Fourth Industrial Revolution (4IR) Summer School

Data **A**nalysis **F**oundations – **D**ay **2** exercises

Question 1 [EDA]

Perform EDA on Iris dataset by doing:

- Descriptive statistics
- Removing duplicate data entries
- Compare between various species based on petal length and width.
    - Write your data insights in a notebook text
- Boxplot petal width distribution over species
    - Write your data insights in a notebook text

Question 2 [Feature Selection]

The Breast Cancer dataset is publicly available dataset, with the following characteristics:

- Target variable.
    - Two categories are: malignant and benign.
- Total observations **569**
    - malignant class has 212 samples
    - benign class has 357 samples
- **30** features
- Get it from either
    - Sklearn

```
from sklearn import datasets
cancer = datasets.load_breast_cancer()
x = cancer.data
y = cancer.target
```

    - https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

Perform the following feature selection on this dataset

- **Removing features with low variance**
- **Univariate Selection**
- **Recursive Feature Elimination**
- **Principal Component Analysis**