Project Overview

- **Motivation:** The United States has an influenza season where more people than usual suffer from the flu. Some people, particularly those in vulnerable populations, develop serious complications and end up in the hospital. Hospitals and clinics need additional staff to adequately treat these extra patients. The medical staffing agency provides this temporary staff.
- **Objective:** Determine when to send staff, and how many, to each state.
- **Scope:** The agency covers all hospitals in each of the 50 states of the United States, and the project will plan for the upcoming influenza season.

Hypothesis

If a state has a larger 65+ population percentage, then it will have more flu deaths (adjusted for each state's total population).

Data Overview

Influenza Deaths

This CDC-sourced dataset details each month the flu deaths that occur in a state, broken down into 10-year age groups.

Population Data by Geography

This dataset is from the US Census Bureau, and it contains annual population data for each state's counties. This data is broken down by gender and 5-year age groups.

Data Limitations

Influenza Deaths

There are two data completeness concerns. One, the data is sourced from death certificates, which only list one cause of death. It can be difficult to nail down the true cause of death for a patient with multiple ailments. Two, if an entry (determined by state, month/year, and age group) had less than 10 deaths, then the data was "suppressed". Suppressed entries were discarded, so there are deaths that are unaccounted for.

Population Data by Geography

The lag presents a limitation, as the latest data will not match the current year. However, it shouldn't be much more than a year behind. There is a note that there are estimates involved, too, as the response rate of the survey is 99.98%.

Descriptive Analysis

Variable	Mean	Standard Deviation
65+ population percentage	14%	1.7%
Flu Deaths (adjusted)	13	5.8

The 65+ population percentage had a weak correlation (0.14) with flu deaths (adjusted).

Results and Insights

While there is no correlation between 65+ population percentage and flu deaths (adjusted), the mean percentage of 65+ year olds in a state with above-median flu deaths (adjusted) is greater than the percentage of 65+ year olds in a state with below-median flu deaths (adjusted).

	Above Median Deaths	*Below Median Deaths*
Mean 65+ Population Percentage	14.1%	13.6%

Remaining Analysis and Next Steps

Analysis techniques implementing visualizations remain, continuing to draw insights from the data. This includes pie charts, column charts, treemaps, time forecasts, distribution visualizations, and correlation visualisations.

Expect a video presentation with final deliverables in 5 weeks.

Appendix A: Business Requirements Document

https://images.careerfoundry.com/public/courses/data-immersion/A1-A2_Influenza_Project/A1-A2%20data%20immersion_project%20brief%20.pdf

Appendix B: Hypothesis Development

Our first hypothesis stated, “if a state has a larger vulnerable population, then it will more influenza deaths.” This evolved to the present hypothesis for two main reasons. First, our census data only provided us with age and gender information. So, we could only base it off of age, and 65+ population was the only population that had non-suppressed deaths; larger vulnerable population changed to 65+ population. Then, we normalized the variables in the hypothesis by taking the state’s total population into account; 65+ population became 65+ population percentage, and we added the parenthetical “adjusted for total population” to influenza deaths. The adjustment for influenza deaths that we made is as follows: we took the flu deaths each year for a state, divided it by the 100,000 population total in the state that year, detailed in the equation below.

$$Flu\ Deaths\ (adjusted) = \frac{flu\ deaths}{total\ population \big/ 100{,}000}$$

Appendix C: Data Profiles

Population data by geography, US Census Data

Data Source

The Census Bureau combines data from several sources, both internal and external. It is a trustworthy source of data, as it is government data.

Data Collection

The combined data that the Census Bureau sources is administrative, from other business, and internal survey data. Surveys are automatically processed and translated into ASCII. As there is a lot of data to process in the surveys, it can take over a year to release population counts.

Contents Overview

Each observation has a county, a year, and population figures for that county during that year: the total population, the male population, the female population, and eighteen age-range populations (e.g., 25 to 29 years of age).

Limitations

The lag presents a limitation, as the latest data will not match the current year. However, it shouldn’t be much more than a year behind. There is a note that there are estimates involved, too, as the response rate of the survey is 99.98%.

Relevancy

This data set will be useful for determining each state’s vulnerable population count.

Counts of influenza laboratory test results by state

Data Source

The data comes from health providers external to the CDC. The CDC is also a government entity, so the data source is trustworthy.

Data Collection

The data is survey data, as it is collected through voluntary reporting to the CDC by public health organizations and healthcare providers. There is little lag time, as reporting is updated weekly. Any past inaccuracies are reconciled in future reports. I could not find this said explicitly, but it's safe to assume that the report compilation is automated to a larger degree than manually. I make this assumption because of the quick turnaround of data that comes from many sources.

Contents Overview

Influenza Visits

Influenza Visits tracks patient visits to a medical provider for influenza. It counts the number of visits, number of providers, and total patients seen by week and state from late 2010 to early 2019. This reporting comes from 3,500 outpatient healthcare providers.

Lab Tests

Lab Tests counts the number of positive influenza laboratory tests by week and state from late 2010 to early 2015. This reporting comes from 100 public health providers and over 300 clinical laboratories located throughout the United States and its territories.

Limitations

The most recent data is "preliminary and may change as more data is received". Also, the outside reporting is voluntary, so it is not necessarily comprehensive.

Relevancy

While these two datasets do not apply directly to the hypothesis, which examines the relationship between vulnerable populations and flu deaths, the datasets may be relevant later in the project. Flu visits and positive flu tests could be used in staffing estimates.

Survey of Flu Shot Rates in Children

Data Source

The data are internal to the CDC. As mentioned before, we can trust the credibility of this source.

Data Collection

As mentioned in the title, the data are collected by survey. It is a phone survey in which recipients may or may not grant permission for the CDC to contact their child(ren)'s vaccination provider(s). Then, the vaccination providers share administrative data. Lag time was not mentioned on the CDC site.

Contents Overview

The data contains flu shot data for children 6 months to 17 years. It's categorized by geographic state and contains family demographics including poverty level, race, and parent marital status.

Limitations

The surveys are used to come up with estimates, so the results are not exact.

Relevancy

This dataset does not apply to the hypothesis. Additionally, it is not complete enough to be relevant; it only includes vaccination estimates for children, which is not the only vulnerable population group.

Data Profile Sources (by order of reference)

https://www.census.gov/about/what/admin-data.html
https://www.census.gov/history/www/innovations/technology/tabulation_and_processing.html

https://www.cdc.gov/flu/weekly/overview.htm#Outpatient

https://www.cdc.gov/vaccines/imz-managers/nis/about.html

Appendix D: Statistical Analyses

H_0 For any given year, the percentage of 65+ year olds in a state with above-median influenza deaths (adjusted) is less than or equal to the percentage of 65+ year olds in a state with below-median influenza deaths (adjusted).

$$\mu_{\substack{above\ median\\ flu\ deaths}} \leq \mu_{\substack{below\ median\\ flu\ deaths}}$$

H_A For any given year, the percentage of 65+ year olds in a state with above-median influenza deaths (adjusted) is greater than the percentage of 65+ year olds in a state with below-median influenza deaths (adjusted).

$$\mu_{\substack{above\ median\\ flu\ deaths}} > \mu_{\substack{below\ median\\ flu\ deaths}}$$

Our t-test gave us grounds to reject the null hypothesis. Additionally, we obtained a one-tailed p-value below 0.005 with α=0.05, so there is a significant difference in the average percentage of 65+ year olds in states with above median flu deaths and below median flu deaths. In fact, there is only a 0.5% probability that sample mean difference is due to chance.