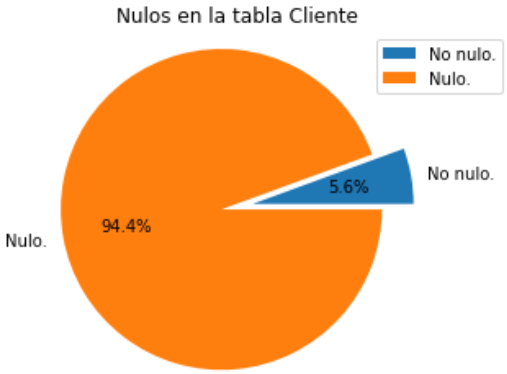
INFORME DE CALIDAD DE DATO

DISCLAIMER: Este reporte y todo mi proyecto fue hecho con base en los datasets que nos proporcionaron el día lunes, estos fueron segmentados usando el mismo criterio que usaron ustedes para crear la segunda camada de datasets. Traté de usar los provistos por ustedes, pero por mucho que trataban de arreglar, aún tenían problemas generados por la cuestión de los separadores.

TABLAS DIMENSIONALES:

TABLA CLIENTES:

TÍTULO	DESCRIPCIÓN	TIPO DE DATO	PROBLEMAS	SOLUCIONES
ID	Índice asociado al cliente.	Entero.	Ninguno.	Ninguna.
Provincia	Nombre de la provincia en la que reside.	String.	25 registros nulos. Comparte registros nulos con 'Localidad'.	Imputar los nulos usando el maestro de localidades.
Nombre_y_Apellido	Nombre y apellido del cliente.	String.	19 registros nulos. Heterogénea en formatos. Algunas entradas están incompletas.	Imputar nulos por "Sin dato". Normalizar el formato de la columna.
Domicilio	Dirección en la que reside.	String.	39 registros nulos. Caracteres raros sin importar el encoding que se use.	Imputar nulos por "Sin dato".
Telefono	Número telefónico asociado al cliente.	String.	31 registros nulos. Campos multivaluados. Heterogénea en formatos.	Imputar nulos por "Sin dato". Dejar solo 1 dato en las columnas multivaluadas. Normalizar la columna.
Edad	Edad.	Entero.	Ninguno.	Ninguna.
Localidad	Nombre de la localidad en la que reside.	String.	25 registros nulos. Comparte registros nulos con 'Provincia'.	Imputar los nulos usando el maestro de localidades.
X	Coordenadas de longitud del domicilio.	Decimal.	53 registros nulos. 40 registros positivos. Coma decimal.	Imputar nulos. Convertir los positivos. Normalizar punto decimal.
Y	Coordenadas de latitud del domicilio.	Decimal.	50 registros nulos. 40 registros positivos. Coma decimal.	Imputar nulos. Convertir los positivos. Normalizar punto decimal.
col10	Columna vacía.	NaN.	Toda la columna está vacía.	Descartar la columna.



Podemos observar que el 5.6% (133/2384) de registros de la tabla clientes posee al menos 1 campo nulo LUEGO de descartar la columna 'col10', la cual es enteramente nula.

TABLA PROVEEDORES:

TÍTULO	DESCRIPCIÓN	TIPO DE DATO	PROBLEMAS	SOLUCIONES
IDProveedor	Índice asociado al proveedor.	Entero.	Ninguno.	Ninguna.
Nombre	Nombre de la empresa.	String.	2 registros nulos.	Imputar los valores faltantes por "Sin dato".
Address	Dirección de la sucursal.	String.	Ninguno.	Ninguna.
City	Nombre de la ciudad en la que se encuentra.	String.	Ninguno.	Ninguna.
State	Nombre de la provincia en la que se encuentra.	String.	Ninguno.	Ninguna.
Country	Nombre del país en el que se encuentra.	String.	Un solo valor para todos los registros.	Dropear la columna.
departamen	Nombre del municipio en el que reside.	String.	Columna irrelevante para nuestro estudio.	Dropear la columna.

En proveedores encontramos 2 (14.3%) registros con el campo 'Nombre' nulo, de 14 registros totales.

TABLA SUCURSALES:

TÍTULO	DESCRIPCIÓN	TIPO DE DATO	PROBLEMAS	SOLUCIONES
ID	Índice asociado a la sucursal.	Entero.	Ninguno.	Ninguna.
Sucursal	Nombre de la sucursal.	String.	Ninguno.	Ninguna.
Direccion	Dirección en la que se localiza la sucursal.	String.	Ninguno.	Ninguna.
Localidad	Nombre de la localidad en la que se encuentra.	String.	Columna sin normalizar.	Normalizar la columna usando Levenshtein y el maestro de localidad.
Provincia	Nombre de la provincia en la que se encuentra.	String.	Columna sin normalizar.	Normalizar la columna usando Levenshtein y el maestro de localidad.
Latitud	Coordenadas de latitud correspondientes.	Decimal.	Coma decimal.	Normalizar a punto decimal.
Longitud	Coordenadas de longitud correspondientes.	Decimal.	Coma decimal.	Normalizar a punto decimal.

La tabla 'Sucursales' no posee campos nulos en ninguno de sus 31 registros.

TABLA TIPOS DE GASTO:

TÍTULO	DESCRIPCIÓN	TIPO DE DATO	PROBLEMAS	SOLUCIONES
IdTipoGasto	Índice asociado al tipo del gasto.	Entero.	Ninguno.	Ninguna.
Descripcion	Especificación del tipo de gasto.	String.	Ninguno.	Ninguna.
Monto_Aproximado	Importe aproximado de los gastos de dicho tipo.	Entero.	Ninguno.	Ninguna.

La tabla 'TiposDeGasto' no posee campos nulos en ninguno de sus 4 registros.

TABLA CANAL DE VENTA:

TÍTULO	DESCRIPCIÓN	TIPO DE DATO	PROBLEMAS	SOLUCIONES
CODIGO	Índice asociado al medio en el que se efectuó la venta.	Entero.	Ninguno.	Ninguna.
DESCRIPCION	Especificación del medio en el que se efectuó la venta.	String.	Ninguno.	Ninguna.

La tabla 'CanalDeVenta' no posee campos nulos en ninguno de sus 3 registros.

TABLA LOCALIDAD:

TÍTULO	DESCRIPCIÓN	TIPO DE DATO
categoría	Tipos posibles de localidad.	String.
centroide_lat	Coordenadas de latitud correspondientes.	Decimal.
centroide_lon	Coordenadas de longitud correspondientes.	Decimal.
departamento_id	Índice correspondiente al departamento del que hace parte.	Decimal.
departamento_nombre	Nombre del departamento del que hace parte.	String.
fuente	Procedencia del dato.	String.
id	Índice correspondiente a la localidad.	Entero.
localidad_censal_id	Índice correspondiente a la localidad cuando esta última es definida por parámetros censales.	Entero.
localidad_censal_nombre	Nombre correspondiente a la localidad cuando esta última es definida por parámetros censales.	String.
municipio_id	Índice correspondiente al municipio del que hace parte.	Decimal.
municipio_nombre	Nombre correspondiente al municipio del que hace parte.	String.
nombre	Nomenclatura de la localidad.	String.
provincia_id	Índice correspondiente a la provincia de la hace parte	Entero.
provincia_nombre	Nombre correspondiente a la provincia del que hace parte.	String.

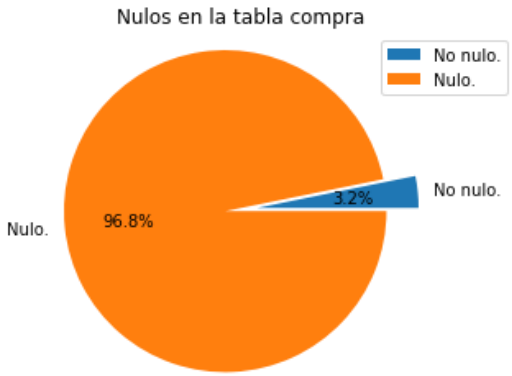
Tabla maestra de localidad, contiene algunos campos irrelevantes a nuestro estudio, además de tener algunos campos en nulo.

TABLAS DE HECHOS:

TABLA COMPRA:

TÍTULO	DESCRIPCIÓN	TIPO DE DATO	PROBLEMAS	SOLUCIONES
IdCompra	Índice asociado a la factura.	Entero	Ninguno.	Ninguna.
Fecha	Fecha en la que se efectuó la compra.	Date.	Fecha en formato distinto a las tablas "Gasto" y "Venta".	Formatear como en las otras tablas.
Fecha_Año	Año en el que se realizó la factura.	Entero.	Redundante con la columna ‘Fecha’.	Dropear la columna para que quede acorde con "Venta" y "Gasto".
Fecha_Mes	Número del mes en el que se realizó la factura.	Entero.	Redundante con la columna ‘Fecha’.	Dropear la columna para que quede acorde con "Venta" y "Gasto".
Fecha_Periodo	Código del período en el que se realizó la factura.	Entero.	Redundante con la columna ‘Fecha’.	Dropear la columna para que quede acorde con "Venta" y "Gasto".
Id_Producto	Índice asociado al producto facturado.	Entero.	Columna poco significativa ya que no tengo la tabla "Producto" para referirla.	Ninguna, dejarla por si en un futuro integran la tabla "Producto".

Cantidad	Número de productos facturados.	Entero.	347 outliers detectados con rango intercuartílico.	Detectar y marcar los outliers.
Precio	Valor unitario del producto.	Decimal.	367 valores nulos. 599 outliers detectados con rango intercuartílico.	Imputar los nulos. Detectar y marcar los outliers.
IdProveedor	Índice asociado al proveedor del producto	Entero.	Ninguno.	Ninguna.



- La única columna con registros nulos de la tabla Compra es 'Precio' con 367/11539 (3.2%).
- En el campo 'Cantidad' encontramos 347/11539 (3.0%) de outliers.
- Por último, en el campo 'Precio' encontramos 599/11539 (5.2%) de outliers.
- Para el cálculo de outliers se usó el rango intercuartílico ya que no es sensible a valores extremos.

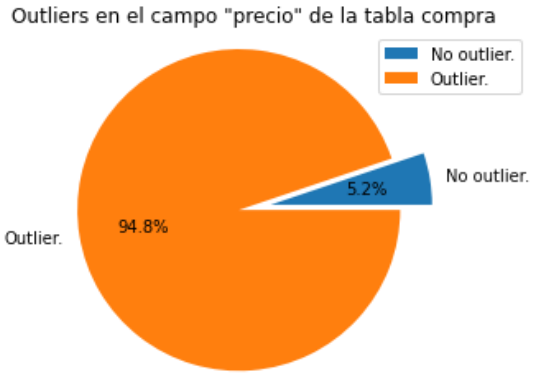


TABLA GASTO:

TÍTULO	DESCRIPCIÓN	TIPO DE DATO	PROBLEMAS	SOLUCIONES
IdGasto	Índice asociado a la factura.	Entero.	Ninguno.	Ninguna.
IdSucursal	Índice de la sucursal en la que se realizó el gasto.	Entero.	Ninguno.	Ninguna.
IdTipoGasto	Índice del tipo de gasto realizado.	Entero.	Ninguno.	Ninguna.
Fecha	Fecha en la que se realizó el gasto.	Date.	Ninguno.	Ninguna.
Monto	Importe correspondiente.	Decimal.	Ninguno.	Ninguna.

Asombrosamente la tabla 'Gasto' no contiene valores nulos u outliers en ninguno de sus 8640 registros.

TABLA VENTA:

TÍTULO	DESCRIPCIÓN	TIPO DE DATO	PROBLEMAS	SOLUCIONES
IdVenta	Índice asociado a la factura.	Entero	Ninguno.	Ninguna.
Fecha	Fecha en la que se efectuó la venta.	Date.	Ninguno.	Ninguna.
Fecha_Entrega	Fecha en la que se realizó la entrega del producto al cliente.	Date.	Ninguno.	Ninguna.
IdCanal	Índice correspondiente al canal en el que se realizó la venta.	Entero.	Ninguno.	Ninguna.
IdCliente	Índice correspondiente al cliente que realizó la orden.	Entero.	Ninguno.	Ninguna.
IdSucursal	Índice correspondiente a la sucursal en la que se realizó la venta.	Entero.	Ninguno.	Ninguna.
IdEmpleado	Índice correspondiente al empleado que realizó la venta.	Entero.	Dato irrelevante ya que no tengo la tabla "Empleado".	Dropear la columna.
IdProducto	Índice correspondiente al producto vendido.	Entero.	Columna poco significativa ya que no tengo la tabla "Producto" para referirla.	Ninguna, dejarla por si en un futuro integran la tabla "Producto".
Precio	Valor unitario del producto vendido.	Decimal.	920 nulos. 2476 outliers calculados con rango intercuartílico.	Imputar los nulos por el último valor no nulo del mismo "IdProducto". Detectar y marcar outliers
Cantidad	Valor unitario del producto vendido.	Decimal.	884 nulos. 910 outliers calculados con rango intercuartílico.	Imputar los nulos por 0.0. Detectar y marcar outliers.

- En total hay 1788/46180 (3.9%) registros con al menos un valor nulo.
- En el campo ‘Cantidad’ encontramos 910/46180 (2.0%) de outliers.
- Por último, en el campo ‘Precio’ encontramos 2476/46180 (5.4%) de outliers.
- Para el cálculo de outliers se usó el rango intercuartílico ya que no es sensible a valores extremos.