Workshop: An introduction to R and building (supervised) machine learning models in R
@ AKB, SPRING/SUMMER 2021

EXERCISE OVERVIEW

DAY 1

- ***01 – What is the role of R in the data science ecosystem? How to get started with R?***

  No exercise.

- ***02 – Why and how to use RStudio as a code editor for R? Why do I need packages in R?***

  1. Install and load the `data.table` package (use RStudio's code editor).
  2. Find and use an R command to list all installed packages (use RStudio's code editor).
  3. Have a look at the help file for the R command `library()`. What are optional arguments when loading a package?

- ***03 – How to access and check data in R? Loading data files and basic data investigation***

  Use the dataset `transactions.csv`. This dataset contains several variables with information on customer transactions.

  1. Set your working directory.
  2. Read in the CSV file (with the help of the `data.table` package). Make the data available for further use and name it `myData`.
  3. Take a closer look at the data (i.e. look at the first / last lines of the dataset, view the complete dataset, and check the data type with `str()`).
  4. Use the lubridate package to format the TransDate column.
  5. Use `str()` to see if the change was made correctly. (How can you tell?)
  6. Use `summary()` to get the summary statistics.
  7. Save the `data.table` object to a csv-file with the name "transactions_backup.csv". Use the `fwrite` function.

- ***04 – How to plot data in R with R base? The basics***

  Use again the dataset `transactions.csv` (stored as `myData`).

  1. Read in `transactions.csv` and call it `myData`. Note: If plotting takes relatively long, please specify the `fread()` argument `nrows` to read in less rows.
  2. Create a histogram for the variable `PurchAmount(x)`.
  3. Create a scatter plot for the variables `Cost(y)` and `PurchAmount(x)`. Can you observe any correlation?

- ***05 – How to plot data in R with R base? An overview of advanced plotting options***

  No exercise.

- ***06 – How to plot data in R with ggplot2? The basics***

  Use again the dataset `transactions.csv` (stored as `myData`).

  1. Install and load the package `ggplot2`.
  2. Create a histogram for the variable `PurchAmount(x)` with `ggplot` .
  3. Create a scatter plot for the variables `PurchAmount(x)` and `Cost(y)` with `ggplot`.

- ***07 – How to plot data in R with ggplot2? An overview of advanced plotting options***

  No exercise.

- ***08 – What are data pipelines?***

  No exercise.

- ***09 – How to wrangle data with R? Select operations***

  Use again the dataset `transactions.csv` (stored as `myData`).

  1. Select rows 10 to 20.
  2. Select all purchases from 2010.
  3. Select all purchases with purchase amount greater than 100 which were made from 01.01.2009 onwards.
  4. From `myData`, create a new column calculating the difference between `PurchAmount` and `Cost`. Call it `Profit`.
  5. Rename `Profit` to `ProfitChange`.
  6. Delete `ProfitChange` again.

- ***10 – How to wrangle data with R? Aggregate operations***

  Use again the dataset `transactions.csv` (stored as `myData`).

  1. Calculate the sum of purchase amount by customer and transaction date.
  2. Determine the highest purchase amount for a single transaction for each customer.
  3. Create a new column in your data table and store, for each customer and transaction, the quantity purchased in the next transaction. *Hint: You can do this by creating an aggregated lead shifting variable for the variable `Cost`. Use an offset of 1 and aggregate the data by customer. You can name the resulting column `CostLead`.*

- ***11 – How to wrangle data with R? Merge operations***

  Use again the dataset `transactions.csv` (stored as `myData`).
  In addition, use the dataset `demographics.csv`. This dataset contains information on customers' gender, birthdate, zip code and join date.

  1. In addition to `transaction.csv` (stored as `myData`), read in demographics.csv and call it `demographics`. Make sure `Birthdate` is in the right format.
  2. Merge the tables `transactions` and `demographics` by the column `Customer` using an outer left join.
  3. Merge the tables `transactions` and `demographics` by the column `Customer` using an inner join. Do this only for the customers born after 1980.

- ***12 – How to wrangle data with R? A comparison with SQL (Connecting to a database)***

  No exercise.

- ***13 – How to wrangle data with R? A comparison with SQL 2 (Using SQL in R)***

  No exercise.

- ***14 - Practicing real-world analytics - RFM (part 1)***

  Use again the dataset `transactions.csv`.

  14A: Preparations
  1. Set your working directory.
  2. Install (if necessary) and load the packages `data.table`, `lubridate`, `ggplot2`, and `Hmisc`.
  3. Load the data `transactions.csv` with `fread()` and name it `transactions`.
  4. Transform the variable `TransDate` to datetime with `lubridate()`.

  14B: Aggregation of variables
  1. Save the latest transaction as the object `now` in your R environment.
  2. Create a new `data.table` called `rfm` that includes the customer ID, as well as the measures for purchase recency, frequency, and monetary value.
  3. Check the structure of the new table and ensure that all the variables are numeric.

  14C: Descriptive Statistics
  1. Inspect the newly created RFM measures by taking a look at the data summary.
  2. Plot the histograms for all 3 measures in `ggplot2` and arrange them in a single figure.
  3. Adjust the title, labels, and colors of your plots in an appealing way.

- ***15 – Getting help & wrap-up day 1***

  No exercise.