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: To assist in preparation of staffing plan in the United States for upcoming influenza season:

- Analyze death trends
- Prioritize states with vulnerable populations

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 larger population of those over 55 then the hospitalization rate in those states would be higher due to flu infection increase.

Hypothesis

If the state has a higher number of vulnerable populations over 65 years old, then the impact of the flu will be increased confirmed by a higher death rate.

Data overview

The Influenza death data set

Data sourcing: External data source. Provider- Centers for Disease Control and Prevention (CDC). CDC is a government agency, the national public health agency of the United States. It is under the US Department of Health and Human Service. Since it is the government data, we can count on this source of data to be trustworthy.

Data collection: The data is administrative data based on death certificates for US residents.

Each death certificate identifies a single underlying cause of death and demographic data. Each of the US 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 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.

Population data by geography

Data sourcing: External data source. Provider- US Census Bureau. US Census Bureau is an agency of the U.S. Federal Statistical System. It is responsible for producing data about the American people and economy. Since it is the government data, we can count on this source of data to be trustworthy.

Data collection: The data is collected through e-mail, phone or online using a survey and it is done manually. The data is administrative data to count citizens and non-citizens living in the US per decade.

Data contents: The data contains the number of the US population by county by gender and by age groups in 5-years increments.

Data limitations

The Influenza death data set

The Death certificate contains one cause of death as a final disease or condition resulting in death. It might influence on the data of vulnerable population, such as people with immune deficiency diseases or cancer, and their health decline might have been initiated by influenza.

Population data by geography

The data is collected every 10 years, the number might be estimated, collected on some samples and then extrapolated to the whole population. Therefore, the estimation might not reflect the real situation. Manually collection the data might be incomplete or contain errors.

Descriptive analysis summary

Core variable	The flu death rate below 65	The flu death rate 65 and above
Variance	0.000008%	0.000027%
Standard Deviation	0.027581%	0.052383%
Mean	0.026920%	0.131652%
Outliers	Two standard deviations are in a range -0.03% and 0.08%. It is no sense to have negative death rates that is why there were no negative outliers. There are 37 outliers from 459, the percentage is 8%	Two standard deviations are in a range from 0.03% and 0.24% There are 28 outliers from 459 rows. Outliers' percentage is 6.1%

A strong positive correlation of 0.79 has been identified between two core variables.

Results and insights

Null Hypothesis: The flu death rate of 65+ years old are the same with other age groups.

Alternative Hypothesis: The flu death rate of 65+ years old is higher than the other age groups.

We can say with 95% confidence (alfa 0.05) that there is a significant difference between the flu death rate of 65+ years and other groups.

Remaining analysis and next steps

Based on the result of the testing we proved that people over 65 years old are more vulnerable to the flu than other age groups. To determine a staffing plan with allocation to various regions across the USA during the flu season we will take into consideration number of vulnerable population in each state.

Next steps:

- Create a data visualization design checklist
- Create a time forecast for a variable and display it in Tableau
- Create visualizations that look at the distribution of a variable and the correlation between variables
- Map a variable and justify spatial visualization choice (heat, density, or choropleth)
- Create a word cloud using qualitative data
- Create a narrative to communicate research findings and insights in relation to research goals
- Record a video presentation for stakeholders

Appendix

Project Brief: Please refer to the Project Management Plan for detailed summary

Data quality

Influenza Deaths

Data Grain: State - Month Code - Ten-Year Age Group Code - Deaths: This presented the most unique combination of variables.

Completeness:

Variable	Count	Notes
State	1296	
Year	7344	
Month	612	
Month Code	612	
Ten Year Age Group	5508	Not Stated (NS) has 5508 counts. There is no impact on the analysis because the death column is surpressed, the way to resolve - doing nothing
Ten Year Age Group Code	5508	
Deaths	variable	The grand total equals the sum of records. There are 54013 records are surpressed, they can be ignored and the way to resolve is by doing

Uniqueness: No duplicate entries were found.

Timeliness: This Influenza Deaths Data contains data from 2009 to 2017 and this is the limitation of the project. An up-to-date version is not necessary.

Data Cleaning/Renaming/Reformatting:

Variables	Changes
Year	year 20133 replaced 2013 (17 cases)
State	Two letter State abbreviation changed into the full name
State	#N/A changed into District of Columbia based on the same State code 11

Variables and Data Types:

Variables	Data Types			
	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	quantitative	continious
Month	time-invariant	structured	quantitative	continious
Month Code	time-invariant	structured	quantitative	continious
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

Census Population

Data Grain: County - State - Year: combination of these variables describes the most unique record.

Completeness:

Variable	Count	Notes
State	25707	The grand total equals the sum of records. There are 51 states plus Puerto Rico.
County	25707	The grand total equals the sum of records.
Year	25707	The grand total equals the sum of records.
Total Population	25707	Variable
Male population	25707	Variable
Female population	25707	Variable
Age groups	25707	Variable

Uniqueness: 3278 duplicate values found and removed; 25707 unique values remain.

Timeliness: This Population Data contains data from 2009 to 2017 and this is the limitation of the project. An up-to-date version is not necessary.

Data Cleaning/Renaming/Reformatting:

Variables	Changes
Population columns (F-W)	the format of columns as number columns rounded to the whole numbe
County	County devided into 2 columns: County and State
County	Some of the names of Municipio in Puerto Rica (State) included ? Instead of spanish letters, changed into ley=tters i, n, o, u as in English

Variables and Data Types:

Variables	Data Types			
	time -variant/-invariant	structured/unstructured	qualitative/quantitative	qualitative: nominal/ordinal quantitative: discrete/continuous
County	time-invariant	structured	qualitative	ordinal
Year	time-invariant	structured	quantitative	continious
Total population	time-variant	structured	quantitative	discrete
Male Total population	time-variant	structured	quantitative	discrete
Female Total population	time-variant	structured	quantitative	discrete
Under 5 years	time-variant	structured	quantitative	discrete
5 to 9 years	time-variant	structured	quantitative	discrete
10 to 14 years	time-variant	structured	quantitative	discrete
15 to 19 years	time-variant	structured	quantitative	discrete
20 to 24 years	time-variant	structured	quantitative	discrete
25 to 29 years	time-variant	structured	quantitative	discrete
30 to 34 years	time-variant	structured	quantitative	discrete
35 to 39 years	time-variant	structured	quantitative	discrete
40 to 44 years	time-variant	structured	quantitative	discrete
45 to 49 years	time-variant	structured	quantitative	discrete
50 to 54 years	time-variant	structured	quantitative	discrete
55 to 59 years	time-variant	structured	quantitative	discrete
60 to 64 years	time-variant	structured	quantitative	discrete
65 to 69 years	time-variant	structured	quantitative	discrete
70 to 74 years	time-variant	structured	quantitative	discrete
75 to 79 years	time-variant	structured	quantitative	discrete
80 to 84 years	time-variant	structured	quantitative	discrete
85 years and over	time-variant	structured	quantitative	discrete

Hypothesis Testing Results:

t-Test: Two-Sample Assuming Unequal Variances

	Flu death rate under 65 years old	*Flu death rate 65+ years old*
Mean	536.6296296	896.7995643
Variance	14183.16384	944307.0209
Observations	459	459
Hypothesized Mean Difference	0	
df	472	
t Stat	-7.881701171	
P(T<=t) one-tail	1.12595E-14	
t Critical one-tail	1.648088336	
P(T<=t) two-tail	2.25189E-14	
t Critical two-tail	1.965002676	

The two tailed test has been conducted because we wanted to reject the null hypothesis that the flu death rate of 65+ years old are the same with other age groups. The p-value is 2.21589E-14 is so small and we can say that the null hypothesis has been rejected. We can state with 95% confidence level that death rate of the population over 65 years old is more likely higher that other age groups, and there is a direct correlation between older people (65+) and higher flu death rate.