Análisis Exploratorio de Datos en Python

Germán Daniel Colón Carrasquilla

Informe de resultados obtenidos del análisis de datos exploratorio realizado al caso de estudio. AA4-EV01

Introducción

El análisis de datos es una herramienta que puede fundamentalmente cambiar la perspectiva de la resolución de problemas basados en la toma de decisiones acertadas en conjunto con la apropiada investigación, análisis y visualización de los datos. Para llevar a cabo el manejo de grandes volúmenes de datos y mejorar el entendimiento de las diversas problemáticas se tiene como eje fundamental diferentes programas y lenguajes de programación que facilitan el manejo de estos datos. No obstante, el análisis debe estar siempre orientado a resolver una problemática y dirigido por un especialista para que el desarrollo del mismo no pierda significado, ya que al no estar dotado de un propósito o un fin estos se vuelven datos sin sentidos para las personas que los vayan a interpretar.

Datos

<https://www.datos.gov.co/Hacienda-y-Cr-dito-P-blico/Inmuebles-Disponibles-Para-La-Venta/72gd-px77/data>

Objetivo

* Realizar el análisis exploratorio según se indica en la guía del curso para el documento suministrado.
* Crear y responder una pregunta problema teniendo en cuenta los datos en el documento.
* Resaltar la importancia de las gráficas en los análisis de datos

Análisis Exploratorio

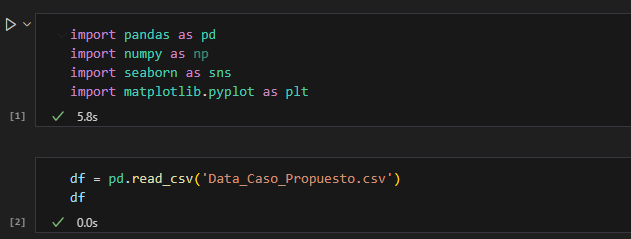
Pregunta problema

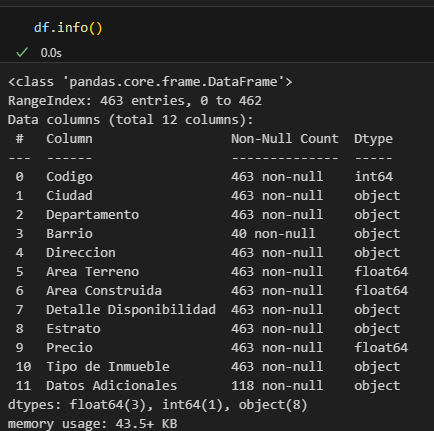
¿Variación del precio de los locales o casas según el estrato?

Esta pregunta está relacionada con la información directa del archivo, en ella se busca entender como los precios se determinan por estrato y que rige esta evaluación en cuanto a la determinación de precios.

Siguiendo con lo anterior se muestran los gráficos y las diferentes interpretaciones a cada uno de ellos para mejorar el entendimiento de la información en los datos.

* Importación de la librerías y lectura del archivo csv:



* Descripción de los datos y tipos de datos: 

El archivo está formado por datos del tipo numérico y de caracteres (letras), el numérico se divide en int (números enteros) y float (números separados por puntos)

* Datos nulos y duplicados:

Texto

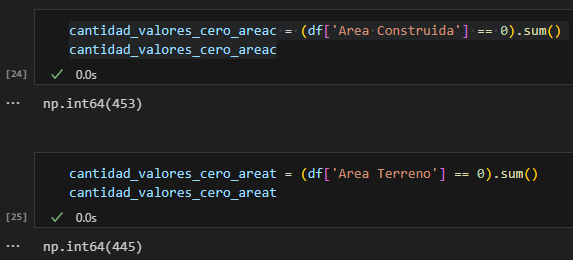
El contenido generado por IA puede ser incorrecto.Interfaz de usuario gráfica, Texto

El contenido generado por IA puede ser incorrecto.

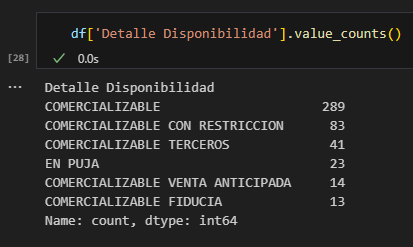
Los datos nulos conforman una cantidad importante, no se pueden descartar porque afectaran significativamente el análisis, cabe destacar que para el análisis las variables en las columnas ‘Barrio’ y ‘Datos adicionales’ no se van a utilizar para el análisis.

No hay valores duplicados entre los datos.

* Variables que se descartan:



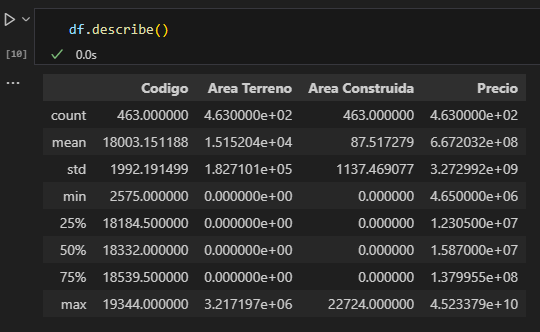
Estas variables, aunque pueden influir en responder la pregunta problema tienen muchos valores en cero por lo que hace imposible que se pueda hacer análisis con ellos.



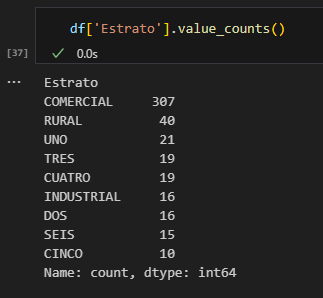
La variable ‘Detalle Disponibilidad’ se descarta porque se sobreentiende que las propiedades están a la venta, por ende, no es necesario ponerlo en el análisis.

Las demás variables diferentes a las mencionadas si se pueden usar para otros tipos de análisis.

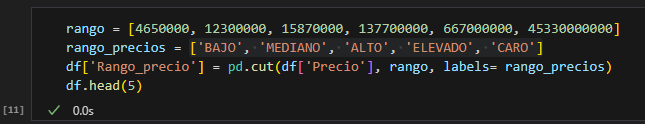
* Descripción de la tendencia de los datos:



* Cantidad de datos por estrato:



* Creación de nueva variable para responder la pregunta



Se crea la variable para determinar como el precio encasilla la propiedad en un término bajo, mediano, alto, elevado y caro dependiendo de su precio.

Los datos para los rangos de precios se tomaron teniendo en cuenta la tendencia de los datos en el apartado anterior.

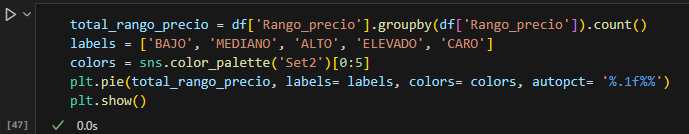
* Cantidad de propiedades por rango de precio:

Texto

El contenido generado por IA puede ser incorrecto.

Se muestran según el rango de precio en que categoría entran las propiedades.

* Gráfico de torta de las categorías:

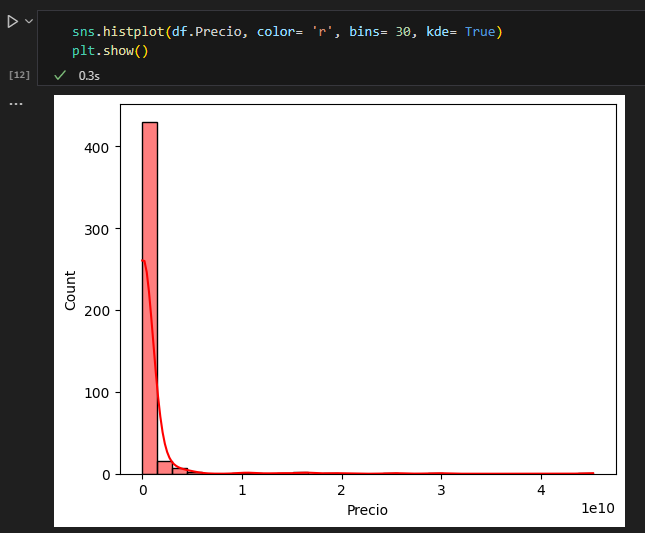


Gráfico, Gráfico circular

El contenido generado por IA puede ser incorrecto.

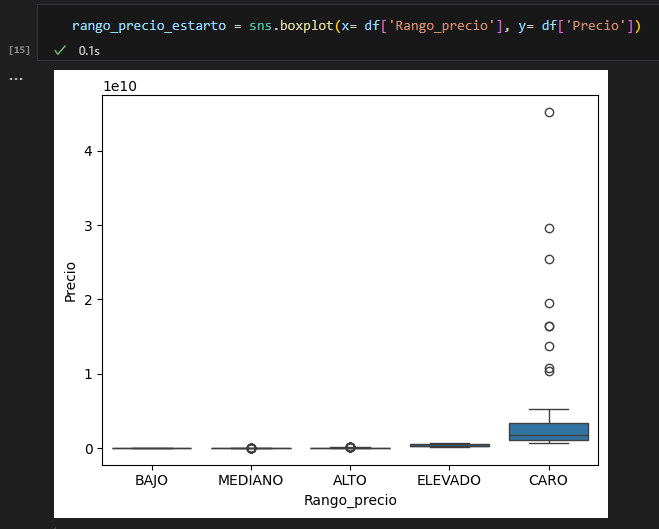
Este grafico ayuda a identificar la cantidad de datos que comparten una característica conocida y la cuantifica en un porcentaje, de esta forma podemos visualizar que porcentaje de cada dato comparte una misma característica.

* Gráfico de frecuencia:



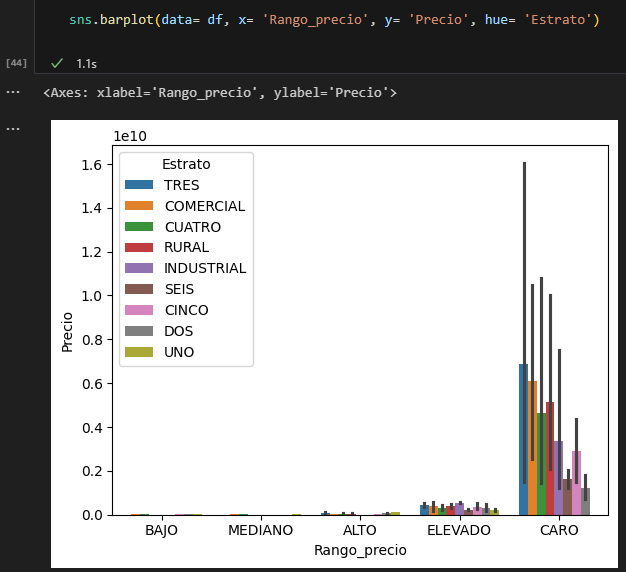
Al ser valores tan grandes y concentrarse en ciertas zonas hace muy difícil apreciar la tendencia del precio por un tipo de propiedad en específico.

* Gráfico para valores atípicos (Caja y bigotes):



Se presentan muchos más valores atípicos de valor ‘Caro’ que, en las demás categorías, la desigualdad en los datos hace muy difícil medir las demás categorías bajo el mismo gráfico.

* Gráfico para análisis multivariado:



En el gráfico se muestran las propiedades con un rango de precio más alto según su estrato, esto nos ayuda a entender que propiedades son más caras y como no depende solo del estrato su precio, aunque también puede ser por la cantidad de las propiedades por estrato no es igualitaria.

Conclusión

* Según los datos podemos afirmar que los precios de las propiedades si varían dependiendo de su estrato, también vemos que la mayoría del valor se concentra en los de estrato 3 y comercial.
* Los precios para las propiedades estrato 3 presentan la mayoría de los valores atípicos ya que en todos los datos solo hay 19 propiedades de este estrato y aun así sustentan la mayor cantidad de los precios altos de todos los datos.
* Dependiendo de que se necesite sustentar el análisis exploratorio de estos datos puede responder más preguntas, sin embargo, en la parte de la recolección de datos donde muchos datos no concuerdan o faltan hace a la vez limitado su desarrollo.
* La importancia de las gráficas en el análisis de datos puede maximizar el entendimiento de los datos y ayudar a tomar las decisiones de forma más consciente de todo lo que rodea a la información dada por los datos.