Google Play Store App Analysis

Data Clean Up

Najja Osiomwan

November 18, 2020

Overview

The purpose of this project is to analyze the Google Play Store data, explore relationships between variables and determine which variables would be good predictors for Installs by using logistic regression. In this document I explain the methods I used to import the data and format it properly for further analysis. I found this data set on Kaggle.com and it is also available for download at:

https://github.com/NajjaOsiomwan/GPS/blob/main/Google-Playstore-Full.csv.zip

A summary of my process is noted below:

- I imported the dataset as 'g_apps', looked at a preview of the data head() and saw that there were 267052 rows and 15 variables. I examined the structure str() and summary summary() of the data and realized that the last 4 columns of data were not named. Further examination showed that there were misaligned rows in the data set. The overflow of data in the misaligned rows were the cause for the extra unnamed columns. By using the which() function I was able to identify which rows were misaligned and remove them from the dataset as well as the, now empty, last 4 columns.

- The variables *Size, Installs, Price, Last.Updated* and *Minimum.Version* needed to be formatted properly for further analysis. I used the gsub() function to remove symbols ($,+,M,k) and phrases ("and up"). I used the strptime() and as.Date() functions to format the *Last.Updated* variable.

- I converted these variables to factors as.factor() for the time being with the intention of changing the data types in the univariate analysis if necessary.