02-05: Vector Operations and NA

1 - Purpose

- mathematical operations on vectors
- dealing with missing values in a vector
- R shortcuts for vector mathematics

2 - Questions about the material...

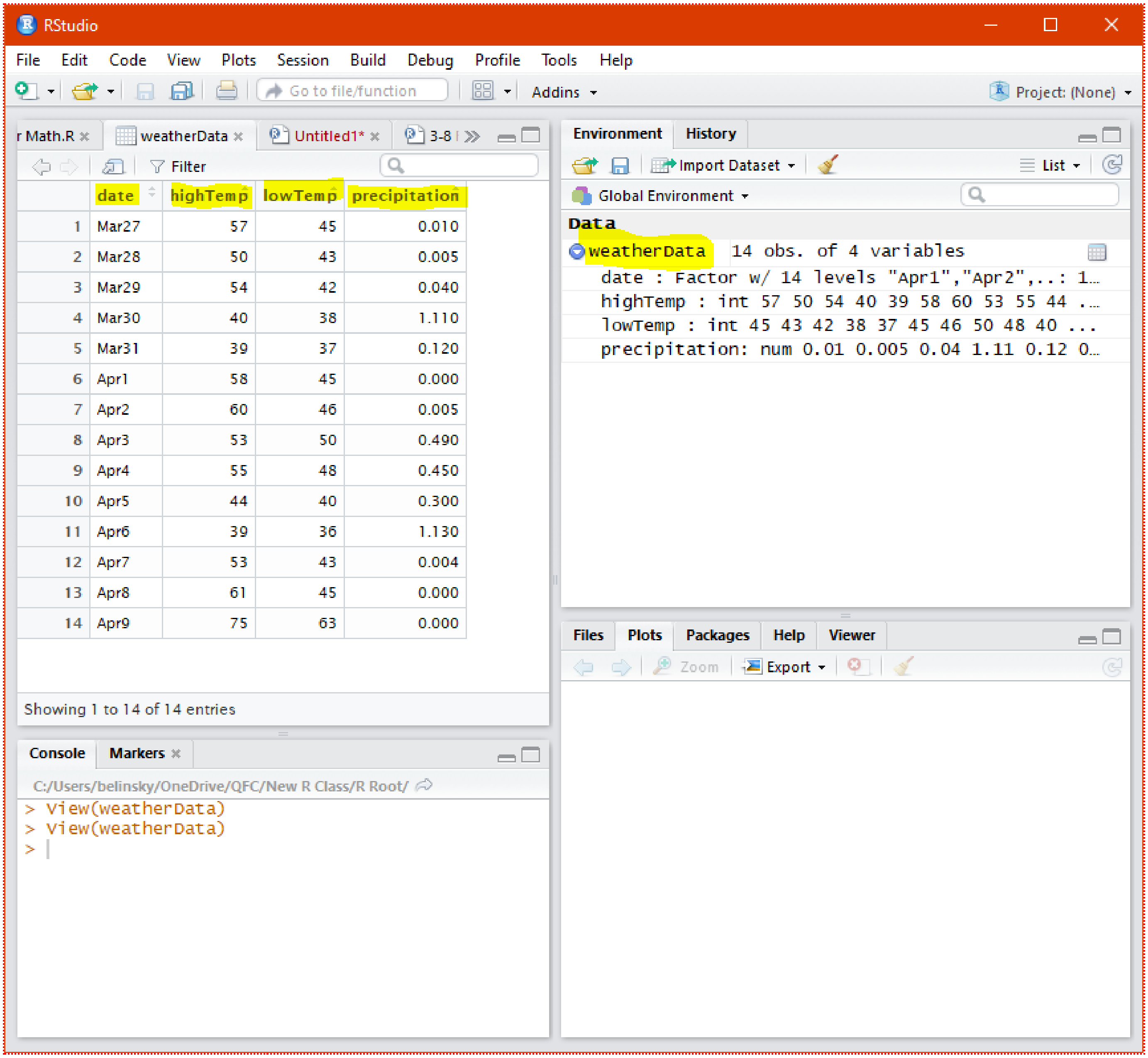
If you have any questions about the material in this lesson or your Class Project *feel free to email them to the instructor here*.

3 - Vector math

For this lesson, we are going to use the data from the file **LansingWeather2.csv**. First we need to open the CSV file and save the data in the file to a data frame, which we will call ***weatherData***.

```
{
  rm(list=ls());  options(show.error.locations = TRUE);

  # read data from LansingWeather2.csv and save to the variable weatherData...
  weatherData = read.csv("data/LansingWeather2.csv");
}
```

The data frame, ***weatherData***, has the date, high temperature, low temperature, and precipitation for 14 days (March 27-April 9, 2017) in Lansing, MI.



	date	highTemp	lowTemp	precipitation
1	Mar27	57	45	0.010
2	Mar28	50	43	0.005
3	Mar29	54	42	0.040
4	Mar30	40	38	1.110
5	Mar31	39	37	0.120
6	Apr1	58	45	0.000
7	Apr2	60	46	0.005
8	Apr3	53	50	0.490
9	Apr4	55	48	0.450
10	Apr5	44	40	0.300
11	Apr6	39	36	1.130
12	Apr7	53	43	0.004
13	Apr8	61	45	0.000
14	Apr9	75	63	0.000

Showing 1 to 14 of 14 entries

```
Data
weatherData    14 obs. of 4 variables
   date : Factor w/ 14 levels "Apr1","Apr2",..: 1…
   highTemp : int 57 50 54 40 39 58 60 53 55 44 ….
   lowTemp : int 45 43 42 38 37 45 46 50 48 40 ...
   precipitation: num 0.01 0.005 0.04 1.11 0.12 0…
```

```
C:/Users/belinsky/OneDrive/QFC/New R Class/R Root/
> View(weatherData)
> View(weatherData)
> 
```

Fig 1: Looking at the values in the ***weatherData*** data frame in the Main Window.

3.1 - Finding change in temperature

We want to find the temperature range for each day -- in other words, the high temperature minus the low temperature. This means subtracting each day's low temperature from the high temperature and saving that value to a vector.

First we will extract the low and high temperature values from ***weatherData*** and save them to a vector. There are two ways we can do this:
1) by column number

```
# get all values from the 3rd column in weatherData
highTemp = weatherData[, 3];
```

2) or by column name

```
# get all values from the column named "lowTemp" in weatherData
lowTemp = weatherData[, "lowTemp"];
```

Note: the row number is left blank to indicate we are getting values *from all rows.*

We will use a ***for()*** to iterate through each value (i.e., the highTemp - lowTemp for each day) and solve for the temperature difference. To do this we also need a vector that will hold the change in temperature values, which we will call ***changeInTemp***.

```
changeInTemp = c();  #declared a vector
```

changeInTemp acts as a state variable because:
1) ***changeInTemp*** gets initialized before the ***for()***
2) ***changeInTemp*** gets populated during the iterations of the ***for()***
3) changeInTemp final state is the full set of temperature differences

3.2 - Subtracting values from two different vectors

We are going to use a ***for()*** loop that iterates through each value in ***highTemp*** and ***lowTemp***, finds the difference, and saves the answer to ***changeInTemp***. We need to know the length of the vectors to do this. In this case, we already know there are **14** values, but we will use ***length()*** to calculate this value so that it works for vectors of any size (i.e., if we wanted to run the same code but use more days).

```
vectorLength = length(lowTemp);
```

Note: since the length of all columns in a data frame are by definition the same, we only need to get the length of one vector.

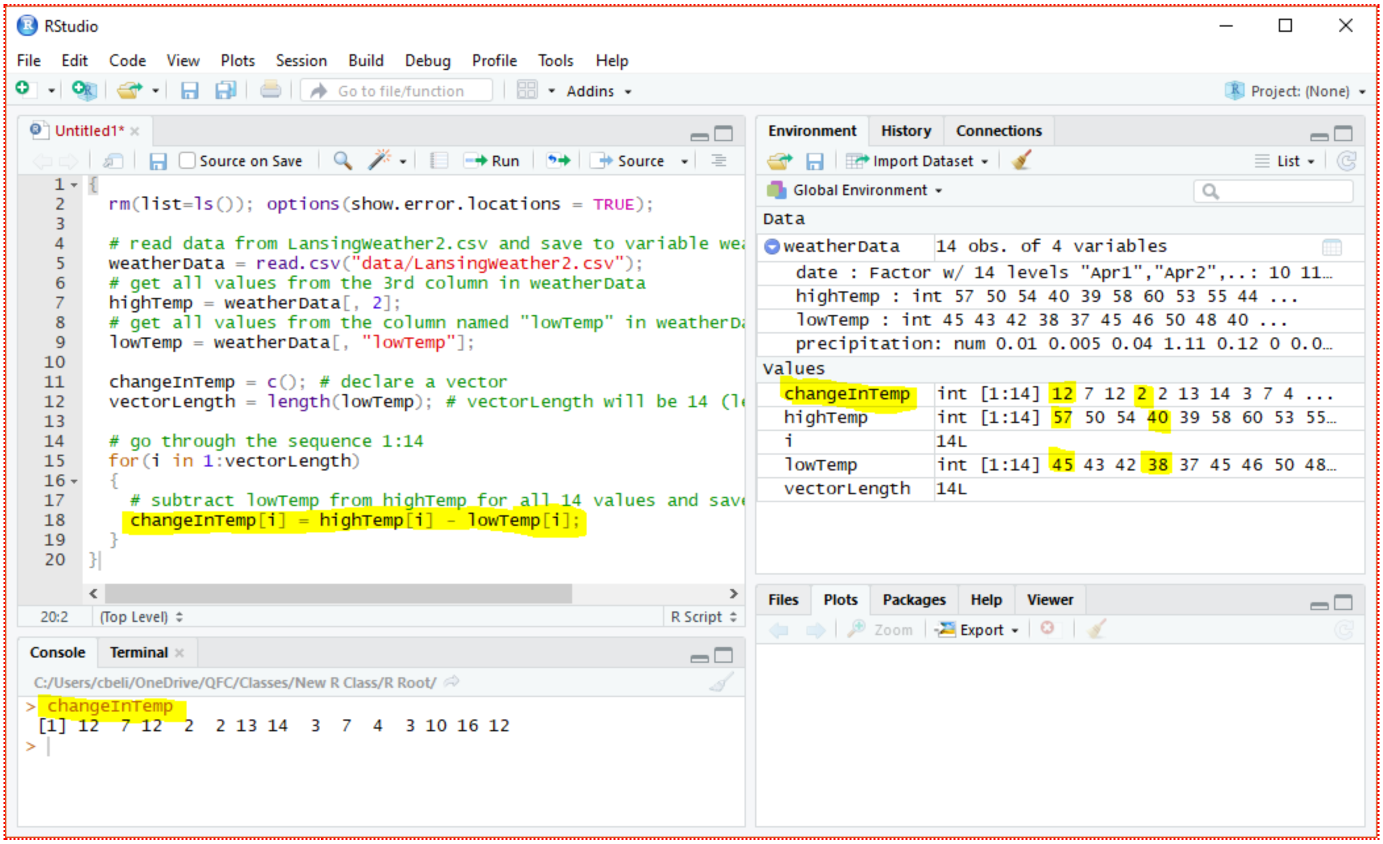
We use ***vectorLength*** to create a sequence that the ***for()*** iterates through. The ***for()*** iterates through the sequence **1:vectorLength**, **(1:14)**, and assigns the value of the **14** ***highTemp*** - ***lowTemp*** difference operations to the **14** ***changeInTemp*** vector elements.

```
for(i in 1:vectorLength)
{
  changeInTemp[i] = highTemp[i] - lowTemp[i];
}
```

Let's put all this code together:

```
{
  rm(list=ls()); options(show.error.locations = TRUE);

  # read data from LansingWeather2.csv and save to variable weatherData
  weatherData = read.csv("data/LansingWeather2.csv");
  # get all values from the 3rd column in weatherData
  highTemp = weatherData[, 2];
  # get all values from the column named "lowTemp" in weatherData
  lowTemp = weatherData[, "lowTemp"];

  changeInTemp = c(); # declare a vector
  vectorLength = length(lowTemp); # vectorLength will be 14 (length of data)

  # go through the sequence 1:14
  for(i in 1:vectorLength)
  {
     # subtract lowTemp from highTemp for all 14 values and save to changeInTemp
     changeInTemp[i] = highTemp[i] - lowTemp[i];
  }
}
```

The Environment Window shows the vector ***changeInTemp*** but it only shows up to 10 values. If you want to see all the values in ***changeInTemp***, you can type *changeInTemp* in the Console Window and the 14 values in ***changeInTemp*** will appear on the next line.



```
{
  rm(list=ls()); options(show.error.locations = TRUE);

  # read data from LansingWeather2.csv and save to variable wea
  weatherData = read.csv("data/LansingWeather2.csv");
  # get all values from the 3rd column in weatherData
  highTemp = weatherData[, 2];
  # get all values from the column named "lowTemp" in weatherDa
  lowTemp = weatherData[, "lowTemp"];

  changeInTemp = c(); # declare a vector
  vectorLength = length(lowTemp); # vectorLength will be 14 (le

  # go through the sequence 1:14
  for(i in 1:vectorLength)
  {
    # subtract lowTemp from highTemp for all 14 values and save
    changeInTemp[i] = highTemp[i] - lowTemp[i];
  }
}
```

```
Data
weatherData        14 obs. of 4 variables
  date : Factor w/ 14 levels "Apr1","Apr2",..: 10 11…
  highTemp : int 57 50 54 40 39 58 60 53 55 44 ...
  lowTemp : int 45 43 42 38 37 45 46 50 48 40 ...
  precipitation: num 0.01 0.005 0.04 1.11 0.12 0 0.0…
Values
 changeInTemp      int [1:14] 12 7 12 2 2 13 14 3 7 4 ...
 highTemp          int [1:14] 57 50 54 40 39 58 60 53 55…
 i                 14L
 lowTemp           int [1:14] 45 43 42 38 37 45 46 50 48…
 vectorLength      14L
```

```
C:/Users/cbeli/OneDrive/QFC/Classes/New R Class/R Root/
> changeInTemp
 [1] 12  7 12  2  2 13 14  3  7  4  3 10 16 12
> 
```

*Fig 2: Using a **for()** to perform iterative mathematical operations on vectors.*

4 - Dealing with missing values

In the previous example, we have an idealized situation where every day had a high and low temperature associated with it. In the real world, especially with large amounts of data, there are often missing data. In R, missing data are represented by ***NA,*** but the CSV file could designate missing data according to a number of different conventions.

First lets create a data set with missing values and save this as ***MissingTemps.csv*** in the ***Data*** directory

```
date,highTemp, lowTemp, precipitation
Mar27,57,45,0.01
Mar28,50,43,0.005
Mar29,54, ,0.04
Mar30,40,38,1.11
Mar31,39,NA,0.12
Apr1,58,45,0
Apr2,60, ,0.005
Apr3,53,50,0.49
Apr4,55,48,0.45
Apr5,44,40,0.30
Apr6,39,36,1.13
Apr7,NULL,43,0.004
Apr8,61,45,0
Apr9,75,63,0
```

In the above data set, the ***lowTemp*** for **Mar29** and **Apr2** are left blank, the ***lowTemp*** for **Mar31** is given as ***NA,*** and the ***highTemp*** for **Apr7** is given as ***NULL***. So, there are four missing values in ***missingTemps.csv***.

4.1 - Viewing blank or NA values

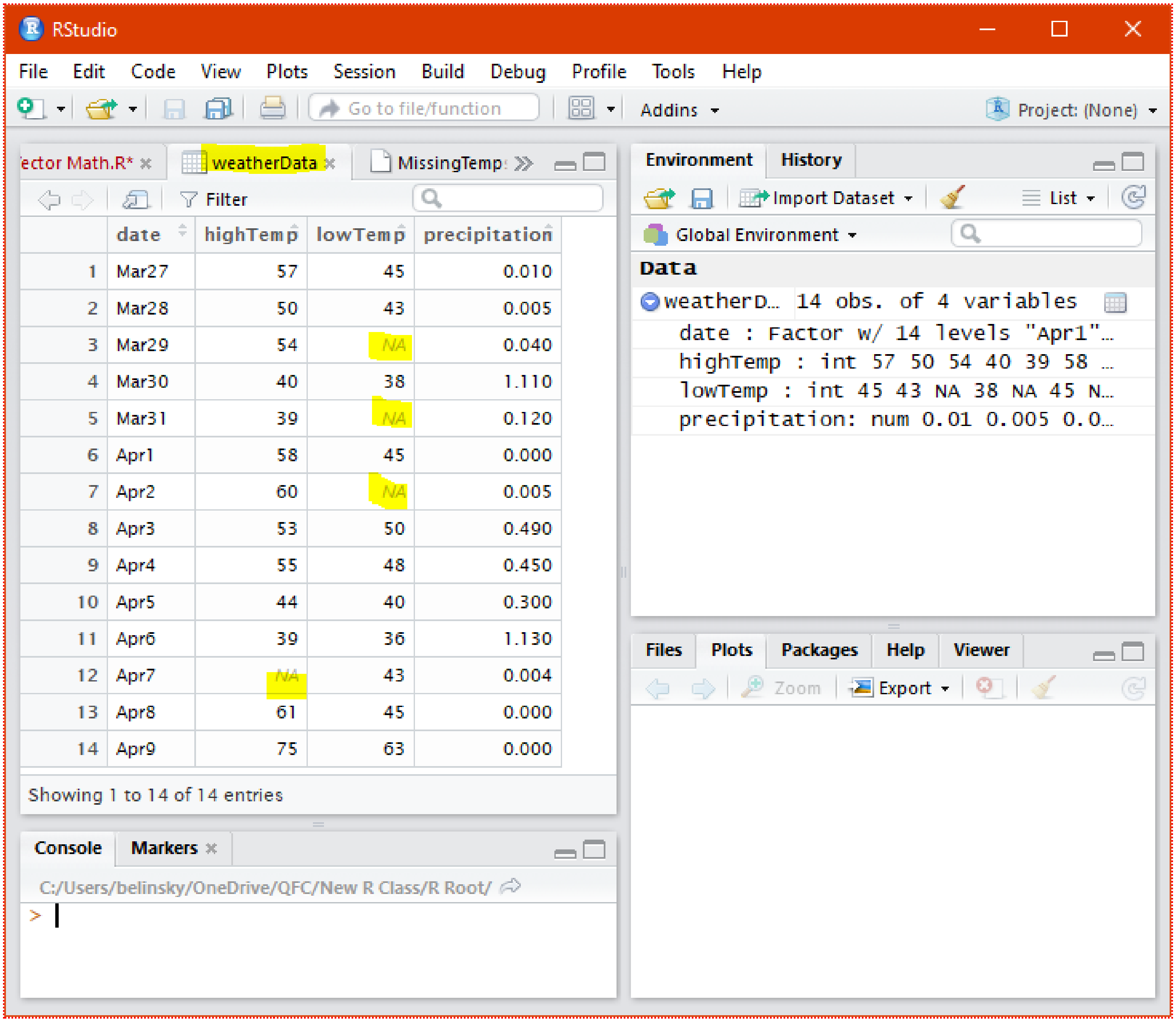
We are going to open a new script and save ***missingTemps.csv*** data to a Data Frame called ***weatherData,*** but we want to standardize the way that missing values are recorded. So, in ***read.csv()*** we add the parameter ***na.strings***:

```
na.strings=c("", " ", "NULL", NULL, "NA", "na", NA, "null")
```

Essentially, ***na.strings*** is a vector that contains all the value that you want R to assign as ***NA***. It is good to be paranoid here and think of all the possible ways in which CSV files will present these values! In this case, we only needed to use " ", "NULL", and "NA".

```
{
  rm(list=ls()); options(show.error.locations = TRUE);

  weatherData = read.csv("data/missingTemps.csv",
                 na.strings=c("", " ", "NULL", NULL, "NA", "na", NA, "null"));
}
```

Execute the code above and double-click on ***weatherData*** in the ***Environment Window*** so that it appears in the Main Window. We see that all four missing values in ***weatherData*** are now labelled as ***NA***.



	date	highTemp	lowTemp	precipitation
1	Mar27	57	45	0.010
2	Mar28	50	43	0.005
3	Mar29	54	NA	0.040
4	Mar30	40	38	1.110
5	Mar31	39	NA	0.120
6	Apr1	58	45	0.000
7	Apr2	60	NA	0.005
8	Apr3	53	50	0.490
9	Apr4	55	48	0.450
10	Apr5	44	40	0.300
11	Apr6	39	36	1.130
12	Apr7	NA	43	0.004
13	Apr8	61	45	0.000
14	Apr9	75	63	0.000

Showing 1 to 14 of 14 entries

```
Data
weatherD…  14 obs. of 4 variables
  date : Factor w/ 14 levels "Apr1"…
  highTemp : int 57 50 54 40 39 58 …
  lowTemp : int 45 43 NA 38 NA 45 N…
  precipitation: num 0.01 0.005 0.0…
```

```
C:/Users/belinsky/OneDrive/QFC/New R Class/R Root/
>
```

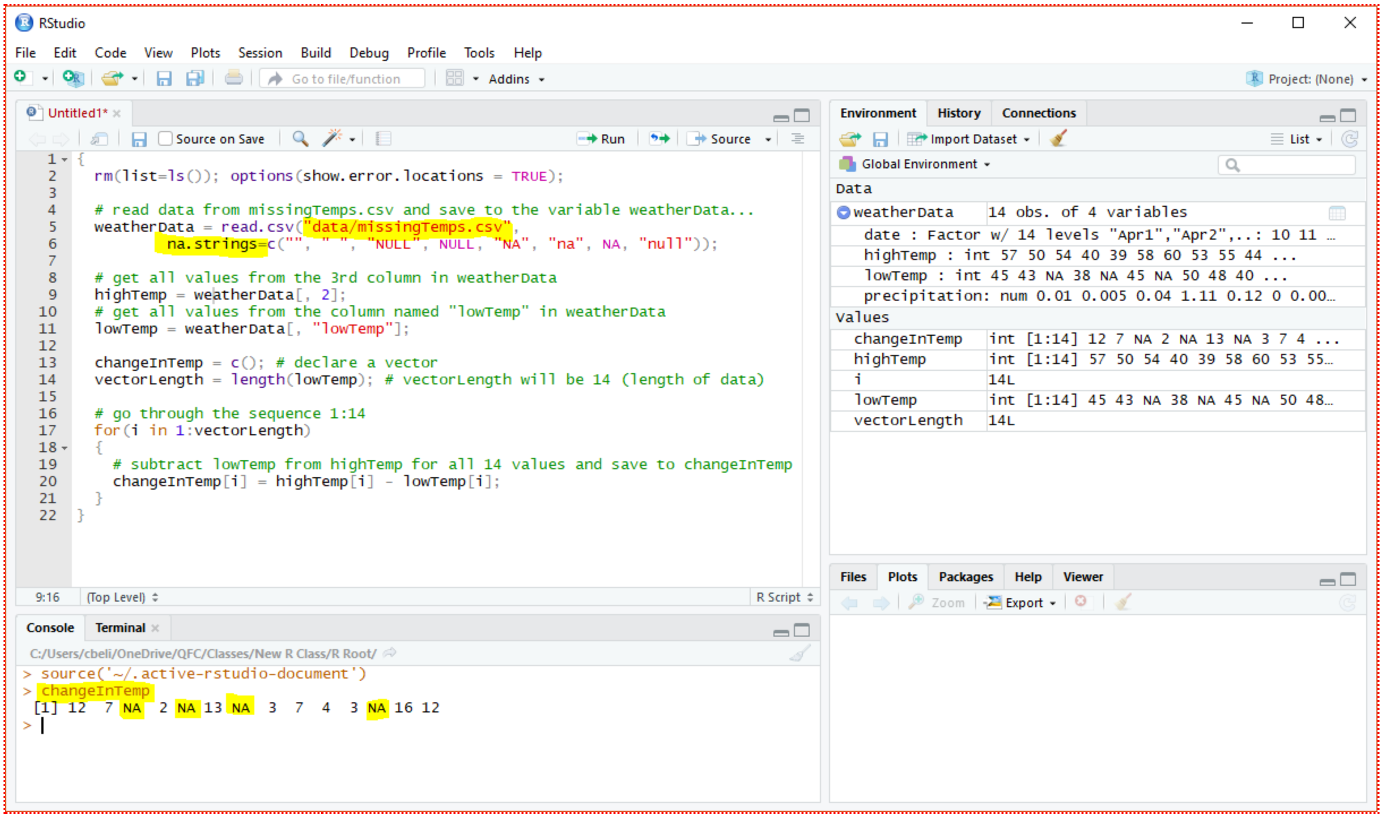
*Fig 3: **NA** values in a data frame*

4.2 - Mathematical operations on vectors with NA values

We are going to run the same change in temperature script as before except with NA values in the data frame:

```
{
  rm(list=ls()); options(show.error.locations = TRUE);


  # read data from missingTemps.csv and save to the variable weatherData...
  weatherData = read.csv("data/missingTemps.csv",
       na.strings=c("", " ", "NULL", NULL, "NA", "na", NA, "null"));

  # get all values from the 3rd column in weatherData
  highTemp = weatherData[, 2];
  # get all values from the column named "lowTemp" in weatherData
  lowTemp = weatherData[, "lowTemp"];

  changeInTemp = c(); # declare a vector
  vectorLength = length(lowTemp); # vectorLength will be 14 (length of data)

  # go through the sequence 1:14
  for(i in 1:vectorLength)
  {
    # subtract lowTemp from highTemp for all 14 values and save to changeInTemp
    changeInTemp[i] = highTemp[i] - lowTemp[i];
  }
}
```

Note: anytime there is an ***NA*** in a calculation (e.g., 3rd element in ***changeInTemp***) , the answer is going to be ***NA***. This is a special feature of ***NA* --** if you used any other value to represent missing data then you would get an error in the calculation.



```
{
  rm(list=ls()); options(show.error.locations = TRUE);

  # read data from missingTemps.csv and save to the variable weatherData...
  weatherData = read.csv("data/missingTemps.csv",
          na.strings=c("", " ", "NULL", NULL, "NA", "na", NA, "null"));

  # get all values from the 3rd column in weatherData
  highTemp = weatherData[, 2];
  # get all values from the column named "lowTemp" in weatherData
  lowTemp = weatherData[, "lowTemp"];

  changeInTemp = c(); # declare a vector
  vectorLength = length(lowTemp); # vectorLength will be 14 (length of data)

  # go through the sequence 1:14
  for(i in 1:vectorLength)
  {
    # subtract lowTemp from highTemp for all 14 values and save to changeInTemp
    changeInTemp[i] = highTemp[i] - lowTemp[i];
  }
}
```

```
Data
weatherData        14 obs. of 4 variables
  date : Factor w/ 14 levels "Apr1","Apr2",..: 10 11 …
  highTemp : int 57 50 54 40 39 58 60 53 55 44 ...
  lowTemp : int 45 43 NA 38 NA 45 NA 50 48 40 ...
  precipitation: num 0.01 0.005 0.04 1.11 0.12 0 0.00…
Values
 changeInTemp      int [1:14] 12 7 NA 2 NA 13 NA 3 7 4 ...
 highTemp          int [1:14] 57 50 54 40 39 58 60 53 55…
 i                 14L
 lowTemp           int [1:14] 45 43 NA 38 NA 45 NA 50 48…
 vectorLength      14L
```

```
C:/Users/cbeli/OneDrive/QFC/Classes/New R Class/R Root/
> source('~/.active-rstudio-document')
> changeInTemp
 [1] 12  7 NA  2 NA 13 NA  3  7  4  3 NA 16 12
> 
```

*Fig 4: Doing mathematical operation with **NA** values*

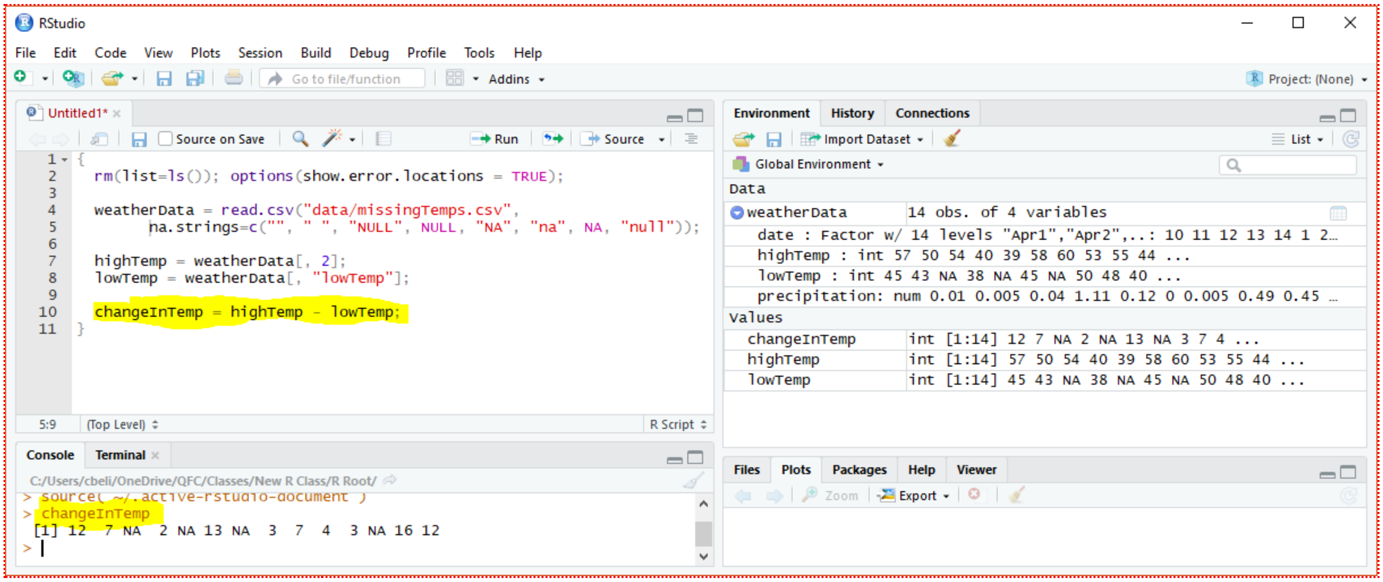
Extension: What counts as NA

5 - Vector Operation Shortcuts

So far we have used ***for()*** loops to perform iterative operations on a vector or multiple vectors. However, R has built in functions that do all these operations using a lot less code. For example, we can perform the above change in temperature calculation without a ***for().***

```
 1 {
 2   rm(list=ls()); options(show.error.locations = TRUE);
 3 
 4   weatherData = read.csv("data/missingTemps.csv",
 5          na.strings=c("", " ", "NULL", NULL, "NA", "na", NA, "null"));
 6 
 7   highTemp = weatherData[, 2];
 8   lowTemp = weatherData[, "lowTemp"];
 9 
10   changeInTemp = highTemp - lowTemp;
11 }
```

Line 9 does all the work of iterating through the values in the vector, subtracting the values, and saving the answer to the vector ***changeInTemp***. The result is the same as the script that uses the ***for()*** loop.



```
 1 {
 2   rm(list=ls()); options(show.error.locations = TRUE);
 3 
 4   weatherData = read.csv("data/missingTemps.csv",
 5          na.strings=c("", " ", "NULL", NULL, "NA", "na", NA, "null"));
 6 
 7   highTemp = weatherData[, 2];
 8   lowTemp = weatherData[, "lowTemp"];
 9 
10   changeInTemp = highTemp - lowTemp;
11 }
```

```
Data
weatherData      14 obs. of 4 variables
  date : Factor w/ 14 levels "Apr1","Apr2",..: 10 11 12 13 14 1 2…
  highTemp : int  57 50 54 40 39 58 60 53 55 44 ...
  lowTemp : int  45 43 NA 38 NA 45 NA 50 48 40 ...
  precipitation: num  0.01 0.005 0.04 1.11 0.12 0 0.005 0.49 0.45 …
Values
  changeInTemp   int [1:14] 12 7 NA 2 NA 13 NA 3 7 4 ...
  highTemp       int [1:14] 57 50 54 40 39 58 60 53 55 44 ...
  lowTemp        int [1:14] 45 43 NA 38 NA 45 NA 50 48 40 ...
```

```
C:/Users/cbeli/OneDrive/QFC/Classes/New R Class/R Root/
> source( ~/.active-rstudio-document )
> changeInTemp
 [1] 12  7 NA  2 NA 13 NA  3  7  4  3 NA 16 12
> |
```

Fig 5: Subtracting a vector in R.

This method is obviously easier than ***for()*** loops for subtracting two vectors -- and it is the method R programmer will usually use. However, this method is really a shortcut and it is important to understand what is going on in with the ***for()*** example because, eventually, you will find a situation where you need to use a ***for()*** to solve a more complex problem.

5.1 - Many other shortcuts

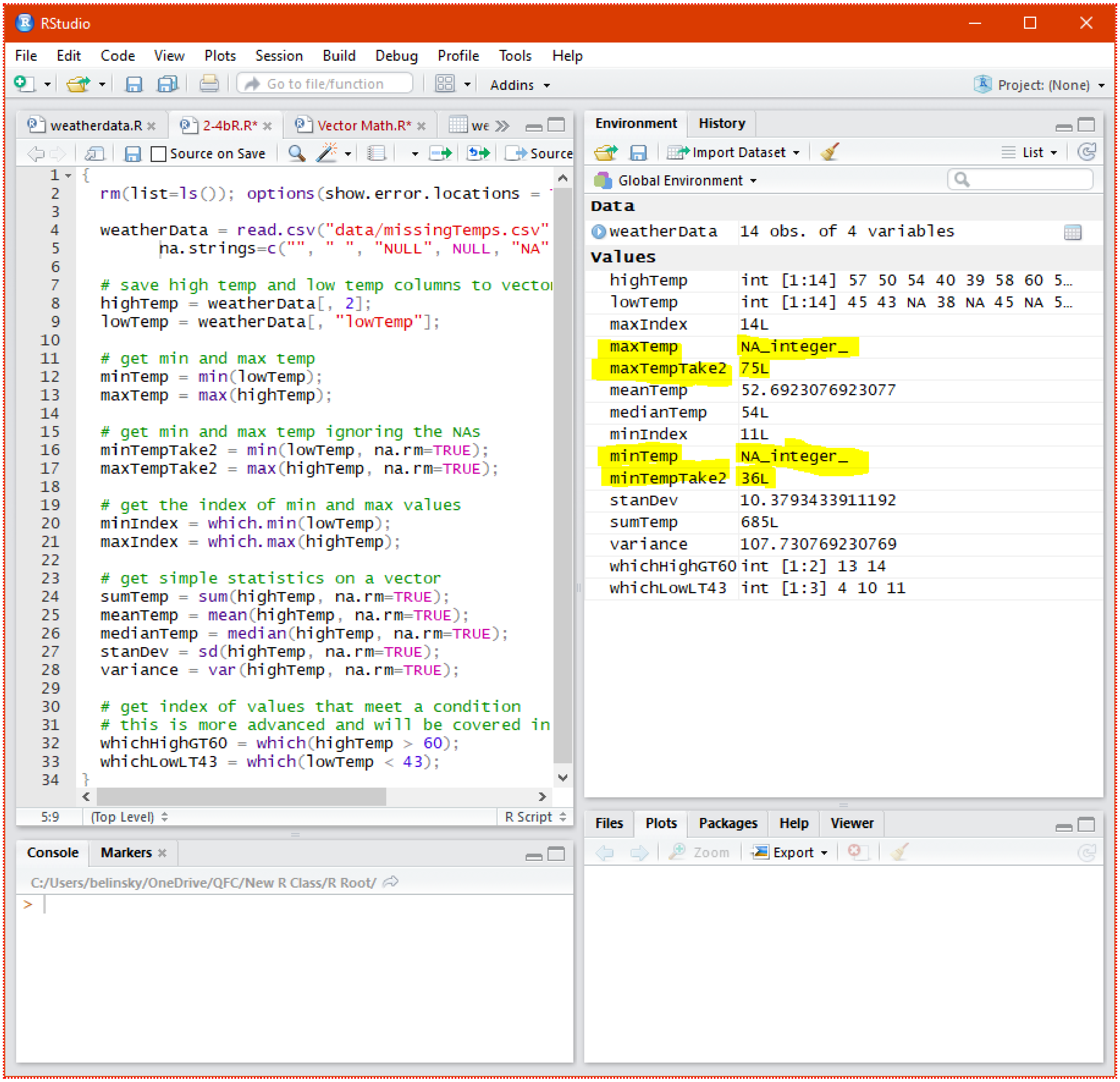
R has many functions that can quickly perform the most common operations on vectors like ***sum(), max(), min(),*** and ***mean()***.

One important parameter used in all of these functions is ***na.rm:***

na.rm = TRUE tells R to exclude the ***NA*** values from the vector before performing the operation.
na.rm = FALSE tells R to return ***NA*** if there are ***NA*** values in the vector.

The default value for ***na.rm*** is ***FALSE***.

```
{
  rm(list=ls()); options(show.error.locations = TRUE);

  weatherData = read.csv("data/missingTemps.csv",
          na.strings=c("", " ", "NULL", NULL, "NA", "na", NA, "null"))

  # save high temp and low temp columns to vectors
  highTemp = weatherData[, 2];
  lowTemp = weatherData[, "lowTemp"];

  # get min and max temp (na.rm = default = FALSE)
  minTemp = min(lowTemp);
  maxTemp = max(highTemp);

  # get min and max temp ignoring the NAs
  minTempTake2 = min(lowTemp, na.rm=TRUE);
  maxTempTake2 = max(highTemp, na.rm=TRUE);

  # get the index of min and max values
  minIndex = which.min(lowTemp);
  maxIndex = which.max(highTemp);

  # get simple statistics on a vector
  sumTemp = sum(highTemp, na.rm=TRUE);
  meanTemp = mean(highTemp, na.rm=TRUE);
  medianTemp = median(highTemp, na.rm=TRUE);
  stanDev = sd(highTemp, na.rm=TRUE);
  variance = var(highTemp, na.rm=TRUE);


  # get index of values that meet a condition
  # this is more advanced and will be covered in detail later
  whichHighGT60 = which(highTemp > 60);
  whichLowLT43 = which(lowTemp < 43);
}
```



```
{
  rm(list=ls()); options(show.error.locations = 
  
  weatherData = read.csv("data/missingTemps.csv"
        na.strings=c("", " ", "NULL", NULL, "NA"
  
  # save high temp and low temp columns to vecto
  highTemp = weatherData[, 2];
  lowTemp = weatherData[, "lowTemp"];
  
  # get min and max temp
  minTemp = min(lowTemp);
  maxTemp = max(highTemp);
  
  # get min and max temp ignoring the NAs
  minTempTake2 = min(lowTemp, na.rm=TRUE);
  maxTempTake2 = max(highTemp, na.rm=TRUE);
  
  # get the index of min and max values
  minIndex = which.min(lowTemp);
  maxIndex = which.max(highTemp);
  
  # get simple statistics on a vector
  sumTemp = sum(highTemp, na.rm=TRUE);
  meanTemp = mean(highTemp, na.rm=TRUE);
  medianTemp = median(highTemp, na.rm=TRUE);
  stanDev = sd(highTemp, na.rm=TRUE);
  variance = var(highTemp, na.rm=TRUE);
  
  # get index of values that meet a condition
  # this is more advanced and will be covered in
  whichHighGT60 = which(highTemp > 60);
  whichLowLT43 = which(lowTemp < 43);
}
```

5:9

```
Data
weatherData     14 obs. of 4 variables
Values
highTemp        int [1:14] 57 50 54 40 39 58 60 5…
lowTemp         int [1:14] 45 43 NA 38 NA 45 NA 5…
maxIndex        14L
maxTemp         NA_integer_
maxTempTake2    75L
meanTemp        52.6923076923077
medianTemp      54L
minIndex        11L
minTemp         NA_integer_
minTempTake2    36L
stanDev         10.3793433911192
sumTemp         685L
variance        107.730769230769
whichHighGT60   int [1:2] 13 14
whichLowLT43    int [1:3] 4 10 11
```

```
C:/Users/belinsky/OneDrive/QFC/New R Class/R Root/
> 
```

Fig 6: Using R functions to perform common vector operations.

6 - Application

If you have any questions regarding this application or your Class Project, feel free to email them to the instructor here. You can attach the whole Root Folder as a zipped file.

1) Create a vector called ***changeInHighTemp*** and, using the data from **lansingWeather2.csv**, find the change in high temperatures from day-to-day.

So, if the four high temperatures are: **40, 45, 35, 42**
then the three ***changeInHighTemp*** would be: (45-40), (35-45), and (42-35) = **5, -10, 7**

2) There will be one less value in ***changeInHighTemp*** than in ***lansingWeather2.csv***. We want the length of ***changeInHighTemp*** to be the same so we can add it to the data frame (remember, all vectors must be the

same size in a data frame). To do this, add an ***NA*** value as the first value in ***changeInHighTemp***. The ***NA*** indicates that there was no previous value to subtract and, hence, this value is not applicable.

3) Add the vector ***changeInHighTemp*** (with the **NA** as the first value) to the data frame.

*Save you script file as **app2-5.r** in the **scripts** folder of your RStudio Project for the class.*

7 - Extension: What counts as NA

R is somewhat savvy when it moves data from a CSV to a data frame. If there is an empty value in a column that is all numbers, then R will convert the empty value to ***NA***. ***NULL***, however, is not considered to be a valid data frame value and if there is a ***NULL*** in a column, it will be treated as the string value "NULL". R also does not consider lowercase ***na*** to be ***NA*** and this will also be treated as the string value "na".

There is now a second problem: if there is even one string in a column, all values in the column are treated as strings. This means that **40** will be seen as the string "40" and mathematical operations on strings will cause an error. Thus, specifying the ***na.strings*** argument can be useful to ensure that you will nor have to convert columns to numeric using an extra step after the data are imported.