CS 4372 Assignment 1

Linear Regression Analysis

September 21 2025

Group Members:
Madison Humphries (mrh210006)
Yeyoung Kim (yxk220011)

1 Pre-processing

- The data contained no missing or null values.
- `Predictors (Features): transaction_date`, `house_age`, `distance_to_nearest_MRT_station`, `n_convenience_store` `latitude`, and `longitude`.
- `Response (Target): price_of_unit_area`.
- The dataset consisted entirely of numerical variables, so no categorical encoding was necessary.

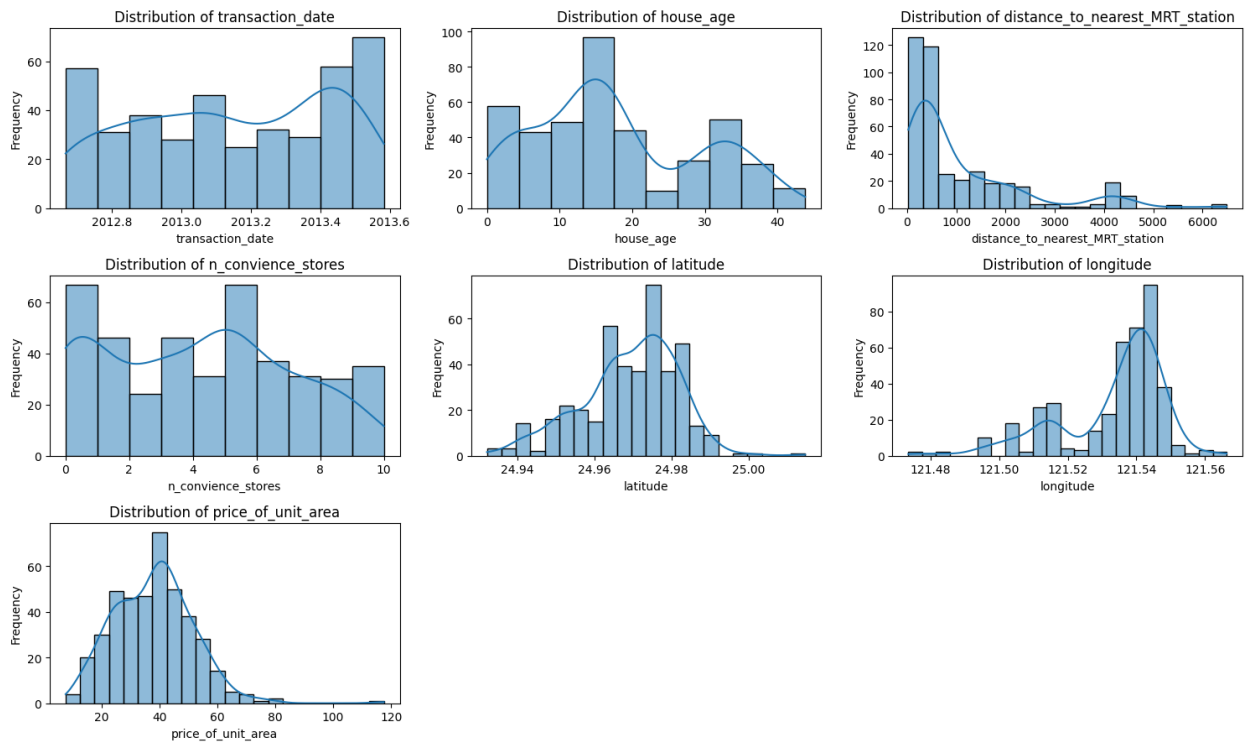

Figure 1: Distribution Plots

The figure above shows that most features aren't perfectly normal: several are skewed, some are discrete counts, and a few have multiple peaks. This is expected in real data. Simple reasons include integer counts (number of stores), a few extreme values (high-price areas), mixing of neighborhoods/time periods, practical geographic limits, and how/when the data were collected. Normality is a useful concept, but real-world features often deviate for these everyday reasons.

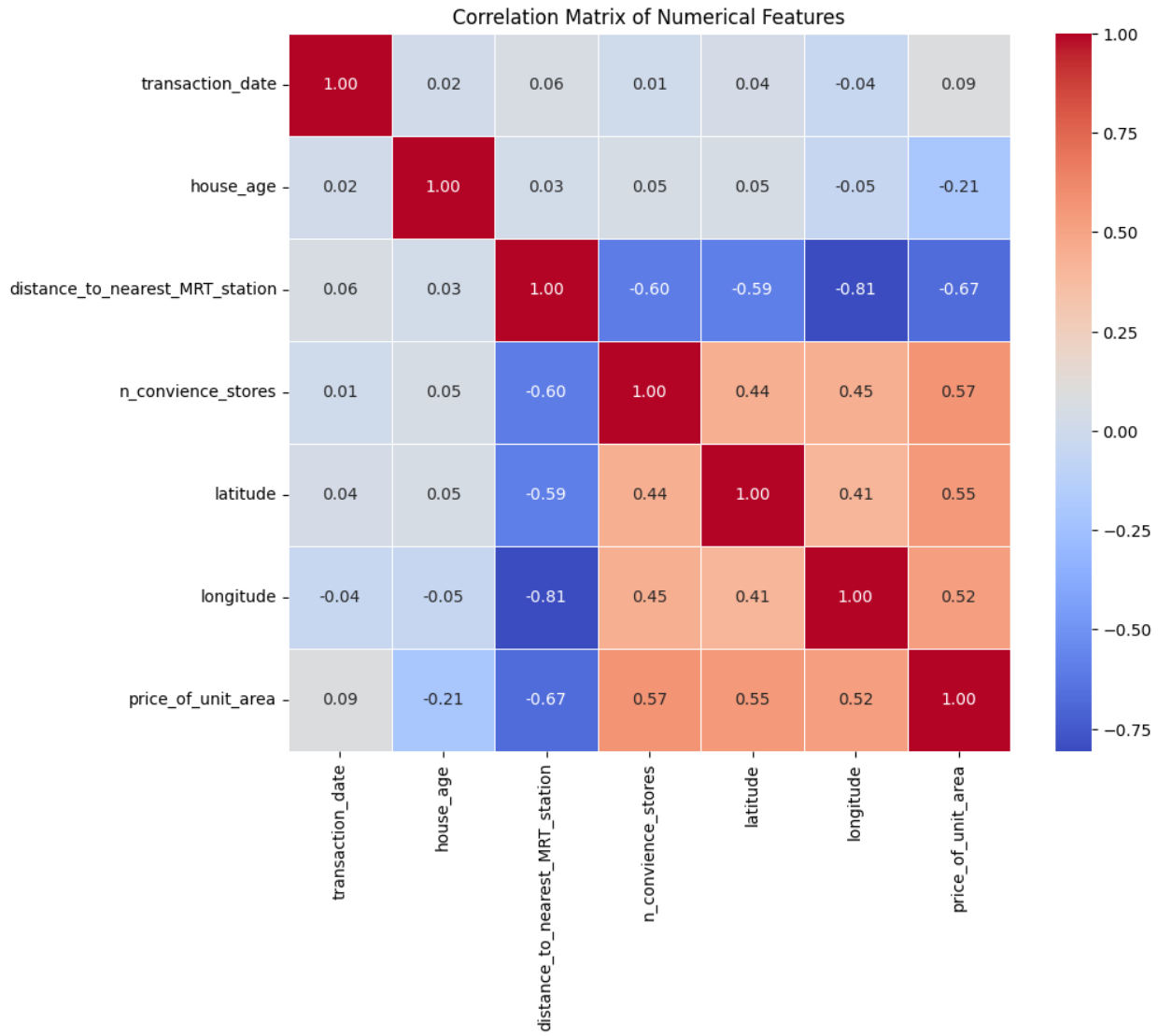

Figure 2: Correlation Matrix

A correlation matrix was computed to evaluate the relationships between each feature and the target variable (`price_of_unit_area`). The analysis revealed that `transaction_date` and `house_age` had very weak correlations with the target compared to other features.

Due to the high correlation between `distance_to_nearest_MRT_station` and `longitude`, we also wanted to check for multicollinearity (See Figure VIF).

	feature	VIF
1	transaction_date	1.014674
2	house_age	1.014287
3	distance_to_nearest_MRT_station	4.323019
4	n_convience_stores	1.617038
5	latitude	1.610234
6	longitude	2.926302

Figure 3: VIF Features

To further assess multicollinearity among the remaining predictors, the Variance Inflation Factor (VIF) was computed. All features produced VIF values well below the common threshold of 10, with the highest being around 4.3 for `distance_to_nearest_MRT_station` and 2.9 for `longitude`. These results indicate that some moderate correlation exists, although it is not severe enough to distort the regression estimates or inflate the standard errors. Therefore, both variables were retained in the model, as they contribute unique predictive information despite their relationship.

Therefore, after preprocessing the data, only two attributes were dropped from the dataset to reduce noise: `transaction_date` and `house_age`. Thus, we could focus on the most relevant predictors: `distance_to_nearest_MRT_station`, `n_convience_stores`, `latitude`, and `longitude`. This step ensures that the model is trained only on features with meaningful predictive power.

2 SGD Regressor Model

An `SGD Regressor` model was developed to predict the unit price of real estate.

2.1 Preprocessing and Data Splitting

As SGD is sensitive to the scale of input features, the predictor variables were first standardized using scikit-learn's `StandardScaler`. This process rescales each feature to have a mean of 0 and a standard deviation of 1, putting them on a comparable scale. Following standardization, the dataset was split into a training set (80% of the data) and a testing set (20%) to allow for a fair evaluation of the model's ability to generalize to new, unseen data.

2.2 Hyperparameter Tuning and Results

The model was then tuned using `GridSearchCV` with 5-fold cross-validation to find the best combination of hyperparameters, including the loss function, regularization penalty, and learning rate.

--- Hyperparameter Tuning Log ---

	param_loss	param_penalty	param_alpha	param_learning_rate	param_eta0	param_max_iter	mean_test_score	mean_test_mse
795	squared_error	l1	0.01	optimal	0.05	2000	-97.489519	97.489519
798	squared_error	l1	0.01	optimal	0.05	3000	-97.489519	97.489519
792	squared_error	l1	0.01	optimal	0.05	1000	-97.489519	97.489519
900	squared_error	l1	0.01	optimal	0.10	1000	-97.489519	97.489519
693	squared_error	l1	0.01	optimal	0.01	5000	-97.489519	97.489519
903	squared_error	l1	0.01	optimal	0.10	2000	-97.489519	97.489519
690	squared_error	l1	0.01	optimal	0.01	3000	-97.489519	97.489519
801	squared_error	l1	0.01	optimal	0.05	5000	-97.489519	97.489519
684	squared_error	l1	0.01	optimal	0.01	1000	-97.489519	97.489519
687	squared_error	l1	0.01	optimal	0.01	2000	-97.489519	97.489519
906	squared_error	l1	0.01	optimal	0.10	3000	-97.489519	97.489519
909	squared_error	l1	0.01	optimal	0.10	5000	-97.489519	97.489519
695	squared_error	elasticnet	0.01	optimal	0.01	5000	-97.976377	97.976377
803	squared_error	elasticnet	0.01	optimal	0.05	5000	-97.976377	97.976377
797	squared_error	elasticnet	0.01	optimal	0.05	2000	-97.976377	97.976377

Figure 4: Hyper-parameters: the cross-validated test MSE

The GridSearchCV process tested 1,296 hyperparameter combinations to find the optimal model. The table above shows a log of the top-performing combinations, sorted by their mean squared error (MSE) on the cross-validation test sets. The optimal parameters found were: **alpha:** 0.01, **eta0:** 0.01, **learning_rate:** `optimal`, **loss:** `squared_error`, **max_iter:** 1000, and **penalty:** `l1`.

2.3 Model Output

The Test R-squared, 0.5951, indicates that the model successfully explains approximately 59.5% of the variance in house prices with new, unseen data. Since the test R^2 exceeds the training R^2, the model appears well-regularized and shows no signs of overfitting.

Metric	Training Set	Test Set
R-squared (R^2)	0.4870	0.5951
Mean Squared Error (MSE)	96.4868	67.9336
Mean Absolute Error (MAE)	7.2292	6.5427

Figure 5: SGD Regressor Performance Metrics

2.4 Actual vs. Predicted

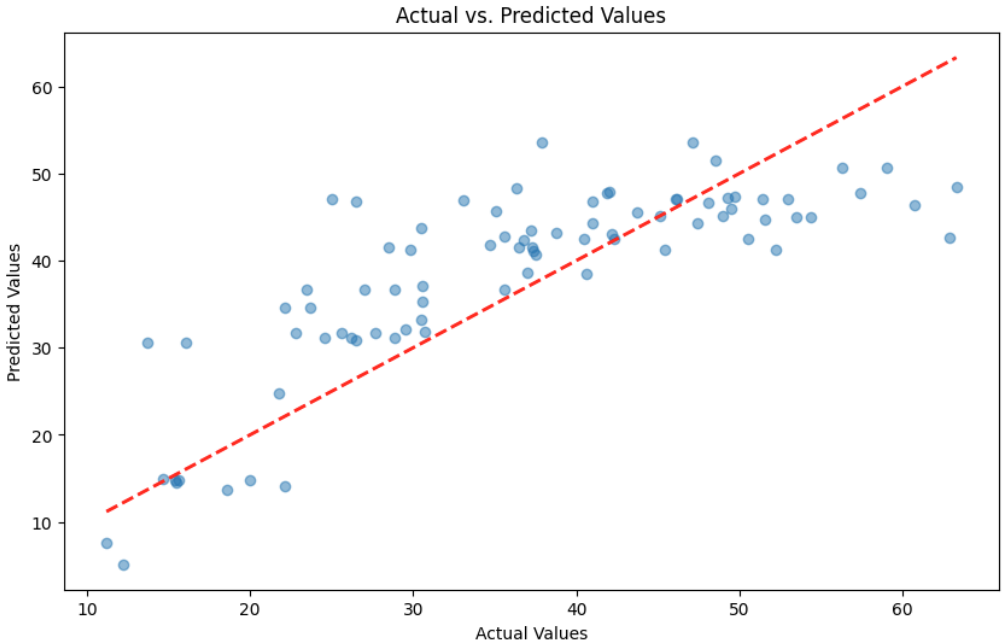

Figure 6: actual vs. predicted

The points form a positive trend that clusters around the red dashed line, which represents a good prediction. This confirms the model's R-squared of approximately 0.60, showing it has successfully learned the underlying patterns in the data.

2.5 Residual Plot

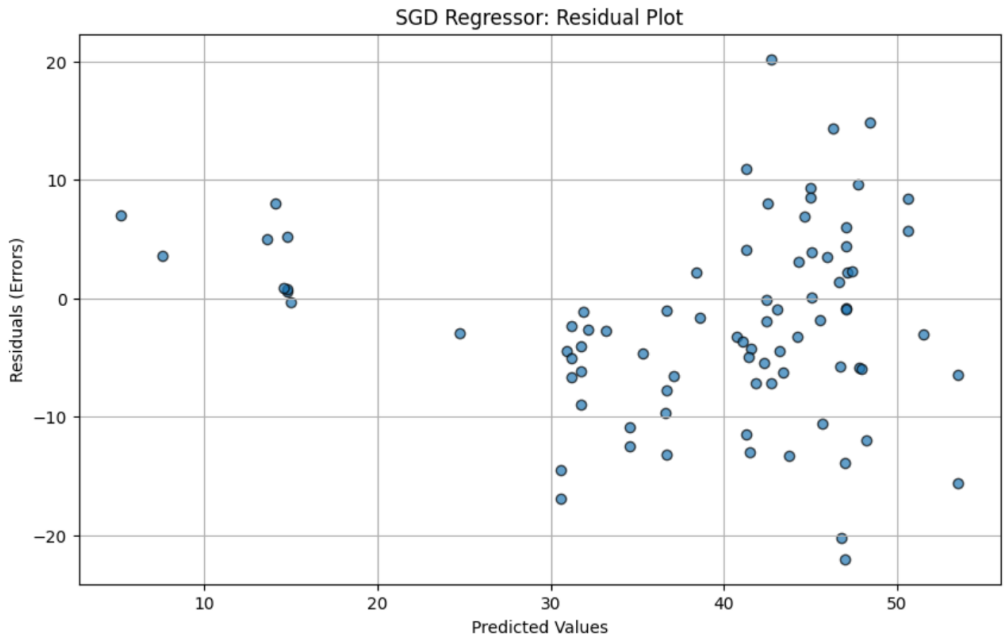

Figure 7: residual

To further diagnose the model's performance, a residuals plot was generated. The resulting scatter of points appears randomly distributed around the horizontal zero line, with no obvious curve or funnel-like pattern. This plot supports that the SGDRegressor is a well-behaved.

2.6 MSE vs. Iterations (Learning Rate)

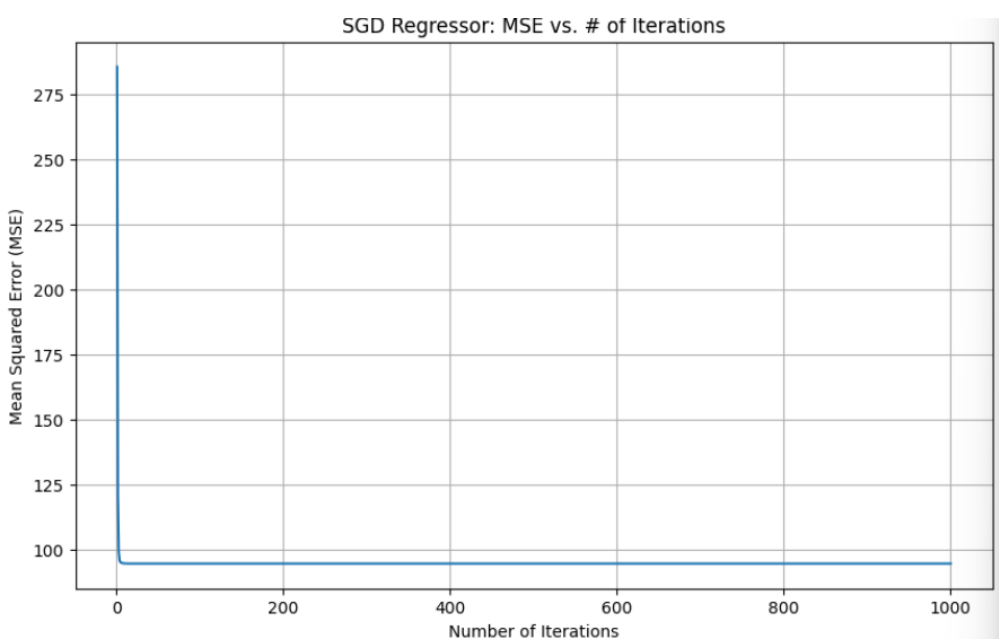

Figure 8: mse vs. iterations

The plot above shows the learning curve for the SGDRegressor, tracking the Mean Squared Error (MSE) at each training iteration. During the initial 40-50 iterations, the MSE drops sharply, indicating that the model is rapidly learning the dominant patterns in the data. The curve then flattens into a plateau, which means the model has converged; its performance is no longer improving with additional training, as the adjustments have become negligible. This visual evidence confirms that the chosen learning_rate was effective, allowing for stable and efficient learning, and that the max_iter value of 1000 was more than sufficient.

3 OLS Model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     price_of_unit_area   R-squared:                       0.521
Model:                            OLS   Adj. R-squared:                  0.516
Method:                 Least Squares   F-statistic:                     111.3
Date:                Sun, 21 Sep 2025   Prob (F-statistic):           4.24e-64
Time:                        01:08:34   Log-Likelihood:                -1515.3
No. Observations:                 414   AIC:                             3041.
Df Residuals:                     409   BIC:                             3061.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                             37.9802      0.465     81.679      0.000      37.066      38.894
distance_to_nearest_MRT_station   -5.6358      0.961     -5.862      0.000      -7.526      -3.746
n_convience_stores                 3.1568      0.589      5.357      0.000       1.998       4.315
latitude                           2.6721      0.586      4.556      0.000       1.519       3.825
longitude                          0.0462      0.795      0.058      0.954      -1.516       1.608
==============================================================================
Omnibus:                      203.952   Durbin-Watson:                   2.143
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2477.776
Skew:                           1.783   Prob(JB):                         0.00
Kurtosis:                      14.442   Cond. No.                         4.20
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 9: OLS Regression Results

The above figure (OLS) was fitted to the original dataset after preprocessing. Our overall interpretation:

1. Coefficients (coef)

- `const` (37.98): This is the intercept. It represents the predicted value of the house price (per unit area) when all predictors are zero.
- `distance_to_nearest_MRT_station` (-5.636): A statistically negative coefficient, which means as the distance to the nearest MRT increases, price decreases.
- `n_convience_stores` (3.157): A statistically positive coefficient, which means more nearby stores increase property value.
- `latitude` (2.672): A statistically positive coefficient, which means the higher latitude slightly increases house prices.
- `longitude` (0.0462): A not-significant positive coefficient, as it's very close to zero, and the p-value> 0.05. This means that longitude does not make a meaningful contribution once the other predictors are accounted for.

2. Standard Error

- The standard error measures how precisely the coefficients are estimated. This means the lower the errors, the more reliable the estimates in the model are.
  - As `longitude` has a large standard error (0.795) relative to its small coefficient (0.0462), it is insignificant to the model

3. t-value and p-value

- t-value tests whether the coefficient is significantly different from 0.
  - Predictors with t-values > 2 mean there is strong evidence that they affect the target; that the coefficient is statistically significant.
  - As `longitude` has a t-value ≈ 0.058, the predictor may not have a meaningful effect.
- p-value illustrates whether the null hypothesis (the predictor is not significant) can be rejected. If the p-value < 0.05, the predictor is significant to the model.
  - In the model, we see that all features except `longitude` have a p-value < 0.05. This means all of the features are significant in the model, while the feature `longitude` does not influence the model in any way.

4. R^2 and Adjusted R^2

- $R^2 = 0.521$: this means about 52.1% of the variation in the house price dataset is explained by the model.
- Adjusted $R^2 = 0.516$: this means that about 51.6% of the variation in the house price dataset is explained by the significant predictors.

5. F-statistic

- The F-statistic illustrates whether the model as a whole is statistically significant.
  - Since the F-statistic is larger, it suggests that the model is effective, as it indicates there is more variation explained by the model than unexplained variance

6. Diagnostics

- Omnibus, JB, Skew, Kurtosis: all of these test for residual normality.
  - As Skew = 1.78 and Kurtosis = 14.44, it suggests the residuals are skewed with heavy tails.
- Durbin-Watson (2.14): This tests for autocorrelation in the residuals.
  - As this is approximately 2, it suggests no serious autocorrelation in the residuals. This means the error terms in the model are independent, which states the model does not violate the OLS assumption. This is an ideal Durbin-Watson score for a regression model.
- Condition No. (4.20): This tests for multicollinearity in the model.
  - As this is a lower number, no multicollinearity issues are in the model, which is expected as we checked VIF during preprocessing.

7. Conclusions:

- The model is statistically significant and moderately predictive.
- The feature `longitude` adds no predictive value and can be removed from the model.
- The features that help predict the house price per unit area are `distance_to_nearest_MRT_station`, `n_convience_stores`, and `latitude`.

Because we see that `longitude` doesn't have a significant influence on the model, we decided to refit the model after excluding the variable and compare the models.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     price_of_unit_area   R-squared:                       0.521
Model:                            OLS   Adj. R-squared:                  0.518
Method:                 Least Squares   F-statistic:                     148.8
Date:                Sun, 21 Sep 2025   Prob (F-statistic):           3.20e-65
Time:                        02:09:59   Log-Likelihood:                -1515.3
No. Observations:                 414   AIC:                             3039.
Df Residuals:                     410   BIC:                             3055.
Df Model:                           3
Covariance Type:            nonrobust
==========================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                             37.9802      0.464     81.778      0.000      37.067      38.893
distance_to_nearest_MRT_station   -5.6768      0.652     -8.700      0.000      -6.959      -4.394
n_convience_stores                 3.1547      0.587      5.370      0.000       2.000       4.309
latitude                           2.6678      0.581      4.590      0.000       1.525       3.810
==============================================================================
Omnibus:                      203.694   Durbin-Watson:                   2.143
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2469.734
Skew:                           1.781   Prob(JB):                         0.00
Kurtosis:                      14.423   Cond. No.                         2.45
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 10: OLS Regression Results 2

Differences in the New Model

1. Coefficients (coef)

- `distance_to_nearest_MRT_station`: -5.636 $->$ -5.677
- `n_convience_stores`: 3.157 $->$ 3.155
- `latitude`: 2.672 $->$ 2.668
- `longitude` was statistically insignificant, contributing almost nothing to the prediction power, so after removing it, the other coefficients barely changed. This means dropping `longitude` did not affect the estimates of the remaining predictors.

2. Adjusted R^2: 0.516 $->$ 0.518

- Removing the `longitude` barely changes the model's explanatory power.
- The adjusted R^2 slightly increased, which makes sense because we removed a predictor that was not significant. This indicates a simpler model without losing predictive performance.

3. F-statistic: 111.3 $->$ 148.8

- As the F-statistic increased, which makes sense because removing an irrelevant variable reduces unnecessary degrees of freedom, improving the overall test statistic slightly.

4. Diagnostic: Cond. No.: 4.20 $->$ 2.45

- The lower condition number in Model 2 indicates reduced multicollinearity. This makes sense again as `longitude` was moderately correlated with `distance_to_nearest_MRT_station`, which we saw with the correlation matrix.

5. Conclusions:

- Model 2 (without the `longitude` feature) is preferable because it is simpler, avoids unnecessary multicollinearity, and retains the same predictive power as Model 1.
- This demonstrates the importance of feature selection: removing irrelevant or redundant variables can improve model interpretability without hindering performance.

4 Data Visualizations

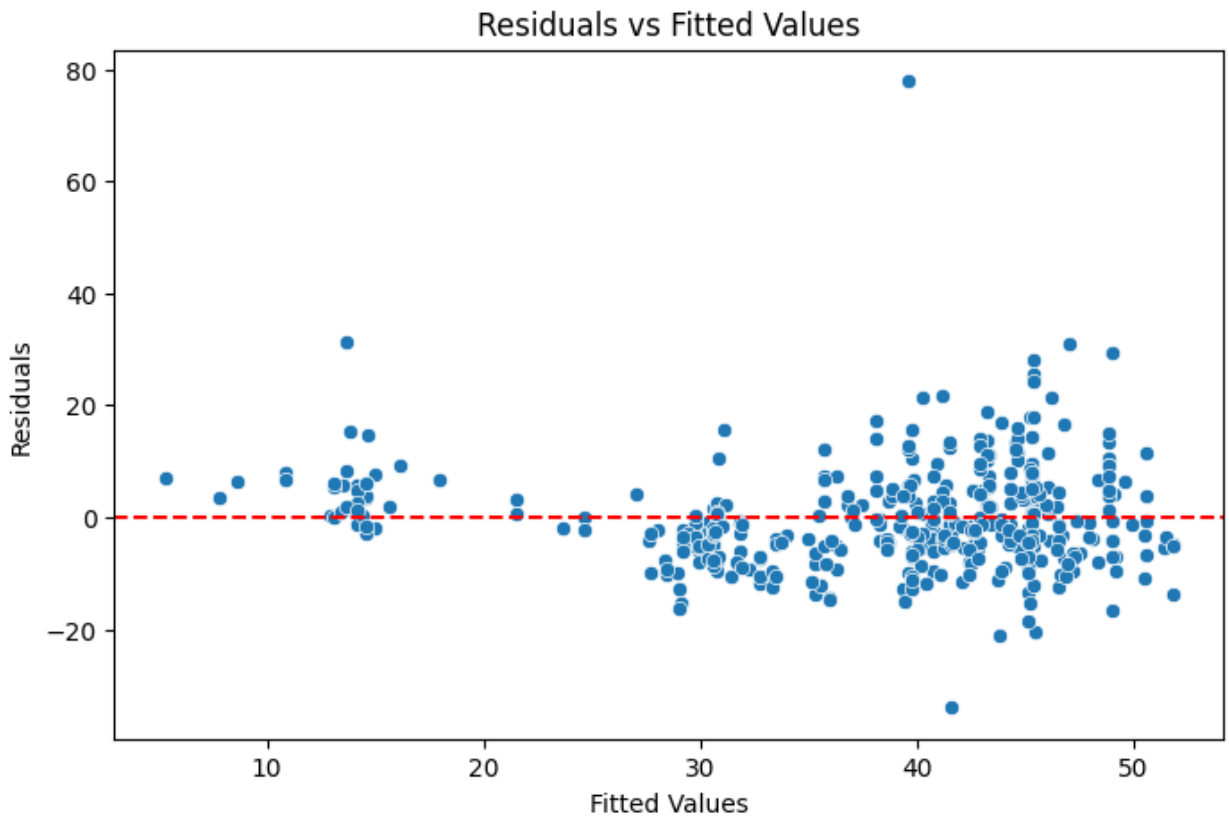

Figure 11: Residuals Vs Fitted Values

We observe that most of the residuals are evenly distributed above and below the horizontal line at 0, which is good. However, there appears to be a slight cone-shaped pattern, suggesting potential heteroscedasticity (variance of residuals changes with fitted values), which may affect inference. There are a few extreme residuals, but not enough to indicate highly influential points. Additionally, the two areas of clustering could suggest missing variables or interaction effects. Overall, these observations are consistent with the model's adjusted $R^2 \approx 0.518$, indicating moderate explanatory power.

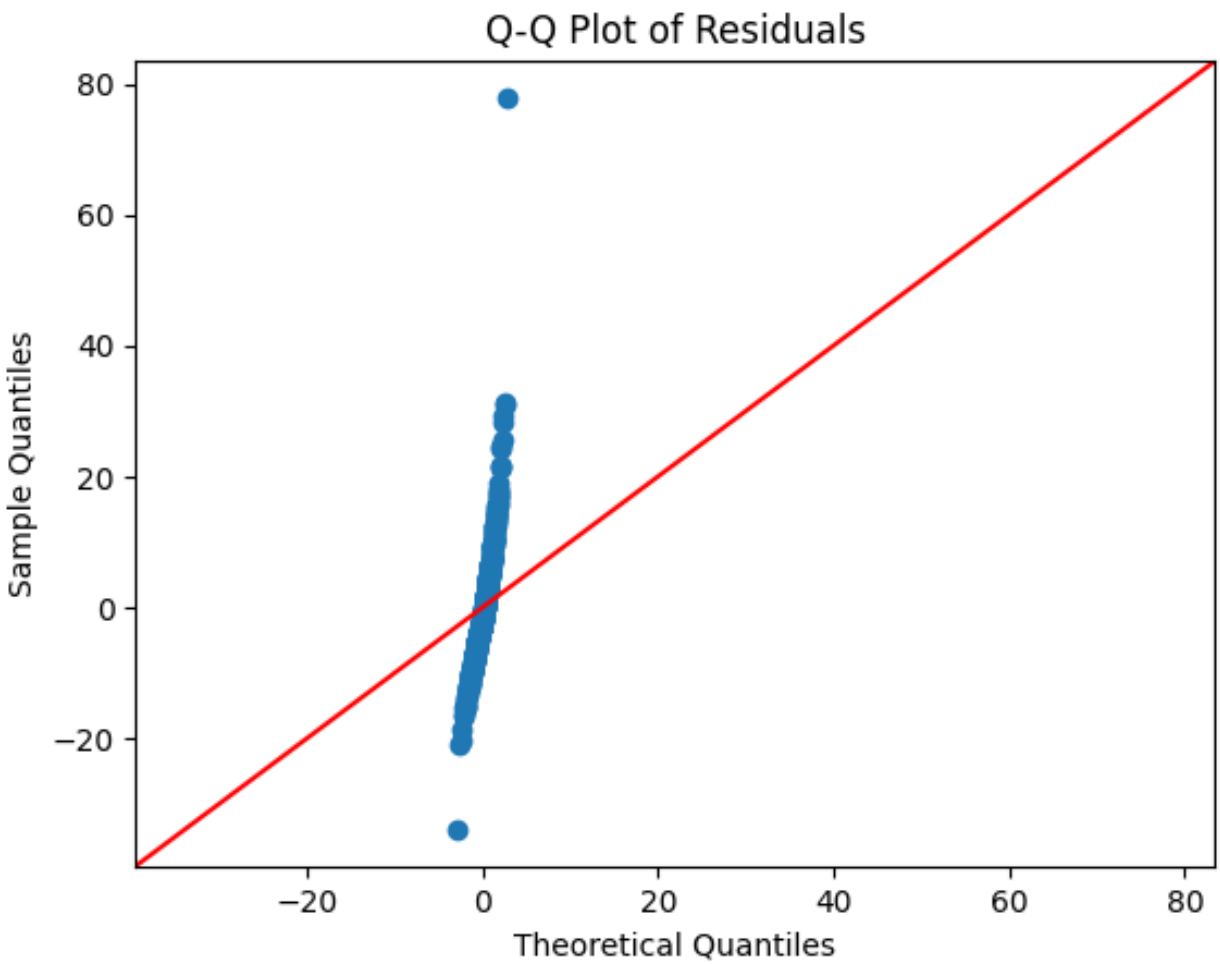

Figure 12: Q-Q Plot

The Q-Q plot suggests that the residuals are not perfectly normally distributed. All of the residuals are tightly clustered near zero, resulting in an almost vertical line where x ≈ 0. This demonstrates a potential violation of the OLS normality assumption, as if it did resemble a normal distribution, the plot would show the points following the diagonal line. This is sensitive to quantile alignment, and right clustering exaggerates deviations from a perfect normal distribution, which corresponds to other results of the model.

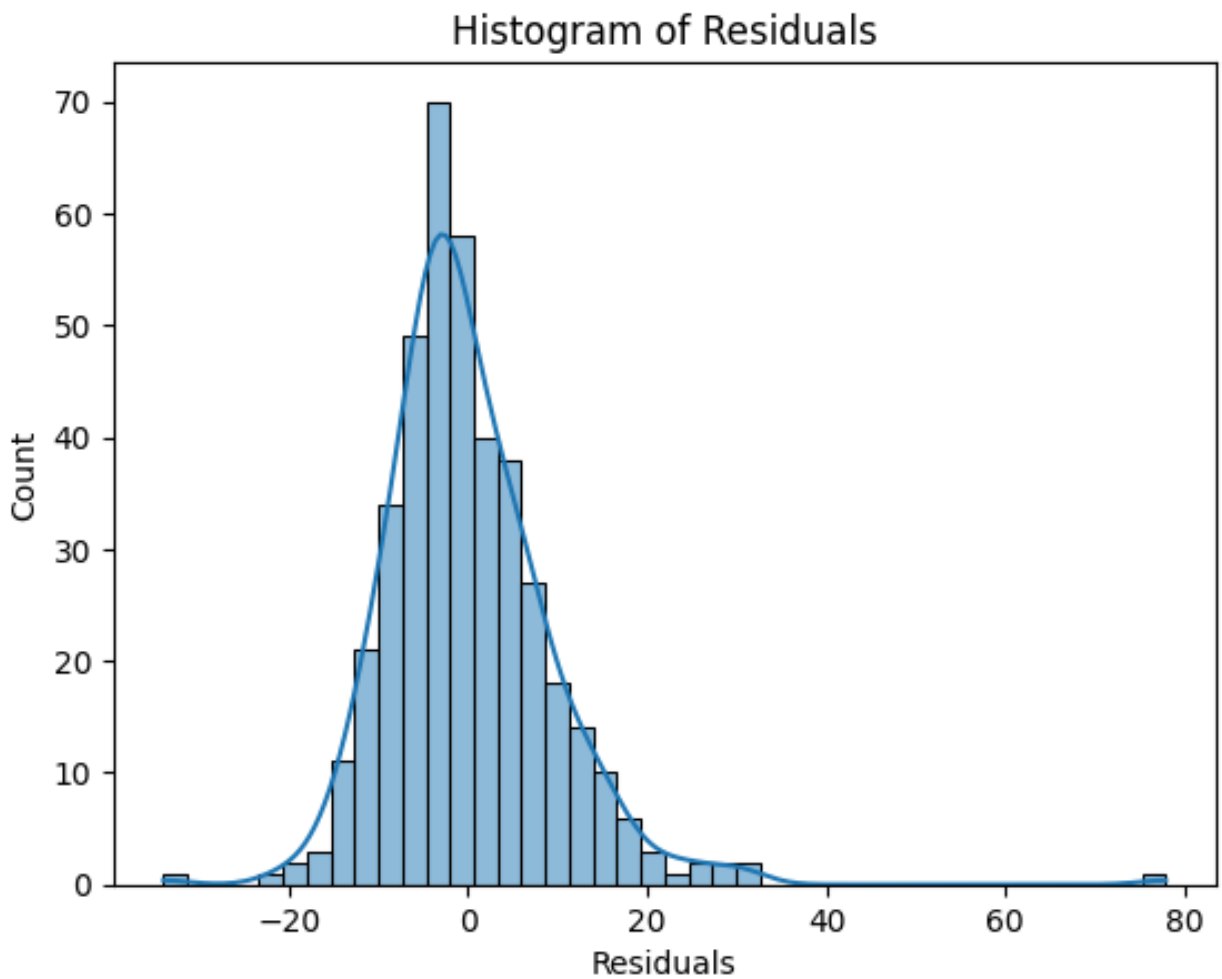

Figure 13: Histogram of Residuals

The histogram shows that residuals are approximately normally distributed with a single central peak around zero. This supports the idea that the majority of predictions are close to the actual values. Even though the Q-Q plot indicated tight clustering around zero and a few outliers, the histogram confirms that the overall shape of the residual distribution is roughly normal.

Conclusion:
The OLS model appears reasonably well-specified with mostly normal residuals, acceptable homoscedasticity, and no problematic multicollinearity. Although a few minor deviations and extreme points exist, they do not substantially undermine the model's validity. The model's adjusted $R^2 \approx 0.518$ reflects its moderate explanatory power given the selected predictors.