Trabajo Práctico N°1: Normalización de Datos

Vidman, Xavier Harry

24 de junio de 2022

1. Introducción

El presente informe tiene como objetivo mostrar un diagnóstico y normalización de los datos recibidos a través de las siguientes tablas:

- Clientes.csv
- Compra.csv
- Gasto.csv
- Localidades.csv
- Proveedores.csv
- Sucursales.csv
- Venta.csv

En este diagnóstico se analizarán posibles incongruencias en los datos para poder definir si los datos son de mala o buena calidad. Después, se procederá a realizar una limpieza y normalización de cada tabla.

2. Diagnóstico y normalización de cada tabla.

2.1. Tabla: cliente

Columna	Descripción
Id_cliente	ID del cliente
Provincia	Provincia del cliente
Nombre_apellido	Nombre y apellido del cliente
Domicilio	Domicilio del cliente
Telefono	Telefono del cliente
Edad	Edad del cliente
Localidad	Localidad del cliente

Figura 1: Tabla de clientes

La tabla cliente fue normalizada de la siguiente forma:

- Se realiza un algoritmo que detecte todas las tablas que contengan el string “Cliente”, para convertirlas en una misma tabla.
- Se borran posibles datos duplicados.
- Se normalizan los nombres de las columnas.
- Se completan valores nulos con el valor de “Sin dato”.
- Se normalizan los datos strings acomodando mayúsculas y minúsculas.
- Se cambian valores strings “Nan” por “Sin dato”.
- Se eliminan columnas instrascendentes para un análisis ejecutivo (columnas: latitud, longitud, col10).

2.2. Tabla: compra

Columna	Descripción
Id_compra	ID de la compra
Fecha	Fecha año-mes-día
Fecha_año	año
Fecha_mes	mes
Fecha_periodo	año-mes
Id_producto	ID del producto
Cantidad	Cantidad de productos
Precio	Precio de la compra
Id_proveedor	ID del proveedor

Figura 2: Diccionario de datos de la tabla Compras

El problema que presenta la tabla de compras es que tiene valores nulos y valores del tipo outlier en la columna “Precio”. Estos valores los denominaremos no confiables y los podemos visualizar en la siguiente figura:

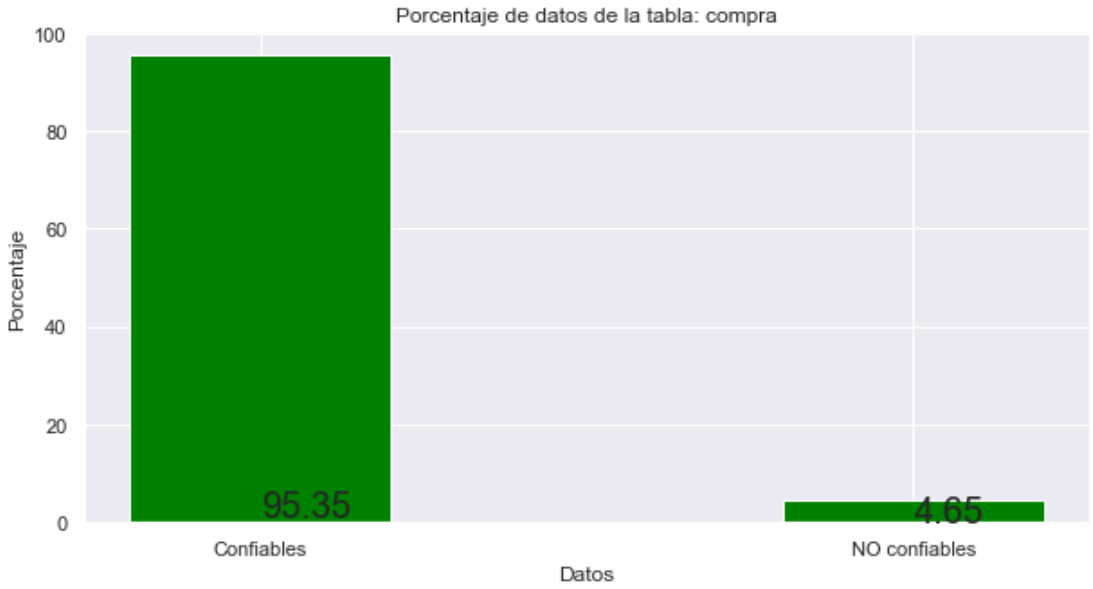

Figura 3: Datos confiables y no confiables de la tabla compras.

Debido a que los datos no confiables son menores a un 5 %, se toma la decisión de descartarlos y solo realizar análisis a futuro con datos confiables.

2.3. Tabla: venta

Columna	Descripción
Id_venta	ID de la venta
Fecha	Fecha año-mes-día
Fecha_entrega	Fecha de entrega
Id_canal	ID del canal de venta
Id_cliente	ID del cliente
Id_sucursal	ID de la sucursal
Id_empleado	ID del empleado
Id_producto	ID del producto
Precio	Precio de la venta
Cantidad	Cantidad de productos

Figura 4: Tabla de ventas.

El problema que presenta la tabla de ventas es que tiene valores nulos y valores del tipo outlier en la columna “Precio”. Estos valores los denominaremos no confiables y los podemos visualizar en la siguiente figura:

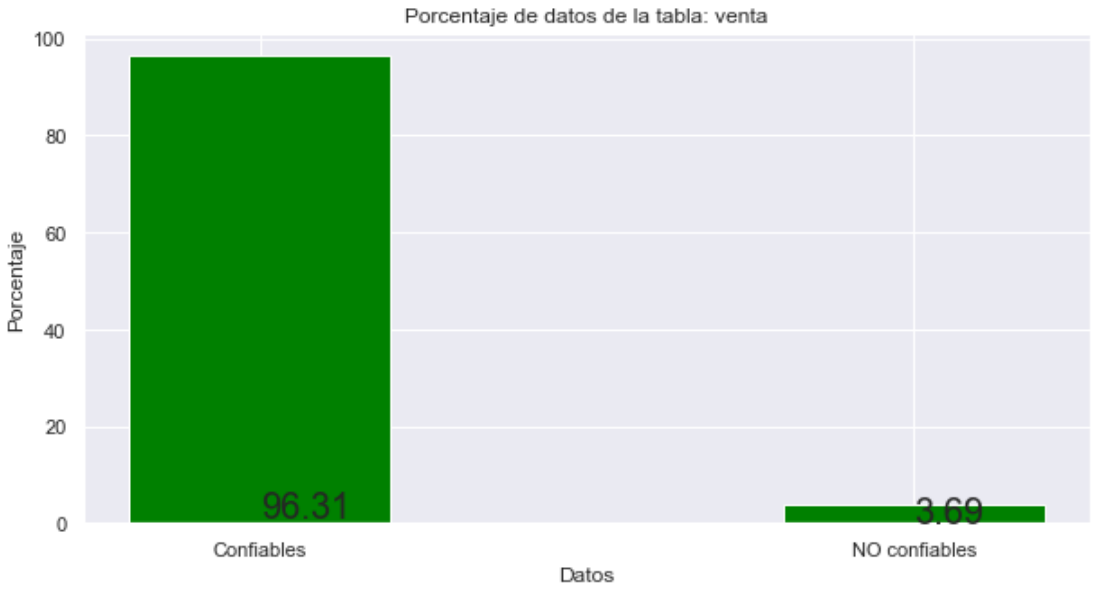

Figura 5: Datos confiables y no confiables de la tabla ventas.

Debido a que los datos no confiables son menores a un 5 %, se toma la decisión de descartarlos y solo realizar análisis a futuro con datos confiables.

2.4. Tabla: gasto

Columna	Descripción
Id_gasto	ID del gasto
Id_sucursal	ID de la sucursal
Id_tipo_gasto	ID del tipo de gasto
Fecha	Fecha año-mes-día
Monto	Monto del gasto

Figura 6: Tabla de gastos

La tabla gastos no presenta valores nulos ni duplicados. En la siguiente figura, se puede ver que tampoco presenta una anomalía en los datos de la columna “Monto” a lo largo del tiempo.

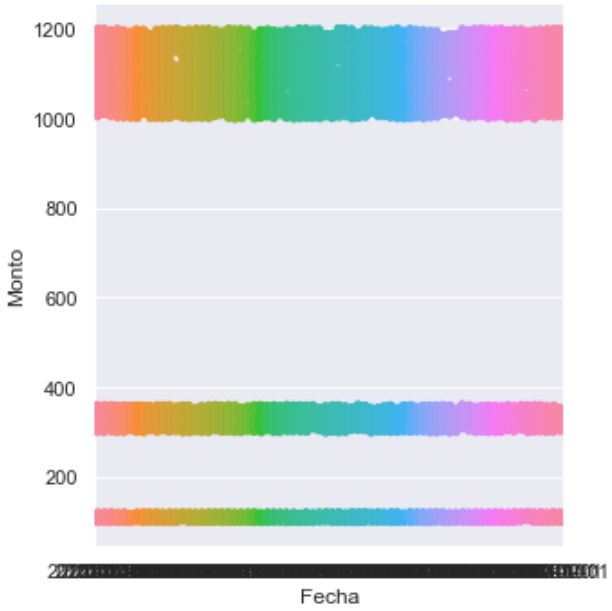

Figura 7: Datos de los Montos de gasto a lo largo del tiempo

2.5. Tabla: sucursal

Columna	Descripción
Id_sucursal	ID de la sucursal
Sucursal	Sucursal
Dirección	Dirección
Localidad	Localidad
Provincia	Provincia

Figura 8: Tabla de sucursales.

La tabla sucursal fue normalizada de la siguiente forma:

- Se borran posibles datos duplicados.
- Se normalizan los nombres de las columnas.
- Se eliminan columnas instrascendentes para el análisis ejecutivo (latitud y longitud).
- Se normalizan los datos strings acomodando mayúsculas y minúsculas.
- Se normalizan las provincias y localidades: a todas las variantes de un mismo lugar se le asignan un mismo nombre.

2.6. Tabla: localidad

Columna	Descripción
Categoría	Categoría de la localidad
Id_departamento	ID del departamento
Departamento	Departamento
Fuente	Fuente
Id_localidad	ID de la localidad
Id_localidad_censal	ID de la localidad censal
Localidad_censal	Localidad censal
Id_municipio	ID del municipio
Municipio	Municipio
Localidad	Localidad
Id_provincia	ID de la provincia
Provincia	Provincia

Figura 9: Tabla de localidades.

La tabla localidad fue normalizada de la siguiente forma:

- Se borran posibles datos duplicados.
- Se normalizan los nombres de las columnas.
- Se completan valores nulos con el valor de “Sin dato”.
- Se normalizan los datos strings acomodando mayúsculas y minúsculas.

2.7. Tabla: proveedor

Columna	Descripción
Id_proveedor	ID del proveedor
Nombre	Nombre
Domicilio	Domicilio
Ciudad	Ciudad
Provincia	Provincia
Pais	Pais
Departamento	Departamento

Figura 10: Tabla de proveedores.

La tabla proveedor fue normalizada de la siguiente forma:

- Se borran posibles datos duplicados.
- Se normalizan los nombres de las columnas.
- Se completan valores nulos con el valor de “Sin dato”.
- Se normalizan los datos strings acomodando mayúsculas y minúsculas.