CSC343 Project Phase 1

Dru Bhatia & Haocheng Hu

**The Dataset:**

**Links**

We procured our data from the following sources: World Health Organization (WHO), Kaggle, and Wikipedia.

*(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, separated by country and year.

**Learning To Do**

There isn’t much learning we have to do on the data. We do have to understand some of the underlying economic implications of the data, but 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 decent amount of cleaning that we need to do to the data. 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 to help with things like maintaining good structure and reducing redundancy.

Another issue with the data is that the data in each column is not properly formatted. As we procured some of our data from things like Wikipedia tables, there was no easy way to download the data in a good 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 one column could sometimes be called “CANADA” or “Canada (Country)” in another, and we need to find a way to make these values consistent.

**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 indicators, such as GDP per capita and income inequality 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], GDPcapita[serial], Gini[real], population[serial], 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]

Domain. The domain you have chosen for your project.

•Dataset. A description of this dataset including:

~~–a link to the dataset that you have identified.~~

–what information in it is relevant to your project (there may be lots of irrelevant extra data too)

–any learning you will have to do in order to interpret the data

–any cleaning up you think you will have to do in order to use the data

•~~Questions. Your three investigative questions that you plan to answer using this dataset. It’s okay if theseevolve as you explore the data.~~

~~•Schema. Your relational schema. It’s also okay if this evolves over the phases of the project~~