Week 3 Tutorial Worksheet

AY23/24 Semester 2

No submission required this week

We will use `R` Markdown (`.Rmd`) for all tutorials and exams throughout the semester. Please make sure your code and texts are in a single `.Rmd` file.

This tutorial describes how to use `R` Markdown. If you have any questions or face any issues, please do not hesitate to approach your tutor for help.

A simple work flow

To use `R` Markdown, you will need to two packages: `rmarkdown` and `knitr`. You can install any necessary packages with `install.packages()` in your **Console**. Note that you only need to install each package once.

1. Download `tutorial_weekX.rmd` from Canvas and store it in your `src` folder. Rename the file as `tutorial_week3.rmd`.
2. In the YAML header, update the title and your name.
3. Click the **Knit** button to render your `R` Markdown file to HTML.

   The HTML now shows the specified title and author information. It will **echo** the commands in the `R` code block, **eval**uate the `R` code, and **include** both the code and the output in the document.
4. Answer each tutorial question in the designated section. At the end of the tutorial, knit the file to HTML again.

Note: The `tutorial_weekX.rmd` template is a partially-filled out `R` Markdown file that you can work with. It aims to help you get started with `R` Markdown. As you gain better familiarity with the syntax and the language, you can generate your own template for future tutorials, exams, and projects.

Question 1 Data manipulation and plotting in Base R

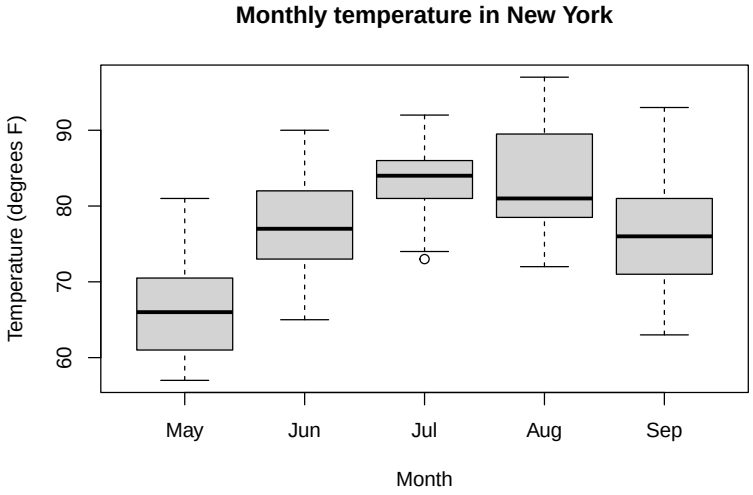
In this question, we will look at daily air quality in New York City in 1973. The data are available in base R and can be loaded via `data(airquality)`. You can find out more about the data set by inspecting its documentation with `?airquality`.

```
data(airquality)
```

1. Examine the structure of this data set. How many months are represented in the data?
2. Are there any missing entries in the `airquality` data frame? Eliminate row(s) that contain missing values and save it in a new data frame named `df`.
3. For the remaining questions, we will work on the data frame `df`. Create two new data frames, one for summer months (May and June) and one for fall months (July, August, and September). Name them as `df_summer` and `df_fall`, respectively.
4. Create a new categorical column `season` in the data frame `df` that takes value of `Summer` for May and June, and `Fall` for July, August, and September. Use the `tapply()` function to compute the summary statistics of temperature in summer and fall. *Hint: Your output should look like the following:*

```
## $Fall
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   63.00   76.25   81.00   81.23   86.00   97.00
##
## $Summer
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   57.00   62.00   68.00   69.67   76.00   90.00
```

5. Recode the `Month` variable from 5 - 9 to May through September. Re-create, as much as you can, the boxplot below comparing temperature across all months.



Requirement

- After answering all questions in the `Rmd` file, hit the "Knit" button again. Make sure your `Rmd` can knit to HTML without error.
- The code in your `Rmd` should create the following objects:
  - `df, df_summer, df_fall`
- The knitted HTML should contain
  - A boxplot for Question 1.5

Reach out to your tutor after the tutorial if you are unsure about any of the requirements above.