Interim Report: Preparing for Influenza Season

Goal

To help a medical staffing agency that provides temporary workers to clinics and hospitals as needed. The analysis will help plan for influenza season when additional staff are in high demand. The final results will examine trends in influenza and how they can be used to proactively prepare for staffing needs across the country.

Business Requirements

As an analyst, you need technical skills to analyze your data and soft skills to communicate your insights to stakeholders. You'll start by distilling business requirements and requests into questions you can answer with an analysis. You'll follow up by sourcing and curating the data to address these questions. After analyzing the data and drawing conclusions or formulating recommendations from your results, you'll present your insights to stakeholders in an easily consumable format. You'll find the requirements for your project below. These requirements are what should guide your approach to the analysis. While this project will use data from healthcare, the steps and framework involved can be used for projects in any domain.

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 severe 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 you're older than 65, you're more likely to develop severe complications from the flu.

Data Overview

Influenza deaths by geography:

This data set shows how many deaths occurred in each state of America. It also consists of variables such as (State, State Code, Year, Month, Month Code, Ten-Year Age Groups, Ten-Year Age Groups Code, and Deaths.

Population data by geography, time, age, and gender:

This data set shows the total population across different counties in America. It also consists of variables such as County, Year, Total Population, Male Total Population, Female Total Population, Under 5 years old, 5 to 9 years, 10 to 14 years, 15 to 19 years, 20 to 24 years, 25 to 29 years, 30 to 34 years, 35 to 39 years, 40 to 44 years, 45 to 49 years, 50 to 54 years, 55 to 59 years, 60 to 64 years, 65 to 69 years, 70 to 74 years, 75 to 79 years, 80 to 84 years, 85 years and over.

Data Limitations

Influenza deaths by geography:

- The data contains mainly death counts for influenza-related deaths in the United States from 2009 to 2017.

Population data by geography, time, age, and gender:

- The US Census Bureau collects data every ten years, so there could be a time lag regarding residence changes, births, deaths, etc.

Descriptive Analysis

	Variance	*Standard Deviation*	*Mean*	*Outlier Lower Bound*	*Outlier Upper Bound*	*Outlier Count*	*Outlier Percentage*
Deaths <64 years	15,786	126	491	240	743	177	39%
Deaths >65 years	951,946	976	890	-1,061	2,841	36	8%
Total Deaths	1,187,722	1,090	1,381	-798	3,561	459	100%

Population <64 years	41,349,384,137,611	6,430,349	5,578,652	-7,282,046	18,439,349	0	0%
Population >65 years	785,085,339,853	785,085,339,853	886,050	806,989	-965,112	30	68%
Total Population	46,231,461,355,325	6,799,372	5,973,849	-7,624,896	19,572,593	0	0%

Correlations Evaluated

- For my first variable, I tested the correlation between Influenza Deaths under 64 years old and Population under 64 years old. The results showed a strong relationship between these variables, as I got a 0.91 Correlation Coefficient. This indicates that there is a strong chance that someone under the age of 64 will die from the flu.
- For my second variable, I tested the correlation between Influenza Deaths over 65 years old and Population over 65 years old. The results showed a strong relationship between these variables, as I got a 0.94 Correlation Coefficient. This indicates that there is a strong chance that someone over the age of 65 will die from the flu.

Results & Insights

Statistical Hypothesis

- **Null Hypothesis:** People younger than 65 will have a higher or equal death rate from the Flu as those older than 65.
- **Alternative Hypothesis:** People older than 65 will have a higher death rate from the Flu than those younger than 65.

T-Test Results

t-Test: Two-Sample Assuming Unequal Variances

	Death % <64 years	*Death % > 65 years*
Mean	0.000224318	0.001274978
Variance	5.21775E-08	2.32539E-07
Observations	459	459

Hypothesized Mean Difference	0
df	654
t Stat	-42.18542223
P(T<=t) one-tail	4.47E-189
t Critical one-tail	1.647186873
P(T<=t) two-tail	8.9447E-189
t Critical two-tail	1.963597919

- I conducted a one-tailed test because we are interested in seeing if people older than 65 will have a higher death rate from the Flu than those younger than 65.
- P-Value: 4.4724E-189
- Because the p-value (4.47236223349919E-189) is smaller than the significance level (0.05), this means I have rejected my null hypothesis.
- At an alpha of 0.05, or a confidence level of 95 percent, we have proven that individuals older than 65 years old will have a higher chance of dying from influenza than people under 65 years old.

Remaining Analysis and Next Steps

Achievement 2: Data Visualization & Storytelling

Throughout this Achievement, you'll continue to draw insights from your data using data visualization techniques. You'll also learn how to apply design principles to make effective visualizations, which you can then use as you develop a narrative for your stakeholders in the form of a presentation and Tableau storyboard.

Exercise 2.1: Intro to Data Visualization

- Explain how data visualizations can be used in your project.
- Install Tableau.

Exercise 2.2: Visual Design Basics & Tableau

- Create a data visualization design checklist.
- Explain how the visualizations in a given example can be improved.
- Connect your project data to Tableau.

Exercise 2.3: Composition & Comparison Charts

- Create a pie, bar, or column chart, as well as a treemap in Tableau.
- Use your visualization design checklist to design your charts.

Exercise 2.4: Temporal Visualizations & Forecasting

- Create a time forecast for a variable and display it in Tableau.
- Use your visualization design checklist to design your chart.

Exercise 2.5: Statistical Visualizations: Histograms & Box Plots

- Create visualizations that look at the distribution of a variable.
- Use your visualization design checklist to design your charts.

Exercise 2.6: Statistical Visualizations: Scatter Plots & Bubble Charts

- Create visualizations that look at the correlation between variables.
- Use your visualization design checklist to design your chart.

Exercise 2.7: Spatial Analysis

- Map a variable and justify your spatial visualization choice (heat, density, or choropleth).
- Use your visualization design checklist to design your chart.

Exercise 2.8: Textual Analysis

- Create a word cloud using qualitative data.
- Use your visualization design checklist to design your chart.

Exercise 2.9: Storytelling with Data Presentations

- Create a narrative to communicate your research findings and insights in relation to your research goals.
- Publish your analysis as a Tableau Storyboard.

Exercise 2.10: Presenting Findings to Stakeholders

- Record a video presentation for your stakeholders.

Next Steps

I will continue the analysis steps and begin developing the project deliverables. I will continue to draw insights from the data from my data using data visualization techniques. This will include charts, spatial analysis, storytelling, and final results to stakeholders.

Appendix

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 be planned for the upcoming influenza season.

Hypothesis Development

Clarifying Questions: Preparing for Influenza Season

- Who is getting most affected by the flu?
- What symptoms do people notice when they start to feel ill?
- Which age range is the most vulnerable to getting the flu?
- Which gender has the most significant risk during influenza season?
- When do people in the United States start to get sick from the flu?
- How are the World Health Organization, CDC, and the government taking measures to stop the spread of the flu?

Funneling Questions: Preparing for Influenza Season

Clarifying Question: Who is getting most affected by the flu?

Funneling:

- Is the same demographic affected by the flu in each state?
- Are different demographics getting affected by the flu in different seasons?

Privacy and Ethics: Preparing for Influenza Season

- Are there privacy laws we must adhere to related to collecting, storing, and analyzing CDC data?
- Are there privacy laws we must adhere to related to collecting, storing, and analyzing data from the US Census Bureau?

Hypothesis

If you're older than 65, you're more likely to develop severe complications from the flu.

Data Overview

CDC Influenza Mortality Data

Data Sourcing*:* This is an external data source. The medical staffing agency doesn't have this

information, so it's relying on government data. The data is provided by the Centers for Disease Control and Prevention (CDC) through their National Center for Health Statistics. As government data, you can verify this as a trustworthy data source.

Data Collection: The data is administrative data collected as part of the National Vital Statistics Cooperative Program. Each of the U.S. states and territories is required to record all births, deaths, marriages, and divorces within their jurisdiction. Death records come from death certificates, in which a doctor codes the primary cause of death as "Influenza" or "Pneumonia" (ICD-10 codes J09-J18).

Data Limitations:

Data Contents: The data contains monthly death counts for influenza-related deaths in the United States from 2009 to 2017. Counts are broken into two categories: state and age.

Data Relevance: The data shows the geographic and monthly spread of influenza across the United States over multiple years. As it was collected via the government vital statistics program, you can assume that it's the most trustworthy and complete version of the data available.

US Census

Data Sourcing: This is an external and government data source, so it is trustworthy. The medical staffing agency doesn't have this information, so it relies on the US Census Bureau to provide it. Hence, the data here is provided by the US Census Bureau through an Excel file, which contains the county, year, total population, male population, female population, and all the age brackets.

Data Collection: The data is administrative data collected by the US Census Bureau. The US Census Bureau states on its website, www.census.gov, that "The Census Bureau collects data about the economy and the people living in the United States from many different sources. Some data is collected from respondents directly (including businesses) through our censuses and surveys. We also collect additional data from other sources. Primary sources for additional data are federal, state, and local governments, as well as some commercial entities. These types of data are generally called "administrative data," as the data is collected and maintained by government agencies.

In addition, the Census Bureau combines administrative data with survey and census data. The website states, "Some data are collected from respondents directly (including businesses) through our censuses and surveys. We also collect additional data from other sources. Moreover, the US Census Bureau automatically reuses data that already exists elsewhere, and

by linking it to census and survey data, the Census Bureau can conduct research that allows us to see the larger picture of the people and economy of the United States.

Data Limitations: Lastly, the US Census Bureau collects data every ten years, so there could be a time lag regarding residence changes, births, deaths, etc.

Data Contents: The data contains population data from 2009 through 2017. Also, it is broken out from Total Population, Male Total Population, Female Total Population, Under 5 Years, 5 to 9 years, 10 to 14 years, 15 to 19 years, 20 to 24 years, 25 to 29 years, 30 to 34 years, 35 to 39 years, 40 to 44 years, 45 to 49 years, 50 to 54 years, 55 to 59 years, 60 to 64 years, 65 to 69 years, 70 to 74 years, 75 to 79 years, 80 to 84 years, 85 years, and over.

Data Relevance: The data is relevant to my project objectives and hypotheses. During influenza season, this dataset will provide an in-depth analysis of the population around counties and help me determine which counties are most vulnerable. Moreover, it showcases the population for different age groups, which will help me dissect whether the flu is more contagious for people getting older or not.

Results & Insights

Statistical Hypothesis

- **Null Hypothesis:** People younger than 65 will have a higher or equal death rate from the Flu as those older than 65.
- **Alternative Hypothesis:** People older than 65 will have a higher death rate from the Flu than those younger than 65.

T-Test Results

t-Test: Two-Sample Assuming Unequal Variances

	Death % <64 years	*Death % > 65 years*
Mean	0.000224318	0.001274978
Variance	5.21775E-08	2.32539E-07
Observations	459	459
Hypothesized Mean Difference	0	
df	654	

t Stat	-42.18542223
P(T<=t) one-tail	4.47E-189
t Critical one-tail	1.647186873
P(T<=t) two-tail	8.9447E-189
t Critical two-tail	1.963597919

- I conducted a one-tailed test because we are interested in seeing if people older than 65 will have a higher death rate from the Flu than those younger than 65.
- P-Value: 4.4724E-189
- Because the p-value (4.47236223349919E-189) is smaller than the significance level (0.05), this means I have rejected my null hypothesis.
- At an alpha of 0.05, or a confidence level of 95 percent, we have proven that individuals older than 65 years old will have a higher chance of dying from influenza than people under 65 years old.