Data Science Project with App Store Dataset

Harshdeep Kahlon

12/4/2019

1. Abstract

The dataset found here contains more than 7,000 applications with details that are currently listed on the Apple iOS App Store. By investigating this data, we can potentially find interesting observations that could assist companies and developers to further grow their mobile applications. This project is mainly designed to find the areas of an app that should be optimized to promote the overall success of an app.

2. Problem Descriptions and Objectives

Data Analytics is used everywhere to analyze website and application performance and trends. The Apple App Store, which is one of the largest digital storefronts in the world, does not provide a public API that can be used to find any app's information (total downloads, size, genre, etc.). However, the App Store's app database is able to be scraped to find such information. The goal of this project is to analyze the App Store data and discover if genre, price, and/or content rating are vital in determining the success of an app. Once the attributes have been found, this project will try to classify whether or not a successful app needs a high user rating. This project will also explore the relationship between an app's size in megabytes and the genre of it. Finally, this project will compare free apps to paid apps ($0.99 and above) and find any major differences. The results of these analyses should help developers and teams find the

3. Data Description

- X - Position of each app in the dataset.
- id - Unique identifier for each app.
- track_name - Name as it appear's on the App Store
- size_bytes - Total downlaad size in bytes
- currency - The primary currency type of each app
- price - Total price in USD
- rating_count_tot - Rating count for all versions
- rating_count_ver - Rating count for its current version
- user_rating - Average rating value for all versions
- user_rating_ver - Average rating value for its current version
- ver - Each app's latest version identifier
- cont_rating - Content rating. Ex. 9+, 12+, 17+
- prime_genre - Primary genre in the App Store
- sup_devices.num - Each app's number of supported devices
- ipadSc_urls.num - Number of screenshots allowed to display
- lang.num - Number of supported languages
- vpp_lic - VPP device-based licensing is or is not enabled

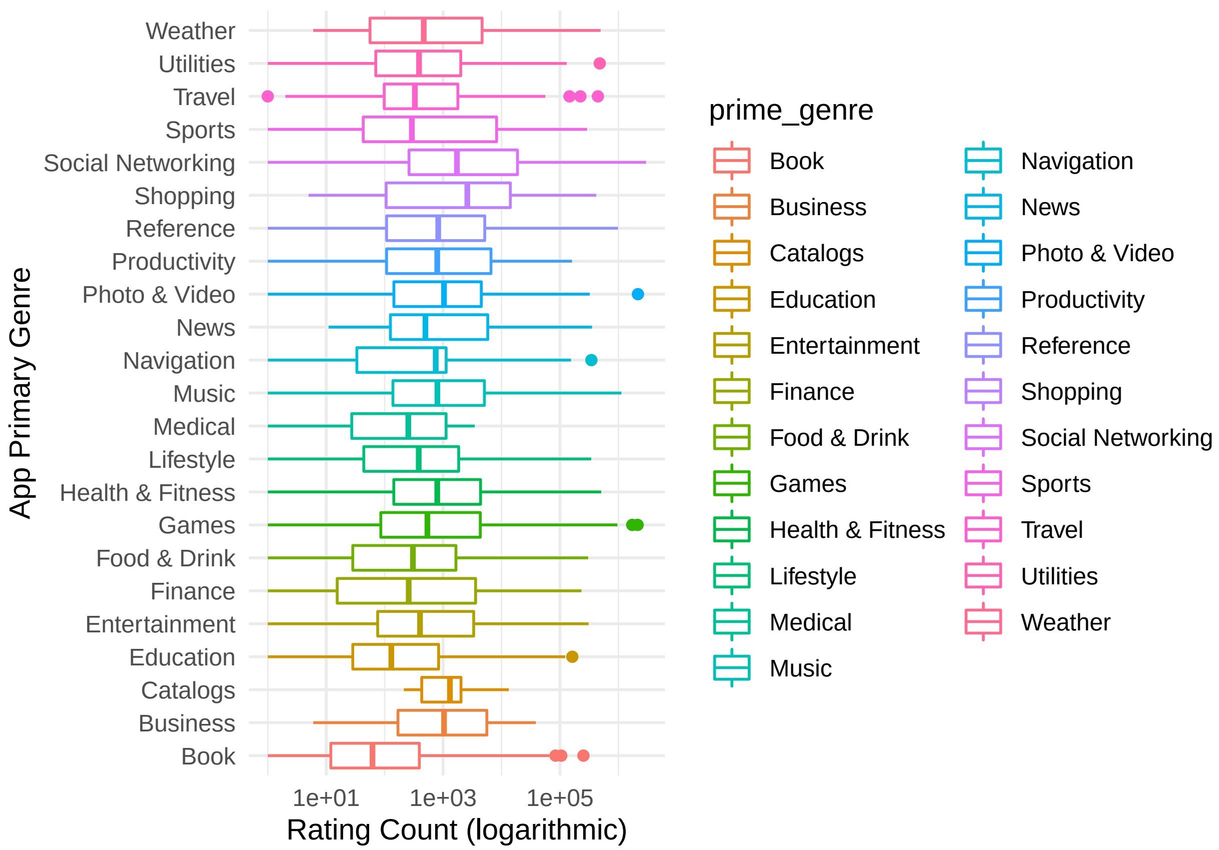
4. Loading and Formatting Data

I first imported the App Store Dataset and found some useful variables to be missing. I created "size_mb" to measure the each app's size in megabytes rather than bytes. I created "is_free" as a Boolean expression to see if each app is free or not. I created "user_rating_string" since there were only 10 possible values for each app's user rating.

```
appstore_df <- read.csv("../Data/AppleStore.csv")
appstore_df <- mutate(appstore_df, size_mb = size_bytes/1000000)
appstore_df <- mutate(appstore_df, is_free = price == 0)
appstore_df <- mutate(appstore_df, user_rating_string = as.character(user_rating))
appstore_df <- appstore_df %>% arrange(X)
```

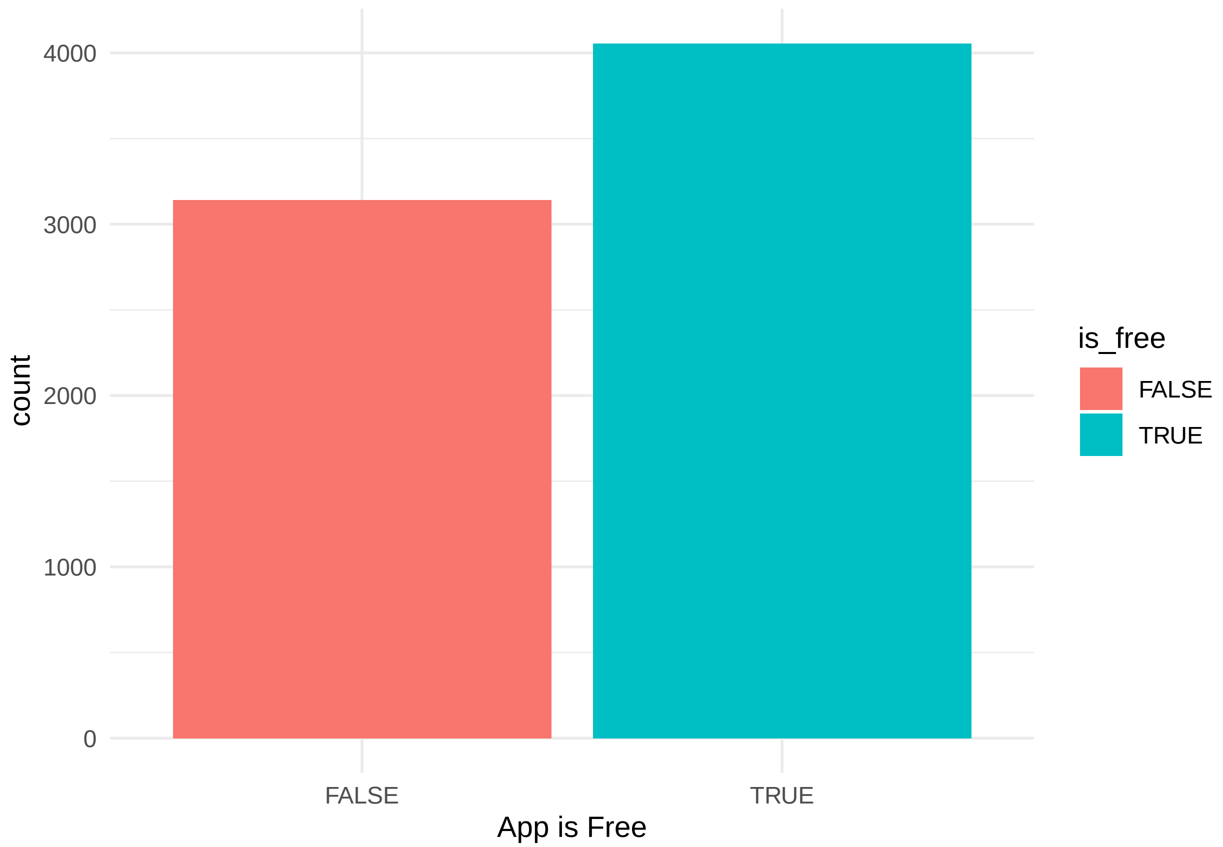
5. Data Visualization and EDA

Figure 1: Rating Count for Each Genre



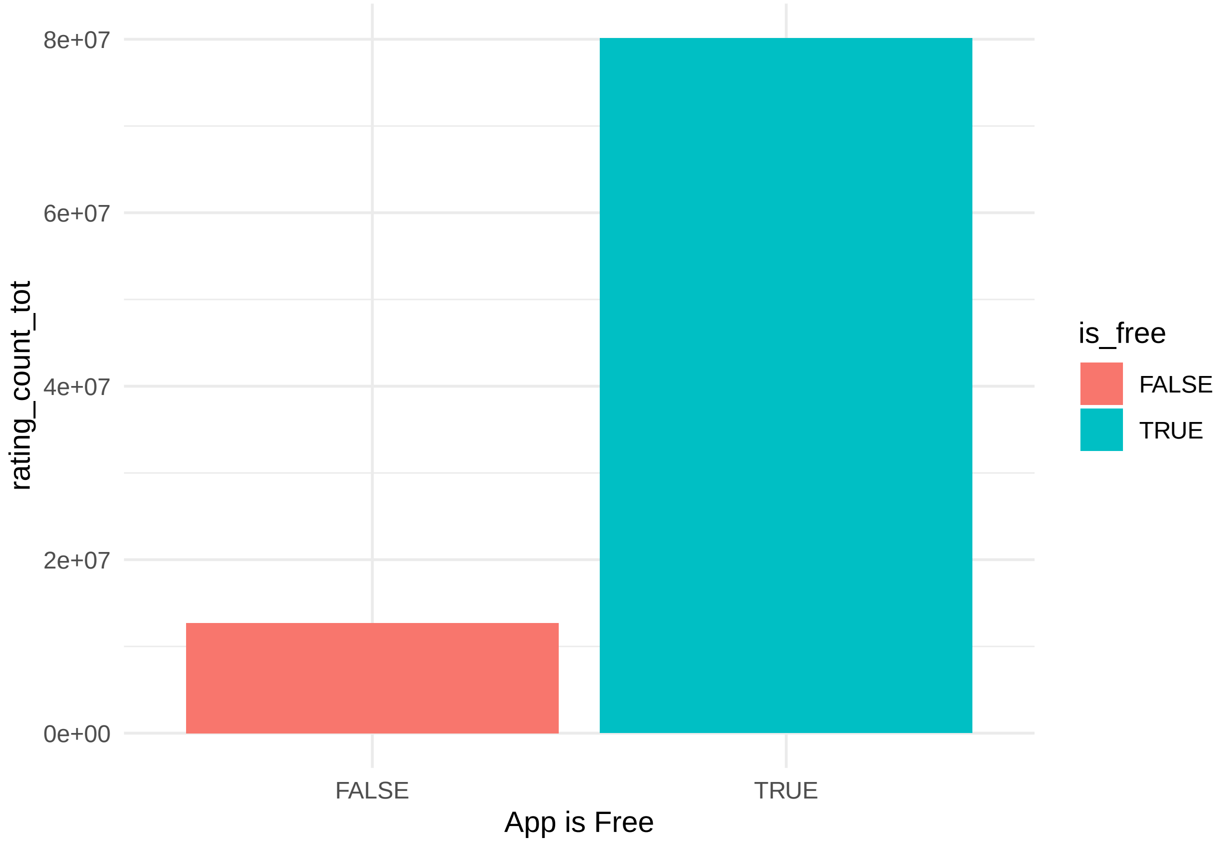
Since a total download count is not available in this dataset due to restrictions from Apple, we will treat the total rating count of an app to be a rough estimate of its total downloads. A higher rating count implies that the app is more successful. We can see how the popularity of every app genre with the graph above. The app genre plays a minor impact in the success of an app (exception is the Book genre).

Figure 2: Number of Free Apps Compared to Paid Apps



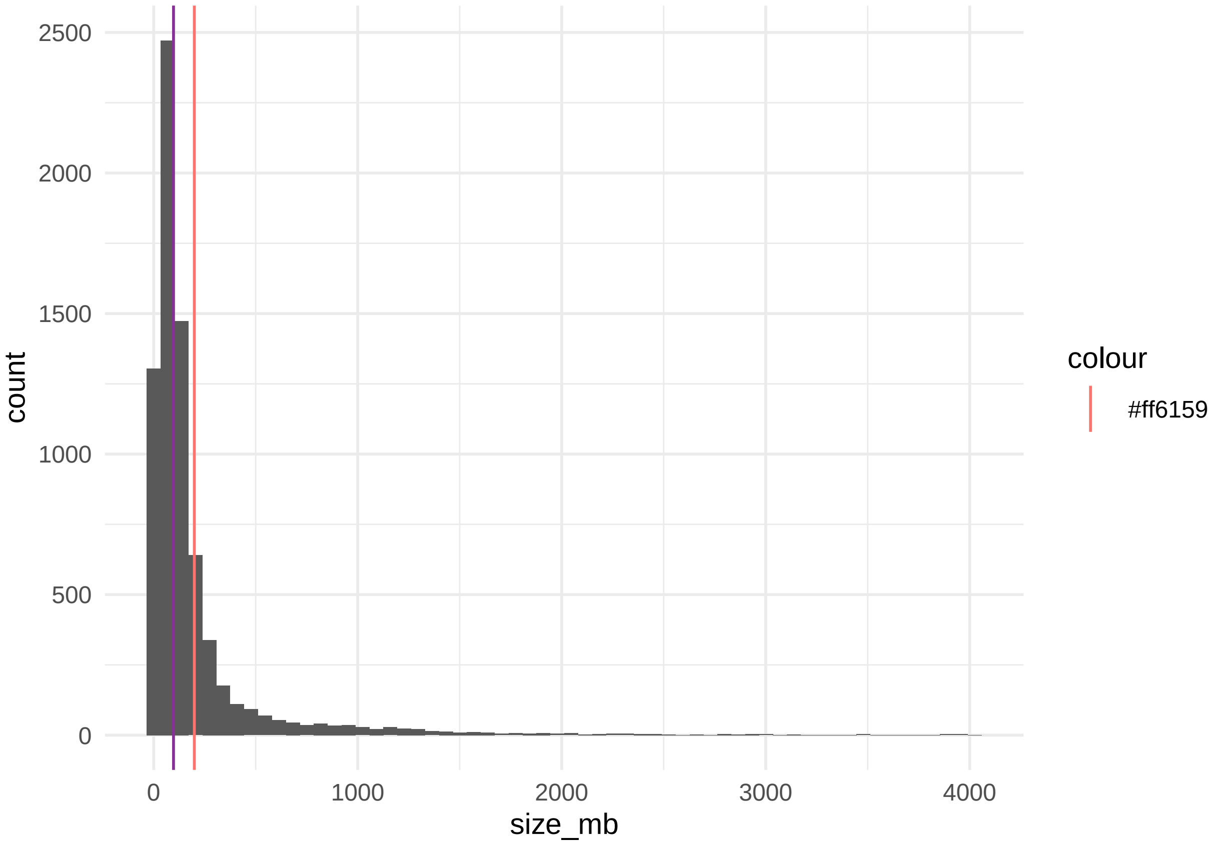
We can see that there are more free apps in the Apple App Store than paid apps, which can range from \$0.99 to \$299.99.

Figure 3: Number of Free Apps Compared to Paid Apps

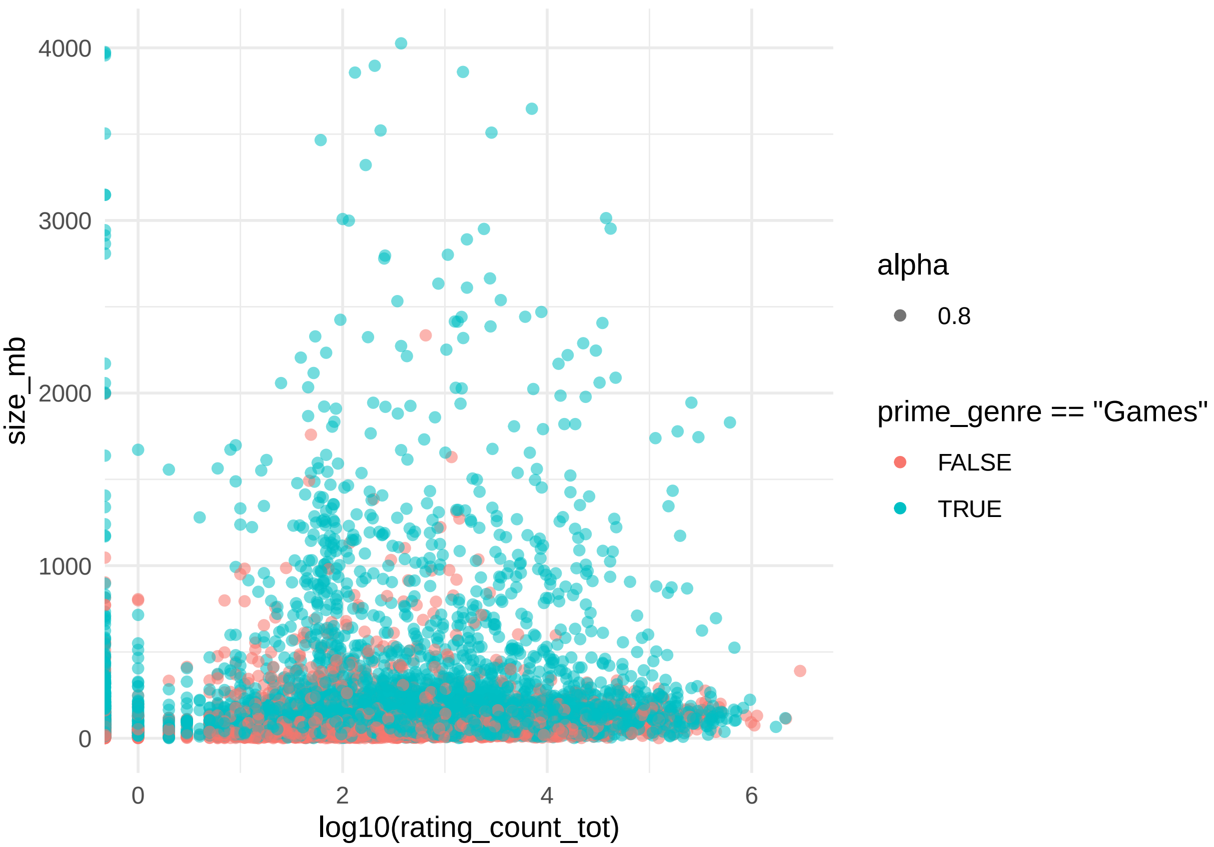


The total rating count, which implies the total download count, of free apps are much higher than paid apps. This ratio is significantly higher than the number of free apps to paid apps, which implies that free apps are typically more successful in the App Store.

Figure 4: Download Size (MB) and Total Rating Count

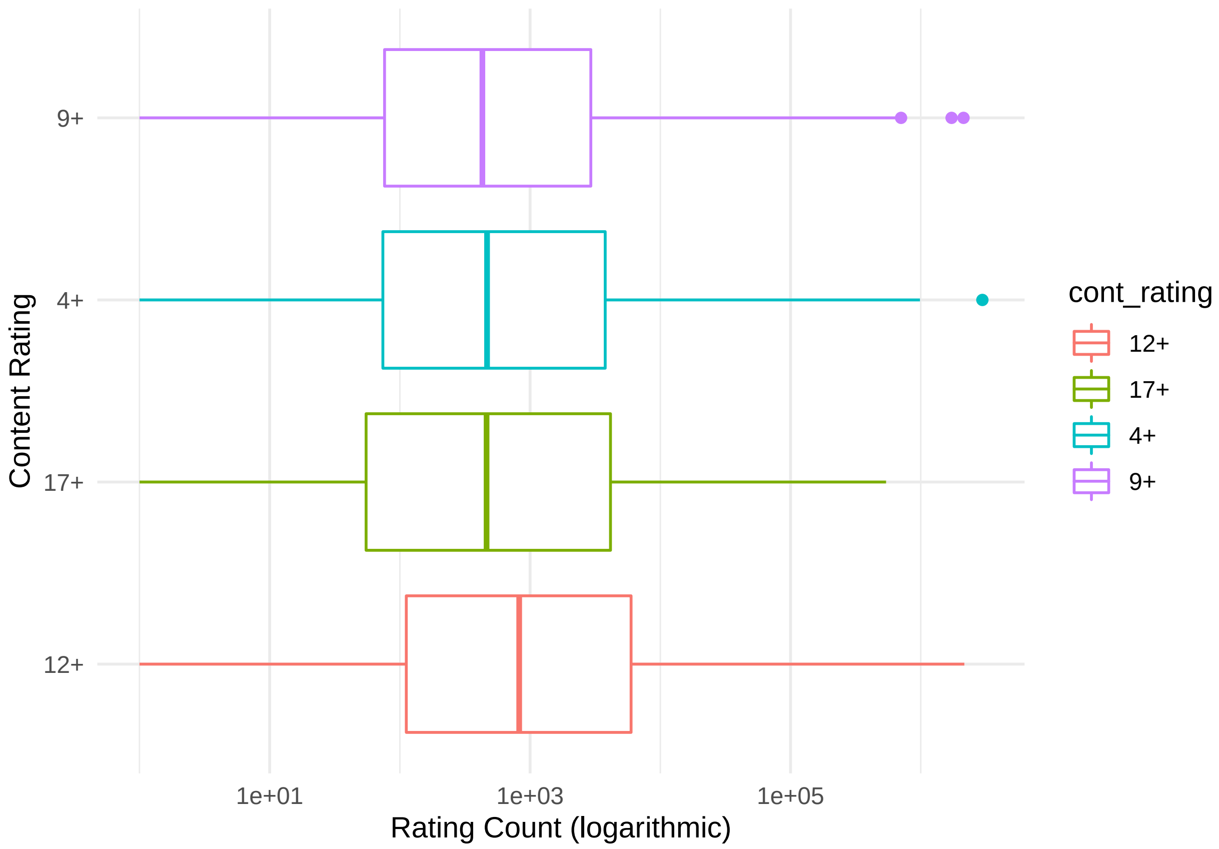


We can see the distribution of the app sizes throughout the ~7000 apps in the data set. The red line represents the mean of the app sizes in megabytes, while the purple line represents the median of the app sizes in megabytes. We can see that most apps are under the 500 MB threshold.



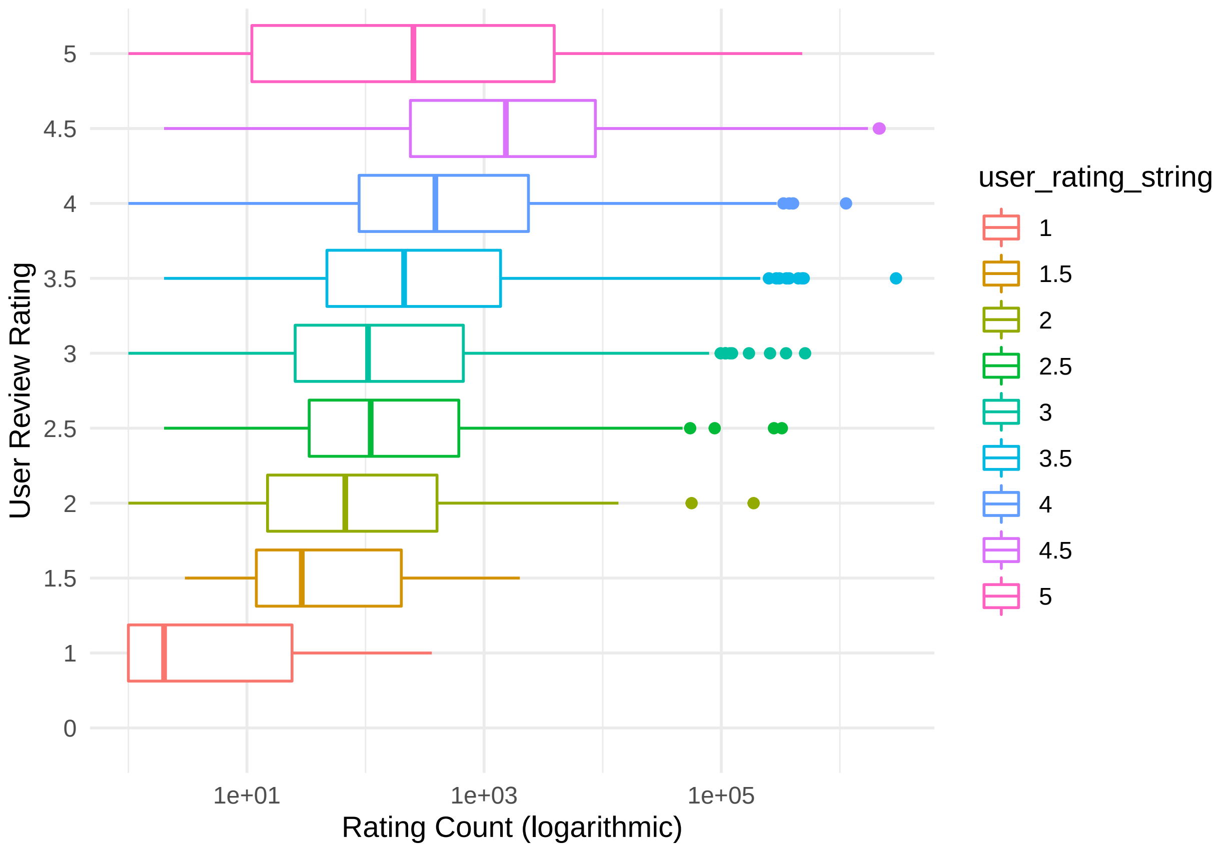
This scatterplot outlines many aspects about the data that we're investigating. The most striking aspect is that most of the apps in the dataset fall into the Games category, as you can see with the blue coloring. Also, most of the apps with greater than 1 GB download size are in fact games. Every other app category mostly falls under the 1 GB download size. In addition, the super-popular apps to the far right of the graph typically do not exceed the 2 GB threshold.

Figure 5: Relation Between Content Rating and Total Rating Count



We could see that content rating plays a very minor roll in determining the success of an app.

Figure 6: Relation Between User Review Ratings and Total Rating Count

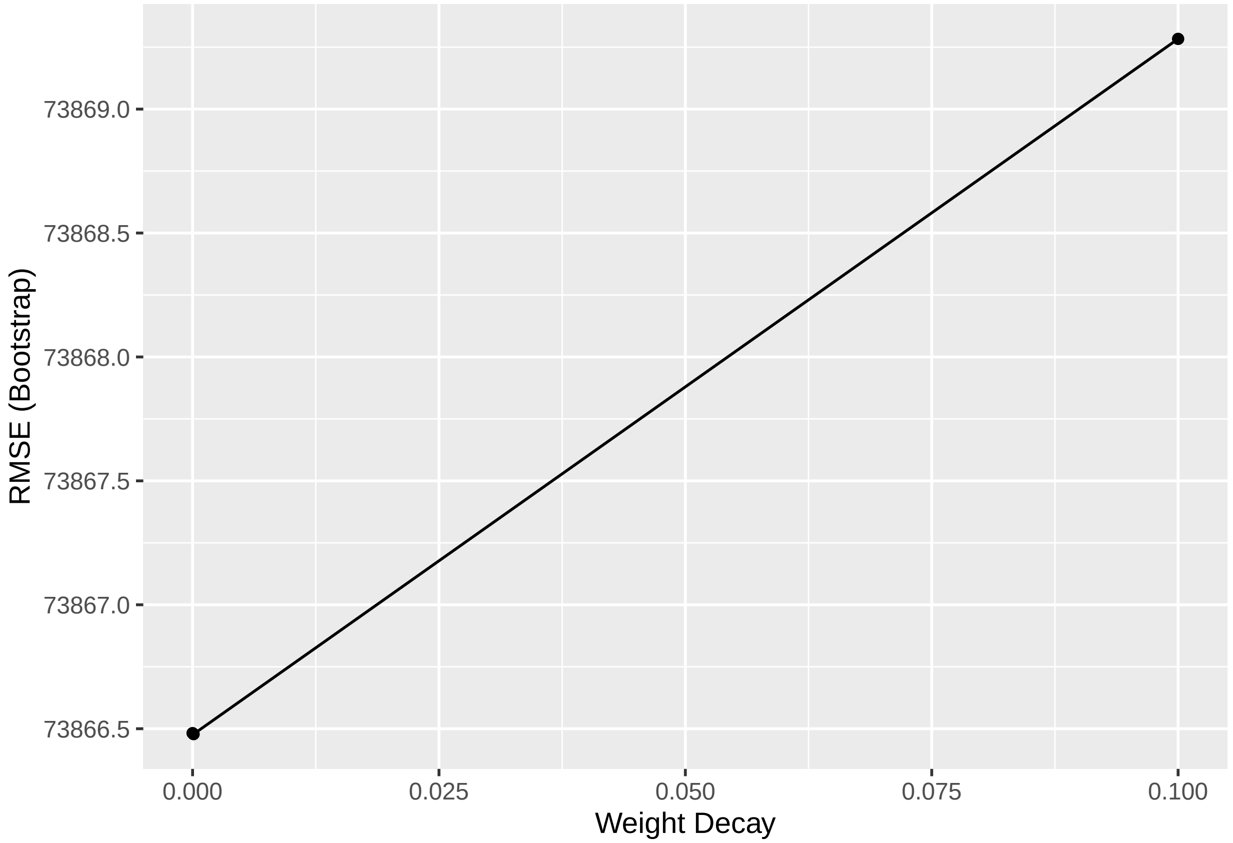


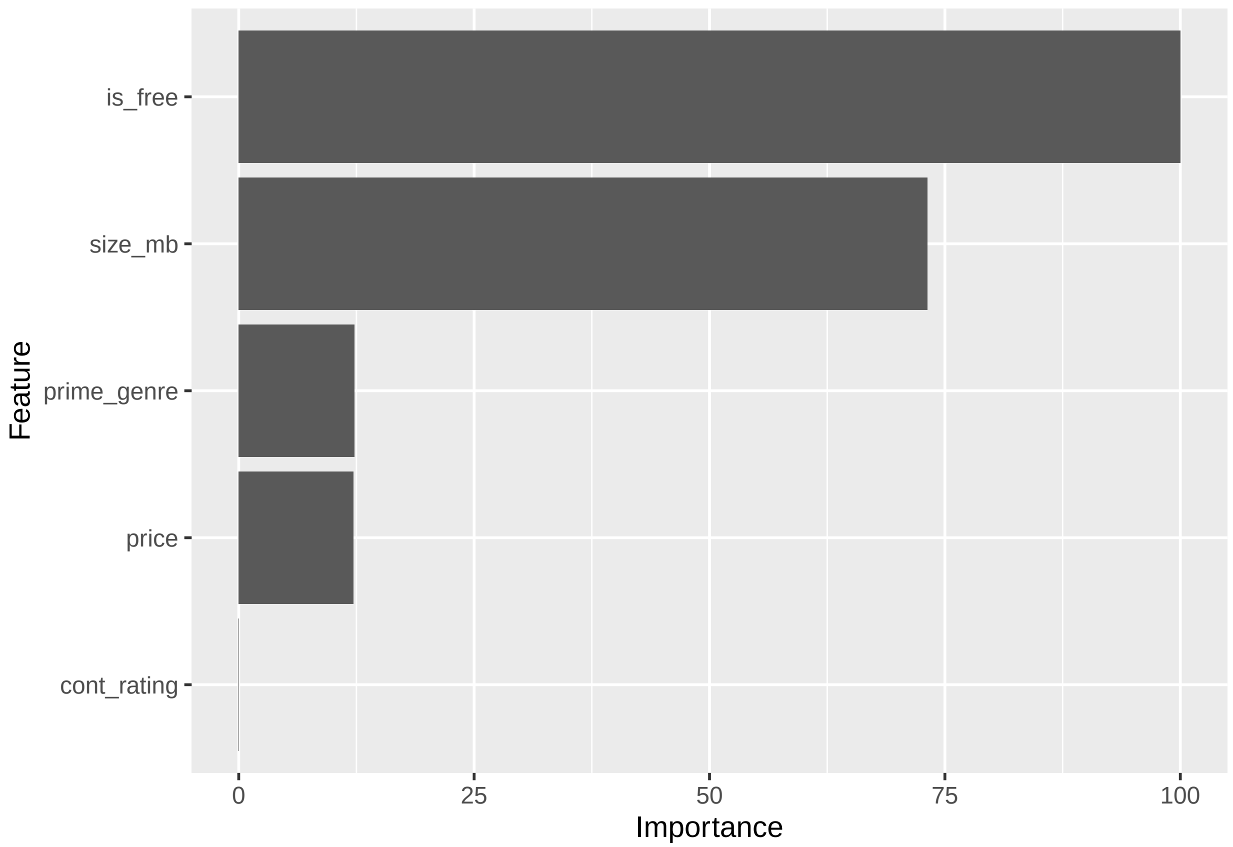
From the boxplot above, we can see a clear growth in total ratings (which gives a rough indication of total downloads) until the user ratings goes above 1.5 stars. However, from 2 - 5 stars, the difference in rating count cannot be determined from the graph.

6. Modeling

Two models were created for this dataset. We will first analyze the Ridge Regression model.

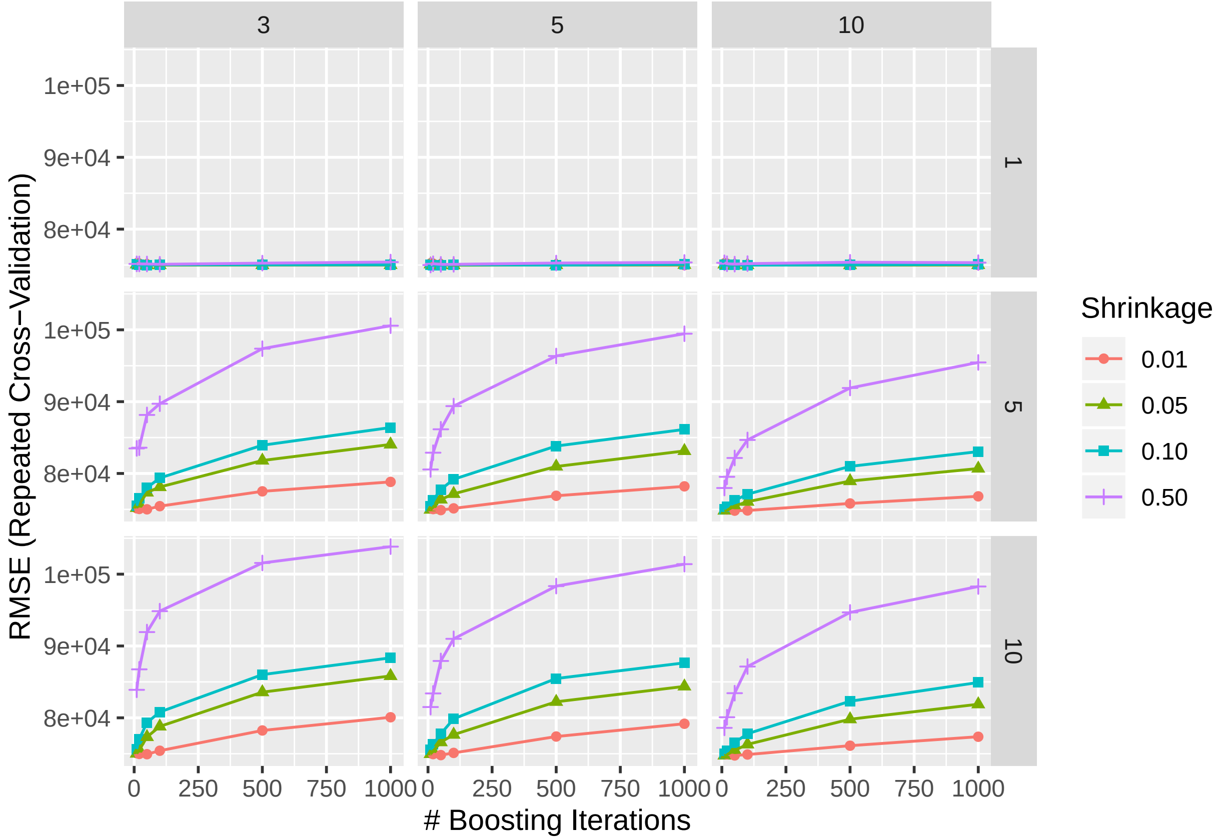
Ridge Regression

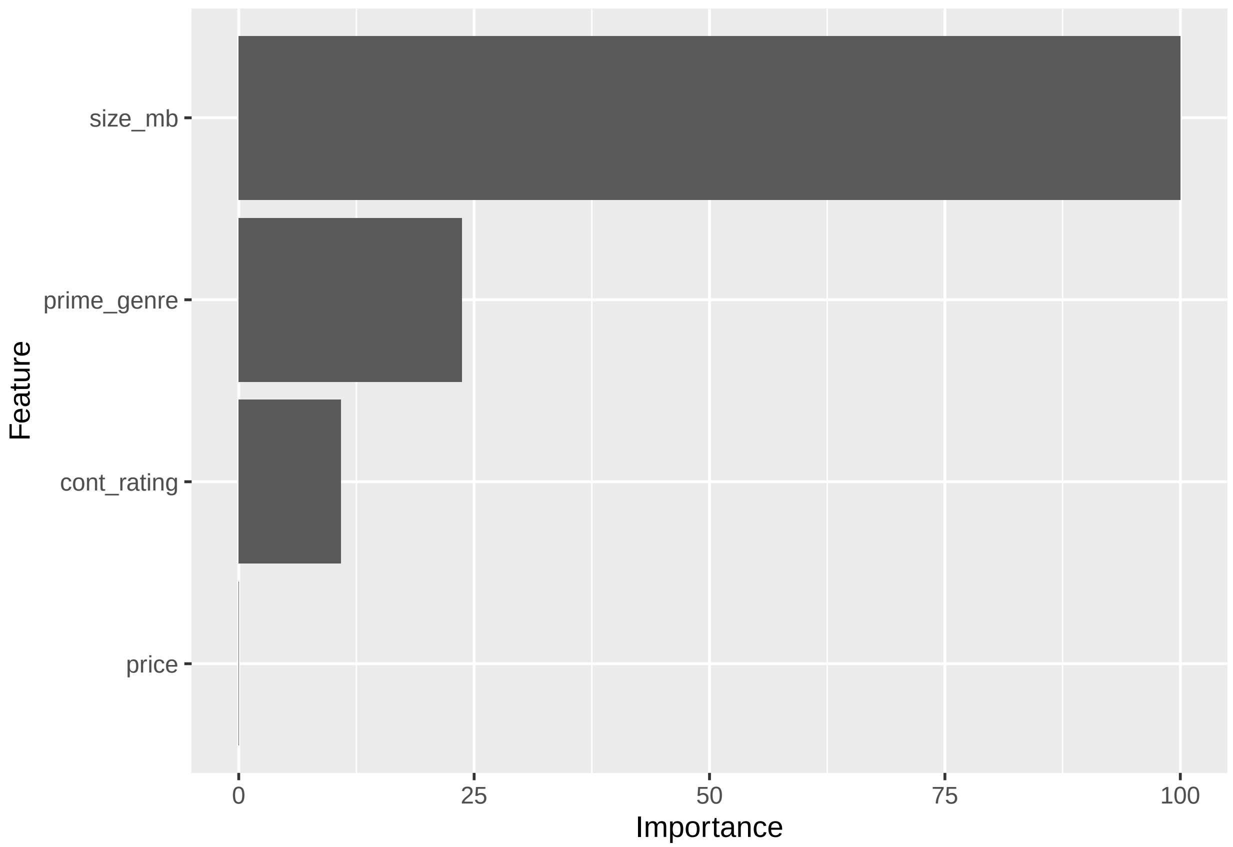




According to the Ridge Regression variable importance plot, the most important factors in determining an app's success (total rating count) is if the app was free or not. Second is the app's total size in megabytes. The actual price of the app has a slight affect of the overall rating count if the app is paid. The prime genre also has slight importance. However, the content rating of the app is not important at all in determining the overall rating count.

Gradient Boosting Machines





The GBM variable importance plot does not contain if the app is free or not, but does contain the rest of the predictors. The most important variable in this is the app's total download size. The prime genre of the app has slight importance as as well. Where this differs from the Ridge Regression model is the importance of content rating and price with both variables having swapped postiion from the Ridge Regression model.

7. Conclusion

After exploring and analyzing the Apple App Store dataset that contains over 7,000 apps, we have discovered what variables affect an app's success (determined by its total rating count) the most. The most impactful variables are if the app is free or not and the total download size of the app. We have also found that the content rating (age rating) and primary genre of an app do not affect the app's success in a significant way. With these findings in mind, developers can use this information to publish apps that have a higher potential to succeed. Although the dataset was based off of 2017 data, it should still provide some relevance to today's App Store markets.

References

[1] Ramanathan. Mobile app statistics (Apple iOS app store), 2017. https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps/kernels,Last accessed on 2019-12-04