Data science en e-Sales

Diciembre 2021

ELÍAS ACTIS GROSSO
CLAUDIO SEBASTIÁN LIO
ALEX SILVA

CoderHouse
Data Science
Comisión 14075

Tabla de contenidos

Versionado

Versión	Fecha	Descripción
1.0	27/09/2021	EDA, Análisis univariado, bivariado y multivariado
1.1	25/10/2021	Implementación de algoritmos k-means y clasificación
1.2	10/11/2021	Agregado de métricas de evaluación de modelos y optimización de hipérparámetros
1.3	01/12/2021	Guardado del modelo, conclusiones, detalles finales

Descripción del caso de negocio

Dash deportes es una compañía de retail deportivo con más de 40 años de historia en el mercado argentino. Cuentan con una extensa cobertura nacional, con más de 70 locales propios distribuidos a lo largo de todo el territorio argentino, y desde 2019 están incursionando fuertemente en la **venta a través de canales de e-commerce** desde sus **tiendas propias (Dash, Grid, Mark)** en la plataforma **Vtex así también como en Mercado Libre.**

Entre las tiendas que venden productos a través de los e-commerce se destacan:

- La tienda **Dash**, que apunta a un sector medio / medio bajo con precios más económicos y variedad de marcas y modelos, en varios casos de lanzamientos pasados o productos de trayectoria y de uso deportivo más general
- La tienda **Grid**, que tiene las marcas más exclusivas de moda deportiva, y es la que tiene la mayor cantidad de productos de lanzamiento
- La tienda **Mark Sports**, que apunta al público deportista de alto perfil y rendimiento dando una vidriera virtual orientada por disciplina

La empresa actualmente **cuenta con gran cantidad de datos** de ventas pasadas, pero no realiza ningún análisis sobre los mismos, por lo que quieren **empezar a explotar esa información para tomar mejores decisiones** de inversión, marketing y publicidad, distribución de los productos, etc.

Objetivos del modelo

- Analizar comportamientos y tendencias de compra
- Analizar recurrencia de compras
- Encontrar patrones entre los consumidores o los productos que adquieren
- Determinar el e-commerce más conveniente para publicar un artículo determinado

Descripción de los datos

Se utilizaron 2 Datasets con **ventas desde el 01/01/2021 al 01/03/2021 de ambas plataformas (vtex y Meli),** que se conectan a través del campo ecomm_order_id

Dataset 1

Contiene la información de ventas a nivel operativo dentro de la base de datos de la empresa. Compuesto por las siguientes columnas:

Columna	Nombre	Descripción
1	ecomm_order_id	ID del pedido
2	ecommerce	ecommerce a través del cual se realizó la venta (vtex o meli)
3	store	Tienda que publicó el artículo vendido
4	ecomm_creation_date	Fecha en la que se realizó el pedido
5	numero_lote	Número interno de armado del pedido
6	fecha_facturado	Fecha en la que se facturó el pedido
7	linea	Categoría del artículo vendido
8	marca	Marca del artículo vendido
9	vArticulo_id	Código interno del artículo vendido
10	vTalle_Codigo	Talle del artículo vendido
11	producto	Descripción del artículo vendido
12	quantity	Cantidad vendida
13	client_price	Precio que pagó el cliente por el artículo
14	PrecioCosto	Costo del producto para la empresa
15	ecomm_tipo_envio	Si es a domicilio o punto de retiro
16	VArticuloTalle_Costo	Costo del producto para la empresa (debería ser igual a PrecioCosto)
17	VArticuloTalle_PrecioRegular	Valor de venta sin promociones ni descuentos
18	ecomm_transporte_nombre	Forma de envío
19	sucursal_original	Sucursal a la cuál se realiza el envío
20	ultima_sucursal	Sucursal a donde se envió el producto por última vez
21	ColorPrimario	Color principal del artículo vendido
22	ProveedorId	Código interno del proveedor
23	Disciplina	Ámbito de uso del artículo vendido
24	Genero	Género del artículo vendido
25	sex	Sexo del comprador

Dataset 2

Contiene la información estandarizada de los json de ambas plataformas (MercadoLibre y Vtex). Compuesto por las siguientes columnas:

Columna	Nombre	Descripción
1	description	ecommerce a través del cual se realizó la venta (vtex o meli)
2	ecomm_order_id	ID del pedido
3	ecomm_creation_date	Fecha en la que se realizó el pedido
4	date_handling	Fecha de cuando la empresa empezó a trabajar el pedido
5	date_invoiced	Fecha en la que se facturó
6	email	email del comprador (encriptado)
7	adress_id	Id interno de la dirección de entrega
8	latitude	Ubicación geográfica de entrega: Latitud
9	longitude	Ubicación geográfica de entrega: Longitud
10	payment	Método de pago
11	client_id	Id único que identifica a cada cliente

Transformaciones en los datos

1. **Se eliminaron** las siguientes columnas que no son de utilidad para el análisis:
   - numero_lote: no propociona información útil
   - fecha_facturado: contiene la misma información que date_invoiced (con minutos de diferencia), pero menos cantidad de registros
   - PrecioCosto: Debería contener la misma información que VArticuloTalle_Costo. Nos quedamos con la segunda porque tiene mayor cantidad de registros
   - sucursal_original: contiene muy pocos datos
   - ultima_sucursal: contiene muy pocos datos
   - description: es la descripción del ecommerce, o sea que contiene la misma información que el campo "ecommerce"
   - ecomm_creation_date_y: es la misma fecha que ecomm_creation_date_x, quedó duplicada al unir los 2 data sources
   - ecomm_tipo_envio: No proporciona información útil

2. Se realizó un proceso de Data Wrangling **normalizando las variables** que presentaban algunas de las siguientes situaciones:
   - Valores nulos
   - Espacios en blanco
   - Valores con el mismo significado escritos de forma distinta
   - Valores negativos en el precio (errores de carga)

Las columnas normalizadas fueron las siguientes:

- client_price
- ecommerce
- store
- línea
- marca

- vTalle_Codigo
- ecomm_transporte_nombre
- ColorPrimario
- Disciplina
- Genero
- payment
- sex

3. **Se agregó al dataset la variable "Ganancia"** (precio – costo) para generar reportes y análisis en base a la misma.
   El cálculo de esta se obtuvo de client_price - VArticuloTalle_Costo

4. **Se seleccionaron las siguientes variables** para utilizar en los algoritmos:
   - client_id
   - client_price
   - store
   - linea
   - ecommerce

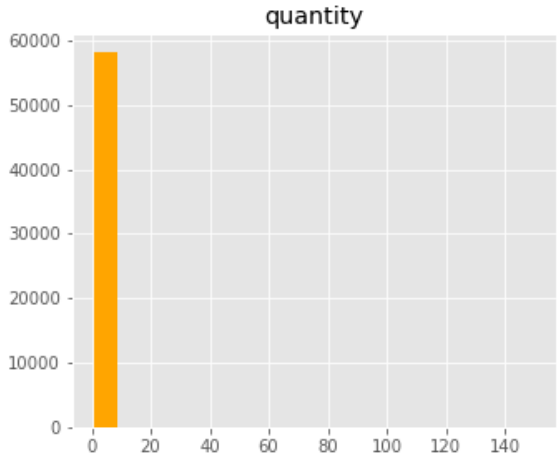
Para el algoritmo "K-means", además, se realizó una agrupación por "client_id" generando un nuevo dataframe con las siguientes columnas:

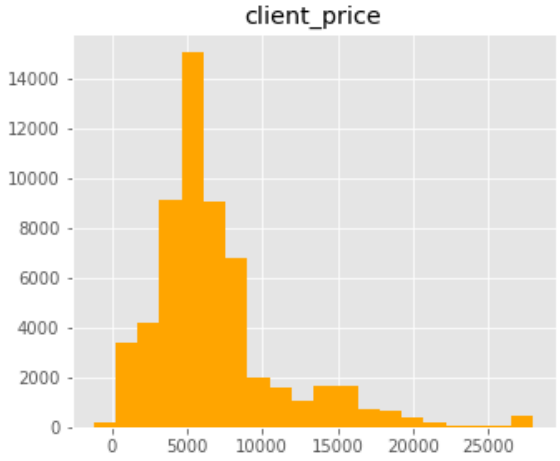
- spent: suma del precio de todos los artículos comprados por el cliente
- transactions: suma de la cantidad de compras del cliente

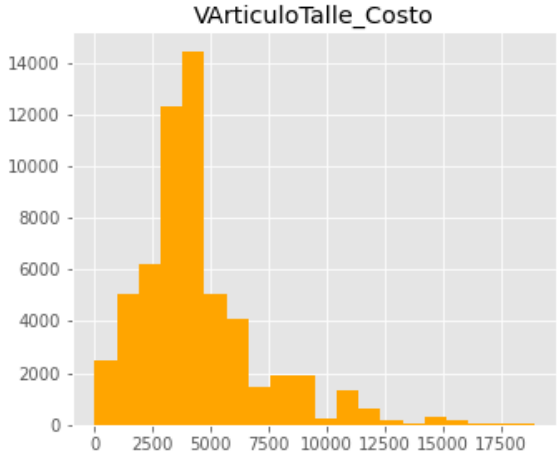
Hallazgos encontrados por el EDA

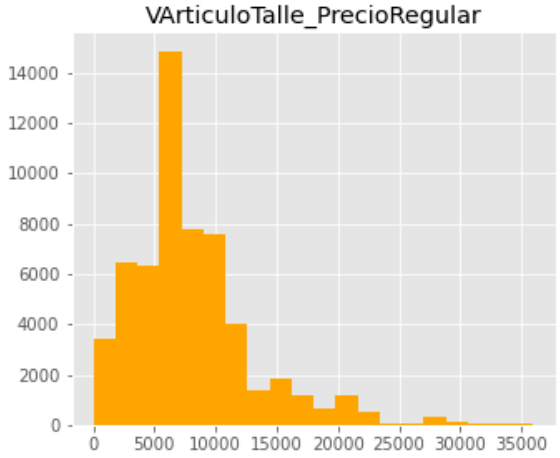
Mediante el **análisis exploratorio de datos** y combinando las distintas variables se pudo observar lo siguiente:

Histograma de variables cuantitativas



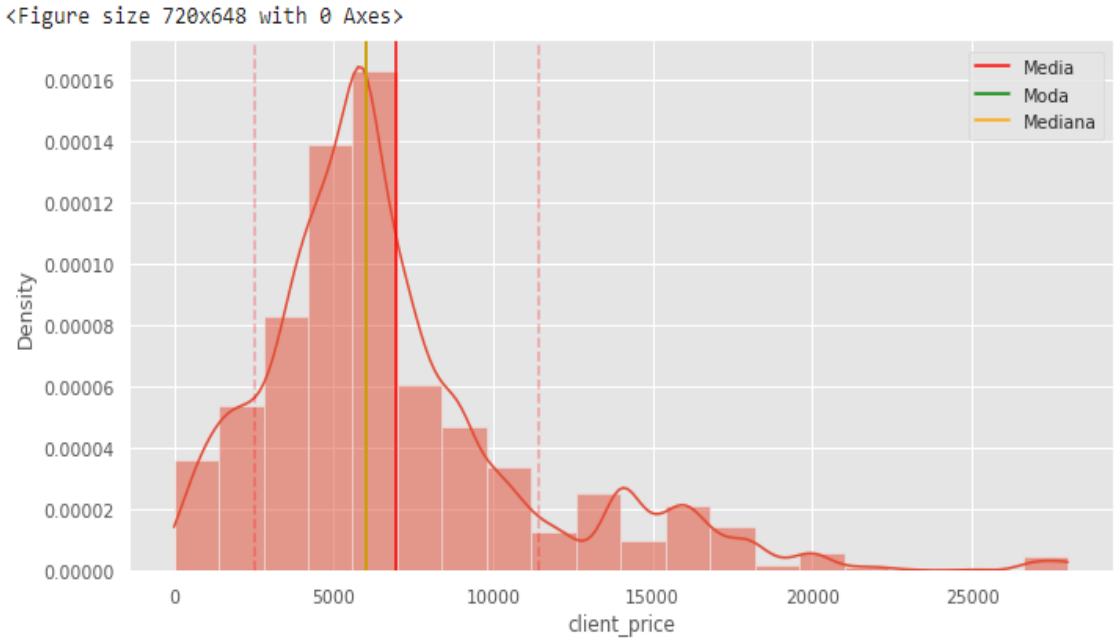




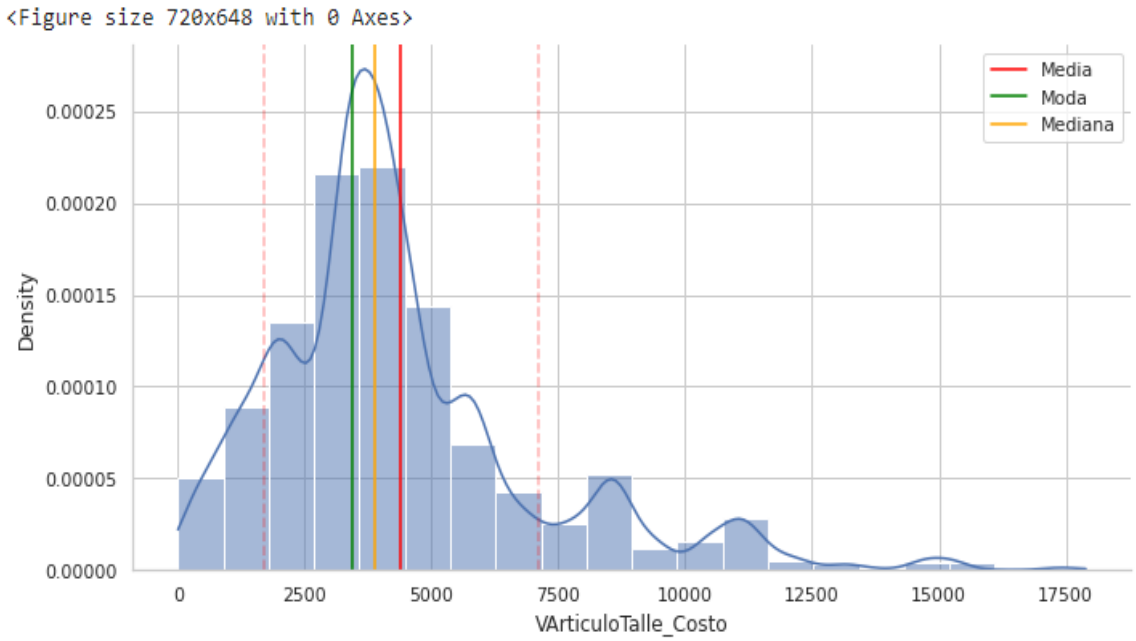


La variable client_price tiene valores negativos, lo cual no tiene sentido y constituye un error de carga en los datos. Por lo tantos, dichos registros se excluyeron del análisis.

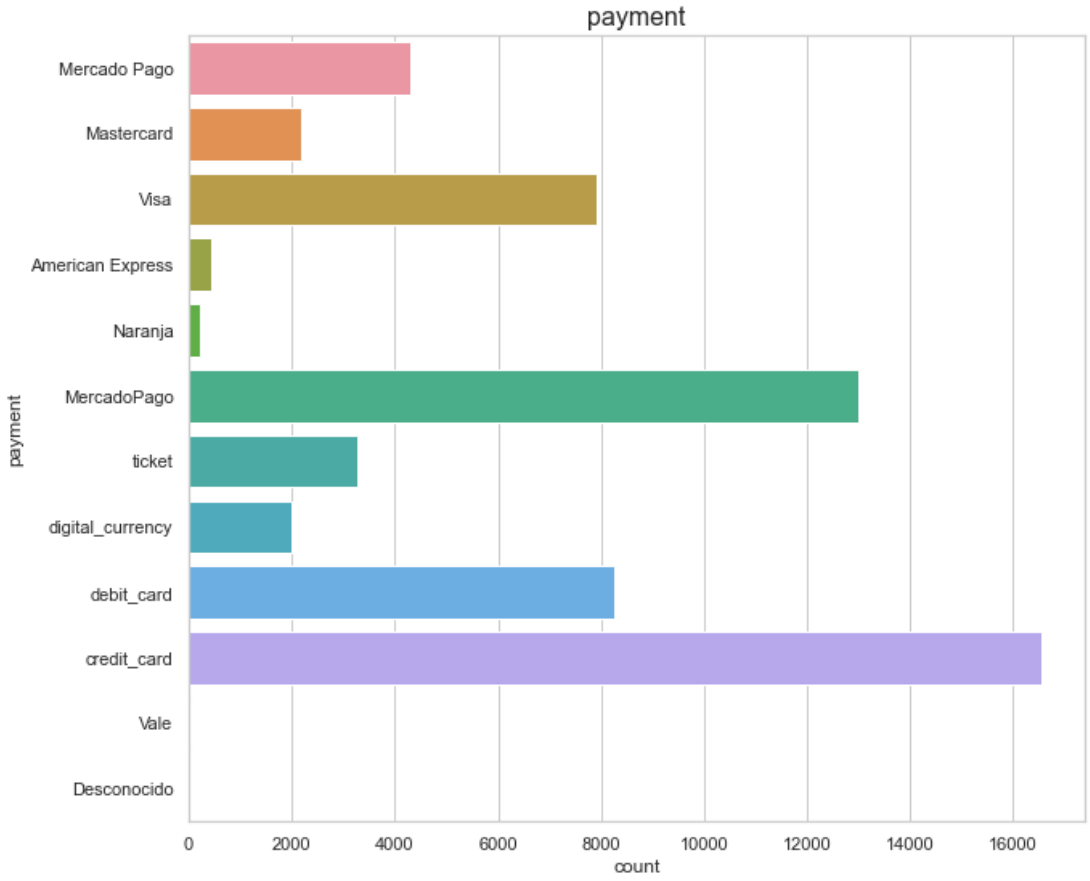
Histograma y curva de densidad del Precio con media, mediana y moda



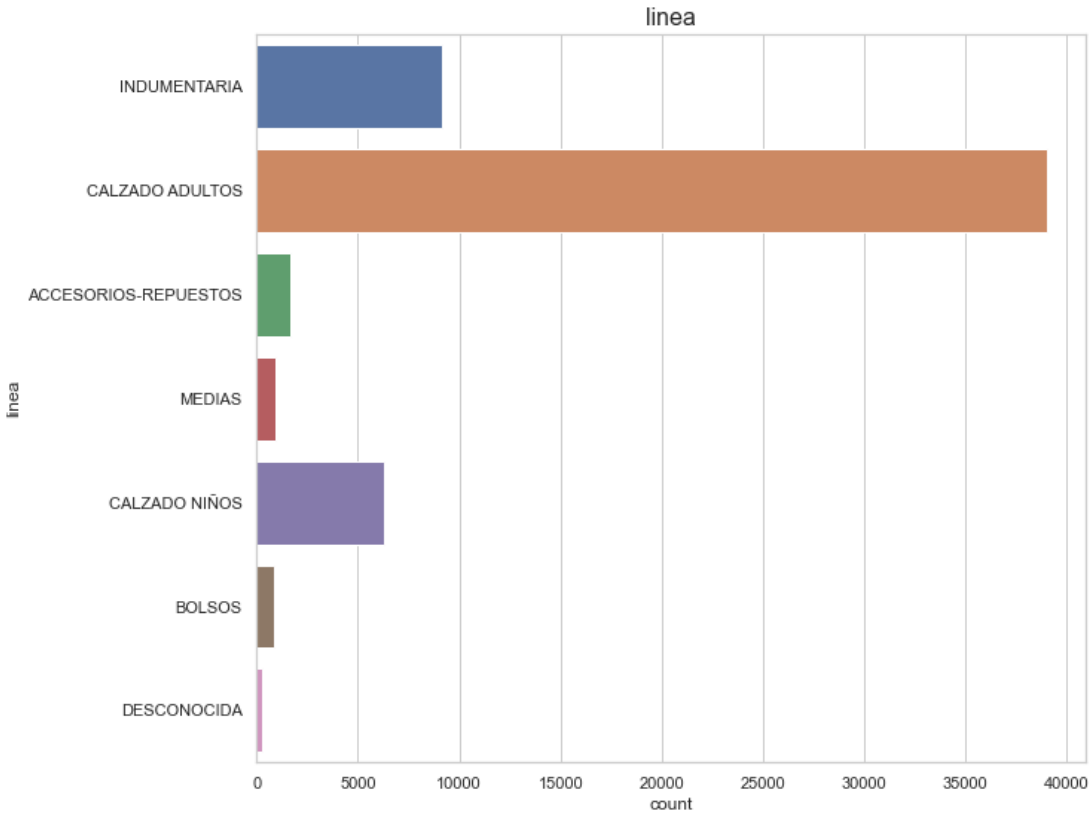
Histograma y curva de densidad del Costo con media, mediana y moda



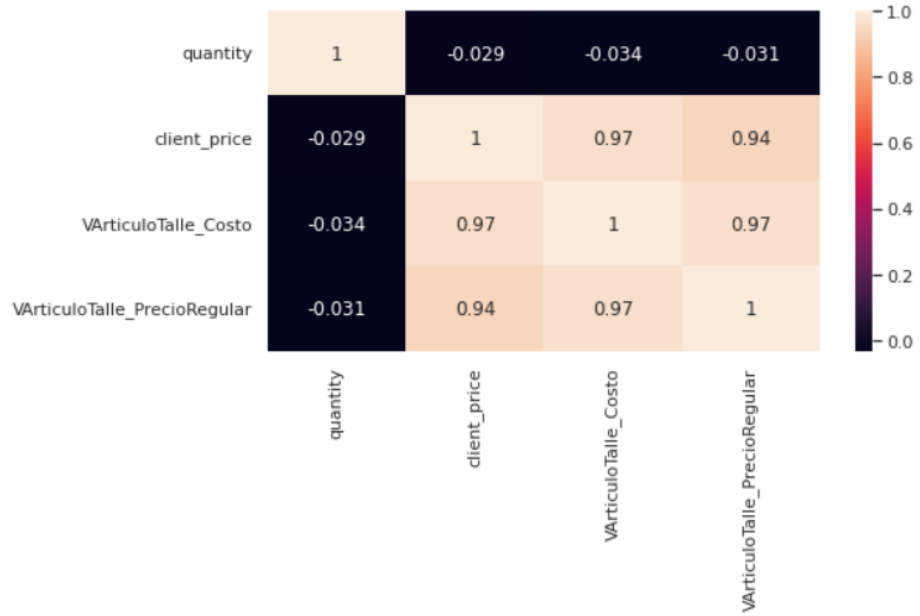
Ventas por medio de pago

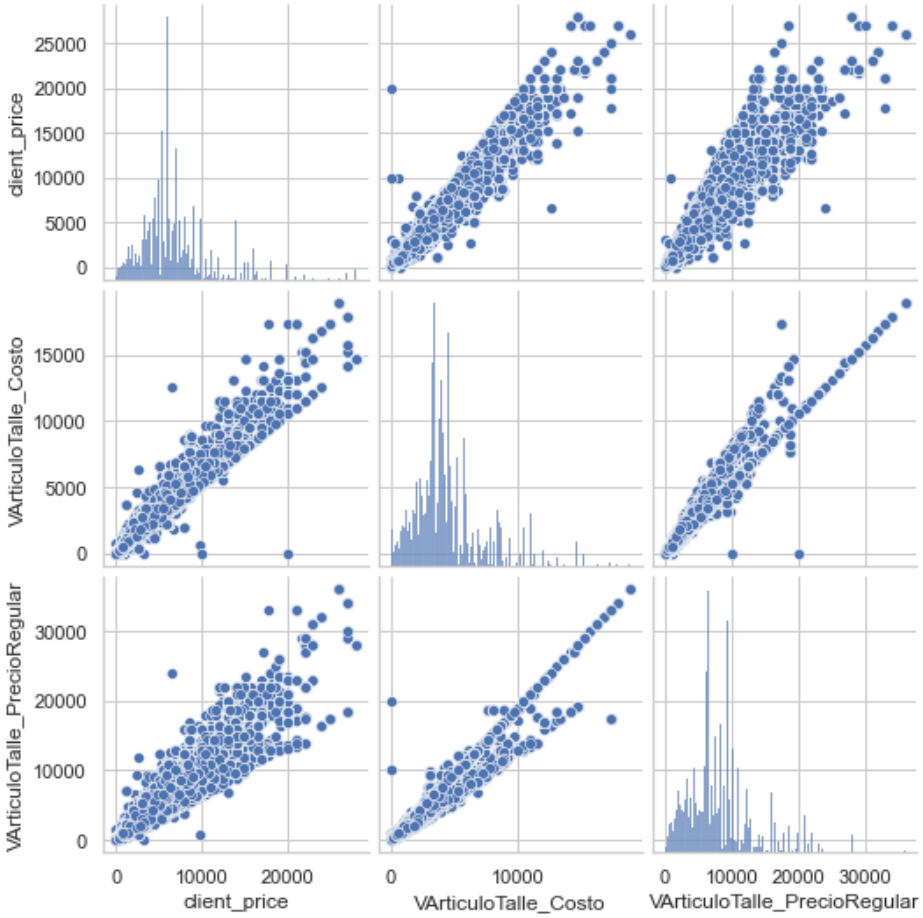


Ventas por línea



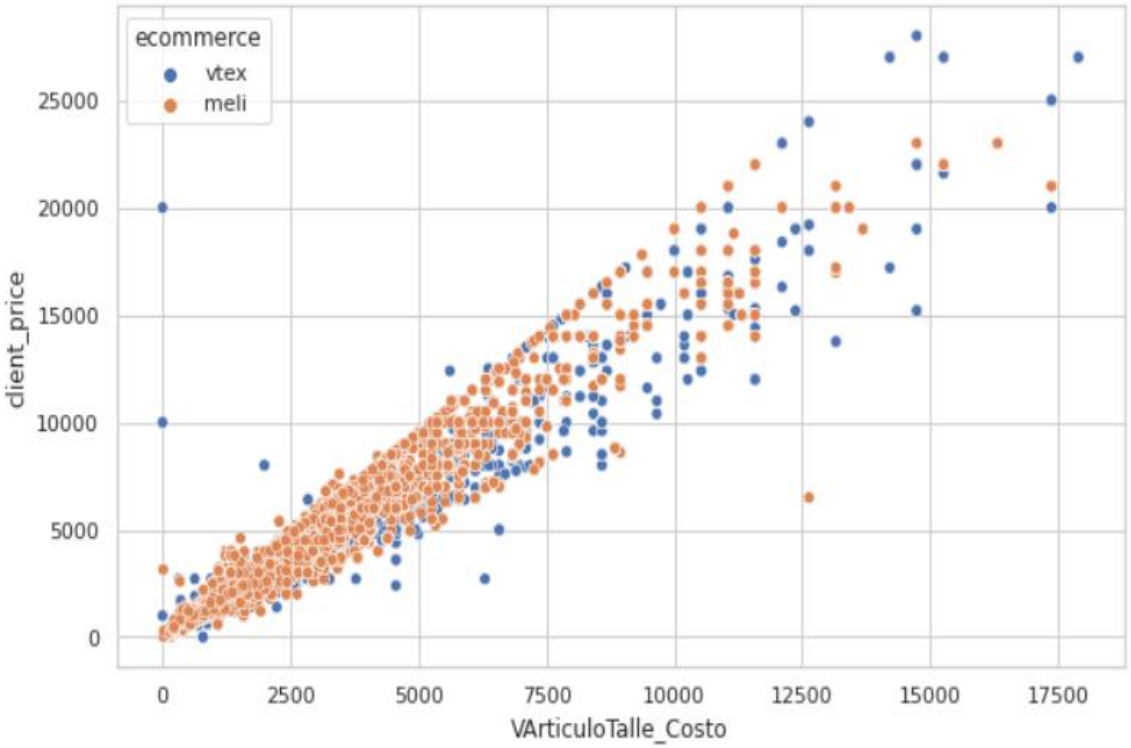
Correlación entre variables cuantitativas





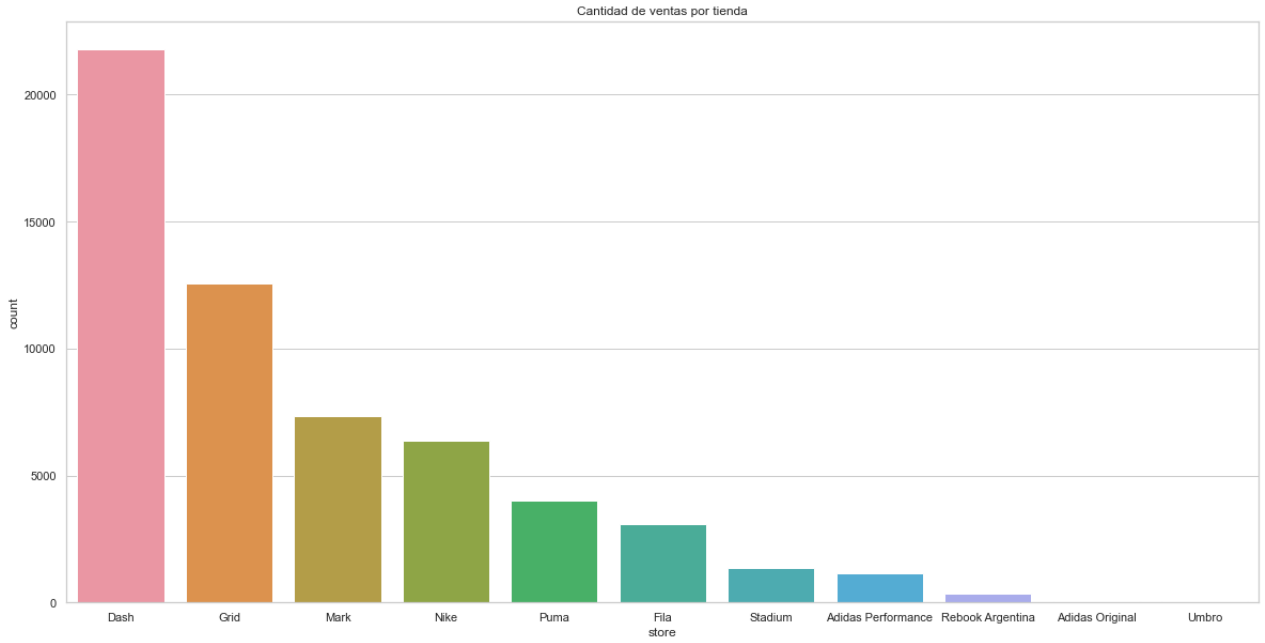
Todas las variables de precio y costo están altamente correlacionadas positivamente (si una crece la otra también), lo cual da sentido al análisis ya que la ganancia se genera vendiendo a un precio superior al costo.

Relación entre costo y precio por e-commerce

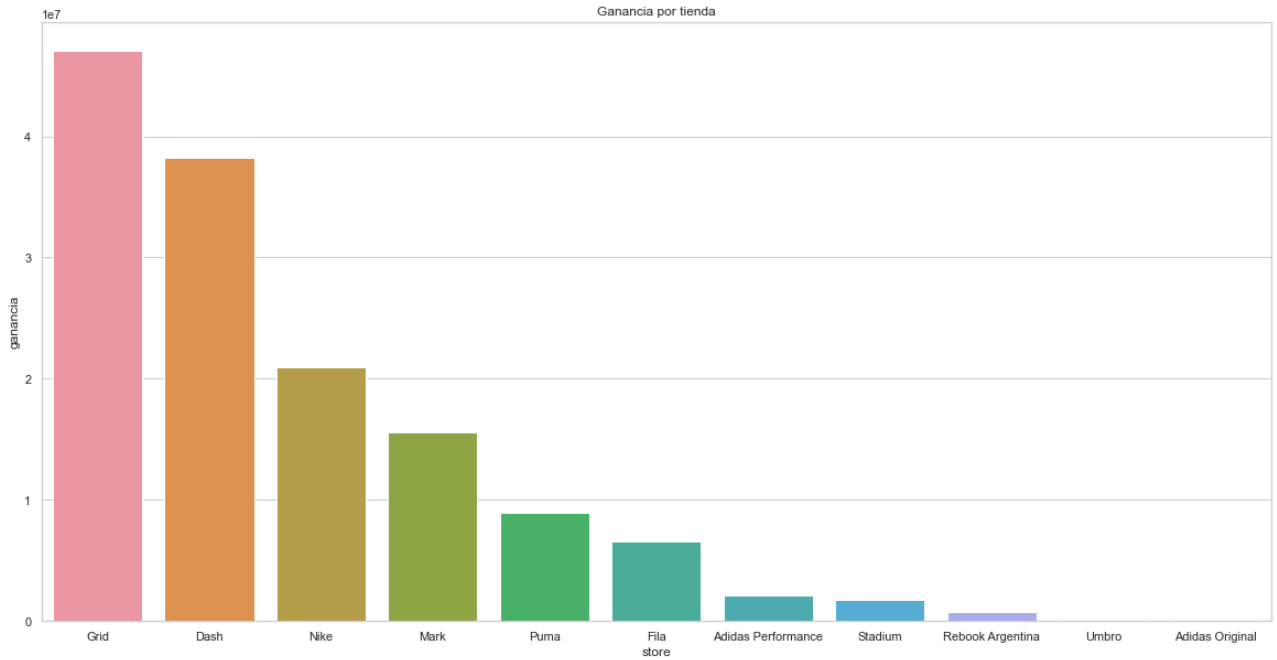


Para costos similares Meli en general vende a un mayor precio, por lo que es probable que genere mayor ganancia. Hay algunos casos atípicos a analizar donde el costo es 0 y el artículo se vendió con un precio determinado. Meli tiene mayor concentración de puntos en precios menores a 10000, por lo que su mayor cantidad de ventas de productos rondan ese rango de precios.

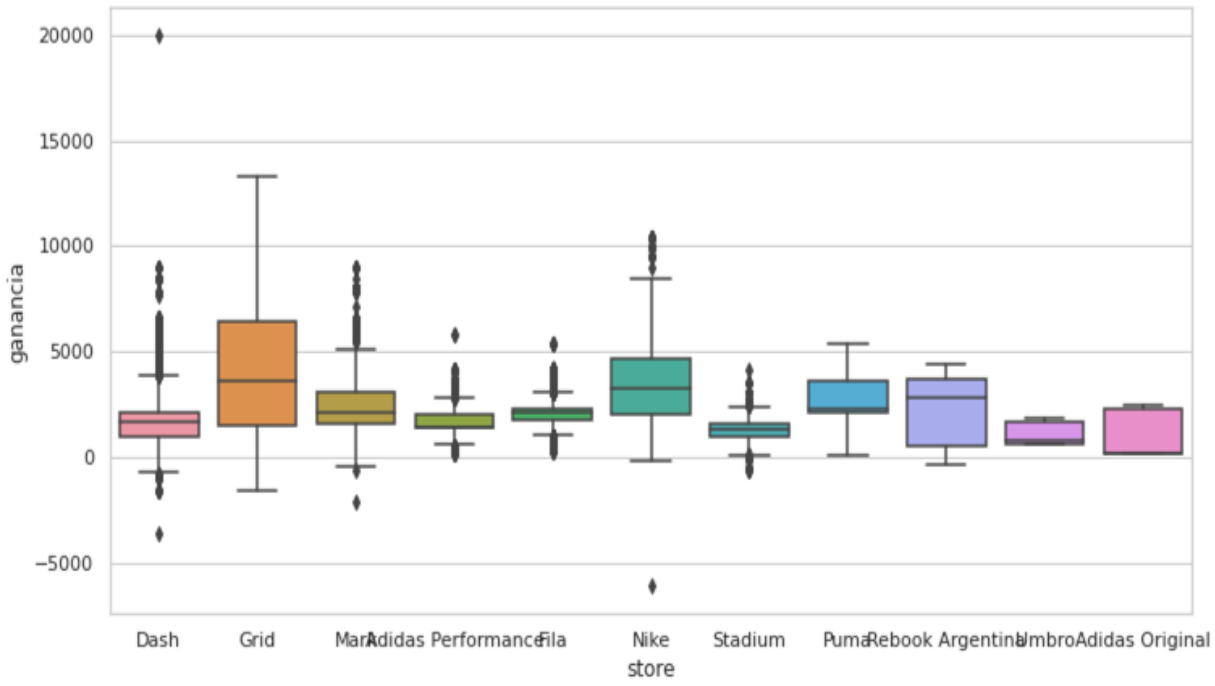
Cantidad de ventas por tienda



Ganancias por tienda



Distribución de ganancias por tienda

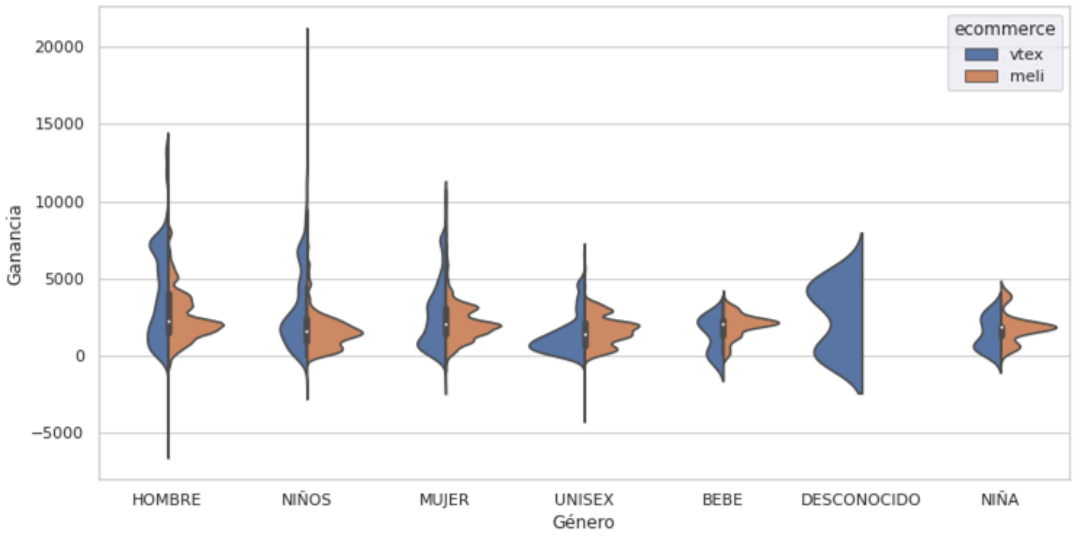


Se puede notar que:

- La tienda "Grid" es la que mayor ganancia genera. La segunda que mayor ganancia genera es "Dash", pero de acuerdo al gráfico tiene en promedio ganancias menores a varias de las tiendas y algunas pérdidas considerables a analizar. Ambas tienen pérdidas por las promociones que ofrecen.

- "Nike" tiene un caso de pérdida mayor a 5000 que habría que analizar para ver si corresponde a un error de datos o hubo realmente una venta con esa pérdida

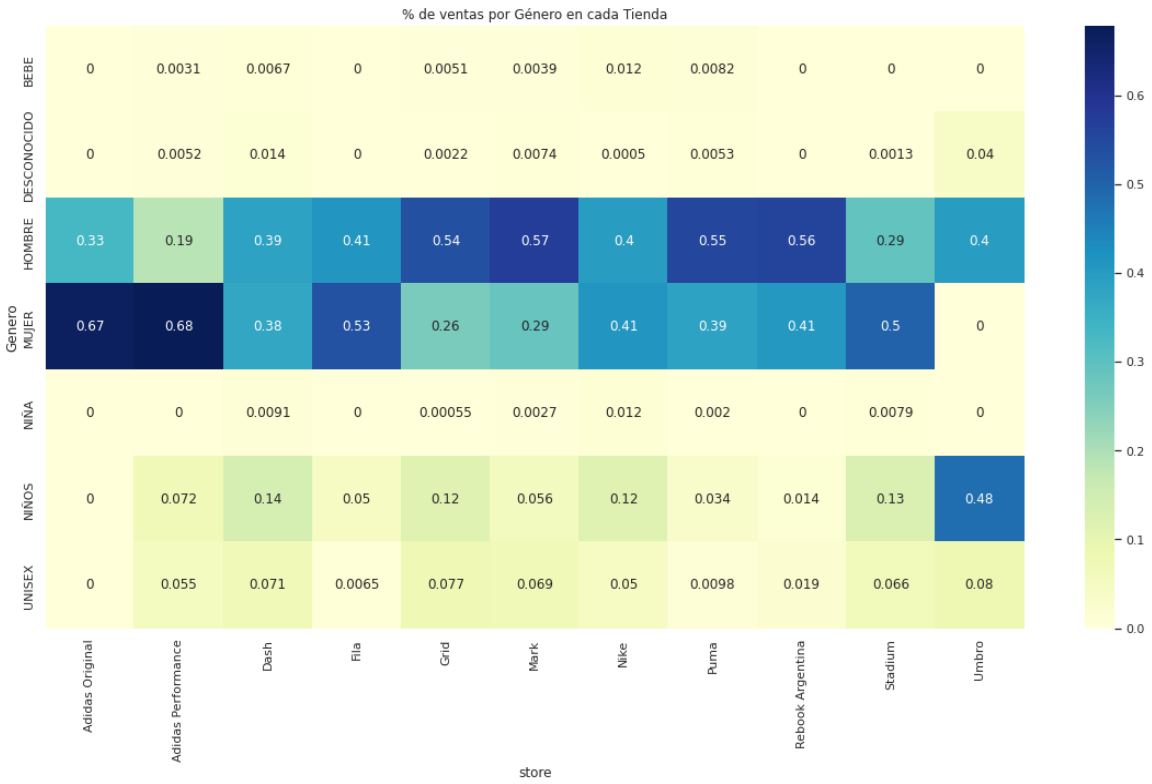
Distribución de ganancias por ecommerce y género



Se puede notar que:

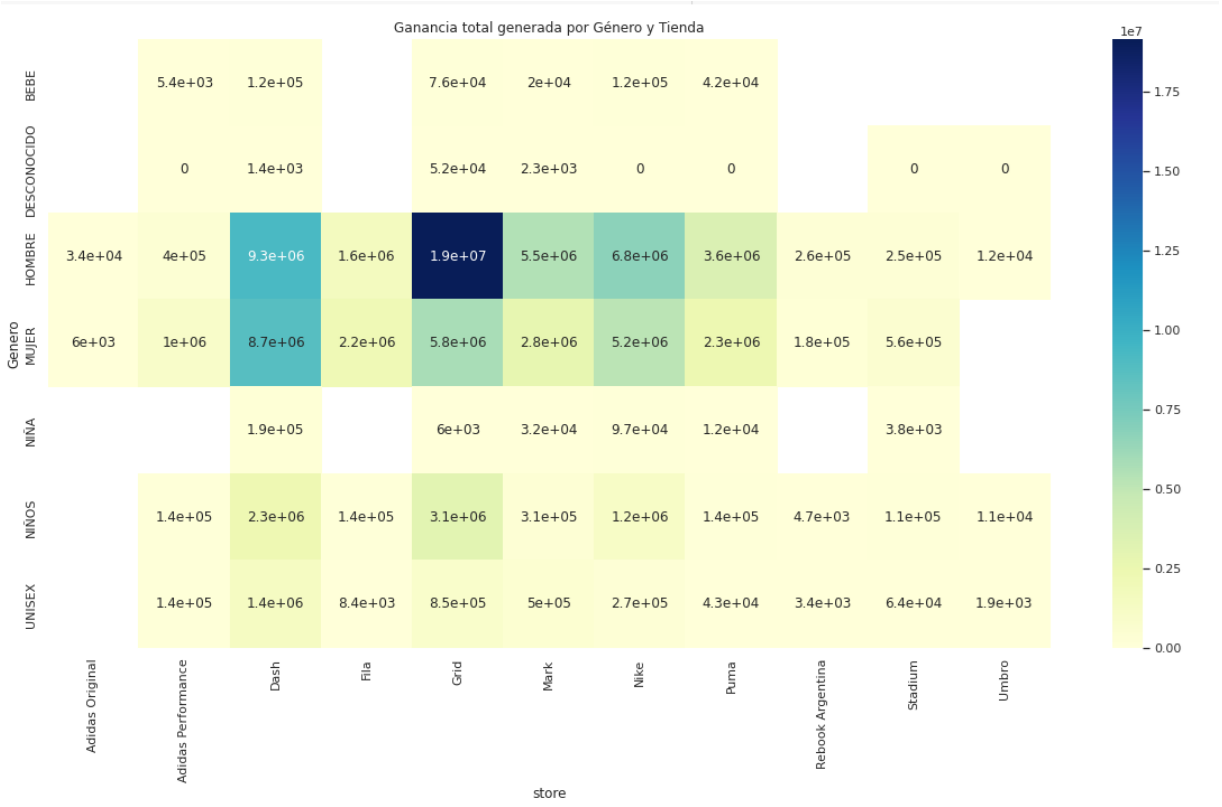
- Ambos e-commerce venden artículos para todos los géneros con una distribución pareja. Para niños y hombres hay valores muy altos o muy bajos a analizar
- En vtex hay un grupo de artículos que no se sabe a qué género corresponden

Distribución de ventas por género y tienda



Casi todas las tiendas venden mayoritariamente productos para hombre y mujer, a excepción de "Umbro" que tiene un 50% de ventas en productos para niños

Ganancia total generada por género y tienda



La mayor ganancia la genera la tienda "Grid" con artículos para hombres, seguida de "Dash" con artículos para hombres y mujeres

Algoritmos Elegidos

A partir del análisis de indicadores, se pensaron los siguientes algoritmos:

1. **K-Means** para segmentar clientes en grupos y encontrar patrones, con el objetivo de:
   - Aplicar una estrategia publicitaria más personalizada
   - Detectar si hay clientes que compran productos para revender

2. **Algoritmos de clasificación** para determinar el mejor e-commerce para vender un producto de acuerdo con:
   - Precio
   - Tienda
   - Línea del producto

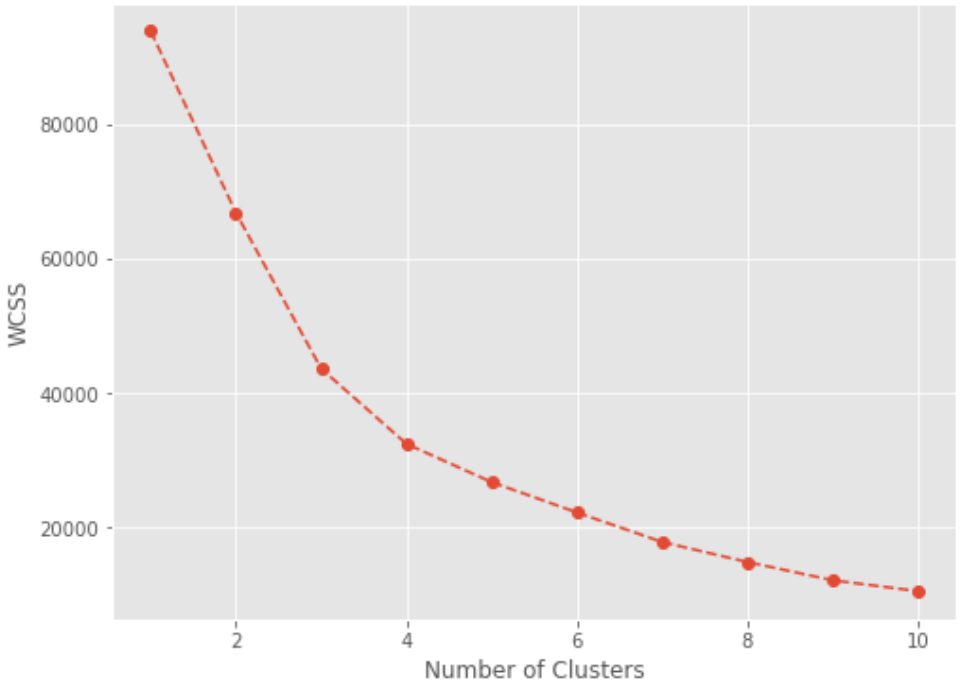
Segmentación de clientes (K-means)

Utilizando el algoritmo **k-means se segmentó a los clientes (identificados por client_id) en grupos** de acuerdo a dos variables:

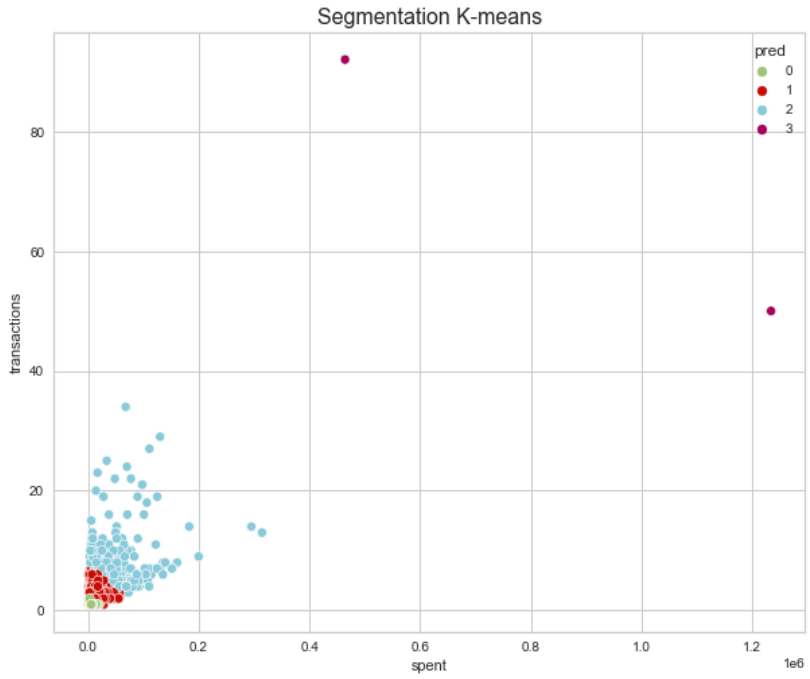
1. Cantidad de dinero que gastaron (suma del precio de los productos que adquirieron)
2. Recurrencia de compra (cantidad de transacciones que realizaron)

client_id	spent	transactions
7527b6831f0f7b7af3cf7c94e44a41ee	464084.2	92
110179867ac351769fb6f665d4d722e0	1233950.0	50
e533ff4a6c8c37c563d1193bf235879d	67641.0	34
cf6a5d84c021d54378f8cd7a31f5c06b	129914.6	29
56aed2cf4b30a7ea7ff39d0fce951ef2	110503.0	27

Mediante el Elbow Method se logró establecer que lo óptimo sería armar **4 clústers (grupos)** de clientes



Luego de aplicar el algoritmo se obtuvo el siguiente resultado:



Cantidad clientes	
Grupo	**Cantidad**
0	40078
1	6580
2	282
3	2

Promedios		
Grupo	**Spent**	**Transactions**
0	7109	1
1	15098	2
2	51487	8
3	849017	71

En conclusión

- El grupo 0 contiene la **mayor cantidad de clientes**, que en promedio solo **compraron 1 vez y gastaron poco**
- El grupo 1 contiene clientes que **compran esporádicamente** o compraron una vez pero **gastaron más que la mayoría**
- El grupo 2 contiene **clientes recurrentes que gastan montos considerables de dinero**
- El grupo 3 contiene **solo 2 clientes** que compraron 71 veces en promedio **gastando montos considerables** de dinero. Es probable que correspondan a **clientes mayoristas,** que compran productos para luego revenderlos

Predicción de e-commerce para ventas (Clasificación)

Se probaron 3 algoritmos de clasificación para predecir **cuál sería el mejor ecommerce (vtex o meli) para publicar un artículo y aumentar las chances de que se venda** en base a la información de precio, tienda y línea del producto en ventas pasadas.

Los algoritmos seleccionados fueron:

- KNN con n_neighbors=3
- Regresión logística
- Random Forest con 200 árboles

Las **métricas de desempeño** de cada modelo fueron las siguientes:

KNN

% de aciertos sobre el set de entrenamiento: 0.87

% de aciertos sobre el set de evaluación: 0.89

Cross Validation KNN: 0.83

Regresión Logística

% de aciertos sobre el set de entrenamiento: 0.84

% de aciertos sobre el set de evaluación: 0.85

Cross Validation Regresion logística: 0.84

Random Forest

% de aciertos sobre el set de entrenamiento: 0.89

% de aciertos sobre el set de evaluación: 0.88

Cross Validation Random Forest: 0.88

Algoritmo	**Accuracy**	**Precision**	**Recall**	**AUC**
KNN	0.89	0.81	0.74	0.84
Regresion Logistica	0.85	0.68	0.80	0.84
Random Forest	0.88	0.75	0.84	0.87

En conclusión

De acuerdo a las métricas el **Random Forest** resulta ser el mejor modelo a elegir para este caso de clasificación

Optimización de modelos

Partiendo del modelo que mejor se adapta a los datos (Random Forest) intentamos **ajustar sus parámetros** para conocer si es posible **mejorar su precisión aplicando GridSearchCV.**
Los resultados fueron:

Algoritmo	Accuracy	Precision	Recall	AUC
Random forest inicial	0.88	0.75	0.84	0.87
Random forest optimizado	0.87	0.80	0.67	0.80

Se puede observar que el modelo optimizado mejoró en precisión 0.5 puntos, **pero bajó considerablemente en recall** y también en las otras métricas, por lo que optamos por **quedarnos con el modelo inicial**.

Conclusiones

Si bien los datasets proporcionados tienen muchos datos para explorar, enfocamos el análisis en generar el mayor valor posible para el negocio, posibilitando:

- Maximizar las ventas al conocer por anticipado en qué e-commerce conviene publicar cada producto según sus características
- Lanzar una estrategia publicitaria diferencial para los clientes en base al grupo al que pertenecen

Futuras líneas

Para complementar el proyecto, se propone:

- Poder evaluar el algoritmo en base a los resultados de las ventas más recientes.
- Ampliar el tamaño de los datasets para cubrir las ventas de todo el año.
- Explorar los datos geográficos de las ventas para predecir volumen de compras por zona geográfica y así lograr mejorar la logística, tanto en tiempos como en costos operativos.
- De acuerdo con la clasificación de los clientes en grupos, profundizar el análisis del grupo 3 que contiene los posibles casos de clientes mayoristas que compran productos para revender, ya que es una operación que no está permitida. Incluso se podría notificarlos y bloquear sus usuarios para que no puedan seguir realizando compras.