STA 518 Final Project

Russell Marvin

2022-12-10

# Loading packages, data, cleaning and exploring

## Loading Packages

library(tidyverse)  
library(lubridate)  
library(knitr)  
library(skimr)  
library(readr)  
library(rnoaa)  
library(broom)  
library(stringr)

## Importing data

#GHG emissions for 194 countries and EU from 1990-2018, including emissions of 6 major GHGs. Non-CO2 emissions are expressed in CO2 equivalents using 100-year global warming potential values from IPCC Fourth Assessment Report:  
ghg\_emissions <- read\_csv("ghg-emissions.csv")  
#Global mean CO2 dry air mole fraction defined as the num of molecules CO2/num all molecules in air WITHOUT WATER VAPOR. (expressed as ppm so .000400 = 400ppm). Data averages from marine surface sites around the globe:  
CO2\_annual\_mean <- read\_csv("CO2AnnualMean-glo.csv")  
#Global mean sea level absolute change in mm:  
sea\_level\_mean <- read\_csv("GlobalMeanSeaLevel93-14.csv")  
#Average global mean temperature anomalies in degrees Celsius relative to a base period. GISTEMP base period: 1951-1980. GISS Surface Temp Analysis. GCAG base period: 20th century average.:  
mean\_temp\_anom <- read\_csv("GlobalMeanTempAnomolies1880-2016.csv")  
#Area and extent of sea ice from NOAA  
#Area defined as area 100% covered by sea ice, extent defined as area at least 15% covered by sea ice  
sea\_ice <- sea\_ice\_tabular()

## let’s tidy up and look at our data

#starting with ghg\_emissions, we want one row for each year, plus add a "sum" column including all countries' totals  
  
#first step is to remove 1990 (it includes several 'false' values), the 'unit' column, and the 'data source' row (full of NAs)  
ghg\_emissions\_2 <- ghg\_emissions %>% select(-('unit':'1990')) %>% na.omit()  
#next, we pivot\_longer to give all country years (eg. China 1995) a row, then use group\_by and summarize to create a 'CO2\_total' for each year  
ghg\_emissions\_tidy <- ghg\_emissions\_2 %>% pivot\_longer(c('1991':'2018'), names\_to = "year", values\_to = "MtCO2e") %>% group\_by(year) %>% summarize(MtCO2e\_total = sum(MtCO2e))  
glimpse(ghg\_emissions\_tidy)

## Rows: 28  
## Columns: 2  
## $ year <chr> "1991", "1992", "1993", "1994", "1995", "1996", "1997", "…  
## $ MtCO2e\_total <dbl> 31918.35, 31788.11, 31879.13, 32061.53, 32755.11, 33090.8…

skim(ghg\_emissions\_tidy)

Data summary

|  |  |
| --- | --- |
| Name | ghg\_emissions\_tidy |
| Number of rows | 28 |
| Number of columns | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| numeric | 1 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| year | 0 | 1 | 4 | 4 | 0 | 28 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| MtCO2e\_total | 0 | 1 | 38706.13 | 5329.96 | 31788.11 | 33862.78 | 38568.9 | 43570.3 | 47260.82 | ▇▂▂▅▅ |

#now on to CO2\_annual\_mean,   
#for now we will leave the uncertainty column, though it's the same for every observation  
#Since we only have one observation per year, let's create a Year column with just the Year  
CO2\_annual\_mean <- CO2\_annual\_mean %>% mutate(YearOnly = year(Year))  
glimpse(CO2\_annual\_mean)

## Rows: 38  
## Columns: 4  
## $ Year <date> 1980-11-01, 1981-11-01, 1982-11-01, 1983-11-01, 1984-11-0…  
## $ Mean <dbl> 338.80, 340.00, 340.77, 342.44, 343.99, 345.47, 346.87, 34…  
## $ Uncertainty <dbl> 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1…  
## $ YearOnly <dbl> 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989…

skim(CO2\_annual\_mean)

Data summary

|  |  |
| --- | --- |
| Name | CO2\_annual\_mean |
| Number of rows | 38 |
| Number of columns | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| Date | 1 |
| numeric | 3 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: Date**

| skim\_variable | n\_missing | complete\_rate | min | max | median | n\_unique |
| --- | --- | --- | --- | --- | --- | --- |
| Year | 0 | 1 | 1980-11-01 | 2017-11-01 | 1999-05-02 | 38 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mean | 0 | 1 | 368.34 | 19.63 | 338.8 | 353.09 | 366.59 | 384.26 | 405.0 | ▇▇▆▆▅ |
| Uncertainty | 0 | 1 | 0.10 | 0.00 | 0.1 | 0.10 | 0.10 | 0.10 | 0.1 | ▁▁▇▁▁ |
| YearOnly | 0 | 1 | 1998.50 | 11.11 | 1980.0 | 1989.25 | 1998.50 | 2007.75 | 2017.0 | ▇▇▇▇▇ |

#next is sea\_level\_mean, which looks good as well, tidy already.   
glimpse(sea\_level\_mean)

## Rows: 22  
## Columns: 2  
## $ Time <date> 1993-03-15, 1994-03-15, 1995-03-15, 1996-03-15, 1997-03-15, 1998…  
## $ GMSL <dbl> 1.4, 2.7, 5.7, 11.4, 16.1, 21.9, 22.1, 25.0, 29.6, 33.3, 35.4, 37…

skim(sea\_level\_mean)

Data summary

|  |  |
| --- | --- |
| Name | sea\_level\_mean |
| Number of rows | 22 |
| Number of columns | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| Date | 1 |
| numeric | 1 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: Date**

| skim\_variable | n\_missing | complete\_rate | min | max | median | n\_unique |
| --- | --- | --- | --- | --- | --- | --- |
| Time | 0 | 1 | 1993-03-15 | 2014-03-15 | 2003-09-14 | 22 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| GMSL | 0 | 1 | 35.72 | 20.95 | 1.4 | 21.95 | 36.5 | 52.27 | 71.3 | ▅▅▇▅▃ |

#For mean\_temp\_anom, we'll take out one measurement of temp but leave all the many observations for now  
mean\_temp\_anom\_tidy <- mean\_temp\_anom %>% filter(Source == "GISTEMP")  
glimpse(mean\_temp\_anom\_tidy)

## Rows: 137  
## Columns: 3  
## $ Source <chr> "GISTEMP", "GISTEMP", "GISTEMP", "GISTEMP", "GISTEMP", "GISTEMP…  
## $ Year <dbl> 2016, 2015, 2014, 2013, 2012, 2011, 2010, 2009, 2008, 2007, 200…  
## $ Mean <dbl> 0.99, 0.87, 0.74, 0.65, 0.63, 0.60, 0.71, 0.64, 0.54, 0.66, 0.6…

skim(mean\_temp\_anom\_tidy)

Data summary

|  |  |
| --- | --- |
| Name | mean\_temp\_anom\_tidy |
| Number of rows | 137 |
| Number of columns | 3 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| numeric | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Source | 0 | 1 | 7 | 7 | 0 | 1 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 0 | 1 | 1948.00 | 39.69 | 1880.00 | 1914.00 | 1948.00 | 1982.00 | 2016.00 | ▇▇▇▇▇ |
| Mean | 0 | 1 | 0.02 | 0.33 | -0.47 | -0.21 | -0.07 | 0.19 | 0.99 | ▇▇▃▃▁ |

#lastly, we have sea\_ice. First let's remove NA observations (represented as -9999). Then we'll look at it  
sea\_ice\_tidy <- sea\_ice %>% filter(data.type != "-9999", extent != -9999.00, area != -9999.00)  
#let's also change the N/S to North/South to make the region column more clear to interpret  
sea\_ice\_tidy <- sea\_ice\_tidy %>% mutate(region\_full = str\_replace\_all(region, pattern = c(N = "North", S = "South"))) %>% select(-region)  
glimpse(sea\_ice\_tidy)

## Rows: 1,053  
## Columns: 6  
## $ year <int> 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1989…  
## $ mo <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…  
## $ data.type <chr> "Goddard", "Goddard", "Goddard", "Goddard", "Goddard", "Go…  
## $ extent <dbl> 15.41, 14.86, 14.91, 15.18, 14.94, 14.47, 14.72, 14.89, 14…  
## $ area <dbl> 12.41, 11.94, 11.91, 12.19, 12.01, 11.68, 11.69, 11.88, 11…  
## $ region\_full <chr> "North", "North", "North", "North", "North", "North", "Nor…

skim(sea\_ice\_tidy)

Data summary

|  |  |
| --- | --- |
| Name | sea\_ice\_tidy |
| Number of rows | 1053 |
| Number of columns | 6 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 2 |
| numeric | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| data.type | 0 | 1 | 7 | 7 | 0 | 2 | 0 |
| region\_full | 0 | 1 | 5 | 5 | 0 | 2 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| year | 0 | 1 | 2000.44 | 12.73 | 1978.00 | 1990.00 | 2000.00 | 2011.00 | 2022.00 | ▇▇▇▇▇ |
| mo | 0 | 1 | 6.51 | 3.45 | 1.00 | 4.00 | 7.00 | 10.00 | 12.00 | ▇▅▅▅▇ |
| extent | 0 | 1 | 11.48 | 4.58 | 2.16 | 7.62 | 12.11 | 15.21 | 19.76 | ▃▅▆▇▃ |
| area | 0 | 1 | 9.03 | 3.98 | 1.35 | 5.43 | 9.75 | 12.56 | 15.75 | ▅▆▅▇▆ |

Thankfully we’re now working with data that does no have any missing values.

## Some EDA

Let’s check out how the data.type variable related with the region variable.

#frequency table using https://www.statology.org/two-way-table-in-r/   
sea\_ice\_freq\_table <- table(sea\_ice\_tidy$region\_full, sea\_ice\_tidy$data.type)  
sea\_ice\_freq\_table

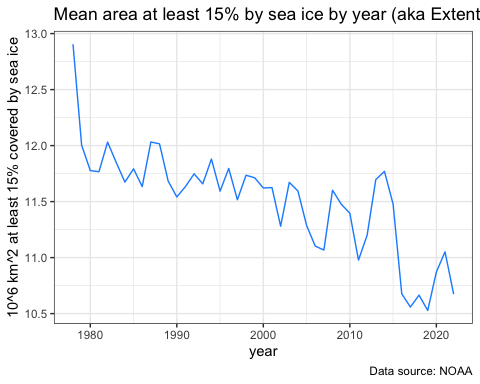
##   
## Goddard NRTSI-G  
## North 515 11  
## South 516 11

There was roughly the same number of each type of observation per region, with one extra Goddard data.type measurement in the Southern region.

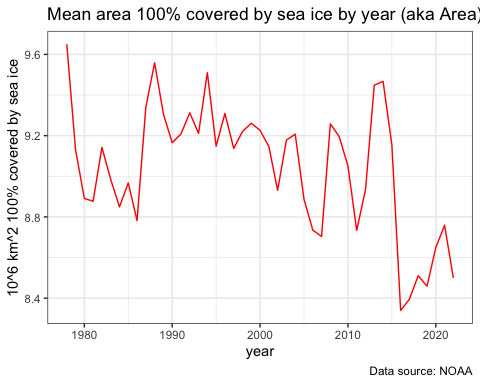
# Visualizations

We will first calculate and plot mean area and extent of sea ice measurements for each year to look at trends.

#let's calculate mean area and extent for each year to look at yearly trends!  
sea\_ice\_yearly\_means <- sea\_ice\_tidy %>% group\_by(year) %>% summarize(extent\_mean = mean(extent), area\_mean = mean(area))   
  
#now to plot each of them  
sea\_ice\_yearly\_means %>% ggplot(aes(x= year,  
 y= extent\_mean))+  
 geom\_line(color = "dodgerblue") +   
 labs(title = "Mean area at least 15% by sea ice by year (aka Extent)",  
 y = "10^6 km^2 at least 15% covered by sea ice",  
 caption = "Data source: NOAA")+  
 theme\_bw()



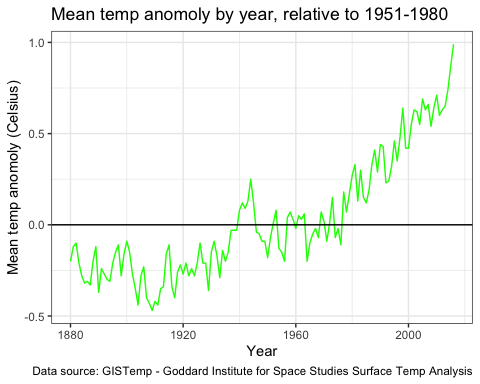
sea\_ice\_yearly\_means %>% ggplot(aes(x= year,  
 y= area\_mean))+  
 geom\_line(color = "red") +   
 labs(title = "Mean area 100% covered by sea ice by year (aka Area)",  
 y = "10^6 km^2 100% covered by sea ice",  
 caption = "Data source: NOAA") +   
 theme\_bw()



Extent seems to have a clearer downward trend, though both do decrease over time.

How about our mean\_temp\_anom\_tidy over time?

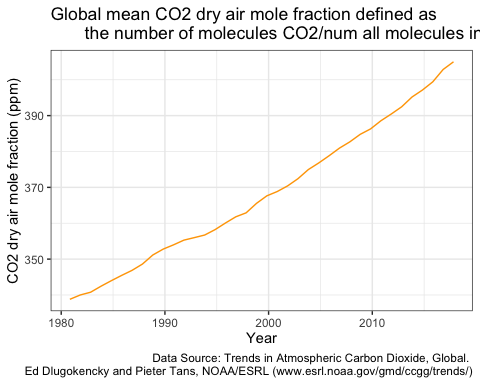
#how about our mean\_temp\_anom\_tidy over time?  
mean\_temp\_anom\_tidy %>% ggplot(aes(x= Year,  
 y= Mean))+  
 geom\_line(color = "green") +   
 geom\_hline(yintercept = 0) +  
 labs(title = "Mean temp anomoly by year, relative to 1951-1980",  
 y = "Mean temp anomoly (Celsius)",  
 caption = "Data source: GISTemp - Goddard Institute for Space Studies Surface Temp Analysis") +   
 theme\_bw()



A clear increase over the baseline period, with a particular spike around the 2010-2020 time period.

How does our CO2 measurement change over time?

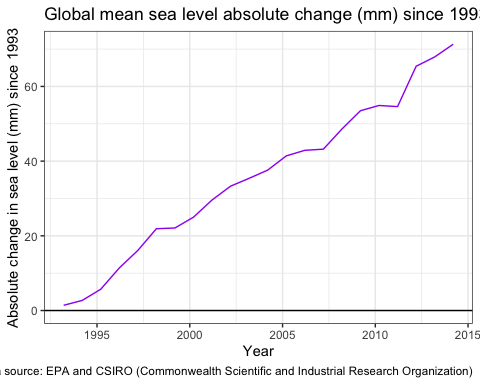
CO2\_annual\_mean %>% ggplot(aes(x= Year,  
 y= Mean))+  
 geom\_line(color = "orange") +  
 labs(title = "Global mean CO2 dry air mole fraction defined as   
 the number of molecules CO2/num all molecules in air w/o water vapor ",  
 y = "CO2 dry air mole fraction (ppm)",  
 caption = "Data Source: Trends in Atmospheric Carbon Dioxide, Global.   
 Ed Dlugokencky and Pieter Tans, NOAA/ESRL (www.esrl.noaa.gov/gmd/ccgg/trends/)") +   
 theme\_bw()



CO2 levels in the atmosphere can be seen increasing over the last 30 years.

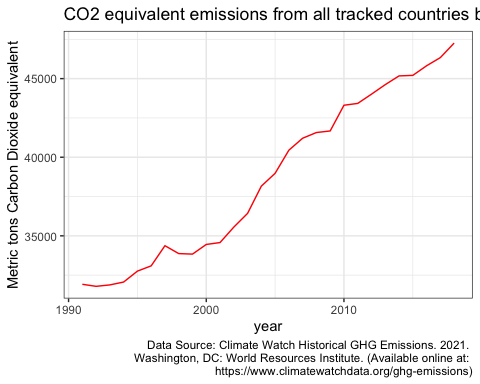
Let’s take a look at sea\_level\_mean, or absolute change in sea level since the beginning of the first year measured.

sea\_level\_mean %>% ggplot(aes(x= Time,  
 y= GMSL))+  
 geom\_line(color = "purple") +   
 geom\_hline(yintercept = 0) +  
 labs(title = "Global mean sea level absolute change (mm) since 1993",  
 x = "Year",  
 y = "Absolute change in sea level (mm) since 1993",  
 caption = "Data source: EPA and CSIRO (Commonwealth Scientific and Industrial Research Organization)") +   
 theme\_bw()



And finally our greenhouse gas measurements (expressed as metric tons of CO2 equivalents, based on the global warming potential of each gas included).

#first we have to coerce our year variable to numeric, as opposed to character  
ghg\_emissions\_tidy$year <- as.numeric(ghg\_emissions\_tidy$year)  
  
ghg\_emissions\_tidy %>% ggplot(aes(x= year,  
 y= MtCO2e\_total))+  
 geom\_line(color = "red") +  
 labs(title = "CO2 equivalent emissions from all tracked countries by year",  
 y = "Metric tons Carbon Dioxide equivalent",  
 caption = "Data Source: Climate Watch Historical GHG Emissions. 2021.   
 Washington, DC: World Resources Institute. (Available online at:   
 https://www.climatewatchdata.org/ghg-emissions)") +   
 theme\_bw()



A clear increase in yearly MtCO2 equivalent emissions over the last 30 years can be seen.

# Merging tables and some statistical tests

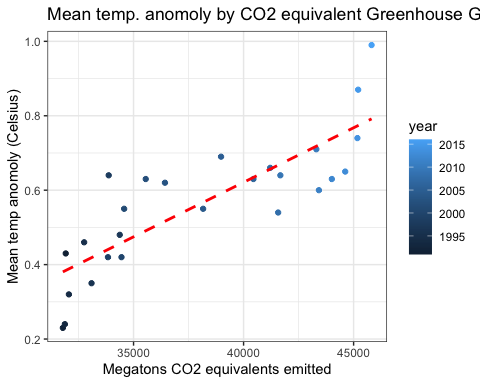
Let’s merge some tables

#first we'll combine our greenhouse gas emissions with mean temperature anomalies ONLY where years overlap   
ghg\_temp\_merged <- ghg\_emissions\_tidy %>%  
 inner\_join(mean\_temp\_anom\_tidy, c("year" = "Year"))  
#check out our resulting data frame  
slice\_head(ghg\_temp\_merged, n=24)

## # A tibble: 24 × 4  
## year MtCO2e\_total Source Mean  
## <dbl> <dbl> <chr> <dbl>  
## 1 1991 31918. GISTEMP 0.43  
## 2 1992 31788. GISTEMP 0.23  
## 3 1993 31879. GISTEMP 0.24  
## 4 1994 32062. GISTEMP 0.32  
## 5 1995 32755. GISTEMP 0.46  
## 6 1996 33091. GISTEMP 0.35  
## 7 1997 34373. GISTEMP 0.48  
## 8 1998 33871. GISTEMP 0.64  
## 9 1999 33838. GISTEMP 0.42  
## 10 2000 34452. GISTEMP 0.42  
## # … with 14 more rows

And visualize the results:

ghg\_temp\_merged %>% ggplot(aes(x = MtCO2e\_total,  
 y = Mean,   
 color = year)) +  
 geom\_point() +  
 geom\_smooth(method = "lm",  
 se = F,   
 color = "red",  
 linetype = "dashed") +  
 labs(title = "Mean temp. anomoly by CO2 equivalent Greenhouse Gas Emissions",  
 x = "Megatons CO2 equivalents emitted" ,  
 y = "Mean temp anomoly (Celsius)") +   
 theme\_bw()

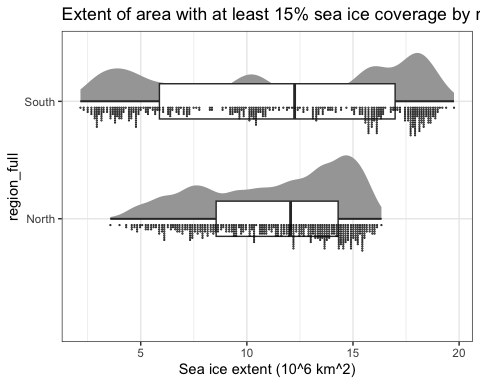


We can see that Mean temp anomoly (Celsius) increases as greenhouse gas emissions increases.

## 2b.

Let’s look at the difference between North and South observations for sea ice coverage (by extent). Our null hypothesis is that the means are the same, the alternative hypothesis will be that there is a difference in means. We’ll use a One-Way ANOVA.

#let's check out the distribution of these data first  
sea\_ice\_tidy %>% ggplot(aes(x = region\_full, y = extent)) +  
 ggdist::stat\_halfeye(adjust = .5, width = 2\*.3, .width = c(0.5, 1)) +   
 geom\_boxplot(width = .3, outlier.shape = NA) +  
 ggdist::stat\_dots(side = "left", dotsize = 1, justification = 1.05, binwidth = .1,  
 color = "black") +  
 coord\_flip() +  
 labs(y = "Sea ice extent (10^6 km^2)",  
 title = "Extent of area with at least 15% sea ice coverage by region") +   
 theme\_bw() +  
 theme(legend.position = "none")



We’ll perform a randomization test for to check if there is a difference between N and S region extents.

#  
myData <- sea\_ice\_tidy %>% select(region\_full, extent)  
  
# Fitting One-Way ANOVA model  
modFit <- aov(extent ~ region\_full, data = myData)  
Fstatistic <- modFit %>% tidy() %>% slice\_head(n = 1) %>% pull(statistic)  
  
  
# Getting number of each observations in each group  
groupCounts <- myData %>% count(region\_full)  
groupCounts

## # A tibble: 2 × 2  
## region\_full n  
## <chr> <int>  
## 1 North 526  
## 2 South 527

N <- nrow(myData)  
N

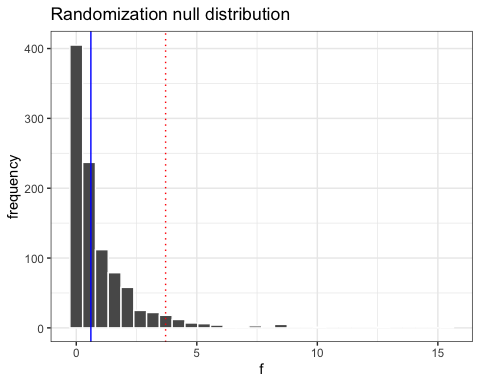
## [1] 1053

#num of permutations   
nperms <- 1000  
  
# Instantiating vector for test statistics  
permFs <- vector(length = nperms)  
  
  
# Create vector of group memberships of individuals  
groups <- rep(groupCounts$region\_full, times = groupCounts$n)  
  
  
for(p in 1:nperms) {  
# Permute individuals keeping group sizes the same as in original data  
permData <- myData %>% mutate(region\_full = groups[sample(1:N, size = N, replace = FALSE)])  
  
# Calculate F test statistic for each permutation  
modFit <- aov(extent ~ region\_full, data = permData)  
permFs[p] <- modFit %>% tidy() %>% slice\_head(n = 1) %>% pull(statistic)  
}  
  
head(permFs)

## [1] 12.958710256 0.288792096 0.217782803 0.003729084 2.157510443  
## [6] 0.058646509

Now, we’ll show the null distribution of the F-statistic for our One-way ANOVA randomization test.

tibble(f = permFs) %>% ggplot(aes(x = f)) +  
 geom\_histogram(color = "white") +  
 geom\_vline(xintercept = quantile(permFs, probs = 0.950),  
 color = "red", linetype = "dotted")+  
 geom\_vline(xintercept = Fstatistic,  
 color = "blue", linetype = "solid") +  
 labs(title = "Randomization null distribution",  
 y = "frequency") +  
 theme\_bw()



Our observed F-statistic is the blue line, and the red dotted line marks the 95th percentile.

To find the p-value of the randomization test, we want to know the proportion of randomized F-statistics that are greater than or equal to the observed F-statistic.

randPvalue <- mean(permFs >= Fstatistic)  
randPvalue

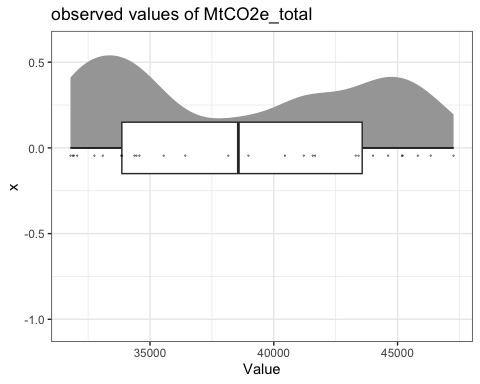
## [1] 0.424

Therefore, we fail to reject the null hypothesis. This was expected, as we didn’t see any difference in the raincloud plots. We do not have evidence at the 5% significance level that the extent North region differs from South region, on average.

Next, we will obtain a parametric and nonparametric bootstrap-estimated standard error for at least one statistic of interest.

Let’s calculate the median year’s megatons CO2 equivalent emissions in our dataset ghg\_emissions\_tidy.

tibble(Value = ghg\_emissions\_tidy$MtCO2e\_total) %>% ggplot(aes(y = Value)) +  
 ggdist::stat\_halfeye(adjust = .5, width = 2\*.3, .width = c(0.5, 1)) +   
 geom\_boxplot(width = .3, outlier.shape = NA) +  
 ggdist::stat\_dots(side = "left", dotsize = 6, justification = 1.05, binwidth = .1,  
 color = "black") +  
 coord\_flip() +  
 labs(title = "observed values of MtCO2e\_total") +   
 theme\_bw() +  
 theme(legend.position = "none")



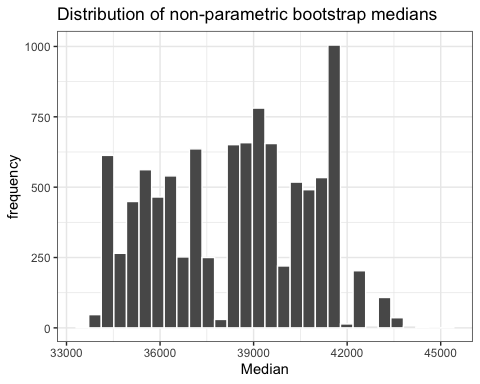
n <- nrow(ghg\_emissions\_tidy)  
#sample median:   
sample\_med <- median(ghg\_emissions\_tidy$MtCO2e\_total)

We can see that the data is bimodal and asymmetric. We have a sample size of 28

Our sample median is 3.8568905^{4}

Now let’s perform a bootstrap

# Number of bootstrap samples  
B <- 10000  
  
# Instantiating matrix for bootstrap samples  
boots <- matrix(NA, nrow = n, ncol = B)  
  
# Sampling with replacement B times  
for(b in 1:B) {  
boots[, b] <- ghg\_emissions\_tidy$MtCO2e\_total[sample(1:n, size = n, replace = TRUE)]  
}  
  
#Using the generated bootstrap samples, let's create a bootstrap distribution of sample medians, and visualize this distribution using a histogram.  
  
# Instantiating vector for bootstrap medians  
bootMedians <- vector(length = B)  
  
# Calculating the median for each of the B resamples  
for(b in 1:B) {  
bootMedians[b] <- median(boots[,b])  
}  
  
#visualizing bootMedians:  
tibble(Median = bootMedians) %>% ggplot(aes(x = Median)) +  
 geom\_histogram(color = "white") +  
 labs(title = "Distribution of non-parametric bootstrap medians",  
 y = "frequency") +  
 theme\_bw()



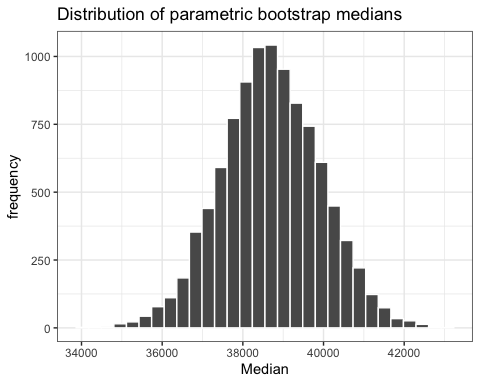
# Using the bootstrap samples to obtain a nonparametric estimate of the standard error of the sample median.   
SEestimate <- sd(bootMedians)  
  
  
#Next, we'll use the bootstrap samples to obtain a nonparametric 95% confidence interval for the population median.  
lowerBoundMed <- quantile(bootMedians, probs = 0.025)  
   
upperBoundMed <- quantile(bootMedians, probs = 0.975)

Our non-parametric bootstrap-estimated standard error for the median year in MtCO2e\_total (total CO2 equivalent emissions in megatons) is 2430.8589418

We are 95% confident that the true median is between 3.441263^{4} and 4.248878^{4}.

Now for our parametric bootstrap-estimated SE…

#we'll assume a normal distribution for the sake of the parametric bootstrap estimate. Using 10000 samples again  
B <- 10000  
  
# Instantiating matrix for bootstrap samples  
paramBoots <- matrix(NA, nrow = n, ncol = B)  
Xbar <- mean(ghg\_emissions\_tidy$MtCO2e\_total)  
s <- sd(ghg\_emissions\_tidy$MtCO2e\_total)  
  
# Simulating a normal set of n values, B times  
for(b in 1:B) {  
paramBoots[, b] <- rnorm(n = n, mean = Xbar, sd = s)  
}  
  
# Instantiating vector for bootstrap medians  
bootParamMedians <- vector(length = B)  
  
# Next we calculate the median for each simulated data set  
for(b in 1:B) {  
bootParamMedians[b] <- median(paramBoots[,b])  
}  
  
#visualizing our distribution iwth a histogram  
tibble(Median = bootParamMedians) %>% ggplot(aes(x = Median)) +  
 geom\_histogram(color = "white") +  
 labs(title = "Distribution of parametric bootstrap medians",  
 y = "frequency") +  
 theme\_bw()



#We find a parametric bootstrap estimate of the standard error of the sample median.  
SEparamEstimate <- sd(bootParamMedians)  
  
  
#Use the bootstrap samples to obtain a parametric 95% confidence interval for the sample median.  
lowerBoundParaMed <- quantile(bootParamMedians, probs = 0.025)  
   
upperBoundParaMed <- quantile(bootParamMedians, probs = 0.975)

Our parametric bootstrap-estimated standard error for the median year in MtCO2e\_total (total CO2 equivalent emissions in megatons) is 1228.4863672

We are 95% confident that the true median is between 3.629137^{4} and 4.109177^{4}.

# Next, to create a data dictionary showcasing the variables used in our analyses.

# Creating data dictionary for ghg\_emissions\_tidy  
dataDictionary\_ghg <- tibble(Variable = colnames(ghg\_emissions\_tidy),  
 Description = c("Year","CO2 equivalent emissions emitted (megatons)"),  
 Type = map\_chr(ghg\_emissions\_tidy, .f = function(x){typeof(x)[1]}),  
 Class = map\_chr(ghg\_emissions\_tidy, .f = function(x){class(x)[1]}))

# Printing nicely in R Markdown   
flextable::flextable(dataDictionary\_ghg, cwidth = 2)

| Variable | Description | Type | Class |
| --- | --- | --- | --- |
| year | Year | double | numeric |
| MtCO2e\_total | CO2 equivalent emissions emitted (megatons) | double | numeric |

# Creating data dictionary for sea\_level\_mean  
dataDictionary\_sea\_level <- tibble(Variable = colnames(sea\_level\_mean),  
 Description = c("Year,month,day","Global mean sea level rise (mm)"),  
 Type = map\_chr(sea\_level\_mean, .f = function(x){typeof(x)[1]}),  
 Class = map\_chr(sea\_level\_mean, .f = function(x){class(x)[1]}))

# Printing nicely in R Markdown   
flextable::flextable(dataDictionary\_sea\_level, cwidth = 2)

| Variable | Description | Type | Class |
| --- | --- | --- | --- |
| Time | Year,month,day | double | Date |
| GMSL | Global mean sea level rise (mm) | double | numeric |

# Creating data dictionary for sea\_ice\_tidy  
dataDictionary\_sea\_ice <- tibble(Variable = colnames(sea\_ice\_tidy),  
 Description = c("Year","month","Technology used for measurement", "area 15% or more covered by sea ice (10^6 km^2)","area 100% covered by sea ice (10^6 km^2)","Region observed"),  
 Type = map\_chr(sea\_ice\_tidy, .f = function(x){typeof(x)[1]}),  
 Class = map\_chr(sea\_ice\_tidy, .f = function(x){class(x)[1]}))

# Printing nicely in R Markdown   
flextable::flextable(dataDictionary\_sea\_ice, cwidth = 2)

| Variable | Description | Type | Class |
| --- | --- | --- | --- |
| year | Year | integer | integer |
| mo | month | integer | integer |
| data.type | Technology used for measurement | character | character |
| extent | area 15% or more covered by sea ice (10^6 km^2) | double | numeric |
| area | area 100% covered by sea ice (10^6 km^2) | double | numeric |
| region\_full | Region observed | character | character |

# Creating data dictionary for mean\_temp\_anom\_tidy  
dataDictionary\_mean\_temp <- tibble(Variable = colnames(mean\_temp\_anom\_tidy),  
 Description = c("Technology used for observation","Year of observation","Average global mean temperature anomalies in degrees Celsius relative to GCIS base period (1951-1980)"),  
 Type = map\_chr(mean\_temp\_anom\_tidy, .f = function(x){typeof(x)[1]}),  
 Class = map\_chr(mean\_temp\_anom\_tidy, .f = function(x){class(x)[1]}))

# Printing nicely in R Markdown   
flextable::flextable(dataDictionary\_mean\_temp, cwidth = 2)

| Variable | Description | Type | Class |
| --- | --- | --- | --- |
| Source | Technology used for observation | character | character |
| Year | Year of observation | double | numeric |
| Mean | Average global mean temperature anomalies in degrees Celsius relative to GCIS base period (1951-1980) | double | numeric |

# Creating data dictionary for CO2\_annual\_mean  
dataDictionary\_CO2\_mean <- tibble(Variable = colnames(CO2\_annual\_mean),  
 Description = c("Year-Month-Day of observation","Global mean CO2 dry air mole fraction defined as the num of molecules CO2/num all molecules in air WITHOUT WATER VAPOR. (expressed as ppm so .000400 = 400ppm).","Statistical uncertainty of measurement","Year of observation only"),  
 Type = map\_chr(CO2\_annual\_mean, .f = function(x){typeof(x)[1]}),  
 Class = map\_chr(CO2\_annual\_mean, .f = function(x){class(x)[1]}))

# Printing nicely in R Markdown   
flextable::flextable(dataDictionary\_CO2\_mean, cwidth = 2)

| Variable | Description | Type | Class |
| --- | --- | --- | --- |
| Year | Year-Month-Day of observation | double | Date |
| Mean | Global mean CO2 dry air mole fraction defined as the num of molecules CO2/num all molecules in air WITHOUT WATER VAPOR. (expressed as ppm so .000400 = 400ppm). | double | numeric |
| Uncertainty | Statistical uncertainty of measurement | double | numeric |
| YearOnly | Year of observation only | double | numeric |