DBS

Dublin Business School

excellence through learning

Course Title	:	Master's in Data Analytics
Module	:	Data Visualization
Student Name	:	YOGESH GUPTA
Student No	:	10386244
Date	:	MAY 2019

Introduction

The health of a country is the backbone of its growth, and there are many types of diseases growing rapidly in today's world and affecting many of the human beings. People die due to unawareness of the disease, that may have been causing death on a large scale. Researches can only be carried out if they have questions to work on.. OECD is Organisation for Economic Co-operation and Development. It is a platform for countries to compare policy experiences and finding out the current problems and provide plethora of data for analysis. In this project, data from the OECD website had been used which had been described in the Dataset section. This project will find out some facts about the different causes of death in the last 25 years. This visualization project contains the three basic charts i.e. a map, a stacked time series and a bubble chart. The map shows the OECD countries worldwide and their population in 2015 is been represented by the size of the bubble and figures will be visible while hovering on the individual bubble. This map is useful to identify a particular country which grabs the attention for the research and later this country can be selected in the selection wizard just below the map. By default, it shows the data for Australia but after selection of the country it automatically reloads and shows a stacked graph describing the percentage of people died by various cause of death like diseases of the circulatory system. From here one can identify the main causes of the death that had been emerged in the last few years and these causes can be selected in the next selection box which contains all causes of deaths, choosing multiple will be an advantage for the comparison among different causes of death or single cause can be selected to analyse its impact on the other countries which can be done in the last graph which is a bubble chart here size of the bubble indicates the total number of deaths and its location describe the country from the x-axis and the percentage it hold from the total deaths is symbolize by y-axis. Further, the year can be shifted from the slider followed by the selection panel which can help in the analysis of causes within the timeline. Hence from a country population to the emerging causes of deaths can be analysed by this project. Moreover, these charts can also be used in a reversed manner like first selecting a cause of death in the last chart and analysing in black box, it impacts on the different countries, then selecting a particular country and comparing the cause with other causes in the particular country and later finding out the population in the country from the map. These graphs are interactive and able to provide all the information needed for an analysis. All graphs had been created in the 'plotly' which is discussed descriptively in the Process section. In the Result section, some of the basic analysis had been shared.

Research Questions

- What is target population in OECD countries.
- We will try to identify the types of diseases which have a large impact in the last few years in OECD countries
- What is their status, along with the diseases, which are emerging in the OECD countries?

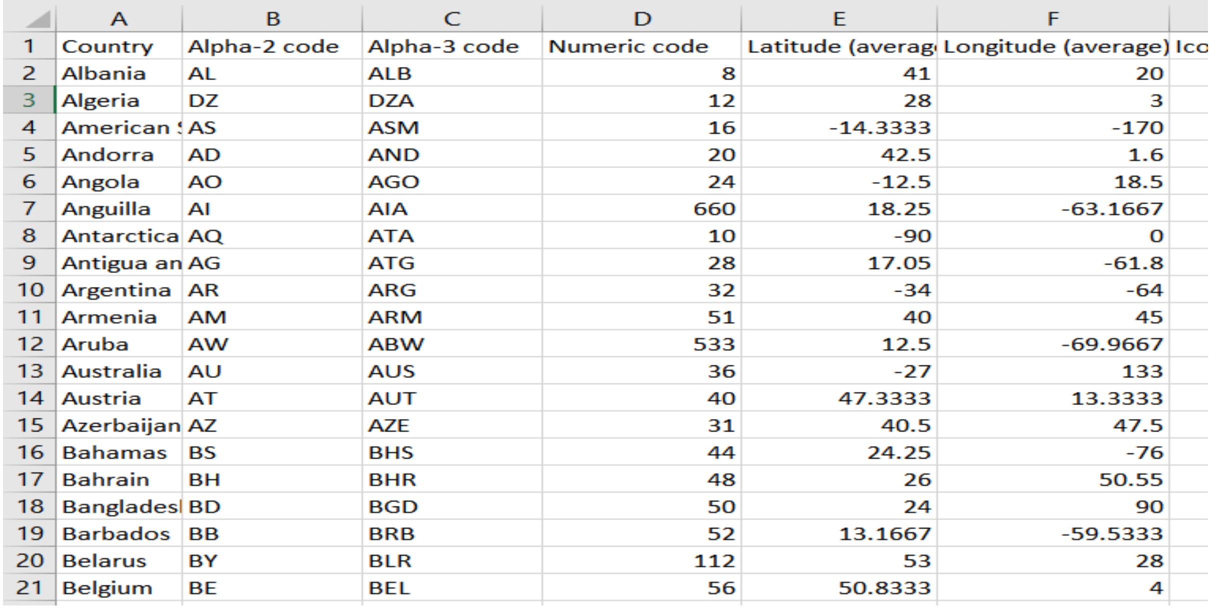
Dataset

As discussed in the previous section, this project uses the OECD data which can be extracted from the OECD website, link shared in the links section. Three CSV files have been used in this project.

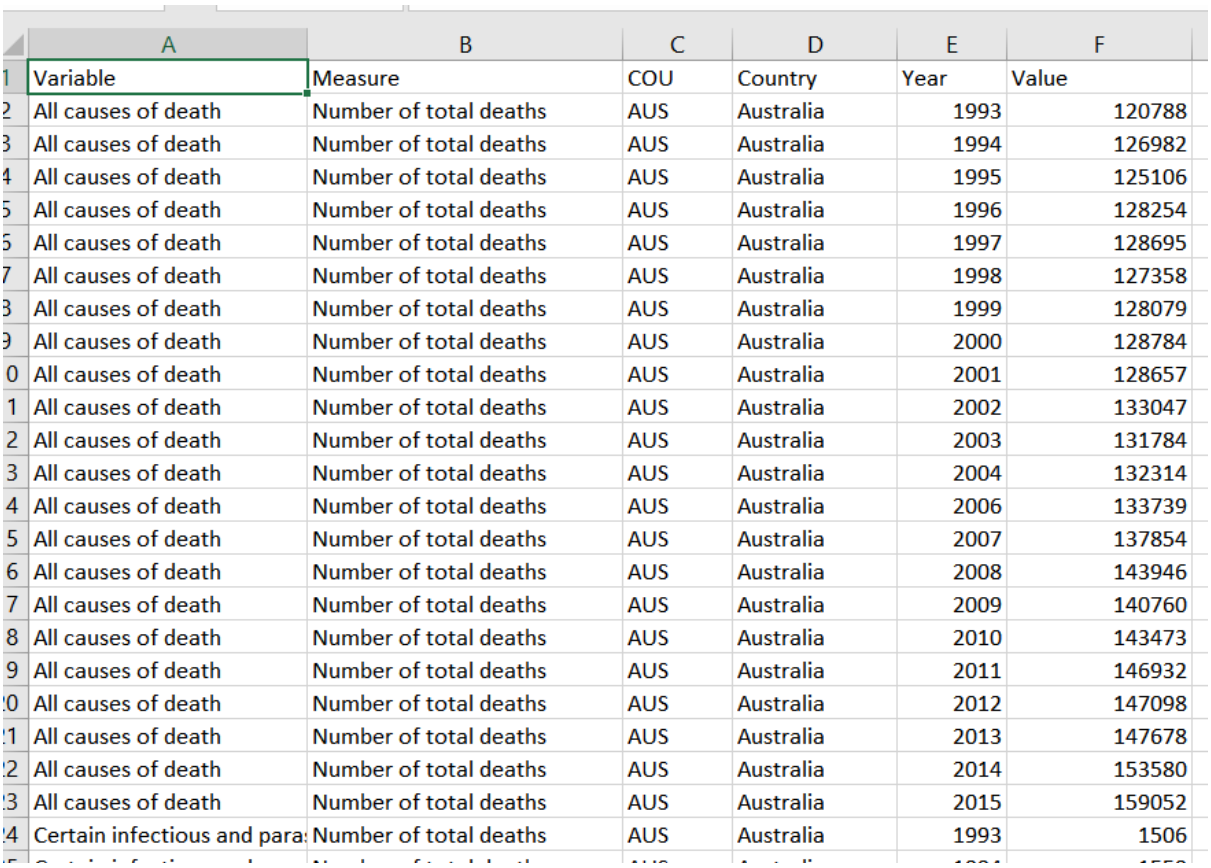
1. The comma separated file containing data for the causes of death in countries from 1993.

2. The sheet containing the population of the countries. These two files had been extracted from the OECD website by creating a customised query containing different filters.

3. The data sheet contains the average latitudinal and longitudinal data for the countries worldwide extracted from the Socrata open data. These both sources are open data sources and had been provided for the analysis openly hence no usage bound applies to the data. The data files are available on the project repository shared in the link section.



	A	B	C	D	E	F	
1	Country	Alpha-2 code	Alpha-3 code	Numeric code	Latitude (averag	Longitude (average)	Ico
2	Albania	AL	ALB	8	41	20	
3	Algeria	DZ	DZA	12	28	3	
4	American !	AS	ASM	16	-14.3333	-170	
5	Andorra	AD	AND	20	42.5	1.6	
6	Angola	AO	AGO	24	-12.5	18.5	
7	Anguilla	AI	AIA	660	18.25	-63.1667	
8	Antarctica	AQ	ATA	10	-90	0	
9	Antigua an	AG	ATG	28	17.05	-61.8	
10	Argentina	AR	ARG	32	-34	-64	
11	Armenia	AM	ARM	51	40	45	
12	Aruba	AW	ABW	533	12.5	-69.9667	
13	Australia	AU	AUS	36	-27	133	
14	Austria	AT	AUT	40	47.3333	13.3333	
15	Azerbaijan	AZ	AZE	31	40.5	47.5	
16	Bahamas	BS	BHS	44	24.25	-76	
17	Bahrain	BH	BHR	48	26	50.55	
18	Banglades	BD	BGD	50	24	90	
19	Barbados	BB	BRB	52	13.1667	-59.5333	
20	Belarus	BY	BLR	112	53	28	
21	Belgium	BE	BEL	56	50.8333	4	



	A	B	C	D	E	F
1	Variable	Measure	COU	Country	Year	Value
2	All causes of death	Number of total deaths	AUS	Australia	1993	120788
3	All causes of death	Number of total deaths	AUS	Australia	1994	126982
4	All causes of death	Number of total deaths	AUS	Australia	1995	125106
5	All causes of death	Number of total deaths	AUS	Australia	1996	128254
6	All causes of death	Number of total deaths	AUS	Australia	1997	128695
7	All causes of death	Number of total deaths	AUS	Australia	1998	127358
8	All causes of death	Number of total deaths	AUS	Australia	1999	128079
9	All causes of death	Number of total deaths	AUS	Australia	2000	128784
10	All causes of death	Number of total deaths	AUS	Australia	2001	128657
11	All causes of death	Number of total deaths	AUS	Australia	2002	133047
12	All causes of death	Number of total deaths	AUS	Australia	2003	131784
13	All causes of death	Number of total deaths	AUS	Australia	2004	132314
14	All causes of death	Number of total deaths	AUS	Australia	2006	133739
15	All causes of death	Number of total deaths	AUS	Australia	2007	137854
16	All causes of death	Number of total deaths	AUS	Australia	2008	143946
17	All causes of death	Number of total deaths	AUS	Australia	2009	140760
18	All causes of death	Number of total deaths	AUS	Australia	2010	143473
19	All causes of death	Number of total deaths	AUS	Australia	2011	146932
20	All causes of death	Number of total deaths	AUS	Australia	2012	147098
21	All causes of death	Number of total deaths	AUS	Australia	2013	147678
22	All causes of death	Number of total deaths	AUS	Australia	2014	153580
23	All causes of death	Number of total deaths	AUS	Australia	2015	159052
24	Certain infectious and para:	Number of total deaths	AUS	Australia	1993	1506

Process

This visualization project consists of three steps.

- Data Cleaning
- Data Visualization
- Reporting.

The complete process had been done in the Jupyter notebook in which the first section contains the importing of different libraries whose explanation had been given in the notebook in comments. In first step data files had been imported i.e. PopulationData(contain the population of countries in last few years), DeathData(contain the figures of death by different causes in last years), CountryData(contains the location variables of the countries) and certain query check had been implemented followed by the cleaning unnecessary data retrieved from the queries. The complete work had been done in the python using its different libraries. The data files had been imported using Pandas library in the data frame. Later on, certain queries had been performed on the data frames which involve null value check, data type check, data scaling check, useful data check and interlinking data check (all mentioned in the notebook cleaning section). Later on, the appropriate action had been performed on the data frame which involved removal of the certain extra column, changing data type of column which had been done using Numpy function and deletion of certain rows creating gap in the data in different file like data available in one file but interlinking data not available in others.

The second step involves the Data Visualization, Plotly has been used for the data visualization. It is a charting library built over D3 which is capable of providing beautiful interactive graphs. Coming on to graphs, as said earlier, this project involves three graphs.

• The first graph is a map which is built using mapbox, CountryData and PopulationData, mapbox is a provider of the custom online maps and it provides an extension for the Plolty which require a map box access token which can be generated from the mapbox site after registration. CountryData and PopulationData create a marker on the map showing the population of the OECD countries.

• The second graph is a stacked time series graph, followed by a widget which can be used for the country selection and automatically update graph when the value changes. The x-axis of the graph contains the time series and Y-axis shows the percentage of the deaths by cause from total population in that year, this had achieved by dividing death cause value by respective population in the year. The different stacks created use a colour palette that has been achieved by using a qualitative colour set from the Colorlover library.

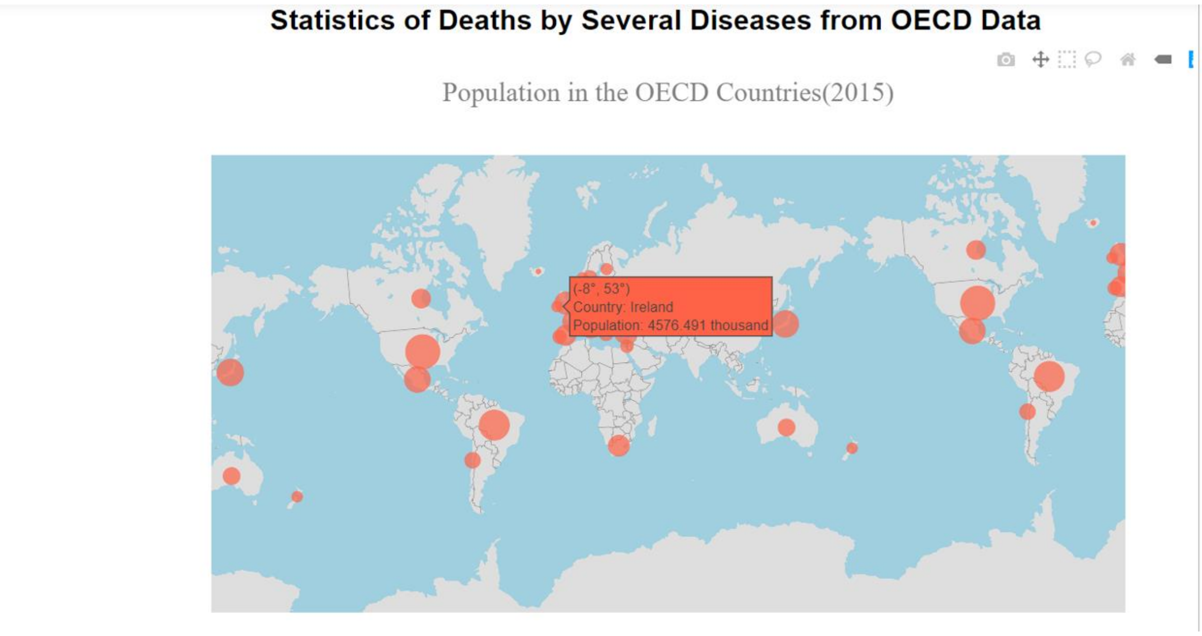
• The last graph is the most interactive graph of the project, handled by a selecting pane and slider. Selecting pane can be used for the selection of the multiple or single cause of death and slider can be used to change the years. The x-axis of this graph contains the countries and Yaxis contains the percentage out of total deaths because of the concern for comparing multiple countries on the same graph and avoiding discrete data. In this bubble chart, bubble size represents the number of deaths. In this project, it is endeavoured to create graphs as more interactive like Tableau dashboard but that's a disadvantage with Plotly, it does not support graph callback while using with widgets in the Jupyter notebook. This can be achieved by Javascript but this project is more on iPython. Moreover, the basic principles of the data visualization had been attempted to achieve like the choice of graphs as best of knowledge and suitability, use of colours suitable to the audience, usage appropriate graph titles and labels on axis plus the important data in the tooltips and HTML headings had been used in middle for describing the flow. Last step is reporting which can be done by exporting the graph state from the interactive pan available on top-right of the graphs or taking screenshot from computer. As a critique, more features for the reporting may have been included like exporting into PDF or HTML. There are certain challenges that had been faced, as being new to Plotly it is hard to do the setup for the Jupyter notebook took time but achieved by finding different solutions on the web. While plotting from multiple files which need calculation like finding the percentage needed a good understanding of Pandas and Numpy. Next, the bubble size and the tooltip needed extra efforts but had achieved by adding new columns in the data frame with the respective size and tooltip. Moreover, while creating the graph like stacked graph Plotly examples generally uses the multi declaration of the trace of the graph but loop have come up with a good solution in spite of writing complete similar code repetitively. At last, this project helped in bolstering the concepts of the data visualization and a good usage of ipython and Plotly.

Results

This visualization can be used in multiple ways, where it had used for the analysis of the impact of the Alzheimer disease i.e. Disease of the Nervous system in last few years. So, firstly a country can be picked from the OECD countries displayed in the map, picked Ireland of them for the stacked graph. Initially it has shown, the total death was around 0.75% of total population in 1993, later this rate dropped to 0.6% in the year of 2004 remained almost constant next four year and after that it started growing and in the year of 2013 it again came to 0.75%, hence it can be concluded that there is no impact on the total death percentage, and can be expected as population in 2015 was around 4561 thousand(graph 1), average 0.73% percentage of people are dying every year, mostly by diseases of circulatory diseases like heart stroke but taking interest in diseases of nervous system, and clicking on nervous diseases. It was found that the rate of death been totally doubled in years as shown in attached figure 1. Later, this can be compared with other countries in the third graph. Choosing the cause of death as Diseases of the nervous system, for reference this can be summed up with other diseases to compare the impacts.

Figure 1

In the figure 1, we can explore the country population by clicking on them. Here in below screenshot we have explored the Ireland.



Statistics of Deaths by Several Diseases from OECD Data

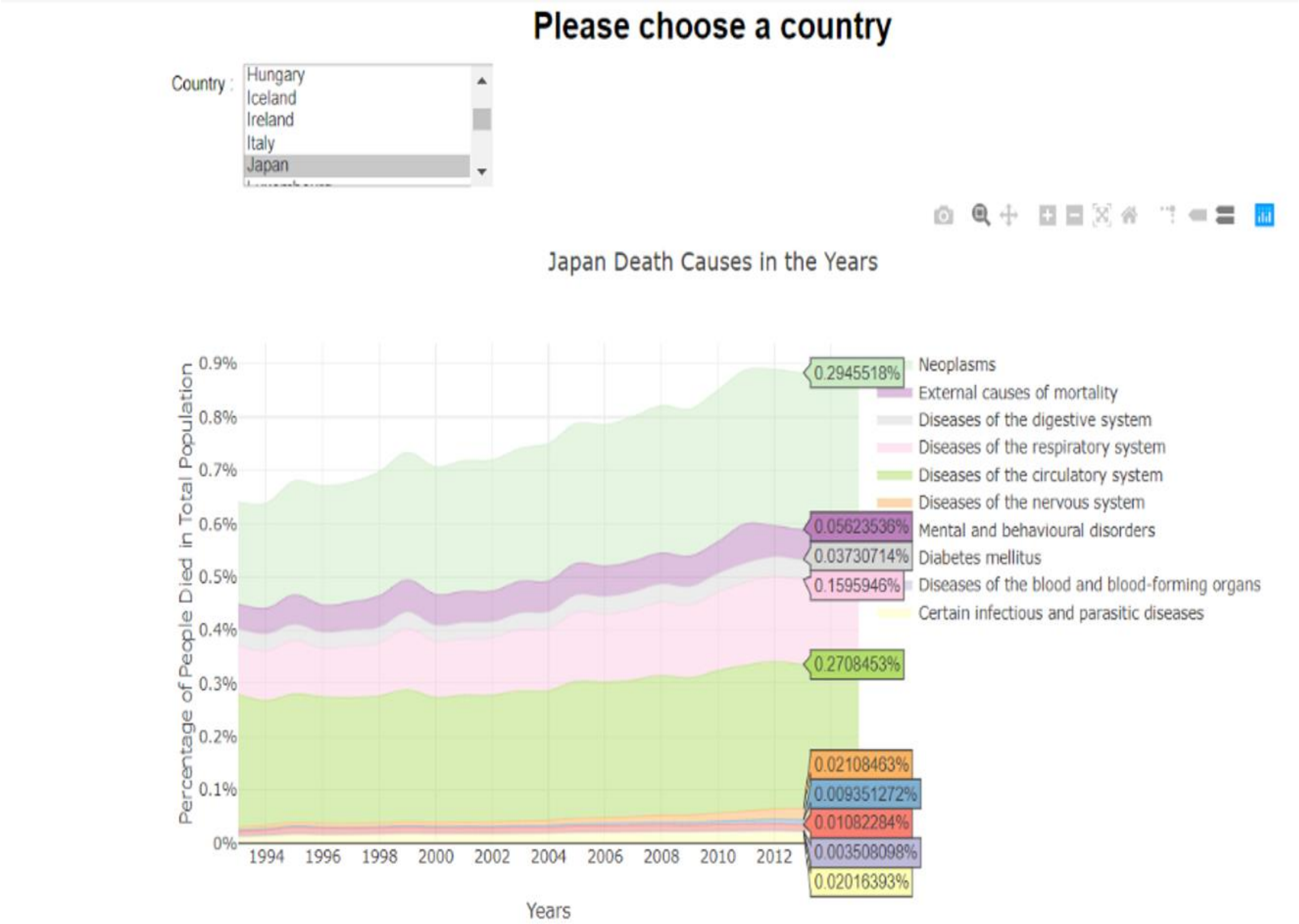
Population in the OECD Countries(2015)

(-8°, 53°)
Country: Ireland
Population: 4576.491 thousand

In second visualisation we can select the country and type of disease as well to explore our visualisation.in below screenshot we can see that in Ireland, how much percentage of population was affected from which type of disease. We can visualise them together as well and separately as well on each country and each type of disease.

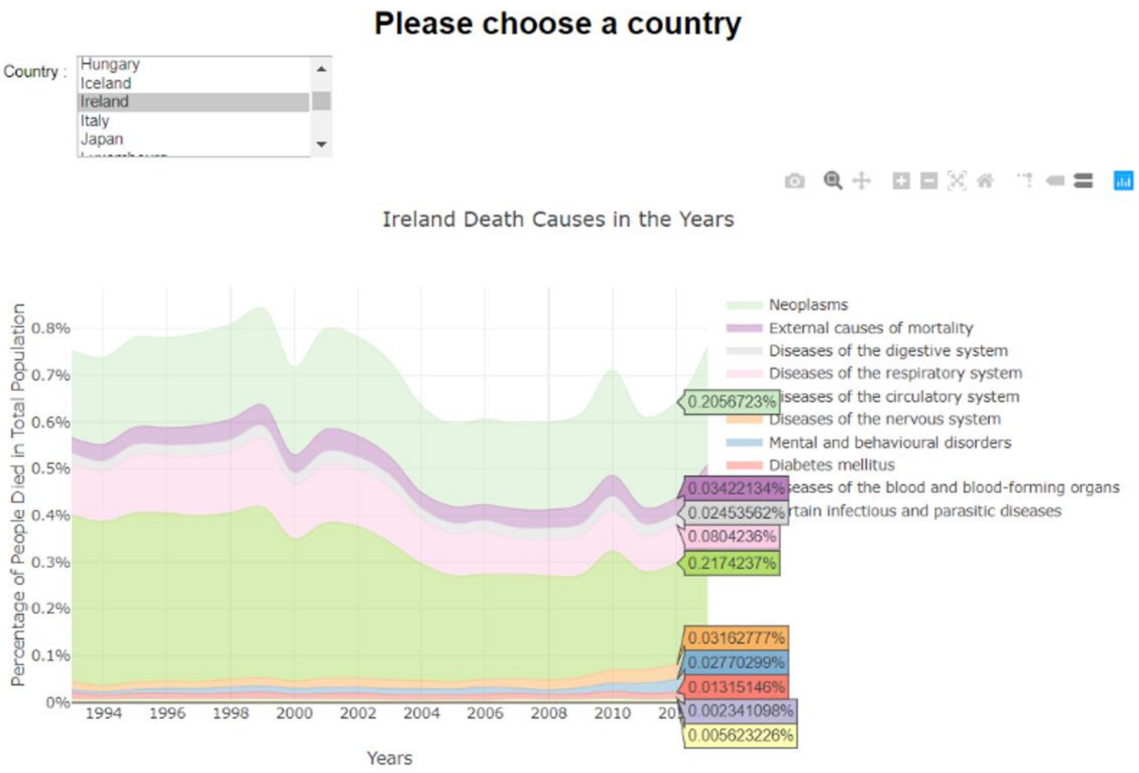
VISUALISATION ON JAPAN

Figure 2


Please choose a country
Country :
Hungary
Iceland
Ireland
Italy
Japan
Japan Death Causes in the Years
0.9%
0.8%
0.7%
0.6%
0.5%
0.4%
0.3%
0.2%
0.1%
0%
Percentage of People Died in Total Population
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
Years
0.2945518%
0.05623536%
0.03730714%
0.1595946%
0.2708453%
0.02108463%
0.009351272%
0.01082284%
0.003508098%
0.02016393%
Neoplasms
External causes of mortality
Diseases of the digestive system
Diseases of the respiratory system
Diseases of the circulatory system
Diseases of the nervous system
Mental and behavioural disorders
Diabetes mellitus
Diseases of the blood and blood-forming organs
Certain infectious and parasitic diseases

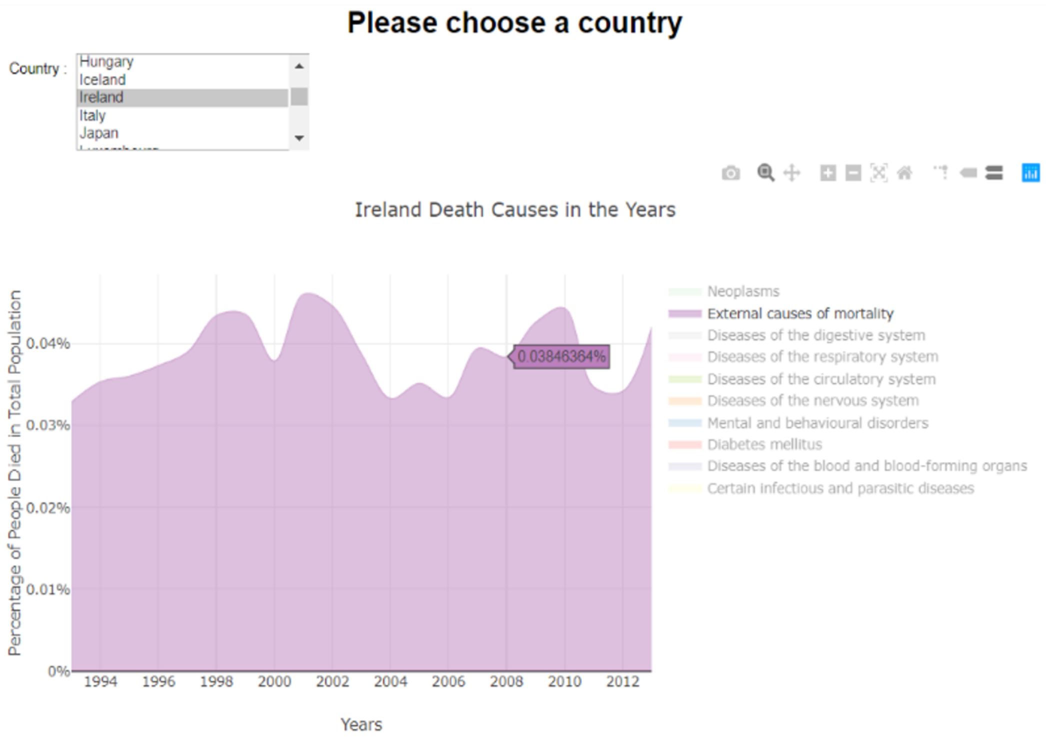
VISUALISATION ON IRELAND

Figure 3


Please choose a country
Country :
Hungary
Iceland
Ireland
Italy
Japan
Ireland Death Causes in the Years
Percentage of People Died in Total Population
0.8%
0.7%
0.6%
0.5%
0.4%
0.3%
0.2%
0.1%
0%
1994
1996
1998
2000
2002
2004
2006
2008
2010
20
Years
Neoplasms
External causes of mortality
Diseases of the digestive system
Diseases of the respiratory system
Diseases of the circulatory system
Diseases of the nervous system
Mental and behavioural disorders
Diabetes mellitus
Diseases of the blood and blood-forming organs
Certain infectious and parasitic diseases
0.2056723%
0.03422134%
0.02453562%
0.0804236%
0.2174237%
0.03162777%
0.02770299%
0.01315146%
0.002341098%
0.005623226%

In below figure we have visualised the death caused by external cause of morality in Ireland, Which resulted in 0.038% in 2008.

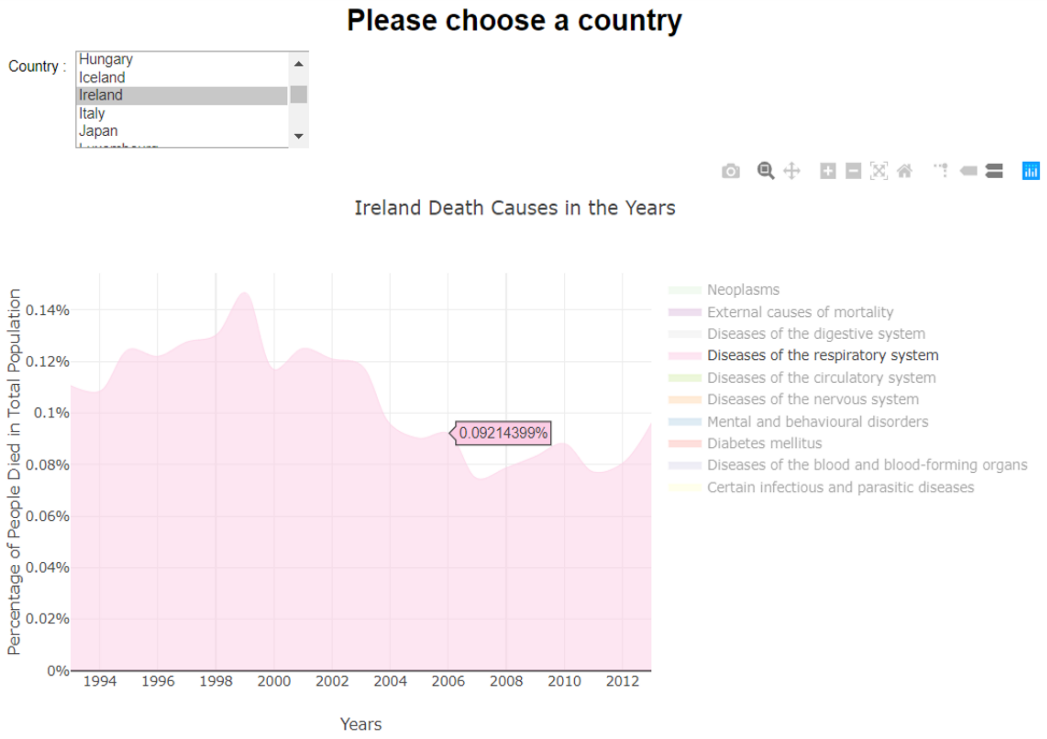
Figure 4


Please choose a country
Country :
Hungary
Iceland
Ireland
Italy
Japan
Ireland Death Causes in the Years
Percentage of People Died in Total Population
0.04%
0.03%
0.02%
0.01%
0%
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
Years
0.03846364%
Neoplasms
External causes of mortality
Diseases of the digestive system
Diseases of the respiratory system
Diseases of the circulatory system
Diseases of the nervous system
Mental and behavioural disorders
Diabetes mellitus
Diseases of the blood and blood-forming organs
Certain infectious and parasitic diseases

In below figure we have visualised the death caused by diseases of respiratory system in Ireland,

Which resulted in 0.092% in 2006.

Figure 5



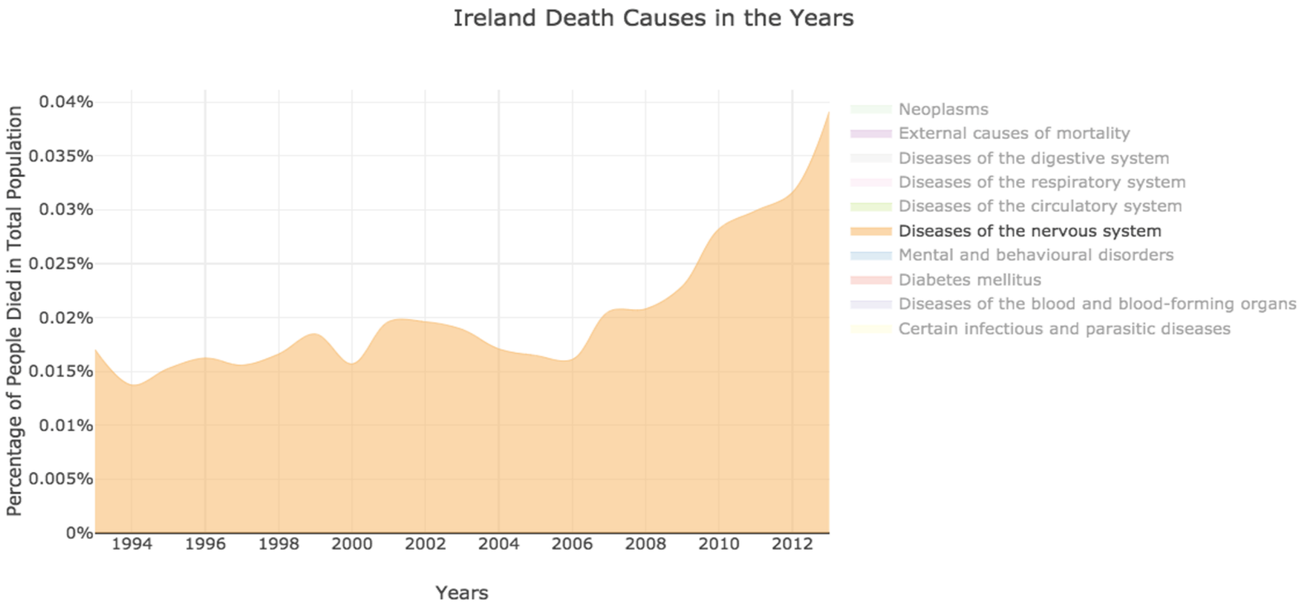
Please choose a country

Country : Hungary, Iceland, Ireland, Italy, Japan

Ireland Death Causes in the Years
Percentage of People Died in Total Population
0.14%
0.12%
0.1%
0.08%
0.06%
0.04%
0.02%
0%
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
Years
0.09214399%
Neoplasms
External causes of mortality
Diseases of the digestive system
Diseases of the respiratory system
Diseases of the circulatory system
Diseases of the nervous system
Mental and behavioural disorders
Diabetes mellitus
Diseases of the blood and blood-forming organs
Certain infectious and parasitic diseases

In below figure we have visualised the death caused by diseases of nervous system in Ireland.

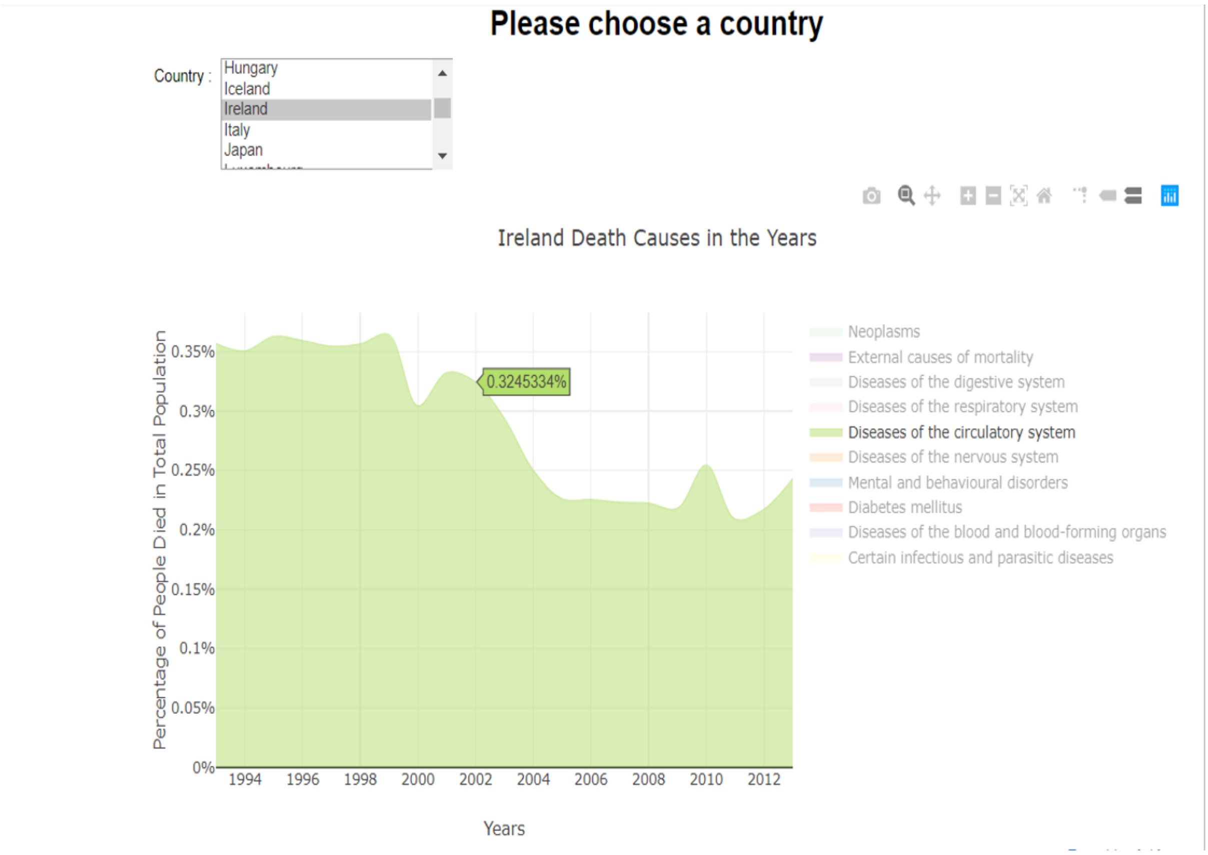
Figure 6



In below figure we have visualised the death caused by diseases of Circulatory system in Ireland,

Which resulted in 0.32% in 2002.

Figure 7

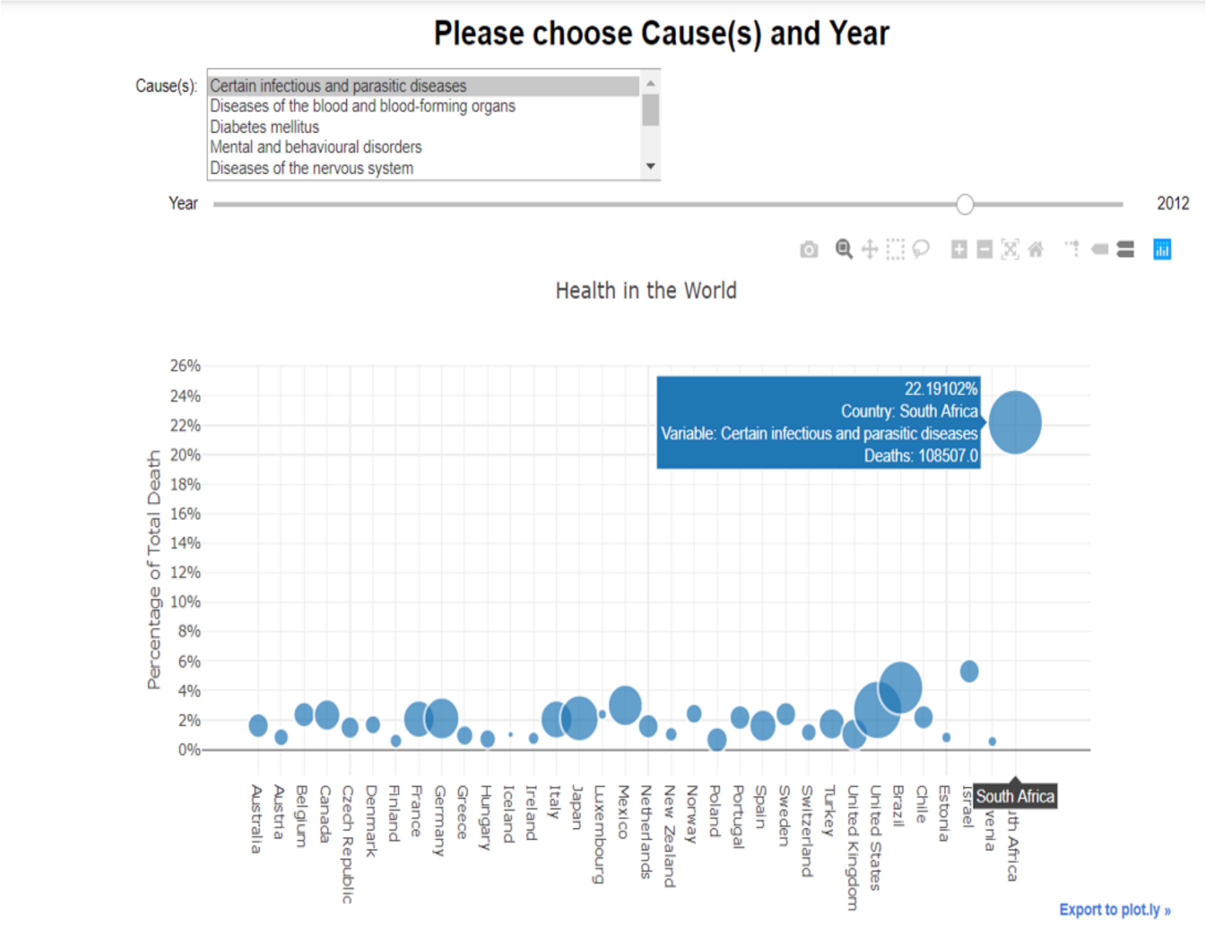

Please choose a country
Country :
Hungary
Iceland
Ireland
Italy
Japan
Ireland Death Causes in the Years
Percentage of People Died in Total Population
0.35%
0.3%
0.25%
0.2%
0.15%
0.1%
0.05%
0%
0.3245334%
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
Years
Neoplasms
External causes of mortality
Diseases of the digestive system
Diseases of the respiratory system
Diseases of the circulatory system
Diseases of the nervous system
Mental and behavioural disorders
Diabetes mellitus
Diseases of the blood and blood-forming organs
Certain infectious and parasitic diseases

Very impressive results were found by comparing with the Diseases of respiratory system. Initially, most of the countries were having 2% people dying by like disease, with the highest rate of 2.82% of total deaths in Belgium and 2.15 % in Ireland with a number of 693 deaths while UK was worst with respiratory diseases with 15.51% in 1993 having 102182 deaths of the total population. Later, sliding to 2003 the average death percentage for the nervous diseases shifted to 3 % while in Ireland it moved to 2.44% with 710 deaths, while this time Ireland was in top for respiratory diseases that 15.31% of total deaths. But when slid to 2013 shown in figure 2, the rate for the respiratory diseases came back to 11.87% while nervous diseases changed to 4.67% with 1378 deaths which is almost double of the initial rate.

Further, we can explore by visualising each type of death cause in every country together by selecting year consecutively. Which explains, which type of death was the reason for the highest number of deaths in particular country.

In below visualisation we can observe that highest number deaths(22.19%) was recorded in the South Africa due to certain infectious and parasitic diseases in year 2012.

Figure 8



Please choose Cause(s) and Year

Cause(s): Certain infectious and parasitic diseases
Diseases of the blood and blood-forming organs
Diabetes mellitus
Mental and behavioural disorders
Diseases of the nervous system

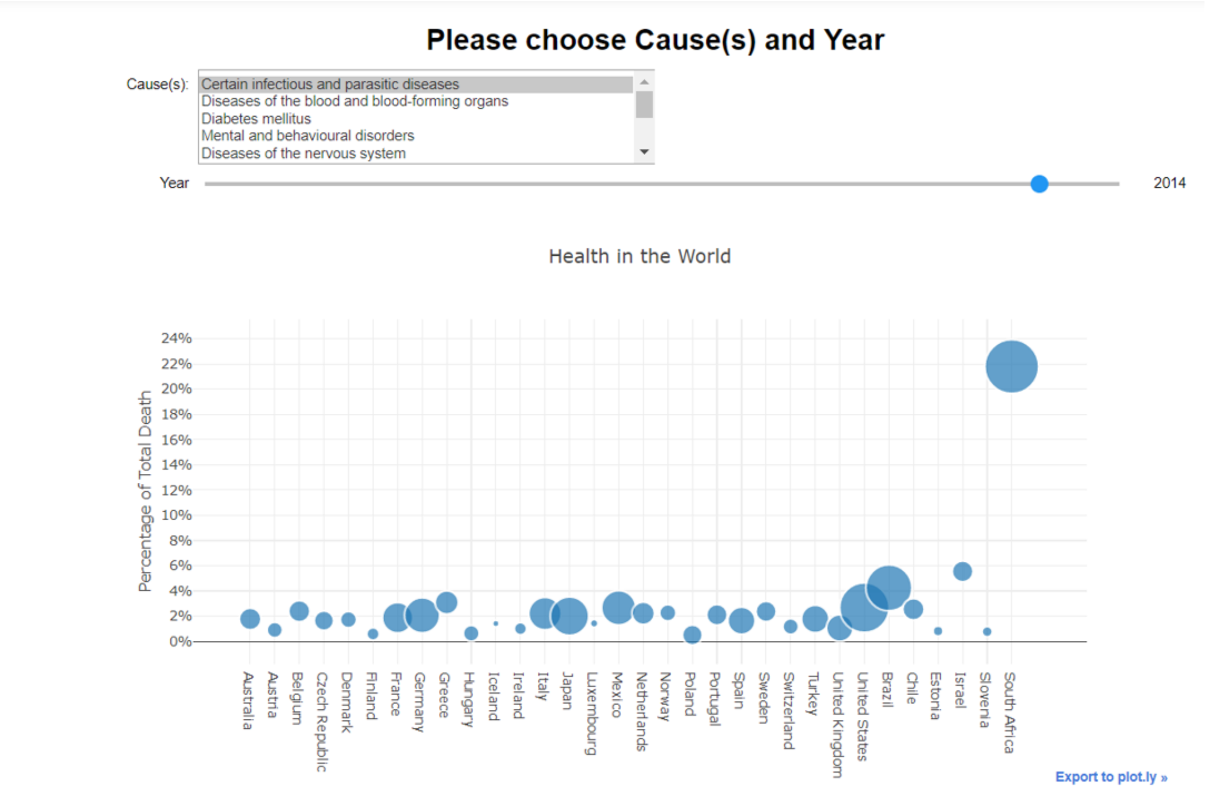
Year 2012

Health in the World

22.19102%
Country: South Africa
Variable: Certain infectious and parasitic diseases
Deaths: 108507.0

Same we have tried to observe in year 2014 as well.But the highest number deaths was recorded in the South Africa only due to certain infectious and parasitic diseases.

Figure 9



Please choose Cause(s) and Year

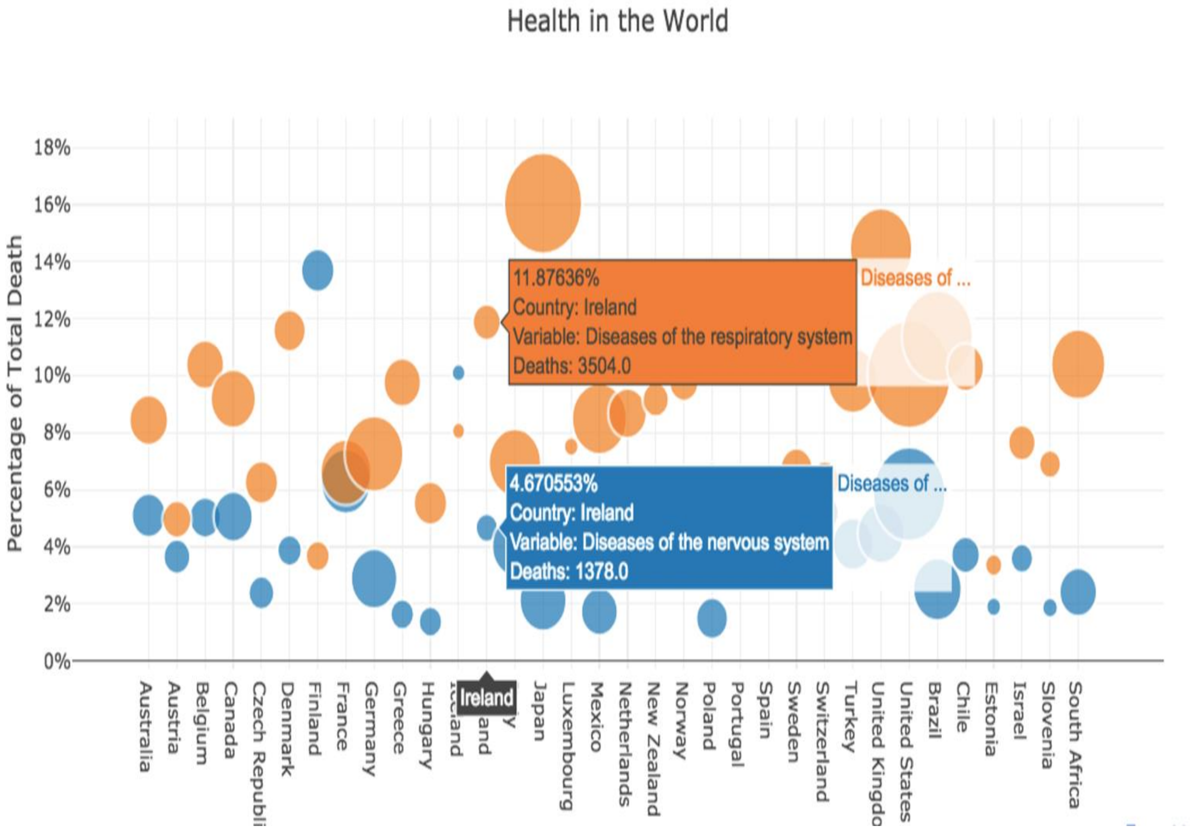
Cause(s): Certain infectious and parasitic diseases
Diseases of the blood and blood-forming organs
Diabetes mellitus
Mental and behavioural disorders
Diseases of the nervous system

Year 2014

Health in the World

Moreover, it can be seen from this graph, that risk of the diseases of the nervous have been arising as for a high income countries like Finland, Australia and United states nervous diseases have a big impact as in Finland out total person die 15.87% are of nervous diseases while 190638 people died in United states in 2015 and 8469 in Australia which is really disturbing. Hence it can be concluded that the nervous diseases have a very big impact not only on Ireland but on the other part of world also. This visualization can be used for the study of the impact of other diseases also in the same manner. Further, analysis for the reason for the growth of these disease can be achieved by creating visualization for the type of labour changed or the change in BMI of people in this span and many more.

Figure 10


Health in the World
Percentage of Total Death
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
11.87636%
Country: Ireland
Variable: Diseases of the respiratory system
Deaths: 3504.0
Diseases of ...
4.670553%
Country: Ireland
Variable: Diseases of the nervous system
Deaths: 1378.0
Diseases of ...
Australia
Austria
Belgium
Canada
Czech Republi
Denmark
Finland
France
Germany
Greece
Hungary
Ireland
Japan
Luxembourg
Mexico
Netherlands
New Zealand
Norway
Poland
Portugal
Spain
Sweden
Switzerland
Turkey
United Kingdo
United States
Brazil
Chile
Estonia
Israel
Slovenia
South Africa

Links

1. https://github.com/yogigupta5292/CA2-DATA-VISUALISATION-

2.OECDDataSource https://stats.oecd.org/index.aspx?r=895785&erroCode=403&lastaction=login_submit#

3. Socrata Open Data(Location point file) https://opendata.socrata.com/dataset/Country-List-ISO-3166-Codes-Latitude-Longitude/mnkm8ram