Ally Fornino

Preparing & Analyzing Data

1.4: Sourcing the Right Data Task

**Hypothesis:** If a state has a higher percentage of individuals that are in the vulnerable population for influenza (ages under 5 or over 65), then that state will have a higher rate of influenza deaths.

**Objective:** To help a medical staffing agency that provides temporary workers to clinics and hospitals on an as-needed basis. The analysis will help plan for influenza season, a time when additional staff are in high demand. The final results will examine trends in influenza and how they can be used to proactively plan for staffing needs across the country.

**“Population Data by Geography” US Census Data**

* **Data Source:** The data set is an external data source as it is government data collected by the US Census Bureau, making it a trustworthy source.
* **Data Collection:** The data set is administrative data, collected manually every ten years in the form of a survey online, over the phone, and through the mail.
* **Data Contents:** The data set includes population estimates by gender and age for every county from 2009 to 2017.
* **Limitations:** The census is only done every ten years, affecting the timeliness of the data. There also may be some inaccuracies due to the data being collected via survey (e.g., users making mistakes when completing it) and some potential bias in the results if certain individuals did not respond to the survey.
* **Relevance:** This data set will be used to normalize other variables, including mortality count in the Influenza Deaths data set. It also provides population data for certain vulnerable populations (e.g., 65 or older).

**Counts of Influenza Laboratory Tests and Patient Visits by Week and State**

* **Data Source:** The data sets are both external data sources provided by the Centers for Disease Control and Prevention (CDC). The CDC is a government organization, making it a trustworthy source.
* **Data Collection:** The data set is administrative data, collected manually via survey. The Influenza Lab Tests data set is collected from 100 public health providers and over 300 clinical laboratories throughout the country and Patient Visits is collected from 3,500 outpatient healthcare providers.
* **Data Contents:** 
  + **Influenza Lab Tests:** This data set counts the number of positive influenza laboratory tests by week and state from 2010 to 2015.
  + **Patient Visits:** This data set tracks patient visits to medical providers for influenza by week and state from 2010 to 2019.
* **Limitations:** These data sets are not complete counts of all influenza lab tests and visits in the country due to them being collected via survey. It also may contain manual errors from entry.
* **Relevance:** Neither of these data sets are relevant to my analysis as I will be focusing on hospitalization rates.

**Children Flu Shots Data Set**

* **Data Source:** The data set is an external data source as it is government data collected by the University of Chicago on behalf of the CDC, making it a trustworthy source.
* **Data Collection:** The data set is administrative data, collected manually through telephone interviews with parents throughout the country. The data comes from a random sampling of parents.
* **Data Contents:** The data set includes flu shot data for children by state and family demographics.
* **Limitations:** The survey was conducted manually and demographics are self-reported, potentially causing inaccuracies. The data set is also a random sample of parents so its data may be misleading.
* **Relevance:** This data set is not relevant to my project as I am focusing on hospitalization rates for vulnerable populations.