



# Visualización de Datos - Entrega Intermedia

Adrián José Zapater Reig

---

## 1. Introducción

**Objetivo:** Obtener información relevante del conjunto de datos para poder visualizar en un *dashboard* el histórico y estado actual de cada tienda.

Fuente de Datos

La fuente de datos que se va a utilizar es la proporcionada por el equipo docente. Esta puede ser descargada desde el siguiente [enlace](#).

La estructura de los datos es la siguiente:

- *Store*: el número de la tienda
- *Department*: el número de departamento
- *Date*: la semana
- *Weekly\_Sales*: ventas para el departamento dado en la tienda dada
- *IsHoliday*: si la semana es una semana especial de vacaciones
- *Temperature*: temperatura promedio en la región
- *Fuel\_Price*: precio del combustible en la región
- *MarkDown1-5*: datos anónimizados relacionados con las promociones que se están llevando a cabo. Los datos de *MarkDown* solo están disponibles desde noviembre de 2011 y no están disponibles para todas las tiendas. Cuando falta un valor se marca como NA
- *CPI*: el índice de precios al consumidor
- *Unemployment*: la tasa de desempleo
- *Type*: una descripción del tipo de tienda que es. Existen tres tipos: A, B y C.
- *Size*: número de metros cuadrados de la tienda.

## 2. Análisis exploratorio

¿Qué información tenemos?

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 282451 entries, 0 to 282450
Data columns (total 16 columns):
 #   Column      Non-Null Count  Dtype  
 --- 
 0   Store        282451 non-null int64  
 1   Dept         282451 non-null int64  
 2   Date         282451 non-null object 
 3   Weekly_Sales 282451 non-null float64 
 4   IsHoliday    282451 non-null bool   
 5   Temperature  282451 non-null float64 
 6   Fuel_Price   282451 non-null float64 
 7   MarkDown1   100520 non-null float64 
 8   MarkDown2   74232 non-null float64  
 9   MarkDown3   91521 non-null float64  
 10  MarkDown4   90031 non-null float64  
 11  MarkDown5   101029 non-null float64 
 12  CPI          282451 non-null float64 
 13  Unemployment 282451 non-null float64 
 14  Type         282451 non-null object 
 15  Size          282451 non-null int64  
dtypes: bool(1), float64(10), int64(3), object(2)
memory usage: 32.6+ MB
```

Como podemos ver en la celda de arriba, nuestra fuente de datos contiene 282451 registros con 16 columnas. La mayor parte de las columnas son cuantitativas continuas, 1 de precisión simple (Size) y 10 de doble precisión. También tenemos variables cuantitativas discretas, *IsHoliday* que es booleana y *Date* que es una cadena de texto. Por último, tenemos 3 variables cualitativas nominales: *Dept*, *Store* y *Type*.

La mayoría de las columnas no tienen valores nulos o NaN, como se puede ver la imagen 1, a excepción de las 5 columnas de *MarkDown*, que como hemos indicado en la descripción de la fuente de datos, sólo contienen información entre el 5/2/2010 y el 26/10/2012. Más adelante tendremos que decidir cómo tratar los valores no informados o si vamos a descartar los registros que no los contengan.

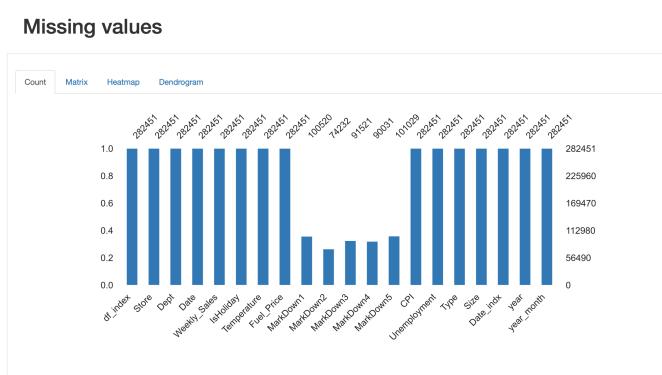


Imagen 1 Grafica que muestra la frecuencia de valores nulos en el dataset

## Modelo de Organización de la información

Ahora que ya conocemos los tipos de datos, tenemos que entender a qué nivel de abstracción están los datos y diseñar un modelo de organización que lo refleje. Este modelo está formado por objetos y relaciones.

Si visualizamos los datos con la estructura mas simple: una tabla, podemos sacar las siguientes conclusiones:

- Las columnas *MarkDown1-5* no varían entre departamentos dada una fecha y una tienda, por lo que su nivel de granularidad es *Store*.
- Las ventas semanales (*Weekly\_Sales*) varían entre departamentos dada una tienda, por lo que su nivel de granularidad es *Store*.
- Temperature, *CPI* y *Unemployment* no varían entre tiendas o departamentos en un mismo día, por lo que su nivel de granularidad es *Date*.

De estas conclusiones podemos sacar el modelo de organización de información con el que estamos trabajando. Dependerá de nosotros decidir si modificarlo para facilitar el análisis y la visualización.

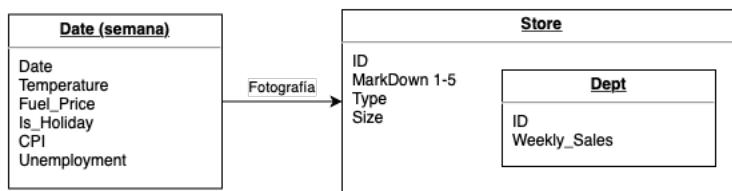


Ilustración 1 - Modelo

El modelo está formado por 2 entidades principales unidas por una relación de composición: *Store*, o tienda, y *Dept*, o departamento. *Store* es un nivel de agrupación superior a *Dept*, y puede contener uno o mas *Depts*. *Store* contiene una serie de atributos que, a primera vista, parecen estar relacionados: *Markdown 1-5*, no considero que tengan tanto significado por si solos como para organizarlos como una entidad u objeto, pero si que los trataremos en conjunto. En el diagrama se muestran de forma aislada al resto de atributos.

La última entidad que queda por presentar es *Date* o tiempo, que nos ofrece una dimensión extra sobre las otras dos. La relación entre *Date* y *Store* se puede ver como una fotografía, cada instancia de *Date* tiene un *Store* distinto que a su vez tiene N *Depts*. Además de la relación con *Store*, *Date* tiene atributos propios.

Este modelo de organización de la información nos ofrece varias perspectivas sobre las que analizar y visualizar los datos, las mas interesantes son:

3. Vista desde la dimensión *Store*
4. Vista desde la dimensión *Dept*
5. Vista desde la dimensión *Date*
6. Vista desde en conjunto desde *Date* y *Store*

## 2.1. Análisis de variables

Ahora que ya conocemos la información con la que estamos trabajando, vamos a analizar con un poco mas de detalle cada variable.

Store

*Store* representa el identificador de tienda de Walmart. Como podemos ver en la imagen 2, hay 45 tiendas y los registros están bastante repartidos (las tiendas 36-38 y 42-44 tienen un 20% menos de registros).

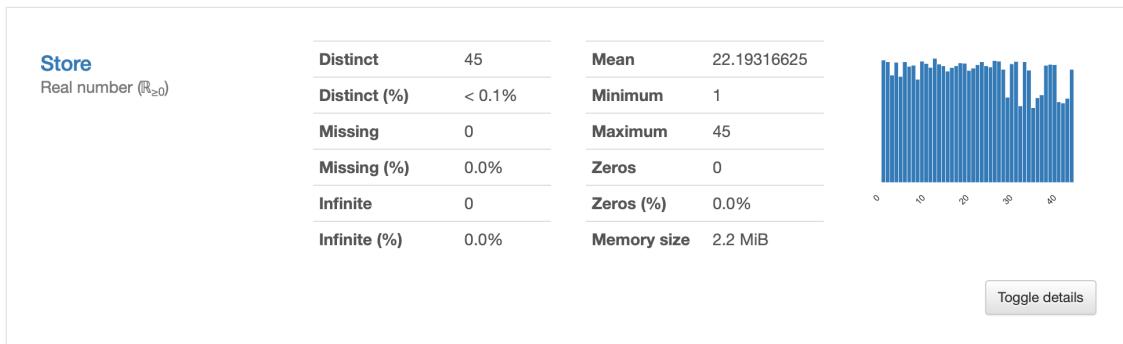


Imagen 2 Resultado de pandas profiling de la variable store

*Store* está en el rango 1-45 y, como hemos comentado antes, es una variable cualitativa nominal. Puesto que Walmart tiene mas de 45 tiendas en EE.UU., podemos concluir que *Store* no es exhaustiva. Para averiguar si es exclusiva desde el punto de vista del *Dept*, debemos comprobar si hay algún departamento que pertenezca a mas de una tienda.

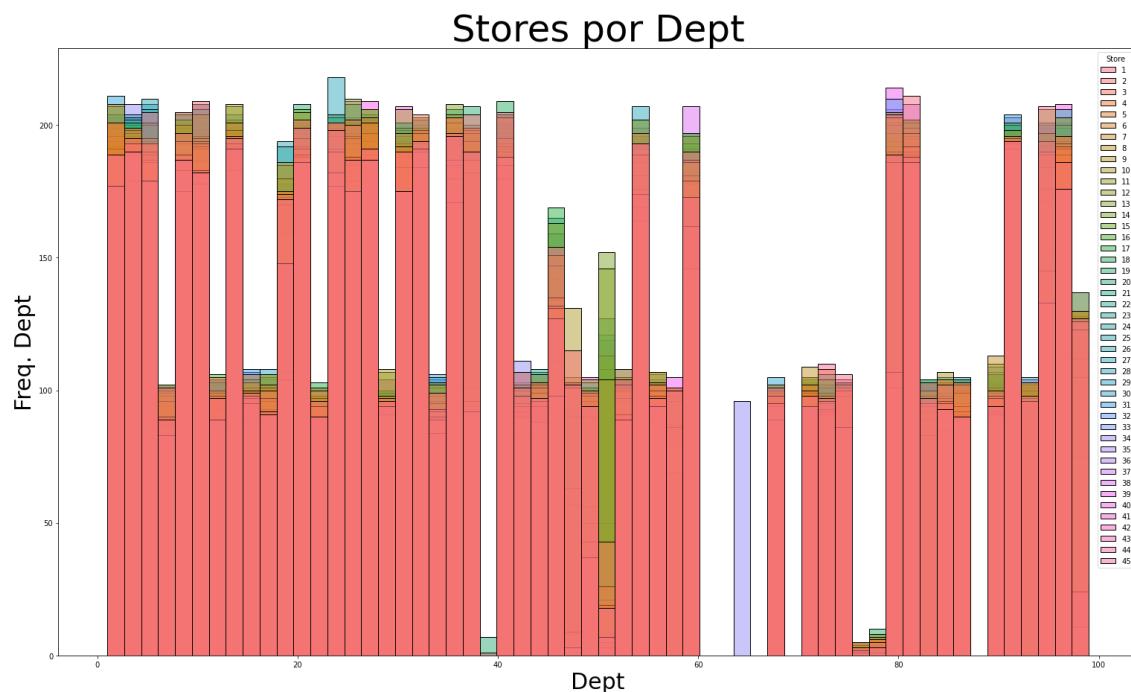


Imagen 3 Grafico que muestra la frecuencia de la variable departamento, y en que stores se encuentra

En la figura 3 podemos ver claramente que *Store* no es exclusiva con respecto de *Dept*, ya que cada departamento puede aparecer en mas de una tienda.

Fuel Price

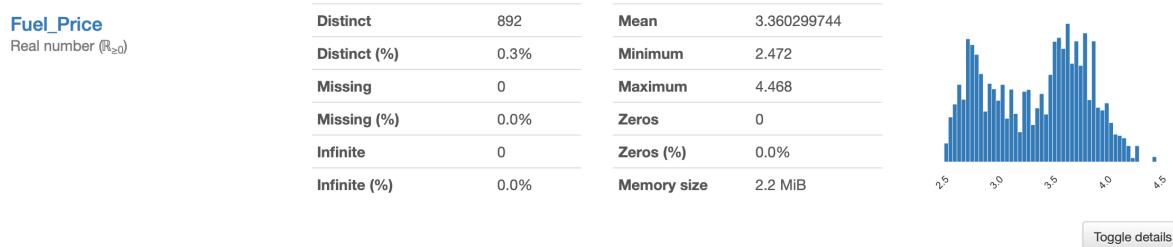


Imagen 4 Resultado de pandas profiling de la variable *store*

Fuel price representa el precio medio en dolares de un galón de gasolina en USA, y es un atributo de *Date*. Esta variable es cuantitativa continua con rango 2,472 – 4,468. En el histograma de valores, de la imagen 4, podemos ver que la mayoría de valores se concentran al rededor de 2,7 y 3,7 y que tenemos observaciones anómalas en la parte superior de 4,4.

```
: draw_box_plot(data, 'Fuel_Price')
```

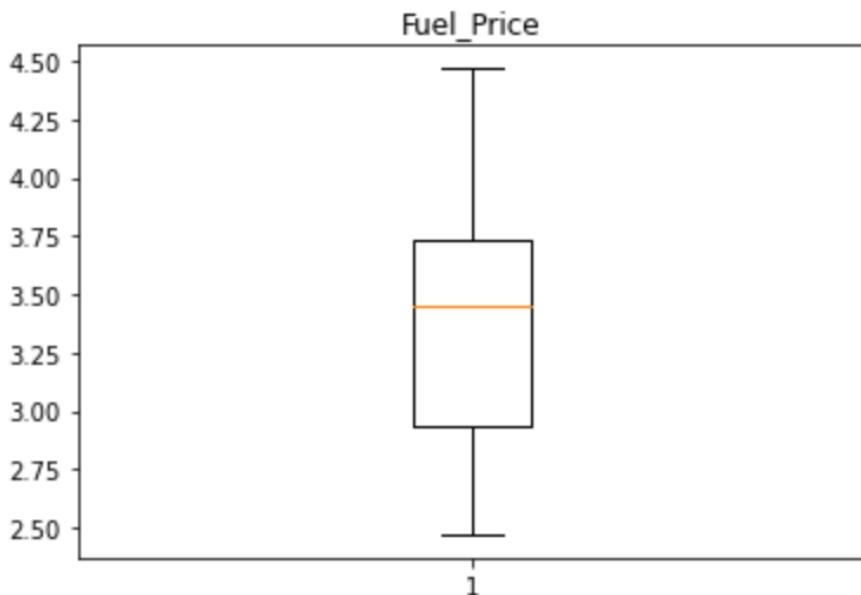
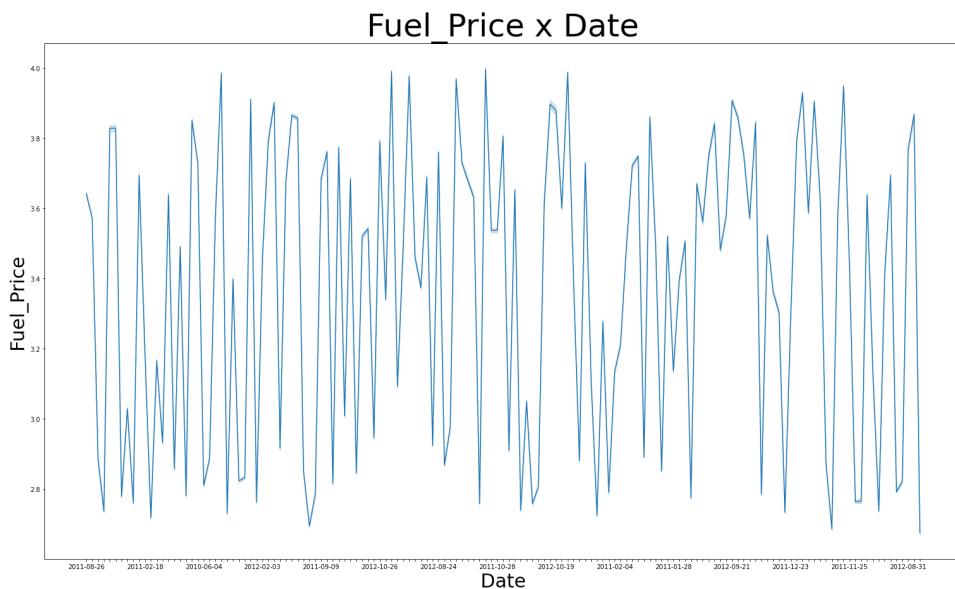


Imagen 5 Box plot variable Fuel\_Price

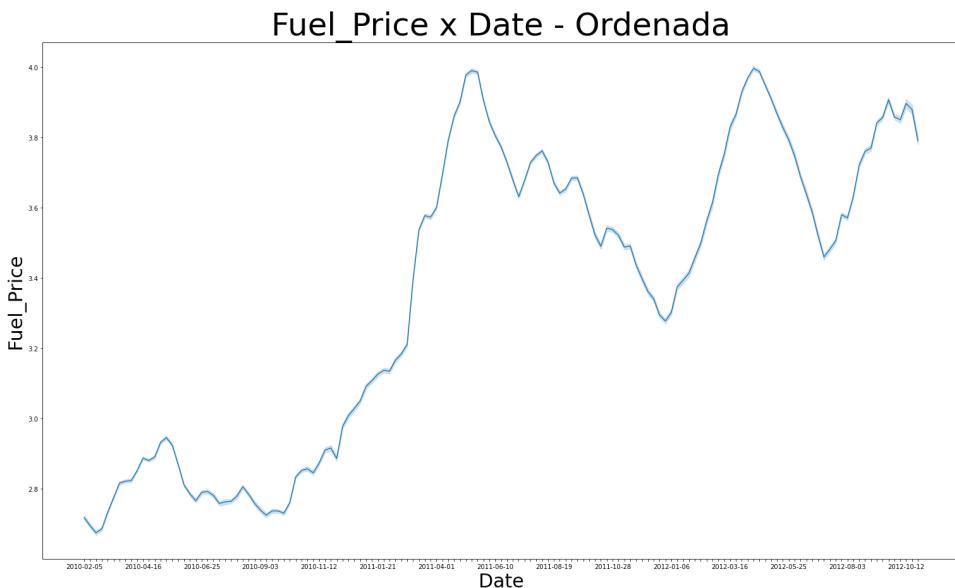
Tras analizar la columna con un boxplot, podemos concluir que no se trata de una anomalía, sino del valor máximo del conjunto.

A continuación, vamos a visualizar como se vería la fluctuación del *Fuel\_Price* con la variable tiempo (date en nuestro caso). Vamos a analizar dicha fluctuación en 3 granularidades distintas, por día, por mes y por año. En todos los casos deberíamos ver una tendencia positiva.



*Imagen 6 Fluctuación de la variable Fuel\_Price con el tiempo*

La imagen 6 muestra el *Fuel\_Price* contra *Date* por días, y como se puede apreciar no parece tener ningún patrón específico. Analizando el dataset vemos que la columna *Date* se ha tratado como cualitativa nominal (sin orden), cuando si que existe un orden (2015-03-05, 2015-03-06,...), por lo que hemos ordenado dicha columna, como se puede ver la imagen 7.



*Imagen 7 Fluctuación de la variable Fuel\_Price con el tiempo ordenado*

La imagen 7 muestra el *Fuel\_Price* contra el *Date* por días ordenado, en el que se puede apreciar el incremento tras cada año, y las fluctuaciones a lo largo de un año. Se puede

ver una gran subida del precio de la gasolina del año 2010 al 2011, y entre el 2011 y 2012 el precio parece que se ha mantenido constante en sus máximos, con una bajada del 2011-2012, y luego otra subida de precio al principio del año 2012.

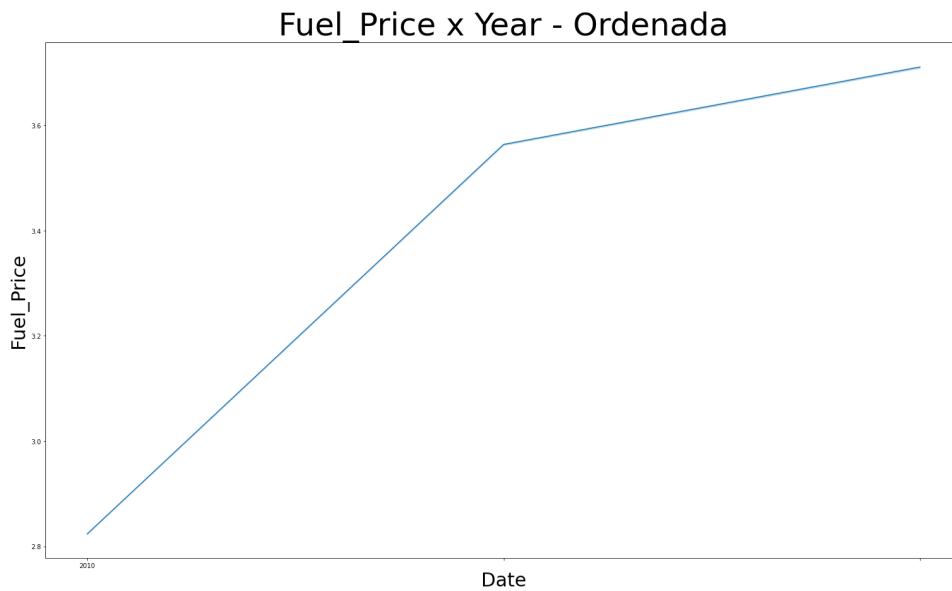


Imagen 8 Fluctuación del Fuel\_Price contra Date por años

En la grafica 8 plasmamos el *Fuel\_Price* contra el año, y como se puede ver perdemos mucha información, por que solo contamos 3 años. En cambio, la imagen 9 que muestra la variación del precio con el tiempo por mes y año, tiene un patrón similar al de la imagen 7, pero mas suavizada ya que no muestra los pequeños picos, eliminamos el ruido, que tiene la imagen 7.

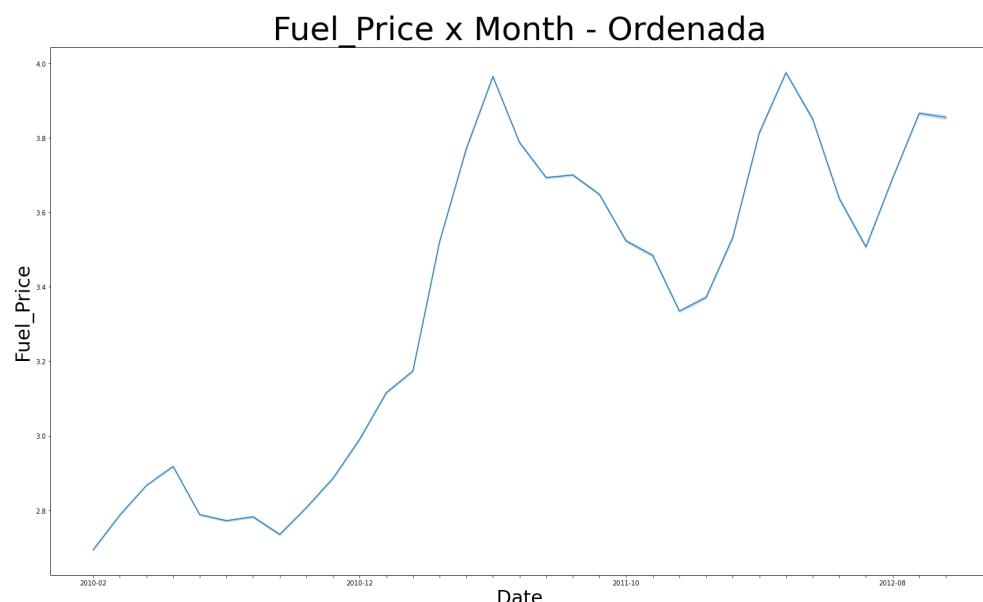


Imagen 9 Fluctuación del Fuel\_price contra Date por mes\_año

De las 3 graficas, la más representativa seria la imagen 7, en la que vemos la fluctuación por días, sin embargo, la imagen 9 también muestra suficiente información para

entender la tendencia. La imagen 8 teniendo tan poco rango de años, no nos aporta la información suficiente para poder sacar conclusiones del comportamiento del precio con el tiempo.

#### Weekly\_Sales , CPI, Unemployment y Temperature

Para estas variables haremos un análisis similar al descrito en *Fuel\_Price*, en la que plasmaremos las variables contra *Date* por días, ya que es la que mas información muestra como se ha concluido previamente. A continuación, mostramos las graficas superpuestas, y en la siguiente sección analizaremos la correlación entre dichas variables.

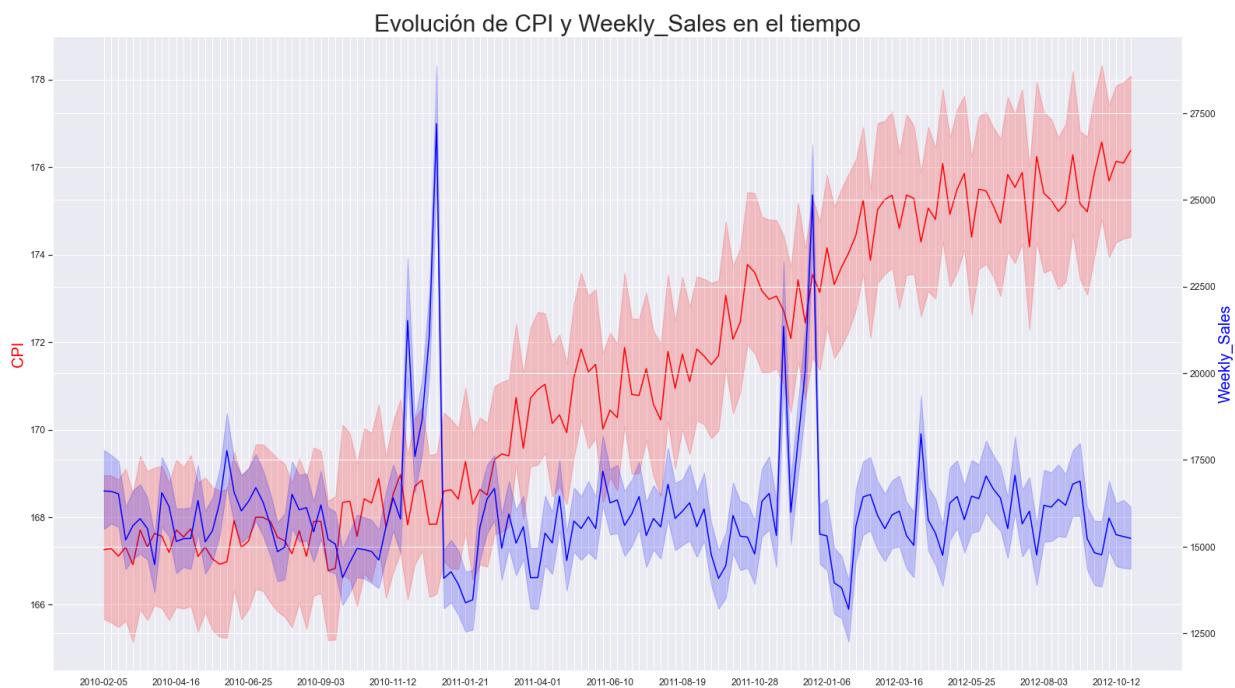


Imagen 10 Fluctuación del CPI y Weekly\_Sales contra Date por días

La imagen 10 muestra las fluctuaciones del *CPI* y *Weekly\_Sales*, como cabe de esperar, el *CPI* conforme avanzamos en el tiempo tiene una tendencia positiva, es decir cada vez el índice de precio del consumidor es mayor. Las *Weekly\_Sales*, muestran dos picos, que se podían atribuir a días festivos, uno es próximo a navidades y el otro coincide con el final de navidades.

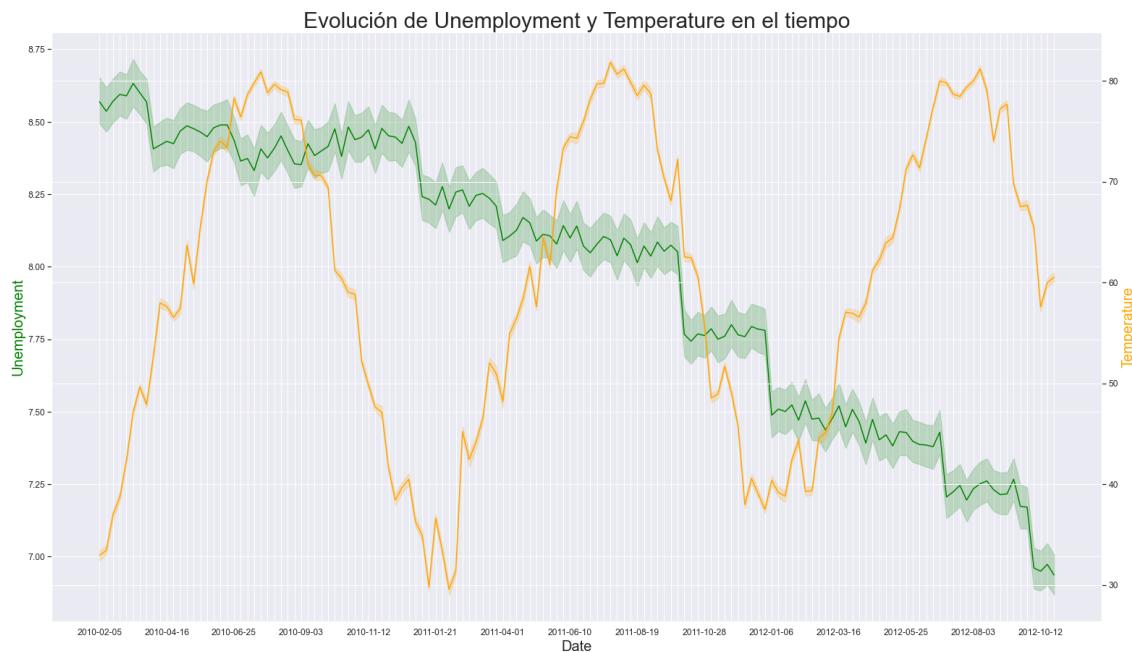


Imagen 11 Evolución de Unemployment y Temperature en el tiempo

La imagen 11 muestra la evolución de la *Temperture* y el *Unemployment* con el tiempo. La temperatura se aprecia que tiene los picos altos de junio a agosto, y los puntos más bajos en enero. La variable *Unemployment* tiene una tendencia negativa conforme al tiempo, es decir ha ido disminuyendo el numero de personas sin empleo desde 2010 hasta finales de 2012.

#### Type

La variable *Type* es cualitativa nominal, teniendo como posibles valores A, B y C. A continuación, mostraremos la distribución de cada tipo de tienda en un grafico de tarta, que muestra la cantidad de tiendas que hay de ese tipo.

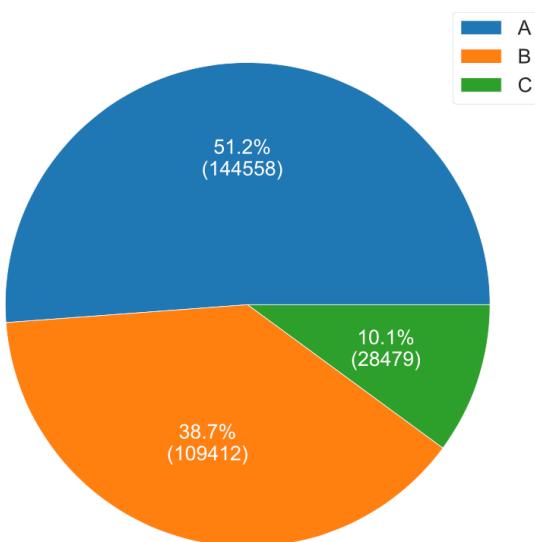


Imagen 12 Grafico que muestra la distribución de la variable *Type*

De la imagen 12 podemos ver que un 50% de las tiendas son del tipo A, y mas de un 25% pertenecen al tipo B, por lo que el tipo de tiendas no está equilibrado.

### Size

La variable *Size* es cuantitativa continua, y describe los metros cuadrados que tiene cada tienda. La imagen 13 muestra un histograma de la distribución de la variable, en la que se pueden distinguir 3 grupos, el primer hasta 50000, el segundo entre 100000 y 150000 y el ultimo entre 150000 - 200000.



Imagen 13 Resultados pandas profiling variable Size

### Markdown

En el dataset encontramos 5 variables llamadas *markdown\_1*, *markdown\_2*, *markdown\_3*, *markdown\_4* y *markdown\_5*, se trata de variables cuantitativas continuas.

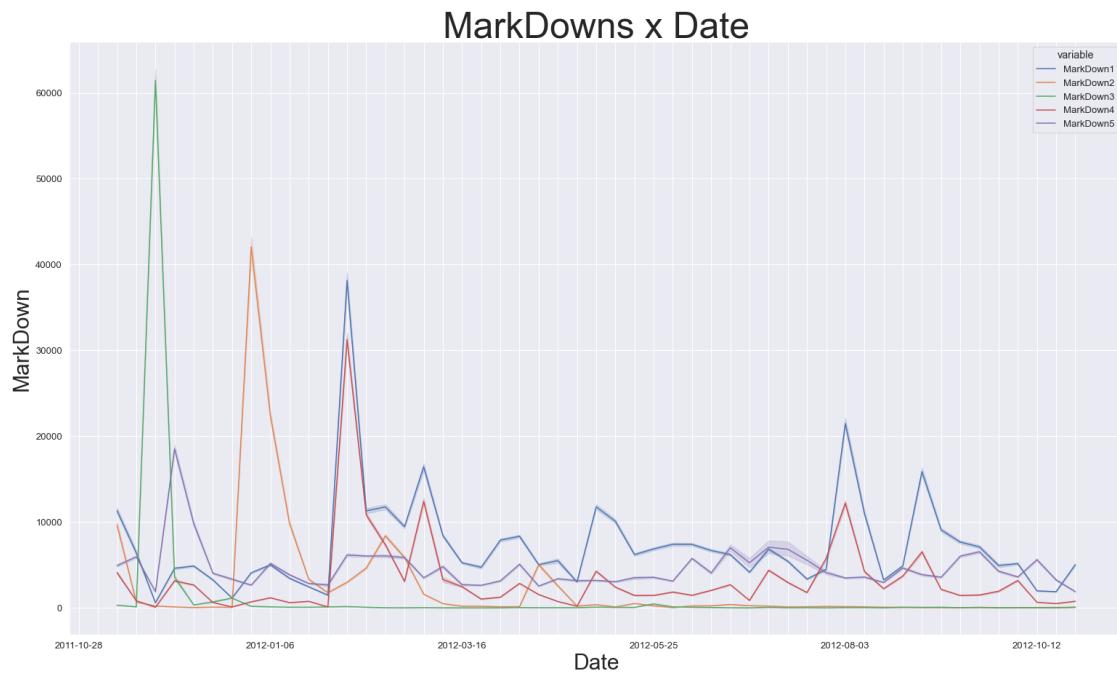


Imagen 14 Evolución de los markdowns con el tiempo

El grafico 14 muestra los 5 *Markdowns* y como evolucionan con el tiempo, el *Markdown* 1 (azul) y 4 (rojo) se puede ver que coinciden en el comportamiento, ya que los tienen

misimo picos y valores similares. En general a finales de 2011 todos los *Markdowns* tienen valores altos, y entre enero y marzo de 2012, empiezan a bajar, mostrando algunos picos en agosto de 2012.

## 2.2. Correlación entre variables

Tras el exhaustivo análisis de las variables vamos a analizar la correlación entre algunas de ellas, para así de esta manera poder descartar las menos relevantes.

En primer lugar, procederemos a pinta la *correlation\_matrix* mostrada en un *heatmap*, para así tener una primera impresión de que variables están relacionadas. En la imagen 15 podemos ver como la diagonal muestra una correlación fuerte, ya que se está relacionando las mismas variables entre si. En general las variables no presentan una clara correlación, si que cabe destacar que el *markdown1* y *markdown4* muestran una fuerte correlación, como habíamos visto en el grafica 14. Para los markdowns, vamos agruparlos en una variables, tomando la media de cada uno de ellos.

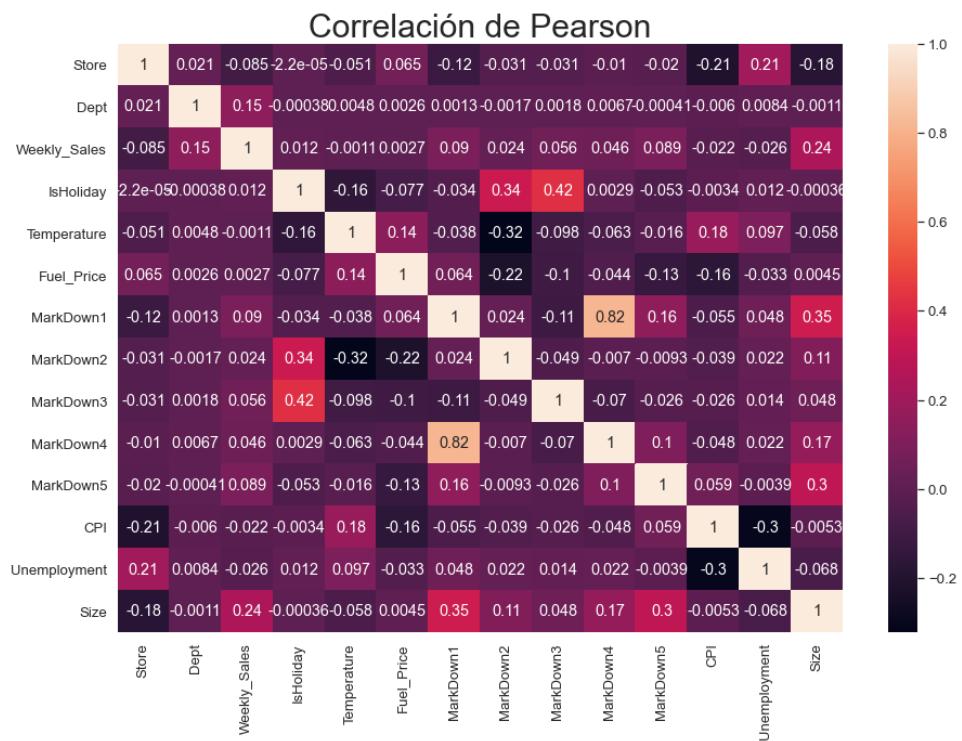


Imagen 15 Matriz de correlación de Pearson

A continuación, vamos a identificar si existe relación entre las siguientes variables:

- *Fuel\_price* y *CPI*
- *Markdown Y Weekly\_Sales*
- *Size y Type*

*CPI y Fuel\_Price*

Teniendo en cuenta el análisis de estas dos variables, creemos que pueden tener cierta correlación, ya que ambas tienen tendencias positivas en el tiempo. Para ello, primero vamos a ver la tendencia que presenta el *CPI* por tienda, como muestra la figura 16, en la que se puede claramente que la tendencia en cada tienda es igual, pero a diferentes escalas. Sabiendo que la tendencia es igual en todo las *Store*, procedemos a seleccionar una, y dibujar el *Fuel\_Price* y el *CPI* para dicha *Store*, como muestra la imagen 17.

### Evolución del CPI en el tiempo para cada tienda

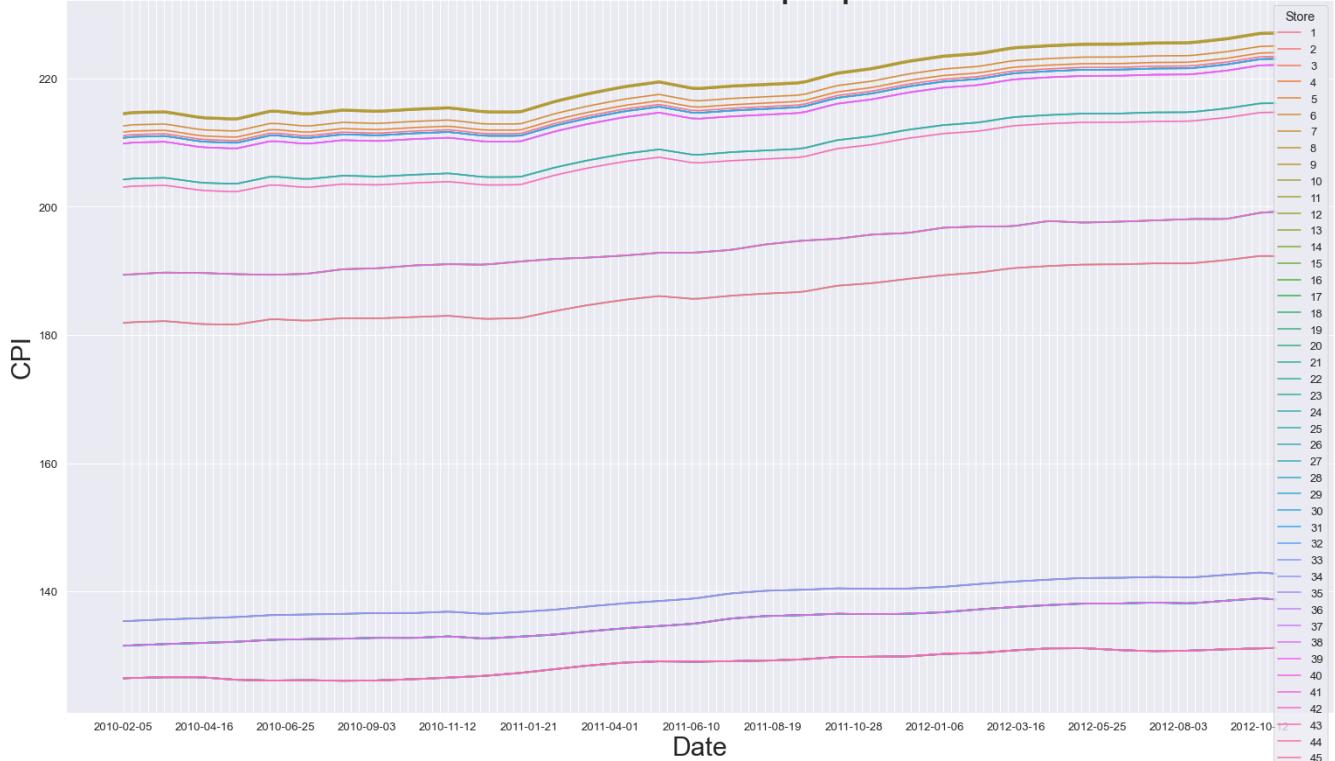


Imagen 16 Evolución CPI en el tiempo para cada tienda

## Evolución de CPI y Fuel\_Price en el tiempo para la tienda 26



Imagen 17 CPI y Fuel\_Price en el tiempo para la tienda 26

Como era de esperar, la tendencia de ambas es positiva, e incluso en enero de 2012, ambas *CPI* y *Fuel\_Price* han presentado una pequeña bajada. Para verificar que efectivamente existe dicha correlación, hemos dibujado la matriz de correlación para esta *Store* (26), imagen 18, en la que se puede ver que *Fuel\_Price* y *CPI* muestran una fuerte relación.

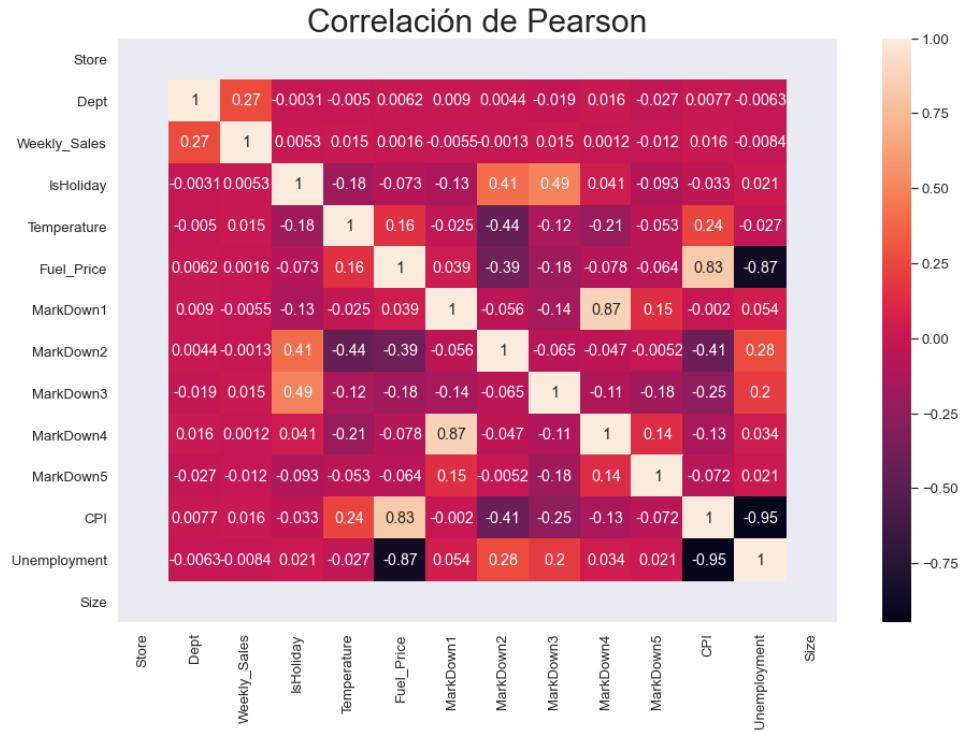


Imagen 18 Matriz de correlación para la Store 26

Por lo que uniremos el *CPI* y *Fuel\_Price* en un KPI para ponderar las ventas en el tiempo, ya que no es lo mismo vender un producto en 2010 que en el 2012, por la inflación.

#### Markdown Y Weekly\_Sales

Teniendo en cuenta que el *Markdown* representa descuentos, creemos que puede existir una correlación con *Weekly\_Sales*, donde en los picos mas altos del *Weekly\_Sales* estén también los picos del *Markdown*.

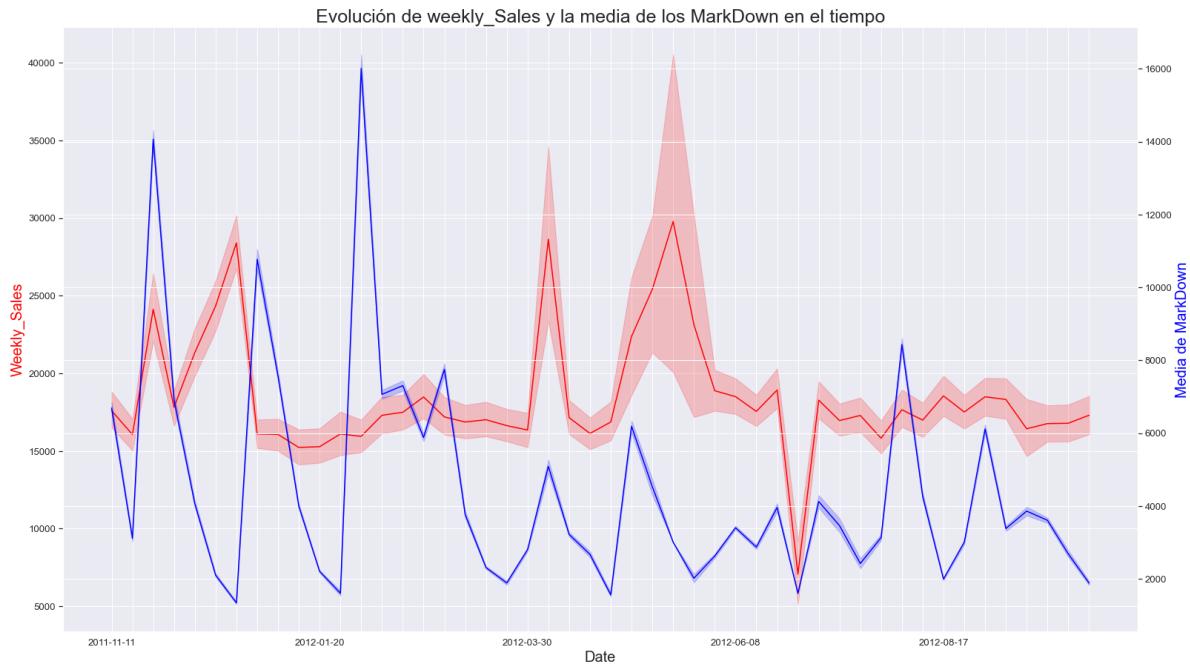


Imagen 19 Evolución de Weekly\_Sales y la media de los Markdowns en el tiempo

La relación entre estas dos variables no es tan significativa como esperábamos, pero si que podemos ver de la imagen 19, que en algunos picos como noviembre 2011 y marzo 2012, cuando el Markdown sube también suben las ventas semanales. Cabe destacar un pico negativo en ambos, en junio de 2012, donde se ve claramente que las ventas semanales se ven afectadas por la bajada de la media de los markdowns.

#### Size y Type

Como hemos visto en apartado anterior, imagen 12, el *Type* viene dado por 3 variables A, B y C, que pueden coincidir con los 3 grupos que hemos visto que hay en histograma de la imagen 13.

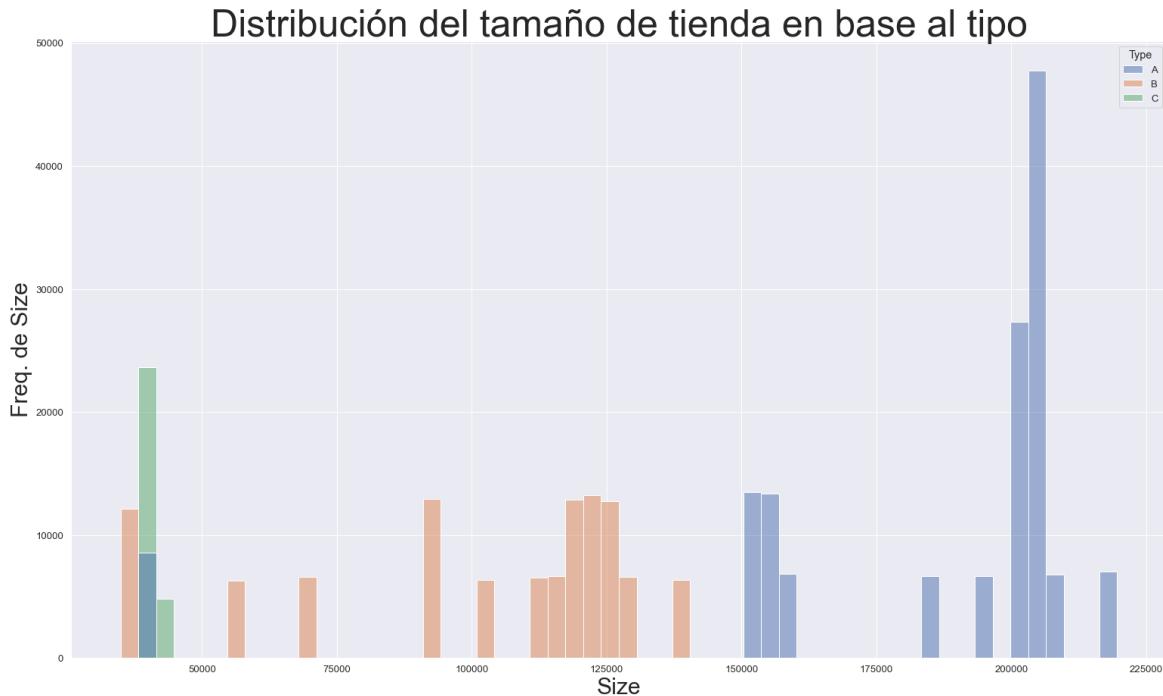


Imagen 20 Distribución del tamaño de la tienda en base al tipo de tienda

De la imagen 20 se puede ver que los 3 grupos que hemos visto en la sección de 2.1 de Size coinciden con los 3 tipos de tienda que hay.

### 2.3 – Preparación de set de datos para la Visualización

Ahora que conocemos los datos ha llegado el momento de decidir con qué datos nos vamos a quedar y qué transformaciones vamos a aplicar. Partiendo del modelo de datos original (Ilustración 1) transformaremos los mismo para conseguir el siguiente modelo:

Estado
Date
Store
store_weekly_sales
date_independent_store_weekly_sales
efficiency_index
CPI
Size

Ilustración 2 - Modelo de datos del Dashboard

El nuevo modelo es mas simple, mas fácil de manejar y está a un nivel de granularidad distinto. La nueva entidad, *Estado*, ya no tiene información a nivel de departamento, toda la información se ha agrupado a nivel de tienda y por fecha.

La nueva entidad tiene 3 columnas nuevas:

- *Store\_weekly\_sales* - Son las ventas agrupadas a nivel de fecha y tienda.
- *Date\_independent\_store\_weekly\_sales* - En un intento de mitigar la inflación en los precios con el tiempo (en el análisis hemos visto que el CPI aumenta con el tiempo), esta métrica intenta reflejar las ventas semanales de cada tienda

eliminando la inflación para que podamos comparar distintas fechas de manera justa. Se calcula dividiendo las store\_weekly\_sales por el CPI.

- *Efficiency\_Index* – Es un KPI que refleja cómo de eficiente es una tienda independientemente de la inflación y el tamaño, y permite comparar las tiendas de una manera justa. Se calcula dividiendo date\_independent\_store\_weekly\_sales por el tamaño (size) y multiplicando por 100. La multiplicación por 100 es para tener un rango de valores más fáciles de interpretar.

En el jupyter notebook adjunto en el paquete de entrega se puede ver el código python para generar el nuevo set de datos. Aquí presento una muestra:

```
[9]: final_data = final_data[['Date', 'Store', 'store_weekly_sales', 'date_independent_store_weekly_sales', 'efficiency_index', 'CPI', 'Size']]
final_data.head()
```

	Date	Store	store_weekly_sales	date_independent_store_weekly_sales	efficiency_index	CPI	Size
0	2010-02-05	1	1112466.82	5811.417603	3.840609	191.427789	151315
1	2010-02-05	2	1506524.45	7869.936029	3.890096	191.427789	202307
2	2010-02-05	3	358646.22	1873.532692	5.010517	191.427789	37392
3	2010-02-05	4	1090558.09	5696.968545	2.767359	191.427789	205863
4	2010-02-05	5	187551.77	979.752059	2.809325	191.427789	34875

Ilustración 3 - muestra de 5 registros del set de datos final

### 3. Visualización

Sobre el set de datos final he construido un dashboard de visualización utilizando Tableau Public que se puede visualizar, [utilizando Google Chrome](#), a través del siguiente enlace: [Visualización Final](#).

Si no fuera posible acceder mediante este enlace, también se puede encontrar la visualización accediendo al perfil de [mi usuario](#) y haciendo click en View:

The screenshot shows a user profile for 'Adrian Jose Zapater Reig'. Below the profile picture, it says '1 viz' and '0 following'. A link 'View' leads to a visualization titled 'VDD\_final' which has '0 views' and '1 star'. The visualization itself is a dark-themed dashboard with a 'FEATURED' section containing a chart and a 'View' button.

Ilustración 4 - Opción 2 de visualización

Por último, si fallan los 2 métodos anteriores, abrir el fichero VDD\_final.twbx, adjunto a la tarea, con la versión de escritorio de Tableau.

#### 3.1. Análisis de la visualización

En primer lugar, recordemos el objetivo de la visualización:

*Obtener información relevante del conjunto de datos para poder visualizar en un dashboard el histórico y estado actual de cada tienda.*

El objetivo busca tener una vista histórica y otra del estado actual, por lo que el resultado tiene 2 secciones principales: Semana en Curso y Datos Históricos.

#### *Semana en Curso:*

Semana en Curso					
Filtrar por Tienda	Total		Media		
	Coef. Eficiencia	Ventas Semanales		Avg. Eficiencia	Avg. Ventas Semanales
(Todo)	120	29,154,001		3	647,867
Tienda Mas Eficiente			Tienda con Mas Ventas		
Store	Coef. Eficiencia	Ventas Semanales	Store	Coef. Eficiencia	Ventas Semanales
43	3	265,616	20	4	1,484,552

*Ilustración 5 - Vista Semana en Curso*

Esta vista muestra el estado actual con 4 métricas y busca dar una visión rápida de todas o unas subselección de las tiendas. La subselección de tiendas se puede configurar con el desplegable de la izquierda.

He optado por utilizar la estructura mas básica, “la tabla”, porque la información que queremos trasmitir es muy poca (4 conjuntos de menos de 3 atributos) y queremos que se trasmite muy rápido.

He optado por utilizar 2 colores similares para dar sensación de unión y con connotación positiva, dorado y verde, para resaltar las métricas de las mejores tiendas. Los valores tienen un tamaño un poco mas grande y están coloreados para hacerlos destacar. Para resaltar el identificador de la tienda, que al final es un de los valores mas importantes, he optado por utilizar un fondo del mismo color. Todo esto hace ver que son métricas similares y positivas, pero indican información distinta.

#### *Datos Histórico*

La vista de Datos históricos esta dividida en 2 gráficas distintas que nos permiten analizar las tiendas en el tiempo en base al KPI que hemos diseñado: Coef. De Eficiencia (*Efficiency\_Index*).

## Datos Históricos



Ilustración 6 - Datos Históricos - visualización 1

La primera nos muestra, en orden, las tiendas mas eficientes en el rango de tiempo seleccionado. He optado por un rango de colores Verde-Rojo para mostrar las mejores tiendas en verde y las peores en rojo. Este rango de colores permite entender casi sin leer los ejes qué tiendas están desempeñando un papel mejor. Esta gráfica tiene como objetivo comparar todas las tiendas en una única visualización. Ha sido una labor difícil ya que no es fácil encontrar una gráfica que soporte mostrar 45 datos sin saturar al receptor. Creo que con esta visualización lo hemos conseguido.

Eficiencia de tienda contra la media de las tiendas en un rango de fechas

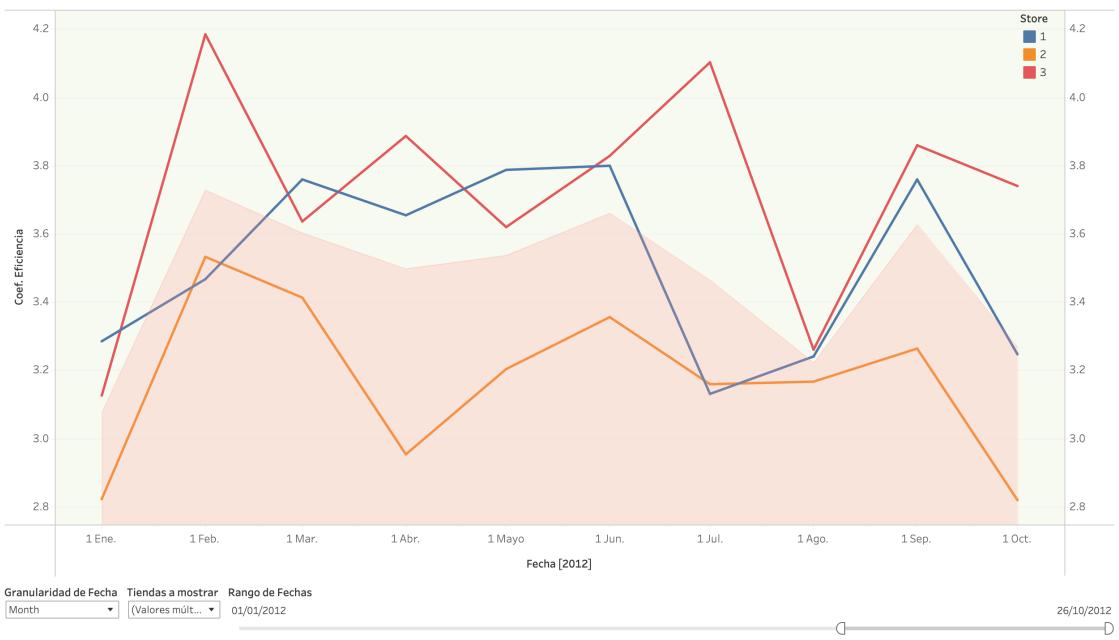


Ilustración 7 - Datos históricos - Visualización 2

La segunda gráfica tiene como objetivo ofrecer una manera de poder comparar N tiendas entre sí. Para poder adaptarse a distintos enfoques (comparativas entre años, comparativas del último mes, comparativas de las 6 tiendas mas eficientes) ofrece 3 parámetros configurables: Rango de tiempo, para seleccionar las fechas que queremos comparar, subselección de tiendas, para elegir qué tiendas mostrar y granularidad de

fecha, que permite comparar agrupando por año, mes o día dinámicamente sobre la misma gráfica. Con esto, tenemos una gráfica muy potente que nos va a permitir analizar comparativas entre tiendas en base a nuestro KPI.

En cuanto a la elección de colores, he optado por mostrar una forma de color rojo que ilumina la sección de la gráfica que está por debajo de la media de las tiendas seleccionadas y verde para la parte superior. Esto nos deja ver, sin tener que pensar mucho, que los puntos que estén por encima son buenos (mas eficientes) y los inferiores son malos (ineficientes).