Andrés Felipe Rivas Luna

Daniel Felipe Galeano Tabares

Jorge Andrés Salas Barrera

LABORATORIO 3

El objetivo de este informe es proporcionar una guía detallada para entender y preparar el conjunto de datos "housing_fincaraiz.csv", siguiendo los lineamientos de IBM en lo que respecta a la comprensión de datos. Primero, se recolectaron los datos iniciales, identificando las variables clave como número de habitaciones, numero de baños, el precio, área construida. Luego, se describieron los datos, confirmando que están en formato .csv, con 31 columnas y cada variable con 8428 registros de datos, incluyendo datos numéricos, alfanuméricos, símbolos y texto. Finalmente, se verificó la calidad de los datos corrigiendo errores de escritura, ajustando nombres de variables y unidades de medida, y eliminando caracteres no deseados. Este proceso asegura que los datos estén limpios y listos para un análisis más profundo, proporcionando una base sólida para la toma de decisiones basada en datos.

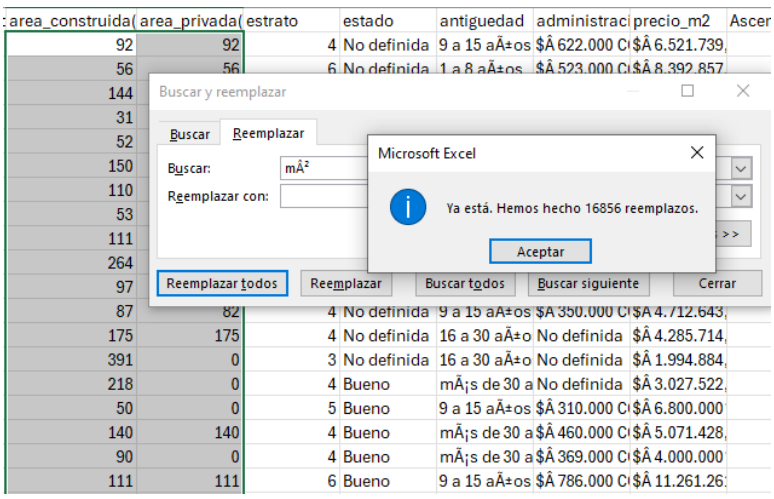
Verificación de datos

- Se corrigen los errores tipológicos de los nombres de cada variable

PorterÃ-a / R	Zonas Verde	SalÃ³n Comu	BalcÃ³n	Barra estilo a	Calentador	Chimenea	CitÃ³fono
1	0	0	1	1	0	0	1
1	0	0	1	1	1	0	0
0	0	0	1	1	1	0	0

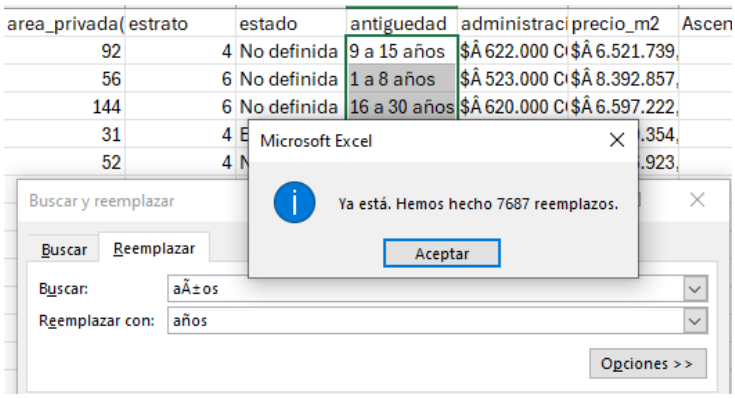
- Para las columnas de área construida y área privada, vamos a generalizar los valores numéricos expuestos en m2, para ello vamos a eliminar “mÂ²” y vamos a especificar en el nombre de la variable que estos valores se encuentran en m2

area_construida	area_privada
92 mÂ²	92 mÂ²
56 mÂ²	56 mÂ²
144 mÂ²	144 mÂ²
31 mÂ²	31 mÂ²
52 mÂ²	52 mÂ²
150 mÂ²	150 mÂ²
110 mÂ²	100 mÂ²
53 mÂ²	47 mÂ²
111 mÂ²	0 mÂ²
264 mÂ²	264 mÂ²
97 mÂ²	0 mÂ²
87 mÂ²	82 mÂ²



- Para la variable antigüedad corregiremos los errores tipográficos para ciertos datos, por ejemplo “aÃ±os” por “años”

antiguedad
9 a 15 aÃ±os
1 a 8 aÃ±os
16 a 30 aÃ±o
menor a 1 aÃ
1 a 8 aÃ±os
mÃ¡s de 30 a
16 a 30 aÃ±o
9 a 15 aÃ±os



antiguedad
9 a 15 años
1 a 8 años
16 a 30 años
menor a 1 año
1 a 8 años
mas de 30 años
16 a 30 años
9 a 15 años
16 a 30 años
mas de 30 años

Datos antigüedad formateados

- En las variables administración y precio m2 se va a eliminar los caracteres “$Â” y todos aquellos caracteres que pueden afectar en la lectura de datos

administracion	precio_m2
$Â 622.000 COP	$Â 6.521.739,13*mÂ²
$Â 523.000 COP	$Â 8.392.857,14*mÂ²
$Â 620.000 COP	$Â 6.597.222,22*mÂ²
$Â 130.000 COP	$Â 7.419.354,84*mÂ²
$Â 219.000 COP	$Â 5.576.923,08*mÂ²
$Â 872.000 COP	$Â 6.533.333,33*mÂ²

administracion(COP)	precio_m2
622.000	6.521.739,13
523.000	8.392.857,14
620.000	6.597.222,22
130.000	7.419.354,84
219.000	5.576.923,08
872.000	6.533.333,33
135.000	4.181.818,18
125.000	3.207.547,17
No definida	3.873.873,87

- Por último, se verifican los datos de la variable ubicación y se corrigen los errores tipográficos

ubicacion
Centro Internacional
Calleja Baja
Cerros de Suba
Mazuren
El plan
La Cabrera
Ciudad jardin sur
Ub. industrial las delicias
Pradera Norte
Nicolas de federman
Pontevedra
Los cedros oriental
Cedro Golf
Sosiego sur
Normandía
Contador
Quinta Paredes
Cedro Golf
La Cabrera
Los Cerezos

CONCLUSIONES

La verificación de la calidad de los datos es esencial para asegurar que el análisis posterior sea preciso y significativo. Al seguir estos pasos, se garantiza que los datos estén limpios y listos para un análisis más profundo, proporcionando así una base sólida para la toma de decisiones basada en datos.