PES University, Bengaluru
(Established under Karnataka Act No. 16 of 2013)
Department of Computer Science & Engineering
UE24MA242A – Mathematics for Computer Science Engineers

Datathon - Set 6

House prices dataset

The Real Estate Price Prediction dataset contains detailed information about residential properties, including attributes such as transaction date, house age, distance to the nearest MRT station, number of convenience stores nearby, latitude, longitude, and unit house price. The dataset is designed for analyzing factors that influence real estate values, studying how location and amenities impact pricing, and building regression models for property valuation. With its mix of numeric variables for age, location, and price, and categorical-like variables such as transaction periods, it is ideal for practicing data preprocessing, exploratory data analysis, and predictive modeling in real estate.

Questions

Unit - 1

1. Manually classify each feature in the dataset by data type: nominal, ordinal, interval, or ratio. Provide a clear explanation for your choice for each feature.

2. For the numeric variables Y house price of unit area and X3 distance to the nearest MRT station:
- Calculate mean, median, mode, standard deviation, and range. Summarize in a table and interpret what these statistics indicate about the data distribution.

3. Identify missing values, non-numeric codes, or other inconsistencies in the dataset and outline the steps to clean the data.

4. Plot histograms and boxplots for Y house price of unit area and X3 distance to the nearest MRT station. Describe the distribution shape (normal, skewed, or multimodal) and identify outliers.

5. Remove outliers from Y house price of unit area using the IQR method or the z-score method—display before-and-after boxplots.

6. Generate a Q-Q plot for Y house price of unit area. Discuss whether the data is approximately normal.

7. Calculate the Pearson correlation between Y house price of unit area and other numeric variables (e.g., X2 house age, X3 distance to the nearest MRT station, X4 number of convenience stores, X5 latitude, X6 longitude), visualize with a heatmap, and identify the features with the strongest positive or negative correlation, explaining why.

8. Create a pairplot of Y house price of unit area, X3 distance to the nearest MRT station, and X2 house age. Describe any patterns in the data.

Unit - 2

1. Divide properties into "near-MRT" and "far-from-MRT" groups based on X3 distance to the nearest MRT station (use a reasonable threshold). Perform a t-test to compare the mean Y house price of unit area between the two groups. State null and alternative hypotheses, show the test statistic and p-value, and visualize group means. Interpret the results in terms of statistical significance.

2. Calculate the margin of error for the mean Y house price of unit area at a 95% confidence level. Interpret its meaning.

3. Construct a 95% confidence interval for the mean of Y house price of unit area.
   - Interpret the result: What does this interval imply about the possible average unit price of houses in the dataset?

Unit - 3

1. Perform hypothesis testing to evaluate whether the mean Y house price of unit area for properties with X4 number of convenience stores above a certain threshold (e.g., more than 5) differs significantly from a given benchmark value (choose a reasonable value from your data). Clearly state:

- Null and alternative hypotheses
- Significance level
- Interpretation of the result

2. Fit a linear regression model to predict Y house price of unit area using X3 distance to the nearest MRT station, X2 house age, and X4 number of convenience stores as predictors. Report fit metrics: R², MSE, and RMSE. Plot predicted vs actual Y house price of unit area and interpret the model's performance.

Deliverables

- Cleaned dataset summary

- Statistical tables and test results
- Plots (histograms, boxplots, Q–Q, heatmaps, pair plots, regression lines)
- Concise interpretations of all hypothesis tests and regression findings