Take-home assignment for TietoEVRY Senior Data Engineer

# Introduction

In this assignment you will be analysing usage statistics from a Helsinki outdoor gym.

You should perform the tasks using Python. In some tasks, we give guidance on how to perform them. Other than that, you are free to select the libraries you use and the form of your solution.

The purpose of this assignment is to assess your capability to:

* Perform data analysis and transformations in Python
* Produce readable, testable and maintainable code
* Justify your design choices in later discussions

The aim is that you would not use more than 2-3 hours on this exercise.

**Expected deliverables:**

Link to a public Git repository in GitHub or similar with the following:

* Coded solutions to the following tasks
* List of Python libraries and versions used
* Any additional documentation required to understand your solutions

Best of luck! :)

# Tasks

The source dataset is usage statistics of gym equipment for Hietaniemi outdoor gym in Helsinki. You can find the original source page for the dataset here:

<https://www.avoindata.fi/data/en_GB/dataset/helsingin-ulkoliikuntapaikkojen-laitteiden-kayttoasteet>

In this exercise you should use the aggregated version of the dataset, which has been provided in file **“hietaniemi-gym-data.csv”**. You can find instructions for interpreting the dataset in attached PDF **“de\_homework\_dataset\_explanation.pdf”**, or through this link:

<https://hkikanslialiikuntapaikat.z6.web.core.windows.net/ulkokuntosali/README.html>.

Section "Processed data" describes the dataset used in this exercise. Instructions are in English, with some Finnish mixed in between. Notably, the names of the different gym devices are in Finnish, however the exact names of the devices don’t matter in this assignment. In any case, if you have any issues understanding the instructions, don't hesitate to be in contact.

You have also been provided a dataset of weather observations for the same time period in file **“kaisaniemi-weather-data.csv”.**

## Task 1

Read in source dataset to a format usable in later data analysis. Aggregate it to hourly precision by summing the usage statistics for each gym device during the hour. Present 10 first rows of the dataset.

## Task 2

Perform the following quality checks on the data:

* There are more than 50,000 rows in the dataset
* There are records from between 2020-04-24 and 2021-05-11
* All values in the numerical columns are positive

Bonus points for using a test automation framework such as Unittest, Pytest or similar.

## Task 3

Analyse the dataset for the following:

* What was the most popular device during the tracking period measured by number of minutes used?
* Did time of day (hour) impact overall popularity of the outdoor gym?
* Was the gym more popular overall on weekends (Saturday and Sunday) than on weekdays?

## Task 4

Add the following new features to the dataset:

* Weekday as number
* Hour as number
* Sum of minutes across all gym devices

## Task 5

In this task we analyse the impact of weather on gym popularity.

In file **“kaisaniemi-weather.csv”** you will find hourly weather data from the closest measuring station in Helsinki for the same period as the gym usage data.

Join this dataset with the gym usage dataset, so that you have both the gym device usage statistics and weather observations for each hour. Then analyse the following:

* Does temperature impact gym popularity?
* What about precipitation?

NOTE: In this dataset, for precipitation and snow depth, value 0 means that there can have been some of either, whereas value -1 means that there really was no precipitation or snow depth. You should consider this in your analysis.

## Bonus task

In this task you'll add inference results to a dataset using a pretrained model file.

The pretrained model is provided in Python joblib pickle file “model.pkl”. The model was trained using Scikit-Learn version 1.0.2. You can read the model using Python joblib’s “load” function.

The model expects the following features in this order:

* Weekday as integer (Monday as 0, Sunday as 6)
* Hour of day as integer (from 0 to 23)
* Precipitation in millimeters as float
* Snow depth in centimiters as float
* Temperature in Celsius as float

Load the model and add predictions produced by the model as a new attribute to the dataset. Do the predictions look believable to you?