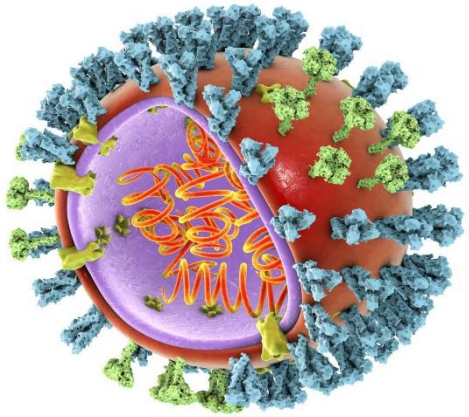
Data Analytics Case Study

Preparing for the 2018 Influenza Season



Eva-Maria Kuck

Project Overview

Problem:
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.

Questions:
1. Who is most vulnerable to influenza complications including death?

2. In which states is additional medical staff most needed?

3. In which months of the year is additional medical staff most needed?

Objective of the project:
Determine **when** to send staff, and **how many, to each state**.

Project Overview

Duration of the project:
6 weeks

Tools used: Tableau, Excel

Role: Data Analyst

Primary Stakeholder: CareerFoundry Data Analytics Course

Used Data

- **US Census Bureau Population Data**: The data includes populations by county, state, gender and age group for each year from 2009 to 2017.

  (https://coach-courses-us.s3.amazonaws.com/public/courses/data-immersion/A1-A2_Influenza_Project/Census_Population_transformed_202101.csv)
- **CDC Influenza Deaths Data**: The data contains monthly death counts for influenza-related deaths in the United States from 2009 to 2017 broken down by state and age group.

  (https://wonder.cdc.gov/ucd-icd10.html)

Used Skills

Excel:

- Profiling and cleaning the data, improving data integrity
- Integrating and transforming data
- Calculating descriptive statistics
- Statistical hypothesis testing

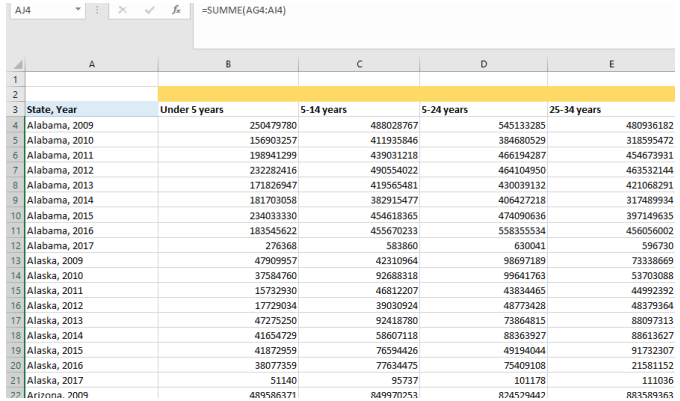
Tableau:

- Composition and comparison charts
- Temporal visualizations and forecasting
- Statistical visualizations
- Spatial analysis

Data Preparation

Before analyzing the data and starting to answer the project questions, I had to prepare and clean the data, because working with “dirty” data can lead to wrong conclusions and hence, hurt the business.The preparation included the following:

- Assessing data accuracy by calculating descriptive statistics
- Assessing data consistency by creating frequency tables and cleaning the data
- Assessing data completeness by creating frequency tables and finding ways to handle missing data
- Assessing data uniqueness by creating a pivot table and removing duplicate records
- Integrating the two datasets by using a key variable



AJ4 =SUMME(AG4:AI4)

	A	B	C	D	E
1					
2					
3	State, Year	Under 5 years	5-14 years	5-24 years	25-34 years
4	Alabama, 2009	250479780	488028767	545133285	480936182
5	Alabama, 2010	156903257	411935846	384680529	318595472
6	Alabama, 2011	198941299	439031218	466194287	454673931
7	Alabama, 2012	232282416	490554022	464104950	463532144
8	Alabama, 2013	171826947	419565481	430039132	421068291
9	Alabama, 2014	181703058	382915477	406427218	317489934
10	Alabama, 2015	234033330	454618365	474090636	397149635
11	Alabama, 2016	183545622	455670233	558355534	456056002
12	Alabama, 2017	276368	583860	630041	596730
13	Alaska, 2009	47909957	42310964	98697189	73338669
14	Alaska, 2010	37584760	92688318	99641763	53703088
15	Alaska, 2011	15732930	46812207	43834465	44992392
16	Alaska, 2012	17729034	39030924	48773428	48379364
17	Alaska, 2013	47275250	92418780	73864815	88097313
18	Alaska, 2014	41654729	58607118	88363927	88613627
19	Alaska, 2015	41872959	76594426	49194044	91732307
20	Alaska, 2016	38077359	77634475	75409108	21581152
21	Alaska, 2017	51140	95737	101178	111036
22	Arizona, 2009	489586371	849970253	824529442	883589363

Descriptive Statistics and Exploratory Data Analysis

With all the data clean and integrated, I once again calculated descriptive statistics. The core variables are (1) total population over the age of 65 years (a vulnerable group for influenza) and (2) total number of influenza deaths over the age of 65 years. The mean and standard deviation of these core variables are shown below:

	Population over the age of 65 years	Number of influenza deaths over the age of 65 years
Mean	518.127.389	350
Standard deviation	632.401.191	495

Also, as part of exploratory data analysis, Pearson's correlation test was performed to determine the relationship between age over 65 years and number of influenza deaths. The test showed that there is a strong positive relationship between the two variables, meaning that age over 65 years goes hand in hand with a high mortality for influenza.

Pearson's Correlation Coefficient: 0.83

Answering Question No. 1: Who is Most Vulnerable to Influenza Complications Including Death?

With the above results of the exploratory data analysis, I developed the null and the alternative hypothesis:

- Null hypothesis: The mortality rate for individuals over 65 years is the same or smaller than the mortality rate for individuals under 65 years.
- Alternative hypothesis: The mortality rate for individuals over 65 years is greater than the mortality rate for individuals under 65 years.

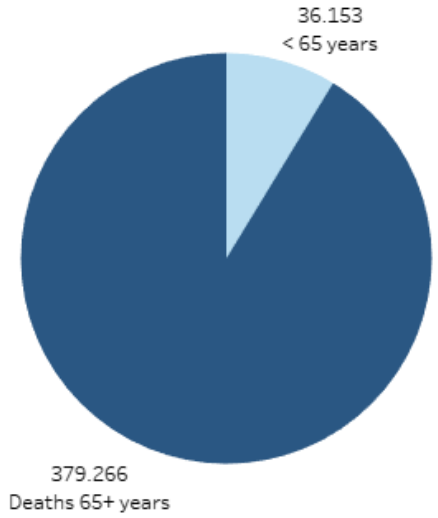
Then, I performed a one-tailed t-test which was significant at an alpha-level of 0.05, because the p-value was smaller than alpha=0.05 (see below). Therefore, we can reject our null hypothesis and state that the mortality rate of individuals over 65 years is greater than the mortality rate for individuals under 65 years with a confidence of 95%. **In other words: If you are 65 years and older, you are significantly more likely to die from influenza.**

- p-value one-tailed: 6.96E-45

Answering Question No. 1: Who is Most Vulnerable to Influenza Complications Including Death?

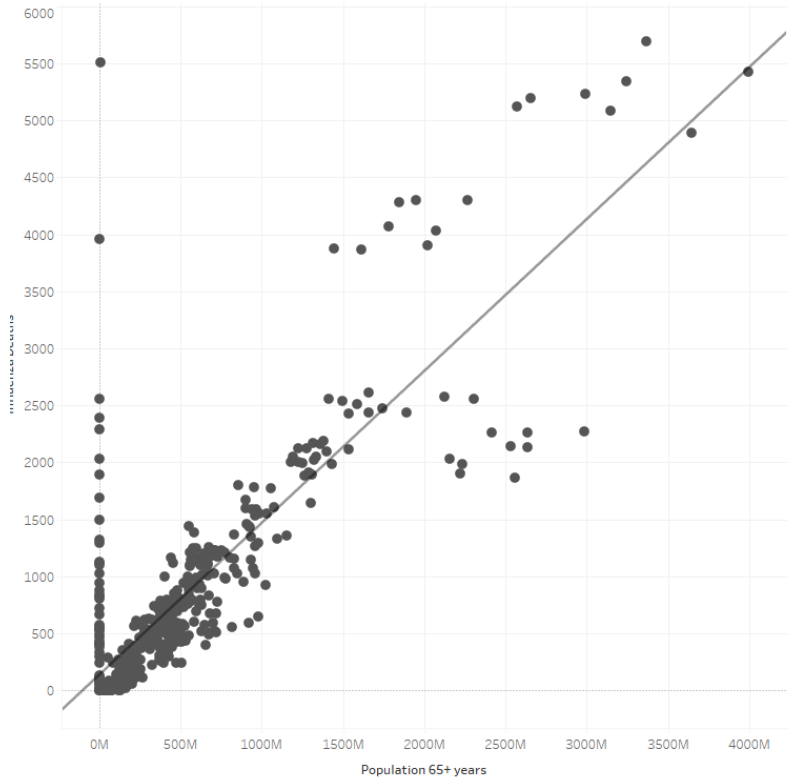
In addition, I visualized my findings with the following graphics in Tableau:

- Pie chart for total US influenza deaths by age group (2009-2017)



Answering Question No. 1: Who is Most Vulnerable to Influenza Complications Including Death?

- Scatterplot for the relationship between US population aged 65+ and influenza deaths (2009-2017)



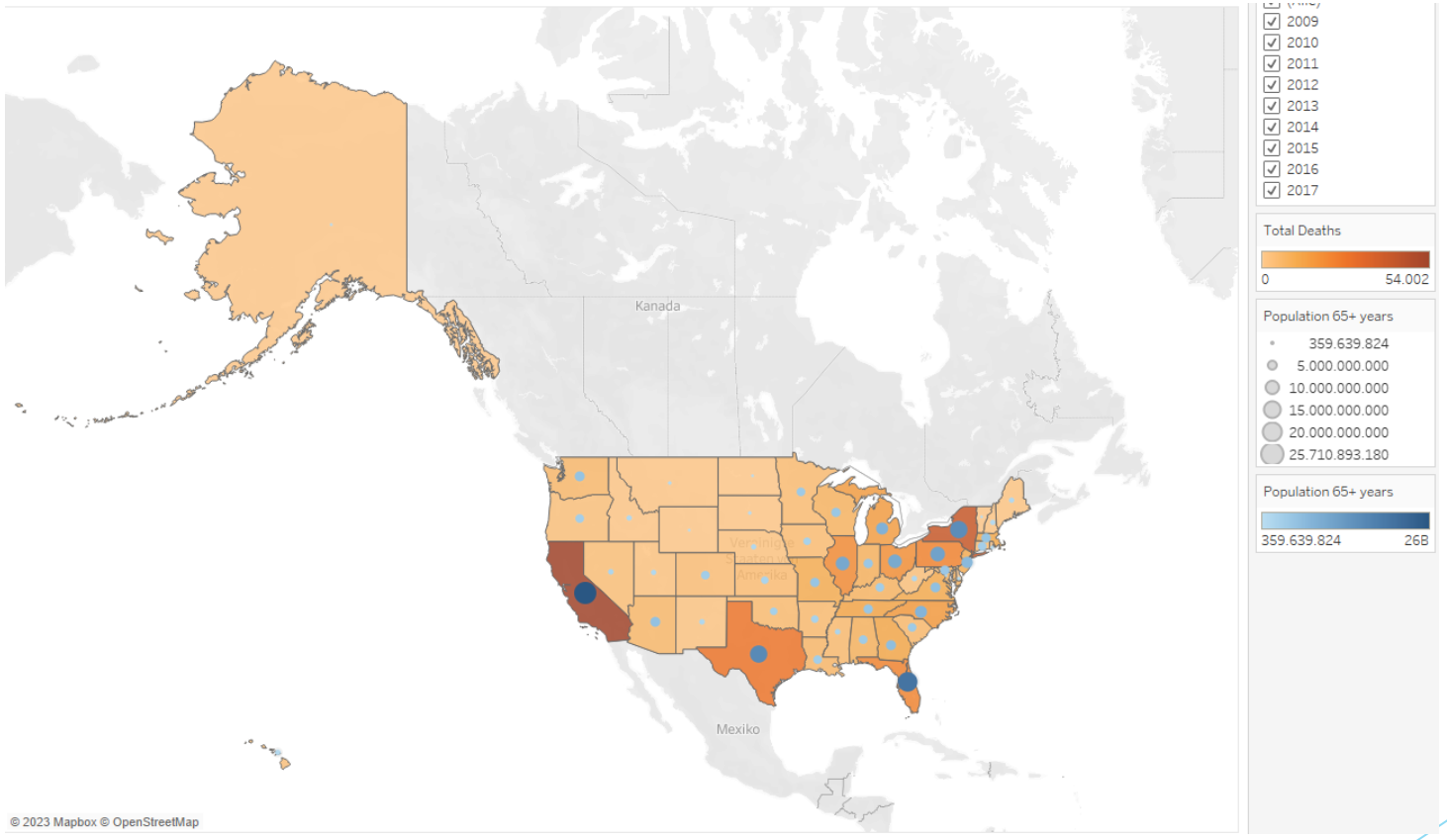
Results:

Influenza deaths of people over 65 years represent the vast majority of total influenza deaths. On the same lines, there is a strong positive correlation between age over 65 and influenza deaths.

Answering Question No. 2: In which States is Additional Medical Staff Most Needed?

For answering this question, I created various visualizations in Tableau:

- Dual axis map on total influenza deaths and population 65+ years by state (2009-2017)

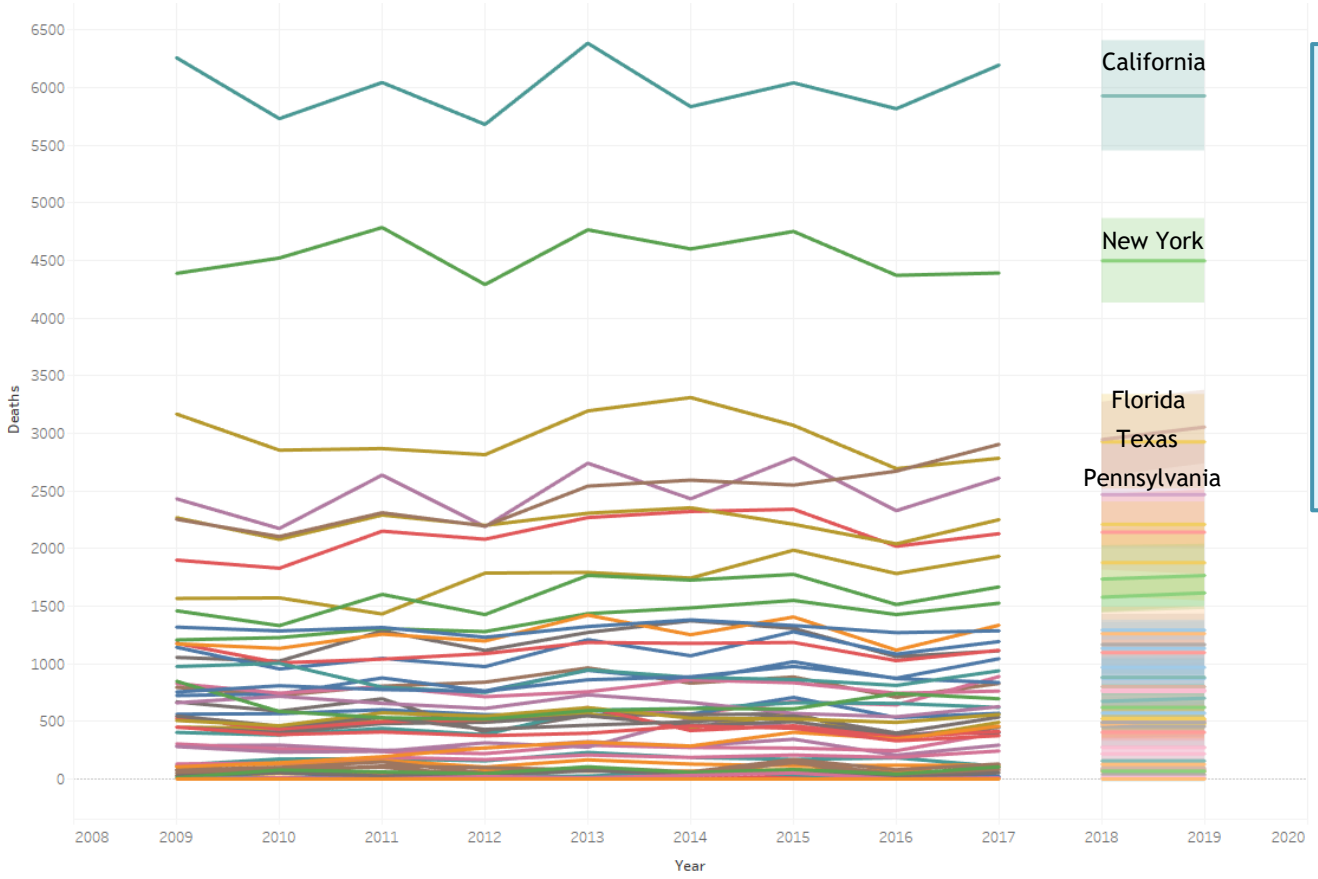


Result:

The top 5 states with the highest number of influenza deaths over time are California, New York, Texas, Pennsylvania and Florida.

Answering Question No. 2: In which States is Additional Medical Staff Most Needed?

- Line chart with 2018 seasonal forecast on influenza deaths by US state (2009-2017)

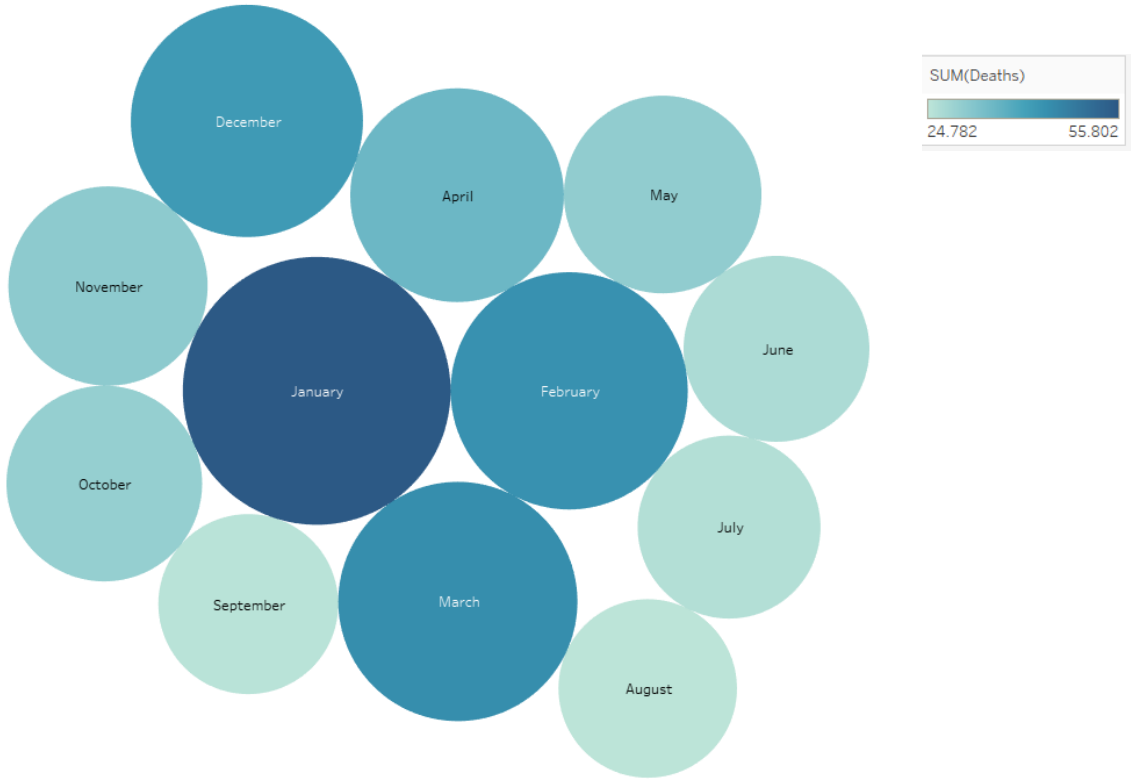


Result:

The top 5 states most affected by influenza deaths in 2018 are California, New York, Florida, Texas and Pennsylvania.

Answering Question No. 3: In which Months of the Year is Additional Medical Staff Most Needed?

For answering this question, I made a bubble chart in Tableau: Influenza deaths by month of year (over all US states, 2009-2017)



Result:

Most influenza deaths occur in January, followed by February, March and December. The typical influenza season is in wintertime.

Conclusion and Recommendations

Finally, I was able to formulate the conclusion and recommenations for the medical staffing agency.

Conclusions:

- Individuals of 65 years and older are a vulnerable group to the influenza virus.
- The US states with the highest number of influenza deaths over the last 9 years are California, New York, Texas, Pennsylvania and Florida.
- Influenza mortality is highest in the months of December until March.

Recommendations:

- The medical staffing agency should send extra staff to the states of California, New York, Florida, Texas and Pennsylvania (in descending order), especially from December to March, to help with high influenza cases and avoid influenza mortalities.

Challenges and Solutions

Working with datasets is never free of challenges and constraints. Here I describe the two biggest challenges and how I handled them.

1. The most notable limitation of the influenza deaths data set was the high percentage of missing values.

- Solution: Unfortunately, I couldn't do much about it, and the best option was to simply ignore the missing values since imputing values would have changes the data set too much due to the high percentage of missing values.

2. Integrating the two datasets was a bit of a challenge for me, since I have never done anything like that before and it required some preparatory work. The two datasets contained data in different formats.

- Solution: First I had to separate the state variable (it was combined with a county variable), then I aggregated the values per state and created a combined key with the values of "state" and "year" in each dataset before I could finally use the VLOOKUP-function to integrate the datasets.