# Exploratory Data Analysis and Visualization - MSc AIDA UoM

Professor George Evaggelidis

Academic Year: 2021-2022

**Exercise 3.1: Reading and Manually Checking**

a. Download the `dirty_iris` dataset from this link. Open the file in a text editor to determine its format and read the file into R ensuring that strings are not converted to factors.

b. Calculate the number and percentage of observations that are complete in the dataset.

c. Check for special values in the data. Replace any special values with NA.

**Exercise 3.2: Checking with Rules**

a. Define data integrity rules based on the following background knowledge:

- Species should be one of the following values: setosa, versicolor, or virginica.
- All measured numerical properties should be positive.
- The petal length should be at least 2 times its petal width.
- The sepal length should not exceed 30 cm.
- Sepals are longer than petals.

Save these rules in a separate text file and read them into R using `editfile` (package `editrules`). Print the resulting constraint object.

b. Determine the frequency of each rule violation using `violatedEdits`. Summarize and plot the results.

c. Calculate the percentage of the data without any errors.

d. Identify observations with excessively long petals using the results from `violatedEdits`.

e. Identify outliers in sepal length using `boxplot` and `boxplot.stats`. Retrieve the corresponding observations for further examination and consider setting outliers to NA or another appropriate value.

**Exercise 3.3: Correcting**

a. Replace non-positive values in `Petal.Width` with NA using `correctWithRules` from the `deducorrect` library.

b. Replace all erroneous values with NA using the result of `localizeErrors`.

**Exercise 3.4: Imputing**

a. Use kNN imputation (package `VIM`) to impute all missing values.

b. Employ sequential hotdeck imputation for `Petal.Width` by sorting the dataset on `Species`. Compare the imputed `Petal.Width` with the original using the sequential hotdeck imputation method. Note the ordering of the data.

c. Repeat the hotdeck imputation but sort the dataset on both `Species` and `Sepal.Length`.