

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

Due: October 16th, 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.

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, each from a different **region** 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 comparing values of the same variable from both regions.

Linear Models

3. For **each subset** (region), do the following: (3%)
 - 3.1. Plot population and GDP against the chosen variable with the best fit line overlayed. Apply transformations to variables as necessary to produce an informative plot.
 - 3.2. Fit **2 linear models** with your chosen variable as response and either population or GDP (or both) as predictors (inputs). Apply transformations to variables if needed to obtain the best performing model. Print summary statistics for both models and plot 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

4. Derive a **new subset** from the original dataset containing countries from **3 regions** and do the following: (3%)
 - 4.1. Consider “region” as target variable and the 3 features: population, gdp, and your chosen variable as predictors. Train multiple **kNN models** and compare them using classification accuracy to find the optimal number of neighbors (k). Plot k with accuracy.
 - 4.2. Train a **Random Forest model** using the same set of predictors and compare its classification accuracy with that of the optimal kNN model. Print confusion matrices.

(optional) Try adding other features to the models and examine their effect on predictions.
 - 4.3. In 1-2 sentences explain which model performs better and why you think that is the case.

The End