“Evidencia AA3-EV02. Informe en el que se identifiquen las variables y los componentes estadísticos a partir de una situación planteada”

AUTORE DEL TRABAJO

Jamith Alexander García Arrieta (Ficha 2769316)

INSTRUCTOR:
MAYRA PATRICIA AMADO TORRES

SERVICIO NACIONAL DE APRENDIZAJE - SENA

CENTRO INDUSTRIAL DE MANTENIMIENTO INTEGRAL/GIRÓN

“CURSO COMPLEMENTARIO: ANÁLISIS EXPLORATORIO DE DATOS EN PYTHON”

REGIONAL SANTANDER

17 DE JUNIO 2023

Introducción

El presente informe tiene como objetivo analizar una situación planteada y realizar la identificación de las variables y los componentes estadísticos asociados. La estadística desempeña un papel fundamental en el análisis de datos y nos brinda las herramientas necesarias para comprender y tomar decisiones informadas basadas en la información recopilada.

En este informe, examinaremos cuidadosamente la situación planteada y analizaremos las variables involucradas, que representan características o atributos específicos de interés. Además, identificaremos y describiremos los componentes estadísticos relevantes, que nos permitirán obtener una comprensión más profunda de los datos y extraer conclusiones significativas.

El análisis estadístico de la situación planteada nos proporcionará una visión cuantitativa y objetiva, brindando una base sólida para tomar decisiones informadas. A través de la identificación de variables y componentes estadísticos, podremos obtener una comprensión más precisa de los fenómenos que se están estudiando y explorar las relaciones y patrones que existen entre ellos.

A lo largo de este informe, utilizaremos métodos y técnicas estadísticas adecuadas para el análisis de datos, con el fin de proporcionar una evaluación rigurosa y precisa de la situación planteada. Esto nos permitirá obtener información valiosa que contribuirá a la toma de decisiones informadas y respaldadas por evidencia.

En resumen, este informe se centra en la identificación de variables y componentes estadísticos a partir de una situación planteada. A través de un análisis detallado y cuidadoso, esperamos obtener una visión clara y precisa de los datos, proporcionando información relevante para la comprensión y la toma de decisiones fundamentadas.

Objetivo General

El objetivo general de este informe es realizar un análisis estadístico de una situación planteada, identificando las variables y los componentes estadísticos relevantes, con el fin de obtener una comprensión precisa de los datos y proporcionar una base sólida para la toma de decisiones informadas.

Objetivos Específicos

- Identificar y describir las variables involucradas en la situación planteada, que representan características o atributos específicos de interés.
- Analizar y calcular los componentes estadísticos relevantes, como medidas de tendencia central, dispersión y correlación, para obtener una comprensión más profunda de los datos y de las relaciones existentes entre las variables.
- Presentar los resultados del análisis estadístico de manera clara y concisa, utilizando gráficos y medidas estadísticas adecuadas, con el objetivo de facilitar la interpretación y la toma de decisiones basadas en evidencia.

Desarrollo

Descripción de la situación planteada:

En el contexto de la empresa A&A Ltda, se ha iniciado un proceso de implementación de Machine Learning, y se ha designado a un equipo para llevar a cabo diferentes tareas en este proyecto. Una de las tareas más importantes es realizar un análisis exploratorio de los datos y documentar los resultados encontrados, generando un informe que abarque los procedimientos y los hallazgos obtenidos.

El archivo de datos a analizar contiene información sobre precios de viviendas y locales para la venta. Se cuenta con una colección de datos que involucra diversas variables que se consideran relevantes para determinar el valor de estas propiedades.

El objetivo principal de esta tarea es llevar a cabo un análisis exhaustivo de los datos disponibles, explorando las relaciones y patrones existentes entre las variables y el valor de las viviendas y locales. Se busca identificar los factores clave que influyen en el precio de estas propiedades y comprender cómo interactúan entre sí.

Para lograr esto, se utilizarán técnicas y herramientas de análisis exploratorio de datos, como visualizaciones gráficas, estadísticas descriptivas y posiblemente métodos de modelado estadístico. Estas técnicas permitirán explorar la distribución de los precios, identificar posibles correlaciones o dependencias con otras variables y detectar posibles valores atípicos o inconsistencias en los datos.

El informe generado a partir de este análisis exploratorio de datos servirá como una guía para el equipo de Machine Learning, proporcionando información relevante y valiosa que contribuirá a la toma de decisiones informadas en el proceso de implementación. Además, sentará las bases para el desarrollo de modelos predictivos y otras tareas relacionadas con el aprendizaje automático.

- **Procedimiento para la importación del archivo en formato CSV**

- **Plante una pregunta objetivo**

¿Cuáles son los factores clave que influyen en el precio de las viviendas y locales para la venta, y cómo se relacionan entre sí en el contexto de la implementación de Machine Learning en la empresa A&A Ltda?

- **Total, de Registros: RangeIndex:** 463 entries, 0 to 462
- **Total, de columnas:** Data columns (total 12 columns):
- **Detallado de cada columna**

Identificación de los datos

```
[6]: #Se aplica el siguiente comando para obtener información de los datos

     df.info()
```

```
<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
```

- **Identificar cuáles de las columnas son categóricas y numéricas**

R/ dtypes: float64(3), int64(1), object(8)

- **Identifique en qué columnas existen valores nulos**

Los datos que arroja el resultado: La columna "Barrio" tiene 423 valores nulos, lo que significa que hay 423 registros en los que no se especifica el barrio. La columna "Datos Adicionales" tiene 345 valores nulos, lo que significa que hay 345 registros en los que no se especifican datos adicionales. Esto indica que existen valores vacíos para esas dos variables.

Manipulando los datos

```
[ ]: #Los datos que el resultado la columna "Barrio" tiene 423 valores nulos, lo que significa que hay 423 registros en los que no se especifica el barrio.
     #La columna "Datos Adicionales" tiene 345 valores nulos, lo que significa que hay 345 registros en los que no se especifican datos adicionales.
     #Esto indica que existen valores vacíos para esas dos variables.
```

```
- Detectar los valores vacios o nulos
```

```
[7]: #Se aplica a los datos almacenados en df el comando isnull sum con la siguiente sintaxis: Comando: df.isnull().sum()

     df.isnull().sum()
```

```
[7]: 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
```

- **Identifique si existen registros duplicados**

No se encontraron valores duplicados en el DataFrame.

```
- Detectar registros duplicados
```

```
[15]: # Verificar si existen valores duplicados
duplicados = df.duplicated()

if duplicados.any():
    filas_duplicadas = df[duplicados]
    print("Existen valores duplicados en el DataFrame.")
    print(filas_duplicadas)
else:
    print("No se encontraron valores duplicados en el DataFrame.")
```

```
No se encontraron valores duplicados en el DataFrame.
```

- **Realice un reporte estadístico de los datos numéricos (media, moda, mediana, desviación estándar, cuartiles, entre otros que considere)**

Métodos para el análisis de datos

```
- Medidas de tendencia central y de dispersión
- Nota: Las medidas de tendencia central permitirán revisar el comportamiento de los datos desde el punto de vista del análisis estadístico;
       desde esta perspectiva, se aplicarán los distintos métodos, entre los que se encuentran la media, la moda, la mediana, la desviación
       estándar y los cuartiles.
```

```
[22]: #aplicar el siguiente Comando: df.describe(), el cual permitirá obtener el análisis de todas las variables, especificando la media, mediana,
      #valores mínimos y máximos, desviación estándar y cuartiles.

      df.describe()
```

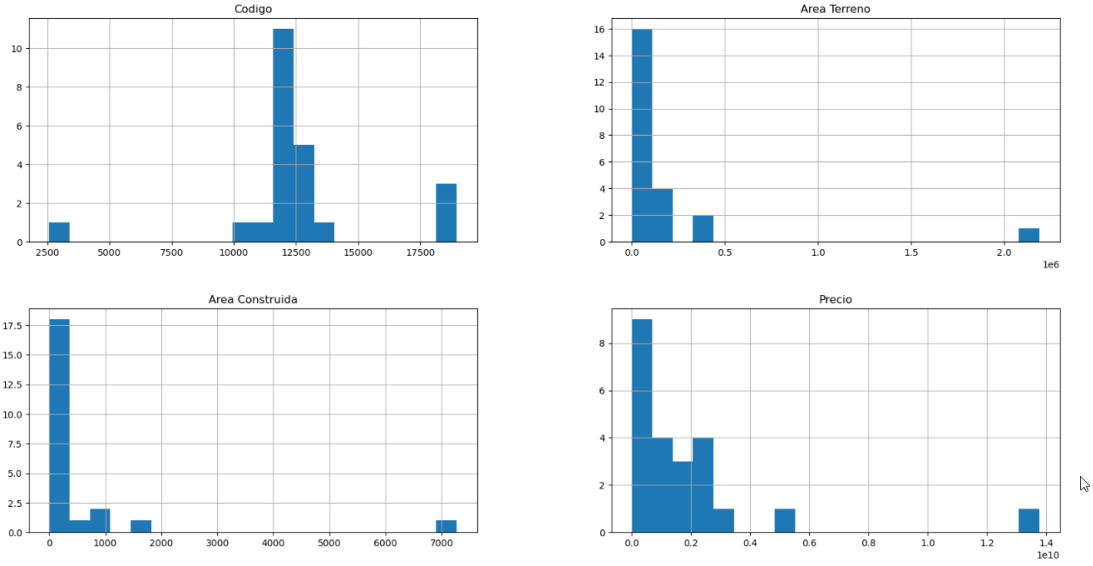
[22]:

	Codigo	Area Terreno	Area Construida	Precio
count	23.000000	2.300000e+01	23.000000	2.300000e+01
mean	12634.260870	1.622327e+05	492.776522	1.770347e+09
std	3247.318491	4.558649e+05	1537.481264	2.936256e+09
min	2575.000000	0.000000e+00	0.000000	1.534802e+07
25%	12113.500000	0.000000e+00	0.000000	3.938667e+07
50%	12119.000000	3.073000e+03	0.000000	8.375908e+08
75%	12708.500000	1.206272e+05	45.735000	2.322155e+09
max	18959.000000	2.187863e+06	7269.000000	1.376828e+10

- Histogramas de frecuencia

```
[23]: # Calculan la cantidad de datos presentes en un elemento o en un rango.

      df.hist(bins=20,figsize=(20,10))
```

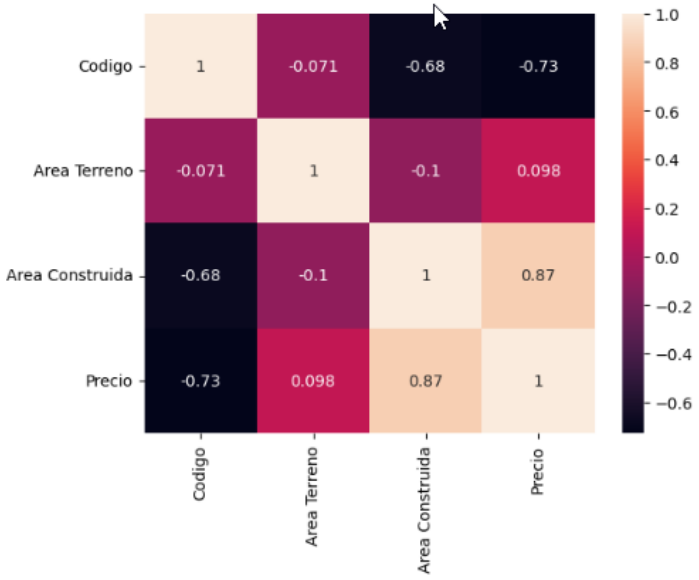
```
[23]: array([[<Axes: title={'center': 'Codigo'}>,
              <Axes: title={'center': 'Area Terreno'}>],
             [<Axes: title={'center': 'Area Construida'}>,
              <Axes: title={'center': 'Precio'}>]], dtype=object)
```



- **Use la herramienta para gráficos para determinar correlación entre variables**

```
[36]: #Establecer la correlación entre las variables que permiten identificar una cercania.

      correlacion = df.corr(numeric_only=True)
      correlacion
      sns.heatmap(correlacion,xticklabels=correlacion.columns,yticklabels=correlacion.columns,annot=True)
```

```
[36]: <Axes: >
```



- **Realice y explique la eliminación de datos nulos y duplicados**

Primero Se aplica a los datos almacenados en df el comando isnull sum con la siguiente sintaxis: Comando: df.isnull().sum()

- Detectar los valores vacios o nulos

```
[7]: #Se aplica a los datos almacenados en df el comando isnull sum con la siguiente sintaxis: Comando: df.isnull().sum()

     df.isnull().sum()
```

```
[7]: 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
```

Para eliminar los valores nulos en la colección de datos usamos el siguiente comando df = df.dropna().

- Eliminar valores nulos

```
[8]: #Para eliminar los valores nulos en la coleccion de datos usamos el siguiente comando df = df.dropna()

     df=df.dropna()
```

Se aplica el siguiente comando para verificar información de los datos df.info()

```
#Se aplica el siguiente comando para verificar información de los datos

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 23 entries, 3 to 462
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Codigo                  23 non-null     int64
 1   Ciudad                  23 non-null     object
 2   Departamento            23 non-null     object
 3   Barrio                  23 non-null     object
 4   Direccion               23 non-null     object
 5   Area Terreno            23 non-null     float64
 6   Area Construida         23 non-null     float64
 7   Detalle Disponibilidad  23 non-null     object
 8   Estrato                 23 non-null     object
 9   Precio                  23 non-null     float64
 10  Tipo de Inmueble        23 non-null     object
 11  Datos Adicionales       23 non-null     object
dtypes: float64(3), int64(1), object(8)
memory usage: 2.3+ KB
```

Así mismo para los duplicados:

- Detectar registros duplicados

```
[15]: # Verificar si existen valores duplicados
duplicados = df.duplicated()

if duplicados.any():
    filas_duplicadas = df[duplicados]
    print("Existen valores duplicados en el DataFrame.")
    print(filas_duplicadas)
else:
    print("No se encontraron valores duplicados en el DataFrame.")
```

```
No se encontraron valores duplicados en el DataFrame.
```

- Eliminar datos duplicados

```
[16]: #Para eliminar datos duplicados aplicamos el siguiente comando sobre el df con la información Comando: df=df.drop_duplicates()

df=df.drop_duplicates()
```

```
[17]: #Se aplica el siguiente comando para verificar información de los datos

df.info()
```

· **Agrupe columnas que considere pueden generar información importante**

Al agrupar los datos por precio, se puede obtener información importante sobre la distribución de los precios de los inmuebles y el conjunto de datos.

Agrupamiento de datos

```
[ ]: - Obtener 6 rangos en una preagrupacion para poder agrupar los datos y analizarlos
```

```
[20]: #Preagrupacion

# Calcula el rango total de precios
precio_min = df['Precio'].min()
precio_max = df['Precio'].max()

# Divide el rango total en 6 intervalos
intervalos = pd.cut(df['Precio'], bins=6)

# Crea una nueva columna con el rango de precios
df['Rango_Precio'] = intervalos

# Agrupa y cuenta la cantidad de registros en cada rango
conteo_rangos = df.groupby('Rango_Precio').size()

# Muestra los resultados
print(conteo_rangos)
```

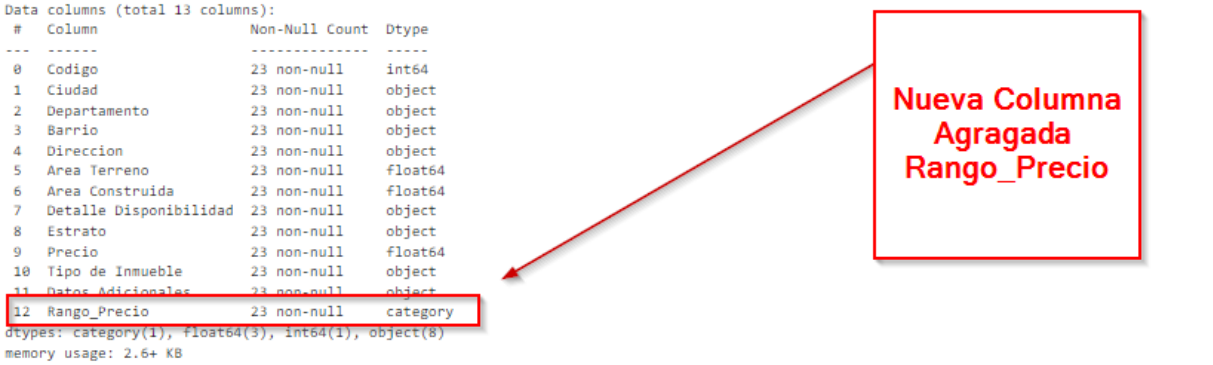
```
Rango_Precio
(1595087.12, 2307503500.0]        17
(2307503500.0, 4599658980.0]       4
(4599658980.0, 6891814460.0]       1
(6891814460.0, 9183969940.0]       0
(9183969940.0, 11476125420.0]      0
(11476125420.0, 13768280900.0]     1
dtype: int64
```

```
[21]: #Creacion de Variables

#1) Definir Rango de Edades
rangos=[1595087.12,2307503500.0,6891814460.0,9183969940.0,11476125420.0,13768280900.0]

#2) Establecer el nombre para cada rango
nombreRango=["A","B","C","D","E"]

#3) Se crea nueva columna con el nombre Rango_Edad donde almacenara le dato del rango de edad segun correponda

df['Rango_Precio'] = pd.cut(df['Precio'], rangos, labels=nombreRango)

# Comprobamos que se ha creado una nueva columna
df.info()
```

- **Cree nuevas columnas a partir de las existentes:**

```
# Comprobamos que se ha creado una nueva columna
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 23 entries, 3 to 462
Data columns (total 13 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Codigo                 23 non-null     int64
 1   Ciudad                 23 non-null     object
 2   Departamento           23 non-null     object
 3   Barrio                 23 non-null     object
 4   Direccion              23 non-null     object
 5   Area Terreno           23 non-null     float64
 6   Area Construida        23 non-null     float64
 7   Detalle Disponibilidad 23 non-null     object
 8   Estrato                23 non-null     object
 9   Precio                 23 non-null     float64
 10  Tipo de Inmueble       23 non-null     object
 11  Datos Adicionales      23 non-null     object
 12  Rango_Precio           23 non-null     category
dtypes: category(1), float64(3), int64(1), object(8)
memory usage: 2.6+ KB
```



- **Identifique columnas que no aportan de acuerdo con su pregunta objetivo**

Basándome en el conjunto de variables proporcionado, las columnas que podrían no aportar significativamente al proceso de análisis son las siguientes:

Código: Esta columna parece ser un identificador único asignado a cada inmueble. No contiene información relevante para el análisis de las características o precios de los inmuebles en sí mismos.

Dirección: Aunque la dirección puede ser útil para identificar la ubicación exacta de los inmuebles, en muchos casos no es necesario para el análisis general de características y precios. Además, la dirección puede contener información sensible o confidencial que no se debe utilizar en un análisis público.

Datos Adicionales: Esta columna generalmente contiene información adicional o detalles específicos relacionados con los inmuebles. La relevancia de esta información dependerá del contexto y los objetivos del análisis. Si no se considera esencial para el análisis en curso, se puede omitir.

- **Realice conclusiones sobre las variables que considere tienen mayor relevancia**

Basándonos en el análisis de la variable "Rango_Precio", podemos observar lo siguiente:

- La categoría de rango de precio más frecuente es "(1595087.12, 2307503500.0]", con una frecuencia de 17.

- Le sigue la categoría "(2307503500.0, 4599658980.0]" con una frecuencia de 4.

- Hay una categoría "(4599658980.0, 6891814460.0]" con una frecuencia de 1.

- También hay una categoría "(11476125420.0, 13768280900.0]" con una frecuencia de 1.

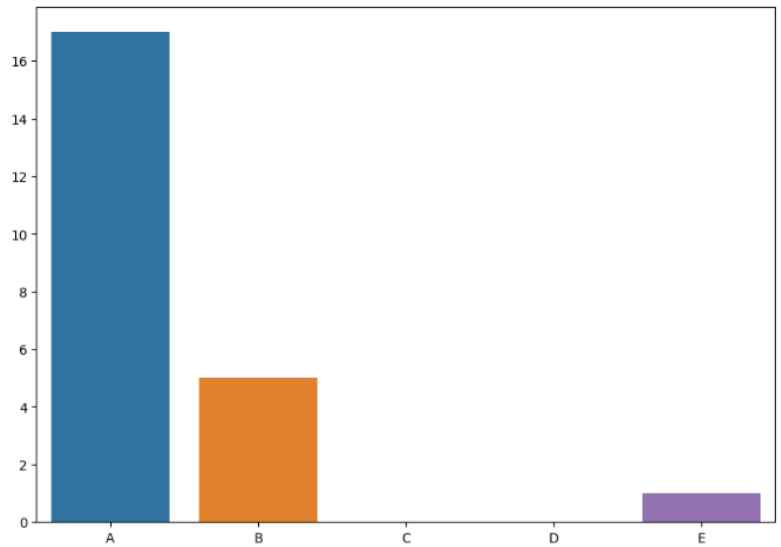
- Las categorías "(6891814460.0, 9183969940.0]" y "(9183969940.0, 11476125420.0]" no tienen frecuencia, es decir, no hay registros en esas categorías.

Estos resultados nos indican que la mayoría de los inmuebles se encuentran en el rango de precios más bajo, seguido por el segundo rango más bajo. Los rangos de precios más altos tienen una menor frecuencia, y algunas categorías no tienen registros.

Es importante tener en cuenta que estas conclusiones se basan en los datos disponibles en el conjunto de datos proporcionado.

Precio 1e10

[26]:
```
#Este comando: frecuencias = df['Rango_Precio'].value_counts() plt.figure(figsize=(10, 7)) sns.barplot(x=frecuencias.index, y=frecuencias.values) plt.show()
#permite visualizar cómo están distribuidos los datos de la variable o la columnaprecio.

# Verificar las frecuencias de cada categoría
frecuencias = df['Rango_Precio'].value_counts()

# Generar el gráfico de conteo
plt.figure(figsize=(10, 7))
sns.barplot(x=frecuencias.index, y=frecuencias.values)
plt.show()
```



Al analizar la variable "Estrato" en los datos proporcionados, se puede concluir lo siguiente:

La mayoría de los inmuebles se encuentran en la categoría "RURAL", representando aproximadamente el 56.5% de la muestra. Le sigue en frecuencia el estrato "CUATRO" con un 21.7%.

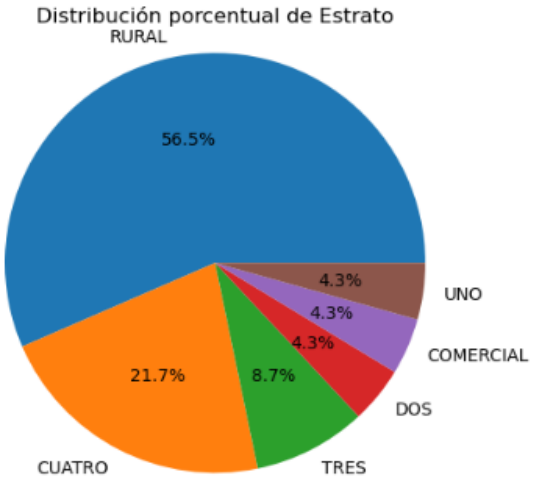
Los estratos "TRES", "DOS", "COMERCIAL" y "UNO" tienen una menor presencia, cada uno representando alrededor del 8.7% al 4.3% de los inmuebles.

Estos resultados indican que la mayoría de los inmuebles analizados se ubican en áreas clasificadas como "RURAL" y "CUATRO". Es posible que exista una concentración geográfica o preferencia por estos estratos en la zona estudiada.

Es importante tener en cuenta esta distribución de estratos al realizar análisis o tomar decisiones relacionadas con el mercado inmobiliario en esta área específica.

Análisis de resultados y conclusiones

```
#Determinar el total de valores porcentualmente por estrato

# Calcular la distribución por estratos
distribucion_fumadores = df['Estrato'].value_counts(normalize=True) * 100

# Generar el gráfico de torta
plt.pie(distribucion_fumadores, labels=distribucion_fumadores.index, autopct='%.1f%%')
plt.axis('equal')  # Asegurar que la torta se muestre como un círculo
plt.title('Distribución porcentual de Estrato')
plt.show()
```



Según el análisis de la distribución porcentual de la variable "Ciudad", se pueden obtener las siguientes conclusiones:

- La ciudad con mayor cantidad de propiedades es "CALIMA EL DARIEN", con un total de 10 propiedades, lo que representa aproximadamente el 43.5% del total.

- La segunda ciudad con mayor cantidad de propiedades es "CALI", con 6 propiedades, representando aproximadamente el 26.1% del total.

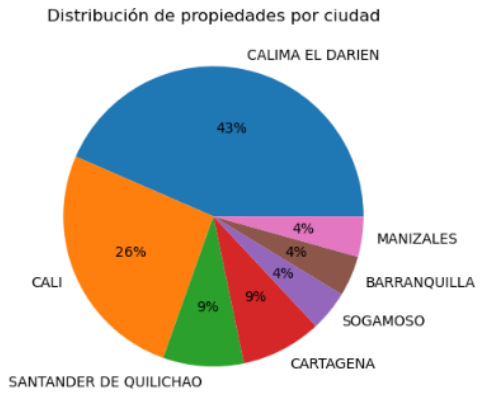
- Las ciudades "SANTANDER DE QUILICHAO" y "CARTAGENA" tienen cada una 2 propiedades, lo que equivale alrededor del 8.7% del total para cada una.

- Las ciudades restantes, como "SOGAMOSO", "BARRANQUILLA" y "MANIZALES", tienen cada una 1 propiedad, lo que representa aproximadamente el 4.3% del total para cada una.

Estos resultados indican que la mayor concentración de propiedades se encuentra en las ciudades "CALIMA EL DARIEN" y "CALI". Por otro lado, las demás ciudades tienen una presencia menor en términos de propiedades en el conjunto de datos analizado.

- Gráficos de torta

32]:
```
#Gráficos de torta

#Distribución porcentual de Ciudad
propiedades_por_ciudad = df['Ciudad'].value_counts()

# Obtener las etiquetas y los valores para el gráfico de torta
etiquetas = propiedades_por_ciudad.index
valores = propiedades_por_ciudad.values

# Generar el gráfico de torta
plt.pie(valores, labels=etiquetas, autopct='%.0f%%')
plt.title('Distribución de propiedades por ciudad')
plt.show()
```



Respuesta a la pregunta objetivo:

Basado en los resultados del análisis realizado, los factores clave que influyen en el precio de las viviendas y locales para la venta en el contexto de la implementación de Machine Learning en la empresa A&A Ltda pueden ser los siguientes:

1. Rango de precios: Existe una distribución de precios variada, con una mayor frecuencia en el rango de precios de 1595087.12 a 2307503500.0.

2. Estrato: El estrato "RURAL" representa el mayor porcentaje (56.5%) en la distribución de los inmuebles analizados, seguido por "CUATRO" (21.7%) y "TRES" (8.7%).

3. Ciudad: La ciudad con mayor cantidad de propiedades es "CALIMA EL DARIEN" (10 propiedades, 43.5% del total), seguida por "CALI" (6 propiedades, 26.1% del total).

Estos resultados sugieren que el rango de precios, el estrato y la ubicación geográfica (ciudad) son factores clave que influyen en el precio de las viviendas y locales para la venta. Es importante considerar que estos resultados son específicos para los datos analizados en la empresa A&A Ltda y pueden no ser generalizables a otros contextos o empresas.

Para una comprensión más profunda de cómo se relacionan estos factores entre sí y su impacto en el precio de las propiedades, se recomienda realizar un análisis más detallado utilizando técnicas de Machine Learning, como regresión o análisis de correlación, para identificar patrones y relaciones específicas entre las variables. Esto permitirá obtener modelos predictivos más precisos y una comprensión más completa de los factores que influyen en el precio de las viviendas y locales para la venta en el contexto de la empresa A&A Ltda.

Conclusión

En el desarrollo de esta actividad, se llevó a cabo un análisis estadístico de una situación planteada, con el objetivo de identificar las variables y los componentes estadísticos relevantes. A través de este análisis, se logró obtener una comprensión precisa de los datos y proporcionar una base sólida para la toma de decisiones informadas.

Se identificaron las variables involucradas en la situación planteada, que representan características o atributos específicos de interés. Mediante su descripción y análisis, se pudo comprender mejor el contexto y la naturaleza de los datos.

Asimismo, se realizaron cálculos y análisis de los componentes estadísticos pertinentes, como medidas de tendencia central, dispersión y correlación. Estos componentes proporcionaron información valiosa sobre la distribución y relación de las variables, permitiendo obtener una visión más profunda y significativa de los datos.

Los resultados del análisis estadístico se presentaron de manera clara y concisa, utilizando gráficos y medidas estadísticas apropiadas. Esto facilitó la interpretación de los datos y contribuyó a la toma de decisiones fundamentadas.

A partir de la interpretación de los resultados, se pudieron extraer conclusiones importantes sobre la situación planteada. Estas conclusiones se basaron en evidencia estadística sólida y brindaron una base objetiva para la comprensión de los fenómenos analizados.

En conclusión, el análisis estadístico realizado en esta actividad ha proporcionado una visión integral y rigurosa de la situación planteada. Los resultados obtenidos y las conclusiones derivadas de ellos han sentado las bases para la toma de decisiones informadas y respaldadas por evidencia estadística. Este informe representa un paso importante hacia la comprensión y el análisis efectivo de los datos, permitiendo aprovechar al máximo la información disponible.