Healthcare Project Report
Mortality rate of ICD 9

Title

In this project, I am developing an interactive dashboard to visualize the **mortality rate** by country, Year, Age, Sex, and Cause of death of **ICD (International Classifications of Diseases) Ninth Revision**.

Abstract

While thinking of an appropriate project in Healthcare Analytics, I initially selected a COVID-19 dataset and calculated the mortality rate based on it but I felt like something was missing and I am not sure if the dataset would be appropriate for my project given that I am relative "newbies" to the world of Healthcare. Finally, I researched the Who mortality datasets which was relatively untouched. WHO provides mortality causes and rates for almost all the ICD clusters and 227 countries globally. But for this project, I have chosen only the ICD-9 cluster, 5 countries, and data from 1991 to 1997. I have provided two dashboards that will enable all of my target audience to go through my analysis.

Introduction

It is very tedious to maintain and track all the data on all the diseases in a healthcare organization without a proper classification system. This is where the ICD system comes in. Many causes of death are subdivided and classified into a cluster and this classification was termed **ICD**(**International Classification of Diseases**) as mentioned above followed by codes for different clusters of diseases. This classification was originally developed by the International Statistical Institute in 1893, but it is now maintained and updated by WHO regularly. Currently, the eleventh revision of the system is in effect and being used. Although the ICD-9 classification is no longer relevant, I felt that it would be appropriate to select this classification as the focus of our project gain more knowledge about ICD codes because of one of my subjects in Data Analytics for Business named Healthcare Analytics in my Postgraduate study.

I gained some knowledge about this coding system in this course in the form of my labs and I was intrigued to find more about the criterion/criteria of the division of ICD codes and so on.

Initially, when I was finding the project domain for my project, I started finding datasets of COVID-19 and Mortality rates and I found this page of mortality rates by age, year, gender, sex, and most important cause of death under the ICD-9 group on the WHO website.

Initially, I developed these questions to complete the initial part of my analysis.

1. Which country has the highest mortality rate?
2. What is the most popular cause of death in the ICD-9 group?
3. In which region or country this cause of death is popular?

4. Which age group has the cause of death found in question 2 and in which country they are located?

Here, I have taken a total of 3 different datasets and one documentation from the Who website.

This course in Healthcare Analytics has helped me get a high-level view of how healthcare organizations function. I determined that the healthcare organizations have divided different causes of death by using a coding system which is termed **ICD (International Classification of Diseases)** first adopted by The International Statistical Institute in 1893. Our project will only look at ICD-9 codes and their subsequent data. One of my objectives is to gain more knowledge about coding and what the different codes stand for as well as how they are used.

This project is centered on the mortality rate that shows the causes of death for the different ICD-9 codes.

Related Work

I tried to find many sites like sci-hub and other literature sites to complete this stage of my report, but I did not find any kind of specific work on ICD-9 cluster analysis, also no one analyzed this specific topic except the WHO (World health organization).

Methods

To complete this project efficiently, I used **six Data Analysis Phases** which are **Ask, Prepare, Process, Analyze, Share, Act,** and they are shown below:

1. Ask 

In this phase, I first define the problem statement as “My project will only look at the ICD-9 codes and their subsequent data. One of my objectives is to gain more knowledge about coding and what the different codes stand for as well as how they are used. The questions that I asked are based on the common thought process of people who are just introduced to healthcare and according to this I have chosen the datasets”.

Then using structured thinking, I asked some questions that can be my Analysis questions to find relevant insights like:

1. Which country has the highest mortality rate?
2. What is the most popular cause of death in the ICD-9 group?
3. In which region or country this cause of death is popular?
4. Which age group has the cause of death found in question 2 and in which country they are located?

2. Prepare 

In this second stage based on my topic, I found my dataset and other documents like how data is collected and who are the sample population of the dataset on Who.
I make sure that the data I am going to use is unbiased and credible by checking the document that they have provided with the dataset.

3. Process 

This phase includes all the cleaning and transformation processes. I have recorded all the changes that have been made during this project in this Changelog.

1. Firstly, I planned to use all records, but that did not work as there were some countries' data that were missing that were related to the Morticd9 dataset. So, I removed those countries.
2. Then for ICD-9, there were some countries under the USSR that had missing records for specific periods after they separated from the USSR. So, I removed those countries as well.
3. Data for China are only for selected urban and rural areas and represent less than 10% of all deaths occurring in the country which would not provide a true sample of the Chinese population. So, I removed China as well.
4. In the Morticd9 dataset, there is one column named “list” which represents the code for specific ICD codes. So, I must keep only 09A and 09B which are for the ICD-9th Revision.
5. After all these processing steps, the data that was remaining had missing data for population and deaths in most of the countries. So, I finally planned to keep only 5 countries that have all the data in a common period.
6. Therefore, I selected 5 countries:
    a. USA
    b. Canada
    c. Australia
    d. Cuba
    e. France

    and their data between 1991 to 1997

4. Analyze 

In this phase, I load the data to find different and interactive insights and to make data-driven decisions. For this, I use Tableau as a visualization tool. In which I made some calculated fields to find the mortality rate per 10,000 people.

I also created a total of 5 calculated fields in the Tableau worksheet that were loaded with the required datasets and the corresponding connections. These were as follows:

Mortality_Rate_1: This was the overall mortality rate which used the fields Deaths1 (Deaths of all ages) from the Morticd9 dataset and Pop1 (Population of all ages) from the Pop dataset.

The calculation formula used was:
(Deaths1/Pop1)*10000

Mortality_Rate_Adolescents: This was the mortality rate of adolescents (10-19 years old) which used the fields Deaths8(Deaths of ages 10-14) + Deaths9(Deaths of ages 15-19) from the Morticd9 dataset and Pop8(Population of ages 10-14) + Pop9(Population of ages 15-19) from the Pop dataset.

Mortality_Rate_Young: This was the mortality rate of young people (25-34 years old) which used the fields Deaths11 (Deaths of ages 25-29) + Deaths12(Deaths of ages 30-34) from the Morticd9 dataset and Pop11(Population of ages 25-29) + Pop12(Population of ages 30-34) from the Pop dataset.

Mortality_Rate_MiddleAged: This was the mortality rate of middle-aged people (45-54 years old) which used the fields Deaths15 (Deaths of ages 45-49) + Deaths16(Deaths of ages 50-54) from the Morticd9 dataset and Pop15(Population of ages 45-49) + Pop16(Population of ages 50-54) from the Pop dataset.

Mortality_Rate_Senior: This was the mortality rate of senior people (65-74 years old) which used the fields Deaths19 (Deaths of ages 65-69) + Deaths16(Deaths of ages 70-74) from the Morticd9 dataset and Pop19(Population of ages 65-69) + Pop20(Population of ages 70-74) from the Pop dataset.

For the bottom 4 calculated fields, the calculation formula used was like the one shown below with only the appropriate fields changed to get the corresponding mortality rate of that age group:

((Deaths8 + Deaths9)/(Pop8 + Pop9))*10000

All the insights and results that I got during this part of the analysis are provided in the next section.

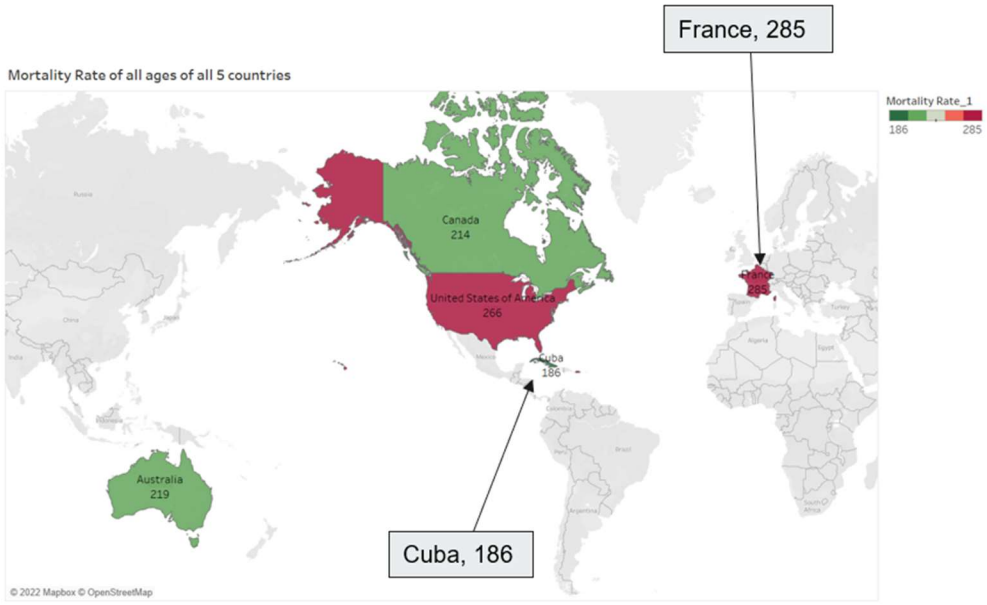
5. Share 

Here I made visuals that can be easy to understand for our stakeholders. I try to bring the data to life. Using the presentation and this report, I try to communicate and help others to understand what the story is.

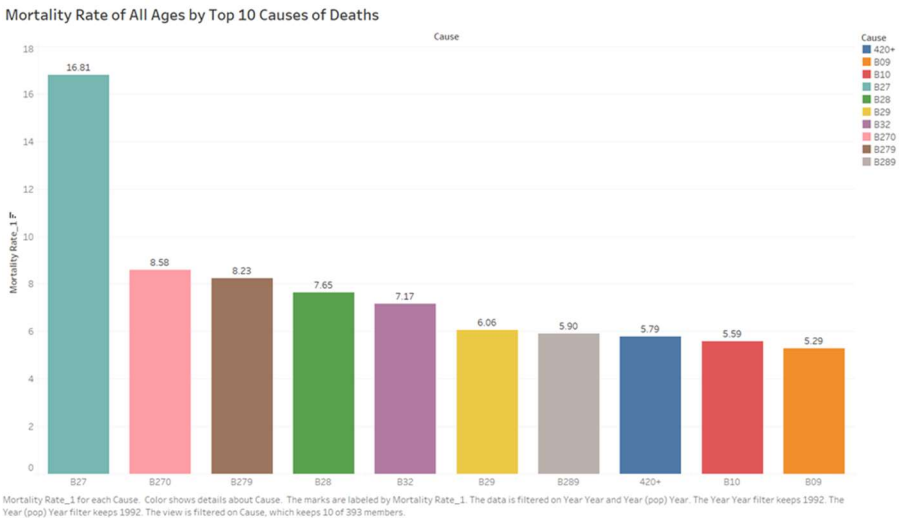
6. Act 

In this phase, I draw my conclusion and answer all the questions that I have asked in Phase-1 of this life cycle.

Results

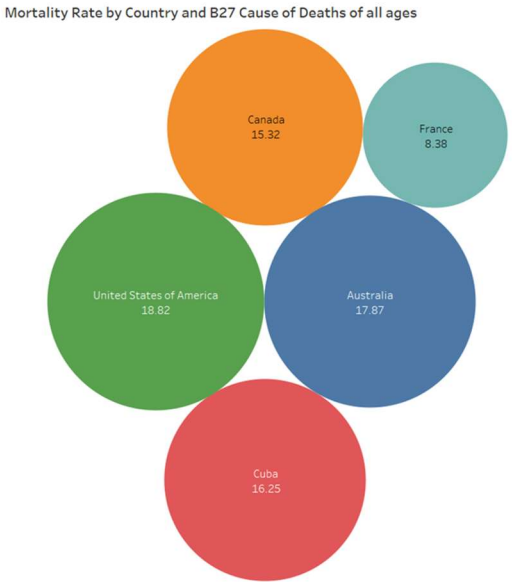


The map above answers the first analysis question that I had. I can see that France had the highest mortality rate of 285 for every 10,000 people in the year 1992. On the other hand, I can see that Cuba had the lowest mortality rate of 186 for every 10,000 people in the same year.

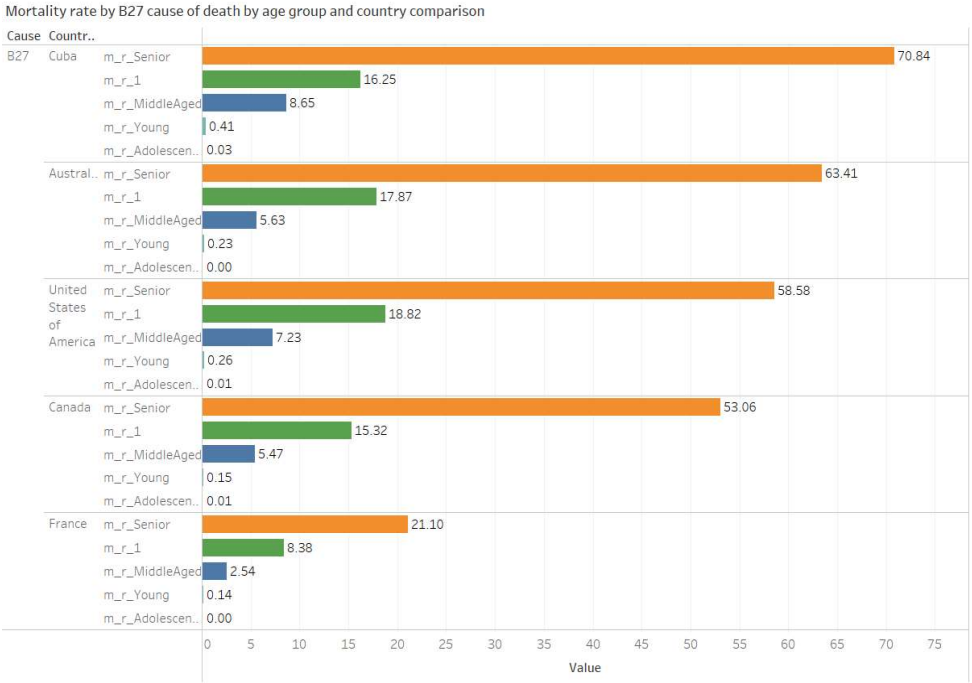


Our next question was to determine the cause responsible for the most deaths (highest mortality rate) in our entire dataset. As I can see, B27 - Ischaemic heart disease was the cause of death that was responsible for the highest mortality rate in 1993.

The bar graph shows me that 16.94 ~ 17 in every 10,000 people died due to Ischaemic heart disease (heart problems caused due to narrowed heart arteries) in 1993.



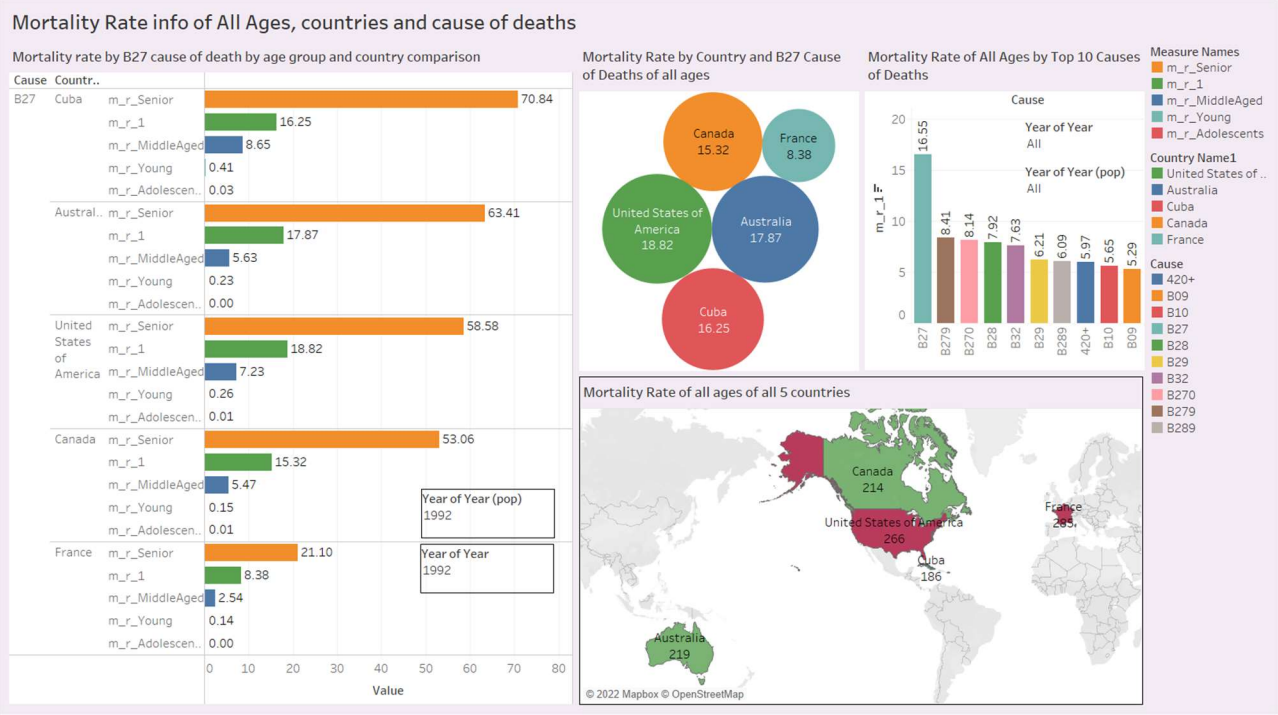
The bubble chart shown above represents the mortality rate by country and the top cause of death (B27) for all ages in 1992. As I can see, the country with the highest mortality rate was the USA with a mortality rate of 18.82~19 in every 10,000 people while the country with the lowest mortality rate was France with a mortality rate of 8.38~9 in every 10,000 people.



Given above is the comparison bar graph of mortality rates of the four age groups that I created separately from the data itself, they are as follows:

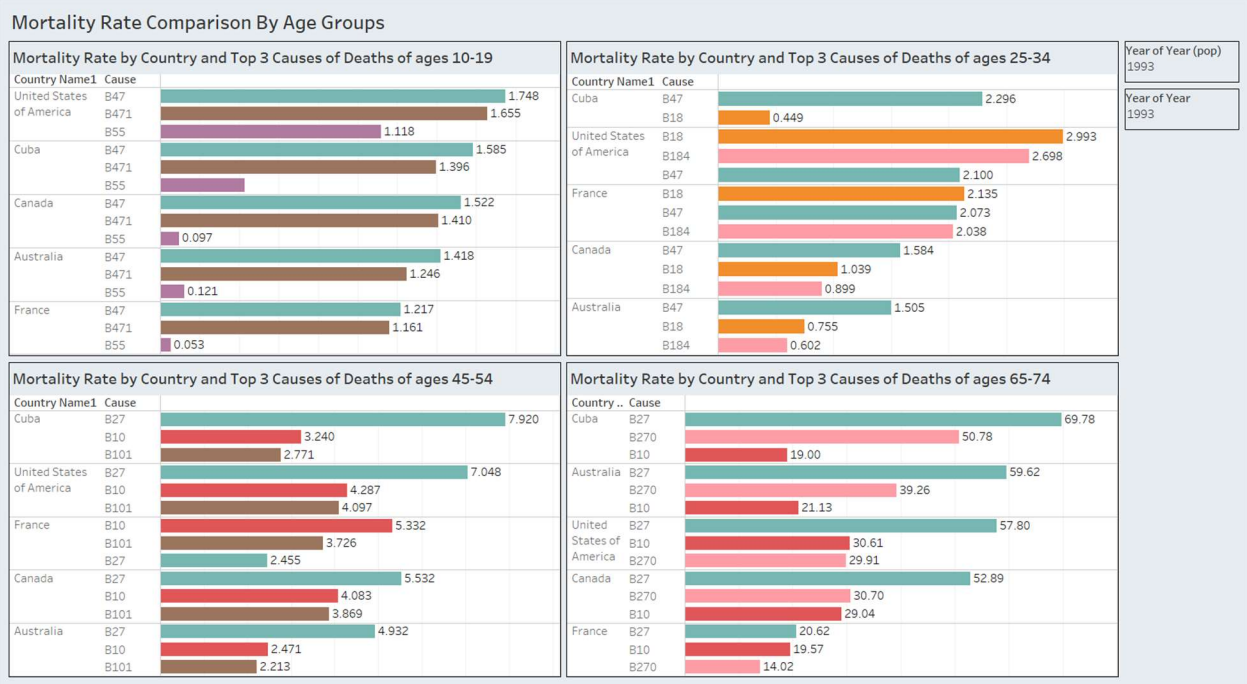
I. Adolescents - 10 to 19 years of age
II. Young - 25 to 34 years of age
III. Middle-Aged - 45 to 54 years of age
IV. Senior - 65 to 74 years of age

Along with this categorization, the above graph is also categorized by country and shows the data for the B27 cause of death only. As I can see, the senior age group has the highest mortality rate regardless of country, followed by the overall mortality rate, middle-aged, young, and finally adolescents.



The above dashboard shows all the charts shown previously together. This dashboard serves well to provide a comprehensive view of the insights I am trying to provide.

I can see that B27 - Ischaemic Heart disease is the top cause of death contributing the most to the mortality rate. Also, I can see that even though France had the lowest mortality rate for this cause of death, its overall mortality rate is the highest as shown on the map.



As explained previously, I created 4 additional calculated fields in Tableau other than the overall mortality rate. This dashboard shows the insights obtained from displaying the 4 age groups separately by country and top 3 causes of death.

I. For the senior age group (65-74 years old), B27 was the top cause of death regardless of country.
II. For the middle-aged group (45-54 years old), the top cause of death was B27 while for France it was B10.
III. For the adolescents' age group (10-19 years old), the top cause of death was B47 regardless of country.
IV. For the young age group (25-34 years old), I saw that in the USA and France B18 was the top cause of death while for the other countries it was B47.

Note: Even though the data shows the insights for only 1 year at a time, the major trends in the data are unchanged for each year in the entire period.

Discussion

From the results above, I saw that:

I. France had the highest overall mortality rate each year and Cuba had the lowest overall mortality rate each year in the entire period from 1991 to 1997.
II. B27 (Ischaemic Heart Disease) was the cause of death responsible for the highest mortality rates in each year in the entire period from 1991 to 1997.
III. When it comes to displaying the data by the top cause of death as well as country, I saw that the USA had the highest mortality rate and France had the lowest mortality rate for B27.

IV. When I categorize the data by age groups, the top cause of death (B27), and country; I can see that the senior age group's (65-74 years old) mortality rate was the highest regardless of country or year. Moreover, for the senior age group, Cuba had the highest mortality rates and France had the lowest mortality rates each year in the entire period from 1991 to 1997.

All of my analysis questions were answered satisfactorily from my analysis. In terms of failures/issues, I faced many issues and all the changes have been appropriately reported and documented in the Changelog.

I overcame these challenges by making changes in the dataset and curating the dataset to a specific set of 5 countries and a period of 1991 to 1997.

Conclusion

Based on the mortality rates of the countries and the responsible diseases, the countries would have had to change their strategies to deal with the prevention of diseases as far as possible as well as make sure that the mortality rate reduces. Improvements would entail the early detection and prevention of these diseases as well as the improvements in the healthcare system with any number of strategies like:

I. Increasing the share of healthcare expenditure of the country.
II. Actively Promoting healthcare as a field and encouraging the upcoming generation of doctors to be more vigilant as well as providing them with the correct knowledge and tools to deal with these and other diseases in the future.
III. Pooling of knowledge and resources with other countries to detect these diseases in their early stages as well as sharing of knowledge to improve healthcare by facilitating communication with other countries as well as global healthcare organizations like WHO.

A second study/project on the next periods of the data as well as other countries would provide a comprehensive view of the countries' performances globally as well as with the global average set by the world health institutions like WHO, CDC, etc.

References

1. WHO Mortality Database - WHO. (2022). Retrieved 13 June 2022, from https://www.who.int/data/data-collection-tools/who-mortality-database
2. Download SQL Server Management Studio (SSMS) - SQL Server Management Studio (SSMS). (2022). Retrieved 13 June 2022, from https://docs.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver15
3. Google Images. (2022). Retrieved 13 June 2022, from https://www.google.com/imghp?hl=en
4. Google Colaboratory. (2022). Retrieved 13 June 2022, from https://colab.research.google.com/?utm_source=scs-index
5. Write your best with Grammarly. (2022). Retrieved 13 June 2022, from https://www.grammarly.com/

6. How to merge multiple CSV files with Python - Softhints. (2022). Retrieved 13 June 2022, from https://blog.softhints.com/how-to-merge-multiple-csv-files-with-python/
7. Google Slides: Online Slideshow Maker | Google Workspace. (2022). Retrieved 13 June 2022, from https://www.google.ca/slides/about/
8. Google Docs: Sign-in. (2022). Retrieved 13 June 2022, from https://docs.google.com/document/u/0/
9. For visualization: Tableau

Appendices

I have added the whole report and required details of the project in this documentation. Code regarding EDA for Pop dataset and Morticd9 are available in these separate documentations.

Some of the cleaning and data extraction SQL code that I used are given below:

```
SELECT min(YEAR(year)) AS min_year, max(YEAR(year)) AS max_year , pop.Country,
country_codes.name
FROM pop
INNER join country_codes ON pop.Country = country_codes.country
WHERE YEAR(year) >= 1991 AND YEAR(year) <= 2000
GROUP BY pop.country,country_codes.name
ORDER BY min_year, max_year DESC
```

```
SELECT * FROM Morticd9
INNER join country_codes ON Morticd9.Country = country_codes.country
WHERE morticd9.Country IN (2450, 2090, 4080, 5020, 2150) AND YEAR(year) >= 1991 AND
YEAR(year) <= 1997
ORDER BY Morticd9.Country, Morticd9.Year
```

```
SELECT * FROM pop
INNER join country_codes ON pop.Country = country_codes.country
WHERE pop.Country IN (2450, 2090, 4080, 5020, 2150) AND YEAR(year) >= 1991 AND YEAR(year)
<= 1997
ORDER BY pop.Country, pop.Year
```