PREPARING FOR INFLUENZA SEASON

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 a state has a larger proportion of vulnerable population, then more deaths from flu will occur. NOTE: it will be considered vulnerable populations, residents 65 years and over.

Data Overview

The following data sets covering influenza in the United States will be used during the project:

1. Influenza deaths by geography

Source: CDC

The Underlying Cause of Death database contains mortality and population counts for all U.S. counties. Data are based on death certificates for U.S. residents.

2. Population data by geography, time, age, and gender

Source: US Census Bureau

Data Limitations

1. Influenza deaths by geography

The Deaths variable in the data set contains a large amount of *'Suppressed'* values as they do not present or publish death counts of 9 or fewer, therefore, exact numbers for that variable is unknown. Also, the deaths listed in this data set does not take into account the possibility of any preexisting health conditions that could be a critical factor causing death from influenza.

2. Population data by geography, time, age, and gender

Errors in censuses can arise from many sources such as flawed data a collection and processing procedures. At the outset it must be pointed out that errors are inevitable in a large data collection exercise such as a census, but we can conclude that since the data is meant to be informative, the likelihood of the data being biased is rare.

During the data cleaning and integration process of both data sets, the age groups were redefined as follows:

Under 5 years, 5-14 years, 15-24 years, 25-34 years, 35-44 years, 45-54 years, 55-64 years, 65-74 years, 75-84 years, 85+ years,

Subsequently those groups were summarised for statistical purposes as follows:

- Under 5 years to 64 years
- 65 years and over

Descriptive Analysis

The CDC & Census data sets were cleaned, merged, profiled, and transformed. Statistical analyses and hypothesis testing were conducted.

	Mortality rate Under 5 years to 64 years	Mortality rate 65 years and over
Mean	0,00549 %	0,47814 %

The table above shows the difference between the mortality rates of both age groups; residents 65 years and over clearly have a higher mortality rate than residents under 65 years old.

I proceeded by calculating the strength of the correlation between the variables. The following table presents a summary of the results:

Variables	"Total vulnerable population" and "average flu deaths per year"
Proposed Hypothesis	If a state has a larger proportion of vulnerable population, then more deaths from flu will occur.
Correlation coefficient	0.94 (94%)
Strength of the correlation	Strong

When the value of vulnerable population variable increases, the value of the number of deaths increases in a similar fashion. The oldest a person is, the higher risk of dead due to influenza tends to be. Age range and deaths variables have a strong positive correlation.

	Population Under 5 years to 64 years	**Population 65 years and over**
Mean	0,004781444	5,48636E-05
Variance	5,20614E-06	4,73835E-09
Observations	459	459
Hypothesized Mean Difference	0	
df	459	
t Stat	44,36063604	
P(T<=t) one-tail	**2,15E-168**	
t Critical one-tail	1,648180137	
P(T<=t) two-tail	4,31E-168	
t Critical two-tail	1,965145755	

Results & Insights

On the basis of the descriptive analysis, statistical hypothesis were created:

Null Hypothesis: Vulnerable populations have less or the same influenza death rates than non-vulnerable populations.

To refute the Null Hypothesis, the two mortality rates were compared in a "one-tailed t-test", the results throw that the p-value (2,15E-168) is lower than the significance level (α= 0.05) which means that the probability of the Null Hypothesis being significant or happening due to chance is less than 1%. The data provided enough proof to prove the Alternative Hyphotesis as statistically significant.

Alternative Hypothesis: Vulnerable populations have higher influenza death rates than non-vulnerable populations.

It can be concluded with 94% confidence that the two groups are significantly different and that the mortality rate of residents 65 years of age and older is significantly higher than the mortality rate of those under 65 years of age.

Remaining Analysis and Next Steps

- Further analysis of the states with vulnerable populations is recommended. For example, the analysis of the correlation between vaccinated and unvaccinated populations. However, in the case of the Labs test data set we have, since it is not broke down into age groups, it is not helpful for the purposes followed.
- In-depth research at state/county level is recommended to maximize staffing efficiency.
- We also need to keep in mind that young children are considered as vulnerable population, therefore their population distribution/death mortality rate should be part of the planning when making the medical staffing distribution.
- The final deliverable for the project will include a video presentation and a Tableau storyboard with statistical visualisations as well as results and recommendations.