Reporte del análisis de datos para la empresa de giro comercial

2022

30 JULIO

HENRY

Creado por: Juan Flores DTS-01

Reporte del análisis de datos para la empresa de giro comercial

Prefacio

Una empresa de giro comercial (ventas de productos al público) tomó la decisión de ser una empresa Data-driven, concepto por el cual los datos son fundamentales a la hora de tomar decisiones.

La gerencia decidió crear una base de datos, teniendo en cuenta las principales entidades que son las ventas, compras y gastos.

La empresa se contacta conmigo y me proporciona una serie de archivos de datos en formato 'csv' (provenientes de su base de datos). Su intención es que los analice para entregar un reporte que contenga todas las incongruencias de los datos o la mala calidad de los mismos.

Una vez analizados ofrecer una serie de KPI´s útiles para ayudar en la toma de decisisones y así encontrar un lugar para establecer una nueva sucursal.

Los archivos que me fueron entregados son los siguientes:

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

Además, se me mencionó que pudieran existir otros archivos, con la misma estructura, de clientes y ventas que se puedan integrar evitando la duplicidad de información.

El trabajo se abordó con la perspectiva presentada en **Presentación storytelling de los datos (PI) [archivo adjunto]** y se dará una descripción más detallada del proceso realizado. Se proporcionan las conclusiones llegadas y los trabajos que se requieren implementar para poder cubrir en su totalidad las peticiones.

Es importante mencionar que el problema se abordó desde un punto de vista general acumulativo, por lo que se crearán dataframes maestros y cada nuevo será agregado a su maestro correspondiente (en caso de existir).

La manera de determinar outliers se realizó con la fórmula de los cuartiles:
$q<Q_1-1.5*IQR$ y $q>Q_3+1.5*IQR$

Introducción

En el siguiente reporte se analizaron y normalizaron los archivos proporcionados por la empresa, encontrando que la incompletitud de los datos oscila entre un 0% a un 14% dependiendo el archivo empleado y un mínimo de 94% de datos válidos. Se contactó a la empresa para definir un margen, previo a la eliminación de registros, para poder realizar la eliminación sin temor a excluir registros relevantes. Se concluye que la sucursal debe abrirse en la localidad de Rosario y se sugiere el cierre de la sucursal localizada en Quilmes. Finalmente hago hincapié en que se necesita una mejora de normalización de los datos, una implementación del modelo para aceptar nuevos archivos, un criterio para definir qué hacer con ciertos datos faltantes para la tabla generada con el archivo 'Clientes.csv'

"La sucursal debe abrirse en Rosario
y
Se sugiere el cierre de la sucursal de Quilmes"

Diccionarios

Clientes	Variable	Definición	Llave
	ID	Código Identificación de cliente	
	Provincia	Provincia	
	Nombre_y_Apellido	Nombre completo	
	Domicilio	Dirección	
	Telefono	Teléfono	
	Edad	Edad en años	
	Localidad	Localidad	
	X	Longitud	Posición Geográfica
	Y	Latitud	Posición Geográfica
	col10	Sin Información'	Sin valores

Compra	Variable	Definición	Llave
	IdCompra	Código Identificación 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 Identificación de producto	
	Cantidad	Cantidad	
	Precio	Precio	Moneda local
	IdProveedor	Código Identificación de proveedor	

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

Localidades	Variable	Definición	Llave
	categoria	Tipo localidad	Localidad simple, Componente de localidad compuesta, Entidad, Localidad simple con entidad, Componente de localidad compuesta con entidad
	centroide_lat	Centroide de Latitud	Posición Geográfica
	centroide_lon	Centroide de Longitud	Posición Geográfica
	departamento_id	Código Identificación de departamento	
	departamento_nombre	Nombre de Departamento	
	fuente	Fuente de obtención del dato	
	id	Código Identificación de localidad	
	localidad_censal_id	Código Identificación de localidad censal	
	localidad_censal_nombre	Nombre de Identificación de localidad censal	
	municipio_id	Código Identificación de municipio	
	municipio_nombre	Nombre de Municipio	
	nombre	Nombre de Localidad	
	provincia_id	Código de Identificación de provincia	
	provincia_nombre	Nombre de provincia	

Proveedores	Variable	Definición	Llave
	IDProveedor	Código Identifiación de Proveedor	
	Nombre	Nombre	
	Address	Dirección Proveedor	
	City	Ciudad Proveedor	
	State	Estado Proveedor	
	Country	País Proveedor	
	departamen	Departamento Proveedor	

Sucursales	Variable	Definición	Llave
	ID	Código Identificación de cliente	
	Sucursal	Nombre Sucursal	
	Direccion	Dirección	
	Localidad	Localidad	
	Provincia	Provincia	
	Latitud	Latitud	Posición Geográfica
	Longitud	Longitud	Posición Geográfica

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

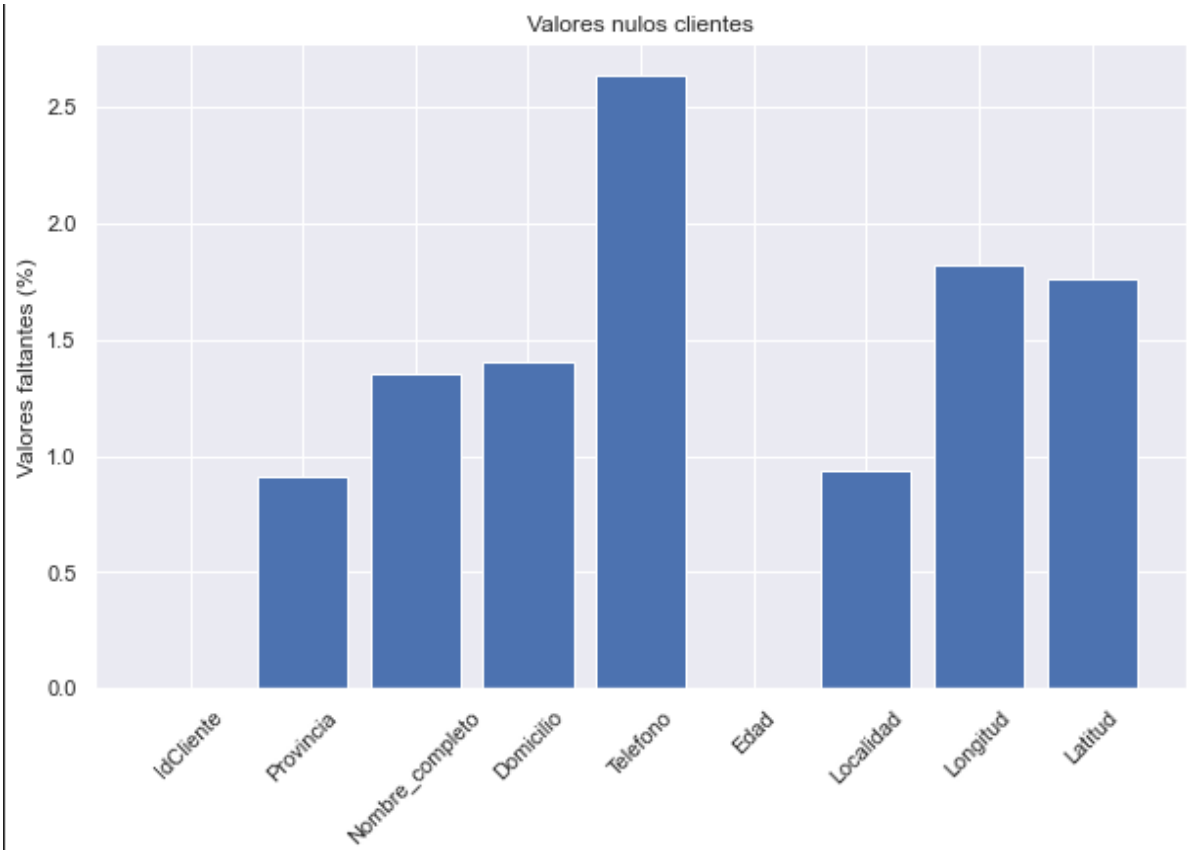
Reporte de calidad

Cada archivo ‘csv’ fue agergado a un dataframe y se normalizaron las leyendas de cada columna (para cada archivo que lo necesitara), se realizó cambio al tipo de dato más conveniente y se obtuvo el porcentaje de elementos faltantes y/o erróneos, si es que los hay.

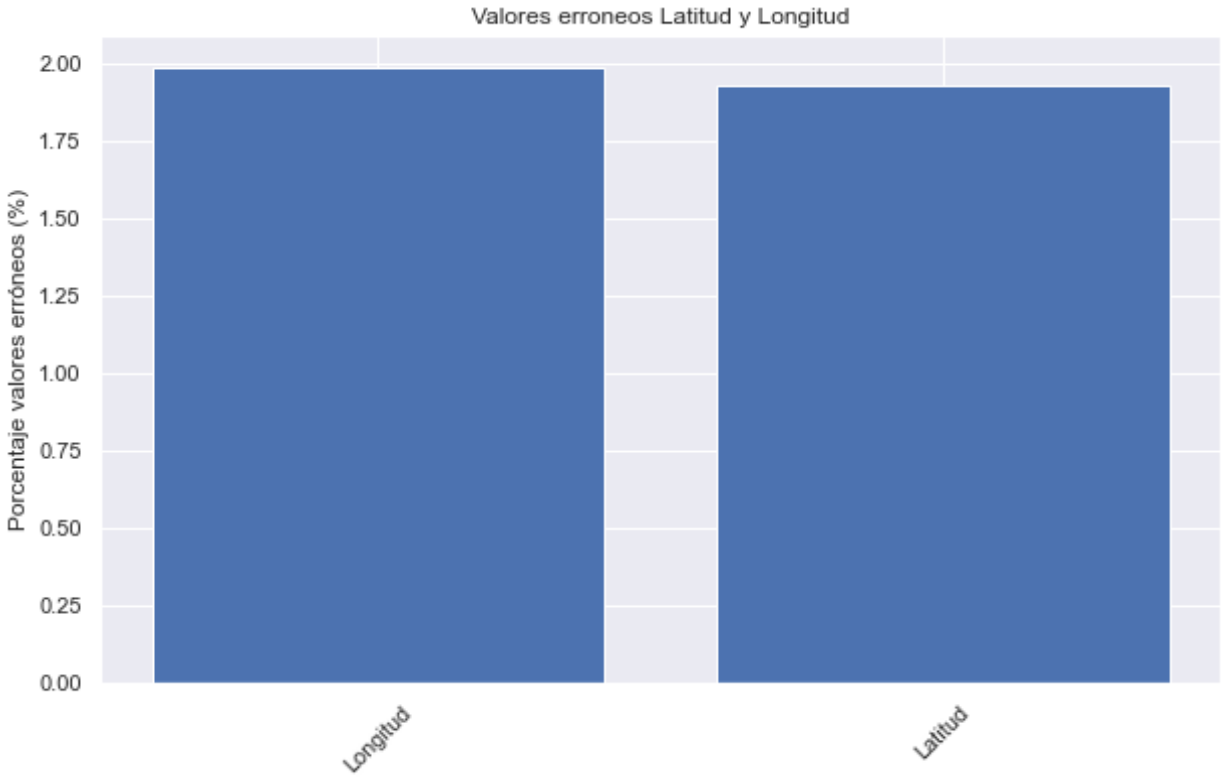
A continuación se mostrarán dichos cambios por dataframe donde los datos no estén completos al 100% o tengan, al menos, un error.

-**cliente** (obtenido de ‘Clientes.csv’).

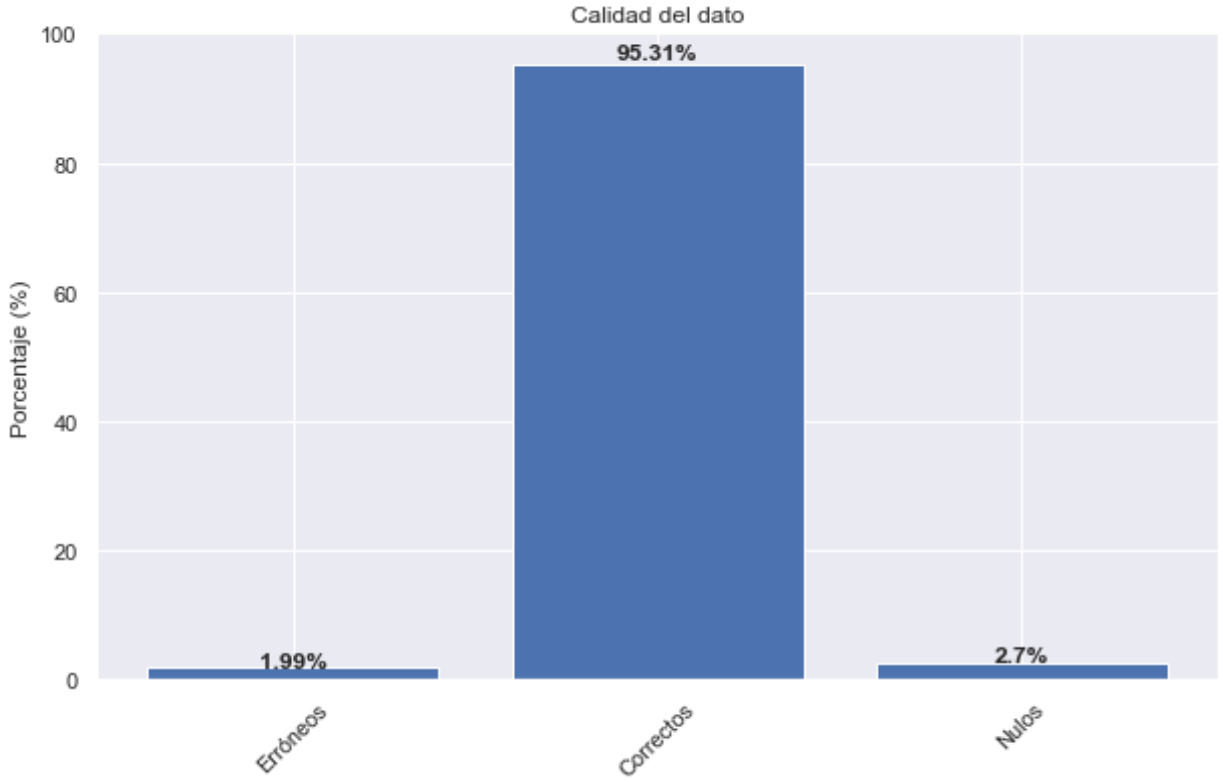
El porcentaje de valores nulos se distribuye de la siguiente manera.



Al analizar los datos de la columna ‘Longitud’ y ‘Latitud’ se observó que existen datos de ‘Longitud’ en ‘Latitud’ y viceversa, además puntos que se encuentran a 180° del original

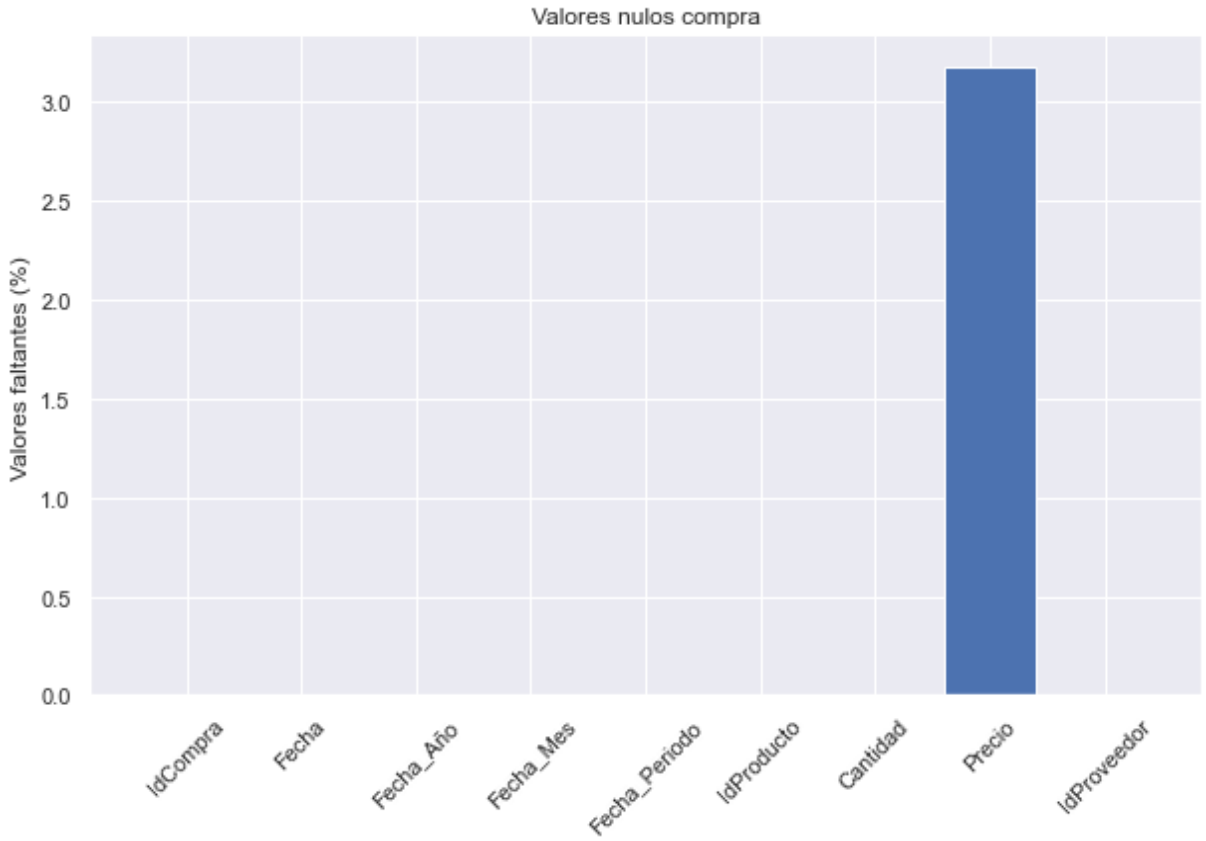


Quedando de la siguiente manera.

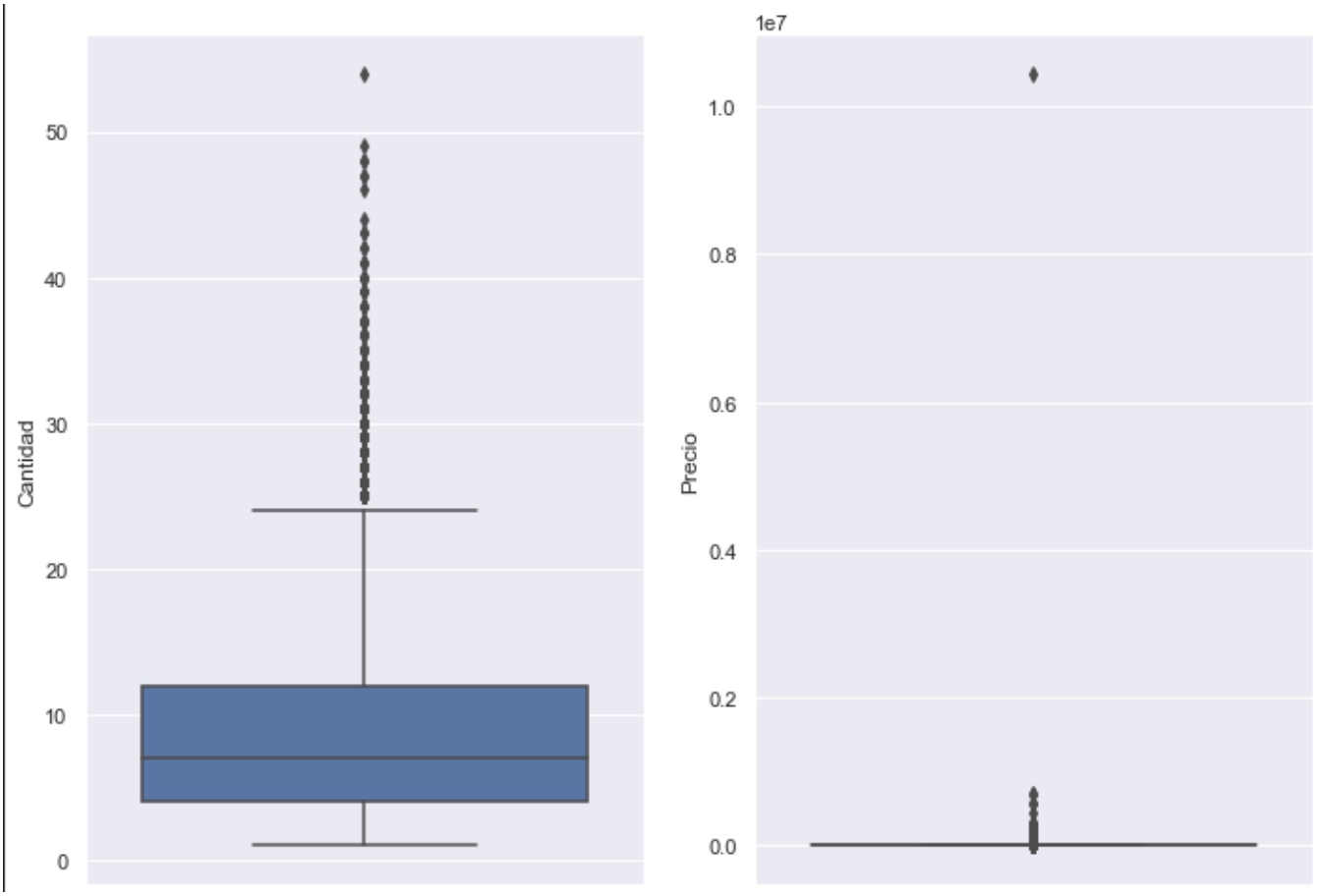


-**compra** (obtenido de 'Compras.csv').

El porcentaje de valores nulos se distribuye de la siguiente manera.

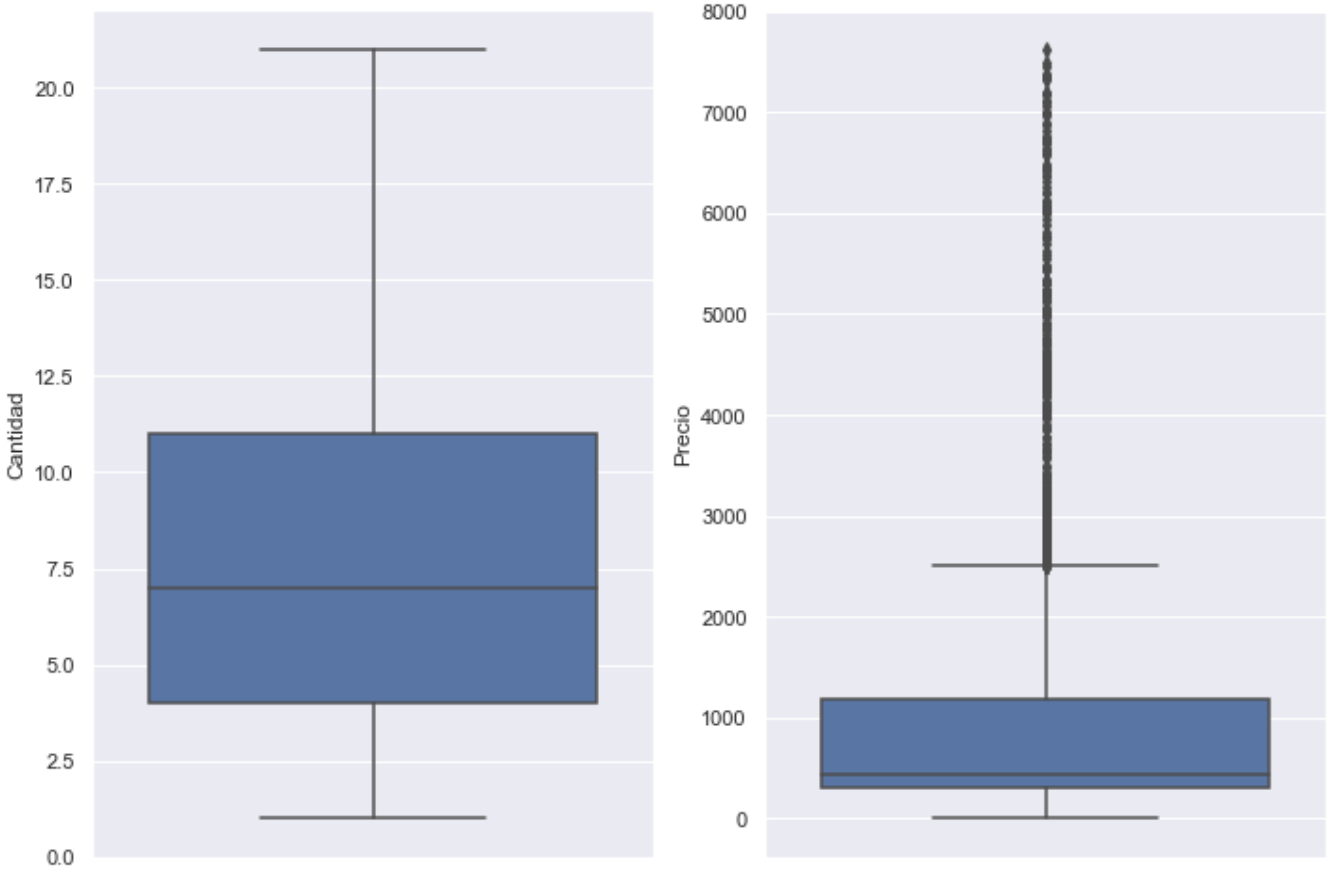


Para determinar el número de datos erróneos sobre las columnas ‘Cantidad’ y ‘Precio’

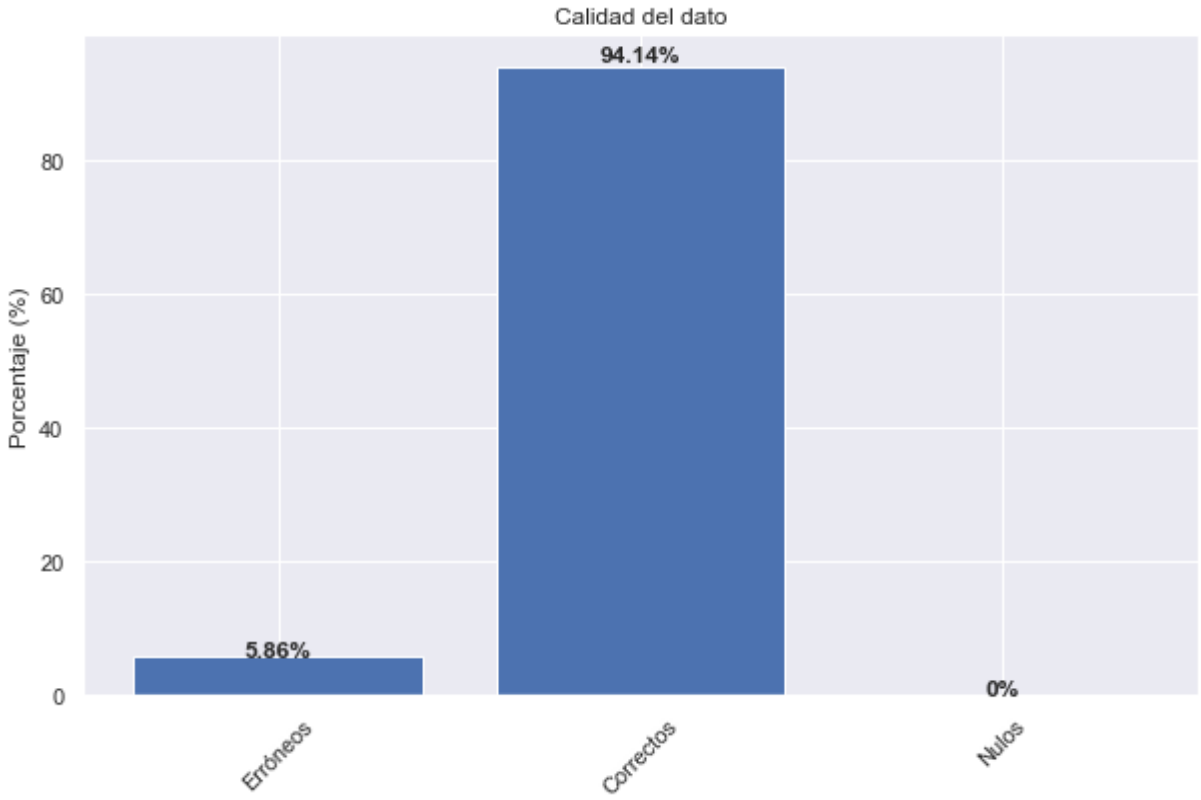


Para los outliers de ‘Precio’ se contactó al negocio y se le preguntó cuál es el artículo más costoso comprado, de ahí se definió el margen para eliminar registros.

Se muestra el resultado en la siguiente gráfica.

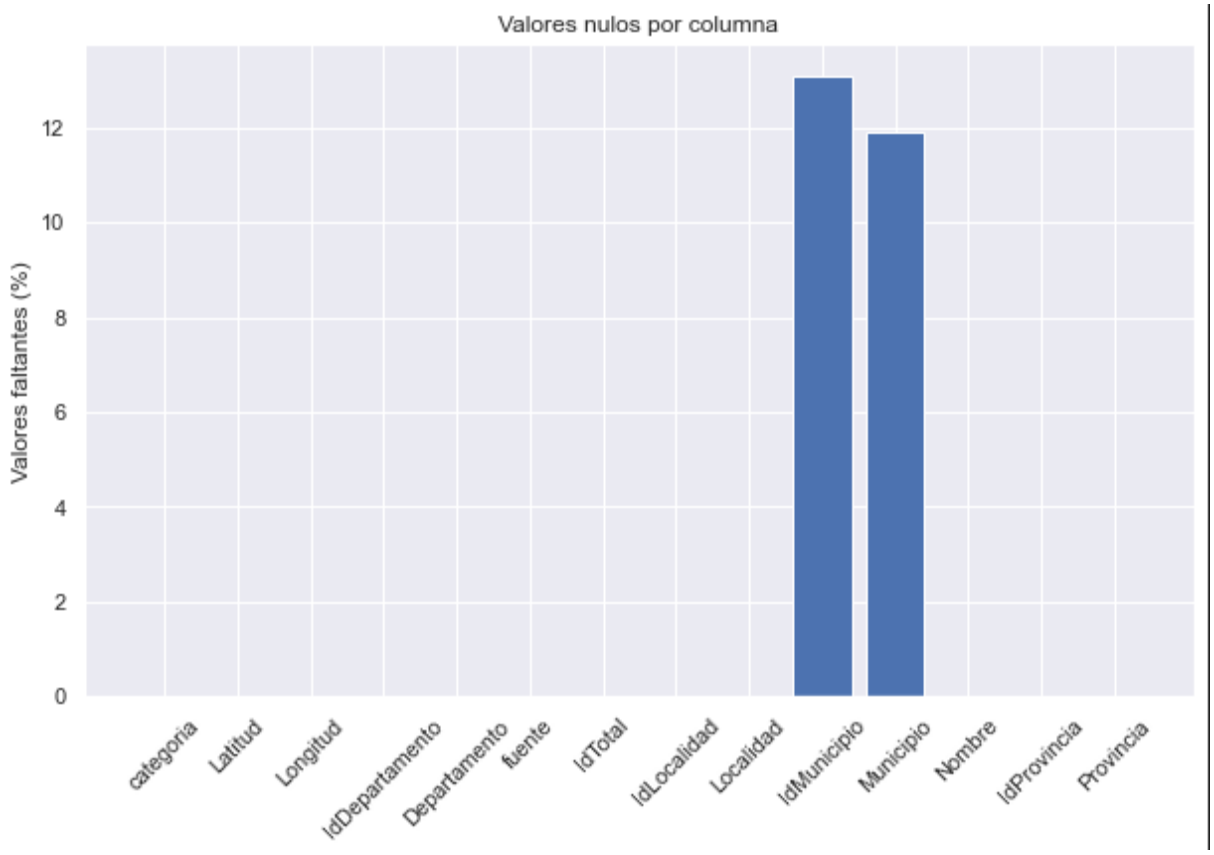


Se observó que aquellos datos nulos quedaron dentro del conjunto de valores erróneos, por lo tanto



-**localidades** (obtenido de 'Localidades.csv')

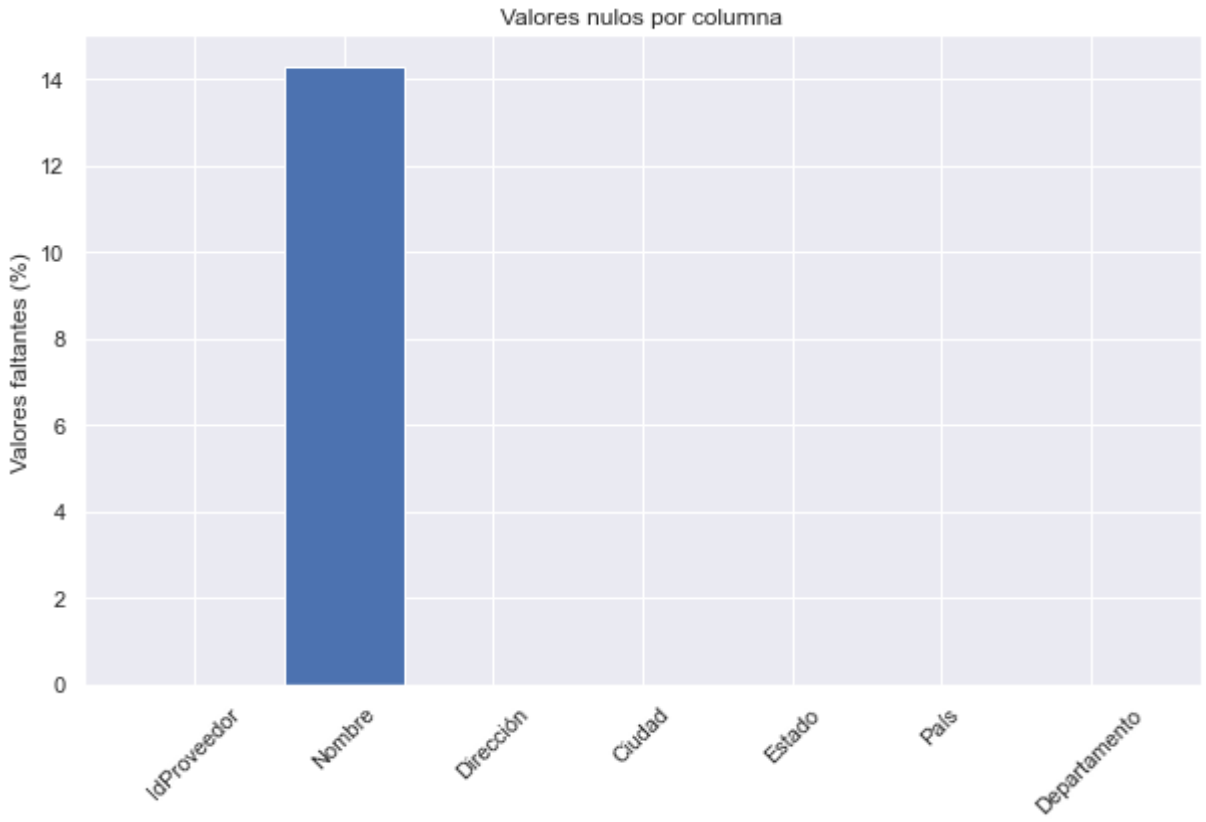
Por lo detallado en la presentación, los valores nulos encontrados fueron.



Las columnas IdMunicipio y Municipio fueron eliminadas por no aportar información relevante.

-**proveedores** (del archivo 'Proveedores.csv')

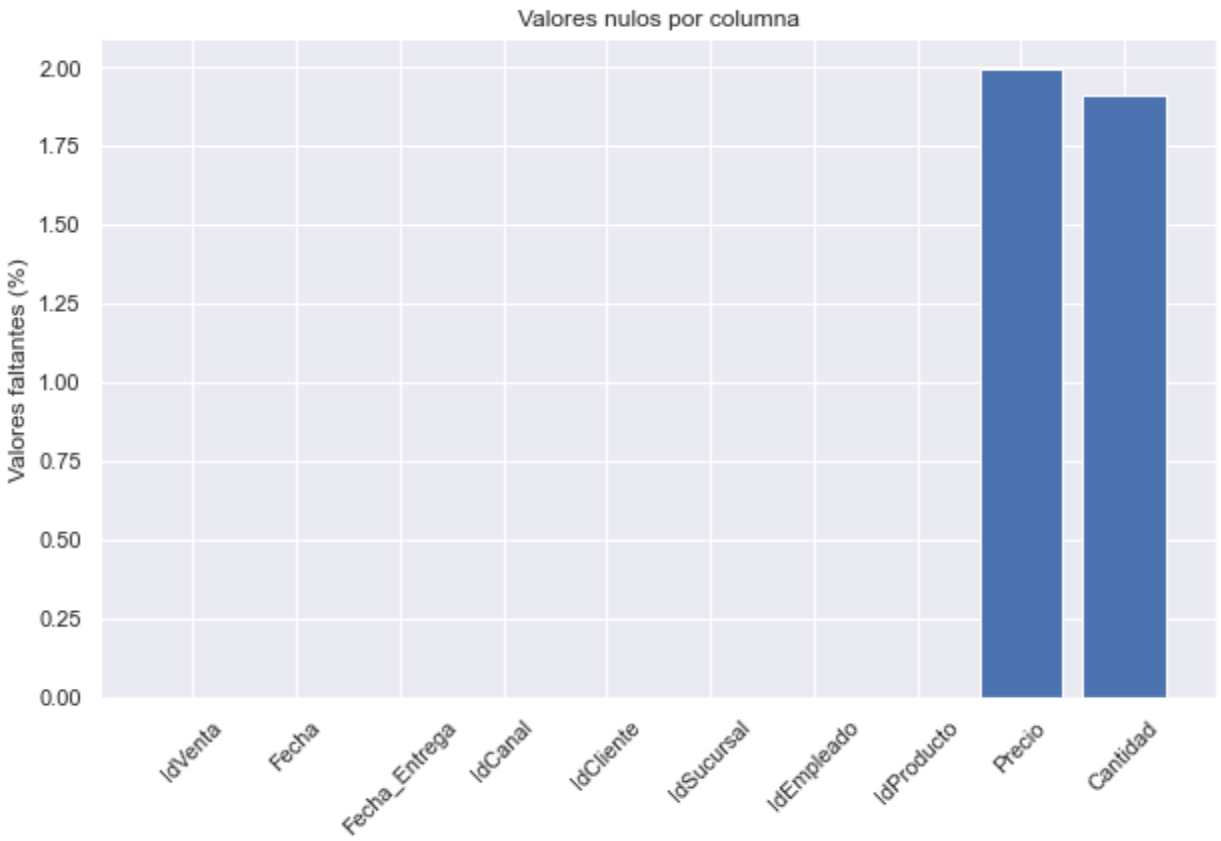
Similar al caso anterior, sólo se cuenta con un par de registros vacíos



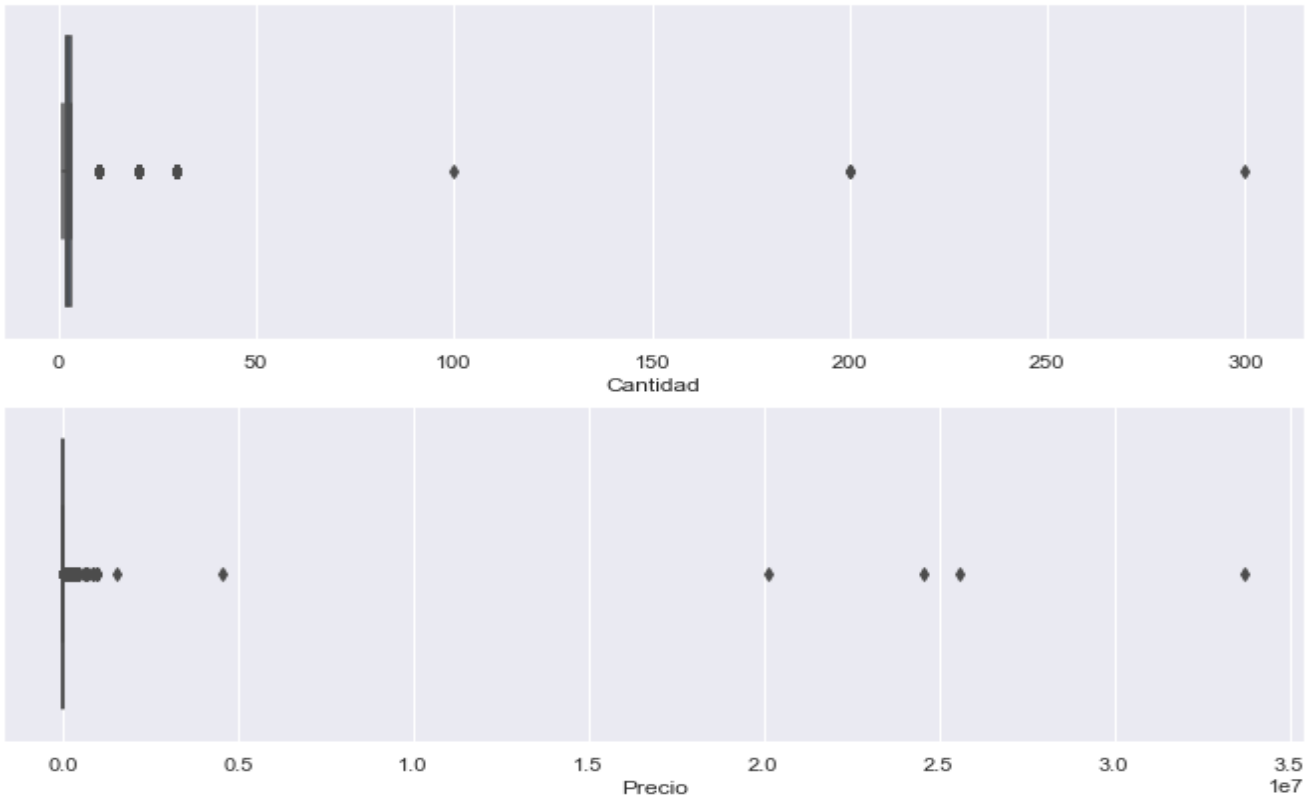
A diferencia del caso anterior, la columna 'Nombre' no fue eliminada por la relevancia del dato, pero aquellos datos nulos fueron cambiados por 'Sin dato'; sin embargo, fue una representación del 14% de datos nulos.

-**venta** (del archivo 'Ventas.csv')

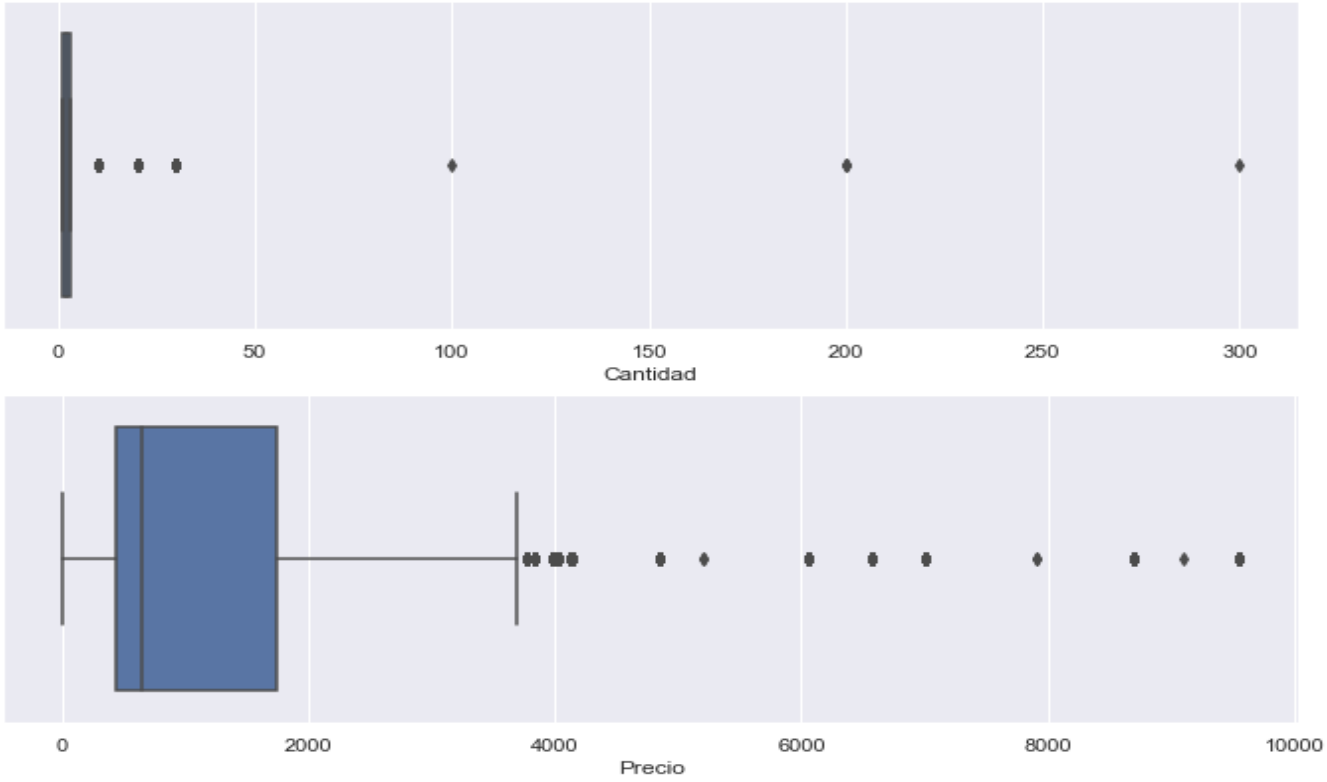
Los valores nulos representaron



Similar al caso del dataframe compra, se realiza un análisis de los valores erróneos utilizando los rangos intercuartílicos para las columnas ‘Precio’ y ‘Cantidad’.



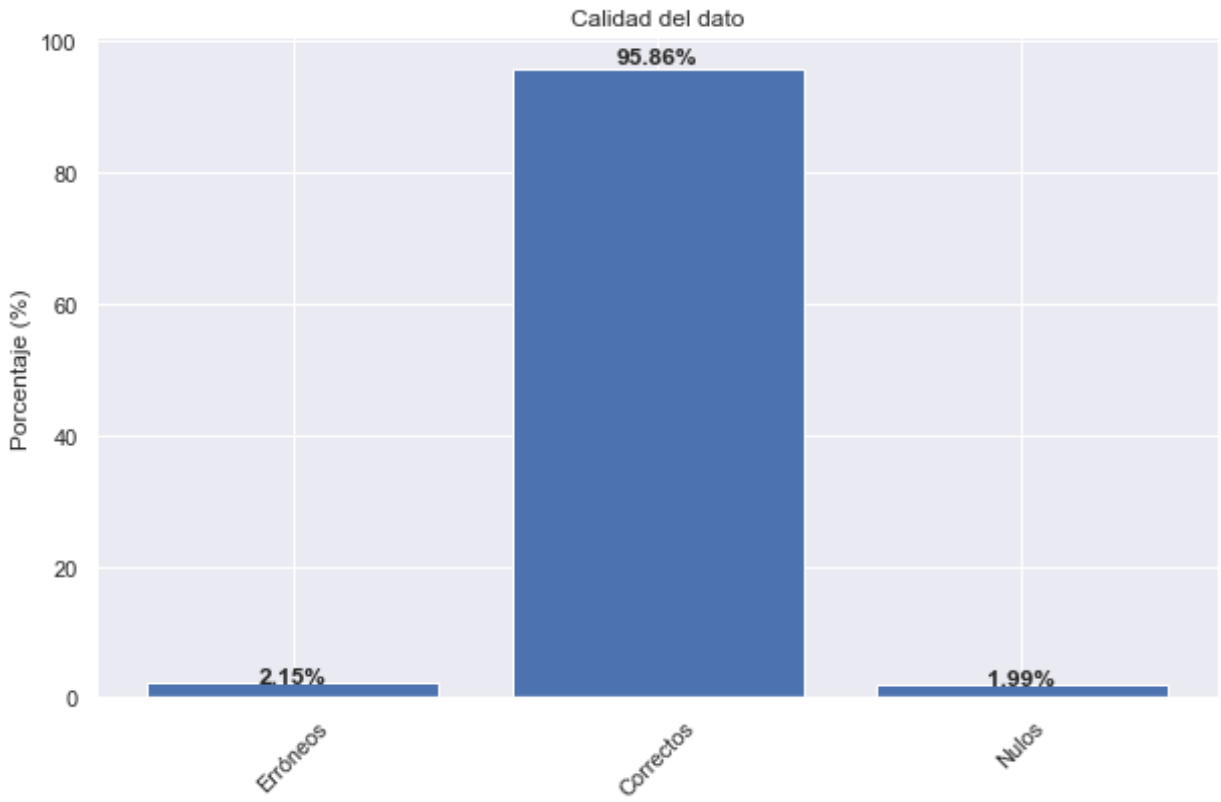
Se consulta de nuevo a la empresa y se nos porporciona el valor del artúclo más costoso (a fin de no eliminarlo en el análisis)



Los valores restantes de Cantidad que fueron nulos se rellenaron con el valor 1 pues existe regitro de venta (tenemos certeza de la transacción) y la cantidad mínima posible para poder vender algún artículo es 1

Por el giro de la empresa, es posible que existan ventas mayoristas de último minuto, por lo que, en conjunto con la tienda, se estableció un margen sobre la cantidad máxima y precio máximo que puede existir en una venta concluyendo que aquellas ventas cuya cantidad supere las 20 pieazas y el valor por producto sea mayor a 400 en moneda local serían eliminadas.

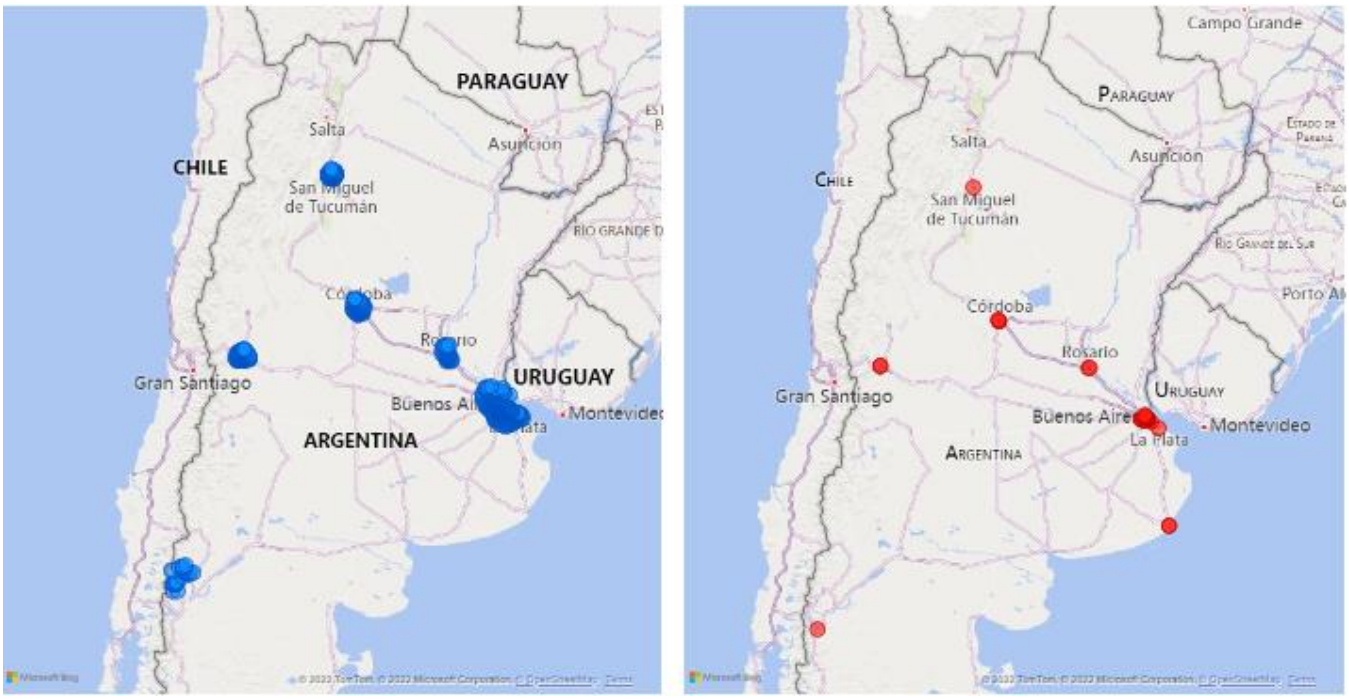
Finalmente, la tabla de nulos, errores y datos correctos quedó de la siguiente manera



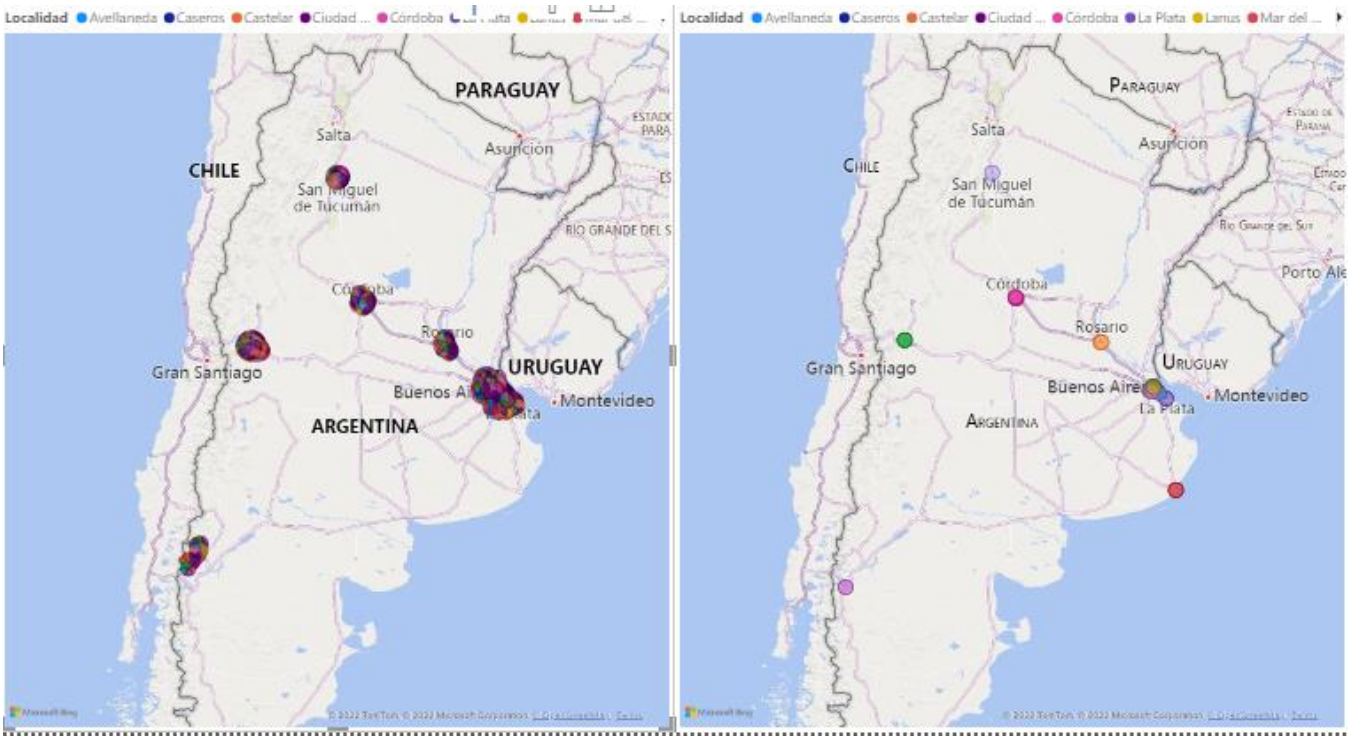
Se concluyó que la calidad de los datos es buena al tener un mínimo de 94% de datos corectos en cada dataframe y se procedió a realizar el análisis con ellos.

Análisis de los datos

Se compara la ubicación de los clientes (azul) con la ubicación de cada local (rojo)

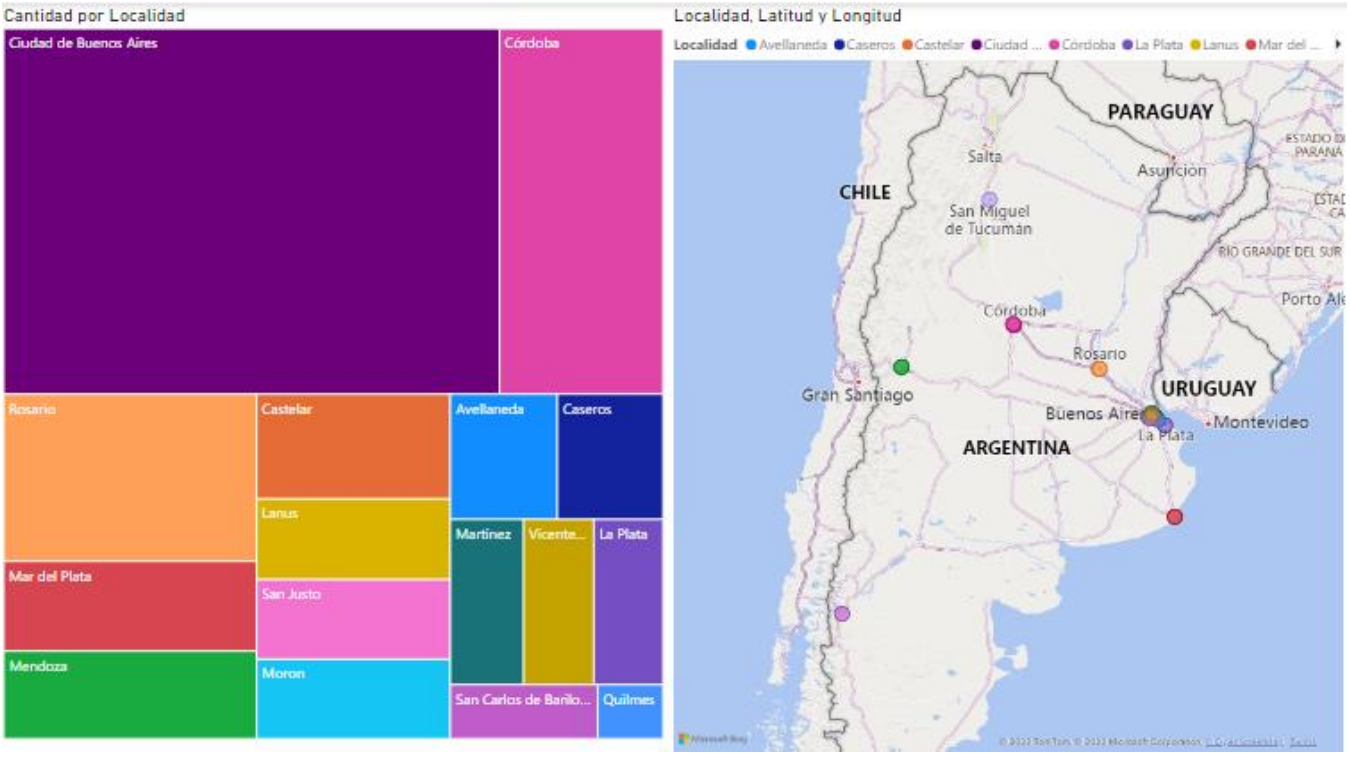


Dividimos cada cliente por las compras realizadas por la(s) sucursal(es) en cada localidad



Es posible apreciar que para un mismo clientes, recibe (realiza compras) en sucursales de diferentes localidades

La cantidad de productos vendidos por localidad queda mostrado a continuación.



Se observó que las localidades de Ciudad de Buenos Aires, Rosario y Córdoba son las localidades donde la mayor cantidad de trabsacciones de productos vendidos fueron realizadas.

KPI´s

Para la toma de decisiones se definieron los KPI´s representatibilidad de la sucursal como:

$$RS(Sucursal) = \frac{Cantidad\ de\ productos\ vendidos\ (sucursal)}{Cantidad\ de\ productos\ vendidos\ total}$$

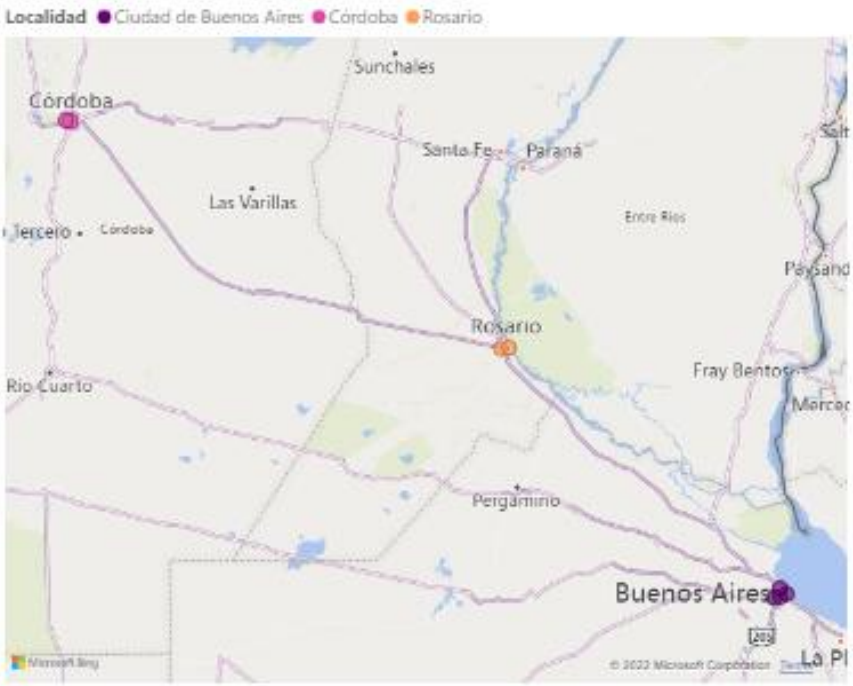
Donde

$$\sum_i RS(Sucursal_i) = 1$$

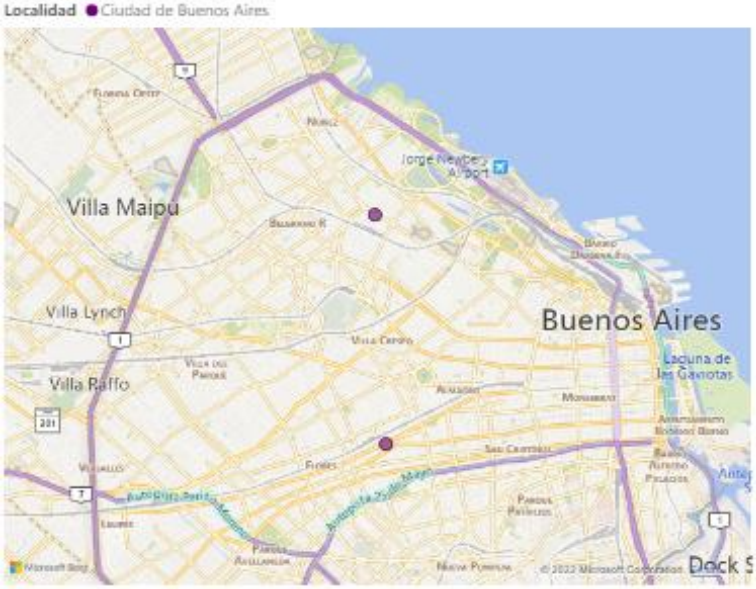
Se determinó entonces que las 8 sucursales con mayor representatibilidad fueron:

1. Sucursal Flores – Id 7 – KPI = 0.07
2. Sucursal Cabildo – Id 1 – KPI = 0.05
3. Sucursal Córdoba Quiroz – Id 26 – KPI = 0.05
4. Sucursal Velez – Id 10 – KPI = 0.05
5. Sucursal Córdoba Centro – Id 25 – KPI = 0.05
6. Sucursal Rosario2 – Id 24 – KPI = 0.05
7. Sucursal Corrientes – Id 4 – KPI = 0.04
8. Sucursal Rosario1 – Id 23 – KPI = 0.04

Estas 8 sucursales representan el 40% de la venta total de la empresa y se distrubuyen como se muestra en el siguiente mapa



Es notable que KPI(Folres) y KPI(Cabildo) son los valores más altos, además se encuentran en la misma localidad



Podría concluirse, en primera instancia colocar la sucursal entre ambos puntos para, en caso de saturación, desahogar las transacciones de ambas sucursales; sin embargo, los altos costos de la renta en la zona dejarían un margen reducido de ganancia.

Es notable también que los últimos valores de KPI son:

1. Sucursal MDQ2 – Id 22 – KPI = 0.014
2. Sucursal Quilmes – Id 19 – KPI = 0.007

La sucursal Quilmes tiene un valor de representación un orden de magnitud menor que el KPI inmediato siguiente y sus transacciones podrían ser absorbidas, sin mayor complicación, por una sucursal de una localidad contigua.

Conclusiones

Después del análisis presentado, se sugiere al negocio que la sucursal a abrir lo haga en la localidad de Rosario (la evaluación de los costos excede el alcance del reporte) convirtiéndose en un centro de distribución para las localidades de Ciudad de Buenos Aires, Córdoba y Rosario. Además de ls dugerencia del cierre de la sucursal Quilmes por baja representatibilidad de sucursal.

Limitaciones del reporte

Es necesario realizar un análisis más exhaustivo en el caso de que existieran nuevas transacciones.

Se necesita un estudio de logística para determinar la razón por la que los clientes obtienen productos de diferentes sucursales aun teniendo que recorrer grandes distancias y con ello buscar una optimización de rutas dependiendo de la ubicación de los productos.

Queda abierta la exploración a nuevos KPI´s que nos brinden una mayor cantidad y más completa información.

Con todo lo anterior se puede utilizar un modelo de ML para determinar, con mayor confianza, la apertura de una nueva sucursal o el cierre de alguna ya existente.