Final Project

App Store Games - Summary

Dasun Wellawalage

November 12 2019

My goal for this project was to analyze a data set of 17000 games on App Store. The original data set consisted of 18 columns which included some useful categories such as the user rating. rating count, game genres, release date, languages, developer etc. Out of those, I chose 6 columns which I thought would have the most impact on game. They were, average user rating, user rating count, price, genres, developer, and age rating. Specifically, I was trying to measure what impact some of these variables have on the average user rating. A primary assumption was made that average user rating of a game is defines how good or successful a game is. Here are some interesting facts that I observed during the initial phase of data exploration.

- More than half the games (only 7561) didn't have a rating
- Most of the games had a very low number of users who rated it. Only 2884 games had a user rating count more than 100
- About 80% of the games were free. Some them had in-app purchases though
- Strategy, puzzle, action, entertainment, simulation, and casual were top game genres
- more than 60% of games had an age rating of 4+

I chose average user rating and price as my two main variables for most of the calculations and comparisons. Plotting the PMFs initially looked like there was not much difference between the free games and priced games. However, when the differences in PMFs were plotted, there was a significant difference in the games that had a rating between 3.5 and 4.5. A vast majority of games in that category were priced games. Plotting the CDF showed the portion of games that different ratings occupied. After doing the regression analysis it seemed that there was no strong relationship between the price of a game and the user rating. Following that I performed the same tests on user rating and Genres. Even that didn't show a significantly strong relationship.

I feel like I couldn't really find the combination of variables that had the most impact on a game's success. Although, price and genres didn't have such a strong correlation with the user rating, when I performed a multiple regression using the "developer" category as well (two explanatory variables), it produced good results. However, it would have been difficult perform these testing with a categorical explanatory variable. I also feel that I could have used the "release date" and "current version" variables too, to get a better understanding on their impact on a game. One of the biggest challenges I faced was, interpreting the final results, once I used a new numeric variable which was made by assigning weights to genres column. Although it produced results, interpretation wasn't that easy. Finally, I found it hard to fit an analytic distribution model to the data set. Justifying the use of an exponential model or a normal probability model was something I didn't fully understand.