CSCI 5388
Big Data Analytics

Project Report

Good Health and Well Being.

Instructor:

Dr. Yalong Wu

Team	**Major**
Suhail Arfaath	(Data Science)
Sri Hari Krishna Perla	(Data Science)
Bhanusree Penukonda	(Data Science)
Vishal Kumar	(Data Science)

Table of Contents:

Abstract:

This report describes the details about the Preprocessing, Transformation, Exploratory Data Analysis, Design, and Implementation of a dataset from the official website of the World Health Organization (WHO) - WHO Mortality Database.

Keywords: Preprocessing, Transformation, EDA, Design and Implementation, Mortality.

Introduction

Project Introduction

Good Health and Well-Being encompass physical health, psychological and emotional stability, and social engagement. Physical wellness involves self-care and a temperate lifestyle. Emotional well-being is psychological happiness that includes good emotions and subjective experiences.

The World Health Organization (WHO) describes health as 'a state of complete physical, mental and social well-being and not simply the absence of sickness or infirmity'. Good health is not only of value to the individual as a major factor of quality of life, well-being, and social participation, it also subsidizes overall social and economic growth. Well-being is the feeling of satisfaction with life, which is defined by health, happiness, and prosperity. The maintenance of the human body and all efforts to keep it free from disease and intoxication while facilitating access to medical care are all part of good health.

Besides the overall accessibility of health care, health can be determined by individual characteristics and behavior, such as smoking, excessive alcohol consumption and unhealthy diets, and by external socio-economic and environmental factors, such as lifestyles, living conditions, air quality and various other factors. These behavioral and external factors are to be addressed by preventive measures. One of the major parts of Sustainable Development Goals (SDG) is Good Health and Well Being. Research is also essential to ensuring good health as well as avoiding and attempting diseases. Thus, the ability to achieve the Sustainable Development Goal (SDG) targets on good health and well-being is strongly linked to other areas related to sustainable development. Ensuring that people live long, and healthy lives also means reducing the causes of premature deaths, such as unhealthy lifestyles or accidents, improving external health determinants and ensuring access to health care for all.

Usage of This Report

The main tenacity of this report is to elaborate the details about our project and make it a very good tutorial for beginners to understand what exactly the meaning of Good Health and Well-Being is. Also explaining the mortality causes through the data recorded by WHO over the years (1955 - 2022) across the globe.

Overview or Background

All nations' progress depends on good health and health equity. In the next WHO strategy for Europe, Health 2020, which the Regional Office is creating in collaboration with the 53 Member States in the European Region, these objectives have been identified as major objectives. This is the justification for doing so.

In Health 2020, addressing socioeconomic determinants of health and lowering associated health disparities are top priorities. For this reason, I warmly appreciate the fifth worldwide HBSC report's focus on social determinants of health. HBSC understands that genes and microorganisms alone cannot account for ill health. It covers the environment in which young people reside, their access to healthcare, educational opportunities, and recreational activities, as well as their residences, neighborhoods, towns, and cities. Additionally, it depicts socioeconomic class, gender, age, ethnicity, morals, and discrimination as well as other personal and cultural traits. In summary, social factors have a significant impact on both individual and population health.

Ever wondered why we are born and why we die. What is the factor which affects our mortality?

Is it diseases or is it the aging which reduces our mortality?

In the evolutionary journey of human beings from Ramapithecus > Australopithecus > Homo Erectus > Homo Sapiens > Modern Humans, there isn’t any way to gauge what is the major reason for death in humans.

Scientist have been using statistical analysis to estimate the cause of death in the past. Which is a way to get observations and draw conclusions. Now in the modern era after the advent of Bigdata Analytics around 2010, numerous ways have evolved to efficiently collect and store data and perform complex computation. It will provide the socio determinants of Health of well beingness of world.

Problem Statement

Human Mortality data plays a huge role in finding challenging insights within the data and extracting patterns. These patterns could help us in understanding human evolution and also analyze the dataset by using Bigdata technologies. This dataset signifies the mortality rate of different age groups. This data set has data right from 1950 and it's a collection of data from past decades that contains the mortality rate of different people from different countries and from different age groups.

The age groups are bucketed based on different ages, age 1 to age 5 fall in bucket 1 or group in a bin. This data has been collected from the National Registry from each country.

What we are trying to solve

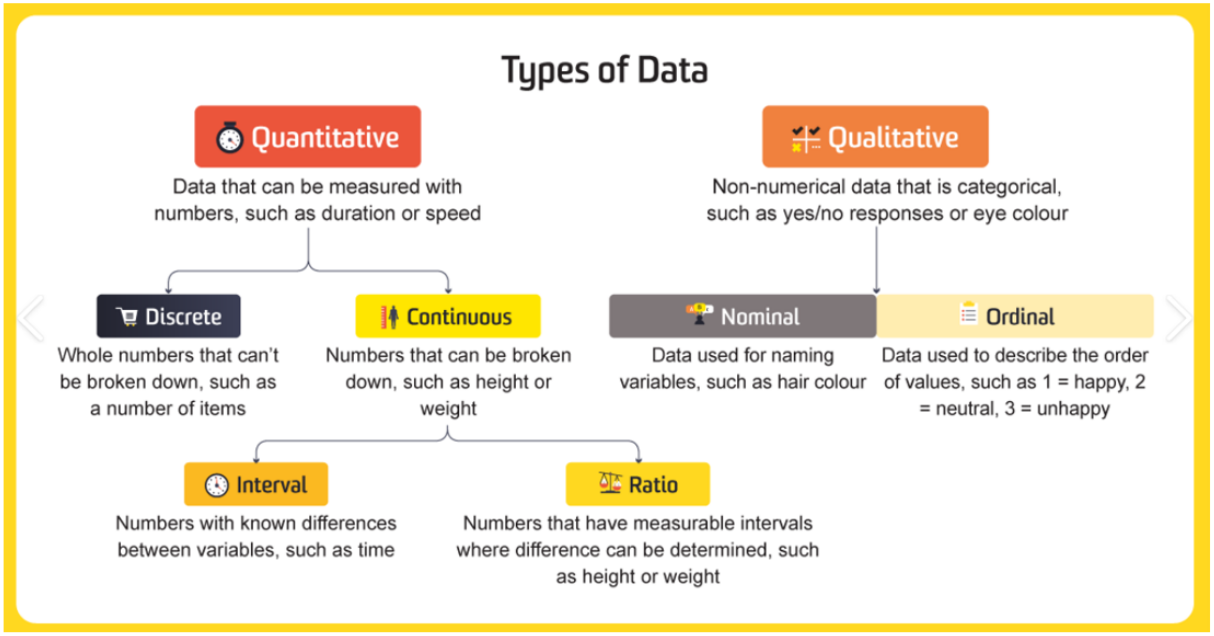
We are trying to solve this problem by identifying the patterns within the data and trying to extract the pattern and cause of problems. The study is important because it provides a snapshot of current health problems. We can also find out persistent problems of risk in certain communities and show specific causes of death overtime. The study can lead to gaining the attention of the respective authorities that can address the issues.

Why we are solving this

This may give us valuable insights through which we can overcome the problem and better help humanity. The importance of mortalities statistics derives both from the significance of death in an individual’s life as well as their potential to improve the public’s health when used so systematically assess and monitor the health status of a specific community. Most of the countries throughout the world that have followed Coding Data collection and data processing set forth buy WHO have seen major impact in addressing the health issues. Therefore, studying the mortality rate Will help not only in identifying the issues but also addressing them thereby minimizing many issues.

Data Gathering

Types of Data



The data available for download from this web site are official national statistics in the sense that they have been transmitted to the World Health Organization by the competent authorities of the countries concerned.

The WHO Mortality Database comprises deaths registered in national vital registration systems, with underlying cause of death as coded by the relevant national authority. Underlying cause of death is defined as “the disease or injury which initiated the train of morbid events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury” in accordance with the rules of the International Classification of Diseases.

The database contains number of deaths by country, year, sex, age group and cause of death as far back from 1950. Data are included only for countries reporting data properly coded according to the International Classification of Diseases (ICD).

Meta Data

Discrete	Continuous	Nominal	Ordinal
Country	Death1 ~ 26	Sex	Admin1
Cause	IM_Deaths1 ~ 4		SubDiv
	Frmat		Year
	IM_Frmat		List

Data Wrangling

We're going to cut the data from different sources and then try to preprocess the data and then try to visualize the data and find the patterns within it. In acquiring the data, we are relying on the data published by the official website of WHO. We have obtained multiple data sets that are related to the subject. The data and its entirety have been merged into one single data set. We then performed Data processing with the aim of exploring the data. Once the exploratory process has been completed, the output is used to visualize and identify patterns. These patterns help us understand the issue better for us to work on the same.

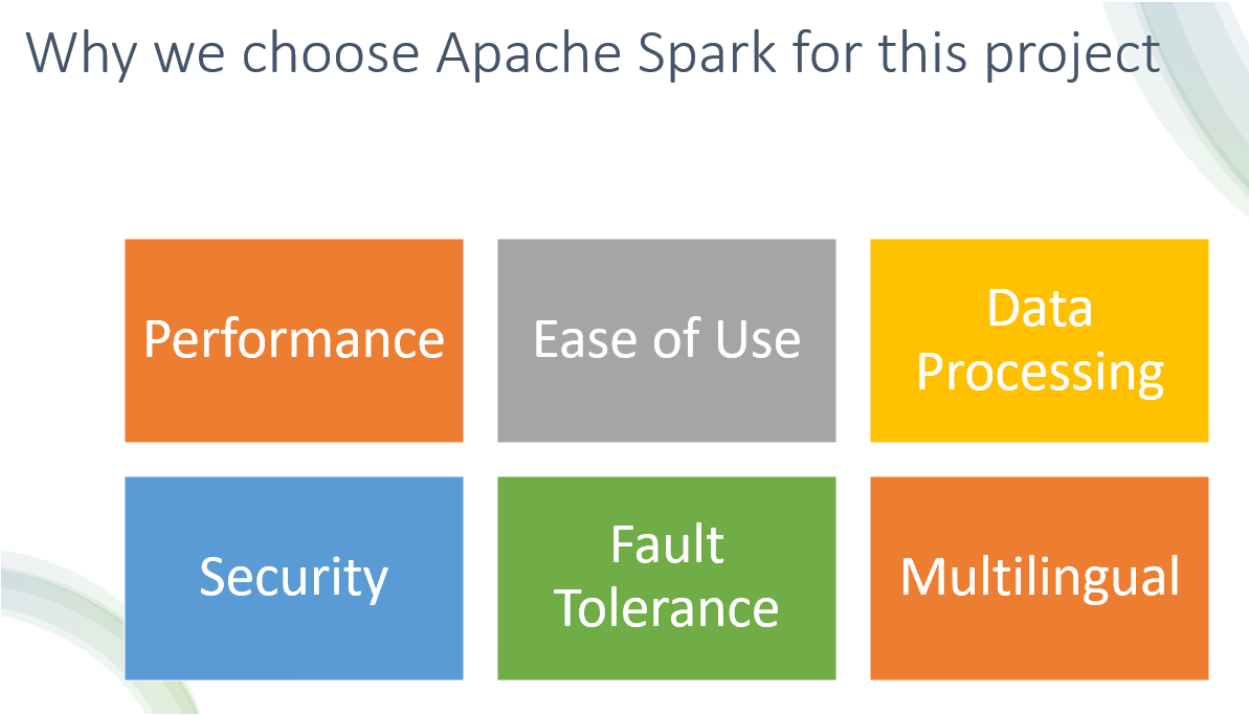
As part of data gathering, we have obtained the data from WHO organizations' website. This data set is named "Human Mortality Data" and it contains 36 columns with more than 6 million rows. As this is a huge dataset which cannot be handled by conventional tools such as MS Excel or pandas and many other similar libraries, we need to rely completely on Bigdata frameworks that can process the information using techniques like min hashing. The primary purpose of big data architecture is to handle injection, processing and analysis of data that is too large are complex for the traditional database systems. As our dataset is about 650 MB, we did not have to work with batch processing. We had to load the entire dataset at once into spark to process the data.

Techniques Performed

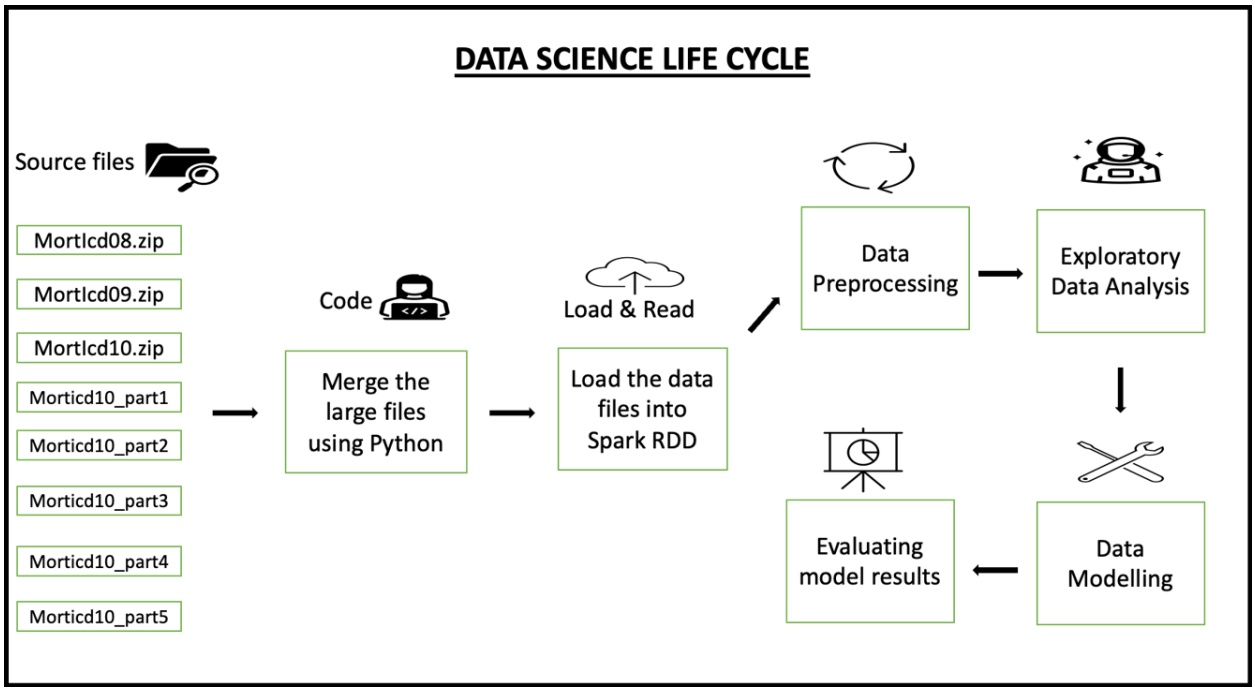
- Data Preprocessing steps
- Remove any null values
- Remove any outliers
- Remove records which have few attributes missing
- Streamlining all the data and rounding to absolute value, such that all records follow the same format
- Set a date and time format on columns, such that all records follow the same format
- Sorting the data

Data Transformation

Architecture:



The big data architecture is also used to transform unstructured data for analysis and reporting. The data is captured, processed and analyzed in real time or with low latency. Some of the benefits of big data architecture are it allows parallelism. We can therefore benefit from a performance boost. One of the major challenges in big data solutions is ensuring security as data is interested in consumed by multiple platforms and applications. However, the major advantage is that it enables interactive exploration of the data.



Spark Implementation:

We initiate with loading the data into SPARK. The following are the actions performed in SPARK:

1. Performed descriptive statistics on the data with the aim to preprocess the data.
2. We imported the required packages into SPARK which are necessary for the preprocessing of the data.
3. Once the packages are imported, we initiate the SPARK session.
4. Later, we are required to code in order to read the data.
5. When we identify the null values, we include a step to deal with it. We can either drop the null values or give them a value. In our project, we have deleted the null values. To accomplish this task, we use "**fillna**" function.
6. We had to drop the rows and columns with null values using "**na.drop".** Some of the columns were blank in the dataset and therefore, we replaced the blank sections with 0 as they were numeric columns.
7. We also performed statistical analysis to explore insights within the data.
8. The output given by spark was converted to pandas and then visualized.
9. Using spark's inbuilt ML library, we have implemented certain algorithms to gain deeper insights of the data.

Research Goal and Objectives

We can try to extract the knowledge from this data in terms of what kind of diseases or pandemics or epidemics occur and how the mortality of populations from different countries.
This includes the cause of death for each age group. With this data we can try to analyze what are the major trends in terms of panic or epidemics and what's the reason why there is a high death toll or mortality for the population.

- Firstly, the main objective of the data is to find the mortality rate for health and wellbeing.
- WHO have provided the major factors affecting the mortality rate.
- Each factor percentage and influence for the occurrence of death.
- In details study of few health conditions affecting the cause of death.
- Age factor and the mortality rate as per death.
- Any possibilities or suggestions for the improvement of the above factors.

What knowledge patterns we are observing

In terms of knowledge, if we try to understand the main reasons and the kind of symptoms or diseases occurring in different countries. We learn the risk factors of diseases and compare contrast health events between different populations. When we study the patterns, the relevant authorities comprehend the information such as the diseases that are majorly impacting the mortality rate. This will enable the authorities to invest in the health sector accordingly.
Main Patterns to find out are:

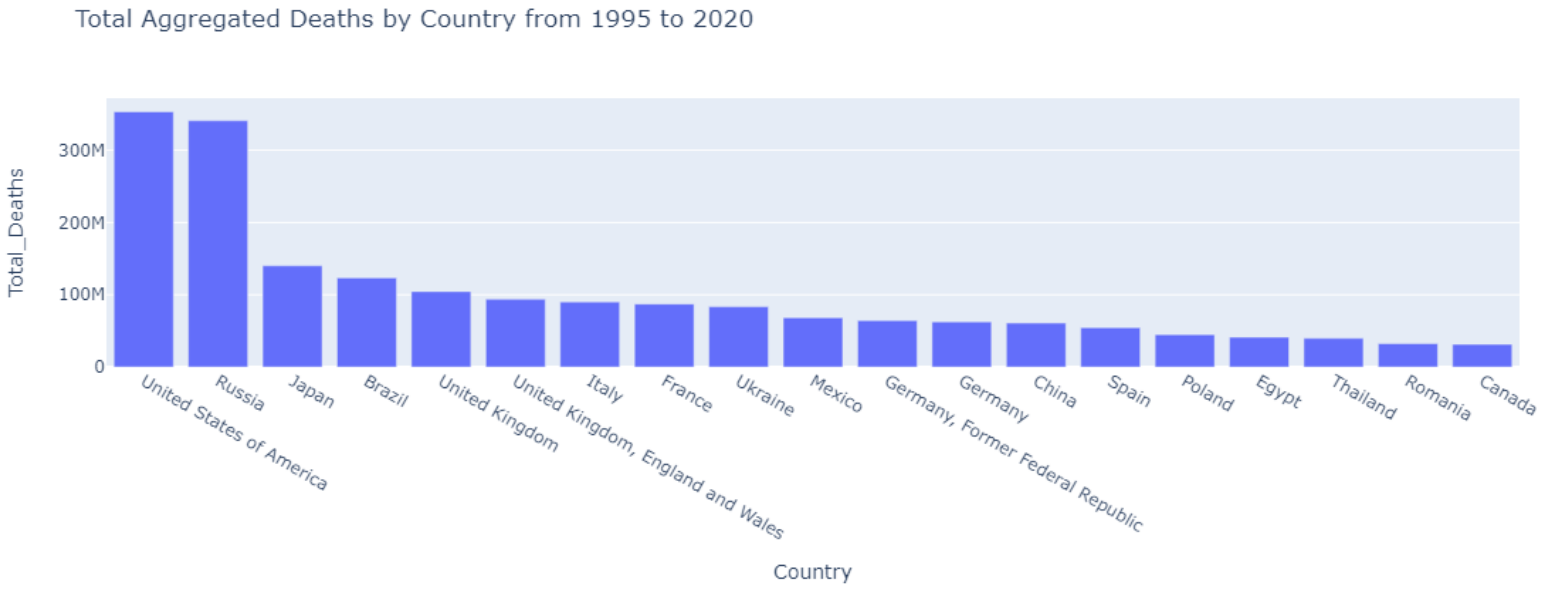
- The top 10 deaths by country.
- Infant deaths vs Overall Deaths.
- Top 10 diseases reasons died in USA.
- How Obesity has changed over the decades.
- Accidental Deaths.

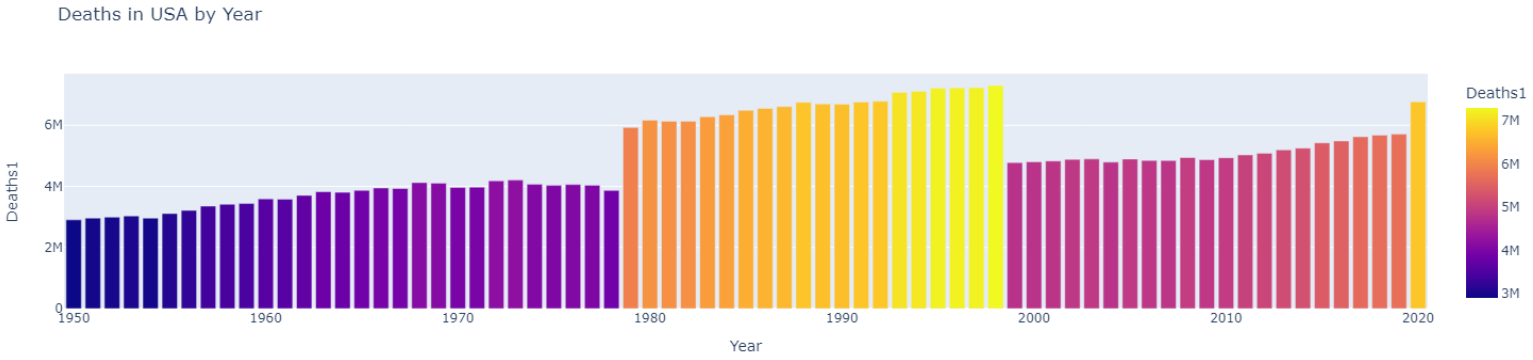
What insights do we gain?

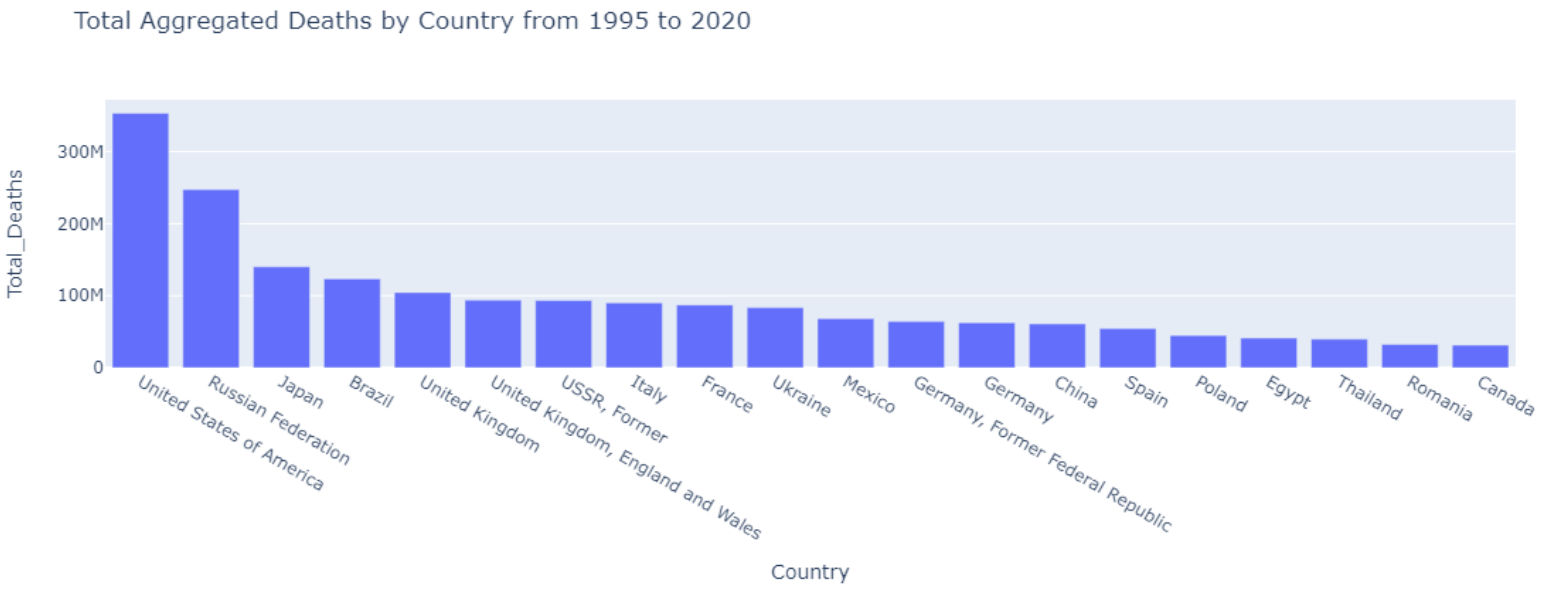
With these insights we can try to better help each country in terms of specific disease aid such as in the case of Africa, if we see we have malaria disease which one is the most prominent, and we have high death toll because of the same.

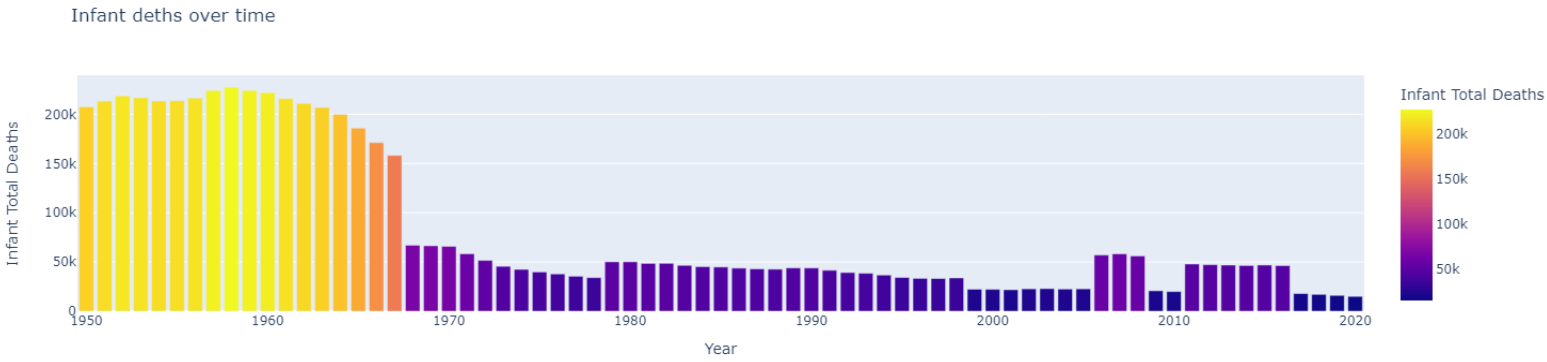
- The possible insights are what are the reasons for high death rate.
- How lifestyle is impacting the health and well-being of the population in different countries.
- The lifestyle changes and that impacts on the mortality rate.
- What are the fatal diseases and how to control the deaths?

Visualizations:

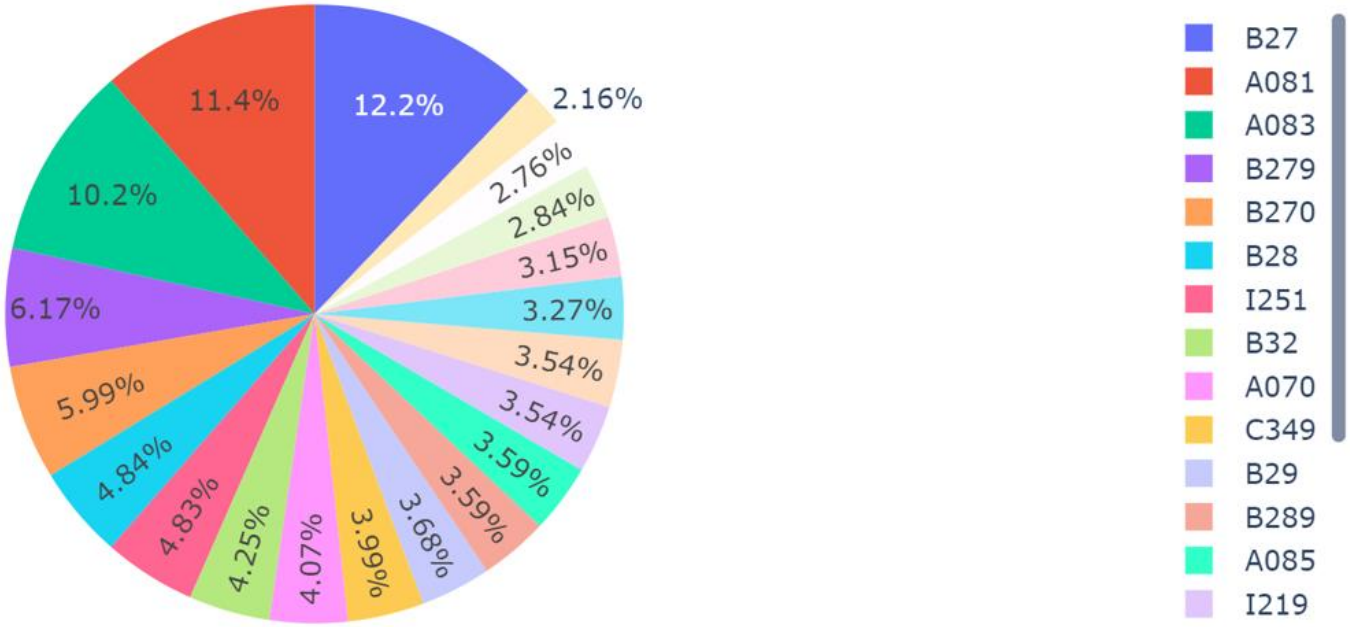





Total Aggregated Deaths by Country from 1995 to 2020
Total_Deaths
300M
200M
100M
0
United States of America
Russian Federation
Japan
Brazil
United Kingdom
United Kingdom, England and Wales
USSR, Former
Italy
France
Ukraine
Mexico
Germany, Former Federal Republic
Germany
China
Spain
Poland
Egypt
Thailand
Romania
Canada
Country


Infant deths over time
Infant Total Deaths
200k
150k
100k
50k
0
1950
1960
1970
1980
1990
2000
2010
2020
Year
Infant Total Deaths
200k
150k
100k
50k

Causes of Deaths



What do we give to society back?

If we were able to address this problem, we could solve the human crisis between different countries in “Third World Countries”. Although the study cannot Address and solve all the issues completely, it can, when carried out properly benefit this society big time. Especially, populations of the third world countries will have their cry heard. For example, people living in the remotest parts of the country who do not have any connection to the outer world suffer from most common diseases such as malaria and Typhoid. Most of them don't even know that the modern medical industry has strong medicine that can cure the disease sends. However, due to lack of knowledge, these diseases are taking up the lives of those people. Another example is that there are many areas where there is a scarcity of freshwater. Due to lack of fresh drinking water, those populations suffer from many diseases caused by drinking contaminated water. If a study is carried out in those areas, this issue can be identified. As these studies are published, the government can address it immediately. As a result, society benefited from this study.

From the data insight the government and the public will get awareness regarding the health and well-being of their countries.

The WHO Forum series has been developed to increase knowledge and understanding around priority public health conditions from the perspective of social determinants of health, allowing researchers, policymakers and practitioners to convene to analyse data, review policies and interventions and

formulate lessons learnt. Beginning with the results of research, the process compares data, experiences and models from throughout Europe. Specific objectives are to document, analyse and increase knowledge and understanding by:
• translating research on young people's health into policies and action within and beyond the health sector;
• scaling up intersectoral policies and interventions to promote young people's health;
• reducing health inequities among young people;
• involving young people in the design, implementation and evaluation of policies and interventions.

This culminates in the development of a synthesis report and policy statement, capacity-building materials and the integration of outcomes into ongoing support to Member States by WHO and partners. Forum meetings usually coincide with regular WHO ministerial conferences on themed areas to ensure that the findings can have the biggest effect during the policy-making cycle.

Are we making any savings?

In terms of savings, we can try to invent better medicines and efficient vaccines and we can also try to deliver the vaccines and medicines to the target location which may aid in solving the problem. We can also identify which medicines are working effectively and those that are not. When we identify this, the medical manufacturing industries can label those medicines obsolete and work on development of the same.

In Terms of monetary $ value

We could try to better address the problem by increasing the supply chain management of delivering the vaccine or medicine to the location where the disease is more prominent especially in terms of covid-nineteen where a lot of people were unable to find the vaccines or the medications due to which a high death toll was recorded.

What is the trend in data?

The trend in data will be dependent on the death rate at that time and the causes of death in that particular year. It gives the top 10 global causes of death.
And, health disaggregated by sex, annual global deaths and DALY's among women were around 15% lower than for men.

How mortality rate differs based on

The mortality rate is a measurement of how frequently people die within a given population during a certain period. The choice of measuring disease or death depends on whether you want to use morbidity or mortality metrics, which are frequently mathematically equivalent.

- Age difference in death rate
- Gender Difference
- Socio Economic Difference
- Country Difference in Health

Literature Review

Our research journey began with an exploration of literature on conditions for well-being in Health and mortality rate.

Some literature already done on this topic:

Yang Yang

Yang (2008) analyzes the impact of changes in cause of death overtime on age, period, and cohort effects in the United States during the 20th century. Table 8 below provides an overview of her findings by cause of death.

Cause of Death	Age	Period	Cohort
Heart disease	Increases exponentially with age	Modest impact	Large monotonic decline from the earliest to the latest cohort
Stroke	Increases exponentially with age	Modest impact	Large monotonic decline from the earliest to the latest cohort
Lung cancer	Increases rapidly with age from early adulthood to peak near ages 80–85, then levels off	Monotonic increase over time	Increases for cohorts through 1905 and decreases for recent cohorts
Breast cancer	Increases with age, but increases slow around menopause	Modest impact	Steady declines in mortality from breast cancer from the earliest to the latest cohort

Yang discovered that significant declines in mortality that began in the late 1960s persisted far into the late 1990s and were primarily related to cohort effects. Although cohort effects vary depending on the precise cause of death, overall survival rates have significantly increased.

Review of the Literature and Evaluation of Mortality Improvement 58 Rates in the US Population

Her data offer further proof of enduring cohort disparities in mortality rates across all investigated causes of death. The predominance of cohort effects in explaining current trends in mortality decreases is a significant conclusion. Birth cohort effects reflect the processes of uneven cohort accumulation of lifelong exposure to risk variables that include education, diet and nutrition, physical activity, and smoking.

Yang discovered that when birth cohort and age effects are jointly controlled, period effects are typically negligible or low. She discovered a very slight decline in stroke mortality since 1975 and almost no period effect for heart disease mortality. Period effects are probably more pronounced at times of war or other significant events with significant social repercussions (Yang 2008).

Research Methodology Contribution to the field

People are more likely to die young or have poor health outcomes when certain behaviors, exposures, and predispositions are present. The seminars were designed to help participants better understand the risk factors, primarily behavioral risk factors, that are most receptive to preventative and health policy measures. There are various ways to define "early." Although his latter study has centered on fatalities before age 80, Michael McGinnis' earlier work concentrated on deaths before age 75. The Global Burden of Disease report from the World Health Organization compares years of life lost to a reference age of 86, which is the maximum average longevity for a nation with a population of over 5 million people.

The methodology used are

- Maternal mortality ratio (per 100 000 live births)
- Age factor as a mortality rate
- Medical health workers working per capital ratio.
- Infectious, parasitic, and respiratory diseases
- Cancer as a health Issue
- Diseases of the circulatory system
- All other causes of the digestive system.

From the data available finding the factors/above methodologies we can find the mortality rates vary based on the different health conditions.

Tools and Packages used

Data Visualization libs: Plotly for Interactive Graphs
BigData Frameworks: Apache Spark, PySpark

References

- Mortality Committee. 2005. “Projecting Future Mortality: Towards a Proposal for a Stochastic Methodology.” CMI Working Paper 15.
- https://www.cdc.gov/csels/dsepd/ss1978/lesson3/section3.html
- Data source and other information:
- https://www.who.int/data/data-collection-tools/who-mortality-database
- Andreev, Kirill F., and James W. Vaupel. 2005. “Patterns of Mortality Improvement Over Age and Time in Developed Countries: Estimation, Presentation and Implications for Mortality Forecasting.” Paper presented at the Population Association of America Annual Meeting, Philadelphia. http://paa2005.princeton.edu/abstracts/51061.