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.

Hypothesis

If there is a greater population of people of higher ages in a state, then the state will have a higher rate of mortality due to influenza.

Data Overview

Census Population Data: This data shows the numbers for total population, male total population, female total population and populations of each age group (groups of 5 years) by county and year for each county in the United States. The data is from the US Census Bureau from 2009-2017.

Influenza Deaths Data: This data shows the number of deaths due to influenza in each age group and state by month and year. The data is from the CDC from 2009-2017.

Influenza Visits Data: This data shows the total number of patients that were seen by medical providers, number of providers that saw the patients, and the percentage of visits by region, year, and week. The data is from the CDC from 2010-2019.

Data Limitations

Census Population Data: The census data is collected once every 10 years, and due to this there is a time lag. Another limitation of the data set is that the numbers are estimates which means that it is possible that some information may be incorrect.

Influenza Deaths Data: The limitation of the data set is that it includes information for people who passed away with influenza, even though influenza may not have been the only or primary cause of their death.

Influenza Visits Data: The limitation of the data set is that the data was obtained through surveys (manual data collection) and as a result, it may have errors. Since the data is obtained through

surveys, all clinics and providers may not have completed the surveys, and due to this, the data does not show the counts for influenza patients visits from every provider in the United States.

Descriptive Analysis

Variable	Mean	Standard Deviation
< 5-64 years Deaths	79	151
65-85+ years Deaths	826	1014
< 5-64 years Population	5278879	5973747
65-85+ years Population	829430	892630
Week 40 Total Patients	14858	16822
Total Deaths	905	1154

The 65-85+ years population has a strong correlation with the number of total deaths with a correlation coefficient of 0.95. This means that the total deaths due to influenza increase when the number of people in the 65-85+ years population group increases. This supports the hypothesis by showing that having higher populations of people of higher ages causes more deaths due to influenza.

Results and Insights

Null Hypothesis: A higher population of people of greater ages in a state leads to a lower rate of mortality due to influenza.

Alternative Hypothesis: A higher population of people of greater ages in a state leads to a higher rate of mortality due to influenza.

The p-value obtained from the statistical test is 3.43445984603435E-64 which is lower than the alpha or significance level of 0.05. Since the p-value is lower than the significance level, the null hypothesis has been rejected. With a confidence level of 95%, a higher population of people of greater ages in a state leads to a higher rate of mortality due to influenza. This means that the research hypothesis is correct, and it is acceptable to proceed with using it to create a plan for how many staff members the medical staffing agency should send to each state.

Remaining Analysis and Next Steps

Remaining analysis:

- Creating composition and comparison charts in Tableau by generating a treemap and a bar, column, or pie chart
- Creating temporal visualizations in Tableau by generating a time forecast
- Creating statistical visualizations such as histograms, box plots, scatterplots, and bubble charts

- Conducting spatial analysis by mapping a variable and creating a spatial visualization
- Conducting textual analysis by using qualitative data to generate a word cloud

The next steps are:

- Publishing a Storyboard in Tableau about the analysis
- Presenting results obtained from the analysis to stakeholders

Appendix

Business Requirements / Project Management Plan

- Stakeholder Communication:
    - Meetings:
        - Prior to the start of data analysis, hold a meeting with the relevant stakeholders which are the representatives of the clinics and hospitals that work with the staffing agency, administrators of the staffing agency, and a few representatives of the frontline staff from the medical agency. Ask clarifying, funneling, and privacy and ethics-related questions. Discuss business requirements with stakeholders. Inform the stakeholders about the plan for communication to them throughout the project.
        - Midway through the analysis, hold a meeting to discuss the project.
        - After the project is completed, hold a meeting to discuss if an adequate staffing plan was developed and if the success factors were met.
    - Calls:
        - While analysis is taking place, hold weekly calls to provide status updates to the stakeholders, allowing questions to be asked by the analyst and stakeholders about the project.
    - Written Communication:
        - Send emails to all stakeholders monthly to provide status updates on the project.
    - Emergency Plan:
        - Send emails immediately to all stakeholders for any urgent issues. Within three days, hold a follow-up call to discuss the urgent issue.
- Schedule and Milestones:
    - Weeks 1-2: Review the business requirements and project objective, motivation, and scope. Generate a list of questions.
    - Week 3: Formulate a hypothesis and design a data research project.
    - Weeks 4-7: Source data, create data profiles, check data integrity, and put data quality measures in place.
    - Week 8: Transform and integrate data.
    - Weeks 9-11: Perform statistical analysis on the data.
    - Week 12: Formulate a statistical hypothesis and perform hypothesis testing.
    - Week 13: Consolidate analytical findings in an interim report.
    - Week 14: Create a data visualization plan and checklist.

    - Weeks 15-18: Create visualizations in Tableau.
    - Weeks 19-20: Perform spatial and textual analysis.
    - Week 21: Create a storyboard in Tableau about the insights and findings.
    - Week 22: Create a video to present project findings.
- Project Deliverables:
    - Interim report on insights obtained from the analysis.
    - Storyboard in Tableau about analysis.
    - Video presentation on insights obtained for stakeholders.
- Audience Definition:
    - Representatives from clinics and hospitals:
        - Can assume that they are familiar with jargon and have a higher level of proficiency with data.
    - Administrators of staffing agency:
        - Can assume that they are familiar with jargon and have a higher level of proficiency with data.
    - Representatives of frontline staff from staffing agency:
        - Can assume that they are not as familiar with jargon and have a lower level of proficiency with data.

Hypothesis Development

Key Questions:

- When are new flu vaccines for the flu season made publicly available?
- Which states have the lowest flu shot administration rates?
- How many doctors, nurses, and physician assistants does the staffing agency have?

Hypotheses Developed:

1. If there is a greater population of people of higher ages in a state, then the state will have a higher rate of mortality due to influenza.
2. If there is more poverty in a state, then the state will have a higher number of influenza cases and in turn, a higher rate of mortality due to influenza.
3. If the population density is higher in a state, then there will be more influenza cases in the state.

Data Overview

Census Population Dataset:

Variables	**Time-variant / -invariant**	**Structured / Unstructured**	**Qualitative / Quantitative**	**Qualitative: Nominal / Ordinal Quantitative: Discrete / Continuous**
County	Time-invariant	Structured	Qualitative	Nominal
State	Time-invariant	Structured	Qualitative	Nominal
Year	Time-invariant	Structured	Qualitative	Ordinal
Total Population	Time-variant	Structured	Quantitative	Discrete
Male Total Population	Time-variant	Structured	Quantitative	Discrete
Female Total Population	Time-variant	Structured	Quantitative	Discrete
Age Groups	Time-variant	Structured	Quantitative	Discrete

Influenza Deaths Dataset:

Variables	**Time-variant / -invariant**	**Structured / Unstructured**	**Qualitative / Quantitative**	**Qualitative: Nominal / Ordinal Quantitative: Discrete / Continuous**
State	Time-invariant	Structured	Qualitative	Nominal
State Code	Time-invariant	Structured	Qualitative	Ordinal
Year	Time-invariant	Structured	Qualitative	Ordinal
Month	Time-invariant	Structured	Qualitative	Ordinal
Month Code	Time-invariant	Structured	Qualitative	Ordinal
Ten-Year Age Groups	Time-invariant	Structured	Qualitative	Ordinal
Ten-Year Age Groups Code	Time-invariant	Structured	Qualitative	Ordinal
Deaths	Time-variant	Structured	Quantitative	Discrete

Influenza Patient Visits Dataset:

Variables	**Time-variant/-invariant**	**Structured/Unstructured**	**Qualitative/Quantitative**	**Qualitative: Nominal/Ordinal Quantitative: Discrete/Continuous**
Region Type	Time-invariant	Structured	Qualitative	Nominal
Region	Time-invariant	Structured	Qualitative	Nominal
Year	Time-invariant	Structured	Qualitative	Ordinal
Week	Time-invariant	Structured	Qualitative	Ordinal
% Weighted ILI	Time-invariant	Structured	Quantitative	Discrete
% Unweighted ILI	Time-invariant	Structured	Quantitative	Discrete
Age 0-4	Time-invariant	Structured	Qualitative	Nominal
Age 25-49	Time-invariant	Structured	Qualitative	Nominal
Age 25-64	Time-invariant	Structured	Qualitative	Nominal
Age 5-24	Time-invariant	Structured	Qualitative	Nominal
Age 50-64	Time-invariant	Structured	Qualitative	Nominal
Age 65	Time-invariant	Structured	Qualitative	Nominal
ILI Total	Time-variant	Structured	Quantitative	Discrete
Num of Providers	Time-variant	Structured	Quantitative	Discrete
Total Patients	Time-variant	Structured	Quantitative	Discrete