```r
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
```

```r
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(FieldName %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>
```

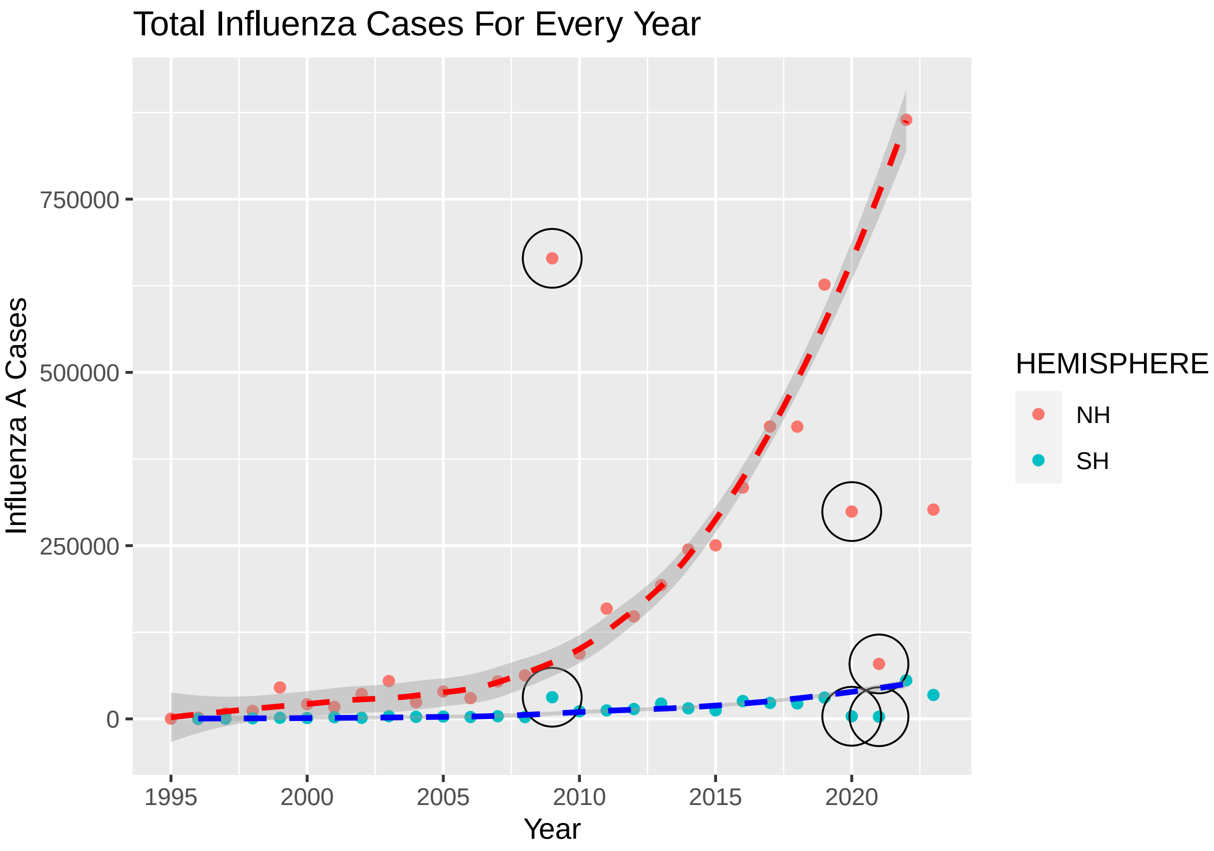
Question 8 [2 marks] Use the data to demonstrate a data visualization skill we have covered during Part I of the course. Choose a visualization relevant to your stated problem. Include your code in your submission. For example, you could visualize the distribution of our outcome with a histogram, or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

Part II

11. [2 marks] Calculate a marginal probability based on your outcome variable. Provide an equation (using probability notation) that describes this probability. For example, if my outcome variable is height in inches, I might calculate the probability that an individual in the dataset has a height of greater than 60 inches. $P(height >= 60)$. This would be a marginal probability. You may need to first add a new variable to your dataset to calculate your probability of interest, such as a binary variable indicating whether height is greater than 60 inches. There is a resource video about how to code such variables that could be helpful!

We calculated the probability that a randomly chosen country in the 2009 flu epidemic had a total number of influenza cases larger than 30. In probability notation this is given by P(INF_A > 30)

```
prob_over_30 <- (fluA %>% filter(ISO_YEAR == 2009, INF_A > 30) %>% nrow())/(
  fluA %>% filter(ISO_YEAR == 2009) %>% nrow())
prob_over_30
```

```
## [1] 0.3250173
```

12. [2 marks] Using any two variables in your dataset (or derived variables), calculate a conditional probability. Provide an equation (using probability notation) that describes this probability and then use R to calculate it.

We calculated the probability that a randomly chosen country in the 2009 flu epidemic had a total number of influenza cases larger than 30, given that the country is in the southern hemisphere.In probability notation this is given by $P(\text{INF\_A} > 30 \mid \text{HEMISPHERE} = \text{SH})$

```
prob_over_30_SH <- (fluA %>% filter(ISO_YEAR == 2009,
                                    HEMISPHERE == "SH",
                                    INF_A > 30) %>% nrow())/
  (fluA %>% filter(ISO_YEAR == 2009,
                   HEMISPHERE == "SH") %>% nrow())
prob_over_30_SH
```

```
## [1] 0.2717584
```

13. [2 marks] Does your dataset contain a continuous variable? If it does, does the distribution of that variable appear to be normal? Justify your answer using a plot. If your data does not contain a continous variable, give an example related to your dataset of a hypothetical variable that is continuous. That is, imagine what a continuous variable could be in relation to your dataset and topic of interest. For this hypothetical variable, describe what you imagine its shape might be, and how you would check whether or not it is normally distributed.

We could include a hypotehtical variable that measures the average temperature of that country for that particular ISO_YEAR. This measurement would be continuous because these temperatures can be measured to arbitrary accuracy. This measurement would be relevant because we are trying to see if there is higher prevalence of Influenza A in the northern vs southern hemispheres and temperature/weather patterns are a key difference between these two regions. If we plotted this temperature data on a histogram, and if it's Q-Q plot showed a linear line, we would consider the data to be normally distributed. We believe this data would be approximately normal because average temperatures for each country would include countries that are both colder and warmer. We have no reason to believe there would be a particular skew for warmer or colder countries, so the data would be approximately normal.

14. [4 marks] Does your dataset contain a binary variable? If so, does this variable meet the criteria to be considered binomially distributed? If so, describe this variable in terms of n and p. Calculate a probability based on this variable, first write the formula for the probability and then using R to calculate the probability (you do not need to calculate the probability by hand). If your data does not contain a binary variable, you can create one based on an underlying continuous variable or a categorical variable with > 2 levels to answer this question.

Yes our data set contains a a binary variable. The column Hemisphere has two possible outcomes either you are a country in the Northern Hemisphere or the Southern Hemisphere, which can be binomially distributed with our chances of success being either a country in the NH or SH, therefore our n would be the number of countries chosen and p would be the probability of the country being in the NH or vice versa p could be the probability of a country lying in the SH.If we choose, the probability of success to be country in the Nh then the formula for our probability would be NH/total # countries

```
knitr::include_graphics("partII_binomial expression.png")
```

$$P(X > 40) = 1 - P(X \leq 40) = 1 - \sum_{i=0}^{40} \mathrm{nCr}(50,i)\,(0.85)^{(i)}\,(1-0.85)^{(50-i)}$$

```
p_NH <- round(filter(fluA, HEMISPHERE=='NH')%>%nrow()/(fluA%>%nrow()),2)
p_greater30 <- pbinom(q = 40,size = 50, prob = p_NH,lower.tail=F)
p_greater30
```

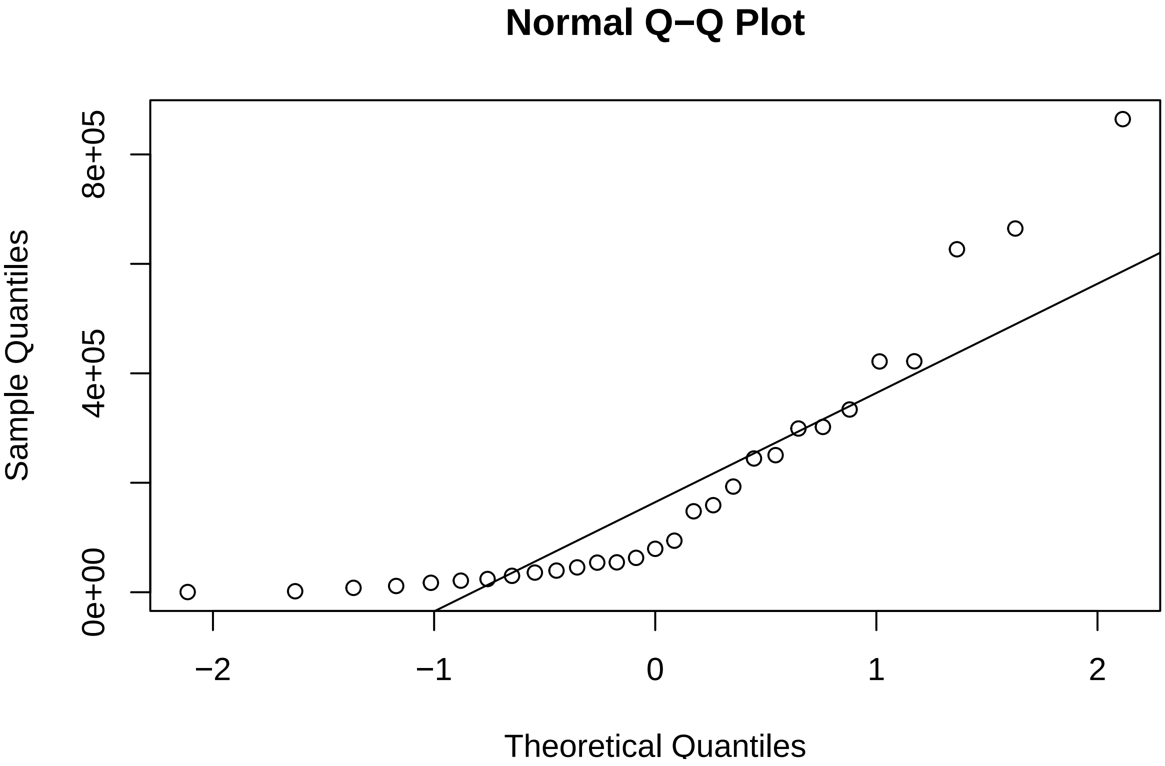
```
## [1] 0.7910937
```

16. [1 mark] Include parts I and II of your project.
17. [2 marks] Identify a statistical test to apply to your data. This must be a statistical test that we cover in part III of the course. Name the statistical test you have chosen and explain why this is the appropriate test for these data. For example, if I have pre- and post-intervention measurements of morning sleepiness recorded as a quantitative variable, I might choose a paired t test, because the paired t-test is appropriate for continuous outcome data in 2 groups that are inherently related.

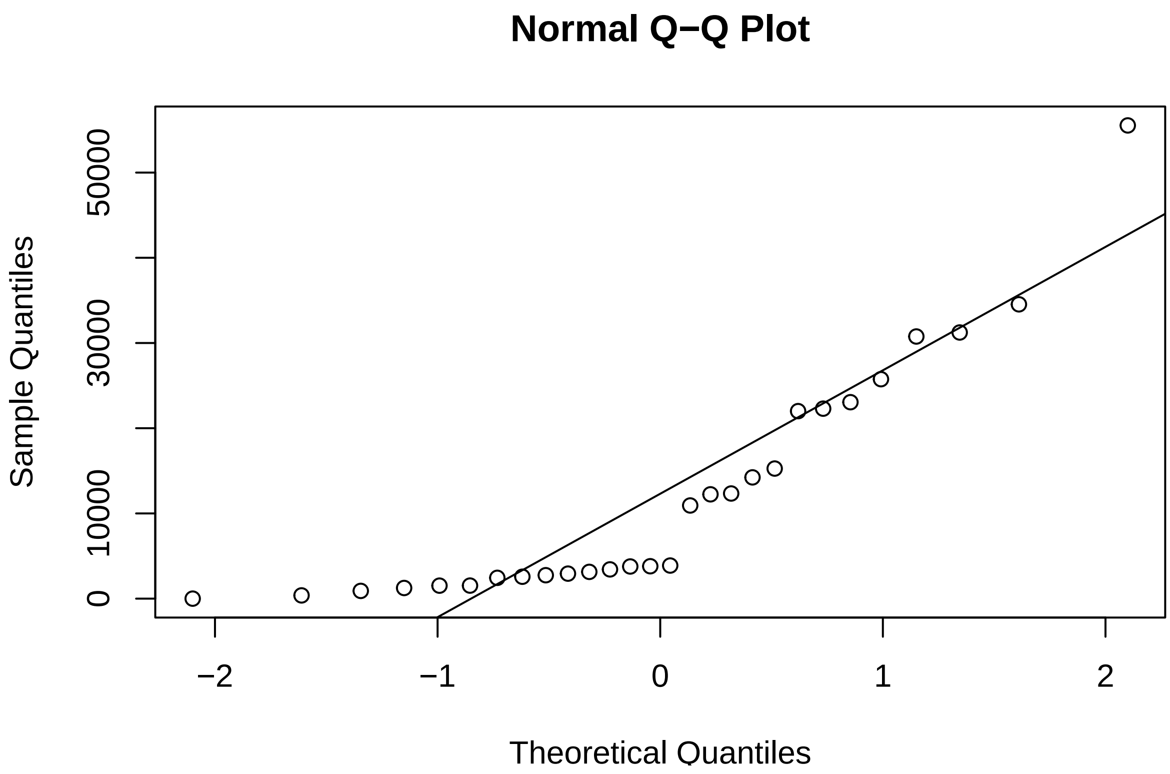
A statistical test to apply to our data is the independent samples t-test. This is because our data has two groups of people infected with influenza, the northern hemisphere and the southern hemisphere. In addition, the groups are independent of each other because they have no relationship with one another. The dependent variable being tested would be the infection rates between the two groups.

18. [2 marks] What assumptions are required by the testing method you chose? Are these assumptions met by your data? How did you assess this? For example, one of the assumptions of the t-test is that the data are normally distributed, so you might choose to assess this with a histogram, or a q-q plot.

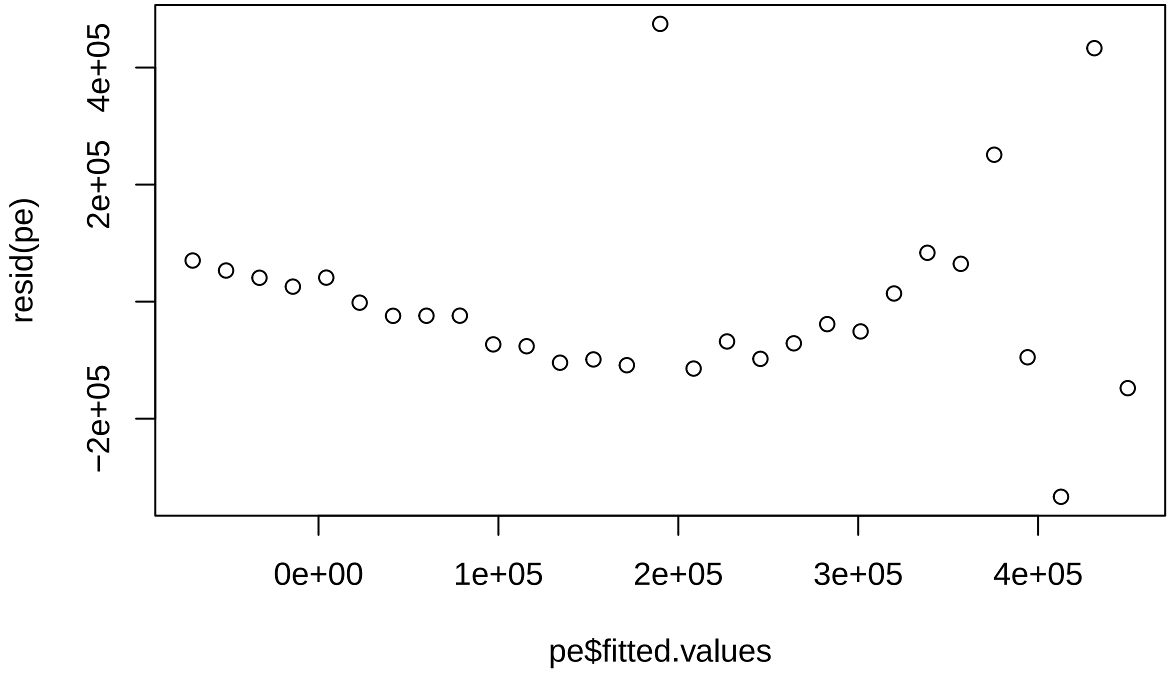
```
fluC <- filter(fluA, HEMISPHERE == "NH")
INFA_AB <- fluC %>% group_by(ISO_YEAR) %>% summarise(INFA_CASES = sum(na.omit(INF_A)))
fluF <- filter(fluA, HEMISPHERE == "SH")
INFA_AO <- fluF %>% group_by(ISO_YEAR) %>% summarise(INFA_CASES = sum(na.omit(INF_A)))

qqnorm(INFA_AB$INFA_CASES)
qqline(INFA_AB$INFA_CASES)
```



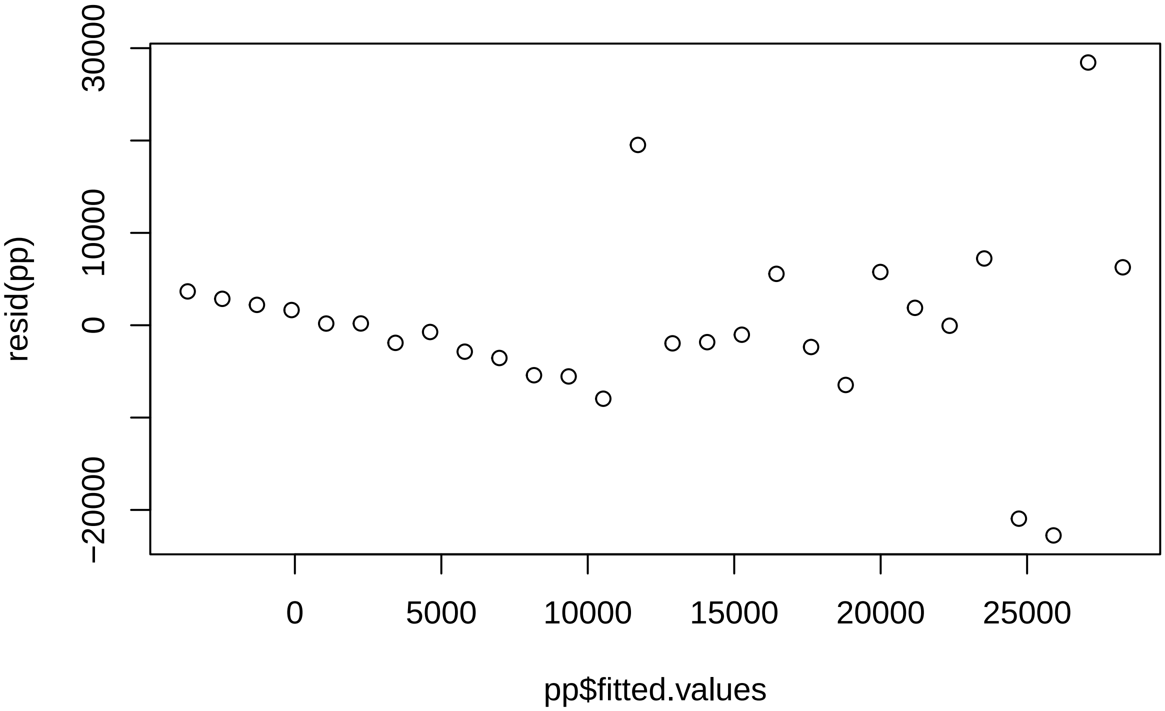
```
qqnorm(INFA_AO$INFA_CASES)
qqline(INFA_AO$INFA_CASES)
```



```
pe <-lm(INFA_CASES~ISO_YEAR,data=INFA_AB)
plot(pe$fitted.values, resid(pe))
```



```
pp <-lm(INFA_CASES~ISO_YEAR,data=INFA_AO)
plot(pp$fitted.values, resid(pp))
```



The first two plots show the QQ plots for the northern and southern hemispheres, respectively. The last two plots show the fitted vs residuals for the northern and southern hemispheres, respectively.

Assumptions required to make the independent samples t-test are that the observations within individuals must belong to one group, meaning the groups must be independent of each other. Also, the data within each group must be normally distributed, and there is near equal variance with no strong outliers. Our data meets the first assumption because each observation of influenza infection belongs to either the northern or southern hemisphere. Using a qq plot, we can see that the dataset for the hemispheres are not normally distributed because the best-fit lines do not resemble a linear relationship line at the 45 degree angle where y=x. In addition, there exists outliers in our data, seen by the outlier data points from the circled points in question 10 part B. We can calculate if there is a constant variance using a fitted vs residual plot. From the plot, both hemisphere datasets show a "fanning out" pattern, showing that the constant variance assumption does not hold. To resolve this issue, we normalized our data so they better fit the criteria for our test method. Specifically, we decided to use the log scale of the influenza case count.

19. [2 marks] Clearly state the null and alternative hypotheses for your test

The null hypothesis would be that there is no significant difference in influenza infection rates between the northern and southern hemispheres, while the alternative hypothesis is that there is a larger rate of influenza in the northern hemisphere vs the southern hemisphere. More precisely, our null hypothesis is that the average difference between the log(influenza cases) will be equal to zero.

20. [2 marks] Conduct the statistical test. Include the R code you used to generate your results. Annotate your code to help us follow your reasoning.

```r
fluA <- fluA %>% replace(is.na(.), 0) %>%
  mutate(addINF_A = INF_A+1) %>%
  mutate(logINF_A = log(addINF_A))
#In order to perform a log transformation (which we will need since the data
#has a skew), we interpret NA values as zero, we add 1 to all observations in
#order to eliminate the zeroes (since log(0)= -inf) and perform the transform

fluA.t <- t.test(fluA %>% filter(HEMISPHERE=="NH") %>% pull(logINF_A),
       fluA %>% filter(HEMISPHERE=="SH") %>% pull(logINF_A),
       alternative="greater")
fluA.t
```

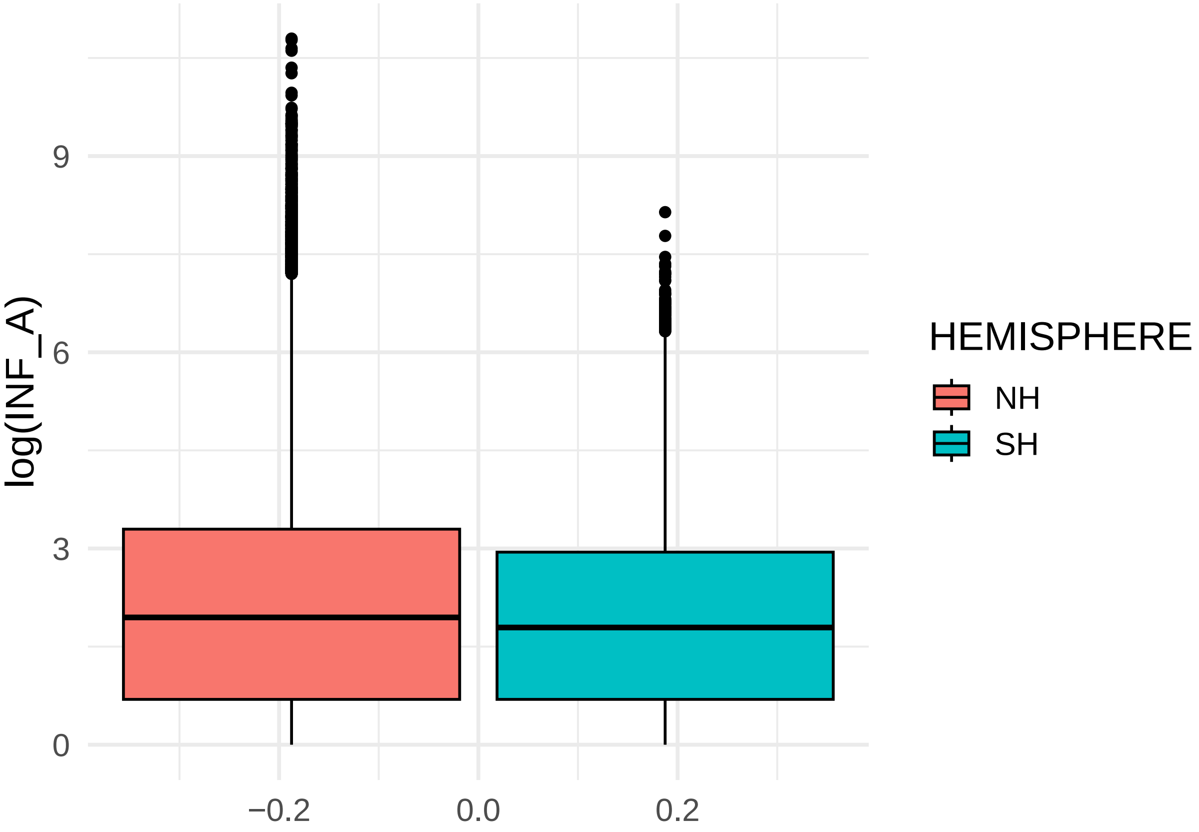
```
##
##  Welch Two Sample t-test
##
## data:  fluA %>% filter(HEMISPHERE == "NH") %>% pull(logINF_A) and fluA %>% filter(HEMISPHERE == "SH")
## t = -3.6917, df = 31585, p-value = 0.9999
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.05907441         Inf
## sample estimates:
## mean of x mean of y
##  1.157220  1.198086
```

```r
#comparing the number of Influenza A cases in the Northern Hemisphere and the
#number of cases in the southern hemisphere.
```

21. [4 marks] Present your results in a clear summary. This should include both a text summary and a table or figure with appropriate labeling. For example, if your outcome and predictor/exposure variables are both binary, this might be a 2x2 table. If your method was regression, you might present your regression line graphically. Include your code and annotations.

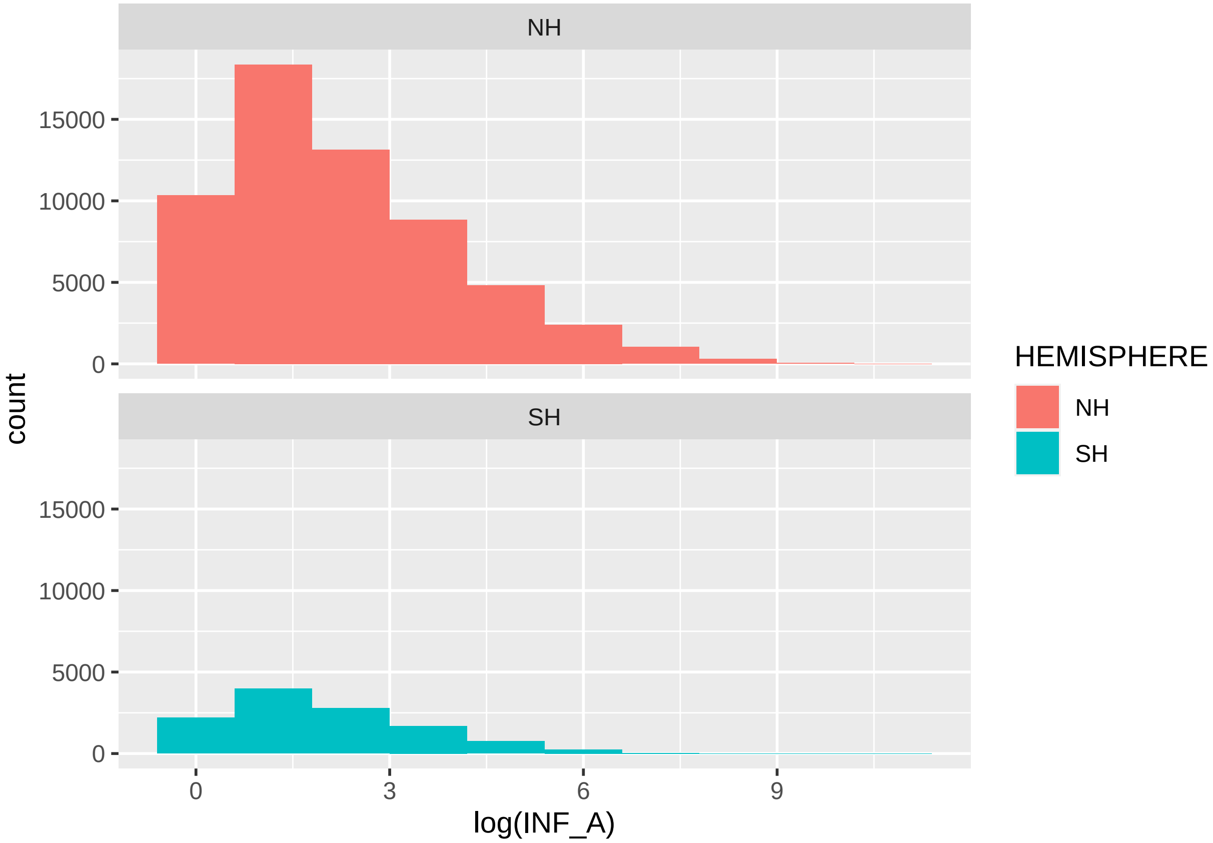
```
ggplot(na.omit(fluA),aes(y=log(INF_A))) +
  geom_boxplot(aes(fill=HEMISPHERE),col="black") +
  theme_minimal(base_size=15)
```

```
## Warning: Removed 76878 rows containing non-finite values (`stat_boxplot()`).
```



```
ggplot(na.omit(fluA),aes(x=log(INF_A))) +
  geom_histogram(bins=10,aes(fill=HEMISPHERE)) +
  facet_wrap(~HEMISPHERE,nrow=2)
```

```
## Warning: Removed 76878 rows containing non-finite values (`stat_bin()`).
```



```
#Box Plot and Histogram with a sqrt transformation on the number of influenza A
#recorded. A transformation was necessary to visualize since values vary widely.

data.frame(Parameter = c("Sample Mean NH","Sample Mean SH",
                         "tstat",
                         "Lower CI","Upper CI",
                         "pvalue"),
           Value = c(fluA.t$estimate,
                     fluA.t$statistic,
                     fluA.t$conf.int,
                     round(fluA.t$p.value,5)))
```

```
##        Parameter       Value
## 1 Sample Mean NH  1.15722038
## 2 Sample Mean SH  1.19808633
## 3          tstat -3.69171546
## 4       Lower CI -0.05907441
## 5       Upper CI         Inf
## 6         pvalue  0.99989000
```

The results show no significant difference between the log number of influenza cases between the northern hemisphere and the southern hemisphere. The p-value is very large (~1) which means we fail to reject the null hypothesis; it is likely that there is no difference in means between the number of cases of influenza in the northern hemisphere and the southern hemisphere. The 95% confidence interval tells us that 95% of our sampled test statistics will be greater than -0.06.

22. [4 marks] Interpret your findings. Include a statement about the evidence, your conclusions, and the generalizability of your findings. Our analyses and conclusions depend on the quality of our study design and the methods of data collection. Any missteps or oversights during the data collection process could potentially change the outcome of what we are trying to find. Consider the methods used to collect the data you analyzed. Was there any potential issue in how the participants were selected/recruited, retained, or assessed that may have impacted the outcome of your analysis/visualization? Were there any potential biases that you might be concerned about? Were there factors that were not measured or considered that you think could be important to the interpretation of these data?

Looking at both graphs it's clear that our data is skewed. For both influenza rates in the Northern and Southern Hemisphere our date is skewed to the right. Indicating some kind of outlier, which we can assume to possibly be from outbreaks of influenza during certain years. The skews can also indicate differences in base population size where countries with larger populations contributes more influenza cases and thus become outliers. To account for this skew we utilized log scale in q18 to normalize our data. The box plots do overlap so there doesn't seem to be a significant difference between influenza rates in the Northern and Southern Hemisphere. We verify this assessment with a t-test, which shows that there is no significant difference in influenza cases. What could have possibly influenced the results is the filtering of our data to only include data from the year 2009, a year that may have experienced influenza outbreak therefore increased influenza rates. What may have also influenced the results is the separation of our data between NH and SH which clearly have a large difference in population size. We did not compensate for these biases when interpreting our data. It is difficult to generalize these findings accross all types of influenza, since we only performed an analysis on Influenza A. There may be different results for different strains and influenza A just happens to be a strain that has no particular prefrence between the two hemispheres.

23. [1 mark] Create a statement of contribution. This is now common in journal articles. For example, the American Journal of Epidemiology provides the following instructions to authors: “Authorship credit should be based on criteria developed by the International Committee for Medical Journal Editors (ICMJE):

1) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data;

2) drafting the article or reviewing it and, if appropriate, revising it critically for important intellectual content;

3) final approval of the version to be published. Authors should meet all conditions. In addition, each author must certify that he or she has participated sufficiently in the work to believe in its overall validity and to take public responsibility for appropriate portions of its content. Author names should be listed in ScholarOne and author contributions should be detailed in the cover letter (e.g., “Author A designed the study and directed its implementation, including quality assurance and control. Author B helped supervise the field activities and designed the study’s analytic strategy. Author C helped conduct the literature review and prepare the Methods and the Discussion sections of the text.”). An example from a recent issue of the BMJ (Woolf, Masters, and Aron BMJ 2021;373:n1343): “Contributors: SHW led the production of this manuscript and had primary responsibility for the composition. He is guarantor. RKM contributed revisions and had primary responsibility for data acquisition and analysis, the modeling results that form the basis for this study, and production of the supplementary 2 material. LYA contributed revisions and had primary responsibility for dealing with the study’s policy implications in the discussion section. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.” For your project, please craft a statement indicating the contributions of each group member. If your group divided the assignment responsibilities by question you may use question numbers to indicate which member had primary responsibility for each question (for example: Member XX had primary responsibility for questions x,x,x. . .).

Ze Yu Li designed the statistical test and the hypotheses, variables and assumptions used, helped implement the data graphs to support/refute the assumptions. He was primarily responsible for questions 17-19 in part three of the data project.

Yosola Olakunle was in charge of the interpretation and analysis of the results in question 22. Analyzed graphs and other data closely to spot any trends or significant differences. Using this information she was able to generate reasonable conclusions and also reflect on our data collection methods to account for any mistakes, biases, or left out information.

Vincent Tolentino performed the initial test in question 20 and the subsequent summarization of the results in question 21. Research into literature was performed to ensure a valid test was performed on the log transformed data and that the statistics generated from the test were correct. He also performed most of the formatting in the published version and the consolidating of work.

Ziqi Shi was responsible for the proofreading of the part. He also advised taking the log transformation on the influenza count so that the data better fits a normal distribution and fit the test method criteria. He also helped spotting potential oversights that might have happened through the project to put in the conclusion.