JSC Final Prject

Yufan Qian

14/04/2022

Introduction

Nowadays, we have so many apps on our phone and computers and everyone's life is associated with them. I want to know what affects people's opinion on different apps and whether they favor some apps with particular characteristics than others. I have narrowed my target to apps on Google Play Store.

The primary question of interest is: What are the factors for different app's rating?

This question can be decomposed into smaller questions:

1. What is the association between number of reviews and rating of the app?
2. What is the association between number of installs and the rating of the app?
3. What is the association between the category of the app and the rating of the app?
4. What is the association between the size of the app and the rating of the app?

Reason for these questions: I think the number of reviews indicate how popular the app is, in both good and bad ways. For example, if the app has many reviews, then it may be the app is super good, or very terrible. Number of installations may indicate good quality and reputation of the apps, while those with lower installations may have more extreme values than those with more installations. App categories may affect ratings, since people may enjoy a certain type of app more than others. Lastly, size of the app may affect people's opinion on it due to their phone's storage limitation. We will find out whether these guesses are correct through analyzing below.

Methods

This dataset is from author Lavanya on Kaggle, and the link is: https://www.kaggle.com/lava18/google-play-store-apps. The reason why I choose to search for dataset on Kaggle is because this website was used mainly for publishing datasets and was also being mentioned in past courses before. According to information on the web page, the "dataset is scraped from the Google Play Store".

checking datasets number of rows in our dataset: 10841. The number of observations is big and should be enough to answer our questions.

After checking all the variables in the playstore dataset, I found the only one variable with missing value is: `Rating`. Since it doesn't make sense to impute and get the rating for this app from other apps (every app is different even within the same category), and we have 10841 observations already, so I choose to remove observations with missing Rating values. The method I have used for removing missing values is called "filter".

Also, there are some duplicate values for the same app in our dataset, so I choose to keep the first occurrence of the same app (For example, there are two observations for Facebook in the raw dataset). The method I have used for keeping only one occurrence is called "unique". After removing observations with missing Rating and duplicates, we have 8197 observations in total, which is still good enough for analyzing.

checking variables we are interested in: `Rating`, `Reviews`, `Installs`, `Catergory` and `Size`

Table 1: summary table for Rating

minimum Rating	maximum Rating	average Rating
1	19	4.175052

Since the maximum value we found for the `Rating` is 19, which is not correct since the rating in Play Store is between 1 and 5, so we modify the original dataset and only keep observations with `Rating` value between 1 and 5.

Table 2: summary table after removing unusual Rating

minimum Rating	maximum Rating	average Rating
1	5	4.173243

After removing unusual observations, we now have 8196 observations in our dataset. Thus, there was only one unusual observation exceed our range (19). The min and max `Rating` are now 1 and 5 respectively, with average rating around 4.17. This means most of the apps got a pretty good rating and we will likely see right-skewed graphs later on.

The variable `Reviews` has type `character`, we first convert it to type `integer` for easier graphing.

Table 3: summary table for Reviews

minimum number of Reviews	maximum number of Reviews	average number of Reviews
1	78158306	255251.5

Here, we could see that the minimum number of `Reviews` for apps in our dataset is 1, and maximum is 78158306(which is from app Facebook). Average number of `Reviews` is 255251.5.

The data for variable `Installs` are written as "10,000+". So first we need to remove unnecessary "+" and "," in the values to make it integer.

Table 4: summary table for Installs

minimum number of Installs	maximum number of Installs	mean number of Installs
1	1e+09	9165090

For variable `Category`, there are in total 33 different values. They look normal so no modification is needed.

Table 5: summary table for Size

minimum Size	maximum Size	average Size
1e+06	9.94e+08	37401281

After checking the variable `Size`, I found some apps has `Size` value "Varies with device". I removed those observations and convert the type of this variable to numeric. Now we have 7027 observations left.

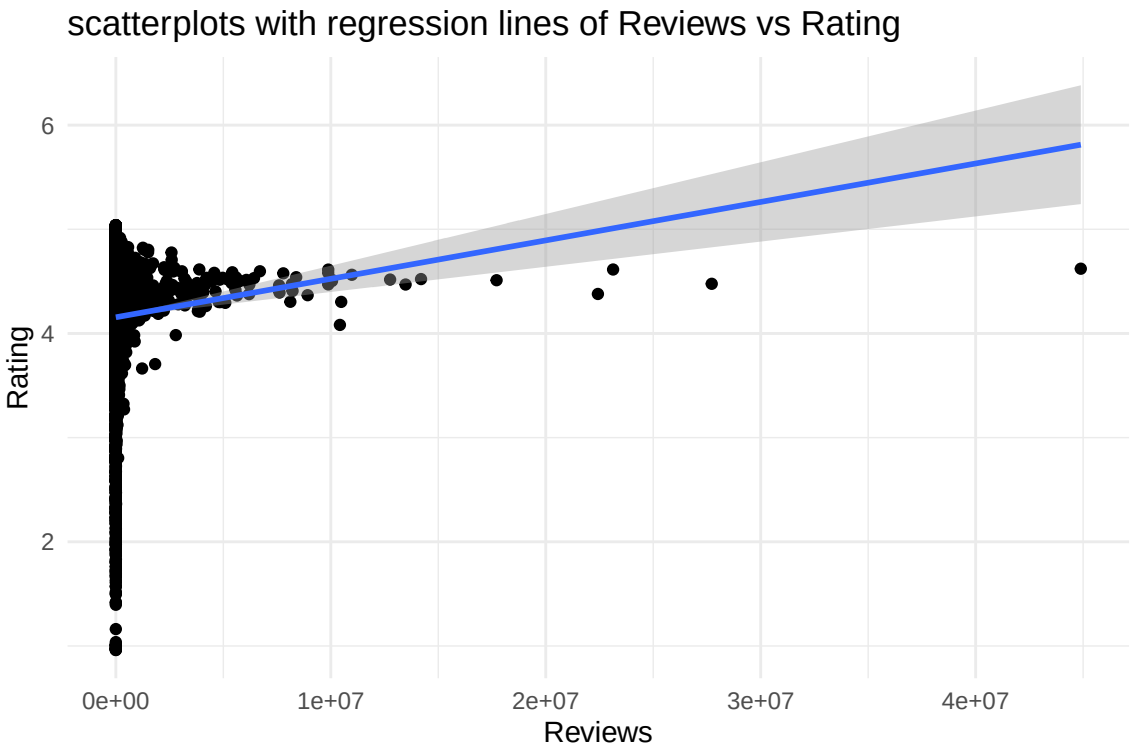
Lastly, I created new variables: `review_level` and `install_level` to make analyzing later on more convenient. They categorize number of reviews and number of installs using thresholds: <1000, 1000-10000, 10000-50000,

>50000, with `review_level` having values of "almost no reviews", "medium reviews", "many reviews", and "lots of reviews", and `install_level` having values of "not popular", "normal", "popular", and "super popular".

The methods I have used for my analyze are mainly graphs: `scatterplot`, `barchart` and `statistical summary graph`. I have also used `summary table` for checking our variables of interests. The reason why I would like to use graphs a lot is because I think graphs tend to be more visualizing than numbers.

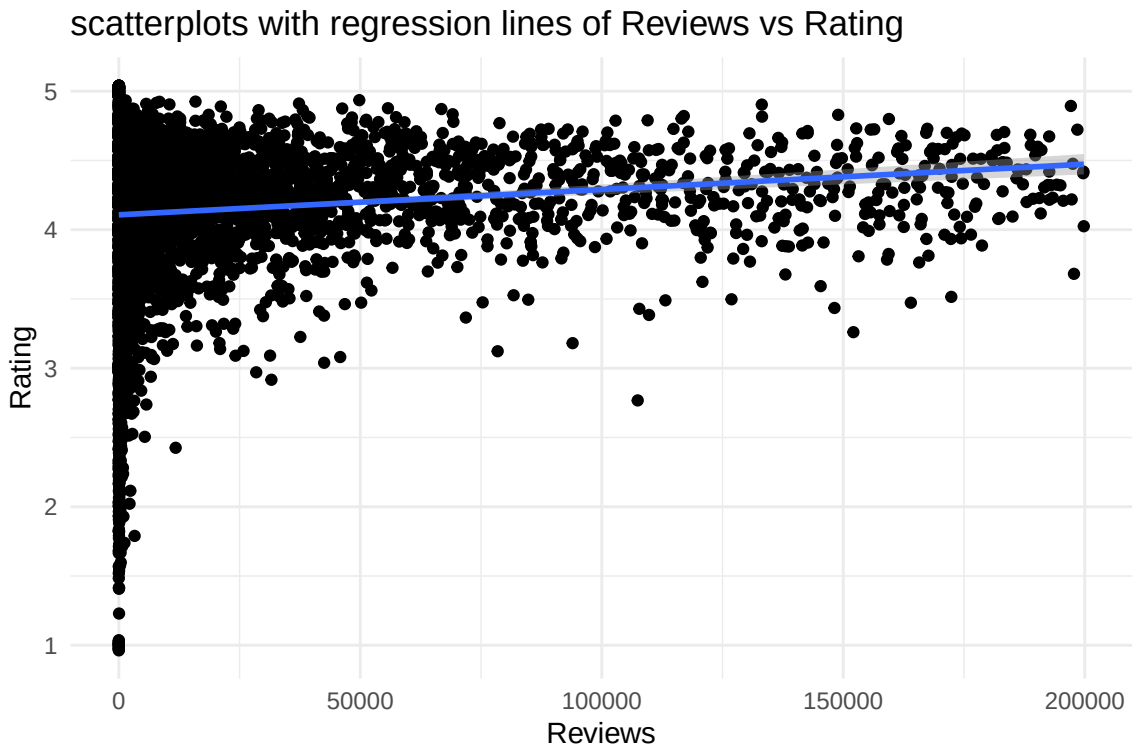
Preliminary Results

1. Reviews, Rating - Scatterplot First of all, we want to see whether apps with more reviews tend to have higher ratings, since number of reviews may indicate how popular the app is.



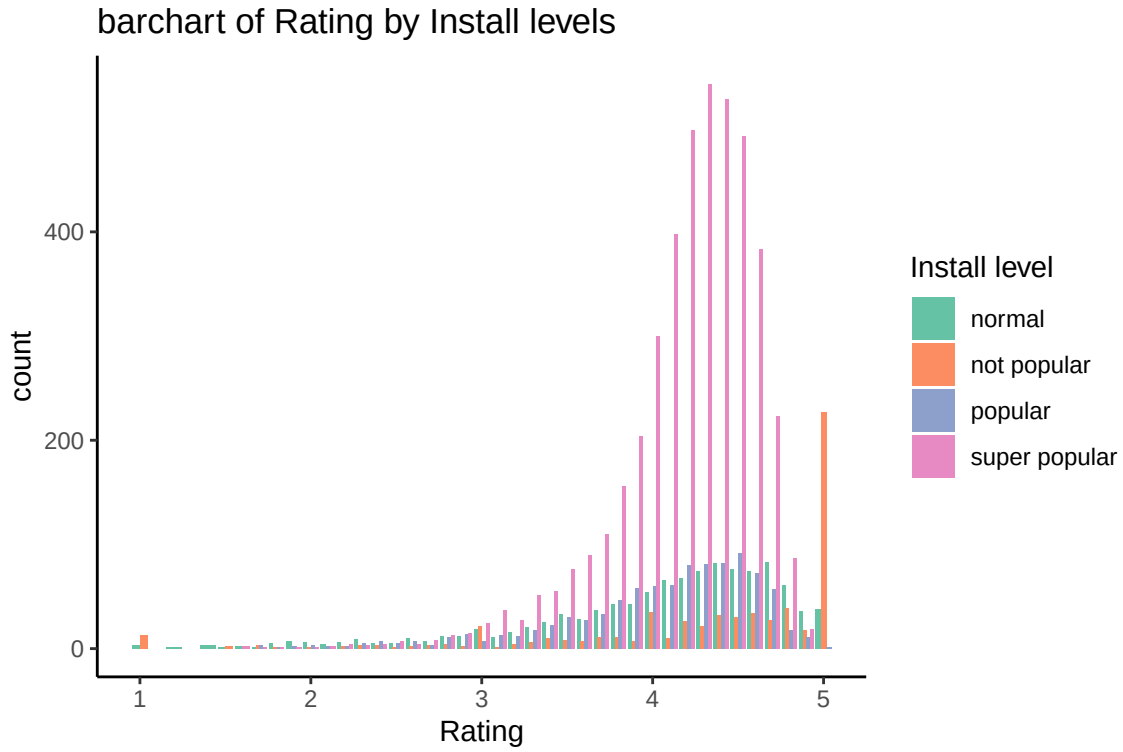
Here, we use the scatter plot with regression line to examine the association between Reviews and Rating. We could see that most of the points are on the left hand side, makes a huge cluster and is hard to see them clearly. Although the regression line has a positive slope, but when the review numbers goes up, the number of observations decreases. So the slope may be influenced by the few observations with large reviews. If we pay attention on the left hand side, we could see that there is no obvious linear pattern on ratings and number of reviews.

In order to see clearly about what is happening when reviews are under 200000 (to zoom in the left hand side cluster), we created a new dataset called less_reviews which only keep apps with Reviews $<$ 200000 to check the left cluster.



Now our new graph looks more clear than the previous one with points all over the graph. Here, we do see a slightly increasing trend through the positive slope of the linear regression line. Thus, we can conclude that there is a weak positive relationship between ratings and number of reviews.

2. Installs, Rating - Barchart After checking the variable `Reivews`, we want to see whether apps with more `Installs` tend to have higher ratings.



From the barplot, we could first see that the number of apps which are super popular(lots of installations) are the largest component of our observations(pink), and most of the apps in this category are centered around 4.3. The apps which are normal(green) has a relatively same trend with apps which are popular(blue) and their trend seems to be more flat than the trend for super popular(although they have relatively the same peak value), thus ratings are less than those super popular apps. The apps which are not popular(orange) has a peak at 5. We can also notice that apps which are not popular also has a local maximum at 1(orange), thus it confirms our hypothesis that apps with less attention will have more ratings on two tails(1 and 5), i.e., more extreme values than others. Through the bar chart, we do see that more installations may result in higher ratings due to the shape of the graph for each installation level.

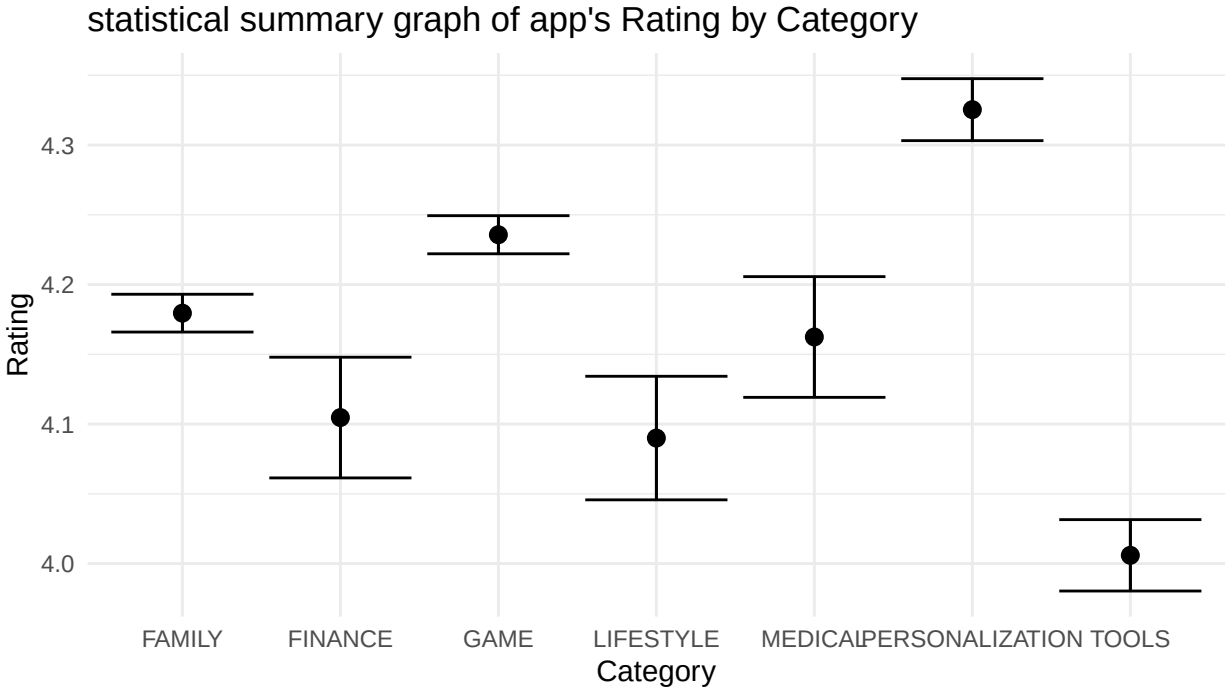
I'm also interested to see whether these two variables: `Installs` and `Reviews` both represent popularity of the app. So I have decreased the number of installs by a factor of 30 for better visualization and using line graph to plot both variables on the same graph. The interactive graph can be viewed from our website's home page, called "Installs vs Reviews".

From the line graph, we clearly see that the peak of the line for number of reviews is near the peak of the line for number of installs. All the high values of number of installs/reviews are between ratings 4 and ratings 5, and the peaks are around 4.5. This is probably different than what people would normally guess, since we would think that higher rating and popularity are linked with each other. But the peak is not at rating of 5. I guess this is because the more popular the app is (more people using it), the more criticizing the app will receive, and there is no app that makes everyone happy.

So, so do see similar trend in the two varibles `Installs` and `Reviews`. Thus, combined with the two graphs above, we can conclude that there is a positive relationship between number of `Installs` and `Rating`, and there is positive relationship between number of `Reviews` and `Rating`.

3. Category, Rating - Statistical summary graphs, Line plot Now, we want to examine the relation between different `Category` of the apps and `Rating`. since there are too many Categories and it will be inefficient to graph all of them, I choose to create a new dataset: playstore_300 for graphs and only keep those categories with more than 250 apps in it, since more observations would give us more accurate results.

In the new dataset playstore_200, we have only 7 Categories left: FAMILY, FIANCE, GAME, LIFESTYLE, MEDICAL, PERSONALIZATION and TOOLS.



Using a statistical summary graph, we could see that the category`Personalization` and `Game` have higher mean rating value than others, which are above 4.25, while `Tool` has the lowest mean rating value, which is 4.0. All the other types have mean ratings between 4.2 and 4.05. We would think that different categories indeed affect the rating of the apps.

Now, we can use line plot with all the categories involved and combined with `Reviews` to confirm our hypothesis. The line plot can be found on the website's home page, called "Line Graph".

From this graph, we could easily see the extreme low rating value with colour pink, and very low numbers of reviews. This is from category `Tools`, with only 3 reviews and rating of 1.0. Most of the apps have rating between 4 and 5 and less than 10,000,000 reviews, even with fluctuation. The apps have more than 10,000,000 reviews are from categories `Game`, `Communication` and `Tool`. I think this result is due to that apps for `Communication` and `Tool` are popular, and are necessary for people in all the age groups, so these apps have many reviews.

`Game` tend to spread quickly and people may have more things to talk about in reviews. Also, `Game` type of apps are for entertainment, thus may be the reason its higher rating than other type, and we could clearly see the increasing trend of `Rating` for `Game` after 30,000,000 reviews. Thus, different categories does affect the app's rating.

4. new variable: Size of the app, Rating - Scatterplot Now we want to take a closer look at the factors affecting rating, with the newly added independent variable: size of the app. Here, I have plotted a scatter plot, with Rating of each app as our y value, and number of Reviews of each app as our x value. The size of the point is linear with the size of the app, and the color represent different categories of the apps. I added the new variable `Size` because I do believe the size of apps may become a constraint for some people. The interactive graph can be viewed from the home page, called "Size graph".

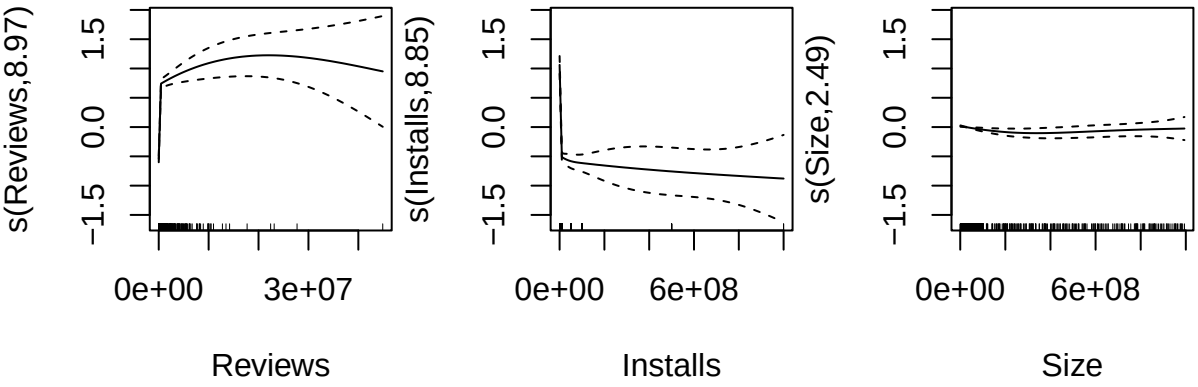
From the scatter plot, first thing we would notice is that most of the apps with large number of reviews are small in size (small circles on the right hand side, some are even hard to see), while most of the large circles (i.e., apps with large size) tends to have smaller number of reviews. I think this observation aligned with my guess and makes sense, since people may encounter storage limitation problem on their phone, so that they would choose to download apps with smaller sizes. Other than extreme values, we see apps with small size and apps with large size tend to behave the same. Thus, there may exist some weak relationship between `Size` and `Rating` or no relationship. We will figure it out in the following section.

Check the correctness of our analysis From our analysize above, we concluded that `Installs`, `Reviews`, `Category` and `Size` of the app affects the `Rating` of the app. To check whether this hypothesis is correct, I'm going to perform additional analysis and build models.

First of all, I'm going to create a linear regression model with only the one variable as the independent variable, and `Rating` as the response.

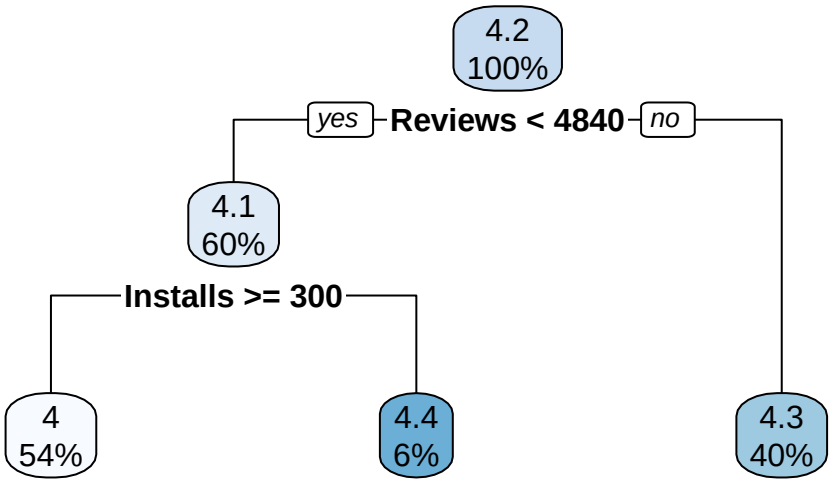
From these four models, the adjusted R-squared values are < 0.005 for `Installs`, `Reviews`, `Category`, but less than 0.001 for `Size`. To see their effects together, I created a new MLR model, and the adjusted R-squared value has increased to 0.02897.

Now, let's build a non-linear association model with cubic regression splines, since during our analysis above, we do see some non-linear characteristics such as extreme values on two tails for `less installs`. We use "s" to denote smooth terms and "bs=cr" to denote cubic regression splines. We don't know for sure whether the relationship between predictor and response are linear or not, and whether the relationship between predictors are linear or not, so I used multiple smooth terms in the model. After testing with multiple cases, it also appears to be the best one.



From the summary table generated, the adjusted r-squared value now is 0.128, which is larger than linear case, and deviance explained is equal to 13.5%. The p-value for variable `Size` is larger than the p-value for other variables, but still is < 0.05 which means the variable is significant and should not be removed. Thus, we can conclude that all of the variables are necessary for our model. Through the increasing values of adjusted r-squared compared with all the linear models above, we can conclude that there is indeed a non-linear relationship between `Installs`, `Reviews`, `Category`, `Size` and `Rating`.

Lastly, let's check our conclusion using pruned regression tree. Since the `Category` variable has too many values, I used the other three: `Installs`, `Size` and `Reviews`.



Here, we do see that two levels of the tree already explains the `Rating` very well. If we have number of `Reviews` < 4840, we get average `Rating` 4.1. If we have number of `Reviews` $>= 4840$, we get average `Rating` 4.3. This aligns with our finding above that `Reviews` has a positive association with `Rating`. Combined with the varaible `Installs`, we get this overall pruned regression tree. After calculating the MSE of the regression tree using the testing data, we get a value of 0.27. This is a small error, and indicates that our model predicts the testing data very well.

Conclusion/summary

In conclusion, we found that the `Rating` of apps on Google Play Store indeed depends on many other factors, and through our analysis, we found that number of `Installs` of the app, number of `Reviews` of the app, `Category` of the app and `Size` of the app all contributes to the variation of `Rating`, with `Size` contributes less than other three predictors. The relationship between predictors and response is non-linear, there are associations and interactions between the four predictors stated above, and the non-linear association model with cubic regression splines we generated could explain 13.5% of the deviance in `Rating`. We also built a regression tree, which aligns with our hypothesis and confirmed the positive relationship between number of `Reviews` and `Rating`.