

Evidencia AA3-EV02. Informe en el que se identifiquen las variables y los componentes

estadísticos a partir de una situación planteada.

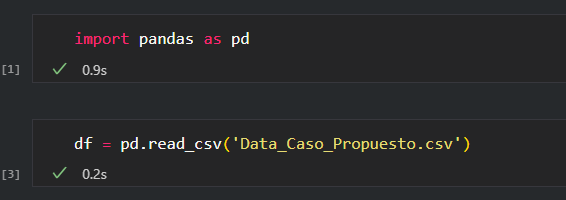
ANALISIS EXPLORATORIO DE DATOS EN PYTHON. (2675625)

Eduar Alejandro Cano Montoya.

Noviembre de 2022.

# Procedimiento para la importación del archivo en formato CSV

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

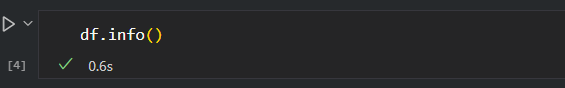


# Plante una pregunta objetivo

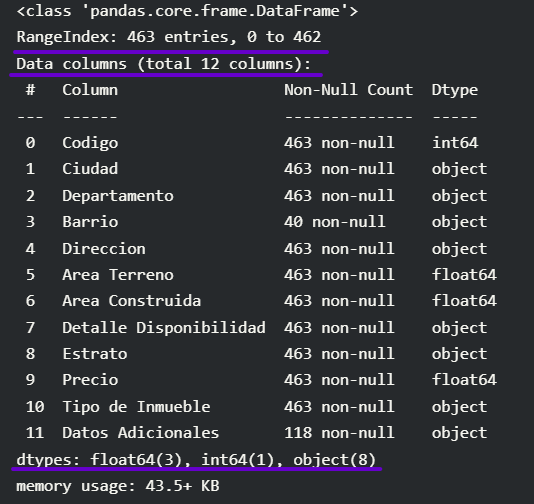
Como pregunta para la base de datos se analizará que departamentos tienen los precios más altos y más bajos para vivienda y locales en el país. A medida que se desarrolle el ejercicio se podrá complementar la pregunta y respuestas

# 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



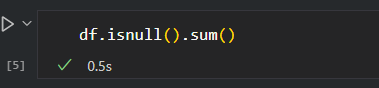
Con el comando anterior podemos saber cuantos registros tiene la base de datos y responder a las preguntas planteadas. El resultado se muestra en la siguiente imagen



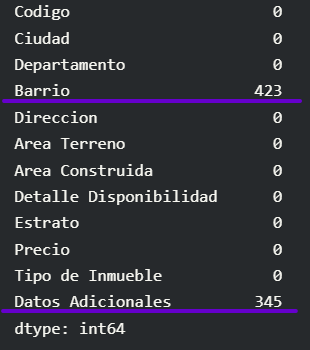
El total de registros es de 463, el total de columnas es de 12, el número de columnas categóricas es 8 y el número de columnas numéricas es de 4.

* Identifique en que columnas existen valores nulos

Para identificar los valores nulos ingresamos el siguiente código



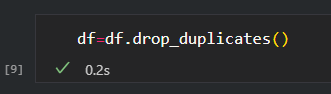
Y el resultado nos muestra que el las columnas de barrio nos faltan 423 datos y en la de datos adicionales nos muestra que nos hace falta 345 datos como se muestra en la siguiente imagen



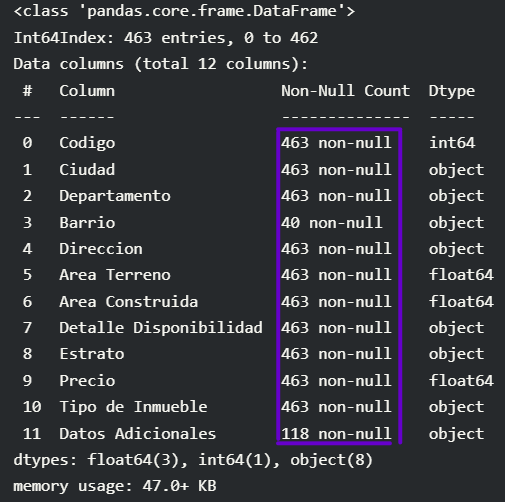
Como desde el inicio las columnas que vamos a evaluar son ciudad y precio las columnas barrio y datos adicionales no afectan a la obtención de los datos y por ende no se eliminaran de registro para obtener la mayor cantidad de datos posibles.

* Identifique si existen registros duplicados

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



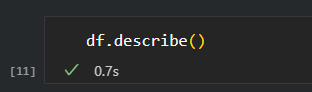
Después de ejecutar el comando podemos obtener la siguiente imagen



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

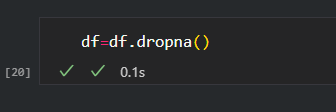


Lo cual nos arroja la siguiente tabla

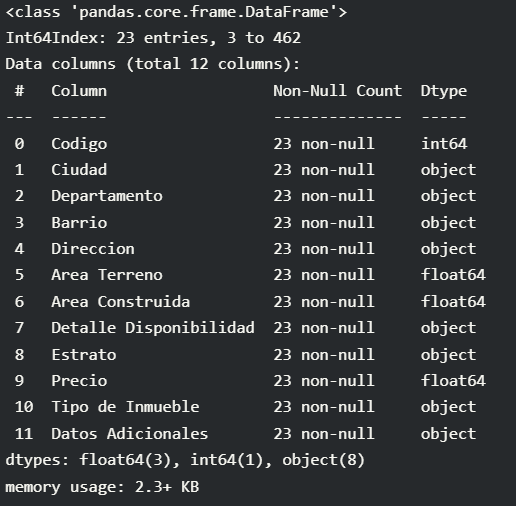


* Identifique columnas con valores erróneos
* Realice y explique la eliminación de datos nulos y duplicados

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



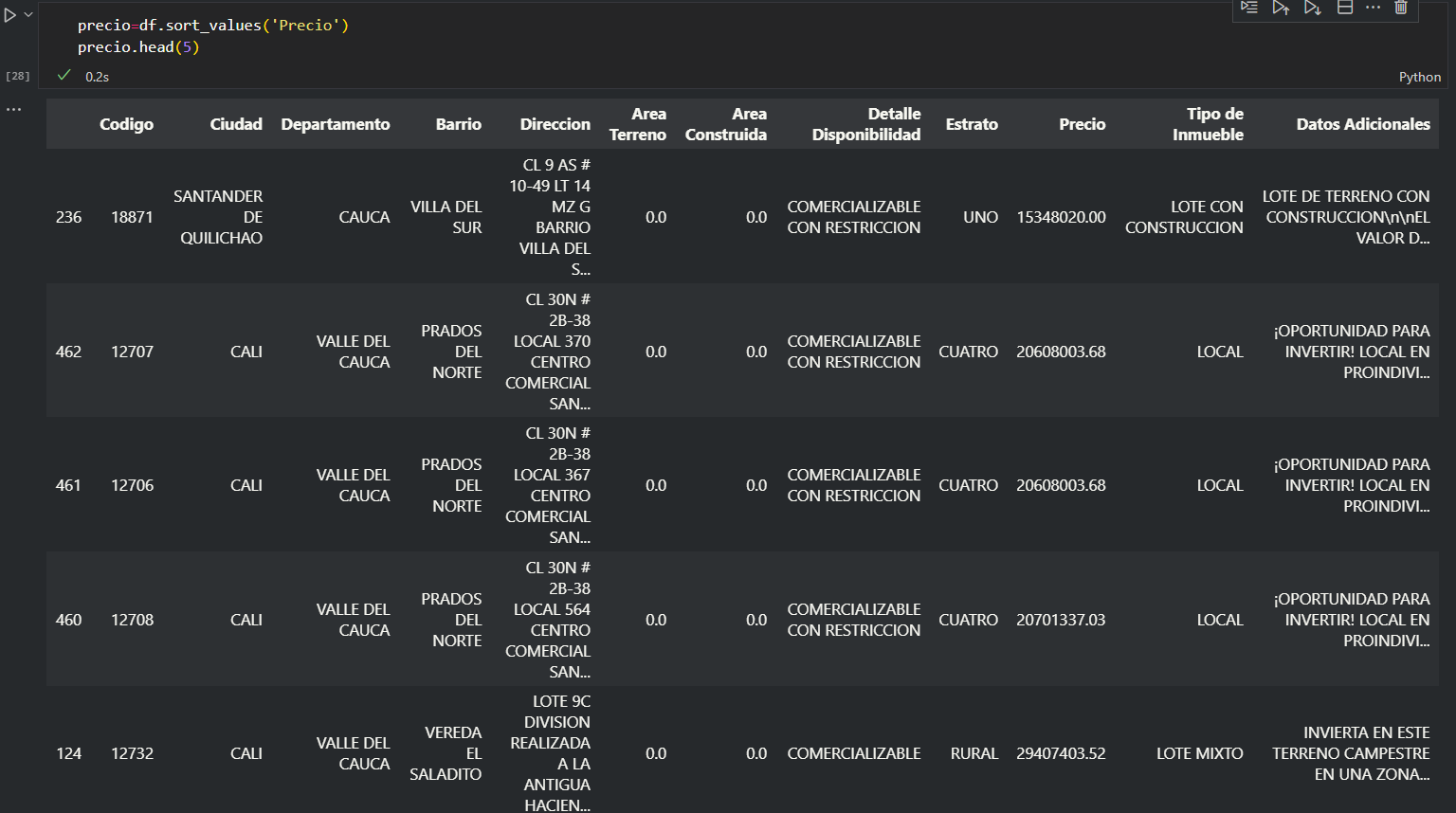
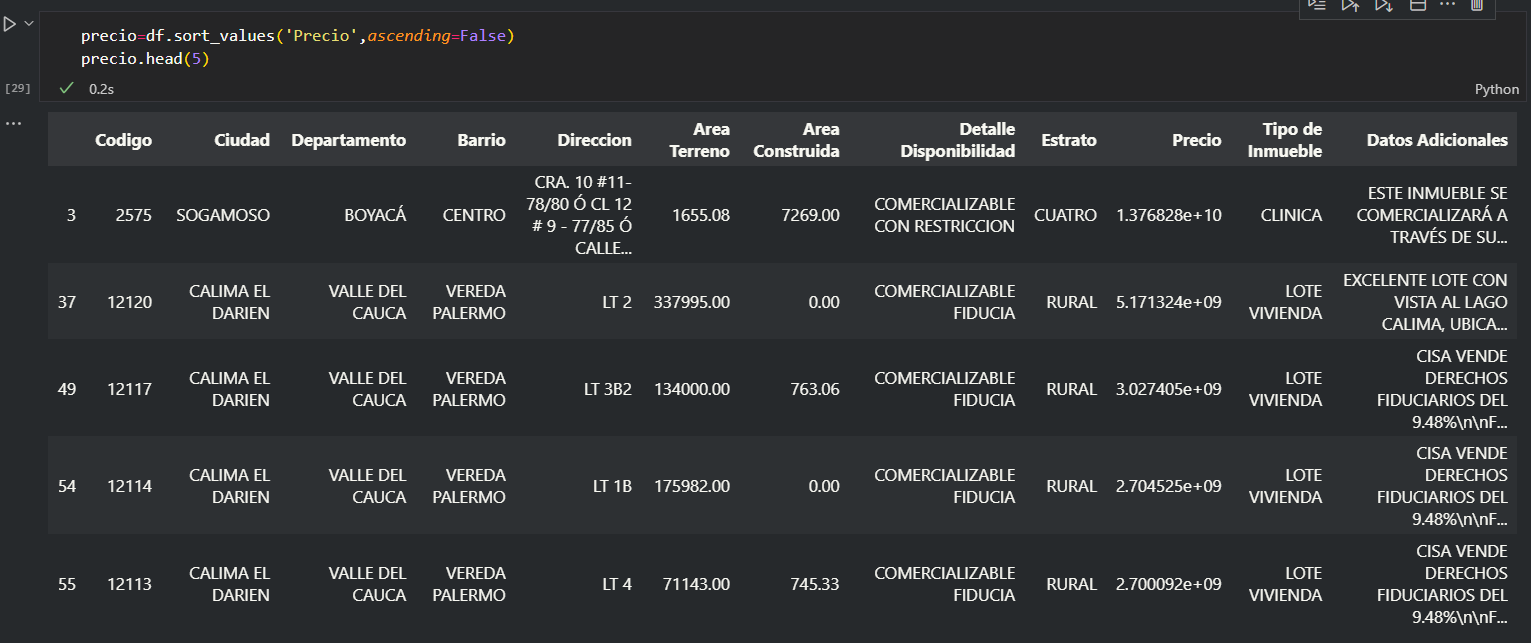
Así ejecutando de nuevo los códigos podemos obtener los siguientes resultados



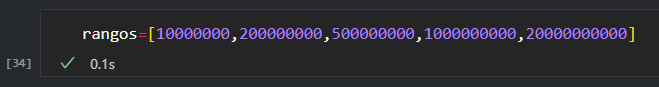


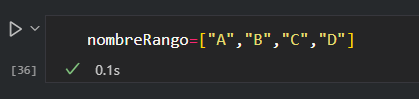
* Agrupe columnas que considere pueden generar información importante
* Cree nuevas columnas a partir de las existentes
* Identifique columnas que no aportan de acuerdo con su pregunta objetivo

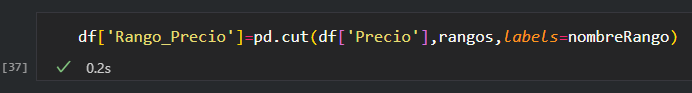
Para esto ordenamos los datos de menor a mayor y de mayor a menor

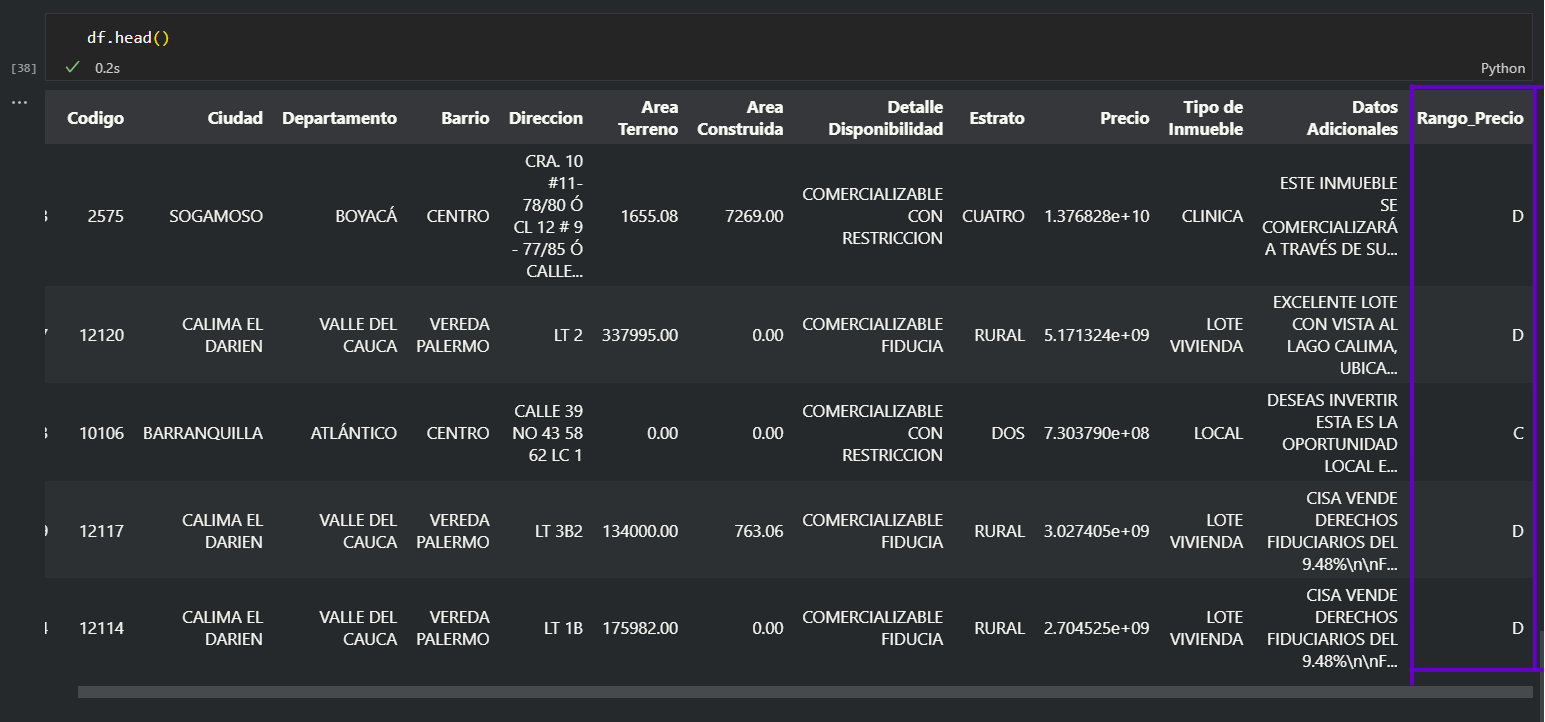


Luego crearemos rangos de valores de las vivienda.creamos nuevas columnas





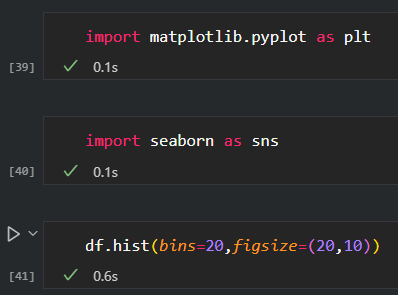


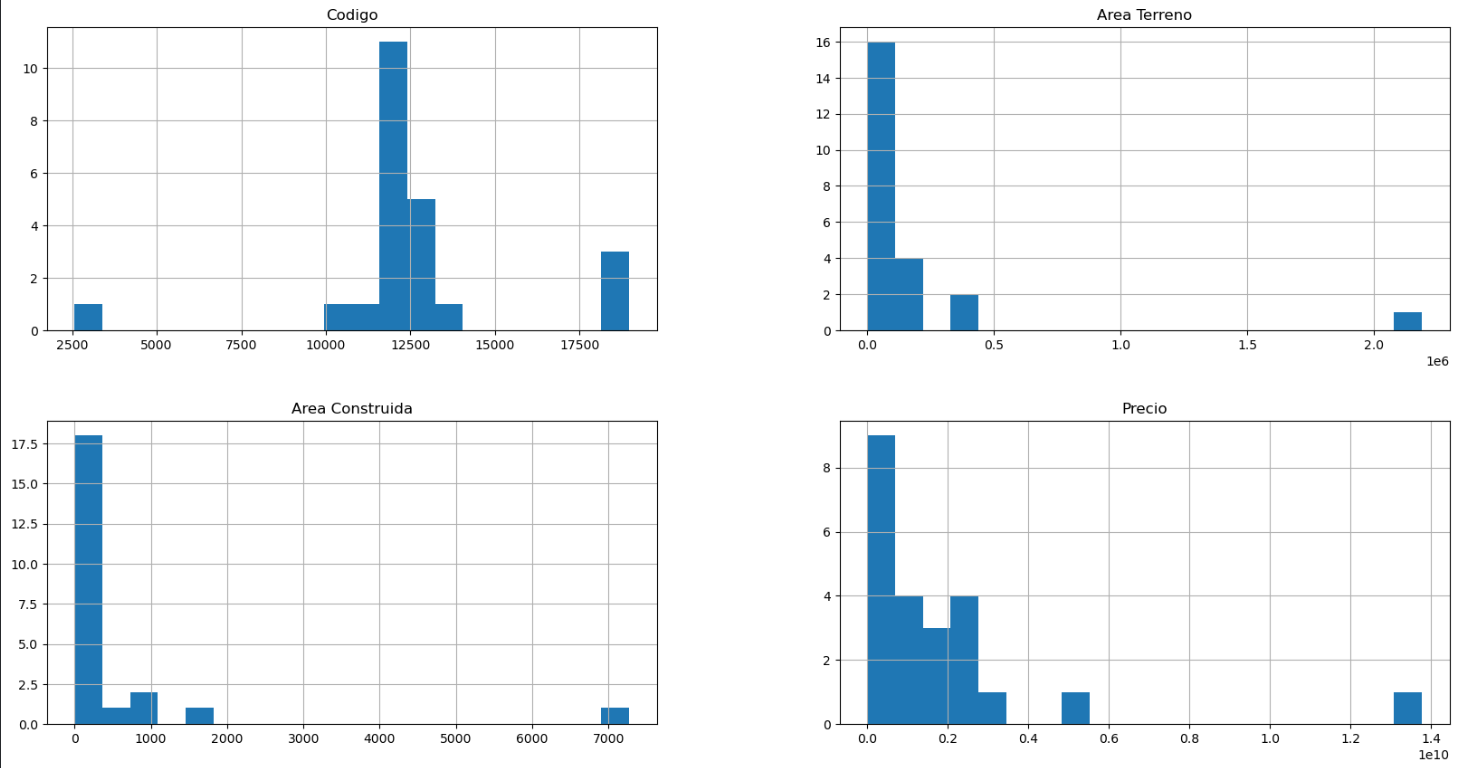
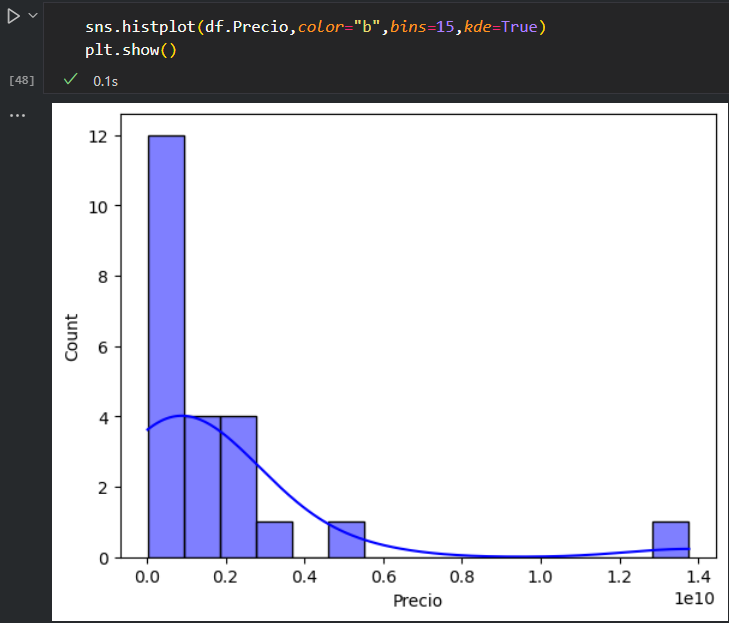
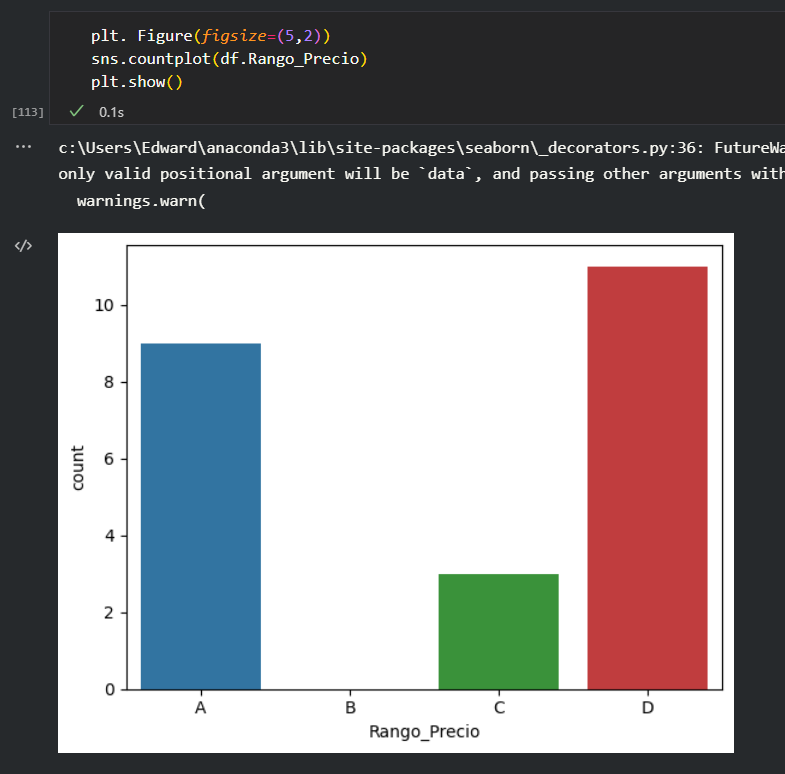


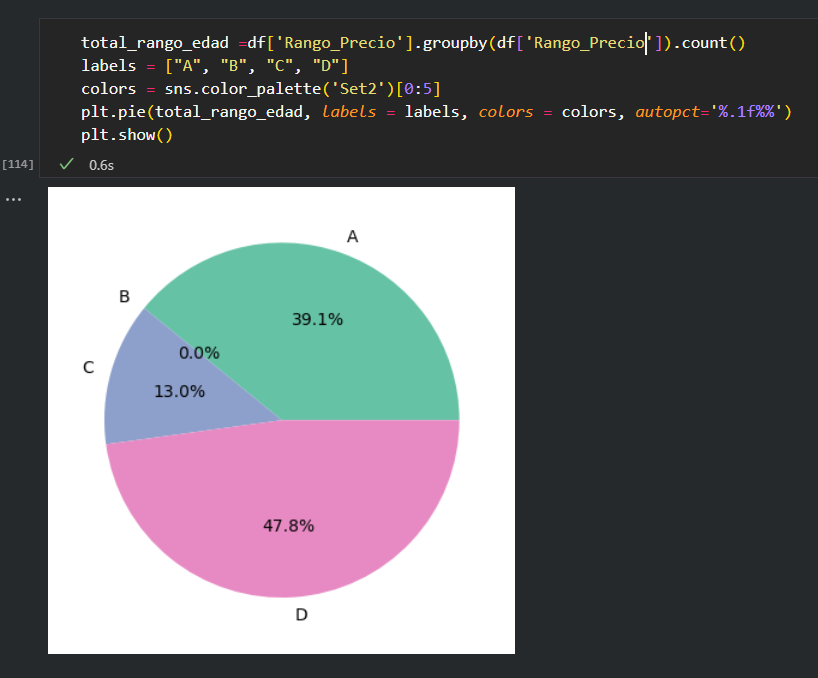
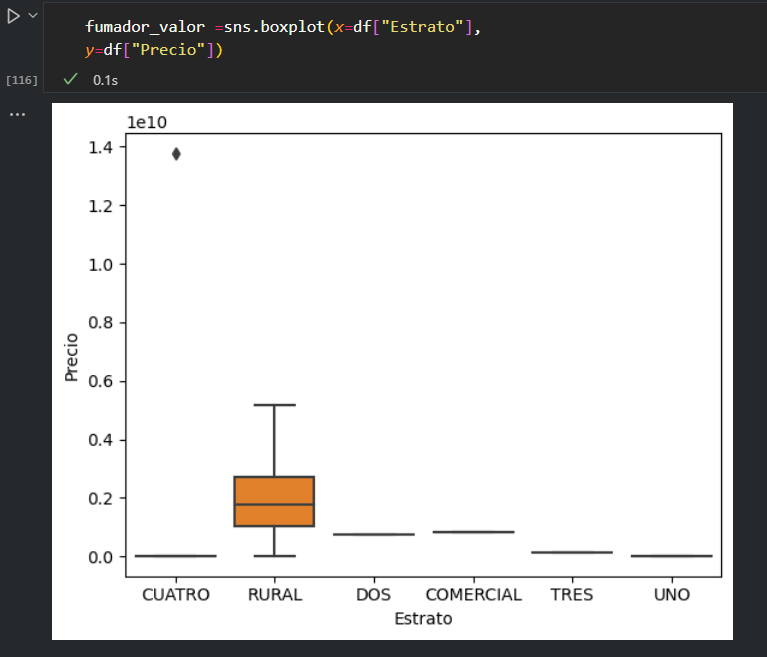
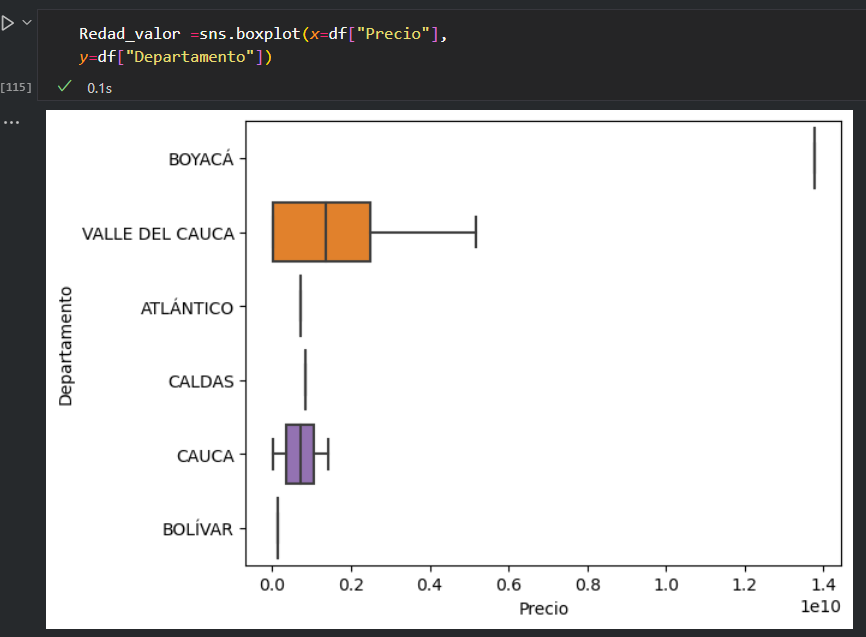
* Utilice gráficos para identificar valores atípicos
* Realice histogramas de frecuencia

Use la herramienta para gráficos para determinar correlación entre variables

Para realizar los gráficos importamos los las librerías necesarias



Grafico de histogramas



* 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.