DATA MANAGEMENT: PART 3

INTRODUCTION

This set of exercises will demonstrate how to import, clean, merge, and reshape a dataset.

Our goals for this lab are:

- Import data
- Subset data
- Create appropriate dtypes
- Create new variables as needed
- Merge data

DATA USED IN THIS EXERCISE:

We will use the following datasets in this exercise:

- data/phishing/Campaign.csv
- data/phishing/Lookup.csv
- data/phishing/User.csv

Concept of the data:

- A company has conducted a phishing campaign to test how susceptible their employees are to falling for a phishing attempt. Multiple campaigns were conducted. There were three types of campaigns:
    - (1) an email requesting users to enter data
    - (2) an email requesting users to open an attachment,
    - (3) an email requesting users to click on a link.
- The User table contains details about all the users (employees) in the system
- The Campaign table contains details about each campaign that was sent, including the time it was sent and the type of campaign it was.
- The lookup table contains the results of each campaign for the users it was sent to.

QUESTIONS

1) Continue the script you developed in the prior exercises.
2) Missing Data:
    a. Check each of the three tables for missing data. Do any of the tables have any missing data? Which columns have missing data?
3) Our goal is to combine these three files into a single dataframe we can use for analysis. **You do not need to implement this analysis yet**, you only need to create the dataframe that would allow you to do this analysis. This final dataframe should be able to answer questions like the following:
        i. What are the overall rates of failure (clicking the links, opening the attachments, or entering data)?
        ii. Is age associated with falling for the phishing schemes?
        iii. Is time of employment (how long you have worked at the company) associated with falling for the phishing schemes?
        iv. Is time of day associated with falling for the phishing schemes?
    b. To do this, we need to merge data from all three files into a single dataframe. You will need to think carefully about what steps to take to create this final analytic dataframe. You will need to do multiple subsets, merges, and create new variables. The final file should have a row for every User and campaign combination, with the following columns:
        i. UserID
        ii. CampaignID
        iii. CampaignType
        iv. CampaignSentMorningAfternoon (if campaign was sent in morning or afternoon)
        v. UserAge (User Age as of 2018-01-01)
        vi. UserTenure (User tenure at company as of 2018-01-01, how long have they worked here?)
        vii. UserOpenedEmail (Did the user open the email? Yes or no)
        viii. UserFail (Did the user fail by either clicking the link, Submitted info, or Opening the attachment? Should be values of 'Yes' and 'No')
        ix. UserReportedSpam (Did the user report the email as spam?)

