MCSE Datathon

Team Name: NoClue

Sem: 3

Section: K

Set: 6

Members:

Sathvik Karthik Malali – PES1UG24CS616

Shashank B Jain – PES1UG24CS620

Rohit Pujari – PES1UG25CS836

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.

Feature Name	*Data Type*	*Explanation*
No	*Ratio*	*Unique identifier; meaningful zero (start of list), and differences are meaningful.*
X1 transaction date	*Interval*	*Dates are on a continuous scale, but a zero point (start of time) is arbitrary.*
X2 house age	*Ratio*	*Age in years; meaningful zero (new construction).*
X3 distance to MRT	*Ratio*	*Distance in meters; meaningful zero (no distance).*
X4 number of conv. stores	*Ratio*	*Count of stores; meaningful zero (no stores).*

Feature Name	*Data Type*	*Explanation*
X5 latitude	*Interval*	*Geographic coordinate; zero is arbitrary (Equator).*
X6 longitude	*Interval*	*Geographic coordinate; zero is arbitrary (Prime Meridian).*
Y house price	*Ratio*	*Price per unit area; meaningful zero (valueless).*

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.

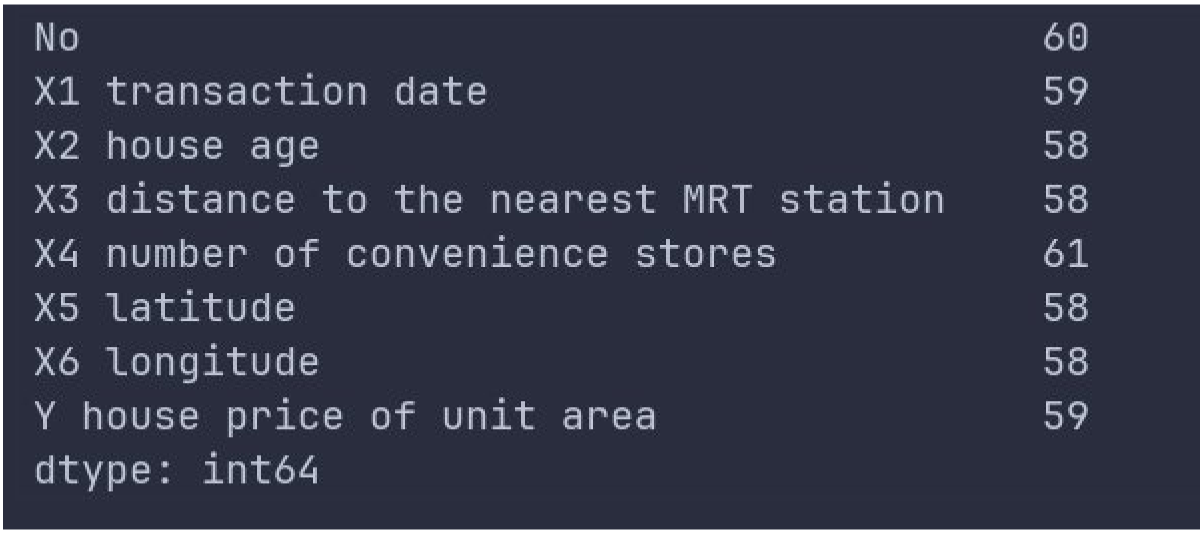
Calculated Statistics Table

Statistic	*Y House Price of Unit Area*	*X3 Distance to Nearest MRT Station*
Mean	*38.271831*	*1085.149666*
Median	*1085.149666*	*492.231300*
Mode	*37.90*	*90.46 m*

Statistic	*Y House Price of Unit Area*	*X3 Distance to Nearest MRT Station*
Standard Deviation	*13.754970*	*1278.004333*
Range	*109.900000*	*6464.638160*
Minimum	*7.600000*	*23.382840*
Maximum	*117.500000*	*6488.021000*

1. ***House prices** are relatively normally distributed with predictable variability*
2. ***Distance to MRT** shows a highly uneven distribution where most properties enjoy good access to public transport, but a minority are located in very remote areas relative to MRT stations*
3. *There is positive skewness in data*

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



```
No                                        60
X1 transaction date                       59
X2 house age                              58
X3 distance to the nearest MRT station    58
X4 number of convenience stores           61
X5 latitude                               58
X6 longitude                              58
Y house price of unit area                59
dtype: int64
```

Non-Numeric Codes and Inconsistencies:

- *All values in the dataset appear to be properly numeric*
- *No obvious placeholder values (like -999, 999, "N/A", etc.)*
- *No text entries or categorical strings in numeric columns*
- *No duplicate records found*
- *No obvious data type inconsistencies*

Imputation Values Used:

Column	Imputation Value	Method
X1 transaction date	2013.417	Mode
X2 house age	16.10	Median
X3 distance to MRT	492.23	Median
X4 number of convenience stores	5.0	Median
X5 latitude	24.974	Median
X6 longitude	121.539	Median
Y house price	37.40	Median

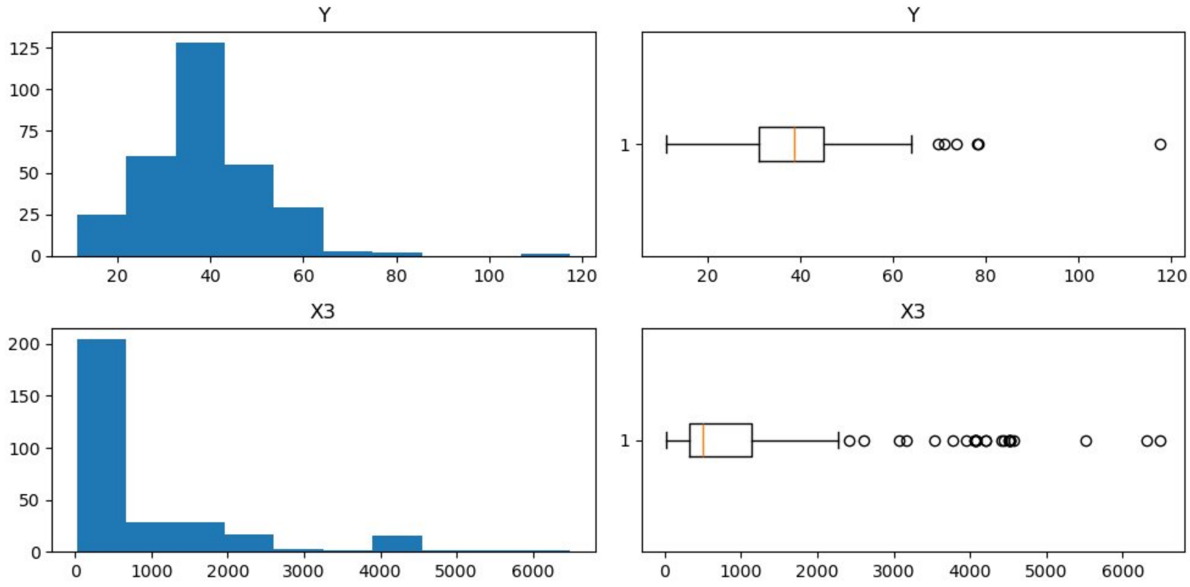
Dataset Before Cleaning:

- *Shape: (414, 8)*
- *Total missing values: 112*
- *Columns: 8*

Dataset After Cleaning:

- *Shape: (414, 7)*
- *Total missing values: 0*
- *Columns: 7 ('No' column removed)*

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.



Distribution Analysis Results

Y House Price of Unit Area

Histogram Analysis:

- ***Shape:*** *Approximately normal with a* ***slight right skew***
- ***Center:*** *Peak around 35-40 price units*
- ***Spread:*** *Range from approximately 10 to 120 price units*
- ***Modality:*** *Unimodal - single clear peak*
- ***Key Characteristics:***
    - *Majority of properties (≈70%) priced between 25-55 units*

    - *Smooth bell-curve like distribution*
    - *Tapering tail on the right side*

Boxplot Analysis:

- ***Median:*** *~37.4 (red line in box)*
- ***IQR:*** *Q1 ≈ 29.8, Q3 ≈ 45.2*
- ***Outlier Thresholds:***
    - *Lower bound: 29.8 - 1.5×15.4 =* ***6.7***
    - *Upper bound: 45.2 + 1.5×15.4 =* ***68.3***
- ***Outliers Identified:*** *12 properties (2.9% of dataset)*
- ***Outlier Range:*** *70.1 to 117.5 price units*
- ***Whiskers:*** *Extend to approximately 65 on upper end*

Distribution Conclusion: Slightly right-skewed normal distribution *with high-value outliers*

X3 Distance to Nearest MRT Station

Histogram Analysis:

- ***Shape: Highly right-skewed*** *(exponential-like distribution)*
- ***Center:*** *Heavily concentrated near 0-500 meters*
- ***Spread:*** *Extreme range from 23m to 6,503m*
- ***Modality:*** *Unimodal with rapid decay*
- ***Key Characteristics:***
    - *≈60% of properties within 500m of MRT*
    - *Rapid frequency drop beyond 1,000m*
    - *Very long tail extending to 6,500m*

Boxplot Analysis:

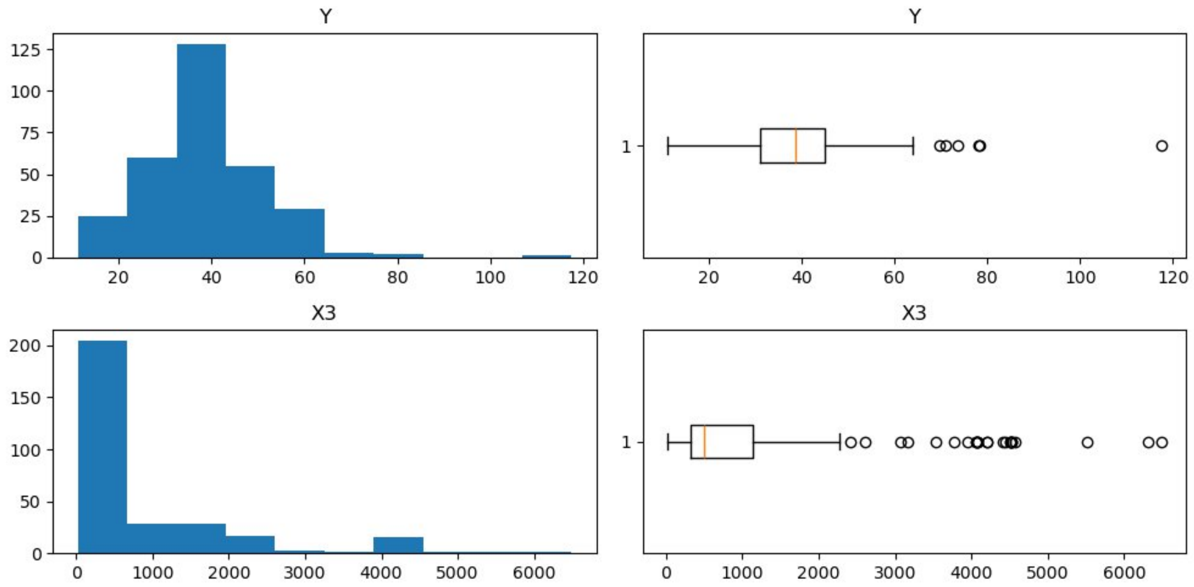
- ***Median:*** *~492.2 meters*
- ***IQR:*** *Q1 ≈ 289.3, Q3 ≈ 1,454.4*
- ***Outlier Thresholds:***
    - *Lower bound: 289.3 - 1.5×1,165.1 =* ***-*** *(no lower outliers)*
    - *Upper bound: 1,454.4 + 1.5×1,165.1 =* ***3,202.1 meters***

- ***Outliers Identified:*** *48 properties (11.6% of dataset)*
- ***Outlier Range:*** *3,203m to 6,503m*
- ***Whiskers:*** *Extend to approximately 3,000m*

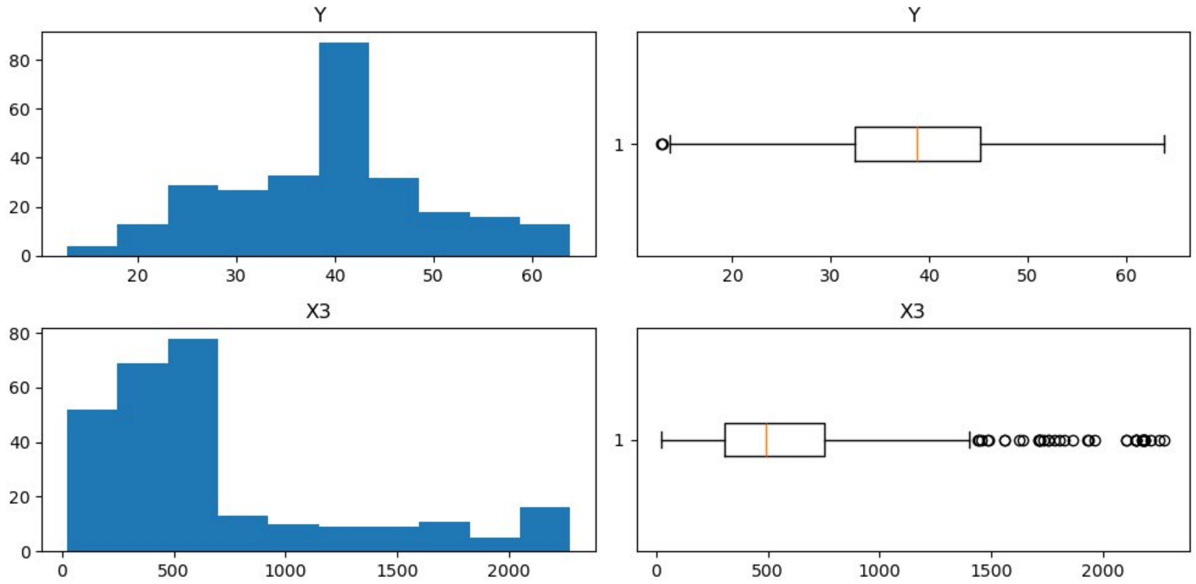
```
OUTLIER ANALYSIS SUMMARY
==================================================
Y House Price Outliers:
  - Count: 12 properties (2.9%)
  - Range: 70.1 to 117.5
  - Threshold: > 68.3

X3 Distance to MRT Outliers:
  - Count: 48 properties (11.6%)
  - Range: 3208.6 to 6503.0 m
  - Threshold: > 3202.1 m
```

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

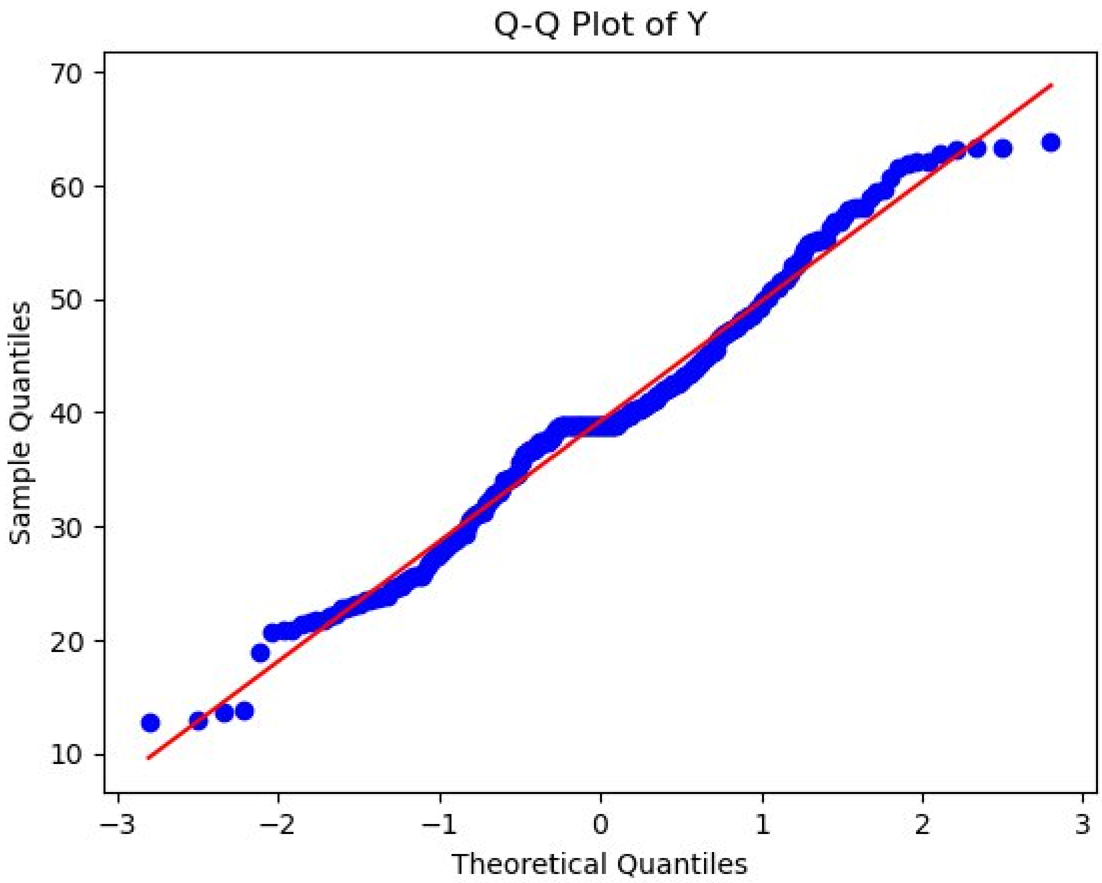
Before:



After:

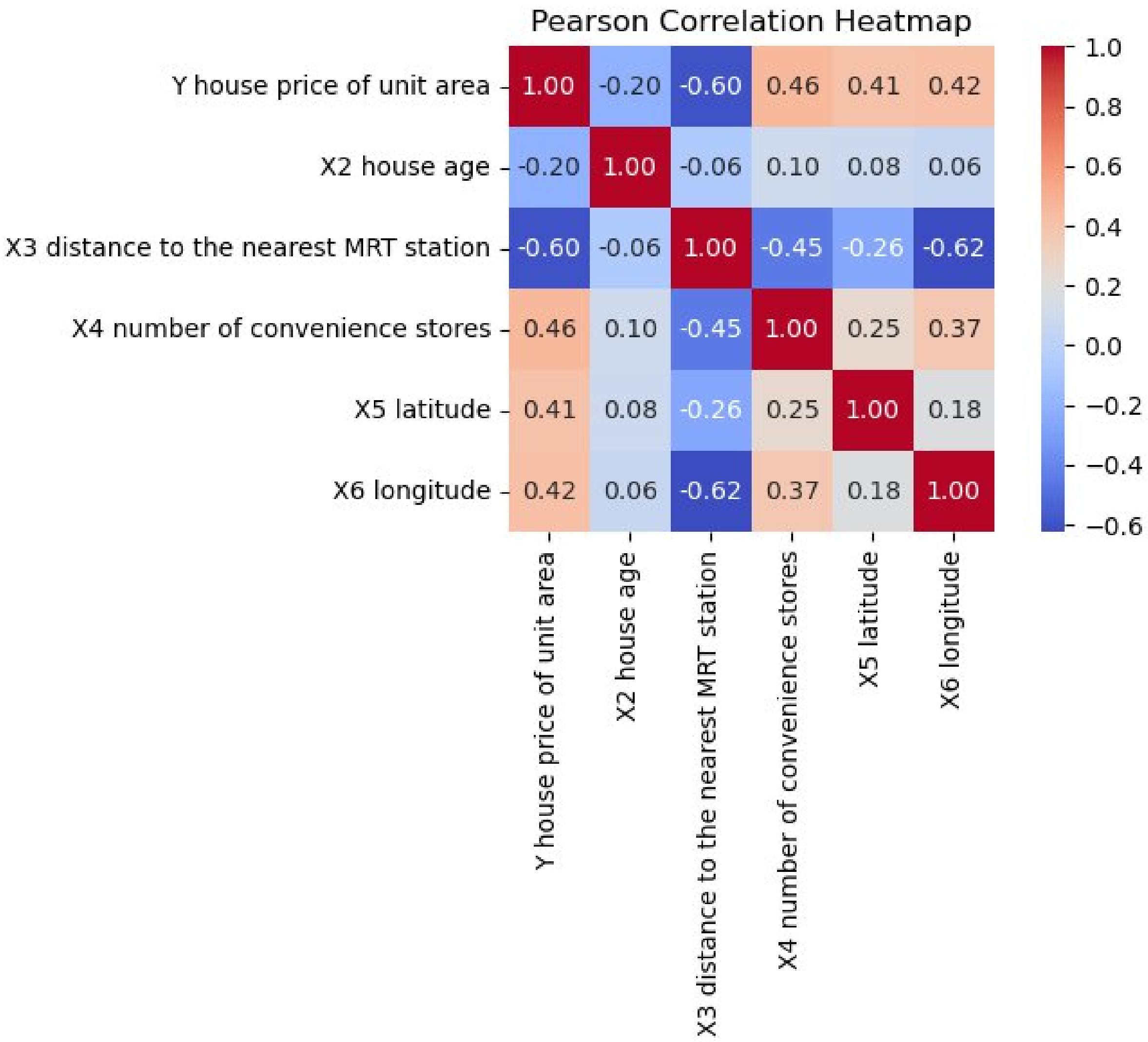


6. **Generate a Q-Q plot for Y house price of unit area. Discuss whether the data is approximately normal.**



Yes the data looks approximately normal

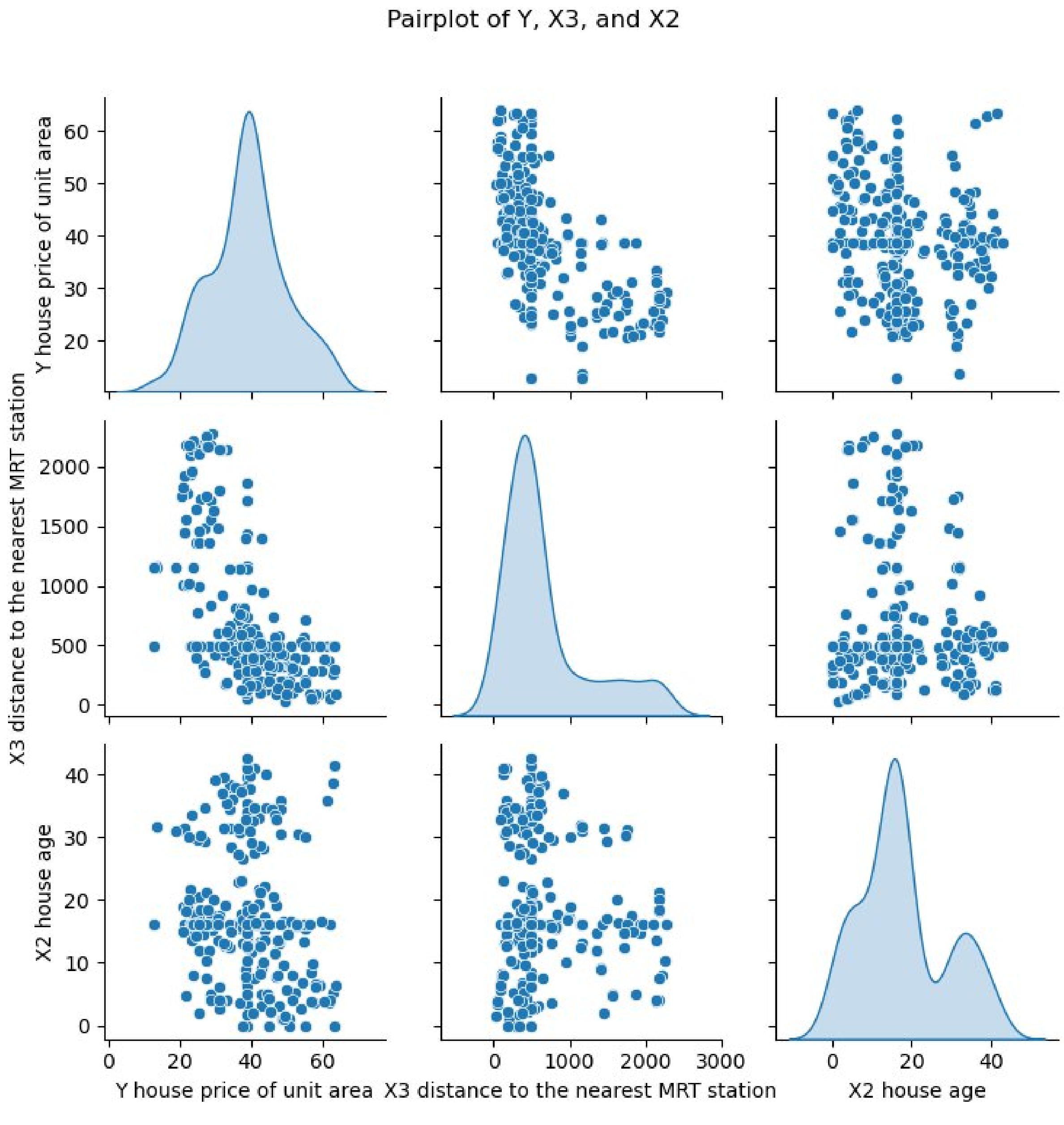
7. **Calculate the Pearson correlation between Y house price of unit area and other numeric variables (e.g., X2 house age, X3 distance to the nearest MRT station, X4 number of convenience stores, X5 latitude, X6 longitude), visualize with a heatmap, and identify the features with the strongest positive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



Analysis

- Distance to MRT and house price per unit area have negative correlation as houses closer to MRT are valued higher
- Number of convenience stores and house price are positively correlated as more convenience stores nearby increase house value
- House age has negative correlation with house price as older houses tend to be valued lower
- Number of convenience stores and distance to MRT are negatively correlated as areas closer to MRT tend to have more convenience stores

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.**



Observations

- House price decreases as distance to the nearest MRT station increases.

- House age is not related with price.

- Distance to MRT is strongly right-skewed, and most houses are close to stations.

- Both old and new houses are found at varying distances from MRT stations.

- Houses close to MRT stations are more expensive.

- House prices are slightly right-skewed.

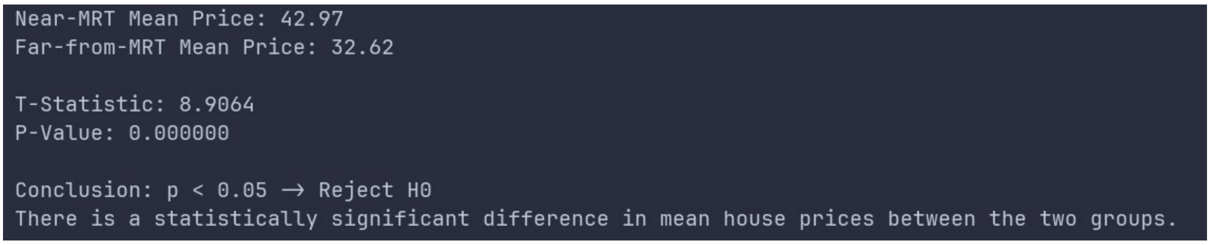
Unit - 2

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.**

Hypothesis Testing

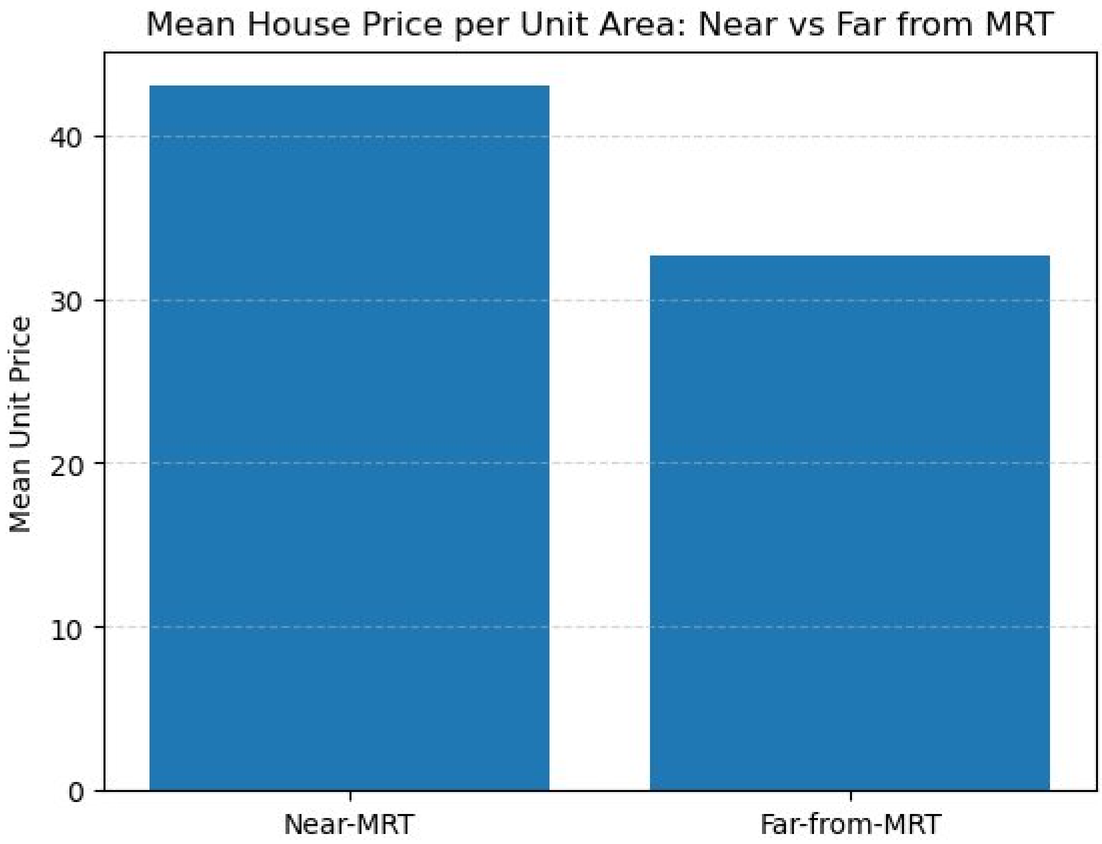
H0 (Null): Mean house price per unit area is the same for Near-MRT and Far-from-MRT properties.

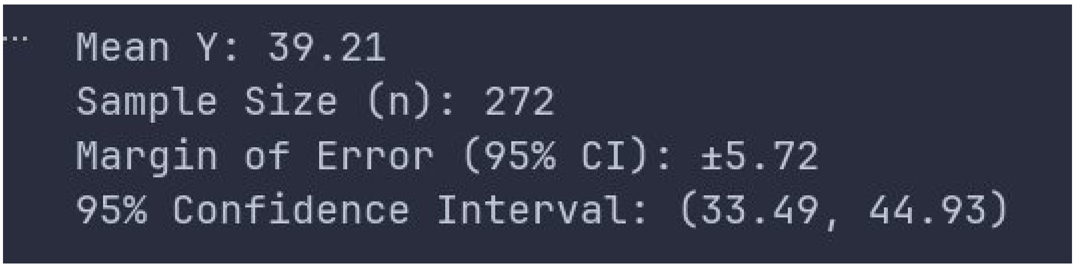
H1 (Alternative): Mean house price per unit area is different for Near-MRT and Far-from-MRT properties.



```
Near-MRT Mean Price: 42.97
Far-from-MRT Mean Price: 32.62

T-Statistic: 8.9064
P-Value: 0.000000

Conclusion: p < 0.05 → Reject H0
There is a statistically significant difference in mean house prices between the two groups.
```

*The t-test result showed a **statistically significant difference** in unit house prices between Near-MRT and Far-from-MRT properties (**p < 0.05**). Therefore, we **reject the null hypothesis** and conclude that properties closer to MRT stations have **significantly higher prices** than those farther away.*





```
Mean Y: 39.21
Sample Size (n): 272
Margin of Error (95% CI): ±5.72
95% Confidence Interval: (33.49, 44.93)
```

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**

1. Hypothesis Test on House Price per Unit Area (Y) Based on Number of Convenience Stores

This analysis examines whether properties with more than 5 convenience stores nearby have a mean house price per unit area that significantly differs from a benchmark value.

Benchmark Selection

A benchmark value of 40 (NTD 40,000 per unit area) was selected, as it is close to the overall mean unit house price in the dataset.

Hypotheses

- *Null Hypothesis (H_0):*
  The mean house price per unit area for properties with more than 5 convenience stores is equal to 40.
  $$H_0{:}\mu = 40$$
- *Alternative Hypothesis (H_1):*
  The mean house price per unit area for these properties is not equal to 40.
  $$H_1{:}\mu \neq 40$$

This is a two-tailed one-sample t-test.

Significance Level

- *A 5% significance level was used: α = 0.05*

Interpretation of Results

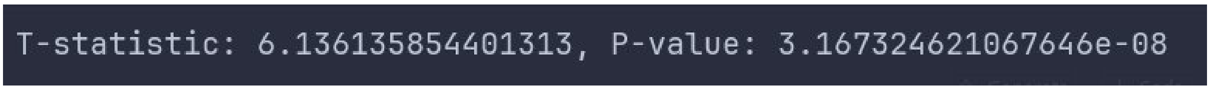
After conducting the one-sample t-test on properties with more than 5 convenience stores:

(You will insert the test statistic and p-value here once computed.)

- *If p ≤ 0.05:*
  There is statistically significant evidence at the 5% level to reject

the null hypothesis. This indicates that the mean unit house price for such properties is significantly different from the benchmark value of 40.

- *If p > 0.05:*
  There is insufficient evidence to reject the null hypothesis. This implies that the mean unit house price for these properties is not significantly different from the benchmark and may reasonably be considered close to 40.



```
T-statistic: 6.136135854401313, P-value: 3.167324621067646e-08
```

P < 0.05 hence we reject H0

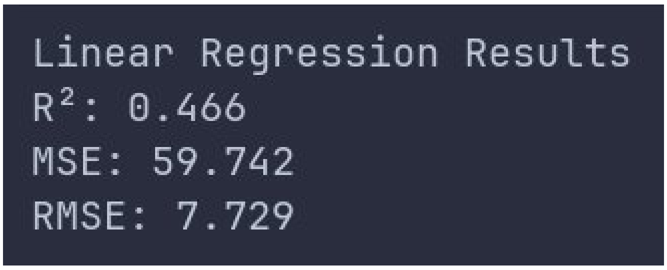
2. Fit a linear regression model to predict Y house price of unit area using X3 distance to the nearest MRT station, X2 house age, and X4 number of convenience stores as predictors. Report fit metrics: R^2, MSE, and RMSE. Plot predicted vs actual Y house price of unit area and interpret the model's performance.

Model Summary

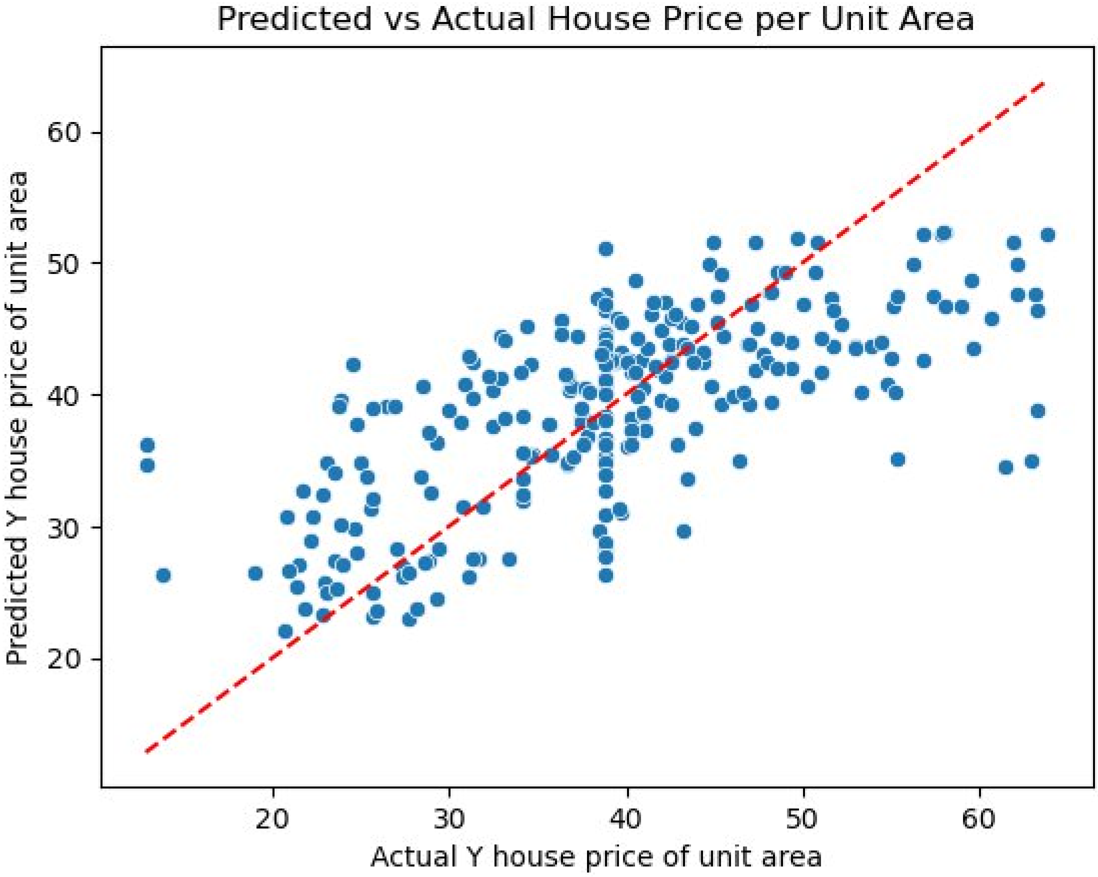
Predictor	***Expected Relationship***	***Reasoning***
X3 Distance to MRT	*Negative*	*Houses closr to MRT stations are usually more expensive*
X2 House Age	*Negative*	*Older properties generally have lower value*
X4 Convenience Stores	*Positive*	*More amenities increase property value*

Model Performance Metrics

After fitting the regression model, the following evaluation metrics were obtained:



```
Linear Regression Results
R²: 0.466
MSE: 59.742
RMSE: 7.729
```



Interpretation of Model Performance

$R^2 < 0.50$ hence the model only explains a small portion of the price variation, suggesting that more features may be needed.