Preparing for Influenza Season 2020-2021: 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 a state has a higher vulnerable population, then they will have more influenza deaths (and therefore require more staff).

Data Overview

- Influenza Deaths Data Set: This data from the CDC shows the numbers of deaths by influenza in each state by month, with separate subcategories for gender and age groups.

- Population Data by Geography: This data from the U.S. Census Bureau shows the population of each county in the United States by year, with separate subcategories for gender and age groups.

Data Limitations

- Influenza Deaths Data Set: There is only one cause of death listed for each person, which could create discrepancies in vulnerable populations e.g. those with cancer, AIDS, etc. Also, the majority of the data for gender and age group is suppressed, providing no useful information.

- Population Data by Geography: This is survey data, collected manually from the American Community Survey (ACS). This survey is conducted every month using a randomly selected sample of about 3.5 million addresses. Although this survey is mandatory, there may be some people who avoided taking the survey for legal/privacy reasons (e.g. immigration status), which can result in bias.

Descriptive Analytics

	Vulnerable Population	% Vulnerable	% Female Vulnerable	% Male Vulnerable	Ratio of Deaths to Population	% Population <5 years
Mean	1219865	0.202149451804	0.215053083986	0.188848806079	0.000178116	0.0642753109
Standard Deviation	0.014426942	0.014426942	0.015570196	0.013802059	4.77971E-05	0.007388638

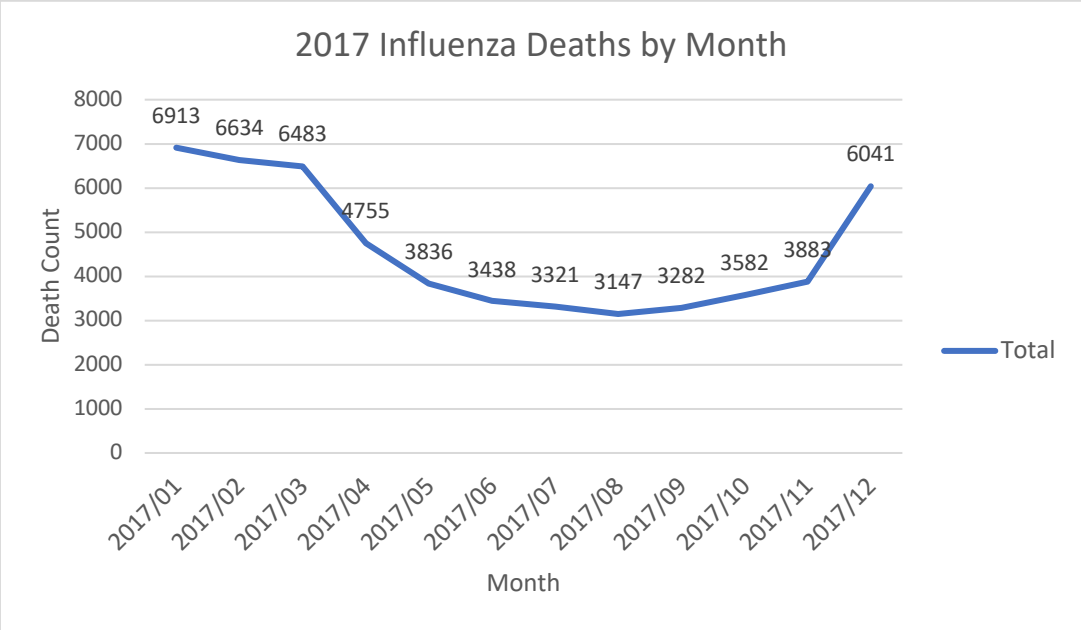
The percent of the population under 5 years old (part of the vulnerable population) had a weak correlation (-0.2) with the ratio of deaths to population, meaning that places with a higher percentage of children under 5 did not tend to have a higher percentage of those who died from the flu.

Results and Insights

Null hypothesis: The percentage of the male population that is vulnerable is equal to the percentage of female population that is vulnerable.

Alternative hypothesis: The percentages of vulnerable male and female populations are different.

Using a two-tailed t-test, we can say with 95% confidence that the differences in the means are not simply a chance occurrence. On average, there is a larger percentage of females that are a part of the vulnerable population in each state. (See Appendix for further details)



The highest death counts for 2017 occurred in January, February, March, and December.

Remaining Analysis and Next Steps

Remaining Analysis:

- I would like to know if the larger percentage of females that are vulnerable come from the adults over 65, the children under 5, or both. According to the Census, women have a longer life expectancy than men, so I would expect the larger vulnerable population to come from the pool of people 65 and older. This is easy to test with the available data.

- The next general step would be to see if there is a correlation between the death rate and the percentage of the vulnerable population, regardless of gender.

Next Steps:

- A meeting will be held with all stakeholders to discuss the forming insights, and to ask/answer questions, and to discuss the interim report.

- 2 weeks before the submission date of the final deliverable, a final meeting will be held with all stakeholders to discuss the data insights and the recommendations for the staffing plan. This will leave some time to make sure the recommendations make sense and are feasible for all parties involved.

- A final video presentation with visualizations will be delivered to all stakeholders discussing the insights and final recommendations for the staffing plan.

Appendix

Contents:

- Business Requirements Document

- Data Overview

- Additional Results/Insights

APPENDIX

BUSINESS REQUIREMENTS DOCUMENT

Goal
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.

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 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.

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

Success Factors
The project's success will be based on:
1. A staffing plan that utilizes all available agency staff per state requirements, without necessitating additional resources

2. 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%)

Assumptions & Constraints

Assumptions:

● Vulnerable populations suffer the most-severe impacts from the flu and are the most likely to end up in the hospital.
● Flu shots decrease the chance of becoming infected with the flu.

Constraints:

● The staffing agency has a limited number of nurses, physician assistants, and doctors on staff.
● There's no money to hire additional medical personnel.

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.

Glossary

Influenza: a contagious viral infection, often causing fever and aches.
Vulnerable populations: patients likely to develop flu complications requiring additional care, as identified by the Centers for Disease Control and Prevention (CDC). These include adults over 65 years, children under 5 years, and pregnant women, as well as individuals with HIV/AIDs, cancer, heart disease, stroke, diabetes, asthma, and children with neurological disorders.

Additional Context

A count of the historical influenza deaths gives an indication of the severity of flu in an area. Deaths can be prevented with flu shots and adequate medical staff. In the United States, each state has a different population composition, meaning that some states will have more vulnerable populations. In this project, you should pay particular attention to influenza deaths, vulnerable populations, and (optionally) flu-shot rates—particularly in vulnerable populations—to determine medical staffing needs.

Stakeholder Quotes

Influenza Patient: "I missed work the day they were providing flu shots."
Hospital Nurse: "The babies really suffer when they have the flu. I just moved to Utah this year, and flu season is so much worse here with the state's high birth rates."
Physician: "Being located in Florida near so many retirement communities, we see a lot of elderly patients during influenza season. These patients have a much higher risk of complications and fatality than normal."

Medical Staffing Agency Administrator: “We do see a big difference between states. States differ a lot in their populations and their efforts at prevention. We should take these into account for our planning.”

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

1. Influenza deaths by geography, time, age, and gender
Source: CDC
Download Data Set

2. Population data by geography
Source: US Census Bureau
Download Data Set

3. Counts of influenza laboratory test results by state (survey)
Source: CDC (Fluview)
Download Influenza Visits Data Set
Download Lab Tests Data Set

4. Survey of flu shot rates in children
Source: CDC
Download Data Set

Analysis Criteria
- You must explain what cleaning you conduct on the data.
- You must select and integrate at least two relevant data sets from different sources.
- You must identify or create a hypothesis that you then test with the data.
- You must look at the relationship between variables with at least one correlation found.
- You must include spatial and temporal visualizations in your final deliverable.
- You must include your conclusions, recommendations, and proposed next steps in your final presentation.
- You must consider the audience when determining which analysis components to include in your final presentation.

Your Project Deliverables
Throughout the next two Achievements, you’ll be working from Exercise to Exercise to complete your project, submitting a deliverable in each Task. At the end of each Achievement, you’ll create a final deliverable that your Mentor will review for your portfolio.

DATA OVERVIEW

Population Data by Geography:

This is an external data source, publicly available from the U.S. Census Bureau, a non-partisan government agency. This data is considered trustworthy, as it collected by the government and used to make important decisions such as the allocation of federal funding. (https://www.census.gov/programs-surveys/acs/about/acs-and-census.html)

This is survey data, collected manually from the American Community Survey (ACS). This survey is conducted every month using a randomly selected sample of about 3.5 million addresses. Response to this survey is required by law and responses can be submitted online, by mail, by phone, or through an in-person interview. There is a time lag, as the Census Bureau must wait for the responses. The Census Bureau uses this sample to estimate demographics for the population of each county. (https://www.census.gov/programs-surveys/acs/about/top-questions-about-the-survey.html)

This data set lists population by county and month from 2009-2017. The population data is then further broken down by gender and age range.

The data could be biased. Even though the Census Bureau is required to keep all survey responses anonymous by law, there may be people choosing to evade the survey, for example, due to their immigration status. Manual errors are also possible—People could report false information mistakenly or on purpose. This data is collected every month, which is frequently.

This data set is relevant and useful to the project, as it is trustworthy data that can give us a count of the number of adults older than 65, and the number of children younger than 5 in each state. According to the CDC, these people are all considered vulnerable to influenza complications.

Influenza Deaths Data Set:

This is an external data source, publicly available from the Centers for Disease Control and Prevention (CDC) through the National Center for Health Statistics (https://wonder.cdc.gov/ucd-icd10.html) Since it is government data, it is trustworthy.

The data is collected as part of the National Vital Statistics Cooperative Program. Each U.S. state and territory is required to record all births and deaths. Death records come from death

certificates, where a doctor lists the primary cause of death as either “Influenza” or “Pneumonia” (ICD-10 codes J09-J18).

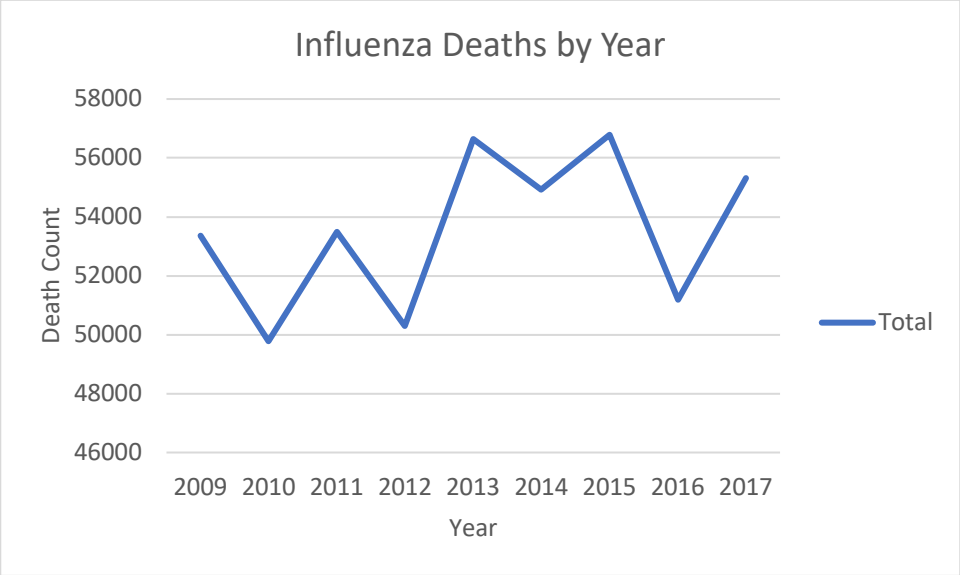
Since this data is part of the government's vital statistics program, an accurate count of deaths can be assumed. However, a death certificate only lists one cause of death. This could create some discrepancies in vulnerable populations e.g. cancer patients who also have the flu.

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

This data set is relevant and useful to this project, as it is trustworthy data that can tell us how severe the flu was in each state.

ADDITIONAL RESULTS/INSIGHTS

There is a very strong correlation between number of deaths and the number of people over 85 in each state. (r = 1) This is not too useful, as more people in general means more deaths.



	% Male Vulnerable Population (Male Vulnerable/Male Total)	*% Female Vulnerable Population (Female Vulnerable/Female Total)*
Mean	0.188848806	0.215053084
Variance	0.000190914	0.000242961
Observations	458	458
Hypothesized Mean Difference	0	
df	901	
t Stat	-26.92294366	
P(T<=t) one-tail	6.4659E-118	
t Critical one-tail	1.646546575	
P(T<=t) two-tail	1.2932E-117	
t Critical two-tail	1.962600396	