PREPARING FOR INFLUENZA SEASON: INTERIM REPORT

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.

Our hypothesis:

" If the old age population (over the age of 60 years) in a state is high then the Mortality due to influenza is high"

DATA OVERVIEW:

The data sets we used were

1) Influenza deaths by Geography, Age , Time and Gender:
   Source: CDC Type: Administrative.
   - The data is administrative data which is collected by CDC across all states which measures the deaths due to influenza by geography, Age, Time and Gender.
2) Population data by Geography:
   Source: US Census Bureau. Type: Administrative.
   - The data is collected by the US census bureau from all the counties in the respective states once yearly from 2009 to 2017.

Data Limitations:

1) **Influenza Deaths by Geography, Age, time and Gender:**
   This dataset is a historical data from year 2009 to 2017. We could be missing current trend. The death numbers between 1-9 in most records are assigned as "suppressed". This prevents us from getting the exact number although insignificant.
2) **Population data by Geography:**
   This is also a historical data from the year 2009 to 2017. We are missing the current census record.
   The age groups in Influenza data set are 10 yr age groups, whereas the age groups in the population data are 5 yr.

Descriptive analysis:

	VARIABLE 1	VARIABLE 2
DATASET NAME: INTEGRATED DATASET	**POPULATION>60**	**DEATHS**
MEAN	6108125	905.05
STANDARD DEVIATION	6834131	1153.24
OUTLIER PERCENTAGE	4%	4.35%

Correlation results:

The correlation coefficient for population over 60 and deaths is 0.95.

This indicates a strong positive relationship between the variables: population over 60 and Deaths (due to Influenza). In other words, There is marked increase in deaths due to influenza with the increase in population over the age of 60 years.

Results and Insights:

Based on the descriptive analysis and correlation results we decided to test the deaths in the population below 60 yrs and population above 60 using statistical analysis. We therefore, made the following **statistical hypotheses**:

Null hypothesis: The number of deaths due to Influenza in younger population (<60 years) is equal to or greater than the number of deaths due to Influenza in older population (>60 Years).

Alternate hypothesis: The number of deaths due to Influenza in Older population (>60 yrs) is greater than the number of deaths due to influenza in Younger Population(<60 yrs)

Interpretation:

After the statistical analysis we came up with the following results:

The P-Value for one tailed 2 sample t-Test was for less than 0.005. Hence, we rejected the null hypothesis and accepted the alternate hypothesis i.e the number of deaths due to Influenza in Older Population (>60 years) is greater than the number of deaths due to Influenza in Younger population (<60 yrs).

Hence, it is clear from the analysis that Older population is at a higher risk of mortality due to influenza than the younger population.

Further analysis and next steps:

- We can further analyse what steps or interventions can help decrease the risk of fatality due to influenza
- The allocation of Healthcare workers can be done based on this observation i.e higher preference to states with higher old age population.
- Presentation of final results to the stakeholders using statistical, spatial and temporal visualizations.
- The stakeholders can have a brainstorming session to come up with allocation plans and their execution based on these insights.

APPENDIX

Project overview:

Requirements:

● Provide information to support a staffing plan, detailing what data can help inform the timing and spatial distribution of medical personnel throughout the United States.

● Determine whether influenza occurs seasonally or throughout the entire year. If seasonal, does it start and end at the same time (month) in every state?

● Prioritize states with large vulnerable populations. Consider categorizing each state as low-, medium-, or high-need based on its vulnerable population count.

● Assess data limitations that may prevent you from conducting your desired analyses

HYPOTHESIS:

"IF THE OLD AGE POPULATION (ABOVE THE AGE OF 60) IS HIGH IN A STATE THEN THE MORTALITY DUE TO INFLUENZA IN THAT STATE IS HIGH"

Questions we asked:

A) WHICH MONTHS OF THE YEAR IS THE FLU SEASON?

B) WHICH STATES HAVE THE HIGHEST POPULATION AT RISK?

C) HOW IS THE STAFF DISTRIBUTED AMONG THE STATES PRESENTLY?

D) WHICH STATES HAVE THE HIGHEST BURDEN OF FLU PATIENTS HISTORICALLY?

E) IS THE FLU SEASON THE SAME LENGTH EVERY YEAR?

F) IS THE SEVERITY OF FLU SEASON THE SAME ACROSS ALL STATES OR

G) ARE THERE ANY OTHER FACTORS THAT CONTRIBUTE TO THE SEVERITY LIKE AGE,SEX, WEATHER AND SO ON?

H) WITHIN THE FLU MONTHS, DOES THE SEVERITY OR INFECTION RATE VARY? OR FROM AMONG THE FLU MONTHS WHICH ARE THE PEAK MONTHS AND DO THEY VARY FROM STATE TO STATE?

DATA REQUIREMENTS:

1)The data about population distribution by age in each state.

2)The data about deaths caused due to influenza by age in these states.

DATA OVERVIEW:

	VARIABLE 1	VARIABLE 2
DATASET NAME: INTEGRATED DATASET	POPULATION	DEATHS
SAMPLE OR POPULATION?	POPULATION	POPULATION
NORMAL DISTRIBUTION?	NORMAL	NORMAL
MEAN	6108125	905.05
VARIANCE	4.67E+13	1329960
STANDARD DEVIATION	6834131	1153.24
OUTLIER PERCENTAGE	4%	4.35%

VARIABLES	POPULATION & Deaths
PROPOSED RELATIONSHIP	positive relationship
CORRELATION COEFFICIENT	0.96

Variables	Population above 60 and Deaths
Proposed Relationship	Positive Relationship
Correlation Coefficient	0.95

t-Test: Two-Sample Assuming Unequal Variances

	<65 YEARS	>65 Years
Mean	78.81659389	826.5633188
Variance	22952.79342	1030699.284
Observations	458	458
Hypothesized Mean Difference	0	
df	477	
t Stat	-15.58973183	
P(T<=t) one-tail	6.96013E-45	
t Critical one-tail	1.648054362	
P(T<=t) two-tail	1.39203E-44	
t Critical two-tail	1.964949728	

----------END----------