Google Analytics Customer Revenue Prediction

Tianying Xu

I. Abstract

This report analyzes and predicts customer revenue data of Google store through linear mixed model, logistic model, Censored Regression with Conditional Heteroscedasticy(CRCH) model and Light Gradient Boosting Machine(LGBM) method. From aspect of Root-Mean-Squared-Error(RMSE), LGBM leads to the least of 1.61. From aspect of association, all models help detect relation between revenue and visit information. Then, interpretation and implication are mentioned to show the result of the analysis. In the end, limitation and future discussion are stated to look forward to an improvement in the future for this analysis.

II. Introduction

1. Background

The 80/20 rule has proven true for many businesses–only a small percentage of customers produce most of the revenue. As such, marketing teams are challenged to make appropriate investments in promotional strategies. Thus, the goals are first to figure out association between factors and customer revenue, second to predict the revenue so that marketing teams can make effective investments.

2. Previous work

Several notes have been put on the kernel on Kaggle.com about the customer revenue prediction for Google store. Shivam Bansal [1] discovered missing data in the whole data set and conducted LGBM in python, where RMSE on test set is 1.64. kxx [2]

created various plots for predictors in the data set, and fitted time series model with RMSE 0.34 on train set, linear mixed model with only random intercept by users, LASSO model, neural network and XGBoost with RMSE 1.696 on test set. Erik Bruin [3] grouped data by workday and by month, also used time series model and LGBM model with RMSE 1.72 on test set.

III. Method

1. Data source

The data set is all from kaggle.com (https://www.kaggle.com/c/ga-customer-revenue-prediction), where it contains train dataset, test dataset and submission file. The size of all three file is more than 30GB, so I use Shared Computer Cluster (SCC) to deal with the data. There is one column "hits" in the data which is pretty large, thus I ignore this column when importing the data. Also, there are several columns with sub-column information in json format, therefore I use "jsonlite" package in r to convert these column into normal columns. Moreover, there are constant columns like "social Engagement Type" and "visits", and I delete them.

2. Model used

I try several different kinds of models, and the best model in that category is displayed below. Then comparison of these models will be in the model choice part.

A. Linear Mixed Model (LMM)

1) Summary

Since the data has repeated observations for one user, I decide to use multilevel models as basic model.

2) Exploratory Data Analysis(EDA)

I group the data by users, and create plots to decide which variables need random intercepts and slopes.

i. Outcome

First, I create plot for revenue, which is the outcome.

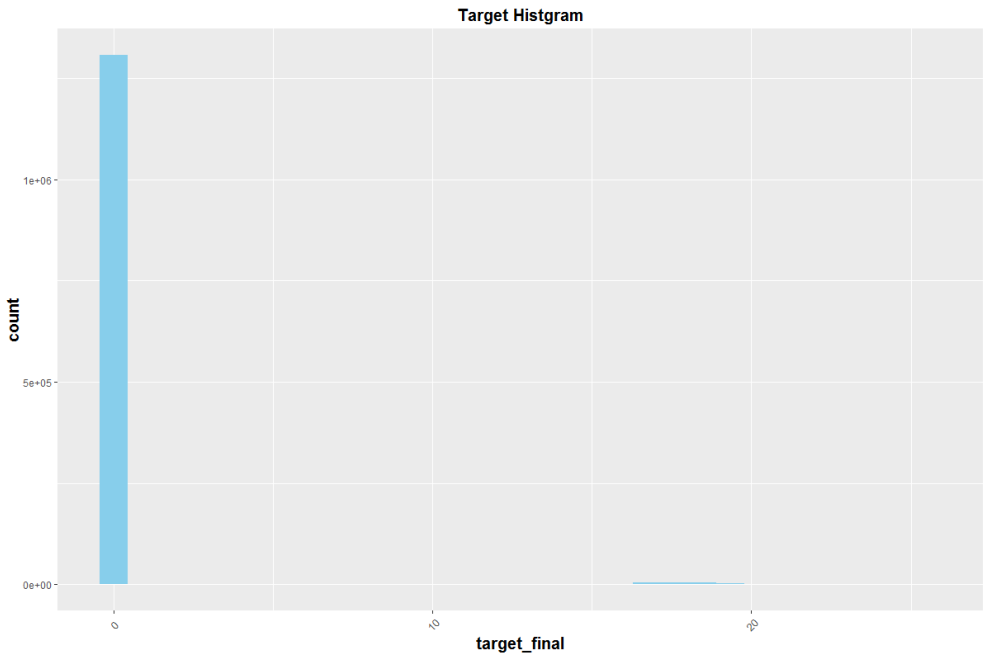


Figure: distribution of all revenue

From the plot, it is clear that most target values are 0, and it is severely right skewed, thus I check the target value without 0s.

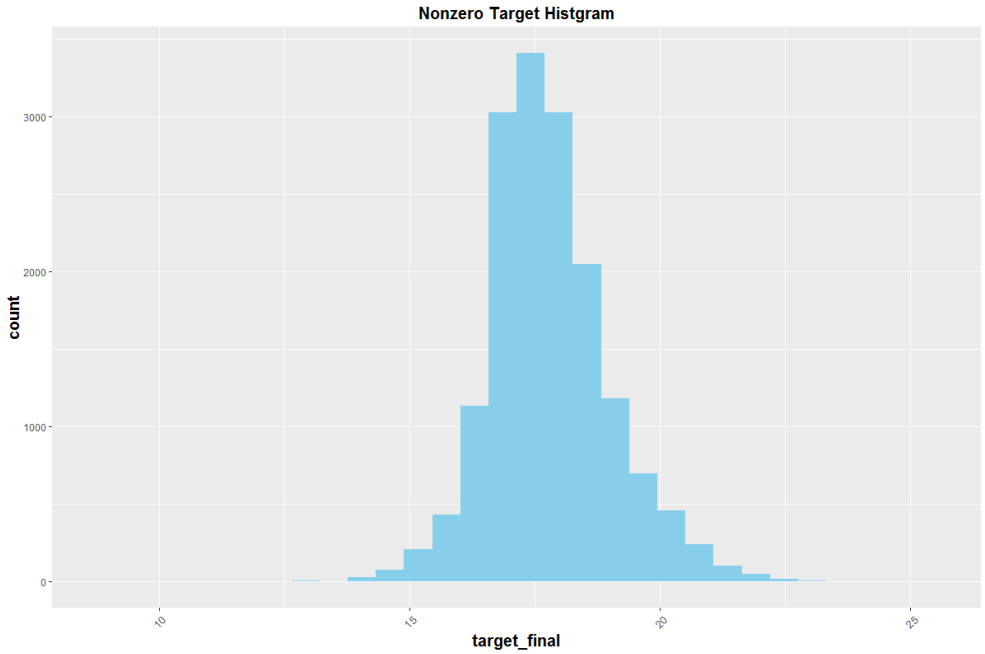


Figure: distribution of nonzero revenue

It seems like a normal distribution, so I create a qqplot.

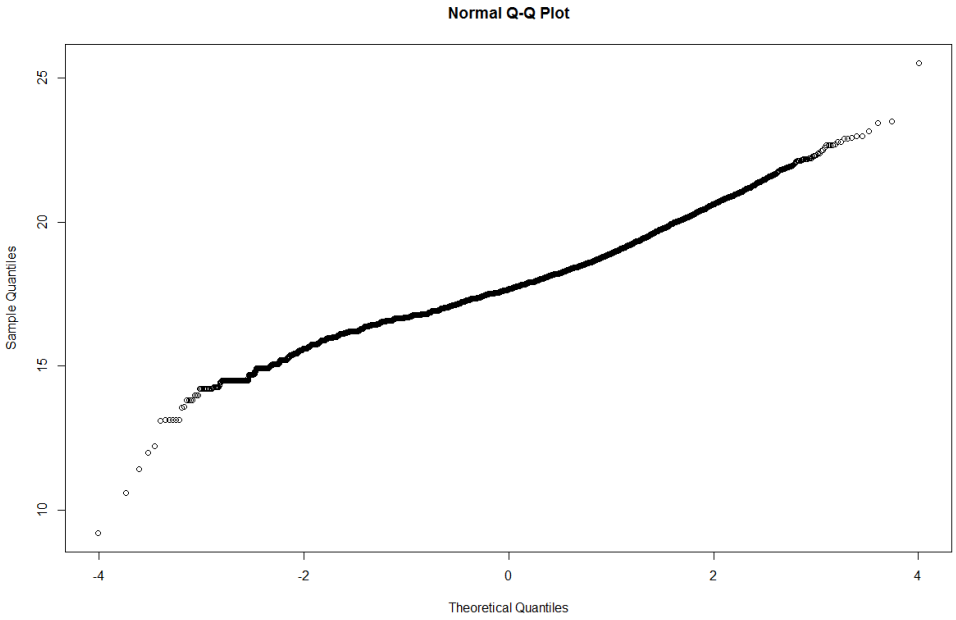


Figure: distribution of nonzero revenue

It seems that except for head and tail, nonzero target values follows a normal distribution approximately.

ii. Predictors and outcome

Then, I draw plots between predictors and outcome.

First, I check the correlation between numerical predictors.

```
            v h p b n t
visitNumber 1
hits          1
pageviews     B 1
bounces       . . 1
newVisits           1
timeOnSite    , , .   1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1
```

Figure: correlation between numerical predictors

From the table we can see that “hits” and “page views” has correlation of 1, thus I select “page views” instead of hits, also correlation between “time on site” and “page views” is more than 0.8. Thus, there may be collinearity. Since the correlation is large as more than 0.8, I choose “page views” instead of “time on site”.

Then, I create plots about “page views”, “bounces”, “new visits” and “visit number”. Since there are like more than 1 million users, I choose top 300 users with most visit number.

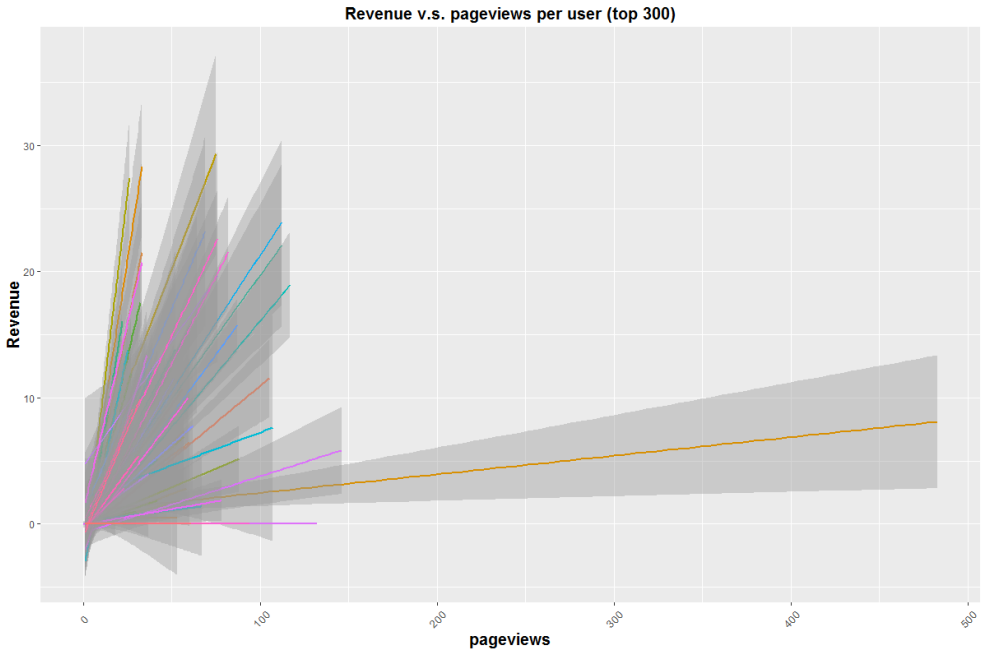


Figure: Revenue versus page views per user (top 300)

From this plot, we can see that the slope of “page views” will change a lot between different users, however there are no correlation between intercept and slope. This may indicate adding (0+pageviews|fullVisitorId) in the linear mixed models.

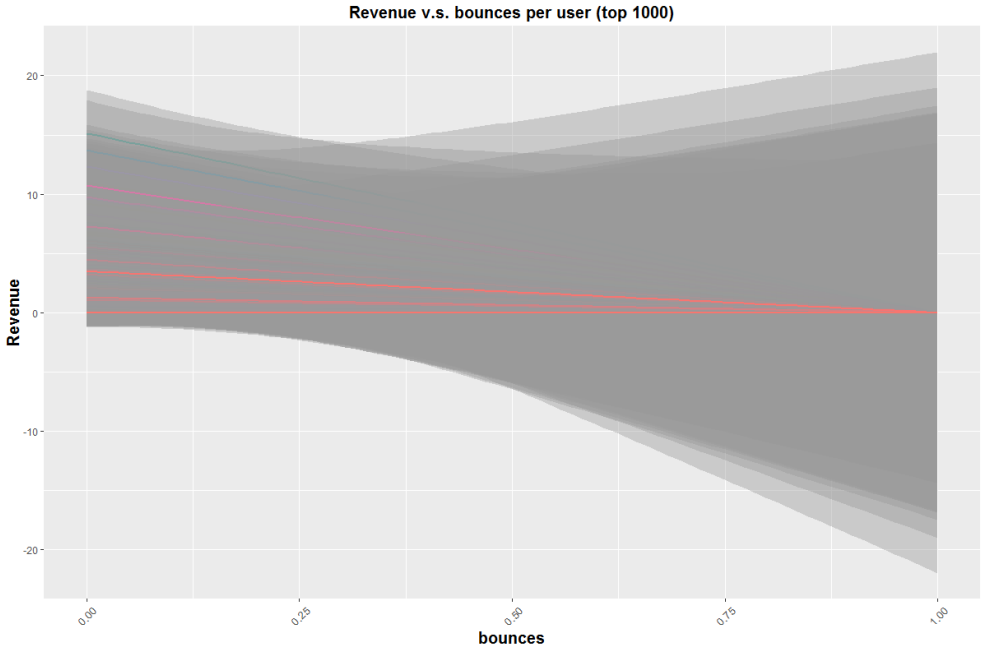


Figure: Revenue versus bounces per user (top 1000)

From the plots of Revenue versus bounces, however, we can see that both slope and intercept don't change a lot between different users. This may indicate no mixed effect for bounces in the linear mixed models. But it can be caused by small sample size, thus this plot is created on 1000 users. As showed in the plot, it still does not change a lot, thus bounces will not be included in mixed effect in the model. Then I check the whole effect of bounces to revenue and it is evident. Thus "bounces" can be fixed effect but not mixed effect.

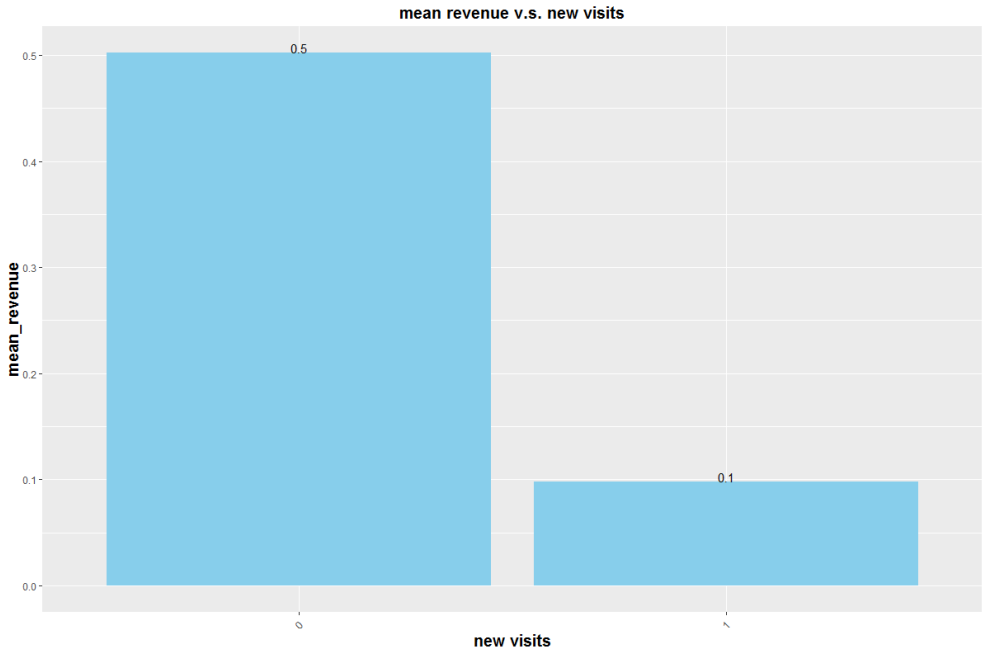


Figure: Revenue versus new visits

From this plot, it is clear that "new visits" has great impact on outcome, thus it should be included in the model.

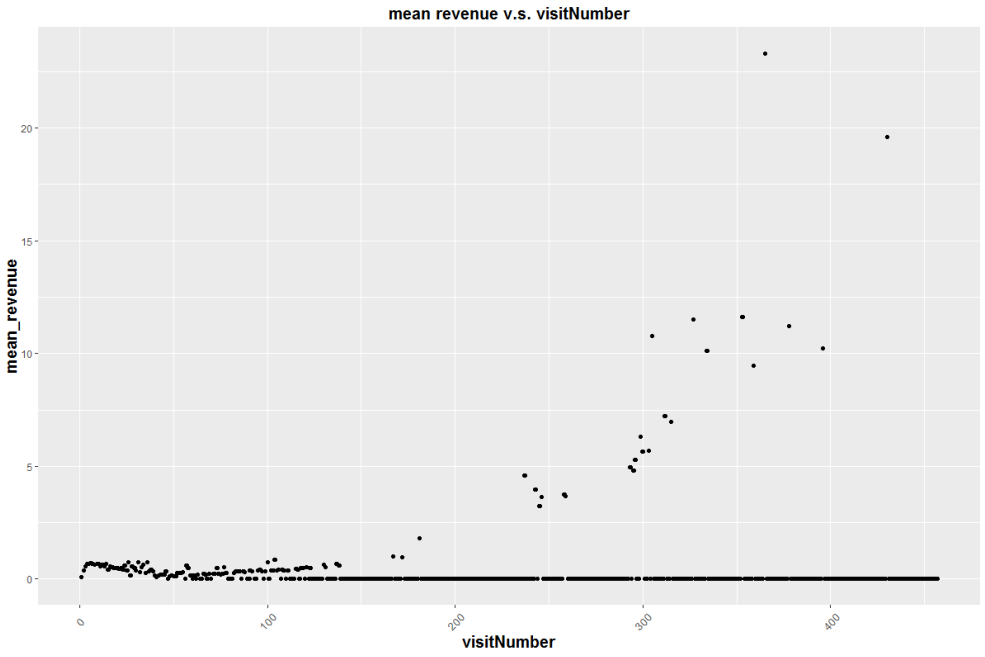


Figure: Revenue versus visit number

It is evident that "visit number" has great impact on revenue.

Next is to deal with "browser", "operating system" and "is mobile". Since there are

too many levels in “browser” and “operating system”, I recode them according to the mean outcome. Here are plots after recoding.

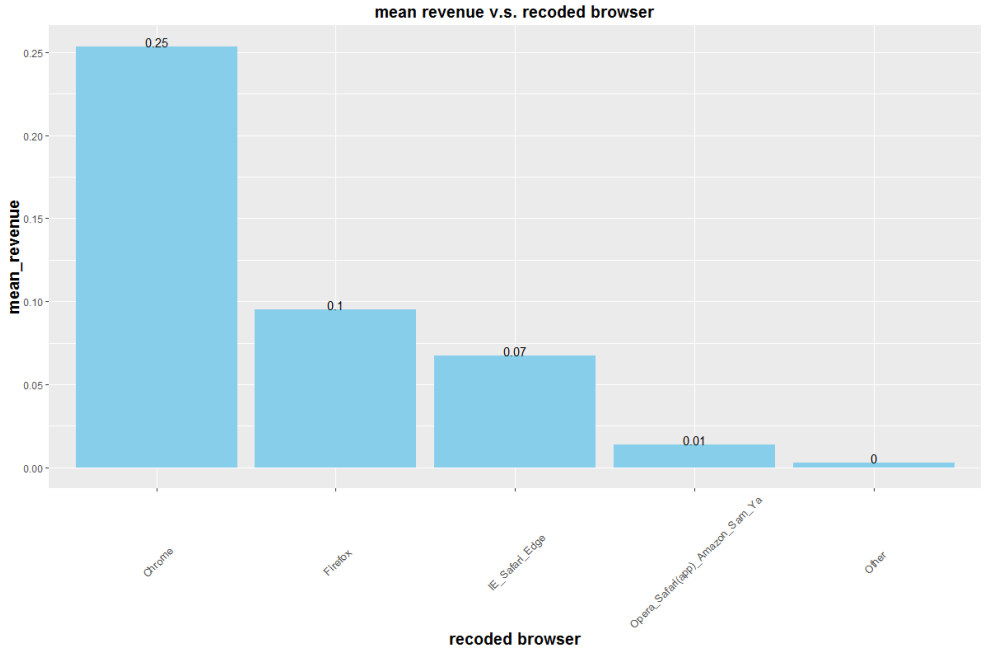


Figure: Revenue versus recoded browsers

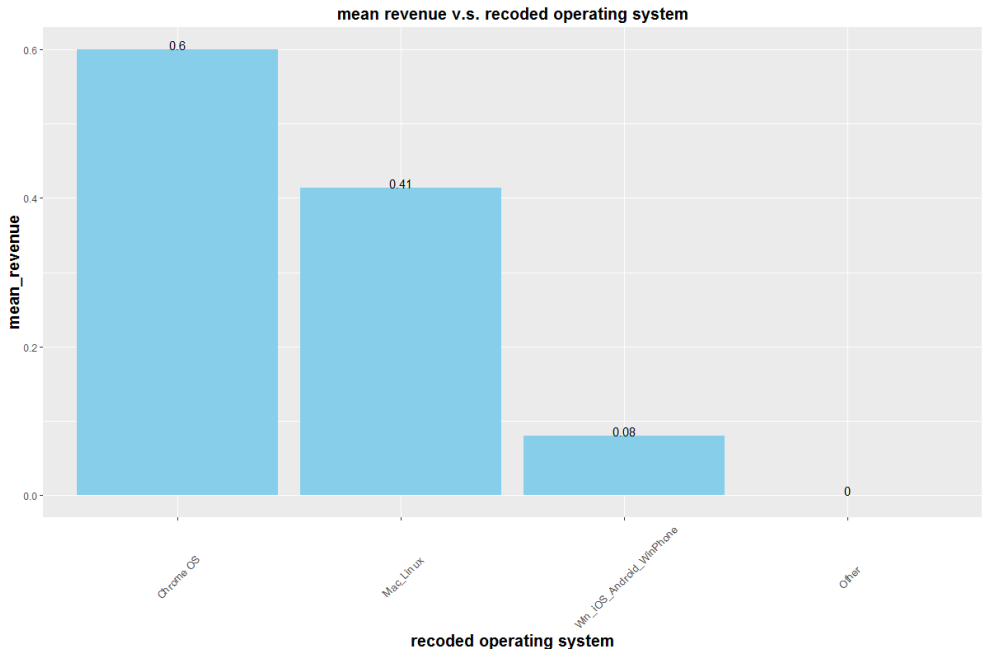


Figure: Revenue versus recoded operating system

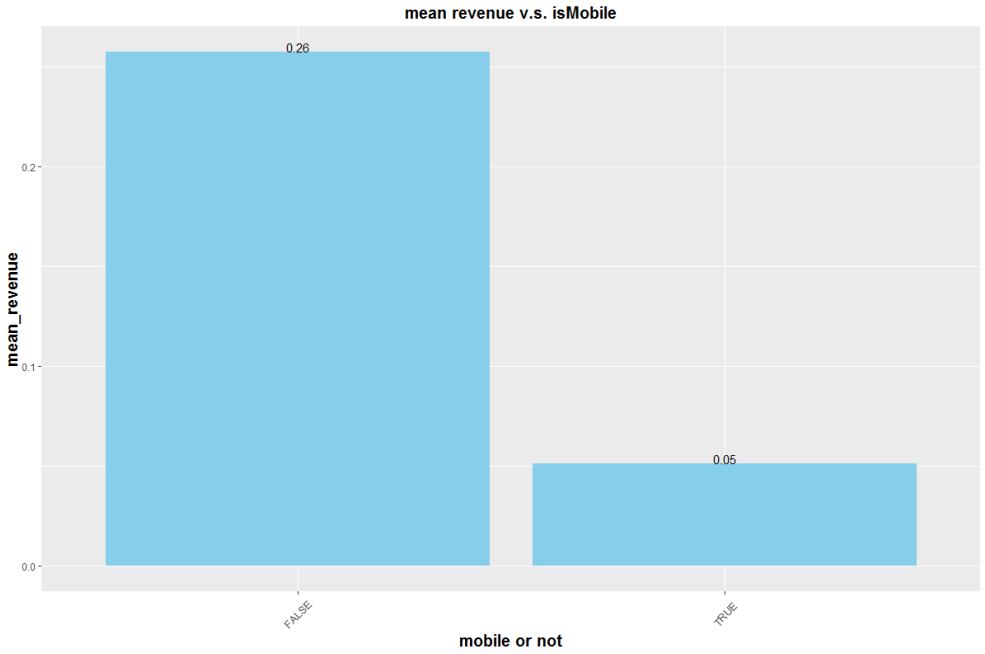


Figure: Revenue versus "is mobile"

From these plots, it is evident that all "browser", "operating system" and "is mobile" have huge impacts on revenue.

Also, for geographical information, we can see that there are huge difference between America and outside America. Thus, I recode continent as "America" and "outside America".

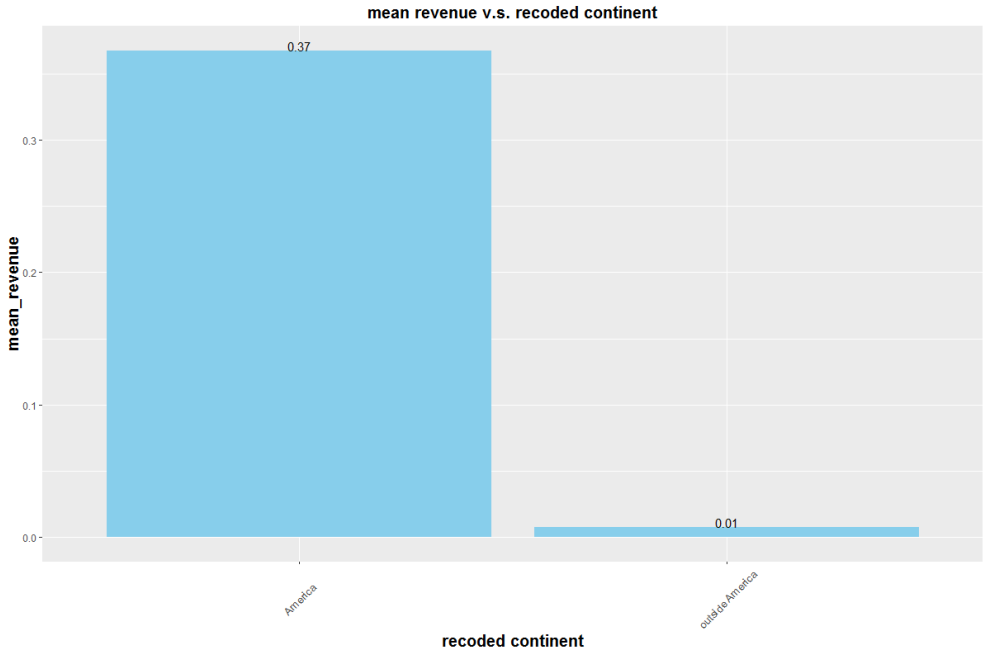


Figure: Revenue versus recoded continent

It is obvious that America has much more revenue than outside America.

Variables of channels like “channel grouping”, “medium” and “is true direct” need to be dealt with.

	channelGrouping <fctr>	medium <fctr>	isTrueDirect <int>
1	Organic Search	organic	0
2	Social	referral	0
3	Social	referral	0
4	Organic Search	organic	0
5	Direct	(none)	1
6	Direct	(none)	1

Figure: variables “channel grouping”, “medium” and “is true direct”

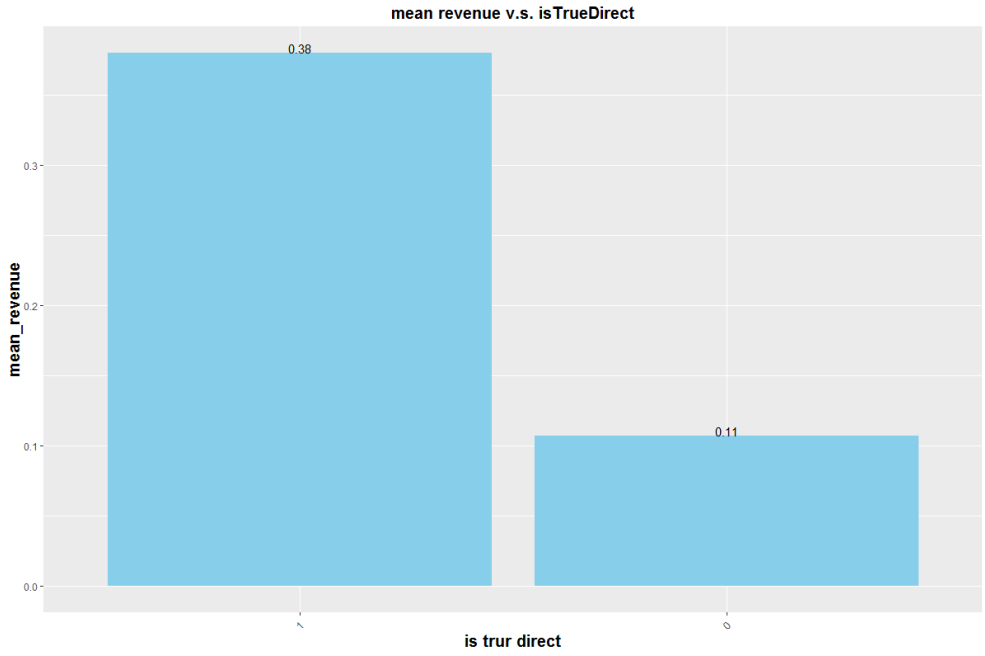


Figure: Revenue versus “is true direct”

From the table, it is clear that the "channel grouping" variable contain similar information with "medium" and "is true direct". From the plot, and also compare it with plots of “channel grouping” and “medium”, we can see that direct or not affect revenue significantly. Thus, "is true direct" should be put in the models. (Plots of “channel grouping” and “medium” are in Appendix I)

Finally, I convert date into weekday or not since there are huge difference between revenue on weekdays and on weekends.

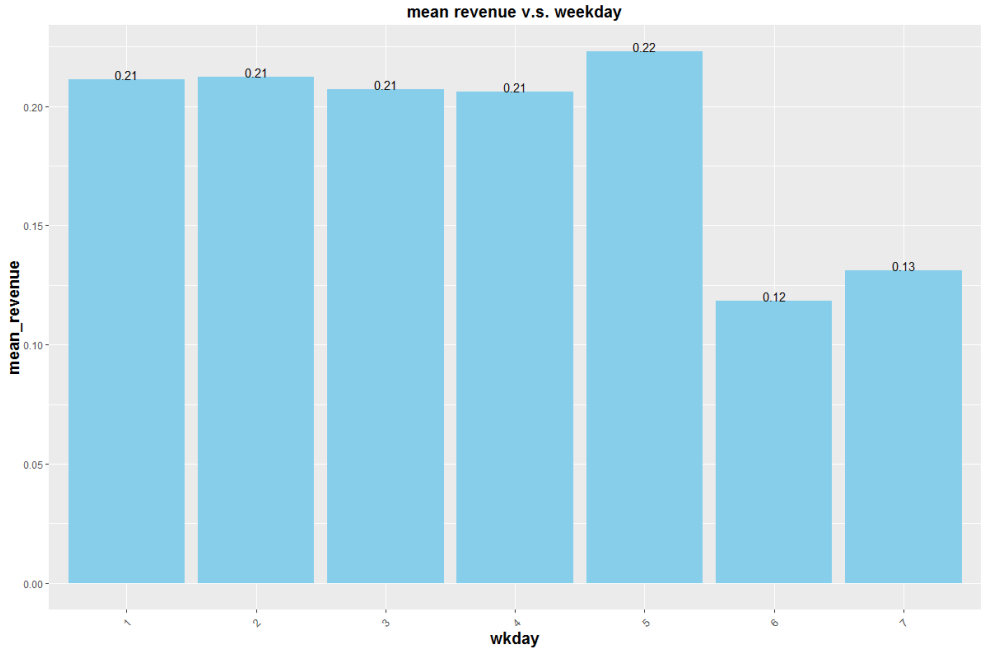


Figure: Revenue versus weekday

3) Modeling

I fit linear mixed model with possible predictors.

```
Linear mixed model fit by REML ['lmerMod']
Formula: log(transactionRevenue + 1) ~ (1 | fullVisitorId) + (0 + pageviews |
    fullVisitorId) + scale(pageviews) + scale(bounces) + factor(newVisits) +
    scale(visitNumber) + factor(browser_n) + factor(operatingSystem_n) +
    factor(isMobile) + factor(continent_n) + factor(isTrueDirect) +      factor(wkday_n)
   Data: T_predictors

REML criterion at convergence: 5237710

Scaled residuals:
     Min       1Q   Median       3Q      Max
-22.8633  -0.0443   0.0069   0.0488  23.4878

Random effects:
 Groups          Name        Variance Std.Dev.
 fullVisitorId   (Intercept) 0.00000  0.0000
 fullVisitorId.1 pageviews   0.02354  0.1534
 Residual                    0.95040  0.9749
Number of obs: 1708337, groups:  fullVisitorId, 1323730

Fixed effects:
                                                 Estimate Std. Error t value
(Intercept)                                     0.1596383  0.0058360  27.354
scale(pageviews)                                0.6060322  0.0030372 199.538
scale(bounces)                                  0.1085692  0.0010301 105.395
factor(newVisits)1                             -0.0703168  0.0025497 -27.579
scale(visitNumber)                              0.0131775  0.0009018  14.613
factor(browser_n)Firefox                        0.0164208  0.0044203   3.715
factor(browser_n)IE_Safari_Edge                 0.0310041  0.0021913  14.149
factor(browser_n)Opera_Safari(app)_Amazon_Sam_Ya  0.0226762  0.0052059   4.356
factor(browser_n)Other                          0.0354069  0.0051255   6.908
factor(operatingSystem_n)Mac_Linux             -0.0010539  0.0055540  -0.190
factor(operatingSystem_n)Other                  0.0162592  0.0111074   1.464
factor(operatingSystem_n)Win_iOS_Android_WinPhone  0.0241365  0.0055019   4.387
factor(isMobile)TRUE                           -0.0177673  0.0022772  -7.802
factor(continent_n)outside America              0.0275513  0.0017948  15.351
factor(isTrueDirect)1                           0.0156247  0.0022582   6.919
factor(wkday_n)weekend                          0.0034554  0.0019933   1.734
```

Figure: result of linear mixed model grouped by users

We can see that the residual is 0.95, which is pretty small, and most coefficients are statistically significant. However, the residual plot looks like this:

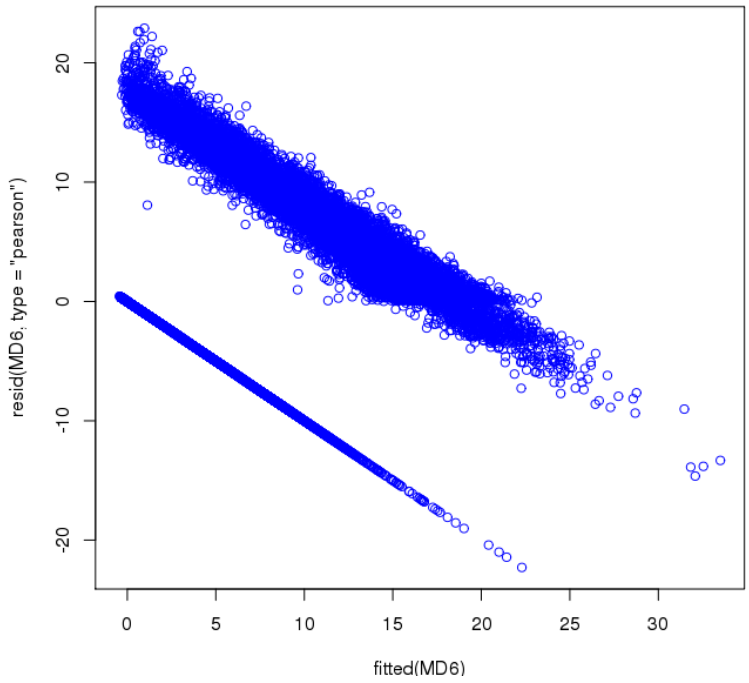


Figure: residual plot of linear mixed model grouped by users

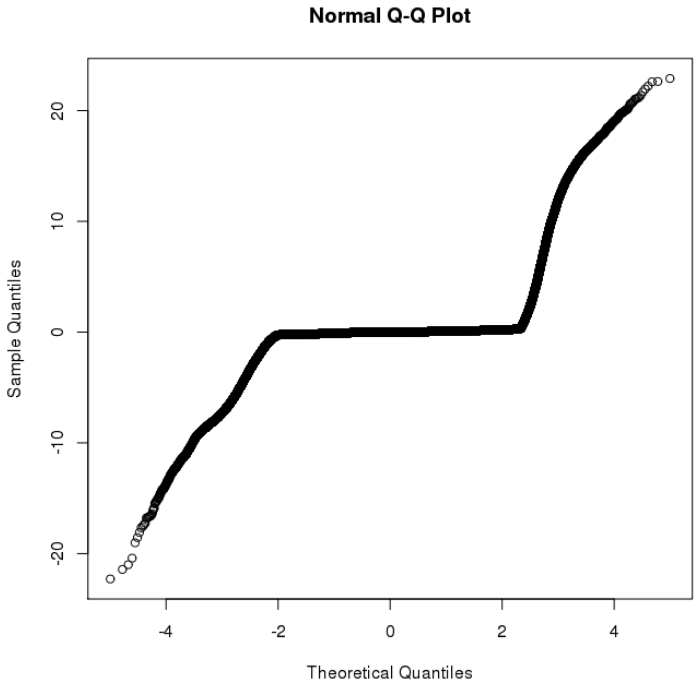


Figure: qq-plot for residuals

The lower straight line in the residual plot is due to too many 0s in the outcome, and the higher one may be due to the violation of normal distribution since the residual plot shows that it does not follow normal distribution.

Also, I predicte revenue on test dataset using this model, the Root Mean Squared Error (RMSE) is 1.97.

B. Linear Mixed Model(LMM) –Grouped by Continent

1) Summary

I also group the data by continent, since there may be similar behavior pattern within one region.

2) EDA

I create plots to see which predictors I should put in the model. According previous plot, different of mean revenue between continents are significant, so there is definitely random intercept. Therefore, I create plots to see if there are random slopes.

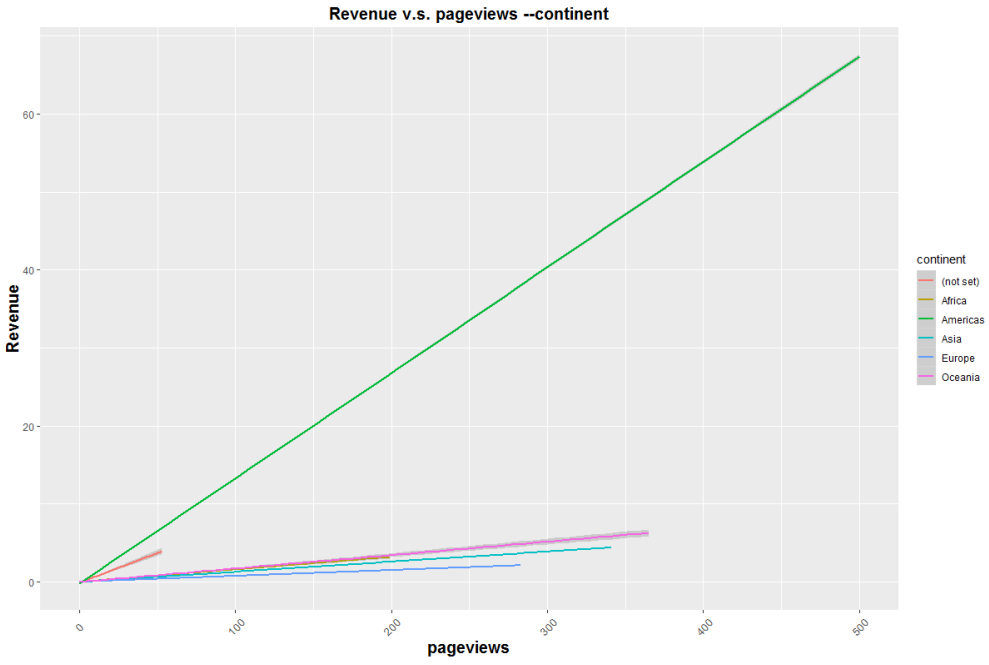


Figure: Revenue versus "page views" --continent

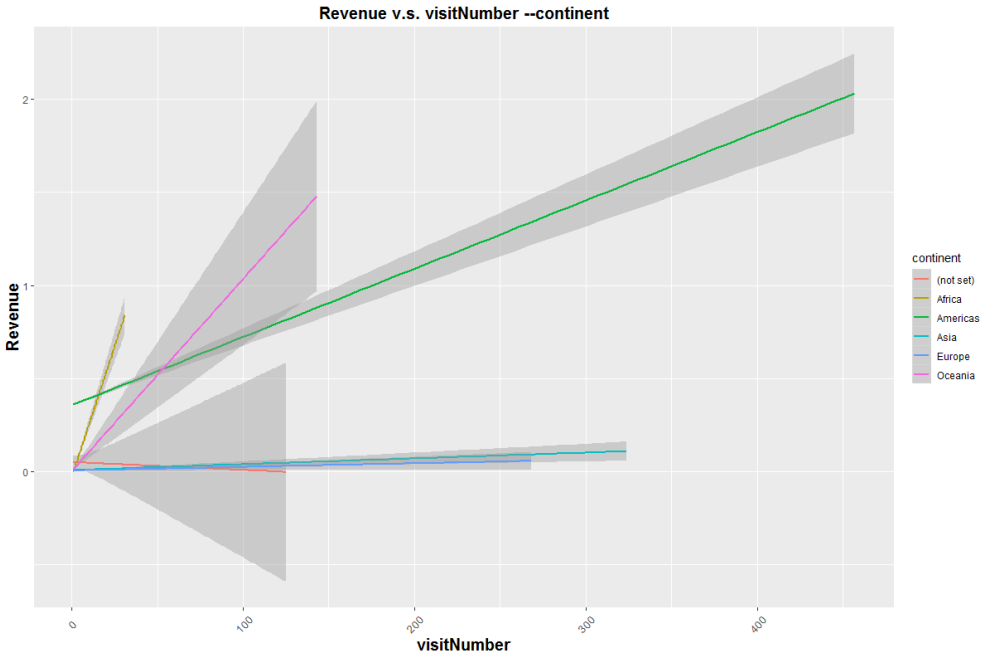


Figure: Revenue versus "visit number" --continent

It is significant that there is random slope of "page views" and "visit number" between continents.

3) Modeling

```
Linear mixed model fit by REML ['lmerMod']
Formula: log(transactionRevenue + 1) ~ (1 | continent) + (0 + pageviews |
    continent) + scale(pageviews) + scale(bounces) + scale(visitNumber) +
    factor(newVisits) + factor(isMobile) + factor(isTrueDirect) +
    factor(browser_n) + factor(operatingSystem_n) + factor(wkday_n)
   Data: T_predictors

REML criterion at convergence: 6591426

Scaled residuals:
    Min      1Q  Median      3Q     Max
-41.322  -0.097   0.001   0.059  13.310

Random effects:
 Groups      Name        Variance Std.Dev.
 continent   (Intercept) 0.012300 0.11091
 continent.1 pageviews   0.002437 0.04937
 Residual                2.774300 1.66562
Number of obs: 1708337, groups:  continent, 6

Fixed effects:
                                                 Estimate Std. Error t value
(Intercept)                                      0.334769   0.087774   3.814
scale(pageviews)                                 0.329438   0.130780   2.519
scale(bounces)                                   0.088995   0.001447  61.491
scale(visitNumber)                              -0.024265   0.001324 -18.334
factor(newVisits)1                              -0.177700   0.003920 -45.336
factor(isMobile)TRUE                            -0.039066   0.003486 -11.205
factor(isTrueDirect)1                            0.074121   0.003509  21.125
factor(browser_n)Firefox                        -0.018232   0.006866  -2.655
factor(browser_n)IE_Safari_Edge                 -0.031621   0.003368  -9.389
factor(browser_n)Opera_Safari(app)_Amazon_Sam_Ya 0.032128   0.008255   3.892
factor(browser_n)Other                           0.055217   0.008374   6.594
factor(operatingSystem_n)Mac_Linux              -0.021734   0.007784  -2.792
factor(operatingSystem_n)Other                  -0.149418   0.017478  -8.549
factor(operatingSystem_n)Win_iOS_Android_WinPhone -0.143249 0.007774 -18.427
factor(wkday_n)weekend                          -0.025220   0.003106  -8.120
```

Figure: Output for LMM by continents

The residual here is 2.77, which is larger than previous one. Most coefficients are statistically significant since t values are mostly large. Also, the residual plot still look like the previous one:

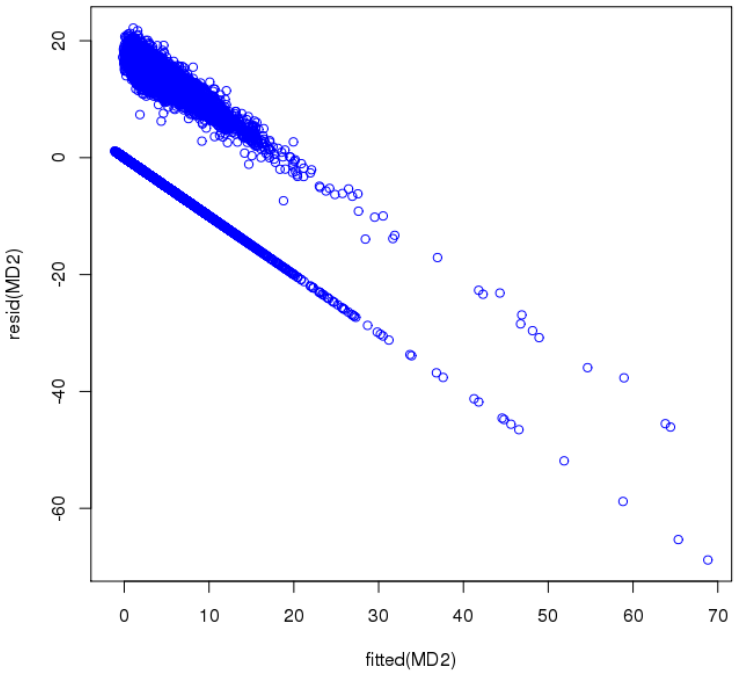


Figure: residual plot

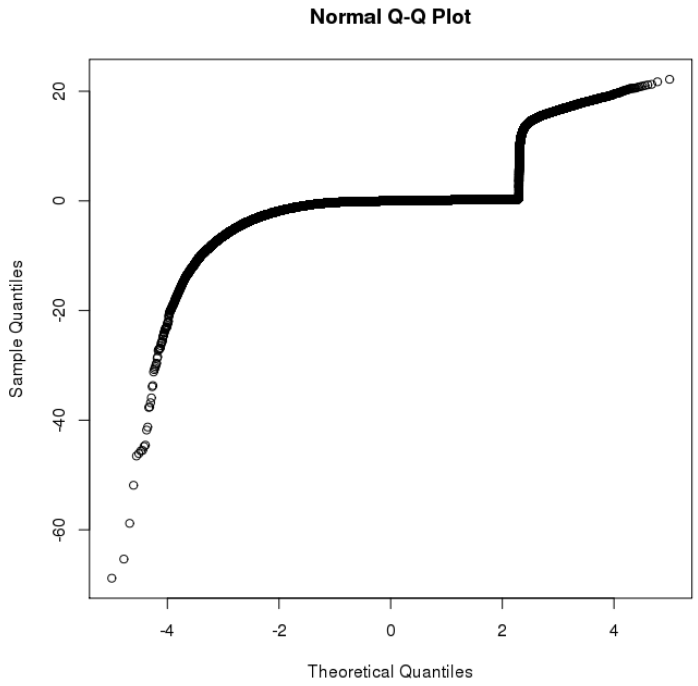


Figure: qq-plot for residuals

Still, the assumption of normal distribution is violated. This may indicate that only mixed linear model here is not enough. Thus, I try first to fit logistic model and then to fit linear mixed model.

Also, RMSE on the test dataset is 1.934, which is a little bit better than before.

C. Logistic & Linear Mixed Model

Since the percentage of nonzero is pretty low in outcome, it may be better to use hurdle model, which is to fit the logistic model first to predict the nonzero value, then fit linear mixed model on the conditional dataset.

Here is the result of the best logistic model:

```
glm(formula = iftransaction ~ scale(pageviews) + scale(bounces) +
    factor(newVisits) + scale(timeOnSite) + factor(browser_n) +
    factor(operatingSystem_n) + factor(isMobile) + factor(continent_n) +
    factor(wkday_n) + factor(isTrueDirect) + scale(visitNumber) +
    pageviews * browser_n + visitNumber * operatingSystem_n +
    visitNumber * isMobile + visitNumber * isTrueDirect + pageviews *
    operatingSystem_n + timeOnSite * newVisits + timeOnSite *
    isTrueDirect + visitNumber * browser_n, family = binomial(link = "logit"),
    data = T_logis)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.1025   0.0000   0.0000   4.0002

Coefficients: (13 not defined because of singularities)
                                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                                     -1.207e+01  2.358e+01  -0.512 0.608609
scale(pageviews)                                 4.774e-01  1.226e-02  38.939  < 2e-16 ***
scale(bounces)                                  -8.439e+00  2.310e+01  -0.365 0.714917
factor(newVisits)1                              -8.915e-01  3.008e-02 -29.635  < 2e-16 ***
scale(timeOnSite)                                1.328e-01  9.957e-03  13.336  < 2e-16 ***
factor(browser_n)Firefox                         4.278e-01  7.316e-02   5.847 4.99e-09 ***
factor(browser_n)IE_Safari_Edge                 -2.887e-01  4.623e-02  -6.246 4.22e-10 ***
factor(browser_n)Opera_Safari(app)_Amazon_Sam_Ya -7.932e-01  2.781e-01  -2.852 0.004344 **
factor(browser_n)Other                          -7.157e-03  1.077e+00  -0.007 0.994698
factor(operatingSystem_n)Mac_Linux              -2.994e-01  4.957e-02  -6.041 1.54e-09 ***
factor(operatingSystem_n)Other                  -1.077e+01  1.264e+03  -0.009 0.993202
factor(operatingSystem_n)Win_iOS_Android_WinPhone -5.020e-01  5.227e-02  -9.604  < 2e-16 ***
factor(isMobile)TRUE                            -4.491e-01  3.790e-02 -11.850  < 2e-16 ***
factor(continent_n)outside America              -3.302e+00  6.274e-02 -52.632  < 2e-16 ***
factor(wkday_n)weekend                          -2.081e-01  2.484e-02  -8.379  < 2e-16 ***
factor(isTrueDirect)1                            5.562e-01  3.092e-02  17.987  < 2e-16 ***
scale(visitNumber)                              -5.019e-02  5.498e-02  -0.913 0.361296
pageviews                                               NA         NA      NA       NA
browser_nFirefox                                        NA         NA      NA       NA
browser_nIE_Safari_Edge                                 NA         NA      NA       NA
browser_nOpera_Safari(app)_Amazon_Sam_Ya                NA         NA      NA       NA
browser_nOther                                          NA         NA      NA       NA
visitNumber                                             NA         NA      NA       NA
operatingSystem_nMac_Linux                              NA         NA      NA       NA
operatingSystem_nOther                                  NA         NA      NA       NA
operatingSystem_nWin_iOS_Android_WinPhone               NA         NA      NA       NA
isMobileTRUE                                            NA         NA      NA       NA
isTrueDirect                                            NA         NA      NA       NA
```

```
timeOnSite                                                 NA         NA      NA       NA
newVisits                                                  NA         NA      NA       NA
pageviews:browser_nFirefox                          -4.160e-02  2.933e-03 -14.184  < 2e-16 ***
pageviews:browser_nIE_Safari_Edge                   -3.183e-03  1.859e-03  -1.713 0.086804 .
pageviews:browser_nOpera_Safari(app)_Amazon_Sam_Ya  -1.809e-03  9.287e-03  -0.195 0.845542
pageviews:browser_nOther                            -4.503e-02  7.263e-03  -6.199 5.67e-10 ***
visitNumber:operatingSystem_nMac_Linux              -1.535e-02  4.670e-03  -3.288 0.001010 **
visitNumber:operatingSystem_nOther                   2.409e-01  5.495e+02   0.000 0.999650
visitNumber:operatingSystem_nWin_iOS_Android_WinPhone  1.791e-03  4.254e-03   0.421 0.673688
visitNumber:isMobileTRUE                            -8.735e-03  6.065e-03  -1.440 0.149827
visitNumber:isTrueDirect                            -4.119e-03  4.356e-03  -0.946 0.344261
pageviews:operatingSystem_nMac_Linux                 2.432e-02  2.020e-03  12.043  < 2e-16 ***
pageviews:operatingSystem_nOther                    -1.164e+00  4.864e+02  -0.002 0.998090
pageviews:operatingSystem_nWin_iOS_Android_WinPhone -6.961e-03  2.061e-03  -3.377 0.000732 ***
timeOnSite:newVisits                                 2.321e-04  2.669e-05   8.698  < 2e-16 ***
isTrueDirect:timeOnSite                             -2.288e-04  2.684e-05  -8.525  < 2e-16 ***
browser_nFirefox:visitNumber                         5.428e-03  1.700e-03   3.192 0.001412 **
browser_nIE_Safari_Edge:visitNumber                  6.759e-03  4.066e-03   1.663 0.096402 .
browser_nOpera_Safari(app)_Amazon_Sam_Ya:visitNumber -7.358e-02  1.065e-01  -0.691 0.489499
browser_nOther:visitNumber                          -8.598e-01  9.214e-01  -0.933 0.350753
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 204369  on 1708336  degrees of freedom
Residual deviance: 116291  on 1708302  degrees of freedom
AIC: 116361

Number of Fisher Scoring iterations: 21
```

Figure: Logistic model output

The accuracy on train dataset is 98.9% and on test dataset is 98.8%, which seem pretty high. However, it only predict 840 outcome as 1 out of 401590, and in fact there are 4594 outcome should be 1.Then I fit linear mixed model using nonzero outcome observations, the result is as follow:

```
Linear mixed model fit by REML ['lmerMod']
Formula: log(transactionRevenue + 1) ~ (0 + visitNumber | fullVisitorId) +
    scale(visitNumber) + browser_n + operatingSystem + isMobile +      isTrueDirect + continent + wkday_n
   Data: T2

REML criterion at convergence: 56634.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.4213 -0.6202 -0.0677  0.5720  4.6681

Random effects:
 Groups        Name        Variance Std.Dev.
 fullVisitorId visitNumber 0.005931 0.07701
 Residual                  1.146752 1.07087
Number of obs: 18514, groups:  fullVisitorId, 16141

Fixed effects:
                                      Estimate Std. Error t value
(Intercept)                           17.61125    0.46967   37.50
scale(visitNumber)                     0.47714    0.04661   10.24
browser_nFirefox                       0.02294    0.06321    0.36
browser_nIE_Safari_Edge               -0.33395    0.03941   -8.47
browser_nOpera_Safari(app)_Amazon_Sam_Ya -0.72349    0.18153   -3.99
browser_nOther                        -0.67686    0.34225   -1.98
operatingSystemChrome OS               0.82765    0.23479    3.53
operatingSystemiOS                     0.19664    0.06423    3.06
operatingSystemLinux                   0.30420    0.23514    1.29
operatingSystemMacintosh               0.73133    0.23370    3.13
operatingSystemWindows                 0.53601    0.23264    2.30
operatingSystemWindows Phone           0.01598    1.11580    0.01
isMobileTRUE                           0.12754    0.22953    0.56
isTrueDirect                           0.18019    0.01798   10.02
continentAfrica                        0.76769    0.53901    1.42
continentAmericas                     -0.51128    0.40915   -1.25
continentAsia                         -0.33251    0.41692   -0.80
continentEurope                       -0.65848    0.42066   -1.57
continentOceania                       0.22954    0.45876    0.50
wkday_nweekend                        -0.23507    0.02328  -10.10
```

Figure: Linear mixed model output

RMSE on train set is 4.76, and on test set is 13.52, which is not better than before. I think it may be due to the prediction error of the logistic model.

D. Censored Regression with Conditional Heteroscedasticy(CRCH) Model

Since the outcome is mostly zero and the nonzero part approximately follows normal distribution, thus I decide to fit censored(tobit) regression with conditional heteroscedasticy. The result is as follow:

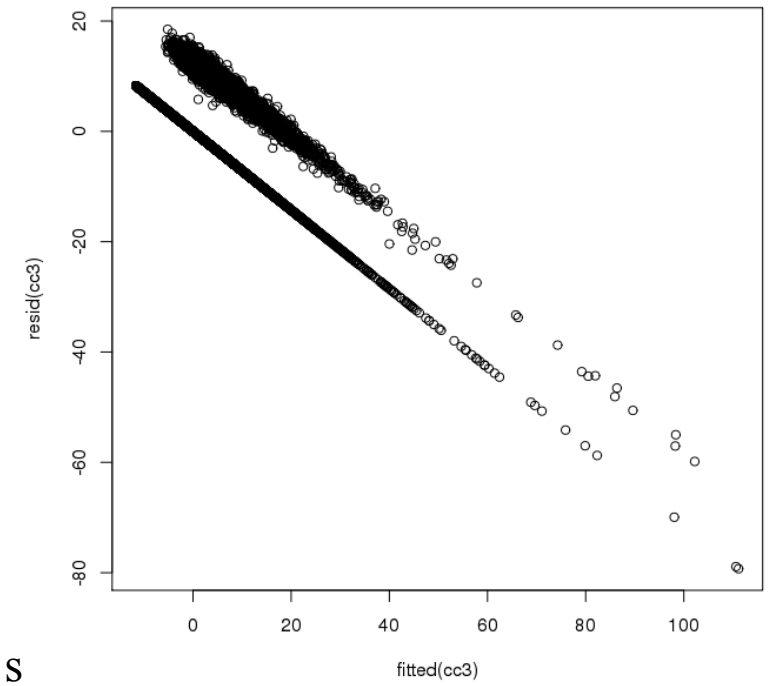
```
Call:
crch(formula = revenue ~ scale(pageviews) + scale(bounces) + factor(newVisits) + scale(timeOnSite) +
    factor(browser_n) + factor(operatingSystem_n) + factor(isMobile) + factor(continent_n) + factor(wkday_n) +
    factor(isTrueDirect) + scale(visitNumber), data = T_logis, link.scale = "log", dist = "gaussian",
    left = 0)

Standardized residuals:
     Min       1Q   Median       3Q      Max
-13.8117   2.5254   5.7373   7.3281  11.3177

Coefficients (location model):
                                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                                 -115.94537  160.98204  -0.720   0.4714
scale(pageviews)                               5.85715    0.06212  94.293  < 2e-16 ***
scale(bounces)                               -58.91452  157.73173  -0.374   0.7088
factor(newVisits)1                            -9.29471    0.30198 -30.780  < 2e-16 ***
scale(timeOnSite)                              2.22823    0.06299  35.373  < 2e-16 ***
factor(browser_n)Firefox                      -1.74694    0.79632  -2.194   0.0283 *
factor(browser_n)IE_Safari_Edge               -3.70071    0.39934  -9.267  < 2e-16 ***
factor(browser_n)Opera_Safari(app)_Amazon_Sam_Ya  -10.55829    1.96595  -5.371 7.85e-08 ***
factor(browser_n)Other                        -19.88718    3.83030  -5.192 2.08e-07 ***
factor(operatingSystem_n)Mac_Linux             0.34929    0.42140   0.829   0.4072
factor(operatingSystem_n)Other               -92.55234 1041.15583  -0.089   0.9292
factor(operatingSystem_n)Win_iOS_Android_WinPhone  -8.38142    0.46935 -17.857  < 2e-16 ***
factor(isMobile)TRUE                          -5.48486    0.42904 -12.784  < 2e-16 ***
factor(continent_n)outside America            -35.73143    0.65869 -54.246  < 2e-16 ***
factor(wkday_n)weekend                        -2.23156    0.31420  -7.102 1.23e-12 ***
factor(isTrueDirect)1                          5.13302    0.29242  17.553  < 2e-16 ***
scale(visitNumber)                            -1.51633    0.10410 -14.567  < 2e-16 ***

Coefficients (scale model with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.408021   0.006652   512.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Distribution: gaussian
Log-likelihood: -1.301e+05 on 18 Df
Number of iterations in BFGS optimization: 78
```

Figure:CRCH model output

From the regression table, we can see that coefficients are all statistically significant.

RMSE on train set is 1.8 and on test set is 1.87, which are better than before. The

residual plot looks like this:



S

Figure: Residual plot of CRCH model

We can still see the trend of the residual plot, and I think it is due to the unbalanced outcome.

E. Light Gradient Boosting Machine(LGBM)

Since there are previous work using LGBM that lead to great result, I decide to use LGBM to fit on the dataset with recoded predictors. LGBM is a mothod of machine learning. Specifically, it is a gradient boosting framework that uses tree based learning algorithm. Also, unlike other boosting algorithm, LGBM is much faster in that it grows tree leaf-wise while other algorithm grows tree level-wise. Thus, it is a suitable method for this dataset. Here is the result of LGBM:

```
[LightGBM] [Warning] learning_rate is set=0.01, learning_rate=0.01 will be ignored. Current value: learning_rate=0.01
[LightGBM] [Info] Total Bins 814
[LightGBM] [Info] Number of data: 1708337, number of used features: 13
[LightGBM] [Info] Start training from score 0.192588
[1]: val's rmse:1.88999
[101]: val's rmse:1.67385
[201]: val's rmse:1.63353
[301]: val's rmse:1.62244
[401]: val's rmse:1.61811
[501]: val's rmse:1.61693
[601]: val's rmse:1.61633
[701]: val's rmse:1.61526
[801]: val's rmse:1.61474
[901]: val's rmse:1.61447
[1000]: val's rmse:1.61401
```

Figure: Light GBM output

It is evident that through iteration, RMSE on test set is reduced to 1.61, which is much better than previous ones. Also, it can calculate the importance of each predictor:

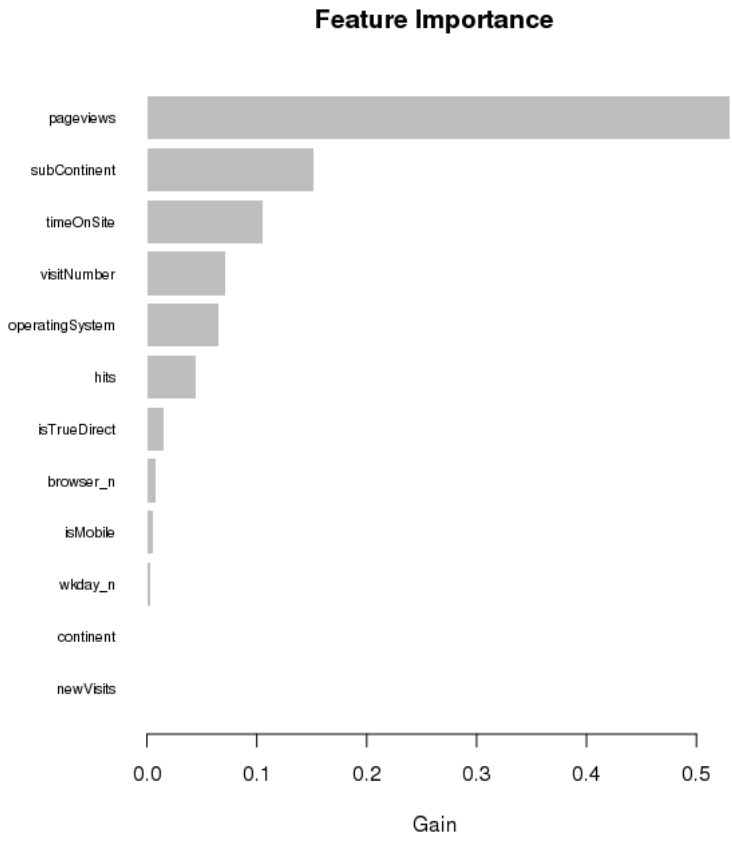


Figure: Light GBM predictors' importance

From the importance plot, we can see that "page views" is the most significant predictor for customer revenue, and then "sub-continent", "time on site", "visit number", "operating system" are also important for prediction of customer revenue. However, predictors like "continent" and "new visits" are not important for prediction.

IV. Result

1. Model choice

Model	RMSE on test set
LMM (by users)	1.97
LMM (by continents)	1.93
Logistic & LMM	13.52
CRCH Model	1.87
LGBM	1.61

Table: Comparison of models

According to RMSE on test set, it is clear that LGBM is the best model, then the CRCH model, and next the LMM models.

2. Interpretation

Although LGBM is the best model, different models explain things through different aspects. I combine results of all models to lead to the result of the whole data analysis process.

A. Important predictors for revenue prediction (every visit)

1) Here are the importance of predictors in LGBM, CRCH and LMM models ordered by importance from high to low:

LGBM		CRCH		
Feature	Gain	predictors	sign	estimate
page views	0.526	page views	+	5.86
subcontinent	0.15	continent_n-outside America	-	35.73
time on site	0.107	time on site	+	2.23
visit number	0.072	new visits	-	9.29

operating system	0.065	operating system-Win_iOS_Android_WinPhone	-	8.38
hits	0.046	is true direct-1	+	5.13
is true direct	0.015	visit number	-	1.52

Table: importance of predictors in LGBM and CRCH models

LMM---by users			LMM---by continents		
predictors	sign	estimate	predictors	sign	estimate
pageviews	+	0.61	bounces	+	0.089
bounces	+	0.11	newVisits-1	-	0.18
Newvisits-1	-	-0.07	isTrueDirect-1	+	0.074
Continent_n-outside America	+	0.027	visitNumber	-	0.024
visitNumber	+	0.013	operatingSystem_n-win_iOS_Android_WinPhone	-	0.143
Browser_n-IE_Safari_Edge	+	0.031s			
isMobile-True	-	0.018			

Table: importance of predictors in LMM models

From the tables, we can see that “page views” is the most frequent predictor that shows on the top of the list, which means it is the most important predictor to predict customer revenue. Also, all of its signs are positive, which means higher

revenue is expected when there are more “page views”. Also, impact of “page views” to prediction of revenue may change between users and continents, thus it is the most indispensable predictor through all predictors.

Geography is also a critical predictor for revenue. Combined with previous EDA part and the result here, we can find out that revenue in America is expected to be higher than which in other continents.

Also, “visit number” is a significant predictor for revenue. Most of its signs are negative, thus more “visit number” may lead to lower expectation of customer revenue.

Next, “operating system” is another crucial predictor. Since the baseline here is “Chrome OS”, thus users with “windows”, “iOS”, “Android” and “WinPhone” are expected to have higher revenue than users with “Chrome OS”.

Finally, “is true direct” is also essential when predicting customer revenue. The baseline is “0”, which means that users are not direct to the store, thus direct users correspond to higher customer revenue than indirect users.

3. Model Checking

I have already done the checking part before.

V. Discussion

1. Implication

The implication of the analysis could be used to maximize effectiveness of marketing investment. For every visit, based on its information like continent, operating system, page views and if it is true direct to the store, an expected

revenue could be predicted. Then a weight according to the expected revenue can be decided so that it can determine how much investment would be made on the visit. Ideally, higher expected revenue would lead to higher investment of marketing, thus users who are going to have transaction revenue will purchase more.

2. Limitation

Due to time limit, customer files have not been built yet. The largest column “hits”, which is also the most interesting column, has not been analyzed. It contains much customer information like users’ favorite product and their interested fields, and these can assist marketing strategies a lot. Also, for every user, there may be a purchase pattern, which can also be a great source of probability prediction for users to purchase things at time level. However, I ran out of time to finish these analysis.

3. Future direction

The whole process to analyze data for customer revenue prediction can be divided into two levels:

A. User level: This includes individual information from three aspects. Firstly, in general, whether it is easy for user to generate revenue. This could be summarized by previous revenue versus visit times. Secondly, from aspect of time, purchase pattern of every user may indicate probability of generating revenue for one visit of a user. Finally, history hits records of a user can show interested products of the user, which could help marketing strategies.

B. Visit level: This includes association between predictors of every visit like hits, page views, is true direct and so forth. Revenue for every visit can be predicted through this analysis, which is what this report mostly about. However, though LGBM may be a great method for prediction, I just give a try due to limit time. Parameters like learning rate can be tuned to improve the performance of the method.

To sum up, future direction of customer revenue prediction is to analyze data on user level, and improve methods in this report on visit level.

VI. Acknowledgement

I would like to express my special thanks of gratitude to Professor Masanao Yajima who assist me in gaining academic knowledge and helped me when I came into problems. Also, I would like to thank Katia Oleinik for helping me install “lightgbm” package in r on SCC. Finally, I want to thank my classmates and friends who discussed about this project with me and gave me help.

VII. Reference

[1] Notes for competition on kaggle.com: https://www.kaggle.com/shivamb/exploratory-analysis-ga-customer-revenue

[2] Notes for competition on kaggle.com: https://www.kaggle.com/kailex/r-eda-for-gstore-glm-keras-xgb

[3] Notes for competition on kaggle.com: https://www.kaggle.com/erikbruin/google-analytics-eda-lightgbm-screenshots

[4] Jakob W. Messner, Georg J. Mayr and Achim Zeileis. Heteroscedastic censored

and truncated regression with crch.

[5] Document of "crch" package in R: https://cran.r-project.org/web/packages/crch/crch.pdf

[6] Notes for interpretation of a LGBM model on kaggle.com: https://www.kaggle.com/slundberg/interpreting-a-lightgbm-model

[7] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye and Tie-Yan Liu. LightGBM: a highly efficient gradient boosting decision tree. 31st Conference on Neural Information Processing Systems (NIPS 2017).

[8] LGBM: https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc

VIII. Appendix

Appendix I

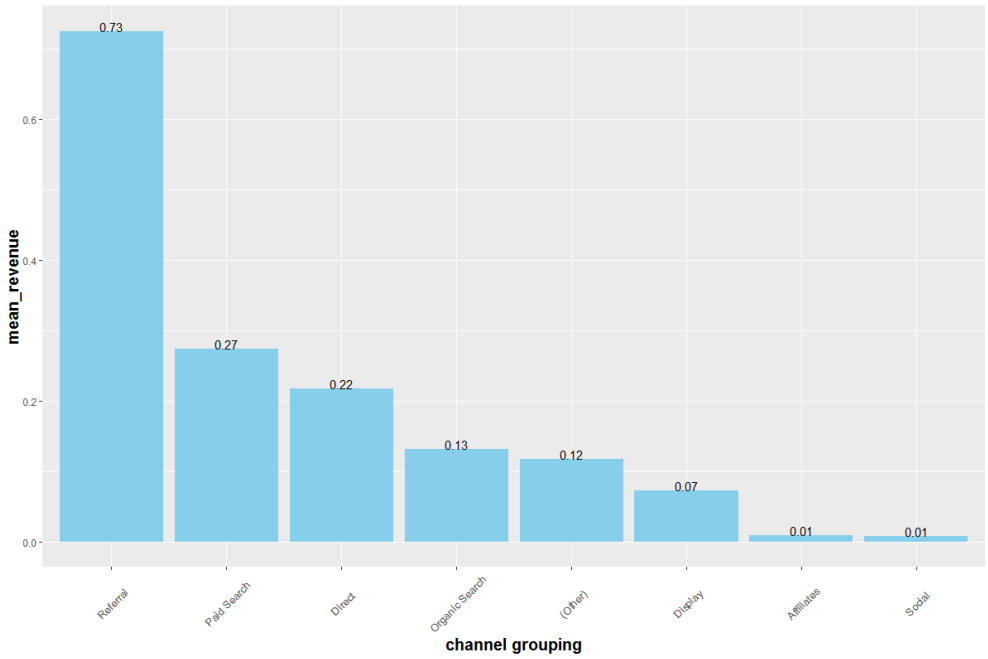


Figure: Revenue versus "channel grouping"

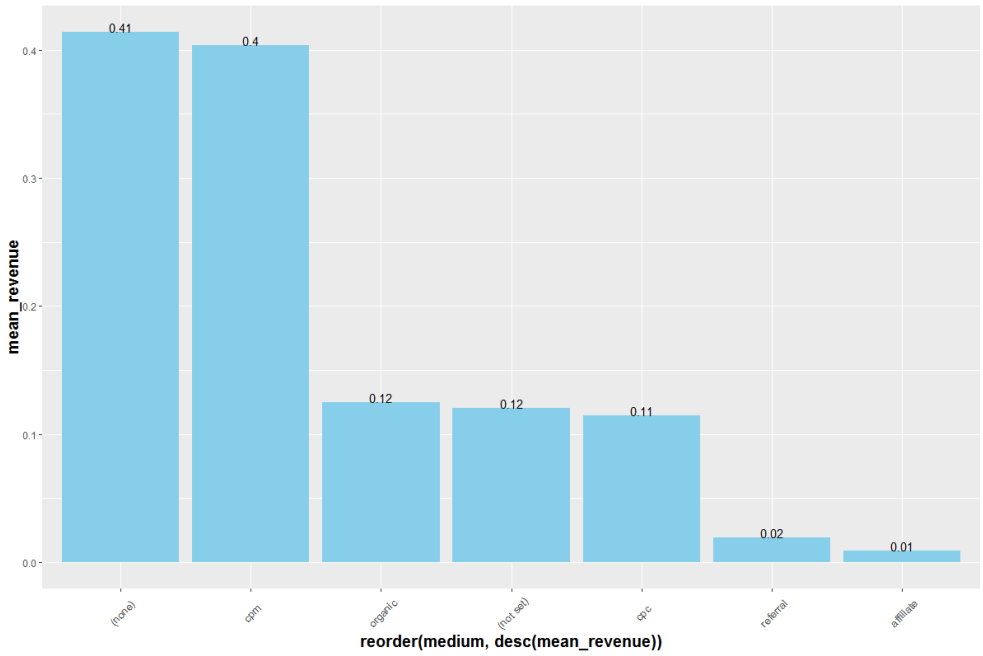


Figure: Revenue versus "medium"