Answers 6.1

Data Sourcing:

The Statistical Performance Indicators dataset was developed internally by the World Bank Group. https://datacatalog.worldbank.org/search/dataset/0037996/Statistical-Performance-Indicators.

National statistical systems are facing significant challenges. These challenges arise from increasing demands for high quality and trustworthy data to guide decision making, coupled with the rapidly changing landscape of the data revolution. To help create a mechanism for learning amongst national statistical systems, the World Bank has developed improved Statistical Performance Indicators (SPI) to monitor the statistical performance of countries. The SPI focuses on five key dimensions of a country's statistical performance: (i) data use, (ii) data services, (iii) data products, (iv) data sources, and (v) data infrastructure.

Data Collection Method:

The data collection method is administrative. The SPI contains indicators along five dimensions: data use, data services, data products, data sources, and data infrastructure. Each dimension of the five pillars incorporates several indicators. These Statistical Performance Indicators embody the granular measures of performance. They are aggregated into levels of dimensions and pillars, and finally to an overall performance score to get a higher level or a more general perspective of a country's performance. The data is to provide an objective, justifiable assessment of country statistical capacity over time with comprehensive, up-to-date information; 2) Provide guidance to the World Bank teams in assessing the progress and sustainability of the Bank supported projects; 3) Inform Systematic Country Diagnostics; 4) Provide a monitoring tool for countries' SDGs data production capacity.

Data Limitations:

The limitation of the dataset is that it is published annually, thus potentially creating a time lag depending on when the data is being analyzed.

Data Contents:

The SPI dataset consists of 4141 rows and 79 columns. The SPI focuses on five key pillars of a country's statistical performance: (i) data use, (ii) data services, (iii) data products, (iv) data sources, and (v) data infrastructure. The SPI are composed of more than 50 indicators and contain data for 186 countries. This set of countries covers 99 percent of the world population. The data extends from 2016-2022, with some indicators going back to 2004.

Data Relevance:

The dataset is developed by World Bank to monitor the statistical performance of countries. The SPI focuses on five key dimensions of a country's statistical performance: (i) data use, (ii) data services, (iii) data products, (iv) data sources, and (v) data infrastructure. It represents data from 2004 to 2022 updated annually. Being collected and developed internally by a reputable government organization, the dataset can be considered reliable and trustworthy.

Data Profile:

Variables	Time-variant / -invariant	Structured / Unstructured	Qualitative / Quantitative	Qualitative: Nominal / Ordinal Quantitative: Discrete / Continuous
country	Time-invariant	Structured	Qualitative	Nominal
iso3c	Time-invariant	Structured	Qualitative	Nominal
date	Time-variant	Structured	Quantitative	Continuous
SPI.INDEX.PIL1	Time-variant	Structured	Qualitative	Ordinal
SPI.INDEX.PIL2	Time-variant	Structured	Qualitative	Ordinal
SPI.INDEX.PIL3	Time-variant	Structured	Qualitative	Ordinal
SPI.INDEX.PIL4	Time-variant	Structured	Qualitative	Ordinal
SPI.INDEX.PIL5	Time-variant	Structured	Qualitative	Ordinal
SPI.INDEX	Time-variant	Structured	Qualitative	Ordinal
SPI.DIM1.5.INDEX to SPI.D5.5.DIFI (66 total columns)	Time-variant	Structured	Qualitative	Ordinal
income	Time-variant	Structured	Qualitative	Ordinal
region	Time-invariant	Structured	Qualitative	Nominal
weights	Time-variant	Structured	Qualitative	Ordinal
population	Time-variant	Structured	Quantitative	Discrete

Data Consistency Checks:

Missing Values	**Duplicates**	**Mixed Data Types**
Pillar 1 - Data Use – Score 1 value replaced with avg 51.01	No duplicates	No mixed type variables
Pillar 2 - Data Services – Score		

2896 values		
Pillar 3 - Data Products - Score 272 values		
Pillar 4 - Data Sources – Score 2903 values		
Pillar 5 - Data Infrastructure - Score 2814 values		
SPI Overall Score 2905 values		
Population 19 values		

Data Wrangling:

Column Renamed	Column Dropped	Column Type Changed	Comments
SPI.INDEX.PIL1			Pillar 1 - Data Use - Score
SPI.INDEX.PIL2			Pillar 2 - Data Services - Score
SPI.INDEX.PIL3			Pillar 3 - Data Products - Score
SPI.INDEX.PIL4			Pillar 4 - Data Sources - Score
SPI.INDEX.PIL5			Pillar 5 - Data Infrastructure - Score
SPI.INDEX			SPI Overall Score
income			Income
region			Region
population			Population
country			Country
date			Date
	SPI.DIM1.5.INDEX to SPI.D5.5.DIFI (66 total columns)		Not needed to analysis
	iso3c		Not needed for analysis
	weights		Not needed to analysis. All scores given on the scale from 0 to 1.

Questions:

- Is there a correlation between county income and their overall SPI score?
- Is there a relationship between regions and scores that countries in that region received?
- Does population have an effect on any of the scores countries receive?
- Did scores change for regions over the years?
- Did scores change for any country over the years?
- Did any country's income level change pre and post COVID?
- How did overall scores change for regions before and after the pandemic.

Since the data is labeled as public by the source there should not be any privacy issues using the data for educational purposes.