Lab 03 - Exploratory Data Analysis

Learning Goals

- Read in and get familiar with the meteorology dataset
- Step through the EDA "checklist" presented in the class slides
- Practice making exploratory graphs

Lab Description

We will work with the meteorological data presented in lecture. Recall the dataset consists of weather station readings in the continental US.

The objective of the lab is to find the weather station with the highest elevation and look at patterns in the time series of its wind speed and temperature.

Steps

1. Read in the data

First download and then read in with data.table:fread()

```
download.file(
  "https://raw.githubusercontent.com/JSC370/jsc370-2023/main/labs/lab03/met_all.gz",
  destfile = "met_all.gz",
  method   = "curl",
  timeout  = 60
  )
met <- data.table::fread("met_all.gz")
```

2. Check the dimensions, headers, footers. How many columns, rows are there?

```
dim(met)
```

```
## [1] 2377343      30
```

```
head(met)
```

```
##    USAFID  WBAN year month day hour min  lat      lon elev wind.dir wind.dir.qc
## 1: 690150 93121 2019     8   1    0  56 34.3 -116.166  696      220           5
## 2: 690150 93121 2019     8   1    1  56 34.3 -116.166  696      230           5
## 3: 690150 93121 2019     8   1    2  56 34.3 -116.166  696      230           5
## 4: 690150 93121 2019     8   1    3  56 34.3 -116.166  696      210           5
## 5: 690150 93121 2019     8   1    4  56 34.3 -116.166  696      120           5
## 6: 690150 93121 2019     8   1    5  56 34.3 -116.166  696       NA           9
##    wind.type.code wind.sp wind.sp.qc ceiling.ht ceiling.ht.qc ceiling.ht.method
## 1:              N     5.7          5      22000             5                 9
## 2:              N     8.2          5      22000             5                 9
## 3:              N     6.7          5      22000             5                 9
## 4:              N     5.1          5      22000             5                 9
## 5:              N     2.1          5      22000             5                 9
## 6:              C     0.0          5      22000             5                 9
##    sky.cond vis.dist vis.dist.qc vis.var vis.var.qc temp temp.qc dew.point
## 1:        N    16093           5       N          5 37.2       5      10.6
## 2:        N    16093           5       N          5 35.6       5      10.6
## 3:        N    16093           5       N          5 34.4       5       7.2
## 4:        N    16093           5       N          5 33.3       5       5.0
## 5:        N    16093           5       N          5 32.8       5       5.0
## 6:        N    16093           5       N          5 31.1       5       5.6
##    dew.point.qc atm.press atm.press.qc       rh
## 1:            5    1009.9           5 19.88127
## 2:            5    1010.3           5 21.76098
## 3:            5    1010.6           5 18.48212
## 4:            5    1011.6           5 16.88862
## 5:            5    1012.7           5 17.38410
## 6:            5    1012.7           5 20.01540
```

```
tail(met)
```

```
##    USAFID  WBAN year month day hour min    lat      lon elev wind.dir
## 1: 726813 94195 2019     8  31   18  56 43.650 -116.633  741       NA
## 2: 726813 94195 2019     8  31   19  56 43.650 -116.633  741       70
## 3: 726813 94195 2019     8  31   20  56 43.650 -116.633  741       NA
## 4: 726813 94195 2019     8  31   21  56 43.650 -116.633  741       10
## 5: 726813 94195 2019     8  31   22  56 43.642 -116.636  741       10
## 6: 726813 94195 2019     8  31   23  56 43.642 -116.636  741       40
##    wind.dir.qc wind.type.code wind.sp wind.sp.qc ceiling.ht ceiling.ht.qc
## 1:           9              C     0.0          5      22000             5
## 2:           5              N     2.1          5      22000             5
## 3:           9              C     0.0          5      22000             5
## 4:           5              N     2.6          5      22000             5
## 5:           1              N     2.1          1      22000             1
## 6:           1              N     2.1          1      22000             1
##    ceiling.ht.method sky.cond vis.dist vis.dist.qc vis.var vis.var.qc temp
## 1:                 9        N    16093           5       N          5 30.0
## 2:                 9        N    16093           5       N          5 32.2
## 3:                 9        N    16093           5       N          5 33.3
## 4:                 9        N    14484           5       N          5 35.0
## 5:                 9        N    16093           1       9          9 34.4
## 6:                 9        N    16093           1       9          9 34.4
##    temp.qc dew.point dew.point.qc atm.press atm.press.qc       rh
## 1:       5      11.7            5    1013.6           5 32.32509
```

```
## 2:       5      12.2           5    1012.8           5 29.40686
## 3:       5      12.2           5    1011.6           5 27.60422
## 4:       5       9.4           5    1010.8           5 20.76325
## 5:       1       9.4           1    1010.1           1 21.48631
## 6:       1       9.4           1    1009.6           1 21.48631
```

There are 2,377,343 rows and 30 columns in the `met` dataset.

3. Take a look at the variables.

```
str(met)
```

```
## Classes 'data.table' and 'data.frame':   2377343 obs. of  30 variables:
##  $ USAFID          : int  690150 690150 690150 690150 690150 690150 690150 690150 690150 690150 ...
##  $ WBAN            : int  93121 93121 93121 93121 93121 93121 93121 93121 93121 93121 ...
##  $ year            : int  2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 ...
##  $ month           : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ day             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ hour            : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ min             : int  56 56 56 56 56 56 56 56 56 56 ...
##  $ lat             : num  34.3 34.3 34.3 34.3 34.3 34.3 34.3 34.3 34.3 34.3 ...
##  $ lon             : num  -116 -116 -116 -116 -116 ...
##  $ elev            : int  696 696 696 696 696 696 696 696 696 696 ...
##  $ wind.dir        : int  220 230 230 210 120 NA 320 10 320 350 ...
##  $ wind.dir.qc     : chr  "5" "5" "5" "5" ...
##  $ wind.type.code  : chr  "N" "N" "N" "N" ...
##  $ wind.sp         : num  5.7 8.2 6.7 5.1 2.1 0 1.5 2.1 2.6 1.5 ...
##  $ wind.sp.qc      : chr  "5" "5" "5" "5" ...
##  $ ceiling.ht      : int  22000 22000 22000 22000 22000 22000 22000 22000 22000 22000 ...
##  $ ceiling.ht.qc   : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ ceiling.ht.method: chr  "9" "9" "9" "9" ...
##  $ sky.cond        : chr  "N" "N" "N" "N" ...
##  $ vis.dist        : int  16093 16093 16093 16093 16093 16093 16093 16093 16093 16093 ...
##  $ vis.dist.qc     : chr  "5" "5" "5" "5" ...
##  $ vis.var         : chr  "N" "N" "N" "N" ...
##  $ vis.var.qc      : chr  "5" "5" "5" "5" ...
##  $ temp            : num  37.2 35.6 34.4 33.3 32.8 31.1 29.4 28.9 27.2 26.7 ...
##  $ temp.qc         : chr  "5" "5" "5" "5" ...
##  $ dew.point       : num  10.6 10.6 7.2 5 5 5.6 6.1 6.7 7.8 7.8 ...
##  $ dew.point.qc    : chr  "5" "5" "5" "5" ...
##  $ atm.press       : num  1010 1010 1011 1012 1013 ...
##  $ atm.press.qc    : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ rh              : num  19.9 21.8 18.5 16.9 17.4 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

4. Take a closer look at the key variables.

```
table(met$year)
```

```
##
##    2019
## 2377343
```

```
table(met$day)
```

```
##
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## 75975 75923 76915 76594 76332 76734 77677 77766 75366 75450 76187 75052 76906
##    14    15    16    17    18    19    20    21    22    23    24    25    26
## 77852 76217 78015 78219 79191 76709 75527 75786 78312 77413 76965 76806 79114
##    27    28    29    30    31
## 79789 77059 71712 74931 74849
```

```
table(met$hour)
```

```
##
##      0      1      2      3      4      5      6      7      8      9     10
##  99434  93482  93770  96703 110504 112128 106235 101985 100310 102915 101880
##     11     12     13     14     15     16     17     18     19     20     21
## 100470 103605  97004  96507  97635  94942  94184 100179  94604  94928  96070
##     22     23
##  94046  93823
```

```
summary(met$temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -40.00   19.60   23.50   23.59   27.80   56.00   60089
```

```
summary(met$elev)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -13.0   101.0   252.0   415.8   400.0  9999.0
```

```
summary(met$wind.sp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    0.00    2.10    2.46    3.60   36.00   79693
```

It looks like the elevation variable has observations with 9999.0, which is probably an indicator for missing. We should take a deeper look at the data dictionary to confirm. The wind speed variable is ok but there are a lot of missing data.

After checking the data we should make the appropriate modifications. Replace elevations with 9999 as `NA`.

```
# BASE R:
met$elev[met&elev == 9999.0] <- NA
summary(met$elev)
```

```
# tidyverse
met <- met %>%
  mutate(elev = ifelse(elev == 9999, NA, elev))
```

```
# data.table:
met[elev == 9999, elev:=NA]
```

At what elevation is the highest weather station?

- *Summarize here*

We also have the issue of the minimum temperature being -40C, so we should remove those observations.

```
# <place your code here>
table(met$temp > -40, useNA = "always")
```

```
##
##   FALSE    TRUE    <NA>
##      36 2317218   60089
```

```
# met <- met[temp > 40] This will remove NA
# sum(is.na(met$temp)) 0
met <- met[! temp %in% c(-40)]
sum(is.na(met$temp))
```

```
## [1] 60089
```

5. Check the data against an external data source.

We should check the suspicious temperature value (where is it located?) and validate that the range of elevations make sense (-13 m to 4113 m).

Google is your friend here.

Fix any problems that arise in your checks.

```
summary(met$temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -17.20   19.60   23.50   23.59   27.80   56.00   60089
```

```
summary(met$elev)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     -13     101     252     413     400    4113     710
```

```
unique(met[met$elev == 4113, .(lat, lon, elev)])
```

```
##     lat      lon elev
## 1: 39.8 -105.766 4113
```

```
met[temp == max(temp, na.rm = TRUE)]
```

```
##    USAFID  WBAN year month day hour min    lat      lon elev wind.dir
## 1: 720267 23224 2019     8  26   11  15 38.955 -121.081  467       NA
##    wind.dir.qc wind.type.code wind.sp wind.sp.qc ceiling.ht ceiling.ht.qc
## 1:           9              C       0          5      22000             5
##    ceiling.ht.method sky.cond vis.dist vis.dist.qc vis.var vis.var.qc temp
## 1:                 9        N    16093           5       N          5   56
##    temp.qc dew.point dew.point.qc atm.press atm.press.qc rh
## 1:       5        NA            9        NA            9 NA
```

- *Summarize anything that was removed*

6. Calculate summary statistics

Remember to keep the initial question in mind. We want to pick out the weather station with maximum elevation and examine its wind speed and temperature.

Some ideas: 1. select the weather station with maximum elevation; 2. look at the correlation between temperature and wind speed; and 3. look at the correlation between temperature and wind speed with hour and day of the month.

```
highest <- met[elev == max(elev, na.rm = TRUE)]
highest <- highest[!is.na(highest$temp)]
highest <- highest[!is.na(highest$wind.sp)]
highest <- highest[!is.na(highest$hour)]
highest <- highest[!is.na(highest$day)]
highest <- highest[!is.na(highest$month)]
highest
```

```
##       USAFID WBAN year month day hour min  lat      lon elev wind.dir
##    1: 720385  419 2019     8   1    0  36 39.8 -105.766 4113      170
##    2: 720385  419 2019     8   1    0  54 39.8 -105.766 4113      100
##    3: 720385  419 2019     8   1    1  12 39.8 -105.766 4113       90
##    4: 720385  419 2019     8   1    1  35 39.8 -105.766 4113      110
##    5: 720385  419 2019     8   1    1  53 39.8 -105.766 4113      120
##   ---
## 1945: 720385  419 2019     8  31   21  12 39.8 -105.766 4113      310
## 1946: 720385  419 2019     8  31   21  36 39.8 -105.766 4113       10
## 1947: 720385  419 2019     8  31   21  54 39.8 -105.766 4113      300
## 1948: 720385  419 2019     8  31   22  11 39.8 -105.766 4113      310
## 1949: 720385  419 2019     8  31   22  35 39.8 -105.766 4113      290
##       wind.dir.qc wind.type.code wind.sp wind.sp.qc ceiling.ht ceiling.ht.qc
##    1:           5              N     8.8          5       1372             5
##    2:           5              N     2.6          5       1372             5
##    3:           5              N     3.1          5       1981             5
##    4:           5              N     4.1          5       2134             5
##    5:           5              N     4.6          5       2134             5
##   ---
## 1945:           5              N     5.7          5       3048             5
## 1946:           5              N     1.5          5       2743             5
## 1947:           5              N     8.2          5       3048             5
```

```
## 1948:           5             N     6.2          5       3048            5
## 1949:           5             N     6.7          5       2438            5
##       ceiling.ht.method sky.cond vis.dist vis.dist.qc vis.var vis.var.qc temp
##    1:                 M        N       NA           9       N          5    9
##    2:                 M        N       NA           9       N          5    9
##    3:                 M        N       NA           9       N          5    9
##    4:                 M        N       NA           9       N          5    9
##    5:                 M        N       NA           9       N          5    9
##   ---
## 1945:                 M        N    16093           5       N          5   12
## 1946:                 M        N    16093           5       N          5   13
## 1947:                 M        N    16093           5       N          5   12
## 1948:                 M        N    16093           5       N          5   12
## 1949:                 M        N    16093           5       N          5   12
##       temp.qc dew.point dew.point.qc atm.press atm.press.qc       rh
##    1:       5         1            5        NA            9 57.61039
##    2:       5         1            5        NA            9 57.61039
##    3:       5         2            5        NA            9 61.85243
##    4:       5         2            5        NA            9 61.85243
##    5:       5         2            5        NA            9 61.85243
##   ---
## 1945:       5         3            5        NA            9 54.39402
## 1946:       5         5            5        NA            9 58.57459
## 1947:       C         4            C        NA            9 58.33755
## 1948:       5         3            5        NA            9 54.39402
## 1949:       5         3            5        NA            9 54.39402
```

```
cor(highest$temp, highest$wind.sp)
```

```
## [1] -0.09373843
```

```
cor(highest$temp, highest$hour)
```

```
## [1] 0.4356801
```

```
cor(highest$temp, highest$day)
```

```
## [1] -0.006130763
```

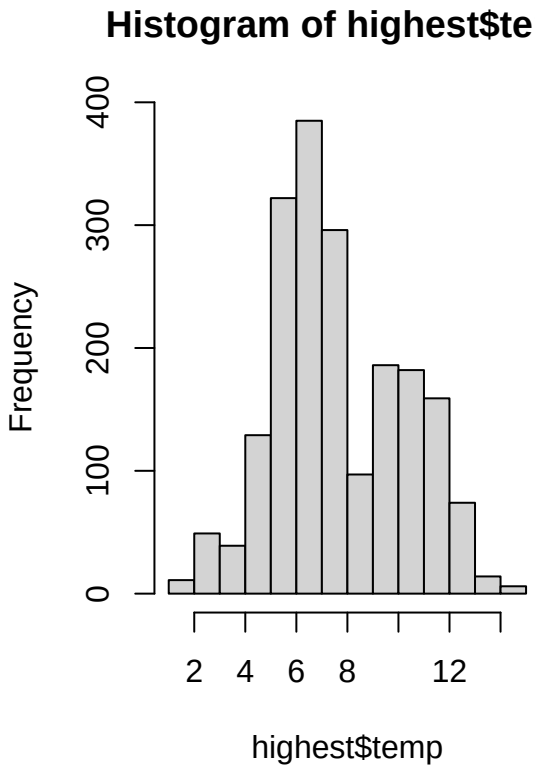
7. Exploratory graphs

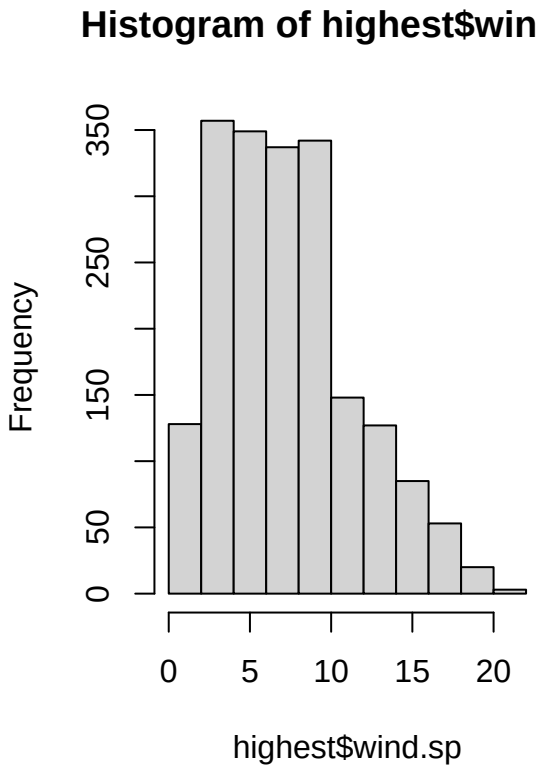
We should look at the distributions of all of the key variables (elevation, temp, wind speed) to make sure there are no remaining issues with the data.

```
par(mfrow = c(1,2))
summary(highest$elev)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4113    4113    4113    4113    4113    4113
```

```
hist(highest$temp)
hist(highest$wind.sp)
```





One thing we should consider for later analyses is to log transform wind speed and elevation as the are very skewed.

Look at where the weather station with highest elevation is located (i.e. make a map!)

```
# <place your code here>
# hint: make use of leaflet
leaflet(highest) %>%
  addProviderTiles('OpenStreetMap') %>%
  addCircles(lng = ~lon, lat = ~lat, fillColor = "orange", fillOpacity = 1, radius = 100)
```

- *Summarize*

Look at the time series of temperature and wind speed at this location. For this we will need to create a date-time variable for the x-axis.

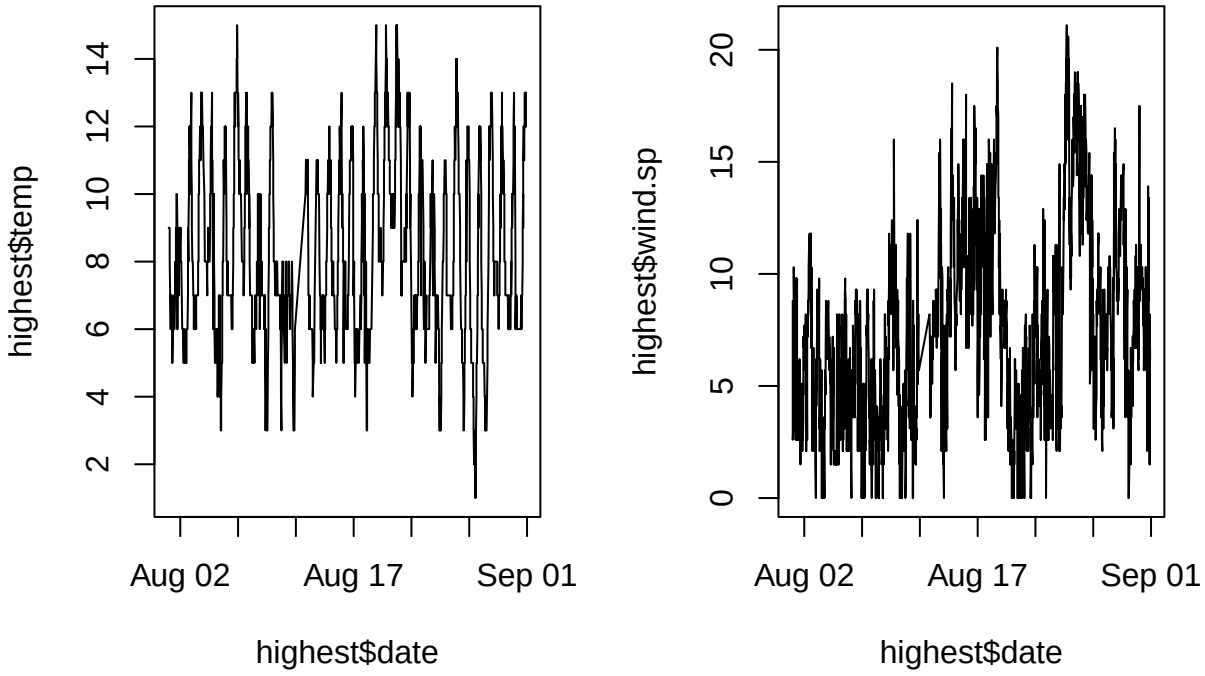
```
library(lubridate)
# highest$date <- ymd_h(paste(highest$year, highest$month, highest$day, highest$hour))
highest$date <- with(highest,
  ymd_h(paste(year, month, day, hour))
)
```

```
str(highest$date)
```

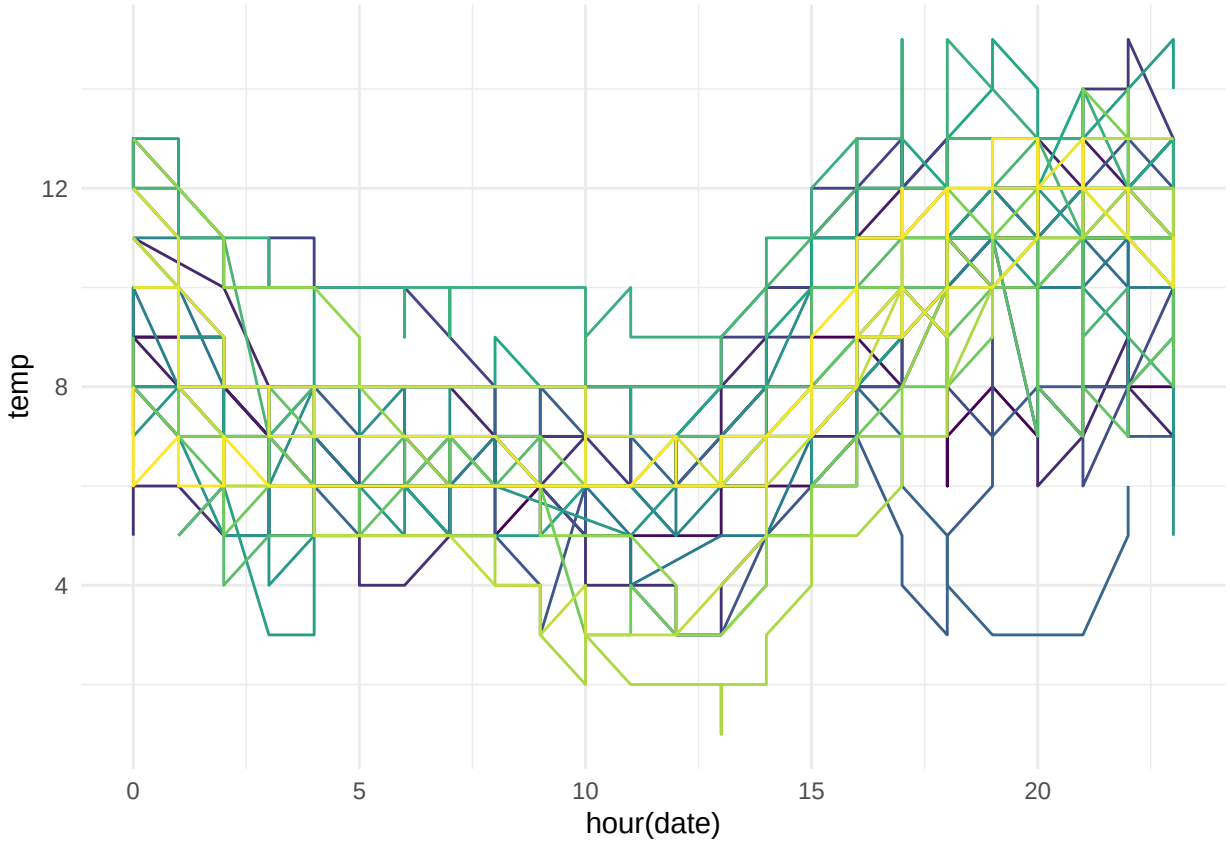
```
##  POSIXct[1:1949], format: "2019-08-01 00:00:00" "2019-08-01 00:00:00" "2019-08-01 01:00:00" ...
```

With the date-time variable we can plot the time series of temperature and wind speed.

```
par(mfrow = c(1, 2))
plot(highest$date, highest$temp, type = 'l')
plot(highest$date, highest$wind.sp, type = 'l')
```




```
ggplot(highest) +
  theme_minimal() +
  geom_line(aes(x = hour(date), y = temp, group = yday(date), color = yday(date))
            ,show.legend = FALSE) + scale_colour_viridis_c()
```



Summarize any trends that you see in these time series plots. The average daily temperature looks stable in August, The average daily wind speed looks highest around August 25. Both temperature and wind varys within a day.