STAT40730 Data Programming with R
Assignment 2

Isabella Gollini

Instructions

- This assignment is due on **Wednesday 15th November 2023** at 11:59pm.
- You should submit it to the "Assignment 2" assignment object in Brightspace.
- You should submit 2 or 3 files only depending on your final format.
  - If you **render to pdf** you have to submit 2 files:
    1. **Qmd** file detailing the commented code you used to obtain your answers.
    2. Rendered document in **pdf** showing all your code and your answers.
  - If you **render to HTML** you have to submit 3 files:
    1. **Qmd** file detailing the commented code you used to obtain your answers.
    2. **A zip file** containing the HTML file showing all your code and your answers. (Notice that the zip file must contain only the HTML file).
    3. a **pdf** file obtained by converting the HTML to pdf. *You can use Google Chrome. File > Print > Destination [Change. . .] > select Save as PDF.*
- Remember that if you decide to produce an HTML output you must have the following on your YAML header:

```
format:
  html:
    embed-resources: true
```

- You may submit it multiple times before the deadline, but only the last version will be marked.
- There is a maximum of 19 marks for this assignment. This assignment is worth 19% of your final grade. The marks available for each question are shown in brackets.
- Late submissions will score 0, unless a "Late Submission of Coursework" form is submitted.
- Assignment 2 consists of 3 tasks: data manipulation, analysis, and creativity.
- This assignment covers up to the material in Topic 6.
- You may have to discover and learn some new functions. Use `help()` and `help.search()` to find what you need.

- A couple of suggestions: create an RStudio Project for this assignment and save the .qmd file and the data set in the same folder. Render your document frequently to fix errors.

Plagiarism

While you are encouraged to ask about the module material, this assignment should be completed individually. Any student who plagiarises will receive a 0 mark. If you are unsure whether a question about the project would be considered as plagiarism, please email the question to the lecturer rather than posting on the discussion forums. The UCD Plagiarism Policy applies to all students. This can be consulted at the following link.

Data

The dataset `dublin-bikes.txt` contains variable concerning bike traffic and weather conditions in Dublin from September 1st 2022 at 12am to August 31st 2023 at 11pm.

There are seven variables concerning bicycle traffic volumes from cycle counters in various locations Dublin city. Passing cyclists are counted and logged every hour, 24 hours per day, 7 days per week. The Bicycle traffic dataset was downloaded from: data.smartdublin.ie

The other variables concern weather condition, and they have been downloaded from Met Éireann.

In detail the dataset `dublin-bikes.txt` consists of the following variables:

- `Time` timestamp for the data collected
- `Clontarf - James Larkin Rd` hourly bicycle traffic volumes from cycle counters in Clontarf (James Larkin Rd)
- `Clontarf - Pebble Beach Carpark` hourly bicycle traffic volumes from cycle counters in Clontarf (Pebble Beach Carpark)
- `Griffith Avenue (Clare Rd Side)` hourly bicycle traffic volumes from cycle counters in Griffith Avenue (Clare Rd Side)
- `Griffith Avenue (Lane Side)` hourly bicycle traffic volumes from cycle counters in Griffith Avenue (Lane Side)
- `Grove Road Totem` hourly bicycle traffic volumes from cycle counters in Grove road
- `Richmond Street Cyclists 1` hourly bicycle traffic volumes from cycle counters in Richmond street (location 1)
- `Richmond Street Cyclists 2` hourly bicycle traffic volumes from cycle counters in Richmond street (location 2)
- `rain` precipitation Amount (mm)
- `temp` air Temperature (°C)
- `wdsp` mean hourly wind speed (kt)
- `clamt` cloud amount (okta):
  - 0 oktas represents the complete absence of cloud
  - 1 okta represents a cloud amount of 1 eighth or less, but not zero
  - 7 oktas represents a cloud amount of 7 eighths or more, but not full cloud cover
  - 8 oktas represents full cloud cover with no breaks
  - 9 oktas represents sky obscured by fog or other meteorological phenomena

Assignment 2

- **Write a scientific report** by completing the three tasks below. [2.5]
  - Complete your assignment using Quarto, check that all the code and output are correctly shown in your final document.
  - Clearly indicate in each code chunk which question it is referring to.
  - In tasks 1 and 2 you must use base R and the packages that we have used in class up to topic 6 only. You are free to use functions from other packages for task 3 if you wish.
  - Do not print the full dataset, it makes the document very hard to read.
  - Save the data file in the same folder as the .Qmd file, so that you don't have to specify the file path that is specific of the computer you are using (and we would not be able to run your code without changing it).

Task 1: Manipulation

1. Load the dataset `dublin-bikes.txt`, save it as a tibble and give meaningful names to the variables related to the weather. [0.5]

2. What is the size (number of rows and columns) this dataset? Write some code to check that the variable `Time` is stored using an appropriate class for a date, and the other variables are numeric, fix them if they aren't. [1]

3. Convert the variable containing the cloud amount information into an ordered factor. Print the levels and the output of a check to confirm it's ordered. [1]

4. Split the information in the column `Time` into two columns: one containing the date (i.e. date only, no time), and the other the hour. Check that there are 24 hours for each date, and that there are 365 different dates. [1]

5. Add two columns one containing the day of the week and the other the month. Check that these two columns are ordered factors. [1]

6. Remove the column `Time` and use `dplyr::relocate()` to put the new columns with the date, hour, day of the week, and month as the first four columns of the dataset. [0.5]

Task 2: Analysis

1. Use functions from *base R* to compute which month had in total the highest and the lowest Precipitation Amount. [1.5]

2. Use `ggplot2` to create a time series plot of the maximum and minimum daily temperatures. [The two time series must be on the same plot.] [1.5]

3. Check if, according to this dataset, there has been on average more rain during the weekend (Sat-Sun) with respect to weekdays (Mon-Fri). [2]

4. Focus on the data for one month of the year of your choice, create a plot of the daily traffic volume in a locations of your choice, and the mode of the Cloud amount each day. Comment on your findings. [Notice that there isn't a built-in function to calculate the mode in R. The mode is defined as the most frequently occurring value in the set of observations.]. [2.5]

Task 3: Creativity

Do something interesting with these data! Create **two plots** or **two tables** or **one plot and one table** showing something we have not discovered above already and outline your findings. [4]

END OF ASSIGNMENT 2

Few tips for troubleshooting

- Be aware that a common error is to give the same label to two different code chunks!

````
```{r}
#| label: cars
summary(cars)
```

```{r}
#| label: cars
plot(cars)
```
````

You can fix this by changing the label to one of them:

````
```{r}
#| label: fig-cars
plot(cars)
```
````

- In case of a code error that you can't fix in time for your submission.

Add the option `error: TRUE` into the R chunk to run the code, show the error message on the rendered file. For example:

````
```{r}
#| error: true
x <- "a"
sum(a)
```
````

Or you can add the option in your YAML header to work on the full document:

```
execute:
  error: true
```