**Nombre:** Sebastian Felipe Valencia Velasco

**Código:**  201815565

**Nombre:** Andrés Francisco Borda

**Código:**  201729184

**Asignatura:** *Analítica computacional para la toma de decisiones*

**Profesor:** Juan Fernando Pérez

**Sección:** 1

Taller 3 Modelos de aprendizaje en Python

**2. Análisis exploratorio**

Una vez cargados los datos, se exploraron las diferentes variables y la información que contienen.

Texto

Descripción generada automáticamente con confianza baja

Como podemos ver tenemos 414 entradas descritas en 7 columnas, fecha de transacción, edad de la casa, distancia hasta la clínica de urgencias más cercana, numero de supermercados cercanos, latitud, longitud y la variable de interés el precio por metro cuadrado.

1. Comportamiento individual de las variables

Para analizar el comportamiento individual de las variables se obtuvieron sus estadísticas descriptivas y se graficaron boxplots de cada una para visualizar mejor su distribución.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

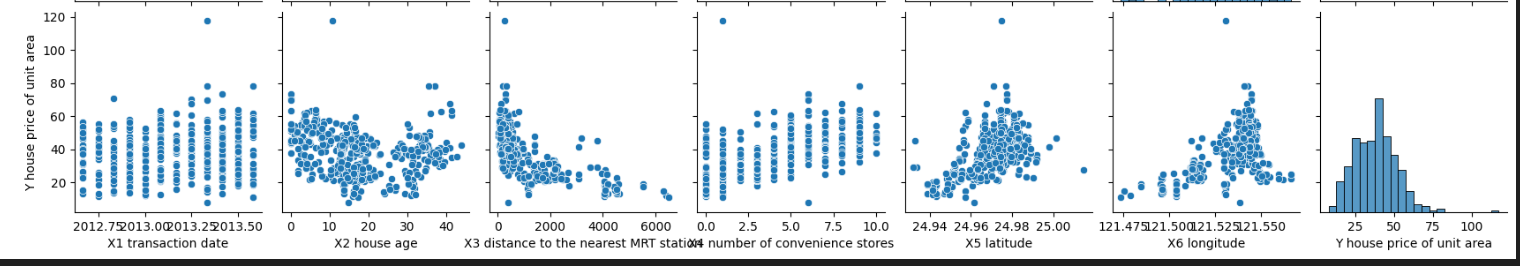
Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Utilizando esto podemos observar los rangos de las variables, su distribución y los datos atípicos que tienen. Viendo la distribución de las variables

1. Correlaciones entre las variables

Para observar las correlaciones entre las variables se realizó un mapa de calor y un gráfico de pares.



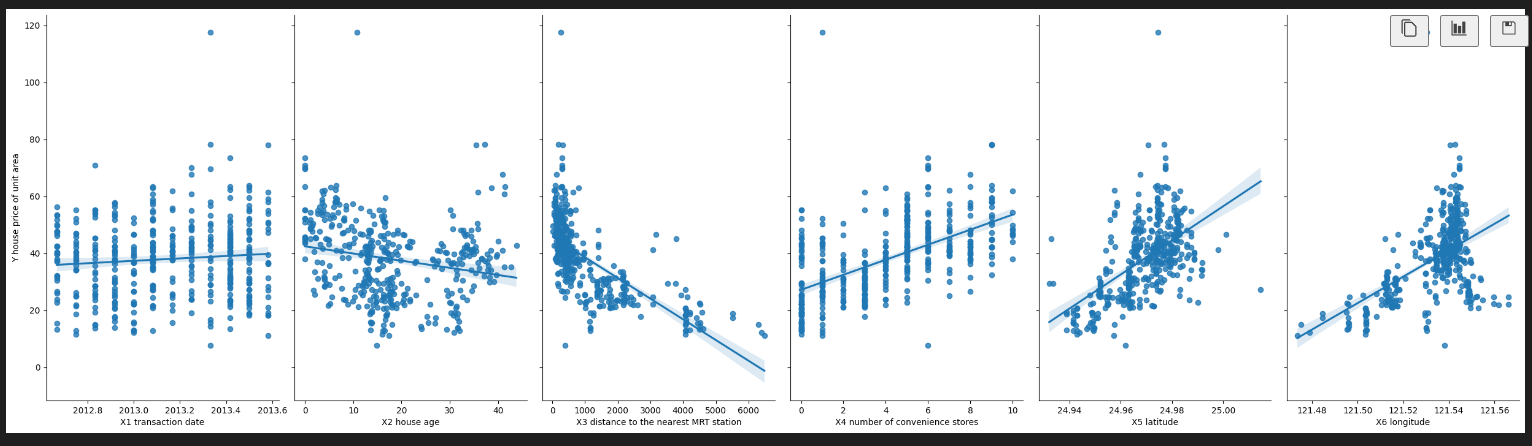
Interfaz de usuario gráfica

Descripción generada automáticamente

Utilizando estas gráficas, podemos comenzar a evidenciar que variables parecen ser buenas candidatas para implementar nuestro modelo lineal.

1. Comparación bivariada de la variable de interés frente a las características

Ahora, realizaremos una comparación bivariada de las características contra la variable de interés utilizando una línea de tendencia que nos permita visualizar de mejor manera la relación entre estas.



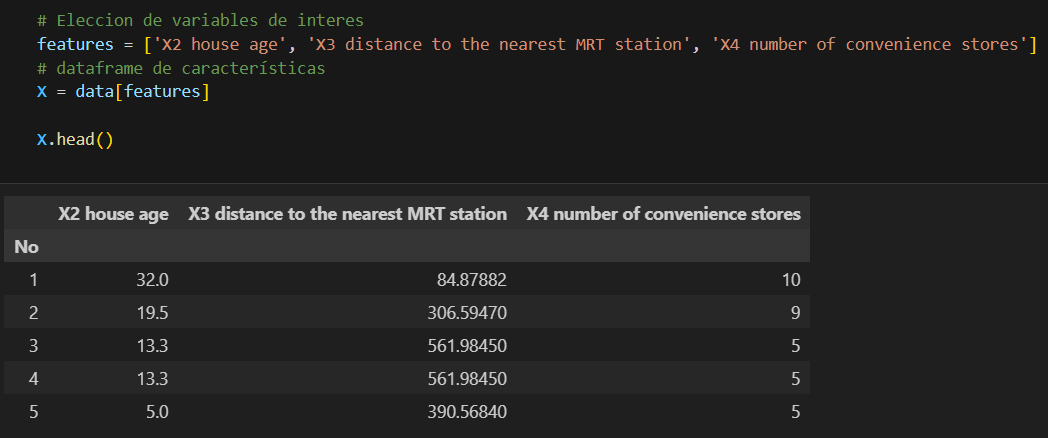
**3. Creación del modelo Lineal:**

Para la creación del modelo se implementarán modelos de regresión basados en la librería de SKLearn , junto con el comportamiento evidenciado en la parte exploratoria del modelo; primero se procede a importar las herramientas que se usarán.

Texto

Descripción generada automáticamente

Basados en el análisis exploratorio, se decidió incluir en el análisis del modelo solo las variables que tenían un comportamiento basado en alguna tendencia, eliminando así los datos aleatorios o sin ningún tipo de comportamiento dependiente aparente. De esta manera, se decidió tomar como los features a analizar la edad de la casa, las distancia a la estación mas cercana y el numero de tiendas de conveniencia. Pues estas variables según en análisis bivariado , muestra que hay un comportamiento que influye en la variable de respuesta.



Se declara como variable de respuesta, los datos correspondientes al precio de la casa por unidad de área

Texto

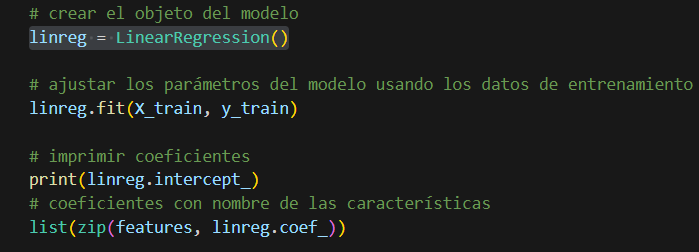
Descripción generada automáticamente

Se procede a realizar una separación del dataset con el fin de tener un conjunto de datos de entreno y otro de validación del modelo , para esto se define una separación de 80% para datos de entreno y 20% para datos de validación , luego se corrobora que todos los datos de entreno para x,y y de validación para x,y Tengan las mismas dimensiones.

Texto

Descripción generada automáticamente

Se declara el objeto del modelo, que para esta instancia se declara como un objeto de regresión lineal y se procede a calcular los coeficientes de cada variable independiente basado en la significancia y afectación de la variable en los datos de respuesta que en este caso sería el precio de la vivienda.



Texto

Descripción generada automáticamente

Se reporta el intercepto del modelo y los coeficientes lineales asociados a las variables independientes que se escogieron para el modelo, de esta manera se puede tener una primera aproximación de la medida en que la variable afecta a la variable de respuesta del modelo , concluyendo que las variables de ‘edad de la casa’ y ‘distancia a la estación más cercana’ tienen un comportamiento de alguna proporcionalidad inversa o negativa con respecto a la variable de respuesta , a diferencia del numero de tiendas cercano que maneja una proporcionalidad con un coeficiente de 1.28 basado en el resumen de los datos.

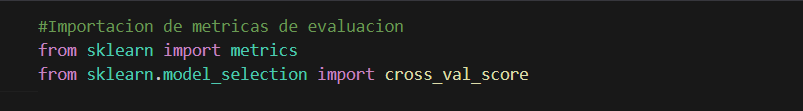
Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

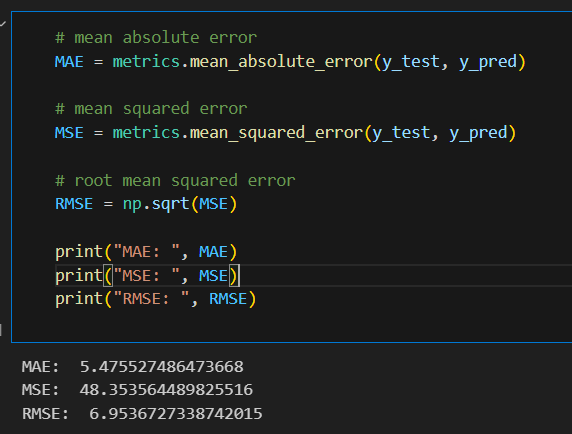
Se procede con la predicción del dataset de validación dado el modelo ya previamente preparado.

1. **Métricas del modelo usando datos de entrenamiento**

Para evaluar las métricas del modelo usando datos de entrenamiento se usará el módulo métrics de la librería de sklearn y también cross\_val\_score para hacer su análisis de validación cruzada.



Se usan varios tipos de métricas para evaluar la exactitud de la regresión con los datos ; para ello se plantea el mean absolute error, mean squared error y RMSE .



Basados en el resultado del análisis no se puede dar un análisis contundente ya que no existe un modelo de referencia para hacer la comparativa relativa entre modelos. No obstante, se puede ver que existe un error considerable entre la regresión y los datos, para ello se harán mas análisis del modelo.

1. **Métricas del modelo usando validación cruzada**

Se complementa el análisis usando la técnica de validación cruzada}

Texto

Descripción generada automáticamente

Se usa la función cross\_val\_score para hacer la validación cruzada del modelo usando (Cv=5) 5 pliegues del dataset y evaluándolos entre si , para proceder a validarlos bajo la métrica de NMSE .

Texto

Descripción generada automáticamente

Se calcula el RMSE basado en la métrica arrojada por la función previa y se procede a sacar el promedio de la métrica:

Texto

Descripción generada automáticamente

Donde se visualiza que no existe una muy buena exactitud entre la regresión y los datos, por ello la gran desviación que aun no se captura usando el modelo de regresión lineal.

1. **Evaluación del modelo y sus parámetros empleando pruebas estadísticas**

Se procede a hacer pruebas estadisticas sacando el resumen estadístico del modelo usando la librería de statsmodels

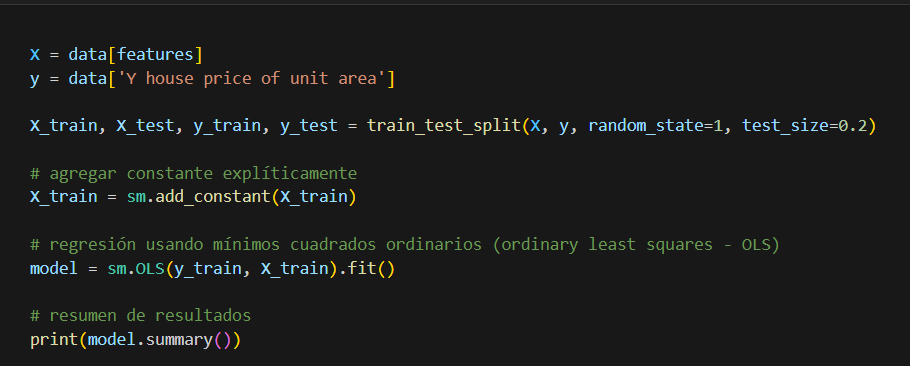


Imagen que contiene Escala de tiempo

Descripción generada automáticamente

Utilizando el análisis estadístico del modelo podemos evaluar la significancia del modelo, así como la significancia individual de cada uno de los factores incluidos en la regresión. En la imagen podemos ver que el P-Value de las tres variables elegidas para el modelo es muy bajo cercano a 0 por lo que podemos afirmar con una significancia del 5% que estas variables son significativas para explicar el cambio en el precio del metro cuadrado. Siguiendo con el análisis de las pruebas estadísticas, podemos ver que el skew cae dentro de los valores aceptables +/- 3 pero la kurtosis es muy alta. Para la prueba de autocorrelación, podemos ver que la prueba de Durbin-Watson nos da un valor muy cercano a 2 por lo que no existe autocorrelación en la regresión construida. Por último, podemos ver que la probabilidad de la prueba de Jarque-Bera nos da un valor muy cercano a 0 por lo que podemos afirmar que los residuos no se distribuyen normalmente por lo que habría un problema con los supuestos del modelo.