Exercise 1.4: Sourcing the Right Data

Hypothesis: If a state has a large *vulnerable population* (like adults over 65 years), then more influenza deaths will occur

1.Population Data by Geography US census Data

Data Source:

i) Data source is an external source.
ii) Owners of Source: US Census data
iii) This data is considered trustworthy as this originates from the US government and holds no conflict of interest.

Data Collection Method:

i) This is administrative data.
ii) The *U.S. Census Bureau* collects population data by geography in the United States in by:
    a. *Decennial Census:* A comprehensive count of the entire U.S. population and housing conducted every 10 years, in years ending in "0" (e.g., 2020, 2030).
        i. *Collected via* Household surveys, In-Person Follow-ups (i.e. Census workers or Enumerators visit homes that did not respond), Group Quarters Enumeration (i.e. nursing Homes, prisons, and Military bases).
    b. *American Community Survey (ACS):* An ongoing survey that collects detailed demographic, social, economic, and housing information every year. It provides more frequent data updates, including estimates for smaller geographic areas.
        i. *Collected via* Ongoing Monthly Surveys and Random Sampling.
iii) There is a time lag with this data.

Overview of the Data Contents:

i) The variables included:
    a. County/State
    b. Years between 2009-2017
    c. Total Population
    d. Male / Female Population
    e. Age in 5-year increments

Limitations of Data set:

i) As the data is administrative and considered trustworthy, it should not be bias.
ii) Data is gathered every 10 years in full but is updated monthly using surveys. Hence, there are timeliness issues in this data.
iii) Due to the various methods the *U.S. Census Bureau* collects the data (detailed above), there is a risk of manual errors occurring in the data set.

Relevancy of the data set to your project:

i) This data I consider relevant to my project.
ii) The data provides populations and age groups for every state, presenting data for states with high populations of vulnerable individuals (like adults over 65 years old) when analysing influenza-associated deaths in the project.
iii) The data presents the total and regional populations between 2009 - 2017, which gives me data to calculate a state's vulnerable populations (like adults over 65 years old) vs a states total population.

2.Influenza Laboratory Tests and Patients visits Data

Data Source:

i) Data source is an external source.

ii) Owners of Source: Centres for Disease Control and Prevention (CDC), specifically through the National Notifiable Diseases Surveillance System (NNDSS) and the Influenza Surveillance System.

iii) This data is considered trustworthy as this originates from the US government and holds no conflict of interest.

Data Collection Method:

i) This is administrative data.

ii) Data is collected weekly from a network of healthcare providers, laboratories, State and Local Departments across the US report on influenza-like illness (ILI). Laboratory-confirmed cases, hospitalizations, and mortality related to influenza are sent through secure electronic reporting systems where the data is aggregated and analysed by the CDC to monitor disease trends, identify outbreaks, and guide public health response efforts.

iii) There is a time lag.

a. The data for Influenza Laboratory Tests only shows from 2010-2015, includes only data prior to the 2015-16 influenza season.
b. The data for Patients visits shows from 2010-2019

Overview of the Data Contents:

i) The variables included in Patient Visits data set:

a. Year
b. Week No.
c. States/Regions
d. No. of providers
e. Total Patients

ii) The variables included in Influenza Lab Tests data set:

f. Year
g. Week No.
h. States/Regions
i. Total specimens collected
j. Type of influenza (if a patient tested positive)

Limitations of Data set:

i) As the data is administrative and considered trustworthy, it should not be bias.

ii) The data has been collected historically and I do not have access currently to more recent data.

iii) The Influenza Lab Tests data set shows errors, i.e. inaccuracies in the percentage positive columns, using a mix of category formats; including dates, integers, and floats. This makes the data unclear and unreliable. This leads to a lack of confidence using this data.

iv) The Patient Visit data set has concerns with inconsistencies in some of the columns, i.e. %UNWEIGHTED ILI data has integers, floats, and 'x' category formats.

Relevancy of the data set to your project:

i) The Patient visit data I consider relevant to my project.
   a. The Patient visit data set presents statistics between 2010-2019. It gives a breakdown by State/Region, total patients, and "ILITOTAL" (total number of Influenza-Like Illness cases reported) which I can use in my project analysis.
   b. This data set also covers the timeframe of Population Data by Geography US census Data, hence allows comparison analysis.

ii) The Influenza Lab Tests data set I would not consider relevant to my project.
   a. The Lab tests data set presents statistics between 2010-2015 and is too small a timestamp to analyse (compared with other data sets).
   b. The percentage positive column is a mix of category formats, including dates, integers, and floats. This makes the data unclear and unreliable. Also, the hypothesis does not specify 'types' of influenza, hence is not relevant.

3.Children Flu Shots Data

Data Source:

i) Data source is an external source.
ii) Owners of Source: The University of Chicago runs the surveys on behalf of the Centres for Disease Control (CDC).
iii) This data is considered trustworthy as this originates from the US government and holds no conflict of interest.

Data Collection Method:

i) This is collected through surveys.
ii) Data is collected through random sampling surveys of parents. The demographics are self-reported, but the flu shot information is verified with health providers and can be considered accurate.
iii) There is a time lag as data is taken from 2017.

Overview of the Data Contents:

The variables included:

a. Age Group
b. Household Income
c. No. of Seasonal flu vaccinations by Region/State
d. Insurance coverage/status
e. Marital & Educational status of parent(s)
f. Race

Limitations of Data set:

i) The data is collected through surveys and could have human error.
ii) The data has been collected for just one year (2017).
iii) The data is limited to a sub-group of vaccination only from one university.

Relevancy of the data set to your project:

i) The Children Flu Shots data set I would not consider relevant to my project due to data collection method is not credible and size of data source does not make this user friendly, i.e. only 2017 data available and a complex 73 columns leading to a higher probability for more human error.