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

This analysis sought to determine if a state has a higher population of citizens considered vulnerable to complications from the flu than that state will have a greater rate of flu deaths.

Data Overview

Data Set: Influenza deaths by geography, time, age, and gender

Source: CDC

'Influenza deaths by geography, time, age and gender' is and external, open data source owned by the United States Centers for Disease Control and Prevention (CDC). This data set contains data collected by the National Center for Health Statistics (NCHS) and can be considered trustworthy.

The mortality data is administrative data based on information from all death certificates filed in the fifty states and the District of Columbia. Mortality information is collected manually by state registries and provided to the National Vital Statistics System. Underlying cause of death and demographic descriptors are indicated on the death certificates.

Data Set: Population data by geography

Source: US Census Bureau

'Population data by geography' is an administrative, external, and open data source owned by the United States Census Bureau. This data includes population estimates broken down by county, gender, and age groups. The data consists of population estimates based on survey results.

Data Limitations

Data Set: Influenza deaths by geography, time, age, and gender

Source: CDC

- A time lag exists in this data set because the most recent data is from 2017.
- A value of 'Suppressed' results when the total deaths for an age group are less than 10 and for all death totals for children under the age of 5 years old.

Data Set: Population data by geography

Source: US Census Bureau

- The data is survey data, so it represents a sample population of all influenza visits in the US.
- Surveys are filled out manually and a prone to human error.
- A time lag exists the data set begins in 2009 and ends in 2017 so it may not completely reflect current population demographics.

Descriptive Analysis

Data Spread	Normalized: Ratio of Flu Deaths/State Population					
	65-74 years		75-84 years		85+ years	
state_pop_flu_deaths						
This is population data						
This is a normal distribution						
Variance	2.3346E-07		8.89145E-07		3.08675E-06	
Standard Deviation	0.000483176		0.000942945		0.001756916	
Mean	0.000491378		0.000873402		0.00154758	
95% of the Data Falls Between:	-0.000474975	0.001457731	-0.001012489	0.002759292	-0.001966252	0.005061413
Outlier Count		25		19		18
% Outliers		**5.45%**		**4.14%**		**3.92%**

Data Spread	Census Data							
	Sum of Total population		Sum of 65-74 years		Sum of 75-84 years		Sum of 85+ years	
state_pop_flu_deaths								
This is population data								
This is a normal distribution								
Variance	4.495E+13		2.32029E+11		80065411023		15009546671	
Standard Deviation	6704475.52		481694.3911		282958.3203		122513.4551	
Mean	6040550.488		449100.4886		259013.3332		109199.8904	
95% of the Data Falls Between:	-7368400.553	19449501.53	-514288.2936	1412489.271	-306903.3074	824929.9738	-135827.0197	354226.8006
Outlier Count		23		28		28		27
% Outliers		**5.01%**		**6.10%**		**6.10%**		**5.88%**

Correlation		
Variables:	Vulnerable Population Ratio	Total Death Rate
Proposed Relationship:	As the vulnerable population ratio increases for each state the total death rate for that state due to flu will increase.	
Correlation Coefficient	0.08090467	
Strength of Correlation	Weak	
Usefulness / Interpretation	Yes, this correlation is still useful. It does not show that states with a larger ratio of more vulnerable populations generally have a higher death rate due to Flu. This data will be helpful in determining which variables have the most impact on a states need for support .	

Results & Insights

Null Hypothesis: States with a low proportion of vulnerable population less than or equal to 0.201725(median) will have a death rate greater than or equal to states with a proportion of vulnerable population greater than 0.201725(median).

Alternative Hypothesis: States with a proportion of vulnerable population greater than 0.201725(median) will have a death rate greater than states with a proportion of vulnerable population less than 0.201725(median).

The p-value of 0.340664001422113 is greater than the significance level of 0.05. This means we cannot reject the null hypothesis. At a confidence level of 95 percent, we could not

find a significant difference in Flu death rates between states with high proportion of vulnerable populations and states with low proportion of vulnerable populations.

Based on the findings of this statistical analysis we cannot assume that states with a higher proportion of the population being vulnerable to death by the flu will need more staffing assistance than states with a low proportion of the population being vulnerable to death by flu. I suggest we look at other variables and see how they impact flu deaths by state. It may be possible that due to such a large proportion of the flu deaths data being suppressed the results do not represent the true population. We may need to use different normalized variables to make any significant conclusions about staffing distribution. Also, we may need to look at a different research hypothesis altogether.

Remaining Analysis and Next Steps

This project seeks to determine when to send staff, and how many, to each state, the reason for the lack of statistical difference must be explored. The next steps will be:

- Holding meetings with project stakeholders and experts in the field.
    - In these meetings we should brainstorm whether new variables should be included in the analysis or if the data needs to be normalized for analysis. The data is cleaned integrated and ready for analysis so there is the option of forming a new research hypothesis and restarting the analysis process.
- Repeating the analysis with additional data
- Perform composition, statistical, spatial, and temporal visualizations.
- Publish analysis as a Tableau Storyboard.
- Record a video presentation for your stakeholders.

Note that any remaining work is on hold until all stakeholders have had a chance to discuss the results.

Appendix

Project Overview:

- See hyperlink Project Brief

Hypothesis Development:

- Which states are most affected by influenza?
    - What percent of the population gets the flu in each state?
    - Of the population contracting the flu in each state, what percent are hospitalized?
    - Which states have the highest mortality rates due to influenza?
- When is flu season?
    - When is flu season in each state?
    - What is the length of the flu season in each state?
    - Is there a single flu season in each state or are the multiple seasons in some states?
- Which states have the most residents in vulnerable populations?
    - Which demographic of vulnerable populations leads to more hospitalizations (i.e., children under 5 or adults over 65)?
    - How are vulnerable populations distributed throughout the states?
    - Does a state having a larger vulnerable population necessarily result in more hospitalizations in that state?
- What percent of the population is vaccinated in each state?
    - Do vaccination numbers have any correlation to hospitalizations due to flu in each state?
    - Does providing vaccinations create an increased need for staffing?
    - How does vaccine rationing differ between states?
- What are each state's current staff-to-patient ratios?
    - Are some states considered understaffed even outside of flu season?
    - Are there states that will not need additional staffing even during flu season due to a high existing staff-to-patient ratio?

- Do hospitals already have systems in place where in house support is share between hospitals during flu season? If so, is data available for this scenario?

- How many nurses, physician assistants, and doctors are currently on staff at the staffing agency?
  - Are staffing agency employees available to all hospitals or clinics unless currently deployed?
  - What is the staffing agencies employee's permanent residence distribution?
  - If staffing is insufficient to meet demands of clients do protocols currently exist for this situation?

Data Profiles:

Census Data Set:

Variables	Description	Data Types			
		time -variant/-invariant	structured/unstructured	qualitative/quantitative	qualitative: nominal/ordinal quantitative: discrete/continuous
State	Text string describing State where survey took place	Time-invariant	Structured	Qualitative	Nominal
Year	The year the survey took place	Time-invariant	Structured	Qualitative	Ordinal
Total population	Number value of total population of observation	Time-variant	Structured	Quantitative	Discrete
Male Total population	Number value of Male Total population of observation	Time-variant	Structured	Quantitative	Discrete
Female Total population	Number value of Female Total population of observation	Time-variant	Structured	Quantitative	Discrete
Under 5 years	Estimated value of county population uder 5 years old	Time-variant	Structured	Quantitative	Continuous
5 to 9 years	Estimated value of county population from 5 to 9 years old	Time-variant	Structured	Quantitative	Continuous
10 to 14 years	Estimated value of county population from 10-14 years old	Time-variant	Structured	Quantitative	Continuous
15 to 19 years	Estimated value of county population from 15-19 years old	Time-variant	Structured	Quantitative	Continuous
20 to 24 years	Estimated value of county population from 20-24 years old	Time-variant	Structured	Quantitative	Continuous
25 to 29 years	Estimated value of county population from 25-29 years old	Time-variant	Structured	Quantitative	Continuous
30 to 34 years	Estimated value of county population from 30-34 years old	Time-variant	Structured	Quantitative	Continuous
35 to 39 years	Estimated value of county population from 35-39 years old	Time-variant	Structured	Quantitative	Continuous
40 to 44 years	Estimated value of county population from 40-44 years old	Time-variant	Structured	Quantitative	Continuous
45 to 49 years	Estimated value of county population from 45-49 years old	Time-variant	Structured	Quantitative	Continuous
50 to 54 years	Estimated value of county population from 50-54 years old	Time-variant	Structured	Quantitative	Continuous
55 to 59 years	Estimated value of county population from 55-59 years old	Time-variant	Structured	Quantitative	Continuous
60 to 64 years	Estimated value of county population from 60-64 years old	Time-variant	Structured	Quantitative	Continuous
65 to 69 years	Estimated value of county population from 65-69 years old	Time-variant	Structured	Quantitative	Continuous
70 to 74 years	Estimated value of county population from 70-74 years old	Time-variant	Structured	Quantitative	Continuous
75 to 79 years	Estimated value of county population from 75-79 years old	Time-variant	Structured	Quantitative	Continuous
80 to 84 years	Estimated value of county population from 80-84 years old	Time-variant	Structured	Quantitative	Continuous
85 years and over	Estimated value of county population from 85+ years old	Time-variant	Structured	Quantitative	Continuous

Flu Data Set:

Variables	Description	Data Types			
		time -variant/-invariant	structured/unstructured	qualitative/quantitative	qualitative: nominal/ordinal quantitative: discrete/continuous
State	Full state name recorded as text string	Time-invariant	Structured	Qualitative	Nominal
Year	Year observation was recorded	Time-invariant	Structured	Qualitative	Ordinal
Month	Month and year observation was recorded	Time-invariant	Structured	Qualitative	Ordinal
Ten-Year Age Groups	10 year age intervals	Time-invariant	Structured	Qualitative	Ordinal
Deaths	Numerical value of recoded deaths greater than or equal to 10	Time-variant	Structured	Quantitative	Discrete

PREPARING FOR INFLUENZA SEASON: INTERIM REPORT

Results and Insights:

t-Test: Two-Sample Assuming Unequal Variances

	Death Rate of States with Low Vulnerable Population	*Death Rate of States with high Vulnerable Population*
Mean	0.000433763	0.000487265
Variance	4.52336E-07	2.69208E-07
Observations	229	230
Hypothesized Mean Difference	0	
df	428	
t Stat	-0.953917173	
P(T<=t) one-tail	0.170332001	
t Critical one-tail	1.648421601	
P(T<=t) two-tail	0.340664001	
t Critical two-tail	1.965522115	