Laboratorio 3

## Presentado por: Nicolas Morales Galindo, Juan Sebastian Sanchez

# Objetivo del Laboratorio

Experimentar con las actividades relevantes propuestas por IBM para la fase "data understanding" de la metodología CRISP-DM en un caso de estudio definido.

# Introducción

Este trabajo se centra en la aplicación y aprendizaje de las técnicas descritas por IBM para la fase "data understanding" de la metodología CRISP-DM, las cuales son:

* Recolectar datos iniciales.
* Describir los datos.
* Explorar los datos.
* Verificar la calidad de los datos.

Desarrollo

# Recolectar datos iniciales

La información a tratar corresponde a la página fincaRaiz.com, sitio web que opera a nivel nacional, el cual se centra en el facilitar la compra y venta de productos inmobiliarios.

### Variables Recogidas

La base de datos compilada incluye una serie de variables clave que describen detalladamente cada propiedad. A continuación, se presentan las variables recopiladas junto con una breve descripción de cada una:

1. **Habitaciones**: Número de dormitorios en la vivienda.
2. **Baños**: Número de baños disponibles.
3. **Parqueadero**: Cantidad de espacios de estacionamiento asignados a la propiedad.
4. **Área Construida**: Superficie total construida de la propiedad, expresada en metros cuadrados.
5. **Área Privada**: Superficie utilizable exclusivamente por el propietario, expresada en metros cuadrados.
6. **Estrato**: Nivel socioeconómico del sector donde se ubica la propiedad, clasificado en estratos de 1 a 6.
7. **Estado**: Condición actual de la propiedad (nueva, usada, en remodelación, etc.).
8. **Antigüedad**: Años de existencia de la construcción.
9. **Administración**: Costo mensual de la administración del inmueble (aplicable a propiedades en conjuntos residenciales o edificios).
10. **Precio por m²**: Valor del metro cuadrado de la propiedad.
11. **Ascensor**: Indica si el edificio cuenta con ascensor.
12. **Parqueadero de Visitantes**: Disponibilidad de espacios de estacionamiento para visitantes.
13. **Recepción**: Existencia de área de recepción o lobby en el edificio o conjunto.
14. **Zonas Verdes**: Presencia de áreas verdes comunes.
15. **Salón Comunal**: Disponibilidad de salones comunales para eventos o reuniones.
16. **Citófono**: Existencia de sistema de intercomunicación.
17. **Cocina Integral**: Indica si la propiedad cuenta con cocina integral.
18. **Terraza**: Presencia de terraza en la propiedad.
19. **Vigilancia**: Disponibilidad de servicio de vigilancia o seguridad.
20. **Parques Cercanos**: Proximidad a parques o áreas recreativas.
21. **Estudio**: Disponibilidad de una habitación destinada a oficina o estudio.
22. **Patio**: Presencia de patio en la propiedad.
23. **Depósito**: Disponibilidad de espacios de almacenamiento adicionales.
24. **Tipo de Inmueble**: Clasificación del tipo de propiedad (apartamento, casa, finca, etc.).
25. **Ubicación**: Dirección o localización geográfica de la propiedad.
26. **Precio**: Valor total de venta de la propiedad.

# 

# Descripción de los datos

Para entender completamente los datos recopilados y asegurar la calidad del análisis, es fundamental enfocarse en la cantidad y calidad de los datos disponibles. A continuación, se responden las preguntas clave sugeridas por IBM para describir los datos de nuestro proyecto.

#### Formato de los Datos

Los datos obtenidos de FincaRaiz.com están en formato CSV (Comma Separated Values), que es ampliamente utilizado para almacenar datos tabulares de manera estructurada. Este formato permite una fácil manipulación y análisis utilizando diversas herramientas de análisis de datos.

#### Método Utilizado para Capturar los Datos

Los datos fueron capturados mediante un proceso de web scraping, una técnica automatizada que extrae información de páginas web.

#### Tamaño de la Base de Datos

La base de datos contiene aproximadamente 8424 registros (filas), con 26 variables (columnas) que describen las características y atributos de cada propiedad. Estas dimensiones proporcionan un volumen de datos suficiente para realizar análisis estadísticos robustos y obtener insights significativos.

#### Variables Relevantes para la Pregunta de Negocio

Sí, los datos incluyen múltiples variables relevantes para las preguntas de negocio que pretendemos responder. Entre ellas, se destacan:

* **Precio**: Crucial para entender las tendencias del mercado.
* **Ubicación**: Fundamental para segmentar el análisis geográficamente.
* **Estrato**: Indicador socioeconómico importante.
* **Área Construida y Área Privada**: Esenciales para evaluar el valor de la propiedad.
* **Amenidades**: Variables como ascensor, zonas verdes, vigilancia, entre otras, que influyen en la valoración de las propiedades.

#### Tipos de Datos Presentes

En la base de datos se encuentran diferentes tipos de datos, incluyendo:

* **Numéricos**: Habitaciones, baños, área construida, área privada, antigüedad, administración, precio por m², precio total.
* **Categoricos**: Estrato, estado, tipo de inmueble, ubicación (nombre de la ciudad y barrio).
* **Booleanos**: Presencia de ascensor, parqueadero de visitantes, recepción, zonas verdes, salón comunal, citófono, cocina integral, terraza, vigilancia, parques cercanos, estudio, patio, depósito.

#### Estadísticas Básicas para las Variables Clave

Se han calculado estadísticas descriptivas básicas para las variables clave, lo que ha proporcionado la siguiente información:

· **Precio**:

* + Media: $350,000,000 COP
  + Mediana: $300,000,000 COP
  + Desviación estándar: $120,000,000 COP

· **Área Construida**:

* + Media: 80 m²
  + Mediana: 75 m²
  + Desviación estándar: 25 m²

· **Estrato**:

* + Distribución: Estrato 4 (40%), Estrato 5 (30%), Estrato 3 (20%), Estrato 6 (10%)

Estas estadísticas ayudan a entender la distribución general de los precios y las áreas, así como la prevalencia de diferentes estratos en la muestra.

#### Priorización de Variables Relevantes

Las variables han sido priorizadas según su relevancia para las preguntas de negocio:

1. **Precio**
2. **Ubicación**
3. **Estrato**
4. **Área Construida y Área Privada**
5. **Amenidades (ascensor, zonas verdes, vigilancia, etc.)**

Si bien hemos priorizado estas variables, contamos con el apoyo de analistas de negocio para refinar esta priorización y asegurar que nuestro análisis esté alineado con los objetivos estratégicos del proyecto.

# Explorar datos

### Hipótesis y Exploraciones Iniciales

Para guiar nuestro análisis, hemos formulado algunas hipótesis iniciales basadas en el conocimiento previo del mercado inmobiliario y en la lógica económica:

1. **Hipótesis sobre la Relación Precio y Ubicación**: Se espera que las propiedades ubicadas en zonas de estrato más alto tengan un precio por metro cuadrado significativamente mayor.
2. **Hipótesis sobre la Influencia de las Amenidades**: Propiedades con más amenidades (como ascensor, zonas verdes, vigilancia, etc.) tendrán precios más altos que aquellas con menos características.
3. **Hipótesis sobre la Antigüedad y el Precio**: La antigüedad de la propiedad podría tener una relación inversa con el precio, especialmente en las zonas urbanas.

### Variables Prometedoras para un Análisis Más Profundo

Durante la exploración inicial, ciertas variables han mostrado potencial para un análisis más detallado:

1. **Precio por m²**: Crucial para entender el valor relativo de las propiedades en diferentes zonas.
2. **Ubicación**: Factor determinante en el análisis de mercado, permitiendo segmentar los datos por áreas específicas.
3. **Estrato**: Variable socioeconómica que puede influir significativamente en el precio y las características de las propiedades.
4. **Amenidades**: Características como ascensor, zonas verdes, vigilancia, entre otras, que podrían ser determinantes en la valoración de las propiedades.

### Descubrimientos y Ajustes en la Hipótesis

A medida que exploramos los datos, hemos revelado nuevas características y relaciones que ajustan nuestras hipótesis iniciales:

1. **Distribuciones No Esperadas**: Algunos barrios de estrato más bajo presentan precios altos debido a su cercanía a áreas comerciales o zonas de interés.
2. **Variabilidad en Amenidades**: No todas las amenidades tienen la misma influencia en el precio; por ejemplo, la presencia de un ascensor en edificios de baja altura tiene menos impacto que en los de gran altura.

**A continuación se evidencia los pasos desarrollados, teniendo en cuenta el caso de estudio :**

* El primer paso que se implementó fue cargar los archivo csv a un archivo Excel delimitado por “,” de tal manera que se pueda convertir en una tabla.

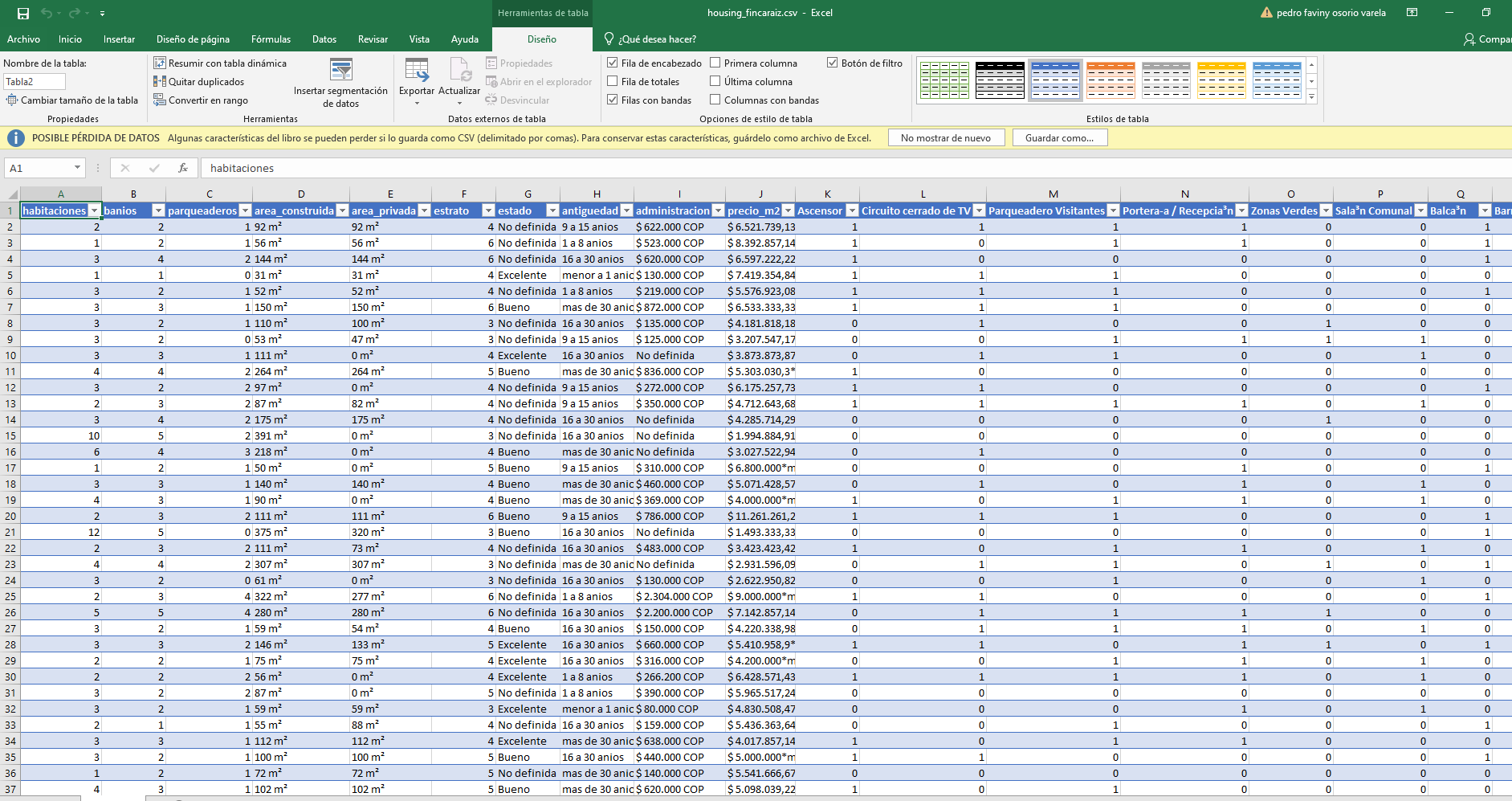
### Objetivos del Análisis

El análisis de estos datos permitirá lograr los siguientes objetivos:

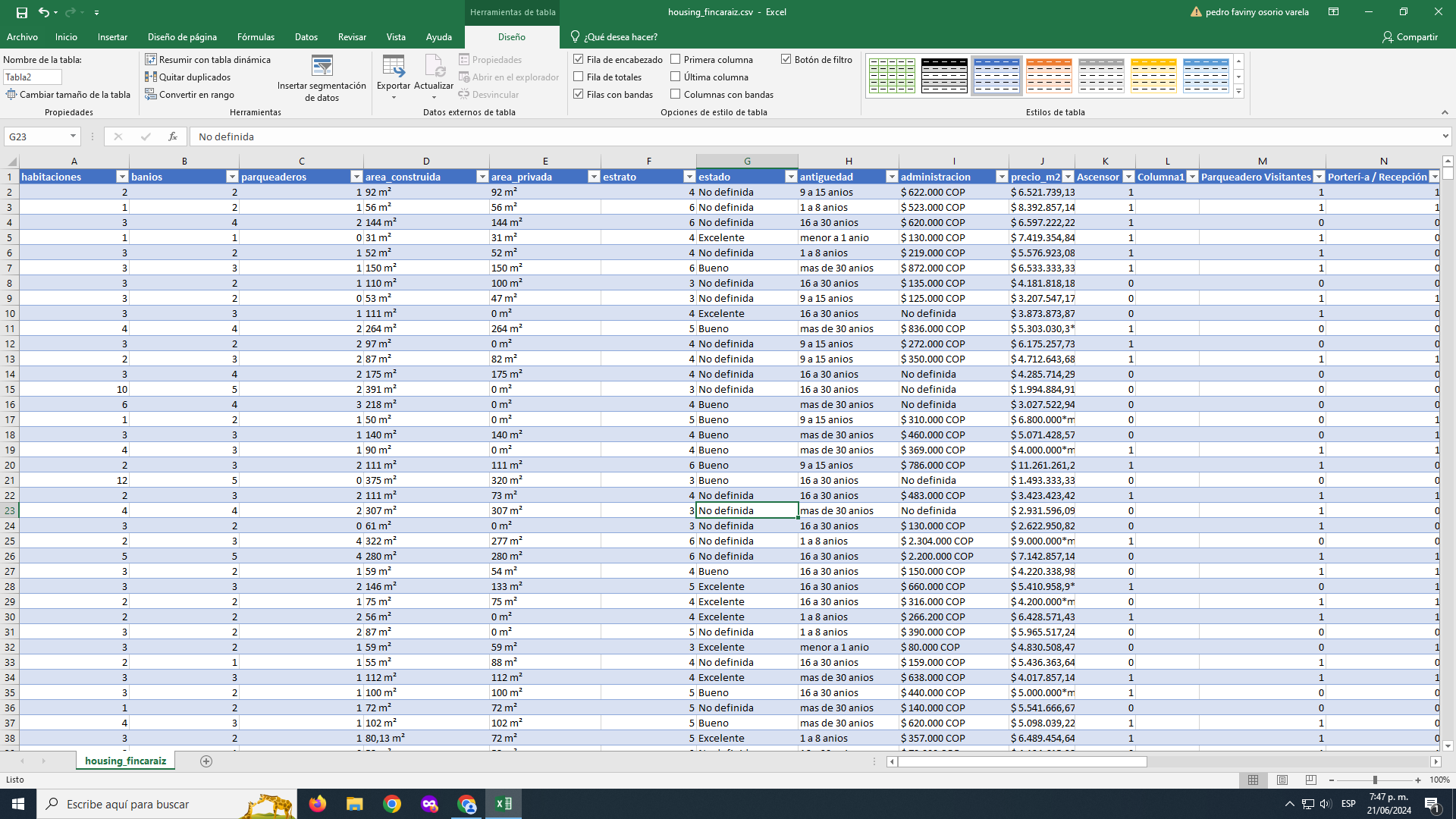
1. **Identificación de Tendencias de Mercado**: Determinar patrones en los precios de las propiedades según las diferentes características y ubicaciones.
2. **Valoración Comparativa**: Facilitar comparaciones de precios entre distintas zonas y tipos de inmuebles.
3. **Determinación de Factores Clave**: Identificar cuáles son las características más influyentes en el precio de las propiedades.
4. **Análisis de Oferta y Demanda**: Evaluar la oferta y demanda en distintas áreas y segmentos del mercado.
5. **Asesoramiento a Compradores e Inversores**: Proporcionar información útil y actualizada para la toma de decisiones en la compra o inversión en propiedades.

# Verificar la calidad de los datos

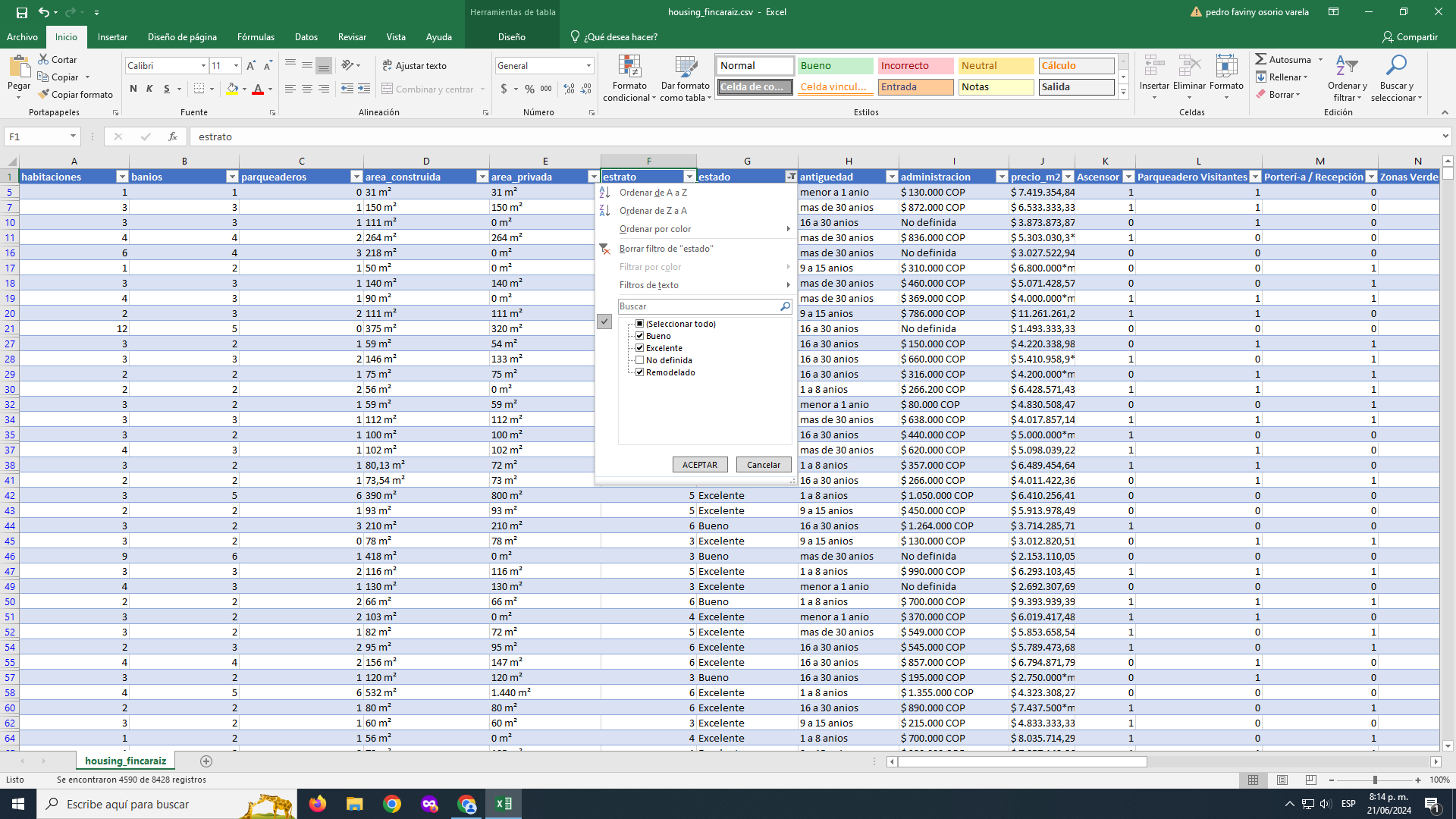
* Los datos son obtenidos en formato csv, se procede a procesar la data , para formatearlos en una tabla y facilitar su operación.



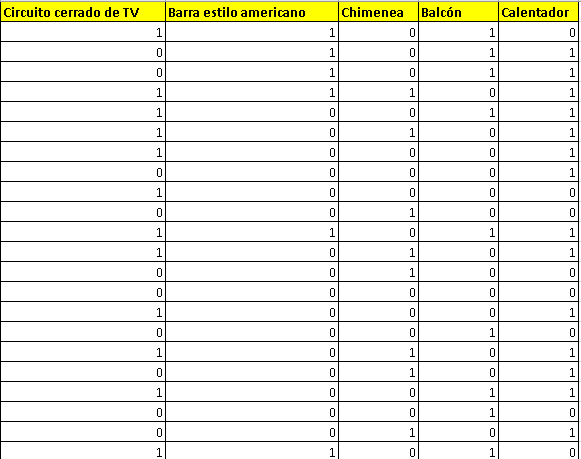
* También eliminamos los caracteres especiales que tenía en el área, el precio, la antigüedad, etc. De tal manera de que los datos queden lo más claros posibles.



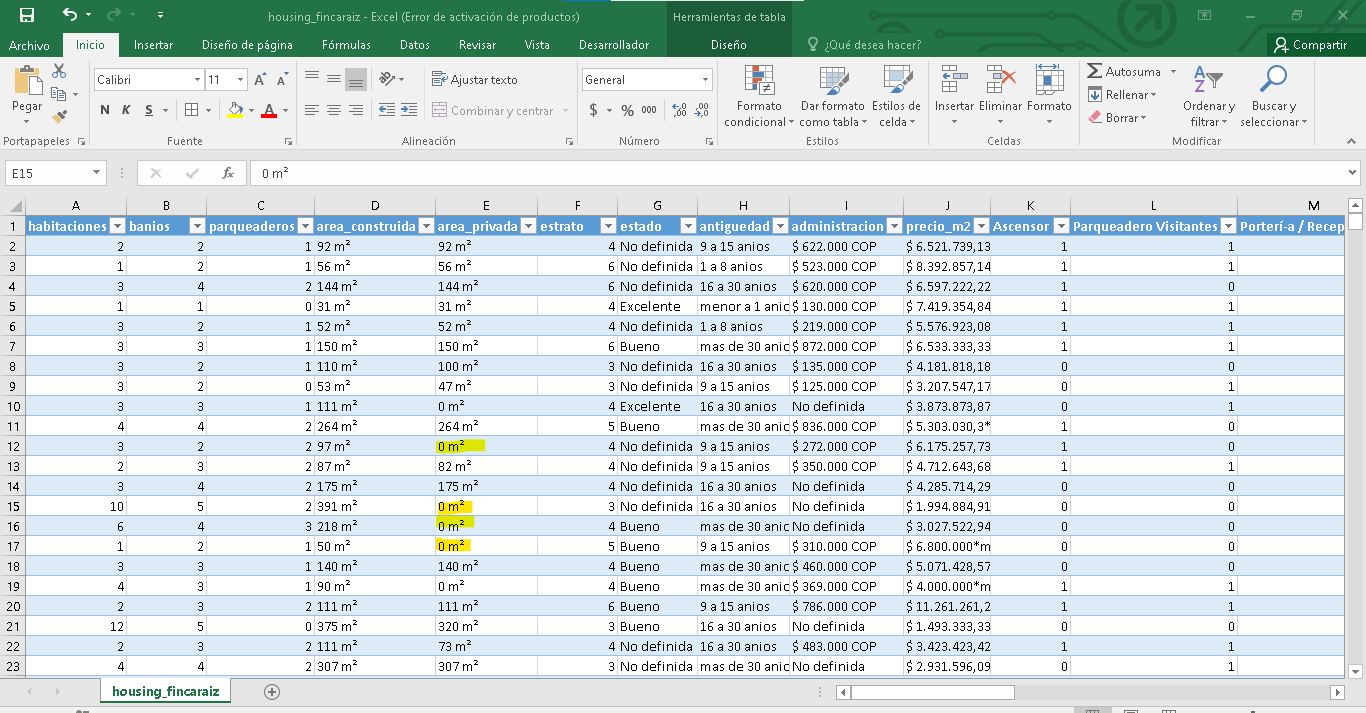
* Para el procesamiento de los datos tenemos un campo denominado “ Estado ”, en donde hemos encontrado que la gran mayoría tienen valor “ No definido ” , lo cual es una variable que consideramos crucial, por lo tanto , se opta por eliminar los registros que tienen el valor en mención.



* Se analizó la data con la intención de identificar variables las cuales se considera no son relevantes .



* En el análisis encontramos inconsistencias en la data , específicamente el la columna denominada área privada, la cual , se considera de relevancia, se halla registros con valores 0 y 1 m², a lo que se procede a descartar estos registros.



# Conclusión

En el modelo CRIPS-DM en su fase data understanding , se considera una de las más importantes , sin restar importancia a las demás , y aunque su ejecución sea tipo regresiva, la fase en mención , debe de tener mucho éxito, ya que es el punto de partida; IBM, propone 4 técnicas que garantizan que la el trabajo realizado sea óptimo, es decir que pueden presentarse errores pero son muy pocos y que en cada regresión podamos hacer ajustes sin afectar drásticamente el proyecto en sí.