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.

**Collection:** 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.

**Contents:** 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 Lab Tests & Patient Visits**

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

**Collection:** This information is collected with a variety of usage data collection methods, such as: hundreds of public health and clinical laboratories reporting influenza tests/results to W.H.O. Collaborating Laboratories Systems and NREVSS (National Respiratory and Enteric Virus Surveillance System and outpatient facilities report to ILINet (U.S. Outpatient Influenza-like Illness Surveillance Network) ([Source](https://www.cdc.gov/flu/weekly/overview.htm)). This is a collection system that may contain both human and computer errors (such as duplicate or error entries) since the CDC is analyzing so many variables regarding positive results.

**Contents:** The first data set, CDC LAB TESTS, contains lab testing specimens versus what type of flu positive results for the year 2010 – 2015 for all fifty states. The second data set, CDC INFLUENZA VISITS, contains information on how many patients were seen for flu-like symptoms (and in what age group) versus number of confirmed flu patients for all fifty states for the year 2010 – 2019. This data set contains several dozen “X” or incomplete data points.

**Relevance:** This data set is not relevant to my hypothesis, as I do not need information relating to testing results of influenza cases or the number of individuals seen / were positive for influenza.

**Data Sets: Children Flu Shots**

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

**Collection:** This data was collected by phone survey to monitor the influenza and vaccination rates amongst children aged 6 moths to 17 years and the COVID-19 vaccination for eligible children/teens aged 18 years and older. Surveys are subject not just to the mercy of the collectors, but to the nature of people themselves. There is much room for error as people may not respond to surveys (skewing one or more age group(s)) or there could be human errors made during collection.

**Contents:** This data sheet includes a variables description page as it is documented on the first workbook by codes. The data set also includes information for age group of patient and some demographics (such as race, language, and education) but for the year 2017 only. It does not include state.

**Relevance:** This data set is not relevant to my hypothesis, as it contains survey-based information on only the year 2017 for certain vaccinations of children.