PM 566: Lab 03

AUTHOR

Tarun Mahesh

Part 1: Checking the dimensions, headers and footers

1. How many columns and rows are there?

```r
met <- read.csv("met_all.gz")
dim(met)
```

```
[1] 2377343      30
```

```r
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
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
```

Taking a look at the variables:

```
str(met)
```

```
'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 ...
```

Taking a closer look at only the key variables i.e. Year, Day, Hour, Temperature, Elevation and Wind Speed

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

```
   2019 
2377343 
```

All data is from the year 2019

```r
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 
```

```r
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 
```

```r
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 
```

Anomalies: The minimum temperature recorded in this dataset is -40, the maximum elevation recorded is 9999, and the wind speed data has 79693 NA values.

Fixes: 1. Replacing all 9999 elevations (impossible) with NA

```
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 
```

Now, the highest weather station is at an elevation of 4113 m.

2. Minimum temperature of -40C looks suspiscious and observations of -40 degC temperatures should be removed.

```
met <- met[met$temp > -40, ]
met2 <-met[order(met$temp), ]
head(met2)
```

```
        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
```

met2 variable is now assigned the ascending order of temperature values from our main dataset that are > -40C, and we notice that the minimum temperature is now -17.2C.

The location of the “suspicious” temperature is lat 38.767 and lon -104.3, which according to Google Earth is at 1838m.

Let us remove all the temperatures colder than -15 degC, and summarise the data.

```r
met <- met[met$temp > -15, ]
met2 <- met[order(met$temp), ]
head(met2)
```

```
        USAFID  WBAN year month day hour min    lat      lon elev wind.dir
2370758 726764 94163 2019     8  27   11  50 44.683 -111.116 2025       NA
2370759 726764 94163 2019     8  27   12  10 44.683 -111.116 2025       NA
2370760 726764 94163 2019     8  27   12  30 44.683 -111.116 2025       NA
```

```
2370761 726764 94163 2019     8  27    12  50 44.683 -111.116 2025       NA
252489  720411   137 2019     8  18    12  35 36.422 -105.290 2554       NA
2370688 726764 94163 2019     8  26    12  30 44.683 -111.116 2025       NA
        wind.dir.qc wind.type.code wind.sp wind.sp.qc ceiling.ht ceiling.ht.qc
2370758           9              C       0          5      22000             5
2370759           9              C       0          5      22000             5
2370760           9              C       0          5      22000             5
2370761           9              C       0          5      22000             5
252489            9              C       0          5      22000             5
2370688           9              C       0          5      22000             5
        ceiling.ht.method sky.cond vis.dist vis.dist.qc vis.var vis.var.qc temp
2370758                 9        N    16093           5       N          5 -3.0
2370759                 9        N    16093           5       N          5 -3.0
2370760                 9        N    16093           5       N          5 -3.0
2370761                 9        N    16093           5       N          5 -3.0
252489                  9        N    16093           5       N          5 -2.4
2370688                 9        N    16093           5       N          5 -2.0
        temp.qc dew.point dew.point.qc atm.press atm.press.qc       rh
2370758       C      -5.0            C        NA            9 86.26537
2370759       5      -4.0            5        NA            9 92.91083
2370760       5      -4.0            5        NA            9 92.91083
2370761       C      -4.0            C        NA            9 92.91083
252489        5      -3.7            5        NA            9 90.91475
2370688       5      -3.0            5        NA            9 92.96690
```

From head(met2), we note that the new minimum temperature is a more reasonable -3 deg C.

Part 2: Calculation of summary statistics

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

```
     USAFID            WBAN            year          month  
 Min.   :720385   Min.   :419     Min.   :2019    Min.   :8  
 1st Qu.:720385   1st Qu.:419     1st Qu.:2019    1st Qu.:8  
 Median :720385   Median :419     Median :2019    Median :8  
 Mean   :720385   Mean   :419     Mean   :2019    Mean   :8  
 3rd Qu.:720385   3rd Qu.:419     3rd Qu.:2019    3rd Qu.:8  
```

```
Max.   :720385   Max.   :419     Max.   :2019    Max.   :8
NA's   :60271    NA's   :60271   NA's   :60271   NA's   :60271
     day             hour            min             lat
Min.   : 1.0    Min.   : 0.00   Min.   : 6.00   Min.   :39.8
1st Qu.: 8.0    1st Qu.: 6.00   1st Qu.:13.00   1st Qu.:39.8
Median :16.0    Median :12.00   Median :36.00   Median :39.8
Mean   :16.1    Mean   :11.66   Mean   :34.38   Mean   :39.8
3rd Qu.:24.0    3rd Qu.:18.00   3rd Qu.:53.00   3rd Qu.:39.8
Max.   :31.0    Max.   :23.00   Max.   :59.00   Max.   :39.8
NA's   :60271   NA's   :60271   NA's   :60271   NA's   :60271
     lon              elev          wind.dir     wind.dir.qc
Min.   :-105.8   Min.   :4113    Min.   : 10.0   Length:62388
1st Qu.:-105.8   1st Qu.:4113    1st Qu.:250.0   Class :character
Median :-105.8   Median :4113    Median :300.0   Mode  :character
Mean   :-105.8   Mean   :4113    Mean   :261.5
3rd Qu.:-105.8   3rd Qu.:4113    3rd Qu.:310.0
Max.   :-105.8   Max.   :4113    Max.   :360.0
NA's   :60271    NA's   :60271   NA's   :60508
wind.type.code        wind.sp       wind.sp.qc          ceiling.ht
Length:62388       Min.   : 0.00   Length:62388       Min.   :   30
Class :character   1st Qu.: 4.10   Class :character   1st Qu.: 2591
Mode  :character   Median : 6.70   Mode  :character   Median :22000
                   Mean   : 7.24                      Mean   :15145
                   3rd Qu.: 9.80                      3rd Qu.:22000
                   Max.   :21.10                      Max.   :22000
                   NA's   :60439                      NA's   :60275
ceiling.ht.qc   ceiling.ht.method    sky.cond           vis.dist
Min.   :5.00    Length:62388       Length:62388       Min.   :    0
1st Qu.:5.00    Class :character   Class :character   1st Qu.:16093
Median :5.00    Mode  :character   Mode  :character   Median :16093
Mean   :5.01                                          Mean   :15913
3rd Qu.:5.00                                          3rd Qu.:16093
Max.   :9.00                                          Max.   :16093
NA's   :60271                                         NA's   :60954
vis.dist.qc          vis.var          vis.var.qc             temp
Length:62388       Length:62388       Length:62388       Min.   : 1.00
Class :character   Class :character   Class :character   1st Qu.: 6.00
Mode  :character   Mode  :character   Mode  :character   Median : 8.00
```

```
                                                          Mean   : 8.13  
                                                          3rd Qu.:10.00  
                                                          Max.   :15.00  
                                                          NA's   :60271  
   temp.qc            dew.point      dew.point.qc        atm.press     
 Length:62388       Min.   :-6.00   Length:62388       Min.   : NA    
 Class :character   1st Qu.: 0.00   Class :character   1st Qu.: NA    
 Mode  :character   Median : 0.00   Mode  :character   Median : NA    
                    Mean   : 0.87                      Mean   :NaN    
                    3rd Qu.: 2.00                      3rd Qu.: NA    
                    Max.   : 7.00                      Max.   : NA    
                    NA's   :60271                      NA's   :62388  
  atm.press.qc          rh       
 Min.   :9       Min.   :53.63  
 1st Qu.:9       1st Qu.:58.10  
 Median :9       Median :61.39  
 Mean   :9       Mean   :60.62  
 3rd Qu.:9       3rd Qu.:61.85  
 Max.   :9       Max.   :70.01  
 NA's   :60271   NA's   :60271  
```

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

```
[1] -0.09373843
```

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

```
[1] 0.4397261
```

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

```
[1] 0.3643079
```

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

```
[1] 0.08807315
```

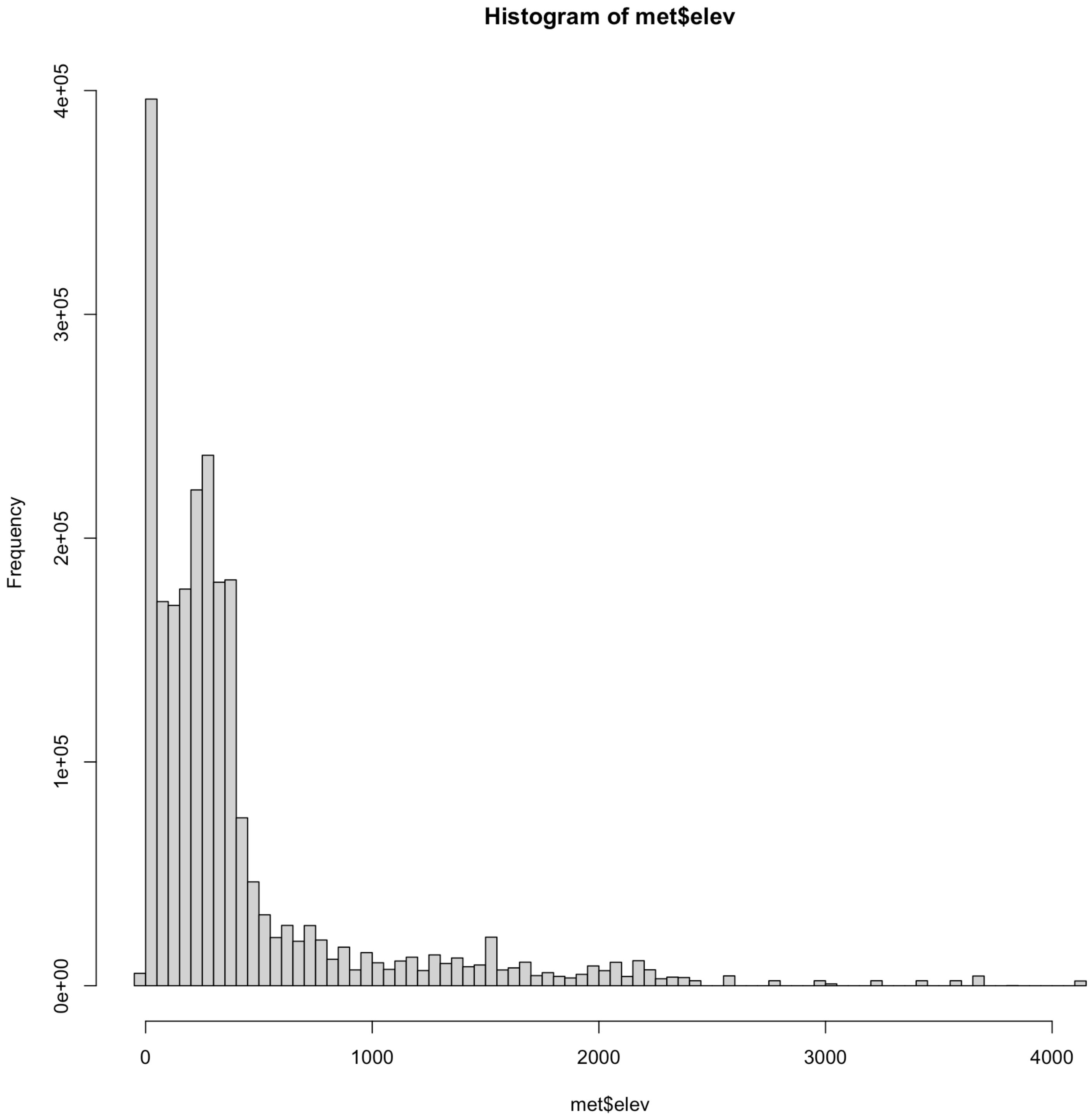
```r
cor(elev$temp, elev$day, use="complete.obs")
```

```
[1] -0.003857766
```

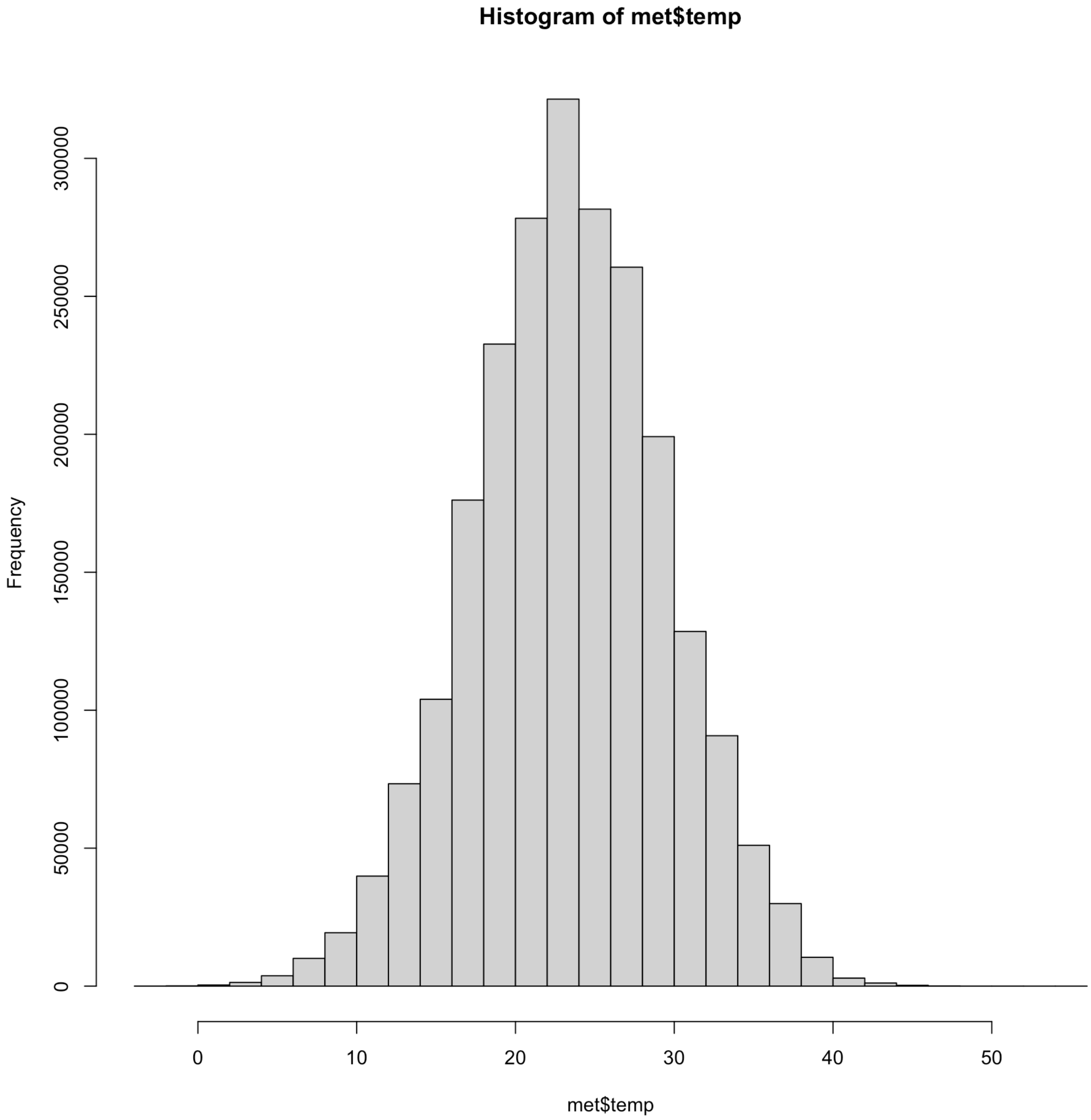
Correlations: Temperature and wind speed have a very weak inverse relationship Temperature and hours have a moderately positive correlation Wind speed and day also have a moderately positive correlation Wind speed and day have a very weak relationship Teperature and day also have no meaningful linear relationship

Part 3: Exploratory Graphs

```r
hist(met$elev, breaks=100)
```

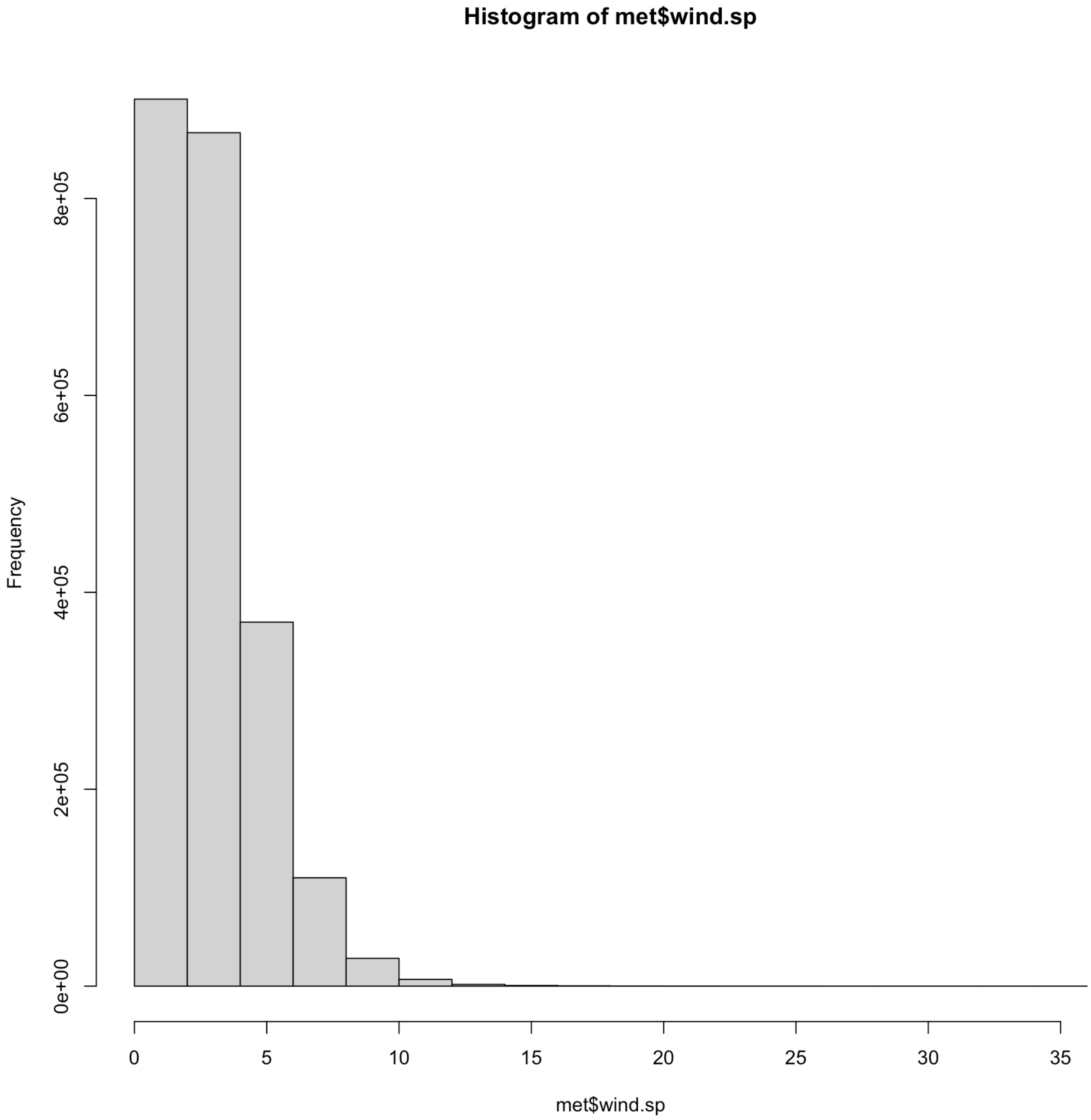

Histogram of met$elev
Frequency
0e+00
1e+05
2e+05
3e+05
4e+05
0
1000
2000
3000
4000
met$elev

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



Histogram of met$temp
Frequency
0
50000
100000
150000
200000
250000
300000
0
10
20
30
40
50
met$temp

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


Histogram of met$wind.sp
Frequency
0e+00
2e+05
4e+05
6e+05
8e+05
0
5
10
15
20
25
30
35
met$wind.sp

Elevation is very skewed to the right, most stations are at low to moderate altitudes. Temperature (after cleaning) shows a normal distribution Windspeed is also right skewed, and many light-wind hours are observed.

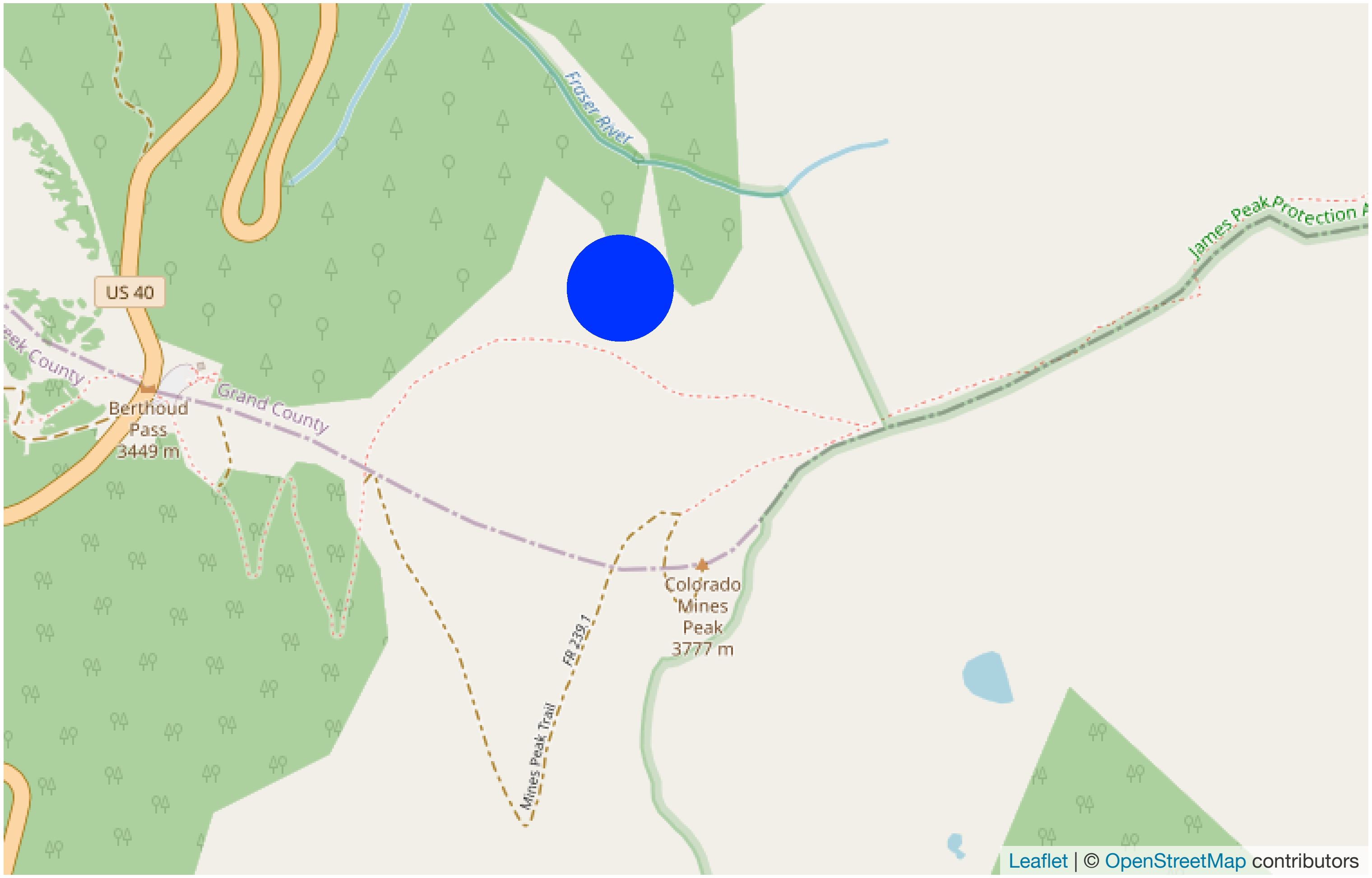
Where is the station with the highest elevation located?

```r
library(leaflet)
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.2     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.2     
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
errors
```

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

```
Warning in validateCoords(lng, lat, funcName): Data contains 60271 rows with
either missing or invalid lat/lon values and will be ignored
```



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

```
Warning: 60271 failed to parse.
```

```
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"
                      NA's
                   "60271"
```

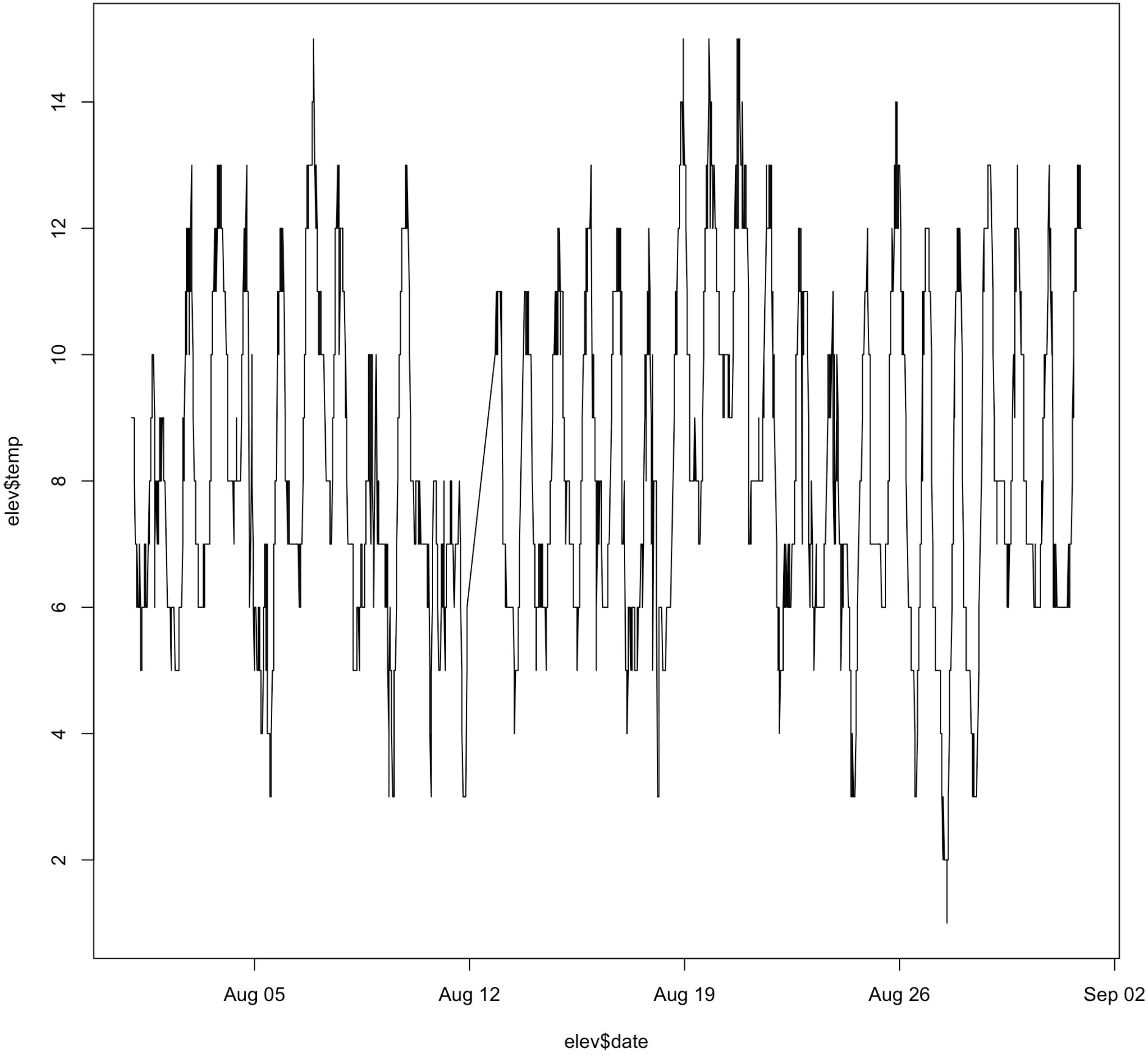
```r
elev <- elev[order(date)]
head(elev)
```

```
   USAFID  WBAN  year month   day  hour   min   lat       lon  elev wind.dir
    <int> <int> <int> <int> <int> <int> <int> <num>     <num> <int>    <int>
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
6: 720385   419  2019     8     1     2    12  39.8 -105.766  4113      120
   wind.dir.qc wind.type.code wind.sp wind.sp.qc ceiling.ht ceiling.ht.qc
        <char>         <char>   <num>     <char>      <int>         <int>
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
6:           5              N     6.2          5      22000             5
   ceiling.ht.method sky.cond vis.dist vis.dist.qc vis.var vis.var.qc  temp
               <char>   <char>    <int>      <char>  <char>     <char> <num>
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
6:                 9        N       NA           9       N          5     9
   temp.qc dew.point dew.point.qc atm.press atm.press.qc       rh
    <char>     <num>       <char>     <num>        <int>    <num>
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
```

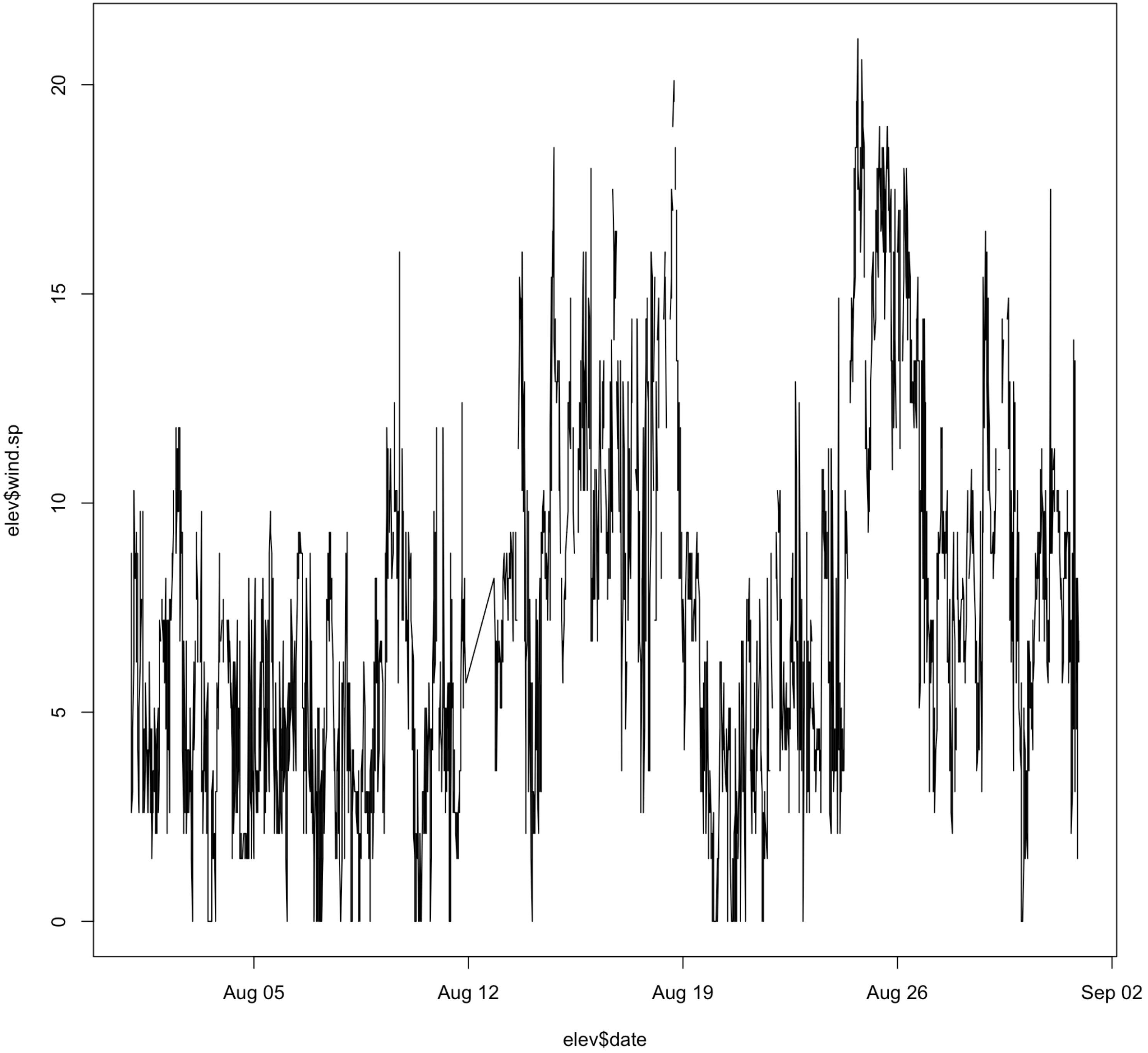
```
6:       5         2            5        NA            9 61.85243
                  date
                <POSc>
1: 2019-08-01 00:00:00
2: 2019-08-01 00:00:00
3: 2019-08-01 01:00:00
4: 2019-08-01 01:00:00
5: 2019-08-01 01:00:00
6: 2019-08-01 02:00:00
```

Now with the date-time variable, we plot the time series of temperature and wind speed

```
plot(elev$date, elev$temp, type='l')
```


elev$temp
2
4
6
8
10
12
14
Aug 05
Aug 12
Aug 19
Aug 26
Sep 02
elev$date

```r
plot(elev$date, elev$wind.sp, type='l')
```

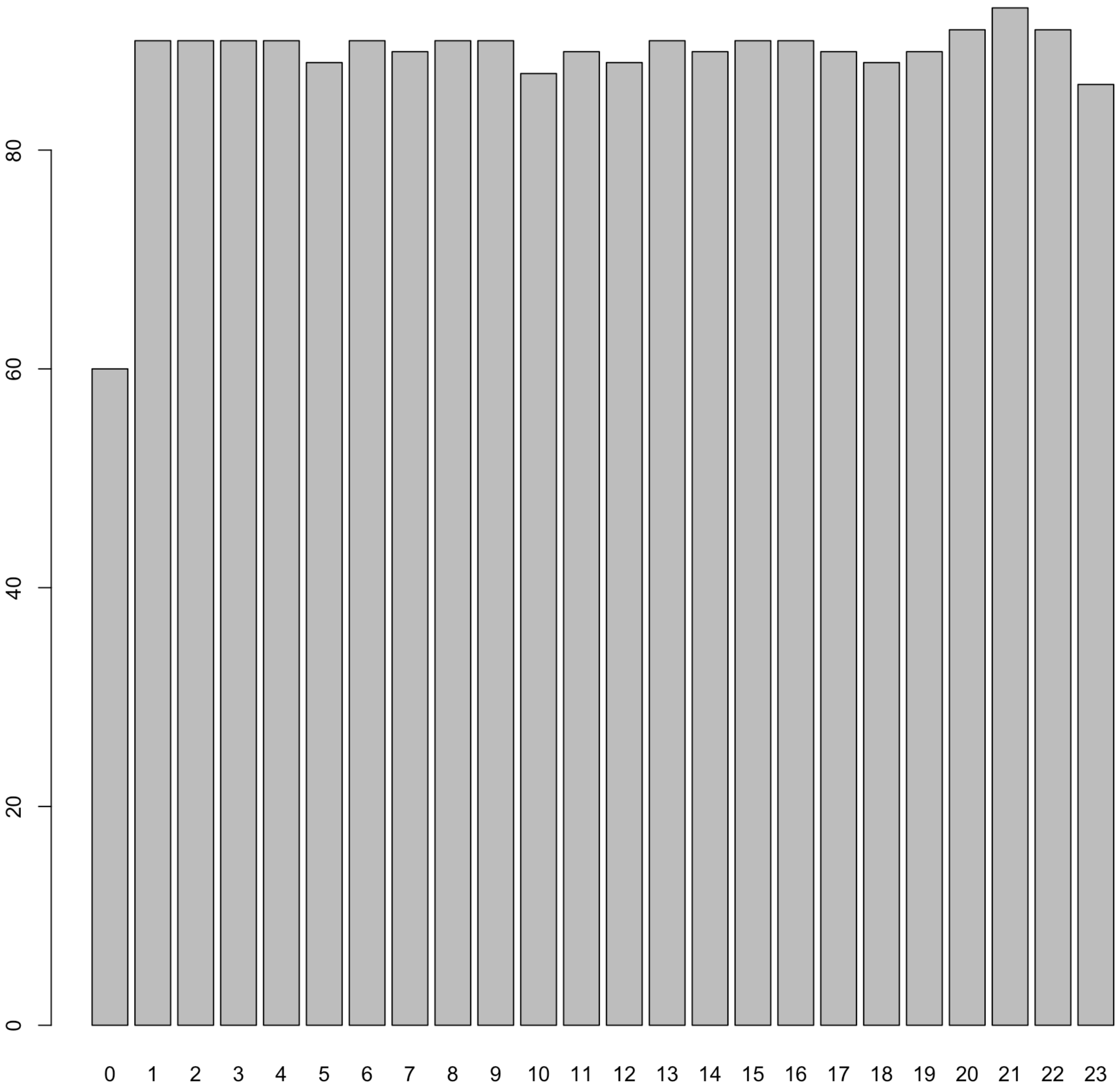

elev$wind.sp
0
5
10
15
20
Aug 05
Aug 12
Aug 19
Aug 26
Sep 02
elev$date

To summarize this data as visualized in the above plots:

I see that the station with max elevation has a pretty cyclical temperature fluctuation every day, but there seems to be a very noticeable spike in wind speeds mid-August, possibly as the Fall season sets in.

A question that I would like to ask, and build an exploratory plot for: Which hours have the most data? For this, I could plot the frequency by hour, and use a barplot to visualize-

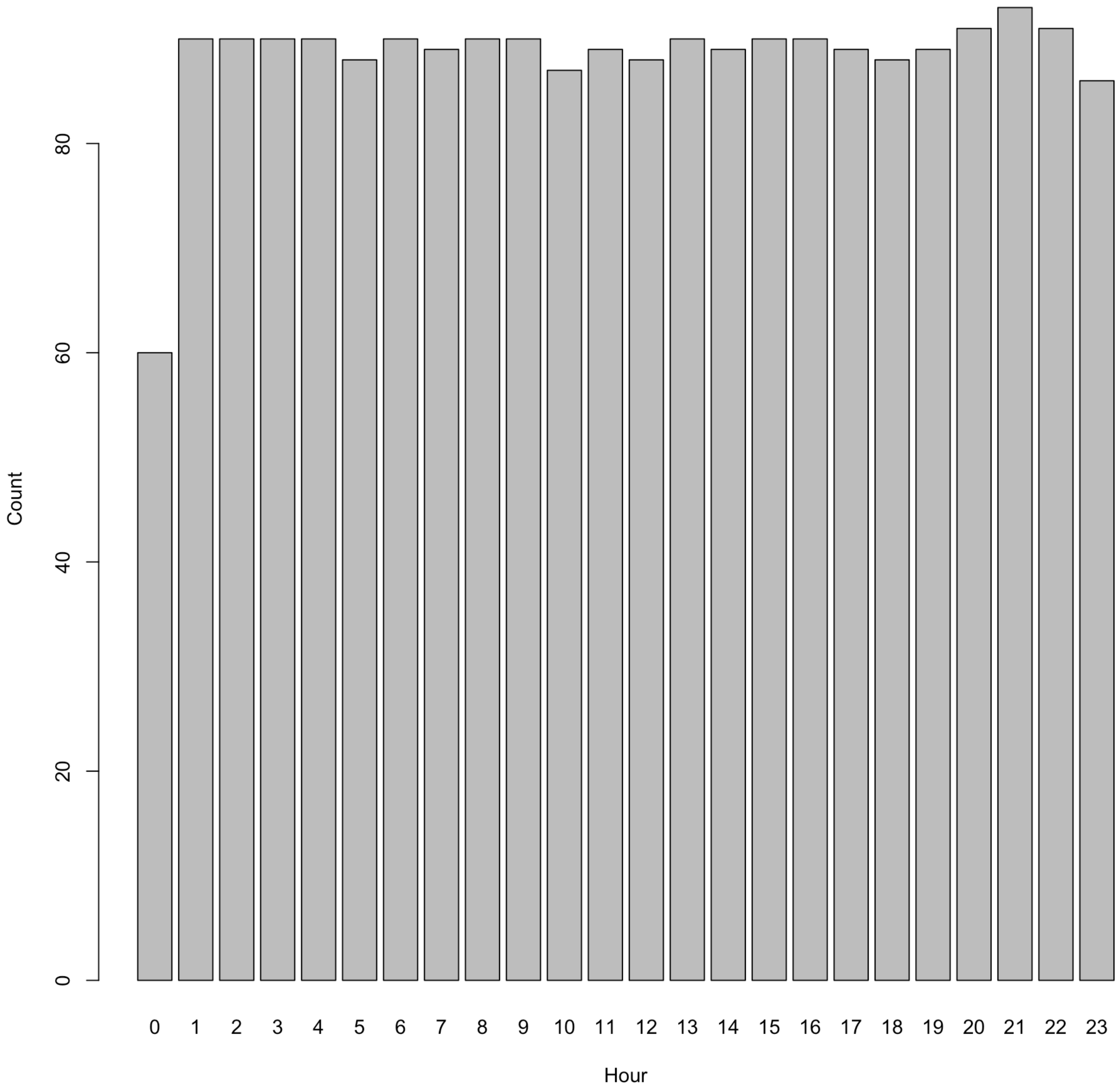
```
barplot(table(elev$hour))
```



0
20
40
60
80
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

Looking at ?barplot, I see that I can use xlab and ylab to label the x and y axes respectively.

```r
barplot(table(elev$hour), xlab="Hour", ylab="Count")
```


Count
0
20
40
60
80
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour

And it seems like the data is well distributed through the hours.