PSS718 - Data Mining

Lecture 4

Asst.Prof.Dr. Burkay Genç

Hacettepe University

October 17, 2016

Why to do?

- Know thy data!
- Improve your understanding
- Shape your data mining process

What can be learned?

- Boundaries of data: min and max values
- Expectations: mean and median values
- Variance: standard deviation and variance
- Distribution: How the data is distributed
- Consistency, Accuracy, Completeness, etc.

Sampling Data

- Often we work with very large datasets
- We can then sample the set to do preliminary analysis

Example

```
> sample(1:100, 20)
```

Example

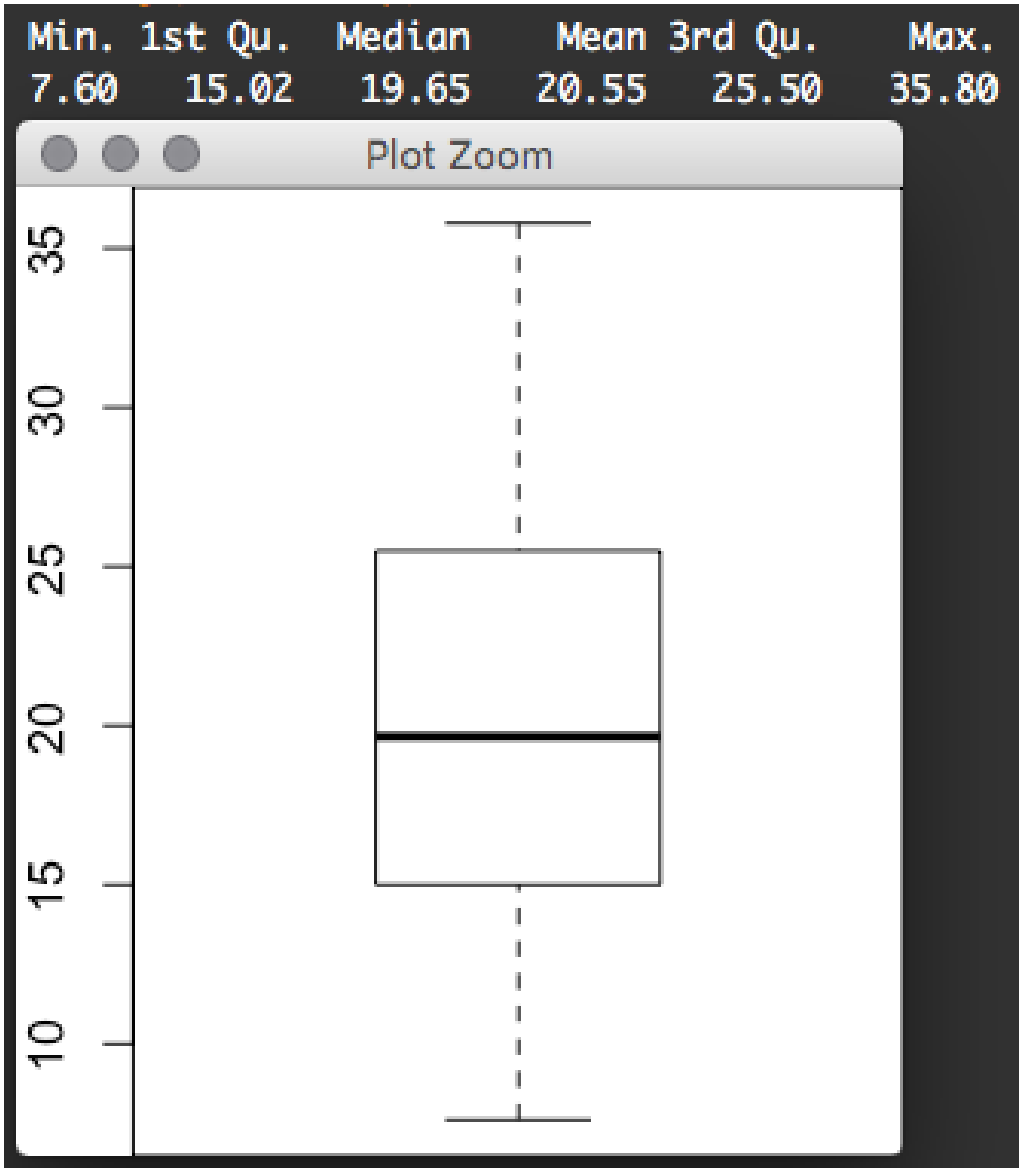
```
> set.seed(42)
> smpl <- sample(nrow(weather), 0.2*nrow(weather))
```

Basic Summaries

```
> summary(weather[7:9])

    Sunshine       WindGustDir   WindGustSpeed
 Min.   : 0.00   NW      : 73   Min.   :13.0
 1st Qu.: 5.95   NNW     : 44   1st Qu.:31.0
 Median : 8.60   E       : 37   Median :39.0
 Mean   : 7.91   WNW     : 35   Mean   :39.8
 3rd Qu.:10.50   ENE     : 30   3rd Qu.:46.0
 Max.   :13.60   (Other):144   Max.   :98.0
 NA's   : 3.00   NA's    :  3   NA's   : 2.0
```

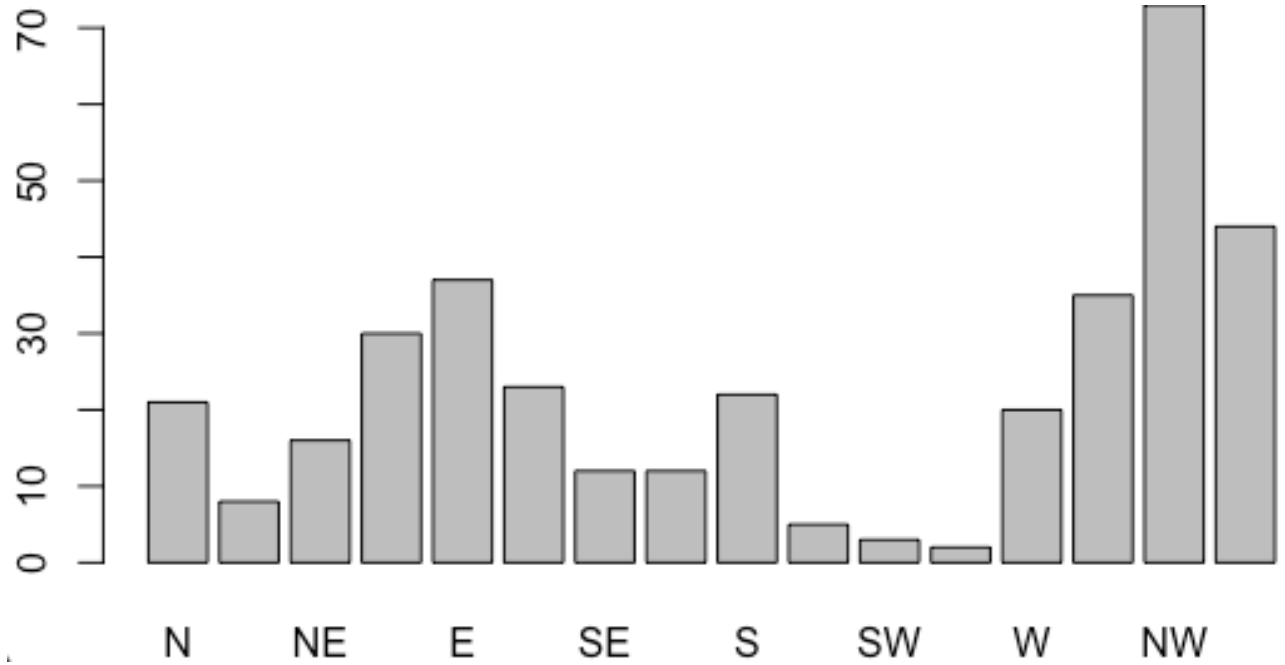
Basic Summaries

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.60   15.02   19.65   20.55   25.50   35.80
```



Categoric Summaries

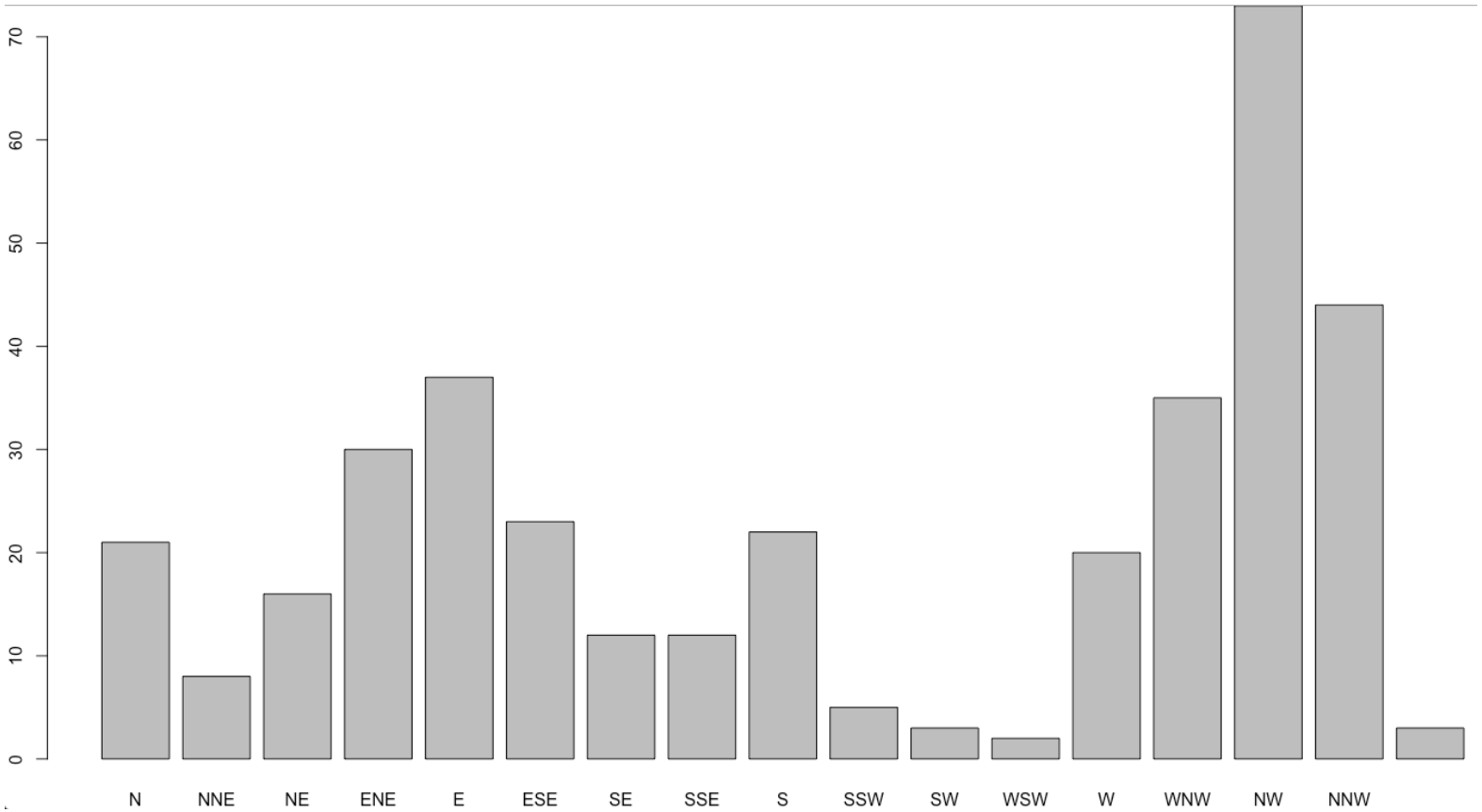
```
> table(df$WindGustDir)

  N NNE  NE ENE   E ESE  SE SSE   S SSW  SW WSW   W WNW  NW NNW
 21   8  16  30  37  23  12  12  22   5   3   2  20  35  73  44
> summary(df$WindGustDir)
   N  NNE   NE  ENE    E  ESE   SE  SSE    S  SSW   SW  WSW    W  WNW   NW  NNW NA's
  21    8   16   30   37   23   12   12   22    5    3    2   20   35   73   44    3
> barplot(table(df$WindGustDir))
```



Barplot Summaries

Example

```
> barplot(table(df$WindGustDir, useNA = "ifany"))
```



More Summaries

```
> library(fBasics)
> basicStats(weather$Sunshine)
```

	X..weather.Sunshine
nobs	366.0000
NAs	3.0000
Minimum	0.0000
Maximum	13.6000
1. Quartile	5.9500
3. Quartile	10.5000
Mean	7.9094
Median	8.6000
Sum	2871.1000
SE Mean	0.1827
LCL Mean	7.5500
UCL Mean	8.2687
Variance	12.1210
Stdev	3.4815
Skewness	-0.7235
Kurtosis	-0.2706

Some Features

Definition (Distribution)

How the values of a variable are distributed along the axis

Example

install.packages("moments")

Definition (Skewness)

Skewness is a measure of how asymmetrically the data is distributed.

- > 1 : Longer right tail
- < -1 : Longer left tail

Definition (Kurtosis)

Tells whether the distribution is skinny or fat.

- High: Sharper peak
- Low: Flatter peak

Missing Values

```
> library(mice)

mice 2.8 2011-03-24

> md.pattern(weather[,7:10])
     WindGustSpeed Sunshine WindGustDir WindDir9am
329              1        1          1          1 0
  3              1        0          1          1 1
  1              1        1          0          1 1
 31              1        1          1          0 1
  2              0        1          0          1 2
                 2        3          3         31 39
```

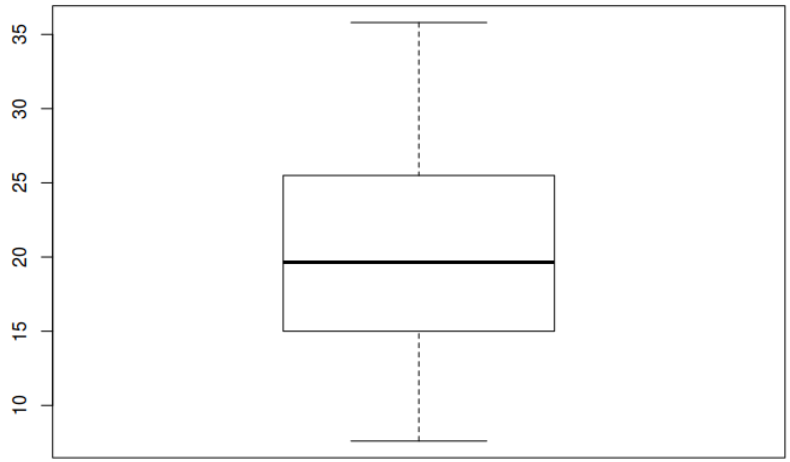
Box Plot

Definition (Box Plot)

A box plot provides a graphical overview of how the observations of a variable are distributed.

Example

```
> boxplot(df$MaxTemp)
```



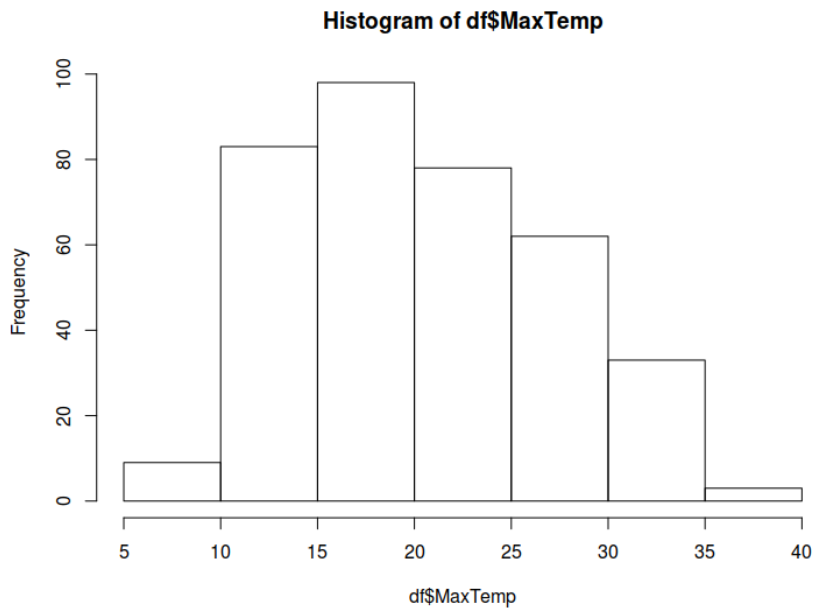
Histogram

Definition (Histogram)

A histogram provides a quick overview of how frequently different values have been observed.

Example

```
> hist(df$MaxTemp)
```



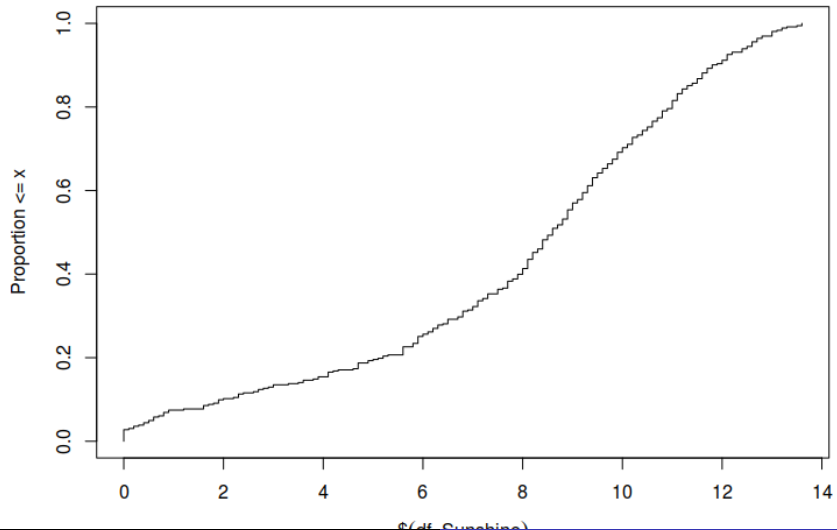
Cumulative Distribution Plot

Definition (Cumulative Distribution)

A cumulative distribution plot displays the proportion of the data that has a value that is less than or equal to the value shown on the x-axis.

Example

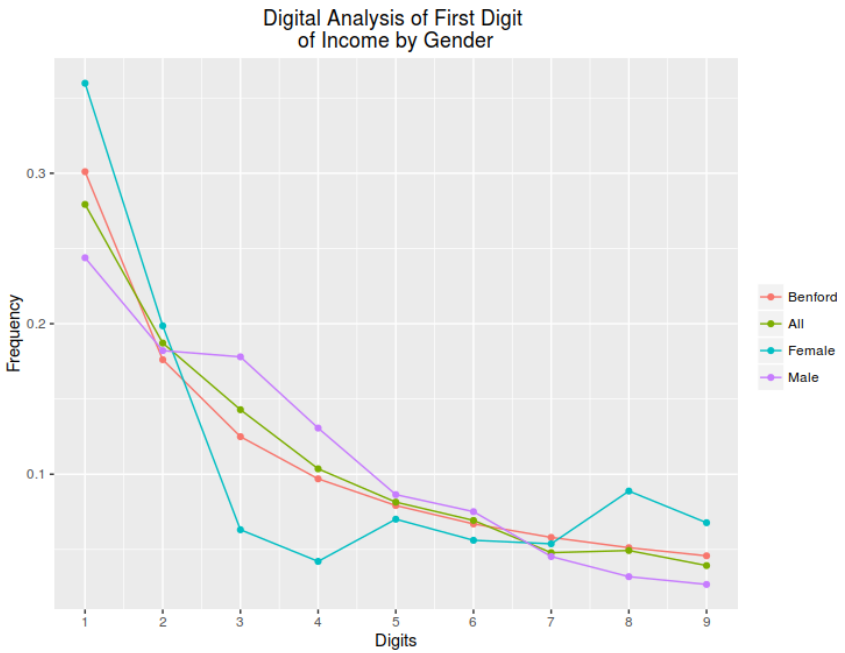
```
> library(Hmisc)
> Ecdf(df$Sunshine)
```



Benford's Law

Definition (Benford's Law)

Benford's law relates to the frequency of occurrence of the first digit in a collection of numbers. The law generally applies when several orders of magnitude (e.g., 10, 100, and 1000) are recorded in the observations.



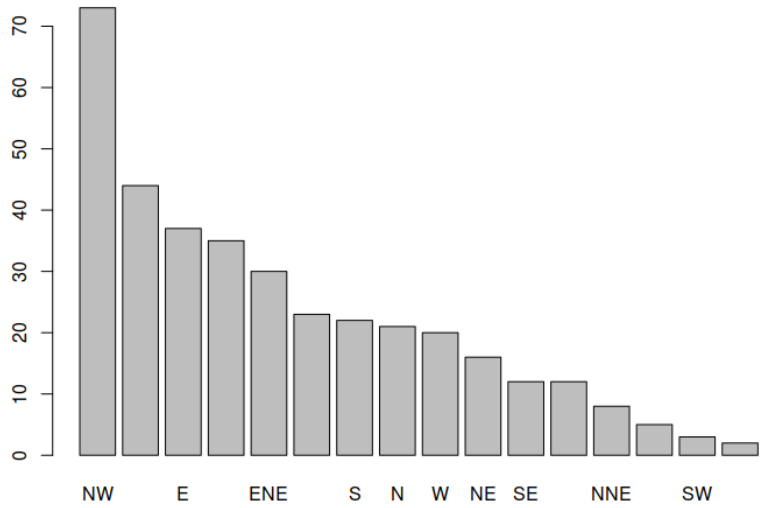
Bar Plot

Definition (Bar Plot)

A bar plot, much like a histogram, uses vertical bars to show counts of the number of observations of each of the possible values of the categoric variable.

Example

```
> barplot(sort(table(df$WindGustDir), decreasing = TRUE))
```



Others

- Dot Plot
- Mosaic Plot
- Pairs and Scatter Plots

What is Correlation?

Definition (Correlation)

A correlation coefficient is a measure of the degree of relationship between two variables - it is usually a number between -1 and 1.

- Magnitude
  - High: The values are closely tied
  - Low: The values are weakly tied (if any)
- Sign
  - Positive: Both increase or decrease at the same time
  - Negative: One decreases while the other one increases

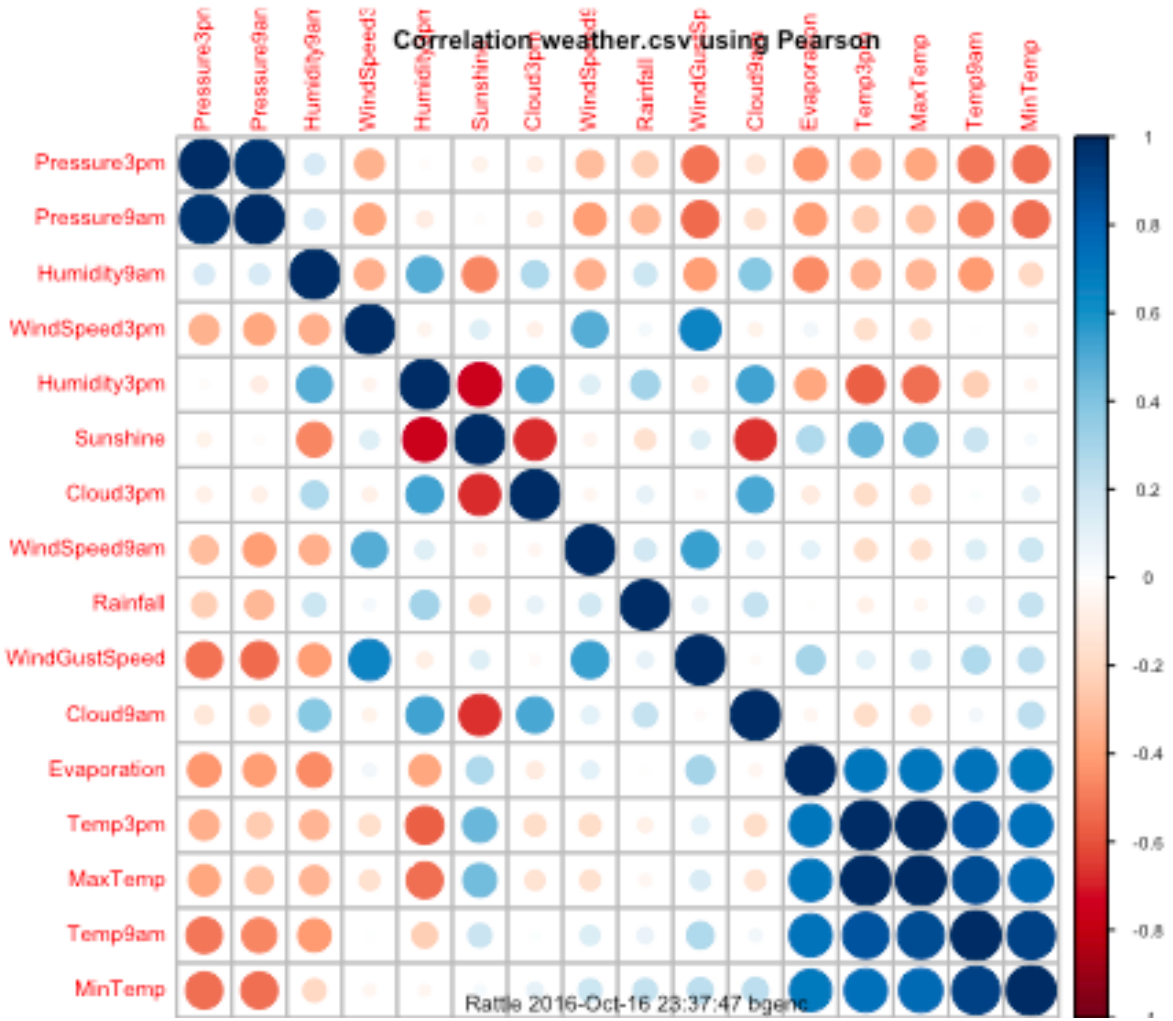
How to?

```
> cor(df[,c(3,4,5,6,7)])
                MinTemp     MaxTemp     Rainfall  Evaporation Sunshine
MinTemp       1.0000000  0.75247079  0.201938716  0.649930175       NA
MaxTemp       0.7524708  1.00000000 -0.073559584  0.690026268       NA
Rainfall      0.2019387 -0.07355958  1.000000000 -0.007292963       NA
Evaporation   0.6499302  0.69002627 -0.007292963  1.000000000       NA
Sunshine             NA          NA           NA           NA        1
```

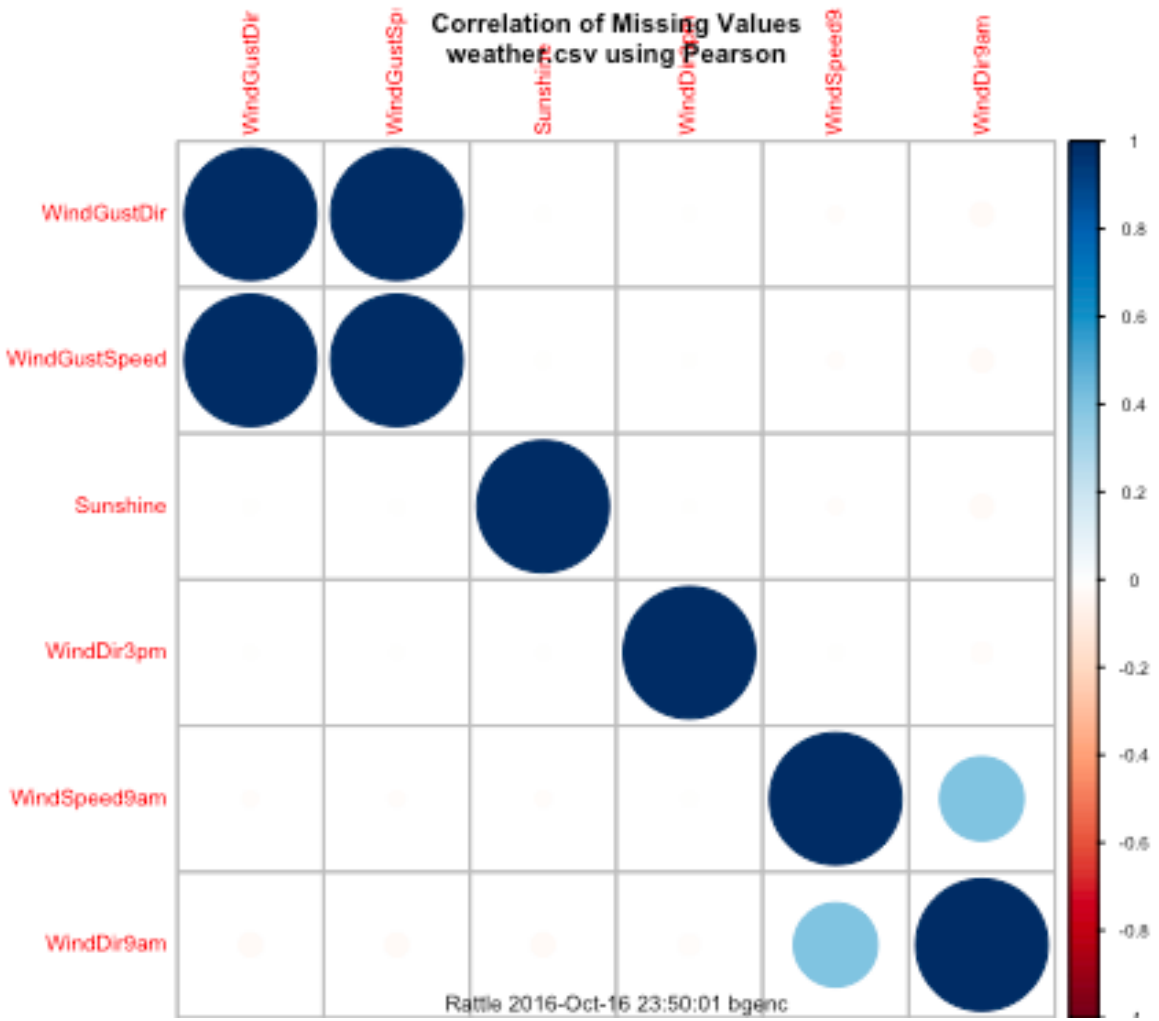
Why do we get NAs for Sunshine?

```
> cor(df[complete.cases(df[,c(3,4,5,6,7)]),c(3,4,5,6,7)])
                MinTemp     MaxTemp     Rainfall Evaporation     Sunshine
MinTemp       1.00000000  0.75331770  0.19764423  0.65110723  0.03571111
MaxTemp       0.75331770  1.00000000 -0.07646213  0.68976877  0.45206352
Rainfall      0.19764423 -0.07646213  1.00000000 -0.01511593 -0.15099036
Evaporation   0.65110723  0.68976877 -0.01511593  1.00000000  0.31802523
Sunshine      0.03571111  0.45206352 -0.15099036  0.31802523  1.00000000
```

Correlation Plots



Missing Values Correlation



Hierarchical Correlation