GROUP (UP TO 3 PEOPLE) WEB ANALYTICS ASSIGNMENT #3 (30 POINTS)

DUE DATE: Wednesday, 6/16/2021

- This assignment will analyze the data (**HotelClickStream.xls**) and interpret the results. This dataset includes clickstream data of online transactions for online hotel booking. Appendix includes the detailed description for the variables.
- **Please follow the instructions very carefully to do this assignment! Please copy/summarize your *key* results for each question to a word file along with your answers to produce the final report for submission.**

1. Please first create the following 2 additional variables into your data

1) ***REF_D*** (create a dummy variable indicating whether the transaction was referenced from other website, if not, the final booking website was directly accessed. If no information provided for the variable *REF_DOMAIN_NAME*, *REF_D* = 0; otherwise *REF_D* = 1)
2) ***LOG_PRICE*** (take the log transformation of the variable *PROD_TOTPRICE* using the LOG function in excel)

a) (**5 points**) Please provide a summary table showing the top 10 *domain names (DOMAIN_NAME)* that generated the most volume of transactions the report should look like the following Table (**Hint: one way to do this is to use the COUNTIF function in excel**). **Please summarize briefly your observations from the results.**

Rank	Domain Names	# of Transactions

b) (**5 points**) Please provide a summary table showing the top 10 *reference domain names (REF_DOMAIN_NAME)* that generated the most volume of transactions the report should look like the following Table. **Please summarize briefly your observations from the results**.

Rank	Reference Domain Names	# of Transactions

2. (**6 points**) Please use the Binary Outcome (Logistic/Logit) regression technique to answer the question on "*what are the factors that influence people's decision on whether to book directly on a hotel website or from other third party website?"* Please use *DIRECT_D* as your Dependent Variable (DV); and *REF_D, LOG_PRICE, TRANS_FREQ, DURATION, HOUSEHOLD_SIZE, CHILDREN_D,* and *CONNECTIONSPEED_D* as your Independent Variables (IV). Please report and *interpret* your regression results, which should include the interpretation of the regression coefficients.

3. **a)** (**6 points**) Please use the Count Data (Poisson) regression model to answer the question on "*what are the factors that influence people's booking frequencies?"* Please use *TRANS_FREQ* as your DV; and *REF_D, LOG_PRICE, PAGES_VIEWED, HOUSEHOLD_SIZE, CHILDREN_D*, and *CONNECTIONSPEED_D* as your IVs. Please report and *interpret* your regression results, which should include the interpretation of the regression coefficients.

b) (**6 points**) Please repeat the analysis in question **a)** using the Negative Binomial Regression model. Please report and *interpret* your regression results and coefficients.

c) (**2 points**) Please summarize your observations by comparing the results from **a)** and **b)**.

Appendix

- **This is a sample data selected from a large online clickstream dataset. The data was collected by tracking over 100,000 unique household online shopping behavior. This small sample data includes transactions for booking hotels online.**

Variable Descriptions

Variable	Description and Measure
ID	Unique transaction ID
DOMAIN_ID	Unique ID for the web domain
MACHINE_ID	Unique ID for the computer (household) on which the transaction was made
SITE_SESSION_ID	Unique ID for the session in which the transaction was made
TRANS_FREQ	Total number of transactions for the household.
DOMAIN_NAME	The website (domain) name where the transaction was made
DIRECT_D	A dummy variable indicating whether the transaction is incurred directly from a hotel website (1) or other third_party travel website (0).
PROD_NAME	The product (e.g., hotel or packages) purchased by the household
PROD_TOTPRICE	Total price paid for this transaction
REF_DOMAIN_NAME	The referring website (domain) name through which the final purchase website was reached
DURATION	Total time spent at a site (mins)
PAGES_VIEWED	Total pages viewed at a site
HOUSEHOLD_SIZE	Total number of people in the household
CHILDREN_D	A dummy variable indicating whether the household has any children.
CONNECTIONSPEED_D	A dummy variable indicating whether the household has high speed internet connection