Case Study 1

Website Visit Duration

ABC corporation has built a website that collects articles on predictive modeling topics. The site has been active for one year and now ABC is trying to determine how certain aspects of their website affect web traffic. Specifically, ABC would like to build a predictive model that predicts how long a first-time visitor will stay when they visit the website. ABC has collected data from two different sources.

1. Information about certain aspects of the website. For example, if there is a news article posted on a specific day, the number of articles the webpage links to, and the number of advertisements ABC puts on the web page.
2. Information about site visitors. For about five months, ABC paid an external company to track visits to the site. The external company also tracks how long visitors stay, if they are first time or repeat visitors, as well as other demographic categories.

ABC wants to build a model that predicts the average time a new visitor stays on the website. Your task is to build a data set for a predictive model for this purpose. There are a few issues with the data:

- The data is in two different files, tracking.csv (data from the external provider) and site_daily.xlsx (daily configuration data collected by ABC). Both files are in the zip file downloaded earlier. ABC has split the site's daily configuration data into two sheets of an Excel spreadsheet, the first half of the year and the second half of the year. The externally collected tracking data covers only part of the year whereas the site daily data is every day in the year.
- The tracking data is recorded every day, but by each visit to the website. ABC needs to find the average visitor duration of all these visitors by day. Also, they think it might be helpful to include the number of visitors each day.
- The variable *headliner_topics* contains the topic on the first article linked on the site. Halfway through the year ABC changed how they were recorded. For example, "machine learning" in the first half of the year is the same as "ML" in the second half of the year.
- From March 30 through April 2 ABC used 70 advertisements on its page as an experiment to see if flooding a page with advertisements increased advertisement revenue. ABC does not plan on having that many advertisements again in the future.
- There are some missing values for when ABC posts a news article. On these days they forgot to record it and couldn't remember if a news article was put up or not. Missing values are indicated by blank cells.
- ABC wants the day of the week to be included as a predictor variable, such as "Sunday," "Monday," etc, but all dates are given in a different form.

ABC has asked you to prepare the data set for use in a predictive model. Your first task is to create that data set.

In addition to preparing the data set, ABC has asked you to provide thoughts on the following questions.

1. Why is it useful information to know that ABC does not plan on having future observations with 70 advertisements?
2. Should the variable for the number of visits on a day be used in the data set? Why or why not? It is likely a very good predictor if it is included!
3. The data ABC got from the tracking company includes the sex of the individual who visited the site. ABC is considering including that in the model somehow, or perhaps building separate models for each sex. Comment on any ethical implications for doing that.