

DATA 621 Final: Predicting A Mobile Application's Success

Group 3

CUNY SPS Spring 2019

Abstract

In an effort to see what developers can focus on to get more downloads of their mobile application, this project will use various modeling techniques to predict an application's success. These models will be built using data about apps on Apple's and Google's stores. This will provide a methodology to scientifically predict how well an app will do in either store when certain metrics are improved.

Keywords: Apps, Apple, Google, Poisson, Quasi

Introduction

In 2014 Dong Nguyen revealed that his game Flappy Bird was earning an average of \$50,000 a day from in-app ads. At that time, Flappy Bird, an addictive and tough game, had been number 1 on the Apple App Store and Google Play Store for almost a month and was gaining more downloads daily. In fact, by January 2014 the app had 50 million downloads, 68,000 reviews, and held the number 1 spot for most downloaded free game in 53 countries.

With millions of apps available for download today, many app developers could only dream of the success that Flappy Bird saw. So, did Nguyen know how to make an app so successful, and is there any way for us to use app statistics to predict the success of an app? Kaggle has a couple of datasets with mobile application statistics that can possibly help us answer these questions. The datasets contain information around Apple apps on the App Store and Google apps on the Play Store; information centered around ratings, app size, installs, and price. Our idea is to take these datasets to see if we can use several regression methods to predict the successfulness of an app.

Methodology

Apple

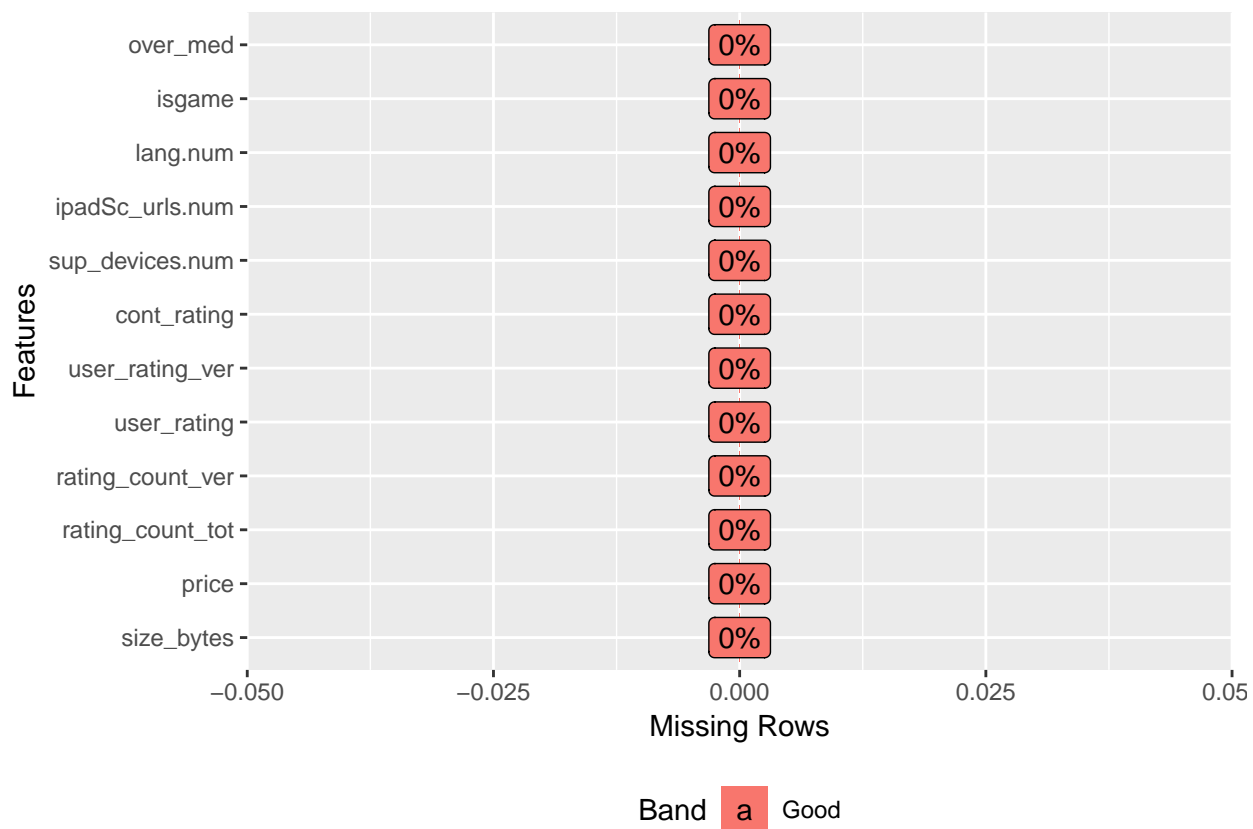
Apple Store Data

For this project, data from the Apple store on 7,197 apps was sourced from Kaggle at this site: <https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps>. The data content on these apps revolved around the apps' size, price, ratings, genre, and number of supporting devices. However, there is no data around how many times the app was downloaded. As a workaround, this study used the total number of ratings as a measure of how many times the app was downloaded and therefore how successful it is.

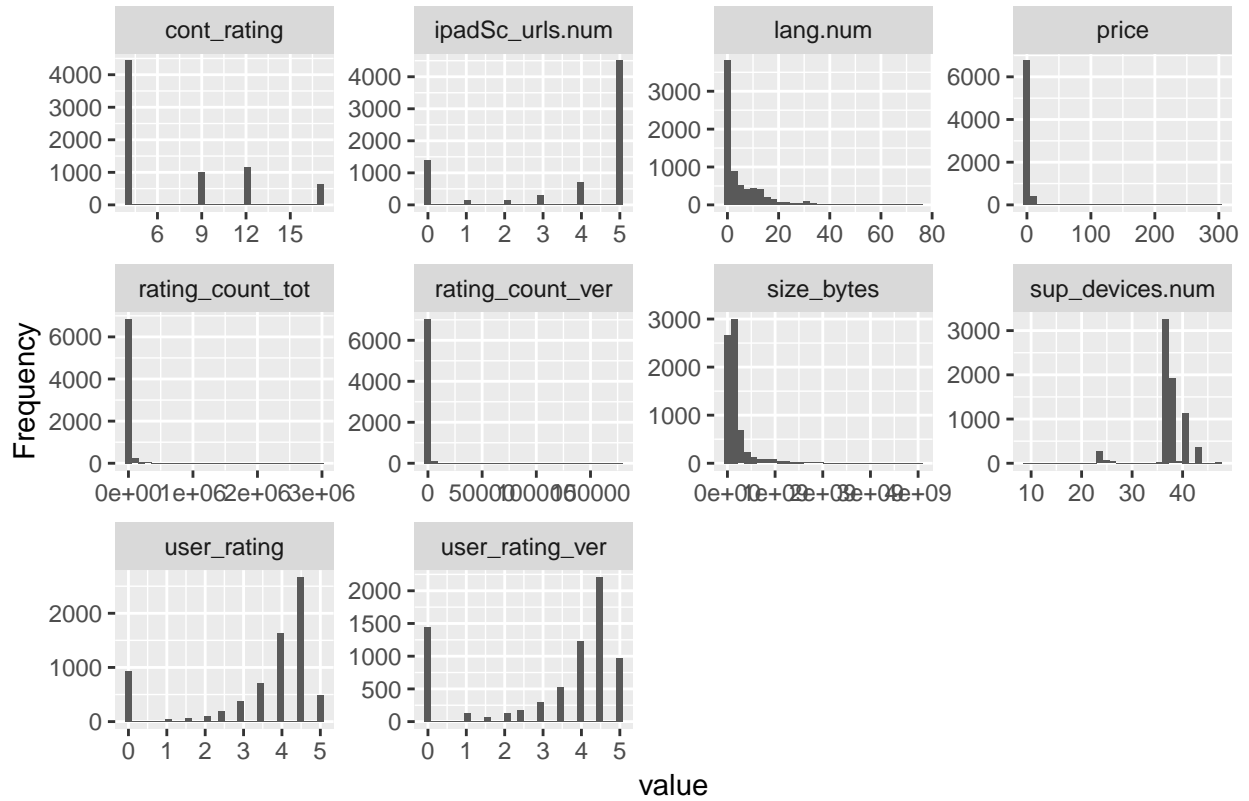
Data Management

Thankfully none of the data in the Apple Store dataset was missing. Two new fields were added to the dataset, one to distinguish if the app was a game or not and another to see if the total ratings count for the

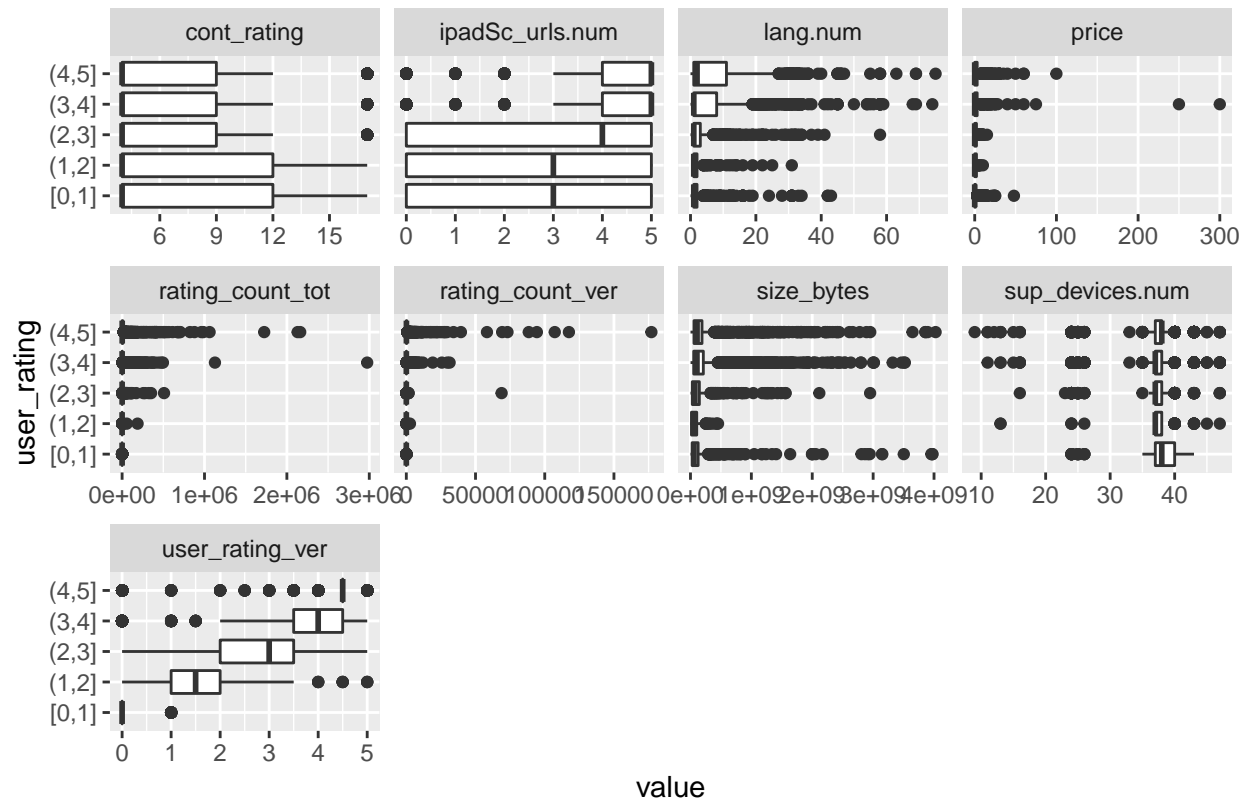
app was over or under the median of the dataset. For the content rating field the '+' symbol was removed and the field was transformed into an integer. Here's a summary of the final dataset.



Histogram of Fields



Boxplot of Fields



NULL

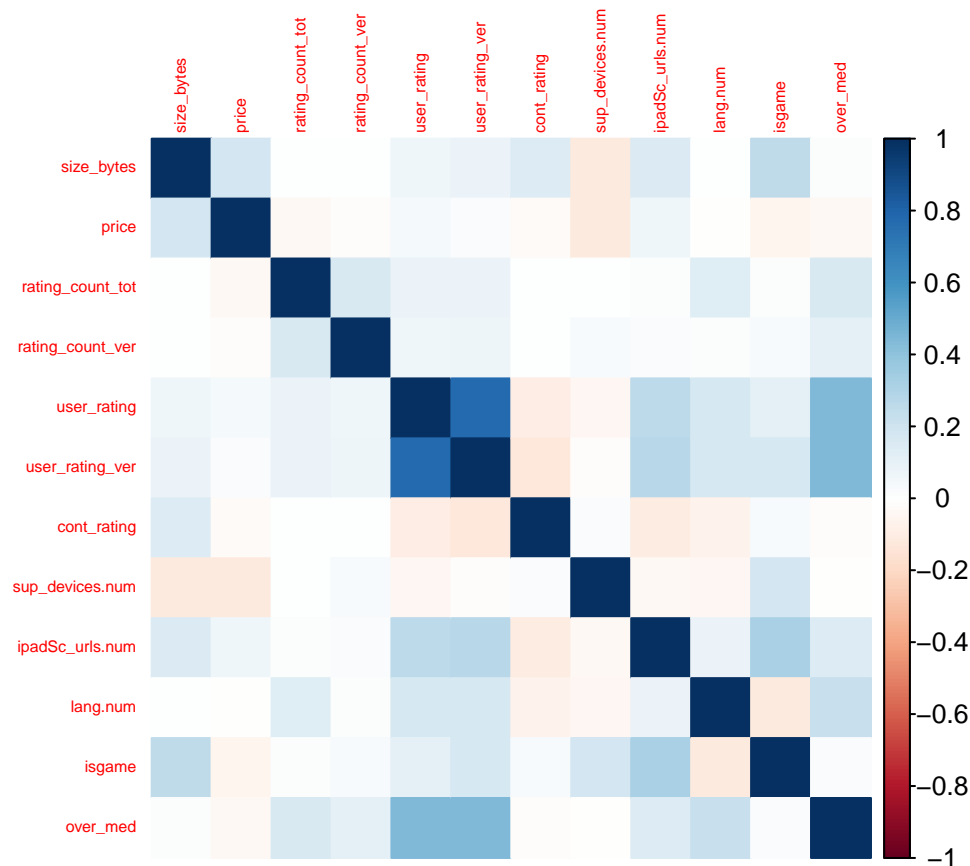
Train – Test Split

Given that the Apple Store dataset is not conveniently split to do model training, it was split by randomly choosing fifty percent of the data to be in the train set and the remaining data to be in the testing set.

Results

Correlation

Before beginning the data analysis, the correlation plot between the metrics was analyzed to see if there was anything worth noting.



Ratings Total Summary

In order to gauge how well the forthcoming models did in predicting the total number of ratings, here's the summary of ratings from the train dataset.

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##         0      26     308   12124    2738 1061624
```

Poisson Model

The first model ran is a Poisson Model.

```
##
## Call:
## glm(formula = rating_count_tot ~ ., family = "poisson", data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -744.94  -122.50   -9.25    2.85   2088.04
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.059e+00  4.270e-03   716.3  <2e-16 ***
## size_bytes    2.904e-10  5.259e-13   552.2  <2e-16 ***
## price        -3.425e-01  1.566e-04 -2187.6  <2e-16 ***
```

```

## rating_count_ver 1.645e-05 6.698e-09 2456.5 <2e-16 ***
## user_rating      5.140e-02 3.837e-04 133.9 <2e-16 ***
## user_rating_ver  1.881e-01 2.353e-04 799.6 <2e-16 ***
## cont_rating      1.494e-02 4.176e-05 357.8 <2e-16 ***
## sup_devices.num  1.529e-02 6.307e-05 242.4 <2e-16 ***
## ipadSc_urls.num -4.476e-02 1.030e-04 -434.3 <2e-16 ***
## lang.num         3.303e-02 1.268e-05 2605.6 <2e-16 ***
## isgame           6.433e-02 4.194e-04 153.4 <2e-16 ***
## over_med         5.365e+00 3.376e-03 1589.2 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 163175599 on 2878 degrees of freedom
## Residual deviance: 96237216 on 2867 degrees of freedom
## AIC: 96257617
##
## Number of Fisher Scoring iterations: 7
##
## TARGET_FLAG
## Min. : 0.00
## 1st Qu.: 49.22
## Median : 203.49
## Mean : 11913.71
## 3rd Qu.: 22609.33
## Max. : 286323.69

```

Running the Poisson model gave the output that every field was statistically significant, so they were all kept for the predictions. The results from this model resembles the train's results.

Zero Inflated

Looking at the data set, we can see that there a large number of zeroes in a lot of fields. A zero inflated count regression should help deal with those.

```

## TARGET_FLAG
## Min. : 0.00
## 1st Qu.: 43.49
## Median : 202.02
## Mean : 11905.23
## 3rd Qu.: 22609.68
## Max. : 286333.79

```

The zero inflated model also resembles the results of the train's results but underestimates the maximum value. Looking at the distribution of the total ratings count in the overall dataset, we can see that there are only six apps with over one million total ratings, so it can be concluded that the zero inflated model's max is acceptable.

The zero inflated model gave the best results in determining a mobile application's success.

Google Play Store

This project also investigated what makes an app successful on the Google Play Store. This data was also taken from Kaggle and has 10,839 apps with 14 fields. This data contains similar information as the Apple

store data but it also includes how many time the app has been downloaded.

```
## [1] ART_AND_DESIGN      AUTO_AND_VEHICLES  BEAUTY
## [4] BOOKS_AND_REFERENCE BUSINESS             COMICS
## [7] COMMUNICATION        DATING             EDUCATION
## [10] ENTERTAINMENT        EVENTS             FINANCE
## [13] FOOD_AND_DRINK       HEALTH_AND_FITNESS HOUSE_AND_HOME
## [16] LIBRARIES_AND_DEMO   LIFESTYLE          GAME
## [19] FAMILY               MEDICAL            SOCIAL
## [22] SHOPPING             PHOTOGRAPHY        SPORTS
## [25] TRAVEL_AND_LOCAL     TOOLS              PERSONALIZATION
## [28] PRODUCTIVITY         PARENTING          WEATHER
## [31] VIDEO_PLAYERS        NEWS_AND_MAGAZINES MAPS_AND_NAVIGATION
## 33 Levels: ART_AND_DESIGN AUTO_AND_VEHICLES BEAUTY ... WEATHER

##                               App
## ROBLOX                        :    9
## CBS Sports App - Scores, News, Stats & Watch Live:    8
## 8 Ball Pool                   :    7
## Candy Crush Saga              :    7
## Duolingo: Learn Languages Free :    7
## ESPN                         :    7
## (Other)                      :10794

##      Category      Rating      Reviews
## FAMILY      :1971  Min.    :1.000  Min.    :    0
## GAME        :1144  1st Qu.:4.000  1st Qu.:   38
## TOOLS       : 843  Median :4.300  Median :  2094
## MEDICAL     : 463  Mean   :4.192  Mean   : 444194
## BUSINESS    : 460  3rd Qu.:4.500  3rd Qu.: 54783
## PRODUCTIVITY: 424  Max.    :5.000  Max.    :78158306
## (Other)     :5534  NA's    :1473

##      Size      size_adjusted      Installs
## Varies with device:1694  Min.    :    8.5  Min.    :0.000e+00
## 11M              : 198  1st Qu.: 5900.0  1st Qu.:3.000e+03
## 12M              : 196  Median :18000.0  Median :1.000e+05
## 14M              : 194  Mean   :20991.0  Mean   :1.547e+07
## 13M              : 191  3rd Qu.:26000.0  3rd Qu.:5.000e+06
## 15M              : 184  Max.    :100000.0  Max.    :1.000e+09
## (Other)         :8182

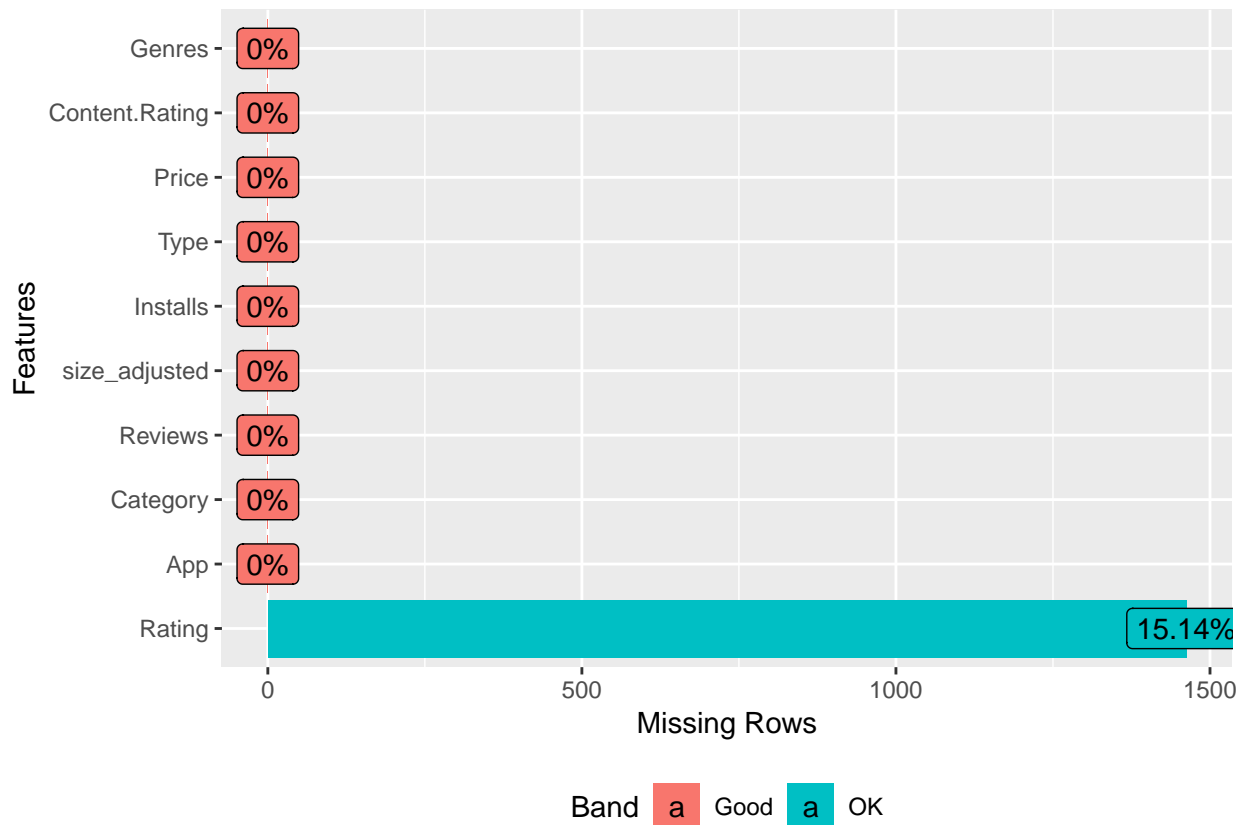
##      Type      Price      Content.Rating      Genres
## Free:10039  0      :10039  Adults only 18+: 3  Tools      : 842
## Paid: 800   $0.99 : 148  Everyone      :8714  Entertainment: 623
##           $2.99 : 129  Everyone 10+ : 413  Education    : 549
##           $1.99 : 73  Mature 17+  : 499  Medical      : 463
##           $4.99 : 72  Teen        :1208  Business     : 460
##           $3.99 : 63  Unrated     : 2   Productivity : 424
##           (Other): 315  (Other)     :7478

##      Last.Updated      Current.Ver      Android.Ver
## 8/3/2018 : 326  Varies with device:1458  4.1 and up      :2451
## 8/2/2018 : 304  1      : 842  4.0.3 and up    :1501
## 7/31/2018: 294  1.1    : 276  4.0 and up      :1375
## 8/1/2018 : 285  1.2    : 185  Varies with device:1361
## 7/30/2018: 211  2      : 165  4.4 and up      : 980
## 7/25/2018: 164  1.3    : 145  2.3 and up      : 652
## (Other)   :9255  (Other) :7768  (Other)         :2519
```

Data Management

There's a lot of missing Rating data and the \$ sign has to be removed from the Price field.

```
##                               App      Category Rating
## 1   Photo Editor & Candy Camera & Grid & ScrapBook ART_AND_DESIGN  4.1
## 2                               Coloring book moana ART_AND_DESIGN  3.9
## 3 U Launcher Lite - FREE Live Cool Themes, Hide Apps ART_AND_DESIGN  4.7
## 4                               Sketch - Draw & Paint ART_AND_DESIGN  4.5
## 5               Pixel Draw - Number Art Coloring Book ART_AND_DESIGN  4.3
## 6               Paper flowers instructions ART_AND_DESIGN  4.4
##  Reviews size_adjusted Installs Type Price Content.Rating
## 1     159           19000   10000 Free      0      Everyone
## 2     967           14000  500000 Free      0      Everyone
## 3   87510            8700 5000000 Free      0      Everyone
## 4  215644           25000 50000000 Free      0          Teen
## 5     967            2800  100000 Free      0      Everyone
## 6     167            5600   50000 Free      0      Everyone
##                               Genres
## 1               Art & Design
## 2 Art & Design;Pretend Play
## 3               Art & Design
## 4               Art & Design
## 5 Art & Design;Creativity
## 6               Art & Design
```



Target Flag

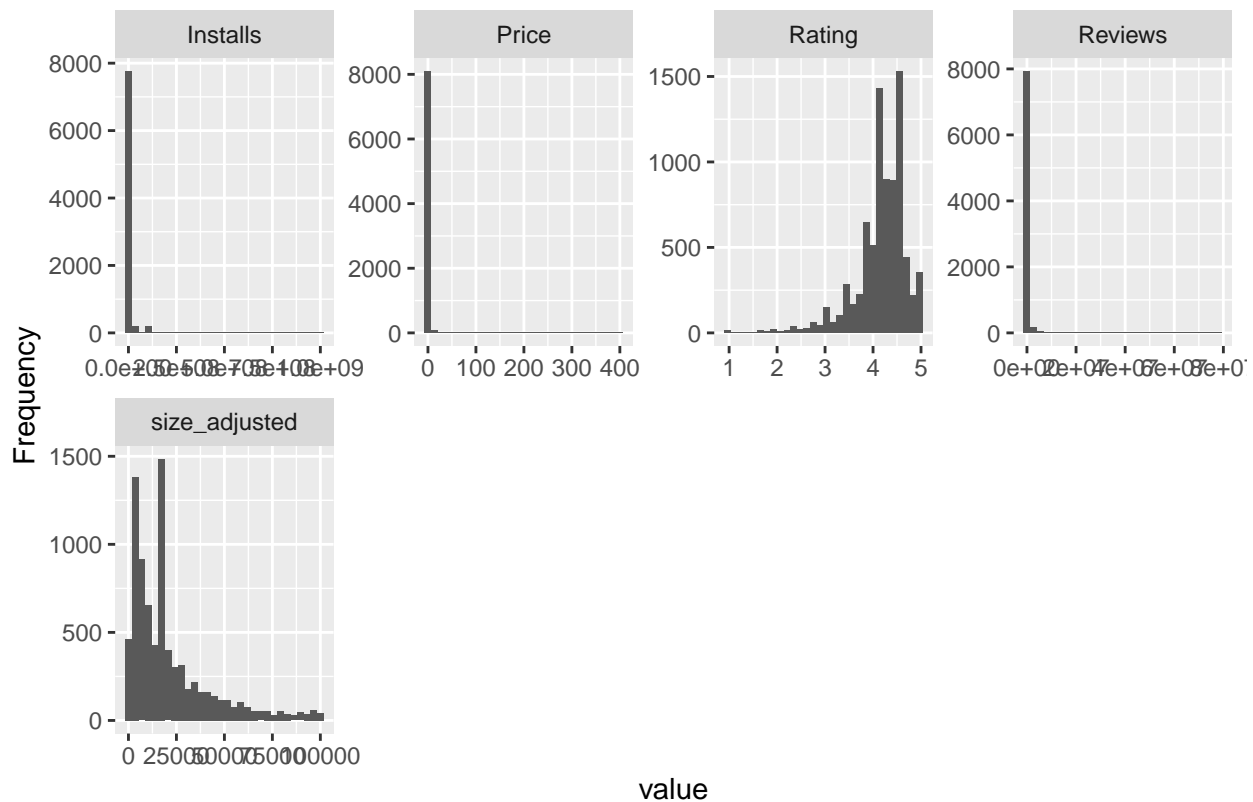
For the analysis of how succesfull an app is on the Google Play store a target falg was put in place to see if an app was downloaded more times than the average of all the apps.

```
##          vars      n      mean      sd      median      trimmed
## App*      1 8196    4940.51    2798.56    5062.5    4964.47
## Category*  2 8196     17.85      8.32     15.0     17.89
## Rating     3 8196      4.17      0.54      4.3      4.24
## Reviews    4 8196   255251.47  1985593.85  3004.0    28416.47
## size_adjusted 5 8196   21240.86   21080.86  18000.0    17348.68
## Installs   6 8196  9165089.72 58250865.45 100000.0 1580781.03
## Type*      7 8196      1.07      0.26      1.0      1.00
## Price      8 8196      1.04     16.86      0.0      0.00
## Content.Rating* 9 8196      2.46      1.00      2.0      2.20
## Genres*    10 8196     65.81     33.04     68.0     67.46
## above_avg  11 8196      0.63      0.48      1.0      0.66
##          mad min      max      range skew kurtosis
## App*      3589.37 2.0      9658     9656.0 -0.07    -1.20
## Category*   8.90 1.0        33      32.0  0.14    -1.09
## Rating      0.44 1.0         5       4.0 -1.74     5.11
## Reviews    4443.35 1.0   78158306  78158305.0 24.50    771.07
## size_adjusted 17049.90 8.5   100000    99991.5  1.70     2.71
## Installs   148111.74 1.0 1000000000 999999999.0 13.85    216.18
## Type*       0.00 1.0         2       1.0  3.26     8.65
## Price       0.00 0.0        400     400.0 22.99    534.13
## Content.Rating* 0.00 1.0         6       5.0  1.90     1.90
## Genres*     43.00 1.0        119     118.0 -0.29    -0.96
## above_avg   0.00 0.0         1       1.0 -0.53    -1.72
##          se
## App*      30.91
## Category*   0.09
## Rating      0.01
## Reviews    21932.57
## size_adjusted 232.86
## Installs   643430.15
## Type*       0.00
## Price      0.19
## Content.Rating* 0.01
## Genres*     0.36
## above_avg   0.01
```

Histogram

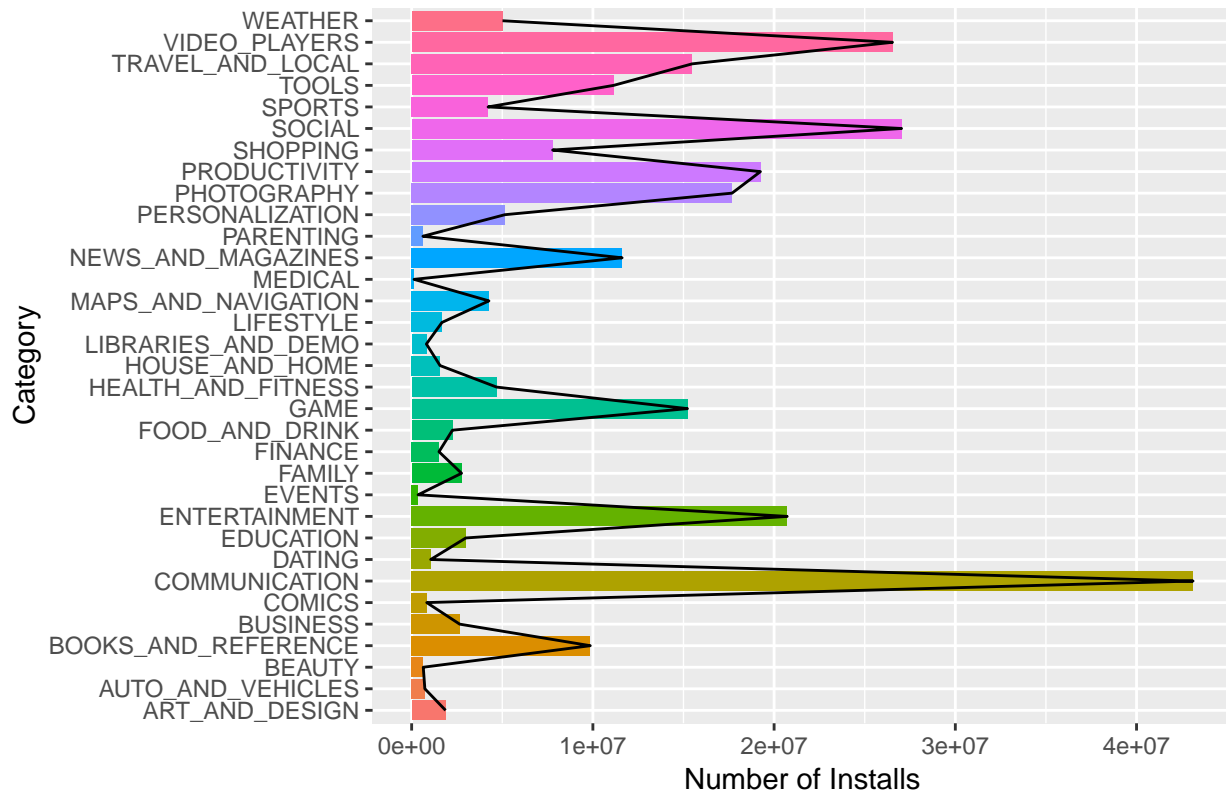
```
##          App      Category Rating
## 1  Messenger - Text and Video Chat for Free  COMMUNICATION  4.0
## 2                WhatsApp Messenger  COMMUNICATION  4.4
## 3              Subway Surfers      GAME  4.5
## 4            Candy Crush Saga      GAME  4.4
## 5              Clash Royale      GAME  4.6
## 6            Clash of Clans      GAME  4.6
## 7                Facebook    SOCIAL  4.1
## 8              Instagram    SOCIAL  4.5
## 9              YouTube VIDEO_PLAYERS  4.3
```

## 10	Clean Master- Space Cleaner & Antivirus	TOOLS	4.7
## 11	Security Master - Antivirus, VPN, AppLock, Booster	TOOLS	4.7
##	Reviews	size_adjusted	Installs
## 1	56642847	18153.77	1000000000
## 2	69119316	18153.77	1000000000
## 3	27722264	76000.00	1000000000
## 4	22426677	74000.00	500000000
## 5	23133508	97000.00	100000000
## 6	44891723	98000.00	100000000
## 7	78158306	18153.77	1000000000
## 8	66577313	18153.77	1000000000
## 9	25655305	18153.77	1000000000
## 10	42916526	18153.77	500000000
## 11	24900999	18153.77	500000000
##	Genres	above_avg	
## 1	Communication	0	
## 2	Communication	1	
## 3	Arcade	1	
## 4	Casual	1	
## 5	Strategy	1	
## 6	Strategy	1	
## 7	Social	0	
## 8	Social	1	
## 9	Video Players & Editors	1	
## 10	Tools	1	
## 11	Tools	1	

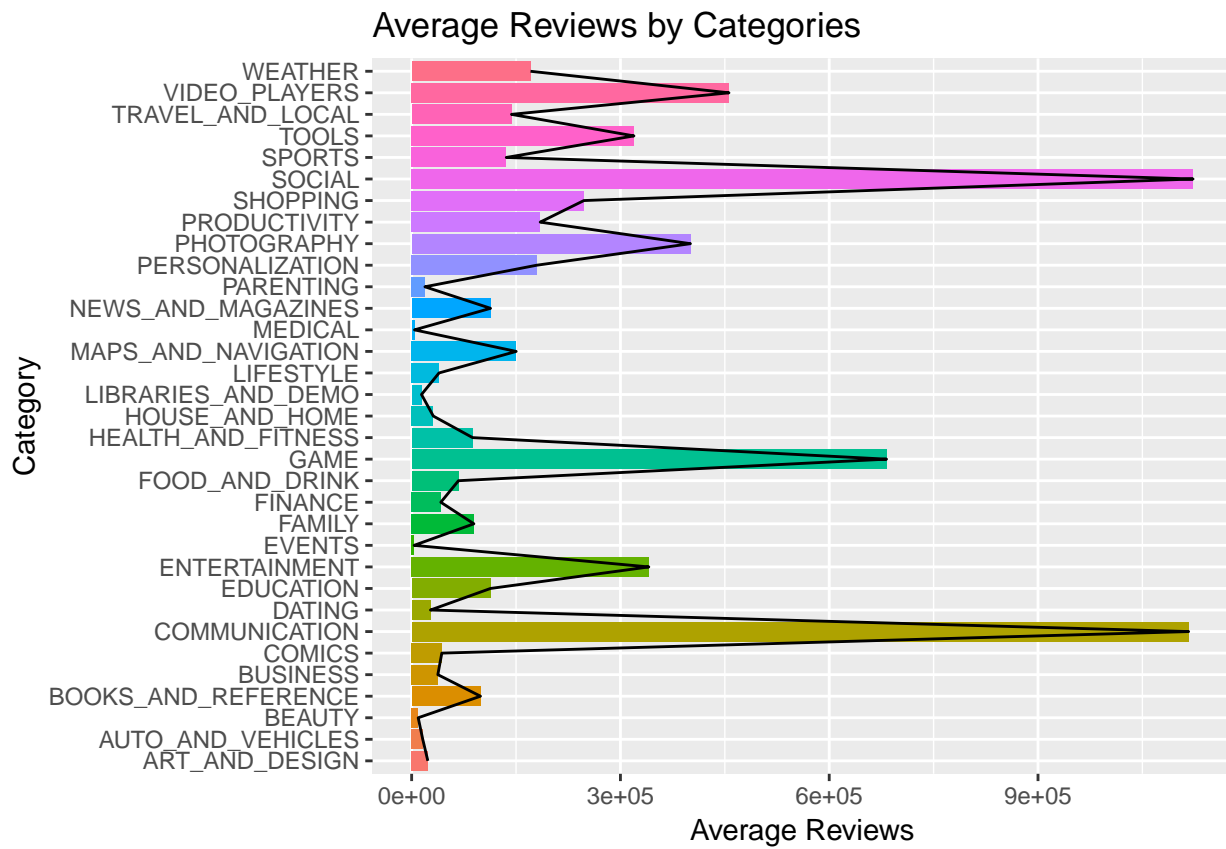


##Downloads by Categories

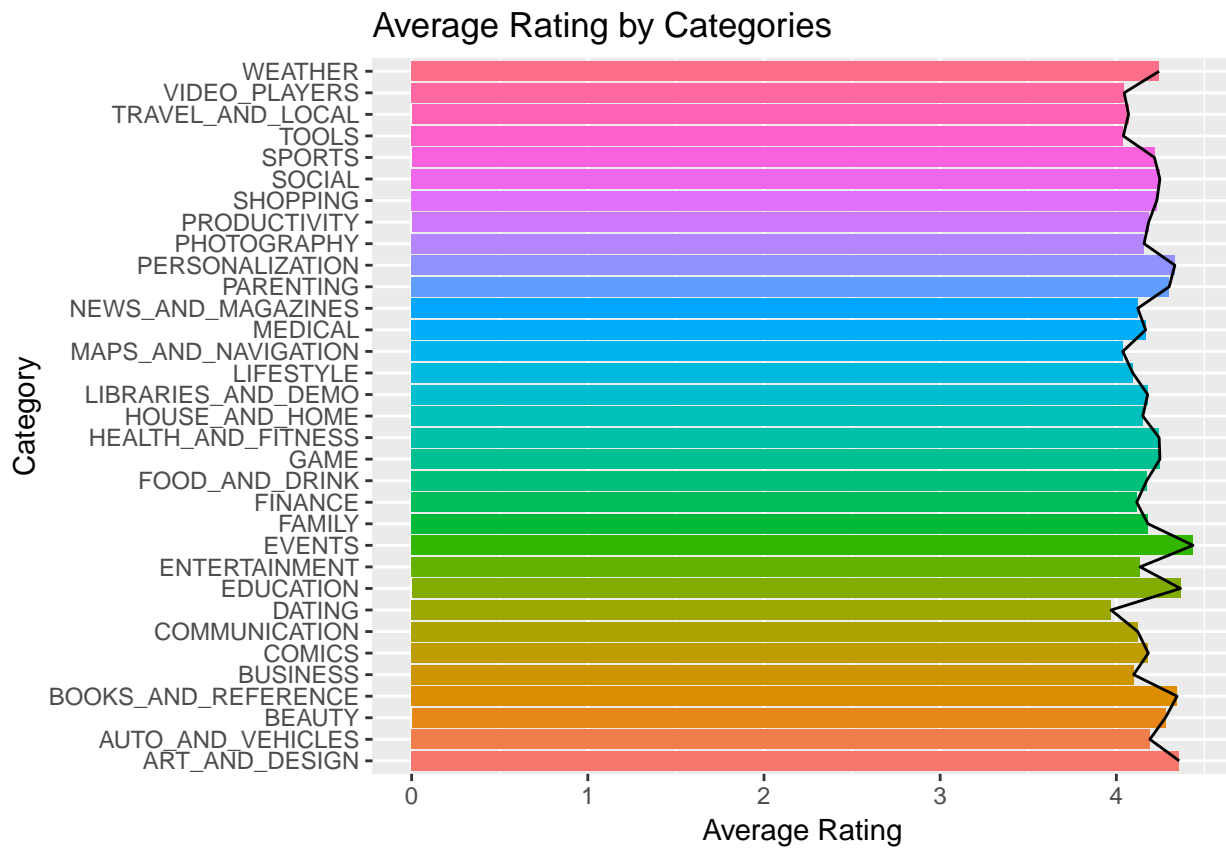
Average Number of Downloads by Categories



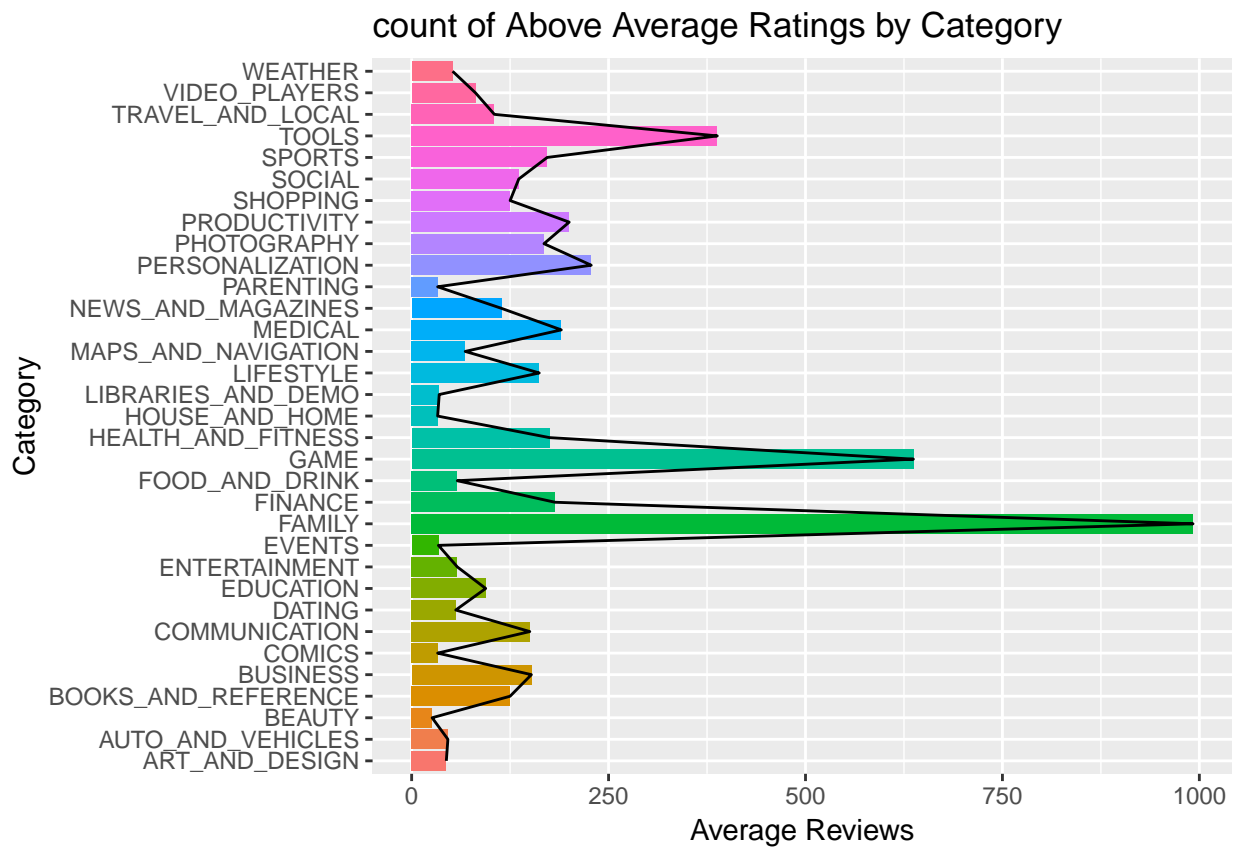
Reviews by Categories



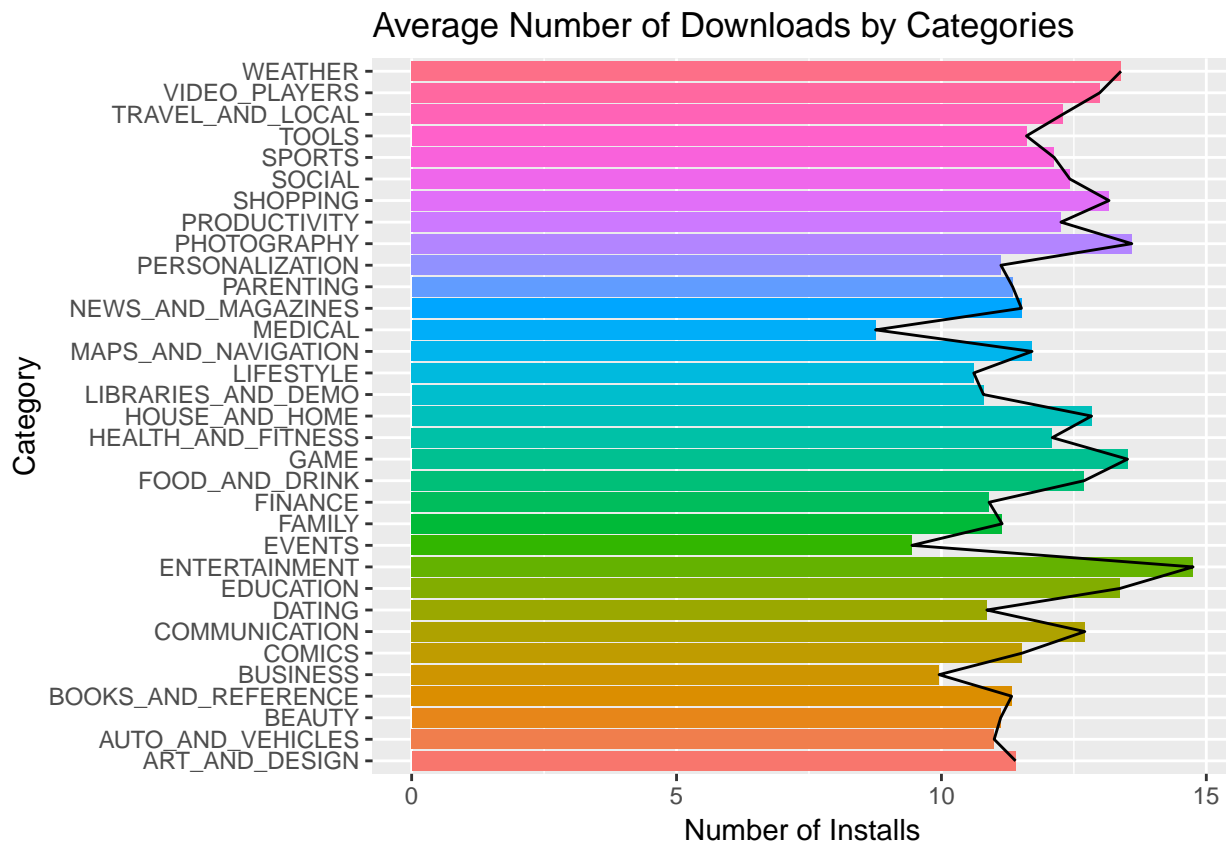
Ratings by Categories



Number Above Average by Categories



Log View of Downloads by Categories



Train-Test Split

Just like the Apple data, the Google data did not come in a training and testing datasets. So this was done by splitting up the total dataset.

Poisson

The first model ran was Poisson.

```
##
## Call:
## glm(formula = above_avg ~ install_log + review_log + Type + Category +
##      Price + Content.Rating, family = "poisson", data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5287  -0.9704   0.1794   0.4534   1.3642
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.526822   0.736588   0.715   0.4745
## install_log   -0.201796   0.015176 -13.297 <2e-16 ***
## review_log     0.238084   0.015220  15.643 <2e-16 ***
```

```

## TypePaid -0.043230 0.063892 -0.677 0.4987
## CategoryAUTO_AND_VEHICLES -0.188760 0.225266 -0.838 0.4021
## CategoryBEAUTY -0.081959 0.264353 -0.310 0.7565
## CategoryBOOKS_AND_REFERENCE -0.109440 0.187340 -0.584 0.5591
## CategoryBUSINESS -0.307656 0.182282 -1.688 0.0914 .
## CategoryCOMICS -0.317223 0.263483 -1.204 0.2286
## CategoryCOMMUNICATION -0.382773 0.181457 -2.109 0.0349 *
## CategoryDATING -0.537897 0.231774 -2.321 0.0203 *
## CategoryEDUCATION -0.094283 0.197590 -0.477 0.6332
## CategoryENTERTAINMENT -0.385009 0.216125 -1.781 0.0748 .
## CategoryEVENTS -0.004081 0.254367 -0.016 0.9872
## CategoryFAMILY -0.268331 0.160640 -1.670 0.0948 .
## CategoryFINANCE -0.289409 0.177648 -1.629 0.1033
## CategoryFOOD_AND_DRINK -0.357987 0.219755 -1.629 0.1033
## CategoryGAME -0.303702 0.164400 -1.847 0.0647 .
## CategoryHEALTH_AND_FITNESS -0.154521 0.177088 -0.873 0.3829
## CategoryHOUSE_AND_HOME -0.321196 0.247950 -1.295 0.1952
## CategoryLIBRARIES_AND_DEMO -0.263189 0.242740 -1.084 0.2783
## CategoryLIFESTYLE -0.342862 0.180578 -1.899 0.0576 .
## CategoryMAPS_AND_NAVIGATION -0.337532 0.208104 -1.622 0.1048
## CategoryMEDICAL -0.119987 0.176356 -0.680 0.4963
## CategoryNEWS_AND_MAGAZINES -0.376348 0.189447 -1.987 0.0470 *
## CategoryPARENTING -0.064951 0.253775 -0.256 0.7980
## CategoryPERSONALIZATION -0.080809 0.172810 -0.468 0.6401
## CategoryPHOTOGRAPHY -0.278518 0.178493 -1.560 0.1187
## CategoryPRODUCTIVITY -0.198151 0.174806 -1.134 0.2570
## CategorySHOPPING -0.190029 0.184886 -1.028 0.3040
## CategorySOCIAL -0.252979 0.185921 -1.361 0.1736
## CategorySPORTS -0.279221 0.177801 -1.570 0.1163
## CategoryTOOLS -0.350636 0.165839 -2.114 0.0345 *
## CategoryTRAVEL_AND_LOCAL -0.317909 0.191332 -1.662 0.0966 .
## CategoryVIDEO_PLAYERS -0.397690 0.201496 -1.974 0.0484 *
## CategoryWEATHER -0.236281 0.227204 -1.040 0.2984
## Price -0.003187 0.001777 -1.794 0.0729 .
## Content.RatingEveryone -0.230735 0.716044 -0.322 0.7473
## Content.RatingEveryone 10+ -0.331805 0.720595 -0.460 0.6452
## Content.RatingMature 17+ -0.335825 0.721327 -0.466 0.6415
## Content.RatingTeen -0.329695 0.717058 -0.460 0.6457
## Content.RatingUnrated -11.980819 284.660069 -0.042 0.9664
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 3848.0 on 6555 degrees of freedom
## Residual deviance: 3507.4 on 6514 degrees of freedom
## AIC: 11793
##
## Number of Fisher Scoring iterations: 10

```

Poisson Reduced

We removed some fields from the Poisson model that were not significant.


```
##
## Call:
## glm(formula = above_avg ~ install_log + review_log + Category +
##       Price, family = "poisson", data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5180  -0.9722   0.1775   0.4555   1.4117
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.266695   0.171236   1.557  0.11936
## install_log     -0.196522   0.014117 -13.920 < 2e-16 ***
## review_log       0.232379   0.014490  16.037 < 2e-16 ***
## CategoryAUTO_AND_VEHICLES -0.179794   0.225195  -0.798  0.42464
## CategoryBEAUTY      -0.078985   0.264303  -0.299  0.76506
## CategoryBOOKS_AND_REFERENCE -0.115132   0.187260  -0.615  0.53867
## CategoryBUSINESS    -0.299099   0.182154  -1.642  0.10059
## CategoryCOMICS      -0.352785   0.260585  -1.354  0.17579
## CategoryCOMMUNICATION -0.383493   0.181418  -2.114  0.03453 *
## CategoryDATING      -0.625897   0.213800  -2.927  0.00342 **
## CategoryEDUCATION   -0.084923   0.197505  -0.430  0.66721
## CategoryENTERTAINMENT -0.434700   0.214407  -2.027  0.04262 *
## CategoryEVENTS      -0.013152   0.254059  -0.052  0.95871
## CategoryFAMILY      -0.283297   0.160468  -1.765  0.07749 .
## CategoryFINANCE     -0.277496   0.177487  -1.563  0.11794
## CategoryFOOD_AND_DRINK -0.355768   0.219736  -1.619  0.10543
## CategoryGAME        -0.341006   0.163475  -2.086  0.03698 *
## CategoryHEALTH_AND_FITNESS -0.150412   0.177032  -0.850  0.39553
## CategoryHOUSE_AND_HOME -0.313647   0.247909  -1.265  0.20581
## CategoryLIBRARIES_AND_DEMO -0.253356   0.242658  -1.044  0.29645
## CategoryLIFESTYLE    -0.344016   0.180498  -1.906  0.05666 .
## CategoryMAPS_AND_NAVIGATION -0.330195   0.208058  -1.587  0.11251
## CategoryMEDICAL     -0.120258   0.176334  -0.682  0.49525
## CategoryNEWS_AND_MAGAZINES -0.402546   0.188513  -2.135  0.03273 *
## CategoryPARENTING    -0.062416   0.253753  -0.246  0.80571
## CategoryPERSONALIZATION -0.087079   0.172641  -0.504  0.61399
## CategoryPHOTOGRAPHY  -0.272968   0.178449  -1.530  0.12610
## CategoryPRODUCTIVITY -0.188451   0.174746  -1.078  0.28084
## CategorySHOPPING     -0.196019   0.184755  -1.061  0.28870
## CategorySOCIAL       -0.306576   0.183050  -1.675  0.09397 .
## CategorySPORTS       -0.273830   0.177705  -1.541  0.12334
## CategoryTOOLS        -0.345782   0.165799  -2.086  0.03702 *
## CategoryTRAVEL_AND_LOCAL -0.313885   0.191308  -1.641  0.10085
## CategoryVIDEO_PLAYERS -0.401416   0.201454  -1.993  0.04631 *
## CategoryWEATHER      -0.231947   0.227160  -1.021  0.30722
## Price              -0.003415   0.001796  -1.901  0.05728 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 3848.0  on 6555  degrees of freedom
## Residual deviance: 3513.7  on 6520  degrees of freedom
```

```
## AIC: 11788
##
## Number of Fisher Scoring iterations: 5
```

Quasi

Next a Quasi model was used.

```
##
## Call:
## glm(formula = above_avg ~ install_log + review_log + Type + Category +
##       Price + Content.Rating, family = "quasi", data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.0407  -0.4740   0.1722   0.3630   1.1298
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.3290074   0.3278915    4.053 5.11e-05 ***
## install_log      -0.1244676   0.0053764  -23.151 < 2e-16 ***
## review_log       0.1473643   0.0053727   27.429 < 2e-16 ***
## TypePaid        -0.0266636   0.0241019   -1.106 0.268644
## CategoryAUTO_AND_VEHICLES -0.1198587   0.0827697   -1.448 0.147638
## CategoryBEAUTY     -0.0577980   0.0968598   -0.597 0.550716
## CategoryBOOKS_AND_REFERENCE -0.0644103   0.0713522   -0.903 0.366713
## CategoryBUSINESS  -0.1921510   0.0675259   -2.846 0.004447 **
## CategoryCOMICS     -0.2041415   0.0930805   -2.193 0.028330 *
## CategoryCOMMUNICATION -0.2422059   0.0675999   -3.583 0.000342 ***
## CategoryDATING     -0.3006443   0.0800058   -3.758 0.000173 ***
## CategoryEDUCATION  -0.0464675   0.0766299   -0.606 0.544278
## CategoryENTERTAINMENT -0.2521986   0.0788436   -3.199 0.001387 **
## CategoryEVENTS     -0.0061205   0.0979021   -0.063 0.950153
## CategoryFAMILY     -0.1711697   0.0611704   -2.798 0.005153 **
## CategoryFINANCE    -0.1878347   0.0667537   -2.814 0.004910 **
## CategoryFOOD_AND_DRINK -0.2291565   0.0797794   -2.872 0.004087 **
## CategoryGAME       -0.1914263   0.0626816   -3.054 0.002268 **
## CategoryHEALTH_AND_FITNESS -0.0946400   0.0678565   -1.395 0.163152
## CategoryHOUSE_AND_HOME -0.2062797   0.0878837   -2.347 0.018945 *
## CategoryLIBRARIES_AND_DEMO -0.1733311   0.0860830   -2.014 0.044099 *
## CategoryLIFESTYLE   -0.2128633   0.0667918   -3.187 0.001445 **
## CategoryMAPS_AND_NAVIGATION -0.2180708   0.0759020   -2.873 0.004078 **
## CategoryMEDICAL    -0.0773091   0.0668120   -1.157 0.247268
## CategoryNEWS_AND_MAGAZINES -0.2416977   0.0694793   -3.479 0.000507 ***
## CategoryPARENTING  -0.0400977   0.0952118   -0.421 0.673665
## CategoryPERSONALIZATION -0.0440051   0.0665140   -0.662 0.508257
## CategoryPHOTOGRAPHY -0.1804969   0.0673954   -2.678 0.007421 **
## CategoryPRODUCTIVITY -0.1246795   0.0665660   -1.873 0.061110 .
## CategorySHOPPING   -0.1232857   0.0703921   -1.751 0.079922 .
## CategorySOCIAL     -0.1585884   0.0701604   -2.260 0.023831 *
## CategorySPORTS     -0.1818067   0.0674896   -2.694 0.007081 **
## CategoryTOOLS      -0.2173114   0.0626083   -3.471 0.000522 ***
## CategoryTRAVEL_AND_LOCAL -0.2040144   0.0702513   -2.904 0.003696 **
## CategoryVIDEO_PLAYERS -0.2430203   0.0730053   -3.329 0.000877 ***
```

```
## CategoryWEATHER          -0.1498150  0.0862034 -1.738 0.082271 .
## Price                    -0.0012202  0.0003499 -3.488 0.000491 ***
## Content.RatingEveryone   -0.2064333  0.3213268 -0.642 0.520610
## Content.RatingEveryone 10+ -0.2710388  0.3226508 -0.840 0.400918
## Content.RatingMature 17+ -0.2770713  0.3227721 -0.858 0.390697
## Content.RatingTeen       -0.2692742  0.3216229 -0.837 0.402491
## Content.RatingUnrated    -0.8082005  0.5540253 -1.459 0.144674
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasi family taken to be 0.2033567)
##
## Null deviance: 1535.7 on 6555 degrees of freedom
## Residual deviance: 1324.7 on 6514 degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 2
```

Quasi Reduced

Quasi with removed fields

```
##
## Call:
## glm(formula = above_avg ~ install_log + review_log + Category +
## Price, family = "quasi", data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.0313  -0.4768   0.1716   0.3639   1.1213
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.1049726  0.0652814  16.926 < 2e-16 ***
## install_log    -0.1213513  0.0050251 -24.149 < 2e-16 ***
## review_log      0.1439350  0.0051270  28.074 < 2e-16 ***
## CategoryAUTO_AND_VEHICLES -0.1151980  0.0828304 -1.391 0.164343
## CategoryBEAUTY    -0.0565079  0.0969364 -0.583 0.559955
## CategoryBOOKS_AND_REFERENCE -0.0681120  0.0714019 -0.954 0.340158
## CategoryBUSINESS  -0.1873830  0.0675576 -2.774 0.005558 **
## CategoryCOMICS    -0.2264302  0.0924609 -2.449 0.014354 *
## CategoryCOMMUNICATION -0.2430465  0.0676505 -3.593 0.000330 ***
## CategoryDATING    -0.3605153  0.0736022 -4.898 9.91e-07 ***
## CategoryEDUCATION -0.0419855  0.0766772 -0.548 0.584012
## CategoryENTERTAINMENT -0.2836527  0.0783184 -3.622 0.000295 ***
## CategoryEVENTS    -0.0120991  0.0979108 -0.124 0.901657
## CategoryFAMILY    -0.1800392  0.0611793 -2.943 0.003264 **
## CategoryFINANCE   -0.1812574  0.0667764 -2.714 0.006657 **
## CategoryFOOD_AND_DRINK -0.2275939  0.0798502 -2.850 0.004382 **
## CategoryGAME      -0.2158595  0.0624129 -3.459 0.000547 ***
## CategoryHEALTH_AND_FITNESS -0.0932919  0.0679096 -1.374 0.169562
## CategoryHOUSE_AND_HOME -0.2024667  0.0879557 -2.302 0.021371 *
## CategoryLIBRARIES_AND_DEMO -0.1674376  0.0861381 -1.944 0.051959 .
## CategoryLIFESTYLE  -0.2139836  0.0668353 -3.202 0.001373 **
```

```
## CategoryMAPS_AND_NAVIGATION -0.2139951 0.0759599 -2.817 0.004859 **
## CategoryMEDICAL -0.0788151 0.0668464 -1.179 0.238422
## CategoryNEWS_AND_MAGAZINES -0.2573843 0.0692395 -3.717 0.000203 ***
## CategoryPARENTING -0.0394365 0.0952929 -0.414 0.679001
## CategoryPERSONALIZATION -0.0488123 0.0665071 -0.734 0.463011
## CategoryPHOTOGRAPHY -0.1779518 0.0674468 -2.638 0.008350 **
## CategoryPRODUCTIVITY -0.1189880 0.0666098 -1.786 0.074090 .
## CategorySHOPPING -0.1276319 0.0704160 -1.813 0.069948 .
## CategorySOCIAL -0.1937799 0.0692961 -2.796 0.005183 **
## CategorySPORTS -0.1788403 0.0675322 -2.648 0.008111 **
## CategoryTOOLS -0.2144281 0.0626456 -3.423 0.000623 ***
## CategoryTRAVEL_AND_LOCAL -0.2010319 0.0703095 -2.859 0.004260 **
## CategoryVIDEO_PLAYERS -0.2459982 0.0730576 -3.367 0.000764 ***
## CategoryWEATHER -0.1476231 0.0862686 -1.711 0.087091 .
## Price -0.0012856 0.0003426 -3.752 0.000177 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasi family taken to be 0.2037328)
##
## Null deviance: 1535.7 on 6555 degrees of freedom
## Residual deviance: 1328.3 on 6520 degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 2
```

Bionmial

```
##
## Call:
## glm(formula = above_avg ~ install_log + review_log + Type + Category +
## Price + Content.Rating, family = "binomial", data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3296  -1.1146   0.6071   0.9153   2.5195
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    14.237464  229.205535   0.062 0.950470
## install_log     -0.617035   0.029084 -21.216 < 2e-16 ***
## review_log       0.728709   0.029665  24.564 < 2e-16 ***
## TypePaid        -0.141471   0.125563  -1.127 0.259872
## CategoryAUTO_AND_VEHICLES -0.591843  0.419383  -1.411 0.158179
## CategoryBEAUTY    -0.295877  0.483547  -0.612 0.540612
## CategoryBOOKS_AND_REFERENCE -0.301180  0.372427  -0.809 0.418691
## CategoryBUSINESS  -0.951116  0.346328  -2.746 0.006027 **
## CategoryCOMICS    -1.017709  0.458315  -2.221 0.026382 *
## CategoryCOMMUNICATION -1.212047  0.347637  -3.487 0.000489 ***
## CategoryDATING    -1.429236  0.403055  -3.546 0.000391 ***
## CategoryEDUCATION -0.150750  0.415391  -0.363 0.716672
## CategoryENTERTAINMENT -1.274439  0.395827  -3.220 0.001283 **
## CategoryEVENTS    -0.068677  0.510372  -0.135 0.892958
```

```

## CategoryFAMILY -0.854903 0.318495 -2.684 0.007271 **
## CategoryFINANCE -0.953736 0.343959 -2.773 0.005557 **
## CategoryFOOD_AND_DRINK -1.149055 0.401958 -2.859 0.004255 **
## CategoryGAME -0.954747 0.326692 -2.922 0.003473 **
## CategoryHEALTH_AND_FITNESS -0.463037 0.355836 -1.301 0.193167
## CategoryHOUSE_AND_HOME -1.026793 0.435961 -2.355 0.018511 *
## CategoryLIBRARIES_AND_DEMO -0.876532 0.424403 -2.065 0.038892 *
## CategoryLIFESTYLE -1.053662 0.342645 -3.075 0.002104 **
## CategoryMAPS_AND_NAVIGATION -1.095832 0.382836 -2.862 0.004204 **
## CategoryMEDICAL -0.385272 0.345604 -1.115 0.264945
## CategoryNEWS_AND_MAGAZINES -1.203165 0.353192 -3.407 0.000658 ***
## CategoryPARENTING -0.183855 0.485874 -0.378 0.705133
## CategoryPERSONALIZATION -0.190532 0.351204 -0.543 0.587466
## CategoryPHOTOGRAPHY -0.914695 0.348259 -2.626 0.008627 **
## CategoryPRODUCTIVITY -0.625541 0.346518 -1.805 0.071040 .
## CategorySHOPPING -0.630691 0.365246 -1.727 0.084212 .
## CategorySOCIAL -0.793406 0.362522 -2.189 0.028628 *
## CategorySPORTS -0.936514 0.349368 -2.681 0.007349 **
## CategoryTOOLS -1.080079 0.324906 -3.324 0.000886 ***
## CategoryTRAVEL_AND_LOCAL -1.021204 0.357272 -2.858 0.004259 **
## CategoryVIDEO_PLAYERS -1.193101 0.371687 -3.210 0.001328 **
## CategoryWEATHER -0.742432 0.444687 -1.670 0.095006 .
## Price -0.005925 0.002045 -2.897 0.003769 **
## Content.RatingEveryone -11.153391 229.205284 -0.049 0.961189
## Content.RatingEveryone 10+ -11.484703 229.205330 -0.050 0.960037
## Content.RatingMature 17+ -11.515534 229.205331 -0.050 0.959930
## Content.RatingTeen -11.470996 229.205294 -0.050 0.960085
## Content.RatingUnrated -24.205913 397.484010 -0.061 0.951441
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 8670.8 on 6555 degrees of freedom
## Residual deviance: 7703.1 on 6514 degrees of freedom
## AIC: 7787.1
##
## Number of Fisher Scoring iterations: 11

```

Binomial backward elimination

```

##
## Call:
## glm(formula = above_avg ~ install_log + review_log + Category +
## Price, family = "binomial", data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3070  -1.1228   0.6083   0.9182   2.4989
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.986227    0.342592   8.717 < 2e-16 ***

```

```

## install_log          -0.599341    0.027256 -21.989 < 2e-16 ***
## review_log           0.709068    0.028270  25.082 < 2e-16 ***
## CategoryAUTO_AND_VEHICLES -0.570105    0.418511  -1.362 0.173128
## CategoryBEAUTY        -0.290754    0.482808  -0.602 0.547032
## CategoryBOOKS_AND_REFERENCE -0.318439    0.372285  -0.855 0.392349
## CategoryBUSINESS      -0.924490    0.345894  -2.673 0.007523 **
## CategoryCOMICS        -1.127037    0.455143  -2.476 0.013278 *
## CategoryCOMMUNICATION -1.215207    0.347374  -3.498 0.000468 ***
## CategoryDATING        -1.732104    0.372242  -4.653 3.27e-06 ***
## CategoryEDUCATION     -0.135369    0.414087  -0.327 0.743736
## CategoryENTERTAINMENT -1.431668    0.392942  -3.643 0.000269 ***
## CategoryEVENTS        -0.097364    0.510130  -0.191 0.848634
## CategoryFAMILY        -0.894225    0.318212  -2.810 0.004952 **
## CategoryFINANCE       -0.919272    0.343476  -2.676 0.007442 **
## CategoryFOOD_AND_DRINK -1.137407    0.401601  -2.832 0.004623 **
## CategoryGAME          -1.077312    0.325058  -3.314 0.000919 ***
## CategoryHEALTH_AND_FITNESS -0.461563    0.355324  -1.299 0.193947
## CategoryHOUSE_AND_HOME -1.006601    0.434881  -2.315 0.020632 *
## CategoryLIBRARIES_AND_DEMO -0.844651    0.423845  -1.993 0.046280 *
## CategoryLIFESTYLE     -1.055644    0.342359  -3.083 0.002046 **
## CategoryMAPS_AND_NAVIGATION -1.073603    0.382395  -2.808 0.004992 **
## CategoryMEDICAL       -0.398203    0.345061  -1.154 0.248497
## CategoryNEWS_AND_MAGAZINES -1.277857    0.351935  -3.631 0.000282 ***
## CategoryPARENTING     -0.182916    0.484728  -0.377 0.705908
## CategoryPERSONALIZATION -0.219705    0.350396  -0.627 0.530647
## CategoryPHOTOGRAPHY   -0.902311    0.347807  -2.594 0.009479 **
## CategoryPRODUCTIVITY  -0.596033    0.346124  -1.722 0.085066 .
## CategorySHOPPING      -0.653339    0.364795  -1.791 0.073297 .
## CategorySOCIAL        -0.967700    0.358309  -2.701 0.006918 **
## CategorySPORTS        -0.921318    0.349054  -2.639 0.008304 **
## CategoryTOOLS         -1.062658    0.324585  -3.274 0.001061 **
## CategoryTRAVEL_AND_LOCAL -1.001418    0.357098  -2.804 0.005042 **
## CategoryVIDEO_PLAYERS -1.204920    0.371211  -3.246 0.001171 **
## CategoryWEATHER       -0.731738    0.443888  -1.648 0.099255 .
## Price                -0.006303    0.002045  -3.081 0.002061 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 8670.8  on 6555  degrees of freedom
## Residual deviance: 7722.5  on 6520  degrees of freedom
## AIC: 7794.5
##
## Number of Fisher Scoring iterations: 3

```

Google Results

Poisson Results

```

##                               2.5 %       97.5 %
## (Intercept)                 -0.068921041  0.6023115169

```

```

## install_log -0.224191354 -0.1688518022
## review_log 0.203979752 0.2607787189
## CategoryAUTO_AND_VEHICLES -0.621167352 0.2615797286
## CategoryBEAUTY -0.597010317 0.4390400035
## CategoryBOOKS_AND_REFERENCE -0.482155110 0.2518912133
## CategoryBUSINESS -0.656115308 0.0579164922
## CategoryCOMICS -0.863522395 0.1579521486
## CategoryCOMMUNICATION -0.739065453 -0.0279197913
## CategoryDATING -1.044936791 -0.2068566248
## CategoryEDUCATION -0.472026861 0.3021800302
## CategoryENTERTAINMENT -0.854929571 -0.0144704579
## CategoryEVENTS -0.511098635 0.4847947538
## CategoryFAMILY -0.597807688 0.0312140409
## CategoryFINANCE -0.625363592 0.0703710336
## CategoryFOOD_AND_DRINK -0.786442893 0.0749077021
## CategoryGAME -0.661412245 -0.0206001009
## CategoryHEALTH_AND_FITNESS -0.497388679 0.1965654687
## CategoryHOUSE_AND_HOME -0.799540157 0.1722459872
## CategoryLIBRARIES_AND_DEMO -0.728956815 0.2222449024
## CategoryLIFESTYLE -0.697785257 0.0097529603
## CategoryMAPS_AND_NAVIGATION -0.737981428 0.0775911943
## CategoryMEDICAL -0.465866241 0.2253510963
## CategoryNEWS_AND_MAGAZINES -0.772024825 -0.0330668878
## CategoryPARENTING -0.559762953 0.4349313984
## CategoryPERSONALIZATION -0.425449365 0.2512922006
## CategoryPHOTOGRAPHY -0.622720873 0.0767853957
## CategoryPRODUCTIVITY -0.530946572 0.1540452146
## CategorySHOPPING -0.558131347 0.1660935643
## CategorySOCIAL -0.665347499 0.0521950157
## CategorySPORTS -0.622126189 0.0744654680
## CategoryTOOLS -0.670742899 -0.0208217423
## CategoryTRAVEL_AND_LOCAL -0.688841538 0.0610710041
## CategoryVIDEO_PLAYERS -0.796258777 -0.0065738766
## CategoryWEATHER -0.677172242 0.2132772850
## Price -0.006934916 0.0001055615

## pred
## actual 0 1
## 0 1031 1424
## 1 743 3358

```

Quasi Results

```

## 2.5 % 97.5 %
## (Intercept) 0.977023452 1.2329217789
## install_log -0.131200410 -0.1115021876
## review_log 0.133886195 0.1539837632
## CategoryAUTO_AND_VEHICLES -0.277542639 0.0471465710
## CategoryBEAUTY -0.246499888 0.1334840112
## CategoryBOOKS_AND_REFERENCE -0.208057074 0.0718331497
## CategoryBUSINESS -0.319793418 -0.0549726108
## CategoryCOMICS -0.407650322 -0.0452100638
## CategoryCOMMUNICATION -0.375638984 -0.1104539845
## CategoryDATING -0.504773027 -0.2162576657

```

```

## CategoryEDUCATION          -0.192270035  0.1082990833
## CategoryENTERTAINMENT      -0.437153913 -0.1301515637
## CategoryEVENTS             -0.204000761  0.1798025196
## CategoryFAMILY             -0.299948358 -0.0601300286
## CategoryFINANCE            -0.312136826 -0.0503779613
## CategoryFOOD_AND_DRINK     -0.384097363 -0.0710904164
## CategoryGAME               -0.338186448 -0.0935325230
## CategoryHEALTH_AND_FITNESS -0.226392338  0.0398085787
## CategoryHOUSE_AND_HOME     -0.374856778 -0.0300766297
## CategoryLIBRARIES_AND_DEMO -0.336265243  0.0013900237
## CategoryLIFESTYLE          -0.344978371 -0.0829887402
## CategoryMAPS_AND_NAVIGATION -0.362873679 -0.0651165296
## CategoryMEDICAL            -0.209831742  0.0522014692
## CategoryNEWS_AND_MAGAZINES -0.393091249 -0.1216773968
## CategoryPARENTING          -0.226207161  0.1473341434
## CategoryPERSONALIZATION    -0.179163856  0.0815392554
## CategoryPHOTOGRAPHY        -0.310145073 -0.0457585560
## CategoryPRODUCTIVITY       -0.249540863  0.0115648704
## CategorySHOPPING           -0.265644656  0.0103808714
## CategorySOCIAL             -0.329597756 -0.0579620445
## CategorySPORTS             -0.311200980 -0.0464795639
## CategoryTOOLS              -0.337211224 -0.0916449346
## CategoryTRAVEL_AND_LOCAL    -0.338835964 -0.0632278801
## CategoryVIDEO_PLAYERS      -0.389188522 -0.1028078137
## CategoryWEATHER            -0.316706403  0.0214601275
## Price                      -0.001957228 -0.0006140706

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.4602  0.5005  0.6215  0.6255  0.7525  1.1898

##      pred
## actual    0    1
##      0  950 1505
##      1  680 3421

```

Binomial Results

```

##              2.5 %      97.5 %
## (Intercept)    2.3147584  3.657694968
## install_log   -0.6527619 -0.545921026
## review_log     0.6536599  0.764476810
## CategoryAUTO_AND_VEHICLES -1.3903714  0.250161876
## CategoryBEAUTY   -1.2370408  0.655533093
## CategoryBOOKS_AND_REFERENCE -1.0481036  0.411224990
## CategoryBUSINESS -1.6024296 -0.246550956
## CategoryCOMICS   -2.0191014 -0.234971715
## CategoryCOMMUNICATION -1.8960472 -0.534367403
## CategoryDATING   -2.4616848 -1.002522784
## CategoryEDUCATION -0.9469645  0.676226240
## CategoryENTERTAINMENT -2.2018206 -0.661516257
## CategoryEVENTS   -1.0972009  0.902472688
## CategoryFAMILY   -1.5179088 -0.270540527
## CategoryFINANCE  -1.5924716 -0.246071937

```



```

## CategoryFOOD_AND_DRINK      -1.9245304 -0.350283799
## CategoryGAME                 -1.7144145 -0.440208758
## CategoryHEALTH_AND_FITNESS  -1.1579850  0.234859830
## CategoryHOUSE_AND_HOME      -1.8589521 -0.154249638
## CategoryLIBRARIES_AND_DEMO  -1.6753729 -0.013929415
## CategoryLIFESTYLE            -1.7266546 -0.384632550
## CategoryMAPS_AND_NAVIGATION -1.8230836 -0.324122494
## CategoryMEDICAL              -1.0745092  0.278104144
## CategoryNEWS_AND_MAGAZINES  -1.9676362 -0.588077913
## CategoryPARENTING            -1.1329659  0.767134745
## CategoryPERSONALIZATION     -0.9064690  0.467058696
## CategoryPHOTOGRAPHY         -1.5840009 -0.220621966
## CategoryPRODUCTIVITY        -1.2744243  0.082358331
## CategorySHOPPING            -1.3683237  0.061645111
## CategorySOCIAL               -1.6699724 -0.265428462
## CategorySPORTS               -1.6054517 -0.237183307
## CategoryTOOLS                -1.6988327 -0.426483284
## CategoryTRAVEL_AND_LOCAL     -1.7013177 -0.301517338
## CategoryVIDEO_PLAYERS        -1.9324795 -0.477360421
## CategoryWEATHER              -1.6017416  0.138265988
## Price                        -0.0103116 -0.002293666

##      pred
## actual  0    1
##      0  988 1467
##      1   700 3401

```

For the results we took all of the reduced models and ran them on the test dataset. The Quasi results gave unexpected negative results and the Poisson had 1 result with a flag of 2. The Binomial reduced results were closest to the training results.

Conclusion

While the Apple store data was difficult to deal with since it did not come with the number of downloads, we were still able to use a zero inflated model to get some predictions. The predictions took into account that there were not that many apps with ratings in the millions so it had a reasonable max number of ratings, around 500,000.

The Google Play store data was easier to work with since that did come with a total number of installs. We were able to get a lot more visualizations with it and run numerous models. We saw that the Binomial reduced model was the biggest help when trying to predict how many times an app would be downloaded.

Appendix

Code for this project can be found here: https://github.com/dquarshie89/DATA-621-FINAL_PROJECT/blob/master/AppPrediction.Rmd