PREPARATION OF THE 2018 INFLUENZA SEASON

Four datasets were analyzed using **Excel** and **Tableau** to examine the influenza trends between 2009 and 2017, to develop a medical resource allocation plan for the 2018 influenza season.

This case study details the analytical process undertaken and the skills developed throughout. For immediate access to the findings, please click the link below:

TABLEAU STORY

PROBLEM

Each year, the U.S. experiences an influenza season and an influx of influenza-related morbidity and mortality.

NEED

To assess how many additional medical professionals are needed in hospitals nationwide to ensure adequate protection and care of citizens during the influx.

QUESTIONS ADDRESSED:

Who is vulnerable?

When is the influenza season and the influenza-related death season?

When are the additional medical professionals needed?

Where are the vulnerable populations?

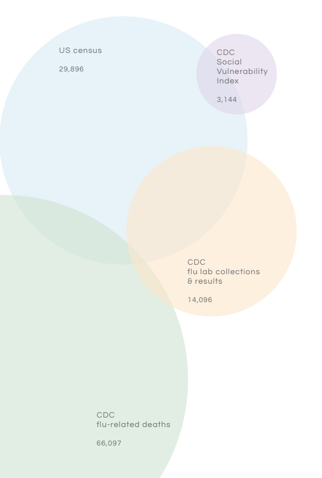
DATASETS:

During the initial review of the variables and the relevance and constraints of each dataset, and the development of the hypotheses, a crucial aspect related to health and the measurement of vulnerable populations was identified missing: the social determinants of health.

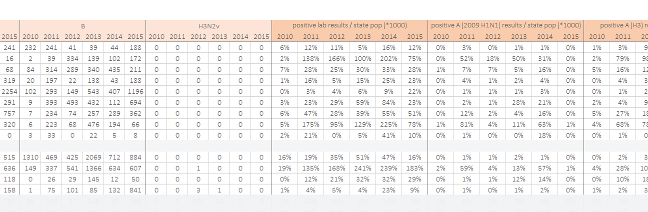
Social determinants of health are social, economic and environmental circumstances in which people are born, live, work and age. These play a significant role in people's health outcomes and accessibility to healthcare.

Through exploring online data repositories, an appropriate dataset was found, called the Social Vulnerability Index (SVI); collected and compiled by the Center for Disease Control and Prevention (CDC) and categorized by: socioeconomic status, minority status and language, household composition and disability, and housing type and transportation.

This analysis was ultimately performed on four datasets as seen on the right*. These took a form of .csv that were imported, cleaned and analyzed in Excel. Final .xlsx files were imported into Tableau to create visuals, Dashboards and Story.



*bubbles signify the relative size of the datasets based on the quantity of rows



DATA PREPARATION & TRANSFORMATION:

Each dataset was checked for quality, and cleaned and transformed.

Integrity check:
- filters and descriptive analysis were used to detect abonormal values, which were corrected or removed; data was checked again post correction or removal

Consistency check:
- miscalculations or incongruent formatting of quantitative values (ie calculated rates entered as decimals, fractions or dates) and spelling/spacing of nominal values were corrected and standardized
- variable groupings were standardized across datasets (ie age groups)
- data points for each variable were counted and standardized using pivot tables

Cleaning:
- duplicates were identified and removed using Remove Duplicates
- missing data were identified and either left as-is, corrected, or removed

Transformation:
- columns including multiple variables were separated to allow for a more granular analysis (ie 'City-County-State' individualized into three using Text-to-Columns, while retaining the original grouped column to be able to distinguish between same-name cities or counties in different states).
- some columns were concatenated to make a combined key (ie 'State' & 'Year')
- datasets were integrated using VLOOKUP
- rates for population, death, and infection were aggregated; weeks were aggregated into months
- standard deviations were used to identify outliers; possible causes were assessed

DATA EXPLORATION:

After transforming the data, we moved into the exploration phase. This stage provided an opportunity to interact with the data, developing and testing research hypotheses through two-tailed t-tests and correlation coefficients to assess relationships between variables.

Pivot tables, slicers, aggregations, and visuals were key tools in this analysis.

I learned that this phase, while resource-intensive second to data preparation and transformation, requires persistence and diligence to uncover meaningful insights. Another learning was that despite significant time spent on a particular hypothesis, the results may sometimes be non-relational or inconclusive.

STORYTELLING:

After thoroughly exploring the available data, a cohesive storyline that addressed each question comprehensively was constructed. This ideation phase required putting pencil to paper.

With a wealth of information at hand, it was a matter of consolidating and writing the findings on Post-Its and placing them in order to construct an insightful narrative.

CHALLENGE:

There were initial expectations for both the data's implications and the presentation of findings. However, when the data revealed a different patterns, I had to pivot and embrace a narrative that more accurately reflected the realities observed. Once I aligned with the actual data, a coherent conclusion emerged naturally. This insight led me to return to the beginning, to tell the story of how we got there.

PRESENTATION HIGHLIGHTS:

Age Group x Death: Who is vulnerable to influenza morbidity?

On average, each state shows a normal distribution of population among each age group (left), with an increase between ages 5 and 54 and a decline starting at 55.

There is an assumption that people over 65 are disproportionately vulnerable to flu-related mortality. This was confirmed with a strong correlation of 0.9 (out of 1.0).

The relationship between population of age groups (left) and flu-related deaths (right) shows a strong negative correlation. Meaning, as the population of increasing age groups decline, deaths of increasing age groups increase.

- Color and annotations used to highlight takeaway messages
- Venn diagram infographics used on each page to show which dataset(s) are being presented

Positive Lab Test Results: When is the influenza season?

Puerto Rico, **Alaska** and **Hawaii** experience uniquely perpetual flu seasons, with **Puerto Rico** and **Alaska** being in HIGH and MODERATE severity throughout the year, and **Hawaii** only dipping into LOW in October and November.

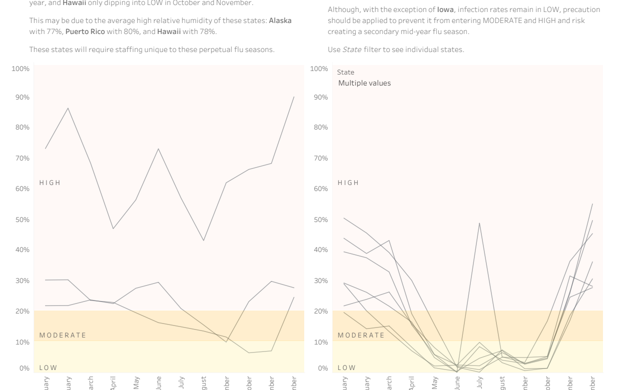
This may be due to the average high relative humidity of these states: **Alaska** with 77%, **Puerto Rico** with 80%, and **Hawaii** with 78%.

These states will require staffing unique to these perpetual flu seasons.

The states below (mostly in **HHS 5**) experience a unique secondary peaks in the summer; **Iowa** with an acute peak. This too could be due to relative humidity.

Although, with the exception of **Iowa**, infection rates remain in LOW, precaution should be applied to prevent it from entering MODERATE and HIGH and risk creating a secondary mid-year flu season.

Use *State* filter to see individual states.



- Drop-down menu used to select specific states and maximize visual real estate and readability
- Colors and annotations used to differentiate between infection severities

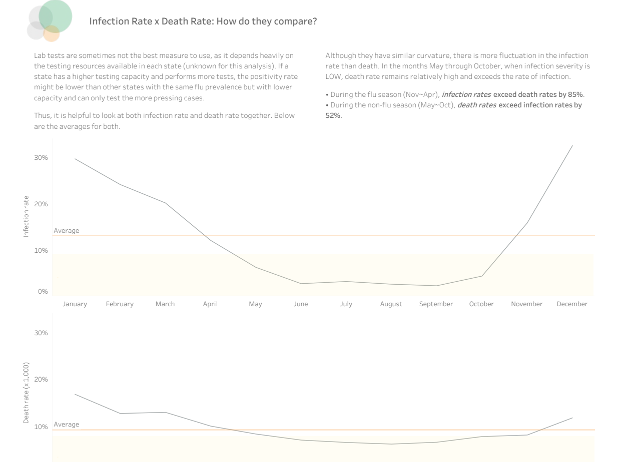
Infection Rate x Death Rate: How do they compare?

Lab tests are sometimes not the best measure to use, as it depends heavily on the testing resources available in each state (unknown for this analysis). If a state has a higher testing capacity and performs more tests, the positivity rate might be lower than other states with the same flu prevalence but with lower capacity and can only test the more pressing cases.

Thus, it is helpful to look at both infection rate and death rate together. Below are the averages for both.

Although they have similar curvature, there is more fluctuation in the infection rate than death. In the months May through October, when infection severity is LOW, death rate remains relatively high and exceeds the rate of infection.

- During the flu season (Nov–Apr), *infection rates* exceed death rates by 85%.
- During the non-flu season (May–Oct), *death rates* exceed infection rates by 52%.



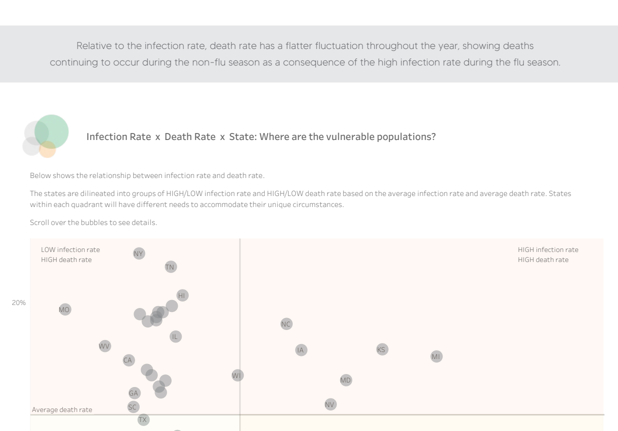
Relative to the infection rate, death rate has a flatter fluctuation throughout the year, showing deaths continuing to occur during the non-flu season as a consequence of the high infection rate during the flu season.

Infection Rate x Death Rate x State: Where are the vulnerable populations?

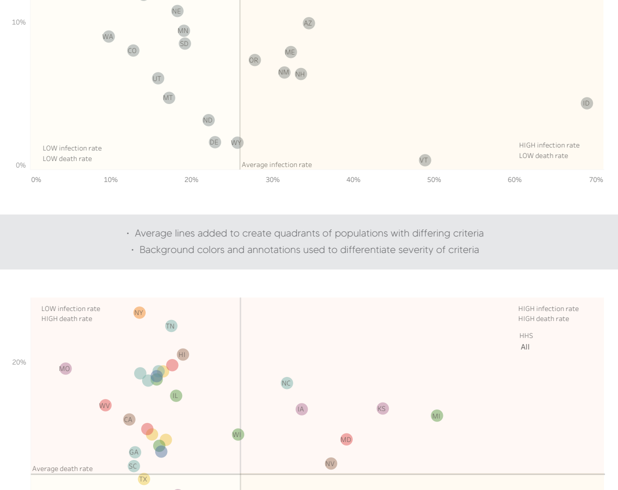
Below shows the relationship between infection rate and death rate.

The states are dilineated into groups of HIGH/LOW infection rate and HIGH/LOW death rate based on the average infection rate and average death rate. States within each quadrant will have different needs to accommodate their unique circumstances.

Scroll over the bubbles to see details.



- Average lines added to create quadrants of populations with differing criteria
- Background colors and annotations used to differentiate severity of criteria



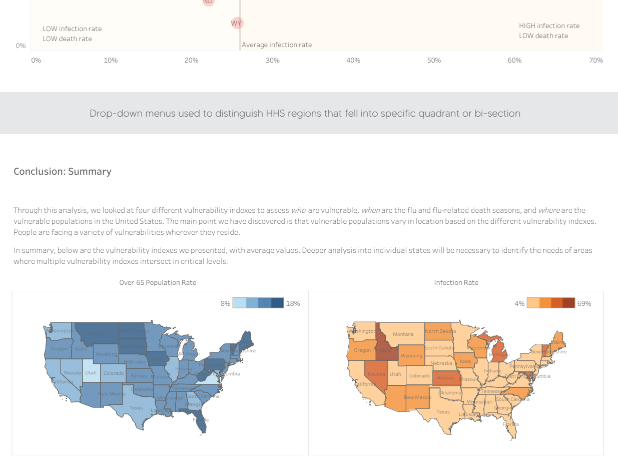
Department of Health and Human Services (HHS) regions added and color-coded to show regional commonalities

Drop-down menus used to distinguish HHS regions that fell into specific quadrant or bi-section

Conclusion: Summary

Through this analysis, we looked at four different vulnerability indexes to assess *who* are vulnerable, *when* are the flu and flu-related death seasons, and *where* are the vulnerable populations in the United States. The main point we have discovered is that vulnerable populations vary in location based on the different vulnerability indexes. People are facing a variety of vulnerabilities wherever they reside.

In summary, below are the vulnerability indexes we presented, with average values. Deeper analysis into individual states will be necessary to identify the needs of areas where multiple vulnerability indexes intersect in critical levels.



Choropleth maps used to show gradients of vulnerability index rates

CONCLUSION:

Initially, when formulating the research hypotheses, there were certain expectations about how the data would be presented. I envisioned using choropleth maps to represent each vulnerability index, with the darker shades representing higher vulnerabilities. The maps would then be overlayed to identify the darkest states, thus the most vulnerable and in need of the highest number of additional medical professionals.

However, the combination of demographics, social vulnerabilities, age distribution, geographical locations and climate resulted in each state exhibiting unique combinations of vulnerabilities.

It was concluded that a more granular and comprehensive analysis was necessary to understand the specific challenges, deficiencies, and needs faced by each community. This would produce a more precise measurement of additional medical professionals, as well as the specific resources to send to each community. This will optimize resource allocation, minimize wastage, and ensure that the provided resources align with each communities' needs.

This project reaffirmed the diversity, complexity and beauty of communities reflected in sociodemographic datasets. Additionally, the insight about crafting storylines and how to be led by - but not limited by - the data was a profound lesson.

SKILLS ACQUIRED:

Excel:
data preparation, transformation & analysis
- filtering
- identifying & removing duplicates
- formatting data consistencies
- text-to-column
- aggregation
- integration
- concatenation
- conditional formatting
- pivot tables, slicers
- data visualizations
- descriptive analysis
- correlation coefficient
- t-test

Tableau:
data visualization & storytelling
- data visualization:
  - line, forecasting (seasonal, non-seasonal)
  - bar, scatterplot, bubble
  - histogram, box plot
  - spatial analysis: point, bubble, heat, choropleth, combination
  - text analysis
- annotations, trendlines, average/custom lines
- Dashboard
- Stories

Akiko Oguchi | Data Analyst | Information Designer | oguchiakiko@gmail.com