Data Science for Everyone – Data Wrangling

Dr. Ab Mosca (they/them)
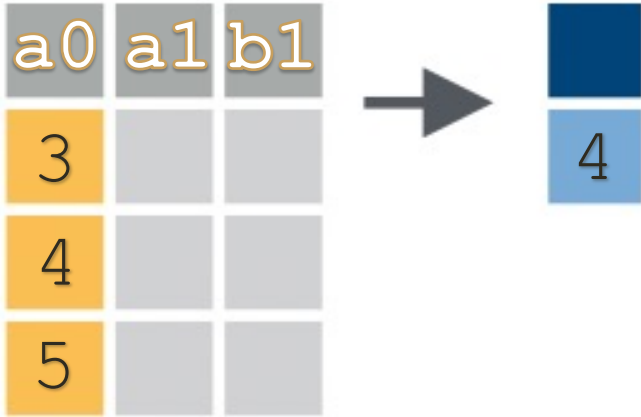
Plan for Today

- Wrangling Data in one Table

Recap: `summarize()`

The 5 Verbs: dplyr

`summarize()` column with a single value(s)

- Apply a summary function to a column. Ex.

```
data %>%
    summarize(mean(a0))
```

The 5 Verbs: `dplyr`

`summarize()` column with a single value(s)


```
data %>%
    summarize(mean(a0))
```

What if I want to summarize by different groups in the data?

What if I want to summarize by different groups in the data?

Ex. What is the average weight of red vs green fruit?

Fruit	Color	Weight
Apple	Green	150
Pear	Green	170
Strawberries	Red	300
Grapes	Green	225
Cherries	Red	190
Apple	Red	165

`summarize()`

summarize()

What if I want to summarize by different groups in the data?

Ex. What is the average weight of red vs green fruit?

Fruit	Color	Weight
Apple	Green	150
Pear	Green	170
Strawberries	Red	300
Grapes	Green	225
Cherries	Red	190
Apple	Red	165

```
data %>%
    summarize(avg_weight =
              mean(Weight))
```

summarize()

What if I want to summarize by different groups in the data?

Ex. What is the average weight of red vs green fruit?

Fruit	Color	Weight
Apple	Green	150
Pear	Green	170
Strawberries	Red	300
Grapes	Green	225
Cherries	Red	190
Apple	Red	165

```
data %>%
    summarize(avg_weight =
                mean(Weight))
```

avg_weight
200

`summarize()`

What if I want to summarize by different groups in the data?

Ex. What is the average weight of red vs green fruit?

Fruit	Color	Weight
Apple	Green	150
Pear	Green	170
Strawberries	Red	300
Grapes	Green	225
Cherries	Red	190
Apple	Red	165

```
data %>%
    summarize(avg_weight =
        mean(Weight))
```

avg_weight

200

We need to group the data and summarize each group…

summarize() by Group

What if I want to summarize by different groups in the data?

Ex. What is the average weight of red vs green fruit?

Fruit	Color	Weight
Apple	Green	150
Pear	Green	170
Strawberries	Red	300
Grapes	Green	225
Cherries	Red	190
Apple	Red	165

```
data %>%
    group_by(Color)
```

summarize()
by Group

What if I want to summarize by different groups in the data?

Ex. What is the average weight of red vs green fruit?

Fruit	Color	Weight
Apple	Green	150
Pear	Green	170
Strawberries	Red	300
Grapes	Green	225
Cherries	Red	190
Apple	Red	165

```
data %>%
    group_by(Color)
```

Fruit	Color	Weight
Apple	Green	150
Pear	Green	170
Grapes	Green	225
Strawberries	Red	300
Cherries	Red	190
Apple	Red	165

summarize() by Group

What if I want to summarize by different groups in the data?

Ex. What is the average weight of red vs green fruit?

Fruit	Color	Weight
Apple	Green	150
Pear	Green	170
Strawberries	Red	300
Grapes	Green	225
Cherries	Red	190
Apple	Red	165

```
data %>%
  group_by(Color) %>%
  summarize(avg_weight =
              mean(Weight))
```

Color	avg_weight
Green	181.67
Red	218.33

group_by()

`group_by()`

- Creates a grouped version of your dataset
- Performs following operations by groups

- Take a look at the `dplyr` cheatsheet for more `summarize()` helper functions: R Cheatsheets

Your turn!

Work with 1-2 other people on lab-group_by