Predicting Life Expectancy of Different Countries

Team Name: Why Axis?

Team member Name	SRN
Adithya M	PES2UG19CS015
Aishwarya Harinivas	PES2UG19CS019
Chetan A Gowda	PES2UG19CS097
R Sharmila	PES2UG19CS309

Dataset Name: Life expectancy dataset(WHO)

Dataset Description:

The dataset contains health factors for 193 countries, which has been collected from the same WHO data repository website. Among all categories of health-related factors only those factors which are more representative were chosen. It has been observed that in the past 15 years, there has been a huge development in the health sector, resulting in improvement of human mortality rates especially in the developing nations, in comparison to the past 30 years. Therefore, for this dataset we have considered data from the years 2000-2015 for 193 countries. As the datasets were from WHO, there were no evident errors. The missing data was from less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc. As finding all the data for these countries was difficult, these countries' data was excluded from the final dataset.

Problem statement:

Predicting life expectancy in different countries using 'Random Forest Regression Model'

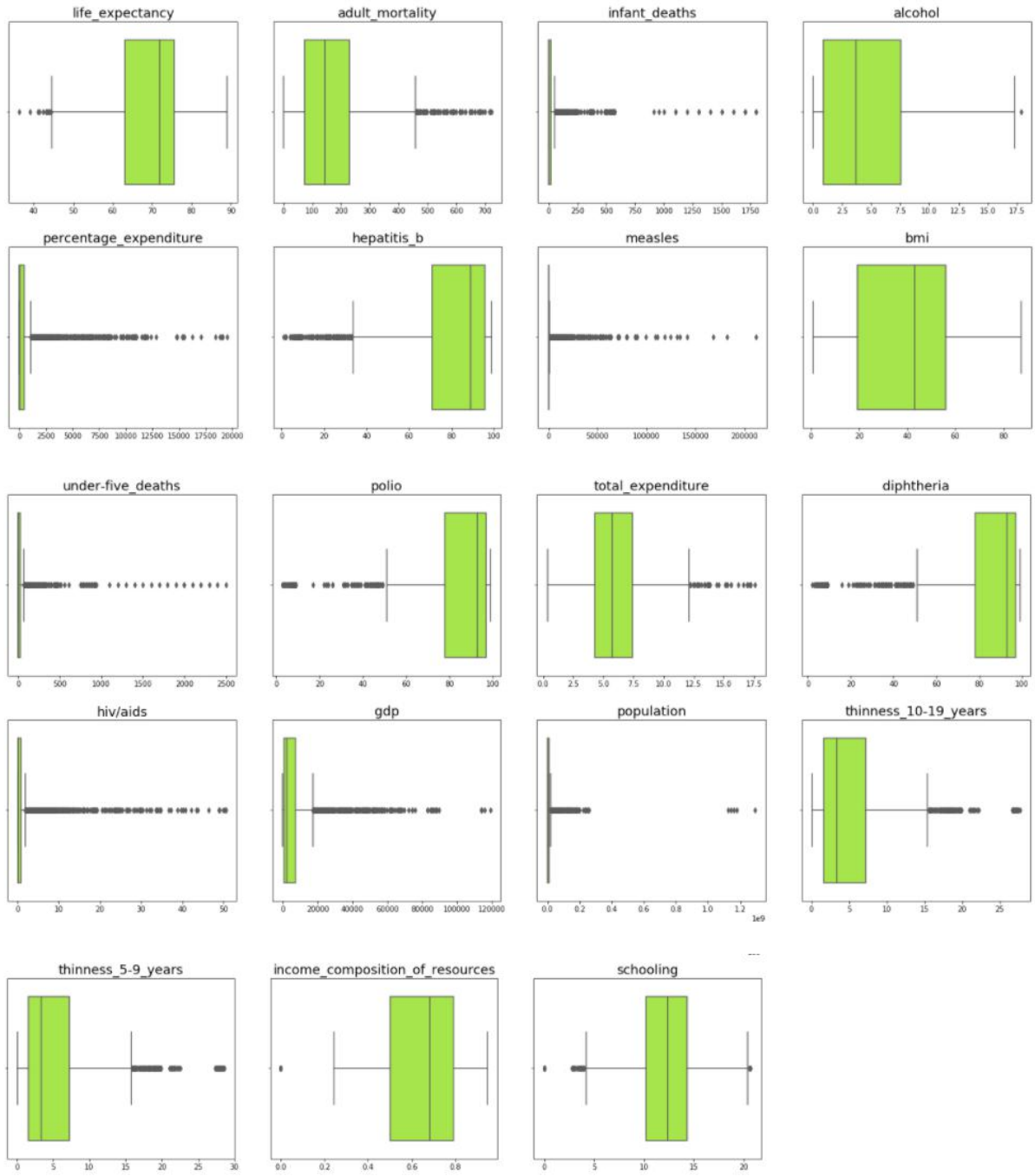
EDA and Visualization:

As part of exploratory data analysis, we started by performing graphical EDA. For this, we first detected the outliers in each of the columns by making use of a boxplot and a histogram for each column. We also plotted the mean in the histograms to check the distribution of the data, that is if the data is symmetric or skewed. Next, we performed non-graphical EDA for which we first removed the non-numerical columns Country and Status from the dataset. We then found some general information about the dataset like count, mean, standard deviation, etc. The last step is to prepare the data for modeling. For this, we first find the number of missing values in each column and then fill the missing values with the mean. Then we perform label encoding on the Country, Year, and Status columns to make it easier for the machine to understand it. Finally, we perform feature selection to only retain those columns that contribute the most to life expectancy.

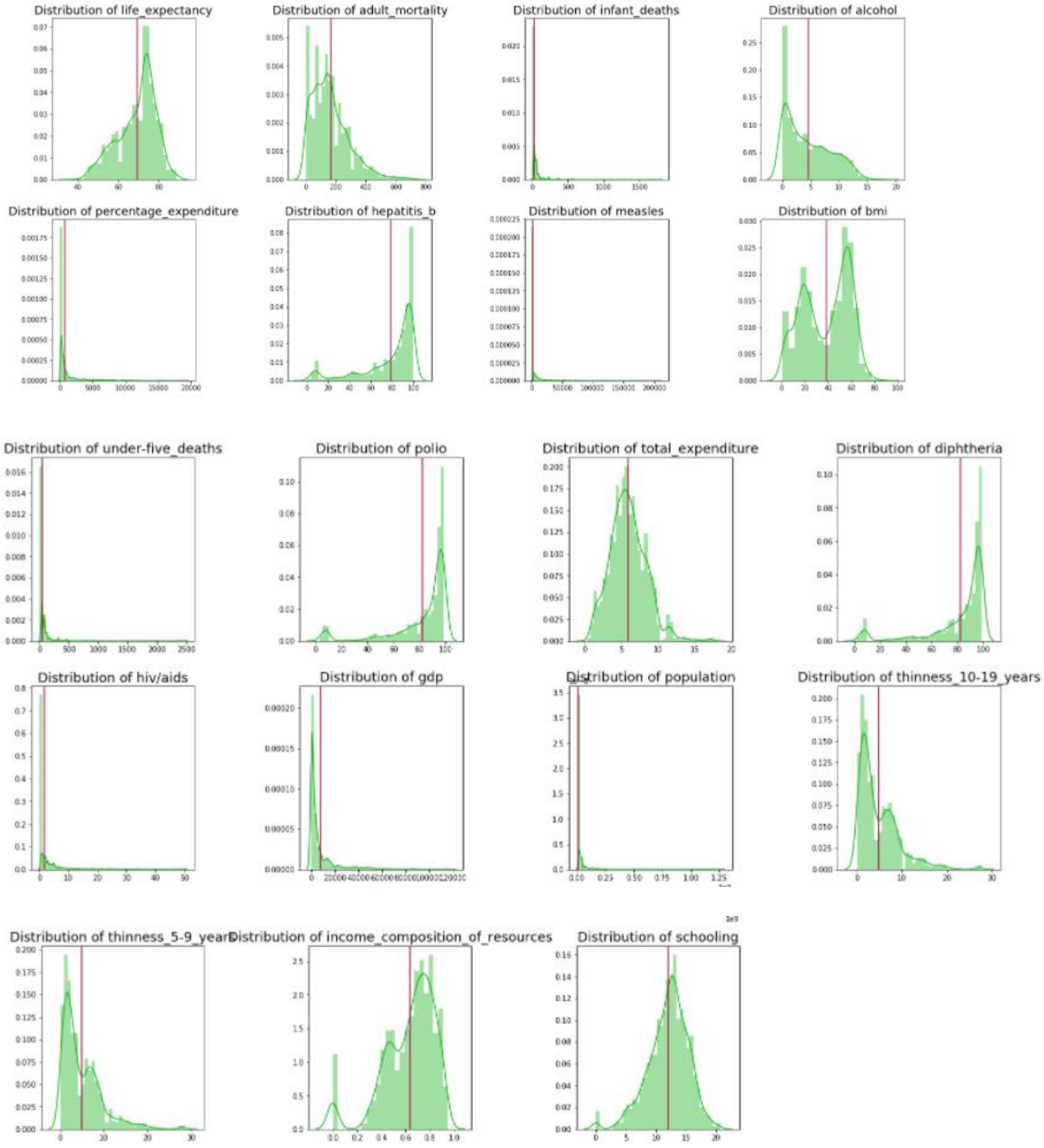
Dataset Insights

polio and dipht… have similar distributions	Similar Distribution
thinn… and thinn… have similar distributions	Similar Distribution
infan… is skewed	Skewed
alcoh… is skewed	Skewed
perce… is skewed	Skewed
hepat… is skewed	Skewed
measle is skewed	Skewed
under… is skewed	Skewed
polio is skewed	Skewed
dipht… is skewed	Skewed
hiv/a… is skewed	Skewed
gdp is skewed	Skewed
popul… is skewed	Skewed
thinn… is skewed	Skewed
schoo… is skewed	Skewed
statu… has constant length 1	Constant Length
statu… has constant length 1	Constant Length
infan… has 848 (28.86%) zeros	Zeros
perce… has 611 (20.8%) zeros	Zeros
measle has 983 (33.46%) zeros	Zeros
under… has 785 (26.72%) zeros	Zeros

Scatter Plot for all the attributes:



Distribution of all attributes via histogram:



How many rows and attributes?

Dataset Statistics

Number of Variables	22
Number of Rows	2938

How many missing data and outliers?

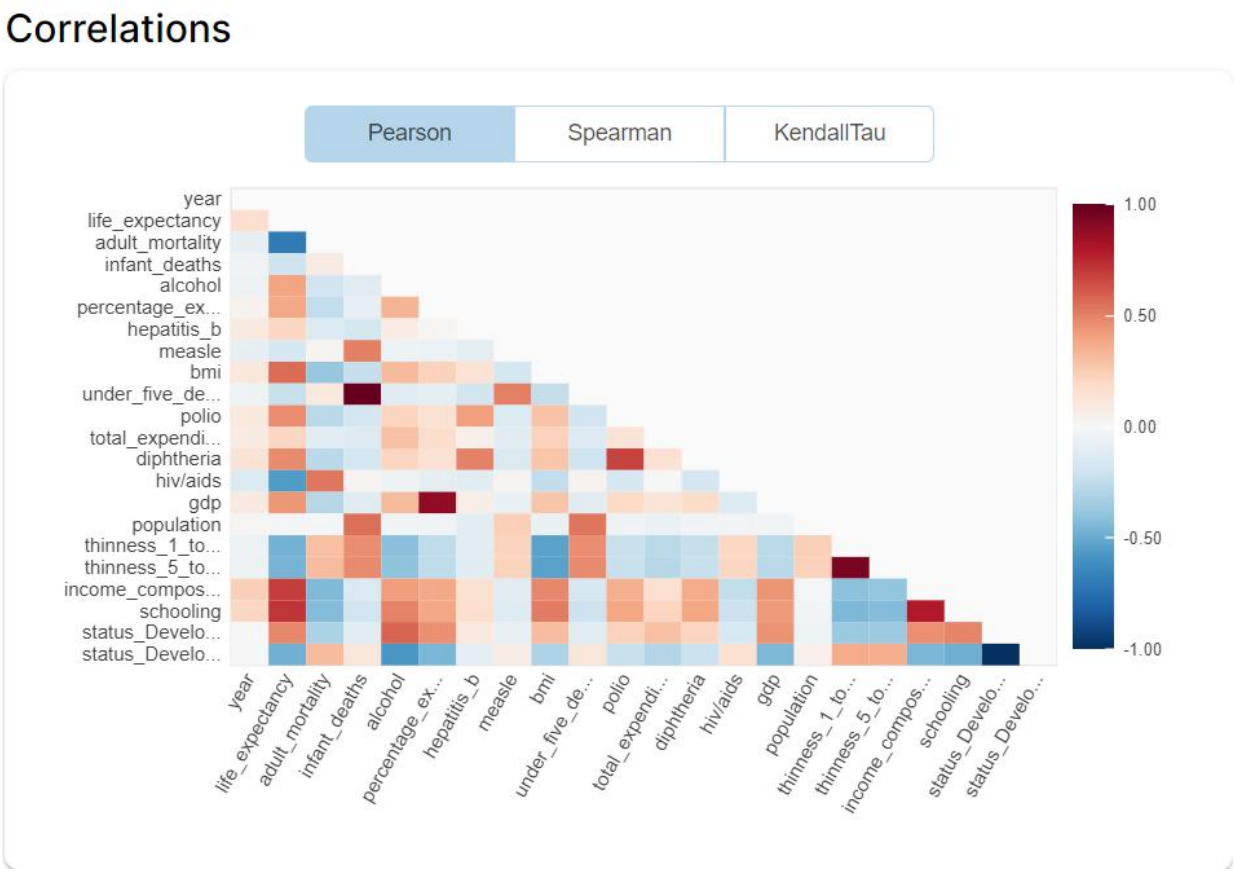
Missing Cells	0
Missing Cells (%)	0.0%
Variable Types	Numerical: 20 Categorical: 2

Any inconsistent, incomplete, duplicate or incorrect data?

Duplicate Rows	0
Duplicate Rows (%)	0.0%

- The random spaces in feature has to be closely observed or it would lead to key error
- The labels of the attributes are wrongly labeled as (per 1000 population) which in turn should be for the whole population of the country
  - The maximum infant death number (per 1000 population) is **1800**
  - The maximum number of reported Measles cases per 1000 population is **212183**
  - The maximum number of under-five deaths per 1000 population is **2500**
  - The minimum population of the country is **34**
- France, Finland have been falsely flagged as developing countries and Greece, Canada are also falsely flagged as developing countries.
- The maximum Average Body Mass Index of the entire population is **87.30** which is not reasonable
- There are percentages greater than 100. Percentage expenditure for Albania is greater than 100% for most of the years. Also, the population of this country varies between almost 2.99 millions (in 2006) and 2941 millions (in 2012).

Are the variables correlated to each other?



Are any of the preprocessing techniques needed: dimensionality reduction, range transformation, standardization, etc.?

- Renaming the label and their description
- Normalization of numerical features
- Dimensionality reduction has its role in predicting accurate life expectancy

Does PCA help visualize the data? Do we get any insights from the histograms/bar charts/line plots, etc.?

Yes, PCA does help in determining the appropriate dimension to predict accurate life expectancy

Data Analytics Literature Survey

Literature Survey By: Adithya M

Determinants of life expectancy and clustering of provinces to improve life expectancy: an ecological study in Indonesia

By: Sekar Ayu Paramita, Chiho Yamazaki & Hiroshi Koyama

Reference Link: ***Determinants of life expectancy and clustering of provinces to improve life expectancy: an ecological study in Indonesia | BMC Public Health | Full Text (biomedcentral.com)***

Summary:

Life expectancy is a population measure of the performance of healthcare systems. Regional disparities in life expectancy in Indonesia have been there for a long and now have become a public health policy challenge. A systematic clustering of provinces can be a valuable alternative for organizing cooperation that aims to increase life expectancy and reduce disparities. Here we aim to identify determinants of life expectancy and designate clusters of Indonesian provinces with similar characteristics. We will also see if there is an alternative method that can be implemented.

We carefully select variables that impact life expectancy and gather 2015 data from Indonesia's Ministry of health. We then perform structural equation modelling (SEM) to select domains that needed to work on from these theoretical models. Then from the results we get from the SEM, we perform cluster analysis to arrange cooperation groups.

Result:

SEM:

We got an adequate fit after 133 iterations where the chi-square value was 0.005 and CFI value was 0.935 and the SRMR value was 0.054. With nine observed variables in the final model out of which six constructs with biblical correlations towards each other and their magnitude of correlation ranged between 0.83 to 0.36.

The result of SEM can be useful to understand the relationship among the variables and is also useful towards designing organized cooperation strategies.

From this SEM we found out the strongest among the six constructs was EXPENDITURE PER CAPITA, which meant enhancing the economy was the most effective approach to improving life expectancy than the other constructs.

Cluster Analysis:

Five clusters of provinces were generated which were sorted based on best to the worst fit of characteristics.

The results of this study show that expenditure per capita is the most influential and core factor in improving life expectancy, as are the health workforce, healthcare facilities, environment, and mean years of schooling.

The result of cluster analysis may be useful to improve coordination between provincial and national governments. These exchanges have the potential to impact provincial integration processes and health policy debates. Cooperation among clusters may strengthen and accelerate health development across the clusters.

Conclusion:

The conclusion we come to here is as the clustering of provinces makes it easier to organize cooperation within and between clusters. Provinces within the same cluster have similar characteristics so they should also have similar goals so these provinces can work together towards the common goal. Also, the national government should support local governments, especially in provinces within the more economically challenged clusters

Pros and Cons:

Pros-

Evaluation criteria were taken initially to ensure the quality of the data. The final model has an adequate fit.

Cons-

The study subject was only the provinces, so the number of observations was only 34.

Abbreviations used:

CFI: Comparative fit index.

SEM: Structural equation modelling.

SRMR: Standardized root mean square residual.

The impact of increasing education levels on rising life expectancy: a decomposition analysis for Italy, Denmark, and the USA

By: Marc Luy, Marina Zannella, Christian Wegner-Siegmundt, Yuka Minagawa, Wolfgang Lutz & Graziella Caselli

Reference Link: ***https://genus.springeropen.com/articles/10.1186/s41118-019-0055-0#Sec2***

Summary:

We see that Life expectancy particularly in the industrialized country has a strong increase due to the significant reduction in mortality rate. Here we question to what extent does life expectancy increase, i.e., its gain related to change in populations due to increasing education levels.

We decompose the change to total population life expectancy at the age of 30 for people in Italy, Denmark, and the US. We call the education mortality change as "M Effect" and population education change as "P Effect". We use the Replacement decomposition analysis to further subdivide the effects into contributions by individual education groups. The P effect ranges from 15% in USA men to 40% in Denmark women.

The mortality rate in Europe began to decrease due to communicable disease at young ages, including neonatal and childhood ages, to non-communicable conditions more prevalent at advanced ages, cardiovascular revolution, new medical advancements, improvement in living conditions, and health-related behaviours.

Naturally, higher education does not automatically lead to better health. It is also unclear to what extent education itself plays a direct role inside the complex network of various health-related socioeconomic factors.

The analysis requires age and education-specific related data for the populations at risk and the deaths for the start and end years of our observation. For Italy information's about the population by age, sex, and education level was taken from Italian Census data which is available online at the data warehouse of the National Statistical Institute of Italy. For Denmark Information about the population by age, sex, and education level comes from nationwide population registers, covering the ages 30 to 100. For the USA the information is collected from data collection IPUMS-USA.

The estimated education-specific differentials in life expectancy cannot be directly compared between the populations, as the characteristics of the underlying data differ between our study populations. So, to compare results across our study populations we construct a life table as the basis for the decomposition analysis. Their life tables are constructed as consistently as possible.

Result:

While the most powerful contributor to increasing life expectancy was the effect of decreasing mortality within education groups, the changing educational structure of the populations also contributed substantially to the increase in the difference in life expectancy at age 30 in all three countries.

The smaller absolute increases in the P effect in the USA can be explained by the smaller overall increase in life expectancy.

We also saw that Danish women experienced the greatest shifts in the educational structure during our study period, with the largest decrease in the proportion of less educated and the largest increase in the proportion of highly educated individuals.

This study examined the extent to which changes in the educational structure is related to change in life expectancy at age 30 in Italy, Denmark, and the USA.

Three main findings emerged from this analysis: -

1. There were considerable changes in the educational composition in all three countries during the study period. The proportion of those with low educational attainment decreased, while the proportion of high and medium-educated individuals increased. In Italy and Denmark, there were particularly large increases.
2. Life expectancy was distributed in a graded fashion across education groups. Life expectancy was highest for those who had more than high school education, followed by those who had medium and then low-level educational attainment levels.
3. The results from decomposition analysis show that the structural change in education accounted for a substantial proportion of improvements in overall life expectancy in all three countries. Specifically, we found the population's changes in educational attainment explained between around 15% (men in the USA) and approximately 40% (women in Denmark) of the increases in life expectancy at age 30.

Conclusion:

The study provides some extension of our understanding of the mechanisms behind recent improvements in life expectancy in Europe. Our results demonstrate the importance of education in the process of increasing life expectancy. Education helps individuals to develop health-related resources, allowing highly educated people to enjoy longer and healthier lives.

This study demonstrates that progress in education has made important contributions to increasing life expectancy in Italy, Denmark, and the USA over the past two decades. These findings are in line with the theoretical heterogeneity approach, which states that mortality levels and differences in mortality are strongly influenced by the specific risk group composition of the populations. In addition to all the other important benefits of education, it can also be viewed as a powerful health policy that allows more people to enjoy both better and longer lives. We tried to provide a piece of evidence to “Enhancing health outcomes through improved educational attainment is attractive, although we still need better evidence that interventions to improve educational attainment will increase life expectancy”.

Pros and Cons:

PROS

- As we use the Replacement decomposition technique we find an advantage that replacement can be performed separately for each education subgroup.
- Due to the Replacement decomposition technique, we can create a cross-country comparison of differentials in life expectancy by education levels.

CONS

- Our results can be affected by the definition used for education levels, the restriction of the decomposition to changes between a start and an end year of the observation period, the chosen observation period itself, and the decomposition technique used.
- There might be further sources for bias connected to some other technical and conceptual issues

Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950–2019: a comprehensive demographic analysis for the Global Burden of Disease Study 2019

By:*The Author(s). Published by Elsevier Ltd*

ReferenceLink: ***https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30977-6/fulltext#seccestitle10***

Summary:

Age-specific mortality rates are a crucial dimension of population health. Fertility rates and population size and composition also have profound effects on the challenges faced by health systems. With rising mean age, for example, diseases such as dementia are a greater burden on individuals, families, and health providers. Assessing the trends in key demographic indicators is a core challenge for global health surveillance.A variety of sources are available on fertility, mortality, population, and migration, but they vary widely in the quality and completeness of registration. In this study, we represent the 2019 version of the demographic for the Global Burden of Diseases, Injuries, and Risk Factor Study (GBD). This incorporates newly released census, surveys, vital registration, and sample registration data. We generate a total of 990 locations at the most detailed level. To better characterize where countries are in the demographic transition, we have developed a seven-category taxonomy.

The analysis can be divided into seven main steps:

1. Age-specific fertility estimation
2. Under-5 mortality estimation
3. Adult mortality estimation
4. Age-specific mortality estimation using a relational model life table system
5. HIV adjustment
6. Accounting for fatal discontinuities such as wars or natural disasters
7. Population estimation.

Below, we provide a low-level description of each analytical component, with an emphasis on new steps for GDB 2019.

Geographical units, age groups, and time periods:

For the demographic analysis, we seek to make the most of rich demographic data, more readily available and robust at the aggregate level, and increase the precision of estimates at the aggregate level by running the modelling process at both the most detailed level, and at the aggregate level.

Fertility estimation:

We used spatiotemporal Gaussian process regression (ST-GPR) to model age-specific fertility rates for the 5-year age group between ages 15 and 49 in each location from 1950 to 2019.

Under-5 mortality estimation:

Follows the analytical framework. We similarly estimate mortality for the more detailed age groups younger than 5 years and constrained these estimates to equal U5MR.

Adult mortality estimation:

We use five different methods to assess completeness, which are – the generalized growth balance method (GGB), the synthetic extinct generations (SEG) method, and a combined method (GCB-SEG), Bayesian model (BCCMP).

HIV adjustment:

We use relational model life table and Estimation and Projection Package Age-Sex Model (EPP-ASM)

HALE:

We use the Socio-demographic Index and HALE to calculate Pearson's correlation coefficient for the analytical method.

Result:

Our paper presents a comprehensive assessment of demographic changes in 204 countries and territories from 1950 to 2019, with a focus on the past two decades. There have been substantial changes in the demographics of most countries and territories since the turn of the millennium. Both life expectancy and HALE have been expanded Almont universally during our study period. While there have been impressive improvements in mortality predictions. Recent slowdowns and reversals in improvements observed especially among adults indicate that process is not a guaranteed thing. In 2019 we see that half the nations in the world had below-replacement fertility and 34 had a cruder birth rate than the crude death rate. 17 countries were in the precarious state of having a negative rate of population change and negative net migration. Demographic transition, which has largely been a story of faster or slower rate of change, is entering a new phase in many countries with high levels of development.

Low fertility and rising average age might lead to inverted population pyramids, which will be fiscally and socially challenging. Our analysis shows that at the global level, all major mortality indicators and HALE have been improving.

Conclusion:

The most recent decade continued the trend of general progress in reducing global fertility and global mortality. While most countries are following this pattern of progress, there is evidence that the world is nearing a demographic infection point. Half of the 204 countries and territories in our analysis had below-par replacement fertility in 2019. Nearly one in five had entered the post-transition state where the natural rate of increase was negative, with negative net immigration. Those demographic shifts, combined with the trend towards stagnation in or reversals or mortality improvements in some high SDI countries, highlight that continued declines in mortality are not guaranteed.

Additionally, sustained low fertility in the setting of slowing or reversing mortality could hasten the number of countries entering the challenging post-transition phase.

Pros and Cons:

PROS

- The comprehensive nature of this study of fertility, mortality, migration, and population helps revise the taxonomy of the demographic transition.
- The data processing steps required to account for known biases, and the data synthesis stage, which deals with the challenges of both missing measurements in given location-years and the common problem of different measurements disagreeing with each other.

CONS

- Our computational resources did not allow us to propagate uncertainty for some covariates through our analytical process.
- Our migration estimates could be improved.

Literature Survey By: Aishwarya Harinivas

A research study on the variables affecting Life Expectancy Descriptive and inferential statistics with Excel and R. By Suresh Kumar Karna and Elisa D'Odorico

Reference link:
https://www.researchgate.net/publication/346345860_A_research_study_on_the_variables_affecting_Life_Expectancy_Descriptive_and_inferential_statistics_with_Excel_and_R_Data_Models_and_Decisions_-Professor_Pompeo_Dalla_Posta

Summary:
This paper aims to analyze how various factors such as GDP, traffic accidents, mortality rates affect life expectancy. It aims to achieve this by performing descriptive, as well as inferential statistics on the data. The dataset has 14 variables, of which 3 are qualitative, and 11 quantitative. Descriptive statistics analysis to analyze the strength of the relationship between life expectancy and the variables is performed by plotting graphs between life expectancy and one of the quantitative variables. One such example is the plotting of a line graph between life expectancy and health expenditure. From the graph it was inferred that there exists a positive relationship between the two variables. In order to perform inferential statistics, a multiple linear regression has been used. MLR allows us to obtain more precise insight into how all the variables affect life expectancy and draw conclusions. Using MLR, the values of the variables are estimated, R2 is calculated to check how much variance is explained by the model and the residual error is also calculated. It also helps to estimate p-value and perform F-statistics to decide whether to accept or reject the null hypothesis which states that there is a significant relationship between the variables and life expectancy.

Pros and Cons:

Pros-
Focuses mainly on the relationship between life expectancy and the variables collected rather than yearly growth or the country.

Cons-
- The dataset is not a collection of observed data, but a prospect based on interpolated data from 5 years.
- Convenience sampling was used to create the dataset, as a result of which the dataset had a lot of missing values.

Results and Conclusions:
Only a few factors related to mortality are statistically significant in their MLR model. Other factors such as GDP, alcohol consumption, etc. are not captured by it as their correlation with life expectancy is weak. However, there are various factors related to mortality that have high correlation with life expectancy. Several initiatives can be taken to lower the mortality rates, for example, mortality due to traffic injuries can be reduced with better road laws and surveillance. Further, improvements in sanitation and access to good quality medical infrastructure can improve the mortality rates from both injuries and diseases.

Trends in life expectancy and healthy life years at birth and age 65 in the UK, 2008–2016, and other countries of the EU28: An observational cross-sectional study

By Claire E.Welsh, Fiona E.Matthews, and CarolJagger

Reference Link: https://www.sciencedirect.com/science/article/pii/S2666776220300235

Summary:

This study aimed to understand the significant deceleration in the increase in life expectancy in the UK, by comparing the trends in Life Expectancy (LE) and Healthy Life Years (HLY) between the UK and other countries of the EU28 using harmonized data in order to identify important differences and provide a benchmark of the UK's performance before the COVID-19 pandemic. The Sullivan method was used to estimate age and sex specific HLY at birth and at the age of 65 for each country of EU28. The statistical analysis was performed by plotting country and sex specific trends in LE and HLY. Whenever non-linearity was suspected, change-point linear analysis was implemented, and the result was compared to simple linear models using automated single knot placement and adjusted R2 values. For those countries with reductions in HLY between 2008 and 2016 as well as the greatest gains, Arriaga decomposition technique was used to examine the extent to which the HLY changes between 2008 and 2016 for each country, with changes in the mortality by specific age groups. As the prevalence of unhealthy life varies over time, the validity of the decomposition conclusions was tested by repeating the analysis on the 2009 to 2015 dataset.

Pros and Cons:

Pros-

- Performs a very large and comprehensive survey on healthy life across the whole population of the countries of interest, over a wide multi-year timespan.
- The direction and timing of the changes match those of the Marmot review.

Cons-

- The study could not identify health inequalities within the countries (only between different countries).
- Data for Croatia prior to 2010 were missing.
- Incomplete harmonization of the EU-SILC question.

Results and Conclusions:

In 2008, LE at birth was highest for French women (84.8 years) and Swedish men (79.2 years). Lithuania and Bulgaria had the lowest LE with 65.9 years for men and 77 years for women. LE in the UK in 2008 was the 17th highest for women (81.8 years) and 10th highest for men (77.7 years). By 2016, the highest LE for women and men were recorded in Spain (86.3 years) and Italy (81 years) respectively, while the lowest remained the same. The UK continued to remain the 17th and 10th LE at birth in 2016 for women and men respectively. Modelling showed that LE increased steadily for most countries, except UK and Germany. The gradient of UK's LE at birth showed one of the steepest increases until 2011, but then dropped below all the countries. Modelling also suggested that the increase in UK men's LE at age 65 slowed down significantly around 2011. The common factor that led to the decline in HLY was a greater increase in the prevalence of unhealthy life, particularly at younger ages, than reduction in mortality.

Predicting life expectancy at birth
By Khalaf Alsalem, Alyx Steinmetz, Nayawiyyah Muhammad, Danielle Frierson and Michael Nashed

Reference Link: https://sci-hub.ee/10.1109/ICCIS49240.2020.9257630

Summary:

Knowing the range, variation and significance of predicting life expectancy allows us to make key and informed planning decisions to better prepare resources for the future. It can also be beneficial in preventive health care and predicting factors that influence mortality. The lifespan of people in different parts of the world varies, not only over regions but over time as well. The fundamental question that drives this project is "What contributes to different life expectancies in different countries?". The goal of this research is to predict life expectancy based on the conditions that a person lives in and the available resources in their region. Excel was used to clean up the data, followed by Azure to perform analysis on the data. The models used in Azure were Clean Missing Data along with Split Data for initial analysis, Linear Regression as the algorithm, Training Model and Pearson Correlation. For visualization, Tableau was used. The methodology involves first analyzing the data from one year to see how the model works, and then showing the correlation between independent variables and the dependent variables, so as to eliminate those variables that have no correlation with the dependent variable and excluding the features that have a very low r score. After this, two regression models (Linear Regression and Boosted Decision Tree Regression) are performed to compare their results and choose the most appropriate.

Pros and Cons:

Pros-

The data is taken from the World Bank and is hence very usable and organized, making data preparation simpler.

Cons-

The data available from early years is insufficient, leading to high variance in the correlation outcomes every year.

Results and Conclusions:

Boosted Decision Tree Regression is chosen because this model gives us the different combination of features that attribute to life expectancy, instead of just producing the higher predictors as in Linear Regression. From the results, we see that there is a lot of variance in what factors affect life expectancy throughout the years. Therefore, the conclusion that we arrive at is that we need to gather as much data as possible before going forth and building the model to get more accurate results.

Literature Survey By : Chetan A Gowda

Factors Explaining Average Life Expectancy: An Examination Across Nations

By: Akansha Maity, Emelie Rhenman, and Elijah Sanders

Reference Link: https://smartech.gatech.edu/bitstream/handle/1853/59089/final_paper_0.pdf

Summary:

The data sets contain variables for nation population, GNI per capita (PPP), poverty headcount ratio at $1.00, life expectancy at birth for males and females as well as the averages between the two, the expenditure on health per capita, the completion rate of secondary education, physicians per 1000 individuals as well as the number of hospital beds per 1000, and the adequacy of social protection (Social Security). However in order to ensure these things, one must first understand what factors may affect them and to what degree.

Although a great many factors can be said to affect the health and well being of a population, it is only realistic to cover a comparatively small number of such factors for the sake of statistical analysis. Regressions between these variables show whether or not the variables are correlated as well as degree of correlation.

Result:

The simple regression models with the average life expectancy in years as the dependent variables and all combinations of Income, Expenditure, Education and Poverty were chosen as independent variables that decide the average life expectancy. The model with poverty and expenditure proved significant since the coefficients have changed. The poverty's coefficient is slightly less negative and expenditure has a large coefficient in comparison to poverty. Furthermore, the R-squared value has increased. However, multiple models similar to these found that these variables were not statistically significant. More than these models were used and in the binary models with the poverty dependent variable and some of the new variables, the P-value was above 0.1 for social protection as well as the dummy Gini variable.

Conclusion:

Though there are models that have statistically significant and possibly economically significant independent variables, the variables chosen are probably not the best measures of average life expectancy. As one knows, average life expectancy can be influenced by gender, genetics, lifestyle, etc. and though these variables might be correlated with some of the variables that were studied in this paper, using the true influencers might have been better suited for modelling. Therefore further analysis should be done in order to study good health and well-being. One recommendation would be to use a dependent variable such as infant mortality, which may be more easily affected by the dependent variables studied in this paper.

Assessing the potential impact of COVID-19 on life expectancy

By: Guillaume Marois, Raya Muttarak, Sergei Scherbov

Reference Link: https://scihub.yncjkj.com/10.1371/journal.pone.0238678

Summary:

The impact of the ongoing Corona virus disease global pandemic which started at the end of 2019 (COVID-19) will last for many years to come, since there is evidence of human-to-human transmission of a novel corona virus, outbreaks of COVID-19 have caused a significant number of deaths worldwide.This paper provides first estimates of the potential direct impact of the COVID-19 pandemic on period life expectancy. At the moment of writing, infection fatality rates are unknown, as the dimension of the contagion is yet to be investigated systematically. Uncertainty around the infection fatality rates is mainly due to the difficulty of identifying the real incidence and prevalence of COVID infection at any point in time. This study shows the impact of life tables on the outcome of the COVID-19 pandemic, all other things being equal and, accordingly, the outcomes can be transposed to any regions sharing similar life tables. While this exercise does not serve as a prediction of what will happen to life expectancy in different contexts, it shows what the potential impact on life expectancy would be if the same age-specific infection rates and fatality rates of Hubei province were replicated elsewhere to regions with different population structures. Given two important features of the actual pandemic. First, the available evidence regarding the prevalence of COVID-19 infection and second, as a consequence, the infection fatality rates of COVID-19 both remain largely uncertain. Rather than providing an estimate, we offer a range of possibilities based on different scenarios of infection rates.

The calculations rely on the assumption that the age-specific infection fatality rates would be the same in all world regions. The study provides six alternative COVID-19 prevalence rates ranging from 1% to 70%. The 1% assumption would be a scenario in which the propagation of the virus is well-contained, while the 70% prevalence would be a scenario in which the virus is spread widely due to limited public interventions to control transmission. In all scenarios, prevalence rates are assumed to be reached within one year.

Result:

The total fatality rate from COVID-19 is 1% in North America and Europe, and ranges from 0.6% under the lower 95% CrI of fx to 2% under the upper one. Total fatality rates are about twice lower in Latin America and the Caribbean and in Southeastern Asia than in North America and Europe and 5 times lower in sub-Saharan Africa, which highlights the role of age structure in vulnerability to mortality risk from COVID-19.

In the absence of COVID-19, life expectancies for men and women combined in 2020 were expected to be 79.2 years in North America and Europe, 76.1 years in Latin America and the Caribbean, 73.3 years in Southeastern Asia and 62.1 years in sub-Saharan Africa.In North America and Europe and in Latin America and the Caribbean, each percentage increase in the prevalence of COVID-19 infection would reduce life expectancy by about 0.1 year. The reduction in life expectancy is slightly steeper when the prevalence is low and becomes flatter when the prevalence gets higher. At an infection prevalence of 10%, a little over one year of life expectancy is lost, and at 50% of infection prevalence, about 5 years are lost.

Given the uncertainty in the estimate of age-specific infection fatality rates, the number of years lost in life expectancy can fall within the upper and lower 95% CrI of fatality rates. With respect to the upper limit, 11 years of life expectancy are lost at 70% prevalence in North America and Europe, 10 years in Latin America and the Caribbean, 8 years in Southeastern Asia and 5 years in sub-Saharan Africa.

Naturally, the years of life lost in the lower limit are lower, equivalent to 4 years under 70% prevalence of COVID-19 infection in North America and Europe and in Latin America and the Caribbean. In Southeastern Asia and sub-Saharan Africa, the loss would be even smaller, 3 years and 1 year, respectively. In general, the loss in life expectancy remains low as long as the prevalence does not exceed a certain threshold. Indeed, with a prevalence of infection below 1%, the years of life lost are likely to be smaller than the annual secular increase, which is about 0.2 in high-income countries, and the trend would remain unaffected. However, with an above-20% prevalence of COVID-19 infection, the effect on the secular trend can become sizable.

Conclusion:

The study relies on scenarios by simulating the pandemic under different assumptions of mortality and infection rates. Our study provides "what if" scenarios that can give policy-relevant information on what could potentially happen to life expectancy under different levels of prevalence that vary with public health measures to control the spread of COVID-19. One advantage of our approach, which uses four different life tables ranging from low to very high life expectancy, is that our estimates can be transposed to any regions or countries that share a similar mortality pattern.Life expectancy is calculated based on annualized mortality rates.

The study assumes a normal distribution of infections and subsequent deaths centered in the middle of the year. The study estimates of the impact of COVID-19 on life expectancy thus smooth the peak of infections for which there would be a higher one-off impact than was suggested by annual life expectancy, even under low prevalence scenarios. The limitation of the study lies in the fact that true case specific fatality rates are unknown. We rely on the estimates of age-specific infection fatality rates adjusting for biases based on data from Hubei province, China. It is highly likely that the true infection fatality rates in other regions differ from that of Hubei province, given country differentials in policy interventions, health infrastructure and population behaviors. In fact, mortality rates are not independent of the prevalence of COVID-19 infection. As the prevalence becomes higher, health infrastructures are likely to be overloaded and to be unable to provide care for everyone who needs it, resulting in higher mortality rates both from the virus itself and from other causes. The risk of mortality is also related to the performance of the health system both in terms of access to it and quality of health care services. This problem is likely to be exacerbated in low-income countries where health care systems lack critical-care resources. On the other hand, the fatality of the virus may decrease as its spreads, given that people who suffer severe symptoms are less likely to contaminate others than those with mild symptoms. It therefore remains to be seen how many years of life will actually be lost following the COVID-19 pandemic.

Life expectancy across high income countries: retrospective observational study

By: Jessica Y Ho, Arun S Hendi

Reference Link:
https://sci-hub.ee/https://www.researchgate.net/publication/327047623_Recent_trends_in_life_expectancy_across_high_income_countries_Retrospective_observational_study

Summary:

Life expectancy is a key summary measure of the health and wellbeing of a population. A nation's life expectancy reflects its social and economic conditions and the quality of its public health and healthcare Infrastructure. It compares recent life expectancy trends in the United States with those in a set of high-income countries, which overlap with those used in recent cross national comparisons of life expectancy. It uses standard life table methods and graduation to parameterize the life table. A life table is a demographic tool used to compute life expectancy and graduation is a recursive smoothing technique to produce estimates of the average years lived by descendants within an age group. It used Arriaga's decomposition to determine the causes of death. Arriaga's decomposition is a method that partitions changes in life expectancy into cause of death contributions. It computed five cause of death contributions for each country. Negative contributions indicate that the cause tended to reduce life expectancy, whereas positive contributions indicate that the cause tended to increase life expectancy.

Result:

Life expectancy in the USA stagnated while life expectancy in other high-income countries exhibited steady increases. Among the comparison countries, the largest gains in life expectancy were observed for Danish women (1.45 years) and men (1.83 years), and the smallest gains were observed for British women (0.37 years) and men (0.68 years). The stagnation in life expectancy in the USA has led to a further deterioration of its standing in international rankings. It is clear that the USA is falling further behind its peer countries and this divergence has been particularly pronounced since 2010. The gap between American women and women in the average of the other countries grew by 0.68 years, from 2.35 to 3.03 years. For men, the gap grew by 1.34 years, from 2.06 to 3.40 years. The 2014-15 declines in life expectancy are more widespread and larger in magnitude than anything observed in decades. The declines in the USA are distinct from those of other high-income countries. Among most other high-income countries, mortality at older ages was the primary driver of the declines in life expectancy. They experienced declines in life expectancy, deaths related to respiratory and cardiovascular diseases and to Alzheimer's disease, other nervous system diseases, and mental disorders. In the USA it appears to be quite distinct from the other Countries, the decline was attributable to drug overdose and external causes.

If life expectancy in the other countries was frozen at their 2016 levels while life expectancy in the USA was allowed to increase at the rate of improvement it experienced in the 2000s—a period of fairly rapid increase in life expectancy for the USA (1.7 years per decade for women and 2.1 years per decade for men), it would take American women 18 years to match the average of the other countries and 34 years to match the world leader, while American men would need 16 years and 2.5 decades, respectively.

Conclusion:

Life expectancy declined across many high-income countries during 2014-15. In some of these countries, life expectancy rebounded in the following year. Though this suggests that these declines may be a fluctuation rather than a new trend, it remains to be seen whether such simultaneous declines across high income countries will become more common in the coming years or whether these countries will continue to achieve robust gains in longevity. The magnitude of these declines is fairly large compared with previous declines. These recent declines were notable both for the number of countries and for the magnitude of the declines. These increases in life expectancy gaps between the USA and other high-income countries are substantial.

Strengths and Limitations:

One indicator of the reliability and accuracy of data on cause specific mortality are the proportion of deaths coded to ill-defined categories.

One limitation of this study is that influenza and pneumonia may be under-reported on death certificates. Influenza often goes undetected owing to lack of diagnostic testing, and influenza infections may increase the risk of dying from cardiovascular diseases and other respiratory diseases, which are ultimately coded as the cause of death-on-death certificates instead of influenza.

Another potential limitation is the issue of correlated causes of death, also known as the competing risks problem. We dealt with this in two ways: firstly, by using broad cause of death categories, which renders the results less sensitive to the competing risks problem, and, secondly, by ensuring that the results are robust by computing cause deleted life tables using an alternative assumption of constant mortality.

An additional potential limitation is the comparability of cause of death coding across countries. This study does not examine how socioeconomic inequality may be contributing to these declines in life expectancy across countries.

Literature Survey By : R Sharmila

Quantifying impacts of the COVID-19 pandemic through life-expectancy losses: a population-level study of 29 countries

By:*Jose´ Manuel Aburto ,Jonas Scholey , Ilya Kashnitsky ,Luyin Zhang , Charles Rahal ,Trifon I Missov, Melinda C Mills ,Jennifer B Dowd and Ridhi Kashyap*

Reference Link: https://academic.oup.com/ije/advance-article/doi/10.1093/ije/dyab207/6375510

Summary :

Variations in the age patterns and magnitudes of excess deaths, as well as differences in population sizes and age structures, make cross-national comparisons of the cumulative mortality impacts of the COVID-19 pandemic challenging. Life expectancy is a widely used indicator that provides a clear and cross-nationally comparable picture of the population-level impacts of the pandemic on mortality.

In a context in which trajectories of life-expectancy progress became more varied, the COVID-19 pandemic triggered a global mortality crisis posing additional chal[1]lenges on population health. Death rates from COVID-19 tend to be higher among males than females, with higher case-fatality rates among older age groups 17,18—precisely those that have accounted for mortality improvements in recent years.

Life tables by sex were calculated for 29 countries, including most European countries, Chile and the USA, for 2015–2020. Life expectancy at birth and at age 60 years for 2020 were contextualized against recent trends between 2015 and 2019. Using decomposition techniques, we examined which specific age groups contributed to reductions in life expectancy in 2020 and to what extent reductions were attributable to official COVID-19 deaths.

This is the first study to assemble a high-quality data set of harmonized mortality estimates, life tables and age by cause decomposition for 29 countries representing most of Europe, Chile and the USA to provide novel evidence of the cumulative, comparative impacts of the pandemic on population health.

· Out of 29 countries analysed, the COVID-19 pandemic led to losses in life expectancy in 27, with large losses of life expectancy of >1 year in 11 countries for males and 8 among females.

· Losses in life expectancy observed in Central and Eastern European countries in 2020 exceeded those observed around the dissolution of the Eastern Bloc (with the exception of Lithuania and Hungary), whereas similar magnitudes of losses in Western Europe were last seen around World War II.

Compared against recent trends, females from 15 countries and males from 10 ended up with lower life expectancy at birth in 2020 than in 2015—a year when life expectancy was adversely impacted already due to an especially bad flu season.

Losses in life expectancy were largely attributable to increased mortality above age 60 years and linked to official COVID-19 deaths.

Result:

Life expectancy at birth declined from 2019 to 2020 in 27 out of 29 countries. Males in the USA and Lithuania experienced the largest losses in life expectancy at birth during 2020 (2.2 and 1.7 years, respectively), but reductions of more than an entire year were documented in 11 countries for males and 8 among females. In some contexts, such as Denmark and Chile, life-expectancy losses due to COVID-19 deaths were larger than total life-expectancy losses, as increased mortality due to COVID-19 was offset by mortality reductions among other causes. Reductions were mostly attributable to increased mortality above age 60 years and to official COVID-19 deaths.

Limitations:

· Model has potential issues by harmonizing and smoothing death counts with a PCLM

· It was recognized nonetheless that late and/or under-registration may affect the estimates by underestimating losses in 2020.

· Migration wasn't taken into consideration as they didn't contribute for many factors.

Conclusion :

The COVID-19 pandemic triggered significant mortality increases in 2020 of a magnitude not witnessed since World War II in Western Europe or the breakup of the Soviet Union in Eastern Europe. Females from 15 countries and males from 10 ended up with lower life expectancy at birth in 2020 than in 2015.

Life Expectancy and Mortality Rates in the United States, 1959-2017

By: Steven H. Woolf; Heidi Schoomaker

Reference Link: *https://sci-hub.ee/10.1001/jama.2019.16932*

Summary:

This report examines longitudinal trends in life expectancy at birth and mortality rates (deaths per 100 000) in the US population.

Mortality rates were stratified by geography, including rates for the 9 US Census divisions (New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, and Pacific), the 50 states, and urban and rural counties.

Excess deaths attributed to the increase in midlife mortality during 2010-2017 were estimated by multiplying the population denominator for each year by the mortality rate of the previous year, repeating this for each year from 2011 to 2017, and summing the difference between expected and observed deaths.24-26 Excess deaths were estimated for each state and census division, allowing for estimates of their relative contribution to the national total.

Result :

Between 1959 and 2016, US life expectancy increased from 69.9 years to 78.9 years but declined for 3 consecutive years after 2014. The recent decrease in US life expectancy culminated a period of increasing cause-specific mortality among adults aged 25 to 64 years that began in the 1990s, ultimately producing an increase in all-cause mortality that began in 2010. During 2010-2017, midlife all-cause mortality rates increased from 328.5 deaths/100 000 to 348.2 deaths/100 000. By 2014, midlife mortality was increasing across all racial groups, caused by drug overdoses, alcohol abuse, suicides, and a diverse list of organ system diseases. The largest relative increases in midlife mortality rates occurred in New England (New Hampshire, 23.3%; Maine, 20.7%; Vermont, 19.9%) and the Ohio Valley (West Virginia, 23.0%; Ohio, 21.6%; Indiana, 14.8%; Kentucky, 14.7%). The increase in midlife mortality during 2010-2017 was associated with an estimated 33 307 excess US deaths, 32.8% of which occurred in 4 Ohio Valley states.

Although all-cause mortality in midlife did not begin increasing in the United States until 2010, midlife mortality rates for a variety of specific causes.

· Drug Overdoses, Alcoholic Liver Disease, and Suicides

major cause of increasing midlife mortality was a large increase in fatal drug overdoses, beginning in the 1990s.30,35,36 Between 1999 and 2017, midlife mortality from drug overdoses increased by 386.5%.

· Organ System Diseases and Injuries

The increase in deaths caused by drugs, alcohol, and suicides was accompanied by significant increases in midlife mortality from organ system diseases and injuries, some beginning in the 1990s.

Analysis was further done based on certain patterns like:

1. Sex-Related Patterns

It was reported that gender-specific influences on mortality and a growing health disadvantage among US women, including smaller gains in life expectancy than among US men, larger relative increases in mortality from certain causes, and inferior health outcomes in comparison with women in other high-income countries.

2. Racial and Ethnic Patterns

3. Socioeconomic Patterns

4. Geographic Patterns

Limitations:

· Mortality data are subject to errors, among them inaccurate ascertainment of cause of death, race misclassification and undercounting, and numerator-denominator mismatching.

· The weak statistical power of annual state mortality rates and their inability to account for substrate variation, the limits of age adjustment, age-aggregation bias, and the omission of cause-specific mortality data from before 1999.190 Purported rate increases may also reflect lagged selection bias.

· Errors in coding, such as the misclassification of suicides as overdoses,192 or changes (or geographic differences) in coding practices could also introduce errors.

· State mortality rates may also reflect demographic changes, such as immigration patterns (and the immigrant paradox 197-199) or the out-migration of highly educated, healthy individuals.

Conclusion :

US life expectancy increased for most of the past 60 years, but the rate of increase slowed over time and life expectancy decreased after 2014. A major contributor has been an increase in mortality from specific causes (e.g., drug overdoses, suicides, organ system diseases) among young and middle-aged adults of all racial groups, with an onset as early as the 1990s.The implications for public health and the economy are substantial, making it vital to understand the underlying causes

Analysis of Life Expectancy using various Regression Techniques

By: Anshu Pandey, Rita Chhikara

Reference Link: *https://sci-hub.ee/https://ieeexplore.ieee.org/document/9362914*

Summary :

In this study they examined trends in life expectancy, and provided an analysis through data visualization of how it will change according to the country, income, education, epidemic, infant death and sexes. Different regression techniques were applied and compared to develop a predictive model.

Around 193 countries data was analyzed through visualization techniques to bring out the relationship between different parameters which have an impact on life expectancy.

Data set includes attributes such as Country, Year, Status, Life Expectancy, Adult Mortality, Infant deaths, Alcohol, Percentage expenditure, GDP, Population, etc.

Following regression techniques were used:

In the study, Life expectancy is the dependent variable and all the other factors are the independent variable.

- Multiple Linear Regression: A multiple linear regression model involves more than one independent variable to find out dependent variable.
- Polynomial Regression: polynomial regression is used to draw relationships between variables which are having nonlinear relations.
- KNN Regression :The dependent value is predicted by local interpolation of the dependent value associated with the nearest neighbours in the training set.
- Decision Tree Regression :A Decision Tree is a flowchart-like tree Structure and non-parametric supervised learning method.
- Gradient Boosting Regression :It is an ensemble Method combining k learned model with aim of creating an improved composite model.
- R square was used as a performance measure. Higher values indicate that the model explains more of the variability of the response data around its mean.

Results :

Histogram was used to observe characteristics of the data. It was observed Adult Mortality, Alcohol and income composition of resources were negatively skewed and life expectancy, population and schooling gave a positively skewed histogram.

Using box and whisker plot it was found that in case of developed nation the range of life expectancy is small and median score is higher than developing nation median score due to same economic conditions.The developing also includes high income per capita countries like Qatar have high GDP per capita but it is lacking in infrastructure and educational opportunities.

Comparison of Regression Technique was done based on testing and training data's r square score.

Conclusion :

This paper compares and examines association between different health, social and economic parameters on life expectancy. Different regression techniques were applied on the data set to develop a predictive model. Results on the train and test data set were evaluated to rule out over fitting. Gradient Boosting regression gave the best results as compared to other regression techniques.

References :

IEEE
Sci Hub
Kaggle