PYTHON PROJECT: LEADING CAUSES OF DEATH

Why do people die ?

Our team : Mathilde Bernard, Romain Fernex, Maksym Verveha, Wenyuan Wang

CONTENT TABLE

INTRODUCTION

OUR RESEARCH QUESTION

Why do people die? In this report we want to figure out what were and currently are the causes of people's death and how those have been changing since 1990. We want to explore the dynamic of change and understand how reasons of death differ by region and countries.

LITERATURE REVIEW

When describing a country's overall health performance, life expectancy and death rates are often hailed as the most popular indicators. However, as analyzed by many leading health institutions, looking at the causes of death has also proved to be a powerful predictor

of a country's overall health status. This term refers, according to the WHO, to "**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**" [1]. There are 3 main identifiable types of "cause of death" : injuries (road accidents, conflicts…), non-communicable disease (chronic diseases such as diabetes, long-term illnesses etc..) and **communicable diseases** (malaria, HIV…). While the wide majority of death - 73% - currently belong to the category of **non-communicable disease** [2] this proportion depends highly on the level of wealth and development of a given country. According to a WHO report on social determinants of health published in 2014 [3], the leading cause of death in **high income** countries is indeed **chronic diseases** but for middle income-countries the proportion of communicable diseases-related deaths is also significant. This proportion even overtakes that of chronic diseases for countries with **low income** where **infectious diseases** such as malaria and tuberculosis still take an important toll on the overall population. In spite of these discrepancies, there is a global tendency, globally and especially in **OECD countries** [4], of death causes **switching rapidly from infectious diseases to ischemic heart, circulatory and respiratory diseases.** This increase in death due to chronic diseases has been strongest in the Western pacific region [5] where changes in living standards have brought about new forms of health risks. Covid also had a notable impact of leading cause of deaths, notably in the developed world, with it maintaining itself as the 3rd leading cause of death in the US for 3 years in a row since 2019 [6].

Finally, as pointed out by the Global Health Observatory of the WHO, monitoring the number of death and their cause is vital to help adapt countries' health systems to react to these risks [7]. This has ramifications in multiple sectors, such as the food industry in countries where an **increasing death rate linked to diabetes** might indicate that tougher regulations on food composition are necessary.

Therefore studying this data is useful both to analyze current health inequalities across the world and to predict future patterns that may have an influence on how future health systems are designed and incorporated in countries' health-related legislation.

DATASET DESCRIPTION

To answer our questions, we selected a free global dataset from the website Kaggle :

- (https://www.kaggle.com/datasets/iamsouravbanerjee/cause-of-deaths-around-the-world?resource=download)

Our dataset contains records from 204 countries over 30 years and 31 causes of death are studied.

Country/Territory	Code	Year	Meningitis	Alzheimer's Disease and Other Dementias	Parkinson's Disease	Nutritional Deficiencies	Malaria	Drowning	Interpersonal Violence	...	Diabetes Mellitus	Chronic Kidney Disease	Poisonings	Protein-Energy Malnutrition	Road Injuries	Chronic Respiratory Diseases	Cirrhosis and Other Chronic Liver Diseases	Digestive Diseases	Fire, Heat, and Hot Substances	Acute Hepatitis
Somalia	SOM	2008	4154	296	89	8756	1634	756	1817	...	1863	1346	494	8598	2729	2973	2851	4535	507	1110
Cote d'Ivoire	CIV	2018	2367	1137	371	809	27758	609	2323	...	3450	3398	377	758	3404	3638	3958	6661	520	234
Turkmenistan	TKM	1999	146	295	54	23	1	481	311	...	372	477	58	11	446	613	1242	1507	352	100
Solomon Islands	SLB	2005	24	14	14	33	220	101	44	...	277	96	14	33	127	282	101	152	17	13
Lebanon	LBN	1994	50	418	90	13	0	73	279	...	462	650	33	12	370	737	377	586	89	163
Papua New Guinea	PNG	2018	609	387	191	206	945	378	972	...	4505	716	150	202	2185	6954	643	1336	708	91
Kuwait	KWT	2019	12	425	58	1	0	23	42	...	365	322	19	0	583	230	204	338	34	1
Saint Kitts and Nevis	KNA	2015	1	8	3	4	0	3	14	...	26	26	0	3	7	9	10	19	2	0
Italy	ITA	2006	216	34184	5876	469	0	419	666	...	18535	11204	100	325	7154	26286	13429	26019	436	97
Uganda	UGA	2004	6737	1048	261	6462	57092	793	1634	...	3518	2372	548	6326	4804	4639	4181	7092	595	392

Extract from the dataset

Data cleaning process: our data cleaning process was nonexistent, since we downloaded the dataset without any missing values or errors. However, we spent some time studying our dataset to think about the best ways of analyzing it. By grouping by countries, we realized the opportunity for a map visualization and by grouping per year, we realized we could draw insights on the evolution of the world.

OBJECTIVES & RESEARCH DESIGN

Our top-level question is understanding most deadly death causes across the world across years. To do that we start our research with generating tables which conveniently show top causes, and we explore what has changed through every 10 years snapshot. To understand the dynamics of the global situation we also group death causes into similar categories and explore categorical change yearly. Afterwards, we dive deeper into geographical differences given the year 2019 and explore main causes of death by country as well as explore differences between OECD and LCD countries.

## OUR FINDINGS

HISTORICAL ANALYSIS

PROJECT N°1 : TOP DEATH CAUSES IN A GIVEN YEAR

Code description : The code first creates a column with the total number of deaths per row. Then it removes geographic indicators and groups the data based on year with a sum to focus on historical data. Then it ranks and sorts the different causes of death based on their weight for the year that one wants to study. Lastly it retains only the top 10 of these causes of deaths and displays them in an orderly fashion on an horizontal bar chart.
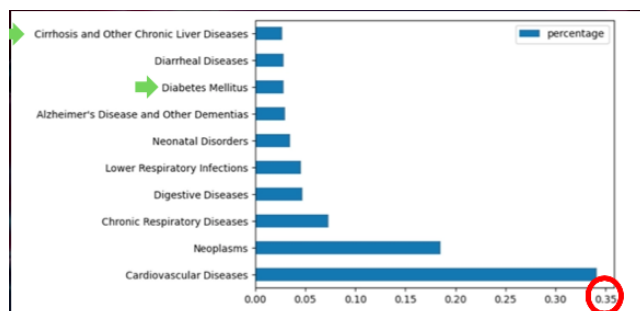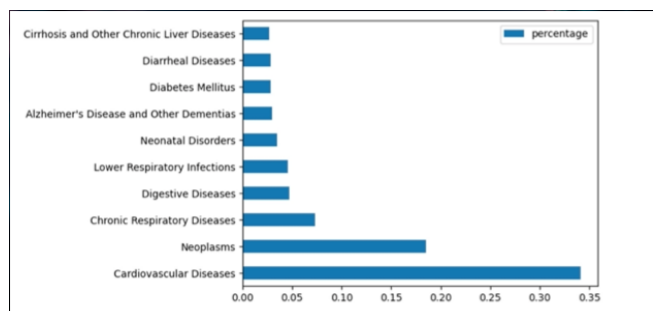
Outcome :
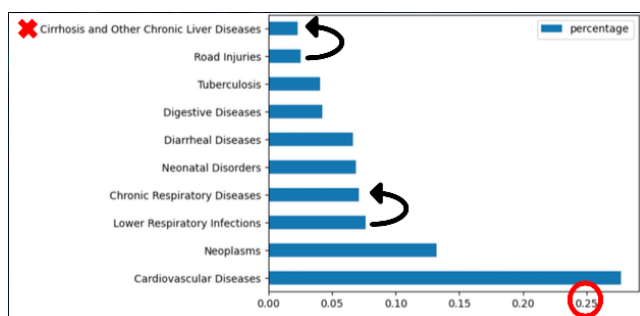
Top 10 causes of death in 1990
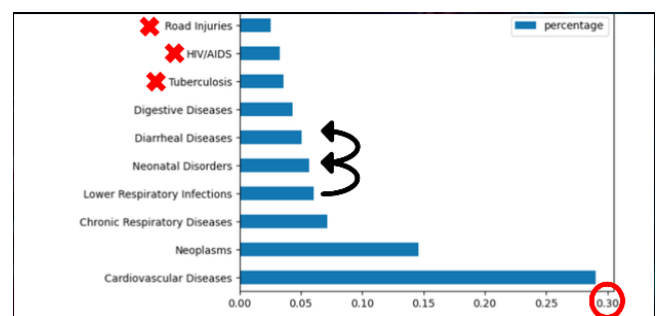
Top 10 causes of death in 2000

Top 10 causes of death in 2010

Top 10 causes of death in 2019

Interpretation : Accidents based cause of deaths such as road injury have slumped since 1990 - most likely due to the implementation of increasing safety practices - and are no longer part of the top 10 causes of death. A similar observation can be made for infectious diseases, notably AIDS and Tuberculosis, which have moved out of the top 10 in the early 2000's with vaccination and awareness efforts in the developing world. Chronic diseases-based deaths, however, rose sharply, even though they were already leading the rankings. Chronic respiratory diseases, neoplasms (cancers) and cardiovascular diseases have cemented themselves as the most likely causes of death by far, causing up to 50% of all deaths worldwide. This is in accordance with data from the OECD and the WHO, which have already identified a significant rise in the proportion of non-communicable diseases related deaths, in relation with higher levels of development across the world.

PROJECT N°2 : DYNAMICS OF GLOBAL CAUSE OF DEATH BY YEAR

Code description : Code groups death causes into 3 groups for easier visual representation using data grouped by year mentioned above and adds a year column to that data. Later the total death count per group is converted into proportion by dividing by total number of deaths. The proportions are moved into a new dataset with only proportions and year as columns. Once the dataset is established the graph is generated by plotting all columns against the year column used as index. The graph is generated using the matplotlib library downloaded before, with help of which visual elements are added to the graph.

Outcomes :
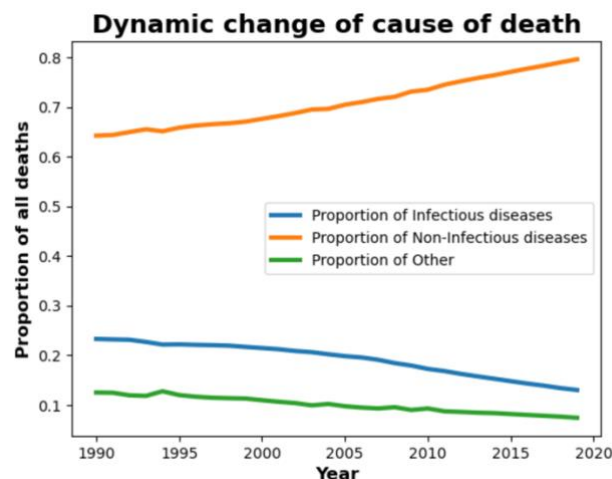
Interpretation : We can clearly see that the proportion of infectious diseases is rapidly decreasing as well as the proportion of other causes of death is decreasing steadily, while the proportion of non-infectious diseases increases significantly reaching almost 80% of all deaths up from 65% 20 years ago. This clearly demonstrates that world is dealing quite efficiently with infectious diseases and their rapid decline is caused hugely by fast development pace of regions which were extremely underdeveloped 20 years ago, however it also means that there is still a big number of death which could be prevented, since most of infectious diseases have cures/vaccines available and that is something we should aim for.

GEOGRAPHIC ANALYSIS

PROJECT N°3 : DEATH BY COUNTRY

Code description: our code starts with downloading some packages specific to geographic visualization. Then, we start formatting the data: we select in one column the main cause of death for each row. Next, we create a dictionary with colors associated with causes of death, and we associate each color with the row. Finally, we select only the rows that belong to the year we want to map. Our next step is to map the data: we create the map, then thanks to an if loop, we change one by one the colors for each country. We repeat that step to obtain 4 maps. Our ultimate step is to draw the legend, to make it broad and specific.

Outcome:
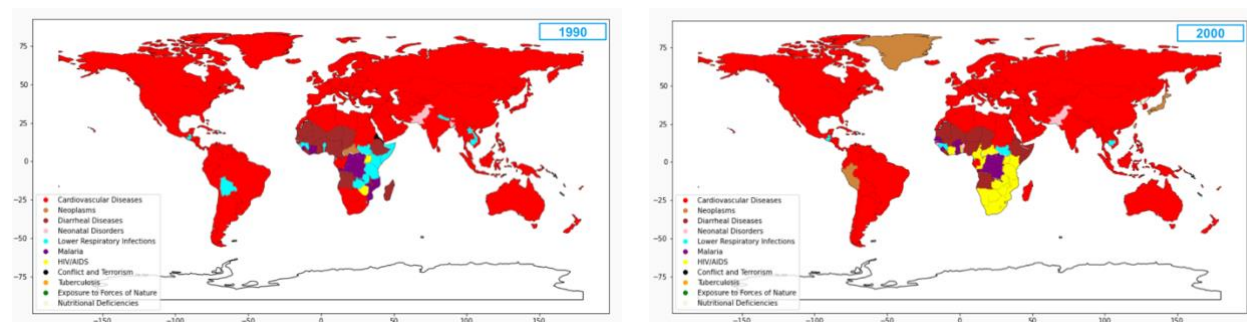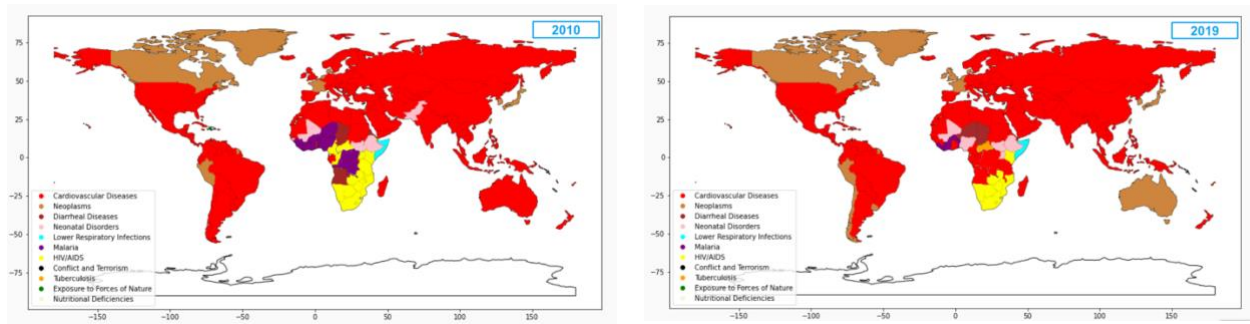
Interpretation: thanks to these maps, we can clearly see the evolution of the causes of death across the world, throughout the years. The first observation we can make is the particularity of African causes of death compared to the rest of the world. This is directly linked with the wealth of the country and the development of modern medecine. A second observation is the increase of neoplasms in the modern world, that our literature review links with modern lifestyle. A third but more specific observation is the outburst of HIV in African countries that can clearly be identify, and the slow process of tackling it.

PROJECT N°4 : DIFFERENCE BETWEEN OECD AND LCD COUNTRIES

Code description : This code first groups data by country using a mean function before removing the year column. It then introduces a definition of the OECD group countries (or the LDC's depending on the subject of the study) and restricts the sample data to this new category. Similarly, it defines three main types of death causes, namely infectious diseases, noninfectious diseases and other (which contains accidents related deaths etc..). It then computes the total number deaths in each of these three categories for each of the OECD(LCD) countries in the study. It then computes a mean of the weight of each of these categories for all OECD countries and uses it to plot the pie chart.

Outcomes :
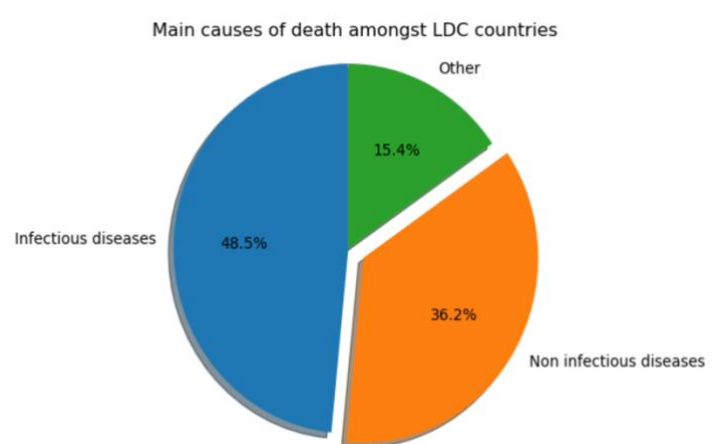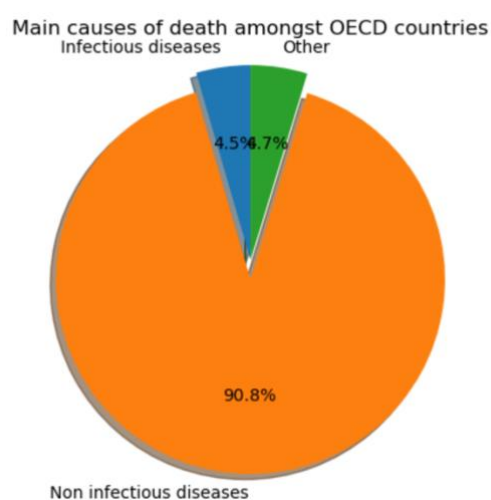
Interpretation : While non infectious diseases indisputably dominate OECD countries, they only come second in least developed countries(LDC) such as Mauritania or Nigeria. This is consistent with OECD and WHO reports which have been highlighting that poor countries tend to lag far behind developed countries when it comes to the dwindling of the share of deaths related to communicable diseases. Fact is that diseases such as HIV are still a major threat in most countries in Sub-Saharan African countries despite the ongoing international endeavor to eradicate them. Besides, some infectious diseases that are extinct in developed countries remains active in those countries and present a genuine threat to international safety.

CONCLUSION

RESULTS SUMMARY

We discovered that causes of death have been dynamically changing, shifting balance further and further into the non-infectious disease side, with infectious diseases remaining dominant only in Africa. Cardiovascular diseases are the main cause of death in most countries in the world, however we can clearly see that the importance of neoplasms is rising, slowly substituting cardiovascular diseases in some parts of the world, for instance Japan, France and Canada. While in developing countries cardiovasculars are rising, it's the neoplasms which are a rising threat in highly developed countries.

In Africa, the infectious diseases remain dominant, however the diseases are changing from diarrheal disorders in 1990s to more HIV dominance in 2010s. Despite the number of deaths caused by infectious diseases falling, the difference is still drastic. The Least Developed Countries have almost half of people dying from infectious diseases, while it's merely 5% in OECD countries. This shows once again how drastic the difference in development is since infectious diseases stopped in developed countries a long time ago.

LIMITATIONS OF OUR STUDY

We identified several limitations to our work: The first major limitation we have encountered is the lack of data at a smaller geographic scale, notably data on cities or counties. This is a problem since we have to assume leading causes of death are homogeneous within one country, which is sometimes not true when comparing people living in metropolitan areas to people living in rural areas.

The time range of our study is also significantly limited as we cannot go beyond 1990. Thus, important changes linked to the invention of vaccines or to other major breakthroughs in the healthcare industry cannot be studied based on this dataset.

Besides these problems linked to data range, we also did not take into account COVID related data as numbers are still evolving as we speak. The pandemic could indeed bring important changes in the hierarchy of causes of death in the short to medium term as COVID probably caused the share of deaths to respiratory infections to rise.

Lastly, as we mainly looked at proportions, we did not track the evolution of the total number of deaths. Fact is that, while the leading causes of deaths have changed over time, the overall number of deaths has risen[8] as world population grew.

FURTHER STUDIES

To tackle the issue of the overall number of deaths it might be interesting to program an anamorphic map that considers both the leading cause of death per country as well as the weight of a given country in the total number of deaths. This could be interesting to track which countries weigh more or less in the total death count over time.

In order to go beyond this descriptive analysis, it would also be interesting to see which external factors have prompted this change in type of leading death cause. Performing a multivariate regression with data on the number of vaccines or on the evolution of people's food diet as independent variable might yield interesting result and explain the rise of deaths due to non-communicable diseases.

Lastly, using data from another dataset centered on the impact of COVID could also be of interest. A regression analysis to estimate the impact of COVID deaths on deaths due to respiratory infections would be valuable to understand the way COVD reshaped the death cause ranking table.

APPENDIX

Sources for the literature review :

[1] https://www.who.int/standards/classifications/classification-of-diseases/cause-of-death

[2]https://ourworldindata.org/causes-of-death

[3]https://www.who.int/news-room/questions-and-answers/item/what-are-the-main-differences-between-rich-and-poor-countries-with-respect-to-causes-of-death

[4]https://www.oecdilibrary.org/sites/ddcd9abfen/index.html?itemId=/content/component/ddcd9abf-en

[5]https://www.paho.org/en/news/9-12-2020-who-reveals-leading-causes-death-and-disability-worldwide-2000-2019

[6] https://www.healthsystemtracker.org/brief/covid-19-leading-cause-of-death-ranking/

[7] https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death

[8] https://ourworldindata.org/grapher/number-of-deaths-per-year