**Taller 3: Modelos de aprendizaje en Python**

**2. Realice un análisis exploratorio y resuma**

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

Inicialmente, se obtuvieron las estadísticas descriptivas de las variables explicativas y la variable de respuesta. Adicionalmente, se graficó el comportamiento de todas las variables como se observa a continuación.

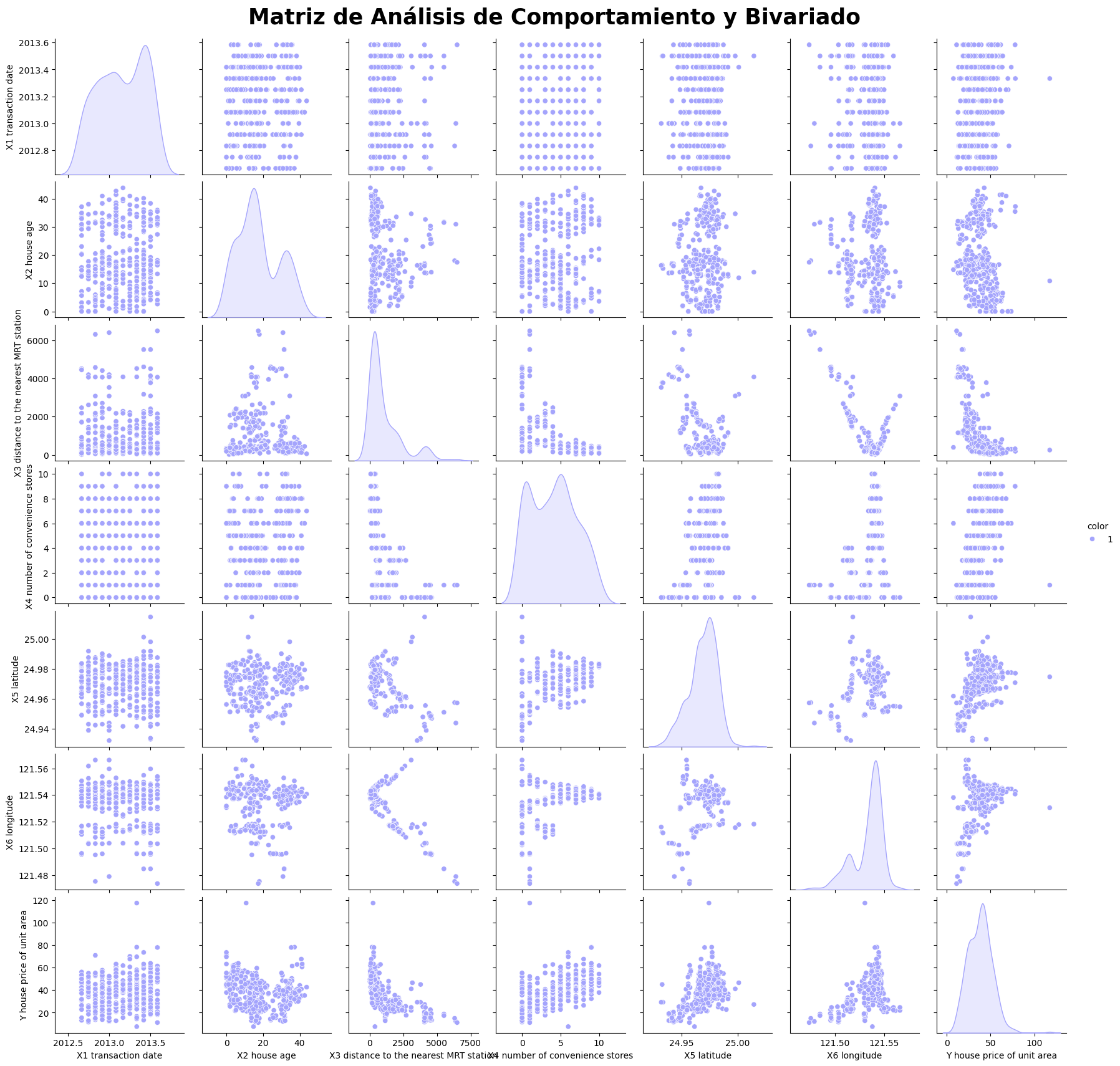


Figura 1. Matriz de Análisis de Comportamiento de Variables

X1: Transaction date

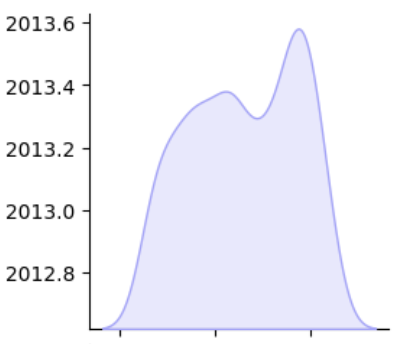


Figura 2. Grafica de Distribución de Fecha de Transacción

La media de la variable de fecha de transacción es de 2013.149, con una desviación estándar de 0.28. Adicionalmente, la variable tiene como valor mínimo 2012.667 y valor máximo 2013.583.

X2: House age

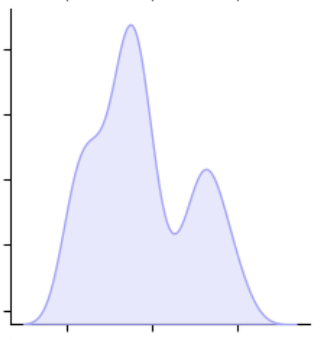


Figura 3. Grafica de Distribución de Antigüedad de Casa

La media de la variable de edad de la casa es 17.713, con una desviación estándar de 11.392. Los valores se encuentran entre 0 y 43.8.

X3: Distance to the nearest MRT station

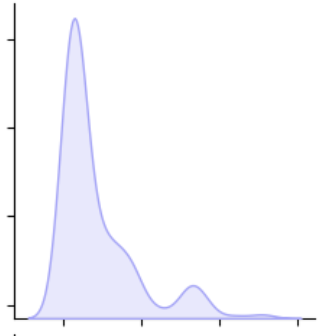


Figura 4. Grafica de Distribución de distancia a estación MRT

La media de la variable de distancia a estación de MRT más cercano es 1083.9 y su desviación estándar de 1262.110. Se puede observar que los valores se encuentran de esta variable se encuentran entre 23.4 como valor mínimo y 6488 como valor máximo. En el gráfico se puede observar que las mayores frecuencias se encuentran en los valores más bajos.

X4: Number of convenience stores

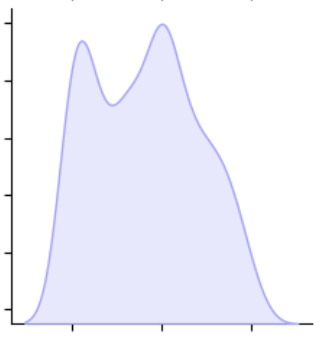


Figura 5. Grafica de Distribución de número de tiendas

La variable de número de tiendas de conveniencia tiene un promedio de 4.94 y una desviación estándar de 0.295. El número de tiendas de conveniencia se encuentra entre 0 y 10.

X5: Latitude

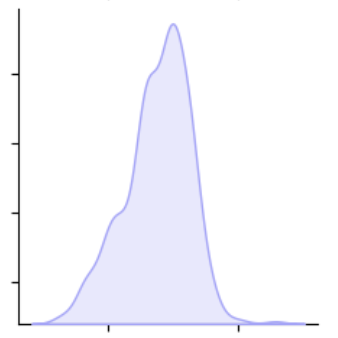


Figura 6. Grafica de Distribución de Latitud

La media de la variable de latitud es 24.969, tiene una desviación estándar de 0.012 y los valores se encuentran entre 24.93 y 25.01. Esto se puede evidenciar en la gráfica pues los datos se distribuyen alrededor de la media.

X6: Longitude

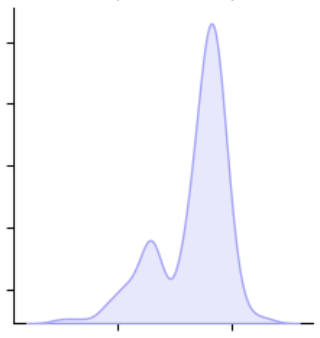


Figura 7. Grafica de Distribución de Longitud

La media de la variable de longitud es 121.5, tiene una desviación estándar de 0.015 y los valores se encuentran entre 121.5 y 121.6. Al igual que la variable de latitud los datos se distribuyen alrededor de la media.

Y: House price of unit area

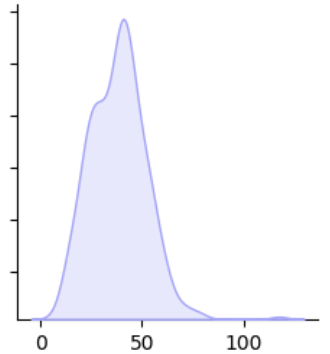


Figura 8. Grafica de Distribución Precio de Casa

La media de la variable de respuesta del precio de la casa por unidad de área es 37.98 y los valores se encuentran entre 7.6 y 117.5. La gráfica al igual que las métricas demuestra que el valor máximo se aleja de la frecuencia de los datos.

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

**Correlación entre características:**

X1 y X2

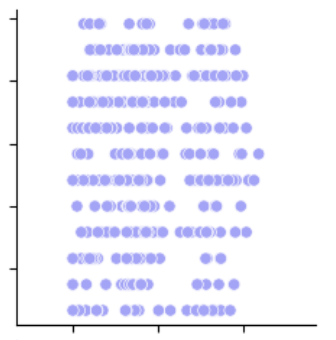


Figura 9. Grafica de Dispersión Transaction date vs House age

La correlación entre X1 y X2 es de 0.018, es positiva pero es débil.

X1 y X3

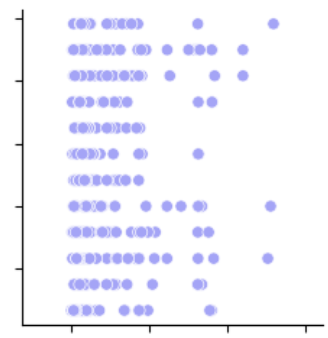


Figura 10. Grafica de Dispersión Transaction date vs Distance nearest MRT station

La correlación entre X1 y X3 es de 0.061, es positiva y débil.

X1 y X4

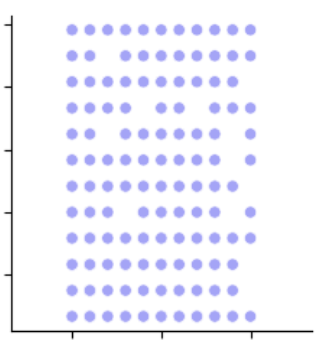


Figura 11. Grafica de Dispersión Transaction date vs Number convenience stores

La correlación entre X1 y X4 es de 0.0096, es positiva y débil.

X1 y X5

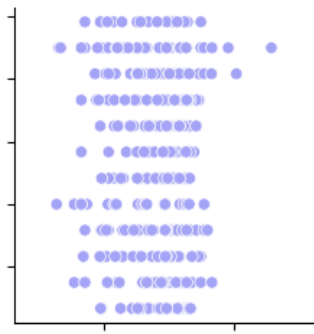


Figura 12. Grafica de Dispersión Transaction date vs Latitude

La correlación entre X1 y X5 es de 0.035, es positiva y débil.

X1 y X6

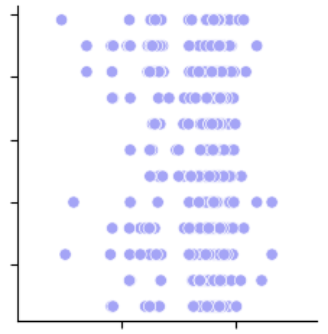


Figura 13. Grafica de Dispersión Transaction date vs Longitude

La correlación entre X1 y X6 es de -0.041, es negativa y débil.

X2 y X3

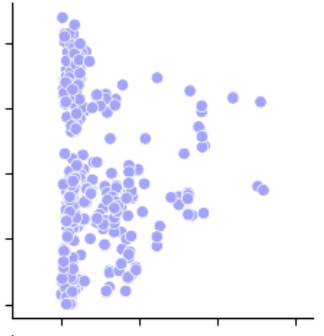


Figura 14. Grafica de Dispersión House Age vs Distance nearest MRT station

La correlación entre X2 y X3 es 0.026, es positiva y débil.

X2 y X4

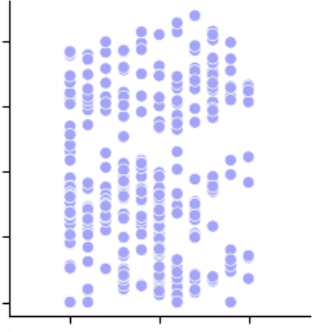


Figura 15. Grafica de Dispersión House Age vs Number of convenience stores

La correlación entre X2 y X4 es 0.05, es positiva y débil.

X2 y X5

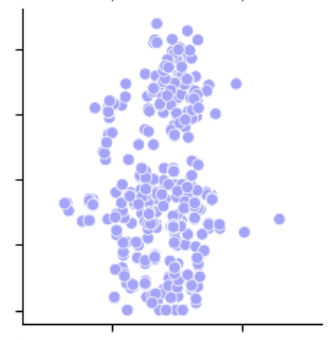


Figura 16. Grafica de Dispersión House Age vs Latitude

La correlación entre X2 y X5 es 0.054, es positiva y débil.

X2 y X6

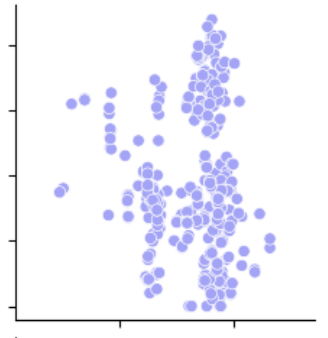


Figura 17. Grafica de Dispersión House Age vs Longitude

La correlación entre X2 y X6 es -0.049, es negativa y débil.

X3 y X4

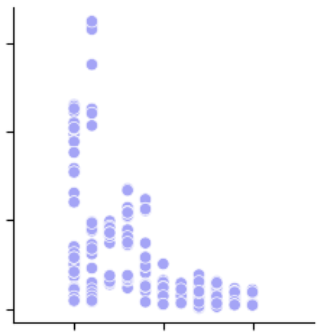


Figura 18. Grafica de Dispersión Distance nearest MRT station vs Number of Convenience Stores

La correlación entre X3 y X4 es -0.6, es negativa y débil.

X3 y X5

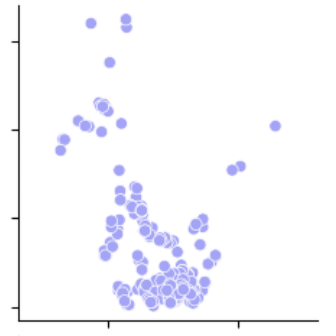


Figura 19. Grafica de Dispersión Distance nearest MRT station vs Latitude

La correlación entre X3 y X5 es -0.59, es negativa y fuerte.

X3 y X6

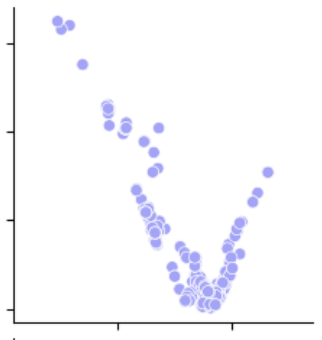


Figura 20. Grafica de Dispersión Distance nearest MRT station vs Logitude

La correlación entre X3 y X6 es -0.81, es negativa y fuerte.

X4 y X5

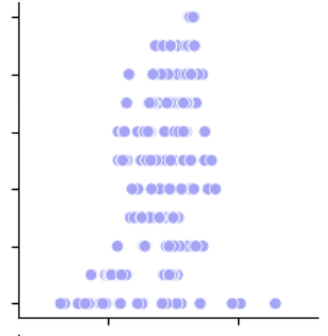


Figura 21. Grafica de Dispersión Number of Convenience Stores vs Latitude

La correlación entre X4 y X5 es 0.44, es positiva y moderada.

X4 y X6

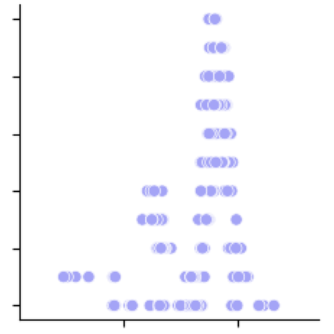


Figura 22. Grafica de Dispersión Number of Convenience Stores vs Longitude

La correlación entre X4 y X6 es 0.45, es positiva y moderada.

X5 y X6

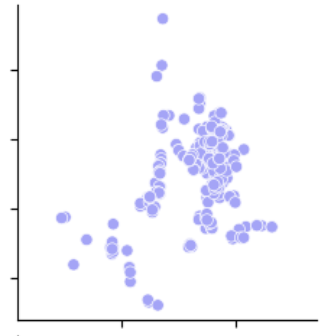


Figura 23. Grafica de Dispersión Latitude vs Longitude

La correlación entre X5 y X6 es 0.41, es positiva y débil.

**Correlación características con variable de respuesta:**

X1 y Y

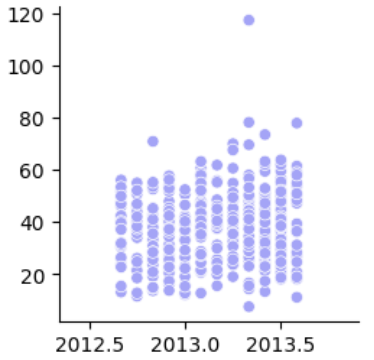


Figura 24. Grafica de Dispersión transaction date vs house price of unit area

La correlación entre X1 y Y es 0.087, es positiva y débil.

X2 y Y

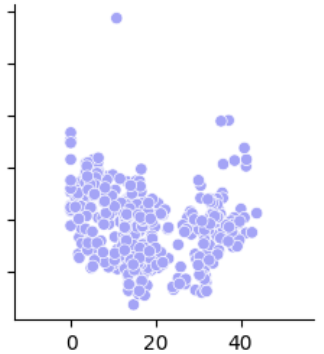


Figura 25. Grafica de Dispersión House Age vs house price of unit area

La correlación entre X2 y Y es -0.21, es negativa y débil.

X3 y Y

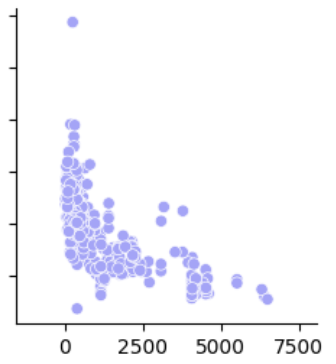
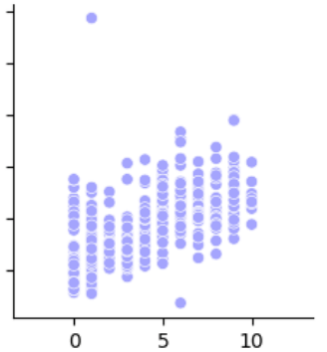


Figura 26. Grafica de Dispersión distance nearest MRT station vs house price of unit area

La correlación entre X3 y Y es -0.67, es negativa y fuerte.

X4 y Y

  
Figura 27. Grafica de Dispersión number convenience stores vs house price of unit area

La correlación entre X4 y Y es 0.57, es positiva y moderada.

X5 y Y

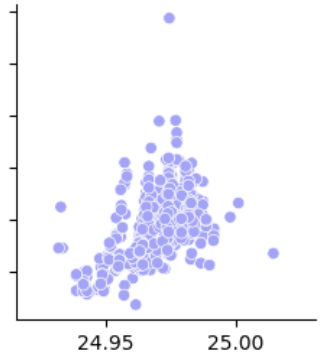


Figura 28. Grafica de Dispersión latitude vs house price of unit area

La correlación entre X5 y Y es 0.55, es positiva y moderada.

X6 y Y

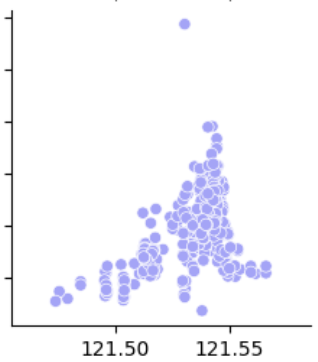


Figura 29. Grafica de Dispersión longitude vs house price of unit area

La correlación entre X6 y Y es 0.52, es positiva y moderada.

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

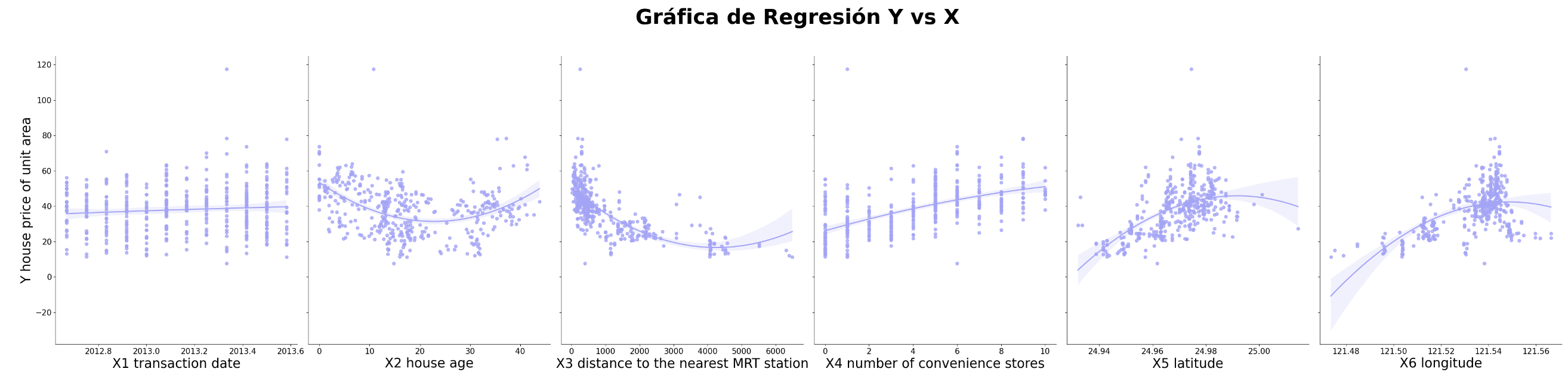


Figura 30. Matriz de graficas de regresión polinomial de segundo orden entre variable explicativa y variable de respuesta

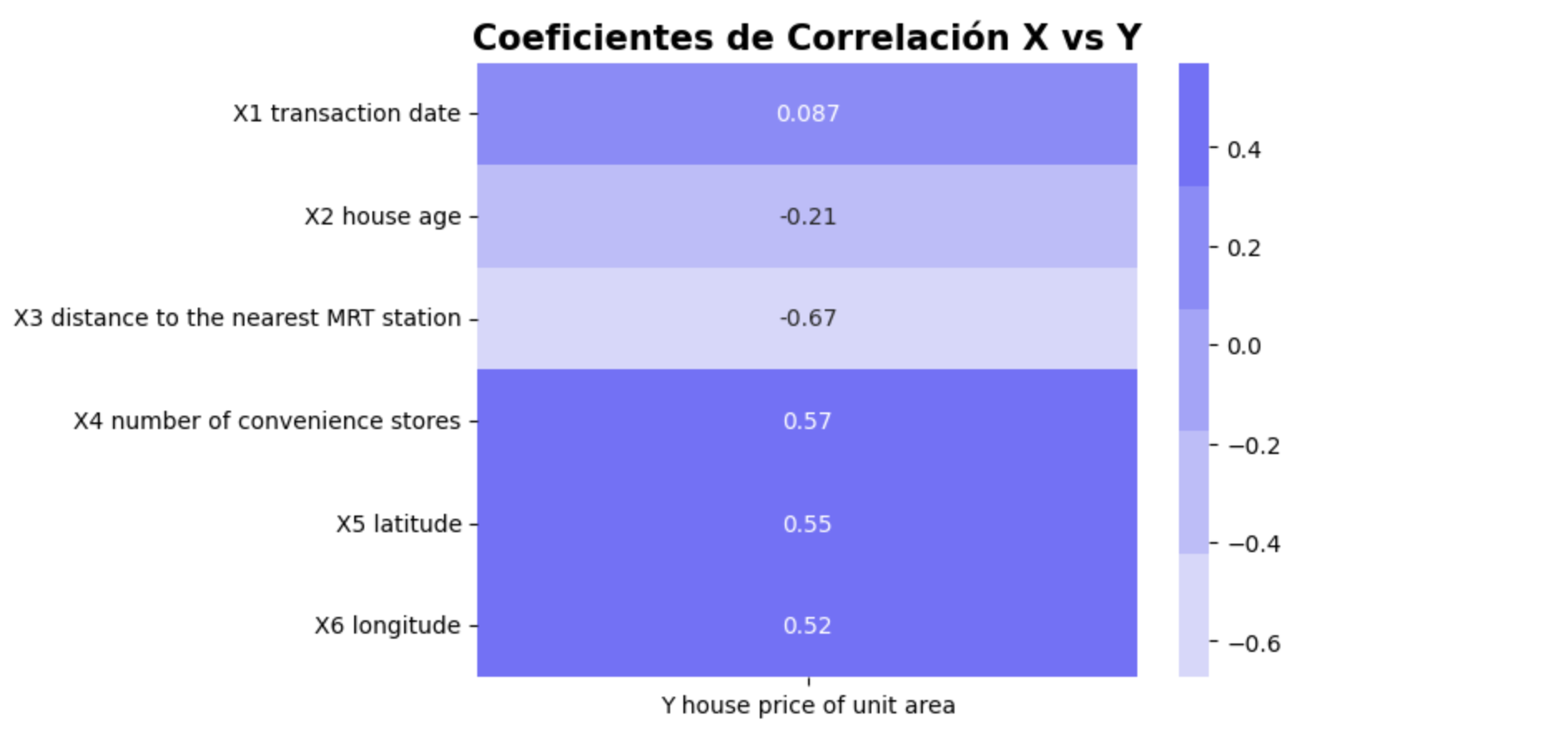


Figura 31. Grafica Coeficientes de Correlación entre variables explicativas y variable de respuesta

Para la variable X1 y Y se puede observar un valor atípico en aproximadamente 120. Se evidencia un aumento en la variable X1 con la variable de respuesta. La correlación baja y positiva con un valor de 0.087, lo que implica que el precio no varía significativamente con la fecha de la transacción.

Para la variable X2 y Y se puede observar un valor atípico en aproximadamente 120. Se evidencia una clara tendencia descendente. Su correlación baja y negativa con un valor de -0.21, lo que implica que las casas más antiguas tienden a tener un precio más bajo por unidad de área.

Para la variable X3 y Y se puede observar un valor atípico en aproximadamente 120. Se evidencia que la relación es negativa. Su correlación es moderadamente alta y negativa con un valor de -0.67, lo que implica que cuanto mayor es la distancia a la estación de MRT más cercana, menor es el precio por unidad de área de la vivienda.

Para la variable X4 y Y se puede observar un valor atípico en aproximadamente 120. Se evidencia que la relación es positiva con una tendencia ascendente. Su correlación es positiva con un valor de 0.57, lo que implica que un mayor número de tiendas de conveniencia cercanas está asociado con precios más altos por unidad de área.

Para la variable X5 y Y se puede observar un valor atípico en aproximadamente 120. Se evidencia que la relación es positiva con una tendencia ascendente moderada. Su correlación es positiva con un valor de 0.55, lo que implica que las casas ubicadas más al norte (si la latitud aumenta hacia el norte) tienen precios más altos.

Para la variable X6 y Y se puede observar un valor atípico en aproximadamente 120. Se evidencia que la relación es positiva con una tendencia ascendente moderada. Su correlación es positiva con un valor de 0.52, lo que implica que moverse hacia el este u oeste afecta de manera positiva el precio de las viviendas.

**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.

Inicialmente, se tomó el conjunto de todos datos y se realizó un *split* de estos en datos de 80% de los datos para entrenamiento y 20% para prueba que esto es lo que sugiere la literatura y la mayoría de artículos de la comunidad académica. Se construyó el siguiente modelo de regresión usando todos los datos como se muestra a continuación:

Sin embargo, estas métricas están sesgadas y no se podría apreciar si hay sobreajuste en el modelo. Esto sucede pues los datos que se usaron para hacer el cálculo de las métricas fueron utilizados para el entrenamiento del modelo. En la práctica se dividen los datos en dos grupos se utilizan unos datos para entrenar el modelo y otros para calcular las métricas de desempeño del modelo, pero por efectos del ejercicio los datos de pruebas también se utilizaron para entrenar el modelo.

b. Métricas del modelo usando validación cruzada.

Utilizando el método de validación cruzada que es un método para reducir la varianza de la estimación de la métrica se utilizaron k-folds donde k=50, es decir, se dividió el conjunto de datos en 50 partes y se ajustaron 50 modelos y para cada modelo se calculó el MSE, posteriormente se calculó un promedio simple para encontrar el valor esperado de cada métrica y se obtuvieron los siguientes resultados para la estimación del MSE, RMSE y el MAE:

El objetivo de validación cruzada es precisamente minimizar la varianza de la estimación de la métrica, ya que la métrica va cambiando, dependiendo de que datos utilizamos para entrenar el modelo y los datos de prueba. Como se puede evidenciar las métricas de la estimación del error utilizando validación cruzada son mayor en comparación a las métricas del inciso a) en donde se utilizaron los mismos datos de entrenamiento para estimarlas. Por consiguiente, las métricas de este ejercicio son más acertadas y tienen menor sesgo.

c. Evaluación del modelo y sus parámetros empleando pruebas estadísticas.

Para el modelo primero se tomaron todas las variables y todos los datos y se ajustó el modelo utilizando la librería statsmodels a través del método de mínimo cuadrados ordinarios (OLS) como se muestra a continuación:

A screenshot of a computer screen

Description automatically generated

Figura 32. Salida de Python Regresión OLS

**Ominbus test (F-test):**

Como se puede ver en la Figura 32 el p-value de la prueba Omnibus que no es más que la prueba F-test es menor a mi nivel de significancia ( por lo tanto se rechaza la hipótesis nula es decir estadísticamente con un nivel de confianza del 95% el modelo es globalmente significativo.

**Jarque-Bera (JB):**

Por otro lado, la prueba de Jarque-Bera (JB) permite establecer si los residuos del modelo se ajustan a una distribución normal. Como el p-value = 0, por lo tanto, se rechaza la hipótesis nula es decir estadísticamente con un nivel de significancia de 5% los residuos no siguen una distribución normal.

**AIC/BIC:**

Estas dos métricas son el *Akaike Information Criterion (AIC)* y el *Bayesian Information Criterion (BIC)* que permiten comparar modelos y son una métrica que mide el error. Por lo tanto, se busca que sea el menor valor posible.

**Durbin-Watson (DW) test:**

La prueba de Durbin-Watson permite validar el supuesto de independencia lineal de los residuos, con este indicador se puede comprobar que efectivamente que los residuos no están correlacionados.

Además, están los betas o coeficientes estimados de cada variable que representan el efecto de aumentar en una unidad la variable explicativa sobre la variable de respuesta teniendo las demás variables constantes como se muestran a continuación:

A partir de los p-values de significancia individual de las variables explicativas se puede concluir que todas las variables son estadísticamente significativas con un p-vlaue < 0.05, excepto la variable longitud con un p-value = 0.798.

Se calculó también el *Variance Inflation Factor (VIF)* que permite determinar si hay problemas de multicolinealidad en el modelo.

A screenshot of a black and white screen

Description automatically generated

Figura 33. Salida de Python Regresión OLS

Como podemos ver en la Figura 33, efectivamente hay problemas de multicolinealidad en el modelo entre las variables X1, X5 y X6.

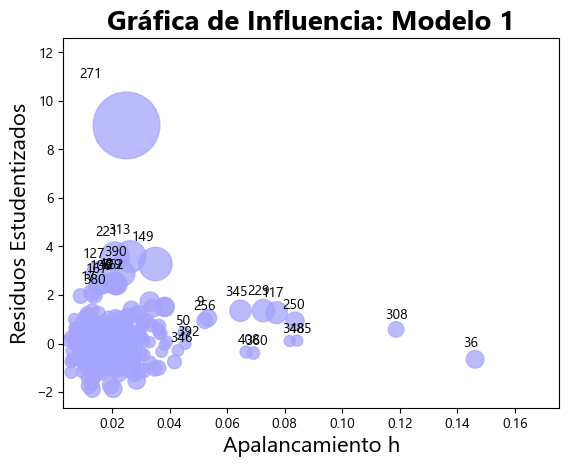


Figura 34. Grafica de influencia Residuos estudentizados vs Apalancamiento

Por último, se hizo un grafico de influencia que permite establecer si hay outliers o valores atípicos en los datos que posiblemente estén afectando la estimación de los coeficientes del modelo, se puede concluir que no pareciese que hubiese datos influyentes.