**Data Selection**

In the data selection process, we analyzed all the datasets, starting with Food\_Insecurity\_Nourishment\_Indicators, Food\_Prices\_for\_Nutrition, Health-Related Indicators and Overall health acheivement. Our analysis identified that the data in Food\_Prices\_for\_Nutrition spans from 2017 to 2022. However, upon analyzing the Food\_Insecurity\_Nourishment\_Indicators dataset, we found that only 37 values exist for 2022 (which is 97.7%), with the rest of the data missing for that year. Therefore, we decided to select data from 2017 to 2022 from all datasets.

Our goal is to obtain insights from different continents and countries with varying income levels. To achieve this, we researched the United Nations Sustainable Development Report, which provided us with information on the different income levels and categorizations of countries. Based on this report and the overall health achievement levels file, we identified four income categories: Low-Income Countries, Lower-Middle-Income Countries, Upper-Middle-Income Countries, and High-Income Countries.

Through our research, we found the World Bank’s income group categorization for each country for each year, and we also discovered continent-wise data mappings for the countries. We choose Asia, Europe, Africa, Oceania, North America, Antarctica, and South America as the seven continents. noting that there are no countries in Antarctica., we selected the other six continents.

After that, we combined these datasets for our analysis. During the data selection process, we took a sample of 15 countries based on data availability, ultimately selecting one country from each income level across the six continents.

**A screenshot of a computer screen

Description automatically generated**

**Data Preparation**

In the dataset preparation process, the main issue we encountered was missing values across the datasets. To address this, we first replaced the missing values to accurately identify and manage them. We then analyzed the percentage of missing values in each column by calculating the proportion of missing data.. The percentage of missing values in each column was computed using the following command:

missing\_pct = round(dataframe\_name.isnull().sum() / len(dataframe\_name) \* 100, 1)

In the transformation phase,Redundant columns, such as 'Time Code', were dropped to remove unnecessary information Afterward, we renamed the columns to more intuitive labels, such as changing 'Affordability of a healthy diet: ratio of cost to the food poverty line [CoHD\_pov]' to 'CoHD\_pov' and 'Cost of a healthy diet [CoHD]' to 'CoHD', among others, to simplify our analysis.

we ensured that the data types were correctly formatted, such as converting key columns like 'CoHD\_pov', 'CoHD', and 'CoHD\_headcount' to float. We also reshaped the Food\_Insecurity\_Nourishment\_Indicators dataset by melting and pivoting it to create a 'Year' column and reorganized other columns accordingly, which helped structure the data set for combining. The datasets were then merged using left and inner joins, where necessary, to combine relevant data from different sources. For data integration, we used left and inner joins to merge relevant data from different sources. Left joins ensured that no data was lost during merging, while inner joins were used when both datasets had complete data for the relevant columns, ensuring meaningful data integration

In the Food\_Prices\_for\_Nutrition and health-related indicators datasets, rows with missing data percentages lower than 10% were dropped. For columns with missing values between 10% and 70%, we used group-wise forward fill with mean imputation, particularly grouped by country. Finally, we ensured that all columns were renamed appropriately, making the datasets ready for further analysis.

In our analysis we did not focus on normalizing the data since all the data are is in the range for our analysis to get correct idea regarding dataset and different indicators related to health and food we did not do the normalization for this analysis at this stage.

Finally, for our analysis, we selected 15 countries representing different income levels that had complete and clean datasets. These countries were chosen based on the availability of all relevant values, ensuring that our analysis would be accurate and robust. By focusing on these countries, we were able to streamline the dataset, eliminate inconsistencies, and ensure that all data was properly cleaned and ready for further in-depth analysis.