Dplyr and Tidyr lab

Fabrice Rossi

This lab is dedicated to dplyr and tidyr. Your work must be submitted as a zip file of a R project containing:

- the R project file (ending with `.Rproj`);
- the data files (csv or rds format);
- a quarto document with all your answers (use one second level section per exercise and one third level section per question);
- the result of rendering your document to html.

The quarto document must be renderable to html **without any modification** upon unzipping. In particular, all file names **must be local** and constructed using the here package. The use of `install.package` in the code is **forbidden**.

Answers can be written in English or French. Grading will take into account the quality of the code and the choice of graphical representations.

Exercise 1

We study in this exercise the Spotify top songs data set[1]. It contains the top 2000 songs on Spotify from 2000 to 2019.

The dplyr verb `distinct` can be used to keep only unique values in a data frame. More precisely, if a data frame `df` contains multiple columns including one named `A`, the expression

```r
df %>% distinct(A)
```

returns a data frame including only `A` and with no duplicates in this column. Multiple columns can be selected in `distinct` leading to uniqueness being defined on the combined values. To keep all columns while enforcing uniqueness for only some of them, use

```r
df %>% distinct(A, B, .keep_all = TRUE)
```

Question 1 Using `distinct` (among other functions) compute the number of different songs, artists and musical genre that have been included in the data set. Be aware that songs may have covers by another artist and thus the title (in the `song` variable) is not sufficient to determine uniqueness (in other words two songs with the same title should be considered distinct if they do not share the same artist). Make sure in all the following questions to consider whether or not duplicate songs may have an effect on the results.

Include the results directly in a presentation text in the markdown document, in the form: the data set contains 1926 songs. Notice that the numerical value cannot be copy-pasted from e.g. the console, but has to be included in the text during knitting.

[1] Available at `https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019`

Question 2 Compute the number of distinct artists per year and include it in the knitted document as a nicely formatted table (using for instance `knitr::kable`). If an artist appears twice in a given year (for two distinct songs) they should only be counted once, but if they appear in different years, they should be counted once per year.

Question 3 Find the most popular artist in the data set, i.e. the artist with the largest number of songs in the data set. Make sure to count each song only once. Include the name of this artist and the number of songs in the text of the knitted document (as in question 1).

Question 4 Compute the minimum, maximum, mean and median `tempo` as well as the number of songs, for each musical genre. Make sure that each pair (artist, song) is used only once in the analysis. Gather the information in a single table included in the knitted result (as in question 2).

Question 5 Compute the mean liveness and the mean danceability per year in a single data frame.

Question 6 Draw *on a single graph* the temporal evolution of the mean annual liveness and the mean annual danceability.

Exercise 2

We study in this exercise the students' dropout data set from the UCI[2]. To ease data loading, the file is available in `Rds` format on the page of the course. It should be loaded as follows:

```
dropout <- readRDS(...) ## replace ... by the file name and access path
```

Some variables have been recoded (based on the documentation) to replace integer codes by readable labels.

Question 1 Compute the median "Admission grade" conditioned both on the Target vraiable and on the "Marital status".

Question 2 Transform the data frame obtained in question 1 in order to have four variables: one for the "Marital status", one for each possible value of the Target variable. Each row should correspond to a specific marital status (given in the corresponding column) while the other columns should contain the corresponding median grade. Include the resulting table in the knitted document as explained in Exercise 1. It should have the following form (only the first 3 rows are shown):

Marital status	Dropout	Graduate	Enrolled
single	123.35	127.30	124.05
married	126.50	130.00	122.95
divorced	126.50	126.00	130.20

Question 3 Compute the conditional median of all variables related to "Curricular units" given the value of the Gender variable.

Question 4 Using the `pivot_*` functions, transform the data in order to include in the knitted result a table of the following form (only the first 3 rows are shown):

[2] Available at https://archive-beta.ics.uci.edu/ml/datasets/predict+students+dropout+and+academic+success.

Units	Male	Female
Curricular units 1st sem (credited)	0.00	0.00
Curricular units 1st sem (enrolled)	6.00	6.00
Curricular units 1st sem (evaluations)	8.00	8.00