Key Assumption:

- Some of the data in our exercises is set in the past. Updates for this data are readily available, but not given. Since we are to use the given datasets, my assumption to complete the exercise is that we are in 2017 and planning for the 2018 flu season.
- If this is an incorrect assumption, the age of the data is a limitation of great significance and would be listed under limitations in the review of each data set.

Data Set Summaries:

Census Data:

- This is an external data source owned by the U.S. Census Bureau. The U.S. Government has many controls in place to help ensure the most complete data capture possible for the census.
    - The controls include regulations making nonparticipation or falsification a crime ($5,000 fine for failure to participate, $10,000 fine for falsification), multiple methods for survey participation (paper, telephone, electronic) all in multiple languages, and assertive data collection attempts.
    - This is the most trustworthy dataset of this type available.
- The data is collected by survey every 10 years.
    - The annual data in this set is extrapolated and estimated by the U.S. Census Bureau.
- The dataset includes:
    - A breakdown of data by state and county.
    - A breakdown of data by binary gender.
    - A breakdown of data by age group.
    - A data element to identify year.
- Limitations:
    - This data becomes less dependable for exact numbers for each year it is removed from the most recent census. However, for our purposes and projections this is the most reliable data available.
        - The numbers for male and female appear to be taken from the census surveys and then extrapolated to whole numbers for each non-census year.
        - The numbers for each age group appear to be estimates, as evidenced by fractional numbers.
    - The survey data is entered manually:
        - Although aggregated by computer, the survey forms are completed either by the participant or by a (temporary) government employee.
        - There are multiple opportunities for error, omission, and/or falsification.

- Relevance of this dataset to our project:
    - This dataset, when combined with the CDC data regarding deaths related to influenza, can help us pinpoint where to allocate more staff in support of our allocation hypothesis:
        - The age group with the most reported deaths throughout the country over the dataset year was 65+. Further, children under five years of age are also a vulnerable population. If we allocate more resources to counties with large numbers of senior citizens and small children, we can reduce deaths in those populations from influenza.
    - This dataset also fits well with one element of our data wishlist:
        - Age-Group Focused Staffing
            - Detailed location data for senior citizen mortality from the period in the dataset.
            - Detailed localized population data by age group.

Conclusion for Census Data:

- This dataset should be included in our project and used in conjunction with CDC data to help guide our allocation plan.

Influenza Laboratory Tests / Patient Visits:

- These two datasets are being considered together. Both are external datasets owned and maintained by the Centers for Disease Control and Prevention (CDC).
    - The CDC also includes data from other sources, including the World Health Organization (WHO) and the National Repository and Enteric Virus Surveillance System (NREVSS).
    - Since all of the providers are governmental agencies, the sources should be considered trustworthy.
- The data is collected via survey weekly from selected providers, which comprise a small sample of total providers.
- The Patient Visits dataset includes:
    - Data indicators for state, year, and week.
    - The number of providers reporting and total patients seen for each week.
    - Extrapolated data to describe the proportion of Influenza Like Illness (ILI) visits compared to total visits.
    - The dataset includes fields for age breakdown; however, no data was reported for those breakdowns.
- The Influenza Laboratory Tests dataset includes:
    - Data indicators for state, year, and week.
    - The number of specimens tested for each state/period.
    - The dataset includes fields for percentage of positive tests and breakdowns of influenza type found.

- Limitations:
    - The CDC lists a warning message on its website indicating that all of the data reported is projected from what is provided by the reporting entities.
        - The reporting entities comprise a very small sample of total entities.
    - The CDC disclosure further indicates that different testing practices are used across the reporting entities.
        - Not all labs report the type of influenza found.
    - The CDC has applied its own normalizing algorithm to the data.
        - This normalization minimizes the effect of vulnerable populations on their presentation to counter possible data skew from the small provider population.
            - This works in direct conflict with the purpose of our project, since we are trying to allocate staff to areas with larger vulnerable patient populations.
    - There is a tremendous amount of data missing from the dataset.
    - The survey data is entered manually and transmitted electronically.
        - There are multiple opportunities for error and/or omission.
    - There are no controls in place to help ensure completeness of data from the reporting entities.
- Relevance of these datasets to our project:
    - Although the source of the datasets is trustworthy, the datasets do not provide reliable data for our purposes.
    - Further, there is no direct tie between the data contained in the datasets and any of our project hypotheses.
    - The unreliable nature of the datasets, combined with the amount of missing data and no direct tie-in, indicates that these datasets would be of no relevance to our project.

Conclusion for Influenza Laboratory Tests / Patient Visits:

- These datasets should not be included in our project.

Children Flu Shots:

- This is an external dataset owned and maintained by the Centers for Disease Control and Prevention (CDC).
    - The CDC directs and uses a non-governmental group (NORC) to gather the data.
    - Since the data is collected under the direction of the CDC and NORC is a nationally recognized research group, the source of the data is considered trustworthy.

- The data is collected via survey.
    - Participants are found by assertive phone campaigns.
    - If the parents or grandparents of children in the appropriate age ranges for the survey agree to participate:
        - The parent/grandparent is asked for the names of all children in their household from the appropriate age groups.
        - The parent/grandparent must provide the name of the vaccination provider(s) for the children and permission to contact the provider(s)
        - A survey is sent to the vaccination provider(s) to collect administrative data regarding the vaccinations given.
        - NORC determines estimates of vaccination coverage data.
- The dataset includes:
    - Data indicators for state and year.
    - Demographic data for each child, limited demographic information for the child's household.
    - Information regarding the type and number of vaccines provided.
- Limitations:
    - The data is for only one year – no projection data is possible.
    - There is no indication whether the vaccinated children contracted the flu, were hospitalized or died from the flu – we cannot determine efficacy.
    - There is a tremendous amount of data missing from the dataset.
    - The survey data is entered manually and transmitted electronically.
        - There are multiple opportunities for error and/or omission.
    - There are no controls in place to help ensure completeness of data from the reporting entities.
- Relevance of these datasets to our project:
    - Although the source of the dataset is trustworthy, the dataset does not provide reliable data for our purposes.
    - The data is not sufficient to provide insights regarding efficacy of vaccinations or the ability to project year over year trends.
    - The data is insufficient to evaluate our hypothesis regarding vaccination of children.
    - If more children are vaccinated, fewer children will become infected.

Conclusion for Children Flu Shots:

- This dataset should not be included in our project.