Phase 1 CSC 343

Connor Burns and Ekagra Luthra

September 30th 2021

1 Domain

The domain that has been chosen for the project is CO2 emissions and pollution.

2 Datasets

Two datasets will be used for the project.

The first dataset is **CO2 and Greenhouse Gas emissions data of the World from years 1750 - 2017.**

Link: `https://www.kaggle.com/yoannboyere/co2-ghg-emissionsdata`

Most of the information in this dataset is relevant to use for the project. The only data that is irrelevant is the data concerning high income deaths, as there is no data regarding low income deaths to compare it to. Some year ranges may be excluded if it is decided later on that only more recent data is relevant to the investigative questions posed. It is likely that the global total, and specific countries' data will be explored.

The second dataset is **Deaths due to Air Pollution from years 1990 - 2017.**

Link: `https://www.kaggle.com/akshat0giri/death-due-to-air-pollution-19902017`

All the information in this dataset is relevant to use for the project. Total pollution deaths, indoor pollution deaths, outdoor pollution deaths, and ozone pollution deaths will all be analyzed through 1990 to 2017. There is also the possibility that socio-economic factors (i.e. income, SDI) can be considered, since the entity attribute contains such data.

The third dataset is **Total Population of the World from 1950 - 2100.**

Link: `https://population.un.org/wpp/Download/Standard/CSV/`

The only relevant data from this dataset is the total population at any given year before 2018 with medium projection.

All of these datasets have very human readable data. They are also all in a very similar format, so little to no cleaning up will need to be done in order to use the data.

3 Investigative Questions

Approximately how many people have died from air pollution in the world since 1990?

Does the amount of emissions affect the amount of people dying from air pollution?

Does population growth affect the amount of emissions?

4 Schema: Relations

- Emissions(entity, year, tonnes)

  A tuple in this relation represents a source of CO2 emissions during a particular year. *entity* is the body which the emissions stemmed from. This can be a country, geographical region, or another source, such as international transport. *year* is the year which the emissions happened. *tonnes* is the amount of emissions, in tonnes.

- PollutionDeathsByLocation(location, year, totalDeaths, outdoorDeaths, indoorDeaths, ozoneDeaths)

  A tuple in this relation represents the deaths due to all types of pollution in a particular year. *location* represents where the deaths occurred. This can be a country or a geographical region. *year* is the year which the deaths occurred. *totalDeaths* is the number of total deaths due to all types of pollution, per 100 thousand people. The other attributes describe the type of pollution specifically, rather then them all.

- PollutionDeathsBySDI(sdiRank, year,totalDeaths, outdoorDeaths, indoorDeaths, ozoneDeaths)

  A tuple in this relation represents the deaths due to all types of pollution, by the Spatial Data Infrastructure of where the death occurred. *sdiRank* is the SDI rank value from the set{”Low”, ”Low-middle”, ”High-middle”, ”High”} year is the year which the deaths occurred. *totalDeaths* is the number of total deaths due to all types of pollution, per 100 thousand people. The other attributes describe the type of pollution specifically, rather then them all.

- Population(country, year, popValue)

  A tuple in this relation represents the population of any country (”World” is a country), in thousands. *country* represents the country, *year* represents the year and *popValue* represents the population in thousands.

5 Schema: Integrity Constraints

- $\Pi_{sdiRank}$PollutionDeathsBySDI ⊆{”Low”, ”Low-middle”, ”High-middle”, ”High”}
- $\sigma_{tonnes<0}$Emissions $= \emptyset$
- $\sigma_{totalDeaths<0 \vee outdoorDeaths<0 \vee indoorDeaths<0 \vee ozoneDeaths<0}$PollutionDeathsByLocation $= \emptyset$
- $\sigma_{totalDeaths<0 \vee outdoorDeaths<0 \vee indoorDeaths<0 \vee ozoneDeaths<0}$PollutionDeathsBySDI $= \emptyset$
- PollutionDeathsByLocation[location] ⊆{the set of all countries or geographic regions in string form}
- PollutionDeathsBySDI[location] ⊆{all countries or geographic regions in string form}
- Emissions[entity] ⊆{all countries or geographic regions in string form} ∪ {”International Transport”}

- $\sigma_{year1<0 \vee year2<0 \vee year3<0 \vee year4<0}(\varphi_{year1}(\Pi_{year}\text{PollutionDeathsBySDI}) \bowtie \varphi_{year2}(\Pi_{year}\text{PollutionDeathsByLocation}) \bowtie \varphi_{year3}(\Pi_{year}\text{Emissions}) \bowtie \varphi_{year4}(\Pi_{year}\text{Population})) = \emptyset$

6 Data Dictionaries

Note: due to restrictions on the sizes of cells in latex, please refer to descriptions under the 'Schema: Relations' section for more details on attributes

Emissions

attribute	description	type	required	default
entity	source of emissions	TEXT	yes	n/a
year	the year which the emissions happened	INT	yes	n/a
tonnes	amount of emissions	INT	yes	0

PollutionDeathsByLocation

attribute	description	type	required	default
location	source of emissions	TEXT	yes	n/a
year	the year which the emissions happened	INT	yes	n/a
totalDeaths	total deaths resulting from all types of pollution	INT	yes	n/a
outdoorDeaths	deaths resulting from outdoor pollution	INT	yes	n/a
indoorDeaths	deaths resulting from indoor pollution	INT	yes	n/a
ozoneDeaths	deaths resulting from ozone pollution	INT	yes	n/a

PollutionDeathsBySDI

attribute	description	type	required	default
sdiRank	ranking of a location's SDI	TEXT	yes	n/a
year	the year which the emissions happened	INT	yes	n/a
totalDeaths	total deaths resulting from all types of pollution	INT	yes	n/a
outdoorDeaths	deaths resulting from outdoor pollution	INT	yes	n/a
indoorDeaths	deaths resulting from indoor pollution	INT	yes	n/a
ozoneDeaths	deaths resulting from ozone pollution	INT	yes	n/a

Population

attribute	description	type	required	default
country	represents a country	TEXT	yes	n/a
year	the year for the estimated population of a country	INT	yes	n/a
popValue	the value of the population of a country in a given year	INT	yes	n/a

7 Justification

It was decided that four tables is all that would be needed. The Emissions data stayed completely the same as in the dataset set since it is nicely organized, and it wouldn't make sense to combine it with anything else. The SDI data was separated from the Location data from the pollution deaths dataset, because was is a geographical identifier, while the other is a socioeconomic factor. The integrity constraints chosen are realistic and do not restrict anything that could happen in a real life situation. For example, we do not allow negative emissions, or negative years. There are no constraints between the years and locations of the different relations, because it is totally plausible that there could be data from one year for one relation, but not have data for the same year in another relation. Same with locations; some of the datasets have geographical regions that are not included in the other data sets, which should be allowed since no assumptions can be made about what data is or isn't present. It is however required that location attributes and the entity attribute of each relation are actually locations or something that is a source of emissions, so that the relations are not nonsensical. The SDI Rank constraint exists for the same reason.