R DPLYR Unique Groups and Count

Fan Wang

Aggregate Table with Groups

Go back to fan's REconTools Package, R4Econ Repository, or Intro Stats with R Repository.

```
rm(list = ls(all.names = TRUE))
options(knitr.duplicate.label = 'allow')
```

```
library(tidyverse)
library(knitr)
library(kableExtra)
library(REconTools)
# file name
st_file_name = 'fs_group_unique_agg'
# Generate R File
try(purl(paste0(st_file_name, ".Rmd"), output=paste0(st_file_name, ".R"), documentation = 2))
# Generate PDF and HTML
# rmarkdown::render("C:/Users/fan/R4Econ/summarize/aggregate/fs_group_unique_agg.Rmd", "pdf_document")
# rmarkdown::render("C:/Users/fan/R4Econ/summarize/aggregate/fs_group_unique_agg.Rmd", "html_document")
```

Aggrgate Groups only Unique Group and Count

There are two variables that are numeric, we want to find all the unique groups of these two variables in a dataset and count how many times each unique group occurs

- r unique occurrence of numeric groups
- How to add count of unique values by group to R data.frame

```
# Numeric value combinations unique Groups
vars.group <- c('hgt0', 'wgt0')

# dataset subsetting
df_use <- df_hgt_wgt %>% select(!!!syms(c(vars.group))) %>%
            mutate(hgt0 = round(hgt0/5)*5, wgt0 = round(wgt0/2000)*2000) %>%
            drop_na()

# Group, count and generate means for each numeric variables
# mutate_at(vars.group, funs(as.factor(.))) %>%
df.group.count <- df_use %>% group_by(!!!syms(vars.group)) %>%
                    arrange(!!!syms(vars.group)) %>%
                    summarise(n_obs_group=n())

# Show results Head 10
df.group.count %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
```

hgt0

wgt0

n_obs_group

40

2000

122

45

2000

4586

45

4000

470

50

2000

9691

50

4000

13106

55

2000

126

55

4000

1900

60

6000

18

Aggrgate Groups only Unique Group Show up With Means

Several variables that are grouping identifiers. Several variables that are values which mean be unique for each group members. For example, a Panel of income for N households over T years with also household education information that is invariant over time. Want to generate a dataset where the unit of observation are households, rather than household years. Take average of all numeric variables that are household and year specific.

A complicating factor potentially is that the number of observations differ within group, for example, income might be observed for all years for some households but not for other households.

- r dplyr aggregate group average
- Aggregating and analyzing data with dplyr
- column can't be modified because it is a grouping variable

- see also: Aggregating and analyzing data with dplyr

```
# In the df_hgt_wgt from R4Econ, there is a country id, village id,
# and individual id, and various other statistics
vars.group <- c('S.country', 'vil.id', 'indi.id')
vars.values <- c('hgt', 'momEdu')

# dataset subsetting
df_use <- df_hgt_wgt %>% select(!!!syms(c(vars.group, vars.values)))

# Group, count and generate means for each numeric variables
df.group <- df_use %>% group_by(!!!syms(vars.group)) %>%
            arrange(!!!syms(vars.group)) %>%
            summarise_if(is.numeric,
                         funs(mean = mean(., na.rm = TRUE),
                              sd = sd(., na.rm = TRUE),
                              n = sum(is.na(.)==0)))

# Show results Head 10
df.group %>% head(10) %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
```

S.country

vil.id

indi.id

hgt_mean

momEdu_mean

hgt_sd

momEdu_sd

hgt_n

momEdu_n

Cebu

1

1

61.80000

5.3

9.520504

0

7

18

Cebu

1

2

68.86154

7.1

9.058931

0

13

18

Cebu

1

3

80.45882

9.4

29.894231

0

17

18

Cebu

1

4

88.10000

13.9

35.533166

0

18

18

Cebu

1

5

97.70556

11.3

41.090366

0

18

18

Cebu

1

6

87.49444
7.3
35.586439
0
18
18
Cebu
1
7
90.79412
10.4
38.722385
0
17
18
Cebu
1
8
68.45385
13.5
10.011961
0
13
18
Cebu
1
9
86.21111
10.4
35.126057
0
18
18
Cebu
1
10

87.67222

10.5

36.508127

0

18

18

```
# Show results Head 10
df.group %>% tail(10) %>%
  kable() %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
```

S.country

vil.id

indi.id

hgt_mean

momEdu_mean

hgt_sd

momEdu_sd

hgt_n

momEdu_n

Guatemala

14

2014

66.97000

NaN

8.967974

NaN

10

0

Guatemala

14

2015

71.71818

NaN

11.399984

NaN

11

0

Guatemala
14
2016
66.33000
NaN
9.490352
NaN
10
0
Guatemala
14
2017
76.40769
NaN
14.827871
NaN
13
0
Guatemala
14
2018
74.55385
NaN
12.707846
NaN
13
0
Guatemala
14
2019
70.47500
NaN
11.797390
NaN
12
0

Guatemala
14
2020
60.28750
NaN
7.060036
NaN
8
0
Guatemala
14
2021
84.96000
NaN
15.446193
NaN
10
0
Guatemala
14
2022
79.38667
NaN
15.824749
NaN
15
0
Guatemala
14
2023
66.50000
NaN
8.613113
NaN
8
0