Final Paper: Google Play Store Apps Rating

Hsin Chen

12/2/2019

1 Introduction

Similar to Apple App Store, Google Play Store is one of the biggest mobile application stores and has innumerable apps available on it. Many companies, studios, and developers launch their applications on the Google Play Store. They are interested in what designs and strategies can make their products get better ratings and comments, so they can possibly to attract and reach more users. Their questions would be:

- What are the types, sizes, prices, genres, and categories of the top-rated apps?
- Is it possible to predict an app's rating given its other attributes?

The dataset provides information about the features of top-rated apps the pricing strategies, and my purpose is to understand why those apps have high ratings and find an approach to get higher user ratings.

Next, The dataset is obtained from Kaggle, and it has more than 10,000 web-scraped records of Apps on Google Play Store collected by Lavanya Gupta. The last update date of this dataset is 2019/02/03. In order to use the dataset for my purpose, I first transformed and cleaned some fields, e.g. Transforming the Size field to numeric data. The procedure and R code of cleaning and modifying data are shown in **Appendix 7.1**.

The fields in the dataset after data cleaning and processing:

Fields' Name	Fields' Description	Data Type
App	The name of the app	Categoric
Category	The category which the app belongs to	Categoric
Rating	Overall user rating of the app	Numeric
Reviews	The number of user reviews for the app	Numeric
Size	The size of the app	Numeric
Installs	Number of user downloads/installs for the app	Numeric
Price	Price of the app	Numeric
Content Rating	Age group which the app is targeted at	Categoric
Genres	An app can belong to multiple genres	Categoric
DaysFromLastUpdate	Numbers of days from the date of last update	Numeric

In the following sections, the research problems and solutions will be addressed. Then, the process of the analysis will be explained and concluded.

2 Problem Statement

The ultimate goal is to get insights into the features of top-rated applications and understand what makes an application have a higher user rating. Therefore, the information can be utilized to enhance an application's user rating.

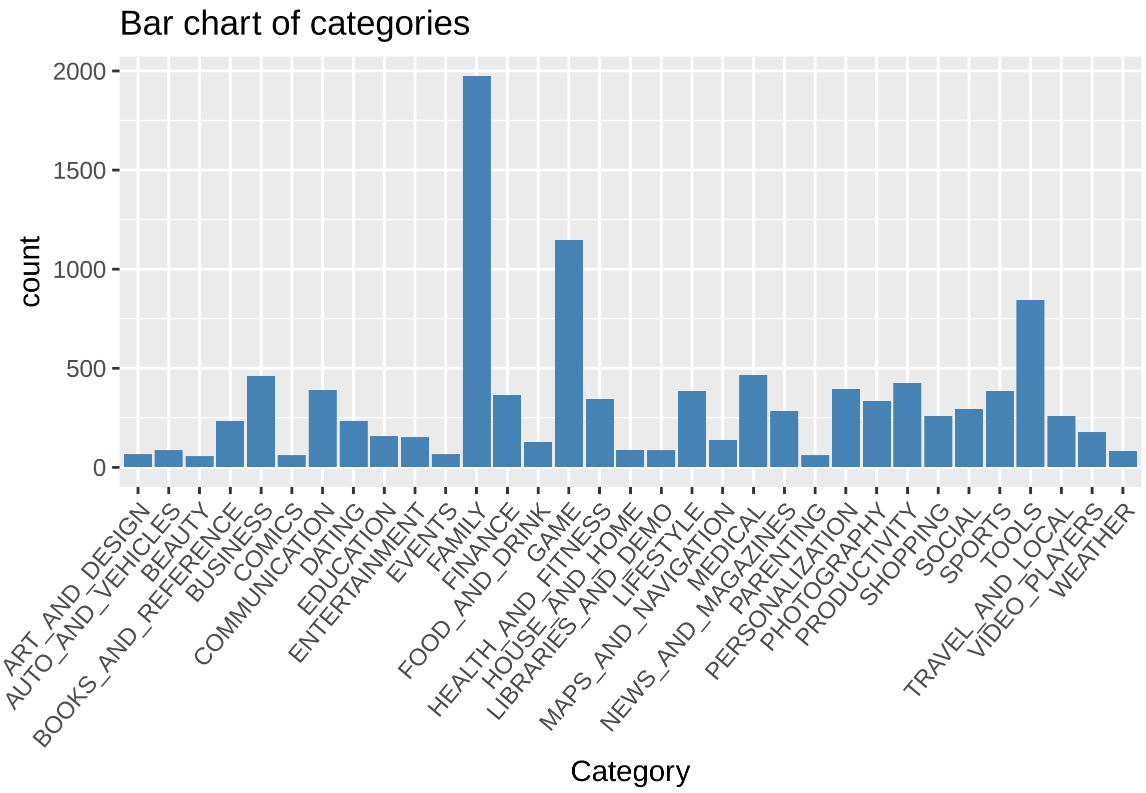
3 Solution

To understand the features of the top-rated applications or what factors affect user ratings, I will do exploratory data analysis first. Using graphs and plots to show the distribution of relationships of attributes can help understand more about the data and have a rough concept about the research. Next, I will use multiple linear regression models to analyze and interpret the data. Manually selecting independent variables will be attempted first, and I will apply 3 variable selection techniques for choosing variables. Inferential tools for multiple regression such as confidence intervals will also be used in the research.

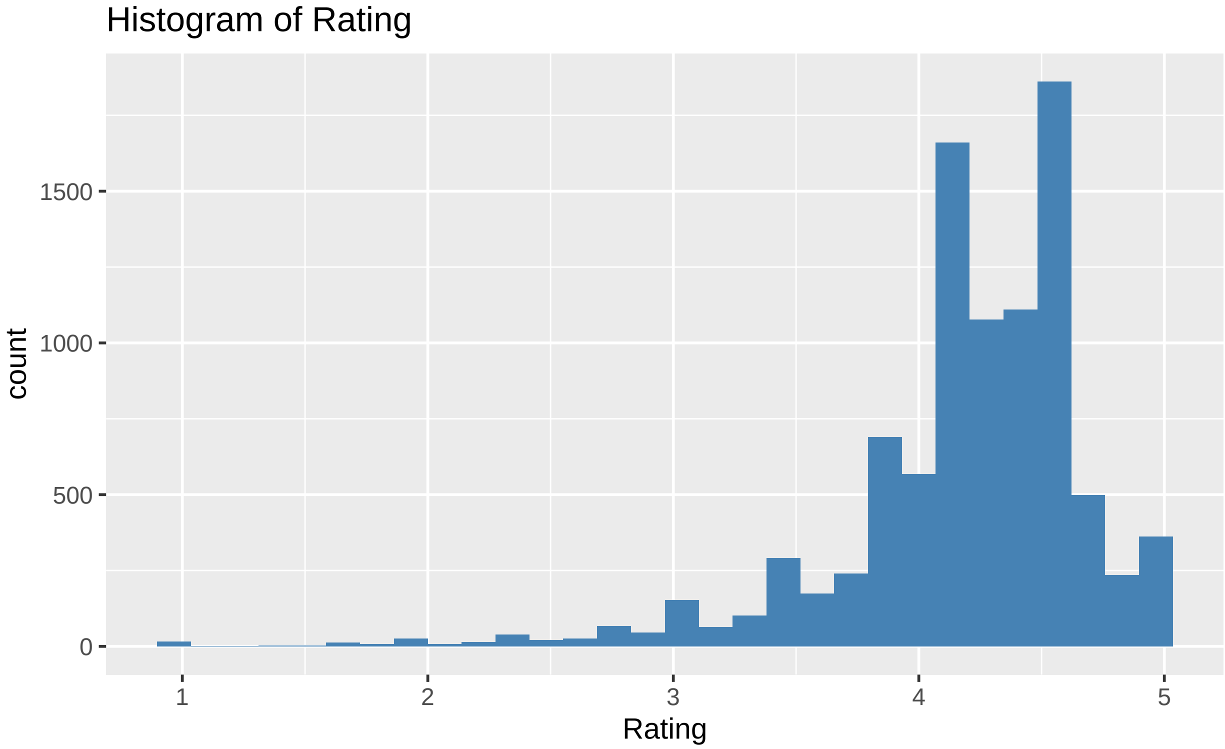
4 Analysis

4.1 Exploratory Anaylsis

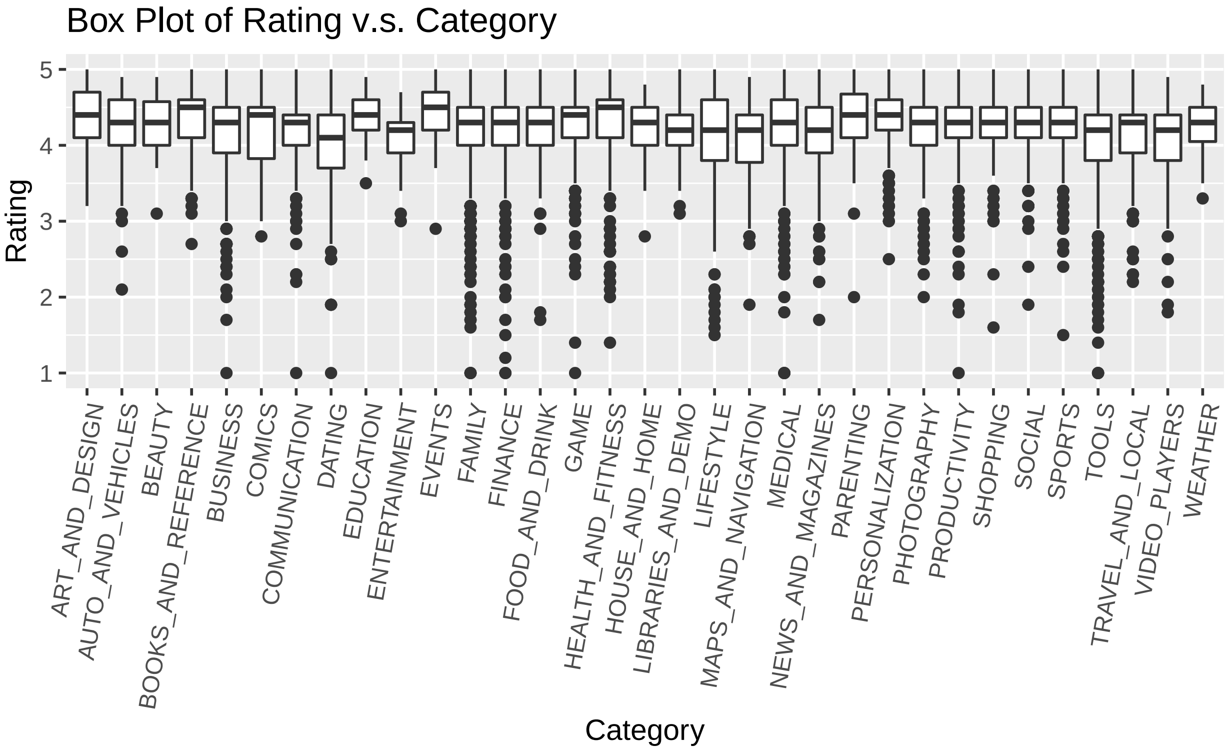
Several techniques and tools will be utilized to do exploratory data analysis and help me know more about the data. First, I plot a bar chart to visualize the number of applications in each category.



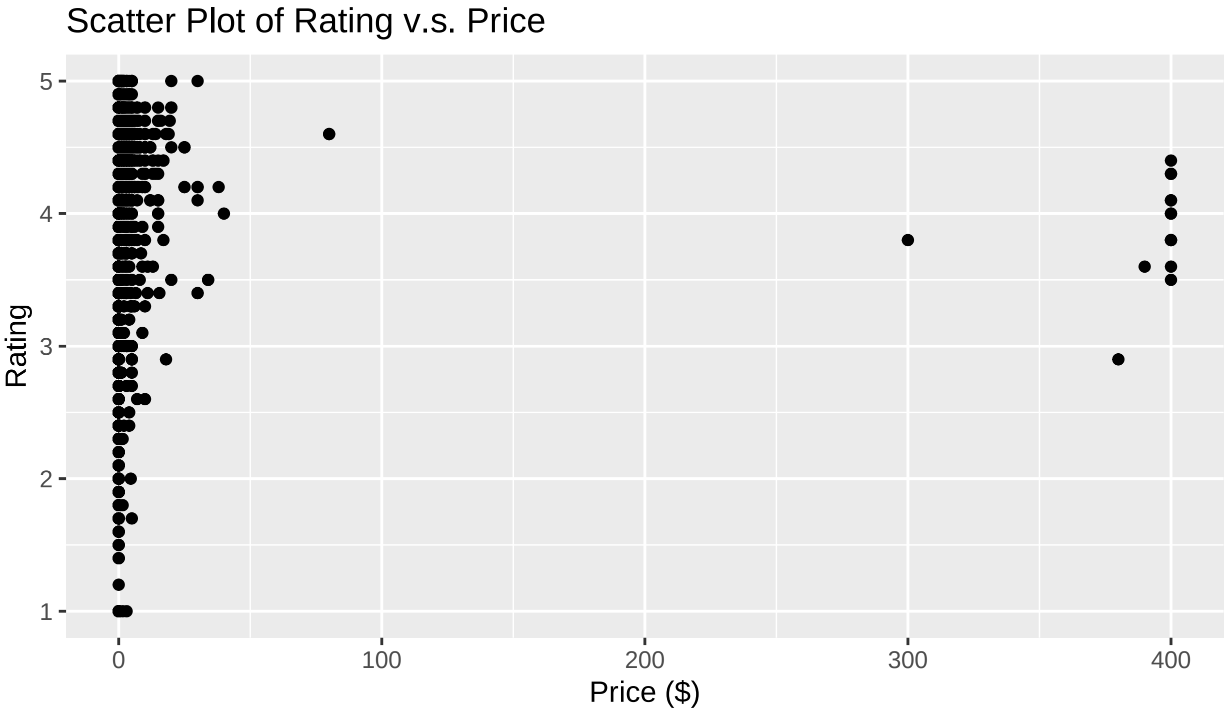
From the bar chart above, it can be seen that *Family*, *Game*, and *Tools* are the top 3 categories that have the most applications.


Histogram of Rating

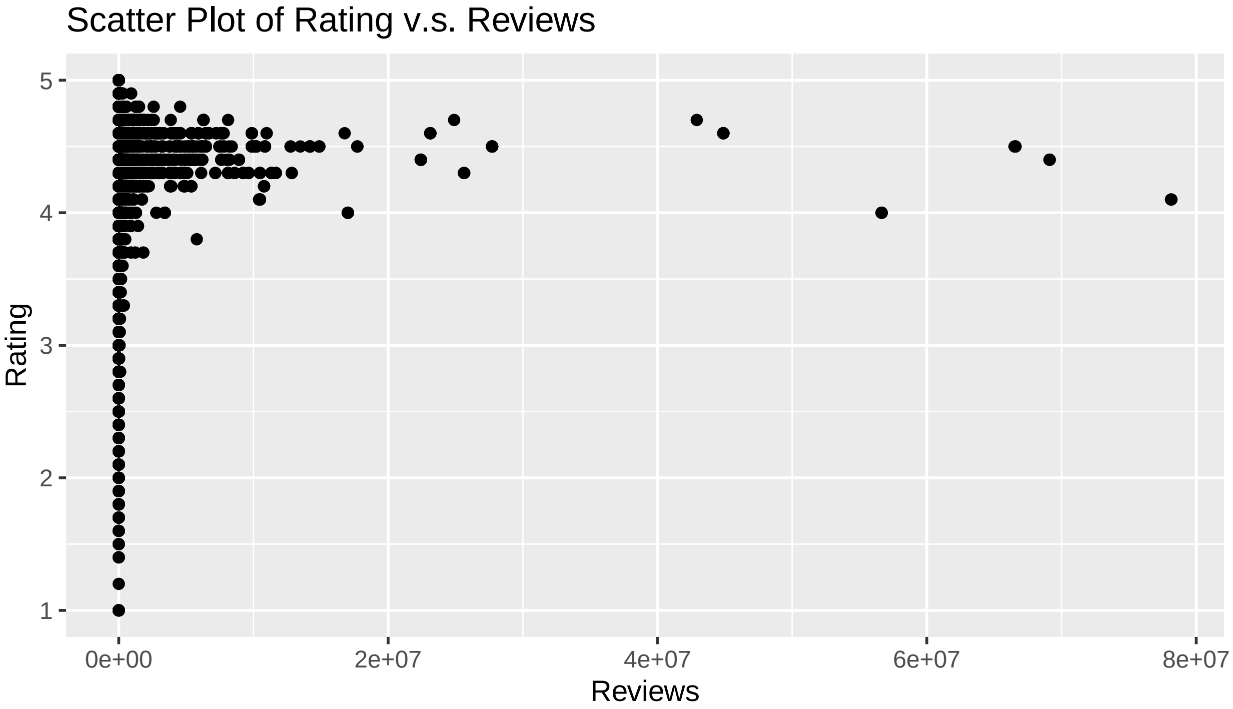
From the histogram above, it shows that most applications' user rating is above 4.0.


Box Plot of Rating v.s. Category

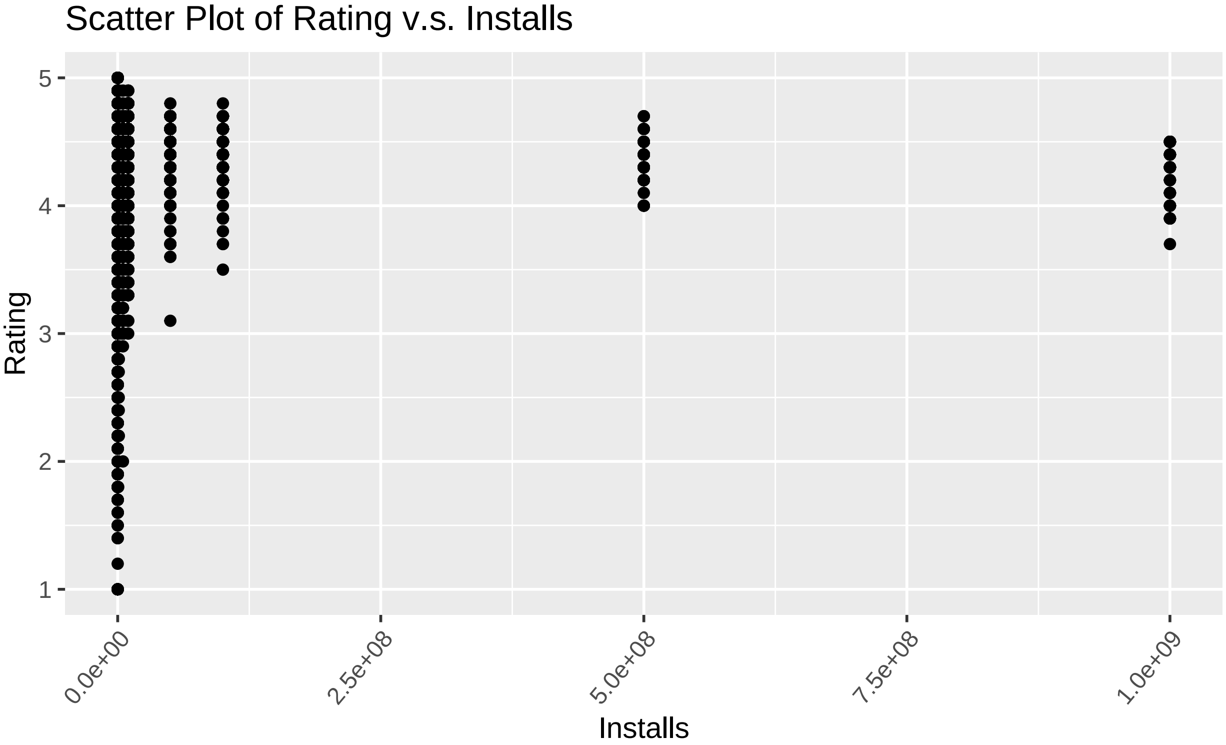
Constructing a box plot can show whether some categories tend to have higher ratings than others. From the box plot above, it seems that the distribution of each category is similar; no obvious differences between categories.



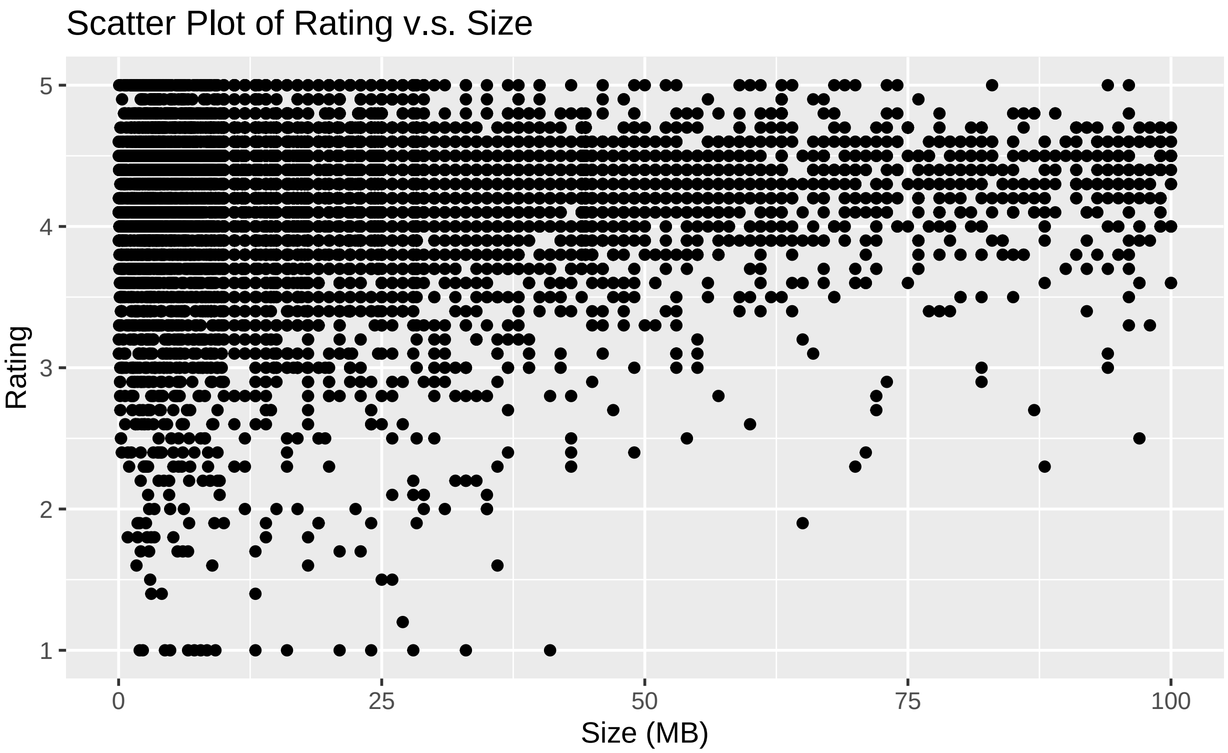
The scatter plot above shows that almost all applications' prices are below \$50; however, there are some applications that cost about \$400, which is extremely expensive. The non-free applications don't necessarily have higher or lower ratings than free applications, but most of the non-free ones have user ratings higher than 3.5.



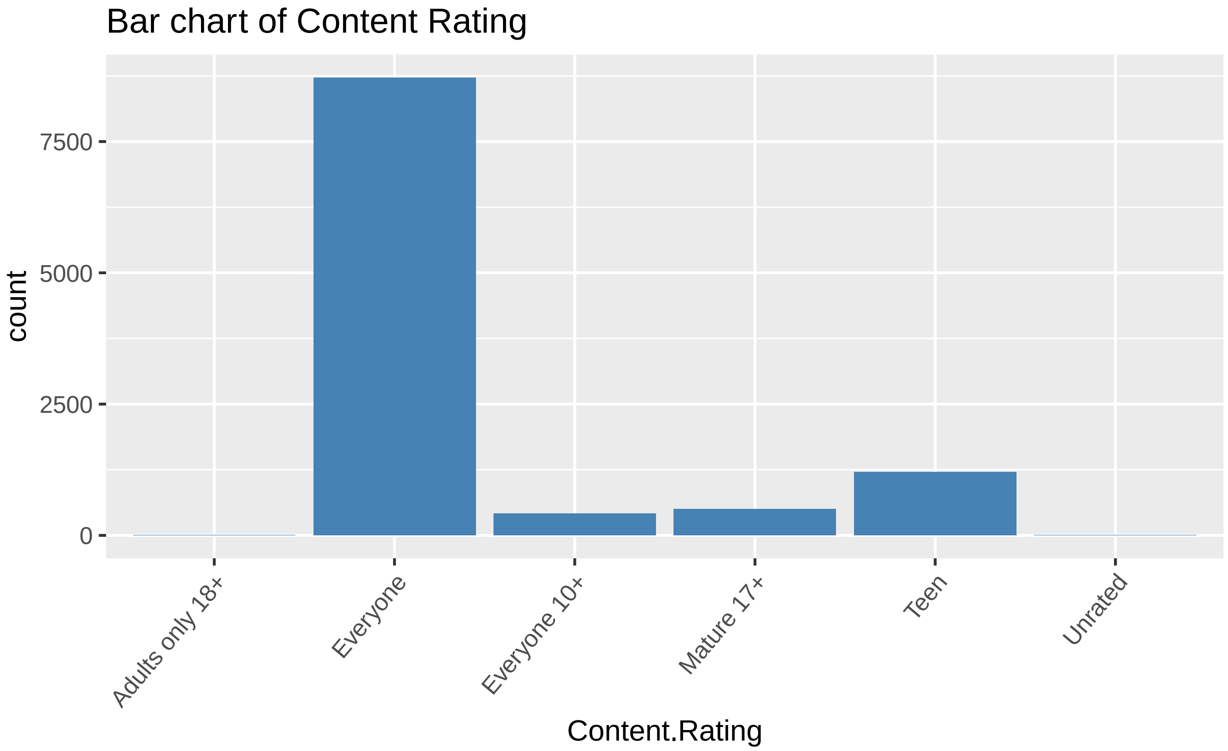
From the scatter plot above, it can be observed that most applications' number of reviews are below 20,000,000. Most applications that have user reviews have good user ratings (above 4.0).



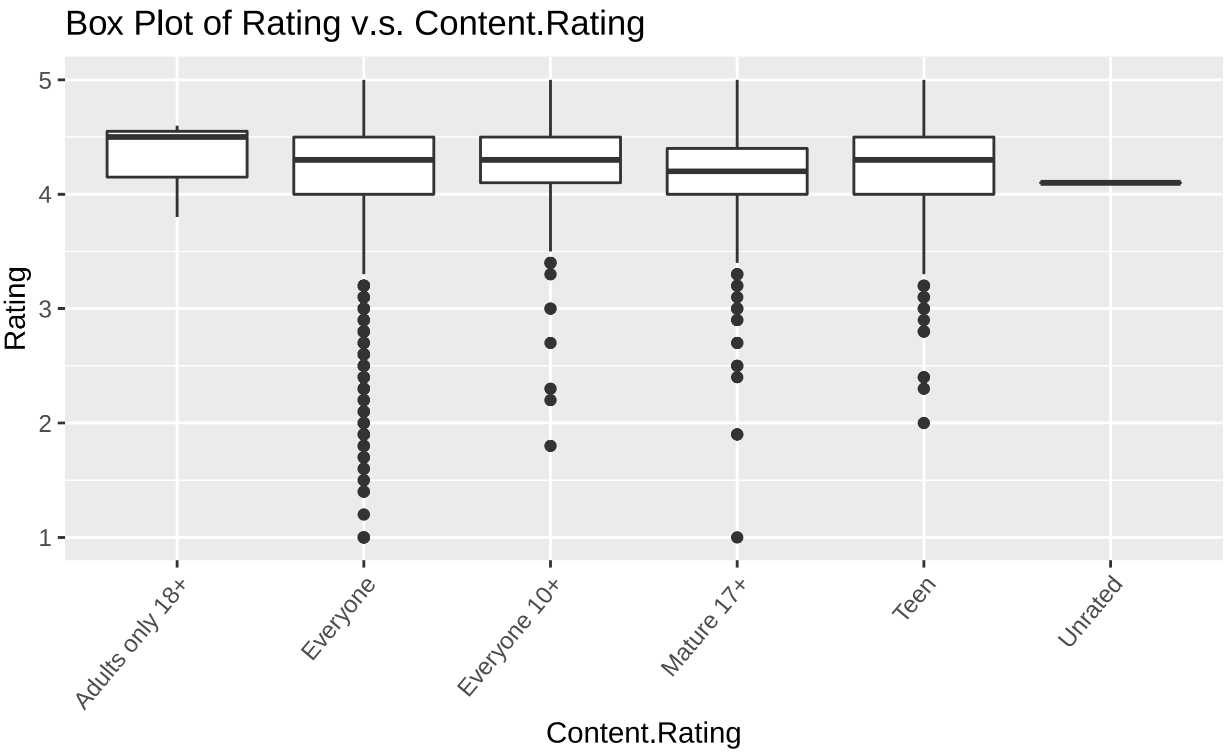
For the Installs field in the dataset, because it was originally a field with factor levels, e.g. 500+, 1000+, 5000+, etc., I transformed those factor levels into a numeric threshold of the number of installs. For example, the points on the line of (Installs = 500,000,000) actually mean the numbers of installs are between 500,000,000 and the next level, which is 1,000,000,000. From this scatter plot, I can tell that most popular applications (Installs > 100,000,000) have user ratings higher than 3.5, and the majority of the ratings are higher than 4.0.



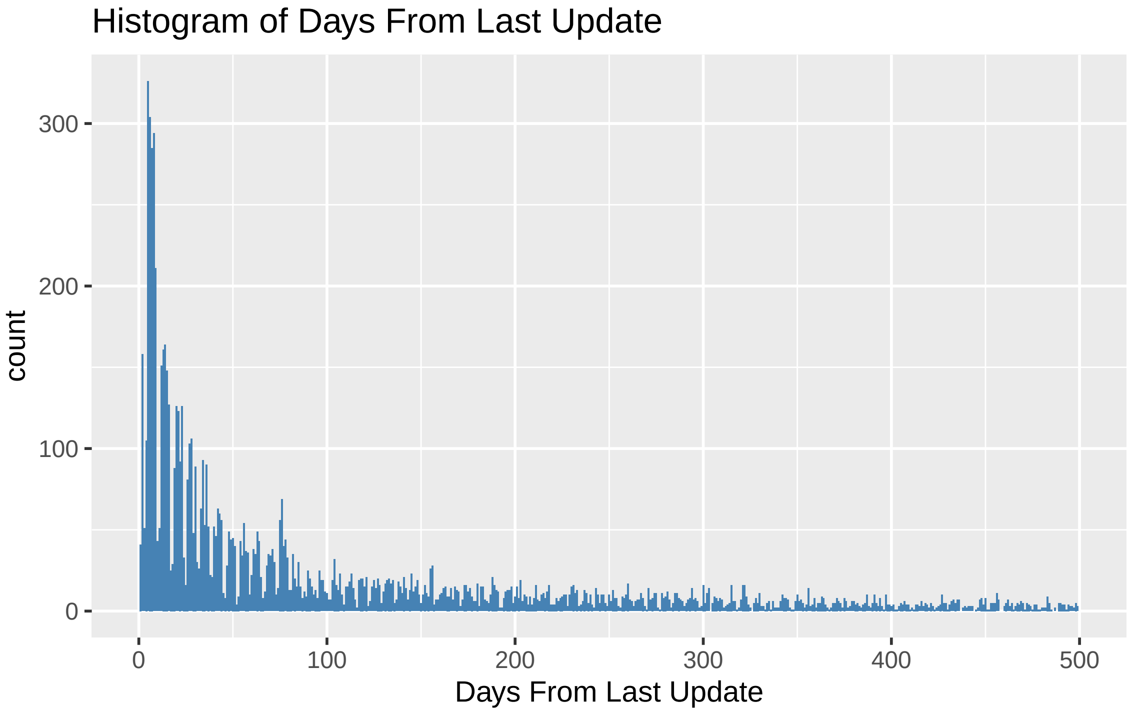
The scatter plot obviously shows that the bigger the sizes of the applications are, the higher the user ratings they have. This relationship may worth trying using regression models to analyze it.



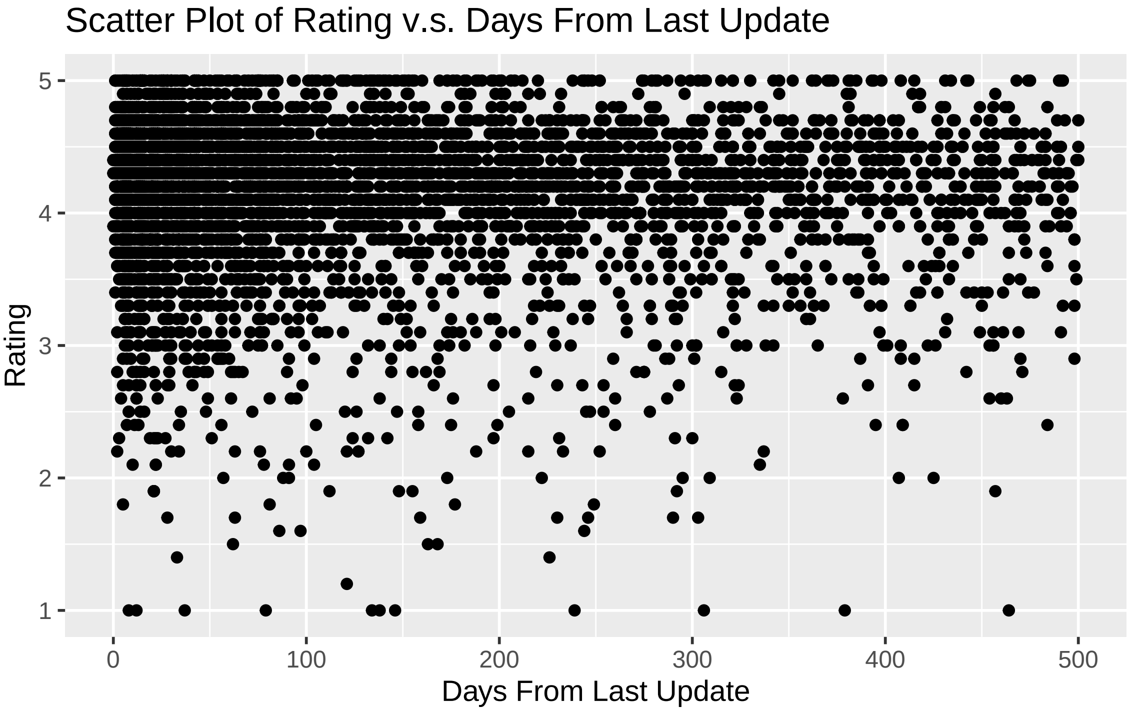
The bar chart shows that most of the applications' content is for everyone, and only very few applications are rated Adults Only or Unrated.



The box plot above implies that there is no obvious difference in user ratings between different content ratings. The content ratings of Adults only 18+ and Unrated have almost no outliers in the box plot, but it is because very few applications are rated in these 2 levels; I can observe that from the previous bar chart of content rating.



After data processing, the field DaysFromLastUpdate has a wide range of data. A portion of applications' last update dates is more than 2000 days ago, which were more than 5 years. Many of these applications are very likely to be out-dated and no longer compatible with current mobile devices. Therefore, I focus on applications whose last update dates within 500 days ago. From the histogram of rating, I can tell that a large part of applications was updated less than 100 days ago. This may indicate that many applications are updated frequently.



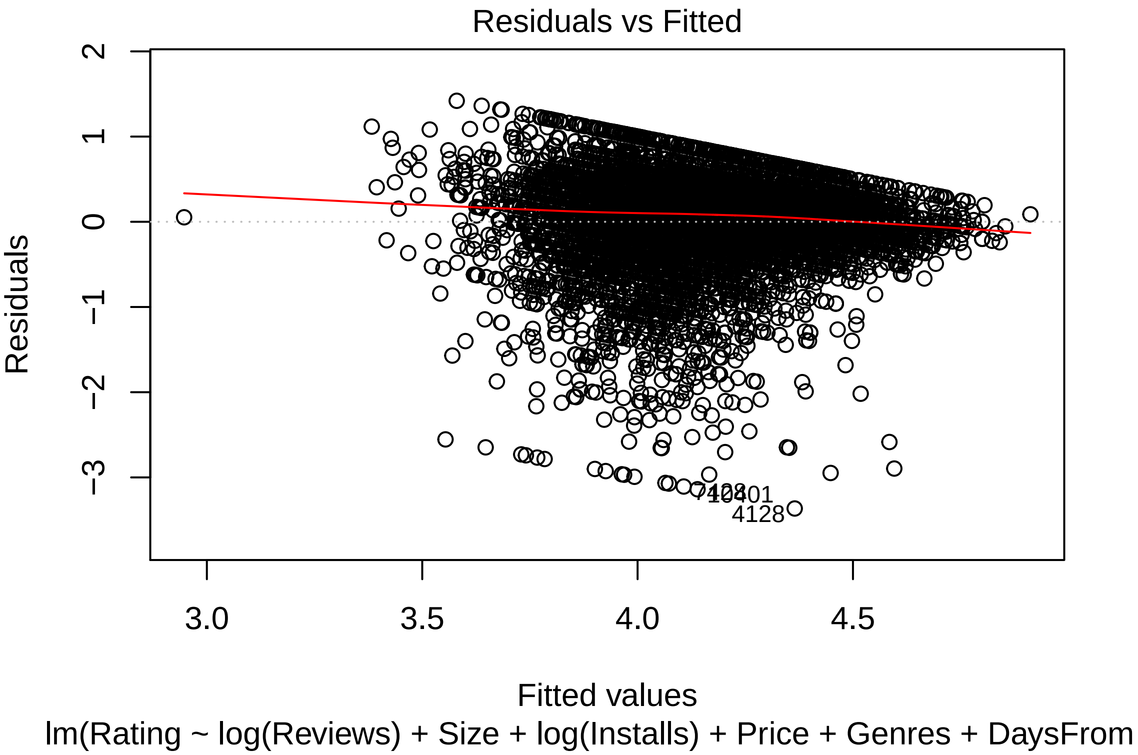
There is a very dense part in the top left corner of the scatter plot, implying that a part of applications that are updated frequently has good user ratings.

4.2 Multiple Regression

I have gained some understanding of the dataset from the previous part. Next, I will use multiple regression models to analyze the data and know more about what the features of a well-rated are. I first attempt to manually select variables for the regression model and then apply some variable selection strategies to select variables.

4.2.1 Manual Variable Selection

```
formula = Rating ~ log(Reviews) + Size + log(Installs) + Price +
  Genres + DaysFromLastUpdate
```



In the beginning, I used the formula with all independent variables. Then I removed the insignificant ones and doing log transformations. After those steps, I got the summary of the model and its scatter plot of residuals v.s. fitted values. The complete summary of this model is shown in **Appendix 7.2**. During the process, I removed *Content Rating* because this variable is not significant in the model. *Category* is also removed because no category has a significant effect on the model. On the other hand, I keep *Genres* because the result shows that some genres' coefficients' p-values are extremely small (<0.001), meaning it is significant that those genres have an influence on user ratings. In addition, because the ranges of *Reviews* and *Installs* are considerably big, I performed log transformations for *Reviews* and *Installs* and then ran the model. The model has slightly higher R-Square (although it is still very small), slightly smaller residual standard error.

Next, I will apply 3 variable selection techniques based on Akaike Information Criterion (AIC): *Forward Selection*, *Backward Elimination*, and *Stepwise Selection*.

4.2.2 Forward Selection

After running Forward Selection in R, the formula I got is:

```
formula = Rating ~ log(Reviews) + log(Installs) + Category +
  DaysFromLastUpdate + Size + Price
```

The complete process of Forward Selection is in **Appendix 7.3**.

4.2.3 Backward Elimination

After running Backward Elimination in R, the formula I got is:

```
formula = Rating ~ Category + log(Reviews) + Size + log(Installs) + Price +
  DaysFromLastUpdate
```

The complete process of Backward Elimination is in **Appendix 7.4**.

4.2.4 Stepwise Selection

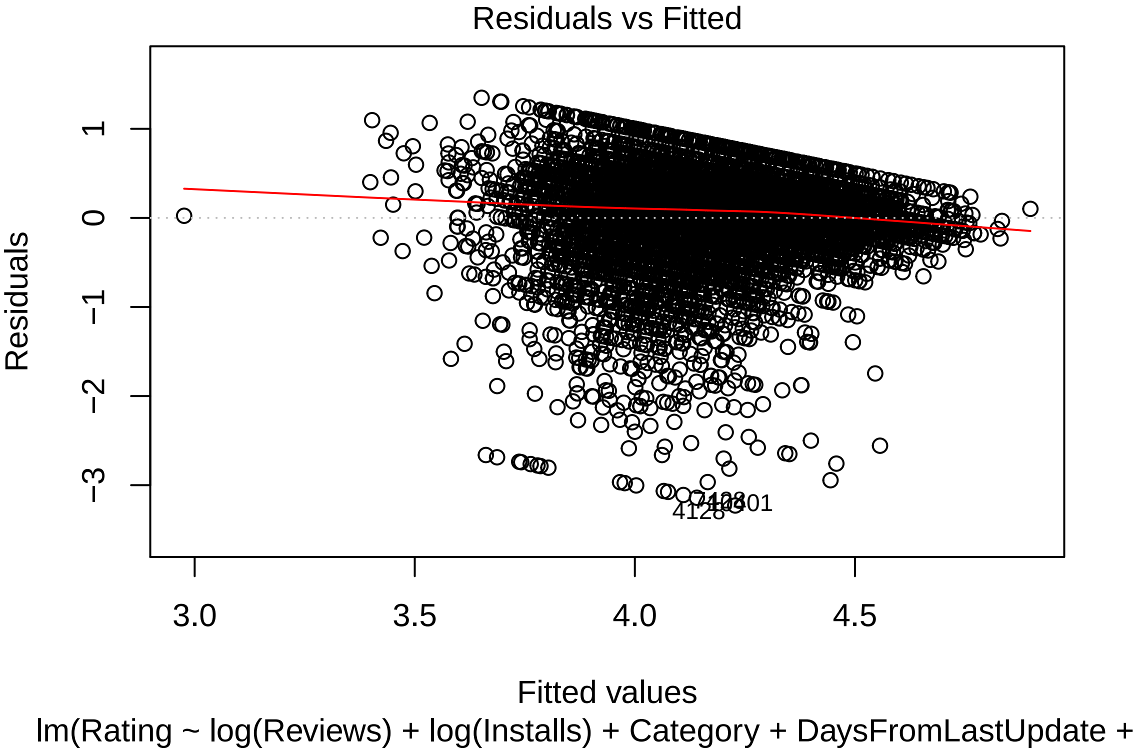
After running Stepwise Selection in R, the formula I got is:

```
formula = Rating ~ log(Reviews) + log(Installs) + Category +
  DaysFromLastUpdate + Size + Price
```

The complete process of Stepwise Selection is in **Appendix 7.5**.

The formulas obtained from 3 variable selection techniques are the same. I use this formula to fit the data and get a summary of the regression model:

```
##
## Call:
## lm(formula = Rating ~ log(Reviews) + log(Installs) + Category +
##     DaysFromLastUpdate + Size + Price, data = app.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2281 -0.1649  0.0525  0.2595  1.3481
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.886e+00  6.489e-02  75.287  < 2e-16 ***
## log(Reviews)                 1.570e-01  4.579e-03  34.286  < 2e-16 ***
## log(Installs)               -1.379e-01  4.558e-03 -30.251  < 2e-16 ***
## CategoryAUTO_AND_VEHICLES   -1.727e-01  8.160e-02  -2.117 0.034326 *
## CategoryBEAUTY              -3.924e-02  9.438e-02  -0.416 0.677636
## CategoryBOOKS_AND_REFERENCE -5.453e-02  6.975e-02  -0.782 0.434378
## CategoryBUSINESS            -2.660e-01  6.589e-02  -4.038 5.44e-05 ***
## CategoryCOMICS              -3.084e-01  8.632e-02  -3.573 0.000355 ***
## CategoryCOMMUNICATION       -3.301e-01  6.563e-02  -5.029 5.02e-07 ***
## CategoryDATING              -4.810e-01  6.892e-02  -6.979 3.17e-12 ***
## CategoryEDUCATION           -8.888e-02  7.116e-02  -1.249 0.211702
## CategoryENTERTAINMENT       -3.887e-01  7.162e-02  -5.428 5.86e-08 ***
## CategoryEVENTS               3.678e-02  9.255e-02   0.397 0.691069
## CategoryFAMILY              -2.129e-01  6.136e-02  -3.470 0.000523 ***
## CategoryFINANCE             -3.123e-01  6.561e-02  -4.761 1.96e-06 ***
## CategoryFOOD_AND_DRINK      -2.876e-01  7.522e-02  -3.824 0.000132 ***
## CategoryGAME                -2.281e-01  6.260e-02  -3.643 0.000271 ***
## CategoryHEALTH_AND_FITNESS  -1.978e-01  6.611e-02  -2.992 0.002776 ** 
## CategoryHOUSE_AND_HOME      -2.162e-01  8.089e-02  -2.673 0.007521 ** 
## CategoryLIBRARIES_AND_DEMO  -1.211e-01  8.396e-02  -1.443 0.149135    
## CategoryLIFESTYLE           -2.781e-01  6.573e-02  -4.230 2.35e-05 ***
## CategoryMAPS_AND_NAVIGATION -3.828e-01  7.351e-02  -5.207 1.96e-07 ***
## CategoryMEDICAL             -1.757e-01  6.528e-02  -2.692 0.007124 ** 
## CategoryNEWS_AND_MAGAZINES  -3.227e-01  6.755e-02  -4.778 1.80e-06 ***
## CategoryPARENTING           -3.246e-02  8.980e-02  -0.361 0.717775    
## CategoryPERSONALIZATION     -9.336e-02  6.587e-02  -1.417 0.156428    
## CategoryPHOTOGRAPHY         -2.703e-01  6.580e-02  -4.108 4.03e-05 ***
## CategoryPRODUCTIVITY        -2.170e-01  6.518e-02  -3.329 0.000874 ***
## CategorySHOPPING            -2.240e-01  6.747e-02  -3.320 0.000905 ***
## CategorySOCIAL              -2.690e-01  6.700e-02  -4.016 5.98e-05 ***
## CategorySPORTS              -2.723e-01  6.579e-02  -4.139 3.52e-05 ***
## CategoryTOOLS               -3.318e-01  6.257e-02  -5.304 1.16e-07 ***
## CategoryTRAVEL_AND_LOCAL    -2.916e-01  6.781e-02  -4.300 1.73e-05 ***
## CategoryVIDEO_PLAYERS       -3.392e-01  7.078e-02  -4.792 1.68e-06 ***
## CategoryWEATHER             -2.349e-01  8.122e-02  -2.893 0.003829 ** 
## DaysFromLastUpdate          -1.583e-04  1.344e-05 -11.785  < 2e-16 ***
## Size                        -1.117e-03  2.744e-04  -4.071 4.72e-05 ***
## Price                       -1.034e-03  3.113e-04  -3.322 0.000898 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4722 on 9328 degrees of freedom
##   (1474 observations deleted due to missingness)
## Multiple R-squared:  0.1633, Adjusted R-squared:  0.1599 
## F-statistic: 49.19 on 37 and 9328 DF,  p-value: < 2.2e-16
```



The scatter plot of residuals v.s. fitted values shows that although the points do not seem to spread randomly, the red line is quite close to the line of 0. From this point of view, the model is not bad; however, the R-Square is extremely low (0.1633), indicating this model is still not a good fit. Compared to the regression model whose variables were selected manually, the regression model using the formula gained by the variable selection approaches above has higher R-Square ($0.1633 > 0.09317$), although it is still very low. The latter model also has slightly lower residual standard error than the former. Thus, I conclude that the latter model is the better one and will use it for further discussion. Nevertheless, the model's R-Square and Adjusted R-Square are both very low; it may imply that there are still many factors influencing an application's user rating but not included in the dataset. Some of the factors may even be difficult to be quantified or categorized.

Next, I will inspect the coefficients of variables in the model. The Intercept, which can be regarded as the base value, is 4.886. Considering the highest rating possible is 5.0, 4.886 is a very high rating score. This may also explain why most variables' coefficients are negative: the model sets a very high base value of the rating score of an application, and this value will be reduced by most of the other variables to its final rating. Besides the *Intercept*, other variables which are significant in the model are listed below:

Variable	Estimate Coefficient
log(Review)	0.157
log(Installs)	-0.1379
DaysFromLastUpdate	-0.0001583
Size	-0.001117
Price	-0.001034

Variable	Estimate Coefficient
Category: AUTO_AND_VEHICLES	-0.1727
Category: BUSINESS	-0.266
Category: COMICS	-0.3084
Category: COMMUNICATION	-0.3301
Category: DATING	-0.481
Category: ENTERTAINMENT	-0.3887
Category: FAMILY	-0.2129
Category: FINANCE	-0.3123
Category: FOOD_AND_DRINK	-0.2876
Category: GAME	-0.2281
Category: HEALTH_AND_FITNESS	-0.1978
Category: HOUSE_AND_HOME	-0.2162
Category: LIFESTYLE	-0.2781
Category: MAPS_AND_NAVIGATION	-0.3828
Category: MEDICAL	-0.1757
Category: NEWS_AND_MAGAZINES	-0.3227
Category: PHOTOGRAPHY	-0.2703
Category: PRODUCTIVITY	-0.217
Category: SHOPPING	-0.224
Category: SOCIAL	-0.269
Category: SPORTS	-0.2723
Category: TOOLS	-0.3318
Category: TRAVEL_AND_LOCAL	-0.2916
Category: VIDEO_PLAYERS	-0.3392
Category: WEATHER	-0.2349

For all coefficients above, their absolute values are smaller than 0.5, which is a very small scale. This is because the scale of the dependent variable, Rating, is small; it is only from 1 to 5. Thus, this makes the coefficients of independent variables very small. Some variables include very big values, e.g. *DaysFromLastUpdate*, and that makes their coefficients even smaller.

log(Reviews)'s coefficient is 0.157, indicating the number of reviews has a positive effect on the user rating. The more reviews an application has, the more likely it has a good rating. On the other hand, it is not intuitive that *log(Installs)* has a negative coefficient. What I expected was that the more installs of an application mean that it was more popular, so it should have a better user rating. But the negative coefficient may be able to explain why the applications which have more than 1,000,000,000 installs don't have user ratings above 4.5. It is possible that as an application has a great number of users, there are usually some people who are not satisfied. This also implies that user ratings may not always be the best standard to evaluate an application: it is because a popular application should be considered successful, but the most popular applications generally don't have very high ratings.

The attribute *DaysFromLastUpdate* which represents the number of days from the date of the last update also has a negative coefficient. Evaluating this model, I want an attribute that can represent how frequent an application is being updated. Considering the data I want, *DaysFromLastUpdate* is not the ideal field for it. I need the data of each application's previous update date and calculate the average update period for analysis. *Size* and *Price* both have a negative coefficient very close to 0. Although these two variables are significant in the summary of the model, they can be regarded as no effect on user ratings because the coefficients are very close to 0.

Furthermore, all coefficients in the attribute *Category* are negative. Hence, generally, the categories have lower coefficients are harder to get better user ratings. From the table above, DATING is the category with the lowest coefficient, meaning that this category has lower ratings in general. Besides DATING, MAPS_AND_NAVIGATION and ENTERTAINMENT also have lower coefficients than others. So, I know that these three categories generally have lower ratings than others. This can be the case that users are more sensitive to these kinds of applications; for example, users may be more likely to get mad and give low ratings if a navigation application leads them to a wrong place than a news application with out-dated news.

Compared with other categories, AUTO_AND_VEHICLES and MEDICAL have higher coefficients, meaning that they usually have higher ratings than other categories.

4.3 Inferential Tools for Multiple Regression

The 95% confidence intervals of independent variables of the regression model:

```
##                                  2.5 %       97.5 %
## (Intercept)                  4.7583183457  5.0127222209
## log(Reviews)                 0.1480096247  0.1659603401
## log(Installs)               -0.1468121490 -0.1289438429
## CategoryAUTO_AND_VEHICLES   -0.3326500150 -0.0127542772
## CategoryBEAUTY              -0.2242450352  0.1457748184
## CategoryBOOKS_AND_REFERENCE -0.1912534564  0.0821980236
## CategoryBUSINESS            -0.3951781318 -0.1368796997
## CategoryCOMICS              -0.4776149905 -0.1392150130
## CategoryCOMMUNICATION       -0.4587507432 -0.2014343499
## CategoryDATING              -0.6160800294 -0.3458989382
## CategoryEDUCATION           -0.2283742010  0.0506130051
## CategoryENTERTAINMENT       -0.5291461698 -0.2483454776
## CategoryEVENTS              -0.1446316525  0.2181906781
## CategoryFAMILY              -0.3331908001 -0.0926328311
## CategoryFINANCE             -0.4409350277 -0.1837288451
## CategoryFOOD_AND_DRINK      -0.4350382897 -0.1401605168
## CategoryGAME                -0.3507606909 -0.1053565806
## CategoryHEALTH_AND_FITNESS  -0.3274027780 -0.0682271550
## CategoryHOUSE_AND_HOME      -0.3748063660 -0.0576917082
## CategoryLIBRARIES_AND_DEMO  -0.2857255946  0.0434522669
## CategoryLIFESTYLE           -0.4069418307 -0.1492345542
## CategoryMAPS_AND_NAVIGATION -0.5268512833 -0.2386665755
## CategoryMEDICAL             -0.3036715109 -0.0477434420
## CategoryNEWS_AND_MAGAZINES  -0.4551349541 -0.1903212832
## CategoryPARENTING           -0.2084749756  0.1435627746
## CategoryPERSONALIZATION     -0.2224899679  0.0357639802
## CategoryPHOTOGRAPHY         -0.3992926683 -0.1413114848
## CategoryPRODUCTIVITY        -0.3447760609 -0.0892358660
## CategorySHOPPING            -0.3562485373 -0.0917188495
## CategorySOCIAL              -0.4003717079 -0.1377066132
## CategorySPORTS              -0.4012579409 -0.1433405601
## CategoryTOOLS               -0.4544874674 -0.2091915751
## CategoryTRAVEL_AND_LOCAL    -0.4245264730 -0.1586621368
## CategoryVIDEO_PLAYERS       -0.4779360905 -0.2004350873
## CategoryWEATHER             -0.3941624565 -0.0757323812
## DaysFromLastUpdate          -0.0001846676 -0.0001319945
## Size                        -0.0016551104 -0.0005792111
## Price                       -0.0016441517 -0.0004238467
```

The confidence intervals of regression model coefficients provide the range of each variable's effect on user rating. For example, DATING's 95% confidence interval is (-0.616, -0.346), which is a wide range considering the range of user ratings is only (1, 5). Although the confidence intervals show us to what degree these variables can affect user ratings, it still doesn't make predicting or interpreting user ratings easier.

5 Conclusions

Generally, predicting whether an application has a good rating based on attributes such as prices, sizes, numbers of installs, and days from the last update is difficult; even though these variables are significant in the model. The low R-Square value reflects this situation. But from the analysis, an app with more reviews is more likely to have a higher rating. Besides that, some categories such as AUTO_AND_VEHICLES and MEDICAL which have higher coefficients in the regression model tend to have higher ratings than other categories, while DATING, MAPS_AND_NAVIGATION, and ENTERTAINMENT tend to have lower ratings than others because of lower coefficients in the model.

The difference between manually selecting variables and 3 variable selection methods demonstrates that the variable selection techniques do a better job on forming formulas since the techniques involves criteria such as AIC in it and are more sophisticated than manual selection.

6 References

Ramsey, F.L., & Schafer, D.W. (2013). *The statistical sleuth: A Course in Methods of Data Analysis* (3rd ed). Boston, MA: Brooks/Cole.

Pimentel, H., Bray, N., Melsted, P., & Pachter, L. (2015). *Sleuth package | R Documentation*. Retrieved from [https://www.rdocumentation.org/packages/sleuth/versions/0.27.3]

7 Appendix

7.1 R Code of Data Cleaning and Processing

```
## Data cleaning and processing

pp.df <- data.frame(stringsAsFactors = FALSE)
app.df <- read.csv("google-play-store-apps/googleplaystore.csv")
View(app.df)
# delete the row with problematic data
app.df <- app.df[-c(10473),]

library(stringr)
library(naniar)
# cleaning Size: get rid of "M" or "k",
# replace "Varies with device" with mean of size in each category,
# and turn numeric

# first, replace "Varies with device" with N/A (and replace it with mean later)
app.df <- replace_with_na(app.df, replace=list(Size=c("Varies with device")))

# remove the unit (M or k), make this columns numeric and use M as the unit
size.temp <- as.character(app.df$Size)
for(i in 1:length(size.temp))
{
  if (is.na(size.temp[i]))
  {
    next
  }
  else if(str_sub(size.temp[i], -1, -1) == "M")
  {
    size.temp[i] <- str_sub(size.temp[i], end=-2)
  }
  else if (str_sub(size.temp[i], -1, -1) == "k")
  {
    size.temp[i] <- str_sub(size.temp[i], end=-2)
```

```r
    size.temp[i] <- as.numeric(size.temp[i])/1024
  }
  else if (str_sub(size.temp[i], -1, -1) == "+")
  {
    size.temp[i] <- str_sub(size.temp[i], end=-2)
  }
}
size.temp <- as.numeric(size.temp)

df.temp <-app.df
df.temp$Size <-  size.temp

# calculate the mean size value of each category
size.meanByCatg <- aggregate(Size ~ Category, df.temp, FUN=mean)

# replace N/A (originally Varies with device) with the mean size value of each categor
for(i in 1:nrow(df.temp))
{
  if (is.na(df.temp[i,]$Size))
  {
    df.temp[i,]$Size <- size.meanByCatg[size.meanByCatg$Category==df.temp[i,]$Category,]
  }
}
app.df <- df.temp
app.df$Size <- round(app.df$Size, digits = 3)

View(app.df$Installs)
# cleaning installs: get rid of "+", and turn numeric
install.temp <- as.character(app.df$Installs)
for (i in 1:length(install.temp))
{
  if(str_sub(install.temp[i], -1, -1) == "+")
  {
    install.temp[i] <- str_sub(install.temp[i], end=-2)
  }
}
View(install.temp)
app.df$Installs <-install.temp
app.df$Installs <- as.numeric(gsub(",","",app.df$Installs))

# cleaning price: remove "$", and turn numeric
price.temp <- as.character(app.df$Price)
for (i in 1:length(price.temp))
{
  if(str_sub(price.temp[i], 1, 1) == "$")
  {
    price.temp[i] <- str_sub(price.temp[i], start=2, end=-1)
  }
}
app.df$Price <- as.numeric(price.temp)

# transform the Last.Updated column into date type
app.df$Last.Updated <- as.Date(app.df$Last.Updated, format="%B %d, %Y")
max(app.df$Last.Updated) # 2018-08-08 (the lastest updated date of all apps)

# add a column "DaysFromLastUpdate"
#(assume today is the day of the lastest updated date of all apps in the dataset (2018
app.df$DaysFromLastUpdate <- as.Date("2018-08-08") - as.Date(app.df$Last.Updated, format

# turn Reviews into numeric
app.df$Reviews <- as.numeric(as.character(app.df$Reviews))

# output data frame to csv file
write.csv(app.df, "apps.csv")
```

7.2 Regression Model Summary of Manual-Selected Formula

```
##
## Call:
## lm(formula = Rating ~ log(Reviews) + Size + log(Installs) + Price +
##     Genres + DaysFromLastUpdate, data = app.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3648 -0.1617  0.0498  0.2568  1.4200
##
## Coefficients:
##                                               Estimate Std. Error t value
## (Intercept)                                  4.6506920  0.0379857 122.433
## log(Reviews)                                 0.1608892  0.0046453  34.635
## Size                                        -0.0008937  0.0002891  -3.092
## log(Installs)                               -0.1410146  0.0046208 -30.517
## Price                                       -0.0009737  0.0003103  -3.138
## GenresAction;Action & Adventure              0.0901725  0.1168184   0.772
## GenresAdventure                             -0.0923274  0.0605748  -1.524
## GenresAdventure;Action & Adventure           0.0173900  0.1329141   0.131
## GenresAdventure;Brain Games                  0.2054064  0.4713775   0.436
## GenresAdventure;Education                   -0.2043285  0.3338538  -0.612
## GenresArcade                                 0.0351236  0.0411334   0.854
## GenresArcade;Action & Adventure              0.0435559  0.1240759   0.351
## GenresArcade;Pretend Play                    0.2595720  0.4713572   0.551
## GenresArt & Design                           0.2446522  0.0688506   3.553
## GenresArt & Design;Creativity                0.2182180  0.1799640   1.213
## GenresArt & Design;Pretend Play              0.0370315  0.3340512   0.111
## GenresAuto & Vehicles                        0.0655055  0.0614449   1.066
## GenresBeauty                                 0.2013396  0.0778024   2.588
## GenresBoard                                  0.0188583  0.0779177   0.242
## GenresBoard;Action & Adventure              -0.1440330  0.2728701  -0.528
## GenresBoard;Brain Games                      0.1226106  0.1243486   0.986
## GenresBoard;Pretend Play                     0.7776989  0.4714358   1.650
## GenresBooks & Reference                      0.1821853  0.0444635   4.097
## GenresBooks & Reference;Education           -0.0585056  0.3338850  -0.175
## GenresBusiness                              -0.0279468  0.0385054  -0.726
## GenresCard                                  -0.1061753  0.0746112  -1.423
## GenresCard;Action & Adventure               -0.1547778  0.3338335  -0.464
## GenresCard;Brain Games                       0.3490376  0.4713537   0.741
## GenresCasino                                 0.0426627  0.0814087   0.524
## GenresCasual                                -0.0932056  0.0427949  -2.178
## GenresCasual;Action & Adventure             -0.0506078  0.1057378  -0.479
## GenresCasual;Brain Games                     0.3222219  0.1332052   2.419
## GenresCasual;Creativity                      0.1051004  0.1796979   0.585
## GenresCasual;Education                       0.0885497  0.2728606   0.325
## GenresCasual;Music & Video                   0.0474985  0.3339385   0.142
## GenresCasual;Pretend Play                   -0.0190673  0.0881917  -0.216
## GenresComics                                -0.0840455  0.0681635  -1.233
## GenresComics;Creativity                      0.6816123  0.4714169   1.446
## GenresCommunication                         -0.0947536  0.0376597  -2.516
## GenresCommunication;Creativity               0.1134431  0.4713265   0.241
## GenresDating                                -0.2439180  0.0433705  -5.624
## GenresEducation                              0.1583983  0.0348805   4.541
## GenresEducation;Action & Adventure           0.3344911  0.1939299   1.725
## GenresEducation;Brain Games                  0.0571309  0.2366153   0.241
## GenresEducation;Creativity                   0.4007271  0.1799011   2.227
## GenresEducation;Education                    0.2027979  0.0712600   2.846
## GenresEducation;Music & Video                0.1926774  0.2120968   0.908
## GenresEducation;Pretend Play                 0.2509015  0.1014086   2.474
## GenresEducational                           -0.2189294  0.0871443  -2.512
## GenresEducational;Action & Adventure         0.0501195  0.2366497   0.212
## GenresEducational;Brain Games                0.2236115  0.1939942   1.153
## GenresEducational;Creativity                -0.0316384  0.2121147  -0.149
## GenresEducational;Education                  0.1990945  0.0806480   2.469
## GenresEducational;Pretend Play               0.0629504  0.1137169   0.554
## GenresEntertainment                         -0.0772822  0.0337151  -2.292
## GenresEntertainment;Action & Adventure        0.0838792  0.2728540   0.307
## GenresEntertainment;Brain Games               0.0754914  0.1684287   0.448
## GenresEntertainment;Creativity                0.2880760  0.2730259   1.055
## GenresEntertainment;Education                 0.3354583  0.4713707   0.712
## GenresEntertainment;Music & Video            -0.0038390  0.0942830  -0.041
## GenresEntertainment;Pretend Play             -0.2930765  0.3338064  -0.878
## GenresEvents                                  0.2761302  0.0755488   3.655
## GenresFinance                                -0.0765069  0.0376906  -2.030
## GenresFood & Drink                           -0.0524815  0.0523212  -1.003
## GenresHealth & Fitness                        0.0359803  0.0378997   0.949
## GenresHealth & Fitness;Action & Adventure    -0.3844143  0.4713004  -0.816
## GenresHealth & Fitness;Education              0.3827459  0.4713086   0.812
## GenresHouse & Home                            0.0202870  0.0603355   0.336
## GenresLibraries & Demo                        0.1174415  0.0645098   1.821
## GenresLifestyle                              -0.0403725  0.0381170  -1.059
## GenresLifestyle;Education                     0.0312104  0.4712646   0.066
## GenresLifestyle;Pretend Play                 -0.1924963  0.4714248  -0.408
## GenresMaps & Navigation                      -0.1461560  0.0501620  -2.914
## GenresMedical                                 0.0614859  0.0371413   1.655
## GenresMusic                                  -0.1149053  0.1057227  -1.087
## GenresMusic & Audio;Music & Video             0.4820077  0.4714688   1.022
## GenresMusic;Music & Video                     0.2441319  0.2729653   0.894
## GenresNews & Magazines                       -0.0853709  0.0412987  -2.067
## GenresParenting                               0.2226772  0.0792613   2.809
## GenresParenting;Brain Games                  -0.2066463  0.4715085  -0.438
## GenresParenting;Education                    -0.1308783  0.2730616  -0.479
## GenresParenting;Music & Video                 0.3345294  0.1940586   1.724
## GenresPersonalization                         0.1423751  0.0379941   3.747
## GenresPhotography                            -0.0357879  0.0375840  -0.952
## GenresProductivity                            0.0196303  0.0370331   0.530
## GenresPuzzle                                  0.1184197  0.0496919   2.383
## GenresPuzzle;Action & Adventure               0.0089543  0.2119459   0.042
## GenresPuzzle;Brain Games                      0.1325062  0.1108576   1.195
## GenresPuzzle;Creativity                       0.0993841  0.3336856   0.298
## GenresPuzzle;Education                        0.5325961  0.4713370   1.130
## GenresRacing                                 -0.0640567  0.0547893  -1.169
## GenresRacing;Action & Adventure               0.1277128  0.1082092   1.180
## GenresRacing;Pretend Play                     0.6269469  0.4714412   1.330
## GenresRole Playing                           -0.0787592  0.0521508  -1.510
## GenresRole Playing;Action & Adventure        -0.0184681  0.1796398  -0.103
## GenresRole Playing;Brain Games                0.0612302  0.4712245   0.130
## GenresRole Playing;Pretend Play              -0.1004224  0.2120035  -0.474
## GenresShopping                                0.0117135  0.0408009   0.287
## GenresSimulation                             -0.0565763  0.0421016  -1.344
## GenresSimulation;Action & Adventure           0.1613515  0.1440606   1.120
## GenresSimulation;Education                 -0.0333502  0.2730253  -0.122
## GenresSimulation;Pretend Play              -0.0769348  0.2367093  -0.325
## GenresSocial                               -0.0354085  0.0395749  -0.895
## GenresSports                               -0.0382074  0.0365409  -1.046
## GenresSports;Action & Adventure            -0.0749078  0.2366733  -0.317
## GenresStrategy                             -0.1123270  0.0526538  -2.133
## GenresStrategy;Action & Adventure           0.2284010  0.3336773   0.684
## GenresStrategy;Creativity                  -0.1377020  0.4712676  -0.292
## GenresStrategy;Education                    0.5178804  0.4713265   1.099
## GenresTools                                -0.0937933  0.0327398  -2.865
## GenresTools;Education                       0.2469349  0.4714528   0.524
## GenresTravel & Local                       -0.0566436  0.0410119  -1.381
## GenresTravel & Local;Action & Adventure     0.0425885  0.4713886   0.090
## GenresTrivia                               -0.1480882  0.0927569  -1.597
## GenresVideo Players & Editors              -0.1011685  0.0460862  -2.195
## GenresVideo Players & Editors;Creativity   -0.2791730  0.3337922  -0.836
## GenresVideo Players & Editors;Music & Video -0.1709597  0.2729816  -0.626
## GenresWeather                               0.0003213  0.0607197   0.005
## GenresWord                                  0.1210428  0.0924192   1.310
## DaysFromLastUpdate                         -0.0001528  0.0000135 -11.318
##                                             Pr(>|t|)    
## (Intercept)                                  < 2e-16 ***
## log(Reviews)                                 < 2e-16 ***
## Size                                        0.001996 ** 
## log(Installs)                                < 2e-16 ***
## Price                                       0.001709 ** 
## GenresAction;Action & Adventure             0.440191    
## GenresAdventure                             0.127496    
## GenresAdventure;Action & Adventure          0.895908    
## GenresAdventure;Brain Games                 0.663023    
## GenresAdventure;Education                   0.540533    
## GenresArcade                                0.393186    
## GenresArcade;Action & Adventure             0.725565    
## GenresArcade;Pretend Play                   0.581859    
## GenresArt & Design                          0.000382 ***
## GenresArt & Design;Creativity               0.225327    
## GenresArt & Design;Pretend Play             0.911733    
## GenresAuto & Vehicles                       0.286413    
## GenresBeauty                                0.009673 ** 
## GenresBoard                                 0.808763    
## GenresBoard;Action & Adventure              0.597620    
## GenresBoard;Brain Games                     0.324148    
## GenresBoard;Pretend Play                    0.099051 .  
## GenresBooks & Reference                     4.21e-05 ***
## GenresBooks & Reference;Education           0.860905    
## GenresBusiness                              0.467987
## GenresCard                                  0.154756
## GenresCard;Action & Adventure               0.642918
## GenresCard;Brain Games                      0.459015
## GenresCasino                                0.600252
## GenresCasual                                0.029434 *
## GenresCasual;Action & Adventure             0.632223
## GenresCasual;Brain Games                    0.015583 *
## GenresCasual;Creativity                     0.558648
## GenresCasual;Education                      0.745549
## GenresCasual;Music & Video                  0.886896
## GenresCasual;Pretend Play                   0.828834
## GenresComics                                0.217608
## GenresComics;Creativity                     0.148245
## GenresCommunication                         0.011885 *
## GenresCommunication;Creativity              0.809802
## GenresDating                                1.92e-08 ***
## GenresEducation                             5.67e-06 ***
## GenresEducation;Action & Adventure          0.084596 .
## GenresEducation;Brain Games                 0.809211
## GenresEducation;Creativity                  0.025939 *
## GenresEducation;Education                   0.004439 **
## GenresEducation;Music & Video               0.363669
## GenresEducation;Pretend Play                0.013373 *
## GenresEducational                           0.012013 *
## GenresEducational;Action & Adventure        0.832277
## GenresEducational;Brain Games               0.249075
## GenresEducational;Creativity                0.881433
## GenresEducational;Education                 0.013579 *
## GenresEducational;Pretend Play              0.579886
## GenresEntertainment                         0.021916 *
## GenresEntertainment;Action & Adventure      0.758535
## GenresEntertainment;Brain Games             0.654012
## GenresEntertainment;Creativity              0.291396
## GenresEntertainment;Education               0.476690
## GenresEntertainment;Music & Video           0.967522
## GenresEntertainment;Pretend Play            0.379976
## GenresEvents                                0.000259 ***
## GenresFinance                               0.042399 *
## GenresFood & Drink                          0.315856
## GenresHealth & Fitness                      0.342465
## GenresHealth & Fitness;Action & Adventure   0.414724
## GenresHealth & Fitness;Education            0.416760
## GenresHouse & Home                          0.736701
## GenresLibraries & Demo                      0.068712 .
## GenresLifestyle                                 0.289549
## GenresLifestyle;Education                       0.947199
## GenresLifestyle;Pretend Play                    0.683042
## GenresMaps & Navigation                         0.003581 **
## GenresMedical                                   0.097866 .
## GenresMusic                                     0.277129
## GenresMusic & Audio;Music & Video               0.306640
## GenresMusic;Music & Video                       0.371147
## GenresNews & Magazines                          0.038747 *
## GenresParenting                                 0.004974 **
## GenresParenting;Brain Games                     0.661203
## GenresParenting;Education                       0.631737
## GenresParenting;Music & Video                   0.084767 .
## GenresPersonalization                           0.000180 ***
## GenresPhotography                               0.341014
## GenresProductivity                              0.596072
## GenresPuzzle                                    0.017189 *
## GenresPuzzle;Action & Adventure                 0.966302
## GenresPuzzle;Brain Games                        0.232007
## GenresPuzzle;Creativity                         0.765834
## GenresPuzzle;Education                          0.258519
## GenresRacing                                    0.242375
## GenresRacing;Action & Adventure                 0.237935
## GenresRacing;Pretend Play                       0.183600
## GenresRole Playing                              0.131021
## GenresRole Playing;Action & Adventure           0.918119
## GenresRole Playing;Brain Games                  0.896618
## GenresRole Playing;Pretend Play                 0.635737
## GenresShopping                                  0.774050
## GenresSimulation                                0.179044
## GenresSimulation;Action & Adventure             0.262732
## GenresSimulation;Education                      0.902782
## GenresSimulation;Pretend Play                   0.745175
## GenresSocial                                    0.370960
## GenresSports                                    0.295769
## GenresSports;Action & Adventure                 0.751628
## GenresStrategy                                  0.032925 *
## GenresStrategy;Action & Adventure               0.493679
## GenresStrategy;Creativity                       0.770144
## GenresStrategy;Education                        0.271896
## GenresTools                                     0.004182 **
## GenresTools;Education                           0.600448
## GenresTravel & Local                            0.167267
## GenresTravel & Local;Action & Adventure         0.928014
## GenresTrivia                                    0.110407
## GenresVideo Players & Editors               0.028174 *
## GenresVideo Players & Editors;Creativity    0.402970
## GenresVideo Players & Editors;Music & Video 0.531154
## GenresWeather                               0.995778
## GenresWord                                  0.190325
## DaysFromLastUpdate                           < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4705 on 9246 degrees of freedom
##   (1474 observations deleted due to missingness)
## Multiple R-squared:  0.1765, Adjusted R-squared:  0.1659
## F-statistic: 16.65 on 119 and 9246 DF,  p-value: < 2.2e-16
```

7.3 Forward Selection

```
## Start:  AIC=-12421.38
## Rating ~ 1
##
##                        Df Sum of Sq    RSS    AIC
## + log(Reviews)          1   109.456 2376.5 -12841
## + Category             32    75.886 2410.1 -12648
## + DaysFromLastUpdate    1    50.687 2435.3 -12612
## + Genres              114   103.012 2382.9 -12590
## + log(Installs)         1    35.549 2450.4 -12554
## + Size                  1    15.498 2470.4 -12478
## + Content.Rating        5     5.998 2479.9 -12434
## + Price                 1     1.193 2484.8 -12424
## <none>                               2485.9 -12421
##
## Step:  AIC=-12841.12
## Rating ~ log(Reviews)
##
##                        Df Sum of Sq    RSS    AIC
## + log(Installs)         1   202.729 2173.8 -13674
## + Category             32    67.183 2309.3 -13046
## + Genres              114    99.938 2276.6 -13016
## + DaysFromLastUpdate    1    19.385 2357.1 -12916
## + Content.Rating        5     3.623 2372.9 -12845
## + Size                  1     0.646 2375.8 -12842
## <none>                               2376.5 -12841
## + Price                 1     0.380 2376.1 -12841
## - log(Reviews)          1   109.456 2485.9 -12421
##
## Step:  AIC=-13674.24
## Rating ~ log(Reviews) + log(Installs)
##
##                         Df Sum of Sq    RSS    AIC
## + Category           32    59.730 2114.0 -13871
## + Genres            114    96.072 2077.7 -13870
## + DaysFromLastUpdate   1    25.190 2148.6 -13781
## + Content.Rating       5     7.240 2166.5 -13696
## + Price                1     2.447 2171.3 -13683
## <none>                             2173.8 -13674
## + Size                 1     0.267 2173.5 -13673
## - log(Installs)        1   202.729 2376.5 -12841
## - log(Reviews)         1   276.636 2450.4 -12554
##
## Step:  AIC=-13871.2
## Rating ~ log(Reviews) + log(Installs) + Category
##
##                        Df Sum of Sq    RSS    AIC
## + DaysFromLastUpdate  1    27.901 2086.1 -13994
## + Price               1     2.204 2111.8 -13879
## + Size                1     0.704 2113.3 -13872
## <none>                            2114.0 -13871
## + Content.Rating      5     2.054 2112.0 -13870
## + Genres             88    39.011 2075.0 -13870
## - Category           32    59.730 2173.8 -13674
## - log(Installs)       1   195.276 2309.3 -13046
## - log(Reviews)        1   267.046 2381.1 -12759
##
## Step:  AIC=-13993.64
## Rating ~ log(Reviews) + log(Installs) + Category + DaysFromLastUpdate
##
##                        Df Sum of Sq    RSS    AIC
## + Size                1     3.566 2082.6 -14008
## + Price               1     2.332 2083.8 -14002
## + Content.Rating      5     2.385 2083.7 -13994
## <none>                            2086.1 -13994
## + Genres             88    37.274 2048.8 -13986
## - DaysFromLastUpdate  1    27.901 2114.0 -13871
## - Category           32    62.441 2148.6 -13781
## - log(Installs)       1   200.007 2286.1 -13138
## - log(Reviews)        1   256.600 2342.7 -12909
##
## Step:  AIC=-14007.66
## Rating ~ log(Reviews) + log(Installs) + Category + DaysFromLastUpdate +
##     Size
##
```

```
##                      Df Sum of Sq    RSS    AIC
## + Price               1     2.461 2080.1 -14017
## <none>                            2082.6 -14008
## + Content.Rating      5     2.005 2080.6 -14007
## + Genres             88    36.016 2046.5 -13995
## - Size                1     3.566 2086.1 -13994
## - DaysFromLastUpdate  1    30.763 2113.3 -13872
## - Category           32    64.455 2147.0 -13786
## - log(Installs)       1   202.024 2284.6 -13142
## - log(Reviews)        1   260.160 2342.7 -12907
##
## Step:  AIC=-14016.73
## Rating ~ log(Reviews) + log(Installs) + Category + DaysFromLastUpdate +
##     Size + Price
##
##                      Df Sum of Sq    RSS    AIC
## <none>                            2080.1 -14017
## + Content.Rating      5     2.012 2078.1 -14016
## - Price               1     2.461 2082.6 -14008
## + Genres             88    35.820 2044.3 -14003
## - Size                1     3.695 2083.8 -14002
## - DaysFromLastUpdate  1    30.969 2111.1 -13880
## - Category           32    64.199 2144.3 -13796
## - log(Installs)       1   204.074 2284.2 -13142
## - log(Reviews)        1   262.130 2342.2 -12907
```

```
##
## Call:
## lm(formula = Rating ~ log(Reviews) + log(Installs) + Category +
##     DaysFromLastUpdate + Size + Price, data = app.df)
##
## Coefficients:
##                 (Intercept)                 log(Reviews)
##                   4.8855203                    0.1569850
##               log(Installs)    CategoryAUTO_AND_VEHICLES
##                  -0.1378780                   -0.1727021
##              CategoryBEAUTY  CategoryBOOKS_AND_REFERENCE
##                  -0.0392351                   -0.0545277
##            CategoryBUSINESS               CategoryCOMICS
##                  -0.2660289                   -0.3084150
##       CategoryCOMMUNICATION               CategoryDATING
##                  -0.3300925                   -0.4809895
##           CategoryEDUCATION        CategoryENTERTAINMENT
##                  -0.0888806                   -0.3887458
##              CategoryEVENTS               CategoryFAMILY
##                   0.0367795                   -0.2129118
##             CategoryFINANCE       CategoryFOOD_AND_DRINK
##                  -0.3123319                   -0.2875994
##                CategoryGAME   CategoryHEALTH_AND_FITNESS
##                  -0.2280586                   -0.1978150
##      CategoryHOUSE_AND_HOME   CategoryLIBRARIES_AND_DEMO
##                  -0.2162490                   -0.1211367
##           CategoryLIFESTYLE  CategoryMAPS_AND_NAVIGATION
##                  -0.2780882                   -0.3827589
##             CategoryMEDICAL   CategoryNEWS_AND_MAGAZINES
##                  -0.1757075                   -0.3227281
##           CategoryPARENTING      CategoryPERSONALIZATION
##                  -0.0324561                   -0.0933630
##         CategoryPHOTOGRAPHY         CategoryPRODUCTIVITY
##                  -0.2703021                   -0.2170060
##            CategorySHOPPING               CategorySOCIAL
##                  -0.2239837                   -0.2690392
##              CategorySPORTS                CategoryTOOLS
##                  -0.2722993                   -0.3318395
##    CategoryTRAVEL_AND_LOCAL        CategoryVIDEO_PLAYERS
##                  -0.2915943                   -0.3391856
##             CategoryWEATHER           DaysFromLastUpdate
##                  -0.2349474                   -0.0001583
##                        Size                        Price
##                  -0.0011172                   -0.0010340
```

7.4 Backward Elimination

```
## Start:  AIC=-13994.88
## Rating ~ Category + log(Reviews) + Size + log(Installs) + Price + 
##     Content.Rating + Genres + DaysFromLastUpdate
## 
##                      Df Sum of Sq    RSS    AIC
## - Genres             88    34.125 2078.1 -14016
## - Content.Rating      5     0.317 2044.3 -14003
## <none>                            2044.0 -13995
## - Category            6     2.800 2046.8 -13994
## - Price               1     2.272 2046.2 -13986
## - Size                1     2.332 2046.3 -13986
## - DaysFromLastUpdate  1    28.551 2072.5 -13867
## - log(Installs)       1   205.678 2249.6 -13099
## - log(Reviews)        1   264.000 2308.0 -12859
## 
## Step:  AIC=-14015.8
## Rating ~ Category + log(Reviews) + Size + log(Installs) + Price + 
##     Content.Rating + DaysFromLastUpdate
## 
##                      Df Sum of Sq    RSS    AIC
## - Content.Rating      5     2.012 2080.1 -14017
## <none>                            2078.1 -14016
## - Price               1     2.468 2080.6 -14007
## - Size                1     3.308 2081.4 -14003
## - DaysFromLastUpdate  1    31.088 2109.2 -13879
## - Category           32    58.061 2136.2 -13822
## - log(Installs)       1   205.389 2283.5 -13135
## - log(Reviews)        1   263.714 2341.8 -12899
## 
## Step:  AIC=-14016.73
## Rating ~ Category + log(Reviews) + Size + log(Installs) + Price + 
##     DaysFromLastUpdate
## 
##                      Df Sum of Sq    RSS    AIC
## <none>                            2080.1 -14017
## - Price               1     2.461 2082.6 -14008
## - Size                1     3.695 2083.8 -14002
## - DaysFromLastUpdate  1    30.969 2111.1 -13880
## - Category           32    64.199 2144.3 -13796
## - log(Installs)       1   204.074 2284.2 -13142
## - log(Reviews)        1   262.130 2342.2 -12907
```

```
## 
## Call:
## lm(formula = Rating ~ Category + log(Reviews) + Size + log(Installs) + 
##     Price + DaysFromLastUpdate, data = app.df)
## 
## Coefficients:
##                 (Intercept)    CategoryAUTO_AND_VEHICLES  
##                   4.8855203                   -0.1727021  
##              CategoryBEAUTY  CategoryBOOKS_AND_REFERENCE  
##                  -0.0392351                   -0.0545277  
##            CategoryBUSINESS               CategoryCOMICS  
##                  -0.2660289                   -0.3084150  
##       CategoryCOMMUNICATION               CategoryDATING  
##                  -0.3300925                   -0.4809895  
##           CategoryEDUCATION        CategoryENTERTAINMENT  
##                  -0.0888806                   -0.3887458  
##              CategoryEVENTS               CategoryFAMILY  
##                   0.0367795                   -0.2129118  
##             CategoryFINANCE       CategoryFOOD_AND_DRINK  
##                  -0.3123319                   -0.2875994  
##               CategoryGAME   CategoryHEALTH_AND_FITNESS
##                 -0.2280586                   -0.1978150
##      CategoryHOUSE_AND_HOME   CategoryLIBRARIES_AND_DEMO
##                 -0.2162490                   -0.1211367
##           CategoryLIFESTYLE  CategoryMAPS_AND_NAVIGATION
##                 -0.2780882                   -0.3827589
##             CategoryMEDICAL   CategoryNEWS_AND_MAGAZINES
##                 -0.1757075                   -0.3227281
##           CategoryPARENTING      CategoryPERSONALIZATION
##                 -0.0324561                   -0.0933630
##         CategoryPHOTOGRAPHY         CategoryPRODUCTIVITY
##                 -0.2703021                   -0.2170060
##            CategorySHOPPING               CategorySOCIAL
##                 -0.2239837                   -0.2690392
##              CategorySPORTS                CategoryTOOLS
##                 -0.2722993                   -0.3318395
##    CategoryTRAVEL_AND_LOCAL        CategoryVIDEO_PLAYERS
##                 -0.2915943                   -0.3391856
##             CategoryWEATHER                 log(Reviews)
##                 -0.2349474                    0.1569850
##                        Size                log(Installs)
##                 -0.0011172                   -0.1378780
##                       Price           DaysFromLastUpdate
##                 -0.0010340                   -0.0001583
```

7.5 Stepwise Selection

```
## Start:  AIC=-12421.38
## Rating ~ 1
##
##                       Df Sum of Sq    RSS    AIC
## + log(Reviews)         1   109.456 2376.5 -12841
## + Category            32    75.886 2410.1 -12648
## + DaysFromLastUpdate   1    50.687 2435.3 -12612
## + Genres             114   103.012 2382.9 -12590
## + log(Installs)        1    35.549 2450.4 -12554
## + Size                 1    15.498 2470.4 -12478
## + Content.Rating       5     5.998 2479.9 -12434
## + Price                1     1.193 2484.8 -12424
## <none>                             2485.9 -12421
##
## Step:  AIC=-12841.12
## Rating ~ log(Reviews)
##
##                       Df Sum of Sq    RSS    AIC
## + log(Installs)        1   202.729 2173.8 -13674
## + Category            32    67.183 2309.3 -13046
## + Genres             114    99.938 2276.6 -13016
## + DaysFromLastUpdate   1    19.385 2357.1 -12916
## + Content.Rating       5     3.623 2372.9 -12845
## + Size                 1     0.646 2375.8 -12842
## <none>                             2376.5 -12841
## + Price                1     0.380 2376.1 -12841
## - log(Reviews)         1   109.456 2485.9 -12421
##
## Step:  AIC=-13674.24
## Rating ~ log(Reviews) + log(Installs)
##
##                      Df Sum of Sq    RSS    AIC
## + Category            32    59.730 2114.0 -13871
## + Genres             114    96.072 2077.7 -13870
## + DaysFromLastUpdate   1    25.190 2148.6 -13781
## + Content.Rating       5     7.240 2166.5 -13696
## + Price                1     2.447 2171.3 -13683
## <none>                             2173.8 -13674
## + Size                 1     0.267 2173.5 -13673
## - log(Installs)        1   202.729 2376.5 -12841
## - log(Reviews)         1   276.636 2450.4 -12554
##
## Step:  AIC=-13871.2
## Rating ~ log(Reviews) + log(Installs) + Category
##
##                     Df Sum of Sq    RSS    AIC
## + DaysFromLastUpdate  1    27.901 2086.1 -13994
## + Price               1     2.204 2111.8 -13879
## + Size                1     0.704 2113.3 -13872
## <none>                            2114.0 -13871
## + Content.Rating      5     2.054 2112.0 -13870
## + Genres             88    39.011 2075.0 -13870
## - Category           32    59.730 2173.8 -13674
## - log(Installs)       1   195.276 2309.3 -13046
## - log(Reviews)        1   267.046 2381.1 -12759
##
## Step:  AIC=-13993.64
## Rating ~ log(Reviews) + log(Installs) + Category + DaysFromLastUpdate
##
##                     Df Sum of Sq    RSS    AIC
## + Size                1     3.566 2082.6 -14008
## + Price               1     2.332 2083.8 -14002
## + Content.Rating      5     2.385 2083.7 -13994
## <none>                            2086.1 -13994
## + Genres             88    37.274 2048.8 -13986
## - DaysFromLastUpdate  1    27.901 2114.0 -13871
## - Category           32    62.441 2148.6 -13781
## - log(Installs)       1   200.007 2286.1 -13138
## - log(Reviews)        1   256.600 2342.7 -12909
##
## Step:  AIC=-14007.66
## Rating ~ log(Reviews) + log(Installs) + Category + DaysFromLastUpdate +
##     Size
##
##                      Df Sum of Sq    RSS    AIC
## + Price               1     2.461 2080.1 -14017
## <none>                            2082.6 -14008
## + Content.Rating      5     2.005 2080.6 -14007
## + Genres             88    36.016 2046.5 -13995
## - Size                1     3.566 2086.1 -13994
## - DaysFromLastUpdate  1    30.763 2113.3 -13872
## - Category           32    64.455 2147.0 -13786
## - log(Installs)       1   202.024 2284.6 -13142
## - log(Reviews)        1   260.160 2342.7 -12907
##
## Step:  AIC=-14016.73
## Rating ~ log(Reviews) + log(Installs) + Category + DaysFromLastUpdate +
##     Size + Price
##
##                      Df Sum of Sq    RSS    AIC
## <none>                            2080.1 -14017
## + Content.Rating      5     2.012 2078.1 -14016
## - Price               1     2.461 2082.6 -14008
## + Genres             88    35.820 2044.3 -14003
## - Size                1     3.695 2083.8 -14002
## - DaysFromLastUpdate  1    30.969 2111.1 -13880
## - Category           32    64.199 2144.3 -13796
## - log(Installs)       1   204.074 2284.2 -13142
## - log(Reviews)        1   262.130 2342.2 -12907
```

```
##
## Call:
## lm(formula = Rating ~ log(Reviews) + log(Installs) + Category +
##     DaysFromLastUpdate + Size + Price, data = app.df)
##
## Coefficients:
##                 (Intercept)                 log(Reviews)
##                   4.8855203                    0.1569850
##               log(Installs)    CategoryAUTO_AND_VEHICLES
##                  -0.1378780                   -0.1727021
##              CategoryBEAUTY  CategoryBOOKS_AND_REFERENCE
##                  -0.0392351                   -0.0545277
##            CategoryBUSINESS               CategoryCOMICS
##                  -0.2660289                   -0.3084150
##       CategoryCOMMUNICATION               CategoryDATING
##                  -0.3300925                   -0.4809895
##           CategoryEDUCATION        CategoryENTERTAINMENT
##                  -0.0888806                   -0.3887458
##              CategoryEVENTS               CategoryFAMILY
##                   0.0367795                   -0.2129118
##             CategoryFINANCE       CategoryFOOD_AND_DRINK
##                  -0.3123319                   -0.2875994
##                CategoryGAME   CategoryHEALTH_AND_FITNESS
##                  -0.2280586                   -0.1978150
##      CategoryHOUSE_AND_HOME   CategoryLIBRARIES_AND_DEMO
##                  -0.2162490                   -0.1211367
##           CategoryLIFESTYLE  CategoryMAPS_AND_NAVIGATION
##                  -0.2780882                   -0.3827589
##             CategoryMEDICAL   CategoryNEWS_AND_MAGAZINES
##                  -0.1757075                   -0.3227281
##           CategoryPARENTING      CategoryPERSONALIZATION
##                  -0.0324561                   -0.0933630
##         CategoryPHOTOGRAPHY         CategoryPRODUCTIVITY
##                  -0.2703021                   -0.2170060
##            CategorySHOPPING               CategorySOCIAL
##                  -0.2239837                   -0.2690392
##              CategorySPORTS                CategoryTOOLS
##                  -0.2722993                   -0.3318395
##    CategoryTRAVEL_AND_LOCAL        CategoryVIDEO_PLAYERS
##                  -0.2915943                   -0.3391856
##             CategoryWEATHER           DaysFromLastUpdate
##                  -0.2349474                   -0.0001583
##                        Size                        Price
##                  -0.0011172                   -0.0010340
```