# CP 318 –Data Science for Smart City Applications

# Project 2: Exploratory Data Analysis:

**Deadline: 10th November 2023, 11:59 PM IST (Hard deadline)**

This project can be completed individually or in a team of up to three people. If working in a team (even if the team is same as Project 1 team) , please inform us of your group members in Project 2 Channel of Teams **by October 23rd, 2023, 11:59 PM. Note that: No more than two students can be from the same department in a team.**

### Description:

This project requires you to select your own dataset from India related to smart city domains and conduct a thorough exploratory data analysis based on exploratory data mining techniques (such as data pre-processing, clustering, dimension reduction, etc.) covered in the course.

You are required to do at least 4 types of analysis e.g.,

1. Data-processing (see Lecture 2 slides for reference, you can use any other data pre-processing techniques which were not taught in the class) [Project weightage =10%]
2. Descriptive Analysis: Generate summary statistics, histograms, or box plots to understand the data distributions. You can also Create visualizations (e.g., scatter plots, bar charts, heatmaps) to identify patterns and correlation. [Project weightage =10%]
3. Clustering Analysis: Apply any two types of clustering algorithms (out of partitioning, hierarchical, density-based, and distribution-based) which are taught in the class, evaluate your results, choose appropriate number of clusters, and interpret the findings. Also, provide their performance comparisons (based on time, clustering accuracy etc.) [Project weightage =30%]
4. Dimension reduction: Use a linear and a non-linear dimensional reduction technique to reduce the dataset's dimensionality, visualize the reduced data in 2D or 3D, and discuss how dimensional reductions helps or affects your analysis. [Project weightage =30%]

## Data sets

In your analysis, you are tasked with selecting a dataset of Indian origin, specifically related to smart city domains encompassing areas such as agriculture, energy, environment (water, air, noise, etc.), healthcare, transportation, parking, education, manufacturing, infrastructure, and more. If you have any doubts regarding whether your chosen dataset fits the criteria for smart city datasets, please don't hesitate to seek guidance from the course instructor.

We strongly encourage you to opt for a dataset that has not undergone extensive analysis. We encourage you to use a dataset for which there is no readily available analysis on the internet. However, if you decide to work with a dataset that has been previously explored in online case studies, you must provide the links to these prior studies and clarify how your analysis distinguishes itself from them. As part of your submission, you are required to provide the dataset source and the entire dataset you used for your analysis.

Ideally, your data set should have more than 1000 observations but less than 1 million, depending on the model and structure of the data. If you have an interesting big data set, you may use a smaller subset of the data to ensure feasible computation times. If you're looking for inspiration or you're not sure where to begin, take a browse over this list of datasets arranged by topic. You can also find datasets from Kaggle datasets or UCI ML Repository. If you do not find Indian datasets from these sources, you can also refer to smart city portal of Govt. of India or DataSmart Cities

Ideally, we aim to ensure that no more than one team works on the same dataset with the identical problem statement. To prevent such conflicts, it is mandatory for each team to publicly disclose their selected dataset and its source in the Project 1 channel. This transparency allows other teams to avoid selecting the same data for their analysis. Thus, we strongly recommend initiating the process of finalizing your dataset early on, granting you the advantage of working with your preferred data while adhering to the project specifications.

In cases where multiple teams intend to work on the same dataset (in such instances, you must obtain permission from the Instructor before commencing data analysis, with appropriate reasons to do so), it is essential that their problem statements, analyses, and approaches substantially differ from one another. This ensures a diversity of perspectives and approaches even when working with the same dataset.

## Submissions:

1. **Code and Data in zip file**. If publicly available data was used, please share the link in code. Also, provide the data in a zipped file.
2. Report in **PDF** format.

The report should be concise and not exceed 5 pages, covering the following topics:

1. Introduction
   - the motivation
   - the problem statement
2. Description of the dataset and the analysis problem. Provide information where the data was obtained, and if it has been previously used in some online case study and how your analysis differs from the existing analyses.
3. Description of all four analysis mention in the project description
4. How the model was run, that is, what options/parameters were used, what similarity measures were chosen, motivation, etc.
5. Comparison of various models in clustering and dimension reductions.
6. Discussion on results with reasonings, highlight interesting patterns, non-obvious findings
7. Provide detailed explanations and interpretations of your analysis throughout the report.
8. Use appropriate visualizations to support your analysis and findings.
9. Discussion of issues/challenges in dataset/analysis and potential improvements.
10. Explain how your analysis could be useful to Indian government, policy makers or to the society.
11. Cite sources and references if you use external information or code.
12. Conclusion, insights gained from the data analysis
13. Self-reflection of what the you/group learned while making the project.

## Evaluation:

The report will be marks using the rubric in Table 1. Please follow the report template as mentioned in Project 1 Spec document.

| Data Analysis mentioned in Report | Report Writing |
|---|---|
| **15-18 marks:** Tried approaches are well motivated and their advantages/disadvantages clearly discussed ;thorough and insightful analysis of why the one approach works/not work for used data; insightful discussion and analysis (as mentioned in Project Report section) | **6-7 marks**: Very clear and accessible description of all that has been done, a postgraduate student can pick up the report and read with no difficulty |
| **11-14 marks:** Tried approach are reasonably motivated and their advantages/disadvantages somewhat discussed; good analysis of why the one approach works/not work for the used data; some discussion and analysis (as mentioned in Project Report section) | **4-5 marks:** Clear description for the most part, with some minor deficiencies/loose ends (e.g., there are no-table gaps and/or unclear sections |
| **6-10 marks:** Advantages/disadvantages discussed; limited analysis of why one approach works/not work for used data; limited discussion and analysis (as mentioned in Project Report section) | **2-3 marks:** Clear description for the most part, with some minor deficiencies/loose ends (e.g., there are no-table gaps and/or unclear sections |
| **0-5 marks:** Approaches ae barely or not motivated and their advantages/disadvantages are not discussed; no analysis of why one approach works/not work for the used data; little or no discussion and analysis (as mentioned in Project Report section). | **1 mark:** The report is unclear on the whole, omits all<br>key reference, and the reader can barely discern what has been done |