Homework 1 - DSA

Please prepare your answers in an R Markdown Document (HTML, including output) and upload on Moodle.

Exercise 1: Data Cleaning

For this exercise, you'll work with data that contains real orders by customers of a clothing store.

On moodle, you can find two data sets, `orders.csv` and `users.csv`. Load them using the `read.csv` function.

```
orders <- read.csv("orders.csv")
users <- read.csv("users.csv")
```

The `orders` data set contains 10 variables:

- `order_item_id` is a unique ID for each order
- `order_date` indicates the date of the order
- `delivery_date` indicates the data of the delivery
- `item_id` is an ID for each item
- `item_size` indicates the size of the product, e.g., "XXL"
- `item_color` indicates the color of the product, e.g., "blue"
- `brand_id` is an ID for each brand
- `item_price` indicates the price of the product in Euros
- `user_id` is an ID for each user
- `return` indicates whether the item was returned (1) or not (0)

The `users` data set contains 5 variables:

- `user_id` is a unique ID for each user
- `user_title` indicates the title of the user, e.g., "Mr."
- `user_dob` is the birth date of the user
- `user_state` indicates the state of residency of the user, e.g., "Berlin"
- `user_reg_date` indicates the date when the user registered with the store

Task 1

Merge the two data sets to form a data set `orders_full`. Which variable can be used as the key?

```
#orders_full <-
```

Task 2

Which types do the variables have? Use the `str()` function.

- Convert all variables that contain dates to a proper date format (you can use the `ymd()` function).
- Convert `return` to a factor variable with labels "Yes" and "No".

Task 3

Compute summary statistics for all variables. Do you notice any peculiarities?

- Which variables have missing values?
- Briefly describe two different methods to impute missing values.

- Compute a variable `delivery_time` as the difference between the delivery date and the order date. What do you notice? Propose and execute a way to clean the data.
- Plot the distribution of user date of birth? What seems odd? Propose and execute a way to clean the data.

Task 4

Assess all values that the variable `item_size` can take. What do you notice? Can you propose a way to use the information in this variable to create product categories?

Exercise 2: Web Scraping

a) Write an R function that computes the point difference between the first and second team in the German football Bundesliga as a function of the season and the matchday.

```
point_difference <- function(season, matchday){
...
}
```

b) Create a plot the shows the difference between the first and second team on matchday 34 for each season from 1995/96 to 2023/24.

Hints

- You can scrape the data from the webpage www.kicker.de. For instance, the table for the 11th matchday of the 2022/23 season is shown on `https://www.kicker.de/bundesliga/tabelle/2022-23/11`.
- You can use the function `paste()` to combine multiple strings, e.g. `paste("test1","test2",-)` returns `"test1-test2"`
- Clean the data you received from the webpage and convert it to the right format
- Compute the difference between the first and second team and return this difference
- You can use a `for-loop` to circle through the different seasons when you apply the function you have written (alternatively you can use the `apply` family if you are familiar with that)