summarise() and group_by() functions

Pema Lama

April 5, 2017

In this session, we discuss about summarise() and group_by() functions.

How do I set current working directory?

Let us load dplyr() function.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

To load the data set, we use read.csv() function.

```
college <- read.csv('College.csv', stringsAsFactors = TRUE, header = TRUE)
```

To see the structure of the data, we use str() function.

```
str(college)
```

```
## 'data.frame':    777 obs. of  19 variables:
##  $ X          : Factor w/ 777 levels "Abilene Christian University",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : int  1660 2186 1428 417 193 587 353 1899 1038 582 ...
##  $ Accept     : int  1232 1924 1097 349 146 479 340 1720 839 498 ...
##  $ Enroll     : int  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : int  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : int  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: int  2885 2683 1036 510 249 678 416 1594 973 799 ...
##  $ P.Undergrad: int  537 1227 99 63 869 41 230 32 306 78 ...
##  $ Outstate   : int  7440 12280 11250 12960 7560 13500 13290 13868 15595 10468 ...
##  $ Room.Board : int  3300 6450 3750 5450 4120 3335 5720 4826 4400 3380 ...
##  $ Books      : int  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : int  2200 1500 1165 875 1500 675 1500 850 500 1800 ...
##  $ PhD        : int  70 29 53 92 76 67 90 89 79 40 ...
```

```
##  $ Terminal   : int  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: int  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : int  7041 10527 8735 19016 10922 9727 8861 11487 11644 8991 ...
##  $ Grad.Rate  : int  60 56 54 59 15 55 63 73 80 52 ...
```

```
college[1:6, ]
```

```
##                              X Private Apps Accept Enroll Top10perc
## 1 Abilene Christian University     Yes 1660   1232    721        23
## 2           Adelphi University     Yes 2186   1924    512        16
## 3               Adrian College     Yes 1428   1097    336        22
## 4          Agnes Scott College     Yes  417    349    137        60
## 5    Alaska Pacific University     Yes  193    146     55        16
## 6            Albertson College     Yes  587    479    158        38
##   Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD
## 1        52        2885         537     7440       3300   450     2200  70
## 2        29        2683        1227    12280       6450   750     1500  29
## 3        50        1036          99    11250       3750   400     1165  53
## 4        89         510          63    12960       5450   450      875  92
## 5        44         249         869     7560       4120   800     1500  76
## 6        62         678          41    13500       3335   500      675  67
##   Terminal S.F.Ratio perc.alumni Expend Grad.Rate
## 1       78      18.1          12   7041        60
## 2       30      12.2          16  10527        56
## 3       66      12.9          30   8735        54
## 4       97       7.7          37  19016        59
## 5       72      11.9           2  10922        15
## 6       73       9.4          11   9727        55
```

The summary statistics of the data set, we use summarise() function.

```
college_summary <- summarise(college,
                             count = n(),
                             avgAccept = mean(Accept, na.rm = TRUE),
                             sdAccept  = sd(Accept, na.rm = TRUE),
                             medAccept = median(Accept, na.rm = TRUE),
                             Q1st = quantile(Accept, 0.25),
                             Q3rd = quantile(Accept, 0.75))
college_summary
```

```
##   count avgAccept sdAccept medAccept Q1st Q3rd
## 1   777  2018.804 2451.114      1110  604 2424
```

Grouping a data set using a variable.

```
groupPrivate <- group_by(college, Private)
groupPrivate[1:6, 1:6]
```

```
## Source: local data frame [6 x 6]
## Groups: Private [1]
##
```

```
##                              X Private  Apps Accept Enroll Top10perc
##                          <fctr>  <fctr> <int>  <int>  <int>     <int>
## 1 Abilene Christian University     Yes  1660   1232    721        23
## 2           Adelphi University     Yes  2186   1924    512        16
## 3               Adrian College     Yes  1428   1097    336        22
## 4          Agnes Scott College     Yes   417    349    137        60
## 5    Alaska Pacific University     Yes   193    146     55        16
## 6            Albertson College     Yes   587    479    158        38
```

```
groupPrivate[1:6, 7:13]
```

```
## # A tibble: 6 × 7
##   Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books Personal
##       <int>       <int>       <int>    <int>      <int> <int>    <int>
## 1        52        2885         537     7440       3300   450     2200
## 2        29        2683        1227    12280       6450   750     1500
## 3        50        1036          99    11250       3750   400     1165
## 4        89         510          63    12960       5450   450      875
## 5        44         249         869     7560       4120   800     1500
## 6        62         678          41    13500       3335   500      675
```

```
groupPrivate[1:6, 14:19]
```

```
## # A tibble: 6 × 6
##     PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate
##   <int>    <int>     <dbl>       <int>  <int>     <int>
## 1    70       78      18.1          12   7041        60
## 2    29       30      12.2          16  10527        56
## 3    53       66      12.9          30   8735        54
## 4    92       97       7.7          37  19016        59
## 5    76       72      11.9           2  10922        15
## 6    67       73       9.4          11   9727        55
```

Grouping by binary variable of summary statistics.

```
college_summary <- summarise(groupPrivate,
                             count = n(),
                             avgAccept = mean(Accept, na.rm = TRUE),
                             sdAccept  = sd(Accept, na.rm = TRUE),
                             medAccept = median(Accept, na.rm = TRUE),
                             Q1st = quantile(Accept, 0.25),
                             Q3rd = quantile(Accept, 0.75))
college_summary
```

```
## # A tibble: 2 × 7
##   Private count avgAccept sdAccept medAccept    Q1st  Q3rd
##    <fctr> <int>     <dbl>    <dbl>     <dbl>   <dbl> <dbl>
## 1      No   212  3919.288 3477.266    2929.5 1563.25  5264
## 2     Yes   565  1305.703 1369.549     859.0  501.00  1580
```