**FORECASTING USING TIME SERIES ANALYSIS**

**GRADED ASSIGNMENT 2**

**DEADLINE: 30th MAY 2023**

**Dataset File =**

1. [**Machine\_temp\_failure\_missing.csv**](https://drive.google.com/file/d/17bqqiJZ7JCXPt0EwseFBPZbBIuV7-YcG/view?usp=sharing)
2. [**machine\_temp\_failure\_original.csv**](https://drive.google.com/file/d/1kulRVMpepw903zqZAIETRHR3CKCw9hCT/view?usp=sharing)

**Tasks =**

* Read the **machine\_temp\_failure\_missing.csv** dataset into a Pandas DataFrame, must use PathLib. Inspect the dataset, and look for missing values if there are any. Find statistics such as mean, max, min. Rename the **value** column to **temperature**.
* Create a new feature **delta\_temperature** which would be **temp[i+1] - temp[i]** where **i** is in **range(Index).** The feature represents the difference between consecutive temperature values.
* Generate two separate plots for **temperature** and **delta\_temperature** using  **matplotlib**.
* Generate a plot in which **temperature** values greater than 70.00 are marked as red, and those below 70.00 are marked as blue.
* Write three functions **read\_datasets(), plot\_dfs(), rmse\_score()** for these datasets (as discussed in 1\_4 notebook).
* Perform univariate imputation on **temperature** and  using Pandas (**mean, ffill, bfill**) and Scikit-Learn (**SimpleImputer**), and **interpolation.** Compare the **RMSE** score for each imputation with **machine\_temp\_failure\_original.csv**  and find the method with the least RMSE score.
* Now the dataset with the least RMSE score  to check the frequency of data and resample the dataset to an hourly frequency.
* Draw **box plot, boxen plot, lag plot** and write your conclusions about outliers (in your own language). Define **iqr\_outliers()** function for this data, and find the outliers. Compute **z-score** and plot **z-score** as discussed in the notebooks.