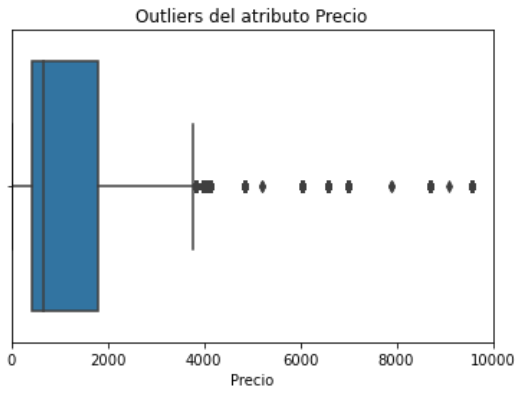
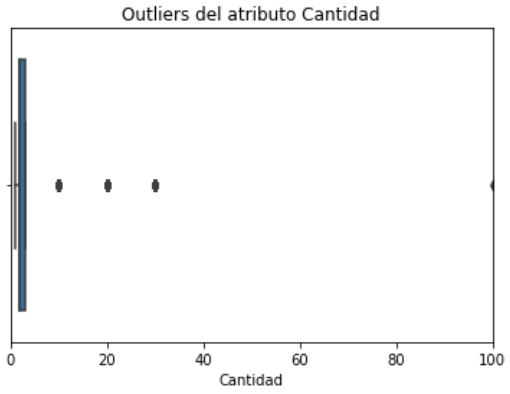
REPORTE DE CALIDAD DE LOS DATOS

VENTA

Al analizar los features cuantitativos ‘Precio’ y ‘Cantidad’ se llega a las siguientes conclusiones. Ambos contienen outliers y valores faltantes. En el caso de ‘Precio’, los outliers son 2476 y representan un 5,36% del total de registros. Mientras que en ‘Cantidad’ los valores faltantes son 910 y representan un 1,97%.

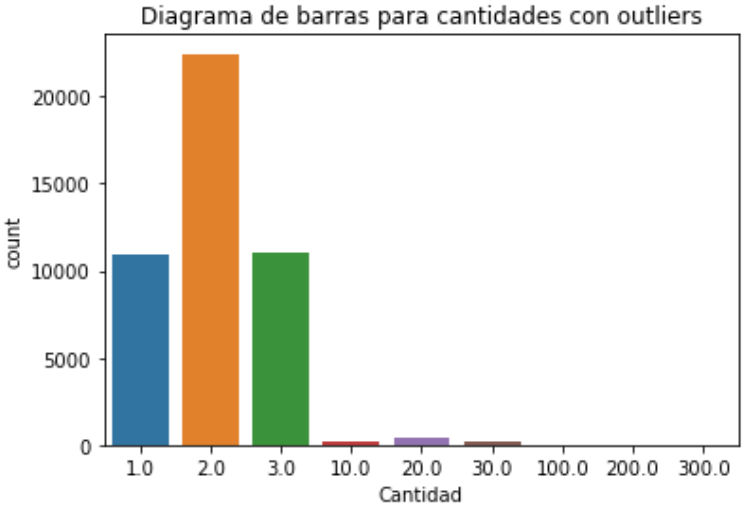




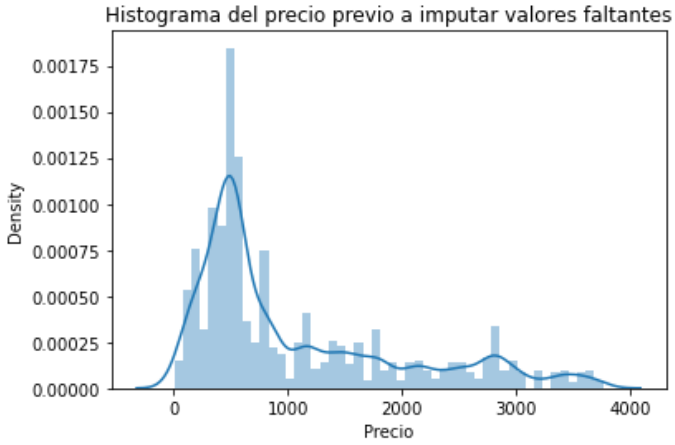
En las gráficas supra se pueden observar los diagramas de caja para ambas variables con sus respectivos outliers. Téngase en cuenta que los ejes ‘x’ están acotados para facilitar su visualización, ya que hay valores muy extremos que dificultan la representación gráfica (en cantidad se estableció un límite 100 para el eje horizontal, pero el máximo valor encontrado es 300).

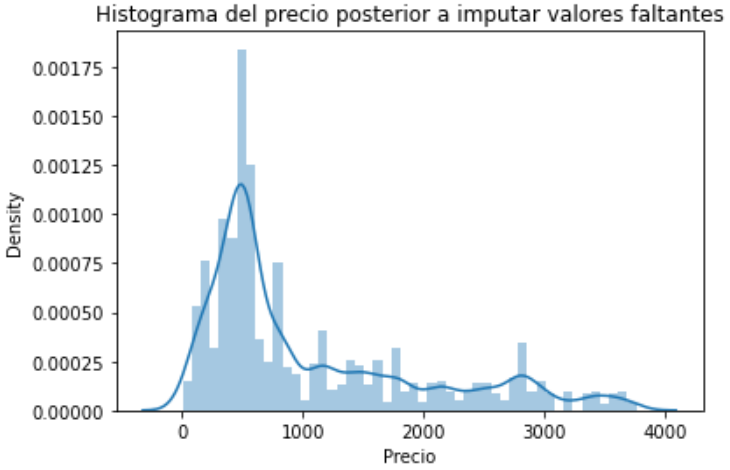
A partir de esta situación, y para evitar la distorsión que suponen unos outliers tan extremos en nuestro conjunto de datos, se procede a realizar la eliminación de los precios que se encuentran por encima del límite superior a partir de la técnica del rango intercuartílico.

Distinto es el caso de los valores extremos de las cantidades, donde se pudo observar que tenemos un error de tipeo generando esos valores atípicos. Todos los outliers son múltiplos de 10 o 100 (tenemos 3 cantidades de 300 y 1 de 100 y 200 que el gráfico de barras no los llega a captar visualmente). Se procede a llevar estas instancias a valores 1, 2 y 3.



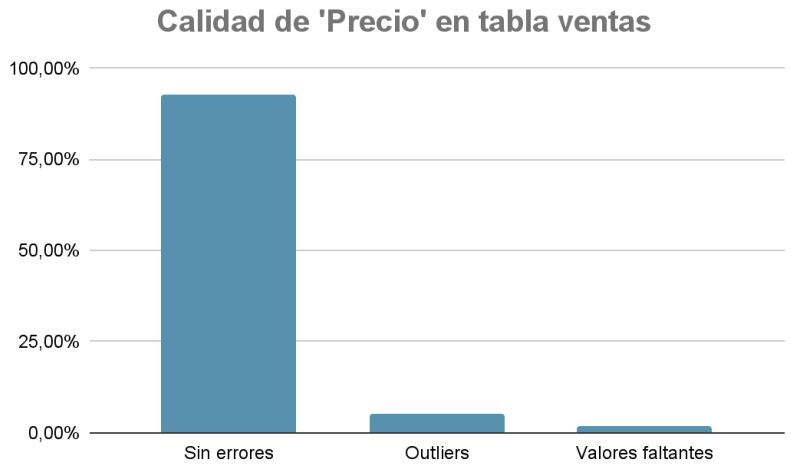
Para el caso de los valores faltantes de ‘Cantidades’, se imputa con la moda (el valor es 2). Esto no modifica la tendencia de nuestros datos. Al considerar los faltantes en precios, se crea una tabla de venta auxiliar con identificadores únicos de los productos y sus precios. De esa manera, luego se hace un merge con la tabla venta original y se reemplaza el valor de precio allí donde es nulo. Como se puede apreciar a continuación, se mantiene inalterable la distribución de la variable ‘Precio’ a posteriori de realizar la imputación. Eso es esencial en todo proceso de preprocesamiento de datos a la hora de lidiar con valores nulos.

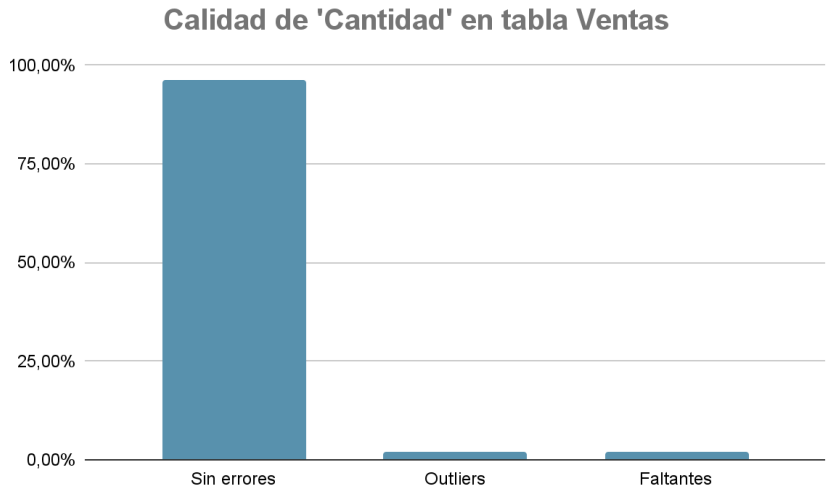




En cuanto a las columnas del dataset, se elige mantener todos sus campos ya que nos aportan información importante y ninguno se puede desestimar. Para este caso, la nomenclatura de los identificadores (IdSucursal, IdCliente, etc.) están correctos y se trabajarán con esa misma denominación en las otras tablas.

Para finalizar el tratamiento de los datos en este dataset, hay que considerar el tipo de dato de cada columna. Por eso se toma la decisión de convertir los campos que contienen fechas (Fecha y Fecha_entrega) y están como string al tipo de dato datetime. Para finalizar, el campo cantidad estaba en flotante y se modifica al tipo de variable entero.





COMPRA

De lo enunciado previamente, aquí nuevamente se convierte la columna que brinda la información de la fecha de compra al tipo de dato correspondiente.

Primeramente, y antes de avanzar en las incongruencias encontradas en este dataset, se debe señalar que se eliminará la columna Fecha_periodo. Esta decisión se sustenta en que este atributo se compone de otros dos ya presentes (Fecha_año y Fecha_mes) y no nos aporta mayor utilidad en nuestros datos.

Respecto a los nombres de cada columna, se mantienen los originales ya que son lo suficientemente descriptivos de los valores que contienen.

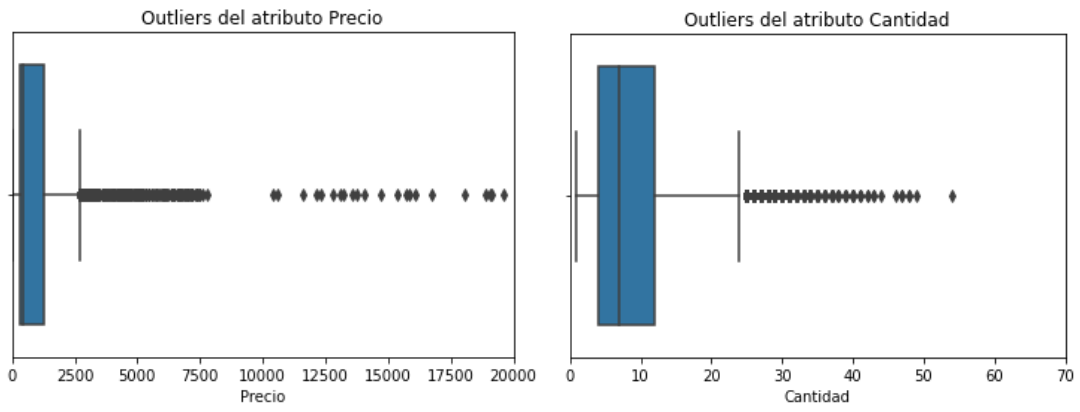
Pasando al estudio de las variables de nuestro dataset, solo en una se observan valores faltantes. Es el caso de ‘Precio’, con un 3,18% del total de datos. Se considera que eliminarlos podría acarrear un problema a la hora de extraer información de nuestros datos, es

por ello que esa estrategia será desestimada. Ahora bien, se analizaron diferentes escenarios con sus respectivas implicancias para el tratamiento de estos valores faltantes. A saber: en un primer momento, se consideró que el valor a imputar en estos registros faltantes debería debatirse con los responsables del área de compras, asumiendo de que lógicamente el valor del precio estaría supeditado a una fecha y un proveedor determinado. En esta línea, se consideró que no contar con un dataset de productos hace a esta tarea de imputación más dificultosa -ténganse en cuenta que la tabla de Compra tiene una columna de 'IdProducto', la cual se podría traer para imputar el valor faltante en el campo Precio-.

Luego de seguir explorando el dataset e intentando emplear una estrategia que no implicase la eliminación de estos faltantes, se avanzó un paso más para intentar responder básicamente a una pregunta central en todo esta problemática: ¿Por qué tenemos valores nulos en ciertas instancias? Después de tratar de responder esta pregunta basándonos en los paradigmas clásicos de MCAR, MAR y MNAR, se llega a la conclusión de que la nulidad en esas celdas responde pura y exclusivamente a una cuestión aleatoria. Es decir, no responde a relaciones que se puedan establecer con otras columnas, ni tampoco a la variable que se quiere medir. La probabilidad de tener un valor faltante es exactamente la misma para todas las instancias.

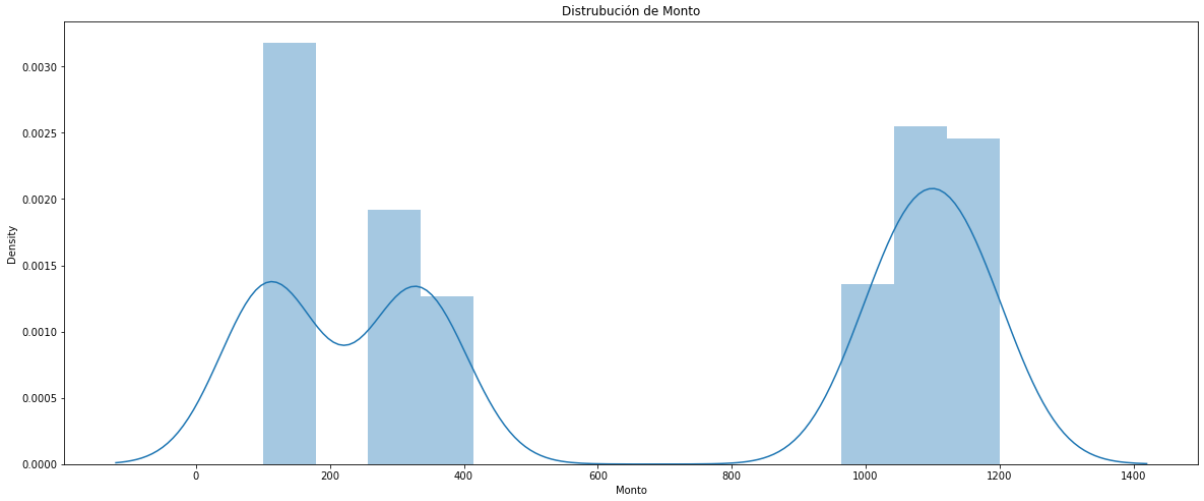
Al llegar a esa conclusión, se sumó el hecho de que las distribuciones de los precios para los IdProducto -independientemente del proveedor que nos lo venda- se manejan en rangos bastante acotados y con una baja desviación estándar. Tampoco, como se había aseverado en esa primera aproximación al problema, se observaron modificaciones de los precios por fecha. En síntesis, se diseñó una estrategia que nos devolviera el precio promedio de cada IdProducto, ya que con todas las consideraciones previas podemos garantizar que ese valor es representativo.

Respecto a los valores atípicos, se realiza el estudio tanto para 'Precio' como para 'Cantidad'. En cuanto a la primera, tenemos un 5.19% de outliers. A simple vista, se puede observar que esto es un error propio de los datos, ya que hay valores de precio que son notoriamente excesivos. Para el caso de 'Cantidad', el porcentaje de valores extremos se ubica en el 3%. Sosteniendo la técnica utilizada en la tabla de Venta para el atributo 'Precio', aquí emplearemos también la regla del rango intercuartílico para la eliminación de estos outliers.



GASTO

En esta tabla vale señalar que no se presentan valores faltantes. Asimismo, el feature de ‘Monto’ no tiene ningún valor atípico. Respecto a esta variable, se puede destacar que su distribución encuentra la mayoría de los datos en los valores del rango comprendido entre 150-400 y 1000-1200.

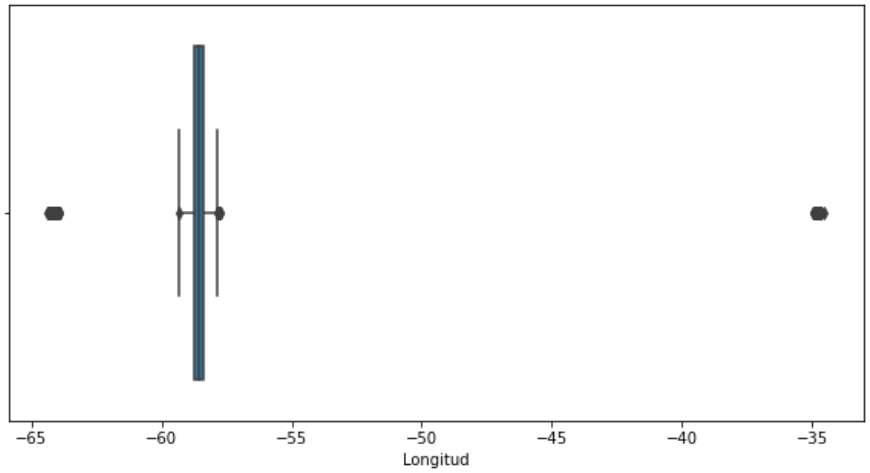


CLIENTE

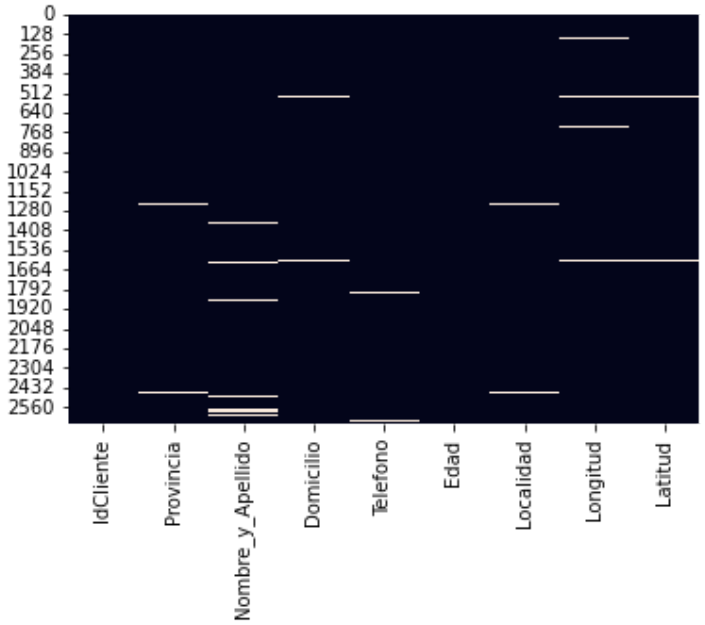
En una primera instancia, eliminamos la columna ‘col10’ que no vamos a utilizar. Luego, renombramos la columna ‘ID’ por ‘IdCliente’ para mantener la consistencia con las otras tablas. Hacemos lo mismo para ‘X’ e ‘Y’ cambiando por ‘Longitud’ y ‘Latitud’ respectivamente. Para estas dos columnas hay que hacer un tratamiento especial ya que se observan muchas inconsistencias. Por una parte, hay que verificar que todas las instancias sean negativas y, caso contrario, cambiar de signo. Hay que señalar en este punto que se debió realizar un reemplazo de ‘,’ por ‘.’ y posteriormente convertir ambas columnas al tipo de dato flotante. A su vez, se aprecia que muchos registros claramente tienen intercambiados los dos campos. Es decir, algunos clientes tienen en el campo latitud su coordenada de longitud y viceversa. Para reemplazarlos, se crea una estructura condicional según la latitud y longitud de Argentina y cada provincia. Cuando se corrobora que la latitud y la longitud están ingresadas en el campo contrario, se realiza su intercambio.

A partir de un riguroso trabajo buscando las localidades, tomando la precaución de que el método no cometa errores, se encontraron 35 registros donde las coordenadas están intercambiadas.

Otra forma en la que uno podría realizar esta modificación sería con un boxplot, poniendo un umbral en torno -55 de Longitud para establecer que lo que sea mayor corresponde en realidad a Latitud.



Luego queda considerar los valores faltantes que encontramos en este dataset de clientes. A continuación se pueden observar en blanco aquellos registros que contienen valores nulos:



En cuanto a los campos Nombre_y_Apellido, domicilio y teléfono de los clientes que están vacíos, eliminarlos sería un error porque estaríamos quedándonos sin esos clientes en nuestra base de datos. Lo más importante dentro de los faltantes están en los campos geográficos. Es decir provincia, localidad, latitud y longitud. Para resolver esta situación, se decide conectar a la API del Estado argentino que realiza normalización de los campos a partir de información en los otros. Esto quiere decir que a partir de una de las funciones provistas se puede inferir la localidad y provincia del cliente con sus coordenadas. Para los clientes que tienen faltantes en sus coordenadas pero poseen su localidad, se coloca en latitud y longitud los centroides.

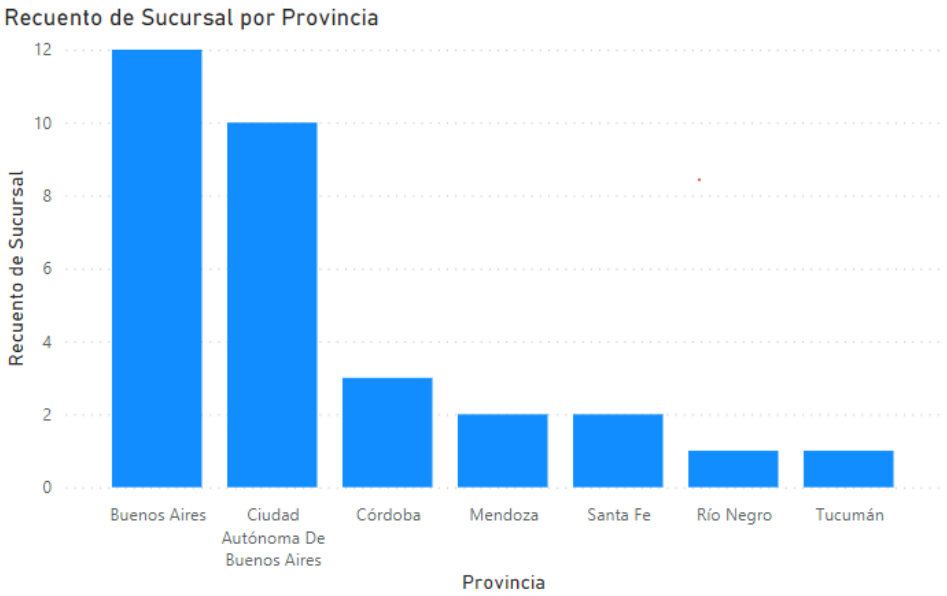
PROVEEDORES

En esta tabla, resulta menester proceder a la normalización de sus campos. En vistas a esta tarea se recurre a la API de Argentina que corrobora la precisión de los valores ingresados según sus propios registros. Luego, se normalizan los registros para que queden debidamente con mayúscula inicial en cada una de las columnas de tipo texto. A su vez, se renombran los campos para que quede más clarificada la información que nos aportan cada uno de ellos.

SOLICITUD DE APERTURA DE NUEVA SUCURSAL

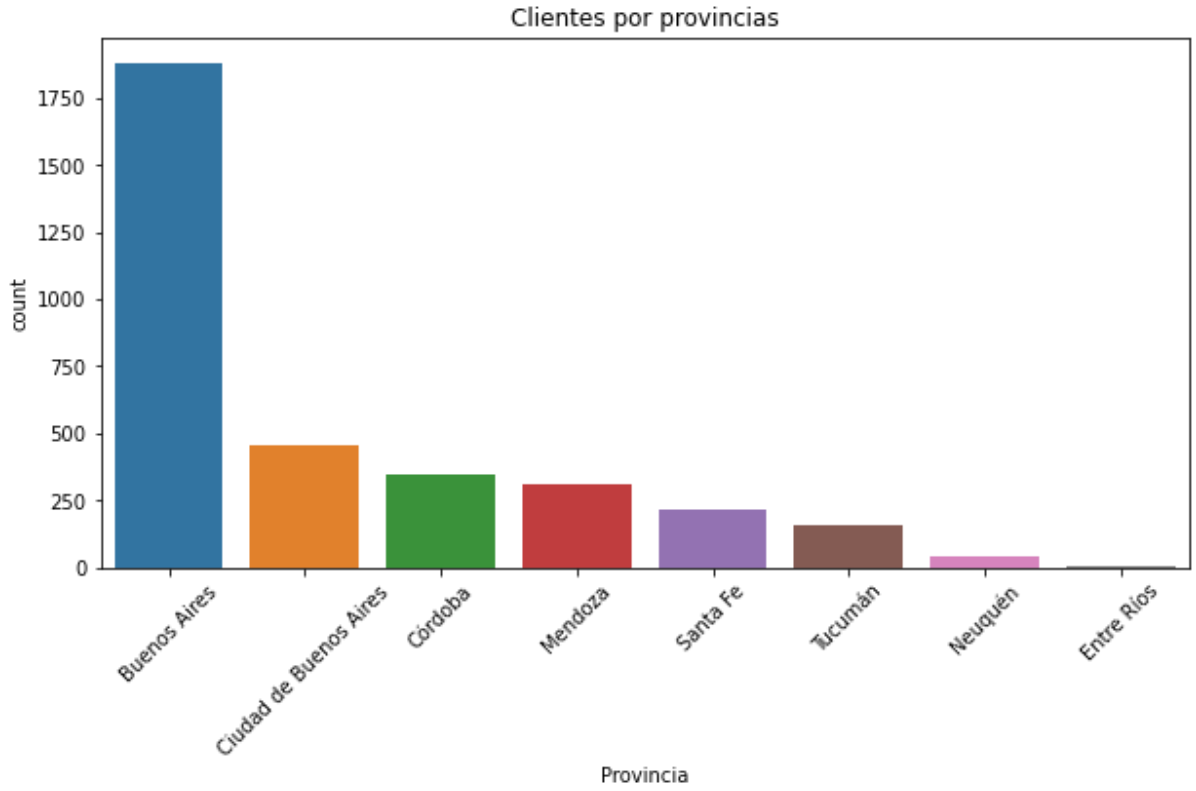
A raíz de la solicitud elevada por el área encargada de la apertura de nuevas sucursales, se detallan las siguientes consideraciones para que puedan tomar una decisión basada y fundamentada en la información que se ha extraído de los datos.

Lo primero a tener en cuenta es cómo se encuentra conformada actualmente la distribución de las sucursales de la empresa.



En el gráfico de barras podemos observar que en la provincia de Buenos Aires se encuentran la mayor cantidad de sucursales de la empresa, representando más del 300% que otras provincias como Córdoba, Mendoza o Santa Fe. Esto cobra sentido por cuestiones poblacionales, de demanda y ubicación de los clientes. Pero, a simple vista, un hecho llama la atención.

Para contemplar más a fondo la representatividad de sucursales por clientes, es esencial analizar la ubicación de estos, ya que nos podría estar señalando una demanda no cubierta y un gran potencial.



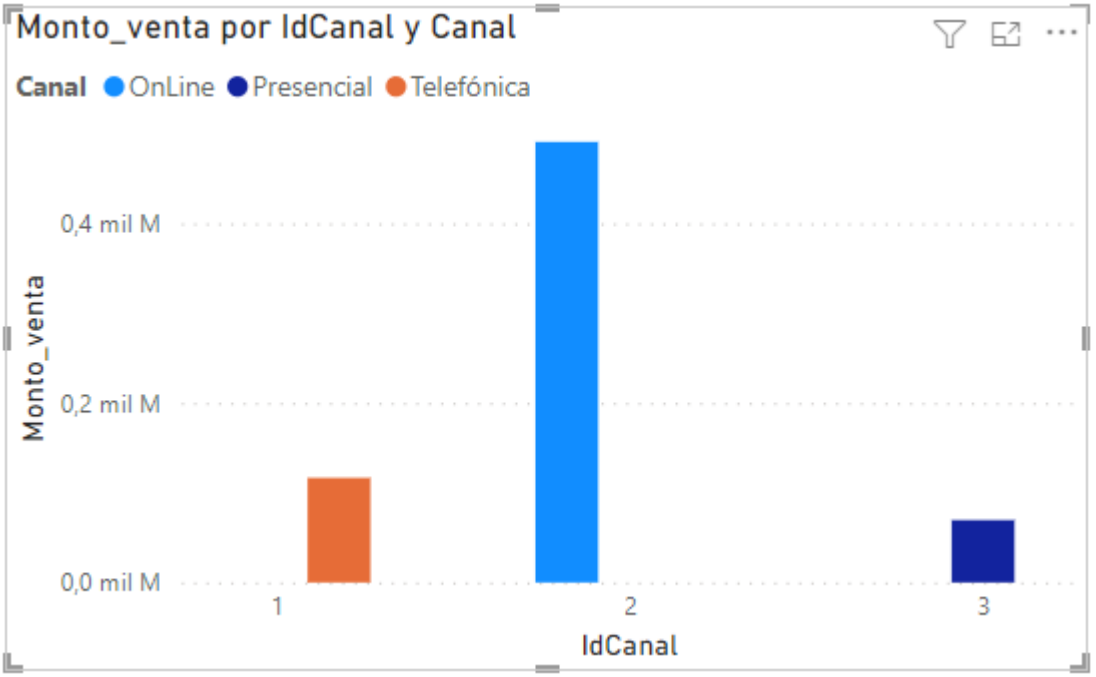
Se destaca que casi el 55% de los clientes se encuentran en la Provincia de Buenos Aires pero solo el 38% de sucursales de la empresa están localizadas allí.

El hecho más destacado es la sucursal de Río Negro, provincia en la que la empresa no tiene absolutamente ningún cliente. En esta provincia hay dos sucursales pero ningún cliente.

Para finalizar, otro hito se puede mencionar. La empresa tiene clientes en Entre Ríos, pero ninguna sucursal instalada. Esto representa un gran potencial de crecimiento en el interior del país. Por ese motivo, desde el área de datos se sugiere una apertura de sucursal en Entre Ríos.

KPI

En este apartado, se presenta un KPI importante para la empresa. Este refiere al monton total de ventas de la empresa durante todo el año 2020 para el canal OnLine.



Como se puede observar, este canal representa la mayor cantidad de ventas para la empresa durante ese año de manera significativa. Esto probablemente estuvo potenciado por la pandemia de COVID.

El total de ventas del año 2020 para el canal de venta OnLine ha sido de 677 millones