UIJIN HWANG

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.

RESEARCH HYPOTHESIS

If states are with higher populations of vulnerable groups, then mortality rates will decrease.

Here vulnerable populations include children under 5 and adults over 65 years old, and this hypothesis suggests that states with higher levels of those populations are more likely to have higher death rates and would therefore need greater numbers of additional medical staff.

DATA OVERVIEW

1. US Census Data:

This data set includes populations of every US state, broken into 5-year age groups (which have been consolidated to focus on the two primary groups being researched), with data from 2009 - 2017. Additional US Census data detailing both urbanization percentage and population density for every US state in 2010 has been integrated as well.

2. CDC Influenza Deaths Data:

This data set includes influenza deaths reported by every US state, broken into 5 and 10-year age groups (which have been consolidated to focus on the two primary groups being researched), with monthly data from 2009 - 2017. In conjunction with the US Census Data, these deaths have been calculated as a percentage of the total population within their corresponding age groups.

DATA LIMITATIONS

All of our data sets are limited by the fact that the information provided is no more recent than 2017. The COVID-19 pandemic has since had major impacts both on the annual flu epidemics and on the demographic makeup of the US population, particularly in regards to our primary age group of concern, those aged 65 and older. In order to expect predicting deaths in 2022, these data sets need to be properly updated.

1. US Census Data:

This data set was missing a substantial number of observations for individual counties of various years, amounting to roughly 13% of all expected data points. Plus, it is *predictive* based on the decennial census and annual data received from states via the National Vital Statistics System, therefore it's not a 100% accurate reflection of the whole population at the given time but it is as close to an accurate population count as we can get.

2. CDC Influenza Deaths Data:

This data is part of the government's vital statistics program, and it can be considered a complete and accurate count of deaths, especially for those states which report flu deaths. However, in this data set, there are no reports of flu deaths for Alaska or for Washington, DC. In addition, for any month in which there were less than 10 deaths for a specific age group in a given state, that data was suppressed for privacy reasons. Therefore, any reporting of "zero deaths" may actually be a marginally small number rather than truly zero. This may be the reason Alaska has no reported deaths, but it does not likely account for Washington, DC.

DESCRIPTIVE ANALYSIS

There is a strong correlation found between vulnerable populations (children under 5 and adults over 65 years) and influenza death rates. This correlation means that states with higher number of vulnerable populations are more likely to have higher rates of influenza related deaths. Therefore, they more likely need additional medical staffing during the influenza season.

Description	Influenza death (0-4, 65+)	Population count (0-4, 65+)
Mean	1017	1,193,271
Standard deviation	972	1,325,775
Correlation coefficient	**0.96**	

In terms of outliers, 96% of influenza death and 93% of population count variables are within two standard deviations of the mean. Outliers are from New York, California, Texas, and Florida. Those are the top 4 states with with high populations and with high elderly populations. So they would not be removed from the dataset.

RESULTS AND INSIGHTS

A key insight of the analysis done so far is that those under 5 years old & aged 65+ suffer from more flu deaths than those between 5 - 64 years old.

A (two-tail) hypothesis was tested to find if deaths by influenza in the age group of under 5 & 65+ years old are equal (null hypothesis) or different (alternative hypothesis) from those in the age group of 5-64 years old. The p-value (8.37872365326141E-34) showed that there is less than 1% probability that the difference is due to chance and therefore the higher number of deaths by influenza in the age groups of under 5 & 65+ is statistically significant, with the confidence level of 95%, which is very significant.

REMAINING ANALYSIS AND NEXT STEPS

Based on the results of the analysis, the next steps would be to identify which states are with large vulnerable populations and ensure that to prioritize those in the distribution of medical staff. The project overview also asks us to understand whether the influenza season differs from state to state, which will require further analysis for us.

In the following analysis, we will conduct to investigate vaccine rates across states, as it has been shown that vaccination deters the impact of influenza. Therefore, states with lower vaccine rates would potentially need more medical staff for the patients.

Further deliverables will include a Tableau storyboard with temporal, statistical, and spatial visualizations. The final research will be presented with visualizations, conclusions, and recommendations.

APPENDIX

PROJECT GOAL:

The goal of this project is to help a medical staffing agency that provides temporary workers to clinics and hospitals on an as- needed basis. The analysis will help plan for influenza season, a time when additional staff are in high demand. The final results will examine trends in influenza and how they can be used to proactively plan for staffing needs across the country.

STAKEHOLDER IDENTIFICATION:

- Medical agency frontline staff (nurses, physician assistants, and doctors)
- Hospitals and clinics using the staffing agency's services
- Staffing agency administrators
- Influenza patients

SUCCESS FACTORS

The project's success will be based on:

- A staffing plan that utilizes all available agency staff per state requirements, without necessitating additional resources
- Minimal instances of understaffing and overstaffing across states (a state can be considered understaffed if the staff-to-patient ratio is lower than 90% of the required ratio and overstaffed if greater than 110%)

DATA SET SOURCES:

1. Influenza deaths by geography, time, age, and gender Source: CDC
2. Population data by geography Source: US Census
3. Counts of influenza laboratory test results by state (survey) Source: CDC (Fluview)

ADDITIONAL STATISTICAL HYPOTHESES TESTS PERFORMED:

t-Test: Two-Sample Assuming Unequal Variances		
	Vulnerable Group(Death Rate)	*Non-Vulnerable Group(Death Rate)*
Mean	1016.799564	416.6296296
Variance	944307.0209	14183.16384
Observations	459	459
Hypothesized Mean Difference	0	
df	472	
t Stat	13.13368947	
P(T<=t) one-tail	4.18936E-34	
t Critical one-tail	1.648088336	
P(T<=t) two-tail	8.38E-34	
t Critical two-tail	1.965002676	

Description	Vulnerable Group (0-4, 65+)	Non-Vulnerable Group (0-4, 65+)
Mean	0.00202161604	0.0013096402
Standard deviation	0.002217136	0.001383049
Correlation coefficient	**0.989597386**	

From this test, since the significance level is much greater than P-value (which is in scientific notification), we can say that we can reject the Null Hypothesis (Vulnerable group has less influenza death cases than or equal cases as non-vulnerable group) and accept the Alternative Hypothesis (Vulnerable group has more influenza death cases than non-vulnerable group). Thus, it confirms that vulnerable group (Under 5 years & 65+ years) has more influenza death cases than non-vulnerable group.

With confidence level of 95%, we found that the two groups are significantly different.