UNIVERSIDAD MAYOR DE SAN SIMÓN
FACULTAD DE CIENCIAS Y TECNOLOGÍA
DIRECCIÓN DE POSGRADO

DIPLOMADO ESTADÍSTICA APLICADA A LA TOMA DE DECISIONES

SEGUNDA VERSIÓN

PRÁCTICA #3

CASO REAL

NOMBRE : HEBERT JUAN DE DIOS DELGADILLO FERNANDEZ

CARLOS ALFREDO ORIHUELA BERRIOS

DOCENTE : DANNY LUIS HUANCA SEVILLA

Cochabamba – Bolivia

Índice

1 ANTECEDENTES

Nova Moda SRL es una importadora y distribuidora de calzados brasileros con presencia a nivel nacional con una trayectoria de más de 30 años. Dedicada a la venta mayorista de calzados Brasileros.

Siendo la distribuidora exclusiva de la fábrica de calzados Beira Rio contando con marcas como:

Nova Moda SRL cuenta con 5 almacenes con capacidad de hasta 120000 pares de calzados, 2 almacenes en Cochabamba, 2 almacenes en La Paz y 1 en Santa cruz.

2 FASE DE ENTENDIMIENTO DEL NEGOCIO

2.1 Desde una perspectiva comercial:

2.1.1 ¿Qué espera obtener de este proyecto?

- Mejora en la eficiencia del aprovisionamiento: Se busca optimizar el proceso de aprovisionamiento de mercadería según la temporada, lo que permitirá una gestión más eficiente de los recursos y una reducción de costos asociados.
- Aumento de la satisfacción del cliente: Al contar con una oferta de productos adecuada a cada temporada, se espera satisfacer las necesidades y preferencias de los clientes, mejorando así su experiencia de compra y fidelidad hacia la empresa.
- Maximización de la rentabilidad: Mediante la adecuada gestión del aprovisionamiento, se busca maximizar la rentabilidad de la organización, optimizando los ingresos y minimizando los costos asociados al almacenamiento y manejo de inventario.
- Toma de decisiones basada en datos: Se espera contar con información y análisis sólidos que respalden la toma de decisiones estratégicas relacionadas con el aprovisionamiento de mercadería. Esto permitirá tomar decisiones informadas y basadas en evidencia, reduciendo la incertidumbre y minimizando los riesgos.
- Ventaja competitiva: Al implementar un enfoque analítico en el aprovisionamiento de mercadería, la empresa podrá diferenciarse en el mercado, ofreciendo productos relevantes y oportunamente según las demandas estacionales. Esto puede resultar en una ventaja competitiva frente a otras organizaciones del mismo sector.

En resumen, se espera obtener una mejora significativa en la gestión del aprovisionamiento, la satisfacción del cliente, la rentabilidad y la capacidad de tomar decisiones estratégicas basadas en datos, lo que contribuirá al crecimiento y éxito de la organización.

2.1.2 ¿Cómo define la finalización de los trabajos?

- Identificación del producto más rentable: A través del análisis de los datos de mercado, se ha identificado el producto que tiene el mayor potencial de rentabilidad durante la temporada específica. Esto permite a la empresa enfocar sus esfuerzos en la promoción y venta de este producto para maximizar sus ganancias.
- Mejor toma de decisiones: La tabla minable proporciona información valiosa sobre las preferencias del mercado y el comportamiento de los consumidores durante la temporada en cuestión. Esto permite a la empresa tomar decisiones más informadas y estratégicas en términos de producción, inventario, marketing y distribución.
- Ventaja competitiva: Al conocer cuál es el mejor producto para la temporada, la empresa puede ganar una ventaja competitiva al enfocar sus recursos en ese producto específico. Esto le permite destacarse en el mercado y captar la atención de los consumidores, aumentando sus posibilidades de éxito.
- Optimización de recursos: Al identificar el producto más rentable, la empresa puede optimizar sus recursos al enfocarse en él. Esto incluye la asignación adecuada de personal, inversión en publicidad y promoción, gestión de inventario y producción, entre otros aspectos. Esto conlleva a una utilización más eficiente de los recursos y a la maximización de los beneficios.

2.1.3 ¿ Dispone de la dotación presupuestaria y de los recursos necesarios para completar los objetivos?

Si, la descripción está realizada en la siguiente tabla, los costos de mantenimiento y supervisión no fueron considerados por solicitud de la empresa contratista. El presupuesto recupera su inversión en un plazo de 2 años y 3 meses,por lo cual el proyecto es viable.

Presupuesto		
Equipo de trabajo	Experto en SQL y BBDD	Analista de datos
Salario total	BOB 74,400.00	BOB 49,600.00
Cuota patronal	BOB 24,955.00	BOB 20,974.60
Indeminzacion de finalizacion	BOB 3,950.00	BOB 3,318.00
Personas/mes	2	3
Duracion laboral	140 horas	
Duracion estimada	910 horas	
Costo del servidor SQL	BOB 5,707.20	
Configuracion y conexión en red del servidor SQL	BOB 2,227.20	
Cables y conectores	BOB 1,392.00	
Costo de instalacion	BOB 765.60	
Costos de transporte	BOB 1,740.00	
Gastos Varios	BOB 8,352.00	
TOTAL	BOB 197,381.60	

Tabla de presupuesto requerido para la instalación y configuración del servidor MySQL.

2.1.4 ¿Dispone de acceso a todos los datos necesarios para el proyecto?

Si, ya que se consultó a la empresa y ésta otorgó varios archivos en formato excel, los cuales fueron unidos para crear un dataset más grande. Los archivos son suficientes y necesarios para realizar un análisis de las ventas de la empresa y también existe disposición de parte de la empresa en caso de requerir más datos en un futuro.

2.1.5 ¿Ha tratado con su equipo los riesgos y contingencias asociadas con el proyecto?

Se trató el riesgo sobre la seguridad y privacidad: Cuando se trabajan con datos sensibles, existe el riesgo de violaciones de seguridad o de privacidad, en este caso se pudo obtener datos sobre las cuentas de saldo de los clientes pero esto son considerados datos muy sensibles por parte de la empresa Nova Moda, el cual podría tener consecuencias legales y reputacionales. Por lo cual no se nos entregó esos datos.

2.1.6 ¿Los resultados del análisis de coste/beneficios hacen que el proyecto sea viable?

El proyecto resulta viable ya que el retorno de inversión anual es de 89718.9 Bs. lo que quiere decir que el retorno de inversión es de 2 años y 3 meses por el ahorro en cajas que ya no se pasaran de moda por un exceso de pedido, en un tiempo de 5 años la relación costo beneficio será de: 2.27, es decir que en 5 años los beneficios serán más del doble veces la inversión inicial.

2.2 Desde una perspectiva de ciencia de datos:

2.2.1 ¿En qué forma puede ayudarle la ciencia de datos a cumplir sus objetivos comerciales?

Para este proyecto en particular, la ciencia de datos puede ayudar de las siguientes formas a cumplir los objetivos comerciales:

Predecir la demanda de mercadería según la temporada: Utilizando técnicas de modelado y análisis de datos, la ciencia de datos puede desarrollar modelos predictivos que estimen la demanda de mercadería en función de la temporada. Esto permitirá planificar y ajustar la importación de productos de acuerdo a las necesidades previstas, evitando así problemas de falta de stock o exceso de inventario.

Identificar patrones de comportamiento de los clientes: Mediante el análisis de datos de los clientes, es posible identificar patrones de comportamiento y preferencias que permitan segmentar y personalizar las ofertas de productos. Esto ayudará a optimizar las estrategias de marketing y ventas, ofreciendo a los clientes los productos más adecuados según sus preferencias y aumentando la probabilidad de compra.

Optimizar la gestión de inventario: La ciencia de datos puede ayudar a optimizar la gestión del inventario, determinando los niveles óptimos de stock para cada producto en función de la demanda histórica, las proyecciones futuras y otros factores relevantes. Esto evitará situaciones de escasez o exceso de inventario, mejorando la eficiencia y reduciendo los costos asociados.

Identificar patrones de fraude: Mediante técnicas de detección de anomalías y análisis de datos, la ciencia de datos puede ayudar a identificar patrones de fraude en las transacciones comerciales. Esto permitirá implementar medidas preventivas y de seguridad para minimizar los casos de fraude y proteger los activos de la empresa.

Mejorar la toma de decisiones estratégicas: La ciencia de datos proporciona información valiosa para la toma de decisiones estratégicas. Mediante el análisis de datos, se pueden identificar oportunidades de crecimiento, evaluar el desempeño de diferentes estrategias y realizar pronósticos y simulaciones que ayuden a tomar decisiones más acertadas y fundamentadas.

2.2.2 ¿Sabe qué técnicas de ciencia de datos producen los mejores resultados?

Modelos de clasificación: Los modelos de clasificación, como la regresión logística, los árboles de decisión o los clasificadores basados en métodos de aprendizaje automático, podrían ser útiles para predecir la popularidad de los artículos en función de diferentes características y variables predictoras.

2.2.3 ¿Cómo se implementarán los resultados de modelado? ¿Ha considerado implementar su plan de proyecto?

La implementación de los resultados de modelado puede variar. Algunas consideraciones generales para implementar los resultados del modelado son:

Integración en sistemas existentes: Si la empresa ya tiene sistemas o plataformas en funcionamiento, es importante evaluar la viabilidad de integrar los modelos de ciencia de datos en estos sistemas. Esto puede implicar el desarrollo de APIs o interfaces que permitan la comunicación entre los sistemas y el uso de los modelos para la toma de decisiones en tiempo real.

Desarrollo de aplicaciones: En algunos casos, puede ser necesario desarrollar aplicaciones o interfaces de usuario que permitan a los usuarios finales interactuar con los modelos de ciencia de datos. Estas aplicaciones pueden incluir paneles de control, herramientas de visualización de datos o incluso aplicaciones móviles para acceder y utilizar los modelos de manera fácil y efectiva.

Automatización de procesos: Si el objetivo es utilizar los modelos de ciencia de datos para automatizar ciertos procesos comerciales, es fundamental integrar los modelos en flujos de trabajo automatizados. Esto puede implicar la programación de tareas o procesos que ejecuten los modelos de forma periódica y generen los resultados esperados.

Monitoreo y mantenimiento: Una vez implementados los modelos, es importante establecer un sistema de monitoreo continuo para asegurarse de que están funcionando correctamente y produciendo resultados precisos. Además, es necesario realizar un mantenimiento regular de los modelos, actualizando y mejorando según sea necesario para garantizar su efectividad a lo largo del tiempo

2.2.4 ¿El plan de proyecto incluye todas las fases de CRISP-DM?

Para esta primera fase solo estamos abarcando hasta la preparación de datos para su posterior modelaje.

2.2.5 ¿Los riesgos y dependencias se incluyen en el plan?

Los riesgos para llevar a cabo el plan son:

Riesgo de presupuesto y recursos: Existe la posibilidad de que el proyecto supere el presupuesto asignado o no se cuente con los recursos adecuados para completar los objetivos. Esto podría afectar la calidad y el alcance del trabajo realizado.

Riesgo de implementación: Si no se planifica adecuadamente la implementación de los resultados del modelado, podría haber dificultades para integrar las soluciones propuestas en el sistema existente o para llevar a cabo los cambios necesarios en los procesos comerciales.

Riesgo de precisión y fiabilidad de los modelos: Los resultados del análisis y los modelos generados pueden estar sujetos a cierto grado de error o incertidumbre. Esto puede influir en la confianza y la validez de las recomendaciones y decisiones basadas en ellos.

Las dependencias del plan para realizar el proyecto serían:

Dependencias de datos: El proyecto depende de la disponibilidad y acceso a conjuntos de datos, ya que muchas veces se necesitan permisos para acceder a ellos y muchas veces están anonimizados las columnas con información sensible. Aparte de la disponibilidad se depende de la calidad, la granularidad y la integridad de los datos utilizados, ya que habrían datos duplicados, nulos, valores faltantes,etc.

Dependencias de recursos: como los humanos, financieros o tecnológicos necesarios para llevar a cabo el proyecto. Por ejemplo, el proyecto puede requerir habilidades del manejo de lenguajes de JAVA, SCALA, PYTHON o conocimientos especializados que no estén disponibles internamente como APACHE SPARK, KAFKA,etc y deban adquirirse o contratar personas externas al equipo original. Dependencias de tiempo: los proyectos muchas veces dependen de entregables, lo que significa que ciertas actividades sólo pueden comenzar una vez que se hayan completado otras. No se puede modelar e implementar un modelo sin haber definido el pipeline de ETL o ELT para la ingesta de datos. Pueden ser proyectos de corto plazo como de 1 año y de largo plazo como 3 años.

3 ENTENDIMIENTO DE LOS DATOS

3.1 Univariado con todas las variables que componen el dataset realizar lo siguiente:

3.1.1 Clasificar las variables entre variables cualitativas (ordinales o nominales) y cuantitativas (discretas o continuas).

Clasificación	
Material	Cualitativa - Nominal
Color	Cualitativa - Nominal
Item	Cualitativa - Nominal
Almacen	Cualitativa - Nominal
fecha_venta	Cuantitativa - Formato Fecha
Marca	Cualitativa - Nominal
Cliente	Cualitativa - Nominal
T_cajas	Cuantitativa - Continua
Cantidad_Pares	Cuantitativa - Discreto
Total	Cuantitativa - Continua
Desc	Cuantitativa - Continua
Pago	Cuantitativa - Continua
Vendedor	Cualitativa - Nominal
Fecha_de_ingreso	Cuantitativa - Continua
hora	Cuantitativa - discreta
dia	Cuantitativa - discreta
dia_de_semana	Cualitativa - Ordinal
nombre_dia	Cualitativa - Ordinal
mes	Cualitativa - Ordinal

3.1.2 Obtención de estadísticas de tendencia central, dispersión.

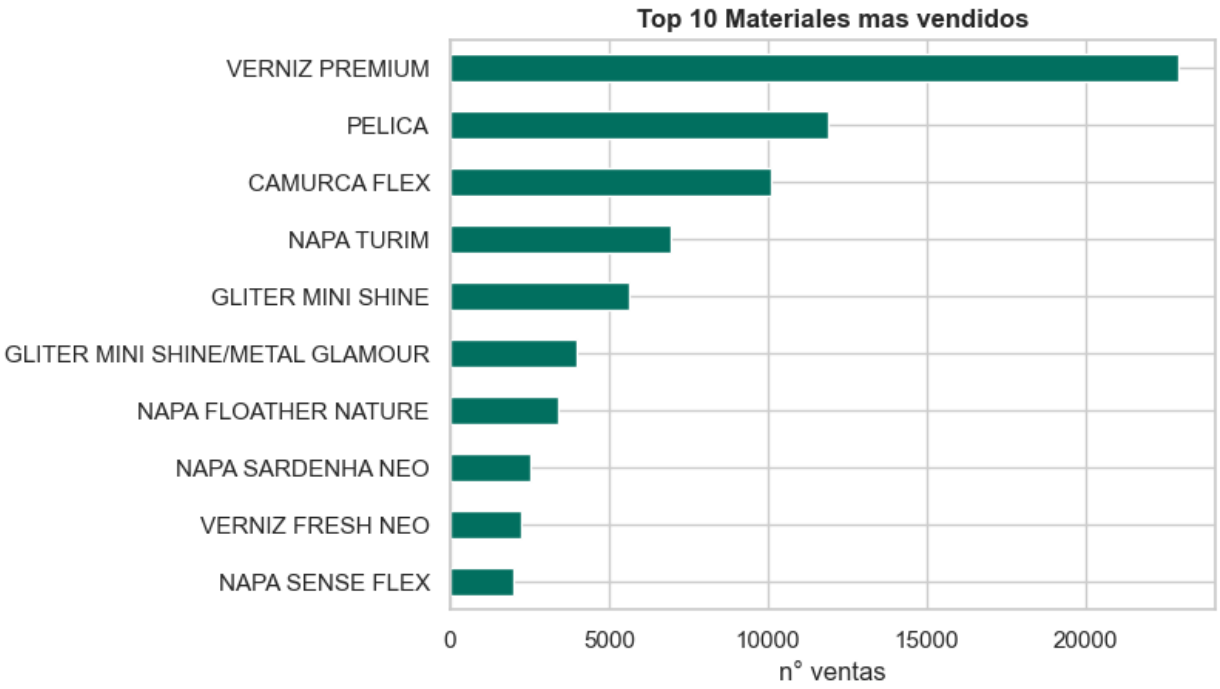
3.1.2.1 Variables Cuantitativas

	count	mean	std	min	25%	50%	75%	max	rango_iqr	rango
T_cajas	141,733.00	0.91	0.27	0.08	1.00	1.00	1.00	4.00	0.00	3.92
Cantidad_Pares	141,733.00	10.95	3.23	1.00	12.00	12.00	12.00	48.00	0.00	47.00
Total	141,733.00	193.62	85.10	0.08	142.00	185.00	244.00	1,569.00	102.00	1,568.92
Desc	141,733.00	723.46	1,555.39	-2,074.00	0.00	131.75	756.50	18,848.00	756.50	20,922.00
Pago	141,733.00	193.62	85.10	0.08	142.00	185.00	244.00	1,569.00	102.00	1,568.92
hora	141,733.00	12.52	3.68	0.00	9.00	12.00	15.00	23.00	6.00	23.00

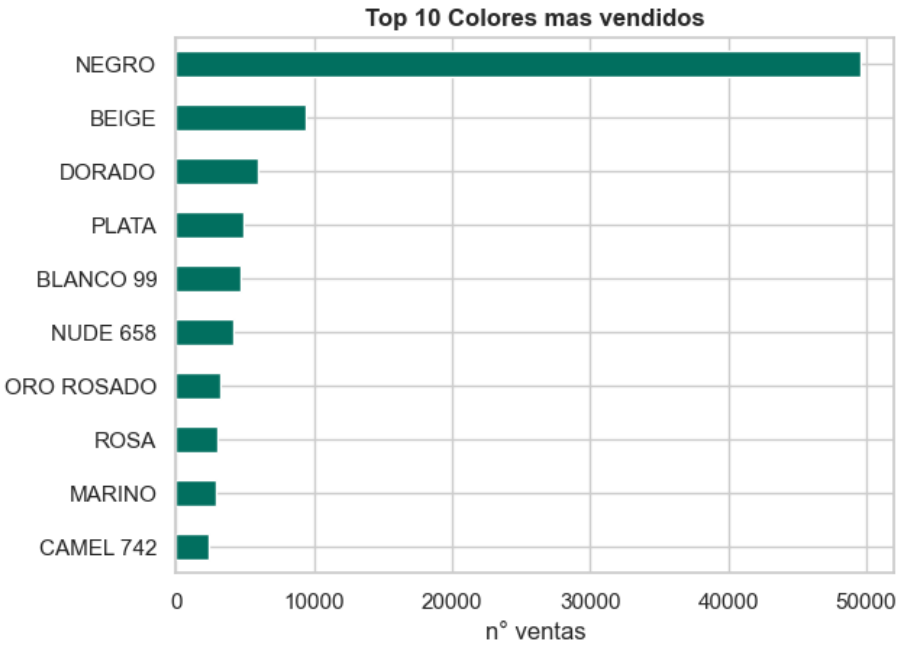
3.1.2.2 Variables Cualitativas

	count	unique	top	freq
Material	141731	4227	VERNIZ PREMIUM	22947
Color	141733	3919	NEGRO	49559
Item	141730	2928	estrategicos	2078
Cliente	141733	1822	LETICIA-FERNANDEZ	3271
Vendedor	141733	136	JUAN CARLOS-BAUTISTA HERBAS	7060
Factura	141703	2494	6171/2021 - SCOF	1252
nombre_dia	141733	7	Friday	26233

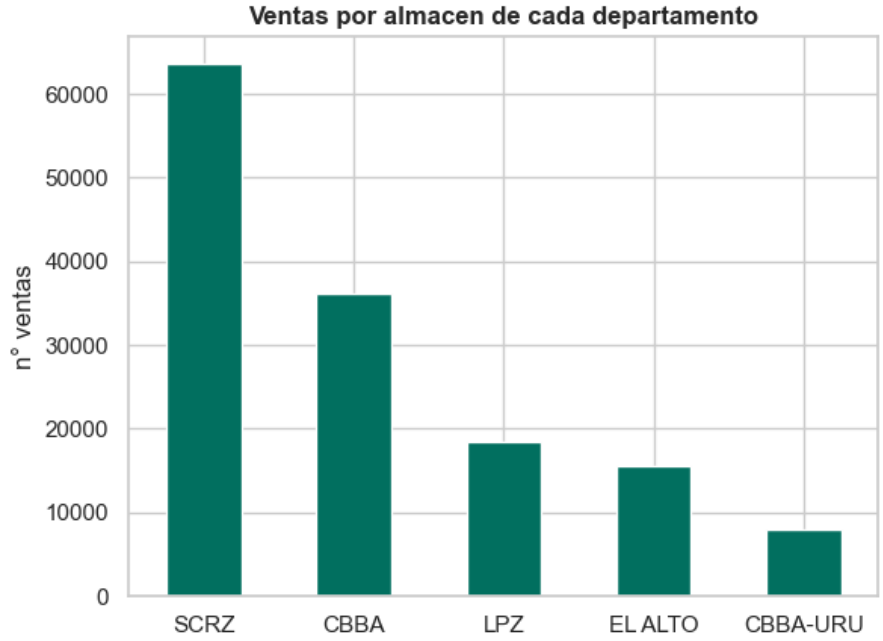
3.1.3 Generación de gráficas dependiendo el tipo, histogramas, barras, boxplot. Las gráficas deben tener algún comentario o descubrimiento que puedan hallar.



- Se observa el Top 10 de materiales de calzados más vendidos.
- El material de calzado más vendido es VERNIZ PREMIUM con 22947 ventas.



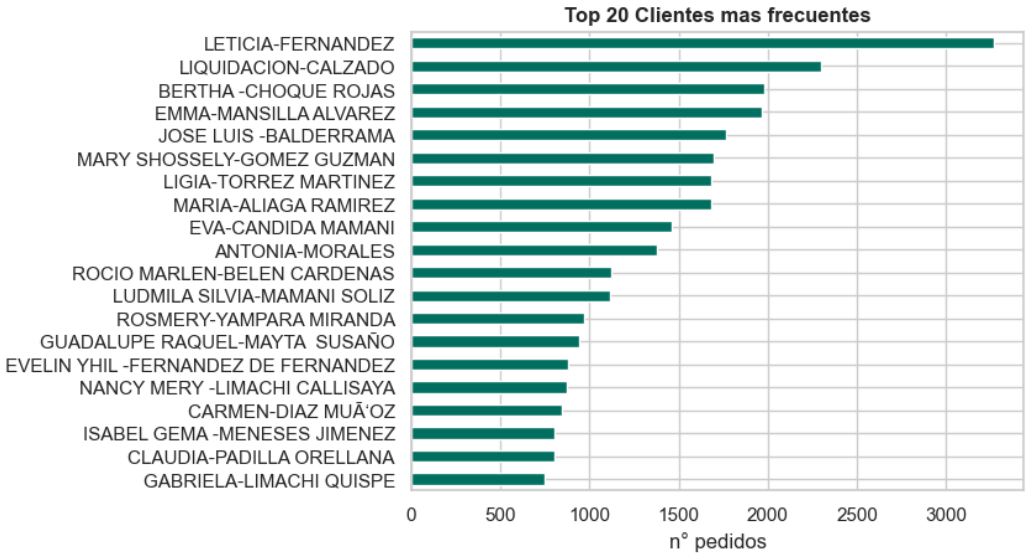
Se observa que el color más vendido y solicitado de acuerdo al número de ventas es el color Negro.



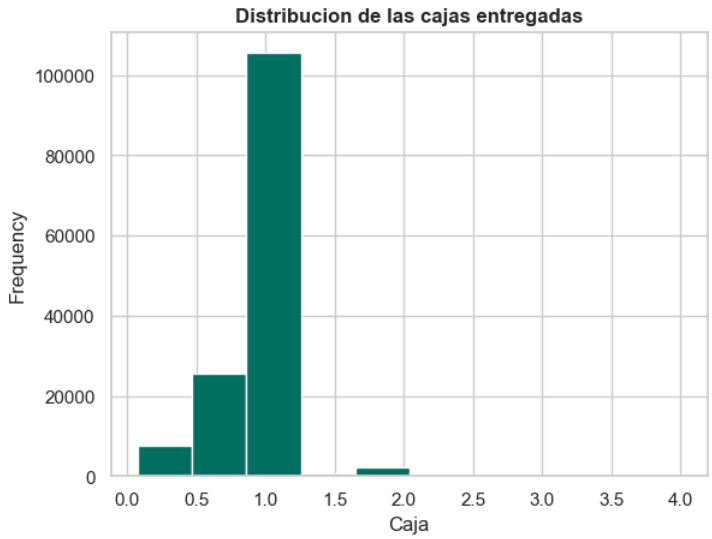
El departamento y el almacén con mayor número de ventas está en Santa Cruz.

Cabe notar que Cochabamba posee dos almacenes en zona Quintanilla (Sacaba) y otra en la calle Uruguay, donde se observa que existe una diferencia del número de ventas. Pero sumando los dos almacenes Cochabamba es el segundo con mayores ventas.

Se puede observar que existe una preferencia por la marca VIZZANO, ya que posee los mayores números de ventas.

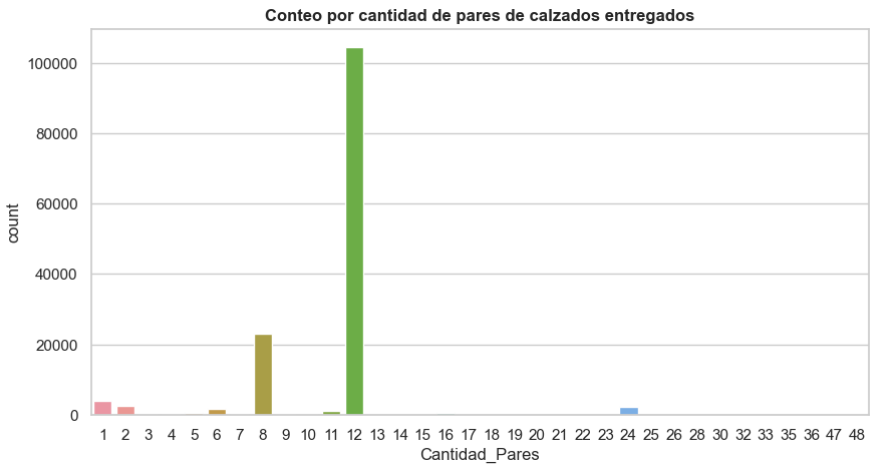


El anterior gráfico muestra los clientes más frecuentes a los que más pedidos se entregó, esto es útil para tomar alguna acción sobre estos clientes , como descuentos u otra promoción de fidelización.



Se puede observar que la mayoría de las cajas que llegan son de tipo 1 (donde un 1 significa que una caja contiene el 100% de 12 pares de calzados) y las que son menores a 1 significan cajas con menos calzados como 0.67 significa que solo contiene 8 calzados.

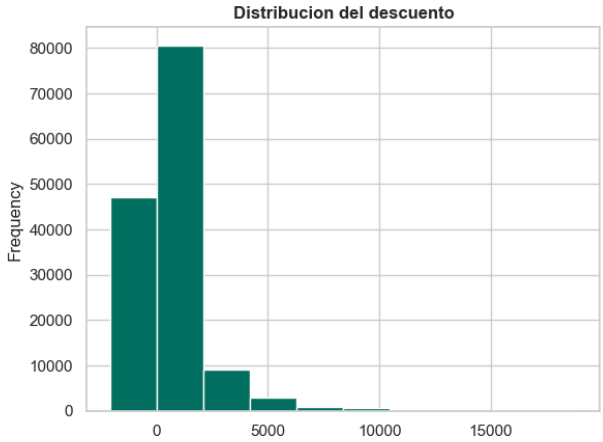
Llegan en menor cantidad al almacén las cajas tipo 2.



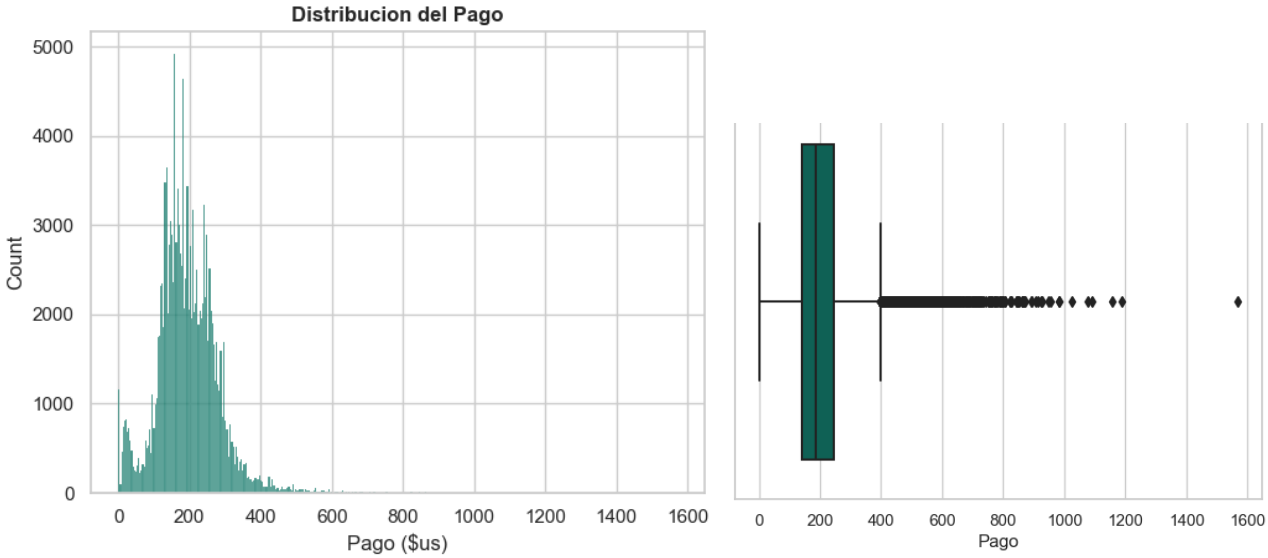
	Cantidad_Pares
12	104371
8	22936
1	3867
2	2453
24	2234
6	1666
11	1003
16	626
5	445
4	411

Se observa que la cantidad más vendida y entregada a los clientes en términos de pares de calzado es 12 (12 pares es el que más se entrega), el cual corresponde también a los pares dentro una caja tipo 1.

Se entregaron 104371 pedidos de 12 pares.

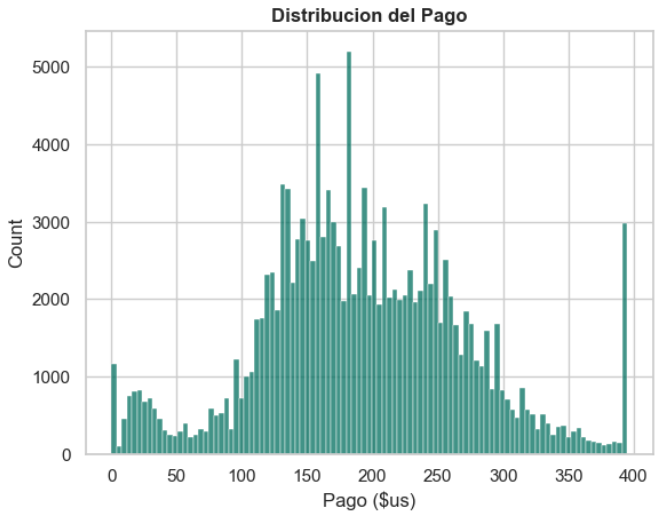


El 80% de los descuentos fueron iguales o menores a 1014.5 Bs.



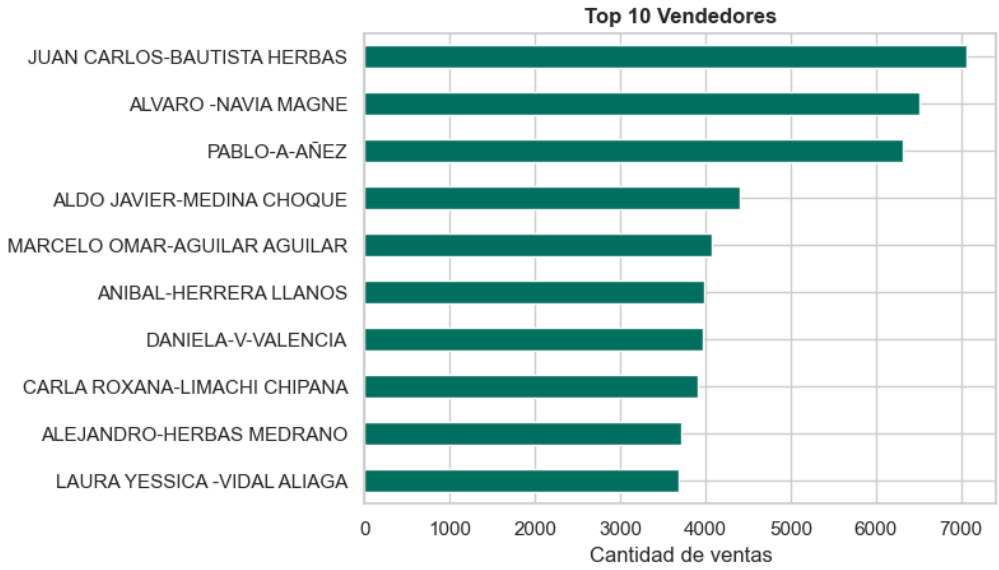
Se observa que hay muchos valores atípicos con respecto al pago.

3.1.4 Quitando los valores atípicos con los cuartiles se tiene la siguiente distribución:

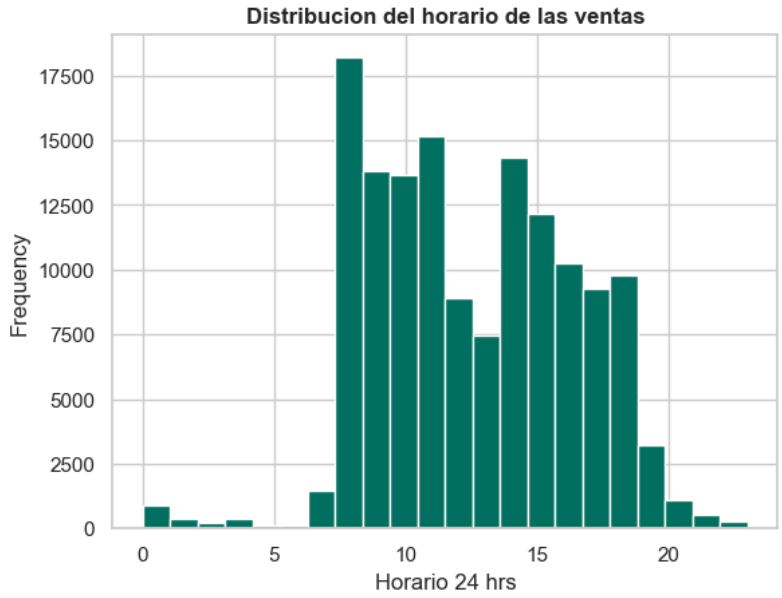


La mayoría o el 98 % de los pagos son menores a 400 $us.

Se puede observar que los pagos tienden al valor de alrededor 200 $us por pedido.

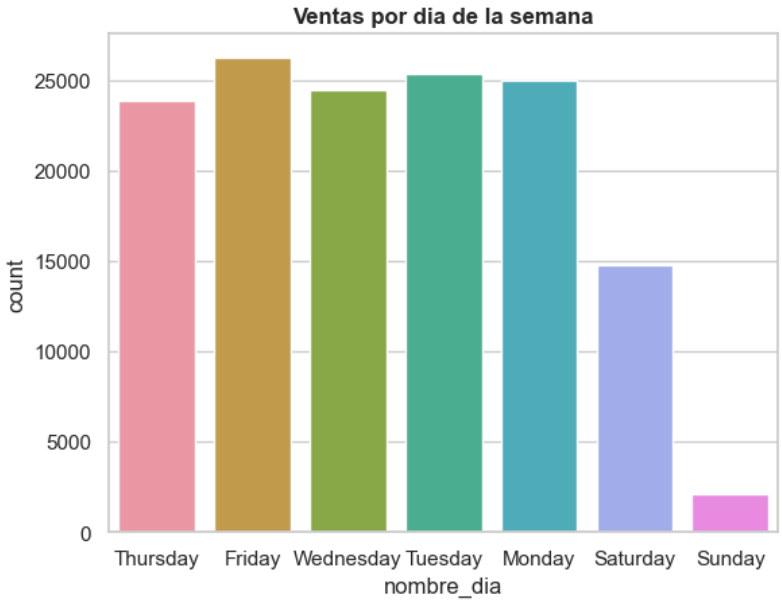


Juan Carlos hizo un poco más de 7000 ventas, es el vendedor con más ventas.

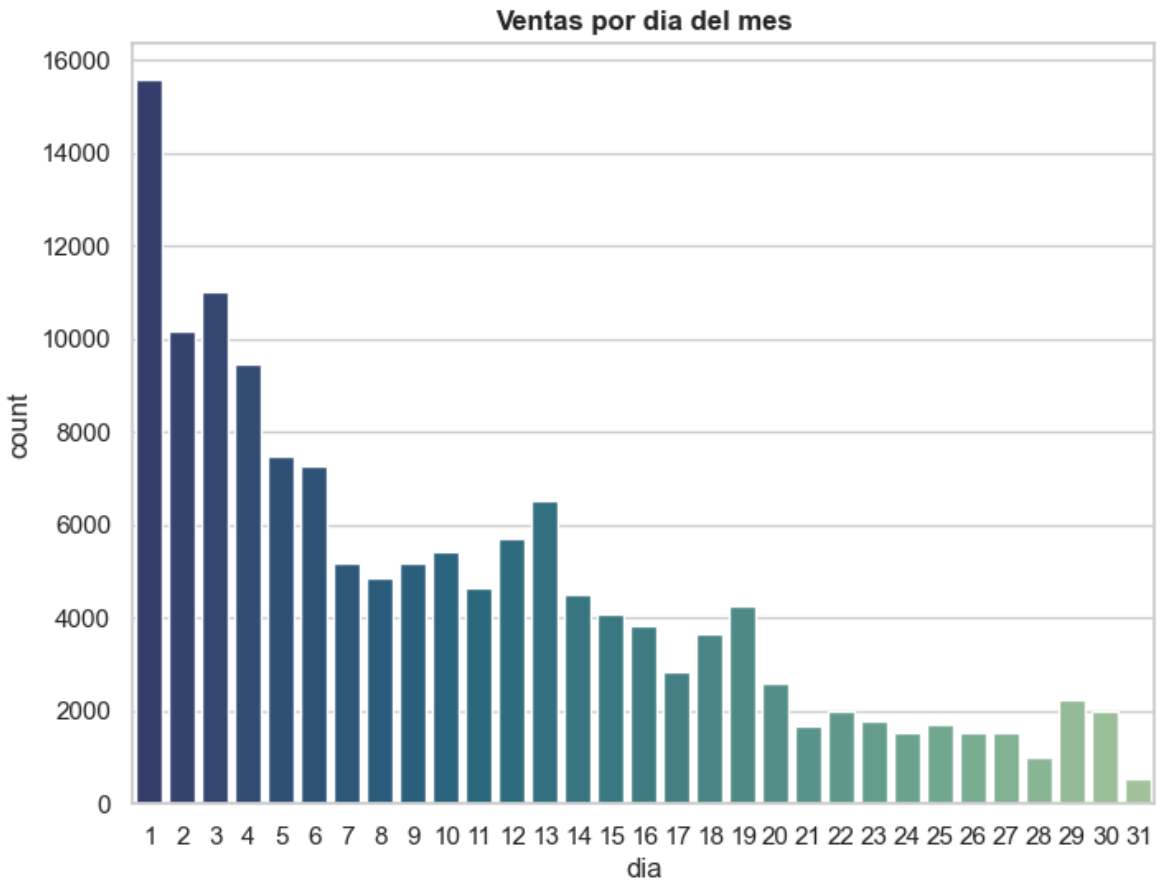


Se observa que el 80% de las ventas fueron realizadas en horario laboral, aunque algunas ventas se hicieron fuera de ese horario.

Existe una distribución aparentemente bimodal, donde hay picos de ventas por la mañana y otro por la tarde, lo cual es lógico debido a la pausa del almuerzo.

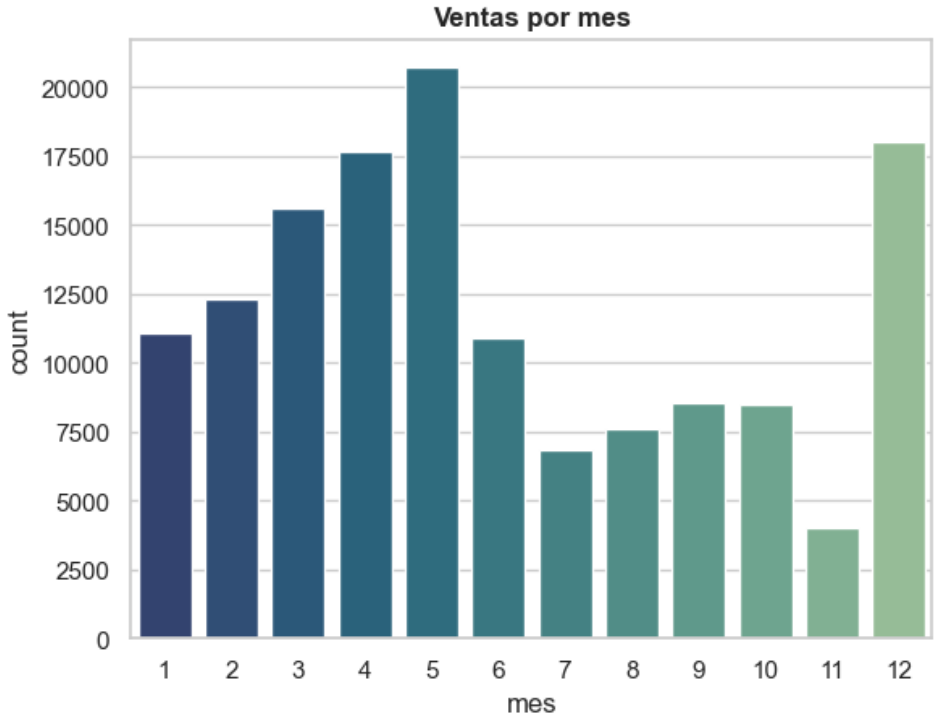


Se observa que en general no parece haber mucha variación entre los días lunes a viernes, a excepción de los días sábado y domingo en donde las ventas se reducen.



Aquí se observa un comportamiento interesante, el día donde más se vende es el 1er día de cada mes.

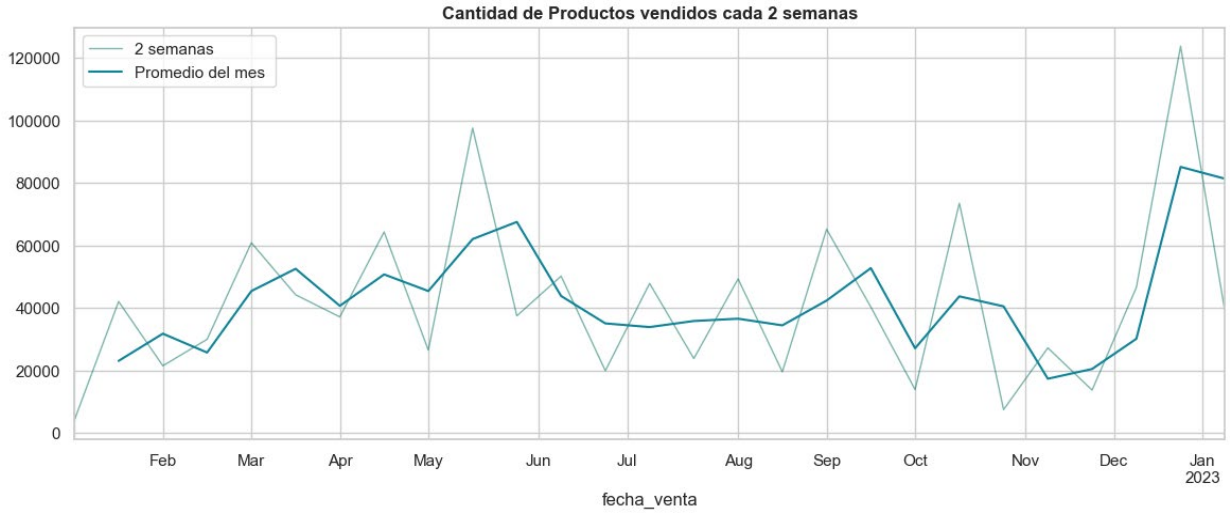
Se vende más durante la 1er y 2da semana del mes , alrededor del 75% de las ventas totales de cada mes.



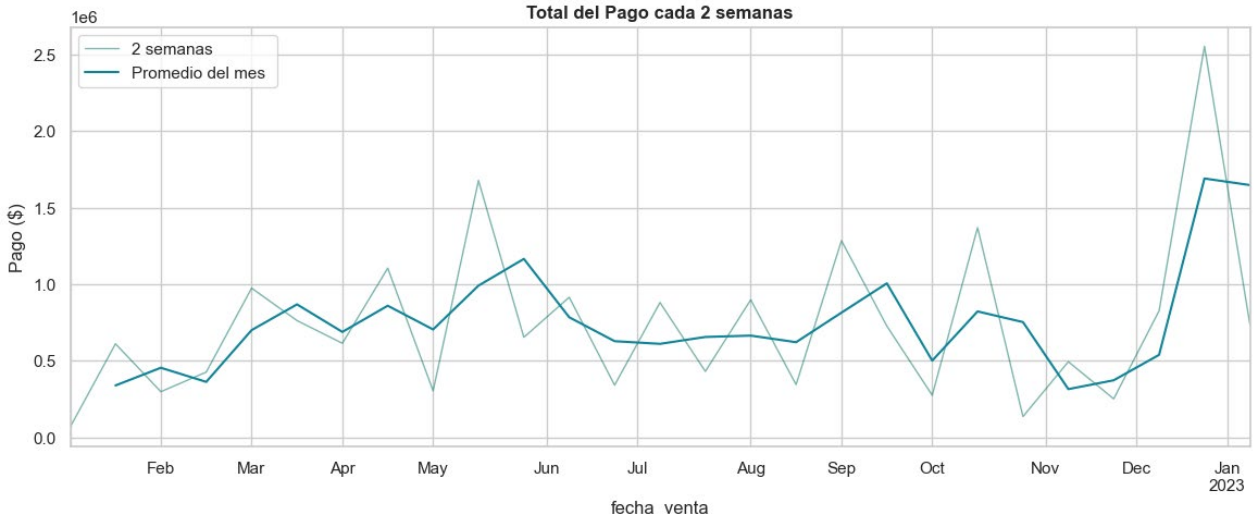
El mes con más ventas es mayo.

Los meses que más ventas tienen son de febrero a mayo y en diciembre.

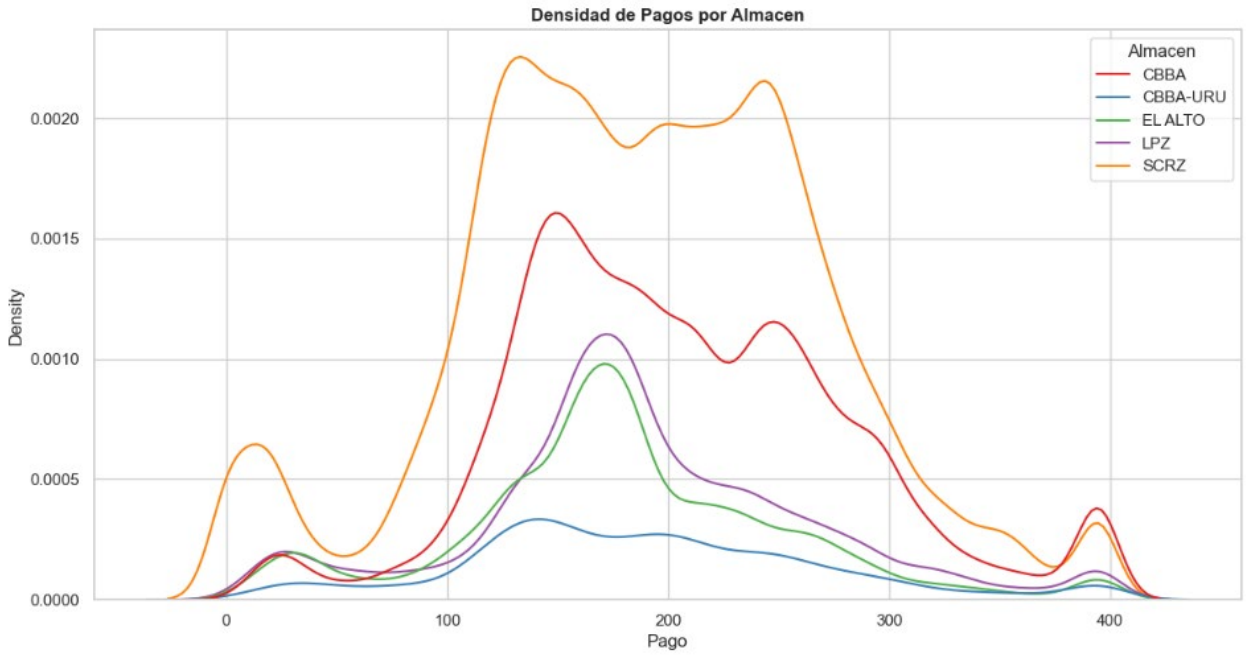
3.2 Bivariado con las variables que ingresen en el estudio, respecto de la variable objetivo o target.



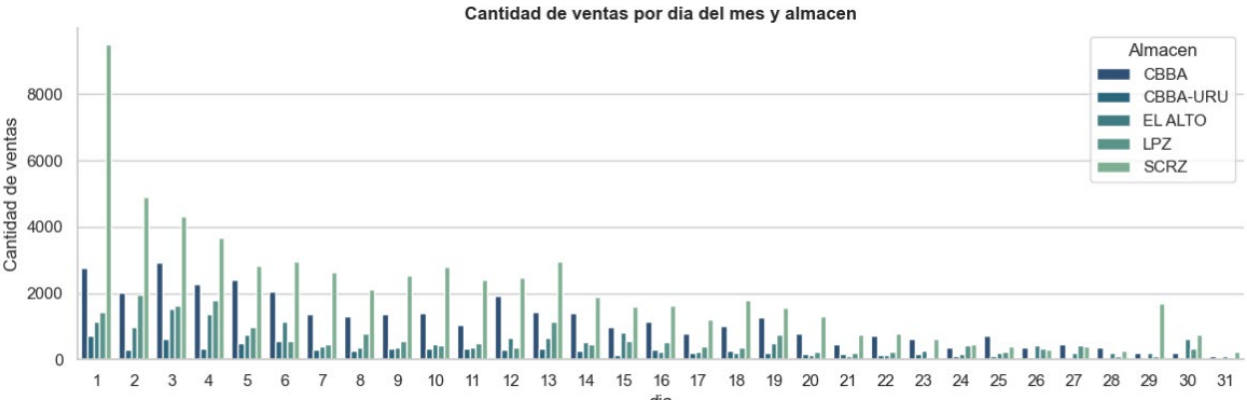
Se puede observar una cierta temporalidad en la cantidad de calzados que se venden durante todo el año de 2022. Las ventas aumentan durante marzo, abril y mayo. Después hay una tendencia baja y posteriormente vuelve a subir por diciembre que es algo esperado por la navidad.



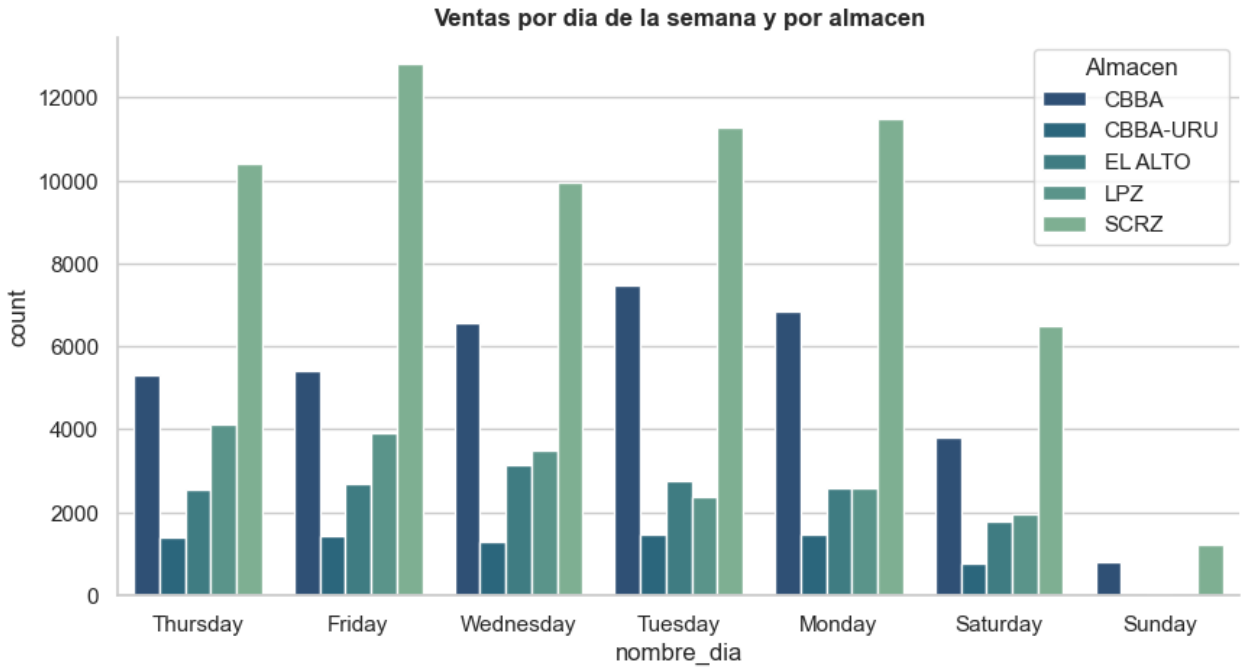
En el anterior gráfico se puede observar que los pagos siguen la misma tendencia que la cantidad y eso es debido a que en su gran mayoría se venden 12 pares por pedido. Por tanto están altamente correlacionados.



Se puede observar que el almacén de Santa cruz concentra la mayor cantidad de pagos de las ventas y que casi todos tienden a un valor alrededor de 200 $us por pedido.



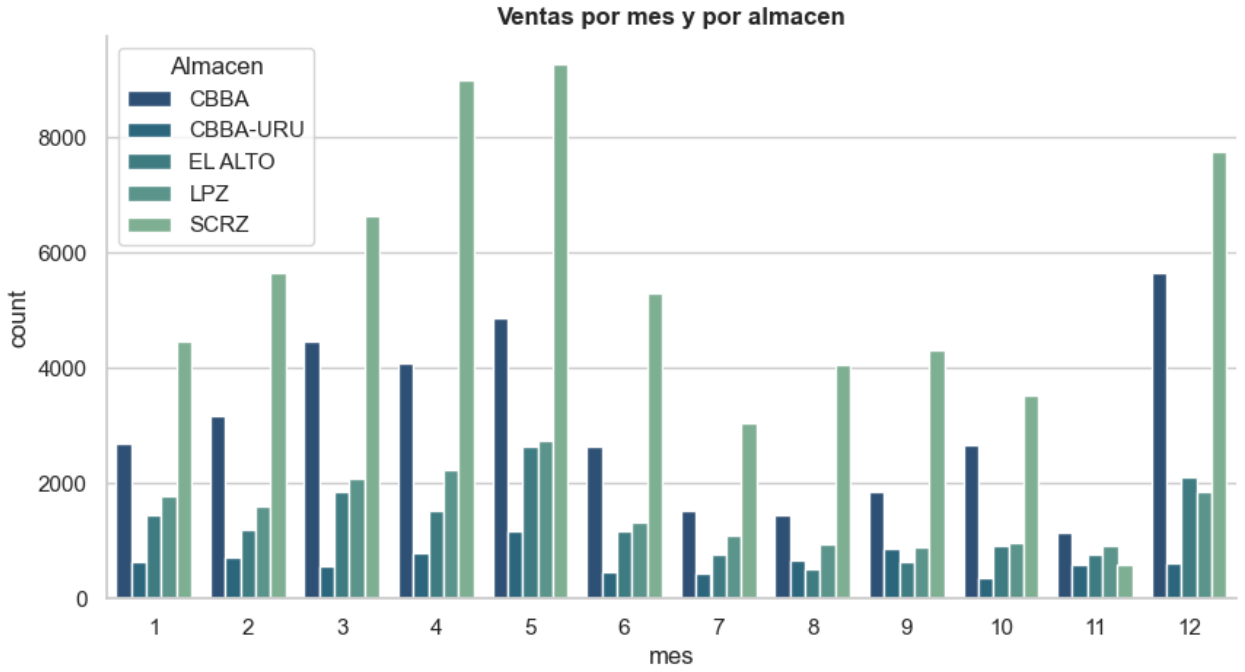
Todos los almacenes venden la mayoría de los calzados la primera semana de cada mes. Especialmente los primeros días.



Existe un incremento de ventas los días viernes en el almacén de Santa Cruz con respecto a los demás días.

Existe un aumento de ventas los martes para el almacén de Cochabamba-Quintanilla.

El incremento de ventas en el almacén de La Paz sucede los jueves.



Se puede observar en el gráfico anterior que en casi todos los meses el Almacén que se encuentra en Santa Cruz es el más ventas realiza.

Se observan grandes picos de ventas se producen alrededor de mayo y diciembre, especialmente para los almacenes de Santa Cruz y Cochabamba. En los demás almacenes no se observan picos tan pronunciados como los otros.

3.3 Generación de un perfilado del total de variables que usará para la construcción de su modelo.

El perfilado de los datos se realizó con la librería Dataprep de Python.

El perfilado completo se encontrará en los archivos adjuntos como Perfilado Nova SRL.

DataPrep Report | Overview | Variables ≡ | Interactions | Correlations | Missing Values

Overview

Dataset Statistics

Number of Variables	19
Number of Rows	141733
Missing Cells	33
Missing Cells (%)	0.0%
Duplicate Rows	17999
Duplicate Rows (%)	12.7%
Total Size in Memory	76.0 MB
Average Row Size in Memory	562.0 B
Variable Types	Categorical: 10 Numerical: 8 DateTime: 1

Dataset Insights

`Total` and `Pago` have similar distributions	Similar Distribution
`T_cajas` is skewed	Skewed
`Cantidad_Pares` is skewed	Skewed
`Total` is skewed	Skewed
`Desc` is skewed	Skewed
`Pago` is skewed	Skewed
Dataset has 17999 (12.7%) duplicate rows	Duplicates
`Item` has a high cardinality: 2928 distinct values	High Cardinality
`Cliente` has a high cardinality: 1822 distinct values	High Cardinality
`Vendedor` has a high cardinality: 136 distinct values	High Cardinality

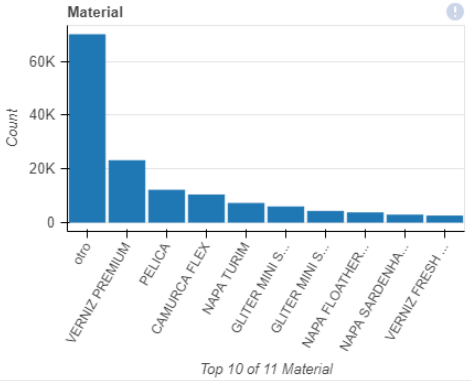
Variables

Sort by Feature order ☐ Reverse order

Material
categorical

Show Details

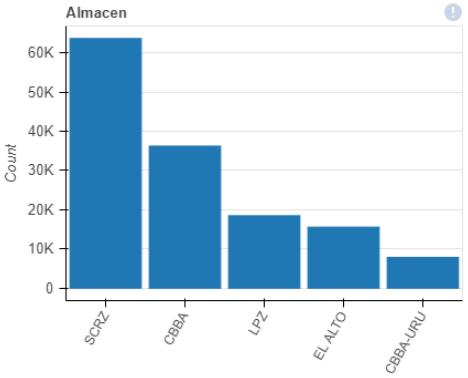
Approximate Distinct Count	11
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	10472493



Almacen
categorical

Show Details

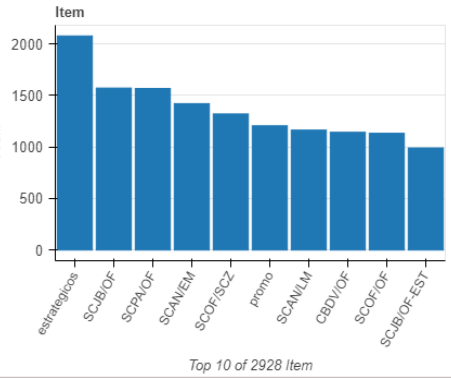
Approximate Distinct Count	5
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	1276080



Item
categorical

Show Details

Approximate Distinct Count	2928
Approximate Unique (%)	2.1%
Missing	3
Missing (%)	0.0%
Memory Size	10341657

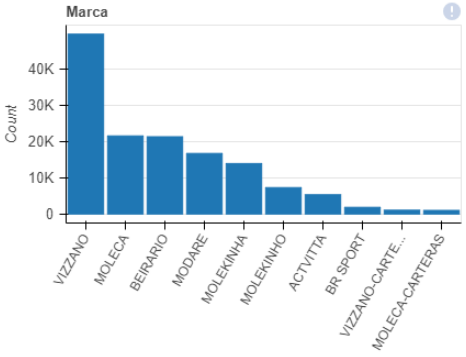


Top 10 of 2928 Item

Marca
categorical

Show Details

Approximate Distinct Count	21
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	1277615

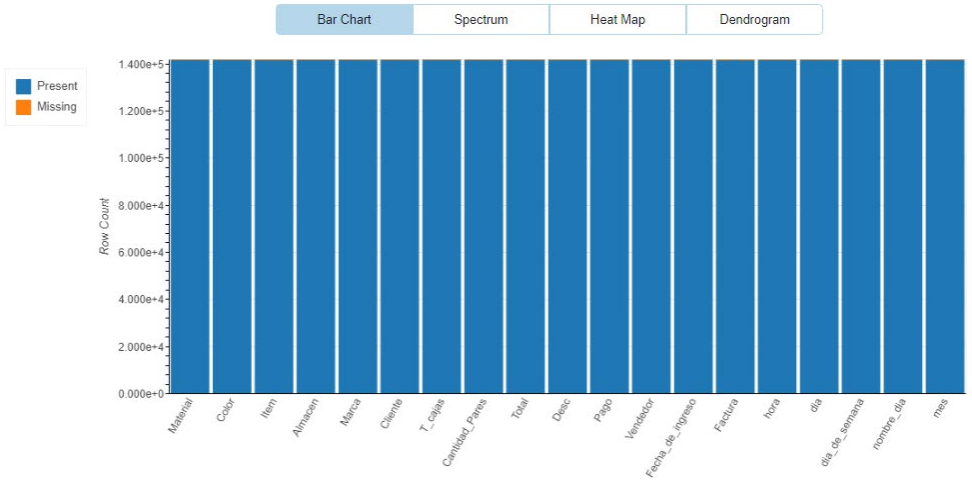


Top 10 of 21 Marca

Fecha_de_ingreso
datetime

Distinct Count	93.066
Approximate Unique (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory Size	2267728
Minimum	2014-08-01 00:00:00
Maximum	2023-06-01 00:00:00

Missing Values



3.3.1 ¿Cuál es su nivel de comprensión de los datos?

3.3.1.1 ¿Ha identificado y accedido correctamente a todos los orígenes de datos? ¿Ha tenido algún problema o restricción de algún tipo?

Si, se accede correctamente a los datos proporcionados por la importadora de calzados brasileros Nova Moda srl. Los cuales fueron cargados a una Base de datos Mysql mediante un ETL en pentaho, para facilitar su análisis correspondiente.

3.3.1.2 ¿Ha identificado atributos clave de los datos disponibles?

Si, mediante el análisis exploratorio se encontraron relaciones interesantes, en especial con las variables de tiempo como fechas, día y mes, ya que se observaron un cierto patrón con el número de ventas que se tuvo. Lo cual es bueno porque son las variables que pueden tener más influencia a la hora hacer predicción por una regresión múltiple.

3.3.2 ¿Le han ayudado estos atributos a formular hipótesis?

Si, ya que pensamos que los datos diarios podrían ser suficientes para predicción, pero también se tiene otra hipótesis donde se podría usar intervalos como de 6 horas para la creación de nuevas columnas con retraso de 6 horas en lugar de días o también se puede probar de manera mensual.

3.3.3 ¿Ha detectado el tamaño de todos los orígenes de datos?

Si, mediante comandos usando pandas para observar el tamaño de los datos, como el comando df.shape que retorna el valor de las dimensiones de un datagrama, así también sus tipos de datos de cada variable

3.3.4 ¿Puede utilizar un subconjunto de datos cuando lo estime conveniente?

En este caso solo por el momento se cuenta con datos de dos años por lo que no sería conveniente tomar muestras pequeñas sino grandes, mientras más datos tengamos más muestras podremos tomar del conjunto. Los datos de más años serán entregados por la empresa los siguientes días.

3.3.5 ¿Ha calculado los estadísticos básicos de cada atributo de su interés? ¿Ha obtenido información de interés?

Si, mediante la función de pandas “describe()” en cual retorna una descripción estadística con la que obtuvimos los valores promedios de cada variable y su variabilidad respecto a ese promedio con la desviación estándar, también otros datos de interés como el rango y el rango intercuartil

3.3.6 ¿Ha utilizado gráficos de exploración para obtener atributos clave? ¿Este conocimiento ha reformulado alguna de sus hipótesis?

Se realizaron varios gráficos de relación con la variable objetivo y otras con una relación relevante, como se puede apreciar en los anteriores entregables de la sección 4.2 fase de entendimiento de los datos

3.3.7 ¿Cuáles fueron los problemas de calidad de datos del proyecto? ¿Tiene una planificación para resolver estos problemas?

Los datos en general no tenían problemas de calidad a excepción de algunas pocas filas que tenían datos erróneos y que el tipo de dato era incorrecto. Para eso se tuvo que convertirlos mediante las funciones de pandas especialmente las fechas.

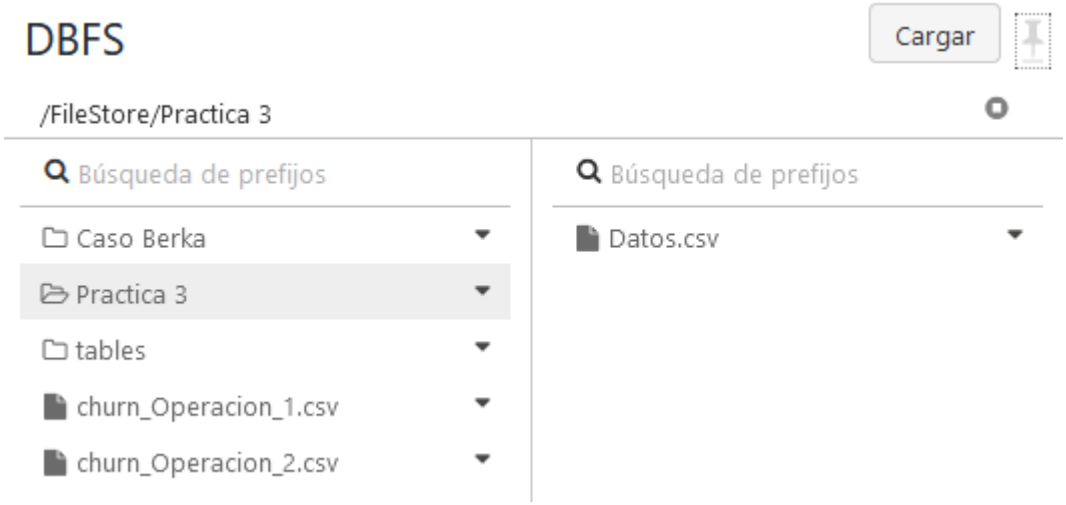
3.3.8 ¿Las fases de preparación de los datos son claras? Por ejemplo, ¿sabe qué orígenes de datos debe fusionar y los atributos que debe filtrar o seleccionar?

Si son claras ya que se realizó el armado del dataset a través de una serie de pasos de transformación , conversión de tipos , guardado de datos y corrección de tipos de datos, posteriormente se realiza el guardado de la tabla transformada paso a paso. Jupyter permite documentar cada paso, por lo cual se hizo un comentario de las líneas de código importantes. Al igual que se establece que todo se ejecute de manera lineal y no con saltos, celda por celda en júpiter notebook, esto para evitar confusiones entre el equipo al momento de compartir un notebook y realizar el proyecto.

4 Preprocesamiento de los datos

4.1 Manejo de datos con DataBricks

Subimos todos los archivos al DBFS



4.1.1 Crear tablas

Con el código siguiente debemos crear toda la tabla, necesaria para trabajar con el caso Nova Moda

```
%sql
DROP TABLE IF EXISTS Nova;
CREATE TABLE Nova
USING csv
OPTIONS (path "/FileStore/Practica 3/Datos.csv",delimiter ",", header "true");
```

```
from pyspark.sql import SparkSession
from pyspark.sql.types import StructType, StructField, StringType, LongType, IntegerType, DoubleType, FloatType

# Create SparkSession
spark = SparkSession.builder \
      .master("local[1]") \
            .appName("SparkByExamples.com") \
                  .getOrCreate()
df=spark.sql("SELECT * FROM Nova;")
```

```
▸ df: pyspark.sql.dataframe.DataFrame = [Codigo: string, Material: string ... 16 campos adicionales]
Comando ejecutado en 0,69 segundos -- por paperoski@hotmail.com el 17/9/2023, 17:16:53 en «Juan de Dios Delgadillo's Personal Compute Cluster»
```

Revisamos la tabla para ver si se subió los datos correctamente:

```
df.show(5)
```

```
▸ (1) trabajos de Spark
+--------+--------------------+--------------------+--------+-------+------+----------+--------+-------+--------------------+-------+-----+--------+-------+----
---+------------------+----------------+----------------+
|  Codigo|            Material|               Color|    Item|Almacen|Boleta|     Fecha|    Hora|  Marca|             Cliente|T.cajas|Pares|Total($)|Desc($)|Pago
($)|          Vendedor|FECHA DE INGRESO|         Factura|
+--------+--------------------+--------------------+--------+-------+------+----------+--------+-------+--------------------+-------+-----+--------+-------+----
---+------------------+----------------+----------------+
|6429.103|GAMUZA FLEX/LINHO...|NEGRO/NATURAL/DORADO| SCRJ/LC|   SCRZ|91090s|28/04/2023|08:13:48|VIZZANO|JUAN CARLOS-BAUTI...|   0.08|    1|   21.75|      0|  2
1.75|   MELINA-VELASQUEZ|          may-20|0570/2020 - SCOF|
|5618.201|       METAL GLAMOUR|          ORO ROSADO|SCOF/SCZ|   SCRZ|90938s|18/04/2023|08:19:28| MOLECA| LIQUIDACION-CALZADO|   0.08|    1|   21.58|      0|  2
1.58|FALLADOS-DE FABRICA|          oct-18|6237/2017 - SCRZ|
|  5605.1|LONA VERAO/LONA V...|       NEGRO/BLANCO| SCBN/OF|   SCRZ|90938s|18/04/2023|08:19:28| MOLECA| LIQUIDACION-CALZADO|   0.08|    1|   14.75|      0|  1
4.75|FALLADOS-DE FABRICA|          feb-18|6532/2017 - SCRZ|
|5285.316|        CAMURCA FLEX|               NEGRO| TDOE/OF|   SCRZ|90938s|18/04/2023|08:19:28| MOLECA| LIQUIDACION-CALZADO|   0.08|    1|   13.08|      0|  1
3.08|FALLADOS-DE FABRICA|          jul-18|1359/2017 - TDOF|
|5291.329|    VERNIZ SIENA NEO|               NEGRO|SCCVE/OF|   SCRZ|90938s|18/04/2023|08:19:28| MOLECA| LIQUIDACION-CALZADO|   0.08|    1|   11.17|      0|  1
1.17|FALLADOS-DE FABRICA|          mar-18|6685/2017 - SCRZ|
+--------+--------------------+--------------------+--------+-------+------+----------+--------+-------+--------------------+-------+-----+--------+-------+----
---+------------------+----------------+----------------+
only showing top 5 rows
```

Obteniendo la siguiente tabla:

Tabla +

	Cuenta	Nro_movimientos	Total_Dinero_movido	Confiabilidad	Frecuencia	Region	Salario_Promedio	Empresarios_en_miles	Habitantes	Urban
1	813	392	1922343.0999999992	Excelente candidato	Uso mensual	south Bohemia	8427	107	93931	56.9
2	544	254	2191771.8999999994	Excelente candidato	Uso mensual	north Bohemia	9272	118	105058	81.0
3	9869	465	3044394.3	Confiable	Uso mensual	central Bohemia	8754	137	107870	58.0
4	2051	224	2893258.8	Dudoso	Uso semanal	south Moravia	9624	145	197099	74.7
5	7819	591	1298770.8999999994	Confiable	Uso mensual	north Bohemia	8965	104	85852	59.8
6	1843	462	2764436.8999999994	Excelente candidato	Uso mensual	central Bohemia	8754	137	107870	58.0
7	7753	500	6346623.200000005	Excelente candidato	Uso mensual	north Moravia	10673	100	323870	100.0
8	6118	408	3032913.2000000007	Inconfiable	Uso semanal	Prague	12541	167	1204953	100.0
9	2912	369	5610280.7	Confiable	Uso mensual	north Moravia	10673	100	323870	100.0
10	6555	438	4243596.7	Excelente candidato	Uso semanal	north Bohemia	8965	104	85852	59.8
11	5700	355	1451880.1999999995	Dudoso	Uso mensual	east Bohemia	8388	87	95907	59.1

682 filas | 3,52 segundos de tiempo de ejecución — Actualizado hace 1 hora

Verificamos que las extensiones sean correctas:

```
1  df.printSchema()
```

```
root
 |-- Codigo: string (nullable = true)
 |-- Material: string (nullable = true)
 |-- Color: string (nullable = true)
 |-- Item: string (nullable = true)
 |-- Almacen: string (nullable = true)
 |-- Boleta: string (nullable = true)
 |-- Fecha: string (nullable = true)
 |-- Hora: string (nullable = true)
 |-- Marca: string (nullable = true)
 |-- Cliente: string (nullable = true)
 |-- T.cajas: string (nullable = true)
 |-- Pares: string (nullable = true)
 |-- Total($): string (nullable = true)
 |-- Desc($): string (nullable = true)
 |-- Pago($): string (nullable = true)
 |-- Vendedor: string (nullable = true)
 |-- FECHA DE INGRESO: string (nullable = true)
 |-- Factura: string (nullable = true)
```

```
Comando ejecutado en 0,08 segundos -- por paperoski@hotmail.com el 17/9/2023, 17:16:53 en «Juan de Dios Delgadillo's Personal Compute Cluster»
```

Renombramos las columnas para no tener problemas con el código posteriormente (evitando caracteres especiales o reservados):

```
1  df=df.withColumnRenamed("T.cajas", "Total_cajas").withColumnRenamed("Total($)", "Total").withColumnRenamed("Desc($)", "Descuento").withColumnRenamed
   ("Pago($)", "Pago")
2  df.printSchema()
```

```
▸ df:  pyspark.sql.dataframe.DataFrame = [Codigo: string, Material: string ... 16 campos adicionales]
oot
 |-- Codigo: string (nullable = true)
 |-- Material: string (nullable = true)
 |-- Color: string (nullable = true)
 |-- Item: string (nullable = true)
 |-- Almacen: string (nullable = true)
 |-- Boleta: string (nullable = true)
 |-- Fecha: string (nullable = true)
 |-- Hora: string (nullable = true)
 |-- Marca: string (nullable = true)
 |-- Cliente: string (nullable = true)
 |-- Total_cajas: string (nullable = true)
 |-- Pares: string (nullable = true)
 |-- Total: string (nullable = true)
 |-- Descuento: string (nullable = true)
 |-- Pago: string (nullable = true)
 |-- Vendedor: string (nullable = true)
 |-- FECHA DE INGRESO: string (nullable = true)
 |-- Factura: string (nullable = true)
```

```
omando ejecutado en 0,09 segundos -- por paperoski@hotmail.com el 17/9/2023, 17:16:53 en «Juan de Dios Delgadillo's Personal Compute Cluster»
```

En este caso no lo son por lo que cambiamos el tipo de variable por las correctas:

```
1  df2= df.withColumn("Total_cajas", df["Total_cajas"].cast('float')).withColumn("Pares", df["Pares"].cast('float')).withColumn("Total", df["Total"].cast
   ('float')).withColumn("Descuento", df["Descuento"].cast('float')).withColumn("Pago", df["Pago"].cast('float'))
```

```
▸ df2:  pyspark.sql.dataframe.DataFrame = [Codigo: string, Material: string ... 16 campos adicionales]
Comando ejecutado en 0,19 segundos -- por paperoski@hotmail.com el 17/9/2023, 17:16:53 en «Juan de Dios Delgadillo's Personal Compute Cluster»
md 7
```

Python

```
1  df2.printSchema()
```

```
root
 |-- Codigo: string (nullable = true)
 |-- Material: string (nullable = true)
 |-- Color: string (nullable = true)
 |-- Item: string (nullable = true)
 |-- Almacen: string (nullable = true)
 |-- Boleta: string (nullable = true)
 |-- Fecha: string (nullable = true)
 |-- Hora: string (nullable = true)
 |-- Marca: string (nullable = true)
 |-- Cliente: string (nullable = true)
 |-- Total_cajas: float (nullable = true)
 |-- Pares: float (nullable = true)
 |-- Total: float (nullable = true)
 |-- Descuento: float (nullable = true)
 |-- Pago: float (nullable = true)
 |-- Vendedor: string (nullable = true)
 |-- FECHA DE INGRESO: string (nullable = true)
 |-- Factura: string (nullable = true)
```

Eliminamos columnas que no son necesarias para el análisis:

```
df3=df2.drop('FECHA DE INGRESO')
df3=df3.drop('Factura')
df3=df3.drop('Fecha')
df3=df3.drop('Hora')
df3=df3.drop('Total')
df3=df3.drop('Descuento')
df3=df3.drop('Vendedor')

```

Realizamos el tratamiento de los valores nulos, como no se tenían la tabla quedo igual:

```
df3.count()
```

(2) trabajos de Spark

Out[11]: 7664

Comando ejecutado en 1,54 segundos -- por paperoski@hotmail.com el 17/9/2023, 17:16:53 en «Juan de Dios Delgadillo's Personal Compute Cluster»

Cmd 12

```
df3 = df3.na.drop()
```

df3: pyspark.sql.dataframe.DataFrame = [Codigo: string, Material: string ... 9 campos adicionales]

Comando ejecutado en 0,11 segundos -- por paperoski@hotmail.com el 17/9/2023, 17:16:53 en «Juan de Dios Delgadillo's Personal Compute Cluster»

Cmd 13

```
df3.count()
```

(2) trabajos de Spark

Out[13]: 7664

Comando ejecutado en 0,49 segundos -- por paperoski@hotmail.com el 17/9/2023, 17:16:53 en «Juan de Dios Delgadillo's Personal Compute Cluster»

4.2 Agrupamiento de variables Cualitativas

Agrupamos variables cualitativas

Agrupamiento de Variables Cualitativas

```
df3.groupBy(F.col('Codigo')).count().show(5)
df3.groupBy(F.col('Material')).count().show(5)
df3.groupBy(F.col('Color')).count().show(5)
df3.groupBy(F.col('Almacen')).count().show(5)
df3.groupBy(F.col('Boleta')).count().show(5)
df3.groupBy(F.col('Marca')).count().show(5)
df3.groupBy(F.col('Cliente')).count().show(5)
```

```
+--------------------+-----+
|            Material|count|
+--------------------+-----+
|NAPA TURIM/TIRA P...|    7|
|NAPA SAFIANO/NAPA...|   37|
|NAPA TURIM/TRANCA...|    5|
|NAPA FLOTER RUSTI...|    1|
|NP SARD NEO MICR/...|    1|
+--------------------+-----+
only showing top 5 rows

+--------+-----+
|Almacen|count|
+--------+-----+
|    SCRZ| 7664|
+--------+-----+

+--------------------+-----+
|               Marca|count|
+--------------------+-----+
|MOLEKINHO-CALCETINES|   10|
|            ACTVITTA|  480|
|             VIZZANO| 1861|
|     MODARE-CARTERAS|   46|
|            BEIRARIO| 1035|
+--------------------+-----+
only showing top 5 rows

+--------------------+-----+
|               Color|count|
+--------------------+-----+
|   ROSA 871/ROSA 942|    8|
|   NEGRO/PLATA/NEGRO|    2|
|   MULTI BEIGE/BEIGE|    3|
|       DORADO/DORADO|    2|
|CARAMELO 876/MARI...|    1|
+--------------------+-----+

+-------+-----+
|Boleta|count|
+-------+-----+
|90134n|    1|
|90814n|    1|
|90438n|    2|
|89656n|    3|
|91310n|    3|
+-------+-----+

+--------------------+-----+
|             Cliente|count|
+--------------------+-----+
|MUHAMMAD-ANWAR HO...|   37|
|EVA LEANDRO-MENAC...|   22|
|MARIA-RIBERA DE G...|   11|
| MIRIAN-TITO CABRERA|   14|
|YESSICA-MOSTASEDO...|   21|
+--------------------+-----+
```

4.3 Agrupamiento de variables cuantitativas

Agrupamos variables cuantitativas y las describimos por el método describe (), para obtener sus valores estadisticos:

Agrupamiento de Variables Cuantitativas

```
1  df3.select(['Total_cajas','Pares','Pago']).describe().show(10)
```

▸ (2) trabajos de Spark

```
+-------+------------------+-----------------+-----------------+
|summary|       Total_cajas|            Pares|             Pago|
+-------+------------------+-----------------+-----------------+
|  count|              7664|             7664|             7664|
|   mean| 0.8937369529896714|10.724034446764092|197.8616354105866|
| stddev|0.27556288613713553|3.3037005429009048|92.25522160223584|
|    min|              0.08|              1.0|             2.83|
|    max|              2.08|             25.0|            600.0|
+-------+------------------+-----------------+-----------------+
```

5 Modelado

El modelado de machine Learning desempeña un papel fundamental en la gestión y el éxito de una empresa en la era actual de la tecnología y la información. La importancia radica en su capacidad para convertir los datos en información procesable y conocimiento valioso. Al aplicar algoritmos de machine Learning a los datos de una empresa, se pueden lograr varios beneficios significativos. Esto puede ayudar a la empresa Nova Moda a obtener conocimiento acerca de las preferencias de los clientes al momento de realizar un compra.

5.1 Aprendizaje No Supervisado

El aprendizaje no supervisado es una rama del machine Learning en la que se utilizan datos no etiquetados para descubrir patrones y estructuras ocultas. A diferencia del aprendizaje supervisado, no hay una salida objetivo conocida, lo que significa que el modelo busca agrupar datos similares o reducir la dimensionalidad para revelar relaciones intrínsecas entre los datos. En este caso nuestra variable objetivo será la marca del zapato de preferencia del cliente, en base a la siguiente tabla:

```
1  df3.sort(F.desc('Total')).show(10)
```

▸ (1) trabajos de Spark

```
+--------+--------------------+--------------------+-----------+-------+------+--------+--------------------+-----------+-----+------+
|  Codigo|            Material|               Color|       Item|Almacen|Boleta|   Marca|             Cliente|Total_cajas|Pares|  Pago|
+--------+--------------------+--------------------+-----------+-------+------+--------+--------------------+-----------+-----+------+
|7377.108|NAPA FLOATHER NATURE|               BEIGE|SCBB/OF-EST|   SCRZ|91481n|  MODARE|    CATALINA-CONDORI|        2.0| 24.0| 600.0|
|3094.101|NAPA GENEBRA/CAMU...|      NEGRO /GRAFITO|    SCMY/NB|   SCRZ|89794n| VIZZANO| NANCI-BENITO COLQUE|        1.0| 12.0|588.96|
|3096.102|NAPA BERLIM/ELAST...|BRANCO OFF 526/CR...|    SCMY/NB|   SCRZ|89794n| VIZZANO| NANCI-BENITO COLQUE|        1.0| 12.0|585.96|
|3096.102|NAPA BERLIM/ELAST...|PRETO 01/PRETO/CR...|    SCMY/NB|   SCRZ|89794n| VIZZANO| NANCI-BENITO COLQUE|        1.0| 12.0|585.96|
|9043.105|NAPA GENEBRA/LYCR...|       CAFE 856/CAFE|    SCAN/LM|   SCRZ|92885n|BEIRARIO|LUDMILA SILVIA-MA...|        1.0| 12.0|584.04|
|9043.105|NAPA GENEBRA/LYCR...|               NEGRO|    SCAN/LM|   SCRZ|92885n|BEIRARIO|LUDMILA SILVIA-MA...|        1.0| 12.0|584.04|
|8290.725|      VERNIZ PREMIUM|               BEIGE|    SCLA/NO|   SCRZ|91343n|BEIRARIO| ROXANA-SALAZAR RIOS|        2.0| 24.0|574.08|
|8290.725|      VERNIZ PREMIUM|               NEGRO|    SCPA/AG|   SCRZ|91343n|BEIRARIO| ROXANA-SALAZAR RIOS|        2.0| 24.0|574.08|
|1389.101|              PELICA|          BLANCO 99|SCRR/OF-EST|   SCRZ|90189n| VIZZANO| ANA-BARRETO RAMIREZ|        2.0| 24.0|571.92|
|  3094.1|NAPA FLOATHER ZUR...|            CAFE 987|     SC/BOL|   SCRZ|91732n| VIZZANO|JOSE ANTENOR-ZEGA...|        1.0| 12.0|552.96|
+--------+--------------------+--------------------+-----------+-------+------+--------+--------------------+-----------+-----+------+
only showing top 10 rows
```

Donde la marca (tipo de zapato) es nuestra variable Y o etiqueta para el aprendizaje supervisado, el cual sugerirá la preferencia de los clientes para poder prever inventario antes de la venta y poder realizar pedidos con una anticipación mayor.

Realizamos la transformación One Hot encoding para las variables cualitativas:

```
Codigo_indexer = StringIndexer(inputCol= 'Codigo', outputCol= 'CodigoIndex')
Codigo_encoder = OneHotEncoder(inputCol = 'CodigoIndex', outputCol= 'CodigoVec')

Material_indexer = StringIndexer(inputCol= 'Material', outputCol= 'MaterialIndex')
Material_encoder = OneHotEncoder(inputCol = 'MaterialIndex', outputCol= 'MaterialVec')

Color_indexer = StringIndexer(inputCol= 'Color', outputCol= 'ColorIndex')
Color_encoder = OneHotEncoder(inputCol = 'ColorIndex', outputCol= 'ColorVec')

Item_indexer = StringIndexer(inputCol= 'Item', outputCol= 'ItemIndex')
Item_encoder = OneHotEncoder(inputCol = 'ItemIndex', outputCol= 'ItemVec')

Boleta_indexer = StringIndexer(inputCol= 'Boleta', outputCol= 'BoletaIndex')
Boleta_encoder = OneHotEncoder(inputCol = 'BoletaIndex', outputCol= 'BoletaVec')

Marca_indexer = StringIndexer(inputCol= 'Marca', outputCol= 'MarcaIndex')
Marca_encoder = OneHotEncoder(inputCol = 'MarcaIndex', outputCol= 'MarcaVec')

Cliente_indexer = StringIndexer(inputCol= 'Cliente', outputCol= 'ClienteIndex')
Cliente_encoder = StringIndexer(inputCol= 'ClienteIndex', outputCol= 'ClienteVec')
```

Comando ejecutado en 0,52 segundos -- por paperoski@hotmail.com el 17/9/2023, 17:16:53 en «Juan de Dios Delgadillo's Personal Compute Cluster»

Definimos los features para el modelo:

```
assembler = VectorAssembler(inputCols = ['CodigoVec','MaterialVec','ColorVec','ItemVec', 'BoletaVec','MarcaVec','ClienteVec','Pares','Total_cajas',
'Pago'], outputCol= 'features')
```

Dividimos los datos en training y testing:

```
train_data, test_data = df3.randomSplit([0.7,0.3])
```

▸ train_data: pyspark.sql.dataframe.DataFrame = [Codigo: string, Material: string ... 9 campos adicionales]
▸ test_data: pyspark.sql.dataframe.DataFrame = [Codigo: string, Material: string ... 9 campos adicionales]

Comando ejecutado en 0,13 segundos -- por paperoski@hotmail.com el 17/9/2023, 17:18:59 en «Juan de Dios Delgadillo's Personal Compute Cluster»

Realizamos un pipeline para ejecutar el modelo:

```
pipeline2 = Pipeline(stages= [Codigo_indexer ,Codigo_encoder ,Material_indexer ,Material_encoder ,Color_indexer ,Color_encoder ,Item_indexer,
Item_encoder  Marca_indexer,Boleta_indexer ,Boleta_encoder     Marca_encoder,Cliente_indexer , Cliente_encoder  assembler,  kmeans])
```

Entrenamos el modelo

```
fit_model2 = pipeline2.fit(train_data)
```

▸ (34) trabajos de Spark

▾ (1) ejecución de MLflow

Se ha registrado 1 ejecución de un experimento en MLflow. Más información

Mostramos los resultados de la predicción:

```
predictions2.show(10)
```

▸ (1) trabajos de Spark

```
+-------+--------------------+--------------------+-----------+-------+------+----------------+--------------------+-----------+-----+-----+----------+--------
-------+--------------------+----------+
| Codigo|            Material|               Color|       Item|Almacen|Boleta|           Marca|             Cliente|Total_cajas|Pares| Pago|MarcaIndex|      Ma
rcaVec|            features|prediction|
+-------+--------------------+--------------------+-----------+-------+------+----------------+--------------------+-----------+-----+-----+----------+--------
-------+--------------------+----------+
|10002.1|   NAPA SARDENHA NEO|  BLANCO 99/NEGRO 01|    SCAC/OF|   SCRZ|90281n|VIZZANO-CARTERAS|  CARMEN-DIAZ MUÃ‘OZ|       0.08|  1.0|27.25|       8.0|(17,[8],
[1.0])|(20,[8,17,18,19],...|         1|
|10002.1|   NAPA SARDENHA NEO|  BLANCO 99/NEGRO 01|SCPA/LT-PRO|   SCRZ|90288n|VIZZANO-CARTERAS| MARTHA ALICIA-LIZME|       0.25|  3.0|81.75|       8.0|(17,[8],
[1.0])|(20,[8,17,18,19],...|         4|
|10003.2|NAPA SOFT STRECH ...|           CREMA 985|    CBAM/BG|   SCRZ|90395n|VIZZANO-CARTERAS|          REMI-PEREZ|       0.08|  1.0|22.25|       8.0|(17,[8],
[1.0])|(20,[8,17,18,19],...|         1|
|10003.2|NAPA SOFT STRECH ...|               NEGRO|    SCLV/OF|   SCRZ|90310n|VIZZANO-CARTERAS|BERTHA CHOQUE (ER...|       0.17|  2.0| 44.5|       8.0|(17,[8],
[1.0])|(20,[8,17,18,19],...|         1|
|10003.2|NAPA SOFT STRECH ...|               NEGRO|    SCMY/OF|   SCRZ|90300n|VIZZANO-CARTERAS|VANESSA -DELGADO ...|       0.17|  2.0| 44.5|       8.0|(17,[8],
[1.0])|(20,[8,17,18,19],...|         1|
|10003.2|NAPA SOFT STRECH ...|               NEGRO|    SCMY/OF|   SCRZ|90302n|VIZZANO-CARTERAS|          REMI-PEREZ|       0.17|  2.0| 44.5|       8.0|(17,[8],
[1.0])|(20,[8,17,18,19],...|         1|
|10017.2|NAPA SOFT STRECH ...|               NEGRO|    SCAC/OF|   SCRZ|90952n|VIZZANO-CARTERAS|BERTHA CHOQUE ROJ...|       0.17|  2.0|41.66|       8.0|(17,[8],
[1.0])|(20,[8,17,18,19],...|         1|
|10018.2|VERSALES BAG/NAPA...|MULTICOLOR/CORAL ...|    SCMO/YM|   SCRZ|90055n|VIZZANO-CARTERAS|YESSICA-MOSTASEDO...|       0.08|  1.0|23.58|       8.0|(17,[8],
```

Comando ejecutado en 0,79 segundos -- por paperoski@hotmail.com el 17/9/2023, 17:21:16 en «Juan de Dios Delgadillo's Personal Compute Cluster»

6 Evaluación del modelo

Finalmente evaluamos el modelo con el coeficiente Silhouette:

7 Conclusiones

El proyecto de modelado mediante el algoritmo de K-Means en el aprendizaje automático no supervisado ha demostrado ser una herramienta valiosa para identificar patrones y agrupar datos de manera eficiente. A través de este enfoque, pudimos comprender mejor la estructura subyacente en nuestros datos y segmentarlos en grupos significativos. Esto no solo nos permitió realizar análisis más profundos y descubrimientos interesantes, sino que también nos proporcionó una base sólida para tomar decisiones a la hora de gestionar el inventario de la empresa Nova Moda y así anticipar los productos a importar en base a las preferencias de los clientes.

8 Recomendaciones

Link del repositorio de GitHub: https://github.com/HebertDelgadillo/cursodatabricksmod4.git