Google Analytics Customer Revenue Prediction

MA678 Final Project

Author: Jingyi Niu

Date:12/09/2020

I. Abstract

This report analyzes and predicts customer revenue data of Google store through linear mixed model, logistic model,XGBoost regression model. From aspect of Root-Mean-Squared-Error(RMSE), XGBoost get the least of 3.019. 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, some future discussion are stated to look forward to an improvement in the future for this analysis.

II. Introduction

1. Project 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.

RStudio, the developer of free and open tools for R and enterprise-ready products for teams to scale and share work, has partnered with Google Cloud and Kaggle to demonstrate the business impact that thorough data analysis can have.

I will do some analysis based on this background, aiming to analyze a Google Merchandise Store (also known as GStore, where Google swag is sold) customer dataset to predict revenue per customer.

2. Data source

The data set is all from kaggle.com (https://www.kaggle.com/c/ga-customer-revenue-prediction/data), where it contains train dataset, test dataset and submission file. The size of this dataset is more than 30GB and it contains several columns with subcolumn information in json format, therefore I use “jsonlite” package in r to convert these column into normal columns. And then I select the variables needed and drop some column that are insignificant. Finally, here are several feathers which could be useful as follow, after deleting constant column. (Appendix 1)

- **fullVisitorId** - an unique identifier for each user of the Google Merchandise Store
- **channelGrouping** - the channel via which the user came to the Store
- **date** - the date on which the user visited the Store
- **device** - the specifications for the device used to access the Store
- **geoNetwork** - this section contains information about the geography of the user
- **totals** - this section contains aggregate values across the session
- **trafficSource** - this section contains information about the Traffic Source from which the session originated
- **visitId** - an identifier for this session
- **visitNumber** - the session number for this user

- **visitStartTime** - the timestamp (POSIX).

3. Main goals

First objective is to do an Exploratory Data Analysis for the Google Analytics Customer Revenue Prediction competition within the R environment. For this EDA in the main we will use tidyverse packages and ggplot2 package.
Secondly, for modelling we will use lme4, xgboost packages.
Our task is to build an algorithm that predicts the natural log of the sum of all transactions per user. Thus, for every user in the test set, the target is:

$$y_{user} = \sum_{i=1}^{n} transaction_{user_i}$$

$$target_{user} = \ln(y_{user} + 1)$$

To visualize our model performance, we would like to use the root mean square error as the evaluation method for time series and regression models. RMSE is defined as:

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where $\hat{y}$ is the predicted revenue for a customer and y is the natural log of the actual revenue value.

III. Method

1. Description of Data Analysis

This section contains the description of Exploratory Data Analysis (EDA) and Data Processing based on EDA in order to help us better understand the dataset.

1) Transaction Revenue

According to Missing Data Analysis (Appendix 2), 98.9% did not make transaction revenue during their visit. This result indicates the imbalance of our target variable. In addition, revenue range is between 0 and 23.8644 and most revenue is lower than 1.(Table 3.1) Therefore, we apply a data transformation on it: "log1p" function transforms x to log(1+x). It converts our data to approximately normal distribution with the median 17.7.(Table 3.1) In natural log scale, most transaction revenue follows between 15 and 20 (See Appendix 3, Figure 3.1)

2) Channel Grouping, Operating System, Device, and Browser

Channels are defined as how users arrive to the website. Most sessions come from "organic search", but "referral" contributes most transaction revenue. Comparison between number of sessions and revenue indicates that desktop is still most important device for G Store online shopping and it generates much more revenue than the other two device types. There are only several main system in the dataset: windows, macintosh, android, ios, linux, chrome os, and windows phone. Macintosh stand out because it generates much

higher revenue with fewer sessions. Finally, analysis on browser gives us roughly the same views on Windows and Mac users, but it seems that Chrome is also popular on Mac because it contributes more revenue than Windows from operating system category (See Appendix 4).

3) geographical information

The column GeoNetwork contain geographical information and according to Missing Data Analysis (Appendix 2), detailed information such as “city”, “region” and "metro" miss approximately 54.6%, so I choose continent and country level data to analysis. From both level,we can see that there are huge difference between America and outside America. Thus, I recode continent as “America” and “outside America”.(Figure 5.1)

4) Visits by Time series and Hits and Pageviews

Customers usually prefer to take multiple views if they are interested in a product, so the page views may be an essential factor for predicting revenue transaction. From the histogram of pageviews, we find that most visitors view less than 5 pages during a session. It hardly leads to a transaction since a shopping cycle on G store needs at least 4 to 6 pages. The distribution of total revenue grouped by pageviews agrees that revenue dramatically increased when pageviews are greater than 10 (Figure 6.8). More Exploratory Data Analysis can be founded in the Appendix 6.

2. Statistical Methods and Analysis

1) Linear Mixed Model

The first model I implemented is linear mixed effects model. It is a simple extension of linear regression but allows to both fixed and variable effects. Therefore, it works better on categorical data. I started with treating Visit ID as the intercept and used this model as the benchmark. Then we added one variable into our formula each time:

m_lmm0: transactionRevenue ~ (1 | fullVisitorId)

m_lmm1: transactionRevenue ~ scale(pageviews) + (1 | fullVisitorId)

m_lmm2: transactionRevenue ~ scale(pageviews) + (1 | fullVisitorId)

m_lmm3: transactionRevenue ~ scale(pageviews) + scale(visitNumber) + (1 | fullVisitorId)

m_lmm4: transactionRevenue ~ scale(pageviews)+scale(visitNumber)+factor(channelGrouping)+(1 | fullVisitorId)

m_lmm5: transactionRevenue ~ scale(pageviews)+scale(visitNumber)+factor(channelGrouping) + factor(browser) + (1 | fullVisitorId)

m_lmm6: transactionRevenue ~ scale(pageviews)+scale(visitNumber)+factor(channelGrouping) + factor(browser) + factor(operatingSystem)+factor(isMobile) + (1 | fullVisitorId)

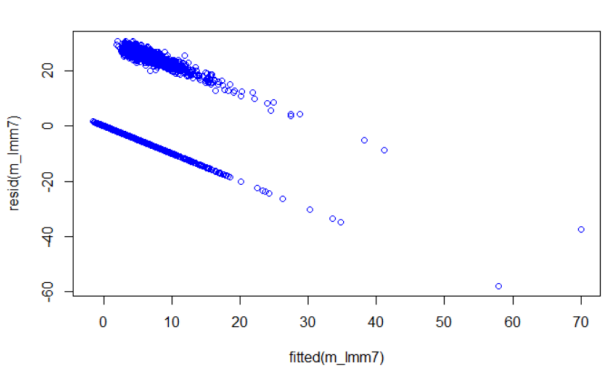
m_lmm7: transactionRevenue ~ scale(pageviews) + scale(visitNumber)+factor(channelGrouping) + factor(browser) + factor(operatingSystem)+factor(isMobile)+factor(country)+(1 | fullVisitorId)

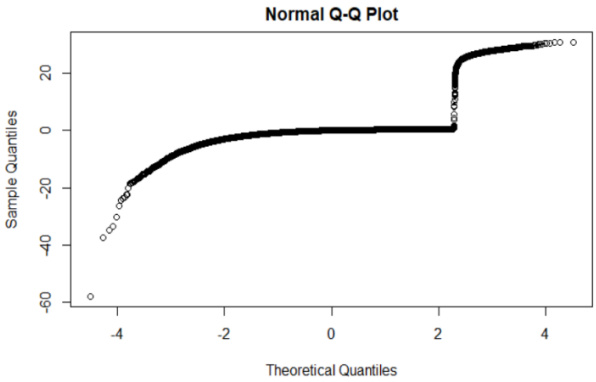
Finally I got 8 models in total. To compare model performance, I introduced ANOVA table and selected our best model with the lowest AIC. Linear Mixed Model 7 that included Page Views, Visit number, Channel Grouping, Browser, and Operating System and isMobile and country has AIC Score 785517. Fitting this model on train set, we got its RMSE as 2.871

```
       npar    AIC    BIC  logLik deviance     Chisq Df Pr(>Chisq)
m_lmm0    3 809332 809362 -404663   809326
m_lmm1    4 786693 786733 -393343   786685 22641.2620  1  < 2.2e-16 ***
m_lmm2    4 786693 786733 -393343   786685     0.0000  0      1.000
m_lmm3    5 786659 786708 -393324   786649    36.3374  1  1.659e-09 ***
m_lmm4   12 786009 786128 -392992   785985   663.7627  7  < 2.2e-16 ***
m_lmm5   16 786016 786176 -392992   785984     0.5542  4      0.968
m_lmm6   21 785774 785983 -392866   785732   252.7945  5  < 2.2e-16 ***
m_lmm7   22 785517 785735 -392736   785473   259.0849  1  < 2.2e-16 ***
---
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: transactionRevenue ~ scale(pageviews) + scale(visitNumber) +
    factor(channelGrouping) + factor(browser) + factor(operatingSystem) +
    factor(isMobile) + factor(country) + (1 | fullVisitorId)
   Data: train_new

REML criterion at convergence: 785570.7

Scaled residuals:
     Min       1Q   Median       3Q      Max
-19.6161  -0.1438   0.0157   0.0736  10.4607

Random effects:
 Groups        Name        Variance Std.Dev.
 fullVisitorId (Intercept) 0.4999   0.707
 Residual                  8.7157   2.952
Number of obs: 155282, groups:  fullVisitorId, 146379

Fixed effects:
                                      Estimate Std. Error t value
(Intercept)                         -0.090701   1.084260  -0.084
scale(pageviews)                     1.168088   0.007936 147.180
scale(visitNumber)                   0.032717   0.009832   3.328
factor(channelGrouping)Affiliates    0.054120   1.073826   0.050
factor(channelGrouping)Direct        0.195447   1.072564   0.182
factor(channelGrouping)Display       0.367189   1.073528   0.342
factor(channelGrouping)Organic Search 0.078779  1.072473   0.073
factor(channelGrouping)Paid Search   0.017192   1.073412   0.016
factor(channelGrouping)Referral      0.523135   1.072616   0.488
factor(channelGrouping)Social        0.267898   1.072622   0.250
factor(browser)Edge                  0.003180   0.069235   0.046
factor(browser)Firefox               0.019565   0.039549   0.495
factor(browser)Internet Explorer     0.051218   0.055063   0.930
factor(browser)Safari               -0.053307   0.031209  -1.708
factor(operatingSystem)iOS           0.054824   0.040805   1.344
factor(operatingSystem)Linux         0.053641   0.164696   0.326
factor(operatingSystem)Macintosh     0.307882   0.161029   1.912
factor(operatingSystem)Windows       0.089339   0.160590   0.556
factor(isMobile)TRUE                -0.008901   0.158647  -0.056
factor(country)1                     0.287621   0.017859  16.106
```

However, predicting this model on test set received a much higher RMSE, 3.45. The results imply a potential overfitting. For residual plot 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.

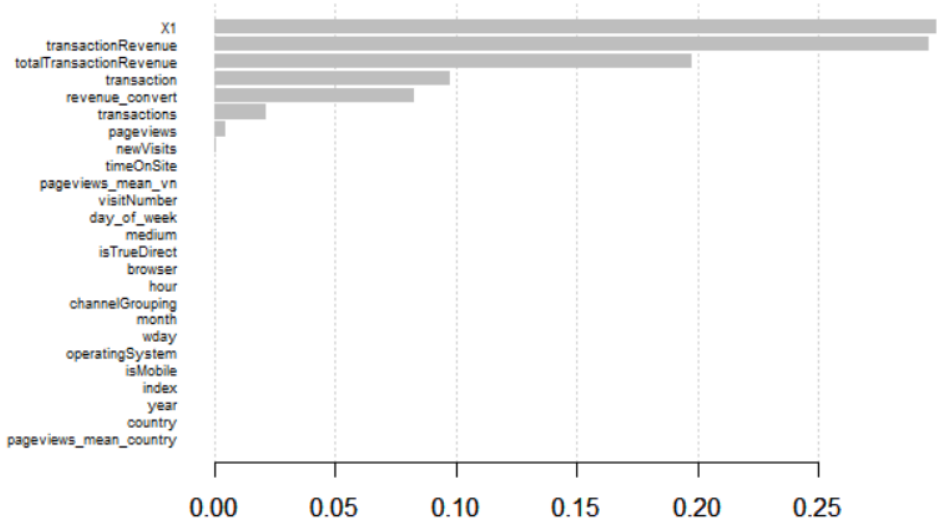




2) XGBoost

The second model I tried to use is a XGBoost regression model. The idea of XGBoost regression method is to use decision tree ensembles. The tree ensemble model consists of a set of classification and regression trees. In other words, this method first classifies inputs and use their average as the output value for regression purpose. As the number of training rounds became more and more, the tree would always adjust their conditions for determining new classification. In this way, a regression model would be built. Using this model on test set, the RMSE is 3.019. This shows that this model is better than linear mixed model.

Also, it can calculate the importance of each predictor:



IV. Result

1. Model choice

According to RMSE on test set, it is clear that XGBoost is better than linear mixed model.

Model	**RMSE**
Linear Mixed Model	3.454
XGBoost Model	3.019

From the importance table we can see that “page views” is the most significant predictor shows on the top of the list, which means it is the most important predictor to predict customer revenue.

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.

V. Discussion

The model I try didn't shown a good result and it might because of the huge amount data that I didn't clean it perfectly. When I do EDA part, I use the whole dataset try to get a overall view but when I move on to the model fitting part, I realized that some of the function cannot deal such amount data and the memory space in R is limited. Then I try to split the dataset randomly, in this procedure it might generate the issue the feature I find before doesn't match with the data fitted with models.
For further thinking of this project, I will try to reduce the dataset at beginning. Besides, since the data is highly imbalanced and there are too many levels of categorical data, we would think further on the improvements on data balancing and dimension reduction.

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-forgstore-glm-keras-xgb
[3]Notes for competition on kaggle.com:
https://www.kaggle.com/erikbruin/googleanalytics-eda-lightgbm-screenshots

VI. Appendix

1-Independent Variables Introduction

After processing train data, we finally obtained 36 variables. All independent variables can be divided into 6 groups: visitor info, visit number, channel, geo networks, devices, and advertisement.

1. Visitor Info: Full Visitor ID, Visitor ID, Visit Number, Visit Start Time, Date
2. Visit Num: Visits, Hits, Pageviews, Bounces, New Visits
3. Channel: Channel Grouping, Campaign, Source, Is True Direct, Referral Path
4. Geo Networks: Continent, Sub Continent, Country, Region, Metro, City, Network Domain
5. Device: Browser, Operating System, Is Mobile, Device Category
6. Advertisement: AD Content, and ADword Click Info Page, ADword Click Info Slot, ADword Click Info Gclid, ADword Click Info AdNetworkType, and ADword Click Info IsVideoAd

Since not all variables have been clearly interpreted, we only analyzed some of variables that potentially might be important features.

Variables	**Data type**	**Description**	**Level**
Date	Quantitative	The date on which a customer visited the store	20160801 - 20180801
Full Visitor Id	Categorical	A unique identifier for each user of the store	Length : 1708337
Visit Number	Quantitative	The session number for the user	1-460
Visit Time	POSIXct	The time on which a customer started to visit the store	ex: “2016-09-02 15:33:05”
Browser	Categorical	The browser category that a visitor using	Main: Chrome; Safari; Internet Explorer; Edge; Firefox; Opera Mini; Safari (in-app)
Operating System	Categorical	The operating system that an user accessing the page	Main: Windows; Android iOS; Macintosh; Chrome OS; Linux
Is Mobile	Binary	An indicator to show whether a customer visted the store from mobile devices	True; False
DeviceCategory	Categorical	The device that the a visitor using	desktop; tablet; mobile
Geo Networks	Categorical	These sections contain information about geography of a visitor, it can be further down into continents, subcontinents, countries,	An example of a visitor from Madrid can be like: Europe, Southern Europe, Spain, Community of Madrid,

		regions, metro and cities.	Madrid.
Source	Categorical	This section contains info from which the session orginated	ex: Google, baidu
Medium	Categorical	An indicator to categorize source path	ex: organic, referral
Hits	Quantitative	The hit times for a visitor	0-500
Page Views	Quantitative	Number of page that a visitor views	0-500
Bounces	Quantitative	A bounce is a single-page session on the site. Bounce rate is defined as the percentage of all sessions on your site in which users viewed only a single page and triggered only a single request to the Analytics server	0,1
Transaction Revenue	Quantitative	The target variable that we need to predict, which is the revenue that a customer made from a single visit.	0-38
Is True Direct	Binary	A binary indicator that displayed whether a visit is direct or through referral path	True; False

Table 1.1: The table shows descriptive statistics of variables in the dataset.

2-Missing Data Analysis

In order to better understand the train dataset, checking the missing value would be important. There are 98.9% users did not make the transaction while looking the product pages. In addition, all advertisement variables have missing rate around 95.6%. This result suggests to remove advertisement variables from the set. Similarly, “keyword”, "isTrueDirect" and “referral path” from channel group also have high missing rate between 61.6% and 68.7%. Also, detailed information on geography of visitors such as “city”,“region” and "metro" miss approximately 54.6%, which implies a better analysis of transaction revenue on country level.

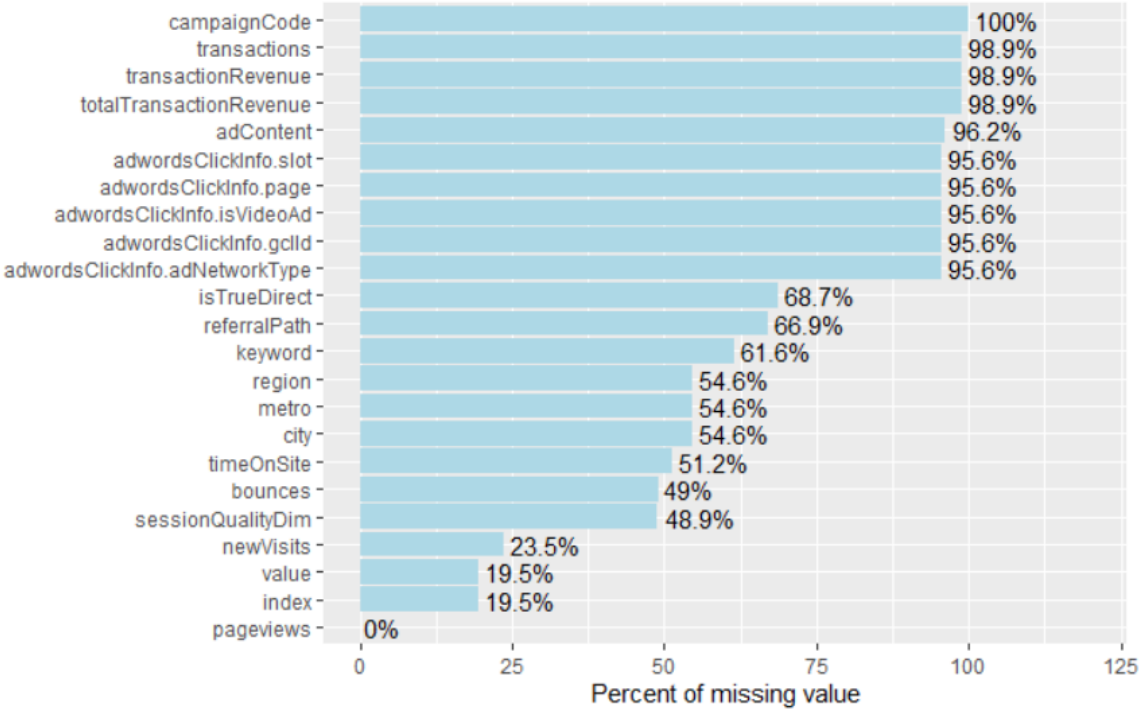


Figure 2.1 The above figure indicates the percentage of missing value, there are 98.9% of users did not make the transaction.

3-Transaction Revenue

For transaction revenue, “NA” value can be treated as 0 since “NA” means no revenue made during the single visit. 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 3.1)

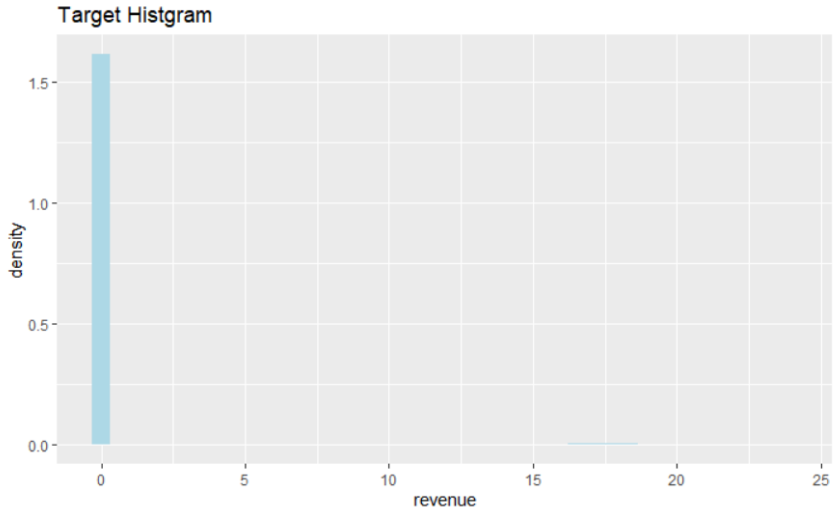


Figure3.1 Histogram of transaction revenue

For nonzero target, we transform the transaction revenue variable into log scale through “log1p” function transforms x to log(1+x) in R. It converts our data to approximately normal distribution with the median 17.7. (Table 3.1) In natural log scale, most transaction revenue follows between 15 and 20. (Figure 3.2)

Variable	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
Transaction Revenue	0.0000	0.0000	0.0000	0.1926	0.0000	23.8644
Log Revenue	9.21	16.95	17.65	17.77	18.42	23.86

Table 3.1 Comparison of transaction revenue and log transaction revenue

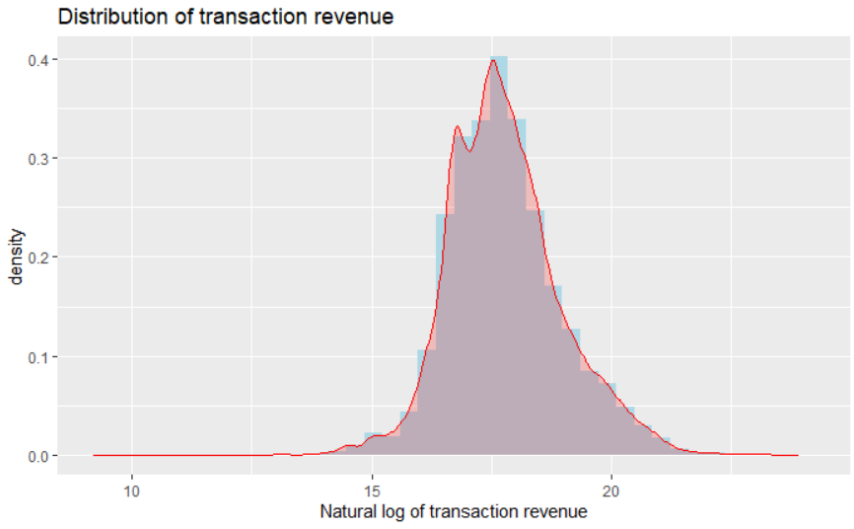


Figure 3.2 distribution of natural log of transaction revenue

4-Device Attributes

Device category contains desktop, mobile, and tablet. See below, we observe that most people prefer to use desktop to visit products and generate transaction revenues. On the other side, customers rarely placed orders on mobile and tablet devices since they are not as userfriendly as computers.

In addition to device category usage, the operating system usage could also be an important factor when predict the revenue transaction. It shows that people use chrome make the most deal. Besides, although most people use windows to visit the system, macintosh users actually makes significantly more revenue transaction. Furthermore, It also confirms that mobile systems are less attractive to customers who plan to purchase products from G Store.

Analysis on browser almostly leads to the same result as that on device and operating system. The main reason is that “Safari” visitors also used macintosh operating system, but “Chrome” outperforms “Windows” from operating system category on transaction revenue. This result let us believe that most Macbook users also Chrome as the dominant browser.

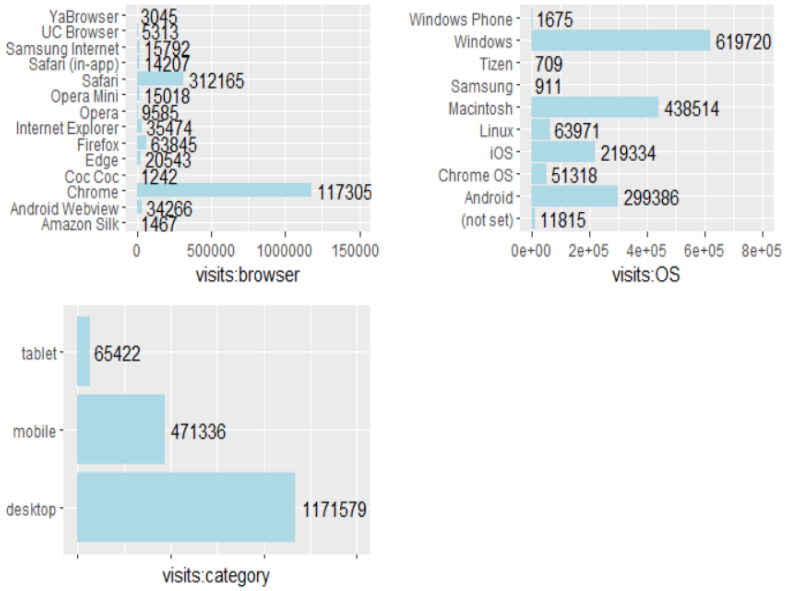


Figure 4.1 visits by browser, device category, operating system

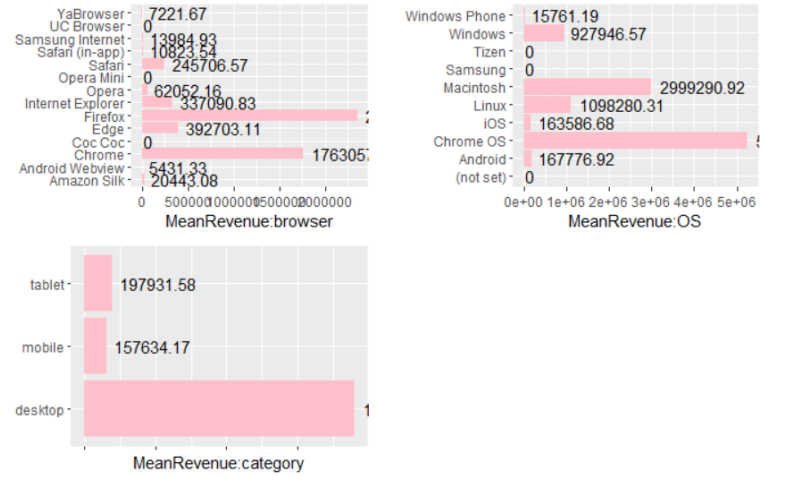


Figure 4.2 mean revenue by browser, device category, operating system

Channels is one of the most important factor feature since it is defined by Google Analytics as how a customer come to the G Store website. It related to other variables such as “source”, “referral path” in the “channel grouping” group (See Appendix 1). Organic Search means that a customer arrive the G Store from a search engine; Social is related to any social media websites or apps; Referral means that a customer come from G Store’s advertisement on other websites. Organic Search is the main access to the G Store and this channel also contribute essential amount of revenue. Social path hardly can generate revenue even though it contributes around half number of sessions than Organic Search. It demonstrates that customer are less interested in placing orders after they redirect from the social media. However, the referral path outperforms any other channels: it generates most transaction revenue. This result indicates that referral path is the most useful channel for selling products from G Store.

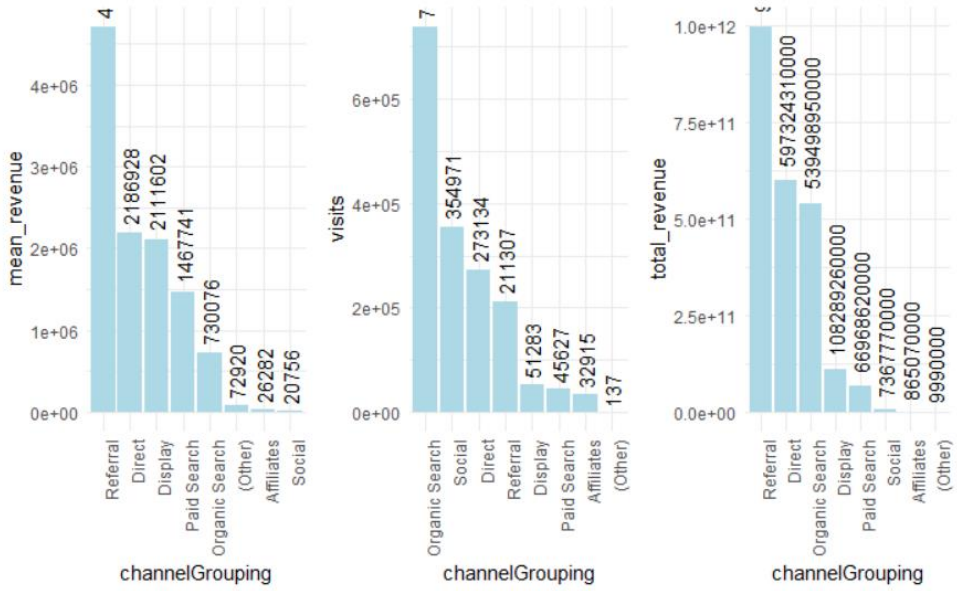


Figure 4.3 mean revenue by channel grouping

We know that the "channelGrouping" variable contain similar information with "is TrueDirect",From the plot, we can see that direct or not affect revenue significantly. Thus, "isTrueDirect" could be consider as a potential predictor.

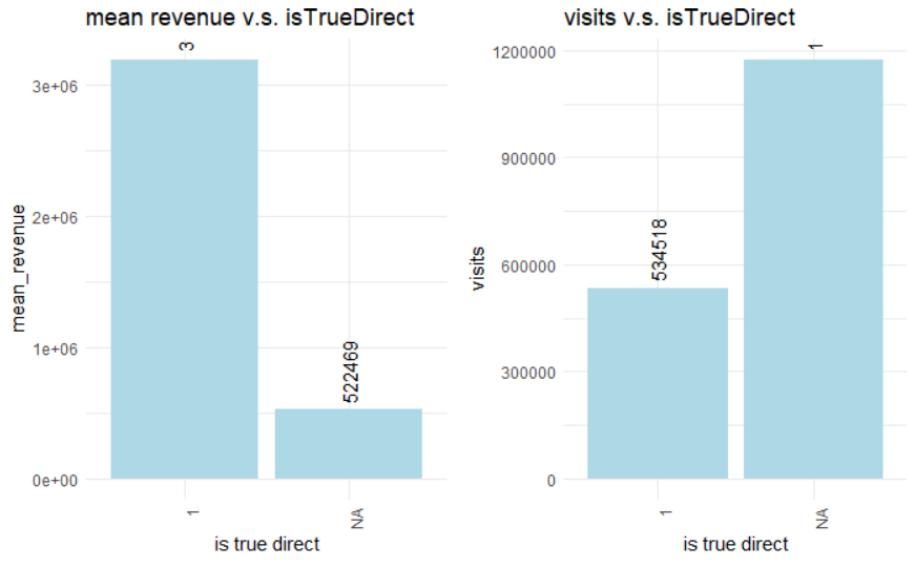


Figure 4.4 mean revenue by is direct or not

5- GeoNetwork Attributes

GeoNetwork conations information about geography of a visitor, it can be further down into continents, subcontinents, countries, regions, metro and cities.(Appendix 1) It shows that where are the majority of customers from, group by continent we see that most purchase occur in Americas and outside America a small amount purchase occur in Africa and Oceania. Besides, the visits information shows that although some session occur in Asia and Europe, no such amount purchase action happened in the end.

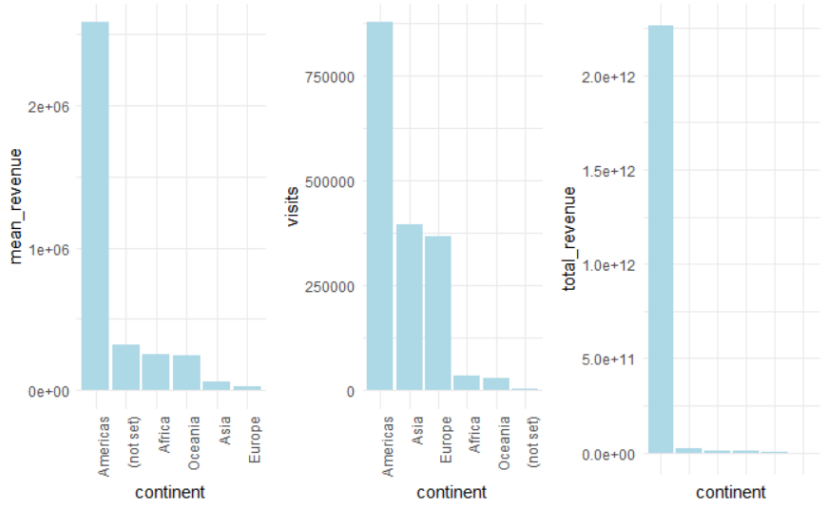


Figure 5.1 mean revenue by continent

Also, group by countries, United States has the largest total revenue and visits. But for average revenue, Anguilla takes the maximum proportion.

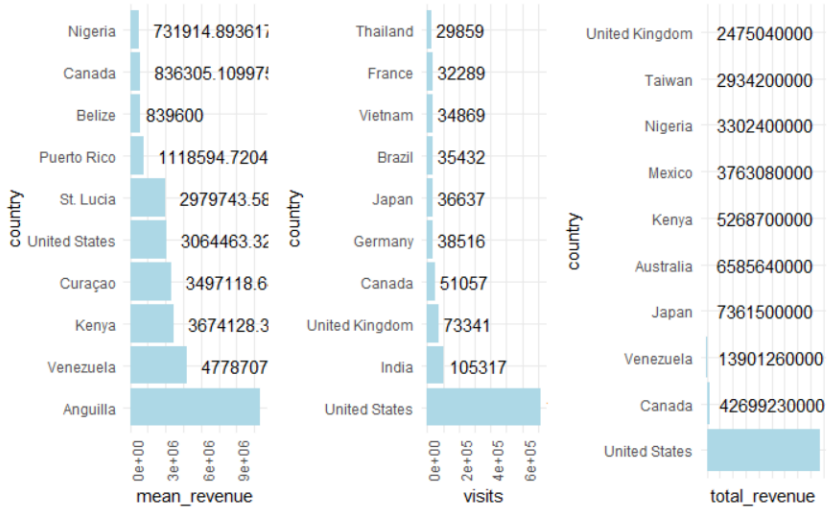


Figure 5.2 mean revenue by country

Following are world map of visit and revenue by countries.

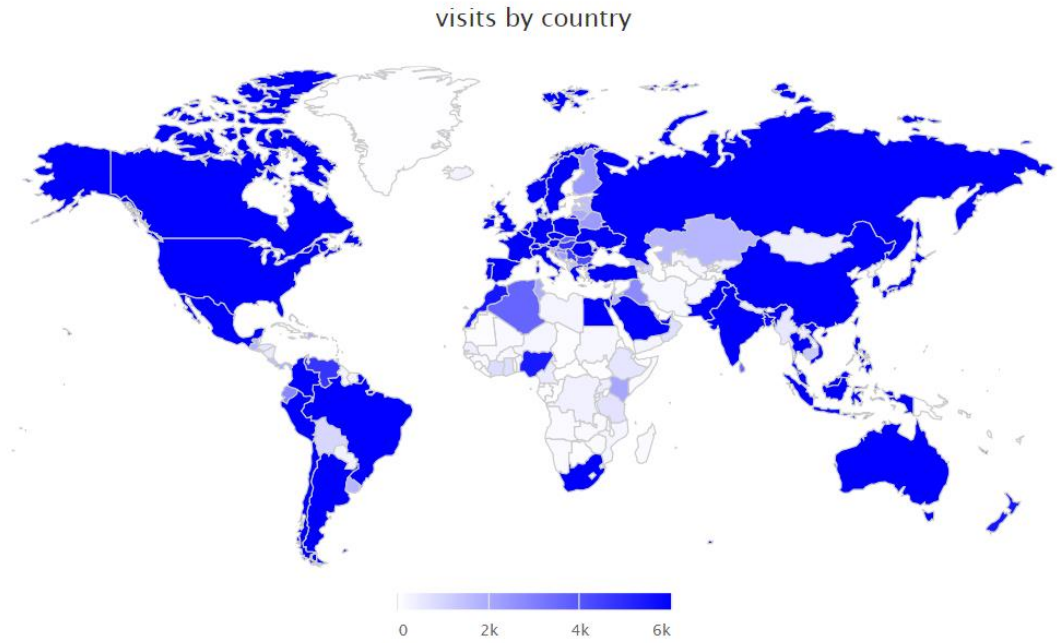


Figure 5.3 map of visits by countries

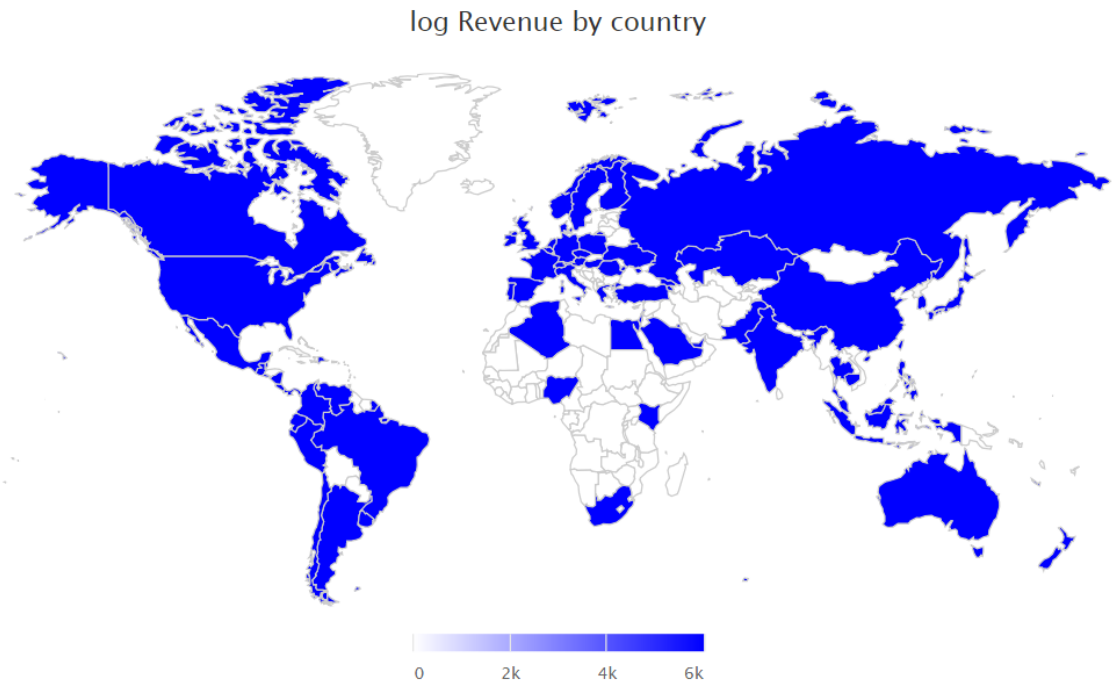


Figure 5.4 map of log revenue by countries

6- Visits, Hits, and Page Views

Figure 6.1 shows the trend of daily visit between 2016-2018.The trend is roughly smoothy, except the date from October 2016 to January 2017, there is a obvious convex curve. At the end of 2016, between November 2016 to December 2016, the daily visits increases significantly. Thus, we believe there were more people visits the product at the end of 2016. And in the end of 2017, there is a abrupt peak occured, which could be some external reason.

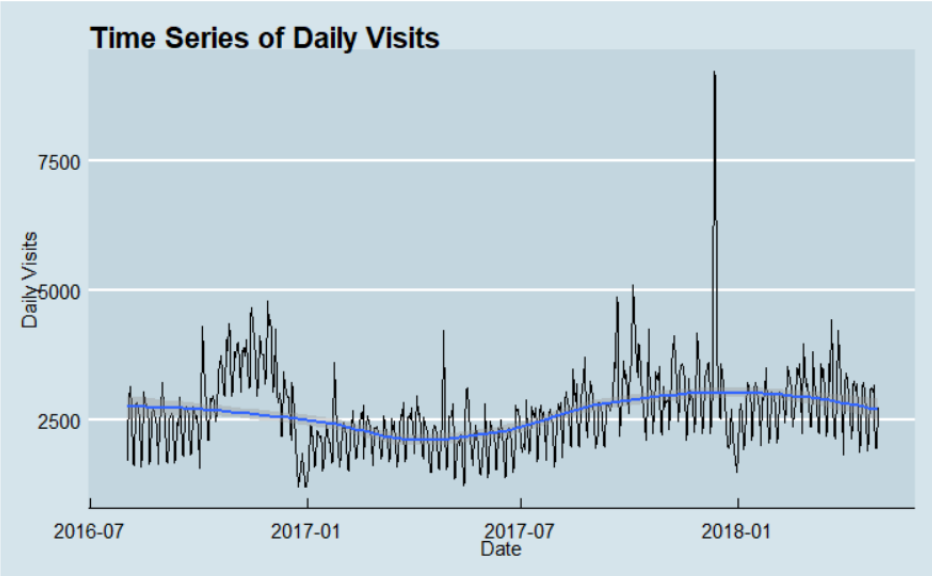


Figure 6.1

Figure 6.2 shows the trend of daily hits, from 2016 to 2018. The trend decreases significantly from 2016 to 2017 and rise a bit when turns to 2018. Therefore, I believe there are less and less customers that can be attracted and that could because of other competitors.

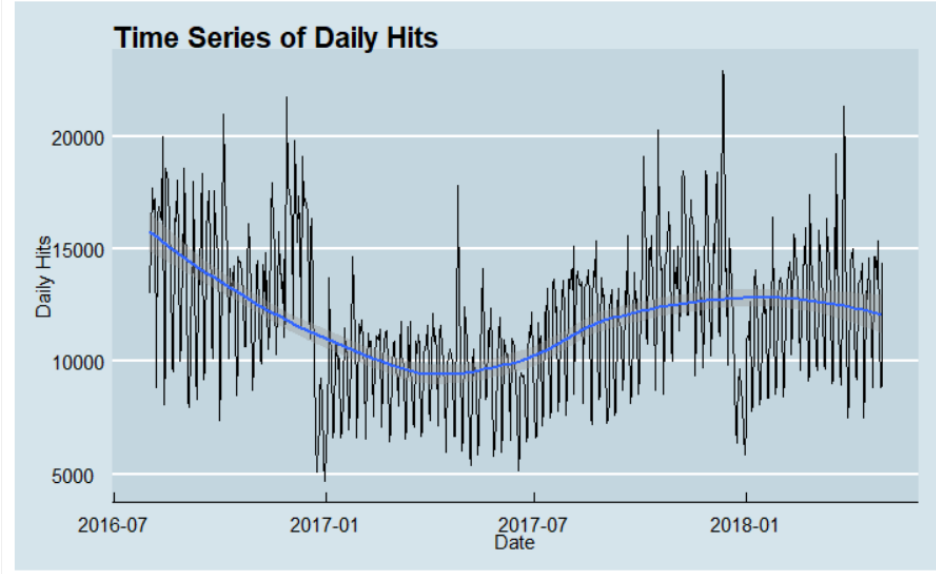


Figure 6.2

Figure 6.3 shows the trend of daily new visit between 2016-2018, it's basically the same with visits.

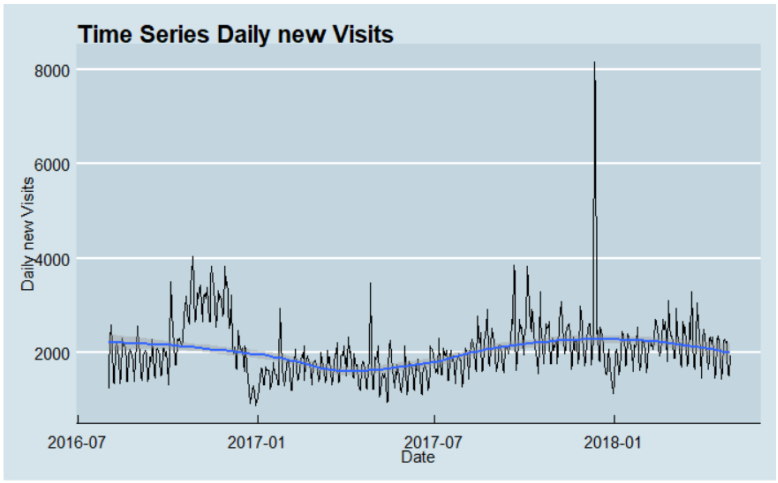


Figure 6.3

Figure 6.4 shows the trend of pageview from 2016 to 2018. Same as daily hits.

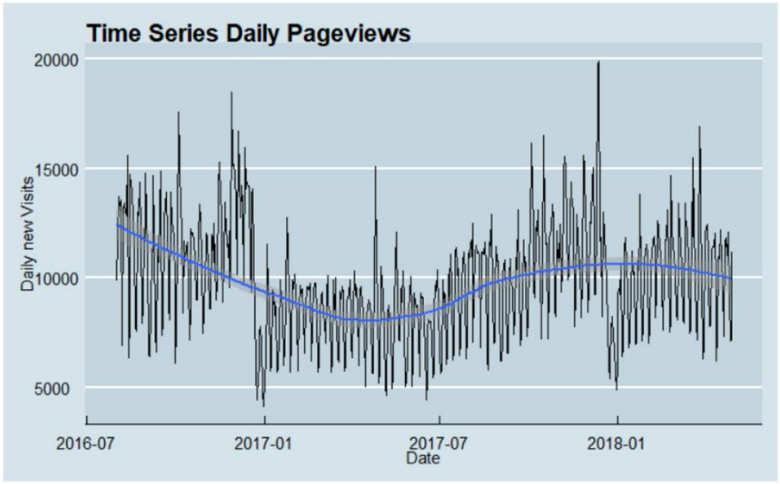


Figure 6.4

See Figure 6.5,although the daily transaction data are very volatile, it seems that a “high and low” pattern is regular over the year. Using the “smooth” function from ggplot2 in R, we observe a relatively stable

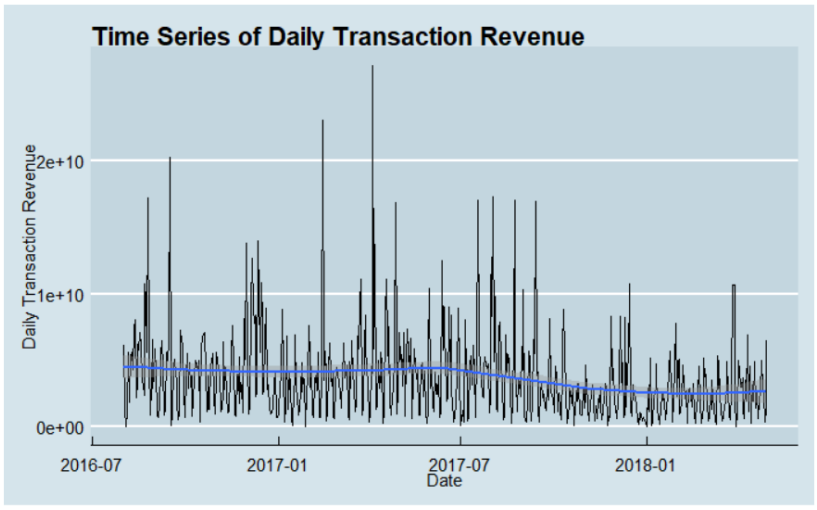


Figure 6.5

From this plot, It is evident that “visit number” has great impact on revenue.

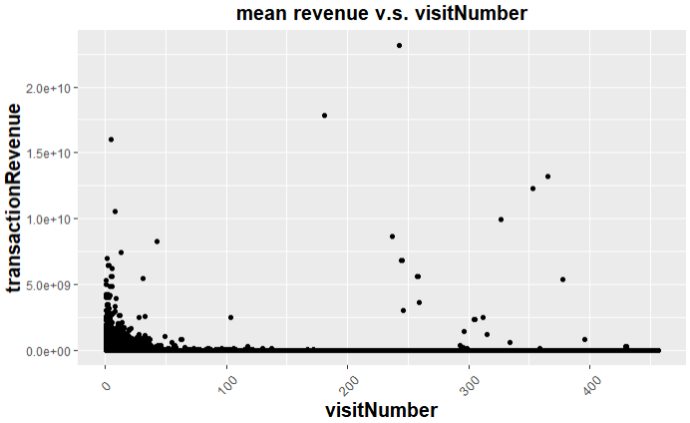


Figure 6.6

The folllowing figure shows that the pageviews of customers. The right tail distribution indicates that most people view the product one to ten times.

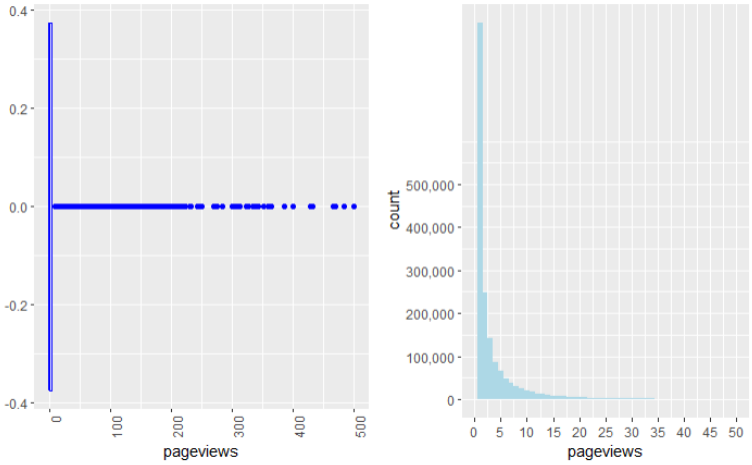


Figure 6.7

These two figure below shows the relationship between page views and revenue transaction. The histograms shows that people who view the page more than 10 times highly likely to make the transaction. If we only count pageviews with any transaction revenue, it leads to the same distribution.

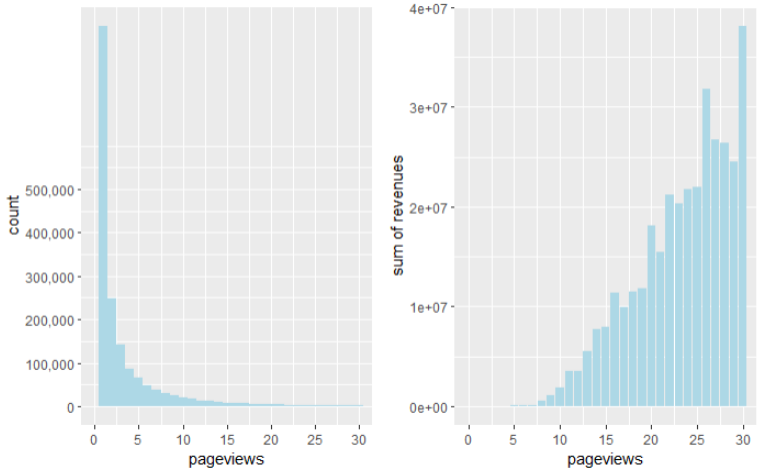


Figure 6.8

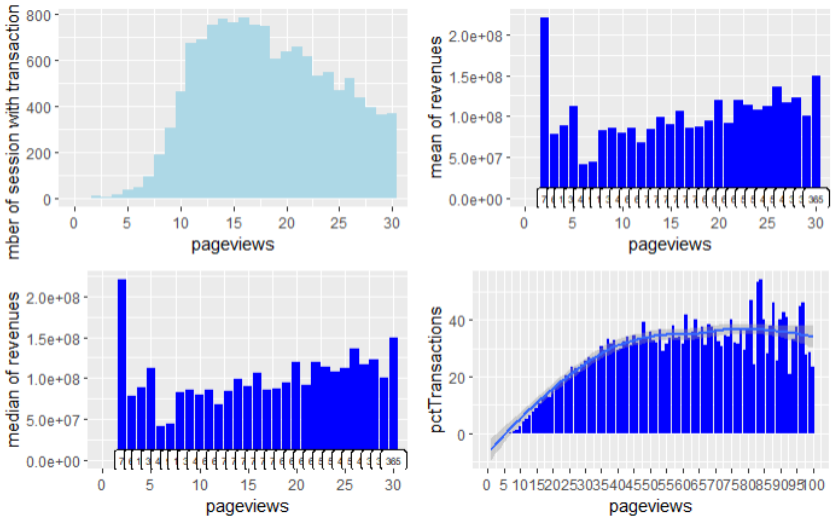


Figure 6.9

We calculated the correlation matrix(Figure 6.10). It shows that “hits” and “page views” have correlation of 0.98 which is near to 1, thus I select “page views” instead of hits. Since there are other number is significant, there may be collinearity.

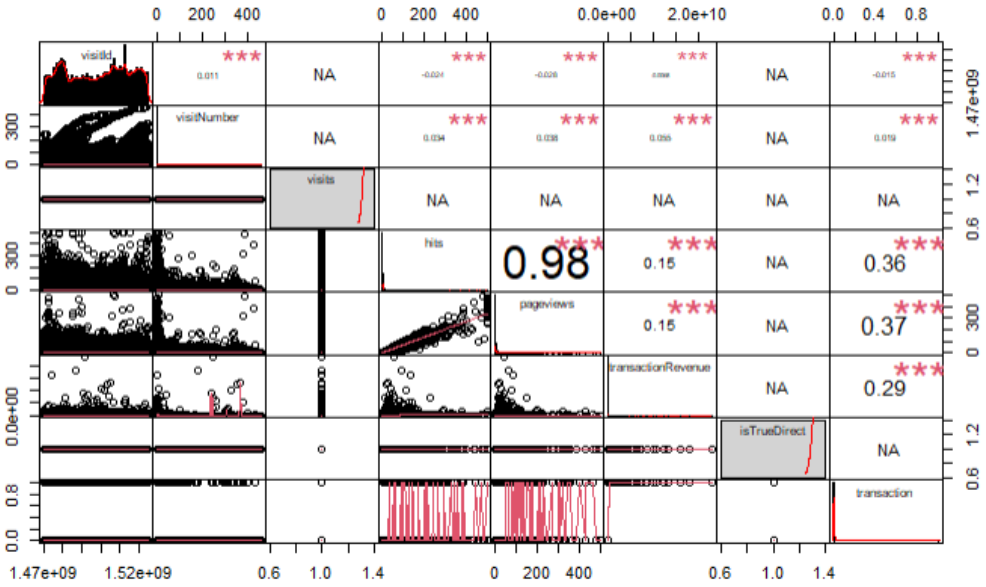


Figure 6.10