Mahindra University

**Machine Learning with Python (CS2202)**

Assignment

**Introduction:** In this assignment, you will explore techniques for handling missing values and detecting outliers in the Gurgaon real estate dataset. The dataset contains information about real estate properties, including various attributes such as property type, location, price, area, etc.

**Objectives:**

1. Identify missing values in the dataset and implement appropriate strategies for handling them.
2. Detect outliers in the dataset using statistical methods and visualization techniques.
3. Apply suitable techniques to handle outliers and ensure data quality for further analysis.

**Tasks:**

**Task 1: Data Exploration**

* Load the dataset into your preferred data analysis environment (e.g., Python).
* Identify the features (columns) in the dataset and their data types.
* Perform initial data exploration to understand the structure and content of the dataset (checking duplicate rows, if any, remove them to decrease the bias), explore the property\_type column (show bar plots), society column is categorical column with 676 categories, you need to remove the societies having very less number of flats or houses (use frequency bins) to reduce the cardinality, explore the price column (check missing values, descriptive statistics, plot histogram, box plot (for outliers), check the skewness and kurtosis to justify the outliers, use price binning), similarly do this for price\_per\_sqft and all the other columns you feel needs to be analyzed such as bathroom, bedroom, balcony etc and write your observations.
* Similarly do the multivariate analysis considering all the columns vs the target column (price) e.g. property\_type vs price (histograms and box plots), scatter plot between price and area
* Check for missing values in the dataset and determine their distribution across features.

**Task 2: Missing Values Handling**

* Choose appropriate strategies for handling missing values based on the characteristics of the dataset and the nature of missingness.
* Implement the selected strategies to handle missing values. This may include imputation, deletion, or other techniques.
* Evaluate the impact of missing values handling on the dataset's integrity and completeness.

**Task 3: Outlier Detection**

* Use statistical methods (e.g., Z-score, IQR) to detect outliers in numerical features (all) of the dataset.
* Visualize the distribution of numerical features using histograms, box plots, or scatter plots to identify potential outliers, use describe function to show the descriptive statistics.

**Task 4: Outlier Handling**

* Decide on appropriate strategies for handling outliers based on their impact on the analysis and the dataset's objectives.
* Implement outlier handling techniques such as winsorization, trimming, or transformation to mitigate the effects of outliers.
* Assess the effectiveness of outlier handling techniques in improving data quality and robustness.

**Task 5: Report and Analysis**

* Prepare a comprehensive report summarizing your findings and actions taken in handling missing values and outliers.
* Include visualizations, tables, and descriptive statistics to support your analysis.
* Discuss the implications of missing values and outliers on the dataset's analysis and potential strategies for future data cleaning and preprocessing.

**Submission Guidelines:**

* Submit your assignment report in a PDF format.
* Include code snippets, visualizations, and explanations to support your analysis and findings.
* Provide clear explanations of the techniques used and the rationale behind your decisions in handling missing values and outliers.

**Note:** Ensure that your analysis is well-documented, and your findings are presented clearly. Use appropriate visualization techniques to communicate your results effectively. Good luck with your assignment!

**Brownie points: Reward will be there for students attempting the feature selection and atleast one machine learning model to predict the price of property considering only the important features (Use Random forest feature importance, RFE, LASSO etc).**

Content of **GroupNumber\_CS2202** folder (Yellow highlights are folders; others are files). For OneDrive submission, zip the folder **GroupNumber\_CS2202**, then upload to

<https://mahindraecolecentrale-my.sharepoint.com/:f:/g/personal/neeraj_choudhary_mahindrauniversity_edu_in/EjLOGj4Q_RdMjTQqdW1uTbkB4yGR0r7XrM-JAhi-vjRpWA>

+--- **GroupNumber\_CS2202**

| +--- Code

| | +--- all\_codes\_R\_py\_Matlab\_here.py (Can be multiple files too)

| | +--- **GroupNumber** \_code.ipynb

| +--- Report

| | +--- **GroupNumber** \_report.pdf

| | +--- **GroupNumber**\_report\_source\_docx\_latex

| | +---**README** file stating the contribution of each group member in percentage terms

**This work is due 17th May 23:59 PM (firm deadline). (Friday)**