Find the evidence of global warming by analyzing data collected by a weather buoy for the last 20 years

setup and install package

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(dplyr)
```

Introduction

For this project, I want to answer 1 question by analyzing the historical data from National Data Buoy Center: Is the global warming a real thing?

Installing the data

Firstly, we need to install and clean the raw data.

After checking the data from 2000-2019, 20 tables online,we found that these tables are similar but different. The tables of the later years have one more column "minute". Besides, the columns name may be different between each year's table. Also, some tables don't even have a header.

To merge these 20 tables into 1. I decide to delete all the original column names and re-title the tables consistently.

```
### make URLs

url1 <- "http://www.ndbc.noaa.gov/view_text_file.php?filename=mlrf1h"
url2 <- ".txt.gz&dir=data/historical/stdmet/"
```

```
years <- c(2000:2019)

urls <- str_c(url1, years, url2, sep = "")

filenames <- str_c("mr", years, sep = "")

###  Read the data from the website

N <- length(urls)
colname_1=colnames(read.table(urls[1],fill=TRUE, header = TRUE))
colname_2=colnames(read.table(urls[7],fill=TRUE, header = TRUE))
for (i in 1:N){
  suppressMessages(  ###  This stops the annoying messages on your screen.  Do this last.
    assign(filenames[i], read.table(urls[i],fill=TRUE ,header = TRUE))
  )

  file <- get(filenames[i])
  if(ncol(file)==17){
    colnames(file)=colname_1
  }else{
    colnames(file)=colname_2
  }

  if(i == 1){
    MR <- file
  }else{
    ###use dplyr package to bind the data
    MR <- dplyr::bind_rows(MR, file)
  }
}
```

Now MR should contain all the data detected by National Data Buoy Center for the last 20 days.

Cleaning the data

Then we want to see what is the data look like now? after using summary(MR), We found that there are still a lot of details need to be improved .

```
summary(MR)
```

```
##       YYYY            MM               DD              hh       
##  Min.   :2000   Min.   : 1.000   Min.   : 1.00   Min.   : 0.00  
##  1st Qu.:2004   1st Qu.: 3.000   1st Qu.: 8.00   1st Qu.: 6.00  
##  Median :2009   Median : 6.000   Median :16.00   Median :12.00  
##  Mean   :2009   Mean   : 6.492   Mean   :15.72   Mean   :11.57  
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:18.00  
##  Max.   :2019   Max.   :12.000   Max.   :31.00   Max.   :23.00  
##                                                                 
##        WD             WSPD             GST              WVHT         DPD    
##  Min.   :  0.0   Min.   : 0.000   Min.   : 0.000   Min.   :99   Min.   :99  
##  1st Qu.: 77.0   1st Qu.: 3.700   1st Qu.: 4.300   1st Qu.:99   1st Qu.:99  
##  Median :114.0   Median : 5.800   Median : 6.500   Median :99   Median :99  
##  Mean   :142.3   Mean   : 6.467   Mean   : 8.073   Mean   :99   Mean   :99  
##  3rd Qu.:186.0   3rd Qu.: 8.100   3rd Qu.: 9.100   3rd Qu.:99   3rd Qu.:99  
##  Max.   :999.0   Max.   :99.000   Max.   :99.000   Max.   :99   Max.   :99
##
##      APD          MWD           BAR              ATMP             WTMP
##  Min.   :99   Min.   :999   Min.   : 982.3   Min.   :  3.20   Min.   : 15.90
##  1st Qu.:99   1st Qu.:999   1st Qu.:1015.0   1st Qu.: 23.50   1st Qu.: 24.70
##  Median :99   Median :999   Median :1017.2   Median : 25.60   Median : 26.60
##  Mean   :99   Mean   :999   Mean   :1834.4   Mean   : 32.49   Mean   : 42.05
##  3rd Qu.:99   3rd Qu.:999   3rd Qu.:1020.0   3rd Qu.: 28.00   3rd Qu.: 29.10
##  Max.   :99   Max.   :999   Max.   :9999.0   Max.   :999.00   Max.   :999.00
##
##      DEWP          VIS          TIDE            mm
##  Min.   :999   Min.   :99   Min.   :99     Min.   : 0
##  1st Qu.:999   1st Qu.:99   1st Qu.:99     1st Qu.: 0
##  Median :999   Median :99   Median :99     Median : 0
##  Mean   :999   Mean   :99   Mean   :99     Mean   : 0
##  3rd Qu.:999   3rd Qu.:99   3rd Qu.:99     3rd Qu.: 0
##  Max.   :999   Max.   :99   Max.   :99     Max.   :14
##                             NA's   :3911   NA's   :42008
```

There are 4 columns give us information about date and time, we need to transform these data into posix numbers in 1 column. using lubridate package.

```r
#transform time data
MR=data.frame(MR)
time=MR%>% select(YYYY,MM,DD,hh)
time=make_datetime(MR$YYYY,time$MM,time$DD,time$hh)
time=data.frame(time)
#add it to MR
MR=dplyr::mutate(time,MR)
#remove the old date&time data
MR=MR[,-c(2,3,4,5)]
MR$time=as.POSIXct(MR$time)
```

Now we have all the information of time in one column! However, from the summary, we also notice that there are some useless variables such as “mm”,“DEWP”,“VIS”. . . .. the observations of these variables: 99;999;9999 or NA are obviously not real number. We exclude them from the data.

```r
#found that:WVHT,DPD,APD,MWD,DEWP,VIS TIDE,MM are all useless, delete them from the dataset
MR=MR[,-c(5,6,7,8,12,13,14,15)]
summary(MR)
```

```
##      time                           WD             WSPD
##  Min.   :2000-01-01 00:00:00   Min.   :  0.0   Min.   : 0.000
##  1st Qu.:2004-08-21 16:00:00   1st Qu.: 77.0   1st Qu.: 3.700
##  Median :2009-03-18 19:00:00   Median :114.0   Median : 5.800
##  Mean   :2009-04-08 10:33:00   Mean   :142.3   Mean   : 6.467
##  3rd Qu.:2013-09-24 17:00:00   3rd Qu.:186.0   3rd Qu.: 8.100
##  Max.   :2019-04-06 01:00:00   Max.   :999.0   Max.   :99.000
##      GST              BAR              ATMP             WTMP
##  Min.   : 0.000   Min.   : 982.3   Min.   :  3.20   Min.   : 15.90
##  1st Qu.: 4.300   1st Qu.:1015.0   1st Qu.: 23.50   1st Qu.: 24.70
##  Median : 6.500   Median :1017.2   Median : 25.60   Median : 26.60
##  Mean   : 8.073   Mean   :1834.4   Mean   : 32.49   Mean   : 42.05
##  3rd Qu.: 9.100   3rd Qu.:1020.0   3rd Qu.: 28.00   3rd Qu.: 29.10
##  Max.   :99.000   Max.   :9999.0   Max.   :999.00   Max.   :999.00
```

The cleaning procedure is not over yet. By summary(MR) again, we notice that there are still lot of unreal observations value for the remaining variables. Because we have huge number of observations (more than 155000), delete these observations won't cause too much damages to our dataset. So we remove observations which has values such as 99, 999.

```
#delete outliers
MR=filter(MR,MR$WD<999&MR$WSPD<99&MR$GST<99&MR$BAR<9999&MR$ATMP<999&MR$WTMP<999)
```

Sampling the data

Finally we finished cleaning the data, However, the size of dataset is still too big—we have to sampling the data to reduce the number of observations in the dataset without reducing the amount of information in the data. So I use the mean value for each day.

```
#sampling the data, reduce data size by using the mean value for each day.
MR2=MR%>%group_by(date(time))%>%summarize(mean(WD),mean(WSPD),mean(GST),mean(BAR),mean(ATMP),mean(WTMP))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

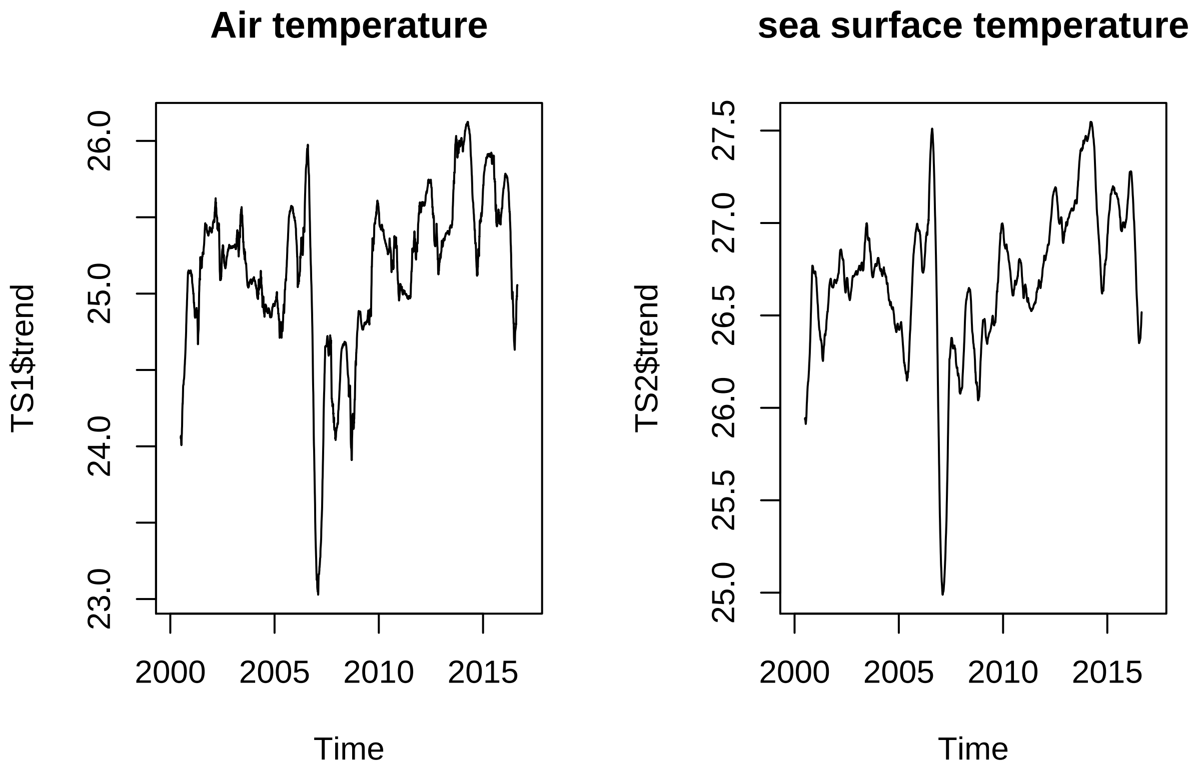
```
summary(MR2)
```

```
##    date(time)            mean(WD)        mean(WSPD)       mean(GST)
##  Min.   :2000-01-01   Min.   : 11.00   Min.   : 0.020   Min.   : 0.1333
##  1st Qu.:2004-06-27   1st Qu.: 90.88   1st Qu.: 3.917   1st Qu.: 4.4542
##  Median :2010-05-16   Median :120.58   Median : 5.747   Median : 6.4208
##  Mean   :2009-09-28   Mean   :137.60   Mean   : 5.938   Mean   : 6.6761
##  3rd Qu.:2014-08-27   3rd Qu.:175.46   3rd Qu.: 7.706   3rd Qu.: 8.5875
##  Max.   :2019-03-31   Max.   :346.28   Max.   :27.525   Max.   :32.6917
##    mean(BAR)         mean(ATMP)       mean(WTMP)
##  Min.   : 990.4   Min.   : 5.867   Min.   :19.61
##  1st Qu.:1014.9   1st Qu.:23.396   1st Qu.:24.66
##  Median :1016.9   Median :25.575   Median :26.49
##  Mean   :1016.9   Mean   :25.122   Mean   :26.67
##  3rd Qu.:1018.9   3rd Qu.:27.994   3rd Qu.:29.00
##  Max.   :1030.4   Max.   :30.350   Max.   :31.32
```

Analyzing the data

It's time to see the trend of the temperature! One thing we need to take into considered is the seasonal fluctuation, to avoid this index, we generate a simple time series plot to show the actual trend of both Air temperature and sea surface temperature

```
#time series plot to ignore seasonal fluctuation
par(mfrow=c(1,2))
ATMP=ts(MR2$`mean(ATMP)`,frequency =365,start=c(2000,1))
TS1=decompose(ATMP)
plot(TS1$trend,main="Air temperature")
WTMP=ts(MR2$`mean(WTMP)`,frequency =365,start=c(2000,1))
TS2=decompose(WTMP)
plot(TS2$trend,main="sea surface temperature")
```



As we can see from the plot, the Air temperature and sea surface temperature has a very similar trend during the last 20 years.(almost the same) Apart from a sharp decrease in 2007(I guess the buoy broke down that year or something else happened), there is an overall upward trend for temperature and it reached to a peak at about 2014.

Further to prove the global warming is real, I biuld a linear regression with time and temperature.

```
fit1=lm(TS1$trend~MR2$`date(time)`)
fit2=lm(TS2$trend~MR2$`date(time)`)
coef(fit1)[2]
```

```
## MR2$`date(time)`
##     8.496768e-05
```

```
coef(fit2)[2]
```

```
## MR2$`date(time)`
##      7.70798e-05
```

The coefficient for date correspound to temperature are positive number(although it's small, it make sense because the temperature won't change fast and the tempertur in 2007 is too low)

Summarize

From my analyze above, I can answer the question mentioned before: yes, Global warming is real,the temperature is not rising fast, but we need to pay attention to it.

reference

R packages: tidyverse; stringr; lubridate; dplyr cheatshit: https://rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf https://evoldyn.gitlab.io/evomics-2018/ref-sheets/R_lubridate.pdf data resourses: https://www.ndbc.noaa.gov/