Sebastián Arango Crispín - 201921816

Tomás Liévano Casas - 201920980

Taller 3 – Modelos de aprendizaje en Python

Analítica Computacional para la toma de decisiones

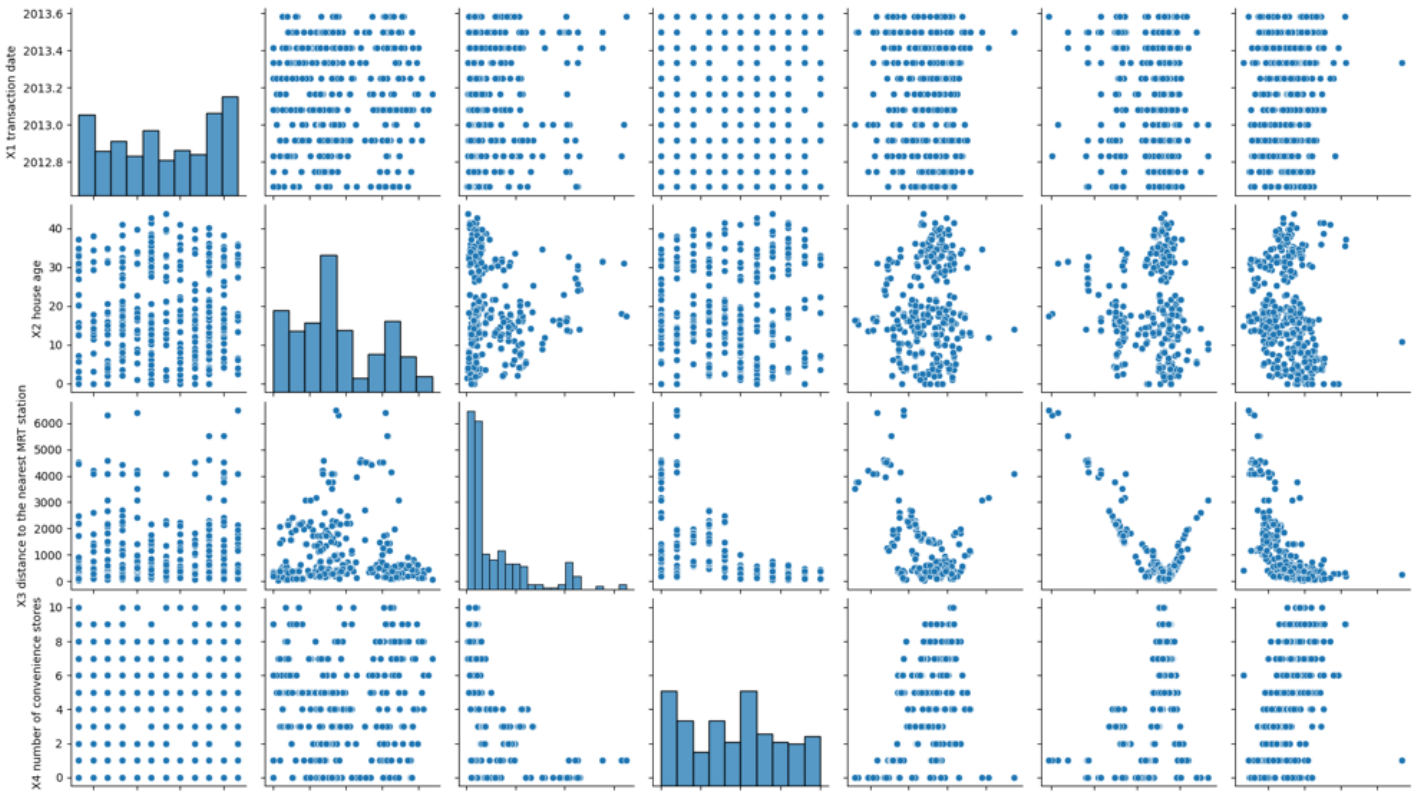
1. Exploración

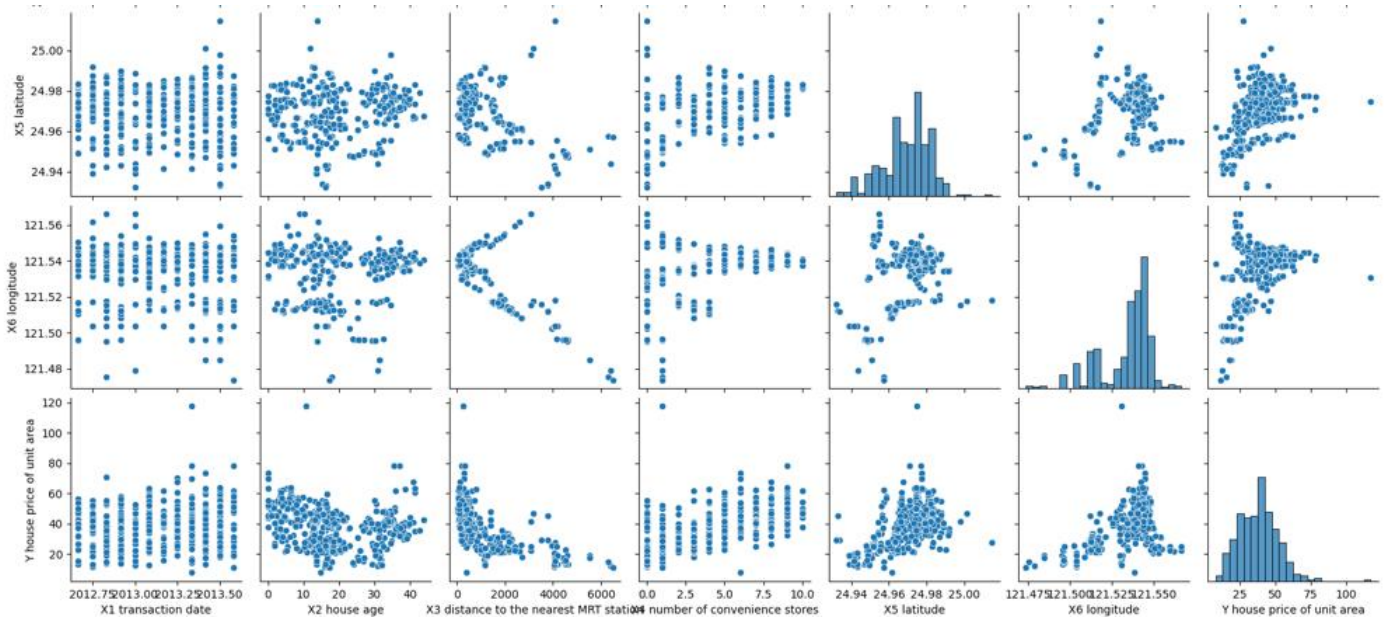
1.1 Comportamiento individual de cada característica y de la variable de respuesta

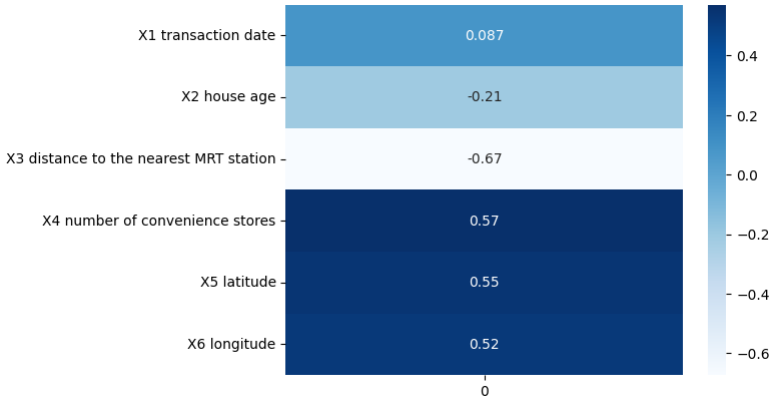
Tomás: Tenemos 414 datos con 6 características y una variable de respuesta,

	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

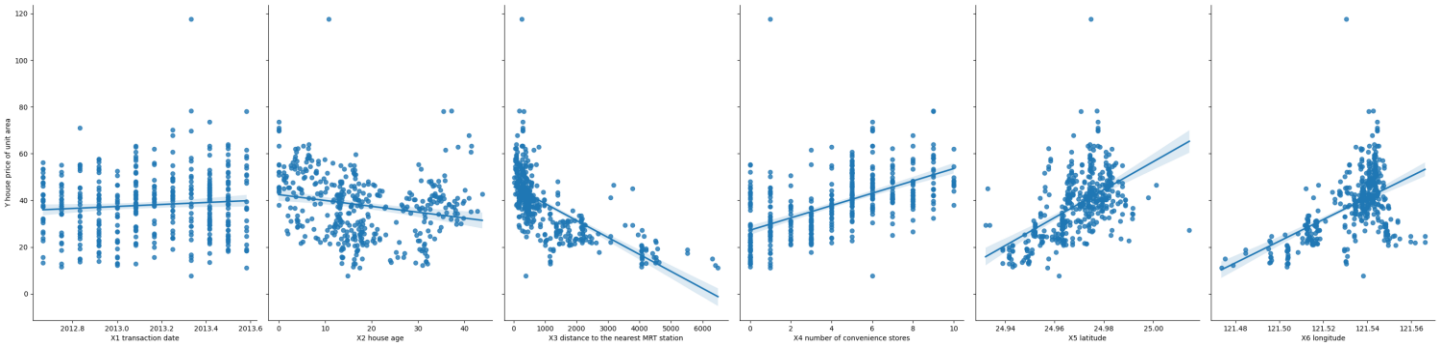
1.2 Correlaciones entre características y con la variable de respuesta







1.3 Exploración bivariada entre cada característica y la variable de respuesta



2. Modelo lineal

2.1 Métricas del modelo usando datos de entrenamiento

MAE: 5.5725

MSE: 53.7308

$RMSE$: 7.3301

2.2 Métricas del modelo usando validación cruzada

$$MSE: 78.3342$$

$$RMSE: 8.6934$$

Como se puede ver, la validación cruzada da un valor de errores mayor que en el caso donde se utilizan los datos de entrenamiento solo

2.3 Evaluación del modelo y sus parámetros empleando pruebas estadísticas

```
                            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:                      mar., 27 ago. 2024   Prob (F-statistic):           1.05e-48
Time:                              13:49:19   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
```

Las pruebas estadísticas muestran que las variables 1 a la 5 son significativas individualmente para el modelo, mientras que la 6 variable no es estadísticamente significativa de manera individual para el modelo, ya que supera el valor de α = 5%. Por su parte, es posible ver que el modelo global es estadísticamente significativo. La prueba de durbin-Watson muestra que no hay evidencia estadística de autocorrelación, mientras qye la prueba de Omnibus muestra que los residuos no son normales.