Time to put it all together!

The challenge

- Historical weather data from Boston, USA
- 12 months beginning Dec 2014
- The data are dirty
    - Column names are values
    - Variables coded incorrectly
    - Missing and extreme values
    - ...
- **Clean the data!**

Understanding the structure of your data

- `class()` - Class of data object
- `dim()` - Dimensions of data
- `names()` - Column names
- `str()` - Preview of data with helpful details
- `glimpse()` - Better version of `str()` from dplyr
- `summary()` - Summary of data

Looking at your data

- `head()` - View top of dataset
- `tail()` - View bottom of dataset
- `print()` - View entire dataset (not recommended!)

Visualizing your data

- `hist()` - View histogram of a single variable
- `plot()` - View plot of two variables

Let's practice!

Let's tidy the data

Column names are values

```
> head(weather)
  X year month            measure X1 X2 X3 X4 X5 X6 X7 X8 X9 ...
1 1 2014    12  Max.TemperatureF 64 42 51 43 42 45 38 29 49 ...
2 2 2014    12 Mean.TemperatureF 52 38 44 37 34 42 30 24 39 ...
3 3 2014    12  Min.TemperatureF 39 33 37 30 26 38 21 18 29 ...
4 4 2014    12    Max.Dew.PointF 46 40 49 24 37 45 36 28 49 ...
5 5 2014    12    MeanDew.PointF 40 27 42 21 25 40 20 16 41 ...
6 6 2014    12     Min.DewpointF 26 17 24 13 12 36 -3  3 28 ...
```

Values are variable names

```
> head(weather2)
  X year month            measure day value
1 1 2014    12  Max.TemperatureF  X1    64
2 2 2014    12 Mean.TemperatureF  X1    52
3 3 2014    12  Min.TemperatureF  X1    39
4 4 2014    12    Max.Dew.PointF  X1    46
5 5 2014    12    MeanDew.PointF  X1    40
6 6 2014    12     Min.DewpointF  X1    26
```

Let's practice!

CLEANING DATA IN R

Prepare the data for analysis

Dates with lubridate

```
# Load the lubridate package
> library(lubridate)

# Experiment with basic lubridate functions
> ymd("2015-08-25")
[1] "2015-08-25 UTC"    year-month-day

> ymd("2015 August 25")
[1] "2015-08-25 UTC"      year-month-day

> mdy("August 25, 2015")
[1] "2015-08-25 UTC"      month-day-year

> hms("13:33:09")
[1] "13H 33M 9S"    hour-minute-second

> ymd_hms("2015/08/25 13.33.09")
[1] "2015-08-25 13:33:09 UTC"   year-month-day hour-minute-second
```

Type conversions

```
> as.character(2016)
[1] "2016"

> as.numeric(TRUE)
[1] 1

> as.integer(99)
[1] 99

> as.factor("something")
[1] something
Levels: something

> as.logical(0)
[1] FALSE
```

CLEANING DATA IN R

Let's practice!

Missing, extreme, and unexpected values

Finding missing values

```
# Create a small dataset
> x <- data.frame(a = c(2, 5, NA, 8),
                  b = c(NA, 34, 9, NA))

# Return data frame of TRUEs and FALSEs
> is.na(x)
         a     b
[1,] FALSE  TRUE
[2,] FALSE FALSE
[3,]  TRUE FALSE
[4,] FALSE  TRUE

# Count number of TRUEs
> sum(is.na(x))
[1] 3

# Find indices of missing values in column b
> which(is.na(x$b))
[1] 1 4
```

Identifying errors

- Context matters!
- Plausible ranges
- Numeric variables in weather data
  - Percentages (0-100)
  - Temperatures (Fahrenheit)
  - Wind speeds (miles per hour)
  - Pressures (inches of mercury)
  - Distances (miles)
  - Eighths (of cloud cover)

Let's practice!

Your data are clean!

Clean weather data

```
# View head of clean data
> head(weather6)
        date    events cloud_cover max_dew_point_f ...
1 2014-12-01      Rain           6              46 ...
2 2014-12-02 Rain-Snow           7              40 ...
3 2014-12-03      Rain           8              49 ...
4 2014-12-04      None           3              24 ...
5 2014-12-05      Rain           5              37 ...
6 2014-12-06      Rain           8              45 ...

# View tail of clean data
          date events cloud_cover max_dew_point_f ...
361 2015-11-26   None           6              49 ...
362 2015-11-27   None           7              52 ...
363 2015-11-28   Rain           8              50 ...
364 2015-11-29   None           4              33 ...
365 2015-11-30   None           6              26 ...
366 2015-12-01   Rain           7              43 ...
```

Summary of your accomplishments

- Inspected the data
- Tidied the data
- Improved date representations
- Dealt with incorrect variable codings
- Found and dealt with missing data
- Identified and corrected errors
- Visualized the result

CLEANING DATA IN R

Congratulations!