PROJECT PLAN FOR UPCOMING INFLUENZA SEASON: INTERIM REPORT

ANUSHMA SHARMA

Project Overview:

Goal: Analyze and compare influenza-related staffing needs across U.S. states by examining differences in influenza death rates between vulnerable populations (below 5 years and above 65 years) and non-vulnerable populations (between 5 and 65 years).

Motivation: To better allocate resources and manage the influenza season by understanding its impact on vulnerable and non-vulnerable populations.

Objective: Identify patterns and trends in influenza death rates between vulnerable and non-vulnerable populations using statistical analysis and visualizations.

Scope: Analyze CDC influenza death data and US Census Bureau population data (2009-2017), focusing on comparing death rates between vulnerable and non-vulnerable populations. Utilize maps, bar charts, and line charts for key findings and trends.

Research Hypothesis:

If a state has a larger vulnerable population (below 5 years and above 65 years), then it will have a higher influenza death percentage compared to the non-vulnerable population (between 5 and 65 years).

Data Overview:

Influenza Deaths (Source: CDC)

- **Description:** This dataset contains records of influenza-related deaths.
- **Variables:** State, year, month, and age group.
- **Time Period:** 2009-2017.

Population Data (Source: US Census Bureau)

- **Description:** This dataset provides state-wise population data.
- **Variables:** State, year, and various age groups.
- **Time Period:** 2009-2017.

Together, these datasets enable a comprehensive analysis of influenza death rates with population demographics across different states and years.

Data Limitations:

Sample Size: Variability in sample sizes, especially in states with smaller populations, may impact the statistical reliability of the analysis.

Data Variability: Inconsistencies in reporting methods across states may introduce variability in the data, affecting the analysis.

Suppressed and Missing Data:

- Replaced entries labelled 'Suppressed' with 0 (zero).
- Removed entries labelled 'Not Stated'.
- Conducted data imputation for missing values using statistical methods.
- Excluded incomplete records from the analysis.

Descriptive Analysis:

Data Spread				
Variable	**Total Non-Vulnerable Population**	**Total Deaths of Non-Vulnerable Population**	**Total Vulnerable Population**	**Total Deaths of Vulnerable Population**
Dataset Name	Integrated Data	Integrated Data	Integrated Data	Integrated Data
Sample or Population?	Sample	Sample	Sample	Sample
Normal Distribution?	Left-Skewed	Left-Skewed	Left-Skewed	Left-Skewed
Variance	31292718528837	24284	1823088779379	1053021
Standard Deviation	5593990	156	1350218	1026
Mean	5253724	85	1317152	897
Median	3733097	20	950028	560
Outlier Lower Bound	-5934255	-226	-1383284	-1156
Outlier Upper Bound	16441703	397	4017588	2949
Outlier Count	18	28	29	18
Outlier Percentage	4%	6%	6%	4%

Correlation:

1. Total Non-Vulnerable Population and Total Deaths of Non-Vulnerable Population:

- **Proposed Relationship:** If the total non-vulnerable population increases, then the total deaths of the non-vulnerable population will also increase. This suggests a direct relationship where larger populations of non-vulnerable individuals lead to higher death counts in this group.
- **Correlation Coefficient:** 0.93
- **Strength of Correlation:** Strong Relationship
- **Usefulness/ Interpretation:** With a correlation coefficient of 0.93, there is a strong positive relationship between the total non-vulnerable population and the total deaths within this group. This implies that as the non-vulnerable population increases, the number of deaths in this group also tends to rise. Understanding this relationship can help in allocating healthcare resources more effectively to manage the health outcomes of the non-vulnerable population.

2. Total Vulnerable Population and Total Deaths of Vulnerable Population:

- **Proposed Relationship:** If the total vulnerable population increases, then the total deaths of the vulnerable population will also increase. This indicates that states with larger vulnerable populations are likely to experience higher numbers of deaths in this demographic.
- **Correlation Coefficient:** 0.94
- **Strength of Correlation:** Strong Relationship
- **Usefulness/ Interpretation:** The correlation coefficient of 0.94 indicates a strong positive relationship between the total vulnerable population and the total deaths within this group. This suggests that states with larger vulnerable populations experience higher death rates among this population. This insight is crucial for public health planning and prioritizing resources for the most at-risk groups, such as the elderly and very young children.

3. Total Non-Vulnerable Population and Total Vulnerable Population:

- **Proposed Relationship:** If a state has a larger total non-vulnerable population, then it will likely also have a larger total vulnerable population. This relationship reflects demographic trends where increases in the general population often correlate with increases in specific age groups.
- **Correlation Coefficient:** 0.99
- **Strength of Correlation:** Strong Relationship
- **Usefulness/ Interpretation:** A correlation coefficient of 0.99 shows a strong positive relationship between the total non-vulnerable and vulnerable populations. This means that states with higher non-vulnerable populations also tend to have higher vulnerable populations. This relationship highlights the need for a balanced approach in healthcare planning to address the needs of both population groups concurrently.

4. **Total Deaths of Non-Vulnerable Population and Total Deaths of Vulnerable Population:**

- **Proposed Relationship:** If the total deaths of the non-vulnerable population increase, then the total deaths of the vulnerable population will also increase. This suggests that factors contributing to higher death rates in non-vulnerable populations might also affect vulnerable populations similarly.
- **Correlation Coefficient:** 0.92
- **Strength of Correlation:** Strong Relationship
- **Usefulness/ Interpretation:** With a correlation coefficient of 0.92, there is a strong positive relationship between the deaths of non-vulnerable and vulnerable populations. This indicates that states with higher death counts in the non-vulnerable population also tend to have higher death counts in the vulnerable population. This suggests that systemic health issues or broader environmental factors may be impacting both groups, necessitating comprehensive public health interventions.

Results and Insights:

t-Test: Two-Sample Assuming Unequal Variances		
Variable	*Vulnerable Population Total Deaths*	*Non-Vulnerable Population Total Deaths*
Mean	896.6099291	85.46808511
Variance	1053020.57	24283.51971
Observations	423	423
Hypothesized Mean Difference	0	
df	441	
t Stat	16.07303295	
P(T<=t) one-tail	2.17586E-46	
t Critical one-tail	1.6483162	
P(T<=t) two-tail	4.35172E-46	
t Critical two-tail	1.965357827	

Hypothesis	If the population is vulnerable, then the number of deaths will be significantly higher compared to the non-vulnerable population.
Dependent variant	Total deaths
Independent variant	Population vulnerability status (Vulnerable vs. Non-Vulnerable)
Null hypothesis	There is no significant difference in the number of deaths between the vulnerable and non-vulnerable populations.
Alternative hypothesis	The number of deaths in the vulnerable population is significantly higher than in the non-vulnerable population.
One-tailed or Two-tailed test	I used a two-tailed test because I am interested in detecting any significant difference in the number of deaths between the vulnerable and non-vulnerable populations, regardless of the direction of the difference.
alpha	0.05
p-value	4.35E-46
p-value vs alpha	Since p < α (i.e. 4.35E-46 < 0.05), we reject the null hypothesis. This indicates that the observed difference in the number of deaths between the vulnerable and non-vulnerable populations is statistically significant. The extremely small p-value provides strong evidence that there is a significant difference in the number of deaths between these two groups, thereby supporting the alternative hypothesis.

Remaining Analysis and Next Steps:

Remaining Analyses:

- Conduct visual analysis to understand the distribution and variability of the data better.
- Perform subgroup analysis considering factors such as geographic regions and healthcare access.

Next Steps:

- Prepare a comprehensive final report with detailed visualizations and interpretations.
- Develop a presentation for stakeholders summarizing key findings and recommendations for resource allocation.

Appendix:

Additional Resources for Stakeholders:

- Influenza deaths by geography: https://www.cdc.gov/nchs/fastats/flu.htm
- Population data by geography, time, age, and gender: https://www.census.gov/data.html
- Access to the analysis script and Excel sheets used in this project.