Fundamentals of Artificial Intelligence
Lab 2: Exploratory Data Analysis - pandas & matplotlib

A story about pandas...

Once upon a time in a bustling city, there lived a brilliant data scientist named Emily. She had always been fascinated by the mysteries hidden within numbers and data. Her life was filled with algorithms, equations, and endless streams of information. Emily was content, but she yearned for a change, a new challenge that would take her beyond the confines of her computer screen.

One sunny morning, as Emily sipped her coffee and browsed through job listings, she stumbled upon a unique opportunity that piqued her interest. The headline read, "Data Scientist Wanted: Zoo of Wonders." Of course she applied and got hired.

Emily's work quickly improved the zoo. She optimized resources, enhanced visitor experiences, and ensured the well-being of animals like Ming and Mei, the pandas.

As months passed, Emily's data-driven insights made the zoo thrive. She found her true calling, turning numbers into a force for good. The zoo became her home, and she relished making a difference, one data point at a time.

You see, data is very important in many fields. With lots of data, you can uncover hidden patterns, make informed decisions, and drive innovation that can transform industries and improve lives. But it's also important to find good data and process it in a useful way.

This lab focuses on data analysis. You will be presented with a set of data and you should process this data, compute and output different statistics, using 2 of the popular Python libraries - **pandas** (`https://pandas.pydata.org/`) and **matplotlib** (`https://matplotlib.org/`). The first one is a Python library for working with tabular data (pandas = panel data) and the second one is a powerful visualisation tool.

General guidelines

- **This lab should be done with Python, pandas and matplotlib. Other programming languages are not allowed.**
- Submit your solution as **.ipynb** file containing both code and explanations.
- Please don't host your solution in public repositories (e.g Github etc). You can use private repositories if you need.
- **Plagiarism will not be tolerated.**

Grading policy

For this lab, your objective is to investigate a real-world dataset provided by Rossman GmbH. This dataset comes from an online machine learning challenge, focusing on predicting sales for multiple stores. To begin, access and familiarize yourself with the dataset attributes available at the following link: `https://www.kaggle.com/c/rossmann-store-sales/data`.

Download the *'train.csv'* and *'store.csv'* files.

- **Task 1**: Find the store that has the maximum sale recorded. Print the store id, date and the sales on that day. ***(0.5p)***
- **Task 2**: Find the store(s) that has/ve the least possible and maximum possible competition distance(s). ***(1 p)***
- **Task 3**: Check if there are any missing values in the dataset and output the number of missing values per each column. ***(0.5p)***
- **Task 4**: Plot the monthly mean of sales across all stores using matplotlib. ***(1p)***
- **Task 5**: Which store type ('a','b' etc.) has had the most sales? ***(1p)***
- **Task 6**: What is the difference in the mean of sales (across all stores) when offering a Promo and not? Plot this data with matplotlib. ***(1p)***
- **Task 7**: For the store with id 1, plot the mean sales per each day of week in a pie chart by using matplotlib. ***(1p)***
- **Task 8**: Plot the mean of sales across all the stores for each day of the week recorded in the dataset, by using matplotlib. ***(1p)***
- **Task 9**: For the first 10 stores (first 10 ids), draw boxplots of their sales by using matplotlib. ***(1p)***
- **Report & Presentation of the solution**:

  Clear explanations, report formatting, code quality, comments in the code, docstrings, visualisations if relevant etc. ***(2p)***

Good Luck!