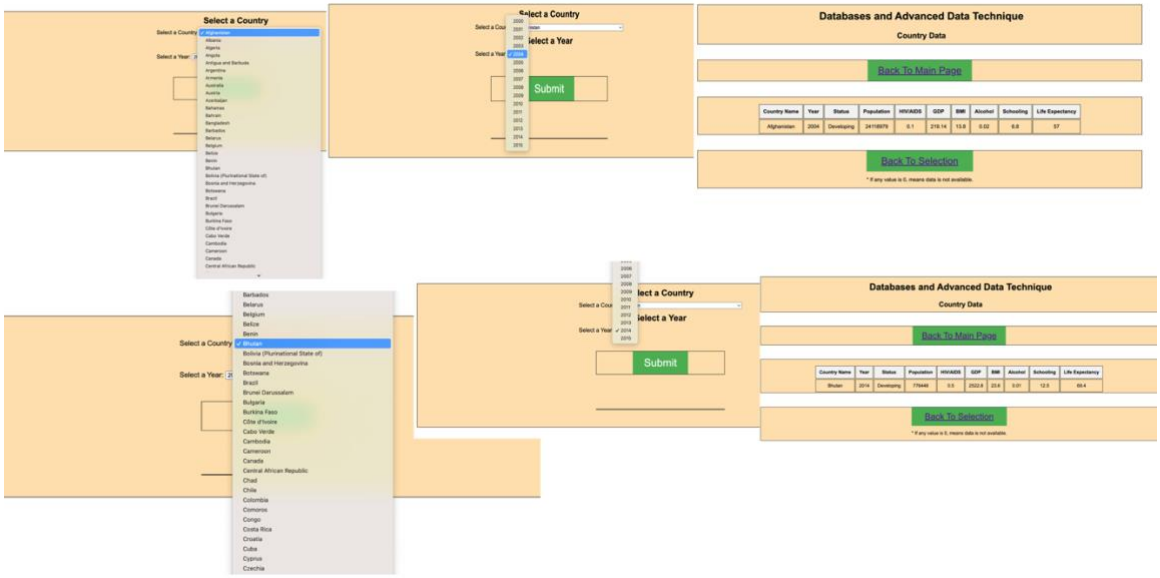
Databases and Advanced Data Technique
Mid-Term Assignment
July -2023

Sandeep Dharnia

Special Note:
Run command: npm start

In the virtual lab, there is an issue that I would like to point out before we proceed. In the HTML file, I used a 'select' tag to display the options to choose from, but it does not work in the virtual lab environment. However, I want to highlight that it works perfectly fine on my local machine, and I have included screenshots to demonstrate this.



For this assignment, I have explored open data sources in the domain of world population, specifically focusing on countries population, healthcare data and basic social healthcare requirements.
The data source I found is the “Countries Life Expectancy” on Kaggel.com by Amirhossein Mirzaei [1]. This dataset recently updated. In this dataset, shown countries population, status of the countries development, etc.

Quality:
The data in the Countries life expectancy Dataset is considered unreliable due to its source. It is available on an open-source data website where no information is provided regarding how the data was collected or its main source. However, upon comparing it with other sources, I have reached the conclusion that the quality of the dataset is fairly good, and we can rely on it to a certain extent."

Additionally, it is important to note that while the dataset may have some limitations, such as the lack of transparency regarding its collection methodology, the overall quality seems acceptable based on comparisons with alternative sources.

Detail:
The dataset 'Countries Life Expectancy' offers a wealth of detailed information about each country. It consists of 18 columns that provide extensive data on various aspects, including the year of study, population, and more.

This level of granularity enables in-depth analysis and exploration of the different levels of development among countries, their correlation with life expectancy, and other factors that may pose a threat to shorter life spans. For instance, it is evident that developed countries have a higher percentage of their population receiving vaccinations compared to developing countries.

The dataset's comprehensive nature allows researchers to delve into the intricate relationships between socio-economic indicators, healthcare access, and life expectancy, shedding light on critical areas for further investigation. By leveraging this dataset, we can gain valuable insights into the factors influencing life expectancy across different countries and potentially identify interventions to improve global health outcomes.

Documentation:
The Countries Life Expectancy' dataset is readily available on the website Kaggle, offering a comprehensive collection of information. The dataset is structured in a way that makes it easy to understand and interpret the data it contains. One notable advantage is the accompanying documentation, which provides detailed explanations of the data structure and the significance of each field.

Accessing the documentation is convenient, as it is located on the same page where the dataset can be downloaded. This ensures that users have immediate access to the necessary information to comprehend and utilize the dataset effectively.

By leveraging the clear documentation provided alongside the dataset, researchers and data analysts can confidently explore and analyze the 'Countries Life Expectancy' data, gaining valuable insights into the factors influencing life expectancy in different countries. This facilitates robust and informed decision-making processes aimed at addressing critical issues related to public health and well-being.

Interrelation:
Connecting the Life Expectancy dataset with other relevant datasets can greatly enhance the value of conducting more holistic analyses. By integrating this dataset with sources such as the vaccination program dataset or the literacy program dataset, it becomes possible to gain valuable insights into the overall Human Development Index (HDI) of the population.

While the Life Expectancy dataset may not directly provide links to other databases, the potential for integration with complementary datasets opens up new avenues for comprehensive research. By combining information from multiple sources, researchers can gain a deeper understanding of the factors influencing life expectancy and the broader socio-economic context in which it operates.

Integrating these datasets allows for a more comprehensive assessment of population health and well-being, researchers to identify correlations and uncover patterns that can inform targeted interventions and policies.

Use:
The Life Expectancy dataset holds immense potential for various applications. Researchers can leverage it to study disease trends, analyze causes of death, and generate valuable recommendations for future policies. Governments and policymakers can also utilize the dataset to inform the design of more effective and responsive programs and policies.

However, it is worth noting that the dataset may have certain limitations. For instance, it may not include information on the impact of the Hunger Index on life expectancy, which could be a crucial factor influencing health outcomes. Additionally, the dataset may not provide insights into per capita income, an important socio-economic indicator closely linked to overall well-being and life expectancy.

Discoverability:
Finding open-source data in the population domain proved to be relatively straightforward. I explored platforms such as Kaggle, U.S. Government platforms, the Singaporean Open Data source, and the World Health Organization. The Life Expectancy dataset was easily accessible and discovered on Kaggle.

Regarding the maintenance of the dataset, it is indicated on Kaggle that it has been recently updated, and it has benefitted from contributions by five individuals. This active involvement and updates from multiple contributors further enhance the dataset's reliability and usefulness.

Among the various options explored, the Life Expectancy dataset stood out as a superior choice due to its ease of access, availability, and up-to-date information. These factors make it a valuable resource for conducting population-based analyses and gaining insights into factors influencing life expectancy and related health outcomes.

The term of use:

1. What have you found:
   The Life Expectancy dataset is governed by the Community Data License Agreement - Sharing, Version 1.0 [2].

2 Where did you find the information:

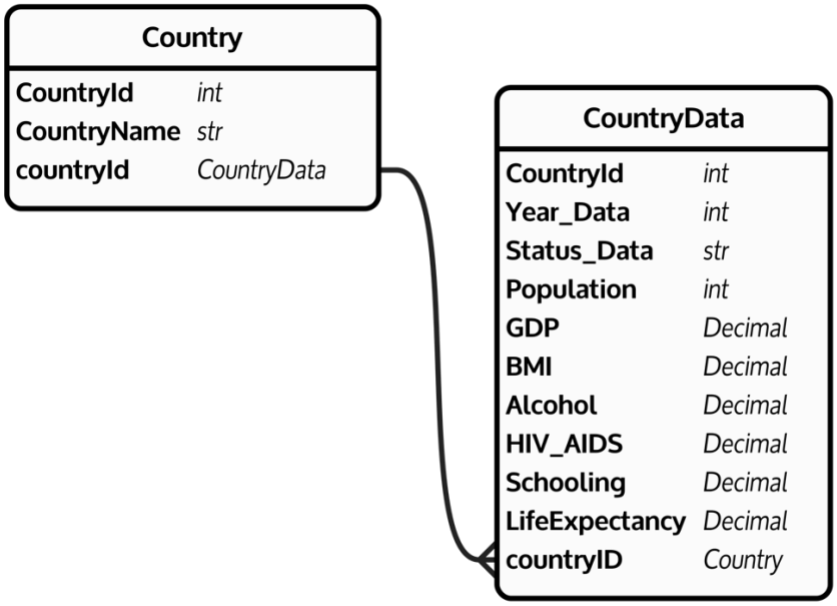
The information regarding the dataset's use and license is prominently displayed on the same page where the download button is located.

3 What is not allowed by terms and conditions:
Since the Life Expectancy dataset does not have a centralized license, it is crucial to refer to the terms of use specified by the individual sources to determine what actions are permitted and what restrictions may apply. Depending on the specific sources contributing to the dataset, there may be limitations on redistribution, commercial use, or modification of the data. These restrictions could impact the extent to which the dataset can be utilized, particularly in commercial applications or when creating derivative works.

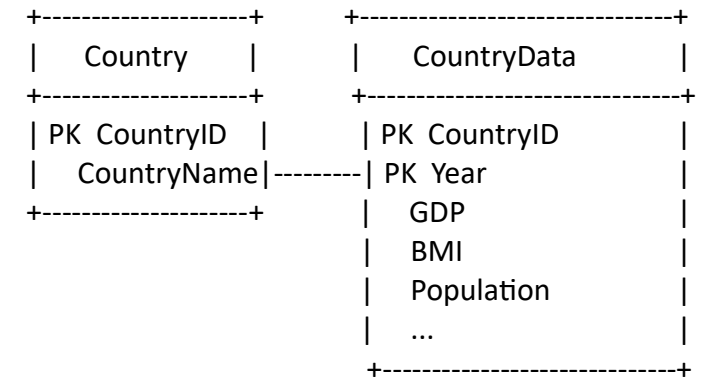
4 How easy is it for the licensing and rights information to get separated from the data:
Obtaining the dataset, along with the associated licensing information, was relatively straightforward as it is provided with the Community Data License Agreement. The licensing and rights information is readily accessible and does not appear to be separate from the dataset itself.

Task 2.

In our dataset there are two tables, "Country" and "CountryData," and it is in Third Normal Form (3NF), let's create an Entity-Relationship Diagram (ERD) to represent the relationship between these tables:



SQL View



In this diagram:
- "Country" table has attributes CountryID (Primary Key) and CountryName.
- "CountryData" table has attributes CountryID and Year (combined Primary Key), along with other attributes like GDP, BMI, Population, etc.

The relationship between the two tables is established through the "CountryID" attribute, which serves as a Foreign Key in the "CountryData" table, referencing the Primary Key of the "Country" table.

This ERD representation shows that the dataset is in Third Normal Form (3NF), as each table has a primary key, and non-key attributes are dependent on the entire primary key without any transitive dependencies.

Task 3.

In this exercise we have just dataset that has two tables, "Country" and "CountryData," and the "Year" data is in the "CountryData" table. We need to examine the dependencies to determine if the database is in Third Normal Form (3NF).

1. Table: Country

CountryID	CountryName
1	Afghanistan
2	Albania
3	…..

2. Table: CountryData

CountryID	Year_Data	Status_Data	Population	HIV/AIDS	GDP	BMI	Alcohol	Schooling	LifeExpectancy
1	2015	…….	….	….	….	….	….	….	….
1	2014	….	….	….	….	….	….	….	….
2	2015	….	….	….	….	….	….	….	….

In this scenario, we need to examine if there are any partial dependencies and transitive dependencies. A partial dependency occurs when a non-prime attribute (an attribute that is not part of the primary key) depends on only part of the primary key.

In the "CountryData" table, the primary key is the combination of "CountryID" and "Year." The other attributes like "GDP," "BMI," and "Population" depend on both "CountryID" and "Year," which means there are no partial dependencies.

A transitive dependency occurs when a non-prime attribute depends on another non-prime attribute. In this case, we don't have any transitive dependencies as well.

Since there are no partial or transitive dependencies in the "CountryData" table, and the primary key is formed by using all the attributes that define the table, the database is already in Third Normal Form (3NF).

let's analyze whether 4th Normal Form (4NF) is needed or not.

4th Normal Form (4NF) addresses multi-valued dependencies. It applies to situations where a table contains non-key attributes that depend on a subset of the primary key and are independent of the other parts of the primary key.

In the given dataset, we have two tables:

1. Country Table: (CountryID, CountryName)
   - Primary Key: CountryID

2. CountryData Table: (CountryID, Year, GDP, BMI, Population, ...)
   - Primary Key: (CountryID, Year)

In this scenario, there are no multi-valued dependencies in the CountryData table. Each non-key attribute (GDP, BMI, Population, etc.) depends on the entire primary key (CountryID, Year) and not on just a subset of it. Therefore, the CountryData table does not violate 4th Normal Form (4NF) requirements.

Since there are no multi-valued dependencies to address, there is no need to further normalize the tables beyond Third Normal Form (3NF). The dataset is adequately structured in 3NF and does not require 4NF for data integrity and redundancy purposes.

Reference:

1. Amirhossein Mirzaei updated Jul 2023 kaggel
Source: https://www.kaggle.com/datasets/amirhosseinmirzaie/countries-life-expectancy

2. Community Data License Agreement – Sharing, Version 1.0
Source: https://cdla.dev/sharing-1-0/