

LABORATORIO 3

SEMINARIO DE BIG DATA Y CIENCIAS DE DATOS

CARLOS EDUARDO CORREA ARCHILA-73139

PROFESOR

ELIAS BUITRAGO

UNIVERCIDAD ECCI

BOGOTA, COLOMBIA
2024

Laboratorio de Comprensi3n de Datos

Introducci3n

En este documento, vamos a trabajar en la limpieza del archivo de datos llamado "housing_fincaraiz.csv". Adem1s, vamos a seguir los lineamientos de IBM para entender los datos, siguiendo el proceso de "Data Understanding".

Fase 1: Recolectar Datos Iniciales

En esta primera fase, nuestro objetivo es reunir y revisar el conjunto de datos que vamos a utilizar. Esto incluye obtener el archivo de datos y hacer una primera evaluaci3n para entender su contenido y estructura. Es el punto de partida para todo el proceso de an1lisis de datos.

habitaciones	baños	parqueadero	area_const	area_priva	estrato	estado	antigüedad	administrat	precio_m2	Ascensor	Círculo cel	Parqueadero	PorterA-a	Zonas Verde	Salañ Con	Balcñ	Barra e
2	2	1 92 m12	92 m12		4	No definida	9 a 15 a1os	\$1 622.000 CI \$1 6.521.739,		1	1	1	1	0	0	1	
1	2	1 56 m12	56 m12		6	No definida	1 a 8 a1os	\$1 523.000 CI \$1 8.392.857,		1	0	1	1	0	0	1	
3	4	2 144 m12	144 m12		6	No definida	16 a 30 a1os	\$1 620.000 CI \$1 6.597.222,		1	0	0	0	0	0	1	
1	1	0 31 m12	31 m12		4	Excelente	menor a 1 a1	\$1 130.000 CI \$1 7.419.354,		1	1	1	0	0	0	0	
3	2	1 52 m12	52 m12		4	No definida	1 a 8 a1os	\$1 219.000 CI \$1 5.576.923,		1	1	1	0	0	0	1	
3	3	1 150 m12	150 m12		6	Bueno	m1s de 30 a1	\$1 872.000 CI \$1 6.533.333,		1	1	1	0	0	0	0	
3	2	1 110 m12	100 m12		3	No definida	16 a 30 a1os	\$1 135.000 CI \$1 4.181.818,		0	1	0	0	1	0	0	
3	2	0 53 m12	47 m12		3	No definida	9 a 15 a1os	\$1 125.000 CI \$1 3.207.547,		0	0	1	1	1	1	0	
3	3	1 111 m12	0 m12		4	Excelente	16 a 30 a1os	No definida \$1 3.873.873,		0	1	1	0	0	1	0	
4	4	2 264 m12	264 m12		5	Bueno	m1s de 30 a1	\$1 836.000 CI \$1 5.303.030,		1	0	0	0	0	0	0	
3	2	2 97 m12	0 m12		4	No definida	9 a 15 a1os	\$1 272.000 CI \$1 6.175.257,		1	1	0	0	0	0	1	
2	3	2 87 m12	82 m12		4	No definida	9 a 15 a1os	\$1 350.000 CI \$1 4.712.643,		1	1	1	1	0	1	0	
3	4	2 175 m12	175 m12		4	No definida	16 a 30 a1os	No definida \$1 4.285.714,		0	0	0	0	0	1	0	
10	5	2 391 m12	0 m12		3	No definida	16 a 30 a1os	No definida \$1 1.994.884,		0	0	0	0	0	0	0	
6	4	3 218 m12	0 m12		4	Bueno	m1s de 30 a1	No definida \$1 3.027.522,		0	1	0	0	0	0	0	
1	2	1 50 m12	0 m12		5	Bueno	9 a 15 a1os	\$1 310.000 CI \$1 6.800.000,		0	0	0	1	0	0	1	
3	3	1 140 m12	140 m12		4	Bueno	m1s de 30 a1	\$1 460.000 CI \$1 5.071.428,		0	1	0	1	0	1	0	
4	3	1 90 m12	0 m12		4	Bueno	m1s de 30 a1	\$1 369.000 CI \$1 4.000.000,		1	0	1	1	0	1	0	
2	3	2 111 m12	111 m12		6	Bueno	9 a 15 a1os	\$1 786.000 CI \$1 11.261.261,		1	1	1	0	0	0	1	
12	5	0 375 m12	320 m12		3	Bueno	16 a 30 a1os	No definida \$1 1.493.333,		0	0	0	0	0	0	1	
2	3	2 111 m12	73 m12		4	No definida	16 a 30 a1os	\$1 483.000 CI \$1 3.423.423,		1	0	1	1	0	1	0	
4	4	2 307 m12	307 m12		3	No definida	m1s de 30 a1	No definida \$1 2.931.596,		0	1	1	0	0	1	0	
3	2	0 61 m12	0 m12		3	No definida	16 a 30 a1os	\$1 130.000 CI \$1 2.622.950,		0	0	1	0	0	1	0	
2	3	4 322 m12	277 m12		6	No definida	1 a 8 a1os	\$1 2.304.000 \$1 9.000.000,		1	1	0	0	0	0	1	
5	5	4 280 m12	280 m12		6	No definida	16 a 30 a1os	\$1 2.200.000 \$1 7.142.857,		0	1	1	1	1	0	0	
3	2	1 59 m12	54 m12		4	Bueno	16 a 30 a1os	\$1 150.000 CI \$1 4.220.338,		0	1	1	1	0	1	0	
3	3	2 146 m12	133 m12		5	Excelente	16 a 30 a1os	\$1 660.000 CI \$1 5.410.958,		1	1	0	1	1	0	1	

An1lisis de Variables en la Base de Datos

Variables M1s Prometedoras

En la base de datos, las siguientes variables parecen ser las m1s 1tiles para nuestro an1lisis:

- Habitaciones: El n1mero de habitaciones puede influir en el precio y en la demanda de una propiedad.
- Baños: La cantidad de baos tambi3n es un factor importante que afecta el valor de una propiedad.
- Parqueadero: La disponibilidad de parqueadero puede ser un aspecto relevante para los compradores.
- 1rea Construida: El tamao de la propiedad en metros cuadrados es una medida clave para determinar su valor.

- Área Privada: Similar al área construida, pero se enfoca en el espacio privado de la propiedad.
- Estrato: En algunas regiones, el estrato socioeconómico puede afectar el precio de las propiedades.
- Antigüedad: La edad de la propiedad puede influir en su valor y condición.
- Administración: Los costos de administración pueden ser importantes para los compradores potenciales.
- Precio por m²: Es una medida directa del costo de la propiedad en función de su tamaño.
- Zonas Verdes: La cercanía a áreas verdes puede ser un atractivo para los compradores.
- Calentador: La presencia de un calentador puede ser una característica buscada por los compradores.
- Cocina Integral: Una cocina bien equipada puede agregar valor a la propiedad.
- Vigilancia: La seguridad proporcionada por la vigilancia puede ser un factor decisivo para algunos compradores.
- Parques Cercanos: La proximidad a parques es un aspecto que puede ser importante para las familias.
- Nombre: El nombre puede no ser directamente relevante para el análisis, pero puede ser útil para identificar propiedades.
- Ubicación: La ubicación es crucial y puede influir en el valor de la propiedad.
- Precio: El precio de la propiedad es la variable dependiente principal en nuestro análisis.

Variables Irrelevantes que Se Pueden Excluir

Algunas variables parecen ser menos importantes para el análisis y podrían ser excluidas:

- Ascensor: Puede no ser crucial para la valoración en ciertos contextos.
- Circuito Cerrado de TV: Puede ser menos relevante dependiendo de la región y el tipo de propiedad.
- Parqueadero de Visitantes: No siempre es un factor determinante en el precio.
- Portería/Recepción: A menudo no afecta significativamente el valor de una propiedad.
- Salón Comunal: Puede ser menos relevante para el análisis del precio de una propiedad.
- Balcón: Dependiendo del mercado, un balcón puede no ser un factor decisivo.
- Barra Estilo Americano: Puede ser una característica menor en el contexto del valor de la propiedad.

- Chimenea: No siempre es un aspecto importante en todas las regiones.
- Citófono: Puede no ser relevante para el análisis del valor de la propiedad.
- Terraza: Similar al balcón, puede ser menos relevante dependiendo del mercado.
- Estudio: Puede no ser una característica clave en la valoración de propiedades.
- Patio: En algunos contextos, puede ser una característica menos relevante.
- Depósito/Bodega: No siempre es un factor determinante en el valor de una propiedad.

Datos Suficientes para Conclusiones y Predicciones

Con un total de 8429 registros en el conjunto de datos, se cuenta con una cantidad suficiente de datos para hacer análisis generalizables y para realizar predicciones precisas.

Número de Variables para el Modelado

El conjunto de datos tiene 31 variables, lo que podría ser un número alto dependiendo del método de modelado elegido. Por lo tanto, se podría considerar reducir el número de variables para mejorar la precisión del análisis.

Fusión de Fuentes de Datos

El conjunto de datos actual es independiente y no implica la fusión de varias fuentes de datos.

Manejo de Valores Faltantes

Durante la revisión de datos, no se encontraron espacios en blanco ni valores faltantes en el archivo, por lo que no es necesario abordar este tema en esta fase.

Resumen

- Variables más prometedoras: Habitaciones, baños, parqueadero, área construida, área privada, estrato, antigüedad, administración, precio por m², zonas verdes, calentador, cocina integral, vigilancia, parques cercanos, nombre, ubicación, precio.
- Variables irrelevantes: Ascensor, circuito cerrado de TV, parqueadero de visitantes, portería/recepción, salón comunal, balcón, barra estilo americano, chimenea, citófono, terraza, estudio, patio, depósito/bodega.
- Datos suficientes: Con 8429 registros, hay datos suficientes para hacer predicciones.

- Número de variables: Con 31 variables, puede ser necesario reducirlas para mejorar el análisis.
- Fusión de datos: No se está fusionando ninguna fuente de datos.
- Valores faltantes: No se encontraron valores faltantes en los datos.

Espero que esta explicación te sea útil. Si necesitas más detalles o ajustes, no dudes en decírmelo.

Fase 2: Describir los Datos

Formato de los Datos

El conjunto de datos contiene diferentes tipos de información, incluyendo datos numéricos, alfanuméricos y caracteres.

Método de Captura de Datos

Los datos se han recopilado a partir de registros de casas en finca raíz, es decir, información sobre propiedades inmobiliarias.

Tamaño de la Base de Datos

La base de datos tiene 8,429 filas y 31 columnas, lo que da un total de 261,299 datos individuales.

Variables Relevantes para la Pregunta de Negocio

Algunas variables clave en el conjunto de datos que son importantes para responder a las preguntas de negocio incluyen:

- Precio: El costo de la propiedad.
- Valor del Metro Cuadrado: Cuánto cuesta cada metro cuadrado de la propiedad.
- Estrato: El nivel socioeconómico del área donde se encuentra la propiedad.
- Valor de la Administración: El costo de la administración del edificio o complejo.
- Ubicación: La dirección o zona donde se encuentra la propiedad.

Estas variables son útiles para analizar el mercado inmobiliario y entender qué factores afectan el valor de las propiedades.

Tipos de Datos Presentes

En el conjunto de datos se encuentran datos numéricos, alfanuméricos, caracteres y simbólicos.

Estadísticas Básicas y Variables Clave

Se han calculado estadísticas básicas para las variables clave, como promedios, medianas y desviaciones estándar. Estas estadísticas ayudan a entender mejor las características generales del mercado inmobiliario, como cuál es el rango de precios o el tamaño promedio de las propiedades.

Priorización de Variables

Es posible priorizar las variables más relevantes para el análisis, basándose en su impacto en el valor de las propiedades. Si hay dudas, se pueden consultar a analistas de negocio para obtener más información y clarificar qué variables deben tener más peso en el análisis.

Fase 3: Describir los Datos

Hipótesis sobre los Datos

La hipótesis inicial es que los datos reflejan los valores de propiedades y si estas incluyen características específicas, como vigilancia o zonas verdes, que podrían influir en el precio.

Variables Prometedoras para un Análisis Más Profundo

Las siguientes variables parecen ser útiles para un análisis más detallado:

- Precio: La variable principal para analizar el mercado.
- Estrato: Puede influir en el valor de la propiedad según el nivel socioeconómico.
- Valor del Metro Cuadrado: Ayuda a entender la relación entre el tamaño y el precio.
- Estado: El estado de la propiedad puede afectar su valor.
- Antigüedad: La edad de la propiedad puede influir en su valor.
- Habitaciones: El número de habitaciones puede ser un factor importante para los compradores.

- Valor de la Administración: Puede afectar la decisión de compra.
- Ubicación: La localización de la propiedad es crucial para su valor.

Exploraciones y Nuevas Características

La exploración de datos ha revelado detalles específicos sobre las propiedades, lo que permite un análisis más exhaustivo de lo que cada propiedad ofrece.

Cambio en la Hipótesis Inicial

La hipótesis inicial se mantiene, ya que el análisis confirma que los datos representan registros de propiedades con características específicas.

Reformulación del Alcance del Proyecto

El análisis actual no ha llevado a una necesidad de cambiar el alcance del proyecto ni ha alterado los objetivos iniciales.

Subconjuntos de Datos para Uso Posterior

Se pueden identificar subconjuntos de datos que contienen detalles sobre cada propiedad. Estos subconjuntos pueden usarse para análisis más detallados sobre cómo obtener el mejor precio para una propiedad.

Fase 4: Verificar la calidad de los datos

Cambios en los Nombres de las Columnas y Limpieza de Datos

Renombrar columnas:

- Se cambia el nombre de la columna “baños” a “banos”.



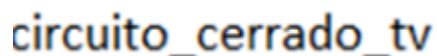
A screenshot of a data table interface showing a column header. The text 'banos' is displayed in a blue font, and a small dropdown arrow is visible to its right.

- Se cambia el nombre de la columna “Ascensor” a “ascensor”.



A screenshot of a data table interface showing a column header. The text 'ascensor' is displayed in a blue font, and a small dropdown arrow is visible to its right.

- Se cambia el nombre de la columna “Circuito cerrado de TV” a “circuito_cerrado_tv”.



A screenshot of a data table interface showing a column header. The text 'circuito_cerrado_tv' is displayed in a blue font.

- Se cambia el nombre de la columna “Parqueadero Visitantes” a “parqueadero_visitantes”.

parqueadero_visitantes

- Se cambia el nombre de la columna “Portería / Recepción” a “porteria-recepcion”.

porteria-recepcion

- Se cambia el nombre de la columna “Zonas Verdes” a “zonas_verdes”.

zonas_verdes

- Se cambia el nombre de la columna “Salón Comunal” a “salon_comunal”.

salon_comunal

- Se cambia el nombre de la columna “Balcón” a “balcon”.

balcon

- Se cambia el nombre de la columna “Barra estilo americano” a “barra_estilo_americano”.

barra_estilo_americano

- Se cambia el nombre de la columna “Calentador” a “calentador”.

calentador

- Se cambia el nombre de la columna “Chimenea” a “chimenea”.

chimenea

- Se cambia el nombre de la columna “Citófono” a “citofono”.

citofono

- Se cambia el nombre de la columna “Cocina Integral” a “cocina_integral”.

cocina_integral

- Se cambia el nombre de la columna “Terraza” a “terrazza”.

terrazza

- Se cambia el nombre de la columna “Vigilancia” a “vigilancia”.

vigilancia

- Se cambia el nombre de la columna “Parques cercanos” a “parques_cercanos”.

parques cercanos

- Se cambia el nombre de la columna “Estudio” a “estudio”.

estudio

- Se cambia el nombre de la columna “Patio” a “patio”.

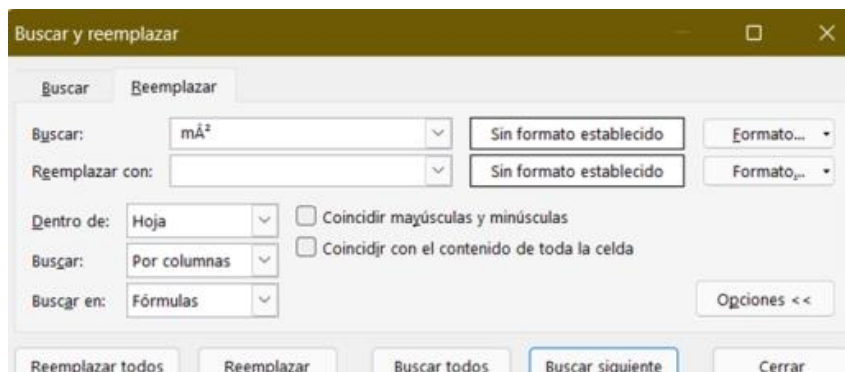
patio

- Se cambia el nombre de la columna “Depósito / Bodega” a “deposito-bodega”.

bodega

Limpieza de Datos:

A partir de estos cambios, se procederá a limpiar los datos en las columnas usando la función “Buscar y reemplazar”. Cuando sea necesario cambiar todos los datos a la vez, se utilizará la opción “Reemplazar todos”.



1. Columna “habitaciones”:

Se cambiaron los valores que estaban como “No definida” a “0” para asegurar que solo haya números en esta columna.

habitacion	Y
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0

2. Columna “banos”:

Se reemplazaron los valores “No definida” por “0” para que esta columna contenga únicamente datos numéricos.

banos	Y
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0
	0

3. Columna “parqueaderos”:

Se reemplazaron los valores “Más de 10” por “11”. Esto indica que si aparece “11”, significa que hay más de 10 parqueaderos disponibles.

C
parqueade
11

4. Columna “area_construida”:

Se eliminaron las unidades “m²” de los valores para que la columna muestre solo los números de metros cuadrados.

area_const
92
56
144
31
52
150
110

5. Columna “area_privada”:

Se eliminaron las unidades “m²” de los valores, dejando únicamente los números para representar el área en metros cuadrados.

area_privada
92
56
144
31
52
150
100

6. Columna “antigüedad”:

Se corrigieron errores en los caracteres, cambiando “años” por “anos” y “más” por “mas” para evitar problemas con los datos.

- ☒ (Seleccionar todo)
- ☒ 1 a 8 anos
- ☒ 16 a 30 anos
- ☒ 9 a 15 anos
- ☒ mas de 30 anos
- ☒ menor a 1 ano
- ☒ No definida

7. Columna “administracion”:

Se eliminaron el símbolo de peso y el formato de texto para dejar solo los números en “622000”. También se cambiaron las entradas con “No definida” a “0” para mantener solo valores numéricos.

- ☒ (Seleccionar tod
- ☒ 1
- ☒ 1.000.000
- ☒ 1.003.800
- ☒ 1.007.400
- ☒ 1.008.000
- ☒ 1.009.000
- ☒ 1.009.800
- ☒ 1.010.000

8. Columna “precio_m2”:

Se transformó el formato de los datos para que muestre solo los números, eliminando el símbolo de dólar, la unidad “m²” y otros caracteres innecesarios, como el asterisco. El resultado es un formato como “6521739,13”.

- ☒ (Seleccionar tod
- ☒ 10000000*
- ☒ 10032362,46*
- ☒ 10033841,95*
- ☒ 10045871,56*
- ☒ 10063291,14*
- ☒ 10064794,82*
- ☒ 10081967,21*
- ☒ 10084033,61*

9. Columna “ubicacion”:

Se corrigieron los caracteres con tilde que estaban mal representados, reemplazándolos por los caracteres correctos.

AD
ubicacion
Alfonso lopez
Alfonso Lopez
Alfonso lopez
AD
ubicacion
Bogota
Bogota
Bogota
Bogota
Bogota
Bogota
Bogota
Bogota

Revisión de Datos

Has encontrado variables faltantes o campos en blanco? ¿Qué podrían significar estos valores faltantes?

- No hay campos en blanco en el conjunto de datos, por lo que no hay valores faltantes que necesiten ser interpretados.

Existen errores ortográficos que podrían causar problemas en futuras fusiones o transformaciones de datos?

- Sí, durante el proceso de limpieza, se encontraron errores ortográficos como tildes incorrectas y el uso de la “ñ” que deben ser corregidos para evitar problemas en el análisis.

Has examinado las desviaciones en los datos para ver si son errores o si hay aspectos interesantes a investigar?

- Se identificaron algunas inconsistencias como propiedades con solo 1 metro cuadrado de área privada o construida, y propiedades con un valor de 1 COP. Estas desviaciones podrían ser errores o requerir un análisis más profundo.

Has comprobado la validez de los valores en los datos? ¿Has encontrado algún conflicto evidente?

- Se encontraron datos problemáticos, como propiedades con 1 metro cuadrado de área privada o construida y un valor de 1 COP, que podrían necesitar ser revisados o excluidos.

Has considerado eliminar datos que no ayudan a tus hipótesis?

- Sí, se está considerando excluir los datos problemáticos que no aportan valor a las hipótesis planteadas.

Los datos están en archivos planos? ¿Son consistentes los delimitadores entre los archivos?

- Los datos están en un archivo plano con campos separados por comas, y no se encontraron problemas con estos delimitadores.

Todos los registros tienen el mismo número de campos?

- Los registros son consistentes, ya que todos tienen el mismo número de campos desde el principio hasta el final de cada fila.