# Lab 2: Data Preparation (2%)

---

In the previous lab exercise, you practised performing exploratory data analysis (EDA) of a dataset without making changes to the dataset. Once you have a good understanding of the dataset, a next step in the data analytics process is to prepare this dataset for more elaborate analysis that typically involves machine learning algorithms. This preparation is known as ***data munging*** or ***data wrangling***, and it typically involves:

- encoding categorical features as numerical,
- filling in missing values,
- rescaling numerical features (i.e. columns) or normalising examples,
- appropriately treating outliers/extreme values,
- creating new features (i.e. columns) from the existing ones.

At the end of this lab exercise, you should be able to apply commonly used techniques for data preparation.

---

**Task 1**

Download the following dataset and example Jupyter notebook:

- loans_train.csv
- Lab 2 - Data Preparation.ipynb

---

**Task 2**

Download the dataset: bank.csv.

---

**Task 3**

Open a new Jupyter notebook. In your notebook, load **bank.csv** into a Pandas dataframe. Follow the examples and process in **Lab 2 - Data Preparation.ipynb** to prepare **bank.csv** for machine learning.

Ensure that you have:

- encoded all categorical attributes as numerical,
- treated/filled in missing values
- applied scaler where necessary
- treated outliers appropriately
- created at least one new feature from the existing ones.

In the end, save the prepared dataset as a CSV file on your hard disk.

---

**Submission**

Save your Jupyter notebook and name *lab2.ipynb*.

Submit your notebook in the Course Tools > Assignment section by 01-Mar-2024 23:59. Late submissions will not be accepted.