WORLD ECONOMIC INDICATORS

# Process Solution :

* Gain a quick understanding of the content within the Google Sheet by skimming through the Data Dictionary.
* Reordered the columns in all the sheets, placing the shared columns ('Region/Country', 'Region', 'Year') as the initial three columns in each individual sheet. This adjustment enhances clarity in visualization.
* The objective involves crafting a distinct identification for every row, serving the purpose of VLookup Merging. The =CONCATENATE() function was employed to generate a column that combines 'Region/Country' and 'Year'.
* The process of introducing the 'Unique ID' column was replicated for each of the Sheets: 'GDP', 'Energy', 'Human Resource', 'Tourism', and 'Business'.
* To further enhance visualization, distinct color schemes were applied to each sheet, establishing a consistent approach for the entire project.

## Handling Missing Values of GDP :

* A new column labeled 'GDP (if missing)' was introduced with the primary aim of addressing any absent GDP values. Due to the critical nature of GDP and its susceptibility to fluctuations over time due to various factors, substituting missing values with metrics like mean, median, zero, etc., is not appropriate. To fill in the current year's GDP, a strategy involves utilizing the preceding year's value for the same region and feature. This approach is grounded in the notion that economic patterns tend to persist for consecutive years, facilitating analysis continuity. In situations where the prior year's data is also absent, the most suitable practice is to leave the respective cell empty.
* The formula applied to this column is as follows: =IF(E4 = "", IF(D4 = 2000, E5, E3), E4). The latter part of the formula ensures that the possibility of overlapping consecutive Country/Region entries is mitigated.

## Misprint in ‘Energy’ Sheet :

* Mistakes occurred in the Project: World Economic Indicators report. Within the 'Energy' sheet, the figures for 2006, 2007, and 2010 were absent, while 2005, 2008, and 2011 values were repeated for each Country/Region.
* A quick analysis of Energy Usage and CO2 Emission trends was conducted. Although a correlation was observed, caution was advised against presuming or interchanging Year values.
* Due to a lack of stakeholders or data engineers, a Copy-Paste method was employed instead of using Vlookup for that specific sheet.

## Merging Sheets using Vlookup :

The following formulas were used for merging data into one Combination Sheet.

Unique ID column was filtered into A to Z and copy pasted from ‘GDP’ into a New Sheet named ‘Combination Sheet’.

1. **GDP (Cell B3) :**

=ARRAYFORMULA(VLOOKUP(A3,GDP!$A$2:$I$2692,{2,3,4,5,6,7,8,9},0))

1. **ENERGY :**

Copy pasted the content after arranging Unique ID in ascending order as there was a content mistake.

1. **HUMAN RESOURCE (Cell M3) :**

=ARRAYFORMULA(VLOOKUP(A3,’HUMAN RESOURCES’!$A$2:$M$2692,{5,6,7,8,9,10,11,12},0))

1. **TOURISM : (Cell U3) :**

=ARRAYFORMULA(VLOOKUP(A3,TOURISM!$A$2:$F$2692,{5,6},0))

1. **BUSINESS (Cell W3) :**

=ARRAYFORMULA(VLOOKUP(A3,BUSINESS!$A$2:$J$2692,{5,6,7,8,9,10},0))

## GDP/ Capita :

* To gain a more profound insight into the connection between GDP (when available) and a country's population, an additional column (Column AC) named 'GDP/Capita' has been introduced in the Combination sheet.
* This column is populated with a numerical value derived from dividing the GDP by the Total Population associated with each Unique ID. (Formula: =F3/S3)

NOTE : (Blanks) are replaced with zero for each rows and columns for better visualisation.

# Final Analysis :

A Copy of ‘Process Solution\_DebdishaBanerjee’ is made to perform the various univariate and bivariate analysis between columns of interest.

## Correlation

To analyze the correlation trends between various columns and GDP as well as Population, a duplicate of the Combination Sheet was generated. The formula =CORREL($F:$F, column 2) was utilized to compute correlations with GDP (when available). This formula was applied in a new row placed beneath the column headings. Subsequently, the entire row was transposed into a column format and color-coded on a new sheet called 'Correlation' for improved visual representation using the 'Paste Special' feature. This process was repeated to determine correlations with Population.

The magnitude of correlation provides insights into our areas of interest among the columns. A higher positive correlation value indicates a stronger relationship between the variables.

## Univariate

Sheet ‘Univariate’ consists of univariate analysis of various columns of interest. In few cases a minimum value (approx. 0.00000001) is added to exclude missing values which are replaced with zeros.

For various charts, bucket size of the histogram is altered for better visualization.

## Pivot Table

Row 1 of 'Combined sheet' is hidden to treat cells in row 2 as column headers.

Table 1: A Pivot Table is utilized for analysis involving Country and GDP (if missing), Population, and GDP/Capita. For the sake of clarity, values are aggregated using the AVERAGE function. This provides us with an understanding of trends among various countries.

Table 2: A similar analysis is conducted using a Pivot Table involving Country and GDP (if missing), Population, and GDP/Capita. Values are summarized using the AVERAGE function to facilitate comprehension. This offers insights into trends across diverse countries.

## Bivariate Analysis

Bivariate analysis is performed using scattered plot graph of various columns containing continuous data, with GDP.

# Insights:

* The relationship between Energy Usage and GDP exhibits the most pronounced correlation coefficient among all the listed factors. This signifies a robust and positive connection between these two variables.
* CO2 Emissions display the second-highest correlation coefficient with GDP, implying that nations with greater economic productivity tend to generate increased CO2 emissions.
* There exists a substantial positive correlation coefficient between Inbound Tourism and GDP. This implies that countries witnessing a higher influx of international tourists tend to enjoy elevated economic outputs.
* Health exp/capita exhibit a stronger correlation coefficient with GDP compared to health expenditures as a percentage of GDP. This indicates that the monetary resources allocated per individual for healthcare serve as a more effective predictor of a nation's economic output than the proportion of GDP allocated towards healthcare.
* Population Total and Population 65+ demonstrate a moderate positive correlation with GDP, implying that nations with larger overall populations and a higher proportion of elderly residents often experience greater economic outputs.
* On the other hand, Birth Rate, Infant Mortality Rate, Population 0-14, and Population 15-64 exhibit either negative or weak correlation coefficients with GDP. This suggests a limited or negative association with economic output.
* Lending Interest, Business Tax Rate, Days to Start Business, Ease of Business, and Mobile Phone Usage showcase low correlation coefficients with GDP, indicating a weak or marginal relationship with economic output.
* Internet Usage displays a moderate positive correlation coefficient with GDP, indicating that countries with elevated levels of internet usage typically boast higher economic outputs
* The region of The Americas displays the most elevated mean GDP per capita compared to all the listed regions, signifying a superior quality of life within this area.
* Europe has the second highest GDP across all the regions listed, signaling a superior overall economic productivity in this particular region.
* In contrast, Africa exhibits the least average GDP per capita among the regions listed, pointing to a comparatively lower standard of living within this geographic area.
* Even though Africa maintains a relatively lower GDP per capita, it boasts the highest population when compared to all the regions mentioned. This could potentially hint at future growth possibilities as the economy undergoes development.