dplyr Exercises

Elena Dubova

6/23/2020

Data from Harvard Database: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/O35FW8

Entrance exam scores of students applying to a university in Brazil (Federal University of Rio Grande do Sul), along with the students' GPAs during the first three semesters at university. In this dataset, each row contains anonymized information about an applicant's scores on nine exams taken as part of the application process to the university, as well as their corresponding GPA during the first three semesters at university. The dataset has 43,303 rows, each corresponding to one student. The columns correspond to: 1) Gender. 0 denotes female and 1 denotes male. 2) Score on physics exam 3) Score on biology exam 4) Score on history exam 5) Score on second language exam 6) Score on geography exam 7) Score on literature exam 8) Score on Portuguese essay exam 9) Score on math exam 10) Score on chemistry exam 11) Mean GPA during first three semesters at university, on a 4.0 scale.

First, we import the data into R and brush it up.

```
data <- read.csv('GPA_Brazil.csv', header = FALSE)
head(data)
```

```
##   V1     V2     V3     V4     V5     V6     V7     V8     V9    V10     V11
## 1  0 622.60 491.56 439.93 707.64 663.65 557.09 711.37 731.31 509.80 1.33333
## 2  1 538.00 490.58 406.59 529.05 532.28 447.23 527.58 379.14 488.64 2.98333
## 3  1 455.18 440.00 570.86 417.54 453.53 425.87 475.63 476.11 407.15 1.97333
## 4  0 756.91 679.62 531.28 583.63 534.42 521.40 592.41 783.76 588.26 2.53333
## 5  1 584.54 649.84 637.43 609.06 670.46 515.38 572.52 581.25 529.04 1.58667
## 6  1 325.99 466.74 597.06 554.43 535.77 717.03 477.60 503.82 422.92 1.66667
```

```
colnames(data) <- c('Gender', 'Physics', 'Biology', 'History', 'Second_Language', 'Geography', 'Literatu
```

```
data$Student_ID <-  paste0('Student_', 1:43303)
data <- data[c(12,1:11)]
```

```
head(data)
```

```
##   Student_ID Gender Physics Biology History Second_Language Geography
## 1  Student_1      0  622.60  491.56  439.93          707.64    663.65
## 2  Student_2      1  538.00  490.58  406.59          529.05    532.28
## 3  Student_3      1  455.18  440.00  570.86          417.54    453.53
## 4  Student_4      0  756.91  679.62  531.28          583.63    534.42
## 5  Student_5      1  584.54  649.84  637.43          609.06    670.46
## 6  Student_6      1  325.99  466.74  597.06          554.43    535.77
##   Literature Portuguese_Essay   Math Chemistry  GPA_3S
## 1     557.09           711.37 731.31    509.80 1.33333
## 2     447.23           527.58 379.14    488.64 2.98333
## 3     425.87           475.63 476.11    407.15 1.97333
## 4     521.40           592.41 783.76    588.26 2.53333
```

```
## 5     515.38           572.52 581.25    529.04 1.58667
## 6     717.03           477.60 503.82    422.92 1.66667
```

```
summary(data)
```

```
##   Student_ID            Gender          Physics         Biology
##  Length:43303       Min.   :0.0000   Min.   :299.3   Min.   :263.0
##  Class :character   1st Qu.:0.0000   1st Qu.:482.8   1st Qu.:492.4
##  Mode  :character   Median :1.0000   Median :565.6   Median :566.4
##                     Mean   :0.5158   Mean   :576.1   Mean   :568.7
##                     3rd Qu.:1.0000   3rd Qu.:662.8   3rd Qu.:634.8
##                     Max.   :1.0000   Max.   :952.1   Max.   :966.6
##     History      Second_Language   Geography       Literature
##  Min.   :265.0   Min.   :222.7   Min.   :224.9   Min.   :239.1
##  1st Qu.:516.1   1st Qu.:517.7   1st Qu.:510.2   1st Qu.:516.8
##  Median :578.9   Median :580.3   Median :575.5   Median :587.1
##  Mean   :580.8   Mean   :574.0   Mean   :574.5   Mean   :583.3
##  3rd Qu.:650.2   3rd Qu.:640.6   3rd Qu.:637.3   3rd Qu.:648.7
##  Max.   :925.8   Max.   :858.4   Max.   :941.8   Max.   :904.8
##  Portuguese_Essay      Math          Chemistry          GPA_3S
##  Min.   :151.6    Min.   : 298.0   Min.   : 300.5   Min.   :0.000
##  1st Qu.:491.9    1st Qu.: 489.4   1st Qu.: 484.5   1st Qu.:2.280
##  Median :553.6    Median : 571.9   Median : 565.5   Median :2.920
##  Mean   :551.0    Mean   : 579.2   Mean   : 571.7   Mean   :2.786
##  3rd Qu.:613.1    3rd Qu.: 665.2   3rd Qu.: 655.4   3rd Qu.:3.430
##  Max.   :825.5    Max.   :1072.1   Max.   :1001.9   Max.   :4.000
```

```
sum(is.na(data))
```

```
## [1] 0
```

Data is clean, so we can proceed with extracting information from it.

Let us start with dplyr.

```
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
```

Special note: in the same fashion as with scalars, you can assign the output of the pipe to a variable or just print it out.

```
data %>%
  select(Student_ID, GPA_3S) %>%
  sample_n(5)
```

```
##      Student_ID  GPA_3S
## 1 Student_19796 2.63333
## 2 Student_27512 0.00000
## 3   Student_746 2.89333
```

```
## 4 Student_32881 2.61000
## 5 Student_25920 2.85333
```

```
my_sample <-data %>%
  select(Student_ID, GPA_3S) %>%
  sample_n(5)

my_sample
```

```
##      Student_ID  GPA_3S
## 1 Student_13440 1.99000
## 2 Student_40591 2.51667
## 3 Student_11964 1.64667
## 4 Student_18892 3.82000
## 5 Student_36424 2.58000
```

```
my_sample <-data %>%
  select(Student_ID, GPA_3S) %>%
  sample_n(5)

my_sample
```

```
##      Student_ID  GPA_3S
## 1 Student_10770 2.17667
## 2 Student_32491 3.86667
## 3 Student_15510 2.60000
## 4 Student_15719 3.32667
## 5 Student_32498 2.83333
```

1. Select.

1a. Select physics, chemistry and math columns. Print out first 4 rows of resulting dataframe. What is the dimensions of the dataframe?

```
data %>%
  select(Physics, Chemistry, Math) %>%
  head(4) %>%
  dim()
```

```
## [1] 4 3
```

1b. Select all columns but Mean_GPA_3S. Print out first 4 rows of resulting dataframe. Can you think about a second way to write this code (if you don't know, google it!)? Which way you prefer?

```
data %>%
  select(-GPA_3S) %>%
  head(4)
```

```
##   Student_ID Gender Physics Biology History Second_Language Geography
## 1  Student_1      0  622.60  491.56  439.93          707.64    663.65
## 2  Student_2      1  538.00  490.58  406.59          529.05    532.28
## 3  Student_3      1  455.18  440.00  570.86          417.54    453.53
## 4  Student_4      0  756.91  679.62  531.28          583.63    534.42
##   Literature Portuguese_Essay   Math Chemistry
## 1     557.09           711.37 731.31    509.80
## 2     447.23           527.58 379.14    488.64
## 3     425.87           475.63 476.11    407.15
```

```
## 4     521.40           592.41 783.76    588.26
```

2. Filter.

2a. Filter the datadrame so that it contains only male students. How many male students are in the dataframe? Female students?

```
data %>%
  filter(Gender == 1) %>%
  count()
```

```
##       n
## 1 22335
```

2b. Filter the dataframe so that it has only students with math exam scores above 1000. What are their IDs?

```
data %>%
  select(Student_ID, Math) %>%
  filter(Math > 1000)
```

```
##      Student_ID    Math
## 1  Student_1665 1072.12
## 2 Student_21017 1019.69
## 3 Student_28500 1045.90
## 4 Student_30490 1045.90
```

2c. Filter the dataframe so that it has only students with both math and Portuguese essay exams scores above 800. What are their IDs?

```
data %>%
  select(Student_ID, Math, Portuguese_Essay) %>%
  filter(Math > 800 & Portuguese_Essay > 800) # you can use comma
```

```
##      Student_ID   Math Portuguese_Essay
## 1  Student_9462 902.16           810.44
## 2 Student_26637 810.85           801.96
```

3. Mutate.

3a. Create a new column and record sum of chemistry and biology scores. Name it 'life_science_score'.

```
data <- data %>%
  mutate(life_science_score = Chemistry + Biology)
head(data)
```

```
##   Student_ID Gender Physics Biology History Second_Language Geography
## 1  Student_1      0  622.60  491.56  439.93          707.64    663.65
## 2  Student_2      1  538.00  490.58  406.59          529.05    532.28
## 3  Student_3      1  455.18  440.00  570.86          417.54    453.53
## 4  Student_4      0  756.91  679.62  531.28          583.63    534.42
## 5  Student_5      1  584.54  649.84  637.43          609.06    670.46
## 6  Student_6      1  325.99  466.74  597.06          554.43    535.77
##   Literature Portuguese_Essay   Math Chemistry  GPA_3S life_science_score
## 1     557.09           711.37 731.31    509.80 1.33333            1001.36
## 2     447.23           527.58 379.14    488.64 2.98333             979.22
## 3     425.87           475.63 476.11    407.15 1.97333             847.15
## 4     521.40           592.41 783.76    588.26 2.53333            1267.88
```

```
## 5     515.38           572.52 581.25    529.04 1.58667            1178.88
## 6     717.03           477.60 503.82    422.92 1.66667             889.66
```

3b. Now update the values of 'life_science_score' column by dividing each value by 2. What summary statistic did you get?

Remember, you can reuse the code above by assigning it to a variable or write the code from scratch.

```
data <- data %>%
  mutate(life_science_score = Chemistry + Biology) %>%
  mutate(life_science_score = life_science_score/2)

head(data)
```

```
##   Student_ID Gender Physics Biology History Second_Language Geography
## 1  Student_1      0  622.60  491.56  439.93          707.64    663.65
## 2  Student_2      1  538.00  490.58  406.59          529.05    532.28
## 3  Student_3      1  455.18  440.00  570.86          417.54    453.53
## 4  Student_4      0  756.91  679.62  531.28          583.63    534.42
## 5  Student_5      1  584.54  649.84  637.43          609.06    670.46
## 6  Student_6      1  325.99  466.74  597.06          554.43    535.77
##   Literature Portuguese_Essay   Math Chemistry  GPA_3S life_science_score
## 1     557.09           711.37 731.31    509.80 1.33333            500.680
## 2     447.23           527.58 379.14    488.64 2.98333            489.610
## 3     425.87           475.63 476.11    407.15 1.97333            423.575
## 4     521.40           592.41 783.76    588.26 2.53333            633.940
## 5     515.38           572.52 581.25    529.04 1.58667            589.440
## 6     717.03           477.60 503.82    422.92 1.66667            444.830
```

3. Mutate. 3c. Create 'life_science_score_1' but this time calculate mean using rowMeans() function wrapped around cbind() function. To see documentation and examples type ?rowMeans

Check if you get the same result. How would you approach this?

```
data %>%
  mutate(life_science_score = Chemistry + Biology) %>%
  mutate(life_science_score = life_science_score/2) %>%
  mutate(life_science_score_1 = rowMeans(cbind(Chemistry,Biology))) %>%
  summarize(sum(life_science_score == life_science_score_1))
```

```
##   sum(life_science_score == life_science_score_1)
## 1                                           43303
```

3d. Create 'high_gpa' column that says 'You are good!' of student's GPA is above 3.0 and 'Work harder!' otherwise. Save result as new_data. Print the last 6 rows and check if your code worked.

You can use ifelse() function. Type ?ifelse to see the documentation.

```
new_data <- data %>%
  mutate(high_gpa = ifelse(GPA_3S >= 3, 'You are good!', 'Work harder!'))

tail(new_data)
```

```
##          Student_ID Gender Physics Biology History Second_Language Geography
## 43298 Student_43298      0  670.08  682.52  784.25          665.91    636.61
## 43299 Student_43299      1  519.55  622.20  660.90          543.48    643.05
## 43300 Student_43300      1  816.39  851.95  732.39          621.63    810.68
## 43301 Student_43301      0  798.75  817.58  731.98          648.42    751.30
## 43302 Student_43302      0  527.66  443.82  545.88          624.18    420.25
```

```
## 43303 Student_43303      0  512.56  415.41  517.36          532.37    592.30
##       Literature Portuguese_Essay   Math Chemistry  GPA_3S life_science_score
## 43298     676.80           649.44 611.36    652.51 3.63333            667.515
## 43299     579.90           584.80 581.25    573.92 2.76333            598.060
## 43300     666.79           705.22 781.01    831.76 3.81667            841.855
## 43301     648.67           662.05 773.15    835.25 3.75000            826.415
## 43302     676.80           583.41 395.46    509.80 2.50000            476.810
## 43303     382.20           538.35 448.02    496.39 3.16667            455.900
##            high_gpa
## 43298 You are good!
## 43299  Work harder!
## 43300 You are good!
## 43301 You are good!
## 43302  Work harder!
## 43303 You are good!
```

4. Group by a variable and summarize.

4a. Using the dataframe you created in the previous example, group by the new column - high_gpa. How many students have to work harder?

```
new_data %>%
  group_by(high_gpa) %>%
  summarize(count=n())
```

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

```
## # A tibble: 2 x 2
##   high_gpa      count
##   <chr>         <int>
## 1 Work harder!  23040
## 2 You are good! 20263
```

4b. Using the dataframe you created in the previous example, group by the new column - high_gpa. Calculate mean GPA for every group? Does the result make sense?

```
new_data %>%
  group_by(high_gpa) %>%
  summarize(count=n(), mean_gpa=mean(GPA_3S))
```

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

```
## # A tibble: 2 x 3
##   high_gpa      count mean_gpa
##   <chr>         <int>    <dbl>
## 1 Work harder!  23040     2.18
## 2 You are good! 20263     3.47
```

4c. What is the median GPA for students with high GPA?

```
new_data %>%
  filter(high_gpa == 'You are good!') %>%
  summarize(count=n(), mean_gpa=mean(GPA_3S), median_gpa=median(GPA_3S))
```

```
##   count mean_gpa median_gpa
## 1 20263 3.471684    3.46333
```

4c. What is the median GPA for students with high GPA?

```
new_data %>%
  filter(high_gpa == 'You are good!') %>%
  summarize(count=n(), mean_gpa=mean(GPA_3S), median_gpa=median(GPA_3S))
```

```
##   count mean_gpa median_gpa
## 1 20263 3.471684    3.46333
```

4d. Who are the students with GPA exactly equal median GPA? Let's collect them in a new dataframe named 'median_club'

```
median_club <- new_data %>%
  filter(GPA_3S == 3.46333)
```

4e. Group 'median_club' by gender and check the mean exam score for every discipline. Talk the results through.

```
median_club %>%
  group_by(Gender) %>%
  summarise_all(mean)
```

```
## Warning in mean.default(Student_ID): argument is not numeric or logical:
## returning NA

## Warning in mean.default(Student_ID): argument is not numeric or logical:
## returning NA

## Warning in mean.default(high_gpa): argument is not numeric or logical: returning
## NA

## Warning in mean.default(high_gpa): argument is not numeric or logical: returning
## NA
```

```
## # A tibble: 2 x 14
##   Gender Student_ID Physics Biology History Second_Language Geography Literature
##    <int>      <dbl>   <dbl>   <dbl>   <dbl>           <dbl>     <dbl>      <dbl>
## 1      0         NA    635.    660.    623.            597.      591.       657.
## 2      1         NA    705.    657.    641.            626.      635.       634.
## # ... with 6 more variables: Portuguese_Essay <dbl>, Math <dbl>,
## #   Chemistry <dbl>, GPA_3S <dbl>, life_science_score <dbl>, high_gpa <dbl>
```

```
data %>%
  group_by(Gender) %>%
  summarise_all(mean)
```

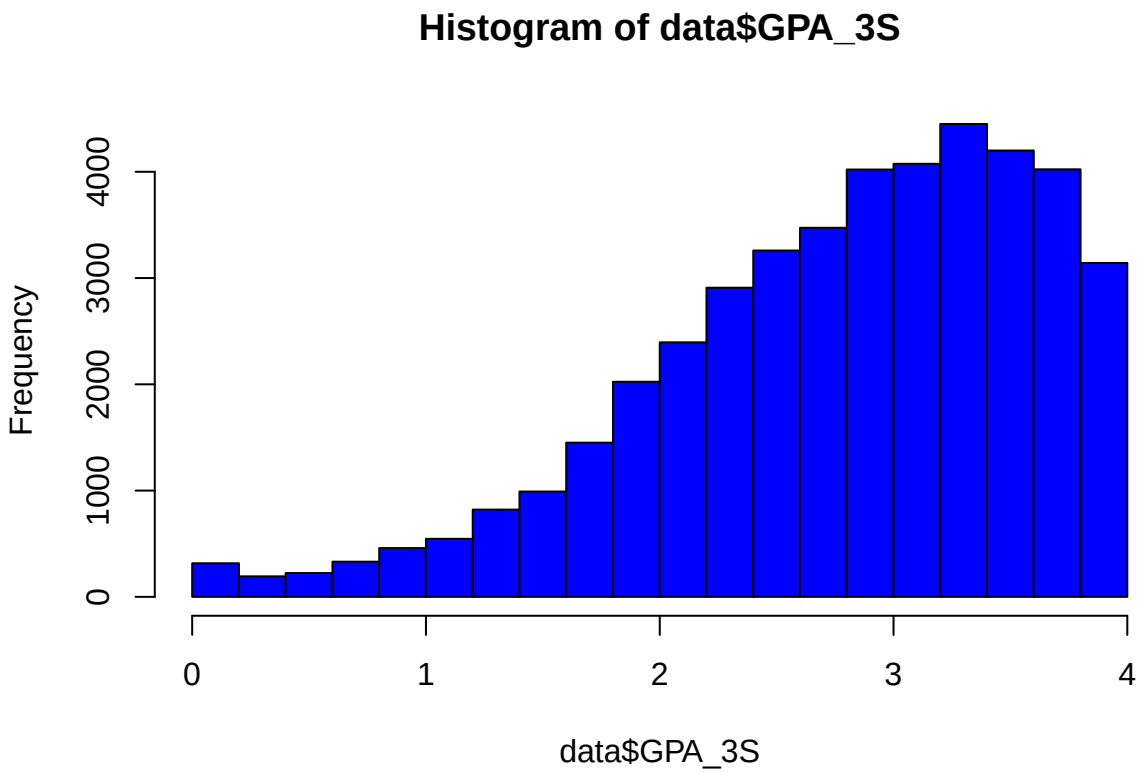
```
## Warning in mean.default(Student_ID): argument is not numeric or logical:
## returning NA

## Warning in mean.default(Student_ID): argument is not numeric or logical:
## returning NA
```

```
## # A tibble: 2 x 13
##   Gender Student_ID Physics Biology History Second_Language Geography Literature
##    <int>      <dbl>   <dbl>   <dbl>   <dbl>           <dbl>     <dbl>      <dbl>
## 1      0         NA    552.    565.    566.            571.      556.       596.
## 2      1         NA    599.    572.    594.            577.      592.       571.
## # ... with 5 more variables: Portuguese_Essay <dbl>, Math <dbl>,
## #   Chemistry <dbl>, GPA_3S <dbl>, life_science_score <dbl>
```

Let's see the distribution of GPA in our Brazil University.

```
hist(data$GPA_3S, col = 'blue')
```

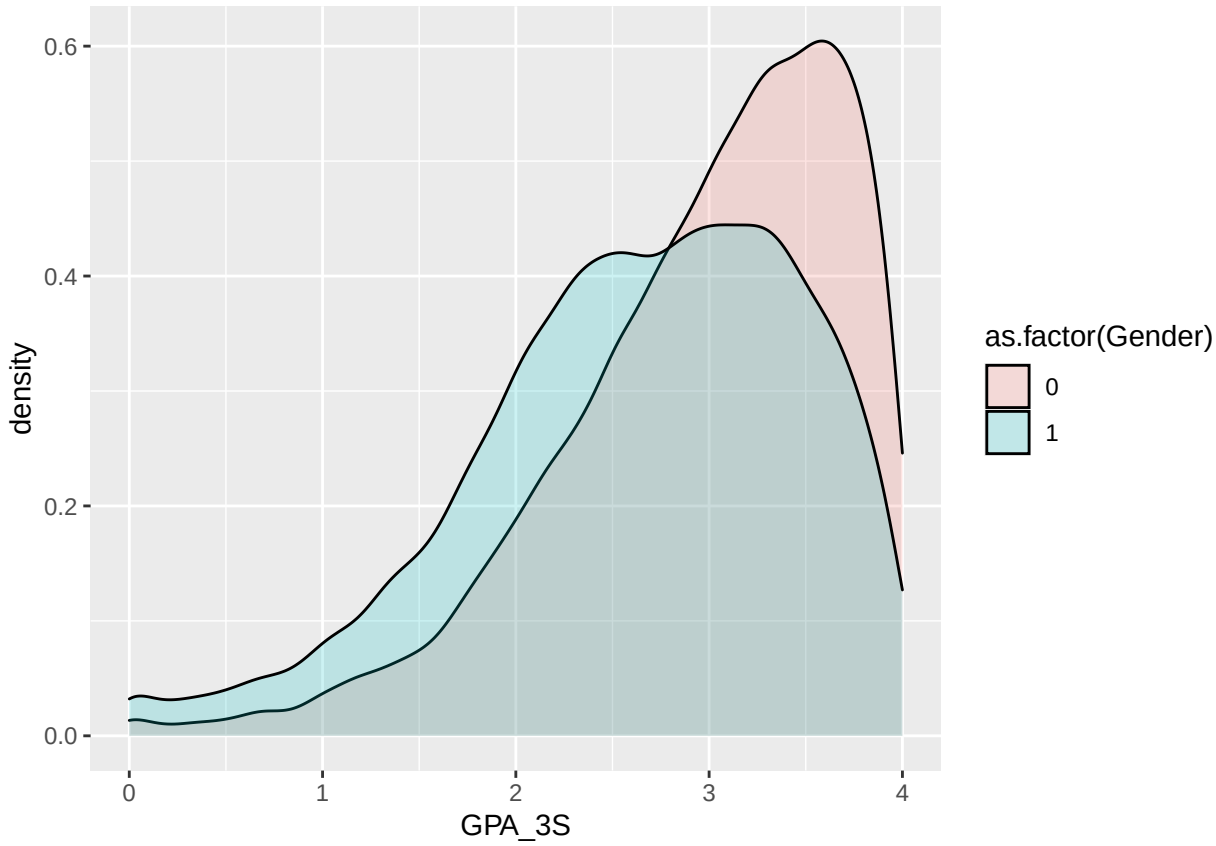



```
head(new_data)
```

```
##   Student_ID Gender Physics Biology History Second_Language Geography
## 1  Student_1      0  622.60  491.56  439.93          707.64    663.65
## 2  Student_2      1  538.00  490.58  406.59          529.05    532.28
## 3  Student_3      1  455.18  440.00  570.86          417.54    453.53
## 4  Student_4      0  756.91  679.62  531.28          583.63    534.42
## 5  Student_5      1  584.54  649.84  637.43          609.06    670.46
## 6  Student_6      1  325.99  466.74  597.06          554.43    535.77
##   Literature Portuguese_Essay   Math Chemistry  GPA_3S life_science_score
## 1     557.09           711.37 731.31    509.80 1.33333            500.680
## 2     447.23           527.58 379.14    488.64 2.98333            489.610
## 3     425.87           475.63 476.11    407.15 1.97333            423.575
## 4     521.40           592.41 783.76    588.26 2.53333            633.940
## 5     515.38           572.52 581.25    529.04 1.58667            589.440
## 6     717.03           477.60 503.82    422.92 1.66667            444.830
##       high_gpa
## 1 Work harder!
## 2 Work harder!
## 3 Work harder!
## 4 Work harder!
## 5 Work harder!
## 6 Work harder!
```

```
library(ggplot2)

ggplot(new_data, aes(GPA_3S, fill = as.factor(Gender))) + geom_density(alpha=.2)
```


density
0.6
0.4
0.2
0.0
0
1
2
3
4
GPA_3S
as.factor(Gender)
0
1