PPGCOMP - FURG | 23148P - Data Visualization and Exploratory Data Analysis | 02/2024

This notebook contains the solution for Task 03 of the course 23148P - Data Visualization and Exploratory Data Analysis - 02/2024 of the Graduate Program in Computing at FURG (PPGCOMP-FURG).

Professor: Dr. Adriano Velasque Werhli.

Student: Vitor Avelaneda.

- **Contact:** avelaneda.vitor@gmail.com

The repository with the notebooks can be accessed here!

Exercises

1. Load the tab delimited file small_file.txt using the function read_delim. The loaded data should be attributed to the variablem my.data. After having the data in the variable:
    - Inspect the data with the function `head()`, `view()` and `glimpse()`
    - Using the function `filter()` from tidyverse library show only the rows that are from category D
    - Using the solution above, show only rows with category D and ordered by lenght
    - Calculate de mean of the Lenght of Category D and of Category A using the filters above and the function `mean()`. Remember that you can attribute the resulto of a pipe to a variable.
2. You have been provided the file student_grade.csv. Load this data and put it in a tidy format. Think about:
    - Which of the columns are annotations and which are measurements?
    - How many different types of measurement are there?
    - Are all of the measurements of the same type in a single column?
    - What is the name of the variable being measured? I has its name in on column?
    - After tidying are there any NA values which should be removed?
    - Are there any columns with repeated information in its rows that should be removed?

- Remove NA
- What is the mean and standard deviation of the grades in questions 1 and 2?

Solution 1:

Importing the data and assigning it to the variable `my.data`.

In [1]:
```
my.data <- read.delim('small_file.txt', header = TRUE)
```

Checking if the dplyr package is installed and importing the package:

In [2]:
```
if (!requireNamespace("dplyr", quietly = TRUE)) install.packages("dplyr")
library(dplyr)
```

```
Anexando pacote: ‘dplyr’


Os seguintes objetos são mascarados por ‘package:stats’:

    filter, lag


Os seguintes objetos são mascarados por ‘package:base’:

    intersect, setdiff, setequal, union
```

Visualizing the data:

In [3]:
```
head(my.data)
View(my.data)
glimpse(my.data)
```

A data.frame: 6 × 3

	Sample	Length	Category
	<chr>	<int>	<chr>
1	x_1	45	A
2	x_2	82	B
3	x_3	81	C
4	x_4	56	D
5	x_5	96	A
6	x_6	85	B

A data.frame: 40 × 3

Sample	Length	Category
<chr>	<int>	<chr>
x_1	45	A
x_2	82	B
x_3	81	C
x_4	56	D
x_5	96	A
x_6	85	B
x_7	65	C
x_8	96	D
x_9	60	A
x_10	62	B
x_11	80	C
x_12	63	D
x_13	50	A
y_1	64	B
y_2	43	C
y_3	98	D
y_4	78	A
y_5	53	B
y_6	100	C
y_7	79	D
y_8	84	A
y_9	68	B
y_10	99	C

Sample	Length	Category
<chr>	<int>	<chr>
y_11	65	D
y_12	55	A
y_13	98	B
z_1	56	C
z_2	83	D
z_3	81	A
z_4	69	B
z_5	50	C
z_6	72	D
z_7	54	A
z_8	56	B
z_9	87	C
z_10	84	D
z_11	80	A
z_12	68	B
z_13	95	C
z_14	93	D

```
Rows: 40
Columns: 3
$ Sample   <chr> "x_1", "x_2", "x_3", "x_4", "x_5", "x_6", "x_7", "x_8", "x_9"…
$ Length   <int> 45, 82, 81, 56, 96, 85, 65, 96, 60, 62, 80, 63, 50, 64, 43, 9…
$ Category <chr> "A", "B", "C", "D", "A", "B", "C", "D", "A", "B", "C", "D", "…
```

Filtering the data by `category == D`.

In [4]:
```
my.data %>%
  filter(Category == "D")
```

A data.frame: 10 × 3

Sample	Length	Category
<chr>	<int>	<chr>
x_4	56	D
x_8	96	D
x_12	63	D
y_3	98	D
y_7	79	D
y_11	65	D
z_2	83	D
z_6	72	D
z_10	84	D
z_14	93	D

Organizing `category == D` by `Length`.

In [5]:

```
my.data %>%
  filter(Category == "D") %>%
  arrange(Length)
```

A data.frame: 10 × 3

Sample	Length	Category
<chr>	<int>	<chr>
x_4	56	D
x_12	63	D
y_11	65	D
z_6	72	D
y_7	79	D
z_2	83	D
z_10	84	D
z_14	93	D
x_8	96	D
y_3	98	D

Calculating the mean of `Length`:

In [6]:

```r
D_data <- data.frame(
  my.data %>%
    filter(Category == "D")
)

mean(D_data$Length)
```

```
78.9
```

Assigning the mean to the variable `D_mean`:

In [7]:

```r
D_mean <- my.data %>%
  filter(Category == "D") %>%
  summarise(mean_length_D = mean(Length)) %>%
  pull(mean_length_D)
```

Calculating the mean of `Length`:

In [8]:

```
A_data <- data.frame(
  my.data %>%
    filter(Category == "A")
)

mean(A_data$Length)
```

68.3

Filtering the data by `category == A`, calculating the `mean` of `Length`, and assigning it to the variable `A_mean`.

In [9]:

```
A_mean <- my.data %>%
  filter(Category == "A") %>%
  summarise(mean_length_A = mean(Length)) %>%
  pull(mean_length_A)
```

Solution 2:

Checking if the `readr` package is installed and importing the package:

In [10]:

```
if (!requireNamespace("readr", quietly = TRUE)) install.packages("readr")
library(readr)
```

Importing the data and assigning it to the variable `my.data2`.

In [11]:

```
my.data2 <- read_csv("student_grade.csv")
```

```
Rows: 43 Columns: 14
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr  (2): Class, Student
dbl (12): Year, Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10, Q11

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Visualizing the data:

In [12]:

```
View(my.data2)
glimpse(my.data2)
```

A spec_tbl_df: 43 × 14

Year	Class	Student	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
<dbl>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2022	Student	Lucca	7.50	6.23	6.50	7.15	NA	5.43	8.58	8.19	7.96	7.92	6.48
2022	Student	Salles	10.00	10.00	10.00	10.00	10.00	NA	10.00	10.00	10.00	10.00	10.00
2022	Student	Bueno	9.50	9.00	9.00	9.25	9.25	8.00	9.75	9.75	7.50	7.25	8.00
2022	Student	Simas	9.50	9.00	9.00	9.25	9.25	8.00	9.75	9.75	7.50	7.25	8.00
2022	Student	Goncalves	1.67	3.17	4.67	1.67	4.00	1.67	4.83	0.83	0.83	1.67	1.67
2022	Student	Dornelles	9.10	8.75	9.83	9.00	9.75	9.00	9.50	9.25	9.00	9.18	9.36
2022	Student	John	9.53	7.07	8.40	7.60	7.67	8.38	7.27	6.87	7.80	8.00	7.93
2022	Student	Ramos	6.25	6.23	7.15	6.38	6.00	2.00	7.23	6.62	6.69	6.62	6.31
2022	Student	Junior	10.00	10.00	10.00	10.00	10.00	NA	10.00	10.00	10.00	10.00	10.00
2022	Student	Freitas	9.68	8.92	9.44	9.68	9.28	8.84	9.60	9.48	9.60	9.68	9.64
2022	Student	Zelira	9.36	8.48	8.84	8.92	8.56	7.92	9.24	9.16	9.08	8.72	8.84
2022	Student	Francisca	9.60	9.50	9.70	9.75	9.75	8.86	9.75	9.85	9.85	9.75	9.65
2022	Student	Vitor	9.61	9.78	9.91	9.17	9.87	9.27	9.91	9.50	9.65	9.43	9.26
2022	Student	Bruno	7.50	6.23	6.50	7.15	6.65	5.43	8.58	8.19	7.96	7.92	6.48
2022	Student	Rafael	9.52	9.17	NA	9.40	9.31	9.13	9.54	9.15	8.98	9.25	9.62
2022	Student	Pinto	9.56	9.35	9.80	8.55	9.15	8.28	9.95	9.83	9.80	9.80	9.85
2022	Student	Nunes	8.08	6.81	8.93	7.19	7.93	6.00	9.22	7.85	8.52	7.56	6.04
2022	Student	Andrade	8.45	6.87	8.55	7.47	7.35	5.75	8.68	7.87	7.68	7.48	8.07
2022	Student	Santos	9.37	8.38	9.38	8.62	8.90	7.67	9.59	9.55	9.45	9.21	9.28
2022	Student	Lima	8.49	8.35	8.68	8.86	8.54	8.03	9.46	8.70	8.70	8.41	8.24
2022	Student	Gabriel	9.83	9.55	9.45	9.68	9.70	9.85	9.93	9.68	9.65	9.60	9.73
2022	Student	Pereira	9.06	7.74	9.42	9.05	7.89	8.17	9.79	9.11	9.61	7.84	8.06
2022	Student	Luciano	10.00	9.36	10.00	9.43	9.21	8.27	10.00	9.93	9.71	9.79	9.57

Year	Class	Student	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
<dbl>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2022	Student	Gleiser	4.50	3.95	4.81	4.00	4.38	3.25	5.33	4.48	4.81	4.33	4.76
2022	Student	Rafaela	8.30	5.03	7.12	6.00	6.63	6.18	8.48	6.61	8.09	7.94	5.73
2022	Student	Silvio	9.05	8.37	9.16	8.68	9.16	8.39	9.26	7.95	9.11	8.53	8.42
2022	Student	Pedro	9.84	9.81	9.91	9.77	9.84	9.39	9.98	9.79	9.95	9.81	9.86
2022	Student	Adriano	7.33	6.94	7.45	7.00	6.97	6.71	9.09	7.50	8.12	6.85	7.13
2022	Student	Carneiro	8.78	7.89	8.19	8.07	8.04	7.62	8.67	8.19	8.48	8.48	8.27
2022	Student	Andre	7.95	8.00	8.54	8.32	8.37	8.20	9.17	8.59	8.10	8.15	8.12
2022	Student	Machado	8.60	7.52	9.03	7.76	8.24	7.50	9.36	8.00	8.85	8.21	7.45
2022	Student	Ribeiro	7.16	7.50	7.85	7.45	7.55	6.75	9.05	7.50	6.60	6.30	6.05
2022	Student	Augusto	8.51	8.45	8.78	8.63	8.78	8.38	9.40	9.05	8.38	8.59	8.41
2022	Student	Marcela	9.31	9.55	9.71	9.36	9.55	8.94	9.69	9.43	9.62	9.30	9.37
2022	Student	Silva	9.54	8.63	9.21	8.88	8.88	8.38	9.71	9.42	8.88	8.50	8.79
2022	Student	Oliveira	7.42	6.68	7.53	7.11	6.89	7.00	8.26	7.16	7.21	7.21	6.67
2022	Student	Cleonice	9.10	9.59	9.91	9.73	9.82	9.29	9.86	9.45	9.91	9.86	9.43
2022	Student	Emanuela	7.88	7.54	7.97	7.94	7.64	4.07	8.42	7.60	7.91	7.63	7.49
2022	Student	Luiza	9.47	8.94	9.44	8.65	8.90	7.70	9.87	9.47	9.27	8.90	9.20
2022	Student	Nunes	9.45	8.78	9.25	8.90	8.95	8.52	9.22	9.20	9.18	8.90	8.97
2022	Student	Samara	5.63	4.93	5.42	5.69	5.09	5.23	7.01	5.31	5.44	5.51	4.97
2022	Student	Marcela	7.87	7.98	9.00	8.08	8.61	7.30	9.73	8.33	8.33	8.29	7.59
2022	Student	Regina	8.66	7.93	8.48	8.48	8.34	7.50	8.93	8.66	8.24	8.24	8.31

```
Rows: 43
Columns: 14
$ Year    <dbl> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 20…
$ Class   <chr> "Student", "Student", "Student", "Student", "Student", "Studen…
$ Student <chr> "Lucca", "Salles", "Bueno", "Simas", "Goncalves", "Dornelles",…
$ Q1      <dbl> 7.50, 10.00, 9.50, 9.50, 1.67, 9.10, 9.53, 6.25, 10.00, 9.68, …
$ Q2      <dbl> 6.23, 10.00, 9.00, 9.00, 3.17, 8.75, 7.07, 6.23, 10.00, 8.92, …
$ Q3      <dbl> 6.50, 10.00, 9.00, 9.00, 4.67, 9.83, 8.40, 7.15, 10.00, 9.44, …
$ Q4      <dbl> 7.15, 10.00, 9.25, 9.25, 1.67, 9.00, 7.60, 6.38, 10.00, 9.68, …
$ Q5      <dbl> NA, 10.00, 9.25, 9.25, 4.00, 9.75, 7.67, 6.00, 10.00, 9.28, 8.…
$ Q6      <dbl> 5.43, NA, 8.00, 8.00, 1.67, 9.00, 8.38, 2.00, NA, 8.84, 7.92, …
$ Q7      <dbl> 8.58, 10.00, 9.75, 9.75, 4.83, 9.50, 7.27, 7.23, 10.00, 9.60, …
$ Q8      <dbl> 8.19, 10.00, 9.75, 9.75, 0.83, 9.25, 6.87, 6.62, 10.00, 9.48, …
$ Q9      <dbl> 7.96, 10.00, 7.50, 7.50, 0.83, 9.00, 7.80, 6.69, 10.00, 9.60, …
$ Q10     <dbl> 7.92, 10.00, 7.25, 7.25, 1.67, 9.18, 8.00, 6.62, 10.00, 9.68, …
$ Q11     <dbl> 6.48, 10.00, 8.00, 8.00, 1.67, 9.36, 7.93, 6.31, 10.00, 9.64, …
```

> Which of the columns are annotations and which are measurements?

- The columns `Year`, `Class`, and `Students` are annotations, and the others are measures.

> How many different types of measurement are there?

- The data contains 11 measurement columns, each associated with a different question and representing each student's score.

> Are all of the measurements of the same type in a single column?

- No. The data contains multiple columns with scores assigned to each question.

> What is the name of the variable being measured? I has its name in on column?

- Analyzing the context of the data, the columns from Q1 to Q11 represent different questions from an assessment. They do not have clear names but are likely scores obtained in the evaluation.

> After tidying are there any NA values which should be removed?

In [13]:

```
if (!requireNamespace("tidyr", quietly = TRUE)) install.packages("tidyr")
if (!requireNamespace("stringr", quietly = TRUE)) install.packages("stringr")

library(tidyr)
library(stringr)

my.data2.tidy <- my.data2 %>%
  pivot_longer(cols = starts_with("Q"),
               names_to = "Question",
               values_to = "Score") %>%
  mutate(Question = str_remove(Question, "Q"))

head(my.data2.tidy)
```

A tibble: 6 × 5

Year	Class	Student	Question	Score
<dbl>	<chr>	<chr>	<chr>	<dbl>
2022	Student	Lucca	1	7.50
2022	Student	Lucca	2	6.23
2022	Student	Lucca	3	6.50
2022	Student	Lucca	4	7.15
2022	Student	Lucca	5	NA
2022	Student	Lucca	6	5.43

- Yes, there are NA values to be removed

> Are there any columns with repeated information in its rows that should be removed?

- There is a column for the year and another for the student's class; although this data may be repetitive, it is still important.

> Remove NA

In [14]:

```
my.data2.tidy <- my.data2.tidy %>%
  drop_na()

head(my.data2.tidy)
```

A tibble: 6 × 5

Year	Class	Student	Question	Score
<dbl>	<chr>	<chr>	<chr>	<dbl>
2022	Student	Lucca	1	7.50
2022	Student	Lucca	2	6.23
2022	Student	Lucca	3	6.50
2022	Student	Lucca	4	7.15
2022	Student	Lucca	6	5.43
2022	Student	Lucca	7	8.58

> What is the mean and standard deviation of the grades in questions 1 and 2?

Mean and standard deviation Question 1:

In [15]:
```
stats_question1 <- my.data2.tidy %>%
  filter(Question == 1) %>%
  summarise(
    mean_score = mean(Score, na.rm = TRUE),
    sd_score = sd(Score, na.rm = TRUE)
  )

stats_question1
```

A tibble: 1 × 2

mean_score	sd_score
<dbl>	<dbl>
8.500233	1.614545

Mean and standard deviation Question 2:

In [16]:
```
stats_question2 <- my.data2.tidy %>%
  filter(Question == 2) %>%
  summarise(
```

```r
    mean_score = mean(Score, na.rm = TRUE),
    sd_score = sd(Score, na.rm = TRUE)
  )

stats_question2
```

A tibble: 1 × 2

mean_score	sd_score
<dbl>	**<dbl>**
7.952326	1.616007