USDA Data Science Training Program
Intermediate Assignment 3: Exploratory Data Analysis

Data: USDA-ERS data on food environment factors (a subset of the Food Environment Atlas), along with county-level food security data from Feeding America. This assignment uses the variables listed below.

- `State` - Name of the state
- `County` - Name of the county
- `CHILDPOVRATE15` - Child poverty rate, 2015
- `PCT_LACCESS_HHNV15` - Percent of households with no car & low access to store, 2015
- `METRO13` - Metro/nonmetro classification, 2013

For reference, a full list of the variables in this dataset is included at the end of the assignment.

DataCamp: The following DataCamp courses correspond to this exercise:

- R: Exploratory Data Analysis in R, Introduction to Statistics in R, Reporting with R Markdown (bonus question)
- Python: Exploratory Data Analysis in Python, Introduction to Statistics in Python, Building Dashboards with Dash and Plotly (bonus question)

Assignment:

Part A: Load, inspect, and prep the data

1. Save `EDA_assignment_dataset.csv` to your computer and import the data into a dataframe called `food_env`.
2. Print the summary statistics for all columns in `food_env`. Do any variables have missing data?
3. Create a new dataframe named `food_env_full` that excludes rows with any missing data. How many rows were dropped?
4. Convert the `METRO13` column to data type "logical."

Part B: Explore the data

5. What are the minimum, maximum, mean, median, and standard deviation of `CHILDPOVRATE15`? Plot a histogram of child poverty rates to view the distribution.
6. Create a density plot to examine the distribution of `PCT_LACCESS_HHNV15`. Does it look like the data are skewed? To investigate further, create a boxplot of the `PCT_LACCESS_HHNV15` data.
7. Filter `food_env_full` to just counties in Virginia, Illinois, Michigan, Arkansas, and Ohio and save the smaller dataset as `food_env_5_states`.

8. What are the minimum, maximum, mean, median, and standard deviation values of CHILDPOVRATE15 and PCT_LACCESS_HHNV15 in each state?
9. Create faceted histograms or density plots of CHILDPOVRATE15 and PCT_LACCESS_HHNV15 by state. What do you notice about the distributions?

Part C: Preliminary analysis

10. Using food_env_full, compute the interquartile range of the PCT_LACCESS_HHNV15 column to identify outliers. Remove rows with outliers from the data then create a new boxplot with the filtered dataframe. How does this boxplot compare to the one you created in question 6?
11. Make a scatterplot with PCT_LACCESS_HHNV15 (outliers removed) on the x-axis and CHILDPOVRATE15 on the y-axis. Add a straight line that shows the linear relationship between the two variables to your plot.
12. Compute the correlation between PCT_LACCESS_HHNV15 and CHILDPOVRATE15. Do you think the relationship is strong? Are there other factors that could affect both the child poverty rate and a household's access to a grocery store in a county?

Bonus: Create a report in R Markdown or Dash using at least one of the plots you made, a table showing summary statistics for either CHILDPOVRATE15 or PCT_LACCESS_HHNV15, and text explaining your analysis.

Deliverables:

- Your code (the .R, .Rmd, .py, or .ipynb file).
- If you've chosen to write your responses or to complete the bonus question in an R Markdown file, the knitted document with your responses.
- These deliverables will be submitted through GitHub by the end of the program.

Data Dictionary:

- FIPS - Numeric code that identifies the geographic area
- State - Name of the state
- County - Name of the county
- Pop2020 - Population of the county, 2020
- POVRATE15 - Poverty rate, 2015
- SNAPS17 - Count of SNAP-authorized stores in the county, 2017
- CHILDPOVRATE15 - Child poverty rate, 2015
- PCT_LACCESS_HHNV15 - Percent of households with no car & low access to store, 2015
- GROC16 - Count of grocery stores in the county, 2016
- FMRKT_SNAP18 - Count of Farmers' markets that report accepting SNAP, 2018
- METRO13 - Metro/nonmetro classification, 2013
- food_insecurity_2016 - Food insecurity rate, 2016