**TALLER 3**

**LINEAMIENTOS DE IBM**

**ESTIVEN SIERRA MARTINEZ**

**UNIVERSIDAD ECCI**

**FACULTAD DE INGENIERIAS**

**ING. DE SISTEMAS**

**BOGOTÁ, D.C.**

**2024**

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.

• ¿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 más 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 esté 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 creó 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