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-Cr-dito-P-blico/Inmuebles-Disponibles-Para-La-Venta/72gd-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.

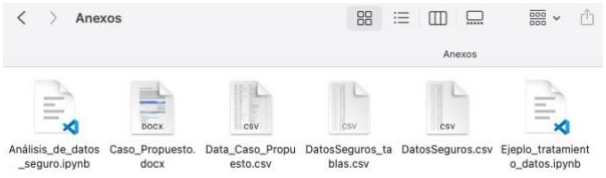

Figura 1. Archivos de la carpeta Anexos

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(5) # Leer los datos
```

	Codigo	Ciudad	Departamento	Barrio	Direccion	Area Terreno	Area Construida	Detalle Disponibilidad	Estrato	Precio	Tipo de Inmueble	Datos Adicionales
0	17180	BOGOTA	CUNDINAMARCA	NaN	AV CR 7 NO. 166 - 51 LT B	0.00	0.0	COMERCIALIZABLE CON RESTRICCION	TRES	2.958081e+10	LOTE COMERCIAL	ESTE INMUEBLE SE COMERCIALIZARÁ A TRAVES DE SU...
1	19292	BOGOTA	CUNDINAMARCA	NaN	CL 72 No. 12 - 77	0.00	0.0	COMERCIALIZABLE	COMERCIAL	1.646059e+10	EDIFICIO	NaN
2	19292	BOGOTA	CUNDINAMARCA	NaN	CL 72 No. 12 - 77	0.00	0.0	COMERCIALIZABLE VENTA ANTICIPADA	COMERCIAL	1.646059e+10	EDIFICIO	NaN
3	2575	SOGAMOSO	BOYACÁ	CENTRO	CRA. 10 #11-78/80 Ó CL 12 # 9 - 77/85 Ó CALLE...	1655.08	7269.0	COMERCIALIZABLE CON RESTRICCION	CUATRO	1.376828e+10	CLINICA	ESTE INMUEBLE SE COMERCIALIZARÁ A TRAVÉS DE SU...
4	11409	BUGA	VALLE DEL CAUCA	VEREDA CHAMBIMBAL	LT A1-A24 B1-B79 C1-C51 D1-D9 STA ROSA LT1-46 ...	3217197.00	22724.0	COMERCIALIZABLE FIDUCIA	RURAL	4.523379e+10	LOTE MIXTO	NaN

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 encontraos 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 dados 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 667 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 asi poder ir respondiendo las preguntas objetivos planteadas.

```
plt.figure(figsize=(10, 5))
sns.boxplot(x='Estrato', y='Precio', data=df)
plt.title("Distribución del Precio por Estrato")
plt.show()
```

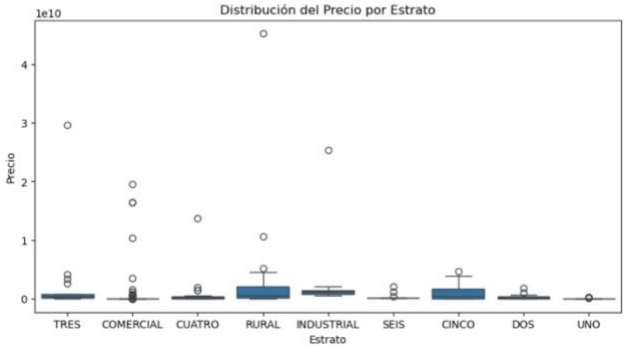

Figura 9. Identificación de valores atípicos, mostrando la relación del precio, con respecto a los estratos

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.

```
# Conteo de estratos
conteo_estrato = df['Estrato'].value_counts()

# Gráfico de barras
plt.figure(figsize=(8, 3))
sns.barplot(x=conteo_estrato.index, y=conteo_estrato.values)

# Personalización
plt.title('Distribución de Estratos')
plt.xlabel('Estrato')
plt.ylabel('Cantidad de Viviendas')
plt.tight_layout()
plt.show()
```

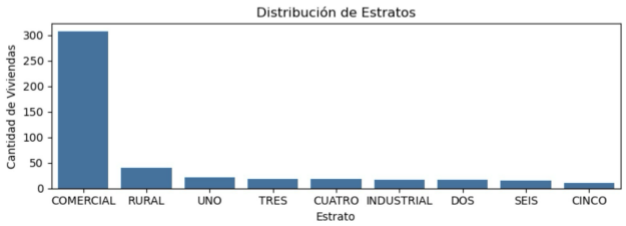

Figura 10. Visualización de la distribución de los estratos

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.

```
# Columnas a graficar
columnas = ['Precio', 'Area Terreno', 'Area Construida']

# Crear la figura y los ejes
fig, axes = plt.subplots(1, 3, figsize=(18, 5))  # 1 fila, 3 columnas

# Generar los histogramas
for i, col in enumerate(columnas):
    sns.histplot(data=df, x=col, bins=30, kde=True, ax=axes[i], color='lightcoral')
    axes[i].set_title(f'Histograma de {col}')
    axes[i].set_xlabel(col)
    axes[i].set_ylabel('Frecuencia')

plt.tight_layout()
plt.show()
```

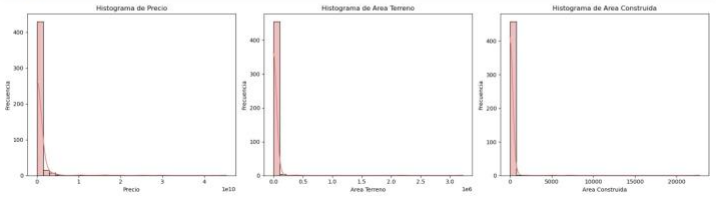

Figura 11. Histograma de las variables numéricas de la base de datos.

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.

```
# Calcular la matriz de correlación
correlacion = df.corr(numeric_only=True)

# Crear el mapa de calor
plt.figure(figsize=(10, 8))
sns.heatmap(correlacion, annot=True, cmap='coolwarm', fmt='.2f', square=True, linewidths=.5)

plt.title('Mapa de Calor de Correlaciones')
plt.show()
```

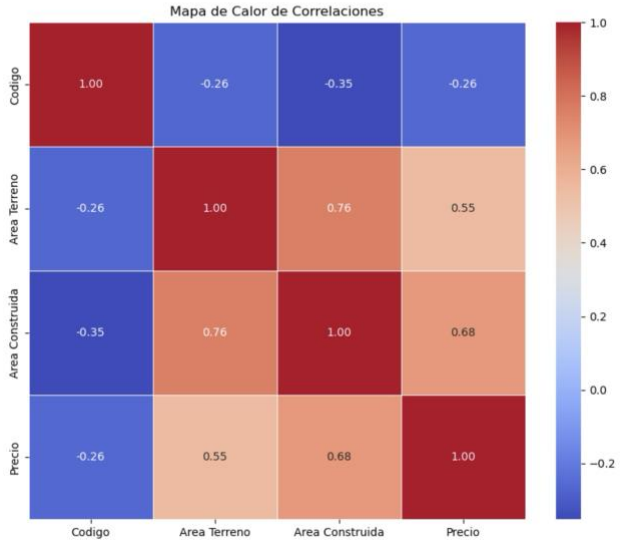

Figura 12. Mapa de calor para encontrar correlaciones entre las variables numéricas

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.

```
# Filtrar el DataFrame por precios entre 50 y 500 millones
df_filtrado = df[(df['Precio'] >= 30_000_000) & (df['Precio'] <= 700_000_000)]

plt.figure(figsize=(8,5))
sns.boxplot(data=df_filtrado, x='Estrato', y='Precio', hue='Estrato', palette='Set2')
plt.title("Distribución de Precio por Estrato (30M - 700M)")
plt.show()
```

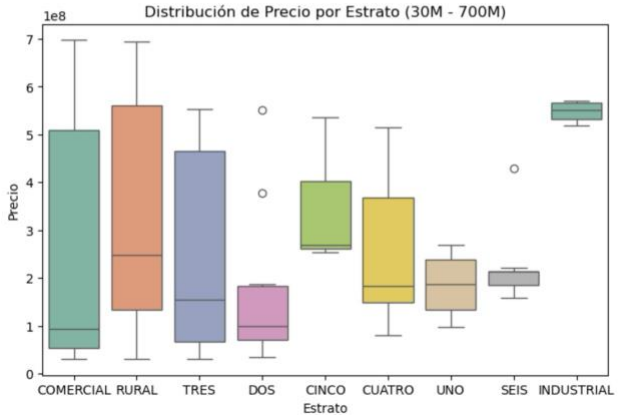

Figura 13. Relación de los estratos, con respecto a un rango de precios

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.

```
plt.figure(figsize=(10,6))
sns.boxplot(data=df_filtrado, x='Ciudad', y='Precio', hue='Ciudad', palette='Set3', legend=False)
plt.title("Distribución de Precio por Ciudad (30M - 700M)")
plt.xticks(rotation=90)
plt.show()
```

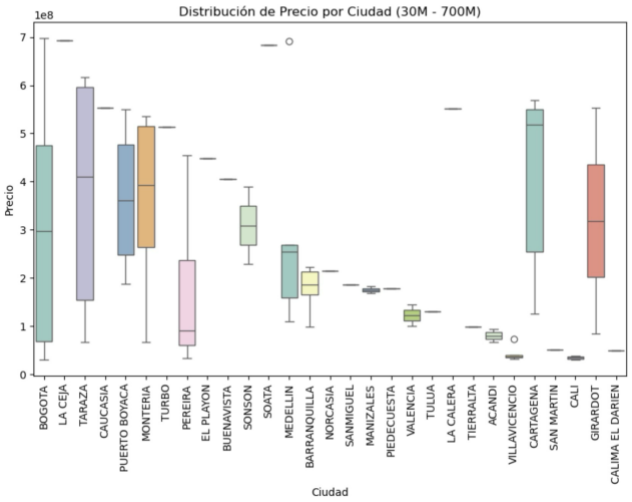

Figura 14. Relación de los precios con las ciudades

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 - CORREGIMIENTO GUADUAL	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
491	17336	PEREIRA	RISARALDA	CL 69 Ó 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
449	18137	VILLAVICENCIO	META	CENTRAL MINORISTA DE ABASTOS DE VILLAVICENCIO ...	0.0	0.0	COMERCIALIZABLE	COMERCIAL	6884500.0	LOCAL

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

```
45233789820.0
```

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

```
4650000.0
```

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 CHIPICHAPE	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	7.350000e+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()
```

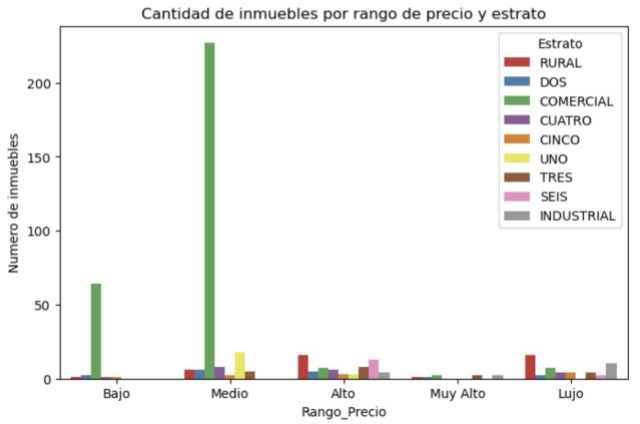

Figura 18. Grafico de relación de rango de precios con respecto al numero de inmuebles con respecto a su categoría

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

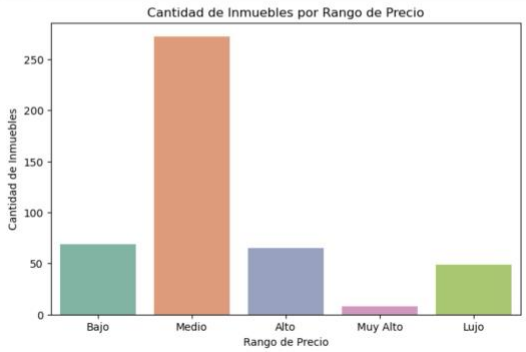

Figura 19. Distribución de los datos con respecto a la cantidad de inmuebles, con respecto a los categorías creadas con respecto los precios

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

plt.figure(figsize=(7, 7))
plt.pie(conteo_rangos, labels=conteo_rangos.index, autopct='%.1f%%', startangle=90, colors=colores)
plt.title("Distribución de Precios por Rangos")
plt.axis('equal')
plt.show()
```

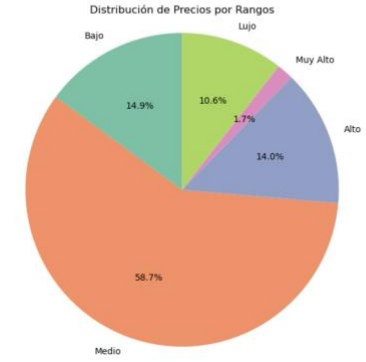

Figura 20. Distribución porcentual de las categorías de los rangos de precios

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.

Análisis y resultados

Mediante el tratamiento de los datos, realizados, se consideró que los datos Barrio y Datos adicionales, no representaban una utilidad relevante para el análisis de este proyecto. Por tanto, se consideran datos que son prescindibles para el análisis, por ello se pueden eliminar. Por otro lado, los datos que informa el código de la base de datos, también no presentan alguna información útil o relevante; por tanto, también se podrían considerar como datos prescindibles para el estudio y se podrían eliminar.

Realizando la primera gráfica (figura 10) para encontrar la distribución de estratos, se encuentra que la mayoría de datos están en la sección comercial, presentado una dominancia de los datos en general. Por tal motivo, es coherente mencionar que el estrato comercial representa un punto clave para realizar comparaciones, con otro tipo de datos y encontrar relaciones, tales como los precios o el área construida.

Siguiendo la misma idea, se ve la distribución de los datos numéricos de la base de datos. Como se puede ver en la figura 11 se encuentra que existe una gran cantidad de parámetros, por lo que no se puede ver una distribución clara de los datos, por lo que es evidente que no se puede encontrar un análisis importante, no obstante, se encontró que realizando los cálculos estadísticos, se pueden llegar a análisis interesantes. Primero como se mencionó en la figura 8 se encuentra que el valor mínimo de los precios esta alrededor de los 667 millones, pero si se observa la desviación estándar, se empieza a detectar valores de los miles de millones, de esta manera explicar por qué la distribución que se observa en la figura 11 se representa y podría explicar dicho comportamiento. Ahora bien, realizando la misma comparativa para el

área de terrero y construcción, se encuentra que, en efecto, existe una media baja, que al ver su distribución se dispara a valores más altos, lo que indica el comportamiento que se observa en la figura 11.

Como se presentan problemas para encontrar información relevante para los valores numéricos, se procede a encontrar una correlación numérica entre dichos parámetros, lo cual se puede ver en la figura 12. Como se observa, las secciones azules que representan los parámetros del código, representan valores de 0.26, lo cual confirma que no se pueden encontrar relaciones importantes y, por ende, se puede despreciar dichos datos. Ahora bien, observando los datos de precio y área construida, se encuentra una clara relación con una puntuación del 0.68, lo cual indica excelentes relaciones. Ahora bien, si se quisiera contrastar con otros parámetros, se podría, realizar un tratamiento de datos más avanzado, para transformar las variables categoricas a variables numéricas, y de esta manera encontrar relaciones con el área y el precio.

Ya que se empezó hablar sobre las variables categóricas, como se puede observar en la figura 13, en donde se ve una relación entre el rango de precios de 30 millones a 700 millones, intentando observar el comportamiento de los datos a partir de la información de la media, se encuentra una clara dominancia en precios para los estratos comercial, rural y estrato tres, el resto de precios se distribuyen para los demás estratos, aunque se observa que el estrato seis y el industrial una clara disminución, entre los rangos. Esto puede implicar que existan rangos de precios más altos para dichos estratos o precios más bajos. No obstante, para la inversión de los estratos comercial, rural y tres, representan, una excelente opción, además de que para el estrato comercial representa una amplia gama de precios disponibles, ya que representa en su mayoría de datos recopilados como se mencionó con la figura 10.

Ahora bien, en la figura 14 se observa una relación de las ciudades que presentan el rango de precios de 30 millones a 700 millones, en donde se encuentran que para las ciudades principales hay una clara ocupación de precios, siendo los más relevantes Bogotá, Taraza, Cartagena y Girardot. Seguido de otras ciudades tales como Puerto Boyacá, Montería y Pereira. Ahora bien, si se quisiera encontrar cómo están distribuidos los estratos para cada ciudad, se podrían realizar diferentes gráficos de tortas, para ver qué ciudad contiene los estratos comercial, rural y tres, los cuales son los estratos, que tienen un gran rango de precios, para poder invertir. Claro está que si se desea también realizar una investigación sobre inmuebles con mayor valoración con respecto a los estratos, se podría considerar dicha posibilidad.

Por último, los datos ordenados y agrupados para identificar mejor el rango de precios específicos, con etiquetas de bajo, medio, alto, muy alto y de lujo, se pueden observar en las figuras 18, 19 y 20. Los resultados gráficos mostrados indican que para el rango de precios bajo, alto y lujo hay una distribución pareja, pero para precios altos, hay pocos datos. Ahora bien, es evidente que el rango de precios medio es el más dominante, el cual esta acotado entre los 10 millones a los 100 millones, como se puede observar en la figura 16. Por lo que si se observa en la figura 18, en donde se hace la relación entre los rangos, el número de inmuebles y los estratos, se encuentra que el estrato comercial es el más predominante. Ya para los demás precios se encuentra que para el rango de precios altos, existe una mejor competencia de precios distribuidas entre los estratos, al igual que se puede observar para el rango de lujo. Ahora bien, se plantea la siguiente pregunta para futuros estudios: ¿Qué tipo de estrato se quiere identificar dependiendo de la ciudad? Ya que estos resultados abren la posibilidad de observar el comportamiento de los precios, el área de terreno o construcción con la distribución de los datos.

Conclusiones

A través del análisis exploratorio realizado, se evidenció la importancia de depurar la base de datos, descartando columnas como *Barrio*, *Datos adicionales* y *Código*, que no aportaban valor significativo al estudio. La visualización de la distribución de estratos permitió identificar al estrato *Comercial* como predominante, lo cual lo posiciona como un foco clave para establecer comparaciones con variables numéricas como *Precio* y *Área*

construida. Asimismo, se observó que los datos de variables numéricas del *Precio* presentan distribuciones con valores atípicos en los estratos lo que indica que hay varios precios un alto valor en ciertos estratos, en especial en el comercial, no obstante indica un riesgo para la compañía si se de invertir en dichos bienes. De modo que fue confirmado por las desviaciones estándar elevadas. Al calcular la correlación entre variables numéricas, se identificó una relación fuerte entre *Precio* y *Área construida* (0.68), lo que sugiere que esta última puede ser un buen predictor del valor de los inmuebles. Por otro lado, al categorizar los precios en rangos (bajo, medio, alto, etc.), se concluyó que el rango *Medio* (10 a 100 millones) domina la base de datos, principalmente en los estratos *Comercial*, *Rural* y *Tres*, lo que representa oportunidades potenciales de inversión. Finalmente, se destaca que ciudades como Bogotá, Cartagena y Girardot concentran una mayor proporción de inmuebles en estos rangos, lo cual abre nuevas posibilidades para enfocarse en nuevos datos específicos para identificar potenciales oportunidades para la inversión en dichas ciudades en los estratos previamente mencionados, realizando estudios como varían los estratos y precios según la ubicación geográfica. Para esto ser requerida una segunda perspectiva para observar dichos comportamientos.

Bibliografía

- Central de Inversiones S. A. (2017). Inmuebles Disponibles Para La Venta [Data set]. Tomado de https://www.datos.gov.co/Hacienda-y-Cr-dito-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.