# Using R for data analysis (SSA)

Boerhaave Nascholing LUMC

January 23rd, 2024

## Introduction

You will analyse the `storms` table which comes with the `tidyverse` package.
Make sure you put `library( tidyverse )` in the R chunk at the top of your R Markdown file as shown here below:

```
library( tidyverse )
```

After the library has been loaded you will have access to the table in the `storms` variable. Each row of `storms` table is an observation of a storm recorded at a certain moment (date and time) at a geographical location (`lat`, `long`). Some additional storm features (`wind` speed, `pressure`, ...), classifications (`status`, `category`) and a `name` are also included.

For more details you may consult the help on storms tibble with `?storms` but the following column description is sufficient for the SSA:

- `name`: Name of the storm.
- `year`, `month`, `day`, `hour`: Date and time of the observation.
- `lat`, `long`: Geographical location of the storm centre (numbers).
- `wind`: Wind speed (number, in knots).
- `pressure`: Pressure at the storm's centre (number, in millibars).
- `tropicalstorm_force_diameter` (or `ts_diameter` in older versions of `tidyverse` library): Storm diameter (number, in nautical miles).
- `status`: Storm classification (a factor, many levels).
- `category`: Storm category (a number, range: -1..5; many values are missing).

Note, that a single storm is usually observed multiple times (so one storm may be described in multiple rows).

Here is a random part of the table (some columns are omitted):

```
# A tibble: 6 x 9
  name   year month   lat  long status         category  wind pressure
  <chr> <dbl> <dbl> <dbl> <dbl> <fct>             <dbl> <int>    <int>
1 Emily  1981     9  42.7 -41   extratropical        NA    30     1008
2 Klaus  1984    11  22.7 -58.7 hurricane             1    80      980
3 Hugo   1989     9  16.6 -62.5 hurricane             4   125      949
4 Ana    1991     7  37.9 -61.1 tropical storm       NA    45     1000
5 Ivan   2004     9  11.6 -59.4 hurricane             3   100      963
6 Kirk   2018     9  11   -46.8 tropical wave        NA    35     1007
```

## Questions

### Question 1: [4p] Percentage of storms with category at least 4.

Out of all storm measurements with non-missing `category` value, calculate the *percentage* of the storm observations that have `category` at least `4`. Find how to use `round` to round the result to 2 decimal places.

Assign the result to the `largeCategoryPercentage` variable.

```
# largeCategoryPercentage <- ...
```

## Question 2: [3p] Changing factor levels, counting occurrences.

Take the data from the `status` column and change the order of levels such that the first three levels are `("tropical storm", "tropical depression", "hurricane")` (in exactly this order).
Then, produce a table of counts of the number of observations for each storm `status` level.
Store the result in `statusCounts` variable.
Note: Do not modify the original `storms` table (a changed table may not work in other questions).

```
# statusCounts <- ...
```

## Question 3: [7p] Table summary in a list.

Create a list with some summaries of the `storms` table and assign this list to the variable `stormsSummary`. The list should have the following three elements:

- `obsNum` – the *number* of observations in the `storms` table,
- `avgWind` – the mean of observed `wind` speeds (force removal of missing values),
- `uniqueNames` – a *character vector* of names from the `name` column with duplicates removed, sorted in alphabetical order.

```
# stormsSummary <- ...
```

## Question 4: [6p] Dropping summer storms

Create a new tibble `stormsNoSummer` that contains all observations from `storms` except those that were made in a summer. Consider 21st of June to be the first day of summer and 22nd of September to be the last day of summer.

```
# stormsNoSummer <- ...
```

## Question 5: [6p] Summarizing storms by month.

Build a *tibble* reporting the fastest wind and the lowest pressure observed over all years in each `month`. Report also the total number of observations for each `month`. During the min/max calculations force omitting possible missing values in the respective columns.
The final table should have four columns: `month`, `fastestWind`, `lowestPressure`, `obsNum` and it should be sorted in descending order of the number of observations (the most frequent at the top row). Store the result in the variable `stormsByMonth`.

```
# stormsByMonth <- ...
```

## Question 6. [4p] Cross-tabulation

Create a *tibble* `stormsByStatusAndMonth` that contains a cross-tabulation of `status` and `month`. The result should be a table with `status` represented by rows, `month` in columns, and table values representing the number of observations for each combination of `month` and `status` values. Some entries in the crosstable will be `NA`: check the manual and fill them with zeros.

```
# stormsByStatusAndMonth <- ...
```

## Question 7. [9p] Adding wind speed in km/h and its category.

Wind speed in the `wind` column is given in knots. Create a new column `windKPH` that expresses wind speed in km/h (1 knot = 1.852 km/h). Then, create a new column `windCategory` that contains a factor with levels

`"low"`, `"medium"`, `"high"` (exactly in that order). The levels should be determined by the `windKPH` column values: `"low"` for `windKPH` $< 75$, `"medium"` for `windKPH` $< 150$, and `"high"` otherwise. The final table should only have columns: `name`, `windCategory` and `windKPH` (exactly in this order). Store the result in the variable `stormsWithWindCategory`.

```
# stormsWithWindCategory <- ...
```

## Question 8: [7p] A box plot.

Based on the `storms` tibble create a box plot:

- The vertical axis should represent `pressure`.
- The horizontal axis: in `aes(...)` instead of `wind` use `factor(wind)` (to make `wind` a categorical variable).
- Use `gray` box fill and `blue` colour.
- Adjust the vertical title to `"Pressure [millibars]"` and horizontal to `"Wind speed [knots]"`.
- Use the black/white theme.

```
# ggplot( ... ) + ...
```

## Question 9: [8p] Scatter plot

For this scatter plot take from `storms` only the rows with a missing `tropicalstorm_force_diameter` (or `ts_diameter`) value. Use `long` for the horizontal axis and `lat` for the vertical. Use transparency level of 0.5 and point size of 0.75. Colour points according to `wind`. Finally, use the colour scale with `green` for low and `red` for high `wind` values.

```
# ggplot( ... ) + ...
```