# Informe de calidad de datos

## Proyecto individual I

Franco Pes – 24 de Junio del 2022

# Contenido

# Introducción

El presente informe tiene como objetivo llevar a cabo un profundo análisis de calidad de datos de la compañía, con el objetivo posterior de obtener información relevante para la toma de decisiones de la empresa.

El trabajo se divide en cuatro secciones. La primera sección esta dedicada a explicar brevemente la naturaleza de las tablas y de sus respectivas columnas, mientras que, en la segunda, se lleva a cabo un análisis general de la calidad de lo datos, mostrando su diagnóstico correspondiente. La tercera sección se enfoca en la resolución de los problemas que presenta cada tabla, explicando detalladamente la ejecución de algunas soluciones propuestas. Por último, la cuarta sección se centra en proponer un análisis con el fin de abrir una nueva sucursal de la empresa.

# Diccionario de datos

A continuación se plantea un diccionario de datos, en el cual se describen las columnas de cada tabla:

## Ventas

| Ventas | |
|---|---|
| Idventa | Identificador unico de cada venta |
| Fecha | Fecha se realizacion de venta |
| Fecha_Entrega | Fecha de entrega |
| IdCanal | Identificador unico de cada canal de venta |
| IdCliente | Identificador unico de cada cliente |
| IdSucursal | Identificador unico de cada sucursal |
| IdEmpleado | Identificador unico de cada empleado |
| IdProducto | Identificador unico de cada producto |
| Precio | Precio de venta |
| Cantidad | Cantidad de productos vendidos |

## Compra

| Compra | |
|---|---|
| IdCompra | Identificador unico de cada compra |
| Fecha | Fecha de realizacion de compra |
| Fecha_Año | Año de realizacion de compra |
| Fecha_Mes | Mes de realizacion de compra |
| Fecha_Periodo | Periodo de realizacion de compra |
| IdProducto | Identificador unico de cada producto |
| Cantidad | Cantidad de productos comprados |
| Precio | Precio de productos comprados |
| IdProveedor | Identificador unico de cada proveedor |

## Gasto

| Gasto | |
|---|---|
| IdGasto | Identificador unico de cada gasto |
| IdSucursal | Identificador unico de cada sucursal |
| IdTipoGasto | Identificador unico de cada tipo de gasto |
| Fecha | Fecha de realizacion de gasto |
| Monto | Monto de compra |

## Proveedores

| Proveedores | |
|---|---|
| IDProveedor | Identificador unico de cada proveedor |
| Nombre | Nombre de cada proveedor |
| Address | Dirección de cada proveedor |
| City | Ciudad del proveedor |
| State | Provincia del proveedor |
| Country | Pais del proveedor |
| departamen | Departamento del proveedor |

## Sucursales

| Sucursales | |
|---|---|
| ID | Identificador unico de cada sucursal |
| Sucursal | Nombre de cada sucursal |
| Direccion | Dirección de cada sucursal |
| Localidad | Localidad de cada sucursal |
| Provincia | Provincia del sucursal |
| Latitud | Latitud geográfica de sucursal |
| Longitud | Longitud geográfica de sucursal |

## Localidades

| Localidades | |
|---|---|
| categoria | Categoria de cada localidad |
| centroide_lat | Latitud de cada localidad |
| centroide_lon | Longitude cada localidad |
| departamento_id | Identificador unico de cada departamento |
| departamento_nombre | Nombre del departamento |
| fuente | Fuente de extraccion |
| id | Identificador unico |
| localidad_censal_id | Identificador unico de cada localidad censal |
| localidad_censal_nombre | Nombre de cada localidad censal |
| municipio_id | Identificador unico de cada municipio |
| municipio_nombre | Nombre de cada municipio |
| nombre | Nombre de cada municipio |
| provincia_id | Identificador unico de cada provincia |
| provincia_nombre | Nombre de cada provincia |

## Clientes

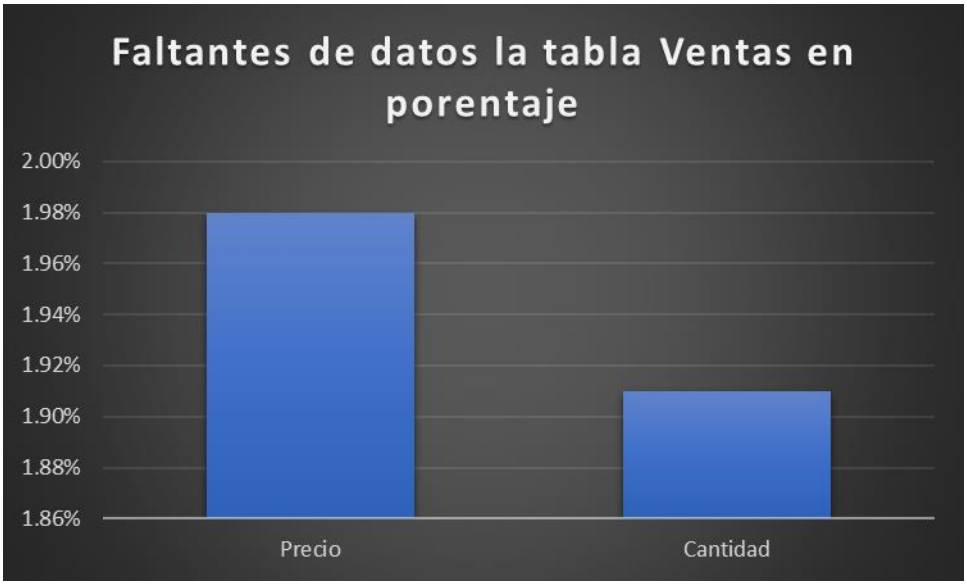
| Clientes | |
|---|---|
| ID | Identificador unico de cliente |
| Provincia | Nombre de la Provincia |
| Nombre_y_Apellido | Nombre del cliente |
| Domicilio | Domicilio del cliente |
| Telefono | Telefono del cliente |
| Edad | Edad del cliente |
| Localidad | Localidad |
| x | Latitud |
| y | Longitud |
| col10 | Columna vacia |

# Diagnóstico de cada tabla

Esta sección se enfoca principalmente en encontrar y cuantificar los problemas generales a resolver que presenta cada tabla:

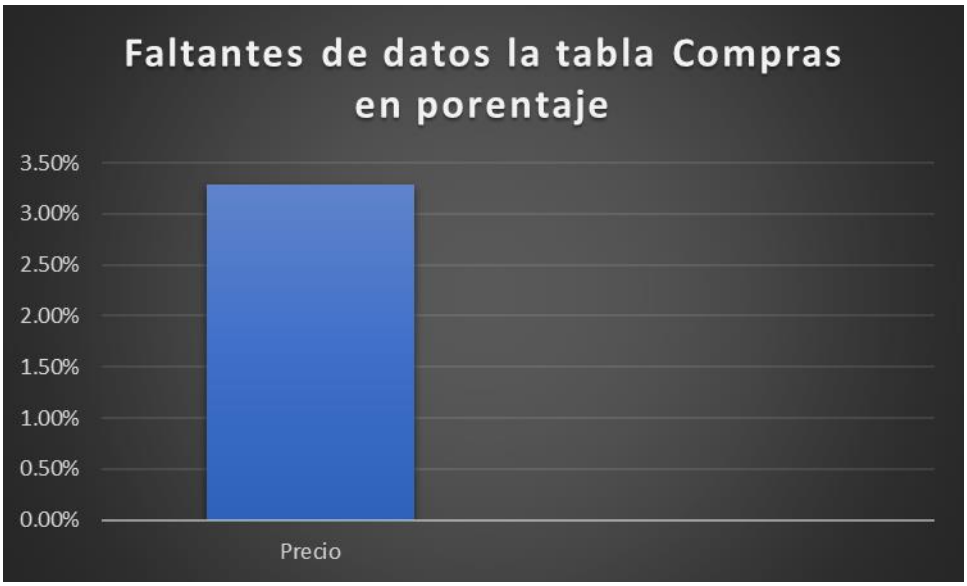
## Ventas

- Valores faltantes en columna Precio y Cantidad
- Tipo de dato inadecuado en la columna Fecha
- Errores de tipeo en la columna Precio (Outliers)



## Compra

- Columnas posiblemente redundantes (Fecha_Año, Fecha_Mes, Fecha_Periodo)
- tipo de dato incorrecto en columna Fecha
- Falta de datos en la columna Precio
- Posibles outliers en la columna precio




## Gasto

- tipo de dato incorrecta en la columna Fecha

## Proveedores

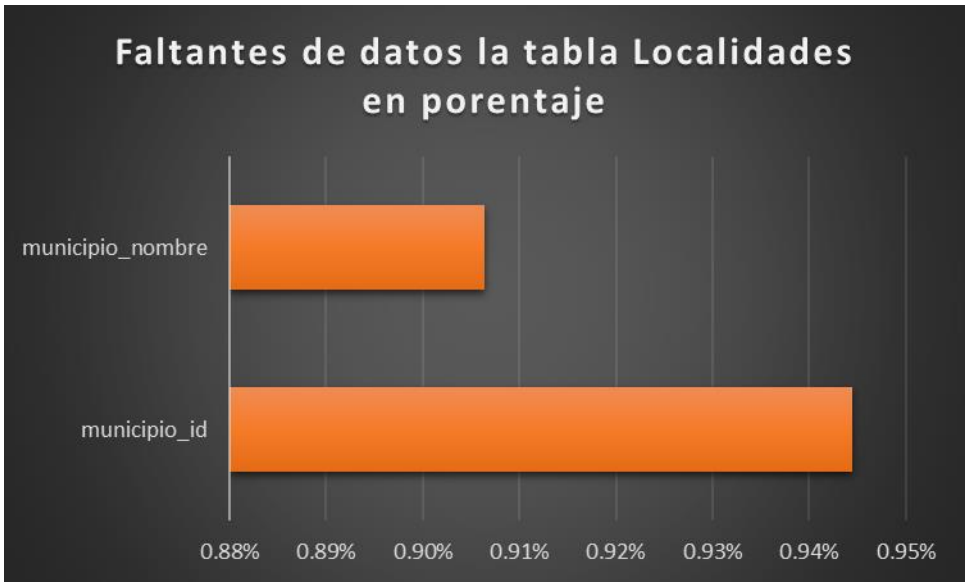
- Nombre de la columna de identificador único de proveedores no normalizado
- Columnas en inglés (no adecuado)
- columna ‘Country’ ya que resulta redundante
- Faltante de datos en la columna nombre
- Nombres de las provincias en la columna State no normalizados

## Sucursales

- nombre de la columna de identificador único de proveedores no normalizado
- nombres de algunas localidades y Provincias no normalizadas

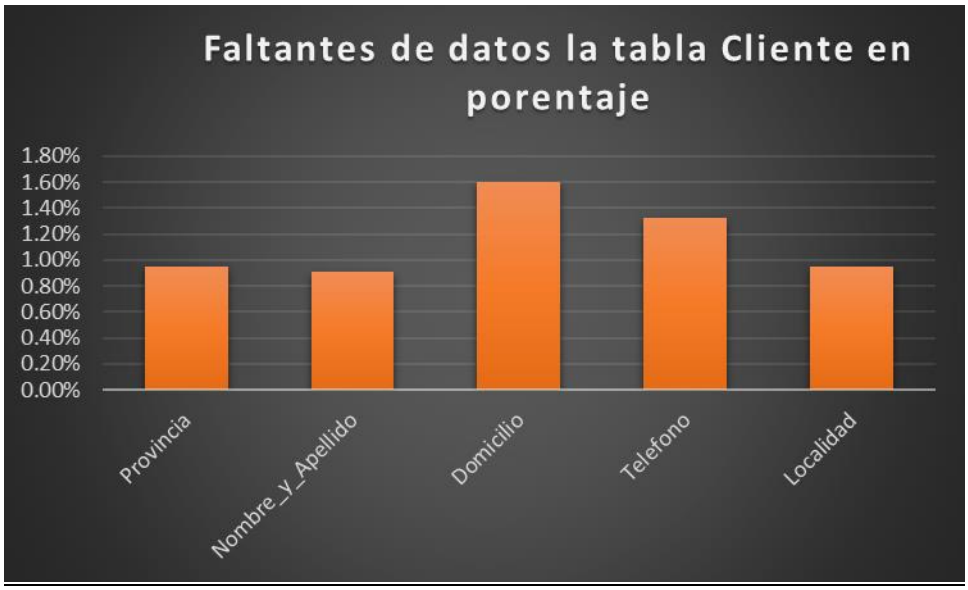
## Localidades

- nombre de todas las columnas con nombres inadecuados
- columna nombre es redundante
- columna IdMunicipio con valores faltantes
- Faltante de datos en la columna Municipio_Nombre
- Columnas con orden inadecuado




## Clientes

- columnas redundantes (Provincia, X, Y)
- nombre de la columna de identificador único no normalizado
- Faltante de datos



# Resolución de problemas

Esta sección se enfoca en la resolución de los problemas descritos en la sección anterior y en las posibles soluciones de estos.

## Modificación del nombre de las columnas

Se modifica el nombre de columnas con nombres inadecuados o no normalizados:

| Tabla | Columna anterior | Columna Modificada |
|---|---|---|
| Cliente | ID | Idcliente |
| Cliente | Nombre_y_Apellido | Nombre |
| Sucursales | ID | IdSucursal |
| Proveedores | Adress | Direccion |
| Proveedores | City | Ciudad |
| Proveedores | State | Provincia |
| Proveedores | Country | Pais |
| Proveedores | Department | Localidad |
| Proveedores | IDproveedor | IdProveedor |
| Localidades | departamento_id | IdDepartamento |
| Localidades | departamento_nombre | Departamento |
| Localidades | municipio_id | IdMunicipio |
| Localidades | municipio_nombre | Municipio |
| Localidades | provincia_id | IdProvincia |
| Localidades | provincia_nombre | provincia |
| Localidades | centroide_lat | Latitud |
| Cliente | centroide_lon | Longitud |

## Eliminación de columnas redundantes

Se eliminan aquellas columnas que, en principio, no aportaran valor para un análisis posterior:

| Tabla | Columna eliminada |
|---|---|
| Clientes | X |
| Clientes | Y |
| Clientes | Col10 |
| Clientes | Provincia |
| Localidades | Nombre |
| Proveedores | Country |
| Localidades | categoria |

## Rellenado de columnas

Se rellenan aquellas columnas con datos tipo string con la cláusula “Sin dato”:

| Tabla | Columna rellenada |
|---|---|
| Clientes | Nombre_y_apellido |
| Clientes | Domicilio |
| Clientes | Telefono |
| Clientes | Localidad |
| Localidades | Municipio |
| Localidades | Departamento |
| Proveedores | Nombre |

## Normalización de nombre de Provincias y localidades

A modo de ejemplo, se exponen las variantes de Buenos Aires (Provincia) que se observaron en la columna Provincia de la tabla Sucursales:

| Variantes de Buenos Aires(Provincia) | | |
|---|---|---|
| Ciudad de Buenos Aires | CABA | C deBuenos Aires |
| Bs As | Bs. As. | B. Aires |
| Buenos Aires | B.Aires | Provincia de Bs AS. |

Luego, se procede a normalizar los nombres de algunas provincias en columnas de las tablas Sucursales y Proveedores:

| Tabla | Columna | Normalizacion |
|---|---|---|
| Sucursales | Provincia | Buenos Aires |
| Sucursales | Provincia | Ciudad de Buenos Aires |
| Sucursales | Localidad | Córdoba |
| Sucursales | Localidad | Ciudad de Córdoba |
| Proveedores | Provincia | Buenos Aires |
| Proveedores | Provincia | Ciudad de Buenos Aires |

## Normalización de strings

Se procede a normalizar los campos de todas las columnas string, de todas las tablas con el método title() de Python, el cual les da un formato de título a todos los campos, es decir que coloca en mayúscula a la primer letra de cada palabra, y en minúscula al resto. A continuación, se muestra las columnas afectadas con este método:

| Tabla | Columna normalizada (title) |
|---|---|
| Cliente | Nombre_y_apellido |
| Cliente | Domicilio |
| Cliente | Localidad |
| Proveedores | Provincia |
| Proveedores | Departamento |
| Proveedores | Localidad |
| Proveedores | Direccion |
| Proveedores | Nombre |
| Localidades | Provincia |
| Localidades | Municipio |
| Localidades | Departamento |
| Sucursales | Sucursal |
| Sucursales | DIreccion |
| Sucursales | Localidad |
| Sucursales | Provincia |

## Búsqueda de Outliers y solución

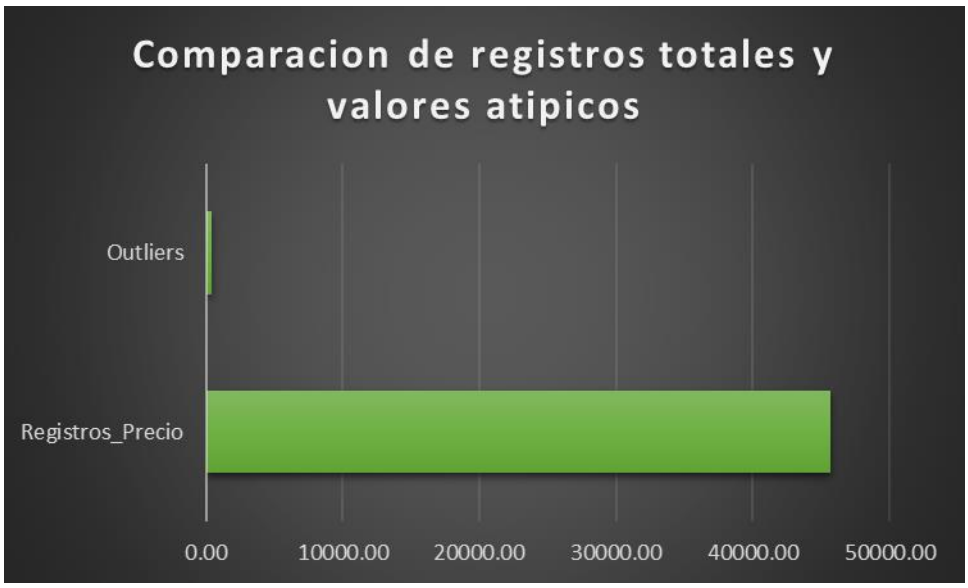
A continuación, se muestra una tabla donde se evidencia la naturaleza de los outliers de Precio en la tabla venta, lo cual puede darnos pistas para luego resolver los mismos:

| IdVenta | Fecha | Precio |
|---|---|---|
| 60 | 18/4/2017 | 1282,82 |
| 308 | 15/7/2019 | 1282,82 |
| 994 | 18/4/2017 | 1282,82 |
| 1250 | 10/3/2017 | 1282,82 |
| 2078 | 10/3/2017 | 1282,82 |
| 48241 | 20/11/2020 | 1282,82 |
| 36371 | 27/4/2016 | 128282 |

Esta tabla corresponde a una muestra de la tabla venta, para un producto en particular (idproducto = 42810). Nótese que la mayoría se vendieron a $1282.82, excepto por uno, el cual se vendió a $128282, lo cual sugiere que los outliers vienen dados por errores de tipeo al ingestarlos, posiblemente confundiendo punto por coma.

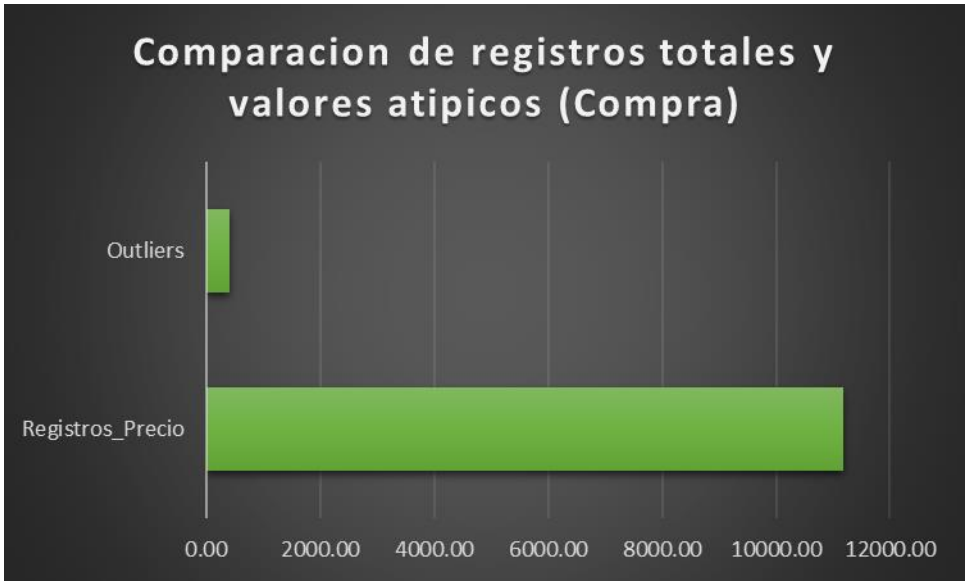
Con lo cual, para solucionar este problema se crea una columna llamada “outliers”, donde se asigna un valor 0 si el precio es mayor al promedio agrupado por idproducto, más dos desvíos (agrupados por

producto), y uno si es menor (No outlier). Como resultado, se encuntran 396 valores correspondientes a valores atípicos.




Luego, se procede a extraer los valores atípicos de precios, multiplicando la columna Outliers por Precio, obteniendo así una columna llamada “Precio2”. Luego, se toma la decisión de eliminar los valores nulos y los valores donde Precio2 es cero, de esta forma se eliminan los outliers. Cabe aclarar que existe una alternativa a esta decisión, la cual consiste en imputar el promedio por idproducto en los valores faltantes, o bien usar el algoritmo k-means. Por último, se renombra la tabla Precio2 a Precio.

Se lleva a cabo exactamente el mismo procedimiento, pero con la columna precio de la tabla Compra. Se obtiene un total de 530 valores atípicos, los cuales se proceden a eliminar.



## Automatización de la ingesta de archivos CSV

Se plantea la ingesta automática de archivos, a modo de facilitar el proceso de ETL en un futuro, a medida que los datos de la empresa se incrementen. Para ello, se creo un script de Python en el cual el proceso de ingesta y de limpieza de datos (sección 2 y 3), se efectúa automáticamente, simplemente ejecutando dicho archivo.

Basta con añadir cada archivo csv de cada tabla a la carpeta con su mismo nombre. Cabe aclarar que los archivos csv de cada tabla deben llamarse de la misma manera, pero incrementando un numero al final del mismo. A modo de ejemplo, se presenta el siguiente cuadro:

| Carpeta "Cliente" | |
|---|---|
| Archivo 1 | cliente1.csv |
| Archivo 2 | cliente2.csv |
| Archivo 3 | cliente3.csv |
| Archivo 4 | cliente4.csv |

Por último, se plantea para trabajos posteriores automatizar la ingesta de los datos “limpios” en una base de datos SQL.

# Estudio de apertura de nueva sucursal

## Problemática

Se desea conocer cuál es la ubicación optima de una nueva sucursal. En principio, es deseable abrir una nueva sucursal en áreas no cubiertas, es decir, áreas donde se observen una gran cantidad de ventas y las mismas no existan sucursales cerca pero si muchas ventas, y de esta manera plantear una posible locación. Posteriormente, una vez planteada la locación, debe hacerse un análisis de rentabilidad para determinar si los beneficios de abrir una nueva sucursal superan a los costos.

## Posible solución

Para determinar la posible locación de la nueva sucursal, se plantea usar principalmente la tabla ventas concatenada con la tabla clientes. Lo que nos interesa de la tabla clientes son las coordenadas geográficas de lo clientes, para llevar a cabo un análisis de clusters con la locación de las ventas realizadas.

Lo que nos interesa es encontrar regiones de ventas “densas”, con lo cual es conveniente realizar un análisis de densidad de latitud y longitud de ventas usando el algoritmo DBSCAN, distingue regiones densas. Una vez encontradas las regiones densas, debe realizarse un análisis mas exhaustivo sobre las regiones densas. Con lo cual, se sugiere tomar las regiones densas (se intuye que Buenos aires, Santa fe, Córdoba, etc. son regiones densas, pero no se descartan otras) y sobre estas aplicar el algoritmo k-means y verificar que los centroides efectivamente coincidan con las sucursales existentes. Si el modelo k-means sugiere un centroide donde no hay ninguna sucursal, entonces es posible tomar dicho centroide como posible locación de la nueva sucursal.

Posteriormente, corresponde realizar un estudio meramente económico sobre la posible nueva locación. Es decir que deben estimarse tanto lo beneficios como los costos de abrir esta nueva sucursal. Entre los costos, se destacan los costos de construcción, los nuevos gastos que afrontara la empresa en materia de costos operativos y resulta conveniente estudiar los costos de transporte de traslado de mercadería de los proveedores hasta la nueva sucursal. Por otro lado, deben estimarse los beneficios, como el aumento de las ventas presenciales o la disminución de los costos de transporte de la mercadería. Cabe aclarar que si bien los algoritmos DBSCAN y K-means nos darán una buena idea de la locación, la decisión final debe basarse en un análisis meramente de rentabilidad de la nueva sucursal.