Paper 4050-2019

Google Store Revenue Probability Prediction

Meera Govindan, Purdue University; Rohit Kaul, Purdue University

ABSTRACT

We design and examine an analytics solution to make probability predictions of earning revenue per customer visit. The motivation for this study is to help marketing teams make better use of their budgets. Specifically, we wish to aid a firm to use their data as a guiding-tool for decision-making. Often the 80/20 rule has been proven right for many firms- A significant portion of their revenue comes from a relatively small percentage of customers. Therefore, marketing teams are challenged to make appropriate investments in promotional strategies. We use customer data from Google's Merchandise Store (G-Store) available on Kaggle along with SAS® University Edition to derive essential insights and generate models to predict the probability of earning revenue per visit. Our analysis shows that the G-store's revenue earning potential is the highest amongst the customers who visited the store 100-500 times since the mean revenues from this segment are the highest even though the number of such transactions was limited. The features highlighted in our model can be used by Google to increase the revenues from its existing customer base rather than expand its resources to acquire new customers, who may or may not make a substantial purchase per visit on its G-Store.

INTRODUCTION

To aid Google's marketing team to make better use of its marketing budget, we analyze a publicly available dataset. We use a combination of descriptive and predictive analytics techniques to create our solution.

Descriptive analytics helps us transform raw data into meaningful information. We use this technique to identify target customer segments and summarize and analyze the characteristics of customers belonging to this segment to aid Google devise better marketing strategies to cater to its target segment.

Predictive analytics is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data. We use this technique to build a model to predict the Google's likelihood of earning revenues from its target segment.

DATA

The publicly available data set used in this analysis came from Kaggle's Google Analytics Customer Revenue Prediction competition. The Data can be found here [1]. The train_v2.csv is a training set containing user transactions between August 1st, 2016 to April 30th, 2018. This data in JSON format was first parsed using Python's jsonlite package. Our code for parsing can be found here [2] The variables in the data set are described in Table 1:

Variable	Type	Description
Device	Character	The specifications for the device used to access the Store.
geoNetwork	Character	This section contains information about the geography of the user.
socialengagementType	Character	Engagement type, either "Socially Engaged" or "Not Socially Engaged"
totals	Numeric	This section contains aggregate values across the session.
trafficsource	Character	This section contains information about the Traffic Source from which the session originated.
visitID	Numeric	An identifier for this session. This is part of the value usually stored as the _utmb cookie. This is only unique to the user. For a completely unique ID, you should use a combination of fullVisitorId and visitId.
visitnumber	Numeric	The session number for this user. If this is the first session, then this is set to 1.
hits	Numeric	This row and nested fields are populated for any and all types of hits. Provides a record of all page visits.
visitstarttime	POSIX time	The timestamp (expressed as POSIX time).

Table 1. Data Dictionary and Data Types for Variables

PROBLEM STATEMENT

We build a model that would help predict the probability of a customer visit generating revenue. We use a logistic regression model to identify which visits have a high probability of generating revenue.

$$\operatorname{logit} p = \ln \frac{p}{1-p} \quad \text{for } 0 < p < 1.$$

The logit function is the link function in this kind of generalized linear model, i.e. we have logit(E(Y))=a+b*x.

DATA CLEANING AND VALIDATION

After converting the data from the json format to standard cell format, we arrive at a total of 58 features. From this, we drop the redundant features which we believe are not essential for predicting customer revenues for Google Store. Some of these features are: geoNetwork_networkLocation, device_mobileDeviceModel, device_language, device_flashVersion, trafficSource_campaignCode, socialEngagementType.

Mean Imputations: For features which have less than 50% missing values, we use mean imputation to fill in the missing values. For e.g. pageviews. For totals_bounces and totals_newvisits, since these were categorical binary variables, the missing values indicated 0. For e.g. 0 for totals_newvisits indicated no new visits. The missing values for variables are displayed in Table 2:

Variable	N	N Miss	% Missing
date	1,048,575	0	0
visitNumber	1,048,575	0	0
totals_bounces	536,702	511,873	48.8%
totals_hits	1,048,575	0	0
totals_newVisits	804,173	244,402	23.3%
totals_pageviews	1,048,427	148	0.014%
totals_sessionQualityDim	520,193	528,382	50.4%
totals_timeOnSite	510,185	538,390	51.4%
totals_transactionRevenuetotals_transactions	11,250	1,037,325	98.9%
	11,277	1,037,298	98.9%

Table 2. Missing Values for Variables

Drop columns with >50% missing values: For session quality and time spent on site features as missing values exceed 50% of the total values, we eliminated these columns.
Define new category for visitNumber feature: As visit numbers ranged from 0 to 500, we decided to examine the effect of these visits on the revenue generating capacity of G-Store. We thus, binned the visits into 5 categories – Less than 100, 100-200, 200-300, 300-400 and 400-500. We found that majority of the visits (99.81%) is in the "Less-than-100 category".
Extract Month feature: From the date column we extract the month to study the seasonality effects of revenue for the company.

ANALYSIS

As a first step in our analysis, we try to identify the target customer segment which Google can use to align its marketing strategy. To identify this target group (from which the maximum revenue is derived), we first choose the visit number metric. We consider the visit number as a key performance metric. This may help us identify whether the website's revenue increases with an increase in a customer's visit count. This may in turn help us identify whether the company should expand its customer base and target fleeting customers who may visit the website only a few times before purchasing and may not necessarily visit again or focus its strategies on converting the existing customer base into loyal customers (customers who view the website regularly). Thus, we conduct our analysis in the following steps:

1. Part 1: Descriptive Analytics: We identify the target customer segments and their characteristics
    - Choose visit number feature as indicator of target group.
    - Test the hypothesis that visit number affects the mean revenue.
    - If the hypothesis is true, we identify the most important target customer group as well as the important features of this group that are directly correlated with the revenues earned from those visits.
1. Part 2: Predictive Analytics: We build a model to predict revenue earning potential
    - Using the identified target segment, we build a predictive model that helps predict the probabilities of google earning revenues from a customer visit.
    - Finally, we suggest ways for G-Store to maximize revenues by identifying important

characteristics of this customer segment.

PART 1: DESCRIPTIVE ANALYTICS

VISUALIZATION

1. The visit number feature provides interesting insights (64.3% of the revenues for G-Store come from customers who have visited the store less than 5 times (most page hits (4.1M) come from this segment too), 18.7% from customers who have visited between 5 to 10 times, whereas only 2.0% come from those who have visited the store more than 300 times). However, the mean of revenue is highest for the >100 visits segment. Thus, although fewer customers visit the store >100 times, these are the loyal customers from whom Google derives the most per capita revenue and this segment could potentially be a good target segment. Figure 1: illustrates the graph between the Visit Number and Revenue.

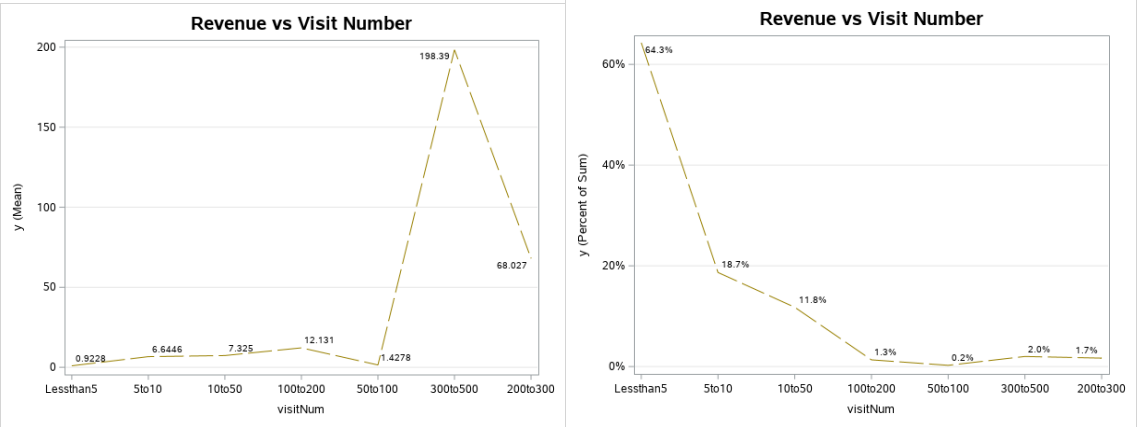

Figure 1: Visit Number vs Revenue

2. To test the hypothesis that visit number does affect revenue, we conduct one-way ANOVA followed by pairwise t-tests (results in appendix).
3. From the output, we can see that since the overall p-value is low and corresponding t-values are low as well, the visit number affects mean revenue significantly. Furthermore, the 200-300 visit number segment generated the maximum mean revenue for G-Store followed by the 300 to 500 segment and then the 100 to 200 group. Therefore, these 3 segments together become the ideal target candidates for our analysis.
4. From the Pearson correlation analysis, we find that visit number has a high correlation (0.34) with transaction revenue for the non-zero revenue segment whereas other numeric features have a less positive correlation.

Analysis of Characteristics of the 100-500 Visit Number (target) Segment

Revenue Seasonality: We observe that maximum revenues for the target segment (33% of the total) were derived in the month April (aggregate of two years). There is a sharp dip in the May turnover with only 0.2% of the total revenues coming in that month. Since the target segment revenues are maximum during the month of April, the company could offer additional

promotional strategies during this period. Figure 2: illustrates the seasonality in the Revenue.

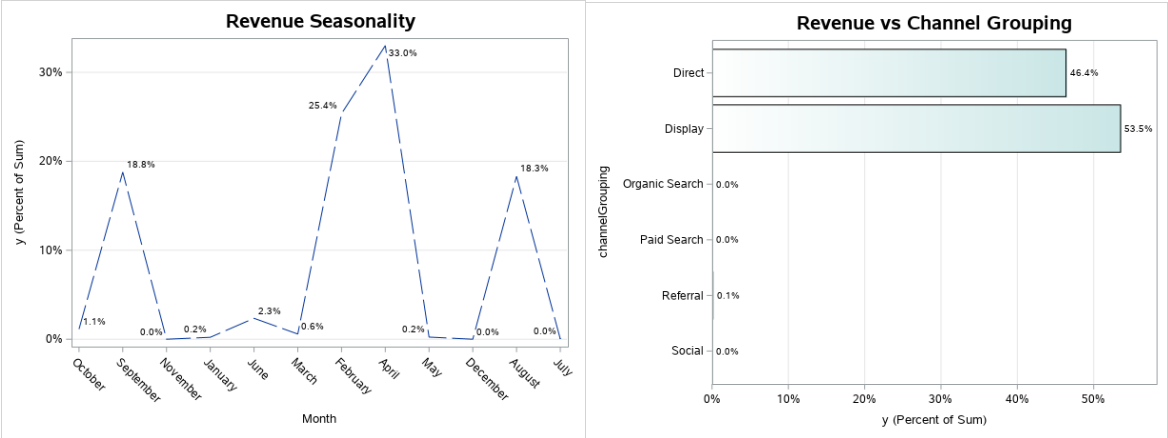

Figure 2: Revenue Seasonality (Left), Revenue vs Channel Grouping (Right)

Channel Grouping: Display (53.5%) and Direct (46.4%) results contributed to the maximum revenues for the business, whereas referrals and social categories hardly contributed to any business for G-Store. This may suggest probable target platforms for Google to advertise its products on. Figure 2: illustrates the graph between Channel Grouping and Revenue.

Revenue vs Device Browser: About 98.9% of the revenues were derived (of the target segment) from the Firefox browser despite the higher number of hits received from Google Chrome (63.6% of the G-Store visits originated from Chrome). This is mainly because mean revenue was the highest for Firefox at $167 as compared to only $0.6 from Chrome. Figure 3: illustrates the graph between Revenue and Device Browser.

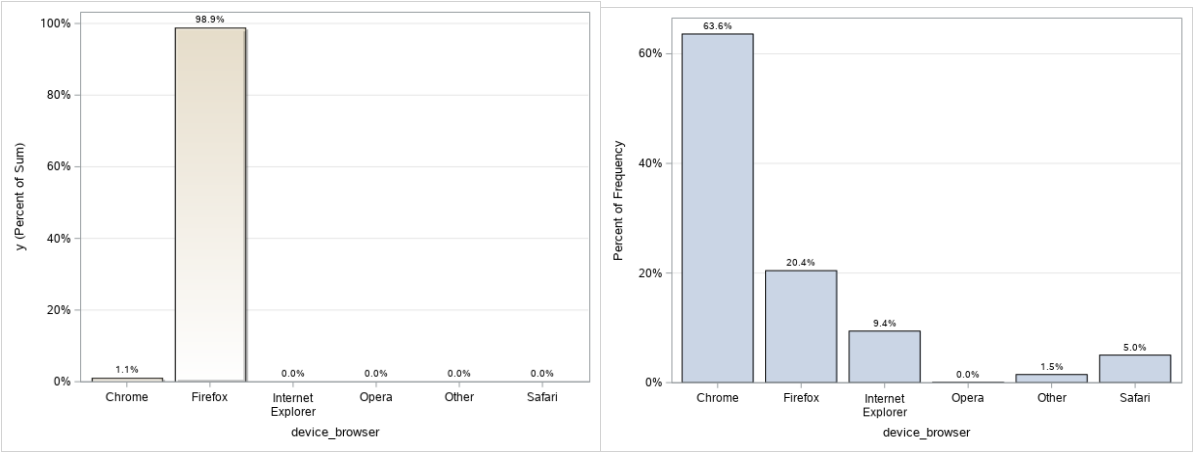

Figure 3: Revenue vs Device Browser

Revenue vs Operating System: If we take a closer look at the frequency of customer visits in this segment using various operating systems, we find that 61.5% use Windows OS followed by 24.6% for Macintosh. However, as far as mean revenues from these OS are concerned, Windows dominates other OS' with mean revenue per visit at $56 followed by just $0.14 for Macintosh. Thus, Google should look to target users using Windows OS as these are high yield customers for them.

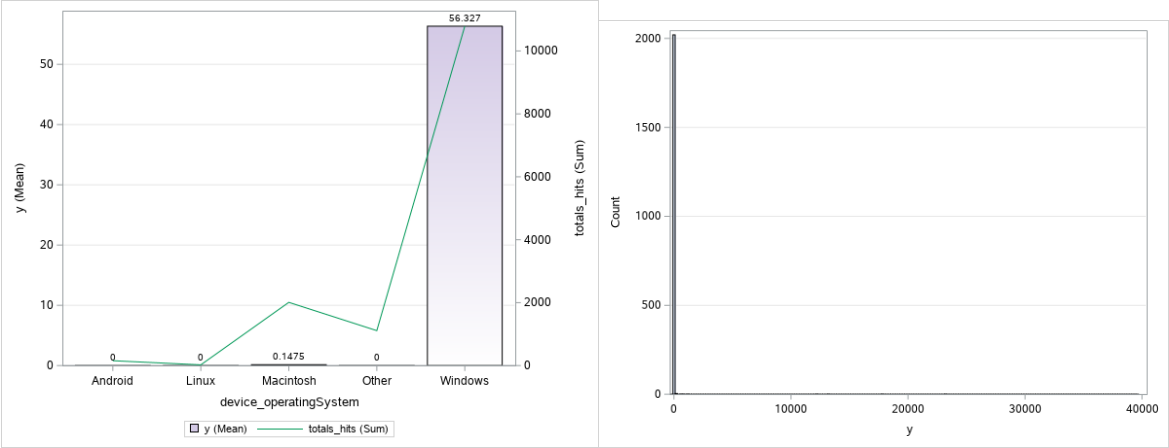

Figure 4: Revenue vs Device OS (Left), Revenue Frequency (Right)

PART 2: PREDICTIVE ANALYTICS

For the 100-500 segment, we find that the distribution of revenues is right skewed. Out of the 2031 visits, we find that 2007 generate no revenues for Google whereas just 24 visits are revenue generating.

MODEL FORMULATION AND INTERPRETATION

Post the data cleaning and validation step, we consider the following features to include in our model

1. Visit Number
2. Geonetwork continent
3. Device Operating System
4. Device Browser
5. Month of Visit
6. Total Pageviews
7. Total Hits
8. Total Bounces

For the analysis, we used a probit model. We formulate the response variable as "0" if the customer does not generate revenue and "1" if the customer does generate revenue. From the model results, we find that only Month and Visit Number are significant in explaining variation in the probability of visits earning revenues.

GENERALIZATION

From this analysis, we identify those visits and subsequently customers who have high chances of generating revenues for Google Store. While our methodology and analysis may seem specific to G-store, it may be extended to analyze customer visit data for any retailer. It is important for retailers (online or physical stores) to possess insight about their potential target segments, customer demographics so that they may strategize and utilize their marketing budgets accordingly.

FURTHER STUDIES

As part of future studies, we can look to expand the model to study how Google can look to acquire and work on new customers (visits less than 5) and understand what metrics it will need to study/alter during the endeavor. We also plan to identify the top products that these customers who visit the store the most often buy and promote those products more often than the others to maximize revenues.

CONCLUSION

Using the insights derived from our analysis, we suggest G-Store to target existing customers who visit their G-Store website between 100-500 times as the mean revenues earned per visit from these customers is the highest. This can be achieved by implementing the following strategies –

We observe that windows OS users form the bulk of the revenue that google generates, hence there should be more emphasis laid on targeting Windows OS users by purchasing ad slots on Microsoft store or on Firefox browsers on these Windows PC's. Maximum revenue was achieved in the month of April which suggests that google should target this month and allocate more advertising budget during April (than at any other time of the year). Even though, G-Store derives 97.7% revenues from the American Continent, Africa shows the highest revenue earning capacity per user as its average revenue per user is 436% higher than Oceania while it is less for all other continents. Thus, G-Store should target users in Africa. The features highlighted in the model can hence be used by Google, to increase the revenues that it can gain from its existing customer base rather than expanding its resources to acquire zero revenue customers (or new customers) who may or may not make a substantial purchase on its G-Store.

APPENDIX

The FREQ Procedure

visitNumber	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Less than 100	1046544	99.81	1046544	99.81
100-200	1543	0.15	1048087	99.95
200-300	346	0.03	1048433	99.99
300-400	105	0.01	1048538	100.00
400-500	37	0.00	1048575	100.00

Pearson Correlation Coefficients, N = 11250				
	visitNumber	totals_hits	totals_pageviews	transactions
Transaction_revenue	0.34830	0.13015	0.11902	0.04013

Logistic Regression Output

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	262.750	222.255
SC	268.367	390.744
-2 Log L	260.750	162.255

R-Square	0.0473	Max-rescaled R-Square	0.3929

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	98.4951	29	<.0001
Score	115.3182	29	<.0001
Wald	31.6051	29	0.3375

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
totals_bounces	1	0.0401	0.8412
totals_hits	1	3.8171	0.0507
totals_pageviews	1	2.4832	0.1151
device_browser	5	4.7877	0.4423
device_deviceCategor	1	0.0002	0.9879
device_operatingSyst	4	2.9761	0.5618
geoNetwork_continent	3	1.4610	0.6913
Month	11	22.6113	0.0200
visitNumber	2	10.0569	0.0065

One-Way ANOVA Outcome

Levene's Test for Homogeneity of y Variance ANOVA of Squared Deviations from Group Means					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
visitNumber	6	1.624E15	2.707E14	558.43	<.0001
Error	1.05E6	5.084E17	4.848E11		

Welch's ANOVA for y			
Source	**DF**	**F Value**	**Pr > F**
visitNumber	6.0000	37.74	<.0001
Error	1439.2		

Pairwise t-test (100-200 and 200-300 visit segments shown here)

visitNum	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
100to200		1543	12130629	4.546E8	11572901	0	1.786E10
200to300		346	68026994	1.2435E9	66850147	0	2.313E10
Diff (1-2)	**Pooled**		-5.59E7	6.7199E8	39972392		
Diff (1-2)	**Satterthwaite**		-5.59E7		67844485		

visitNum	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
100to200		12130629	-1.057E7	34830916	4.546E8	4.391E8	4.7123E8
200to300		68026994	-6.346E7	1.9951E8	1.2435E9	1.1572E9	1.3438E9
Diff (1-2)	**Pooled**	-5.59E7	-1.343E8	22498368	6.7199E8	6.5122E8	6.9414E8
Diff (1-2)	**Satterthwaite**	-5.59E7	-1.893E8	77517659			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1887	-1.40	0.1622
Satterthwaite	Unequal	365.92	-0.82	0.4105

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	345	1542	7.48	<.0001

Final table showing visits having >20% chance of generating a revenue:

channel	visitNun	device_	device_	device_	device_	geoNetv	totals_b	totals_h	totals_n	totals_p	trafficSc	y	transact	date2	visitNun	Month	logy	date_ne	custom	FROM_	INTO_	IP_0	IP_1
Social	100to200	Chrome	desktop	FALSE	Windows	Americas	0	301	0	166	1	0	0	02Aug2016	100to200	August	0	AUG16	0	0	1	0.224	0.776
Display	300to500	Firefox	desktop	FALSE	Windows	Americas	0	42	0	30	1	1196.74	1	30Jun2017	300to500	June	7.088192	JUN17	1	1	0	0.617	0.383
Display	300to500	Firefox	desktop	FALSE	Windows	Americas	0	38	0	33	1	25.77	2	09Jun2017	300to500	June	3.287282	JUN17	1	1	0	0.664	0.336
Display	300to500	Firefox	desktop	FALSE	Windows	Americas	0	22	0	18	1	23.96	1	08Jun2017	300to500	June	3.217275	JUN17	1	1	0	0.683	0.317
Display	300to500	Firefox	desktop	FALSE	Windows	Americas	0	37	0	37	1	0	0	16Jun2017	300to500	June	0	JUN17	0	0	0	0.695	0.305
Display	300to500	Firefox	desktop	FALSE	Windows	Americas	0	18	0	16	1	0	0	08Jun2017	300to500	June	0	JUN17	0	0	0	0.698	0.302
Display	300to500	Firefox	desktop	FALSE	Windows	Americas	0	19	0	19	1	0	0	09Jun2017	300to500	June	0	JUN17	0	0	0	0.709	0.291
Direct	100to200	Chrome	desktop	FALSE	Windows	Americas	0	24	0	17	1	123.25	1	01May2017	100to200	May	4.822296	MAY17	1	1	0	0.710	0.290
Display	300to500	Firefox	desktop	FALSE	Windows	Americas	0	15	0	15	1	0	0	22Jun2017	300to500	June	0	JUN17	0	0	0	0.712	0.288
Display	300to500	Firefox	desktop	FALSE	Windows	Americas	0	13	0	13	1	0	0	20Jun2017	300to500	June	0	JUN17	0	0	0	0.714	0.286
Display	300to500	Firefox	desktop	FALSE	Windows	Americas	0	11	0	11	1	0	0	23Jun2017	300to500	June	0	JUN17	0	0	0	0.715	0.285
Direct	100to200	Chrome	desktop	FALSE	Windows	Americas	0	31	0	26	1	29.99	2	22May2017	100to200	May	3.433665	MAY17	1	1	0	0.716	0.284
Display	300to500	Firefox	desktop	FALSE	Windows	Americas	0	3	0	3	1	0	0	13Jun2017	300to500	June	0	JUN17	0	0	0	0.721	0.279
Organic Se	100to200	Chrome	desktop	FALSE	Windows	Americas	0	38	0	29	1	0	0	30Jun2017	100to200	June	0	JUN17	0	0	0	0.734	0.266
Direct	100to200	Chrome	desktop	FALSE	Windows	Americas	0	61	0	32	1	0	0	10Apr2017	100to200	April	0	APR17	0	0	0	0.743	0.257
Direct	300to500	Firefox	desktop	FALSE	Windows	Americas	0	386	0	386	1	0	0	05Jan2018	300to500	January	0	JAN18	0	0	0	0.748	0.252
Organic Se	100to200	Chrome	desktop	FALSE	Windows	Americas	0	25	0	18	1	0	0	30Jun2017	100to200	June	0	JUN17	0	0	0	0.754	0.246
Direct	100to200	Chrome	desktop	FALSE	Windows	Americas	0	10	0	10	1	16.99	1	03May2017	100to200	May	2.889816	MAY17	1	1	0	0.759	0.241
Display	100to200	Chrome	desktop	FALSE	Windows	Americas	0	2	0	1	1	0	0	23May2017	100to200	May	0	MAY17	0	0	0	0.759	0.241
Direct	100to200	Chrome	desktop	FALSE	Windows	Americas	0	3	0	3	1	0	0	15May2017	100to200	May	0	MAY17	0	0	0	0.764	0.236
Display	100to200	Chrome	desktop	FALSE	Windows	Americas	0	9	0	7	1	0	0	01Jun2017	100to200	June	0	JUN17	0	0	0	0.790	0.210
Organic Se	100to200	Chrome	desktop	FALSE	Windows	Americas	0	6	0	4	1	0	0	06Jun2017	100to200	June	0	JUN17	0	0	0	0.792	0.208
Organic Se	100to200	Chrome	desktop	FALSE	Windows	Americas	0	6	0	5	0	0	0	01Jun2017	100to200	June	0	JUN17	0	0	0	0.797	0.203

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Meera Govindan
+1-7657015262
gmeera26@gmail.com

Rohit Kaul
+1-7656378057
rhtk.1234@gmail.com