LABORATORIO 3

Gabriel Nicolas Aguilera Niño

1. DESCRIPCIÓN DE LOS DATOS

- **¿Cuál es el formato de los datos?**

- **Habitaciones (int)**
- **Baños (int)**
- **Parqueaderos (int)**
- **area_construida (float)**
- **area_privada (float)**
- **estrato (int)**
- **administración (int)**
- **ubicación (string)**
- **precio (int)**

- **¿Qué tamaño tiene la base de datos (en número de filas y columnas)?**

Un archivo CSV de 8429 filas con información sobre la finca principal de Bogotá, que pesa solo 1.0 MB.

.¿**Incluyen los datos una o más variables relevantes para la pregunta de negocio?**

Si se pueden usar los datos de las columnas que no utilizamos para clientes que quieren hacer preguntas comerciales, como si tiene zonas verdes cerca, colegios cerca, aparcamiento para visitantes, etc. Si se le da la misma ubicación, se puede determinar cuán cerca está de las áreas clave de la ciudad, por ejemplo, una vivienda en SUBA de puede estar a muchos kilómetros de la zona bancaria de la 72 o de la universidad.

- **¿Qué tipos de datos están presentes (simbólicos, numéricos, etc.)?**

- ***Int***
- ***Float***
- ***String***

- **¿Ha calculado estadísticas básicas para las variables clave? ¿Qué información le ha proporcionado sobre la cuestión de negocio?**

Se ha aplicado el análisis estadístico más simple disponible a la columna de habitaciones. Fue saber la moda; no creo que me ayude a crear un modelo de negocio, pero si me ayuda a comprender que se venden principalmente casas con cuatro habitaciones.

```
from collections import Counter

fileFR['habitaciones'] = datos_divididos.iloc[:, 1].replace('No definida',

0).fillna(0).astype(int) contador = Counter(fileFR['habitaciones']) moda =

contador.most_common(1)[0][0]
```

2. EXPLORACIÓN DE LOS DATOS

- **¿Qué tipo de hipótesis se ha formado sobre los datos?**

- La relación entre el precio y el tamaño (en metros cuadrados)

- El impacto de la ubicación en el precio

- Influencia de cualidades adicionales

-El impacto que tiene el número de habitaciones en el precio

El precio, que se puede evaluar en comparación con otras variables, demuestra que es una variable crucial.

- **¿Qué variables parecen prometedoras para un análisis más profundo?**

Para hacer un análisis más profundo, podríamos utilizar una variable llamada nombre, que podría ser categórica y decirnos si la vivienda es una casa, un apartamento, un lote, un edificio, etc., y ahora podríamos usar variables dumi para determinar si hay lugar cerca, por ejemplo:

Zonas verdes si, Zonas verdes no

Estudio cerca si, Estudio cerca no

Chimenea si, Chimenea no

Cocina integral si, cocina integral no

Entre otros datos en los que podríamos evaluar con un poco más de precisión una vivienda o comparar su precio con la demanda.

- **¿Sus exploraciones han revelado nuevas características sobre los datos?**

Analiza la distribución de otros factores, como el tamaño de la casa y el número de habitaciones, entre otros. Esto nos podría decir, por ejemplo, que la mayoría de las casas tienen entre 3 y 4 habitaciones.

Calcule las medidas de tendencia central y dispersión. Esto, por ejemplo, nos ayudaría a calcular la división del precio. La casa de $300.000.000 tiene una desviación de $50.000.000.

Por ejemplo, al analizar patrones y tendencias, podemos concluir que las casas más grandes, nuevas y ubicadas en áreas céntricas tienen precios significativamente más altos o que las casas con más habitaciones tienen una alta correlación positiva con los precios.

¿Cómo han cambiado estas exploraciones su hipótesis inicial?

Puedo decir que las hipótesis que hice al principio aún se mantienen, pero he demostrado que se puede realizar un análisis mucho más profundo para ser mucho más preciso con los modelos que se utilizan al analizar o evaluar los datos.

- **¿Considera que debería reformular el alcance del proyecto?**

Con total seguridad, diría que si es posible reformular el alcance si realizamos un análisis con los datos iniciales y luego agregamos más datos a ese modelo, podemos requerir un poco más de análisis.
Si al principio el propósito del proyecto era estimar los precios de la demanda utilizando un modelo de recesión, ahora se podría utilizar un modelo predictivo para analizar los patrones, tendencias y variabilidad en los precios y su impacto en los clientes.

- **¿Esta exploración ha alterado los objetivos?**

Si se considera reformular el alcance del proyecto, entonces claramente afectaría todos los objetivos del proyecto.

- **¿Puede identificar subconjuntos particulares de datos para su uso posterior?**

Se podría utilizar un balcón, un citófono o una barra de estilo americano para realizar otro análisis un poco más pequeño para evaluar luje de detalles de la vivienda, para hacer un análisis más preciso de la influencia de los clientes o incluso para determinar la relación entre los detalles de interiores y el precio.

3. VERIFICACIÓN DE CALIDAD DE LOS DATOS

- **¿Ha identificado variables faltantes y campos en blanco? Si es así, ¿Hay algún significado detrás de tales valores faltantes?**

Si hay datos n registrados en cada una de las 30 columnas del archivo CSV, en este caso pasé a tomar 2 opciones. En primer lugar, hice un conteo de las columnas en las que aparecían los datos con esa anotación. Si son pocos, se pueden eliminar, pero si son suficientes, se debe considerar si esa columna se considera o no.

También se evidenciaron columnas corridas en el archivo CSV; sin embargo, no se pueden tener en cuenta estos datos porque agregarlos podría afectar la precisión del análisis.
Para lograrlo, es recomendable eliminar los datos o, dependiendo del cálculo, dejar el valor en cero; sin embargo, cero también es un valor y puede tener un impacto en la precisión del modelo.

- **¿Hay inconsistencias ortográficas que puedan causar problemas en fusiones o transformaciones posteriores?**

Si. si los caracteres especiales no se pueden leer en el formato CSV, se debe realizar un remplazo antes de realizar una transformación. Por ejemplo, si la palabra es "organización", se deja como "organización" y se hace un remplazo de carácter especial, luego se aplica la transformación y luego la función del modelo.

- **¿Ha explorado las desviaciones para determinar si son "ruido" o fenómenos que vale la pena analizar más a fondo?**

El CSV contiene 30 variables totales, por lo que podemos comenzar a identificar cuáles se pueden usar como ruido. Por lo tanto, si necesito realizar un primer análisis con las características y el precio, no tendría que usar las variables restantes. Sin embargo, podría comenzar a analizar cada variable para determinar si es ruido o no. Por último, si solo tengo la variable citófono y no la uso porque la interpretaría como ruido, pero si uso otras variables, debo tenerla en cuenta.

- **¿Ha considerado excluir datos que no tienen impacto en sus hipótesis?**

Si claro, solo necesito 10 variables de las 30 variables del CSV para mi hipótesis; si se necesita un análisis más profundo, podría excluirlas o tenerlas aparte.

- **¿Los datos se almacenan en archivos planos? Si es así, ¿Son los delimitadores coherentes entre los archivos?**

Si, actualmente se encuentran delimitados por comas (,)

- **¿Cada registro contiene el mismo número de campos?**

No, algunos registros están corridos y no concuerdan con el numero de campos del archivo, por mi parte procedí a eliminar esos registros para que todos quedaran similares.

LINEAMIENTOS DE IBM

• ¿Qué variables (columnas, atributos) de la(s) tabla(s) o base(s) de datos parecen más prometedores?

- **estrato**
- **administración**
- **ubicación**
- **habitaciones**
- **area_privada**
- **baños**
- **parqueaderos**
- **area_construida**

Podría decir que estas son las características principales de la vivienda, lo que le permite segmentar las viviendas por características e identificar patrones, lo que le permite hacer marketing o ventas.

Además, se podría hacer una predicción de los precios de los bienes raíces con estos datos, pero en nuestro caso, ya tenemos el precio:

- **precio**

Estas son las características principales de la vivienda, lo que le permite segmentar las viviendas por características e identificar patrones, lo que le permite hacer marketing o ventas.

Además, se podría hacer una predicción de los precios de los bienes raíces con estos datos, pero en nuestro caso, ya tenemos el precio:

• ¿Qué variables parecen irrelevantes y pueden ser excluidos?

Aunque cualquier información es útil para realizar un análisis adecuado de mis datos, algunos campos pueden ser menos relevantes o incluso eliminarse:

- **Parqueadero Visitantes**
- **Portería / Recepción**
- **Zonas Verdes**
- **Salón Comunal**
- **Barra estilo americano**
- **Calentador**
- **Chimenea**
- **Citófono**
- **Cocina Integral**
- **Terraza**
- **Vigilancia**
- **Parques cercanos**
- **Circuito cerrado de TV**
- **Estudio**
- **Depósito / Bodega**

Aunque cualquier dato puede ser útil para realizar un análisis en este caso, debemos seleccionar los que sean más significativos para aplicar un modelo

.**¿Hay suficientes datos para sacar conclusiones generalizables o hacer predicciones precisas?**

Aunque el archivo pesa 1.00MB, aún es considerado un archivo pequeño con 8429 datos, pero con estos datos podemos realizar un análisis para pequeños sectores de Bogotá.

Con los datos actuales, se puede aplicar una presión precisa sobre las necesidades del cliente y la variabilidad del precio de la vivienda.

¿Hay demasiadas variables para el método de modelado de su elección?

Todo depende de la cantidad de datos que se utilizarán, por ejemplo, si timo todos los campos del CSV y los limpio, tendré 30 variables que usar.

```
# Asignar cada dato a columnas específicas en fileFR
fileFR['habitaciones'] = datos_divididos.iloc[:, 1].replace('No definida', 0).fillna(0).astype(int) fileFR['baños'] = datos_divididos.iloc[:, 2] fileFR['parqueaderos'] = datos_divididos.iloc[:, 3] fileFR['area_construida'] = datos_divididos.iloc[:, 4] fileFR['area_privada'] = datos_divididos.iloc[:, 5] fileFR['estrato'] = datos_divididos.iloc[:, 6] fileFR['estado'] = datos_divididos.iloc[:, 7] fileFR['antiguedad'] = datos_divididos.iloc[:, 8] fileFR['administracion'] =
```

```
datos_divididos.iloc[:, 9] fileFR['precio_m2'] = datos_divididos.iloc[:, 10]
fileFR['Ascensor'] = datos_divididos.iloc[:, 11]
fileFR['Circuito cerrado de TV'] = datos_divididos.iloc[:,
12] fileFR['Parqueadero Visitantes'] =
datos_divididos.iloc[:, 13] fileFR['Portería / Recepción'] =
datos_divididos.iloc[:, 14] fileFR['Zonas Verdes'] =
datos_divididos.iloc[:, 15] fileFR['Salón Comunal'] =
datos_divididos.iloc[:, 16] fileFR['Balcón'] =
datos_divididos.iloc[:, 17]
fileFR['Barra estilo americano'] =
datos_divididos.iloc[:, 18] fileFR['Calentador'] =
datos_divididos.iloc[:, 19] fileFR['Chimenea'] =
datos_divididos.iloc[:, 20] fileFR['Citófono'] =
datos_divididos.iloc[:, 21] fileFR['Cocina Integral'] =
datos_divididos.iloc[:, 22] fileFR['Terraza'] =
datos_divididos.iloc[:, 23] fileFR['Vigilancia'] =
datos_divididos.iloc[:, 24] fileFR['Parques cercanos']
= datos_divididos.iloc[:, 25] fileFR['Estudio'] =
datos_divididos.iloc[:, 26] fileFR['Patio'] =
datos_divididos.iloc[:, 27] fileFR['Depósito / Bodega']
= datos_divididos.iloc[:, 28] fileFR['nombre'] =
datos_divididos.iloc[:, 29] fileFR['ubicacion'] =
datos_divididos.iloc[:, 30] fileFR['precio'] =
datos_divididos.iloc[:, 31]
```

Sin embargo, si solo utilizo 10 columnas de todo el CSV, solo tendría que limpiar 10 columnas. Esto ocurre cuando se utiliza un modelo de variabilidad o análisis, ya que se debe pasar el dataFrame completo.

```
# Data Frame de Ejemplo
data = {
```

```
    'habitaciones': [2, 2, 3, 3, 4, 4],
    'precio': [200, 220, 250, 270, 300, 320]
}
# Crear DataFrame df
= pd.DataFrame(data)

# Formulación del modelo ANOVA
# Aquí, 'precio' es la variable dependiente y 'habitaciones' es la variable
independiente model = ols('precio ~ C(habitaciones)', data=df).fit()
# Realizar el ANOVA
anova_table = sm.stats.anova_lm(model, typ=2)

print(anova_table)
```

- **¿Ha considerado cómo se manejan los valores que faltan en cada uno de sus orígenes de datos?**

Para este caso, los datos que no se proporcionaron se mostraron como "No definida", así que usé un remplazo en Excel y los cambié por número 0. Esto también se puede usar con Python de la siguiente manera:

```
Str.remplace(“No definida”, 0). .fillna(0).astype(int)
```

Por lo tanto, la palabra "No definida" no se tomará al aplicar el modelo. Si no hay un número 0, no debería tener ningún impacto en el cálculo.