# GROUP 8 - PROJECT STEP 1

Ayşe Sena Acar 29506

Barış Bakırdöven 29028

Ahmet Ozan Döm 29234

Mustafa Aybars Çebi 29242

Ebrar Berenay Yiğit 28824

## GITHUB LINK

https://github.com/Barisos22/Database-Project

# PROJECT NAME

Healthcare Expenses of Countries and their Consequences

# PROBLEM DEFINITION

The purpose of this project is to create a database that is designed for data analysis which will aim to find possible strong correlations between healthcare expenses of countries and major health concerns(mental health, cancer, air pollution) of today's world while taking population distribution into account.

## WORLD PROBLEMS ADDRESSED

1) How does the healthcare spending rate of countries affect the death rates of different types of cancer in those countries?
2) How does the healthcare spending rate of countries affect the air pollution death rate of that country?
3) How does the Air Pollution Death rate of a country relate with death rates of different types of cancer, with which cancer type is it correlated with the strongest(if it is correlated)?
4) How does the Air Pollution Death rate of a country relate with mental disorder rates of that country?
5) How does the distribution of the population to urban and rural regions affect the mental disorder rates of that population, which type of mental disorder is seen strongest in rural majority populations and which in urban majority populations?
6) How does the distribution of the population to urban and rural regions relate with the air pollution death rate of that country?

# DATABASE APPLICATION

MySQL Database Application will be used for this project.

# UTILIZED TABLES

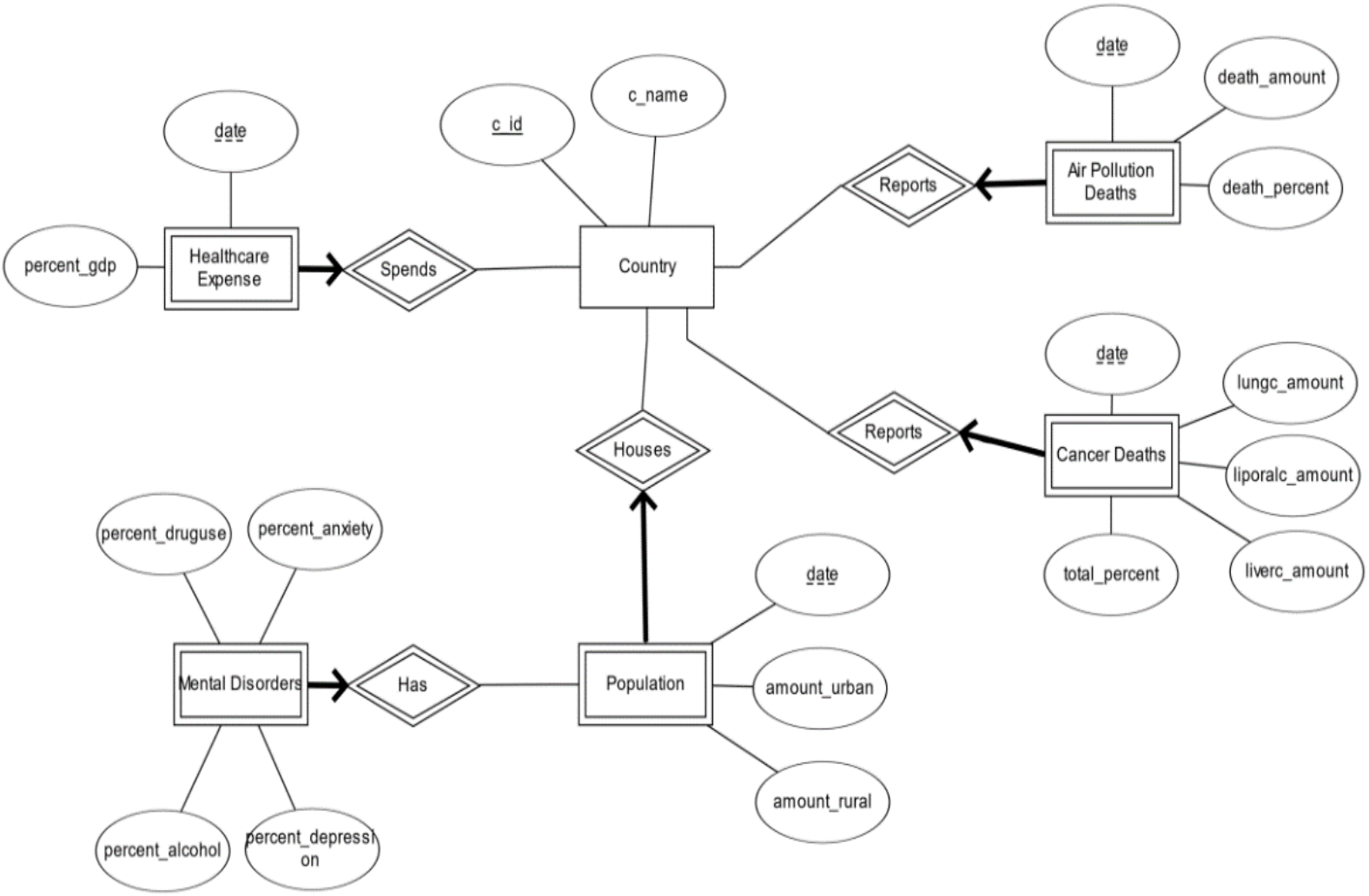
1) CANCER: https://ourworldindata.org/cancer
2) MENTAL HEALTH: https://ourworldindata.org/mental-health
3) HEALTHCARE EXPENSE: https://ourworldindata.org/financing-healthcare
4) AIR POLLUTION: https://ourworldindata.org/air-pollution
5) URBANIZATION: https://ourworldindata.org/urbanization

# FINAL TABLES

1) Country
    a. Name
    b. Id
2) Healthcare Expense
    a. Date (in years)
    b. Percentage of GDP that Countries spend on Healthcare
3) Population
    a. Date (in years)
    b. Urban Population amount
    c. Rural Population amount
4) Mental Disorders
    a. Drug use percentage
    b. Anxiety percentage
    c. Alcohol percentage
    d. Depression percentage
5) Cancer Deaths
    a. Date (in years)
    b. Total death percentage due to cancer
    c. Total death amount due to Lung Cancer
    d. Total death amount due to Lip/Oral Cancer
    e. Total death amount due to Liver Cancer
6) Air Pollution Deaths
    a. Date (in years)
    b. Total death amount due to air pollution
    c. Total death percentage due to air pollution

# STEPS INCLUDED

## 1) ER Diagram



## TABLES

1) **Country:**
    a. Strong entity with c_name and c_id attributes, c_id being the primary key
    b. Connected to weak entities called Healthcare Expense, Population, Cancer Deaths, Air Pollution Deaths with different relations.
2) **Healthcare Expense:**
    a. Weak Entity connected to Country entity with “Spends” Relation.
    b. Contains date and percent_gdp attributes.
    c. Date as its partial key, c_id as its foreign key (referencing Country).
    d. Key and Participation Constraints in order, as a requirement of being a weak entity.
3) **Population:**
    a. A weak Entity connected connected to the strong entity “Country” with “Houses” Relation.
    b. Connected to another weak entity called “Mental Disorders” with relation “Has”.
    c. Contains Date, amount_urban, amount_rural attributes.
    d. Date as its partial key, c_id as its foreign key(referencing Country).
    e. Key and Participation Constraints in order, as a requirement of being a weak entity.
4) **Mental Disorders:**
    a. A weak entity connected to “Population” weak entity with “Has” relation.
    b. Contains percent_druguse, percent_anxiety, percent_alcohol and percent_depression attributes
    c. No partial key, as it can be uniquely defined with Population’s partial key and c_id.
    d. Key and Participation Constraints in order, as a requirement of being a weak entity.
5) **Cancer Deaths:**
    a. A Weak entity connected to Strong entity “Country” with “Reports” Relation
    b. Contains Date, lungc_amount, liporalc_amount, liverc_amount, total_percent attributes
    c. Date as its partial key, c_id as its foreign key(referencing Country).
    d. Key and Participation Constraints in order, as a requirement of being a weak entity.
6) **Air Pollution Deaths:**
    a. A Weak entity connected to Strong entity “Country” with “Reports” Relation
    b. Contains Date, death_amount, death_percent attributes.
    c. Date as its partial key, c_id as its foreign key(referencing Country).
    d. Key and Participation Constraints in order, as a requirement of being a weak entity.

## 2) Excel Operations

- After designing the ER diagram, we gathered the various datasets that provided the information we required.

- Certain tables had extra information that was unnecessary for our research, which we deleted.

- As each table contained its independent country name fields, we first deleted this redundant information so that it is only contained in the Country table.

- Certain fields were associated with regions (such as Africa) or groups of people (such as middle-income). As we were interested only in country data, we located all rows in which the country code field was blank and deleted these rows.

- While there were none, we performed checks to ensure that no duplicate field was found, duplicate being defined as having the same country code and year.

- As the data we required to have in one table sometimes came from different data sets, we had to perform merge operations to acquire a single CSV file: this was accomplished by making use of the Transform Data function, allowing a merge of rows according to their country ID and date. No further cleanup was necessary as the two files merged without any issue or blank fields.