Programming and Database Fundamentals
Final Project

EDA on Google Merchandise Store

Contributors:
Narendra Badam
Abhishek Kumar

December 5, 2019

1. Abstract

In sales and marketing there is a general conception that almost 80% of the revenue come from just around 20% of the customer base. This is known as the 80/20 rule and is a base assumption for many strategies. Hence, it is very imperative to make appropriate investments in advertisements and promotional activities which is precise and targeted. To target the correct set of population it is important to analyze and determine the traits of people who buy products versus those who don't.

Here we did some exploratory data analysis to unearth some of those features that implicitly determine the likelihood of transaction from the customer visiting the Google Merchandise Store (GStore).

2. Introduction

The dataset is part of the competition held by R Studio, the developer of free and open tools for R and enterprise-ready products, in collaboration with Google Cloud. The objective of the competition is to predict the revenue of customer for the next financial year for which we have been given training data for just one task. We have used several plots and diagrams to explore the datasets and identify patterns within it and discover traits that are characteristic to one of the two classes i.e. revenue generating customers vs non-revenue generating customers.

3. Data

Our dataset has 903653 rows and 12 columns. Some of them are Json dictionaries which contains more specific information regarding the customer store visit. We extracted these key-value pairs from Json dictionaries and created new columns against each key type. The original parameters are listed and explained below:

1. ***fullVisitorId*** - `A unique identifier for each user of the Google Merchandise Store.`

2. ***channelGrouping*** - `The channel via which the user came to the Store.`

3. ***date*** - `The date on which the user visited the Store.`

4. ***device(Json)*** - `The specifications for the device used to access the Store.`

5. ***geoNetwork(Json)*** - `This section contains information about the geography of the user.`

6. ***sessionId*** - `A unique identifier for this visit to the store.`

7. ***socialEngagementType*** - `Engagement type, either "Socially Engaged" or "Not Socially Engaged".`

8. ***totals(Json)*** - `This section contains aggregate values across the session.`

9. ***trafficSource*** - `This section contains information about the Traffic Source from which the session originated.`

10. ***visitId*** - `An identifier for this session. This is only unique to the user.`

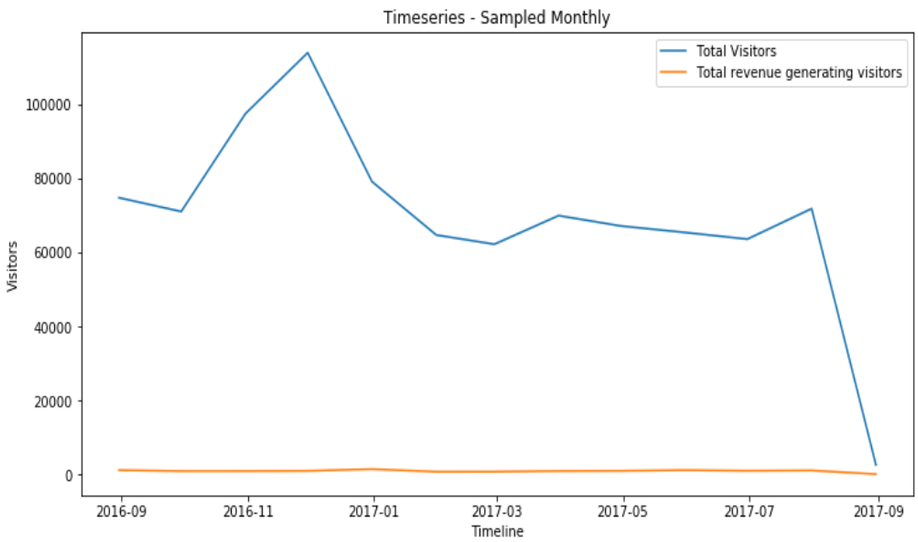
11. ***visitNumber*** *-* `The session number for this user. If this is the first session, then this is set to 1.`

12. ***visitStartTime*** *-* `The timestamp (expressed as POSIX time).`

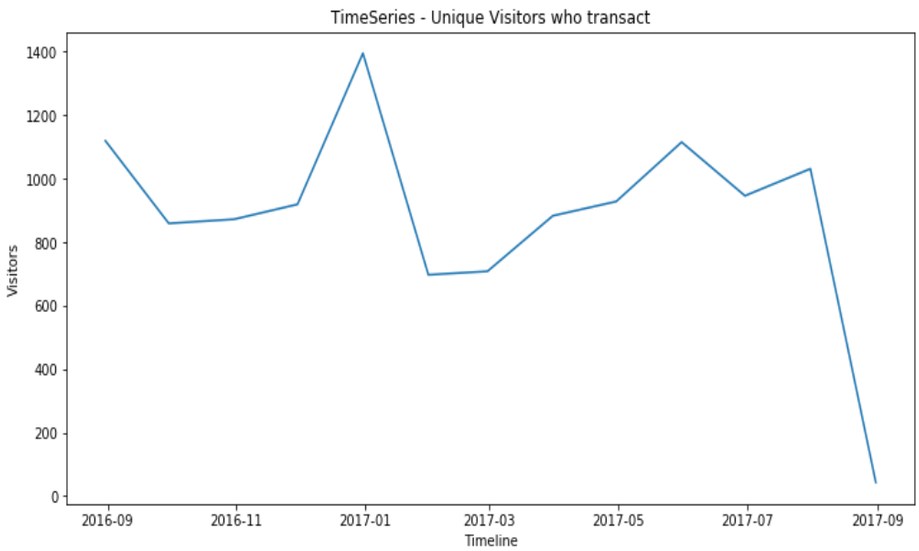
4. Analysis

This a time series dataset with one year of training data from Sep 2016 – Sep 2017. So, let's look at some of the plots.

A. Timeseries for total visitors and number of revenue generating visitors



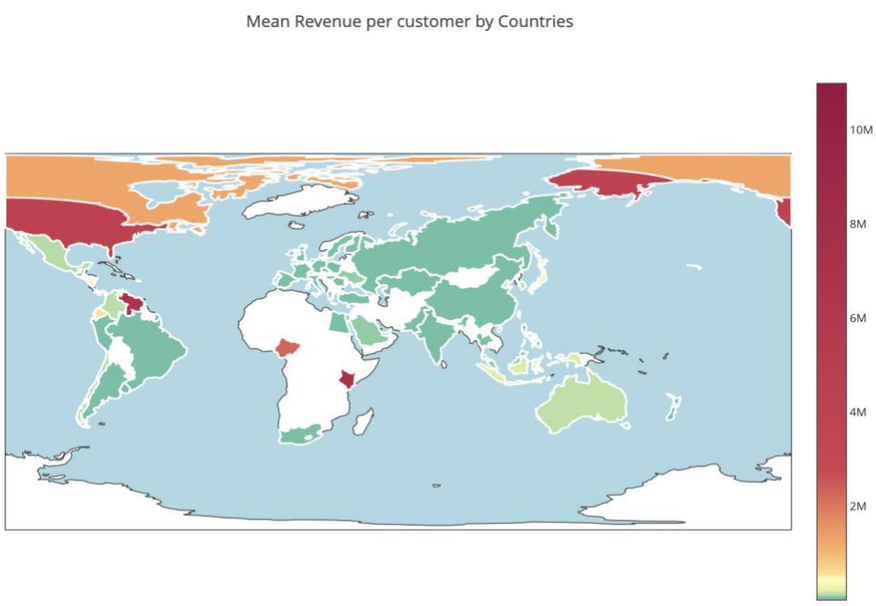
B. Timeseries for unique visitors who transact on visit



Inference : From the 1st time-series plot we can infer that the total revenues generating visitors are very less compared to total number of unique visitors to the store. It approximately ~1.4%

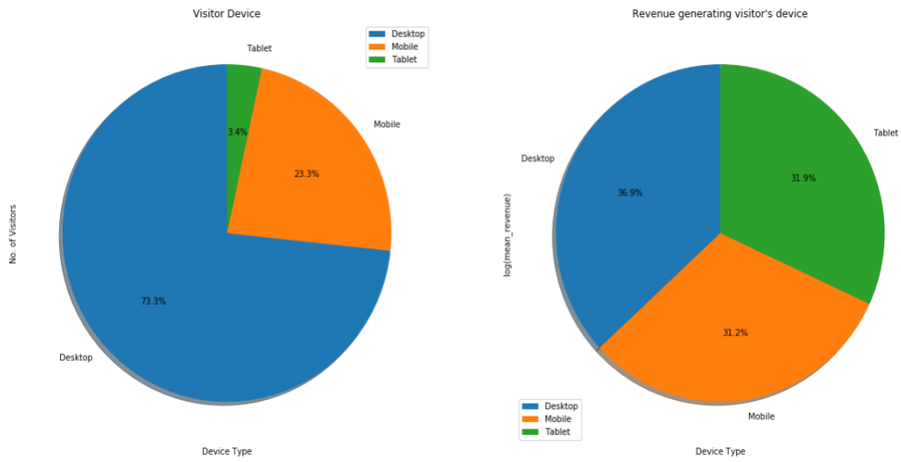
C. Visitor Count by countries

D. Mean Revenue per customer



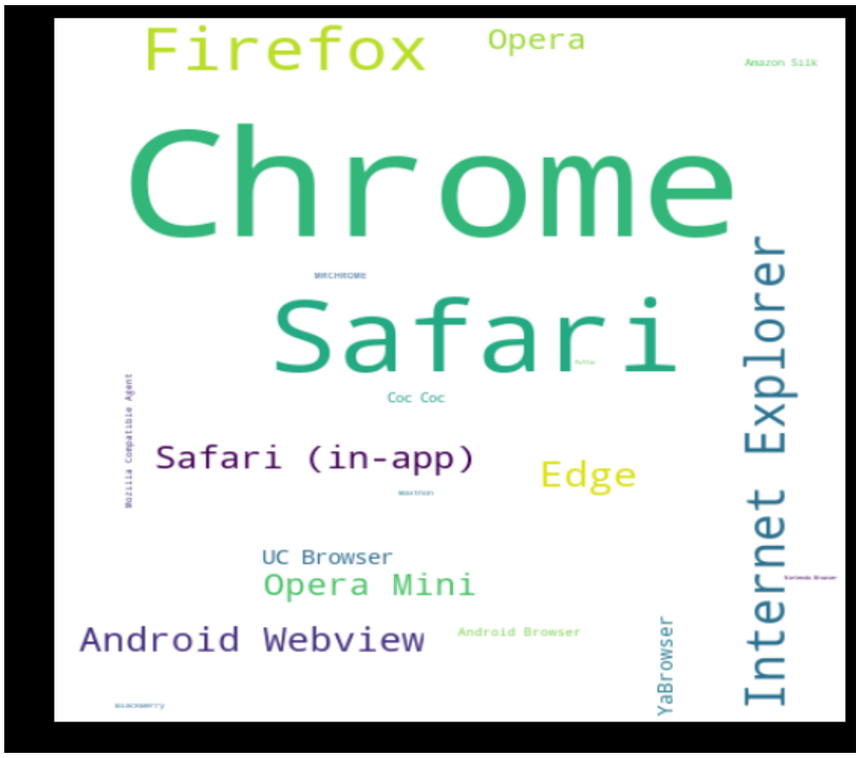
Inference : From the above geography plot we see that although most of the visitors come from US, India, Brazil but high mean revenue is generated by visitors from US, Venezuela, Nigeria, Kenya. The client should spend more in these countries as it may result in more revenue.

E. Visitor device and revenue generating visitor device



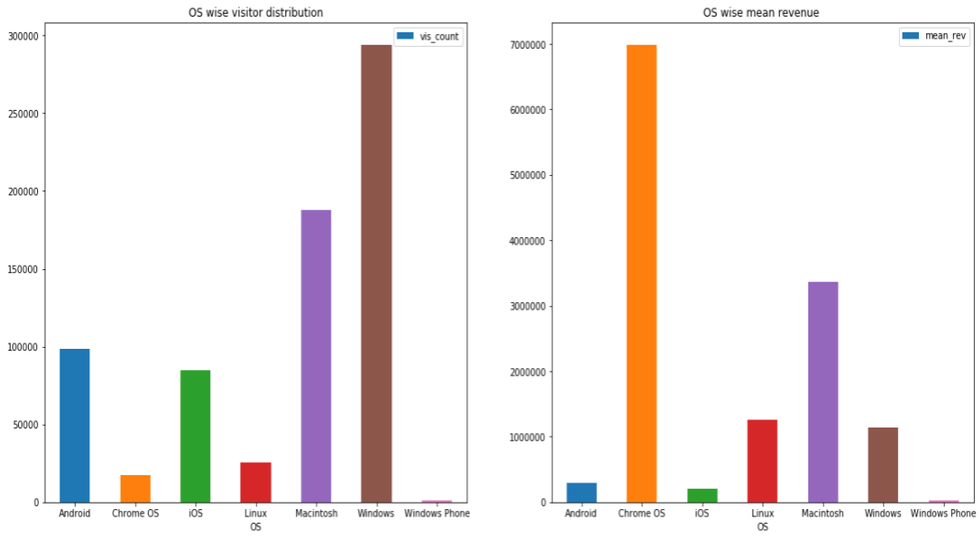
Inference : Although very less proportion of visitors use tablet but the high proportion of those visitors transact and generate revenue.

F. Browser Information of visitors



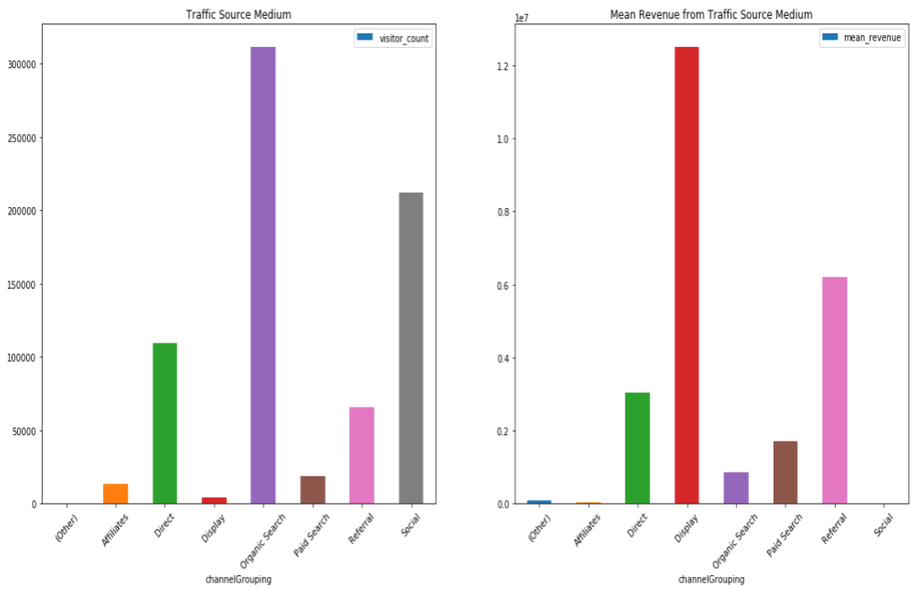
Inference : High proportion of users use Chrome and Safari browser.

G. OS wise visitor and mean revenue distribution



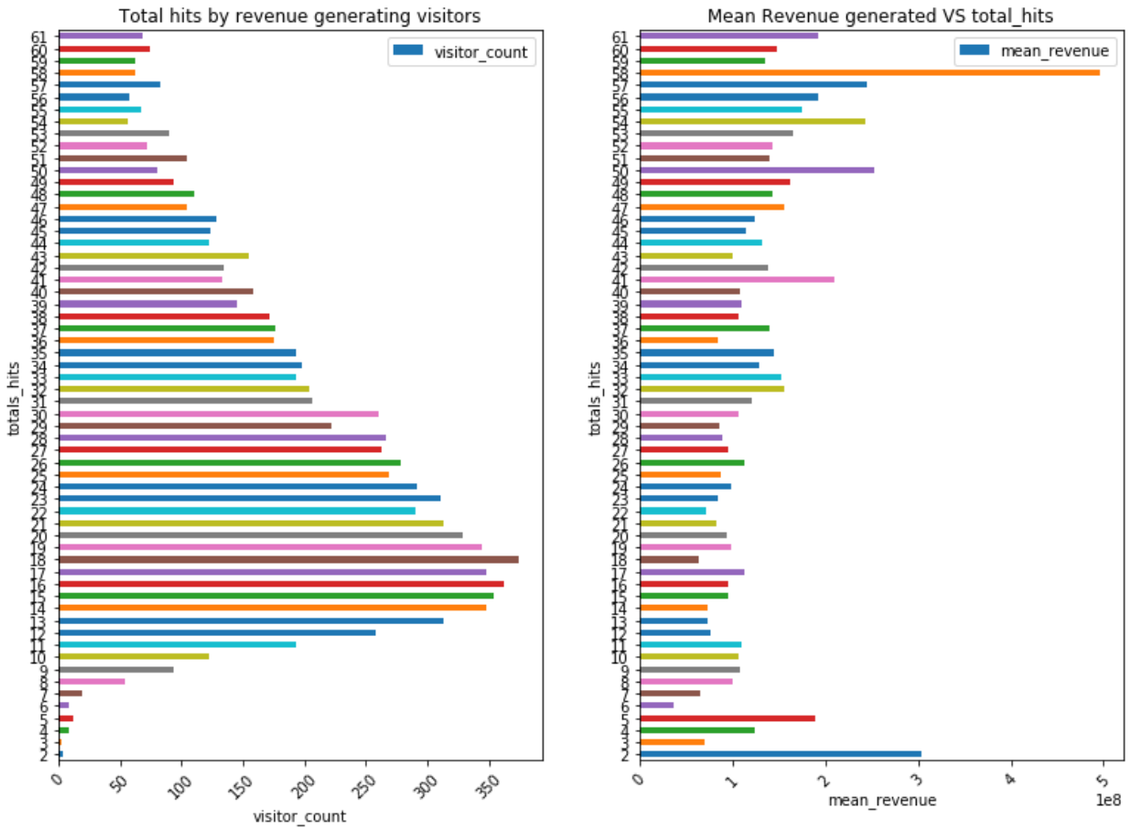
Inference : Chrome OS users generate very high mean revenue than other users although most of the users don't use chrome indicating chrome users are loyal Google customers.

H. Traffic Source



Inference : Most of the users land on the website from organic search but most of the revenue is generated by visitors who land via display ads and referrals.

I. Total hits by revenue generating visitors



Inference : From the left plot we can see that most of the revenue generating customers have total hits in range 10-40 while on the right plot of mean revenue generated there is no definite pattern. It indicates that total hits cannot be used to predict which customer would generate high revenue

5. Future Research Direction

This exploration can help managers to take necessary steps to increasing the efficiency of their marketing and advertising campaign. We could implement neural networks to predict the revenue generated by customers. This method would give high accuracy as we have around 50 variables and neural networks tend to work better on high dimensional dataset than other algorithms. But using neural networks we may lose on the interpretability of the model. I think it would be more logical to use regression trees algorithms like XGBoost which also work better for high dimensional datasets without compromising much on interpretability of the model.