

# HW04

*Oksana Ivanova*

*13 04 2019*

## Exploring dataset

```
set.seed(42)
raw <- readRDS('weather.rds')
head(raw)
```

```
##      X year month      measure X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12
## 1 1 2014      12 Max.TemperatureF 64 42 51 43 42 45 38 29 49 48 39 39
## 2 2 2014      12 Mean.TemperatureF 52 38 44 37 34 42 30 24 39 43 36 35
## 3 3 2014      12 Min.TemperatureF 39 33 37 30 26 38 21 18 29 38 32 31
## 4 4 2014      12   Max.Dew.PointF 46 40 49 24 37 45 36 28 49 45 37 28
## 5 5 2014      12   MeanDew.PointF 40 27 42 21 25 40 20 16 41 39 31 27
## 6 6 2014      12   Min.DewpointF 26 17 24 13 12 36 -3  3 28 37 27 25
##      X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 X29 X30
## 1  42  45  42  44  49  44  37  36  36  44  47  46  59  50  52  52  41  30
## 2  37  39  37  40  45  40  33  32  33  39  45  44  52  44  45  46  36  26
## 3  32  33  32  35  41  36  29  27  30  33  42  41  44  37  38  40  30  22
## 4  28  29  33  42  46  34  25  30  30  39  45  46  58  31  34  42  26  10
## 5  26  27  29  36  41  30  22  24  27  34  42  44  43  29  31  35  20  4
## 6  24  25  27  30  32  26  20  20  25  25  37  41  29  28  29  27  10 -6
##      X31
## 1  30
## 2  25
## 3  20
## 4   8
## 5   5
## 6   1
```

```
summary(raw)
```

```
##           X           year           month           measure
## Min.      : 1.00    Min.    :2014    Min.      : 1.000    Length:286
## 1st Qu.: 72.25    1st Qu.:2015    1st Qu.: 4.000    Class :character
## Median :143.50    Median :2015    Median : 7.000    Mode  :character
## Mean     :143.50    Mean     :2015    Mean     : 6.923
## 3rd Qu.:214.75    3rd Qu.:2015    3rd Qu.:10.000
## Max.     :286.00    Max.     :2015    Max.     :12.000
##           X1           X2           X3
## Length:286    Length:286    Length:286
## Class :character    Class :character    Class :character
## Mode  :character    Mode  :character    Mode  :character
##
##
##           X4           X5           X6
## Length:286    Length:286    Length:286
## Class :character    Class :character    Class :character
```

```

## Mode :character Mode :character Mode :character
##
##
##
##      X7          X8          X9
## Length:286      Length:286      Length:286
## Class :character Class :character Class :character
## Mode :character Mode :character Mode :character
##
##
##
##      X10         X11         X12
## Length:286      Length:286      Length:286
## Class :character Class :character Class :character
## Mode :character Mode :character Mode :character
##
##
##
##      X13         X14         X15
## Length:286      Length:286      Length:286
## Class :character Class :character Class :character
## Mode :character Mode :character Mode :character
##
##
##
##      X16         X17         X18
## Length:286      Length:286      Length:286
## Class :character Class :character Class :character
## Mode :character Mode :character Mode :character
##
##
##
##      X19         X20         X21
## Length:286      Length:286      Length:286
## Class :character Class :character Class :character
## Mode :character Mode :character Mode :character
##
##
##
##      X22         X23         X24
## Length:286      Length:286      Length:286
## Class :character Class :character Class :character
## Mode :character Mode :character Mode :character
##
##
##
##      X25         X26         X27
## Length:286      Length:286      Length:286
## Class :character Class :character Class :character
## Mode :character Mode :character Mode :character
##
##
##
##      X28         X29         X30

```

```
## Length:286      Length:286      Length:286
## Class :character Class :character Class :character
## Mode :character Mode :character Mode :character
##
##
##
##      X31
## Length:286
## Class :character
## Mode :character
##
##
##
```

```
str(raw)
```

```
## 'data.frame':   286 obs. of  35 variables:
## $ X      : int  1 2 3 4 5 6 7 8 9 10 ...
## $ year   : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
## $ month  : int  12 12 12 12 12 12 12 12 12 12 ...
## $ measure: chr   "Max.TemperatureF" "Mean.TemperatureF" "Min.TemperatureF" "Max.Dew.PointF" ...
## $ X1     : chr   "64" "52" "39" "46" ...
## $ X2     : chr   "42" "38" "33" "40" ...
## $ X3     : chr   "51" "44" "37" "49" ...
## $ X4     : chr   "43" "37" "30" "24" ...
## $ X5     : chr   "42" "34" "26" "37" ...
## $ X6     : chr   "45" "42" "38" "45" ...
## $ X7     : chr   "38" "30" "21" "36" ...
## $ X8     : chr   "29" "24" "18" "28" ...
## $ X9     : chr   "49" "39" "29" "49" ...
## $ X10    : chr   "48" "43" "38" "45" ...
## $ X11    : chr   "39" "36" "32" "37" ...
## $ X12    : chr   "39" "35" "31" "28" ...
## $ X13    : chr   "42" "37" "32" "28" ...
## $ X14    : chr   "45" "39" "33" "29" ...
## $ X15    : chr   "42" "37" "32" "33" ...
## $ X16    : chr   "44" "40" "35" "42" ...
## $ X17    : chr   "49" "45" "41" "46" ...
## $ X18    : chr   "44" "40" "36" "34" ...
## $ X19    : chr   "37" "33" "29" "25" ...
## $ X20    : chr   "36" "32" "27" "30" ...
## $ X21    : chr   "36" "33" "30" "30" ...
## $ X22    : chr   "44" "39" "33" "39" ...
## $ X23    : chr   "47" "45" "42" "45" ...
## $ X24    : chr   "46" "44" "41" "46" ...
## $ X25    : chr   "59" "52" "44" "58" ...
## $ X26    : chr   "50" "44" "37" "31" ...
## $ X27    : chr   "52" "45" "38" "34" ...
## $ X28    : chr   "52" "46" "40" "42" ...
## $ X29    : chr   "41" "36" "30" "26" ...
## $ X30    : chr   "30" "26" "22" "10" ...
## $ X31    : chr   "30" "25" "20" "8" ...
```

## Problems of the dataset:

- 1) "X" column is just an index which is not needed for further analysis - have to be deleted.
- 2) Split [measure] column into multiple columns
- 3) It is better to assign our samples as [Year+month+day] from X number of columns.

```
# Packages included
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 3.5.2
```

```
# Get rid of [X] column
```

```
tidy <- raw[,-1]
```

```
head(tidy)
```

```
##   year month      measure X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13
## 1 2014    12 Max.TemperatureF 64 42 51 43 42 45 38 29 49 48 39 39 42
## 2 2014    12 Mean.TemperatureF 52 38 44 37 34 42 30 24 39 43 36 35 37
## 3 2014    12 Min.TemperatureF 39 33 37 30 26 38 21 18 29 38 32 31 32
## 4 2014    12   Max.Dew.PointF 46 40 49 24 37 45 36 28 49 45 37 28 28
## 5 2014    12   MeanDew.PointF 40 27 42 21 25 40 20 16 41 39 31 27 26
## 6 2014    12   Min.DewpointF 26 17 24 13 12 36 -3  3 28 37 27 25 24
##   X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 X29 X30 X31
## 1  45  42  44  49  44  37  36  36  44  47  46  59  50  52  52  41  30  30
## 2  39  37  40  45  40  33  32  33  39  45  44  52  44  45  46  36  26  25
## 3  33  32  35  41  36  29  27  30  33  42  41  44  37  38  40  30  22  20
## 4  29  33  42  46  34  25  30  30  39  45  46  58  31  34  42  26  10   8
## 5  27  29  36  41  30  22  24  27  34  42  44  43  29  31  35  20   4   5
## 6  25  27  30  32  26  20  20  25  25  37  41  29  28  29  27  10  -6   1
```

```
# Make dataframe from wide to long
```

```
tidy <- gather(tidy, day, value, X1:X31)
```

```
# Make dataframe from long to wide
```

```
tidy = spread(tidy, measure, value)
```

```
# Assign [Year+month+day] as samples names
```

```
tidy$day = str_sub(tidy$day, 2, 3)
```

```
tidy = unite(tidy, Date, year, month, day, sep = '_')
```

```
head(tidy)
```

```
##      Date CloudCover   Events Max.Dew.PointF Max.Gust.SpeedMPH
## 1 2014_12_1         6     Rain             46                 29
## 2 2014_12_10        8     Rain             45                 29
## 3 2014_12_11        8 Rain-Snow             37                 28
## 4 2014_12_12        7     Snow             28                 21
## 5 2014_12_13        5             28                 23
## 6 2014_12_14        4             29                 20
##   Max.Humidity Max.Sea.Level.PressureIn Max.TemperatureF
## 1           74                30.45             64
## 2          100                29.58             48
## 3           92                29.81             39
## 4           85                29.88             39
## 5           75                29.86             42
## 6           82                29.91             45
```

```
##      Max.VisibilityMiles Max.Wind.SpeedMPH Mean.Humidity
## 1             10             22             63
## 2             10             23             95
## 3             10             21             87
## 4             10             16             75
## 5             10             17             65
## 6             10             15             68
##      Mean.Sea.Level.PressureIn Mean.TemperatureF Mean.VisibilityMiles
## 1             30.13             52             10
## 2             29.5             43             3
## 3             29.61             36             7
## 4             29.85             35             10
## 5             29.82             37             10
## 6             29.83             39             10
##      Mean.Wind.SpeedMPH MeanDew.PointF Min.DewpointF Min.Humidity
## 1             13             40             26             52
## 2             13             39             37             89
## 3             13             31             27             82
## 4             11             27             25             64
## 5             12             26             24             55
## 6             10             27             25             53
##      Min.Sea.Level.PressureIn Min.TemperatureF Min.VisibilityMiles
## 1             30.01             39             10
## 2             29.43             38             1
## 3             29.44             32             1
## 4             29.81             31             7
## 5             29.78             32             10
## 6             29.78             33             10
##      PrecipitationIn WindDirDegrees
## 1             0.01             268
## 2             0.28             357
## 3             0.02             230
## 4             T             286
## 5             T             298
## 6             0.00             306
```

- 4) Then it was noticed that some samples have all NAs data because of unexisting days (February 30, etc). Also, the measurements in the end of the table are NAs as well.

```
tidy = na.omit(tidy)
```

- 5) Formatting of the numeric columns:

```
tidy$PrecipitationIn = as.numeric(tidy$PrecipitationIn, na.string='T')
```

```
## Warning: NAs introduced by coercion
```

```
# Make columns with values as numbers numeric
tidy[,c(2,4:23)] = lapply(tidy[,c(2,4:23)], as.numeric)
# Check the dataframe after formatting
row.names(tidy) = 1:nrow(tidy)
summary(tidy)
```

```
##      Date          CloudCover      Events      Max.Dew.PointF
## Length:360      Min.    :0.000  Length:360      Min.    : -6.00
## Class :character 1st Qu.:3.000  Class :character 1st Qu.:31.00
## Mode  :character Median :5.000  Mode  :character Median :47.00
```

```

##           Mean    :4.733           Mean    :45.33
##           3rd Qu.:7.000           3rd Qu.:61.00
##           Max.    :8.000           Max.    :75.00
##
## Max.Gust.SpeedMPH Max.Humidity    Max.Sea.Level.PressureIn
## Min.    : 0.00    Min.    : 39.00    Min.    :29.58
## 1st Qu.:21.00    1st Qu.: 73.00    1st Qu.:30.00
## Median :25.50    Median : 86.00    Median :30.14
## Mean    :26.99    Mean    : 85.64    Mean    :30.16
## 3rd Qu.:31.25    3rd Qu.: 93.00    3rd Qu.:30.31
## Max.    :94.00    Max.    :1000.00    Max.    :30.88
##
## Max.TemperatureF Max.VisibilityMiles Max.Wind.SpeedMPH Mean.Humidity
## Min.    :18.00    Min.    : 2.000    Min.    : 8.00    Min.    :28.00
## 1st Qu.:42.00    1st Qu.:10.000    1st Qu.:16.00    1st Qu.:56.00
## Median :60.00    Median :10.000    Median :20.00    Median :66.00
## Mean    :58.75    Mean    : 9.906    Mean    :20.71    Mean    :65.97
## 3rd Qu.:76.00    3rd Qu.:10.000    3rd Qu.:24.00    3rd Qu.:76.25
## Max.    :96.00    Max.    :10.000    Max.    :38.00    Max.    :98.00
##
## Mean.Sea.Level.PressureIn Mean.TemperatureF Mean.VisibilityMiles
## Min.    :29.49    Min.    : 8.00    Min.    : -1.000
## 1st Qu.:29.87    1st Qu.:36.00    1st Qu.: 8.000
## Median :30.03    Median :53.00    Median :10.000
## Mean    :30.04    Mean    :51.23    Mean    : 8.847
## 3rd Qu.:30.19    3rd Qu.:68.00    3rd Qu.:10.000
## Max.    :30.77    Max.    :84.00    Max.    :10.000
##
## Mean.Wind.SpeedMPH MeanDew.PointF    Min.DewpointF    Min.Humidity
## Min.    : 4.00    Min.    : -11.00    Min.    : -18.00    Min.    :16.00
## 1st Qu.: 8.00    1st Qu.: 24.00    1st Qu.: 16.00    1st Qu.:35.00
## Median :10.00    Median : 40.50    Median : 35.00    Median :46.00
## Mean    :10.72    Mean    : 38.76    Mean    : 31.99    Mean    :48.28
## 3rd Qu.:13.00    3rd Qu.: 56.00    3rd Qu.: 51.00    3rd Qu.:60.25
## Max.    :22.00    Max.    : 71.00    Max.    : 68.00    Max.    :96.00
##
## Min.Sea.Level.PressureIn Min.TemperatureF Min.VisibilityMiles
## Min.    :29.16    Min.    : -3.00    Min.    : 0.000
## 1st Qu.:29.76    1st Qu.:30.00    1st Qu.: 2.000
## Median :29.94    Median :45.00    Median :10.000
## Mean    :29.92    Mean    :43.15    Mean    : 6.689
## 3rd Qu.:30.09    3rd Qu.:60.00    3rd Qu.:10.000
## Max.    :30.64    Max.    :74.00    Max.    :10.000
##
## PrecipitationIn WindDirDegrees
## Min.    :0.0000    Min.    : 1.0
## 1st Qu.:0.0000    1st Qu.:114.0
## Median :0.0000    Median :223.0
## Mean    :0.1195    Mean    :201.4
## 3rd Qu.:0.0700    3rd Qu.:275.5
## Max.    :2.9000    Max.    :360.0
## NA's    :49

```

6) Make [Events] column as factor

```

tidy$Events = as.factor(tidy$Events)
levels(tidy$Events)[1] = "Ordinary day"
head(tidy)

```

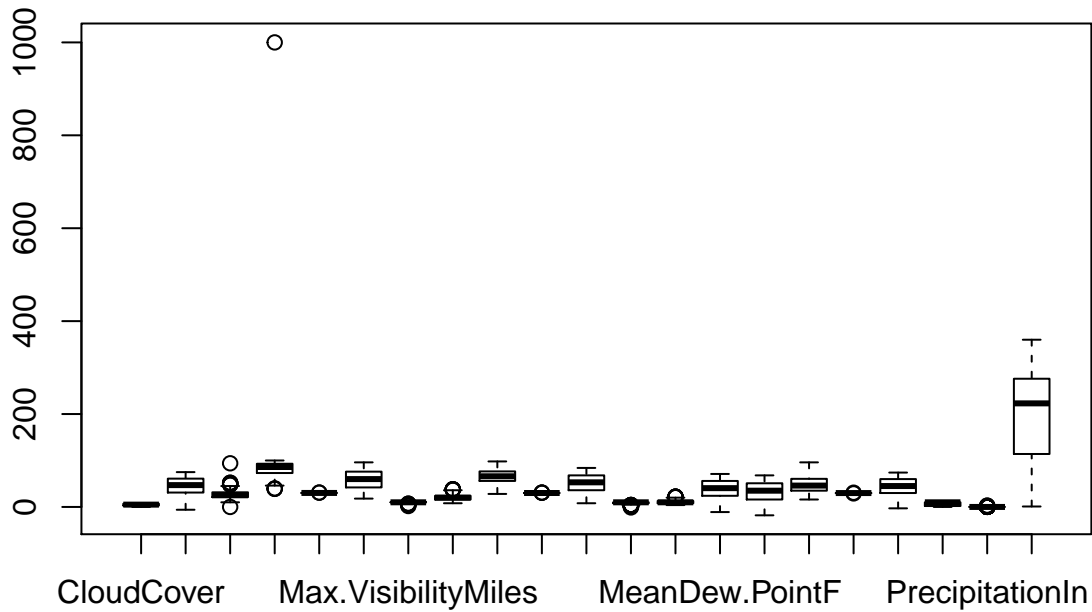
```

##      Date CloudCover      Events Max.Dew.PointF Max.Gust.SpeedMPH
## 1  2014_12_1         6      Rain             46                29
## 2  2014_12_10        8      Rain             45                29
## 3  2014_12_11        8  Rain-Snow             37                28
## 4  2014_12_12        7      Snow             28                21
## 5  2014_12_13        5 Ordinary day             28                23
## 6  2014_12_14        4 Ordinary day             29                20
##      Max.Humidity Max.Sea.Level.PressureIn Max.TemperatureF
## 1             74                30.45                64
## 2            100                29.58                48
## 3             92                29.81                39
## 4             85                29.88                39
## 5             75                29.86                42
## 6             82                29.91                45
##      Max.VisibilityMiles Max.Wind.SpeedMPH Mean.Humidity
## 1                  10                22                63
## 2                  10                23                95
## 3                  10                21                87
## 4                  10                16                75
## 5                  10                17                65
## 6                  10                15                68
##      Mean.Sea.Level.PressureIn Mean.TemperatureF Mean.VisibilityMiles
## 1                30.13                52                10
## 2                29.50                43                 3
## 3                29.61                36                 7
## 4                29.85                35                10
## 5                29.82                37                10
## 6                29.83                39                10
##      Mean.Wind.SpeedMPH MeanDew.PointF Min.DewpointF Min.Humidity
## 1                  13                40                26                52
## 2                  13                39                37                89
## 3                  13                31                27                82
## 4                  11                27                25                64
## 5                  12                26                24                55
## 6                  10                27                25                53
##      Min.Sea.Level.PressureIn Min.TemperatureF Min.VisibilityMiles
## 1                30.01                39                10
## 2                29.43                38                 1
## 3                29.44                32                 1
## 4                29.81                31                 7
## 5                29.78                32                10
## 6                29.78                33                10
##      PrecipitationIn WindDirDegrees
## 1                0.01                268
## 2                0.28                357
## 3                0.02                230
## 4                 NA                286
## 5                 NA                298
## 6                0.00                306

```

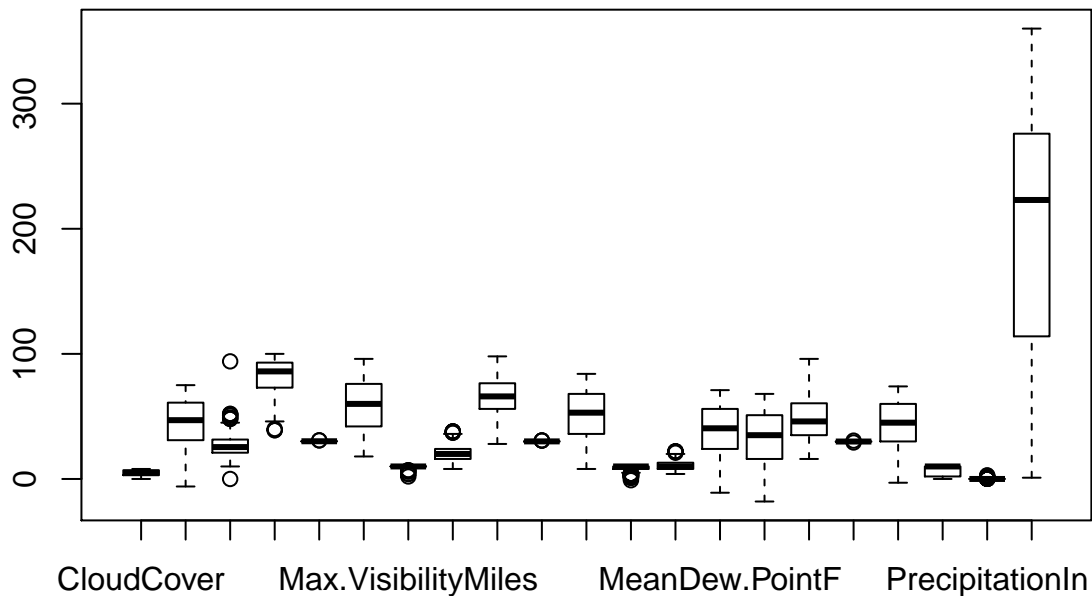
7) Plot the numeric columns data

```
boxplot(tidy[, c(2, 4:23)])
```



8) Obvious outlier was detected - the [Humidity] was 1000% at one point of measurements that seems to be unreal and also [speedMPH] looks like outlier. I decided to change only [Humidity] data for this row for more realistic value = 100%.

```
tidy[135,6] = 100  
# Check the plot after adding  
boxplot(tidy[, c(2, 4:23)])
```





Finally, the dataset seems to be clear and tidy after processing and ready for further analysis.

Difference between “raw” and “tidy” dataframes.

```
head(raw)
```

```
##      X year month          measure X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12
## 1 1 2014      12 Max.TemperatureF 64 42 51 43 42 45 38 29 49 48 39 39
## 2 2 2014      12 Mean.TemperatureF 52 38 44 37 34 42 30 24 39 43 36 35
## 3 3 2014      12 Min.TemperatureF 39 33 37 30 26 38 21 18 29 38 32 31
## 4 4 2014      12   Max.Dew.PointF 46 40 49 24 37 45 36 28 49 45 37 28
## 5 5 2014      12   MeanDew.PointF 40 27 42 21 25 40 20 16 41 39 31 27
## 6 6 2014      12   Min.DewpointF 26 17 24 13 12 36 -3  3 28 37 27 25
##      X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 X29 X30
## 1  42  45  42  44  49  44  37  36  36  44  47  46  59  50  52  52  41  30
## 2  37  39  37  40  45  40  33  32  33  39  45  44  52  44  45  46  36  26
## 3  32  33  32  35  41  36  29  27  30  33  42  41  44  37  38  40  30  22
## 4  28  29  33  42  46  34  25  30  30  39  45  46  58  31  34  42  26  10
## 5  26  27  29  36  41  30  22  24  27  34  42  44  43  29  31  35  20  4
## 6  24  25  27  30  32  26  20  20  25  25  37  41  29  28  29  27  10 -6
##      X31
## 1  30
## 2  25
## 3  20
## 4   8
## 5   5
## 6   1
```

```
head(tidy)
```

```
##      Date CloudCover      Events Max.Dew.PointF Max.Gust.SpeedMPH
## 1 2014_12_1         6        Rain              46                 29
## 2 2014_12_10        8        Rain              45                 29
## 3 2014_12_11        8   Rain-Snow              37                 28
## 4 2014_12_12        7        Snow              28                 21
## 5 2014_12_13        5 Ordinary day              28                 23
## 6 2014_12_14        4 Ordinary day              29                 20
##      Max.Humidity Max.Sea.Level.PressureIn Max.TemperatureF
## 1              74              30.45              64
## 2              100              29.58              48
## 3              92              29.81              39
## 4              85              29.88              39
## 5              75              29.86              42
## 6              82              29.91              45
##      Max.VisibilityMiles Max.Wind.SpeedMPH Mean.Humidity
## 1              10              22              63
## 2              10              23              95
## 3              10              21              87
## 4              10              16              75
## 5              10              17              65
## 6              10              15              68
##      Mean.Sea.Level.PressureIn Mean.TemperatureF Mean.VisibilityMiles
## 1              30.13              52              10
## 2              29.50              43              3
```

## 3	29.61	36	7
## 4	29.85	35	10
## 5	29.82	37	10
## 6	29.83	39	10
##	Mean.Wind.SpeedMPH	MeanDew.PointF	Min.DewpointF Min.Humidity
## 1	13	40	26 52
## 2	13	39	37 89
## 3	13	31	27 82
## 4	11	27	25 64
## 5	12	26	24 55
## 6	10	27	25 53
##	Min.Sea.Level.PressureIn	Min.TemperatureF	Min.VisibilityMiles
## 1	30.01	39	10
## 2	29.43	38	1
## 3	29.44	32	1
## 4	29.81	31	7
## 5	29.78	32	10
## 6	29.78	33	10
##	PrecipitationIn	WindDirDegrees	
## 1	0.01	268	
## 2	0.28	357	
## 3	0.02	230	
## 4	NA	286	
## 5	NA	298	
## 6	0.00	306	