Exploratory Data Analysis – Beginners

Assignment:

Using RMarkdown (R) or Jupyter Notebook (Python), create a reproducible report. Explain your analysis throughout the notebook as if it would be presented to a stakeholder.

1. Using the Food Environment Atlas, create a histogram of diabetes rates (PCT_DIABETES_ADULTS13) and add a reference line for the national mean.
2. Separate the data into groups by metro status (METRO13) and persistent-poverty status (PERPOV10). Using histograms, how do the distributions differ depending on whether the diabetes rates are for metro vs. non-metro areas, or for areas with/without persistent poverty?
3. Take the correlation between diabetes rates and median household income (MEDHHINC15). Then, make a scatterplot to visualize the relationship.
    a. The relationship does not look linear from the scatterplot. Take the log of median household income and recalculate the correlation. What happens?
4. Is the relationship between diabetes rates and median household income affected by metro status?

- Make comparative scatterplots and calculate correlation coefficients across the two groups of metro vs. non-metro counties.

Deliverables:

- PDF or HTML file of your notebook and visualizations

Datasets:

Food Environment Atlas:

- USDA ERS - Data Access and Documentation Downloads
- Or here: https://www.ers.usda.gov/webdocs/DataFiles/80526/FoodEnvironmentAtlas.xls
- Or utilize the dataset created in a past assignment or find a copy of the dataset saved in **General > Files > Datasets**

Complete (up to and including) the following courses for this exercise:

R:

- Reporting with R Markdown
- Exploratory Data Analysis in R
- Introduction to Statistics in R

Python:

- Exploratory Data Analysis in Python
- Introduction to Statistics in Python