

Informe sobre identificación de las variables y los componentes estadísticos a partir de una situación planteada. AA3-EV02

Nicolas Cortes Parra
Ingeniero físico

TUTOR(A):
Mayra Patricia Amadotorres

Análisis exploratorio de datos en python. (3176972)
Servicio Nacional de Aprendizaje
SENA
2025

Introducción

Para la actividad, el propósito del análisis que se lleva a cabo en presentar un análisis exploratorio de datos como un caso propuesto de un proyecto de implementación de las técnicas de Machine Learning que tiene la empresa A&A Ltda. La actividad plantea revisar y documentar la base de datos, identificando patrones, relaciones o incoherencias. Por lo que se debe de realizar diferentes actividades, además de plantear preguntas sobre la investigación de los datos. La base de datos objeto de análisis, en este caso, contiene la información del precio de los inmuebles tanto de viviendas como de locales en venta, en donde a su vez, están presentes numerosas variables que podrían influir de una u otra manera en estos valores.

Importación de los datos

Para esto bajamos los datos de la carpeta anexos del curso del sena o de la pagina:

<https://www.datos.gov.co/Hacienda-y-Credito-Publico/Inmuebles-Disponibles-Para-La-Venta/72gdp-px77/data> en donde estan alojados los datos. Dichos datos estan marcados con el nombre de **Data_Caso_Propuesto.csv**

Como se puede evidenciar en la siguiente figura, se observa el contenido de la carpeta, más los archivos notebook para el análisis de los datos. Esto se hace ya que para poder importar la base de datos se requiere el notebook y la base de datos estén en la misma locación o carpeta.



Figura 1. Archivos de la carpeta Anexos

Teniendo la base de datos y el notebook creado, abrimos la aplicación Anaconda y entramos a JupyterLabs para abrir el entorno, para esto primero importamos las librerías pertinentes, como se puede ver en la siguiente figura.

```
# importamos librerias
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

Figura 2. Importación de librerías siendo pandas, matplotlib, seaborn y numpy

Las bibliotecas más relevantes para la analítica en datos en el lenguaje Python son las siguientes: la biblioteca Pandas, que se utiliza para manipular y visualizar grandes volúmenes de datos (import pandas as pd); la biblioteca Matplotlib, que permite crear sencillamente diferentes tipos de gráficos (import matplotlib.pyplot as plt); la biblioteca Seaborn, que es una biblioteca construida sobre la biblioteca Matplotlib que permite mejores visualizaciones estadísticas (import seaborn as sns); y la biblioteca NumPy, que se utiliza para trabajar con funciones matemáticas, vectores y matrices (import numpy as np). Teniendo claro estos conceptos, ahora si partamos a leer la base de datos, como se ve en la siguiente figura.

```

# Importamos la base de datos: Data_Caso_Propuesto.csv
df = pd.read_csv('Data_Caso_Propuesto.csv')
df.head(3) # Leer los datos

```

Codigo	Ciudad	Departamento	Barrio	Direccion	Area Terreno	Area Construida	Detalle Disponibilidad	Estrato	Precio	Tipo de Inmueble	Datos Adicionales	
0	17180	BOGOTÁ	CUNDINAMARCA	N/A	AV CB 7 NO. 106 - 51 LT 8	0.00	0.0	COMERCIALIZABLE CON RESTRICCION	TRES	2.958891e+10	LOTE COMERCIAL	ESTE INMUEBLE SE COMERCIALIZARA A TRAVES DE SU...
1	19292	BOGOTÁ	CUNDINAMARCA	N/A	CL 72 No. 12 - 77	0.00	0.0	COMERCIALIZABLE VENTA ANTICIPADA	COMERCIAL	1.646059e+10	EDIFICIO	N/A
2	19292	BOGOTÁ	CUNDINAMARCA	N/A	CL 72 No. 12 - 77	0.00	0.0	COMERCIALIZABLE VENTA ANTICIPADA	COMERCIAL	1.646059e+10	EDIFICIO	N/A
3	2575	BOGOTÁ	BOYACÁ	CENTRO	CRA 10 #75-7800 O CL 12 # 9 - 7708 O CALLE...	1655.08	7269.0	COMERCIALIZABLE CON RESTRICCION	CUATRO	1.376828e+10	CLINICA	ESTE INMUEBLE SE COMERCIALIZARA A TRAVES DE SU...
4	11409	BUGA	VALLE DEL CAUCA	VEREDA CHAMIMBAL	LT 45-434 85-879 C1-151 01-09 STA 905A LT1-48 ...	32719700	22724.0	COMERCIALIZABLE FIDUCIA	RURAL	4.523379e+10	LOTE MIXTO	N/A

Figura 3. Lectura y visualización de la base de datos.

Podemos identificar que las columnas están divididas por el código, ciudad, departamento, barrio, dirección, área terreno, área construida, detalle disponibilidad, estrato, precio, tipo de inmueble y datos adicionales. Ahora bien, para saber que tipo de datos son, vamos a sacar el comando info, para analizar y de esta manera plantear las preguntas que permitirán el desarrollo y tratamiento de datos a escoger en el análisis, como se puede ver en la siguiente figura.

```
df.info() # observamos la informacion del contenido de los datos
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 463 entries, 0 to 462
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Codigo                 463 non-null    int64
1   Ciudad                 463 non-null    object
2   Departamento           463 non-null    object
3   Barrio                 40 non-null     object
4   Direccion              463 non-null    object
5   Area Terreno           463 non-null    float64
6   Area Construida        463 non-null    float64
7   Detalle Disponibilidad 463 non-null    object
8   Estrato                463 non-null    object
9   Precio                 463 non-null    float64
10  Tipo de Inmueble        463 non-null    object
11  Datos Adicionales       118 non-null    object
dtypes: float64(3), int64(1), object(8)
memory usage: 43.5+ KB
```

Figura 4. Información de la base de datos

Podemos identificar que existen 4 datos numéricos siendo **código**, **área de terreno**, **área construida** y por último el **precio**. Ahora bien para las variables categoría se encontró que los datos son 8, siendo **ciudad**, **departamento**, **barrio**, **dirección**, **detalle disponibilidad**, **estrato**, **tipo de inmueble** y **datos adicionales**. Adicionalmente también encontramos que existe 463 datos o registros en la base de datos, por tanto, para saber si existe datos duplicados y datos nulos o vacíos hacemos los siguientes comandos como se puede observar en la siguiente figura.

```
df.isnull().sum() # Revisar valores nulos
Codigo                0
Ciudad                0
Departamento         0
Barrio                423
Direccion              0
Area Terreno          0
Area Construida       0
Detalle Disponibilidad 0
Estrato               0
Precio                0
Tipo de Inmueble      0
Datos Adicionales     345
dtype: int64

# miramos que datos duplicados hay
df.duplicated()
0    False
1    False
2    False
3    False
4    False
...
458  False
459  False
460  False
461  False
462  False
Length: 463, dtype: bool

df.duplicated().sum() # permite sumar cuantos datos repetidos hay
0
```

Figura 5. Información de datos nulos o vacíos y datos duplicados.

Como se puede observar en la figura 5 se encuentra que para datos nulos o vacíos están en la columna de Barrio y de datos adicionales, esto se puede realizar llamando el comando `isnull().sum()` para conocer la cantidad de datos nulos o vacíos de la base de datos `df`, de modo que encontramos 423 datos de barrio representa un 91.34% de datos faltantes y 345 de

datos adicionales representa 74.51% de datos que indica que existan valores faltantes, además también se puede observar que para los valores duplicados, se marcan como cero, por tanto, es evidente que no existen datos duplicados. Por tanto ya podemos plantear que los datos de barrio y datos adicionales se pueden descartar.

Planteamiento de preguntas sobre el análisis

- ¿Existen datos que se puedan eliminar o se consideren un error para los resultados?
- ¿Los datos me permiten tomar decisiones con respecto a relaciones que encuentre entre ellos mismo y sus pares?
- ¿Cuál es la distribución de precios de los inmuebles en venta?
- ¿Cómo varía el precio de los inmuebles según el estrato socioeconómico?
- ¿Qué ciudades concentran los inmuebles de mayor y menor precio?
- ¿Qué relación existe entre el precio y el área construida?
- ¿Existen valores atípicos en los precios o en el área construida? ¿Cómo impactan en el análisis?

Tratamiento de datos

Teniendo nuestras preguntas objetivos, pasamos a la etapa de análisis de los datos, por lo que primero vamos a detectar o analizar cuales son nuestros datos o valores que pueden presentarse como un error para el análisis y exploración. Para esto vamos a partir de los datos nulos o vacíos, ya que por lo general se pueden considerar como variables o datos no deseados en el análisis por qué pueden presentar errores estadísticos a futuro, por ende vamos a llamar las columnas de barrios y datos adicionales y vamos a utilizar el comando `sample` para mirar datos aleatorios y tomar una decisión.

```
df[['Barrio', 'Datos Adicionales']].sample(10)
```

	Barrio	Datos Adicionales
437	NaN	NaN
70	NaN	NO PIERDAS LA OPORTUNIDAD DE UBICAR TU OFICINA...
454	VEREDA PALERMO	LOTE DE TERRENO QUE HACE PARTE DEL PROYECTO CA...
323	NaN	NaN
438	NaN	NaN
109	NaN	NaN
117	NaN	NaN
383	NaN	NaN
52	NaN	EL EDIFICIO LANCASTER HOUSE PH. DEL QUE HACE P...
187	NaN	NaN

Figura 6. Datos Barrio y Datos adicionales

Como se puede observar llamamos las columnas y buscamos un orden aleatorio para ver como se presentan los datos y encontramos que hay varios parámetros NaN lo cual indica que son variables nulas o que no hay información, por lo tanto, se puede considerar como datos que se puede eliminar. De modo que se utiliza la variable drop para eliminar dichos valores, además de mandar un .info para observar si hay algún cambio en los datos de la base de datos, adicionalmente también se descarta eliminar datos duplicados, ya que la base de datos no presenta dichos datos.

```
# no se eliminan datos duplicados, por que no hay, no obstante se pone el codigo
df.drop(columns=['Barrio','Datos Adicionales'], inplace=True)
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 463 entries, 0 to 462
Data columns (total 10 columns):
#   Column                Non-Null Count  Dtype
---  ---                ---
0   Codigo                463 non-null   int64
1   Ciudad                463 non-null   object
2   Departamento          463 non-null   object
3   Direccion             463 non-null   object
4   Area Terreno          463 non-null   float64
5   Area Construida       463 non-null   float64
6   Detalle Disponibilidad 463 non-null   object
7   Estrato               463 non-null   object
8   Precio               463 non-null   float64
9   Tipo de Inmueble      463 non-null   object
dtypes: float64(3), int64(1), object(6)
memory usage: 36.3+ KB
```

Figura 7. Información de la base de datos, en donde se observa que los datos barrio y datos adicionales ya no hacen parte de la base de datos.

Reporte estadístico de datos

Una vez realizado todos los tratamientos de los datos pasamos a observar un reporte estadístico para calcular parámetros como media, moda, mediana, desviación estándar, cuartiles, entre otros y posterior realizar gráficos que permitan identificar parámetros como variables atípicas y plantear respuestas a las preguntas objetivo. Para esto vamos a utilizar la variable describe, para realizar dicho

reporte, como se puede observar en la siguiente figura

```
df.describe()
```

	Codigo	Area Terreno	Area Construida	Precio
count	463.000000	4.630000e+02	463.000000	4.630000e+02
mean	18003.151188	1.515204e+04	87.517279	6.672032e+08
std	1992.191499	1.827101e+05	1137.469077	3.272992e+09
min	2575.000000	0.000000e+00	0.000000	4.650000e+06
25%	18184.500000	0.000000e+00	0.000000	1.230500e+07
50%	18332.000000	0.000000e+00	0.000000	1.587000e+07
75%	18539.500000	0.000000e+00	0.000000	1.379955e+08
max	19344.000000	3.217197e+06	22724.000000	4.523379e+10

Figura 8. Reporte estadístico de las variables numéricas de la base de datos.

Como se puede apreciar en la figura 8, se puede prescindir de la variable código, debido a que no nos brinda información que sea fundamental para el objetivo de este análisis. Sin embargo, encontramos datos interesantes como que el promedio del precio es de aproximadamente 60 millones de pesos, lo que nos va a permitir en un futuro determinar rangos de precios y buscar su asociación con otras variables como la del estrato o la ciudad.

En el caso de las variables área de terreno y área construida, se puede apreciar que existen en muchos casos situaciones sin información sobre construcciones y, en otros, valores extremadamente altos. En esos casos, resulta útil observar la desviación típica ya que nos indica que, en el caso del área de terreno, hay muy muy altos, lo que da la impresión de que se trata de propiedades con extensas dimensiones. En el área construida podemos observar que la desviación pasaba por los 1137 m2, lo que da la impresión de que podría encontrarse con edificaciones significativas, implicando a su vez que podría ser una oportunidad: el fenómeno de valorización en zonas de cierto desarrollo estructural podría implicar una oportunidad de inversión inmobiliaria.

Graficas

Ahora teniendo una noción de los datos reportados, pasamos a la creación de gráficos, para comprender los datos presentados, además de crear relaciones

entre los datos y así poder ir respondiendo las preguntas objetivos planteadas.

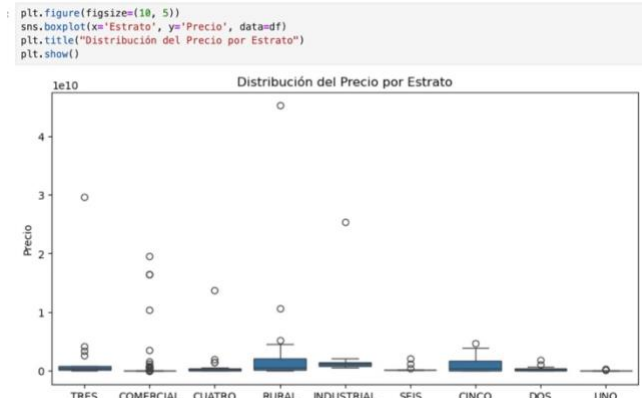


Figura 9. Identificación de valores atípicos, mostrando la relación del precio, con respecto a los estratos

Como se puede observar en la figura 9 se encuentran que los datos presentan una gran cantidad de datos atípicos en la parte de estrato comercial, como también en otros estrados, lo que indican que hay precios muy elevados, lo cual esto podría generar dificultades para encontrar posibles acciones económicas como toma de decisiones importantes al momento de invertir para la compañía.

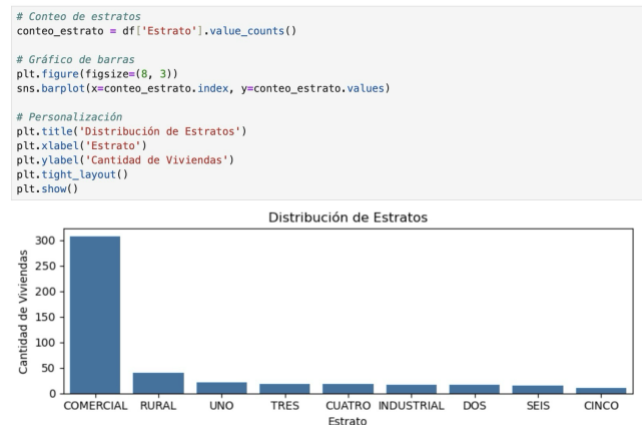


Figura 10. Visualización de la distribución de los estratos

Como se puede observar en la figura, se encuentra que la distribución de los estratos, se encuentra que la gran mayoría de los inmuebles, están en el grupo de comercial, lo cual esto podría generar gran interés para la inversión en dichas propiedades.

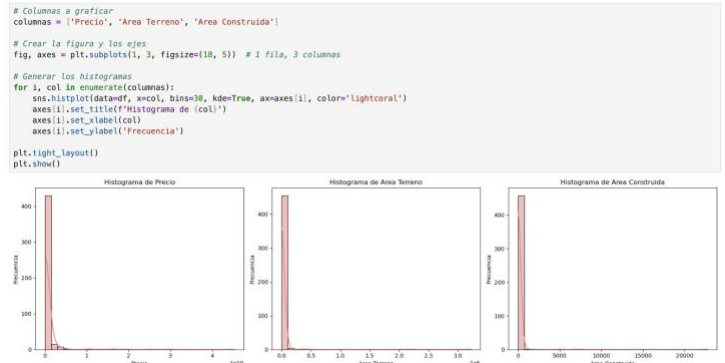


Figura 11. Histograma de las variables numéricas de la base de datos.

Encontramos que la tendencia de los datos presenta una gran cantidad de datos con precios y áreas cuadradas con una gran cantidad de valores numéricos, por lo que estas gráficas no nos permitiría encontrar una información relevante para toma de decisiones, por lo tanto se necesitan encontrar otro tipo de relaciones gráficas que reflejen mas información.

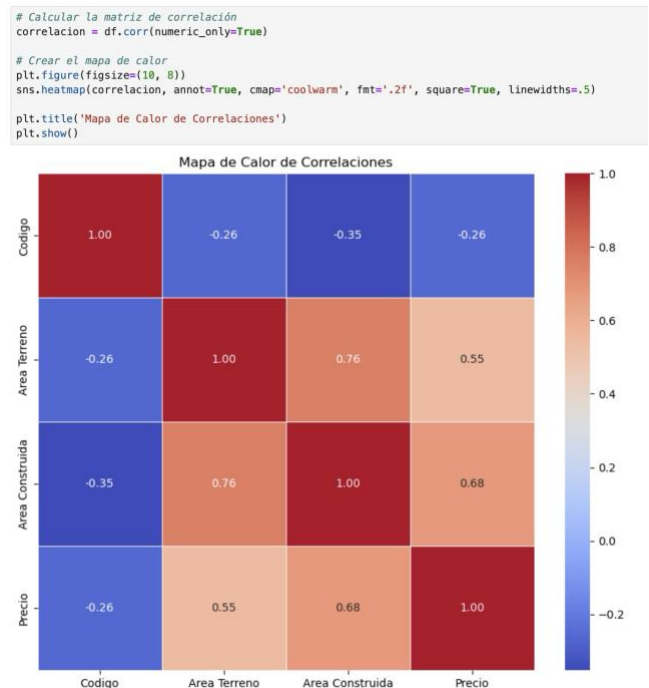


Figura 12. Mapa de calor para encontrar correlaciones entre las variables numéricas

Como las distribuciones no reflejo información relevante, se procede a realizar una correlación de los datos numéricos, en donde se encuentra que existe una fuerte relación con el precio y el área construida, con un valor de 0.68 el cual le sigue el

área del terreo con un 0.55, esto es congruente ya que los bienes aumentan su valor a mayor sección de terreno y de construcción.

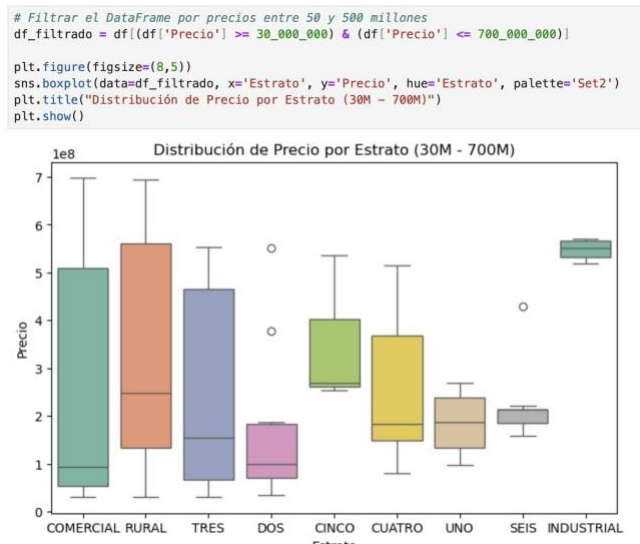


Figura 13. Relación de los estratos, con respecto a un rango de precios

Se puede encontrar que gran parte de los precios, entre el rango de 30 millones a 700 millones, están entre los estratos comerciales, rula y el estrato tres, esto podría indicar posibles opciones de inversiones para estos estratos.

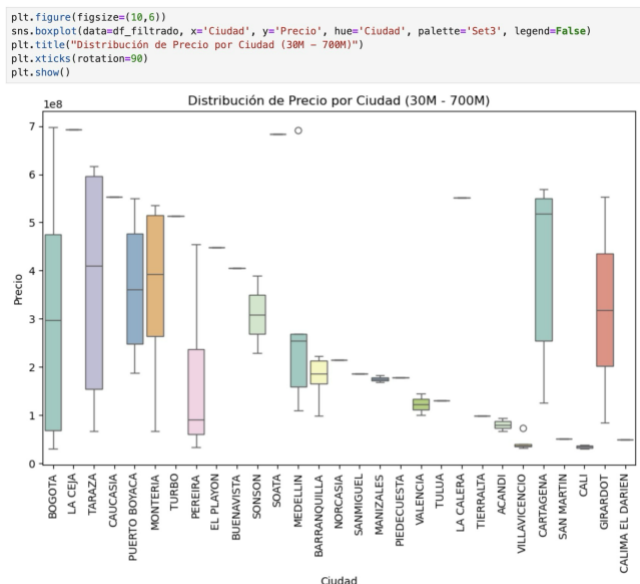


Figura 14. Relación de los precios con las ciudades

Se puede identificar que para las ciudades principales, albergan la gran mayoría de los precios,

entre los rangos de 30 a 700 millones, lo cual se pueden encontrar relaciones claras con los estratos.

```
df.sort_values(by='Precio', ascending=True, inplace=True)
df.head(5)
```

Codigo	Ciudad	Departamento	Direccion	Area Terreno	Area Construida	Detalle Disponibilidad	Estrato	Precio	Tipo de Inmueble
455	17009	VALENCIA	CÓRDOBA CALLE PRINCIPAL - CORREJIMIENTO GUADAL	0.0	0.0	COMERCIALIZABLE	RURAL	4650000.0	LOTE AGRICOLA
453	17337	PEREIRA	RISARALDA CL 69 O ACCESO A CUBA AV 30 DE AGOSTO LOTE...	0.0	0.0	COMERCIALIZABLE CON RESTRICCION	DOS	6333900.0	LOTE VIVIENDA
451	17336	PEREIRA	RISARALDA CL 69 O ACCESO A CUBA AV 30 DE AGOSTO LOTE 13	0.0	0.0	COMERCIALIZABLE CON RESTRICCION	DOS	6832720.0	LOTE VIVIENDA
450	18447	VILLAVICENCIO	META CENTRAL MINORISTA DE ABASTOS DE VILLAVICENCIO ...	0.0	0.0	COMERCIALIZABLE	COMERCIAL	6835500.0	LOCAL

```
df['Precio'].max()
4523789829.0

df['Precio'].min()
4050000.0
```

Figura 15. Ordenamiento de datos, utilizando el precio, así mismo encontrando los valores mínimos y máximos de la base de datos

Se plantea un ordenamiento de datos, tomando como parámetro el precio, para posterior encontrar rangos de precios y crear gráficas, que nos permita analizar los parámetros de los precios con respecto a cada uno de los estratos. Para esto, primero se tiene que crear variables para crear los rangos de los precios, luego crear las etiquetas que en este caso las llamares bajo, medio, alto, muy alto y lujo. Esto se puede evidenciar en la siguiente figura.

```
# Crear rangos personalizados
Rangos_Precios = [0, 10_000_000, 100_000_000, 700_000_000, 1_000_000_000, df['Precio'].max()]
Etiquetas = ['Bajo', 'Medio', 'Alto', 'Muy Alto', 'Lujo']

# Crear nueva columna con los rangos de precio
df['Rango_Precio'] = pd.cut(df['Precio'], bins=Rangos_Precios, labels=Etiquetas, include_lowest=True)

# Contar cuántos datos hay por cada rango
conteo_rangos = df['Rango_Precio'].value_counts().sort_index()
```

Figura 16. Creación de la variable rangos de precios y de etiquetas para posterior, agregarlo en la base de datos como un nuevo parámetro categórico

```
df.sample(5)
```

Codigo	Ciudad	Departamento	Direccion	Area Terreno	Area Construida	Detalle Disponibilidad	Estrato	Precio	Tipo de Inmueble	Rango_Precio
302	18185	VILLAVICENCIO	META CENTRAL MINORISTA DE ABASTOS DE VILLAVICENCIO ...	0.00	0.0	COMERCIALIZABLE	COMERCIAL	1.230500e+07	LOCAL	Medio
216	18454	VILLAVICENCIO	META CENTRAL MINORISTA DE ABASTOS DE VILLAVICENCIO ...	0.00	0.0	EN PUJA	COMERCIAL	1.587000e+07	LOCAL	Medio
117	18445	VILLAVICENCIO	META CENTRAL MINORISTA DE ABASTOS DE VILLAVICENCIO ...	0.00	0.0	COMERCIALIZABLE	COMERCIAL	3.133550e+07	LOCAL	Medio
452	11335	CALI	VALLE DEL CAUCA LT B PARTE DEL PREDIO CHIPCHAPPE	430399.68	0.0	COMERCIALIZABLE FIDUCIA	RURAL	1.667784e+09	LOTE VIVIENDA	Lujo
426	18314	VILLAVICENCIO	META CENTRAL MINORISTA DE ABASTOS DE VILLAVICENCIO ...	0.00	0.0	COMERCIALIZABLE	COMERCIAL	7.350500e+06	LOCAL	Bajo

Figura 17. Observación de los datos, con la nueva columna en donde se evidencia el uso correcto de las etiquetas categóricas de los datos con respecto a los precios

```
# Visualizar con estrato
plt.figure(figsize=(8,5))
sns.countplot(data=df, x='Rango_Precio', hue='Estrato', palette='Set1')
plt.title("Cantidad de inmuebles por rango de precio y estrato")
plt.ylabel("Numero de inmuebles")
plt.show()
```

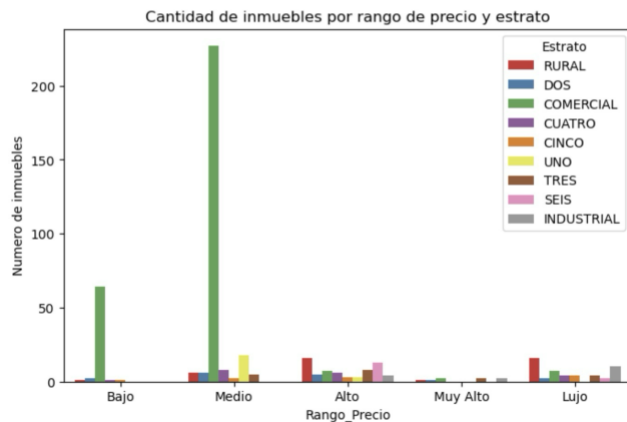


Figura 18. Grafico de relación de rango de precios con respecto al numero de inmuebles con respecto a su categoría

```
import warnings
warnings.filterwarnings('ignore') # La gráfico actual no necesita hue

# Gráfico de barras
plt.figure(figsize=(8, 5))
sns.barplot(x=conteo_rangos.index, y=conteo_rangos.values, palette='Set2')
plt.title("Cantidad de Inmuebles por Rango de Precio")
plt.xlabel("Rango de Precio")
plt.ylabel("Cantidad de Inmuebles")
plt.show()
```

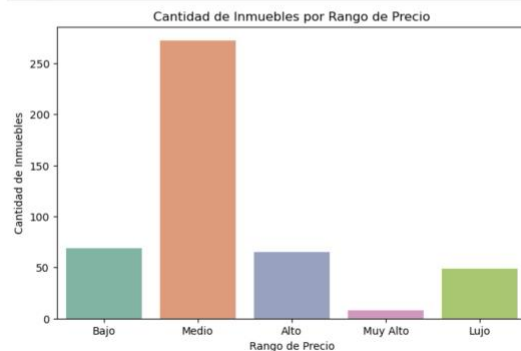


Figura 19. Distribución de los datos con respecto a la cantidad de inmuebles, con respecto a los categorías creadas con respecto los precios

```
# Gráfico diagrama de torta
colores = sns.color_palette('Set2')

plt.figure(figsize=(7, 7))
plt.pie(conteo_rangos, labels=conteo_rangos.index, autopct='%1.1f%%', startangle=90, colors=colores)
plt.title("Distribución de Precios por Rangos")
plt.axis('equal')
plt.show()
```

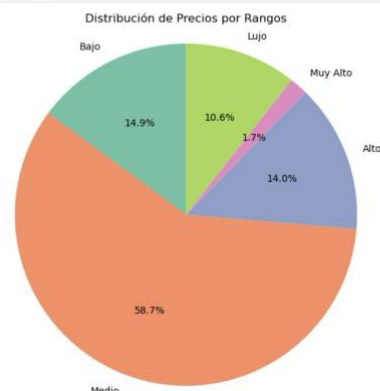


Figura 20. Distribución porcentual de las categorías de los rangos de precios

Se puede observar que en las figuras 18 , 19 y 20 se ve reflejado como los valores de la categoría medio, abarcan una gran cantidad de los datos, esto implica que gran parte de los precios de los inmuebles no superan los 100 millones de pesos, adicional mente se puede observar que gran parte de los inmuebles que presentan la categoría medio, son de tipo estrato comercial y otra parte están en la categoría bajo.

Conclusiones

Se puede concluir que para la exploración de los datos, las herramientas estadísticas y graficas como los diagramas de barras, los bigotes y los mapas de calor, permiten la identificación de parámetros que encaminan a la observación y descubrimiento para entender, comprender y encontrar parámetros que permitan la identificación y respuesta de la interacción de los datos, lo que permite explotar conclusiones fuerte entre la relación y la influencia de los datos. Adicional mente también el tratamiento y compresión de los datos en una fase preliminar, permite el manejo de los mismo sin que se presenten problemas al momento de realizar análisis y graficar los datos. Por último la utilización de las librerías y el manejo del programa Jupyter como el lenguaje Python permiten el desarrollo y facilidad para realizar operaciones matemáticas, gráficas y de gestión de los datos para su máxima aplicación y eficiencia.

Bibliografía

- Central de Inversiones S. A. (2017). Inmuebles Disponibles Para La Venta [Data set]. Tomado de https://www.datos.gov.co/Hacienda-y-Credito-P-blico/Inmuebles-Disponibles-Para-La-Venta/72gd-px77/about_data
- Sena, I. E. (2025). Recursos y herramientas para el análisis efectivo de datos. Repositorio.