Simone van der Graaf

10 October 2023

Data Immersion task 1.4

Objective: This project will analyze the impact of influenza in each state, for vulnerable patients of 65 years and older with low vaccination rates.

Hypothesis: If a state has a higher proportion of vulnerable residents (65+ years) with low vaccination rates, then there are more hospitalizations (or deaths) due to the flu in that state.

Data source: population data by geography US Census data

The data covers information about the US population. It comes from US Census Bureau and is therefore external. Even though the numbers are estimates, it is government data and I can verify it as a trustworthy source.

De data is a complete count of the population and is administrative. It is collected manually. The data covers the years 2009 until 2017. This source shows data about the US population per county and state year and divided by gender and age groups per 5 years. Starting with Under 5 years and ending with 85 years and over.

Every 10 years, the US Census Bureau conducts a census counting every resident in the United States. Because it's only done every 10 years it has a risk that it is not up to date and therefore has limitations. Because the data is collected manually there is also a risk of errors.

This data set is very useful for my project, as it clearly shows the population numbers per county and state, per different age groups. This will allow me to compare the number of influenza patients with the actual population. It will show the more vulnerable states for my age group.

Data sources: Influenza Visits data set and Lab Tests data set

Like the Influenza Deaths source, these data sources come from CDC. They are trustworthy governmental sources and are external. CDC has collected the data from healthcare providers and clinical laboratories.

The data is about patient visits and lab tests and is administrative. The information is collected weekly in the years 2010 until 2019. However, in the lab test data source the collection of data changed from season 2015-2016 and therefore only shows results form 2010 until 2015. The Influeza Visits data source shows visits reported by sentinel providers, categorized by the number of visits, number of providers, and the total number of patients

seen by week and state. The Lab Tests data counts the number of positive lab tests by week and state.

Looking at the sources from the data sets, there is a big difference. The influenza visits data set has data from around 3500 outpatient healthcare providers. The lab test data set has data from 300 health providers and 100 clinical laboratories. These numbers are a lot lower and more importantly, results from partly 2015 until 2019 are missing. This limits the information.

Both data sets are not useful for my project as they do not give information about my specific age group.

Data source: Children Flu Shots data set

The data comes from the NORC at the University of Chicago but is used by CDC and therefore it is trustworthy. It is an external source.

The data comes from surveys of completed household interviews via telephone that are done with parents of patients throughout the country. All interviews are from 2017. The data shows provider data, the age group of the child stated by month, family demographics, race, poverty level, marital status parents and vaccination information.

This data set is created from surveys conducted via telephone. It is subject to human error. There is a risk of the data being biased and containing errors or missing information.

My hypothesis concerns the age group 65 years and older. Therefore, I cannot use this data set.