GOTIndex Data Manipulation Protocol
March 2025

Initial data manipulation procedure:

1. Enter all the raw sub-pillar data **(Sheet 1)**.
2. Sub-pillar standardization process: Fit all the data points on a 0 - 100 scale **(Sheet 2)**.
    a. For each sub-pillar, for every year, find the following information:
        i. Highest value data point using the =MAX() function
        ii. Lowest value data point using the =MIN() function
        iii. Difference between the highest and lowest value data points (MAX - MIN)
    b. Create a two-part "if" calculation
        i. If the raw data cell on sheet 1 is empty (data was not available), the corresponding cell on sheet 2 remains empty
        ii. If the raw data cell on sheet 1 is filled, the following calculation occurs:
            - =(Raw data point *Sheet 1* - sub-pillar MIN) / (Difference between sub-pillar Max and Min)
    c. Apply this process to each sub-pillar, for every year
3. Inversion process: This ensures that each sub-pillar's data points are displaying risk - the higher the value, the greater the risk **(Sheet 3)**
    - Identify if the raw sub pillar data source is displaying risk
        i. If the raw sub-pillar data is displaying risk…
            - Create a two-part "if" function
                a. If the standardized data cell on sheet 2 is empty, the corresponding cell on sheet 3 remains empty
                b. If the standardized data cell on sheet 2 is filled, the following calculation occurs on sheet 3:
                    - =Standardized data point * 100
        ii. If the raw sub-pillar data is not displaying risk…
            - Create a two-part "if" function
                a. If the standardized data cell (sheet 2) is empty, the corresponding cell on sheet 3 remains empty
                b. If the standardized data cell on sheet 2 is filled, the following calculation occurs on sheet 3:
                    - (Standardized data point - 1) * (-100)
                        1. -1: inverts the value, ensuring the value displays risk
                        2. * (-100): makes the value positive
4. Calculate each country's yearly risk profile for the given pillar **(Sheet 3)**
    - A country's yearly average risk level for each pillar is calculated by taking an average of all the standardized and inverted (if necessary) sub-pillar data points.
        i. Formula: =AVERAGE()
5. Re-standardize each country's average risk level for each pillar using the highest and lowest risk values as global maximum and minimum benchmarkers for each year **(Sheet 3)**

1. Using the country risk averages, find the following information:
    a. Highest risk value for the given pillar =MAX()
    b. Lowest risk value for the given pillar =MIN()
    c. Difference in risk between the highest and lowest-risk values
2. Re-standardize the data, using the highest and lowest risk values as benchmarkers
    a. Formula = (Country's Average Risk Score - Lowest Risk Score) / (Highest Risk Score - Lowest Risk Score)
3. Apply this formula to every country, for every year, and to every pillar.

Addressing the missing data points:

1. Since many online datasets did not include data for every country, regional averages were used as proxy data points to estimate a nation's risk level when exact figures were unavailable. The process for calculating these regional averages is as follows…
    a. Highlight all empty cells on the raw data sheet **(Sheet 1)** in yellow, creating a differentiation between regional average data and the real raw data.
    b. Apply the =AVERAGE() formula to every missing data cell for each country, ensuring that the formula includes the raw data points from all countries within the same region that have available data for the specific sub-pillar.
    c. Once the blank cells are filled with the regional averages, the calculations on sheets 2 and 3 will update automatically as the formulas have already been programmed.
2. Exceptional circumstances concerning the sources of raw data and the regional average calculation process include:
    a. Social Data
        i. Transport Data: Data was only available up to 2019, 2020-2022 uses the 2019 data
        ii. Gender Gap: 2019 uses 2020 report data
        iii. Food affordability: Data point used - Food Prices for Nutrition 3.0 - Affordability of a healthy diet: ratio of cost to the food poverty line
        iv. Childhood: No data for 2022, 2021 data is used
    b. Economic Data
        i. No data points for Tourism Balance in any North African country in 2021 - countries are using Western Asia (Middle East) averages.
        ii. No data points for IT as a % of GDP in any Western African country in 2021 - sub saharan African average applied (Eastern + Southern Africa)
        iii. No data points for Informal Labour in any North African country in 2019 - the region uses the Western Asian (Middle East) average
        iv. IT as a % of GDP, Tourism Balance & Informal Labor all use 2021 data for 2022
    c. Political Data
        i. Rule of Law: Data for 2017 and 2018 are combined, same data used for both years
    d. Environmental Data

i. Air pollution: No data is available for 2022, 2021 data is used as a placeholder.

Developing a composite risk summary sheet:

1. A new file called “GOTIndex - Total Composite Risk Data” was created, which includes 5 sheets: Total Composite Risk, Social Composite Risk, Economic Composite Risk, Political Composite Risk, and Environmental Composite Risk.
    a. The composite risk sheets for each pillar include a column for each year (2017-2022)
    b. Each country’s average risk score for each pillar, for each year is directly imported from the corresponding primary data sheet where the initial data manipulation procedure occurred.
        i. Using the formula =IMPORTRANGE
    c. On the summary sheet for each pillar, a 6-year average is calculated for each country.
    d. Once the summary sheets for each individual pillar were created, the total composite risk for each country and each year can be calculated.
        i. Formula = (Social Composite Risk Score + Economic Composite Risk Score + Political Composite Risk Score + Environmental Composite Risk Score)/4

Assessing data accuracy & the rationale for removing countries from the final analysis:

- Although the dataset is complete, the accuracy of risk representation varies by country. Some countries have precise data, while others rely heavily on regional estimates, which may affect the consistency of the final analysis.
- An analysis was needed to see which countries were relying too heavily on regional averages
    a. For each pillar and each year, the number of regional average data points each country used was tabulated
    b. A summary table called TRA (total regional averages) was created to calculate each country’s reliance on regional averages across all pillars and all years
        - A percentage was calculated for each country using the following formula:
            - # of regional average data points used / 126
                - 126 - Total data points used per country across all pillars and years

2. Conclusion: The top 15 countries using the highest amount of regional averages were removed from consideration
    - All countries from Melanesia, Polynesia, and Micronesia were subsequently excluded from the analysis because a significant number of countries initially removed were from these regions, making meaningful regional analysis difficult.
    - This altered the total country count from 174 to 155.
3. Eleven more countries were thereafter also removed from the analysis due to missing data related to the independent variable of the research, IT arrivals/Population.

- This brought the final country count from 155 to 144, which was the final country list which was used for the final analysis.

Data Analysis

- Independent variable = IT Arrivals / Population
- Dependent variables = Total Composite Risk, Social Composite Risk, Economic Composite Risk, Political Composite Risk, Environmental Composite Risk
- For each year, the IT Arrivals / Population value was analyzed in relation to each risk pillar from the same given year
- Analysis methods:
    - Correlation Analysis
    - Regression Analysis - R Squared value
- A regional analysis was also conducted to see if the correlation between risk and ARR/POP was being driven by a certain region and was less prominent in other regions
    - The macro regional analysis is as follows:
        - Northern Europe, Southern Europe, Western Europe, Eastern Europe
        - Australia/New Zealand, South Eastern Asia, Eastern Asia
        - Caribbean
        - South America
        - North America, Central America
        - Sub Saharan Africa: Southern Africa, Eastern Africa, Western Africa
        - Western Asia, Northern Africa
        - Southern Asia, Central Asia