**Topics of exercises**

* R data types
* Elements of descriptive statistics

**Exercise 1.1**

1. Create a factor **colors** that contains names of 4 arbitrary colors.
2. Create a list **pattern** that contains 500 colors randomly selected from the factor **colors**.
3. Create a list **pattern1** as a result of representation of **pattern** in a form of a data frame and **pattern2** - in the form of a matrix. Compare the result to be achieved.
4. Represent **colors** in a form of the list and denote it as **colors1**.
5. Expand the list **pattern** by the factor **colors** (at first) and by the list **colors1** (secondly). Compare the results (are there any differences between them?) with the earlier ones.

**Exercise 1.2**

Load the **cars** data set

data(cars)

1. Create a copy of the data, say, **cars2**.
2. Modify **cars2**: transform speed units into **m/s** (meters per second), and distance units into **m** (meters).
3. Create a scatterplot for the variables **speed** and **dist**.
4. Use linear regression to find the best line to fit **speed** and **dist** variables.
5. Extend the data-frame with the third column (**dist2**), which contains the speed in **km/h** (kilometers per hour)

**Exercise 1.3**

Suppose that the data for analysis includes the attribute **age**. The **age** values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

1. What is the mean of the data? What is the median?
2. What is the mode of the data? Comment on the data’s modality (i.e., bimodal, trimodal, etc.).
3. What is the midrange of the data?
4. Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
5. Give the five-number summary of the data.
6. Show a boxplot of the data.

**Exercise 1.4**

Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result:

* age: 23, 23, 27, 27, 39, 41, 47, 49, 50, 52, 54, 54, 56, 57, 58, 58, 60, 61
* %fat: 9.5, 26.5, 7.8, 17.8, 31.4, 25.9, 27.4, 27.2, 31.2, 34.6, 42.5, 28.8, 33.4, 30.2, 34.1, 32.9, 41.2, 35.7

1. Define R data-frame with the above data.
2. Calculate the mean, median, and standard deviation of **age** and **%fat**.
3. Draw the boxplots for **age** and **%fat**.
4. Draw a scatter plot based on these two variables.
5. Calculate the correlation coefficient (Pearson's product moment coefficient). Are these two variables positively or negatively correlated?

**Exercise 1.5**

Assume, values that are more than two standard deviations away from the mean for a given attribute are outliers.

1. Find outliers for the **age** attribute (Exercise 1.3).
2. Find outliers for the **speed** and **dist** attributes (**cars2** dataset).
3. Find outliers for the **Sepal.Length**, **Sepal.Width**, **Petal.Length**, **Petal.Width** attributes (**iris** dataset).

**Exercise 1.6**

Let us consider the age of the presidents of the United States at the time of their inauguration: 57, 61, 57, 57, 58, 57, 61, 54, 68, 51, 49, 64, 50, 48, 65, 52, 56, 46, 54, 49, 51, 47, 55, 55, 54, 42, 51, 56, 55, 51, 54, 51, 60, 61, 43, 55, 56, 61, 52, 69, 64, 46, 54, 47.

1. Create the histogram for the **pres.age** attribute.
2. Find outliers (the same assumption as previously). Can you point out the outliers using the histogram?
3. Create the boxplot for the **pres.age** attribute. Compare the boxplot and histogram.

**Exercise 1.7**

Consider the following data regarding the Winter Olimpic Games since 1992 (their years, locations, number of disciplines and the heads of state who opened them):

* 1992, Albertville, 57, F. Mitterand,
* 1994, Lillehammer, 61,King Harald V,
* 1998, Nagano, 68, Emperor Akihito,
* 2002, Salt Lake City, 78, President G.Bush,
* 2006, Torino, 84, President C. Ciampi,
* 2010, Vancouver, 86, Governor General M.Jean.

1. Define R-data frame with the above data,
2. Define the subset of the data with the names of heads of state that opened these Olympic Games,
3. Redefine the subset of the data taking into account the Olympic Games with more than 78 disciplines only,
4. Create the subset of the data taking into account the Olympic Games with than mean value of the number of disciplines.

**Exercise 1.8**

Taking into account the number of disciplines (Exercise 1.7), create a list of their normalized values obtained as follows:

1. N1(x) = (x - mean(x)) / 100, where x is an initial value (a number of disciplines),
2. N2(x) = (x - mean(x)) \* (max(x) - min(x))/1000.
3. N3(x) is a difference between the initial values and their standard deviation, divided by their midrange. (Is it a real normalization?)

**Exercise 1.9**

Consider the numbers of passengers of „Queen Elisabeth” in the following years:

* 1994-34567,
* 1995-34678,
* 1996-36789,
* 1997-38102,
* 1998-39024.

1. Create the (frequency) histogram of the number of passengers with 5 green breaks and blue border with all values oriented horizontally.
2. Create the (probability) histogram of the number of passengers with red breaks and green border and density function with all values oriented vertically.
3. Formulate the condition of a drawing of the density line.

**Exercise 1.10**

Load the **iris** data set, consider the values (140-150) of **Sepal.Length** and treat it as the vector. Plot:

1. the histogram with the density line and 8 breaks,
2. the pie chart and boxplot for iris data set.