Project proposal

After 6: Angela Yoon, Ki Hyun, Sara Mehta, Tharun Raj

2020-10-22

Load packages & data

```r
library(tidyverse)
library(broom)
library(knitr)
library(patchwork)
library(kableExtra)
```

```r
apps <- read_csv("data/appData.csv")
```

Section 1. Introduction

In this project, we are investigating a dataset called "Google Play Store Apps." The dataset contains information on 9,660 unique applications (10,841 apps, including repeats) on the Google Play Store obtained using web scraping. Google Play Store is a digital distribution service operated and developed by Google, providing access to content, including apps, books, magazines, music, movies, and television programs. Play Store is Google's official pre-installed app store on Android-certified devices.[1]

The Google Play Store is growing at an extremely fast pace, and its popularity is not unknown over the Apple App Store. In fact, the number of Google Play Store apps surpassed 1 million in as early as July 2013, and is now placed at 2.87 million.[2] This makes this data not only extremely interesting and relevant to look at, but also highlights its enormous potential in gauging success in the app-making businesses.

With all four of us team members being extremely interested in application development and some of us being Android users, we are interested in exploring the different factors of a given application on Play Store that relates to the app's number of downloads. As per Forbes, an average app download brings in around 2 cents to its developer on Androids.[3] Thus, the average number of downloads for an app is a good measurement we can use to get a sense as to what an average developer can expect to make on an app today, which is why it is a very important measurement for developers that is worth looking into.

There is never a dull moment when it comes to technology, in that it continues to thrive in our technologically advancing world. However, it is difficult to study what links to an application getting more downloads, because various elements could affect an application being downloaded. Moreover, there is immense research on techniques that can be used to get the first 100,000 downloads for an application. As per Mobisoft, a top IT outsourcing company specializing in product development, a top IT outsourcing company specializing in product development, custom software development, mobile app development, and more, getting ratings and reviews as well as strategically pricing the app can all be techniques to get the first 100,000 downloads for an application, which can in turn lead to even more downloads and consequently the success of an app.[4]

[1] *Google Play*, Retrieved from https://play.google.com/store

[2] Sharma S (2020 May 6). Top Google Play Store Statistics 2020-21. *Appventurez.* https://www.appventurez.com/blog/google-play-store-statistics/

[3] Louis T (2013 Aug 10). How Much Do Average Apps Make?. *Forbes.* https://www.forbes.com/sites/tristanlouis/2013/08/10/how-much-do-average-apps-make/#7c1bfb4446c4

[4] Mobisoft Team (2015 July 14). Top 34 Techniques to Get First 100,000 Downloads For Your App. *Mobisoft.* https://mobisoftinfotech.com/resources/blog/top-34-techniques-to-get-first-100000-downloads-for-your-app/

Thus, this literature has inspired our investigation to delve deeper into the relation between such application features like ratings, size, and broad category, and the applications' number of downloads. These features are easier to grapple with.

All in all, actionable insights can be drawn for developers to work on and capture the Android market based on our results. Moreover, these results will be insightful for prediction of future applications' downloads as well.

With the above motivations and dataset in mind, we arrived at a research question to probe our interest on the relationship between the facets of an application and its number of downloads, for Google Play Store applications in the dataset. One thing that is worth noting is that the number of downloads in this dataset is recorded as a categorical data, with values of "1,000+", "5,000+", "10,000+", "50,000+", and so on. Because these values are not exact and the increments between the rounded values are not constant, we decided that converting this to a continuous variable would lead to inaccurate analysis results. Instead, we decided to separate these values into two bigger groups: "1 million +" and "less than 1 million."

At the proposal stage, we chose 1 million as a cut-off as a temporarily working number based on the above mentioned Mobisoft article, as well as guides such as the one by AppInventiv that illustrate one million downloads as a prominent, industry-standard goal for one's application.[5] After we do some Elementary Data Analysis and ask for advice and guidance from instructors, we will adjust the cut-off if needed and provide grounds as to why we chose such a cut-off.

The general research question our team will be exploring will therefore be as follows: "For Google Play Store Apps, to what extent do the aspects of the application such as ratings, genre, pricing, size, and other variables in the dataset relate to the odds of the app having more than 1 million downloads?"

For example, to answer this general question, we will ask questions such as but not limited to:

- Do applications that tap into a certain genre, such as news or games, receive more downloads than those in other genres?
- Are ratings even significant in relating to downloads?
- Is there a pattern between the number of reviews and the number of downloads?
- Does it matter whether an application is paid or free?

We will try and narrow down and identify the set of predictor variables that contribute the most to an app having 1 million + downloads.

We think apps in the dataset will have a set of characteristics that will be similar among them because we believe an app must be of certain quality and practicality for people to actively download it. Accordingly, our formal null hypothesis and alternative hypothesis based on this research question are as follows:

Null Hypothesis (H_0): None of the aspects of an app will have a relationship with the odds that it has 1 million+ downloads.

Alternative Hypothesis (H_a): At least one of the aspects of the app will have a linear relationship with the odds that it has 1 million+ downloads.

Section 2. Data description

- Observations:

There are 10841 observations in the dataset. Each row represents an application available in the United States in August 2018. Each case also contains 13 different variables. The 13 variables and their descriptions[6] are shown below.

[5] Srivastav S (2018 March 23). How to Get Million Downloads On Your App. *appinventive.* https://appinventiv.com/blog/get-million-downloads-app/

[6] Lavanya Gupta (Retrieved 2020 Oct 21). "Google Play Store Apps". *Kaggle* https://www.kaggle.com/lava18/google-play-store-apps

```r
column_names <- colnames(apps)
data_description <- data.frame(Variables = column_names)
data_description["Description"] = c("Application Name (available in the US, Aug 2018)","Category the app
data_description%>%
  kable(col.names = c("Variables", "Description"), align=rep('l', length(data_description[,1]))) %>%
  column_spec(2, width = "40em")
```

Variables	Description
App	Application Name (available in the US, Aug 2018)
Category	Category the application
Rating	Overall user rating of the app (up to Aug 2018)
Reviews	Number of user reviews (up to Aug 2018)
Size	Size of the application (in Aug 2018)
Installs	Number of user downloads/installs (up to Aug 2018)
Type	Paid or Free (in Aug 2018)
Price	Price (in Aug 2018)
Content Rating	Target age group (e.g. Children / Mature/ Adult)
Genres	Genre (multiple could be assigned to one application)
Last Updated	Date last updated on Play Store (until Aug 2018)
Current Ver	Version of the application available on Play Store (in Aug 2018)
Android Ver	Minimum Android version requirement (in Aug 2018)

- Response Variable:

The response variable in our analysis will be derived from the variable "Installs". This study focuses on exploring which set of characteristics are related with applications, that were available in the Unites States in August 2018, with one million or more downloads. Therefore, "Installs" would provide sufficient information to construct a response variable in this study.

In this dataset, "Installs" is given as a categorical variable, and each category indicates the maximum arbitrary threshold (set up in units of 5×10^n) individual app installation has surpassed. For instance, an app that has been installed more than 100 times but less than 500 times has the "Installs" value of "100+", whereas, an app that has more than 500 installs but less than 1000 installs has the "Installs" value of "500+".

From the "Installs" variable, we are able to construct a new variable to fit the purpose of the study. The new categorical variable will be a simplified version of "Installs" that only indicates whether each apps have been installed more than a million times or not. This two-level categorical variable constructed from "Installs" would be the response variable in our analysis.

- Predictors for Analysis:

The predictor variables that we are interested in looking at are related to several aspects of the application on the Play Store, i.e. not only the ones associated with the actual features of an application, but also the features of users' feedback for the application on the Play Store. Namely, `Category`, `Rating`, `Reviews`, `Size`, `Type`, `Price`, `Content Rating`, `Genres`, `Last Updated`, `Current Ver` and `Android Ver`. The description for these variables can be found in the table above. (* Note that this data was collected in August 2018, and are limited to the applications available in the United States.)

- Source:

This specific dataset was obtained from Kaggle.[7] The data itself was scraped from the Google Play Store. As mentioned, the data was scraped in August 2018 from NYC.[8] Therefore, the scraping would be specific to the geographical location and might not include the same apps if the scraping procedure was carried out from

[7] Lavanya Gupta (Retrieved 2020 Oct 21). "Google Play Store Apps". *Kaggle* https://www.kaggle.com/lava18/google-play-store-apps

[8] Lavanya Gupta (Retrieved 2020 Oct 22). "When was this dataset scraped?". *Kaggle* https://www.kaggle.com/lava18/google-play-store-apps/discussion/67452

another location. There could be other caveats from Google's end making the list of apps different when scraping is performed by different users.[9]

Section 3. Glimpse of data

```
glimpse(apps)
```

```
## Rows: 10,841
## Columns: 13
## $ App              <chr> "Photo Editor & Candy Camera & Grid & ScrapBook", ...
## $ Category         <chr> "ART_AND_DESIGN", "ART_AND_DESIGN", "ART_AND_DESIG...
## $ Rating           <dbl> 4.1, 3.9, 4.7, 4.5, 4.3, 4.4, 3.8, 4.1, 4.4, 4.7, ...
## $ Reviews          <dbl> 159, 967, 87510, 215644, 967, 167, 178, 36815, 137...
## $ Size             <chr> "19M", "14M", "8.7M", "25M", "2.8M", "5.6M", "19M"...
## $ Installs         <chr> "10,000+", "500,000+", "5,000,000+", "50,000,000+"...
## $ Type             <chr> "Free", "Free", "Free", "Free", "Free", "Free", "F...
## $ Price            <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", ...
## $ `Content Rating` <chr> "Everyone", "Everyone", "Everyone", "Teen", "Every...
## $ Genres           <chr> "Art & Design", "Art & Design;Pretend Play", "Art ...
## $ `Last Updated`   <chr> "January 7, 2018", "January 15, 2018", "August 1, ...
## $ `Current Ver`    <chr> "1.0.0", "2.0.0", "1.2.4", "Varies with device", "...
## $ `Android Ver`    <chr> "4.0.3 and up", "4.0.3 and up", "4.0.3 and up", "4...
```

[9] Lavanya Gupta (Retrieved 2020 Oct 22). "Why are there only 10K apps in google play?". *Kaggle* https://www.kaggle.com/lava18/google-play-store-apps/discussion/66148