CSC343 PROJECT PHASE 1

DHRUOV BHATIA & HAOCHENG HU

THE DOMAIN

Our domain is suicide rate and its relationship with economic indicators and political climate.

THE DATASET:

LINKS

We procured our data from online data community Kaggle (Szamil, 2018), (Rajarshi, 2017), (psterk, 2020) and Wikipedia *(List of countries by system of government*, 2020)*, (List of countries and dependencies by area*, 2020).

(Note: A full, formatted list of references is at the end of the document.)

RELEVANT INFORMATION

We expect to use the following information out of the data we have acquired:

- Statistics on suicide rates and life expectancy, separated by country, year, and demographic group (age, sex).
- Information on government type and land area for each country.
- Data on GDP per capita, population, income inequality, and political climate, separated by country and year.

LEARNING TO DO

There isn't much learning we have to do in order to interpret the data. We must do some research in order to fully understand the underlying implications of the data, such as on differences between government types, or the different components of the Gini coefficient and democracy index. However, the columns seem to be structured in a way that makes intuitive sense. This is not to say that there is no work to be done however, which we will detail in the next section.

CLEANING TO DO

There is a considerable amount of cleaning we will need to do. For starters, the data in its current form is not stored in a way that represents a good schema. We need to move quite a few columns around and form new tables with the existing data in order to maintain good structure and reduce redundancy.

Another issue is that the data in each column is not properly formatted. As not all of our data was drawn from the same source, and we procured some of our data from places like Wikipedia tables, there was no easy way to download the data in a convenient .csv format, so we had to enlist the help of some scripts to scrape the data (yes we collected the data already). As a result, not all data is consistent. For example, "Canada" in a column of one table could sometimes be called "CANADA" or "Canada (Country)" in another, and we need to find a way to make these values consistent.

Several entries in our tables also contain irrelevant information, like notes, thumbnails, and links, which we will have to remove to make the data easier to work with.

Investigative Questions:

1. Do suicide rates vary across demographic groups (e.g. sex, age group) within countries?
2. Do suicide rates differ relative to economic or quality-of-life indicators, such as GDP per capita, income inequality, and life expectancy across countries?
3. Is there any link between suicide rates and differences in political regimes and political environments between countries?

The Schema:

Tables:

Continent(conID[smallserial], conName[varchar(255)])

Country(cID[smallserial], cName[varchar(255)], conID[smallserial], gID[smallserial], landArea[real])

Government(gID[smallserial], government[varchar(255)])

Economy(cID[smallserial], year[smallserial], GDP[serial], population[serial], GDPcapita[serial], Gini[real], lifespan[real], demoIndex[real])

Age(aID[smallserial], ageGroup[varchar(255)], population[serial])

Suicide(cID[smallserial], year[smallserial], aID[smallserial], suicides[integer], population[serial], sRate[real])

Referential Integrity Constraints:

Country[conID] ⊆ Continent[conID]

Country[gID] ⊆ Government[gID]

Economy[cID] ⊆ Country[cID]

Suicide[cID] ⊆ Country[cID]

Suicide[aID] ⊆ Age[aID]

Suicide[year] ⊆ Economy[year]

References

List of countries and dependencies by area. (2020, September). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_area

List of countries by system of government. (2020, October). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/List_of_countries_by_system_of_government

psterk. (2020, April). *GapMinder - Income Inequality*. Retrieved from Kaggle: https://www.kaggle.com/psterk/income-inequality

Rajarshi, K. (2017). *Life Expectancy (WHO)*. Retrieved from Kaggle: https://www.kaggle.com/kumarajarshi/life-expectancy-who

Szamil. (2018). *WHO Suicide Statistics*. Retrieved from Kaggle: https://www.kaggle.com/szamil/who-suicide-statistics