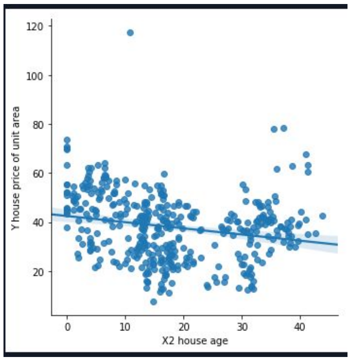
AIAP Question 2

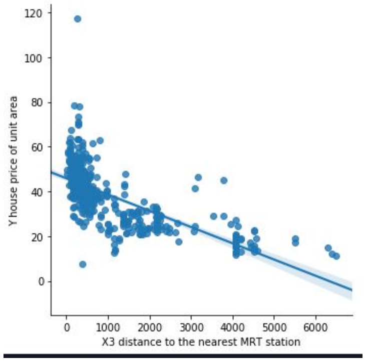
Data Cleaning

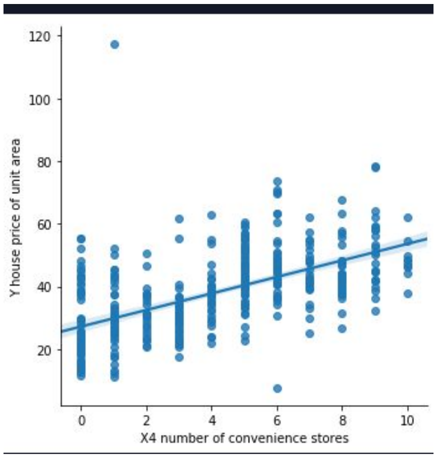
The data set provided is fairly clean as there aren't any missing values and we learn this from the .info and .describe methods. The main issue is one outlier data point with exceptionally high price that we find during EDA. This data point is removed to ensure that the data is not biased and heavily influenced by this single point.

Exploratory Data Analysis

I formed 3 simple hypotheses that younger houses, houses nearer to the MRT and higher number of convenience stores are likely to be more expensive. These are proven by the below plots:





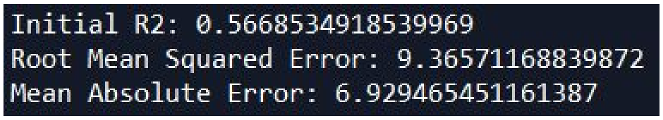


The outlier point is also clearly visible in the plots.

Modelling and Evaluation

Since we are required to predict the house price of unit area, this is a regression problem. I have chosen the metrics of R2, RMSE and MAE as they are common metrics for a regression problem. The goal is to achieve higher R2 scores and lower RMSE and MAE scores. R2 explains how well the price variable explains the variability in the other dependant variable X1-6. MAE is the average of the absolute difference between the predicted values and observed value and RMSE represents the sample standard deviation of the differences between predicted values and observed values.

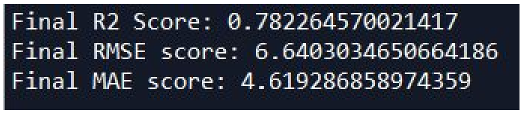
The initial model I chose was a simple linear regression model to get a baseline score for the metrics chosen. These are given as follows:



```
Initial R2: 0.5668534918539969
Root Mean Squared Error: 9.36571168839872
Mean Absolute Error: 6.929465451161387
```

Following this, I use a Ridge regression and Lasso visualization to find that the R2 score does improve with regularization and the lasso visualization shows that the feature of transaction date has the highest influence. This then calls for regularization and standardization as transaction date as a feature has very high coefficients in comparison to the other features and thus may be unduly influencing the data.

A decision tree model with tuned parameters has significantly improved the R2 score and this points in the right direction of the final model which is a random forest regressor. By tuning the parameters and with a randomized search validation, we get a R2 score of roughly 0.75. However, we have yet to apply scaling. This is implemented through a simple pipeline which does the scaling and applies a Random Forest regressor which does improve the R2 score to roughly 0.78. The final RMSE and MAE metrics are also printed to show their decrease which does indicate a better model than the initial linear regression.



```
Final R2 Score: 0.782264570021417
Final RMSE score: 6.6403034650664186
Final MAE score: 4.619286858974359
```

Conclusion

In conclusion, this dataset has been relatively clean to use and did not require much cleaning. The main task was in finding the metrics for evaluation and to tune the models progressively to find the best parameters required to boost the model performance. This also included a random search validation to increase the reliability of the predictions as well as simple pipeline process to scale the data before modelling.

Overall, it was an interesting problem to work on as it is always fun to get hands on experience with real life datasets to draw interesting insights from them.