Taller 3 de Analítica Exploración y Regresión en Python

Isabela Castillo 201813093

Carlos Molano 201922691

Sección 1:

1. Exploración y regresión en Python

1. Adjunto a este taller encontrará 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.
   b) Correlaciones entre características y con la variable de respuesta.
   c) Exploración bivariada entre cada característica y la variable de respuesta.
3. Cree un modelo lineal que permita predecir la variable de respuesta a partir de las características. En su **reporte** resuma y comente:
   a) Métricas del modelo usando datos de entrenamiento.
   b) Métricas del modelo usando validación cruzada.
   c) Evaluación del modelo y sus parámetros empleando pruebas estadísticas.
4. Incluya todo el código de exploración y análisis como **soporte**.

Punto 2

Literal 2a:

	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

El conjunto de datos presenta un resumen estadístico de varias características de las propiedades (X1 a X6) y la variable de respuesta, que es el precio por unidad de área (Y). Este análisis exploratorio tiene como objetivo comprender el comportamiento individual de cada característica y su posible influencia en la variable de respuesta.

La fecha de transacción (X1) oscila entre los años 2012.67 y 2013.58, con un promedio en 2013.15. Esto indica que la mayoría de las transacciones se realizaron en el año 2013. La baja desviación estándar (0.28) sugiere que las fechas de transacción están bastante concentradas alrededor de este año, lo que podría implicar un mercado activo durante ese período.

La edad de las casas (X2) varía considerablemente, desde casas nuevas (0 años) hasta propiedades con más de 40 años de antigüedad, con una media de 17.71 años. La amplia desviación estándar (11.39) muestra que hay una gran diversidad en la antigüedad de las casas, lo cual podría influir en la percepción del valor de estas propiedades por parte de los compradores.

En cuanto a la distancia a la estación de MRT más cercana (X3), se observa un amplio rango de entre 23.38 metros y 6488.02 metros. La media es de 1083.89 metros, pero la elevada desviación estándar (1262.11) indica que las distancias están muy dispersas, lo que sugiere que la proximidad al transporte público varía significativamente entre las propiedades, un factor que podría ser crucial en la determinación del precio de estas.

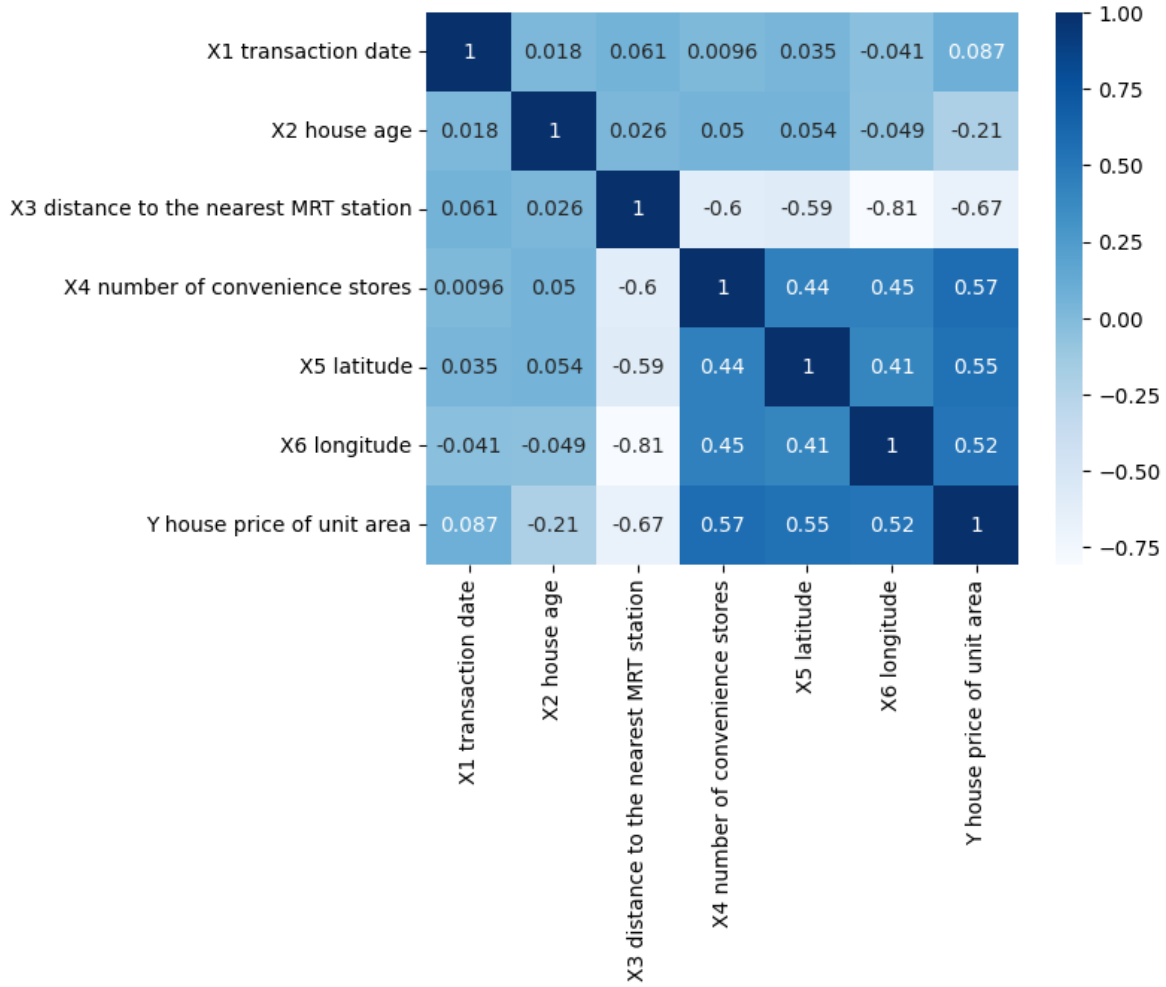
El número de tiendas de conveniencia cercanas (X4) varía entre ninguna y hasta 10 tiendas, con un promedio de 4.09 tiendas por propiedad. La desviación estándar (2.95) revela que existe una variación significativa en la cantidad de tiendas cercanas, lo cual podría ser un indicador importante del nivel de accesibilidad y conveniencia de la ubicación.

La latitud (X5) y longitud (X6) de las propiedades muestran un rango estrecho, lo que sugiere que todas las propiedades están ubicadas dentro de una región geográficamente limitada. La latitud varía entre 24.93 y 25.01, mientras que la longitud oscila entre 121.47 y 121.57. Ambas variables tienen desviaciones estándar muy bajas (0.01), lo que implica que, geográficamente, las propiedades están concentradas en un área pequeña.

Finalmente, la variable de respuesta, el precio por unidad de área (Y), presenta un rango amplio que va desde 7.6 hasta 117.5, con una media de 37.98. La desviación estándar (13.61) sugiere una considerable variabilidad en los precios, probablemente influenciada por las características mencionadas anteriormente, como la proximidad al transporte público, la edad de la casa y la accesibilidad a servicios.

Este análisis revela que las propiedades en el conjunto de datos son diversas en términos de antigüedad, proximidad a estaciones de MRT y número de tiendas cercanas, lo cual podría estar contribuyendo a la variabilidad observada en los precios por unidad de área.

Literal 2b:



El análisis de las correlaciones entre las características (variables X) y la variable de respuesta (Y) revela algunas relaciones significativas. La fecha de transacción (X1) muestra una correlación casi nula con la edad de la casa (X2), lo que indica que el año en que se realizó la transacción no está significativamente relacionado con la antigüedad de la propiedad. Sin embargo, la correlación entre X1 y el precio de la unidad de área (Y) es débil y positiva (0.21), lo que sugiere que las propiedades vendidas en fechas más recientes tienden a tener precios ligeramente más altos.

La edad de la casa (X2) tiene una correlación moderada (0.26) con la distancia a la estación de MRT más cercana (X3), lo que implica que las casas más antiguas tienden a estar más alejadas de las estaciones de MRT. Además, la edad de la casa tiene una correlación negativa y débil (-0.21) con el precio (Y), sugiriendo que las propiedades más antiguas tienden a tener precios ligeramente más bajos.

En cuanto a la distancia a la estación de MRT más cercana (X3), existe una correlación moderadamente fuerte y negativa (-0.67) con el precio de la unidad de área (Y). Esto indica que las propiedades más cercanas a las estaciones de MRT tienden a tener precios

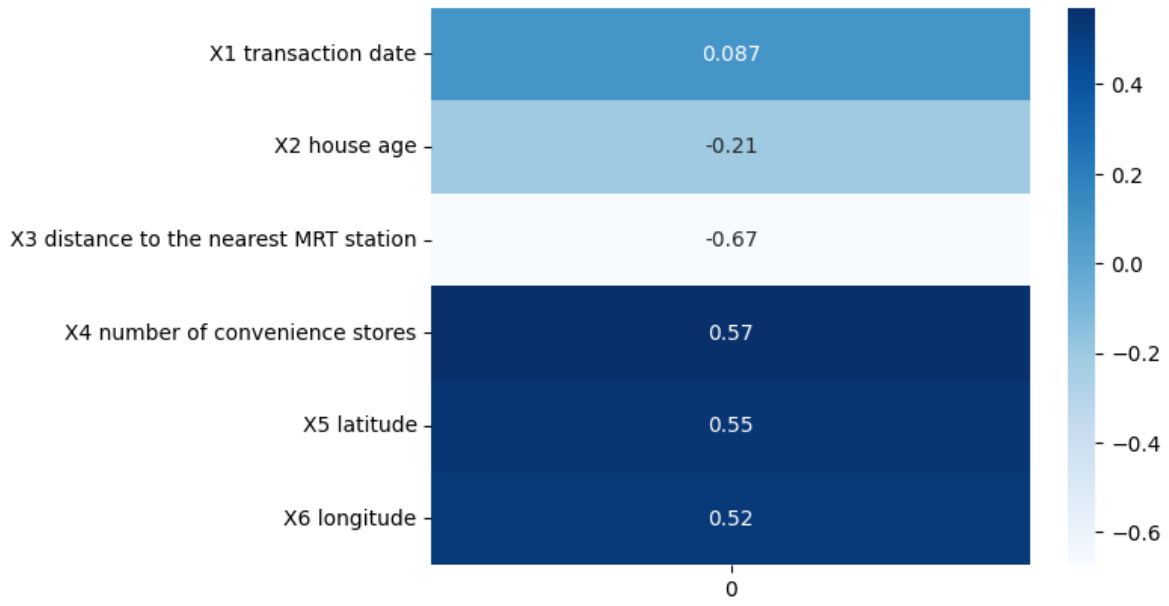
significativamente más altos, lo cual subraya la importancia de la proximidad al transporte público en la valorización de las propiedades. Sin embargo, la correlación entre la distancia a la MRT y el número de tiendas de conveniencia (X4) es de (-0.6), lo que sugiere que la disponibilidad de tiendas está directamente relacionada con la proximidad al MRT.

El número de tiendas de conveniencia (X4) muestra una correlación positiva (0.57) con el precio de la unidad de área (Y), lo que sugiere que las propiedades cercanas a más tiendas tienden a ser más valiosas. Sin embargo, las correlaciones entre X4 y las coordenadas geográficas (latitud (X5) y longitud (X6)) son positivas, lo que indica que puede existir una relación con la ubicación específica en términos de latitud y longitud.

Finalmente, las coordenadas geográficas muestran correlaciones moderadas con el precio de la unidad de área (Y). La latitud (X5) tiene una correlación positiva (0.55), y la longitud (X6) también tiene una correlación positiva (0.52) con el precio. Esto sugiere que ciertas ubicaciones geográficas específicas dentro del área de estudio están asociadas con precios más altos de las propiedades.

Las relaciones más relevantes se encuentran entre la distancia al MRT y el precio de la propiedad, así como las coordenadas geográficas, las tiendas de conveniencia y el precio. La antigüedad de la casa y la fecha de transacción influyen en menor medida sobre el valor de las propiedades.

Literal 2c:



El análisis bivariado entre cada característica (Variables X) y la variable de respuesta (Y) revela varias relaciones significativas que pueden ayudar a comprender cómo estas características individuales afectan el precio de la unidad de área (Y).

Primero, la **fecha de transacción** (X1) muestra una correlación mínima y positiva (0.087) con el precio de las propiedades (Y). Esto sugiere que las transacciones realizadas en fechas más recientes o en fechas más antiguas no tienen un mayor impacto con respecto al comportamiento de los precios.

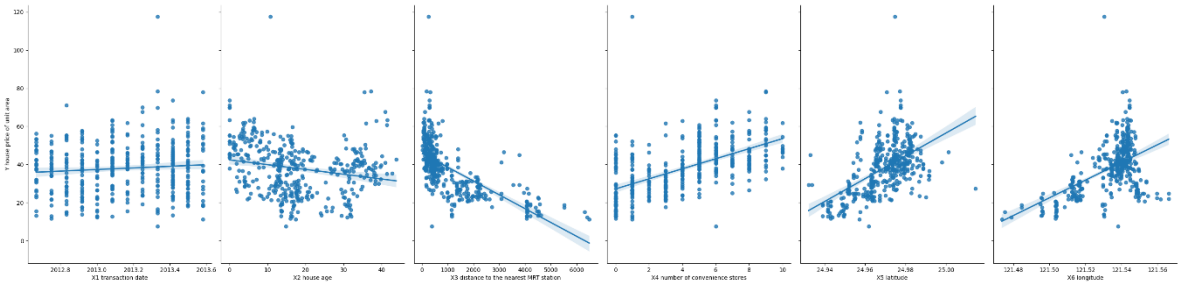
La **edad de la casa** (X2), por otro lado, tiene una correlación débil y negativa (-0.21) con el precio. Esto indica que las casas más antiguas suelen tener precios ligeramente más bajos, lo que es consistente con la expectativa de que las propiedades más nuevas, o en mejor estado, podrían tener un valor mayor en el mercado.

La **distancia a la estación de MRT más cercana** (X3) presenta una correlación moderadamente fuerte y negativa (-0.67) con el precio de la unidad de área. Este es un hallazgo significativo, ya que subraya la importancia de la proximidad al transporte público en la valoración de las propiedades. Las propiedades más cercanas a una estación de MRT tienden a ser considerablemente más caras, lo que refleja la alta demanda por la conveniencia de estar cerca del transporte público.

El **número de tiendas de conveniencia** cercanas (X4) muestra una correlación moderada y positiva (0.57) con el precio de las propiedades. Esto sugiere que la disponibilidad de tiendas de conveniencia en las cercanías es un factor importante que puede aumentar el valor de una propiedad, posiblemente debido a la comodidad adicional que estas tiendas ofrecen a los residentes.

Las coordenadas geográficas, tanto la **latitud** (X5) como la **longitud** (X6), también están moderadamente correlacionadas de manera positiva con el precio de la unidad de área, con coeficientes de 0.55 y 0.52, respectivamente. Esto indica que la ubicación geográfica específica dentro del área de estudio es un factor importante en la determinación del precio, con ciertas áreas posiblemente siendo más deseables o de mayor valor.

El análisis bivariado muestra que las características relacionadas con la ubicación, como la proximidad al transporte público, la cantidad de tiendas cercanas y la ubicación geográfica específica, son factores clave en la determinación del precio de las propiedades. Mientras tanto, la antigüedad de la casa tiene un impacto negativo, y las transacciones recientes parecen no estar asociadas con el comportamiento de los precios.



Con base en los gráficos de dispersión para las diferentes características (X1 a X6) en relación con el precio de la unidad de área (Y), se pueden hacer varias observaciones clave.

Primero, la dispersión de los datos en función de la **fecha de transacción** (X1) sugiere una ligera tendencia ascendente en los precios a lo largo del tiempo. Esto indica que las propiedades más recientemente vendidas tienden a ser más caras, lo que podría reflejar una apreciación del mercado inmobiliario. Además, la homogeneidad en la dispersión de los datos sugiere que no hubo fluctuaciones extremas en los precios durante el periodo analizado.

En cuanto a la **edad de la casa** (X2), aunque hay una correlación negativa con el precio, la relación no parece ser estrictamente lineal. Algunas casas antiguas aún alcanzan precios elevados, posiblemente debido a factores como renovaciones o ubicaciones privilegiadas. Además, los precios de las casas más antiguas muestran una mayor variabilidad, lo que indica que la antigüedad no es el único factor que influye en su valor.

La **distancia a la estación de MRT más cercana** (X3) muestra una fuerte relación negativa con los precios. Los datos claramente indican que a medida que aumenta la distancia a la estación de MRT, los precios de las propiedades tienden a disminuir. Este patrón confirma la importancia del acceso al transporte público en la valoración de las propiedades. Además, se observan agrupaciones de precios más altos en distancias cortas a las estaciones, reforzando la idea de que la proximidad al transporte es un factor clave en el mercado inmobiliario.

El **número de tiendas de conveniencia** cercanas (X4) también tiene una relación positiva con los precios, aunque menos marcada que la distancia al MRT. Las propiedades en áreas con más tiendas de conveniencia tienden a tener precios más altos, lo que puede estar relacionado con la comodidad y accesibilidad que estas tiendas ofrecen a los residentes. Sin embargo, hay algunas excepciones, donde propiedades con pocas tiendas cercanas aún alcanzan precios elevados, lo que sugiere la influencia de otros factores.

En cuanto a las coordenadas geográficas, tanto la **latitud** (X5) como la **longitud** (X6) muestran tendencias positivas en relación con los precios. Esto indica que la ubicación exacta dentro de la región de estudio es un factor importante en la determinación del valor de las propiedades. La latitud parece estar ligeramente más correlacionada con los precios que la longitud, aunque ambas variables muestran agrupaciones de propiedades con precios altos, sugiriendo zonas geográficas específicas de mayor valor.

Los gráficos destacan la importancia de múltiples factores en la determinación del precio de las propiedades (Variable Objetivo), incluyendo la proximidad al transporte público, la cantidad de tiendas de conveniencia cercanas, y la ubicación geográfica precisa. Aunque la antigüedad de la casa tiene un impacto negativo, su influencia es más compleja y está mediada por otros factores.

Punto 3

Se realizó la preparación y construcción de un modelo de regresión lineal para predecir el precio por unidad de área de una propiedad. Primero, se seleccionaron las características clave del dataset, como la fecha de transacción, la edad de la casa, la distancia a la estación de MRT más cercana, el número de tiendas de conveniencia cercanas, la latitud y la longitud, definiendo estas como las variables predictoras, mientras que la variable de respuesta fue el precio por unidad de área. Posteriormente, se dividieron los datos en conjuntos de entrenamiento y prueba utilizando la función *train_test_split,* donde el 80% de los datos se destinó al entrenamiento del modelo y el 20% se reservó para la evaluación. Con estos datos, se creó y ajustó un modelo de regresión lineal usando *LinearRegression* de la librería *sklearn*, lo que permitió al modelo aprender la relación entre las características y la variable de respuesta. Finalmente, se inspeccionaron los coeficientes obtenidos por el modelo, junto con la intersección, para entender la influencia de cada característica en la predicción del precio, identificando cuáles tienen un impacto positivo o negativo en la variable de respuesta.

Literal 3a:

La evaluación del modelo de regresión lineal se realizó siguiendo una metodología rigurosa que incluyó la división del dataset en conjuntos de entrenamiento y prueba, el uso de validación cruzada y la comparación de métricas clave como el *Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE), el R-cuadrado (R^2) y el Error Absoluto Medio (MAE)*. Esta metodología permitió no solo entrenar el modelo de manera efectiva, sino también evaluar su capacidad para generalizar a nuevos datos no vistos, lo cual es esencial para asegurar que el modelo sea robusto y aplicable en escenarios reales.

Los resultados de la evaluación en el conjunto de entrenamiento indicaron que el modelo fue capaz de capturar la relación entre las características y la variable de respuesta con un alto grado de precisión, como lo sugiere un R^2 relativamente alto y errores bajos (MSE, RMSE, MAE). Al evaluar el modelo en el conjunto de prueba, se observó que las métricas se mantuvieron en un rango similar a las obtenidas en el conjunto de entrenamiento, lo que refuerza la idea de que el modelo es sólido y capaz de predecir con precisión los precios por unidad de área en datos no vistos. Esta coherencia entre las métricas en ambos conjuntos demuestra que la metodología seleccionada para la evaluación fue adecuada y efectiva, permitiendo obtener un modelo equilibrado que no solo se ajusta bien a los datos con los que

fue entrenado, sino que también es capaz de ofrecer predicciones fiables en aplicaciones prácticas.

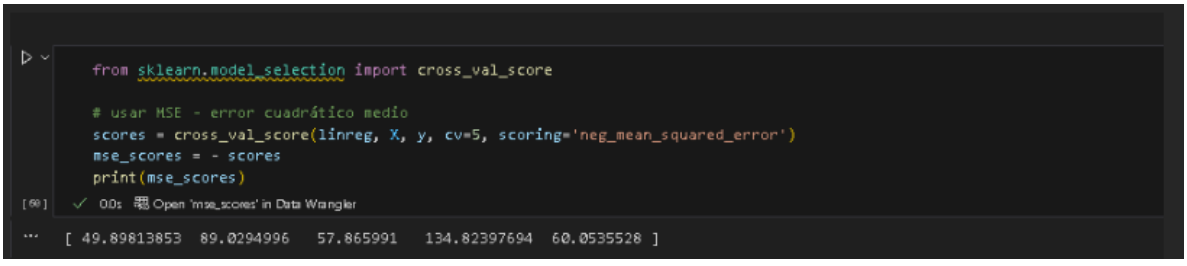
	Entrenamiento	Prueba
MAE	5.34 %	5.57%
MSE	45.01%	53.73%
RMSE	6.70%	7.33%

Punto 3b.

Sin embargo, para garantizar que este buen desempeño no fuera simplemente un reflejo de un sobreajuste a los datos de entrenamiento, se utilizó la validación cruzada, que mostró consistencia en los resultados a través de múltiples particiones del *dataset.* Esto sugiere que el modelo tiene una buena capacidad de generalización, minimizando el riesgo de sobreajuste.

En el análisis realizado, se empleó la validación cruzada k-fold con 5 particiones (folds) para evaluar la robustez y capacidad de generalización del modelo de regresión lineal. Este método dividió el conjunto de datos en cinco subconjuntos de igual tamaño, donde en cada iteración, el modelo fue entrenado con cuatro de estos subconjuntos y evaluado en el restante. El Error Cuadrático Medio (MSE) se calculó para cada fold, y posteriormente se promediaron estos resultados para obtener una medida representativa del error de predicción en datos no vistos. La consistencia en los valores de MSE entre los distintos folds indicó que el modelo tiene un desempeño estable y generaliza bien, lo que refuerza la confianza en su capacidad predictiva más allá del conjunto de datos de entrenamiento.

Validación cruzada (5 – folds) MSE



Validación cruzada (5 – folds) RMSE

```
# calcular RMSE
rmse_scores = np.sqrt(mse_scores)
print(rmse_scores)
```

[61] ✓ 0.0s Open 'rmse_scores' in Data Wrangler

```
[ 7.06386145  9.43554448  7.6069699  11.61137274  7.74942274]
```

Valores Promedio

```
# RMSE promedio a través de todos los grupos
print(rmse_scores.mean())
print(mse_scores.mean())
```

[60] ✓ 0.0s

```
8.693434260346503
78.33423177467634
```

Punto 3c:

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:	Sun, 25 Aug 2024	Prob (F-statistic):	1.05e-48
Time:	16:33:24	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

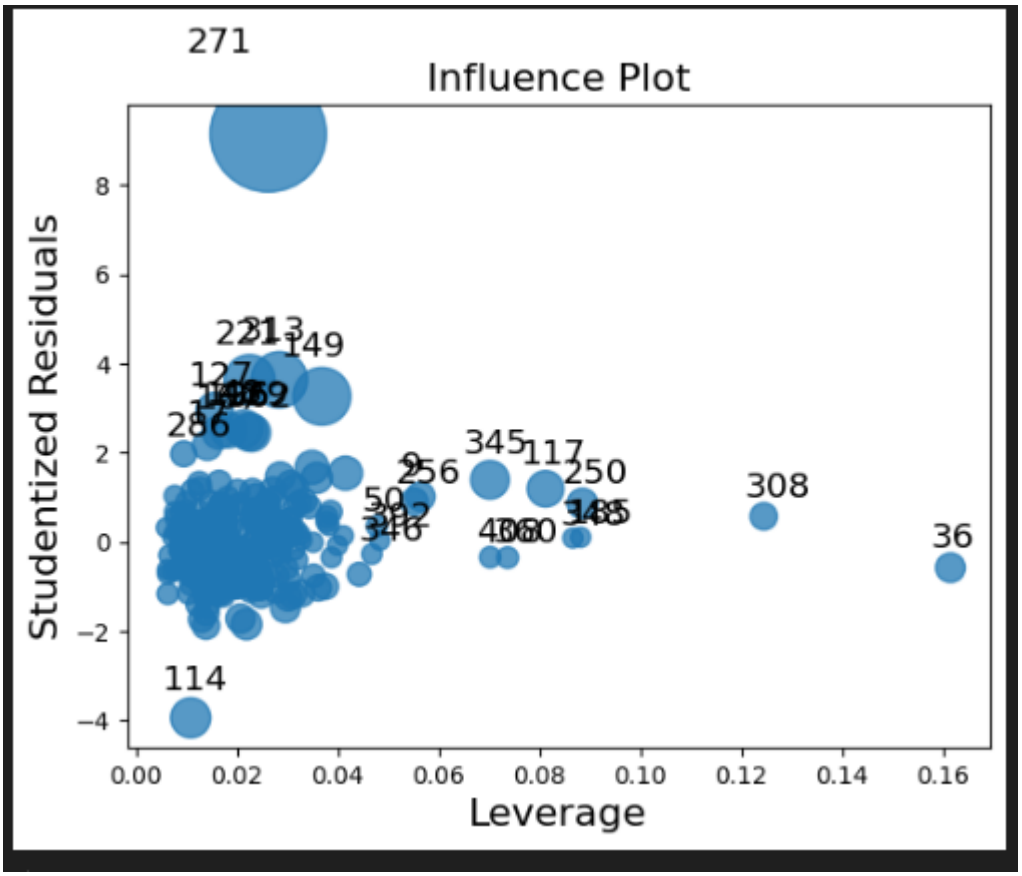
Omnibus:	189.462	Durbin-Watson:	2.086
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2953.563
Skew:	2.181	Prob(JB):	0.00

...

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.



Resultados del Modelo de Regresión:

- R-cuadrado (R^2): El modelo tiene un R^2 de 0.543, lo que significa que aproximadamente el 54.3% de la variabilidad en el precio por unidad de área es explicada por las variables incluidas en el modelo. Este valor sugiere que el modelo tiene un nivel moderado de capacidad explicativa.

Significancia de los Coeficientes:

- La mayoría de las variables, como `X1 transaction date` (p=0.007), `X2 house age` (p=0.000), y `X3 distance to the nearest MRT station` (p=0.000), tienen coeficientes que son estadísticamente significativos, lo que indica que estas variables tienen un impacto significativo en el precio por unidad de área.
- Sin embargo, variables como `X6 longitude` (p=0.512) no son estadísticamente significativas, lo que sugiere que su efecto sobre la variable dependiente no es concluyente en este modelo.
- Multicolinealidad: El estadístico de Durbin-Watson (2.086) sugiere que no hay problemas serios de autocorrelación en los residuos. Sin embargo, la nota sobre el número de condición ("condition number") elevado (3.79e+07) indica que podría

haber problemas de multicolinealidad entre las variables independientes, lo cual podría afectar la estabilidad y la interpretación de los coeficientes.

2. Gráfico de Influencia:

- El gráfico de influencia muestra los residuos estudentizados en relación con el apalancamiento de los datos.
- Puntos como el 271 se destacan significativamente, lo que sugiere que tienen un alto impacto en la estimación del modelo y podrían considerarse como puntos de influencia elevados.
- La presencia de estos puntos indica que ciertos datos podrían estar influyendo de manera desproporcionada en el modelo, y podría ser necesario analizarlos más a fondo para determinar si deberían ser excluidos o si requieren un tratamiento adicional.

3. Conclusiones:

- Aunque el modelo tiene un nivel moderado de capacidad explicativa (R^2=0.543), la presencia de multicolinealidad potencial y la influencia significativa de algunos puntos de datos sugiere que podría ser beneficioso refinar el modelo. Esto podría incluir la eliminación de variables no significativas, la transformación de variables para reducir la multicolinealidad, o la consideración de modelos alternativos que manejen mejor las características del conjunto de datos.

- Además, los puntos de influencia altos, como los identificados en el gráfico de influencia, deberían revisarse para garantizar que no estén sesgando indebidamente los resultados del modelo.

Link Repositorio: https://github.com/IsabelaCastillo/Analitica-Computacional/blob/main/Taller_3/Taller%203.ipynb