Joep van der Kamp – 202416832
Wilmar Andrés Calderón Guevara 201630701
13 de febrero de 2024
Analítica Computacional para la Toma de Decisiones – IIND 4130

Taller 3: Modelos de Aprendizaje en Python

Github repository: https://github.com/joepie2/actd-taller2

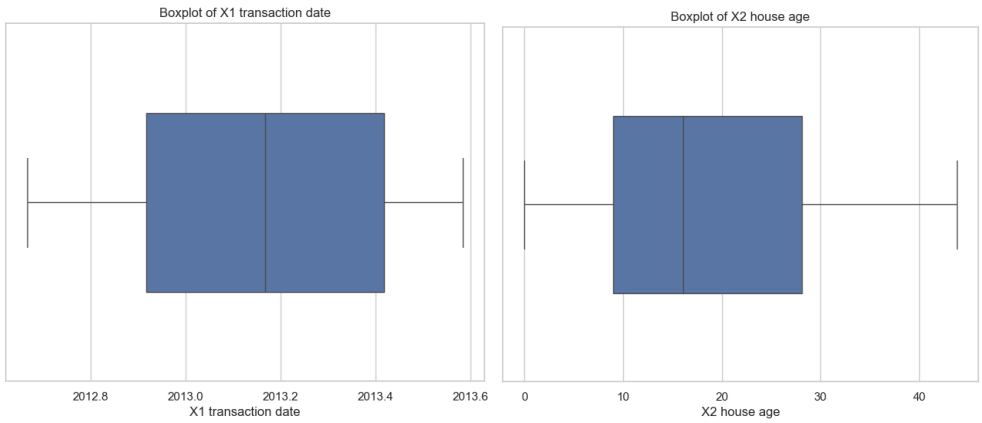
Section 1: Exploratory analysis

(a) Individual behaviour of each characteristic and of the response variable. / *Comportamiento individual de cada característica y de la variable de respuesta.*

Estadísticas descriptivas / *Descriptive Statistics*

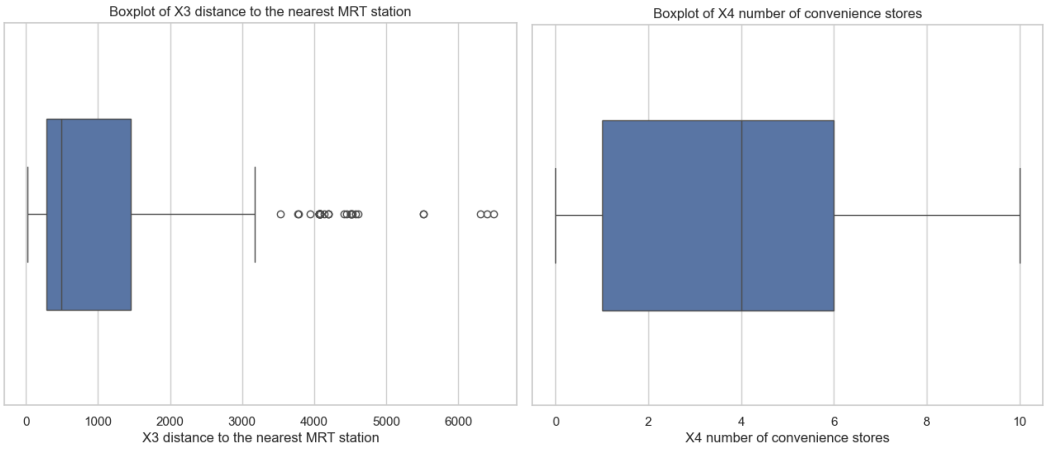
	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
count	414.000000	414.000000	414.000000	414.000000	414.000000	414.000000	414.000000
mean	2013.148971	17.712560	1083.885689	4.094203	24.969030	121.533361	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. / *Todas las variables tienen un recuento de 414, lo que significa que no faltan datos.*



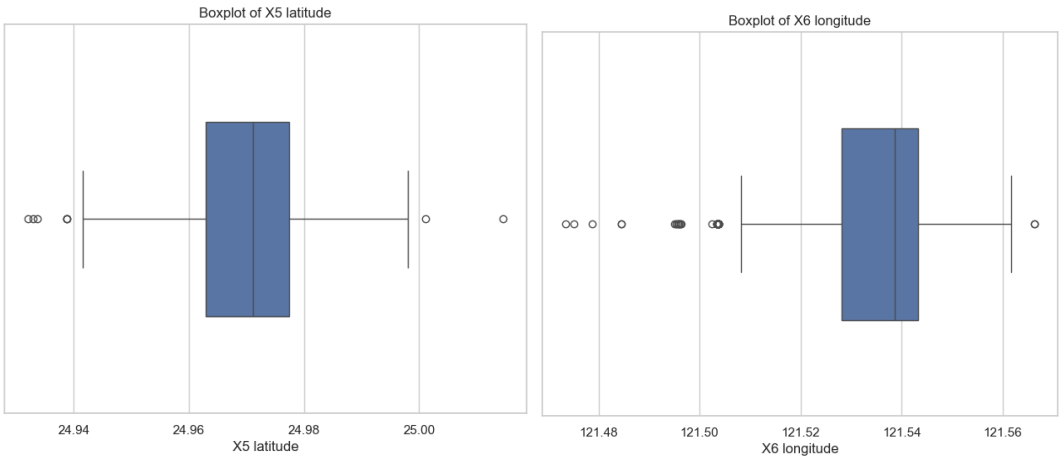
1. **X1**: Transaction date is a number from 2012.667 to 2013.583 indicating the year of the transaction. The distribution of the data looks evenly distributed in the boxplot with the mean being roughly in the center of the min and the max and the quarters also look roughly the same distance away from 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.
   / *La fecha de la transacción es un número de 2012.667 a 2013.583 que indica el año de la transacción. La distribución de los datos se ve distribuida uniformemente en el diagrama de cajas y bigotes, con la media aproximadamente en el centro entre el mínimo y el máximo y los cuantiles también se ven aproximadamente a la misma distancia de la media de 2013.15.*
2. **X2**: 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.

/ La antigüedad de la casa es un número entre 0 y 43,8 que probablemente indica años. La distribución tiene una media de 16,1 y está sesgada hacia la derecha.



3. **X3**: 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. /
La distancia a las estaciones de metro más cercanas es un valor de 23.38 a 6488.02 y tiene una media de 1083.9. El diagrama de cajas y bigotes muestra que hay bastantes valores atípicos, ya que hay bastantes valores fuera del RIC (Rango Intercuartil) 1,5 veces. Tanto el diagrama de caja como el histograma de la sección b muestran que la mayoría de los valores son bastante bajos.

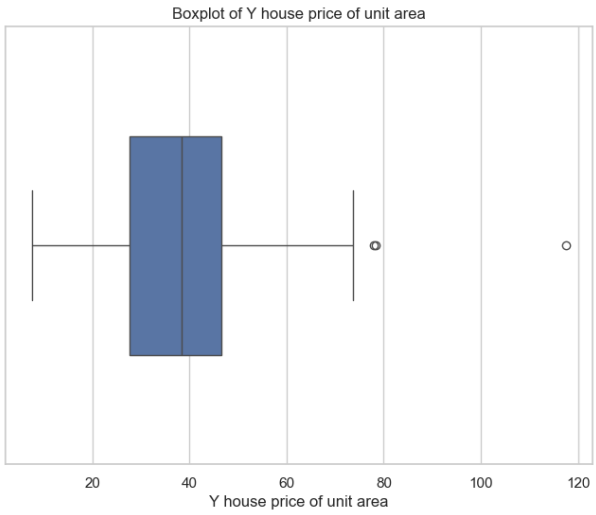
4. **X4**: The number of convenience stores are integers from 0-10 with a mean of 4.09. / *El número de tiendas de conveniencia es entero de 0 a 10 con una media de 4,09.*



5. **X5**: 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. / *La latitud se distribuye de manera bastante uniforme con algunos valores atípicos. La media es de 24,97 y los valores oscilan*

entre 24,9321 y 25,0146, por lo que están muy juntos. El histograma de la parte b muestra que la mayoría de los valores están cerca de la media con un pico grande en el medio.

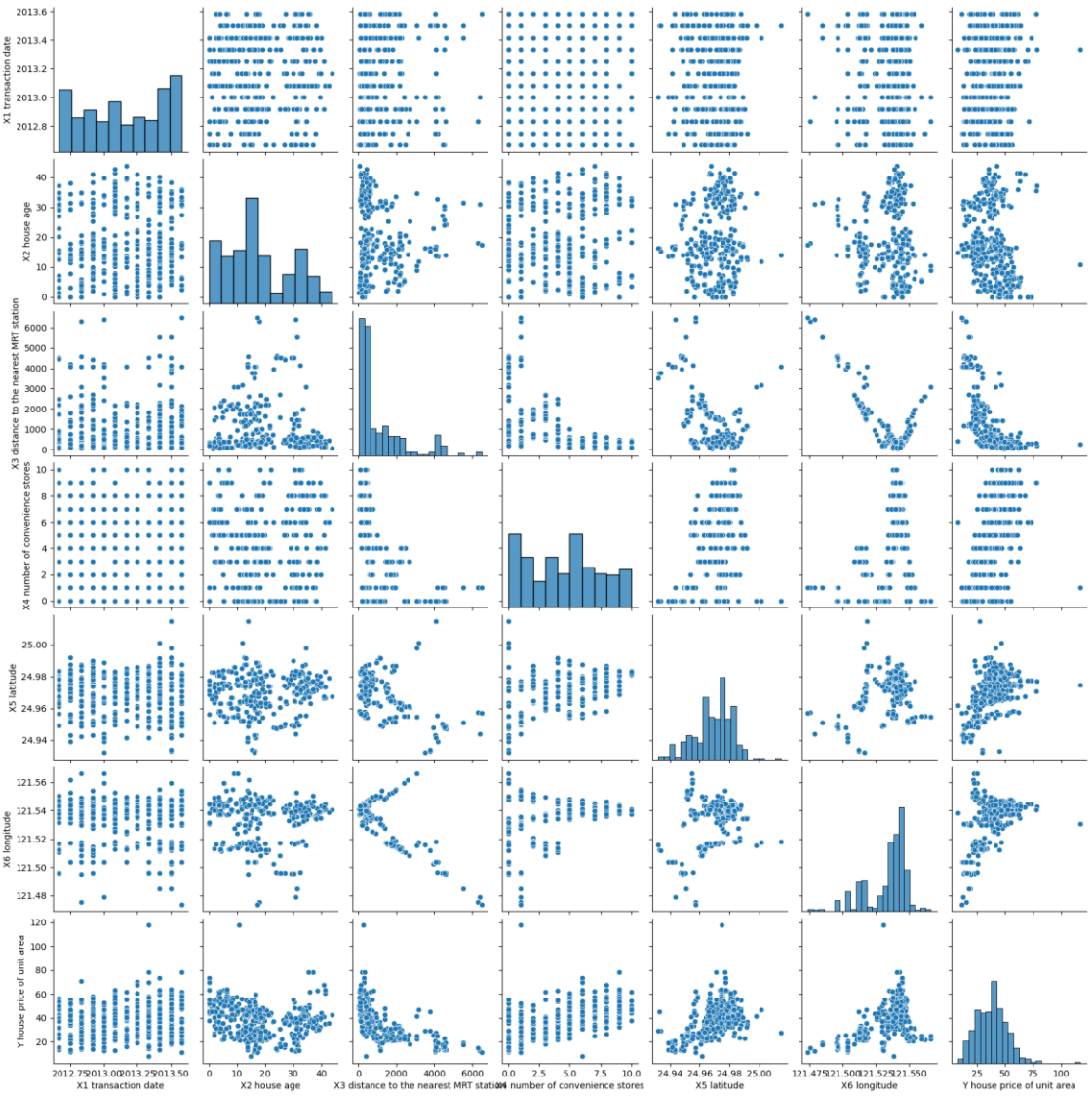
6. **X6**: Longitude is similar to the latitude but with a slight left skewness. The mean is 121.53 and the values range from 121.4735 and 121.5663. / *La longitud es similar a la latitud pero con una ligera inclinación a la izquierda. La media es de 121,53 y los valores oscilan entre 121,4735 y 121,5663.*



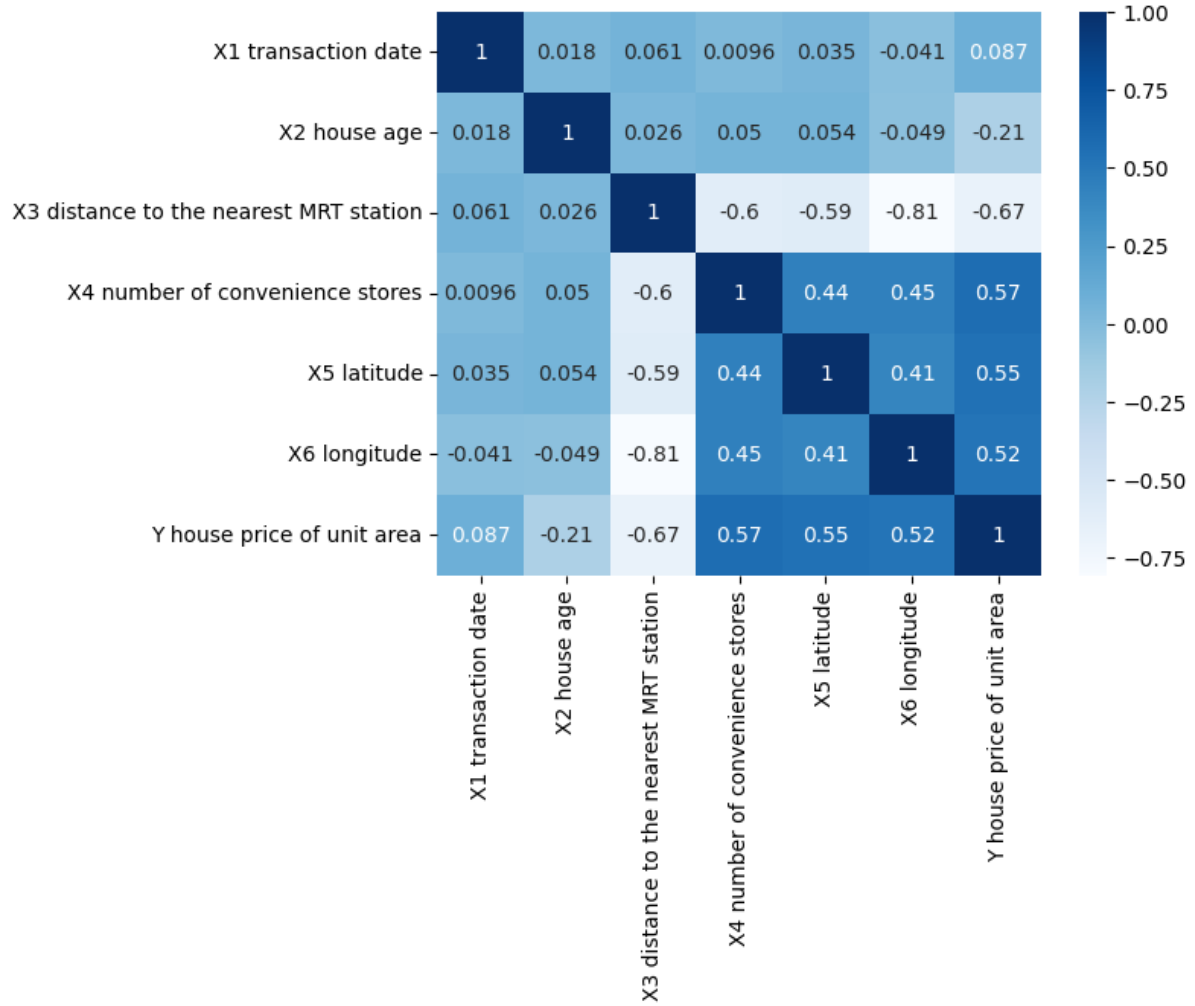
7. **Y**: The house price of unit area ranges from 7.6 to 117.5 with a mean of 37.98. The histogram in section b shows a fairly normal distribution with a slight right skewness and the boxplot shows a few outliers to the right. /
El precio de la vivienda por unidad de superficie oscila entre 7,6 y 117,5 con una media de 37,98. El histograma de la sección b muestra una distribución bastante normal con una ligera asimetría a la derecha y el diagrama de caja muestra algunos valores atípicos a la derecha.

(b) Correlations between characteristics and with the response variable. / *Correlaciones entre características y con la variable respuesta.*

Joep van der Kamp – 202416832
Wilmar Andrés Calderón Guevara 201630701
13 de febrero de 2024
Analítica Computacional para la Toma de Decisiones – IIND 4130



Joep van der Kamp – 202416832
Wilmar Andrés Calderón Guevara 201630701
13 de febrero de 2024
Analítica Computacional para la Toma de Decisiones – IIND 4130



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. /

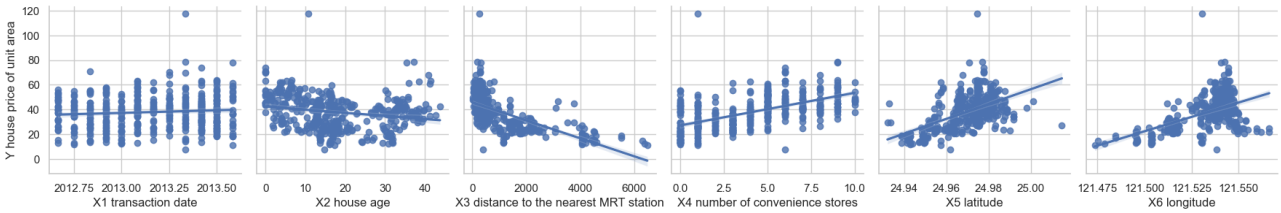
La matriz de correlación muestra las medidas de correlación de Pearson entre las características y la variable respuesta. La distancia a la estación de MRT más cercana tiene fuertes correlaciones negativas con el precio de la vivienda de la unidad de área (-0.67), longitud (-0.81), latitud (-0.59) y número de tiendas de conveniencia (-0.6). Esto indica que las casas con una gran distancia a la estación más cercana tienen precios más bajos de área unitaria. También parece haber un número menor de tiendas de conveniencia si la distancia a la estación de MRT más cercana es mayor.

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). /

Joep van der Kamp – 202416832
Wilmar Andrés Calderón Guevara 201630701
13 de febrero de 2024
Analítica Computacional para la Toma de Decisiones – IIND 4130

El precio de la vivienda por unidad de superficie tiene algunas correlaciones positivas moderadas/fuertes con el número de tiendas de conveniencia (0,57), latitud (0,55) y longitud (0,52). Además, la longitud tiene una correlación positiva moderada con el número de tiendas de conveniencia (0,45) y la latitud (0,41), y la latitud también tiene una correlación positiva moderada con el número de tiendas de conveniencia (0,44).

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



- X1 y Y: se observa que no hay una correlación fuerte entre las variables, también asociado a que los valores están entre 2012.6 y 2013.6 (1 año), por lo que su variabilidad no es muy representativa en el precio de venta. */- X1 and Y: it is observed that there is no strong correlation between the variables, also associated to the fact that the values are between 2012.6 and 2013.6 (1 year), so their variability is not very representative in the sale price.*

- X2 y Y: se observa una ligera tendencia negativa, de modo que los inmuebles más viejos tienen un menor precio / - *X2 and Y: a slight negative trend is observed, so that older properties have a lower price.*

- X3 y Y: se observa una tendencia negativa marcada del precio con respecto a la distancia al transporte público. / - *X3 and Y: there is a marked negative trend in price with respect to distance to public transportation.*

- X4 y Y: se evidencia una tendecia creciente del valor del inmueble conforme el número de tiendas aumenta. / - *X4 and Y: there is an increasing trend in the value of the property as the number of stores increases.*

- X5 y Y: se evidencia que mayores valores de latitud implican mayores valores del inmueble, por lo que vivir al norte es más caro. / - *X5 and Y: it is evident that higher latitude values imply higher property values, so living in the north is more expensive.*

- X6 y Y: se evidencia que mayores valores de longitud implican mayores valores del inmueble, por lo que vivir al oriente es más caro. / - *X6 and Y: it is evident that higher longitude values imply higher property values, so living in the east is more expensive.*

Section 2. Create a linear model that allows predicting the response variable from the characteristics. In your report summarize and comment on: / *Cree un modelo lineal que permita predecir la variable de respuesta a partir de las características. En su reporte resuma y comente:*

(a) Metrics of the model using training data. / *Métricas del modelo usando datos de entrenamiento.*

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:

Intercept: -12796.118

Linear Regression Coefficients:

[('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. / *Métricas del modelo usando validación cruzada.*

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 (using Statsmodels) / *Evaluación del modelo y sus parámetros empleando pruebas estadísticas.*

(i) 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. /

El R^2 es de 0,543, lo que indica que el modelo explica aproximadamente el 54,3% de la varianza. El R^2 ajustado, que tiene en cuenta el número de variables predictoras, es de 0,534. El modelo global es significativo, ya que tiene un elevado estadístico F de 60 y un valor p muy bajo, 1,05e-48, muy por debajo de 0,05. Todas las variables, excepto la longitud X6, tienen valores p inferiores a 0,05, lo que significa que valdría la pena descartar esta variable.

(ii) 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. *El valor del R2 ajustado incrementó un poco a 0.535, significando que este modelo tiene ligeramente un mejor ajuste.*