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

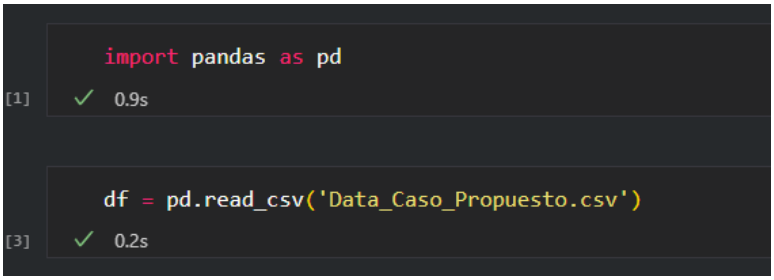
ANALISIS EXPLORATORIO DE DATOS EN PYTHON. (2675625)

Eduar Alejandro Cano Montoya.

Noviembre de 2022.

1. Procedimiento para la importación del archivo en formato CSV

Primero cargamos la librería panda y luego importamos los archivos .csv

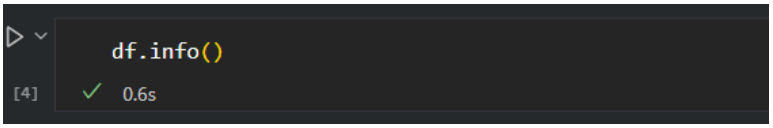


2. Plante una pregunta objetivo

Como pregunta para la base de datos se analizará que departamentos tienen los precios más altos y más bajos para vivienda y locales en el país. A medida que se desarrolle el ejercicio se podrá complementar la pregunta y respuestas

3. Análisis de la base de datos

- Total, de Registros
- Total, de columnas
- Detallado de cada columna
- Identificar cuáles de las columnas son categóricas y numéricas



Con el comando anterior podemos saber cuantos registros tiene la base de datos y responder a las preguntas planteadas. El resultado se muestra en la siguiente imagen

```
<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
```

El total de registros es de 463, el total de columnas es de 12, el número de columnas categóricas es 8 y el número de columnas numéricas es de 4.

- Identifique en que columnas existen valores nulos

  Para identificar los valores nulos ingresamos el siguiente código

```
df.isnull().sum()
```
```
[5]  ✓ 0.5s
```

Y el resultado nos muestra que el las columnas de barrio nos faltan 423 datos y en la de datos adicionales nos muestra que nos hace falta 345 datos como se muestra en la siguiente imagen

```
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
```

Como desde el inicio las columnas que vamos a evaluar son ciudad y precio las columnas barrio y datos adicionales no afectan a la obtención de los datos y por ende no se eliminaran de registro para obtener la mayor cantidad de datos posibles.

- Identifique si existen registros duplicados

  Para saber si existen registros duplicados es necesario escribir el siguiente código

```
[9]  df=df.drop_duplicates()
     ✓ 0.2s
```

  Después de ejecutar el comando podemos obtener la siguiente imagen

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 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: 47.0+ KB
```

  En donde podemos ver que no existen datos duplicados en la base de datos

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

  Para el reporte estadístico escribimos el siguiente comando

```
[11]  df.describe()
      ✓ 0.7s
```

  Lo cual nos arroja la siguiente tabla

	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

- Identifique columnas con valores erróneos
- Realice y explique la eliminación de datos nulos y duplicados

Como se puede observar en la siguiente tabla, el valor de mínimo y los cuartiles muestran valores errados y esto puede concluir que la obtención de los datos es errónea. Por esta razón se decide manipular de nuevo la base de datos y eliminar los valores que presenten inconvenientes. Para esto utilizamos el siguiente comando

```
[20]  df=df.dropna()
      ✓ ✓ 0.1s
```

Así ejecutando de nuevo los códigos podemos obtener los siguientes resultados

```
<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
```

	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

- Agrupe columnas que considere pueden generar información importante
- Cree nuevas columnas a partir de las existentes
- Identifique columnas que no aportan de acuerdo con su pregunta objetivo

Para esto ordenamos los datos de menor a mayor y de mayor a menor

```
precio=df.sort_values('Precio')
precio.head(5)
```

[28] ✓ 0.2s Python

	Codigo	Ciudad	Departamento	Barrio	Direccion	Area Terreno	Area Construida	Detalle Disponibilidad	Estrato	Precio	Tipo de Inmueble	Datos Adicionales
236	18871	SANTANDER DE QUILICHAO	CAUCA	VILLA DEL SUR	CL 9 AS # 10-49 LT 14 MZ G BARRIO VILLA DEL S...	0.0	0.0	COMERCIALIZABLE CON RESTRICCION	UNO	15348020.00	LOTE CON CONSTRUCCION	LOTE DE TERRENO CON CONSTRUCCION\n\nEL VALOR D...
462	12707	CALI	VALLE DEL CAUCA	PRADOS DEL NORTE	CL 30N # 2B-38 LOCAL 370 CENTRO COMERCIAL SAN...	0.0	0.0	COMERCIALIZABLE CON RESTRICCION	CUATRO	20608003.68	LOCAL	¡OPORTUNIDAD PARA INVERTIR! LOCAL EN PROINDIVI...
461	12706	CALI	VALLE DEL CAUCA	PRADOS DEL NORTE	CL 30N # 2B-38 LOCAL 367 CENTRO COMERCIAL SAN...	0.0	0.0	COMERCIALIZABLE CON RESTRICCION	CUATRO	20608003.68	LOCAL	¡OPORTUNIDAD PARA INVERTIR! LOCAL EN PROINDIVI...
460	12708	CALI	VALLE DEL CAUCA	PRADOS DEL NORTE	CL 30N # 2B-38 LOCAL 564 CENTRO COMERCIAL SAN...	0.0	0.0	COMERCIALIZABLE CON RESTRICCION	CUATRO	20701337.03	LOCAL	¡OPORTUNIDAD PARA INVERTIR! LOCAL EN PROINDIVI...
124	12732	CALI	VALLE DEL CAUCA	VEREDA EL SALADITO	LOTE 9C DIVISION REALIZADA A LA ANTIGUA HACIEN...	0.0	0.0	COMERCIALIZABLE	RURAL	29407403.52	LOTE MIXTO	INVIERTA EN ESTE TERRENO CAMPESTRE EN UNA ZONA...

```
precio=df.sort_values('Precio',ascending=False)
precio.head(5)
```

[29] ✓ 0.2s Python

	Codigo	Ciudad	Departamento	Barrio	Direccion	Area Terreno	Area Construida	Detalle Disponibilidad	Estrato	Precio	Tipo de Inmueble	Datos Adicionales
3	2575	SOGAMOSO	BOYACÁ	CENTRO	CRA. 10 #11-78/80 Ó CL 12 # 9 - 77/85 Ó CALLE...	1655.08	7269.00	COMERCIALIZABLE CON RESTRICCION	CUATRO	1.376828e+10	CLINICA	ESTE INMUEBLE SE COMERCIALIZARÁ A TRAVÉS DE SU...
37	12120	CALIMA EL DARIEN	VALLE DEL CAUCA	VEREDA PALERMO	LT 2	337995.00	0.00	COMERCIALIZABLE FIDUCIA	RURAL	5.171324e+09	LOTE VIVIENDA	EXCELENTE LOTE CON VISTA AL LAGO CALIMA, UBICA...
49	12117	CALIMA EL DARIEN	VALLE DEL CAUCA	VEREDA PALERMO	LT 3B2	134000.00	763.06	COMERCIALIZABLE FIDUCIA	RURAL	3.027405e+09	LOTE VIVIENDA	CISA VENDE DERECHOS FIDUCIARIOS DEL 9.48%\n\nF...
54	12114	CALIMA EL DARIEN	VALLE DEL CAUCA	VEREDA PALERMO	LT 1B	175982.00	0.00	COMERCIALIZABLE FIDUCIA	RURAL	2.704525e+09	LOTE VIVIENDA	CISA VENDE DERECHOS FIDUCIARIOS DEL 9.48%\n\nF...
55	12113	CALIMA EL DARIEN	VALLE DEL CAUCA	VEREDA PALERMO	LT 4	71143.00	745.33	COMERCIALIZABLE FIDUCIA	RURAL	2.700092e+09	LOTE VIVIENDA	CISA VENDE DERECHOS FIDUCIARIOS DEL 9.48%\n\nF...

Luego crearemos rangos de valores de las vivienda.creamos nuevas columnas

```
[34]
rangos=[10000000,200000000,500000000,1000000000,20000000000]
✓ 0.1s
```

```
[36]
nombreRango=["A","B","C","D"]
✓ 0.1s
```

```
[37]
df['Rango_Precio']=pd.cut(df['Precio'],rangos,labels=nombreRango)
✓ 0.2s
```

```
[38]
df.head()
✓ 0.2s
Python
```

	Codigo	Ciudad	Departamento	Barrio	Direccion	Area Terreno	Area Construida	Detalle Disponibilidad	Estrato	Precio	Tipo de Inmueble	Datos Adicionales	Rango_Precio
5	2575	SOGAMOSO	BOYACÁ	CENTRO	CRA. 10 #11-78/80 Ó CL 12 # 9 - 77/85 Ó CALLE...	1655.08	7269.00	COMERCIALIZABLE CON RESTRICCION	CUATRO	1.376828e+10	CLINICA	ESTE INMUEBLE SE COMERCIALIZARÁ A TRAVÉS DE SU...	D
7	12120	CALIMA EL DARIEN	VALLE DEL CAUCA	VEREDA PALERMO	LT 2	337995.00	0.00	COMERCIALIZABLE FIDUCIA	RURAL	5.171324e+09	LOTE VIVIENDA	EXCELENTE LOTE CON VISTA AL LAGO CALIMA, UBICA...	D
8	10106	BARRANQUILLA	ATLÁNTICO	CENTRO	CALLE 39 NO 43 58 62 LC 1	0.00	0.00	COMERCIALIZABLE CON RESTRICCION	DOS	7.303790e+08	LOCAL	DESEAS INVERTIR ESTA ES LA OPORTUNIDAD LOCAL E...	C
9	12117	CALIMA EL DARIEN	VALLE DEL CAUCA	VEREDA PALERMO	LT 3B2	134000.00	763.06	COMERCIALIZABLE FIDUCIA	RURAL	3.027405e+09	LOTE VIVIENDA	CISA VENDE DERECHOS FIDUCIARIOS DEL 9.48%\n\nF...	D
10	12114	CALIMA EL DARIEN	VALLE DEL CAUCA	VEREDA PALERMO	LT 1B	175982.00	0.00	COMERCIALIZABLE FIDUCIA	RURAL	2.704525e+09	LOTE VIVIENDA	CISA VENDE DERECHOS FIDUCIARIOS DEL 9.48%\n\nF...	D

- Utilice gráficos para identificar valores atípicos
- Realice histogramas de frecuencia

Use la herramienta para gráficos para determinar correlación entre variables

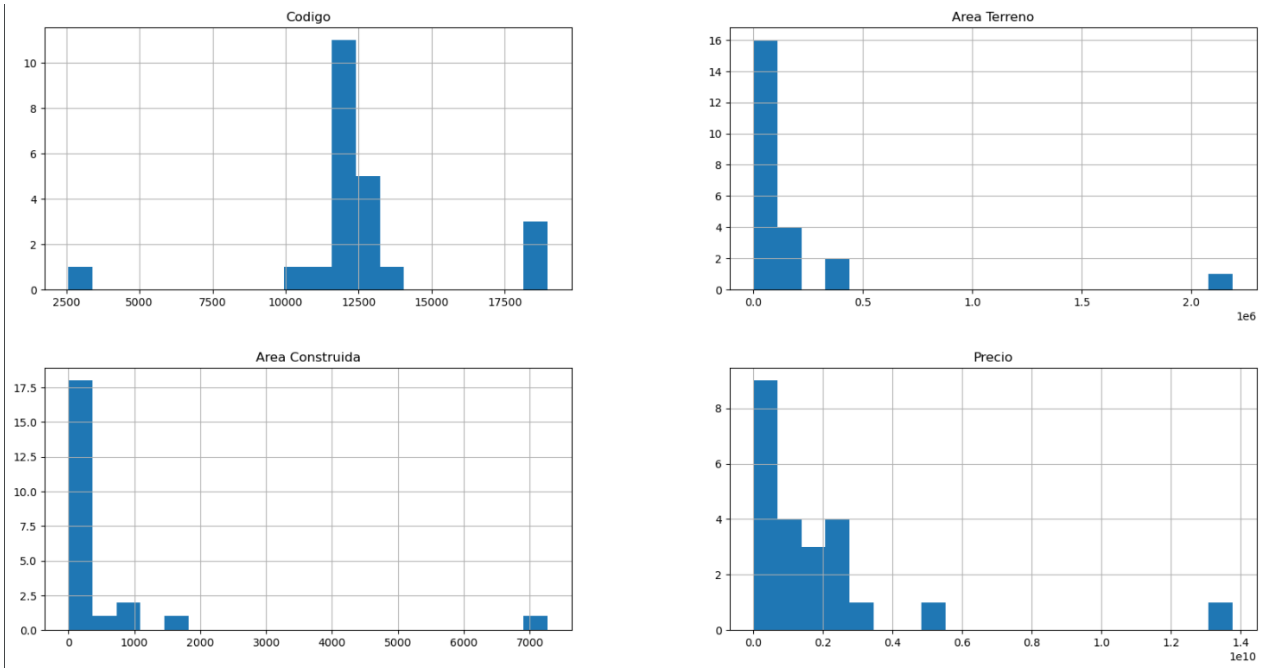
Para realizar los gráficos importamos los las librerías necesarias

```
[39]
import matplotlib.pyplot as plt
✓ 0.1s
```

```
[40]
import seaborn as sns
✓ 0.1s
```

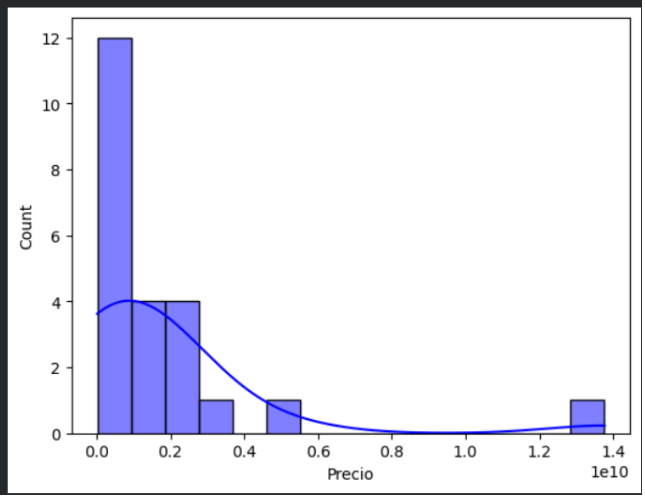
```
[41]
df.hist(bins=20,figsize=(20,10))
✓ 0.6s
```

Grafico de histogramas




```
sns.histplot(df.Precio,color="b",bins=15,kde=True)
plt.show()
```

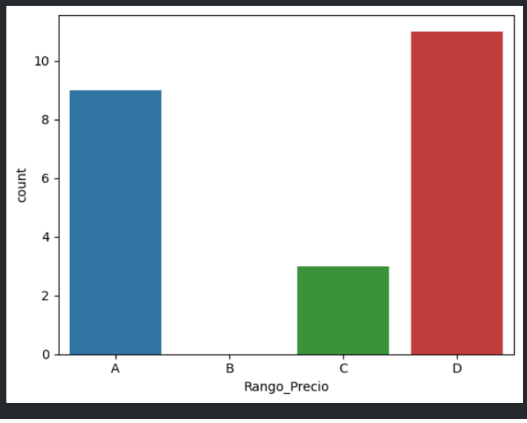
[48] ✓ 0.1s




```
plt. Figure(figsize=(5,2))
sns.countplot(df.Rango_Precio)
plt.show()
```

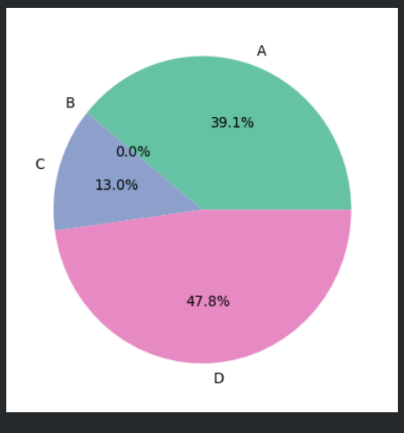
[113] ✓ 0.1s

```
c:\Users\Edward\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWa
only valid positional argument will be `data`, and passing other arguments with
  warnings.warn(
```



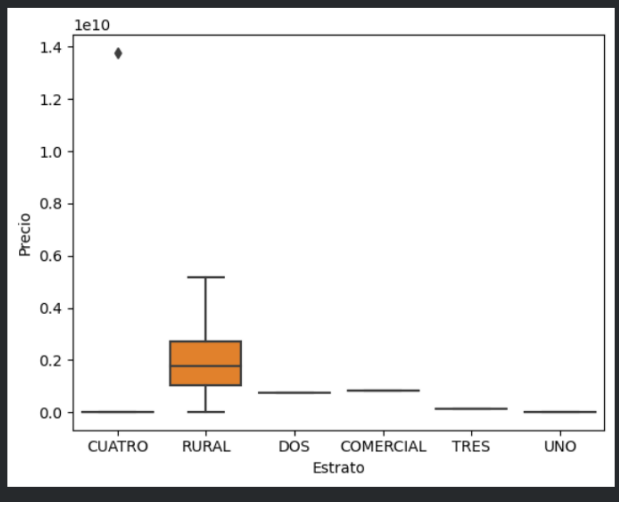
```
total_rango_edad =df['Rango_Precio'].groupby(df['Rango_Precio']).count()
labels = ["A", "B", "C", "D"]
colors = sns.color_palette('Set2')[0:5]
plt.pie(total_rango_edad, labels = labels, colors = colors, autopct='%.1f%%')
plt.show()
```

[114] ✓ 0.6s



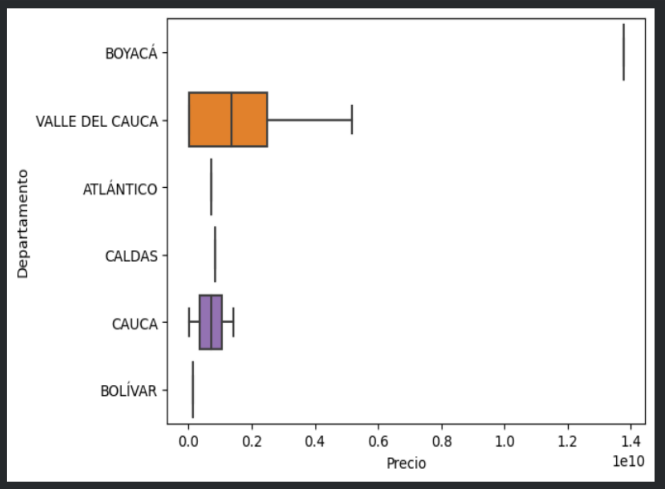

```
fumador_valor =sns.boxplot(x=df["Estrato"],
y=df["Precio"])
```

[116] ✓ 0.1s




```
Redad_valor =sns.boxplot(x=df["Precio"],
y=df["Departamento"])
```

[115] ✓ 0.1s



- Realice conclusiones sobre las variables que considere tienen mayor relevancia

  Podemos observar según los gráficos que la gran mayoría de precios de los inmuebles se encontraba en el rango D que se encuentra entre 1000 a 2000 millones con un 47.8% y en el rango A que se encuentra entre 10 y 200 millones 39.1% del total de los inmuebles. En el rango B no se tienen inmuebles que están ubicados en el rango de 201 millones a 500 millones.

  También podemos concluir que según la estratificación la gran mayoría de los datos se encuentran en el entorno rural y en estos se encuentran los precios mas elevados siguiendo por los inmuebles comerciales, luego de los estratos dos, tres, cuatro, tres y uno respectivamente.

La gran mayoría de datos de inmuebles se encuentran en el Valle del Cauca y Cauca

- Conclusiones generales

En un inicio la base de datos tenía muchos valores nulos, esto disminuía el total de datos que desearíamos para realizar el análisis. También los títulos contenían espacios lo cual es una mala practica en una base de datos pues no permite ejecutar el código para el análisis y por ultimo los datos no eran del todo confiables debido a los valores arrojados al analizarlos. Con solo 23 datos después de la depuración no es posible hacer un análisis confiable para dar respuestas a las preguntas surgidas. Para esto se recomendaría volver a tomar los datos, tratar de completar los registros nulos y volver a realizar el análisis de los datos para generar valores confiables para la solución de las preguntas propuestas.