Video 21: dplyr - group_by, summarise

Stats 102A

Miles Chen, PhD

Summarize with `summarise()`

Hadley is from New Zealand where they spell it with an s. He later added `summarize()` to have the same functionality, but I'm accustomed to using the original function.

Summary functions take multiple values and summarize them with a single value. For example, `mean()` and `var()` are summary functions.

```
starwars %>%
  select(height, mass) %>%
  summarise(
    avg_height = mean(height, na.rm = TRUE),
    var_height = var(height, na.rm = TRUE),
    avg_mass = mean(mass, na.rm = TRUE),
    min_height = min(height, na.rm = TRUE),
    max_mass = max(mass, na.rm = TRUE),
    count = n())
```

```
# A tibble: 1 × 6
  avg_height var_height avg_mass min_height max_mass count
       <dbl>      <dbl>    <dbl>      <int>    <dbl> <int>
1       175.      1209.     97.3         66     1358    87
```

Create groups using `group_by()`

We can create groups using the `group_by()` function.

```
starwars %>%
  group_by(species) %>%
  select(name, height, mass, species)
```

```
# A tibble: 87 × 4
# Groups:   species [38]
   name               height  mass species
   <chr>               <int> <dbl> <chr>
 1 Luke Skywalker        172    77 Human
 2 C-3PO                 167    75 Droid
 3 R2-D2                  96    32 Droid
 4 Darth Vader           202   136 Human
 5 Leia Organa           150    49 Human
 6 Owen Lars             178   120 Human
 7 Beru Whitesun Lars    165    75 Human
 8 R5-D4                  97    32 Droid
 9 Biggs Darklighter     183    84 Human
10 Obi-Wan Kenobi        182    77 Human
# ℹ 77 more rows
```

group_by() + summarise()

The power of `group_by()` is realized when combined with `summarise()`

```
starwars %>%
  group_by(species) %>%
  select(name, height, mass, species) %>%
  summarise(
    mean_ht = mean(height, na.rm = TRUE),
    sd_ht = sd(height, na.rm = TRUE),
    mean_mass = mean(mass, na.rm = TRUE),
    sd_mass = sd(mass, na.rm = TRUE),
    count = n())
```

```
# A tibble: 38 × 6
   species   mean_ht sd_ht mean_mass sd_mass count
   <chr>       <dbl> <dbl>     <dbl>   <dbl> <int>
 1 Aleena         79   NA        15      NA       1
 2 Besalisk      198   NA       102      NA       1
 3 Cerean        198   NA        82      NA       1
 4 Chagrian      196   NA       NaN      NA       1
 5 Clawdite      168   NA        55      NA       1
 6 Droid         131.  49.1      69.8    51.0     6
 7 Dug           112   NA        40      NA       1
 8 Ewok           88   NA        20      NA       1
 9 Geonosian     183   NA        80      NA       1
10 Gungan        209.  14.2      74      11.3     3
# ℹ 28 more rows
```

group_by() + summarise()

```
starwars %>%
  group_by(species) %>%
  select(name, height, mass, species) %>%
  summarise(
    mean_ht = mean(height, na.rm = TRUE),
    sd_ht = sd(height, na.rm = TRUE),
    mean_mass = mean(mass, na.rm = TRUE),
    sd_mass = sd(mass, na.rm = TRUE),
    count = n()
    ) %>%
  filter(count > 1) %>%
  arrange(desc(count)) %>%
  head()
```

```
# A tibble: 6 × 6
  species  mean_ht sd_ht mean_mass sd_mass count
  <chr>      <dbl> <dbl>     <dbl>   <dbl> <int>
1 Human        178  12.0       81.3   19.3     35
2 Droid        131. 49.1       69.8   51.0      6
3 <NA>         175  12.4       81     31.2      4
4 Gungan       209. 14.2       74     11.3      3
5 Kaminoan     221  11.3       88     NA        2
6 Mirialan     168   2.83      53.1    4.38     2
```

group_by() + mutate()

Note that C-3PO is above average when compared to other droids in the data set, but below average when compared to all characters in the data set.

```
starwars %>%
  filter(species %in% c("Human","Droid") |
      is.na(species)) %>%
  select(name, species, height) %>%
  group_by(species) %>%
  mutate(z_height = (height - mean(height, na.rm = TRUE))/sd(height, na.rm = TRUE)) %>%
  head()
```

```
# A tibble: 6 × 4
# Groups:   species [2]
  name           species height z_height
  <chr>          <chr>    <int>    <dbl>
1 Luke Skywalker Human      172   -0.498
2 C-3PO          Droid      167    0.728
3 R2-D2          Droid       96   -0.716
4 Darth Vader    Human      202    1.99
5 Leia Organa    Human      150   -2.32
6 Owen Lars      Human      178    0
```

Without group_by()

Note that C-3PO is above average when compared to other droids in the data set, but below average when compared to all characters in the data set.

```
starwars %>%
  filter(species %in% c("Human","Droid") |
      is.na(species)) %>%
  select(name, species, height) %>%
  # group_by(species) %>%
  mutate(z_height = (height - mean(height, na.rm = TRUE))/sd(height, na.rm = TRUE)) %>%
  head()
```

```
# A tibble: 6 × 4
  name           species height z_height
  <chr>          <chr>    <int>    <dbl>
1 Luke Skywalker Human      172   0.0123
2 C-3PO          Droid      167  -0.188
3 R2-D2          Droid       96  -3.03
4 Darth Vader    Human      202   1.21
5 Leia Organa    Human      150  -0.867
6 Owen Lars      Human      178   0.252
```

Multiple group_by() on some toy data

```
1 toy_cases <- read_csv("https://raw.githubusercontent.com/rstudio/EDAWR/master/data-raw/toyb.csv")
2 print(toy_cases)
```

```
# A tibble: 12 × 4
   country      year sex    cases
   <chr>       <dbl> <chr>  <dbl>
 1 Afghanistan  1999 female     1
 2 Afghanistan  1999 male       1
 3 Afghanistan  2000 female     1
 4 Afghanistan  2000 male       1
 5 Brazil       1999 female     2
 6 Brazil       1999 male       2
 7 Brazil       2000 female     2
 8 Brazil       2000 male       2
 9 China        1999 female     3
10 China        1999 male       3
11 China        2000 female     3
12 China        2000 male       3
```

Multiple `group_by()` + `summarise()`

We can provide `group_by()` two variables and it will create a hierarchy of groups

```
summary1 <- toy_cases %>% group_by(country, year) %>%
  summarise(cases = sum(cases))
print(summary1)
```

```
# A tibble: 6 × 3
# Groups:   country [3]
  country      year cases
  <chr>       <dbl> <dbl>
1 Afghanistan  1999     2
2 Afghanistan  2000     2
3 Brazil       1999     4
4 Brazil       2000     4
5 China        1999     6
6 China        2000     6
```

Multiple group_by() + summarise()

```
summary2 <- summary1 %>% summarise(cases = sum(cases))
summary2
```

```
# A tibble: 3 × 2
  country     cases
  <chr>       <dbl>
1 Afghanistan     4
2 Brazil          8
3 China          12
```

```
summary3 <- summary2 %>% summarise(cases = sum(cases))
summary3
```

```
# A tibble: 1 × 1
  cases
  <dbl>
1    24
```

Chanign the order of the multiple `group_by()`

```r
summary_a <- toy_cases %>% group_by(year, country) %>%
  summarise(cases = sum(cases))
print(summary_a)
```

```
# A tibble: 6 × 3
# Groups:   year [2]
   year country     cases
  <dbl> <chr>       <dbl>
1  1999 Afghanistan     2
2  1999 Brazil          4
3  1999 China           6
4  2000 Afghanistan     2
5  2000 Brazil          4
6  2000 China           6
```

Multiple group_by() + summarise()

```
summary_b <- summary_a %>% summarise(cases = sum(cases))
summary_b
```

```
# A tibble: 2 × 2
   year cases
  <dbl> <dbl>
1  1999    12
2  2000    12
```

```
summary_c <- summary_b %>% summarise(cases = sum(cases))
summary_c
```

```
# A tibble: 1 × 1
  cases
  <dbl>
1    24
```