Introduction to R

First course - Questions

Jean-Baptiste Guiffard and Florence Lecuit

October 11, 2024

Contents

Create R project

Load dataset

Exercise 1: Manipulating vectors

1. What is the length of vector a?
2. Try doing a[1:3], what is the result?
3. Create new vector *agrumes* with only values *orange* and *lemon*
4. Try doing a[-1], what is the result?
5. Sort vector a alphabetically.
6. Combine vectors b and c into a data-frame. What is the problem?
7. Combine vectors a and b into a data-frame. Why does this work?

Exercise 2: Describing a data-frame

1. What type of object is *data_pollution* ?
2. How many observations and variables does this dataset contain?
3. How many missing values are there in this dataset?
4. What type of variables are in this dataset?

Exercise 3: Subsetting, Selecting Columns, and Dropping Duplicates

1. Create a new data frame that contains only the variables: `country`, `year`, `population`, `gdp`, and `co2`.
2. Filter the dataset to include only data for the country "France".
3. Subset the data to include only countries with a population greater than 50 million. Which variable should you use to do this?
4. Check if there are any duplicate rows in the dataset and drop them if they exist.

Exercise 4: Creating Variables

1. Create a new variable in the dataset called `gdp_per_capita`, which calculates GDP per capita (GDP divided by population).

2. Similarly, create a new variable called `co2_per_capita`, which calculates CO_2 emissions per capita.
3. Create a new variable that groups countries into quartiles based on GDP per capita.
4. Are there any missing values in the new variables you created? If so, filter out the rows where these values are missing.

Exercise 5: Basic Statistics

1. For the new dataset, calculate the mean, minimum, and maximum for the `gdp_per_capita` and `co2_per_capita` columns.
2. Group the data by `country` and calculate the average `co2_per_capita` for each country.
3. Group the data by `quartiles` and calculate the average `co2_per_capita` for each quartile.

Exercise 6: Merging Datasets

1. Load a new dataset that contains additional information on countries. Assign it to a variable called `df2`.
2. Which variable(s) would you use as the key(s) to merge the two datasets (`df1` and `df2`)? Explain why.
3. Perform an inner join between `df1` and `df2` based on the common key(s).
4. After merging the datasets, check how many new columns were added. How many columns and rows does the new data frame now contain?
5. Check for any missing values in the merged dataset after the join. Which countries or years might be missing from one of the datasets?
6. Group the data by `continents` and calculate the average `co2_per_capita` for each continent.