STAT30340 Data Programming with R
Assignment 2

Isabella Gollini

Instructions

- This assignment is due on **Friday 15th November 2024** at 11:59pm.
- You should submit it to the "Assignment 2" assignment object in Brightspace.
- You should submit 2 files only
    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.
- 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 datasets `pedestrian_2023.csv` and `weather_2023.txt` contain hourly data concerning pedestrian traffic and weather conditions in Dublin in 2023 (from January 1st 2023 at 12am to December 31st 2023 at 11pm).

- `pedestrian_2023.csv` consists of a column `Time` containing a timestamp for the data collected about pedestrian footfall counts of people at a number of locations in Dublin city centre (information contained in the following columns). See: data.smartdublin.ie.
- `weather_2023.txt` contains variables concerning weather condition, and they have been downloaded from Met Éireann. In detail the dataset consists of the following variables:
  - `Time` timestamp for the data collected
  - `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** (i.e. write some sentences in markdown answering the questions and explaining what you are doing/what you have done) by completing the 3 tasks below.[2.5]
  - Complete your assignment using Quarto, check that all the code and output are correctly shown in your final document.
  - Do not print the full dataset, it makes the document very hard to read.
  - Clearly indicate in each code chunk which question it is referring to.
  - You must use base R and the packages that we have used in class up to topic 6 only.
  - 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. We need to be able to render your file without making any change to it (we'll have the data saved in the same directory as your qmd file).

Task 1: Manipulation

1. Load the dataset `pedestrian_2023.csv`, save it as a tibble. Use a function from `dplyr` to remove the columns with a column name ending with "IN" or "OUT". What is the size (number of rows and columns) of this dataset? [1.5]
2. 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. [0.3]
3. Load the dataset `weather_2023.txt`, save it as a tibble. Give meaningful names to the variables related to the weather. What is the size (number of rows and columns) of this dataset? [0.7]
4. 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. [0.8]
5. Use the function `skim_without_charts()` from the package `skimr` on this weather dataset, and briefly explain in your own words what the function is doing. [1]
6. Check that the variable `Time` in the weather dataset is of an appropriate class for a date (fix it if it isn't), and confirm that the range of `Time` in the two dataset is the same. [0.3]
7. Join the two dataset. What is the size (number of rows and columns) this dataset? [0.7]
8. Add two columns one containing the name of the day of the week and the other the month. Check that these two columns are ordered factors. [1]
9. Use `dplyr::relocate()` to put the new columns with the month and day of the week as the second and third columns of the dataset. Print the column names. [0.7]

Task 2: Analysis

1. Use functions from *base R* to compute which months had the highest and the lowest overall pedestrian traffic (i.e. total pedestrian traffic in all location in the whole month). [1.5]
2. Use `ggplot2` to create a plot displaying three time series of **daily** pedestrian footfall in three locations of your choice. Add two vertical bars to mark St. Patrick's day and Christmas Day [The three time series must be in the same plot. You can use any package covered up to Topic 6 to prepare the data] [2.5]
3. Create a table displaying the minimum and maximum temperature, the mean daily precipitation amount, and the mean daily wind speed by season (Winter: December to February, Spring: March to May, Summer: June to August, and Autumn: September to November). [2.5]

Task 3: Creativity

Do something interesting with these data! Create **one plot and one table** showing something we have not discovered above already and outline your findings (the plot and the table must display different findings). [3]

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
```