**TALLER #3: COMPRENSIÓN DE LOS DATOS**

**Kevin Steven Olivera Caicedo (64178)**

1. **Limpieza de datos**Para llevar a cabo la limpieza de los datos, se utilizó la herramienta de Excel, en donde se siguió los siguientes pasos:  
     
   Se delimitaron los datos por, comas, puntos, doble puntos:  
    - Seleccionar las columnas a ordenar / Pestaña de “Datos”  
    - Herramientas de Datos / “Texto en Columnas”  
    - Limitar los datos por Comas, Puntos, etc.  
    - Finalizar y Guardar Documento  
      
   Los datos se encuentras delimitados, pero con caracteres extraños, de lo cual se debe hacer lo siguiente:  
    - Abrir un nuevo Excel.  
    - Seleccionar la pestaña de “Datos”  
    - En el campo “Obtener y transformar datos” seleccionar “Desde el texto/CSV”  
    - Seleccionar el CSV delimitado  
    - Cargara una ventana de transformación de datos  
    - Hay que cambiar el Origen de los Datos.  
    - Cambiar “Europeo Occidental” por “Unicode (UTF-8)”  
    - Presionar Cargar, y los datos se cargaran de manera correcta.
2. **Datos**Se mostrarán los datos antes de organizar y limpiar.  
     
     
     
     
     
     
     
     
     
     
     
     
     
     
   Después de la limpieza y el orden:  
     
   
3. **Recolectar Datos Iniciales**

* **¿Qué variables (columnas, atributos) de la(s) tabla(s) o base(s) de datos parecen más prometedoras?**
  + En el archivo CSV de "housing fincaraiz": habitaciones, baños, parqueaderos, área construida, área privada, estrato, estado, antigüedad, administración, precio\_m2, 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, precio.
* **¿Qué variables parecen irrelevantes y pueden ser excluidas?**
  + Variables que no se relacionen directamente con la predicción de precios de propiedades o automóviles pueden ser consideradas de poca relevancia. Como "nombre" y "ubicación" podrían ser menos significativas.
* **¿Hay suficientes datos para sacar conclusiones generalizables o hacer predicciones precisas?**
  + La suficiencia de datos dependerá de la cantidad y diversidad de los datos recolectados. Por lo que pueden ser suficientes o demaciados Justos para hacer predicciones generalizables, dependiendo lo que se requiera en concreto, y el caso de uso.
* **¿Hay demasiadas variables para el método de modelado de su elección?**
  + La cantidad de variables es manejable para la mayoría de los métodos de modelado.
* **¿Ha considerado cómo se manejan los valores que faltan en cada uno de sus orígenes de datos?**
  + Sí, se debe identificar y manejar los valores faltantes en el conjunto de datos. Esto puede incluir imputación, eliminación de registros incompletos o el uso de técnicas de aprendizaje automático para predecir valores faltantes.

1. **Describir los Datos**

* **¿Cuál es el formato de los datos?**
  + El conjuntos de datos están en formato CSV.
* **¿Cuál es el método utilizado para capturar los datos?**
  + Los datos de propiedades fueron obtenidos de un archivo CSV preexistente.
* **¿Qué tamaño tiene la base de datos (en número de filas y columnas)?**
  + El tamaño se basa de la siguiente manera: Cantidad de columnas (Variables): 30. Cantidad de Filas:8.428. Datos Totales: 101.148
* **¿Incluyen los datos una o más variables relevantes para la pregunta de negocio?**
  + Sí, el conjunto de datos incluyen múltiples variables relevantes, como precio, ubicación, características varias, etc.
* **¿Qué tipos de datos están presentes (simbólicos, numéricos, etc.)?**
  + Los datos incluyen tanto variables numéricas (habitaciones, baños, precio)
* **¿Ha calculado estadísticas básicas para las variables clave? ¿Qué información le ha proporcionado sobre la cuestión de negocio?**
  + Las estadísticas básicas, como media, mediana, desviación estándar, pueden proporcionar información sobre la distribución de los precios, la antigüedad de las propiedades, etc.
* **¿Es capaz de priorizar las variables relevantes? Si no es así, ¿hay analistas de negocio disponibles para proporcionar más información?**
  + Sí, se pueden priorizar variables relevantes utilizando técnicas de análisis de correlación y pruebas de hipótesis.

1. **Explorar los Datos**

* **¿Qué tipo de hipótesis se ha formado sobre los datos?**
  + La hipótesis podría ser que ciertas características de las propiedades (como ubicación, estrato, área) tienen una relación significativa con el precio.
* **¿Qué variables parecen prometedoras para un análisis más profundo?**
  + Para propiedades: precio, área construida, estrato, antigüedad.
* **¿Sus exploraciones han revelado nuevas características sobre los datos?**
  + La relación entre ciertas comodidades y el precio de las propiedades.
* **¿Cómo han cambiado estas exploraciones su hipótesis inicial?**
  + La exploración podría refinar las hipótesis iniciales al revelar patrones inesperados o variables adicionales que afectan el precio.
* **¿Esta exploración ha alterado los objetivos?**
  + Los objetivos podrían ajustarse para centrarse en las variables que muestran una mayor influencia en los precios.
* **¿Puede identificar subconjuntos particulares de datos para su uso posterior?**
  + Sí, podrían identificarse subconjuntos basados en ubicación, tipo de propiedad, etc.

1. **Verificar la Calidad de los Datos**

* **¿Ha identificado variables faltantes y campos en blanco? Si es así, ¿Hay algún significado detrás de tales valores faltantes?**
  + Sí, la identificación de valores faltantes es crucial. La ausencia de datos puede deberse a errores en la recolección o a la falta de disponibilidad de ciertas características en algunas propiedades, como el “Estado” y la “Administración”.
* **¿Ha explorado las desviaciones para determinar si son "ruido" o fenómenos que vale la pena analizar más a fondo?**
  + Existen información, que conviene analizar a fondo, ya que hay parámetros que no tienen un valor valido, por lo que puede afectar el resultado final, como puede ser el estado de la propiedad.
* **¿Ha realizado una comprobación de plausibilidad de los valores? Tome notas sobre cualquier conflicto aparente (como adolescentes con altos niveles de ingresos).**
  + Es importante verificar la plausibilidad de los datos para evitar conclusiones erróneas basadas en datos incorrectos.
* **¿Ha considerado excluir datos que no tienen impacto en sus hipótesis?**
  + Sí, la exclusión de datos irrelevantes puede simplificar el análisis y mejorar la precisión de los modelos, como puede ser el “Precio en Metros cuadrados”.
* **¿Los datos se almacenan en archivos planos? Si es así, ¿Son los delimitadores coherentes entre los archivos?**
  + Los datos se almacenan en archivos CSV, y es importante asegurarse de que los delimitadores sean coherentes, y/o ordenados correctamente, para su análisis.
* **¿Cada registro contiene el mismo número de campos?**
  + Sí, se cuenta con el mismo número de campos.