Computer Science & Information Systems

# Data Management for IoT - Lab sheet 2

Exercise 1 – Data exploration using pandas and visualizations

# Objective

The objective is to

* + understand the steps involved in the usual data exploration step
    - Load the structured data into Pandas Dataframe
    - Identifying and fixing the outliers
    - Getting familiarity with Pandas functions for data exploration
    - Exploring the data with visualizations

# Steps to be performed

**Tool** Python3 and Jupyter Notebook

**Libraries required** Pandas, NumPy, MatplotLib and Seaborn

**Input**

cleaned\_beach\_water\_data.csv

**Analytics Technique** Exploratory Data Analysis

**Implementation** DMIoT - Lab2 - Beach Water Quality Analysis -2.ipynb

**Steps**

* + Import required Python library and convert to as data frame
  + Get quick understanding of data using Pandas functions
  + Identify the presence of outliers in the given dataset
  + Using the different techniques get rid of the outliers
  + Perform the univariate analysis of categorical and numeric variables
  + Perform the bivariate analysis of categorical and numeric variables
  + Make use of visualizations to find out hidden insights from the dataset
  + Plot the visualizations using date time axis

# Results

* + Outliers are identified and removed / smoothened out
  + Meaningful observations are noted by performing univariate analysis of categorical and numerical attributes using Pandas and Visualizations
  + Insights are obtained by comparing two or more variables using Pandas and Visualizations
  + Correlation between the variables is noted down

# Observation

* + The effect of outliers is noted down
  + Meaningful insights after univariate and bivariate analysis of data are noted down.