



# Memoria TFM DS Market

## Grupo 3



**Trabajo realizado por:**

Fernando Fracchia  
Cristian Alarcón  
Jesús González  
Helena Roig

**Tutor:**

Matías Hermida

**Master Data Scientist & AI 0224**

**Nuclio Digital School**

# Índice

Índice .....	2
Introducción .....	3
Tarea 1: Análisis .....	5
Visión Global: .....	5
Visión por Categoría: .....	7
Visión por Department: .....	8
Visión por Región: .....	10
Visión por Store: .....	10
Ítems sin venta: .....	12
Tarea 2: Clustering .....	13
Cluster 0: .....	15
Cluster 1: .....	15
Cluster 2: .....	16
Cluster 3: .....	16
Delete: .....	16
Tarea 3: Time Series .....	17
Estimación de la venta en Boston: .....	19
Estimación de la venta en Nueva York: .....	19
Estimación de la venta en Philadelphia: .....	20
Tarea 4: Caso de uso de abastecimiento de tiendas y API.....	22
Desarrollo de la API.....	23
Annexo 1: .....	26

## Introducción

DSMarket es una cadena de hipermercados en Estados Unidos que quiere empezar a usar sus datos para mejorar su toma de decisiones.

Se requiere de la incorporación de un Data Scientist, que reportando al Director Financiero, aporte las primeras conclusiones en previsión de ventas como principal prioridad, dentro de su estrategia de convertirse en una empresa Data driven.

Como fuente de datos se aportan las siguientes tablas con la siguiente información:

### FILE 1. daily\_calendar\_with\_events.csv

Name	Table	Description
date	calendar	date in y-m-d format
weekday	calendar	day of the week
weekday_int	calendar	numeric day of the week (Saturday day 1, Friday day 7)
d	calendar	day identifier
event	calendar	if the date includes an event, the name of this event (only a few are included)

### FILE 2. item\_prices.csv

Name	Table	Description
item	prices	product id
category	prices	product category
store_code	prices	alphanumeric code of the store
yearweek	prices	date period for the price (year-week format)
sell_price	prices	price for the product "item" for the period in "yearweek". Prices are provided per week (average across 7 days). If not available, there were no sales for the product during that week

### FILE 3. item\_sales.csv

Name	Table	Description
id	sales	sales series id (combination of item + store_code)
item	sales	product id
category	sales	product category
department	sales	department id (different identifier for different stores)
store	sales	store name
store_code	sales	store id
region	sales	region
d_1,d_2,d_...	sales	number of units sold per day

A partir de aquí, nosotros como Data Scientist pasaremos por las siguientes etapas/tareas para responder este caso práctico:

- **Tarea 1: Análisis**

En este apartado analizaremos los datos y la información de la que disponemos y haremos un business understanding de la empresa. Como también utilizaremos PBI para generar una serie de dashboards que permita a la empresa actualizar de forma regular los datos y monitorizar sus KPI's.

- **Tarea 2: Clustering**

Lanzaremos un modelo que nos permita entender si hay grupos de productos similares y a qué responden, con el fin de entender qué campañas de marketing pueden ser más efectivas.

- **Tarea 3: Predicción de ventas**

Lanzaremos un modelo de predicción que nos calcule la serie temporal de ventas de los productos.

- **Tarea 4: Caso de uso de abastecimiento de tiendas**

Se detalla el modelo que vamos a utilizar para responder al abastecimiento de las tiendas, además de la creación de una API para consultar información acerca de los stocks.

## Tarea 1: Análisis

Inicialmente recibimos tres tablas con información por separado.

- File 1. daily\_calendar\_with\_events.csv  
Esta tabla en formato .csv contiene información de fechas.
- File 2. item\_prices.csv  
Esta tabla en formato .csv contiene información referente a los precios que cada producto ha tenido en cada tienda por cada semana del año.
- File 3. item\_sales.csv  
Esta tabla contiene información de la cantidad de productos vendidos al día en cada tienda.

Como se puede observar, la primera tarea era crear una tabla que contenga toda la información. Es decir, una única tabla con la información de las ventas, precios y fechas. Finalmente nos queda una tabla de 58327370 registros.

Column	id	item_x	category_x	department	store	store_c	region	unique_id	value	date	week	d	even	year	item_store_yearweek	sell	revenue	Ubicación	
77019	SUPERMARKET_3_121_NYC_4	SUPERMARKET_3_121	SUPERMARKET	SUPERMARKET_3	Brooklyn	NYC_4	New York	23457	1	martes, 20 de octubre de 2013	Tuesday	4	d_1005	0	201343	SUPERMARKET_3_121_NYC_4_201343	\$2.58	2.576	United States, New York, Brooklyn
77194	SUPERMARKET_3_159_NYC_4	SUPERMARKET_3_159	SUPERMARKET	SUPERMARKET_3	Brooklyn	NYC_4	New York	23837	1	martes, 20 de octubre de 2013	Tuesday	4	d_1005	0	201343	SUPERMARKET_3_159_NYC_4_201343	\$2.58	2.576	United States, New York, Brooklyn
77258	SUPERMARKET_3_183_NYC_4	SUPERMARKET_3_183	SUPERMARKET	SUPERMARKET_3	Brooklyn	NYC_4	New York	24077	1	martes, 20 de octubre de 2013	Tuesday	4	d_1005	0	201343	SUPERMARKET_3_183_NYC_4_201343	\$2.58	2.576	United States, New York, Brooklyn
77819	SUPERMARKET_3_305_NYC_4	SUPERMARKET_3_305	SUPERMARKET	SUPERMARKET_3	Brooklyn	NYC_4	New York	25297	1	martes, 20 de octubre de 2013	Tuesday	4	d_1005	0	201343	SUPERMARKET_3_305_NYC_4_201343	\$2.58	2.576	United States, New York, Brooklyn
78169	SUPERMARKET_3_384_NYC_4	SUPERMARKET_3_384	SUPERMARKET	SUPERMARKET_3	Brooklyn	NYC_4	New York	26087	1	martes, 20 de octubre de 2013	Tuesday	4	d_1005	0	201343	SUPERMARKET_3_384_NYC_4_201343	\$2.58	2.576	United States, New York, Brooklyn
78288	SUPERMARKET_3_411_NYC_4	SUPERMARKET_3_411	SUPERMARKET	SUPERMARKET_3	Brooklyn	NYC_4	New York	26357	1	martes, 20 de octubre de 2013	Tuesday	4	d_1005	0	201343	SUPERMARKET_3_411_NYC_4_201343	\$2.58	2.576	United States, New York, Brooklyn
78332	SUPERMARKET_3_423_NYC_4	SUPERMARKET_3_423	SUPERMARKET	SUPERMARKET_3	Brooklyn	NYC_4	New York	26477	1	martes, 20 de octubre de 2013	Tuesday	4	d_1005	0	201343	SUPERMARKET_3_423_NYC_4_201343	\$2.58	2.576	United States, New York, Brooklyn
78809	SUPERMARKET_3_559_NYC_4	SUPERMARKET_3_559	SUPERMARKET	SUPERMARKET_3	Brooklyn	NYC_4	New York	27837	1	martes, 20 de octubre de 2013	Tuesday	4	d_1005	0	201343	SUPERMARKET_3_559_NYC_4_201343	\$2.58	2.576	United States, New York, Brooklyn
79085	SUPERMARKET_3_596_NYC_4	SUPERMARKET_3_596	SUPERMARKET	SUPERMARKET_3	Brooklyn	NYC_4	New York	28207	1	martes, 20 de octubre de 2013	Tuesday	4	d_1005	0	201343	SUPERMARKET_3_596_NYC_4_201343	\$2.58	2.576	United States, New York, Brooklyn
79159	SUPERMARKET_3_616_NYC_4	SUPERMARKET_3_616	SUPERMARKET	SUPERMARKET_3	Brooklyn	NYC_4	New York	28407	1	martes, 20 de octubre de 2013	Tuesday	4	d_1005	0	201343	SUPERMARKET_3_616_NYC_4_201343	\$2.58	2.576	United States, New York, Brooklyn
79183	SUPERMARKET_3_623_NYC_4	SUPERMARKET_3_623	SUPERMARKET	SUPERMARKET_3	Brooklyn	NYC_4	New York	28477	1	martes, 20 de octubre de 2013	Tuesday	4	d_1005	0	201343	SUPERMARKET_3_623_NYC_4_201343	\$2.58	2.576	United States, New York, Brooklyn
79241	SUPERMARKET_3_637_NYC_4	SUPERMARKET_3_637	SUPERMARKET	SUPERMARKET_3	Brooklyn	NYC_4	New York	28617	1	martes, 20 de octubre de 2013	Tuesday	4	d_1005	0	201343	SUPERMARKET_3_637_NYC_4_201343	\$2.58	2.576	United States, New York, Brooklyn
79530	SUPERMARKET_3_702_NYC_4	SUPERMARKET_3_702	SUPERMARKET	SUPERMARKET_3	Brooklyn	NYC_4	New York	29257	1	martes, 20 de octubre de 2013	Tuesday	4	d_1005	0	201343	SUPERMARKET_3_702_NYC_4_201343	\$2.58	2.576	United States, New York, Brooklyn

Peso del fichero csv: 3,41 GB

El siguiente paso será exportar estos datos como un fichero .csv y los cargamos en PowerBI para iniciar nuestro análisis de los datos y el business understanding.

Tras cambiar el formato de algunas de las columnas, como también del precio organizamos el PowerBI de la siguiente manera para sacar insights:

1. Visión Global.
2. Visión por Categoría.
3. Visión por Departamento.
4. Visión por Región.
5. Visión por Store.
6. Conclusiones.

### Visión Global:

Hemos generado un primer dashboard con todos los datos que nos permita entender los datos de negocio desde una perspectiva global. Para los primeros KPIs respecto a las tendencias de ventas de la empresa, hemos tomado como punto de referencia 2015 y hacer comparación con el año anterior para extraer los siguientes insights.

- Como KPIs tenemos Ventas con 52,32 M\$, y su comparativa respecto al año anterior con 46,58 M\$ (+12,33%).
- Como KPIs tenemos Items Vendidos con 13,8 M\$, y su comparativa respecto al año anterior con 13,09 M\$ (+5,43%).
- Como KPIs tenemos Promedio de la Venta con 3,79 \$, y su comparativa respecto al año anterior con 3,56 \$ (+6,55%).
- Como KPIs tenemos Venta por Día con 143.349 \$, y su comparativa respecto al año anterior con 127,611 M\$ (+12,33%).



Estos KPIs son interesantes para la empresa con el fin de poder monitorizar y entender la evolución a futuro.

Como se puede leer de los KPIs hay una tendencia positiva que se ha mostrado en una serie de gráficos.

Con un gráfico de líneas podemos observar la evolución de ventas, ítems vendidos y precio promedio. Este gráfico ha sido configurado de forma que se puede observar la comparativa con el mismo mes del año anterior con el fin de comparar la tendencia anual. Del mismo, podemos observar cómo Julio y Agosto son meses en los que se ve un mayor porcentaje de ventas, que se ve explicado por una mayor venta de ítems, si bien el precio promedio en esos meses es precisamente inferior. En cambio, se ve un aumento del precio promedio a principios y a finales de año, cuando las ventas son relativamente inferiores, sobre todo a principios de año.

Seguidamente hemos plasmado estos mismos conceptos en gráficos de barras agrupados por trimestre, donde podemos corroborar estas conclusiones.



A continuación, pasamos a hacer zoom en el comportamiento de las ventas por cada categoría.

## Visión por Categoría:

Tras volcar los datos en el PBI y filtrando por categoría, nos damos cuenta que existen 3 categorías diferentes:

- Accessories
- Home & Garden
- Supermarket

Esta slide del PBI contiene un gráfico de barras para ver la evolución de las ventas y los ítems vendidos por año, al que se le ha incorporado una línea de regresión, claramente con tendencia positiva año a año.

También podemos encontrar un gráfico de donut para entender el porcentaje que pesa cada categoría en el total de ventas.

Se incorporan también dos cuadros para poder acceder al dato de total de ventas por año y también por día de la semana.

Tal y como veníamos diciendo la tendencia de cada año es positiva por lo que vemos una evolución del total de ventas que incrementa cada año. Y en el caso de los días de la semana, vemos claramente una mayor venta en fines de semana y en segundo lugar sus días colindantes, es decir los viernes y los lunes.

Por último, hemos añadido un gráfico de barras que mide el porcentaje de cada categoría por cada región. En él, podemos observar cómo Supermarket es la categoría principal, seguido de Home & Garden y por último, Accessories.

En la parte superior de la slide se incorporan unos selectores para filtrar por: año, mes y categoría.

Si hacemos el foco por Categoría:

- **Supermarket** representa el 56% de la venta, con un precio promedio de 3,11\$ (+5,6% respecto al año anterior), una venta diaria de 76.917\$ y 9M de ítems vendidos.



- **Home&Garden** representa un 30% de la venta, con un precio promedio de 4,83\$ (en tendencia negativa respecto al año anterior -0,55%), una venta diaria de 45 k\$ y 3,4 M de ítems vendido.



- En cambio **Accessories**, es la categoría con un precio promedio superior de 5,64\$, pero con una venta diaria inferior de 21.546\$, debido a las pocas unidades vendidas 1,39 M.



## Visión por Departamento:

Siguiendo el mismo ejemplo que en la slide por categorías, hemos bajado un nivel más dentro de cada una y ahora encontramos los siguientes departamentos:

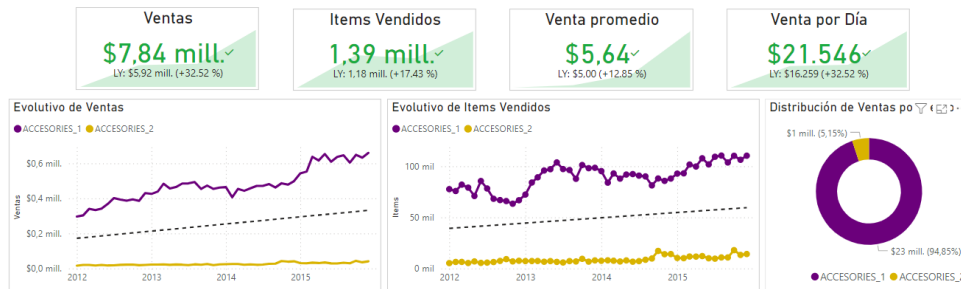
- Accessories 1.
- Accessories 2.
- Home & Garden 1.
- Home & Garden 2.
- Supermarket 1.
- Supermarket 2.
- Supermarket 3.

Dentro de la categoría Accessories, podemos observar como el 94% viene dado por Accessories 1.

Respecto a estos dos departamentos, Accessories 1 representa la mayor venta en unidades y en precio de venta, motivo por el cual encontramos tanta diferencia. Si analizamos en Accessories 2 vemos como el precio de venta ha ido a la baja

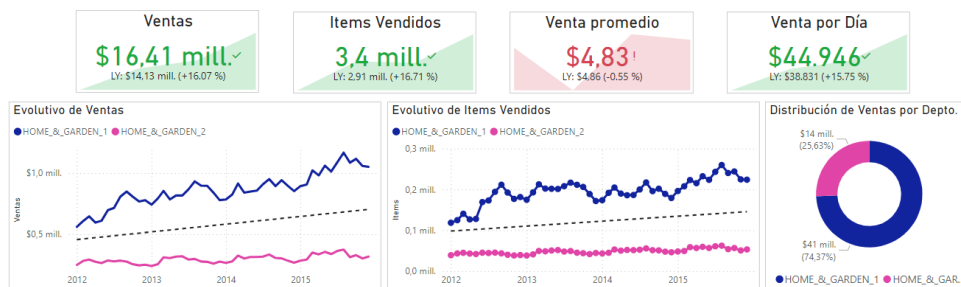


respecto al año anterior, pero ha tenido un fuerte aumento de ítems vendidos (+23,5 % respecto año anterior).

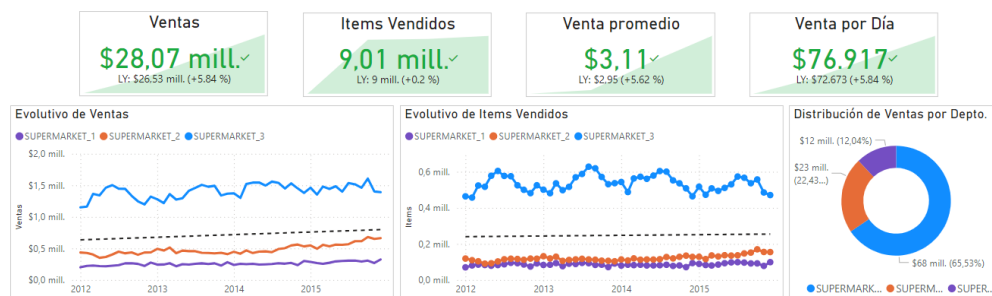


Aplicando el filtro de categoría Home & Garden, vemos como Home & Garden 1 representa el 74,4% de las ventas. Con un precio promedio de 4,55 \$, y una tendencia de ventas importante (18% más que el año anterior).

En cambio, Home & Garden 2 representa un 25,6% de las ventas con un comportamiento altamente estacional dentro del año, si bien con un ticket promedio similar, tiene muchos menos ítems vendidos.



Por último, respecto a la categoría de Supermarket, tenemos Supermarket 3 con un 66% de las ventas, a pesar de que el número de ítems vendidos, 6,22 millones, ha sufrido un descenso del 5% en el último año. Por otro lado, el revenue de este departamento se ha mantenido prácticamente igual gracias a un incremento de casi un 5% del precio de sus productos. El departamento Supermarket 2, representando un 22% de las ventas, tiene un incremento del 22% en las ventas llegando a 7,04 millones de \$ de revenue y un aumento del 18% de las ventas, todo esto solo con un incremento del 3,5% en el precio. Finalmente, Supermarket 1 representa solo un 12% de las ventas, se aprecia un aumento del 13% en el revenue y del 10% en las ventas, con solo un aumento del 2,5% del precio.



## Visión por Región:

En este apartado vemos los datos desglosados por las siguientes 3 regiones: New York, Boston y Philadelphia.

- New York: representa el 44,9% de la venta.
- Boston: representa el 28,7% de la venta.
- Philadelphia: representa el 26,4% de la venta.

Seguimos observando una parecida estacionalidad donde los meses de mayor venta siguen siendo en verano, este comportamiento es fácilmente reconocible en New York y Boston. En cambio, no podemos decir lo mismo de Philadelphia, su comportamiento en ventas no deja vislumbrar ninguna estacionalidad.

También se ha incluido un mapa donde se puede sacar el detalle de cada una de las regiones.



## Visión por Store:

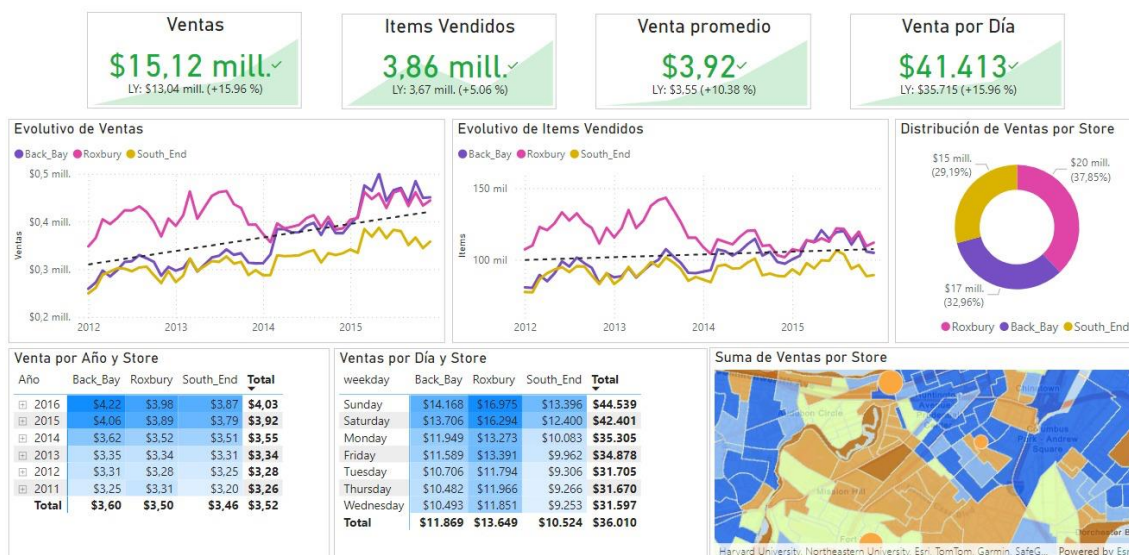
En esta sección podemos analizar el comportamiento de las tiendas de cada región. Se han incluido selectores en formato botón en la parte superior de la slide para seleccionar las regiones.

En Boston, encontramos 3 stores con porcentaje de ventas muy parecido:

- Roxbury con un 37,9%.
- Back bay con un 33,9%.
- South End con un 29,2%.

La evolución de las tiendas tiene una tendencia positiva. Principalmente gracias a Back Bay y debido a un incremento de precio de los ítems, puesto que la evolución de unidades vendidas se mantiene bastante constante.

En cambio, en Roxbury, si bien tiene una tendencia de venta positiva, tiene una tendencia de unidades vendidas muy a la baja. Lo que nos lleva a pensar que han compensado la falta de ventas con incrementos de precio (+10,6% respecto al año anterior).

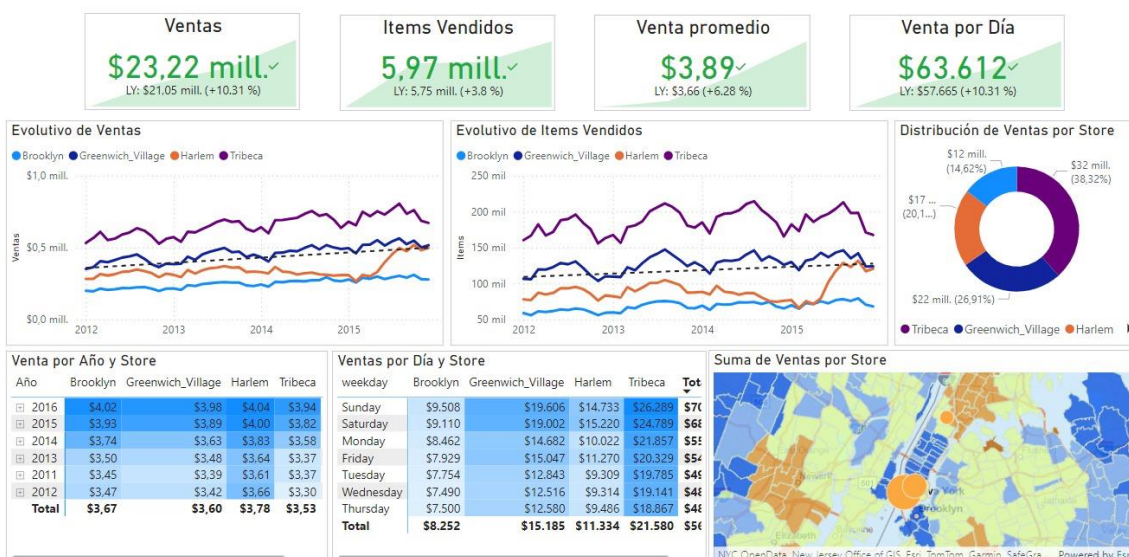


En New York encontramos 4 stores con la siguiente distribución de ventas:

- Tribeca con un 38,3%.
- Greenwich Village con un 26,9%.
- Harlem con un 20,1%.
- Brooklyn con un 14,6%.

Observamos una tendencia menos significativa de crecimiento, pero muy estable.

En cambio, vemos en Harlem una bajada importante de ventas en 2015.



Por último, en Philadelphia tenemos 3 stores con la siguiente distribución de ventas:

- Yorktown con un 36,7%
- Queen Village con un 32,7%
- Midtown con un 30,6%

La evolución de las ventas tiene una tendencia estable y positiva en Yorktown y en Midtown. No observamos lo mismo en Queen Village, éste tiene una tendencia claramente negativa.



## Ítems sin venta:

Con nuestra base de datos trabajada y tras el manejo de nulos (ítems sin venta), exportamos una lista de los ítems no vendidos en los últimos 28 días, para que el departamento de marketing o ventas pueda evaluar y decidir eliminar de los stocks:

ACCESORIES\_1\_335', 'ACCESORIES\_2\_110', 'HOME\_&\_GARDEN\_1\_209',  
 'HOME\_&GARDEN\_1\_366', 'HOME&\_GARDEN\_2\_158',  
 'HOME\_&GARDEN\_2\_202', 'HOME&\_GARDEN\_2\_210',  
 'HOME\_&GARDEN\_2\_456', 'HOME&\_GARDEN\_2\_502', 'SUPERMARKET\_1\_004',  
 'SUPERMARKET\_1\_043', 'SUPERMARKET\_1\_120', 'SUPERMARKET\_1\_126',  
 'SUPERMARKET\_2\_292', 'SUPERMARKET\_3\_002', 'SUPERMARKET\_3\_008',  
 'SUPERMARKET\_3\_073', 'SUPERMARKET\_3\_077', 'SUPERMARKET\_3\_205',  
 'SUPERMARKET\_3\_210', 'SUPERMARKET\_3\_271', 'SUPERMARKET\_3\_419',  
 'SUPERMARKET\_3\_441', 'SUPERMARKET\_3\_444', 'SUPERMARKET\_3\_647'

Seguiremos nuestro trabajo de Clustering y de Time Series eliminando estos ítems de nuestro data frame.



## Tarea 2: Clustering

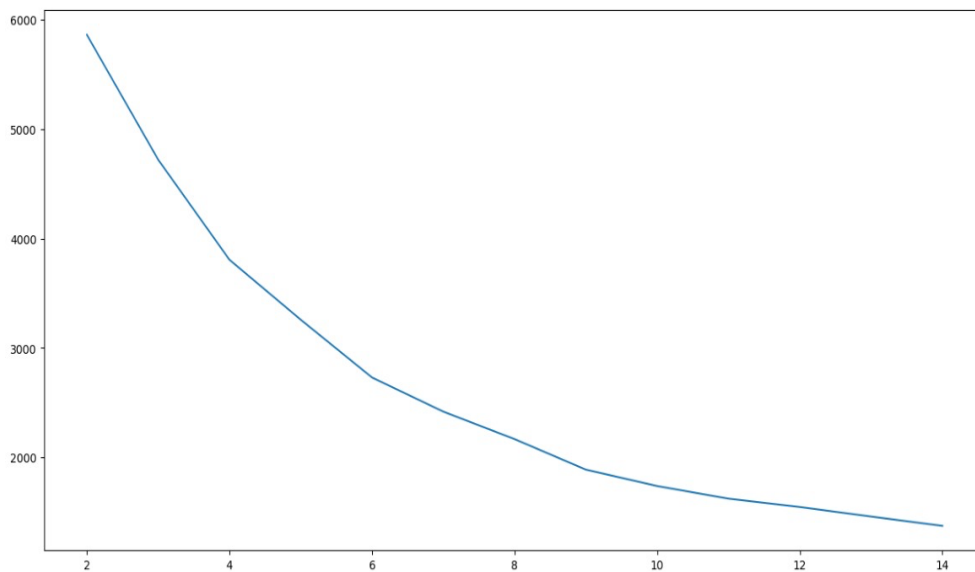
Tras hacer el análisis de negocio, la empresa nos solicita que identifiquemos grupos de productos que se comporten de manera similar, para así evaluar esos ítems en las diferentes campañas de Marketing.

Para ello vamos a hacer nuestro cluster por ítem, ya que tienen un comportamiento bastante similar entre tienda y tienda. Hemos trabajado un poco la base de datos para quedarnos, de cada ítem, con la información más relevante mediante la creación de nuevas variables como por ejemplo el mínimo, el máximo, la media y la mediana del precio, de las ventas y de la facturación. De la misma manera, también se han creado esas medidas para fines de semana y para fechas coincidentes con eventos especiales.

```
Data columns (total 31 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Unnamed: 0                            3025 non-null   int64
1   id                                     3025 non-null   object
2   item                                  3025 non-null   object
3   category                             3025 non-null   object
4   store_code                           3025 non-null   object
5   region                               3025 non-null   object
6   department                           3025 non-null   int64
7   max_price                            3025 non-null   float64
8   min_price                            3025 non-null   float64
9   mean_price                           3025 non-null   float64
10  std_price                             3025 non-null   float64
11  total_sales                           3025 non-null   int64
12  max_sales                             3025 non-null   int64
13  min_sales                             3025 non-null   int64
14  mean_sales                           3025 non-null   float64
15  std_sales                             3025 non-null   float64
16  total_revenue                         3025 non-null   float64
17  max_revenue                           3025 non-null   float64
18  min_revenue                           3025 non-null   float64
19  mean_revenue                          3025 non-null   float64
20  std_revenue                           3025 non-null   float64
21  total_sales_holiday                  3025 non-null   int64
22  max_sales_holiday                    3025 non-null   int64
23  min_sales_holiday                    3025 non-null   int64
24  mean_sales_holiday                   3025 non-null   float64
25  std_sales_holiday                    3025 non-null   float64
26  total_sales_weekend                   3025 non-null   int64
27  max_sales_weekend                     3025 non-null   int64
28  min_sales_weekend                     3025 non-null   int64
29  mean_sales_weekend                   3025 non-null   float64
30  std_sales_weekend                     3025 non-null   float64
```

Tras varias iteraciones y pruebas hemos lanzado el modelo KMeans para la estimación de hasta 14 clusters. Lo cual nos ha dado una variación de la dispersión siguiente:

Variación de la dispersión de los clústers en función de la k



Tras inicialmente analizar el resultado para 8 clusters, hemos observado que la mejor estimación la encontrábamos con 4 clusters. Puesto que al representar gráficamente el comportamiento del precio y las ventas, encontrábamos grupos próximos al comportamiento de los 4 clusters.

Así pues, hemos creado un listado de los ítems por cada cluster y a continuación, procedemos a analizar la segmentación en PowerBI.

Hemos añadido una columna extra dentro de los cluster donde vamos a calificar una problemática que nos comentaba la empresa que se correspondía con aquellos productos que no se estaban vendiendo bien. Esta columna será calificada como Delete e incluye esos ítems que mencionamos antes de eliminar.

Comparativa de los KPIs por cada cluster:

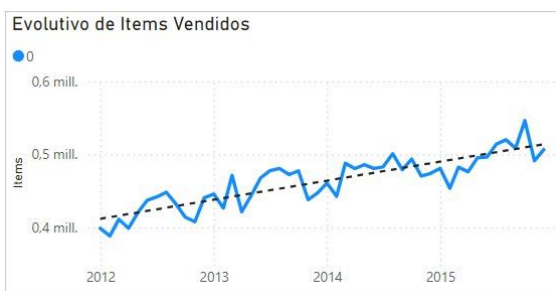
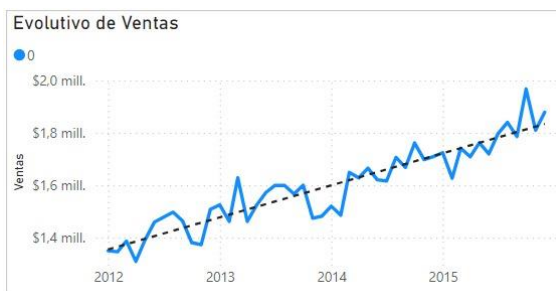
	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Delete	Total
Nº ítems	1346	84	561	1033	25	<b>3024</b>
Precio medio	3,78 \$	1,89 \$	7,4 \$	6,81 \$	4,49 \$	<b>5,56 \$</b>
Items vendidos	28,25 M	17,34 M	5,97 M	13,44 M	0,7 M	<b>65,70 M</b>
Revenue total	97,53 M\$	32,5 M\$	30,2 M\$	67,8 M\$	3 M\$	<b>231,0 M\$</b>
Revenue al día	51 k\$	17 k\$	16 k\$	36 k\$	1,61 k\$	<b>121 k\$</b>

Podemos observar en el siguiente gráfico la interpretación de esta tabla de forma visual, comparando el revenue con los Items vendidos y el tamaño del cluster:



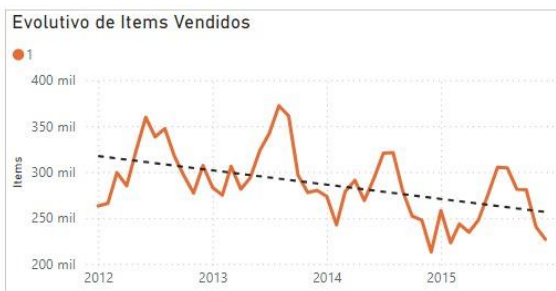
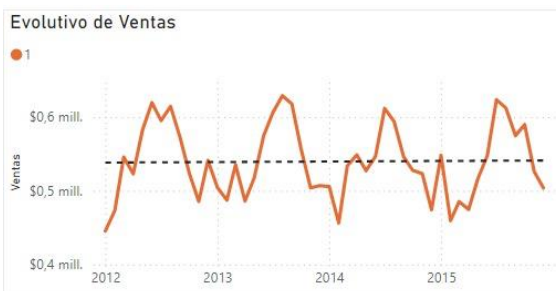
## Cluster 0:

El cluster 0 se compone de un gran número de ítems (44,5%), con un precio promedio parecido a la media y con una tendencia creciente respecto a los años anteriores. Con un crecimiento del 20% en las ventas el fin de semana. Todos ellos pertenecientes a la categoría Supermarket. Este cluster supone prácticamente la mitad de la facturación total.



## Cluster 1:

El cluster 1 en cambio se compone de 84 ítems (2,7%), repartido en departamentos de todas las categorías. Con un precio promedio muy por debajo de la media y una tendencia de las ventas muy estacional, prácticamente sin crecimiento a lo largo de los años. Si bien sigue la tónica de venderse más los fines de semana, su característica principal es su alta demanda en los meses de verano. También es resaltable su tendencia a la baja de ventas, que no se traduce en facturación debido al incremento de los precios, pero que entre 2010 y 2015 ha bajado aproximadamente un 15%.



## Cluster 2:

El cluster 2 se compone también de un número limitado de ítems (18,5%), con un precio promedio muy por encima de la media. Estos ítems sólo se componen de la categoría Accesorios. Principalmente destaca el departamento Accesorios 1 con un 95% de los ítems.

Sus ventas y unidades vendidas desde el 2012 han aumentado de forma constante, en 2015 han duplicado sus ventas respecto a 2012.



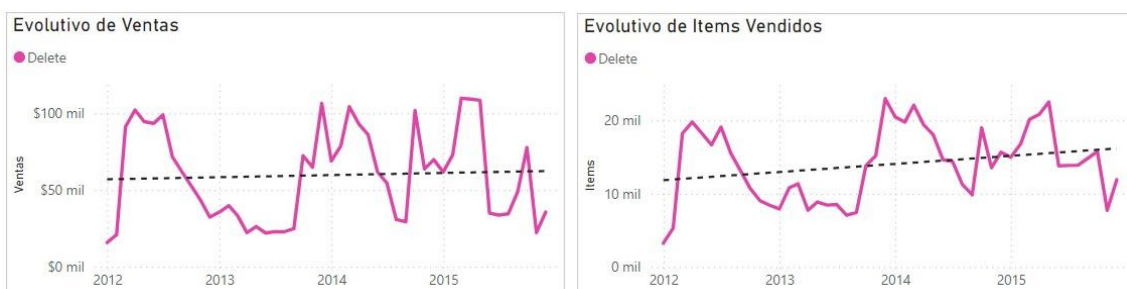
## Cluster 3:

El cluster 3 por último, se compone también de un gran grupo de ítems (34%), pero únicamente formado por ítems de la categoría Home&Garden. Estos tienen un precio promedio superior a la media con una tendencia creciente respecto a los años anteriores. De hecho, a pesar de venderse un 25% menos que el cluster 1, genera el doble de beneficios que este cluster.



## Delete:

Estos ítems tienen una clara caída en las ventas durante las últimas semanas. Se caracterizan por representar menos del 2% de los beneficios en todos estos años y haber vendido tan solo 0,7 millones de productos.





## Tarea 3: Time Series

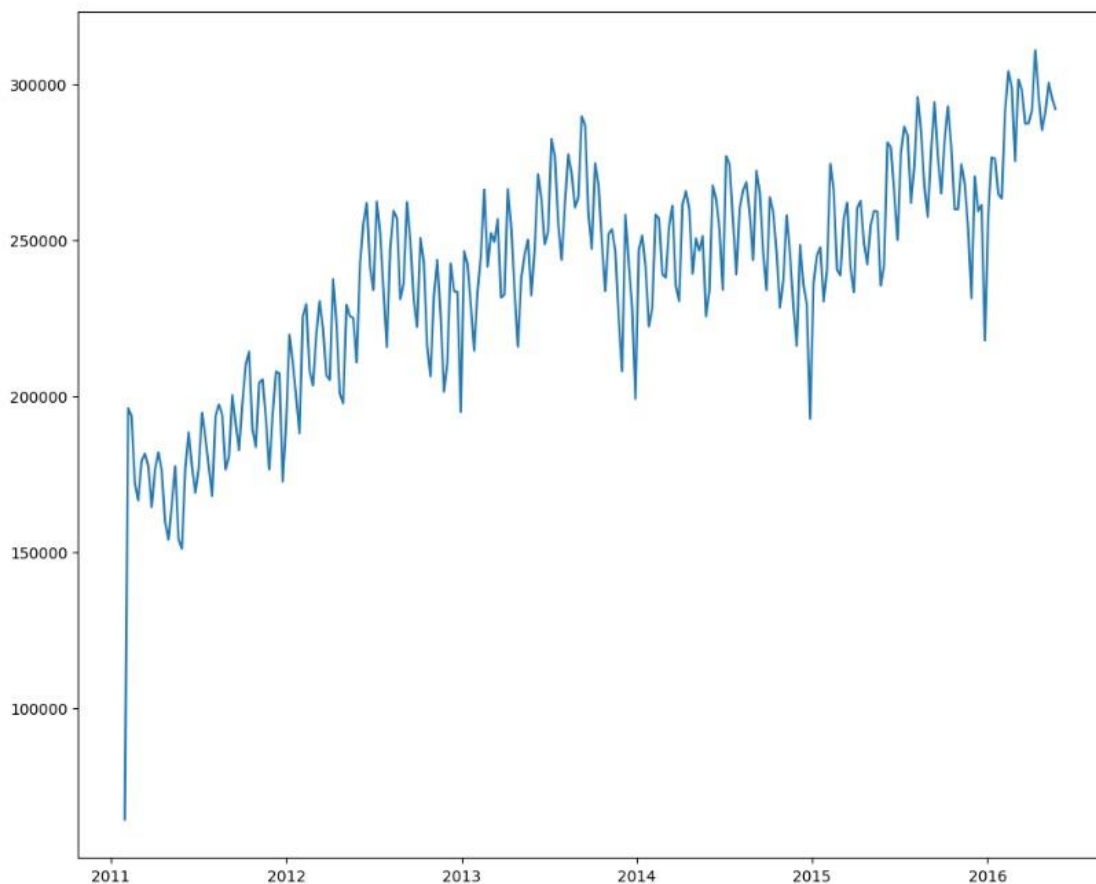
Para el modelo de Time Series se vieron varias opciones, se entrenaron modelos como ARIMA o PROPHET, pero finalmente vimos que la mejor opción para un dataframe multi-variante era hacer la predicción con un modelo XGBOOST.

Este modelo XGBOOST también se ha entrenado de distintas maneras. Inicialmente se optó por hacer un entrenamiento por días. Pero observamos que era mucho más lento y requería mucha RAM y memoria. Además de que la predicción era más inexacta a nivel de semanas

Finalmente optamos por hacer el entrenamiento con las ventas por semanas y conseguir predicciones más realistas. Para ello conservamos el sell price de la última semana para cada id.

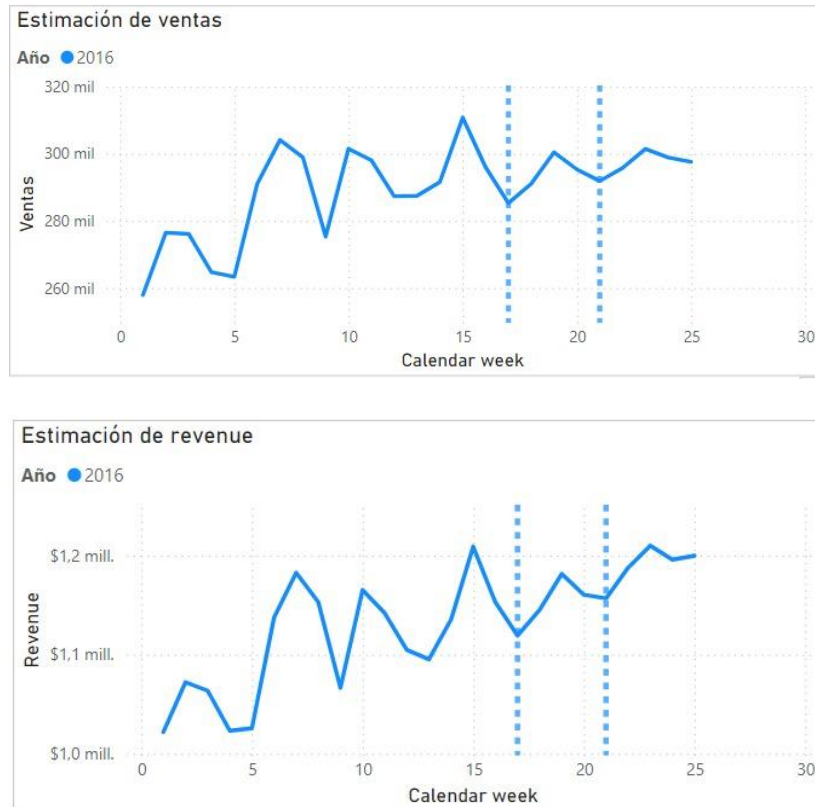
Este sería el evolutivo de ventas incluyendo una predicción de 4 semanas.

Monthly Sales for all items in all shops



Nuestro modelo de Time Series nos aporta predicciones de 4 semanas, pero repitiendo este proceso 3 veces pudimos ampliar nuestras predicciones a 12

semanas y graficar. Aunque realmente la principal idea es hacer el estudio de las 4 primeras semanas debido a que es la parte entrenada con datos totalmente reales. Lo ideal es iterar nuestro modelo cada vez que tengamos nuevos datos reales y no hacerlo sobre datos que se correspondan con predicciones de nuestro modelo.



Viendo este gráfico global más detallado de 2016, podemos apreciar nuestras primeras 4 semanas de predicciones, las cuales se corresponden con la zona delimitada en el gráfico. Se caracterizan por tener un aumento en las ventas respecto a nuestro punto de partida que es la semana 17. Esta subida se sitúa en torno a un 5,5% durante las dos primeras semanas, descendiendo luego un 1,5% en las dos posteriores.

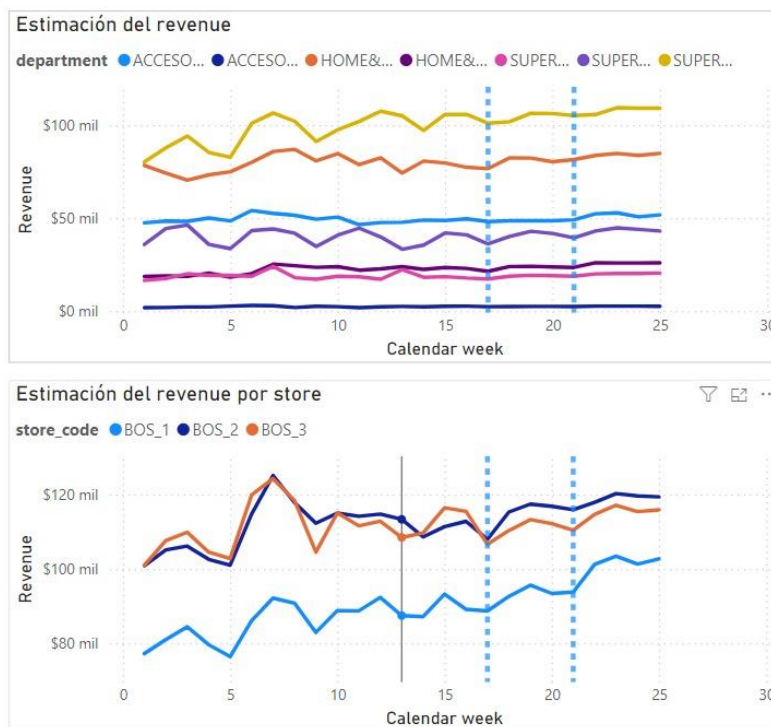
Así podemos concluir que al final de nuestras 4 semanas de predicciones hay un aumento en torno al 3% en el revenue total. En las semanas posteriores vemos como se espera también un crecimiento exponencial, lo que cuadra con las tendencias de otros años, donde el verano es la época del año donde se esperan mejores ventas.

En cuanto a las predicciones a nivel global por categoría tenemos que la categoría Supermarket sigue siendo la que más beneficios aporta, con un crecimiento a las 2 semanas del 6% respecto a la semana 17, pero con tan sólo un incremento del 2% al final de las 4 semanas. Por otro lado, la categoría que más crece es la de Accessories con un aumento en torno al 7% en la segunda semana y que llega hasta el 7,5% al final de las 4 semanas de predicciones respecto a la semana 17.

## Estimación de la venta en Boston:

Respecto a nuestras predicciones en la región de Boston, tenemos que durante las primeras 2 semanas se llega a un aumento del 8,5% en revenue total y un 9,5% en ítems vendidos, descendiendo a solo un aumento del 5,5% en la semana 21, respecto a la semana 17. Este incremento de las ventas está impulsado por un incremento del casi 10% en el número de ítems dentro de la categoría de Home & Garden y un incremento del 9% del revenue en la categoría Supermarket. Además, tenemos como gran insight, un aumento del departamento Supermarket 2 del 18%, lo que sería muy interesante de cara a campañas de marketing de estos productos para estas fechas.

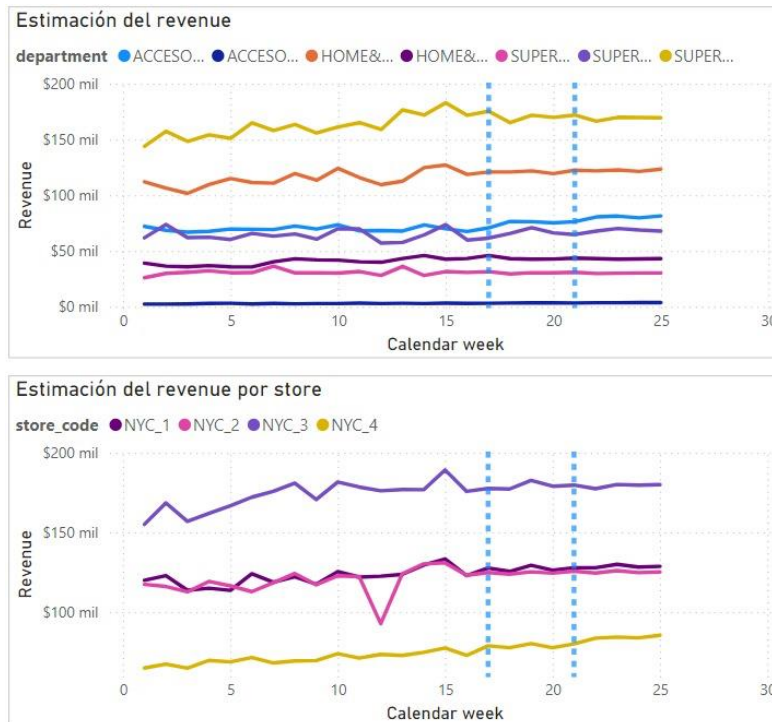
En cuanto a las tiendas, tenemos que es Roxbury (BOS\_2), la tienda que termina las 4 semanas con un mayor incremento tanto en ventas con un 6,8% como en revenue con un 7,34%.



## Estimación de la venta en Nueva York:

En la región de Nueva York es donde menos se prevé un incremento alto de las ventas, tenemos sólo un incremento del casi 2% en las primeras dos semanas, reduciéndose a casi un 1% a final de las 4 semanas. Destacar que en la única categoría en la que se prevé un mayor incremento de ventas es Accesorios con un casi 8% más de revenue.

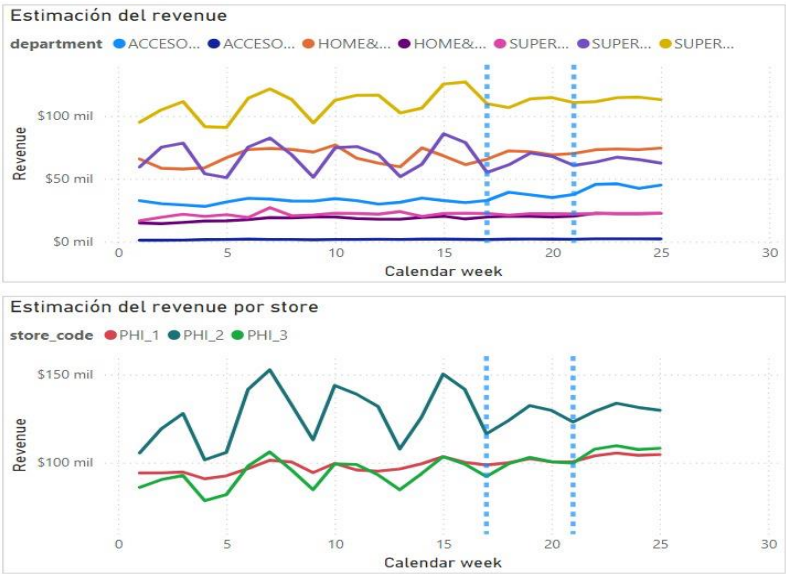
Es la tienda de Brooklyn (NYC\_4), la que mejores ventas se esperan en estas semanas pero sin pasar del 2%. Aunque decir que de cara a verano, con nuestras predicciones a más semanas se ve una clara tendencia al alza en esta tienda.



### Estimación de la venta en Philadelphia:

Philadelphia es la región que destaca por ser la que mayor incremento se prevé de las ventas respecto a la semana 17, con prácticamente un 10% más de revenue. Esto se debe a la subida en el revenue de la categoría Accesorios llegando a un 14% más, lo que supone la mayor subida de cualquier categoría en cualquier región durante estas 4 semanas. Además, se ve un incremento del 10% en Supermarket, que está mayormente potenciado por un incremento exponencial del 28% del departamento Supermarket 2, que al igual que en Boston supone el mayor crecimiento dentro de un departamento. Por ello sería interesante enfocar esas campañas de marketing que comentábamos antes en estos productos.

Entre las tiendas que más destacan, tenemos por un lado Yorktown (PHI\_2) con un 24% más de revenue y casi un 9,5% más de ítems vendidos. Y por otro lado la de Queen Village (PHI\_3) con un 20,5% más de revenue y casi un 13% más de ítems vendidos.



## Tarea 4: Caso de uso de abastecimiento de tiendas y API.

Para llevar a cabo el abastecimiento de tiendas vamos a utilizar las predicciones creadas con nuestro modelo de time series combinado con un modelo EOQ (modelo de cantidad económica de pedido). Este enfoque va orientado a lograr el mejor ajuste del stock por cada uno de los productos que tenemos por cada tienda, es decir por cada id. En el notebook del proyecto podemos apreciar todos los detalles en cuanto a código, para nuestro stock crearemos el data frame 'df\_stock'.

El modelo EOQ es una herramienta muy utilizada para el control de inventario y el abastecimiento de tiendas. Este modelo nos permite hacer un buen control tanto en la optimización del tamaño de pedidos como en el stock de seguridad que debemos tener en cada una de nuestras tiendas en función de cada id. De hecho, este modelo es muy útil a la hora de lidiar con productos en los que hay una alta variabilidad en el número de ventas, desde productos que se venden muy poco a otros que se venden asiduamente.

Para aplicar las fórmulas en la que se basa este modelo debemos tener ciertas variables entre las que podemos encontrar:

- Cálculo de venta promedio (D): en nuestro caso va a ser el valor de la columna `df_stock['n_sales']`, que son los valores aportados por nuestro modelo de time series.
- Desviación estándar de la demanda del producto ( $\sigma$ ): una nueva columna creada a continuación que muestra esta desviación, `df_stock['std_desviation']`.
- Coste de realizar un pedido (S): una columna que exprese el costo de realizar un pedido por cada producto, en nuestro caso hemos escogido un 10% del `sell_price`, aunque el equipo de logística podría darnos una estimación al respecto. Así esta variable será: `df_stock['sell_price']*0.10`.
- Coste de mantenimiento de un producto (H): variable que corresponde a lo que le cuesta a la empresa mantener un producto en stock, elegimos un 5%, aunque como en el caso de la variable S, logística podría ajustar este valor. Corresponde esta variable a: `df_stock['sell_price']*0.05`.

Una vez tenemos estas variables, el modelo nos va a aportar 3 datos para hacer el abastecimiento de nuestras tiendas utilizando las fórmulas detalladas en el notebook. Los cuales son:

1. Cantidad económica de pedido (EOQ), el número de unidades que se deben demandar en un pedido.
2. Stock de seguridad (SS), nos aporta la cantidad de unidades en inventario usadas para mitigar la desviación de la demanda. Quiere decir que es el stock extra para enfrentar posibles imprevistos relacionados con cambios

en la demanda o problemas de abastecimiento por parte de los proveedores.

3. Punto de reorden (ROP): representa el nivel de inventario en el cual se debe realizar un nuevo pedido para reabastecer el stock antes de que se agote. Así este valor indicará si fuera necesario reponer el producto. Por tanto, lo vamos a tomar cómo el stock preciso mínimo.

Hay que tener en cuenta que este modelo de abastecimiento se puede modificar en función de los productos que tenemos en stock, por ejemplo, se podrían ajustar las fórmulas para productos más perecederos o aquellos con una demanda muy alta. Además, hay que añadir que aquellos productos donde el stock preciso sea 0, serán productos que se podrán vender bajo demanda del cliente.

Finalmente, nuestro data frame final de stock acaba formándose de la siguiente manera.

	id	week	n_sales	n_productos_pedido	stock_extra	stock_preciso
0	ACCESORIES_1_00 1_BOS_1	17	3	7	1	4
1	ACCESORIES_1_00 1_BOS_2	17	4	8	1	5
2	ACCESORIES_1_00 1_BOS_3	17	4	8	1	5
3	ACCESORIES_1_00 1_NYC_1	17	7	10	2	8
4	ACCESORIES_1_00 1_NYC_2	17	7	10	2	8

## Desarrollo de la API

Con todo lo anterior, nuestro objetivo será optimizar el proceso de reposición de stock desarrollando una API, de manera que el modelo pueda ser accesible y utilizable para consultar el stock necesario.

En el caso de uso con el que estamos trabajando, la API tiene la función principal de proporcionar información precomputada sobre el stock necesario para un producto en una tienda concreta y en la semana que se especifique. Esta información es esencial para optimizar la reposición de inventario en las tiendas, minimizando el stock remanente y asegurando que los productos estén

disponibles para los clientes sin necesidad de excederse en el número de pedidos.

Será, por tanto, una API sencilla que recibirá una solicitud con los siguientes parámetros:

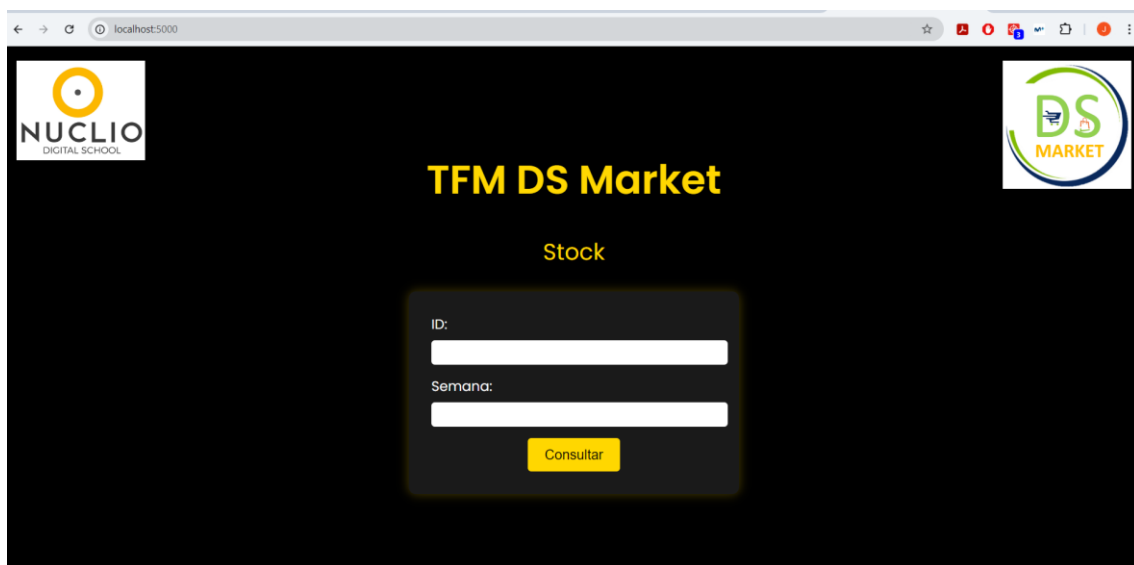
- 'id' : Es la clave única que identifica el artículo en cuestión en una tienda específica.
- 'week' : El número de la semana para la cual se desea conocer el stock necesario.

Una vez recibida la consulta, la API buscará la entrada correspondiente en el data frame 'df\_stock', mencionado anteriormente, que contiene la información precomputada.

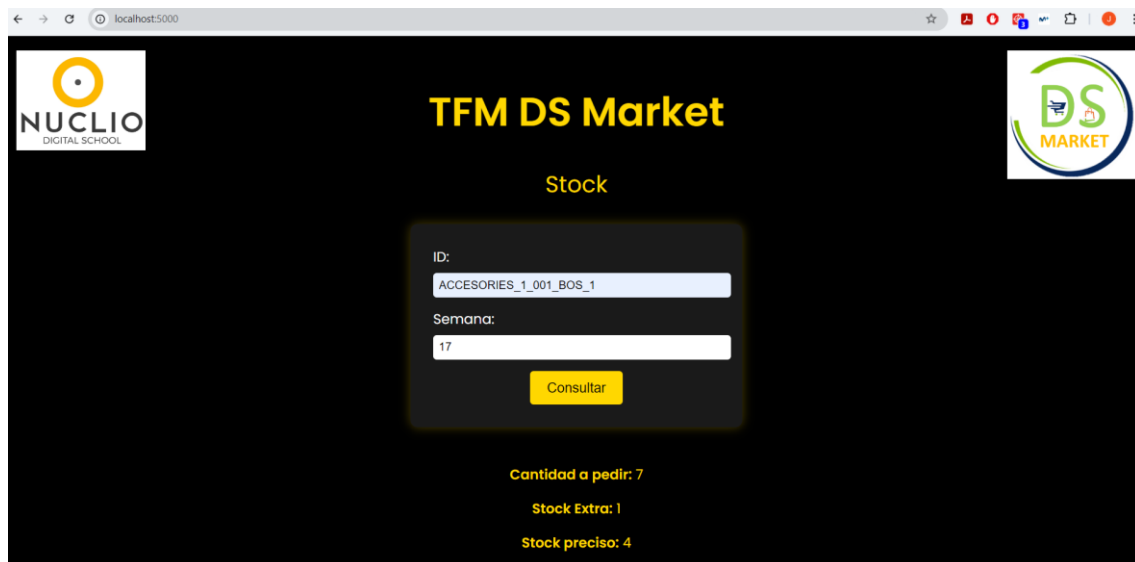
Por último, la API devolverá la información de stock en formato JSON, permitiendo que el usuario que haga la solicitud pueda utilizar estos datos para tomar decisiones de reposición de inventario y hacer este proceso más eficiente.

En este punto, la API puede devolver dos tipos de respuesta, dependiendo de si el dataframe de stock contiene la información necesaria para el 'id' y 'week' seleccionados, de manera que, si hay registro en el data frame de stock de un producto en una tienda concreta y en una semana dada, la API devolverá los valores de stock correspondientes, es decir, la fila del 'df\_stock' que contiene los valores de la consulta. Por el contrario, si la pareja de valores de la consulta ('id' y 'week') no contiene ningún registro en el data frame de stock, la API devolverá un mensaje de error indicando que no se han encontrado datos para el 'id' y 'week' especificados.

A continuación, se muestran algunos ejemplos de la pantalla de inicio y de las consultas que permite hacer la API.







NUCLIO DIGITAL SCHOOL

# TFM DS Market

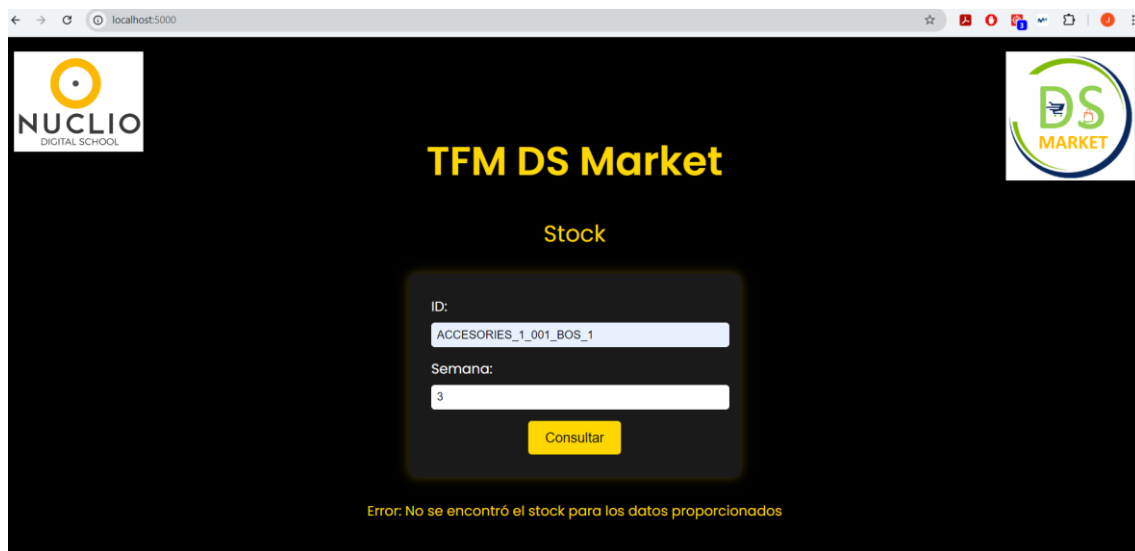
Stock

ID:  
ACCESORIES\_1\_001\_BOS\_1

Semana:  
17

Consultar

Cantidad a pedir: 7  
Stock Extra: 1  
Stock preciso: 4



NUCLIO DIGITAL SCHOOL

# TFM DS Market

Stock

ID:  
ACCESORIES\_1\_001\_BOS\_1

Semana:  
3

Consultar

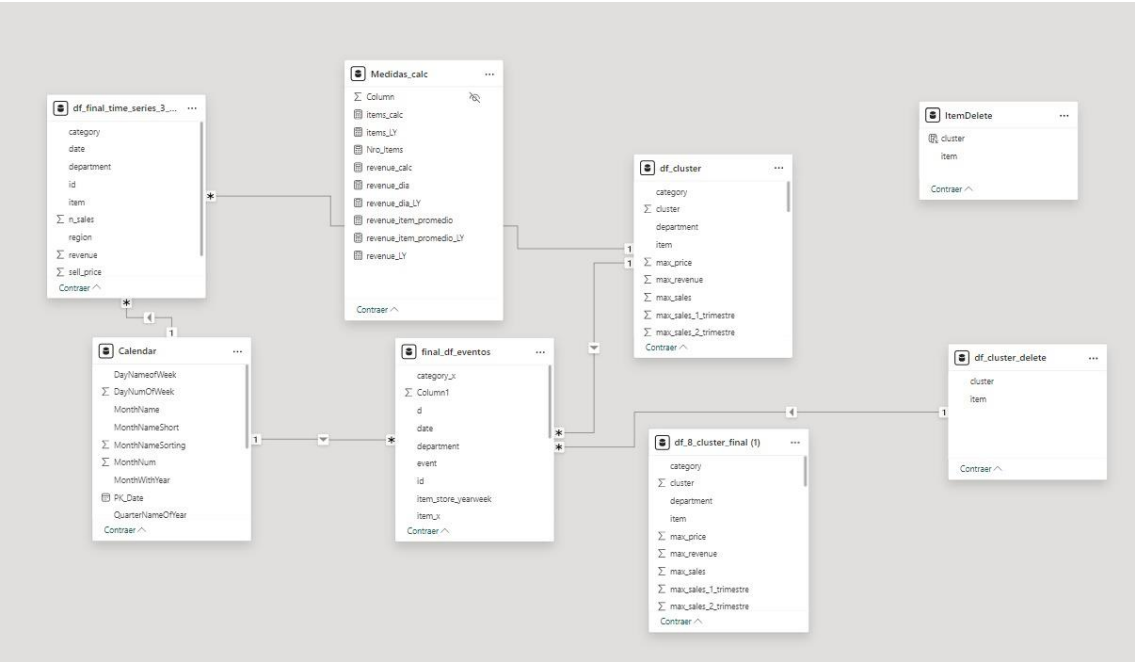
Error: No se encontró el stock para los datos proporcionados

Para el despliegue de esta API se ha trabajado con Flask, que es un framework escrito en Python que permite crear aplicaciones web, y Docker, cuya labor es empaquetar la aplicación en un contenedor que luego puede ser desplegado fácilmente en cualquier servidor que soporte Docker, evitando así configurar manualmente cada servidor de destino. Todo ello se ha configurado trabajando en un entorno de VS Code.



# Annexo 1:

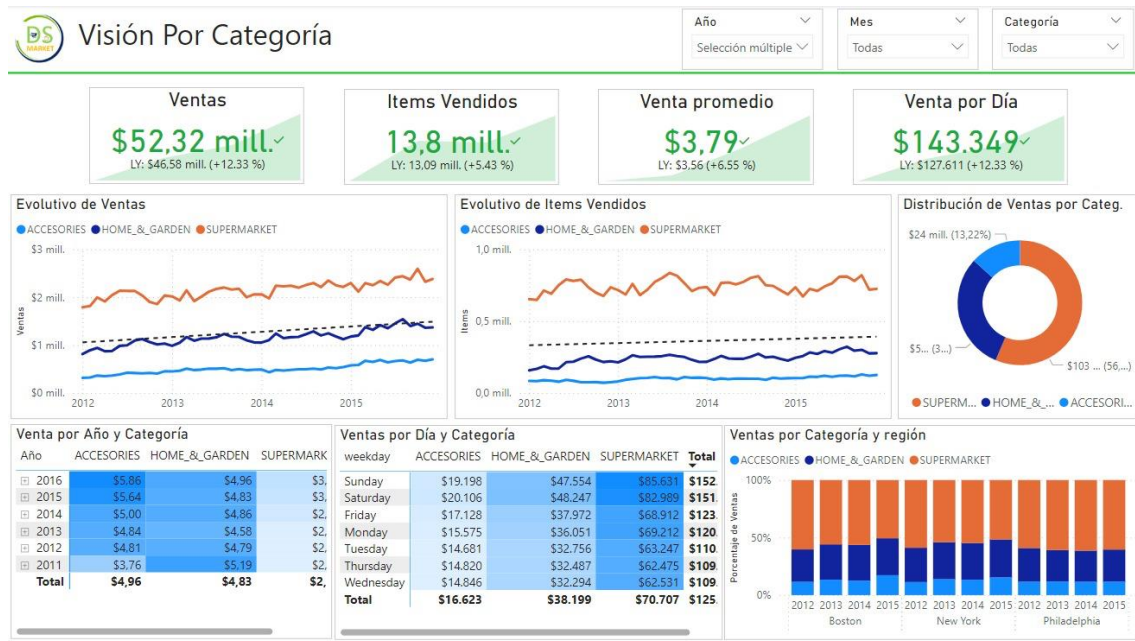
## Relación entre tablas



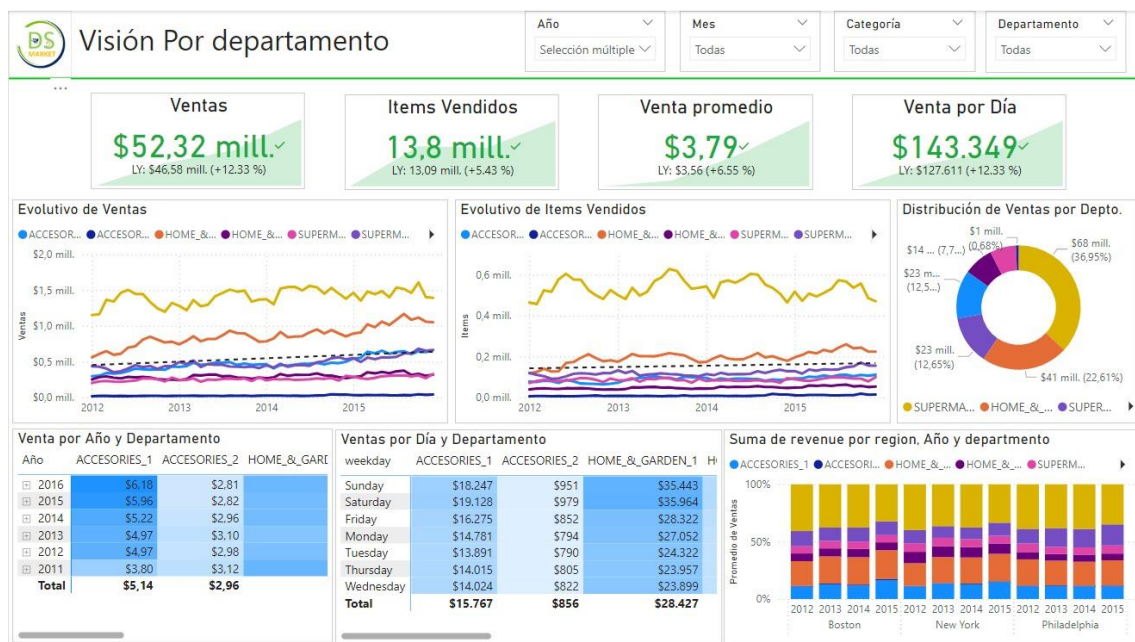
## Visión Global



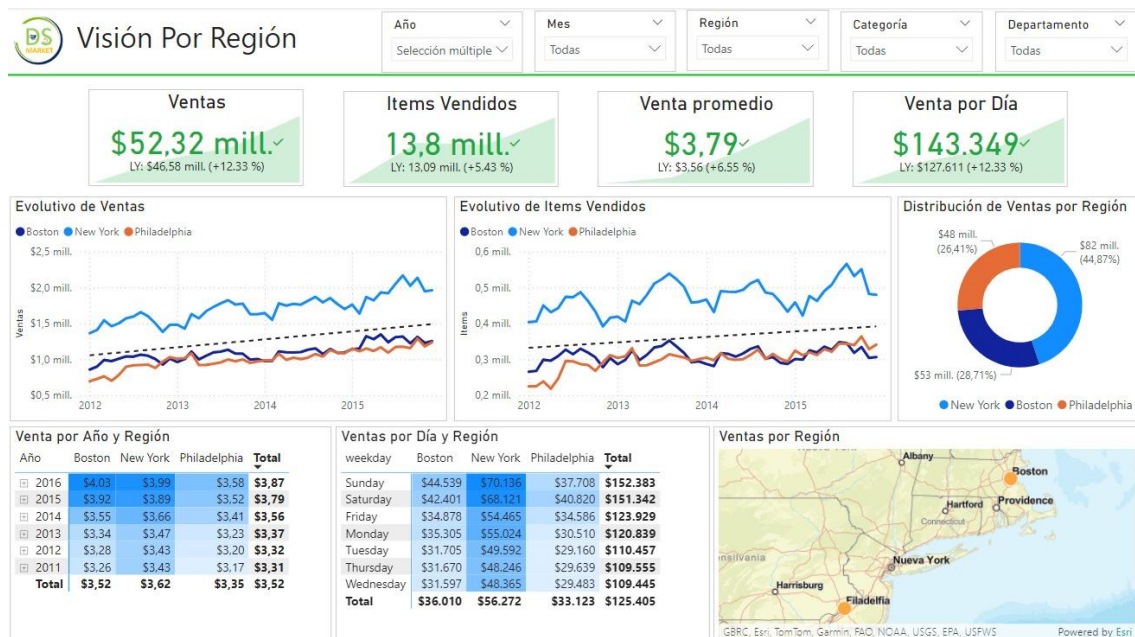
## Visión por Categoría



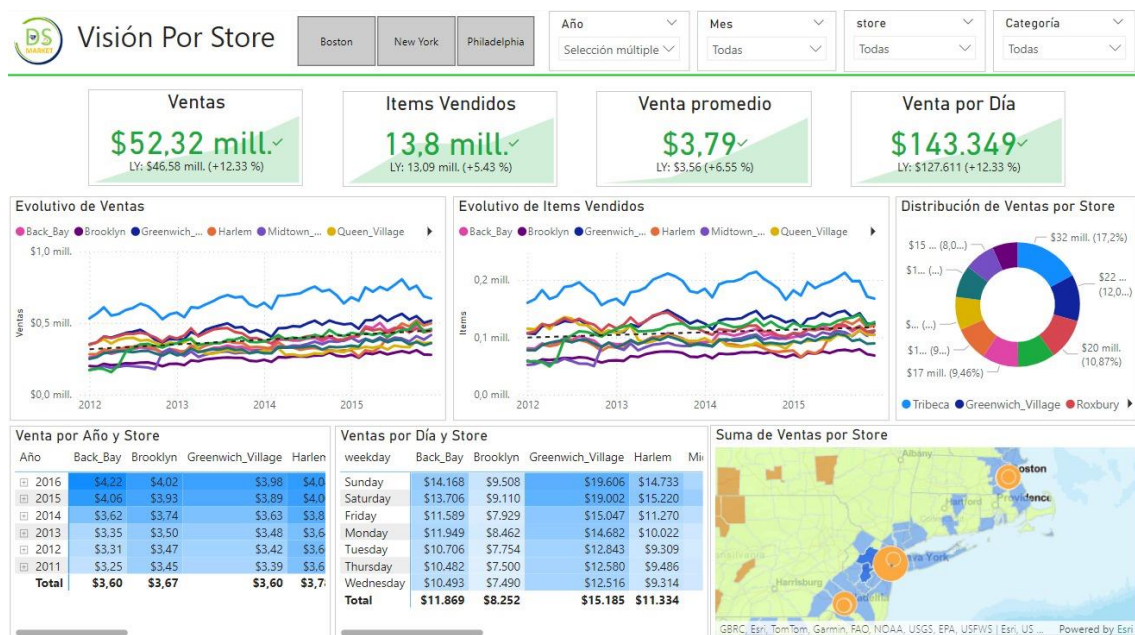
## Visión por Departamento



## Visión por Región

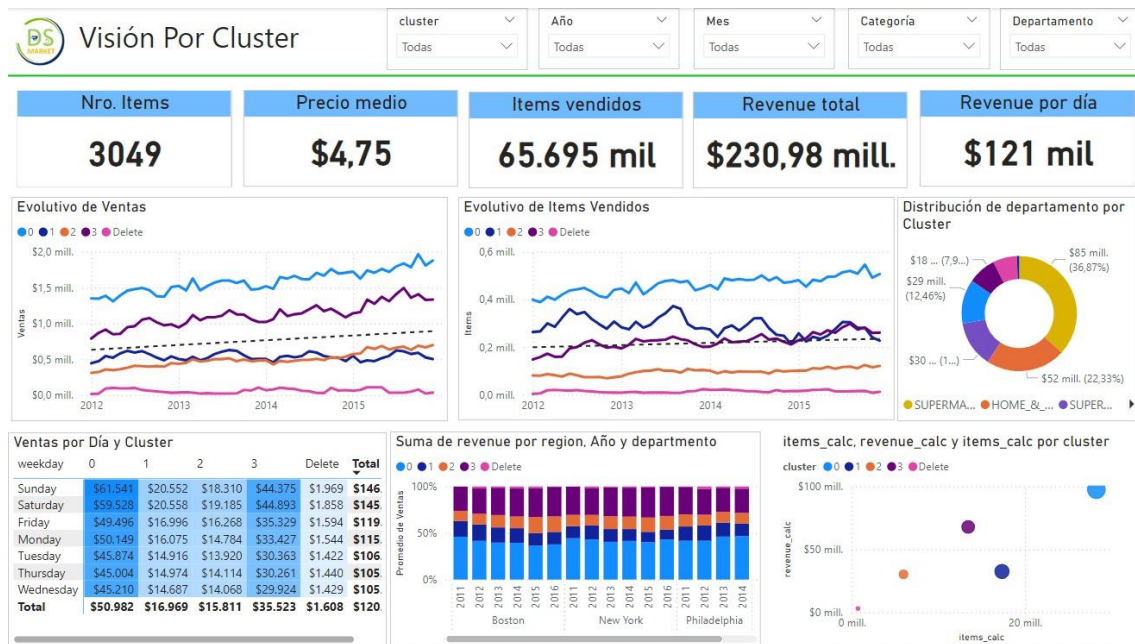


## Visión por Store

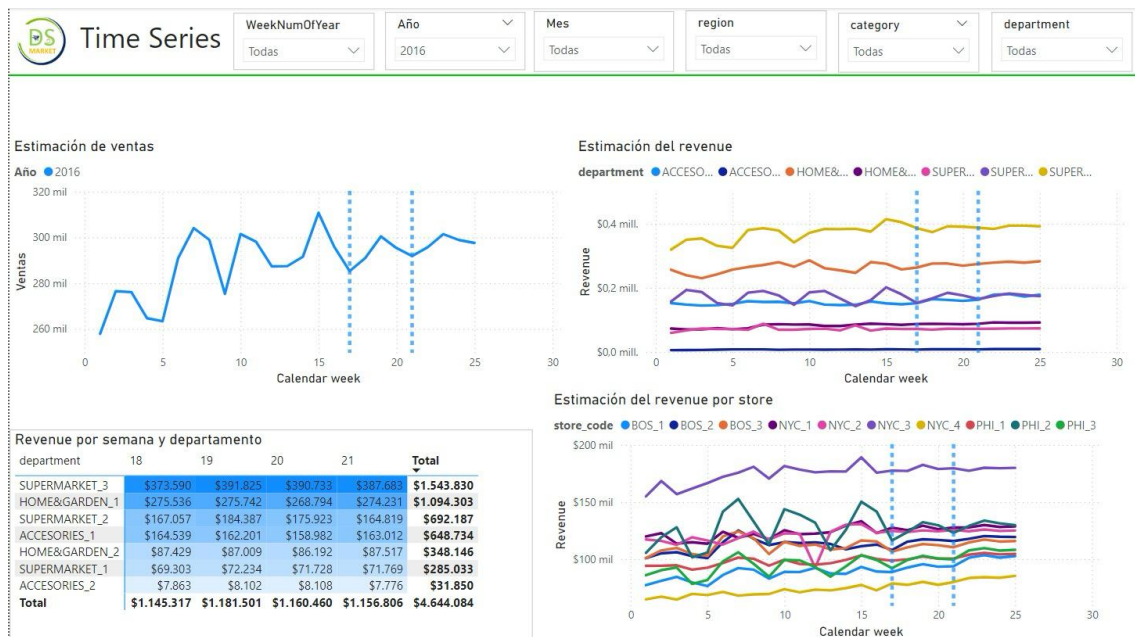




## Visión por Cluster

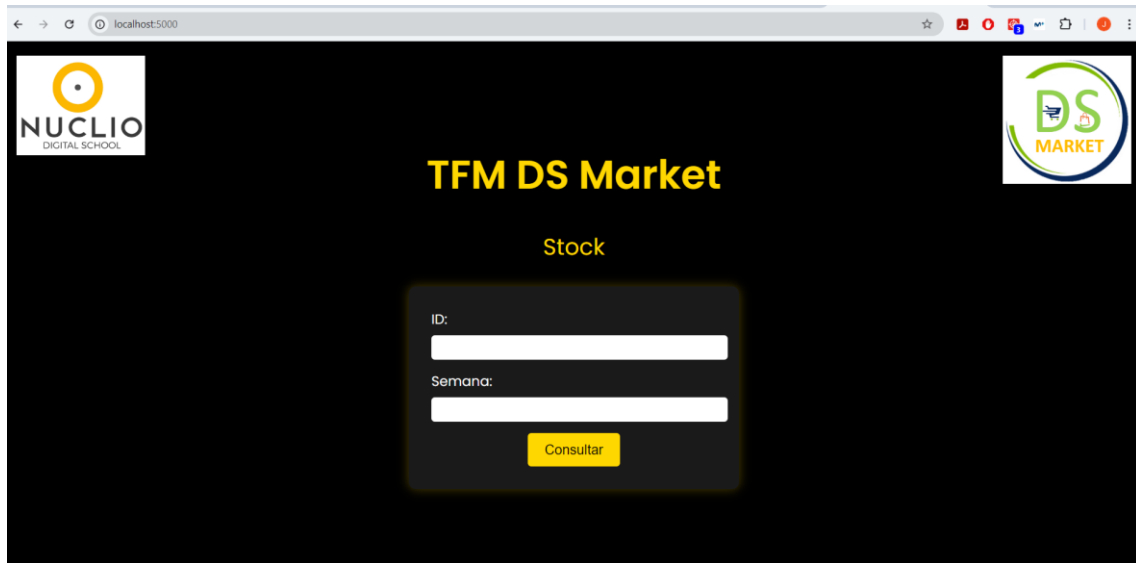


## Visión del Time Series





## API

A screenshot of a web browser showing the 'TFM DS Market' Stock API interface. The browser's address bar shows 'localhost:5000'. The page has a black background. In the top left corner is the 'NUCLIO DIGITAL SCHOOL' logo. In the top right corner is the 'DS MARKET' logo. The main heading 'TFM DS Market' is in large yellow text. Below it, the word 'Stock' is in smaller yellow text. In the center, there is a dark gray rounded rectangle containing two input fields: 'ID:' and 'Semana:'. Below these fields is a yellow button labeled 'Consultar'.