Dplyr graded lab

Fabrice Rossi

Instructions

Your work must be submitted as a github project and as a zip file. You must:

- create a github repository called **grades-test**. It should be public. You may create a private repository but you have then to invite me as a contributor;
- create a R project on your computer from the github project and make an initial commit with the classical R project configuration files;
- write all your answers in a quarto document **named after your last name**: commit the initial version of this document but dot not include the rendering of the document in the repository;
- each time you are satisfied by your answer to a question, commit the modifications;
- push the commits on a regular basis;
- at the end of the session, make a final commit with a push and then prepare to upload a zip file on moodle with at least:
  - the R project file (ending with `.Rproj`);
  - the data file;
  - the quarto document;
  - the result of rendering your document to html (including the directory with figures)

  The simplest way to produce the zip file is to compress the full directory of the project.

All graphical representations must be done with ggplot2 and all calculations must be done with dplyr and tidyr.

Individual instructions

The instructions contained in this document are student specific. Your data set is unique to you and parts of the instructions depend on the data set. In particular, most names are unique. Any failure to use those specific personal instructions will lead to an automatic fail of the assessment (0/20).

1 Introduction

You are an independent contractor and you have been selected by the Dean of The Clockwork University to analyse the performances of their students. For confidentiality issues, you had to travel to London to get access to the data. But strangely, you are allowed to use github to store your work. The ways of The Clockwork University are mysterious…

1.1 Study organisation

At The Clockwork University each student follows 10 different courses that are organised in 2 modules. Students are divided into 19 groups.

The number of grades per course depends on the course but also on the student as they may fail to attend to an exam. The following table give for each course the number of exams and the assignment to modules.

course	module	number of exams
Airship Piloting and Navigation	1	5
Alchemy and Chemical Engineering	2	5
Clockwork Automata and Robotics	1	5
Cryptography and Codebreaking	1	3
Electrical Engineering and Telegraphy	1	3
Etiquette and Social Graces	1	5
Fashion Design and Textile Innovation	1	10
Historical Archaeology and Antiquarian Studies	1	3
Steam Engine Mechanics and Engineering	2	3
Victorian Literature and Social Commentary	1	8

1.2 Data set

The data set is stored in the `grades.csv` CSV file. The first five rows of the data set are given in the following table:

group	id	course	module	grade
1	8	Airship Piloting and Navigation	1	13.5
1	8	Airship Piloting and Navigation	1	11.0
1	8	Airship Piloting and Navigation	1	10.5
1	8	Airship Piloting and Navigation	1	11.5
1	8	Airship Piloting and Navigation	1	9.5

The data set uses a long format with a small number of columns and a large number of rows. Each row gives the `grade` of a student for a `course`. Grades are between 0 (the worst) and 20 (the best). Students are uniquely identified by the `id` column. To ease data processing each row gives the `group` of the student and the module of the course.

Question 1

Load the data set.

2 Simple analyses

> **!** Important
>
> Did you commit after the first question? Did you push your modifications? Now is a good time to do both.

Question 2

Give the number of grades in the data set directly in the text of your quarto file, in a sentence of the form "The data set contains xxx grades." where xxx is replaced by the number of grades. This must be computed from the data set.

Question 3

The dplyr function `distinct()` can be used to keep only distinct (!) values in a data frame, according to the specified variables. For instance if the data frame `df` has a variable `foo`, then

```
df |> distinct(foo)
```

gives a new data frame with only the column `foo` and such that each value of `foo` appears only once. This applies to multiple variables in a similar way.

Use `distinct()` compute the number of students on the data set and report the value directly in the text as in the previous question.

> **i** Note
>
> Most of the data frames that will be produced during your work will be too long to be included directly in the quarto output. To display part of a data frame, it is recommend to select a few lines with `slice()`, `slice_sample()`, `slice_head()` or `slice_tail()`, and to pass the result to `knitr::kable()`. For instance, assuming the data set was loaded in the `grades` variable, the following code
>
> ```
> grades |>
>   slice_tail(n = 5) |>
>   knitr::kable()
> ```
>
> includes in the quatro render the five last rows of the grade data set, as follows:

group	id	course	module	grade
19	843	Victorian Literature and Social Commentary	1	6.8
19	843	Victorian Literature and Social Commentary	1	10.8
19	843	Victorian Literature and Social Commentary	1	5.2
19	843	Victorian Literature and Social Commentary	1	6.7
19	843	Victorian Literature and Social Commentary	1	6.6

Question 4

Create a `students` data frame that contains one row per student and two variables, `id` and `group` that gives the assignment of each student to their group. Make sure that each student

appears on only once in the data frame. Include a short extract of the data frame in the quarto rendering.

Question 5

Use a graphical representation to display the number of students per group.

> ❗ Important
>
> Don't forget to commit after each question! Now is a good time to push your commits!

Question 6

Compute the average of all the grades in *Electrical Engineering and Telegraphy* in each group and display graphically this average as a function of the group. It is recommend to use `geom_col()` for this task (read the documentation!).

Question 7

Compare graphically the distribution of the grades of the 2 modules.

3 Attendance analysis

Question 8

Compute the number of grades per student and include in your quarto rendering an extract of the resulting data frame. Make sure to keep in the data frame the `id` of the students but also their `group`. Include in the text a markdown table with the minimum, maximum, average and median number of grades per student.

Question 9

Create a data frame that gives for each student their id, their group and the number of grades they obtained in *Fashion Design and Textile Innovation* and include in the quarto rendering a small extract of the result.

Question 10

Compute from the previous data frame the distribution of the number of grades, that is for each number of grades (e.g. 10) the number of students who have exactly this number of grades in *Fashion Design and Textile Innovation*. Represent graphically the results.

Question 11

Using a graphical representation to study whether the number of grades per student in *Fashion Design and Textile Innovation* depends on the group.

> **!** Important
>
> Now is a good time to make sure that your quarto document properly renders to html (and possibly also to pdf). You should in fact do that before each commit.

4 Grade analysis

Question 12

Create a data frame that gives for each student their `id`, their `group` and the average of grades they obtained in each course. Using an adapted pivoting method, create a new data frame with one row per student and 12 columns: one for the `id`, one for the `group` and one per course. Include in the quarto rendering a small extract of the data frame with the `id` and `group` columns and with two of the course columns. You should obtain something like this:

id	group	Airship Piloting and Navigation	Alchemy and Chemical Engineering
534	9	7.5	13.4
559	5	7.2	12.4
457	7	8.9	12.4
461	6	11.3	12.7
256	18	8.6	14.1

Question 13

Show the average grades in *Clockwork Automata and Robotics* as a function of the average grades in *Airship Piloting and Navigation*. Make sure to maximise the readability of the proposed representation.

Question 14

The `cor()` function computes the correlation coefficient between two vectors. It can be used as a summary function in dplyr. Using it, compute the correlation between the average grades in *Electrical Engineering and Telegraphy* and the average grades in *Alchemy and Chemical Engineering* **group by group**.

Question 15

Display the average grades in *Electrical Engineering and Telegraphy* as a function the average grades in *Alchemy and Chemical Engineering* for the students of the group in which those grades are the most correlated (positively or negatively).

Question 16

Let us assume that the final grade of a student is the average of the averages of their grades for each course. Create a data frame with three columns, `id`, `group` and `final grade` based on this definition for the last column. Sort the data frame in decrease order of `final grade` and include in the quarto rendering its first five rows.

Question 17

Find a way to study differences in final grades between groups.

Question 18

To pass the year, a student must fulfil the following conditions:

- have no average grade in a course lower than 5;
- have an average grade in each module larger or equal to 10 (the average in a module is simply the average of the average grades of the courses in the module).

Create a data frame that gives for each student their `id`, their `group`, their `final grade` (as defined before) and a `pass` variable equal to `TRUE` if the student pass the year (and `FALSE` if they do not).

Question 19

Compute and display the number of students who do not pass and yet have a final grade larger or equal to 10.

Question 20

Compute the pass rate per group and represent it graphically.

> **!** Important
>
> Do not forget to:
>
> - make a final rendering test
> - commit the remaining modifications
> - push everything to github
> - zip your work and upload it on moodle