## Background

https://shop.googlemerchandisestore.com/

We are challenged to analyze a Google Merchandise Store (also known as GStore, where Google swag is sold) customer dataset to predict revenue per customer. Through various models, we are prediction whether a customer will make a purchase based on the features generated from user data.

Hopefully, the outcome will be more actionable operational changes and better use of marketing budgets for those companies who choose to use data analysis on top of GA data.

Link to Kaggle Dataset
https://www.kaggle.com/c/ga-customer-revenue-prediction/overview

## Goal

test_v2.csv, for their transactions in the future time period of **December 1st, 2018 through January 31st, 2019.**

## File Description:

train_v2.csv - the updated training set - contains user transactions from August 1st, 2016 to April 30th, 2018. **Less than 2 years**
test_v2.csv - the updated test set - contains user transactions from May 1st, 2018 to October 15th, 2018.

## Data Fields

- fullVisitorId- A 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.
- 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. 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 - The session number for this user. If this is the first session, then this is set to 1.
- visitStartTime - The timestamp (expressed as POSIX time).
- hits - This row and nested fields are populated for any and all types of hits. Provides a record of all page visits.
- customDimensions - This section contains any user-level or session-level custom dimensions that are set for a session. This is a repeated field and has an entry for each dimension that is set.
- totals - This set of columns mostly includes high-level aggregate data.

## Meaningful Columns after the Primitive Cleaning

*A column takes a large set of memories, therefore we deleted a

| Column Names | |
|---|---|
| Channel Group | How users accessed the online Gstore for swags |
| Date | |
| ID | |
| Social Engagement | Details unknown. Maybe refer to whether a user has left any comment. |
| Visit ID | An identifier for this session |
| Visit Number [1] | The session number for this user. If this is the first session, then this is set to 1. |
| Device Browser | Type of browser used<br>E.g. Chrome, Safari, etc |
| Device Category | Type of device used<br>E.g. Desktop, Laptop |
| Device Language | This information is not available |
| Device Mobile Device Branding | This information is not available |
| Device Mobile Device Model | This information is not available |
| Device Operating System | Type of operating system that the device is on |

[1] In tabulating statistics for Web site usage, a user session (sometime referred to as a visit) is the presence of a user with a specific IP address who has not visited the site recently (typically, anytime within the past 30 minutes). The number of user sessions per day is one measure of how much traffic a Web site has. A user who visits a site at noon and then again at 3:30 pm would count as two user visits.

| | |
|---|---|
| Device Screen Resolution | This information is not available |
| geoNetwork.city | According to the city & country, we can map data points |
| geoNetwork.country | According to the city & country, we can map data points |
| geoNetwork.continent | By country the tableau |
| geoNetwork.latitude | This information is not available |
| geoNetwork.longitude | This information is not available |
| Totals.Bounces[2] | Bounce rate is an Internet marketing term used in web traffic analysis. It represents the percentage of visitors who enter the site and then leave ("bounce") rather than continuing to view other pages within the same site. |
| Totals.Hits | Basic terminology |
| totals.newVisits | Basic terminology |
| Totals.Pageviews | Basic terminology |
| totals.sessionQualityDim | Basic terminology |
| totals.timeOnSite | Basic terminology |
| **totals.totalTransactionRevenue** | this is the total revenue (This is what we predict) |
| totals.transactionRevenue | revenue per transaction |
| Totals.Transactions | |
| Totals.Visits | |
| trafficSource.adwordsClickInfo.isVideoAd | |
| trafficSource.isTrueDirect | |
| trafficSource.medium | |
| trafficSource.source | |

[2] Bounce rate is an Internet marketing term used in web traffic analysis. It represents the percentage of visitors who enter the site and then leave ("bounce") rather than continuing to view other pages within the same site.

Now, based on at least three examples I can conclude that the field transactionRevenue corresponds only to the revenue resulting from one certain transaction during a visit, whereas totalTransactionRevenue represents the whole revenue resulting from the whole visit. That is why any aggregations of transactionRevenue on the user level represent some (random) fractions of the aggregate user revenue over a certain period of time, and if this is really the case, this number is unfortunately fairly difficult to model and senseless to predict.