Katherine Huerta
DATA 606 SP'20
Deliverable III

Exploring Confirmed Influenza Cases in New York State at the County-Level.

Methods Overview

For more details, please refer to my presentation or my code provided on GitHub (Delivery 3 Link). A summary of my methods are as follows:

1. Created subsets for each influenza season.
    a. Each resulting subset has 62 rows representing 62 unique counties, and the "Count" column containing the sum of all confirmed influenza cases for that county.
2. Created an interactive map visualizing the total number of confirmed influenza cases per county for each season.
    a. Used Python/Jupyter Notebook (Pandas and NumPy)
    b. For the map, I used Plotly Express:
        i. Mapbox Choropleth Maps [1].
    c. Based code off of Plotly Express documentation [1]. Resulting map can be seen in fig. 1(a).
3. Added population data to each season subset to make an interactive map - illustrating county-level prevalence rates.
    a. Extracted population data from "Annual Population Estimates for New York State and Counties: Beginning 1970" dataset [2]. Estimates are based on census counts (base population), intercensal and postcensal estimates [2].
    b. Added the population data as a new column in my season subsets.
    c. Defined and applied a function to calculate the prevalence rate of confirmed influenza cases (per 10,000 people) for each county in each subset.
4. Created an interactive map visualizing the prevalence rate of confirmed influenza cases (per 10,000 people) at a county-level.
    a. Followed the same code format as before, but used the subset containing the prevalence rates.

Results/Figures

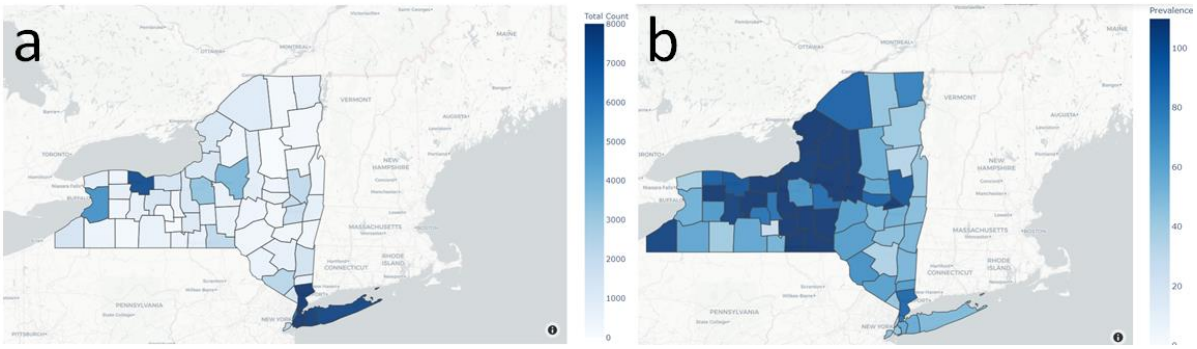

Figure 1. Side by side comparison of the two density maps from the 2017-2018 influenza season. The density map illustrating the total count of confirmed cases **(a)** is quite different from the map illustrating the prevalence rate of confirmed influenza cases per 10,000 people **(b)**.

Table 1. Comparison between the sum of confirmed influenza cases and prevalence rates in the 2017-2018 influenza season for three counties with either a relatively high sum of cases, or high prevalence rate.

County	Count (confirmed cases)	Prevalence Rate (confirmed cases/10,000 people)
Queens	13,511	59.90
Bronx	11,749	84.08
Oneida	3,745	159.89

Discussion/Conclusion

After creating interactive maps for all influenza seasons (visualizing the sum of all confirmed influenza cases per county - fig. 1(a)), it was brought to my attention that perhaps looking at the rate of confirmed cases with respect to the population may show something different and more meaningful. I defined a function to calculate the prevalence rate for each row (based on census-based population estimates [2]), and added this new column to all of the seasonal subsets. I chose the prevalence rate to be per 10,000 residents (rather than 100,000), because some counties had populations fewer than 50,000 people. After visualizing the prevalence rates across all counties for the 2017-2018 season, the counties that once seemed to be concerning (Queens, Bronx, etc.) had relatively low prevalence rates in comparison to many counties clustered in the center of the state (fig. 1(b)). Table 1 shows a brief comparison between the sum and prevalence rate of laboratory confirmed influenza of a few counties of interest.

The previous goal of my project was to make an interactive dashboard, but now the focus has shifted to attempting to answer some questions. I will try to determine whether there is a trend with the prevalence rates over all the seasons (in each county). I will also try to implement inferential statistics to identify what the relationship may be. Although there are not many features in my dataset to help identify potential factors affecting each county - there is data about the flu types (A or B). Influenza type A viruses are the only influenza viruses known to cause pandemics because of their ability to change in two ways (antigenic drift and shift) rather than one [3]. Thus, I want to see if there is a relationship between the number of each type of influenza (type A and B), and the prevalence rates in counties. My hypothesis is: If there are more influenza type A cases, the prevalence rate will be higher. This is because research indicates that influenza type A viruses are responsible for approximately 75% of confirmed influenza cases [4]. It will be interesting to see if New York State reflects a similar trend.

References Cited

1. Plotly Graphing Libraries, "Mapbox Choropleth Maps in Python", 2020. [Online], Available: https://plotly.com/python/mapbox-county-choropleth/. [Accessed March 15, 2020]
2. New York State Department of Labor. "Annual Population Estimates for New York State and Counties: Beginning 1970", 2020. Data.ny.gov [Online]. Available: https://data.ny.gov/Government-Finance/Annual-Population-Estimates-for-New-York-State-and/krt9-ym2k. [Accessed March 30, 2020].
3. Centers for Disease Control and Prevention, "How Flu Viruses Can Change", 2019. [Online], Available: https://www.cdc.gov/flu/about/viruses/change.htm. [Accessed March 29, 2020].
4. M. Nyirenda, R. Omori, H. Tessmer, H. Arimura, and K. Ito, "Estimating the Lineage Dynamics of Human Influenza B Viruses", PLoS One. 2016;11(11): e0166107. [Online], Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5102436/. [Accessed April 3, 2020].

Presentation Link