Course of Study Bachelor Computer Science	Exercises Statistics WS 2021/22
Sheet I	

1 Descriptive Statistics - Variables

1. Categorize the following variables as being qualitative or quantitative:

   (a) Rating of the quality of a movie on a 7-point scale
   (b) Age
   (c) Country you were born in
   (d) Favorite Color
   (e) Time to respond to a question
   (f) Intelliquence quotient
   (g) Immatrikulation number

   Specify the level of measurement used for the items.

2. It is possible to transform a variable "downwards", from a scale with more information contained, to a scale with less information contained. Give an example for the variable *Price for a bottle of wine* for the transformation from a ratio to an ordinal scale.

3. Is it possible to transform a variable "upwards", from a scale with less information contained, to a scale with more information contained? Give an example (showing if it is possible or not)!

4. Consider the question of describing students attitudes towards to legalisation of Marihuana, what proportion of them wants to legalize the drug and whether this proportion differs by gender and age.

   (a) Which data collection method is most suitable here: survey or experiment?
   (b) How could you capture the attitudes towards legalisation in a single variable?
   (c) Which variables are needed to answer the questions? Describe the type and the scale of the variables.
   (d) How would an appropriate data set look? Try to describe the question in more details.

Introduction to R and RStudio

Some useful hints for the first steps

- Open a new script file via File→New→R Script.
- Save the script file at any time via File→Save.
- Comments to the code that are not evaluated can be made with the # icon.
- Send R code to the R console:
  - Click with the mouse on a line (any location). Then click on the Run-button. Only the selected line will be sent to the R console. The cursor will automatically move to the next line. You can now click again to send this line to the R console, and so on.
  - Mark the code you want to send to the R console. Then click the Run-button. So all the marked text will be sent to the R console.
  - $< Ctrl > + < Enter >$ on your keyboard instead of prerssing the Run-button

Useful shortcuts

- Assignment arrow $< -$: $< Alt > +-$
- Complete code:Tab key
- Comment in/out marked region:$< Ctrl > + < Shift > +C$
- Delete R console: $< Ctrl > +L$
- Cancel evaluation (if it takes too long): click or press $< Esc >$ in the R console.
- In the R console, retrieve previous code: Arrow keys (up and down).
- Switch to the editor with the cursor: $< Ctrl > +1$
- Move the cursor to the R console: $< Ctrl > +2$
- Save: $< Ctrl > +S$

Start now RStudio, open a new script file and solve the following tasks.

1. Calculate the following quantities:

   - the sum of 52.3, 74.8, 3.17
   - the square root of 144
   - the 10-based logarithm of 200 multiplied with sin of $\pi/4$
   - the cumulative sum of the numbers 1,3,18,20,2 (use the cumsum() command)
   - find 10 numbers between 0 and 20 rounded to the nearest integer value (hint use the command sample() or a combination of the commands round() and runif()).

   Hint: If you do not know command() use the ?command().

2. Assigning Variables

   - Assign the number 5 to x and the number 10 to y.
   - Calculate the product of x and y.
   - Store the result in a new variable z.
   - Inspect your workspace by clicking the "environment" tab in RStudio, and find the three objects.
   - Make a vector myvec of the objects x,y,z.
   - Find the minimum, the maximum and the mean of the vector.
   - Remove myvec from the workspace.

3. The numbers below are the first ten days of rainfall in a year

   ```
   0.1  0.5  2.3  1.1  11.3  14.7 23.4  15.7  0  0.9
   ```

   - Read them into a vector using the c() command.
   - Calulate the mean and the standard deviation.
   - Calculate the cumulative rainfall over these ten days. What is total sum of the rainfall?
   - Which day saw the highest rainfall? Find an appropriate R command.
   - Take a subset of the rainfall data where rain is larger than 10.
   - What is mean rainfall for days where the rainfall was at least 5?
   - Subset the vector where it is either exactly 0 or 1.1 and find the corresponding days.

4. The length of five cylinders are 2.5, 3.4, 4.8, 3.1, 1.7 and their diameters are 0.7, 0.4, 0.5, 0.5, 0.9.

   - Read these vectors into two vectors with appropriate names.
   - Calculate the volumes of each cylinder and store it in a new vector.
   - Assume the values are given in centimeter. Recalculate the volumes so that their units are cubic millimeter.

5. Inspect the R commands union(), setdiff() and intersect() implying set operations. Make two vectors

```
x <- c(1,2,3,4,5)
y <- c(3,5,7,9)
```

   - Find values that are contained in both x and y.
   - Find values that are in x but not y and vice versa.
   - Construct a vector that contains all values contained in either x or y. Compare the result with c(x,y).

6. Construct a matrix with 8 rows and 10 columns. The first row should contain the numbers 0, 2, 4, ..., 18 and the other rows should random integer numbers between 0 and 100. Use runif() to create the random numbers and as.integer() to to transform to integers.

   - Calculate the row means of this matrix (use rowMeans()) and the standard deviation across the row means.
   - Store the rows 2,3,..,8 in a other matrix and calculate the column means (use colMeans()). Use the command hist() to create a histogram of the column means.

7. The R dataset mpg

   (a) Inspect the dataset mpg.

   (b) Determine the types and the scales of measurement of all variables in the dataset mpg. Further more determine whether the variables are discret or continous.

   (c) Create an empty tibble str_mpg with variables name, type, level and dc of type character(). Add for every variable in the dataset mpg a row in str_mpg containing for every variable the name, the type, the level of measurement and discrete/continous.

(d) Display the structure of the tibble str_mpg.

(e) Use the tibble to display all variables which are quantitative and discrete applying the R function subset().

Hint: The dataset mpg is part of the package ggplot2 and tibbles are part of the tidyverse package.