Final Project

Entong Li, Jingyu Xie, Xindi Lu, Zijia Cao

12/1/2019

Introduction

It's quite common for people to evaluate the effectiveness of Apps based on their ratings. Therefore, our group decides to analyze does any factors show a traceable effect on ratings and what kind of factors indeed change ratings. We think the findings may help software developers to improve their applications in the future. This study aims to evaluate the Google Apps' ratings, and how do category, size, and price affect the final performance of ratings. After doing exploratory data analysis and building linear model, we finally conclude that Apps which have a lower price and larger size are more likely to receive higher ratings on the Google Play store.

Research Question

This project is mainly focused on: What kind of apps would always have a higher rating as well as the prediction of app ratings.

Data Description

This dataset initially collects the basic information of different apps from the Google Play Store. The original dataset includes 9366 observations and 11 variables(includes one dependent variable Ratings). All the variables are: App, Category, Rating, Reviews, Size, Installs, Type, Price, Content Rating, Genres, Last Updated, Current Version and Android Version.

This dataset is cited from: https://www.kaggle.com/lava18/google-play-store-apps.

Data Cleaning

As only a few of the apps have `Size` is counted by `kb`, which is around 200, so we decide to not include this kind of apps. Also, for the apps that have `Size` larger than 100Mb, its size is too large and only a few of them are in this dataset. Thus, we will also exclude the apps with `Size` larger than 100Mb. What's more, for the `Price` of the apps, there some extreme value, less than 180 of them have `Price` larger than `$10`, so we will include the apps with `Price` less than or equal to `$10` only. The `Category` of apps we mainly focus on the top five popular categories: Family, Game, Tools, Fitness, and Media.

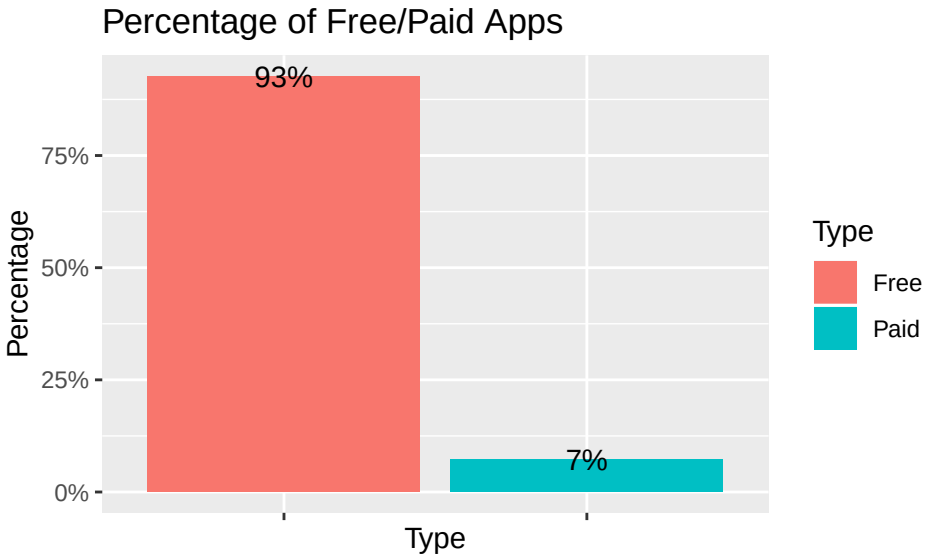
Overall, after data cleaning, the current total observations would be used in the following EDA and models are 4234.

Also, mainly focus on the following variables:

1. Category: Family, Game, Tools, Fitness, Media.
2. Rating(Response variable): From 1.0 to 5.0.
3. Price: From $0.0 to $10.0
4. Type: Free or Paid.
5. Size: 1.0M - 100M
6. Reviews: From 1 to around 44 million.

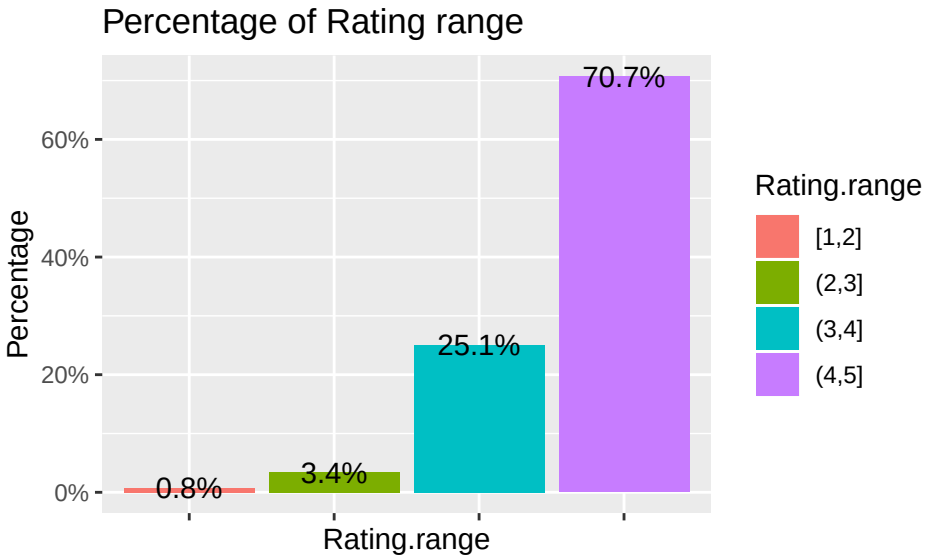
EDA:

Free vs Paid:



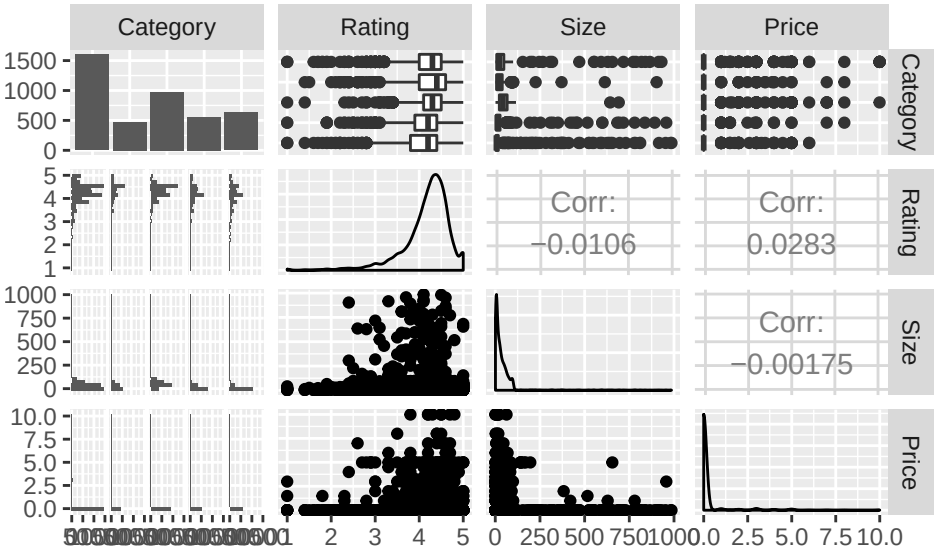
The ratio of free over paid is 3925:309, which is about 12.7 : 1. The size of `Paid` apps is very small compared to the size of `Free` apps, so it is hard and maybe insufficient to draw a solid conclusion about the influence of price on ratings if we build the model to describe the relationship only between Price and Rating.

Rating range distribution:



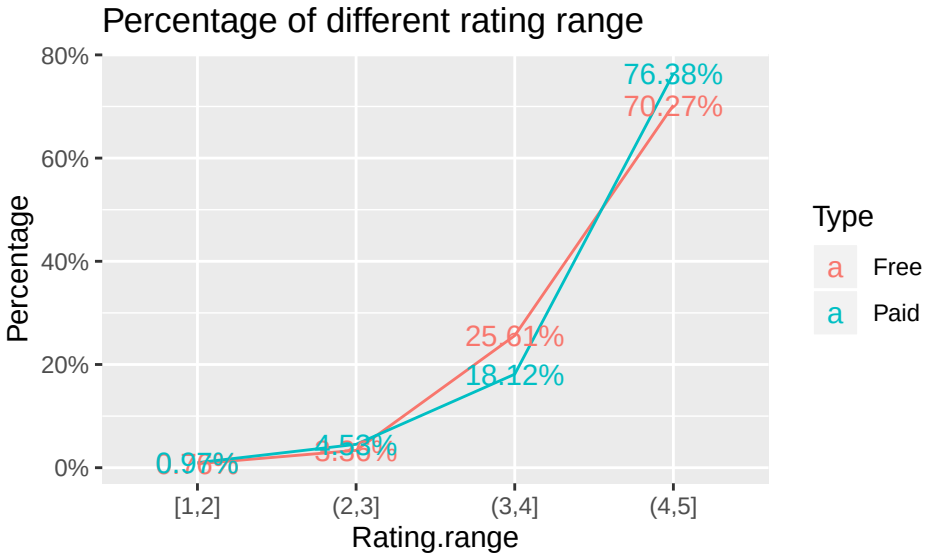
Most of the Apps have rating higher than 4, the distribution of Rating is left-skewed, and when we build the model later, most of the predicted value of Rating may be in the range from 4 to 5, as over 50% of the rating is higher than 4. It may be more clearly to show how rating changes with explanatory variables if limit the rating in a range from 4 to 5.

Rating explanatory:

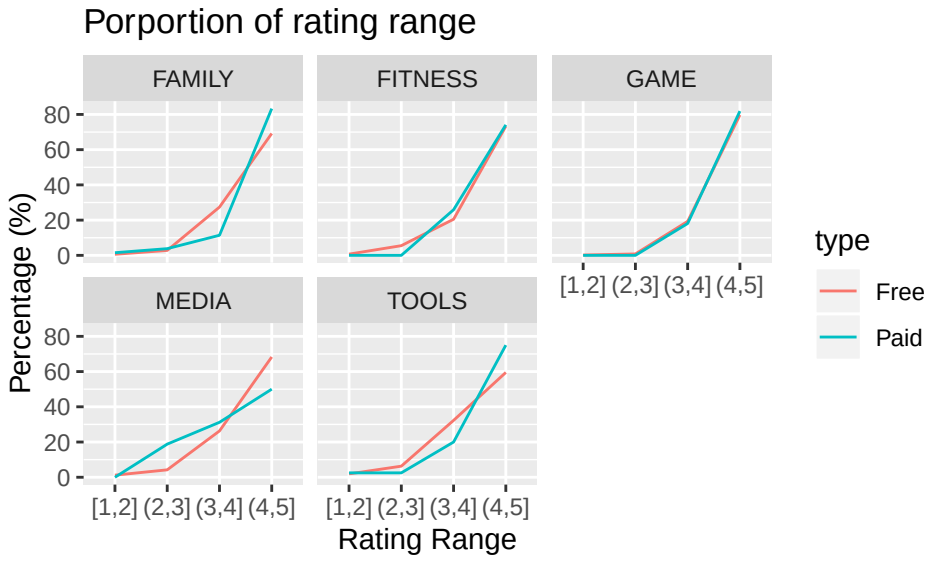


It's true that the Rating is left-skewed, Size and Price is right-skewed, but the Price is more extremely right-skewed. Thus, we will need to log Size and log Price in the following models.

Rating and Free or Paid:

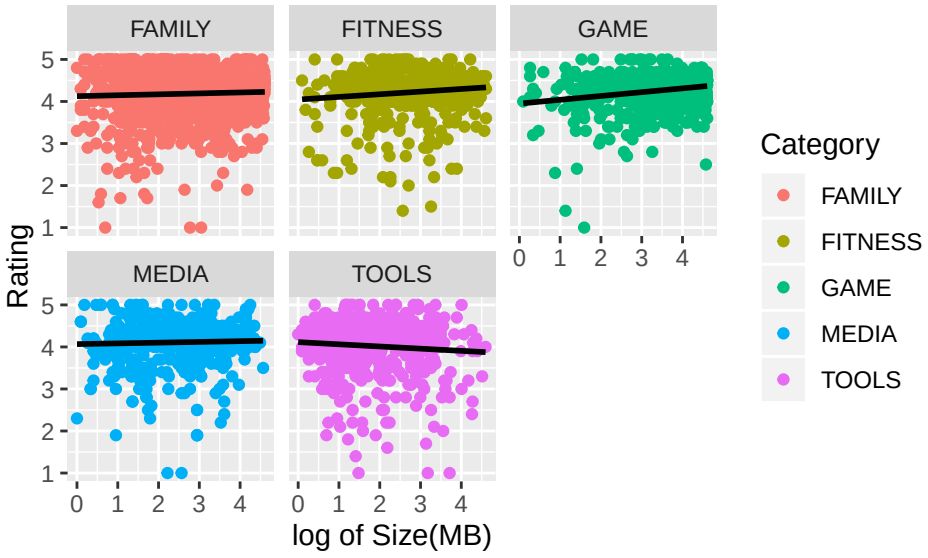


For paid apps, near 80% of them have a rating higher than 4, but for free apps, only 70% of them have a rating over 4. We could say that paid apps have a higher probability to have a rating higher than 4. However, this relation will be changed if we consider category.



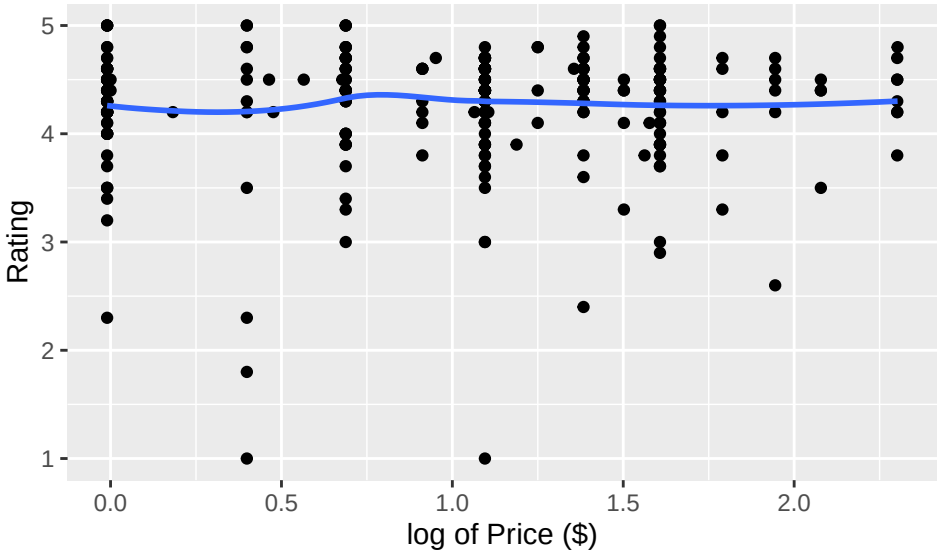
When we consider the category of apps, it's not always true that the paid apps have a higher probability of having a rating over 4. For Media apps, free one has a higher percentage of rating over 4, for game apps, the rating is not affected by the type of apps. So the interaction between category and type will be considered when we build the model.

Rating and Size:

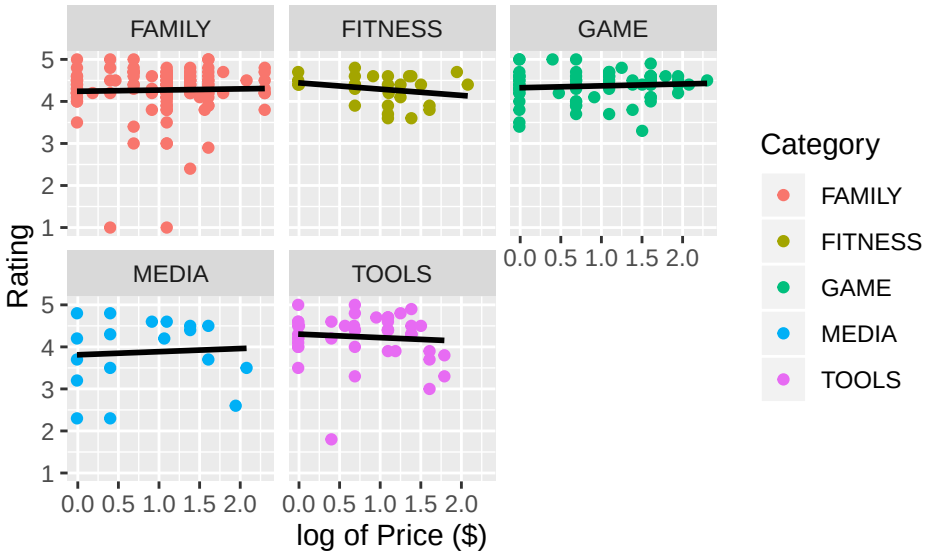


The size only affects Tools apps significantly, larger tools apps have a lower rating. While for the other four kinds of apps, the increasing of size seems to increase the rating slightly. We will take a look at how the size affects rating after we fitting the model.

Rating VS. Price:

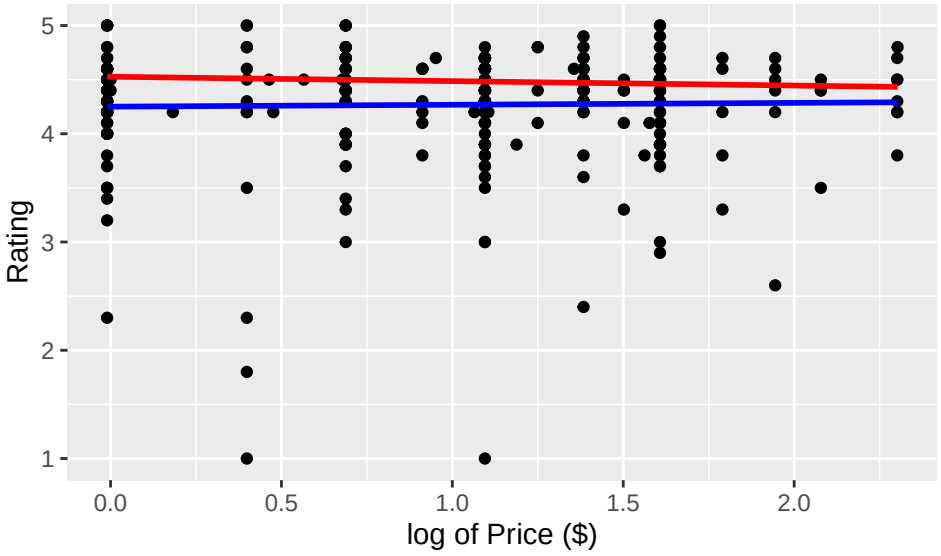


From the plot above, it seems that the price of apps does not have any relationship with the rating of apps, as the loess smoother looks almost flat. This is not what we expect to see, so in order to comfirm that whether Price has a relation with Rating, we group the Price by apps' Category to see whether the relationship between Price and Rating will be different.



It's clear that if we group the apps by their categories, the relationship between Price and Rating will be different based on their category. For Fitness, and Tools apps, their price has a negative relationship with Rating, the higher price will have lower Rating for these three kinds of apps, especially for Fitness apps, this negative relationship is stronger. However, for Family and Game, their Price has a weak positive relationship with Rating, the increasing of Price will slightly increase the Rating of apps. Only for Media apps, the relationship between Price and Rating is still weak or not existed. Thus, we also need to include the category in the model if we want to include the Price, it will be not sufficient enough if we only include the Price in the model.

Weighted by Review or not:



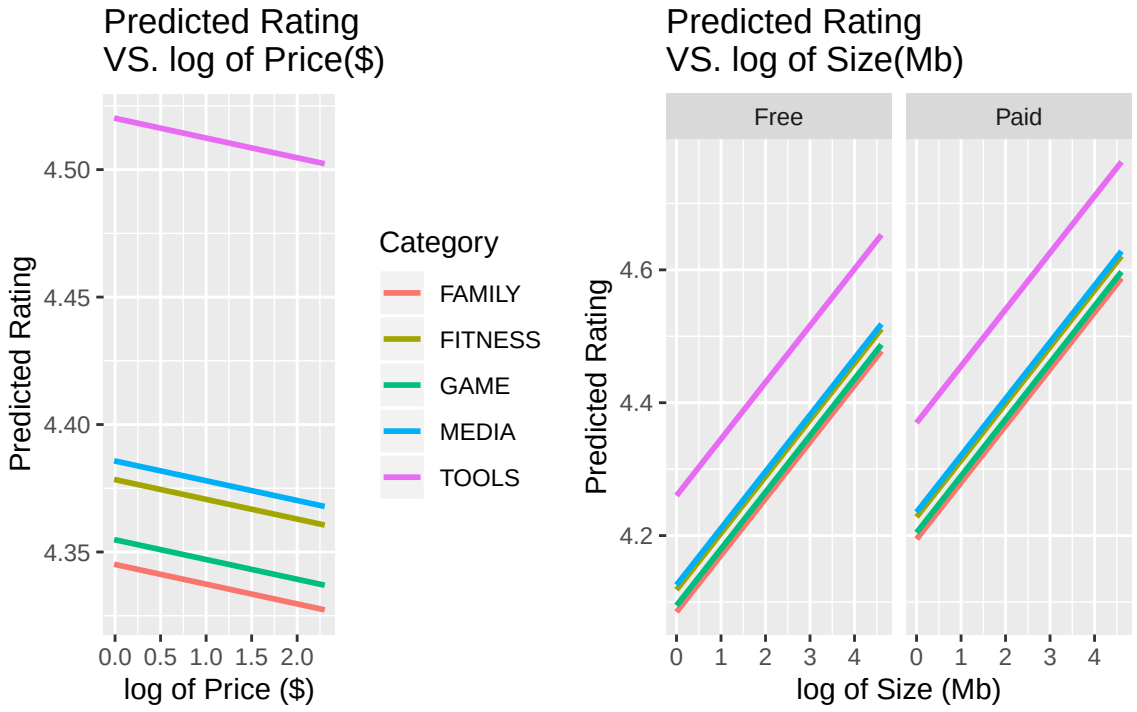
For the plot above, the red line is the weighted linear smoother, which is weighted by `Reviews`, while the blue line is not weighted. It's clear that the weighted (red) one will decrease as the price increase, while the blue one maintains the same value although the price increases. `Reviews` weights will affect the relationship between `Price` and `Rating` of apps, so we will use the model weighted by `Reviews`.

Model building and analyses:

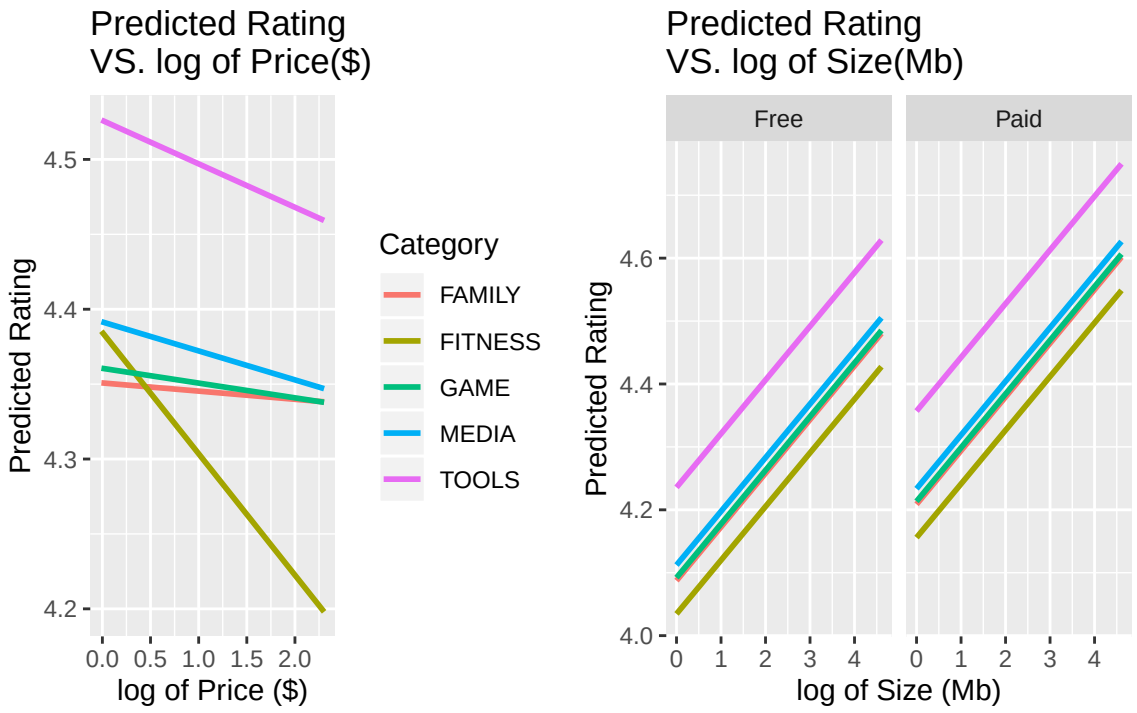
There will be a totally five linear models that we are going to build and compare. All five models will use the same set of explanatory variables: log of Price, log of Size, Type, Category, and response variable Rating, and weighted by Reviews. The **Null Model** will be the only one model without any interaction. **I1 Model** will be the model with the interaction between the log of Price and Category. **I2 Model** will be the model with the interaction between Type and Category. **I3 Model** will be the model with interaction between the log of Size and Category. Lastly, **I4 Model** will be the model with two interactions: between the log of Pirce and Category, and between the log of Size and Category.

Model building:

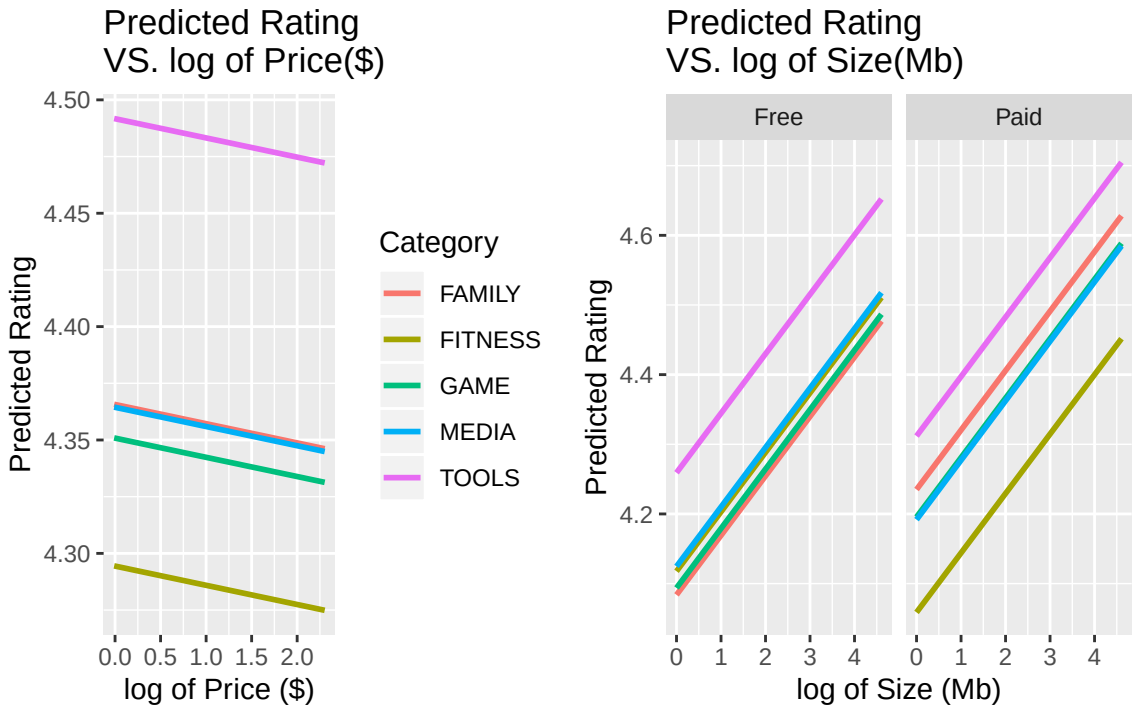
Null Model



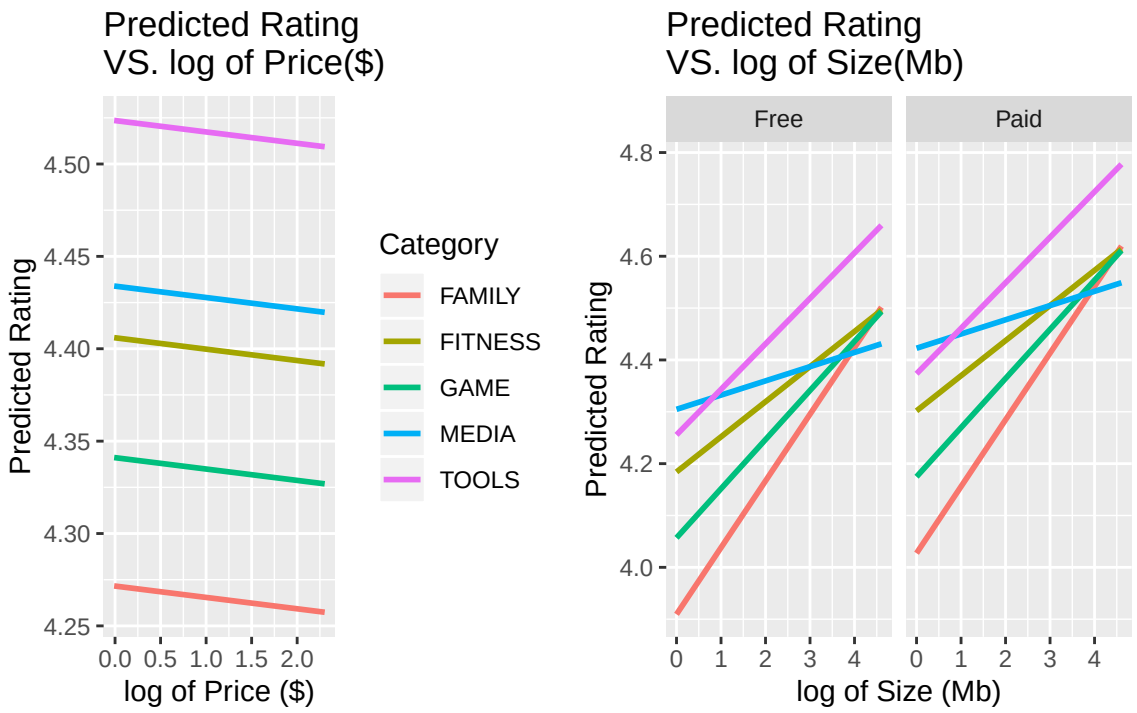
I1 Model



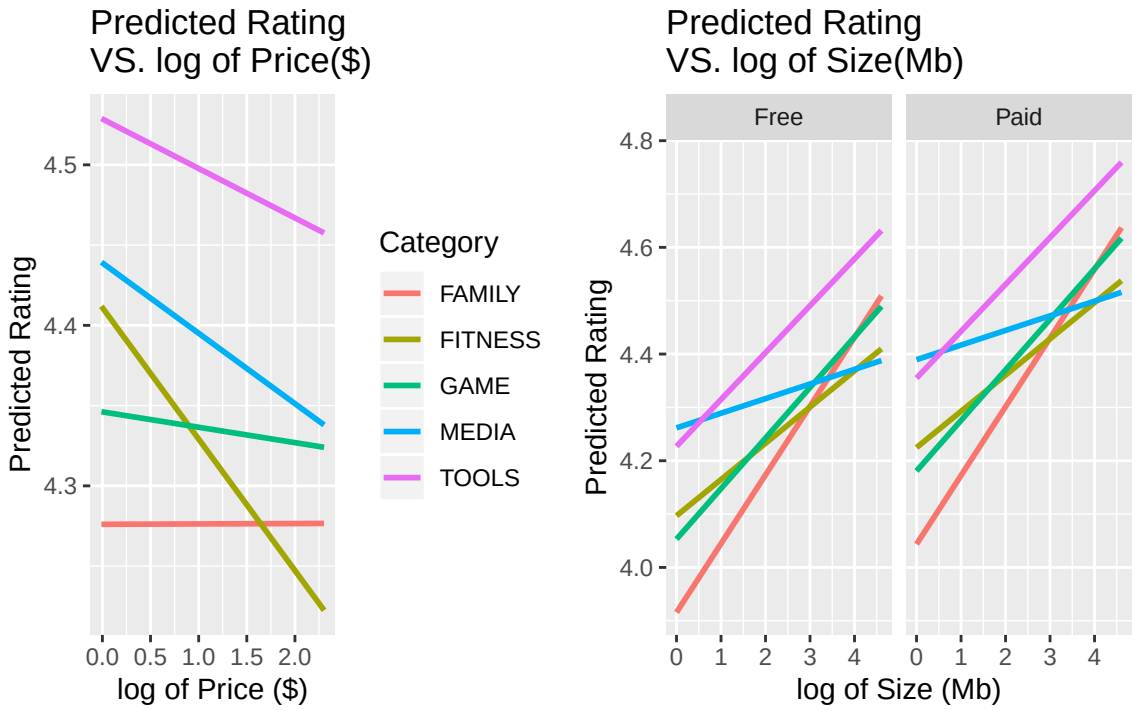
I2 Model



I3 Model



I4 Model



For the **Null Model**, we could tell, while the higher the `Price` of an app (exclude free app) the lower the `Rating` will be. For the `Size`, it doesn't matter an app is free or not, As `Size` goes larger, the `Rating` goes higher.

For the **I1 Model**, while the `Price` increasing the `Rating` still decreasing, however, for this model we use interaction between log of `Price` and `Category`, the slope of the `Price-Rating` plot changed dramatically that slopes become sharper compare to the Null model, but `log Size-Rating` stay similar, so the interaction between log of `Price` and `Category` is significant.

For the **I2 Model**, compare to with the Null model, we could see that there's no big change of their slopes of both of the plots, the tendency for both plot for both models looks very similar to each other, so that the interaction between Type and Category does not need to take into account.

For the **I3 Model**, compare with the Null model, we could easily discover that, the slop changed for the model adding interaction. It is very obvious in the plot `log. Size-Rating` that slope of five `Categories` changed dramatically that some of them become sharper than before. And slopes of the plot `Price-Rating` changed as well, they become flatter. Anyway, the interaction between the log of Size and Category we also need to take into account.

Lastly, for the **I4 Model**, compare to the Null model, we could find out that both plots' slopes changed so obviously which all slops becomes sharper that we need to take these two interactions model into account.

Comparison:

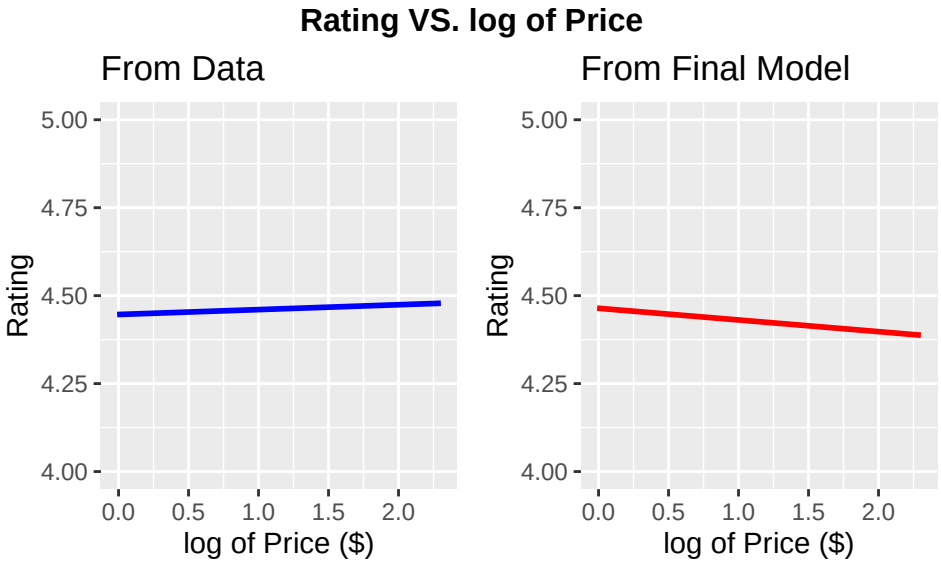
As the interaction in **I2 Model** is not sufficient, so we will not consider **I2 Model**. We will compare the `R-squared` from **Null Model**, **I1 Model**, **I3 Model**, and **I4 Model**, to see which one fits the data better.

From the table `R-squared Comparison`, **I4 Model** has the largest value of `R-squared`, so we decide to use **I4 Model** as the final model.

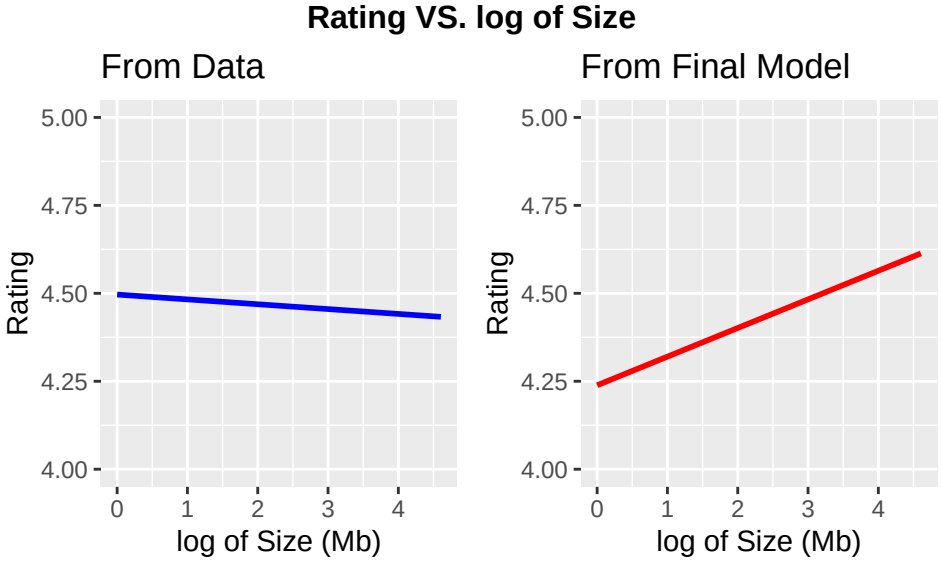
Then we take a look at how the Rating changes with the log of Price from the final model compared to that from data.

Table 1: R-sqaured Comparison

	Null Model	I1 Model	I3 Model	I4 Model
R-squared	0.1191549	0.1192676	0.1348074	0.1349587



By comparing with the relationship between the log of Price and Rating from data and from the final model, we could see that from data, the increase of price will slightly increase the rating of apps, while from the final model, the increase of price will decrease the rating of apps significantly.



From the plot above, we could tell that from the final model, the relationship between `Size` and `Rating` is stronger than that from data. For the final model, the increasing of Size will increase obviously the Rating of apps, for the data the increasing of Size will decrease the Rating of apps weakly.

Conclusion:

Based on all the analysis and model selection, the **I4 Model** is the most precise one to answer our research question: what kind of apps would have higher ratings. And the answer is those with lower prices as well as those with larger size are more likely to gain higher ratings. To predict the rating of apps in the google store, `Type`, `Price`, `Category` and `Size` are all crucial variables. We pick one null model and four alternative models which include interactions among variables in all. And the best model is the one with both the interaction between `Category` and `Size` and the interaction between `Price` and `Category`. Although `Category` is not as crucial as `Price`, it can not be eliminated from the model.

Being different from our original observations or hypothesis that the relationship between rating and price is positive, the final model pulls the relationship between rating and price downwards. Although the slope does not change a lot in absolute values, it means a lot to the `Rating` since it has a small range.

While there are some deficiencies we find which might weaken our conclusion. The first one is that the data set does not include sufficient paid app observations, the ratio of free over paid is around 12.7 : 1. Besides the data size, here is nothing related to fake ratings in this data set. Although fake ratings and reviews violate the Google Play Developer Program Policies, it is hard to remove all of these fake ratings from the store. In this data set, most of the ratings are from 4 to 5, which is relatively high. There might be some fake ratings but we have no evidence about this. What's more, the R squared for all our models ranges from 0.12 to 0.135 which is really low.