Google GStore Customer Transaction Analysis

Duo Zhang, Meixing Dong, Yu Tian, Zishen Li

1. Summary

Inspired by the prevalence of e-commerce and successful implementation of 80/20 rule[1] on manifold businesses, exploration of online customer behavior attracts increasing attention. Google Analytics Customer Revenue Prediction[2], is one of those analyses aiming to obtain business insights and predict revenue per customer for Google Merchandise Store(GStore). In this research, we explored the relationship between features transaction revenue and built several prediction models. We found the main source of revenue is the United States, and desktop and chrome are the most frequently used device and browser. We also implemented logistic regression and linear regression model to predict logarithm transaction revenue for each record and analyzed significant feature for each feature.

1.1 Data introduction

The GStore dataset has 903,653 visit records and 12 features including a different type of format such as numeric, character and JSON (4 variables). The features contain identity, channel, geographical, transaction and other information for each visit to Gstore from July 2016 through August 2017 (Appendix I). After tidy the data, we got 55 features. The details can be found in table.2. The target variable is the revenue (revenue, transaction, and sales used in our report refer to the target variable), while the competition calls participants to predict log revenue for each customer. Thus, all the graphs in our report except special annotations plot the log revenue when revenue is mentioned.

1.2 Workflow

To tidy the data, we first flattened JSON formatted data as well as removed columns with constant value. EDA is conducted, which includes exploration on missing value, relationship among channel-related, time-related and geographical features, as well as the user behaviors. According to the EDA, we built a classification model to classify users who will buy or not and employed linear regression model to predict the logarithm revenue for each record.

2. Methods

2.1 Packages

The packages we used are listed below (Table.1).

ggplot2	tidyverse	dplyr	stringr	jsonlite	tidyr	lubridate	maps

Table.1 Packages used in the report

2.2 Data tidy

2.1.1 Parse the JSON format

Taking advantage of package ‘jsonlite’ and ‘stringr’, we created a function to take care of the parsing process. The function basically transforms the input string into a usable format for method ‘fromJSON’ which can automatically parse the string to separate columns. Last, we combined all columns together and from a dataframe in RStudio.

2.1.2 Data transformation

According to the request of the competition, we first changed the data type of fullvisitorId from numeric to character in order to have a unique visitor ID for each customer. Second, we converted date to date format visitId, channelGrouping, sessionId to character, hits, pageviews, bounces and newVisits to numeric, which are their natural representation. Last, we noticed that many columns only have one distinct value which is useless for data analysis and modeling. 20 columns were removed by this step (see Appendix II).

2.3 Data exploration

2.3.1 Missing value

To calculate the missing value rate for each feature, we transformed all kinds of missing values, such as ‘not available in demo dataset’, ‘(not provided)’, ‘(not set)’, ‘<NA>’, ‘unknown.unknown’, ‘(none)’, into NA, Using visualization and transformation methods provided by package ‘dplyr’ and ‘ggplot’, such as ‘mutate’, ‘rename’, ‘sapply’, ‘geom_col’, ‘coord_flip’, we got a plot showing missing value rate for each feature.

2.3.2 Distribution

We plotted the distribution of log revenue facilitated by ‘geom_histgram’. This method will remove all NAs without showing in the graph. We also plotted another graph where the x-axis is the index of the revenue and y-axis is the value of revenue.

2.3.3 Relationship exploration

In this section, we will mainly discuss these four most related features that channel, geographical, time, user behaviors) and the relationship between them and revenue or visits. The methods used for the plot are ‘geom_col’, ‘geom_polygon’, ‘facet_wrap’, ‘borders’, ‘geom_line’, ‘geom_smooth’, ‘geom_point’.

For channel related features, we focused on channelGrouping, deviceCategroy, and browser. (Appendix III) These three variables can reflect how people reach GStore. We plotted the revenue and visits with each of these three variables individually by adopting the ‘group_by’ method.

For geographical features, we explored the distribution of revenue, visits, page views(total and average), hits(total and average), devices, and browsers. To visualize the distribution more directly, we imported the package ‘maps’ and illustrated the distribution by color. Since the United States was the main source of the visits and revenue, we also plotted the top 10 cities with most visits and revenue on the United States map by colors and sizes.

Since the revenue and visits can be viewed on a time series, we analyzed the yearly, monthly and weekly patterns of these two variables. Furthermore, we included the analysis of visits with different kinds of devices over time in the report as the recent trend of mobile devices becomes more popular. The transformation methods we used are ‘mday’, ‘wday’.

Last, we plotted the average page views and hits against revenue and visits with the method 'geom_point'. By it, we could find the possible related features behind the purchase behaviors with the non-purchase records. The different purchase behaviors are separated by colors.

2.4 Modeling

In this section, we will discuss the strategies we used to formulate the prediction problem and the utilized models. There are three main challenges to build up our model: imbalanced data, high dimension, and missing value.

From our previous exploratory analysis, the data was found to be highly imbalanced. It means most of the user visit records do not have any purchase behaviors and the total transaction revenue is zero. If we use the regression model directly, the parameters of the regression model will be dominant by records without any purchases. As a result, we decided to use a classification model first to predict whether a certain user visit record would lead to a purchase behavior or not. If the result predicted by classification model is positive, the record will go through a regression model which is used for predicting non-zero transactions. We chose the linear regression model and logistic regression model because we concerned with not only the model accuracy but also the explanation of the model.

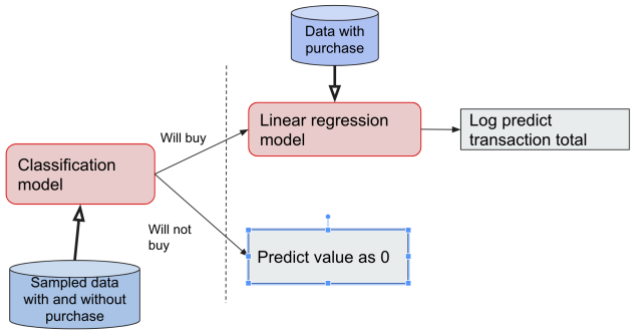


Fig.1 model strategy

To prepare the dataset for model training and testing, we did several preprocessing. The missing value plot (Appendix V Fig.3) shows that there are 19 variables with more than 12% missing values, and most of them do not provide valid information for model prediction. After removing high missing rate variables, we removed records with the missing value which is about 1% of our total data. Besides, we manually removed some of the non-valid variables such as vistId and sessionId. The duplicated information also get removed such as continent, which is covered in subcontinent.

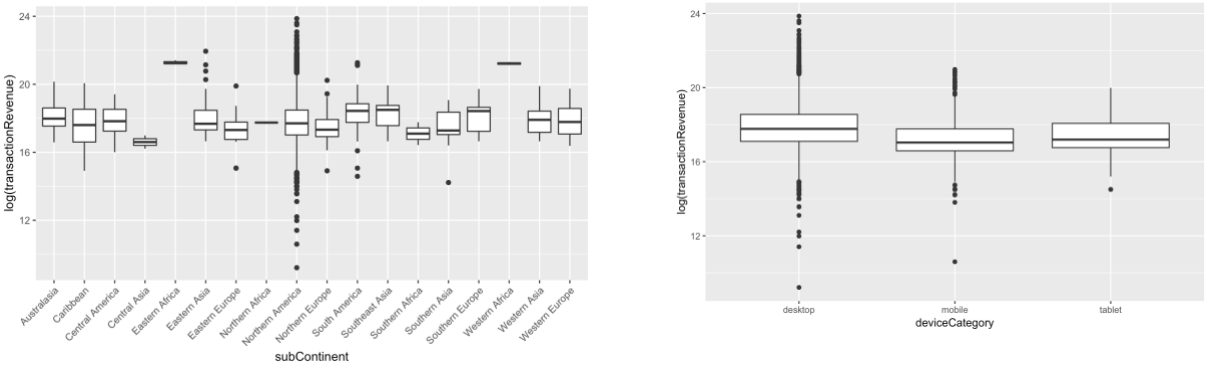
2.4.1 Classification

The objective of the classification model is to predict whether each user behavior record will result in a purchase behavior. We mutated a new column call buy_or_not, which was generated based on total transaction revenue: if the transaction is 0, the buy_or_not value is false and if the transaction is greater than 0, the value for buy_or_not is true. As we mentioned before, the data is highly imbalanced and 98% of total data is without buying behavior. Since we have enough number of data, we took down sampling method, which is sampling from the majority class, to eliminate the side effect of imbalanced data. We split the dataset into a training set with

80% and a test set with 20%. We also set the threshold for predicting positive response as 0.3 instead of 0.5. Since we have another regression dealing with positive predictions, it is acceptable that we predict more users who actually do not purchase as positive instead of predicting users who made the transaction as negative.

2.4.2 Regression

For the records which are predicted as positive for buying behavior, the regression tries to predict the logarithm total transaction revenue. For building up the regression model, we first plotted the relationship between response and each variable to check if there is a systematic relation between them. Following graphs (Fig. 2) show examples of the relationship between response and subcontinent, deviceCategory, respectively.



(a) **(b)**

Fig. 2 Relationship between log revenue and subContinet (a) and deviceCategory(b)

We can observe there exists an obvious difference in the relationship between logarithm across different subContinent. However, we did not see the obvious differences between response across the different device, so we removed the device category variable from our regression model. Based on above criteria, we selected 10 variables and add them into our linear regression model.

3. Results

3.1 Data tidy

After tidy the data, we got 35 features, 29 of which are categorical variables and 6 of which are continuous variables. Details of the dataset are shown below. (Appendix IV)

3.2 Data exploration

3.2.1 Distribution and missing value

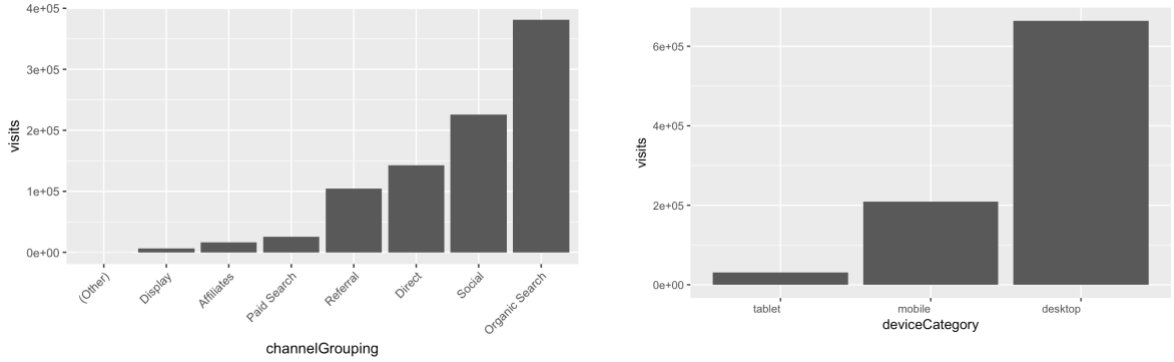
There are 15 variables have more than 50% missing value rate(see Appendix V Fig.3). These incomplete records will be dealt with extensive care in the modeling process.

The distribution of log revenue is approximately normal distributed with a little right skew(see Appendix V Fig.4). As the 'ggplot' will drop the NAs when plotting, it (Appendix V Fig.4) only illustrates the data with purchase. Since we will model the purchase behavior before predicting the value of revenue, it (Appendix V Fig.4) still provides efficient information about the revenue. Appendix V Fig.5 demonstrates the whole distribution of revenue. Combined with the calculations, we discovered that 98.7% of the revenue is missing. This phenomenon is

consistent with our intuition because most of the users will visit the website multiple times before making a purchase.

3.2.2 Channel related features

In this part, we pick some features, channelGrouping, deviceCategory and browser, to see their individual distributions and then find the relationship between these features and log revenue. As for the individual distributions, we found Organic Search and Social are the two most frequent channels. Desktop and Mobile are the two most frequent used devices. And Chrome and Safari are the two most popular browsers (Fig. 3, Appendix V Fig.16-19).



(a) (b)

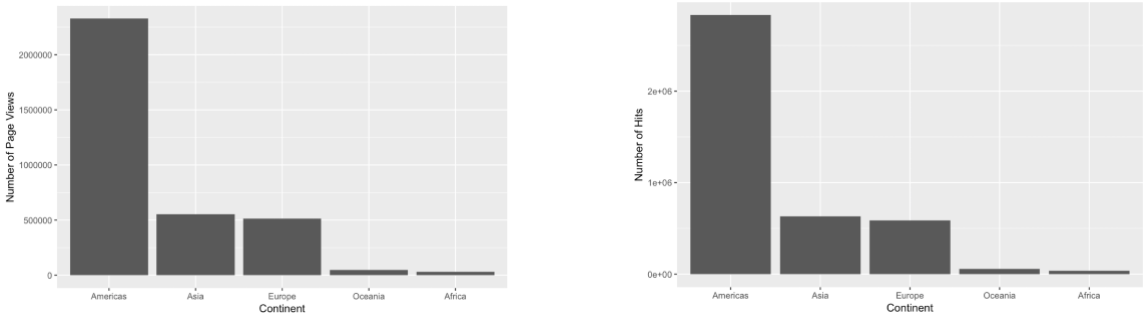
Fig.3 Visits with different channelGrouping (a), devices (b)

After, we explored the relationship between these features and log revenue. For the channelGrouping, Direct and Referral contribute to the revenue most though the Organic Search and Social are the most popular channels. For devices, Desktop and Mobile contribute to the revenue most. Also, users from Chrome produce the highest log revenue. However, Safari contributes less than Firefox while Safari is more popular than Firefox (Appendix V Fig.16-19).

3.2.3 Geographical features

In this section, we analyzed how the site visits, page views & hits revenue are distributed and what browsers and devices are the most popular globally.

It is found that the Americas had the highest number of pages views and site hits among all five continents (Fig. 4, Appendix V Fig.1-2). However, Oceania's rank of the average page views and hits is increased from top four to top two instead (Fig. 4, Appendix V Fig.1-2). It implies that Oceania has a low population compared to other continents even though its average page views and hits are top two.



(a) (b)

Fig.4 (a) Number of Page Views over Continent (b) Number of Hits over Continent

Since the Americas has the highest number of network traffic, we took a look at sites visits in Americas specifically. According to the data we analyzed, the United States had the highest number of visits all over the world.The top ten cities in the USA hold the most visits can be seen in Appendix V Fig.13. However, the sales distribution shows the revenue is distributed differently than the visits ranking. See Appendix V Fig.15 to find the top ten cities with the most revenue in the United State.

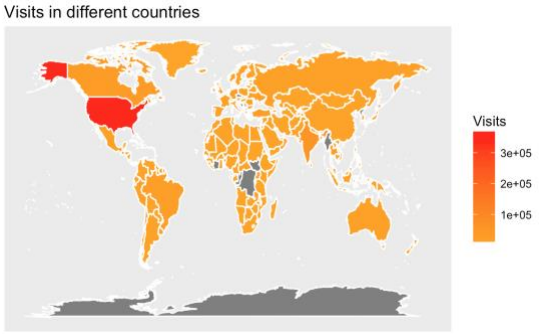
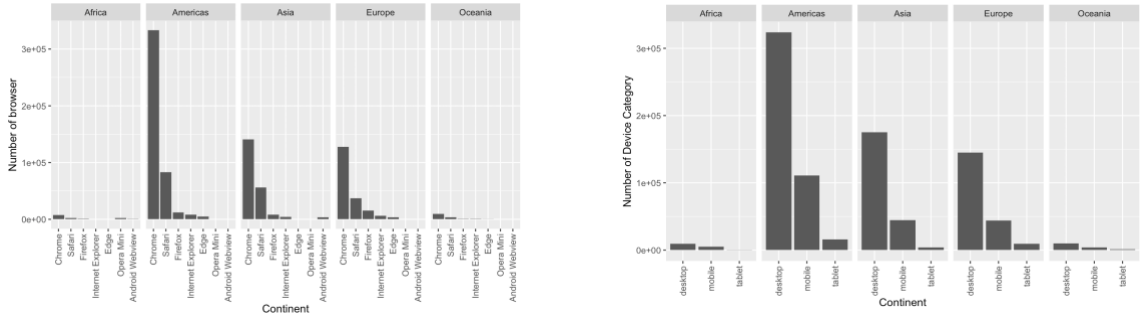


Fig.5 Visits distribution worldwide

Next, we tried to explore the data of browsers and devices distribution in different continents. According to the chart below, the top seven most popular browsers are Chrome, Safari, Firefox, IE, Edge, Opera Mini, and Android Webview, and Chrome is the most popular browser among all the continents. For the devices part, it is found the users prefer to use Desktop the most and tablet the least (Fig. 6).



(a) (b)

Fig.6 (a) Number of browsers over Continent (b) Number of Device Category over Continent

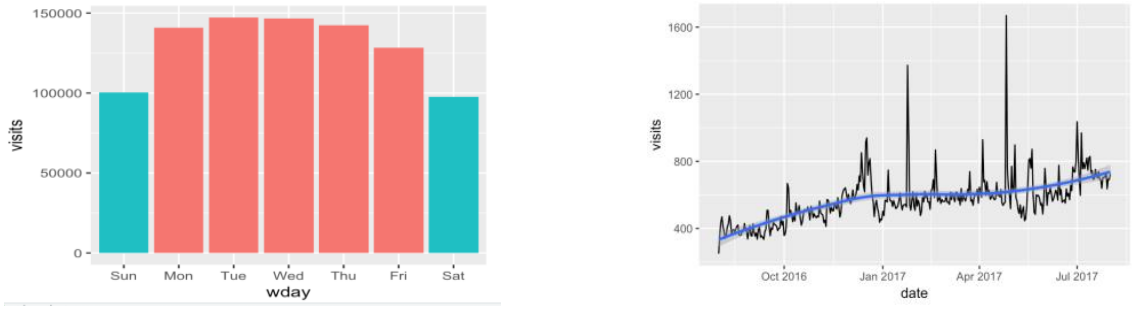
3.2.4 Time-related features

We analyzed these three fields in this part: trends of user accesses and transactions, weekly and monthly behaviors, visits by different devices over periods.

We found there is a visit peak around November 2016. However, there exists an opposite pattern during the same period in the trend of revenue. We supposed this may be due to the delay between customers' visits and transactions (Appendix V Fig.6-7).

Surprisingly, we found people visit Gstore more often on weekdays and generates more revenue on weekdays than weekends (Fig. 7(a)). We speculated the reason was that these transactions may come from the company's behavior. Then, we found the monthly trends of the number of visits and revenue are opposite. The trend of visits is decreasing while the trend of revenue is increasing (Appendix V Fig.9-10). We supposed this may be also due to the delay

between visits and transactions. These two features can help build the logistic regression model in the model section.



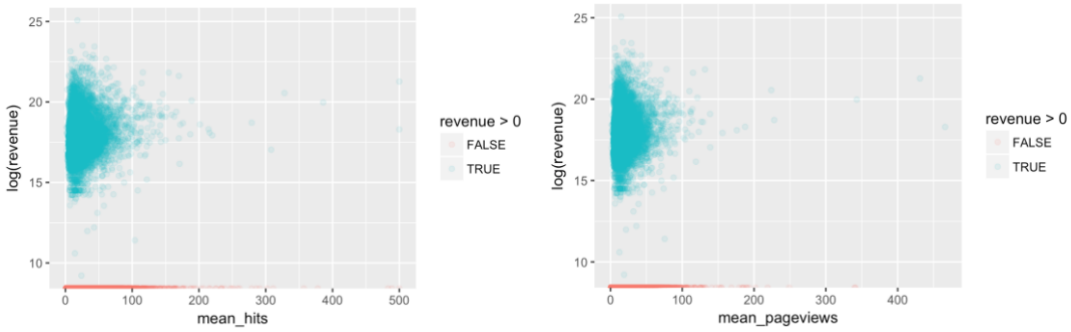
(a) **(b)**

Fig.7 (a) Visits over week (b) Visits from mobile

The visits from mobile and tablet are increasing while that of desktop is decreasing(Fig. 7(b), Appendix V Fig.11). Also, we noticed that mobile devices are becoming more and more popular. Thus, we suggest the company could pay attention to attract more visits from mobile devices.

3.2.5 User behaviors

In this part, we focus on average hits and average page views. We separated the data according to if the hits or page views created revenue or not. Finally, we found that the average page views and hits for users who buy or not are quite different. We included these features in the classification model in the next modeling part (Fig.8).



(a) **(b)**

Fig.8 (a) Revenue over mean_hits (b) Revenue over mean_pageviews

3.3 Modeling

For the classification model, we finally achieved 95.6% accuracy and 98% recall on the test set. According to the summary of the logistic regression model, the number of hits and the number of page view are statistically significant, which is consistent with our exploratory analysis. The continent of America also stands out in terms of the significance of the feature. For regression, we get 1.097742 root mean square error (RMSE) and 1.112894 RMSE on the test set, which means we do not overfit on our training dataset. We also implemented residual analysis using QQ plot shown as following (Appendix V Fig.20), which indicates the residuals basically follow normal distribution and are consistent with our model assumption. For feature significance analysis, we found that Africa's number of hits and number of pageview are the most significant feature in terms of the p-value.

4. Discussion

4.1 Conclusion

On the data exploration side, we found that more page views and hits bring more revenue which is consistent with the worldwide distribution of revenue where the United States contributes most to the GStore revenue. Inside of United States, metropolis generates more revenue than countryside leading to a possible advertising strategy focusing on big cities. From the time view, there is a delay between customers' visits and transactions both monthly and yearly, which indicates multiple visits before making the purchase. Additionally, the weekly pattern shows that people tend to visit GStore and make purchases during the weekdays. It implies the customer purchase behavior. All the patterns we discovered with time could help create an accurate advertising plan. Moreover, mobile devices could be a potential revenue point as user habits are transferring from desktop to mobile device, which is also confirmed by our research. Since most visits are coming from organic search, Gstore also needs to increase its brand awareness.On the model side, we found that hits, page views, and visitnumber are useful on both classification and regression model.

4.2 Future work

To confirm the assumption that increasing mobile device usage could help improve the revenue, we should check the page views, hits and revenue in different devices over the period. Under the 80/20 rule, we also want to check if there is a potential to increase the page views and hits as well as the revenue from other areas instead of just focusing on the main source of the revenue. To provide a solid development plan, we should dig deeper and understand the whole market more comprehensively.

We did not include regularization in the linear regression model. For the next step, we could explore LASSO and Ridge Regression to produce better results along with a better understanding of the importance of each feature. As the revenue can be viewed as a time series, we consider include time as a feature and implement the time series model in the future.

5. Statement of contribution

Duo Zhang: Data exploration (channel related, geographic features and user behavior)
Meixing Dong: Data exploration (channel related, geographic features)
Yu Tian: Classification model and regression model building
Zishen Li: Data tidy, data exploration (time-related and geographic features, user behavior, missing value and distribution)

6. Reference

[1] Kaggle, https://www.kaggle.com/c/ga-customer-revenue-prediction
[2] Pareto principle, https://en.wikipedia.org/wiki/Pareto_principle

7. Appendix

I.

Column names	Instruction

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
sessionId	An unique identifier for this visit to the store
socialEngagementType	Engagement type, either 'Socially Engaged' or 'Not Socially Engaged'
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)

II.

socialEngagementType	browserVersion	browserSize	operatingSystemVersion
mobileDeviceBranding	mobileDeviceModel	mobileInputSelector	mobileDeviceInfo
mobileDeviceMarketingName	flashVersion	language	screenColors
screenResolution	cityId	latitude	longitude
networkLocation	visits	campaignCode	adwordsClickInfo.criteriaParameters

III. Explanation of different levels in channelGrouping

Organic Search	Any number of traffic sources as long as the medium of the traffic sources is 'organic'.
Referral	Any number of traffic sources as long as the medium of the traffic sources is 'referral'.
Paid Search	Any number of traffic sources as long as the medium of the traffic sources is 'CPC', 'ppc' or 'paidsearch' and 'Ad Distribution Network' does not matches 'content'.

Affiliates	Any number of traffic sources as long as the medium of the traffic sources is ‘email’.
Direct	Any number of traffic sources as long as the traffic sources are unknown to Google Analytics.
Display	Any number of traffic sources as long as the medium of the traffic sources is ‘display’, ‘cpm’ or ‘banner’ and ‘Ad Distribution Network’ matches ‘content’.
Social	Any number of traffic sources as long as the medium of the traffic sources is ‘social’, ‘social media’, ‘social-media’, ‘social network’, or ‘social-network’.
(Other)	All those traffic sources whose medium is: (1) Not pre-defined but is in fact defined by a user via custom tracking parameters ‘utm_medium’ (2) (not set).

IV.Table.Variables in the dataset after data tidy

Categorical	channelGrouping	date	fullVisitorId	sessionId
	visitId	browser	operatingSystem	isMobile
	deviceCategory	continent	subContinent	country
	region	metro	city	networkDomain
	campaign	source	medium	keyword
	isTrueDirect	referralPath	adContent	adwordsClickInfo.slot
	adwordsClickInfo.gclId	adwordsClickInfo.adNetworkType	adwordsClickInfo.isVideoAd	adwordsClickInfo.page
Continuous	visitNumber, visitStartTime, hits, pageviews, bounces, newVisits			

V.

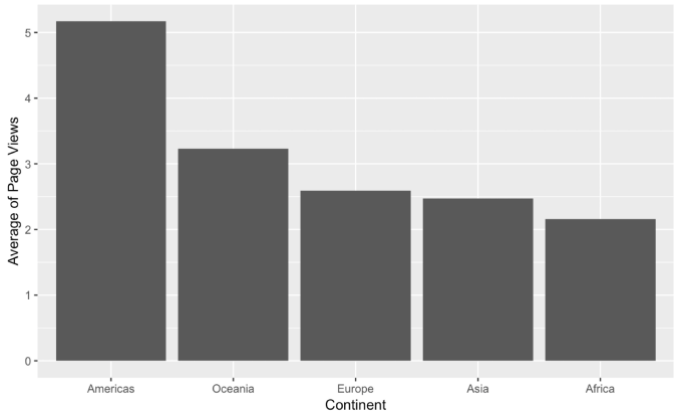


Fig.1 Average page views over different continents

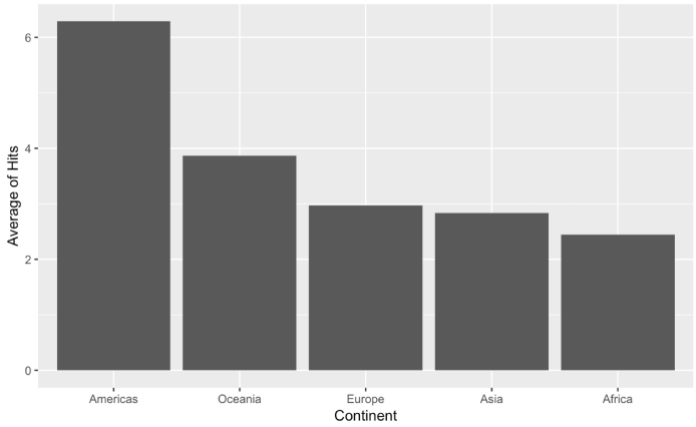


Fig.2 Average hits over different continents

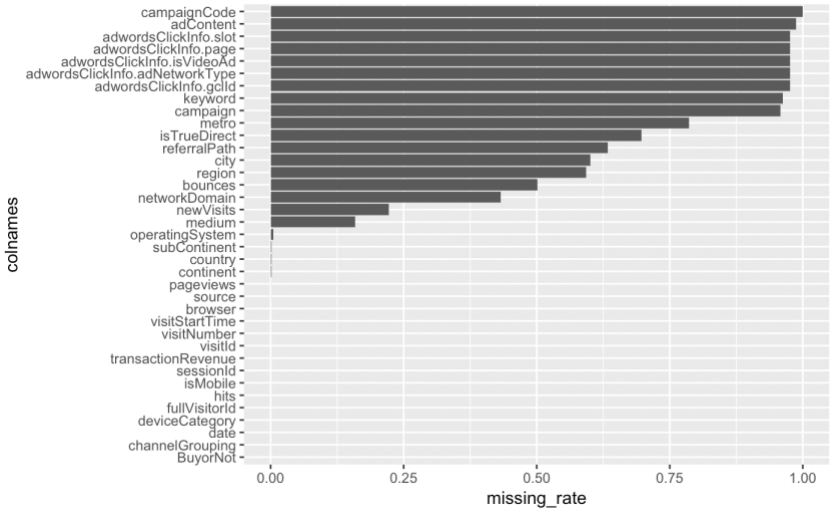


Fig.3 Missing value rate for each features

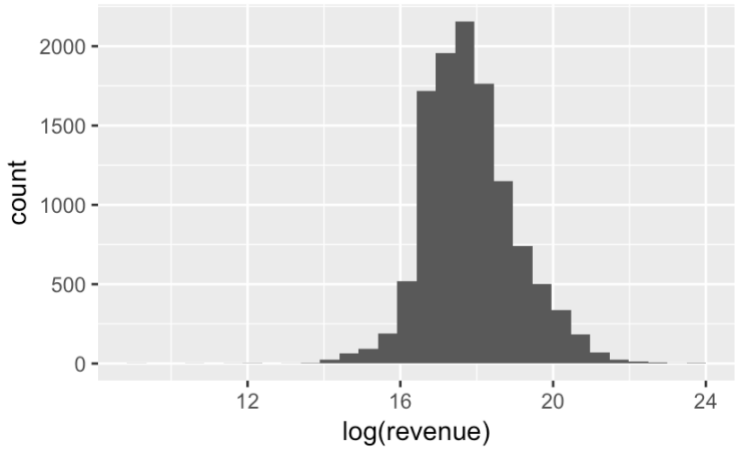


Fig.4 Distribution of log revenue with purchase

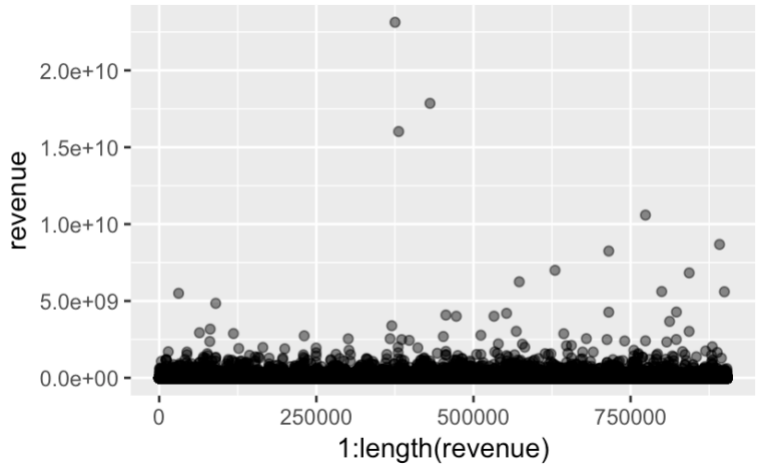


Fig.5 Distribution of all revenue records

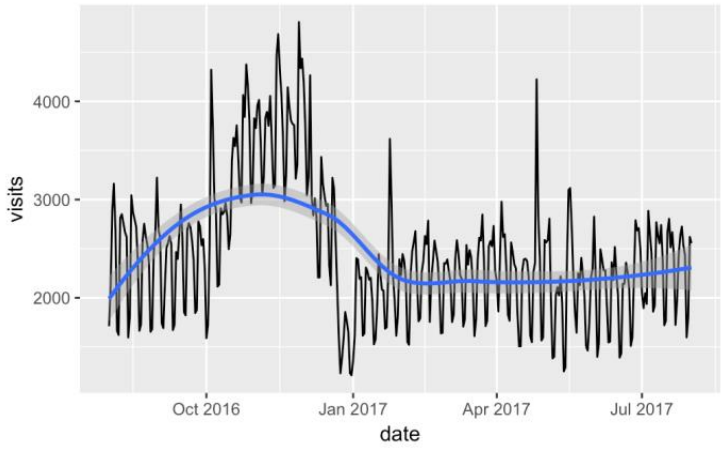


Fig.6 Visits over time

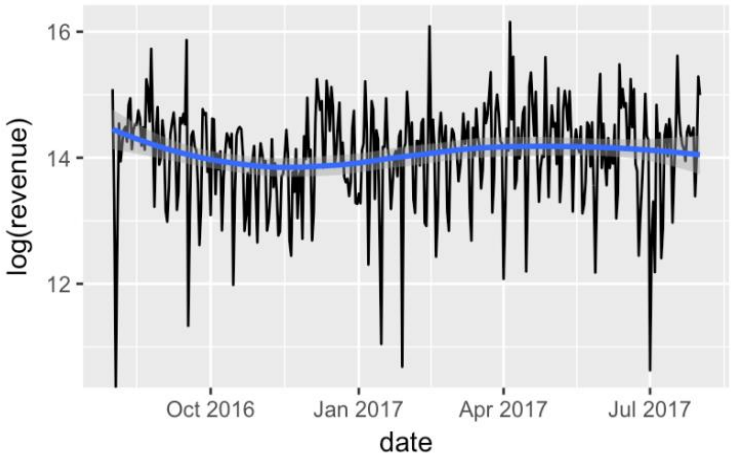


Fig.7 Log revenue over time

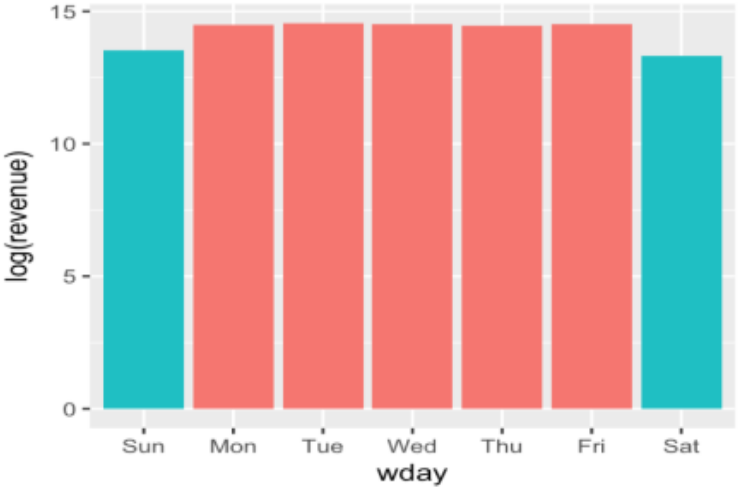


Fig.8 Log revenue over week

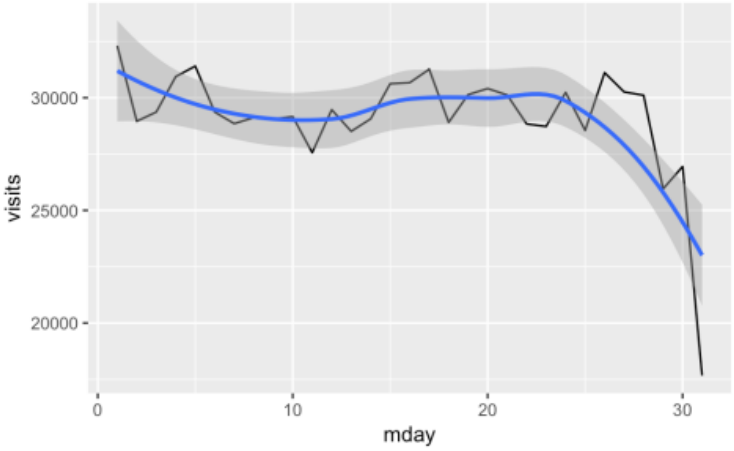


Fig.9 Total visits over month

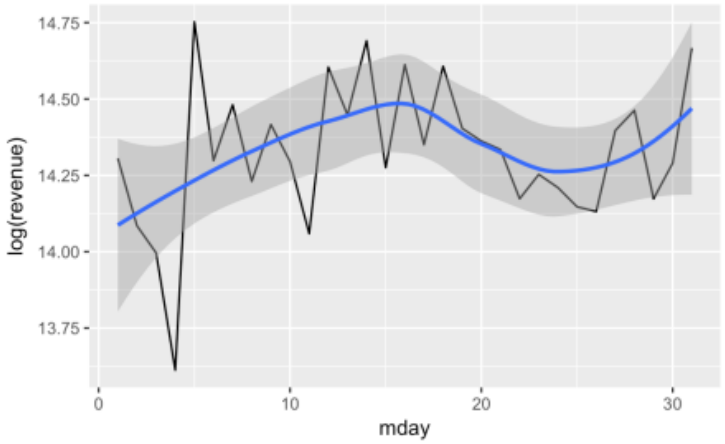


Fig.10 Log revenue over month

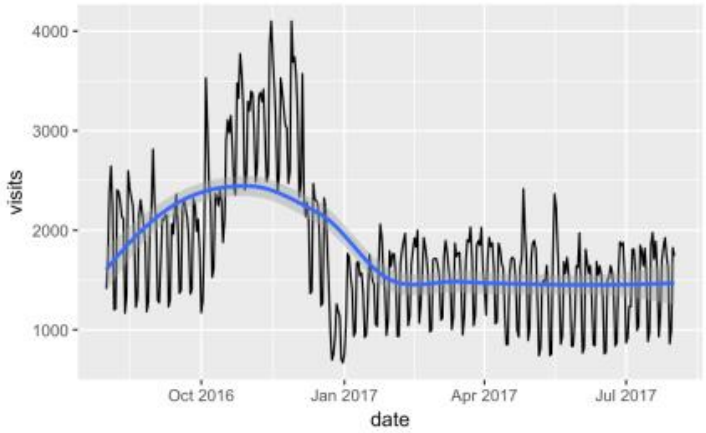


Fig.11 Visits from desktop over time

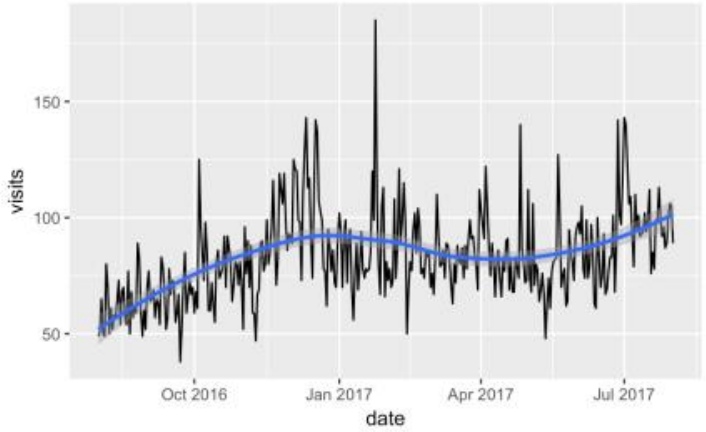


Fig.12 Visits from tablet over time

Fig.13 Top 10 cities with most visits in United States

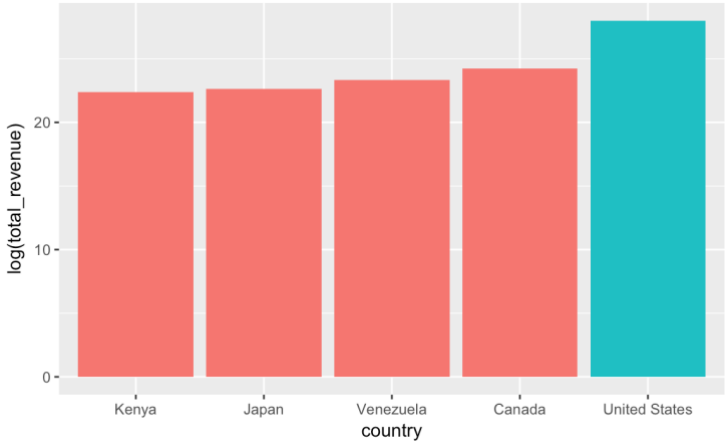


Fig.14 Log revenue in different countries

Fig.15 Top 10 cities with most revenue in United States

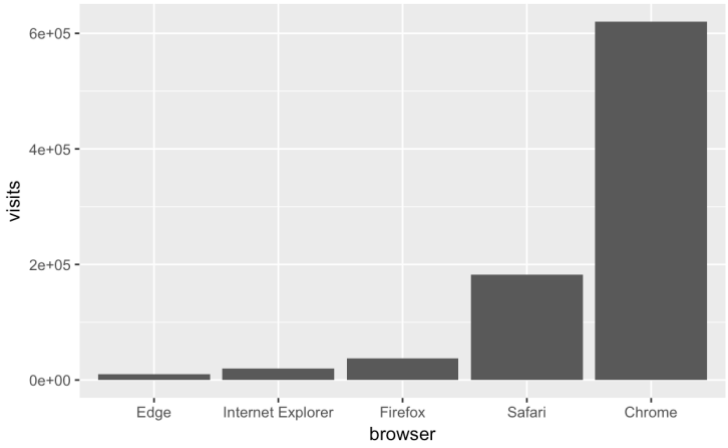


Fig.16 Visits from different browsers

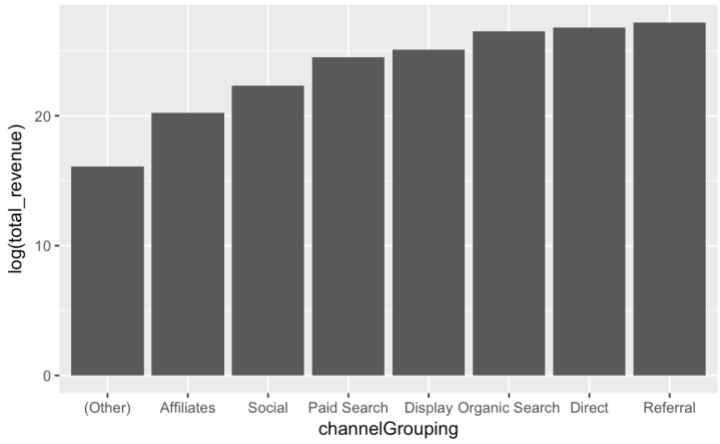


Fig.17 Log revenue from different channels

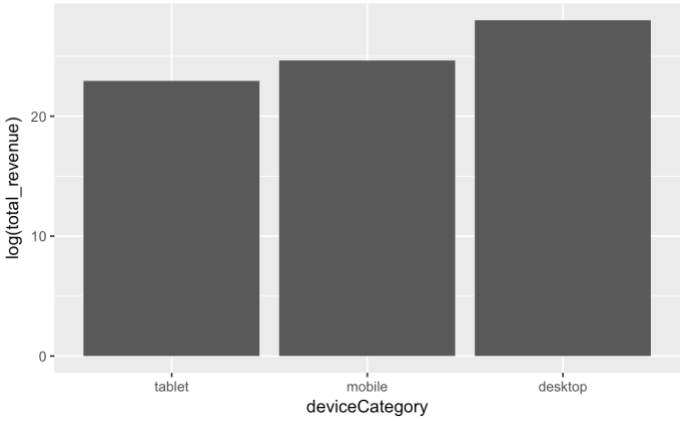


Fig.18 Log revenue from different devices

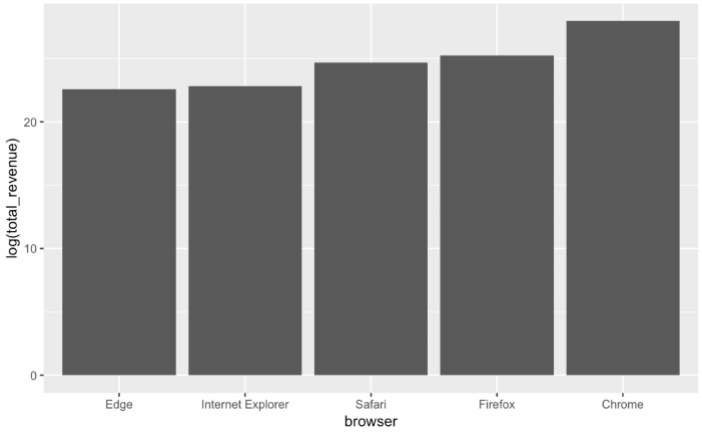


Fig.19 Log revenue from different browsers

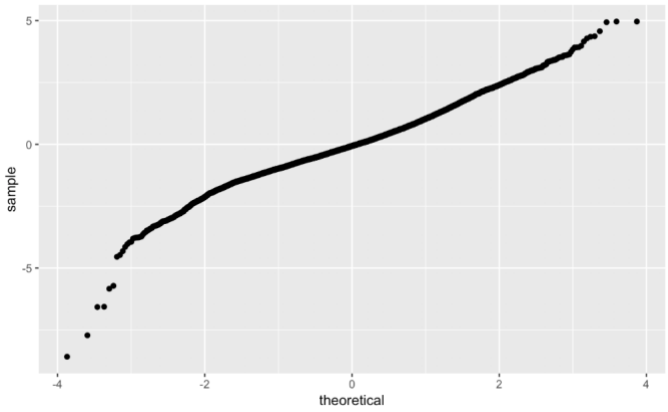


Fig.20 QQplot

VI. Codes related to the report can be found at

https://github.com/tonytontian/ds5110_data_menagement_course_project