Digital Marketing Analytics – Individual Homework 1

Author: Siow Meng Low
Date: 5th May 2017

Query to identify the top 5 customer locations by average spend

The “average spend” can be defined in two ways:

1. Average spend per order, breakdown by locations

For this definition, the total spending is firstly aggregated over the entire location (differentiate locations by using first 3 digits of ZIP code) and then divided by the total number of orders in that region to provide the average spend per order. The SQL query to perform this can be found in the submitted file, *IndividualHW1-AveSpendPerOrder.sql*. The top 5 customer locations are as below (note that *GeoCode* refers to the first 3 digits of ZIP code):

GeoCode	AveAmtPerOrder
823	950.430
202	268.725
36	265.534
788	260.675
586	234.200

2. Average spend per customer, breakdown by locations

For this definition, the total spending is firstly aggregated over the entire location and then divided by the total number of customers in that region to provide the average spend per customer. The SQL query to perform this can be found in the submitted file, *IndividualHW1-AveSpendPerCustomer.sql*. The top 5 customer locations are as below:

GeoCode	AveAmtPerCustomer
823	1188.038
511	904.307
36	840.858
873	708.530
811	554.476

Loading and Cleaning Data

To load the data into PostgreSQL database, *COPY* command is used. Before issuing the *COPY* command, *CREATE TABLE* is used to create the four table schemas. For the *Summary* table with over 100 fields, the column name was copied to a text editor so it’s easier to construct the *CREATE TABLE* command. Indexes are also created for the tables to enable faster query.

The SQL query to perform the above can be found in the submitted file, *IndividualHW1-Loading.sql*. To summarise in plain words, the loading and cleaning performed for each table are:

Summary Table

- The *Cust_ID* column is the unique identifier and used as the primary key, as well as the index.

- Some customers have missing values in fields such as first 3 digits of ZIP code, age group, interests etc. These will be populated as *NULL* in the database since we cannot be sure what the correct values are for each unique customer.

Contacts Table

- The *Cust_ID* column is defined as a foreign key (in reference to the *Summary* table).
- There is no missing value for this table.
- The (*Cust_ID*, *ContactDate*) combination are not unique since a customer can be contacted more than once in a day.
- *Cust_ID* column is used as index for faster table query.

Orders Table

- The *Cust_ID* column is defined as foreign key (in reference to the *Summary* table).
- There is no missing value for this table.
- In the data, the same *OrderNum* is reused for different customer orders and hence each row (i.e. each unique order) can only be uniquely identified by the combination (*Cust_ID*, *OrderNum*, *OrderDate*). This combination is hence used as the primary key.
- *Cust_ID* column is used as index for faster table query.

Lines Table

- The combination of three columns (Cust_ID, OrderNum, OrderDate) is defined as foreign key (in reference to the *Orders* table).
- Only *Gift* and *RecipNum* columns have empty values. *Gift* is unknown for store purchases, whereas *RecipNum* is empty when *Gift* is unknown or when it is not a gift item. Both of these two columns are populated as *NULL* in such situations.
- *Cust_ID* column is used as index for faster table query.