Assignment 2: Data Analytics (Spring 2026) / written + figures 10%

Due: February 27th, 2026

Submission method: 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. Using the **full dataset**, produce the following plots: (2%)

   1.1. A histogram of a variable of your choice with density lines overlayed.

   1.2. Boxplots of the same variable, one for each region, i.e. multiple boxplots in one figure.

2. Derive **2 subsets** from the dataset for any **2 regions** of your choice. Produce the following plots: (2%)

   2.1. Histograms of the same variable from (1) for each region.

   2.2. A QQ plot for the variable between the 2 subsets.

Linear Models

3. For each subset (region), do the following: (3%)

   3.1. Plot **population** and **GDP** separately against your chosen variable with the best fit line overlayed. Apply transformations (e.g. log) as necessary to obtain an informative plot.

   3.2. Fit **2 linear models** with that variable as response. Choose either population or GDP (or both) as predictors (inputs). Apply transformations to variables if needed to obtain the best performing model.

For each model:

- Print the model summary stats.
- Plot the residuals.

3.3. Compare the models for both regions and very briefly describe which one is a better fit and why you think that is the case.

Classification (kNN)

4. Derive **a new subset** from the original dataset containing **2-3 regions** and do the following: (3%)

4.1. Train a kNN model using “region” as the class with 3 input variables (*population* + *gdp* + *your chosen variable*). Evaluate the model using a confusion matrix and calculate the accuracy of correct classifications.

- *Accuracy = correctly classified/total data points*.
- Apply variable transformations as needed!
- You may try several values for *k* and compare the accuracy. Report only on the best performing model.

4.2. Using the best performing *k* value from above, train another model with a new 3^{rd} variable (*population* + *gdp* + *new chosen variable*).

4.3. In 1-2 sentences explain which model performs better and why you think that is the case.

EOF