Summarizing Data

(dimensions shrink)

slides/10-Summarizing.pdf

Summary

Summarize whole columns

```
summary(data.frame)
```

Summarize columns by factor level

```
group_by(factor) +
summarize(summarystat =)

ungroup()
complete()   (tidyr)
```

Summarizing whole columns

`summary(`*data.frame*`)`

```
summary(iris)
```

```
skimr::skim(iris)
```

Skimr::skim()

```
skimr::skim(iris)
#> Skim summary statistics
#>  n obs: 150
#>  n variables: 5
#>
#> ── Variable type:factor ────────────────────────────────────────────
#>  variable missing complete   n n_unique                       top_counts
#>   Species       0      150 150        3 set: 50, ver: 50, vir: 50, NA: 0
#>  ordered
#>    FALSE
#>
#> ── Variable type:numeric ───────────────────────────────────────────
#>      variable missing complete   n mean   sd  p0 p25  p50 p75 p100
#>  Petal.Length       0      150 150 3.76 1.77 1   1.6 4.35 5.1  6.9
#>   Petal.Width       0      150 150 1.2  0.76 0.1 0.3 1.3  1.8  2.5
#>  Sepal.Length       0      150 150 5.84 0.83 4.3 5.1 5.8  6.4  7.9
#>   Sepal.Width       0      150 150 3.06 0.44 2   2.8 3    3.3  4.4
#>      hist
#>  ▇▁▁▂▅▅▃▁
#>  ▇▁▁▅▃▃▂▂
#>  ▂▇▅▇▆▅▂▂
#>  ▁▂▅▇▃▂▁▁
```

group_by() / summarize()

```
iris %>%
  group_by(Species) %>%
  summarize(SLmean = mean(Sepal.Length))
#> # A tibble: 3 x 2
#>   Species    SLmean
#>   <fct>       <dbl>
#> 1 setosa       5.01
#> 2 versicolor   5.94
#> 3 virginica    6.59
```

Reorder results

```
iris %>%
  group_by(Species) %>%
  summarize(SLmean = mean(Sepal.Length)) %>%
  arrange(desc(SLmean))
```

```
#> # A tibble: 3 x 2
#>   Species    SLmean
#>   <fct>       <dbl>
#> 1 virginica    6.59
#> 2 versicolor   5.94
#> 3 setosa       5.01
```

group_by multiple groups

```
mtcars %>%
   group_by(gear, am) %>%
   summarize(mean_mpg = mean(mpg))
```

```
#> # A tibble: 4 x 3
#> # Groups:   gear [3]
#>    gear    am mean_mpg
#>   <dbl> <dbl>    <dbl>
#> 1     3     0     16.1
#> 2     4     0     21.0
#> 3     4     1     26.3
#> 4     5     1     21.4
```

Add missing combinations

```
mtcars %>%
  group_by(gear, am) %>%
  summarize(mean_mpg = mean(mpg)) %>%
  ungroup() %>%
  complete(gear, am)
#> # A tibble: 6 x 3
#>    gear    am mean_mpg
#>   <dbl> <dbl>    <dbl>
#> 1     3     0     16.1
#> 2     3     1       NA
#> 3     4     0     21.0
#> 4     4     1     26.3
#> 5     5     0       NA
#> 6     5     1     21.4
```

Percentages

```
mtcars %>%
   group_by(gear) %>%
   summarize(count = n()) %>%
   mutate(percent = count/sum(count))
```

```
#> # A tibble: 3 x 3
#>    gear count percent
#>   <dbl> <int>   <dbl>
#> 1     3    15   0.469
#> 2     4    12   0.375
#> 3     5     5   0.156
```

More on percentages

```
mtcars %>%
   group_by(gear, am) %>%
   summarize(count = n()) %>%
   mutate(percent = count/sum(count))
```

```
#> # A tibble: 4 x 4
#> # Groups:   gear [3]
#>    gear    am count percent
#>   <dbl> <dbl> <int>   <dbl>
#> 1     3     0    15   1
#> 2     4     0     4   0.333
#> 3     4     1     8   0.667
#> 4     5     1     5   1
```

More on percentages

```
mtcars %>%
   group_by(am, gear) %>%
   summarize(count = n()) %>%
   mutate(percent = count/sum(count))
```

```
#> # A tibble: 4 x 4
#> # Groups:   am [2]
#>      am  gear count percent
#>   <dbl> <dbl> <int>   <dbl>
#> 1     0     3    15   0.789
#> 2     0     4     4   0.211
#> 3     1     4     8   0.615
#> 4     1     5     5   0.385
```

More on percentages

```
mtcars %>%
   group_by(gear, am) %>%
   summarize(count = n()) %>%
   ungroup() %>%
   mutate(percent = count/sum(count))
```

```
#> # A tibble: 4 x 4
#>    gear    am count percent
#>   <dbl> <dbl> <int>   <dbl>
#> 1     3     0    15   0.469
#> 2     4     0     4   0.125
#> 3     4     1     8   0.25
#> 4     5     1     5   0.156
```

Common summarize functions

`mean()`

`median()`

`min()`

`max()`

`sum()`

`n()`

all reduce input to a single value

Practice

`labs/10-Summarizing.Rmd`

Summarize whole columns

```
summary(data.frame)
```

Summarize columns by factor level

```
group_by(factor) +
summarize(summarystat =)

ungroup()
complete()   (tidyr)
```