Preparing for Influenza Season- Interim Report

Alexandra Lindsay

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 higher percentage of the American population get vaccinated then influenza-related cases and deaths would decrease.

Data overview:

CDC Influenza Deaths

This is an external data source by the Centers for Disease Control and Prevention (CDC). The data contains monthly death counts for influenza-related deaths in the United States from 2009 to 2017 segmented by state and age.

US Census Bureau Population Data

This is an external data source that is provided by the US Census Bureau that is responsible for the national census. The data contains the annual American population in each county by state. The data starts in 2009 and ends in 2017 and includes the gender and an age range over under 5 to over 85.

CDC Influenza Laboratory Test Results and CDC Influenza Visit

These are external data sources that are provided by the CDC, the Centers for Disease Control and Prevention.

The lab data set, include all health and clinical laboratories that report total number of respiratory specimens tested and the number of positive cases of influenza every week with age or age group. The clinical laboratories include a weekly total number of specimens tested, number of positive influenza cases tested, the type and age group.

As for the influenza visits data set, it is manually entered through a network ILINet that they entered every week the data of the total number of patients seen, those with influenza and their age group.

Data limitations

CDC Influenza Deaths

Using this data set as a primary source is concerning due to large number of suppressed data (over 80%). To be able to draw conclusions for this analysis, we imputed the average 5 deaths for the suppressed data. This will influence the results in determining the seasonality of influenza cases.

US Census Bureau Population Data

The US Census Bureau does offer population estimates and demographic components of change every year but they are estimates therefore not 100% accurate. This data set inconsequence offers population counts with decimals.

CDC Influenza Laboratory Test Results and CDC Influenza Visit

The influenza visit data set is full of medical terms, which were not clearly defined and is missing data from Florida, New Jersey and Rhode Island. The data is also questionable, in the lab test, the percent positive is inaccurate and although there were age categories available, none were filled out.

Descriptive Analysis

The data grain below was state, year. Each data record is segmented by a particular year in each state in the United States. Unfortunately, since the data from under 5 to 14 years old are inputted, even if the study identified them as a vulnerable group we cannot accurately analyze the data. Which is why the data below is segmented between under 65 years of age and 65 years and above.

	Influenza deaths		Census Population		Death/pop ratio	
	>65+	<=65+	>65+	<=65+	>65+	<=65+
standard deviation	118.8233	971.7546	5944643	887017.2	0.028%	0.052%
mean	537.8061	896.7996	5167038	806988.9	0.027%	0.132%

By grouping the age groups to identify the vulnerable groups, we can see that there is a larger population sample of the 65+ age in the United States and that 89% of the deaths caused by influenza are citizens aged 65 years and above. The correlation between the deaths by influenza in both age groups demonstrate a correlation of 0.91. This means that even though we have an older population and that the death rates show they are the population that is more impacted, they are both affected by the spread of influenza.

Results and Insight

Null Hypothesis: The Death rate for deaths caused by Influenza for the age group of 65 years of age or older is less than the rest of the population

Alternative Hypothesis: The Death rate for deaths caused by Influenza for the age group of 65 years of age or older is higher than for the rest of the population.

With the established one-tailed null hypothesis above, since the p-value (1.12595E-14) is smaller than the established significance level of 0.05, we can establish a 95% level of confidence that null hypothesis is false. Therefore, the death rate cause by influenza for the age group of 65 + is higher than the rest of the population.

Remaining Analysis and Next Steps

Having established that the vulnerable population of 65 years and above should be our targeted group, the next step would be to identify when this population is most vulnerable throughout the year and segmented it by each state. If needed some data sets included data by county instead of state, therefore that could help establish trends.

By segmenting, the data above, a visual analysis of the targeted population growth by month and state. Same for the progressing death rates. By establishing the trends, we can then establish if the positivity rates and the clinic visits follow or deviate from the trend.

After the submitting of this interim report, a meeting will be sent to the stakeholders to discuss the current data points above and answer any questions. The next steps can be adjusted on the feedback received from the stakeholders. If all agreed upon, the next steps will follow the schedule as previously established.

Appendix

Schedule and milestones

week	**Tasks**
9	• **Halfway point deliverable:** consolidate findings of the analysis and present a report to stakeholders
10	• Explain how data visualizations can be used in this project
11	• Create data visualization design checklist • Explain how the visualizations can be improved
12	• Create a chart and treemap in Tableau • Use visualization design checklist to design chart
13	• Create a time forecast for a variable and display it • Use visualization design checklist to design chart
14	• Create visualizations that look at the distribution of a variable • Use visualization design checklist to design chart

15	• Create visualizations that look at the correlation between variables • Use visualization design checklist to design chart
16	• Map a variable and justify spatial visualization of choice • Use visualization design checklist to design chart
17	• Create a word cloud using qualitative data • Use visualization design checklist to design chart
18	• Create a narrative to communicate research findings and insights
19	• **Final deliverable:** record a video presentation for stakeholders

CDC Influenza Deaths

Data Cleaning/Renaming/Reformatting

Variables	Changes
State	cleaned up state column (a) due to state abbreviation and state name in description
State	there was a #NA in the state but with state code 11 attached to it, it is the district of Columbia therefore changed all the NA to the district of Columbia.
State code	missing code 3,7,14,43,52, verified state code data and these do not exist. No action required
year	Changed all 20133 to 2013, looking at column D month, the year 2013 was included in the data confirming assumption in data integrity exercise.
ten-Year age groups	there are 5508 counts of suppressed age groups, they are also suppressed deaths. Since do valid data was recorded, they have need omitted.

data uniqueness
state-year-month-age group
no duplicates found
since month variable includes year, I have separated them to leave only the month in the variable
removed state code/month code and ten-year age group code due to them being duplicate values of other variables

Completeness
each data count states the same total therefore is complete
There are still 48505 suppressed deaths but, due to this being too big of a data segment, we will need to leave it in. and input a mean of 5.

Timeliness	
Does the data need to be updated with any frequency?	No, it is historical data from 2009-2017.
What is the most recent data you need?	If data from 2017- today is available it would be determining the staffing needs with more up-to-date data and make this analysis more accurate to the population needs.
When was the data collected?	from January 2009- December 2017
Does the data have a suitable level of timeliness to be used in your analysis?	it will help to determine the influenza rates of the time but this data is 5 years old therefore would need more updated data to be accurate.

US Census Bureau Population Data

Data Cleaning/Renaming/Reformatting

Variables	Changes
County	Multiple county names in Puerto Rico has a character that didn't transpose therefore fixed the ? To the letter without the accent

data uniqueness and completeness
county-state-year
using county, state, year as unique identifiers, 3277 duplicated removed
removed Puerto Rico data since the agency is looking for deployment only in the United States
Missing 3372 data sets of years when filtering through county unique identifiers. Since no consistencies or pattern cannot fill in the dates that are missing.

Timeliness	
Does the data need to be updated with any frequency?	No, it is historical data from 2009-2017.
What is the most recent data you need?	If data from 2017- today is available it would be determining the staffing needs with more up-to-date data and make this analysis more accurate to the population needs.
When was the data collected?	from 2009- 2017 with some years missing
Does the data have a suitable level of timeliness to be used in your analysis?	it will help to determine the influenza rates of the time but this data is 5 years old therefore would need more updated data to be accurate.

CDC Influenza Laboratory Test Results

Data Cleaning/Renaming/Reformatting

Variables	Changes
weeks	there are 53 weeks in 2014 but there are only 52. will change this to 52
Percent Positive	due to having dates in the data set that were supposed to be percentages, I redownloaded the file from the CDC and copied the variable data once more.

data uniqueness and completeness
removed region type since only 1 type: state
region - year - week
Week: there are 53 weeks in 2014 but there are only 52. the data is different therefore not duplicates. Will keep the 53.
percent positive: due to having dates in the data set that were supposed to be percentages, I redownloaded the file from the CDC and copied the variable data once more
Year: not every year in data set has a full 52 weeks. Since data is missing cannot enter new data. Verified original data set from CSC fullview and they are missing that data as well.

Timeliness	
Does the data need to be updated with any frequency?	no, it is historical data from 2010 week 40 - 2015 week 39
What is the most recent data you need?	If data from 2015 - today is available it would be determining the staffing needs with more up-to-date data and make this analysis more accurate to the population needs.
When was the data collected?	from 2010 week 40 - 2015 week 39
Does the data have a suitable level of timeliness to be used in your analysis?	It will help to determine the influenza rates of the time but this data is 7 years old therefore would need more updated data to be accurate.

CDC Influenza Visit

Data Cleaning/Renaming/Reformatting

Variables	Changes
% weighted	have certain cells in the unweighted column to general since they had years in it
Region	The Commonwealth of the Northern Mariana Islands are a commonwealth of the US but it is not present in the rest of the data sets and there is no information entered in the sheet. All are "x" therefore it has been removed.
All age variables	there is no data entered for the age variables therefore are now removed.
% weighted/unweighted, ili total, num of providers and total patients	there are 465 lines that are only "x"s and have been removed
% of weighted ili	the data in this variable are either "X"s or 0s therefore have been removed

data uniqueness and completeness
removed region type since only 1 type: state
region - year - week
Week: there are 53 weeks in 2014 but there are only 52. the data is different therefore not duplicates. Will keep the 53.
%unweighted ili: due to having dates in the data set that were supposed to be percentages, I redownloaded the file from the CDC and copied the data set once more. It also includes the full week data set from 2019.
no need to Virgin Islands and Puerto Rico since agency only looking to staff within the US
Data set for NY state and NY city. Chose to not combine since the data set would be offset.
Florida has no data. Will keep for consistency.

Timeliness	
Does the data need to be updated with any frequency?	no, it is historical data from 2010 week 40 - 2019 week 52
What is the most recent data you need?	If data from 2019 - today is available it would be determining the staffing needs with more up-to-date data and make this analysis more accurate to the population needs.
When was the data collected?	from 2010 week 40 - 2019 week 52
Does the data have a suitable level of timeliness to be used in your analysis?	It will help to determine the influenza rates of the time but this data is 3 years old therefore would need more updated data to be accurate.