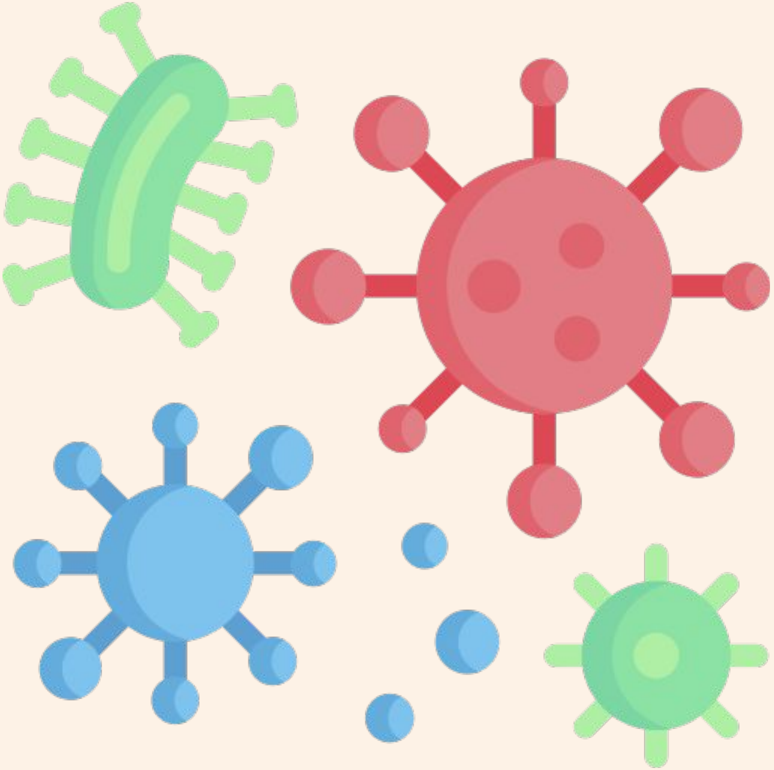
INFLUENZA SEASON

CASE STUDY



Business Problem

The United States has an influenza season where more people than usual suffer from the flu. Hospitals and clinics need additional staff to adequately treat these extra patients. The goal is to help a medical staffing agency that provides temporary workers to clinics and hospitals on an as-needed basis. The final results will examine trends in influenza and how they can be used to proactively plan for staffing needs across the country.

Key business questions:

- Determine whether flu occurs seasonally or throughout the entire year. If seasonal, does it start and end at the same time in every state?
- Who is most affected by the flu?
- Which states are at high risk?
- Establish the final staff planning guide.

Data Overview

Influenza Deaths by geography, time, age, and gender from CDC.

Population Data by geography from US Census Bureau. Dataset can be downloaded here.

Both datasets consist of totally around 100,000 observations.

Tools

- Tableau
- Excel

Analytical Methods

Data profiling
Data transformation & integration
Statistical hypothesis testing
Forecasting
Tableau storyboard

Data cleaning
Statistical analyses
Temporal & statistical visualizations
Spatial analysis
Presenting results

1. Data Profiling, Cleaning & Integration

- Created a data profile for each of the data sets.
- Included information on data types, data integrity issues (accuracy and consistency), cleaning that was conducted, as well as summary statistics in each profile.
- Integrated data from two sources into one cohesive data set for final analysis using data transformations.

CHALLENGES	SOLUTIONS
Each dataset had different measures of time: week, month, year	Aggregated and transformed data into yearly records
Data varied drastically state by state	Normalized population counts per state by turning them into a percentage rather than raw data

2. Statistical Hypothesis Testing

- Proposed hypothesis:
  If the individual affected with influenza is aged 65 years or above (vulnerable population), then they are at a greater risk of mortality.
- Null hypothesis:
  The mortality rate of individuals aged 65 years or above (vulnerable population) is the same or less as the mortality rate of individuals aged less than 65 years (non-vulnerable population).
- Alternative hypothesis:
  The mortality rate of individuals aged 65 years or above (vulnerable population) is more than the mortality rate of individuals aged less than 65 years (non-vulnerable population).

- Dependent variable: Mortality rate
- Independent variable: Vulnerable population

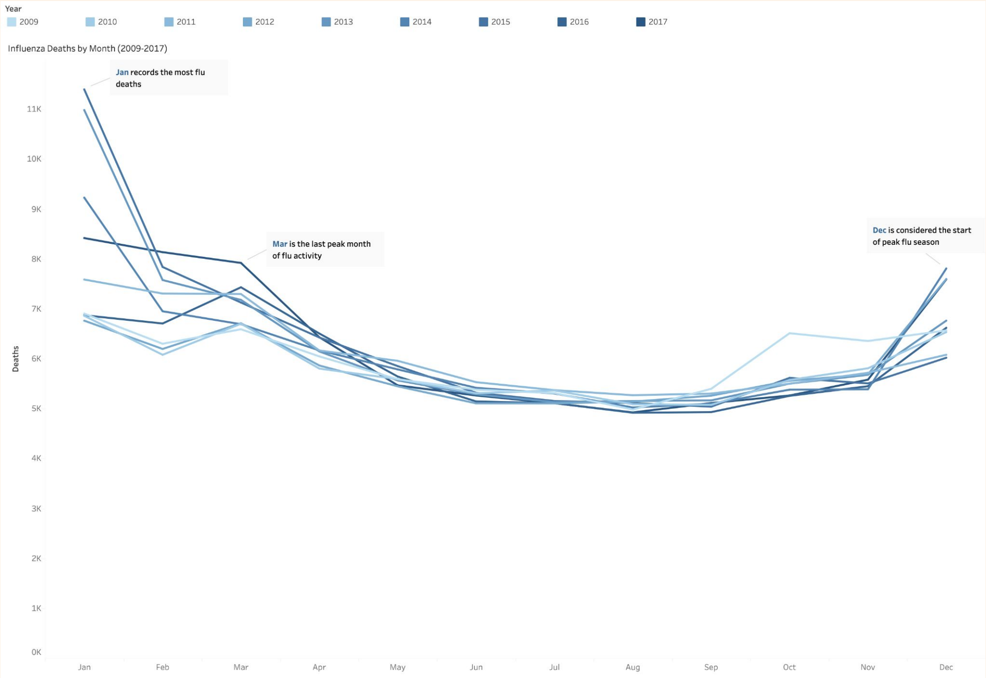
- A one tailed T-test was used to understand the relationship between the two variables.

	MORTALITY RATE BELOW 65 YEARS	*MORTALITY RATE 65 YEARS AND ABOVE*
Mean	0.000268932	0.00131388
Variance	7.58101E-08	2.74343E-07
Observations	459	459
Hypothesized Mean Difference	0	
df	693	
t Stat	-37.83313364	
P(T<=t) one-tail	**4.97E-171**	
t Critical one-tail	1.647055388	
P(T<=t) two-tail	9.9396E-171	
t Critical two-tail	1.96339306	

- Interpretation:
  - At an alpha of 0.05 or confidence level of 95%, a significant difference is observed in the independent variable (age) between the 2 groups: mortality rate below 65 years, mortality rate 65 years and above.
  - Since p-value is significantly smaller than alpha, it is clear that mortality rate (dependent variable) is affected by age (independent variable); thereby rejecting the null hypothesis.

3. Temporal Visualization

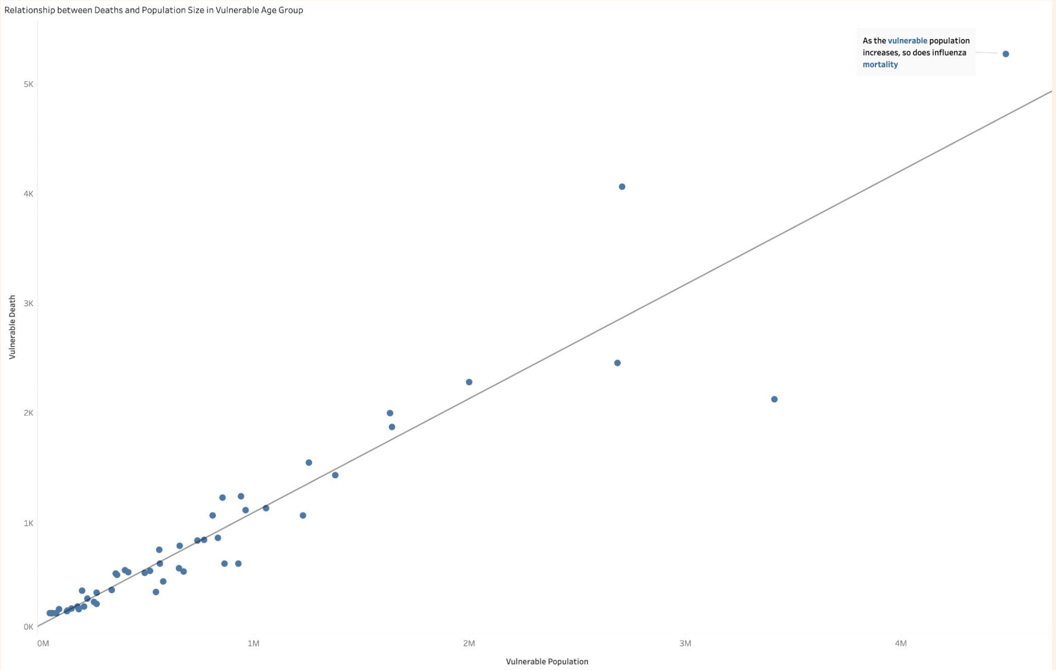
- Created a line chart to determine whether influenza occurs seasonally or throughout the year.



- Most flu fatalities occur from December to March, peaking during the colder winter months and early spring.

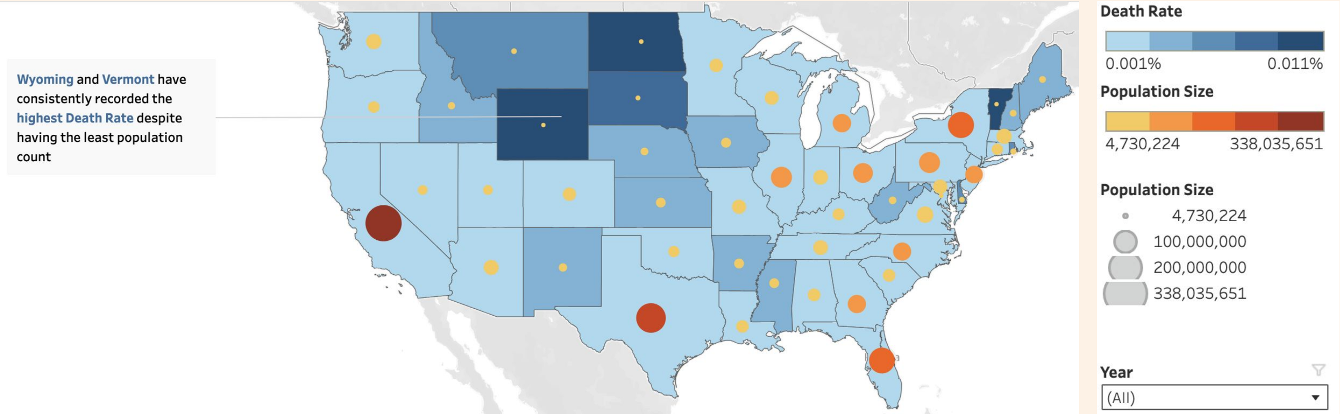
4. Statistical Visualization

- Created a scatterplot to visualize the relationship between Number of Deaths and Population Size in Vulnerable Age Group.

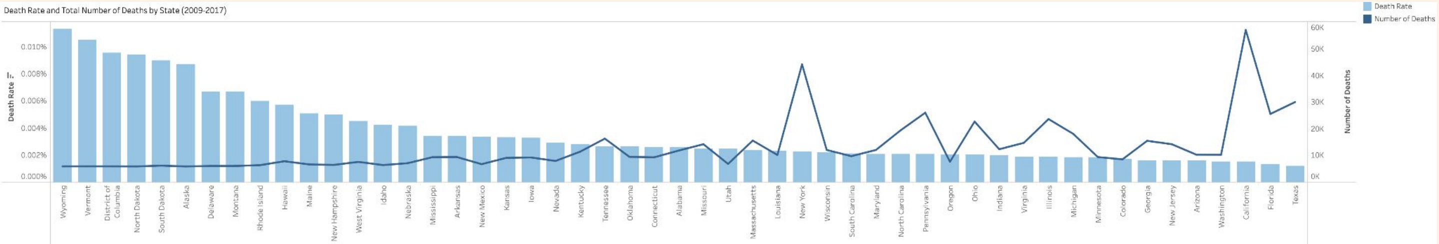


- The trend shows that as the population size increases, so does influenza mortality among the vulnerable group.
- A strong correlation is thus identified between the two variables.

5. Spatial Analysis



- Population Size has an impact on the number of influenza deaths.
- The more populous a state is, the higher are the deaths in it.
- However, deriving a new variable of Death Rate led to discover interesting insights.
- Created a map chart by integrating Death Rate and Population Size to understand whether Population Size has an impact on the Death Rate of a state or not.
- Wyoming, Vermont, District of Columbia have a higher death rate than California, Texas and New York.



- Death Rate highlighted several other states as a concern, that would need deployment of extra medical staff.

Conclusions & Recommendations

- Peak flu season occurs in the colder months from December through March. Send the first batch of medical personnel in October, when flu activity begins to increase. Send additional staff between December and March, when flu season peaks.
- Adults aged 65 years and older are at higher risk of complications from the flu.
- States with large populations, particularly large vulnerable populations, are of high priority in terms of agency staffing needs.
- Number of deaths is not the only factor to consider. States with higher death rates will also require additional resources.
- Key focus on below states:
  - Higher than average death rate: Wyoming, Vermont, District of Columbia, North and South Dakota.
  - High vulnerable populations: California, Florida, New York, Texas, Pennsylvania.

Next steps and further research

- Monitor influenza death rates on a regular basis after allocating agency staff to each state in order to check the efficacy of staffing project.
- Send out survey forms to patients and medical staff to gather feedback.
- Additional studies needed on vaccination rate of individuals and average staff-to-patient ratio in hospitals in every state.

Final Project Deliverables

- Tableau Storyboard
- Presentation

Reflections

This project was my favorite since I absolutely love working with SQL. What I like the most with SQL is its efficiency to manipulate data. In addition to that, I quickly learned that SQL queries are repeatable and scalable. I can write a query once and then reuse it again and again. In fact, even if new tables are added, I can use the previous query and make a few changes that shall work just fine. It was fairly simple to learn SQL yet it could solve incredibly challenging problems. Rockbuster's database had large sets of data in the form of numerous tables but SQL queries made pulling complex information simultaneously from multiple tables fast and easy.