Data Wrangling Demo

Art Tay

Loading Dependencies

```
# Option 1: Individual Packages (not recommended).
library(dplyr)

# Option 2: tidyverse (okay).
library(tidyverse)

# Option 3: tidymodels (preferred).
library(tidymodels)
tidymodels_prefer() # Can be used to avoid conflicts with other packages.
```

```
data <- read.csv("./Cholesterol_R.csv")
head(data)
```

```
  ID Before After4weeks After8weeks Margarine
1  1   6.42        5.83        5.75         B
2  2   6.76        6.20        6.13         A
3  3   6.56        5.83        5.71         B
4  4   4.80        4.27        4.15         A
5  5   8.43        7.71        7.67         B
6  6   7.49        7.12        7.05         A
```

```
str(data)
```

```
'data.frame':   18 obs. of  5 variables:
 $ ID         : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Before     : num  6.42 6.76 6.56 4.8 8.43 7.49 8.05 5.05 5.77 3.91 ...
 $ After4weeks: num  5.83 6.2 5.83 4.27 7.71 7.12 7.25 4.63 5.31 3.7 ...
 $ After8weeks: num  5.75 6.13 5.71 4.15 7.67 7.05 7.1 4.67 5.33 3.66 ...
 $ Margarine  : chr  "B" "A" "B" "A" ...
```

```
# Quick Intro to pipes plus renaming.

# rename(new_name = old_name).
#data <- data %>% rename(ID = ï..ID)

# Think %>% mean "and then".
```

Subseting

```
# Selecting columns manually.
measurements <- data %>% select(Before, After4weeks, After8weeks)
head(measurements)
```

```
  Before After4weeks After8weeks
1   6.42        5.83        5.75
2   6.76        6.20        6.13
3   6.56        5.83        5.71
4   4.80        4.27        4.15
5   8.43        7.71        7.67
6   7.49        7.12        7.05
```

```
# Selecting based on a pattern.
after <- data %>% select(starts_with("After"))
head(after)
```

```
  After4weeks After8weeks
1        5.83        5.75
2        6.20        6.13
3        5.83        5.71
4        4.27        4.15
5        7.71        7.67
6        7.12        7.05
```

```
# Lots of options.
tidyselect::starts_with()
tidyselect::ends_with()
tidyselect::contains()
tidyselect::matches()
tidyselect::num_range()
tidyselect::everything()
tidyselect::one_of()
tidyselect::all_of()
tidyselect::any_of()
```

```
# Slicing rows.
first3 <- data %>% slice(1:3)
first3
```

```
  ID Before After4weeks After8weeks Margarine
1  1   6.42        5.83        5.75         B
2  2   6.76        6.20        6.13         A
3  3   6.56        5.83        5.71         B
```

```
# Filter rows based on a condition.
A_above_6 <- data %>% filter(After8weeks >= 6 & Margarine == "A")
head(A_above_6)
```

```
  ID Before After4weeks After8weeks Margarine
1  2   6.76        6.20        6.13         A
2  6   7.49        7.12        7.05         A
```

```
3 14   7.67        7.11        6.96        A
4 15   7.34        6.84        6.82        A
```

Sorting

Pivoting

Joins

Summarizing

Feature Engineering

Extensions for Modeling

Resampling

Missing Data