

Primer paso:

Se realizo la respectiva acomodación de los datos seleccionándolos todos.

Ingresamos al apartado de datos y se selecciono texto en columnas.

,area_construida,area_privada,estrato,estado,antiguedad,administracion,precio_m

Asistente para convertir texto en columnas - paso 1 de 3

?

X

El asistente estima que sus datos son Ancho fijo.

Si esto es correcto, elija Siguiente, o bien elija el tipo de datos que mejor los describa.

Tipo de los datos originales

Elija el tipo de archivo que describa los datos con mayor precisión:

☒ Delimitados - Caracteres como comas o tabulaciones separan campos.

☐ De ancho fijo - Los campos están alineados en columnas con espacios entre uno y otro.

Vista previa de los datos seleccionados:

1	habitaciones,baños,parqueaderos,area_construida,area_privada,e	^
2		
3		
4		
5		

<

III

>

Cancelar

< Atrás

Siguiente >

Finalizar

Se selecciono el campo de delimitados.

Asistente para convertir texto en columnas - paso 3 de 3

Esta pantalla permite seleccionar cada columna y establecer el formato de los datos.

Formato de los datos en columnas

☒ General
☐ Texto
☐ Fecha: DMA
☐ No importar columna (saltar)

'General' convierte los valores numéricos en números, los valores de fechas en fechas y todos los demás valores en texto.

Avanzadas...

Destino: \$A\$1

Vista previa de los datos

General	General	General	General	General	General
habitaciones	baños	parqueaderos	area_construida	area_privada	est...

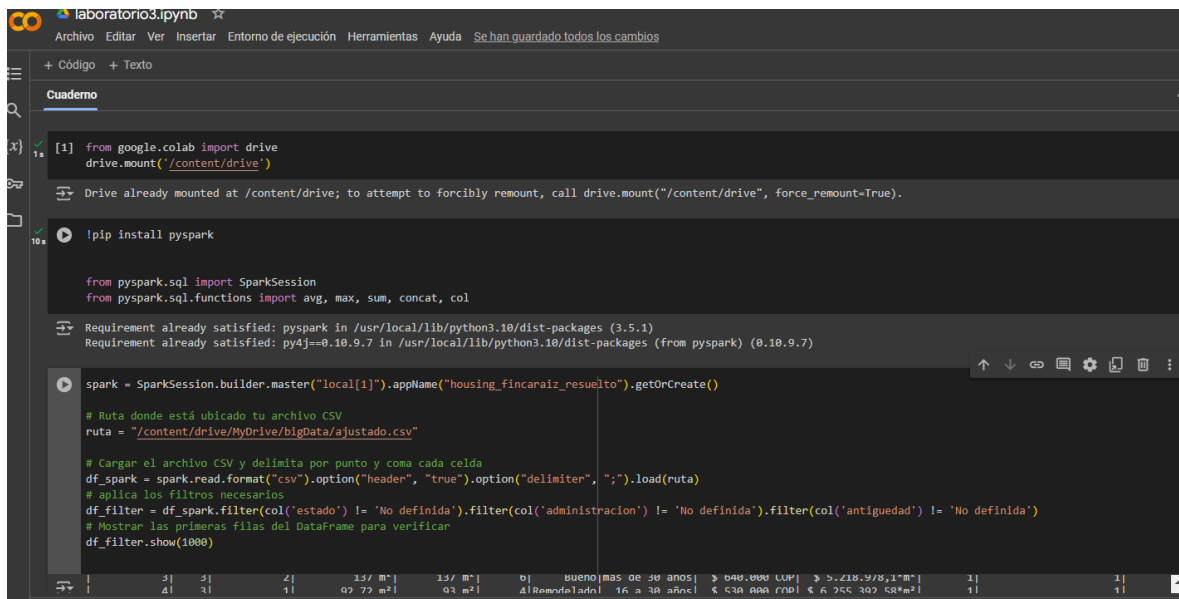
Cancelar < Atrás Siguiente > Finalizar

Se deja todos lo campos en general, y se presiona finalizar.

#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
289	3	3	2 135 m²	135 m²	6 No definida	16 a 30 años	SA 907.000 C	SA 5.918.518	1	1	1	1	0	0	0	0	0	1	1	1	0	1	0
290	3	3	2 106 m²	106 m²	4 Excelente	1 a 8 años	SA 355.000 C	SA 7.075.471	1	1	1	1	0	0	0	0	0	1	1	1	0	1	0
291	3	2	2 132 m²	132 m²	6 Excelente	más de 30 a	SA 659.000 C	SA 5.681.818	1	1	1	1	1	0	0	0	0	1	1	1	1	1	0
292	1	2	1 80 m²	80 m²	5 Bueno	16 a 30 años	SA 600.000 C	SA 6.250.000	1	0	1	0	0	0	0	0	0	0	1	0	1	1	0
293	2	2	1 62 m²	62 m²	4 Bueno	1 a 8 años	SA 257.000 C	SA 5.903.225	0	0	0	1	0	1	1	1	1	0	0	0	0	0	0
294	4	6	4 330 m²	330 m²	6 Bueno	1 a 8 años	SA 1.090.000 C	SA 8.181.818	1	1	1	0	0	0	0	0	1	0	0	0	1	1	0
295	3	2	1 73 m²	73 m²	3 No definida	1 a 8 años	SA 365.000 C	SA 4.931.506	1	0	1	1	0	0	0	0	0	0	1	0	0	1	0
296	3	4	2 151 m²	0 m²	6 Excelente	9 a 15 años	SA 900.000 C	SA 7.748.344	1	0	0	1	0	1	1	1	1	0	1	0	0	0	0
297	3	5	3 220 m²	220 m²	6 Bueno	16 a 30 años	SA 1.100.000 C	SA 8.181.818	1	0	0	0	0	0	0	0	1	0	1	0	1	0	0
298	1	1	0 31 m²	31 m²	3 No definida	menor a 1 a	SA 142.000 C	SA 7.096.774	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0
299	3	4	2 189 m²	189 m²	6 Bueno	16 a 30 años	SA 1.000.000 C	SA 8.465.608	1	1	1	1	0	0	0	1	0	0	0	0	1	1	0
300	2	2	1 75 m²	0 m²	5 Bueno	más de 30 a	SA 650.000 C	SA 5.333.333	1	0	0	1	1	0	1	0	0	0	1	1	0	0	0
301	1	2	1 80 m²	80 m²	5 Bueno	16 a 30 años	SA 600.000 C	SA 6.250.000	1	0	1	0	0	0	1	0	0	1	0	1	1	0	0
302	3	2	0 76 m²	74 m²	3 No definida	16 a 30 años	SA 182.000 C	SA 9.453.846	0	1	0	1	0	1	0	1	1	0	1	0	1	1	0
303	3	4	4 168 m²	168 m²	5 No definida	16 a 30 años	No definida	SA 8.273.809	1	0	1	1	0	1	0	1	0	0	1	1	0	0	0
304	3	2	0 57 m²	0 m²	2 Bueno	9 a 15 años	No definida	SA 3.245.614	1	1	1	1	0	0	0	0	0	1	1	0	1	1	0
305	2	3	2 120 m²	120 m²	6 Bueno	9 a 15 años	SA 730.000 C	SA 10.250.00	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0
306	4	5	3 257 m²	297 m²	6 No definida	16 a 30 años	SA 1.300.000 C	SA 7.575.757	1	1	1	1	0	0	0	0	0	0	1	1	1	0	0
307	3	3	1 96,86 m²	94 m²	4 Excelente	9 a 15 años	SA 286.000 C	SA 5.337.882	1	1	1	1	0	0	0	0	1	0	1	1	1	0	0
308	2	3	2 82 m²	82 m²	6 Bueno	1 a 8 años	SA 900.000 C	SA 8.048.780	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0
309	3	4	2 190 m²	0 m²	6 Excelente	menor a 1 a	SA 661.000 C	SA 7.789.473	1	1	1	1	0	0	0	1	1	0	1	1	1	0	1
310	3	4	2 144 m²	144 m²	6 No definida	16 a 30 años	SA 650.000 C	SA 6.180.555	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0
311	2	1	0 45 m²	39 m²	3 No definida	menor a 1 a	SA 116.000 C	SA 4.111.111	1	1	1	1	0	0	0	0	0	0	1	0	1	1	0
312	2	3	1 75 m²	75 m²	5 No definida	1 a 8 años	SA 530.000 C	SA 6.640.000	1	1	1	1	0	0	0	0	1	1	1	0	1	1	0
313	1	1	1 49 m²	49 m²	6 No definida	16 a 30 años	SA 273.000 C	SA 7.244.897	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0
314	4	2	0 80 m²	80 m²	2 No definida	9 a 15 años	SA 140.000 C	SA 2.625.000	0	1	0	1	0	1	0	1	0	0	1	0	0	1	0
315	3	2	2 132 m²	132 m²	6 Excelente	más de 30 a	SA 659.000 C	SA 5.681.818	1	1	1	1	0	0	0	0	0	1	1	1	1	0	0
316	3	1	0 48 m²	48 m²	1 No definida	No definida	SA 58.300 C	SA 2.666.666	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0
317	3	3	4 352 m²	352 m²	6 Bueno	más de 30 a	SA 2.800.000 C	SA 6.761.363	0	1	0	0	0	0	0	0	1	0	1	1	1	0	0
318	3	2	0 55 m²	52 m²	3 Excelente	menor a 1 a	No definida	SA 4.727.272	0	1	0	0	0	0	0	0	0	1	0	1	1	1	1
319	2	4	2 119 m²	119 m²	6 No definida	más de 30 a	SA 740.000 C	SA 4.789.915	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
320	2	2	1 77 m²	77 m²	4 Excelente	9 a 15 años	SA 798.000 C	SA 6.626.571	1	1	0	1	0	0	0	1	1	0	1	1	1	0	0
321	11	8	2 590 m²	0 m²	6 No definida	más de 30 a	SA 10.000 C	SA 2.883.355	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
322	1	1	1 48 m²	0 m²	4 Bueno	1 a 8 años	SA 256.000 C	SA 6.875.000	1	1	0	1	0	1	0	1	0	0	0	0	1	0	0
323	3	2	1 90 m²	90 m²	6 No definida	más de 30 a	SA 447.000 C	SA 5.777.777	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0
324	5	3	2 324 m²	0 m²	6 Bueno	16 a 30 años	SA 330.000 C	SA 4.135.802	0	0	0	1	0	0	1	0	1	0	0	1	1	0	0
325	3	3	1 89 m²	89 m²	6 Excelente	16 a 30 años	SA 680.000 C	SA 5.449.438	1	0	0	0	1	1	0	1	0	0	0	0	1	1	0

Se ajusto el documento respectivamente con los pasos previos.

Nota: solo se ajusta, las filas y las columnas. La limpieza de los datos se hace en python con la librería pyspark



```
laboratorio3.ipynb
Archivo  Editar  Ver  Insertar  Entorno de ejecución  Herramientas  Ayuda  Se han guardado todos los cambios

+ Código + Texto

Cuaderno

[1] from google.colab import drive
drive.mount('/content/drive')

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

!pip install pyspark

from pyspark.sql import SparkSession
from pyspark.sql.functions import avg, max, sum, concat, col

Requirement already satisfied: pyspark in /usr/local/lib/python3.10/dist-packages (3.5.1)
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)

spark = SparkSession.builder.master("local[1]").appName("housing_fincaraiz_resuelto").getOrCreate()

# Ruta donde está ubicado tu archivo CSV
ruta = "/content/drive/MyDrive/bigData/ajustado.csv"

# Cargar el archivo CSV y delimita por punto y coma cada celda
df_spark = spark.read.format("csv").option("header", "true").option("delimiter", ";").load(ruta)
# aplica los filtros necesarios
df_filter = df_spark.filter(col('estado') != 'No definida').filter(col('administracion') != 'No definida').filter(col('antiguedad') != 'No definida')
# Muestran las primeras filas del DataFrame para verificar
df_filter.show(1000)
```

1	3	3	2	13/ m²	13/ m²	0	BUENO	mas de 30 años	3 040.000 COP	3 5.218.9/8,1"m²	1	1
1	21	21	11	02 72 m²	02 m²	41Remodelada	16 a 20 años	5 320 000 COP	5 6 255 202 50"m²	1	1	1

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

RTA: los campos más relevantes son:

Habitaciones, Baños, Parqueaderos, Área construida, Área privada, Estrato, Estado, Antigüedad, Administración, Precio, Precio m2, Ubicación.

¿Qué variables parecen irrelevantes y pueden ser excluidas?

RTA: Los campos irrelevantes son:

Ascensor, Circuito cerrado de TV, Parqueadero Visitantes, Portería / Recepción, Zonas Verdes, Salón Comunal, Balcón, Barra estilo americano, Calentador, Chimenea, Citófono, Cocina Integral, Terraza, Vigilancia, Parques cercanos, Estudio, Patio, Depósito / Bodega, nombre, ubicación.

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

RTA: si hay suficiente información .

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

RTA: si hay suficientes variables para aplicar el método.

¿Está fusionando varias fuentes de datos? Si es así, ¿hay áreas que podrían plantear un problema al fusionar?

RTA: no, se utiliza únicamente una fuente de datos que contiene toda la información necesaria.

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

RTA: se omiten principalmente

¿Cuál es el formato de los datos?

RTA: El formato de los datos es un archivo CSV .

¿Cuál es el método utilizado para capturar los datos?

RTA: Se usa el método de web Scraping para capturar la información de la página de finca raíz.

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

RTA: Tiene 8.429 filas y 31 columnas.

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

RTA: más de una variable se puede aprovechar para resolver 1 o varias preguntas del negocio.

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

RTA: El precio de la administración y el precio del apartamento y textos descriptivos.

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

RTA: No se han usado estadísticas básicas para las variables claves.

¿Es capaz de priorizar las variables relevantes? Si no es así, ¿hay analistas de negocio disponibles para proporcionar más información?

RTA: Se utilizó la librería pyspark para primero limpiar la información y luego llamar a las variables relevantes como Habitaciones, Baños, Parqueaderos, Área construida, Área privada, Estrato, Estado, Antigüedad, Administración, Precio, Precio m2.

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

RTA: Que el CSV tenía información inconsistente e incompleta la cual había campos que decía "No definida" y no era información válida.

¿Qué variables parecen prometedoras para un análisis más profundo?

RTA: Las variables prometedoras podrían ser el Área construida, Precio m2, Antigüedad

¿Sus exploraciones han revelado nuevas características sobre los datos?

RTA: No se revelaron nuevas características de los datos

¿Cómo han cambiado estas exploraciones su hipótesis inicial?

RTA: como no se revelaron nuevas características no se ha cambiado la hipótesis.

¿Considera que debería reformular el alcance del proyecto?

RTA: Por el momento no se considera cambiar el alcance.

¿Esta exploración ha alterado los objetivos?

RTA: Por el momento no se considera cambiar el objetivo

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

RTA: si los Datos Demográficos y Socioeconómicos como ingresos promedio, tasas de desempleo, tasas de crecimiento poblacional, edad promedio de los residentes, etc.

Datos de Listados de Propiedades: Información detallada sobre las propiedades disponibles para la venta o alquiler, incluyendo características como tamaño del lote, número de habitaciones, características especiales (piscina, garaje, etc.), ubicación geográfica precisa, precio

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

RTA: Se ha identificado que en las filas de las columnas de las variables, se encontró el apartado o el texto “No definida” que se podría definir como faltantes o campos que no aportan ninguna información

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

RTA: En el documento CSV se identificó errores que podían causar problemas como lo son el m*2 y las tildes de las palabras o la letra ñ.

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

RTA:

¿Ha realizado una comprobación de plausibilidad de los valores? Tome notas sobre cualquier conflicto aparente (como adolescentes con altos niveles de ingresos).

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

RTA: hay campos en los que las características de las casas son normales pero sus precios totales son demasiados elevados solo por la ubicación en las que están pero no se considere que valgan su valor por lo que se plantea limitar este tipo de apartamentos

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

RTA: no los datos quedan en memoria temporal al ser procesados y limpiados

¿Cada registro contiene el mismo número de campos?

RTA: si todos los campos de cada fila tienen información