**Data Immersion Exercise 1.4: Sourcing the Right Data**

**#1: Data Set: Population Data by Geography US Census**

* **Data Source:** 
  + This data is from an external data source.
  + The data is owned by the government. It is provided by the United States Census Bureau.
  + Because this is government data, the data is trustworthy and a creditable source.
* **Data Collection Method:**
  + This data is administrative data.
  + This data is collected manually by various data providers such as Department of Veterans Affairs (DVA), Federal-State Cooperative for Population Estimates (FSCPE), National Center for Health Statistics (NCHS), etc. After the data is collected manually, it is aggregated automatically.
  + There is a two-year time lag between the full detail of births and deaths data. The data is collected every 10 years.
* **Overview of data contents:**
  + The data contains the annual population of each county in each state of the US by gender and age from 2009 to 2017.
* **Data limitations:**
  + Since the data is meant to be informative, the likelihood of the data being biased is rare.
  + However, the data is not up to date since it is collected every 10 years and the last data recorded was in 2017.
  + Besides birth and death, emigration and immigration affect population, which is not reflected in the data. So, the data is not entirely accurate.
  + There is a discrepancy between the metric for male and female population and for the grouped ages.
  + The data can contain errors since it was manually collected.
* **Relevance of data set:**
  + Objective: Determine when to send staff, and how many, to each state.

Hypothesis: If a state has higher proportion of vulnerable populations, then more medical staff will be required.

This data set is significant to the project objective because it can help determine the priority of each state based on its vulnerable population account. By identifying which states have a higher proportion of vulnerable populations, the necessary number of medical staff can be sent to those states.

**#2: Data Sets: Influenza Laboratory Tests and Patient Visits**

* **Data Source:** 
  + This data is from an external data source.
  + The data is owned by the government. It is provided by the Centers for Disease Control and Prevention (CDC).
  + Because this is government data, the data is trustworthy and a creditable source.
* **Data Collection Method:**
  + This data is survey data.
  + This data is collected manually by outpatient healthcare staff.
  + There is no time lag as the data is updated on a weekly basis. The data is collected weekly.
* **Overview of data contents:**
  + Influenza visits:
    - The data contains information on patient visits at outpatient healthcare providers for flu related illness by age.
    - The timeframe of this data is from the late 2010 to early 2019.
    - Each week, the number of providers, number of visits, and total patients seen are recorded in each state.
  + Lab tests:
    - The data counts the number of positive influenza laboratory tests each week in each state.
    - The timeframe of this data is from the late 2010 to early 2015.
    - Each week, the total number of specimens tested, the total number of positive influenza tests, and the total number of the influenza virus type (influenza A/influenza B) are recorded in each state.
* **Data limitations:**
  + The data must be approved before it is made public.
  + Some states have no data (e.g., New Jersey, Florida, Idaho, Kansas, etc.) in the Laboratory tests.
  + There are some inconsistencies in the data in the Laboratory tests. For example, in the Percent Positive column (column F) there are dates like 2-May or March-28, blanks, and decimal values like 0.93 or 0.31, etc.
  + Therefore, this leads to a bias and skewed data because of the states that do not want to make their data public.
  + Also since this is a survey data, the data can contain errors due to deceit or no responses.
  + The “%weighted” and “%unweighted” variables in the Influenza visits are vague and need more clarification on what they mean.
* **Relevance of data set:**
  + These data sets are significant to the project objective because it shows the total number of patients for influenza like illnesses by each state and positive tests. However, these data sets are not relevant to any of the hypotheses because there was no breakdown of the age of the patients.

**#3: Data Set: Children Flu Shots**

* **Data Source:** 
  + This data is from an external data source.
  + The data is owned by the government. It is provided by the National Immunization Surveys (NIS) under the direction of the CDC.
  + Because this is government data, the data is trustworthy and a creditable source.
* **Data Collection Method:**
  + This data is survey data.
  + This data is collected manually through randomized telephone interviews with parents or guardians for the information of their children’s vaccination providers.
  + Time-lag is not applicable since the data set only contains flu shot data for 2017.
* **Overview of data contents:**
  + The data contains the flu shot vaccinations for children 6 months to 17 years old.
  + The data is categorized by location, race, social and marital status of the family.
* **Data limitations:**
  + Since the data is manually collected through phone surveys, the data can contain manual errors or no responses.
  + However, the data cannot be bias because of the random sampling. Random sampling removes an unconscious bias. The control would be the healthcare providers who verify the flu shot information, which adds validity to the data.
  + As previously stated, the data set is only for 2017.
* **Relevance of data set:**
  + This data set would be significant if the data set wasn’t only for 2017. Although this data set is pertinent to the younger age group, there isn’t any way to draw insight or make correlations based on only one year.