Lab 3

Jazmin Hernandez

Lab 03 - Exploratory Data Analysis

1. Read in the data

```
met <- read.csv(file.path("~", "Github", "met_all.gz"))
```

2. Check the dimensions, headers, footers

```
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
```

```r
dim(met)
```

```
[1] 2377343      30
```

```r
tail(met)
```

```
        USAFID  WBAN year month day hour min    lat      lon elev wind.dir
2377338 726813 94195 2019     8  31   18  56 43.650 -116.633  741       NA
2377339 726813 94195 2019     8  31   19  56 43.650 -116.633  741       70
2377340 726813 94195 2019     8  31   20  56 43.650 -116.633  741       NA
2377341 726813 94195 2019     8  31   21  56 43.650 -116.633  741       10
2377342 726813 94195 2019     8  31   22  56 43.642 -116.636  741       10
2377343 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
2377338           9              C     0.0          5      22000             5
2377339           5              N     2.1          5      22000             5
2377340           9              C     0.0          5      22000             5
2377341           5              N     2.6          5      22000             5
2377342           1              N     2.1          1      22000             1
2377343           1              N     2.1          1      22000             1
        ceiling.ht.method sky.cond vis.dist vis.dist.qc vis.var vis.var.qc temp
2377338                 9        N    16093           5       N          5 30.0
2377339                 9        N    16093           5       N          5 32.2
2377340                 9        N    16093           5       N          5 33.3
2377341                 9        N    14484           5       N          5 35.0
2377342                 9        N    16093           1       9          9 34.4
2377343                 9        N    16093           1       9          9 34.4
        temp.qc dew.point dew.point.qc atm.press atm.press.qc       rh
2377338       5      11.7            5    1013.6            5 32.32509
2377339       5      12.2            5    1012.8            5 29.40686
```

```
2377340       5      12.2            5    1011.6            5 27.60422
2377341       5       9.4            5    1010.8            5 20.76325
2377342       1       9.4            1    1010.1            1 21.48631
2377343       1       9.4            1    1009.6            1 21.48631
```

There are 30 columns and 6 rows.

3. Take a look at the variables

```r
str(met)
```

```
'data.frame':   2377343 obs. of  30 variables:
 $ USAFID           : int  690150 690150 690150 690150 690150 690150 690150 690150 690150 69
 $ 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 ...
```

The key variables related to our question of interest are in the time variables, wind speed, temperature and elevation. More specifically, the variables for time series include: year, month, day, hour and minute. Variables for wind speed include wind.sp. Variables for temperature include temp. Variables for elevation include elev.

4. Take a closer look at the key variables

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

```
   2019
2377343
```

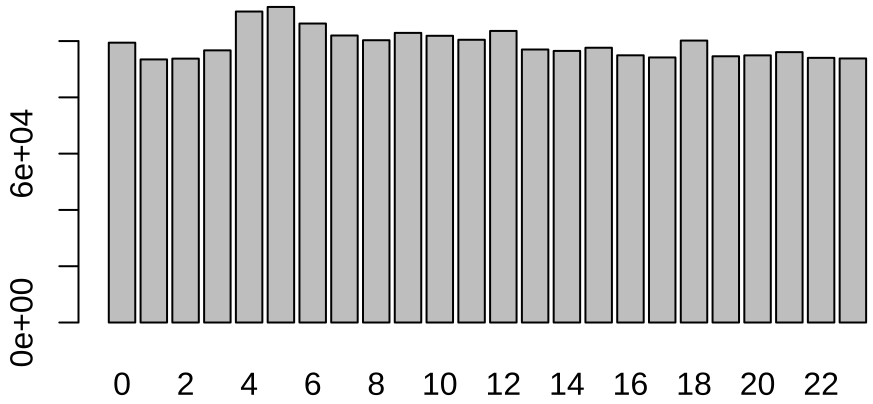
```
table(met$day)
```

```
    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
```

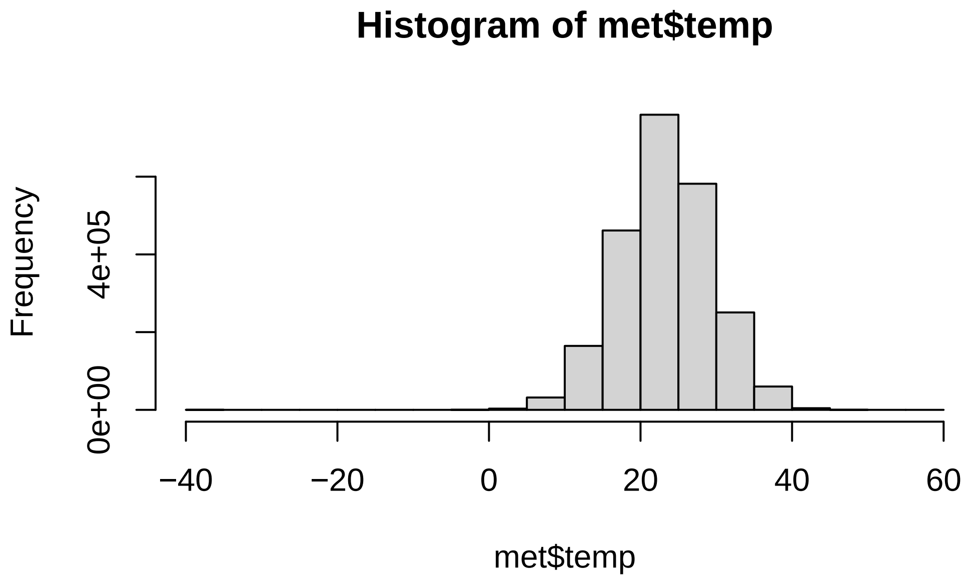
```
barplot(table(met$hour))
```



```
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 
```

```
hist(met$temp)
```



```
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 
```

```
met$elev[met$elev == 9999.0] <- NA
summary(met$elev)
```

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

The highest weather station is at 4113m.

```
met <- met[met$temp > -40, ]
head(met[order(met$temp), ])
```

```
        USAFID WBAN year month day hour min    lat    lon elev wind.dir
1203053 722817 3068 2019     8   1    0  56 38.767 -104.3 1838      190
1203055 722817 3068 2019     8   1    1  56 38.767 -104.3 1838      180
1203128 722817 3068 2019     8   3   11  56 38.767 -104.3 1838       NA
1203129 722817 3068 2019     8   3   12  56 38.767 -104.3 1838       NA
1203222 722817 3068 2019     8   6   21  56 38.767 -104.3 1838      280
1203225 722817 3068 2019     8   6   22  56 38.767 -104.3 1838      240
        wind.dir.qc wind.type.code wind.sp wind.sp.qc ceiling.ht ceiling.ht.qc
1203053           5              N     7.2          5         NA             9
1203055           5              N     7.7          5         NA             9
1203128           9              C     0.0          5         NA             9
1203129           9              C     0.0          5         NA             9
1203222           5              N     2.6          5         NA             9
1203225           5              N     7.7          5         NA             9
        ceiling.ht.method sky.cond vis.dist vis.dist.qc vis.var vis.var.qc
1203053                 9        N       NA           9       N          5
1203055                 9        N       NA           9       N          5
1203128                 9        N       NA           9       N          5
1203129                 9        N       NA           9       N          5
1203222                 9        N       NA           9       N          5
1203225                 9        N       NA           9       N          5
         temp temp.qc dew.point dew.point.qc atm.press atm.press.qc rh
1203053 -17.2       5        NA            9        NA            9 NA
1203055 -17.2       5        NA            9        NA            9 NA
1203128 -17.2       5        NA            9        NA            9 NA
1203129 -17.2       5        NA            9        NA            9 NA
1203222 -17.2       5        NA            9        NA            9 NA
1203225 -17.2       5        NA            9        NA            9 NA
```

```
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   91832 
```

```
mean(is.na(met$wind.sp))
```

```
[1] 0.03862858
```

There are 91,853 missing values in the wind.sp variable. 3.8% of the data are missing.

5. Check the data against an external data source

Using Google to find the location of the coordinates (38.767, -104.300) where the temperature is -17.2C, we find that this is in Yoder, Colorado. The elevation of Yoder, Colorado is about 1,873m which is about the same as we have recorded in the data set (1838m). The elevation range does make sense because Yoder, Co falls between this. The temperature does not make sense given the month of August that it was recorded in. According to Google, temperatures in Yoder, Colorado during August are in the 80F range.

```r
met <- met[met$temp > -17.2, ]
met <- met[!is.na(met$temp), ]
met <- met[order(met$temp), ]
```

6. Calculate summary statistics

```r
elev <- met[which(met$elev == max(met$elev, na.rm = TRUE)), ]
summary(elev)
```

```
     USAFID            WBAN          year          month        day      
 Min.   :720385   Min.   :419   Min.   :2019   Min.   :8   Min.   : 1.0  
 1st Qu.:720385   1st Qu.:419   1st Qu.:2019   1st Qu.:8   1st Qu.: 8.0  
 Median :720385   Median :419   Median :2019   Median :8   Median :16.0  
 Mean   :720385   Mean   :419   Mean   :2019   Mean   :8   Mean   :16.1  
 3rd Qu.:720385   3rd Qu.:419   3rd Qu.:2019   3rd Qu.:8   3rd Qu.:24.0  
 Max.   :720385   Max.   :419   Max.   :2019   Max.   :8   Max.   :31.0  
                                                                         
      hour            min             lat             lon              elev     
 Min.   : 0.00   Min.   : 6.00   Min.   :39.8   Min.   :-105.8   Min.   :4113  
 1st Qu.: 6.00   1st Qu.:13.00   1st Qu.:39.8   1st Qu.:-105.8   1st Qu.:4113  
 Median :12.00   Median :36.00   Median :39.8   Median :-105.8   Median :4113  
 Mean   :11.66   Mean   :34.38   Mean   :39.8   Mean   :-105.8   Mean   :4113  
 3rd Qu.:18.00   3rd Qu.:53.00   3rd Qu.:39.8   3rd Qu.:-105.8   3rd Qu.:4113  
 Max.   :23.00   Max.   :59.00   Max.   :39.8   Max.   :-105.8   Max.   :4113  
                                                                               
    wind.dir      wind.dir.qc        wind.type.code        wind.sp     
 Min.   : 10.0   Length:2117        Length:2117        Min.   : 0.000  
 1st Qu.:250.0   Class :character   Class :character   1st Qu.: 4.100  
 Median :300.0   Mode  :character   Mode  :character   Median : 6.700  
 Mean   :261.5                                         Mean   : 7.245  
 3rd Qu.:310.0                                         3rd Qu.: 9.800  
```

```
Max.   :360.0                                         Max.   :21.100  
NA's   :237                                           NA's   :168     
 wind.sp.qc          ceiling.ht    ceiling.ht.qc   ceiling.ht.method
Length:2117        Min.   :   30   Min.   :5.000   Length:2117      
Class :character   1st Qu.: 2591   1st Qu.:5.000   Class :character 
Mode  :character   Median :22000   Median :5.000   Mode  :character 
                   Mean   :15145   Mean   :5.008                    
                   3rd Qu.:22000   3rd Qu.:5.000                    
                   Max.   :22000   Max.   :9.000                    
                   NA's   :4                                        
  sky.cond            vis.dist     vis.dist.qc          vis.var         
Length:2117        Min.   :    0   Length:2117        Length:2117       
Class :character   1st Qu.:16093   Class :character   Class :character  
Mode  :character   Median :16093   Mode  :character   Mode  :character  
                   Mean   :15913                                        
                   3rd Qu.:16093                                        
                   Max.   :16093                                        
                   NA's   :683                                          
 vis.var.qc             temp         temp.qc            dew.point      
Length:2117        Min.   : 1.00   Length:2117        Min.   :-6.0000  
Class :character   1st Qu.: 6.00   Class :character   1st Qu.: 0.0000  
Mode  :character   Median : 8.00   Mode  :character   Median : 0.0000  
                   Mean   : 8.13                      Mean   : 0.8729  
                   3rd Qu.:10.00                      3rd Qu.: 2.0000  
                   Max.   :15.00                      Max.   : 7.0000  

dew.point.qc         atm.press     atm.press.qc       rh       
Length:2117        Min.   : NA    Min.   :9     Min.   :53.63  
Class :character   1st Qu.: NA    1st Qu.:9     1st Qu.:58.10  
Mode  :character   Median : NA    Median :9     Median :61.39  
                   Mean   :NaN    Mean   :9     Mean   :60.62  
                   3rd Qu.: NA    3rd Qu.:9     3rd Qu.:61.85  
                   Max.   : NA    Max.   :9     Max.   :70.01  
                   NA's   :2117                                
```

```
cor(elev$temp, elev$wind.sp, use="complete")
```

```
[1] -0.09373843
```

```
cor(elev$temp, elev$hour, use="complete")
```

```
[1] 0.4397261
```

```r
cor(elev$wind.sp, elev$day, use="complete")
```

```
[1] 0.3643079
```

```r
cor(elev$wind.sp, elev$hour, use="complete")
```

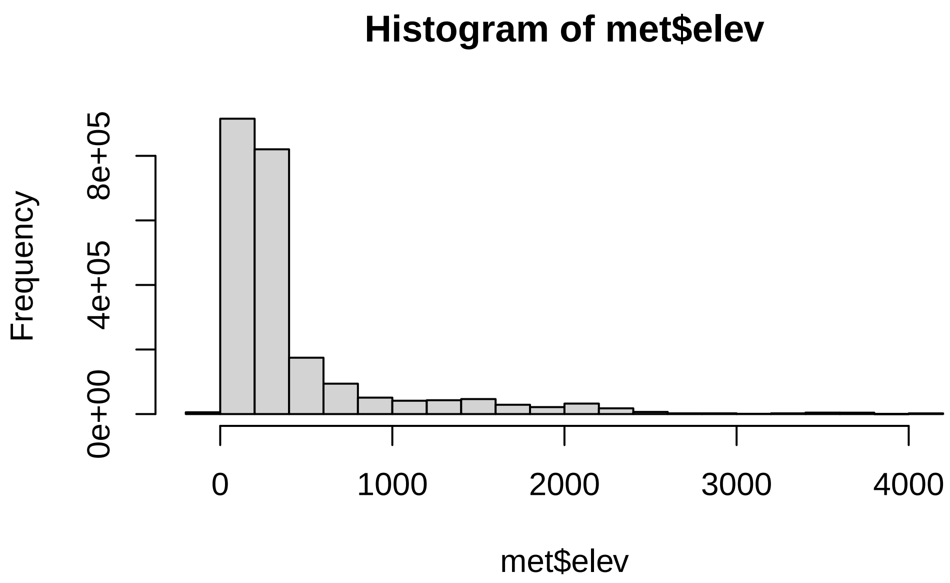
```
[1] 0.08807315
```

```r
cor(elev$temp, elev$day, use="complete")
```

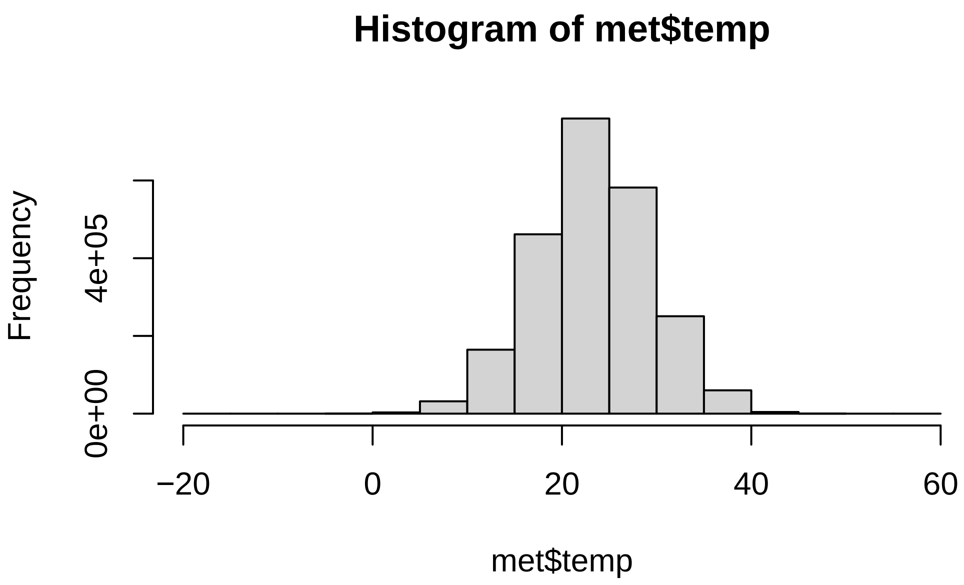
```
[1] -0.003857766
```

7. Exploratory graphs

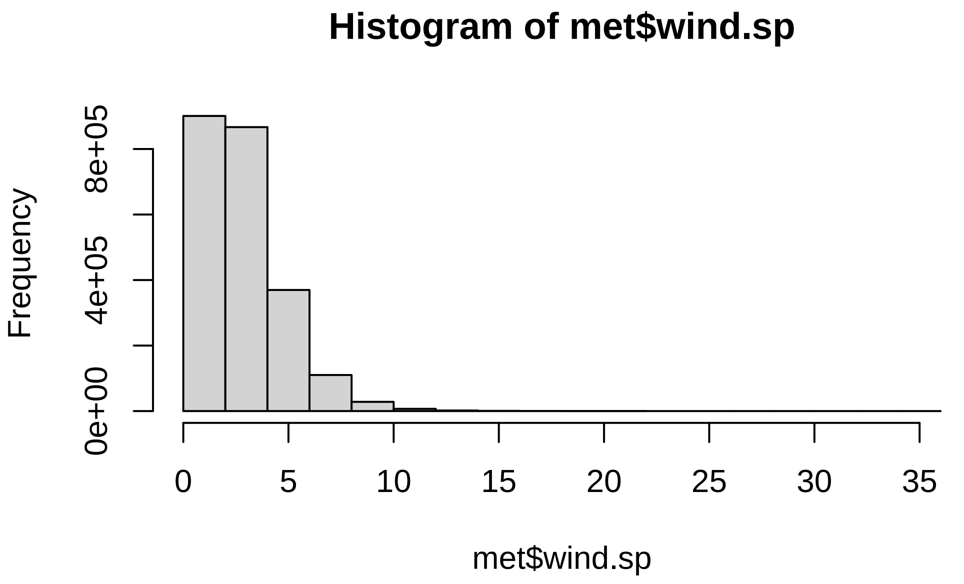
```r
hist(met$elev)
```



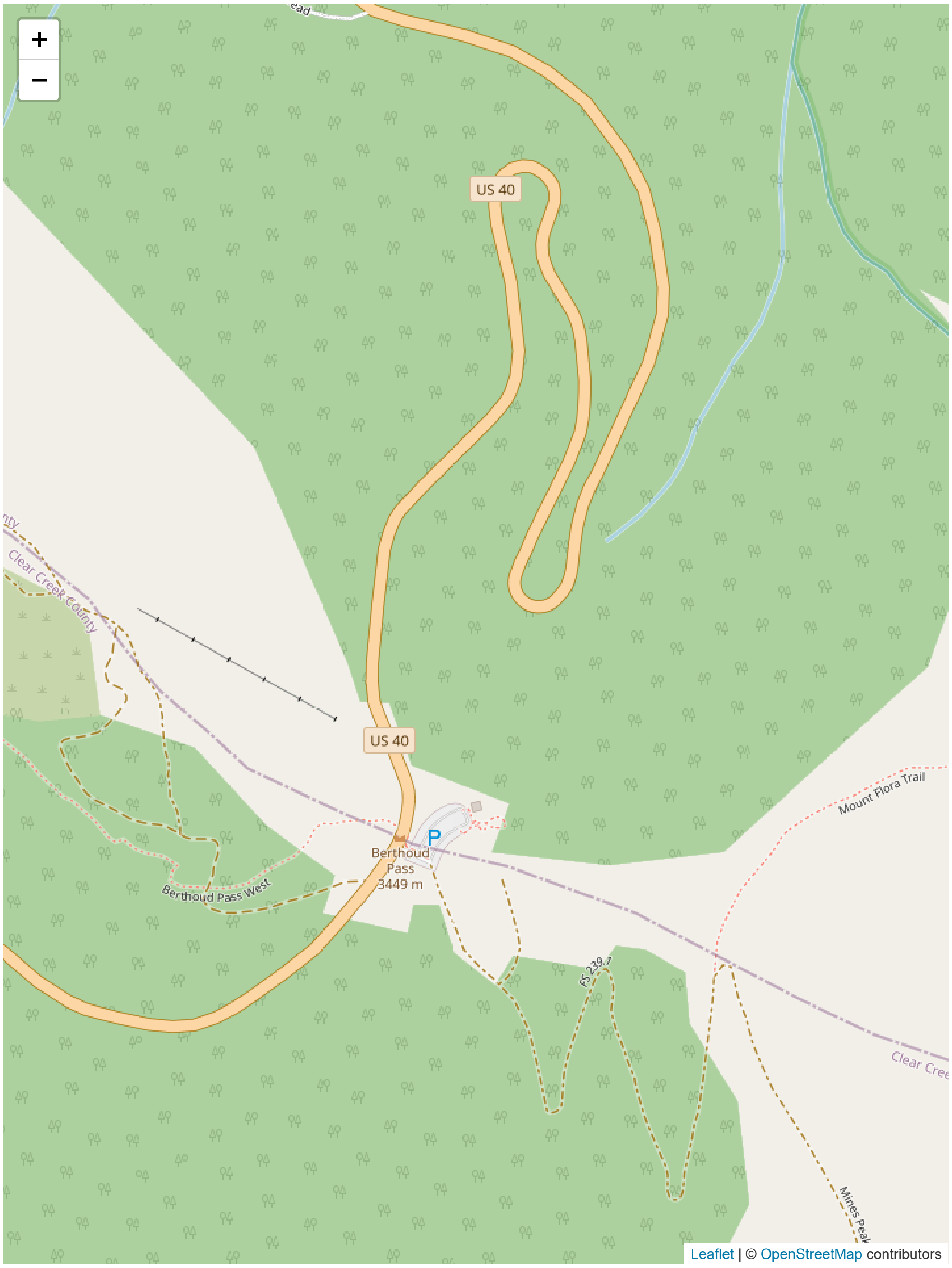
```
hist(met$temp)
```



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



```r
library(leaflet)
leaflet(elev) %>%
  addProviderTiles('OpenStreetMap') %>%
  addCircles(lat = ~lat, lng = ~lon, opacity = 1, fillOpacity = 1, radius = 100)
```


+
−
ead
US 40
nty
Clear Creek County
US 40
P
Berthoud
Pass
3449 m
Berthoud Pass West
Mount Flora Trail
FS 239.1
Clear Cree
Mines Peak
Leaflet | © OpenStreetMap contributors

```r
library(lubridate)
```

```

Attaching package: 'lubridate'


The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```r
elev$date <- with(elev, ymd_h(paste(year, month, day, hour, sep= ' ')))
summary(elev$date)
```

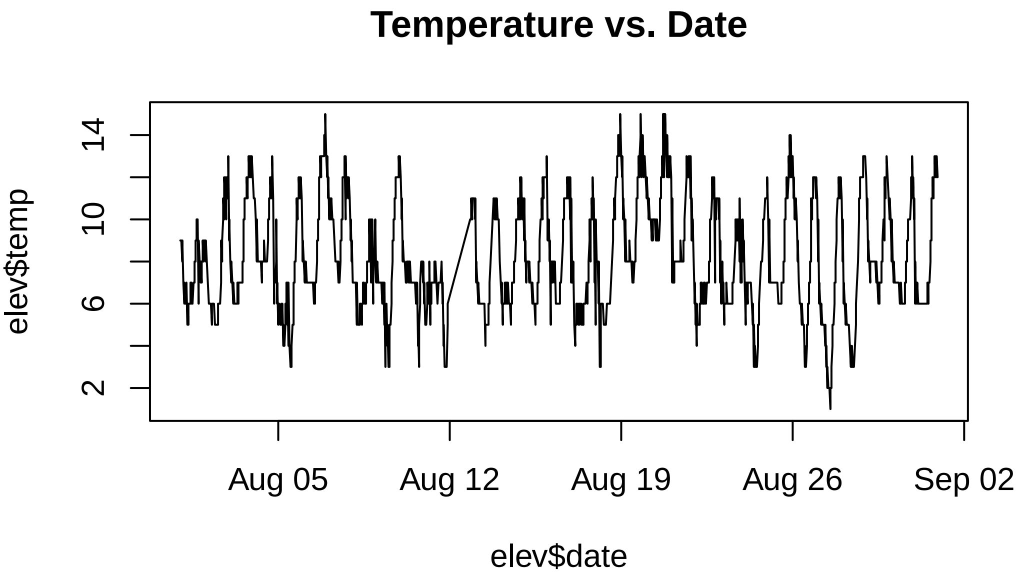
```
                      Min.                    1st Qu. 
"2019-08-01 00:00:00.0000" "2019-08-08 11:00:00.0000" 
                    Median                       Mean 
"2019-08-16 22:00:00.0000" "2019-08-16 14:09:56.8823" 
                   3rd Qu.                       Max. 
"2019-08-24 11:00:00.0000" "2019-08-31 22:00:00.0000" 
```

```r
elev <- elev[order(elev$date), ]
head(elev)
```

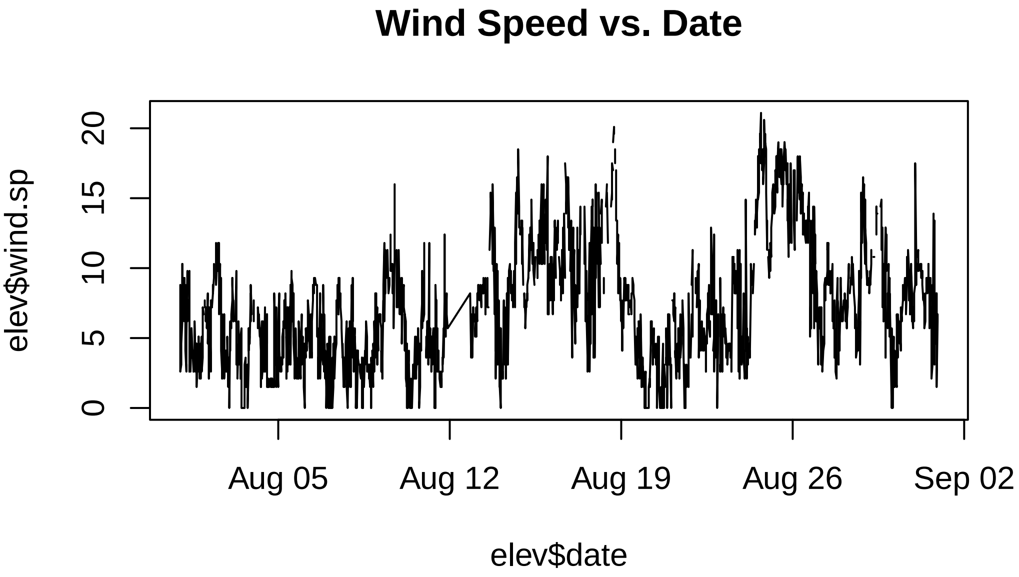
```
       USAFID WBAN year month day hour min  lat      lon elev wind.dir
221697 720385  419 2019     8   1    0  36 39.8 -105.766 4113      170
221698 720385  419 2019     8   1    0  54 39.8 -105.766 4113      100
221699 720385  419 2019     8   1    1  12 39.8 -105.766 4113       90
221700 720385  419 2019     8   1    1  35 39.8 -105.766 4113      110
221701 720385  419 2019     8   1    1  53 39.8 -105.766 4113      120
221703 720385  419 2019     8   1    2  36 39.8 -105.766 4113      110
       wind.dir.qc wind.type.code wind.sp wind.sp.qc ceiling.ht ceiling.ht.qc
221697           5              N     8.8          5       1372             5
221698           5              N     2.6          5       1372             5
221699           5              N     3.1          5       1981             5
221700           5              N     4.1          5       2134             5
221701           5              N     4.6          5       2134             5
221703           5              N     6.2          5      22000             5
       ceiling.ht.method sky.cond vis.dist vis.dist.qc vis.var vis.var.qc temp
221697                 M        N       NA           9       N          5    9
221698                 M        N       NA           9       N          5    9
221699                 M        N       NA           9       N          5    9
```

```
221700                 M        N       NA           9       N          5    9
221701                 M        N       NA           9       N          5    9
221703                 9        N       NA           9       N          5    8
       temp.qc dew.point dew.point.qc atm.press atm.press.qc        rh
221697       5         1            5        NA            9 57.61039
221698       5         1            5        NA            9 57.61039
221699       5         2            5        NA            9 61.85243
221700       5         2            5        NA            9 61.85243
221701       5         2            5        NA            9 61.85243
221703       5         1            5        NA            9 61.62158
                      date
221697 2019-08-01 00:00:00
221698 2019-08-01 00:00:00
221699 2019-08-01 01:00:00
221700 2019-08-01 01:00:00
221701 2019-08-01 01:00:00
221703 2019-08-01 02:00:00
```

```r
plot(elev$date, elev$temp, type="l",
main = "Temperature vs. Date")
```



```
plot(elev$date, elev$wind.sp, type="l",
main = "Wind Speed vs. Date")
```



From the time series plots of temperature versus date, we can see that the highest peak in temperature seems to occur right around August 5th with the lowest peak occurring near the start of September, around August 26th. The wind speed versus date plot shows the highest peak in wind speed around August 26th which can relate to lower temperatures or a change in temperature as seen in the temperature versus date plot.

8. Ask questions

I did have a question as to what the variable wind.type.code represented. Upon looking through the data dictionary, I did find out that it meant "Wind-observation type code" where a value of N for example, meant normal winds. I would like to know however, what exactly characterizes wind observations to be "normal" versus "Beaufort."