Week 2:

How do we typically clean data?

- Common data wrangling issues
- Resources for help with coding
- Commonly used pandas methods
- Creating a reproducible workflow

Coding Tasks:

1. Delete unecessary cells from notebook.
2. Verify that notebook runs error free from top to bottom.
3. Annotate steps with markdown.
4. Comment code as needed to clarify steps in a code block.
5. Look at the distribution of analysis values within the state of Tennessee for both `ha_costs_df` and `cancer_costs_df`. Does there appear to be a difference in these distributions for urban counties compared to rural counties?
6. Create `income_dict`, a dictionary that uses the numerical codes in the income bucket column as keys and the matching descriptions ('Total', 'Under $1', 'Between 1 and $10,000', 'Between 10,000 and $25,000','Between 25,000 and $50,000', 'Between 50,000 and $75,000', 'Between 75,000 and $100,000', 'Between 100,000 and $200,000', '$200,000 or more') as values.
7. Pass the dictionary as an argument to the pandas `replace()` method to change the `income_bucket` column so that it uses descriptive text instead of the numeric code.
8. Use the pandas `groupby()` method to group the data by county and get the `sum()` of all numeric columns for that county. Be sure to also `reset_index()` so that our aggregated data is re-indexed to begin at 0. Save this as a DataFrame called `income_county_agg`, and look at the first few rows.