Data Splitting Documentation

This section documents how the dataset was split into training and testing sets.

Splitting Strategy:

A **temporal split** was applied to the dataset. This strategy divides the data based on a specific point in time, ensuring that the training data precedes the test data. This is particularly important for time-series or time-sensitive data to simulate a realistic prediction scenario where the model is trained on historical data and evaluated on future data.

Split Point:

The dataset was split at the beginning of the year **2020**.

- **Training Data:** Includes all data from the earliest year in the dataset (**1995**) up to and including **2019**.
- **Test Data:** Includes all data from **2020** up to and including the latest year in the dataset (**2025**).

Rationale for Temporal Split:

Choosing a temporal split aligns with the project's goal of predicting economic freedom scores over time. By training on historical data and testing on later data, we can evaluate the model's ability to generalize to unseen future periods.

The split resulted in approximately 80% of the data for training and 20% for testing. Note that in a temporal split, the exact percentage in each set depends on the distribution of data points over time, not a predefined ratio.

Conclusion:

The temporal splitting strategy at the beginning of 2020 provides a suitable framework for evaluating the predictive performance of the regression models on future economic freedom data.