Dataframe basics

Bryan Clair

Monday, February 1, 2021

Recall

- Using RStudio IDE to code in R
- Variables, assign with `<-`
- Vectors and dataframes, `str()` for structure.
- Libraries: install once, load every session
- `?` for help

RMarkdown

Homework due Friday. RMarkdown required.

- `markdown-demo.Rmd`
- `homework-template.Rmd`

Find on our course data page:

https://mathstat.slu.edu/~clair/stat

RMarkdown lets you display R commands followed by the output of those commands.

Also formatted text, math with LaTeX.

`Knit` button converts to HTML or PDF (if you have LaTeX installed)

Reading data

R can read data from almost any source: Excel, wikipedia tables, PDFs, …

We will primarily read data from `.csv` files:

```
hot_dogs <-
  read.csv("https://mathstat.slu.edu/~clair/stat/data/hot-dogs.csv")
str(hot_dogs)
```

```
## 'data.frame':    54 obs. of  3 variables:
##  $ type    : Factor w/ 3 levels "Beef","Meat",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ calories: int  186 181 176 149 184 190 158 139 175 148 ...
##  $ sodium  : int  495 477 425 322 482 587 370 322 479 375 ...
```

Hot dogs

The `$` operator selects one variable from a data frame:

```
hot_dogs$sodium
```

```
##  [1] 495 477 425 322 482 587 370 322 479 375 330 300 386 401 645 440 317 319 298
## [20] 253 458 506 473 545 496 360 387 386 507 393 405 372 144 511 405 428 339 430
## [39] 375 396 383 387 542 359 357 528 513 426 513 358 581 588 522 545
```

Explore the hot dogs data using `summary`, `hist`, and `plot`.

- Make a histogram of sodium content.
- How many hot dogs of each type are there?
- What is the mean calorie content of hot dogs in this data?
- What is the maximum sodium content of any hot dog?
- Plot sodium and calories together. Color by hot dog type.

Element selection with [ ]

```
head(hot_dogs,n=4)
```

```
##   type calories sodium
## 1 Beef      186    495
## 2 Beef      181    477
## 3 Beef      176    425
## 4 Beef      149    322
```

```
hot_dogs$calories[3]  # 3rd entry
```

```
## [1] 176
```

```
hot_dogs[3,2]         # row 3 column 2
```

```
## [1] 176
```

```
hot_dogs[3,]          # row 3
```

```
##   type calories sodium
## 3 Beef      176    425
```

Filtering

```
hot_dogs$calories > 180
```

```
##  [1]  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
## [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [49] FALSE FALSE FALSE FALSE FALSE FALSE
```

```
hot_dogs[hot_dogs$calories > 180, ]
```

```
##    type calories sodium
## 1  Beef      186    495
## 2  Beef      181    477
## 5  Beef      184    482
## 6  Beef      190    587
## 15 Beef      190    645
## 22 Meat      191    506
## 23 Meat      182    473
## 24 Meat      190    545
## 34 Meat      195    511
```

Filtering with `dplyr`

```
library(dplyr)
```

```
filter(hot_dogs, calories > 180)
```

```
##   type calories sodium
## 1 Beef      186    495
## 2 Beef      181    477
## 3 Beef      184    482
## 4 Beef      190    587
## 5 Beef      190    645
## 6 Meat      191    506
## 7 Meat      182    473
## 8 Meat      190    545
## 9 Meat      195    511
```

Practice

- How many hot dogs of each type have over 500mg of sodium?

Get the `cereal.csv` data from our data page.

- How many variables are there? How many observations?
- Look at the sugars variable. What do you observe?
- What is the type variable about?
- How many cereals are there from each manufacturer?
- How are sugar content and rating related?
- How are shelf and manufacturer related? (hint: use `table`)
- Which cereal has the most sugar per cup?