Lab Report #2: Clustering by Investment Risk
Course: RSM338H1S, Winter 2026

Instructions:

- This assignment may be completed in **groups of up to 3 students**. Group members may be from either section. If you work in a group, submit one report with all names listed.
- Submit your lab report as a **PDF** to Crowdmark via Quercus. Export your Jupyter Notebook to PDF before uploading. Only one group member should upload the submission, being sure to select their group mates at the time of submission.
- There is no page limit, but be concise. A good report is thorough but not padded.

Marking:

- **75%** — Coding and results (correct implementation, complete answers to all parts, appropriate choice of methods, accurate numerical output, properly formatted tables and figures)
- **25%** — Overall quality (clear and professional writing, thoughtful interpretation of results, demonstrated understanding of the underlying concepts, logical flow and narrative structure)

Writing Expectations: Your report should read as a **coherent narrative**, not just code with scattered comments. Use section headers to indicate which problem you're working on. Before each code block, briefly explain what you are about to do and why. After results appear, interpret what you see. A reader should be able to understand your analysis even if they skipped the code cells.

You may use AI coding assistants (ChatGPT, Copilot, Claude, etc.) to help write code, but you must be able to explain what every line does. The text you write around the code is what demonstrates your understanding. You are ultimately responsible for your own work. **If you use an AI tool, you must disclose this in a note at the end of your report. Mention which tool you used, which tasks you asked it to complete, and discuss your (dis)satisfaction with its assistance.**

Assignment: You are a risk analyst at a global investment firm. The emerging markets team is expanding their coverage and has asked you to develop a data-driven framework for categorizing countries by investment risk. Rather than relying on subjective assessments or pre-defined categories (e.g., "developed" vs. "emerging"), they want you to use clustering algorithms to discover natural groupings based on objective risk indicators. Your analysis will help the team identify peer groups, spot outliers, and make more informed allocation decisions.

Data: You are provided with `countryriskdata.csv`, which contains risk indicators for 122 countries. The data includes four measures used by international investors to assess country risk:

- **Corruption** — Corruption Perceptions Index from Transparency International. Ranges from 0 (highly corrupt) to 100 (very clean).
- **Peace** — Global Peace Index from the Institute for Economics and Peace. Ranges from 1 (very peaceful) to 5 (not at all peaceful).
- **Legal** — Legal Risk Index from the Property Rights Association. Ranges from 0 to 10, with higher values indicating stronger property rights protection.
- **GDP Growth** — Annual GDP growth rate from the IMF, expressed as a percentage.

The data for this assignment are adapted from: *Machine Learning in Business: An Introduction to the World of Data Science* (2025) by John C. Hull, Jacky Chen, Zissis Poulos, and Jun Yuan.

Your goal is to use clustering to group countries by their overall risk profile. This is an **unsupervised learning** problem—we have no predefined "risk categories." Instead, we want to discover natural groupings in the data.

Problem 1: Exploratory Data Analysis

Before clustering, you should understand your data.

Data Preparation:

(i) Load the data into a pandas DataFrame. Inspect the first few rows and check for missing values.

(ii) Note the scales of the four features. Which feature has the largest range? Which has the smallest?

Tasks:

(a) Create a **scatter plot matrix** (also called a pairplot) showing the relationships between all pairs of features. You can use `pandas.plotting.scatter_matrix()` or `seaborn.pairplot()`.

Briefly describe what you observe:

- Are any pairs of features strongly correlated?
- Do you see any obvious clusters in any of the scatter plots?
- Are there any outliers that stand out?

(b) Compute the **correlation matrix** for the four features and present it as a table. Which pair of features has the strongest correlation? Does this make intuitive sense from a country risk perspective?

(c) Explain why we need to **standardize** the features before clustering. In particular, what would happen if we ran K-Means on the raw (unstandardized) data? Which feature would dominate the distance calculations, and why?

Problem 2: K-Means Clustering

Now apply K-Means clustering to group the countries.

Data Preparation:

(i) Extract the four numeric features (Corruption, Peace, Legal, GDP Growth) into a feature matrix $\mathbf{X}$.

(ii) Standardize the features using `sklearn.preprocessing.StandardScaler`. All subsequent analysis should use the standardized features.

Tasks:

(a) **The Elbow Method.** Run K-Means for $K = 1, 2, \ldots, 10$ clusters. For each K, record the **within-cluster sum of squares** (WCSS), which sklearn calls `inertia_`.

Plot WCSS against K. Based on the "elbow" in this plot, what number of clusters would you recommend? Explain your reasoning.

(b) Using your recommended K from part (a), fit a K-Means model to the data. Report:

- The **cluster centers** (centroids) in standardized units. Present these as a table with one row per cluster and one column per feature.
- A brief interpretation of each cluster. For example, if one cluster has high standardized Corruption and high standardized Legal, what does that suggest about the countries in that cluster?

(c) List the countries in each cluster. Do the groupings make intuitive sense? Identify 2–3 countries in each cluster and briefly explain why their placement seems reasonable (or surprising) given what you know about those countries.

(d) **The Silhouette Score.** Compute the silhouette score for $K = 2, 3, \ldots, 10$ clusters and plot the results.

According to the silhouette analysis, what is the optimal number of clusters? Does this agree with your elbow method result? If the two methods disagree, which would you trust more in this context, and why?

Problem 3: Hierarchical Clustering

Now apply hierarchical clustering to the same data.

Tasks:

(a) Using **Ward's linkage**, perform agglomerative hierarchical clustering on the standardized data. Create a **dendrogram** showing the full merge history.

Looking at the dendrogram:

- At what height would you cut the tree to get the same number of clusters as your K-Means solution?
- Are there any large "gaps" in the dendrogram that suggest a natural number of clusters?

(b) Cut the dendrogram to produce the same number of clusters as your K-Means analysis. Compare the cluster assignments:

- How many countries are assigned to the same cluster by both methods?
- Identify any countries that are assigned to different clusters. Can you explain why these countries might be "borderline" cases?

(c) Briefly discuss the tradeoffs between K-Means and hierarchical clustering. When might you prefer one over the other?

Problem 4: Interpretation and Investment Implications

(a) Based on your clustering analysis, create a **risk ranking** of the clusters from lowest risk to highest risk. Justify your ranking by referring to the cluster centers.

(b) Suppose you are advising an institutional investor who wants to diversify their emerging markets portfolio. How might they use your clustering results? What are the limitations of relying solely on these four indicators?

