Week 1

What is Data Science?

- Coding Skills/Statistics/Domain Expertise
- The Data Science Process
- Examples of Data Science in the world

Understanding the Question + Getting Data

- Critical thinking and active questioning
- Acquiring domain expertise
- Articulating the question
- Review of code for today (pandas API)

Coding tasks: Get and begin cleaning data.

1. Create a DataFrame, `ha_costs_df`, from the `mmd_heart_attack_data.csv` file.
    - Examine the first 5 rows and the last five rows of the TN heart attack data.
    - Check the data types of each column by using `.info`.
    - Keep only the rows where facilities are in TN.
    - Print the dimensions of the resulting dataframe.
    - How many TN counties are represented in the heart attack costs data?
    - How many counties are classified as urban? How many as rural?
2. Create a DataFrame, `cancer_costs_df`, from the `mmd_cancer_data.csv` file.
    - Look at the head and tail of the DataFrame.
    - Keep only the rows where facilities are in TN.
    - Print the dimensions of the data. How many TN counties are represented in the cancer costs data?
    - You should have found that the datasets have a different number of counties. Bonus: Can you figure out which counties are missing from one of the datasets?
3. Create a DataFrame, `income_df`, from the `irs_county_2016.csv` file.
    - Keep only the data that pertains to Tennessee.
    - Look at the head and the tail.
    - Print the shape.
    - Keep only the following columns: `['STATE', 'COUNTYNAME', 'agi_stub', 'N1', 'mars1', 'MARS2', 'MARS4', 'N2', 'NUMDEP', 'ELDERLY', 'A00100', 'N02650', 'A02650', 'N02300', 'A02300']`
    - rename those columns: `['state', 'county', 'income_bucket', 'return_count', 'single_returns', 'joint_returns', 'head_of_house_returns', 'exemptions', 'dependents', 'elderly', 'agi', 'returns_with_total_inc','total_inc_amt', 'returns_with_unemployment', 'unemployment_comp']`