1005

Real Estate Valuation

Team- Falcon

- Mahit Alva
- Rani Dynna Parida
- Anushka Mukherjee

Guide’s Name

Nikhil Maurya

GOAL

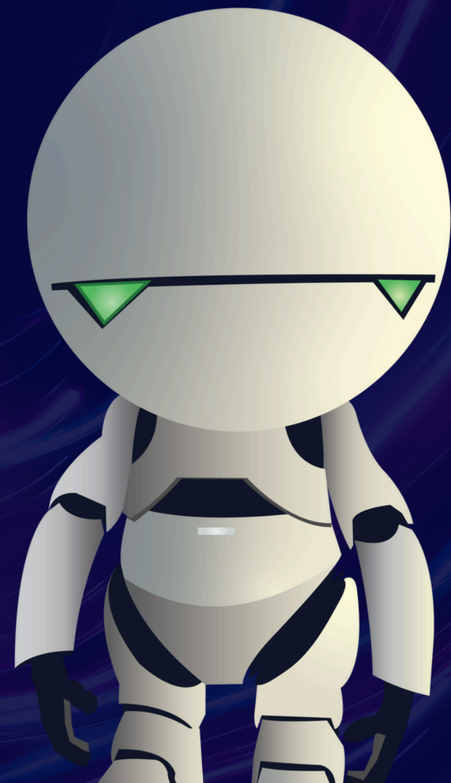
1. Informed Investment Decisions: Analyzing historical market data enables investors to make educated decisions about property investments, identifying undervalued properties or emerging market trends.

2. Accurate Pricing Models: Developing robust valuation models helps real estate professionals set fair market prices, benefiting both sellers and buyers by ensuring transactions are based on accurate assessments.

3. Neighborhood Analysis: Insights gained from the dataset can reveal valuable information about neighborhood dynamics, such as growth potential, amenities, and safety, influencing consumer preferences and investment strategies.

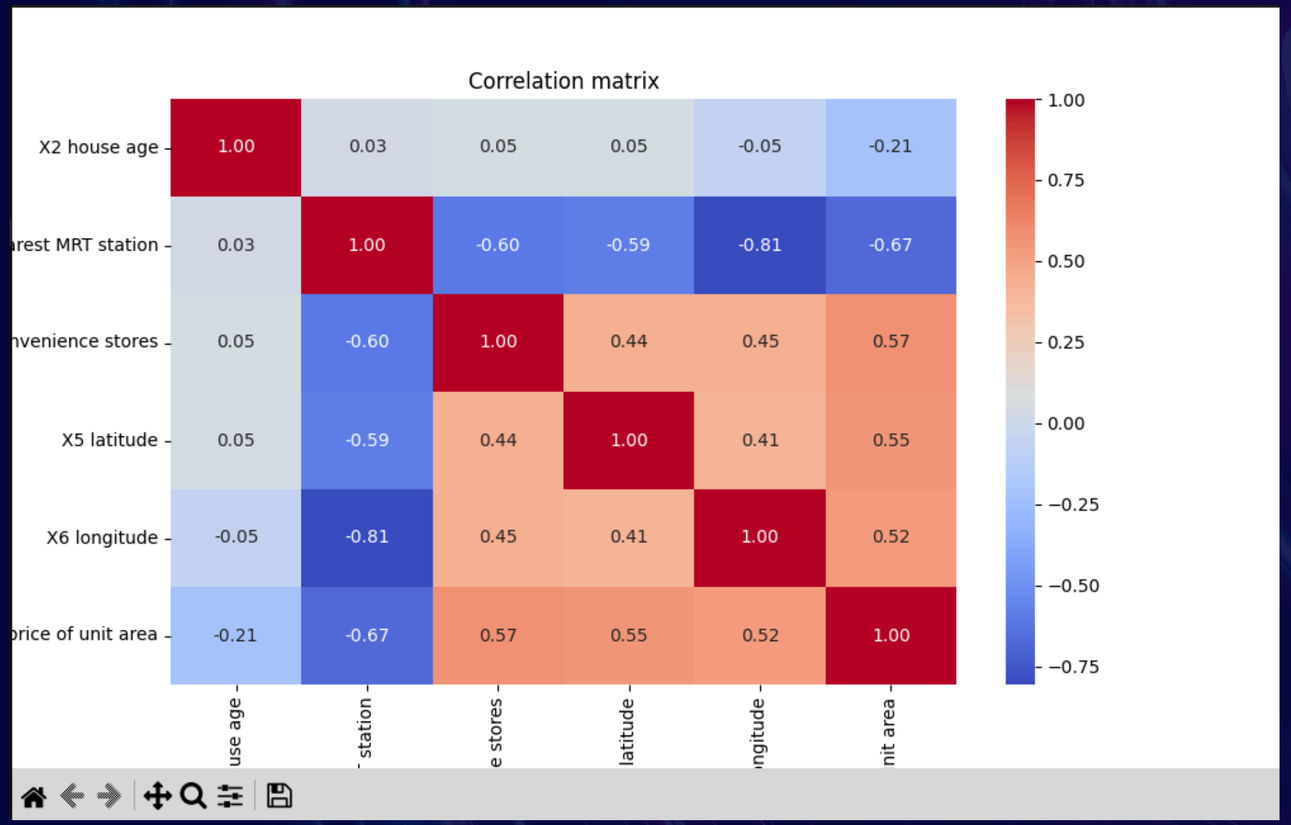
4. Policy Impact: Understanding valuation trends and factors influencing real estate prices can aid policymakers in making informed decisions regarding zoning, housing policies, and community development.

5. Market Forecasting: Leveraging historical data allows stakeholders to predict future market trends, helping agents, developers, and investors anticipate changes in property values and adapt their strategies accordingly.



EXPLORATORY DATA ANALYSIS

BASIC INFORMATION



This correlation matrix visualizes relationships between several features from a dataset. Here are some key points:

-X2 house age and other features: There are weak correlations between house age and most other features, with the strongest negative correlation being with "price of unit area" (-0.21), indicating that older houses might be slightly cheaper.

-Nearest MRT station and other features: This feature has strong negative correlations with X5 longitude (-0.81) and X6 latitude (-0.59), suggesting that properties further away from the MRT station are likely to have specific geographic coordinates.

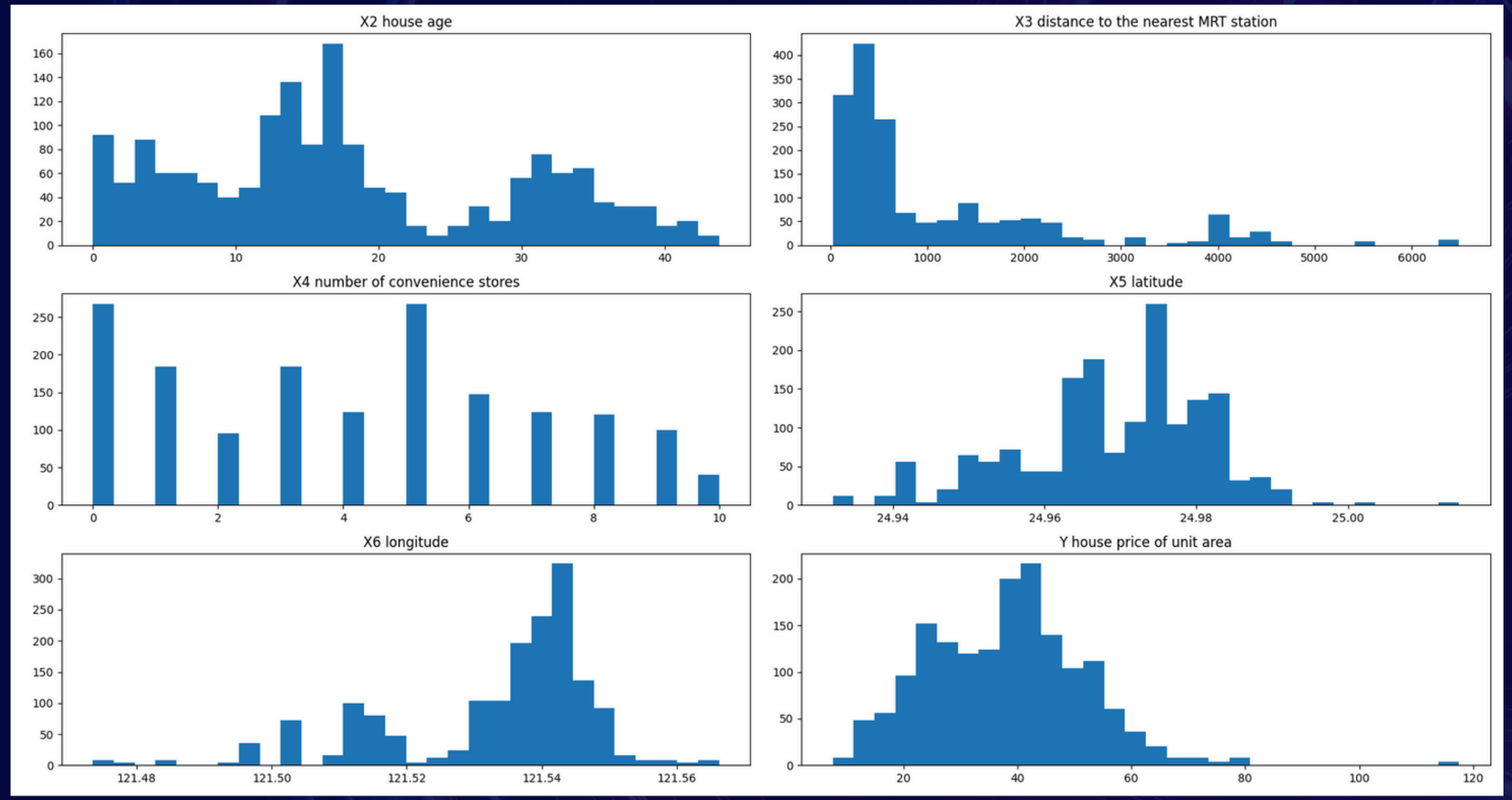
-Number of convenience stores and price of unit area (0.57): A moderately strong positive correlation suggests that areas with more convenience stores tend to have higher property prices.

-X5 latitude and X6 longitude: There's a positive correlation (0.41) between latitude and longitude, which likely reflects the geographic positioning of properties in the dataset.

-Price of unit area and other features: Price of unit area is moderately correlated with the number of convenience stores (0.57) and geographical coordinates (latitude: 0.55, longitude: 0.52), indicating that property prices are higher in areas with more convenience and certain geographic locations.

-MRT station distance, and geographic features are influencing the price of the unit area the most.

EXPLORATORY GRAPHS



The image shows several histograms representing the distribution of features in a dataset. Key takeaways include:

-X2 house age: The distribution is spread, with a peak around 15 to 20 years, indicating many houses fall within this age range.

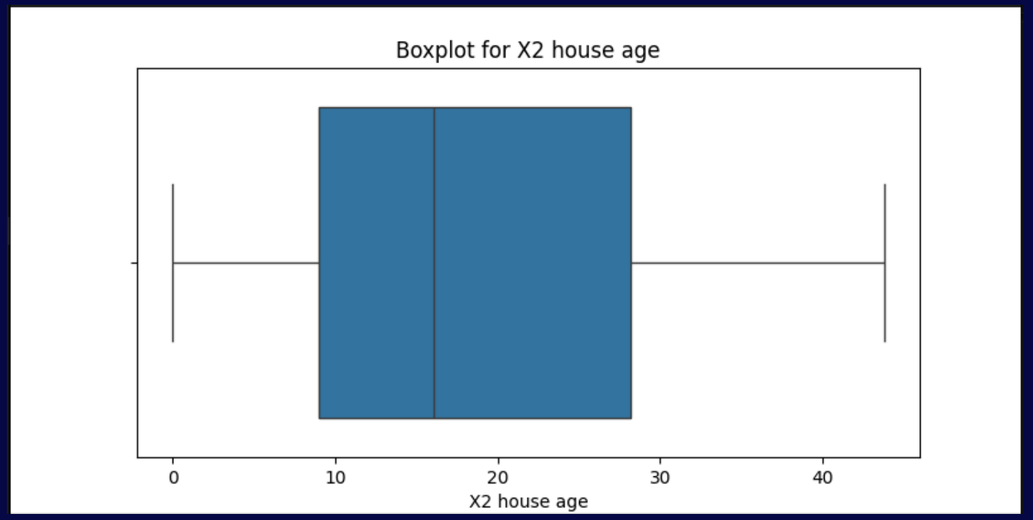
-X3 distance to the nearest MRT station: Skewed distribution with most properties located within 500 meters of an MRT station, while few are located farther away.

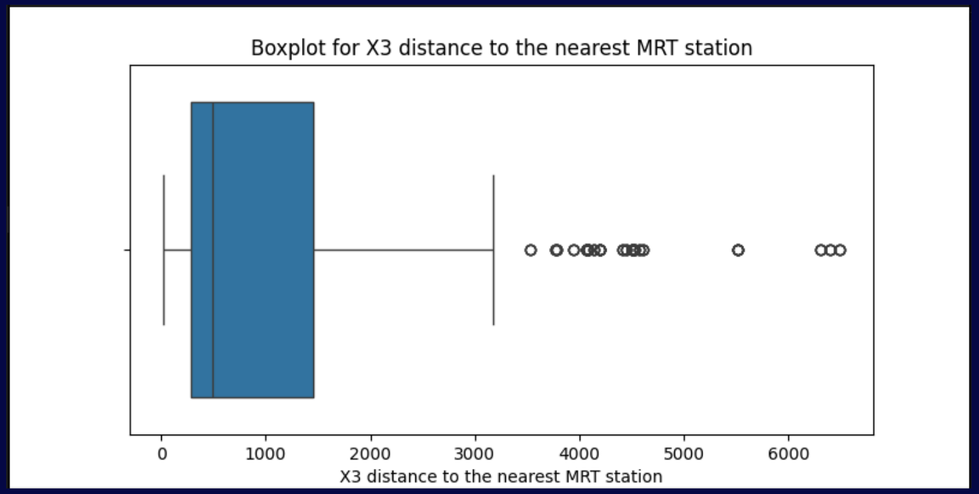
-X4 number of convenience stores: The number of stores mostly falls between 0 and 6, with some areas having as many as 10.

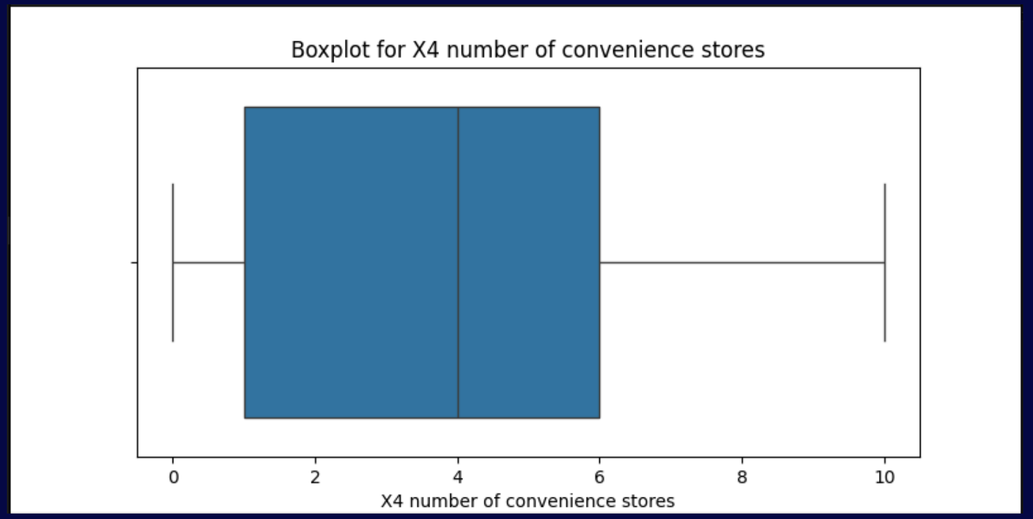
-X5 latitude: A relatively even distribution with a peak around 24.98, indicating properties are mainly clustered in a specific geographic band.

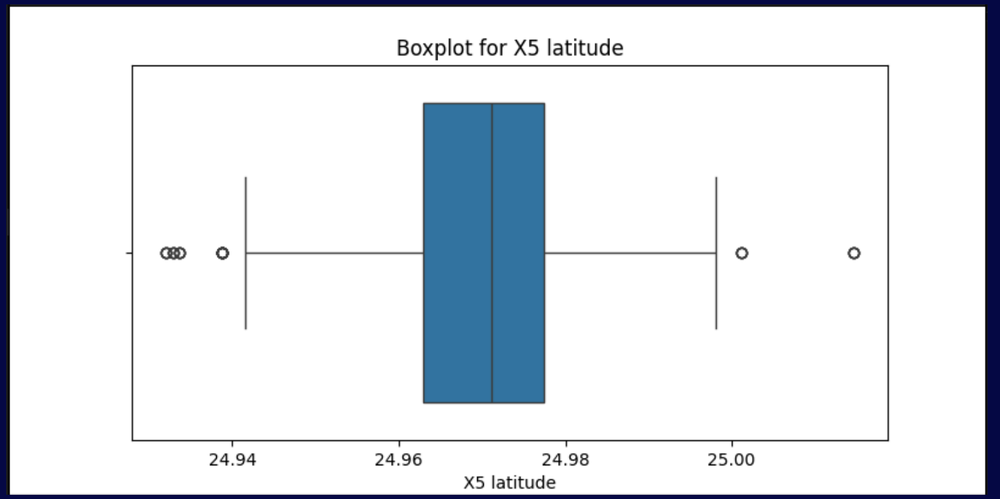
-X6 longitude: There's a peak between 121.52 and 121.54, showing most properties are concentrated in this range.

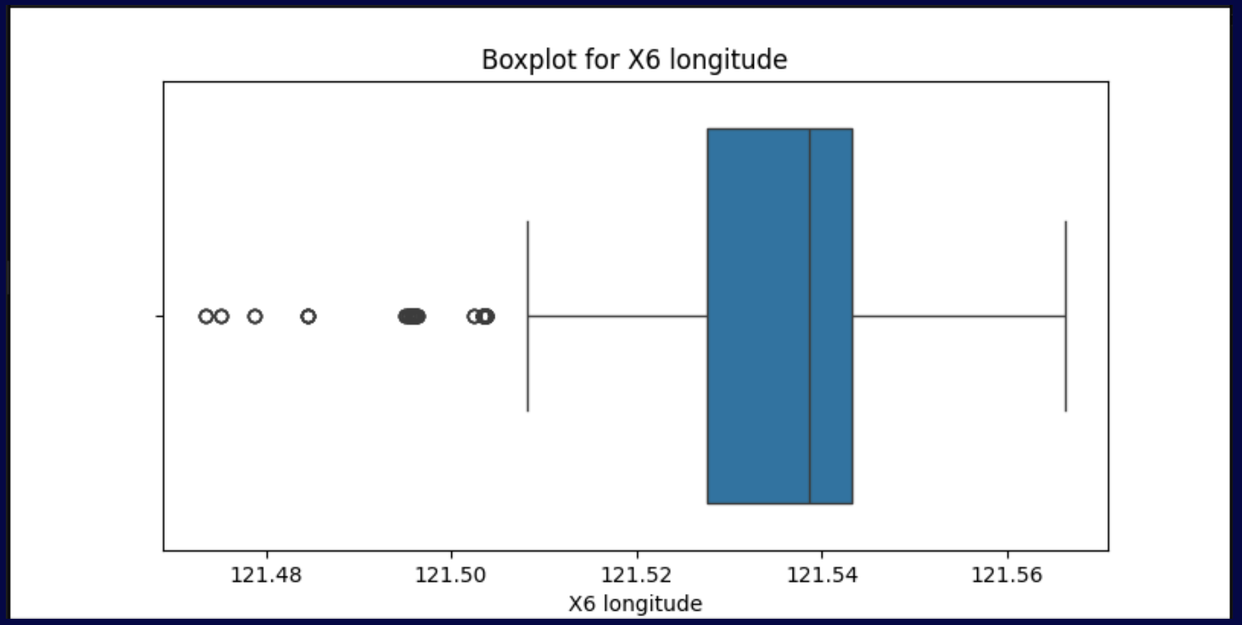
-Y house price of unit area: A normal distribution, with most property prices centered around 40 to 60 units, and fewer properties priced very high or low.

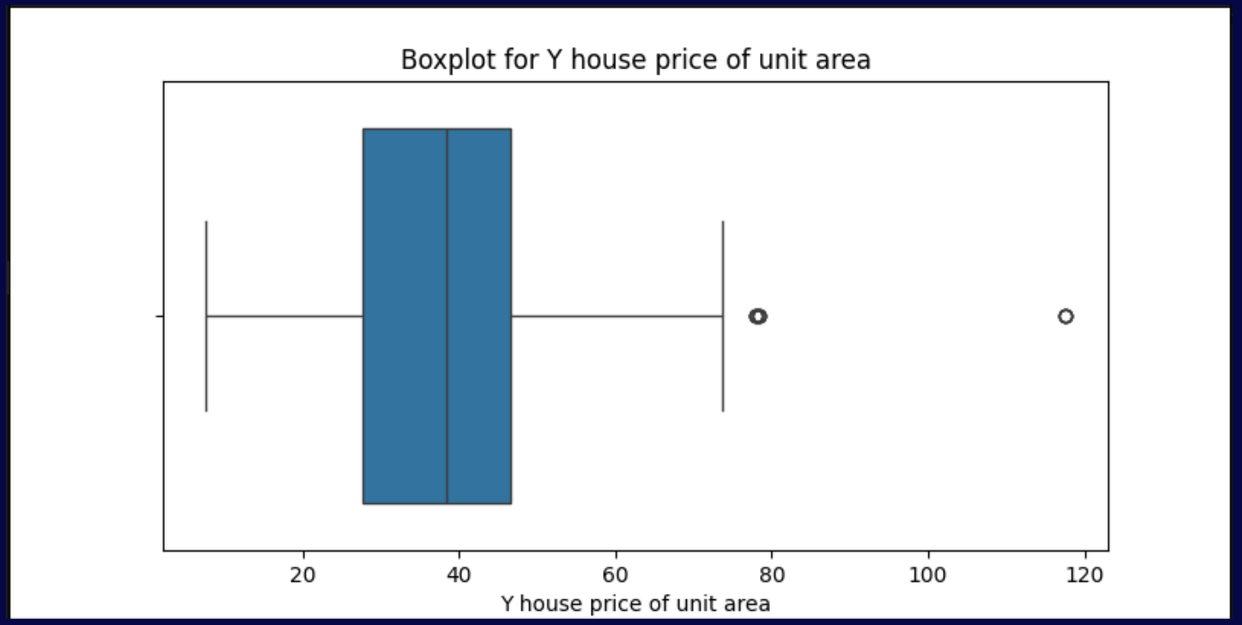

Boxplot for X2 house age
0
10
20
30
40
X2 house age


Boxplot for X3 distance to the nearest MRT station
0
1000
2000
3000
4000
5000
6000
X3 distance to the nearest MRT station


Boxplot for X4 number of convenience stores
0
2
4
6
8
10
X4 number of convenience stores


Boxplot for X5 latitude
24.94
24.96
24.98
25.00
X5 latitude


Boxplot for X6 longitude
121.48
121.50
121.52
121.54
121.56
X6 longitude


Boxplot for Y house price of unit area
20
40
60
80
100
120
Y house price of unit area

TRAINING OF MODEL

To create a well performing model we have tried to train the model on various classes present in scikit learn module of Python and XGBoost module of Python. The following models were used:

- Linear regressor
- KNN regressor
- Random forest regressor
- XGBoost regressor
- Voting regressor

PARAMETERS USED

1.Linear regression: Proceeded with default parameters

2.KNN regression: n_neighbors=2

3.Random forest regression: n_estimators=350, max_depth=20

4.XGB regression: n_estimators=500, learning_rate=0.01, max_depth=50, gamma=0.1

5.Voting regressor: estimators=[('knn', knn_model), ('rf', rf_model), ('xgb', xgb_model)],

n_jobs=-1,weights=[1,1,1]

TRAINING

```
Linear regression results on training data:
Mean absolute error: 0.45973340
Mean square error: 0.43114203
Root mean squared error: 0.65661406
R2 score: 0.56885797

Linear regression results on testing data:
Mean absolute error: 0.47889295
Mean square error: 0.50201410
Root mean squared error: 0.70852953
R2 score: 0.57606562
```

```
Voting regressor results on training data:
Mean absolute error: 0.06382860
Mean square error: 0.01316121
Root mean squared error: 0.11472233
R2 score: 0.98683879

Voting regressor results on testing data:
Mean absolute error: 0.07090721
Mean square error: 0.01296747
Root mean squared error: 0.11387481
R2 score: 0.98577000
```

```
KNN results on training data :
Mean absolute error: 0.03811718
Mean square error: 0.02076255
Root mean squared error: 0.14409216
R2 score: 0.97923745

KNN results on testing data:
Mean absolute error: 0.07869216
Mean square error: 0.08225454
Root mean squared error: 0.28680052
R2 score: 0.90833309
```

```
Random forest results on training data:
Mean absolute error: 0.04814250
Mean square error: 0.01136121
Root mean squared error: 0.10658896
R2 score: 0.98863879

Random forest results on testing data:
Mean absolute error: 0.07632463
Mean square error: 0.02044215
Root mean squared error: 0.14297606
R2 score: 0.97850986
```

```
XGBoost results on training data:
Mean absolute error: 0.11063254
Mean square error: 0.02252972
Root mean squared error: 0.15009902
R2 score: 0.97747028

XGBoost results on testing data:
Mean absolute error: 0.11627595
Mean square error: 0.02483971
Root mean squared error: 0.15760618
R2 score: 0.97278419
```

EVALUATION PROCESS

1. Mean Absolute Error (MAE): This measures the average absolute differences between predicted and actual values. It's easy to interpret since it's in the same units as the target variable, but it doesn't emphasize larger errors as much as other metrics.

2. Mean Squared Error (MSE): This calculates the average of the squared differences between predicted and actual values. Squaring the errors means MSE heavily penalizes larger mistakes, making it sensitive to outliers and useful for minimizing significant errors.

3. Root Mean Squared Error (RMSE): This is the square root of the MSE, which brings the error scale back to the same unit as the target variable, making it more interpretable. RMSE still penalizes larger errors but provides a balanced view of overall error magnitude.

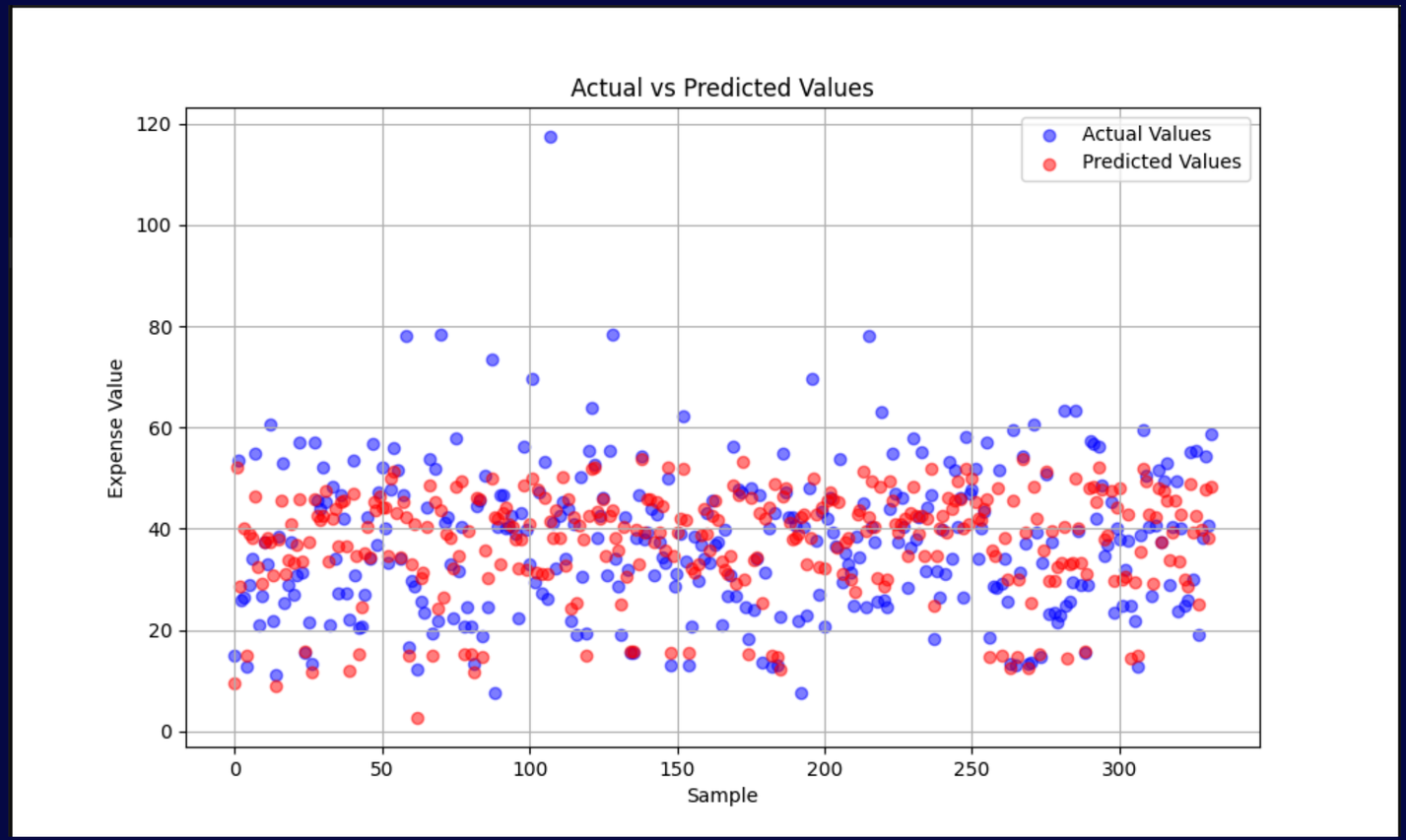
4. R^2 Score (Coefficient of Determination): This evaluates how well the model explains the variance in the target variable. An R^2 score of 1 indicates perfect predictions, while 0 means the model performs no better than using the average of the data.

Each of these metrics offers unique insights into model performance.

EVALUATING MODELS

Linear regression model

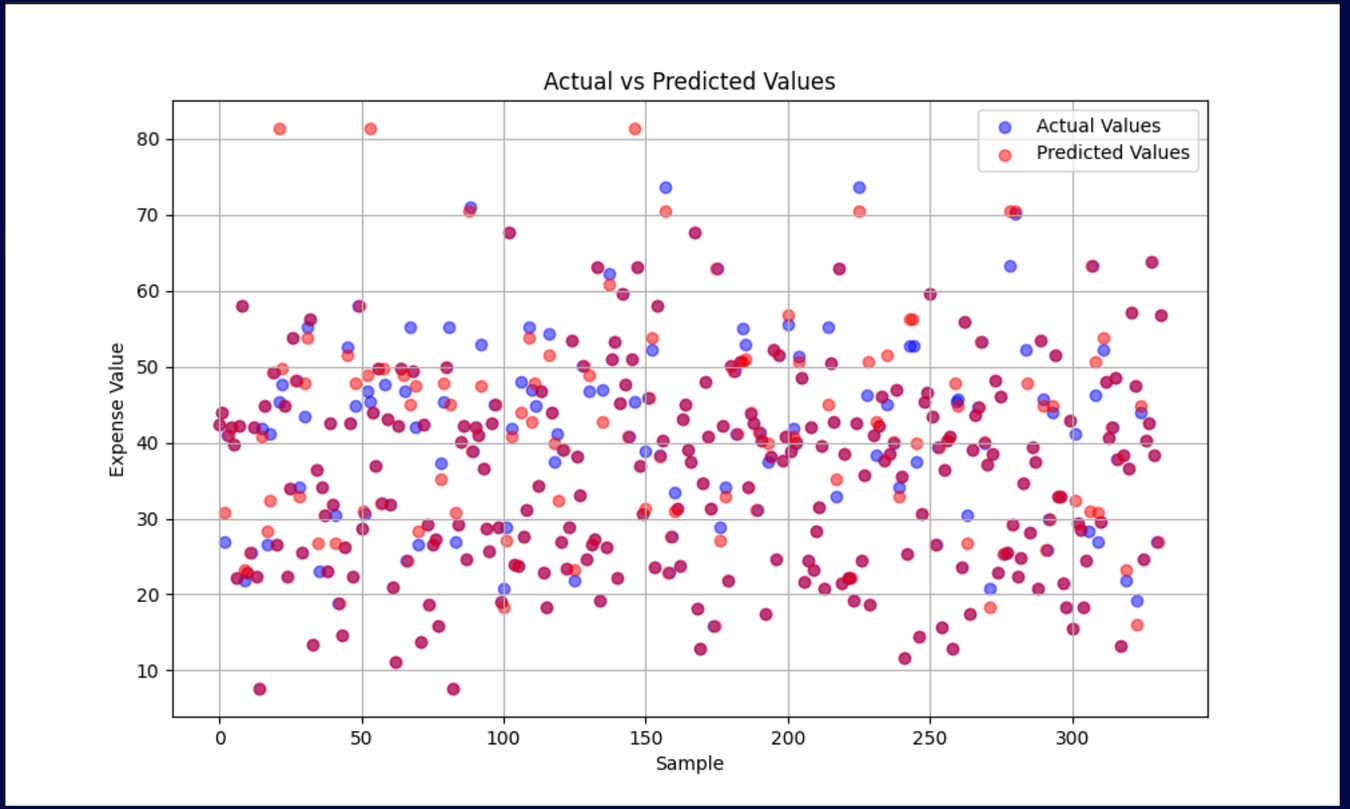
```
Linear regression results on testing data:
Mean absolute error: 0.47889295
Mean square error: 0.50201410
Root mean squared error: 0.70852953
R2 score: 0.57606562
```



EVALUATING MODELS

KNN regression model

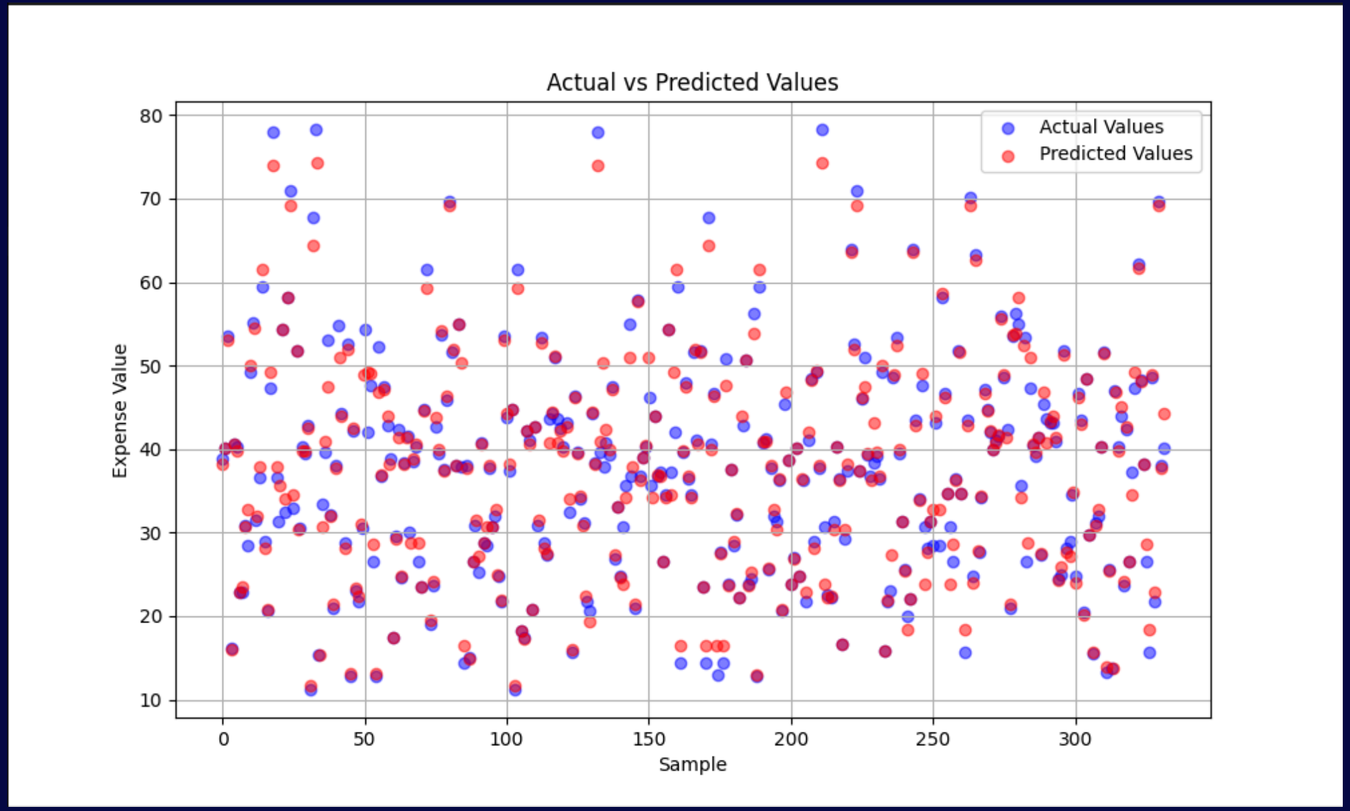
```
Mean absolute error: 0.07869216
Mean square error: 0.08225454
Root mean squared error: 0.28680052
R2 score: 0.90833309
```



EVALUATING MODELS

Random forest regression model

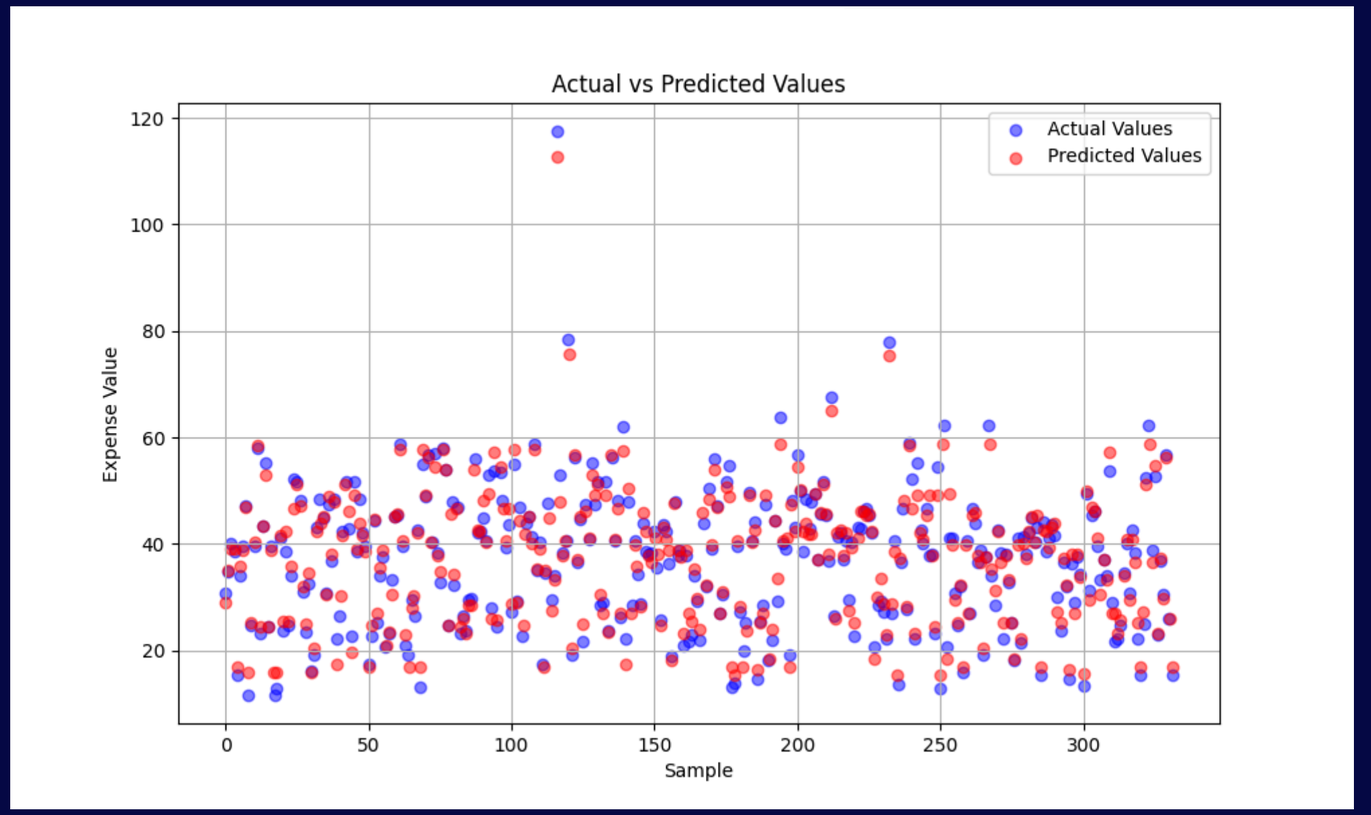
```
Mean absolute error: 0.07632463
Mean square error: 0.02044215
Root mean squared error: 0.14297606
R2 score: 0.97850986
```



EVALUATING MODELS

XGB regression model

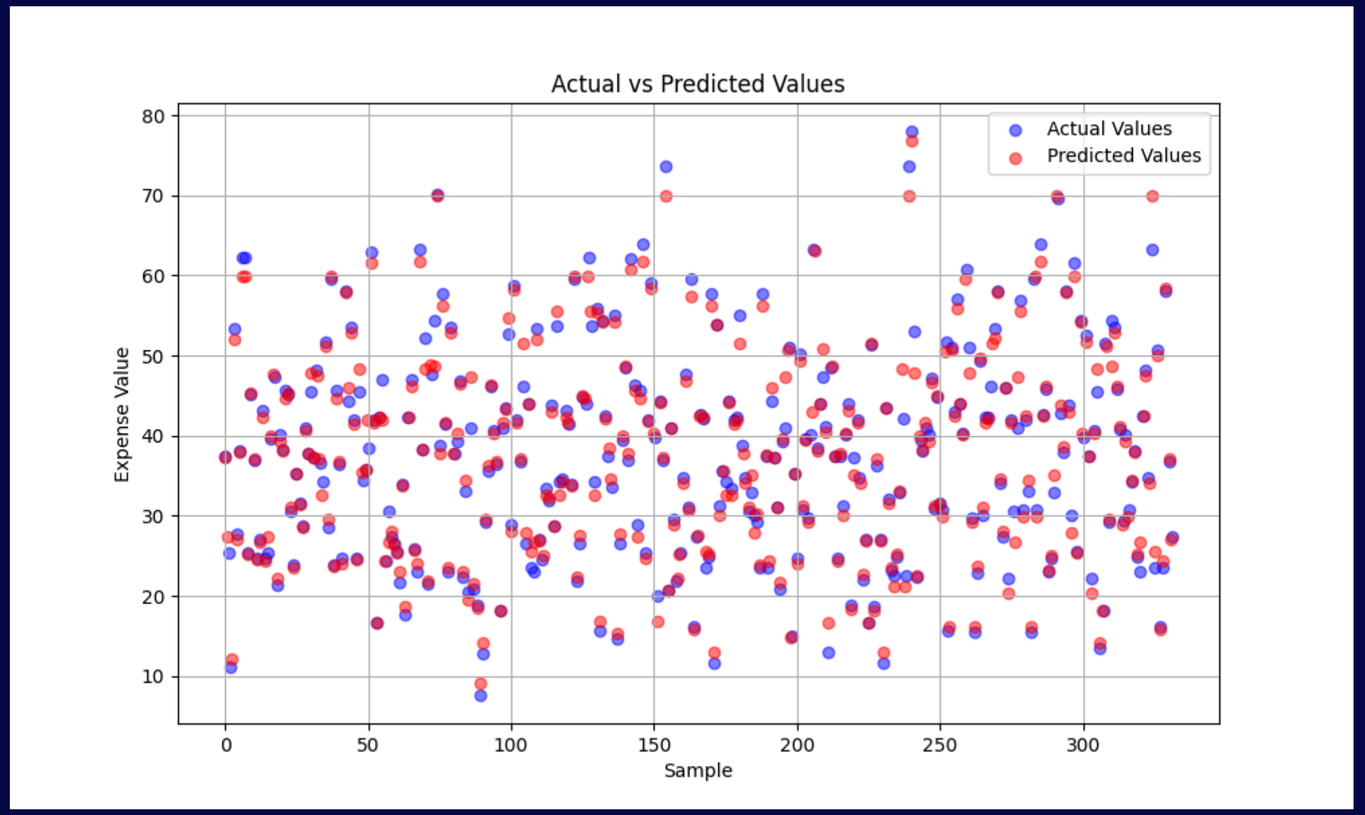
```
XGBoost results on testing data:
Mean absolute error: 0.11627595
Mean square error: 0.02483971
Root mean squared error: 0.15760618
R2 score: 0.97278419
```



EVALUATING MODELS

Voting regression model

```
Voting regressor results on testing data:
Mean absolute error: 0.07090721
Mean square error: 0.01296747
Root mean squared error: 0.11387481
R2 score: 0.98577000
```



COMPARING MODELS

Model	Mean Absolute Error	Mean Square Error	Root Mean Squared Error	R2 Score
Voting Regressor	0.070907	0.012967	0.113875	0.985770
Random Forest	0.076325	0.020442	0.142976	0.978510
XGB	0.116276	0.024838	0.157606	0.972784
KNN	0.078692	0.082255	0.286801	0.908333
Linear Regression	0.478893	0.502014	0.708530	0.576066

CHOOSING BEST MODEL

The Voting Regressor performs the best with the highest R^2 score (0.985770) and the lowest errors across the board. This model combines predictions from multiple individual models, leading to a more robust and accurate overall prediction by leveraging the strengths of different algorithms. The low Mean Square Error and Root Mean Squared Error also indicate minimal prediction variance, making it highly reliable.

TESTING BEST MODEL

	Actual	Predicted
0	40.3	40.028435
1	46.7	47.232858
2	38.1	37.799499
3	48.2	47.800693
4	28.1	25.654790
5	37.9	38.807471
6	25.7	25.585632
7	42.4	41.595723
8	18.8	18.418025
9	34.6	34.432404
10	39.3	40.487914
11	12.9	13.384405
12	54.4	49.834675
13	43.1	44.049979
14	28.9	27.768876
15	58.1	58.842609
16	42.5	42.524922
17	34.7	34.826102
18	30.5	31.033108
19	44.3	44.772855

As seen, the voting regressor model is able to give predicted values which are very close to the actual values, indicating that the model is able to perform very well on unseen data.


THANK YOU