

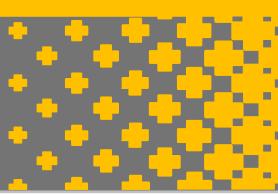


YOUR NEXT GENERATION STORE



# TRANSFORMACIÓN DIGITAL: Implementación de Estrategias de Ciencia de Datos en DMarket

MEMORIA





**Título:**

TRANSFORMACIÓN DIGITAL: Implementación de Estrategias de Ciencia de Datos en  
DSMarket

Capstone Project – Grupo 1 Retail

**Autores:**

- Juan José Guerrero López
- Adrián Ortega Fernández
- Ariana Puentes Lafée
- Cristina Sánchez García

**Tutor:**

Daniel Pegalajar Luque

**Fecha:** 21-marzo-2025

## ÍNDICE

1.	Alcance de proyecto .....	4
2.	Datos de partida.....	5
3.	Metodología utilizada .....	6
3.1.	Preprocesamiento de datos .....	7
	Tabla price.....	7
	Tabla event .....	8
	Tabla sales.....	8
3.2.	Análisis de los datos .....	10
3.2.1.	Análisis de tendencia, variación y estacionalidad .....	10
3.2.1.	Correlación de los datos .....	20
3.3.	Clustering .....	21
3.3.1.	Metodología.....	21
3.3.2.	Creación de clusters.....	23
3.3.3.	Resumen de los resultados .....	27
3.3.4.	Selección de modelos de clúster.....	29
3.3.5.	Representación gráfica del modelo de clúster final.....	31
3.3.6.	Conclusiones .....	33
3.4.	Time Series.....	34
3.4.1.	Análisis temporal.....	34
3.4.2.	Modelos SARIMAX .....	35
3.4.3.	Modelo PROPHET.....	38
3.4.4.	Modelos ML .....	39
3.4.5.	Cálculo de pesos por ID.....	43
3.5.	Abastecimiento de tiendas - MLOps .....	46
3.5.1.	Aplicación de ML para el abastecimiento de tiendas.....	47
3.5.2.	Productivización del modelo.....	50
3.5.3.	Preprocesamiento y enriquecimiento de datos.....	51
3.5.4.	Extensiones en modelo de predicción .....	53
3.5.5.	Diseño de prueba piloto.....	54
4.	Métricas y resultados.....	56
4.1.	Mejor modelo de predicción de ventas .....	56

4.2.	Cuadro de mando .....	57
5.	Conclusiones .....	61
6.	Pasos futuros .....	61
6.1.	Propuestas .....	61
6.2.	Beneficios Esperados: .....	65

## 1. Alcance de proyecto

El proyecto se centra en la transformación digital de una cadena de supermercados, con el objetivo de modernizar procesos y adoptar un enfoque basado en datos.

Los objetivos principales de la transformación digital en DSMarket son los siguientes:

- **Modernización de Procesos:** Reconfigurar y optimizar todos los procesos internos de la empresa para adaptarse a las tecnologías digitales y mejorar la eficiencia operativa.
- **Estandarización de Datos:** Implementar un sistema que permita la estandarización y transformación de las fuentes de datos existentes, garantizando la calidad y coherencia de la información.
- **Migración de Datos a la Nube:** Trasladar las fuentes y procesos de datos a la nube, lo que facilitará el acceso y el análisis de información en tiempo real, mejorando la toma de decisiones.
- **Mejoras en la Precisión de Predicciones:** Desarrollar modelos predictivos avanzados para mejorar la precisión en las proyecciones de ventas y reducir márgenes de error, lo que impacta positivamente en la planificación de inventarios y estrategias de marketing.
- **Optimización del Inventario y Procesos Internos:** Utilizar inteligencia artificial para optimizar procesos críticos como la gestión de stock, la fijación de precios y la logística, minimizando errores y aumentando la eficiencia.
- **Fomento de una Cultura Basada en Datos:** Impulsar una cultura organizacional que valore el uso de datos en la toma de decisiones, fomentando un enfoque más analítico y estratégico en todas las áreas de la empresa.
- **Colaboración Interdisciplinaria:** Trabajará en estrecha colaboración con otros departamentos, como marketing y finanzas, para entender sus necesidades y traducirlas en requisitos de datos y analítica. Esto suavizará la transición hacia un enfoque impulsado por datos en toda la organización.
- **Adaptación a la Transformación del Sector Retail:** Posicionar a DSMarket como un competidor relevante en el sector retail, aprovechando las oportunidades que brinda la digitalización y la analítica avanzada.

Estos objetivos buscan no solo transformar tecnológicamente a DSMarket, sino también mejorar su competitividad y capacidad de adaptación en un entorno de retail en evolución.

## 2. Datos de partida

Los datos proporcionados para este proyecto están formados por:

- Tabla **daily\_calendar\_with\_events** donde se registran los eventos.
- Tabla **item\_prices** de productos/tienda con registro diario de sus ventas.
- Tabla **item\_sales** con el precio medio semanal de cada producto/tienda.

Se han observado las siguientes características de los datos:

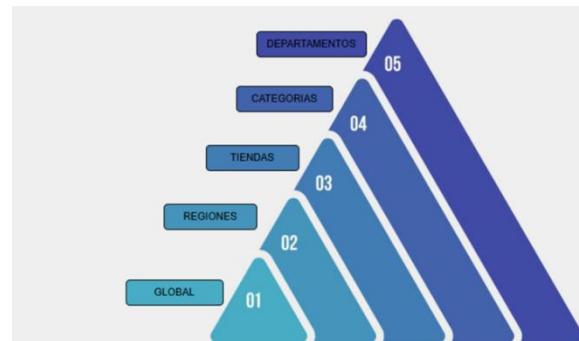
- Se cuenta con **6 años de estudio**, de los cuales **4 de ellos están completos**. El primer año, el mes de enero no se encuentra registrado y su último año solo cuenta con datos hasta abril. El número de días registrados son de 1913 (Fecha mínima: 2011/01/29 - Fecha máxima: 2016/04/24), no se ha encontrado en el dataset fechas con falta de recogida de datos. No se ha encontrado ningún valor duplicado ni error en las fechas. Los nulos existentes son aquellos días que no hay eventos.
- Con respecto los eventos se han registrado 5 tipo ('SuperBowl', 'Ramadan starts', 'Thanksgiving', 'NewYear', 'Easter'), el más usual es el de la Superbowl.
- En la **tabla de prices**, se recogen las semanas del año para registrar los precios medios semanales. Se recogen los datos diferenciados por tienda y producto. La dispersión de los datos en este conjunto es amplia, reflejando una distribución altamente asimétrica. El valor mínimo registrado es 0.1, mientras que el máximo alcanza 134.15, lo que indica una gran variabilidad en los datos.

	mean	std	min	Q0,25	Q0,5	Q0,75	max
sell_price	5,52	4,39	1,20	2,62	4,20	7,18	134,15

Las columnas categóricas se han revisado para comprobar que no hay errores tipográficos que puedan dar una inconsistencia en los datos, asegurando la calidad y estandarización de los datos.

Se observa que los datos recopilados por tienda son similares y no hay un desbalanceo de los datos recogidos. Se ha observado que todos los valores duplicados tienen la columna yearweek nula. También se ha localizado un 0.4% de nulos adicionales en esa columna.

- Se cuenta con un total de **3049 artículos** vendidos en **10 tiendas**, distribuidas en **3 ciudades** (BOSTON - 3 tiendas, NEW YORK - 4 tiendas y PHILADELPHIA - 3 tiendas). Estos artículos están divididos en **3 categorías**, a su vez estas están divididas en **7 departamentos**:



- SUPERMARKET (1473 artículos asociados)
    - SUPERMARKET\_1 - 216 artículos
    - SUPERMARKET\_2 - 398 artículos
    - SUPERMARKET\_3 - 823 artículos
  - HOME&GARDEN (1047 artículos asociados)
    - HOME&GARDEN\_1 - 532 artículos
    - HOME&GARDEN\_2 - 515 artículos
  - ACCESORIES (565 artículos asociados)
    - ACCESORIES\_1 - 416 artículos
    - ACCESORIES\_2 - 149 artículos
- En la **tabla de sales**, se recogen por fila el producto por tienda. Las ventas se representan en columna por día por producto/tienda.
- Al igual que en la tabla de prices, se comprueban las columnas categóricas para revisar errores tipográficos y asegurar la calidad de los datos.
- Se comprueba también que todos los productos estén registrados en todas las tiendas. No existen duplicados, ni valores nulos.

### 3. Metodología utilizada

Para el desarrollo de este proyecto, se utilizó un enfoque metodológico mixto, combinando técnicas cualitativas y cuantitativas para garantizar un análisis completo de los desafíos en las predicciones de ventas y abastecimiento en DSMarket.

Se emplearon herramientas como Python y bibliotecas de ciencia de datos. Con las cuales se llevó a cabo fases de investigación, diseño, implementación y validación. La recopilación de datos se realizó mediante la extracción de información de las bases de datos internas de la empresa, así como el análisis de tendencias históricas de ventas. Los resultados fueron analizados a través de técnicas de modelado estadístico y aprendizaje automático.

Finalmente, la solución propuesta fue validada mediante métricas de evaluación de modelos predictivos, como el error cuadrático medio (RMSE), asegurando su efectividad y aplicabilidad en el contexto de la empresa.

El sistema MLOps de DSMarket optimiza la predicción de ventas en la cadena de supermercados, mejorando la estimación de stock, minimizando pérdidas por

sobreabastecimiento y facilitando la planificación logística. Está diseñado para ser escalable, automatizado y de alto rendimiento, utilizando tecnologías modernas en Google Cloud Platform (GCP).

Los notebooks utilizados se encuentran en un repositorio de github ([DSMarketNPF01](#)) organizado por las tareas propuestas en el proyecto.

- Preprocesamiento y Análisis exploratorio de los datos  
(CP-N-G1\_retail\_T1Analysis)
- Clustering (CP-N-G1\_retail\_T2Cluster)
- Modelos de predicción de ventas (CP-N-G1\_retail\_T3TimeSeries)
- MLOPS app ([dsmarket-web](#))
- Cuadro de Mando Negocio ([PowerBi](#))

### 3.1. Preprocesamiento de datos

El objetivo del preprocesamiento de datos es garantizar que la información utilizada en el análisis y modelado sea de alta calidad, coherente y estructurada adecuadamente. Para ello, se han aplicado diversas técnicas que permiten la limpieza y transformación de los datos, eliminando inconsistencias y mejorando su utilidad.

Entre las tareas fundamentales desarrolladas en este proceso se incluyen la gestión de valores faltantes mediante la imputación o eliminación de registros incompletos, la detección y eliminación de datos duplicados para evitar redundancias, y la conversión de tipos de datos para asegurar su correcta interpretación en los análisis. A su vez se ha aplicado transformaciones y extracciones de características que optimicen la información contenida en los datos.

En conjunto, estas técnicas permiten mejorar la calidad de los datos, garantizando que sean fiables y adecuados para su posterior análisis o modelado predictivo.

Las operaciones realizadas en el procesamiento de datos han sido las siguientes:

#### Tabla price

Se han realizado las siguientes transformaciones:

- > Creación del campo `id`, para posteriormente unirlo con la tabla de sales.
- > Optimización de las columnas, las categóricas 'category', 'store\_code', 'departament' se cambian a tipo *category*.
- > Eliminación de registros duplicados y nulos.

item	category	store_code	yearweek	# sell_price	departament	id
0 ACCESORIES_1_001	ACCESORIES	NYC_1	201328	12.7414	ACCESORIES_1	ACCESORIES_1_001_NYC_1
1 ACCESORIES_1_001	ACCESORIES	NYC_1	201329	12.7414	ACCESORIES_1	ACCESORIES_1_001_NYC_1
2 ACCESORIES_1_001	ACCESORIES	NYC_1	201330	10.9858	ACCESORIES_1	ACCESORIES_1_001_NYC_1
3 ACCESORIES_1_001	ACCESORIES	NYC_1	201331	10.9858	ACCESORIES_1	ACCESORIES_1_001_NYC_1
4 ACCESORIES_1_001	ACCESORIES	NYC_1	201332	10.9858	ACCESORIES_1	ACCESORIES_1_001_NYC_1

## Tabla event

Para poder relacionar el precio medio semanal con las ventas en la tabla de sales, es necesario sacar el yearweek para cada día registrado. El campo yearweek en la tabla de prices no sigue el estándar ISOCALENDAR, sino que inicia la semana los sábados y asigna 00 a aquellas semanas con días del año anterior al cambiar de año, comenzando en 01 para aquellos días que sí que pertenecen al año nuevo.

date	weekday	# weekday_int	d	event	yearweek
0 2011-01-29 00:00:00	Saturday		1 d_1	Missing value	201105
1 2011-01-30 00:00:00	Sunday		2 d_2	Missing value	201105
2 2011-01-31 00:00:00	Monday		3 d_3	Missing value	201105
3 2011-02-01 00:00:00	Tuesday		4 d_4	Missing value	201105
4 2011-02-02 00:00:00	Wednesday		5 d_5	Missing value	201105

Para alinear los datos correctamente, se creó el df\_days con días de referencia en intervalos de 7 días, al que se le asignó una lista ordenada de yearweek excluyendo las semanas 00.

Luego, se realizó un merge con la **tabla de event** y se llenaron valores nulos con la última semana válida mediante ffill(), asegurando una correcta correspondencia de semanas en los análisis posteriores.

## Tabla sales

Se ha pivotado la tabla para tener registro de venta diaria por producto/tienda. Se ha pasado de tener una dimensión de 30.490 filas a 58M de filas.

Se crean las siguientes columnas adicionales:

- > **Event, event\_boolean:** Para añadir los eventos ocurridos en el periodo de ventas. Event\_boolean se ha creado para transformar el evento a un valor que entienda los modelos de predicción.
- > **year, week, quarter, week, day:** Esto permite un análisis más granular de las ventas por diferentes periodos de tiempo (año, semana, mes, trimestre y día).
- > **w:** determina la semana relativa del histórico de datos.
- > **holiday, holiday\_boolean:** Para añadir los festivos ocurridos en el periodo de ventas por región. holiday\_boolean se ha creado para transformar el evento a un valor que entienda los modelos de predicción.

Optimización del dataset según la tipología de columnas.

- > categorical = ['category', 'department', 'store', 'store\_code', 'region', 'weekday','event','boolean']
- > string = ['item', 'yearweek']

```
> integer = ['daily_sales', 'year', 'month', 'day', 'd', 'w', 'event_boolean',
 'holiday_boolean']
```

Adicionalmente, se ha añadido una **tabla holidays** con el registro de los días festivos, tanto del país como locales de las ciudades donde se encuentran las tiendas analizadas.

	date	holiday	region
0	2016-01-01 00:00:00	New Year's Day	New York
1	2016-05-30 00:00:00	Memorial Day	New York
2	2016-07-04 00:00:00	Independence Day	New York
3	2016-09-05 00:00:00	Labor Day	New York
4	2016-11-11 00:00:00	Veterans Day	New York
5	2016-11-24 00:00:00	Thanksgiving	New York
6	2016-12-25 00:00:00	Christmas Day	New York
7	2016-12-26 00:00:00	Christmas Day (observed)	New York
8	2016-01-18 00:00:00	Martin Luther King Jr. Day	New York
9	2016-02-15 00:00:00	Susan B. Anthony Day; Washington's	New York

Finalmente, se ha creado un dataset con los datos post-procesados, en el cual se ha unido los precios medios semanales por producto/tienda de la tabla prices. Rellenando los precios nulos (semanas que no se han vendido los productos) con 0 y comprobando que no se vendieron ningún producto en esas semanas, siendo un 20% del total de datos. En este nuevo dataset, se ha creado el campo `revenue` para saber las ganancias (ventas\*precio).

Este dataset serán los datos de partida para el análisis exploratorio de los datos (EDA).

id	item	category	department	store	store_code	region
0 ACCESORIES_1_001_NYC_1	ACCESORIES_1_001	ACCESORIES	ACCESORIES_1	Greenwich_Village	NYC_1	New York
1 ACCESORIES_1_002_NYC_1	ACCESORIES_1_002	ACCESORIES	ACCESORIES_1	Greenwich_Village	NYC_1	New York
2 ACCESORIES_1_003_NYC_1	ACCESORIES_1_003	ACCESORIES	ACCESORIES_1	Greenwich_Village	NYC_1	New York
3 ACCESORIES_1_004_NYC_1	ACCESORIES_1_004	ACCESORIES	ACCESORIES_1	Greenwich_Village	NYC_1	New York
4 ACCESORIES_1_005_NYC_1	ACCESORIES_1_005	ACCESORIES	ACCESORIES_1	Greenwich_Village	NYC_1	New York
5 ACCESORIES_1_006_NYC_1	ACCESORIES_1_006	ACCESORIES	ACCESORIES_1	Greenwich_Village	NYC_1	New York
6 ACCESORIES_1_007_NYC_1	ACCESORIES_1_007	ACCESORIES	ACCESORIES_1	Greenwich_Village	NYC_1	New York
7 ACCESORIES_1_008_NYC_1	ACCESORIES_1_008	ACCESORIES	ACCESORIES_1	Greenwich_Village	NYC_1	New York
8 ACCESORIES_1_009_NYC_1	ACCESORIES_1_009	ACCESORIES	ACCESORIES_1	Greenwich_Village	NYC_1	New York
9 ACCESORIES_1_010_NYC_1	ACCESORIES_1_010	ACCESORIES	ACCESORIES_1	Greenwich_Village	NYC_1	New York

# d	# daily_sales	date	weekday	# weekday_int	event	yearweek
1	0	2011-01-29 00:00:00	Saturday		1 No	201105
1	0	2011-01-29 00:00:00	Saturday		1 No	201105
1	0	2011-01-29 00:00:00	Saturday		1 No	201105
1	0	2011-01-29 00:00:00	Saturday		1 No	201105
1	0	2011-01-29 00:00:00	Saturday		1 No	201105
1	0	2011-01-29 00:00:00	Saturday		1 No	201105
1	0	2011-01-29 00:00:00	Saturday		1 No	201105
1	12	2011-01-29 00:00:00	Saturday		1 No	201105
1	2	2011-01-29 00:00:00	Saturday		1 No	201105
1	0	2011-01-29 00:00:00	Saturday		1 No	201105

# year	# month	# quarter	# week	# day	# w	# event_boolean	holiday
2011	1	1	1	5	29	1	0 No
2011	1	1	1	5	29	1	0 No
2011	1	1	1	5	29	1	0 No
2011	1	1	1	5	29	1	0 No
2011	1	1	1	5	29	1	0 No
2011	1	1	1	5	29	1	0 No
2011	1	1	1	5	29	1	0 No
2011	1	1	1	5	29	1	0 No
2011	1	1	1	5	29	1	0 No
2011	1	1	1	5	29	1	0 No
2011	1	1	1	5	29	1	0 No

holiday_boolean	sell_price	revenue
0	0.0	0.0
0	0.0	0.0
0	0.0	0.0
0	0.0	0.0
0	0.0	0.0
0	0.0	0.0
0	0.6118	73416
0	2.0748	41496
0	4.2161	0.0

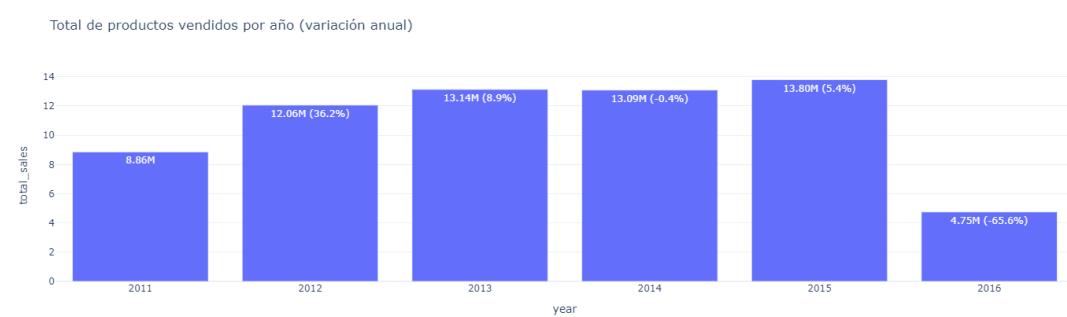
## 3.2. Análisis de los datos

Este apartado es un resumen de lo extraído del EDA que se encuentra recogido al final del documento en el CP-A-G1\_Retail\_Memoria\_Anexo EDA.

### 3.2.1. Análisis de tendencia, variación y estacionalidad

#### TENDENCIAS

La empresa DSMarket presenta un crecimiento sostenido en sus primeros años de operación, con un desempeño destacado en el segundo año. A pesar de que en 2014 experimentó una pequeña caída en el volumen de ventas, logró recuperarse en 2015 con un repunte de 5,4%. Estas cifras revelan un negocio con potencial de estabilidad a largo plazo, aunque sujeto a variaciones que requieren un monitoreo constante de los factores internos y externos que influyen en las ventas.

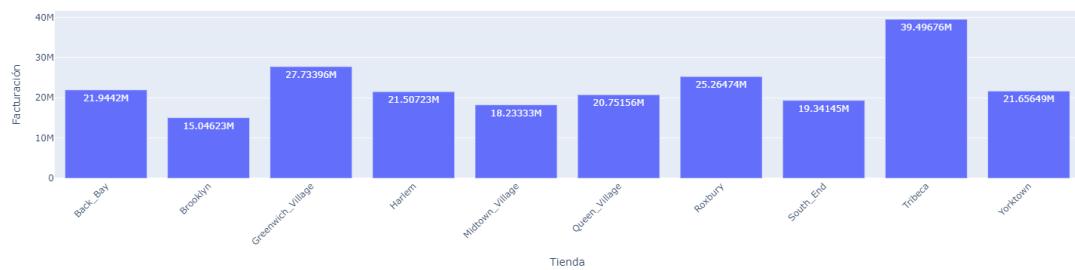


En cuanto al desempeño regional, New York encabeza los niveles de facturación, impulsada principalmente por la tienda Tribeca, que se ha posicionado como el referente de ventas más alto en todo el histórico. Green Village y Roxbury, esta última ubicada en Boston, también muestran resultados notables, evidenciando su capacidad para adaptarse a la demanda de los consumidores.

Facturación por Región

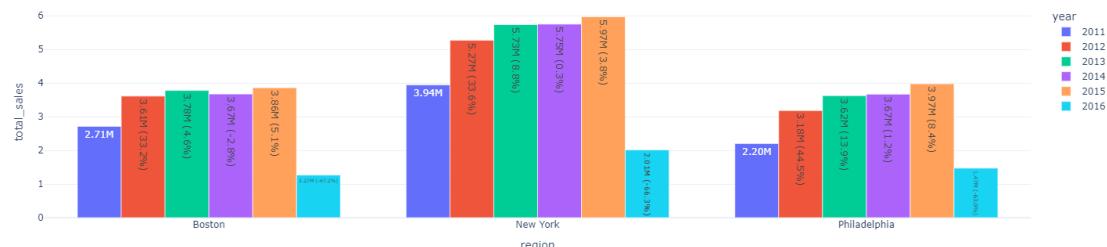


Facturación por Tienda



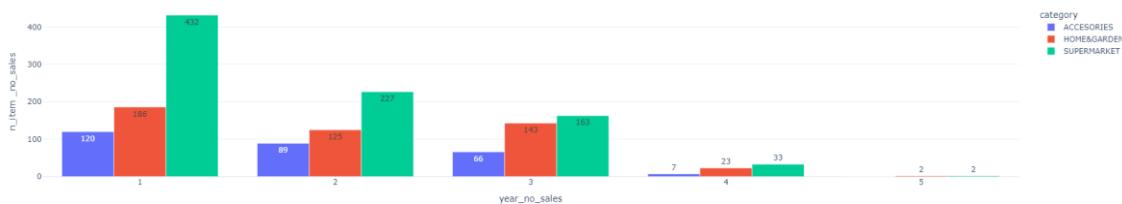
Por otro lado, Philadelphia destaca con un crecimiento más pronunciado, pero con menor volumen de ventas.

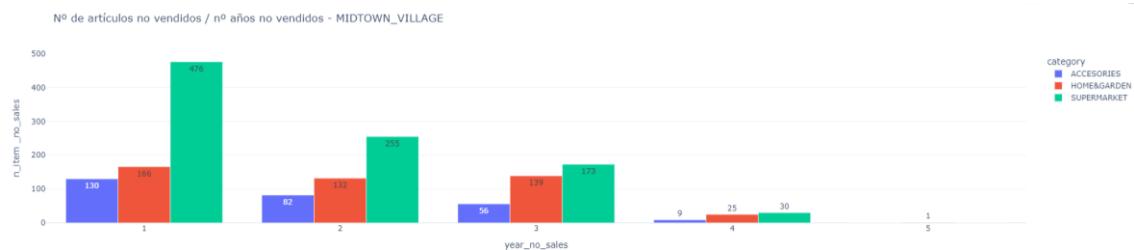
Total de productos vendidos por REGIÓN/AÑO



Es importante destacar que en esta región (Philadelphia) se ha observado un mayor número de productos no vendidos al menos en un año en alguna de sus tiendas(Yorktown y Midtown), lo que sugiere oportunidades de mejora en la gestión de inventarios y en la ejecución de estrategias de marketing específicas.

Nº de artículos no vendidos / nº años no vendidos - YORKTOWN

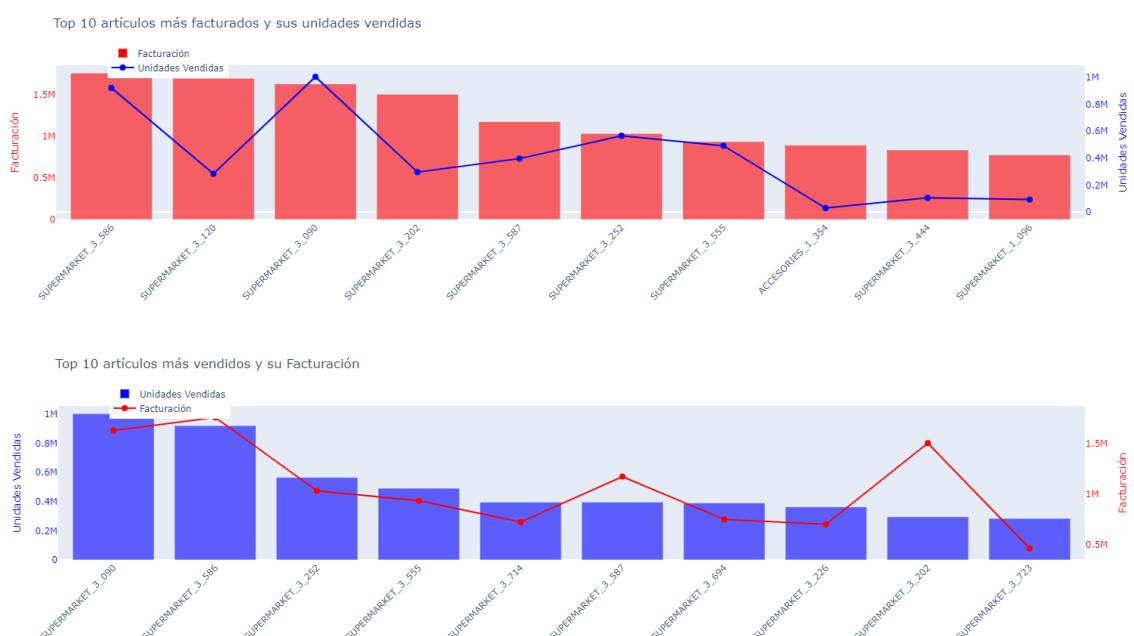




Se ha observado una correlación directa entre el volumen de ventas y la facturación. A mayor volumen de ventas, se incrementa la facturación como se puede observar en el siguiente gráfico, lo que indica que las estrategias implementadas para mejorar el volumen de ventas no solo son efectivas, sino que también son predecibles y consistentes a lo largo del tiempo.



Si bien los productos más económicos suelen registrar una demanda más alta, se identifican artículos de valor elevado que logran una facturación considerable gracias a factores como la marca o la percepción de calidad.



## VARIABILIDAD

Con respecto la variabilidad de la venta de los productos, se observa que la mayoría de los productos tienen una variabilidad baja entre años en las diferentes tiendas, con pocos productos alcanzando niveles medios o altos, siendo similar a los valores detectados a nivel global.

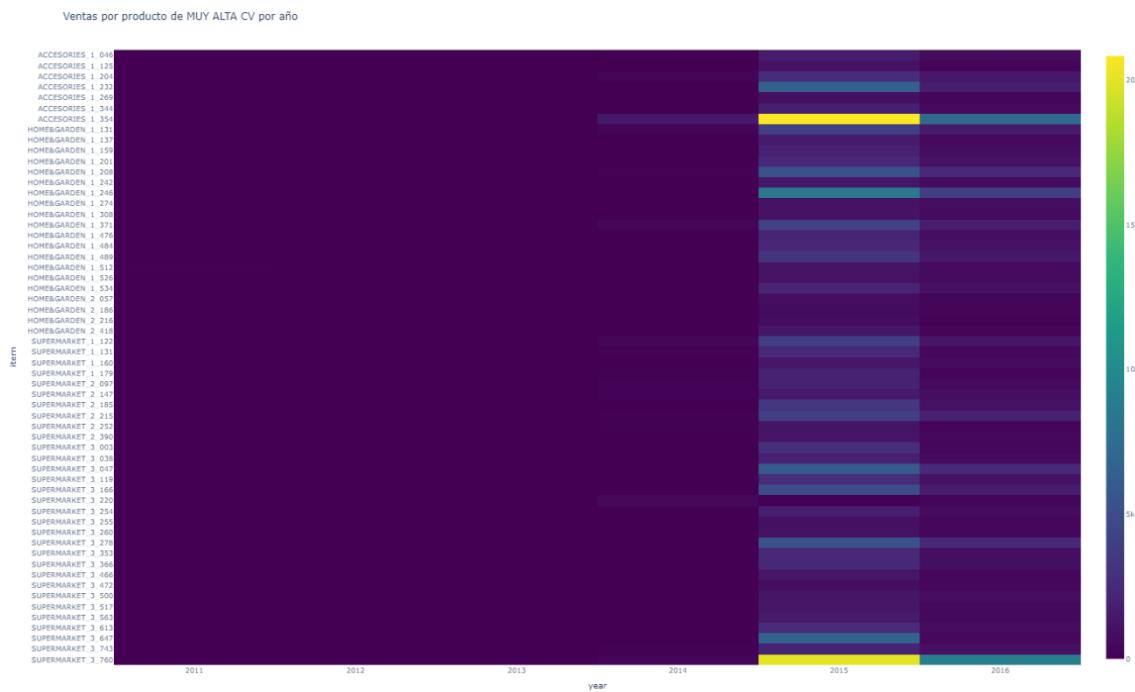
Harlem exhibe fluctuaciones de ventas más marcadas, lo que indica la necesidad de un análisis más detallado para comprender cuáles factores locales o de surtido pueden estar afectando su desempeño. No obstante, las tiendas con mayor estabilidad y altos volúmenes de venta sirven como modelos a seguir para implementar mejores prácticas en aquellas con resultados inferiores.



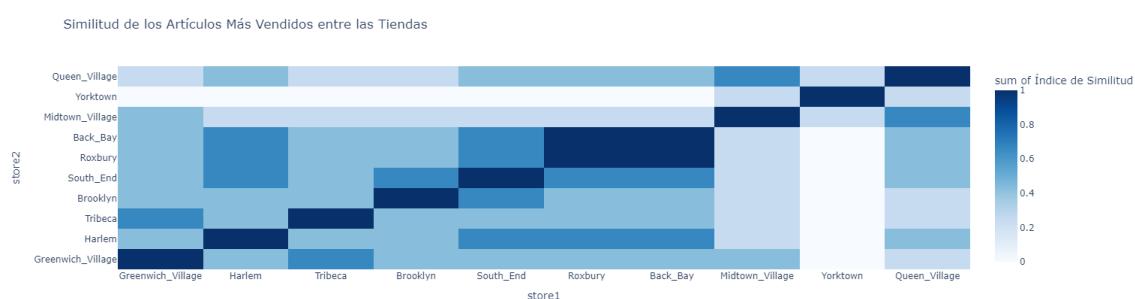
Distribución de CV por Tienda



El análisis de la variabilidad en las ventas por producto a lo largo de los años revela que algunos productos han experimentado un crecimiento explosivo en ventas después de un período de baja o nula demanda. Identificamos estos productos analizando la varianza en sus ventas anuales y detectando aquellos cuyo comportamiento ha sido significativamente inestable.

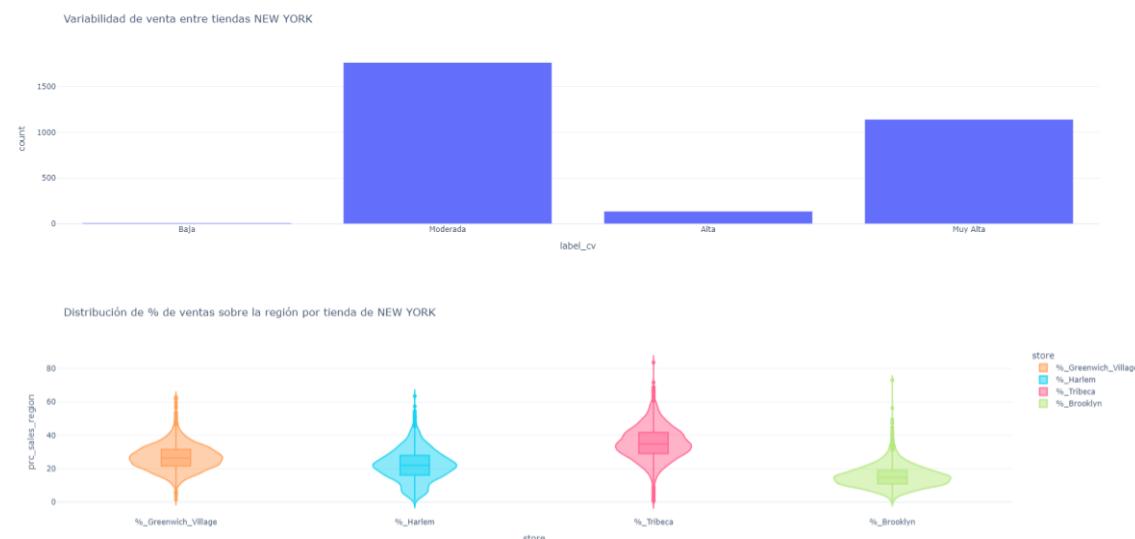


Con respecto los productos estrella, se observa que algunas tiendas, como Back Bay y Hidden Village, presentan un alto grado de similitud en sus productos más vendidos, lo que indica que tienen patrones de consumo similares. En contraste, Yorktown tienen una menor similitud con la mayoría de las demás tiendas, lo que sugiere diferencias significativas en la demanda de productos. También se pueden identificar agrupaciones de tiendas con comportamientos de venta similares, lo que podría ser útil para estrategias de marketing y gestión de inventario.



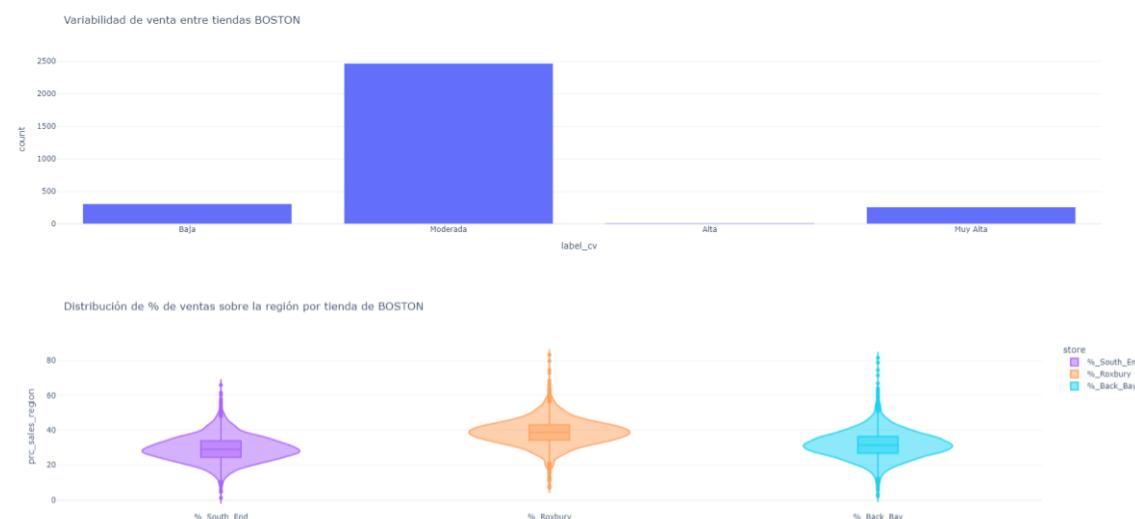
Se ha analizado si existen artículos que se venden más en algunas tiendas y menos en otras, con el fin de entender los factores que influyen en esta variabilidad y proponer estrategias de optimización.

## > NEW YORK



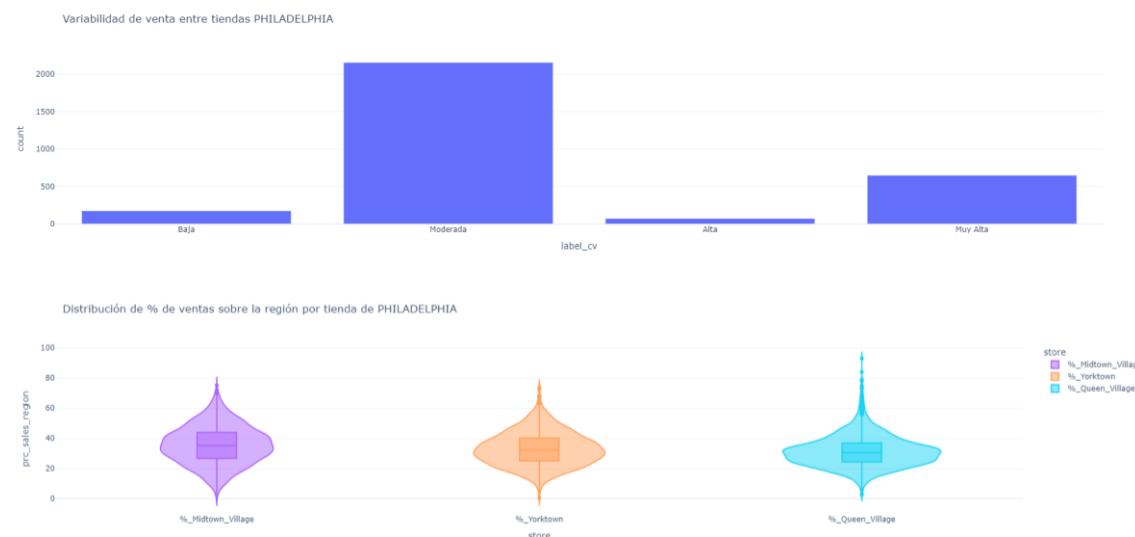
La variabilidad en la región de New York de los productos entre sus tiendas mayoritariamente es MODERADA (1761 productos) y MUY ALTA (1140 productos). Se observa que la tienda con mayor peso en ventas es Tribeca, llegando alcanzar en producto hasta el 83% sobre el total y con una media de 35% en sus productos. En cambio, Brooklyn destaca por ser la tienda con el menor porcentaje de ventas en los productos sobre el total, la gran parte de sus productos no alcanza el 20%.

## > BOSTON



En este caso, el 80% de sus productos tiene una variabilidad MODERADA entre tiendas. En el gráfico de distribuciones se observa una distribución más homogénea de las tiendas, Roxbury es la tienda que sobresale sobre el resto de tiendas, la media de productos alcanza el 38% de ventas sobre el resto.

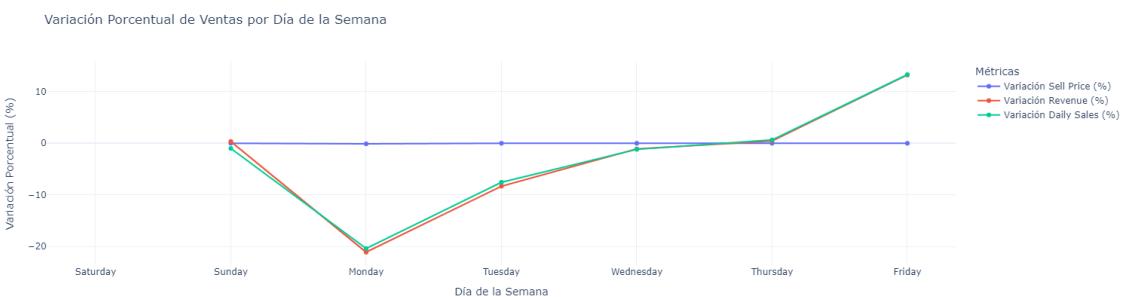
## > PHILADELPHIA



En el caso de Philadelphia el 70% de sus productos tiene una variabilidad MODERADA, pero destaca que la variación MUY ALTA tiene un mayor peso que en el caso anterior, llegando alcanzar el 20% de los productos. La distribución del peso de ventas por tienda es más homogéneo que en el resto de regiones. Queen\_village destaca por tener una concentración entre el 20-40% de ventas sobre el total. Mientras que Midtown\_village tiene una distribución mucho más homogénea.

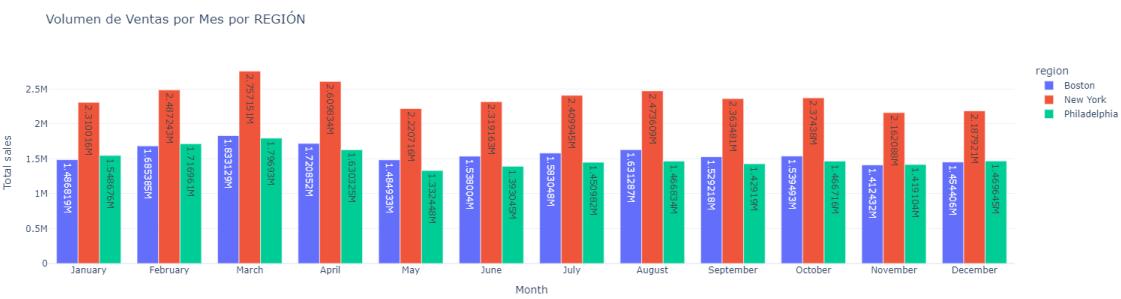
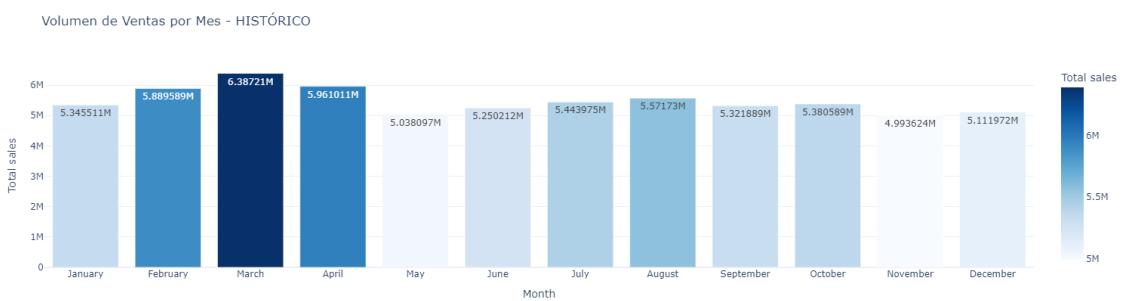
Con respecto la correlación entre las variaciones de precio y el volumen de ventas, se ha observado que pequeñas variaciones en el precio no siempre derivan en cambios significativos en la cantidad vendida, por lo que adoptar estrategias de precios dinámicos —que consideren la elasticidad de la demanda y las características específicas de cada mercado— puede resultar beneficioso para maximizar los ingresos y mantener el posicionamiento deseado.



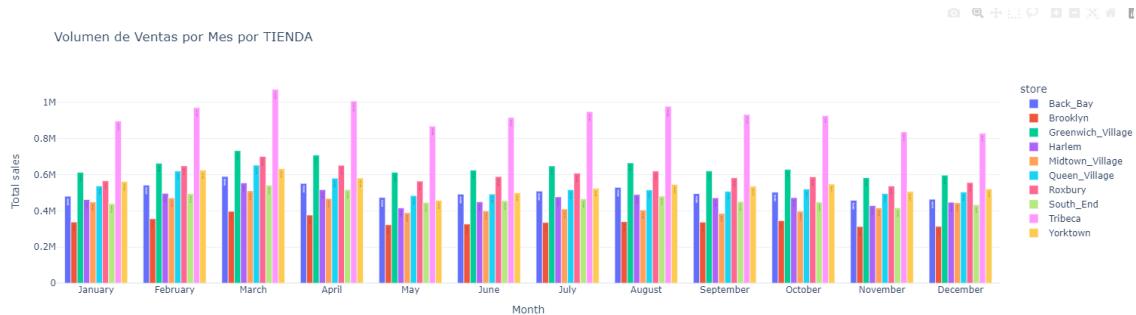


## ESTACIONALIDAD Y TEMPORALIDAD

La estacionalidad en las ventas está presente de forma general en la empresa, con marzo siendo el mes de mayor facturación. Las diferencias en la estacionalidad se ven en el comportamiento por región, donde New York presenta las fluctuaciones más notables, Boston tiene un comportamiento más estable, y Philadelphia está en proceso de normalización.



A nivel de tienda, Tribeca y Greenwich Village se benefician significativamente de las tendencias estacionales, mientras que las tiendas en Boston y Philadelphia experimentan patrones menos marcados. Esto refuerza la idea de que las estrategias deben ser adaptadas según la región y la tienda para maximizar las ventas durante los períodos de alta demanda.

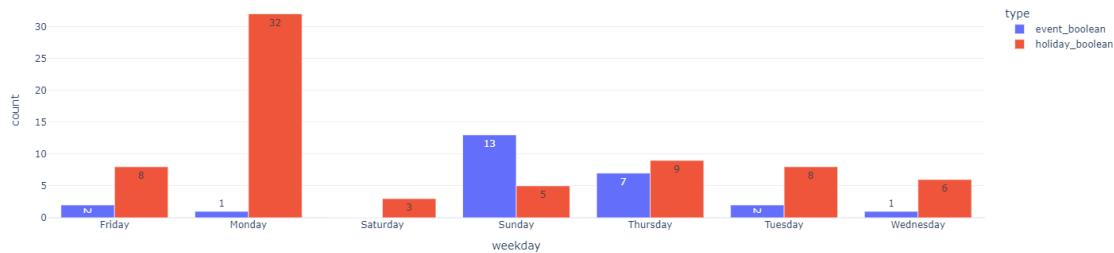


La estacionalidad en las ventas por día de la semana es clara y consistente. Los sábados y domingos son responsables de los picos en las ventas, mientras que los días de la semana (especialmente martes, miércoles y jueves) muestran un volumen significativamente menor. Esta información puede ser crítica para las estrategias de marketing y promociones, sugiriendo que las campañas podrían orientarse hacia el fin de semana para maximizar las ventas y captar la atención de los consumidores que están más disponibles para comprar.

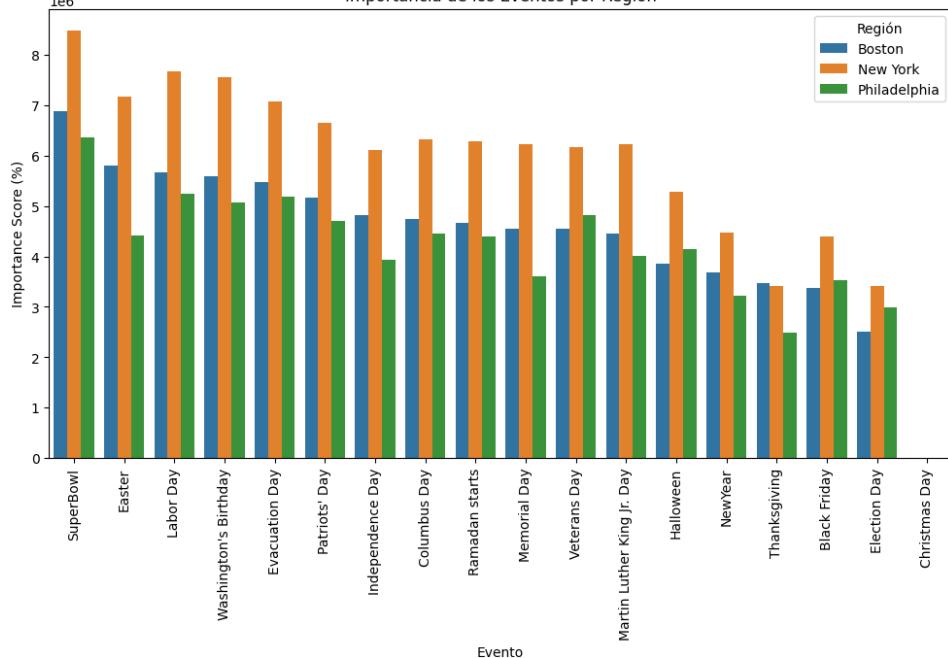


El análisis de la influencia de eventos y la temporalidad aporta otra perspectiva estratégica. Los fines de semana y ciertas fechas festivas (Super Bowl, Independence Day, etc.) generan picos de ventas importantes, con New York siendo la región más susceptible a capitalizar estos incrementos.

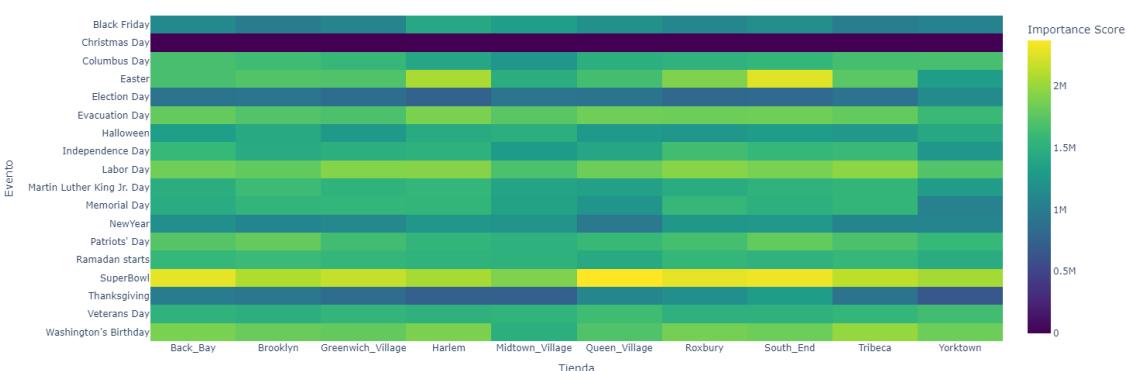
Distribución de los festivos/eventos



Importancia de los Eventos por Región



Promociones y campañas dirigidas en los días previos a dichos eventos, así como un análisis post-evento que ayude a refinar las estrategias en el futuro, pueden optimizar la explotación de la demanda estacional. Philadelphia, por su parte, puede requerir un abordaje diferenciado para elevar su rentabilidad, tanto en la planificación de acciones de marketing como en la adecuación de la oferta de productos al perfil de consumo local.



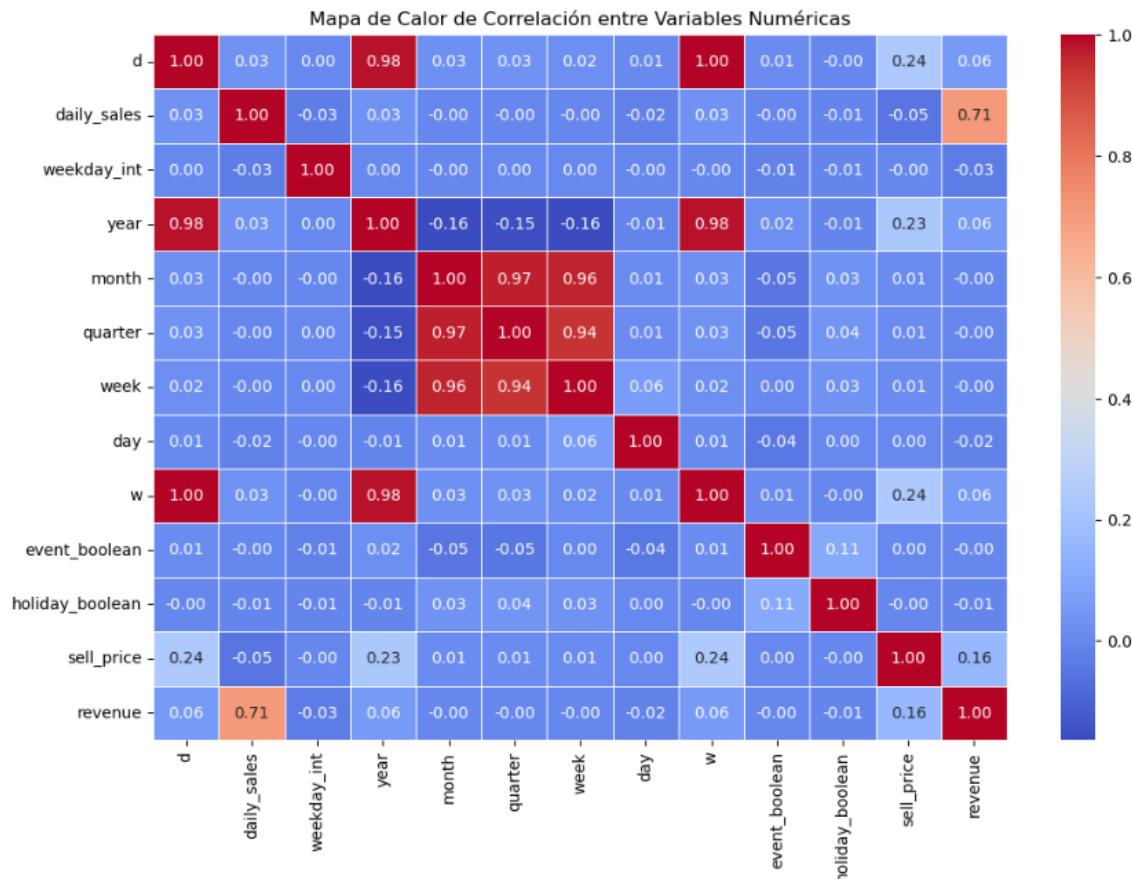
Uno de los hallazgos más relevantes recae en la gestión de inventarios y productos. El 39% de los productos se han mantenido sin vender durante al menos un año. Esto revela una oportunidad de optimización significativa: ajustar de manera más precisa la oferta de productos, priorizar los artículos de alta rotación y diseñar estrategias de comercialización más focalizadas en aquellos con menor salida. Este enfoque ayudaría a reducir costos y maximizar utilidades, dirigiendo la atención hacia las referencias que representan un mayor potencial de ingreso.



En síntesis, la compañía cuenta con una base sólida y un potencial de crecimiento estable, siempre que se preste atención a ciertos puntos de mejora que han surgido del análisis. Asegurar un inventario equilibrado, aplicar precios flexibles basados en el valor real y percibido, y personalizar las acciones comerciales para cada región y cada fecha de alta demanda constituyen pilares esenciales para consolidar el éxito de la empresa en un mercado cada vez más competitivo.

### 3.2.1. Correlación de los datos

En el análisis de datos, la identificación de patrones es crucial para comprender las relaciones entre diferentes variables y para la toma de decisiones informadas. Un aspecto fundamental en este proceso es el estudio de la correlación, que nos permite observar cómo se relacionan entre sí distintas variables. La correlación puede revelar patrones ocultos y ayudar a segmentar productos o identificar tendencias en el comportamiento de los clientes.



### 3.3. Clustering

En este trabajo, se utilizan técnicas de clusterización para agrupar productos y tiendas según su comportamiento en términos de ventas diarias, precios y temporalidad. El objetivo principal es identificar patrones en los datos que sean útiles para segmentar productos o tiendas, lo que facilita la toma de decisiones estratégicas basadas en sus características y desempeño. Además, estos clusters se emplearán en un modelo predictivo, ya que agrupar series homogéneas de datos mejora tanto la precisión como la eficiencia de las predicciones. Así, se establece una conexión directa entre los análisis de clustering y la parte predictiva del proyecto.

#### 3.3.1. Metodología

##### Análisis Inicial con dos Variables utilizando K-Means:

El proceso comenzó con un análisis preliminar utilizando únicamente dos variables del conjunto de datos. Para determinar el número óptimo de clusters, se aplicaron tres técnicas clásicas de evaluación de clustering:

- **Método del Codo:** Esta técnica calcula la inercia (o distorsión) para diferentes valores de k (número de clusters) y observa el punto donde la disminución de la inercia se estabiliza. Este punto se considera el valor óptimo de k.
- **Coeficiente de Silueta:** Mide qué tan bien se asignan los puntos a sus respectivos clusters. Un valor cercano a +1 indica una asignación correcta, mientras que valores cercanos a -1 sugieren que los puntos podrían estar mal asignados.
- **Índice de Calinski-Harabasz:** Evalúa la dispersión dentro de los clusters en comparación con la dispersión entre los clusters. Un valor alto indica una buena separación entre los clusters.

Estas tres técnicas permitieron seleccionar el valor de k que mejor representaba la estructura subyacente de los datos.

---

#### Reducción de Dimensionalidad con PCA y Aplicación de K-Means y HDBSCAN:

---

Con el objetivo de explorar la estructura de los datos en un espacio de mayor dimensión, se aplicó la técnica de Análisis de Componentes Principales (PCA). El PCA permitió reducir la dimensionalidad de los datos, conservando la mayor parte de la variabilidad del conjunto de datos original. Una vez realizada la reducción de dimensionalidad, se aplicaron dos algoritmos de clustering:

- **K-Means:** Se utilizó nuevamente el algoritmo K-Means para realizar el agrupamiento en el espacio reducido, aplicando los mismos métodos de evaluación (Codo, Silueta, Calinski-Harabasz) para determinar el número adecuado de clusters.
- **HDBSCAN:** Este algoritmo se utilizó para detectar clusters de forma jerárquica y también para identificar puntos de ruido, es decir, aquellos que no pertenecen a ningún cluster.

---

#### Medición de Pertenencia al Cluster, Estabilidad y Ruido

---

Para evaluar la calidad de los clusters obtenidos con K-Means, se utilizaron dos métricas principales:

- **Silhouette Score:** Mide la calidad de la asignación de los puntos a sus clusters. Un valor cercano a +1 indica que los puntos están bien asignados, mientras que valores cercanos a -1 indican que los puntos están mal asignados. Es útil para evaluar la **cohesión** y la **separación** de los clusters.
- **Calinski-Harabasz Score:** Compara la dispersión interna dentro de los clusters con la dispersión entre los clusters. Un valor alto indica que los clusters están bien separados.

En el caso de **HDBSCAN**, se realizaron mediciones diferentes:

- **Pertenencia al Cluster:** Se calculó la "fuerza de pertenencia" de cada punto a su cluster asignado, lo que mide cuán seguro está el modelo de que un punto pertenece a un determinado cluster.
- **Estabilidad de los Clusters:** Se evaluó la persistencia de los clusters en diferentes ejecuciones del modelo. Un valor alto indica que los clusters son consistentes y estables.
- **Ruido:** HDBSCAN identifica puntos considerados como "ruido", es decir, aquellos que no pueden ser asignados a ningún cluster. La proporción de ruido fue una métrica clave para evaluar la calidad del agrupamiento.

---

#### Series Temporales usando Dynamic Time Warping (DTW)

---

Para el último modelo, se utilizó un enfoque diferente al resto. El algoritmo **Dynamic Time Warping (DTW)** se aplicó para analizar y comparar series temporales, una técnica que mide la similitud entre dos series temporales que pueden no estar alineadas en el tiempo. A diferencia de otras métricas, el DTW ajusta las series, permitiendo compararlas incluso cuando los eventos ocurren en momentos distintos. En nuestro caso, nos permite identificar patrones similares en las ventas de productos a lo largo del tiempo, a pesar de las variaciones temporales. Esto nos proporciona una segmentación más precisa, incluso si los picos de ventas ocurren en momentos diferentes.

### 3.3.2. Creación de clusters

En este ejercicio, se intentaron crear 9 modelos de clusters, siguiendo la metodología descrita anteriormente siempre que fue posible. Para obtener información más detallada sobre la creación de cada modelo, por favor consulte el Notebook de Clusters Sección 1.

Los modelos creados son los siguientes:

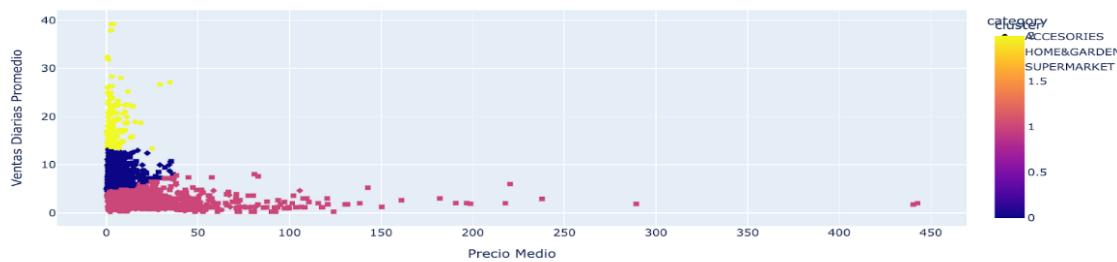
---

#### Cluster 1 por Precio Medio y Unidades Vendidas

---

Este modelo agrupa los productos por su identificador único, calculando la media de las ventas diarias (columna `daily_sales`) y el precio medio de venta (`sell_price`). El objetivo es obtener una visión más clara del comportamiento de cada producto en términos de ventas y precios promedio. Se utilizó únicamente el K-Means con las dos variables descritas:

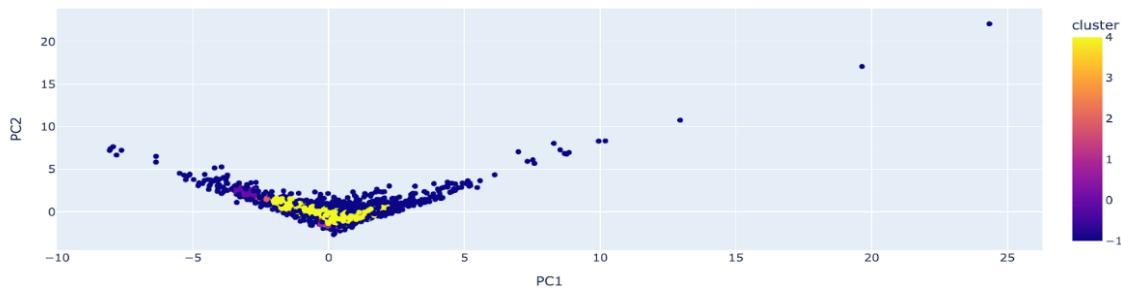
Clustering de Productos basado en Precio Medio y Ventas Promedio



### Clúster 2 por Ventas y Precio Medio según Evento

En este análisis, se comparan los productos en dos escenarios: días con evento y días sin evento. Se calcula el promedio de ventas diarias y el precio medio de venta en ambos escenarios, lo que permite evaluar el impacto de los eventos en las ventas y precios. Se utilizaron todas las metodologías:

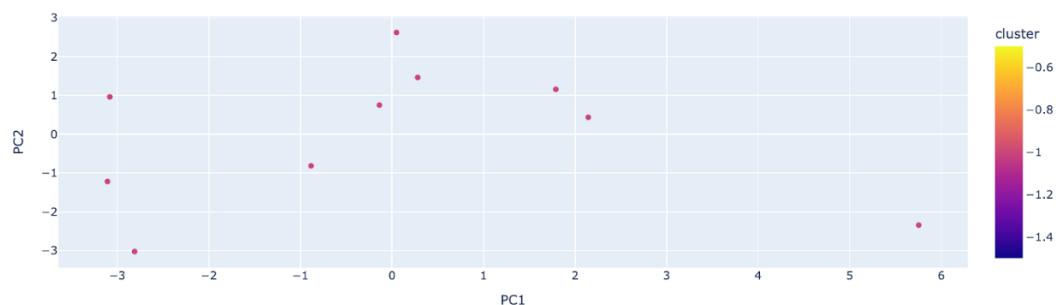
Clustering de Productos con HDBSCAN (PC1 vs PC2)



### Clúster 3 por Tiendas

En este análisis, se agrupan las tiendas y se calculan varias variables clave, como ventas diarias promedio, precio medio de venta, facturación total, número de productos únicos vendidos y la proporción de días con evento. Esta agrupación permite analizar el desempeño de cada tienda y cómo los eventos impactan en las ventas. Se utilizaron todas las metodologías:

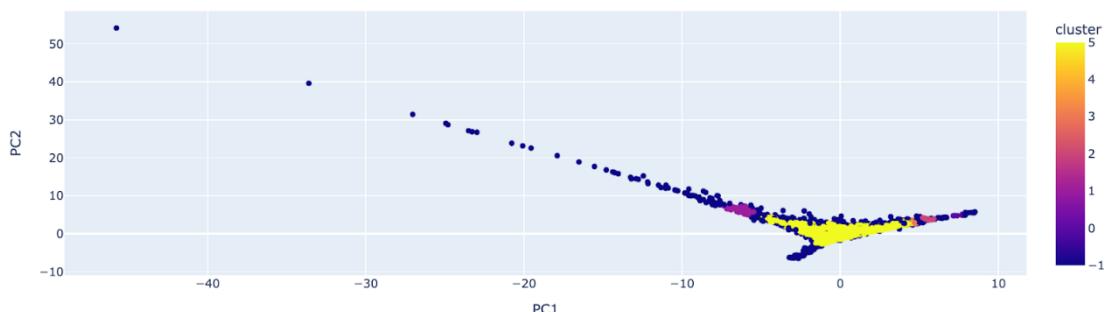
Clustering de Productos con HDBSCAN (PC1 vs PC2)



### Clúster 4 por ID

En este análisis, se evalúan los productos a nivel de ID, que incluye la tienda donde se ha vendido. Se analizan dos escenarios: días con evento y días sin evento, calculando el promedio de ventas diarias y el precio medio de venta en ambos contextos. Se utilizaron todas las metodologías:

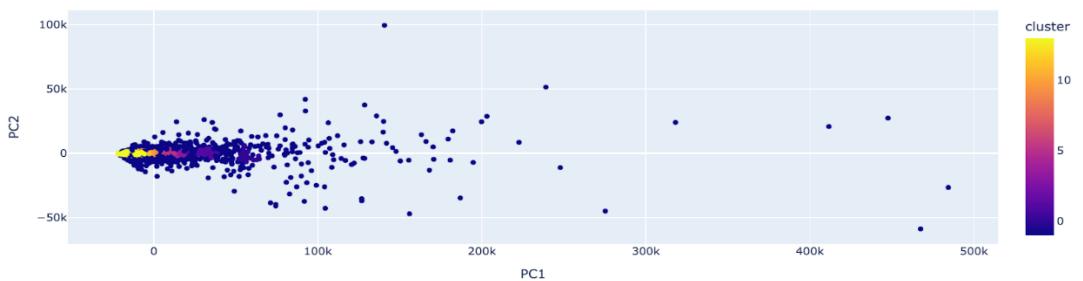
Clustering de Productos con HDBSCAN (PC1 vs PC2)



### Clúster 5 por Ítem y Mes

Este análisis agrupa las ventas de los productos durante los 12 meses del año, y las ventas se estandarizan para asegurar que todos los meses contribuyan de manera equilibrada al clustering. Esto permite identificar patrones estacionales o tendencias dentro del comportamiento de compra. Se utilizaron todas las metodologías:

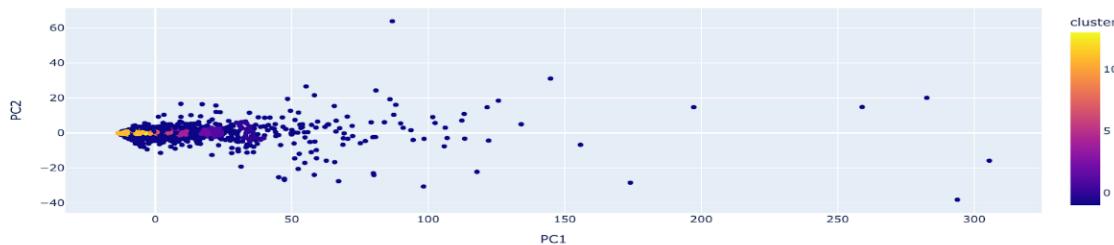
Clustering de Productos con HDBSCAN (PC1 vs PC2)



### Clúster 6 por Ítem y Mes (Media)

Similar al anterior, pero utilizando las ventas promedio diarias de cada producto durante los 12 meses del año. Se utilizaron todas las metodologías:

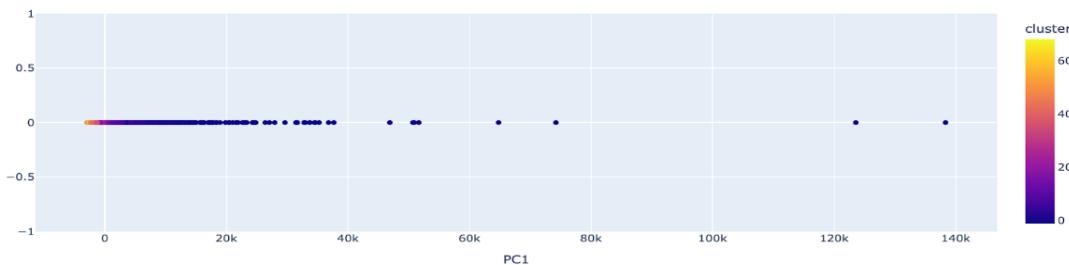
Clustering de Productos con HDBSCAN (PC1 vs PC2)



### Clúster 7 por ID y Días Próximos al Evento

En este modelo, las ventas se agrupan según los días que faltan para el evento. Se crea una tabla pivotada que muestra las ventas diarias por cada producto según los días previos al evento. Se utilizaron metodologías de PCA con K-Means y HSBCA:

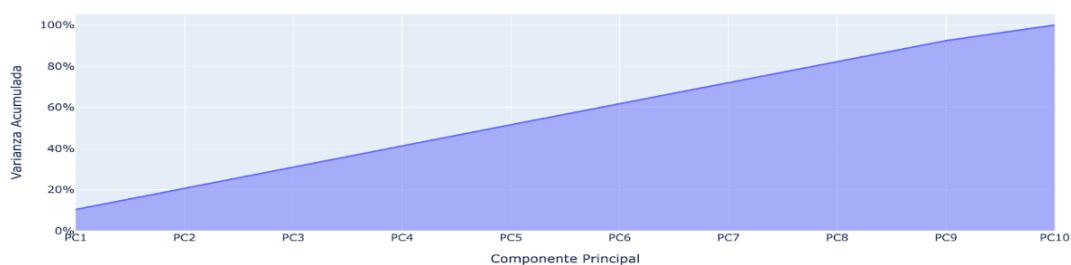
Clustering de Productos con HDBSCAN (PC1 en 1D)



### Cluster 8 por ID y Tienda

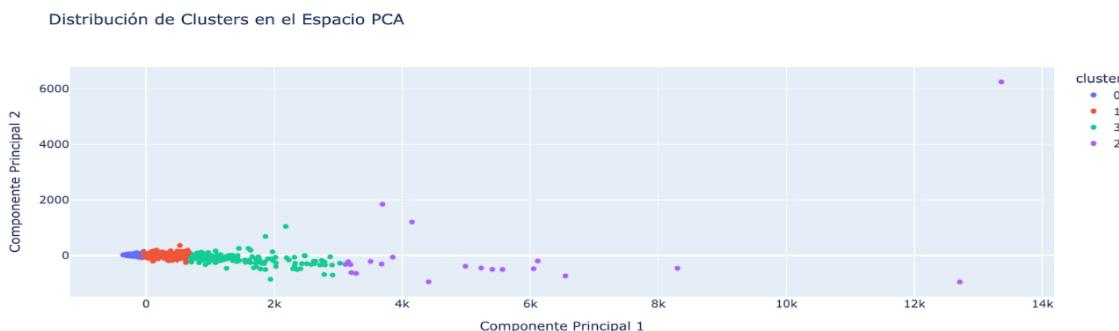
Este análisis agrupa las ventas por producto y tienda, proporcionando una visión detallada de las ventas en cada ubicación. Se intentaron utilizar las metodologías de PCA con K-Means y HSBCA, pero al necesitar tantas componentes para llegar a explicar al menos el 80% de la variabilidad descartamos continuar con esta opción:

Varianza Acumulada



### Cluster 9 por Series Temporales

Se utiliza el algoritmo **Dynamic Time Warping (DTW)** para crear un modelo basado en las series temporales. Este modelo es diferente a los anteriores, ya que se enfoca en las variaciones en el tiempo, comparando las series temporales de ventas de productos:



### 3.3.3. Resumen de los resultados

Las siguientes tablas muestran los resultados obtenidos para los clusters usando K-Means, HDBSCAN y DTW:

Tabla Scores para Clusters con 2 variables usando K-Means

Cluster	Silhouette Score	Calinski-Harabasz Score
1	0.358031	1134.330246
2	0.411744	1627.807307
3	0.274229	7.712628
4	0.355740	15223.509400

Aunque el Silhouette Score y el Calinski-Harabasz Score para el cluster 2 son relativamente buenos, la variabilidad en los resultados de otros clusters (en especial el cluster 3) sugiere que este modelo no ofrece una segmentación sólida para los productos. Sin embargo, al comparar estos resultados con los de HDBSCAN, se observa que los modelos de HDBSCAN tienen una mejor capacidad de adaptación a la complejidad de los datos, ofreciendo resultados más estables y con menos ruido.

---

**Tabla Scores para Clusters con varias variables usando K-Means**

---

Cluster	Silhouette Score	Calinski-Harabasz Score
2	0.407595	1195.429897
3	0.258454	4.240204
4	0.335786	13188.538175
5	0.680830	5563.855477
6	0.680162	5564.023881
7	0.634603	13072.520374

Aquí, el uso de múltiples variables mejora el Silhouette Score y el Calinski-Harabasz Score para ciertos clusters, especialmente el clúster 5 y 6. Sin embargo, la variabilidad sigue siendo alta en comparación con los modelos de HDBSCAN, que tienden a agrupar más efectivamente y a manejar el ruido de manera más eficiente, produciendo resultados más consistentes.

---

**Tabla de Scores para Clusters usando varias variables con HSBCA**

---

Cluster	% de Pertenencia	Número de Clusters	% de Ruido
2	96.795540	3	16.562807
3	0.000000	0	100.000000
4	99.115361	6	3.073139
5	97.169370	14	29.058708
6	97.724732	14	34.732699
7	94.085322	69	6.297147

Los modelos basados en HDBSCAN tienen una notable ventaja en cuanto a la baja proporción de ruido (6.3%), en comparación con los modelos de K-Means que a menudo generan más clusters y mayor ruido, lo que hace que la interpretación sea más difícil. Además, la estabilidad de los clusters en HDBSCAN, especialmente en el modelo con 14 clusters, es un indicio de que estos clusters son más consistentes a lo largo del tiempo, lo cual es crucial para una segmentación robusta.

---

**Tabla de Comparación de Modelo Global con DTW**

---

El uso de DTW es una excelente opción. El modelo que utiliza DTW muestra una reducción en el error absoluto en comparación con el modelo global, lo que indica que la segmentación en clusters a nivel individual mejora la precisión del modelo. Además, DTW es ideal para nuestros datos, ya que muchas veces las series temporales de diferentes productos o tiendas pueden tener patrones de venta similares pero con variaciones en el tiempo.

	mae	abs_error	bias	elapsed_time
model				
Global model	4	441946	-8549	0:01:05
Global model per cluster (DTW)	4	428269	-5026	0:01:45

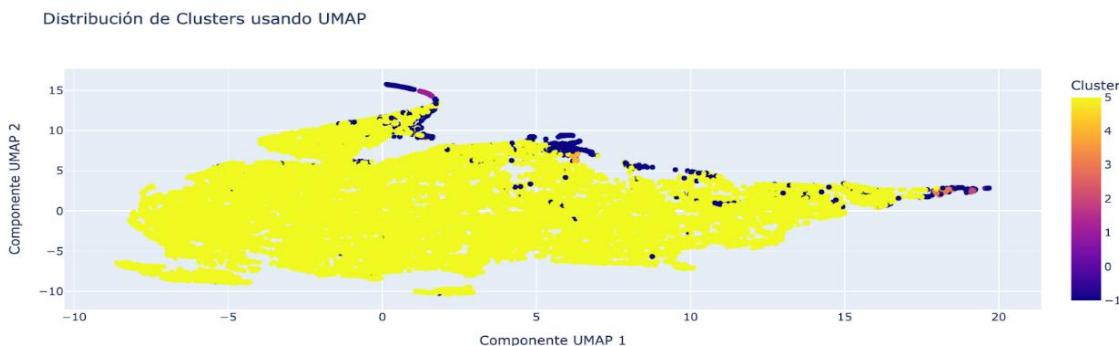
El modelo con DTW presenta una menor cantidad de sesgo negativo (menos diferencia sistemática entre las predicciones y los valores reales) en comparación con el modelo global. Esto significa que, al trabajar con clusters, las predicciones son más equilibradas y cercanas a los valores reales.

### 3.3.4. Selección de modelos de clúster

Finalmente, seleccionamos tres modelos de clustering para utilizarlos en el siguiente paso de análisis de series temporales basándonos en los resultados y los tipos de variables.

#### Modelo 4: Cluster por ID realizado con PCA utilizando el método HSBCA

Este modelo obtuvo los mejores resultados (99% de pertenencia, 6 clusters y 3% de ruido). Además, este incluye los IDs de los productos, proporcionando el nivel más detallado de información:



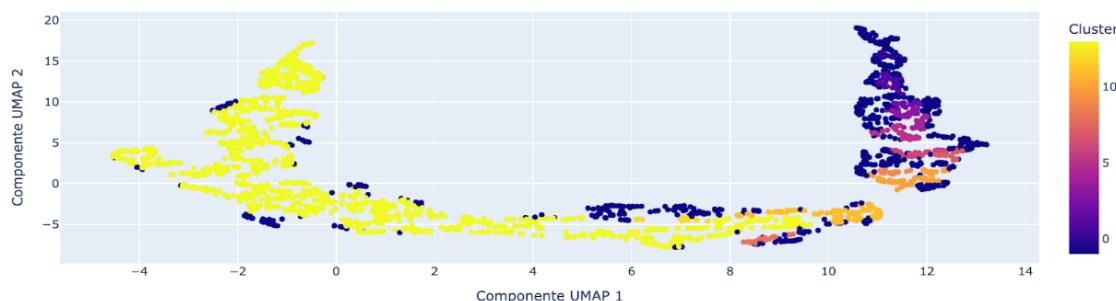
Algunos clusters, como el **cluster 0** (amarillo), tienen una distribución más densa en el gráfico, con muchos puntos agrupados en una región, mientras que otros clusters, como el **cluster 5** (morado), están más dispersos o en las fronteras de la visualización.

Esto podría indicar que los datos del cluster 0 están más "concentrados" en el espacio UMAP, mientras que los del cluster 5 podrían estar menos definidos o abarcar más diversidad.

#### Modelo 5: Cluster por Item y Mes realizado con PCA utilizando el método HSBCA

Es el segundo mejor modelo (97% de pertenencia, 14 clusters y 29% de ruido). Este modelo se diferencia de los anteriores porque las variables son los meses, permitiendo un análisis estacional:

Distribución de Clusters usando UMAP

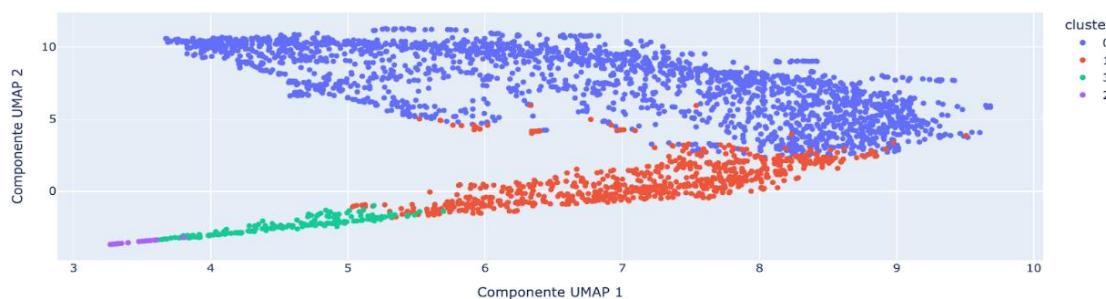


Los clusters están claramente separados en el gráfico, con grupos de puntos organizados a lo largo del plano 2D. Se puede observar que algunos clusters, como el cluster 0, tienen una distribución más dispersa (más áreas amarillas dispersas), mientras que otros, como el cluster 10, están más concentrados en áreas específicas. Algunos de los otros clusters (por ejemplo, cluster 2 y cluster 5, representados en colores rojos y morados) tienen una distribución más extendida, lo que podría sugerir que los datos en esos clusters tienen una mayor variabilidad o dispersión.

#### Modelo 9: Cluster por Series Temporales

Este modelo mostró una mejora significativa en el error absoluto y fue seleccionado por su capacidad para trabajar con series temporales no alineadas:

Distribución de Clusters usando UMAP



La separación visible de los clusters en el gráfico indica que UMAP ha podido reducir eficazmente las dimensiones mientras conserva la estructura de los clusters.

Los productos en el Cluster 0 (color azul) parecen tener características muy diferentes de los productos en el Cluster 1 (rojo), lo que es útil para la segmentación de los productos en función de sus comportamientos en las series temporales.

El Cluster 2 y Cluster 3 parecen estar más cerca el uno del otro en términos de sus características, lo que sugiere que estos productos comparten más similitudes.

### 3.3.5. Representación gráfica del modelo de clúster final

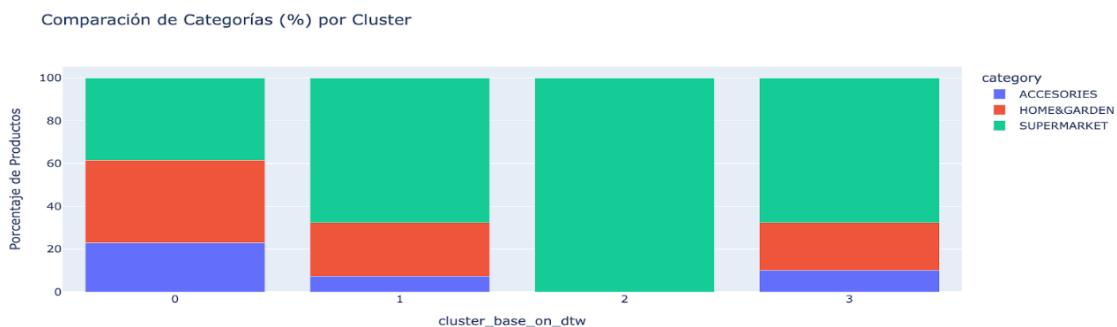
En las siguientes fases de este trabajo, el modelo de **Series Temporales (Modelo 9)** se utilizará para realizar predicciones con Time Series, ya que ha demostrado ser el más efectivo para capturar variaciones a lo largo del tiempo y adaptarse a los patrones de comportamiento de ventas.

*En el notebook se pueden consultar las representaciones gráficas de los otros modelos.*

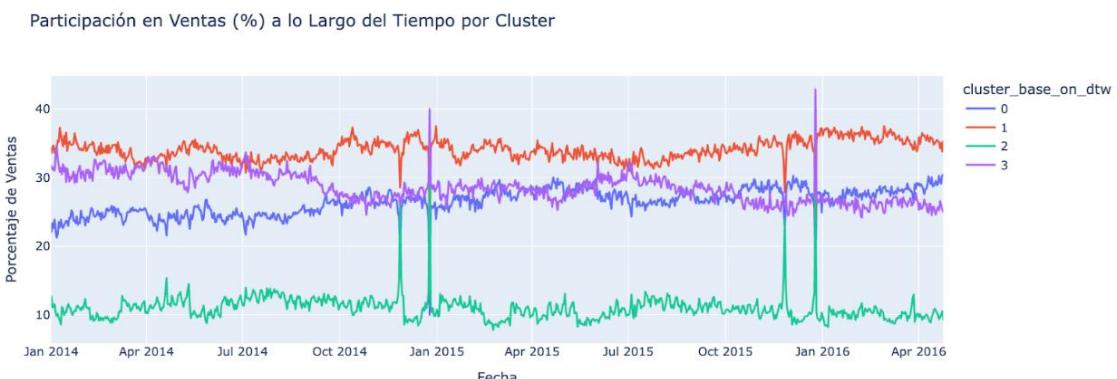
**Número de productos por clúster:** el cluster 0 tiene la mayor cantidad de productos, mientras que los clusters 1, 2 y 3 contienen significativamente menos productos, lo que refleja su diversidad y tamaño.

cluster	number_of_products
0	2168
1	680
2	21
3	180

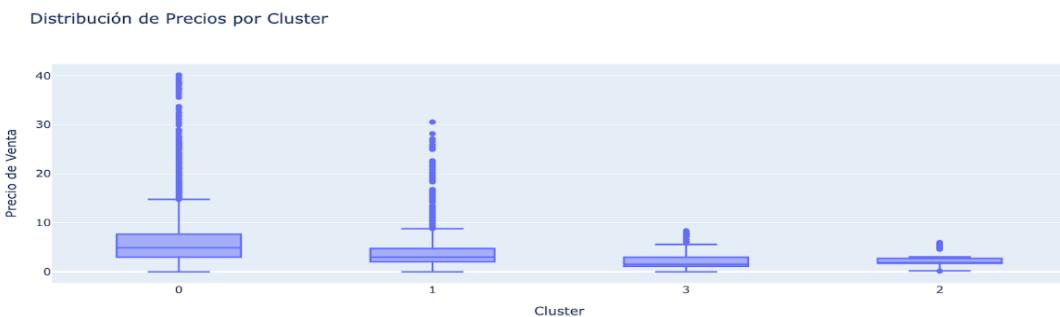
**Comparación de categorías por clúster:** Se observa que "SUPERMARKET" predomina en todos los clusters, especialmente en los clusters 0 y 1. "ACCESORIES" es menos representada, lo que ayuda a identificar las características dominantes de cada clúster. Estos datos están en concordancia con la oferta de productos totales del supermercado.



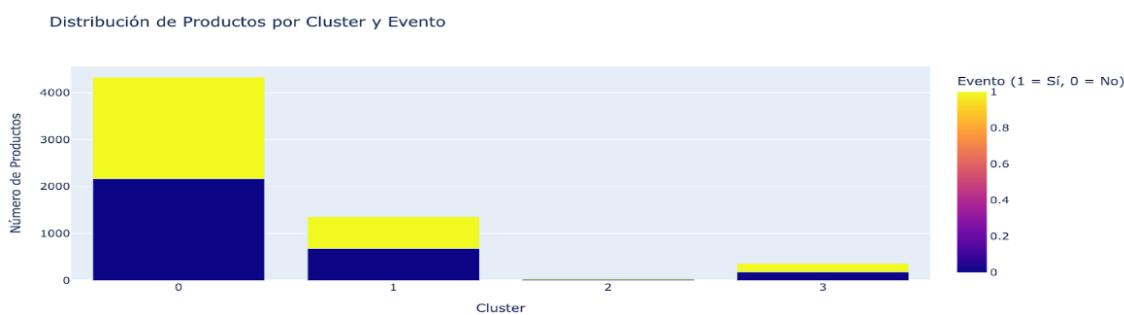
**Participación en ventas a lo largo del tiempo:** Los clusters 0 y 1 muestran picos de ventas similares en el tiempo, mientras que los clusters 2 y 3 tienen una variabilidad menor, lo que sugiere distintos patrones estacionales o de eventos:



**Distribución de precios por cluster:** El cluster 0 presenta la mayor variabilidad en precios, con productos económicos y premium, mientras que los clusters 1, 2 y 3 tienen precios más concentrados, indicando productos de precios más bajos o medios:



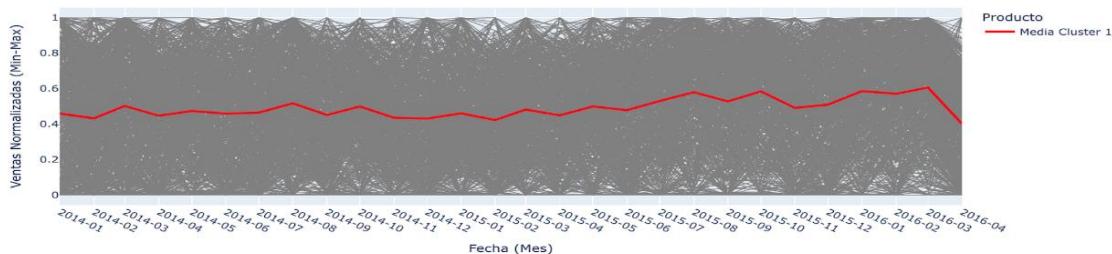
**Distribución de productos por evento:** Los clusters 0 y 1 muestran una mayor proporción de productos vendidos durante eventos, mientras que los clusters 2 y 3 están más concentrados en días sin evento:



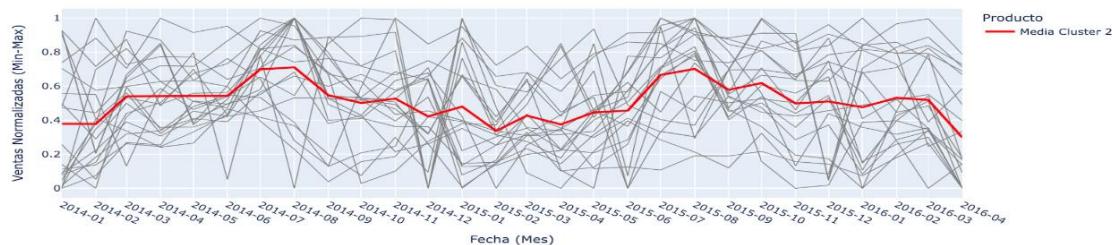
**Evolución de las ventas por clúster:** Este gráfico muestra las ventas normalizadas (min-max) a lo largo del tiempo para los productos dentro de un clúster específico. Cada línea gris representa un producto individual, mientras que la línea roja refleja la media del clúster. Debido a la gran cantidad de productos dentro de cada clúster, las variaciones individuales de las líneas dificultan la identificación clara de patrones o tendencias comunes, ya que las diferencias entre productos pueden ser bastante amplias:



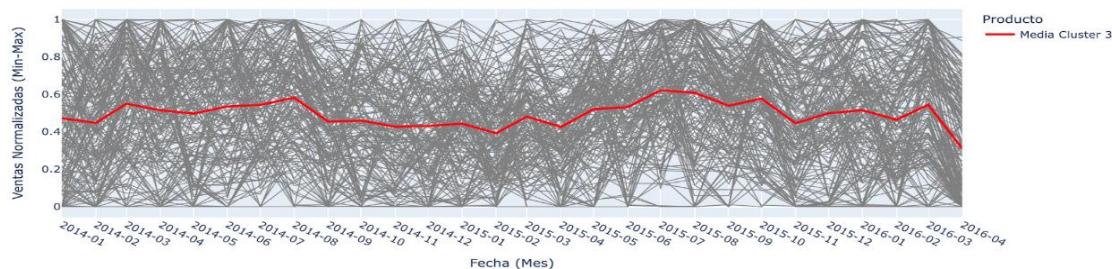
Cluster 1 - Spaghetti Chart de Ventas Normalizadas por Mes



Cluster 2 - Spaghetti Chart de Ventas Normalizadas por Mes



Cluster 3 - Spaghetti Chart de Ventas Normalizadas por Mes



### 3.3.6. Conclusiones

Este trabajo ha demostrado cómo el uso de técnicas de clusterización, combinadas con métodos avanzados como PCA y DTW, puede ofrecer una visión clara y estructurada de los datos de ventas de productos y tiendas. Los distintos modelos de clustering propuestos permitieron segmentar los productos según diversas características, lo que facilitó el análisis y la toma de decisiones estratégicas. A lo largo del proceso, se utilizaron métricas robustas como el Silhouette Score y el Índice de Calinski-Harabasz, que permitieron evaluar la calidad de los clusters generados, y se compararon dos enfoques: K-Means y HDBSCAN, con el objetivo de identificar los mejores métodos de agrupación.

Los modelos basados en HDBSCAN destacaron por su capacidad para manejar el ruido y adaptarse mejor a la complejidad de los datos, mientras que el uso de DTW permitió una segmentación más precisa de las series temporales de ventas, mejorando la precisión de las predicciones. En particular, el modelo 9, basado en series temporales con DTW, ha mostrado ser el más efectivo para capturar las variaciones a lo largo del tiempo y adaptarse a patrones de comportamiento dinámicos, lo que lo convierte en una opción ideal para las predicciones futuras.

Si bien los resultados obtenidos en este análisis son sólidos, siempre existe un margen para la mejora continua. Con más tiempo y recursos computacionales, habríamos optimizado aún más los parámetros de HDBSCAN, ajustando detalles como `min_samples` y `min_cluster_size`, lo que podría haber incrementado la estabilidad y reducido el ruido, permitiendo una segmentación más precisa. Además, habríamos explorado otras técnicas avanzadas de reducción de dimensionalidad, como t-SNE, para obtener representaciones visuales aún más diferenciadas y detalladas de los clusters. Finalmente, la incorporación de variables adicionales, tales como factores climáticos, eventos globales o información más detallada sobre la ubicación de las tiendas, habría enriquecido los resultados y proporcionado una segmentación aún más refinada, contribuyendo a una toma de decisiones más precisa y fundamentada. En definitiva, aunque el trabajo realizado es valioso, el camino hacia una optimización completa siempre está abierto a nuevas posibilidades.

## 3.4. Time Series

En este estudio se ha desarrollado un análisis de series temporales con el objetivo de predecir las ventas. Se han empleado diversas metodologías y modelos de aprendizaje automático para optimizar la predicción de las ventas y entender los patrones temporales presentes en los datos.

La predicción de tendencias es una de las aplicaciones más avanzadas del análisis de datos en supermercados. Mediante el uso de modelos estadísticos y técnicas de Machine Learning, se pueden anticipar cambios en el comportamiento del consumidor, lo que permite adaptar la oferta de productos y las estrategias comerciales con mayor precisión. Identificar tendencias emergentes en el consumo, como el aumento de la demanda de productos orgánicos o sin gluten, brinda una ventaja competitiva al supermercado al permitirle adelantarse a las necesidades del mercado. Además, la capacidad de prever cambios en la demanda ayuda a optimizar el stock y a desarrollar estrategias de marketing más efectivas. Con estos análisis, se pueden detectar oportunidades para lanzar nuevos productos o adaptar las campañas promocionales de manera más efectiva.

Antes del modelado, se ha realizado una exploración inicial con el fin de comprender la estacionalidad, tendencia y variabilidad en las ventas.

### 3.4.1. Análisis temporal.

Gracias a esa exploración hemos confirmado que existían patrones diarios, semanales, y mensuales. Para analizar la relaciones y patrones, hemos realizado una autocorrelación (ACF) y una autocorrelación parcial (PACF) que han servido para analizar la correlación entre los valores pasados de la serie temporal y ayudar en la identificación de patrones estacionales y de tendencia. Han ayudado a definir el número óptimo de lags en modelos de predicción como ARIMA o XGBoost.

A continuación, hemos aplicado la descomposición de la serie temporal que ha permitido desglosar la serie original en sus componentes principales para entender mejor su comportamiento y facilitar el modelado. Posibilitó aplicar transformaciones apropiadas, como diferenciar la serie si la tendencia era fuerte o eliminar la estacionalidad antes de ciertos modelados.

La estacionariedad es una propiedad clave en el análisis de series temporales que implica que las características estadísticas de la serie (como la media, la varianza y la autocorrelación) se mantienen constantes en el tiempo. Hemos evaluado la estacionariedad mediante pruebas estadísticas como las pruebas Dickey-Fuller Aumentada (ADF). Como vemos que no era estacionaria hemos aplicado una diferenciación la cual nos ha servido para mejorar modelos como ARIMA o modelos de machine learning que requieren que sea estacionaria, y de la cual hemos sacado todos los valores para realizar esos modelos.

### 3.4.2. Modelos SARIMAX

#### SARIMAX

El primer modelo que hemos realizado ha sido el modelo SARIMAX que ha sido implementado con el objetivo de capturar patrones de tendencia y estacionalidad en las ventas, además de permitir la inclusión de factores externos (variables exógenas) que pueden influir en las predicciones. El modelo ha sido configurado con los siguientes parámetros:

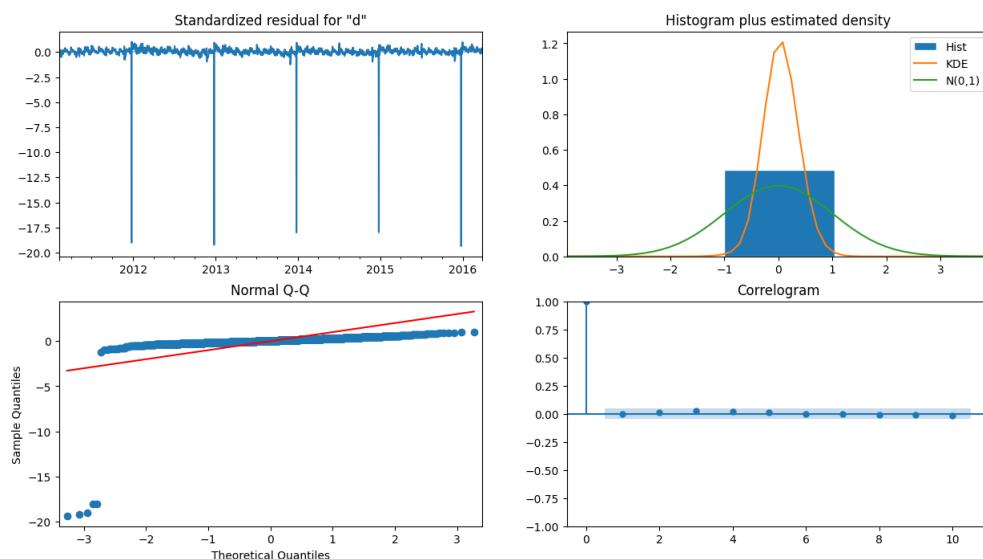
- (p, d, q): Representando la parte autorregresiva, de diferenciación e integración de la serie.
- (P, D, Q, m): Definiendo la estacionalidad del modelo, con un período de estacionalidad determinado según el análisis de la serie.

Esta estacionalidad semanal que hemos reflejado en SARIMAX es debido a que hemos calculado la fuerza de la estacionalidad basada en la varianza. Y vemos que se observa un patrón repetitivo cada 7 días, por eso los lags aparecen a los 7, 14, 21 días. También se detecta una estacionalidad mensual, pero en este modelo no tiene tanta fuerza.

Basándonos en los gráficos de ACF y PACF, hemos podido obtener los parámetros óptimos:

<b>p = 1</b>	La PACF tiene un corte claro en lag 1.
<b>d = 1</b>	Aplicamos una diferenciación para hacer la serie estacionaria.
<b>q = 1</b>	La ACF muestra un decrecimiento en estos lags.
<b>P = 1</b>	Debido a la estacionalidad semanal.
<b>D = 1</b>	Para capturar la estacionalidad diferenciada.
<b>Q = 1</b>	Siguiendo la estructura de la ACF.
<b>S = 7</b>	Estacionalidad semanal.

Tras lanzar el modelo podemos hallar estos resultados con transformación logarítmica:



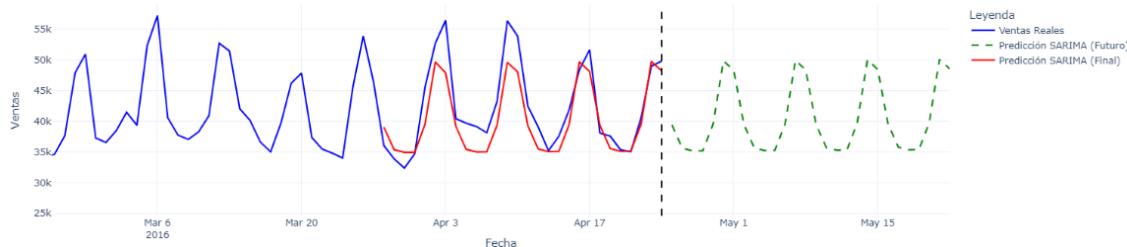
Los residuos presentan algunos valores extremos, lo que sugiere que el modelo podría no estar capturando completamente la variabilidad de los datos. La distribución de los residuos no es perfectamente normal, lo que podría afectar la validez de los intervalos de predicción. No parece haber autocorrelación fuerte en los residuos, lo cual es positivo, pero hay un pico en el primer rezago que podría revisarse.

Sarimax es muy útil para capturar tendencias y patrones estacionales en las ventas de los productos, incorporar variables exógenas como eventos especiales o festivos y mejorar la precisión de las predicciones en comparación con modelos más simples.

A continuación, se muestra las métricas obtenidas y predicciones del test y de 28 días a futuro:

Métrica	Valor
0 MAE	2,773.9337
1 RMSE	3,463.2741
2 MAPE	6.2707

Predicción de Ventas para los Próximos 28 Días con SARIMA

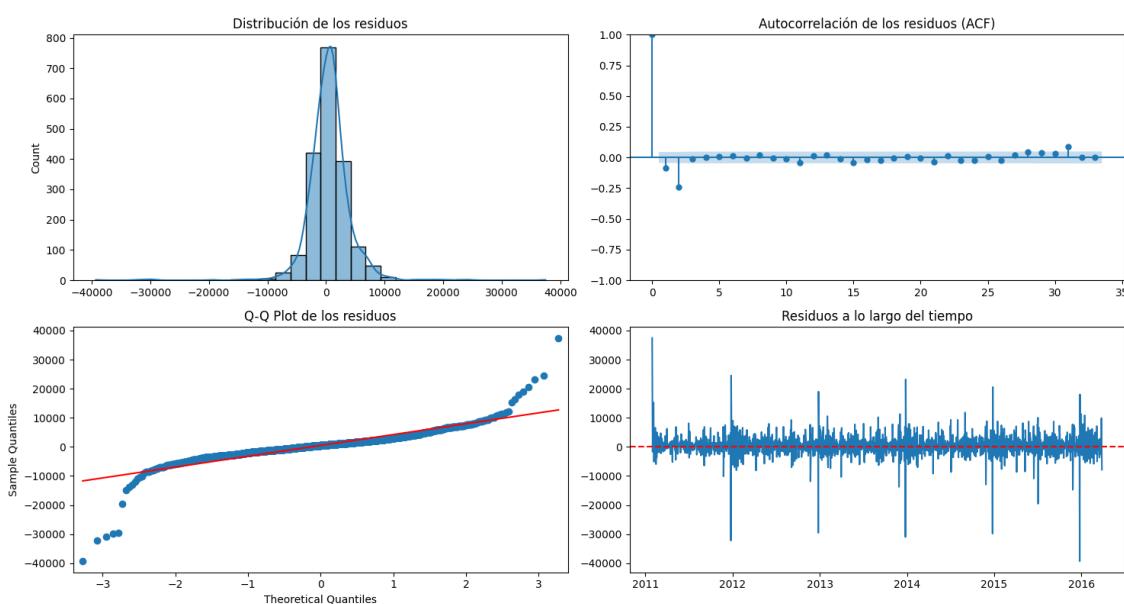


Un MAPE del 6.27% significa que, en promedio, las predicciones se desvían en un 6.27% respecto a las ventas reales. Se observan errores algo grandes (diferencia entre RMSE y MAE), lo que indica la presencia de algunos valores atípicos. La predicción sigue la estacionalidad correctamente, por lo que se espera que los patrones futuros sean confiables.

Aunque SARIMA es un modelo potente, tiene algunas limitaciones que podrían afectar la precisión de la predicción. Solo usa datos pasados de la misma serie, lo que limita su capacidad para captar factores externos, supone que la estacionalidad es constante en el tiempo, pero en la realidad, los patrones estacionales pueden cambiar.

También probamos el Auto Arima, pero nos da peores resultados que nuestro propio SARIMAX por lo que lo descartamos al instante:

Métrica	Valor
0 MAE	8,582.8394
1 RMSE	9,400.4989
2 MAPE	19.9633



Hay margen de mejora en la reducción de errores grandes (RMSE alto), lo que sugiere la necesidad de revisar outliers o incluir más factores en el modelo. Aunque este modelo tiene unas métricas buenas preferimos seguir valorando otros modelos a continuación.

### 3.4.3. Modelo PROPHET

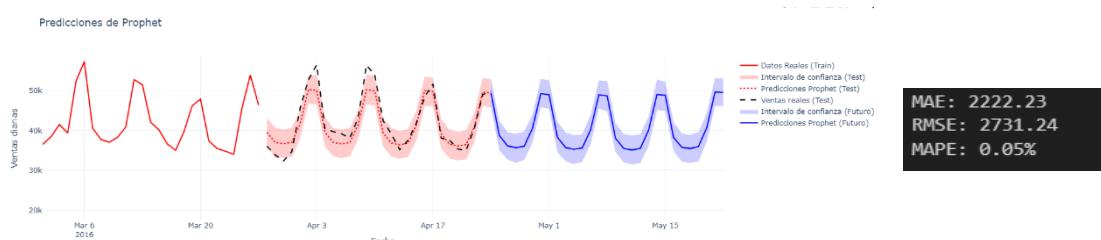
#### PROPHET

El siguiente modelo que usamos fue Prophet, ha sido utilizado como una alternativa para modelar la serie temporal de ventas debido a su capacidad de capturar tendencias, estacionalidad y eventos especiales mediante una regresión multiplicativa.

Ha demostrado ser una herramienta útil y efectiva para la predicción de ventas, especialmente por su facilidad de uso y su capacidad para modelar estacionalidades y tendencias. Sin embargo, en situaciones donde la serie temporal presenta cambios estructurales o múltiples variables influyentes, los modelos de machine learning se ajustan mejor.

Su configuración es propia de este modelo, puesto que se tiene que tratar haciendo dos columnas: 'ds' e 'y'. Donde una es las fechas y la otra el target objetivo que en nuestro caso han sido las ventas diarias. Hemos dividido el dataset en train y test, como la predicción a futuro que buscamos es de 28 días, el test lo hemos establecido en los últimos 28 días.

Le hemos introducido las estacionalidades adecuadas, y hemos añadido tanto las vacaciones precisas en Estados Unidos, como los eventos que teníamos en el dataset:



El modelo Prophet sigue bien la estacionalidad, las predicciones en la fase de test están bastante alineadas con los datos reales, con intervalos de confianza relativamente estrechos, lo que indica confianza en la predicción. Las predicciones futuras mantienen el mismo patrón cíclico, lo cual es esperable en un modelo de series temporales.

Gracias a Prophet pudimos añadir la influencia de eventos especiales y vacaciones para que capte patrones el modelo que antes no podía. La predicción (línea azul) sigue un patrón periódico similar a los datos de entrenamiento.

Si las ventas tienen grandes fluctuaciones diarias, Prophet podría estar subestimando los valores reales en ciertos momentos. Suaviza demasiado los datos, lo que puede hacer que no detecte picos o caídas bruscas en las ventas.

Prophet ha mostrado un desempeño sólido en tu serie de ventas, con una buena precisión y baja tasa de error, aunque con algunas limitaciones en la captura de picos o cambios bruscos, aunque el MAPE es bastante bueno, vemos que la gráfica tiende hacer una línea muy repetitiva, por lo que no parece real, por lo que probamos a seguir modelos, y a continuación pasamos a modelos de machine learning, como XGBoost, CatBoost o LightGBM.

### 3.4.4. Modelos ML

#### MACHINE LEARNING

Para los modelos de machine learning, usamos una función donde ajustamos el dataframe para los tres modelos por igual. Por cuestión de tiempo y recursos, se hacía imposible realizar un modelo por cada id, ya que son 30000 id, por lo que es inviable lanzar tantos modelos. Por ello optamos por un enfoque “Middle-Out”, la predicción en un nivel intermedio y desglosada o agregada según sea necesario. Realizamos varias agregaciones, las cuales están en la gráfica de comparación de modelos.

Después de esta agregación, agrupamos por fecha ya que es requerido para una serie temporal, y añadimos las variables que aportan información a nuestros modelos. Para cada data frame de cada agregación vamos a dividir en dos, “Train” y “Test”, con el cual entrenaremos y después usaremos para calcular el error de nuestras predicciones diarias de ventas.

Se seleccionan los modelos previamente dichos:

- **XGBoost:** Algoritmo basado en árboles de decisión con boosting, optimizado para rendimiento y precisión.
- **LightGBM:** Variante eficiente de boosting desarrollada por Microsoft, enfocada en grandes volúmenes de datos.
- **CatBoost:** Algoritmo de boosting especializado en manejo de datos categóricos, desarrollado por Yandex.

Cada modelo se inserta dentro de un pipeline de preprocesamiento y predicción que incluye:

1. **Deseasonalizer:** Elimina estacionalidad en la serie temporal mediante un modelo multiplicativo.
2. **LogTransformer:** Aplica una transformación logarítmica para estabilizar la varianza.

3. **Detrender:** Usa un modelo de regresión polinómica para eliminar tendencias lineales.
4. **Forecasting Reducer:** Implementa la estrategia de predicción recursiva con una ventana de 28 días.

En este estudio, se ha implementado un proceso de validación cruzada especializado para series temporales junto con la optimización de hiper parámetros de modelos basados en Gradient Boosting. Para ello, se ha utilizado una estrategia de validación cruzada con ventanas deslizantes (SlidingWindowSplitter) y un proceso de búsqueda de hiper parámetros mediante ForecastingRandomizedSearchCV.

La validación cruzada tradicional no es adecuada para series temporales, ya que los datos presentan dependencia temporal y no pueden ser reorganizados aleatoriamente sin comprometer la estructura de la información. Para abordar esta limitación, se emplea SlidingWindowSplitter, que permite evaluar el modelo de manera progresiva manteniendo la coherencia temporal.

- **window\_length=28\*6:** Se define un periodo de entrenamiento de 6 meses (28 días por mes).
- **fh=fh:** Se establece un horizonte de predicción de 28 días.
- **step\_length=validation\_size:** Determina el desplazamiento de la ventana en cada iteración, simulando predicciones mensuales sobre nuevos datos.

Este método garantiza que cada iteración del modelo se entrena en un conjunto de datos históricos y se valide en el siguiente período cronológico, replicando condiciones reales de predicción.

Para optimizar el rendimiento del modelo, se emplea una búsqueda aleatoria de hiper parámetros (RandomizedSearchCV). Se ajustan diferentes configuraciones de parámetros según el modelo seleccionado: XGBoost, LightGBM o CatBoost.

Cada modelo presenta hiper parámetros específicos:

- **max\_depth:** Controla la profundidad máxima de los árboles de decisión: [3, 4, 5]
- **learning\_rate:** Ajusta la velocidad de aprendizaje del modelo: [0.05, 0.1]
- **n\_estimators / iterations:** Define el número de árboles que se entran: [100, 200] / [25, 50, 100]
- **subsample:** Regula la cantidad de datos utilizados en cada iteración: [0.7, 0.8, 0.9]
- **colsample\_bytree / colsample\_bylevel:** Controlan el número de características utilizadas en cada división del árbol: [0.6, 0.8] / [0.7, 0.8]

Para encontrar la mejor combinación de hiper parámetros, se aplica ForecastingRandomizedSearchCV, una variante optimizada para series temporales.

- **n\_iter=10:** Se establecen 10 iteraciones de búsqueda aleatoria.
- **random\_state=42:** Se fija una semilla para reproducibilidad.

- **error\_score='raise'**: Se evita que errores silenciosos afecten la evaluación del modelo.
- **n\_jobs=-1**: Se permite la ejecución en paralelo para acelerar el proceso.

Este enfoque combina validación cruzada con ventanas deslizantes y búsqueda aleatoria de hiperparámetros para mejorar el desempeño de modelos de predicción de series temporales. La combinación de estos métodos permite evaluar distintas configuraciones de modelos de Gradient Boosting de manera eficiente y en condiciones realistas de predicción, asegurando una mejor capacidad de generalización sobre nuevos datos.

Una vez entrenados, se generan dos conjuntos de predicciones:

1. **Predicciones de Validación**: Se comparan con los datos reales para evaluar el rendimiento.
2. **Predicciones a Futuro**: Se extienden 28 días más allá de los datos de test. Para esto es necesario crear un data frame que contenga las mismas variables que X\_test, por lo que lo creamos con datos que tenemos previamente.

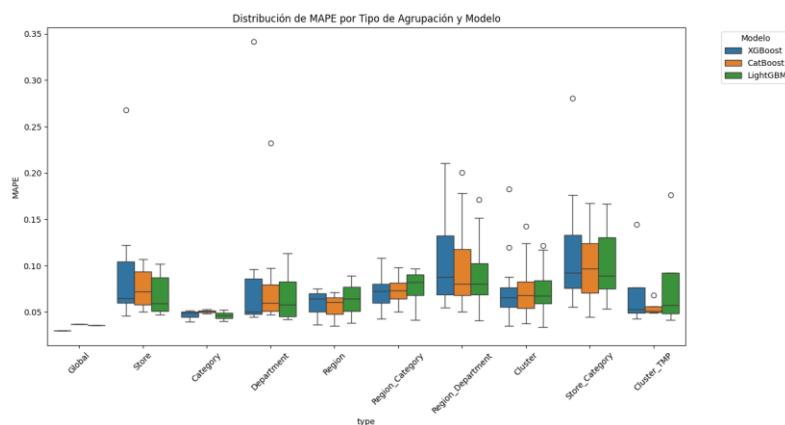
Los modelos se evalúan con métricas estándar:

- **MAPE** (Mean Absolute Percentage Error): Error porcentual medio absoluto.
- **MAE** (Mean Absolute Error): Error medio absoluto.
- **RMSE** (Root Mean Squared Error): Raíz del error cuadrático medio.

Los resultados se almacenan en un DataFrame para comparar el desempeño entre modelos y detectar cuál proporciona mejores predicciones. Gracias a graficar estos resultados buscamos la mejor opción para elegir cual será nuestro modelo, que dependerá de sus resultados y de cuanta profundidad tengan sus agregaciones.

Ya que, aunque la agrupación de manera global tenga los mejores resultados, al ser tan general, cuando especifiquemos con pesos esas predicciones a los "id" de manera individual el error va a ser mucho mayor que si elegimos un modelo que agregue de manera más específica.

A continuación, se muestra la gráfica con la cual elegimos nuestra mejor opción mediante una media de los resultados:

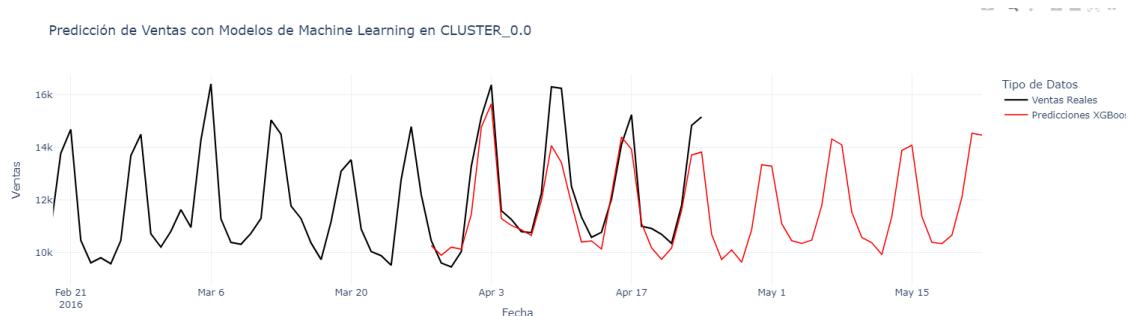


Como podemos apreciar, teniendo en cuenta el grado de especificación de las agrupaciones y su MAPE podemos ver que las agrupaciones por cluster son significativamente mejores. La agrupación llamada "Cluster", es una agrupación que hemos hecho por ítems respecto de las ventas medias al mes, y "Cluster\_TMP" son cluster de series temporales.

Como podemos ver los MAPE son mejores en CatBoost en "Cluster\_TMP" pero vamos a seleccionar XGBoost ya que la diferencia es insignificante y XGBoost es más eficiente por computación, por lo que vamos a seleccionar esa agregación como la mejor para nuestro modelo de predicción de ventas:

294	XGBoost	0.051382	684.040251	966.760570	CLUSTERTMP_0.0	Cluster_TMP
295	CatBoost	0.049727	656.513470	885.939341	CLUSTERTMP_0.0	Cluster_TMP
296	LightGBM	0.050402	677.310837	1007.258028	CLUSTERTMP_0.0	Cluster_TMP
297	XGBoost	0.042700	636.770780	794.585402	CLUSTERTMP_1.0	Cluster_TMP
298	CatBoost	0.048696	719.392396	860.516838	CLUSTERTMP_1.0	Cluster_TMP
299	LightGBM	0.041398	629.537723	760.918434	CLUSTERTMP_1.0	Cluster_TMP
300	XGBoost	0.144105	579.473294	622.182371	CLUSTERTMP_2.0	Cluster_TMP
301	CatBoost	0.068342	262.305584	312.857632	CLUSTERTMP_2.0	Cluster_TMP
302	LightGBM	0.176014	717.861448	768.170081	CLUSTERTMP_2.0	Cluster_TMP
303	XGBoost	0.054114	577.972024	693.313875	CLUSTERTMP_3.0	Cluster_TMP
304	CatBoost	0.052004	564.043960	731.579766	CLUSTERTMP_3.0	Cluster_TMP
305	LightGBM	0.064519	691.591765	797.615874	CLUSTERTMP_3.0	Cluster_TMP

Aquí podemos ver la gráfica de la predicción de ventas diarias para los clusters seleccionados:





### 3.4.5. Cálculo de pesos por ID

Una vez elegido el mejor modelo nos quedamos con sus métricas y el best\_model optimizado para posteriormente poder utilizarlo. Se guardará también las predicciones y las predicciones futuras, ya que vamos a tener que utilizar estas predicciones de ventas diarias por clúster, para junto con los pesos de cada "id" poder calcular el error real y así obtener los verdaderos resultados de nuestra predicción de ventas y poder utilizarlo para el abastecimiento del stock.

#### MACHINE LERANING

Nuestra primera opción fue hacer un modelo de machine learning mediante árboles de decisión (XGBoost), al que nosotros le diéramos las variables temporales, y características de ese producto y nos devolviese el peso que tiene sobre esa predicción de venta diaria.

Para ello se agrupan las ventas por 'clúster' y 'date' para obtener la suma de ventas diarias en cada clúster, calculamos la proporción de ventas de cada producto respecto al total del clúster en cada fecha y reemplazan valores nulos por 0 para evitar problemas en el modelado.

Se obtiene la fecha más reciente y se filtran los datos para considerar sólo los últimos dos años. Hacemos un "feature engineering" para la creación de variables del modelo:

- 1. Media de Ventas Mensuales por Producto:**  
Se calcula la media de ventas de cada producto por mes.

**2. Media del Precio de Venta Mensual:**

Se calcula el precio promedio de venta de cada producto por mes.

**3. Cálculo de Eventos y Festivos:**

Se determina si hay al menos un evento o un festivo en el mes.

**4. Cálculo del Total de Eventos y Festivos por Mes:**

Se cuentan los días con eventos y festivos en cada mes.

Se seleccionan las variables numéricas y categóricas a utilizar en el modelo y se filtran los datos para trabajar solo con un cluster para sacar las predicciones de los pesos en ese determinado cluster. Se define el modelo XGBoost y los rangos de hiperparámetros a optimizar y se busca la mejor combinación de hiperparámetros mediante validación cruzada.

Por último, se calculan las métricas para ver el rendimiento del modelo. Hemos analizado las métricas del clúster y hemos llegado a la conclusión que el modelo no está funcionando correctamente, y es debido a que se complica por la gran cantidad de pesos en 0, y la gran cantidad de ventas intermitentes.

Debido a la limitación de recursos computacionales no hemos podido profundizar en este modelo, pero se ha pensado en un futuro probar a agregarlo semanalmente que será más sencillo y romperlo a nivel diario (proporción de venta para los artículos de esa categoría-tienda a nivel día de la semana).

---

**DISTRIBUCIÓN PROPORCIONAL A VENTAS PASADAS**

---

Se ha optado por un sistema razonado y simple, hemos realizado un ponderador para cada artículo y día a repartir en función de dos pesos,  $w_{cluster}$  (pondera las ventas del ítem sobre el cluster) y  $w_{item}$  (ponderan las ventas del ítem a nivel global y lo distribuye entre las tiendas). De tal forma que, construyas un peso basado en el histórico estacional y para tener en cuenta los artículos que no se han vendido en los últimos 30 días, se calcula el peso repartido para esos artículos poniéndolos su peso a cero y se reparte para el resto de artículos.

Para la realización de hallar los pesos y con ello las ventas individuales por producto hemos realizado la función `prediction_id_daily_TS_x`, que tiene como objetivo calcular predicciones de ventas diarias a nivel de tienda e ítem utilizando datos pasados y genera métricas de error, distribuye pesos de ventas y estima ingresos para cada tienda y producto.

Y la que mejores resultados nos ha dado ha sido con 30 días repartiendo el peso de aquellas ventas que han tenido 0 ventas en los últimos 30 días. La función para calcular los pesos elegida es `prediction_id_daily_TS_30`.

Sus parámetros de entrada son:

- **df:** DataFrame con los datos históricos de ventas.
- **df\_filter:** DataFrame filtrado para el cluster en estudio.
- **y\_pred\_filter:** DataFrame con predicciones de ventas.
- **cluster:** Nombre del cluster en análisis.
- **dict\_metric\_cluster:** Diccionario para almacenar métricas de error a nivel de cluster.
- **dict\_weight:** Diccionario para almacenar los pesos de ventas.
- **dict\_metric\_id:** Diccionario para almacenar métricas de error a nivel de producto.
- **dict\_metric\_daily:** Diccionario para almacenar métricas de error a nivel diario.

Se definen las características (columns\_X) y la variable objetivo (columns\_y) y se agrupan las ventas a nivel de cluster para calcular métricas de predicción.

Se extraen los últimos 28 días para evaluar las predicciones y se compara la predicción del cluster con los valores reales y se calcula la raíz del error cuadrático medio (RMSE) y se guarda la métrica en el diccionario correspondiente.

Se hace el filtrado de los últimos 30 días para el cálculo de pesos, y se calcula la suma de ventas por ítem en los últimos 30 días. Se hace el cálculo de los pesos relativos de cada producto respecto al total del cluster y se eliminan los pesos de productos con ventas nulas para redistribuirlos.

Por último, se ajustan las predicciones según los pesos calculados y gracias a nuestro modelo de machine learning que nos ofrece las predicciones de ventas diarias en un cluster y estos pesos calculados podemos hallar la predicción individual de los productos. Con ello se estiman los ingresos reales y predichos por tienda y se cuantifican los errores en términos de ingresos.

A continuación, se muestran los resultados generales de las predicciones para el cluster 0, en el notebook se puede comprobar el resto de clusters:

date	id	weight_cluster	weight_item	weight	y_pred	daily_sales	sell_price	store	revenue_pred	revenue	
0	2016-03-28	ACCESORIES_1_001_BOS_1	0.000	0.064	0.000	0	1	10.986	South_End	0.000	10.986
1	2016-03-28	ACCESORIES_1_001_BOS_2	0.000	0.098	0.000	0	0	10.986	Roxbury	0.000	0.000
2	2016-03-28	ACCESORIES_1_001_BOS_3	0.000	0.094	0.000	0	0	10.986	Back_Bay	0.000	0.000
3	2016-03-28	ACCESORIES_1_001_NYC_1	0.000	0.147	0.000	0	1	10.986	Greenwich_Village	0.000	10.986
4	2016-03-28	ACCESORIES_1_001_NYC_2	0.000	0.132	0.000	0	0	10.986	Harlem	0.000	0.000
...	...	...	...	...	...	...	...	...	...	...	...
607035	2016-04-24	SUPERMARKET_3_826_NYC_3	0.001	0.102	0.000	1	3	1.536	Tribeca	1.536	4.608
607036	2016-04-24	SUPERMARKET_3_826_NYC_4	0.001	0.151	0.000	1	4	1.536	Brooklyn	1.536	6.144
607037	2016-04-24	SUPERMARKET_3_826_PHI_1	0.001	0.119	0.000	1	2	1.536	Midtown_Village	1.536	3.072
607038	2016-04-24	SUPERMARKET_3_826_PHI_2	0.001	0.094	0.000	0	1	1.536	Yorktown	0.000	1.536
607039	2016-04-24	SUPERMARKET_3_826_PHI_3	0.001	0.058	0.000	0	3	1.536	Queen_Village	0.000	4.608

	date	store	y_pred	daily_sales	revenue_pred	revenue	rsme	error_coste
0	2016-03-28	Back_Bay	166	1008	926.695	6076.882	842.000	-5150.188
1	2016-03-28	Brooklyn	111	820	450.240	4870.672	709.000	-4420.433
2	2016-03-28	Greenwich_Village	416	1153	2070.821	7076.464	737.000	-5005.644
3	2016-03-28	Harlem	341	1229	1831.806	7035.266	888.000	-5203.460
4	2016-03-28	Midtown_Village	227	1078	1042.098	5928.037	851.000	-4885.939
...	...	...	...	...	...	...	...	...
275	2016-04-24	Queen_Village	365	1088	1645.843	6058.006	723.000	-4412.162
276	2016-04-24	Roxbury	531	1217	2954.362	7831.356	686.000	-4876.995
277	2016-04-24	South_End	277	1058	1543.102	6750.200	781.000	-5207.097
278	2016-04-24	Tribeca	1307	2370	7166.119	14567.983	1063.000	-7401.864
279	2016-04-24	Yorktown	434	1343	1958.593	7076.156	909.000	-5117.563

A pesar de que las predicciones subestiman las ventas, los ingresos reales son mayores de lo esperado, lo que es una señal de que la empresa tiene más oportunidades de crecimiento de lo previsto. Si se ajustan los modelos de predicción, se puede optimizar el inventario, mejorar la planificación financiera y aumentar la eficiencia operativa.

### 3.5. Abastecimiento de tiendas - MLOps

A continuación, se redactan los puntos principales de la propuesta de integración del modelo de predicción de ventas con MLOps. Todos los detalles de esta propuesta se encuentra en el anexo **CP-A-G1\_Retail\_Memoria\_Anexo\_PropuestaMLOps**.

El objetivo de esta propuesta es definir cómo aplicar los modelos predictivos de ventas al abastecimiento de tiendas, establecer un esquema de productivización mediante una API e identificar mejoras en la predicción.

Actualmente, en DSMarket se utilizan métodos tradicionales para estimar la demanda de productos, basándose principalmente en la experiencia. Este enfoque es rudimentario y tiene limitaciones significativas en cuanto a precisión y capacidad para ajustarse rápidamente a cambios en la demanda, lo que resulta en sobreabastecimiento o desabastecimiento de productos.

El modelo de Machine Learning (ML) propuesto busca reemplazar estos métodos con predicciones basadas en datos históricos, tendencias, estacionalidad, etc. Esta propuesta permite una gestión más precisa y dinámica del inventario.

Para gestionar de manera eficiente la implementación y mantenimiento de este modelo, se empleará un sistema de MLOps, que automatiza los flujos de trabajo de desarrollo, despliegue y monitorización del modelo. Este sistema MLOps se apoya en **Google Cloud Platform (GCP)** y herramientas como **Cloud Run**, **FastAPI** y **AlloyDB**, lo que garantiza que las predicciones y recomendaciones de reabastecimiento se realicen de manera ágil, escalable y continua, sin intervención manual.

### 3.5.1. Aplicación de ML para el abastecimiento de tiendas

La integración del modelo de predicción con el proceso de abastecimiento actual de DSMarket nos permite mejorar la eficiencia del inventario, reducir las pérdidas y minimizar las roturas de stock, lo que mejora la experiencia del cliente y optimiza los márgenes operativos.

#### INTEGRACIÓN CON EL PROCESO DE ABASTECIMIENTO

Actualmente, el modelo utiliza datos históricos provenientes de las tiendas, los cuales han servido para su entrenamiento. Sin embargo, es necesario combinar esta predicción con información de stock y abastecimiento de la empresa. Para completar el proceso de abastecimiento, se requiere incorporar datos logísticos y de inventario que, hasta ahora, no han sido considerados en el modelo. Además, es fundamental verificar si estos datos están disponibles a nivel de producto o de tienda para calcular un stock óptimo.

En próximos pasos se puede implementar un nuevo modelo de Machine Learning que procese datos logísticos y sea capaz de predecir el stock óptimo a nivel producto-tienda.

#### FLUJO DE DATOS

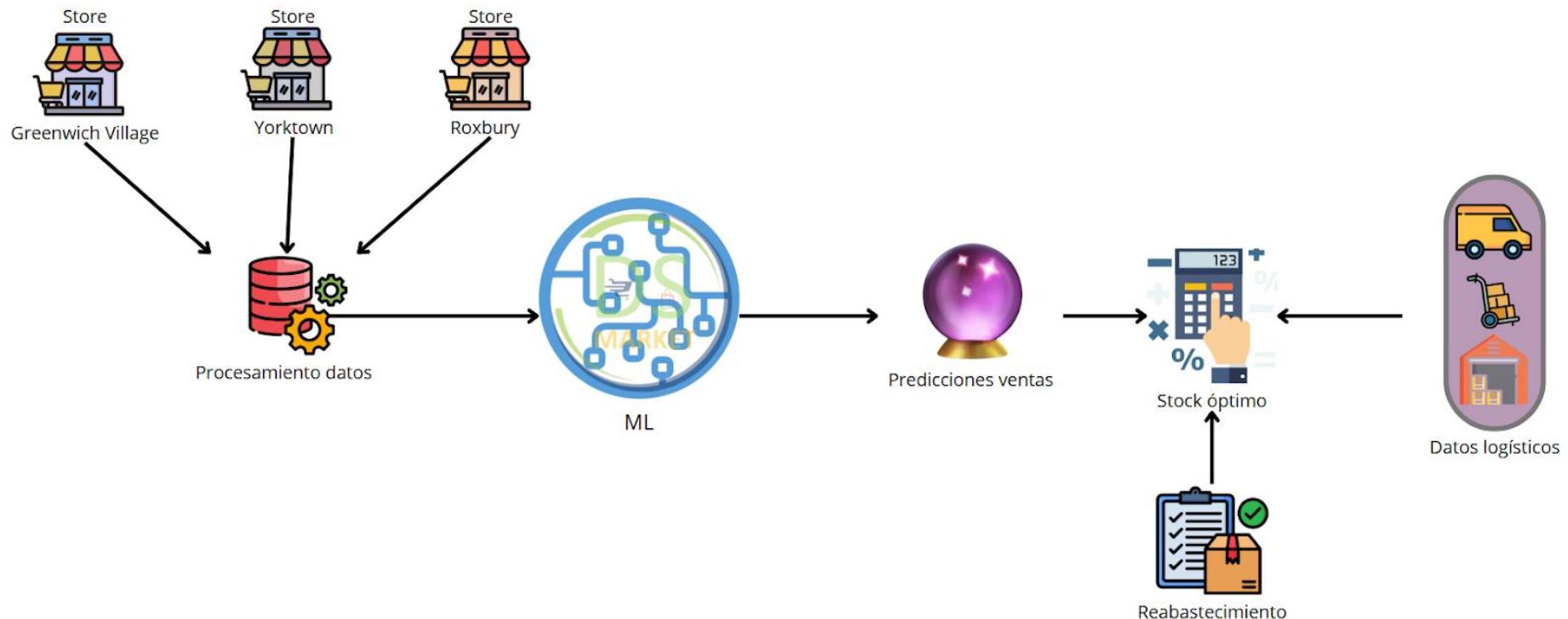
Para llevar a cabo una buena integración entre las diferentes áreas de negocio afectadas es necesario tener un flujo de datos claro y bien definido.

- Entrada de Datos
  - Datos históricos de ventas: Se utilizan los registros de ventas de cada tienda para analizar patrones y comportamientos de la demanda de productos.
  - Calendario de eventos: Datos sobre festividades, promociones, y otros eventos que pueden afectar la demanda de productos.
  - Precios: Cambios de precios y descuentos aplicados a productos que pueden modificar el comportamiento de compra de los clientes.
- Predicción
  - Usando los datos de entrada, el modelo genera predicciones de la demanda futura para cada producto, por tienda, para las siguientes cuatro semanas.
- Optimización
  - Tiempo de entrega: Cuánto tiempo toma recibir los productos desde el almacén.
  - Capacidad de almacenamiento: El espacio disponible en cada tienda para mantener los productos.
  - Nivel de stock: Determinado para evitar tanto el desabastecimiento como el exceso de inventario.

- Recomendaciones de reabastecimiento

- Con todos estos datos somos capaces de crear recomendaciones de cuánto pedir de cada producto en cada tienda para satisfacer la demanda proyectada. Estas recomendaciones se pueden integrar con las plataformas internas de gestión de pedidos para una ejecución eficiente.

- Diagrama flujo de datos



---

### BENEFICIOS ESPERADOS

---

Al aplicar un modelo de predicción de ventas existen numerosos beneficios a diferentes niveles dentro de la empresa. Aquí se recogen las ventajas principales de implementar el modelo a este caso de uso concreto, el abastecimiento de tiendas.

- **Optimizar la gestión del inventario:** Se pueden realizar pedidos más ajustados a las necesidades reales de las tiendas, reduciendo tanto el exceso de inventario como los productos agotados.
  - **Reducir pérdidas por sobreabastecimiento:** Evitar tener productos en exceso, lo que puede llevar a pérdidas económicas por deterioro de productos o almacenamiento innecesario.
  - **Minimizar desabastecimientos:** Garantizar que las tiendas tengan suficiente stock de los productos más demandados, evitando la falta de productos que afecten las ventas y la satisfacción del cliente.
  - **Mejor planificación logística:** Con las predicciones de demanda y las recomendaciones de reabastecimiento, se optimiza la logística, lo que puede reducir costos de transporte y mejorar la eficiencia en la distribución de productos.
- 

### 3.5.2. Productivización del modelo

En este punto explicamos cómo hemos pasado de un modelo experimental de predicción de ventas a una solución operativa que se puede usar de manera continua dentro de la infraestructura de la empresa.

Para integrar la solución dentro de la operación de DSMarket, se ha desplegado una **API** en **Google Cloud Run**. Esta API permitirá que los sistemas internos consulten predicciones en tiempo real.

---

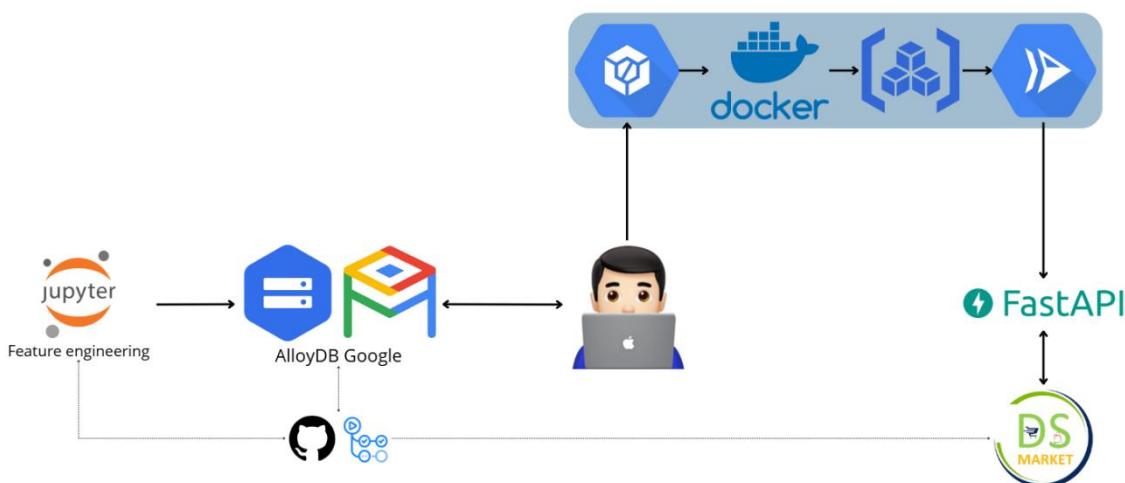
### ARQUITECTURA DE LA SOLUCIÓN

---

La solución MLOps se despliega en Google Cloud Platform usando **Google Cloud Run** para ejecutar la API de predicción con escalabilidad automática y Docker para contenerizar el modelo, asegurando consistencia. **AlloyDB** almacena datos históricos y predicciones con alto rendimiento y con CI/CD automatizado vía Google Cloud Build y GitHub Actions.

Esta arquitectura se encuentra explicada con todo detalle en el anexo y el código correspondiente en esta organización de GitHub: [DSMarketNPF01](#)

A continuación se muestra un esquema de las herramientas usadas a días de hoy.



Una de las primeras mejoras a implementar en el futuro próximo será el uso de MLFlow para gestionar el ciclo de vida del modelo.

#### ENDPOINTS DE LA API

/predict → Devuelve la predicción de ventas para un producto/tienda en los 28 días posteriores.

Esto permitirá que el equipo de DS Market acceda a recomendaciones optimizadas de reabastecimiento en tiempo real.

En cuanto a las extensiones y mejoras del modelo se proponen dos líneas de mejoras complementarias.

En primer lugar, se abordarán las relacionadas con **los datos**, incluyendo su obtención, procesamiento y calidad. A continuación, se presentarán las extensiones directamente vinculadas **al modelo** en sí, como su arquitectura, entrenamiento y ajuste.

### 3.5.3. Preprocesamiento y enriquecimiento de datos

Se proponen las siguientes mejoras con el objetivo de mejorar la calidad y cantidad de los datos.

#### INCORPORACIÓN DE DATOS LOGÍSTICOS

Para mejorar el sistema de abastecimiento se deben considerar diversos factores logísticos:

- Tiempo de entrega.
- Disponibilidad en almacenes.
- Capacidad de almacenamiento de las tiendas.

---

## SEGMENTACIÓN AVANZADA DE PRODUCTOS Y DEPARTAMENTOS

---

Mejorar la información sobre las categorías y departamentos de los productos mejora la segmentación de la demanda y permite ajustar mejor las estrategias de abastecimiento.

Sería interesante incluir información sobre la caducidad de los productos, por ejemplo:

- Productos frescos: demanda a corto plazo, minimizar el riesgo de caducidad.
  - Electrodomésticos y artículos no perecederos: La demanda puede depender más de promociones.
- 

## INCORPORACIÓN DE FACTORES INTERNOS

---

- Eventos especiales

En el análisis efectuado se ha demostrado el impacto que tienen los eventos en el número de ventas por eso consideramos que el modelo debe estar preparado para incluir información más detallada sobre estos datos.

- Precios y promociones

Actualmente existen datos de precios pero no se registran si los cambios que sufren los precios se deben a promociones o rebajas.

---

## INCORPORACIÓN DE FACTORES EXTERNOS

---

Los cambios en el entorno social, económico y cultural pueden tener un impacto significativo sobre el comportamiento de los consumidores y, por lo tanto, sobre la demanda de productos.

- Festivos nacionales y locales.
  - Datos demográficos.
  - Factores económicos.
  - Condiciones climáticas.
- 

## AUTOMATIZACIÓN DE CÁLCULO DE STOCK ÓPTIMO:

---

Definir umbrales de reabastecimiento según las predicciones.

- **Definir umbrales de stock mínimo y máximo:** estos umbrales son calculados en función de las predicciones de demanda realizadas por el modelo, los tiempos de entrega y la capacidad de almacenamiento de cada tienda.

- **Cálculo Automático de las cantidades de reabastecimiento:** Una vez definidos los umbrales, el sistema puede calcular automáticamente las cantidades necesarias para reponer el stock.
- **Ajuste en tiempo real:** De manera que el modelo esté siempre optimizado para las condiciones cambiantes del mercado.

---

#### PREDICCIÓN DE ROTURAS DE STOCK:

---

Identificación de productos con alto riesgo de agotamiento.

El modelo de predicción de ventas también debe ser capaz de detectar posibles stock-outs (agotar existencias) y optimizar las decisiones de reabastecimiento.

### 3.5.4. Extensiones en modelo de predicción

A medida que el sistema de predicción de ventas evoluciona, se han identificado varias líneas de mejora orientadas a aumentar la precisión, escalabilidad y flexibilidad de la solución actual. Estas extensiones permiten anticipar nuevas necesidades del negocio y garantizar una integración más fluida dentro de la infraestructura tecnológica de DSMarket.

---

#### MODELOS MÁS PROFUNDOS Y COMPUTACIÓN ESCALABLE

---

Con más tiempo de entrenamiento y acceso a mayores recursos computacionales (Cloud Run + AlloyDB), se abre la posibilidad de emplear modelos más complejos y profundos que puedan capturar mejor las interacciones no lineales entre variables, especialmente en contextos de alta estacionalidad o con patrones de venta muy específicos.

---

#### MEJORA EN LA ASIGNACIÓN DE PESOS A PRODUCTOS

---

Actualmente, el modelo **middle-to-bottom** calcula las ventas a nivel de producto (id) distribuyendo las predicciones del clúster a los productos individuales según su histórico reciente (30 días). Esta aproximación puede resultar limitada si no se consideran dinámicas del mercado.

Esto se podría mejorar implementando un modelo de regresión auxiliar que calcule estos factores de asignación utilizando variables como:

- Media de precios y variación
- Tendencias recientes de ventas
- Fechas y eventos
- Información regional
- Categoría y estacionalidad del producto

Esto permitirá ajustar los pesos dinámicamente según el contexto actual y no solo en base al histórico.

---

### MEJORAS EN EL SISTEMA MLOPS

---

Aunque el sistema MLOps ya se encuentra operativo, se han identificado varias áreas de expansión:

- Migración completa a la nube: se procederá a subir todo el histórico a AlloyDB, incluyendo actualizaciones en tiempo real mediante pipelines automatizados.
- Ejecución del modelo online: se desplegará una versión del modelo entrenado con seguimiento a través de MLFlow.
- Automatización y mantenimiento: se reforzará el pipeline de CI/CD en Google Cloud Build, y se habilitará el logging centralizado mediante Google Cloud Logging.

Estas medidas permitirán una mayor trazabilidad y fiabilidad del sistema en producción.

---

### PERFILES PERSONALIZADOS EN LA API

---

Se plantea la incorporación de una capa de autenticación basada en perfiles de usuario dentro de la API:

- **Administrador**: acceso total, posibilidad de cambiar fecha de predicción y parámetros del modelo.
- **Gestor de tienda**: acceso restringido a su tienda y fechas específicas.
- **Analista de datos**: acceso a logs, históricos y predicciones por lote.

Esto permitirá una personalización del acceso según los distintos roles dentro de DSMarket, alineándose con las necesidades operativas y de seguridad.

### 3.5.5. Diseño de prueba piloto

Antes de implementar la aplicación de modelos de predicción en DS Market a gran escala, se realizará un piloto para evaluar su efectividad.

Con esta prueba queremos demostrar el impacto económico positivo que tendrá en la empresa el uso de modelos de predicción para el caso de uso de reabastecimiento de productos.

---

### DESCRIPCIÓN ENFOQUES

---

A continuación, se definen los dos enfoques actuales:

- Rudimentario: Enfoque basado en reglas simples y procesos manuales para prever ventas e inventarios, con márgenes de error altos que afectan a la toma de decisiones. media móvil(\*\*\*)�.
- Machine Learning: modelo de predicción de ventas basado en datos históricos de las tiendas de New York, Boston y Philadelphia.

### METODOLOGÍA DE LA PRUEBA PILOTO

- Selección de KPIs

Ventas por producto, margen de ganancia, niveles de inventarios, precisión de las predicciones, costos operativos, y eficiencia en la reposición de inventarios.

- Selección de productos y tiendas

Se probarán productos en las 10 tiendas distribuidas entre Nueva York, Boston y Filadelfia. Las categorías y departamentos serán utilizados para segmentar los resultados y analizar los beneficios de ML en distintos segmentos de productos.

- Ejecutar ambos enfoques

Se ejecutarán ambos enfoques (tradicional y ML) durante un período determinado de 12 semanas, y se recolectarán los datos necesarios para la comparación.

#### *Resultados y análisis comparativo*

Durante las semanas de prueba, se recopilarán datos sobre ventas, márgenes de ganancia y predicciones de ventas para los productos de las 10 tiendas.

Posteriormente, se llevará a cabo un análisis estadístico utilizando métricas como t-test, RMSE y R2 para comparar la precisión de los enfoques tradicional y de Machine Learning, con el fin de evaluar la efectividad del modelo en comparación con el método anterior.

#### *Evaluación de la mejora y ROI*

Se da por hecho que después de analizar los resultados y comparar ambos enfoques el modelo de predicción de ventas obtendrá un mejor resultado que el enfoque antiguo.

Por lo tanto, una vez conocemos cómo mejora el sistema de Machine Learning al sistema previo, podemos evaluar cómo va a afectar a la empresa implementar este nuevo sistema a nivel global.

Se calculará el retorno de inversión (ROI) comparando los costos de implementación y operación del modelo de ML con los beneficios tangibles obtenidos, que incluyen la optimización de procesos, la mejora de márgenes y la capacidad de toma de decisiones más precisas.

### Presentación de resultados finales

Se presentará un informe ejecutivo con gráficos y tablas comparativas para mostrar los resultados de ambos enfoques en función de los KPIs clave, facilitando una comparación directa entre el enfoque tradicional y el de Machine Learning.

## 4. Métricas y resultados

### 4.1. Mejor modelo de predicción de ventas

Para evaluar el desempeño del modelo XGBoost en la predicción de ventas por clúster, se han calculado las siguientes métricas: MAPE (Mean Absolute Percentage Error), MAE (Mean Absolute Error) y RMSE (Root Mean Squared Error). Los resultados obtenidos son los siguientes:

	model	MAPE	MAE	RMSE	GROUP	type
0	XGBoost	0.051382	684.040251	966.760570	CLUSTER_0.0	cluster_base_on_dtw
1	XGBoost	0.042700	636.770780	794.585402	CLUSTER_1.0	cluster_base_on_dtw
2	XGBoost	0.144105	579.473294	622.182371	CLUSTER_2.0	cluster_base_on_dtw
3	XGBoost	0.054114	577.972024	693.313875	CLUSTER_3.0	cluster_base_on_dtw

- Los valores de **MAPE** en los clústeres 0.0, 1.0 y 3.0 son relativamente bajos (entre 4% y 5.4%), lo que indica que el modelo **predice las ventas con una buena precisión** en estos grupos.
- Un MAPE menor a 10% suele considerarse un buen resultado en problemas de predicción de series temporales.
- Las gráficas muestran que **las predicciones de XGBoost siguen la tendencia de las ventas reales**, capturando bien los picos y caídas en la mayoría de los casos.
- Esto sugiere que el modelo ha logrado aprender patrones estacionales y de demanda en los datos.
- Los clústeres con mayores volúmenes de ventas (por ejemplo, CLUSTER\_1.0) tienen **errores absolutos bajos** en términos de MAE y RMSE, lo cual es positivo porque estos representan la mayor parte de las ventas totales.
- La capacidad de predecir ventas con esta precisión puede ayudar en **la optimización de inventarios, estrategias de marketing y planificación logística** en las diferentes tiendas.
- Si se mejora aún más la precisión en clústeres con mayores errores (como CLUSTER\_2.0), se podrá lograr una planificación más efectiva.

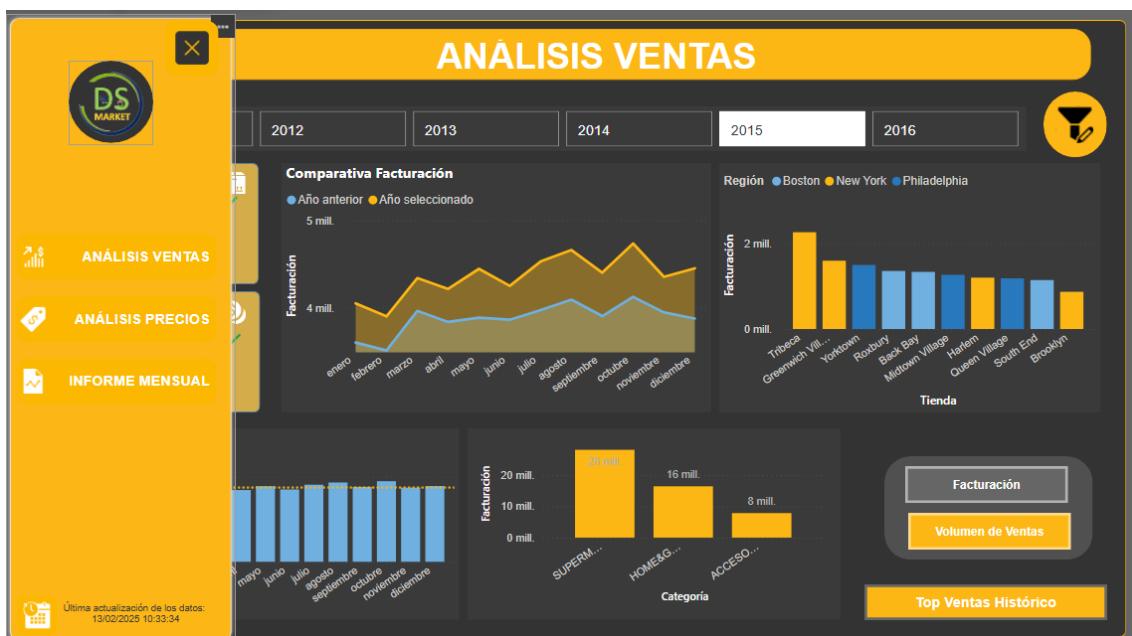
En general, estos resultados muestran que el modelo XGBoost **es una buena base para predecir ventas**, aunque aún hay margen de mejora, especialmente en algunos clústeres donde el error es más alto.

## 4.2. Cuadro de mando

Con este dashboard de Power BI hemos reunido y analizado información clave de nuestro desempeño comercial, permitiéndonos visualizar de forma interactiva y dinámica los indicadores que respaldan nuestras decisiones estratégicas. Se ha publicado en la web para poner en disposición al equipo de negocio y marketing ([PowerBi](#)).

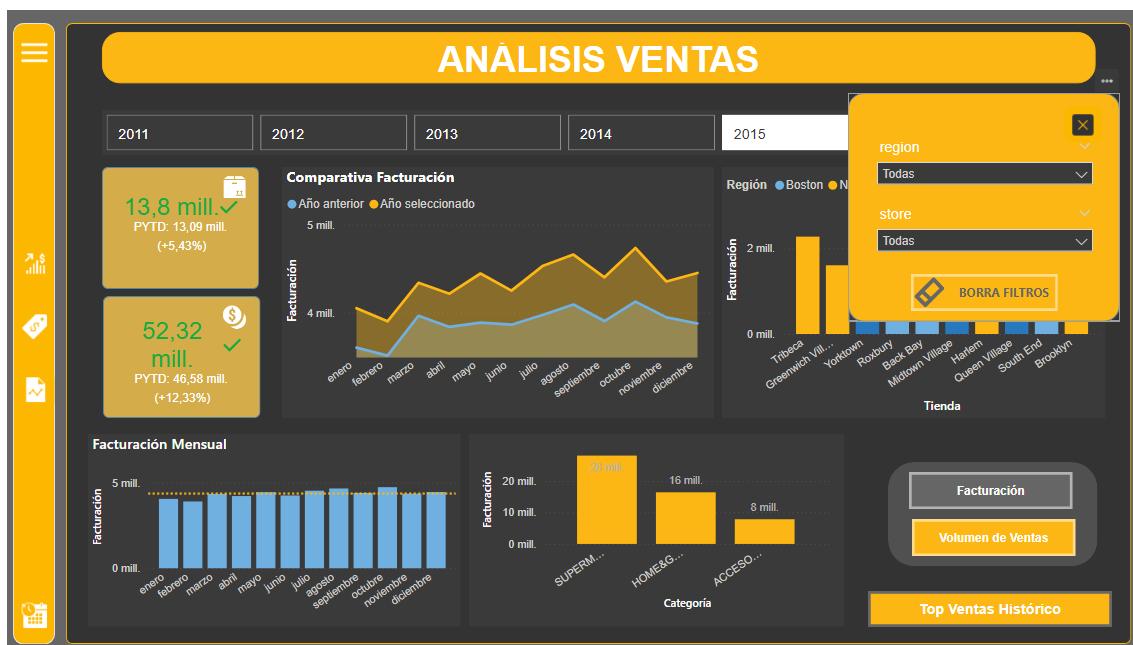
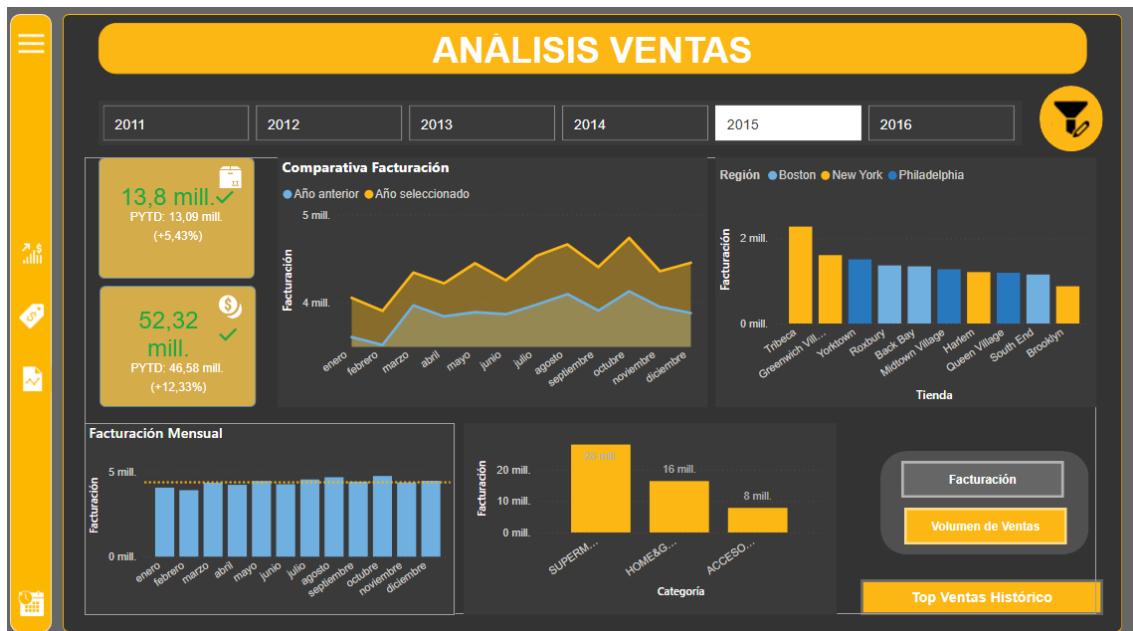


Cuenta con un menú desplegable en todas las páginas para poder navegar entre los contenidos de forma sencilla, incluso es posible sin desplegar el menú pulsando sobre los iconos de la barra.



## ANÁLISIS DE VENTAS

En la página de Análisis de Ventas hemos desarrollado una vista integral de nuestros resultados, combinando indicadores clave y gráficos que comparan la facturación actual con períodos anteriores. Hemos incorporado un gráfico de líneas para visualizar la tendencia en el tiempo, así como gráficos de barras que destacan el desempeño por tienda o región y una representación mensual que nos ayuda a identificar patrones estacionales, lo que nos permite evaluar nuestro rendimiento de manera rápida y precisa.



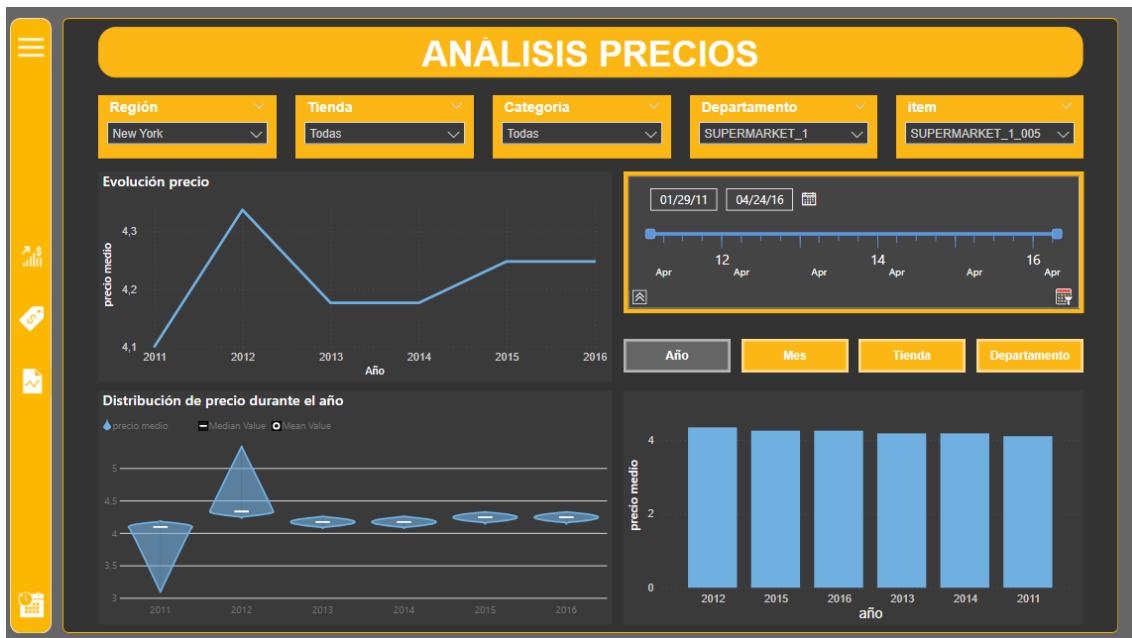
Se puede acceder a la Evolución Histórica Top 10, mediante el botón de la parte inferior derecha. Se ha implementado filtros dinámicos para región, tienda, categoría y

departamento, lo que nos permite analizar con detalle el desempeño de nuestros establecimientos. A través de un gráfico de barras horizontales, hemos plasmado el ranking de los 10 principales contribuyentes en ventas, facilitando la comparación de la evolución de cada uno a lo largo del tiempo y permitiéndonos ajustar nuestras estrategias de mercado de forma oportuna.



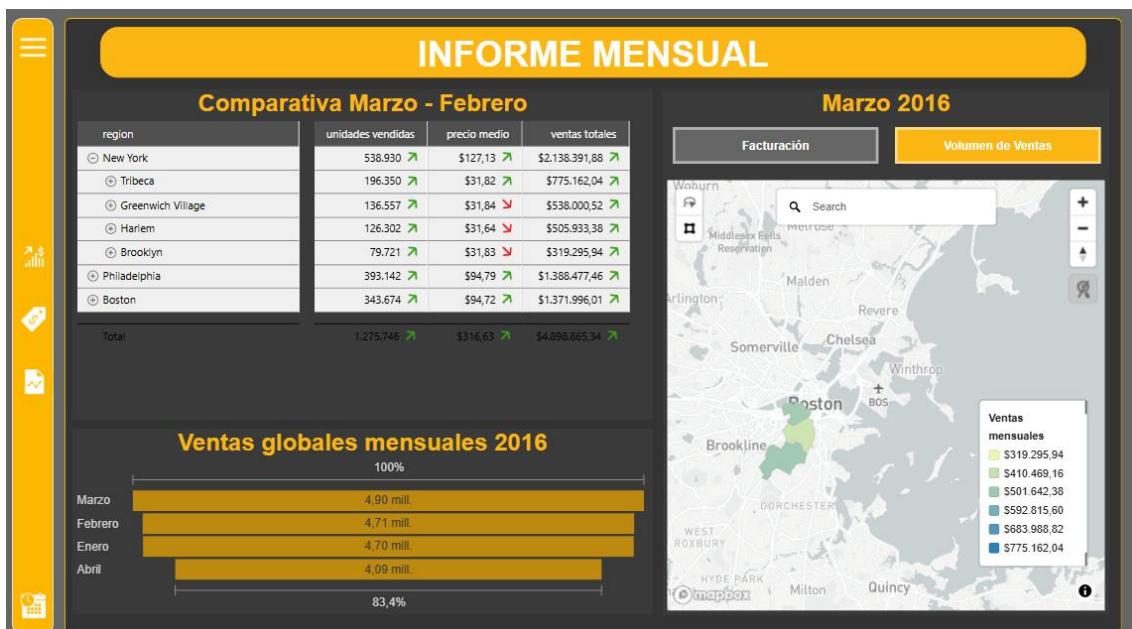
## ANÁLISIS DE PRECIOS

En la página de Análisis de Precios nos hemos enfocado en examinar la evolución de nuestros precios a lo largo del tiempo. Hemos integrado un gráfico de líneas que refleja la variación histórica de los precios y complementado esta visualización con una representación de la distribución de precios y un gráfico de barras que muestra el precio medio anual. Con esta información, podemos identificar tendencias, detectar variaciones estacionales y ajustar nuestras estrategias de pricing de manera informada.



## INFORME MENSUAL

En la página del Informe Mensual correspondiente a marzo de 2016, hemos desarrollado un análisis comparativo cuyo objetivo es mostrar el último mes y contrastarlo con el mes anterior. Para ello, hemos utilizado una tabla que detalla las variaciones por región, un mapa geográfico que ilustra la distribución de nuestras ventas y un gráfico que resume el comportamiento mensual a lo largo del año, lo que nos permite identificar tendencias y áreas de mejora para optimizar nuestras decisiones estratégicas.



Todas estas páginas son dinámicas y se pueden filtrar por región, categoría, fecha, tienda, entre otros, lo que nos brinda la flexibilidad de profundizar en el análisis de cada aspecto según nuestras necesidades y objetivos.

## 5. Conclusiones

Gracias a la exploración de diversos modelos de clustering, seleccionamos los tres que consideramos más óptimos y los incorporamos en los modelos de series temporales que se desarrollaron posteriormente. Finalmente, el modelo que arrojó mejores resultados fue el de clustering basado en series temporales utilizando Dynamic Time Warping (DTW).

El proyecto ha demostrado que es posible predecir con éxito las ventas de las tiendas mediante machine learning, utilizando XGBoost y técnicas de clustering. La estructura de ventas varía entre ciudades y tipos de productos, lo que resalta la necesidad de estrategias comerciales diferenciadas. Aunque los resultados son positivos, hay oportunidades claras de mejora que pueden aumentar la precisión y utilidad del modelo para la toma de decisiones estratégicas.

Este modelo no solo predice las ventas, sino que transforma la manera en que la empresa toma decisiones estratégicas. Su implementación garantiza una mayor eficiencia operativa al reducir costos asociados al sobre stock y desabastecimiento, además de un aumento en los ingresos gracias a la optimización de promociones y estrategias de venta basadas en predicciones precisas. También permite una mejor planificación financiera al generar proyecciones confiables que facilitan la toma de decisiones a largo plazo. Asimismo, otorga una ventaja competitiva al anticiparse a las tendencias del mercado y responder con rapidez a cambios en la demanda. Finalmente, su escalabilidad y automatización lo convierten en un modelo adaptable y preparado para crecer, integrándose fácilmente en nuevos entornos de negocio.

## 6. Pasos futuros

Tras haber analizado la integración del modelo en el caso de uso de abastecimiento de tiendas y presentado una propuesta de prueba piloto con una duración de 12 meses, consideramos oportuno destacar una serie de acciones adicionales que, si bien no forman parte directa de este caso específico, representan pasos clave para el crecimiento y la mejora continua de DSMarket a nivel global.

### 6.1. Propuestas

A continuación, se detallan estas iniciativas estratégicas orientadas a reforzar la infraestructura analítica, operativa y tecnológica de la empresa.

---

## OPTIMIZACIÓN CONSTANTE BASADA EN DATOS

---

Establecer un ciclo de feedback continuo basado en datos permitirá optimizar constantemente los modelos de predicción y mejorar la precisión de las decisiones estratégicas.

Es esencial analizar de manera periódica los resultados obtenidos, comparándolos con el rendimiento real de las ventas para identificar desviaciones y áreas de mejora. Este proceso debe incluir la reevaluación de las variables utilizadas en los modelos, la incorporación de nuevas fuentes de datos y la actualización de los algoritmos para adaptarse a cambios en las tendencias del mercado.

Además, la retroalimentación entre los equipos de análisis de datos, marketing y operaciones facilitará una respuesta ágil ante variaciones en la demanda, permitiendo ajustar estrategias de precios, promociones y abastecimiento de productos en función de información actualizada y precisa.

---

## CAPACITACIÓN CONTINUA DEL PERSONAL

---

Implementar programas para formar a los empleados en técnicas de análisis de datos y el uso de herramientas digitales, ayudando a que todos se sientan cómodos utilizando datos para la toma de decisiones.

La empresa puede adaptar estos cursos a diferentes niveles de conocimiento de manera que poco a poco todo el personal de la empresa comenzará a darle más valor a los datos y será la propia empresa la que se beneficie de ello.

---

## IMPLEMENTACIÓN DE MÉTRICAS DE ÉXITO

---

Evaluar si la empresa actualmente calcula este tipo de métricas y comenzar a introducir como variables a analizar el impacto de las iniciativas de ciencia de datos en los resultados del negocio.

La definición de indicadores clave de rendimiento (KPIs) permitirá medir con precisión la efectividad de los modelos predictivos y su contribución al negocio de DS Market. La optimización de estrategias comerciales. Algunos KPIs que se pueden incluir serían: precisión de las predicciones de ventas, la reducción de costos operativos, el incremento en la rotación de inventario y la mejora en la rentabilidad de los productos.

---

## INTEGRACIÓN DE DATOS EXTERNOS

---

Tal y como se mencionó en la propuesta del caso de uso de abastecimiento de tiendas, consideramos de vital importancia integrar los datos de la empresa con datos externos.

Esto nos ayudará a obtener una visión más completa del consumidor y su comportamiento.

En primer lugar, el uso de **datos meteorológicos** permitirá establecer correlaciones entre las condiciones climáticas y la demanda de productos específicos. Además, el análisis de eventos climáticos extremos, como tormentas o huracanes, facilitará la planificación de inventarios y la logística de distribución, reduciendo el impacto en la cadena de suministro.

Por otro lado, la incorporación de **datos demográficos** permitirá segmentar con mayor precisión el mercado según características socioeconómicas y de consumo de la población en cada ubicación. Factores como edad, ingresos y composición familiar pueden influir en las preferencias de compra.

Además, el análisis de **datos de clientes**, obtenidos a partir de programas de fidelización y compras en línea, permitirá una mayor personalización de las estrategias de marketing. Mediante la segmentación basada en el comportamiento de compra, se podrán diseñar promociones específicas para diferentes grupos de clientes, aumentando la efectividad de las campañas y la fidelización.

Finalmente, el **monitoreo de la competencia**, a través del seguimiento de precios y promociones de tiendas similares en la misma área geográfica, ayudará a optimizar la estrategia de precios y mantener la competitividad en el mercado.

Integrando estos datos la empresa podrá mejorar significativamente su capacidad de anticipación a la demanda, abastecimiento, estrategias de marketing, etc

### ANÁLISIS DE LA CESTA DE LA COMPRA

El análisis de la cesta de la compra es una herramienta clave para entender los patrones de compra de los clientes y optimizar tanto la disposición de los productos en las tiendas como las estrategias de marketing. Mediante técnicas como el análisis de asociación es posible identificar combinaciones frecuentes de productos comprados juntos y extraer reglas de asociación que permitan mejorar la experiencia de compra y aumentar la rentabilidad.

Esta información resulta valiosa para la creación de promociones cruzadas efectivas. Con base en las asociaciones detectadas, se pueden diseñar estrategias de descuentos y paquetes promocionales que fomenten el aumento del ticket de compra..

Aplicar un análisis sistemático de la cesta de la compra permitirá optimizar la distribución de productos en las tiendas, diseñar campañas promocionales más efectivas y personalizar la oferta para cada cliente, generando así un impacto positivo en la rentabilidad y la satisfacción del consumidor.

---

## IMPLEMENTACIÓN DE MODELOS DE PRICING DINÁMICO

---

Adoptar un sistema de precios dinámicos basado en algoritmos de Machine Learning permitiría ajustar los precios de los productos en tiempo real según factores como la demanda, la competencia, la estacionalidad y las tendencias del mercado. Este enfoque optimizaría la rentabilidad y respondería de manera más ágil a cambios en el comportamiento del consumidor.

---

## MODELOS DE SEGMENTACIÓN TIENDAS

---

Aprovechando los datos históricos de ventas y patrones de compra, se puede construir un sistema de clustering dinámico para segmentar las tiendas en función de su comportamiento y demanda de productos. Esto permitiría adaptar estrategias comerciales personalizadas para cada grupo de tiendas. Aseguraría un ajuste preciso de surtido, precios y promociones según el perfil de cada tienda.

---

## INCORPORACIÓN DE GEOMARKETING

---

El geomarketing es una herramienta clave para mejorar la eficiencia operativa y optimizar las estrategias comerciales mediante el uso de datos geoespaciales. La integración de información geográfica y análisis de comportamiento del consumidor permitirá a DSMarket adaptar su oferta de productos, mejorar la logística de distribución y maximizar la rentabilidad en cada una de sus tiendas.

DS Market podrá identificar las zonas con mayor densidad de pedidos y diseñar rutas de entrega óptimas que minimicen costos y reduzcan los tiempos de transporte. El uso de herramientas GIS permitirá evaluar factores clave como el tráfico, la geografía y la distancia a las tiendas, garantizando una distribución más eficiente.

Para potenciar esta estrategia, la empresa puede implementar tecnología avanzada de ruteo, integrando software de optimización de rutas que tenga en cuenta variables en tiempo real, como condiciones del tráfico, disponibilidad de productos y demanda proyectada. Esto no solo reducirá el tiempo de entrega, sino que también disminuirá los costos operativos y mejorará la experiencia del cliente.

Asimismo, la gestión de flotas mediante el monitoreo continuo de la logística permitirá reaccionar de manera ágil ante imprevistos y optimizar la planificación de entregas. La integración de datos geoespaciales en la toma de decisiones garantizará un control más preciso sobre las operaciones y facilitará ajustes en tiempo real.

Otro aspecto clave es la integración del geomarketing con las predicciones de ventas. Al utilizar modelos de Machine Learning para prever la demanda en distintas regiones, DSMarket podrá anticiparse a las necesidades de abastecimiento y asegurar que cada

tienda cuenta con los productos adecuados en el momento oportuno, evitando quiebres de stock y maximizando la eficiencia en la cadena de suministro.

Para asegurar una integración efectiva del geomarketing en la estrategia de DSMarket, se recomienda seguir un enfoque escalonado:

- Fase Inicial: Analizar la distribución actual de las tiendas en relación con los datos demográficos, patrones de compra y volumen de ventas por región. Identificar áreas estratégicas donde la optimización geoespacial pueda generar mayor impacto.
- Pruebas Piloto: Implementar la optimización de rutas en zonas específicas y medir su efectividad en términos de reducción de costos, mejora en tiempos de entrega y satisfacción del cliente. Ajustar los modelos y estrategias según los resultados obtenidos.
- Feedback y Retroalimentación: Recoger datos operativos y opiniones de clientes sobre la experiencia de compra y entrega. Utilizar esta información para ajustar y perfeccionar continuamente las rutas, la gestión de inventario y las estrategias de marketing geoespacial.

En conclusión, la adopción de herramientas de geomarketing permitirá a DSMarket no solo optimizar su logística y operaciones, sino también fortalecer su estrategia comercial a través de un enfoque basado en datos, asegurando una mayor eficiencia y rentabilidad a largo plazo.

## 6.2. Beneficios Esperados:

La incorporación de técnicas avanzadas de análisis de datos en la estrategia de DSMarket traerá consigo una serie de beneficios clave que impactarán directamente en la eficiencia operativa, la satisfacción del cliente y la rentabilidad del negocio.

### MEJORA EN LAS PREDICIONES DE VENTAS

El uso de datos más contextualizados y relevantes permitirá una mayor precisión en los modelos de predicción de ventas. Al integrar variables externas como datos meteorológicos, tendencias de consumo por ubicación y patrones de compra históricos, las estimaciones serán más específicas para cada tienda. Esto contribuirá a una gestión más eficiente del inventario, asegurando que los productos más demandados estén disponibles en el momento y lugar adecuados, reduciendo así el desperdicio y optimizando los costos de almacenamiento.

---

### AUMENTO EN LA SATISFACCIÓN DEL CLIENTE

---

Un conocimiento más profundo de los hábitos y preferencias de los clientes permitirá personalizar la oferta de productos y promociones de manera más efectiva. Al adaptar las estrategias de venta a las necesidades específicas de cada segmento de consumidores, se fortalecerá la fidelización y se incrementará la satisfacción del cliente. Además, la optimización de los tiempos de entrega y la disponibilidad de productos mejorará la experiencia de compra, generando una percepción más positiva de la marca.

---

### EVALUACIÓN Y AJUSTES:

---

Establecer métricas para evaluar el impacto de estas integraciones en las decisiones de negocio y ajustar los modelos según los resultados observados. Esto ayudará a calibrar continuamente el enfoque y maximizar el ROI de las iniciativas de análisis de datos.

Con estas adiciones, tu propuesta no solo enriquecerá el modelo predictivo sino que también proporcionará un marco más robusto y flexible para la toma de decisiones en DSMarket. Es importante mencionar la viabilidad técnica y la necesidad de asegurar que los sistemas actuales puedan manejar la integración y el procesamiento de estos nuevos datos.