Population Data by Geography

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

This is population data collected by the US Census Bureau on an annual basis. It is external data and can be considered trustworthy as the Bureau is a government agency, and there is no profitable gain from not providing accurate information.

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 current dataset with only the age group breakdown is administrative data, but the way the US Census Bureau collects its other data is a mix between administrative and survey data.[1] The method of collection is a combination of manual and automated practices; "*the data collection for the state and local finance survey is comprised of three modes to obtain data: mail canvass, Internet collection, and central collection from state sources. Collection methods vary by state and type of government.*"[1] There is a time lag with the population data as it is only collected on an annual basis.

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

The dataset contains the county and respective state, the year the data was collected in, the population of that respective year, biological sex information, and age group categories in five-year increments up until 85 where it is just "85 years and over."

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 dataset is dependent on a variety of methods for its collection, so these are as accurate as possible, but are still an estimate of reality. The dataset of course does not include other variables that we would need to make recommendations for our project such as vaccination and influenza rates of these people – it is simply a breakdown of the age group of the people in each respective county, nothing more, nothing less; thus, it needs to be used in conjunction with other datasets.

The dataset should not be bias as it is simply a general demographic breakdown; however, the collection method may be considered biased, which in turn make the data itself subsequently bias. It is not so much the case of the data being collected infrequently as census data is collected on an annual or "regular" basis, but with a survey-style collection method, the Bureau is dependent on people responding, which not only causes a lag, but also manual errors. The other concern is people, such as undocumented immigrants, may be hesitant in filling out census surveys for fear of retaliation or deportation. While they contribute to society and of course live among us, they may not be accounted for in the census due to these fears.

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

The project objective is to analyze and understand the influenza trend data enough to provide effective staffing recommendations. The hypothesis for this project is that if the state has a large

[1] US Census Bureau. 2022. Source: https://www.census.gov/programs-surveys/gov-finances/technical-documentation/methodology/how-the-data-are-collected.html

[2] Centers for Disease Control and Prevention (CDC). 2022. Source: https://www.cdc.gov/flu/weekly/overview.htm

population under the age of 5 and over the age of 65 (vulnerable population), then the influenza rates are likely higher; subsequently, these states will likely need additional staffing to combat the higher risk.

This dataset is relevant to our project. We cannot begin to get a picture of what each state is like and what they may need without first understanding the basic information: how many people are estimated to live in that state, and of those, how many fall into the vulnerable age population. We definitely need the census / population data to perform our analysis and calculations.

Influenza Laboratory Tests and Patient Visits Data Sets

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

One of the datasets collected data on when patients visited a participating facility, and the other collected data on when lab tests were conducted to test whether the patient did in fact show influenza-like illness (ILI). These two datasets were collected through collaborative effort between the CDC and a variety of partners throughout the country, including state, local, and territorial health departments, public health and clinical laboratories, vital statistics offices, health care providers, hospitals, clinics, emergency departments, and long-term care facilities.[2] This data should be deemed trustworthy as there are many trustworthy entities participating and it is being managed / owned by the Centers for Disease Control and Prevention (CDC), a historically trustworthy organization not aiming for profit. Both are considered external data.

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 is survey data that is being manually collected on a weekly basis by participating facilities contributing to the database, so technically there is not a lag, but the CDC website does point out that "*the data presented each week are preliminary and may change as more data are received.*"[1] It is important to include when the data was extracted when providing analytical findings as it may be different from another report due to when the respective datasets were extracted.

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

Both datasets include the region type, region, year, and week of when the data was collected. The influenza visit data includes age group of the patients as well as total number of participating providers and patients. The lab test data includes how many specimens were collected and if they tested positive for one of the ILI.

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 datasets do not include population data for each respective year thus it needs to be used in conjunction with the US Census (Population) dataset; the datasets also do not include biological sex information, the type of hospital or medical clinic the patient visited, and while the influenza visit dataset has age categories, it is unclear what they represent. The cell in these columns has a "X" in them, which is unclear whether that means the data is crossed out or included in the

[1] US Census Bureau. 2022. Source: https://www.census.gov/programs-surveys/gov-finances/technical-documentation/methodology/how-the-data-are-collected.html

[2] Centers for Disease Control and Prevention (CDC). 2022. Source: https://www.cdc.gov/flu/weekly/overview.htm

total patient calculation; these columns do not indicate how many of the total patients fall into these categories. It is also unclear what the "% UN/WEIGHTED ILI" mean.

The data in these datasets could potentially be bias because the participating facilities can choose not to provide information. Facilities may decide not to provide information as it may look unfavorably towards them (e.g., high influenza rates, low vaccination rates).

These are collected on a weekly basis by different participating entities.

As it is being manually inputted by different participating entities, there is bound to be clerical errors; however, it should hopefully be miniscule enough to not affect the overall numbers.

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

The project objective is to analyze and understand the influenza trend data enough to provide effective staffing recommendations. The hypothesis for this project is that if the state has a large population under the age of 5 and over the age of 65 (vulnerable population), then the influenza rates are likely higher; subsequently, these states will likely need additional staffing to combat the higher risk.

While the dataset does have age category columns, they don't exactly provide much insight as there is no indication of how many are in each respective category. It is nice to know how many people are visiting the participating facilities to get a quick picture of whether people in the state trusts medical facilities enough to visit and likely receive the vaccine. This dataset would be relevant to this project, though the most useful thing in this dataset would probably be the TOTAL PATIENTS column to be used in conjunction with the population data.

Children Flu Shots Data Set

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

This is data on children receiving their flu shots. It is external data collected and owned by the CDC. It should be deemed trustworthy as the CDC has been a historically trusted source.

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 is survey data collected manually by phone calls to monitor vaccination coverage of children. There is a time lag for this as it takes quite awhile to make that many calls to collect the desired data.

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

The dataset contains a variety of data including the age of the child, the demographic of the parents, the education and socio-economic status of the family, the state of residence, whether the child is/was breastfed, the type of insurance the family has, and of course the vaccination information.

[1] US Census Bureau. 2022. Source: https://www.census.gov/programs-surveys/gov-finances/technical-documentation/methodology/how-the-data-are-collected.html

[2] Centers for Disease Control and Prevention (CDC). 2022. Source: https://www.cdc.gov/flu/weekly/overview.htm

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 most notable limitations are the method of collecting the data and the timeframe of this dataset – there is only one year! Regarding the first limitation: phone surveys take a lot of manpower and hours to conduct that may not provide the best return on investment. The recipients may not pick up a call from an unknown caller, and even if they did, many may not want to invest the time divulging so much information about themselves. The data may be bias or limited as people may not be comfortable divulging that they did not go to college or that they are living in poverty, or they may lie to sound better.

This data is not so much collected infrequently as is inherently infrequently collected. Even if phone surveyors consistently make ten calls in an hour, not all ten recipients are going to pick up their phone or even provide valid responses. Some recipients may talk very little, while others talk more, creating a delay for the next call. And as these responses are being manually entered by the surveyors, there are likely issues simply due to human errors.

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

The project objective is to analyze and understand the influenza trend data enough to provide effective staffing recommendations. The hypothesis for this project is that if the state has a large population under the age of 5 and over the age of 65 (vulnerable population), then the influenza rates are likely higher; subsequently, these states will likely need additional staffing to combat the higher risk.

As this hypothesis is a two-parter, focusing on both vulnerable populations of those under the age of 5 and those above the age of 65, this dataset only provides data for half of the desired population. Even so, this data set is relevant to this project. It is definitely beneficial to know the age, race, and biological sex of the child being vaccinated along with the state of residence as those are basic components of this analysis. The additional information on the socio-economic status of the family, education level, and insurance coverage provide context on what type of family is more likely to be (or can be) vaccinated in each respective state.

[1] US Census Bureau. 2022. Source: https://www.census.gov/programs-surveys/gov-finances/technical-documentation/methodology/how-the-data-are-collected.html

[2] Centers for Disease Control and Prevention (CDC). 2022. Source: https://www.cdc.gov/flu/weekly/overview.htm