Report of Buoy Project

Zhe Yu

2020/9/25

Introduction

In this buoy project we aim to find whether there is an evidence of global warming, so we intended to use the data in National Data Buoy Center from 1987 to 2016 to see if there are upward trends with air temperature and water temperature as time goes by. We have three steps: * Import and clean data * Transform data and make analysis * Draw a conclusion

Import and clean data

At first, we have to import the data from the website. Because there are some differences in some years so I separated the 20 years into three parts, making them all have 16 variables. And then combine all the data into one data frame.

```r
#get each data frame from 1987 to 2016
url1 <- "http://www.ndbc.noaa.gov/view_text_file.php?filename=mlrf1h"
url2 <- ".txt.gz&dir=data/historical/stdmet/"
years <- c(1987:2016)
urls <- str_c(url1, years, url2, sep ="")
filenames <- str_c("mr", years, sep = "")
N <- length(urls)

for (i in 1:N){
  suppressMessages(
    assign(filenames[i], read.table(urls[i], header = TRUE,fill = T))
  )
  file <- get(filenames[i])
}

# add 19 in front of year
for (i in 1:12){
  file <- get(filenames[i])
  file$YY <- file$YY + 1900
  assign(filenames[i],file)
}

#throw out the last column
for (i in 14:18){
  file <- get(filenames[i])
  assign(filenames[i],file[,1:16])
}

#throw out the "mm" column and the last column
for (i in 19:30){
```

```
  file <- get(filenames[i])
  assign(filenames[i],file[,c(1:4,6:17)])
}

#combine all the data frame and unify the column name
for (i in 1:30){
  file <- get(filenames[i])
  colnames(file) <- c("YYYY", "MM", "DD", "hh", "WD", "WSPD", "GST", "WVHT", "DPD", "APD", "MWD", "BAR"

  if(i==1){
    MR <- file
  }
  else{
    MR <- rbind.data.frame(MR, file)
  }
}

#select columns that are not 99, 999
MR <- MR[,c(1:7,12:14)]

#filter rows that ATMP and WTMP are larger than 100
MR <- filter(MR,MR$ATMP<100&MR$WTMP<100)
```

Then I got the data frame MR.csv

```
##       YYYY            MM               DD              hh
##  Min.   :1987   Min.   : 1.000   Min.   : 1.00   Min.   : 0.00
##  1st Qu.:1994   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 6.00
##  Median :2002   Median : 7.000   Median :16.00   Median :12.00
##  Mean   :2002   Mean   : 6.536   Mean   :15.73   Mean   :11.52
##  3rd Qu.:2009   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:18.00
##  Max.   :2016   Max.   :12.000   Max.   :31.00   Max.   :23.00
##        WD             WSPD              GST              BAR
##  Min.   :  0.0   Min.   : 0.000   Min.   : 0.000   Min.   : 982.3
##  1st Qu.: 81.0   1st Qu.: 3.800   1st Qu.: 4.300   1st Qu.:1014.8
##  Median :118.0   Median : 5.800   Median : 6.500   Median :1017.1
##  Mean   :144.9   Mean   : 6.496   Mean   : 7.285   Mean   :1485.1
##  3rd Qu.:186.0   3rd Qu.: 8.000   3rd Qu.: 9.000   3rd Qu.:1019.4
##  Max.   :999.0   Max.   :99.000   Max.   :99.000   Max.   :9999.0
##       ATMP            WTMP
##  Min.   : 1.80   Min.   :15.90
##  1st Qu.:23.50   1st Qu.:24.70
##  Median :25.60   Median :26.60
##  Mean   :25.16   Mean   :26.68
##  3rd Qu.:28.00   3rd Qu.:29.00
##  Max.   :32.20   Max.   :32.10
```

For in this data frame, the date information is separated into different columns, so I use the lubricate package to transfer these data into posix type.

```
#combine all the time into one column using lubridate package
MR_DATE <- MR %>%
  mutate(DATETIME = make_datetime(YYYY,MM,DD,hh))

#change the order of each column, put the time into first place.
```

```
MR_DATE <- MR_DATE[,5:11]
cols <- colnames(MR_DATE)
new_cols <- c(cols[7],cols[1:6])
MR_DATE <- MR_DATE[,new_cols]
```

Then I got the MR_DATE.csv

```
##     DATETIME                            WD             WSPD
##  Min.   :1987-12-04 12:00:00   Min.   :  0.0   Min.   : 0.000
##  1st Qu.:1994-12-15 18:45:00   1st Qu.: 81.0   1st Qu.: 3.800
##  Median :2002-04-28 02:30:00   Median :118.0   Median : 5.800
##  Mean   :2002-03-29 20:44:55   Mean   :144.9   Mean   : 6.496
##  3rd Qu.:2009-06-24 17:15:00   3rd Qu.:186.0   3rd Qu.: 8.000
##  Max.   :2016-12-31 23:00:00   Max.   :999.0   Max.   :99.000
##       GST              BAR              ATMP            WTMP
##  Min.   : 0.000   Min.   : 982.3   Min.   : 1.80   Min.   :15.90
##  1st Qu.: 4.300   1st Qu.:1014.8   1st Qu.:23.50   1st Qu.:24.70
##  Median : 6.500   Median :1017.1   Median :25.60   Median :26.60
##  Mean   : 7.285   Mean   :1485.1   Mean   :25.16   Mean   :26.68
##  3rd Qu.: 9.000   3rd Qu.:1019.4   3rd Qu.:28.00   3rd Qu.:29.00
##  Max.   :99.000   Max.   :9999.0   Max.   :32.20   Max.   :32.10
```

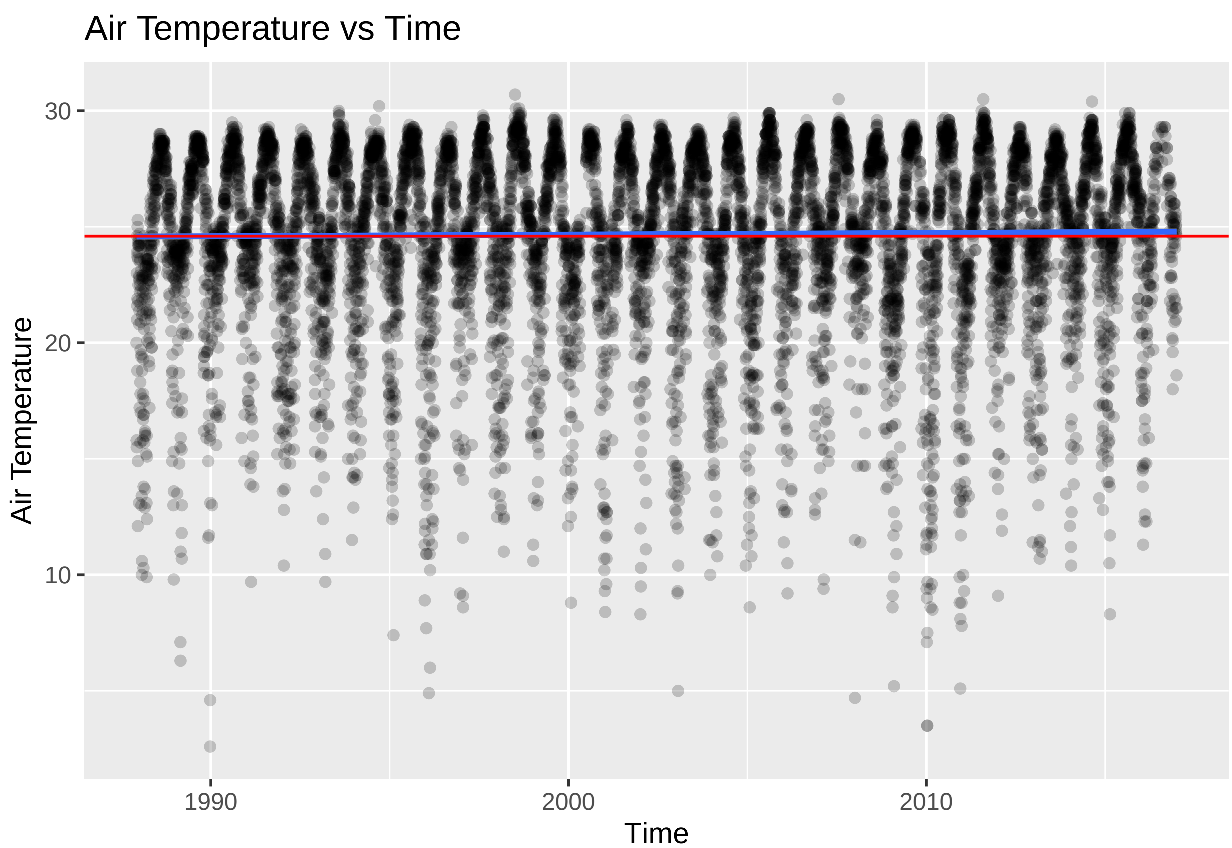
Transform data and make analysis

For there are too many data to draw a plot so I will reduce the amount of data by selecting the data at 12:00:00 each day and become a new data frame.

Next I draw two plot:(a)Air Temperature vs Time. (b)Water Temperature vs Time. And fit a linear regression model. As can be seen that the slope is really small so I add horizontal lines to make it more visible.

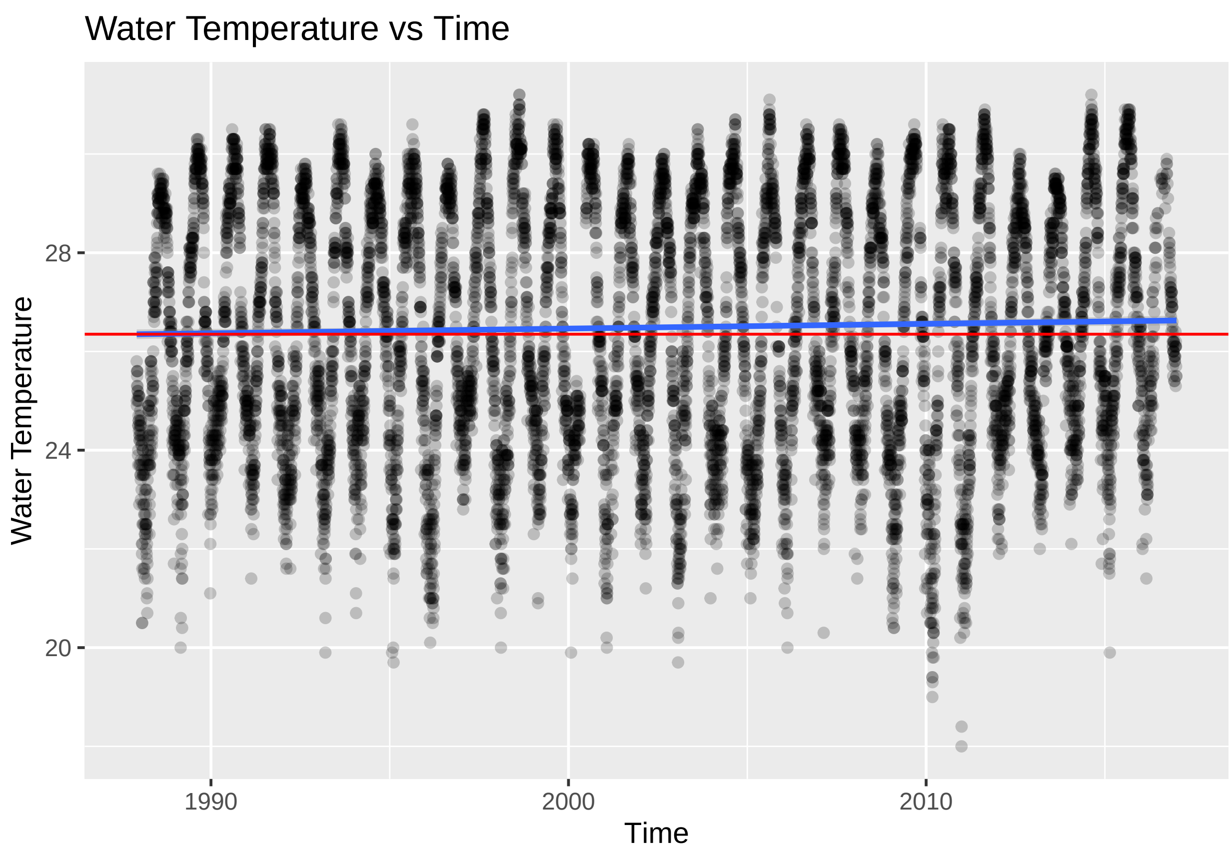
```
ggplot()+
  geom_point(aes(MR_DAY$DATETIME,MR_DAY$ATMP),alpha=0.2)+
  geom_smooth(method=lm,aes(MR_DAY$DATETIME,MR_DAY$ATMP))+
  geom_abline(intercept = 24.6,slope = 0,color="red")+
  labs(title="Air Temperature vs Time",x="Time",y="Air Temperature")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(data=)+
  geom_point(aes(MR_DAY$DATETIME,MR_DAY$WTMP),alpha=0.2)+
  geom_smooth(method=lm,aes(MR_DAY$DATETIME,MR_DAY$WTMP))+
  geom_abline(intercept = 26.35,slope = 0,color="red")+
  labs(title="Water Temperature vs Time",x="Time",y="Water Temperature")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



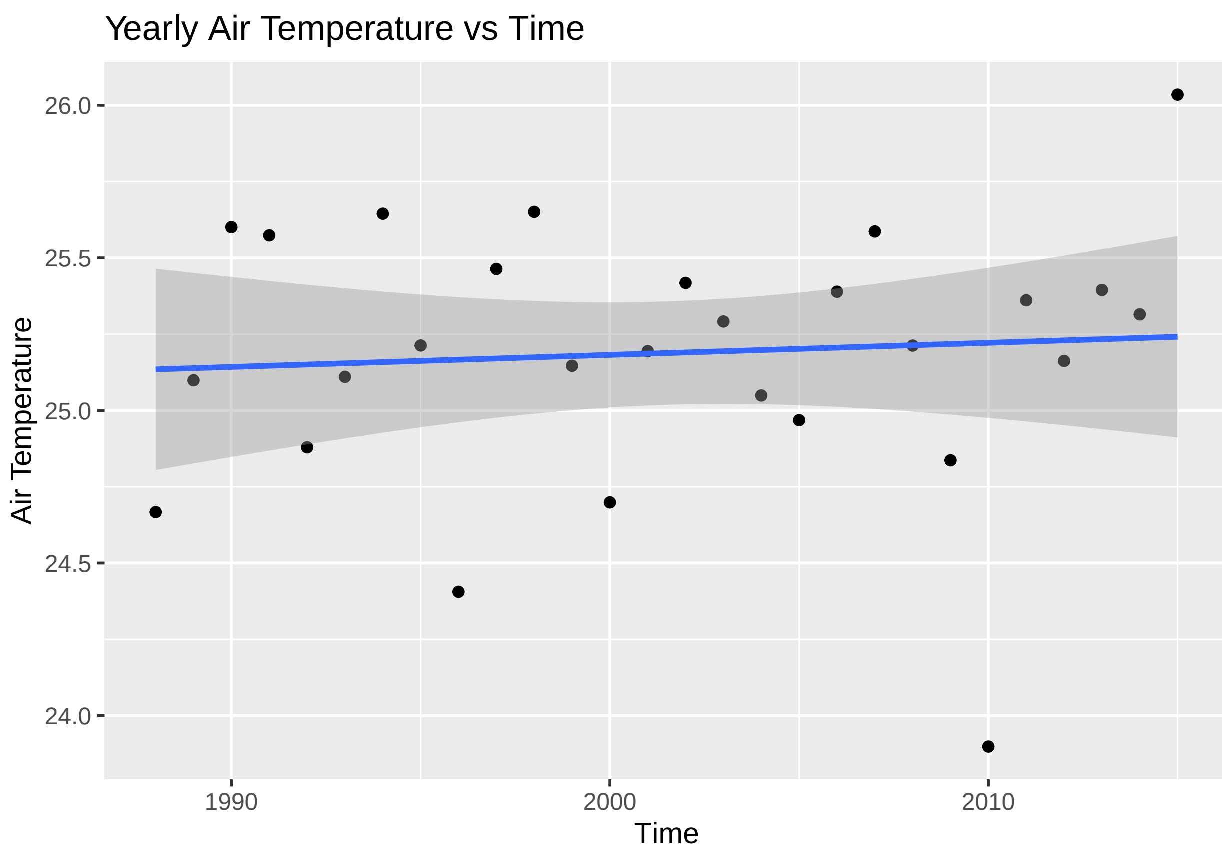
But the density of dots in those two plots are too high so it is a little bit dirty. Thus I choose using the mean value to draw another two plots.

```
mean_atemp <- rep(NA,28)
for (i in 1:28){
  file2 <- filter(MR, MR$YYYY==1987+i)
  mean_atemp[i] <- mean(file2$ATMP)
}

mean_wtemp <- rep(NA,28)
for (i in 1:28){
  file2 <- filter(MR, MR$YYYY==1987+i)
  mean_wtemp[i] <- mean(file2$WTMP)
}
```

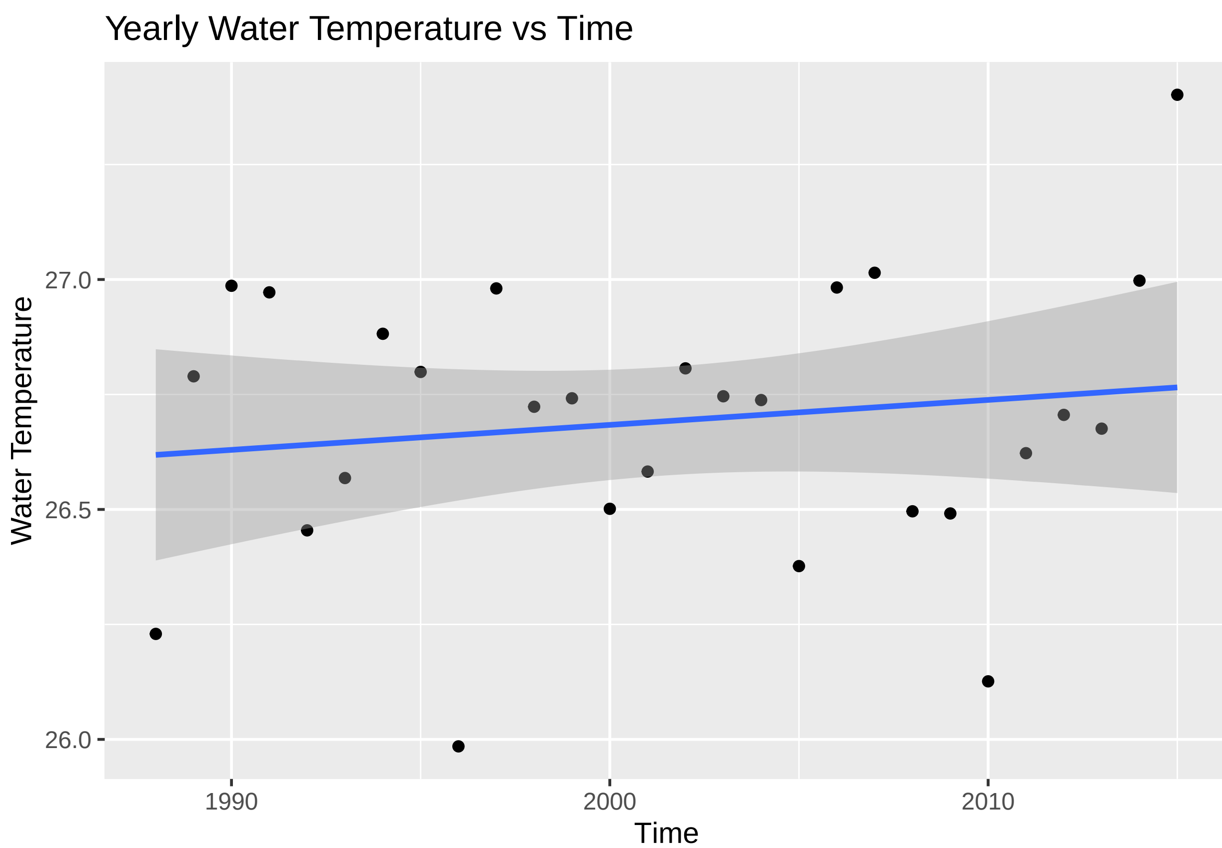
```
ggplot(mapping=aes(1988:2015,mean_atemp))+
  geom_point()+
  geom_smooth(method=lm)+
  labs(title="Yearly Air Temperature vs Time",x="Time",y="Air Temperature")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```r
ggplot(mapping=aes(1988:2015,mean_wtemp))+
  geom_point()+
  geom_smooth(method=lm)+
  labs(title="Yearly Water Temperature vs Time",x="Time",y="Water Temperature")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



It is much clear than the plots before!

Conclusion

Using the linear regression model gets the coefficients.

```
model_Atemp <- data.frame(1988:2015,mean_atemp)
colnames(model_Atemp) <- c("year","mean_atemp")
fit_Atemp <- stan_glm(mean_atemp~year,data=model_Atemp,refresh=0)
coef(fit_Atemp)
```

```
##  (Intercept)         year
## 16.875672273  0.004142911
```

```
model_Wtemp <- data.frame(1988:2015,mean_wtemp)
colnames(model_Wtemp) <- c("year","mean_wtemp")
fit_Wtemp <- stan_glm(mean_wtemp~year,data=model_Wtemp,refresh=0)
coef(fit_Wtemp)
```

```
##  (Intercept)         year
## 15.834357486  0.005428556
```

It is clear that by each year, both air temperature and water temperature increase no more than 0.01 celsius degree, although it is a small number but also an evidence of global warming.