2017 SDSC Summer Institute
Machine Learning Overview

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UC San Diego

Data Exploration

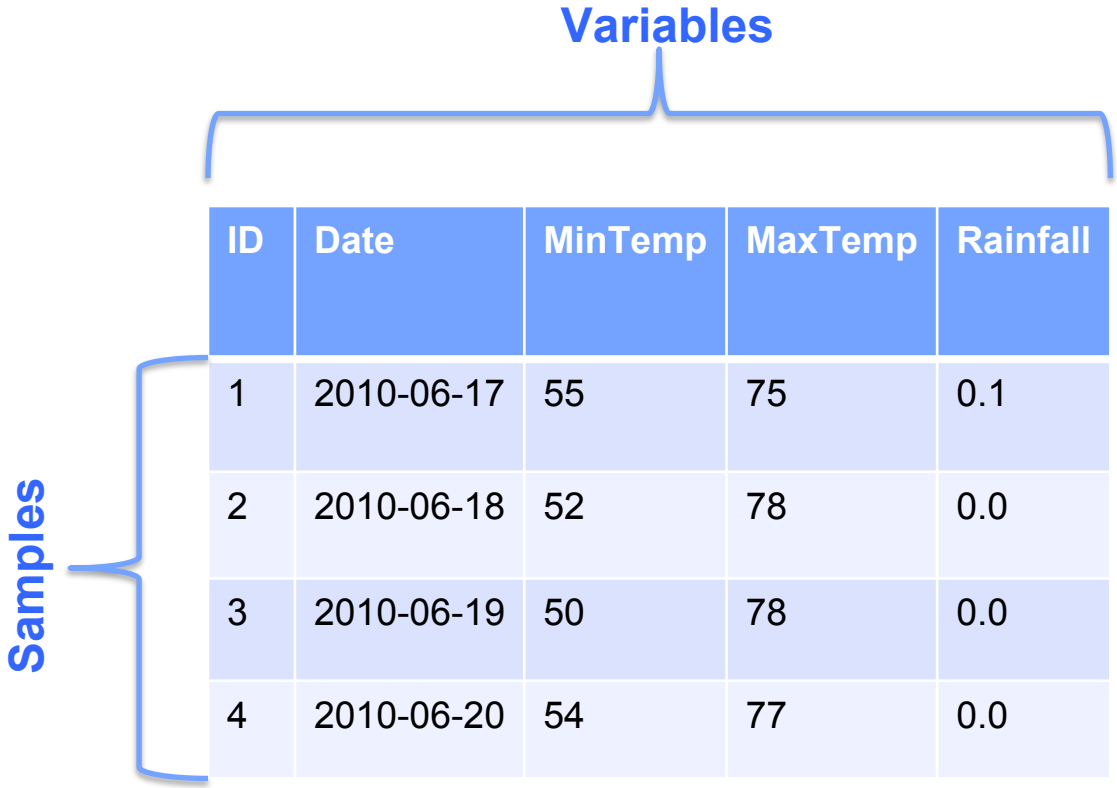
Mai H. Nguyen

Overview

- **Terminology**
  - Structure of data
  - Data types
- **Motivation**
  - Why explore data before modeling?
- **Overview of Dataset**
- **Data Exploration Techniques**
  - Data validation
  - Summary statistics
  - Data visualization
- **Hands-on**
  - Explore weather data

Data Terminology

- **Row = sample**
  Also: observation, instance, record, example, feature vector, etc.
- **Column = variable**
  Also: feature, attribute, field, etc.



ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

Data types

- **Most common**
  - Numeric
  - Categorical
- **Other types are typically engineered to be either numeric or categorical**
  - Strings, ordinals, etc.

Data Types: Numeric

- **A person's height**
  - Positive number
- **Score on an exam**
  - Range between 0 and 100%
- **Number of transactions per hour**
  - Positive integer
- **Change in stock price**
  - Can be positive or negative

Data Types: Categorical

- **Values are just labels or names**

- **Examples:**
  - Color of an item
  - Gender
  - Person's blood type
  - Type of customer
  - Product categories
  - Label for handwritten digits

Data Exploration

- **Definition**
  - Preliminary investigation of your data
- **Purpose**
  - To gain better understanding of specific characteristics of the data
  - Things to look for: Correlations, general trends, outliers, etc.
- **Techniques**
  - Data validation
  - Summary statistics
  - Visualization

Hands-on: Overview of Weather Dataset

- **Description of dataset**
  - Contains 1 year of daily observations from weather station in Canberra, Australia
  - Data source:
    - Australian Commonwealth Bureau of Meteorology
  - Processed to include various useful attributes
  - Attributes described in “data_dictionary_weather.txt”
  - Available in R package ‘rattle’

Weather Data in Detail – 1

Date

The date of observation.

Location

The common name of the location of the weather station.

MinTemp

The minimum temperature in degrees Celsius.

MaxTemp

The maximum temperature in degrees Celsius.

Rainfall

Amount of rainfall for the day in mm.

Evaporation

Evaporation (mm) in the 24 hours to 9am.

Sunshine

The number of hours of bright sunshine in the day.

WindGustDir

Direction of strongest wind gust in the 24 hours to midnight.

WindGustSpeed

Speed (km/hr) of strongest wind gust in the 24 hours to midnight.

Weather Data in Detail – 2

WindDir9am / WindDir3pm

Wind direction averaged over 10 minutes prior to 9am / 3pm

WindSpeed9am / WindSpeed3pm

Wind speed (km/hr) averaged over 10 minutes prior to 9am / 3pm.

Humidity9am / Humidity3pm

Relative humidity (%) at 9am / 3pm.

Pressure9am / Pressure3pm

Atmospheric pressure (hectopascals) at 9am / 3pm, reduced to mean sea level.

Cloud9am / Cloud3pm

Fraction of sky obsured by cloud at 9am / 3pm.

Temp9am / Temp3pm

Temperature (degrees Celsius) at 9am / 3pm.

RainToday

1 if precipitation(mm) recorded for the day exceeds 1mm; otherwise 0.

RISK_MM

Rainfall for NEXT day

RainTomorrow

1 if precipitation(mm) recorded for the NEXT day exceeds 1mm; otherwise 0.

Samples from Weather Dataset

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed
1	2007-11-01	Canberra	8.0	24.3	0.0	3.4	6.3	NW	30
2	2007-11-02	Canberra	14.0	26.9	3.6	4.4	9.7	ENE	39
3	2007-11-03	Canberra	13.7	23.4	3.6	5.8	3.3	NW	85
4	2007-11-04	Canberra	13.3	15.5	39.8	7.2	9.1	NW	54
5	2007-11-05	Canberra	7.6	16.1	2.8	5.6	10.6	SSE	50
6	2007-11-06	Canberra	6.2	16.9	0.0	5.8	8.2	SE	44

	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm
1	SW	NW	6	20	68	29	1019.7	1015.0
2	E	W	4	17	80	36	1012.4	1008.4
3	N	NNE	6	6	82	69	1009.5	1007.2
4	WNW	W	30	24	62	56	1005.5	1007.0
5	SSE	ESE	20	28	68	49	1018.3	1018.5
6	SE	E	20	24	70	57	1023.8	1021.7

	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RISK_MM	RainTomorrow
1	7	7	14.4	23.6	No	3.6	Yes
2	5	3	17.5	25.7	Yes	3.6	Yes
3	8	7	15.4	20.2	Yes	39.8	Yes
4	2	7	13.5	14.1	Yes	2.8	Yes
5	7	7	11.1	15.4	Yes	0.0	No
6	7	5	10.9	14.8	No	0.2	No

Source: http://www.mapsofworld.com/australia/

Hands-on Setup

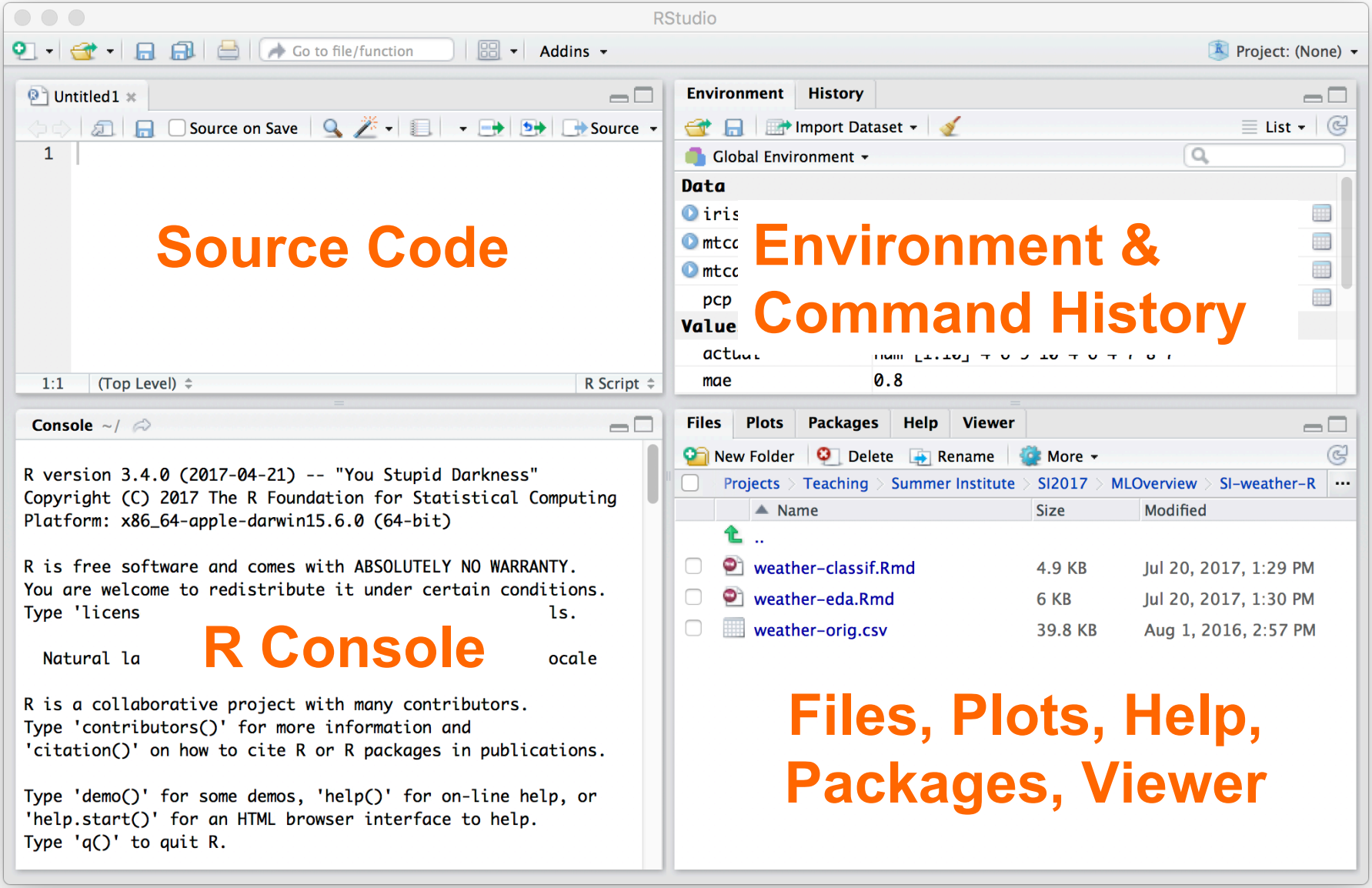
- **Clone github repo**
  - git clone https://github.com/sdsc/sdsc-summer-institute-2017
- **Create directory on laptop**
  - mkdir SI-weather-R
  - cd SI-weather-R
- **Copy files to newly created directory**
  - cp sdsc-summer-institute-2017/datasci2_machine_learning_overview/weather* .
  - cp sdsc-summer-institute-2017/datasci2_machine_learning_overview/data-dictionary-weather.txt .
  - 3 files: weather-classif.Rmd, weather-eda.Rmd, weather-orig.csv

R & RStudio

- **R**
  - Open source programming language and software environment for statistical computing and graphics
  - Widely used for statistical analysis and data mining
- **RStudio**
  - IDE (integrated development environment) for R

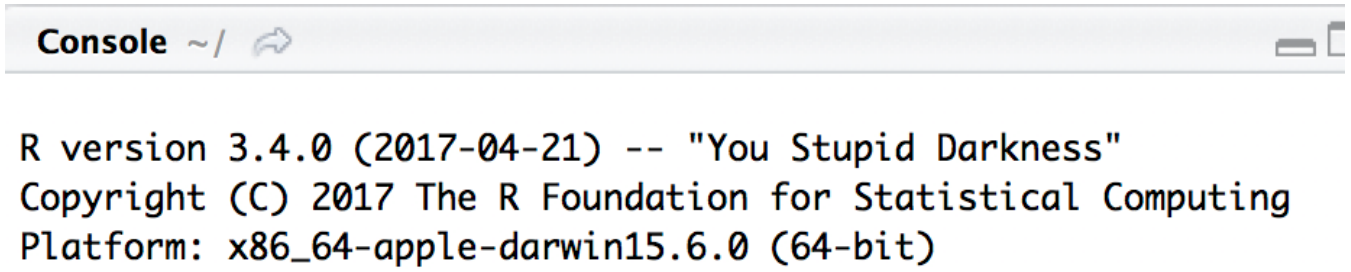
- **Start up Rstudio**
  - Click on RStudio icon or Applications/RStudio.app
  - Has source code editor, build automation tools, debugger

RStudio


RStudio
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Source
1
Source Code
1:1
(Top Level)
R Script
Environment
History
Import Dataset
List
Global Environment
Data
iris
mtcc
mtcc
pcp
Value
mae
0.8
Environment &
Command History
Console ~/
R version 3.4.0 (2017-04-21) -- "You Stupid Darkness"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
R Console
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Projects
Teaching
Summer Institute
SI2017
MLOverview
SI-weather-R
Name
Size
Modified
weather-classif.Rmd
4.9 KB
Jul 20, 2017, 1:29 PM
weather-eda.Rmd
6 KB
Jul 20, 2017, 1:30 PM
weather-orig.csv
39.8 KB
Aug 1, 2016, 2:57 PM
Files, Plots, Help,
Packages, Viewer

Getting Started with R & RStudio

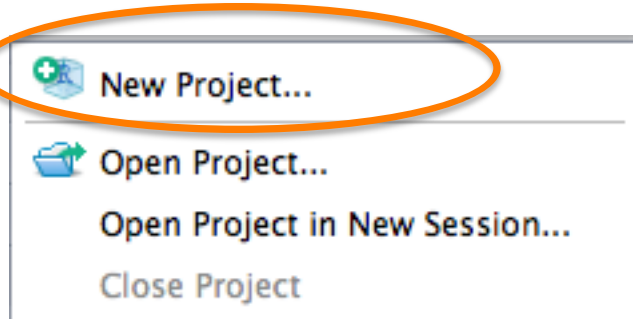
- **R version**
  - => 3.4.0



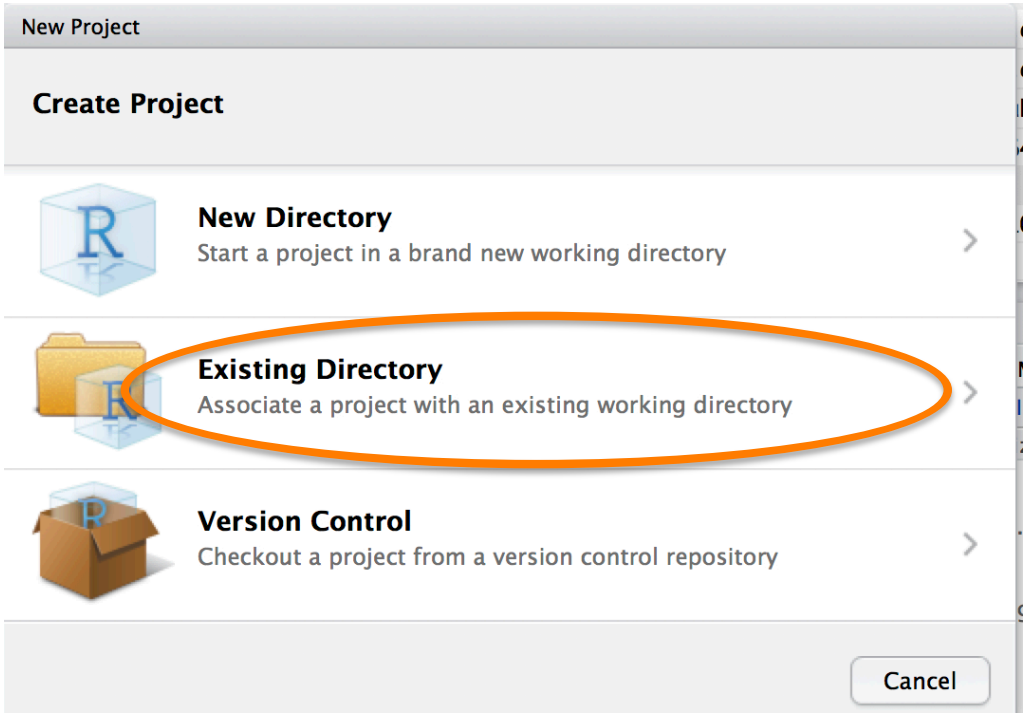
- **RStudio version**
  - Rstudio.Version() % type in Console
  - Or click on RStudio -> About RStudio
  - => 1.0.143

Start New Project

- Start new project

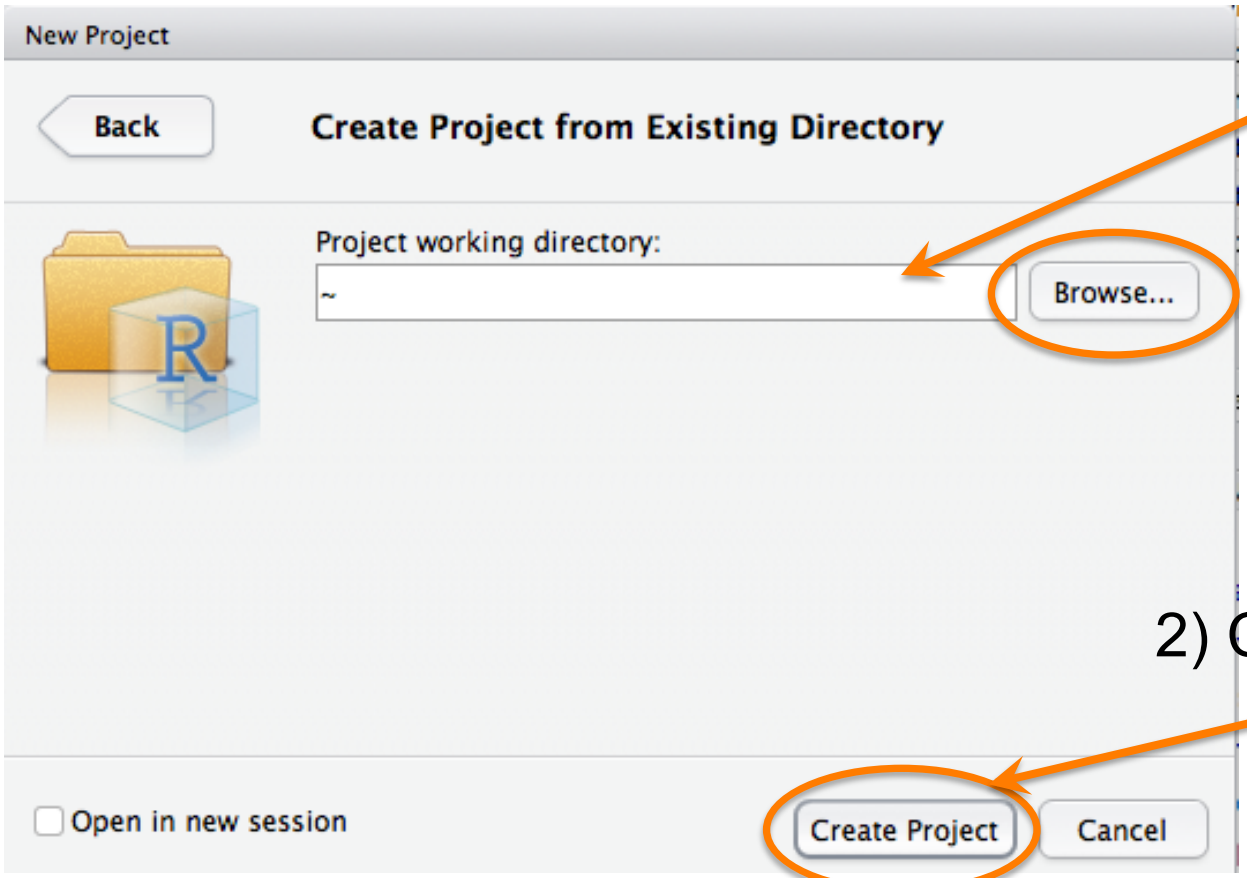


- From Existing Directory



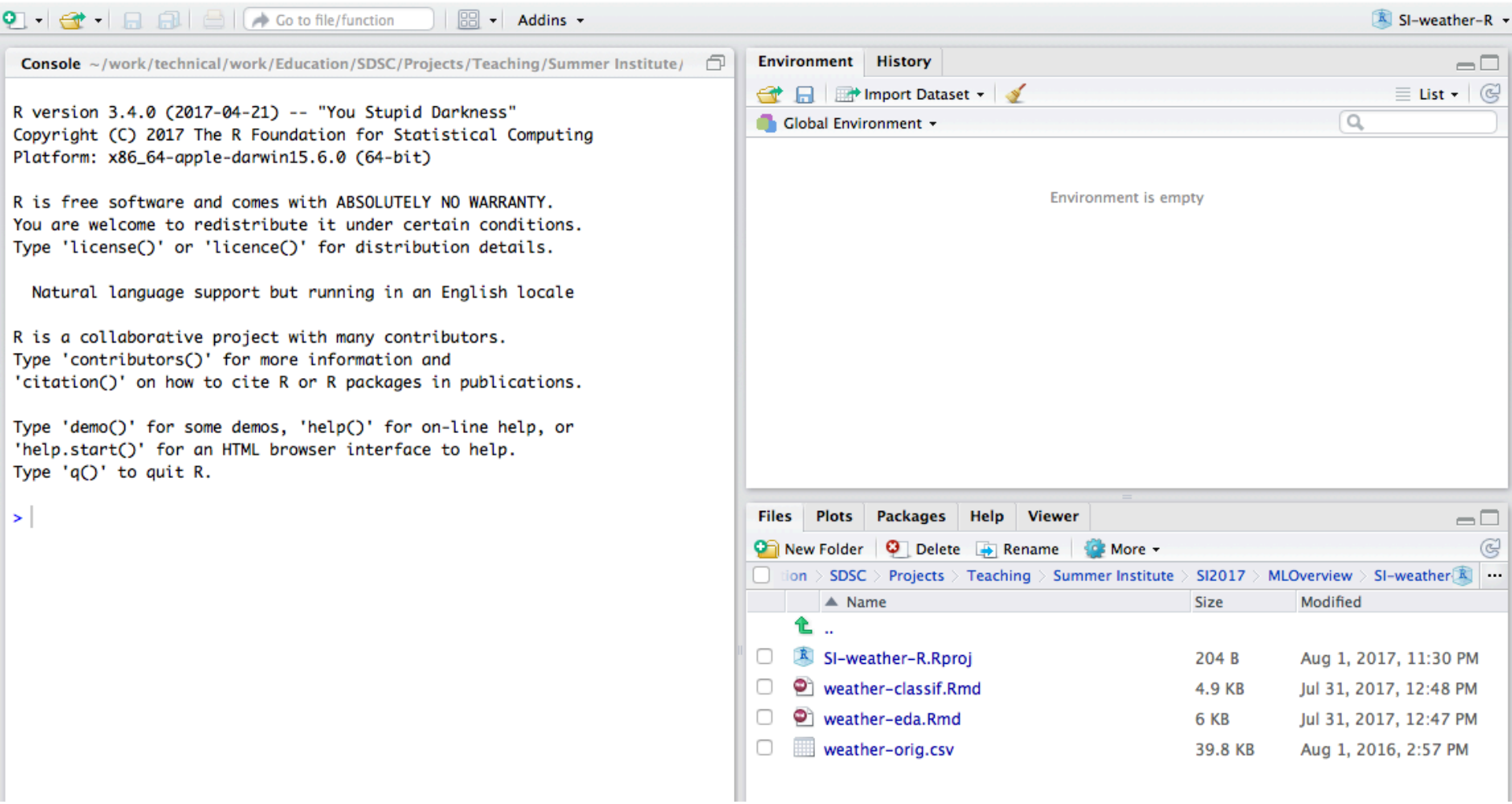
Select Directory & Create Project

1) Browse to directory you created:
<path>/SI-weather-R

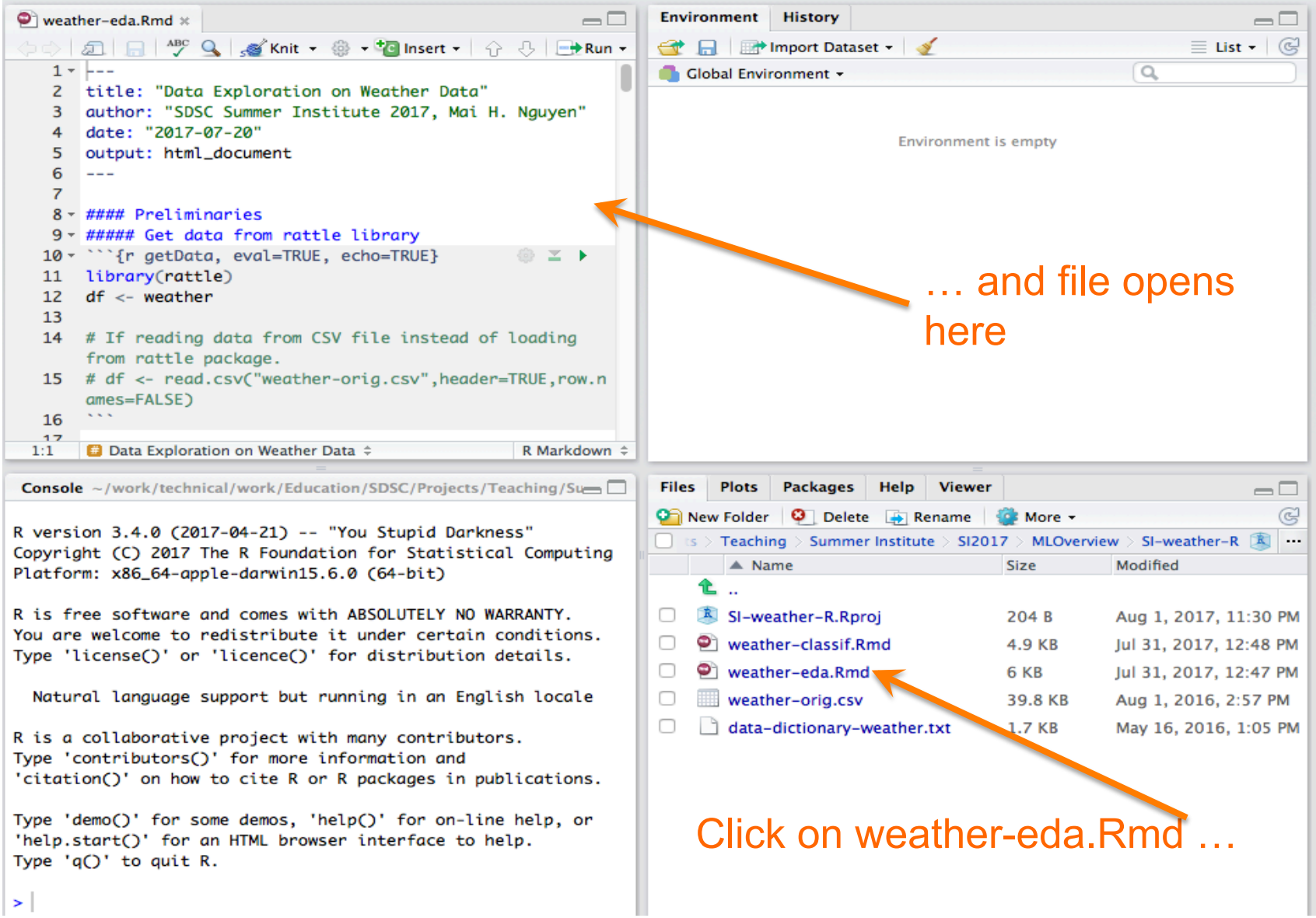


2) Create Project

R Project Created



Open EDA Script



Data Validation

- **Things to Check**
  - Check number of rows and number of columns
  - Check structure of data: data types of columns
  - Look at first and last few rows of data
  - Check that values in data are reasonable.

Check Dimensions

- Get data from ‘rattle’ R library

```
> library(rattle)
> df <- weather
> dim(df)
[1] 366  24
```

- Remove variable RISK_MM

```
> df$RISK_MM <- NULL
> dim(df)
[1] 366  23
```

Data Validation – Structure of Data

```
> str(df)
'data.frame':   366 obs. of  23 variables:
 $ Date          : Date, format: "2007-11-01" "2007-11-02" "2007-11-03" ...
 $ Location      : Factor w/ 49 levels "Adelaide","Albany",..: 10 10 10 10 10 10 10 10 10 10 ...
 $ MinTemp       : num  8 14 13.7 13.3 7.6 6.2 6.1 8.3 8.8 8.4 ...
 $ MaxTemp       : num  24.3 26.9 23.4 15.5 16.1 16.9 18.2 17 19.5 22.8 ...
 $ Rainfall      : num  0 3.6 3.6 39.8 2.8 0 0.2 0 0 16.2 ...
 $ Evaporation   : num  3.4 4.4 5.8 7.2 5.6 5.8 4.2 5.6 4 5.4 ...
 $ Sunshine      : num  6.3 9.7 3.3 9.1 10.6 8.2 8.4 4.6 4.1 7.7 ...
 $ WindGustDir   : Ord.factor w/ 16 levels "N"<"NNE"<"NE"<..: 15 4 15 15 8 7 7 5 9 5 ...
 $ WindGustSpeed: num  30 39 85 54 50 44 43 41 48 31 ...
 $ WindDir9am    : Ord.factor w/ 16 levels "N"<"NNE"<"NE"<..: 11 5 1 14 8 7 7 7 5 9 ...
 $ WindDir3pm    : Ord.factor w/ 16 levels "N"<"NNE"<"NE"<..: 15 13 2 13 6 5 6 5 4 6 ...
 $ WindSpeed9am : num  6 4 6 30 20 20 19 11 19 7 ...
 $ WindSpeed3pm : num  20 17 6 24 28 24 26 24 17 6 ...
 $ Humidity9am  : int  68 80 82 62 68 70 63 65 70 82 ...
 $ Humidity3pm  : int  29 36 69 56 49 57 47 57 48 32 ...
 $ Pressure9am  : num  1020 1012 1010 1006 1018 ...
 $ Pressure3pm  : num  1015 1008 1007 1007 1018 ...
 $ Cloud9am      : int  7 5 8 2 7 7 4 6 7 7 ...
 $ Cloud3pm      : int  7 3 7 7 7 5 6 7 7 1 ...
 $ Temp9am       : num  14.4 17.5 15.4 13.5 11.1 10.9 12.4 12.1 14.1 13.3 ...
 $ Temp3pm       : num  23.6 25.7 20.2 14.1 15.4 14.8 17.3 15.5 18.9 21.7 ...
 $ RainToday     : Factor w/ 2 levels "No","Yes": 1 2 2 2 2 1 1 1 1 2 ...
 $ RainTomorrow : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 1 1 1 2 1 ...
```

Data Validation – Check values

```
> head(df,3)
        Date Location MinTemp MaxTemp Rainfall Evaporation Sunshine WindGustDir WindGustSpeed
1 2007-11-01 Canberra      8.0     24.3      0.0         3.4      6.3          NW            30
2 2007-11-02 Canberra     14.0     26.9      3.6         4.4      9.7         ENE            39
3 2007-11-03 Canberra     13.7     23.4      3.6         5.8      3.3          NW            85
  WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm Humidity9am Humidity3pm Pressure9am
1         SW         NW            6           20          68          29      1019.7
2          E          W            4           17          80          36      1012.4
3          N        NNE            6            6          82          69      1009.5
  Pressure3pm Cloud9am Cloud3pm Temp9am Temp3pm RainToday RainTomorrow
1      1015.0        7        7    14.4    23.6        No          Yes
2      1008.4        5        3    17.5    25.7       Yes          Yes
3      1007.2        8        7    15.4    20.2       Yes          Yes
>
> tail(df,3)
           Date Location MinTemp MaxTemp Rainfall Evaporation Sunshine WindGustDir WindGustSpeed
364 2008-10-29 Canberra     12.5     19.9        0         8.4      5.3         ESE            43
365 2008-10-30 Canberra     12.5     26.9        0         5.0      7.1          NW            46
366 2008-10-31 Canberra     12.3     30.2        0         6.0     12.6          NW            78
    WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm Humidity9am Humidity3pm Pressure9am
364        ENE        ENE           11            9          63          47      1024.0
365        SSW        WNW            6           28          69          39      1021.0
366         NW        WNW           31           35          43          13      1009.6
    Pressure3pm Cloud9am Cloud3pm Temp9am Temp3pm RainToday RainTomorrow
364      1022.8        3        2    14.5    18.3        No          No
365      1016.2        6        7    15.8    25.9        No          No
366      1009.2        1        1    23.8    28.6        No          No
```

Check Data Frame

- Get column headers

```
> names(df)
 [1] "Date"          "Location"      "MinTemp"       "MaxTemp"       "Rainfall"
 [6] "Evaporation"   "Sunshine"      "WindGustDir"   "WindGustSpeed" "WindDir9am"
[11] "WindDir3pm"    "WindSpeed9am"  "WindSpeed3pm"  "Humidity9am"   "Humidity3pm"
[16] "Pressure9am"   "Pressure3pm"   "Cloud9am"      "Cloud3pm"      "Temp9am"
[21] "Temp3pm"       "RainToday"     "RainTomorrow"
```

Summary Statistics on Data Frame

```
> summary(df)
      Date                    Location      MinTemp          MaxTemp          Rainfall        Evaporation        Sunshine     
 Min.   :2007-11-01   Canberra     :366   Min.   :-5.300   Min.   : 7.60   Min.   : 0.000   Min.   : 0.200   Min.   : 0.000  
 1st Qu.:2008-01-31   Adelaide     :  0   1st Qu.: 2.300   1st Qu.:15.03   1st Qu.: 0.000   1st Qu.: 2.200   1st Qu.: 5.950  
 Median :2008-05-01   Albany       :  0   Median : 7.450   Median :19.65   Median : 0.000   Median : 4.200   Median : 8.600  
 Mean   :2008-05-01   Albury       :  0   Mean   : 7.266   Mean   :20.55   Mean   : 1.428   Mean   : 4.522   Mean   : 7.909  
 3rd Qu.:2008-07-31   AliceSprings :  0   3rd Qu.:12.500   3rd Qu.:25.50   3rd Qu.: 0.200   3rd Qu.: 6.400   3rd Qu.:10.500  
 Max.   :2008-10-31   BadgerysCreek:  0   Max.   :20.900   Max.   :35.80   Max.   :39.800   Max.   :13.800   Max.   :13.600  
                      (Other)      :  0                                                                      NA's   :3       
  WindGustDir  WindGustSpeed     WindDir9am    WindDir3pm    WindSpeed9am     WindSpeed3pm     Humidity9am     Humidity3pm   
 NW     : 73   Min.   :13.00   SE     : 47   WNW    : 61   Min.   : 0.000   Min.   : 0.00   Min.   :36.00   Min.   :13.00  
 NNW    : 44   1st Qu.:31.00   SSE    : 40   NW     : 61   1st Qu.: 6.000   1st Qu.:11.00   1st Qu.:64.00   1st Qu.:32.25  
 E      : 37   Median :39.00   NNW    : 36   NNW    : 47   Median : 7.000   Median :17.00   Median :72.00   Median :43.00  
 WNW    : 35   Mean   :39.84   N      : 31   N      : 30   Mean   : 9.652   Mean   :17.99   Mean   :72.04   Mean   :44.52  
 ENE    : 30   3rd Qu.:46.00   NW     : 30   ESE    : 27   3rd Qu.:13.000   3rd Qu.:24.00   3rd Qu.:81.00   3rd Qu.:55.00  
 (Other):144   Max.   :98.00   (Other):151   (Other):139   Max.   :41.000   Max.   :52.00   Max.   :99.00   Max.   :96.00  
 NA's   :  3   NA's   :2       NA's   : 31   NA's   :  1   NA's   :7                                                       
  Pressure9am      Pressure3pm       Cloud9am        Cloud3pm        Temp9am          Temp3pm       RainToday RainTomorrow
 Min.   : 996.5   Min.   : 996.8   Min.   :0.000   Min.   :0.000   Min.   : 0.100   Min.   : 5.10   No :300   No :300    
 1st Qu.:1015.4   1st Qu.:1012.8   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 7.625   1st Qu.:14.15   Yes: 66   Yes: 66    
 Median :1020.1   Median :1017.4   Median :3.500   Median :4.000   Median :12.550   Median :18.55                        
 Mean   :1019.7   Mean   :1016.8   Mean   :3.891   Mean   :4.025   Mean   :12.358   Mean   :19.23                        
 3rd Qu.:1024.5   3rd Qu.:1021.5   3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:17.000   3rd Qu.:24.00                        
 Max.   :1035.7   Max.   :1033.2   Max.   :8.000   Max.   :8.000   Max.   :24.700   Max.   :34.50                        
```

Summary Statistics on Variable

- **Summary statistics on individual variables**

```
> mean(df$MinTemp)
[1] 7.265574
>
> var(df$MinTemp)
[1] 36.31026
>
> sd(df$MinTemp)
[1] 6.0258
>
> summary(df$MinTemp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -5.300   2.850   7.900   7.743  12.800  20.900
>
```

More Statistics on Variable

```
> min(df$Date)                                  # First date
[1] "2007-11-01"
>
> max(df$Date)                                  # Last date
[1] "2008-10-31"
>
> sum(df$Rainfall)                              # Total rainfall amount for year
[1] 522.8
>
> sum(df$Rainfall) / length(df$Rainfall)        # Average daily rainfall
[1] 1.428415
>
```

Missing Values

- Determine number of samples with NAs

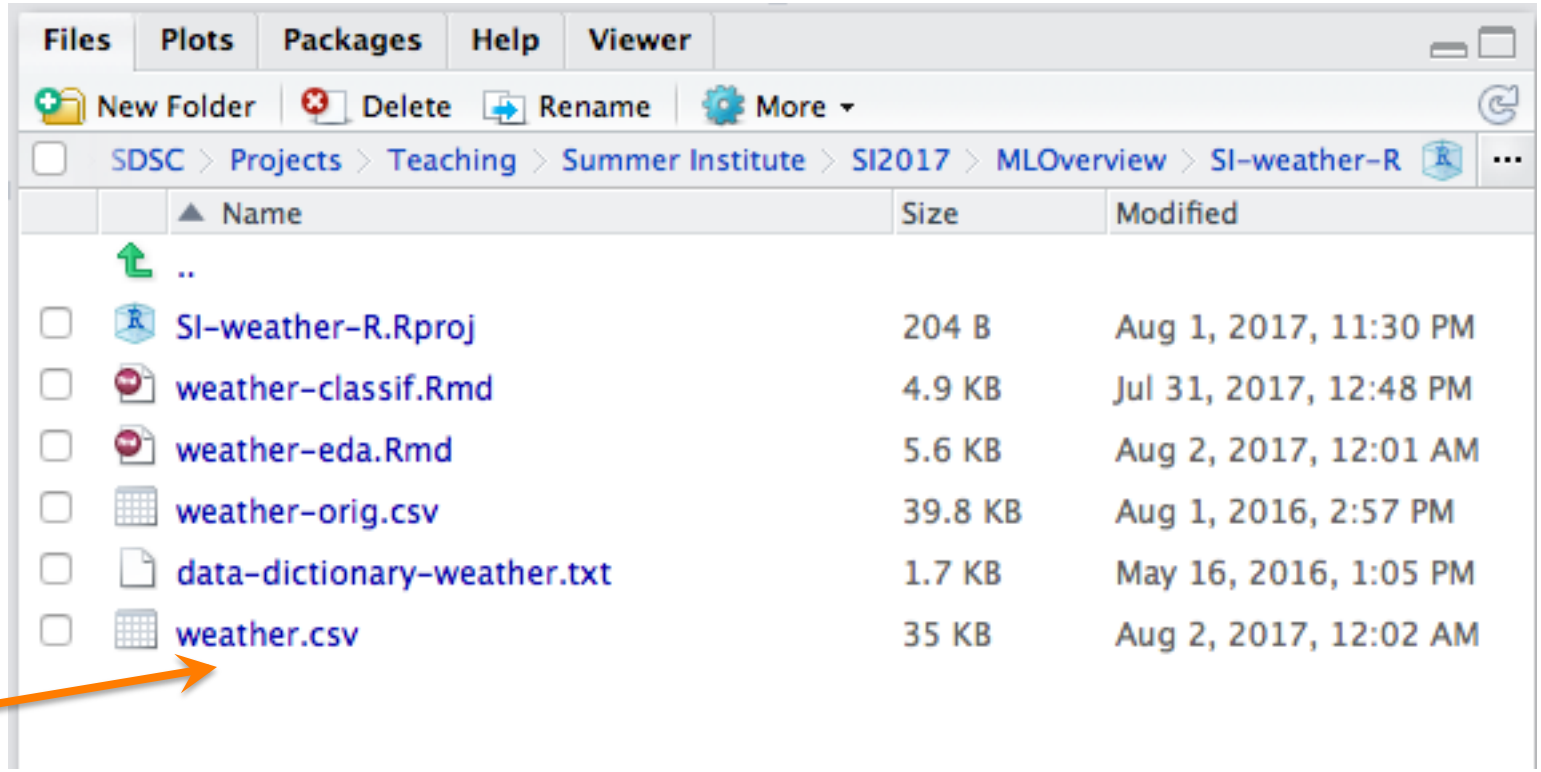
```
>
> # Find number of places in dataset where there are missing values
> sum(is.na(df))
[1] 47
>
> # Find number of samples (i.e., rows) with any missing value
> sum(complete.cases(weather))
[1] 328
>
```

- Remove samples with NAs

```
> dim(df)
[1] 366  23
> df <- na.omit(df)
> dim(df)
[1] 328  23
>
```

Save Data to File

```
> write.csv(df, "weather.csv", row.names=FALSE)
>
```



New file

Correlation – Setup

- Extract numerical variables

```
> numeric.vars <- c("Rainfall","Evaporation","Sunshine","WindSpeed9am","Humidity9am","Pressure9am",
"Cloud9am","Temp9am")
> df.num <- df[numeric.vars]
> dim(df.num)
[1] 328   8
> head(df.num,3)
  Rainfall Evaporation Sunshine WindSpeed9am Humidity9am Pressure9am Cloud9am Temp9am
1      0.0         3.4      6.3            6          68      1019.7         7    14.4
2      3.6         4.4      9.7            4          80      1012.4         5    17.5
3      3.6         5.8      3.3            6          82      1009.5         8    15.4
>
```

Correlation

- Compute pairwise correlation for numerical variables

```
> cor(df.num, use = "pairwise")
```

	Rainfall	Evaporation	Sunshine	WindSpeed9am	Humidity9am	Pressure9am	Cloud9am	Temp9am
Rainfall	1.00000000	-0.011767012	-0.15806222	0.238705011	0.1463208	-0.34873128	0.17260983	0.07189336
Evaporation	-0.01176701	1.000000000	0.31012411	0.006259292	-0.4922181	-0.36393637	-0.11426073	0.68874886
Sunshine	-0.15806222	0.310124113	1.00000000	-0.103840777	-0.5015955	0.02562989	-0.69760347	0.19965882
WindSpeed9am	0.23870501	0.006259292	-0.10384078	1.000000000	-0.2223374	-0.34429047	0.11762541	-0.01784294
Humidity9am	0.14632082	-0.492218142	-0.50159550	-0.222337388	1.0000000	0.10225017	0.41749552	-0.39564694
Pressure9am	-0.34873128	-0.363936368	0.02562989	-0.344290471	0.1022502	1.00000000	-0.16831577	-0.45366889
Cloud9am	0.17260983	-0.114260735	-0.69760347	0.117625413	0.4174955	-0.16831577	1.00000000	0.01001183
Temp9am	0.07189336	0.688748862	0.19965882	-0.017842944	-0.3956469	-0.45366889	0.01001183	1.00000000

Categorical Variables

```
> str(df)
'data.frame':   366 obs. of  23 variables:
 $ Date          : Date, format: "2007-11-01" "2007-11-02" "2007-11-03" ...
 $ Location      : Factor w/ 49 levels "Adelaide","Albany",..: 10 10 10 10 10 10 10 10 10 10 ...
 $ MinTemp       : num  8 14 13.7 13.3 7.6 6.2 6.1 8.3 8.8 8.4 ...
 $ MaxTemp       : num  24.3 26.9 23.4 15.5 16.1 16.9 18.2 17 19.5 22.8 ...
 $ Rainfall      : num  0 3.6 3.6 39.8 2.8 0 0.2 0 0 16.2 ...
 $ Evaporation   : num  3.4 4.4 5.8 7.2 5.6 5.8 4.2 5.6 4 5.4 ...
 $ Sunshine      : num  6.3 9.7 3.3 9.1 10.6 8.2 8.4 4.6 4.1 7.7 ...
 $ WindGustDir   : Ord.factor w/ 16 levels "N"<"NNE"<"NE"<..: 15 4 15 15 8 7 7 5 9 5 ...
 $ WindGustSpeed: num  30 39 85 54 50 44 43 41 48 31 ...
 $ WindDir9am    : Ord.factor w/ 16 levels "N"<"NNE"<"NE"<..: 11 5 1 14 8 7 7 7 5 9 ...
 $ WindDir3pm    : Ord.factor w/ 16 levels "N"<"NNE"<"NE"<..: 15 13 2 13 6 5 6 5 4 6 ...
 $ WindSpeed9am : num  6 4 6 30 20 20 19 11 19 7 ...
 $ WindSpeed3pm : num  20 17 6 24 28 24 26 24 17 6 ...
 $ Humidity9am  : int  68 80 82 62 68 70 63 65 70 82 ...
 $ Humidity3pm  : int  29 36 69 56 49 57 47 57 48 32 ...
 $ Pressure9am  : num  1020 1012 1010 1006 1018 ...
 $ Pressure3pm  : num  1015 1008 1007 1007 1018 ...
 $ Cloud9am      : int  7 5 8 2 7 7 4 6 7 7 ...
 $ Cloud3pm      : int  7 3 7 7 7 5 6 7 7 1 ...
 $ Temp9am       : num  14.4 17.5 15.4 13.5 11.1 10.9 12.4 12.1 14.1 13.3 ...
 $ Temp3pm       : num  23.6 25.7 20.2 14.1 15.4 14.8 17.3 15.5 18.9 21.7 ...
 $ RainToday     : Factor w/ 2 levels "No","Yes": 1 2 2 2 2 1 1 1 1 2 ...
 $ RainTomorrow : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 1 1 1 2 1 ...
```

Statistics on Categorical Variables

- Categories (levels) of WindGustDir

```
> levels(df$WindGustDir)
 [1] "N"   "NNE" "NE"  "ENE" "E"   "ESE" "SE"  "SSE" "S"   "SSW" "SW"  "WSW" "W"   "WNW" "NW"  "NNW"
```

- Number of occurrences for each level

```
> table(df$WindGustDir)

  N NNE  NE ENE   E ESE  SE SSE   S SSW  SW WSW   W WNW  NW NNW
 21   7  15  29  34  23  11  12  21   4   3   2  15  32  64  35
```

```
> sort(table(df$WindGustDir))

WSW  SW SSW NNE  SE SSE  NE   W   N   S ESE ENE WNW   E NNW  NW
  2   3   4   7  11  12  15  15  21  21  23  29  32  34  35  64
```

Statistics on Categorical Variables

- Percentages for each WindGustDir level

```
> round (table(df$WindGustDir) / sum(table(df$WindGustDir)) * 100, digits=2)

    N   NNE    NE   ENE     E   ESE    SE   SSE     S   SSW    SW   WSW     W   WNW    NW   NNW
 6.40  2.13  4.57  8.84 10.37  7.01  3.35  3.66  6.40  1.22  0.91  0.61  4.57  9.76 19.51 10.67
```

- Sorted percentages for each WindGustDir level

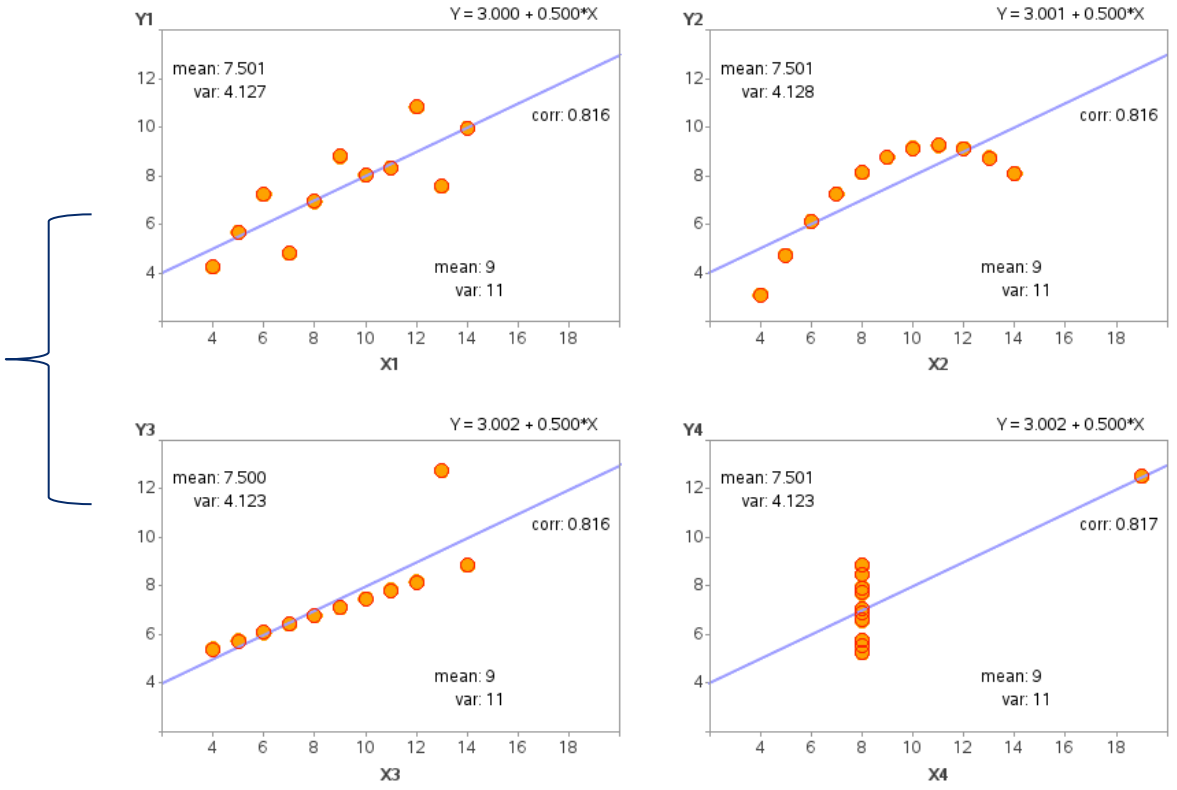
```
> sort (round (table(df$WindGustDir) / sum(table(df$WindGustDir)) * 100, digits=2))

  WSW    SW   SSW   NNE    SE   SSE    NE     W     N     S   ESE   ENE   WNW     E   NNW    NW
 0.61  0.91  1.22  2.13  3.35  3.66  4.57  4.57  6.40  6.40  7.01  8.84  9.76 10.37 10.67 19.51
```

Data Visualization

- **Not just because graphs are pretty...**

Anscombe's quartet: 4 data sets that have "identical" summary statistics yet are very different



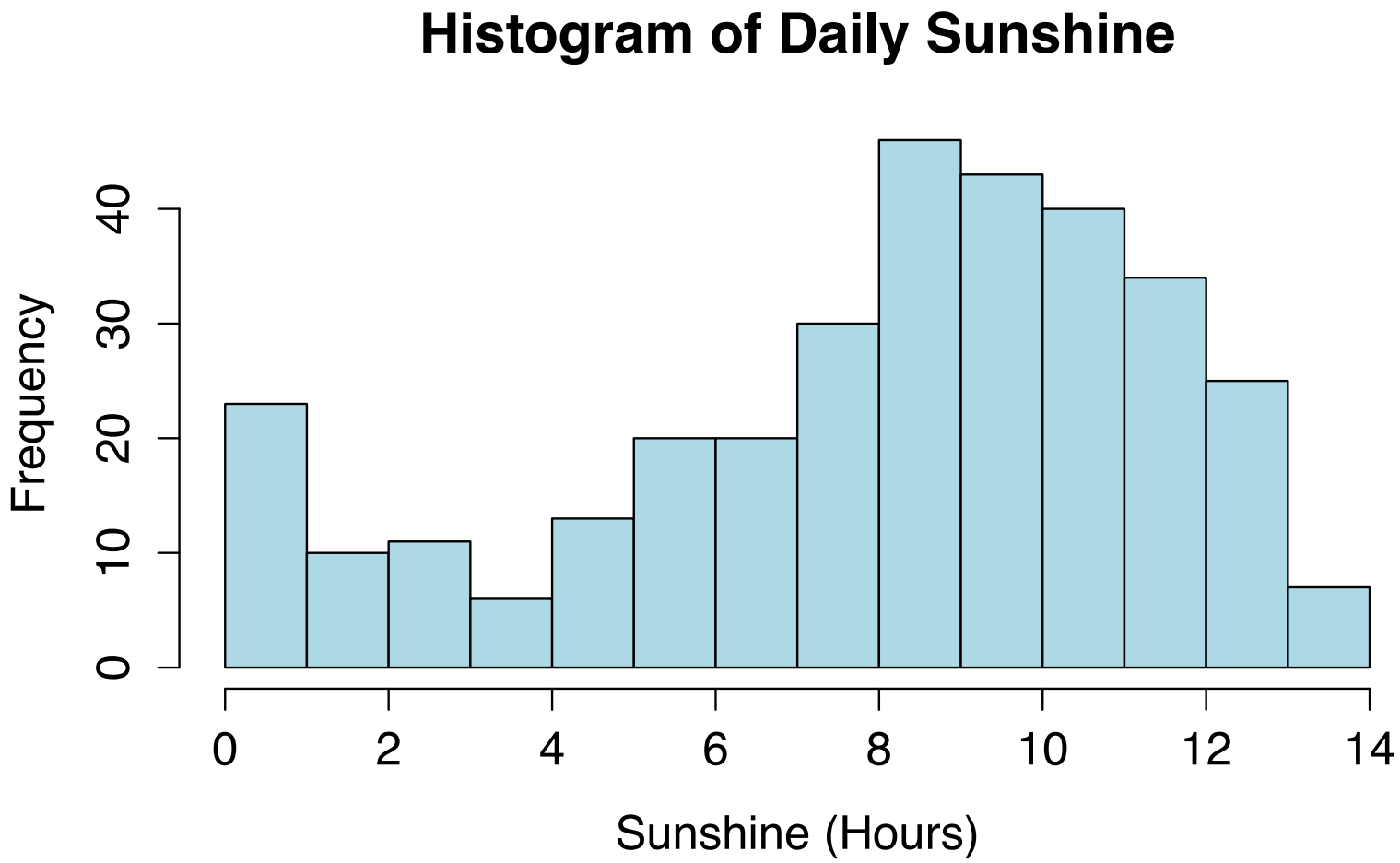
Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Data Visualization

- **Types of graphs**
  - Histogram
  - Bar plot
  - Scatter plot
  - Line plot
  - Box plot
  - etc.

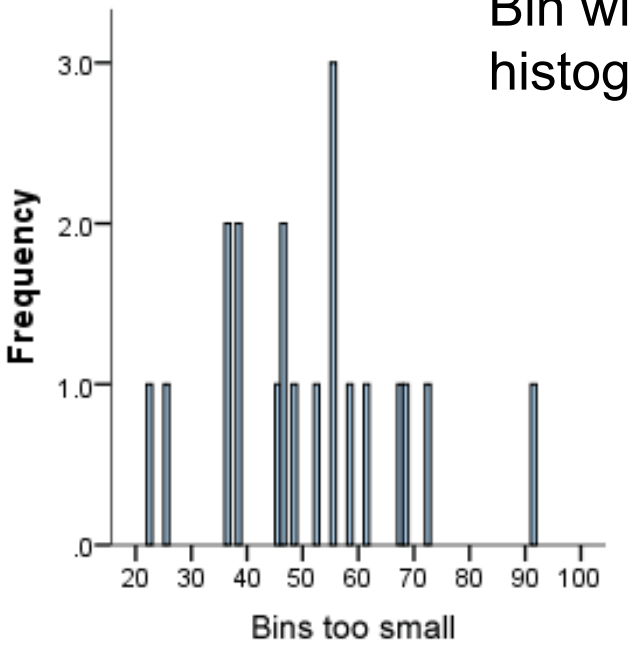
Histogram

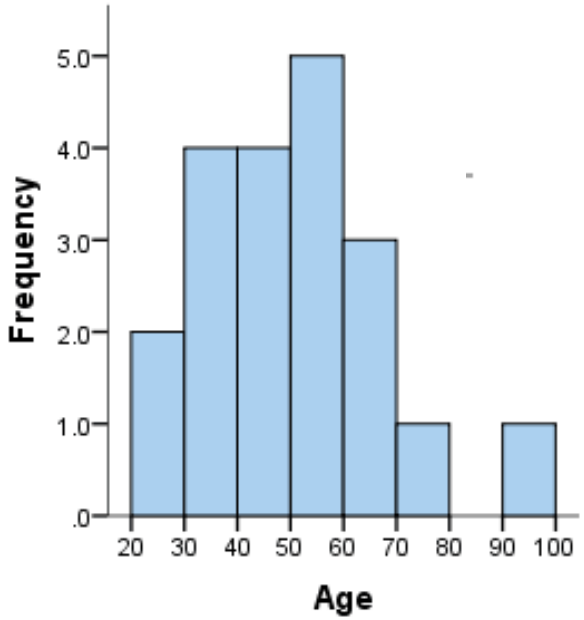
```
hist(df$Sunshine,col="lightblue",main="Histogram of Daily Sunshine",xlab="Sunshine (Hours)")
```

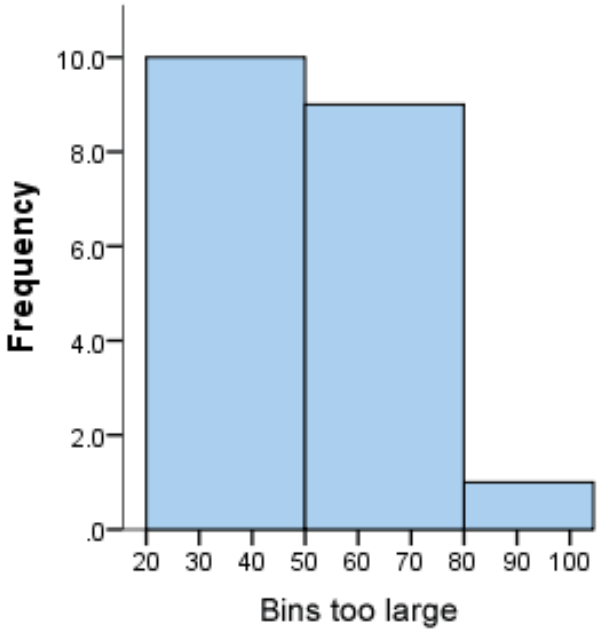


Histogram – Bin Width

Bin width can affect shape of histogram



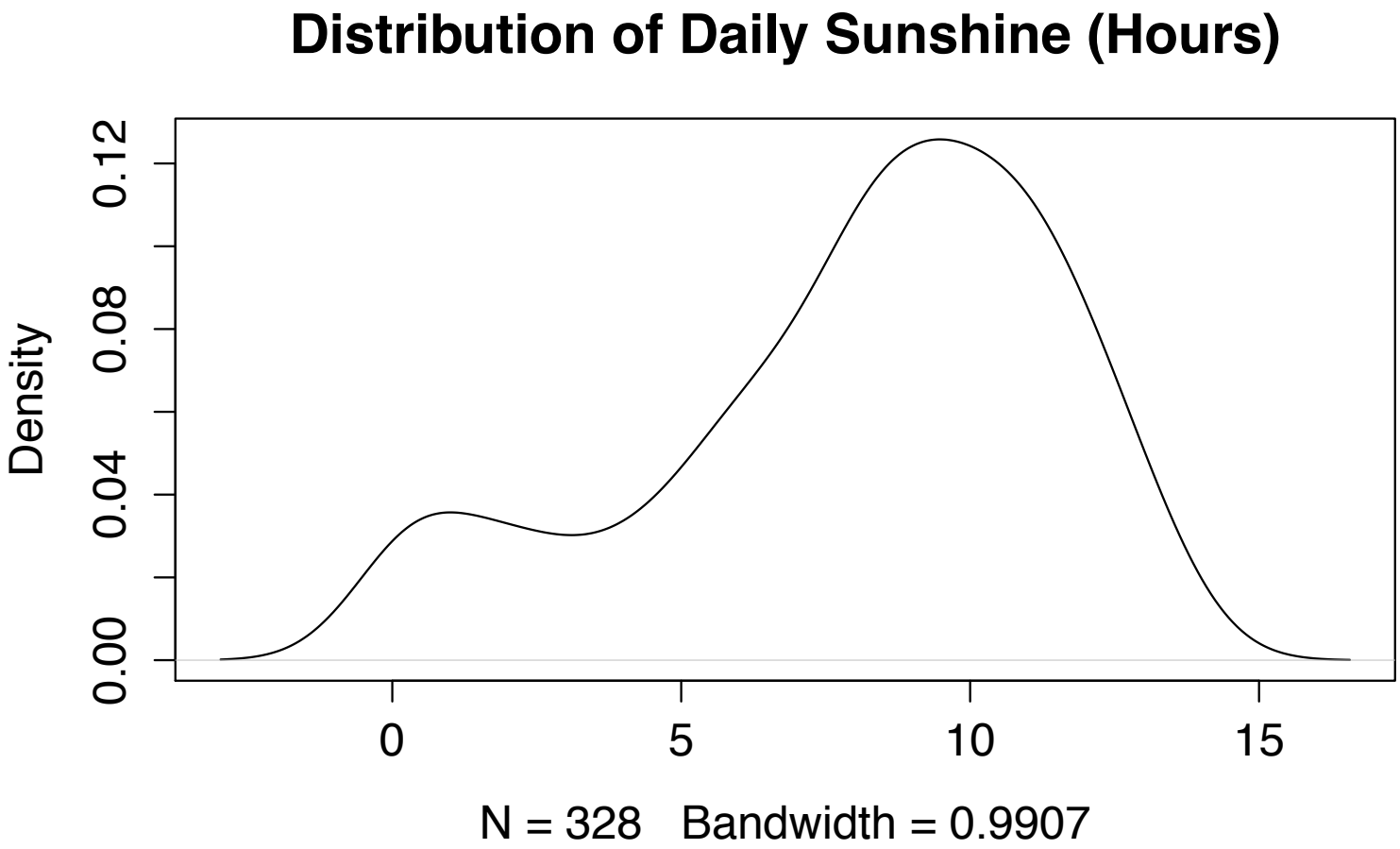




Source: https://statistics.laerd.com/statistical-guides/understanding-histograms.php

Density Plot

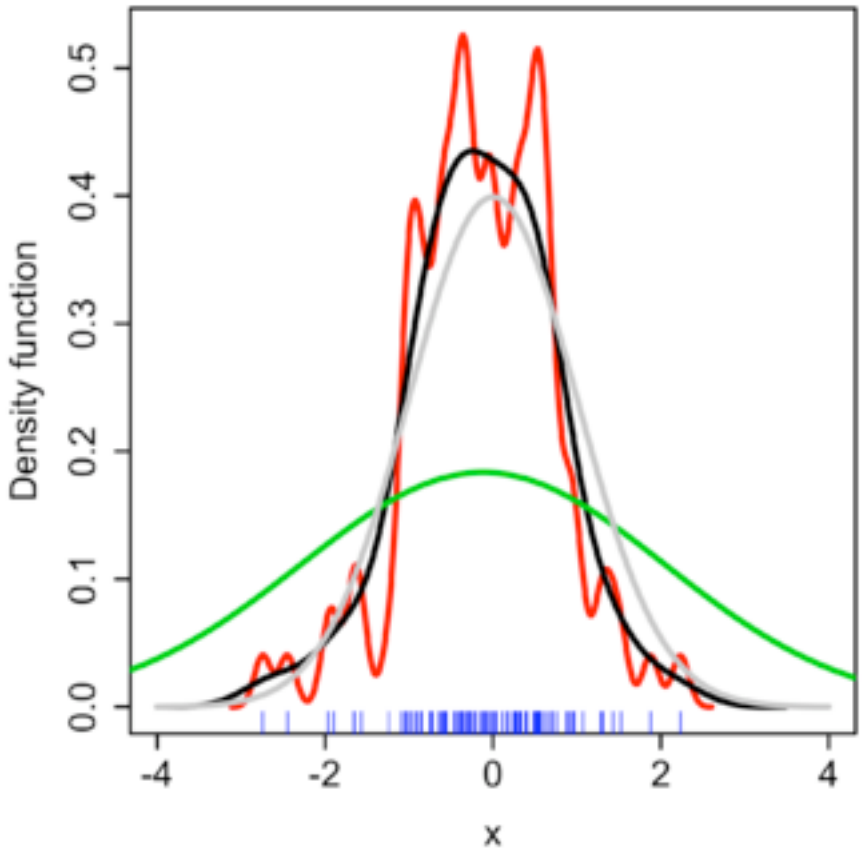
```
plot(density(df$Sunshine), main="Distribution of Daily Sunshine (Hours)")
```



Density Plot - Bandwidth

Bandwidth can affect shape of density curve

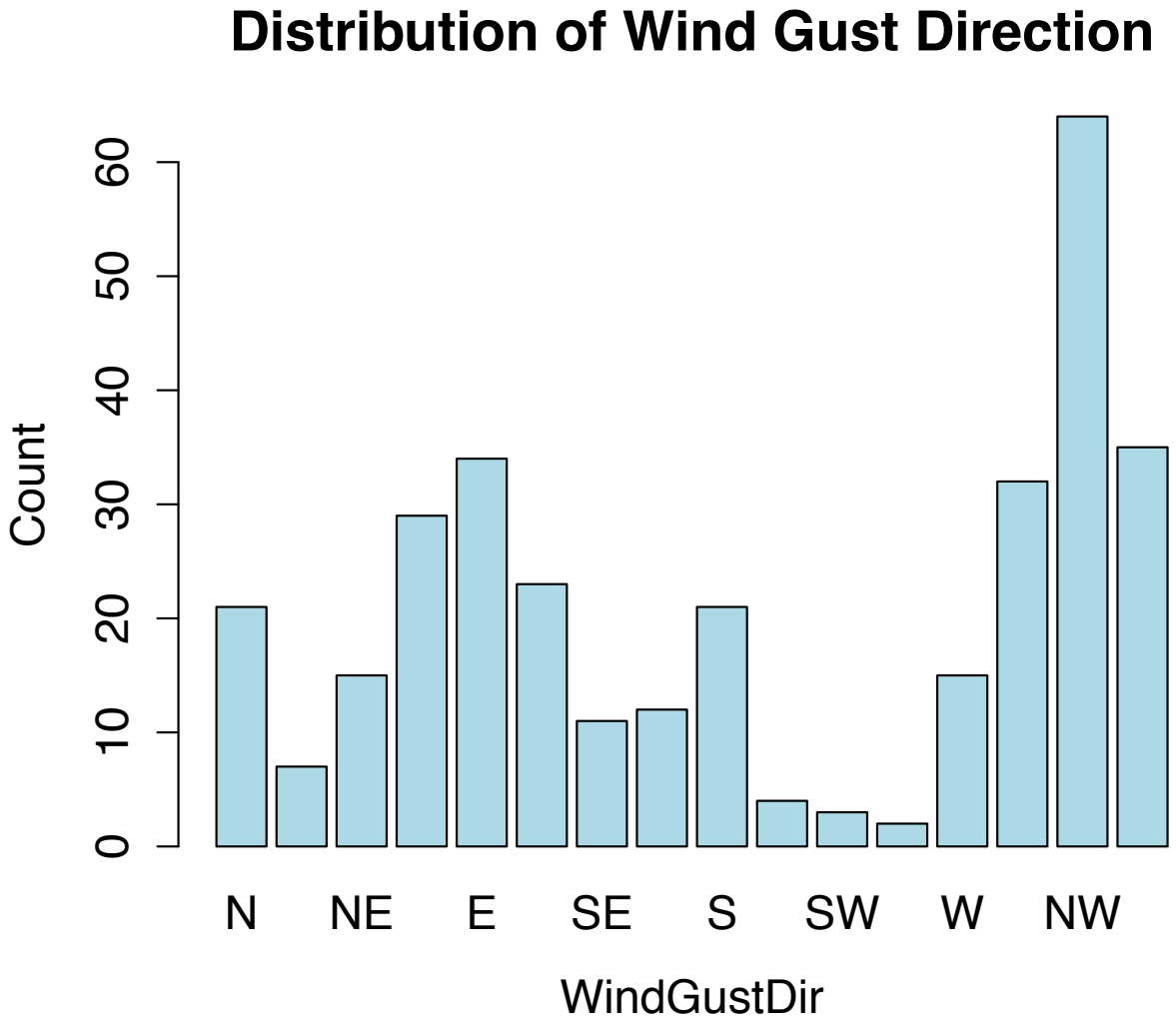
- True density (normal)
- Bandwidth optimal
- Bandwidth too small
- Bandwidth too large



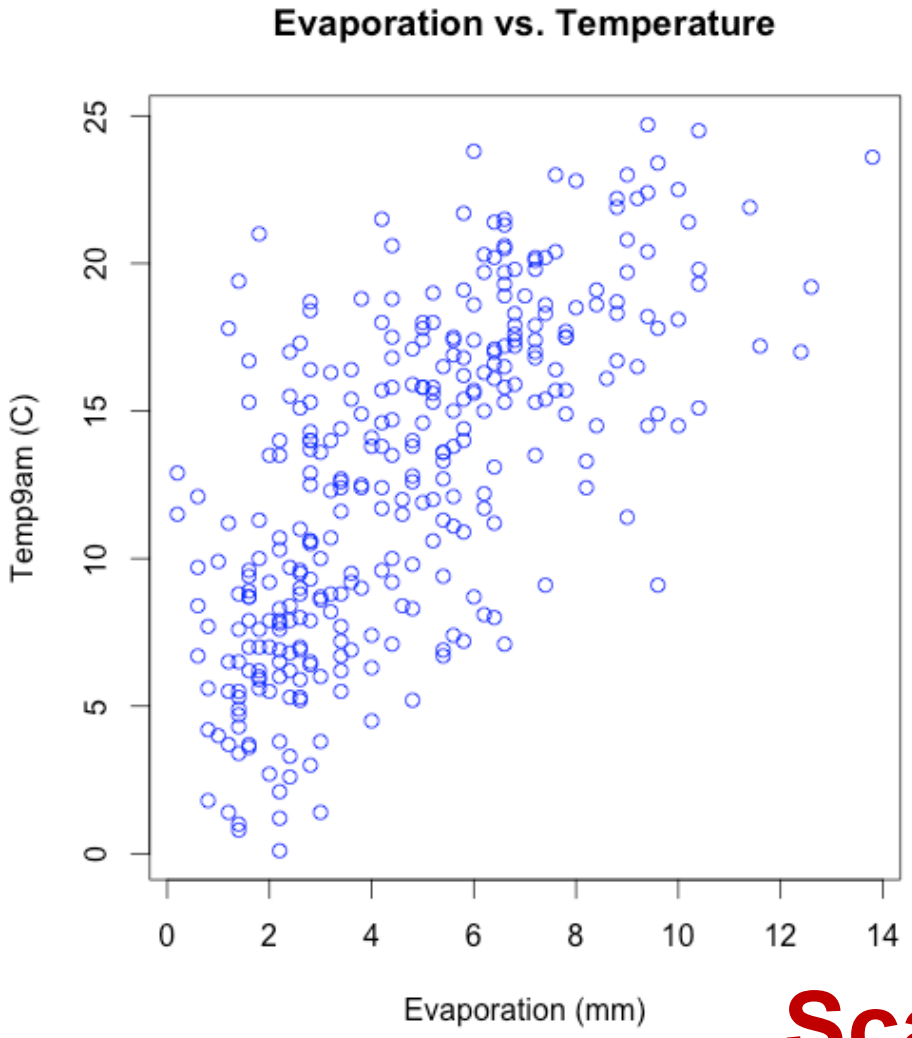
Source: https://en.wikipedia.org/wiki/Kernel_density_estimation

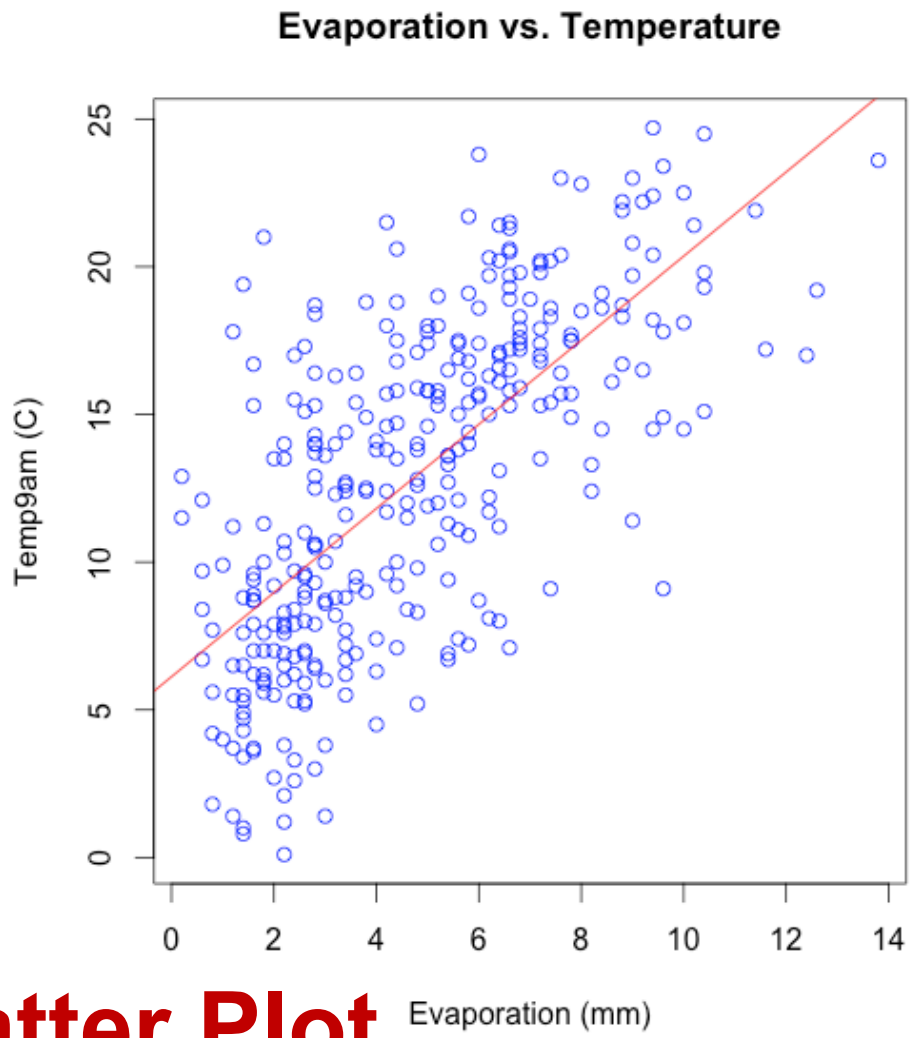
```
plot(df$WindGustDir,col="lightblue",main="Distribution of Wind Gust Direction",xlab="WindGustDir",ylab="Count")
```

Bar Plot



```
plot(df$Evaporation,df$Temp9am,col="blue",main="Evaporation vs. Temperature",xlab="Evaporation (mm)",ylab
="Temp9am (C)")
abline(lm(df$Temp9am ~ df$Evaporation), col="red")
```

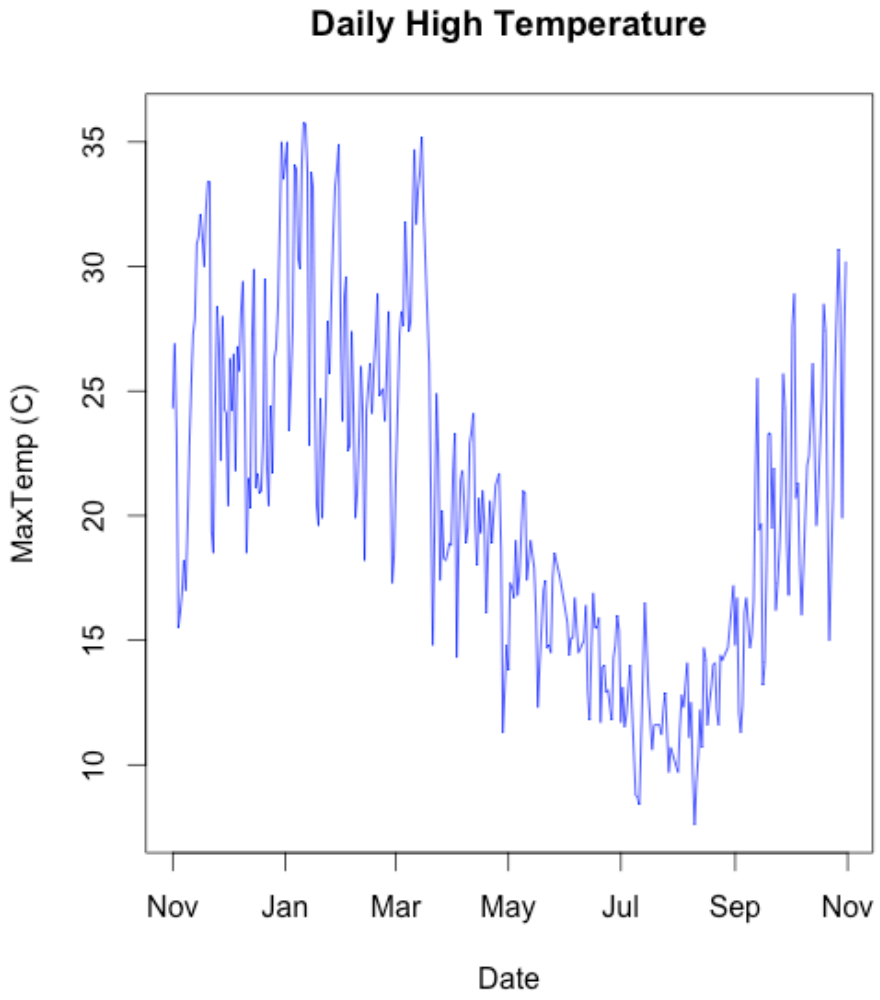




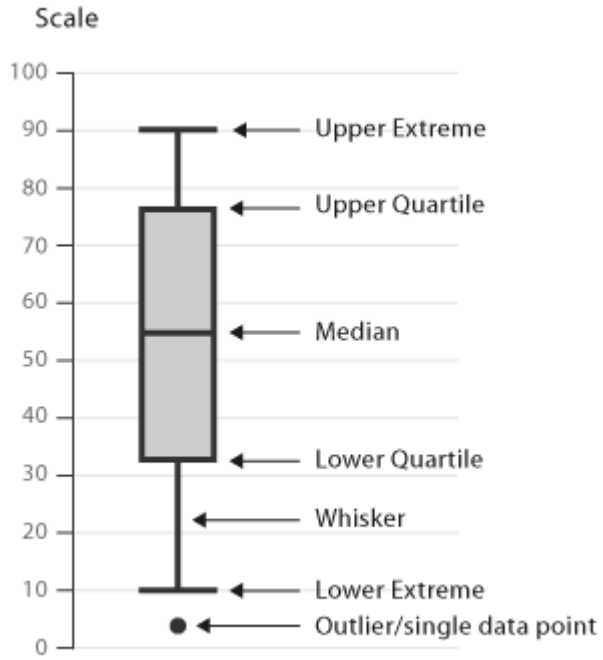
Scatter Plot

Line Plot

```
plot(df$Date,df$MaxTemp,type='l',col="blue",main="Daily High Temperature",xlab="Date",ylab="MaxTemp (C)")
```



Box Plot



Quantiles:

- 25% quantile (lower quartile)
- 50% quantile (median)
- 75% quantile (upper quartile)

Robust maximum:

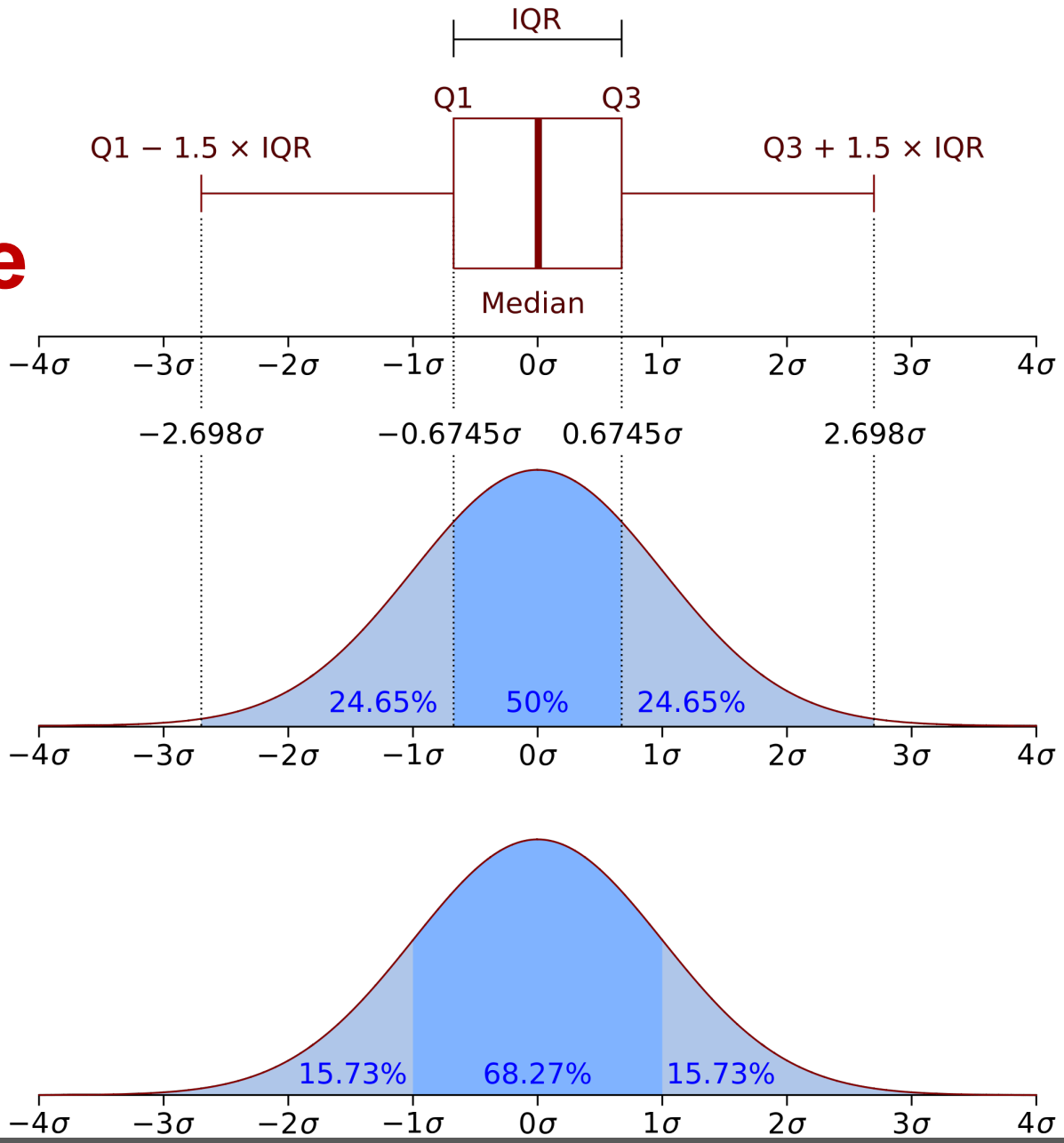
- 95% quantile

Robust minimum:

- 5% quantile

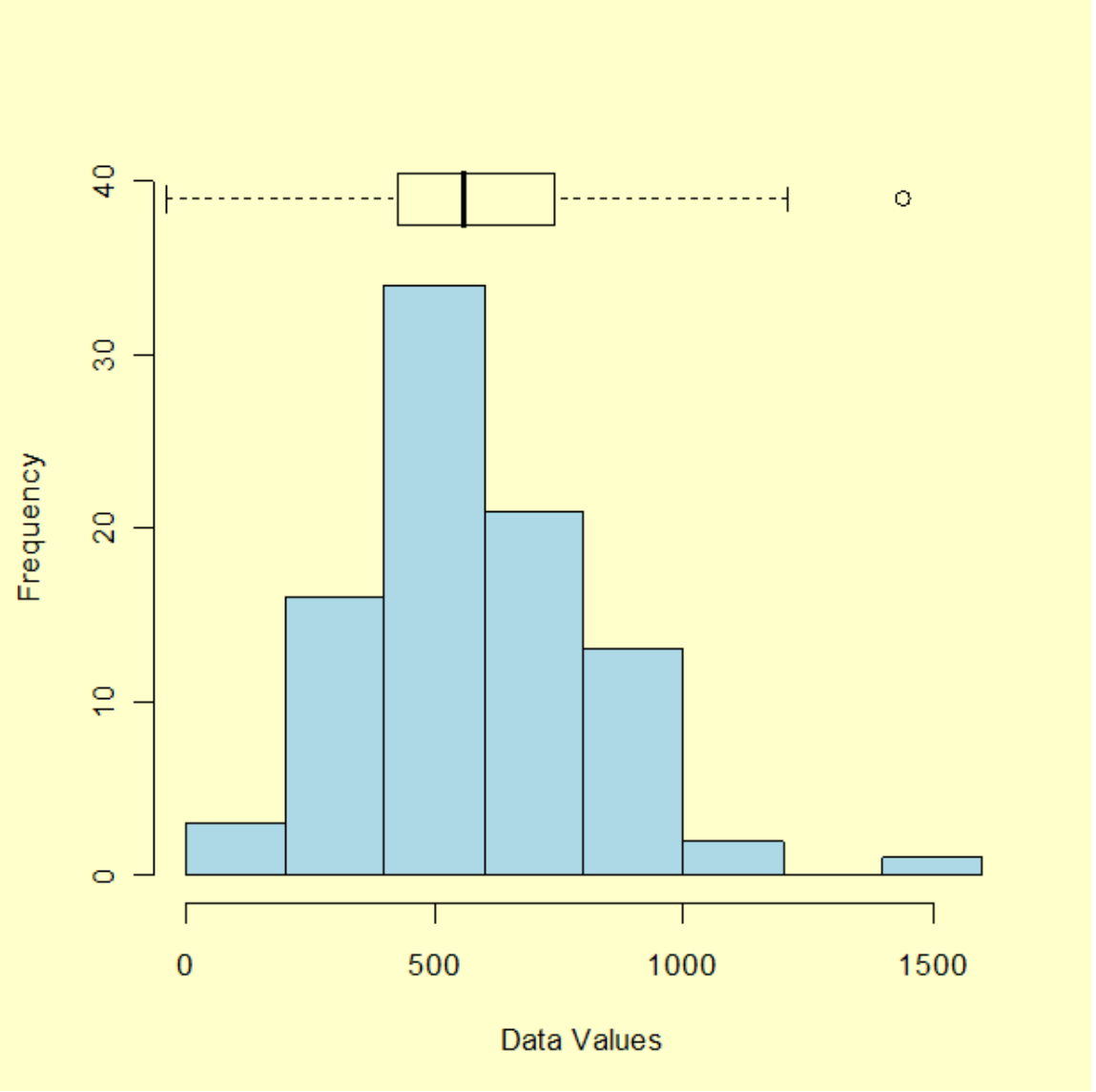
Source: http://www.datavizcatalogue.com/methods/box_plot.html

Box Plot vs. Density Curve



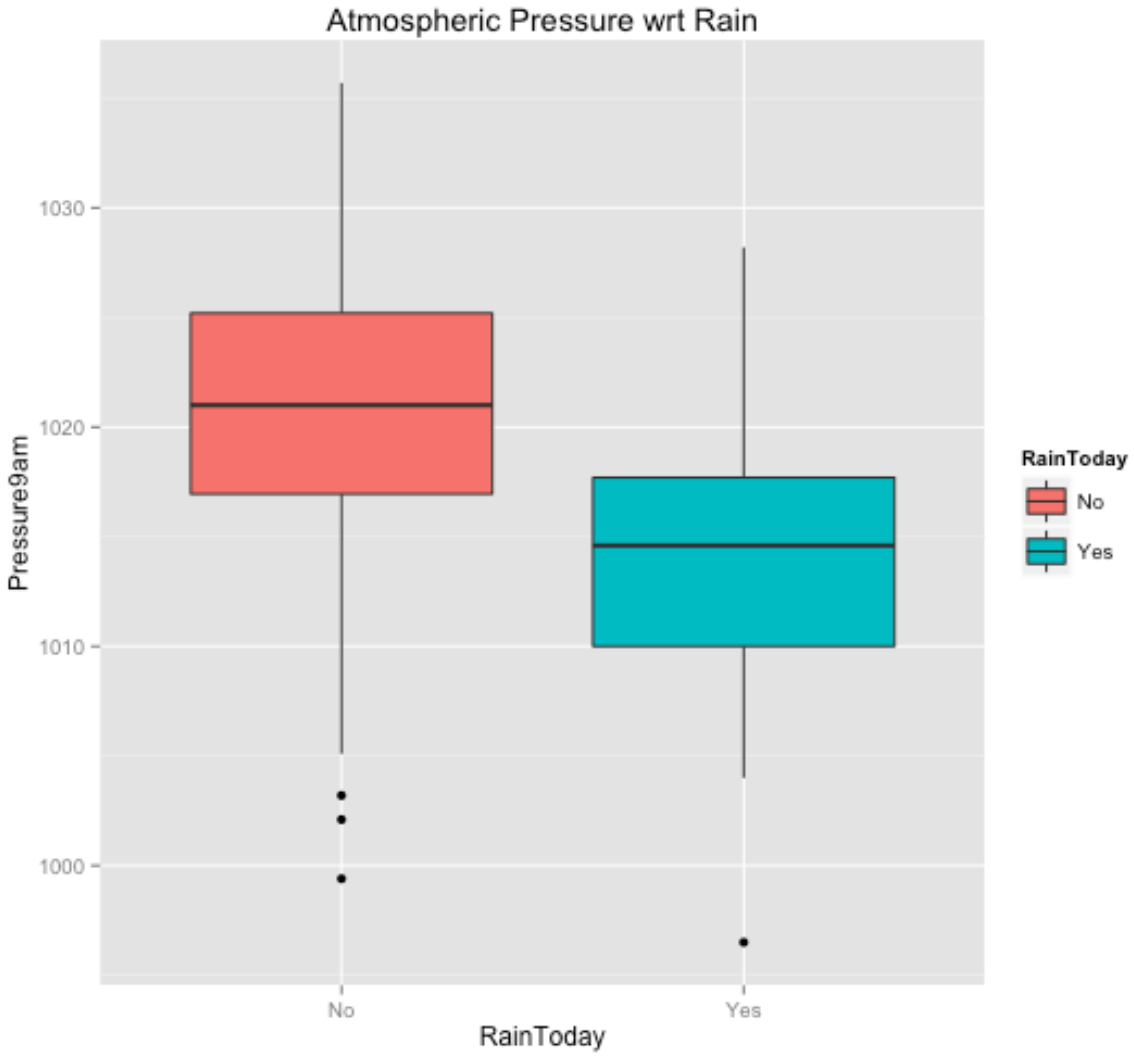
Source: https://en.wikipedia.org/wiki/Box_plot

Box Plot vs. Histogram



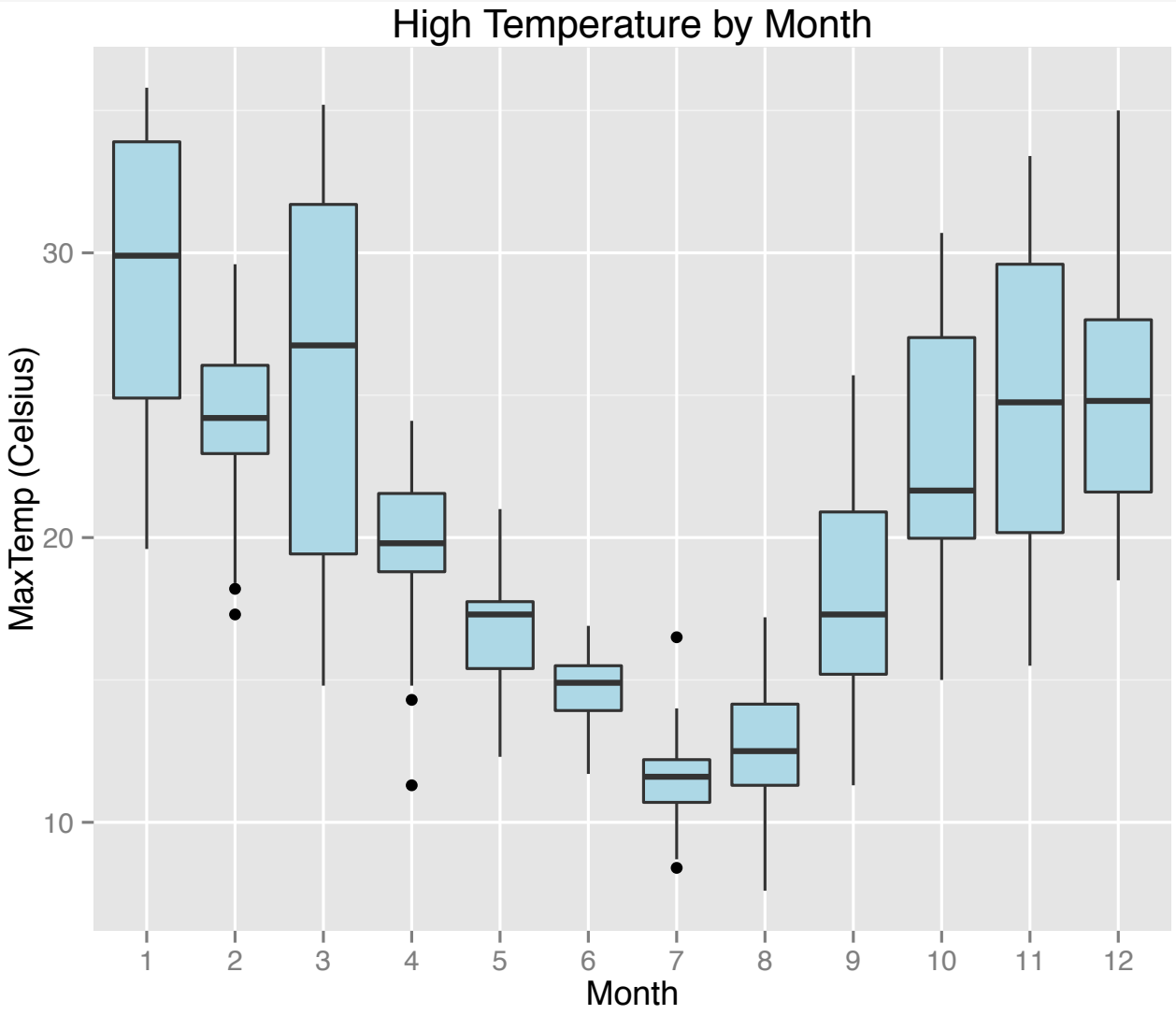
Source: http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

```
ggplot(df, aes(x=RainToday, y=Pressure9am, fill=RainToday)) + geom_boxplot() +
  ggtitle("Atmospheric Pressure wrt Rain")
```



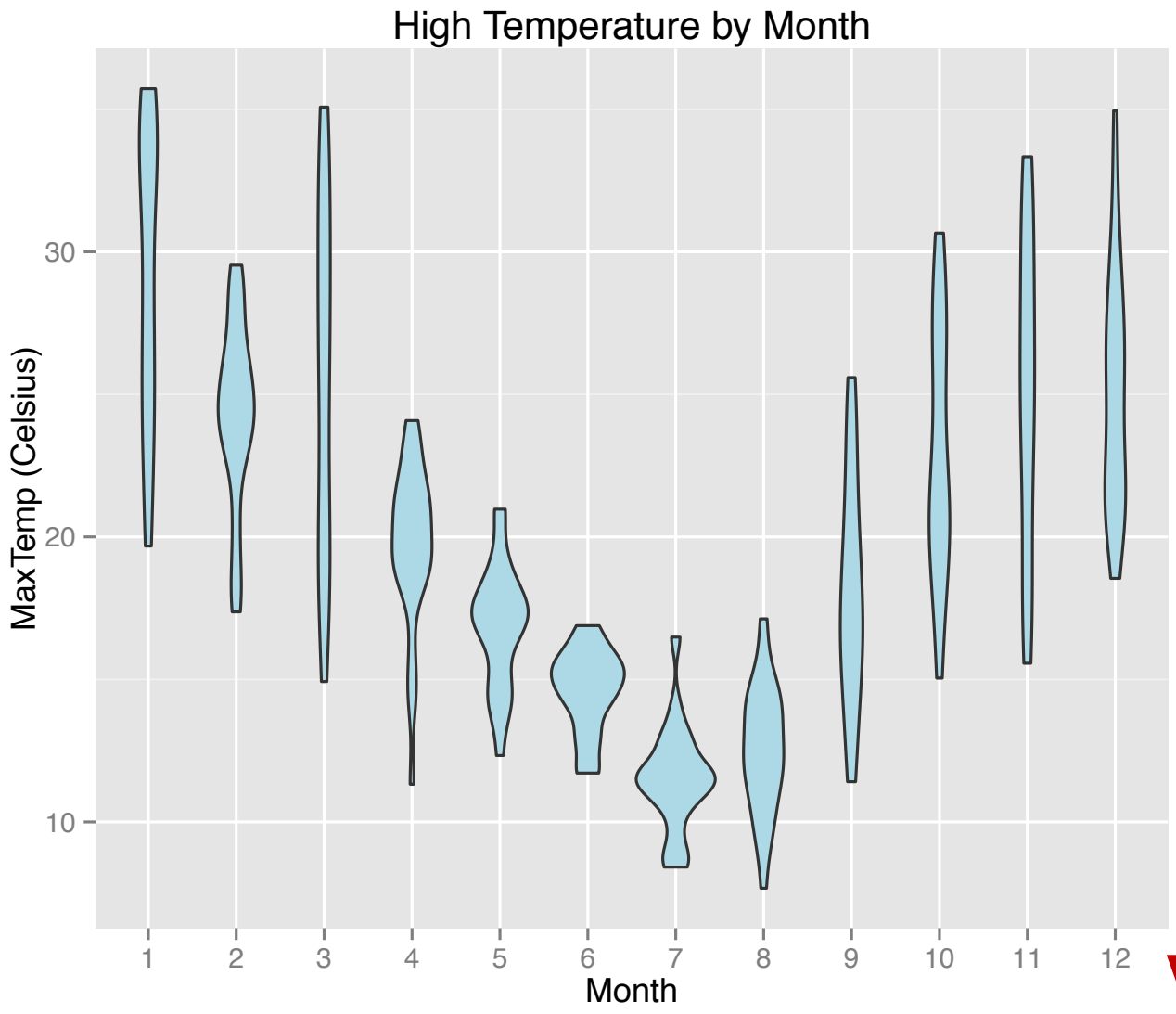
Box Plot

```
library(lubridate)    # R package for handling dates
ggplot(df, aes(x=as.factor(month(Date)), y=MaxTemp)) + geom_boxplot(fill="lightblue") +
  xlab("Month") + ylab("MaxTemp (Celsius)") + ggtitle("High Temperature by Month")
```



Box Plot

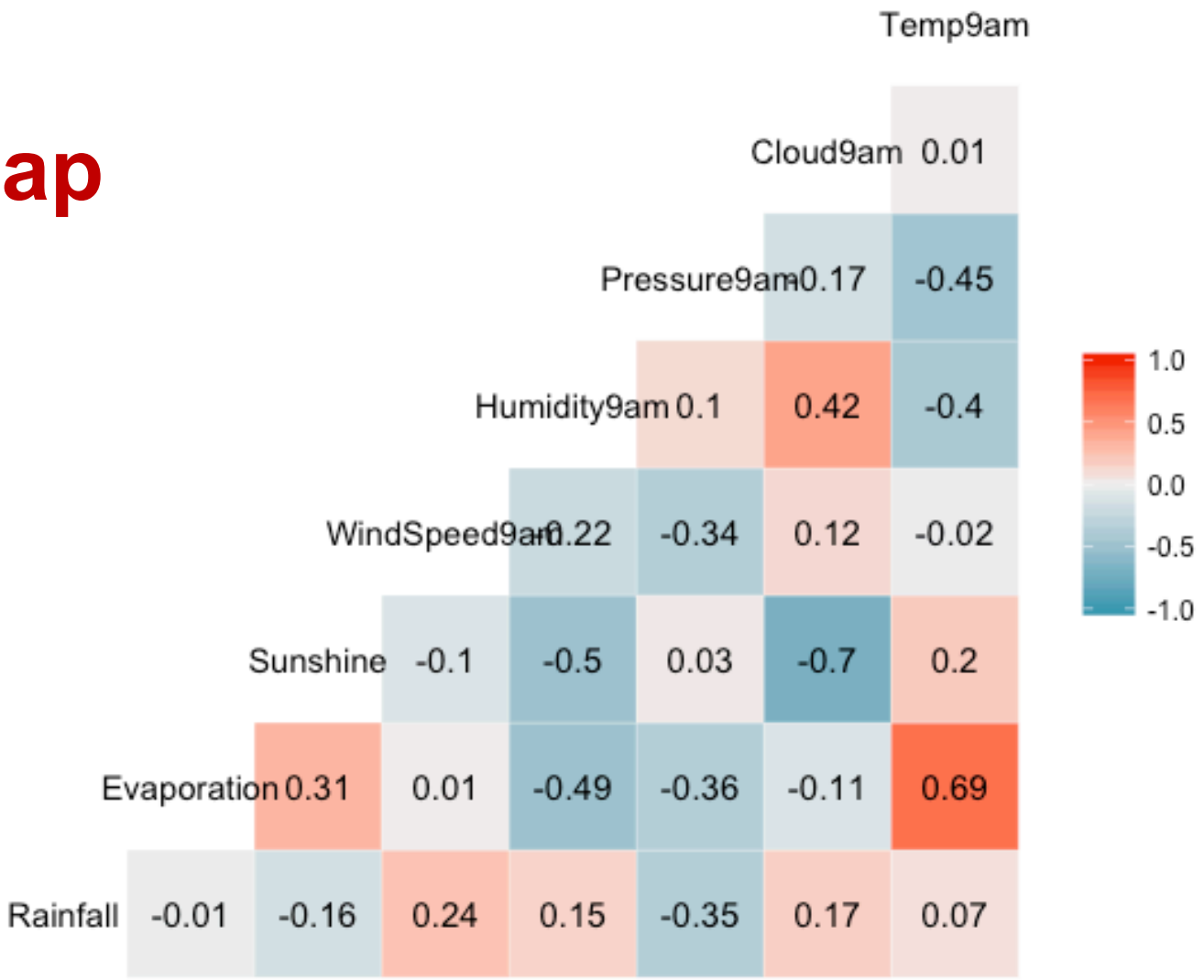
```
ggplot(df, aes(x=as.factor(month(Date)), y=MaxTemp)) + geom_violin(fill="lightblue") +
  xlab("Month") + ylab("MaxTemp (Celsius)") + ggtitle("High Temperature by Month")
```



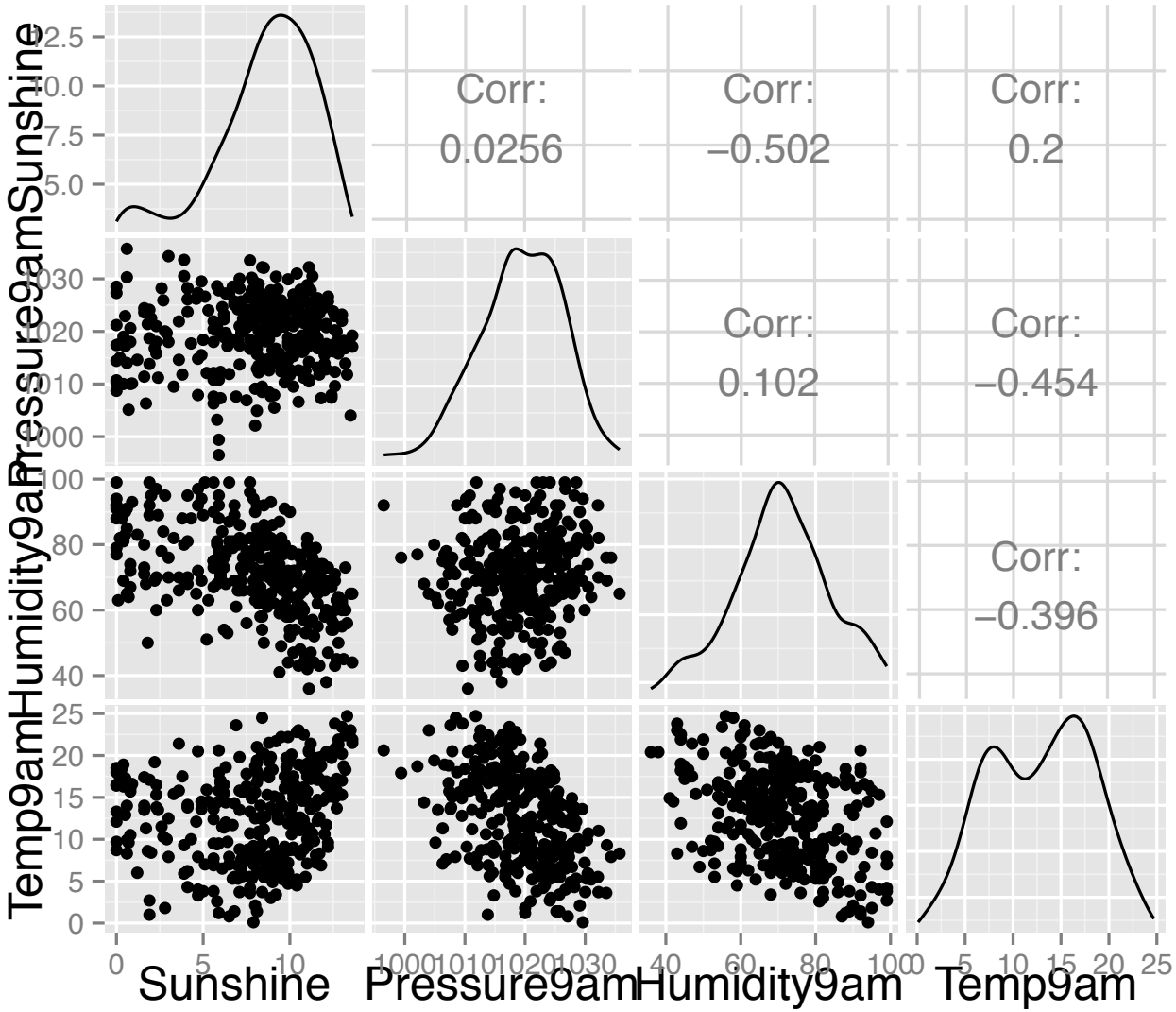
Violin Plot

```
ggcorr(df.num, use="pairwise", label=TRUE, label_round=2)
```

Heat Map



```
ggpairs(df[c("Sunshine","Pressure9am","Humidity9am","Temp9am")])
```



Pairwise Correlation Plot

Data Exploration – Key Points

- **Purpose**
  - To understand the data you have
- **Methods**
  - Data validation
  - Summary statistics
  - Visualization

References

- R. Cabacoff. Quick-R: Basic Graphs. Retrieved from http://www.statmethods.net/graphs/
- CRAN: The Comprehensive R Archive Network. Retrieved from https://cran.r-project.org/
- B. Schloerke et al. GGally. Retrieved from https://cran.r-project.org/web/packages/GGally/index.html
- H. Wickham. ggplot2. Retrieved from http://ggplot2.org/
- G. Williams et al. Package 'rattle'. Retrieved from https://cran.r-project.org/web/packages/rattle/index.html

Questions?