Introduction to R

Second course - Questions

Jean-Baptiste Guiffard and Florence Lecuit

October 18, 2024

Load libraries

```
library(tidyverse) #tidyverse contains ggplot2 library
```

For the first two exercises, we will use a built-in dataset in R (already loaded), called **iris**. This dataset gives petal and sepal lengths and widths for different flowers.

```
head(iris) #to see the first rows of the dataset
```

Exercise 1 : Analyzing numeric variables

1. How many variables are in the dataset? What type of variables?
2. What graphs can you use to observe the distribution of a numeric variable?
3. Create a histogram of the variable *Sepal.Length*. What does this graph show us?
4. Create a boxplot of the variable *Sepal.Width*. Group by flower species. What do you observe?

Exercise 2 : Analyzing more than one numeric variable

1. Plot petal length (y) against petal width (x). What is represented in this plot? What do you observe?
2. Change the color and symbol of the points. Change the points by flower species, to visualize the different groups.
3. Add axis titles and a main title to the plot.

Exercice 3 : GGplot2

Load dataset

```
library(tidyverse)
data_pollution <- read.csv2('DATA/co2_clean.csv', sep=";")
```

1. Using the dataset data_pollution only for the year 2015, create a histogram showing the distribution of CO2 emissions per capita. (Hint: Use ggplot() with geom_histogram().)

2. Create the GDP per capita variable and then create a scatter plot that shows the relationship between GDP per capita (log-transformed) and CO2 emissions per capita (log-transformed). (Hint: Use geom_point() and log-transform the axes inside aes().)

3. Modify the scatter plot by changing the size, shape, and color of the points. Make the points red with a black outline. (Hint: Look into geom_point() parameters like size, shape, and colour.)

Exercice 4

```
Metadata_Country <- read.csv2('DATA/Metadata_Country.csv', sep=",")
join_pollution_wb_data <- data_pollution %>%
  dplyr::inner_join(Metadata_Country, by = c("iso_code" = "Country.Code"))
join_pollution_wb_data <- join_pollution_wb_data %>%
  filter(country != "") %>%
  filter(IncomeGroup !="")
```

1. From this database :

- Create two variables GDP per capita and CO2 per capita in kg ;
- Create a new database that, for the period [1990;2020], gives the average of these two variables by country;
- Delete the columns with missing values.

2. Create a bar chart showing the number of countries by their income group in 2015 (IncomeGroup). Color the bars by IncomeGroup. (Hint: Use geom_bar() with aes(fill=IncomeGroup).)

3. From the dataset in 2015, create a bar chart showing the average CO2 emissions per capita for each income group, and color the bars using a custom palette (e.g., Reds). (Hint: Use geom_bar() and scale_fill_brewer().)

4. Use facet_wrap() to create multiple scatter plots of GDP per capita vs. CO2 emissions per capita, one for each income group. (Hint: Facet by IncomeGroup using facet_wrap(~ IncomeGroup).)

5. Add a linear regression line to the scatter plot of GDP per capita vs. CO2 emissions per capita. Display the regression line without the confidence interval. (Hint: Use geom_smooth(method="lm", se=FALSE).)