MA22019 2025 - Problem Sheet 3

Data visualization with ggplot2

Overview

This week's tutorial questions revise some of the important R functions in dplyr and ggplot2 we may use for analysing real-world data sets. Tutorial Question 1 will help you to practice your skills in analysing complex data sets and you should focus on this question. Tutorial Question 2 provides some additional training on the dplyr R package. Finally, Tutorial Question 3 considers one aspect of data visualization we did not consider in the lectures, but which may nevertheless be useful in certain applications.

Exercises 1-3 help you with revising the techniques we will cover in the lectures in Week 3. There is no homework question as Coursework 1 will be released at 15:00 on Friday, 21 February 2025.

Before starting the questions, make sure to load the different packages we considered so far:

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

Tutorial Question 1 - Utopia's Ambulance Service

The country *Utopia* has collected data on their ambulance service and the patients admitted to the country's hospitals. The health department of Utopia has given you access to their data in 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 location
- **PatientID** - Health insurance number of the patient
- **Hospital** - Time patient arrived at the hospital (missing values correspond to the case that no hospital treatment 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

You are asked by the Utopian government to consider the following two tasks which are aimed towards analyzing the performance of their ambulance service and the factors influencing health outcomes.

a) How does the length of stay in hospital vary across the four ambulance response categories?

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?

Exercise 1 - CO_2 emissions across the globe

The file "CO2 Emissions.csv" contains daily CO_2 emissions across six different sectors between 1 January 2019 and 31 May 2023 for 11 countries, the EU, the rest of the world (ROW) and earth.

a) Which two countries had the highest total CO_2 emissions across all sectors for the considered time window?

b) Create a pie chart that illustrates the total emissions per sector for the UK between 1 January 2019 and 31 May 2023. Which sector recorded the most CO_2 emmissions?

Exercise 2 - Traffic between Minneapolis and St Paul

The file "Traffic Minnesota.csv" contains hourly data on the traffic volume for westbound I-94, a major interstate highway in the US that connects Minneapolis and Saint Paul, Minnesota. The data was collected by the Minnesota Department of Transportation (MnDOT) from 2012 to 2018 at a station roughly midway between the two cities.

The variables are

- **traffic_volume:** Hourly I-94 reported westbound traffic volume.
- **holiday:** Indicates whether the date is a US national holiday or a regional holiday (such as the Minnesota State Fair).
- **date_time:** Shows the hour of the data collected in local CST time.

a) Use the function **dmy_hm()** in the lubridate R package to convert the variable **date_time** to its correct type. Apply the function **weekdays()** to extract the day of the week and store this information as a separate variable within the data frame.

b) Explore the differences in traffic volume per hour for workdays and weekend / holidays. USe the results from your analysis to answer the questions in the Moodle quiz. **Hint:** You can use the function **hour()** to extract the hour from a date-time object.

Exercise 3 - Weather in Brisbane and Gold Coast

The file "BrisbaneGoldCoast.csv" contains daily weather measurements for Brisbane and Gold Coast in Australia.

a) Explore how the daily maximum temperature for Gold Coast varies throughout the year. Use the results from your analysis to answer the Moodle quiz.

b) Create a scatter plot of the daily maximum temperature for Brisbane and Gold Coast. What do you conclude?