Introduction to the R package dplyr

Stat 133 with Gaston Sanchez

Creative Commons Attribution Share-Alike 4.0 International CC BY-SA

First contact with tabular data

Game Plan

We'll use the R package **dplyr** to manipulate tables in a modern-syntactic way.

We'll be using a toy data table to illustrate dplyr concepts.

Toy Data

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50

4

Toy Data

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50

```
dat <- data.frame(
   name = c('Anakin', 'Padme', 'Luke', 'Leia'),
   gender = c('male', 'female', 'male', 'female'),
   height = c(1.88, 1.65, 1.72, 1.50)
)</pre>
```

dplyr verbs

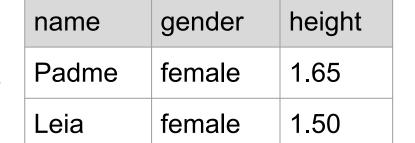
- filter
- select
- slice
- mutate
- group_by
- arrange
- summarise

Structure of dplyr verbs

- First argument is a data frame (or tibble)
- Subsequent arguments say what to do with data frame
- Always return a data frame (or tibble)
- Never modify in place

filter

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



name	gender	height
Luke	male	1.72

filter(dat, name == "Luke")

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



name	gender	height
Luke	male	1.72
Leia	female	1.50

filter(dat, name %in% c("Luke", "Leia"))

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72

filter(dat, name != "Leia")

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



name	gender	height
Padme	female	1.65
Leia	female	1.50

filter(dat, height < 1.70)</pre>

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



name	gender	height
Padme	female	1.65
Luke	male	1.72

filter(dat, height > 1.6 & height < 1.8)</pre>

select

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



select(dat, name)

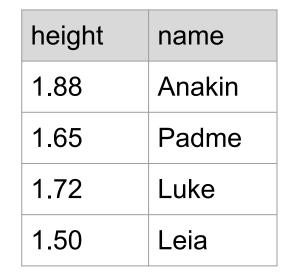
name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



name	height
Anakin	1.88
Padme	1.65
Luke	1.72
Leia	1.50

select(dat, name, height)

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



select(dat, height, name)

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



name	gender
Anakin	male
Padme	female
Luke	male
Leia	female

select(dat, -height)

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50

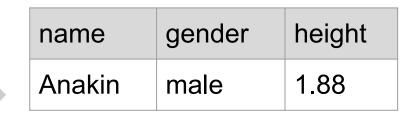


name	gender
Anakin	male
Padme	female
Luke	male
Leia	female

select(dat, name:gender)

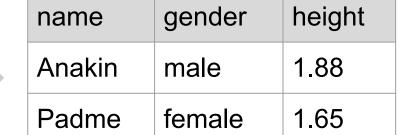
slice

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



slice(dat, 1)

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



slice(dat, 1:2)

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



name	gender	height
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50

slice(dat, -1)

arrange

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



name	gender	height
Anakin	male	1.88
Leia	female	1.50
Luke	male	1.72
Padme	female	1.65

arrange(dat, name)

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



name	gender	height
Padme	female	1.65
Leia	female	1.50
Anakin	male	1.88
Luke	male	1.72

arrange(dat, gender)

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



name	gender	height
Leia	female	1.50
Padme	female	1.65
Luke	male	1.72
Anakin	male	1.88

arrange(dat, height)

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



name	gender	height
Anakin	male	1.88
Luke	male	1.72
Padme	female	1.65
Leia	female	1.50

arrange(dat, desc(height))

mutate

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



name	gender	height
Anakin	male	0.188
Padme	female	0.165
Luke	male	0.172
Leia	female	0.150

mutate(dat, height = height / 10)

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50

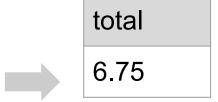


name	gender	height	ht10
Anakin	male	1.88	18.8
Padme	female	1.65	16.5
Luke	male	1.72	17.2
Leia	female	1.50	15.0

mutate(dat, ht10 = height * 10)

Grouped Summarise

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



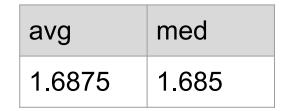
summarise(dat, total = sum(height))

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



summarise(dat, avg = mean(height))

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



```
summarise(dat,
  avg = mean(height),
  med = median(height))
```

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



gender	avg
female	1.58
male	1.8

by_gender <- group_by(dat, gender)</pre>

summarise(by_gender, avg = mean(height))

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



gender	min
female	1.5
male	1.72

by_gender <- group_by(dat, gender)</pre>

summarise(by_gender, min = min(height))

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



gender	min	max
female	1.5	1.65
male	1.72	1.88

```
by_gender <- group_by(dat, gender)</pre>
```

```
summarise(by_gender,
min = min(height),
max = max(height))
```

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



gender	avg	sd
female	1.58	0.106
male	1.8	0.113

```
summarise(
  group_by(dat, gender),
  avg = mean(height),
  sd = sd(height))
```

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



gender	avg	sd
male	1.8	0.113
female	1.58	0.106

```
arrange(
   summarise(group_by(dat, gender),
   avg = mean(height),
   sd = sd(height)),
   desc(avg))
```

Other Functions

dat

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50

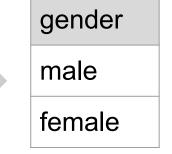


gender	n
female	2
male	2

count(by_gender)

dat

name	gender	height
Anakin	male	1.88
Padme	female	1.65
Luke	male	1.72
Leia	female	1.50



distinct(select(dat, gender))

n_distinct(select(dat, gender)) ----> 2