Katherine Huerta
DATA 606 SP'20
Final Deliverable

Exploring influenza cases in New York State with Python and Interactive Visualizations

Introduction

Project and Purpose.

The aim of this project was to explore seasonal influenza data to glean insights regarding trends at the county level in New York State. This was done primarily through interactive visualizations and analysis of New York State datasets. Additional goals of this project were to determine statistical significance of trends/correlations, as well as to create an interactive dashboard geared towards healthcare professionals and policy makers.

Dataset and Data Source.

The main data set I explored was *Influenza Laboratory-Confirmed Cases By County: Beginning 2009-10 Season* [1]. The data set can be acquired from HealthData.gov [1], or from the original source, the New York State Department of Health (health.data.ny.gov) [2]. The dataset contains 62,286 rows and nine columns. The nine columns are: Season (flu season ranging from October through the following May), Region of New York State (region of lab-confirmed cases, such as Central, Western, Capital District, etc.), County (county in New York State, such as Madison County, NY), CDC Week (week number in season), Week Ending Date, Disease (influenza strain - A or B), Count (number of cases), County Centroid (map coordinates), and FIPS (Federal Information Processing Standard - five digit FIPS code that uniquely identifies county and county equivalents in the United States). I used additional datasets to supplement my research. Primarily, I used population data from health.data.ny.gov [3] and data.ny.gov [4] - this allowed me to extract valuable features such as prevalence rates for confirmed influenza.

Motivation and Rationale.

Seasonal flu is deadly and widespread - more deadly and widespread than many viruses [5]. So far in this 2019 - 2020 flu season, an estimated 10,000 people have died and 180,000 hospitalized in the US [5]. Additionally, seasonal flu and pandemic flu are quite similar (except pandemic flu is much more deadly), thus, seasonal flu data can help us better understand and prepare for the pandemic flu for the next pandemic. New York State is an ideal state to analyze because it has locations that range from highly populated to rural. New York State also has high influenza-like illness activity [5]. Lastly, plentiful, and up-to-date data is available for this State.

Research Questions and Hypothesis.

The previous goal of my project was to make an interactive dashboard, but now the focus has shifted to attempting to answer some questions. 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 most commonly causes illness in humans [17]. Influenza type A viruses are the only influenza viruses known to cause pandemic flu because of their ability to change in two ways (antigenic drift and antigenic shift) [15]. Influenza strains of type A most commonly causes illness in humans. Therefore, I wanted to see if there was a relationship between the prevalence rates of influenza type A vs. influenza type B, and the overall prevalence rates in each county. I hypothesized that if there is a higher prevalence rate of influenza A than B, then the overall prevalence rate for the respective county will be higher. This is because research indicates that influenza type A viruses are responsible for approximately 75% of confirmed influenza cases [16]. With this trend in mind, I also decided to see if the percentage of confirmed cases of influenza A and B follow the same percentage pattern mentioned above (75% A and 25% B).

Related Work and Literature Review

As stated earlier – one of the proposed outcomes of the project was an interactive dashboard. Thus, the majority of research done on similar projects was on dashboards. There are various data dashboards that visualize flu data [6-8] as well as the spread of other viruses [9]. From these dashboards, I learned about what makes a dashboard intuitive, visually appealing, and easy to navigate. The importance of properly disseminating surveillance data is of greater focus in literature. This results in the optimization of a dashboard's impact on health officials' and policy makers' decisions and actions [10-11]. For example, Cheng et al. developed and implemented an influenza surveillance dashboard that displayed intuitive figures from multiple surveillance data streams per panel [10]. Their dashboard was applied to the influenza surveillance data in Hong Kong, while the proposed dashboard in this project will be data from New York State. The current New York State Flu Tracker Dashboard [6] only displays cases and trends in the more recent seasons (2016 - 2020). My dashboard will include all seasons (2010 - 2020). I will also attempt to implement supplemental data in the dashboard, such as the vaccination rate of healthcare workers with patient contact [12], or overall vaccine effectiveness per year [13]. Because the developers of the NYS Flu Tracker dashboard do not provide their methods, the main challenge will be developing a dashboard that works as seamlessly as theirs. I will provide the methods and code for creating the dashboard, allowing others to replicate my work and make improvements. Additionally, providing methodology will provide transparency for users who seek a deeper understanding of the data and dashboard.

Preliminary Exploratory Data Analysis

The preliminary exploratory data analysis revealed that the sum of all confirmed influenza cases (influenza type A, B, and unspecified) were highest in the 2017-2018 flu season (128,247 cases). The 2018-2019 season had the second highest number of cases (107,805 cases). Previous flu seasons (2009-2010 to 2016-2017) had a lower number of cases, with the highest being 64,765 cases in the 2016-2017 season. There are a variety of factors that may contribute to this difference. Factors include (but are not limited to) virulence, vaccine effectiveness, and vaccination rates. Another important factor to consider is the number (or availability) of laboratory tests for the flu. It is often the case that the flu is diagnosed based on symptoms alone, meaning there are more flu cases than those reported in this dataset. Regardless, this preliminary exploration revealed one insight that will be represented in my interactive dashboard.

Methods Overview

Part 1. Interactive Maps of Influenza Cases by County

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 [14].
    c. Based code off of Plotly Express documentation [14]. 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 [4]. Estimates are based on census counts (base population), intercensal and postcensal estimates [4].
    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.

Part 2. Analyzing Influenza Types A, B and Unspecified Prevalence Rates and Adding Male and Female Population Data.

The methods below are summarized – to see exactly what was done, see the following code provided on GitHub:

- Part 1 (corresponds to Step 1 below): https://github.com/kathuerta/DATA606/blob/master/Deliverable%204/Huerta_Delivery4_Part1_Flucases_Datasets.ipynb
- Part 2 (Step 2): https://github.com/kathuerta/DATA606/blob/master/Deliverable%204/Huerta_Delivery4_Part2_Analysis1.ipynb
- Part 3 (Step 3): https://github.com/kathuerta/DATA606/blob/master/Deliverable%204/Huerta_Delivery4_Part3_FandM_Population_Data.ipynb
- Part 4 (Step 4): https://github.com/kathuerta/DATA606/blob/master/Deliverable%204/Huerta_Delivery4_Part4_Analysis_2017to2018_Season%20(2).ipynb
- Part 5 (Step 5): https://github.com/kathuerta/DATA606/blob/master/Deliverable%204/Huerta_Delivery4_Part5_MapAvsB_GitHubFriendly.ipynb

1. Added influenza type (A, B, and Unspecified) counts to seasonal subsets.
    a. Followed same methods for making subsets by year, without grouping or sorting the data because it would affect the column of interest ('Disease').
    b. Split each influenza type into separate columns showing the count of each type of flu for each county (was originally all under one column).
    c. Extracted and added these features (the counts for type A, B, and Unspecified) to each respective seasonal subset and saved the new data frames as csv files.
2. Analyzed the new subsets (containing the counts of each influenza type for each county).
    a. Made stacked bar plots for each season – comparing the count of influenza type A vs B for each county for each influenza season (fig. 2).
    b. Calculated and visualized the percentage of type A and type B influenza for each county for each season – looking at the distribution via boxplots and histograms.
        i. Made boxplots and histograms for the data normalized over population data (fig. 4-5)
    c. Chose a single season of interest for remaining analysis (due to time constraints).
        i. Chose the 2017-2018 influenza season
        ii. Performed a paired sampled t-test on the percentage of influenza A vs. B (H_0 = difference of means between the two samples are 0, and H_1=The mean difference between the two

samples are not 0). Results indicated acceptance of the null hypothesis. However, further analysis was not done thus this evidence is inconclusive and not applicable to the project.

3. Added county-level gender population data to 2017-2018 season subset.
    a. Used dataset "New York State Population Data: Beginning 2003" from health.data.ny.gov [3].
    b. Did not have time to do much analysis on gender, but kept and shared methods/code for future use and possible correlations for remaining parts of delivery 4.
4. Performed further analysis on 2017-2018 season (that now contains features not in the original dataset).
    a. Calculated and added type A, B and Unspecified influenza type prevalence rates as new columns.
    b. Calculated/Looked at the mean, standard deviation, variance, and other summary statistics related to prevalence rates of the three influenza types. However, more focus was on the prevalence rates of influenza A and B (table 2).
    c. Visualized the distribution of data using boxplots (fig. 4), and histograms with kernel density estimates (KDE) using seaborn distribution plots (fig. 5)
    d. Visualized the percentage of the three influenza types as well as just influenza type A vs. B via pie chart (fig. 6)
    e. Briefly explored potential correlations with a heatmap.
    f. Looked at the stacked bar chart of prevalence rates of influenza A vs. B for each county (fig. 3).
5. Visualized influenza A vs B prevalence rates per county for the 2017-2018 season using interactive maps (fig. 6-8).
    a. Followed same methods from part 1.

Results/Figures

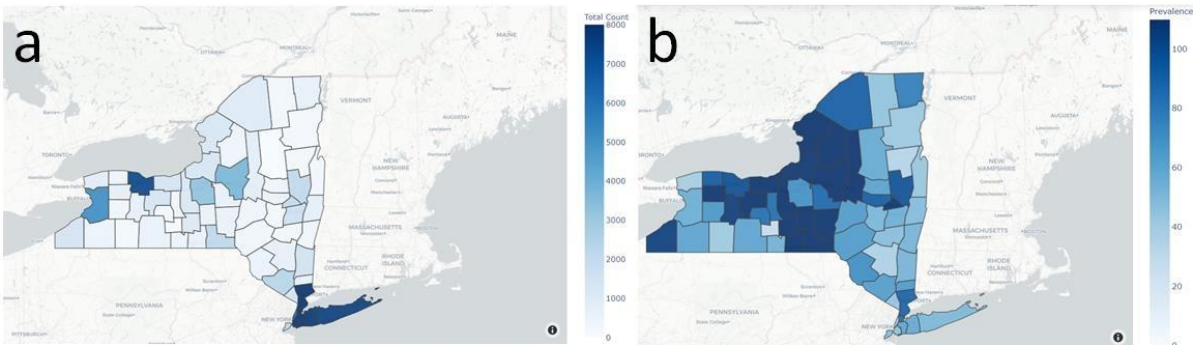

Figure 1. Side by side comparison of the two density maps from the 2017-2018 influenza season. The density map illustrating the total count of confirmed cases **(a)** is quite different from the map illustrating the prevalence rate of confirmed influenza cases per 10,000 people **(b)**.

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

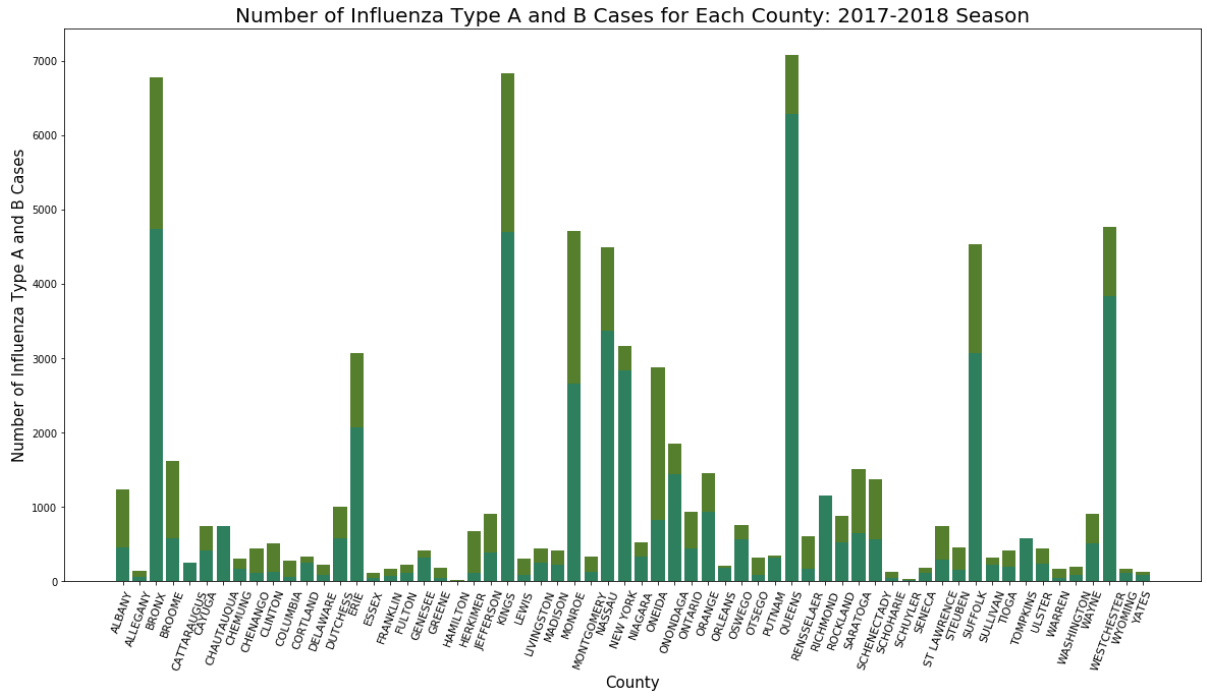

Figure 2. Stacked bar chart of the count of influenza A (green) and influenza B (blue). Interestingly, the counties with higher counts of confirmed flu cases seem to have more influenza B cases than influenza A cases. Although it may not be apparent in this visualization, there are more influenza A cases than B for this season (Sum of A = 77,031 and sum of B = 50,215).

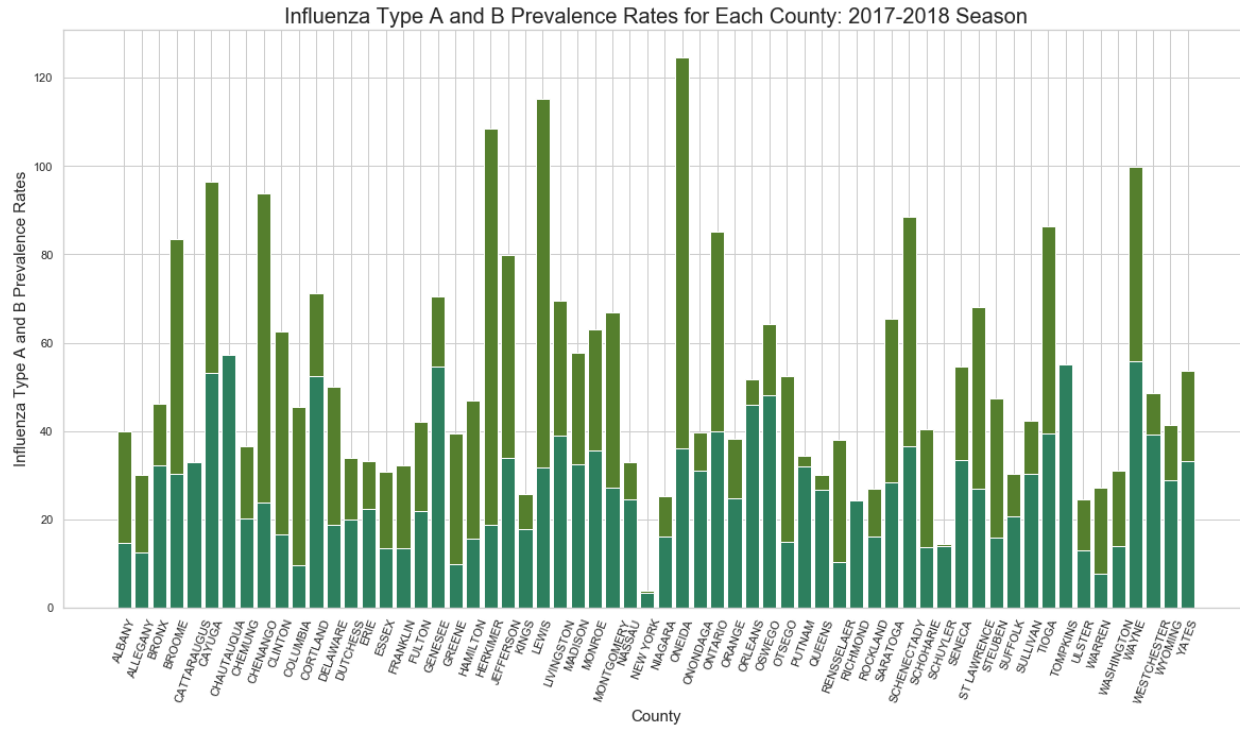

Figure 3. Stacked bar chart of the prevalence rate of influenza A (green) and influenza B (blue). As expected, the counties with higher prevalence rates of confirmed flu cases seem to have higher prevalence rates for influenza A than influenza B.

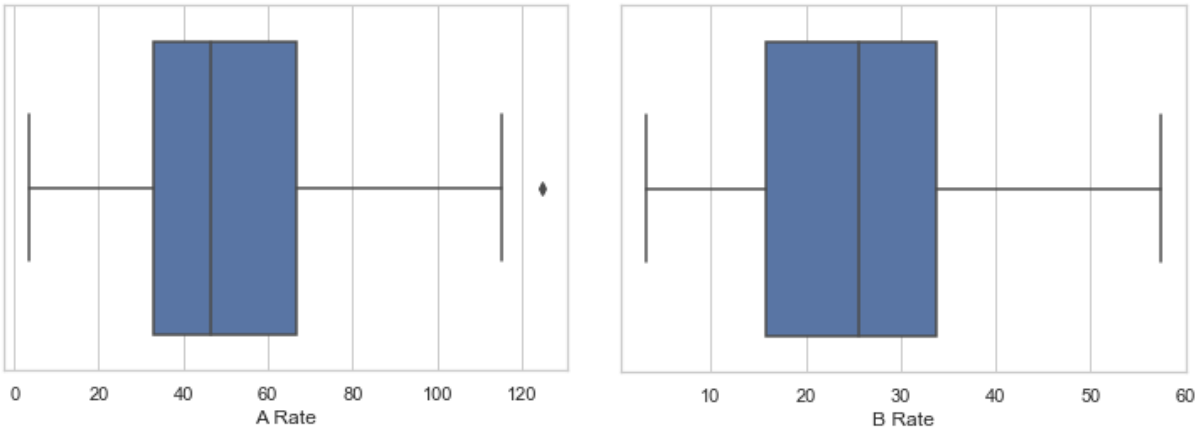

Figure 4. Box plots visualizing the summary statistics of the prevalence rates of influenza A (left) and influenza B (right). The x-axis shows the prevalence rate of the respective influenza type (count of cases per population of 10,000). The boxplot for influenza A prevalence rates indicates a potential outlier, which was found to be Oneida County (prevalence rate = 159.894). Although this county affects the distribution, the data is not incorrect and therefore should not be excluded.

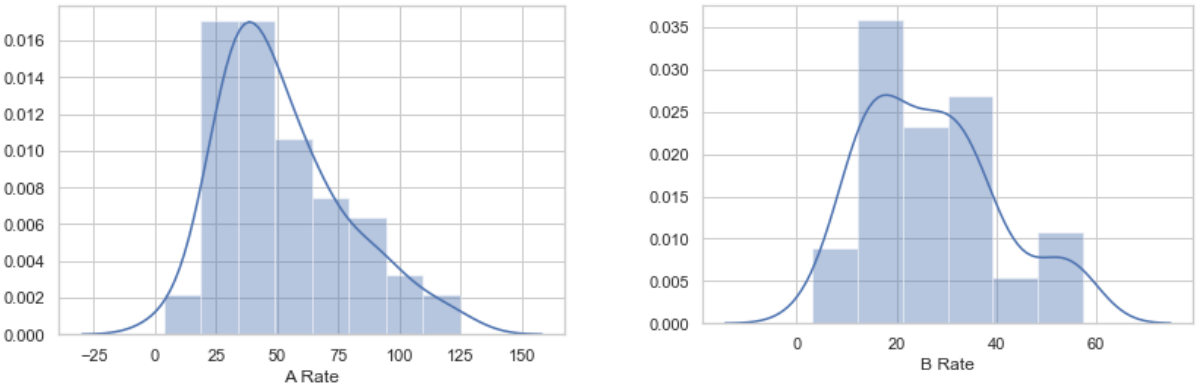

Figure 5. Distribution plots of the prevalence rate of influenza A (left) and influenza B (right). The distribution plot is a histogram showing the frequency (y-axis) of observed prevalence rates (x-axis) with a kernel density estimate (KDE) fit. Neither prevalence rates have a "normal" distribution. This is likely due to the grouping of many observations into only 62 rows (grouping by counties).

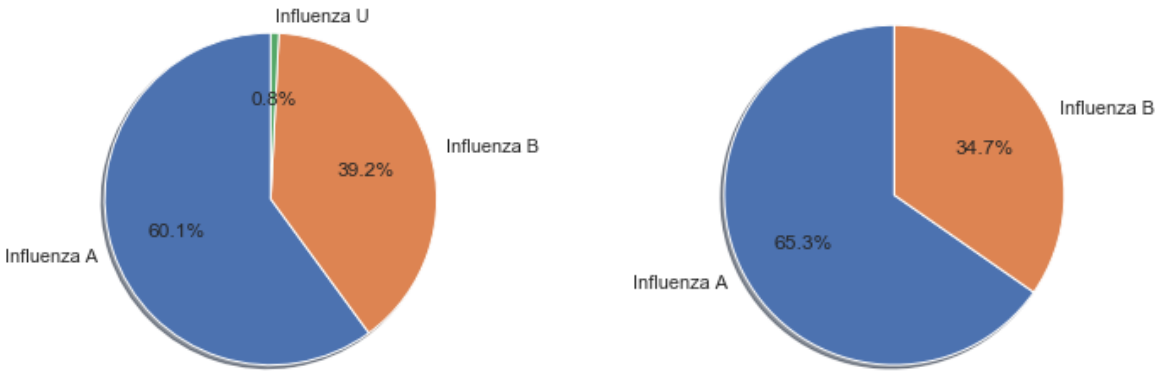

Figure 6. Pie charts visualizing the percentage (out of all three influenza categories) of each influenza type (left) and the percentages of influenza A and B only (calculated based on the total count of A+B, thus, excluding the additional unspecified cases). Influenza U (left) represents cases that did not have a diagnosis that specified the type of influenza.

Table 2. Prevalence Rate* Summary Statistics for Influenza A and B (2017-2018 Season).

	A	B
Mean	52.2840	27.0950
Standard Deviation	25.6870	13.4580
Variance	659.811	181.114
Minimum	3.67800	3.28300
25%	32.9310	15.8270
50%	46.4680	25.6440
75%	66.6030	33.6870
Maximum	124.669	57.3440

Numbers are rounded up or down to the third decimal.
*Prevalence Rate = Number of cases per population of 10,000

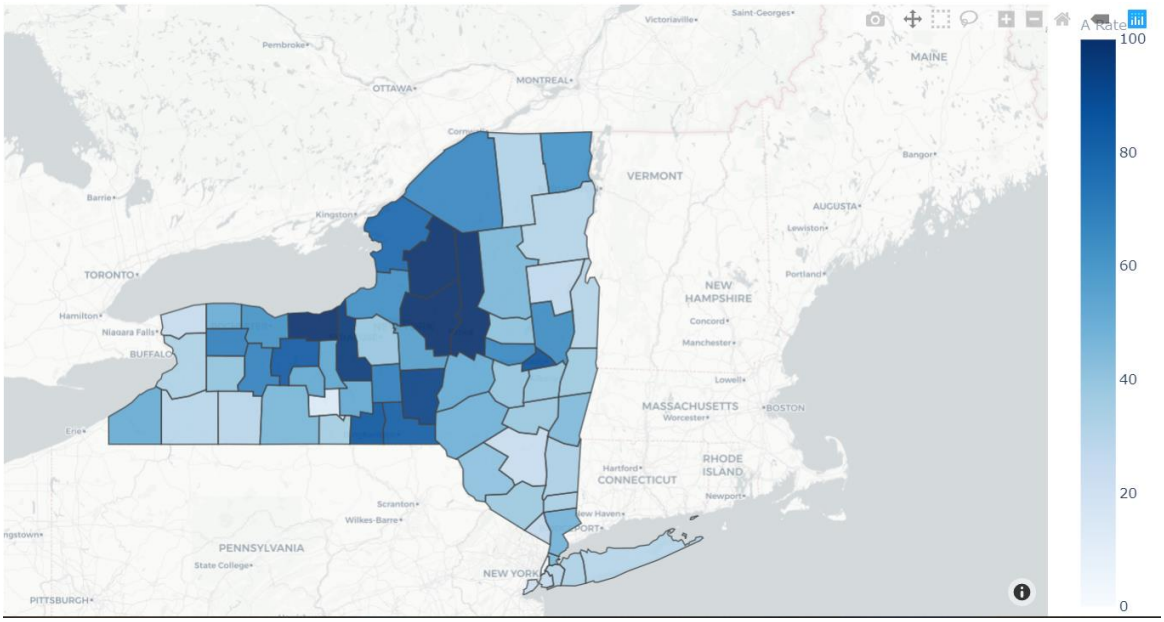

Figure 7. Density map of influenza A prevalence rates per county in the 2017-2018 season. The scale of coloring is from 0 to 100, meaning the lower the prevalence rate, the lighter the color blue. The higher the prevalence rate, the darker the color blue will be.

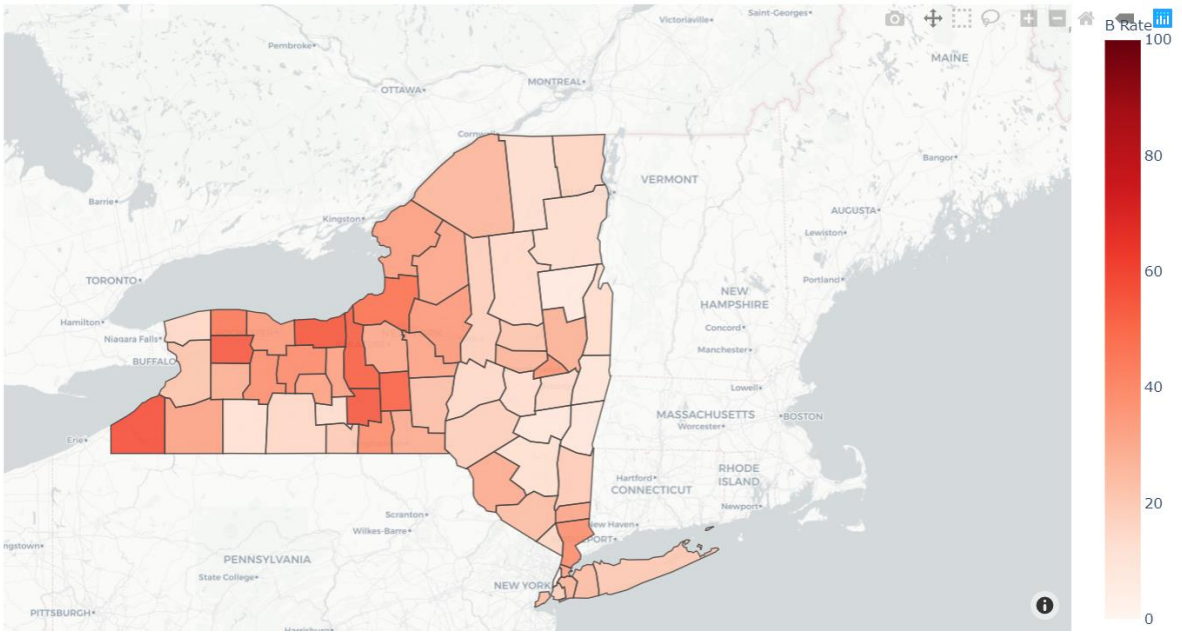

Figure 8. Density map of influenza B prevalence rates per county in the 2017-208 season. The scale of coloring is from 0 to 100, meaning the lower the prevalence, the lighter the color red. The higher the prevalence rate, the darker the color red will be.

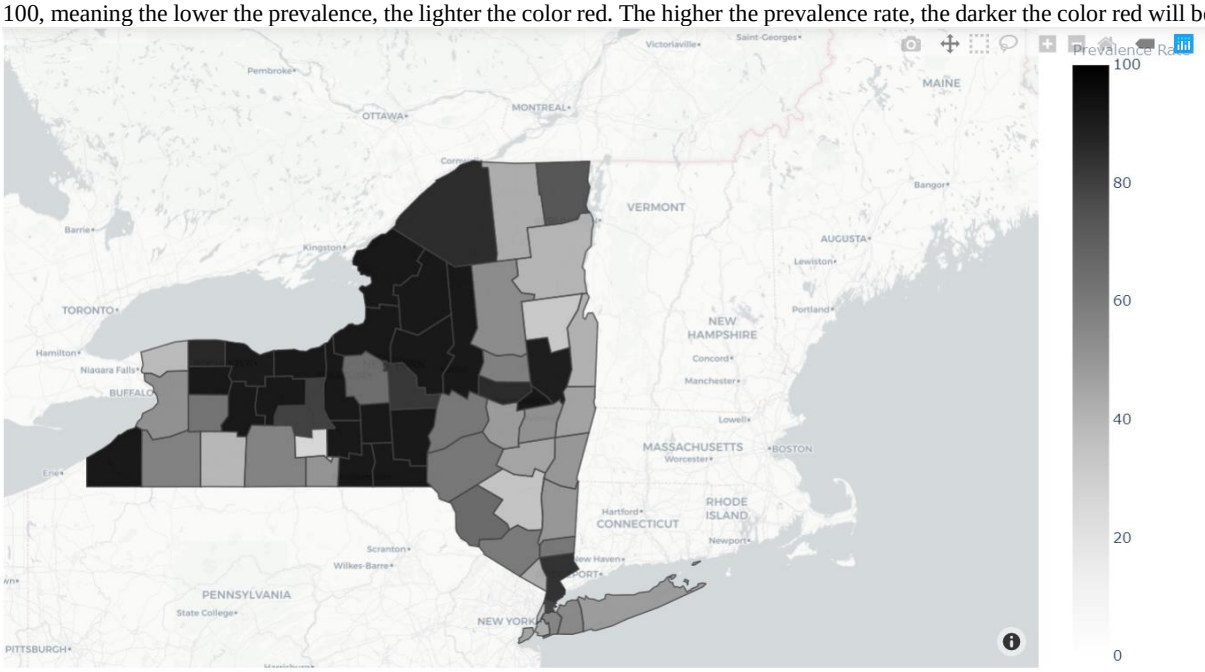

Figure 9. Density map of the total prevalence rate (of any type of influenza) per county in the 2017-208 season. The scale of coloring is from 0 to 100, meaning the lower the prevalence, the lighter the color grey. The higher the prevalence rate, the darker the color grey will be.

Discussion

Part 1. Interactive Maps

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 [4]), 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.

Part 2. Exploring Influenza Types

The main subset explored was the 2017-2018 season. After analyzing the 2017-2018 influenza season, results indicated that for this season in particular, influenza type A cases were responsible for approximately 65% of confirmed influenza cases and influenza type B cases were responsible for approximately 35% of confirmed influenza cases (excluding confirmed influenza cases that did not have a specified type of influenza), seen in figure 6. More analysis of remaining seasons, and statistical analysis would need to be performed to be able to confirm the actual percentages. However, we can say that this season deviates 10% from the general claim that influenza type A viruses are responsible for 75% of confirmed influenza cases, and type B responsible for 25% [16].

Seen in figure 3, it appears that counties with the highest prevalence rates overall tend to have higher influenza A prevalence rates than influenza B. This supports my initial hypothesis for the 2017-2018 season only. However, with non-normalized data (just the counts of confirmed cases), the counties with higher counts of influenza cases (in the 2017-2018 season) have a higher count of influenza type B than type A (fig. 2). Future work should be done to investigate why more populated counties (such as Bronx, New York, Kings, and Queens) have a higher count of type B cases as well as higher prevalence rates of influenza type B (figure 3, and figures 7-9). In contrast, future work should be done to investigate why in the 2017-2018 season, counties with high overall prevalence rates have a greater difference between the prevalence rate of influenza type A and B (where, it appears that influenza A prevalence rates tend to be greater than influenza B prevalence rates).

Conclusion

Although many goals in this project were not achieved, a lot was learned not only about the data, but also in the challenges and solutions associated with preparing the data, exploring the data, and interpreting results. More questions have surfaced than answers – providing various avenues for future work geared towards making claims about influenza cases in New York State, that are backed by strong statistical evidence. Unfortunately, I was unable to make any substantial conclusions. This project has inspired me to spend more time improving my knowledge and ability to take the questions and observations raised from this project – and discover answers validated by statistical analysis and hypothesis testing.

References Cited

1. HealthData.gov. Influenza Laboratory-Confirmed Cases By County: Beginning 2009-10 Season. (2020). [Online], Available: https://healthdata.gov/dataset/influenza-laboratory-confirmed-cases-county-beginning-2009-10-season. [Accessed January 10, 2020].
2. New York State Department of Health. Influenza Activity, Surveillance, and Reports. (2020).
3. Health Data NY, "New York State Population Data: Beginning 2003", 2020. [Online], Available: https://health.data.ny.gov/Health/New-York-State-Population-Data-Beginning-2003/e9uj-s3sf. [Accessed April 15, 2020].
4. 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-forNew-York-State-and/krt9-ym2k. [Accessed March 30, 2020].
5. Centers for Disease Control and Prevention. Weekly U.S. Influenza Surveillance Report. (2020).
6. NYS Health Connector, "New York State Flu Tracker", 2020. [Online], Available: https://nyshc.health.ny.gov/web/nyapd/new-york-state-flu-tracker. [Accessed Feb.18, 2020].
7. FluView Interactive, "National, Regional, and State Level Outpatient Illness and Viral Surveillance", Centers for Disease Control and Prevention, 2020. [Online], Available: https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html. [Accessed Feb. 18, 2020].
8. L. VanWhy, and P. Galebach, "The athenaInsight Flu Dashboard," athenahealth Inc., September 20, 2019. [Online], Available: https://www.athenahealth.com/insight/flu-dashboard-2017-2018 [Accessed Feb. 18, 2020].
9. L. Gardner, "Mapping 2019-nCoV", Johns Hopkins Whiting School of Engineering| Center for Systems Science and Engineering, January 23, 2020. [Online], Available: https://systems.jhu.edu/research/public-health/ncov/. [Accessed Feb. 18, 2020]. Full dashboard Available: https://gisanddata.maps.arcgis.com. [Accessed Feb. 14, 2020].
10. C. Cheng, D. Ip, B. Cowling, L. Ho, and E. Lau, "Digital dashboard design using multiple data streams for surveillance with influenza surveillance as an example", J Med Internet Res. 2011 Oct-Dec; 13(4): e85. Published online 2011 Oct 14, doi: 10.2196/jmir.1658. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3222192/. [Accessed Feb. 19, 2020].
11. S. Hamid, L. Bell, and E. Dueger, "Digital dashboards as tools for regional influenza monitoring", WPSAR Vol 8, No 3, 2017, doi: 10.5365/wpsar.2017.8.2.003.
12. Health Data NY, "Influenza Vaccination Rates for Health Care Personnel: Beginning 2012-13", 2019. [Online Dataset], Available: https://health.data.ny.gov/Health/Influenza-Vaccination-Rates-for-Health-Care-Person/jpkp-z76p. [Accessed Feb. 19, 2020].
13. Centers for Disease Control and Prevention. "Vaccine Effectiveness Studies", 2019. [Online], Available: https://www.cdc.gov/flu/vaccineswork/effectiveness-studies.htm. [Accessed January 28, 2020].
14. Plotly Graphing Libraries, "Mapbox Choropleth Maps in Python", 2020. [Online], Available: https://plotly.com/python/mapbox-county-choropleth/. [Accessed March 15, 2020].
15. 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].
16. 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].
17. Niad.nih.gov. "Influenza Basic Research", 2017. [Online], Available: https://www.niaid.nih.gov/diseases-conditions/influenza-basic-research. [Accessed May 6, 2020].

Project Links

- GitHub Repository: https://github.com/kathuerta/DATA606
- Presentation Links
  - Presentation I: https://www.youtube.com/watch?v=U-xCLGhcbqE&t=97s
  - Presentation II: https://www.youtube.com/watch?v=3ER1HsfmJqI
  - Presentation III: https://www.youtube.com/watch?v=aiwdoGZqZIA&feature=youtu.be
  - Presentation IV: https://youtu.be/cWpPs0SXmmw
- DATA 606 Website: https://sites.google.com/umbc.edu/data606/home/katherine-huerta?authuser=0