# Taller 3-Preprocesamiento de Datos

## Juan José Toro Villegas CC. 1017271652

Inicialmente se importan los datos a través de la URL brindada la cual contiene la base de datos: precios

**Datos Faltantes**

Después de analizar y codificar, vemos que solo las variables longitud y latitud presentan datos faltantes, presentan un total de 2135 cada una y ambas también presentan igual porcentaje de datos faltantes, cada una con 23,11%. Debido a que tenemos muchos datos, se decidió eliminar las filas con al menos un dato faltante.

Vemos que se eliminaron, sin embargo hay algunos datos "n.d.", miremos cuáles son esas variables: Tienen n.d. dos de las columnas, precio y variabilidad.

Para precio, se decide imputar con la variable anterior y para variabilidad se imputaron los datos n.d. de variabilidad con un 0.

**Datos Duplicados**

Después de codificar, vemos que no hay presencia de estos.

**Redefinición**

Se uso la función describe para identificar diferencias entre las medidas descriptivas y posteriormente nos damos cuenta de que dos de las variables numéricas no estaban en el formato correcto, por tanto se convierte la columna 'precio' a tipo numérico al igual que la variable variabilidad.

**Outliers**

Para identificar datos atípicos se conto con la ayuda de histogramas y boxplots de las variables numéricas. Vemos que hay presencia de datos que se salen de los límites del boxplot, sin embargo no hay presencia de datos atípicos, ya que hay gran cantidad entre 7000 y 100000.

**Estandarización y Normalización**

No se requiere estandarización o normalización, ya que la mayoría de las variables son categóricas o nominales, no continuas. La estandarización y normalización son usadas en variables numéricas continuas para permitir comparaciones significativas al igualar sus escalas.

**Categorización**

Se decide agregar una columna que se llame información, esta dira si la información de la fila es completa o no, teniendo en cuenta la imputación pasada y lo q se ha hecho, una fila contara con información incompleta cuando en sus filas haya al menos un valor 0.

Luego se calcula valor máximo, mínimo y media de la variable precios para si categorizar nuevamente, una nueva variable que se llama “tipo de pdto”, este clasifica el producto en 3: muy económico, moderado y alto, según su precio.

**Preguntas (Filtrado)**

**NOTA:** Ver colab para visualizar mejor cada pregunta.

1) Cuales productos se denominan altos o costosos?

2) Productos de precio superior a 9000 pesos

3)Cuales productos se denominan muy económicos?

4) Cual será la frecuencia de cada tipo de producto?

5) Que productos presentan variabilidad negativa y en qué ciudad?

**Herramientas Graficas**

**NOTA:** Ver mejor en el colab