Preparing for Influenza Season: Interim Report

1. **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.

2. **Research Hypothesis**

   If individuals are infected with influenza in USA, then the likelihood of influenza-related deaths is higher among those aged 65 years or older compared to children under the age of 5 years.

3. **Data overview**

Data set:	Influenza Deaths	US Census data
Data Source:	Centers for Disease Control and Prevenion (CDC)	US Census Bureau
Data Content:	The data contains monthly death counts for influenza-related deaths in the United States from 2009 to 2017. Counts are broken into categories of state and age.	The data set contains count of population in the United States listed by states for the period of 2009-2017. Counts are broken into categories of state, sex and age.

Table 1. Represent the data overview

4. **Data limitations**

   1) Influenza Deaths

   Single Cause of Death: Death certificates only list one cause of death, potentially leading to underrepresentation or overlooking of cases where influenza may have contributed to the decline in health of vulnerable populations.

   2) US Census data

   • Sampling Bias: The US Census uses sampling techniques, which can result in biased population estimates if certain groups are underrepresented or missed in the sample.

   • Non-Response Bias: Some individuals or households may not respond to the census survey, leading to bias in the data.

   • Coverage Bias: The census may not accurately capture all segments of the population, such as homeless individuals or those in remote areas, resulting in undercounting or omission.

• Timing and Frequency: The decennial census is conducted every ten years, providing a comprehensive count. However, for more frequent updates, other surveys like the American Community Survey are used, which may have a delay between data collection and release. This limits the availability of up-to-date population data.

5. Descriptive analysis

	Sum of population < 5	Sum of deaths < 5	Sum of population +65	Sum of deaths +65
Data Set Name	USA Census Dataset	Influenza Deaths Dataset	USA Census Dataset	Influenza Deaths Dataset
Sample/Population	Sample	Sample	Sample	Sample
Normal Distribution	yes	yes	yes	yes
Variance	2,10831E+11	0	7,96789E+11	944307,021
Standard Deviation	459163,585	0	892630,263	971,755
Mean	394351,026	120	829429,917	896,800
Outlier %	4%	0%	7%	4%

Table 2. Represent the descriptive analysis of the core variables

Variables	Sum of population < 5	Sum of deaths < 5	Sum of population +65	Sum of deaths +65
Correlation Coefficient	/		0,94	
Strength of Correlation	no relationship		strong relationship	
Usefulness and interpretation	There is no variability in the data of the variable "Sum of deaths < 5". Therefore, without variability in the variable of interest, it is not meaningful to calculate the correlation coefficient or assess the relationship with other variables.		The high correlation coefficient suggests that there is a consistent pattern in the data, indicating that as the population size of individuals above 65 years increases, the number of deaths from influenza tends to increase as well. In terms of usefulness, a correlation coefficient of 0.94 is considered very high and indicates a strong relationship between the variables. This information can be valuable for understanding the variation in the number of deaths influenced by influenza among the population above 65 years old.	

Table 3. Represent the correlation results between the core variables

6. Results and insight

Statistical Hypothesis:	
Null Hypothesis:	Mortality ratio of influenza-related cases for the individuals aged 65 years or older is less than or equal to mortality ratio of influenza-related cases for children under the age of 5.
Alternative Hypothesis:	Mortality ratio of influenza-related cases for the individuals aged 65 years or older is greater to mortality ratio of influenza-related cases for children under the age of 5.
Standard-level significance level, alpha:	0,05
P-Value:	4,07E-15
Significance Level Assessment:	This p-value is significantly less than 0.05. Therefore, we can reject our null hypothesis and state that mortality ratio of influenza-related cases for the individuals aged 65 years or older is greater to mortality ratio of influenza-related cases for children under the age of 5, with 95% confidence level.

Table 4. Representation of statistical test parameters and interpretation of the statistical test results

7. Remaining analysis and next steps

- Statistical visualizations and spatial analysis in order to:
  - map the vulnarible population and prioritize resource allocation and staffing plans for states with higher vulnerability levels
  - examine the monthly distribution of influenza deaths to determine if influenza occurs seasonally or throughout the entire year

- Publishing the analysis as a Tableau Storyboard
- Power point presentation of the final results of the conducted analysis.

Author: Sanja Ilin

Date: Juni 2nd, 2023

Appendix

1. Project management plan

1.1. Stakeholder communication

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

Meetings (with all stakeholders):
In week **two**, I hold a meeting with the stakeholders listed above. We define a few medical agency staff representatives and parent representatives. During the meeting, business requirements are discussed, and I ask the clarifying, funneling, and privacy- and ethics-related questions I brainstormed in week one. Finally, I relay my communication plan.
During **week ten**, I hold the final meeting with all stakeholders to present the results of the conducted study.

Written Communication:
In **week five** an interim report will be sent to staffing agency administrators; hospitals and clinics using the staffing agency's services and the above chosen stakeholder representatives.

Calls:
In **week 6**, after sending the report, I hold a call with the stakeholders to answer any questions they have (or ask more of my own).
In week ten I hold a call with the stakeholders for presenting the final results – video presentation showing the results to stakeholders.

Emergency/Contingency Plan:
Any urgent issues are communicated via email with a follow-up call scheduled within three days.

1.2. Schedule and milestones

- Project research phase - create a list of the data questions you need to answer for the analysis (**by week 2)**
- Design a data research project (**by week 2**)
- Sourcing the Right Data – explain the relevance and limitations of each data set to your project (**by week 3**)
- Data Profiling & Integrity - create a data profile for each of the data sets in your analysis, including information on data types, data integrity issues (accuracy and consistency), data cleaning, as well as summary statistics in each profile (**by week 3**)
- Data Quality Measures - Implement additional data quality measures to your data profiles related to completeness, uniqueness, and timeliness. (**by week 4**)

- Data Transformation & Integration (**by week 4**)
- Conducting Statistical Analyses (**by week 4**)
- Statistical Hypothesis Testing (**by week 5**)
- Consolidating Analytical Insights - Interim report (**by week 5**)
- Create a data visualization design checklist & Connect the project data to Tableau (**by week 6**)
- Design the charts- pie, bar, or column chart, a treemap in Tableau, as well as time forecast for a variable (**by week 7**)
- Statistical Visualizations- create visualizations that look at the distribution of a variable& create visualizations that look at the correlation between variables (**by week 8**)
- Spatial Analysis - Map a variable and justify your spatial visualization choice (heat, density, or choropleth) (**by week 9)**
- Textual Analysis - Create a word cloud using qualitative data (**by week 9**)
- Storytelling with Data Presentations - Create a narrative to communicate the research findings and insights in relation to the research goals & publish the analysis as a Tableau Storyboard (**by week 11)**
- Final presentation (**by week 11)**

1.3. Project deliverables

- Interim report consolidating the findings of the analysis.
- Power point presentation for the final results of the conducted analysis.

2. Hypothesis development

The project brief defines the vulnerable population as patients who are at higher risk of developing flu complications and requiring additional care, as determined by the CDC. This includes adults over 65 years, children under 5 years, pregnant women, individuals with certain health conditions such as HIV/AIDS, cancer, heart disease, stroke, diabetes, asthma, and children with neurological disorders.

However, according to the research article" Influenza (Seasonal) (who.int)" from the WHO, in developed countries deaths related to influenza predominantly occur among individuals aged 65 years or older. Considering the USA's status as a developed country, and in alignment with the business requirements and the need to prioritize states with large vulnerable populations, the research hypothesis has been formulated as follows*: If individuals in the USA become infected with influenza, then the likelihood of influenza-related deaths will be higher among those aged 65 years or older compared to children under the age of 5 years*.

3. Data Overview

For any additional information related to the profiles of the data please refer to the following documents: **[Data Set Profiles]**

4. Results and Insights

For any additional information related to the statistical analyses and statistical hypothesis testing, please refer to the following documents: **[Results and insights]**