Project Data Review, A1-4

Haley Bustle July 5, 2023

**Project Hypothesis:** Influenza is a seasonal epidemic that requires an increased need for staffing. This is in part due to a need for illness care and care arising from complications in at risk populations, such as those 65 years of age and older, five years and younger, and pregnant women. If a state has a larger population, then it is more likely to have influenza related deaths in these at-risk groups.

**Data Sets: Population Data by Geography**

**Source:** This is an external data source provided by the US Census Bureau and is a trusted government source.

**Summary:** As indicated from the US Census website, the information and data collected by them is collected both through surveys and administrative data ([Source](https://www.census.gov/about/what/admin-data.html#:~:text=The%20Census%20Bureau%20uses%20data%20from%20a%20variety%20of%20sources.&text=Some%20data%20are%20collected%20from,additional%20data%20from%20other%20sources.)). However, this collection only happens even ten years ([Source](https://www.census.gov/programs-surveys/censuses.html#:~:text=The%20U.S.%20census%20counts%20every,takes%20place%20every%2010%20years.&text=Learn%20about%20all%20the%20decennials%20from%201790%20to%20present.)), which can cause massive discrepancies in a data set. In addition to a time lag, this data has further room for error, as it is collected both through survey (where human error can occur) and administratively (where computer error can occur). Surveys are also at the mercy of people themselves, as not everyone will fill out or respond to them, which can also skew data collection for one or more populations. This data set includes population data from all fifty states, District of Columbia, and Puerto Rico for gender, age categories, and year data was collected. However, this data set is from 2009 – 2017 and the most recent census was in 2020, which will cause a data discrepancy.

**Relevance:** This data set would be useful to my hypothesis as it examines the age and population density of the United States, however it is seven years old, and the most recent census was in 2020. I will need to take this into consideration when testing my data and will need to test if adjusting the ages of the entire data set by 7 (years) will be appropriate for a more accurate result.

**Data Sets: Influenza Deaths, Geography**

**Source:** This is a data set owned and provided by CDC (Centers for Disease Control and Prevention) Wonder, which is a trusted government source.

**Summary:** This dataset contains information on influenza related deaths categorized by state, year, gender, and age group from 2009 to 2017. This dataset contains mostly data, however, there are a notable number of zeros reported. These zeros, as indicated by the CDC, are marked as an unreliable zero if there are reports of less than or equal to nine deaths, and a true zero report of deaths. However, this dataset does contain enough information regarding deaths related to influenza to study against a total population.

**Relevance:** While this dataset will allow me to study the relationship between the number of influenza deaths in relationship to the total population by state, year, gender, and age group, it does contain several zeros that must be accounted for and can affect the outcome of the study. This dataset studies specifically between 2009 and 2017, and the data is now seven years old which can cause a skewed result when attempting to put this project result to real work practice. These factors will be considered and handled during the analysis process.