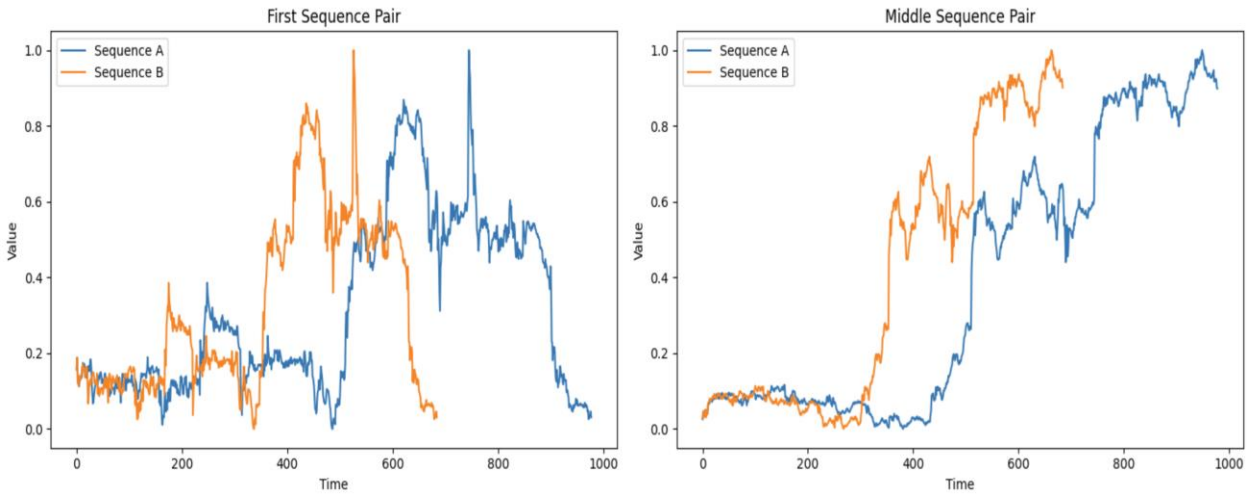
Part 3. Time Series Similarity.

In this section, we implement the Dynamic Time Warping (DTW) algorithm to measure the similarity between time series of varying resolutions. We test the approach on a sample dataset, where each row includes a pair of time series. The objective is to compute the distance between these time series using the DTW algorithm with Euclidean distance.

The dataset.

The dataset comprises 1001 rows, each representing a unique instance consisting of a pair of time series stored in the columns series_a and series_b. These columns contain time series encoded as comma-separated strings, with each time series potentially varying in length across instances. We can see how some instances look like in the below figure.

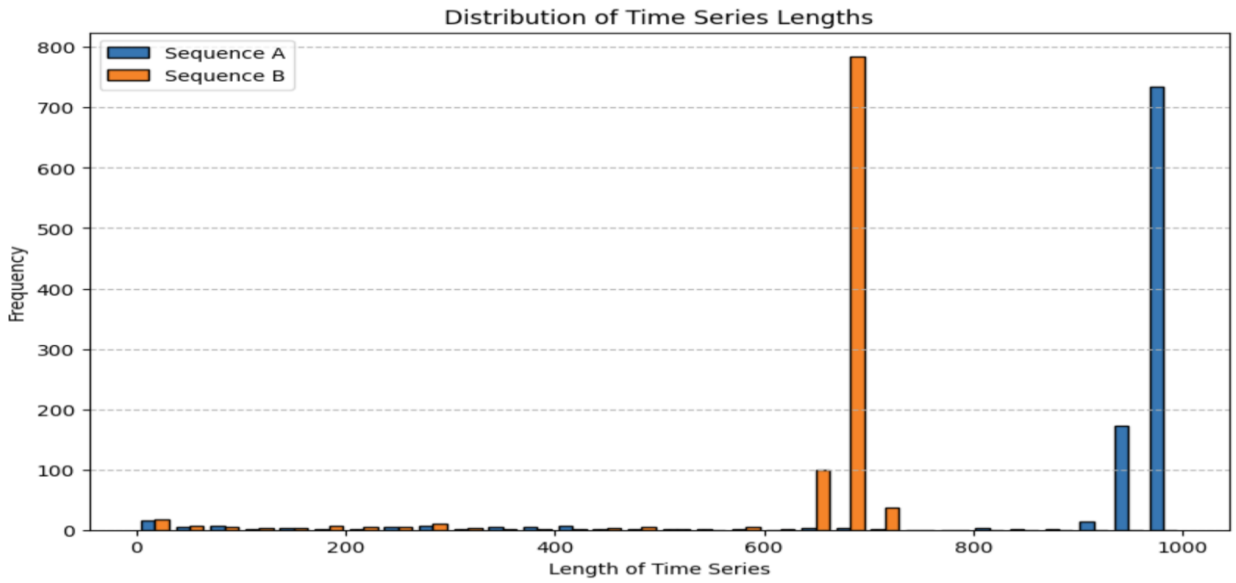


Some preprocessing is required for the dataset to ensure the time series data is suitable for numerical analysis and comparison.

To facilitate numerical computations, the time series in series_a and series_b must be converted into numerical arrays. This transformation is essential for applying mathematical operations and algorithms, such as Dynamic Time Warping (DTW), which aligns points between time series to compute their similarity. Without this conversion, the raw string format would render numerical comparisons and distance calculations impossible.

Additionally, scaling the time series is crucial due to potential differences in their ranges or magnitudes. For instance, one time series might span from 0.7 to 1, while another ranges from 70 to 77. Without proper scaling, such differences could skew the DTW results, prioritizing magnitude over the patterns or trends within the data. By applying techniques like min-max normalization, we ensure that all time series are represented on a consistent scale, enabling fair and meaningful similarity comparisons.

In the next figure, we observe the distribution of time series lengths for series_a and series_b in the dataset.



The figure highlights significant variability in the lengths of time series across both Sequence A and Sequence B. While most sequences are concentrated at specific lengths (e.g., around 1000 for Sequence A and 700 for Sequence B), there are many shorter sequences with varying distributions. This variability makes normalization crucial to ensure fair comparison during similarity calculations. After applying DTW, we normalize the results to eliminate biases caused by differences in sequence lengths, ensuring that the computed distances reflect the true patterns and trends in the data rather than being dominated by sequence magnitudes or durations.

Dynamic Time Warping (DTW) algorithm.

Dynamic Time Warping (DTW) is an algorithm designed to measure the similarity between two time series, even if they differ in length or are misaligned. It achieves this by aligning the sequences non-linearly, allowing for stretching, shrinking, insertions, or deletions to find the best alignment. DTW calculates a cumulative cost matrix that represents the optimal alignment path between the two series while minimizing the overall cost, which, in our case, is measured using the Euclidean distance.

DTW algorithm:

- Inputs:
    - Two sequences, x (length N) and y(length M).
- Define the DTW Cost matrix:
    - Construct a cost matrix D, where i and j correspond to the indices of sequences x and y, respectively. The matrix has dimensions (N+1, M+1).
- Initialization:
    - Set the boundary values:

        - D[i, 0] = infinity for i=1 to N.
        - D[0, j] = infinity for j=1 to M.
    - Set D[0, 0] = 0 as the starting point.

- Compute the Cost Matrix:

    - For i=1 to N:
      For j=1 to M:

      Compute D[I, j] = d(x[i], y[j]) + min{D[i-1, j-1], D[i-1, j], D[i, j-1]}.
      Here, d(x[i], y[j]) is the Euclidean distance between elements x[i] and y[j].

- Find the optimal alignment (Backtracking):
    - Once the cost matrix D[I, j] is fully computed, backtracking is performed from D[N, M] to D[0, 0]. This step identifies the optimal alignment path by following the direction with the minimum cost (match, insertion, or deletion) at each step.

    - The overall alignment cost is D[N, M], representing the minimal cumulative cost of aligning the two sequences.

Importance of DTW in Similarity Measurement.

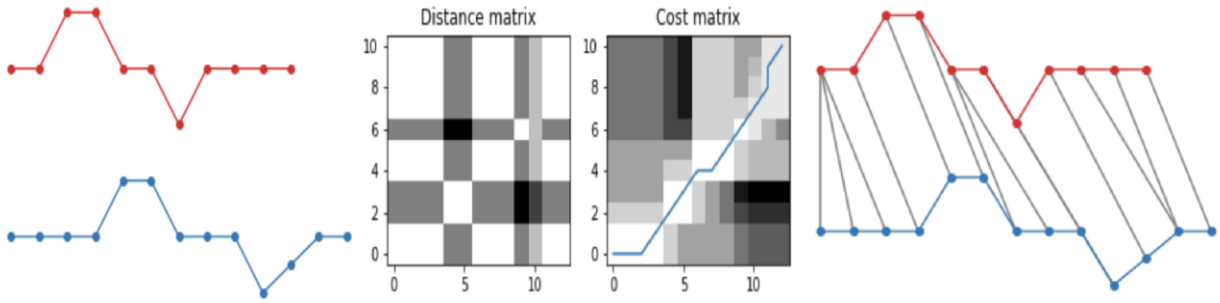
The value of D[n, m] provides a quantitative measure of similarity:

- A small alignment cost indicates that the sequences are highly similar after optimal warping.
- A large alignment cost suggests significant differences between the sequences, even after alignment.

DTW is particularly effective for handling time series of varying lengths or those that are misaligned in time, making it a powerful tool for comparing patterns in datasets like the one in this project.

DTW on a dummy example.

In the below graph, we see how the Dynamic Time Warping (DTW) algorithm works step by step on a dummy example. On the left side, we can see the two input time series of different, one in red and one in blue. In the middle, the distance matrix and the cumulative cost matrix D are shown. The cost matrix highlights the optimal alignment path in blue, tracing the minimum cumulative cost. And, on the right, the alignment of the two time series is illustrated, with gray lines connecting matched points to show how DTW aligns the sequences despite their different lengths. This effectively demonstrates the DTW process for sequence comparison and alignment which we applied in the provided dataset.



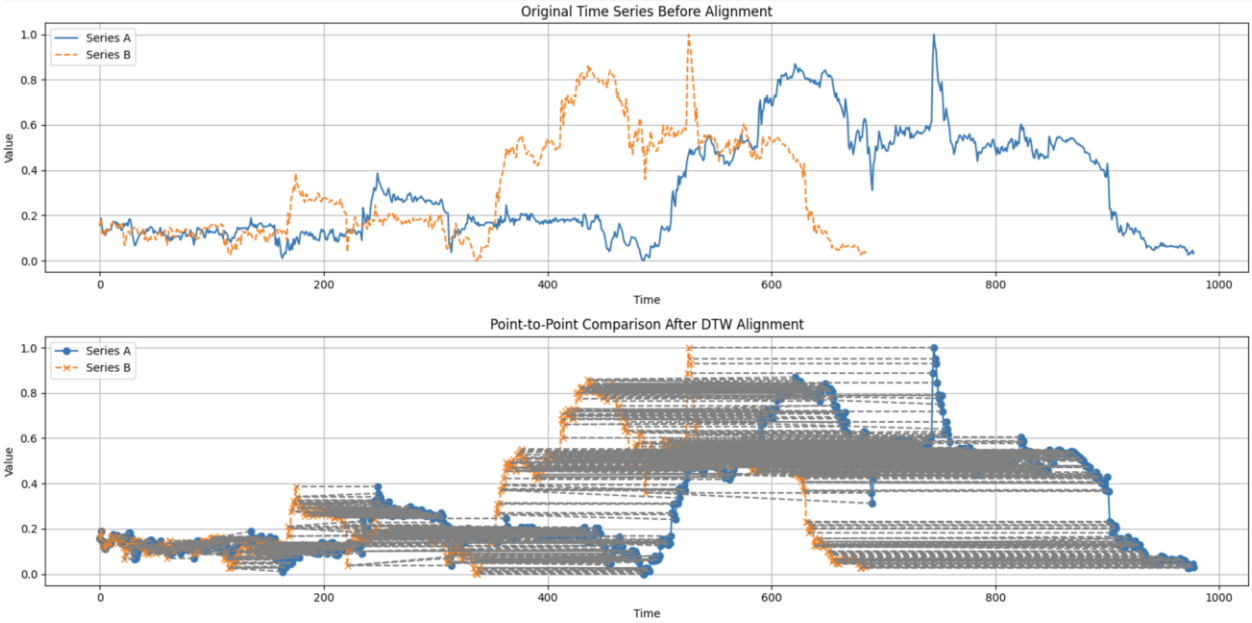
The method.

We utilize the Dynamic Time Warping (DTW) algorithm to measure the similarity between pairs of time series in the dataset. The process begins with preprocessing, where the raw time series data is converted into numerical arrays and scaled using min-max normalization to ensure fair comparisons of magnitude. Using DTW, we compute the optimal alignment and corresponding similarity scores for each pair of time series, even when they differ in length or are misaligned.

To account for variations in sequence lengths, the resulting DTW similarity scores are normalized by dividing the raw DTW cost by the combined lengths of the two sequences. This length normalization ensures that longer sequences, which naturally accumulate higher costs, do not bias the comparisons. The combination of magnitude and length normalization allows for meaningful and unbiased comparisons across all pairs in the dataset, focusing the analysis on the underlying patterns and trends within the time series rather than being influenced by their magnitudes or durations.

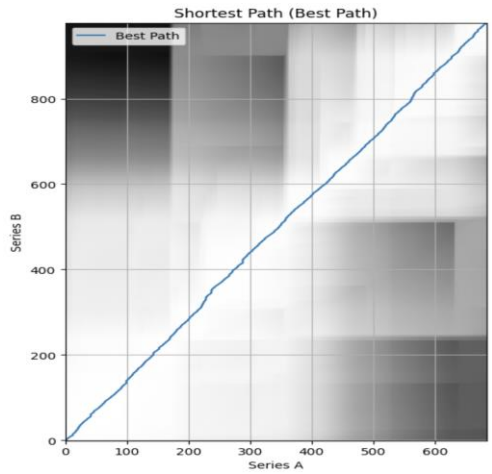
Results.

The Dynamic Time Warping (DTW) algorithm was applied to all pairs of time series in the dataset, yielding similarity scores that reflect the optimal alignment between sequences. To ensure fair comparisons, the results were normalized both in terms of magnitude and sequence length. The raw DTW distances and their normalized counterparts are presented, with the normalized values allowing for unbiased analysis of alignment quality across all pairs. The final normalized results, stored in the file **dtw.csv**, highlight meaningful similarities by eliminating biases due to differences in magnitude and length.



The figure above demonstrates the effect of Dynamic Time Warping (DTW) on aligning two time series, Series A (blue) and Series B (orange). In the top plot, we observe the original time series before alignment**,** where the two sequences exhibit significant temporal misalignment despite having similar trends in some regions. Differences in timing and scaling are apparent, making direct point-to-point comparison ineffective for similarity analysis. In the bottom plot, we see the point-to-point comparison after DTW alignment**.** The algorithm warps the sequences to align their key features by stretching or compressing segments as needed. The gray lines indicate the optimal alignment path between points in the two series, showing how DTW minimizes the cumulative distance by allowing non-linear alignments.

This alignment is crucial for calculating meaningful similarity scores between time series, particularly when sequences differ in length or exhibit temporal distortions. By aligning the series optimally, DTW ensures that similarity is evaluated based on patterns rather than misalignment or magnitude differences, which enhances the robustness of the analysis.



The additional figure visualizes the cost matrix and the shortest path (or "best path") derived using DTW. The grayscale heatmap represents the alignment cost at each point in the matrix, with lighter regions indicating lower costs. The blue line traces the optimal path through the cost matrix, aligning the two series with the minimal cumulative cost. This best path highlights the sequence of alignments that minimizes the total distance between Series A and Series B. It demonstrates DTW's ability to handle non-linear alignments by allowing deviations from the diagonal (perfect alignment), reflecting the temporal warping needed to align the two sequences effectively. The visualization underscores DTW's capability to provide meaningful similarity measures even for misaligned or length-varying time series.

Applications of DTW.

Having completed this part of the assignment, we explored the practical applications of Dynamic Time Warping (DTW) and discovered its extensive use across various domains. DTW is widely applied in speech recognition to align spoken words with reference templates, regardless of differences in speed or timing. In bioinformatics, it is used to compare gene expression profiles and align protein sequences. DTW also plays a crucial role in finance, where it helps identify similar patterns in stock price movements or detect anomalies in time series data (a field that we are interested in seeing more things!). Moreover, in healthcare, DTW is utilized for analyzing physiological signals such as ECG and EEG data, enabling more accurate diagnostics. These diverse applications highlight the versatility and importance of DTW in solving real-world problems where alignment of time series with varying lengths and distortions is essential.