Econometrics 673 Fall 2022 Homework #3

Due in Canvas by midnight Friday, October 7th

How Far do My Customers Travel

The purpose of the analysis in this exercise is to perform a type of geographic market analysis commonly used in U.S. antitrust matters. Starting from a specific physical location (that is, a geographic sales point) how large of an area is necessary to capture 75% (or 90%) of sales for a location based on the distance customers travel to a location.

From the class website, download the "HW3_Data.zip" ZIP file. There are two files in the compressed zip file, "hw3_data.sas7bdat" which is a SAS dataset, and "Location Lat Long Data.xlsx". The first few rows of the SAS dataset should look like the following:

VIEWTABLE: Hw4.Hw4_data

	location_id	customer_id	latitude	longitude	quantity	date
1	1001	1101057303678972	34.049572	-84.584294	1	2012-06-01
2	1001	1101057303554777	33.956095	-84.50519	1	2015-11-21
3	1001	1101057303554777	33.956095	-84.50519	2	2015-07-01
4	1001	1101057303554777	33.956095	-84.50519	1	2015-08-01
5	1001	1101057303554777	33.956095	-84.50519	1	2014-03-28
6	1001	1101057303554777	33.956095	-84.50519	4	2014-08-08
7	1001	1101057303554777	33.956095	-84.50519	2	2014-08-24
8	1001	1101057303554777	33.956095	-84.50519	3	2014-04-13

There are 27 unique values for "location_id" and in total there should be 5,217,038 observations in these data. This data is transactions sales data for a retail establishment based on a loyalty card program. The variable "Customer_ID" is a unique identifier for a specific loyalty card member. The variables "latitude" and "longitude" are the physical latitude and longitude of customer's home address; "quantity" is the quantity purchased; and "date" is the transaction date (stored as a SAS date value).

The first few rows of the "Location Lat Long Data" look like:

	location_id	ref_latitude	ref_longitude
1	1001	34.00723	-84.571644
2	1002	33.554955	-84.329283
3	1003	33.690267	-84.097121
4	1004	33.97305	-84.077583

1. To determine how far customers travel to reach the specific retailing location from which these data were recorded, merge the "Location Lat Long Data.xlsx" data into the transaction data by location_id. (See the Lecture 6 PP slides for accessing Excel data with a LIBNAME statement, or the Lecture 7 slides for using PROC IMPORT.) The "ref_latitude" and "ref_longitude" variables are the physical latitude and longitude for the retail location – location_id; latitude and longitude from the sales data are the customer's home location. Now use the GEODIST function to calculate the straight-line distance (measured in miles) traveled by each customer to the retail location:
    a. Miles=GEODIST(latitude, longitude, ref_latitude, ref_longitude, 'M');
2. Next use accumulator variables and by-group processing to aggregate (sum) the quantity variable for each location_id/customer_id combination and also get an overall total for each location_id.
    a. Create a subset data set where, for each location_id, each observation on customer_id is unique and includes the customer level aggregate of the quantity variable.
    b. Create another subset data set of these data that includes the location_id and the overall total quantity for each location_id.

3. Using the distances (Miles), create a distance class variable for purposes of constructing frequency distributions and cumulative distributions (counts and percentages) for each location_id to show the proportion of sales associated with travel distances in half mile increments up to 100 miles. For example:
    a. Dist_class=0;
    b. If 0<miles<=0.5 then dist_class=0.5;
    c. If 0.5<miles<=1 then dist_class=1;
    d. Etc. (A DO loop will make this much easier)
4. Next, for each location_id use an accumulator variable and by-group processing to aggregate (sum) the quantity variable for each distance class (dist_class).
5. Merge in the overall total sales for each location_id created in 2.b to calculate the percentage of sales attributable to each distance class and create a cumulative percentage sales variable.
6. Given your analysis, for each of the 27 locations, how large must the draw area be to account for at least 75% of sales? At least 90% of sales?
7. What is the average distance traveled per unit (quantity) sale? Is this a simple average or weighted average? Which do you think is more appropriate?

Prepare a (short) summary report that includes a summary table like that below and answers to the questions posed above. Clearly identify any assumptions you must make in conducting your analysis. Upload your report in either a PDF or word processor format to Canvas using the file name 'Lname Fname UIN HW4'

Location ID	Average Distance Traveled	Distance Category Cut-off for 75% of Sales	Distance Category Cut-off for 90% of Sales
1001			
1002			
1003			
1004			
1005			
1006			
1007			
1008			
…			

Some coding notes: Keep in mind that to use by-group processing within a DATA step, the input data set must be sorted in the order of the BY variables. It is possible to complete all of this analysis using SAS coding. Nonetheless, it may be easier to complete the last step or two in Excel. Suppose you have a SAS data set in your WORK library named h5. You can create a new Excel workbook and write a SAS data set that workbook using a LIBNAME statement and DATA step such as the following:

```
libname hw3out excel "D:\ECMT673\HW3\HW3_Output.xlsx";
data hw3out.summary;
set h5;
run;
libname hw3out clear;
```

The last LIBNAME statement is needed so that you can open the new workbook in Excel without closing SAS.