MA22019 2025 - Problem Sheet 2

Working with dplyr and ggplot2

Overview

Exercises 1-3 help you with revising the techniques covered in the lectures in Week 2 (and Sections 1.3-2.2 in the lecture notes). You can check your solutions to these questions yourself by entering them on Moodle.

The tutorial questions ask you to apply functions from the dplyr, lubridate and ggplot2 packages. You should prioritize Tutorial Question 1, and only after completing it consider Tutorial Questions 2 and 3. Please make sure to use R Markdown for all your analyses.

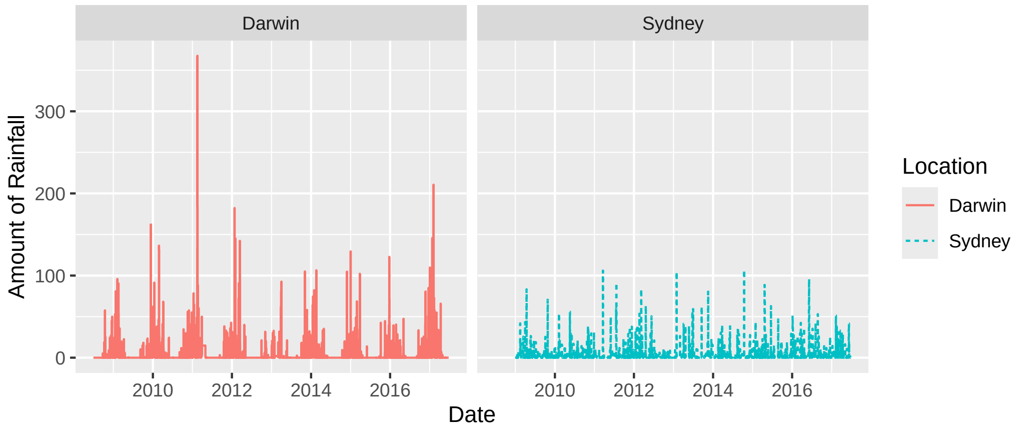
Your answer to the Homework Question can be submitted on Moodle to your tutor for feedback. The submission deadline is 17:00 on Thursday 20 February 2025. You should submit a single Word, PDF or HTML file that provides your R code, any created R output and all your comments.

Before starting the questions, please make sure to load the relevant R packages:

```r
library(dplyr)
library(lubridate)
library(ggplot2)
library(patchwork)
```

Exercise 1 - Identifying visual cues

An analysis of the Australian weather data, considered in the lecture, produced the following plot:



Identify the visual cues used in the plot and submit your answers via the quiz on Moodle.

Exercise 2 - Temperatures for Rio de Janeiro and Sao Paulo

The files “Rio Temperature.csv” and “Sao Paulo Temperature.csv” contain monthly average temperature measurements for Rio de Janeiro and Sao Paulo in Brazil. Missing values in the data matrix are stored as `NA`.

a) Join the temperature data for Sao Paulo and Rio de Janeiro into a single data frame using full_join(). Each row in this new data frame should provide the year, month and the monthly average temperatures for Rio de Janeiro and Sao Paulo. Rename any variables that have uninformative names.

b) Using the data frame produced in part a), answer the following two questions and submit your answers via the quiz on Moodle:

(i) What was the monthly average temperature for Rio de Janeiro when Sao Paulo recorded its highest monthly average temperature?

(ii) What was the largest difference in the monthly average temperatures for Sao Paulo and Rio de Janeiro?

Exercise 3 - Weather in New York City in 2019

The file “NYC Weather 2019.csv” provides daily weather data for 2019 for New York City. We are given the following variables:

- **Date:** Day of the recording
- **Tmin, Tmax:** Minimum and maximum temperature in degree Fahrenheit
- **Precipitation:** Amount of precipitation in inches

Consider the following tasks. You can submit your answers to parts c) and e) via the quiz on Moodle.

a) Create a density plot to visualize the distribution of the maximum daily temperature.

b) Similar to Section 2.2.3. in the lecture notes, create a line plot of day of year against amount of precipitation on that day.

c) Identify the date for which the difference in recorded daily minimum and maximum temperatures was at its highest in 2019. Enter your answer in the Day/Month format via the quiz on Moodle.

d) Create a plot which per month provides a box plot for the maximum daily temperature. **Hint:** Use the function month() to extract the month from a date and then apply the function factor() to convert it to a categorical variable.

e) Answer the following two questions and submit your questions via the quiz on Moodle:

   (i) Which month had the highest median daily maximum temperature according to your plot in part d)?

   (ii) We learned that box plots should usually be ordered based on a criteria, such as the median. Should this approach also be taken for the box plot you created in part d)?

Tutorial Question 1 - Utopia's Ambulance Service

The health department of *Utopia* has collected data on the response time of their ambulance service, including whether patients were admitted to the country's hospitals and how long they stayed. All the data are stored across the files "Ambulance.csv" and "Hospital.csv". The variables in "Ambulance.csv" are

- **Call** - Day and time at which the ambulance service was notified
- **Category** - Ambulance response category (1="Life threatening", 4="non-urgent")
- **Arrival** - Time of arrival of the ambulance at the patient's location
- **PatientID** - Health insurance number of the patient
- **Hospital** - Time patient arrived at the hospital (missing values indicate that no hospitalization was required)

The file "Hospitals.csv" provides the following information:

- **PatientID** - Health insurance number of the patient
- **Age** - Age of the patient
- **Length** - Number of days the patient stayed in hospital

We are asked by the Utopian government to perform an analysis which only considers the patients who were brought to hospital by the ambulance service. For these patients, they ask us to address the following two questions:

a) When considering patients within the different response categories, are there any differences in terms of time spent in hospital?

b) What is the relation between a patient's waiting time for the ambulance and their length of stay in hospital? **Hint:** You may want to consult the lubridate cheat sheet to help with converting dates to their correct type.

Tutorial Question 2 - 2021/22 UEFA Champions League

The files "UEFA goals.csv" and "UEFA attacking.csv" contain all the player stats on goals and assists for the UEFA Champions League season 2021-22. Use the data to extract the following information:

a) Which two teams scored the most goals in the 2021/22 UEFA Champions League? How many goals did they score?

b) When ranking attackers, the sum of goals and assists is a widely used metric. Extract the top 10 players in terms of this metric. **Hint:** Not every player in the data recorded at least one goal and one assist.

Tutorial Question 3 - Working with discrete variables

When plotting values of discrete variables (such as number of sales) some information may not be visible because points may lie on top of each other. We will explore how the functions **geom_count()** and **geom_jitter()** facilitate handling such situations.

NOAA in the United States has recorded 332 billion-dollar natural disasters between 1980 and 2022, that is, each event caused overall damages in excess of $1 billion US dollars. The types of natural disasters include flooding, severe storms, droughts, etc. In this analysis we focus on the number of natural disasters per type and year and the data are provided in the file "US Billion Dollar Disasters.csv".

a) Create a scatter plot for number of flood events against severe storms per year. Discuss which information can be drawn from the plot?

b) Replace the function geom_point() in part a) by geom_count(). Which conclusions can be drawn from this new plot?

c) Replace the function geom_point() in part a) by geom_jitter(). What does this function do? Which conclusions can be drawn from this new plot?

Homework Question - Optimizing growing conditions for orchids

A research institute set up an experiment in 2024 to determine the best growing conditions for two types of Orchids: Cymbidium and Dendrobium. The design of the experiment was as follows:

- All orchids were planted in March or April 2024. Each individual orchid was exposed to constant temperature and phosphate levels, but levels differed across orchids.
- The height (in inches) and quality of each orchid were measured on 20 October 2024. Plant quality was assessed using a score between 1 and 10 (1=very poor, 10=excellent). Any orchid with a score above 6 is considered of "good" quality.

The research institute approached us to analyze their collected data which provides:

- **Type** - Type of the orchid.
- **Height** - Height of the orchid on 22 October 2024.
- **Quality** - Quality as measured on 22 October 2024.
- **Phosphate** - The level of phosphate in ppm (parts per million) the orchid was exposed to since it had been planted.
- **Temperature** - The temperature (in degree Celsius) the orchid was exposed to since it had been planted.
- **Planting** - The date the orchid was planted.

The full data are provided in the file “Orchids.csv” and the research institute is interested in:

a) How do the two types of orchids compare in terms of the heights measured on 22 October 2024?

b) What are the effects of phosphate and temperature on the height of the orchids?

c) Did the date when the orchid was planted have any effect on whether their quality was at least “good” on 22 October 2024?

Perform an analysis which considers the three aspects above. Make sure to clearly state your approach and conclusions.