**UNIVERSIDAD CATÓLICA BOLIVIANA”SAN PABLO”  
MAESTRÍA EN CIENCIA DE DATOS V2  
PRIMERA VERSIÓN   
MATERIA MINERÍA DE DATOS 1  
DATASET DE PRACTICA**

**BIMBO**

1. Quitar filas duplicadas
2. Quitar columnas duplicadas
3. Identificar columnas que tienen valores constantes
4. Identifique el porcentaje de valores faltantes en las columnas
5. Analice el caso de los atributos siguientes por medio de la obtención de gráficas: Agencia\_ID, Town, Product\_ID, Cliente\_ID, NombreCliente, Semana, Canal\_ID, Ruta\_SAK, Venta\_uni\_hoy, Venta\_hoy. Cuales decidiría usted de transformar o en su caso quitar, con el propósito de mejorar el dataset.
6. Obtenga la matriz de correlación de las variables numéricas e identifique si algunas se pudieran quitar
7. Discretice Ruta\_SAK, Venta\_hoy, Dev\_uni\_próxima y Dev\_próxima.
8. Obtenga la tabla de relación de frecuencias entre los atributos NombreProducto y State.
9. Aísle los atributos numéricos e identifique outliers. Quite los outliers que tengan rangos muy lejanos.
10. Genere variables dummy utilizando Hot-encoding. Decida en que atributos aplicar este procedimiento, justificándolo.
11. En Ruta\_SAK el valor 9999 es null, cambie este valor por el que considere se adecua más a este campo (i.e. media, mediana, algún otro)
12. El atributo State tiene valores faltantes, cambiar estas marcas con el State que sea más común.
13. Identifique atributos que no serían de utilidad por su nivel de variación.
14. Normalice y estandarice los atributos numéricos.
15. Realice el cálculo con los campos Venta\_uni\_hoy multiplicado por Venta\_hoy, para lo cual cree un campo con el resultado.
16. Obtenga por campo State y Product\_ID, el total en monto agrupado del resultado del campo de la respuesta a la pregunta 15.
17. Obtenga los nombres de clientes y los nombres de productos de State “BAJA CALIFORNIA NORTE” y Canal\_ID “4”.