CA01: Exploratory Data Analysis – House Price Analysis

Exploratory Data Analysis is an approach analyzing data sets to summarize their main characteristics such as mean, standard deviation, and count, so on, often with visual methods. It's where the researcher takes a bird's eye view of the data and tries to make some sense of it. It's often the first step in data analysis, implemented before any formal statistical techniques are applied.

House Price Dataset

As always, learning by doing is the best practice to understand deeper. So now, we are going to make our hands dirt by analyzing **House Price Dataset**.

Once the dataset is made "analytics ready", you can use this dataset for Hour Price Prediction using various algorithms later. But for now you first need to analyze the structure, domain, and contents of it thoroughly.

Dataset: Basic Info

What is it about?
This data provides values of various features that "might" be predictive of house prices. In other words, if a ML model is developed using this dataset, potentially such a model can be used to determine the "listing price" of a property that the owners are planning to sell.

What information it has?
The dataset comes with the following files. You can view them at:
github.com/ArinB/MSBA-CA-Data/CA01

- Data Description text file
- house-price-train.csv
- house-price-test.csv

However, for this assignment you will use only house-price-train.csv for the purpose of EDA and Data Cleaning. The goal is to make the training data ready for building a ML Model. The model building is not part of this assignment. If and when you would build a ML model to predict house prices, you will have to do the same EDA+Data Cleaning for the house-price-test.csv as well. But for this assignment, just focus on the house-price-train.csv

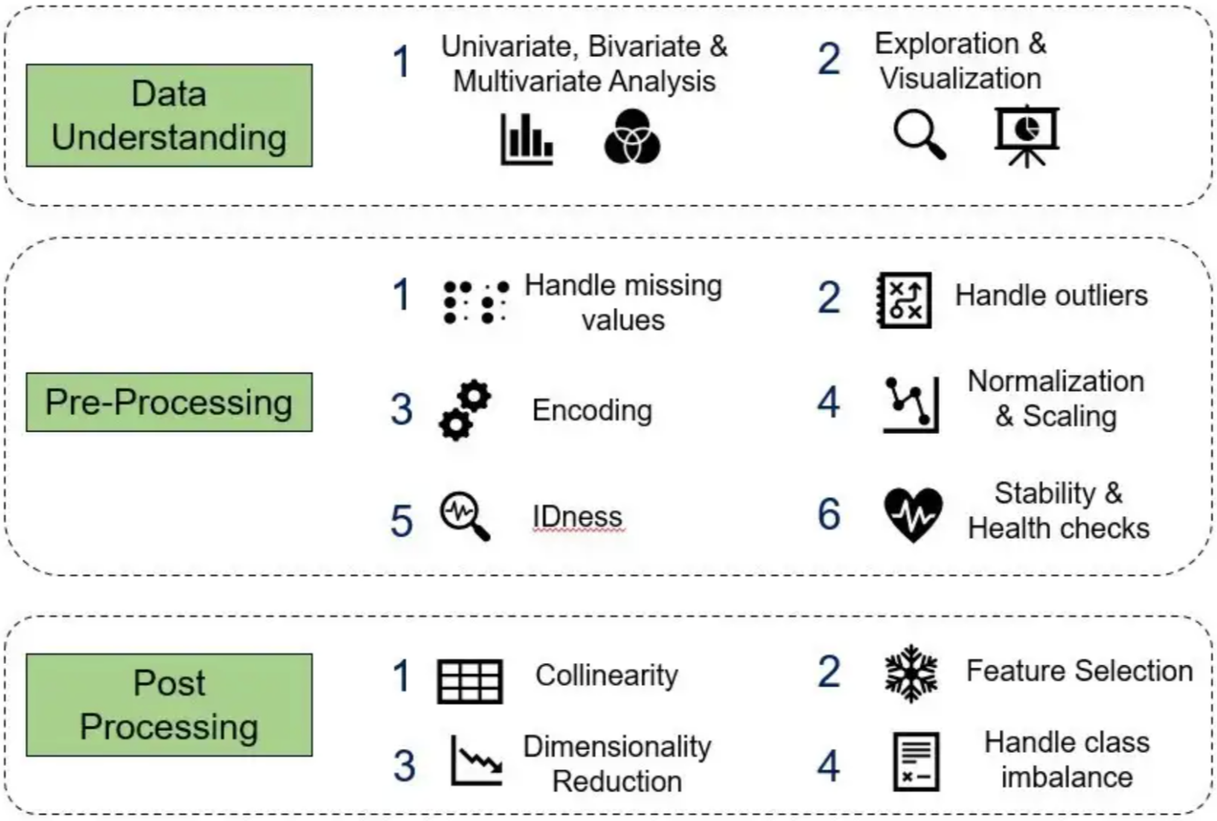
How can I access the data files in my code?

Copy the following path exactly and paste it in your pd.read_csv() line of the code.

https://github.com/ArinB/MSBA-CA-Data/raw/main/CA01/house-price-train.csv

What am supposed to do for this assignment?

As you have learned in during the lecture sessions, EDA has three Parts as in the picture below.



Data Understanding
1 Univariate, Bivariate & Multivariate Analysis
2 Exploration & Visualization

Pre-Processing
1 Handle missing values
2 Handle outliers
3 Encoding
4 Normalization & Scaling
5 IDness
6 Stability & Health checks

Post Processing
1 Collinearity
2 Feature Selection
3 Dimensionality Reduction
4 Handle class imbalance

Part 1: Data Understanding – Output of this part will be various visualization of the variable analysis and finally the Data Quality Report that will identify the data problems that exist in this dataset.

Part 2: In this part you will fix the data problems identified in the Data Quality Report learning from python techniques and code samples shared with you (links provided in the slide deck).

Part 3: For this part, ONLY do the “collinearity” visualization and identification of the features that needs to be dropped (feature selection), if any. You DO NOT NEED to do the “class imbalance” step, as this has not been covered yet in class.

All work, descriptions, explanation, observation, conclusion etc. must be in the Notebook itself. No separate report is needed. Your final deliverable is ONLY your notebook, which you will upload @BrightSpace Submission folder for CA01.