Exercise: Introduction to Data Science (2018)

The entire exercise is split in two parts.

Part one is a simple introduction to a predefined problem and to learn basic skillsets.

Part two applies data science knowledge (business, method, interpretation, coding) on a new data set with a more realistic problem statement

Part 1

Task	Short Description	Goal
Publish results in GITHUB	GITHUB is the most important hosting platform for version control and collaboration	lean version control system and publish own (lecture) activities
Participate in kaggle competetion	kaggle is one important source to learn and promote your skills.	get a score number at kaggle's titanic example https://www.kaggle.com/c/titanic
Hacking skills	Python hacking skills and practice is a minimal prerequisite to become a data scientist	

Part 2

You are working as a data scientist for a company which maintaining a larger car fleet for their logistic services. The company hat a truck fleet (100) with an average milage per truck and year of 220 000 kms and operates in Gemany. The company would like to know:

Can we define an added value for the company by knowing more about the gas prices?

The provided data show an extraction of the gasoline prices in Germany. The full historic of the data can be accessed at: https://creativecommons.tankerkoenig.de. There is as well a description of the data set

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
- Derive a possible business model and judge on what is possible or not

Procedure:

- Understand the data
- Define a possible business case
- Develop a model
- Analyze the result
- Present the result

Task1: understand/analyze the data

Questions to be answered (understand the data):

- How many different locations are present in the data

- How many different brands are there

- What is the min, max price for each gasoline type, per month

- Mandatory Homework: Find 5 more questions which might be of interest and present the analytics results (visual plots)

Task 2: define a possible business potential

- Describe a possible business potential for the customer
- Do high level calculations of the business case in €

Task 3: develop a predictive model

- write a predictor on the gasoline prices (define your horizon carefully)
- start with a trivial predictor, enhance to a (possible) stronger model (scikit-learn)

Task 4: Analyse the result

- interpret the results from a mathematical perspective
- interpret the result from a business perspective (think about the realistic assumption that a truck diver will not always follow the best decision)

Task 5: Present the result

Imagine you have 3-5 slides to present the result to a management board

- prepare the presentation

Deadline: all tasks have to be finished one week before the oral exam

- source code has to be uploaded to your GITHUB account
- 5 min presentations (3 slides) with the key findings has to be given within the oral. The presentation should address (business challenge, key findings)