9 Oct

-----------------------

Received the data.

Data size = 690.5 mb

dim(citi\_data)  
[1] 18532355 5

> head(citi\_data)  
 Cust\_map Merch\_Map\_final TXN\_MTH SPND\_CATGY NumTrans  
1 1 1 201309 HOUSEHOLD & UTILITIES 2  
2 1 1 201310 HOUSEHOLD & UTILITIES 2  
3 1 1 201311 HOUSEHOLD & UTILITIES 3  
4 1 1 201312 HOUSEHOLD & UTILITIES 2  
5 1 1 201401 HOUSEHOLD & UTILITIES 2  
6 1 1 201402 HOUSEHOLD & UTILITIES 3

> range(citi\_data$TXN\_MTH)  
[1] 201309 201408

Hence the data is given for the 12 months.

> colSums(is.na(citi\_data))  
 Cust\_map Merch\_Map\_final TXN\_MTH SPND\_CATGY NumTrans   
 0 0 0 0 0

No missing data

10th Oct

----------------------------------------------------------------------------------

Strategy for data :

Take first 100 customers, work with them for building the model.

Cust = 0.3M; Merch = 10K; Records = 18M; Transactions = 65M

----------------------------------------------------------------------------------

(Jaideep - I need customer list arranged by no. of transactions similar to merchant list given by Gautamji)

1. Take first 1500 merchants by total no. of transactions.
2. Or take top 100 merchants by total no. of transactions from each category.

We’ll get 1500 merchants in this way. Or take a weighted no. of merchants from each category depending on no. of merchants in each category divided by total no. of merchants.

1. For each customer take a weighted no. for category of buying based on no. of transactions in each category divided by total no. of transactions.
2. Suggest top Merchants from each category sequentially based on outputs of points 2 & 3 after discarding merchants with whom customer has already transacted.
3. The top 1500 merchants (as per serial no.) are roughly the top as per no. of transactions. (Barring very few ~10 outliers)

11th october

Have seen dependency of the 3 integral variables (cust no, merch no, trans no) with 1st categorical variable (buy category). 2nd categorical variable (trans month) dependency has to be done yet with aggregate.ts function.

Uploaded plots hence removed pasted image here.

Uploaded Summaries file of 1st categorical variable.

Strategy for weight < 10% (<1 out of 10) for any category: Discard the suggestion

Gautam: 11th october.

Have done the code for creating a table, which shows category wise weightage for each customer. Tested for 500 customer it took 8 minutes to do this for 500 customers.

Code as below:

-----------------------------------------------------------------------------------------

## category weight for each customer = no of transactions in one category

##divided by total number of transaction by that customer

All\_Cust\_numbers <- unique(city\_data$Cust\_map)

test\_customers <- subset(city\_data, city\_data$Cust\_map < 500 )

no\_of\_cust <- nrow(test\_customer\_numbers)

## list of categories

category\_list <- unique(city\_data$SPND\_CATGY)

cust\_cat\_weight <- matrix(,nrow = no\_of\_cust,ncol = no\_of\_cat, byrow = TRUE)

## test with 6 categories

no\_of\_cat = length(category\_list)

## for each customer, calculate number of transactions in a particular category divided by

## total number of transactions for that customers.

for (i in 1:no\_of\_cust) {

records\_of\_cust <- subset(city\_data,city\_data$Cust\_map==i)

total\_trans\_by\_cust <- sum((records\_of\_cust$NumTrans))

for( j in 1:(no\_of\_cat)) {

records\_cust\_cat<- subset(records\_of\_cust,records\_of\_cust$SPND\_CATGY==category\_list[j])

no\_trans\_cust\_cat <- sum(records\_cust\_cat$NumTrans)

cust\_cat\_weight[i,j] <- no\_trans\_cust\_cat / total\_trans\_by\_cust

}

}

##create a data frame from this matrix.

cust\_cat\_weight\_df <- as.data.frame(cust\_cat\_weight)

## write this data frame into a CSV file.

write.csv(file = "cust\_category\_3.csv",x=cust\_cat\_weight\_df)

View(cust\_cat\_weight)

-----------------------------------------------------------------------------------------------------------------------

12th Oct

Binning carried out - generated lot more insights. top 10 merchants from each category identified. 2 important observations :-

1. data seems to be time (quarter) independant.
2. market share for any merchant is <26%. This means no merchant is “globally” popular. this hints towards customer clustering based on common merchants. But don’t know about implementation feasibility as of now.

Jayanta completed the recommendation code and is working well. We have to check it for at least 100 customers.

I’ll generate data of how many merchants each customer has and see if customers can be grouped in some manner.

13 Oct

Approaches not taken.

1. Find Similar customer to a particular customer: (clustering/bin based on customers)Take a customer, find customers which spends similar to it. Suggest the merchants of those customers for the customer.

Problem is that there are 3 lac customers. If we try find the nearest customers, then we will need to do some sampling. (can’t compare each with everyone). If we do the sampling then it could be biased and would induce some errors.

Even if we can do computation for complete dataset, then again it will be difficult to decide which particular merchants to suggest from these.

2) Find Similar customers to a particular merchant: Take a merchant, cluster the customers for each individual merchant. And suggest the merchants with which these customer buys.

The problem is how to incorporate this relation for each customer. A customer has many merchants, and he can’t take suggestion from such a large number. The relationship is therefore not clear in prediction.

4) Cluster/Bin based on months: We could have clustered based on months creating 15\*12 bins. But we need to provide some weightage to the first quarter as we need to predict for next 3 months. So effectively we are given data for 4 quarters and we need to predict for the next quarter.

Thus a detailed time series analysis doesn’t seems to be appropriate. (¼ th of the total time duration we need to predict). However we should give some importance to transactions in the first quarter of data. There could be cyclicity in terms of annual payments. At the same time we also can not neglect the rest of the quarters as they can also have some cyclicity.

Approach taken:

# Given: Each merchant belongs to only 1 category.

Total 15 categories.

We will take top 100 merchants of each category. Will will be based on the number of transactions of the complete population.  
  
We will take a particular customer. We will find how much is the proportion of number of transactions in each category. Based on this proportion we will assign weights to each category.

For the transactions which were done in the first quarter, we will assign more weights as there is more chance of these transactions in the upcoming quarter.

So for each quarter we have decided to assign the weights in ration 2:1:1:1

Once we have the weightage of these customers, then we will multiply the weightage by 10. We will round of this number and take those many top merchants from the particular category. We thought about whether to use ceil/floor/round. The round function seems to be most appropriate as ceil/floor could result in overfitting/underfitting.

**TO DO :**

**Check if there is any outlier in the merchants. While considering the top merchant we also need to check the merchant had sufficient frequency of transactions. it shouldn’t happen that a merchant did transaction only 5-10 times, and the number of transaction is 10000 each and this makes him top merchant.**