PRESENTACIÓN STORYTELLING DE LOS DATOS (PI)

MODELO DE EMPRESA DE VENTA DE PRODUCTOS

POR: JUAN FLORES DTS-01

DICCIONARIO TABLAS

Compra	Variable	Definición	Llave
	IdCompra	Código Identifiación de compra	
	Fecha	Fecha completa	aaaa-mm-dd
	Fecha_Año	Año del evento	aaaa
	Fecha_Mes	Mes del evento	número mes
	Fecha_Periodo	Año y mes	aaaamm
	IdProducto	Código Identifiación de producto	
	Cantidad	Cantidad	
	Precio	Precio	Moneda local
	IdProveedor	Código Identifiación de proveedor	

Venta	Variable	Definición	Llave
	IdVenta	Código Identifiación de cliente	
	Fecha	Fecha completa	aaaa-mm-dd
	Fecha_Entrega	Fecha de producto entregado	aaaa-mm-dd
	IdCanal	Código Identifiación de cliente	
	IdCliente	Código Identifiación de cliente	
	IdSucursal	Código Identifiación de cliente	
	IdEmpleado	Código Identifiación de cliente	
	IdProducto	Código Identifiación de cliente	Posición Geográfica
	Precio	Precio unitario de producto	moneda local
	Cantidad	Cantidad de productos vendidos	

Gasto	Variable	Definición	Llave
	IdGasto	Código Identificación de cliente	
	IdSucursal	Código Identifiación de sucursal	
	IdTipoGasto	Código Identifiación de tipo de gasto	1, 2, 3 : Teléfono, Presencial, Internet
	Fecha	Fecha completa de evento	aaaa-mm-dd
	Monto	Gasto realizado	moneda local

EVENTOS ENCONTRADOS EN CADA TABLA

Análisis de la calidad de los datos originales

NORMALIZACIÓN TÍTULOS DE COLUMNAS

- Para un mejor análisis y referencia de los elementos entre tablas, se normalizaron los nombres de las columnas bajo los siguientes criterios:
    1. Columnas en español
    2. Identificador respectivo por cada tabla diferente de los demás identificadores
    3. Nombre de columna representativo de los valores que ésta contiene
    4. Cambio del tipo de dato al más pertinente
    5. Eliminación de columnas sin registro alguno

TRANSFORMACIÓN DE TABLA OBTENIDA DEL ARCHIVO CLIENTES

COLUMNAS EN BRUTO

#	Column	Non-Null Count	Dtype
0	ID		
1	Provincia	729 non-null	object
2	Nombre_y_Apellido		
3	Domicilio	729 non-null	object
4	Telefono	680 non-null	object
5	Edad	735 non-null	int64
6	Localidad	728 non-null	object
7	X	729 non-null	object
8	Y	729 non-null	object
9	col10	0 non-null	fl 64

NORMALIZACIÓN DE TABLA

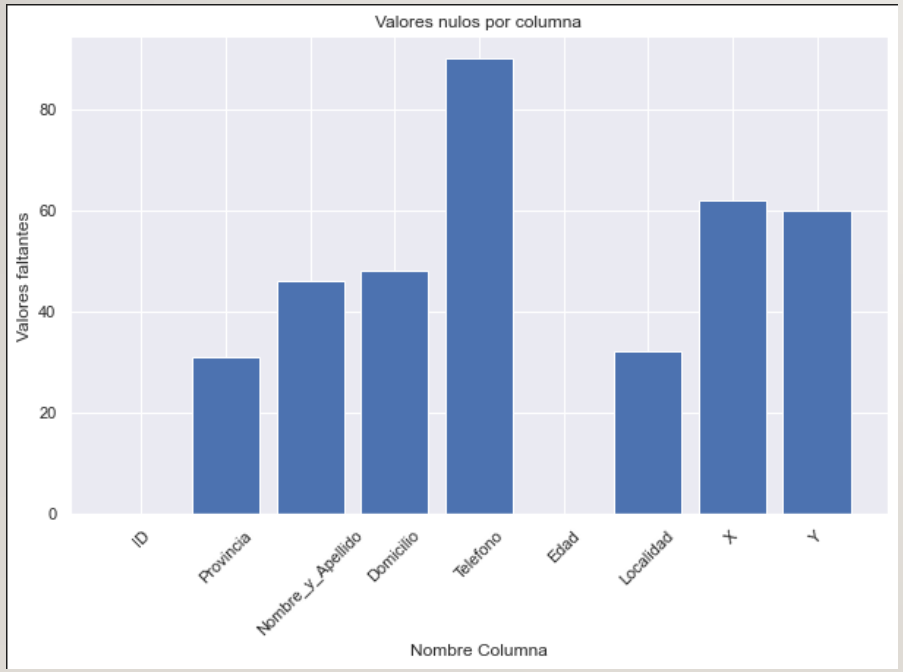
#	Column	Non-Null Count	Dtype
0	IdCliente	3407 non-null	int64
1	Provincia	3376 non-null	object
2	Nombre_completo	3361 non-null	object
3	Domicilio	3359 non-null	object
4	Telefono	3317 non-null	object
5	Edad	3407 non-null	int64
6	Localidad	3375 non-null	object
7	Longitud	3345 non-null	float16
8	Latitud	3347 non-null	float16

IDENTIFICACIÓN DE ELEMENTOS NULOS

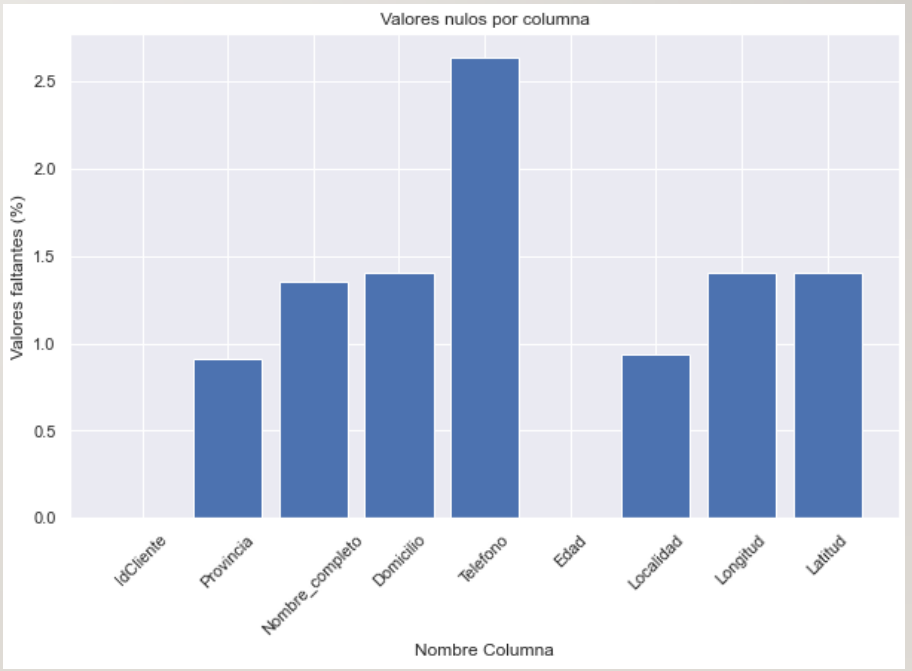
- Se identificaron los elementos nulos de cada una de las tablas
- Por la naturaleza distinta de cada tabla y el tipo de valor nulo representado se tomaron decisiones diferentes para cada conjunto de datos
- Se prefirió tomar la perspectiva de incompletitud de los datos sobre datos erróneos

ELEMENTOS NULOS CLASIFICADOS POR COLUMNA DE LA TABLA CLIENTES

CANTIDAD DE VALORES FALTANTES



PORCENTAJE DE VALORES FALTANTES



LOS VALORES DE INCOMPLETITUD VAN DE: 14%* A 0%**

La estrategia para lidiar con los elementos nulos se definirá más adelante

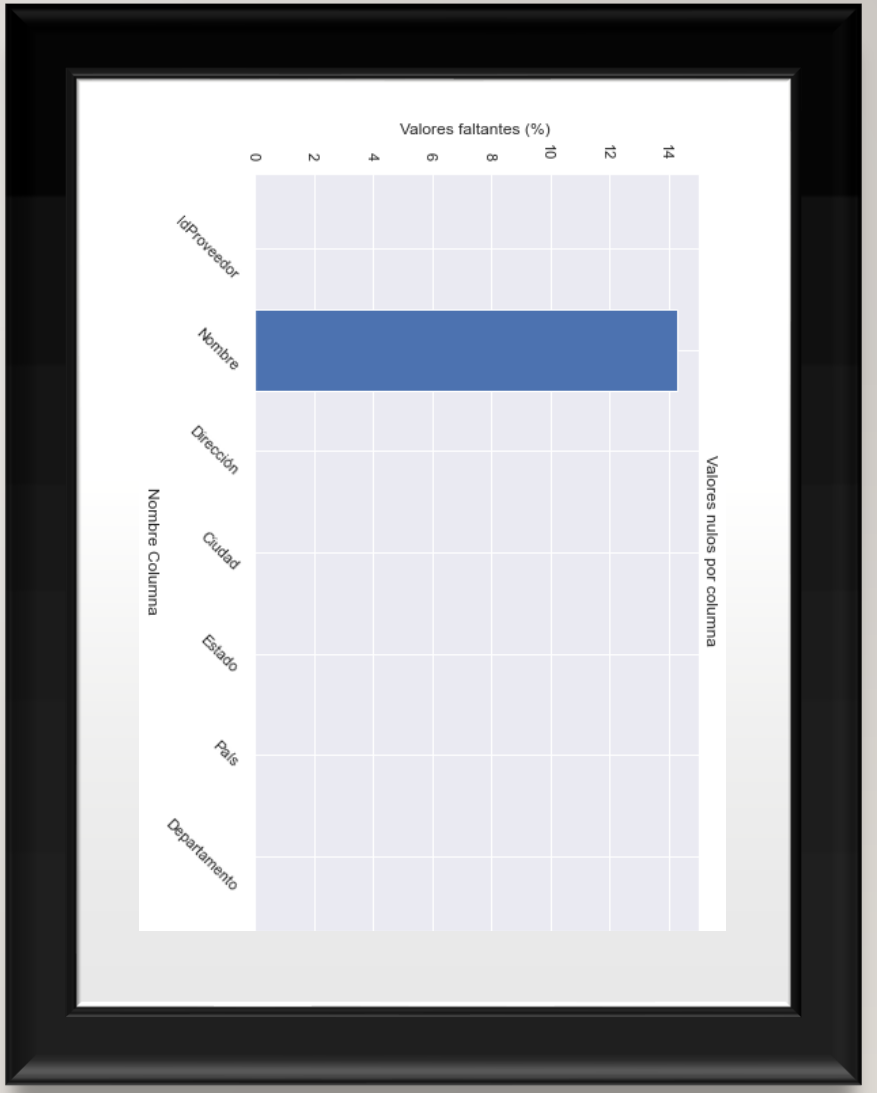


*Tabla proveniente del archivo Proveedores.csv
**Tabla proveniente del archivo Gasto.csv entre otros

DATOS ERRÓNEOS

- Se encontró que la tabla proveniente del archivo Clientes.csv contenía ~2% de datos cuya columna no correspondía a la que estaban alojados la mezcla fue Latitud con Longitud, además, de las mismas columnas la posición geográfica sugería una región rotada 180° del punto original.

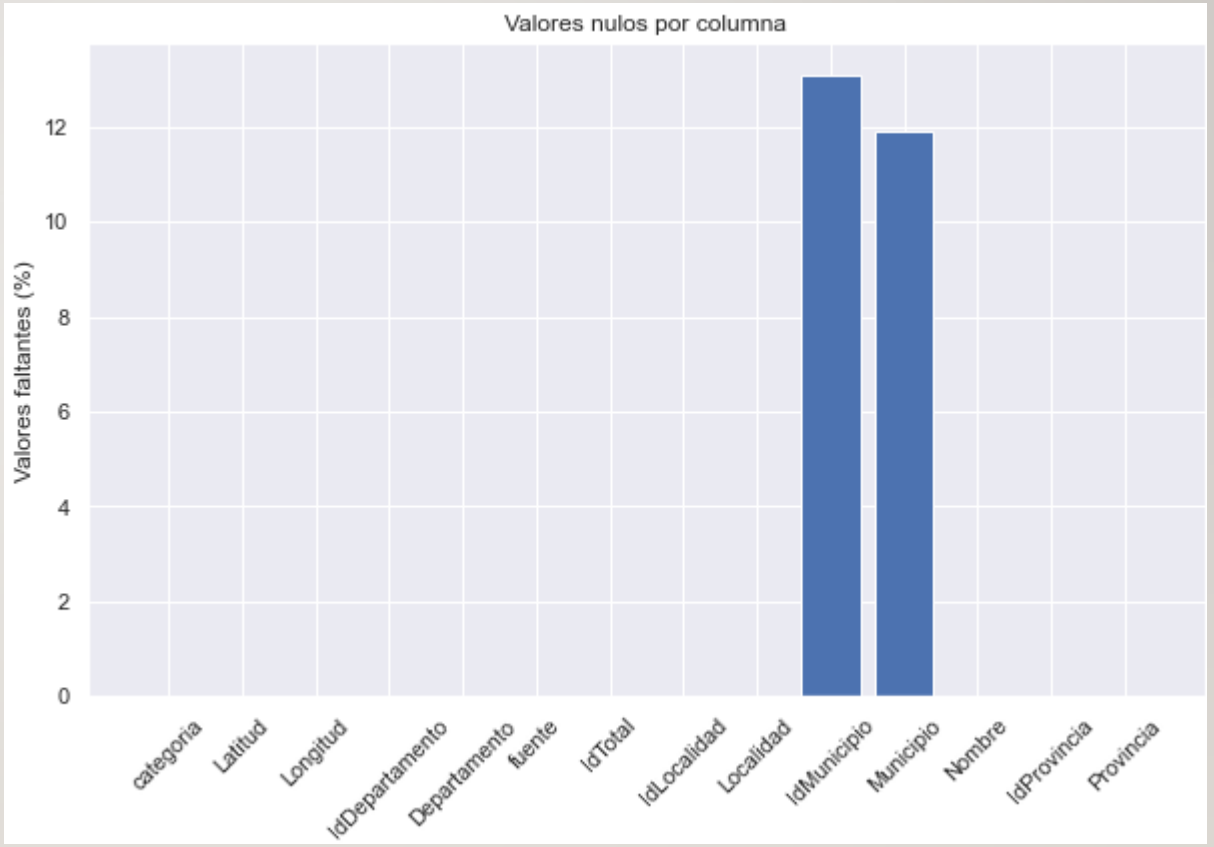
	58.24998664	34.72714735
	58.23370523	34.75063453
	57.94472391	34.83803456
	57.91147674	34.87030133
	57.89268291	34.92541294
	57.89034227	34.86153325
	57.89011046	34.88376252
	57.88193471	34.86948863
	-26.73824371	-65.25201214
	-32.86297422	-68.83853887
	-32.8670104	-68.82421367

CONSIDERACIONES

- Se tomaron las siguientes consideraciones respecto de los datos:
  - Los valores de Latitud y Longitud se consideraron verdaderos
  - Dada la constitución geográfica de la región en donde se presentó el caso de estudio, la columna de mayor importancia geográfica fue 'Localidad' para cada tabla
  - La tabla proveniente del archivo Localidades.csv, por petición de quien proporcionó los datos, ya se encuentra normalizada y sería usada como referencia
  - Cualquier columna de cualquier tabla que tenga Id se considera como verdadero y único de manera semántica, en caso de encontrar alguno repetido me reservé mi derecho de elegir cuál fue correcto

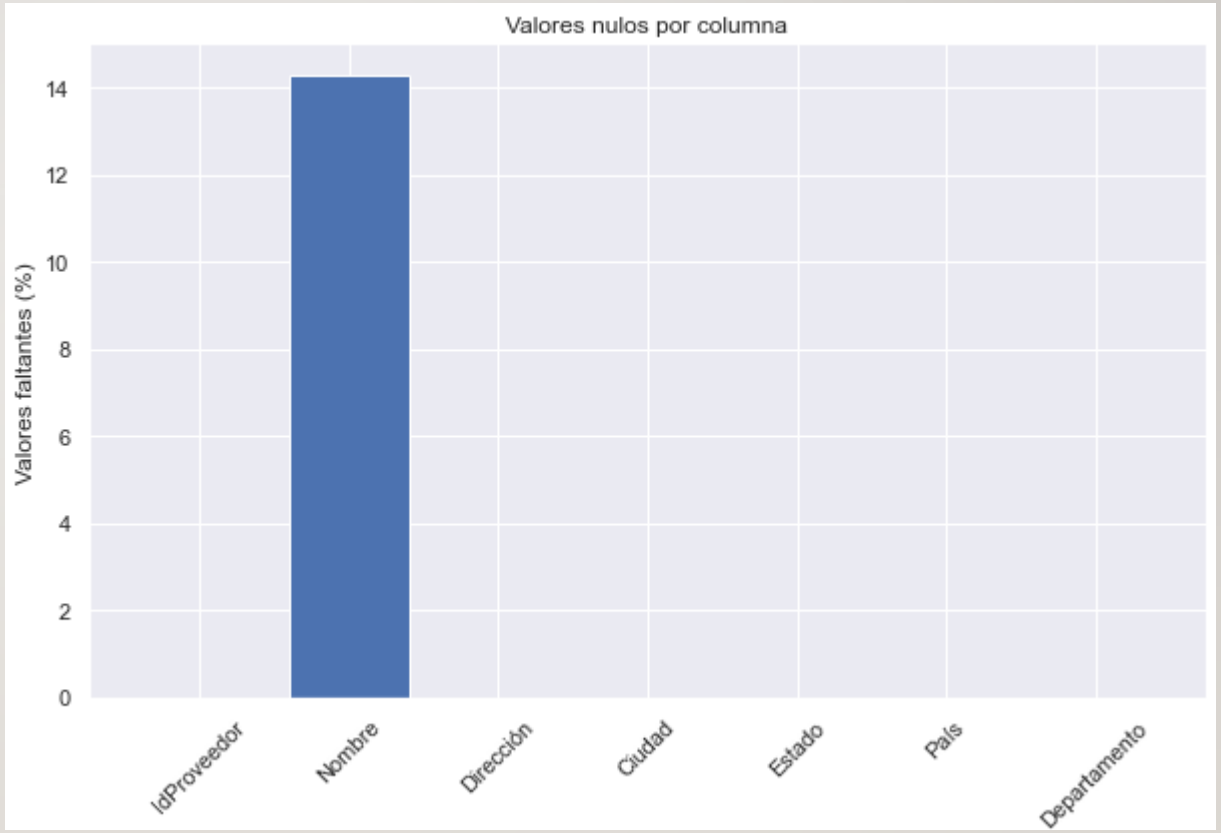
TRATAMIENTO DE VALORES NULOS

- Si los valores nulos se referían al nombre o identificador geográfico que no fuera dentro de la columna ‘Localidad’, se eliminó toda la columna



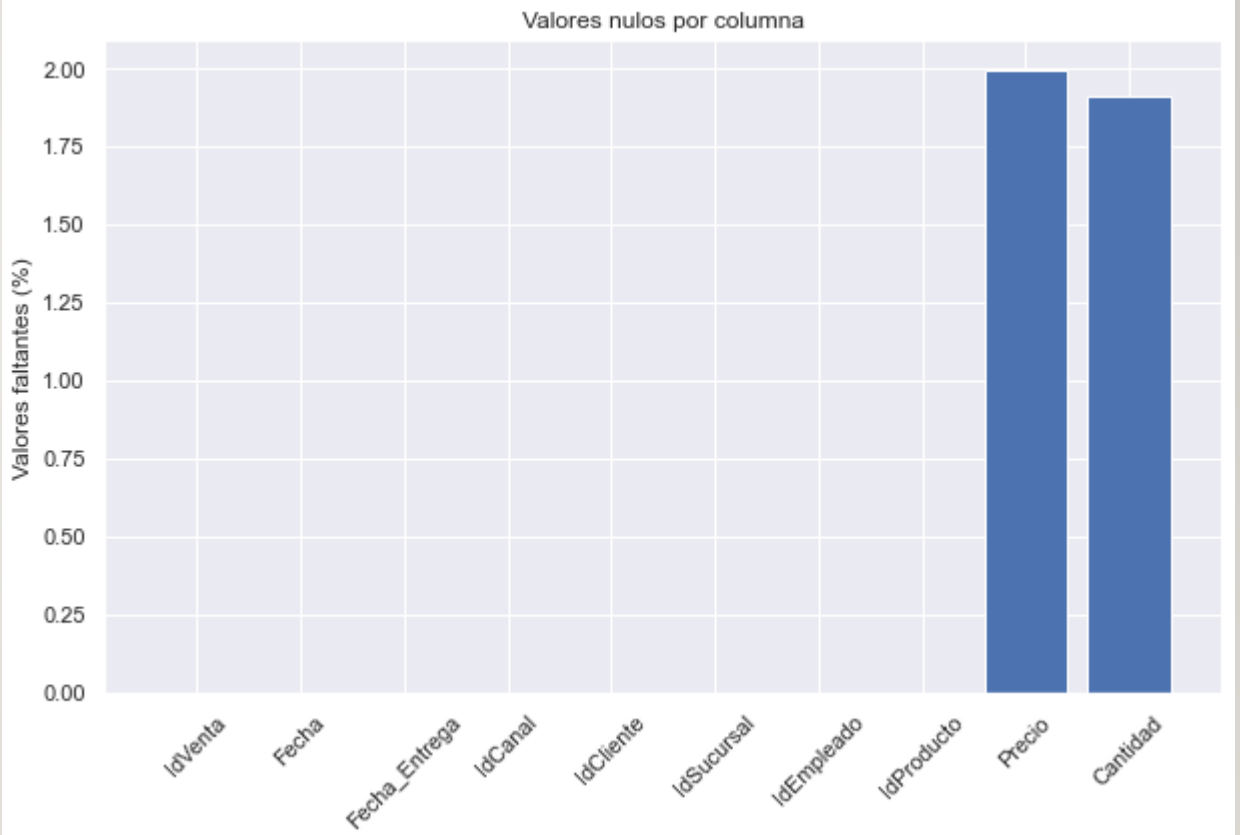
TRATAMIENTO DE VALORES NULOS

- Si los valores nulos se referían al nombre del tipo de alguna tabla, se cambiaron por 'Sin dato'



TRATAMIENTO DE VALORES NULOS

- Si la importancia de los valores nulos se encontraron en columnas de alto valor para el negocio, previo a su eliminación, se filtraron haciendo un tratamiento de outliers y, de seguir existiendo posterior al filtrado, se eliminaron



OUTLIERS

- Por el giro del negocio es imperativo siempre analizar los datos que correspondan a dinero o estén directamente relacionados; es decir, aquellos que representen gastos, precios, cantidades de producto comprado/vendido
- Se abordó desde el punto de vista de los cuartiles y se emplearon gráficas para un mejor entendimiento
- Se utilizó el método de los cuartiles para determinar la eliminación de los valores OUTLIERS y para los casos muy restrictivos se consultó a la empresa para decidir un margen de eliminación (realmente existen productos muy costosos que se venden)
- Se encontraron errores por outliers de hasta el 6%

LOS VALORES ERRÓNEOS POR OUTLIERS VAN DE: 6%* A 0%**

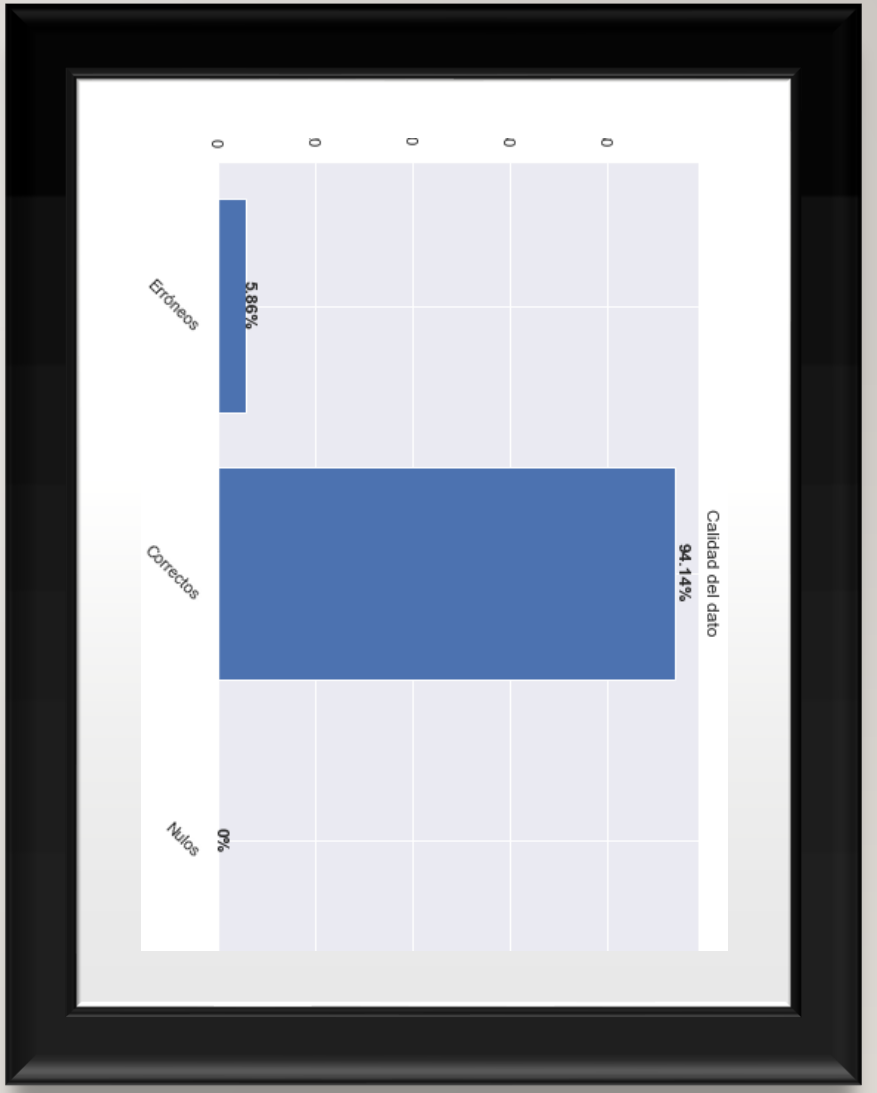
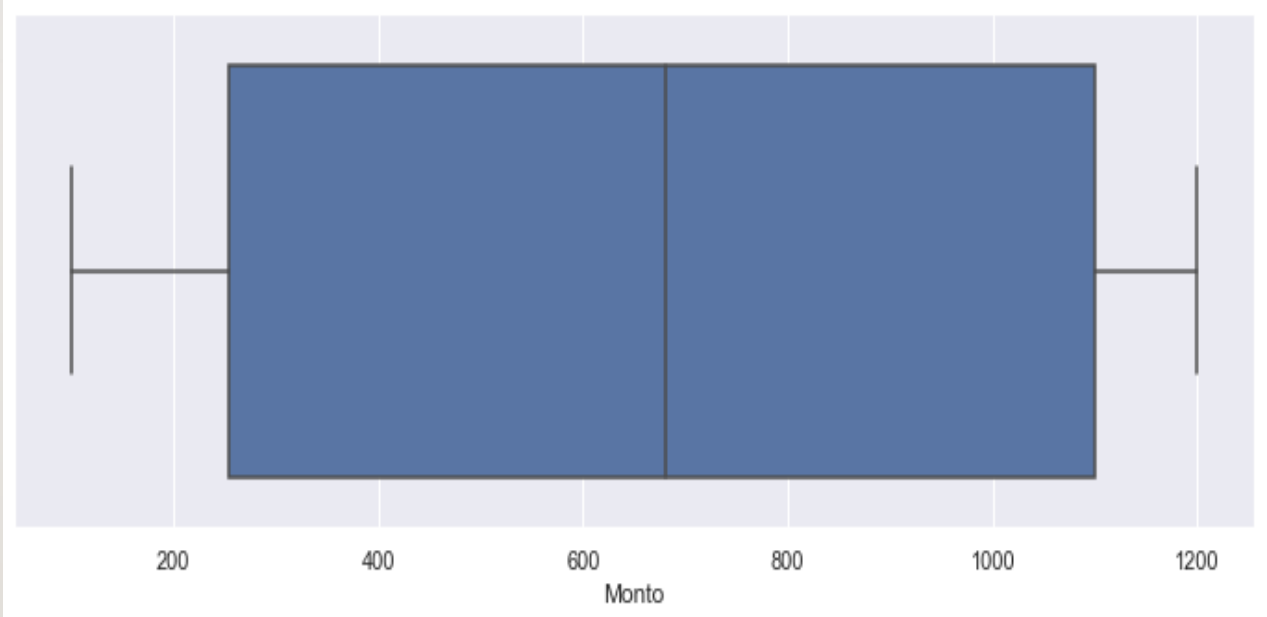


*Tabla proveniente del archivo Compras.csv
**Tabla proveniente del archivo Gasto.csv entre otros

TRATAMIENTO DE OUTLIERS

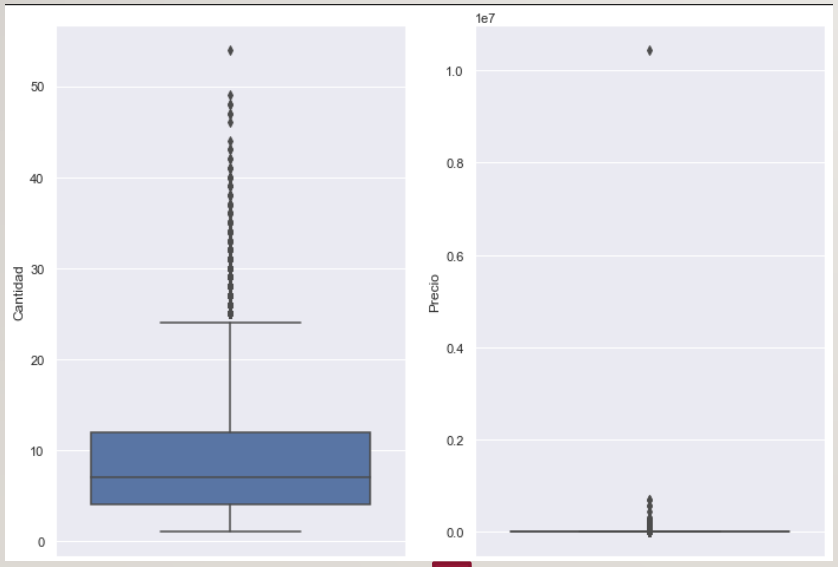
- Si analíticamente no se encontraron, la gráfica boxplot se visualizó de la siguiente manera

Boxplot de monto gastado de la tabla gasto del archivo Gasto.csv

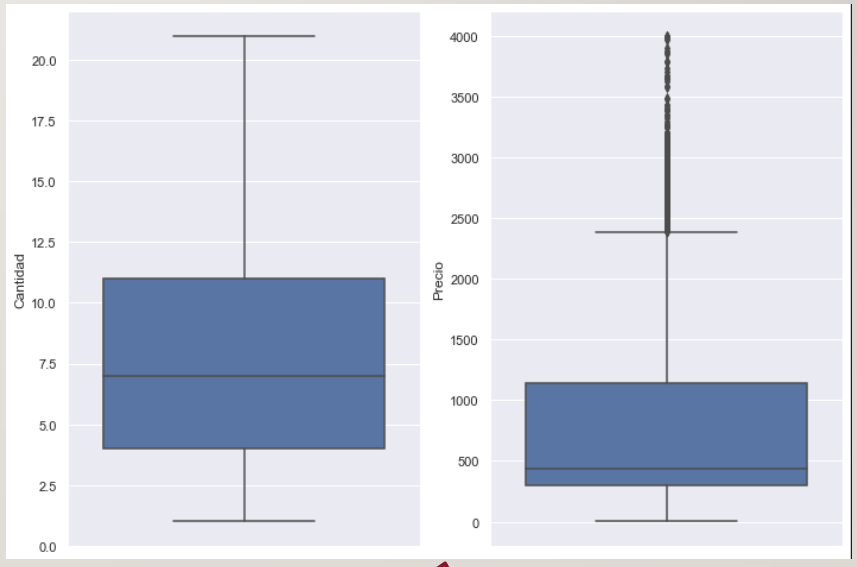


TRATAMIENTO DE OUTLIERS

COLUMNAS 'CANTIDAD' Y 'PRECIO' DE TABLA COMPRA



COLUMNAS 'CANTIDAD' Y 'PRECIO' DE TABLA COMPRA SIN OUTLIERS

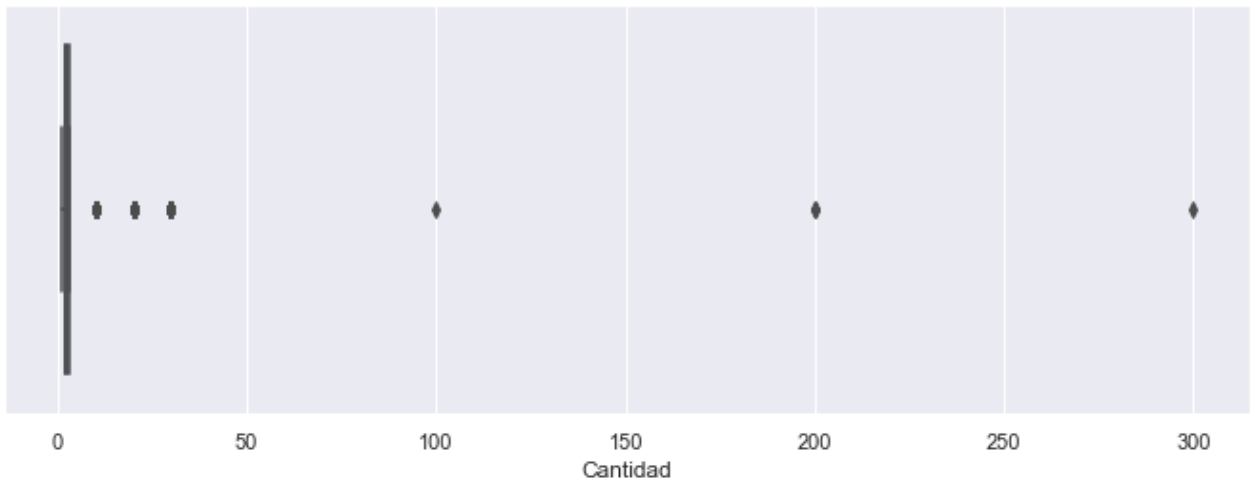


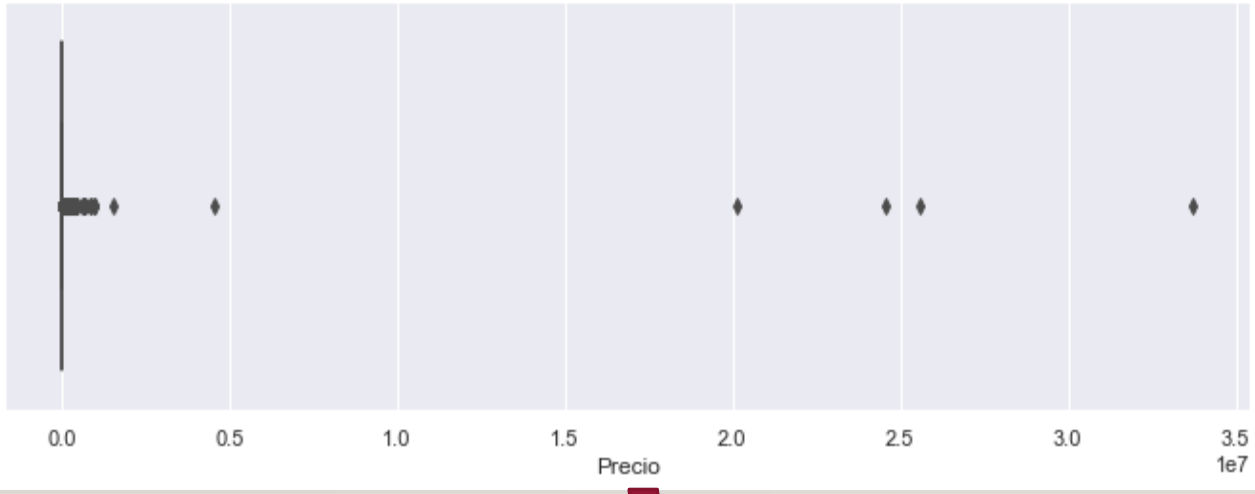
TRATAMIENTO DE OUTLIERS Y VALORES NULOS

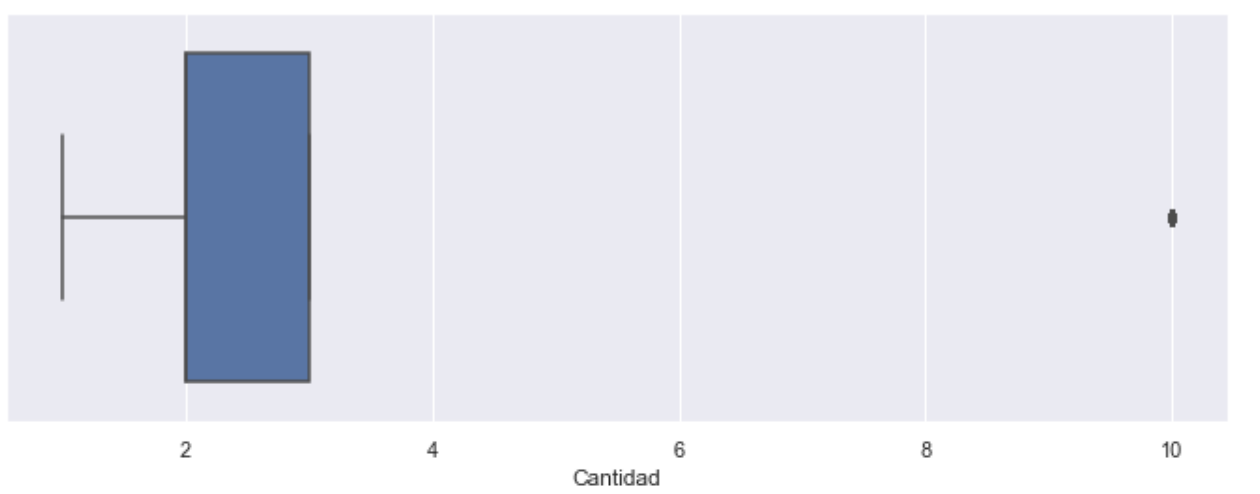
- Para el caso específico de una venta se toman, además, las siguientes consideraciones:

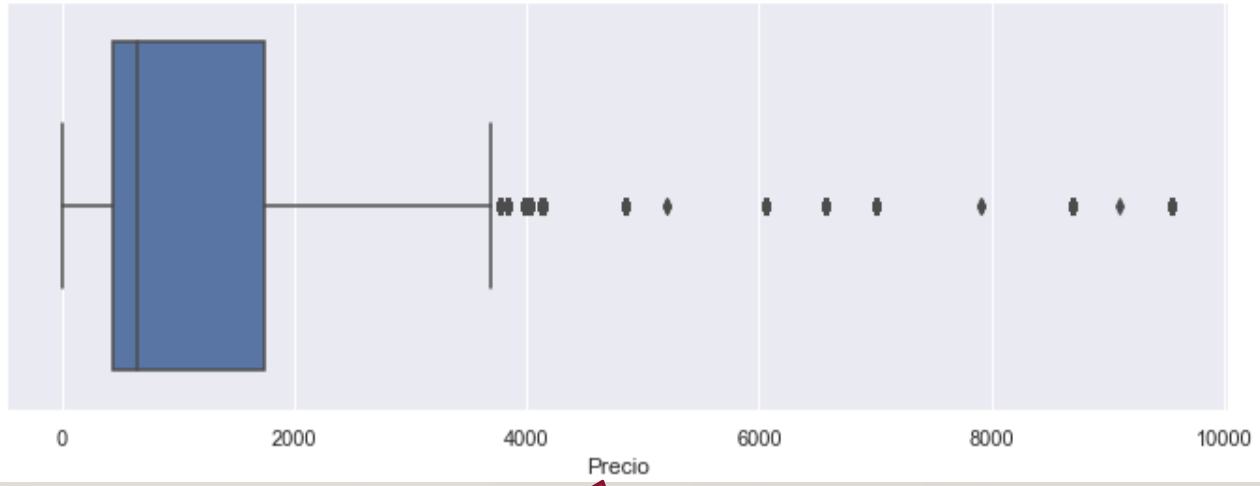
1. Si de la idea anterior aún quedan registros nulos o vacíos, en caso de ser un precio se rellena con el valor promedio
2. De un modo similar, si el dato corresponde a una cantidad se cambia al valor 1, pues se tiene la certeza de que la venta existió y la cantidad de producto mínimo vendido posible es 1
3. Existen registros que requieren una condicional para ser detectados debido a que, por la naturaleza del negocio, podrían existir ventas de ultimo momento mayoristas, por lo que se estableció un margen de precio-cantidad para detectarlas y no eliminarlas

Tratamiento de la tabla venta sobre las columnas ‘Precio’ y ‘Cantidad’









FINALIZACIÓN

Se considera que a partir de éste punto las tablas contienen sólo datos representativos del negocio para poder modelar y/o sacar conclusiones de ellos

TRABAJO FUTURO

- Solicitar al negocio los datos de la dirección de sus clientes para los datos faltantes correspondientes, aunque se pudieran usar los datos de la tabla de Localidades, sólo darían una ubicación aproximada y sería la misma para diferente clientes
- Implementar una normalización de los datos de texto más extensa y sobre más campos que sólo 'Localidad'
- Crear un rango etario efectivo para los clientes y distribuirlos en dichos rangos
- Implementación de modelo para 'producción'