Exercise: Introduction to Data Science

You are working as a data scientist for a company which maintaining a larger car fleet for their maintenance services. The company would like to know:

Do we have an added value for the company by knowing more about the gas prices?

The provided data (Olat System) show an extraction of the gasoline prices in Germany. The full historic and the description of the data can be accessed at: https://creativecommons.tankerkoenig.de.

Exercise Goal:

- Learn a systematic approach to deal with a high level business request
- Learn a data driven approach to ask and answer the correct questions
- Learn a systematic approach to derive a machine learning model

Procedure:

- Exercise 1: Understand the data, visual explanation
- Exercise 1: Define a possible business case

- Exercise 2: Develop a simple predictive reference model for reference
- Exercise 2: Develop a machine learning model related to a business case
- Exercise 2: Analyze and compare the result and understand the modelling procedure

Deadlines: **(4 days later than initially announced)**

- Wednesday. 03.07.2019 deadline programming exercise 1
- Wednesday 21.08.2019 deadline programming exercise 2

Acceptance criteria for each exercise

- For each question a representative visual chart including a text for interpretation has to be performed
- The results have to be submitted via a 2 IPyhton notebooks, one for each exercise
- At least 3 test have to be included in each exercise
- At least 3 function have to be included in each exercise
- Tests have to be written for each of the 3 functions (import unittest)
- The results have to be submitted via the GITLAB: https://classroom.github.com/g/hjbNYUyB

Exercise 1: understand/analyze the data

1. How many different stations exist in the data set and what is the existing history in days (bar chart)?
2. What is the min, mean, max price for each gasoline type and station weekly (time series graph)
3. What is cheapest station (in average) and why?
4. At which day of a weak is the price most likely the cheapest (week profile)
5. At which hour during a day is the price the cheapest in average (hour profile)
6. How many different station locations are present in the data (visualize via a map)
7. What is the gas station which has most price data points, choose one and draw the time series for all 3 gasonline types
8. At which hour during a day do we have the most price changes
9. Select 20 gas stations having the longest time history and visualize the average price per month. Use heatmap and only the prices between 12:00-13:00 of e10 and diesel.
10. Describe a possible business potential in € for the customer (textual description in the ipyhton file). Define the constraints of the business case 5 lines, the answer max 15 lines (high level summary)

Exercise 2: time series predictive model and evaluation

Build a model which predicts the average daily price of the next day.
Apply a time series back testing procedure! (two third of the data for training and one third for evaluation of the model

Step 0: prepare data set

- Attention for this the data set has to be cleaned and normalized to daily data!
- The entire analysis has to be performed on 100 gas stations
- Choose the 100 gas stations with the longest time history
- Ensure that all stations have all identical time horizon and steps (open and close hour identical for all 100 stations, take 7:00 to 20:00). Align all data before calculating the average price of the day to ensure comparable daily prices!
- ensure a perfect data set, clean outlier

Step 1: start with small (3 stations) data set and develop a simple reference model

- Extract a small good data set you can start with
- Start with a trivial model (reference model), use a simple moving average
- Use the Time Series Cross Validation (TimeSeriesSplitt) of scikit-learn
- Check and visualize the results
- Use an appropriate metric for the forecast error: mean absolute deviation (MAD) and mean absolute percentage error (MAPE)
- Do a visualization of the predicted output (time series

Step 2: go with a more complicated model

- Use a more elaborated time series modeling.
- Use the Facebook Prophet package (https://facebook.github.io/prophet/docs/quick_start.html)
- Compare the prediction result against the trivial reference model, visually on MAD / MAPE

Step 3: Do the comparison on the full 100 data set, are the results for e5, e10 and diesel