# Preparing for Influenza Season:

## Interim Report

Gregor Gurski 23 . 09 . 2023

# 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 people are 65 years old or older, then they are more likely to die from influenza.

# Data Overview

1) **Influenza death data**: This data set is an external data source. It is owned by the Centers for Disease Control and Prevention (CDC) through their National Center for Health Statistics. This data counts monthly influenza death rates divided by state, months, and age groups from 2009 to 2017.

2) **US Census Population data**: This data set is an external data source. It is owned by the US Census Bureau. The data contains population data for every county in the United States from 2009 to 2017 and is segregated by total population, male total population, female total population, and is also broken down into 18 age groups starting from under 5 years up to 85 years and above.

# Data Limitations

### Influenza death data:

Since the data was collected manually, it can be subject to human error during the data collection process. Because it is collected annually, there will be discrepancies with actual rates. The data set contains such missing values as suppressed deaths and unknown age group.

### US Census Population data:

Due to the decennial census, there is a large time gap, i.e., 10 years, between when data is collected, and counts can change for various reasons rapidly, such as births, deaths, change in living situations, etc. The data may also have some bias due to individuals unwilling or unable to participate in the US Census surveys. So, while the data is deemed trustworthy, it does have its limitations which are beyond the analyst's control.

# Descriptive Analysis

| Variable | Mean | Standard Deviation |
|---|---|---|
| under 65 population | 5278879 | 5973747 |
| over 65 population | 829430 | 892630 |
| under 65 deaths | 79 | 151 |
| over 65 deaths | 826 | 1014 |

The correlation coefficient of the variables between the over 65 population and over 65 deaths is 0.94. This is a strong positive relationship between the two variables. This correlation supports the hypothesis that as age increases the death rate also increases.

# Results & Insights

**Null Hypothesis**: People under 65 years have an influenza death rate that is the same or greater than the influenza death rate of people 65 years or older.

**Alternative Hypothesis**: People under 65 years have an influenza death rate that is smaller than the influenza death rate of people 65 years or older.

Due to the P-value (4.96E-45) being much smaller than the significance level (or alpha value, i.e., α=0.05), I can reject the null hypothesis. Therefore, I can say with a 95% confidence level, that people under 65 years have an influenza death rate that is smaller than the influenza death rate of people 65 years or older.

## Remaining Analysis & Next Steps

Given the above analysis, I can begin to approach the project goal of helping the staffing agency send medical workers in preparation for the influenza season. The analysis has shown a significant impact of age on influenza death rates and therefore the elderly population in different locations should be considered as an important criterium for where and by how much to staff up hospitals and clinics. Thus, further analysis should include:
1) Examine the US Census Population data again to determine what counties in each state have the highest 65+ populations to help distribute staff accordingly to avoid understaffing and overstaffing across states.
2) Further research influenza season to determine whether influenza occurs seasonally or throughout the entire year by utilizing the cdc.gov website.
3) Create a data visualization design checklist.
4) Create composition, comparison, temporal, forecasting, statistical, spatial, and textual visualizations.
5) Create a visual storyboard and a video presentation to stakeholders.

## Appendix

### Sources:

Project brief

Influenza deaths by geography

US Census Population Data

## Hypothesis Development:

### Clarifying questions:

1. How long is the time period of the flu?
2. How many staff are available and required on average across states?
3. Which states have the largest need for flu shots?

### Questions concerning privacy and ethics:

1. Are there privacy laws we need to adhere to related to collecting, storing, and analyzing data from influenza patients?
2. Is it ethical to analyze historical influenza deaths?
3. Likewise, is it ethical to categorize certain groups of people as vulnerable populations?

## Profiles of data:

US Census Population data: This is an external data source. The medical staffing agency does not have this information, so they're relying on government data. The data is provided by the US Census Bureau which is the federal governments largest statistical agency. As government data, I verify this as a trustworthy data source. The data collection method is a combination of administrative and survey data. As per https://en.wikipedia.org/wiki/United_States_Census_Bureau, "in addition to the decennial census, the Census Bureau continually conducts over 130 surveys and programs a year, including the American Community Survey, the U.S. Economic Census, and the Current Population Survey." Because this data is collected by the government and is mandatory by law, I assume that the data set is trustworthy.

Influenza death data: The data source is external as it comes from the government agency of the CDC. The CDC owns the data. Since this is run by the government this is trustworthy. This data is administrative data. Each US state is required to record all birth and death records. Death records come from the providers that code the cause of death as influenza. The deaths on a death certificate only list one cause of death. This could cause issues to the elderly and vulnerable population as the cause of death may be related to another health issue such as AIDs, diabetes, or asthma. The decline in health may have been initiated by

influenza. Only being able to list one cause of death on a death certificate can lead to bias from the providers. Nevertheless, since it is governmental data, I assume that the data set is trustworthy.

## Results and insights:

Results of the two-sample t-test with a one-tailed test:

| | Over 65 Deaths | Under 65 Deaths |
|---|---|---|
| Mean | 826.2875817 | 78.76470588 |
| Variance | 1028483.747 | 22903.91395 |
| Observations | 459 | 459 |
| Hypothesized Mean Difference | 0 | |
| df | 478 | |
| t Stat | 15.61886217 | |
| P(T<=t) one-tail | 4.96E-45 | |
| t Critical one-tail | 1.648047653 | |
| P(T<=t) two-tail | 9.91009E-45 | |
| t Critical two-tail | 1.964939272 | |