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

Due: October 14th, 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 one for a different region and use them to do the following:

   1.1. Plot boxplots and histograms for a variable of your choice for each region separately with density lines overlayed. (2 boxplots & 2 histograms)

   1.2. Plot a QQ plot for the same variable between the 2 regions. (1 QQ plot)

Linear Models

2) Using the full dataset, fit linear models as follows:

2.1. Choose 1 variables and fit 2 linear models with that variable as response. Choose either population or GDP (or both) as predictors (inputs). Apply transformations (e.g. log) to variables if needed to obtain the best performing model.

For each model:

- Print the model summary stats.

- Plot the most significant predictor vs the response.

- Plot the residuals.

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) Using the full dataset, do the following:

3.1. Train a kNN model using “region” as the class label and choose 3 variables (not population or gdp) as inputs to the model. 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 and the same *k* value. Evaluate with a confusion matrix and in 1-2 sentences explain which model is better.