Report of MA678 Midterm Project

Tong Sun

12/1/2021

Abstract

Google Play is the official app store of Android, Google's mobile platform. It allows people to view applications and load of content before downloading anything on their devices. The Play Store apps data has enormous potential to drive app-making businesses to success. In this report, I explored this question – what factors might influence the number of users' reviews and how they drive to future app-making businesses probably? Based on setting up a multilevel regression model, I discovered several factors that may make sense, making some suggestions for future app-making businesses at the same time. Additionally, it indexed important information about Android apps. The model shows that the number of reviews is slightly different between categories. This report consists 4 main parts: Introduction, Method, Result and Conclusion.

Introduction

Google Play indexes important information about Android apps, including ratings, alternative suggestions,user reviews and other descriptions of apps. It uses sophisticated modern-day techniques (like dynamic page load) using JQuery making scraping more challenging. The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market! So each app has its own features and some similarities with others as well, such as the number of installs, types and the price of apps. And some features may lead the apps to be more out-standing.

So here I would use multilevel regression to see how different their number of reviews are between each category.Before that, I cleaned the data and added with some new columns to see which factors are more significant.

Method

Data Cleaning and Processing

The main data set is published on Kaggle: Google Play Store Dataset.

Firstly I made lots of steps to data cleaning. In order to have a clear format of data, I checked the types of the data and variables and changed all factor variables to be numeric, eliminating lots of symbols (such as "M"and currency symbol) as well.

Description of Data

column names	explanation
App	Application name
Rating	Overall user rating of the app (as when scraped)
Reviews	Number of user reviews for the app (as when scraped)
Size	Size of the app (as when scraped)
Installs	Number of user downloads for the app (as when scraped)
Type	Paid or Free
Content Rating	Age group the app is targeted at-Children/Mature 21+/Adult
Genres	An app can belong to multiple genres (apart from its main category)

EDA

Here I would like to draw some plots to see the relationship between reviews and other variables and analyze features of variables, since my question is how some factors effect the number of reviews.

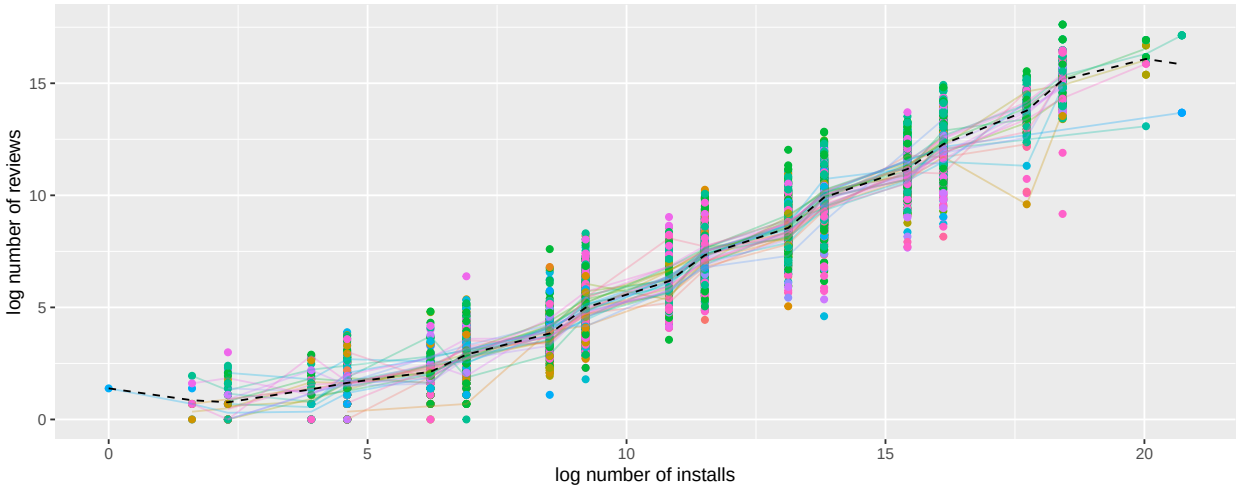


Figure 1: The relationship between number of installs and reviews.

Figure 1 shows the relationship between the number of installs and reviews. But it seems like there is not too much significant relationship between these two variables, because all the lines in the plot have the similar trend by default. So I guess the variable 'log_installs' should not be included in our model.

Figure 2 shows the relationship between size and the number of reviews. Unlike the 'installs', the relationship here varies obviously between categories. There should be random intercept and random slope at the same time. For the plot on the right, I found that for some categories of apps, the number of reviews varies dramatically with the size of the application.

Figure 3 shows the relationship between rating and the number of reviews. The result is the same as the size's. The differences between categories are so significant. Also different categories have different slopes and intercepts. Therefore, I think the variables 'Size' and 'Rating' should be taken into account when I fit the model.

Model Fitting

After deleting lost data and duplicated rows, transforming log of column 'Installs' and 'Reviews' (because I found they all have a large scale with a long tale from the histogram plots showed above), I got 7369

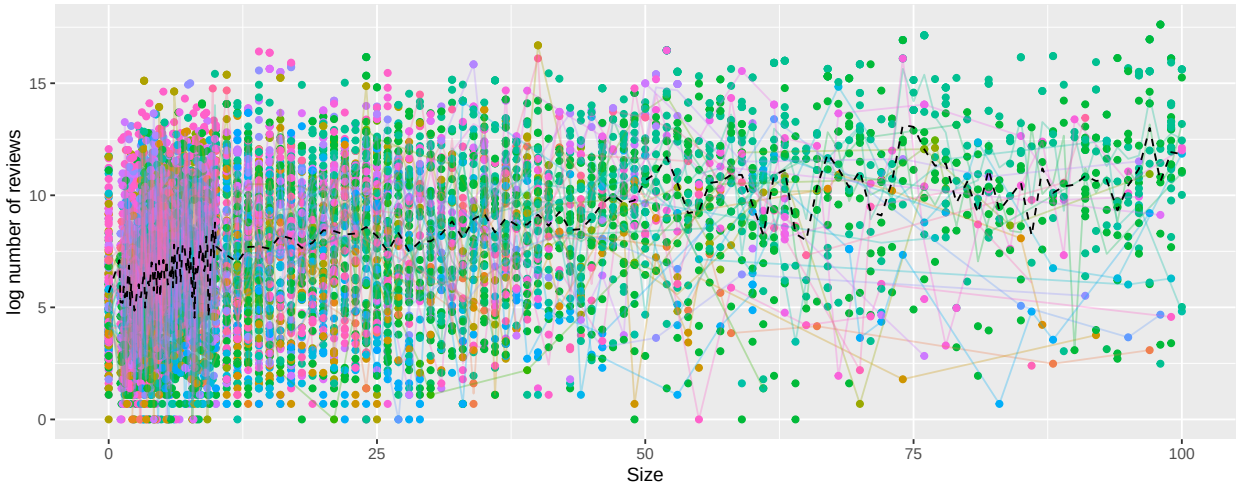


Figure 2: The relationship between size and the number of reviews.

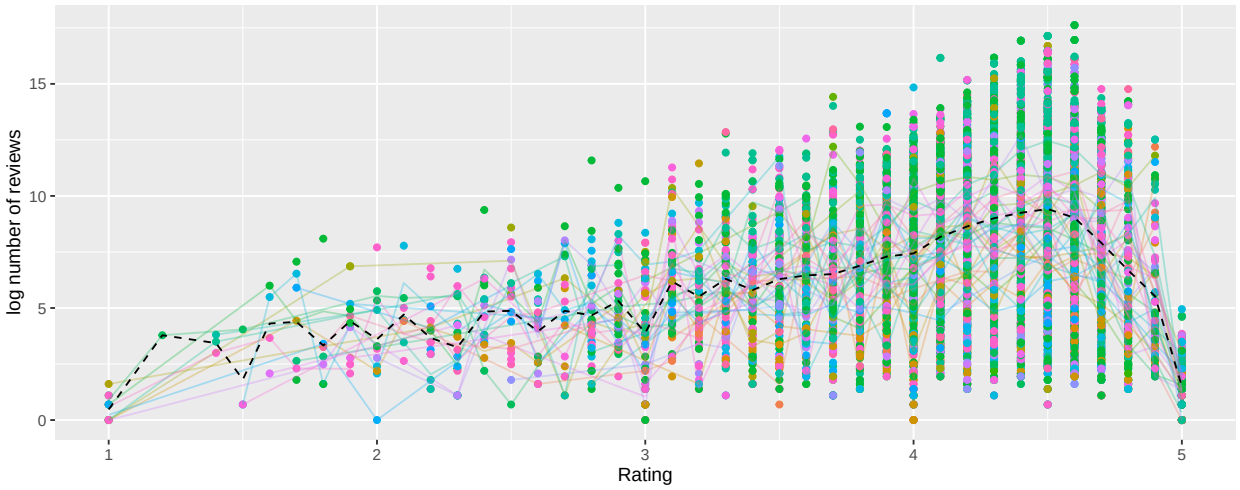


Figure 3: The relationship between rating and the number of reviews.

observations of 15 variables. But I will choose only several variables to use. Here I only analysis those 'Type' == "Free",which has a larger proportion.

Considering different categories, I used multilevel model to fit the data. From the EDA part, I found that for the 'log_installs' predictor, the differences between each category was not so significant. So I did not choose 'log_installs' as one of model's predictor. Here I chose continuous variables – 'Size' and 'Rating' as model's predictors. From EDA part above, I guessed that for 'Size' and 'Rating' predictors, it should be random intercept and random slope model. To be more convinced, I still did three different models and used anova for model selecting (choosing the one has the smallest AIC, model.2 shown as follow).

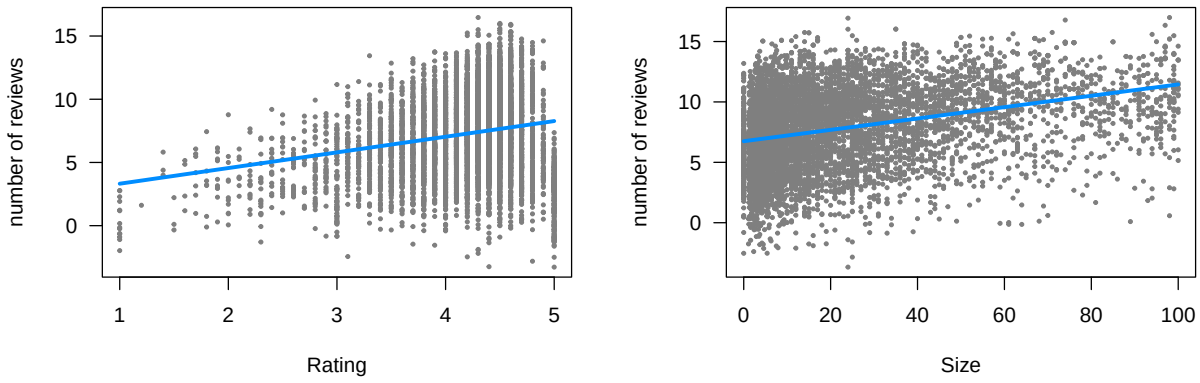


Figure 4: How the expected value of the y(Reviews) changes as a function of x,with all other variables held fixed.

To see the fixed effects below:

	Estimate	Std. Error	t value	p value
(Intercept)	1.42	0.52	2.75	5.93e-03
Rating	1.24	0.16	7.72	1.20e-14
Size	0.05	0.01	8.48	0.00e+00

Model results

As we can see the results of fixed effects above, here I only take one of those categories into account. For ART_AND_DESIGN category, we can conclude this formula:

$$log(reviews) = 4.70 + 0.64 \cdot Rating + 0.07 \cdot Size$$

All the parameters of three predictors are all bigger than 0 (except for the parameter of 'Size' under AUTO_AND_VEHICLES category), which means they all have positive impact on number of reviews. For each 1% difference in the number of installs, the predicted difference in reviews is 0.95%. About the reason of the negative parameter of 'Size' under AUTO_AND_VEHICLES category, I think it does not have so much meaning. Because the abstract value of this parameter is almost 0.

In addition, we can see differences between each category. Here I took three of them as example. Let's look at their parameters of these predictors as follow:

	(Intercept)	Rating	Size
ENTERTAINMENT	-7.25	2.84	0.05
EVENTS	7.29	0.18	0.02
MEDICAL	6.42	0.27	0.01

For different categories, the influence of each predictor is always not the same. For ENTERTAINMENT, I think whether people will download the app depends more on how people rate it, and because it's entertainment, people are more likely to decide whether to download a non-essential app because of the size of the device's memory. So the parameters of 'Rating' and 'Size' are bigger than other categories. For EVENTS, the parameter of 'Rating' is small, this might because these apps are mostly nonfiction apps, and people rarely rate them. And for MEDICAL, since apps in this category are essential to people's lives, the number of times people review them is less dependent on other users' ratings and the size of the app itself.

Model Validation

In this part, based on residual checking I discovered the normality of the model is good and there are not obvious point, which means the coefficients in the regression model would not change if a particular observation was removed from the data.

Conclusion

From the estimates above, I would draw the conclusion as follow. The parameters of variables shows that "Rating" and "Size" both have a positive influence on the number of reviews. The higher level from people's rating, which means people who used this app before have more sense about it, they would like to recommend it to others, so they are willing to share their feelings on social platforms. In addition, for the "Size" factor, the larger size the app has, the more probably people will write reviews for it. By default, the reason of it is those apps with larger size may include more information and more utility features, in that case there will be more people downloading and using them. Therefore there will be more reviews of them, but the influence of size is not so significant comparing with rating.

In addition, app-making businesses should pay more attention on the creation of entertainment apps because from the distribution of number of reviews by categories, those are at the top. The more comments people make about software, the more social impact it has, and the more likely software developers are to find new ideas from reviews. By default, apps designers should explore apps that are free. Certainly,the importance of this is somewhat diminished for categories of software that people need to use in their daily lives, such as those under the medical category.

For the purpose of giving suggestions for future app-making businesses, I would like to suggest them pay more attention to people's rating factor when they create new apps. Obviously, if the app has higher rating that means people like this production more, probably they will write more reviews on it so this will attract more and more people to use it. And another suggestion is businesses could care about the app's size when they design a new production. For those apps include more utility features, people have to use them so they probably will not care about their sizes so much. But for those apps only for entertainment, if the application occupies a large amount of device memory, people will consider not downloading and using it, or have a high chance to uninstall it after downloading. Of course, there are not many comments on these software, which indicates that these software is not popular. Finally, software developers should take better advantage of the momentum of young people, who are more likely to post reviews on social platforms. The development industry should take advantage of the drive effect while doing what it wants to do to drive adoption of other applications.

Appendix

More EDA

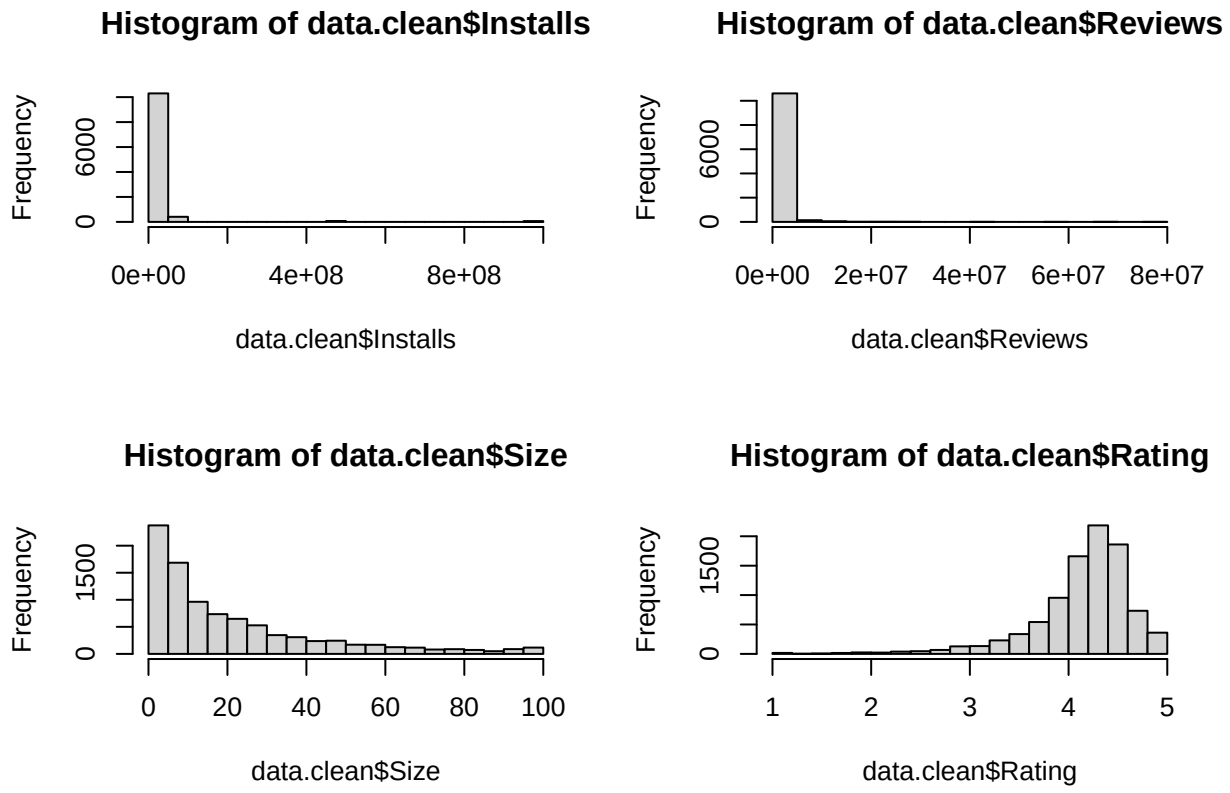
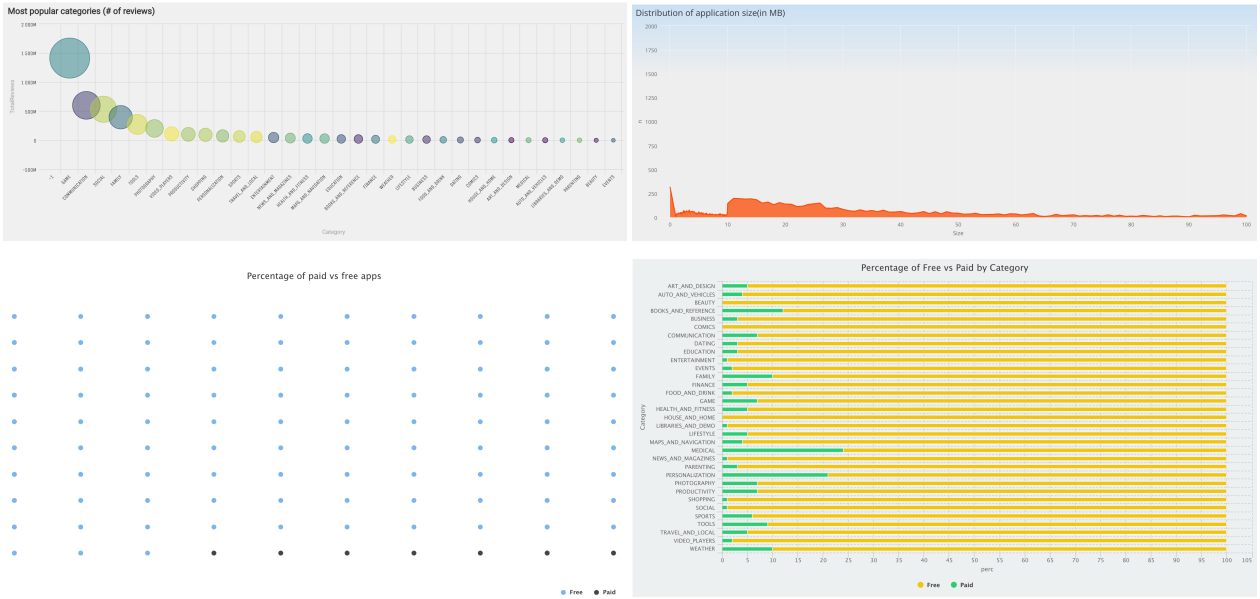


Figure 5: Distribution plots for number of installs, number of reviews, size and rating



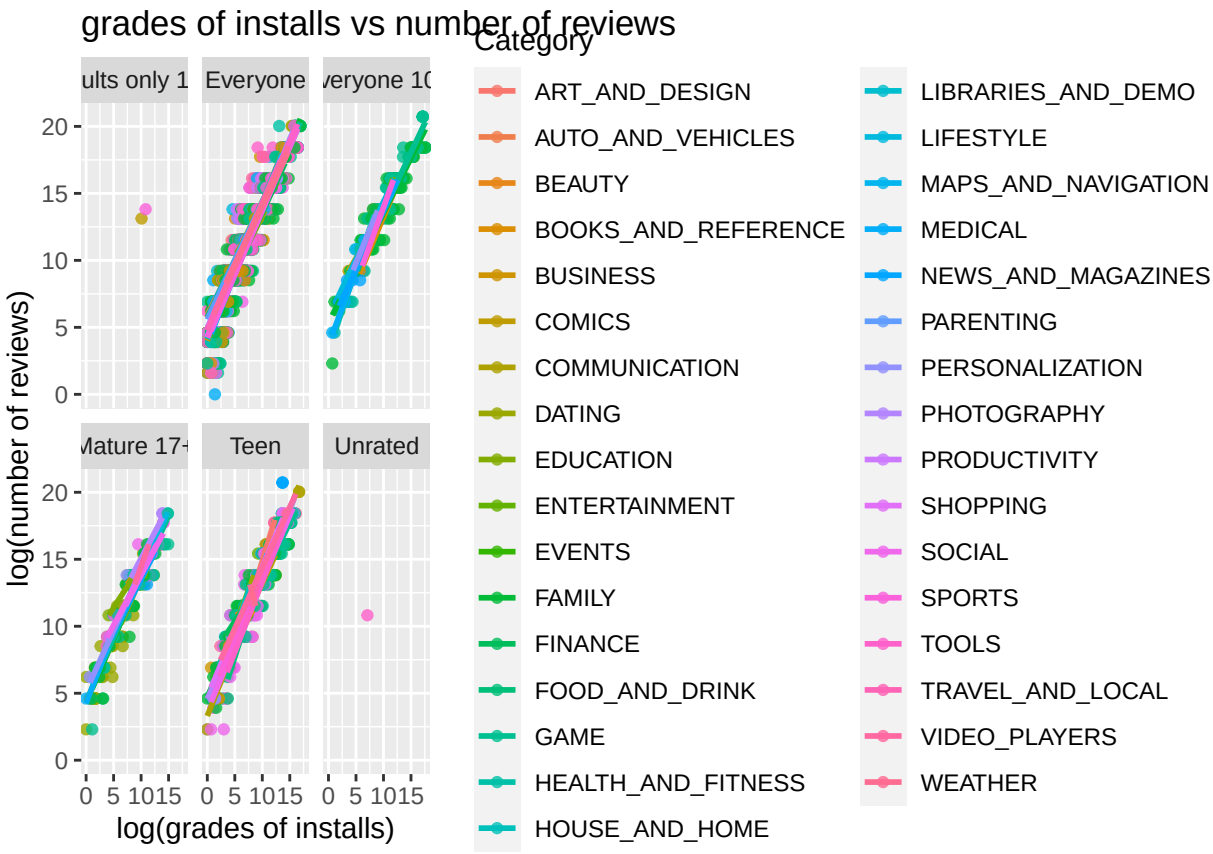


Figure 6: 'log_reviews' vs 'log_installs'

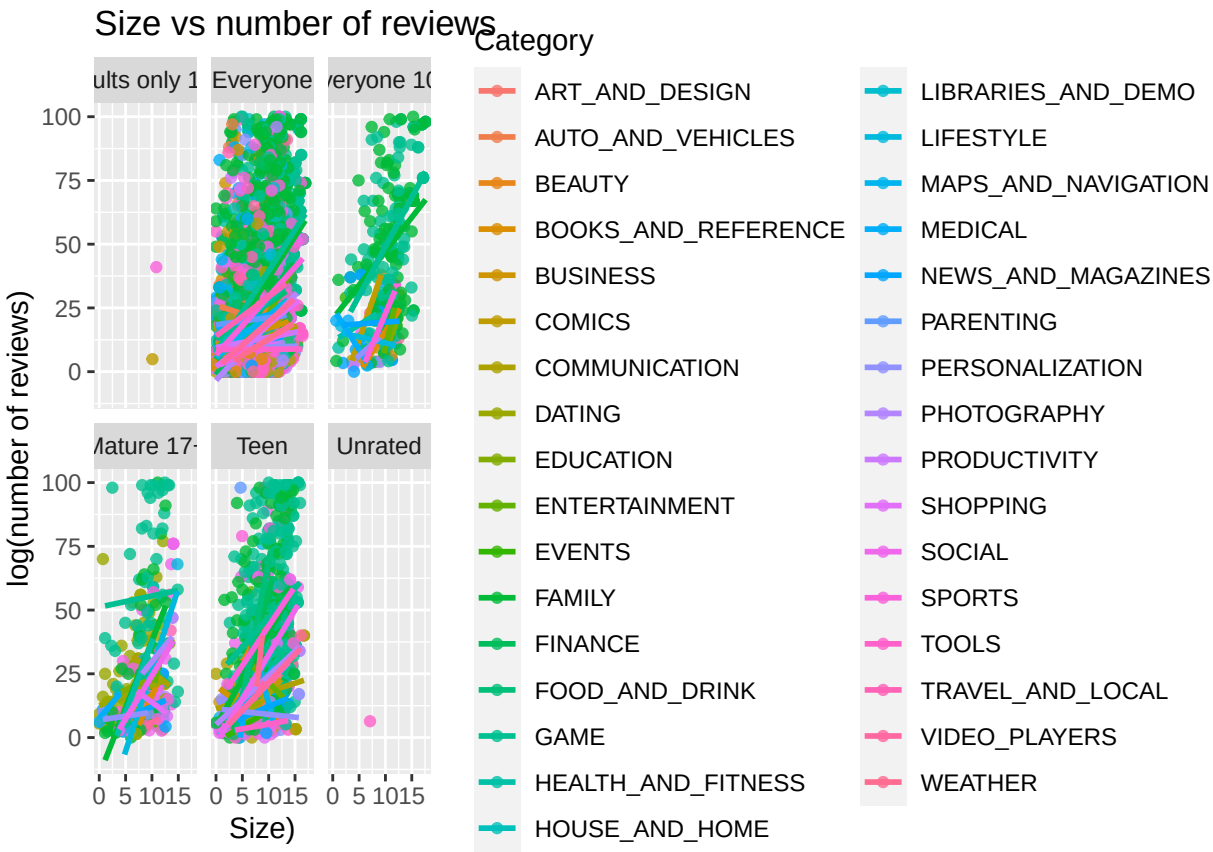


Figure 7: 'log_reviews' vs 'Size'

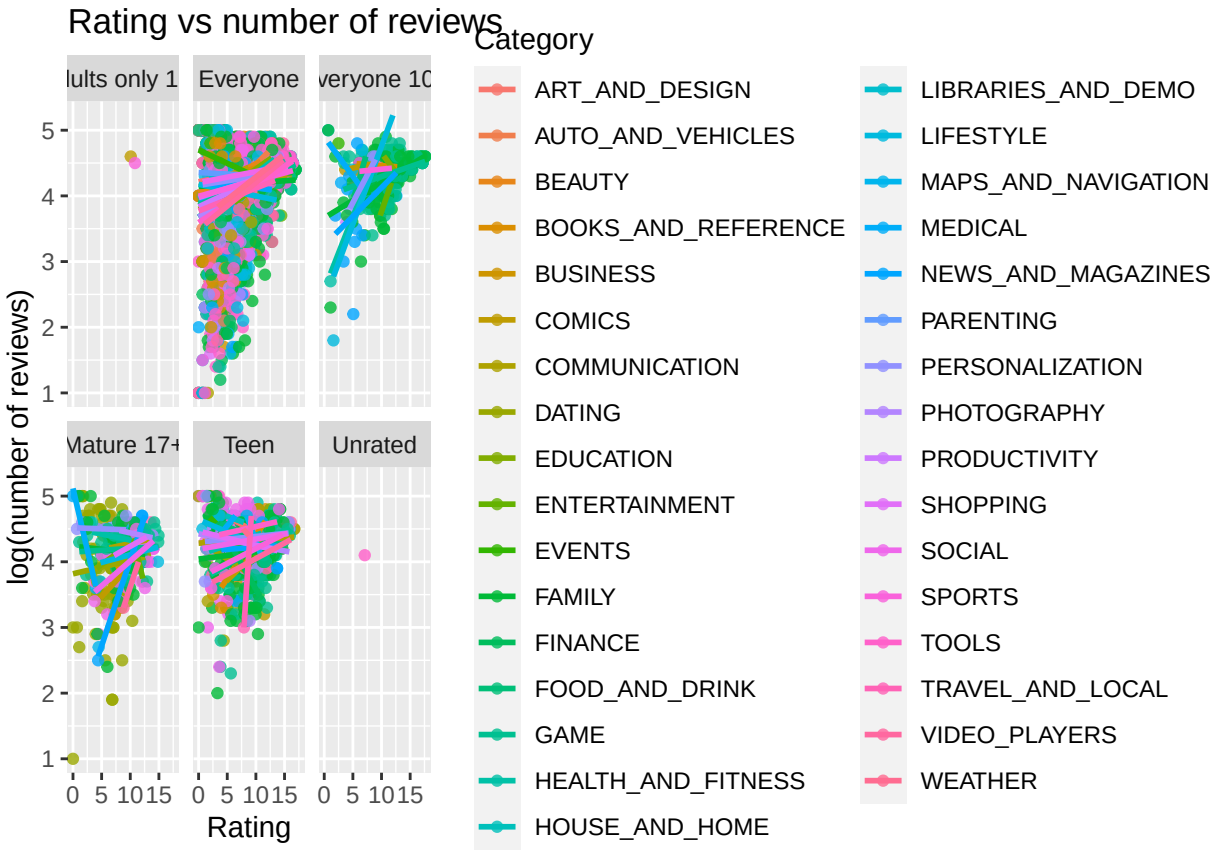


Figure 8: 'log_reviews' vs 'Rating'

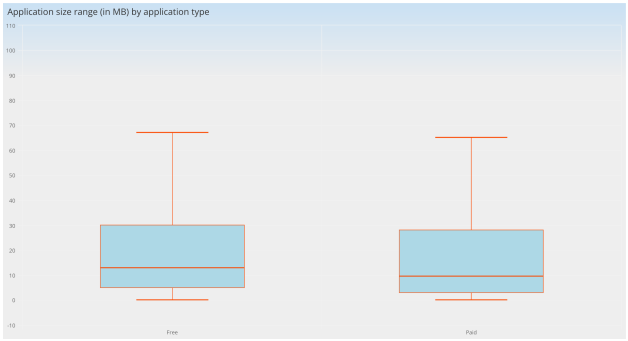
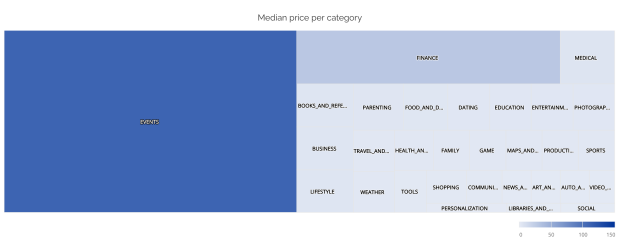
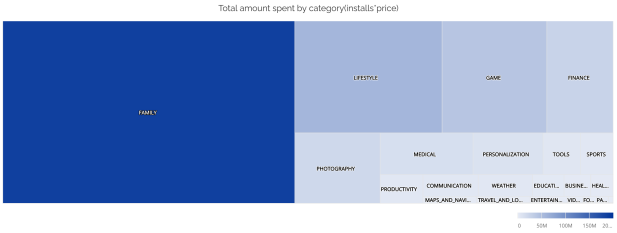
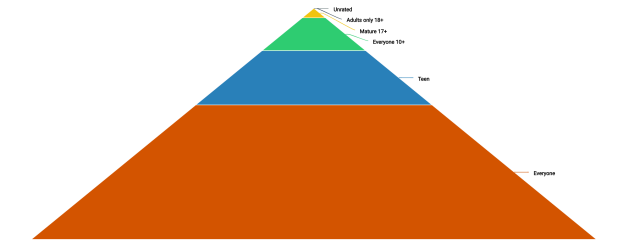
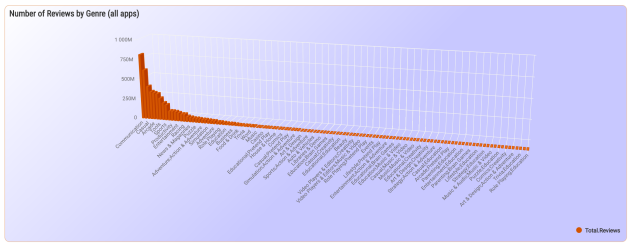


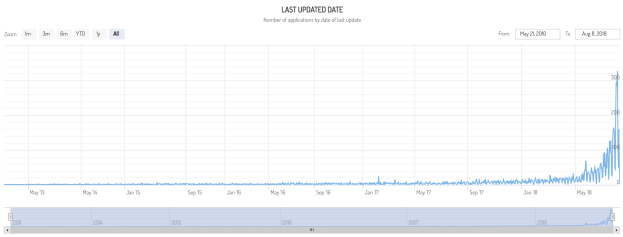
Figure 9: hchart

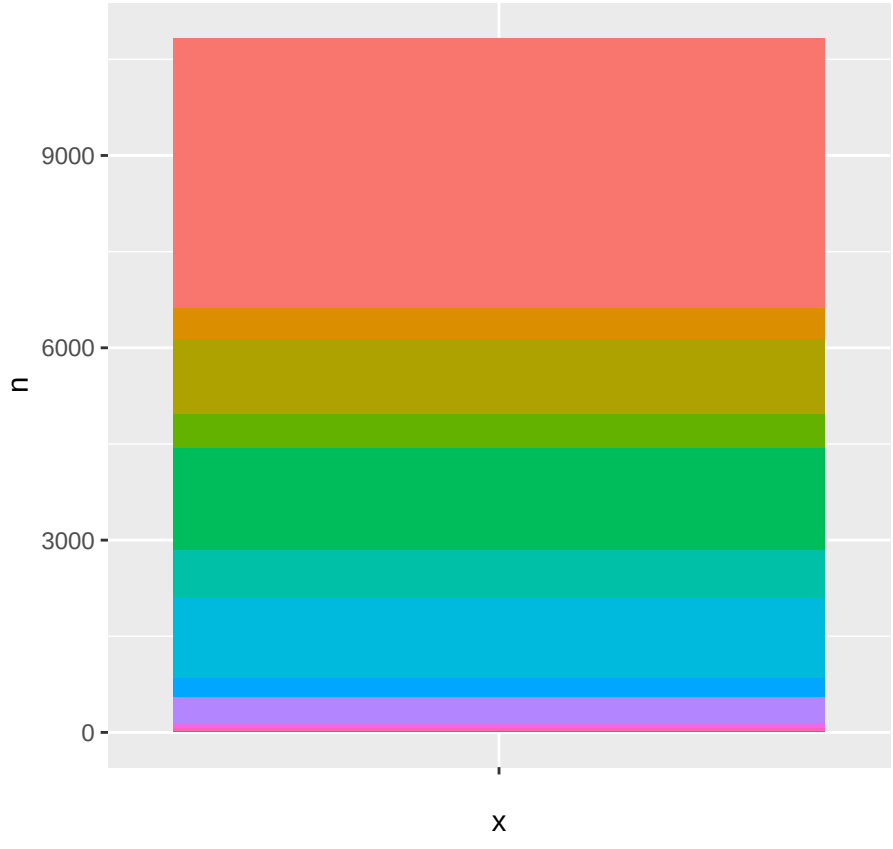

Median price per category
FINANCE
MEDICAL
EVENTS
BOOKS_AND_REFE...
PARENTING
FOOD_AND_D...
DATING
EDUCATION
ENTERTAINM...
PHOTOGRAP...
BUSINESS
TRAVEL_AND...
HEALTH_AN...
FAMILY
GAME
MAPS_AND...
PRODUCTI...
SPORTS
LIFESTYLE
WEATHER
TOOLS
SHOPPING
COMMUNI...
NEWS_A...
ART_AN...
AUTO_A...
VIDEO_...
PERSONALIZATION
LIBRARIES_AND_...
SOCIAL
0
50
100
150


Total amount spent by category(installs*price)
FAMILY
LIFESTYLE
GAME
FINANCE
PHOTOGRAPHY
MEDICAL
PERSONALIZATION
TOOLS
SPORTS
PRODUCTIVITY
COMMUNICATION
WEATHER
0
50M
100M
150M


Unrated
Adults only 18+
Mature 17+
Everyone 10+
Teen
Everyone


Number of Reviews by Genre (all apps)
1 000M
750M
500M
250M
0
Total Reviews


LAST UPDATED DATE
Number of applications by date of last update
Zoom
1m
3m
6m
YTD
1y
All
From
May 21, 2010
To
Aug 8, 2018
300
200
100
0
May '13
May '14
Jan '15
Sep '15
Jan '16
May '16
Sep '16
Jan '17
May '17
Sep '17
Jan '18
May '18


9000
6000
3000
0
n
x
Installs.Group
(0,10000]
(40000,50000]
(90000,1e+05]
(4.9e+05,5e+05]
(9.9e+05,1e+06]
(4.99e+06,5e+06]
(9.99e+06,1e+07]
(4.999e+07,5e+07]
(9.999e+07,1e+08]
(4.9999e+08,5e+08]
(9.9999e+08,1e+09]
NA

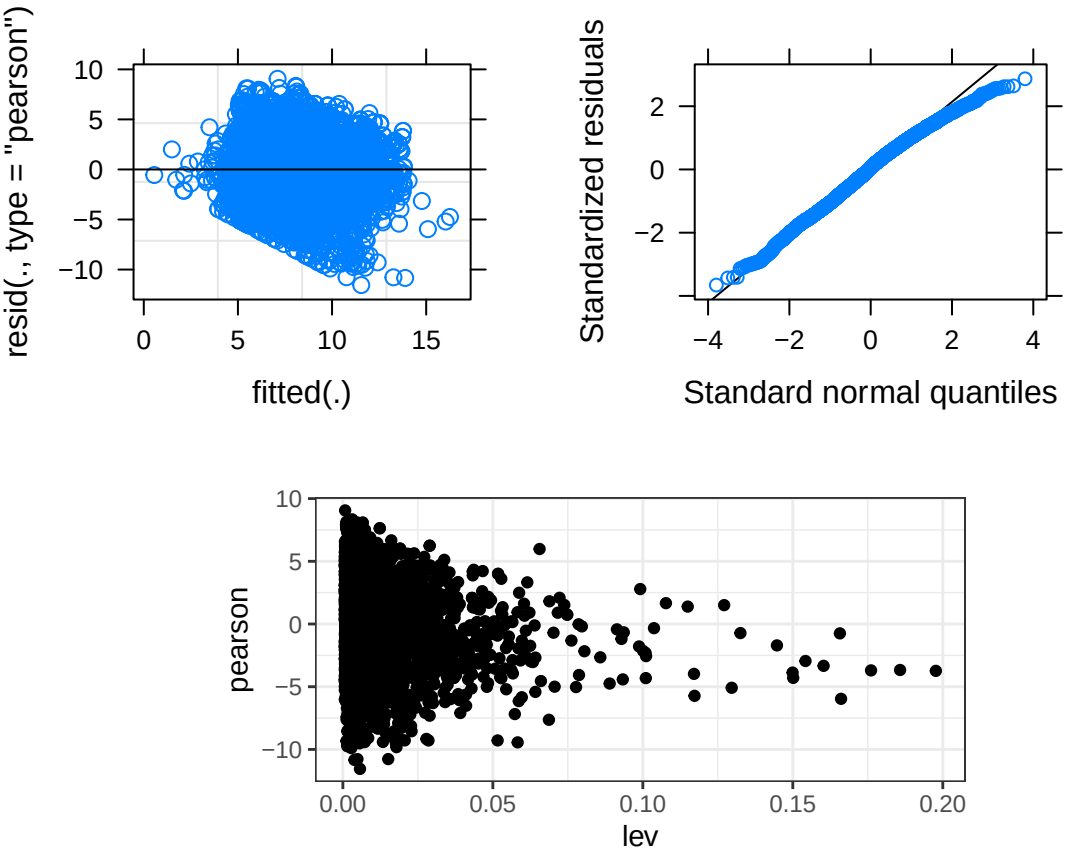


Figure 10: Residuals vs Leverage.

Full Results

Random effects of model

```
## $Category
##                     (Intercept)      Rating  (Intercept)         Size
## ART_AND_DESIGN       1.63897212 -0.59986912  0.061376687  0.022303412
## AUTO_AND_VEHICLES   -0.43334202  0.13526397 -0.206370059 -0.048043999
## BEAUTY               2.05745877 -0.76646123 -0.028194079  0.004884401
## BOOKS_AND_REFERENCE  0.66560986 -0.26229620 -0.064574464 -0.007521832
## BUSINESS             2.59056273 -0.96851933 -0.117352468 -0.014783223
## COMICS               0.98600743 -0.34134926  0.036446778  0.005013509
## COMMUNICATION       -2.06952393  0.73836924  0.041997505  0.012622017
## DATING               1.19512703 -0.38972805  0.055130783 -0.001098583
## EDUCATION           -3.47085376  1.25382716 -0.049095402 -0.018102059
## ENTERTAINMENT       -4.33507668  1.60013895  0.080034626  0.001622436
## EVENTS               2.93444859 -1.05385444 -0.099307882 -0.025856487
## FAMILY               2.79043051 -0.95720180  0.129915250  0.017775264
## FINANCE              1.63736173 -0.66016946 -0.015482770  0.027056416
## FOOD_AND_DRINK      -0.05336559  0.02676756  0.071502173  0.016322857
## GAME                -3.61751823  1.30145040 -0.002658751 -0.003321852
## HEALTH_AND_FITNESS  -0.12477898  0.03435678 -0.027680664 -0.003202457
## HOUSE_AND_HOME      -1.18648161  0.43812173 -0.012703675 -0.009100660
## LIBRARIES_AND_DEMO   1.83273668 -0.64821509 -0.009339331 -0.005958533
## LIFESTYLE            3.04519269 -1.08903289  0.076551260  0.020372617
## MAPS_AND_NAVIGATION -0.91486792  0.30056556 -0.063483465 -0.005926984
## MEDICAL              2.49698084 -0.96471710 -0.246336299 -0.037497378
## NEWS_AND_MAGAZINES   0.95181140 -0.33423055  0.085039273  0.020482338
## PARENTING            1.08955673 -0.38851953 -0.109257539 -0.030584475
## PERSONALIZATION     -0.92831539  0.36581633 -0.038898119 -0.024798169
## PHOTOGRAPHY         -1.38308314  0.51988167  0.201722734  0.044765412
## PRODUCTIVITY        -1.11490937  0.41198886 -0.049914853 -0.019069826
## SHOPPING            -0.78296489  0.28673547  0.182421702  0.047218794
## SOCIAL               0.41524547 -0.12365197  0.123147637  0.023095816
## SPORTS              -0.48478622  0.20513118  0.067307342  0.005057910
## TOOLS               -1.45025148  0.51749661 -0.097303576 -0.025867252
## TRAVEL_AND_LOCAL     0.22799339 -0.10326363 -0.049599060 -0.004412828
## VIDEO_PLAYERS       -1.63214491  0.57392048  0.030630545  0.012830667
## WEATHER             -2.57323184  0.94124769  0.044328159  0.003722729
##
## with conditional variances for "Category"
```

Fixed effects of model

```
## (Intercept)      Rating        Size
##  1.42202388  1.23866126  0.04701581
```

Coefficients of model

```
## $Category
##                     (Intercept)    Rating         Size
## ART_AND_DESIGN        4.6999681 0.6387921  0.069319225
## AUTO_AND_VEHICLES     0.5553398 1.3739252 -0.001028187
```

```
## BEAUTY                5.5369414 0.4722000  0.051900214
## BOOKS_AND_REFERENCE   2.7532436 0.9763651  0.039493981
## BUSINESS              6.6031493 0.2701419  0.032232589
## COMICS                3.3940387 0.8973120  0.052029321
## COMMUNICATION        -2.7170240 1.9770305  0.059637830
## DATING                3.8122779 0.8489332  0.045917230
## EDUCATION            -5.5196836 2.4924884  0.028913754
## ENTERTAINMENT        -7.2481295 2.8388002  0.048638249
## EVENTS                7.2909211 0.1848068  0.021159326
## FAMILY                7.0028849 0.2814595  0.064791076
## FINANCE               4.6967473 0.5784918  0.074072228
## FOOD_AND_DRINK        1.3152927 1.2654288  0.063338669
## GAME                 -5.8130126 2.5401117  0.043693960
## HEALTH_AND_FITNESS    1.1724659 1.2730180  0.043813355
## HOUSE_AND_HOME       -0.9509393 1.6767830  0.037915153
## LIBRARIES_AND_DEMO    5.0874972 0.5904462  0.041057279
## LIFESTYLE             7.5124093 0.1496284  0.067388430
## MAPS_AND_NAVIGATION  -0.4077120 1.5392268  0.041088829
## MEDICAL               6.4159856 0.2739442  0.009518434
## NEWS_AND_MAGAZINES    3.3256467 0.9044307  0.067498150
## PARENTING             3.6011373 0.8501417  0.016431337
## PERSONALIZATION      -0.4346069 1.6044776  0.022217643
## PHOTOGRAPHY          -1.3441424 1.7585429  0.091781224
## PRODUCTIVITY         -0.8077949 1.6506501  0.027945986
## SHOPPING             -0.1439059 1.5253967  0.094234607
## SOCIAL                2.2525148 1.1150093  0.070111629
## SPORTS                0.4524514 1.4437924  0.052073723
## TOOLS                -1.4784791 1.7561579  0.021148560
## TRAVEL_AND_LOCAL      1.8780107 1.1353976  0.042602985
## VIDEO_PLAYERS        -1.8422659 1.8125817  0.059846480
## WEATHER              -3.7244398 2.1799089  0.050738541
##
## attr(,"class")
## [1] "coef.mer"
```

Confidence intervals for both fixed and random effects

```
##                   2.5 %      97.5 %
## .sig01       1.15815430  3.14576095
## .sig02      -1.00000000 -0.96741392
## .sig03       0.49167761  1.04475642
## .sig04       0.00000000  0.45216507
## .sig05      -0.99458915  1.00000000
## .sig06       0.01755679  0.03522591
## .sigma       3.10644987  3.21272425
## (Intercept)  0.42292842  2.43140994
## Rating       0.92196389  1.56445002
## Size         0.03570815  0.05847952
```