Four ways to gain installation of andriod application in terms of rating, review, price and category

songyuan luo

22 December 2020

Code and data supporting this analysis is available at: https://github.com/songyuan-l/304ps5

Keywords:

Andriod, installation number, google play store, brm

Abstract

Several UOFT students decide to establish a startup company that design applications based on the Andriod operating system. Before we start to code a program, we want to decide what kind of application to build. Therefore, we will build a brm model to predict category, rating, reviews, size, type which will affect the popularity of the applications in the market.According to our model,rating,review, price and category has a major impact on the installation number of applications, so we have four method to increase the installation number.

Introduction

For the newly established company, the first and also the most important task is to figure out which kind of product we need to build. Nowadays, almost everyone has mobile devices, so the market for a mobile app is pretty large. Since we have a limited labor force, we have to choose which operating system will we build our application on. Andriod and IOS are the most popular operating system. Andriod OS shared 85% of the operating system market, while IOS has 13% of the market. Compare to Andriod and IOS, Android os has more consumers than IOS.(Nick Galov (2020)) Also, the apple store's application review mechanism is more strict than the play store. Therefore we decide to develop an android application first. There are many popular apps on the play store such as TikTok, Spotify, Facebook, etc. Our second task is to find out what kind of application we need to develop. Hence we find a dataset of Google Play Store Apps from Kaggle(Lavanya Gupta (2018)). As we are a newly established company, we don't care about how much profit we earn, we only care about how many people are willing to use our application.So the installation number will indicate if the application is popular or not.Also, if we want to find an investor, the number of the user plays a significant role in the level of investment.

In order to find the causal inference,We will fit a brm model to predict the influence of the APP's rating,review, price and category on the installation number.We will determine whether those factors have a positive or negative impact on the the installation number.By the model result, we will get some methods to increase our app'S installation number.

data

The data in use is from the Google Play Store Apps from Kaggle(Lavanya Gupta (2018)). In this study, we aim to find out factors that influence the popularity of the application. More specifically, what is the influence of different features on the installation number? In our case, we use these data to predict how popular the applications are: category, rating, reviews, size, type.

“Installs” is the total installation number of the application on the google play store “Category” shows which kind of application it is. for example BUSINESS, photograph, etc “Rating” is the overall review rate of the application on the google play store “Reviews” is the total number of reviewers of the application on the google play store “Type” shows the application is free or not

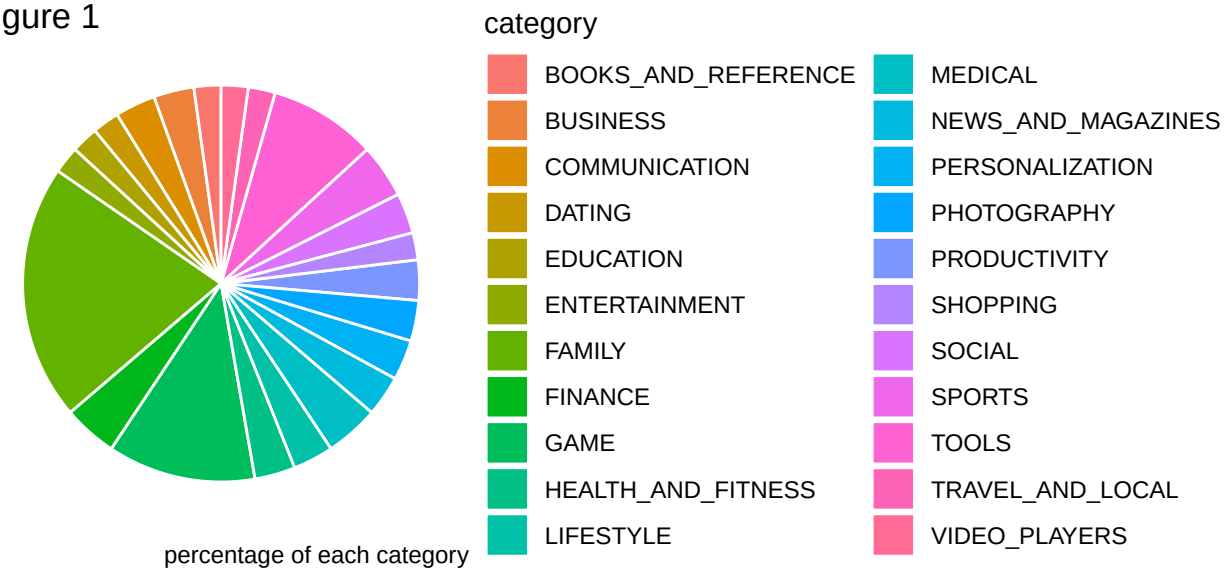
Among these predictor variables, we filter out all non-responses. For the type variable, we denote the ‘free’ as ‘0’, and paid as ‘1’.

The dependent variable for this study is set as the installation number. In order to study the influence on the installation number, we mutate the installation number “Installs” to numeric, for example,“1,000,000+” is convert to 1000000. Lastly, we removed all missing values.

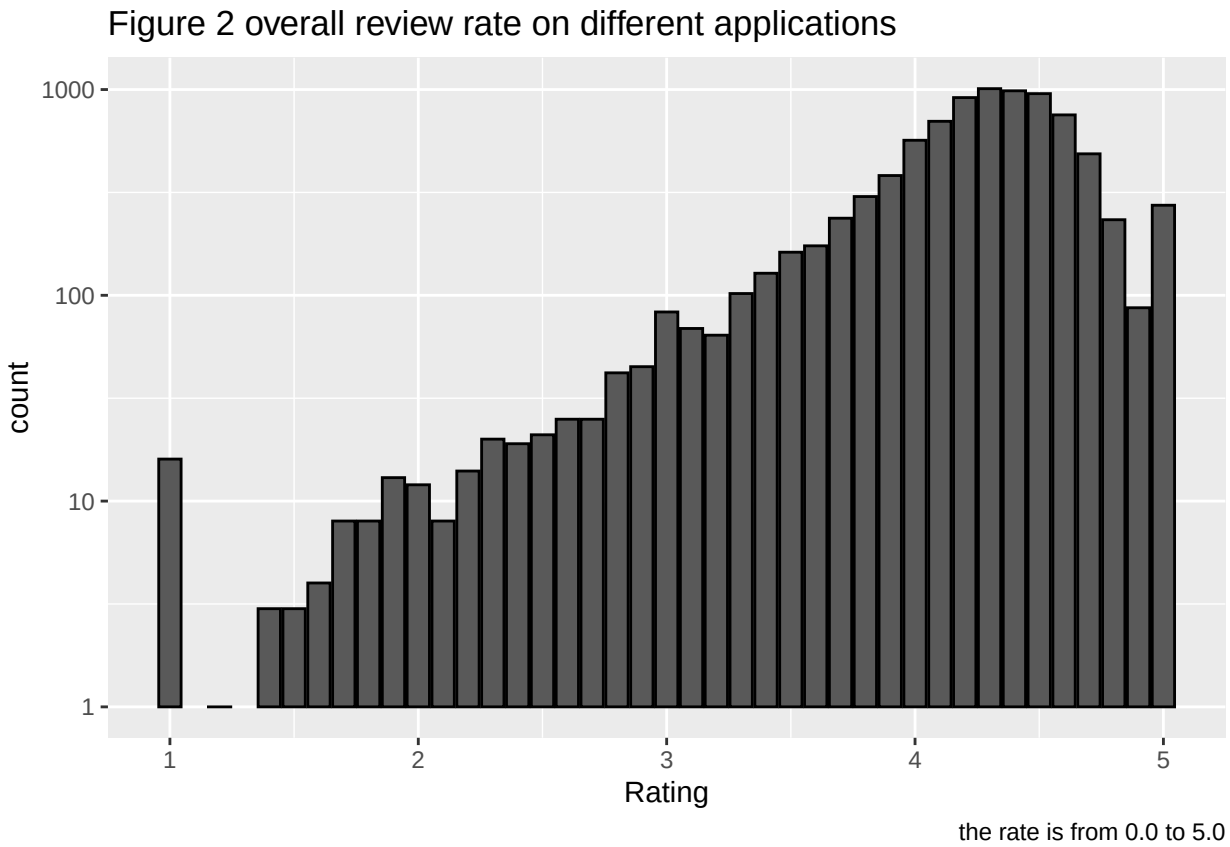
we will have plots by using ggplot.@ggplot2

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

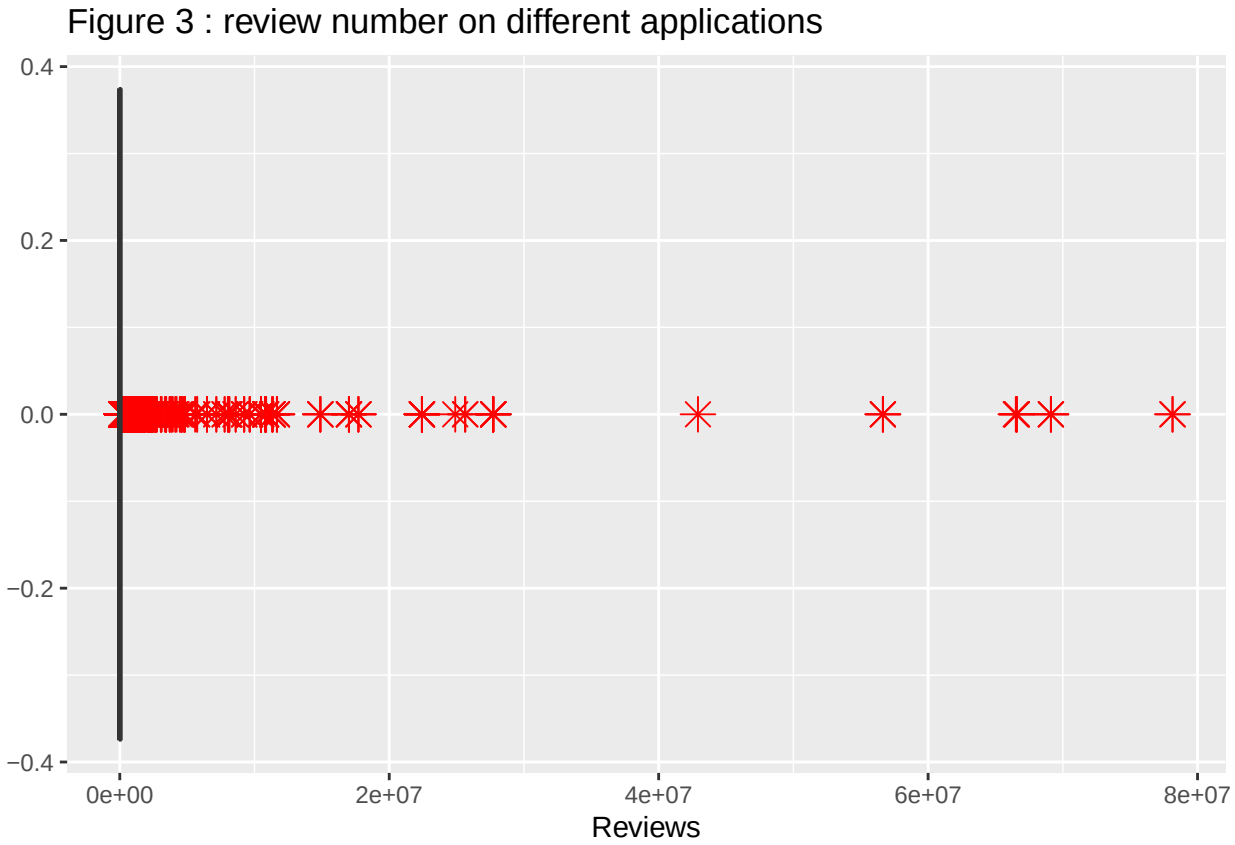
Figure 1



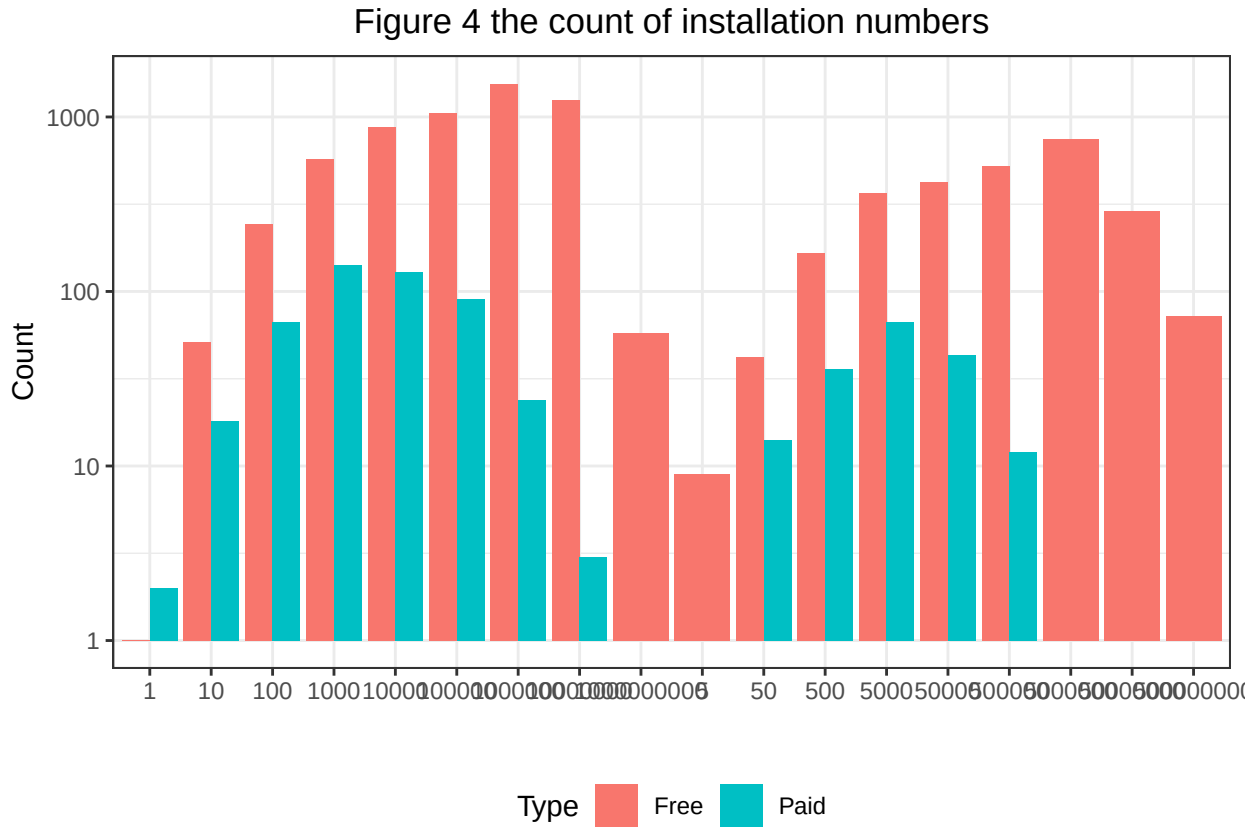
From figure 1, we can see that “FAMILY’ is the largest category in the play store. It occupies almost a quarter of the pie chart.”Game" and “TOOLS” are the second and third largest category, which take almost ??? of the pie chart each. Most of the other categories take around 1/30 of the pie chart. So most developers worked on the family, game, and tools kind of application, and there must be many competition and repetition in these three categories.



From Figure 2, we can see the histogram plot is a left-skewed distribution. The peak is around 4.3. The mean of the rating is lower than the median. Most apps are in a good review rate.



From figure 3, we can see that most of the applications' review numbers are lower than 10^7.



From figure 4, the plot seems a normal distribution.The free application gets more installation numbers than those paid apps. Also, almost all applications with an installation number of more than 5000000 are free. So, free applications will have more users.

model

The purpose of the model is to find the influencing factors on the installation number and their effect.In order to find out the causality of the installation number, we fitted a Bayesian Generalized Multilevel Model on our dataset to verify whether those influencing factors have siginfcant impact.As predictors we choose the rating, reviews,pay, and category.Since the dependent variable is the installation number, we choose gaussian model.

$$install.numb\tilde{N}ormal(\mu_i, \sigma_i)$$

$$\mu_i = \beta_0 + \beta_1 * Rating_i + \beta_2 * Reviews + \beta_3 * pay + \beta_4 * category$$

The β_0 is the intercept The β terms are the coeffcients. And they are free parameters of the model.

The model is fitted with the function brm from the Burkner (2017) package.

Bayesian model consists of a prior distribution function

$$P(\theta|D) = \frac{P(\theta) \times P(D|\theta)}{\int P(\theta' \times P(D|\theta'))\, d\theta'}$$

θ is a vector of parameter values of the model.

Result

Table 5

```
summary(fit_model)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: as.numeric(install.numb) ~ Rating + Reviews + pay + Category
##    Data: popular_data (Number of observations: 8957)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##                                Estimate   Est.Error     l-95% CI    u-95% CI
## Intercept                   -9468780.17 10526294.41 -30106313.69 10508035.58
## Rating                       2600422.05  1395307.61   -199319.25  5333174.35
## Reviews                           21.09        0.25        20.58       21.59
## pay                         -7233586.58  2774276.46 -12553008.43 -1835888.53
## CategoryAUTO_AND_VEHICLES   -1109141.61 11666905.11 -23836694.40 21171367.81
## CategoryBEAUTY              -1062957.31 13312825.07 -26279919.07 24288173.25
## CategoryBOOKS_AND_REFERENCE  4610167.27 10016606.61 -14963496.50 23992391.44
## CategoryBUSINESS              461646.31  9293176.09 -17948280.53 17933562.14
## CategoryCOMICS              -1788397.53 12156576.12 -25752786.38 21415009.35
## CategoryCOMMUNICATION       50136520.34  9410903.43  31695365.67 67967437.81
## CategoryDATING               -352321.47  9762459.43 -19494651.80 18407856.17
## CategoryEDUCATION            -736144.69 10148059.57 -20840456.51 18426251.51
## CategoryENTERTAINMENT        7954265.98 10297999.60 -12281855.46 28066112.71
## CategoryEVENTS              -2152717.89 13137726.19 -27569424.55 23046473.79
## CategoryFAMILY                769762.15  8692835.65 -15814809.93 17339964.61
## CategoryFINANCE              -704725.66  9358362.13 -18640551.87 17497212.01
## CategoryFOOD_AND_DRINK       -591832.53 10534160.78 -21470214.91 19226121.61
## CategoryGAME                 6150806.59  8786413.73 -10693581.48 22898612.48
## CategoryHEALTH_AND_FITNESS   1192833.23  9439351.16 -17032104.21 19519318.81
## CategoryHOUSE_AND_HOME       -529351.25 11524175.60 -23095851.20 22155082.09
## CategoryLIBRARIES_AND_DEMO   -919946.52 11818250.16 -24000993.72 22478235.97
## CategoryLIFESTYLE            -202869.07  9311677.93 -18259845.57 17721360.84
## CategoryMAPS_AND_NAVIGATION   724041.21 10343791.30 -19355197.79 21217227.21
## CategoryMEDICAL               307125.75  9279470.76 -17644937.76 18227334.33
## CategoryNEWS_AND_MAGAZINES  25912681.43  9684491.37   7255866.94 44794189.21
## CategoryPARENTING           -1465009.00 12794877.50 -26083721.39 23793470.90
## CategoryPERSONALIZATION      1384332.99  9297548.30 -16934930.99 19335833.71
## CategoryPHOTOGRAPHY         14405450.65  9431598.54  -4065995.43 32934261.83
## CategoryPRODUCTIVITY        28527392.81  9323696.63  10116810.49 46643138.08
## CategorySHOPPING             1434197.78  9613911.78 -16707430.78 19939284.66
## CategorySOCIAL                512176.46  9553615.66 -18197326.88 19488974.99
## CategorySPORTS                243470.87  9418340.82 -18336967.65 18248703.22
## CategoryTOOLS                7099971.02  8832204.07 -10008307.26 24155780.17
## CategoryTRAVEL_AND_LOCAL    21499172.18  9723240.72   2685191.38 40595905.96
## CategoryVIDEO_PLAYERS       21767760.16 10018773.47   2394562.67 40961453.22
## CategoryWEATHER               570572.89 11547861.42 -21089952.77 23474525.71
##                            Rhat Bulk_ESS Tail_ESS
## Intercept                  1.01      641      972
```

```
## Rating                       1.00     3266     2705
## Reviews                      1.00     3296     2597
## pay                          1.00     3701     2696
## CategoryAUTO_AND_VEHICLES    1.01      768     1581
## CategoryBEAUTY               1.01     1033     2093
## CategoryBOOKS_AND_REFERENCE 1.01      601     1073
## CategoryBUSINESS             1.01      530      940
## CategoryCOMICS               1.01      916     1813
## CategoryCOMMUNICATION        1.01      554      927
## CategoryDATING               1.01      591     1101
## CategoryEDUCATION            1.01      627     1256
## CategoryENTERTAINMENT        1.01      634     1253
## CategoryEVENTS               1.00      936     1739
## CategoryFAMILY               1.01      469      649
## CategoryFINANCE              1.01      526      834
## CategoryFOOD_AND_DRINK       1.01      683     1277
## CategoryGAME                 1.01      489      788
## CategoryHEALTH_AND_FITNESS   1.01      545     1015
## CategoryHOUSE_AND_HOME       1.01      757     1322
## CategoryLIBRARIES_AND_DEMO   1.01      811     1513
## CategoryLIFESTYLE            1.01      532      962
## CategoryMAPS_AND_NAVIGATION 1.01      672     1350
## CategoryMEDICAL              1.01      518      821
## CategoryNEWS_AND_MAGAZINES   1.01      571     1048
## CategoryPARENTING            1.01      887     1394
## CategoryPERSONALIZATION      1.01      533      870
## CategoryPHOTOGRAPHY          1.01      544      930
## CategoryPRODUCTIVITY         1.01      544      946
## CategorySHOPPING             1.01      560      884
## CategorySOCIAL               1.01      560     1081
## CategorySPORTS               1.01      528      981
## CategoryTOOLS                1.01      500      773
## CategoryTRAVEL_AND_LOCAL     1.01      561      992
## CategoryVIDEO_PLAYERS        1.01      618     1012
## CategoryWEATHER              1.01      734     1432
##
## Family Specific Parameters:
##          Estimate Est.Error    l-95% CI    u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma 66955606.62 512167.68 65952695.28 67960087.66 1.00     4587     2983
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Table 5 shows parameter estimates and their corresponding analysis(β_i). The variable Intercept refers to the coefficient β_0 which is the mean of the reference level.

In terms of rating, it has most significant impact on installation numbers. If the review rate increase the installation number also increases dramatically. This is reasonable. When the application's review rate increase, google play store tend to put the application on the top of the searching list. Most people are willing to installation tp 3 apps from their search list, and they will not scroll down to install other apps. Also, if the review rate is very high, the google play store will recommend this app on the main page.

In terms of reviews, it has a positive impact on the installation number, but not very tremendous compare to the rating. The possible reason is that more feedback would give other users a better idea of what the

app is.

In terms of payment, it has a huge negative on the installation number. People tend to use free apps. If there is a demand for a certain kind of APP, people will prefer a free one instead of a paid one. If the paid one is the only App in the google play store, many people choose to not install the App when the demand is elastic. The payment creates a barrier to stop customers to use it.

In terms of category, some categories have a positive impact on the installation number, such as PHOTOGRAPHY, TRAVEL_AND_LOCAL, and COMMUNICATION. These categories are highly related to most people's life. People will use this kind of Apps every day. On the other hand, there are some categories that have a negative impact on the installation number, such as COMICS, BEAUTY, and FINANCE. Those apps have limited target customers, and some of them are professional.

Discussion

In this study, we explore the influencing factors of the installation number. For the application we are trying to develop, we want to have the installation number as high as possible. So, we need to improve our App's quality so that people will give a high rate. Also, we can invite some people to test our APP and write their use experience comments. For our App, we need to set it free in order to have more customers or give customers a free trial. Finally, when we are deciding which kind of App to develop, we need to focus on the apps that the majority of people will use every day.

weakness

The data is biased and not very comprehensive. It only contains the application data from the google play store. Different OS systems and platforms might have different target customers. The result from our model might not fit the situation in which we develop an application based on IOS. Also, the data is gathered in 2018 which is not up to date. Some applications were popular back then, but not 2020. The raw data contains the variable "size" which has information about how much storage will the APP use in the cell phone memory. This is also a very important factor that affects the installation number since most phones have limited storage. However, the "size" contains too many observations of "Varies with the device". It covert the size data from numeric to categorical. It is hard to manipulate the data. So I did not choose "size" as a factor. Furthermore, there will be some missing data. It will affect us analyzing the result.

The next step

We could gather the most up-to-date data that contains information about APPs on both Android and IOS. For observations of the "Varies with the device" question, we could list the type of phones with their installation size. Also, we could fit other causal inference models to see the differences.

Reference

The data source is : Lavanya Gupta (2018)

R Core Team (2019)

Xie (2020)

Nick Galov (2020)

Allaire et al. (2020)

Harrell Jr (2020)

Firke (2020)

Wickham (2016)

Burkner (2017)

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2020. *Rmarkdown: Dynamic Documents for R.* https://github.com/rstudio/rmarkdown.

Burkner, Paul-Christian. 2017. *Brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software.*

Firke, Sam. 2020. *Janitor: Simple Tools for Examining and Cleaning Dirty Data.* https://CRAN.R-project.org/package=janitor.

Harrell Jr, Frank E. 2020. *Rms: Regression Modeling Strategies.* https://CRAN.R-project.org/package=rms.

Lavanya Gupta. 2018. *Google Play Store Apps.* https://www.kaggle.com/lava18/google-play-store-apps.

Nick Galov. 2020. *Mobile and Desktop Operating Systems Market Share.* R Foundation for Statistical Computing. https://hostingtribunal.com/blog/operating-systems-market-share/#gref .

R Core Team. 2019. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Xie, Yihui. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in R.* https://yihui.org/knitr/.