Geographic Spread of Seasonal Flu

Clarke Miller MADA Project

Clarke Miller

2024-04-06

Warning: package 'here' was built under R version 4.3.3

Warning: package 'knitr' was built under R version 4.3.3

# 1. Abstract:

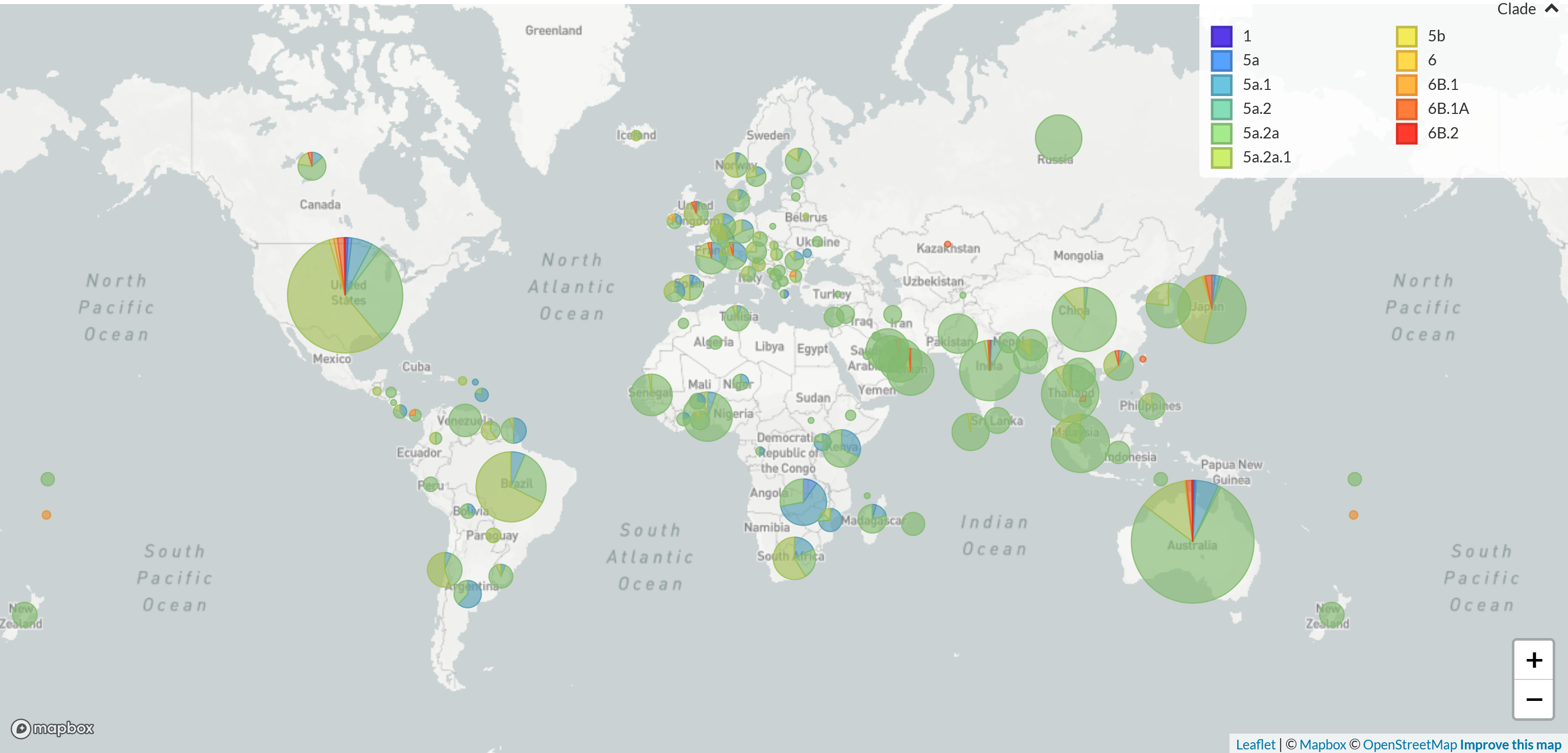
# 2. Introduction:

## 2.1 General Background Information:

Influenza is an acute respiratory virus caused by one of several strains of virus. Seasonal influenza is typically caused by circulating “A” and “B” strains of the disease. Infection is accompanied by typical respiratory symptoms. These may include cough, sore throat, and nasal discharge. Symptoms can also include abrupt onset of high-grade fever, muscle ache, loss of appetite, headache, and malaise. After a typical course, influenza can go on to affect other organs such as the lungs, brain, and heart. Severe infections may require hospitalization. Influenza infection can be much more severe for children, elderly adults, patients with chronic illnesses, and immunocompromised hosts. According to data provided by the World Health organization, seasonal respiratory illnesses are linked to over 650,000 deaths each year worldwide. In the United States the flu is responsible for approximately 42,000 deaths per year and ranks 13th most frequent cause of death. As such, influenza is the subject of extensive public health surveillance and mediation efforts.Moghadami (2017)

## 2.2 Description of Data and Data Source:

Detailed data on covid and flu strains, outbreaks, and general epidemiological information is available via several sources. Data on these infectious diseases is available from sites such as GISAID (<https://gisaid.org/>) and the Global Health Data Exchange (<https://ghdx.healthdata.org/>). These databases are highly detailed and can provide data that includes information about lineage and geographic data. (See **Figure 1** Below.) Yearly surveillance data is also available from the World Health Organization Global Influenza Programme.



***Figure 1:*** *World Flu Burden by Clade.* Figure generated at GISAID.org.

## 2.3 Questions/Hypotheses to be Addressed:

The research questions that I would like to answer are:

1. Can the progression of flu outbreaks be traced geographically or temporally over the course of a year?
2. Is there a specific location from which the annual predominant strain of flu originates?
3. If a pattern can be established, does it repeat on a yearly basis? Does the pattern shift from year to year? Can other factors, such as weather, account for any changes to this pattern?

I am certain that efforts like this already exist. The best of these is software produced by Nextstrain (<https://nextstrain.org/>) which is capable of resolving complex time/geographic coupled with genetic data for reported viral infections. Nextstrain offers real-time tracking of pathogen evolution for viruses such as SARS-CoV-2, seasonal flu, avian flu, mpox, RSV, and dengue. In fact, the graphic used in Figure 1 (above) was made by GISAID using Nextstrain software. My purpose for this project is to use this project as a learning tool and to see if I can replicate, in a small way, those prior efforts. Data to support this project is readily available from multiple sources such as GISAID (<https://gisaid.org/>), the World Health Organization Global Influenza Programme (<https://www.who.int/tools/flunet>).

# 3. Methods:

## 3.1 Schematic of Workflow:

## 3.2 Data Aquisition:

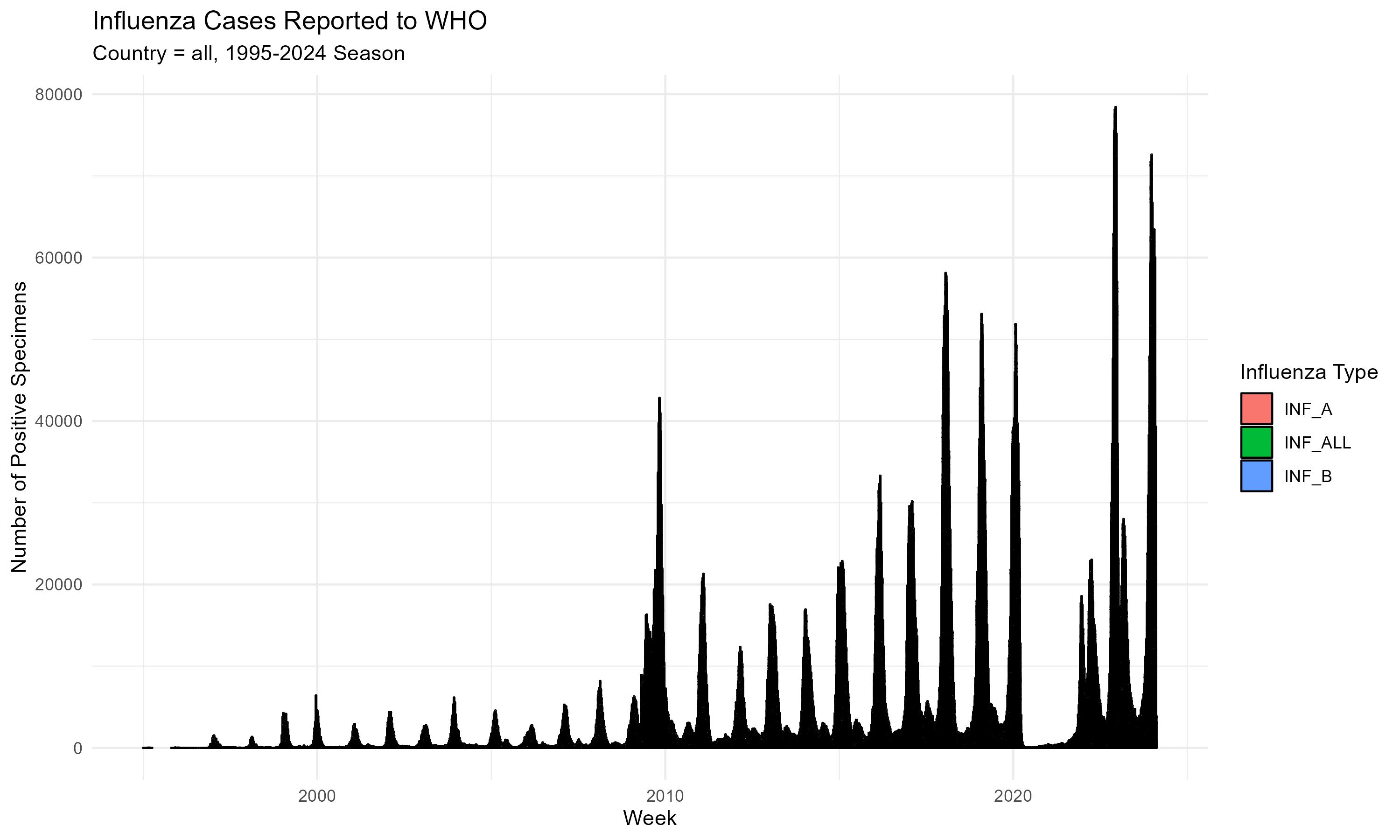
Data was acquired from the World Health Organization Global Influenza Programme. The data includes the excel files containing raw data (VIW\_FNT\_CM.xlsx) and a file containing the legend for the data (VIW\_FLU\_METADATA.csv). This data set is one example of the data available for a project of this sort. This specific data set includes information on 89 different parameters associated with influenza outbreaks as reported world-wide since about 1995. The data includes time, region, type of flu, and many other variables that are important for answering the questions posed above.

## 3.3 Data import and cleaning:

Data was imported directly from the repository as an xlsx file. (VIW\_FNT\_CM.xlsx) The file was cleaned to remove missing data. A .rds version of the data was saved (VIW\_FNT\_CM.rds). Data that was filtered or summarized for graphs, tables, and for eventual statistical analysis were saved as individual .rds or csv files. All data analysis and statistical work was performed and documented in a separate .qmd file (workspace.qmd).

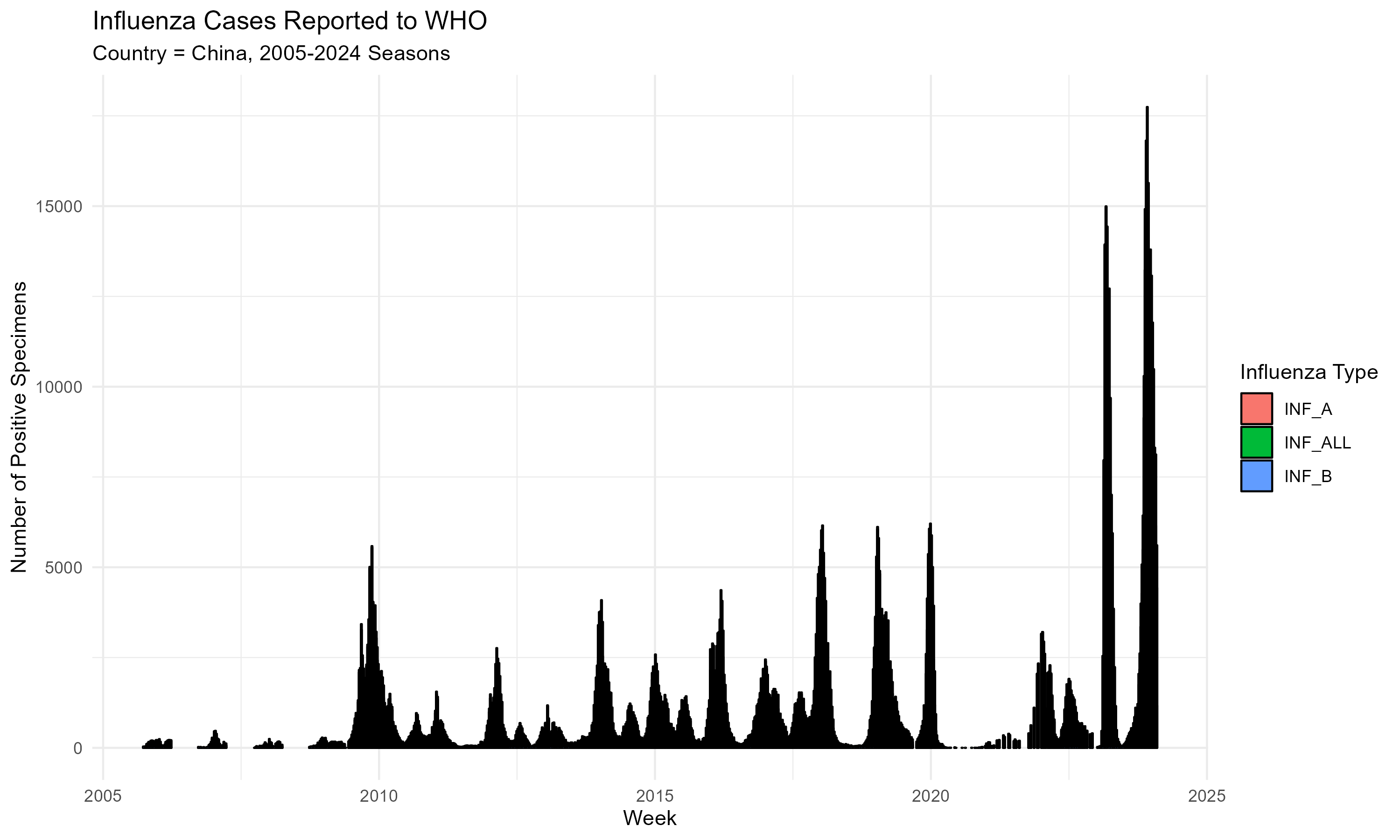
## 3.4 Exploratory/Descriptive analysis:

In order to get a feel for the scale of data in the VIW\_FNT\_CM.xlsx one often needs to summarize or tabulate the data. In this case, the data set is very large and summary tables exceedingly cumbersome. Given the size of the data set, it is best to begin with a graph of the incidence of influenza type A, type B, and the total number of reported cases for the duration of reported. (See **Figure 2.**)



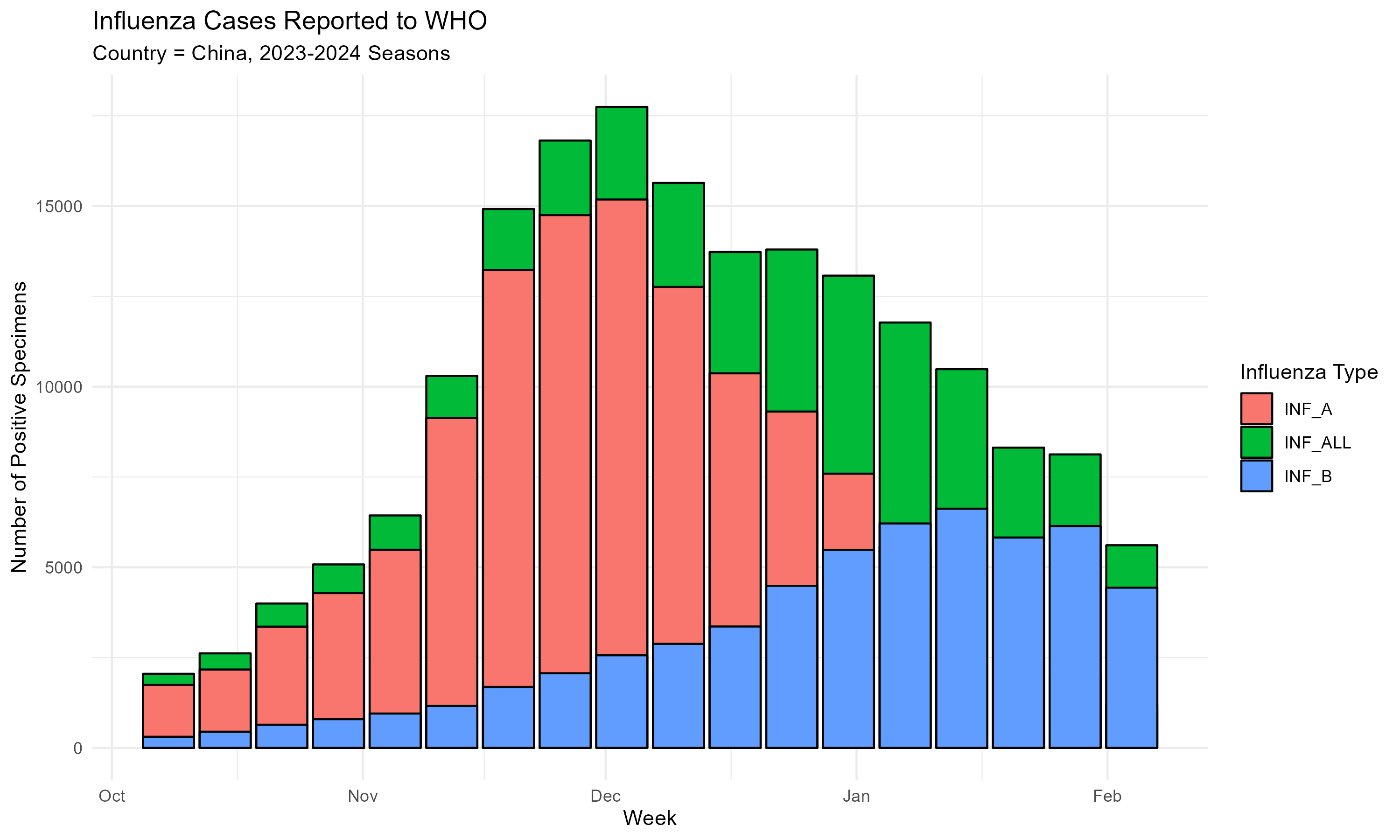
***Figure 2:*** *World Flu Burden by Year.*

The graph reveals the extent of the data available. The graph also reveals interesting features of the data set that may not have been readily apparent from if viewing the data in a tabular format. Since 1995 the number of seasonal flu cases has followed a pattern that peaked during the winter months in the Northern hemisphere. Notable features of this graph include the large jump in reported cases in the year 2010 and the large gap in reported cases in the year 2021. The jump in cases on or about 2010 is due to increased participation in surveillance efforts and improved diagnostic techniques. Hand in hand with these factors was an increasing awareness of the issue and consequent increased rate of reporting. (Perdue & Nguyen, 2012) The gap in the 2021 flu season was likely due to the measures taken to curb the SARS-CoV-2 pandemic through the 2020-2021 flu season. Basic measures such as maintaining proper social distance, hand washing, public wearing of masks to prevent aerosol droplet spread, and other efforts were extremely effective at preventing the spread of seasonal flu. (Rubin, 2021) Never the less, the amount of data reported worldwide since 1995 is massive. In fact, the data is so dense that none of the categories of flu selected are visible as separate itmes in the bar graph. Further filtering is obviously required. For example, we might want to look at the total flu cases compared to the number of total cases of the A strains and the total number of the B strains in countries such as China, Australia, Canada, and France. It is also a good idea to clean up the data by omitting any missing values. **Figure 3** depicts the data set filtered for information from China.



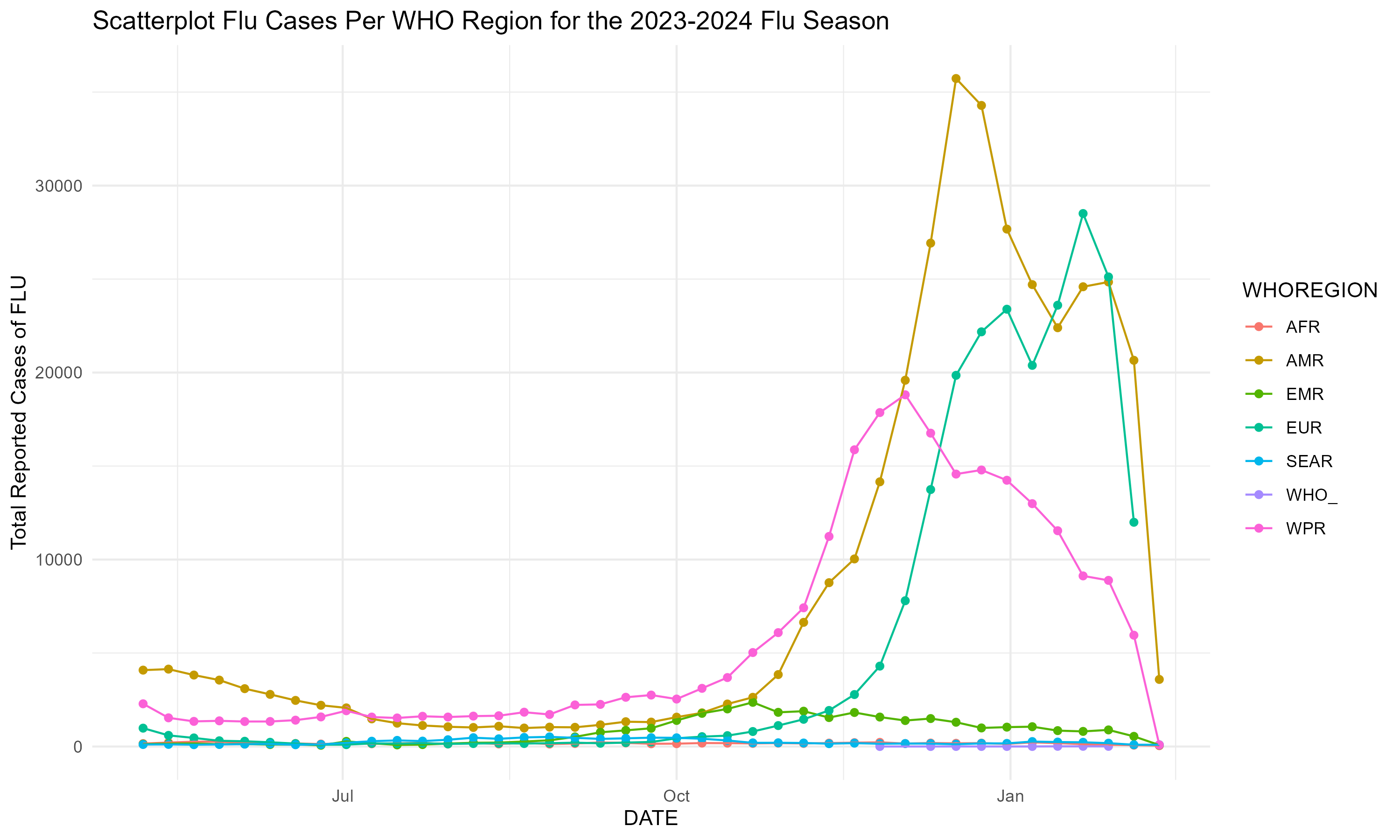
***Figure 3:*** *World Flu Burden by Year for China Since Reporting Began.*

Restricting the data set to a single country doesn’t seen to resolve the data in a useful manner. Overall the data still shows the same basic structure and features as the entire data set. It may be helpful to focus in on a single year. **Figure 4** depicts the influenza surveillance data for the country of China for a single flu season.



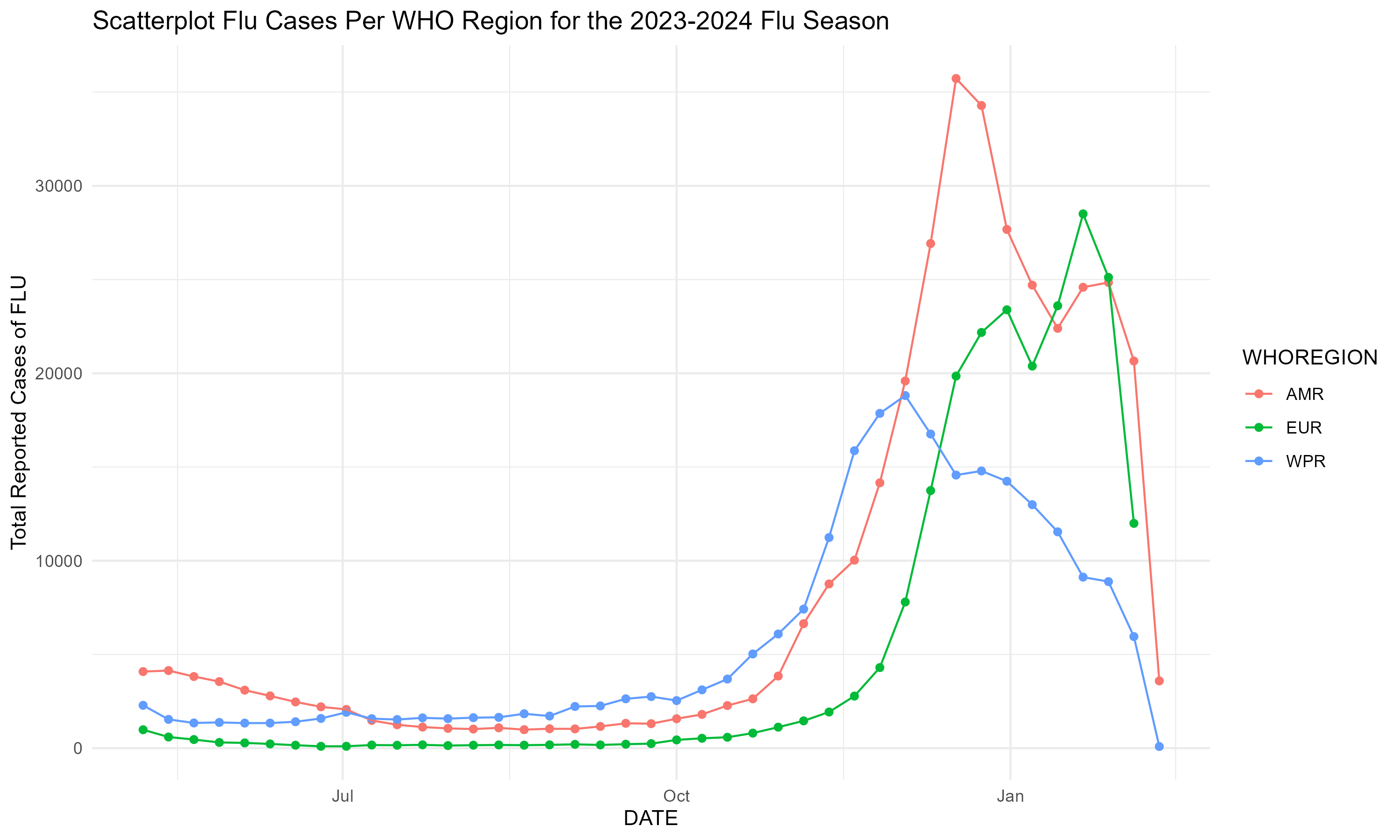
***Figure 4:*** *World Flu Burden by Year: China, 2023-2024 Flu Season.*

These graphs help whittle down the data a bit. For example, we can look at the stacked graph in **Figure 4** to see that over the course of the year the major flu variant shifted from influenza A to influenza B. That being said, this format is not helpful for looking at the incidence over time per country or per region and comparison of those regions. To look at the data in that manner and, perhaps, look for time and geographical patterns a simple scatter plot may be helpful. For example, a graph of total reported cases of flu for each WHO region over a single flu season may be revealing. (See **Figure 5**)

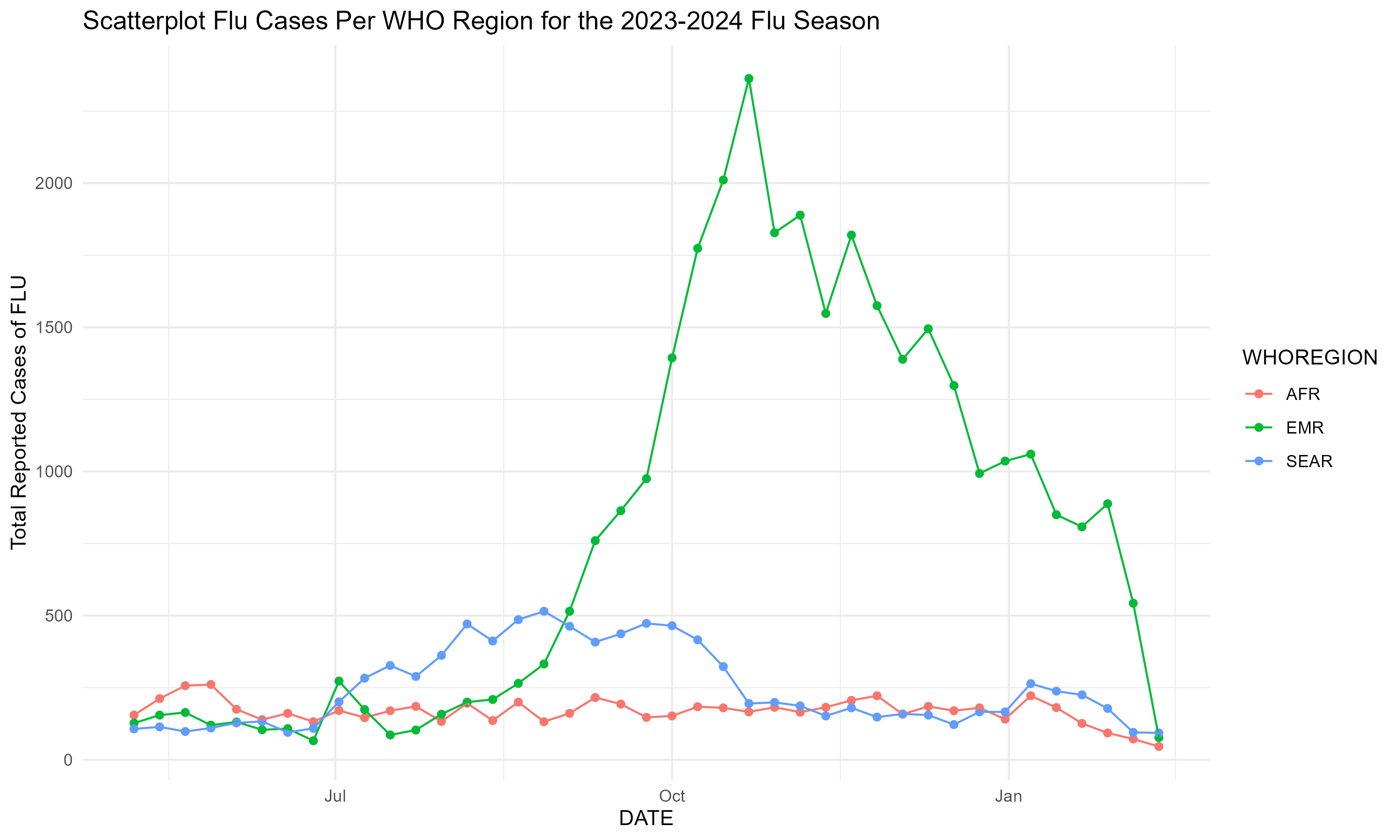


***Figure 5:*** *World Flu Burden by Region Over the 2023-2024 Flu Season.*

The resulting graph shows that the maxima for the three largest WHO flu regions, WPR, AMR, and EUR, do follow each other by approximately 3-4 weeks. The region encompassing Europe shows two distinct maxima. The first maxima is preceded by the maxima in the People’s Republic of China WHO region by 3-4 weeks and corresponds to the largest maxima in the AMR data. The second is preceded by the maxima in the data from the WHO Americas Region by the same 3-4 week time frame. The data from the AMR region also show a second peak in the number of flu cases reported that follows the first EUR region peak by approximately 3-4 weeks. Perhaps this indicates spread from EUR back to the AMR region? As such there certainly seems to be a geographic/time component to the spread of the flu worldwide. (West to east in the northern hemisphere.) Further analysis of the available phylogeographic data may be able to pinpoint the origin of each specific strain. In **Figure 5** it is also apparent that the relative magnitudes of the values of the WPR, AMR, and EUR regions overwhelm the data from the other WHO regions. In this case it may be prudent to separate the graphs into separate graphs for WPR, AMR, EUR and AFR, EMR, SEAR. (See **Figure 6** and **Figure 7.**)



***Figure 6:*** *World Flu Burden by Region Over the 2023-2024 Flu Season.* WHO Regions: WPR, AMR, and EUR



***Figure 7:*** *World Flu Burden by Region Over the 2023-2024 Flu Season.* WHO Regions: AFR, EMR, SEAR

By splitting the graphs a great deal more detail can be discerned. In addition to the putative west to east pattern that appeared for the WPR, AMR, and EUR WHO regions, there is a similar pattern for the SEAR and EMR regions. An easy interpretation of this data would point to the seasonal flu beginning in the southeast Asia WHO region progressing to the eastern Mediterranean WHO region, then progressing to the WPR, AMR, and the EUR WHO regions respectively. Of course, it is impossible to draw any conclusions from this data set and a single season. To make a solid determination the data set would need to track not just time and geographic information, but connect the actual flu strains as they move around the globe. This can be done with genetic data, i.e. DNA or RNA sequence data, from the flu cases reported. In fact this is exactly the approach that the Nextstrain software takes. In addition to presenting strong evidence of the central question of this research. Further, graphing the data as opposed to attempting to generate a summary table is a far more effective way to explore the data set. As can be seen from **Table 1** (below), a tabular format doesn’t highlight the patterns in the data in the same explicit manner as the graphic exploration.

WHOREGION 2023-08 2023-09 2023-10 2023-11 2023-12 2024-01 2024-02  
1 AFR 664 717 864 775 834 622 118  
2 AMR 4128 4817 12122 39591 144201 96535 24244  
3 EMR 1006 3114 9370 6832 6211 3606 620  
4 EUR 666 829 3470 10463 86959 97608 11993  
5 SEAR 1884 1781 1598 667 768 905 188  
6 WPR 6819 9862 20454 52385 79170 42541 6042

***Table 1:*** *World Flu Burden by Region Over the 2023-2024 Flu Season.*

## 3.5 Statistical Analysis:

A cursory look at the data presented in **Figure 6** (above) shows a putative dependence of the AMR region flu season on the flu season peak approximately 3 week prior in the WPR region. Likewise the peaks in reported flu cases in the EUR region seem to correspond to the WPR maxima approximately 3 weeks prior followed by a second maxima at a 3 week delay from the AMR region maxima. A simple multiple linear regression allowing the EUR data to be dependent data, while using offset data from the WPR and AMR regions as independent data should reveal a strong relationship.

# 4. Results

[Table 1](#tbl-summarytable) shows a summary of the data.

Note the loading of the data providing a **relative** path using the ../../ notation. (Two dots means a folder up). You never want to specify an **absolute** path like C:\ahandel\myproject\results\ because if you share this with someone, it won’t work for them since they don’t have that path. You can also use the here R package to create paths. See examples of that below. I generally recommend the here package.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Data summary table.   | skim\_type | skim\_variable | n\_missing | complete\_rate | factor.ordered | factor.n\_unique | factor.top\_counts | numeric.mean | numeric.sd | numeric.p0 | numeric.p25 | numeric.p50 | numeric.p75 | numeric.p100 | numeric.hist | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | factor | Gender | 0 | 1 | FALSE | 3 | M: 4, F: 3, O: 2 | NA | NA | NA | NA | NA | NA | NA | NA | | numeric | Height | 0 | 1 | NA | NA | NA | 165.66667 | 15.97655 | 133 | 156 | 166 | 178 | 183 | ▂▁▃▃▇ | | numeric | Weight | 0 | 1 | NA | NA | NA | 70.11111 | 21.24526 | 45 | 55 | 70 | 80 | 110 | ▇▂▃▂▂ | |

***Table 1.*** \*

## 4.1 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

[Figure 1](#fig-result) shows a scatterplot figure produced by one of the R scripts.

|  |
| --- |
| Figure 1: Height and weight stratified by gender. |

## 4.2 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example [Table 2](#tbl-resulttable2) shows a summary of a linear model fit.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Linear model fit table.   | term | estimate | std.error | statistic | p.value | | --- | --- | --- | --- | --- | | (Intercept) | 149.2726967 | 23.3823360 | 6.3839942 | 0.0013962 | | Weight | 0.2623972 | 0.3512436 | 0.7470519 | 0.4886517 | | GenderM | -2.1244913 | 15.5488953 | -0.1366329 | 0.8966520 | | GenderO | -4.7644739 | 19.0114155 | -0.2506112 | 0.8120871 | |

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

These papers (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, Shen, & Handel, 2020) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like.

# 6. References

Krammer, S., F. (2018). Influenza. *Nat Rev Dis Primers*, *4*(1), 3. <https://doi.org/10.1038/s41572-018-0002-y>

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. <https://doi.org/10.1093/ofid/ofaa494>

McKay, B., Ebell, M., Dale, A. P., Shen, Y., & Handel, A. (2020). Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients. *Proceedings. Biological Sciences*, *287*(1927), 20200496. <https://doi.org/10.1098/rspb.2020.0496>

Moghadami, M. (2017). A narrative review of influenza: A seasonal and pandemic disease. *Iran J Med Sci*, *42*(1), 2–13.

Perdue, M. L., & Nguyen, T. (2012). The WHO research agenda for influenza: Two years later. *Bull World Health Organ*, *90*(4), 246. <https://doi.org/10.2471/blt.11.090175>

Rubin, R. (2021). Influenza’s unprecedented low profile during COVID-19 pandemic leaves experts wondering what this flu season has in store. *JAMA*, *326*(10), 899–900. <https://doi.org/10.1001/jama.2021.14131>