Dplyr graded lab

Fabrice Rossi

! Instructions

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

- Create a GitHub repository named **grades-lab**. It should be public. If you create a private repository, you must invite me as a contributor;
- Create an R project on your computer from the GitHub repository and make an initial commit with the standard 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 do not include the rendered output in the repository;
- Commit your changes each time you are satisfied with your answer to a question;
- Push your commits regularly;
- At the end of the session, make a final commit and push, then prepare to upload a zip file to Moodle containing at least:
  - The R project file (with the `.Rproj` extension);
  - The data file;
  - The Quarto document;
  - The rendered HTML output of your document (including the figures directory).

  The easiest way to create the zip file is to compress the entire project directory.

All graphical representations must be created using **ggplot2**, and all calculations must be performed using **dplyr** and **tidyr**. Grading will consider your proper use of Git.

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 NeoTokyo Institute of Technology to analyse the performances of their students. For confidentiality issues, you had to travel to NeoTokyo to get access to the data. But strangely, you are requested to use github to store your work. The ways of The NeoTokyo Institute of Technology are mysterious…

1.1 Study organisation

At The NeoTokyo Institute of Technology each student follows 10 different courses that are organised in 3 trimesters. 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 names of the courses, their assignment to trimesters and their number of exams are given in the `courses.csv` file which as the following columns:

- `course`: the name of the course
- `course_id`: a unique key to identify the course
- `trimester`: the trimester of the course
- `nb_exams`: the number of exams in the course

Question 1

Load the course description file.

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

Question 2

Display the content of the file in a nicely formatted table in your document. Courses should be sorted in alphabetical order and column names should be written in English. This has to be done in `R` for instance using `knitr::kable()`.

You are expected to produce something like this:

course name	identifier	trimester	number of exams
Artificial Intelligence and Machine Learning	6	3	8
Bioengineering and Genetic Modification	3	1	4
Corporate Espionage and Industrial Sabotage	9	3	10
Cybernetic Implants and Augmentations	1	1	8
Darknet Culture and Subversive Technologies	8	1	3
Hacking and Network Security	2	1	10
Neurotechnology and Mind Control	5	1	3
Post-Human Philosophy and Ethics	10	2	7
Urban Planning and Megacity Design	7	2	5
Virtual Reality and Augmented Reality	4	1	3

> **!** Important
>
> Did you verify the rendering before committing?

1.2 Students

Students are described in the `students.csv` CSV file. The first five rows of this data set are given in the following table:

id	group	birth_date	sex
1	7	2002-08-06	M
2	16	2000-04-16	M
3	1	2002-11-06	F
4	7	2002-10-12	F
5	8	2002-04-16	F

Students are uniquely identified by the `id` column. Their group is given by the `group` column which contains the unique identifier of the group. The `birth_date` column contains the birth date of the students. The `sex` gives the sex of each student (`F` or `M`). It is also highly recommend to have `group` loaded as a factor and `id` as an integer.

Question 3

Load the student data set. Make sure the birth dates are correctly identified as `Date` objects by including the `class` of the corresponding column in the rendered text. For instance, include a sentence of the following form:

The `birth_date` column of the `students` data frame is of class `Date`.

Make also sure that the `sex` is identified as a factor.

1.3 Grades

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

id	course_id	grade
3	6	14.5
3	6	13.5
3	6	7.0
3	6	15.0
3	6	13.0

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 identified by the `id` column which contains their unique identifier, while the course is identified by the `course_id` column which contains the unique identifier of the course.

Unfortunately, The NeoTokyo Institute of Technology does not maintain very high data management standards and the `grades.csv` file is not perfectly encoded:

- rather than using a comma as the separator, the file uses the following character: :.
- rather than using `NA` to represent missing data, the file uses this: unknown.

Question 4

Load the grade data set despite the encoding issues.

2 Student population analysis

Question 5

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

Question 6

Show the gender balance in each group on a single graphical representation.

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

> **i** Note
>
> Manipulating dates is difficult. Fortunately, the `lubridate` package can be used to simplify those manipulations. In particular, if an event took place at a certain date stored in variable `event_date`, one can compute the number of years from this event to today with the following code:

```
event_date <- date("2024-02-01")
time_length(today() - event_date, unit="year")
```

```
[1] 1.828884
```

> The value is fractional and can be rounded to the nearest integer when needed (using the `round` function).

Question 7

Show graphically the distribution of the age of the students.

Question 8

Compute the median age of the students in each group and include in your rendering a table giving the median age for each group. The median should be reported as an integer.

Question 9

Build a table with the id, sex and age of the oldest student in each group. Sort the table by decreasing age order and include it in the rendering.

3 Simple grade analysis

Question 10

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. Note that missing grades should obviously not be included in the total.

Question 11

Compute the average of all the grades in *Bioengineering and Genetic Modification* 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 12

Compare graphically the distribution of the grades of the 3 trimesters.

4 Attendance analysis

> **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:

id	course_id	grade
863	7	9.0
863	7	8.0
863	4	7.7
863	4	9.9
863	4	11.3

Question 13

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` and their `sex`. Include in the text a markdown table with the minimum, maximum, average and median number of grades per student.

Question 14

Compare graphically the distribution of the number of grades for female and male students.

Question 15

Create a data frame that gives for each student their id, their group and the number of grades they obtained in *Corporate Espionage and Industrial Sabotage* and include in the quarto rendering a small extract of the result.

Question 16

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 *Corporate Espionage and Industrial Sabotage*. Represent graphically the results.

Question 17

Using a graphical representation to study whether the number of grades per student in *Corporate Espionage and Industrial Sabotage* depends on the group.

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

Question 18

Take the analysis of the previous question one step further to see if there is a dependency of the number of grades with respect to both the group and the sex of the students.

5 Grade analysis

Question 19

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	Artificial Intelligence and Machine Learning	Bioengineering and Genetic Modification
436	14	13.12	9.00
529	10	13.88	10.12
633	19	12.31	12.75
551	14	12.57	11.88
75	1	12.25	12.62

Note that the table displays only 2 courses but the data frame must include all the courses.

Question 20

Show the average grades in *Hacking and Network Security* as a function of the average grades in *Neurotechnology and Mind Control*. Make sure to maximise the readability of the proposed representation.

Question 21

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 *Darknet Culture and Subversive Technologies* and the average grades in *Artificial Intelligence and Machine Learning* **group by group**.

Question 22

Display the average grades in *Darknet Culture and Subversive Technologies* as a function the average grades in *Artificial Intelligence and Machine Learning* for the students of the group in which those grades are the most correlated (positively or negatively).

Question 23

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 four columns, `id`, `group`, `sex` 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 24

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

Question 25

Include in the analysis conducted in the previous question the effect of `sex`.

Question 26

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 trimester larger or equal to 10 (the average in a trimester is simply the average of the average grades of the courses in the trimester).

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 27

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

Question 28

Compute the pass rate per group and represent it graphically.

Question 29

Determine the course that is the most responsible of student failure owing to grades below than 5.

Question 30

Display the effect of age on success. The goal **is not** to display the average age for students who passed but to show the success rate as a function of student age.

> **!** 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