1. **Introduction**

**1.1 Problem Description**

In our rapidly evolving world and ever-changing global landscape the availability and quality of statistical data are pivotal for decision-making in societal and economic contexts. To address this the World Bank has developed the Statistical Performance Index (SPI) which acts as a comprehensive too to measure a country's statistical capacity. It covers key areas such as data infrastructure, statistical methodologies, and information dissemination. Significant challenges are posed by differences in statistical performance across different regions and income classifications. Countries with lower SPI scores often lack robust data systems which limit their ability to make informed decisions on economic policies, resource allocation, and sustainable development initiatives.

Understanding the key contributors to SPI and identifying regional and income-based disparities is crucial for policymakers and development agencies. Insights derived from this analysis aim to enable targeted interventions to strengthen statistical systems, reduce disparities, and promote evidence-based decision-making globally.

**1.2 Aims and Objectives**

This project aims to analyse the Statistical Performance Index (SPI) and its underlying pillars (Pillar 1 to Pillar 5) to address the following questions:

* Which SPI pillars (Pillar 1 to Pillar 5) contribute the most to the overall SPI scores, and how does this vary across regions?
* How does the Statistical Performance Index (SPI) vary across income classifications and regions, and what are the key disparities?

Additionally, a global Choropleth Map will be presented to visualize the distribution of SPI scores and its five pillars over the years 2016-2023. This analysis will provide valuable insights for addressing statistical gaps and promoting data-driven growth across countries.

**2.0 Data**

**2.1 Selection of Data**

The dataset has been obtained directly from the World Bank a highly reliable and reputable source. This ensures credibility and accuracy of the analysis. The SPI dataset includes data covering key variables such as the overall SPI scores and its five pillars: Data Use, Data Services, Data Products, Data Sources, and Data Infrastructure. These pillars form the basis for understanding and evaluating statistical performance across different regions and income classifications.

This dataset is publicly available and has been chosen due to its alignment with the project objectives. Community-based platforms such as Kaggle primarily host user-generated datasets. In sharp contrast the World Bank dataset is curated and verified by experts. The focus on publicly available official data ensures transparency and eliminates concerns about accuracy or credibility, which can sometimes arise with platforms like Kaggle.

**2.2 Ethical, Privacy, and Security Considerations**

This publicly available dataset does not contain sensitive or personally identifiable information which guarantees compliance with data ethics standards and data privacy regulations. Using public data ensures that the project avoids ethical concerns related to unauthorized data access or misuse. Additionally, the analysis has been conducted in a responsible manner thus ensuring that the data is used in its intended context distorting findings.

**2.3 Data Reliability and Integrity**

Several measures were undertaken to ensure the reliability of the data. The dataset was validated by cross-checking it to confirm its completeness and consistency. Since this is a real-world data there are columns where significant chunks of data missing. Any columns with 70% or more data missing were excluded, however the number of such columns were extremely small. Missing values were identified and handled appropriately to maintain the integrity of the analysis. For example, missing SPI pillar scores for specific countries in certain years were excluded from statistical calculations or replaced using the median wherever justified. Outliers, which could skew the results were carefully reviewed to assess their validity. Unjustifiable outliers were flagged and excluded to ensure accurate findings.

**2.4 Data Nature, Structure, and Preparation**

The SPI dataset is structured as a time series dataset containing data for multiple years across various countries. It includes cross-sectional components, allowing analysis across different regions and income classifications.

Key attributes of the dataset include:

* Country/Region - Representing geographical divisions
* Income Classification - Categorizes countries based on income levels
* SPI Overall Scores and Pillars - Metrics measuring statistical performance of countries
* Years - Annual data

The dataset underwent several cleaning and preparation steps:

* Handling Missing Data: Missing values were identified and treated through imputation methods or exclusion based on their impact on the analysis.
* Outlier Treatment: Extreme values were analysed and addressed to avoid skewing the results.
* Aggregation: Data was aggregated to calculate regional averages and trends across income classifications for specific years.

These steps were to prepare the dataset for a robust and accurate analysis ensuring a reliable and insightful analysis.