Boston-Buoy-Data-Analysis

RickyZhao

9/24/2020

Download Data

Download Historical data from NOAA National Data Buoy Center. Read data ofNDBC Station 44013, years from 1987 to 2019.

Because these txt files have different column names, the initial data is separated into 5 data frames. Take 1987 - 1998 as example. Their original column names are:

```
colnames(merge.data1)
```

```
##  [1] "YY"   "MM"   "DD"   "hh"   "WD"   "WSPD" "GST"  "WVHT" "DPD"  "APD"
## [11] "MWD"  "BAR"  "ATMP" "WTMP" "DEWP" "VIS"
```

Combine its date variables into DATE, and time variables into TIME, Add an new variable as MONTH for future group.

```
colnames(merge.data1)
```

```
##  [1] "DATE"  "MONTH" "TIME"  "WD"    "WSPD"  "GST"   "WVHT"  "DPD"   "APD"
## [10] "MWD"   "BAR"   "ATMP"  "WTMP"  "DEWP"  "VIS"
```

Do the same with the next 4 data frames.

Combine data frames.

After change the date and time variables, we can combine 5 data frames into 3. Print the column names to find the difference.

```
##  [1] "DATE"  "MONTH" "TIME"  "WD"    "WSPD"  "GST"   "WVHT"  "DPD"   "APD"
## [10] "MWD"   "BAR"   "ATMP"  "WTMP"  "DEWP"  "VIS"
```

```
##  [1] "DATE"  "MONTH" "TIME"  "WD"    "WSPD"  "GST"   "WVHT"  "DPD"   "APD"
## [10] "MWD"   "BAR"   "ATMP"  "WTMP"  "DEWP"  "VIS"   "TIDE"
```

```
##  [1] "DATE"  "MONTH" "TIME"  "WDIR"  "WSPD"  "GST"   "WVHT"  "DPD"   "APD"
## [10] "MWD"   "PRES"  "ATMP"  "WTMP"  "DEWP"  "VIS"   "TIDE"
```

From the definition on NOAA National Data Buoy Center Website, we find the “BAR” is the historic name of “PRES”. Change the column name for data before 2007.

```
names(data_87_99)[names(data_87_99)=="BAR"]="PRES"
names(data_00_06)[names(data_00_06)=="BAR"]="PRES"
```

Add NA for TIDE, WDIR and WD to the data frame as a missing value to make their names as each other. Then combine the data frames into one.

```
data_87_19 <- rbind(data_87_99,data_00_06,data_07_19)
```

Convert the data from character into numeric.

```
str(data_87_19)
```

```
## 'data.frame':    276411 obs. of  17 variables:
##  $ DATE : Date, format: "1987-01-01" "1987-01-01" ...
##  $ MONTH: Date, format: "1987-01-01" "1987-01-01" ...
##  $ TIME : num  0 3600 7200 10800 14400 18000 21600 25200 28800 32400 ...
##  $ WD   : num  290 290 290 300 290 340 10 10 20 20 ...
##  $ WDIR : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ WSPD : num  8 7 6 6 5 6 5 4 7 5 ...
##  $ GST  : num  10 8 8 7 6 7 6 6 8 6 ...
##  $ WVHT : num  2.7 2.4 2.5 2.6 2.7 2.4 2.4 2.6 2.5 2.5 ...
##  $ DPD  : num  11.1 10 11.1 11.1 12.5 14.3 12.5 12.5 12.5 12.5 ...
##  $ APD  : num  8.6 8 8.3 8.6 8.7 8.4 8.8 9.5 9.2 9.1 ...
##  $ MWD  : num  999 999 999 999 999 999 999 999 999 999 ...
##  $ PRES : num  1024 1024 1024 1024 1025 ...
##  $ ATMP : num  2.8 2 1.6 1.3 1 0.7 0.7 0.5 0.4 0.4 ...
##  $ WTMP : num  5.9 5.9 5.9 6 5.9 5.9 5.9 5.9 5.9 5.9 ...
##  $ DEWP : num  999 999 999 999 999 999 999 999 999 999 ...
##  $ VIS  : num  99 99 99 99 99 99 99 99 99 99 ...
##  $ TIDE : chr  NA NA NA NA ...
```

There are some data like “9999”, “999”, “99”, which seems impossible value. They are actually missing value. Replace them with NA.

```
summary(data_87_19,na.rm = TRUE)
```

```
##       DATE                MONTH                 TIME             WD
##  Min.   :1987-01-01   Min.   :1987-01-01   Min.   :    0   Min.   :  0.0
##  1st Qu.:1995-03-23   1st Qu.:1995-04-01   1st Qu.:21000   1st Qu.:136.0
##  Median :2003-06-28   Median :2003-07-01   Median :42600   Median :224.0
##  Mean   :2003-07-08   Mean   :2003-07-08   Mean   :42451   Mean   :270.4
##  3rd Qu.:2011-07-04   3rd Qu.:2011-07-01   3rd Qu.:64200   3rd Qu.:299.0
##  Max.   :2019-12-31   Max.   :2020-01-01   Max.   :85800   Max.   :999.0
##                                                            NA’s   :107545
##       WDIR             WSPD             GST             WVHT
##  Min.   :  1.0    Min.   : 0.000   Min.   : 0.00   Min.   : 0.000
##  1st Qu.:125.0    1st Qu.: 3.700   1st Qu.: 4.40   1st Qu.: 0.400
##  Median :204.0    Median : 5.700   Median : 6.80   Median : 0.660
##  Mean   :197.7    Mean   : 9.788   Mean   :11.14   Mean   : 2.095
##  3rd Qu.:281.0    3rd Qu.: 8.600   3rd Qu.:10.40   3rd Qu.: 1.090
##  Max.   :999.0    Max.   :99.000   Max.   :99.00   Max.   :99.000
##  NA’s   :168866
##       DPD              APD              MWD             PRES
##  Min.   : 0.000   Min.   : 0.000   Min.   :  0.0   Min.   : 964.6
##  1st Qu.: 4.550   1st Qu.: 3.900   1st Qu.:122.0   1st Qu.:1010.2
##  Median : 7.690   Median : 4.800   Median :999.0   Median :1015.7
##  Mean   : 9.473   Mean   : 6.186   Mean   :700.9   Mean   :1051.8
##  3rd Qu.:10.000   3rd Qu.: 5.960   3rd Qu.:999.0   3rd Qu.:1021.2
##  Max.   :99.000   Max.   :99.000   Max.   :999.0   Max.   :9999.0
##
##       ATMP             WTMP             DEWP            VIS
##  Min.   :-19.70   Min.   : -1.80   Min.   :-24.9   Min.   : 0.00
##  1st Qu.:  3.60   1st Qu.:  5.30   1st Qu.:  6.7   1st Qu.:99.00
##  Median :  9.70   Median : 10.30   Median : 20.6   Median :99.00
##  Mean   : 13.87   Mean   : 54.43   Mean   :491.7   Mean   :92.98
##  3rd Qu.: 16.70   3rd Qu.: 16.50   3rd Qu.:999.0   3rd Qu.:99.00
##  Max.   :999.00   Max.   :999.00   Max.   :999.0   Max.   :99.00
##
##      TIDE
##  Length:276411
##  Class :character
##  Mode  :character
##
##
##
##
```

```
summary(data_87_19,na.rm = TRUE)
```

```
##       DATE                MONTH                 TIME             WD
##  Min.   :1987-01-01   Min.   :1987-01-01   Min.   :    0   Min.   :  0.0
##  1st Qu.:1995-03-23   1st Qu.:1995-04-01   1st Qu.:21000   1st Qu.:127.0
##  Median :2003-06-28   Median :2003-07-01   Median :42600   Median :211.0
##  Mean   :2003-07-08   Mean   :2003-07-08   Mean   :42451   Mean   :197.8
##  3rd Qu.:2011-07-04   3rd Qu.:2011-07-01   3rd Qu.:64200   3rd Qu.:280.0
##  Max.   :2019-12-31   Max.   :2020-01-01   Max.   :85800   Max.   :360.0
##                                                            NA's   :122835
##       WDIR             WSPD             GST             WVHT
##  Min.   :  1.0    Min.   : 0.000   Min.   : 0.0    Min.   :0.000
##  1st Qu.:124.0    1st Qu.: 3.600   1st Qu.: 4.3    1st Qu.:0.400
##  Median :203.0    Median : 5.500   Median : 6.6    Median :0.650
##  Mean   :195.8    Mean   : 6.079   Mean   : 7.4    Mean   :0.864
##  3rd Qu.:281.0    3rd Qu.: 8.100   3rd Qu.: 9.8    3rd Qu.:1.060
##  Max.   :360.0    Max.   :25.700   Max.   :32.4    Max.   :9.100
##  NA's   :169114   NA's   :11033    NA's   :11295   NA's   :3467
##       DPD              APD              MWD              PRES
##  Min.   : 0.000   Min.   : 0.000   Min.   :  0.0    Min.   : 964.6
##  1st Qu.: 4.550   1st Qu.: 3.900   1st Qu.: 78.0    1st Qu.:1010.2
##  Median : 7.690   Median : 4.780   Median : 94.0    Median :1015.7
##  Mean   : 7.378   Mean   : 5.007   Mean   :124.3    Mean   :1015.5
##  3rd Qu.:10.000   3rd Qu.: 5.900   3rd Qu.:129.0    3rd Qu.:1021.1
##  Max.   :25.000   Max.   :12.100   Max.   :360.0    Max.   :1045.8
##  NA's   :6319     NA's   :3467     NA's   :182225   NA's   :1117
##       ATMP              WTMP            DEWP             VIS
##  Min.   :-19.700   Min.   :-1.80   Min.   :-24.90   Min.   : 0.00
##  1st Qu.:  3.600   1st Qu.: 5.10   1st Qu.: -0.50   1st Qu.: 8.10
##  Median :  9.700   Median : 9.80   Median :  7.00   Median : 9.40
##  Mean   :  9.671   Mean   :10.49   Mean   :  6.28   Mean   :12.48
##  3rd Qu.: 16.700   3rd Qu.:15.70   3rd Qu.: 14.50   3rd Qu.:11.60
##  Max.   : 32.100   Max.   :27.80   Max.   : 26.10   Max.   :36.00
##  NA's   :1172      NA's   :12288   NA's   :135148   NA's   :257172
##      TIDE
##  Length:276411
##  Class :character
##  Mode  :character
##
##
##
##
```

Save data as BuoyData.Rdata

```
save(data_87_19,file='BuoyData.Rdata')
```

Data Select

According to the summary, there are over 10,000 NAs in variables “WD”, “WDIR”, “MWD”, “VIS”, “TIDE” and “DEWP”, so we select other variables from the data frame.

Remove rows with NA, group the data by day(month), calculate the mean value of variables by day(month).

```
tmpdata  <-  na.omit(tmpdata)
planes3 <-  group_by(tmpdata, DATE)
delay3  <-  summarise(planes3, WSPD = mean(WSPD, na.rm = TRUE),
                      GST = mean(GST, na.rm = TRUE),
                      WVHT = mean(WVHT, na.rm = TRUE),
                      DPD = mean(DPD, na.rm = TRUE),
                      APD = mean(APD, na.rm = TRUE),
                      PRES = mean(PRES, na.rm = TRUE),
                      ATMP = mean(ATMP, na.rm = TRUE),
                      WTMP = mean(WTMP, na.rm = TRUE))
```

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

Delay3 is the data frame grouped by day.

```
head(delay3,n=3)
```

```
## # A tibble: 3 x 9
##   DATE        WSPD   GST  WVHT   DPD   APD  PRES  ATMP  WTMP
##   <date>     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 1987-01-01   4    5.25  2.08 11.9   8.44 1025.  1.7   5.89
## 2 1987-01-02  11.7 14.6   2.76  9.46  6.87 1005.  3.06  5.84
## 3 1987-01-03  10.0 12.3   1.92 10.4   6.08  999. -1.12  5.72
```

```
planes4 <-  group_by(tmpdata, MONTH)
delay4  <-  summarise(planes4, WSPD = mean(WSPD, na.rm = TRUE),
                      GST = mean(GST, na.rm = TRUE),
                      WVHT = mean(WVHT, na.rm = TRUE),
                      DPD = mean(DPD, na.rm = TRUE),
                      APD = mean(APD, na.rm = TRUE),
                      PRES = mean(PRES, na.rm = TRUE),
                      ATMP = mean(ATMP, na.rm = TRUE),
                      WTMP = mean(WTMP, na.rm = TRUE))
```

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

Delay4 is the data frame grouped by month.

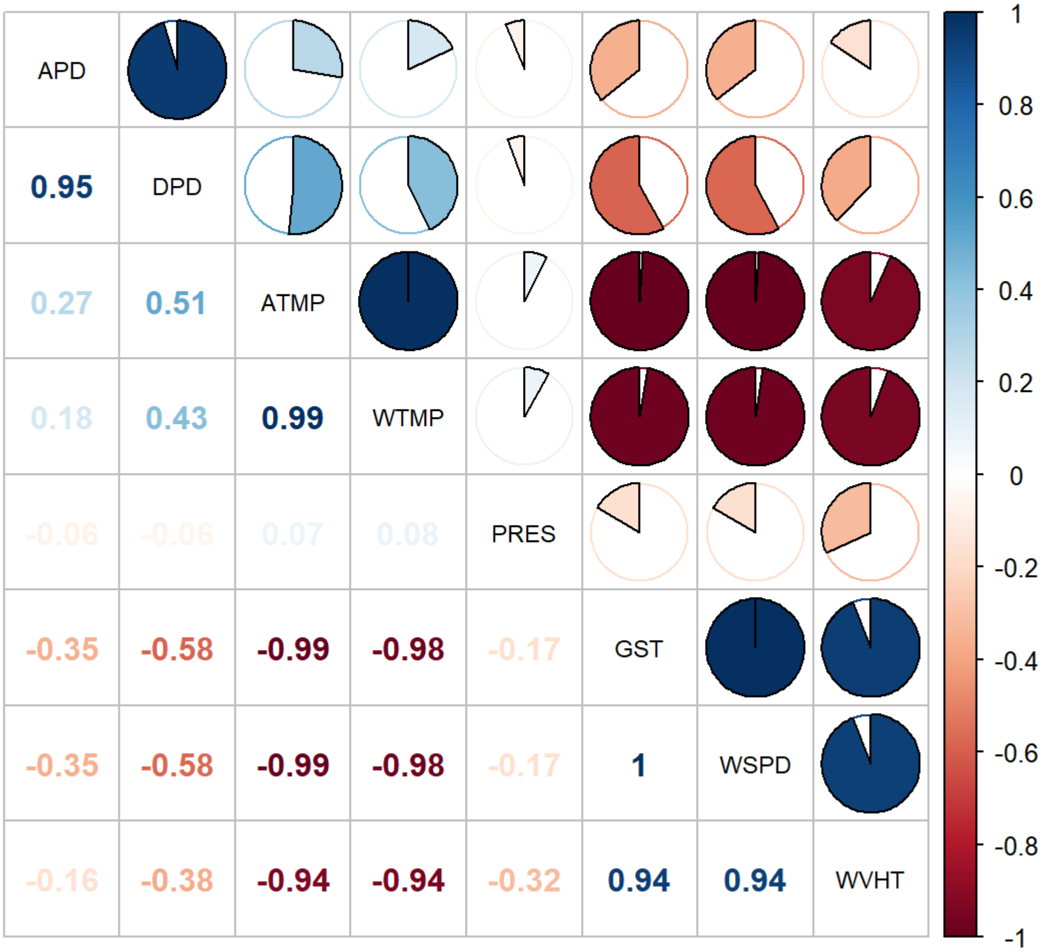
```
head(delay4,n=3)
```

```
## # A tibble: 3 x 9
##   MONTH       WSPD   GST  WVHT   DPD   APD  PRES   ATMP  WTMP
##   <date>     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1 1987-01-01  7.48  9.24 0.890  6.07  4.50 1010.  1.75   5.33
## 2 1987-02-01  7.41  9.31 1.11   7.71  5.36 1011. -2.23   3.69
## 3 1987-03-01  6.19  7.67 0.907  6.68  4.94 1019. -0.480  2.40
```

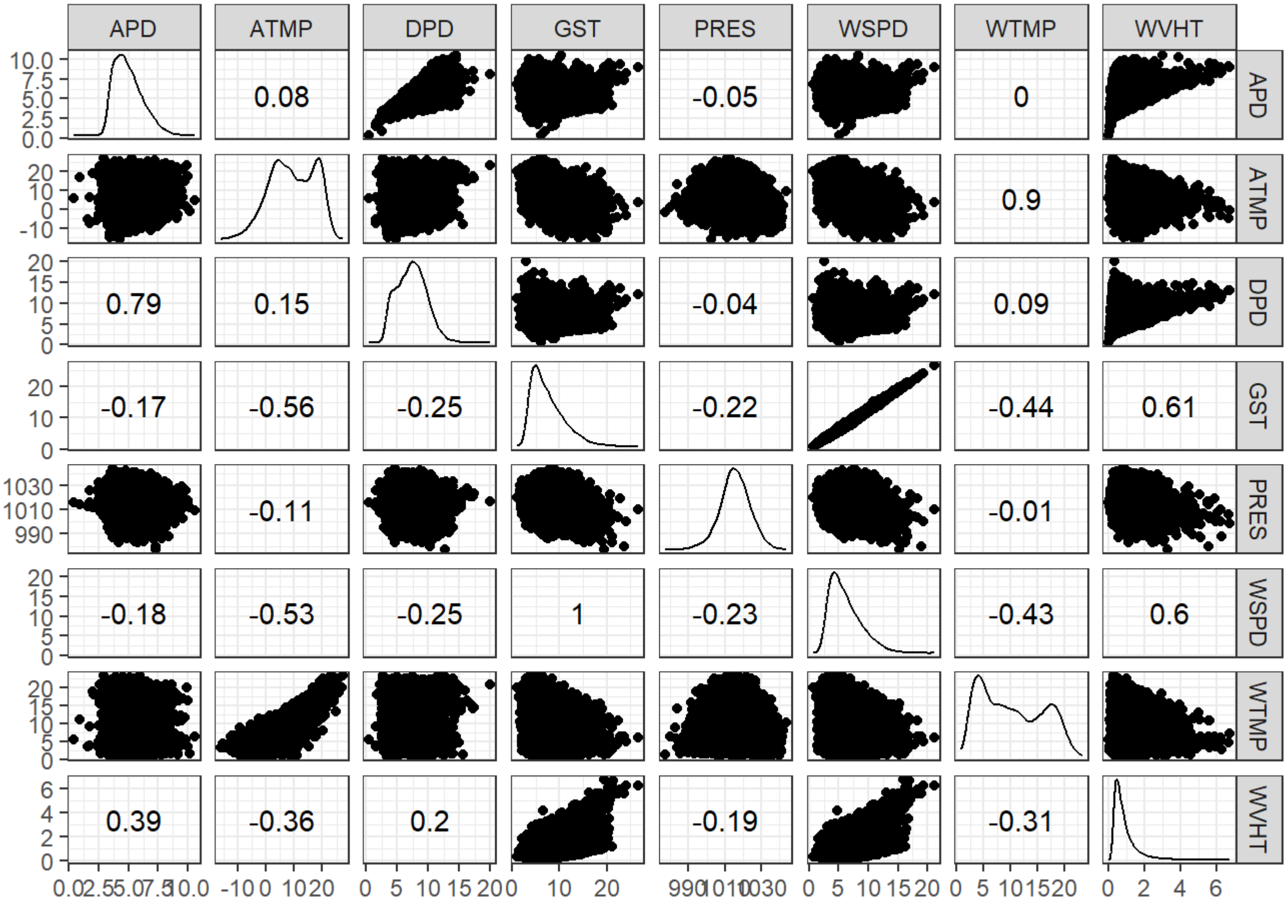
Correlation between different variables

We make plot correlation between these variables.

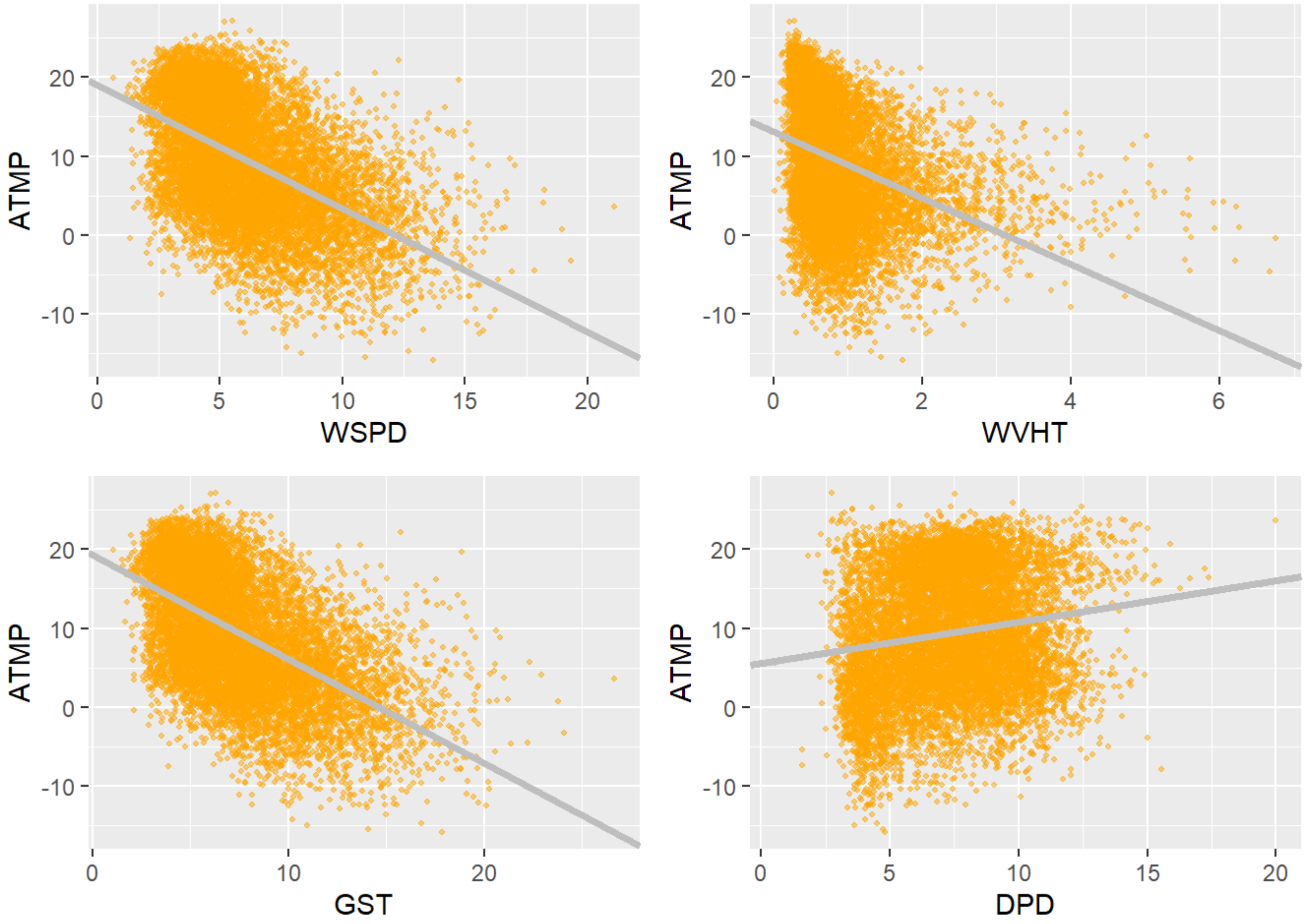
```
corrplot(corr = res_cor,order = ″AOE″,type=″upper″,method=″pie″,tl.pos = ″d″,tl.cex = 0.75,tl.col
 = ″black″)
corrplot(corr = res_cor,add=TRUE, type=″lower″, method=″number″,order=″AOE″,diag=FALSE,tl.pos=″n″,
cl.pos=″n″)
```



Plot variables in matrix, we can find ATMP have high correlation with WSPD, GST, WVHT and DPD.



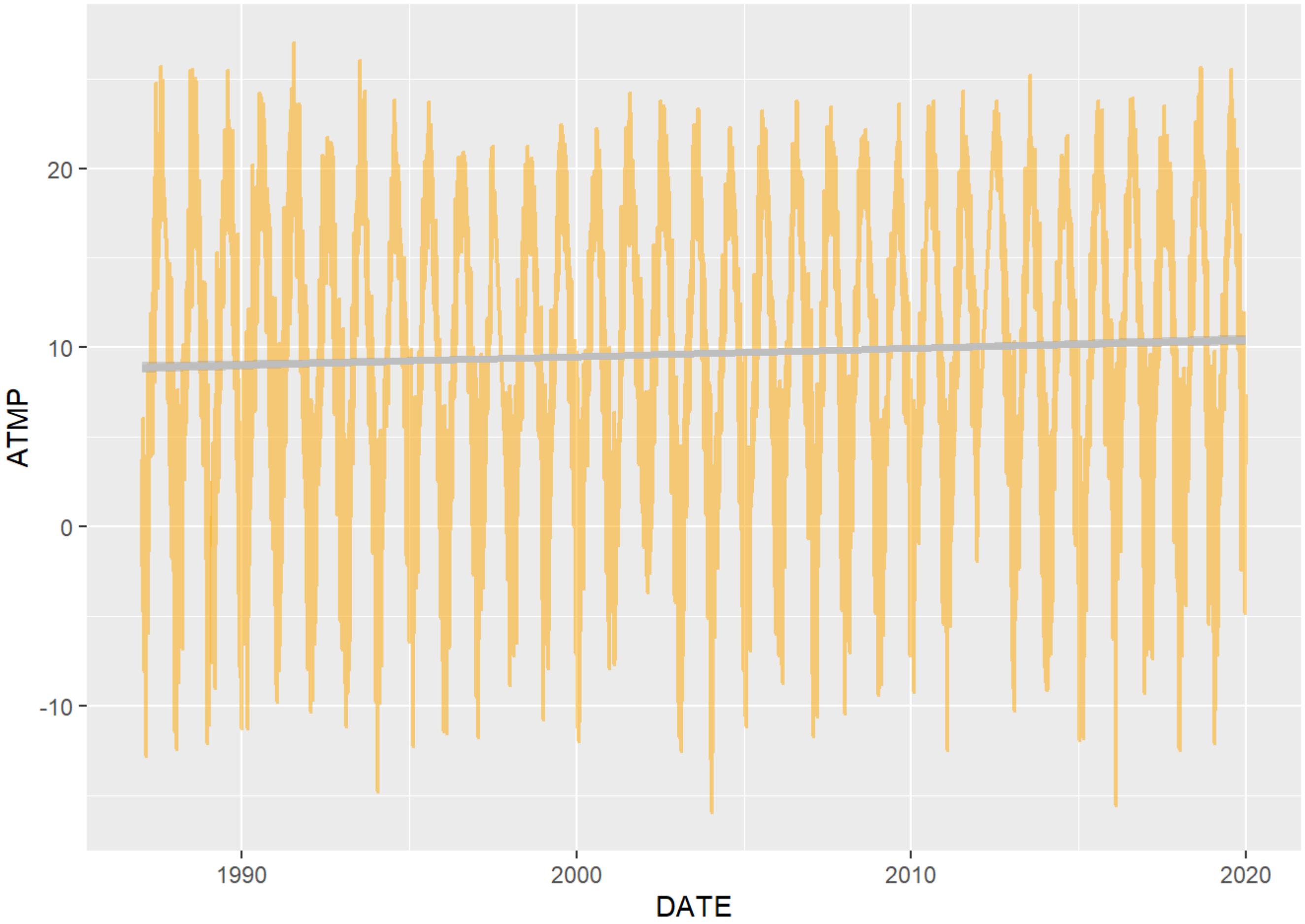
Plot the relationship between ATMP and WSPD, GST, WVHT and DPD with ggplot.



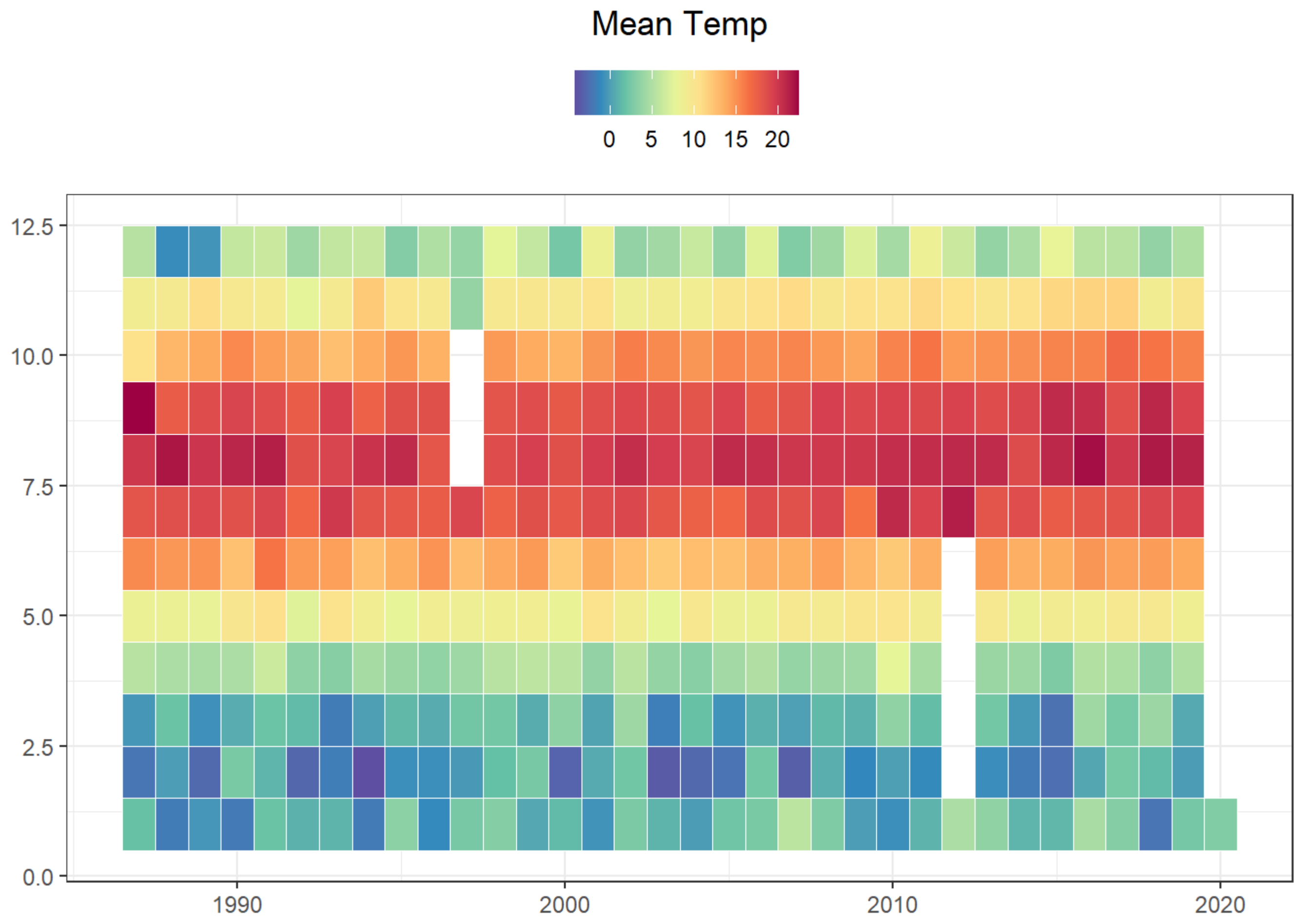
Time and Temperature

Plot the Date and Temperature, add smooth line to show the trends.

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```



Plot a heatmap to show the trends of temperature with DATE.



Do regression of Temperature of DATE:

```
fit_date<-lm(data=delay3, ATMP~DATE)
summary(fit_date)
```

```
##
## Call:
## lm(formula = ATMP ~ DATE, data = delay3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.942  -5.951  -0.201   7.068  18.721
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.389e+00  2.883e-01   22.16   <2e-16 ***
## DATE        2.402e-04  2.219e-05   10.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.829 on 10657 degrees of freedom
## Multiple R-squared:  0.01087,    Adjusted R-squared:  0.01078
## F-statistic: 117.2 on 1 and 10657 DF,  p-value: < 2.2e-16
```

```
coef(fit_date)
```

```
##  (Intercept)         DATE 
## 6.3887626487 0.0002401496
```

We can see the slop is small, but the p-value shows this coefficient is significantly different from zero!

Conclusion

The temperature is raising by time!

It can be seen from the figure and the slop coefficient that although the daily temperature increase is tiny, after 30 years of continuous accumulation, the global average temperature has risen by over 2 degrees, which means that the global temperature will rise by nearly 7 degrees within a century. This is a significant increment.