PREDICCIÓN DE VENTAS DIARIAS Y SEMANALES DE LA CADENA DE SUPERMERCADOS TOSCOS.

PUNTOS CLAVE

---

20 Enero 2023

Mario Sabater Pascual

CONTEXTO

En el fichero '*datos.csv*' se encuentran las ventas ('*Sales*') de las diez tiendas de la cadena den supermercados TOSCOS.

Son datos diarios y además de las ventas tiene información sobre el número de clientes que han entrado en cada tienda ('*Customers*'), si está abierta ('*Open*'), si tienes promociones ('*Promo*'), si es fiesta ('*StateHoliday*') y si es día lectivo ('*SchoolHoliday*').

Las tiendas están divididas en 3 zonas (1, 2 y 3) y tienen 3, 3 y 4 tiendas respectivamente. Así la tienda T2c, es la tienda c de la zona 2.

OBJETIVO E INFORMACIÓN ESENCIAL

El objetivo es predecir las ventas desde el 1/08/2015 al 10/09/2015. Para esto se debe:

- Graficar ventas diarias de las 10 tiendas y de las 3 zonas.
- Graficar y predecir el horizonte establecido de ventas totales diarias.
- Graficar y predecir el horizonte establecido de ventas totales semanales.

PUNTOS CLAVE

1. Al tratarse de una predicción conjunta de las 10 tiendas, la información solicitada es altamente improbable que sea certera, ya que existen muchos factores que puedan alterar las ventas de una tienda o de varias.

2. Para la predicción diaria los días de cierre de las tiendas, los modelos utilizados no son capaces de predecir un cero absoluto, en su lugar predicen valores cercanos al cero o incluso negativos. Para evitar esto, una vez terminada la predicción, cambiamos estos valores a cero, ya que no tiene sentido dejar valores de venta negativos.

3. Ninguna de las dos series (ventas semanales y ventas diarias) presentan una tendencia clara. Nuestros modelos de predicción, para el horizonte solicitado, presentan una línea continuista de las ventas respetando las estacionalidades y los, mencionados anteriormente, días de cierre.

PREDICCIÓN DE VENTAS DIARIAS Y SEMANALES DE LA CADENA DE SUPERMERCADOS TOSCOS.

RESUMEN EJECUTIVO

20 Enero 2023

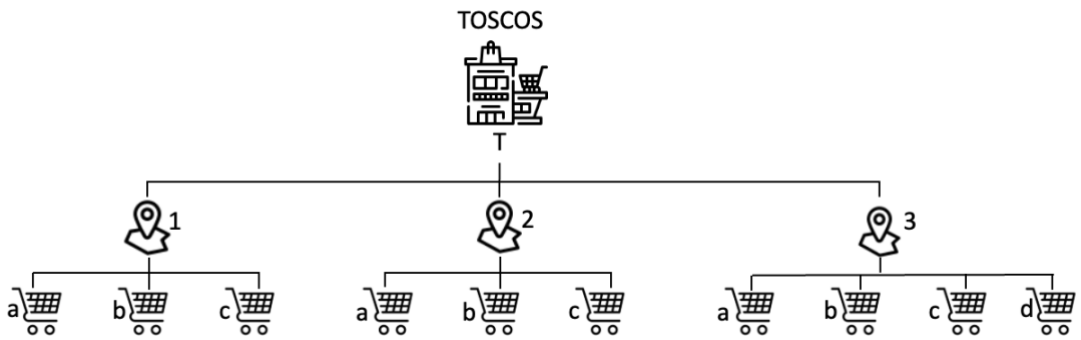
Mario Sabater Pascual

CONTEXTO

En el fichero '*datos.csv*' se encuentran las ventas ('*Sales*') de las diez tiendas de la cadena den supermercados TOSCOS. Este fichero, comprende un intervalo de tiempo desde el 1 de enero de 2013 hasta el 31 de julio de 2015.

Son datos diarios y además de las ventas tiene información sobre el número de clientes que han entrado en cada tienda ('*Customers*'), si está abierta ('*Open*'), si tienes promociones ('*Promo*'), si es fiesta ('*StateHoliday*') y si es día lectivo ('*SchoolHoliday*').

Las tiendas están divididas en 3 zonas (1, 2 y 3) y tienen 3, 3 y 4 tiendas respectivamente. Así la tienda T2c, es la tienda c de la zona 2. Por lo que el diagrama de la red de supermercados quedaría de la siguiente forma:



OBJETIVO E INFORMACIÓN ESENCIAL

El objetivo es predecir las ventas desde el 1/08/2015 al 10/09/2015. Para esto se debe:

- Graficar ventas diarias de las 10 tiendas.
- Graficar ventas diarias de las 3 zonas.
- Graficar las ventas totales diarias.
- Graficar las ventas totales semanales.
- Predecir las ventas totales diarias.
- Predecir las ventas totales semanales.

PRINCIPALES CONCLUSIONES

1. GRAFICAS DE VENTAS DIARIAS POR TIENDA Y POR ZONA.

Debido al número de tiendas y a la similitud de la serie entre ellas, es necesario graficar de manera individual cada una de ellas. Al hacer esto descubrimos que todas las tiendas cierran una vez por semana y que algunas de ellas están cerradas durante un periodo de tiempo largo. Aunque no sabemos a qué se debe, podemos entender que es a reformas si se encuentra a mitad de la serie, o a que aún no estaba abierta al público si es al principio (como es el caso de T1c).

Observamos también que tras un periodo de cierre el primer día o primeros días aumentan las ventas exponencialmente, pero se trata de un aumento puntual.

Con respecto a la tendencia, la gran mayoría de las tiendas no tienen una tendencia marcada, salvo T2b y T3c que tienen una tendencia creciente lineal no muy pronunciada, a simple vista para el resto diríamos que no tienen tendencia.

Existe también una estacionalidad más o menos clara con un patrón semanal de lunes a domingo, siendo el primer día de la semana el de mayor venta y el sábado el de menos venta. Como hemos mencionado anteriormente, los domingos las tiendas se encuentran cerradas.

2. GRAFICAR Y PREDECIR LAS VENTAS TOTALES DIARIAS.

De manera global, las ventas totales diarias no se observa una tendencia clara. Se observa un ligero crecimiento durante el año 2013, posiblemente también producido por la apertura de la tienda T1c. Esta tendencia de crecimiento vemos que no continua en 2014, donde principalmente a partir del segundo semestre se observan los valores de ventas más bajos. Durante este periodo, la tienda T3b permaneció cerrada, y siendo esta una con las ventas medias más altas, es posible que sea el detonador de esta bajada en los niveles de ventas.

A la hora de predecir, debido a la longitud de la serie y a la frecuencia de los datos hemos decidido prescindir de los valores de 2013, lo cual hace la serie algo más compacta y evitamos posibles problemas de *overfitting*.

Con respecto a la predicción, el modelo *Prophet* ofrece la mayor precisión llegando a niveles de R^2 mayores del 0.88, con un *split train-test* de 80-20 respectivamente. Pero para llegar a estos niveles hemos de añadir manualmente la estacionalidad y dos regresores o variables exógenas a este, siendo *StateHoliday* y *Promo*.

Sobre la estacionalidad, además de la semanal ya comentada, al modelo le incluimos una estacionalidad mensual (donde la segunda y cuarta semana del mes disminuyen los valores de venta) y una anual.

Con respecto a este primer regresor, *StateHoliday* queríamos corregir principalmente el efecto de las tiendas los domingos y los días festivos en los que se encuentran cerradas las tiendas. También hemos probado directamente con la variable *Open*, tomando como *Holiday* los días que estaban las tiendas cerradas.

Pero a pesar dar un buen resultado, perdíamos información ya que había día que había abiertas 9 de 10 tiendas. El otro regresor utilizado, *Promo*, ayuda al modelo a predecir los picos de mayores ventas. Observando esta variable, vemos que las promociones suelen ser semanales intermitentes. Debido a que durante la última semana de la serie había promociones, decidimos establecer los valores de promoción intermitentemente, comenzando nuestra predicción con sin promoción los primeros 9 días (ya que los sábados no hay promoción).

3. GRAFICAR Y PREDECIR LAS VENTAS TOTALES SEMANALES.

Las ventas semanales, aunque a priori pueda parecer que ofrecen menos información, también ofrecen menos ruido a la hora de observar las tendencias y los cambios en la serie. Podemos observar mejor esta tendencia alcista amortiguada, aunque a final de año encontremos un valor atípico siendo un máximo en la serie. Este valor corresponde la última semana del año y está principalmente provocado por las ventas de las zonas 1 y 3.

También en este gráfico semanal observamos el ya comentado efecto de la tienda T3b que, siendo una de las tiendas con mayor volumen de ventas, permanece cerrada durante el segundo semestre de 2014.

Esta vez, debido a que la serie semanal ofrece menos valores que la diaria, hemos seleccionado el total de la serie, teniendo un total de 135 valores. Para el *train-test split*, decidimos 120 y 15 valores respectivamente. A los cuales le aplicamos un modelo ETS. El mejor modelo en nuestro caso es un ANA, es decir, error aditivo, tendencia nula y estacionalidad aditiva anual.

El score de referencia, el R^2, es peor que el obtenido en las ventas diarias, pero hemos de tener en cuenta que en este caso no solo estamos añadiendo las ventas de las diferentes tiendas, sino también de una misma semana. En estos casos, se suele pierde precisión.

CONCLUSIONES FINALES.

Tal y como acabamos de comentar, las predicciones y la precisión de estas varía drásticamente a pesar de tratarse de los mismos datos. De igual forma las predicciones son continuistas, en ambos casos.

PREDICCIÓN DE VENTAS DIARIAS Y SEMANALES DE LA CADENA DE SUPERMERCADOS TOSCOS.

REPORTE COMPLETO

20 Enero 2023

Mario Sabater Pascual

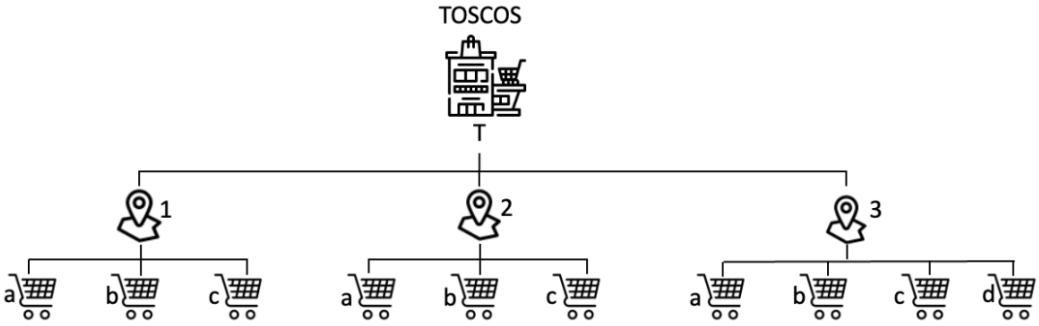
TABLA DE CONTENIDO

I. CONTEXTO

En el fichero '*datos.csv*' se encuentran las ventas ('*Sales*') de las diez tiendas de la cadena den supermercados TOSCOS. Este fichero, comprende un intervalo de tiempo desde el 1 de enero de 2013 hasta el 31 de julio de 2015.

Son datos diarios y además de las ventas tiene información sobre el número de clientes que han entrado en cada tienda ('*Customers*'), si está abierta ('*Open*'), si tienes promociones ('*Promo*'), si es fiesta ('*StateHoliday*') y si es día lectivo ('*SchoolHoliday*').

Las tiendas están divididas en 3 zonas (1, 2 y 3) y tienen 3, 3 y 4 tiendas respectivamente. Así la tienda T2c, es la tienda c de la zona 2. Por lo que el diagrama de la red de supermercados quedaría de la siguiente forma:



II. OBJETIVO E INFORMACIÓN ESENCIAL

El objetivo es predecir las ventas desde el 1/08/2015 al 10/09/2015. Para esto se debe:

- Graficar ventas diarias de las 10 tiendas. Debido a la forma de la serie y a la similitud de los datos entre tiendas, no podemos hacer una misma grafica para todas las tiendas. Para poder visualizar correctamente los datos hemos de realizar una gráfica por tienda. *(Información incluida en 1. FINAL_ST_HW_MSP)*

- Graficar ventas diarias de las 3 zonas. Similar al caso anterior, a pesar de tratarse de únicamente 3 series, al tener valores de cero de manera repetida, dificulta su visualización, por tanto, realizamos una gráfica por zona. *(Información incluida en 1. FINAL_ST_HW_MSP)*.

- Graficar las ventas totales diarias. Sumadas todas las ventas de las tiendas, y previamente a predecirlas, graficaremos las ventas totales, analizando

los valores nulos, los outliers, así como la tendencia y estacionalidad. *(Información incluida en 2. FINAL_ST_HW_MSP).*

- Graficar las ventas totales semanales. Sumando las tiendas y cambiando la serie de una diaria a semanal, graficaremos las ventas totales, analizando los outliers, así como la tendencia y estacionalidad. *(Información incluida en 3. FINAL_ST_HW_MSP).*

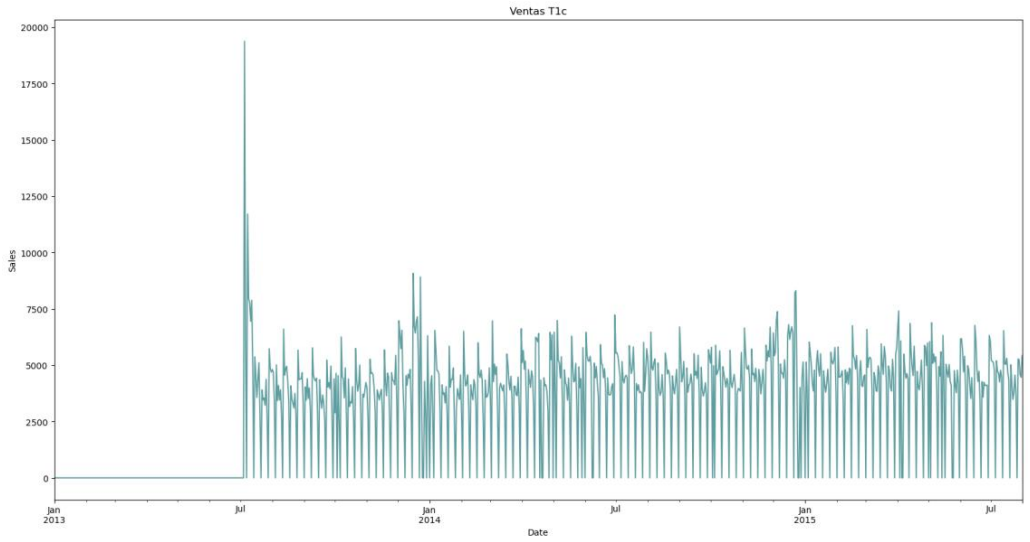
- Predecir las ventas totales diarias. Probamos diferentes modelos buscando aquellos que tienen el mejor score en la fase de entrenamiento. Posteriormente aplicaremos este modelo para predecir nuestro horizonte de 41 días. *(Información incluida en 2. FINAL_ST_HW_MSP).*

- Predecir las ventas totales semanales. Probamos diferentes modelos buscando aquellos que tienen el mejor score en la fase de entrenamiento. Posteriormente aplicaremos este modelo para predecir nuestro horizonte de 6 semanas. *(Información incluida en 3. FINAL_ST_HW_MSP).*

III. VISUALIZACIÓN Y ESTADISTICAS DE LOS DATOS.

3.1 VENTAS DIARIAS POR TIENDA.

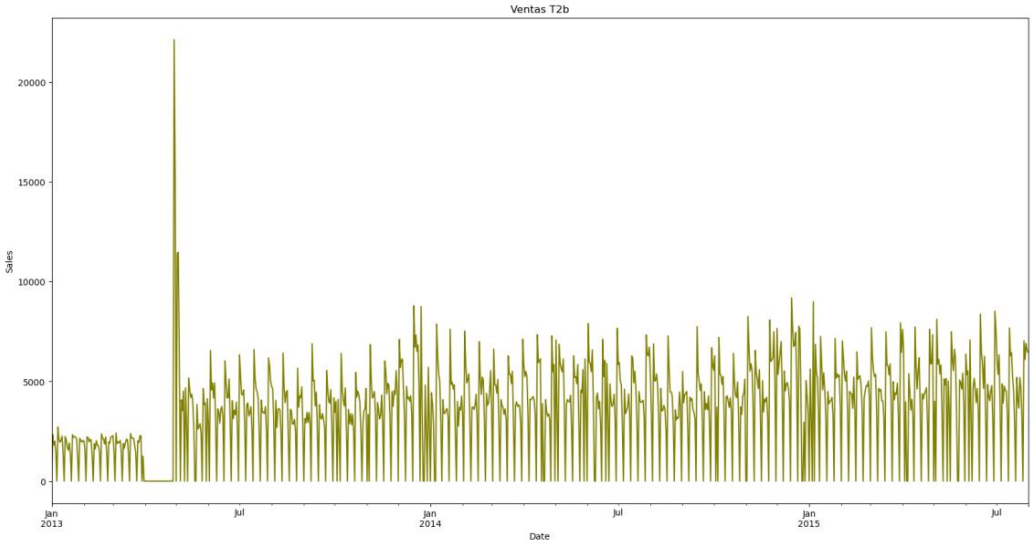
Graficamos las ventas diarias por tiendas, a modo de simplificación plotearemos y comentaremos las que a priori parecen más interesantes. En el anexo 1, podremos encontrar las gráficas de las ventas diarias de todas las tiendas.



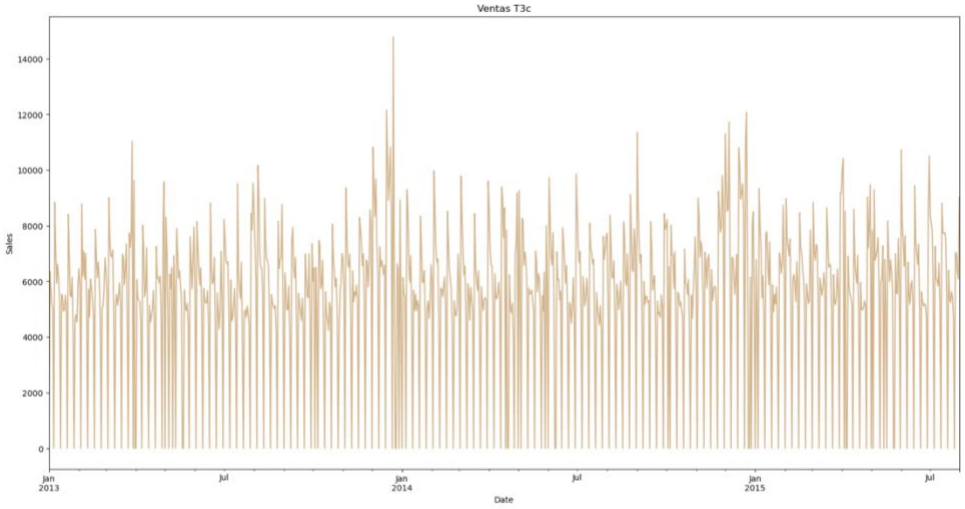
En la gráfica de las ventas correspondientes a la tienda T1c, observamos que se encontraba cerrada en los primeros seis meses de 2013. En el momento de apertura, provocado también por la novedad la tienda llega a facturar cerca de los

20.000 u.m. (unidades monetarias). Tras la primera semana las ventas de la tienda se estabilizan y podemos observar ya una estacionalidad semanal, así como los valores 0 cuando la tienda está cerrada.

Tanto en está gráfica como en el resto de las gráficas de ventas diarias podemos observar que no solo las tiendas tienen un cierre más o menos semanal (posteriormente sabremos que se trata de los domingos) si no que también existen días festivos en los que las tiendas no abren (algo especialmente visible en época de navidad).



A diferencia de la tienda T1c, la T2b estaba operativa desde el principio, pero en abril de 2013, se cierra. Posterior al cierre y al igual que en el gráfico anterior, observamos que las ventas de los primeros días se multiplican exponencialmente, en este caso superando cómodamente las 20.000 u.m. Cabe destacar que, tras el periodo de cierre, las ventas de la tienda aumentan considerablemente comparadas con los valores previos al cierre, por lo que podemos entender que se trata de una ampliación o una renovación de la tienda.

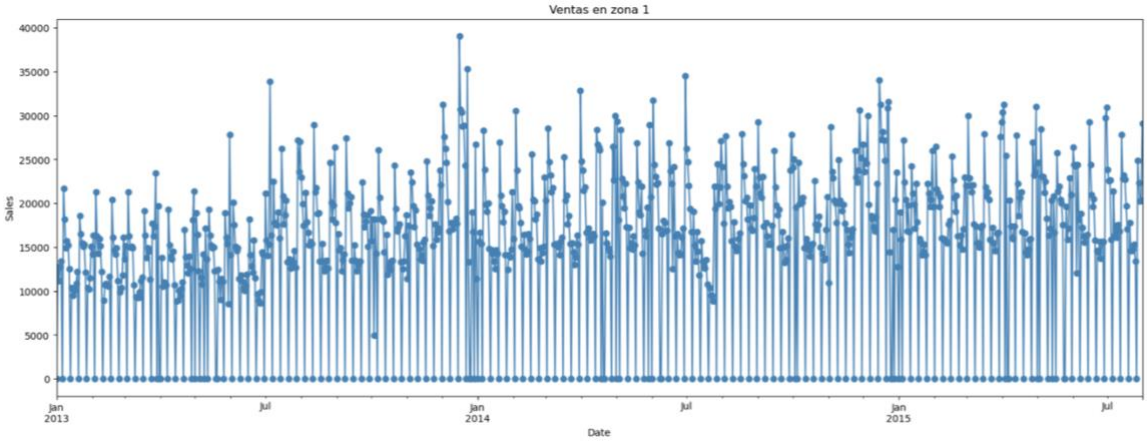


En este caso nos encontramos con una tienda que no ha sufrido ningún periodo continuado de cierre, de esta manera podemos observar de mejor manera lo que

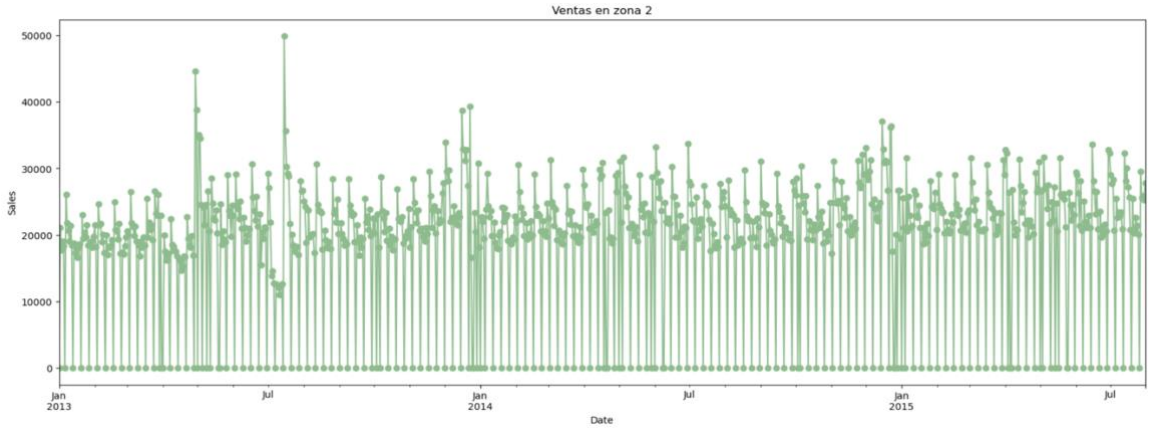
sería una serie "normal" o "típica". En este caso no observamos tendencia ninguna en la serie y podemos observar los días festivos de mejor forma. De manera consistente solo podemos ver los periodos de abril-mayo (probablemente coincidiendo con Semana Santa / *Easter*) así como el ya mencionado el periodo de Navidad.

3.2 VENTAS DIARIAS POR ZONA.

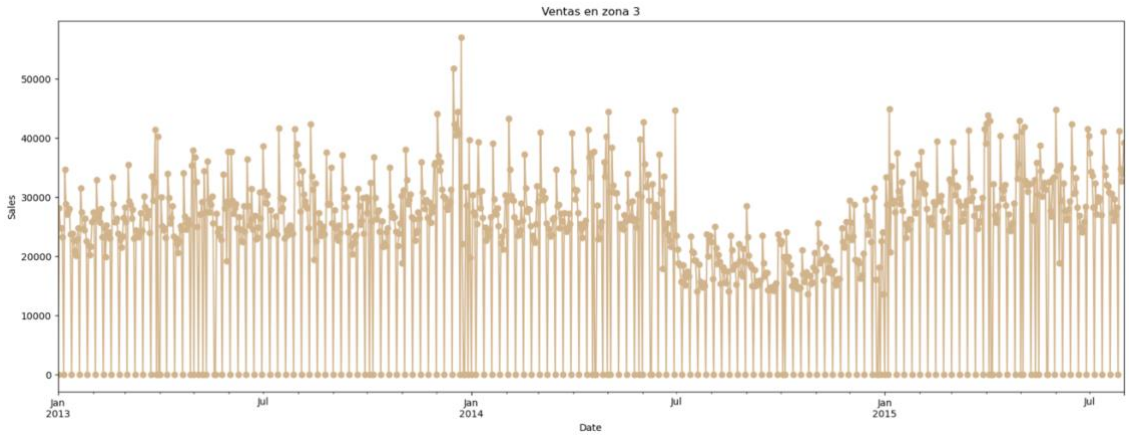
En este caso de ventas por zona, debido a que tenemos únicamente tres zonas, procedemos a hacer un breve estudio de las tres. A modo de datos relativos hemos de tener en cuenta que las dos primeras zonas cuentan con tres tiendas cada una, mientras que la zona 3 cuenta con cuatro tiendas diferentes.



Como ya hemos visto anteriormente, en la zona 1, la tienda T1c se abre en julio de 2013, de ahí que podamos observar unos valores de ventas menores en los seis primeros meses de la zona 1. Con respecto al resto, no observamos valores excesivamente notales, a excepción de la bajada en julio de 2014 correspondiente al cierre de la tienda T1b. A excepción de esto, observamos una serie bastante estable con una estacionalidad muy marcada, mientras que la tendencia, a excepción del primer año donde si se observa una tendencia creciente, debido por la apertura de la nueva tienda, en el resto de la serie, es bastante estable.

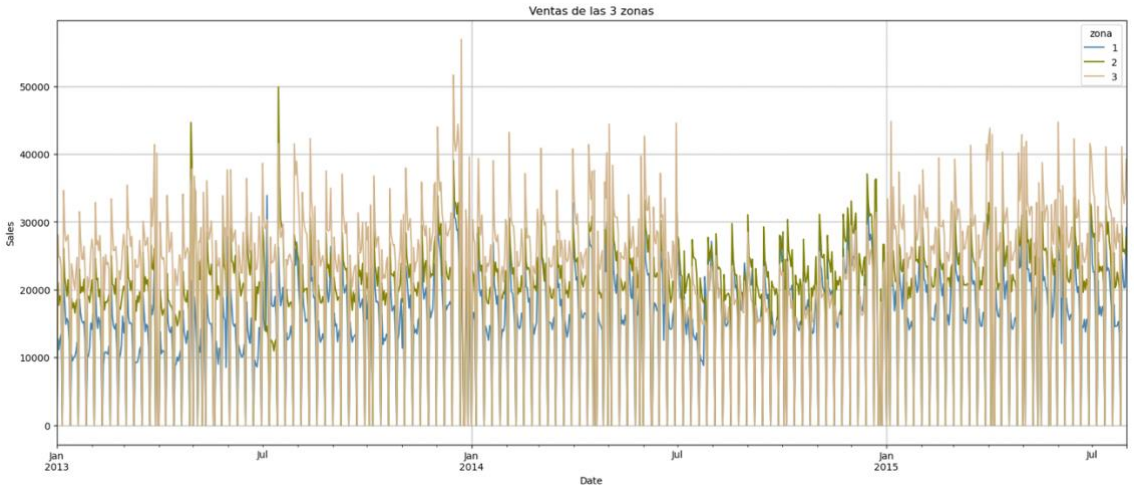


En la serie de ventas totales diarias de la zona 2, podemos observar los efectos de la ya mencionada T2b. También en la zona 2, la tienda T2c pasa por un proceso similar en Julio de 2013 de ahí la bajada a principios de ese mes y la subida puntual automáticamente después. A parte de estos eventos, de la serie no podemos destacar más, ya que se observa una estacionalidad muy similar a la de la zona 1 y sin tendencia.



Finalmente, en la zona 3 observamos casos similares a los anteriores, pero a diferencia de estos, durante el segundo semestre de 2014, una de las tiendas con mayor facturación media, T3b (*ver anexo 1*), cierra sus puertas. Salvo esta excepción, nos encontramos una serie análoga a las anteriores.

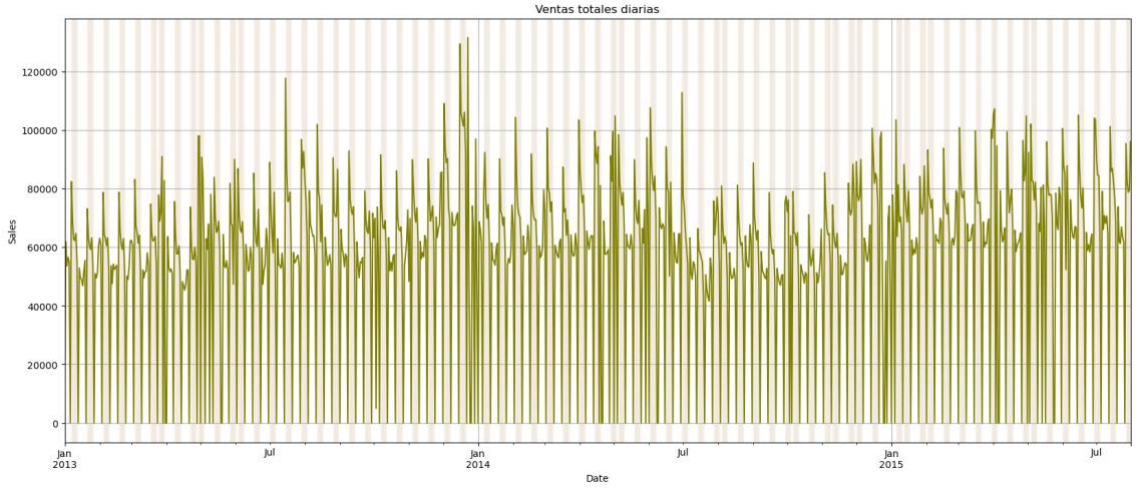
Globalmente y comparando los niveles de ventas por zonas, observamos que la zona 3 es la de mayor volumen, pero como ya hemos comentado anteriormente, también tiene una tienda más que las otras dos. La zona 2, que factura algo por debajo de la zona 3 con una tienda menos factura alrededor de un 50% que la zona 1 con el mismo número de tiendas. Sin saber las características de las zonas ni de las tiendas no podemos llegar a mayores conclusiones en este aspecto.



3.3 VENTAS DIARIAS TOTALES.

A la hora de graficar las ventas totales diarias hemos tenido en cuenta no solo las ventas brutas si no también otra de las variables que nos proporciona el dataset, *Promo*. Desde el primer momento de realizar el análisis era evidente que las variables *extra* que nos eran proporcionadas iban a ser de gran utilidad (al menos algunas de ellas) ya que la gran mayoría están correlacionadas de una manera u otra con las ventas.

Para la representación gráfica de las ventas diarias hemos escogido representar también la variable promo a través de un sombreado en aquellos días en los que Toscos activa la promoción. Esta promoción es activada de manera simultánea en todas las tiendas que forman parte del estudio al mismo tiempo.



Observando el dataset podemos obtener al menos tres conclusiones:

I. Que los días de promoción corresponden con los picos de mayores ventas.

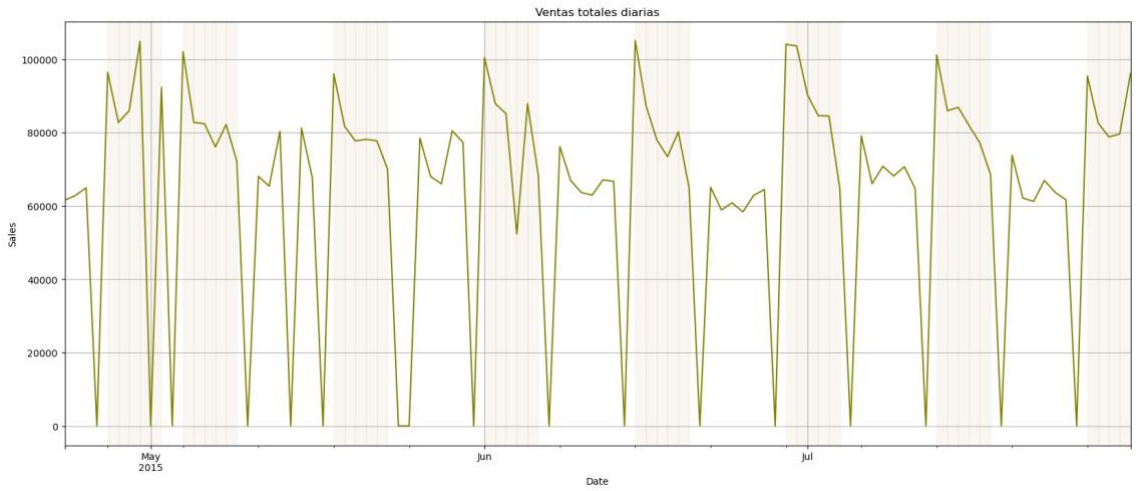
II. Las promociones tienen una periodicidad más o menos regular, aunque cada x tiempo hay dos semanas de promoción seguidas.

III. A pesar de no ser evidente a primera vista, existe una tendencia alcista general, aunque también puede venir dado porque, en los últimos 7 meses de la serie, todas las tiendas estaban disponibles.

El hecho de observar que los mayores picos de ventas correspondan con valores de una variable exógena nos ayuda a entender la relación entre las mismas, así como el poder determinar que estos valores no los podemos considerar como atípicos. Un valor atípico en este caso correspondería más con las ventas de una

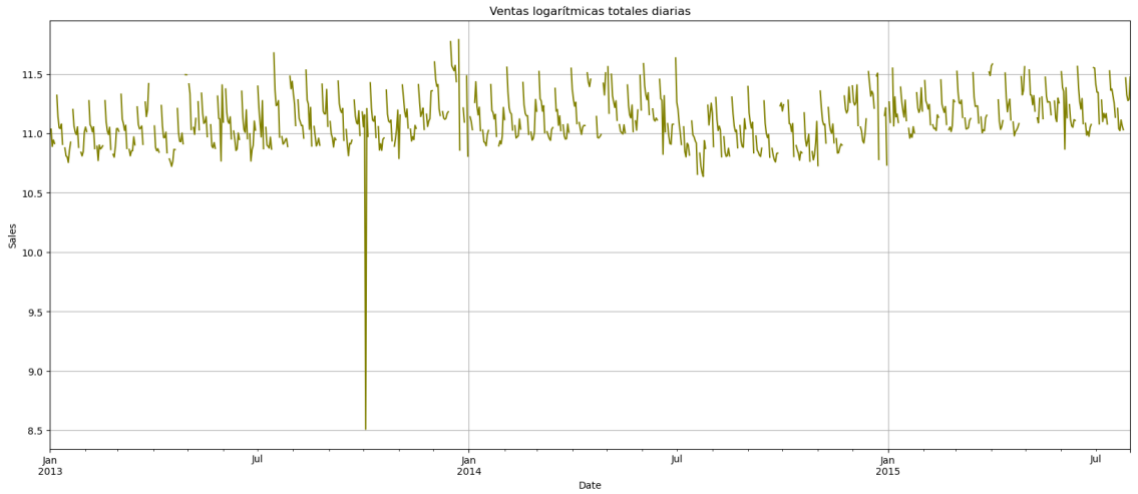
de las tiendas tras un periodo de renovación, pero en la serie global es más difícil de observar estos valores.

A continuación, procedemos a mostrar un zoom de los últimos 100 días, los cuales nos proporcionará una mejor visión, así como entendimiento, de los valores de ventas diarios.



Como hemos comentado ya previamente, observamos que las promociones salvo una vez cada x tiempo, ocurren en semanas intermitentes, lo que ahora es más fácil de ver es que estas no duran de lunes a domingo, sino solo los primeros cinco días de la semana. También gracias al zoom observamos la diferencia de valores medios que encontramos en el valor de las ventas en semanas contiguas con y sin promoción, donde de manera aproximada hay unas 10.000 u.m. de ventas más en semanas de Promo por día.

También, se intuye la existencia de una estacionalidad semanal y otra mensual. Más adelante cuando tratemos los modelos observaremos estas estacionalidades. Para el estudio de la estacionariedad, sin embargo, hemos procedido a realizar la gráfica de la serie logarítmica de ventas totales diarias.



Observamos en la serie logarítmica un claro outlier, el 3 de octubre de 2013. Esta fecha es considerada *State Holiday* de tipo "a", pero sin conocer muy bien el motivo T1b es la única tienda de las 10 que forman parte del estudio de Toscos, que ha abierto en un día festivo. Ese día, la tienda en cuestión alcanzó unas ventas de casi 5.000 u.m.

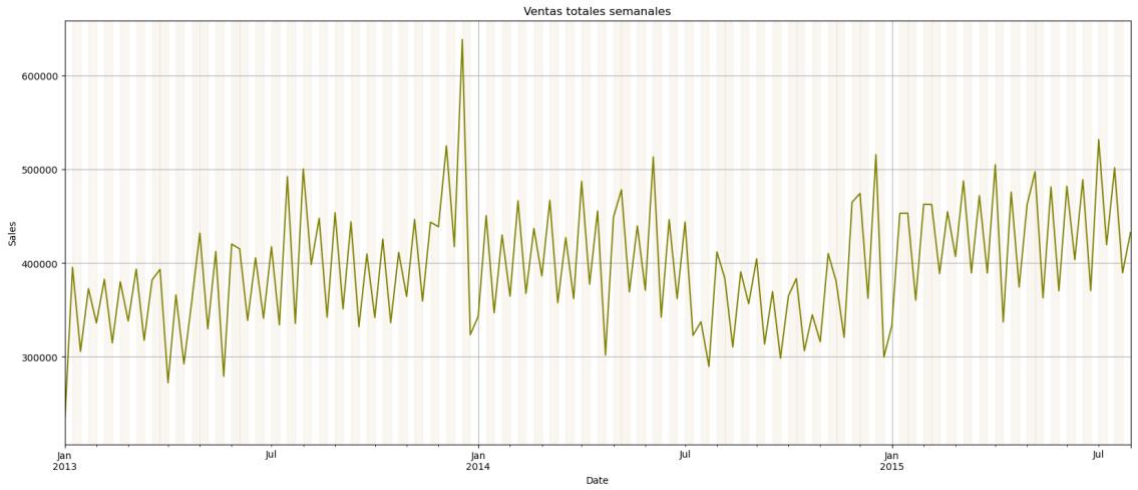
A pesar de formar parte de la serie, como ya hemos mencionado lo consideramos outlier y no lo tendremos en cuenta para nuestro estudio de la estacionariedad. Con respecto al resto de la serie, se observa bastante estable, y salvo en algunos casos puntuales como los ya mencionados como las navidades de 2013-14 o el cierre de la tienda T3b (que también se puede apreciar en la serie logarítmica) establecemos que nos encontramos ante una varianza y media prácticamente constantes. Por lo que existe estacionariedad.

3.4 VENTAS SEMANALES TOTALES.

Para la serie temporal de ventas semanales totales, ya no estudiaremos la estacionariedad de esta. En su lugar haremos una breve observación de los datos. Ya que no deja de ser la misma serie, pero acumulada por semanas.

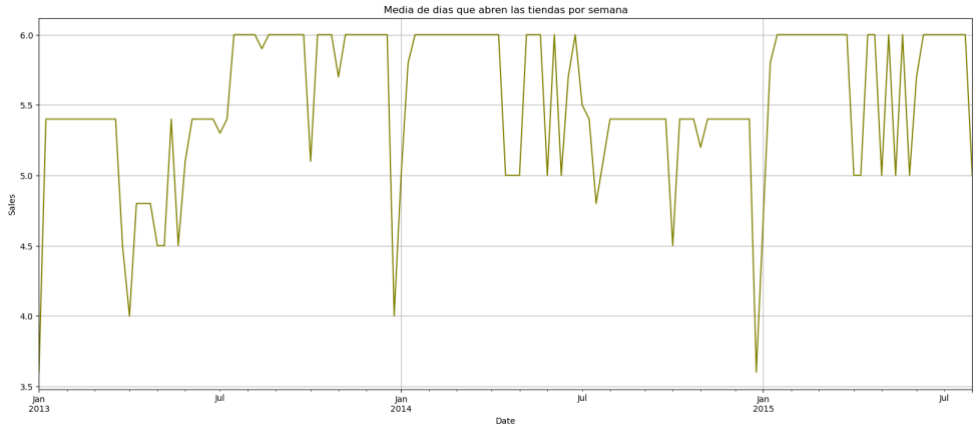
Para una mayor comprensión del trabajo realizado en las ventas semanales totales, hemos acumulado las ventas por semana y se ha elegido a modo de referencia el domingo. A pesar de que estos días las tiendas no se encuentran abiertas y no reciben ventas, se entiende como mejor opción para el entendimiento de una serie temporal semanal.

De nuevo en esta serie temporal hemos graficado las ventas con respecto a los días de promoción. Aunque la promoción no dura la semana completa, el impacto que esta tiene en 5 de los 6 días en los que está activada, se entiende como suficiente como para ser graficada.



En esta serie semanal, debido a la existencia de menos ruido por los días en los que no había ventas podemos observar de manera más clara las tendencias y variaciones a nivel general de la serie. Destacamos una vez más las navidades de 2013, así como el cierre de T3b que sigue siendo patente en las ventas semanales durante el segundo semestre de 2014.

Con respecto a la ya mencionada tendencia, seguimos encontrándonos dificultades para determinarla, ya que no hay grandes periodos de todas las tiendas abiertas a la vez. A continuación, mostramos un gráfico que nos muestra la media de tiendas abiertas por semana. El valor 6 es el máximo, ya que la serie está dividida por 10 (el nº de tiendas) y todos los domingos las tiendas están cerradas.



Es por este motivo por el cual no podemos determinar una tendencia clara de la serie.

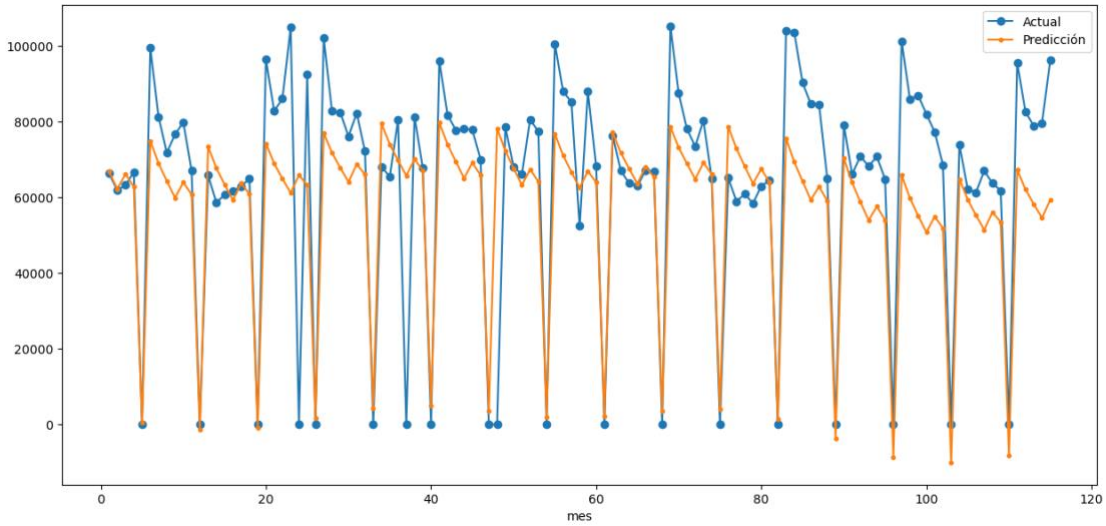
IV. PREDICCIÓN DEL HORIZONTE ESTABLECIDO.

4.1 PREDICCIÓN VENTAS DIARIAS TOTALES.

A la hora de predecir, debido a la longitud de la serie y a la frecuencia de los datos hemos decidido prescindir de los valores de 2013, lo cual hace la serie algo más compacta y evitamos posibles problemas de *overfitting*. En lo que respecta a la selección del modelo a utilizar para las predicciones diarias, debido al tipo de dataset, existían varias opciones.

La primera y que a priori parecería más lógica sería predecir la serie a través de un modelo jerárquico. Este nos permitiría realizar un modelo de abajo hacia arriba prediciendo tienda por tienda y sumando todas las predicciones. El problema que encontramos es la ineficiencia de este modelo y la limitación de este a la hora de usar modelos para la predicción individual.

Otra opción sería utilizar el *Neural Prophet*. Este modelo basado en redes neuronales nos permitiría incluir varias estacionalidades al modelo, mejorando así su rendimiento. Tras probar este obtenemos un R^2 de 0,62. Tras graficar los resultados del *Neural Prophet* un train-test de 80-20, observamos que predice por debajo de lo real y que los días de promoción predice muy por debajo.



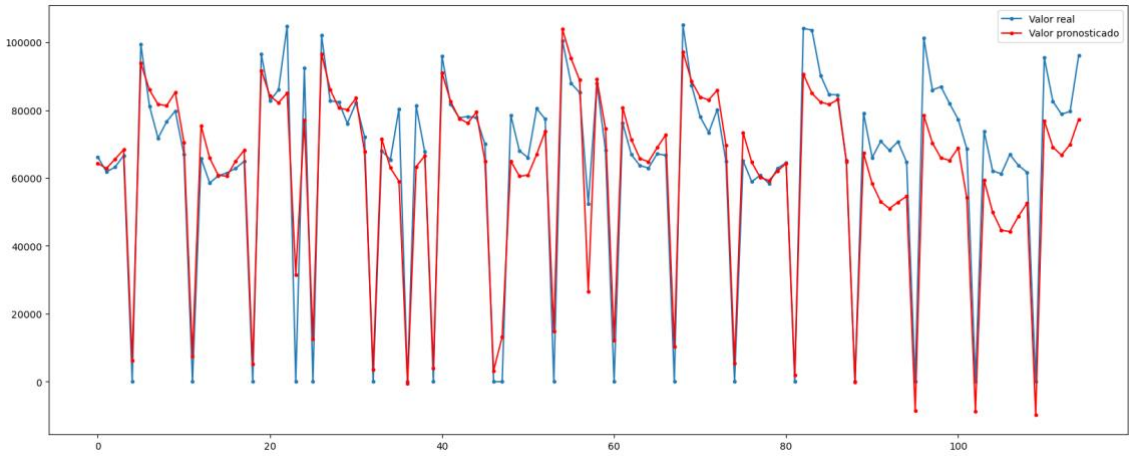
Tras este hallazgo sobre los días con promoción, se entiende la importancia de las variables exógenas a la hora de realizar una predicción más consistente con el resto de variable. A través de un *Prophet*, podemos incluir no solo varias estacionalidades como en el *Neural Prophet*, sino que además podemos incluir variables exógenas como *Promo* y *StateHoliday*.

Nuestro modelo *Prophet* con mejor R^2 (>0.88) y de nuevo un *train-test split* de 80-20 es el siguiente:

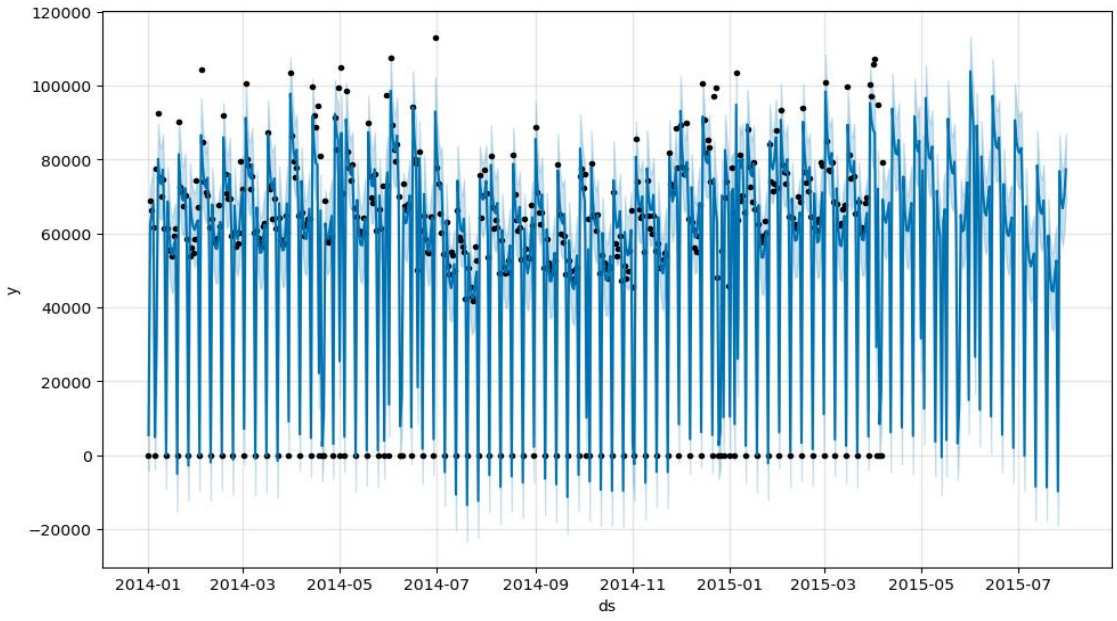
```
m5 = Prophet(holidays=holidays)
m5.add_regressor('Promo')
m5.add_seasonality(name='weekly', period=7, fourier_order=9)
m5.add_seasonality(name='monthly', period=30.5, fourier_order=5)
m5.add_seasonality(name='yearly', period=365.25, fourier_order=9)

m5.fit(train5)

future5 = m5.make_future_dataframe(periods=len(test))
future5 = future5.merge(X_promo, left_on = 'ds', right_on = 'fecha',
                        how = 'left')[['ds', 'Promo']].fillna(0)

forecast5 = m5.predict(future5)

print(r2_score(list(test['y']), list(forecast5.loc[round(len(y)*0.8):,'yhat'] )))
print(mean_absolute_error(test['y'].values, forecast5['yhat'][round(len(y)*0.8):].values))
```

En este caso, hemos utilizado como regresor la ya mencionada columna *Promo* y como la variable holidays (que permite introducir *Prophet* en el modelo) la columna *StateHoliday*, indistintamente de si esta tienes valores de "a", "b" o "c". Con respecto a la estacionalidad, y en base a lo aprendido observando las gráficas, el modelo mejora cuando le incluimos tres estacionalidades: Semanal, Mensual y Anual. (*Ver anexo 2*)

Las predicciones del test de este modelo son las siguientes:



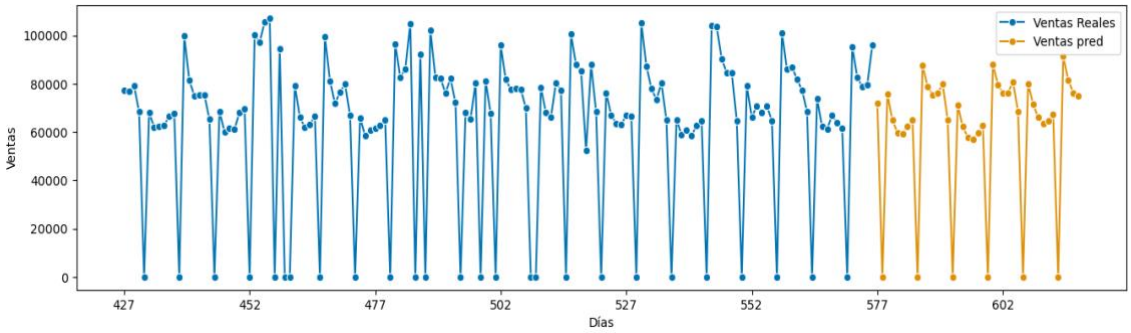
Como es natural en cualquier predicción, las primeras predicciones son más certeras, pero a lo largo que pasa el tiempo la predicción es menos certera. Observando la predicción completa, podemos observar cómo se comporta el modelo a lo largo de toda la serie.



Finalmente, utilizamos este modelo para predecir nuestro horizonte hasta el 10 de septiembre de 2015. Para ello, incluimos todo el dataset (desde 2014) como entrenamiento y predecimos 41 días. Para los valores de *Promo*, debido a que no conocemos esos datos en el horizonte de predicción procedemos a estimarlos.

Previamente, se ha comentado que, en la gran mayoría de la serie, los valores de Promo aplicaban para los días de diario (lunes a viernes) de semanas intermitentes. Se observa también que la última semana se la serie incluye promoción, por lo que parece lógico y una buena estimación que la promoción comience la segunda semana de agosto.

Tras incluir estos datos procedemos a predecir, obteniendo lo siguiente:



Los valores de esta predicción son los siguientes:

DS	yhat
1-8-2015	71.721,76
2-8-2015	0,00
3-8-2015	75.597,21
4-8-2015	64.861,21
5-8-2015	59.740,97
6-8-2015	59.271,10
7-8-2015	62.298,72
8-8-2015	65.025,12
9-8-2015	0,00
10-8-2015	87.643,38
11-8-2015	78.752,71
12-8-2015	75.415,42
13-8-2015	76.134,43
14-8-2015	79.838,66

DS	yhat
15-8-2015	65.126,14
16-8-2015	0,00
17-8-2015	71.108,55
18-8-2015	62.180,28
19-8-2015	57.864,17
20-8-2015	57.060,83
21-8-2015	59.527,37
22-8-2015	62.530,40
23-8-2015	0,00
24-8-2015	87.973,89
25-8-2015	79.731,79
26-8-2015	76.000,22
27-8-2015	76.259,23
28-8-2015	80.750,29

DS	yhat
29-8-2015	68.511,73
30-8-2015	0,00
31-8-2015	80.094,19
1-9-2015	71.606,92
2-9-2015	66.043,74
3-9-2015	63.480,21
4-9-2015	64.725,50
5-9-2015	67.202,63
6-9-2015	0,00
7-9-2015	91.264,17
8-9-2015	81.552,82
9-9-2015	76.261,57
10-9-2015	75.144,20

Finalmente, descartamos el uso de *Prophet* con una predicción jerárquica por dos motivos.

El primero es que *Prophet*, a pesar de ser capaz, no fue a priori diseñado como algoritmo para series jerárquicas, y a pesar de que existen ejemplos de series temporales jerárquicas realizadas con *Prophet*, no son oficiales. Además, en la guía de Greg Rafferty sobre este algoritmo, menciona las series jerárquicas, pero en lo que a tiempo se refiere, no a este caso de jerarquía de tiendas.

"*A hierarchical time series is an example case where this may be useful: you may find good results by forecasting the more reliable daily values of one time series, for instance, and using those values to forecast hourly values of another time series that is more difficult to predict.*"

Forecasting Time Series Data with Facebook Prophet. Gerg Raffery (2021).

El Segundo motivo es que ya contamos con un R^2 elevado sin necesidad de utilizar la jerarquía como otro valor añadido. Entendemos en este caso que, ya que se solicitan datos totales y no de carácter individualizado, la eficiencia del modelo *Prophet* total y el resultado que arroja es una buena predicción.

4.2 PREDICCIÓN VENTAS SEMANALES TOTALES.

Finalmente, en lo respectivo a la predicción de las ventas semanales, el uso de regresores cobraba menos sentido que en las predicciones diarias. En la serie semanal, no existen valores nulos como en las ventas diarias y los valores máximos que encontrábamos los lunes de promoción se suavizan ya que se ven compensados con el resto de la semana.

Esta improbable necesidad de uso de variable exógenas lleva en primer lugar al uso del mencionado previamente *Neural Prophet*, al que le incluimos una estacional mensual y un modelo relativamente complejo con 10 *changepoints* (ya que encontramos este número de puntos de inflexión en la tendencia de la serie), 30 capas ocultas en la red neuronal o 190 neuronas en estas capas ocultas. A la hora de probar este modelo, con un *train-test split* de 90-10 obtenemos un score de R^2 negativo, por lo que arroja un resultado peor que el resultante de aplicar la media a predicción. Por este motivo y debido a que nos situamos lejos de un primer ajuste aceptable, se decide no continuar mejorando el modelo *Neural Prophet*.

Como segunda opción a la predicción de ventas semanales totales decidimos realizar un pipeline en el que probamos tres modelos diferentes (y a su vez con estrategias diferentes gracias a un grid) y validación cruzada:

1) *Naive Forecaster* con estrategias "drift", "last" y "mean".

2) Exponential Smoothing con diferentes tendencias y estacionalidad y se establece si se debe o no utilizar una tendencia amortiguada.

3) Theta Forecaster. Se le incluye una estacionalidad (al igual que al resto de modelos de 52 valores, es decir, anual).

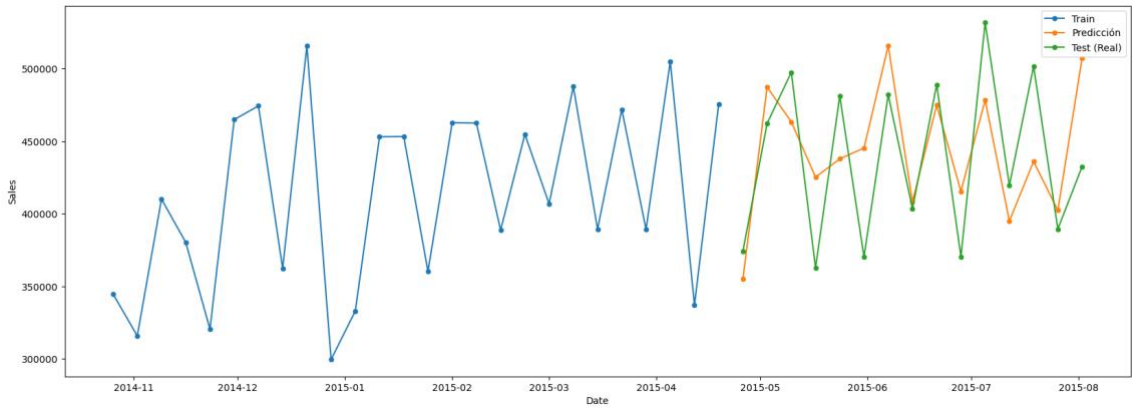
El mejor modelo resultante es el *Exponential Smoothing* con tendencia nula pero amortiguada, estacionalidad aditiva anual. En este caso es posible tener una tendencia nula, pero tendencia amortiguada positiva, ya que en el *Exponential Smoothing*, la tendencia amortiguada no necesariamente se relaciona con la tendencia general de la serie temporal, sino con la forma en que la tendencia se suaviza a medida que se acerca al final de la serie temporal. Es decir, se refiere a la forma en la que se va acomodando la tendencia a medida que se va acercando al final de la serie temporal, es independiente de si la tendencia general es positiva o negativa.

Este mejor modelo arroja un R^2 de 0.229, lo cual supone una gran mejora, ya que nos encontramos en una predicción mejor que la media, pero aún lejos de ser optima.

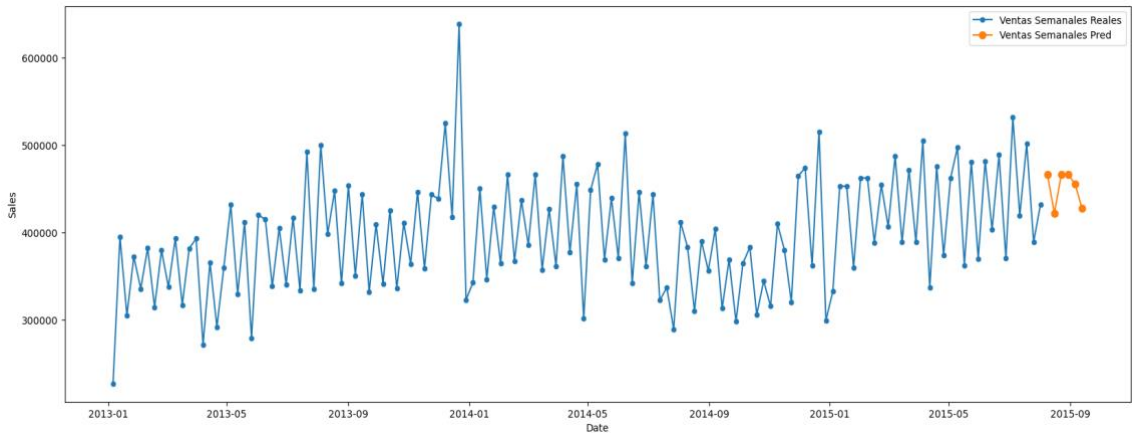
Finalmente, como último modelo, se decide realizar un bucle con varios modelos ETS, dando a este los siguientes valores de error, estacionalidad y tendencia.

```
error = ['add', 'mul']
trend = ['add', 'mul', None]
damped_trend = [True, False]
seasonal = ['add', 'mul', None]
```

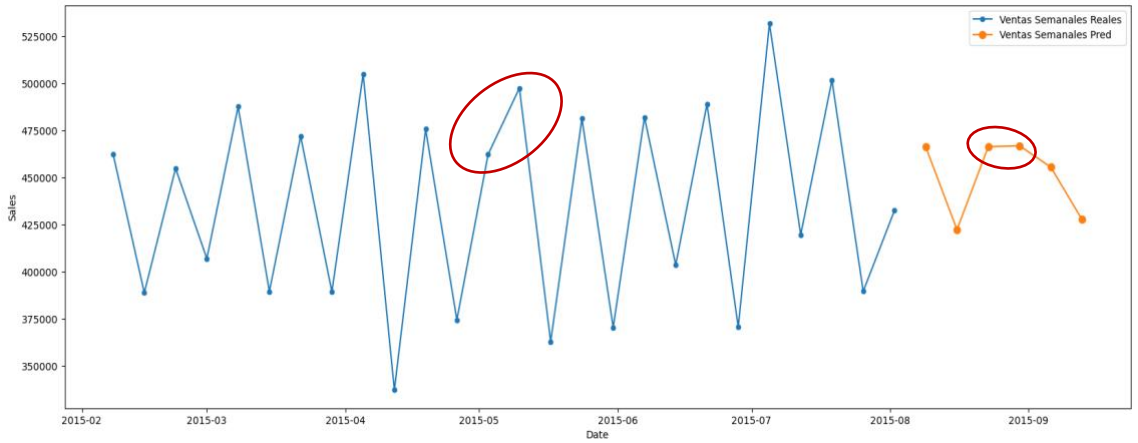
El mejor modelo resultante es un modelo con Error aditivo, tendencia nula y estacionalidad aditiva anual. El modelo arroja un score en R^2 de 0.340. Dado este score podemos decir que no se trata de un modelo preciso, pero debido a que la serie se trata de una doble suma de datos diarios y de 10 tiendas diferentes, procedemos a utilizar este como modelo para predecir.



Al igual que en las predicciones diarias, se procede a utilizar toda la serie de datos como entrenamiento del modelo y se establece un horizonte predictivo de 6 datos hasta el 13 de septiembre (a pesar de que se pide predecir hasta el 10 de Septiembre, debido a que se trata de datos semanales, las opciones son hasta el 6 o hasta el 13). El resultado de la predicción es el siguiente:



Observamos que se trata de una predicción bastante conservadora, ya que la varianza de esta es bastante menor y se encuentra alrededor de la media de los últimos datos de la serie.



Aunque en el grafico global también se observa, en este último comprobamos que la predicción incorpora un elemento que observamos únicamente en algunos momentos de la serie. Esta anomalía (marcada en rojo) puede ser determinante para la precisión del resultado.

	9-8-2015	16-8-2015	23-8-2015	30-8-2015	6-9-2015	13-9-2015
yhat (ETS)	466.234	422.272	466.366	466.871	455.514	427.638

4.3 COMPARATIVA SEMANAL CON DIARIA TRANSFORMADA

A la hora de mostrar los datos de la predicción también hemos realizado una comparativa sumando los datos de la predicción diaria realizada con *Prophet,* siguiendo la cita mencionada anteriormente de Greg Raffery sobre el algoritmo:

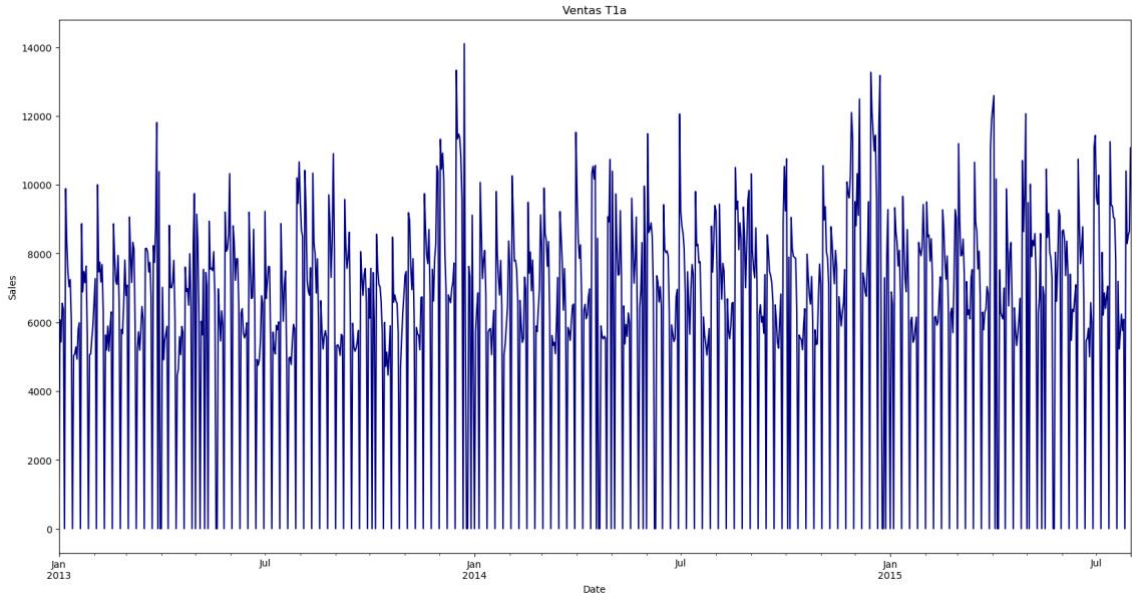
	9-8-2015	16-8-2015	23-8-2015	30-8-2015	6-9-2015	13-9-2015
yhat (ETS)	466.234	422.272	466.366	466.871	455.514	427.638
yhat (suma pred. diaria)*	386.794	462.911	370.272	469.227	413.153	324.223
Diff (ABS)	79.439	-40.639	96.095	-2.356	42.361	
Diff (%)	20,54%	-8,78%	25,95%	-0,50%	10,25%	

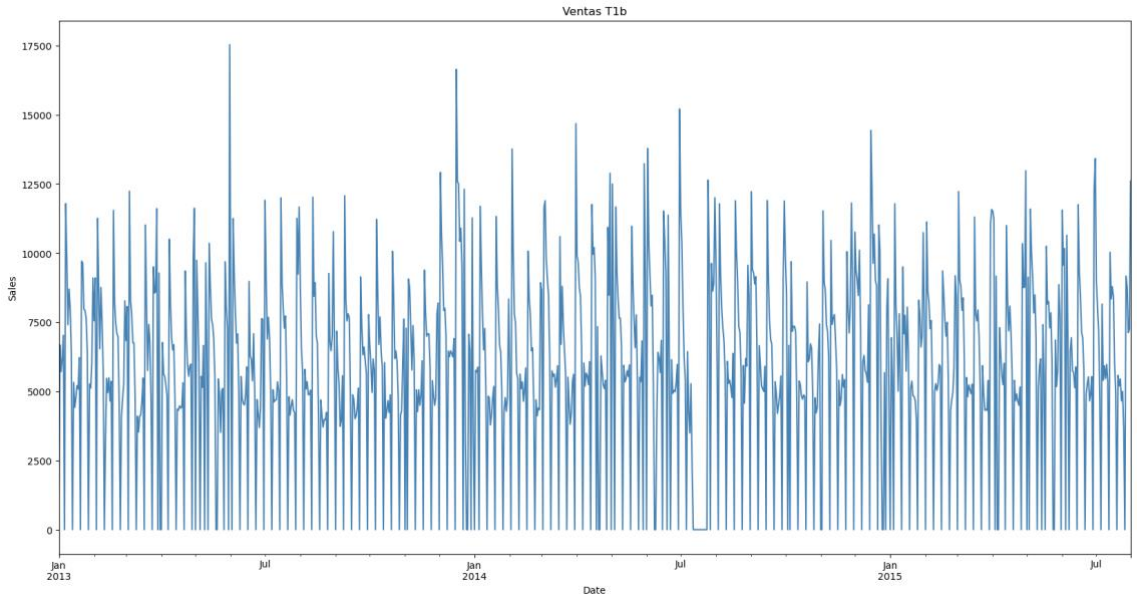
** La predicción diaria solo llega hasta el 10/9/2015, por lo que los datos no son comparables. No es posible extrapolarlos debido a la diferencia entre las ventas dentro del dia de la semana*

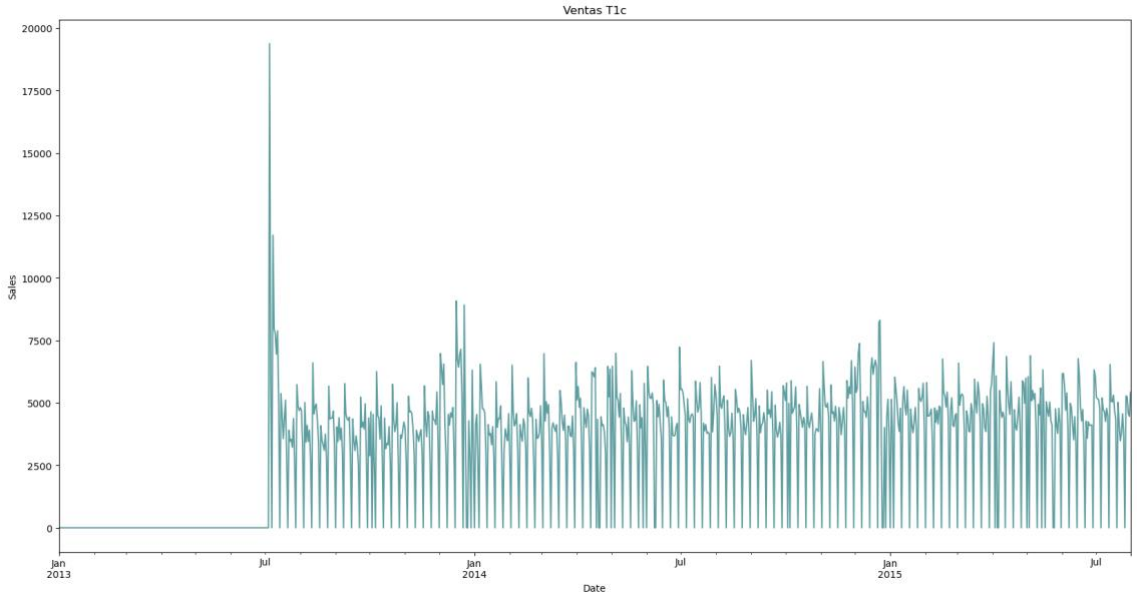
Se observa gran variación entre las series (a excepción de la cuarta semana), lo cual no nos genera gran confianza. Teniendo en cuenta el R^2 del modelo elegido y la forma de realizarlo, diríamos sin embargo que la serie que mejor predice es la diaria sumada.

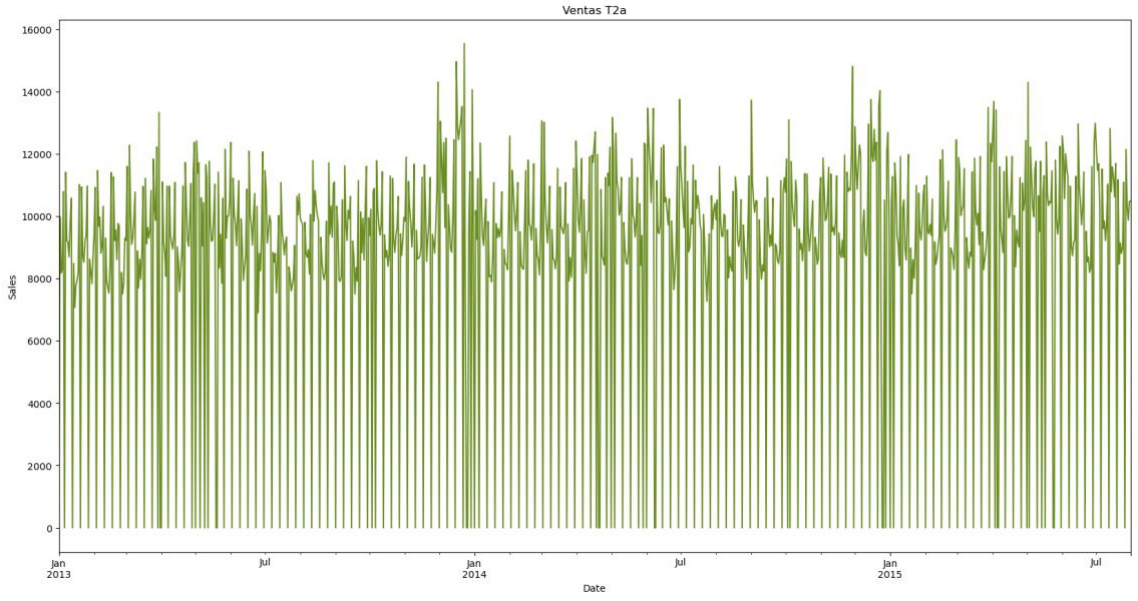
V. ANEXOS

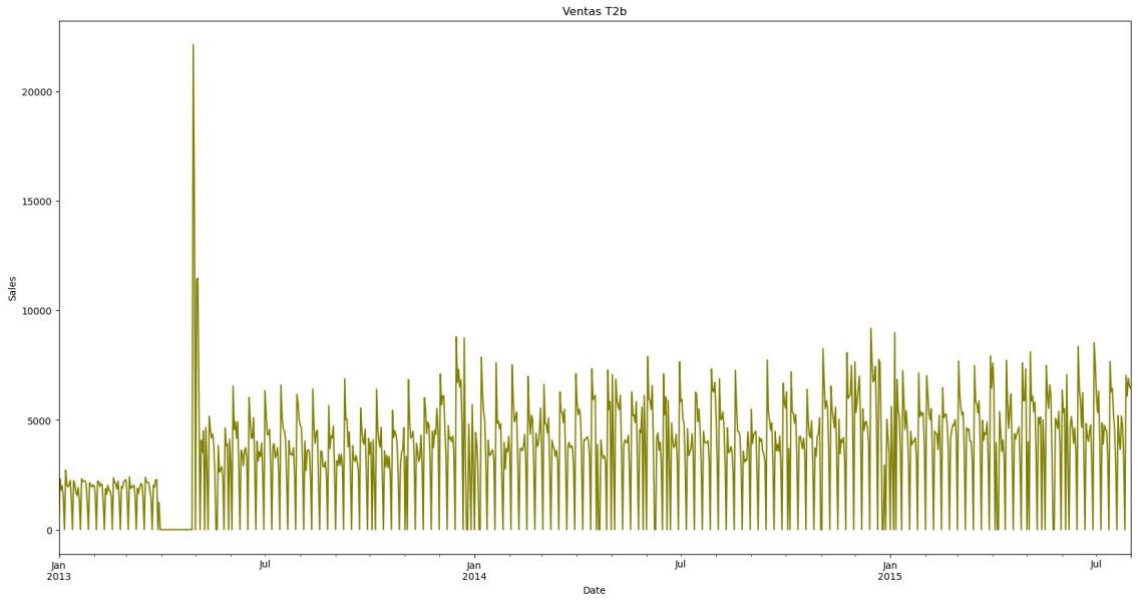
Anexo I. Gráficas de ventas diarias por tienda

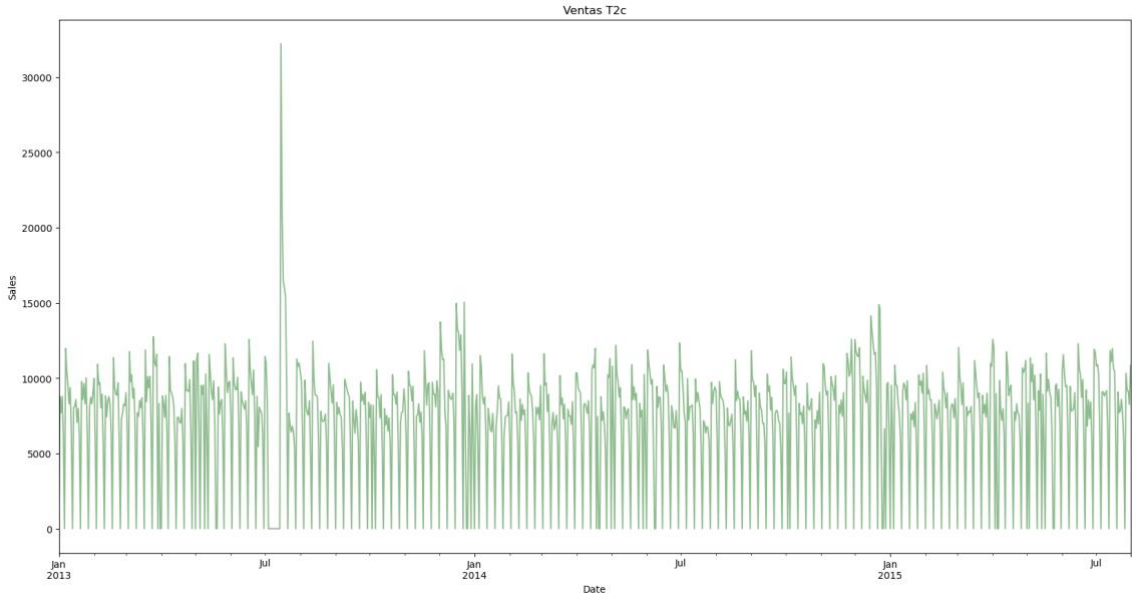


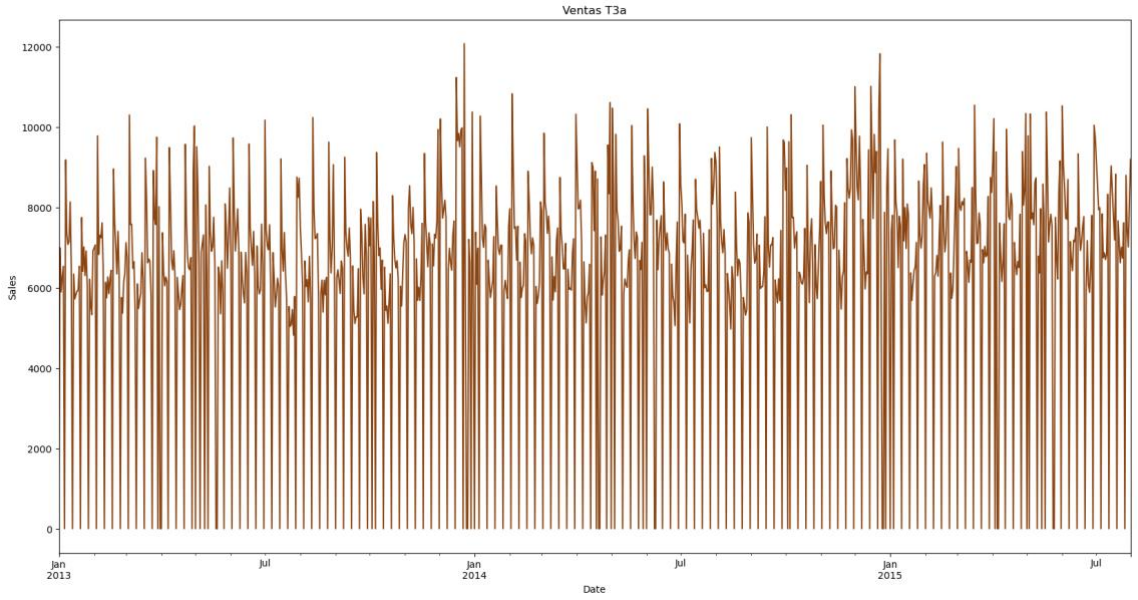

Ventas T1b
Sales
Date
Jan 2013
Jul
Jan 2014
Jul
Jan 2015
Jul

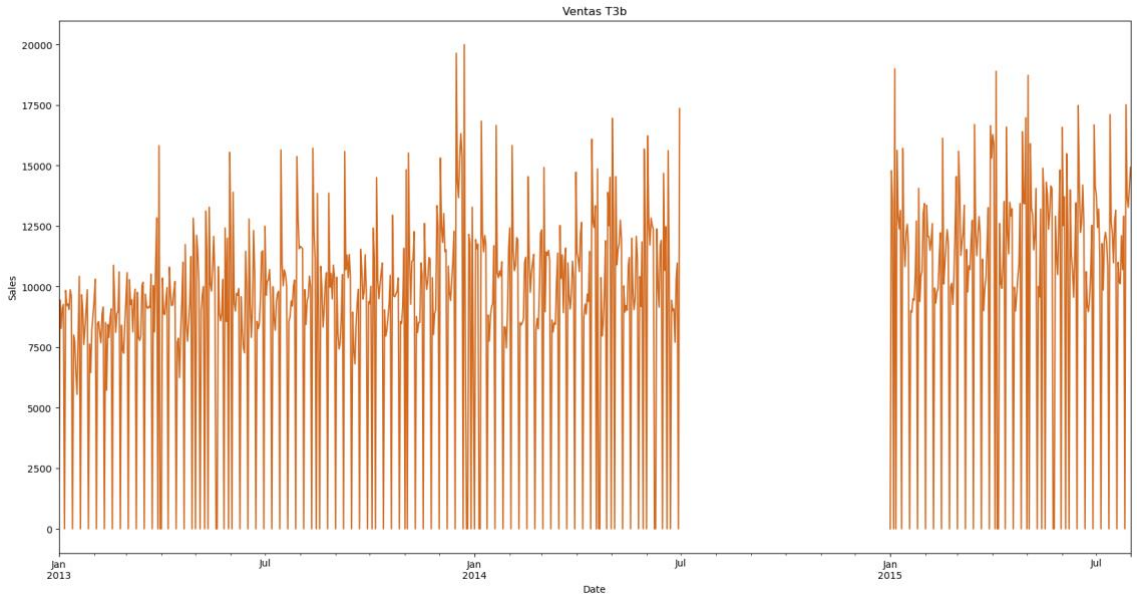

Ventas T1c
Sales
Date
Jan 2013
Jul
Jan 2014
Jul
Jan 2015
Jul

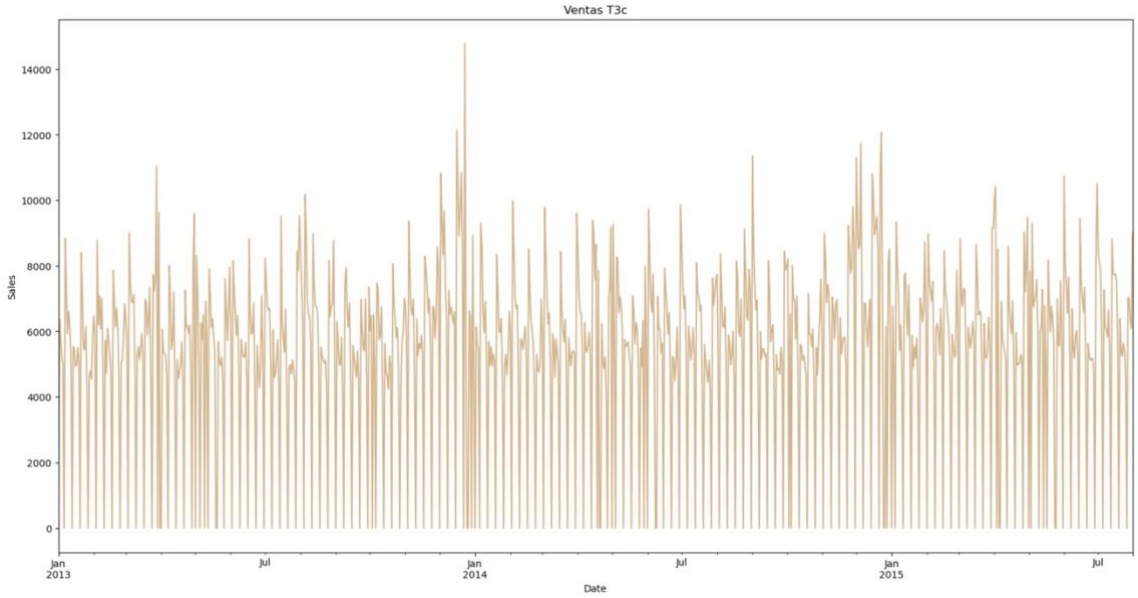

Ventas T2a
Sales
Date
Jan 2013
Jul
Jan 2014
Jul
Jan 2015
Jul

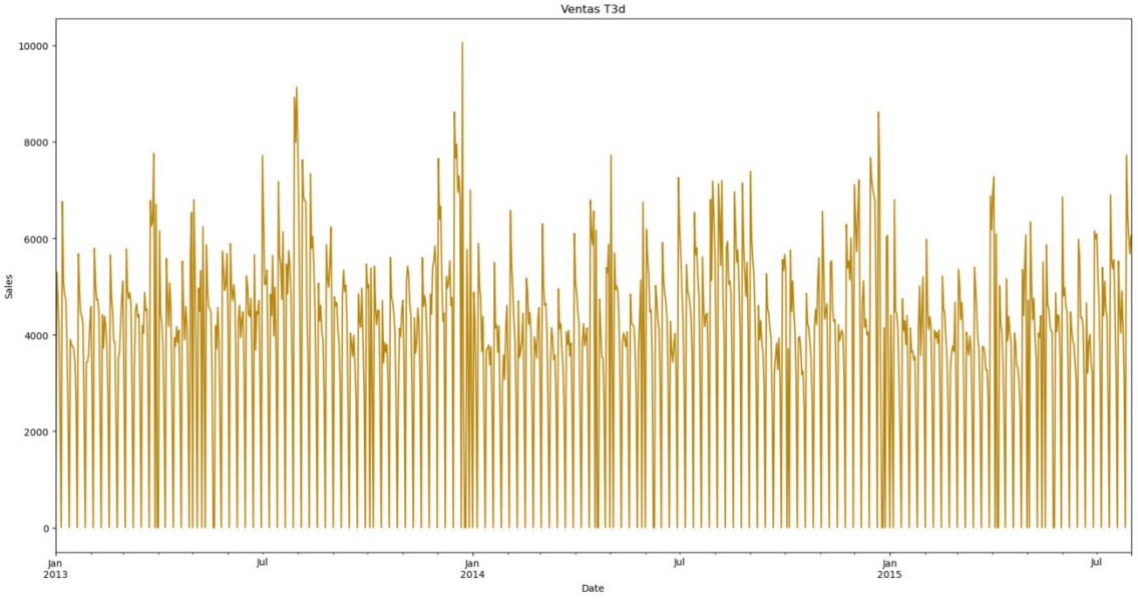

Ventas T2b
Sales
Date
0
5000
10000
15000
20000
Jan 2013
Jul
Jan 2014
Jul
Jan 2015
Jul


Ventas T2c
Sales
Date
0
5000
10000
15000
20000
25000
30000
Jan 2013
Jul
Jan 2014
Jul
Jan 2015
Jul


Ventas T3a
Sales
Date
0
2000
4000
6000
8000
10000
12000
Jan 2013
Jul
Jan 2014
Jul
Jan 2015
Jul


Ventas T3b
Sales
Date
Jan 2013
Jul
Jan 2014
Jul
Jan 2015
Jul


Ventas T3c
Sales
Date
Jan 2013
Jul
Jan 2014
Jul
Jan 2015
Jul


Ventas T3d
Sales
Date
Jan 2013
Jul
Jan 2014
Jul
Jan 2015
Jul

Anexo II. Descomposición de elementos en la serie diaria Prophet.