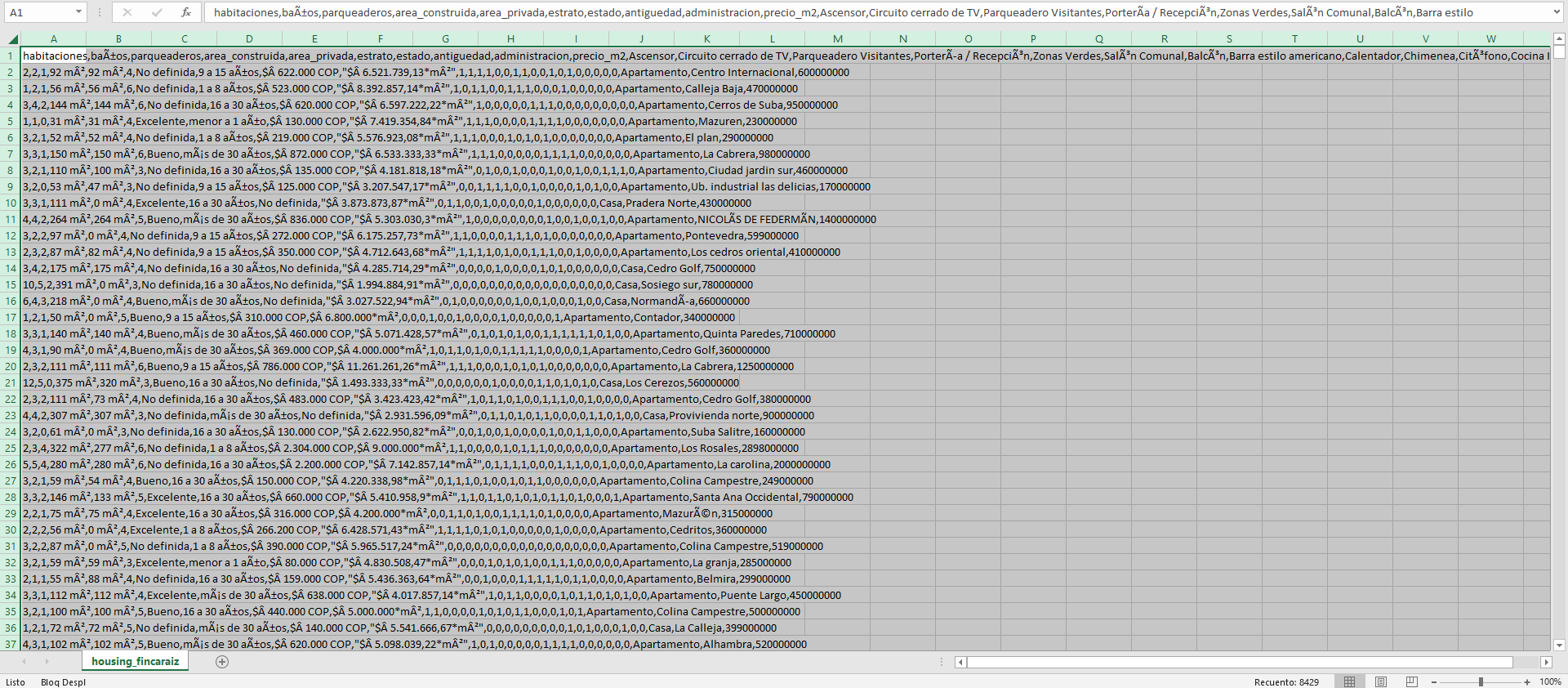
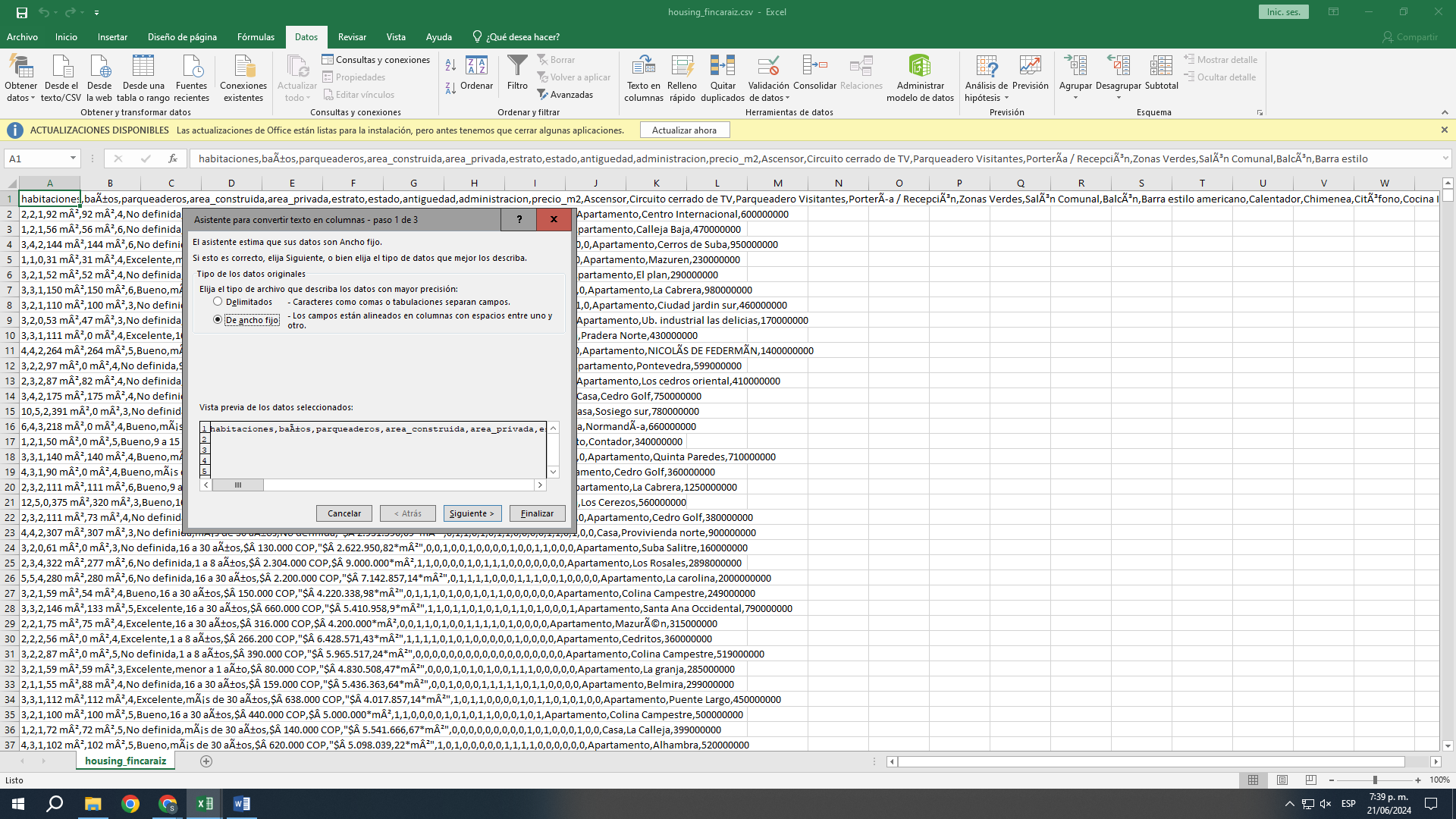
**Documentación laboratorio 3.**

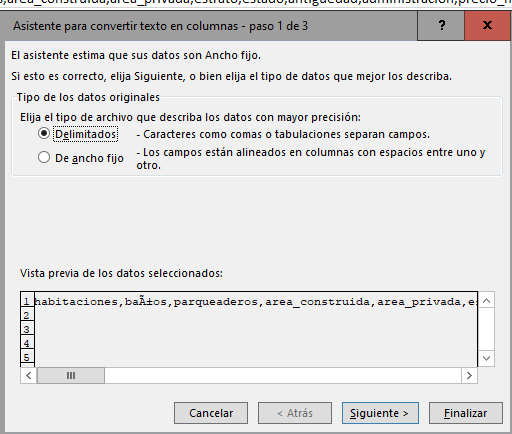
Primer paso:



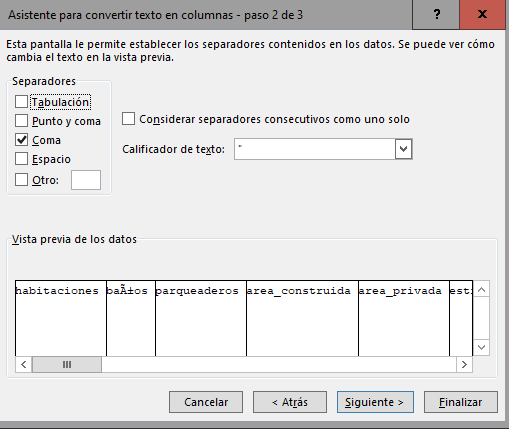
Se realizo la respectiva acomodación de los datos seleccionándolos todos.



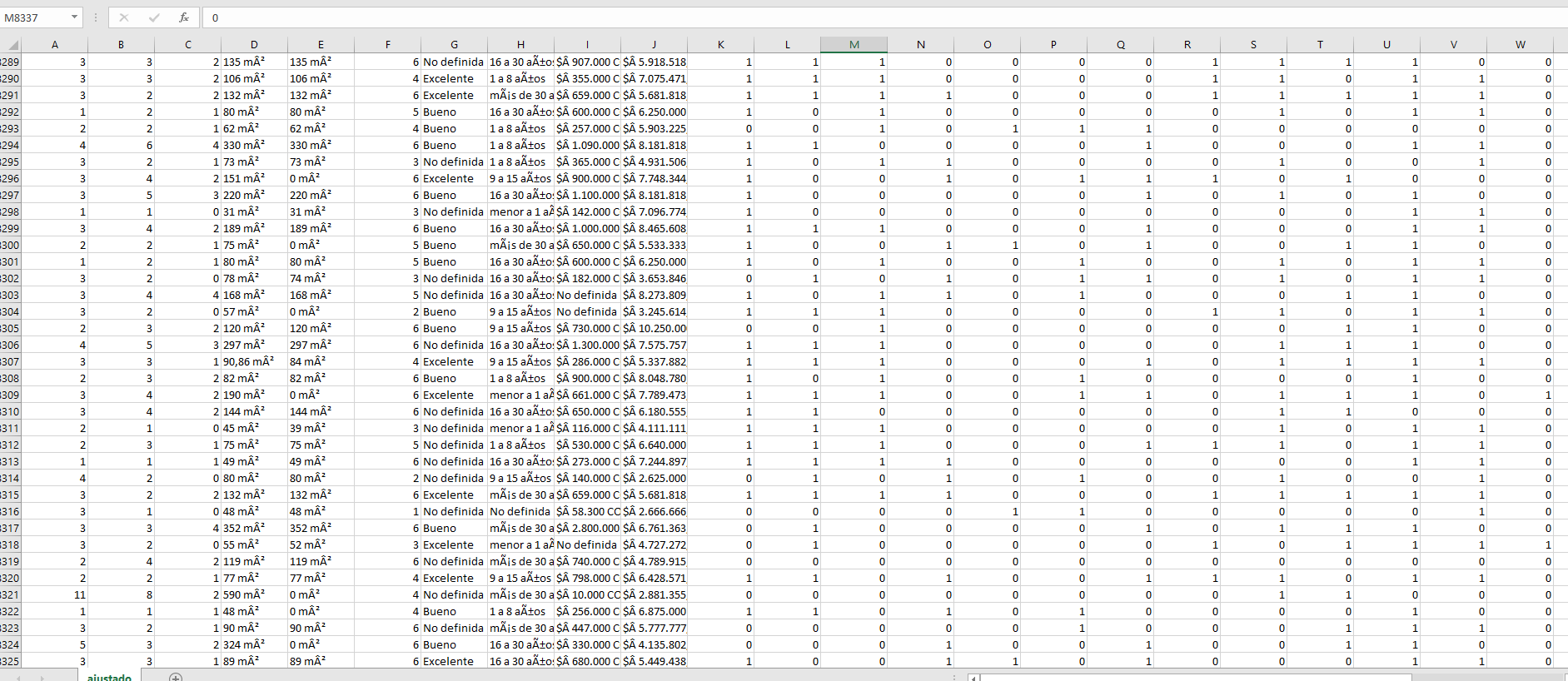
Ingresamos al apartado de datos y se selecciono texto en columnas.

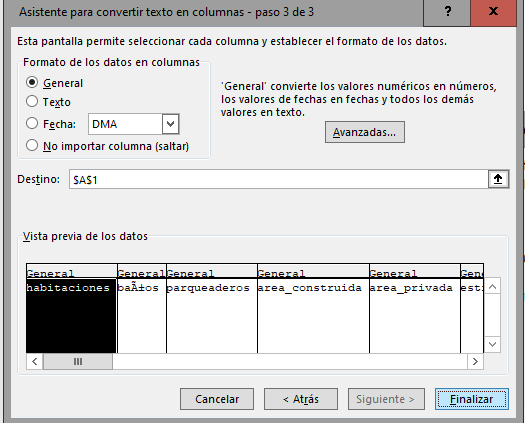


Se selecciono el campo de delimitados.

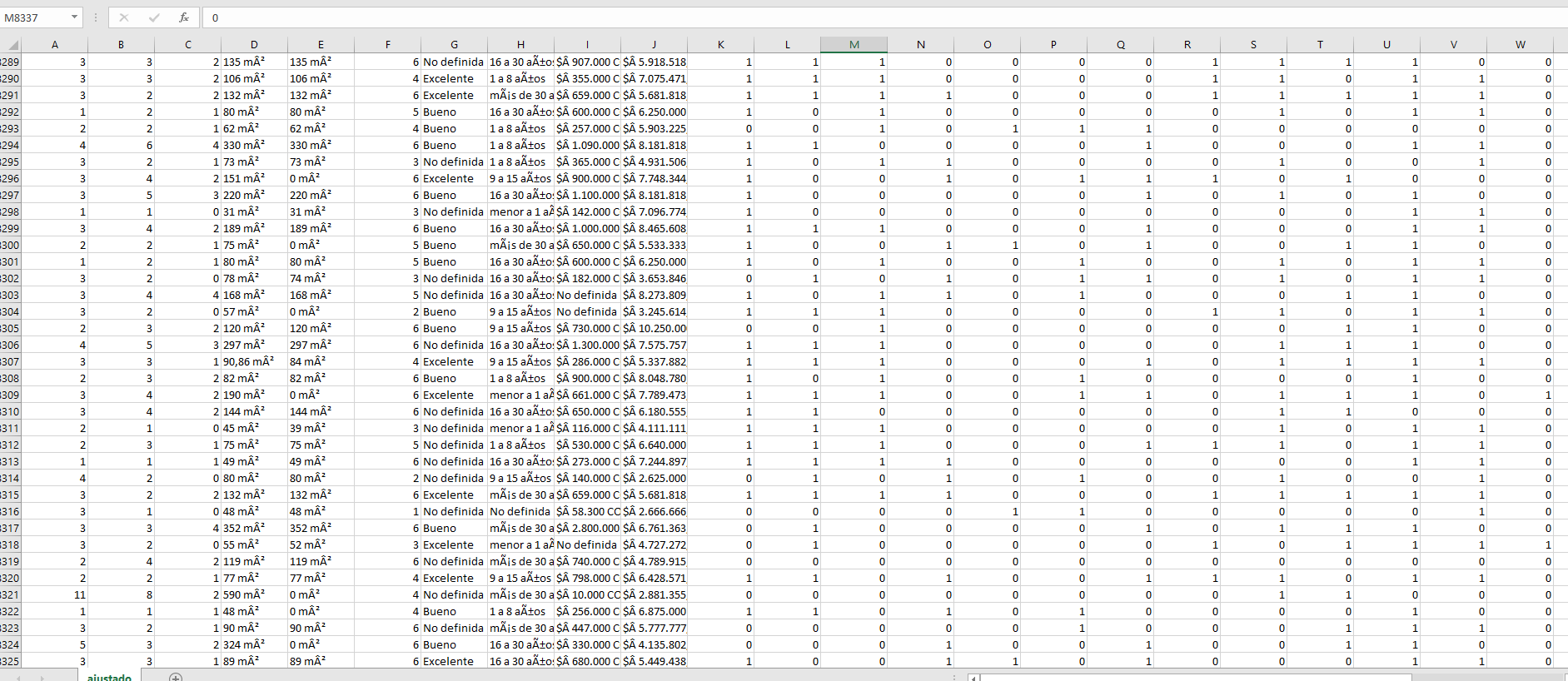


Se selección el campo “coma” para separar las columnas



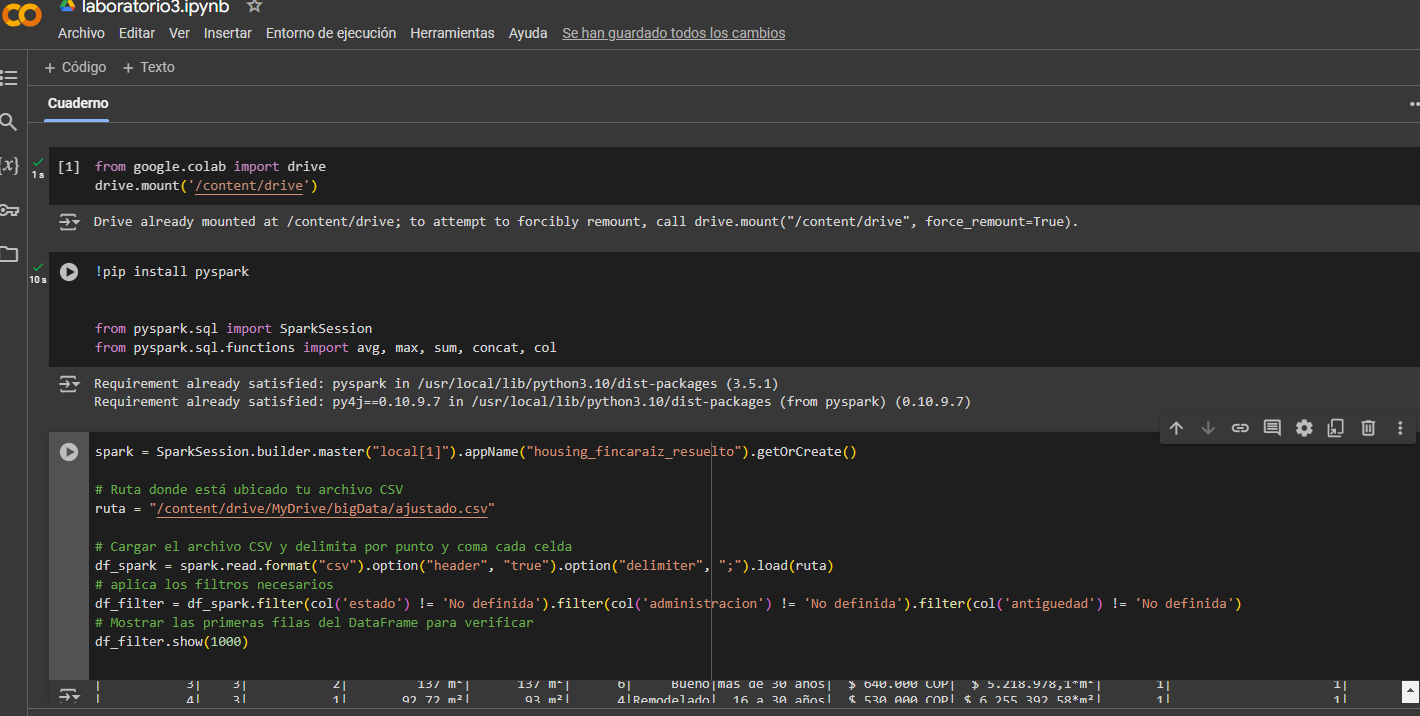


Se deja todos lo campos en general, y se presiona finalizar.



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 pyton con la librería pyspark



**Preguntas**

**¿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: Al momento de limpiar los datos nos dimos cuenta que habian muchos datos indefinidos, otros con valor en 0 y área privado en 0, habían otros que excedían el valor normal del precio por metro cuadrado y valor de administración por lo cual todo esto se podría considerar ruido o inconsistencias lo cual afectará el análisis de los datos, más sin embargo si vale la pena estudiar más a fondo, ya que estos datos también nos pueden arrojar información valiosa, como lo podría ser sector donde más hubo ruido, margen de error en el tipeo de la información entre otros.

**¿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