BUMK776: Action Learning Project

Group member names:
Ziying Yan
Zidi Wang
Danli Hu
Junfei Hao
Yinan QI
Haoning Deng

Date: 2020/3/10

Honor Pledge:

We pledge our honor that we have not given or received any unauthorized assistance on this assignment.

1 Executive Summary

Based on the keywords customers entered in the search engine, we inferred that the dataset contains transaction information related to Google Merchandise Store, an e-commerce site that sells Google-branded merchandise. The dataset contains approximately 450,000 session-level records generated by more than 350,000 unique customers from August 2016 to August 2017.

Statistics descriptive of this report was based on session-level behavior and the prediction was based on the customer level. We discovered that there are some underlying patterns of behaviors behind the dataset. We assumed that a substantial amount of customers are likely to be working at silicon valley or some major cities in the United States. On workdays, they visited the online store with their laptops, most of which are Macintosh or Windows systems. They purchased during the break of work, usually from 10 am. To 11:00 a.m. Or from 1:00 p.m. to 3:00 p.m. and visited the website through referral.

Three models were applied in the project: Logistic Model, Linear Regression Model, and Random Forest. Random Forest cooperated with the Logistic Model and Linear Regression Model in that it worked as a feature selection tool.

We suggest that the client company adopt different promotion strategies for internal customers and external customers. To increase the purchase rate, the business owner could focus on the referral channel and Mactonish users. To increase the purchase amount, web-design for tablet users should be improved. What's more, an intense promotion could be applied on weekdays to motivate purchase. Geographically, advertisements could be focused on the United States, Canada users. Lastly, improving user experience to retain customers and motivate purchase is recommended.

2 Objective of analysis

This project was commissioned to analyze how customer characteristics and behaviors affect their purchases, predict the purchase revenue of each customer in the testing dataset, and build insights for the company in terms of decision making and advertising.

From the descriptive statistics, we aimed at analyzing the relationship between variables and discovering the underlying pattern of customer behaviors. To determine which features affect the probability of purchasing, we built a Logistic Model. Linear Regression to explain which factors affect the quantity of purchasing. For accurate prediction, we applied machine learning, and the random forest was chosen as the model.

3 Methodology

3.1 Inferential Statistics

With descriptive statistics methods, we were able to tell about the basic features of a set of data and to describe relationships between two variables. In the modeling results of Inferential Statistics, we cared more about the interpretation of coefficients and significance level than the prediction capability of the whole model. We chose the Logistic Regression model and the Multiple Linear Regression model for different purposes.

Logistic Regression model: The purpose of this model is to see what factors have an impact on consumers' purchase decisions. Because of the extremely heavy proportion of the records in "no transaction" class, we handled the class imbalance by downsampling. We also used the Variance Inflation Factor (VIF) to detect multicollinearity in the analysis. We fit the model with downsampling data, tested VIFs for all variables in the model and removed variables with the highest VIF. Then we refit the model and repeated the process until VIFs of all variables were under 10. Also, we removed all variables under the same category if no variable in that category appeared statistically significant.

Multiple Linear Regression model: The purpose of this model is to examine what factors have an impact on consumers' purchase amount. In this case, we only explore those records whose "transaction revenue" was greater than 0. First of all, we chose the semi-log model to normalize the dependent variable "transaction revenue" as possible. Then we obtained all influential observations by Cooks Distance and deleted all records with Cooks Distance of 4 times higher than the mean value.
The same variable selection process was repeated as in the Logistic Regression model.

Random Forest: Based on the unbalanced distribution of purchase incidence and high dimension characteristic of the dataset, we chose random forest as our prediction model.

3.3 Data Preparation

Original dataset: The original Train data contains 12 variables **(Exhibit 1).** We read the original dataset as JSON and split all the dictionary formatted variables (device, geonetwork, totals, and traffic source) into 55 individual variables.

How we deal with NA: The following statements are replaced with NA: "Not Socially Engaged", "not available in demo dataset","(not set)", "(not provided)".

Which variable is deleted: Variable "device is mobile" was deleted, because it is overlapped with variable "device.deviceCategory". Besides, variables solely consist of missing **(Exhibit 2)** were removed.

Newly created variable: For the "Local time" variable, the "date" variable in original data consists of POSIX time. In order to measure hour-level customer behaviors, POSIX time was first converted to human-readable time. With the help of Google API Service, we were able to derive coordinates and hence timezone according to the city name of each record. Next, UTC time was converted to local time according to the corresponding timezone. To notice that there was a small number of mismatch between country and city, e.g. the city for a record is HongKong while the country indicates the United States. We assume that the use of a VPN leads to a mismatch.

Dummy Variables: After the data cleaning process based on frequency, 83 variables were finally chosen. "Weekday" indicates whether a specific day is weekday(1) or weekends(0). "Month" included 12 dummy variables for the month of January to December. Seven dummy variables were created for each individual weekday. "TransactionRevenue_dummy" was a dummy variable created to indicate purchase incidence, "TransactionRevenue_dummy" equaled 1 if there was a purchase.

Categorical variables: we counted the frequency of each category and kept 5 to 10 most frequently appeared categories, classifying the rest categories into "others".

Aggregation: The dataset was aggregated based on "Fullvisitorid". During the aggregation process, we summed dummy variables and changed them into count variables.

3.4 Sampling

Our team weighed between two sampling methods, random sampling and disproportionate stratified random sampling.

The proportion of purchase incidence was highly unbalanced (1:0.014), which meant the number of purchases was approximately 70.704 times larger than the number of non-purchase. Due to the extremely disproportional purchase instance, we used random oversampling to scale down the ratio to reach a more balanced distribution and avoid purchased being ignored in the model building process. However, during the model training process, we found the RMSE of the disproportional sample (between 1.8 and 2.0) was always larger than the RMSE of the random sample (between 1.6 and 1.7) in the training dataset. Based on this we chose random sampling as our sampling method to build the random forest model.

Bootstrap sampling was used for records that "TransactionRevenue_dummy" equals 1 and no replacement sampling was used for those that "TransactionRevenue_dummy" equals 0. The proportion between 0 and 1 achieved 1:10 with the disproportional stratified random sampling method.

The conclusion above was two sampling methods for parameter tuning. After the parameters were decided, we applied default parameters to the test dataset and derived the corresponding RMSE. We found that random sampling resulted in lower RMSE (1.638) than that with disproportional stratified sampling(1.686).

3.5 Model Optimization

Feature Selection: The random forest model generated feature importance for all the 82 features in the dataset. Referring to descriptive statistics, we had a quick selection and kept features which have importance more than 0.001. After trying different combinations of features, we finally decided to keep 32 important features for our model (**Exhibit 3** and **Exhibit 4**).

Parameter Adjustment (Grid Search): To perform hyperparameter optimization for the random forest model, we used the grid search, which is an exhaustive search over specified parameter values for an estimator. From the grid search result we adjusted "n_estimators" to 1200, "min_samples_split" into 38, "min_samples_leaf" into 29, "max_features" into "auto", "max_depth" into 18, and "bootstrap" into "True".

***10-fold cross validation*:** We applied the cross-validation technique to determine the hyperparameters of the random forest model, to test which parameters will result in the lowest test error. In the 10-fold cross-validation, the original sample was randomly partitioned into k equal sized subsamples. From the k subsamples, a single subsample was retained as the validation data for testing the model, and the remaining (k - 1) subsamples were used as training data. The cross-validation process was then repeated k times.

4 Results and findings

4.1 Descriptive Statistics

The original dataset consists of 451,626 records containing session level online store purchase information. Among records generated by 357,300 unique customers, 5733 have transaction revenue. “Bounces” shows whether the session was wrongly clicked. Among all the records, 222,558 of them were indicated to be miss clicked. We explored the data both at the session level and the customer level.

Keywords: “google store” has the highest frequency among all keywords. It is reasonable to infer that the records were from online google store related activities.

Trafficsource: For records that have transaction revenue, 2,534 of them are from *mall.googleplex.com*, which is an internal website for google employees. This internal website also produced the highest amount of revenue compared with other sources. So it could be inferred that the majority of the transaction revenue are from internal customers. (**Graph 1**)

Geographic: Among all countries, the United States contributes to 54.27% of the total transaction revenue with counts of 5,451, followed by Canada with 106 counts and 1.53% of the total transaction revenue **(Graph 2)**. City-level revenue **(Graph 3)** shows a positive relationship between sales and the social and economic status of the cities. Four cities in the U.S. contribute to more than one third of its revenue. New York ranks the first with 13.33% of total revenue, followed by Mountain View (7.05%), San Francisco (6.71%) and Chicago (5.12%). Mountain View is the headquarter of the Alphabet. Inc, somehow supporting our assumption that the sales is probably mostly driven by the internal needs of Google.

Time: The number of sessions reached the crescendo in October 2016 to December 2017 (**Graph 4.1** and **Graph 4.2**), while the month-level (**Graph 5**) transaction revenue peaked in the middle of September, contributing to the monthly sales peak. Our hypothesis is that purchase occurred in the online store is mostly for internal use, such as celebrating Google Anniversary. It is also worth noticing that compared with weekends, people preferred to purchase on weekdays **(Graph 6)**. Specifically, people spent more on Tuesday and Wednesday in March and April. People’s spending behavior changed from hour to hour **(Graph 7)**, among which 10:00 a.m., 2:00 p.m., and 3:00 p.m. contributed the most to the transaction revenue while the average transaction reached the zenith at 1:00 a.m., September 2016.

Pageviews: The distributions of “pageview” and “hits” are highly correlated (**Graph 8**) As we can see from **Graph 9,** the distribution in the pageview frequency table is right-skewed, and most revenues were generated when pageviews equal 10 to 70. The total revenue peaked when pageviews is 89.

Channel Grouping: For all the records, 42.10% are from organic search, 25.10% are from social, 15.80% are from direct, and 17.00% are from other channels. **(Graph 10)**. Among all the channels, referral contributed to the largest amount of transaction revenue, and the display channel appears to have the largest amount of average transaction revenue.

Devices: For all records with Transaction Revenue, 95.12% are from desktop, with 7.74% from mobile and 2.77% from tablet **(Graph 11.1)**. The bar chart of the operating system indicates that the largest amount of total revenue comes from Macintosh, and the highest average transaction revenue is from Windows **(Graph 11.2)**. For browsers, Google Chrome produces the largest amount of revenue, and Firefox leads to the highest average revenue **(Graph 11.3)**.

4.2 Coefficient Interpretation

All coefficient estimates **(Exhibit 5 & Exhibit 6)** in the Logistic Regression model and Multiple Linear Regression model enable us to interpret the factor impact in terms of its direction and degree. Since we already removed variables of high VIF, those removed categorical variables combined acted as baseline. For example, in the logistic regression model, we compared the probability of consumer purchase among different channels based on that through Affiliate, Direct and Other channels together.

The same factor can cause the opposite impact on these two models. For purchase incidence, consumers acquired through channel Referral are 41.00% more likely to make a purchase than consumers through baseline channel. Interestingly, at the same time, those who visited through Referral tend to spend 23.94% less than those who visited by baseline channel. Oppositely, some factors could influence these two dependent variables in the same direction. For example, each unit increase in pageviews can lead to a 69.35% increase in the chance of purchase and a 1.82% increase in the purchase amount. While if consumers are new visitors, generally they will be 64.41% less likely to make the purchase and spend 29.57% less amount of dollars in the store.

4.3 Important Features & RMSE

Generated by the random forest model, the important features were mostly consistent with descriptive statistics findings. The most important feature was “totals.pageviews”. The features listed in **Exhibit 5** were features relatively more important than others.

Different combinations of features yielded different RMSE results in the random forest model. Feature selection resulted in the RMSE reduction from over 1.8 to 1.614 (**Graph 12**) for the random forest model. To verify whether our model has an overfitting problem, we compared the RMSE of training and testing data sets. In the beginning, the difference of RMSE was relatively large (between 0.6-1.2), indicating an overfitting problem of our model (**Graph 12**). We decided the model was valid after parameter adjustment, in which the RMSE difference shrunk to 0.04.

5 Recommendation

We assume that the dataset contains transaction information of Google Merchandise Store and that the revenue was mostly contributed by Google employees for internal use. For example, employees got a lot of random swag for office events. Also, the peak of revenue could result from the preparation of a celebration for Google Anniversary in September. Based on the assumptions, we recommend that our client make an effort to advertise towards the general public, increase revenue by meeting the external need, and having a better promotion plan around Google Anniversary and other celebrations events inside the company should be promoted towards employees that could increase the revenue.

Since our customers mainly come from the United States and Canada, and they are more likely to purchase during weekdays compared with weekends, we recommend our client to invest more in marketing activities or promotions in these countries, and make more promotions on weekdays. Besides, our client should pay attention to the 1:00 a.m. which has an upsurge of average transaction revenues in a day, we recommend our client having a special investigation to find reasons for the consumer purchase pattern in the hours. Also, there is a chance for the client company to explore the Asia and South American markets.

We found that pageview is significantly important both for the purchase conversion rate and the transaction revenue amount, users who have macintosh devices are more likely to make a purchase, at the same time, tablet users generate the most revenue on average. We recommend that the client invest more in the referral channel to increase the purchase conversion rate, and applying such information to improve segmentation strategies and target different user devices with different strengths of advertising push, meanwhile, improve the web-design and user-experience design for both websites and tablet users, which can retain users and stimulate them to purchase or purchase more.

Moreover, based on the situation where over half of the observations are bounce behavior, we suggest our client have a deeper look at the bounce rate to improve cost-effectiveness, such as which channels or websites have a higher bounce rate.

Appendix

Exhibit 1. Original train data

variable	meaning
channel grouping	The channel via which the user came to the Store.
Date	The date on which the user visit the store
Device	The specifications for the device used to access the store
Fullvisitor id	A unique identifier for each user in the dataset
Geonetwork	This section contains information about the geography of the user
Session id	The unique identifier for the session
Social engagement type	either "Socially Engaged" or "Not Socially Engaged".
Totals	This section contains summary values for the entire session.
Traffic source	This section of information displays information related to the source of the traffic that initiated the session.
Visit id	Unique identifier of a session
Visit number	The session number for this user. If this is the first session, then this is set to 1
Visit start time	The timestamp.Unix timestamp (the number of seconds since January 1, 1970 (midnight UTC / GMT)

Exhibit 2. Deleted variables

	Variable name

1	device is mobile
2	socialEngagementType
3	device.browserVersion
4	device.browserSize
5	device.operatingSystemVersion
6	device.mobileDeviceBranding
7	device.mobileDeviceModel
8	device.mobileInputSelector
9	device.mobileDeviceInfo
10	device.mobileDeviceMarketingName
11	device.flashVersion
12	device.language
13	device.screenColors
14	device.screenResolution
15	geoNetwork.cityId
16	geoNetwork.latitude
17	geoNetwork.longitude
18	geoNetwork.networkLocation
19	trafficSource.adwordsClickInfo.criteriaParameters'

Exhibit 3. Variable interpretation in model

variable	meaning
ChannelGroupingDirect	the traffic which you will get when someone visits your website directly tying in browser.

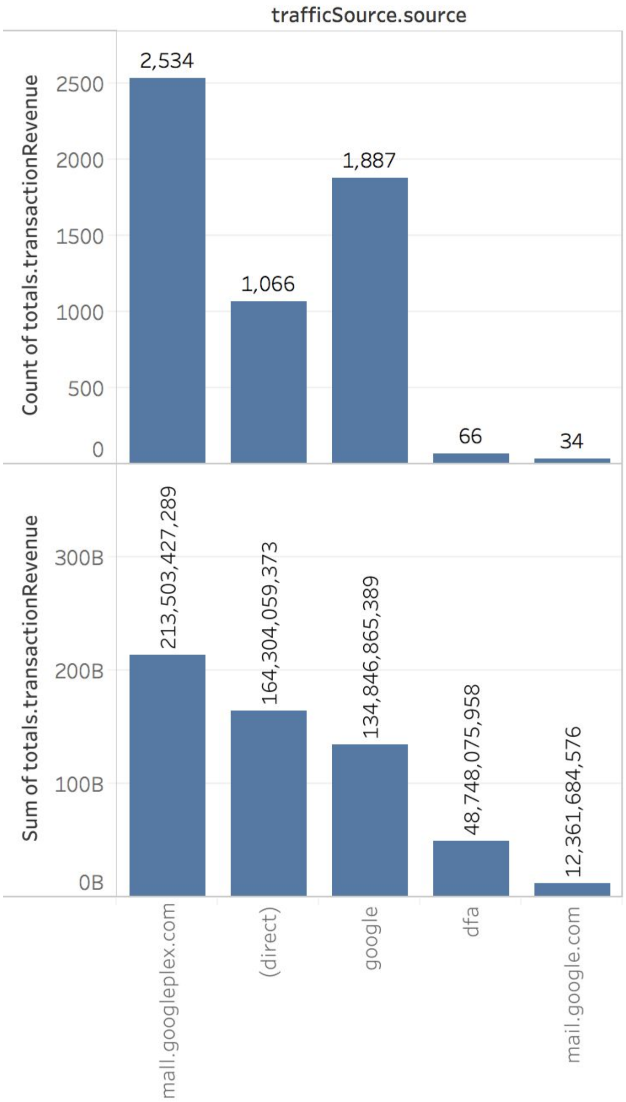
ChannelGroupingDisplay	True if the source of the session is "Direct" (meaning the user entered the name of your website URL in a browser or visited your site via a bookmark); if 2 consecutive but different sessions have the exact same ad Series details, this field is also true.
ChannelGroupingOrganic.Search	this type of traffic you will get from search engines such as Google, Yahoo, Bing, Yandex, etc.
ChannelGroupingReferral	what you get from other sites who placed your link on their website.
fullVisitorId	A unique identifier for each user in the dataset
visitNumber	The session number for this user. If this is the first session, then this is set to 1
totals.hits	The number of files your website serves
totals.pageviews	number of pages a user views in a session
trafficSource.isTrueDirect	True if the source of the session is "Direct" (meaning the user entered the name of your website URL in a browser or visited your site via a bookmark); if 2 consecutive but different sessions have the exact same ad Series details, this field is also true. Otherwise, it is NULL.
Month (1，2，3…12)	The month in which the user visits the store
Day (Sunday，Monday…Saturday)	The day on which the user visits the store
device.operatingSystem.Macintosh	The user used Macintosh system to access the store
device.operatingSystem.Windows	The user used Windows system to access the store
device.operatingSystem.Chrome.OS	The user used Chrome system to access the store
geoNetwork.country.Canada	The user is located in Canada
geoNetwork.country.United.States	The user is located in the united states
log transaction	The logarithm of transaction

Exhibit 4. Feature importance while building models

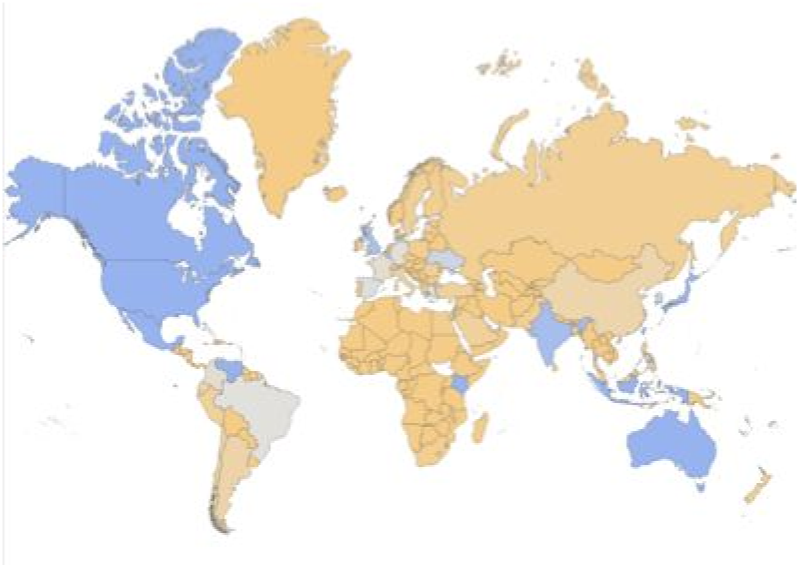
Features	Importances
totals.pageviews	0.640
geoNetwork.country.United.States	0.139
totals.hits	0.044
visitNumber	0.031
device.operatingSystem.Macintosh	0.028
channelGroupingReferral	0.023
channelGroupingOrganic.Search	0.014
month9	0.007
device.operatingSystem.Windows	0.006
channelGroupingDirect	0.006
dayMonday	0.005
trafficSource.isTrueDirect	0.005
dayWednesday	0.005
month8	0.005
month5	0.004
dayThursday	0.004
dayFriday	0.004
dayTuesday	0.004
month10	0.003
device.operatingSystem.Chrome.OS	0.003
geoNetwork.country.Canada	0.003
daySunday	0.003
month6	0.002
month7	0.002
month12	0.002
month4	0.002
daySaturday	0.001
month3	0.001
month11	0.001
month1	0.001
month2	0.001
channelGroupingDisplay	0.001

Exhibit 5
Exhibit 6

Graph 1

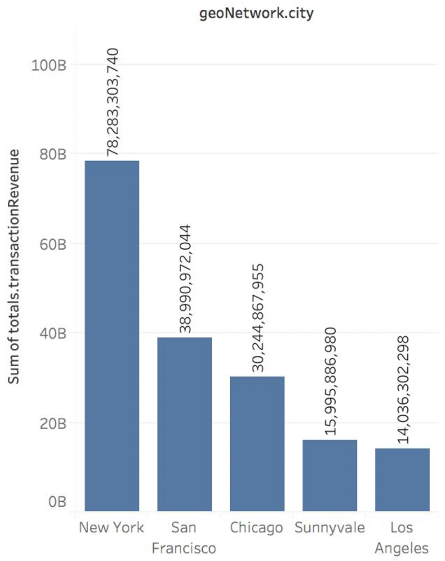


Graph 2. Total transaction revenue distribution among countries



In this map, we used extreme colors to show the range of transaction revenue, blue stands for the largest amount or total transaction revenue while yellow stands for the least amount of total transaction revenue.

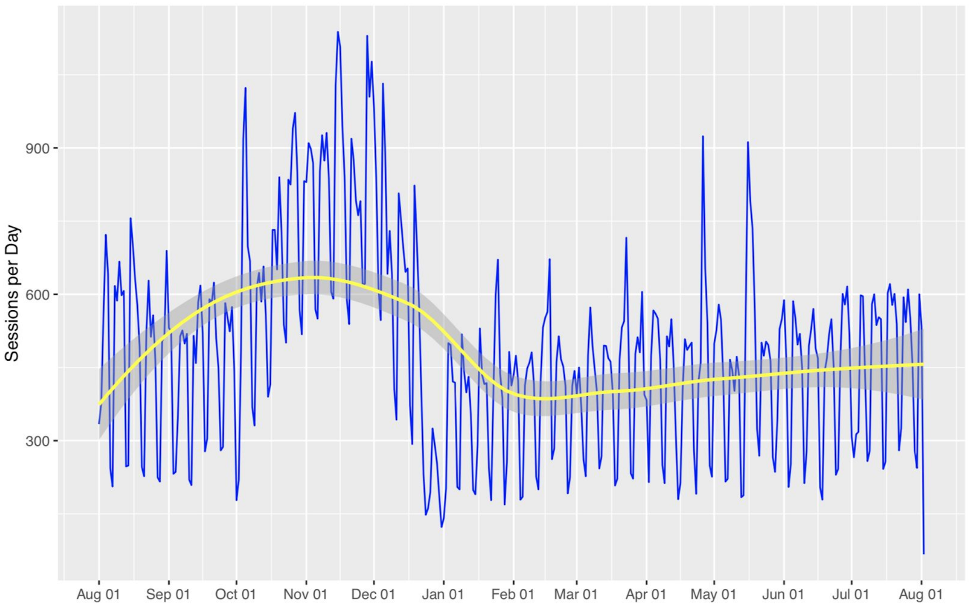
Graph 3. Total transaction revenue distribution among countries



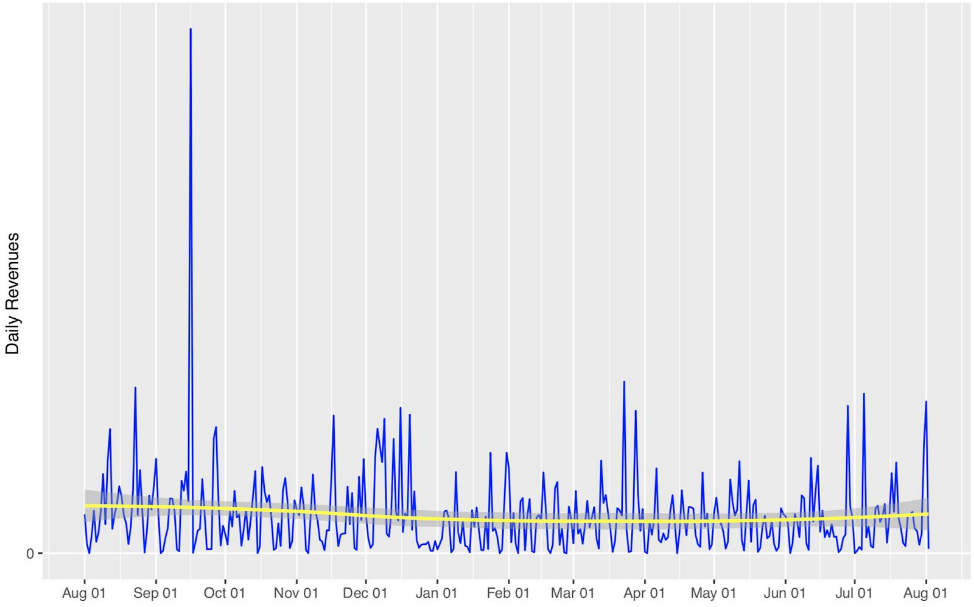
Sum of totals.transactionRevenue for each geoNetwork.city. The view is filtered on geoNetwork.city, which keeps Chicago, Los Angeles, New York, San Francisco and Sunnyvale.

Graph 4. Total session and transaction revenue distribution among year by day

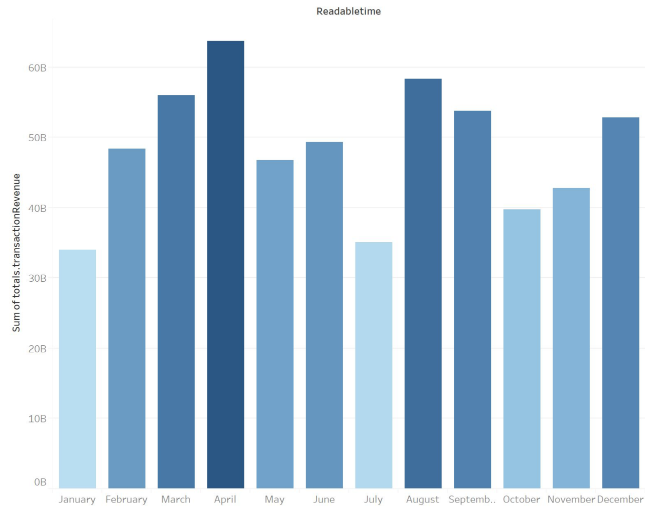
4.1. Total sessions distribution among year by daily level



4.2. Total transaction revenue distribution among year by day

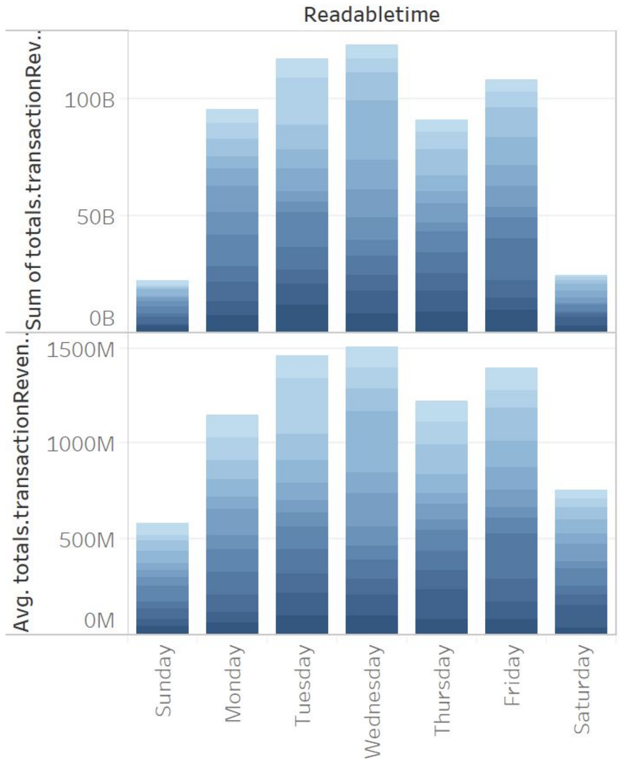


Graph 5. Month-level transaction revenue distribution



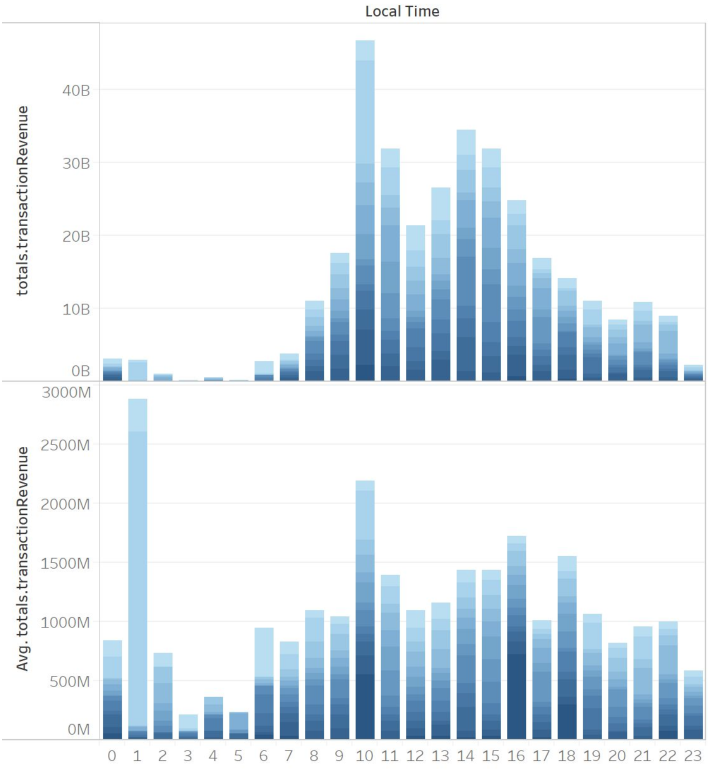
Sum of totals.transactionRevenue for each Readabletime Month. Color shows sum of totals.transactionRevenue.

Graph 6. Weekday transaction revenue and average transaction revenue distribution



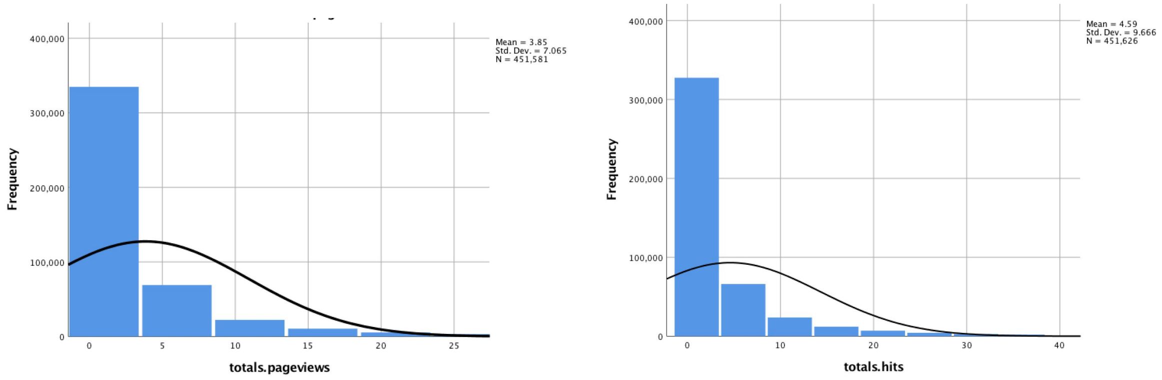
Sum of totals.transactionRevenue and average of totals.transactionRevenue for each Readabletime Weekday. Color shows details about Month.

Graph 7. Total transaction revenue distribution and average transaction revenue at hour level

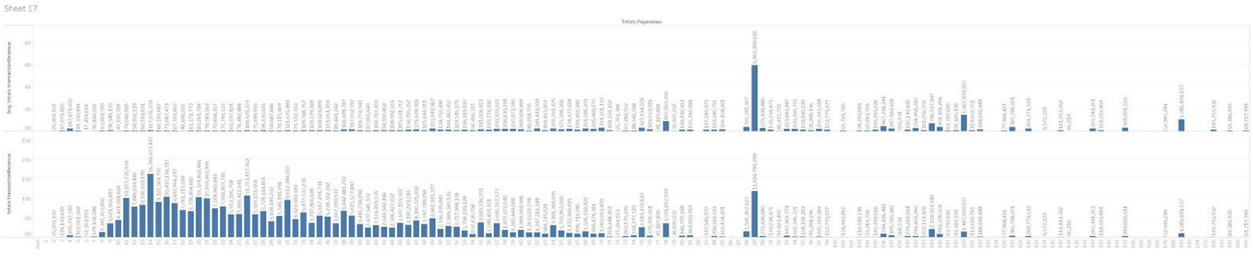


Sum of totals.transactionRevenue and average of totals.transactionRevenue for each Local Time Hour. Color shows details about Local Time Month. The data is filtered on Exclusions (HOUR(Local Time),YEAR(Local Time)), which keeps 48 members.

Graph 8. Frequency distribution of Page view and hits

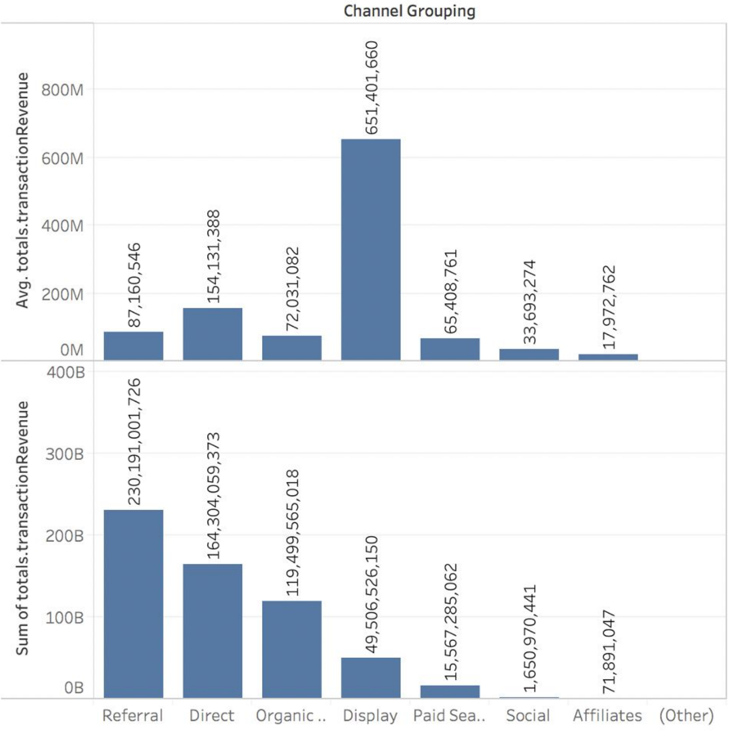


Graph 9. Transaction Revenue at different page view amounts.



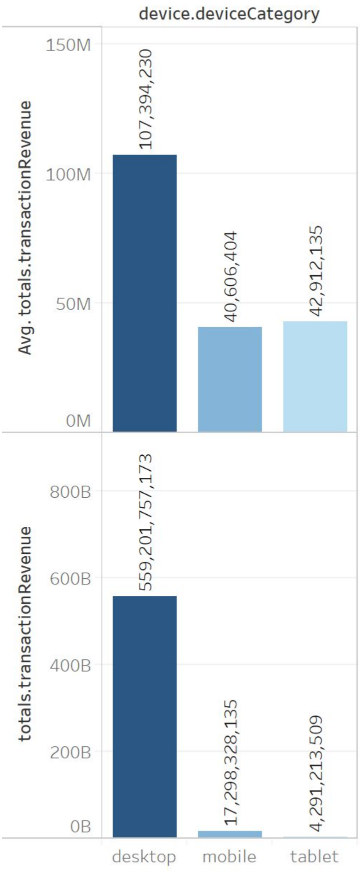
Average of totals.transactionRevenue and sum of totals.transactionRevenue for each Totals.Pageviews. The data is filtered on Device.Browser and geoNetwork.city. The Device.Browser filter has multiple members selected. The geoNetwork.city filter excludes (not set) and not available in demo dataset.

Graph 10. Channel grouping

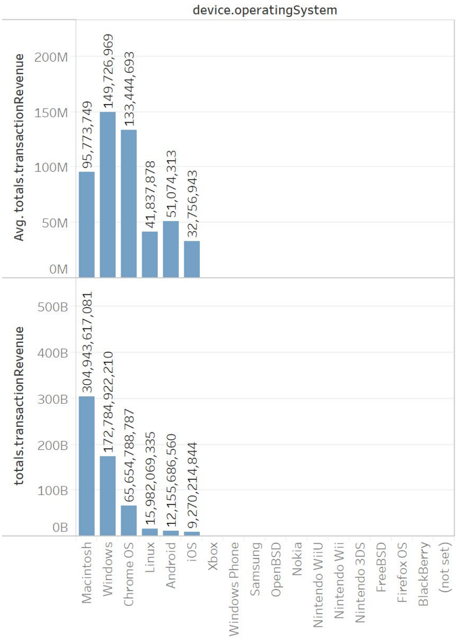


Graph 11. Transaction revenue distribution among different device category, operating system and browser

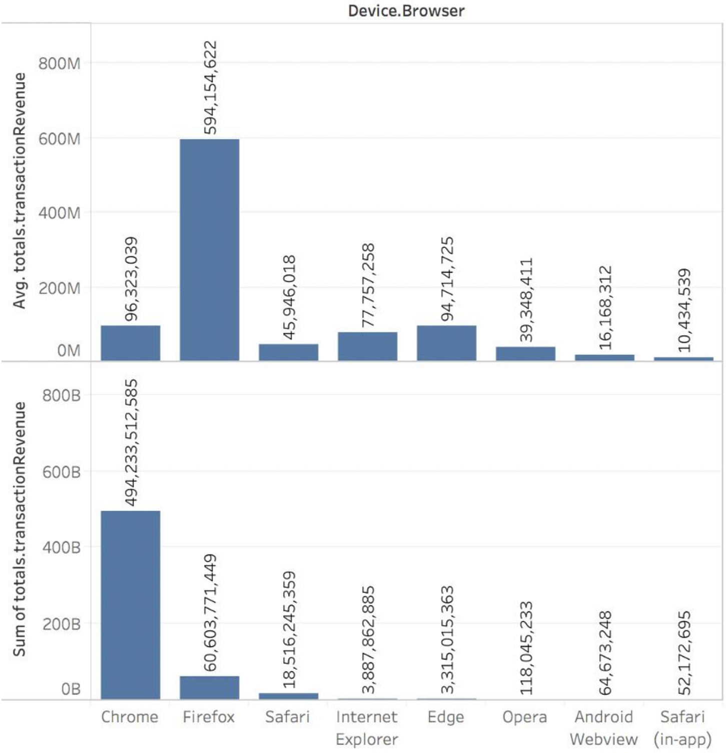
11.1. Transaction revenue distribution among different device categories



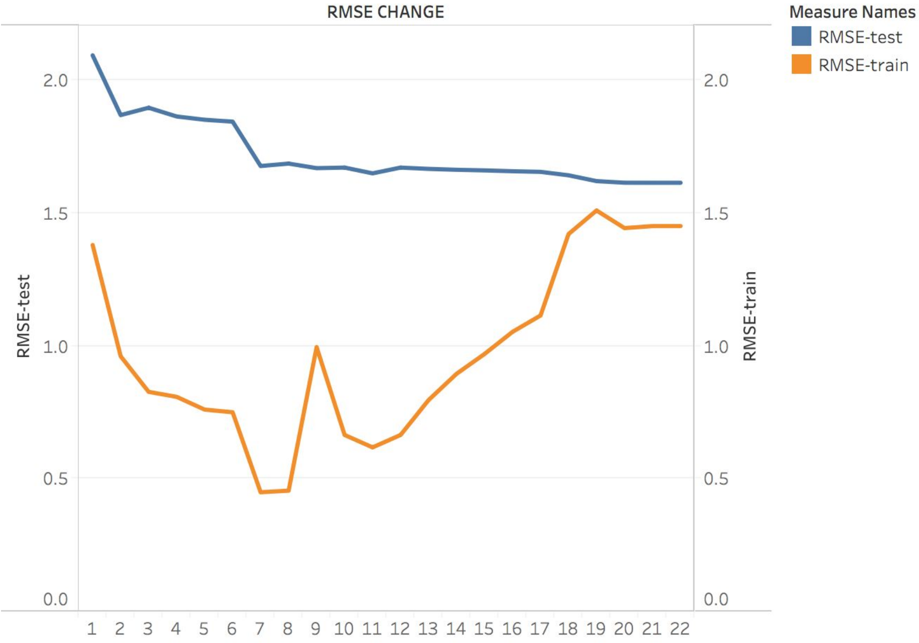
11.2. Transaction revenue distribution among different operating systems



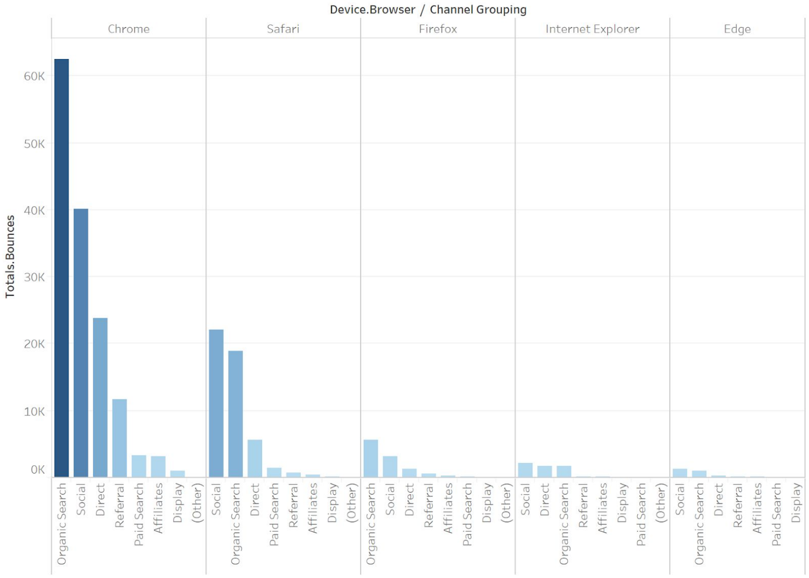
11.3. Transaction revenue distribution among different browsers

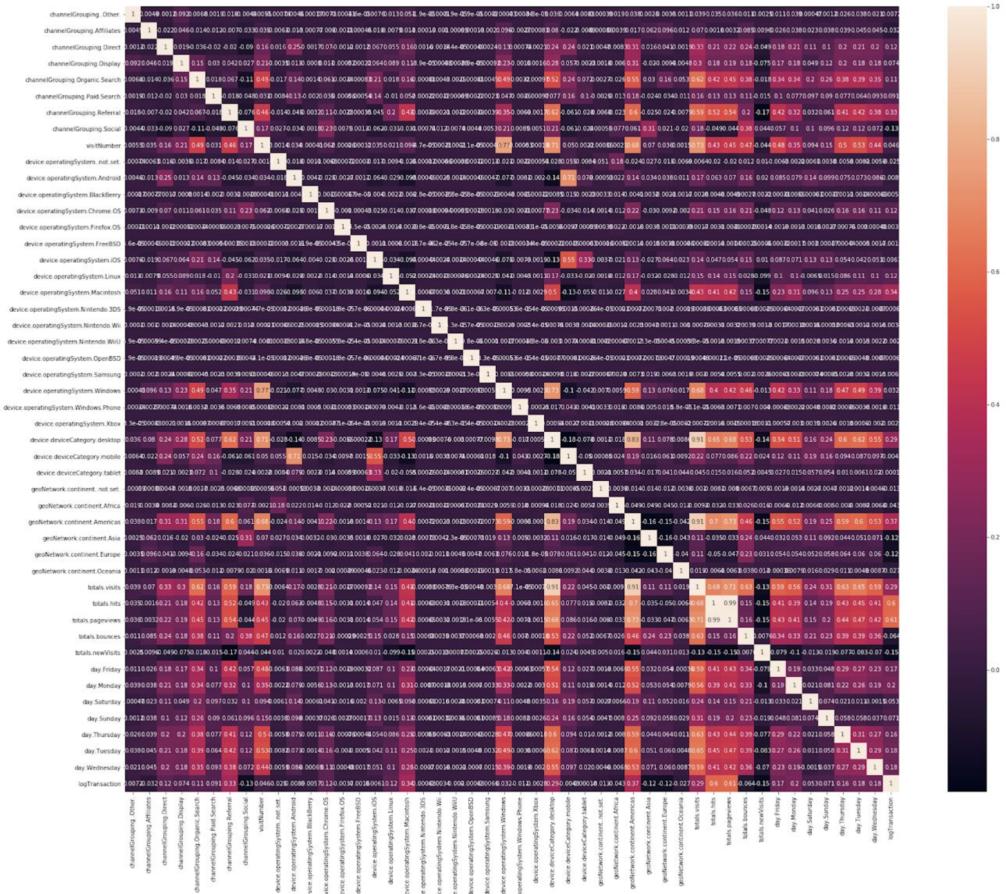


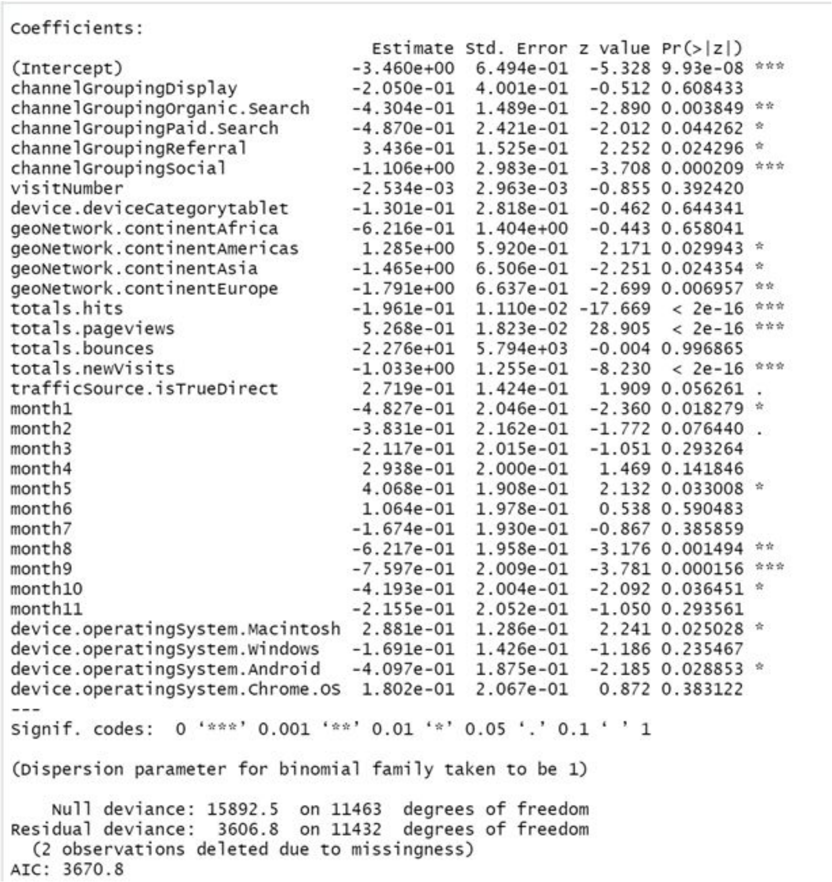
Graph 12 Change of RMSE Results



Graph 13 Bounce among device and channel grouping




channelGrouping..Other.
channelGrouping.Affiliates
channelGrouping.Direct
channelGrouping.Display
channelGrouping.Organic.Search
channelGrouping.Paid.Search
channelGrouping.Referral
channelGrouping.Social
visitNumber
device.operatingSystem..not.set.
device.operatingSystem.Android
device.operatingSystem.BlackBerry
device.operatingSystem.Chrome.OS
device.operatingSystem.Firefox.OS
device.operatingSystem.FreeBSD
device.operatingSystem.iOS
device.operatingSystem.Linux
device.operatingSystem.Macintosh
device.operatingSystem.Nintendo.3DS
device.operatingSystem.Nintendo.Wii
device.operatingSystem.Nintendo.WiiU
device.operatingSystem.OpenBSD
device.operatingSystem.Samsung
device.operatingSystem.Windows
device.operatingSystem.Windows.Phone
device.operatingSystem.Xbox
device.deviceCategory.desktop
device.deviceCategory.mobile
device.deviceCategory.tablet
geoNetwork.continent..not.set.
geoNetwork.continent.Africa
geoNetwork.continent.Americas
geoNetwork.continent.Asia
geoNetwork.continent.Europe
geoNetwork.continent.Oceania
totals.visits
totals.hits
totals.pageviews
totals.bounces
totals.newVisits
day.Friday
day.Monday
day.Saturday
day.Sunday
day.Thursday
day.Tuesday
day.Wednesday
logTransaction
1.0
0.8
0.6
0.4
0.2
0.0



```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                   -3.460e+00  6.494e-01  -5.328 9.93e-08 ***
channelGroupingDisplay        -2.050e-01  4.001e-01  -0.512 0.608433    
channelGroupingOrganic.Search -4.304e-01  1.489e-01  -2.890 0.003849 ** 
channelGroupingPaid.Search    -4.870e-01  2.421e-01  -2.012 0.044262 *  
channelGroupingReferral        3.436e-01  1.525e-01   2.252 0.024296 *  
channelGroupingSocial         -1.106e+00  2.983e-01  -3.708 0.000209 ***
visitNumber                   -2.534e-03  2.963e-03  -0.855 0.392420    
device.deviceCategorytablet   -1.301e-01  2.818e-01  -0.462 0.644341    
geoNetwork.continentAfrica    -6.216e-01  1.404e+00  -0.443 0.658041    
geoNetwork.continentAmericas   1.285e+00  5.920e-01   2.171 0.029943 *  
geoNetwork.continentAsia      -1.465e+00  6.506e-01  -2.251 0.024354 *  
geoNetwork.continentEurope    -1.791e+00  6.637e-01  -2.699 0.006957 ** 
totals.hits                   -1.961e-01  1.110e-02 -17.669  < 2e-16 ***
totals.pageviews               5.268e-01  1.823e-02  28.905  < 2e-16 ***
totals.bounces                -2.276e+01  5.794e+03  -0.004 0.996865    
totals.newVisits              -1.033e+00  1.255e-01  -8.230  < 2e-16 ***
trafficSource.isTrueDirect     2.719e-01  1.424e-01   1.909 0.056261 .  
month1                        -4.827e-01  2.046e-01  -2.360 0.018279 *  
month2                        -3.831e-01  2.162e-01  -1.772 0.076440 .  
month3                        -2.117e-01  2.015e-01  -1.051 0.293264    
month4                         2.938e-01  2.000e-01   1.469 0.141846    
month5                         4.068e-01  1.908e-01   2.132 0.033008 *  
month6                         1.064e-01  1.978e-01   0.538 0.590483    
month7                        -1.674e-01  1.930e-01  -0.867 0.385859    
month8                        -6.217e-01  1.958e-01  -3.176 0.001494 ** 
month9                        -7.597e-01  2.009e-01  -3.781 0.000156 ***
month10                       -4.193e-01  2.004e-01  -2.092 0.036451 *  
month11                       -2.155e-01  2.052e-01  -1.050 0.293561    
device.operatingSystem.Macintosh  2.881e-01  1.286e-01   2.241 0.025028 *  
device.operatingSystem.Windows   -1.691e-01  1.426e-01  -1.186 0.235467    
device.operatingSystem.Android   -4.097e-01  1.875e-01  -2.185 0.028853 *  
device.operatingSystem.Chrome.OS  1.802e-01  2.067e-01   0.872 0.383122    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15892.5  on 11463  degrees of freedom
Residual deviance:  3606.8  on 11432  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 3670.8
```

Table: Results of Logistic Regression model

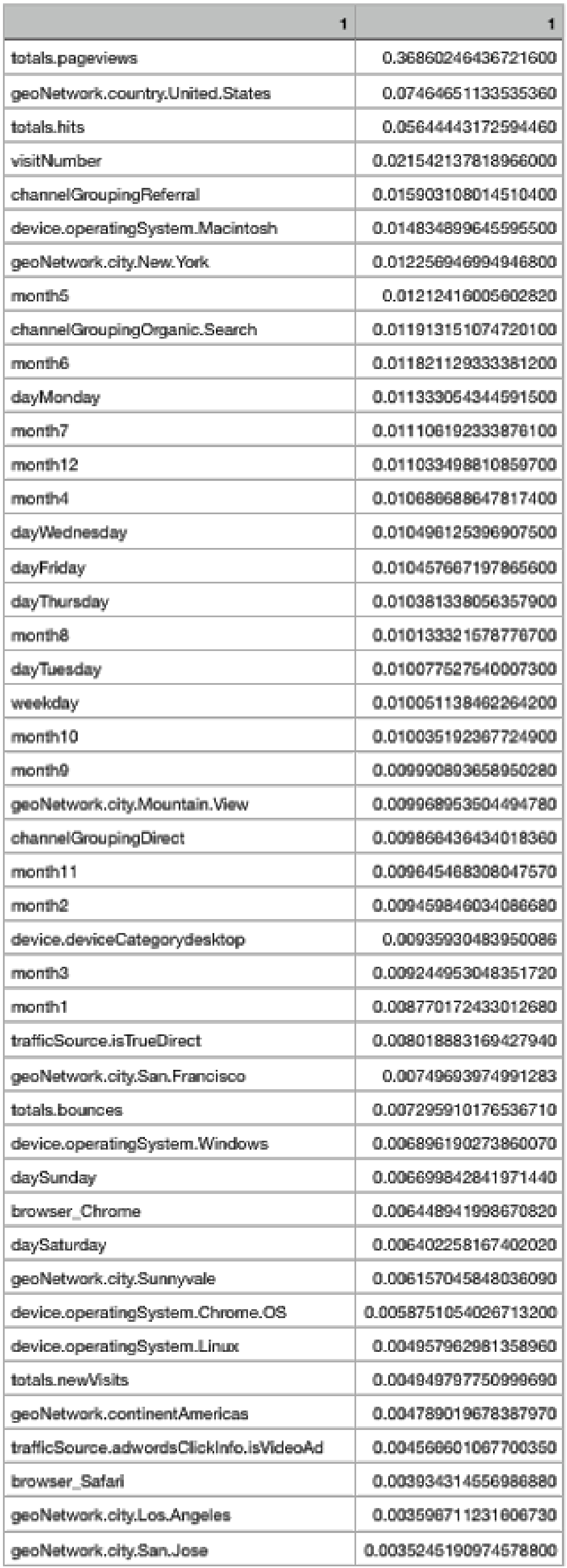
Table: Results of Linear Multiple Regression model

	Estimate	Std.Error	t-Value	Pr(>|t|)	
(Intercept)	17.74752	0.348114	50.982	2.00E-16	***
channelGroupingAffiliates	-1.06199	0.998638	-1.063	0.287632	
channelGroupingDisplay	0.112087	0.14276	0.785	0.432401	
channelGroupingOrganic.Search	-0.4035	0.04785	-8.433	2.00E-16	***
channelGroupingPaid.Search	-0.30264	0.083564	-3.622	0.000295	***
channelGroupingReferral	-0.27367	0.045231	-6.051	1.54E-09	***
channelGroupingSocial	-0.60245	0.16456	-3.661	0.000254	***

visitNumber	0.007837	0.001397	5.61	2.12E-08	***
device.deviceCategorytablet	0.318494	0.117945	2.7	0.006948	**
geoNetwork.continentAmericas	-0.5036	0.333641	-1.509	0.131252	
geoNetwork.continentAsia	-0.32222	0.360526	-0.894	0.37149	
geoNetwork.continentEurope	-0.69416	0.391103	-1.775	0.075974	.
totals.pageviews	0.017961	0.000766	23.458	2.00E-16	***
totals.newVisits	-0.35055	0.041102	-8.529	2.00E-16	***
trafficSource.isTrueDirect	-0.01328	0.043554	-0.305	0.760427	
month1	0.079754	0.068087	1.171	0.241507	
month2	0.093225	0.066889	1.394	0.163458	
month3	0.170646	0.061274	2.785	0.005372	**
month4	0.134131	0.061534	2.18	0.029316	*
month5	0.139968	0.057923	2.416	0.015705	*
month6	0.081585	0.060198	1.355	0.17538	
month7	-0.02764	0.058648	-0.471	0.637506	
month8	0.184493	0.057468	3.21	0.001333	**
month9	0.152979	0.062071	2.465	0.013747	*
month10	0.113734	0.061157	1.86	0.062981	.
month11	0.118029	0.062228	1.897	0.057918	.
dayFriday	0.356414	0.057317	6.218	5.39E-10	***
dayMonday	0.343472	0.056258	6.105	1.10E-09	***
daySaturday	0.108782	0.069054	1.575	0.115239	
dayThursday	0.327891	0.057274	5.725	1.09E-08	***
dayTuesday	0.353533	0.056866	6.217	5.44E-10	***
dayWednesday	0.296759	0.056844	5.221	1.85E-07	***

browser_Safari	-0.28955	0.073524	-3.938	8.31E-05	***
browser_Firefox	-0.30649	0.114267	-2.682	0.007336	**
browser_IE	-0.04307	0.161268	-0.267	0.789408	
browser_AndroidWebview	-0.47791	0.999921	-0.478	0.63271	
browser_Safariapp	-0.90061	0.582999	-1.545	0.122455	
device.operatingSystem.iOS	-0.58944	0.100184	-5.884	4.25E-09	***
device.operatingSystem.Linux	-0.55941	0.069953	-7.997	1.54E-15	***
device.operatingSystem.Macintosh	-0.13567	0.049708	-2.729	0.006367	**
device.operatingSystem.Windows	-0.34788	0.05639	-6.169	7.35E-10	***
device.operatingSystem.Android	-0.78919	0.085536	-9.226	2.00E-16	***

全部feature importance



1	1
totals.pageviews	0.36860246436721600
geoNetwork.country.United.States	0.07464651133535360
totals.hits	0.05644443172594460
visitNumber	0.021542137818966000
channelGroupingReferral	0.015903108014510400
device.operatingSystem.Macintosh	0.014834899645595500
geoNetwork.city.New.York	0.012256946994946800
month5	0.01212416005602820
channelGroupingOrganic.Search	0.011913151074720100
month6	0.011821129333381200
dayMonday	0.011333054344591500
month7	0.011106192333876100
month12	0.011033498810859700
month4	0.010686688647817400
dayWednesday	0.010496125396907500
dayFriday	0.010457667197865600
dayThursday	0.010381338056357900
month8	0.010133321578776700
dayTuesday	0.010077527540007300
weekday	0.010051138462264200
month10	0.010035192367724900
month9	0.009990893658950280
geoNetwork.city.Mountain.View	0.009968953504494780
channelGroupingDirect	0.009866436434018360
month11	0.009645468308047570
month2	0.009459846034086680
device.deviceCategorydesktop	0.00935930483950086
month3	0.009244953048351720
month1	0.008770172433012680
trafficSource.isTrueDirect	0.008018883169427940
geoNetwork.city.San.Francisco	0.00749693974991283
totals.bounces	0.007295910176536710
device.operatingSystem.Windows	0.006896190273860070
daySunday	0.006699842841971440
browser_Chrome	0.006448941998670820
daySaturday	0.006402258167402020
geoNetwork.city.Sunnyvale	0.006157045848036090
device.operatingSystem.Chrome.OS	0.0058751054026713200
device.operatingSystem.Linux	0.004957962981358960
totals.newVisits	0.004949797750999690
geoNetwork.continentAmericas	0.004789019678387970
trafficSource.adwordsClickInfo.isVideoAd	0.004566601067700350
browser_Safari	0.003934314556986880
geoNetwork.city.Los.Angeles	0.003596711231606730
geoNetwork.city.San.Jose	0.0035245190974578800