Data Immersion
Exercise 1.4
Jordan Novelli

Sourcing the Right Data

CDC Influenza Deaths Data Set & Census Population Data Set

1. Data Source Summary/Collection Method/Data Contents:
    a. When looking at the CDC Influenza Deaths Data Set and the Census Population Transformed Data Set, these are considered External Data sources. This is because this data is collected for numerous states and multiple companies are contributing their collected data to the CDC (Center of Disease Control). This data is owned by the CDC's but there are many contributors that have collected the data to be used by the CDC. Since the CDC is a government organized department this is trustworthy information.
    b. The collection method used for the CDC Death and Census is most likely Interviews and Survey's method. It sounds odd but as mentioned in the reading, deaths are indicated by the doctor. When filing a death certificate, they typically put the cause of death there. It's not considered a normal type of survey, but it is a documented case. That information is pulled from those hospital records and used to create these data sets. This information is logged manually by physicians in medical records that are more than likely stored in a digital format. From there this information could be extracted automatically depending on how it's stored. I would suspect there is a time lag, as you cannot predict when an individual is going to pass.
    c. Just focusing on the CDC Influenza Death Data Set and the Census Population Transformed Data Set, there are a few variables to summarize.
        i. CDC Influenza Death Data Set
            1. You can indicate what states had deaths.
            2. You can indicate what month and year X amount of people died.
            3. You can indicate approximately what age X amount of people were when they died.
        ii. Census Population Transformed Data Set
            1. You can indicate the total population of specific counties in a state.
            2. You can indicate the total male population of specific counties in a state.
            3. You can indicate the total female population of specific counties in a state.
            4. You can indicate the population in relation to age range of specific counties in a state.
            5. You can indicate various populations and age ranges of specific counties in a state that year.

Data Immersion
Exercise 1.4
Jordan Novelli

Sourcing the Right Data

2. There are limitations to the data. As an example, in the Deaths Data Set, ages are indicated by Ten-Year Age Groups. This means each specific age is not collected individually but grouped together. This doesn't allow you to ask specific questions about age. If data indicates that more 85+ year olds died you could not determine if the majority were 85 or 87 or even 92. The data could be slightly biased. If someone did ultimately die of Influenza but had prior health conditions involved immunocompromising diseases, Influenza wouldn't be the only things that killed them. Only one cause of death can be placed on a death certificate. As far as information being collected infrequently, this data is collected each year and depending on when the data is pulled and submitted could lead to some infrequencies but overall, I believe this data is kept regularly up-to-data via the CDC. Manual errors can come into play when doctors are creating their documentations or if there is misinformation from patients.
3. The project objective states, "Determine when to send staff, and how many to each state."
    a. My Exercise 1.3 Hypothesis:
        i. If people 85 years + are dying from Influenza, then flu vaccines should be provided more readily for them prior to their most potent flu season (based on state).
        ii. If there is one type of influenza that infects people most often, then more vaccines related to that virus should be provided.
        iii. If hospitals/clinics are staffed appropriately in states that are hit the hardest by the flu, then more patients will be treated sooner and recover quicker.
    b. I believe the relevancy of the Deaths Data Set is important. You can get a better understanding of the individuals that are most susceptible to death by the Influenza Virus. The Census Population Data Set doesn't give me much more than just how many people are in each County and State each year by gender or age. I believe later down the line this will come in handy if I ask a question relevant to entire population but currently it's a bit irrelevant to me.

Data Immersion
Exercise 1.4
Jordan Novelli

Sourcing the Right Data

Influenza Laboratory Tests & Patient Visits Data Sets

1. Data Source Summary/Collection Method/Data Contents:
    a. Both the Lab Test and Patient Visits are internal data. They are collected by the individual hospitals/clinics and therefore are also owned by those practices. I would consider them trustworthy if the surveys used to collect the data were completed by the staff and not clientele. However, this data is not a good representation of influenza for those states.
    b. This data collection is survey collected as stated in the Task. If the answer wasn't already provided my explanation would be simple. The Lab Test Data indicates what type of influenza they tested for. Therefore, staff would have completed this survey to put that specific detail in their data set. For the Visits Data, there is data indicating the number of staff available at that clinic that week for those percentage of patients. This would be another clinic conducted survey for the clinic/hospital.
    c. Just focusing on the Influenza Laboratory Tests and Patient Visits Data sets, there are a few variables to summarize.
        i. Influenza Laboratory Tests
            1. You can indicate X amount of people that were tested by State, year and week.
            2. You can indicate how many times a specific strain of influenza was detected.
            3. You can compare total number of people tested to X amount of influenza detected.
        ii. Patient Visits Data
            1. You can indicate the total number of providers available by state/year.
            2. You can indicate the total number of patients seen by state/year.
2. I think there are more limitations to the Patient Visits Data set. When looking through the data set there are columns to indicate people by age that were seen, however, those columns all have an "X" as their information. This doesn't allow me to see how many people based on age were visiting due to potential flu symptoms. I can only see the total number of patients and providers available. This is a small amount of data that is useful but there is a lot of information missing. The limitations I see with the Lab Test Data is that in the Percent Positive column there are not percents but dates. This is an issue in data collection, and I'd say that's a manual error. I also don't understand what %Unweighted ILI and % Weighted ILI means in this data sheet so that information is not useful to me.

Data Immersion
Exercise 1.4
Jordan Novelli

Sourcing the Right Data

3. The project objective states, "Determine when to send staff, and how many to each state."
    a. My Exercise 1.3 Hypothesis:
        i. If people 85 years + are dying from Influenza, then flu vaccines should be provided more readily for them prior to their most potent flu season (based on state).
        ii. If there is one type of influenza that infects people most often, then more vaccines related to that virus should be provided.
        iii. If hospitals/clinics are staffed appropriately in states that are hit the hardest by the flu, then more patients will be treated sooner and recover quicker.
    b. The Lab Test data sheet is slightly useful in that it gives me a number that I can total for how many people did test positive for a specific influenza. This information can be useful to me for one of my hypothesis questions. Also, the same data sheet can be compared to the Visits Data sheet, and I can get a general (not complete) understanding of how staffed a state might be compared to the patient load it sees. Because there is no information identifying age, I cannot use it to answer one of my hypothesis questions. I can also use some of this information to help answer my 3rd hypotheses question when trying to gauge appropriate staffing. I cannot use this as a sole source of information because it's not a representation of the entire US. There is some relevancy but it's not as trustworthy in the long run.

Child Flu Shots Data Set

1. Data Source Summary/Collection Method/Data Contents:
    a. This data is external. It is a survey conducted by the University of Chicago on behalf of the CDC. It is collected throughout all states in the US; therefore, UC owns it, but they share this data with the CDC under the guidelines of the National Immunization Surveys. This is trustworthy data; however, not every single household is surveyed but a random sampling of parents.
    b. As stated previously, this is a survey through National Immunization Survey. The data is collected manually as households are called and individuals are spoken to. There could be a time lag. Without knowing the process fully, if a person doesn't answer their phone there may be a period when they are called back several times or completely removed from the survey.

Data Immersion
Exercise 1.4
Jordan Novelli

Sourcing the Right Data

c. Just focusing on the Children Flu Shot data set, there are a few variables to summarize.
    i. Children Flu Shot Data Set
        1. You can indicate the poverty level of the household.
        2. You can indicate that age of the child in question.
        3. You can indicate the number of children in the home.
        4. You can indicate the income of the household.
        5. You can indicate race and sex of the child in question.
        6. You can indicate age the child was at time of vaccination.
        7. You can indicate child provider.
        8. You can indicate the state in which the child lives.

2. At this time the only limitations I can mention are that this is collected on a sampling of parents and not every household in the US. Further limitations are that the ages of children only go from 19 months – 35 months making the children as young as 1.5 years to 3 years. The task indicated that the children's ages were 6months- 17 years and I'm not seeing that data.
3. The project objective states, "Determine when to send staff, and how many to each state."
    a. My Exercise 1.3 Hypothesis:
        i. If people 85 years + are dying from Influenza, then flu vaccines should be provided more readily for them prior to their most potent flu season (based on state).
        ii. If there is one type of influenza that infects people most often, then more vaccines related to that virus should be provided.
        iii. If hospitals/clinics are staffed appropriately in states that are hit the hardest by the flu, then more patients will be treated sooner and recover quicker.
    b. Due to age limitations, I wouldn't be able to answer my first hypothesis. This data set does not indicate the type of influenza that is tested for only if vaccines were given. This data set also does not give any information on hospital/clinic staff. This data set doesn't now help to answer the objective or any hypotheses I created.
