Tara Perrige
Task 1.4

1. Summary of population data by geography data set:
    a. Data source: This is an external data source. It is provided by the US government, specifically the US Census Bureau. The data is trustworthy because it comes from the government.
    b. Data collection: This is survey data collected by the American Community Survey, and it is only sent to a sample of addresses, unlike the US census that is held every 10 years that records information from every household. It is collected manually every month of every year. This specific data set appears to use the 5-year estimates, which means that each year's data is based on the previous 60 months of collected data. That's a huge time lag.
    c. Data contents: This data contains yearly US population estimates from 2009 to 2017 broken down by age group and by county in every state, including Washington DC and Puerto Rico. Ages groups are arranged in groups of five, except for 85 and older, which gets its own group.
    - Resources: https://www.census.gov/programs-surveys/acs/about/acs-and-census.html, https://www.census.gov/programs-surveys/acs/guidance/estimates.html

2. Limitations of population by geography data set: There is some bias inherent in this data set because it is a result of a survey. Even though households are legally required to answer the survey and reassurances are made that responses remain confidential (see resources above), people may be hesitant to answer such personal questions (especially about their employment) or be unwilling to take the time to fill out everything as accurately as possible. (As someone who has received this survey at their household in the past, I can confirm that it is extremely detailed and time-consuming to fill out.) Also, although the data is collected frequently, it is only published yearly from a collection period of 60 months, so it is not very current. Manual errors from putting the survey data into the system are also very likely, especially since there is quantative as well as qualitative data. (From previous experience, I can tell you that this isn't something that can be easily scanned into a computer.)

3. Relevance of population by geography data set: The health staffing agency needs to know which states will need more staff to handle the influx of patients dealing with severe influenza cases. Most people who develop these serious cases are part of the "vulnerable population", which includes people over the age of 65. My hypothesis is that states with a larger percentage of people over 65 will have higher death rates from influenza. This particular data set is relevant because it will show demographic data by state and age group, allowing me to calculate the percentage of people over 65 for each state.

4.

a. Influenza visits data set:
    i. Data Source: This is an external data source. It is provided by the Centers for Disease Control and Prevention (CDC), a US government agency. The data is trustworthy because it comes from the government.
    ii. Data Collection: This survey data is collected from approximately 3,500 outpatient healthcare providers throughout the US (including Puerto Rico, Washington DC, and the US Virgin Islands) and is collected manually. Data is collected each week; participants gather their data from Sunday to Saturday, submit by Tuesday, and the CDC updates on Friday. Since the CDC performs a pretty quick turnaround, I would say there isn't much of a time lag.
    iii. Data Contents: This data contains information on weekly US outpatient visits for influenza-like illnesses from 2010 to 2019. It counts the number of influenza-related visits, the number of providers, and the total number of patients per week per state.
    iv. Limitations: There is some bias inherent in this data because it's entirely voluntary to respond to this survey. Also, data is only collected from a small sample of providers throughout the country. We basically have to trust that there's enough responses for each state to make it more representative. Since data is collected frequently, I don't have to worry about it being out of data. Also, it's possible that there are manual errors.
    v. Relevance: This data set is not relevant, because it does not help to determine the percentage of people over 65 in each state, nor does it help to determine influenza deaths.

b. Influenza lab tests data set:
    i. Data Source: This is an external data source. It is provided by the Centers for Disease Control and Prevention (CDC), a US government agency. The data is trustworthy because it comes from the government.
    ii. Data Collection: This survey data is collected from approximately 100 public health laboratories and over 300 clinical laboratories throughout the US (including Puerto Rico, Washington DC, and Guam) and is collected manually. Data is collected each week; participants gather their data from Sunday to Saturday, submit by Tuesday, and the CDC updates on Friday. Since the CDC performs a pretty quick turnaround, I would say there isn't much of a time lag.
    iii. Data Contents: This data contains information on weekly US positive influenza tests from 2010 to 2015. It shows the percentage of positive influenza tests and the total number of tests by state.
    iv. Limitations: There is some bias inherent in this data because it's entirely voluntary to respond to this survey. Also, data is only collected from a small sample of providers throughout the country. We basically have to trust that there's enough responses for each state to make it more

representative. Since data is collected frequently, I don't have to worry about it being out of data. Also, it's possible that there are manual errors. However, the most worrying fact is that testing practices are different between public health labs and clinical labs. Apparently, public health labs often get samples that have already tested positive from a clinical lab, so this combined data could suffer from duplication. In fact, this is why the CDC began to separate the public health lab data from the clinical lab data starting in 2015, after the latest date in this data set.

v. Relevance: This data set is not relevant, because it does not help to determine the percentage of people over 65 in each state, nor does it help to determine influenza deaths.

5. Flu shot rates in children data set:
   a. Data Source: This is an external data source. It is provided by the National Center for Immunization and Respiratory Diseases (NCIRD) of the Centers for Disease Control and Prevention (CDC), a US government agency. The University of Chicago runs this survey on their behalf. This data is trustworthy because it comes from the government.
   b. Data Collection: This survey data is collected by calling randomly selected cell phone numbers (throughout the US, including Washington DC and some US territories) and asking the parent or guardian questions about whether their eligible children have received flu vaccines. Eligible children include ages 6 months to 17 years. Also, with permission, a questionnaire is mailed to each child's vaccination provider to confirm. This data is collected manually. It is conducted annually, so there is a significant time lag.
   c. Data Contents: This data contains information about flu shot vaccination coverage from 2017 for children (6 months through 17 years) throughout the US, broken down by state and family demographic information. It records how old the child was when they received a flu vaccine and each dose, or if the child did not receive a flu vaccine.
   d. Limitations: There is some bias inherent in this data because it's entirely voluntary to respond to this survey. Some people may be hesitant to reveal personal information (especially regarding their incomes). Responses are further limited to people who have cell phones and people who decide to answer a phone call from an unfamiliar number. It's possible that people from certain states are more willing to talk to strangers, skewing the data so that some states are disproportionately represented. Also, a second response is required from the vaccination provider. Responses from the vaccination providers are also voluntary, which will also lead to a lack of responses. This data is collected more infrequently. It is also prone to manual errors, since phone calls have to be turned into data on a spreadsheet.
   e. Relevance: This data set is not relevant, because it does not help to determine the percentage of people over 65 in each state, nor does it help to determine influenza deaths.