# Homework assignment week 4

# GitHub Instructions

1. Use the same repository as in week 2 (ds5\_assignment\_ group#), make sure all the group members and teachers are still collaborators.
2. Create one .ipynb (or .py) file that contains the answers to all the exercises below! However, you can call other files in your main file.
3. Make sure that you push and commit all your changes to the file before the deadline. If changes are made after the deadline, it will not be taken into account for the bonus point.

# Exercise 1: Project 4 data cleaning

In the lecture, we discussed pre-processing by making use of data cleaning. Data cleaning is a very wide concept, and we can apply a lot of techniques for data cleaning. A few steps you can take are as follows:

Fix Structural Errors/Data Standardization:

* Converting all text to lowercase or uppercase for consistency.
* Formatting dates and times into a standardized format.
* Normalizing numerical values to a specific scale or range.
* If categories are present, clean the data in such a way that they can be analysed.

Handling Missing Data:

* Removing rows or columns with a high percentage of missing values.
* Imputing missing values using techniques like mean, median, or mode.
* Using advanced imputation methods such as regression or nearest neighbours.

Removing Duplicates:

* Identifying and removing duplicate records based on key fields.
* Handling duplicate values within a single record or field.
* Resolving conflicts when duplicate records contain conflicting information.

Handling Outliers:

* Detecting and handling outliers using statistical methods or domain knowledge.
* Applying techniques like winsorization, truncation, or clustering to manage outliers.
* Assessing the impact of outliers and deciding whether to remove, transform, or treat them separately.

Validate data:

* Does the data make sense?
* Does the data follow the appropriate rules for its field?
* Does it prove or disprove your working theory, or bring any insight to light?
* Can you find trends in the data to help you form your next theory?
* If not, is that because of a data quality issue?

Deciding which techniques to use, depends on the data itself. For example, you are not going to standardize dates if your data set does not contain dates.

In Project 4 you received data from Royal Swinkels Family Brewers. You can find a part of these data in the file *‘dataProject4.xlsx’*. For this exercise, you need to read the Excel file, analyse the data and apply at least three data techniques. In particular:

1. Read the excel file using the pandas library
2. Apply *at least three* data cleaning techniques of which you think are suitable for this data set. You can choose from the ones above, or think of a technique yourself.
3. Structure your code by combining all functionalities into a function. This function should return the cleaned data set.

# Exercise 2: Hotel Bookings

You have been given a file about hotel bookings, you can find these data in the file *‘hotelBookings.xlsx’*. These data contain the number of bookings of the hotel ‘Resort Hotel’, which is spread over different countries. This file contains a lot of errors, and needs to be cleaned. For this exercise, you need to read the Excel file, analyse the data, find as many mistakes as possible and correct them. In particular:

1. Read the excel file using the pandas library
2. Inspect the data and find as many mistakes you can.
3. Clean the data, which decisions did you make to solve the errors?

# Exercise 3: Retail 2.0

You have been given the file ‘*detailedRetail.xlsx’*, which contains the sales data for a company including the responsible sales manager. For this exercise, you need to read the Excel file, analyse the data and generate a sales report with key metrics. In particular, perform the following steps:

1. Read the excel file using the pandas library
2. Perform data analysis on the sales data to calculate the following metrics:

* Total sales for each category
* Percentage that a category contributes to the total sales
* Sales for each month
* Percentage that the month contributes to the total
* Sales for each sales manager
* Percentage that the sales manager contributes to the total

1. Generate a report (dataframe) that displays the metrics calculated in step 2.
2. Store the report into an excel file called ‘reportRetail.xlsx’
3. After applying these techniques, structure your code by adding docstrings and breaking down your code.

# Exercise 4: Post-processing of tweets

Everyone knows the social media platform Twitter, a place where people can post Tweets about their life, politics, interests and so on. As Twitter is very popular these days, the company wants to analyse the sentiment of all the tweets. To be specific, they want to know whether it is positive, negative or neutral. A part of the tweets can be found in ‘*tweets.xlsx’*. However, not all these tweets are in English, which means that we have to post-process this file to make it useable for Twitter.

# Exercise 4.1: Language detection

To find out the sentiment of a tweet, we first have to find out in which language this tweet is written. In this exercise, detect the language of a tweet, and store it in the column ‘language’. There are already multiple packages available that you can use. One of the most famous packages is the **langdetect** package in Python. An example of how your output can look like is as follows:

|  |  |
| --- | --- |
| Tweet | Language |
| Hello Wolrd | en |
| Bonjour! | fr |
| Hola! | es |
| Ciao! | it |
| привет! | ru |

Handle any exceptions by assigning ‘Unknown’ to the language column and install any missing libraries using pip (or conda) install.

# Exercise 4.2: Sentiment detection

In the previous exercise, we have established the language of the tweet. In this exercise, you will be asked to define the sentiment of a tweet based on the language. As this is not a straightforward approach, we have defined a few steps that should be taken:

1. Create a function analyze\_sentiment\_english(tweet) that takes a tweet as input and performs sentiment analysis using the TextBlob library. The function should return 'positive' if the polarity score is greater than 0, 'negative' if the polarity score is less than 0, and 'neutral' otherwise. Hint: Use the TextBlob library to create a TextBlob object from the tweet text and access its sentiment.polarity property.
2. Define a function analyze\_sentiment\_other(tweet) that takes a tweet as input and performs sentiment analysis using the SentimentIntensityAnalyzer from nltk. The function should return 'positive' if the compound score is greater than or equal to 0.05, 'negative' if the compound score is less than or equal to -0.05, and 'neutral' otherwise. Hint: Use the SentimentIntensityAnalyzer from nltk.sentiment to create a SentimentIntensityAnalyzer object and calculate the sentiment scores using its polarity\_scores() method.
3. Apply language-specific sentiment analysis to each tweet in the DataFrame. Iterate over each tweet and use the appropriate sentiment analysis function based on the identified language. Store the sentiment result in a new column 'sentiment' in the DataFrame.

An example of how your output can look like is as follows:

|  |  |  |
| --- | --- | --- |
| Tweet | Language | Sentiment |
| Hello Wolrd | en | positive |
| Bonjour! | fr | positive |
| Hola! | es | positive |
| Ciao! | it | positive |
| привет! | ru | positive |