An Investigation into the High Death Rate in New York

The effect of Pneumonia and Influenza

As submitted by

Samuel Bergin

Marticulation number: 22208243

Email: samuel.bergin@stud.th-deg.de

Overview

Introduction

Diseases are a well known part of life and can have major effects on the general populace. Serious infections like COVID-19 which rocked the world in 2019 for 3 years can greatly affect the livelihoods of many people.

Since the beginning of the century there have been 6 separate pandemics which have shocked the world and left societies reeling from the impact (Bhadoria, Gupta and Agarwal, 2021).

Because of how harmful diseases are, it is always important for governments and companies to understand the patterns of infection and death rates of diseases within their region. This is where Project Tycho comes in. Project Tycho is a collaborative project run by the University of Pittsburgh in Pennsylvania. It is a collection of data sets which present diseases within certain regions like the US, Australia and the UK.

Each of these data sets provides information on the diseases prevalent within that region, the number of cases and the number of deaths associated with those diseases. Which can be very useful for governments to understand where their public health systems are having the most trouble disease wise and aid in decision making around healthcare.

Problem Definition

The aim of this data analysis is to determine the reason as to why New York state in the US has the highest death rate. This analysis includes analyzing the data from each region in New York that is provided, each disease prevalent in New York and the population distribution of these regions in comparison to each other. The data provided accounts for events from 1888 to 2014.

Objectives

The objective of this analysis is to determine if the higher death rate in New York can be attributed to a certain disease or if there is another underlying reason within the population density or the healthcare system itself within New York state.

Methods

Before the analysis can be performed the following libraries were loaded:

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(openair)
```

Once these libraries were loaded the first step was to load the data into R, reorder it and break it down into usable parts. The initial data is too massive and trying to analyse the whole thing will lead to memory space issues.

```
US_data <- read.csv("C:/Users/sambe/OneDrive/Documents/Data Vis/Data/ProjectTycho_Level2_v1.1.0_0/Proje
```

```
US_data <- US_data %>% arrange(epi_week)
```

```
US_1888_to_1940 <- US_data[c(1:2282844),c(1,9,10)]
```

```
US_1940_to_present <- US_data[c(2282844:3659360),c(1,9,10)]
```

After separating the data by date the next step was to do the same by disease.

```
US_diseases <- US_data[,c(1,3,5,7,8)]
US_diseases <- US_diseases %>% arrange(state)
```

Following this data it was clear in Fig. that New York had the highest number of deaths related to disease. This lead to subsetting the data to just those elements that are located in the state of New York.

```
NY_Data <- subset(US_data, state == 'NY')
```

```
NY_1888_1929 <- NY_Data[c(1:95928),]
```

```
NY_1930_Present <- NY_Data[c(95929:186258),]
```

```
NY_P_I <- subset(NY_1930_Present, disease == "PNEUMONIA AND INFLUENZA")
NY_P_I <- NY_P_I[,c(1,4,6,7,8,9)]
```

```
NY_1975_2014 <- NY_P_I[c(3626:16451),]
```

```
NY_2000s <- NY_1975_2014[c(8388:12826),]
```

Results

General Visualisation

Initially the total deaths and cases for every recorded disease across America and across the entire data set. Fig. 1 clearly depicts New York state as having the greatest total cases and deaths across the years from diseases.

```
ggplot(US_diseases) +
  aes(x = state, y = number, fill = event) +
  geom_col() +
  scale_fill_hue(direction = 1) +
  labs(
    x = "States",
    y = "Total Count",
    title = "Total Deaths Vs Cases"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom", text = element_text(size = 6))
```

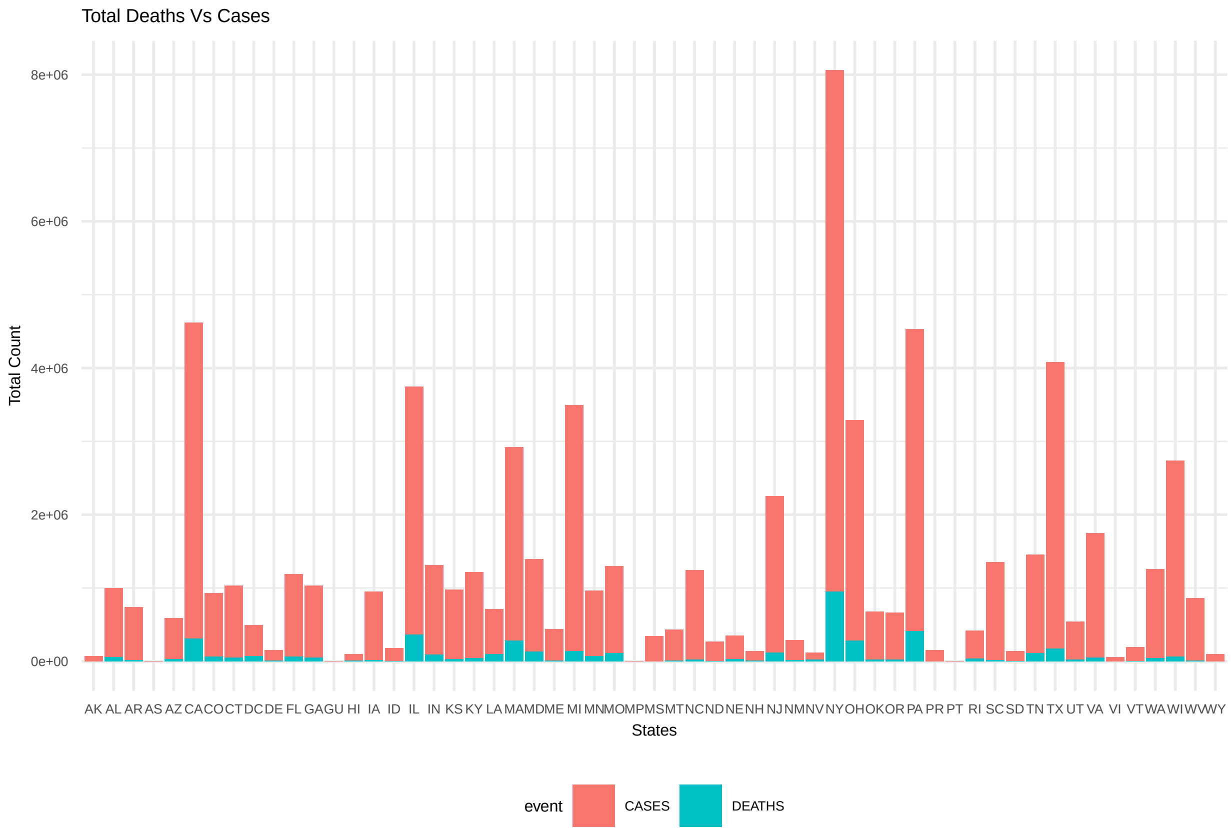

Fig. 1: The total number of events and deaths across the US. Events and Deaths are presented seperately with cases being the greater proportion of the total count from the data.

Following this revelation, the data split into two year ranges was used to compare what events occured at each region in New York state for each date recorded. Fig. 2.1 presents something well known and understood. Until the early 1900s it was much more likely for people to die from disease because of a lack of treatment for these now very treatable diseases.

What's more interesting is that in Fig. 2.2 it is observed that deaths suddenly spike in the major cities of Buffalo, New York City and Rochester.

It is important to note that after the 1920's quite a few of the regions in New York state stopped providing data! This means that we cannot be sure if this is a wide spread issue across the entire state or if this is a problem only in the major cities of New York. Why does the death events take over later on the timeline?

```
ggplot(NY_1888_1929) +
  aes(x = from_date, y = loc, fill = event) +
  geom_tile() +
  scale_fill_manual(
    values = c(CASES = "#EF0000",
    DEATHS = "#0100EB")
  ) +
  labs(
    x = "1888 - 1930",
    y = "Cities in New York",
    title = "New York Locations",
    subtitle = "1888 - 1930"
  ) +
  theme_minimal()
```

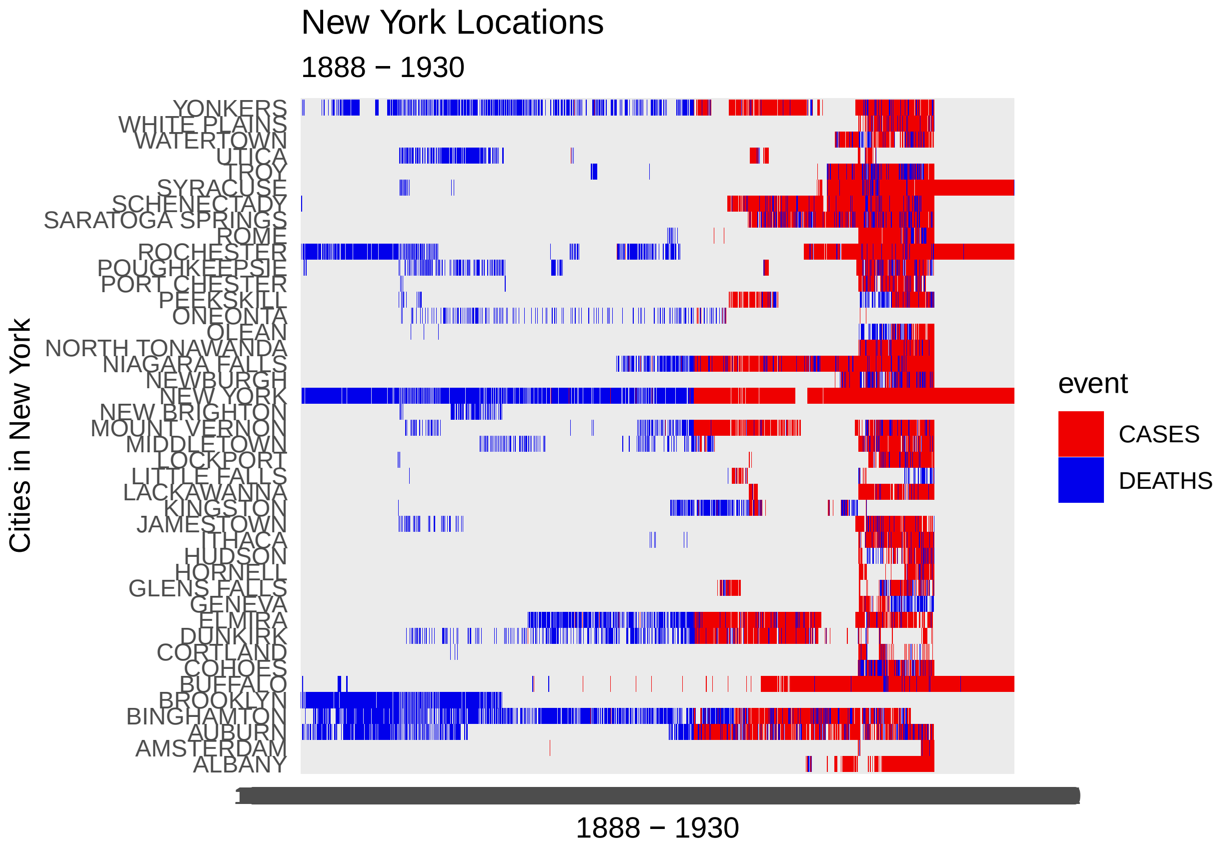

Fig. 2.1: The total number of cases and deaths across New York state, between the years of 1888 and 1930.

```
ggplot(NY_1930_Present) +
  aes(x = from_date, y = loc, fill = event) +
  geom_tile() +
  scale_fill_manual(
    values = c(CASES = "#F81000",
    DEATHS = "#000DFF")
  ) +
  labs(
    x = "1930 - 2014",
    y = "Cities in New York",
    title = "New York Locations",
    subtitle = "1930 - 2014"
  ) +
  theme_minimal()
```

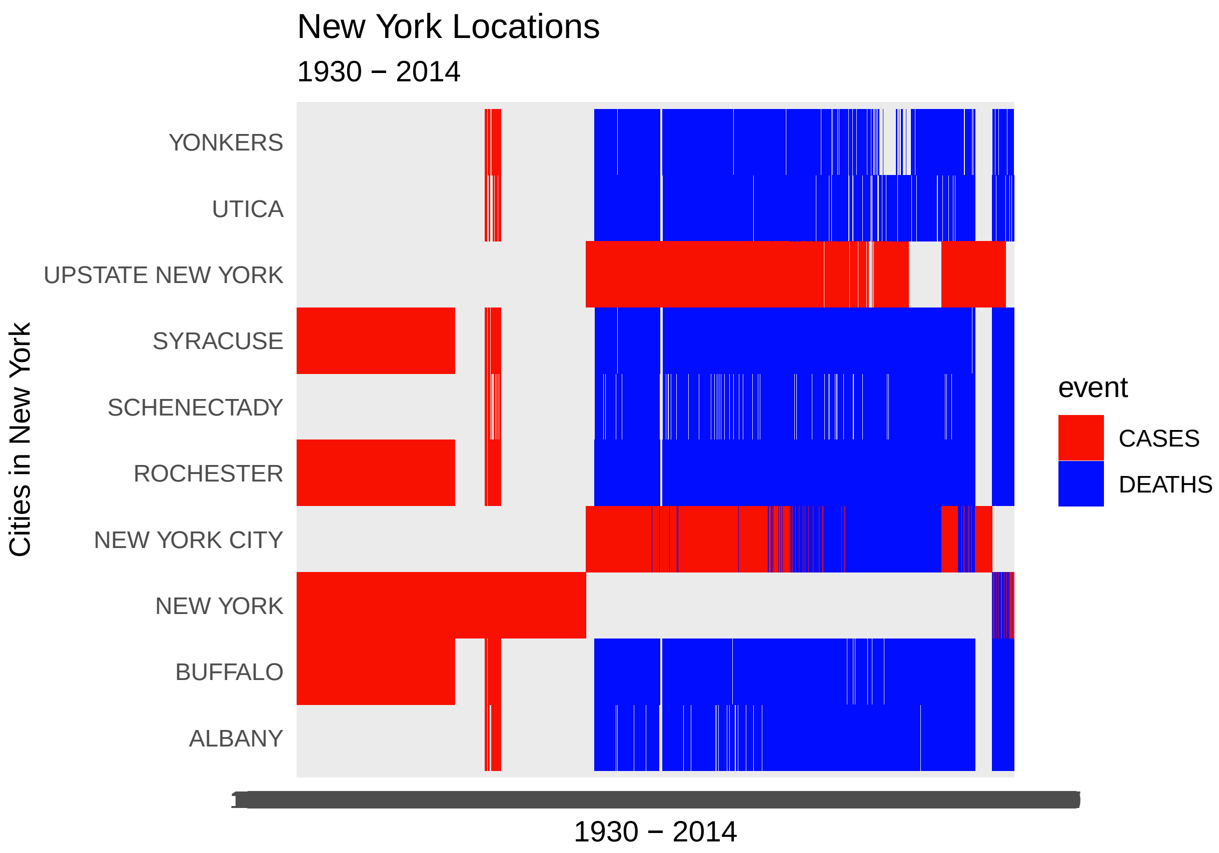

Fig. 2.2: The total cases and deaths across NY state between the years of 1930 and 2014.

Pneumonia and Influenza

To analyze more this spike in deaths I analyzed each disease against the outcome of deaths or cases. The observation in Fig. 3 is that Pneumonia and Influenza suddenly appears in the late 90s and has only deaths related to it suggesting that anyone who has gotten the combination of Pneumonia and Influenza has dies from it.

```
ggplot(NY_1930_Present) +
  aes(x = from_date, y = disease, fill = event) +
  geom_tile() +
  scale_fill_manual(
    values = c(CASES = "#FF1100",
    DEATHS = "#0136F9")
  ) +
  labs(
    x = "1930 - 2014",
    y = "Diseases",
    title = "New York Diseases",
    subtitle = "1930 - 2014"
  ) +
  theme_minimal()
```

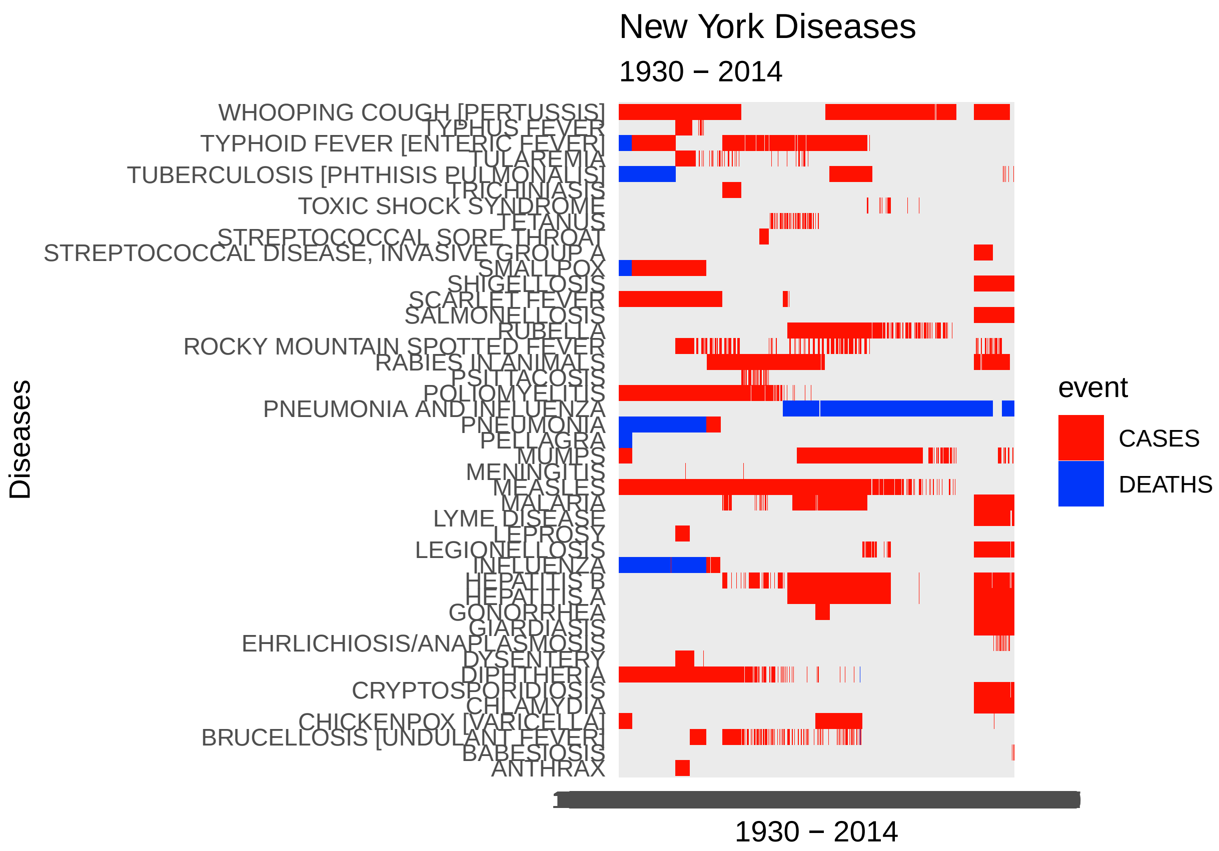

Fig. 3: The total cases and deaths in relation to diseases in New York state between the years of 1930 and 2014.

Table 1: All cases of Pneumonia and Influenza in New York state between 1930 and 2014

```
head(NY_P_I)
```

```
##         epi_week           loc                 disease  event number  from_date
## 3024646   196501        ALBANY PNEUMONIA AND INFLUENZA DEATHS      2 1965-01-03
## 3024648   196501       BUFFALO PNEUMONIA AND INFLUENZA DEATHS      6 1965-01-03
## 3024654   196501 NEW YORK CITY PNEUMONIA AND INFLUENZA DEATHS    104 1965-01-03
## 3024659   196501     ROCHESTER PNEUMONIA AND INFLUENZA DEATHS     18 1965-01-03
## 3024662   196501         UTICA PNEUMONIA AND INFLUENZA DEATHS      8 1965-01-03
## 3024663   196501       YONKERS PNEUMONIA AND INFLUENZA DEATHS      7 1965-01-03
```

```
tail(NY_P_I)
```

```
##         epi_week       loc                 disease  event number  from_date
## 3659245   201431    ALBANY PNEUMONIA AND INFLUENZA DEATHS      1 2014-07-27
## 3659247   201431   BUFFALO PNEUMONIA AND INFLUENZA DEATHS     10 2014-07-27
## 3659251   201431  NEW YORK PNEUMONIA AND INFLUENZA DEATHS     27 2014-07-27
## 3659256   201431 ROCHESTER PNEUMONIA AND INFLUENZA DEATHS      5 2014-07-27
## 3659258   201431  SYRACUSE PNEUMONIA AND INFLUENZA DEATHS      6 2014-07-27
## 3659259   201431     UTICA PNEUMONIA AND INFLUENZA DEATHS      2 2014-07-27
```

According to the data presented in Table 1 above the combination of Pneumonia and Influenza started being recorded in 1965. We can see in Fig. 5 that since that point deaths have fluctuated but have still remained prevalent across the years until the end of the data set in 2014.

```
ggplot(NY_P_I) +
  aes(x = loc, y = number) +
  geom_col(fill = "#112446") +
  labs(
    x = "Cities and Counties",
    y = "Number of Deaths",
    title = "Levels in New York",
    subtitle = "Pneumonia and Influenza"
  ) +
  theme_minimal()+
  theme(text = element_text(size = 9))
```

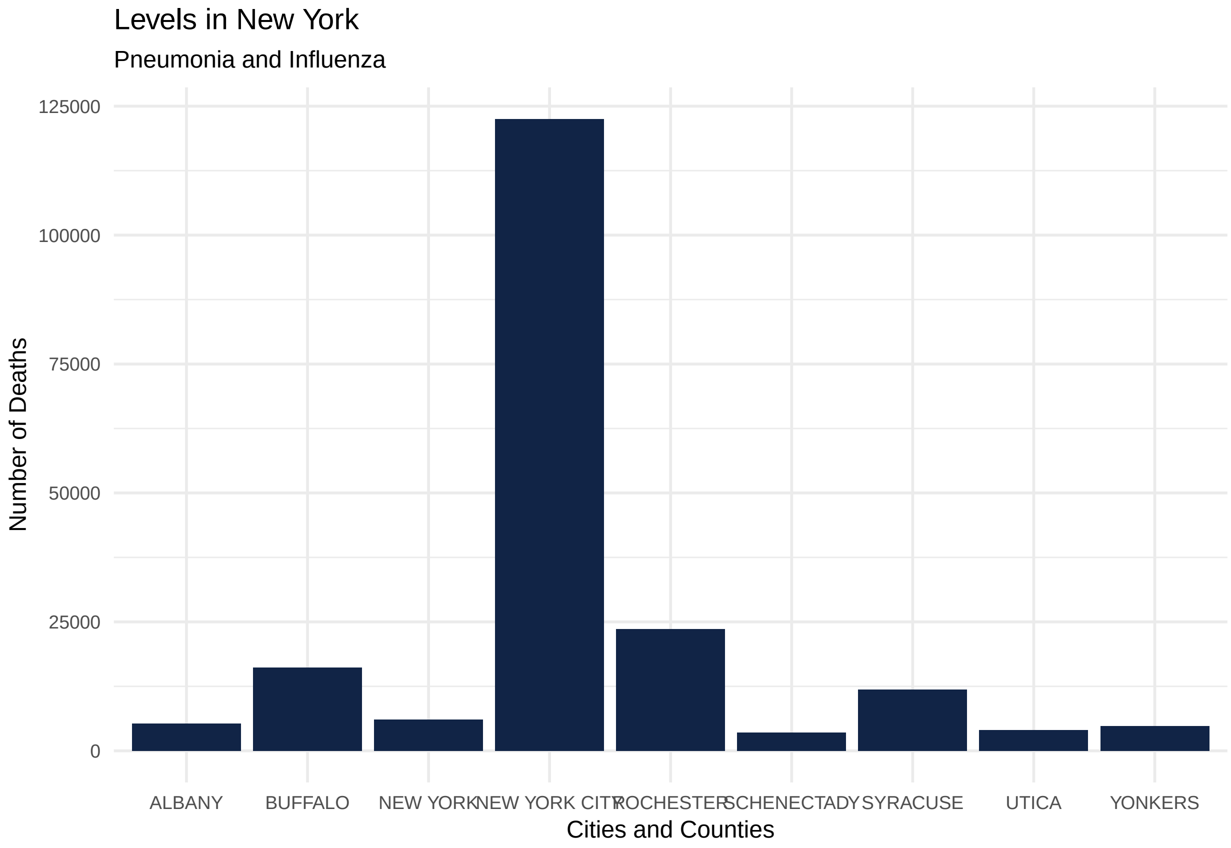

Fig. 4: Cases of Pneumonia and Influenza recorded in each region of New York state.

In Fig. 4 there is clearly a higher number of deaths reported in New York city than in other regions within New York State.

This is not entirely surprising with the difference in population density. With 13,309,214 people in a 40 km radius. In comparison Rochester, with the next highest number of deaths, has 1,161,192 people within a 40 km radius (Big radius tool, 2023).

```
NY_2000s %>%
 filter(epi_week >= 201000L & epi_week <= 201440L) %>%
 ggplot() +
  aes(x = from_date, y = number) +
  geom_line(colour = "#112446") +
  labs(
    x = "Months",
    y = "Number of Deaths",
    title = "Number of Deaths",
    subtitle = "2012-2014"
  ) +
  theme_minimal()+
  theme(axis.text.x = element_text(size=6,angle=90))
```

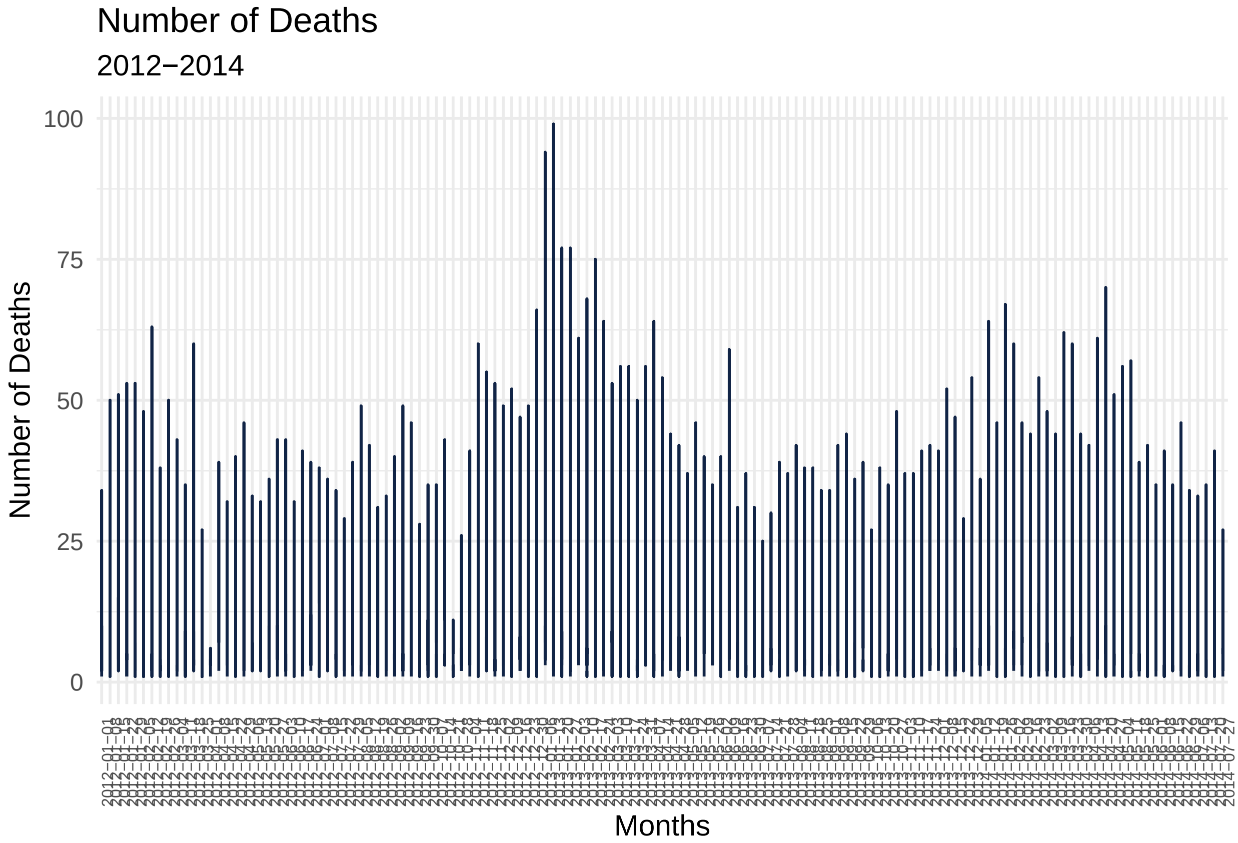

Fig. 5: The number of deaths in New York state in relation to Pneumonia and Influenza between 2012 and 2014.

In Fig. 5 there seems to be a pattern that agrees with what is known about these diseases that the deaths and cases increase in the winter months in the US because of decreased temperature and humidity levels (Enfield K, *et al*, 2012).

Discussion

Following the visualizations presented above it seems clear that this spike in deaths in New York that caused such a difference between New York and the other US states is because this combination of Pneumonia and

Influenza is highly lethal and has caused a major problem within such a densely populated place like New York City.

The greatest evidence of this can be seen in figures 4 and 5. Figure 4 clearly shows deaths in New York City and other densely populated areas like Buffalo and Rochester having higher mortality rates than those of the less populated areas within New York state. While, figure 5 illustrates the pattern of mortality linked to Pneumonia and Influenza and how the winter months have higher mortality than non-winter months. This aligns with outside research which suggests that Pnuemonia and Influenza mortality is higher when the temperature and the humidity levels are lower (Davis *et al*, 2012).

The effect of population density may also be seen in a report written by Lippert *et al* in 2021. In this report it is seen that the top 3 cities with the highest total mortality rate in relation to Pneumonia and Influenza are New York, Los Angeles and Baltimore. These three cities are all areas of high population density with 13,309,214 people in a 40 km radius in New York, 9,721,138 people in a 40 km radius in Los Angeles, and 2,783,961 people in a 40 km radius in Baltimore (Big Radius Tool, 2023). With population densities this high it is very likely that diseases like Influenza spread quickly and easily especially in the winter months when the climate is much more favorable.

It should be noted that in this report by Lippert *et al.* It is stated that mortality in relation to Influenza is higher in populations that are majority black. While this visualization does not cover the racial disparity, it is still evident within the cities with the highest mortality rates. New York, Los Angeles and Baltimore are all cities with a large Black population which may account for a majority percentage of the deaths in the data. This could be another cause for the high mortality rate that is observed in the Project Tycho data and should be considered when drafting a response to this disease.

A 2012 report by Robert E. Davis *et al* shows that Pneumonia and Influenza mortality in New York is correlated with Dew Point Temperature (DPT). In this report it was found that days that had a lower DPT had higher mortality rates in relation to Pneumonia and Influenza.

DPT is an important factor when researching respiratory diseases like Pneumonia and Influenza. In his report Davis confirms that areas with warmer and more humid climates (thus having a higher DPT) had much lower rates of mortality in relation to Pneumonia and Influenza (2012). Thus, mortality rates must be higher in areas with lower temperatures and lower humidity.

New York winters have lower humidity and temperatures. In his report Davis mentions that between 1975 and 2002 cases of Pneumonia and Influenza that lead to death were preceded by 2 to 3 weeks of low DPT (2012). This suggests that this span of time with lower humidity and temperatures is an important factor in the virality of these two diseases.

However, that may not actually be the case. In a report written by Brown *et al.* in 2020, Brown states that many of the deaths attributed to Pneumonia and Influenza in coroners reports in New York were erroneously reported. Brown stated that in a report in 2017 it was found that reports relating to pneumonia-related deaths were more likely to be low in quality and incorrectly completed than other diseases like cancer or diabetes.

Errors like this could lead to this sudden spike in deaths related to Pneumonia and Influenza. Coroners reports are reported with three important factors, the immediate cause of death (COD), intermediate CODs which are conditions that could have lead to the final COD, and the underlying COD which can be infections or a genetic disorder that lead to the COD (Brown *et al.*, 2020).

Pneumonia and Influenza are considered underlying CODs as alone they cannot cause death but their actions in weakening the immune system of the patient can lead to death via another disease. In Browns study it was found that out of 188 deaths reported with Pneumonia and/or Influenza as the underlying COD, 163 were erroneously reported, meaning that in reality Pneumonia and/or Influenza had nothing to do with the death of the patient (2020).

Because of this report I believe that this random spike in deaths related to Pneumonia and Influenza seen in Fig. 3 was caused not by the lethality of the mix of these diseases but instead by the erroneous reporting of Pneumonia and Influenza as the COD.

That is not to say that Davis *et al* were incorrect. These reports when considered together suggests that the erroneous reporting of deaths in relation to Pneumonia and Influenza is influenced by the weather. For a coroner's report to include Pneumonia and Influenza as the COD there must be evidence of these diseases within the cadaver. I suggest that these reports while inaccurate about the COD, do present evidence to the pattern that Pneumonia and Influenza flourish in the colder months in areas with lower humidity.

Conclusions

In conclusion, data sets like this one provided by Project Tycho, are very helpful in identifying patterns within populations related to disease but they can be misleading without proper research and visualization.

As can be seen, at the beginning of this visualization, in Fig. 3, I was mislead into believing that there was a major problem with Pneumonia and Influenza linked deaths in New York state. However, after further external research it was clear that this was, in fact, not the case and that erroneous reports have skewed the data presented.

It is always important to be skeptical of statistics that have been provided or that you yourself have produced. The inclusion of other data sets or other published reports are important in research. Without these steps data can be misleading and lead to the wrong conclusion.

Bibliography

Bhadoria, P., Gupta, G. and Agarwal, A. (2021) 'Viral pandemics in the past two decades: An overview', Journal of Family Medicine and Primary Care, 10(8), p. 2745. Available at: https://doi.org/10.4103/JFMPC.JFMPC_2071_20.

Big Radius Tool: StatsAmerica (no date). Available at: https://www.statsamerica.org/radius/big.aspx (Accessed: 9 June 2023).

Brown, T.S. et al. (2020) 'Erroneous Reporting of Deaths Attributed to Pneumonia and Influenza at 2 New York City Teaching Hospitals, 2013-2014', Public Health Reports, 135(6), pp. 796–804.

Davis, R.E., Rossier, C.E. and Enfield, K.B. (2012) 'The Impact of Weather on Influenza and Pneumonia Mortality in New York City, 1975–2002: A Retrospective Study', PLoS ONE, 7(3), p. 34091. Available at: https://doi.org/10.1371/JOURNAL.PONE.0034091.

Lippert, J.F. et al. (2022) 'Influenza and Pneumonia Mortality Across the 30 Biggest U.S. Cities: Assessment of Overall Trends and Racial Inequities', Journal of Racial and Ethnic Health Disparities, 9(4), p. 1152. Available at: https://doi.org/10.1007/S40615-021-01056-X.