**Project Title: Observing Mortality Data from Northern Ireland, UK, Scotland, and England and Wales**

By: Faisal Khan

04th October 2020

**Introduction**

In the twentieth century the life expectancy in the UK as well as overall is increasing rate. This increase can be attributed to a variety of different factors, such as public health and medicine. Basic public health improvements such as safer food and water, improved nutrition, decreased maternal and infant mortality and decreases in infectious diseases, have caused life expectancy to increase. Medical improvements have also factored into reducing life expectancy, such as the immunisation of millions of people against smallpox, polio, pneumonia, and other major childhood killers like measles.

The general definition for life expectancy is the number of years a person is expected to live, and is calculated using statistics. However, mathematically, life expectancy is the expected number of years remaining for the person, or for a group of people, for a specific age. The reason that in the mathematical sense it is either for a person, or for a group of people, is because of the application that it is used for. Depending on the application, it may be required for an individual person, while other times a group of people.

Life expectancy is interlinked with mortality rate and mortality data. Mortality data consists of a wide variety of information. This includes the number of deaths, mortality rate (death rate), population exposure to risk, as well as more. Mostly these are calculated using statistics, and statistical methods, as well as various statistical models specifically used for mortality data.

As life expectancy is increasing, once would expect mortality rates (as well as mortality data in general) to decrease. However, this is not the case overall. In general mortality rates are increasing, and some instances there is more of a steady growth, and in other instances there is more of an exponential growth. Some examples include viruses becoming resist to medicines, an increase in crime and drug related deaths, as well as others. In the current climate, mortality data is becoming more vital and sort out, due to the coronavirus.

As a result of this, I will be analysing mortality data for Northern Ireland, UK, Scotland, and England and Wales.

**Data Description**

The data that has been used is The Human Mortality Database (HMD).

The database began in the year 2000, but it was launched in May 2002, after its first stage of development. It was created to provide detailed mortality and population data to people interested in human mortality, which includes students, researchers, journalists, police analysts, and many others. Data is provided free of charge.

The data types that HMD provides includes:

* Live birth counts;
* Death counts;
* Population size (an estimate of the population size on January 1st);
* Exposure-to-risk;
* Death rates;
* Life tables

Given the data types that the HMD provides, some are fairly obvious as to what they are, while others are not as clear, so a brief description and explanation will now be given. The live birth counts are the birth counts. Death counts, population size and death rates are self-explanatory. The exposure-to-risk is an estimate of the population exposed to the risk of death for an age interval, and a time interval.

Overall, each of these data sets is organised by calendar year, from 1922 to 2018, gender, and age intervals. However not all of the data is used, a subset of five-year age intervals, and annual year intervals, has been used for UK mortality data. This is denoted by ‘5x1’ where the 5 refers to the age interval, and the 1 refers to yearly intervals. Age intervals are broken down into the following classes, [0, 1-4, 5-9, 10-14, …, 90-94, 95-99, 100-104].

**Data Selection**

For this project, the locations that I have chosen are:

* Northern Ireland
* United Kingdom
* Scotland
* England and Wales

For each location, I collected each of the data types below:

* Exposure-to-risk (ETR)
* Death Rates
* Death Counts

For example, for Northern Ireland, I have 3 tables, Exposure-to-risk (ETR), Death Rates and Death Counts. I chose these data types because they are most related to mortality data and from them, I will be able to draw conclusions and make inferences at a later stage, if required.

Despite the fact that the entire data set could be used, I decided to take a subset.

|  |  |  |
| --- | --- | --- |
|  | **Years** | **Age Groups** |
| Entire Data Set | 1922 to 2018 | [0, 1-4, 5-9, 10-14, …, 90-94, 95-99, 100-104] |
| Subset chosen for my project | 1990 to 2018 | [0, 1-4, 5-9, 10-14, …, 90-94] |
| Reasoning | * Yearly interval was chosen to improve accuracy as it will decrease large fluctuations. * I believe this sample size provided enough data. * From a processing point of view, it would also be quicker to run the data if there were less data. | * A five-year age interval was chosen as it is not too small of an interval. * I removed the age groups 95-99 and 100-104, since for the elderly population there are some extreme values (and in some case no values). * From a processing point of view, it would also be quicker to run the data if there were less data. |

For example, for the location Northern Ireland, I have 3 tables, Exposure-to-risk (ETR), Death Rates and Death Counts, and each table has the yearly interval 1990 to 2018, and each year has the age groups [0, 1-4, 5-9, 10-14, …, 90-94]. Each table (Exposure-to-risk (ETR), Death Rates and Death Counts) consists of 580 rows.

**Data Cleaning**

The HMD provided a vast amount of reliable, accurate information however in order to ingest the data, cleaning was required. The data files provided are tab-delimited text (ASCII) files, and as such I had to convert the files to csv. Although there were other data sets available online, they did not provide as much detailed information, and so I decided to persevere with this data set. In addition to this, I have used this data before in a project. The actual data set itself is very detailed, and it would result in more accurate inferences and conclusions.

**Databases Used**

In this project I used 3 databases in total.

1. MySQL
2. PostreSQL
3. Microsoft SQL Server

Each of these have their own advantages and disadvantages.

|  |  |  |
| --- | --- | --- |
|  | Advantages | Disadvantages |
| MySQL | * The most secure and reliable database management system. * MySQL offers unmatched scalability. * Easy and simple to get up and running with the average download and installation time being less than 30 minutes. | * MySQL does not support a very large database size as efficiently. * There are a few stability issues. * The development is not community driven so it has lagged behind. * There are some limitations. |
| PostgreSQL | * The world’s most advanced open source database. * A powerful relational database. * Easy to learn. * Very robust and powerful. | * Users need to tweak PostgreSQL more for faster speed. * The low budget hosting sites do no support PostgreSQL. * It is not owned by a single company. |
| Microsoft SQL Server | * It includes enterprise-level professional management software database. * Has a number of features that promote restoration and data recovery. * High level security. | * For more of the advanced databased applications and features, you need you need to pay for the higher versions of the software. * Restricted compatibility. * Hardware restrictions. |

**Method – MySQL**

|  |
| --- |
| **For Death Rates** |
| Created the Death Rates database  create database deathrates; |
| Used the database  use deathrates; |
| Created the tables  create table deathratesenw(Year int, Age varchar(7), Female decimal(5,5), Male decimal(5,5), Total decimal(5,5));  create table deathratesni(Year int, Age varchar(7), Female decimal(5,5), Male decimal(5,5), Total decimal(5,5));  create table deathratessco(Year int, Age varchar(7), Female decimal(5,5), Male decimal(5,5), Total decimal(5,5));  create table deathratesuk(Year int, Age varchar(7), Female decimal(5,5), Male decimal(5,5), Total decimal(5,5)); |
| Found the directory of the files  /home/faisal/Documents/Data/DeathRates/DeathRatesENW.csv  /home/faisal/Documents/Data/DeathRates/DeathRatesNI.csv  /home/faisal/Documents/Data/DeathRates/DeathRatesSCO.csv  /home/faisal/Documents/Data/DeathRates/DeathRatesUK.csv |
| Loaded the data from the files and inserted them into the created table  LOAD DATA LOCAL INFILE '/home/faisal/Documents/Data/DeathRates/DeathRatesENW.csv' replace INTO TABLE deathratesenw FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n';  LOAD DATA LOCAL INFILE '/home/faisal/Documents/Data/DeathRates/DeathRatesNI.csv' replace INTO TABLE deathratesni FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n';  LOAD DATA LOCAL INFILE '/home/faisal/Documents/Data/DeathRates/DeathRatesSCO.csv' replace INTO TABLE deathratessco FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n';  LOAD DATA LOCAL INFILE '/home/faisal/Documents/Data/DeathRates/DeathRatesUK.csv' replace INTO TABLE deathratesuk FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n'; |

|  |
| --- |
| **For Deaths** |
| Create the Deaths database  create database deaths; |
| Used the database  use deaths; |
| Created the tables  create table deathsenw(Year int, Age varchar(7), Female int, Male int, Total int);  create table deathsni(Year int, Age varchar(7), Female int, Male int, Total int);  create table deathssco(Year int, Age varchar(7), Female int, Male int, Total int);  create table deathsuk(Year int, Age varchar(7), Female int, Male int, Total int); |
| Found the directory of the files  /home/faisal/Documents/Data/Deaths/DeathsENW.csv  /home/faisal/Documents/Data/Deaths/DeathsNI.csv  /home/faisal/Documents/Data/Deaths/DeathsSCO.csv  /home/faisal/Documents/Data/Deaths/DeathsUK.csv |
| Loaded the data from the files and inserted them into the created table  LOAD DATA LOCAL INFILE '/home/faisal/Documents/Data/Deaths/DeathsENW.csv' replace INTO TABLE deathsenw FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n';  LOAD DATA LOCAL INFILE '/home/faisal/Documents/Data/Deaths/DeathsNI.csv' replace INTO TABLE deathsni FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n';  LOAD DATA LOCAL INFILE '/home/faisal/Documents/Data/Deaths/DeathsSCO.csv' replace INTO TABLE deathssco FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n';  LOAD DATA LOCAL INFILE '/home/faisal/Documents/Data/Deaths/DeathsUK.csv' replace INTO TABLE deathsuk FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n'; |

|  |
| --- |
| **For ETR (Exposure to Risk)** |
| Create the ETR database  create database etr; |
| Used the database  use etr; |
| Created the tables  create table etrenw(Year int, Age varchar(7), Female decimal(15,2), Male decimal(15,2), Total decimal(15,2));  create table etrni(Year int, Age varchar(7), Female decimal(15,2), Male decimal(15,2), Total decimal(15,2));  create table etrsco(Year int, Age varchar(7), Female decimal(15,2), Male decimal(15,2), Total decimal(15,2));  create table etruk(Year int, Age varchar(7), Female decimal(15,2), Male decimal(15,2), Total decimal(15,2)); |
| Found the directory of the files  /home/faisal/Documents/Data/ETR/ETRENW.csv  /home/faisal/Documents/Data/ETR/ETRNI.csv  /home/faisal/Documents/Data/ETR/ETRSCO.csv  /home/faisal/Documents/Data/ETR/ETRUK.csv |
| Loaded the data from the files and inserted them into the created table  LOAD DATA LOCAL INFILE '/home/faisal/Documents/Data/ETR/ETRENW.csv' replace INTO TABLE etrenw FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n';  LOAD DATA LOCAL INFILE '/home/faisal/Documents/Data/ETR/ETRNI.csv' replace INTO TABLE etrni FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n';  LOAD DATA LOCAL INFILE '/home/faisal/Documents/Data/ETR/ETRSCO.csv' replace INTO TABLE etrsco FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n';  LOAD DATA LOCAL INFILE '/home/faisal/Documents/Data/ETR/ETRUK.csv' replace INTO TABLE etruk FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n'; |

**Method – PostreSQL**

Create a table with the name <table name>

tools=# CREATE TABLE <table name> (Year int, Age varchar(7), Female numeric(5,5), Male numeric (5,5), Total numeric (5,5));

Select Table Name

SELECT \* FROM <table name>

Import table from csv file

COPY <table name>

FROM ‘file directory/nameofcsvfile.csv’

DELIMITER ‘,’

CSV HEADER;

|  |
| --- |
| **For Death Rates** |
|  |

|  |
| --- |
| **For Deaths** |
|  |

|  |
| --- |
| **For ETR (Exposure to Risk)** |
|  |

**Method - Microsoft SQL Server**