The 5 verbs of dplyr

teachingr.com

Getting started

As always, the first thing we will do is load the tidyverse.

Note: If you haven't yet installed the tidyverse, you'll first have to run the code install.packages("tidyverse").

```
library(tidyverse)
```

Here's the dataframe that we'll analyze in this exercise:

```
scores <-
  data_frame(
  name = c("munir", "raeesa", "rafiq", "maria", "pervaiz", "jamila", "bobby", "saima", "al
  school = c("rawalpindi", "rawalpindi", "rawalpindi", "rawalpindi", "islamabad", "islamab
  teacher = c("jamil", "jamil", "jamil", "jamil",  "sami", "sami", "sami", "fareeha", "far
  sex = c("male", "female", "male", "female", "male", "female", "male", "female", "female"
  math_score = c(4, 3, 2, 4, 3, 4, 5, 4, 5),
  reading_score = c(1, 5, 2, 4, 5, 4, 1, 5, 4)
  )
```

```
Warning: `data_frame()` was deprecated in tibble 1.1.0.
i Please use `tibble()` instead.
```

Let's first take a look at it:

Before we get started, I want to make sure you understand the difference between doing something and assigning it to a new name and just doing it without naming it. For example, make sure you understand the following:

In this exercise we'll typically just print the results and not save them, but it's an option if you want it!

Now we can get to the exercise. Most sections will begin with an example for you to look at. When questions require a written answer, there will be an "Answer" line for you to complete.

Arrange

Example

Question: Sort the data by math_score from high to low. Who had the best math score?

Q1

Question: Sort the data by name from first to last in the alphabet.

Q2

Question: Sort the data by sex so females show up first. Which sex appears to have better reading scores?

Q3

Question: Sort the data by school, then teacher, then sex, then math_score, and finally by reading_score.

Select

Example

Question: Select only the name, math_score, and reading_score columns.

```
scores %>%
  select(name, math_score, reading_score)
```

```
# A tibble: 9 x 3
  name    math_score reading_score
  <chr>        <dbl>         <dbl>
1 munir            4             1
2 raeesa           3             5
```

```
3 rafiq            2             2
4 maria            4             4
5 pervaiz          3             5
6 jamila           4             4
7 bobby            5             1
8 saima            4             5
9 alina            5             4
```

Q1

Question: Select all of the columns except the sex column.

Q2

Question: Select all of the columns except the math_score and reading_score columns.

Q3

Question: Keep all of the columns but rearrange them so sex is the first column.

Filter

Example

Question: Filter to students who are male and went to rawalpindi.

Q1

Question: Filter to students who did well in math (you decide what "well" means).

Q2

Question: Use filter to figure out how many students had a math score of 4 or more and a reading score of 3 or more.

Q3

Question: Explain the errors in each of the following code blocks, then fix it to make it right!

Q4

Question: You are creating a remediation program. Filter to students who got a 3 or worse in either math or reading.

Q5

Question: Filter to students who got a reading score of 2, 3, or 4.

Challenge

Question: Filter to students who have a name that starts with an "m". Hint: type "?substr" in the console and then scroll to the bottom of the help file to see useful examples.

Filter with groups

Example

Question: Filter to teachers whose best math student got a score of 5.

```
scores %>%
  group_by(teacher) %>%
  filter(max(math_score) == 5)
```

```
# A tibble: 5 x 6
# Groups:   teacher [2]
  name    school     teacher sex    math_score reading_score
  <chr>   <chr>      <chr>   <chr>       <dbl>         <dbl>
1 pervaiz islamabad  sami    male            3             5
2 jamila  islamabad  sami    female          4             4
3 bobby   islamabad  sami    male            5             1
4 saima   rawalpindi fareeha female          4             5
5 alina   rawalpindi fareeha female          5             4
```

Q1

Question: Filter to the sex with a mean math score of 4.

Q2

Question: Explain why the following code removes students who have fareeha as their teacher.

Mutate

Example

Question: Both the math and reading scores were actually out of 50 – replace both variables to be 10 times their original values.

Q1

Question: Create a new column called "math_reading_avg" which is the average of a students math and reading scores.

Q2

Question: Create a new column "high_math_achiever" that is an indicator of if a student got a 4 or better on their math_score.

Q3

Question: Create a new column "reading_score_centered" that is a students reading score with the mean of all students' reading scores subtracted from it.

Q4

Question: Create a new column "science_score". You can make up what the actual scores are!

Mutate with groups

Q1

Question: munir and saima both got a 4 for their math score. Explain why why munir has a higher "math_score_centered_by_sex" score.

Q2

Question: Create a "reading_score_centered_by_teacher" column. What can you learn from it?

Q3

Question: Make a "number_of_students_in_class" column that is number of students in a student's class. For example, it should be 4 for munir and 3 for pervaiz.

Summarize

Example

Question: Use the summarize command to find the mean math score for all students.

```
scores %>%
  summarize(math_score_mean = mean(math_score))
```

```
# A tibble: 1 x 1
  math_score_mean
            <dbl>
1            3.78
```

Q1

Question: Use the summarize command to find the mean reading score for all students.

Q2

Question: Use the summarize command to find the median for both math scores and reading scores.

Q3

Question: Look closely at the following code. Why is it throwing an error? How can Rstudio help you see this error?

Summarize with groups

Example

Question: Find the minimum math score for each school.

Q1

Question: Find the maximum math score for each teacher.

Q2

Question: If we grouped by sex, and then summarized with the minimum reading score, how many rows would the resulting data frame have?

Q3

Question: Remember that mutate always keeps the same number of rows but summarize usually reduces the number of rows. Why doesn't the following use of summarize reduce the number of rows?

Q4

Question: Create a data frame with the mean and median reading score by sex, as well as the number of students of that sex.

Combining verbs

Example

Question: Select just the name and math_score columns. Then create a new column "math_score_ec" that is a students math score plus 5 extra credit points. Finally, arrange the data frame by math_score_ec from low to high.

```
scores %>%
  select(name, math_score) %>%
  mutate(math_score_ec = math_score + 5) %>%
  arrange(math_score_ec)
```

```
# A tibble: 9 x 3
  name    math_score math_score_ec
  <chr>        <dbl>         <dbl>
1 rafiq            2             7
2 raeesa           3             8
3 pervaiz          3             8
4 munir            4             9
5 maria            4             9
6 jamila           4             9
7 saima            4             9
8 bobby            5            10
9 alina            5            10
```

Q1

Question: Select every column except the teacher column. Create a new variable called "mean_score" that is the mean of a student's math and reading score. Finally, arrange the data frame by mean_score from low to high.

Q2

Question: Remove any students with sami as a teacher, then find the mean math_score by sex.

Q3

Question: Find the min, max, and median reading_score for female students at rawalpindi school.

Q4

Question: Inspect each of the following code blocks. They both do about the same thing. Which one do you think is preferred from a computer efficiency standpoint?

Challenge

Play around with these tools. Write a question or two that you think best exposes a misunderstanding you had or drills down on an important thing to remember. I'd love to add these questions in the future! Feel free to email what you came up with to zasghar@qau.edu.pk