Investigation of Causes of Death

Shane, Waris, Shridevi
Webpage: https://shane-poldervaart.shinyapps.io/DiseaseDeathVisualization/

Introduction

In this project, we are looking at data sets from CDC(Center for Disease Control and Prevention) for COVID and a data set from IHME, Institute of Health Metrics and Exchange, an independent Global Health Research center for infectious diseases. Both data sets contain information about the mortality related to different counties and regions in the United States in the last few years and some additional attributes such as gender, few types of infectious diseases, age, occupation, education. With this information there is plenty of different opportunities for engaging and informative visualizations such as cumulative bar graphs for showing loss of years due to preventable disease, economic predictions based on current resource demand for population growth, impacts of the different diseases, choropleth maps showing potential geological factors, and finally a stream graph showing deaths by different causes and how those are affected by age groups.

Motivation

The primary motivation for this project is to understand what happens within each year in the context of death in the United States. These highlighted questions and answers could be used in the future to help guide policy making, and generally inform the public as to what happens to individual people. In particular, a goal of these graphs would be to help others understand these statistics as not just numbers, but real people with real families who die over the year. In each of their individual deaths, there is an impact both emotionally and economically to the people around them and long term as demonstrated later.

Visualization Design

1. In the wake of the covid pandemic what was the real comparison between covid deaths and deaths from other diseases with control for education, sex, or usual occupation?

The use of a bar graph with filter options is a clear and concise way for any user to explore the large feature space that exists, and see what changes as they select different filtering options. Ultimately, they can come to understand that some aspects such as sex, education, and occupation all have vast impacts on what the discrepency between people dying of certain causes is.

2. Across age groups how does death by different causes change? (Streamgraph)

The use of a Streamgraph is useful to isolate death to age only, and see what is relativistically the cause of deaths in people of all ages. This helps users to see the breakdown at each age of what could cause death and understand what could happen to cause them such as lifestyle or sudden accidents. Through this users can personally move to see the individuality of death as it relates to them in these statistics instead of simple absolute numbers.

3. Across the US States how the mortality rate prevailed with different infectious diseases (choropleth)

Choropleth map helps to visualize the variation in mortality rate across different geographic locations: states in united states at a glance by gradients in color. The hover feature enables to obtain the details of the geographic location and the analysis criteria, mortality rate being that criterion. The drop down selection for different diseases enables to quickly switch the dataset(subset of data) to obtain the same view as before for different diseases. This helps researchers quickly compare across regions and diseases and address the high need areas. A dynamic update of these maps in the event of a pandemic helps the resources to be reallocated, evaluate for the commonality of highly affected areas to less affected areas. In short, Visualization goes with the saying, “A picture is worth a thousand words “.

4. If we could stop preventable diseases and their deaths in 2020, how would it have affected the population, in terms of years lost?

The density plot enabled us to visualize the distribution of the number of deaths that occured due to preventable diseases at various ages versus the expected number of deaths at those ages if we could stop preventable diseases. The area between these two curves (above the actual and below the predicted) would depict the amount of lives (in years) lost due to the preventable diseases and would have added up to the existing population, having further effects on economic factors like GDP. We add a range slider on the “Age” axis, so it would be easier to narrow down the range of the Age parameter.

5. If we stopped preventable disease/death in 2020, how would it affect GDP?

We use a bar graph to represent the GDP the United States lost(or will lose) over the next 10 years because of the death of people due to preventable diseases in 2020. The lost GDP essentially represents how much these people could have contributed to the economy. We add another axis to the graph and stack it with a line graph to visualize the cumulative lost GDP over these years.

Methodology

1. In the wake of the covid pandemic what was the real comparison between covid deaths and deaths from other diseases with control for education, sex, or usual occupation?

This data analysis can be accomplished by summary analysis of the CDC data frame, ultimately building a summary of all the different combinations of factors to quickly display information in a web format.

This information will be shown in a bar graph with filter dropdowns, allowing for each individual variable such as sex, education, or usual occupation to be filtered down and see what industries affect death rate. Ultimately utilizing the data matrix to quickly pull column counts. This informations is important because it will help display to the viewer how much covid affected certain occupations and the influence occupation has on our life outcomes

2. Across age groups how does death by different diseases change?

Analysis can be accomplished similarly by building a data matrix which only contains age as the key, with the values containing the counts of deaths for diseases previously mentioned in order to be immediately accessible on a web format as opposed to querying the large original data set.

This analysis will be shown in the form of a steam graph, with age on the x axis and the respective deaths being colored on the y axis. It is important to understand how prevalent certain causes of death are at different times in our lives in order to help understand broadly how our body effectively functions as we age and the impact these have on our lifespan.

3. What is the geographical, state makeup of individuals dying from diseases? Why is it different?

Mortality rate data is available in hierarchical grouping by state and region. Choropleth map will be used to visually present the data analysis. Analysis will involve using the intensity of color to represent the high and low mortality rates at the State level. Analysis would also involve looking for significant patterns and anomalies across the geographical regions by states and different diseases such as tuberculosis, lower respiratory disease, hepatitis, meningitis, Hiv_aids, Diarrhea.

The Analysis done with hierarchical grouping with interaction is important to get a visual of how the rate compares across regions and states. It is important to know this, look for patterns to delve into the overarching reasons behind these patterns, for example socio-economic aspect, population density, rural vs urban access to health care and address the cause of higher rates by looking at combined CDC and IHME data.

4. If we could stop preventable diseases and their deaths in 2020, how would it have affected the population, in terms of years lost?

From the analysis and data used for Q2, we obtained the age of all people who died due to one of the preventable diseases. This data was aggregated to obtain the number of people who died at each age. The age range we considered for our project was between 1 to 108. We used the death probability estimation at each age from the “Actuarial Life Table”, to calculate the number of deaths that would have happened if we removed the preventable disease factor from

the equation. The excess deaths at each age (when estimated death is subtracted from deaths to the preventable diseases) then sum up to form the total years los

5. If we stopped preventable disease/death in 2020, how would it affect GDP?

Using the historical per capita income dataset from *World Bank Open Data* for the United States from 1970 to 2019, we make predictions about how much the per capita income is going to be for the next 10 years. For this purpose we used the Autoregressive Integrated Moving Average (ARIMA) model. Once we had our prediction, we multiplied it with the number of years lost as obtained in Q4 to obtain the lost GDP. We only considered the ages between 25-60 to calculate the years lost for this story point, essentially because people between these ages typically contribute to GDP. For every subsequent year, we keep moving this window backward, for example to calculate the lost GDP for year 2025, the window we considered is 20-55, to consider the aging of those people.

Evaluation Plan

The goal for the project was to visit sociodemographic factors such as age, gender, occupation, education level and geographic location contributing to epidemic spread. In order to showcase this, we drew inspiration from an arc graph with projected lifespans, streamgraph and choropleth graphs showing diseases spread across geographic regions or by education and occupations. We decided to project each lost life as lost years and loss of GDP.

The intention for this project was to see that the spread of disease had several contributing factors rather than just human contact/interaction. The line of work, occupation could lead to greater risk of getting affected. Economic and/or education level could be another direct influencing factor. The purpose was to use this information to provide appropriate support protocols to prevent or control the spread of disease. The goal was to evaluate the observations from initial visualization to make claims and test for any statistical significance to further implement relevant programs to stop the spread of diseases.

On the implementation end of the project, the challenge was to get data cleaned and formatted for use. Data coming from two different sources had to be merged into our story line. Some challenges with data were that data from two sources were available over different time frames and geographic information was available in different formats. Each one of us had to evaluate the data we took up and think of the analysis with that data and appropriate visualization technique. Mortality rate for different diseases was done with Choropleth, mortality rates with different occupations, gender was done using bar graphs. The experience from this project is a true steppingstone transporting one into an adept data analyst, using visual analytics.

Discussions & Future Work

Our project focuses on understanding what happens within each year in the context of death in the United States and the impact they can have. Our first question answered how the

distribution of deaths changed with sex, level of education and occupation as well as what impact the COVID-19 pandemic in terms of death as compared to other diseases. Moreover, using the streamgraph we dived into how the distribution changes for individual age groups for different diseases. A potential future work can be to add filters for top X number (Say 10) of diseases and try to find connections when factors like age, sex and occupation changes. We restricted our study for the United States and could be extended to developing countries which could potentially offer more insights into the problem and help us study the effects of economical and social factors. We analyzed the lost year and lost death while considering deaths from the year 2020 only, and it can be extended to include multiple years in future works.

References.

Holtz, Y. (n.d.). *Help and inspiration for R charts*. The R Graph Gallery. Retrieved March 1, 2022, from https://r-graph-gallery.com/

US Data. GHDx. (n.d.). Retrieved February 7, 2022, from https://ghdx.healthdata.org/us-data

Gemignani, Z. (2022, May 2). *20 best data storytelling examples (updated for 2022) - juice analytics*. Data Analytics and Visualization Made Easy - Juice Analytics. Retrieved March 1, 2022, from https://www.juiceanalytics.com/writing/20-best-data-storytelling-examples

United States Gun Death Data Visualization by periscopic. U.S. Gun Deaths. (n.d.). Retrieved March 13, 2022, from https://guns.periscopic.com/?year=2013

Centers for Disease Control and Prevention. (2021, December 22). *NVSS - Public Use Data File Documentation*. Centers for Disease Control and Prevention. Retrieved February 7, 2022, from https://www.cdc.gov/nchs/nvss/mortality_public_use_data.htm

Guardian News and Media. (n.d.). *Bussed out: How america moves thousands of homeless people around the country*. The Guardian. Retrieved May 1, 2022, from https://www.theguardian.com/us-news/ng-interactive/2017/dec/20/bussed-out-america-moves-homeless-people-country-study

Okada, S. (2021, August 12). *How to create an animated choropleth map with less than 15 lines of code*. Medium. Retrieved March 7, 2022, from https://towardsdatascience.com/how-to-create-an-animated-choropleth-map-with-less-than-15-lines-of-code-2ff04921c60b#ced3

United States Counties Database. simplemaps. (n.d.). Retrieved March20 1, 2022, from https://simplemaps.com/data/us-counties

"Actuarial Life Table." *Social Security*, https://www.ssa.gov/oact/STATS/table4c6.html Accessed 3 May 2022.

World Bank Open Data | *Data*, https://data.worldbank.org/. Accessed 3 May 2022.

Work Done

Shane Poldervaart

- Project Lead
- Acquiring and cleaning CDC Dataset
- Site Design and Shiny App hosting: Final integration
- Question 1 and Question 2 in entirety (Streamgraph and Filtered Bar Graph)
- General debugging
- Data for Actual Deaths vs Predicted

Shridevi

- Initial topic/Project idea originator.
- Acquiring and cleaning the data set from IHME
- Question 3 and 4 for initial project outline and on Final choropleth implementation
- Meet every week brainstorming, checking feasibility.
- Choropleth shiny app object was implemented, which was taken into final integration for shiny app webpage hosting.

Waris

- Acquiring/Cleaning Data from World Bank and Actuarial Life Table.
- Question 4: Processing the cleaned data and Data Analysis and Visualization implementation
- Question 5: in the entirety (Ideation, Data Processing, Predictive Modeling and Visualization)
- Shiny App unit development for Question 4 and 5.