Analysis of New York State Influenza Data with Python and Interactive Visualizations

Katherine Huerta

Deliverable 4

Data 606 SP 2020

https://www.nfid.org/infectious-diseases/influenza-flu/

Introduction

Project and Purpose: Explore seasonal influenza data to glean insights regarding trends at the county level in NYS.

Main Dataset: https://healthdata.gov/dataset/influenza-laboratory-confirmed-cases-county-beginning-2009-10-season)

Motivation and Rationale:

- Why Influenza?
  - Seasonal flu = deadly and widespread
  - Seasonal flu and pandemic flu are similar
- Why NYS?
  - Diverse populations
  - High influenza-like illness activity
  - Data!

Preliminary Exploratory Data Analysis

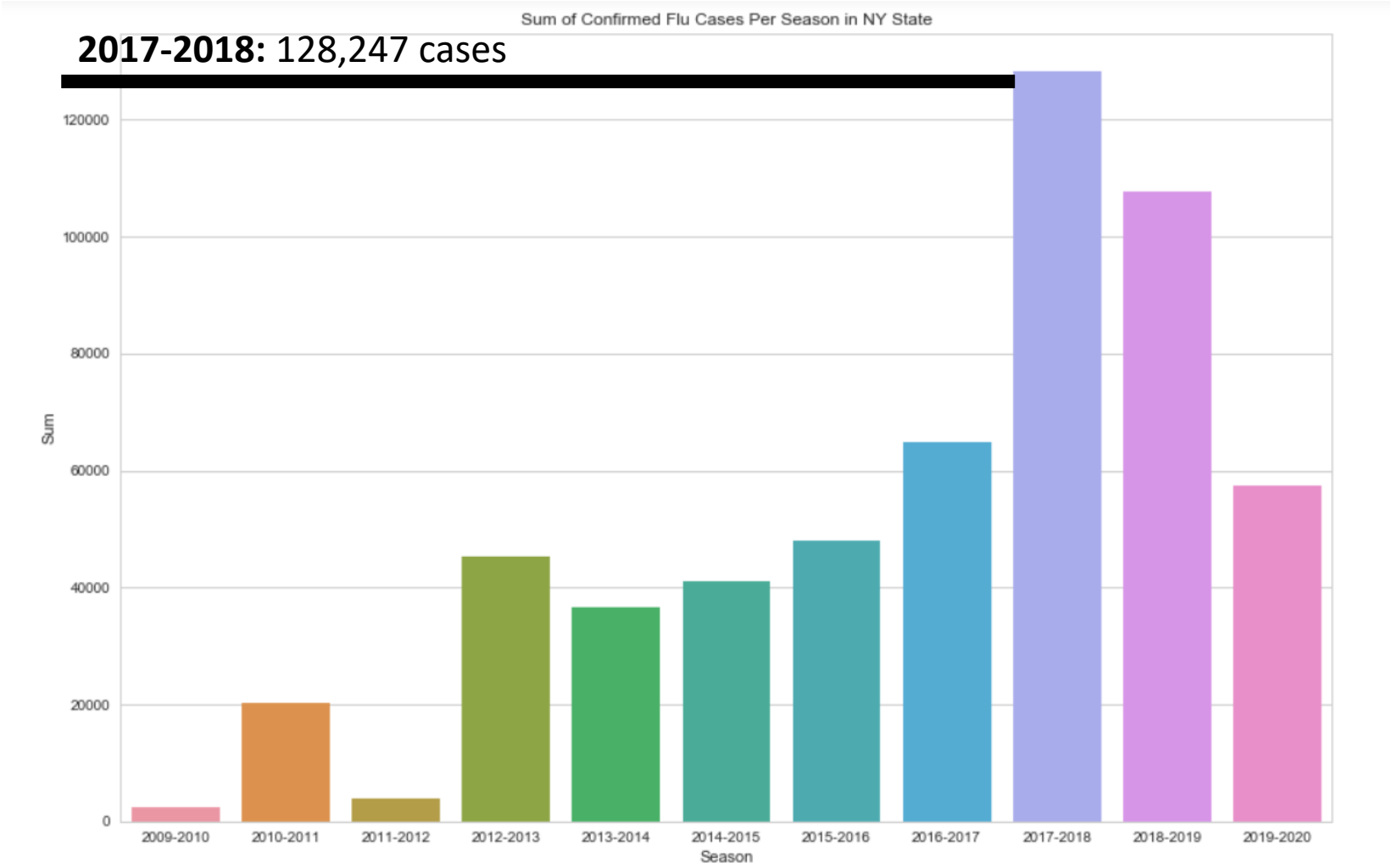


Figure 1. Bar plot displaying the sum of confirmed flu cases (influenza A, B, and unspecified) for each flu season in New York State. The x-axis represents the sum (ranging from 0 – 140,000 cases, increments of 20,000 cases) the y axis represents the season (2009 → present). The 2019 – 2020 flu season is not over until May, so the number of confirmed flu cases will change each week.

Interactive Maps: Number of Lab Confirmed Flu Cases vs. County

2016-2017: 64,765 cases

2017-2018: 128,247 cases

2018-2019: 107,805 cases

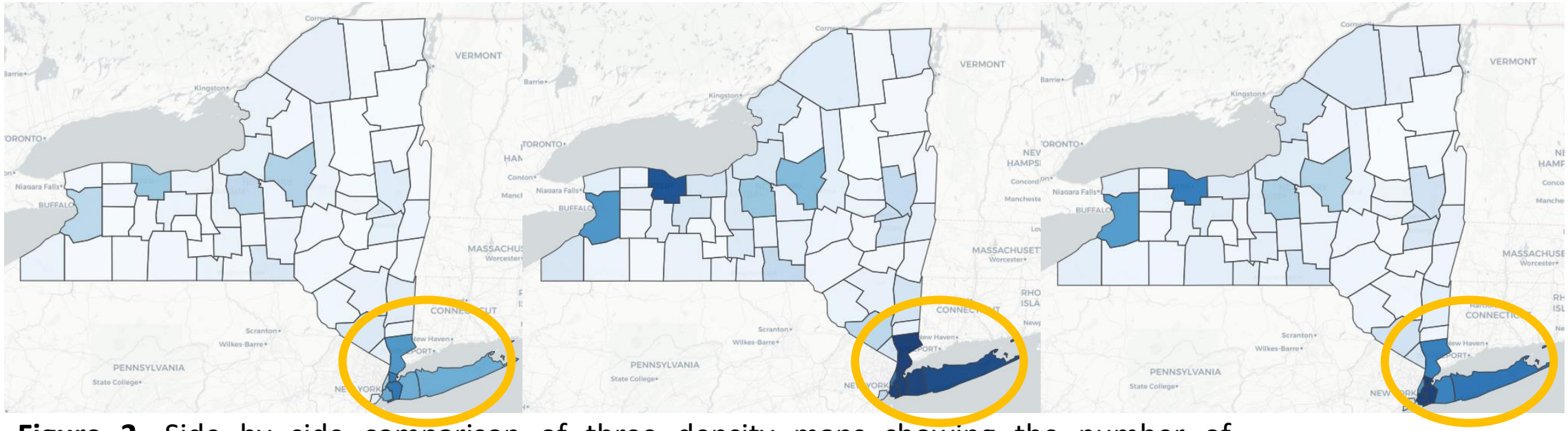


Figure 2. Side by side comparison of three density maps showing the number of confirmed influenza cases for each county in the 2016-2017 (left), 2017-2018 (middle), and 2018-2019 (right) influenza seasons. All coloring is based on the same scale. The darker the color – the higher the case count, while the lighter the color – the lower the case count.

Adding Population Data to Subsets

Population data from: https://data.ny.gov/Government-Finance/Annual-Population-Estimates-for-New-York-State-and/krt9-ym2k

```
def CalcPrevalenceNumerator(df):
    numerator=[]

    for index, row in df.iterrows():
        numerator.append(row.Count*10000)

    return numerator
```

```
def CalcPrevalenceRate(df):
    rate=[]

    for index, row in df.iterrows():
        rate.append(row.x/row.Population)

    return rate
```

$$Rate = \left(\frac{10{,}000 x Count}{Population}\right)$$

	County	Count	FIPS	Population
0	ALBANY	1708	36001.0	304596
1	ALLEGANY	205	36003.0	48800
2	BRONX	11749	36005.0	1397335
3	BROOME	2214	36007.0	199363
4	CATTARAUGUS	492	36009.0	79815

Adding Population Data to Subsets


```
def CalcPrevalenceNumerator(df):
    numerator=[]

    for index, row in df.iterrows():
        numerator.append(row.Count*10000)

    return numerator
```

```
def CalcPrevalenceRate(df):
    rate=[]

    for index, row in df.iterrows():
        rate.append(row.x/row.Population)

    return rate
```

$$Rate = \left(\frac{10{,}000 x Count}{Population}\right)$$

County	Count	FIPS	Population	Rate
ALBANY	1708	36001.0	304596	56.074275
ALLEGANY	205	36003.0	48800	42.008197

Interactive Maps: The Power of Normalization

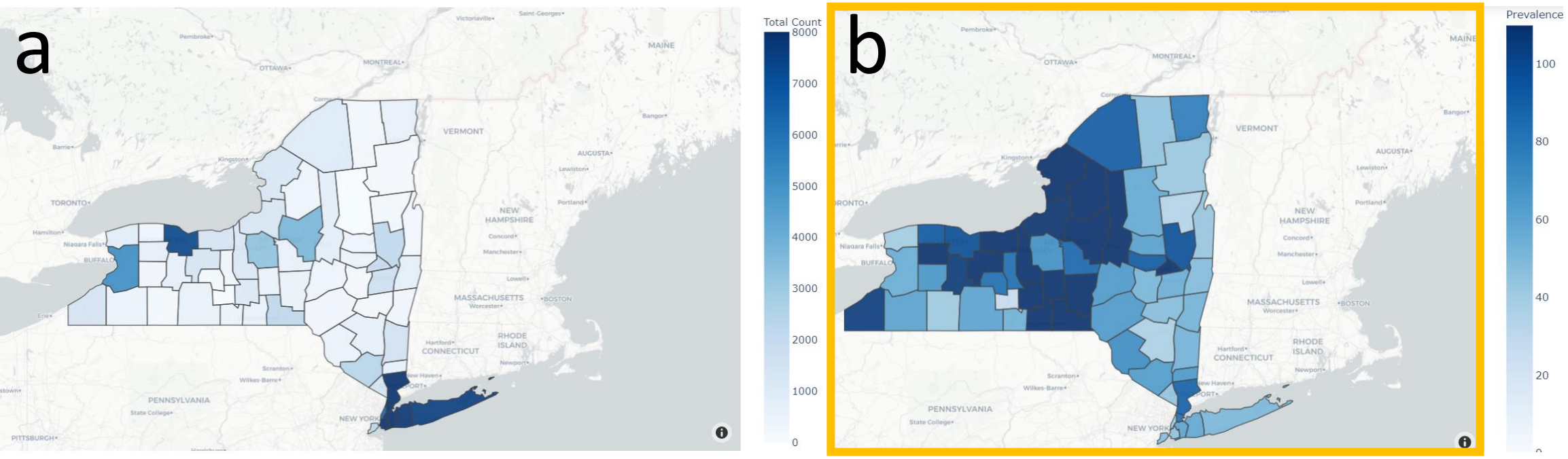


Figure 3. Side by side comparison of the two density maps from the 2017-2018 influenza season. The density map illustrating the total count of confirmed cases (a) is quite different from the map illustrating the prevalence rate of confirmed influenza cases per 10,000 people (b).

New Questions

Prevalence Rates of Influenza A vs. Influenza B:

- Influenza A
  - Responsible for ~75% of confirmed influenza cases [1].
  - Influenza A viruses most commonly cause illness in humans [1].
  - Only influenza virus known to cause pandemics [1].
- Influenza B
  - ~25% of confirmed influenza cases [1].

Does this coincide with NYS prevalence rates of influenza A and B?

- Is there a relationship between the prevalence rates of influenza type A and B, and the overall prevalence rate in counties?
- Does the 2017-2018 Season for NYS follow this 75% vs. 25% trend?

Hypothesis: If there is a higher prevalence rate of type A cases in a county, the prevalence rate will be higher.

1. 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].

Data Preparation

County	CDC_Week	Week_Ending_Date	Disease	Count
DUTCHESS	52	12/28/2019	INFLUENZA_B	21
ALBANY	47	11/23/2019	INFLUENZA_A	4
LIVINGSTON	47	11/23/2019	INFLUENZA_B	0
CORTLAND	50	12/14/2019	INFLUENZA_UNSPECIFIED	0
NIAGARA	40	10/05/2019	INFLUENZA_UNSPECIFIED	0

← Before

County	Count	Influenza A	Influenza B	Influenza Unspecified	Prevalence Rate	Total Population	A Rate	B Rate	U Rate
ALBANY	1708	1234	456	18	56.074275	309612	39.856336	14.728111	0.581373
ALLEGANY	205	141	58	6	42.008197	46894	30.067813	12.368320	1.279481
BRONX	11749	6784	4740	225	84.081484	1471160	46.113271	32.219473	1.529405
BROOME	2214	1618	584	12	111.053706	193639	83.557548	30.159214	0.619710
CATTARAUGUS	492	235	255	2	61.642548	77348	30.382169	32.967885	0.258572

← After

Results

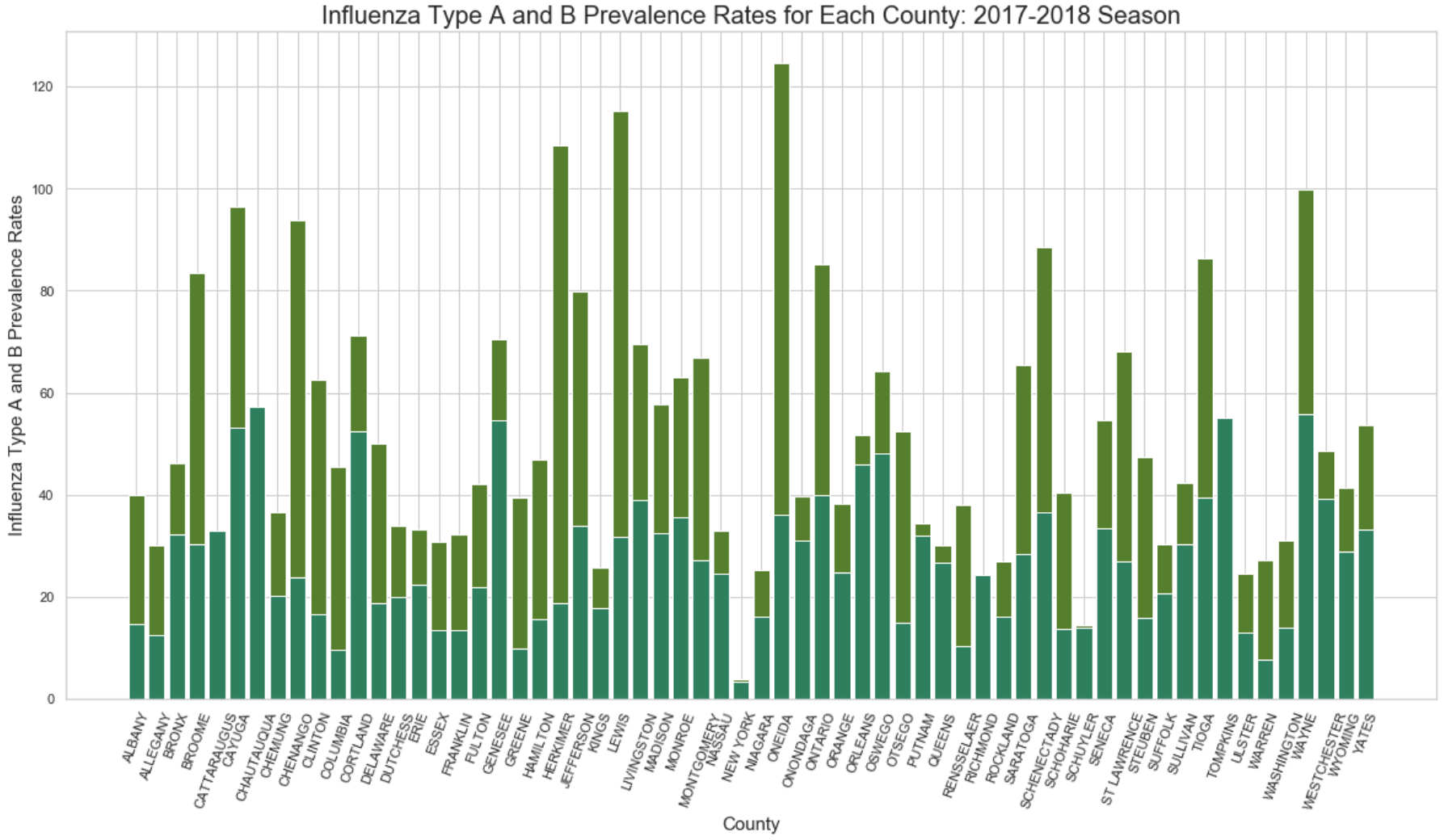


Figure 4. Stacked bar chart of the prevalence rate of influenza A (green) and influenza B (blue). As expected, the counties with higher prevalence rates of confirmed flu cases seem to have higher prevalence rates for influenza A than influenza B.

Results

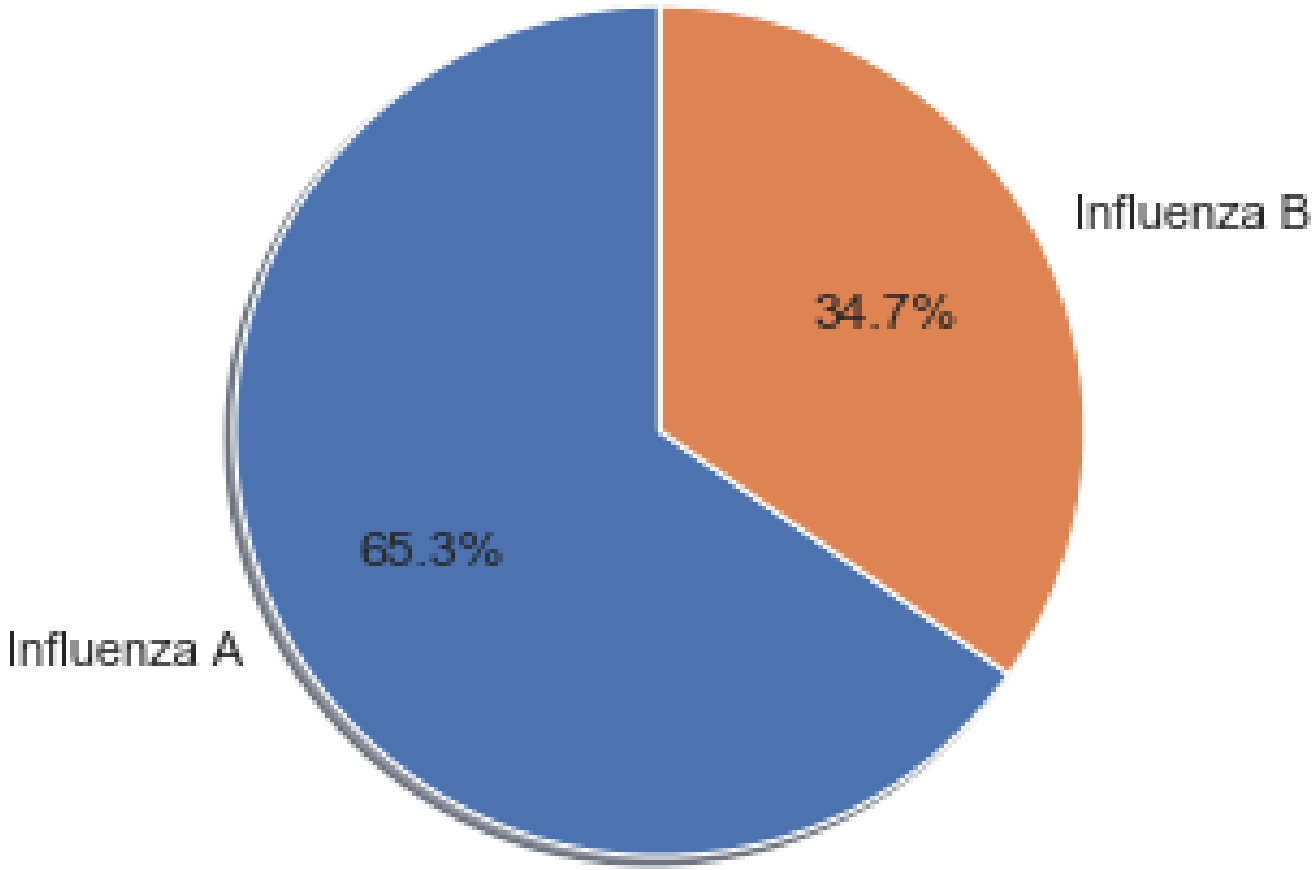


Figure 5. Pie chart visualizing the percentage of influenza A and influenza B confirmed cases across all counties in the 2017-2018 season.

Results

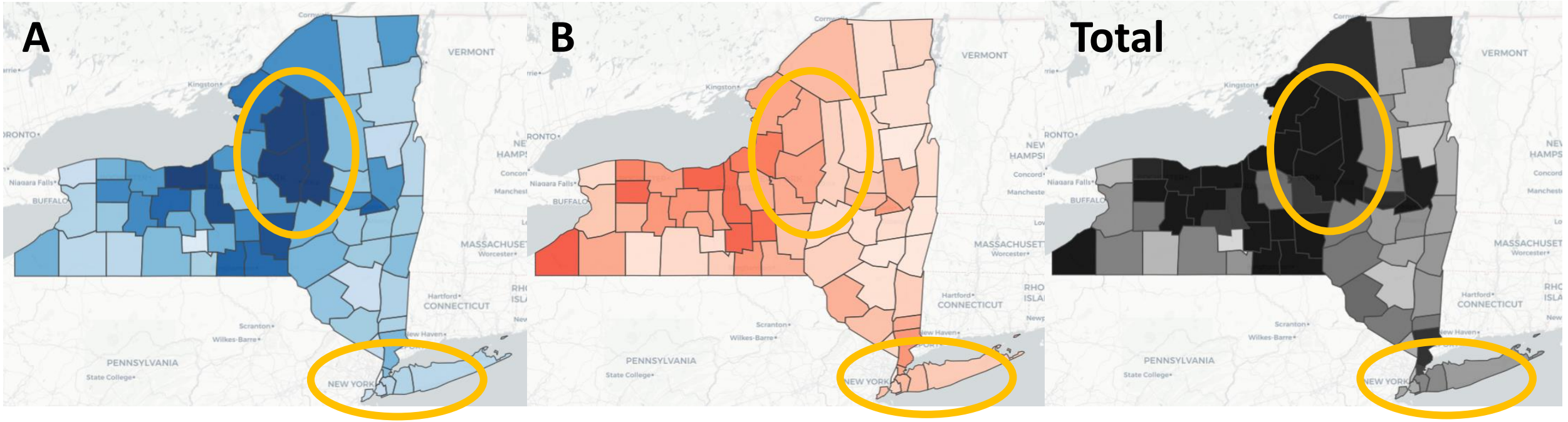


Figure 6. Density map of prevalence rates for influenza A (blue), B (red), and total (grey) per county in the 2017-2018 season. The scale of coloring from 0 to 100 (0 prevalence rate – prevalence rate of 100), meaning the lower the prevalence rate, the lighter the color. The higher the prevalence rate, the darker the color.

Discussion

2017-2018 Season:

- Results support my hypothesis that counties with higher prevalence rates tend to have prevalence rates of influenza A > influenza B
- Influenza A is responsible for approximately 65% of confirmed influenza cases, while influenza B is responsible for approximately 35% of confirmed influenza cases (~10% deviation)
- More statistical testing is required to make any valid conclusions that can be generalized to any population!
  - Standard deviation from mean across counties need to be investigated – variance appears to be high.

Table 1. Prevalence Rate* Summary Statistics for Influenza A and B (2017-2018 Season).

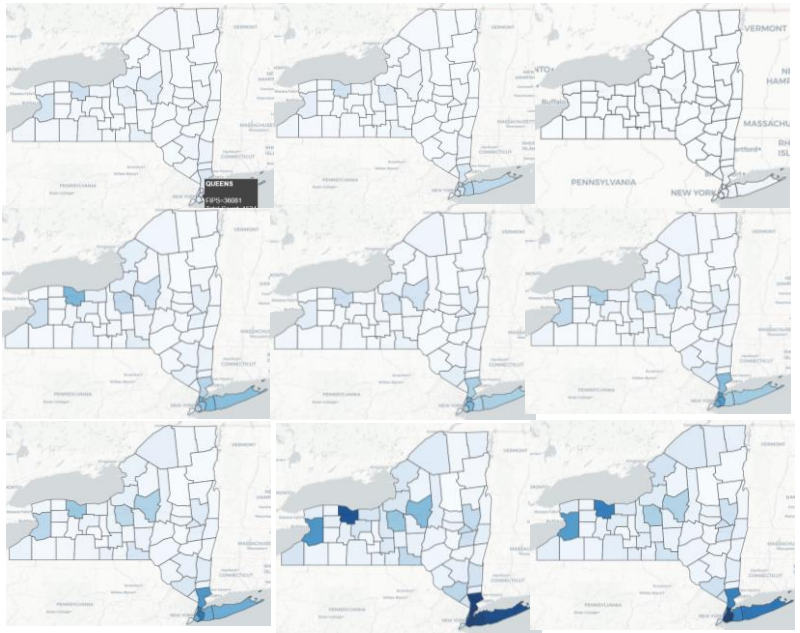
	A	B
Mean	52.2840	27.0950
Standard Deviation	25.6870	13.4580
Variance	659.811	181.114
Minimum	3.67800	3.28300
25%	32.9310	15.8270
50%	46.4680	25.6440
75%	66.6030	33.6870
Maximum	124.669	57.3440

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

Discussion

2017-2018 Season:

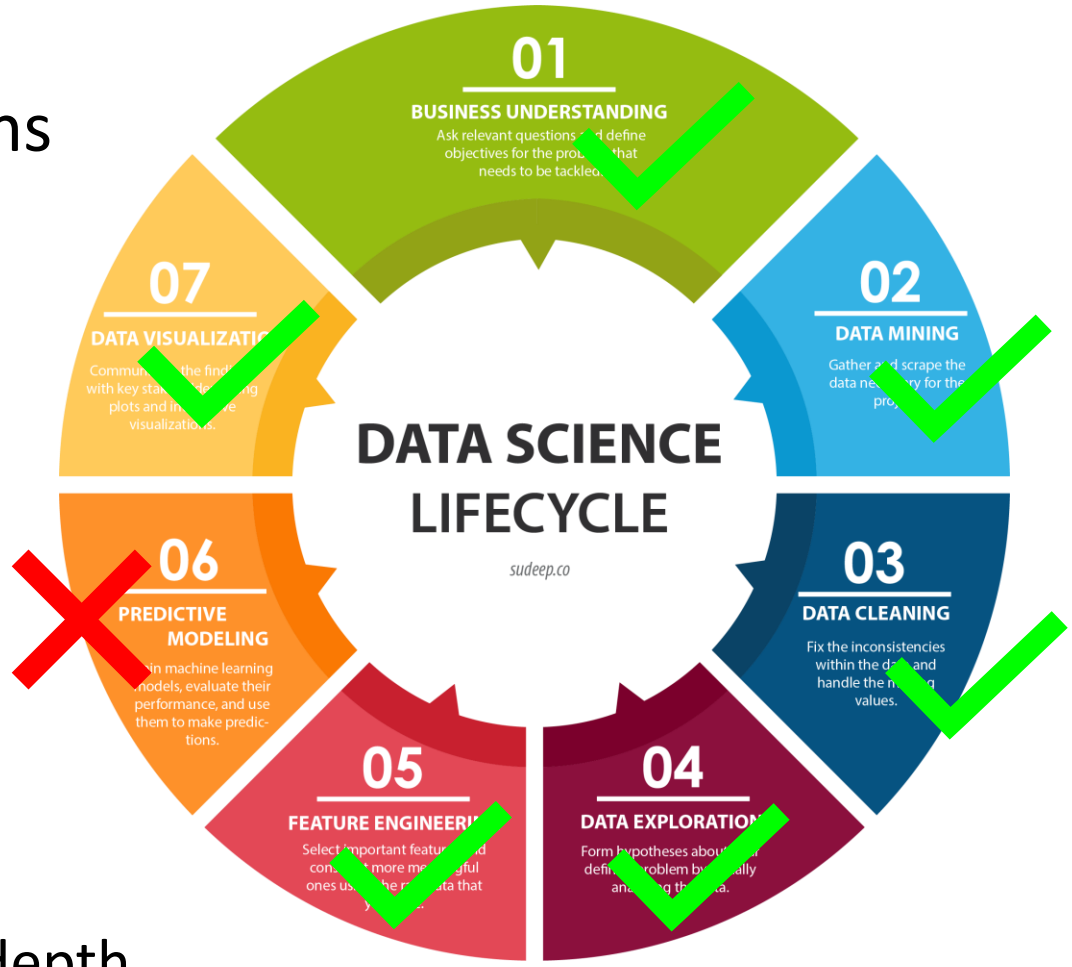
- Results support my hypothesis that counties with higher prevalence rates tend to have prevalence rates of influenza A > influenza B
- Influenza A is responsible for approximately 65% of confirmed influenza cases, while influenza B is responsible for approximately 35% of confirmed influenza cases (~10% deviation)
- More statistical testing is required to make any valid conclusions that can be generalized to any population!
  - Standard deviation from mean across counties need to be investigated – variance appears to be high.
  - Need to look at all seasons and compare trends.



Conclusion

What was achieved and valuable lessons learned:

- Feature selection and extraction
- Interactive maps
  - Normalize data!
- Powerful visualizations
  - Identify trends
    - Prevalence rates: Influenza A vs. B
  - Raised more questions than answers
- Ambitious goals vs. time constraints
  - Most seasons were not investigated in depth
  - Quantifying significance of results required – requires more time in the library
- Statistics is key for future work



http://sudeep.co/data-science/Understanding-the-Data-Science-Lifecycle/

Thank you!
Questions or feedback?

uv94014@umbc.edu

https://github.com/kathuerta/DATA606

https://www.nfid.org/infectious-diseases/influenza-flu/