US Census Data

Data Source
The data comes from the US Census Bureau, a government agency. It is an external, open source data set, collected by a trustworthy source for data collection.

Data Collection Method
The data collected by the US Census Bureau here is administrative data, providing a core understanding for the subject at hand. It is collected annually, and the majority of it is from direct respondents to the census, with some others coming from surveys.

Data Contents Variables
The variables included are the location (which US state or DC), age ranges (following 10 year intervals between 5-84, as well as, infants [<1 years old], 1-4, and 85+), gender, Hispanic Origin, Race, Year (1999-2020), Weekday, whether there was an Autopsy, the place of death, and the cause of death.

Limitations in the Data Set
Since it involves manual collection, there is some potential for biased response (like missing data from those who chose not to respond) and the possible human error involved, however no real world data can completely avoid these scenarios, so this provides near optimal flu data, which will be useful for the project.

Relevance
My hypothesis was “If we allocate more resources to training hospital staff for Influenza season, then more people will fully recover and survive the flu.” So the dataset is partially relevant, as it contains information on patients, such as deaths and patient numbers, but does not contain information on Hospital funding or employee training (which admittedly could be very difficult to find.

U.S. Influenza Surveillance

Data Source

The data comes from the Influenza Division of the CDC, a government agency. It is an external, open source data set, collected by a trustworthy source for data collection.

Data Collection Method

The data collected by the CDC and other localized public health partners is mostly survey data, providing detailed coverage on the transmission, characterization, and changing nature of the virus throughout each region in the country.

Data Contents Variables

The Influenza Visits page's variables are the number of visits, numbers of providers, and number of patients seen, provided weekly from the years 2010-2019, provided by 3500 outpatient healthcare providers.

The Lab Test page's variables are the number of positive influenza tests and the variant type, provided weekly between 2010 and 2015, provided by 100 public health providers and over 300 clinical laboratories across the U.S.

Limitations in the Data Set

As this data involves surveying, there is potential for inaccurate, incomplete and untimely responses. It is still useful, however, as it provides detailed data (with a fairly high degree of accuracy) from a reliable source that could lead to important conclusions.

Relevance

My hypothesis was "If we allocate more resources to training hospital staff for Influenza season, then more people will fully recover and survive the flu." So the dataset is partially relevant, as it provides localized information on how many patients are seen, and the number of positive tests, which could provide useful analysis to my hypothesis. Once again it does not contain Hospital funding or training data.

Children Flu Shots

Data Source
The data comes from the National Immunization Surveys (NIS), a national intelligence agency. It is an external, open source data set, collected by a trustworthy source for data collection.

Data Collection Method
The data collected by the NIS, as well as partners such as The University of Chicago's data collection for the CDC is collected via telephone call questionnaires, asking of parents' children's vaccination status, as well as the date and dosage quantity of the inoculation.

Data Contents Variables
The variables included are the location (which US state or DC), the age (from 6 months-17 years old), the family income status, race, and parental marital status.

Limitations in the Data Set
The data is collected annually and is subject to error as parents may not want to disclose the information, or may not want to provide accurate information, or possibly don't know, depending on their level of involvement. However it still provides a reasonable level of accuracy that can be used to look at trends in vaccination data.

Relevance
My hypothesis was "If we allocate more resources to training hospital staff for Influenza season, then more people will fully recover and survive the flu." So this data is not relevant to my hypothesis.