CLEANING DATA IN R

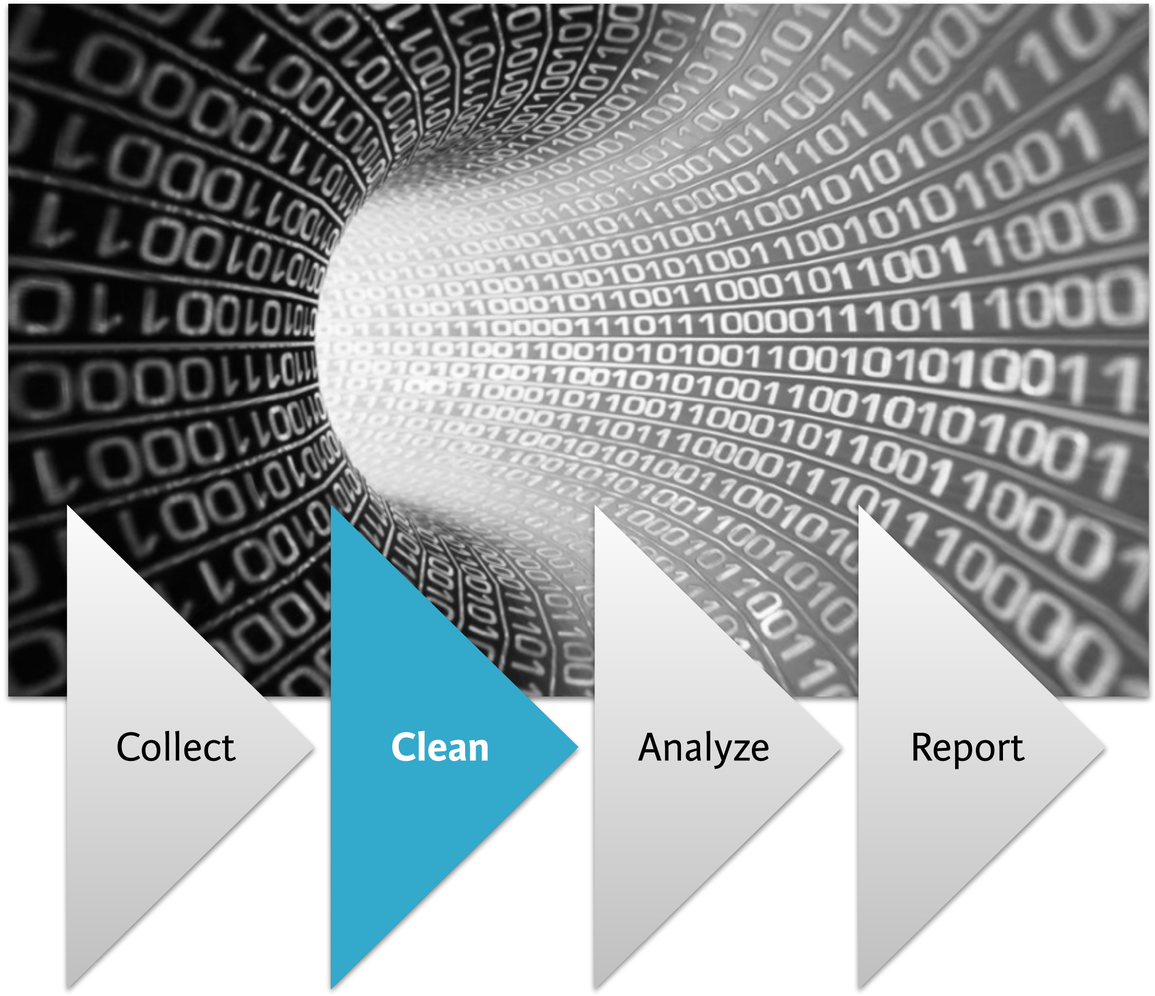
Introduction to Cleaning Data in R

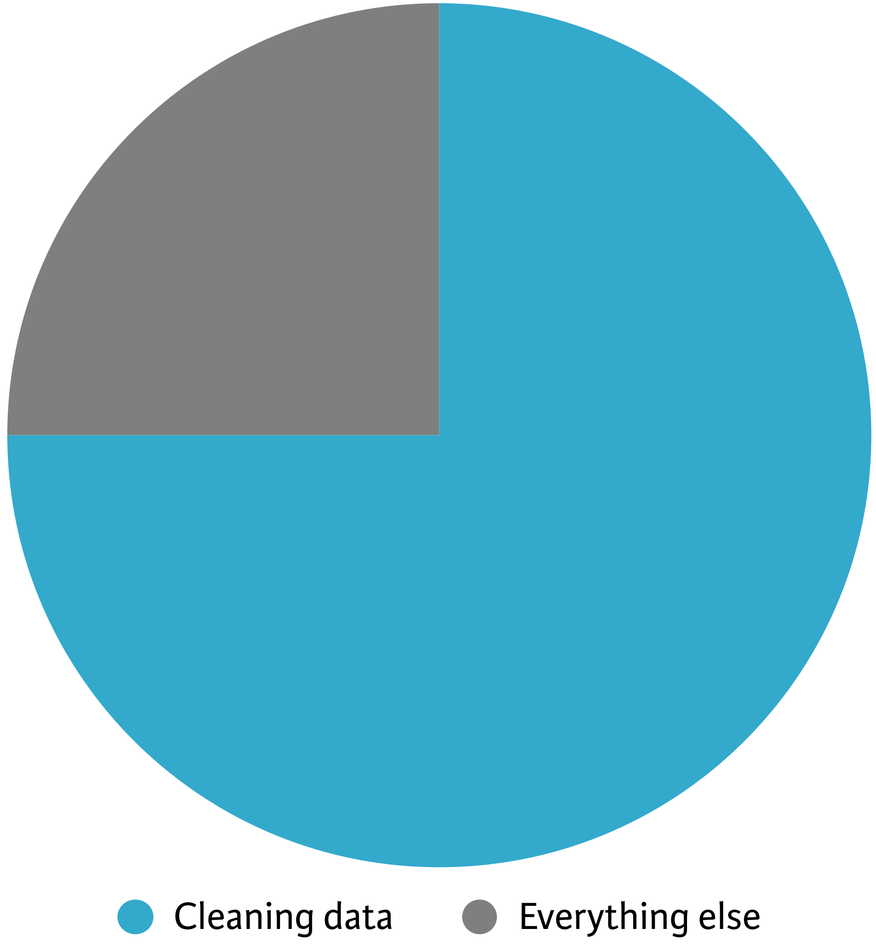
A look at some dirty data

```
> 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 ...

> tail(weather)
      X year month             measure   X1   X2   X3   X4 ...
281 281 2015    12 Mean.Wind.SpeedMPH    6 <NA> <NA> <NA> ...
282 282 2015    12  Max.Gust.SpeedMPH   17 <NA> <NA> <NA> ...
283 283 2015    12    PrecipitationIn 0.14 <NA> <NA> <NA> ...
284 284 2015    12         CloudCover    7 <NA> <NA> <NA> ...
285 285 2015    12             Events Rain <NA> <NA> <NA> ...
286 286 2015    12     WindDirDegrees  109 <NA> <NA> <NA> ...
```

Why care about cleaning data?





What we'll cover in this course

1. Exploring raw data
2. Tidying data
3. Preparing data for analysis

4. Putting it all together

CLEANING DATA IN R

Let's practice!

CLEANING DATA IN R

Exploring raw data

Exploring raw data

- Understand the structure of your data
- Look at your data
- Visualize your data

Understanding the structure of your data

```
# Load the lunch data
> lunch <- read.csv("datasets/lunch_clean.csv")

# View its class
> class(lunch)
[1] "data.frame"

# View its dimensions
> dim(lunch)
[1] 46  7
   Rows Columns
# Look at column names
> names(lunch)
[1] "year"           "avg_free"       "avg_reduced"   "avg_full"
[5] "avg_total"      "total_served"   "perc_free_red"
```

Understanding the structure of your data

```
> str(lunch)
'data.frame':   46 obs. of  7 variables:
 $ year          : int  1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 ...
 $ avg_free      : num  2.9 4.6 5.8 7.3 8.1 8.6 9.4 10.2 10.5 10.3 ...
 $ avg_reduced   : num  0 0 0.5 0.5 0.5 0.5 0.6 0.8 1.3 1.5 ...
 $ avg_full      : num  16.5 17.8 17.8 16.6 16.1 15.5 14.9 14.6 14.5 14.9 ...
 $ avg_total     : num  19.4 22.4 24.1 24.4 24.7 24.6 24.9 25.6 26.2 26.7 ...
 $ total_served : num  3368 3565 3848 3972 4009 ...
 $ perc_free_red: num  15.1 20.7 26.1 32.4 35 37.1 40.3 43.1 44.8 44.4 ...
```

Understanding the structure of your data

```
# Load dplyr
> library(dplyr)

# View structure of lunch, the dplyr way
> glimpse(lunch)
Observations: 46
Variables: 7
$ year           (int) 1969, 1970, 1971, 1972, 1973, 1974...
$ avg_free       (dbl) 2.9, 4.6, 5.8, 7.3, 8.1, 8.6, 9.4,...
$ avg_reduced    (dbl) 0.0, 0.0, 0.5, 0.5, 0.5, 0.5, 0.6,...
$ avg_full       (dbl) 16.5, 17.8, 17.8, 16.6, 16.1, 15.5...
$ avg_total      (dbl) 19.4, 22.4, 24.1, 24.4, 24.7, 24.6...
$ total_served   (dbl) 3368, 3565, 3848, 3972, 4009, 3982...
$ perc_free_red (dbl) 15.1, 20.7, 26.1, 32.4, 35.0, 37.1...
```

Understanding the structure of your data

```
# View a summary
> summary(lunch)
      year          avg_free      avg_reduced
 Min.   :1969   Min.   : 2.90   Min.   :0.00
 1st Qu.:1980   1st Qu.: 9.93   1st Qu.:1.52
 Median :1992   Median :10.90   Median :1.80
 Mean   :1992   Mean   :11.81   Mean   :1.86
 3rd Qu.:2003   3rd Qu.:13.60   3rd Qu.:2.60
 Max.   :2014   Max.   :19.20   Max.   :3.20
    avg_full      avg_total     total_served  perc_free_red
 Min.   : 8.8   Min.   :19.4   Min.   :3368   Min.   :15.1
 1st Qu.:11.4   1st Qu.:24.2   1st Qu.:4006   1st Qu.:45.6
 Median :12.2   Median :25.9   Median :4252   Median :52.4
 Mean   :12.8   Mean   :26.4   Mean   :4367   Mean   :51.1
 3rd Qu.:14.2   3rd Qu.:28.3   3rd Qu.:4751   3rd Qu.:58.3
 Max.   :17.8   Max.   :31.8   Max.   :5278   Max.   :71.6
```

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

CLEANING DATA IN R

Let's practice!

CLEANING DATA IN R

Exploring raw data

Looking at your data

```
# View the top
> head(lunch)
  year avg_free avg_reduced avg_full avg_total total_served
1 1969      2.9          0.0     16.5      19.4          3368
2 1970      4.6          0.0     17.8      22.4          3565
3 1971      5.8          0.5     17.8      24.1          3848
4 1972      7.3          0.5     16.6      24.4          3972
5 1973      8.1          0.5     16.1      24.7          4009
6 1974      8.6          0.5     15.5      24.6          3982
  perc_free_red
1          15.1
2          20.7
3          26.1
4          32.4
5          35.0
6          37.1
```

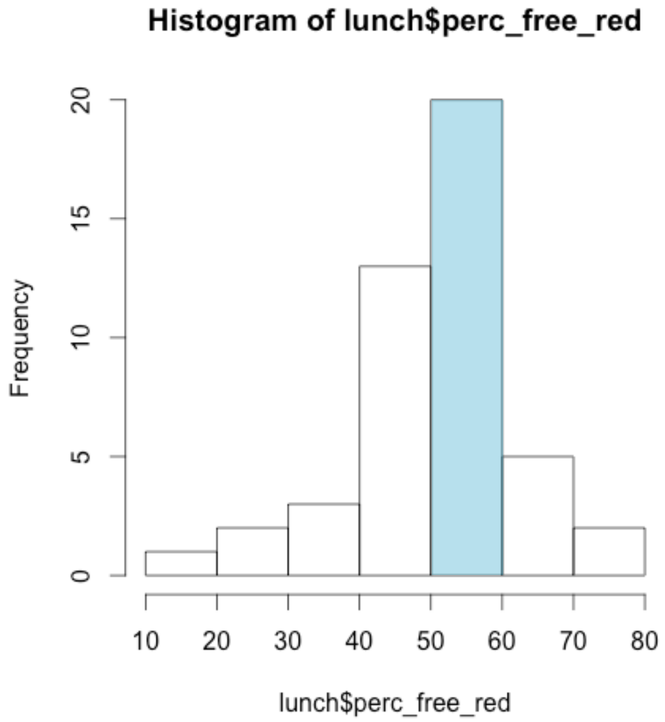
head(lunch, n = 15)

Looking at your data

```
# View the bottom
> tail(lunch)
   year avg_free avg_reduced avg_full avg_total total_served
41 2009     16.3          3.2     11.9      31.3         5186
42 2010     17.6          3.0     11.1      31.8         5278
43 2011     18.4          2.7     10.8      31.8         5274
44 2012     18.7          2.7     10.2      31.7         5215
45 2013     18.9          2.6      9.2      30.7         5098
46 2014     19.2          2.5      8.8      30.5         5020
   perc_free_red
41          62.6
42          65.3
43          66.6
44          68.2
45          70.5
46          71.6
```

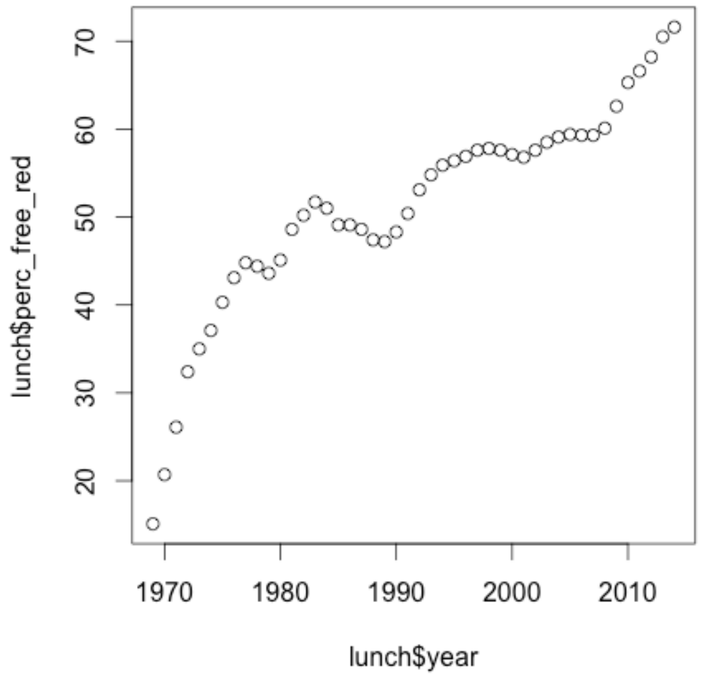
Visualizing your data

```
# View histogram
> hist(lunch$perc_free_red)
```



Visualizing your data

```
# View plot of two variables
> plot(lunch$year, lunch$perc_free_red)
```



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

CLEANING DATA IN R

Let's practice!