Course Bachelor Computer Science	Exercises Statistics WS 2024/25
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 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. The R statements

```
no <- 30
exercise.results <- tibble(
  stud.id = 1:no,
  group = sample(x=c("A","B","C"), size=no, replace = TRUE),
  ex1 = sample(x=1:10, size=no, replace = TRUE),
  ex2= sample(x=1:10, size=no, replace = TRUE),
  ex3 = sample(x=1:10, size=no, replace = TRUE),
  ex4 = sample(x=1:10, size=no, replace = TRUE),
  ex5 = sample(x=1:10, size=no, replace = TRUE)
)
```

creates a tibble containing the scores of 30 students in 5 exercises.

(a) Apply n() and count() to get the number of students in the different groups. What are the difference between n() and count()?

(b) Add the variables sum.scores and mean.scores containing the sum and the of the scores in the exercises for every student by applying the the functions sum() and mean(). What is the result if rowwise() is appplied before the mutate()?

5. 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.

  Example: departure time 722 given by HMinutes is in minutes after midnight 442!
- Add a column speed which denotes the average speed of the flight and determine the carrier, flight of the top 10 values of speed.

  Hint: The first 5 lines of the output are

```
  carrier flight speed
  <chr>    <int> <dbl>
1 DL        1499  703.
2 EV        4667  650.
3 EV        4292  648
4 EV        3805  641.
5 DL        1902  591.
```

- 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.

  Hint: The first 3 lines of the output are

```
  carrier   nof  ndel del_ratio
  <chr>   <int> <dbl>     <dbl>
1 HA        342   277     0.810
2 AS        709   574     0.810
3 VX       5116  3942     0.771
```

  with nof = number of flights, ndel = number of flight with arrival delay less than 10 minutes, del_ratio = ratio of nof and ndel.

The following evaluations are more challenging and need more time!

- Find a list which denotes for every month the carrier with highest ratio which denotes the number of flights with arr_delay less than 10 minutes to the total number of flights. The list should have the columns month, carrier, number of flights of the carrier in that month and ratio.
  Hint: The first 3 lines of the output are

  ```
    month carrier   nof  ndel del_ratio
    <int> <chr>   <int> <dbl>     <dbl>
  1     1 VX        314   290     0.924
  2     2 HA         28    25     0.893
  3     3 VX        303   257     0.848
  ```

- 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.

  Hint: Join the evaluations of

  – no of cancelled flights per month and day

  – no of flights with no delay per month and day

  – averages of departure and arrival delays per month and day

  to one table. The first 3 lines out the output are

  ```
    month   day nof_canc nof_no_delay mean_dep_del mean_arr_del
    <int> <int>    <int>        <int>        <dbl>        <dbl>
  1     1     1        4          471         11.5         12.7
  2     1     2        8          488         13.9         12.7
  3     1     3       10          496         11.0          5.73
  ```

- Determine a table that shows, for each airline (carrier), the flight connection given by the airports of dest und origin that occurred most frequently in 2013. The table should contain only the columns names of airline, destination, origin and frequency and be sorted by frequency in descending order. You can find the names of the carrier from the dataset airlines and the names of the airports from the dataset airports.
  Hint: Evaluate for every carrier the frequency of most frequently flight connection and join the names of the airports of origin and destination from the table airports. The first 3 lines of the output are

  ```
       n Airline                     Airport.Origin      Airport.Destination
    <int> <chr>                        <chr>               <chr>
  1  5694 American Airlines Inc.       La Guardia          Chicago Ohare Intl
  2  5544 Delta Air Lines Inc.         La Guardia          Hartsfield Jackson Atlanta Intl
  3  4716 US Airways Inc.              La Guardia          Ronald Reagan Washington Natl
  ```

6. Applications of pivot_longer() and pivot_wider(). The use of the pivot_longer() and pivot_wider() functions is described https://cran.r-project.org/web/packages/tidyr/vignettes/pivot.html. Additionally you find there hints how to solve the following problems.

(a) Consider the dataset who of the package tidyr.
country, iso2, iso3, and year are already variables, so they can be left as is. But the columns from new_sp_m014 to newrel_f65 encode four variables in their names:

- The new_/new prefix indicates these are counts of new cases. This dataset only contains new cases, so we'll ignore it here because it's constant.
- sp/rel/ep describe how the case was diagnosed.
- m/f gives the gender.
- 014/1524/2535/3544/4554/65 supplies the age range.

Break these variables up by specifying multiple column names and make the dataset tidy.

(b) The anscombe dataset contains four pairs of variables (x1 and y1, x2 and y2, etc.) that underlie Anscombe's quartet, a collection of four datasets that have the same summary statistics (mean, sd, correlation etc), but have quite different data.
Produce a dataset with columns set, x and y.

(c) Apply the following R statements

```
production <-
  expand_grid(
    product = c("A", "B"),
    country = c("AI", "EI"),
    year = 2000:2014
  ) %>%
  filter((product == "A" & country == "AI") | product == "B") %>%
  mutate(production = rnorm(nrow(.)))
```

The data set production contains the combination of product, country, and year.
Widen the data so we have one column for each combination of product and country.

(d) The data set warpbreaks gives the number of warp breaks per loom, where a loom corresponds to a fixed length of yarn for every combination of wool (A and B) and tension (L, M, H).
Produce a data set with the columns tension, A, B with the means of the breaks.