

```
library(dplyr)
```

```
##  
## Attaching package: 'dplyr'  
## The following objects are masked from 'package:stats':  
##  
##   filter, lag  
## The following objects are masked from 'package:base':  
##  
##   intersect, setdiff, setequal, union
```

```
library(ggplot2)  
library(readr)  
library(broom)  
library(testthat)
```

```
##  
## Attaching package: 'testthat'  
## The following objects are masked from 'package:readr':  
##  
##   edition_get, local_edition  
## The following object is masked from 'package:dplyr':  
##  
##   matches
```

Question 1 [2 marks] The first part of our PPDAC framework is to identify the problem you are addressing with these data. State the question you are trying to answer and let us know what type of question this is in terms of the PPDAC framework. A question statement should be as specific as possible. For example: Do students who regularly get 8 hours of sleep have fewer visits to the health center? This question is an example of an etiologic or causal question.

---

Our question is as follows: does influenza A have a higher prevalence in the northern or southern hemisphere? This is an etiological/causal question that aims to see whether a country will have a higher prevalence of influenza (response variable) depending on whether they are in the northern or southern hemisphere (explanatory variable).

Question 2 [2 marks] Why is this question interesting or important? You could talk here about how existing data/studies suggest this might be important, how the findings might make an impact, how the findings might be used, or why you are personally interested in this question.

---

Part of the reason we are interested in this question is because of the historical context of “third world countries” and the modern discourse on “the Global South.” The history of imperialism in the global south has led to the majority of countries being classified by international ranking organizations as “poor” and “developing.” Given this knowledge, it is valuable to consider how the health of people today is affected by their geographic location, and if it indeed corresponds to the historical exploitation of peoples and comparative lack of resources. Each country comes with its own history of infectious disease, and it will be interesting to see how this develops into the current trends. Another aspect of this question is how the findings can show where the virus frequents and reveal more about its genetic disposition and what climate it prefers to proliferate in, allowing scientists to advance antiviral drugs and vaccines to combat Influenza. Additionally, this question is important given that the data we analyze can also be used to guide fund allocation for influenza treatment around the world. Like any other disease, influenza’s spread is highly relevant to the hygiene and economic conditions of different countries.

Question 3 [2 marks] What is the target population for your project? Why was this target chosen? (i.e., what was your rationale for wanting to answer this question in this specific population?)

---

Our target population is the entire world. This population was chosen to shed light on hemispheric differences to create an understanding of our world as a whole. Taking data from a city, country, or one region would not be enough to generalize to the world.

Question 4 [2 marks] What is the sampling frame used to collect the data you are using? It may be helpful here to read any protocol papers, trial registration records, 'Readme' files or documentation that are associated with your dataset. If you have trouble identifying how the records/individuals were sampled, confirm with your supporting GSI that your dataset will be usable for the purposes of the class. Describe why you think this sampling strategy is appropriate for your question. To what group(s) would you feel comfortable generalizing the findings of your study and why?

---

As the WHO explains, "The data are provided remotely by National Influenza Centres (NICs) of the Global Influenza Surveillance and Response System (GISRS) and other national influenza reference laboratories collaborating actively with GISRS, or are uploaded from WHO regional databases."

According to FluNet, "FluNet is a global web-based tool for influenza virological surveillance first launched in 1997 [...].The data is provided remotely by National Influenza Centres (NICs) of the Global Influenza Surveillance and Response System (GISRS) and other national influenza reference laboratories collaborating actively with GISRS, or are uploaded from WHO regional databases."

Overall, the sampling method used by FluNet resembles convenience sampling, where the data is voluntarily reported by the national institutions registered with WHO. This sampling method is appropriate as it is the most cost-effective and timely approach to acquire such large quantities of data across a vast geographic landscape.

This sampling strategy also creates bias where only severe cases of influenza that reports to governmental institutions are recorded in the data. There's also the bias that the governments which do not work with WHO (like DPRK) would not be registered on the dataset. With this in mind, we can comfortably generalize our findings to the tested cases of influenza in countries that are WHO members.

Question 5 [2 marks] Write a brief description (1-4 sentences) of the source and contents of your dataset. Provide a URL to the original data source if applicable. If not (e.g., the data came from your internship), provide 1-2 sentences saying where the data came from. If you completed a web form to access the data and selected a subset, describe these steps (including any options you selected) and the date you accessed the data

---

The data is publically available at <https://www.who.int/tools/flunet> under the view/download filtered data section.

Each row of the data represents a reported outbreak from the aforementioned influenza centers, containing crucial information like the country the outbreak occurred in, the specific break down of the influenza genotypes, as well as the time stamp of the outbreak event.

We accessed this data on Oct 5 2023.

Question 6 [1 mark] Write code below to import your data into R. Assign your dataset to an object. Make sure to include and annotate this code in your submission (you can use a # to comment out regular text within code chunks to annotate).

```
fluA <- read_csv("VIW_FNT.csv") %>%
  select(COUNTRY_AREA_TERRITORY, WHOREGION, HEMISPHERE, ISO_YEAR, AH1N12009,
         AH1, AH3, AH5, AH7N9, AOTHER_SUBTYPE, INF_A)

## Rows: 147936 Columns: 49
## -- Column specification -----
## Delimiter: ","
## chr (12): WHOREGION, FLUSEASON, HEMISPHERE, ITZ, COUNTRY_CODE, COUNTRY_AREA...
## dbl (35): ISO_YEAR, ISO_WEEK, MMWR_YEAR, MMWR_WEEK, SPEC_PROCESSED_NB, SPEC...
## date (2): ISO_WEEKSTARTDATE, MMWR_WEEKSTARTDATE
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

*#From our influenza data, we pick out region data and country as well as the  
#year the data was collected. For simplicity sake, we are only importing  
#relevant data for the influenza A virus (and not B strains).*

```
flu.metaA <- read_csv("VIW_FLU_METADATA.csv") %>%
  select(-DatasetName, -TableName, -Comments) %>%
  filter(FieldNames %in% c("COUNTRY_AREA_TERRITORY", "WHOREGION", "HEMISPHERE",
                          "ISO_YEAR", "AH1N12009", "AH1", "AH3", "AH5", "AH7N9",
                          "AOTHER_SUBTYPE", "INF_A"))
```

```
## Rows: 89 Columns: 6
## -- Column specification -----
## Delimiter: ","
## chr (6): DatasetName, TableName, FieldName, DataType, Description, Comments
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

*#Variable descriptions for the fluA dataset. WHO region codes can be found  
#online on the WHO website.*

```
fluA
```

```
## # A tibble: 147,936 x 11
##   COUNTRY_AREA_TERRITORY WHOREGION HEMISPHERE ISO_YEAR AH1N12009 AH1 AH3
##   <chr>                  <chr>      <chr>      <dbl>      <dbl> <dbl> <dbl>
## 1 Democratic Republic of t~ AFR      SH      2010          0      0    10
## 2 Honduras              AMR      NH      2012          1      0      0
## 3 Malta                  EUR      NH      2016          0      0      1
## 4 Bangladesh            SEAR     NH      2014          0      0     21
## 5 United Kingdom, Northern~ EUR      NH      2016          0      0      0
## 6 Netherlands (Kingdom of ~ EUR      NH      2015          0      0      1
## 7 United Republic of Tanza~ AFR      SH      2009         30      1      0
## 8 Cameroon              AFR      NH      2019         13     NA      0
## 9 Malaysia              WPR      NH      2014          0      0      1
## 10 Mongolia             WPR      NH      2022          0      0      1
## # i 147,926 more rows
## # i 4 more variables: AH5 <dbl>, AH7N9 <dbl>, AOTHER_SUBTYPE <dbl>, INF_A <dbl>
```

```
flu.metaA
```

```
## # A tibble: 15 x 3
##   FieldName      DataType Description
##   <chr>          <chr>    <chr>
## 1 WHOREGION      string    WHO regions (AFR, AMR, EMR, EUR, SEAR, WPR)
## 2 HEMISPHERE      string    Hemisphere (NH=northern hemisphere, SH=south-
## 3 COUNTRY_AREA_TERRITORY string    Country, area or territory name
## 4 ISO_YEAR        integer    Year (ISO 8601)
## 5 AH1N12009       integer    Number of A(H1N1)pdm09 detections
## 6 AH1             integer    Number of A(H1) detections (other than A(H1N-
## 7 AH3             integer    Number of A(H3) detections
## 8 AH5             integer    Number of A(H5) detections
## 9 AH7N9           integer    Number of A(H7N9) detections
## 10 AOTHER_SUBTYPE  integer    Number of other influenza A subtype detectio-
## 11 INF_A          integer    Number of influenza A detections (all subtyp-
## 12 WHOREGION      string    WHO regions (AFR, AMR, EMR, EUR, SEAR, WPR)
## 13 HEMISPHERE      string    Hemisphere (NH=northern hemisphere, SH=south-
## 14 COUNTRY_AREA_TERRITORY string    Country, area or territory name
## 15 ISO_YEAR        integer    Year (ISO 8601)
```

Question 7 [3 marks] Write code in R (included in your submission with annotation) to answer the following questions:

7i) What are the dimensions of the dataset?

```
dim(fluA)
```

```
## [1] 147936      11
```

7ii) What are the variable names of the variables in your dataset?

```
names(fluA)
```

```
## [1] "COUNTRY_AREA_TERRITORY" "WHOREGION"      "HEMISPHERE"
## [4] "ISO_YEAR"              "AH1N12009"      "AH1"
## [7] "AH3"                   "AH5"            "AH7N9"
## [10] "AOTHER_SUBTYPE"        "INF_A"
```

7iii) Print the first six rows of the dataset.

```
fluA %>% head(6)
```

```
## # A tibble: 6 x 11
##   COUNTRY_AREA_TERRITORY WHOREGION HEMISPHERE ISO_YEAR AH1N12009 AH1 AH3
##   <chr>                 <chr>    <chr>      <dbl>    <dbl> <dbl> <dbl>
## 1 Democratic Republic of th~ AFR      SH        2010         0     0    10
## 2 Honduras               AMR      NH        2012         1     0     0
## 3 Malta                   EUR      NH        2016         0     0     1
## 4 Bangladesh              SEAR     NH        2014         0     0    21
## 5 United Kingdom, Northern ~ EUR      NH        2016         0     0     0
## 6 Netherlands (Kingdom of t~ EUR      NH        2015         0     0     1
## # i 4 more variables: AH5 <dbl>, AH7N9 <dbl>, AOTHER_SUBTYPE <dbl>, INF_A <dbl>
```



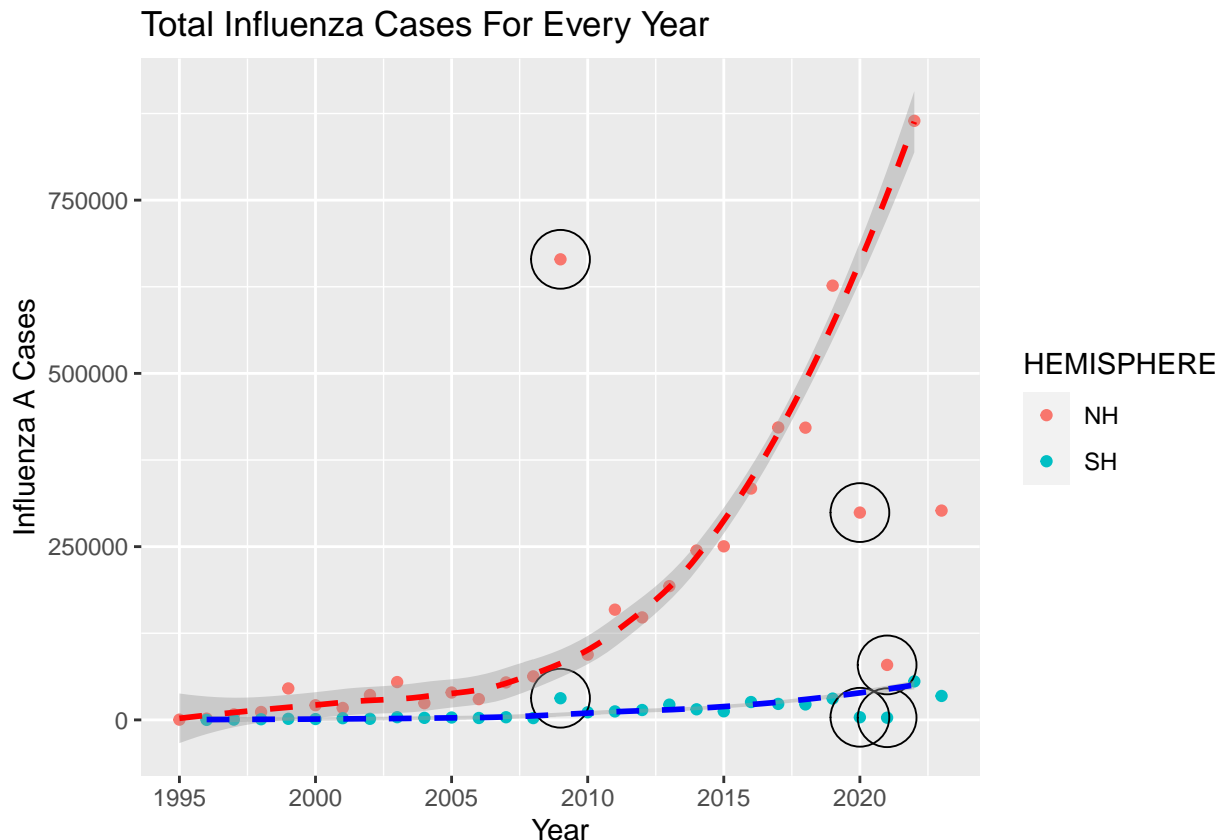
Question 8 [2 marks] Use the data to demonstrate a data visualization skill we have covered during Part I of the course. Choose a visualization relevant to your stated problem. Include your code in your submission. For example, you could visualize the distribution of our outcome with a histogram, or use a bar graph to represent the distribution of your exposure variable.

```
plotA <- fluA %>% group_by(ISO_YEAR, HEMISPHERE) %>%
  summarise(INFA_CASES = sum(na.omit(INF_A)))
```

```
## `summarise()` has grouped output by 'ISO_YEAR'. You can override using the
## `.groups` argument.
```

```
ggplot(plotA, aes(x=ISO_YEAR, y=INFA_CASES, col=HEMISPHERE)) +
  geom_point() +
  geom_point(data=plotA %>% filter(ISO_YEAR %in% c(2009, 2020, 2021)),
    pch=21,
    size=10,
    colour="black") +
  geom_smooth(data=plotA %>% filter(HEMISPHERE=="NH") %>%
    filter(!(ISO_YEAR %in% c(2009, 2020, 2021, 2023))),
    colour="red", linetype="dashed") +
  geom_smooth(data=plotA %>% filter(HEMISPHERE=="SH") %>%
    filter(!(ISO_YEAR %in% c(2009, 2020, 2021, 2023))),
    colour="blue", linetype="dashed") +
  labs(title="Total Influenza Cases For Every Year") +
  xlab("Year") +
  ylab("Influenza A Cases")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Question 9 [2 marks] Describe the skill that you are demonstrating and interpret your findings. For example, if you have created a histogram, describe the central tendency, shape of the distribution, etc.

Since our question is etiological, it is useful to divide our visualization between the northern and southern hemispheres as a form of comparison. In our visualization, we can observe some key events such as the 2009 influenza pandemic and the 2020 COVID-19 quarantine. We may consider these events outliers to the general trend of Influenza A and, as such, we will omit them in our loess line computation. Beyond these events, the general trend for influenza in the 21th century shows a positive and curved increase in influenza A over time in the Northern hemisphere. In the Southern hemisphere, the the number of influenza cases has not had the same rapid increase over time and has remained mostly constant relative to Northern hemisphere levels.