Understanding Trends in Influenza Strains Over Time

Ashkan Nikfarjam

Objectives

I wanted to understand in my project how influenza trends have changed over time in California.

Dataset

The dataset comes from the California Government Influenza Surveillance data. It is updated annually. The current data from public health laboratories runs from 2009-2020.

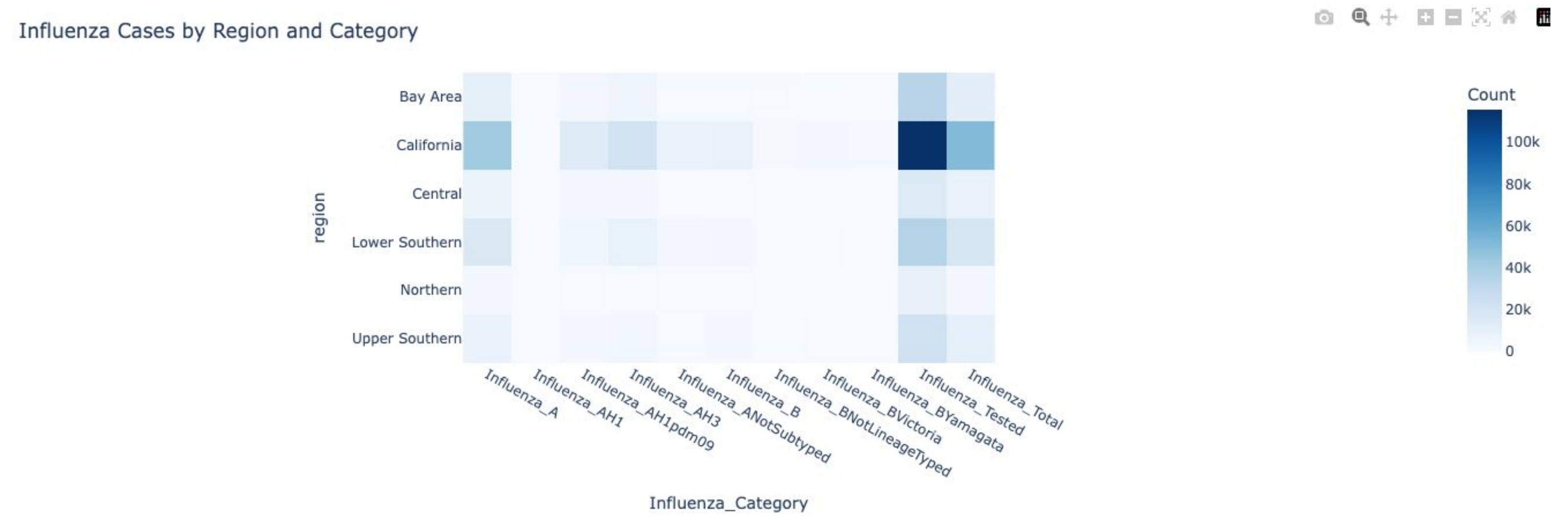
The raw dataset contains 30073 entries, with each row containing the following columns:

- `season`
  - string
  - formatted as `{YEAR_1}-{YEAR_2}` (e.g. 2009-2010)
  - a season runs from the 40th week to the 39th week
- `date_code`
  - number
  - formatted as `{YEAR}{WEEK}` (e.g. 200952)
- `weekending`
  - date
  - formatted as `MM/DD/YYYY` (e.g. 10/04/09)
  - this is the last day in the date_code week
- `region`
  - string
  - one of: “Bay Area”, “Lower Southern”, “Central”, “Northern”, “Upper Southern”, “California”
  - each region is defined as a set of counties
- `Influenza_Category`
  - text
  - formatted as `Influenza_{TYPE}` (e.g. Influenza_B)
- `Count`
  - number
  - number of specimens meeting the criteria for the category lineage

Results

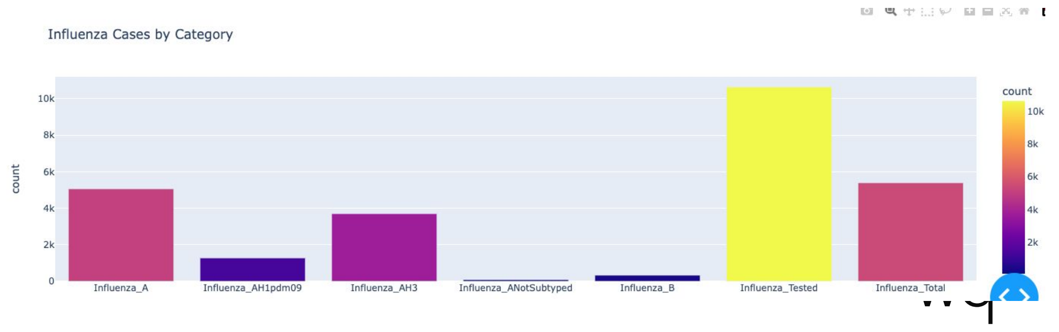
I was interested in finding out if i could use my visualization to be able to assess each region total performance in management of this virus. I used bash script, employing awk, to group and aggregate the data. I used python libraries and packages to create a heatmap that represented the death rate based on region and the type of Influenza.

Based on the result of my visualization I found out, the Bay Area and the Lower Southern california region had the highest death rate. But it also because the density of the population in these regions.

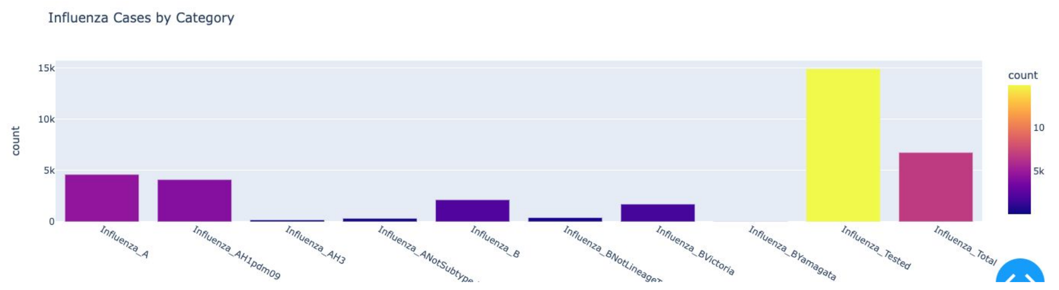


I also wanted to analyze the year-over-year trends for each influenza strain. To do this, I grouped data by year and flu strain. This helped me find that while strain A is overall more prevalent than strain B, strain B has been increasing in ratio with A.

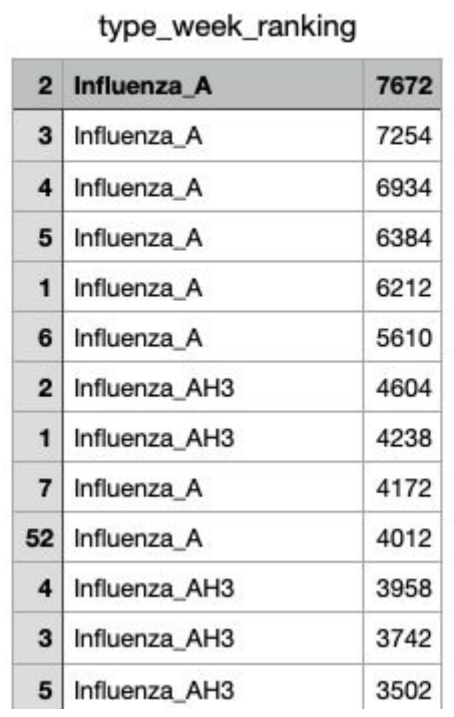
2012



2020



Top week-type ranking



type_week_ranking

2	**Influenza_A**	**7672**
3	Influenza_A	7254
4	Influenza_A	6934
5	Influenza_A	6384
1	Influenza_A	6212
6	Influenza_A	5610
2	Influenza_AH3	4604
1	Influenza_AH3	4238
7	Influenza_A	4172
52	Influenza_A	4012
4	Influenza_AH3	3958
3	Influenza_AH3	3742
5	Influenza_AH3	3502

Tools and Commands

I used bash scripts to preap the data for visualization

- from the original data set i groups and aggregate data by year. Using awk, and sort in my yearly.sh bash script. The result data set is acting as the master data set that i aggregate most of the rest from it.
- type_week_ranking.sh organize data based on which weeks and types with hughes rate.
- yrly_data.sh creates different data set one for each year from the master dataset. It includes region, type, count. I dropped the year column because the name of the file was sufficient indication of year.
- type_year.sh create several datasets from the yearly one that i created in previous task, and it shows the total death count for each individual type for each year.
- total_death.sh and total_type.sh creates two data sets that shows the total death based on region and on based on influanzatype. The output are stored in total_Deat.csv and type_data.csv in the root file.

For visual representation I created a python Flask application.

I employed Pandas and Plotly libraries.

References

[1] https://data.ca.gov/dataset/influenza-surveillance