Optimizing Google's e-commerce ROI

Data Analytics and Statistics Project

Miguel Chacón

IronHack Barcelona

November, 2020

Table of Contents

Introduction

This technical report, and its accompanying presentation, serve as my project for the Statistics module for the IronHack DAFT course. In addition to fulfilling the requirements of the assignment, it is also meant to join my GitHub portfolio as proof of my knowledge of inferential statistical concepts, SQL, the BigQuery API, Tableau, and Python. Thus, I will follow-up on this project by creating a Jupyter notebook and *Medium* article to summarize and showcase my work.

Preamble

The project is based on a Kaggle dataset from the Google Merchandise store. My goal here is to analyze the dataset in order to optimize the e-commerce store's ROI.

Business Questions

1. What Markets should the Marketing Manager focus on scaling or improving?

2. Which channels and channel sub-categories should be scaled?

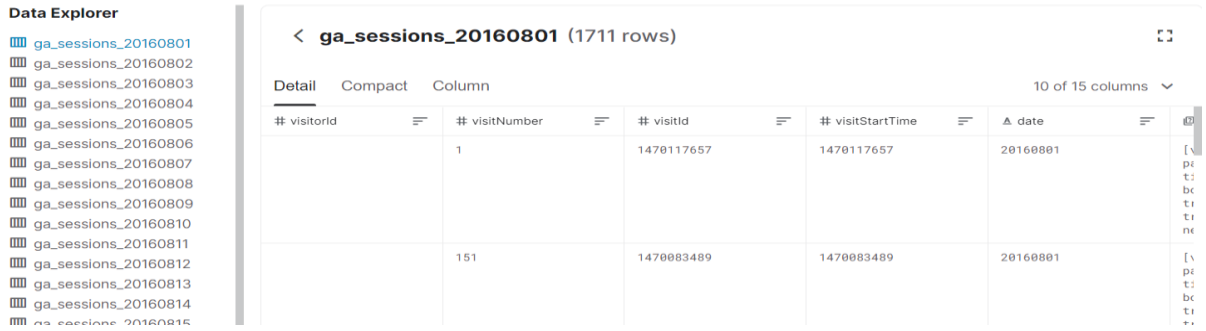
3. Are there any opportunities for cost-cutting?

The Project

The project is split into five parts. The first is understanding and exploring the data. Here, I seek to understand how the data was collected, what it represents, and whether it is clean and organized. Each of the subsequent stages is made up of an analysis of the data to answer each business question in order. Finally, in the fifth part, I present my findings and offer potential follow-ups for the project.

The data

The dataset contains Google Analytics metrics for the Google Merchandise Store collected between August 2016 and August 2017. The dataset contains 366 tables, each storing data for a single day. Within the tables, the unique key is a visit id. Each row then contains the metrics for a unique visit to the page. Note that a single user may have multiple visits and, thus, several rows in a single table.



Each table has 15 columns. The relevant ones are:

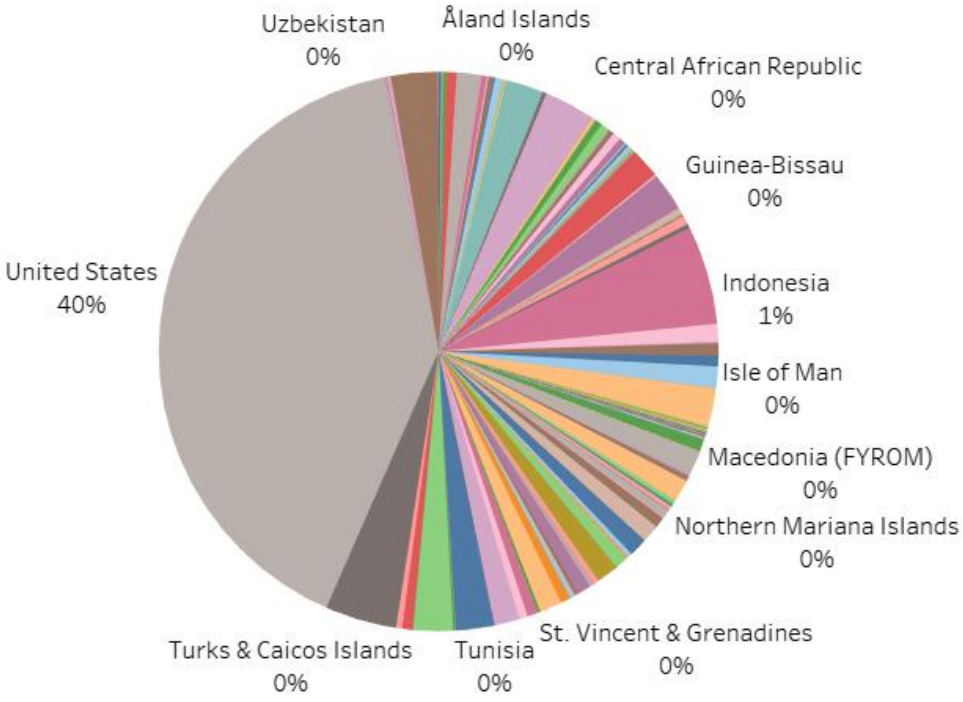
1. VisitNumber: This counts the number of times a user has visited the site.
2. VisitID: This should be a unique identifier for a given visit, and there should be no duplicates.
3. Date: The exact date of the visit
4. Totals: A dictionary of session metrics per visit.
5. Traffic Source: A dictionary of source categories per visit.
6. Device: A dictionary of device categories per visit
7. Geo Network: A dictionary of geographical categories per visit.
8. Full Visitor ID: A unique identifier for the user.
9. Channel Grouping: Denotes the channel category. It is similar to 'totals.medium,' with the key difference being that channel groupings can be customized by the media team, whereas the medium is automatically determined by Google Analytics.

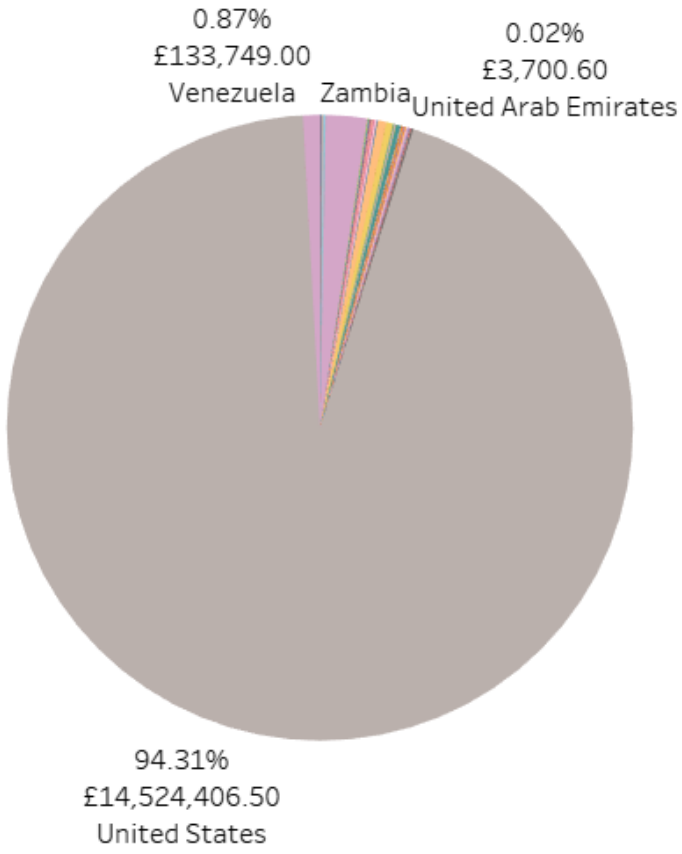
Moreover, though the columns need to be unnested, the data is clean, and the data types make sense.

Markets

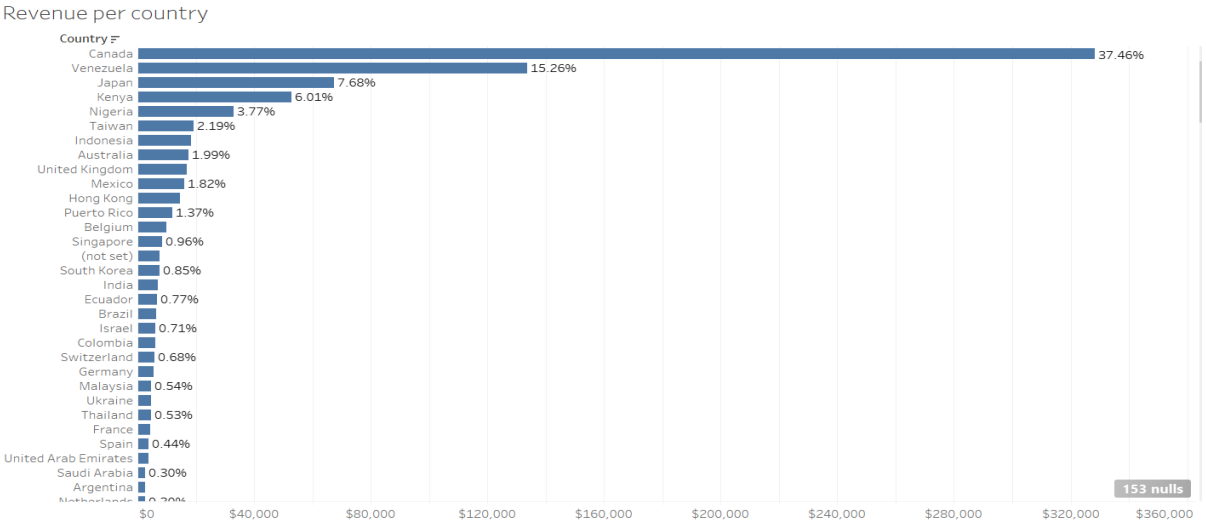
Markets are the first point of analysis. The data provider has made geographical data available down to the national unit. Information for regions and cities is unavailable.

We first explor the distribution of (1) visits and (2) revenue per country:





The United States is both the market with the greatest number of visits and the highest Revenue. Moreover, the country's percentage of total revenue far surpasses that of its total visits, indicating that it considerable revenue per visit when compared to other markets. We will establish the United States as our primary market and move on to explore secondary markets:



Though Canada dominates in this arena, other markets also contribute a considerable share of revenue. We will limit the scope of our analysis to the top 50% of secondary market revenue.

Revenue per country

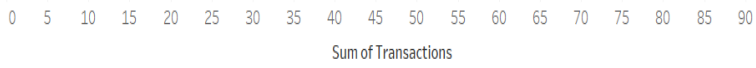
Coun..	% of Total Revenue Adjus..	Revenue Adjusted	Conversion Rate	ROI	Revenue per transaction	Visits
Canada	53.37%	328,245	0.77%	13	165	25,869
Venezuela	21.75%	133,749	7.18%	63	87	2,132
Japan	10.94%	67,290	0.09%	3	374	19,731
Kenya	8.57%	52,687	0.39%	68	1,756	771
Nigeria	5.37%	33,024	0.14%	23	1,651	1,446

The table above offers the following observations:

1. Canada and Japan have the largest share of visits. Nevertheless, Japan's ROI and revenue is far below Canada's.
2. Nigeria and Kenya offer an exceptionally high Revenue per transaction despite a low number of visits. These unusual numbers point to a small percentage of visitors making a high degree of the transactions. We hypothesize that users are buying in bulk, and suggest to the marketing team that they conduct a customer analysis for potential partner marketing opportunities.
3. Venezuela has a disproportionately high conversion rate, both in terms of e-commerce benchmarks and against other markets in this dataset, as well as the lowest revenue per transaction. We will analyze this market further by breaking down the channels:

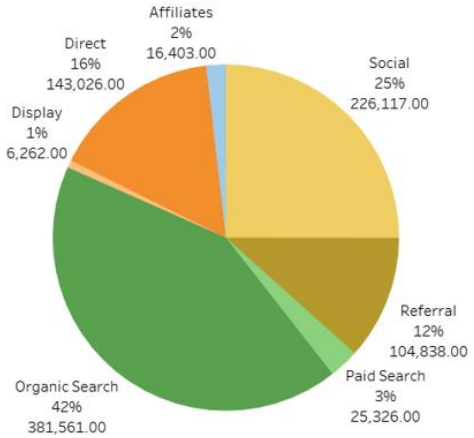
Venezuela channel breakdown

Channel Group..	Medium	Source	Campaign	Ad Conte..
Direct	(none)	(direct)	(not set)	Null
Organic Search	organic	google	(not set)	Null
Referral	referral	mail.google.com	(not set)	Null
Social	referral	facebook.com	(not set)	Null
		l.facebook.com	(not set)	Null



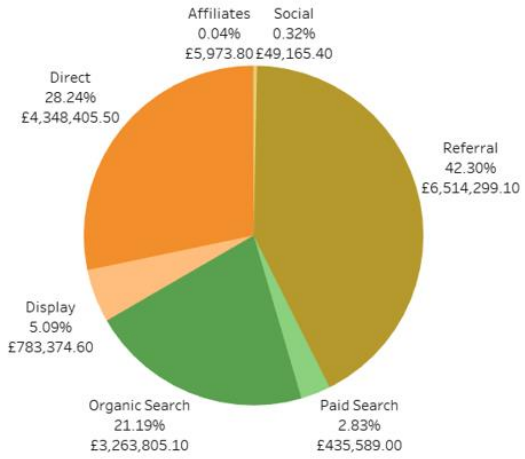
The table above shows that Direct and Organic search traffic lead in terms of transactions. Thus, purchases are attributable to consumers with previous brand awareness rather than to marketing campaigns. We will need to analyze channels for other countries in order to determine whether this is the benchmark, or whether Venezuela is an outlier.

Markets



Visits per Channel

Channel G..	Visits	Conversion Rate
Organic Search	382K	0.94%
Social	226K	0.06%
Direct	143K	1.55%
Referral	105K	5.29%
Paid Search	25K	1.89%
Affiliates	16K	0.05%
Display	6K	2.43%
(Other)	0K	0.83%



Revenue per Channel

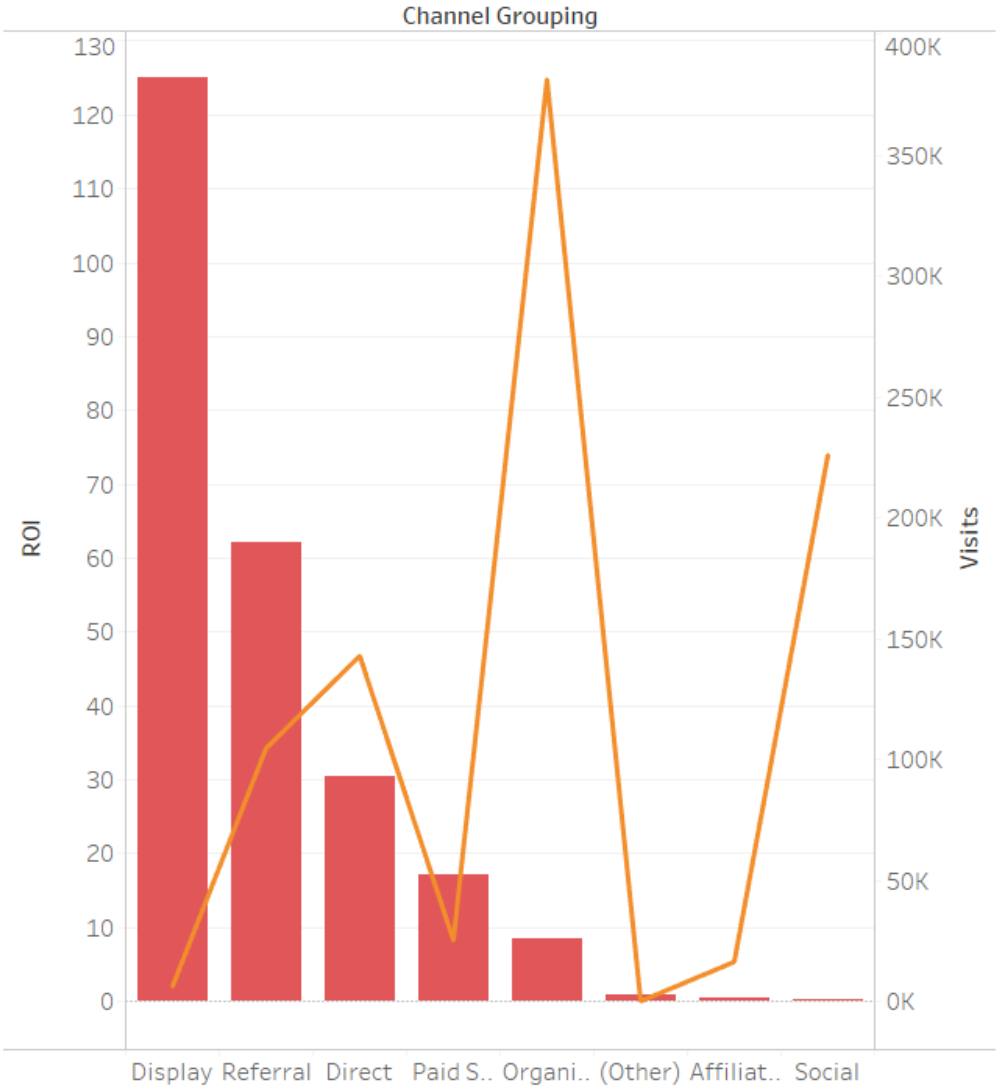
Channel G..	Revenue Adjusted	Conversion Rate
Referral	$6.51M	5.29%
Direct	$4.35M	1.55%
Organic Search	$3.26M	0.94%
Display	$0.78M	2.43%
Paid Search	$0.44M	1.89%
Social	$0.05M	0.06%
Affiliates	$0.01M	0.05%
(Other)	$0.00M	0.83%

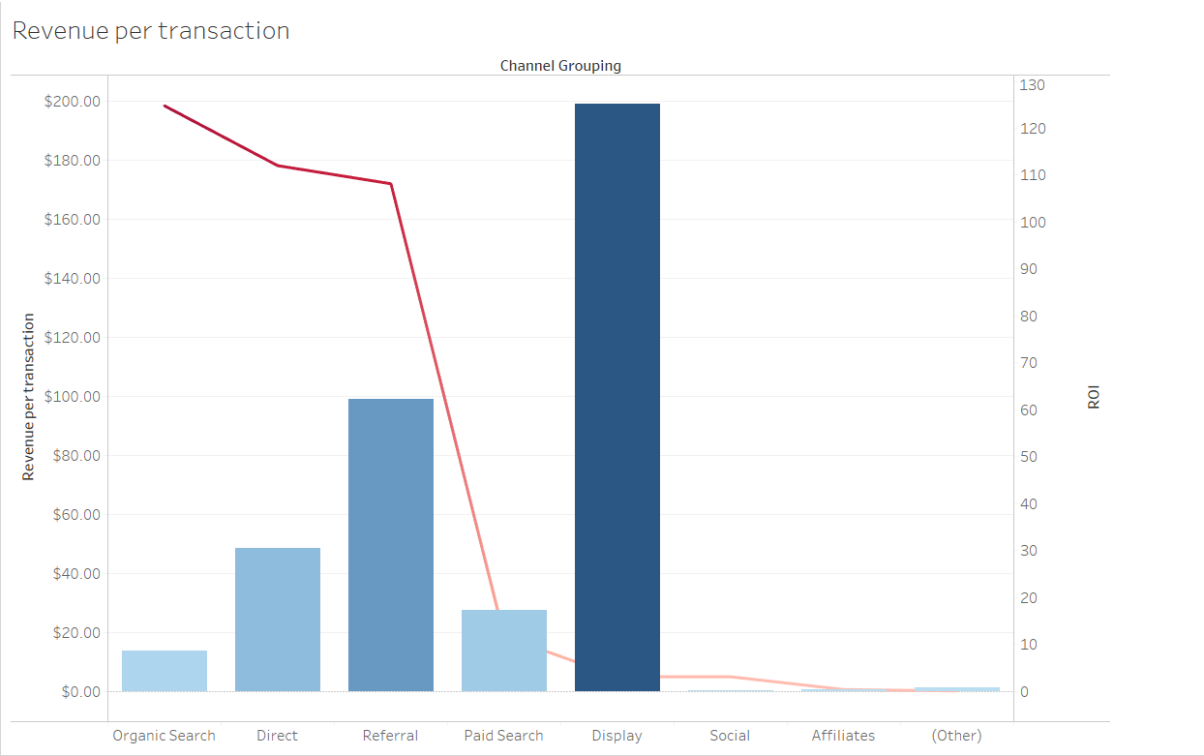
The graphics above show that, though Organic Search and Social lead in terms of traffic, Referrals and Direct lead in terms of revenue. We hypothesize that this is due to a larger conversion rate and conduct a z-test to confirm.

```
(52.755213923906915, 0.0)
(18.906751895295656, 1.0035062351924274e-79)
(-11.96888978106944, 5.1715988011352805e-33)
(2.71439867013498, 0.006639620919980044)
(56.24322498021343, 0.0)
(0.15788699004137305, 0.8745458499183396)
(-3.454800685445845, 0.0005506997682332232)
```

The following shows the z-statistic and p-value for each conversion rate per channel in descending order. The difference foe each with the exception of the penultimate two is significant. Now, we explore whether conversion rate is a good predictor of ROI, our optimization goal:

Visits per Channel

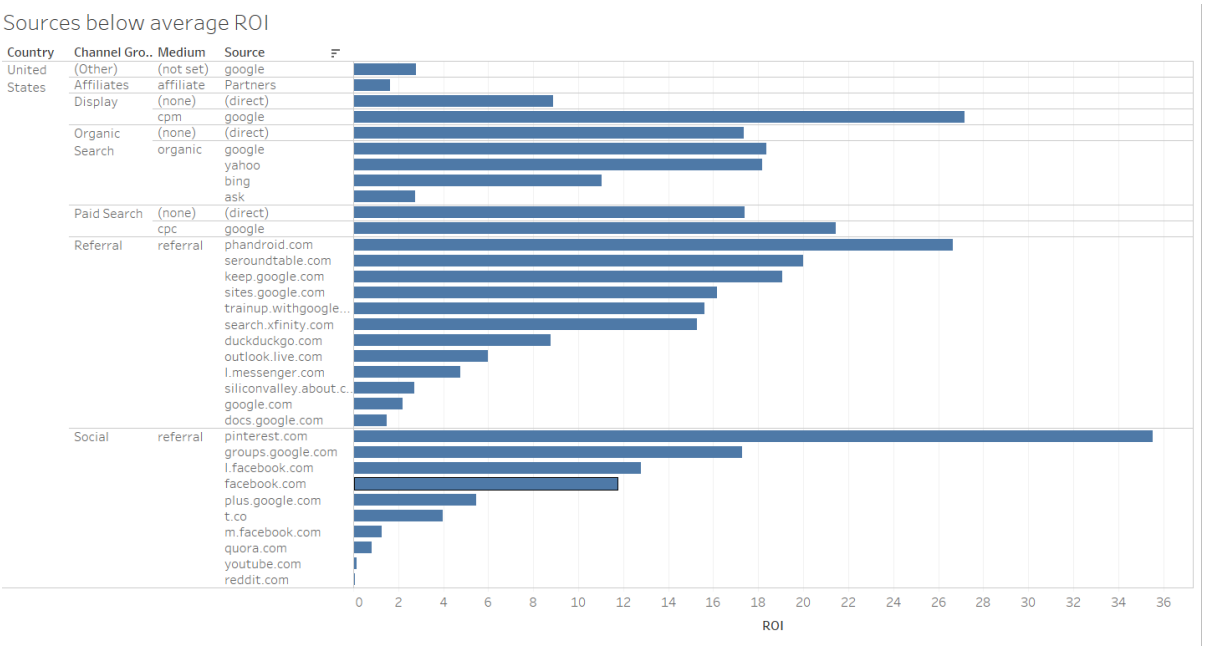




The second graph shows an interesting finding: Not only is a higher Roi characteristic of channels with a high conversion rate, but channels with high ROIs also tend to hae a large number of small transactions rather than a small number of large ones. This tells us that the marketing team should strive to convert visitors into transactions rather than on upselling visitors who do choose to make a purchase.

Cost Optimization

Finally, we conduct a short analysis of the source, or channel sub-categories, in the United States that have an ROI below the market's average.



Interestingly, there are a large number of affiliate sites performing poorly, despite the channel as a whole having a high ROI. Suggesting to the marketing team that they cut the sources above should lead to an improvement in overall ROI.

Conclusion

We recommend to the marketing team that they:

1. Maximize traffic from Display and Referral channels across markets.
2. Maximize high conversion rates and small transactions rather than upselling.
3. Research Kenya, Nigeria, and Venezuela for potential partner marketing.
4. Cut marketing spending in the listed channels for the US market.

As a follow-up a new project should be set in motion to benchmark conversion rates in terms of (1) time and (2) competitors and dedicate a budget to either cut or improve markets and further channel sub-groups that underperform. Moreover, a project to build and deploy a logistic regression model that predicts the likelihood of a visit turning into a transaction would add tremendous value, as the conversion rate is a excellent determinant of a category's ROI.