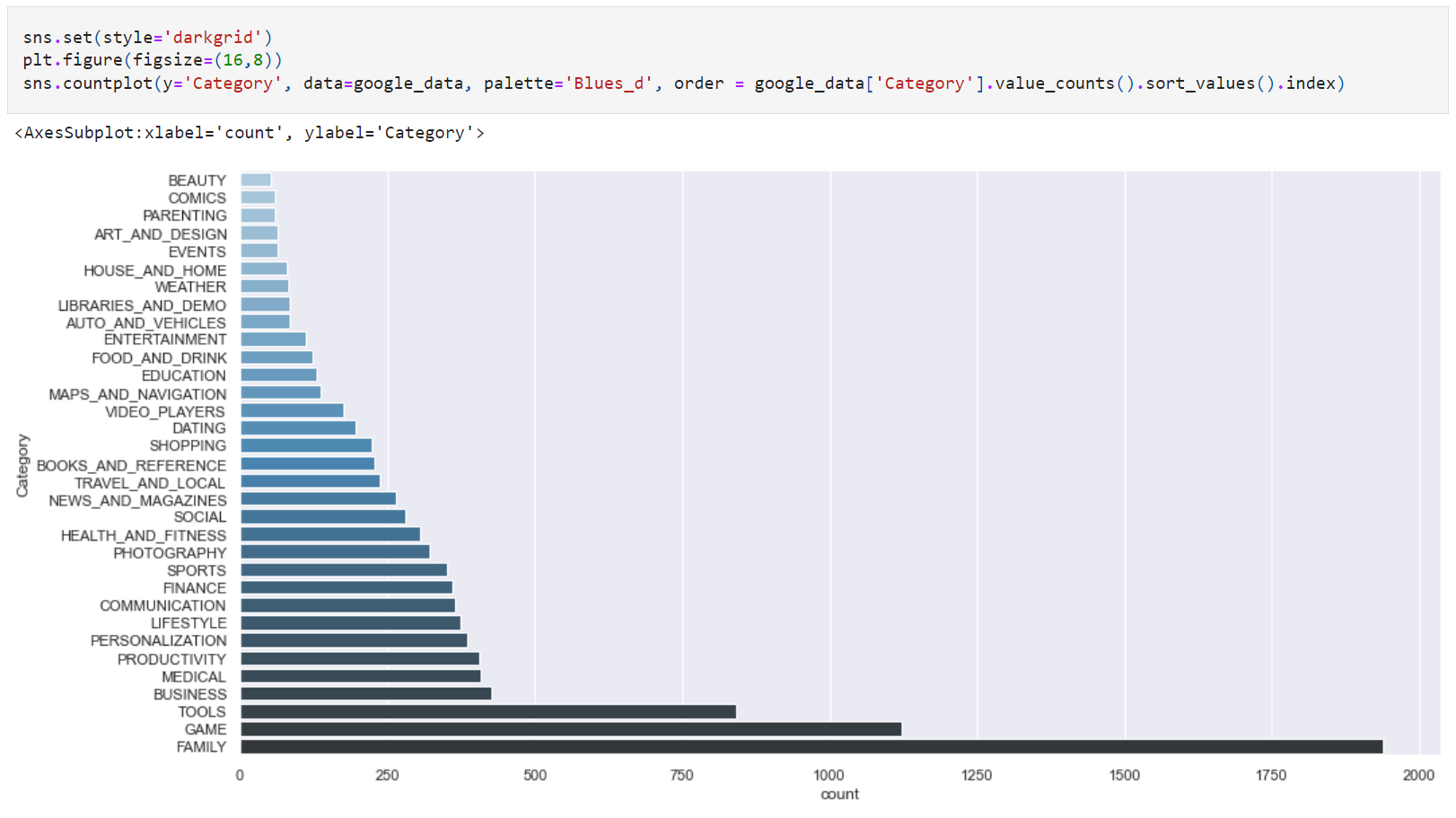
*Categories of the apps*



We will use a count plot to help us understand these Figures better:

**Figure 22**

*Count plot of the apps in categories*

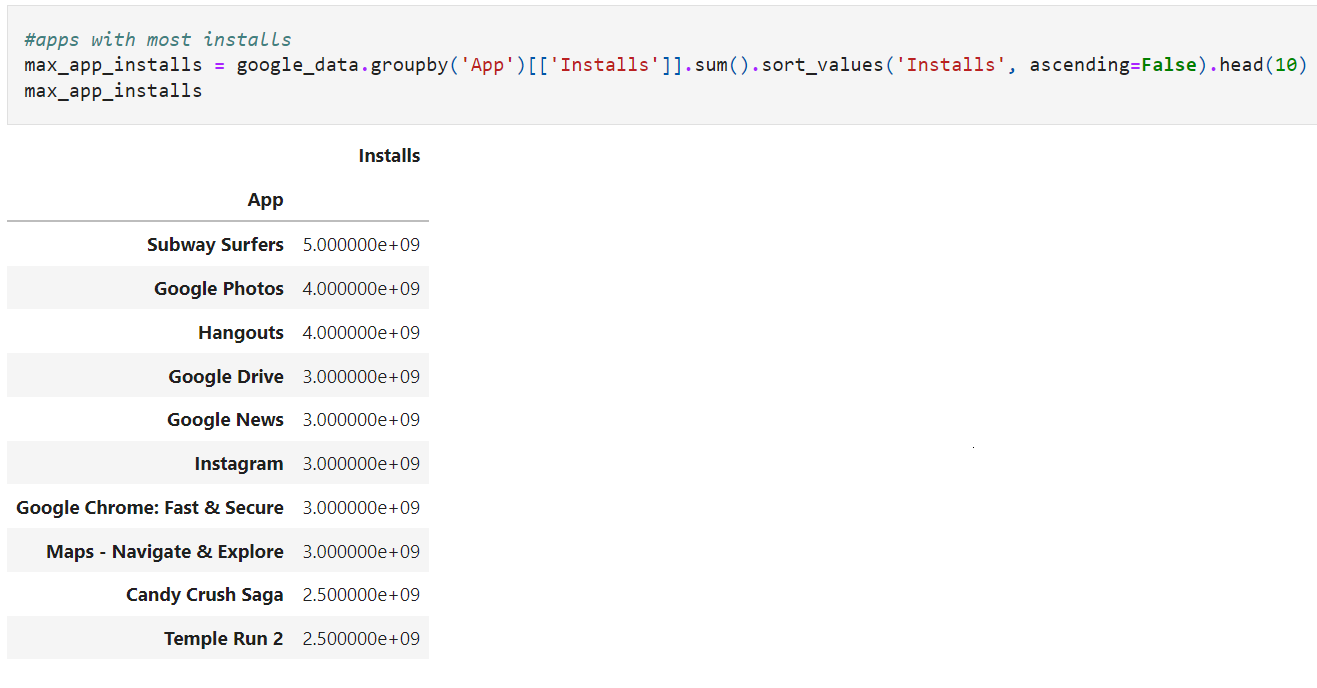


In the Google Play Store, there are **33** different sorts of categories. With 1939 apps, the most popular category is *“Family”*, which is significantly larger than any other category. The *“Beauty”* category, on the other hand, has the least number of apps in the store with the number of 53.

Next up, we will use the *“.sort\_values”* function to find out the top 10 apps with the most installations on the Google Play Store:

**Figure 23**

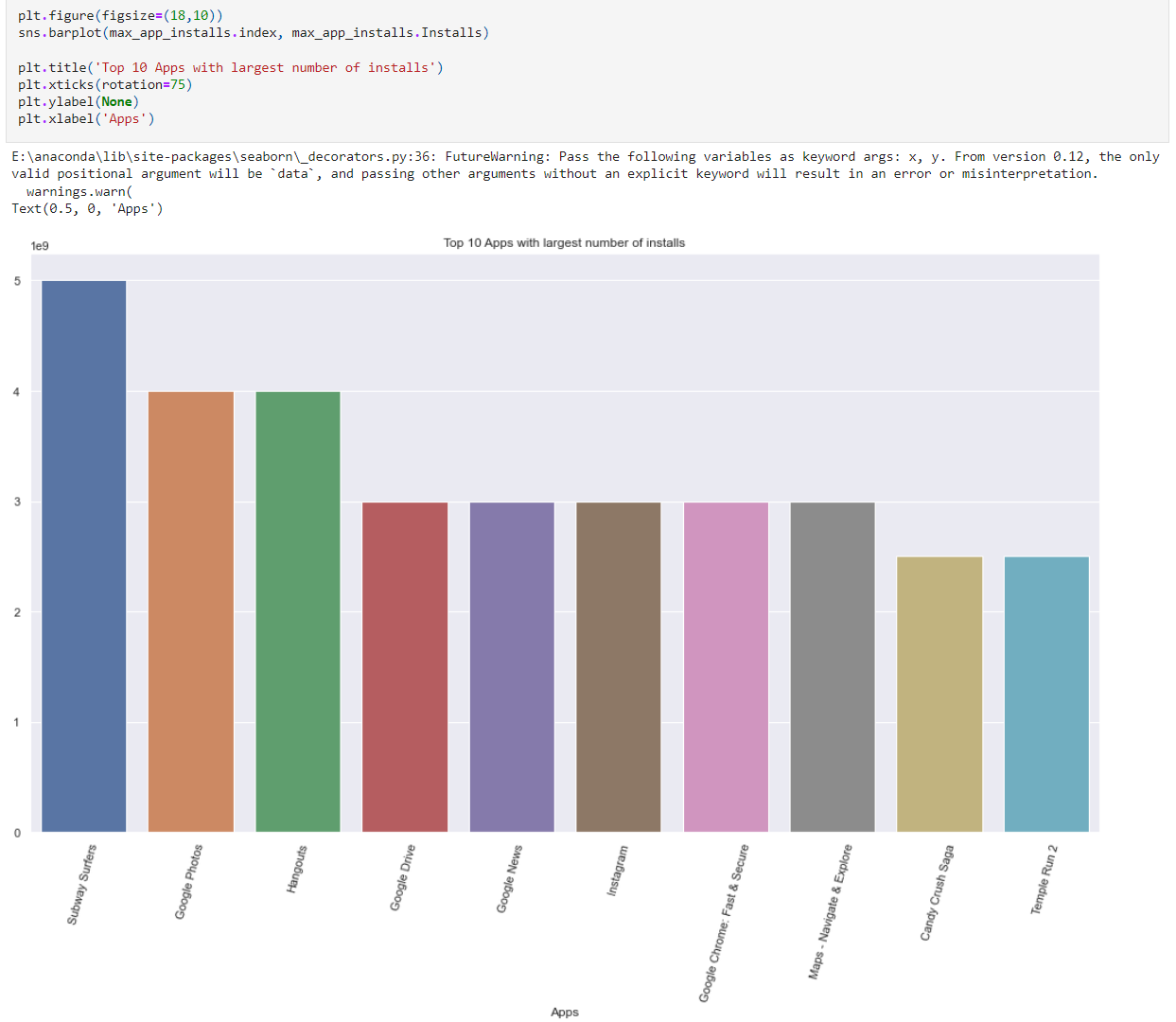
*Top 10 apps with most installations*



To properly illustrate these numbers, we will create a bar plot :

**Figure 24**

*Bar plot of top 10 apps with most installations:*

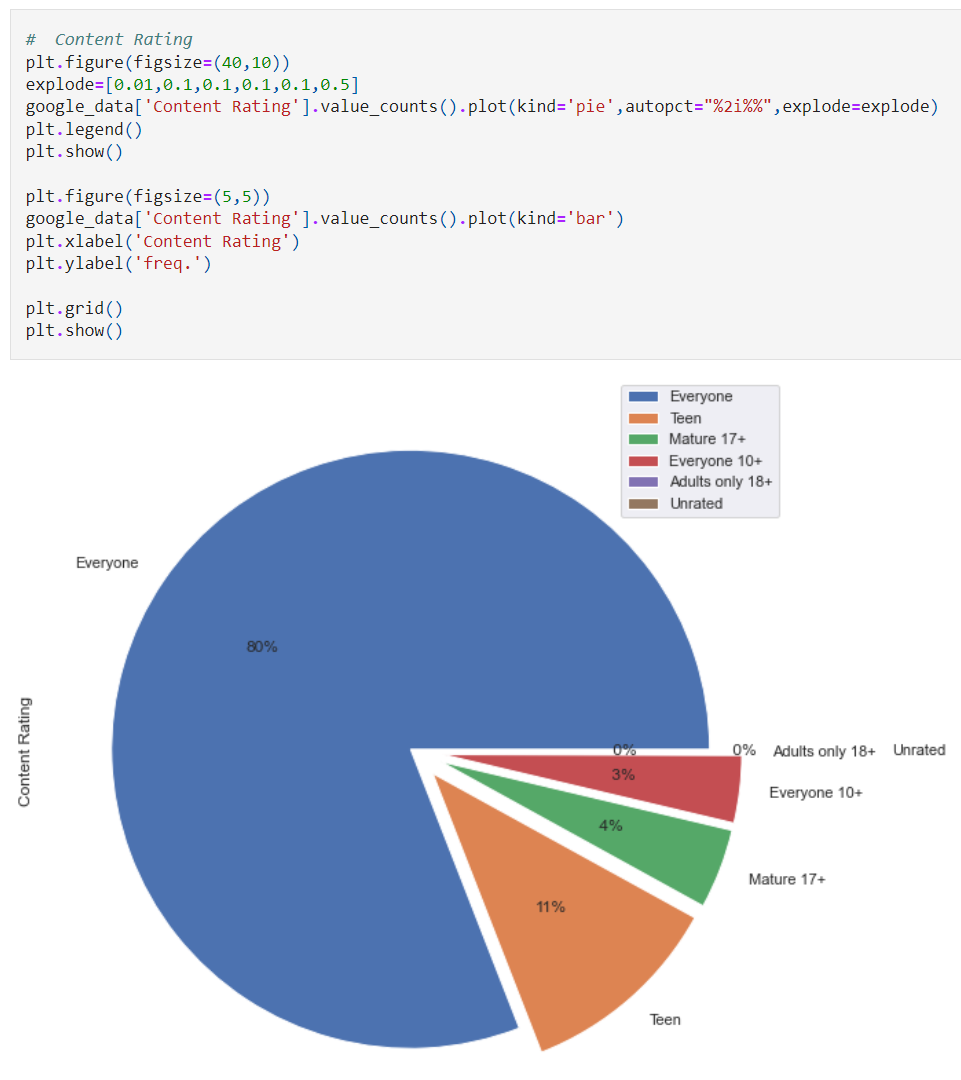


The most popular app on the Play Store is the game *“Subway Surfers”.* There are also 2 other games in the top 10 which are *“Candy Crush Saga”* and *“Temple Run 2”.* Surprisingly, 6 out of 10 most downloaded apps on the Play Store are released or made by Google. As a result, we can conclude that Google has a big influence on consumers who own an Android phone. In addition to games and Google-based apps, *“Instagram”* makes an appearance in the top 10 also.

It is time to move on to the next part, in which we will look at the proportion of each age group in the *“Content Rating”* attribute using a pie chart and a bar plot:

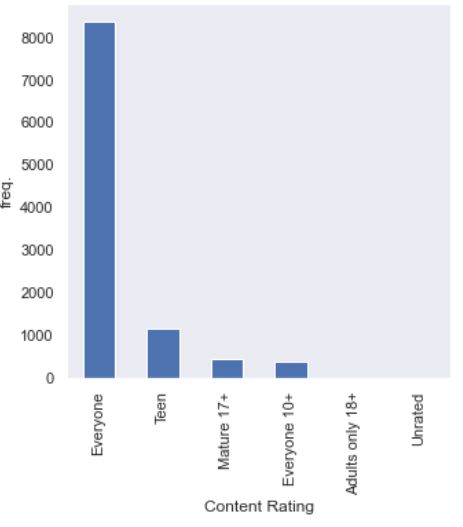
**Figure 25**

*Pie chart of “Content Rating” column*



**Figure 26**

*Bar chart of “Content Rating” column*

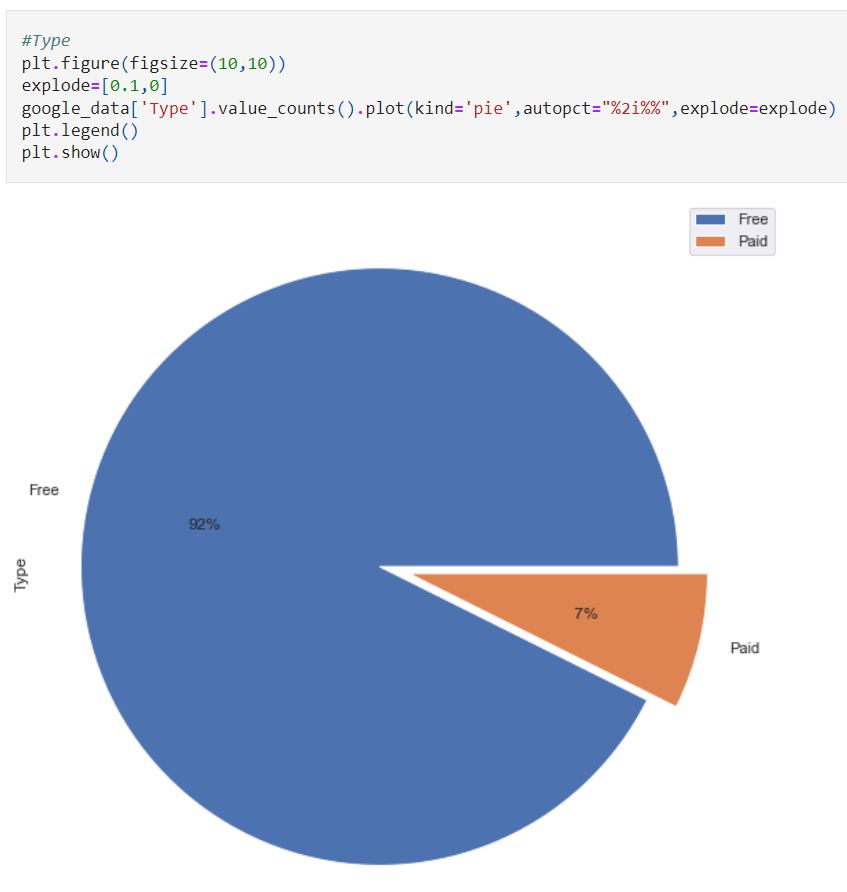


There are **5** separate age groups in the *“Content Rating”* attribute, apps that do not have a valid age group fall into the *“Unrated”* type. The majority of the apps are made suitable for *“Everyone”* to download, whereas the other age groups have much fewer apps. There are also less than 10 apps in the *“Adults only 18+”* and *“Unrated”* content rating.

We will see how many apps are free and paid in the Play Store by looking at the next pie chart:

**Figure 27**

*Pie chart of “Type” column*

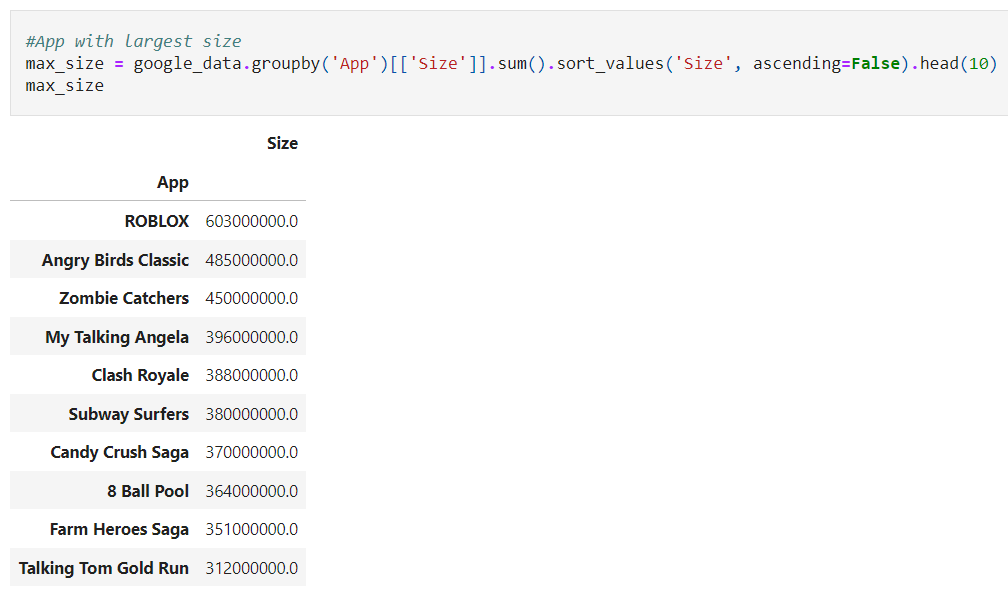


Up to **92%** of apps in the Play Store are free to download, while only **7%** remaining apps are paid with money.

The *“.sort\_values”* function will be used one more to display the top 10 apps in the Play Store which have the largest size:

**Figure 28**

*Top 10 apps with the largest size*



Interestingly, all 10 apps with the largest size on the Play Store are games. This is reasonable, though, since games will require more size to contain better graphics or content in them.

And for the final section in the **Descriptive Analysis**, we will see the top 8 apps with the most reviews by building a bar plot:

**Figure 29**

*Top 8 apps with the most reviews*



The app with the most reviews on the Play Store is *“Facebook”* with more than 78 million reviews. The others app in the top 8 are mainly games or social media platforms.

So far, we have analyzed the values in each attribute in the Google Play Store Apps dataset. In the next section, we will see the relationship between these columns.

# DIAGNOSTIC ANALYSIS

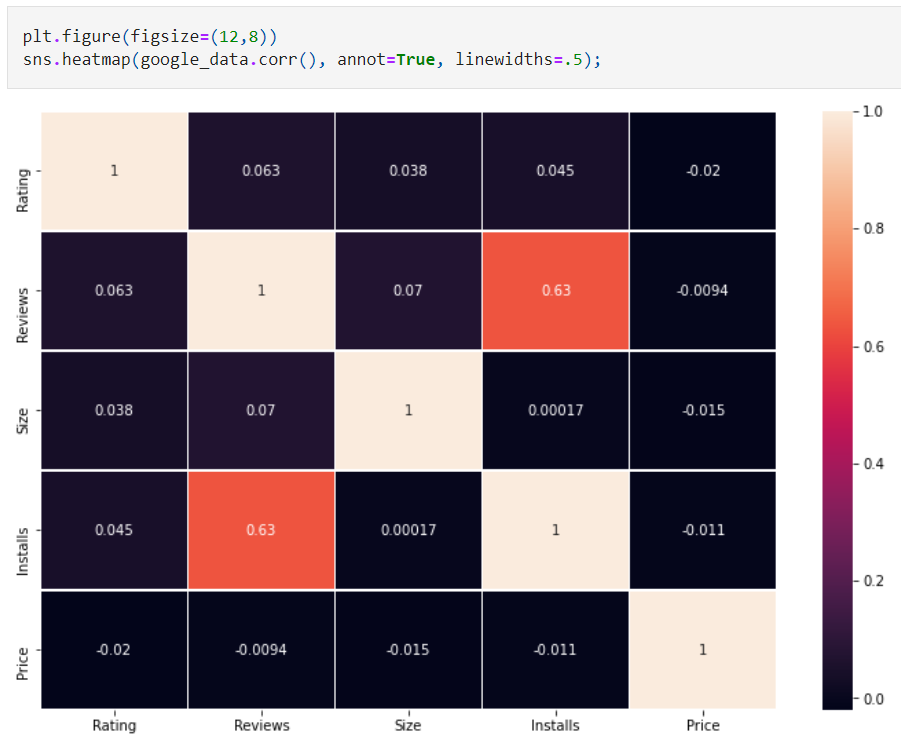
The **Diagnostic Analysis** will be used in this section to explain why we got those results in the previous part. We will also find out the *“Installs”* column depends on which attribute and how.

## Relationship between installs and other attributes

First, we will create a heatmap to visualize the link between the number of installations and other attributes:

**Figure 30**

*Heatmap between attributes*



According to the heatmap, Rating, Size, and Price have little effect on the number of installations, while the connection between Reviews and Installations is around 0.63, implying that the greater the Reviews, the higher the Installations.

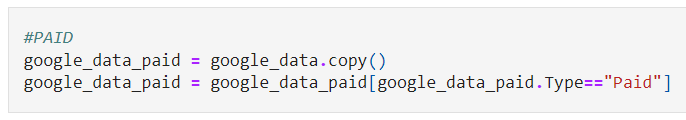
## Differences in free and paid apps

Before continuing, we will divide the values in the dataset into 2 different data frames: **Free** and **Paid** in order to observe the differences between these two types of apps:

**Figure 31**

*Free and Paid dataset*

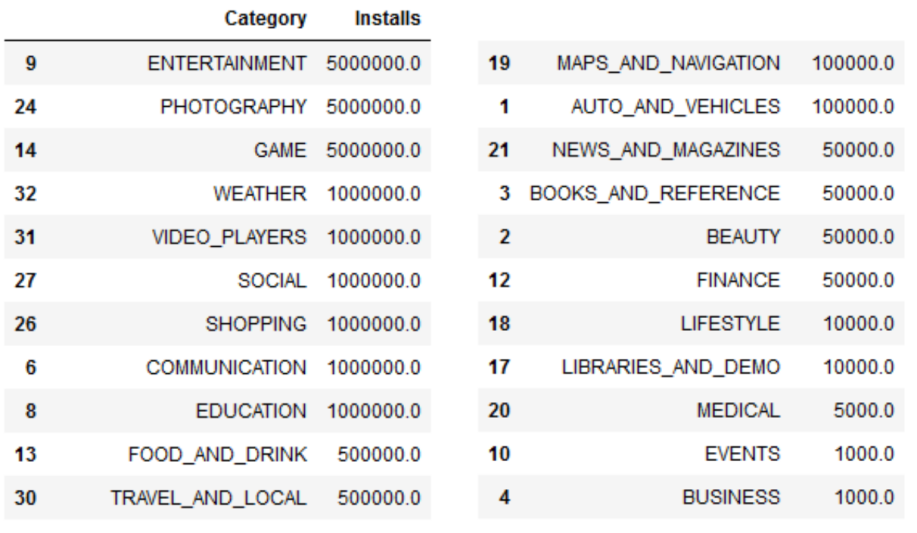




The first aspect we are going to look at are the number of downloads in different categories between **Free** and **Paid** apps:

**Figure 32**

*Installation in categories for Free apps*

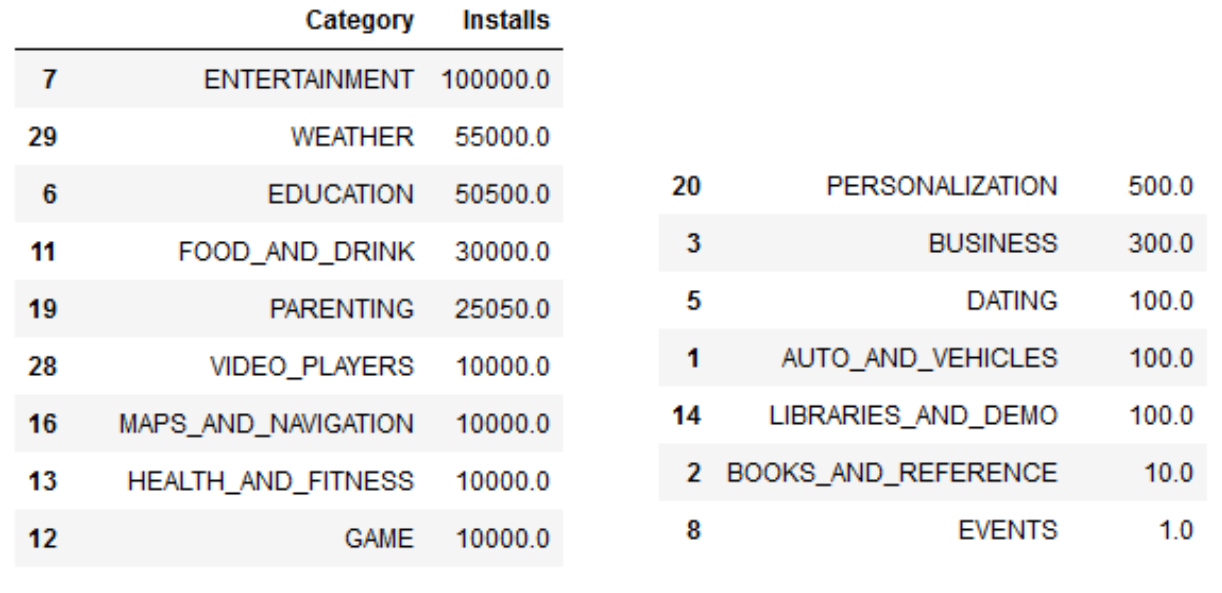


It reveals that free apps in the categories of *“Entertainment”*, *“Photography”*, and “*Game”* have the highest number of installs, with a median of around 5 million.

Furthermore, because free apps may reach a larger number of users, they may become more popular and profitable.

**Figure 33**

*Installation in categories for Paid apps*

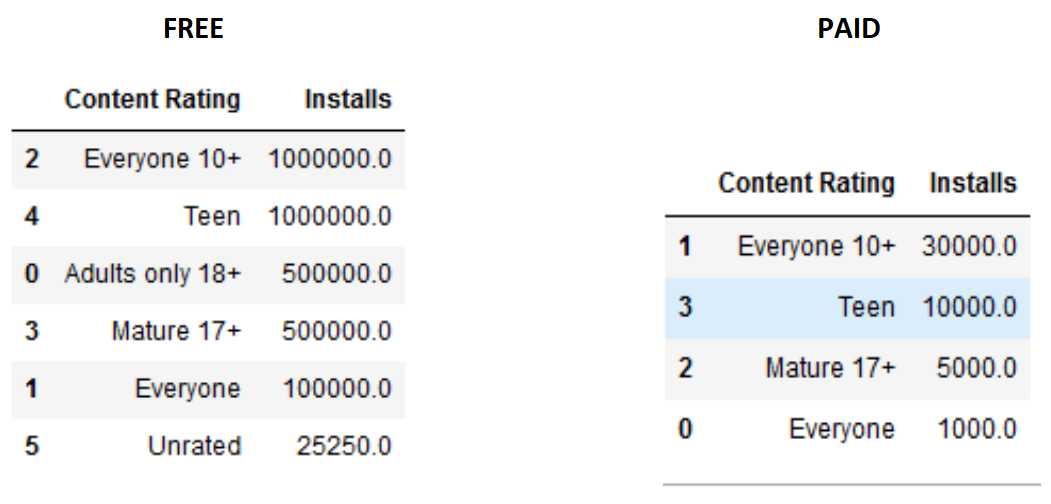


The most popular paid apps are in “Entertainment”, “Weather”, and “Education”, while there are the least amount of people purchasing apps in “Events” and "Books and reference". It suggests that individuals do not spend too much money on unnecessary applications, instead, they will spend their money on more relevant apps such as entertainment or education. Developers should focus on these categories to increase the number of installs.

The second aspect we are going to compare between **Free** and **Paid** apps are *“Content Rating”:*

**Figure 34**

*Installation in Content rating for Free and Paid apps*

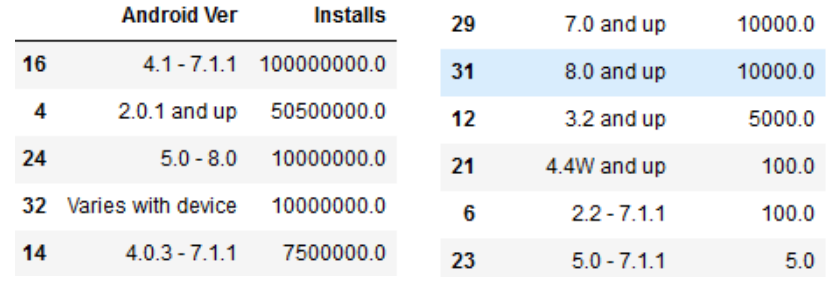


It suggests that content rating has little effect on installation. *“Everyone”* and *“Teens”* are still the most installed in both scenarios.

And the final factor to consider when distinguishing between free and paid apps is the app's *“Android Ver”:*

**Figure 35**

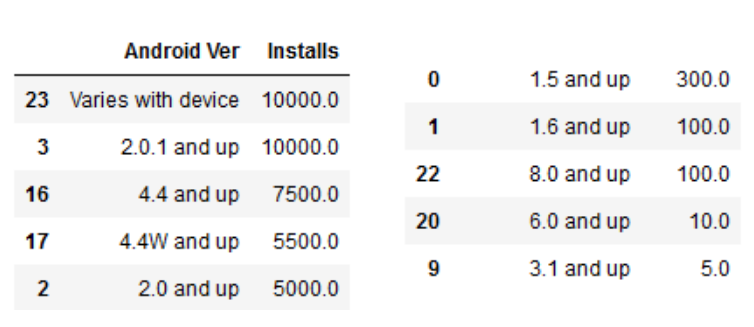
*Installation in Android Ver for Free apps*



People often installed free apps with Android versions 4.1 - 7.1.1 and 2.0.1 and up. It might indicate that the majority of customers' phones are medium-type.

**Figure 36**

*Installation in Android Ver for Paid apps*



While in Paid apps, Varies with device and 2.0.1 and up are the most popular to be installed.

Up to this point, we have done these following steps to analyze our dataset: **Data Preprocessing, Descriptive Analysis** and **Diagnostic Analysis**. Now, we will move to the last part which is **Predictive Analysis.**

# PREDICTIVE ANALYSIS

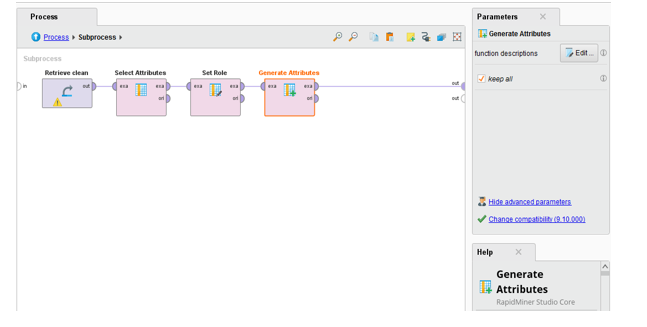
## Goal

Our main goal is to predict what factors affect the installations most and find the way to get the apps approached to the customers. We decided to export the cleaned dataset in Python to .csv file to use in RapidMiner for applying Decision Tree model in order to achieve the goal.

## Implementation

**Figure 37**

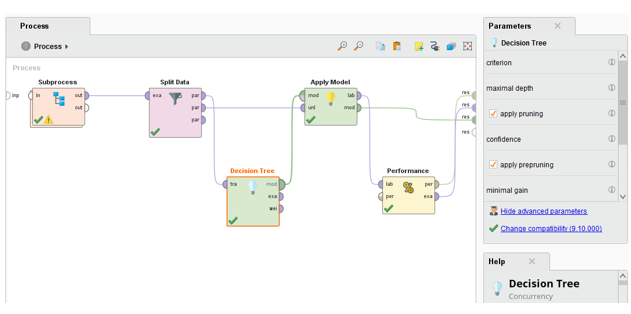
*Process in Rapidminer for preparation the data*



We drop the “att1” column since it is the order number (unnecessary) by “Select Attributes” operator. Then, set the “Installs” column as label and split the number of installations into 2 types – Popular (installations ≥ 100,000) and Unpopular (installations < 100,000) to apply the model.

**Figure 38**

*Process in Rapidminer for testing and training the data*

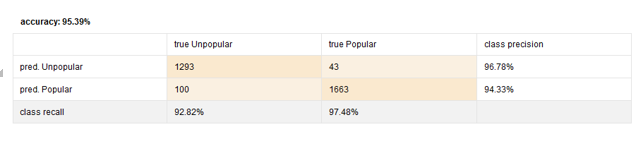


For applying the model, the dataset is split into 2 for training and testing (70% for training and 30% for testing).

## Result

**Figure 39**

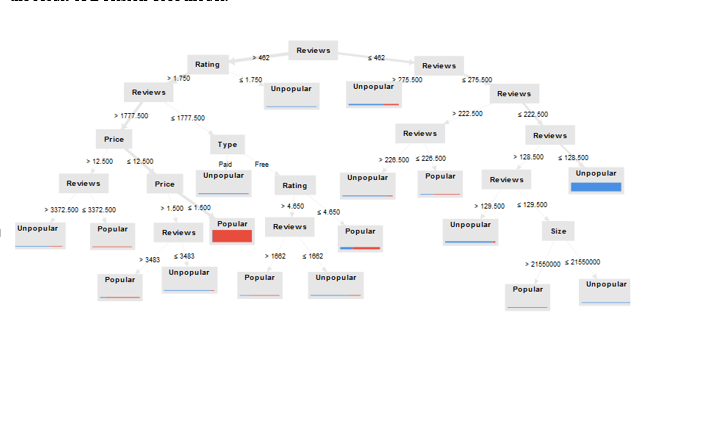
*Result of the data after training and testing*



All of the Accuracy, Precision, Recall gets high percentage (>90%). Therefore, we decided to use the result of Decision Tree model.

**Figure 40**

*Decision Tree of the data*



According to the Tree, we can see that Reviews has a huge impact on the Installs. The apps will be popular if they have more than 462 reviews (the tag “Popular” account for around 92%). For more details, Rating does have impact on Installation, but it is not too much as the Rating need to be less than 1.75 then the apps may become unpopular. In addition, customers would prefer to install the apps whose price is less than 1.5. In contrast, the apps that have less than 462 reviews tend to less popular.

# CONCLUSION

* While preprocessing the dataset, we noticed a number of problems with the raw data that may have a negative influence on the accuracy of our findings: **missing values**, **duplicated values**, **unnecessary symbols in values**, and **incorrect types of values in various attributes**.
* We have discovered the distribution of several numerical attributes in the **Descriptive Analysis** section, as well as some various types of values in some attributes and their proportions, as well as top values in those attributes.
* We have uncovered the connection between attributes as well as the differences between the free and paid apps in the **Diagnostic Analysis** section. And in the final part, **Predictive Analysis**, we constructed a model to predict app installations number.
* Only the value distribution in the *"Rating"* column is somewhat symmetric. While the distribution of values in other attributes is unequal or has an excessive number of outliers.
* There are 33 different sorts of categories, in which games and Google-related apps are downloaded the most. The majority of apps are designed for everyone to use, and free apps are the most common in the Play Store. Games are often the largest in size, and social networking platform apps typically have a high number of reviews.
* The attributes *"Reviews"* and *"Installs"* are significantly dependent on one another. Whether free or paid, *"Entertainment"* and *"Everyone"* apps will have high installations.
* Free apps would attract more customers than usual. For paying apps, the price should be less than 1.5$ to have more installations.

# REFERENCE

Lavanya. (2019). *Google Play Store Apps*. Kaggle. https://www.kaggle.com/lava18/google-play-store-apps