**Exercise – 5: DS203-2023-Sem2**

**Submissions due by: March 24, 2024, 11:55pm**

In this Exercise you will be processing the following datasets:

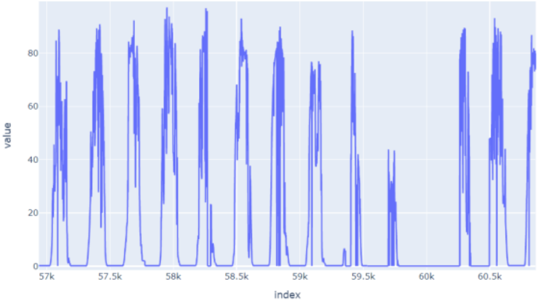
- e5-htr-current.csv: This data set tracks the current flowing through a transformer.
- e5-Run2-June22-subset-100-cols.csv: This data set contains process parameters of a chemical plant
- mnist_test.csv: This data set is a very small subset of the MNIST dataset

---

**Problem 1: Dealing with outliers / missing / incorrect data**

A. Perform EDA on the data file **e5-htr-current.csv** to understand the overall nature / quality of the data. Create a description of the data and document your EDA observations in detail. You might want to use **plotly** as **one** of the EDA tools …

B. Identify a 2-week period, where the data seems to be relatively **unstable**. For example, see the image below: there are wild fluctuations and missing observations.



C. In your report, clearly state the start and end dates you have chosen for such detailed analysis.

D. On this data segment, implement at least 3 methods to remove outliers (if any) / smoothen the data / impute missing data, thereby converting the relatively bad data into good data.

E. Can you think of a method that uses other **good data** – elsewhere in the data set – to guide you into treating the bad region that you have identified? That is, can you use the global data trend information to make local changes?

F. For each of the methods implemented, describe the steps, and clearly show your final results (visually + using statistical measures). Adequately justify your decisions!

---

**Problem 2: Outliers, missing values, scaling / normalization, correlation analysis, VIF analysis, PCA analysis**

A. Perform EDA on the data file **e5-Run2-June22-subset-100-cols.csv** to understand the overall nature / quality of the data.

B. Process each of the columns to resolve each of the following matters, and implement the solutions:

   a. Do you want to keep or discard the column? What is your basis for these decisions?

   b. Are there any outliers / missing values / wrong values in the data? If so, how will you fix them? Fix them!

   c. Do the columns need to be standardized / normalized? If so, do it!

   d. Which of the columns are correlated? Create a correlation heat map and state your observations.

   e. You must deal with the correlated columns. Which columns will you remove from the consideration of further analysis? (as you may not have enough information about the domain, take appropriate calls in case of conflicts)

   f. Which of the columns have a **multi-collinearity** relationship with other columns? Perform **VIF** analysis to understand this aspect and document your observations.

   g. Which of these columns will you discard? Decide, by stating the applicable reasons, and discard those columns.

   h. Perform PCA on the data at the two stages i and ii mentioned below, and analyze the results.

      i. After step ‘c’

ii. After step ‘g’

iii. Do not forget to create and understand the **elbow** diagram in the context of PCA.

iv. Interpret the diagram and decide how many principal axes can adequately describe the dataset.

C. For each of the above tasks, describe the steps, and clearly show your final results (visually + using statistical measures). Adequately justify all your decisions!

---

**Problem 3: PCA and t-SNE**

A. Subject the file **mnist_test.csv** to **PCA analysis**

B. Create the elbow diagram and make your conclusions

C. Create the scatter plot PC2 v/s PC1 and interpret the results

D. Subject the file **mnist_test.csv** to **t-SNE analysis**, to map the data to 2 dimensions

E. Visualize the mapped data by creating a scatter plot of these 2 dimensions, and interpret the results

F. Subject the file **e5-Run2-June22-subset-100-cols.csv** also to t-SNE analysis and visualize and analyze the results. Is any important aspect of the data emerging from this exercise?

G. Document all your steps, outputs, results, analysis and interpretations

---

**Problem 4:**

List your major learnings from this exercise.

---

Your submissions should be as follows: A single ZIP file containing the following:

- A PDF document with all the above analyses and comments. Ensure that you include the required figures / Tables / Metrics in your report, along with the explanations and analysis.
- Python source files (.py files), one for each Problem, as part of your submission. Please DO NOT upload Jupyter Notebooks – they get bulky!
- Name of the PDF should be **E5-your-roll-number.pdf** and the name of the Python source file should be **E5-Problem-X-your-roll-number.py** (replace **X** with the problem number)
- Upload the PDF and the source files to the assignment submission point E5.

**oooOOOooo**