This is a submission for Group 1 including the students Göktuğ Aygün 30608, Kemal Yılmaz 31097, Steven El Khaldi 30048, Kerim Demir 28853 and Berk Bilgiç 29157.

In this step of the project, we started by filtering the excel sheets according to the data we needed for each table. In other words, if an excel file contained information about diet and substance deaths together, what we did was create 2 new excel files for substance and diet deaths and we copied to each the relevant information along with the country code column to both. Because, if we created a new table for each data that contains approximately 6000 rows it would be a waste of space. As for the files, each excel file is self-explanatory that each excel file contains the data about their related field.

Filling in missing data

Some of the countries in our csv file for countries were actually regions and didn't have an iso code value, and since iso code is a primary key in our countries table and can't be null, we decided to come up with our own unique codes for them and manually filled in the missing iso code columns for the regions in both the countries csv file and the continent-country mapping csv file.

Use of Pandas library

Some of the excel (and csv) files we had for the non-location data (e.g., emissions) didn't have a country code column and only a country name column, and the ones that had a country code column has missing cells for the regions as aforementioned. Hence, we decided to make use of Pandas library in Python on Google Colab where we uploaded the countries csv file and each of the other files that had the missing country code column (or missing cells in the country code column) in order to add a new country code column to each of them by merging them with the countries csv file on the country name column. This was completely successful.

Creating the tables and Importing the data

After our csv files for all 8 tables (3 location and 5 non-location) became ready and finalized, we came up with the SQL create table commands mentioned at the end of this report and executed them in SQL Workbench in order to create all 8 tables. After that, we imported each csv file into each table one by one using SQL Workbench's data import wizard, where we had to manually map each source column from the csv file to the correct destination column in the SQL table. All was successful except for an illegal character issue we encountered while importing countries csv file into our countries table, and the reason was that Madagascar's and Slovenia's iso codes seemed to accidentally have a special "é" character at the end of them that wasn't supported. After removing them, importing countries was successful too.

As mentioned before there are 8 files in total. They are

- continents
  Showing the entity set of continents with their codes as the primary key and their names
- countries
  Showing the entity set of countries with their iso_codes as the primary key and their names
- locatedin
  Showing the relation between countries and continents, taking the iso_code of the country as a primary key
- health_deaths
- diet_deaths
- nature_deaths
- substance_deaths
- emissions

Last 5 tables which cover the weak entity sets have the same format except "emissions" table. Other 4 have attributes in the format "deaths_by_X", X being the specific cause belonging to their category. Their primary keys are the country of interest and the year.

Emissions table focuses on the emission rates of countries of 4 gases being

- Nitrogen Oxide (NO_x)
- Sulphur Dioxide (SO_2)
- Carbon Monoxide (CO)
- Ammonia (NH_3)

The primary key for this table is again the country of interest and the year.

ER Diagram

Some changes need to done on the ER diagram in order to match the format of the tables created, explained in more detail at the beginning of this documentation. The updated version of the ER diagram can be found in the Github repository previously shared as well as all the other files. The link for the same repository can also be found at the end of this file.

The GitHub repository link is https://github.com/kerimdemir9/Cs306-Project