Sourcing the Right Data (Task1.4)

1. The population data by geography US Census data

A. Summarize the data source. Include whether it's internal or external data, who owns the data, and how trustworthy it is?

- the population data is from **external** sources by US Census Bureau from 2009 to 2017.
- This dataset is owned by **US Government** and has been made **publicly**.
- The dataset is **trustworthy** because it owns by government.

B. Summarize the data collection method. Is it administrative data, usage data, or survey data? Is it collected manually or automatically? Is there a time lag?

- This data has been collected **manually** by US citizens and it is **administrative** data .
- There is a **time lag** because it is collected **annually** and there could be discrepancies during that time gap such as birth , death.

C. Write an overview of the data contents. What variables are included?

- The dataset contains county, year, gender of population
- Age with 5 years of group and total population from 2009 to 2017.

D. Be sure to note any limitations of the data set. Could the data be biased? Is it collected infrequently? Could it contain manual errors?

- This dataset could contain **missed** information and prone to typos because it is collected manually with surveys.
- This dataset is **not biased** because it collected by **trustworthy** resources and collected annually.

E. Use the project objective and your hypothesis to determine the relevancy of the data set to your project.

- This dataset is **relevant** to my project because the project objective is to send staff and how many staff to each state. So for that i would need population of every state, on the basis of gender, age and year. For example if I want to know the influenza cases over 60 years age in every state then I will definitely get answer from this dataset

2. Influenza Laboratory Tests and Patient Visits data

A. Summarize the data source. Include whether it's internal or external data, who owns the data, and how trustworthy it is.

- This dataset is from an **external** source by CDC(Centre for Disease Control).
- It is highly **trustworthy** because this is owned by US Government .

B. Summarize the data collection method. Is it administrative data, usage data, or survey data? Is it collected manually or automatically? Is there a time lag?

- Data is collected by the Influenza Division at CDC on Influenza activity in the United State. FluView, a weekly influenza surveillance report, and FluView Interactive, an online application which allows for more in-depth exploration of influenza surveillance data, are updated each week. So this would considered as **survey data** for both the dataset and collected **automatically**(FluView and FluView Interactive) and they're not complete counts of all influenza visits or laboratory tests in the United States..
- The data presented each week are preliminary and may change as more data are received. So there has been a **time lag** because data is collected **weekly**, so there could be changes during the gap.

C. Write an overview of the data contents. What variables are included?

- Data contents are for **Patients visit data** is the number of influenza tests by state from 2010 to 2019 on a ***weekly*** basis.

- The variables are region, year, week of the year, number of Influenza like illness(ILI), number of all outpatient healthcare providers, and age group that is divided into 6 groups.
  - Data contents are for **Lab tests** are the number of influenza positive tests by week and state from late 2010 to early 2015. The variables are public health providers, over 300 clinical laboratories, age group (divided into 6 groups), week of the year, ILI total number, and total patients.

D. Be sure to note any limitations of the data set. Could the data be biased? Is it collected infrequently? Could it contain manual errors?

- There is limitations of the dataset.
- I don't think the dataset is biased, because it is collected **automatically** , although there is still a chance of user error during the manual input of the data so it could contain **manual errors**.
- The data is collected frequently after 7 days so it makes the dataset error free.

E. Use the project objective and your hypothesis to determine the relevancy of the data set to your project.

- I believe that this dataset is **relevant** to my project objectives and hypothesis.
- In order to distribution of accurately staffing needs and geographical information with rates of Flu symptoms with age group .

- One of my hypothesis states that patients with influenza symptom and with aged over 65 are more likely to end up in hospital. We can have all the required information for this dataset.

3. Children Flu Shots data set

A. Summarize the data source. Include whether it's internal or external data, who owns the data, and how trustworthy it is.

The surveys are sponsored and conducted by the National Center for Immunization and Respiratory Diseases (NCIRD) of the Centers for Disease Control and Prevention (CDC) and authorized by the Public Health Service Act [Sections 306]. Data collection for the first survey began in April 1994 to check vaccination coverage after measles outbreaks in the early 1990s.

- This is an **external** data source and owns by **CDC**.
- As government data, we can verify this as a **trustworthy** data source.

- This dataset is provided by The University of Chicago.

B. Summarize the data collection method. Is it administrative data, usage data, or survey data? Is it collected manually or automatically? Is there a time lag?

The data is collected from **Survey** provided by NIC(The National Immunization Surveys) conducted through telephone interviews with parents or guardians across all United States and it's territories.
The data is collects in two parts:
a) a household telephone survey
b) A mail survey of vaccination providers .
So the data is collected **manually** that include asking questions to their parents or guardians (for the names of their children's vaccination providers and permission to contact their vaccination providers) to ask (types of vaccinations, number of doses, dates of administration, and other administrative data about the health care facility).
The demographics are **Manually** collected so there could be an error while collecting, but the flu shot information is verified with health providers so can be considered accurate.

C. Write an overview of the data contents. What variables are included?

- The dataset contains Flu shots for children and their age(6 months to 17 years) vaccination details, vaccination providers
- date of interviews, regions
- financial status of family from child.

D. Be sure to note any limitations of the data set. Could the data be biased? Is it collected infrequently? Could it contain manual errors?

- The data is not biased.
- Data is collected from parents or guardians , so some information could be wrong . The data is collected and stored manually so that could be error prone.

E. Use the project objective and your hypothesis to determine the relevancy of the data set to your project.

This data is not relevant to my project objectives because my hypothesis are based on elderly people, so Vaccination information from children would not be helpful for my project.