Assignment 2: Data Analytics (Fall 2025) / written + figures 10%

Due: February 21th, 2025

Submission method: email (eleisa2@rpi.edu) or LMS

Please use the following file naming for electronic submission:
DataAnalytics_A2_YOURFIRSTNAME_YOURLASTNAME.xxx

Late submission policy: first time – no penalty, otherwise 20% of score deducted each late day

Take care to avoid plagiarism ("copying"), and include references to all web resources, texts, and class presentations. You may discuss the project with other students, but do not take written notes during these discussions, and do not share your presentations before class.

General assignment: Exploratory data analysis. Using the EPI results dataset, perform the following:

Variable Distributions

1) Derive 2 subsets, each for a different region.

    1.1. Plot histograms for a variable of your choice for both regions with density lines overlayed.

    1.2. Plot QQ plots for both variables compared to known probability distributions.

Linear Models

2) Fit linear models as follows:

    2.1. Choose 2 variables and fit linear models with these variables as response and choose either population or gdp (or both) as predictors. For each model print the model summary stats and plot the most significant predictor vs the response as well as the residuals. Apply transformations (e.g. log) to variables if needed.

    2.2. Repeat the previous models with a subset of 1 region and in 1-2 sentences explain which model is a better fit and why you think that is the case.

Classification (kNN)

3) Train 2 kNN models using ”region” as the class label as follows:

3.1. Choose 3 variables (not population or gdp) and create a subset by region keeping 2 regions out of 8 (representing 2 classes) and the 3 chosen variables. Train a kNN model to predict the region based on the 3 chosen variables. Evaluate the model using a confusion matrix and calculate the accuracy of correct classifications. Accuracy = correctly classified/total data points. You may try several values for k.

3.2. Repeat the previous model with 3 other variables. In 1-2 sentences explain which model is better.