# CO7093 - Big Data & Predictive Analytics

CW Assignment

Regression & Clustering

## Assessment Information

| | |
|---|---|
| **Assessment Number** | 2 |
| **Contribution to overall mark** | 70% |
| **Submission Deadline** | 27/03/2023 at 6:00 pm |

## Assessed Learning Outcomes

This second assessment aims at testing your ability to

- carry out data cleansing and visualization
- develop a predictive model and evaluate its performance
- perform appropriate clustering for local regressions
- communicate your findings on the data

## How to submit

For this assignment, you need to submit the followings:

1. A short report (about 8 pages in pdf including all the graphs) on your findings in exploring the given dataset, a description of your model and its evaluation, a description of your clusters and its justification, local regressors based on your clusters as well as their evaluation).
2. The Python source code written in order to complete the tasks set in the paper. You should submit the Python code file, group1_solution.py or group1_solution.ipynb. Note that even if you decide to work on your own, you must enrol yourself into a group.
3. A signed coursework cover – this should include the names of all the students involved in the work submitted.

Please put your source code, report and signed coursework cover into a zip file CW2_GroupID.zip (e.g., CW2_Group1.zip) and then submit your assignment through the module's Blackboard site by the deadline. Note that to submit, you need to click on the Coursework link on Blackboard and then upload your zipped file. Remember it is **1 submission per group**!

## Problem Statement

Consider the Housing dataset that records sale prices of properties in Manhattan from August 2012 to August 2013. For your convenience, the dataset can be downloaded from Blackboard along with a glossary of terms used in the data. The data include sale prices, neighborhood, building type category, square footage, and other types of variables.

**Objective:**

Using the given dataset, we would like to build up a model that can predict the sale prices of houses in terms of some relevant features in the dataset and use a K-Means approach to propose a nontrivial set of houses' clusters, which may improve the performance of the regression model by proposing clusters-based (or local) regression models.

## Exploring the data

Your first task is to prepare the data and carry out data munging or cleansing, bearing in mind the question you would like to answer. For example, what is the impact of age, neighborhood, square footage, or other features on the house prices? Address the following questions:

## 1 Part 1 - Building up a basic predictive model

Load the dataset `Manhattan12.csv` into a pandas dataframe and carry out the following tasks.

Organise your code bearing in mind robustness and maintainability:

1. **Data cleaning and transformation:**

   If you have a closer look at the dataset, you will see that there are lots of missing values. They need be treated appropriately but in the first instance, we will take an aggressive approach to dealing with them.

   - Show the shape of the dataset
   - Rename incorrectly formatted column names (e.g. SALE\nPRICE)
   - Create list of categorical variables and another for the numerical variables
   - For each numerical column, remove the ',' the '$' for the sale price, and then convert them to numeric.
   - Convert the 'SALE DATE' to datetime.
   - For each categorical variable, remove the spaces, and then replace the empty string '' by NaN.
   - Replace the zeros in Prices, Land squares, etc. by NaN
   - Show a summary of all missing values as well as the summary statistics
   - Drop the columns 'BOROUGH', 'EASE-MENT', 'APARTMENT NUMBER'
   - Drop duplicates if any
   - Drop rows with NaN values
   - Identify and remove outliers if any

- Show the shape of the resulting dataframe.
- Consider the log of the prices and normalise the data.

2. **Data Exploration**. Consider the resulting dataframe. This first aggressive cleaning should give a smaller dataset, which you can start by exploring relationships between the various features of the dataset.

    - Visualise the prices across neighborhood
    - Visualise the prices over time
    - Show the scatter matrix plot and the correlation matrix
    - Any further plots, which demonstrate your understanding of the data

3. **Model building**. Consider the resulting dataframe.

    - Select the predictors that would have impact in predicting house prices.
    - Build up a first linear model with appropriate predictors and evaluate it. Split the data into a training and test sets; build up the model; and then show a histogram of the residuals. Evaluate your model by using a cross-validation procedure.

# 2 Part 2 - Improved model

This is an open-ended question and you are free to push your problem-solving skills in order to build up a useful model with higher performance.

1. Consider the entire datasets given in this assignment. Develop an improved predictive model that predicts the sales prices of houses. Make sure to validate your model. You should aim for a model with a higher performance while using a maximum of data points. This implies treating missing values differently for example through imputation rather than dropping them.

2. Use the K-Means algorithm to cluster your cleansed dataset and compare the obtained clusters with the distribution found in the data. Justify your clustering and visualise your clusters as appropriate.

3. Build up local regressors based on your clustering and discuss how this clusters-based regression compares to your regression model obtained in Part 2. 1.

## Marking Criteria

The following areas are assessed:

1. Cleansing, visualizing, and understanding the data **[30 marks]**
2. Building up and evaluating the predictive model **[20 marks]**
3. Clustering and evaluation of clusters-based regressors **[20 marks]**
4. Coding style **[10 marks]**
5. Writing the report (up to 8 pages) interpreting the results. **[20 marks]**

Indicative weights on the assessed learning outcomes are given above and can be found in the **marking rubric on Blackboard**. The following is a guide for the marking:

- **First++ (≥ 90 marks)**: As in **First+** plus a predictive model with excellent performance, excellent justification and visualisation of the clusters, great insights from the data, and a report of professional standards.

- **First+ (≥ 80 marks)**: As in **First** plus a comprehensive coverage of data cleansing techniques demonstrating an excellent understanding of the data, a sound comparison of the global predictive model against the clusters-based model and a well-structured, maintainable, and robust code usefully using functions.

- **First (≥ 70 marks)**: As in **Second Upper** plus a well-justified predictive model by the data cleansing with good performance using sound evaluation techniques; a well-justified clusters and a concise report on the results obtained from the dataset.

- **Second Upper (60 to 69 marks)**: A good coverage of data cleansing techniques exploring the dataset, a good visualisation of the clusters, a predictive model with an appreciable accuracy with a rationale behind it, a working code and a well-structured report on the results obtained from the dataset.

- **Second Lower (50 to 59 marks)**: Some techniques used for data cleansing are overlooked, a predictive model partially justified with an appreciable accuracy, a working clustering, a partially commented code with very few functions, and a narrative of the findings about the dataset with few deficiencies.

- **Third (40 to 49 marks)**: Essential data cleansing techniques are covered, a predictive model is given with some justification, a working but basic block code with no clustering, and a written report describing some of the work done.

- **Fail (≤ 39 marks)**: Not satisfy the pass criteria and will still get some marks in most cases.

- **None-submission**: A mark of 0 will be awarded.

## Marking Group Work

Normally, a group will be given the same mark unless some members made little or no contributions. Any group can be called for an interview during marking. All group members **must attend**, explain their contributions, and defend the work submitted.