Exercises for Lecture 6

Statistical Computing with R, 2023-24

Exercise 1

1. Write a function that takes a dataframe as input. In the function, generate the following 3 objects as output:

- A vector with the dimensions of the dataframe;
- A vector with the column names of the dataframe;
- A dataframe that contains (1) the column names of all the numeric variables, (2) a column with the mean of these variables. Round the mean to 3 digits.

  Let the output of the function be a list with these generated objects. Name the objects "dimensions", "names", and "summary".

2. Test your function on the built-in `iris` dataframe in `R` (check `View(iris)`) and assign the output to a variable.
3. Select and print the object in your list that contains the means of the numeric variables.
4. Select and print the first item of the second object in your list.

Exercise 2

Let's consider again the `heights` dataset from the `brolgar` package.

1. Write a function that given an input country, selects data from that country and produces a scatter plot with year on the x axis, and mean height on the y axis.
2. Edit the `type` argument of `plot( )` in such a way that two consecutive points are joined by a (straight) line.
3. Add to the function an argument that allows to change the shape of points and their colour. Set a default value for these two arguments.
4. Create a barplot that compares mean height in 1990 in the following countries: Argentina, Brazil, Honduras, Nicaragua, Panama, Sri Lanka and Vietnam. Make sure that the bars in the barplot are horizontal rather than vertical.
5. Use histograms and density plots to compare the distribution of mean heights in Asian countries in 1950 and 1990.

Exercise 3

Consider again the data on population by country that you downloaded from the World Bank website during lecture 3.

1. Select the top 20 countries by population in 1990
2. Make a scatter plot that compares the population in 1960 and 1970 in such countries. Make sure that `xlim` and `ylim` are the same so that it is easier to compare x and y

As you can notice, the presence of two countries with much larger population make it difficult to distinguish most of the points from countries with smaller population. To make such points visible, for the rest of the exercise we will use the log_{10} of population.

3. Redo the chart you made at point 2, with log_{10}(population) on both axes
4. Add the line $y = x$ to the plot to make comparisons easier. How many countries experienced a population growth?
5. Create a chart with 4 panels. Each panel should compare the population in 1960 in the countries selected at point (1) to the population in the same countries in: 1980, 2000, 2010 and 2020

Exercise 4

In the exercises of lecture 4 you created a function that computes

$$f(x) = \begin{cases} -5 & \text{if } x \leq -3 \\ \log(x+5) & \text{if } -3 < x < 1 \\ 2 & \text{if } x = 1 \\ \sqrt{(x+3)} & \text{if } 1 < x \leq 14 \\ \log(x) & \text{if } x > 14 \end{cases}$$

1. Draw a plot of $f(x)$ when $x \in [-5, 17]$
2. Change the colour and width of the line
3. Use the function `expression( )` to display the formula $f_X(x)$ as label on the y axis

Exercise 5

1. Use the `split` function to split the built-in `iris` data according to the variable `Species` using the code below. This results in a list with a dataframe for each species in the `iris` dataset.

```
splitiris <- split(iris, iris$Species)
```

2. Use a for loop to create a scatter plot between the sepal length and sepal width for each species. Make sure that the axes have informative titles, and that the main title of the plot indicates the species.
3. On each plot, add a black triangle that indicates the mean values of the sepal length and sepal width per species.
4. Remove the `Setosa` dataset from the list `splitiris`.