CS 306 - Term Group Project Step 1

Group Name: Hasta la Vista Baby!

Group Members: Ahmet Emre Eser, Barkın Var, Beste Bayhan, Sadig Qara, Ecem Akın

Github Repo link: https://github.com/barkinvar/CS306-Hasta-La-Vista-Baby

Project Description:

The primary focus of our database is tobacco use in cigarette form and its effects on the overall life expectancy of the population. We are planning to analyze the data by finding correlations between geographical location and time data.

ER Diagram:

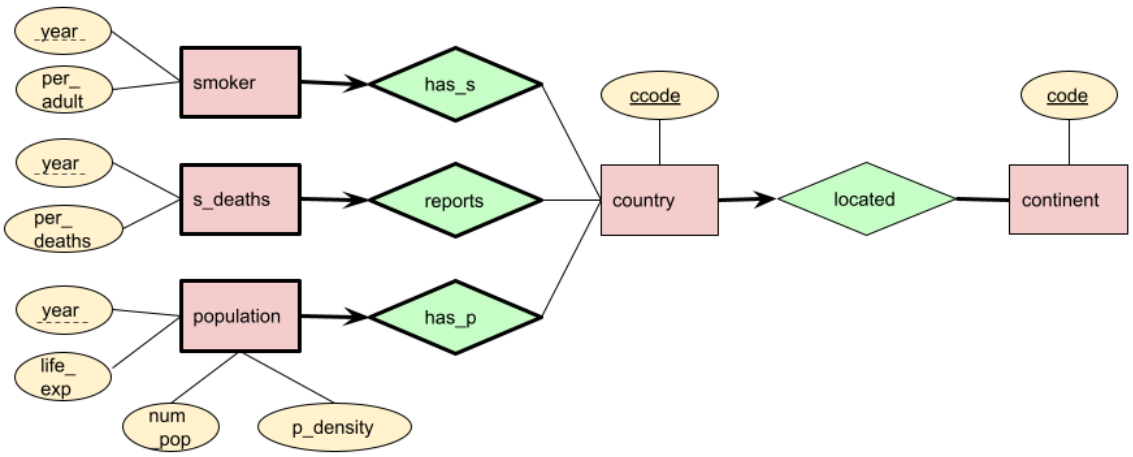

Table 1: Description of Database Entities and Attributes

Entity Name	**Description**	**Attributes**			
continent	Represents a real life continent	code (primary key)			
country	Represent an independent nation	ccode (primary key)			
population (weak)	Represents the population of a given country at a given year	year (partial key)	num_pop: total number of people living in a country	p_density: population density	life_exp: life expectancy of a country
smoker (weak)	Represents the smokers contained within a populace	year (partial key)	per_adult: percentage of smokers among the populace (older than 15)		
s_deaths (weak)	Represents smoking related deaths in a given country at a given year	year (partial key)	per_deaths: the percentage of smoking related deaths among all deaths recorded at a given year		

Clearing the Data:

We collected the datasets from the pages listed below:

- **https://ourworldindata.org/grapher/share-deaths-smoking**
- **https://ourworldindata.org/grapher/prevalence-of-tobacco-use-sdgs**
- **https://ourworldindata.org/covid-cases**
- **https://ourworldindata.org/grapher/number-of-deaths-by-risk-factor**
- **https://ourworldindata.org/world-population-growth**

- **https://ourworldindata.org/grapher/population-density?tab=table**
- **https://ourworldindata.org/life-expectancy**

We used excel software, especially the power pivot extension to clear and condense data we collated from different ourworldindata.com datasets. We decided to restrict the datasets to the year range 1980 - 2021 as most of the datasets don't contain any data prior to 1980. We also removed all data pertaining to any dependent county of an independent power such as Bahamas (US territory) as well as data related to any general region (such as world or Africa). This is a measure to prevent data redundancy as any data pertaining to offshore territories will also be contained within the data of the parent country and any data pertaining to a general geographical region could be calculated from the data of all countries contained within the region.

Rationale for ER Diagram:

Our ER diagram is centered around the country entity which contains the country code attribute as the primary key. Connected to the country entity via the "located" relationship is the continent entity. The only attribute within the continent entity is the code attribute which will be the primary key of the continent entity. The "located" relationship is bound with a thick line on the continent entity side signifying a total participation constraint on the continent entity. Hence each continent in the database cannot be bound with any less than 1 country via the "located" relationship.

Another constraint on the "located" relationship is a key constraint on the country entity side which is denoted by an arrow. The key constraint ensures that any instance of the country entity can at most engage in a "located" relationship with 1 continent. There are many exceptions to this rule, such as Turkey, however we imposed this constraint as a simplifying measure as recording every offshore territory of an independent nation would impose a storage burden on our database. Instead every nation will be set up in the relationship with the continent its capital is located on. The "located" relationship also has a participation constraint on the country entity side which ensures that every instance of the country entity is engaged in the relationship with at least one continent. This bolsters the integrity of the data as logically every country has to be located on at least one continent.

The country entity has 3 weak entities connected to itself which is denoted by the bold frame around the entities as well as the relationships that bind the weak entities to the country entity. The weak entities belonging to the country entity could be listed as: population, smokers and smoking caused deaths. We decided to bind these entities to the country entity since these notions cannot exist without belonging to a country. We decided not to keep these as attributes of the country entity as our dataset involves a lot of different values which could be logically gathered around the weak entities listed above.