DATA MANAGEMENT 2

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, and (3) and 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 exercise, where you already imported this data.
2) Check the dtypes for each variable in each of the three tables, and modify it if you need to.
    a. Inspect each dtype
    b. Modify it as needed. For example, you need to convert each date to an actual date variable (or datetime), and you need to convert Gender to be categorical.
    c. For each string variable, do you want to turn it into a category? Why or why not?
    d. Calculate a new column in the user table, age as of 1/1/2018. You can use code like this:

```
# Calculate age in years as of 2015-01-01
df_1['Age_years'] = \
  ((pd.to_datetime('2015-01-01') - df_1['DateOfBirth']).dt.days/365.25)
```

    e. For the campaign table, we want to identify campaigns that were sent early in the morning versus later in the afternoon. Perhaps people are more susceptible to phishing attacks later in the day? To support this analysis, we need to create a new variable that is 'time of day'. You can use string formatting on the DateTimeSent to identify the time of day sent. Save this as a new variable
    f. Now discretize this variable into 'before 12 (noon)' and 'after 12 (noon)', and save that as a new variable.

