dplyr_summarise

October 12, 2020

Using group_by() and summarise() in dplyr

1 1. Why use `group_by()` and `summarise()` from *dplyr*?

- Being able to aggregate and summarize by grouping is hugely common
- *split-apply-combine* pattern
- These operations can be "chained" with other *dplyr* functions
- Often makes for concise, intuitive, and readable code

1.1 1.1 Example of `group_by()` and `summarise()`

```
In [3]: library(dplyr)

        arrests <- read.csv("data/pvd_arrests_2020-10-03.csv")
```

```
In [4]: gender_tbl <- arrests %>%
            group_by(gender) %>%
            summarise(
                n_rows = n(),
                mean_age = mean(age)
            )

        head(gender_tbl)
```

```
`summarise()` ungrouping output (override with `.groups` argument)
```

A tibble: 5 Œ 3

gender <chr>	n_rows <int>	mean_age <dbl>
	21	29.47619
Female	1906	31.99895
Male	6804	33.20988
NULL	20	28.15000
Unknown	4	34.50000

2 2. Chaining `filter()` with `group_by()` and `summarise()`

```
In [7]: gender_tbl <- arrests %>%
            filter(
                from_city == "Providence",
                year == 2019
            ) %>%
            group_by(gender) %>%
            summarise(
                n_rows = n(),
                mean_age = mean(age),
                mean_cnts = mean(counts, na.rm = TRUE)
            )

        head(gender_tbl)
```

```
`summarise()` ungrouping output (override with `.groups` argument)
```

A tibble: 4 Œ 4

gender <chr>	n_rows <int>	mean_age <dbl>	mean_cnts <dbl>
	9	23.88889	1.000000
Female	515	33.46602	1.064039
Male	2039	33.38941	1.098027
Unknown	1	49.00000	1.000000

2.1 2.1 More Interesting Example of Chaining

```
In [8]: is_summer <- function(month_num) {
            chk <- month_num %in% c(6, 7, 8)
            return(chk)
        }
```

```
In [10]: is_summer(6)   # TRUE
         is_summer(2)   # FALSE
         is_summer(8)   # TRUE
```

TRUE
FALSE
TRUE

2.1.1 2.1.1 More Interesting Example (cont.)

```
In [11]: vio_tbl <- arrests %>%
             filter(
                 statute_desc != "",
                 statute_desc != "NULL",
                 year == 2020
```

```
    ) %>%
    group_by(statute_desc) %>%
    summarise(
        n_vios = n(),
        prop_male = mean(gender == "Male"),
        mean_age = mean(age),
        prop_summer = mean(is_summer(month))
    ) %>%
    arrange(desc(n_vios))

head(vio_tbl, 10)
```

```
`summarise()` ungrouping output (override with `.groups` argument)
```

A tibble: 10 Œ 5

statute_desc <chr>	n_vios <int>	prop_male <dbl>	m <d
Driving after Denial, Suspension or Revocation of License	457	0.7374179	30
DOMESTIC-SIMPLE ASSAULT/BATTERY	364	0.8104396	33
DISORDERLY CONDUCT	216	0.7453704	31
SIMPLE ASSAULT OR BATTERY	199	0.6381910	31
BENCH WARRANT ISSUED FROM SUPERIOR COURT	141	0.8014184	36
RESISTING LEGAL OR ILLEGAL ARREST	123	0.7642276	30
POSSESSION OF SCHEDULE I II III	116	0.8189655	36
BENCH WARRANT ISSUED FROM 6TH DISTRICT COURT	101	0.7821782	36
SHOPLIFTING-MISD - SHOPLIFTING	99	0.4343434	33
WARRANT OF ARREST ON AFFIDAVIT - ALL OTH OFFENSE	93	0.8709677	33