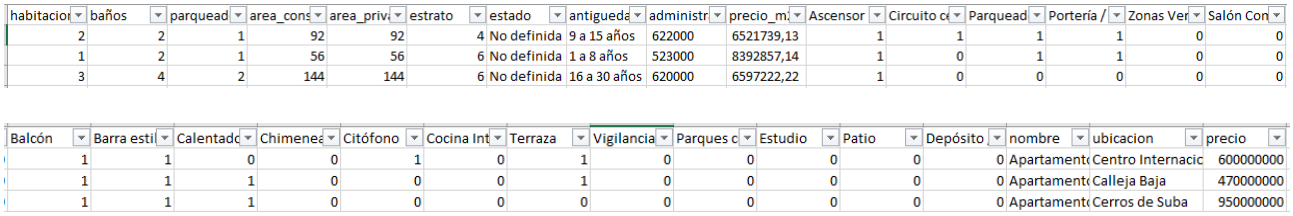
# LINEAMIENTOS DE IBM – LABORATORIO 3

## 1. RECOLECCIÓN DE DATOS INICIALES

A continuación, mostrare dos fotos de mis datos iniciales ya limpios para poder aplicar los lineamientos de IBM en mi data recolectada.



| habitacior | baños | parquead | area_cons | area_priv | estrato | estado | antigueda | administr | precio_m | Ascensor | Circuito c | Parquead | Portería / | Zonas Ver | Salón Con |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 2 | 1 | 92 | 92 | 4 | No definida | 9 a 15 años | 622000 | 6521739,13 | 1 | 1 | 1 | 1 | 0 | 0 |
| 1 | 2 | 1 | 56 | 56 | 6 | No definida | 1 a 8 años | 523000 | 8392857,14 | 1 | 0 | 1 | 1 | 0 | 0 |
| 3 | 4 | 2 | 144 | 144 | 6 | No definida | 16 a 30 años | 620000 | 6597222,22 | 1 | 0 | 0 | 0 | 0 | 0 |

| Balcón | Barra estil | Calentado | Chimenea | Citófono | Cocina Int | Terraza | Vigilancia | Parques c | Estudio | Patio | Depósito | nombre | ubicacion | precio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | Apartament | Centro Internacic | 600000000 |
| 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | Apartament | Calleja Baja | 470000000 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Apartament | Cerros de Suba | 950000000 |

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

Analizando los datos pueden ser variables muy prometedoras serian:

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

Estas se podrían decir que son las características principales de la vivienda, con esto podría segmentar las viviendas por características identificar patrones, con ello podría hacer marketing o ventas.

Con estos datos también se podía hacer una predicción de precios del inmueble, pero en nuestro caso como ya tenemos el precio:

- **precio**

Podríamos hacer un análisis y una variabilidad de los precios y ver que tan significativa es.

También podríamos hacer un análisis para verificar la satisfacción de los clientes en lo que quiere para una vivienda o en lo que desea adquirir.

- **¿Qué variables parecen irrelevantes y pueden ser excluidos?**

A pesar que cualquier dato sirve para hacer un buen análisis para mis los campos que pueden ser menos relevantes o talvez quitarlos pueden ser:

- **Circuito cerrado de TV**
- **Parqueadero Visitantes**
- **Portería / Recepción**
- **Zonas Verdes**
- **Salón Comunal**
- **Barra estilo americano**
- **Calentador**
- **Chimenea**

- **Citófono**
- **Cocina Integral**
- **Terraza**
- **Vigilancia**
- **Parques cercanos**
- **Estudio**
- **Depósito / Bodega**

A pesar que todo dato puede servirnos para hacer un análisis en este caso debemos tomar los que sean mas importantes para aplicar un modelo.

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

Pues mi archivo pesa 1.00MB se considera un archivo pequeño a pesar de ser 8429 datos aun sigue siendo muy pequeño, pero con estos datos podemos hacer un análisis ya sea para pequeños sectores de Bogotá.

Con los datos actuales se podría dar una presión precisa en la variabilidad del precio de las viviendas o en exigencias del cliente.

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

Todo depende de la cantidad de datos que se vayan a utilizar, es decir: Si timo todos los campos del CSV y los limpio eso quiere decir que tengo 30 variables que puedo utilizar.

```
# 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]
```

Pero si de todo el CSV solo voy a utilizar solo 10 entonces solo tendría que limpiar 10 columnas, cuando se este aplicando un modelo de análisis o de variabilidad pues se debe pasar el dataFrame completo, por ejemplo.

```
# 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 mi caso los datos que no se proporcionaron aparecían como **"No definida"** lo que hice fue aplicar un remplazar con Excel y cambiarlos por **número 0,** esto también se puede aplicar con Python de la siguiente manera:

Str.remplace("No definida", 0). .fillna(0).astype(int)

Así al momento de aplicar el modelo no va tomar la palabra No definida si no el numero 0 entonces no debería afectar al momento del cálculo.

## 2. DESCRIPCIÓN DE LOS DATOS

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

En este caso para estas columnas que se están utilizando en el CSV se dejaron de la siguiente manera

- **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)?**

En este caso se creo un CSV de 8429 filas con información de finca raíz en Bogotá, el peso del archivo solo es 1.0 MB, pero se podría generar una mayor cantidad de datos.

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

Si, se podrían utilizar datos de columnas que no utilizamos para clientes que quieran preguntas de negocios es decir: podemos brindar datos si tiene zonas verdes cerca, si hay colegios cerca, si cuenta con parqueadero de visitantes ETC, con la misma ubicación que se le brinda, se podría saber que tan cerca esta de las zonas importantes de la ciudad por ejemplo si una vivienda esta en SUBA de podría tener a cuentos kilómetros esta del la ZONA BANCARIA DE LA 72 o de universidades.

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

Para este caso se manejan 3 tipos de datos:

- Int
- Float
- String

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

He aplicado el análisis más básico en estadística que existe y lo aplique a la columna de habitaciones y fue saber la moda, no creo que me ayude a generar un modelo de negocio, pero si me ayudo a identificar que se venden casas con 4 habitaciones mayormente.

```
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]
```

## 3. EXPLORACIÓN DE LOS DATOS

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

Sin la necesidad de aplicar un modelo de análisis de datos sobre los datos que he limpiado usando un análisis lógico se pueden concluir las siguientes hipótesis

- **Relación entre el tamaño (metros cuadrados) y el precio**
- **Impacto de la ubicación en el precio**
- **Influencia de características adicionales**
- **Influencia del número de habitaciones en el precio**

Se evidencia que una variable muy importante es el precio ya que como variable primordial se puede evaluar frente a las otras.

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

Para hacer un análisis más profundo podríamos utilizar una variable llamada **nombre,** esta podría ser categórica y decirnos si la vivienda es casa, apartamento, casa lote, edificio etc, ahora podríamos usar variables **dumi** para identificar si existen 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?**

Examina la distribución de otras variables, como el tamaño de la casa, el número de habitaciones, etc. Esto nos podría decir que la mayoría de casas tienen entre 3 y 4 habitaciones, por ejemplo.

Calcula medidas de tendencia central y dispersión (desviación estándar, rango). Esto nos ayudaría a medir la división en el precio, por ejemplo. La casa de $300.000.000 tienen una desviación de $50.000.000.

Analizar patrones y tendencias por ejemplo podemos decir que las casas más grandes, nuevas y ubicadas en zonas céntricas tienen precios significativamente más altos o que las casas con más habitaciones tienen precios más altos se valida con una alta correlación positiva.

- **¿Cómo han cambiado estas exploraciones su hipótesis inicial?**

Frente a las hipótesis que hice al inicio puedo decir que aún se mantienen, pero logro evidenciar que se puede hacer un análisis mucho mas profundo para ser mucho mas precisos con los modelos que se implementen al momento de analizar o evaluar los datos.

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

Con plena seguridad me atrevería decir que si es posible reformular el alcance si hacemos un análisis con los datos iniciales y luego agregamos mas datos a ese modelo podemos precisar un poco más en el análisis.

Si al inicio el alcance del proyecto era aplicar un modelo de recesión para estimar los precios de la demanda ahora se podría aplicar un modelo predictivo para analizar los posible patrones, tendencias y variabilidad en precios y influencia 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 balcón, citófono o barra estilo americano para crear otro análisis un poco más pequeño para evaluar luje de detalles de la vivienda o hacer un análisis mas preciso en la influencia de clientes o incluso una relación entre detalle de interiores y el precio.

## 4. 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, en cada una de las 30 columnas que se evidencian del archivo CSV se evidencia que hay datos que n se registran en este caso aparece la palabra “No definida”, en este caso pasé a tomar 2 opciones, como primer paso hice un conteo por columnas de que datos aparecían con esa anotación. Si son pocos pueden eliminarse, si son bastantes de debe tener en cuenta si esa columna se tiene en cuenta o no se tiene en cuenta.

Para el archivo CSV también se evidenciaron columnas corridas, tampoco se puede tener en cuanta ese dato porque si tratamos de agregar un dato, puede dañarnos la precisión en el análisis.

Para ello lo mejor es eliminar el dato o en su defecto dependiendo el cálculo se podría dejar en cero 0 pero cero también es un valor y podría afectar la precisión del modelo.

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

Si, realmente en el formato que viene el CSV no se pudieron leer los caracteres especiales, antes de hacer una transformación toco hacer un remplazo, es decir, si la palabra es **organización** entonces se deja como **organización** se hace el remplazo de carácter especial, luego si se aplica la transformación y posteriormente 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 trae un total de 30 variables desde aquí podríamos empezar a determinar cuales se pueden determinar como ruido. Es decir, si necesito hacer un primer análisis con las características mas el precio, no sería necesario utilizar las variables sobrantes. Pero si quiero hacer un análisis más profundo podría empezar a tomar cada variable y determinar si es ruido o no es ruido. En conclusión, si tengo la variable citófono y no la voy a utilizar pues la determinaría como ruido, pero si voy a complementarla con otras variables entonces debo tenerla en cuenta.

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

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

- **¿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.