Práctica 2

Autor:

John Mauricio Olivos Domínguez

Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

Deberéis trabajar en grupos de 2 personas y entregar un solo archivo con el enlace al repositorio Git donde se encuentren las soluciones, incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos que corresponden a vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github.

Además, se debe entregar un vídeo explicativo de la práctica, donde ambos integrantes del equipo expliquen con sus propias palabras el desarrollo de la práctica, basándose en las preguntas del enunciado para justificar y explicar el código desarrollado. Este vídeo se deberá entregar a través de un enlace a Google Drive que se deberá proporcionar junto con enlace al repositorio Git.

Es importante que la entrega de esta práctica se realice en el formato especificado en el apartado Formato y fecha de entrega.

Aunque no se trata exactamente del mismo enunciado ni de una solución que obtuviera la máxima nota, el siguiente ejemplo de una edición anterior os puede servir de guía para la realización de la práctica: https://github.com/Bengis/nba-gap-cleaning

Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.

- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.

- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Descripción de la Práctica a realizar

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la Práctica 1 o bien cualquier dataset libre disponible en Kaggle https://www.kaggle.com.

Un ejemplo de dataset con el que podéis trabajar es el "Heart Attack Analysis & Prediction dataset": https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

Importante: si se elige un dataset diferente al propuesto es importante que este contenga una amplia variedad de datos numéricos y categóricos para poder realizar un análisis más rico y poder responder a las diferentes preguntas planteadas en el enunciado de la práctica.

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y **justificar**) son las siguientes:

1. **Descripción del dataset**. ¿Por qué es importante y qué pregunta/problema pretende responder?
2. **Integración y selección** de los datos de interés a analizar. Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.
3. **Limpieza de los datos**. 3.1. ¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos. 3.2. Identifica y gestiona los valores extremos.
4. **Análisis de los datos**. 4.1. Selección de los grupos de datos que se quieren analizar/comparar (p.ej., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?) 4.2. Comprobación de la normalidad y homogeneidad de la varianza. 4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.
5. **Representación de los resultados** a partir de tablas y gráficas. Este apartado se puede responder a lo largo de la práctica, sin necesidad de concentrar todas las representaciones en este punto de la práctica.
6. **Resolución del problema**. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?
7. **Código**. Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.
8. **Vídeo**. Realizar un breve vídeo explicativo de la práctica (máximo 10 minutos), donde ambos integrantes del equipo expliquen con sus propias palabras el desarrollo de la práctica, basándose en las

preguntas del enunciado para justificar y explicar el código desarrollado. Este vídeo se deberá entregar a través de un enlace al Google Drive de la UOC (https://drive.google.com/...), junto con enlace al repositorio Git entregado.

Desarrollo de la práctica

---

1. Descripción del dataset.

El dataset utilizado para este ejercicio contiene información detallada acerca de los inmuebles que se encontraban publicados el día 20/04/2023 para venta en la página web https://fincaraiz.com.co/. Este dataset permite consultar atributos de diversos inmuebles que se encuentran en venta (para este ejercicio inmuebles ubicados en la localidad de Engativá, Bogotá, Colombia), de igual manera los datos allí registrados permiten hacer diversas comparativas para determinar precio por metro cuadrado, precio del inmueble, entre otras.

Durante los últimos años el mercado de bienes raíces ha crecido de manera significativa, por lo tanto, es un mercado que diariamente presenta una gran dinámica, motivo por el cual en ocasiones se hace complicado evaluar rápidamente aquellas opciones que son más favorables para el comprador. En este ejercicio se pretende realizar la limpieza de los datos extraídos originalmente de tal manera que facilite el análisis de este conjunto de datos e identificar características de los diferentes inmuebles que pueden influir en su precio, así como identificar aquellos inmuebles que podrían ser una buena opción de inversión.

Variables presentes en el dataset

- **TITULO** (str): Título del anuncio, facilita la identificación del tipo de inmueble
- **CODIGO INMUEBLE** (str): Identificador asignado por la página al inmueble
- **ÁREA PRIVADA** (str): Área privada del inmueble
- **ÁREA CONSTRUÍDA** (float): Área construida del inmueble
- **PRECIO POR M²** (int): Valor por cada m^2 construido
- **ESTRATO** (str): Estrato social de la zona en la que se encuentra ubicado el inmueble
- **ANTIGUEDAD** (str): Tiempo transcurrido desde el año de construcción del inmueble
- **HABITACIONES** (int): Número de habitaciones construidas en el inmueble
- **BAÑOS** (int): Número de baños construidos en el inmueble
- **PARQUEADEROS** (int): Número de zonas de parqueo con las que cuenta el inmueble
- **BARRIO** (str): Dirección de ubicación del inmueble
- **PRECIO** (int): Precio total del inmueble

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Se leen los datos y se da un vistazo a estos

2. Lectura del dataset y limpieza de los datos

In [2]:
```
df = pd.read_csv('Data/InmueblesVentaEngativa_original.csv')
df.head()
```

Out[2]:

	TITULO	CODIGO INMUEBLE	ÁREA PRIVADA	ÁREA CONSTRUÍDA	PRECIO POR M²	ESTRATO	ANTIGUEDAD	HABITACIONES	BAÑOS
0	Apartamento en venta	7802171	0 m²	51.0m²	3039215	2	16 a 30 años	3	1
1	Apartaestudio en venta	7743515	0 m²	36.0m²	7777777	3	1 a 8 años	0	0
2	Apartamento en venta	7834229	0 m²	44.0m²	4318181	3	16 a 30 años	3	2
3	Apartamento en venta	7711987	0 m²	37.0m²	7567567	3	menor a 1 año	1	1
4	Casa en venta	10040921	0 m²	128.0m²	1835937	2	9 a 15 años	4	3

In [3]:
```
df.count()
```

Out[3]:
```
TITULO              538
CODIGO INMUEBLE     538
ÁREA PRIVADA        538
ÁREA CONSTRUÍDA     538
PRECIO POR M²       538
ESTRATO             538
ANTIGUEDAD          538
HABITACIONES        538
BAÑOS               538
PARQUEADEROS        538
BARRIO              538
DIRECCIÓN           538
PRECIO              538
dtype: int64
```

In [4]:
```
df.isnull().sum()/len(df)*100
```

Out[4]:
```
TITULO              0.0
CODIGO INMUEBLE     0.0
ÁREA PRIVADA        0.0
ÁREA CONSTRUÍDA     0.0
PRECIO POR M²       0.0
ESTRATO             0.0
ANTIGUEDAD          0.0
HABITACIONES        0.0
BAÑOS               0.0
PARQUEADEROS        0.0
BARRIO              0.0
DIRECCIÓN           0.0
PRECIO              0.0
dtype: float64
```

Se evidencia que el dataset consta de 13 columnas y 538 registros. Asi mismo, se evidencia que no existen registros con valores nulos.

A continuación se ejecuta la función describe, la cual permite hacer una descriptiva rápida de las variables de tipo numérico. En concreto, muestra la media, la desviación estándard, el mínimo, el màximo y los

cuartiles de las variables.

In [5]:
```
df.describe().T
```

Out[5]:

	count	mean	std	min	25%	50%	75%	
CODIGO INMUEBLE	538.0	8.111135e+06	1.220433e+06	3370112.0	7.473828e+06	7835443.5	8.022842e+06	1.00
PRECIO POR M²	538.0	3.528825e+06	1.369657e+06	800000.0	2.689344e+06	3420000.0	4.148246e+06	1.18
ESTRATO	538.0	2.762082e+00	7.317570e-01	0.0	2.000000e+00	3.0	3.000000e+00	6.00
HABITACIONES	538.0	3.927509e+00	2.469996e+00	0.0	3.000000e+00	3.0	4.000000e+00	1.90
BAÑOS	538.0	2.420074e+00	1.462800e+00	0.0	2.000000e+00	2.0	3.000000e+00	1.00
PARQUEADEROS	538.0	6.505576e-01	8.415478e-01	0.0	0.000000e+00	0.5	1.000000e+00	6.00
PRECIO	538.0	3.292941e+08	2.232629e+08	119899978.0	1.852500e+08	265799984.0	3.900000e+08	2.50

A pesar de que no existen registros nulos, se evidencia que existen casos con las variables ESTRATO, HABITACIONES y BAÑOS con valores en cero (0), motivo por el cual se procede a revisar y cuantificar dichos registros.

In [6]:
```
df[(df['ESTRATO'] == 0) | (df['HABITACIONES'] == 0) |(df['BAÑOS'] == 0)]
```

Out[6]:

	TITULO	CODIGO INMUEBLE	ÁREA PRIVADA	ÁREA CONSTRUÍDA	PRECIO POR M²	ESTRATO	ANTIGUEDAD	HABITACIONES	BAÑ
1	Apartaestudio en venta	7743515	0 m²	36.0m²	7777777	3	1 a 8 años	0	
6	Casa en venta	7849146	0 m²	217.0m²	1612903	0	Sin información	5	
122	Casa en venta	7737397	91 m²	91.0m²	6043956	2	16 a 30 años	0	
125	Apartamento en venta	8022852	0 m²	47.0m²	3723404	3	Sin información	0	
205	Casa en venta	7595380	0 m²	90.0m²	6666666	4	Sin información	0	
262	Apartamento en venta	8022852	0 m²	47.0m²	3723404	3	Sin información	0	

Teniendo en cuenta que son pocos registros se procede a eliminarlos del dataset

In [7]:
```
df = df[(df['ESTRATO'] != 0) & (df['HABITACIONES'] != 0) & (df['BAÑOS'] != 0)]
df.head()
```

Out[7]:

	TITULO	CODIGO INMUEBLE	ÁREA PRIVADA	ÁREA CONSTRUÍDA	PRECIO POR M²	ESTRATO	ANTIGUEDAD	HABITACIONES	BAÑOS
0	Apartamento en venta	7802171	0 m²	51.0m²	3039215	2	16 a 30 años	3	1

2	Apartamento en venta	7834229	0 m²	44.0m²	4318181	3	16 a 30 años	3	2
3	Apartamento en venta	7711987	0 m²	37.0m²	7567567	3	menor a 1 año	1	1
4	Casa en venta	10040921	0 m²	128.0m²	1835937	2	9 a 15 años	4	3
5	Casa en venta	7584617	80 m²	127.0m²	3031496	2	16 a 30 años	5	2

A continuación se procede a realizar un conteo de datos únicos por variale

In [8]:
```
df.nunique()
```

Out[8]:
```
TITULO               3
CODIGO INMUEBLE    437
ÁREA PRIVADA       113
ÁREA CONSTRUÍDA    162
PRECIO POR M²      350
ESTRATO              5
ANTIGUEDAD           6
HABITACIONES        14
BAÑOS                9
PARQUEADEROS         7
BARRIO              52
DIRECCIÓN           92
PRECIO             358
dtype: int64
```

A partir de la observación generada y el conteo de datos únicos de este dataset se puede evidenciar varios puntos:

- Teniendo en cuenta que el conteo de registros únicos de la variable "CODIGO INMUEBLE" es inferior al total de registos, se puede concluir que existen registros de inmuebles duplicados.
- La variable título puede servir para categorizar el tipo de inmueble.
- Teniendo en cuenta que en parte este ejercicio se analizará la relación entre el área construida y el valor por m2 y que la variable "ÁREA CONSTRUIDA" no es númerica, se debe hacer una transformación de este dato, bien sea quitando el cáracter de "m2" o realizando su cálculo mediante la formula PRECIO/PRECIO POR m2.
- Las variables "ESTRATO", "ANTIGUEDAD" y "PARQUEADEROS" se utilizarán como variables categóricas, teniendo en cuenta que para esta última el valor será "NO" para aquellos inmuebles que su valor actual sea cero (0) y "SI" para los registros que contengan valor mayor o igual a 1.
- La variable "ÁREA PRIVADA" no se utilizará teniendo en cuenta que existen varios registros con valor cero(0).

inicialmente se eliminarán aquellos registros que se encuentren duplicados

In [9]:
```
df = df.drop_duplicates(subset=['CODIGO INMUEBLE'])
df.count()
```

Out[9]:
```
TITULO             437
CODIGO INMUEBLE    437
ÁREA PRIVADA       437
ÁREA CONSTRUÍDA    437
PRECIO POR M²      437
```

```
ESTRATO              437
ANTIGUEDAD           437
HABITACIONES         437
BAÑOS                437
PARQUEADEROS         437
BARRIO               437
DIRECCIÓN            437
PRECIO               437
dtype: int64
```

Se realizá el cálculo para obtener la variable númerica "ÁREA CONSTRUIDA M2", se ajusta variable "PARQUEADEROS" para que indique si el inmueble cuenta con parquadero(s), se categorizan las variables a las que haya lugar y por último y con el fin de simplificar la lectura de los análisis se crea una variable que represente en millones de pesos el valor por metro cuadrado(m2) redondeado a 2 decimales.

In [10]:
```
import pandas as pd
pd.options.mode.chained_assignment = None
#Cálculo de variable númerica área construida m2
df['ÁREA CONSTRUIDA M2'] = df['PRECIO'] / df['PRECIO POR M²']

#Clasificación variable parqueaderos
df['PARQUEADEROS'] = np.where(df['PARQUEADEROS']==0, 'NO', 'SI')

#Creación variable valor por m2 en millones de pesos
df['PRECIO M2 MMP'] = round(df['PRECIO POR M²'] / 1000000,2)

# Convertir datos en categóricos
df['TIPO_INMUEBLE'] = df['TITULO'].astype('category')
df['ESTRATO'] = df['ESTRATO'].astype('category')
df['ANTIGUEDAD'] = df['ANTIGUEDAD'].astype('category')
df['PARQUEADEROS'] = df['PARQUEADEROS'].astype('category')

df.head()
```

Out[10]:

	TITULO	CODIGO INMUEBLE	ÁREA PRIVADA	ÁREA CONSTRUÍDA	PRECIO POR M²	ESTRATO	ANTIGUEDAD	HABITACIONES	BAÑOS
0	Apartamento en venta	7802171	0 m²	51.0m²	3039215	2	16 a 30 años	3	1
2	Apartamento en venta	7834229	0 m²	44.0m²	4318181	3	16 a 30 años	3	2
3	Apartamento en venta	7711987	0 m²	37.0m²	7567567	3	menor a 1 año	1	1
4	Casa en venta	10040921	0 m²	128.0m²	1835937	2	9 a 15 años	4	3
5	Casa en venta	7584617	80 m²	127.0m²	3031496	2	16 a 30 años	5	2

A continuación se procede a generar un dataset con las variables que se tendrán en cuenta para el desarrollo del ejercicio

3. Selección de los datos a utilizar

In [11]:
```
df_w=df[['TIPO_INMUEBLE', 'ESTRATO', 'ANTIGUEDAD', 'PARQUEADEROS', 'HABITACIONES', 'BAÑO
df_w
```

Out[11]:

	TIPO_INMUEBLE	ESTRATO	ANTIGUEDAD	PARQUEADEROS	HABITACIONES	BAÑOS	ÁREA CONSTRUIDA M2	PRECIO M2 MMP
0	Apartamento en venta	2	16 a 30 años	NO	3	1	51.00	3.04
2	Apartamento en venta	3	16 a 30 años	NO	3	2	44.00	4.32
3	Apartamento en venta	3	menor a 1 año	SI	1	1	37.00	7.57
4	Casa en venta	2	9 a 15 años	SI	4	3	128.00	1.84
5	Casa en venta	2	16 a 30 años	NO	5	2	127.00	3.03
...	...	...	...	...	...	...	...	...
532	Apartamento en venta	4	1 a 8 años	SI	3	2	60.00	4.17
533	Apartamento en venta	3	9 a 15 años	SI	3	2	65.00	4.00
534	Apartamento en venta	3	9 a 15 años	SI	3	2	64.00	3.91
535	Casa en venta	2	Sin información	NO	10	6	300.00	2.00
537	Apartamento en venta	3	16 a 30 años	NO	2	2	56.29	3.38

437 rows × 8 columns

Luego de haber realizado revisión de completitud de la información y la adecuación del tipo de variables, el dataset de trabajo cuenta con 8 variables; 4 númericas y 4 categóricas y 437 registros. A continuación se procede a realizar los respectivos análisis

4. Análisis de los datos

Análisis de Variables Númericas

In [12]:
```
df_w.describe()
```

Out[12]:

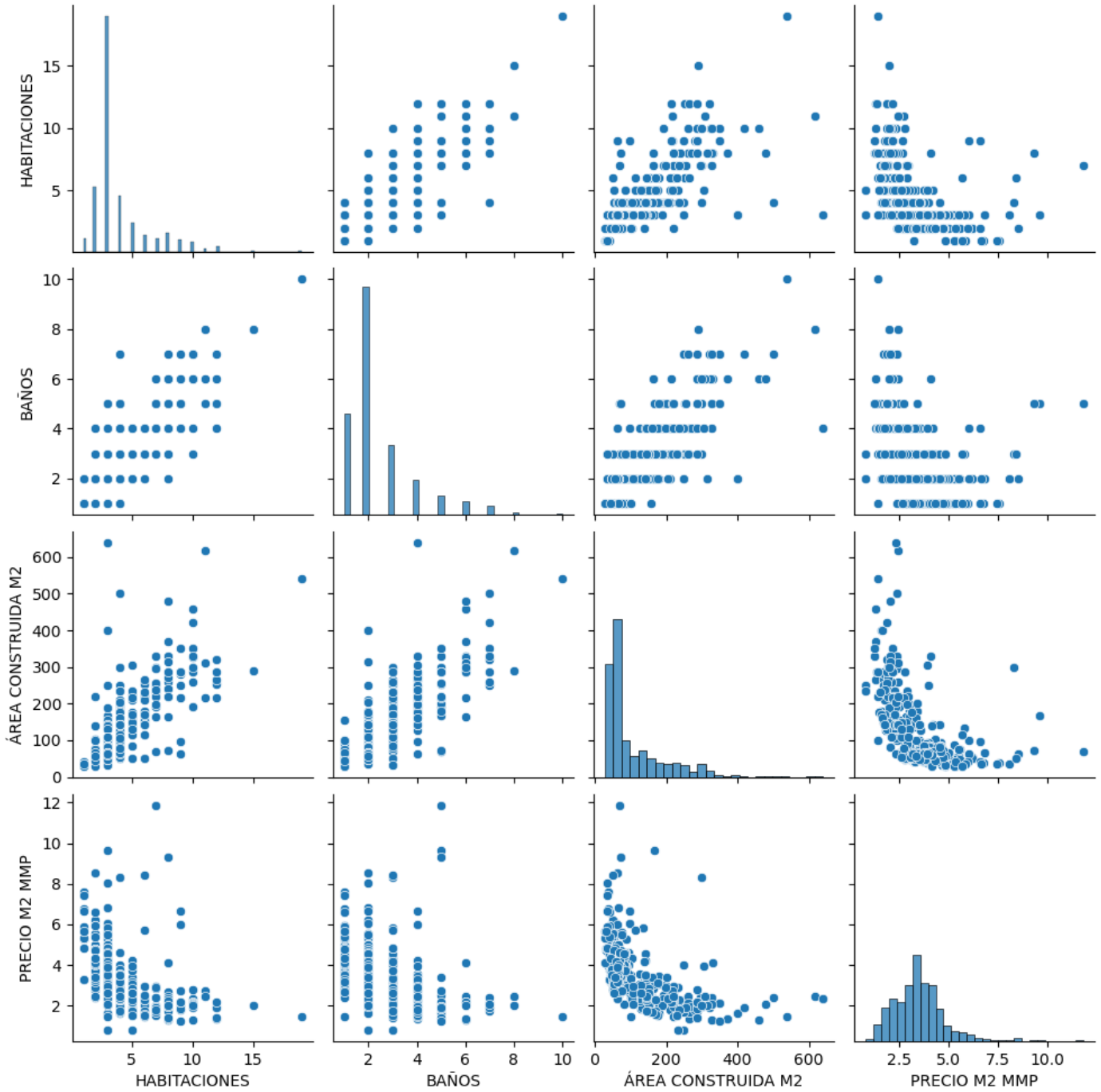
	HABITACIONES	BAÑOS	ÁREA CONSTRUIDA M2	PRECIO M2 MMP
count	437.000000	437.000000	437.000000	437.000000
mean	3.970252	2.446224	113.513776	3.491991
std	2.364014	1.419060	95.702424	1.352848
min	1.000000	1.000000	28.000000	0.800000
25%	3.000000	2.000000	52.000000	2.630000
50%	3.000000	2.000000	68.000000	3.400000

75%	4.000000	3.000000	144.000000	4.130000
max	19.000000	10.000000	640.000000	11.860000

Como primer análisis se puede observar que en la variable "PRECIO M2 MMP" se evidencia que el valor máximo se encuentra muy por encima de la media + 2 desviaciones estandar por lo cual pueden presentarse algunos valores atipicos.

A continuación se procede a realizar una representación gráfica de nuestras variables númericas, para ello se usará pairplot y una matriz de correlación

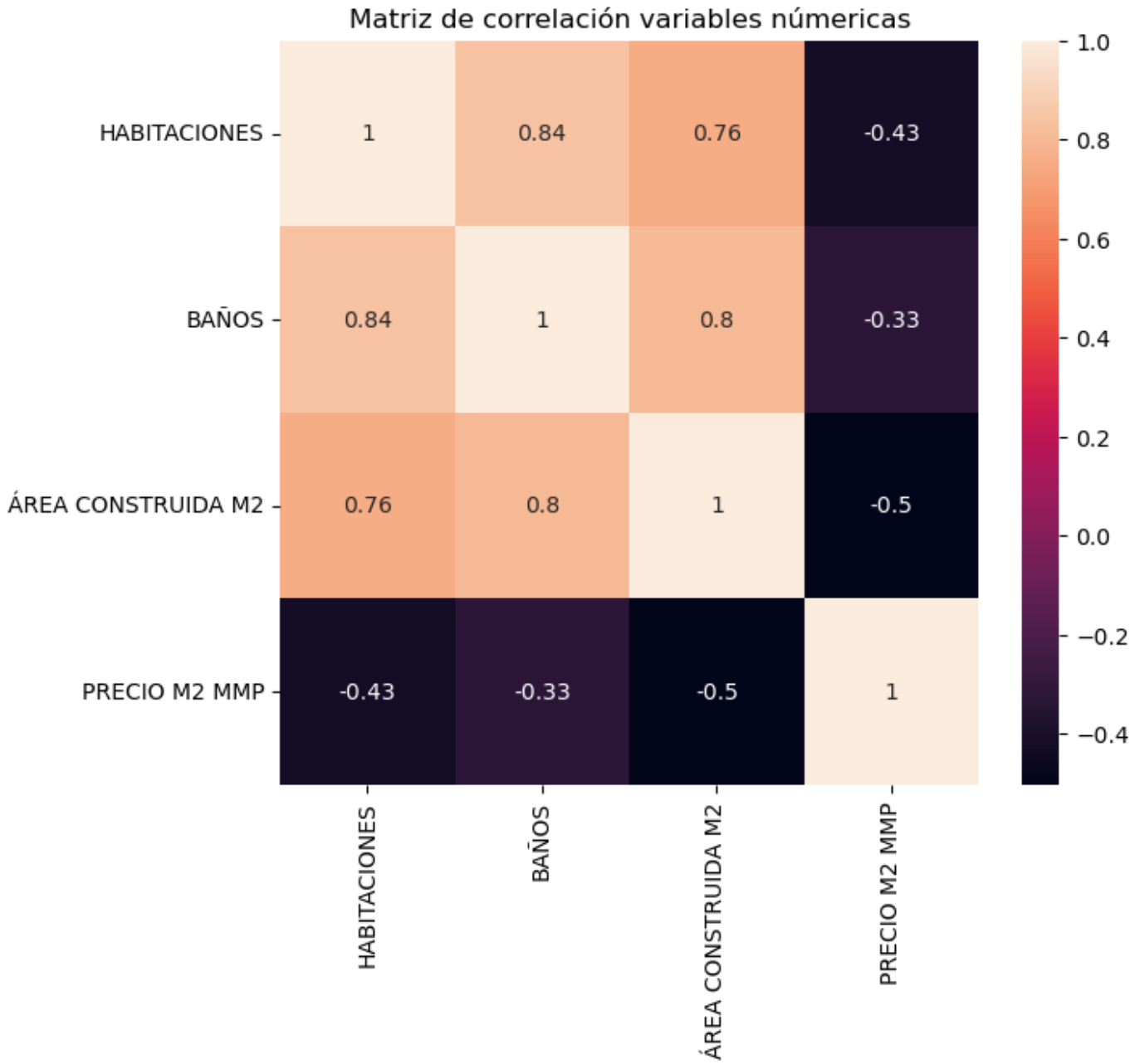
In [13]:
```
par= sns.pairplot(df_w)
```



In [14]:
```
correlation_matrix = df_w.corr(numeric_only=True)
correlation_matrix

plt.figure(figsize=(7,6))
plt.title('Matriz de correlación variables númericas')
```

```
sns.heatmap(correlation_matrix, annot=True,);
#df_w.corr('HABITACIONES','BAÑOS','ÁREA CONSTRUIDA M2', 'PRECIO M2 MMP')
```



De los gráficos podemos concluir que:

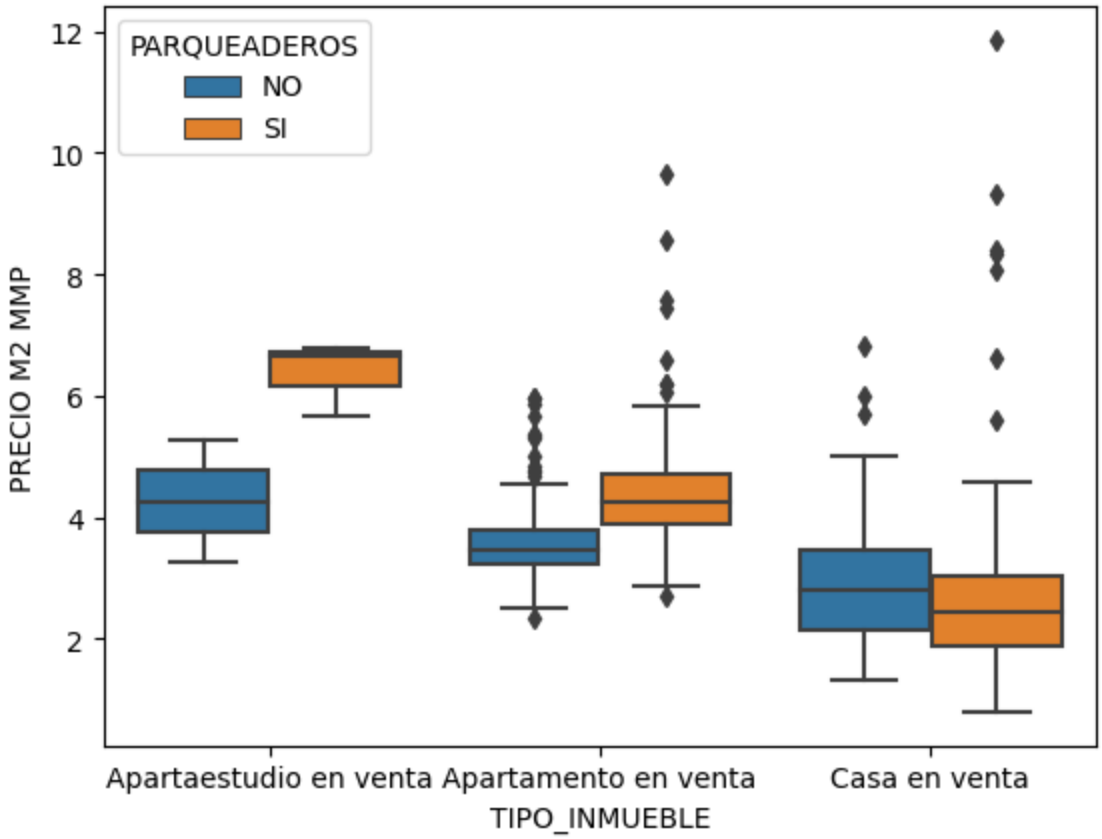
- La variable objetivo "PRECIO M2 MMP" presenta una distribución que se acerca a una distribución normal, con algunos valores atipicos hacia la derecha.
- Existe una tendencia a disminuir el precio por m2 a medida que el area construida aumenta, esto se ratifica en el diagrama de correlación.

Análisis de Variables Categóricas

Para el análisis de las variables categóricas se hará uso de diagramas de caja y bigotes

In [15]:
```
sns.boxplot(x="TIPO_INMUEBLE", y="PRECIO M2 MMP", data=df_w, hue="PARQUEADEROS")
```

Out[15]:
```
<Axes: xlabel='TIPO_INMUEBLE', ylabel='PRECIO M2 MMP'>
```

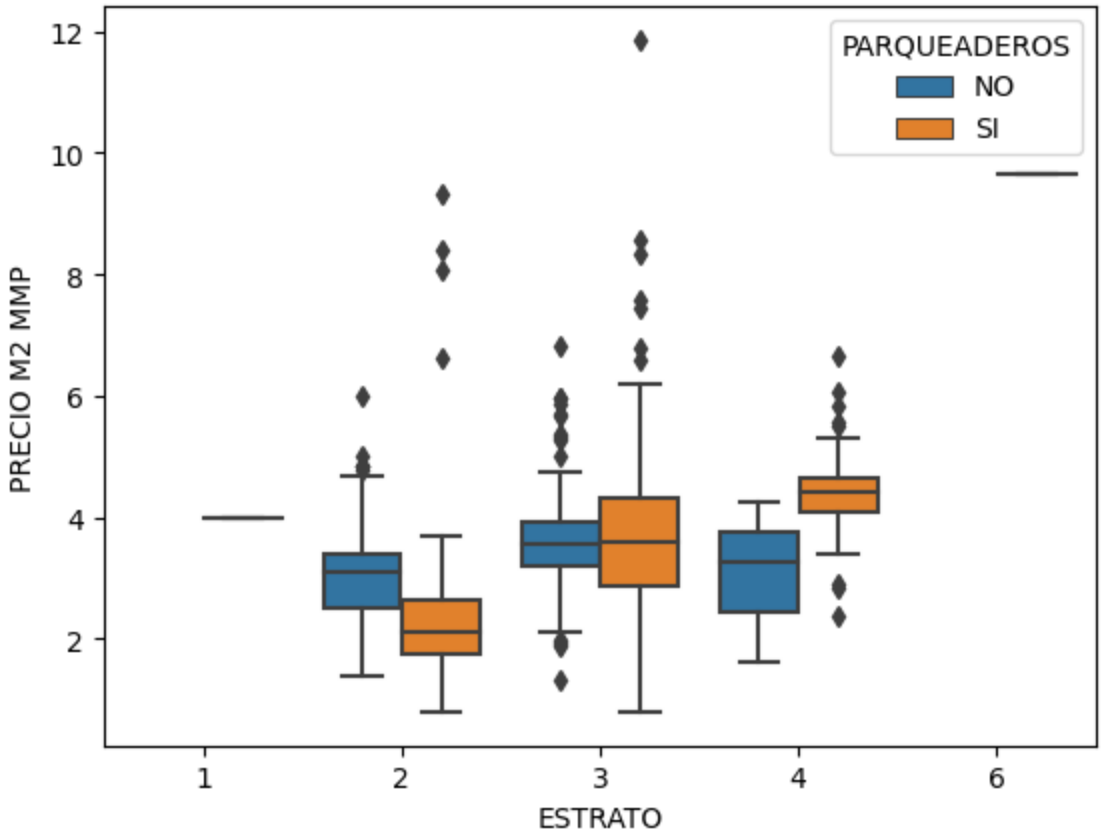


Este diagrama nos indica que:

- A nivel general el precio por metro cuadrado es más alto en los apartaestudios, seguido por los inmuebles tipo apartamento y por último inmuebles tipo casa.
- En los inmuebles tipo apartaestudio y tipo apartamento se evidencia un aumento del precio dependiendo si cuenta con parqueadero.
- La mediana en el precio por metro cuadrado de un apartaestudio sin parqueadero es similar a la de una apartamento con parqueadero.
- Los valores atípicos más altos se encuentran en los inmuebles tipo apartamento y tipo casa en los que se cuenta con parqueadero

In [16]:
```
sns.boxplot(x="ESTRATO", y="PRECIO M2 MMP", data=df_w, hue="PARQUEADEROS")
```

Out[16]:
```
<Axes: xlabel='ESTRATO', ylabel='PRECIO M2 MMP'>
```

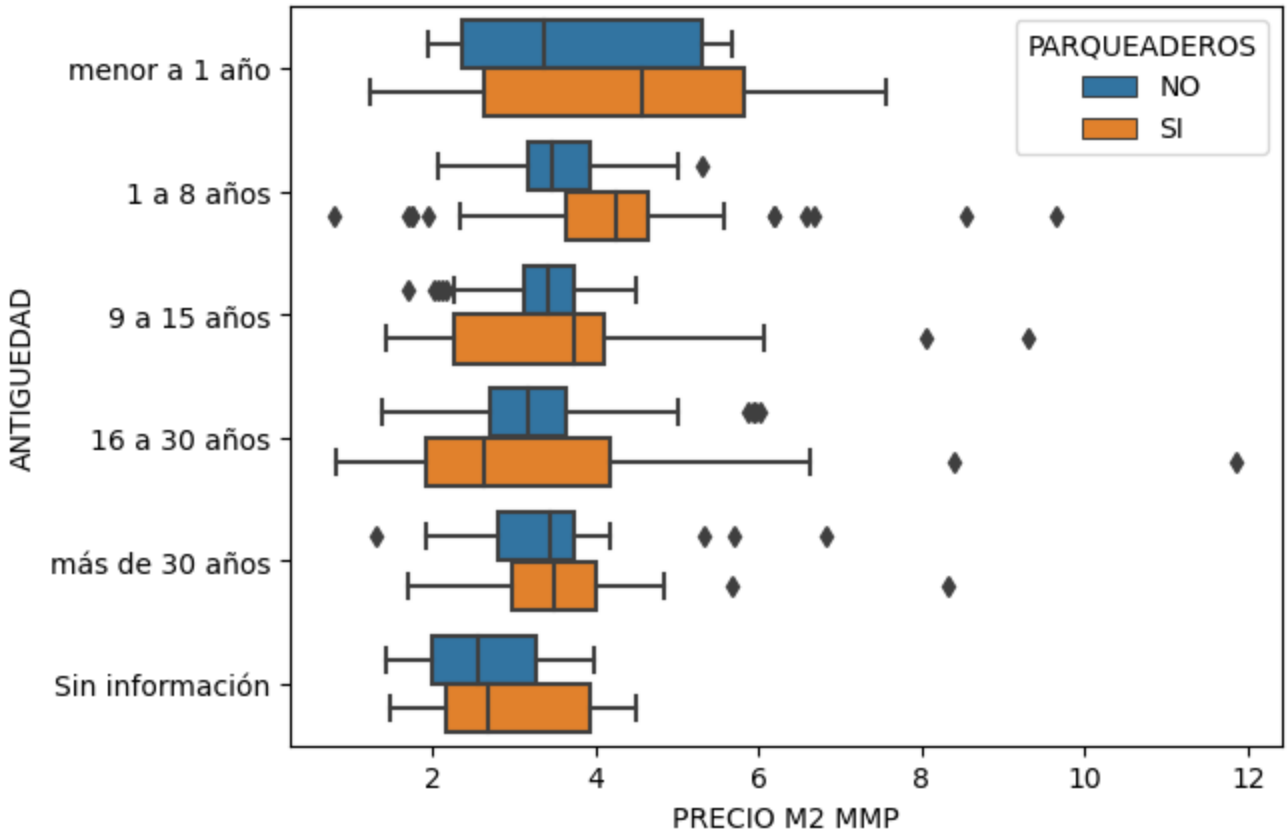


Este diagrama nos indica que:

- A nivel general el precio por metro cuadrado incrementa a medida que aumenta el estrato donde se encuentra ubicado el inmueble
- Los inmuebles de estrato 4 presentan incremento en el valor por m2 cuando estos cuentan con parqueadero
- Los valores atípicos más altos se encuentran en los inmuebles ubicados en estrato 2 y 3

In [17]:
```
sns.boxplot(y="ANTIGUEDAD", x="PRECIO M2 MMP",
            data=df_w, hue="PARQUEADEROS",
            order=["menor a 1 año","1 a 8 años", "9 a 15 años", "16 a 30 años", "más de
```

Out[17]:
```
<Axes: xlabel='PRECIO M2 MMP', ylabel='ANTIGUEDAD'>
```



Este diagrama nos indica que:

- A nivel general el precio por metro cuadrado más costoso se encuentra en inmuebles con tiempo de construcción menor a un año
- El único grupo en el cual se presenta una variación marcada de acuerdo a si tiene parqueadero, es en los inmuebles que se han construido entre 1 a 8 años.
- Los valores atípicos más altos se encuentran en los inmuebles construidos entre 16 a 30 años

5. Resolución del problema

Como resultado del análisis de los datos presentes en este dataset podemos concluir que:

- A nivel general el precio más favorable por m2 se presenta en inmuebles tipo casa.
- Es mas favorable el precio por metro cuadrado en inmuebles con mayor área construida
- En cuanto al estrato en el cual se encuentra construido en inmueble, este no tiene mayor inferencia en cuanto al precio por m2.
- El precio por m2 promedio en esta zona se encuentra alrededor de los 3.5 millones
- El dataset permite hacer el análisis y responder al problema planteado

Se exporta dataset con las adecuaciones realizadas a un archivo .csv

In [18]:
```
# Se exporta dataFrame de url a archivo CSV
df_w.to_csv('Data/archivo_clean_inmuebles.csv', index = False)
```

Tabla contribuciones

Contribuciones	--- Firma ---
Investigación previa	
Redacción de las respuestas	
Desarrollo del código	
Participación en el vídeo