Example Manuscript Template for a Data Analysis Project

Clarke Miller Project Proposal

Clarke Miller

2024-03-15

The structure below is one possible setup for a data analysis project (including the course project). For a manuscript, adjust as needed.

You don’t need to have exactly these sections, but the content covering those sections should be addressed.

This uses MS Word as output format. [See here](https://quarto.org/docs/output-formats/ms-word.html) for more information. You can switch to other formats, like html or pdf. See [the Quarto documentation](https://quarto.org/) for other formats.

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

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

# 1. Summary/Abstract

# 2. Introduction

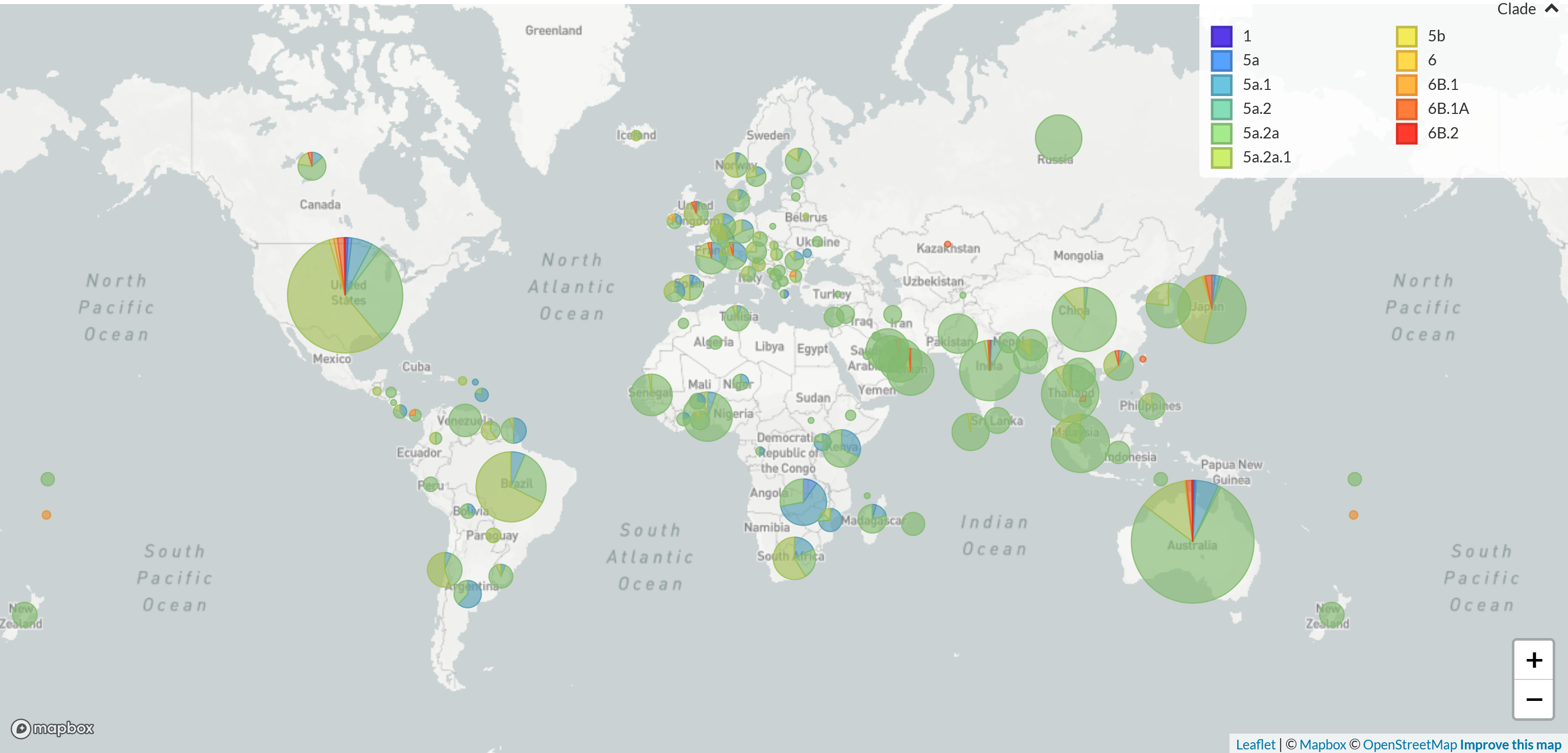
## 2.1 General Background Information

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.

## 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.)

#| label: fig-schematic  
#| fig-cap: "Figure 1: World Flu Burden by Clade."  
#| echo: FALSE  
knitr::include\_graphics(here("assets","World Flu by Cladep.png"))



## 2.3 Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis.*

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. My purpose for this project is to use it as a learning tool and to see if I can replicate those prior efforts. Data to support this project is readily availble 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 opther variables that are important for answereing the questions posed above.

## 3.3 Data import and cleaning

#Call a bunch of libraries.  
suppressPackageStartupMessages(library(tidyverse))  
suppressPackageStartupMessages(library(ggplot2))  
suppressPackageStartupMessages(library(broom))   
suppressPackageStartupMessages(library(here))  
suppressPackageStartupMessages(library(readxl))   
suppressPackageStartupMessages(library(dplyr))   
suppressPackageStartupMessages(library(tidyr))   
suppressPackageStartupMessages(library(skimr))  
suppressPackageStartupMessages(library(gt))  
suppressPackageStartupMessages(library(dslabs))  
suppressPackageStartupMessages(library(plotly))  
suppressPackageStartupMessages(library(gapminder))  
suppressPackageStartupMessages(library(kableExtra))

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

#Load excel data file  
#Data file from WHO "VIW\_FNT\_CM.xlsx"  
  
data\_location <- here::here("data","raw-data","VIW\_FNT\_CM.xlsx")  
WHO\_data\_flu <- readxl::read\_excel(data\_location)  
  
  
#Save a rds version  
save\_data\_location <- here::here("data","raw-data","VIW\_FNT\_CM.rds")  
saveRDS(WHO\_data\_flu, file = save\_data\_location)  
  
summary(WHO\_data\_flu)

WHOREGION FLUSEASON HEMISPHERE ITZ   
 Length:152584 Length:152584 Length:152584 Length:152584   
 Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character   
   
   
   
   
 COUNTRY\_CODE COUNTRY\_AREA\_TERRITORY STARTDATE   
 Length:152584 Length:152584 Min. :1995-01-02 00:00:00.00   
 Class :character Class :character 1st Qu.:2011-12-12 00:00:00.00   
 Mode :character Mode :character Median :2016-12-12 00:00:00.00   
 Mean :2015-08-26 12:42:11.68   
 3rd Qu.:2020-11-30 00:00:00.00   
 Max. :2024-02-19 00:00:00.00   
   
 YEAR ISO\_WEEK MMWR\_WEEKSTARTDATE MMWR\_YEAR   
 Min. :1995 Min. : 1.00 Min. :1995-01-01 00:00:00.00 Min. :1995   
 1st Qu.:2011 1st Qu.:12.00 1st Qu.:2011-12-11 00:00:00.00 1st Qu.:2011   
 Median :2016 Median :25.00 Median :2016-12-11 00:00:00.00 Median :2016   
 Mean :2015 Mean :26.05 Mean :2015-08-25 12:42:11.68 Mean :2015   
 3rd Qu.:2020 3rd Qu.:41.00 3rd Qu.:2020-11-29 00:00:00.00 3rd Qu.:2020   
 Max. :2024 Max. :53.00 Max. :2024-02-18 00:00:00.00 Max. :2024   
   
 MMWR\_WEEK ORIGIN\_SOURCE SPEC\_PROCESSED\_NB SPEC\_RECEIVED\_NB  
 Min. : 1.00 Length:152584 Min. : 0.0 Min. : 0   
 1st Qu.:12.00 Class :character 1st Qu.: 3.0 1st Qu.: 11   
 Median :25.00 Mode :character Median : 26.0 Median : 38   
 Mean :26.05 Mean : 478.6 Mean : 847   
 3rd Qu.:41.00 3rd Qu.: 98.0 3rd Qu.: 117   
 Max. :53.00 Max. :191785.0 Max. :191785   
 NA's :7651 NA's :100024   
 AH1N12009 AH1 AH3 AH5   
 Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.00   
 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00   
 Median : 0.00 Median : 0.00 Median : 0.00 Median : 0.00   
 Mean : 14.39 Mean : 0.83 Mean : 12.42 Mean : 0.01   
 3rd Qu.: 2.00 3rd Qu.: 0.00 3rd Qu.: 2.00 3rd Qu.: 0.00   
 Max. :10575.00 Max. :906.00 Max. :14991.00 Max. :14.00   
 NA's :33191 NA's :48676 NA's :17256 NA's :106265   
 AH7N9 ANOTSUBTYPED ANOTSUBTYPABLE AOTHER\_SUBTYPE   
 Min. :0 Min. : 0.00 Min. : 0.0 Min. : 0.00   
 1st Qu.:0 1st Qu.: 0.00 1st Qu.: 0.0 1st Qu.: 0.00   
 Median :0 Median : 0.00 Median : 0.0 Median : 0.00   
 Mean :0 Mean : 22.87 Mean : 0.2 Mean : 0.02   
 3rd Qu.:0 3rd Qu.: 0.00 3rd Qu.: 0.0 3rd Qu.: 0.00   
 Max. :6 Max. :48835.00 Max. :58.0 Max. :95.00   
 NA's :134698 NA's :21411 NA's :139375 NA's :110408   
 AOTHER\_SUBTYPE\_DETAILS INF\_A BVIC\_2DEL BVIC\_3DEL   
 Length:152584 Min. : 0.00 Min. : 0.00 Min. :0.00   
 Class :character 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.:0.00   
 Mode :character Median : 1.00 Median : 0.00 Median :0.00   
 Mean : 46.41 Mean : 0.42 Mean :0.01   
 3rd Qu.: 8.00 3rd Qu.: 0.00 3rd Qu.:0.00   
 Max. :48835.00 Max. :175.00 Max. :5.00   
 NA's :12756 NA's :145171 NA's :146336   
 BVIC\_NODEL BVIC\_DELUNK BYAM BNOTDETERMINED   
 Min. : 0.0 Min. : 0.00 Min. : 0.00 Min. : -1.000   
 1st Qu.: 0.0 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.000   
 Median : 0.0 Median : 0.00 Median : 0.00 Median : 0.000   
 Mean : 2.7 Mean : 1.22 Mean : 1.38 Mean : 9.994   
 3rd Qu.: 0.0 3rd Qu.: 0.00 3rd Qu.: 0.00 3rd Qu.: 1.000   
 Max. :6596.0 Max. :129.00 Max. :2641.00 Max. :11264.000   
 NA's :40115 NA's :144705 NA's :41404 NA's :18438   
 INF\_B INF\_ALL INF\_NEGATIVE ILI\_ACTIVITY   
 Min. : 0.00 Min. : 1.0 Min. : 0.0 Min. :0.00   
 1st Qu.: 0.00 1st Qu.: 3.0 1st Qu.: 2.0 1st Qu.:1.00   
 Median : 0.00 Median : 8.0 Median : 25.0 Median :3.00   
 Mean : 12.92 Mean : 99.3 Mean : 688.7 Mean :2.57   
 3rd Qu.: 2.00 3rd Qu.: 30.0 3rd Qu.: 87.0 3rd Qu.:3.00   
 Max. :11264.00 Max. :49007.0 Max. :147198.0 Max. :6.00   
 NA's :13408 NA's :69127 NA's :110435 NA's :82134   
 ADENO BOCA HUMAN\_CORONA METAPNEUMO   
 Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.0   
 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.0   
 Median : 0.00 Median : 0.00 Median : 0.00 Median : 0.0   
 Mean : 6.59 Mean : 0.65 Mean : 4.79 Mean : 6.5   
 3rd Qu.: 3.00 3rd Qu.: 0.00 3rd Qu.: 1.00 3rd Qu.: 1.0   
 Max. :376.00 Max. :226.00 Max. :658.00 Max. :723.0   
 NA's :132263 NA's :137596 NA's :138236 NA's :134195   
 PARAINFLUENZA RHINO RSV OTHERRESPVIRUS   
 Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.00   
 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00   
 Median : 0.00 Median : 0.00 Median : 0.00 Median : 0.00   
 Mean : 9.16 Mean : 16.38 Mean : 17.52 Mean : 9.37   
 3rd Qu.: 2.00 3rd Qu.: 4.00 3rd Qu.: 2.00 3rd Qu.: 0.00   
 Max. :463.00 Max. :1416.00 Max. :3523.00 Max. :732.00   
 NA's :132228 NA's :135313 NA's :81099 NA's :136023   
 OTHER\_RESPVIRUS\_DETAILS LAB\_RESULT\_COMMENT WCR\_COMMENT   
 Length:152584 Length:152584 Length:152584   
 Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character   
   
   
   
   
 ISO2 ISOYW MMWRYW   
 Length:152584 Min. :199501 Min. :199501   
 Class :character 1st Qu.:201150 1st Qu.:201150   
 Mode :character Median :201650 Median :201650   
 Mean :201543 Mean :201543   
 3rd Qu.:202049 3rd Qu.:202049   
 Max. :202408 Max. :202408

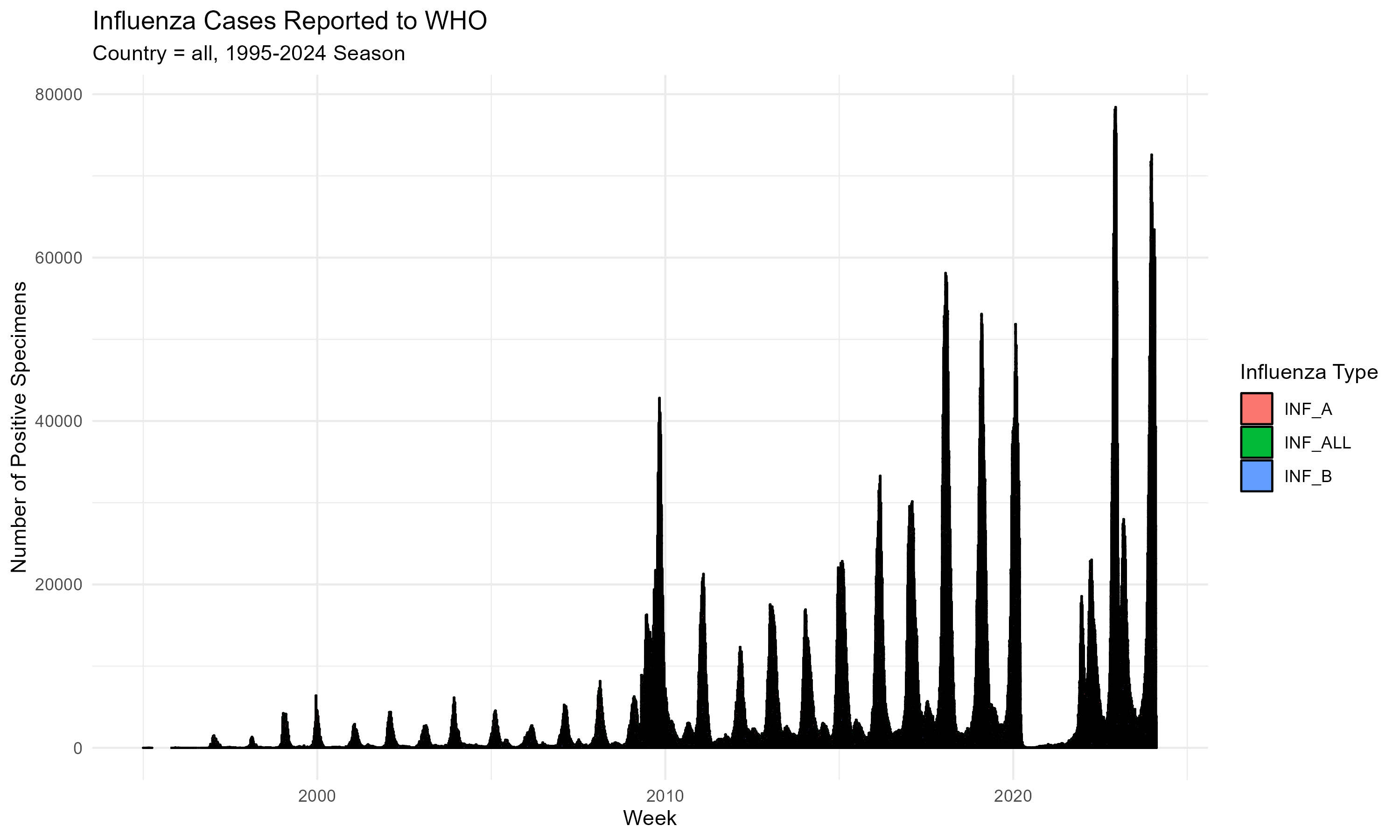
Wow. That’s a lot of data. Okay, let’s filter the data a bit. I am still working out what my final analysis will look like and, frankly, doubting my coding abilities to produce a meaningful analysis. That being said, let’s take a look.

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. We also want to clean up the data by omitting any missing values. So we filter!

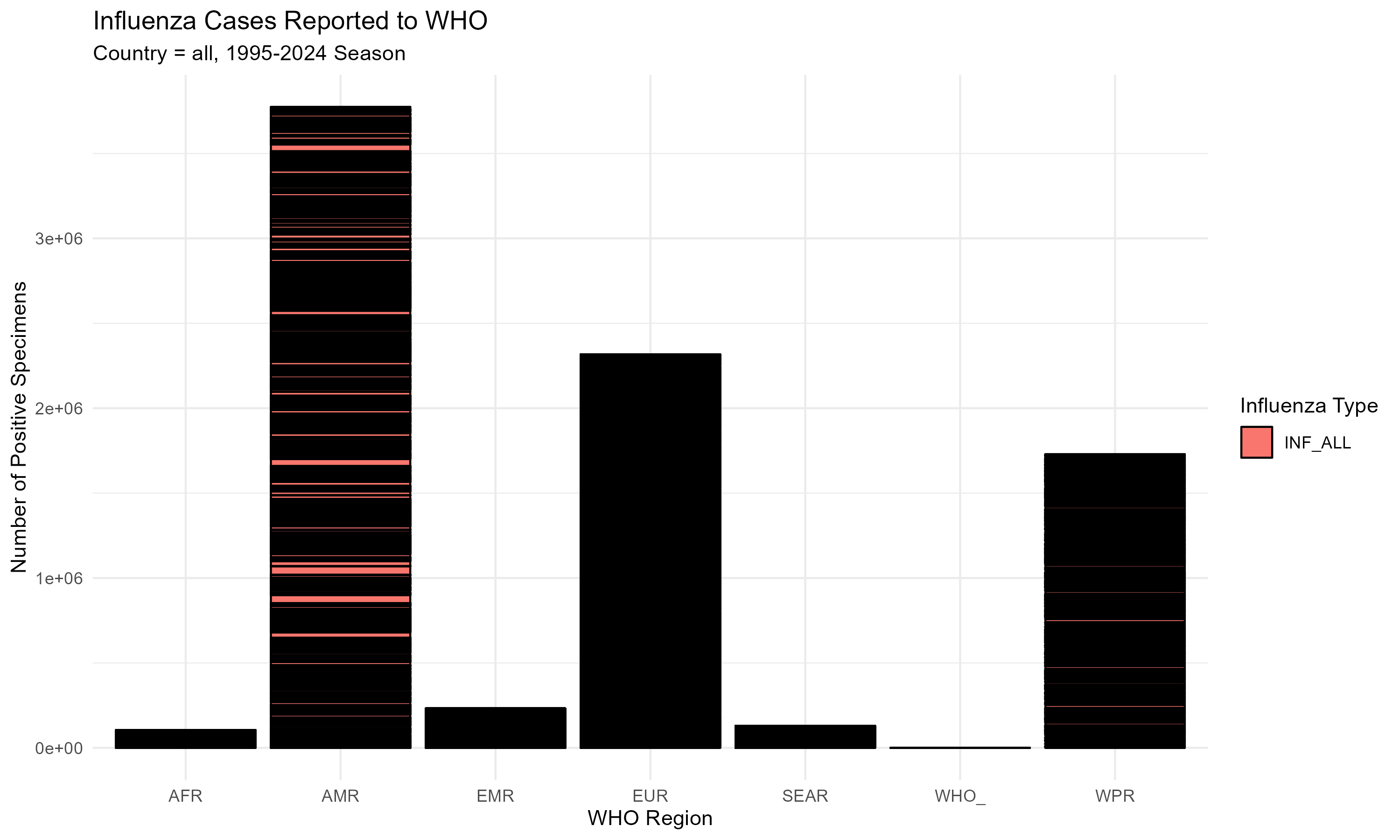
WHO\_data\_flu <- WHO\_data\_flu %>%   
 select(WHOREGION, MMWR\_WEEKSTARTDATE, COUNTRY\_AREA\_TERRITORY, INF\_ALL, INF\_A, INF\_B)  
  
summary(WHO\_data\_flu)

WHOREGION MMWR\_WEEKSTARTDATE COUNTRY\_AREA\_TERRITORY  
 Length:152584 Min. :1995-01-01 00:00:00.00 Length:152584   
 Class :character 1st Qu.:2011-12-11 00:00:00.00 Class :character   
 Mode :character Median :2016-12-11 00:00:00.00 Mode :character   
 Mean :2015-08-25 12:42:11.68   
 3rd Qu.:2020-11-29 00:00:00.00   
 Max. :2024-02-18 00:00:00.00   
   
 INF\_ALL INF\_A INF\_B   
 Min. : 1.0 Min. : 0.00 Min. : 0.00   
 1st Qu.: 3.0 1st Qu.: 0.00 1st Qu.: 0.00   
 Median : 8.0 Median : 1.00 Median : 0.00   
 Mean : 99.3 Mean : 46.41 Mean : 12.92   
 3rd Qu.: 30.0 3rd Qu.: 8.00 3rd Qu.: 2.00   
 Max. :49007.0 Max. :48835.00 Max. :11264.00   
 NA's :69127 NA's :12756 NA's :13408

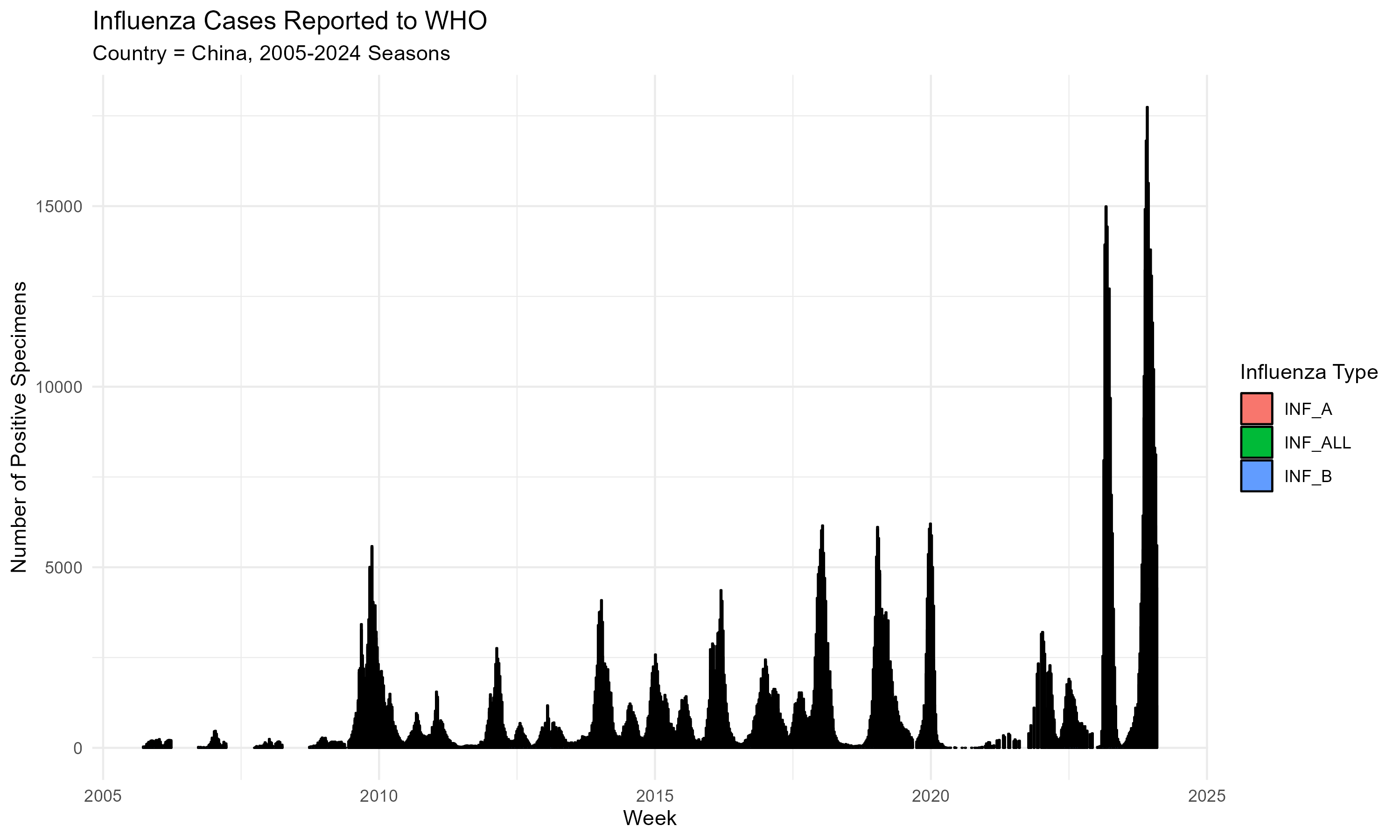
#| label: fig-schematic  
#| fig-cap: "Figure 2: World Flu Burden by Year."  
#| echo: FALSE  
knitr::include\_graphics(here("assets","plot1.png"))



#| label: fig-schematic  
#| fig-cap: "Figure 3: World Flu Burden by Region."  
#| echo: FALSE  
knitr::include\_graphics(here("assets","plot1B.png"))

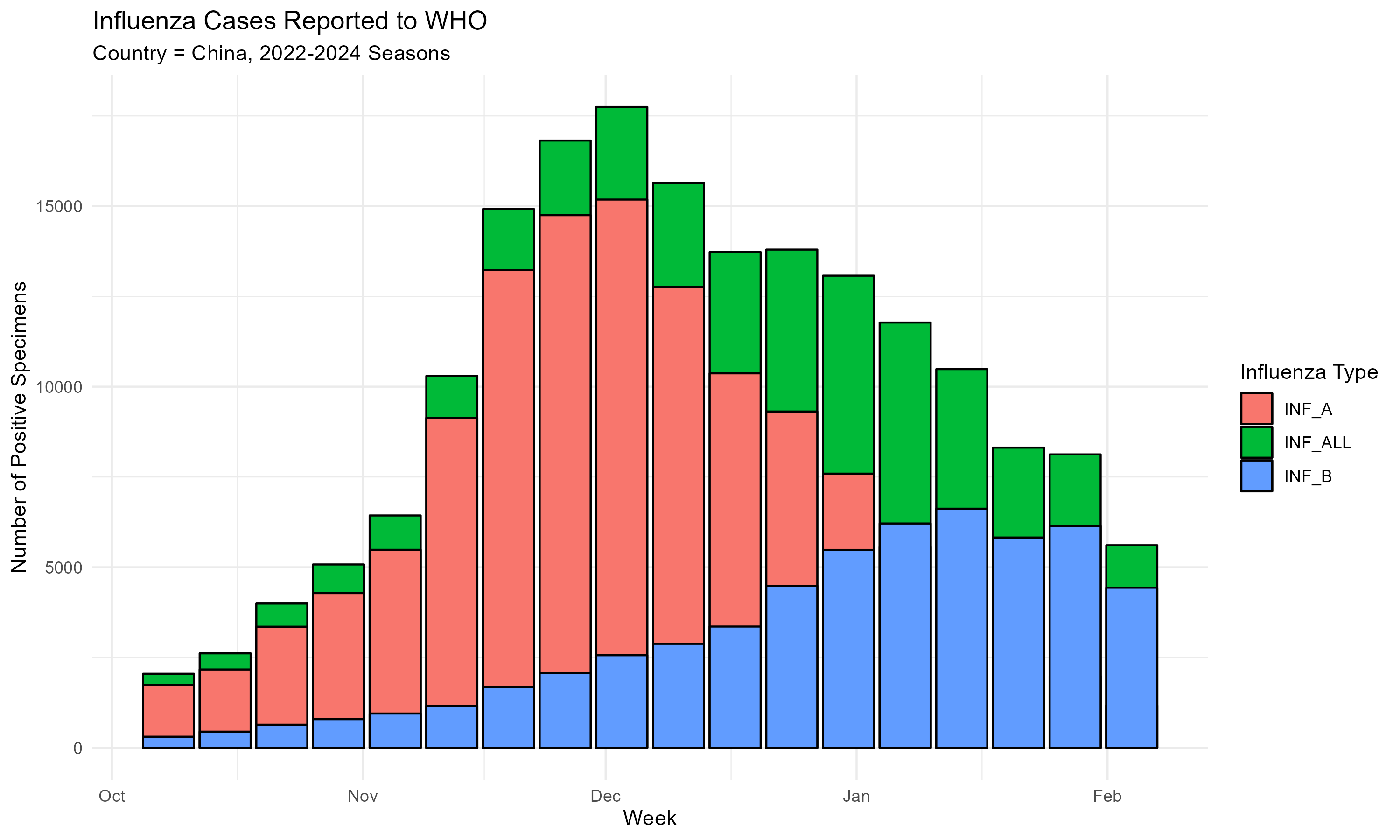


#| label: fig-schematic  
#| fig-cap: "Figure 4: World Flu Burden by Year for China Since Reporting Began."  
#| echo: FALSE  
knitr::include\_graphics(here("assets","plot2.png"))

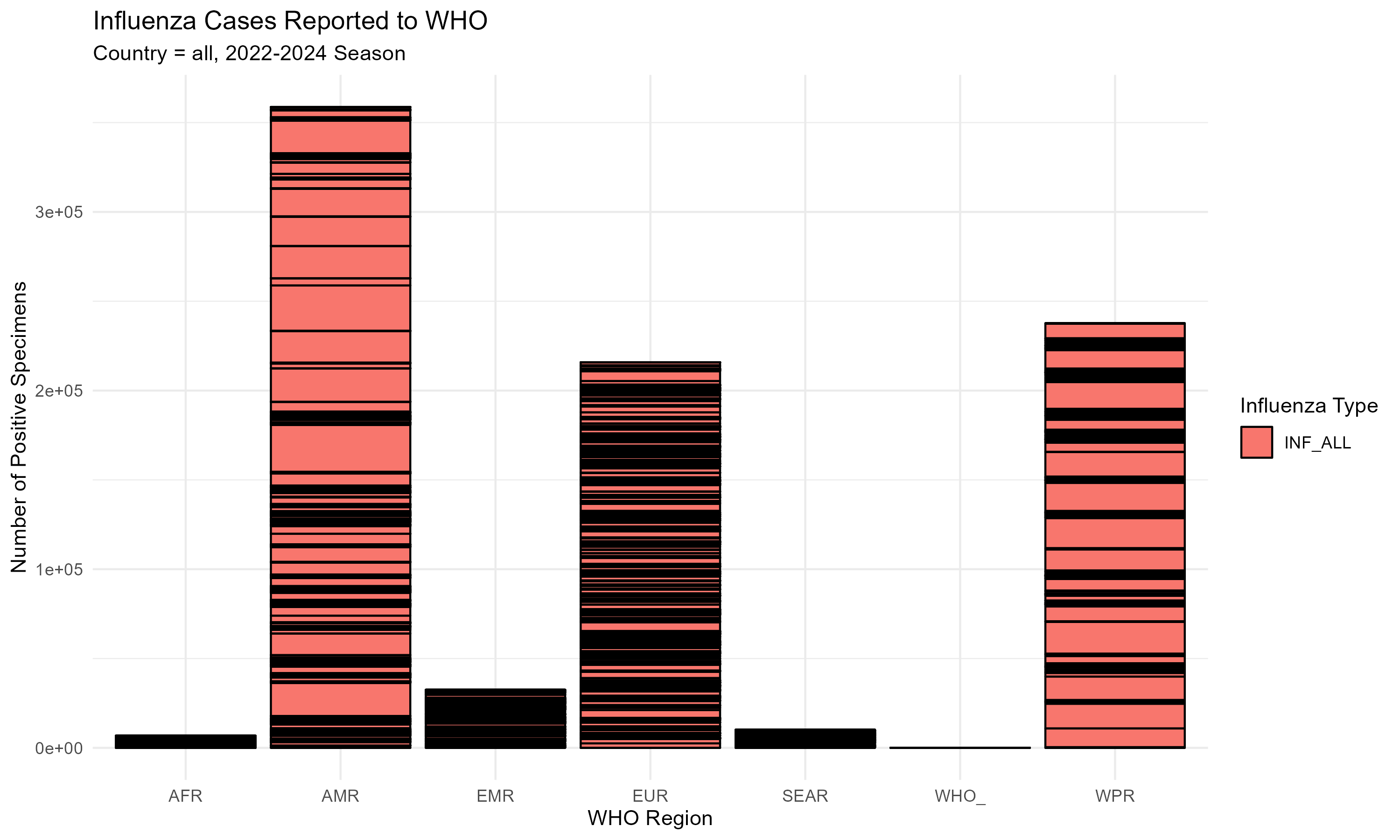


That’s still a LOT of data. Notice the big gap around 2020-2022. That is due to the restrictions imposed by the COVID pandemic. Let’s narrow this down to a more reasonable range… say to the beginning of the 2023-2024 flu season.

#| label: fig-schematic  
#| fig-cap: "Figure 4: World Flu Burden by Year."  
#| echo: FALSE  
knitr::include\_graphics(here("assets","plot3.png"))



#| label: fig-schematic  
#| fig-cap: "Figure 5: World Flu Burden by Region."  
#| echo: FALSE  
knitr::include\_graphics(here("assets","plot1C.png"))



These graphs help whittle down the data a bit. For example, we can look at the stacked grpah 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 region. To look at the sort of data we could do a stacked bar graph of the total flu counts per region/country or perhaps create a graph of the data using a simple scatter plot.

#| label: fig-schematic  
#| fig-cap: "Figure 6: World Flu Burden by Region Over the 2023-2024 Flu Season."  
#| echo: FALSE  
knitr::include\_graphics(here("assets","plot2B.png"))

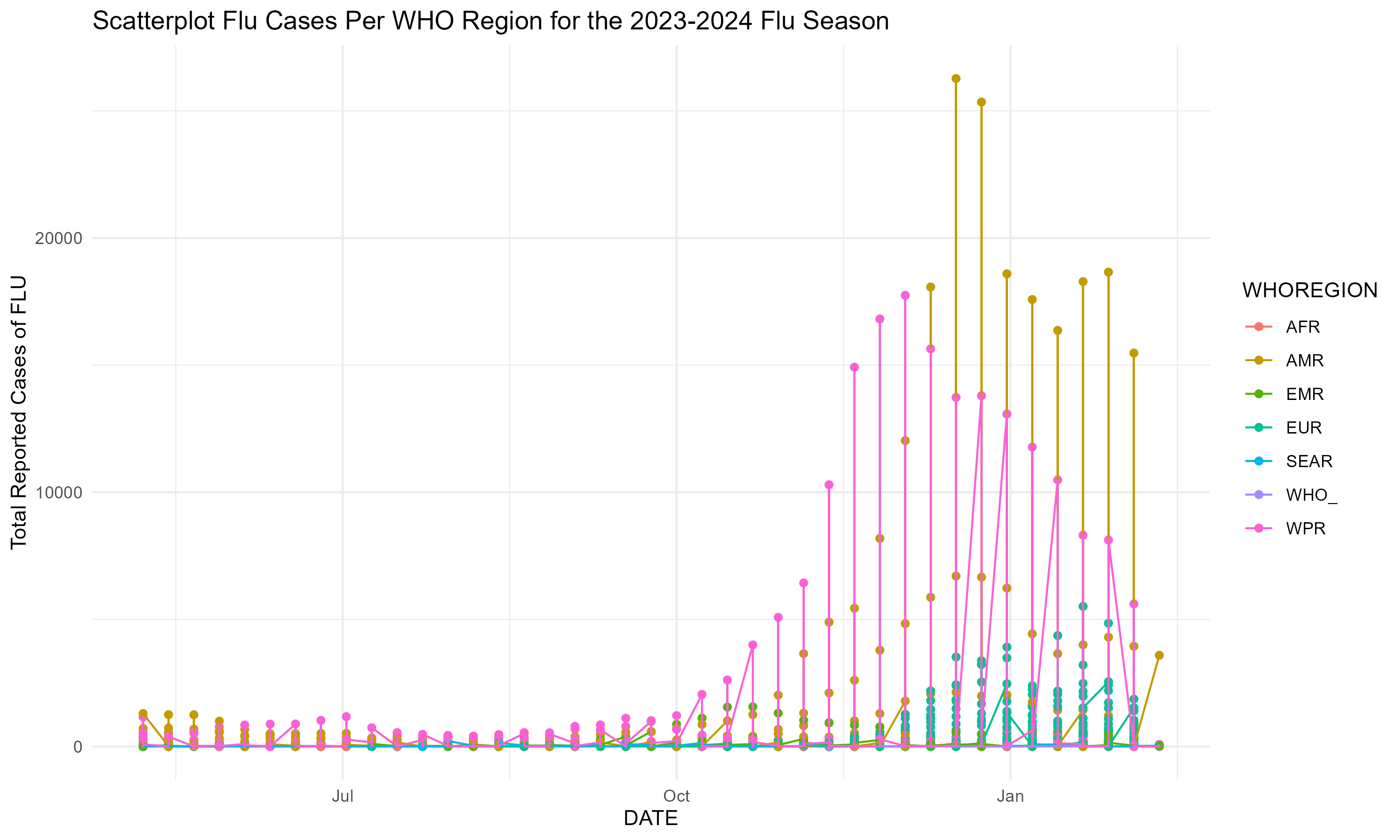


Figure 6

## 3.4 Statistical analysis

*Explain anything related to your statistical analyses.*

# 4. Results

## 4.1 Exploratory/Descriptive analysis

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

[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 | ▇▂▃▂▂ | |

## 4.2 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. |

## 4.3 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

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>