Course of Study Bachelor Computer Science	Exercises Statistics WS 2020/21

Sheet II

1. Tidy Data
   Consider the following datasets

```
student1  <- tibble(
  student = c("Adam","Bernd","Christian","Doris"),
  algebra = c(NA, 5, 3, 4),
  analysis = c(2, NA, 1,3),
  diskrete.math = c(3,NA,2,4),
)

student2 <- tibble(
   name = rep(c("Adam", "Bernd", "Christian", "Doris"), each = 2),
   type = rep(c("height", "weight"), 4),
   measure = c(1.83, 81, 1.75, 71, 1.69, 55, 1.57, 62))

student3 <- tibble(
    name = c("Adam", "Bernd", "Christian", "Doris"),
    ratio = c("81/1.83", "71/1.75", "55/1.69", "62/1.57"))
```

   (a) Describe for every dataset describe what the dataset contains? What are the variables and what are the observations?

   (b) Why are these datasets are not tidy?

   (c) Make a tidy version of all datasets.

2. Using the $\% > \%$-operator.

   - Calculate the value of $\sin(\log(\sqrt{5+3}))$ directly and using the $\% > \%$-operator.
   - Define a vector v with values 0.5,1,1.5,...,5 and calculate the by 2 digits rounded sum of the logarithms of the squared values of v with nested operations and using the $\% > \%$-operator.

3. Create a tibble df with the data of 10 students, i.e. with 10 rows and the columns id (values 1,2,..., 10), sex (values are “f” and “m”, age (integer values between 20 and 35) and score1 (integer values between 0 and 25). You can choose arbitrary values in the columns. If you do not like coding the values by hand you can use:

```
df <- tibble(id = 1:10,
                  sex = sample(x =c("f","m"), size = 10,
                               replace = TRUE),
                  age = round(runif(10,20,35)),
                  score1 = round(runif(10,0,25))
                )
```

- Select the date of all male students.
- Add the data of a new student with id = 11, sex = “m”, age = 25 and score1 = 4.
- Add two columns score2 and score3 with random integer numbers between 0 and 25.
- Add a column containing sum of all scores.
- Add a column which denote the grades according to the scheme
$$\text{grad} = \begin{cases} 5 & \text{if} \quad \text{score sum} \leq 37 \\ 4 & \text{if} \quad 37 < \text{score sum} \leq 45 \\ 3 & \text{if} \quad 45 < \text{score sum} \leq 55 \\ 2 & \text{if} \quad 55 < \text{score sum} \leq 65 \\ 1 & \text{if} \quad \text{score sum} \geq 65 \end{cases}.$$
- Find the values of the variables id, sex and grade sorted by the values of sex of all students who have passed.
- Calculate the mean, minimum, maximum and median of the variable sum of scores grouped by the variable sex.

4. Some data manipulations with the data set flights.

- Load the libraries tidyverse and nycflight13 and inspect the variable of flights.
- Find all flights with more than 2 hours arrival delay.
- Find all flights with more tahn 2 hours arrival delay and no departure delay.
- Find all flights from United, American and Delta with no arrival delay.

- Find all flights from United, American and Delta in the month May with more than 5 hours arrival delay sorted by carrier and flight number.
- Exchange the values of departure time and arrvial time in minute after midnight.
- Add a column speed which denotes the average speed of the flight and determine the carrier, flight of the top 10 values of speed.
- Find a list of carriers with a column ratio which denotes the number of flights with arr_delay less than 10 minutes to the total number of flights. The list should be sorted by ratio.
- Find a list which denotes for every month the carrier with highest ratio. The list sould have the columns month, carrier, number of flights of the carrier in that month and ratio.
- Find a table with the number of cancelled flights (dep_delay = NA), the number of flights with no dep_delay (—dep_delay— ≤ ± 5 minutes and the means of dep_delay, arr_delay per month and day.