**Data Immersion: Preparing & Analyzing Data**

**1.4: Sourcing the Right Data**

**Cody Boyd**

***Population Data by Geography US Census Data Set***

1a.) This data source is external, as it is provided by United States Census Bureau. Since the data is collected and provided by the government, this data is considered complete and trustworthy.

1b.) This data is administrative. The United States Census Bureau collects this data manually via surveys, and uses the data to address issues such as neighborhood improvements, education, public health, transportation, etc.

There is a time lag present with the data. While the data has been collecting, it is processed and published. This all takes a considerable amount of time, and a lot could change during that time.

1c.) The data set shows the number of the people by age ranges, the population by state between 2009 and 2017, and genders.

2.) There are some limitations present in the data.

* This data is only conducted every 10 years. Therefore, a lot can change as the data gets older and older.
* The data is entered manually which can cause human error such as a typo or mis-categorization of the data.
* Not every single person in the US is counted, so the data is not 100% complete.
* The age groups are marked as ranges, not individual age numbers.

3.) Hypothesis: If a state has a high “vulnerable population” present, then the influenza death/hospitalization rates will increase.

This data is relevant because I can collect data from the ages <5 and >65 to help determine the medical staffing counts for each state based on the vulnerable populations.

***Influenza Laboratory Tests and Patient Data Sets***

1a.) This data source is external, as it is provided by the CDC (Center of Disease Control). Since the data is collected and provided by the U.S. government, this data is considered accurate and trustworthy.

1b.) Both sets of data are administrative. The data is automatically collected from hospitals, healthcare providers, laboratories, and health departments across the United States. However, some of the information is collected manually via surveys.

There are time lags present with the data as all the data gets sent weekly to the CDC, which then must go through it all before uploading it for the public.

1c.) The data set shows location of patients by state, month/season/year, different types of the influenza virus, hospitalization, and death rates, as well as the influenza positivity rates recorded by age and gender.

2.) There are some limitations present in the data.

* This data is survey based and not a 100% count of the population.
* Many people do not go to the hospital when they have the influenza virus.
* The influenza season changed year to year based on the calendar method that the CDC uses. The current year will begin on October 3, 2021.
* Only 1 cause of death can be listed on a death certificate, which means that many influenzas deaths go unrecognized because of another diagnosis at the same time, such as having pneumonia.
* Human error can be present with the manually entered data or data processing.
* Influenza viruses can change abruptly (referred to as antigenic shift) that results in a virus that is different than currently circulating influenza viruses.

3.) Hypothesis: If a state has a high “vulnerable population” present, then the influenza death/hospitalization rates will increase.

This data is relevant because I can collect data for positive cases, hospitalization and death rates for the vulnerable population and sort by state.

***Children Flu Shots Data Set***

1a.) This data source is external, as it is provided by the National Immunizations Surveys. Since the data is collected via phone calls and direct mail, this data is considered untrustworthy because can say whatever they want without the opinion of a medical professional.

1b.) This data is collected via survey data, manually with phone calls and direct mail surveys.

There is a major time lag present, as direct mail can take quite some time to get to the patient and back to the data agency.

1c.) The data set shows the number of the kids who received the flu shots between the ages of 6 months and 17 years old by state, as well as demographics of the kids’ families, such as their race, marital status, and poverty level.

2.) There are some limitations present in the data.

* Human error can be present with the manually entered data or data processing.
* There could be bias present in the data since people can answer questions subjectively and it is not just processed by a doctor and sent automatically to the NIS.
* Direct mail can get lost or thrown away. Phone calls are not always answered, or the patient doesn’t always call back to respond, therefore the data is not 100% complete.
* Since the data is related to minors, many parents may choose to not share the information of their child, leaving the NIS with incomplete or missing records.

3.) Hypothesis: If a state has a high “vulnerable population” present, then the influenza death/hospitalization rates will increase.

This data is irrelevant because the data is not consistent or trustworthy and will not provide accurate information to predict our staffing needs.