Taller 3

Section 1: Exploratory analysis

(a) Individual behaviour of each characteristic and of the response variable.

Estadísticas descriptivas / Descriptive Statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **X1 transaction date** | **X2 house age** | **X3 distance to the nearest MRT station** | **X4 number of convenience stores** | **X5 latitude** | **X6 longitude** | **Y house price of unit area** |
| count | 414.000000 | 414.000000 | 414.000000 | 414.000000 | 414.000000 | 414.000000 | 414.000000 |
| mean | 2013.148971 | 17.712560 | 1083.885689 | 4.094203 | 24.969030 | 121.533361 | 37.980193 |
| std | 0.281967 | 11.392485 | 1262.109595 | 2.945562 | 0.012410 | 0.015347 | 13.606488 |
| min | 2012.667000 | 0.000000 | 23.382840 | 0.000000 | 24.932070 | 121.473530 | 7.600000 |
| 25% | 2012.917000 | 9.025000 | 289.324800 | 1.000000 | 24.963000 | 121.528085 | 27.700000 |
| 50% | 2013.167000 | 16.100000 | 492.231300 | 4.000000 | 24.971100 | 121.538630 | 38.450000 |
| 75% | 2013.417000 | 28.150000 | 1454.279000 | 6.000000 | 24.977455 | 121.543305 | 46.600000 |
| max | 2013.583000 | 43.800000 | 6488.021000 | 10.000000 | 25.014590 | 121.566270 | 117.500000 |
|  |  |  |  |  |  |  |  |

All variables have a count of 414 meaning there is no missing data.

A blue square with black lines

Description automatically generated with medium confidenceA blue square with black lines

Description automatically generated

1. Transaction date is a number from 2012.667 to 2013.583 indicating the year of the transaction. The distribution of the data looks evenly distributed in the boxplot with the mean being roughly in the center of the min and the max and the quarters also look roughly the same distance away from the mean of 2013.15. In the histogram in part b we see that there are a bit more values near the min and especially the max.
2. House age is a number between 0 and 43.8 most likely indicating years. The distribution has a mean of 16.1 and is skewed to the right.

A graph of a box with a blue rectangle and black lines

Description automatically generatedA blue square with black lines

Description automatically generated

1. Distance to the nearest MRT stations is a value from 23.38 to 6488.02 and has a mean of 1083.9. The boxplot shows there are quite a few outliers as there are quite a few values outside the 1.5 time IQR. Both the boxplot and the histogram in section b show that most values are quite low.
2. The number of convenience stores are integers from 0-10 with a mean of 4.09.

A graph with a blue rectangle

Description automatically generatedA graph of a diagram

Description automatically generated

1. Latitude is quite evenly spread with a few outliers. The mean is 24.97 and the values range from 24.9321 and 25.0146, so very close together. The histogram in part b shows that most values are close to the mean with a large peak in the middle.
2. Longitude is similar to the latitude but with a slight left skewness. The mean is 121.53 and the values range from 121.4735 and 121.5663.

A graph with a blue rectangle and white lines

Description automatically generated

7. The house price of unit area ranges from 7.6 to 117.5 with a mean of 37.98. The histogram in section b shows a fairly normal distribution with a slight right skewness and the boxplot shows a few outliers to the right.

b) Correlations between characteristics and with the response variable.

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

The correlation matrix shows Pearson correlation measures between the characteristics and the response variable. Distance to the nearest MRT station has some strong negative correlations with house price of unit area (-0.67), longitude (-0.81), latitude (-0.59) and number of convenience stores (-0.6). This indicates that houses with a large distance to the nearest station have lower prices of unit area. There also seems to be a smaller number of convenience stores if the distance to the nearest MRT station is larger.

The house price of unit area has some moderate/strong positive correlations with the number of convenience stores (0.57), latitude (0.55) and longitude (0.52). Furthermore longitude has a moderate positive correlation with number of convenience stores (0.45) and latitude (0.41) and latitude also has a moderate positive correlation with number of convenience stores (0.44).

c) Bivariate exploration between each characteristic and the response variable.

A graph with blue lines and black text

Description automatically generated

3. Create a linear model that allows predicting the response variable from the characteristics.

characteristics. In your report summarize and comment on:

(a) Metrics of the model using training data.

Random\_state=1 is used to be able to replicate the data.

Shuffle=False and test\_size=0.2 gives

MAE: 5.847764972264446

MSE: 59.40924577639948

RMSE: 7.70773934279043

Shuffle=False and test\_size=0.3 gives

MAE: 6.177093207255384

MSE: 65.24836536193719

RMSE: 8.077646028512094

Shuffle=True and test\_size=0.2 gives

MAE: 5.343030944663055

MSE: 45.01050719519454

RMSE: 6.708987046879323

Shuffle=True and test\_size=0.3 gives

MAE: 6.274984907782299

MSE: 105.56582053294484

RMSE: 10.274522885903016

Meaning with shuffle and a test\_size of 0.2 creates a model with the lowest RMSE.

The metrics of the model are:  
[('X1 transaction date', 5.72), ('X2 house age', -0.25), ('X3 distance to the nearest MRT station', -0.005), ('X4 number of convenience stores', 1.076), ('X5 latitude', 227.04), ('X6 longitude', -35.70)]

b) Model metrics using cross-validation.

RMSE (cv=7) gives [ 7.3910315 8.39310153 9.2236802 7.62232252 12.46389036 7.74150419 8.34920442]

Mean = 8.741

c) Evaluation of the model and its parameters using statistical tests.

A model with all 6 independent variables gives us the following results:

OLS Regression Results

======================================================================================

Dep. Variable: Y house price of unit area R-squared: 0.543

Model: OLS Adj. R-squared: 0.534

Method: Least Squares F-statistic: 60.00

Date: Wed, 07 Feb 2024 Prob (F-statistic): 1.05e-48

Time: 13:08:56 Log-Likelihood: -1129.0

No. Observations: 310 AIC: 2272.

Df Residuals: 303 BIC: 2298.

Df Model: 6

Covariance Type: nonrobust

==========================================================================================================

coef std err t P>|t| [0.025 0.975]

----------------------------------------------------------------------------------------------------------

const -1.093e+04 8496.772 -1.287 0.199 -2.77e+04 5786.448

X1 transaction date 5.1272 1.897 2.702 0.007 1.393 8.861

X2 house age -0.2389 0.047 -5.135 0.000 -0.330 -0.147

X3 distance to the nearest MRT station -0.0049 0.001 -5.539 0.000 -0.007 -0.003

X4 number of convenience stores 1.0709 0.231 4.630 0.000 0.616 1.526

X5 latitude 216.8963 52.484 4.133 0.000 113.618 320.175

X6 longitude -39.1702 59.720 -0.656 0.512 -156.689 78.349

==============================================================================

Omnibus: 189.462 Durbin-Watson: 2.086

Prob(Omnibus): 0.000 Jarque-Bera (JB): 2953.563

Skew: 2.181 Prob(JB): 0.00

...

Notes:

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

[2] The condition number is large, 3.79e+07. This might indicate that there are

strong multicollinearity or other numerical problems.

The R-squared is 0.543 indicating tht approximately 54.3% of the variance is explained by the model. The adjusted R-squared which takes into account the number of predictor variables is 0.534. The overall model is significant as it has a high F-statistic of 60 and a very low p-value 1.05e-48 way below 0.05. All variables except X6 longitude have p-values below 0.05, meaning it might be worth looking into leaving this variable out.

A model with 5 independent variables (excluding X6 longitude):

OLS Regression Results

======================================================================================

Dep. Variable: Y house price of unit area R-squared: 0.542

Model: OLS Adj. R-squared: 0.535

Method: Least Squares F-statistic: 72.05

Date: Wed, 07 Feb 2024 Prob (F-statistic): 1.46e-49

Time: 13:56:27 Log-Likelihood: -1129.2

No. Observations: 310 AIC: 2270.

Df Residuals: 304 BIC: 2293.

Df Model: 5

Covariance Type: nonrobust

==========================================================================================================

coef std err t P>|t| [0.025 0.975]

----------------------------------------------------------------------------------------------------------

const -1.586e+04 3963.736 -4.002 0.000 -2.37e+04 -8062.018

X1 transaction date 5.1497 1.895 2.717 0.007 1.420 8.879

X2 house age -0.2373 0.046 -5.112 0.000 -0.329 -0.146

X3 distance to the nearest MRT station -0.0045 0.001 -7.597 0.000 -0.006 -0.003

X4 number of convenience stores 1.0873 0.230 4.733 0.000 0.635 1.539

X5 latitude 221.7681 51.907 4.272 0.000 119.626 323.910

==============================================================================

Omnibus: 192.288 Durbin-Watson: 2.072

Prob(Omnibus): 0.000 Jarque-Bera (JB): 3076.312

Skew: 2.217 Prob(JB): 0.00

Kurtosis: 17.782 Cond. No. 1.77e+07

...

Notes:

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

[2] The condition number is large, 1.77e+07. This might indicate that there are

strong multicollinearity or other numerical problems.

The adjusted R-squared has very slightly increased to 0.535 meaning this model has a slightly better fit.