Causes and Conditions of COVID-19 Deaths in the United States

Overview and Background

COVID-19 and its impacts on public health and general wellbeing are global concerns. During the last 3 years, there have been more than 97 million confirmed cases and 1 million death cases in the United States. Fortunately, as the world starts to understand about the virus and adopt effective vaccines, more countries are opening up and reflecting on the situation.

In this project, we utilize the CDC death and COVID-19 datasets to study the types of diseases affecting people's health and their relations to COVID-19, as well as to visualize interesting patterns and generate predictive time series models. Beginning by conducting an exploratory analysis of the datasets, we study the general data tidiness and plot distributions and correlation among multiple variables. Followed by data preparation where we tidy and clean the dataset, and transform variables into the most suitable format, we will produce the conclusive data visualization findings, along with interactive widgets and predictive models in the R Shiny app. The goal of our study is to help the disease experts, epidemiologist researchers and the general public to get a deeper understanding of how COVID-19 affects the public health and who are most vulnerable to it.

Data Preparation

1. Collect and Import Data: CDC's Open Data can be retrieved via Socrata Open Data API in CSV format.
   a. count_death.csv : shows causes of deaths (including covid-19) monthly in the U.S by sex, age and race.
   b. condition_covi.csv : shows covid-19 death cases involving some health conditions in the U.S..
   c. covi_data.csv : shows all daily cases of covid-19 (include deaths) in the U.S by states.
2. Exploratory Data Analysis
   a. We found and inspected some categorical columns, like column state in count_death.csv.
   b. We inspect ambiguous categorical values. For example, column 4 of condition_covi.csv has 3 categorical values: 'By Total', 'By Year' and 'By Month'. We found they represent different time spans of study.
   c. We conducted research to ensure every medical term is clearly communicated and processed, like 'U071 Multiple Cause of Death', 'U071 Underlying Cause of Death', 'Natural Causes' and 'ICD-10 Code'.
3. Data Tidiness Issues
   a. Multiple columns of diseases can be reshaped into one single "Disease" column. To illustrate, Column 11-22 are the names of Underlying Causes of death in nature but they should be values of a column named 'Underlying Cause'.
   b. Some columns have the wrong datatype, like column 1 of covi_data.csv should be date/time rather than string.
   c. Some NA values are misleading. For example, Column 3 to 13 of covi_data.csv contain NA values, but some of them mean no diagnosis or death of Covid-19 on a specific date, which should be replaced with numeric 0.
   d. Some variables are not consistent. For example, column 'State' has different unique values in 2 datasets since they cover different ranges of U.S. overseas territories.
4. Data Cleaning and Transformation
   a. We will fix the above-mentioned issues using tidyverse's functions like: **pivot_longer()**, **str_to_lower()**, **replace_na()** and **mdy()**.
   b. For modeling and visualizing, we will perform transformations, such as replacing abbr. of state names with full names, extract month & year information and combine data of US's overseas territories.

Modeling

Time series models, for example, the ARIMA model, will be constructed to analyze the patterns of daily confirmed cases for the COVID-19 outbreak during 2020-2021. Predictions for 2022 will be generated and compared with actual observations.

Visualization

The visualization section comprises three elements: KPI cards, visuals, and interactive filters.

1. Visuals
   a. A geographical map showing distribution of covid-19 deaths in each individual state in the United States, filtered by condition group, condition, and age group.
   b. A few plots showing the total percentage of each individual underlying health conditions contributing to causes of deaths in the United States, investigating the importance level of how each individual underlying health condition causes deaths, filtered by gender, ethnicity, and age groups.
   c. One plot using statistical strengths to investigate correlations among all causes, natural causes, select underlying causes of death, and covid-19/non-covid 19 deaths
   d. A time series plot with 2020-2021 covid-19 death data to predict 2022 deaths and compare 2022's against 2020-2021
2. KPI Cards
   a. Death rate: Total causes*100%/Total US population.
   b. % of deaths involved with Covid-19: (column 23 *100%) /column 10
   c. % of deaths not involved with Covid-19: (1 - % of deaths involved with Covid-19)
   d. % of deaths from natural causes: (natural causes*100%)/All causes
   e. % of other underlying causes: (All cause - natural cause) *100% / All causes
3. Interactive Filters**:** age groups, ethnicity groups, gender, state

Disclaimer: This proposal may not include all visualization plans and some parts may change if required.