Loading Data

```
westNile <- read.csv("./data/SwaddleWestNile2002NCEAS-BAD.csv")
```

Loading Data

```
westNile <- read.csv("./data/SwaddleWestNile2002NCEAS-BAD.csv")
```

Note:

- File path (./ is this directory, ../ is back one directory)
- Quotes
- Our data is now an object in R

Look at Your Data

```
head(westNile)
```

```
##   State                      Infected.County WNV.incidence
## 1    AL       Autauga, AL                            2.290
## 2    AL     Calhoun , AL                             0.891
## 3    AL Chambers, AL                                 2.734
## 4    AL Dallas , AL                                  2.157
## 5    AL Marengo , AL                                 8.874
## 6    AL Marion, AL                                   3.204
##   Species.Richness Corvid.Abundance
## 1               66                8
## 2               67               64
## 3               41               69
## 4               60               66
## 5               69               64
## 6               NA    NOT AVAILABLE
```

Look at Columns 3 through 4

```
head(westNile[, 3:4])

##   WNV.incidence Species.Richness
## 1         2.290                66
## 2         0.891                67
## 3         2.734                41
## 4         2.157                60
## 5         8.874                69
## 6         3.204                NA
```

Look at Columns 3 through 4

```
head(westNile[, 3:4])

##   WNV.incidence Species.Richness
## 1         2.290                66
## 2         0.891                67
## 3         2.734                41
## 4         2.157                60
## 5         8.874                69
## 6         3.204                NA
```

- Data Frame is treates as a Matrix.
- $[rows, columns]$

Look at Your Individual Columns

```
names(westNile)

## [1] "State"             "Infected.County"  "WNV.incidence"
## [4] "Species.Richness" "Corvid.Abundance"
```

Look at Your Individual Columns

```
names(westNile)

## [1] "State"            "Infected.County"  "WNV.incidence"
## [4] "Species.Richness" "Corvid.Abundance"
```

(Note that spaces are now .s)

Look at Your Individual Columns

```
names(westNile)

## [1] "State"            "Infected.County"  "WNV.incidence"
## [4] "Species.Richness" "Corvid.Abundance"
```

(Note that spaces are now .s)

```
westNile$Species.Richness

##   [1] 66 67 41 60 69 NA 56 65 54 52 81 51 47 59 49 51 72 53
##  [19] 54 49 61 81 62 70 71 57 87 64 50 62 71 70 59 63 58 51
##  [37] 46 66 53 59 58 56 58 43 65 51 51 63 54 60 53 39 62 67
##  [55] 68 82 70 76 58 60 72 59 72 62 82 63 68 39 67 66 63 47
##  [73] 59 61 65 79 54 56 30 48 56 68 58 42 51 64 73 55 61 65
##  [91] 61 74 65 61 51 93 42 63 68 58 68 61 56 60 81 66 53 49
## [109] 68 72 76 57 76 55 76 56 73 59 73 57 90 50 73 64 78 75
## [127] 61 80 59 69
```

Missing Data is NA

```
westNile$Species.Richness

##   [1] 66 67 41 60 69 NA 56 65 54 52 81 51 47 59 49 51 72 53
##  [19] 54 49 61 81 62 70 71 57 87 64 50 62 71 70 59 63 58 51
##  [37] 46 66 53 59 58 56 58 43 65 51 51 63 54 60 53 39 62 67
##  [55] 68 82 70 76 58 60 72 59 72 62 82 63 68 39 67 66 63 47
##  [73] 59 61 65 79 54 56 30 48 56 68 58 42 51 64 73 55 61 65
##  [91] 61 74 65 61 51 93 42 63 68 58 68 61 56 60 81 66 53 49
## [109] 68 72 76 57 76 55 76 56 73 59 73 57 90 50 73 64 78 75
## [127] 61 80 59 69
```

Note the NA. This is missing data.

Missing Data is NA

```
westNile$Species.Richness
```

```
##   [1] 66 67 41 60 69 NA 56 65 54 52 81 51 47 59 49 51 72 53
##  [19] 54 49 61 81 62 70 71 57 87 64 50 62 71 70 59 63 58 51
##  [37] 46 66 53 59 58 56 58 43 65 51 51 63 54 60 53 39 62 67
##  [55] 68 82 70 76 58 60 72 59 72 62 82 63 68 39 67 66 63 47
##  [73] 59 61 65 79 54 56 30 48 56 68 58 42 51 64 73 55 61 65
##  [91] 61 74 65 61 51 93 42 63 68 58 68 61 56 60 81 66 53 49
## [109] 68 72 76 57 76 55 76 56 73 59 73 57 90 50 73 64 78 75
## [127] 61 80 59 69
```

Note the NA. This is missing data.

```
westNile$Species.Richness[6]
```

```
## [1] NA
```

Let's look at another

```
westNile$Corvid.Abundance
```

```
##   [1] 8              64             69
##   [4] 66             64             NOT AVAILABLE
##   [7] 59             129            54
##  [10] 100            62             82
##  [13] 102            35             31
##  [16] 13             51             60
##  [19] 10             87             53
##  [22] 9999           34             86
##  [25] 75             102            216
##  [28] 71             43             57
##  [31] 98             84             44
##  [34] 109            165            44
##  [37] 68             48             34
##  [40] 63             9999           52
##  [43] 24             39             41
##  [46] 32             47             23
##  [49] 135            40             32
```

Cleaner Data

```
westNile <- read.csv("./data/SwaddleWestNile2002NCEAS-BAD.csv",
    na.strings = "NOT AVAILABLE")
```

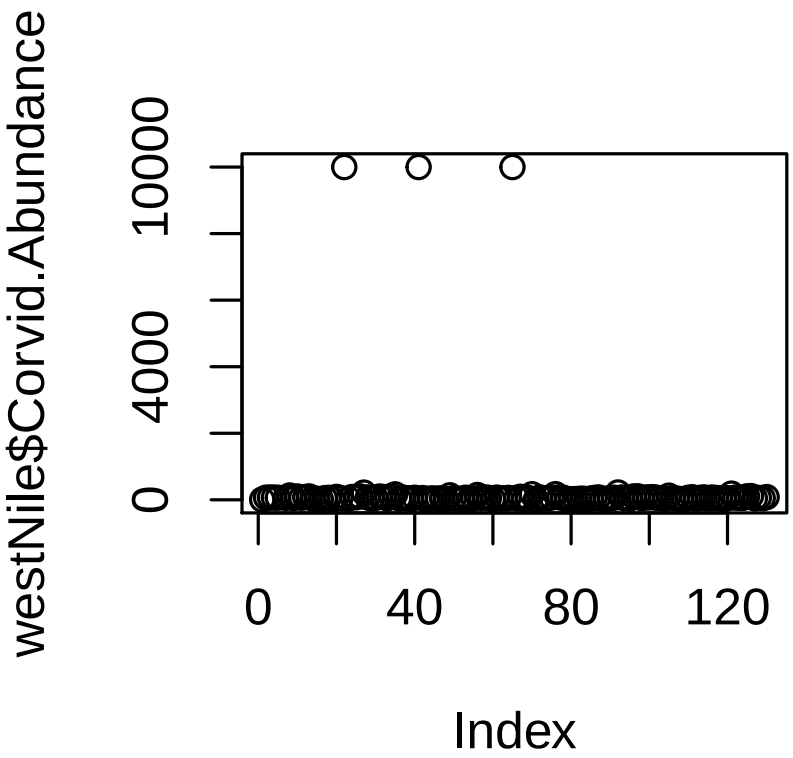
And, Fixed!

```
westNile$Corvid.Abundance

##   [1]    8.00   64.00   69.00   66.00   64.00      NA
##   [7]   59.00  129.00   54.00  100.00   62.00   82.00
##  [13]  102.00   35.00   31.00   13.00   51.00   60.00
##  [19]   10.00   87.00   53.00 9999.00   34.00   86.00
##  [25]   75.00  102.00  216.00   71.00   43.00   57.00
##  [31]   98.00   84.00   44.00  109.00  165.00   44.00
##  [37]   68.00   48.00   34.00   63.00 9999.00   52.00
##  [43]   24.00   39.00   41.00   32.00   47.00   23.00
##  [49]  135.00   49.00   32.00   27.00   63.00   15.00
##  [55]   45.00  144.00   61.00   71.00   57.00   29.00
##  [61]   66.00   36.00   46.00   57.00 9999.00   54.00
##  [67]   91.00   19.00   56.00  168.00   14.00   71.00
##  [73]   43.00   48.00   70.00  170.00   75.00   63.00
##  [79]    6.00   18.00   21.00   29.00   34.00   18.00
##  [85]   39.00   57.00   71.00   26.00   31.00   47.00
##  [91]   63.00  220.00   70.00   42.00   36.00  101.00
##  [97]  106.00   53.00   69.00   73.77   70.00   58.00
```

What about fixing many bad values?

```
plot(westNile$Corvid.Abundance)
```



What about fixing many bad values?

```
which(westNile$Corvid.Abundance == 9999)

## [1] 22 41 65
```

What about fixing many bad values?

```
which(westNile$Corvid.Abundance == 9999)

## [1] 22 41 65
```

$==$ is makes a COMPARISON and returns a logical value
Can also use $<$, $>$, and more.

What about fixing many bad values?

```
which(westNile$Corvid.Abundance == 9999)
```

```
## [1] 22 41 65
```

$==$ is makes a COMPARISON and returns a logical value
Can also use $<$, $>$, and more.

```
westNile$Corvid.Abundance == 9999
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE    NA FALSE FALSE FALSE
##  [10] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [19] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
##  [28] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [37] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
##  [46] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [55] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [64] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [82] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

Replace the 9999s

```
westNile$Corvid.Abundance[which(westNile$Corvid.Abundance ==
    9999)] <- NA
```

The which approach is often good, as once you spot a single problem observation, there may be others like it.

Exercise

1. Is everything OK with West Nile Virus Incidence?

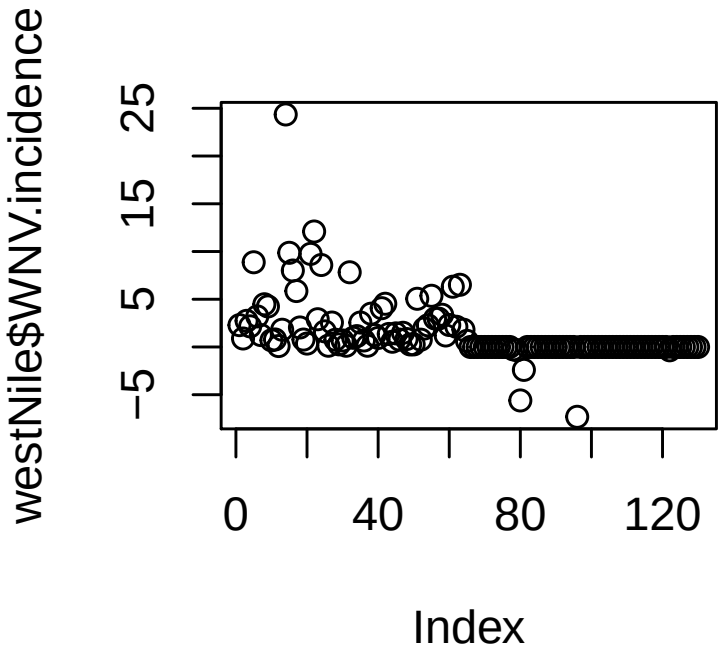
Exercise

1. Is everything OK with West Nile Virus Incidence?

2. Let's say a database overwrote some 0 values - fix these values!

The Fix

```
plot(westNile$WNV.incidence)
```




```
westNile$WNV.incidence[which(westNile$WNV.incidence < 0)] <- 0
```