



Máster en Data Science & AI
Tutor: Mati Hermida

Fecha de entrega: 29-04-2024
Fecha de la Defensa: 30-04-2024

TFM DATA SCIENCE

PROYECTO: DSMarket

Grupo 4:

Cristopher Pazmiño
Daniel González
Simeon Milenov



ÍNDICE

<i>Objetivos y Procesos.....</i>	4
<i>Tarea 1: Análisis (Datos).....</i>	5
<i>Estructura de la empresa</i>	6
<i>Análisis.....</i>	9
<i>Tarea 2: Clustering.....</i>	16
<i>Tarea 3: Predicciones</i>	20
<i>Comparativa de uso</i>	22
<i>Tarea 4: Caso de uso de abastecimiento de tiendas (MLOps).....</i>	27



Objetivos

El objetivo principal es impulsar el negocio a través de una transformación digital.

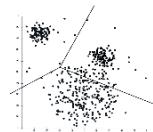
Para lograrlo, aplicaremos conocimientos y herramientas de Data Science en un análisis a fondo de “DS Market”, utilizando los datos disponibles. Este análisis tiene como finalidad identificar oportunidades clave de optimización operativa que mejoren la gestión de inventarios, reconocer los principales grupos de clientes para así tomar decisiones estratégicas más inteligentes con el objetivo de destacar la posición competitiva de la empresa en el mercado.

Además, mediante la implementación de modelos predictivos de ventas, buscamos que la compañía se adapte de manera más eficaz a las tendencias actuales y se adelante a los cambios futuros en el mercado. Esto se integrará en un sistema automatizado que equipará a los empleados con recursos precisos para tomar decisiones más informadas y efectivas.

Procesos



Tarea 1: Análisis



Tarea 2: Clustering



Tarea 3: Predicción de ventas



Tarea 4: Caso de uso de abastecimiento de tiendas (MLOps)

Tarea 1: Análisis

Datos

Identificamos las dimensiones los 3 archivos disponibles (número de filas y columnas)

- Data set de Ventas(pd_sales): (30490, 1920) - 446.6+ MB
- Data set de Calendario(pd_calendar): (1913, 5) - 74.9+ KB
- Data set de Precios(pd_prices): (6965706, 5) - 265.7+ MB

1. Ventas:

Se registran 1913 días de ventas con sus respectivos artículos, id, categoría, tienda, código de la tienda y región el cual son 3 ciudades reconocidas como New York, Boston y Philadelphia.

2. Calendario:

Indica las fechas el cual van desde el 2011-01-29 hasta el 2016-04-24 y eventos. Los eventos son: SuperBowl, Ramadan starts (Inicio del Ramadán), Thanksgiving (Acción de Gracias), NewYear y Easter (Semana Santa). Se ha creado una nueva columna "yearweek" que convierte las fechas de "date" en año semana, por ejemplo el 29 de enero corresponde a la semana 05 así que "2011-01-29" en yearweek sería "201105"

3. Precios:

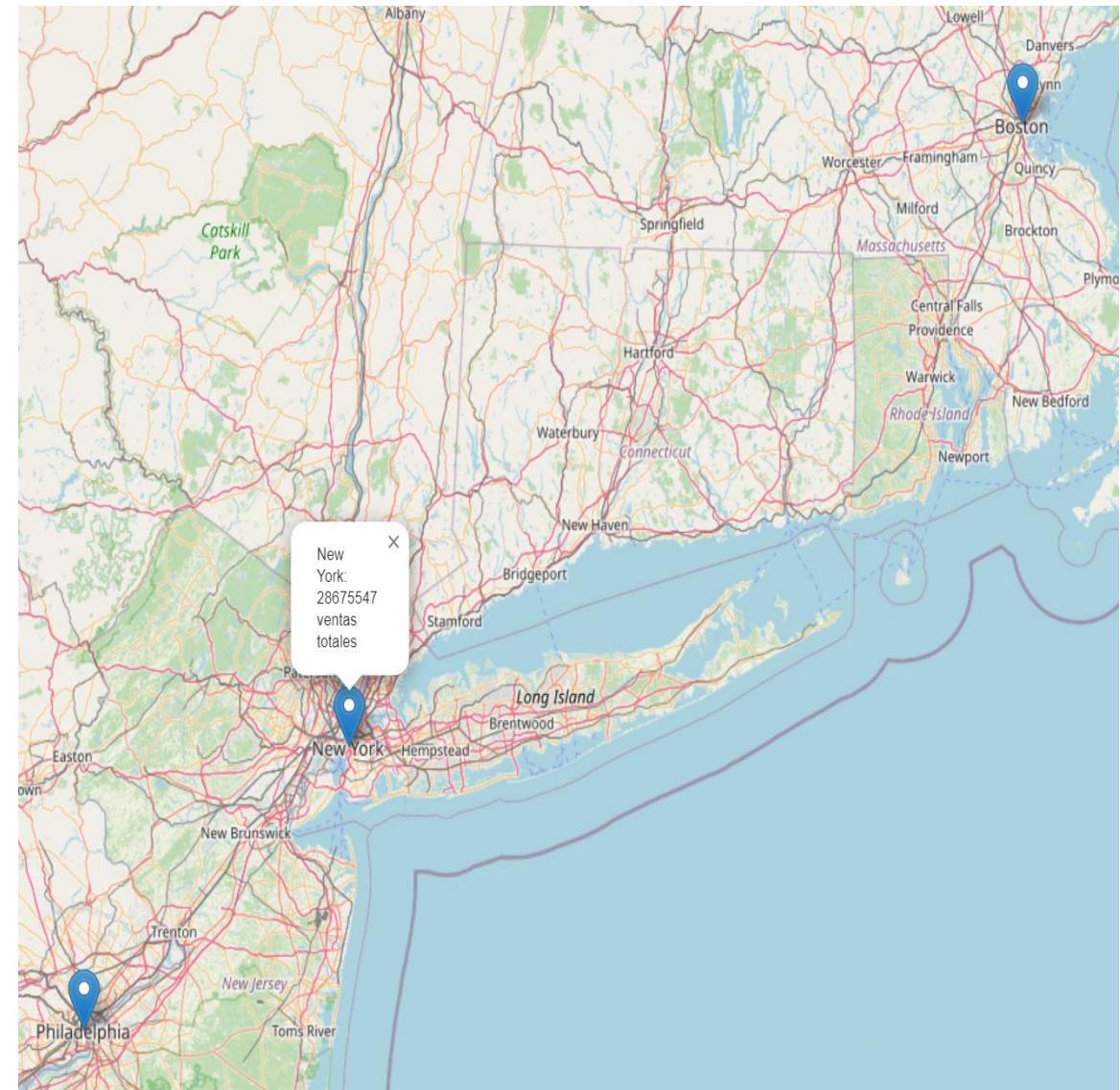
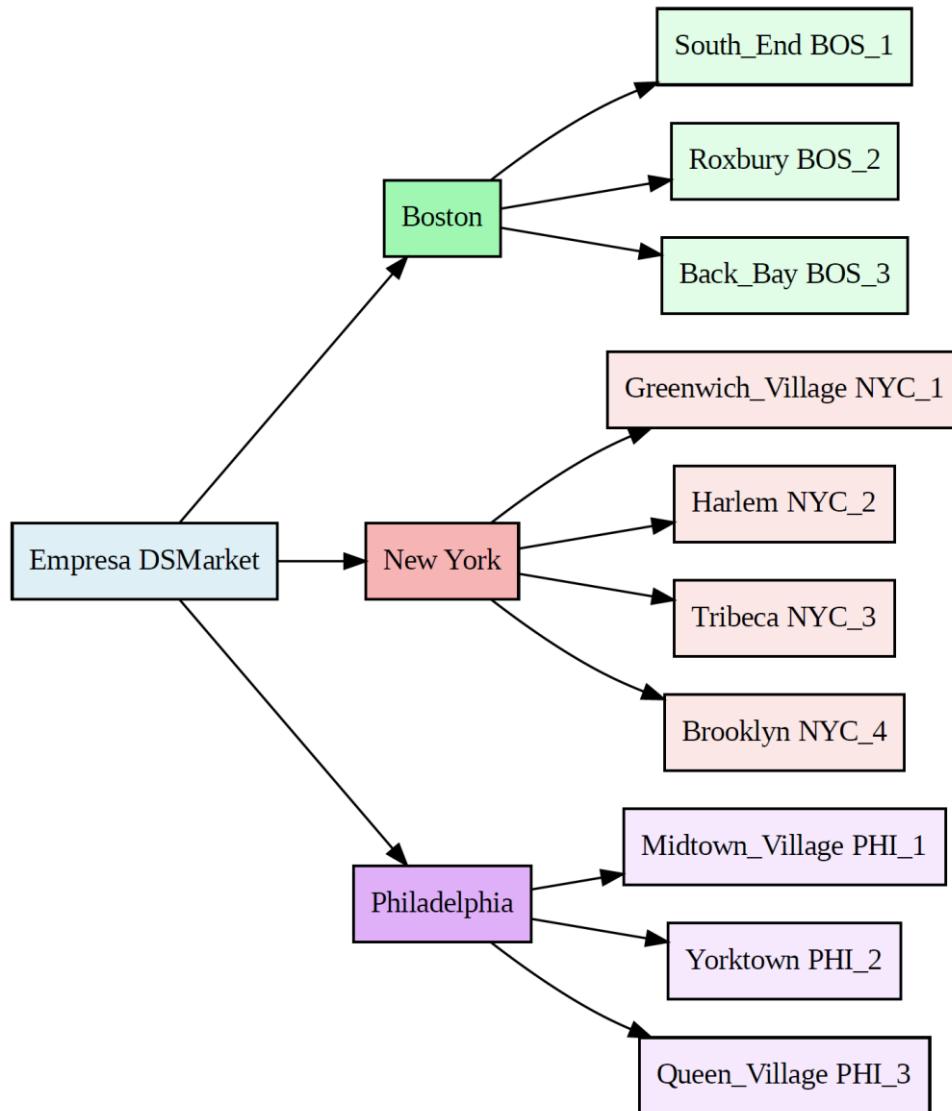
Contiene el precio de todos los productos. Hay un mismo número de artículos en los datos de precios, la cantidad de productos listados en el registro de precios coincide con la que tenemos en el historial de ventas, en este caso con 3049 valores únicos.

The screenshot shows three data frames from a Jupyter Notebook:

- pd_calendar.head()**: Displays the first 5 rows of a calendar dataset. The columns are date, weekday, weekday.int, d, and event. The data shows dates from January 29, 2011, to February 2, 2011, with corresponding weekdays and event codes.
- pd_prices.head()**: Displays the first 5 rows of a prices dataset. The columns are item, category, store_code, yearweek, and sell_price. The data shows multiple entries for 'ACCESORIES_1_001' across different categories and stores, with various sell prices.
- pd_sales.head()**: Displays the first 5 rows of a sales dataset. The columns include id, item, category, department, store, store_code, region, and three additional columns d_1, d_2, and d_3. The data shows purchases for 'ACCESORIES_1_001' across different stores and regions, with some values being 0.

Each data frame's head() method output includes a timestamp (e.g., 0.0s) and a green checkmark icon.

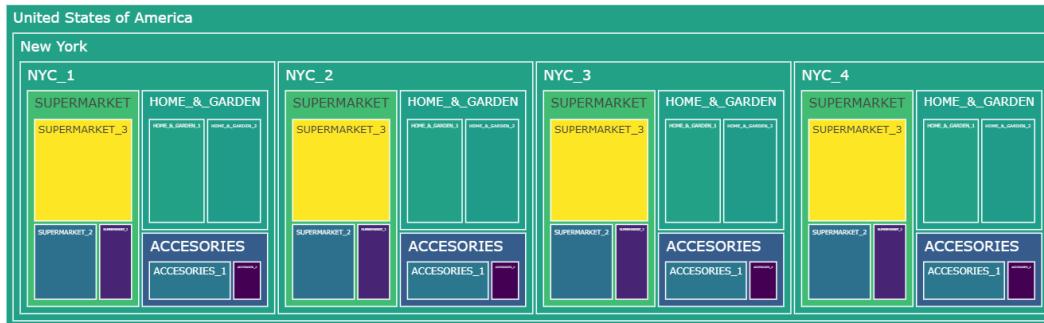
Estructura de la empresa



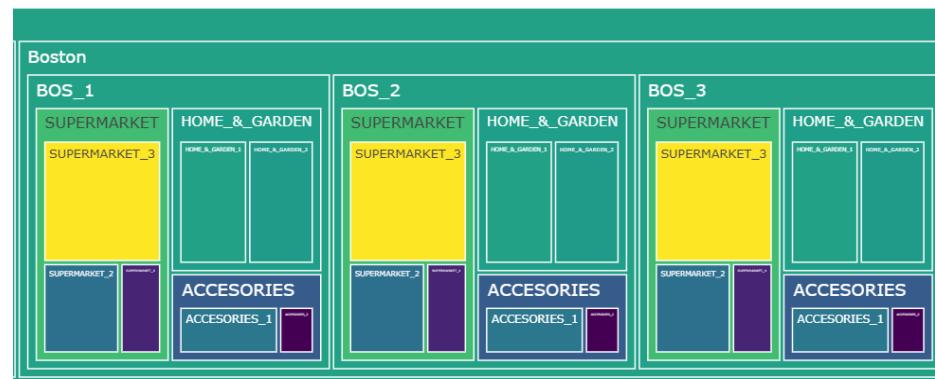


Distribución de Items

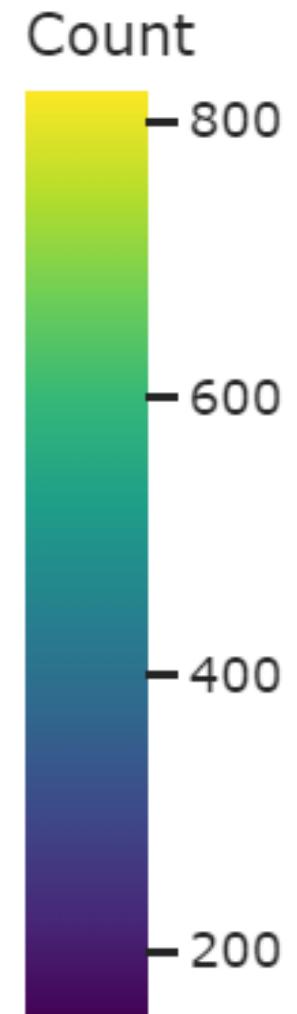
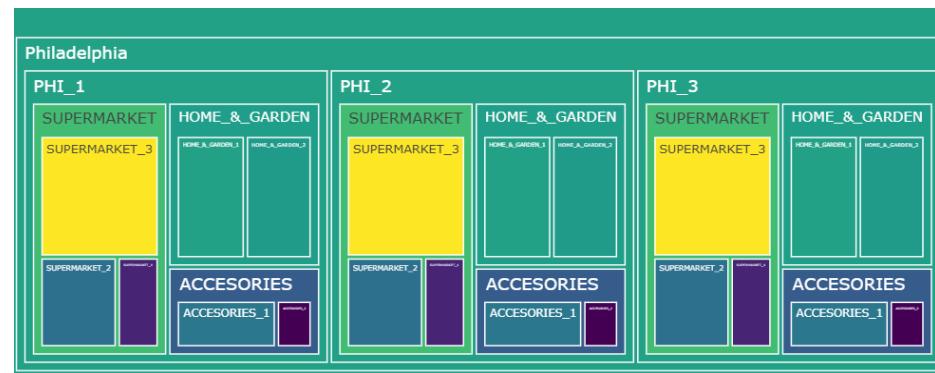
12196 ITEMS

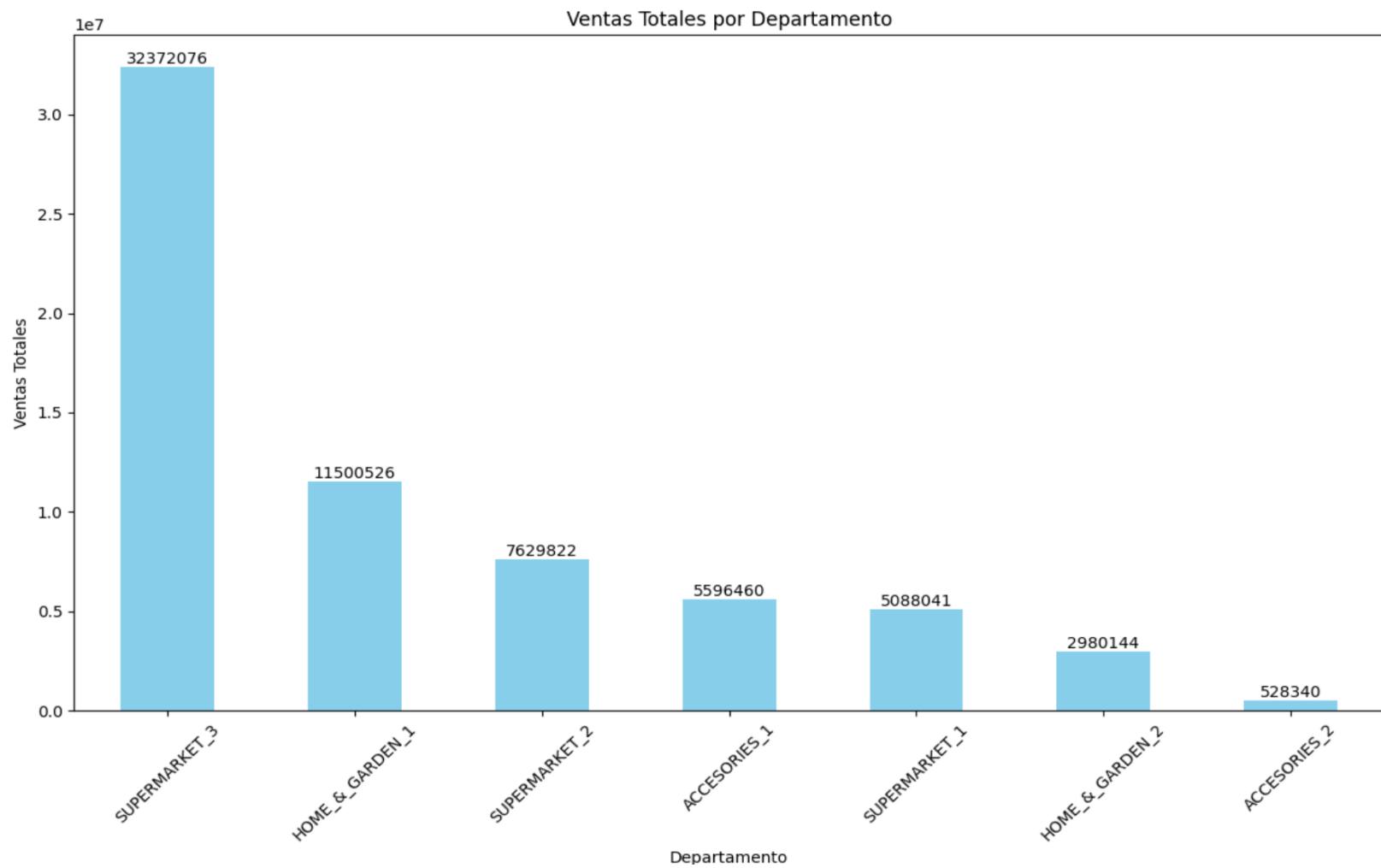


9147 ITEMS

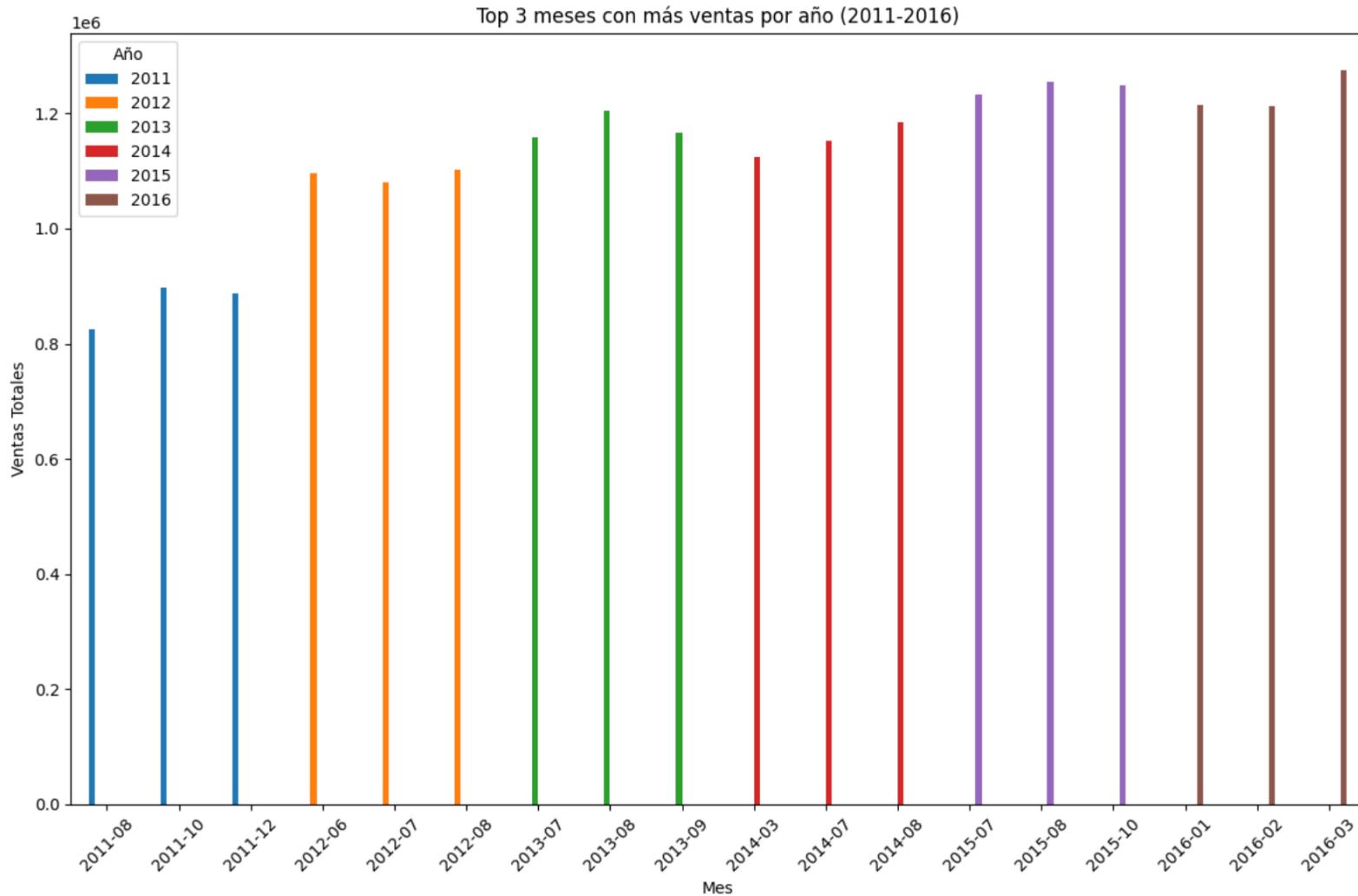


9147 ITEMS





Análisis



- Aquí podemos ver un índice con el que podemos interactuar pinchando en cada botón, para navegar a través de todo el informe creado para las necesidades que se nos han pedido en los correos electrónicos.
 - Análisis Comparativo de Ventas y Facturación por Tienda y Ciudad.
-
- Vamos a examinar el conjunto de datos ya que, son críticos para comprender el desempeño de nuestras tiendas a lo largo del tiempo.
 - La diapositiva se divide en dos paneles comparativos que reflejan la suma de ventas y la facturación total en las ciudades de Boston, Nueva York y Philadelphia.
-
- En cada comparativa, en la parte superior, separada por una línea podrás encontrar un botón con unos deslizantes referidos a la fecha con el que podrás interactuar con las fechas que van desde el 29 de enero de 2011. Hasta el 24 de abril de 2016, este botón irá apareciendo en algunas de las diapositivas de la presentación y tendrá siempre la misma función, la de poder acotar las búsquedas de información a través de fechas concretas de tiempo.
 - Este análisis comparativo sirve como una herramienta valiosa para evaluar el impacto de nuestras estrategias de venta a lo largo del tiempo y cómo las diferentes regiones contribuyen al éxito general de la compañía. Con estos datos, podemos profundizar en las ventas exactas y la facturación, así como mediante una gráfica lineal poder ver de una manera más visual el resultado de estas estrategias.



Primera comparativa

29/01/2011 24/04/2016

65.695.409

Suma de Ventas

\$230.981.146,22

Facturación

Tiendas

- Back_Bay
- Brooklyn
- Greenwich_Village
- Harlem
- Midtown_Village
- Queen_Village
- Roxbury
- South_End
- Tribeca
- Yorktown

Facturación por mes y año

● Boston ● New York ● Philadelphia



Segunda comparativa

29/01/2011 24/04/2016

65.695.409

Suma de Ventas

\$230.981.146,22

Facturación

Tiendas

- Back_Bay
- Brooklyn
- Greenwich_Village
- Harlem
- Midtown_Village
- Queen_Village
- Roxbury
- South_End
- Tribeca
- Yorktown

Facturación por mes y año

● Boston ● New York ● Philadelphia



CANTIDAD DE VENTAS Y FACTURACIÓN

Análisis de Datos por Tienda, Eventos Especiales y Categorías de Producto

La presentación a continuación se centra en el desempeño financiero de nuestras tiendas, examinando la facturación total y las ventas por categoría de producto durante un periodo específico, del 29 de enero de 2011 al 24 de abril de 2016. Se destaca la contribución de eventos especiales a nuestras ventas y cómo varían las ventas durante los fines de semana y días laborales.

Facturación Total por Ciudad: Nueva York lidera con una facturación impresionante de más de \$103 millones. Boston y Filadelfia siguen con aproximadamente \$66 y \$60 millones, respectivamente.

Facturación por Tienda: Tribeca resalta con una facturación que supera los \$39 millones, seguida de Greenwich y Roxbury, ambos con ventas sustanciales superiores a \$25 millones.

Número de Ventas por Tienda: La tienda de Tribeca no solo es líder en facturación, sino también en número de ventas, con más de 11 millones de transacciones. Greenwich Village y Roxbury muestran también un rendimiento fuerte en términos de volumen con 7 y 7.2 millones de ventas, respectivamente.

Facturación en Días de Eventos Especiales: Se evidencia el impacto de las festividades en nuestras ventas, siendo la Pascua la más rentable en las categorías de Accesorios y Hogar y Jardín. El Día de la Independencia es significativo para el Supermercado. El total combinado en estos días especiales asciende a más de \$3.5 millones.

Media de Facturación: Los fines de semana presentan una media superior de facturación excepcional a los \$146 mil, mientras que los días laborales mantienen una media sólida de más de \$111 mil. Esto puede deberse a que en los fines de semana las personas tienen más disponibilidad para ir a hacer las compras, por razones de trabajo.

Ventas por Categoría: En la distribución de ventas por categoría, el Supermercado domina claramente con un 68.63%, seguido por Hogar y Jardín con un 22.04%, y Accesorios con un 9.32%.

Este análisis detallado revela patrones importantes de consumo y preferencias de compra que se pueden aprovechar para optimizar estrategias de ventas y marketing. Por ejemplo, la alta facturación en eventos específicos podría impulsarnos a diseñar campañas promocionales dirigidas que coincidan con estas fechas. Además, la significativa contribución del Supermercado a nuestras ventas totales sugiere un enfoque en la expansión y promoción de esta categoría.





PRODUCTOS

- Los tres productos más vendidos que tenemos son de Supermarket_3_090, en segundo lugar, el Supermarket_3_586 y, en tercer lugar, el Supermarket_3_252.
- Esto sería pudiendo incluir todo el rango de fechas, pero podemos ajustarlo por ejemplo para ver los productos más vendidos en el último año.
- Además de eso, podemos mirar por una tienda en concreto y también por ciudad. Al ir variando y filtrando entre diferentes casos de fechas, tiendas y ciudades, podemos apreciar que los productos más vendidos se suelen mantener en el top más vendido, lo que nos puede sugerir lo demandados que son esos productos. También podemos pensar que son de primera necesidad.
- En la dashboard de abajo, podemos ver cada producto con rango de precio, y lo que se ha vendido con ese precio.



Tarea 2: Clustering

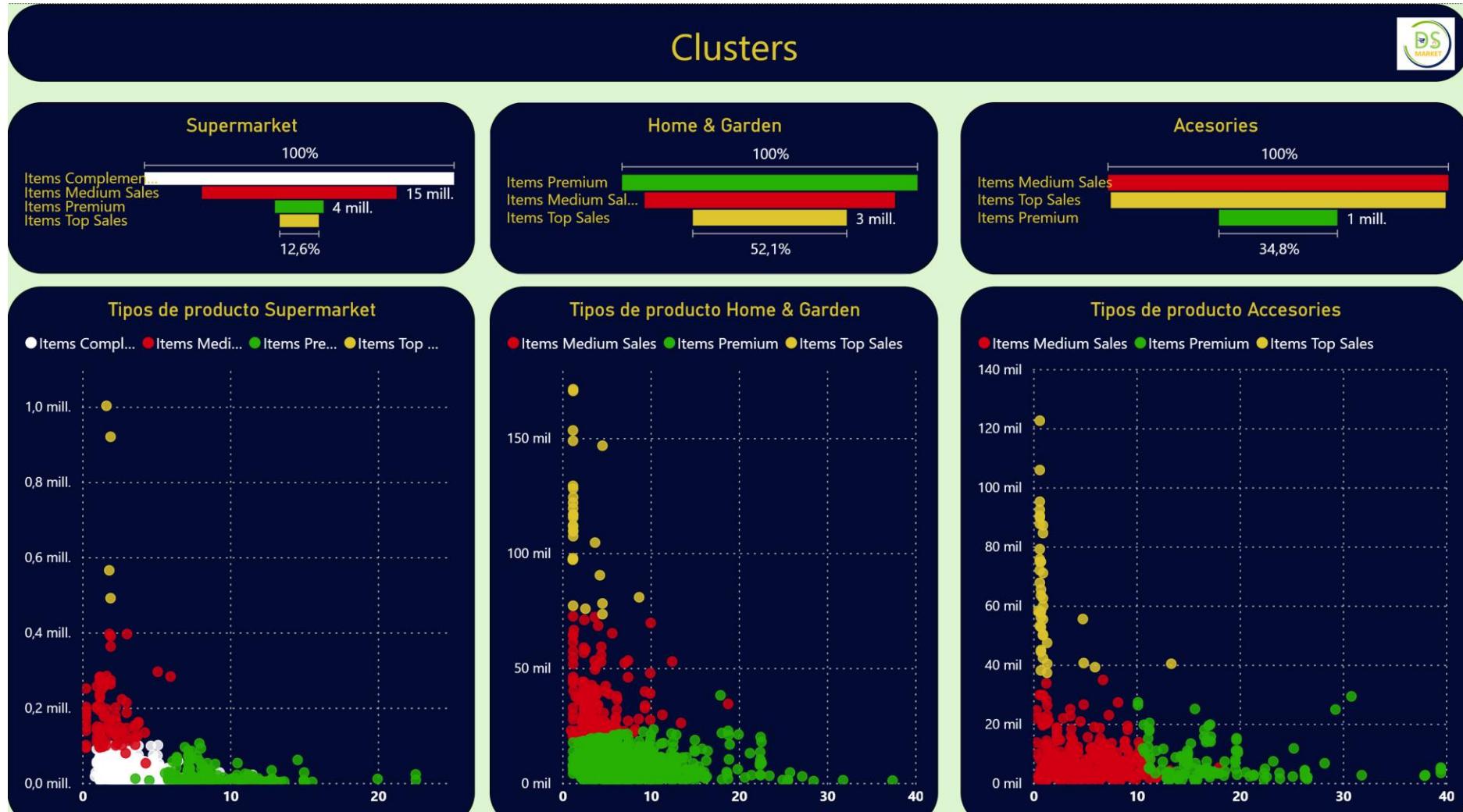
A continuación, vamos a ver los grupos que hemos identificado y segmentado. Estos constan de tres categorías: productos de supermercado, de casa y jardinería y accesorios.

Además de estos grupos, cada uno se divide en varios tipos de productos: Ítems Top Sales, Ítems Medium Sales y Ítems Premium. Para la categoría de Supermercado, hemos añadido un tipo adicional el cuál serían los ítems complementarios.

Las gráficas presentadas muestran el análisis de estos tres grupos de categoría. Cada gráfica consta de dos partes principales: una barra de distribución porcentual y un scatter plot (diagrama de dispersión) que muestra la dispersión de los tipos de productos basada en las ventas y el precio. El eje de las X representa el precio, y el eje Y, la cantidad de ventas.

- El grupo de productos de supermercado representan la mayoría de nuestro inventario, siendo la categoría más amplia. La mayoría de las ventas provienen de productos complementarios, seguidos por los ventas medianas y premium.
- El grupo de productos de casa y jardinería representa el segundo lugar de nuestro inventario. Muestra una predominancia en productos premium, seguido por las ventas medias.
- El grupo de accesorios, constituyen una porción pequeña de nuestro inventario, pero tienen un buen desempeño, especialmente los de ventas medianas, lo que sugiere una oportunidad para potenciar esta categoría.

Clusters





Tipos de Producto (información de Clusters)

- Los ítems de Top Sales se caracterizan por tener muchas ventas de estas, pero su precio es bajo.
- Los ítems Premium son los ítems que tienen un precio elevado, pero tienen menos ventas.
- Los ítems de Medium Sales son ítems que se han mantenido a nivel intermedio entre el precio y ventas.
- Los ítems Complementarios tienen un precio más elevado que el promedio, pero se han generado menos ventas, igualándose a los ítems Premium.

Tipología de los Clusters: Esta visualización detalla la tipología de los ítems en términos de precio medio, cuántos productos componen el tipo y el total de ingresos. Sería el precio medio y total de ingresos de todos los productos de un tipo, no de una categoría en concreto.

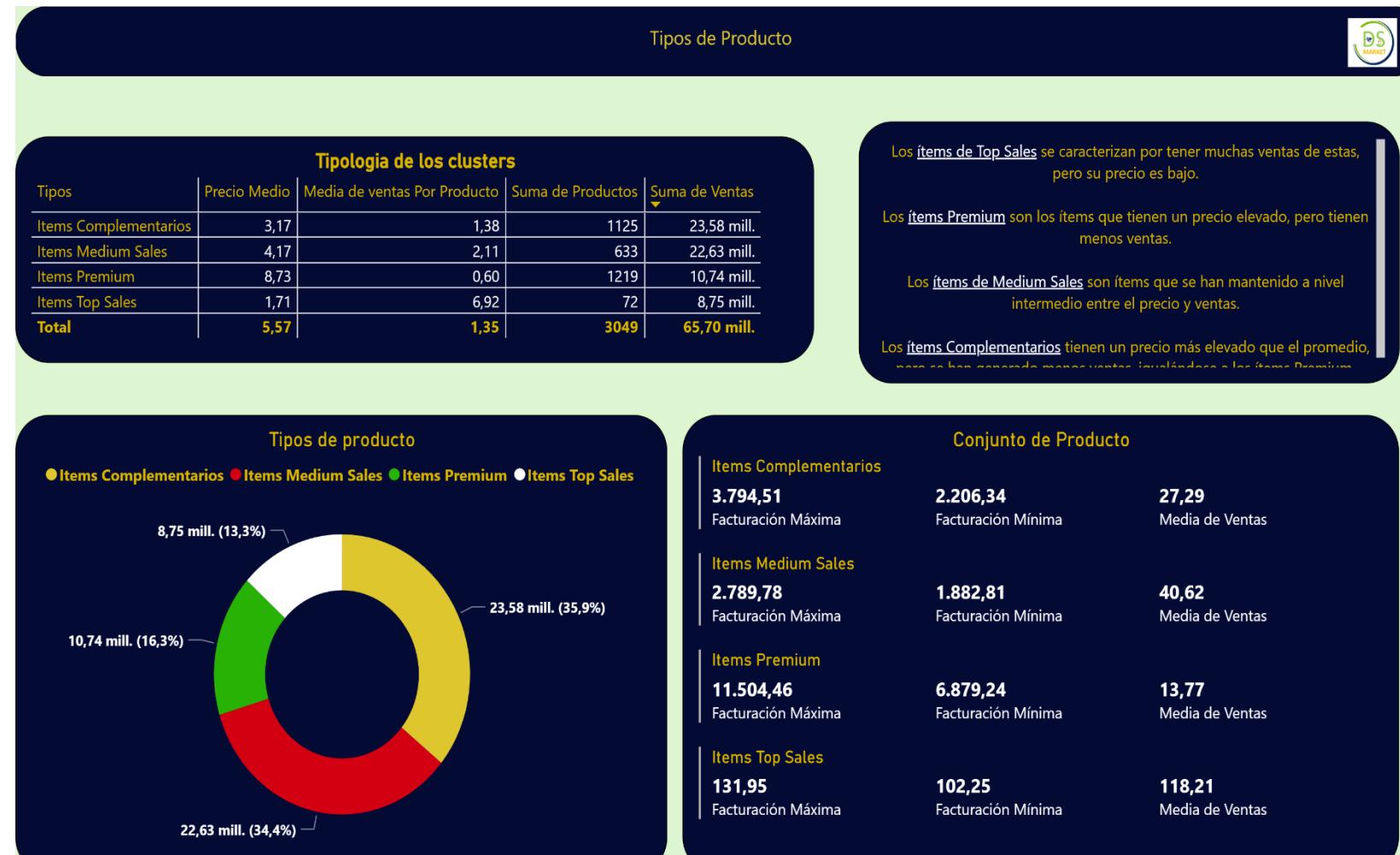
- Los Ítems Complementarios, con un precio medio de \$3.17, son los líderes en ventas, lo cual resalta su importancia en la estrategia de precios y promociones.
- Los Ítems de Medium Sales, se quedan en segundo lugar de ventas totales, con un precio medio de 4,17\$.
- Los Ítems Premium, aunque menos en cantidad con un precio medio de \$8.73, contribuyen significativamente a las ventas totales, lo que refleja un mercado dispuesto a invertir en productos de alto valor.
- Los Ítems Top Sales, tienen el menor volumen, pero con un precio medio similar a los complementarios, lo que indica una fuerte lealtad de los consumidores hacia ciertos productos específicos. Tener en cuenta que tenemos solamente 72 ítems en este grupo.

Visualización: Conjunto de Productos

Los Ítems Complementarios tienen una facturación máxima diaria de más de \$3794,51 y una media de ventas diarias que alcanza las 27,29 unidades al día

Por otro lado, los ítems Premium han alcanzado una facturación máxima impresionante de \$11504,46 con menos unidades vendidas, lo que enfatiza su alto valor y demanda. Los ítems Top Sales tienen ventas medias diarias que superan las del resto, con un 118,21 por producto, pero con la menor facturación que el resto.

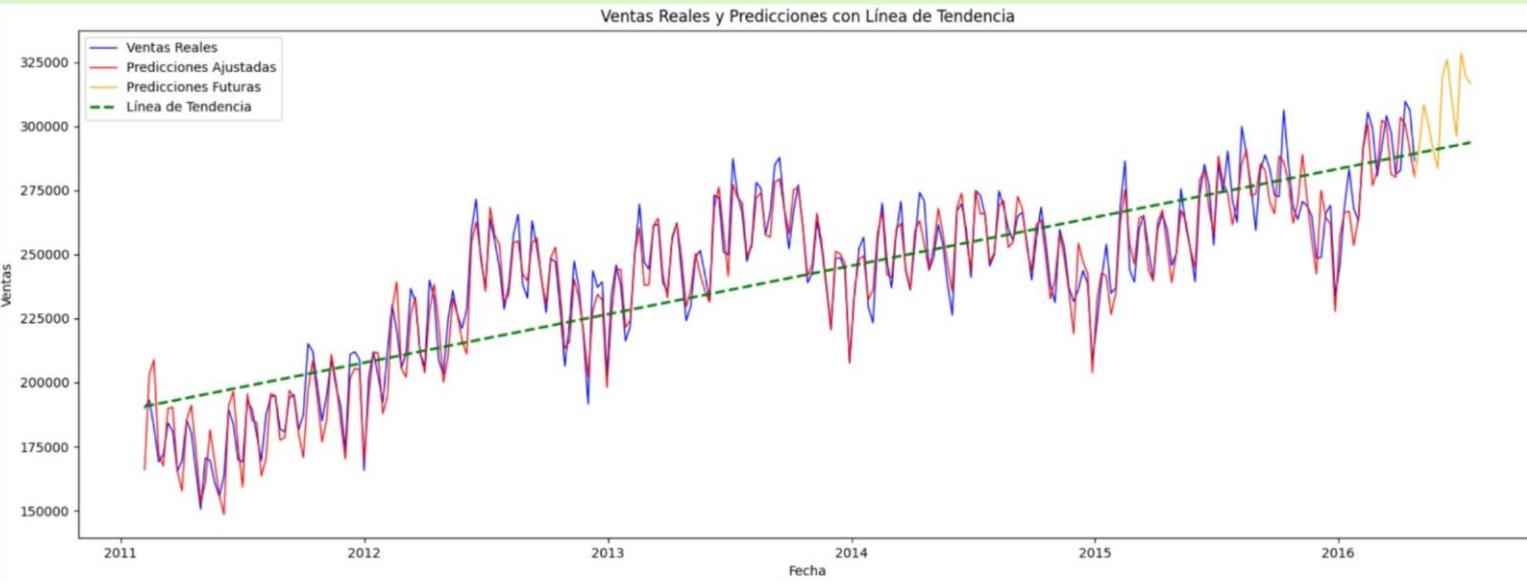
Los ítems de Medium Sales tienen una facturación máxima de \$2789,78 y una media de ventas de 40,62 por día, vendiéndose más que los Ítems Premium y Complementarios.



Tarea 3: Predicción

- Esta gráfica representa el rendimiento histórico de las ventas reales observadas cada semana (línea azul) comparadas con las estimaciones generadas por el modelo Holt Winter (línea roja). Y la línea amarilla representa las predicciones de las 12 semanas posteriores a las últimas ventas reales.
- La cercanía nos indica la precisión de las predicciones. Gracias a un modelo más preciso podemos optimizar el marketing e implementar mejores estrategias de marketing para prevenir la escasez o exceso de stock mediante una aplicación donde estarán implementados estos modelos.
- Esta aplicación funcionará mediante una API, la API actúa como un puente entre nuestra base de datos y la aplicación en las tiendas. Cada vez que se registran nuevas transacciones o cambios en el inventario, la API automáticamente actualiza las predicciones en tiempo real, asegurando que el personal correspondiente siempre tenga acceso a la información más reciente y relevante para tomar decisiones. Esto permite una respuesta rápida a las últimas tendencias.
- El RMSE del modelo representa solo el 3.29% del promedio de ventas semanales, indicando un error muy bajo y demostrando la alta precisión del modelo.

Predicción



Sin embargo continuaremos perfeccionando el modelo para maximizar su precisión y tener el menor riesgo posible de pérdidas, pobraremos con otros modelos como:

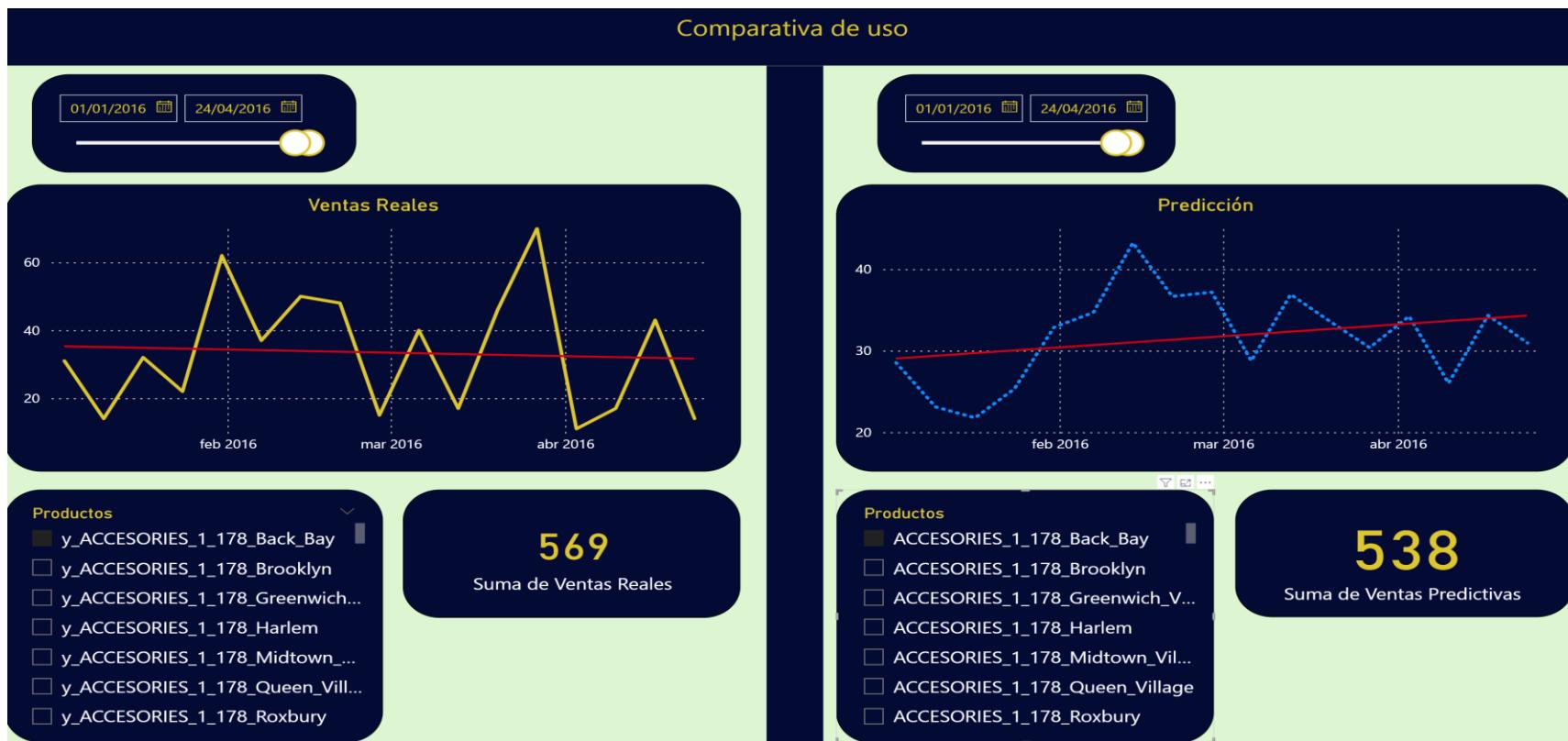
1. XGBoost
2. Red Neuronal Recurrente (LSTM) - Modelo de Deep Learning
3. CatBoost
4. Random Forest

Esta gráfica representa el rendimiento histórico de las ventas reales observadas cada semana y las estimaciones generadas por el modelo. Un modelo más preciso nos permite optimizar el marketing y la gestión de inventario mediante una aplicación, asegurando un abastecimiento óptimo para cumplir con la demanda y aprovechar oportunidades crecimiento.

El RMSE del modelo representa solo el 3.29% del promedio de ventas semanales, indicando un error muy bajo y demostrando la alta precisión del modelo. Este resultado se ha conseguido utilizando el algoritmo de suavizado exponencial llamado "Holt Winters", además, hemos empleado otros enfoques y modelos como Prophet. En el gráfico la línea amarilla corresponde a las predicciones de 12 semanas que finalizarían el '2016-07-17', las últimas ventas reales fueron el 24 de abril del 2016

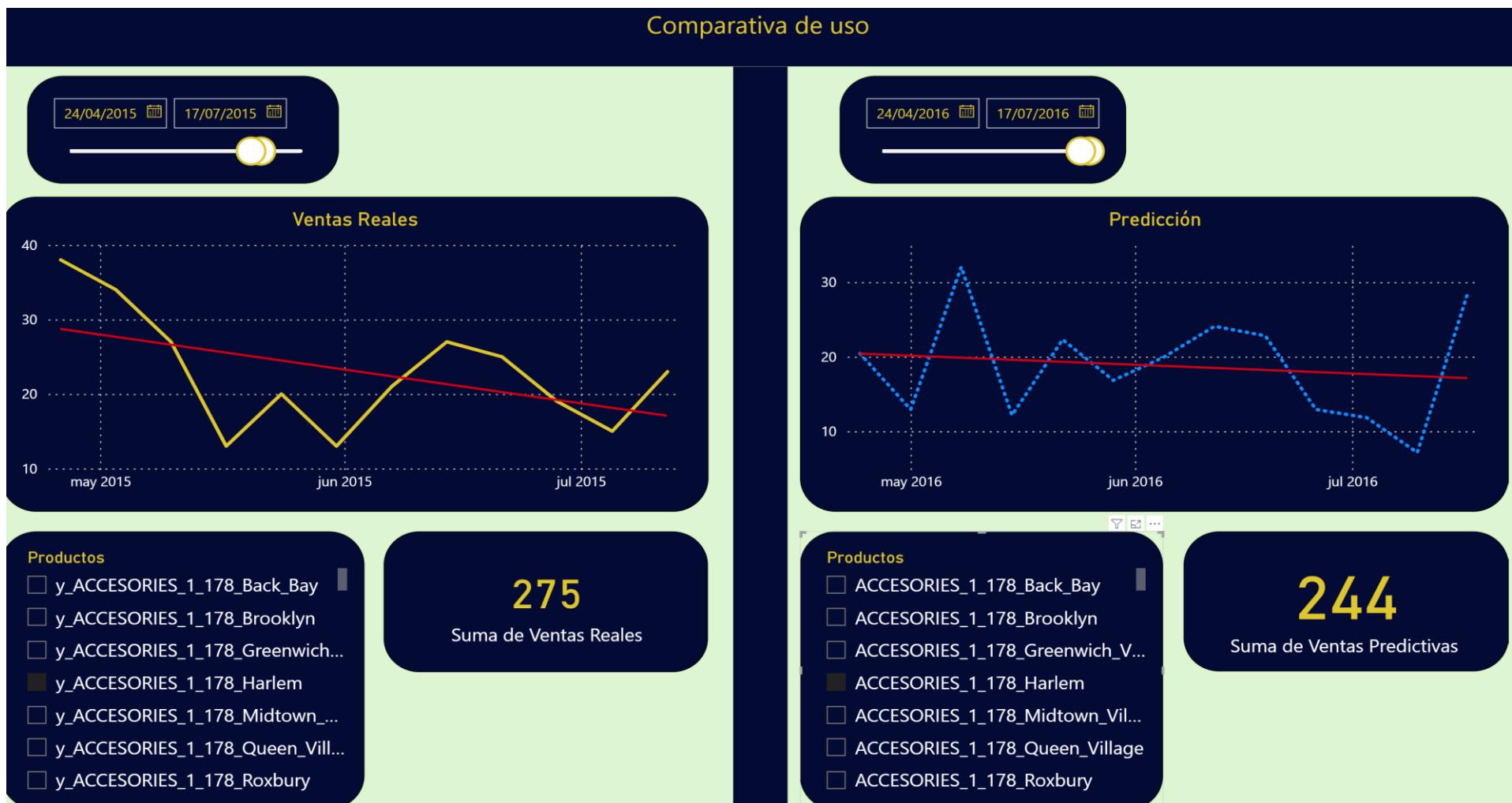
Comparativa de uso

Este enfoque Producto-Tienda nos permite comparar las predicciones de un modelo con las ventas reales de un producto en específico a lo largo del tiempo que filtramos. La cercanía de las ventas nos indica un modelo preciso, aunque podemos mejorarlo.





También ver las ventas predichas de las nuevas semanas a partir de las últimas ventas reales que es el 24/04/2016, comparando con las del año anterior.



Este enfoque es más general ya que vemos las ventas de todas las tiendas con sus predicciones, la misma dinámica que el anterior ejemplo.





store	rmse	media_ventas	rmse_percentage
Tribeca	1991.909943	40926.886447	4.87%
Greenwich_Village	1492.680541	28155.772894	5.30%
Brooklyn	891.543286	15008.978022	5.94%
South_End	1379.976059	20467.249084	6.74%
Midtown_Village	1564.428146	18825.699634	8.31%
Roxbury	2276.144648	26394.172161	8.62%
Queen_Village	2284.271618	23511.981685	9.72%
Back_Bay	2713.901844	22275.787546	12.18%
Yorktown	4500.276890	23932.164835	18.80%
Harlem	13894.452968	20782.084249	66.86%

Para evaluar los modelos hemos utilizado varios años para el train y los últimos 12 meses para el test/prueba, lo ideal es que el rmse siempre sea mucho menor a la media de ventas, y el porcentaje lo más pequeño posible ya que por ejemplo un rmse del aproximado del 20% para nuestro caso, serían pedidos superiores o insuficientes que nos repercutiría en pérdidas.

La tabla está ordenada desde el mejor modelo por lo que podemos ver que hay modelos por mejorar utilizando otros métodos. Respondiendo a Paul (el director financiero) este enfoque de ver las predicciones por Tienda o por producto-tienda seguiría siendo válido ya que este último nos da información más precisa.

Tarea 4: Tarea 4: Caso de uso de abastecimiento de tiendas.

MLOps

Ahora, daremos una propuesta para aplicar los modelos de ventas a un caso de uso de abastecimiento de las tiendas. Podríamos migrar los datos de ventas que se generen en un espacio como el de Google Cloud Storage para que almacenen y se traten de una manera más segura y controlada, disminuyendo la pérdida o modificación de estos datos.

Los datos nuevos se almacenan en Google Cloud Storage para su posterior elaboración.

Dataflow se usa para extraer, validar y transformar los datos crudos (ETL). Este proceso limpia los datos y los prepara para el análisis, asegurando que sean consistentes y estén en el formato adecuado.

Usando Google AI Platforms, se desarrolla y entrena el modelo de predicción de ventas.

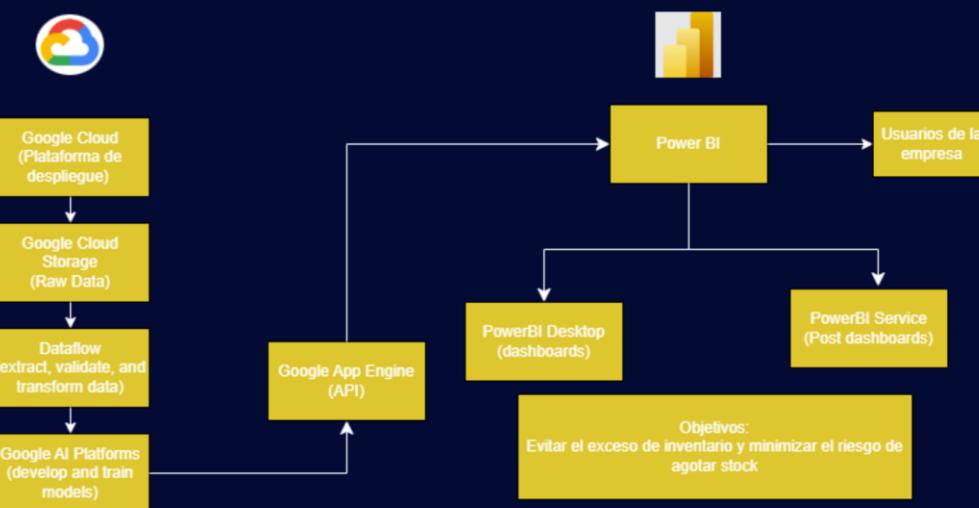
Los modelos entrenados y los insights generados se integran con Power BI a través de Google App Engine, que actúa como una API para servir las predicciones y facilitar la visualización de los datos. Crea archivos en formato csv con las predicciones, para que sean fácil de visualizar y utilizar en PowerBI.

PowerBI Desktop se utiliza para crear dashboards detallados que permiten visualizar las predicciones de las ventas. Los dashboards desarrollados en PowerBI Desktop se publican a través de PowerBI Service. Esto permite compartir las visualizaciones con los usuarios de la empresa en una plataforma accesible y segura. Para darnos de alta en la suscripción de PowerBI Service, necesitamos el correo electrónico de la empresa para crear la cuenta.

Los usuarios finales de la empresa acceden a los dashboards de Power BI para tomar decisiones informadas. Esto incluye evaluar las predicciones de ventas para cada tienda y optimizar los niveles de inventario. Pueden hacerlo a través de la aplicación móvil o portátil.

Google Cloud Scheduler es un servicio completamente administrado que permite programar la ejecución de tareas recurrentes. Con Google Pub/Sub, además con la ayuda de Cloud Scheduler, podemos programar que nos entre una notificación, en el momento en donde las predicciones estén listas, y enviadas al PowerBI.

MLOPS : Caso de Uso



Presupuesto

- Power BI Premium: 224,40 euros /año
- Google Cloud: 201,84 euros / año
- Total: 426,24 euros / año

CONCLUSIÓN

Respondiendo al reciente pedido recibido por email para disminuir el stock en los almacenes regionales, aplicar las conclusiones obtenidas a partir del modelo de datos puede optimizar considerablemente el ciclo completo, abarcando la adquisición de artículos de venta al detalle hasta la entrega de estos.

Esto nos llevaría a una gestión más eficaz de los activos, que es esencial para nuestra empresa dado que es imprescindible recortar gastos operativos con el fin de fortalecer nuestra posición en el mercado.