GitHub repository for the project: https://github.com/ihagverdi/Database-Systems-Project

Group 2:

Hagverdi Ibrahimli 30014
Amil Kazımoğlu 27891
Nuri Kaan Özyer 26837
Raman Afravi 30061
Ayçelen Kaptan 28826

Project title and description:

Our project title is “Smoking kills” and the main idea behind the project is to examine the relation between smoking and health. The data we shall utilize in this database application covers the following topics: Affordability of cigarettes, consumption per smoker per day, cigarette sales per day, share of cancer deaths due to tobacco, share of deaths due to smoking, share of lung cancer deaths. The real world problems which this database application addresses narrow down to uncovering the negative relationship between the smoking rate of the individuals and their health.

Identifying our CSV files:

For the first step of the project, we downloaded 7 different CSV files from the source ourworldindata.org and analyzed them. Some of the CSV files were highly unorganized and contained much duplicate information. At this step, we tried to clean up all the CSV files and keep the information that we would need for our database application.
Firstly, from the “world_population.csv” file, we extract the information regarding the country populations for the dates ranging from 1950 to 2020. After removing the duplicates and some empty rows, we only kept the most important three columns we would use: Iso code, year, and population.

Firstly, from the “affordability_cigarettes.csv” file, we get the information about the percentage of GDP per capita required to purchase 2000 cigarettes of the most sold brand. Aftering removing the duplicates, we decided to keep all the columns and saved the file.

Secondly for the “share_deaths_smoking.csv” file, we processed the file by removing the empty dataset rows and deleting the duplicates from the dataset so that we would have a more concise dataset to make use of. Consequently, all of the columns were kept to be used in the later steps of the project.

Next, we started working on the “share_of_cancer_deaths_due_to_tobacco.csv” dataset. We removed the duplicates and made use of the percentage of cancer deaths subject to country and year. We reckoned that this dataset would be a major contribution to proving our point at the later steps of the project.

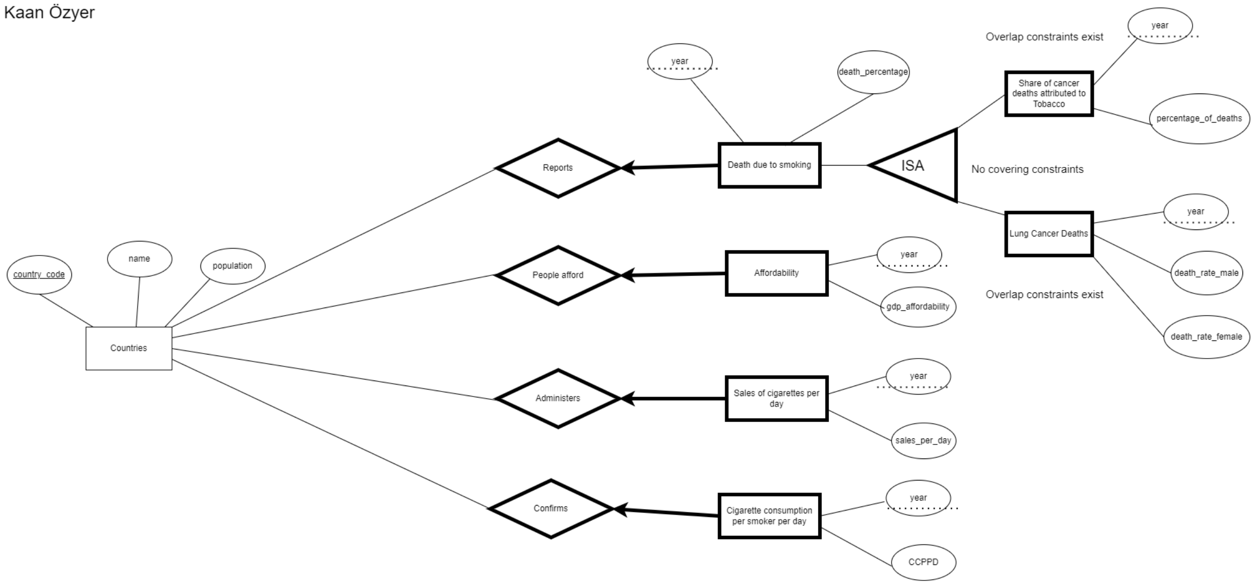
Then for the “consumption_per_smoker_daily.csv” file, we organized the data. After that we delete duplicates. We believe that we can find correlation between consumption of cigarettes and human health for our next phases.

Also, for the other dataset called “lung_cancer_deaths_per_100k_by_sex.csv”, there were some missing data and blank cells. Firstly we organized the data, then we deleted the data which contained blank fields and/or missing information. Following up on that, we removed the duplicates from the dataset.

Lastly, we utilized a dataset named “sales_of_cigarettess_per_adult_per_day.csv” which indicated the average count of cigarettes consumed per adult on a daily basis. Later we reorganized our data, removed the duplicates and made it ready for usage for the later steps.

<u>Explanation for the ER (Entity-Relationship) diagram</u>:

Group members:
Hagverdi Ibrahimli
Ayçelen Kaptan
Amil Kazımoğlu
Raman Afravi
Kaan Özyer



From the ER diagram, we can deduce the following structure: there are 7 entity sets, 4 relationship sets and 1 ISA hierarchy. In addition, the ER diagram shows 4 participation and 4 key constraints.

There is a relationship between “Countries” and the “Affordability” entity sets, that relationship is “People afford”, there is a participation constraint because all the records in affordability must belong to some country, and also there is a one to many relationship, because each country can have many affordability entities, but an affordability entity can only belong to one country.

Similarly, there is a relationship between “Countries” and the “Sales of cigarettes per adult per day” entity sets, that relationship is “Administers” and there is a participation constraint because all the records in sales must belong to some country, and also there is a one to many relationship, because countries can administer many different sales, but a given record in the sales entity set can only belong to one country.

There is also a relationship between the “Countries” entity set and the “Cigarette consumption per smoker per day” entity set and that relationship is “Confirms”. There is a participation constraint because without a country there will be no cigarette consumption, and also there is a one to many relationship, because a country can have many cigarette consumption entities, but a given cigarette consumption entity belongs to only one country.

The relationship between “Death due to smoking” and “Countries” is “Reports” and similarly there would be a participation constraint because without a country there will be no deaths due to smoking, moreover there is a one to many relationship, because countries can have many deaths due to smoking, however deaths due to smoking would only have a single country.

There is also an ISA hierarchy as the “Death due to smoking” entity set can be further divided into two sub entity sets: “Share of cancer deaths attributed to tobacco” and “Lung cancer deaths”. Furthermore, there exists an overlap constraint due to the fact that lung cancer death records should also appear in the more general cancer deaths entity set. However, there is no covering constraint as the reason of death may not be due to cancer.