Wenjun Yang

INFO201 - Project Proposal

GitHub repository: <https://github.com/Katherineeeywj/INFO201---Apple-Store>

(Sorry for turning in this assignment late. I had trouble connecting with my group members.)

1. Project Description

The dataset I will be working with is a dataset called “Apple Store” which I found on Kaggle, and it is extracted from the iTunes Search API at the Apple Inc website in July 2017. This dataset provides contains more than 7000 Apple iOS mobile application details, including their price amounts, primary genres, user rating counts, numbers of supporting devices, numbers of screenshots showed for display, etc.. The target audience will be mobile app marketing teams who want to develop marketing strategies in order to drive growth of future users and rating values. Investigation on factors that affect people to download app will serves as fundamental basis for their strategic decisions. This project aims to visualize variations in user rating values over different product groups (by size, by price, by genre, by number of supporting devices, etc.). Also, by constructing regression models, this project will provide prediction on user rating value based on specific circumstances.

1. Technical Description

This project will answer three questions:

1. Which type of product receives the highest rating?
2. How does number of screenshots showed for display affects rating? Negatively or Positively?
3. Based on this dataset, which combination of factors (price, genre, size, number of supporting devices, number of screenshots showed for display, number of supported languages) will receive the highest predicted rating?

The dataset will be read in as a csv file and the primary libraries that will be using are dplyr, glmnet, and ggplot2. It is necessary to reformat data before analyzing. For example, instead of leaving app sizes as their precise sizes, it will be more convenient and understandable to fit them into categories such as apps less than 100MB, apps between 100MB and 200MB and so on.

Apart from building up codes, one major challenge for this project will be understanding regression models. The ability to correctly interpret statistics results from models is required in the project.