Global Burden of Disease (GBD) - SQL/Tableau

Perusing the site ourworldindata.org, a section called “Burden of Disease” stood out to me and I was instantly interested in exploring the datasets.

Burden of disease is a figure that accounts for disease impact on a population. It accounts for both mortality and morbidity, meaning the death count as well as the years of suffering caused by a disease. The burden of disease is measured in DALYs (Disability Adjusted Life Years), which you’ll see as a measure in my visualisations for this project.

The Data

Initially, I used two datasets from ourworldindata.org -
One dataset lists DALYs per country, broken down by the cause of the burden (communicable, non-communicable, and injury).
Another dataset lists the DALYs for each country per year, over the span of 20 years.

Over the span of the project, I became interested in comparing the burden of disease with the diets of each country, so I enlisted another dataset from outworldindata.org. This dataset breaks down caloric intake of fats, carbohydrates, animal protein, and plant protein by country.

Cleaning

First, I noticed that there were entries in the “country” columns of both GBD datasets that were actually designated groups of countries, such as “African Region (WHO)” and “Eastern Mediterranean Region (WHO)”, and not actual countries. The countries that made up these groups were already listed individually, so I wanted to delete these groups from the dataset. I noticed that each of these country groups have blank cells in the “code” column (this column lists the country code for each country), so I used the DELETE FROM statement to delete every row that had a null in the “code” column.

```
--Delete rows where the "country" is actually a region or group of countries

DELETE FROM GBD_Project.dbo.dalys_rate_from_all_causes_with_averages
WHERE Code IS NULL;

SELECT *
FROM GBD_Project.dbo.dalys_rate_from_all_causes_with_averages;
```

Thinking about what questions I wanted to answer with the data, I decided to add a column to the dataset with only the total DALYs listed, which classified each country by their income level. This way I could later compare DALY figures of countries grouped by their income level. To do this, I found world income groups on worldbank.org, used the ALTER TABLE statement to add a “wealth classification” column, and used the SET and WHERE IN statements to classify the countries in my dataset into one of four classifications: high income, upper-middle income, lower-middle income, and low income.

```sql
--Add column in avg dalys table to input one of four wealth classifications

ALTER TABLE GBD_Project.dbo.Avg_DALYs_Per_Country
ADD Wealth_Classification nvarchar(20);

--designate countries as 'High Income' according to World Bank 2023

UPDATE GBD_Project.dbo.Avg_DALYs_Per_Country
SET Wealth_Classification = 'High Income'
WHERE Country IN ('Australia', 'American Samoa', 'Andorra', 'Antigua and Barbuda', 'Aruba', 'Austria', 'Bahamas', 'Bahra
'Cayman Islands', 'Channel Islands', 'Chile', 'Croatia', 'Curacao', 'Cypress', 'Czech Republic', 'Denmark', 'Estonia', '
'Finland', 'France', 'French Polynesia', 'Guyana', 'Germany', 'Gibralter', 'Greece', 'Greenland', 'Guam', 'Hong Kong', '
'Ireland', 'Isle of Man', 'Israel', 'Italy', 'Japan', 'Korea', 'Kuwait', 'Latvia', 'Liechtenstein', 'Lithuania', 'Luxemb
'Netherlands', 'New Caledonia', 'New Zealand', 'Northern Mariana Islands', 'Norway', 'Oman', 'Panama', 'Poland', 'Portug
'Romania', 'San Marino', 'Saudi Arabia', 'Seychelles', 'Singapore', 'Sint Maarten', 'Slovak Republic', 'Slovenia', 'Spai
'Turks and Caicos Islands', 'United Arab Emirates', 'United Kingdom', 'United States', 'Virgin Islands');

SELECT *
FROM GBD_Project.dbo.Avg_DALYs_Per_Country;

--Designate countries a 'Upper-Middle Income' according to World Bank 2023

UPDATE GBD_Project.dbo.Avg_DALYs_Per_Country
SET Wealth_Classification = 'Upper-Middle Income'
```

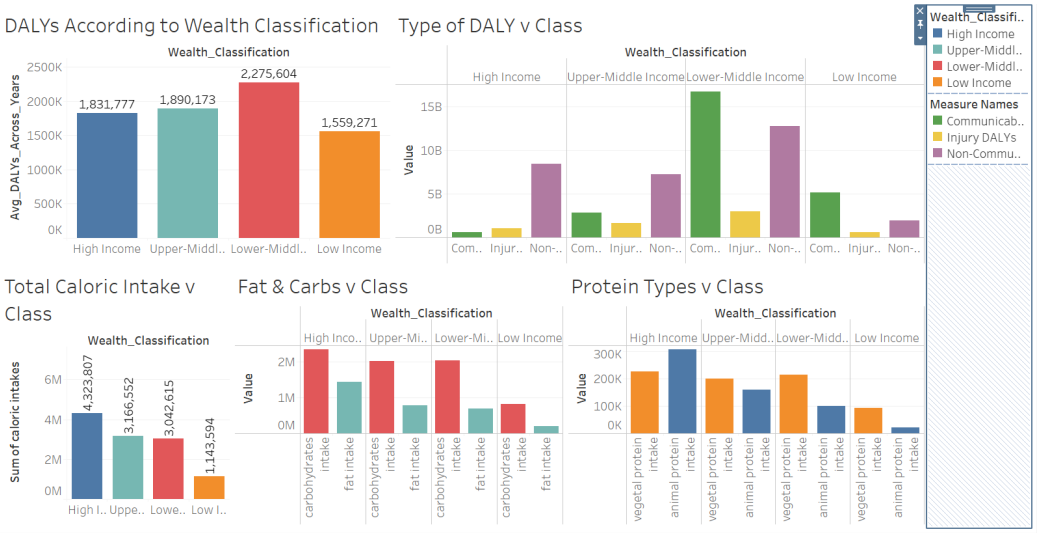
To check my work, I selected to view all “wealth classification” null values and filled in those values accordingly. Mostly, the null values were due to a country being spelled or referenced differently across the Our World in Data dataset and the World Bank classification lists.

```sql
SELECT *
FROM GBD_Project.dbo.Avg_DALYs_Per_Country
WHERE Wealth_Classification IS NULL;

--Fill in leftover High Income classifications

UPDATE GBD_Project.dbo.Avg_DALYs_Per_Country
SET Wealth_Classification = 'High Income'
WHERE Country IN ('Brunei', 'England', 'Niue', 'Northern Ireland', 'Scotland', 'Slovakia', 'South Korea', 'Tokelau', 'U

--Fill in Upper-Middle Income classifications

UPDATE GBD_Project.dbo.Avg_DALYs_Per_Country
SET Wealth_Classification = 'Upper-Middle Income'
WHERE Country IN ('Kyrgyzstan', 'Laos', 'Micronesia (country)', 'Palestine', 'Russia', 'Turkey');

--Fill in Lower-Middle Income classifications

UPDATE GBD_Project.dbo.Avg_DALYs_Per_Country
SET Wealth_Classification = 'Lower-Middle Income'
WHERE Country IN ('Sao Tome and Principe', 'Venezuela');

--Fill in Low Income classifications
```

Visualisations

I joined four different datasets in Tableau to create visualisations that compare the total DALYs across the wealth classifications, the breakdown of DALY causes across wealth classifications (communicable, non-communicable, and injury causes), and I brought in datasets from ourworldindata.org to show the distribution of dietary caloric intake across the wealth classifications to compare the classification's DALYs as well.

The graphs show that the lower-middle income countries had higher DALYs than the other wealth classifications, which was accounted for by a higher occurrence of communicable disease. The diet data was much more inconclusive, suggesting that the difference in DALYs across the classifications is probably due to more than just a difference in dietary intake.



There are so many more interactions that could be explored with these datasets. If I were to continue this project, I would look into other factors that could possibly contribute to the difference in DALYs, such as the number of hospitals or medical workers in each country. I would also look into what types of communicable diseases are accounting for the increase in communicable DALYs within the lower-middle income countries.