Taller 3 - Modelos de aprendizaje en Python

Jerónimo Vargas Rendon – 202113305

Carlos Gómez – 202111593

1. Adjunto a este taller encontrar un archivo CSV con datos sobre propiedades, sus características y su valor por metro cuadrado.

2. Realice un análisis exploratorio y resuma (comentarios breves, precisos, enumerados) en su reporte:

a) Comportamiento individual de cada característica y de la variable de respuesta.

Tabla 1: Estadísticas descriptivas de las variables y la variable de respuesta.

```
data.describe()
```

	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

Comportamiento individual de cada variable:

X1 Transaction date: Las fechas de transacción o el día en el que se realiza la valorización del inmueble están concentradas en el año 2013, con un promedio de 2013.15. El rango de esta variable varía entre finales del 2012 y 2013, lo que sugiere que las transacciones se realizaron relativamente en un corto periodo de tiempo. Por esto, la desviación de esta variable no es alta.

X2 House age: La edad de las casas en los datos recopilados varían considerablemente, con una media de 17.71 y una desviación de 11.39. El rango de esta variable es amplio, donde se pueden ver casas construidas el mismo año y casas construidas hace casi 44 años. Por otro lado, el 50% de las casas tienen una edad de hasta 16 años, mientras que el 75% de ellas tienen una antigüedad de hasta 28.15 años.

X3 Distance to the nearest MRT station: La distancia a la estación de transporte publico más cercana varía considerablemente, con un promedio de 1083.89 metros. El rango va desde 23.38 metros hasta más de 6 kilómetros, es por eso que la desviación es relativamente alta con un valor de 1262.11. Esto indica una amplia dispersión en la accesibilidad de las estaciones de MRT.

X4 Number of convenience stores: El número de tiendas de conveniencia cercanas al inmueble varía entre 1 a 10, con una media de 4.09 tiendas. El 50% de las casas tienen al menos 4 tiendas de conveniencia en las cercanías, lo que sugiere que la mayoría de las áreas tienen un acceso razonable a estos servicios.

X5 Latitude: La latitud tiene un rango estrecho, con una media de 24.96903 grados. La latitud mínima y máxima varía solo ligeramente, lo que indica que todas las propiedades están ubicadas dentro de una región geográfica limitada en términos de latitud.

X6 Longitude: Similar a la latitud, la longitud tiene una variación pequeña, con un promedio de 121.533361 grados. Esto nuevamente sugiere que las propiedades están dentro de una zona geográfica específica.

Y House price of unit area: El precio por unidad de área de las casas varía significativamente, con un promedio de 37.98. El rango va desde 7.6 hasta 117.5, lo que indica una gran variabilidad en los precios, posiblemente debido a factores como la ubicación, la proximidad a estaciones de MRT, y otros servicios cercanos.

b) Correlaciones entre características y con la variable de respuesta.

Gráfico 1: Correlación entre variables del modelo de regresión.

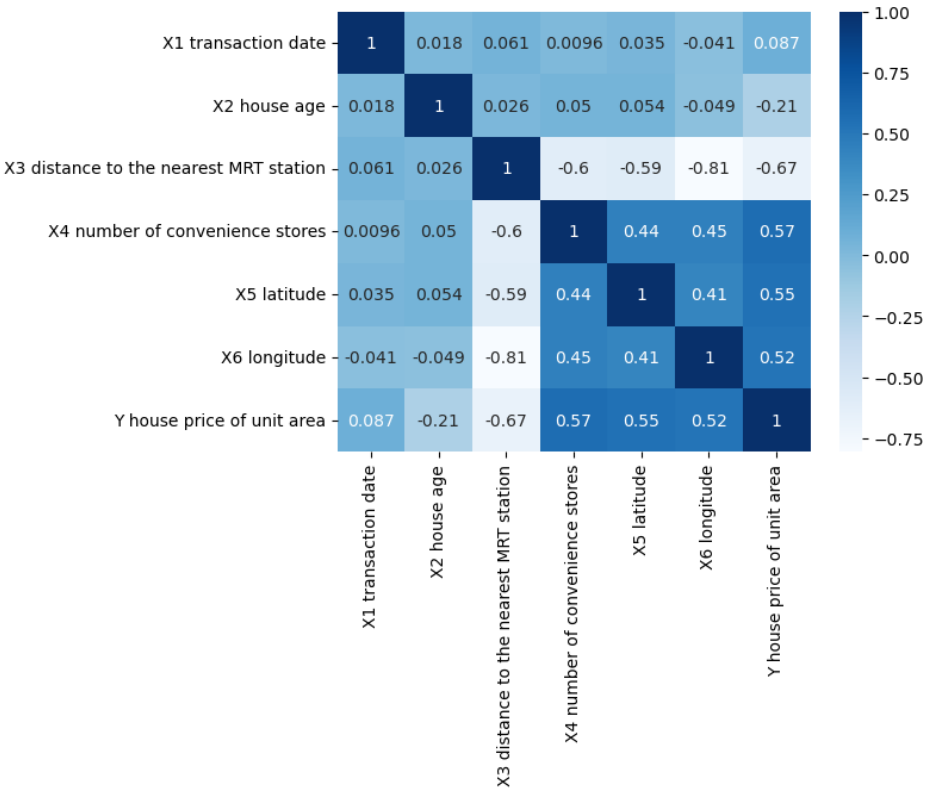
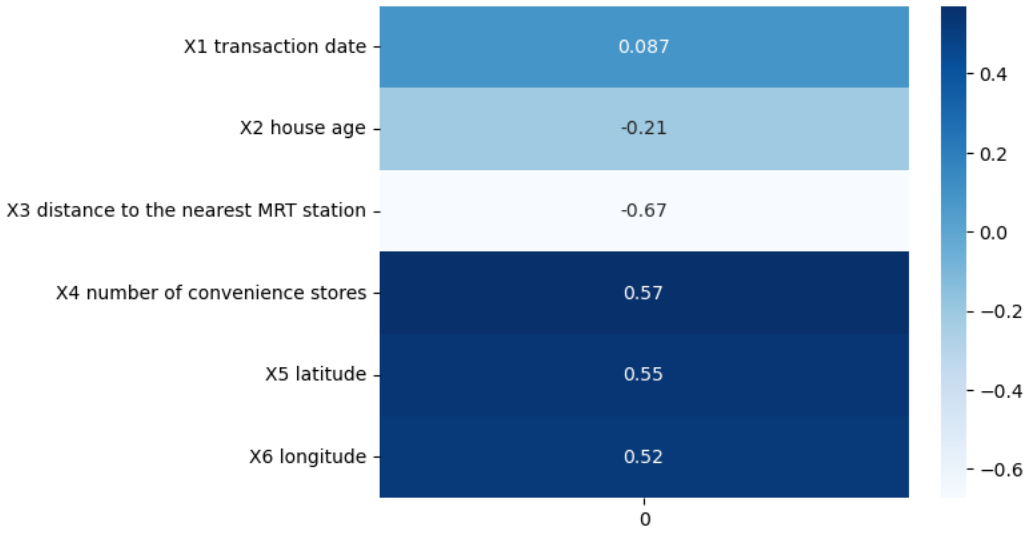


Gráfico 2: Correlación entre variables y variable de respuesta del modelo de regresión.



Análisis de la correlación entre el comportamiento individual de cada característica y la variable de respuesta:

X1 Transaction date: La correlación de X1 con la variable de respuesta Y es positiva pero muy baja (0.087). Esto indica que la fecha de transacción tiene un efecto mínimo sobre el precio de la unidad de área, sugiriendo que otros factores son más determinantes en la variación de los precios.

X2 House age: X2 tiene una correlación negativa moderada con el precio de la unidad de área (Y), con un valor de -0.21. Esto significa que a medida que la edad de la casa aumenta, es probable que el precio disminuya, lo cual es coherente con la depreciación de las propiedades más antiguas.

X3 Distance to the nearest MRT station: La variable X3 muestra una fuerte correlación negativa con la variable de respuesta (Y), con un valor de -0.67. Esto indica que a mayor distancia de la estación de MRT, menor es el precio de la unidad de área. La proximidad al transporte público es un factor importante en la valoración de las propiedades.

X4 Number of convenience stores: La variable sobre el número de tiendas de conveniencia cercanas al inmueble presenta una correlación positiva significativa con el precio de la unidad de área (Y), con un valor de 0.57. Esto sugiere que la disponibilidad de tiendas de conveniencia cercanas tiene un impacto considerable en el valor de las propiedades, haciendo que los precios tiendan a ser más altos en áreas con más servicios.

X5 Latitude: La latitud tiene una correlación positiva moderada con el precio de la unidad de área (Y), con un valor de 0.55. Esto indica que las propiedades situadas en latitudes más altas dentro del área de estudio tienden a ser más caras, aunque este efecto es menos determinante en comparación con otras variables.

X6 Longitude: La correlación entre X6 y el precio de la unidad de área (Y) es también positiva, con un valor de 0.52. Esto sugiere que las propiedades situadas más al este son más valiosas.

c) Exploración bivariada entre cada característica y la variable de respuesta.

Tabla 2: Diagramas de dispersión entre características y precio por unidad de area.

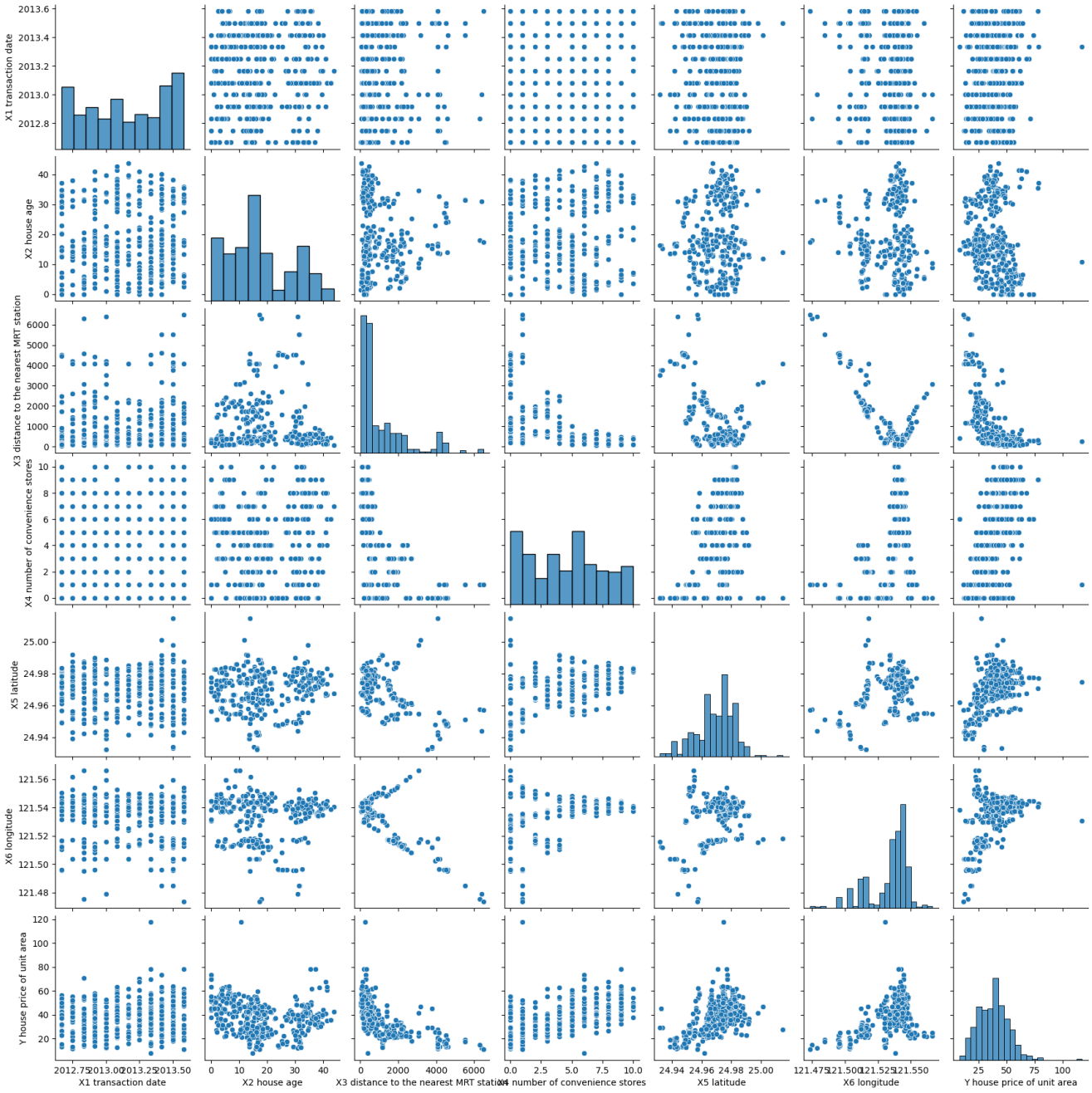


Gráfico 3: Regresión lineal y Dispersión para X1, X2 y X3 con respecto a Y.

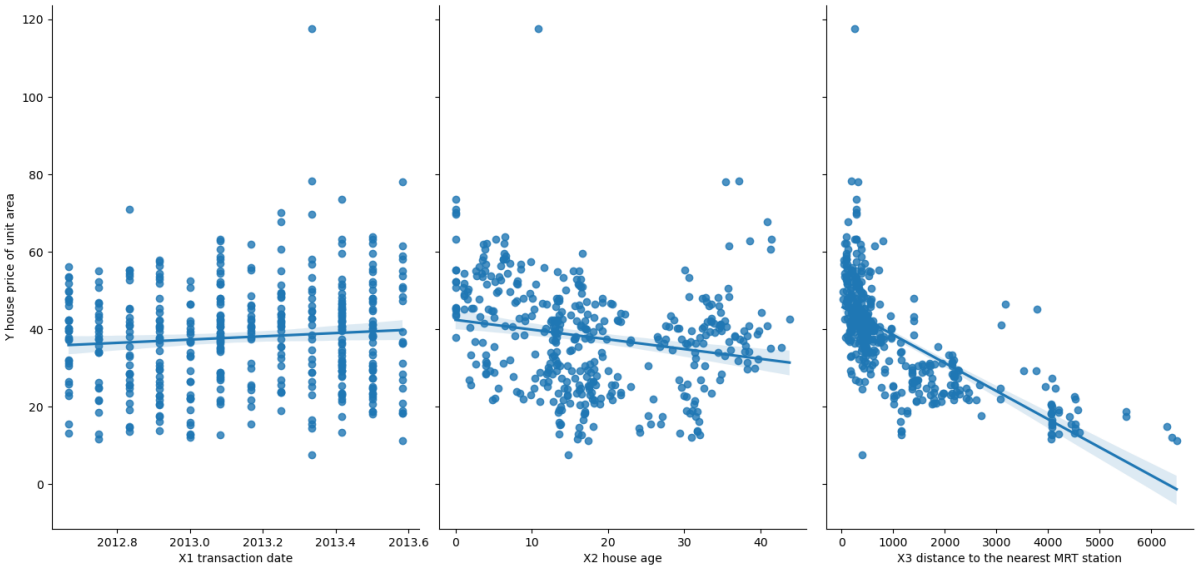
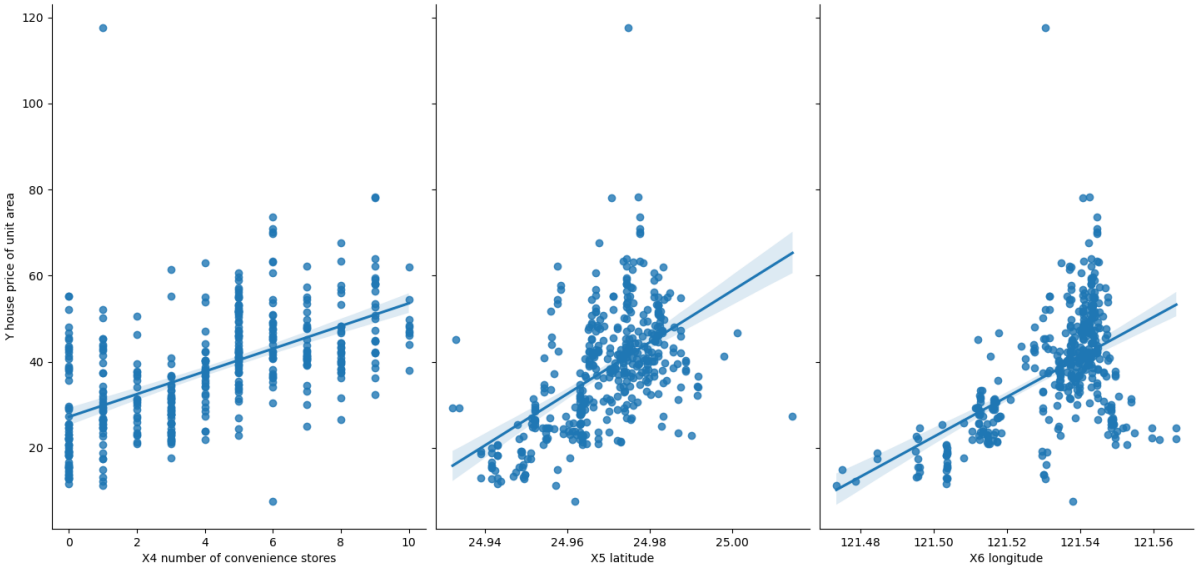


Gráfico 4: Regresión lineal y Dispersión para X4, X5 y X6 con respecto a Y.



Análisis del comportamiento individual de cada característica y la variable de respuesta:

X1 Transaction date: La variable X1 muestra un patrón ligeramente ascendente en relación con el precio de la unidad de área (Y). Esto sugiere que las propiedades vendidas más recientes tienden a tener precios marginales más altos. Sin embargo, la dispersión de los puntos indica que este efecto no es muy fuerte, sugiriendo que la fecha de transacción no es un determinante significativo del precio.

X2 House age: La variable X2 tiene una relación inversa con el precio de la unidad de área (Y). En general, a medida que la edad de la casa aumenta, el precio tiende a disminuir. Esto es consistente con la expectativa de que las casas más nuevas, que pueden requerir menos mantenimiento o tener características más modernas, suelen ser más valiosas.

X3 Distance to the nearest MRT station: La variable X3 muestra una fuerte relación negativa con el precio de la unidad de área (Y). Las propiedades ubicadas más lejos de la estación de MRT tienden a tener precios significativamente más bajos. Esto sugiere que la proximidad al transporte público es un factor importante en la valoración de las propiedades, con las ubicaciones más accesibles siendo más costosas.

X4 Number of convenience stores: La variable X4 presenta una relación positiva con el precio de la unidad de área (Y). A medida que aumenta el número de tiendas de conveniencia cercanas, el precio de la propiedad también tiende a aumentar, es decir, una relación proporcional. Este comportamiento es esperado, ya que una mayor cantidad de servicios cercanos hace que una propiedad sea más atractiva para los compradores.

X5 Latitude: La variable X5 muestra un patrón ascendente con respecto al precio de la unidad de área (Y). Esto podría indicar que las propiedades ubicadas en latitudes más altas dentro del área de estudio son más valiosas. Este comportamiento puede estar relacionado con la percepción de las ubicaciones o con características específicas de esas áreas.

X6 Longitude: La variable X6 también tiene una relación positiva con el precio de la unidad de área (Y). Similar a la latitud, las propiedades situadas más al este parecen ser más valiosas. La dispersión de los puntos sugiere que esta relación es significativa, aunque puede estar influenciada por otros factores geográficos o socioeconómicos no contemplados directamente en el modelo.

Y House price of unit area: La variable Y es la variable de respuesta en este análisis y muestra una considerable variabilidad, lo que indica que hay múltiples factores que influyen en el precio de las propiedades. Las gráficas de dispersión con cada variable independiente sugieren que algunos factores, como la proximidad a la estación de MRT y el número de tiendas cercanas, tienen una influencia más directa en los precios, mientras que otros, como la edad de la casa y las coordenadas geográficas, también desempeñan un papel, pero con diferentes grados de impacto.

3. Cree un modelo lineal que permita predecir la variable de respuesta a partir de las características. En su reporte resuma y comente:

A continuación, se presenta el primer modelo para poder predecir la variable de interés.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

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

Coeficientes:

B0:	-12796.1176
B1 (X1 transaction date):	5.7171
B2 (X2 house age):	0.2493
B3 (X3 distance to the nearest MRT station):	-0.00493
B4 (X4 number of convenience stores):	1.0761
B5 (X5 latitude):	227.0371
B6 (X6 longitude):	-35.6988

Lo anteriores coeficiente reflejan la proporción en la que cambia la variable de interés de acuerdo con un cambio en determinada variable.

En cuanto a las métricas del modelo usando los datos de entrenamiento obtuvimos lo siguiente:

Las métricas utilizadas para evaluar el rendimiento del modelo de regresión lineal proporcionan una visión detallada de la precisión y calidad del ajuste. El Mean Absolute Error (MAE), que es de 5.34, mide el error promedio entre los valores predichos y los valores reales, indicando que, en promedio, las predicciones del modelo se desvían de los valores observados en aproximadamente 5.34 unidades monetarias. El Mean Squared Error (MSE), que es de 45.01, calcula el promedio de los errores al cuadrado, penalizando más los errores grandes y sugiriendo que el modelo tiene un error cuadrático medio relativamente bajo. El Root Mean Squared Error (RMSE), que es de 6.71, toma la raíz cuadrada del MSE y devuelve la métrica a la misma unidad que la variable objetivo, mostrando que el error típico en las predicciones es de alrededor de 6.71 unidades monetarias. Por último, el R-squared (R^2), con un valor de 0.703, indica que el modelo explica el 70.3% de la varianza en la variable dependiente, lo que sugiere un ajuste razonablemente bueno. En conjunto, estas métricas muestran que el modelo tiene una precisión decente y explica una parte significativa de la variabilidad en los datos, aunque podría haber margen para mejorar aún más la precisión de las predicciones.

A continuación, se realizo la validación cruzada a al modelo utilizando un enfoque de validación cruzada de 5 folds. En este proceso, el conjunto de datos se dividió en 5 subconjuntos de tamaño similar. Para cada iteración, el modelo se entrenó utilizando 4 de estos subconjuntos y se evaluó en el subconjunto restante, repitiendo este procedimiento 5 veces, cada vez con un subconjunto diferente como conjunto de prueba. Los errores del modelo en cada iteración se midieron utilizando el Mean Squared Error (MSE), obteniendo los siguientes valores: 49.90, 89.03, 57.87, 134.82, y 60.05. Posteriormente, se calculó el Root Mean Squared Error (RMSE) para cada fold, resultando en 7.06, 9.44, 7.61, 11.61, y 7.75, lo que permitió obtener una métrica de error en la misma unidad que la variable objetivo. Finalmente, se promedió el RMSE a lo largo de las 5 iteraciones, obteniendo un valor promedio de 8.69.

De los resultados obtenidos indican que el modelo presenta un rendimiento razonablemente bueno, con un error típico de aproximadamente 8.69 unidades. Sin embargo, la variabilidad en los valores de MSE y RMSE entre los diferentes folds sugiere que el modelo tiene un desempeño ligeramente inconsistente dependiendo del subconjunto de datos en el que se evalúe. Esto podría señalar que, si bien el modelo es adecuado en general, aún podría beneficiarse de ajustes adicionales o mejoras para reducir la variabilidad y mejorar su precisión general

Evaluación del modelo y sus parámetros empleando pruebas estadísticas.

OLS Regression Results

Dep. Variable:	Y house price of unit area	R-squared (uncentered):	0.948
Model:	OLS	Adj. R-squared (uncentered):	0.947
Method:	Least Squares	F-statistic:	933.0
Date:	Sun, 25 Aug 2024	Prob (F-statistic):	1.70e-192
Time:	21:44:21	Log-Likelihood:	-1129.8
No. Observations:	310	AIC:	2272.
Df Residuals:	304	BIC:	2294.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
X1 transaction date	4.0106	1.689	2.374	0.018	0.687	7.335
X2 house age	-0.2408	0.047	-5.172	0.000	-0.332	-0.149
X3 distance to the nearest MRT station	-0.0056	0.001	-8.561	0.000	-0.007	-0.004
X4 number of convenience stores	1.0545	0.231	4.561	0.000	0.600	1.509
X5 latitude	199.8353	50.836	3.931	0.000	99.800	299.871
X6 longitude	-107.1264	27.916	-3.838	0.000	-162.059	-52.194

Omnibus:	184.800	Durbin-Watson:	2.117
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2744.014
Skew:	2.124	Prob(JB):	0.00
Kurtosis:	16.943	Cond. No.	2.28e+05

De la anterior evaluación del modelo se puede evidenciar como se realizan pruebas estadísticas como: La prueba t de coeficientes y Prueba F. Además, se hace una diagnostico con los test de Durbin-Watson, Prueba de Omnibus y Jarque-Bera, y Número de Condición.

Evaluación del modelo en general:

- **R-squared (R^2) sin centrar**: Con un valor de **0.948**, el R^2 sin centrar indica que el modelo es capaz de explicar el 94.8% de la variabilidad en el precio por unidad de área de las casas. Esto sugiere que el modelo tiene un ajuste fuerte y es eficaz en capturar las relaciones entre las variables independientes y la variable dependiente.
- **F-statistic y su p-valor**: El F-statistic de 933.0 y el p-valor extremadamente bajo **(1.70e-192)** confirman que el modelo es globalmente significativo. Esto significa que

las variables independientes, cuando se consideran juntas, tienen un impacto estadísticamente significativo en el precio por unidad de área.

Significancia de los Coeficientes (Prueba t):

De los resultados se evidencia que todos los coeficientes de las respectivas variables del modelo son estadísticamente significativos con un nivel de significancia del 5%, según la prueba t. Dado que los p-valores de todos los coeficientes fueron menores al 5%, se rechaza la hipótesis nula, la cual indica que el coeficiente es igual a 0. Esto confirma que cada uno de los coeficientes es significativamente diferente de 0.

Diagnóstico del Modelo:

- **Prueba de Omnibus y Jarque-Bera:** Los altos valores de estas pruebas (Omnibus = 184.800 y Jarque-Bera = 2744.014) y los p-valores de 0.000 indican que los residuos del modelo no siguen una distribución normal. Esto podría afectar la validez de las inferencias y sugiere la necesidad de considerar transformaciones o el uso de métodos robustos.
- **Durbin-Watson:** El valor de 2.117 indica que no hay una fuerte autocorrelación en los residuos, lo que sugiere que los errores son independientes y apoya la validez de las inferencias.
- **Condición del número:** El valor alto de 2.28e+05 sugiere posibles problemas de multicolinealidad, lo que indica que algunas variables independientes podrían estar altamente correlacionadas entre sí. Esto podría hacer que las estimaciones de los coeficientes sean inestables, aunque no necesariamente invalidas.

4. Incluya todo el código de exploración y análisis como soporte.

Adjunto a la entrega está el código de exploración y análisis con el nombre de “Taller3.ipynb”