MA22019 2025 - Problem Sheet 1

The dplyr R package

Overview

Exercises 1-3 will help you with revising the techniques covered in the lectures in Week 1 (Sections 1.1 and 1.2 of the lecture notes). You can check your solutions to these questions yourself by entering them on Moodle via the Quiz provided in the "Problem Sheets" section.

This week's tutorial questions ask you to apply functions from the dplyr and lubridate R packages to analyse real-world data set, and also introduce some important functions we did not use so far.

Your answer to the Homework Question can be submitted on Moodle to your tutor for feedback. The submission deadline is 17:00 on Thursday 13 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 exercises, make sure to load the dplyr and lubridate R packages:

```r
library(dplyr)
library(lubridate)
```

If you are using your own laptop, you may have to first run the code in "Load packages.R" (available from Moodle) to install all the packages essential for this course.

Exercise 1 - Functions for data wrangling

We described that data may not be in the format required for the analysis because

- Variable names are uninformative;
- Data types are incompatible with available R functions;
- We are only interested in a subset of the data.

For each of the following R functions, decide which issue they may address:

a) **as_date()**

b) **as.numeric()**

c) **filter()**

d) **group_by()**

e) **rename()**

f) **slice_head()**

If you are unsure about a function, you can use the **help()** function in R, or look at the cheat sheets provided on Moodle.

Exercise 2 - Flights from New York City airports

The file "NYCFlights.csv" provides data for flights in 2013 that departed from any of the three main airports servicing New York City: John F Kennedy (JFK), LaGuardia (LGA) and Newark (EWR). To load the data into your R Workspace, you can use

```r
NYCFlights <- read.csv( "NYCFlights.csv" )
```

Answer the following questions using the provided data:

a) How many flights does the data set contain?

b) For which month were the most flights recorded?

c) Which plane (specified by the **tailnum** variable) departed most often from New York City airports in 2013?

d) How often did the plane identified in part c) depart from New York City airports in January?

Exercise 3 - Earthquakes around the world

The data set "Earthquakes.csv" contains the location and size of all significant earthquakes, as recorded by the National Earthquake Information Center (NEIC), which occurred worldwide between 1965 and 2016.

We can load the data using

```r
Earthquakes <- read.csv( "Earthquakes.csv" )
```

Use the R functions covered in Week 1 to answer the following questions.

a) In which year between 1965 and 2016 were the most earthquakes recorded?

b) Were more earthquakes recorded in the northern or the southern hemisphere between 1965 and 2016?

Tutorial Question 1 - Honey production in the United States

In 2006, global concern was raised over the rapid decline in the U.S. honeybee population, an integral component to American honey agriculture. Large numbers of hives were lost to Colony Collapse Disorder, a phenomenon of disappearing worker bees causing the remaining hive colony to collapse. Speculation to the cause of this disorder points to hive diseases and pesticides harming the pollinators, though no overall consensus has been reached.

The data collected by the National Agricultural Statistics Service (NASS) for 1998-2012 is provided in the file "HoneyProductionUS.csv". The following information is provided:

- state - Initial of the U.S state
- numcol - Number of honey producing colonies
- yieldpercol - Honey yield per colony in pounds
- priceperlb - Average price per pound in dollars based on expanded sales
- year - Year to which the data relates

We are asked to extract some information on the amount of honey produced and its price.

a) For each state calculate the total amount of honey produced (in pounds) for the period 1998-2012. Identify the states which had the lowest and highest production.

b) How has the yield per colony of honey bees in the U.S. evolved from 1998 to 2012?

c) By which factor has the average price of honey increased in the state of Alabama between 1998 and 2012?

Tutorial Question 2 - Price of Bitcoin

The file "Bitcoin.csv" provides hourly data on the price of Bitcoin and the level of trade for the period 17 August 2017 - 19 October 2023. There are four variables:

- **date** - Day and time of the observation
- **close** - The value of one Bitcoin token at the end of the hour
- **volume.usdt** - The trading volume during the hour in Tether (USDT). This provides a stable reference point for trading volume because the value of USDT is supposed to be roughly equivalent to $1 USD.
- **tradecount** - The total number of individual trades or transactions that have occurred within the hour. This variable counts the actual number of separate buy and sell transactions.

Use the data to address the following questions:

a) Load the data and use the function as_datetime() in the lubridate package to convert the variable representing the date and time to its correct type. Explore how the value of one Bitcoin token has evolved between 17 August 2017 and 19 October 2023 (which is the period covered by the data).

b) Does the number of tokens traded per day follow the same pattern as that found in part a) for the price of one token?

Homework Question - Salaries for data scientists

The file "Salary Data Scientists.csv" provides data for 603 employees working in Data Science / Machine Learning / AI. For each employee, we are given the following information

- **work_year** - Year the data were collected for the employee
- **experience_level** - Experience Level: Entry-level (EN), Junior (MI), Senior (SE) or Expert (EX)
- **employment_type** - Full-time (FT), Part-time (PT) or Contingent Workers (CT)
- **salary, salary_currency** - Nominal salary and currency in which it was paid
- **salary_in_usd** - Nominal salary in U.S. Dollars
- **employee_residence** - Country the employee is residing
- **remote_ratio** - Proportion of time the employee may work from home
- **company_location** - Location of the company's headquarters
- **company_size** - Size of the company: Small (S), Medium (M) or Large (L)

An undergraduate is considering to take units in Data Science / Machine Learning / AI. However, they are unsure about the working conditions they are likely to experience if they decide to work in the sector. Specifically, the student plans to work full-time for an U.S. company, and they want to get some insight into the differences between small-/medium-sized and large companies. After a discussion with the student, the following three aspects have been identified as important for them:

a) Opportunity to work remotely for at least 50% of the time.

b) Salaries for Entry-Level positions.

c) Structure of the workforce in terms of experience level.

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