Problem Statement

Download the file OnlineRetailPromotions.xlsx from Canvas. The variables are explained in the worksheet titled “Data Definitions.” The file contains "big data" from 64,000 customers of an online retailer who purchased from that retail website within the last twelve months. These customers are randomly chosen by the retailer for an e-mail campaign: a third of these customers received an e-mail campaign featuring Men’s merchandise, one-third received an e-mail campaign featuring Women’s merchandise, and the remaining third received no e-mail (control group). Customer’s purchase behaviors (whether they visited the retailer website, whether they made a purchase, and and how much money they spend) were tracked for two weeks following the e-mail campaign. Your job is to tell the retailer if the Men’s or Women’s e-mail campaign was successful, and under what circumstances.

Note that this is a classic retail analytics problem. The two most frequent types of retail analytics problems are pricing and promotion analyses, and you are looking at a promotional analysis problem. Also what you have here is a simple experimental design with three groups: two treatment groups (men's and women's promotion) and one control group (no promotion). If you work in a data analytics position, you will be asked to design online experiments like this and analyze their results. Essentially, this assignment is a good indicator of how suitable you are for a data analytics job. Of course, in the real world, you won't have a clean Excel spreadsheet with 12 columns, but more likely a messy dataset with multiple tables that you may have to join using SQL commands, which may result in a merged data set with hundreds of columns, most of which will not be pertinent to this exercise!

There are three dependent variables of interest here: number of visits to the website, number of conversions, and amount spent. As you can see from the data, the number of visits and conversions are binary variables that requires logistic regression (you can try them if you are comfortable with logistic regression, or if not, wait until our logistic regression class to try them out). In this assignment, we will just examine the amount spent. Obviously, if someone receiving the promotion did not visit the website, they cannot be converted into a customer or spend money on the website. So the first step in analysis may be to limit the analysis to converted customers only. There are also other ways to address the problem without subsetting the data. But it is important to understand the data generating process before you can analyze it meaningfully. (Note: If I asked this question in an exam, I will exclude this paragraph and expect you to figure out from common sense that not every person receiving the promotion visits the website, not every visitor converts, etc. I also won't tell you what regression model to run. Making these choices is your responsibility as a data analyst.

1. Examine the "spend" variable that we want to predict and explain step-by-step what you would do to create a model to explain customer spend (bullet points are fine). What model(s) is(are) appropriate for this analysis and why. Run appropriate visualizations if necessary and document your work in your answer. Be sure to read Question 4 below to get a sense of the analysis the client is looking for.

2. Create a table of predictors for our dependent variable, listing all relevant predictors, the sign of their hypothesized effects, and a short 1-sentence rationale for each effect.

3 Run alternative models to test for the effects of the hypothesized predictors. Be sure to test the assumptions of these models and modify them as necessary. Present the best 3 models and their output

in a nice, compact manner. Also justify your choice of these models and include your assumptions testing results.

4. Based on your analysis, answer the following questions (using marginal effects, not statistical significance).

- How did the promotion campaigns work relative to the control group? Did the men's promotions work better than the women's promotion (or vice versa) and by how much?
- Should we target these promotions to new customers (who joined over the last 12 months) rather than to established customers, or vice versa?
- Should we target these promotions to customers who have a higher (or lower) history of spending over the last year?
- Did the promotions work better for phone or web channel?
- Will the promotions work better if the men's promotion is targeted at customers who bought men's merchandise over the last year (compared to those who purchased women's merchandise), and if the women's promotion would work better if targeted at customers who bought women's merchandise over the last year?

5. Reflect on the quality of your analysis, and comment on things you can do to further improve this analysis.