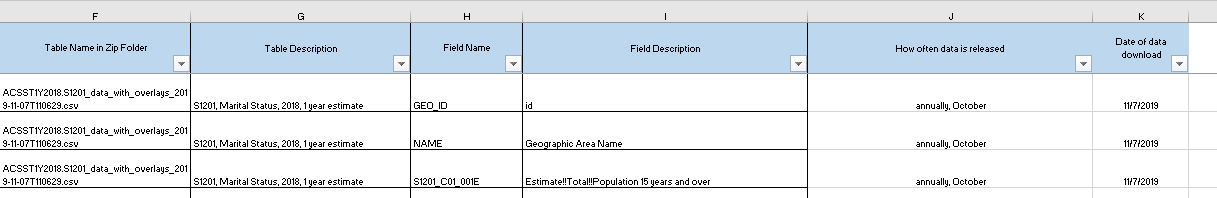
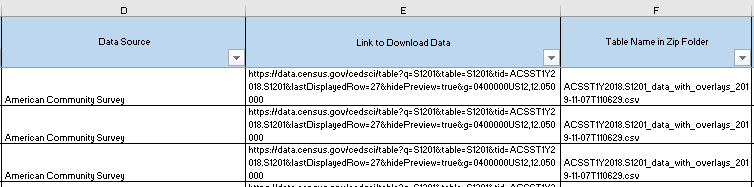
**Center for Advanced Analytics**

**Social Determinants of Health (SDOH) Data Integration Project**

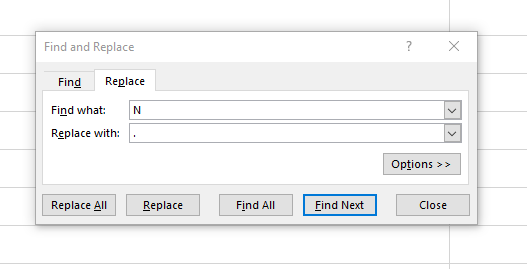
**How-To: Data Download & Clean**

Free, publicly available, SDOH data sources were download to the Baptist Health South Florida (BHSF) Data Warehouse. In the data dictionary, we have outlined when the data is available for download (e.g., annually). In this document, we outline how to check when the data is available, download the data, and clean the data.

1. **Check the data dictionary for data availability**
   1. Open the shared drive: CO CE Machine Learning🡪 Data Management🡪 SDOH🡪 Data Dictionaries🡪 SDOH\_Data Dictionary\_IT\_02\_25\_2020
   2. Scroll to the left of the excel file for the following two columns:
      1. “how often is data released”- this column describes when the data will be available for download in the following year. We suggest setting a calendar reminder (e.g., on October 31, 2020)
      2. “date of data download”- this column describes when the CAA downloaded the data
         1. Note: this column should be updated once new data is downloaded.
2. **Download the data** 
   1. In the same Data Dictionary, there is a column labeled “Link to Download Data”



* 1. Click the link
     1. This will bring you to the data for the row. For example, Marital Status for 2018. Once you’re at the website, you’ll need to update the date range to be specific for your data (i.e., 2019, 2020, etc.)
        1. For American Community Survey, they have helpful tutorials: <https://www.census.gov/data/academy/data-gems/2019/customize-download-tables.html>
        2. For American Community Survey, they will offer you a data dictionary of the fields and field descriptions. This is what you can use to copy and paste into our data dictionary. For the other data sets, data dictionaries may be available in the link. If they are not available, you can follow our naming conventions.
  2. Once the csv is downloaded, save to the share drive.
  3. Add to the data dictionary

1. **Clean the Data**
   1. Depending on the data source, “missing” data is presented differently (i.e., “N”, “(X)”, blank, “-“, “\*”, etc.). For consistency, all missing data has been replaced with a period “.”
   2. To replace:
      1. Open the data excel file
      2. Search for the missing data to understand what you need to replace (e.g., “N”)
      3. Select all the contents of the excel file and press “ctrl + F” to bring you to the “replace” field
         1. In the “find what” field, enter the missing data
         2. In the “replace with” field, enter a period
         3. Select “replace all”
      4. Repeat this step with all missing data (i.e., the same excel file may have (X) that needs to be replaced with a period)
   3. Save to the share drive