

TRABAJO DE FIN DE MÁSTER:

Predicción de ventas en supermercados

Walmart



MÁSTER EN
BIG DATA & BUSINESS ANALYTICS
IMF BUSINESS SCHOOL
UNIVERSIDAD DE NEBRIJA

Autora: Olga Hernández Martínez

Tutor: Juan Manuel Moreno Lamparero

Fecha: 4 de Julio de 2022

Resumen

En este trabajo de fin de máster (TFM) se va a realizar un análisis de datos sobre una base de datos de los grandes supermercados Walmart. El objetivo es poder estudiar cómo influyen ciertas variables sobre el volumen de ventas total que tienen estas tiendas.

Para ello se aplicarán técnicas de clustering para la segmentación de datos como los algoritmos K-Means, Agglomerative Clustering y DBSCAN proporcionados por Scikit-learn. También se estimará un modelo de regresión lineal múltiple, determinando cómo influyen las variables regresoras sobre el total de ventas semanales. Por último, se ajustará un modelo de series temporales utilizando la metodología Box-Jenkins, con el fin de poder predecir las ventas futuras que tendrán estos supermercados.

Índice

1. Introducción y antecedentes.....	5
1.1. Introducción.....	5
1.2. Antecedentes.....	5
1.2.1. Aparición de los supermercados	5
1.2.2. Fundación de Walmart	6
1.2.3. Importancia datos para ventas	6
1.3. Estructura del proyecto.....	7
2. Hipótesis del trabajo y objetivos.....	8
2.1. Hipótesis del trabajo	8
2.2. Objetivos.....	8
3. Material y métodos	9
3.1. Material.....	9
3.1.1. Descripción de las variables.....	9
3.2. Métodos.....	10
3.2.1. Obtención del dataset.	10
3.2.2. Clusterización.....	10
3.2.3. Regresión lineal.....	10
3.2.4. Series temporales.....	10
4. Análisis de datos.....	11
4.1. Limpieza de datos.....	11
4.2. Clusterización.....	11
4.2.1. Selección del número de clústers	12
4.2.2. Aplicación de técnicas.....	14
4.3. Regresión lineal	17
4.3.1. Especificación del modelo.	17
4.3.2. Cumplimiento de hipótesis.....	20
4.4. Series temporales.....	22
4.4.1. Fase de estimación y de identificación.....	22
4.4.2. Fase de contraste	24
4.4.3. Fase de predicción.....	29
5. Discusión de los resultados.....	31
5.1. Clusterización	31

5.2.	Regresión lineal	31
5.3.	Series Temporales.....	32
6.	Conclusiones	33
7.	Referencias Bibliográficas.....	36

1. Introducción y antecedentes.

1.1. Introducción

A lo largo de este trabajo de fin de máster se pretende analizar una base de datos sobre la multinacional de supermercados Walmart, una de las cadenas más grandes del mundo. Para ello, se cuenta con una base de datos con el registro de ventas obtenidas en un periodo de algo más de dos años, y se quiere analizar cómo influyen ciertas variables como la temperatura exterior, los días festivos o el precio del carburante en las ventas que obtenga la empresa. Además, se quiere predecir cuáles serían las ventas futuras que tendría Walmart, para analizar qué campañas de marketing de ofertas deberían realizarse en base a las ubicaciones de los locales, y así maximizar el beneficio en los lugares donde el volumen de ventas se esperaba inicialmente que fuera menor.

Para llevar a cabo este estudio, se implementarán técnicas y conocimientos adquiridos a lo largo de la realización del Máster de Big Data & Business Analytics, tales como la implementación de Machine Learning, Regresión lineal, Forecasting, visualización de datos, etc.

1.2. Antecedentes.

1.2.1. Aparición de los supermercados

Históricamente los clientes acudían a los pequeños comercios con un listado de compra que entregaban a los trabajadores, teniendo que realizar grandes colas para ser atendidos. Esta situación provocaba grandes pérdidas para los comercios. Fue en el año 1916 cuando Clarence Saunders fundó el primer supermercado de la historia, llamándolo ‘Piggly Wiggly’, donde los clientes tenían que coger los productos y el personal los esperaba en la caja para cobrarles (Güemes, 2020). Los analistas del momento pensaron que este nuevo modelo de mercado sería un fracaso pues se eliminarían las comodidades de los clientes y podrían producirse peleas por conseguir los mejores productos.

Sorprendentemente este modelo funcionó a la perfección. Suponía un ahorro de tiempo tanto para los vendedores como para los compradores, se maximizaba el beneficio para el comercio y se optimizaba el trabajo de los comerciantes, pudiendo reponer los productos con mayor frecuencia. Fue tal la revolución que supuso este nuevo modelo que empezaron a abrir supermercados de este tipo por todo el mundo.

1.2.2. Fundación de Walmart

En el año 1945, Sam Walton, tras luchar en la segunda guerra mundial, se veía en la obligación de mantener a su familia, por lo que decidió abrir su primer establecimiento. En el local, Sam se centró en vender productos a bajo precio, consiguiendo proveedores con precios más bajos para poder obtener un mayor volumen de ventas aunque con un margen menor de ganancia (Casino, 2020). Reemplazó las cajas de los mostradores por una línea de cajas a la salida de la tienda, ofrecía promociones especiales, y era consciente que la clave del éxito dependía del número de asociados que consiguiera en su comercio. Gracias a esta innovación y a la política de precios, Walton triplicó el número de ventas en poco tiempo. Como consecuencia, Sam empezó a abrir más locales tanto en zonas urbanas como en zonas rurales. A día de hoy, **Walmart Inc.** es una multinacional que venden al por menor y funciona como una cadena de hipermercados. Es conocida como la minorista de alimentos más grande de Estados Unidos, que cuenta ya con un total de 11.000 tiendas repartidas en 28 países del mundo.

1.2.3. Importancia datos para ventas

Con el paso del tiempo, cada vez han sido más los supermercados en recurrir a almacenar y analizar grandes datos acerca sus clientes. Estudian cuáles son los hábitos de consumo, las tendencias sociales, los días y horarios en los que se compra más, qué clase de productos son más buscados en una determinada época, cómo influyen las previsiones meteorológicas, etc. Gracias al Big Data, los supermercados pueden conocer en tiempo real cuáles son las ventas que se están produciendo en el momento para tomar decisiones en los precios, ofertas, y almacenamiento de stock (InStoreView, 2020).

Siendo conscientes del gran poder que otorgan los datos para poder aumentar las ventas de cualquier tipo de comercio, en este proyecto se llevará a cabo una predicción de ventas sobre la cadena de supermercados Walmart, con el fin de analizar la importancia de gestionar todo tipo de información sobre los clientes.

1.3. Estructura del proyecto

Este proyecto se estructura en seis apartados, donde se realizará lo siguiente:

- Apartado 1. Se introducirá el trabajo, y se explicará la evolución histórica de los supermercados, así como de la aparición de las cadenas de supermercados Walmart.
- Apartado 2. Se mostrarán las hipótesis de trabajo de las que se parten, y se especificarán los objetivos a lograr.
- Apartado 3. Se detallará el dataset con el que se va a trabajar a lo largo del proyecto, indicando las características principales, y explicando los materiales y métodos que se utilizarán para el análisis posterior.
- Apartado 4. Se aplicarán las técnicas de análisis de datos con el fin de estudiar a fondo la información proporcionada en el dataset, y con el propósito de lograr los objetivos inicialmente propuestos.
- Apartado 5. Se analizan los resultados obtenidos en el anterior apartado.
- Apartado 6. Se explican las conclusiones a las que se llega tras la realización completa del trabajo.

2. Hipótesis del trabajo y objetivos

2.1. Hipótesis del trabajo

Existen ciertas variables que son muy influyentes en el número de ventas de los comercios, y si las empresas son conscientes del gran poder que otorgan, pueden aumentar considerablemente el número de ingresos. Según Weather Unlocked, el clima tiene la mayor influencia en el comportamiento del consumidor, ya que afecta en su estado de ánimo e impulsa en las decisiones de compra del cliente (Solares, s.f.) Por otro lado, el aumento del precio del carburante contribuye a que los supermercados tengan que subir el precio de sus productos por el encarecimiento de los costes generados. Esta situación la tenemos muy presente debido a la inflación que se está produciendo en los últimos meses. Es por esto, que el precio del carburante tendrá una influencia directa sobre el volumen de compra que se genere en los supermercados. Asimismo, tener presentes los días festivos, o temporadas de cambio estacional, que permita dedicar pasillos enteros a épocas como Halloween, Navidad, Pascua, vuelta al cole, etc., posibilitará el aumento de ingresos producidos por estas temáticas.

La idea principal de este proyecto es poder determinar la influencia de ciertas variables sobre el volumen de ventas de los supermercados Walmart, haciendo uso de técnicas analíticas propias de Big Data & Business Analytics.

2.2. Objetivos.

Con el análisis de estos datos se pretende lograr los siguientes objetivos:

- a) Segmentar las tiendas según las ventas que tienen.
- b) Determinar en qué locales habría que realizar más campañas con descuentos para aumentar el número de ventas.
- c) Analizar el impacto de la temperatura y del precio del combustible sobre las ventas.
- d) Analizar el impacto que tiene la tasa de desempleo sobre el volumen de ventas.
- e) Analizar el impacto que tienen las semanas con festivos sobre el número de ventas.
- f) Predecir el volumen de ventas que tendrá la cadena de supermercados en unos meses.

3. Material y métodos

3.1. Material

Para la realización de este proyecto se ha hecho uso del dataset público Walmart en formato csv (Walmart Analysis Dataset, s.f.) extraído de Kaggle. Se trata de una base de datos que contiene la información del número de ventas semanal por cada una de las tiendas que tiene distribuidas en EEUU. Cuenta además con información como la fecha, la tasa de desempleo o el coste del combustible entre otros.

Se trata de una base de datos históricos que abarcan las ventas desde 2010-02-05 hasta 2012-11-01. El dataset consta de:

- 17 variables (columnas)
- 423.325 registros

3.1.1. Descripción de las variables.

De todas las variables presentes, no se van a tener en cuenta las de tipo `Markdown_i` (con $i \in [1,5]$) ya que no se especifica qué tipo de información almacenan. Tampoco se utilizarán las variables `X` (número de registro), `Type`, ni `Size` ya que no aportan información relevante para el análisis que se quiere llevar a cabo. En la siguiente tabla se describen las características, y el tipo de formato inicial de cada una de ellas:

Variable	Definición	Tipo
<code>Store</code>	El número de la tienda	<code>int</code>
<code>Date</code>	La fecha de la semana de ventas	<code>chr</code>
<code>IsHoliday</code>	Especifica si la semana es una semana especial festiva	<code>logical</code>
<code>Dept</code>	La región donde está situada la tienda	<code>int</code>
<code>WeeklySales</code>	Ventas para la tienda dada	<code>num</code>
<code>Temperature</code>	Temperatura el día de la venta	<code>num</code>
<code>Fuel_Price</code>	Coste del combustible de la región	<code>num</code>
<code>CPI</code>	Índice de precios del	<code>num</code>

	consumidor predominante	
Unemployment	Tasa de desempleo predominante	num

3.2. Métodos.

En este proyecto se va a utilizar el lenguaje de programación R, y Jupyter Notebook para la aplicación de técnicas analíticas, y para la consecución de los objetivos planteados inicialmente. A continuación se describen los métodos que se van a llevar a cabo durante este proyecto.

3.2.1. Obtención del dataset.

Como ya se ha detallado previamente en el apartado 3.1., el dataset es un dataset público en formato csv obtenido de Kaggle. Esta base de datos contiene información acerca de las ventas de la multinacional Walmart (Walmart Analysis Dataset, s.f.).

3.2.2. Clusterización.

En la clasificación de las tiendas según el volumen de ventas que tengan, se aplicará una segmentación de datos haciendo uso de los algoritmos K-Means, Agglomerative Clustering y DBSCAN, todas ellas proporcionadas por Scikit-learn.

3.2.3. Regresión lineal.

Para poder encontrar el grado de influencia de algunas variables sobre el número de ventas, se va a realizar un análisis de regresión lineal múltiple. Para ello se implementarán varios modelos, y se hará uso de la función StepAIC. Posteriormente, se comprobarán las hipótesis de linealidad del modelo.

3.2.4. Series temporales.

Con motivo de poder predecir el número de ventas que se producirán en un futuro, se realizará un análisis de series temporales, donde se implementará la metodología Box-Jenkins, basada en el empleo de modelos SARIMA, es decir modelos del tipo:

$$(1 - B)^d (1 - B^s)^D X_t = \frac{\theta_q(B)\Theta_Q(B)}{\phi_p(B)\Phi_P(B)} a_t$$

4. Análisis de datos

El análisis realizado para la clusterización se encuentra en el notebook `Clusterización.ipynb` (Hernández, 2022). El resto de implementaciones sobre el dataset como la limpieza de datos, la regresión lineal, y el ajuste de series temporales se encuentra en `data_Analysis.Rmd` (Hernández, 2022)

4.1. Limpieza de datos.

Para poder realizar el análisis de datos, previamente hay que realizar una limpieza de estos, para ello se ha realizado lo siguiente:

- Primeramente se han eliminado las columnas: `X`, `MarkDowni` (con $i \in [1,5]$), `Type` y `Size`, ya que no aportan información relevante.
- Se ha convertido en formato fecha la variable "Date" cuyo formato inicial era "character".
- Se ha convertido en factor la variable `Is_Holiday`, asignándole el valor 1 a TRUE, y 0 a FALSE.
- Se han eliminado los registros en los que el número de ventas era 0, ya que no tenía sentido para el análisis que iba a realizar.
- Se ha hecho una agrupación de datos según los locales y sus fechas, sin tener en cuenta la segmentación interna en departamentos.

Finalmente se obtiene un dataset cuyo nombre es "walmart_new.csv", que contiene un total de 6.435 registros y 8 variables.

4.2. Clusterización.

Las técnicas de clusterización permiten agrupar y relacionar datos, cuya relación no resulta aparentemente significativa o determinada. Con este análisis se buscará agrupar los locales según las ventas, para poder determinar aquellos en los que sería necesario realizar ofertas o descuentos con el fin de aumentar la demanda en esas tiendas.

Es por esto que de todas las variables de la base de datos, solo nos quedaremos con `Store`, y `TotalSalesWeek` para el análisis de clusterización. Asimismo, se seleccionan los datos

del último año, y se contabilizan las ventas que se han obtenido por cada local. Obteniendo como conjunto de datos final, un dataset con el registro de las 45 tiendas de Walmart, y el volumen de ventas que tienen a partir del año 2012, ya que son los últimos registros del dataset inicial.

En la siguiente figura se muestra un histograma de las variables, que nos permiten observar la distribución de las frecuencias de cada variable.

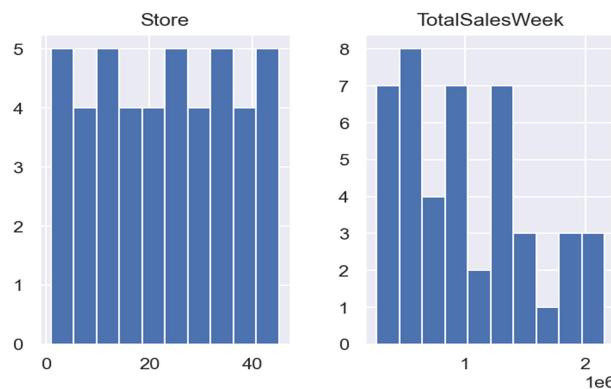


Figura 1. Histograma de las variables Store y TotalSalesWeek

4.2.1. Selección del número de clústers

Una de las desventajas de aplicar los algoritmos de clusterización es poder determinar el número de clústeres en que se van a segmentar los datos. Esta elección resulta de gran relevancia, ya que si seleccionamos pocos clústeres, realizaremos agrupaciones de datos muy heterogéneos; o si seleccionamos muchos clústeres podremos agrupar datos muy similares en diferentes conjuntos (Moya, 2016).

Para poder elegir el número óptimo de clústeres se han aplicado los siguientes métodos:

- El método de la silueta (sillhouette method).

Este método mide la distancia de separación entre los clústeres, siendo el rango de esta medida: [-1,1]. El coeficiente de la silueta para una observación se calcula con la siguiente expresión (González, 2019):

$$S(x) = \frac{b(x) - a(x)}{\max(a(x), b(x))}$$

Los valores cercanos a 1 indican que la observación se encuentra lejos de los clústeres de alrededor. El valor cercano a 0 señala que la observación está muy cerca o en la frontera entre dos clústeres. Los valores negativos expresan la posible asignación errónea de una observación en un clúster determinado.

El coeficiente de la silueta para todo el agrupamiento es:

$$SC = \frac{1}{N} \sum_{i=1}^N S(x)$$

- El método del codo (elbow method).

Este método utiliza los valores de la inercia obtenidos tras aplicar el algoritmo K-Means en diferente número de clústeres (de 1 a N). Para ello se utiliza la inercia, que es la suma de las distancias al cuadrado de cada objeto del clúster a su centroide (Moya, 2016):

$$\text{inercia} = \sum_{i=0}^N ||x_i - \mu||^2$$

Tras la implementación de este método se obtiene una gráfica con forma de codo, donde el punto en que se produzca el cambio de sentido más brusco, será el valor del clúster a seleccionar.

Con los datos que tenemos, se han aplicado ambas técnicas y se han obtenido las siguientes representaciones gráficas:

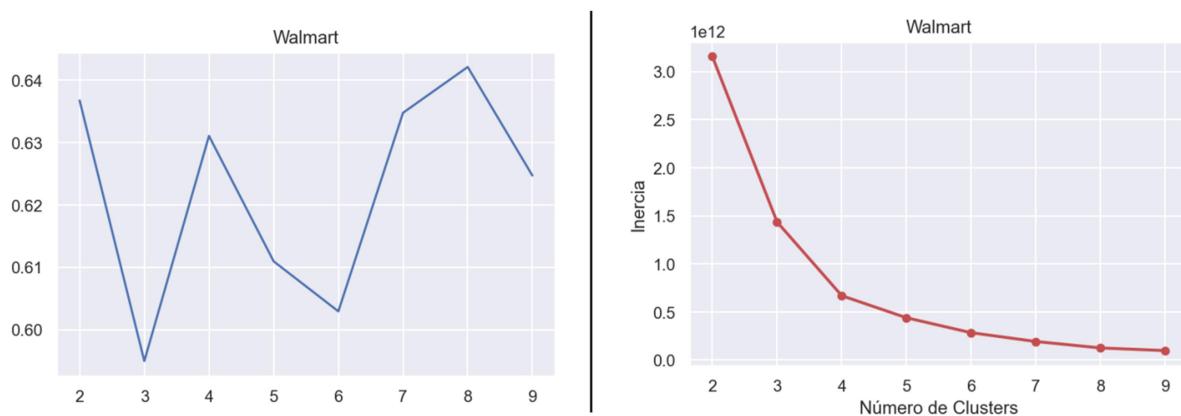


Figura 2. Aplicación del método de la silueta (izq.) y del método del codo (dcha.)

Se elige continuar con el algoritmo de clusterización tomando 4 clústeres de segmentación.

4.2.2. Aplicación de técnicas.

Existen varias técnicas de aprendizaje automático no supervisado para aplicar en la segmentación de datos. Las que se utilizarán son las siguientes:

- **K-Means:** El algoritmo realiza el agrupamiento comenzando con la selección al azar de un primer grupo de centroides, que se utilizan como puntos iniciales. Posteriormente se realizan cálculos iterativos para optimizar y estabilizar los centroides, hasta obtener los grupos finales (Garbade, 2018).
- **Agglomerative Clustering:** Realiza un agrupamiento jerárquico recursivo donde agrupa los datos de una muestra. Cada observación comienza en su propio grupo y estos se van fusionando en cada iteración. (Documentación sklearn.)
- **DBSCAN:** Significa *Clustering espacial basado en densidad de aplicaciones con ruido*. Este algoritmo necesita que se establezcan el radio donde se determina si los puntos están próximos o no al centro, y los puntos mínimos que es la cantidad mínima de datos necesaria para formar un clúster. Al contrario que los otros dos métodos, este no necesita la especificación del número de clústeres (Moreno, s.f.).

Los resultados de aplicar estos tres algoritmos diferentes sobre nuestros datos son los que muestran la figura 3.

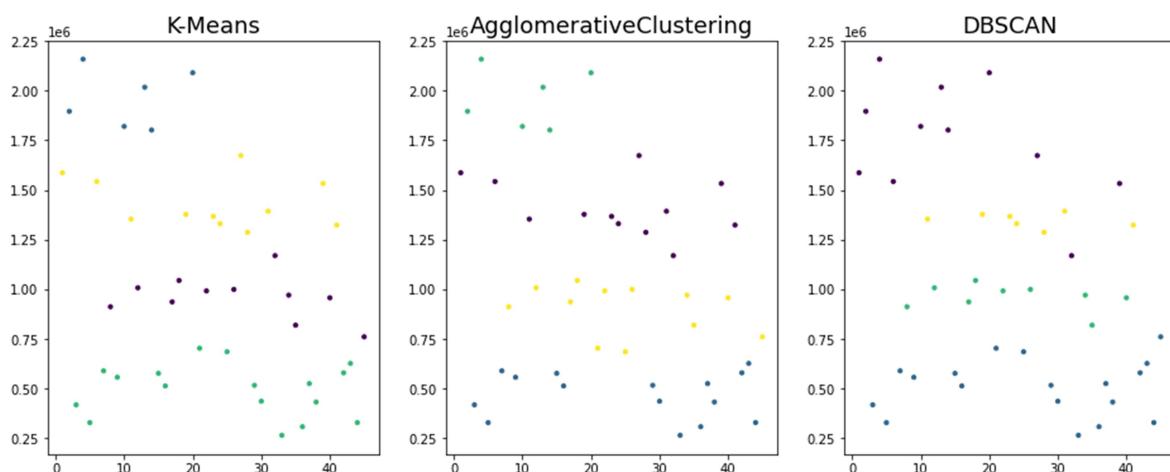


Figura 3. Aplicación de K-Means, Agglomerative Clustering y DBSCAN.

Para comparar estos tres algoritmos y elegir el que dé mejores agrupaciones de datos, se utiliza el índice Dunn. Se trata de un valor numérico que se obtiene al calcular la proporción entre la mínima distancia inter-grupo, y la máxima distancia intra-grupo (PyShark, 2022). Para m clústeres, el índice Dunn se calcula mediante la siguiente expresión:

$$DI_m = \frac{\min_{1 \leq i \leq j \leq m} \delta(C_i, C_j)}{\max_{1 \leq k \leq m} \Delta_k}$$

Se ha realizado en el Jupyter notebook la implementación del índice Dunn (Rodrigues, 2018), obteniendo como resultados la siguiente tabla:

	K-Means	Agglomerative Clustering	DBSCAN
0	0.268249	0.247808	0.126201

Tabla 1. Comparativa índice Dunn.

Ya que los valores más altos del índice de Dunn representan una mejor agrupación, continuaremos el análisis utilizando el algoritmo K-Means.

Para segmentar los locales según el volumen de ventas, se aplica K-Means utilizando los 4 clústeres elegidos mediante los métodos del codo y de la silueta:

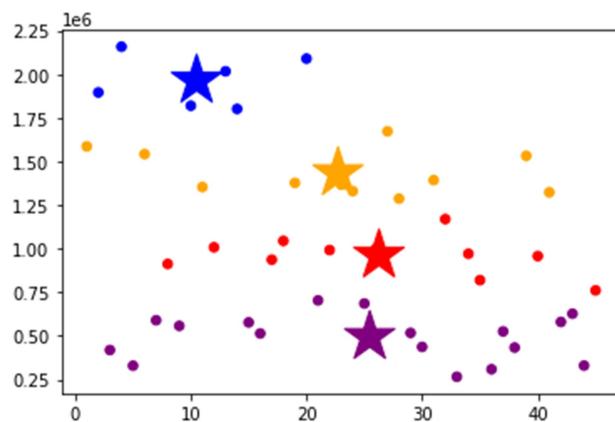


Figura 4. Segmentación según K-Means.

En la siguiente tabla se muestra la clasificación de cada local en su clúster correspondiente, y se obtiene además, el valor de la media del número de ventas que existe en cada agrupación de datos.

Clúster	Stores	Media de ventas del clúster
0	2, 4, 10, 13, 14, 20	1.963.087
1	3, 5, 7, 9, 15, 16, 21, 25, 29, 30, 33, 36, 37, 38, 42, 43, 44	494.300
2	1, 6, 11, 19, 23, 24, 27, 28, 31, 39, 41	1.432.937
3	8, 12, 17, 18, 22, 26, 32, 34, 35, 40, 45	961.029

Tabla 2. Clasificación de las tiendas en los clústers.

Observamos que en el clúster número 1 se produce un menor volumen de ventas medio que en el resto de agrupaciones.

4.3. Regresión lineal

4.3.1. Especificación del modelo.

Para llevar a cabo el análisis de regresión lineal múltiple, se toma como variable respuesta TotalSalesWeek, con la finalidad de poder crear una ecuación que nos ayude a estudiar el grado de relación o de predicción que puedan tener las variables regresoras sobre esta. La ecuación de regresión lineal múltiple de nuestro modelo es:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + \beta_6 \cdot X_6 + \beta_7 \cdot X_7 + Error$$

Siendo nuestras variables las siguientes:

$$Y = \text{TotalSalesWeek}$$

$$X_1 = \text{Store}$$

$$X_2 = \text{Date}$$

$$X_3 = \text{IsHoliday}$$

$$X_4 = \text{Temperature}$$

$$X_5 = \text{Fuel_Price}$$

$$X_6 = \text{CPI}$$

$$X_7 = \text{Unemployment}$$

Primeramente se realiza el análisis de regresión lineal con todas las variables:

```
Call:
lm(formula = TotalSalesWeek ~ ., data = datos_new)

Residuals:
    Min      1Q  Median      3Q     Max 
-1033409 -392826 - 37826   371809  2711772 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.961e+06 5.346e+05 3.669 0.000245 ***
Store       -1.539e+04 5.225e+02 -29.457 < 2e-16 ***
Date        2.555e+00 3.934e+01  0.065 0.948214    
IsHoliday    7.288e+04 2.606e+04  2.796 0.005187 **  
Temperature -9.767e+02 3.765e+02 -2.594 0.009504 **  
Fuel_Price   8.325e+03 2.454e+04  0.339 0.734485    
CPI         -2.322e+03 1.903e+02 -12.206 < 2e-16 ***  
Unemployment -2.181e+04 3.948e+03 -5.524 3.45e-08 *** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 523200 on 6427 degrees of freedom
Multiple R-squared:  0.1416,    Adjusted R-squared:  0.1406 
F-statistic: 151.4 on 7 and 6427 DF,  p-value: < 2.2e-16
```

Figura 5. Resumen del primer modelo

Observamos que todas las variables son estadísticamente significativas excepto Date y Fuel_Price. Así que se decide realizar un nuevo modelo eliminando estas variables como predictoras.

```

Call:
lm(formula = TotalSalesWeek ~ Store + IsHoliday + CPI + Unemployment +
    Temperature, data = datos_new)

Residuals:
    Min      1Q   Median      3Q     Max 
-1035875 -392198 -40396  371093 2711784 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2031946.3    50655.1 40.113 < 2e-16 ***
Store        -15373.6     521.3 -29.489 < 2e-16 ***
IsHoliday     72222.2    25911.1  2.787 0.00533 **  
CPI          -2346.0     180.2 -13.019 < 2e-16 ***
Unemployment -22196.1    3756.0 -5.910 3.61e-09 ***
Temperature   -929.0     369.1 -2.517 0.01185 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 523100 on 6429 degrees of freedom
Multiple R-squared:  0.1415, Adjusted R-squared:  0.1408 
F-statistic: 211.9 on 5 and 6429 DF, p-value: < 2.2e-16

```

Figura 6. Resumen del segundo modelo

Al eliminar estas dos variables, se obtiene un nuevo modelo donde todas las variables son significativas. Para poder asegurarnos de que se ha elegido el mejor modelo, se implementa la función *StepAIC* de la librería *MASS*.

```

Step: AIC=169472.6
TotalSalesWeek ~ Store + CPI + Unemployment + IsHoliday + Temperature

          Df Sum of Sq    RSS    AIC
<none>             1.7593e+15 169473
+ Fuel_Price  1 1.149e+11 1.7592e+15 169474
+ Date       1 8.456e+10 1.7593e+15 169474

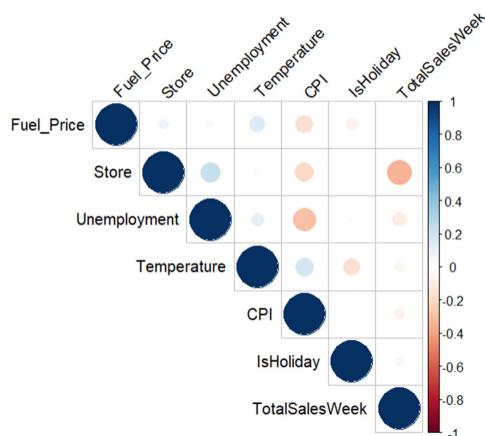
Call:
lm(formula = TotalSalesWeek ~ Store + CPI + Unemployment + IsHoliday +
    Temperature, data = datos_new)

Coefficients:
(Intercept)      Store      CPI Unemployment      IsHoliday      Temperature
 2031946       -15374     -2346     -22196        72222       -929

```

Figura 7. Implementación de StepAIC

El modelo con el menor valor de AIC es justamente el anterior modelo que habíamos ajustado. En la siguiente imagen estudiamos la matriz de correlaciones de las variables. A través de ella podremos tener una primera impresión de multicolinealidad en el modelo.



Consideramos que las variables regresoras son linealmente independientes, debido a que no se observan coeficientes de Pearson mayores que 0,5.

Aun así, es importante analizar la multicolinealidad de las variables utilizando el VIF (Factor de Inflación de Varianza, del inglés “Variance Inflation Factor”). Este valor se obtiene a través de la siguiente expresión:

$$VIF = \frac{1}{1 - R^2}$$

Si el valor que se obtiene es próximo a 1, implicaría que el $R^2 = 0$, es decir, la variable sería independiente del resto. Si por el contrario, este valor tiende a infinito, la variable no sería independiente, sino que se podría calcular a partir del resto de variables independientes (Rodríguez, 2020).

Los valores de VIF que se obtienen en el modelo son los siguientes:

Store	IsHoliday	CPI	Unemployment	Temperature
1.077985	1.026819	1.182447	1.167162	1.089627

Figura 8. Valores VIF

Como los valores son próximos a 1 en todos los casos, aceptaremos que se cumple la hipótesis de no multicolinealidad.

El valor del R cuadrado es de 0,1415, es decir, en torno a un 15% de la variabilidad de la variable respuesta es explicada por el modelo. Por su parte el R cuadrado ajustado tiene un valor de 0,1408. Esto nos muestra cómo funcionaría nuestro modelo si se generalizase.

```
Residual standard error: 523100 on 6429 degrees of freedom
Multiple R-squared:  0.1415,    Adjusted R-squared:  0.1408
F-statistic: 211.9 on 5 and 6429 DF,  p-value: < 2.2e-16
```

Figura 9. R cuadrado y R ajustado

Que estos valores sean tan bajos puede deberse a la alta variabilidad de datos que hay. Aun así queremos focalizar este análisis en el grado de influencia de ciertas variables sobre la variable dependiente. Para ello, obtenemos la estimación del modelo:

$$\hat{Y} = 2.031.946 - 15.374X_1 - 2.346X_2 - 22.196.X_3 + 72.222.X_4 - 929.X_5$$

Siendo:

X_1 = Store

X_2 = CPI

X_3 = Unemployment

X_4 = IsHoliday

X_5 = Temperature

4.3.2. Cumplimiento de hipótesis

Para saber si la relación entre las variables de estudio es lineal o no, debemos estudiar además la hipótesis de linealidad del modelo, para poder verificar o descartar este tipo de relación. Para ello tendremos que comprobar si se cumplen otras tres hipótesis: normalidad, homocedasticidad e independencia de los residuos, veámoslo:

- Normalidad.

El histograma de los residuos se asemeja a una distribución Normal desplazada hacia la izquierda.

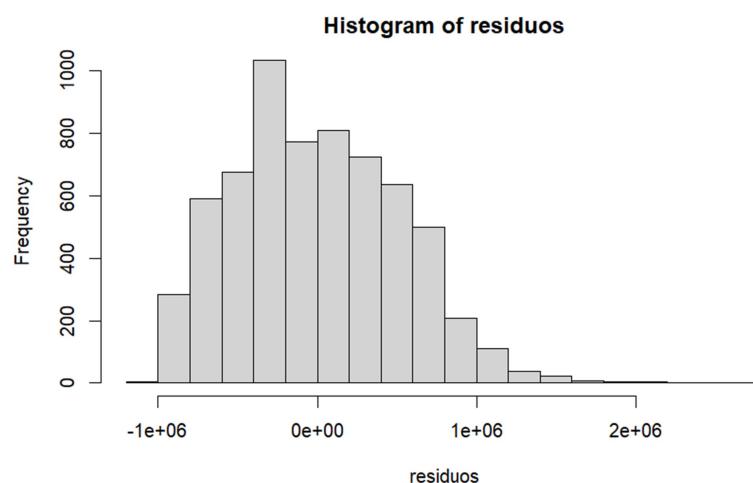


Figura 10. Histograma de los residuos

Aun así, estudiamos también el gráfico de probabilidad normal de los residuos:

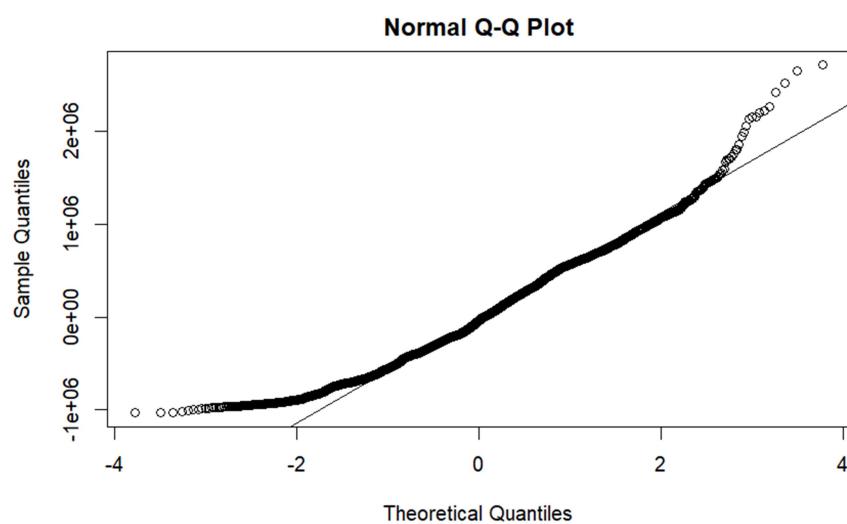


Figura 11. Gráfico Q-Q Plot.

Aceptamos la hipótesis de normalidad ya que los puntos se distribuyen a lo largo de la recta.

- Homocedasticidad e independencia.

En la siguiente figura se muestra el gráfico de los residuos estandarizados frente a las observaciones ajustadas:

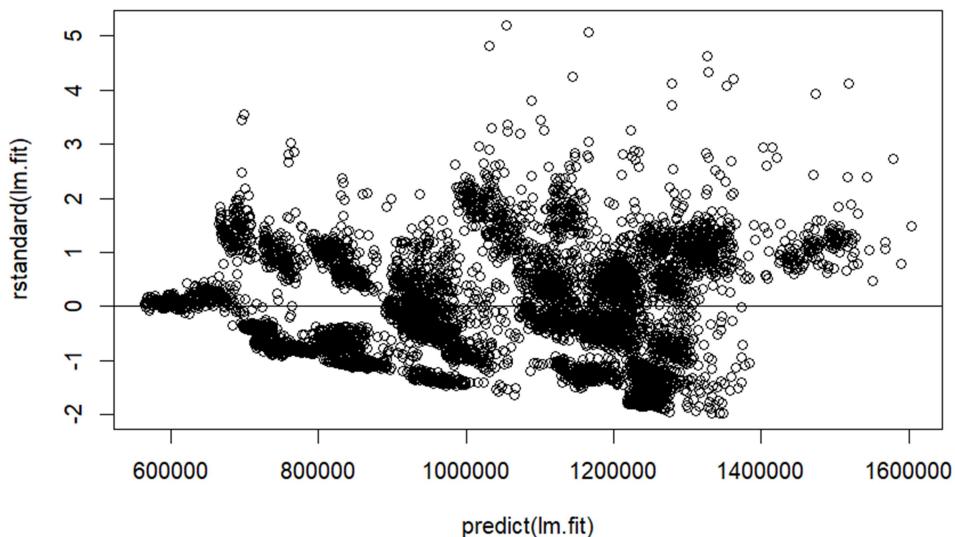


Figura 12. Residuos estandarizados vs observaciones ajustadas

Podemos apreciar que los residuos se distribuyen aleatoriamente por todo el gráfico, y que la varianza de estos no presenta una tendencia de aumento o disminución a lo largo del gráfico, por lo que podríamos dar como válida la hipótesis de homocedasticidad y de independencia.

Dado que se cumplen todas las hipótesis anteriores, podemos dar como válida la hipótesis de linealidad. Es decir, el modelo de regresión lineal múltiple que hemos planteado es válido, al existir una relación de dependencia lineal entre las variables explicativas y la variable respuesta.

4.4. Series temporales.

4.4.1. Fase de estimación y de identificación.

Durante este apartado se van a aplicar técnicas de forecasting, analizando las ventas que se han producido en el periodo que comprende el dataset. En la siguiente figura se muestra el volumen de ventas producidas en ese tiempo:

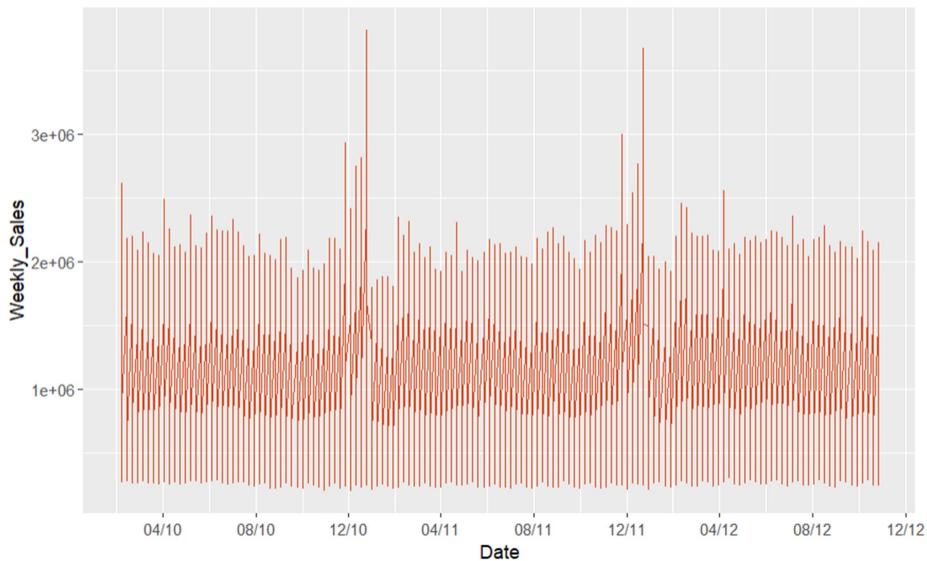


Figura 13. Evolución de ventas.

Es importante analizar la periodicidad en que se repiten los datos, así como su tendencia y evolución. Con la siguiente figura podemos tener una idea predeterminada de cómo será el comportamiento de los datos a lo largo del análisis.

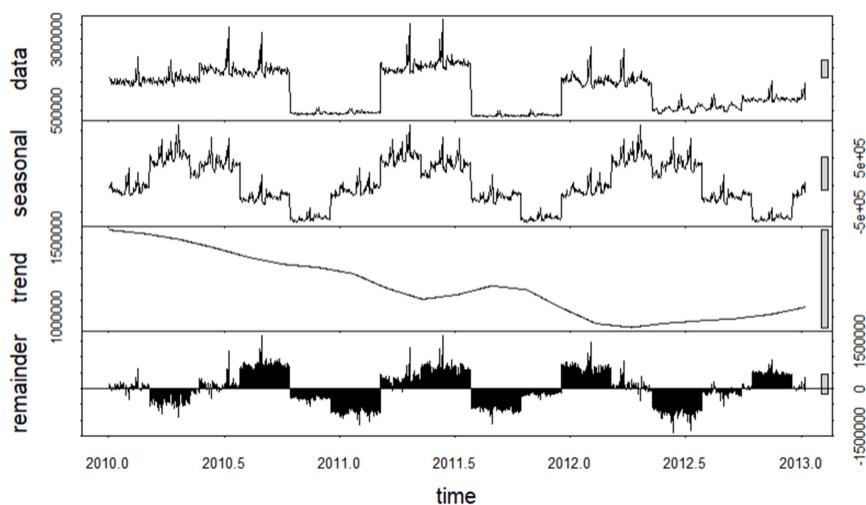


Figura 14. Resumen de los datos

Para comenzar con el análisis, primero se divide la muestra en entrenamiento y validación. Siendo la muestra de entrenamiento la comprendida entre el "2010-02-05" y el "2012-02-05", y la muestra de validación del "2012-02-05" en adelante.

Antes de ajustar ningún modelo es necesario asegurarnos si la serie es estacionaria en media y varianza.

I. Estacionaria en media:

Para estudiar la estacionariedad en media se utiliza el test de Dickey-Fuller que contrasta si en la parte AR hay una raíz unitaria, porque en caso de existir, habrá que diferenciar.

Test de Dickey-Fuller

$$\begin{cases} H_0: |\phi| = 1 \\ H_1: |\phi| \neq 1 \end{cases}$$

$$(1 - \phi B)X_t = a_t \rightarrow 1 - \phi z = 0 \Rightarrow |z| = \left| \frac{1}{\phi} \right| > 1 \Leftrightarrow |\phi| < 1$$

Si el p-valor $\leq 0.05 \Rightarrow$ Rechazamos $H_0 \Rightarrow$ La serie es estacionaria en media

Si el p-valor $> 0.05 \Rightarrow$ No rechazamos $H_0 \Rightarrow$ La serie No es estacionaria en media \Rightarrow

Diferencio

Aplicando el test de Dickey-Fuller sobre los datos, obtenemos:

```
Augmented Dickey-Fuller Test
data: datos.train.ts
Dickey-Fuller = -5.579, Lag order = 30, p-value = 0.01
alternative hypothesis: stationary
```

Figura 15. Estacionariedad en media

Como el p-valor es menor que 0.05, rechazamos la hipótesis nula, y corroboramos que la serie es estacionaria en media.

II. Estacionaria en varianza:

Se evalúa la necesidad de transformar la serie para hacerla estacionaria en varianza. Para ello, se utiliza la transformación Box-Cox que viene dada para diferentes valores de λ por la siguiente expresión (Soage, s.f.):

$$\begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log(x) & \text{si } \lambda = 0 \end{cases}$$

Donde según el valor que tome λ se proponen las siguientes transformaciones:

λ	-2	-1	-0.5	0	0.5	1	2
Transformación	$1/x^2$	$1/x$	$1/\sqrt{x}$	$\log(x)$	\sqrt{x}	x	x^2

En la siguiente gráfica se muestran los valores óptimos de λ

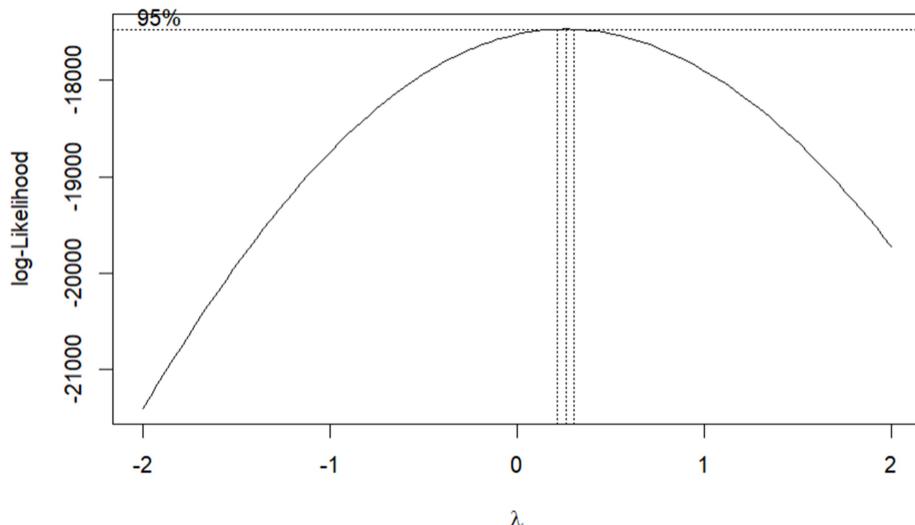


Figura 16. Representación gráfica de λ

El valor que maximiza la transformación Box-Cox es $\lambda = 0.2626$, y aunque esté igual de próximo a $\lambda=0$ y $\lambda=0.5$, se tomará como transformación de la serie: $y = \log(x)$

4.4.2. Fase de contraste

Para ajustar la serie tendremos que determinar cuáles son los parámetros p, P, d, D, q, Q , del modelo SARIMA $(p, d, q)x(P, D, Q)_S$. Las funciones de autocorrelación parcial y simple de los datos de entrenamiento, una vez realizada la transformación logarítmica, son las siguientes:

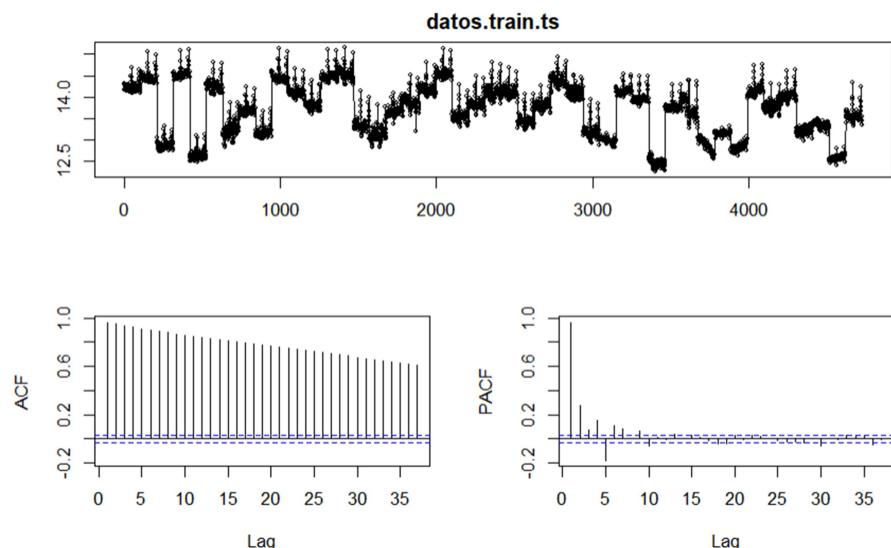


Figura 17. FAS y FAP para la muestra de entrenamiento.

Debido a que hay infinitos “palitos” en la parte autorregresiva, se propone un MA(1)

- **SARIMA ($0, 0, 1$) $x(0, 0, 0)_{12}$**

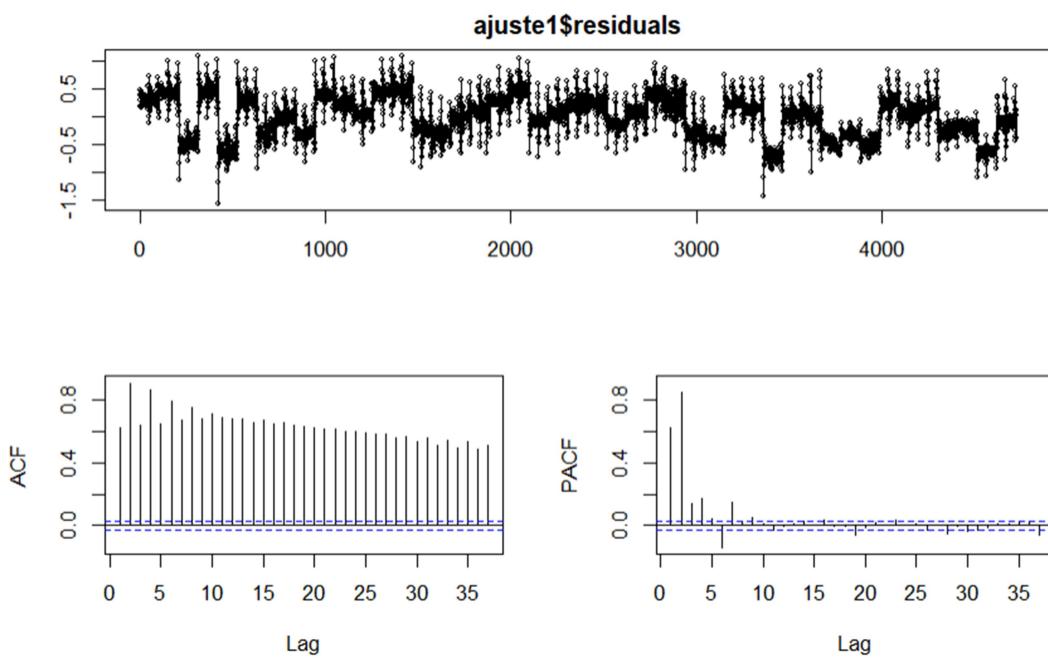


Figura 18. FAS y FAP para el ajuste 1.

Observamos que sobresalen de la banda de Bartlett aún muchas correlaciones. Se propone una diferenciación:

- **SARIMA $(0, 1, 1)x(0, 0, 0)_{12}$**

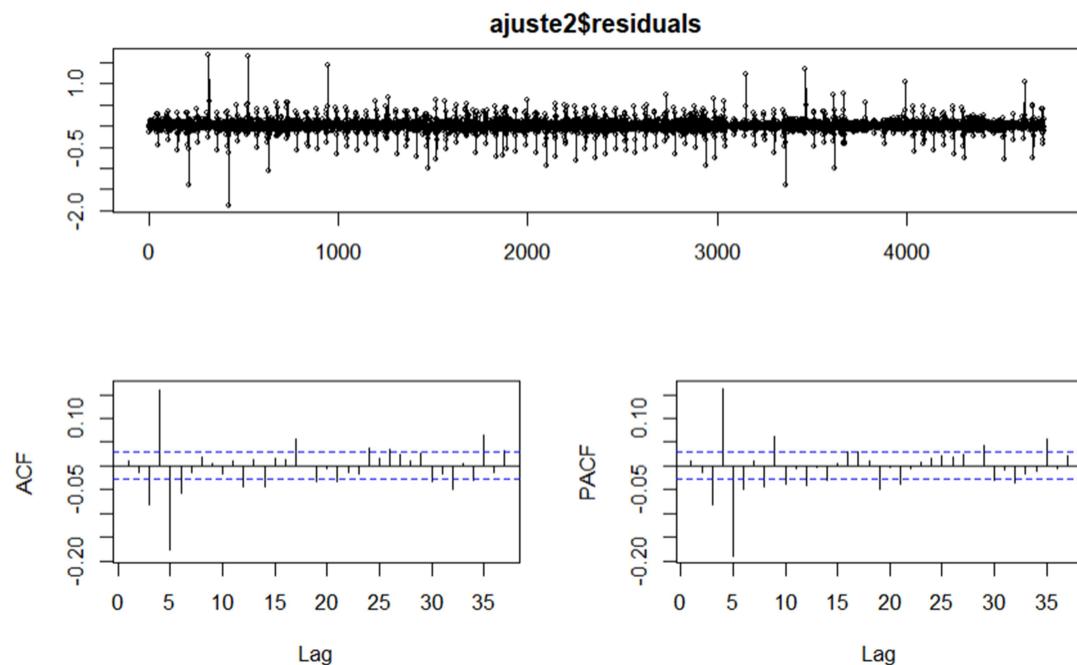


Figura 19. FAS y FAP para el ajuste 2.

Si nos fijamos ya son muchas las correlaciones que han quedado dentro del interior de las bandas. Observamos que el valor de $p=4$ y $q=4$ queda fuera, por lo que se propone un AR(4):

- **SARIMA $(4, 1, 1)x(0, 0, 0)_{12}$**

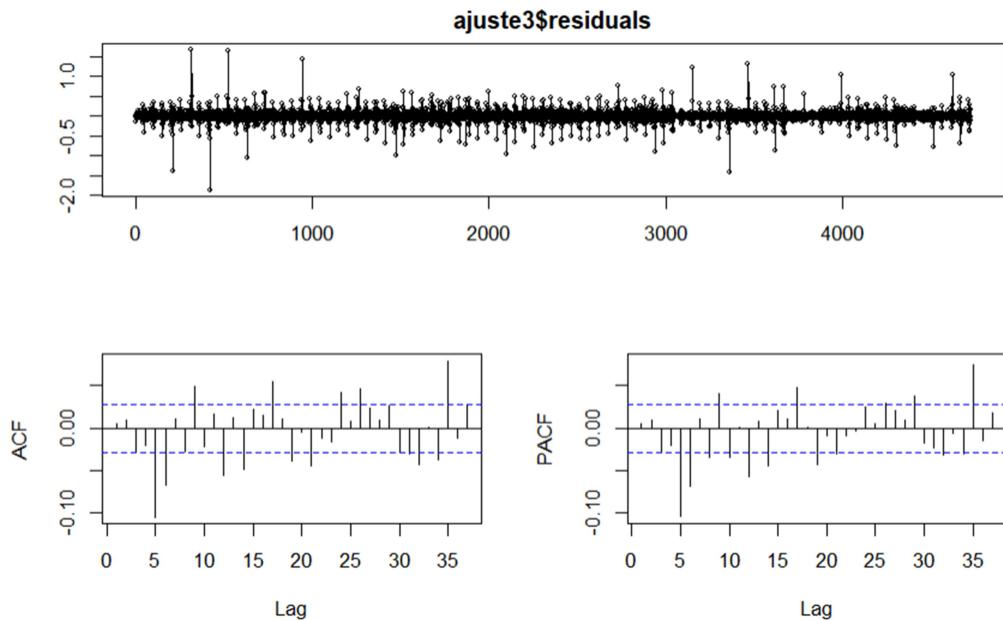


Figura 20. FAS y FAP para el ajuste 3.

En este caso el modelo es peor que el propuesto anteriormente, así que se propone aplicar el modelo Auto Arima para obtener el modelo óptimo.

- **SARIMA (0, 1, 5)x(0, 0, 0)₁₂**

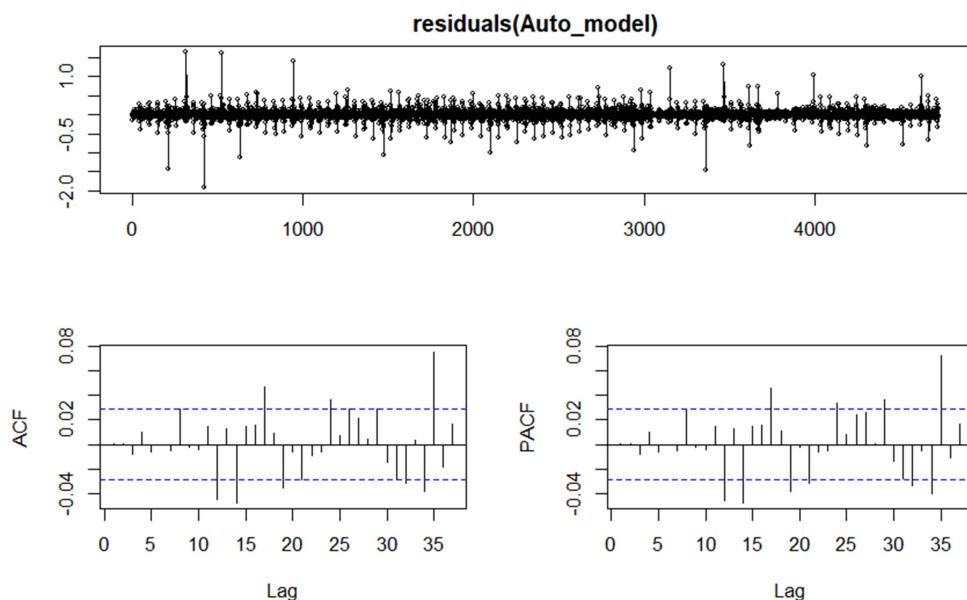


Figura 21. FAS y FAP para el auto modelo.

Observamos que las correlaciones con $p, q=4$ y $p, q=5$ ya no aparecen, pero sí lo hacen otras cuando el valor del *lag* aumenta. Así que se propone añadir un AR(12) y una diferenciación estacional:

- **SARIMA (0, 1, 5)x(1, 1, 0)₁₂**

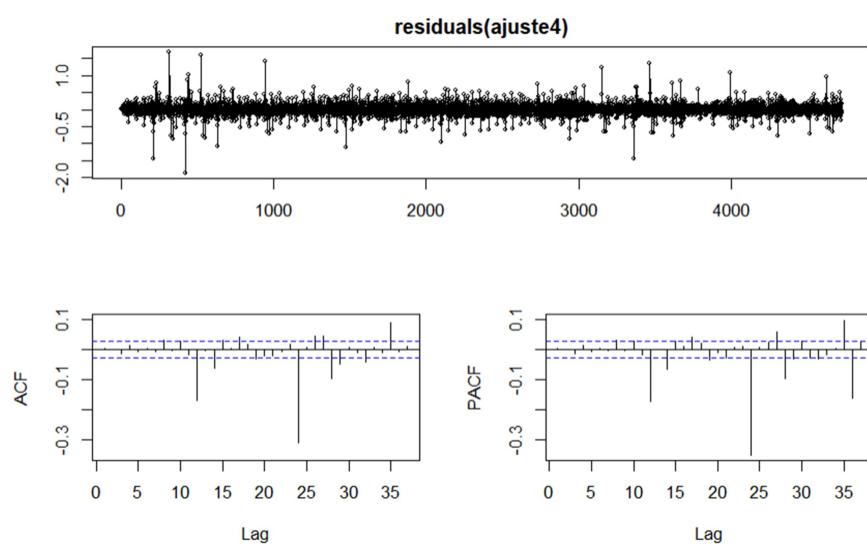


Figura 21. FAS y FAP para el ajuste 4.

Este modelo es mucho más bueno que el propuesto por el Auto Arima, ya que quedan muchas menos correlaciones fuera de la banda de Bartlett.

Para intentar solventar que aun queden algunas correlaciones fuera, se tendrá en cuenta como variable regresiva `IsHoliday`, junto con los outliers que se detecten, y se introducirán en el modelo.

- **SARIMA (0, 1, 5)x(1, 1, 0)₁₂ + wIt^{festivos}**

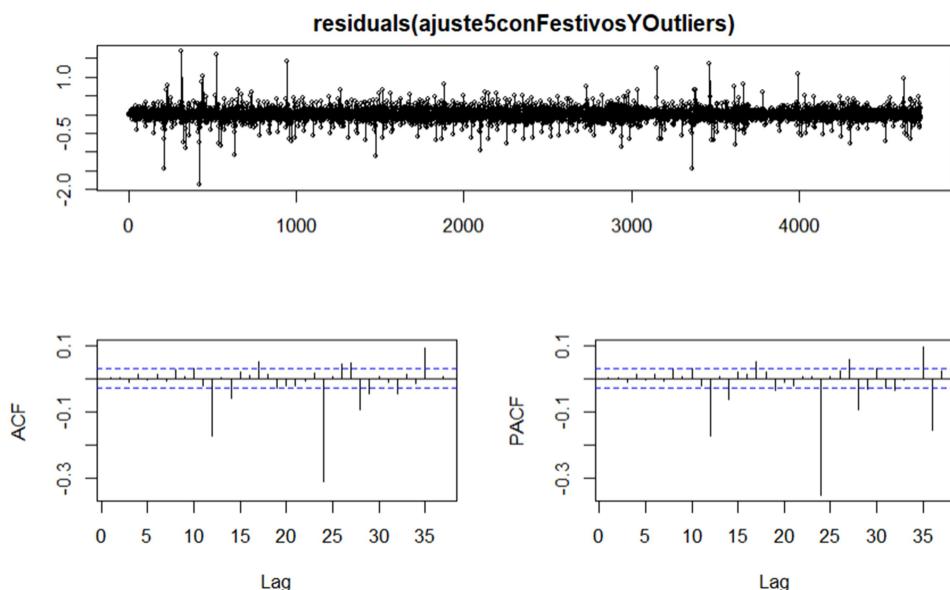


Figura 22. FAS y FAP para el ajuste 5 con festivos y outliers.

Se vuelven a obtener las mismas funciones de autocorrelación parcial y simple que antes.

z test of coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
ma1	-0.2956365	0.0143890	-20.5460	< 2.2e-16	***
ma2	-0.0312413	0.0148943	-2.0975	0.0359462	*
ma3	-0.0720590	0.0155928	-4.6213	3.814e-06	***
ma4	0.1787501	0.0138571	12.8996	< 2.2e-16	***
ma5	-0.2417841	0.0145224	-16.6491	< 2.2e-16	***
sar1	-0.5386461	0.0123281	-43.6927	< 2.2e-16	***
xreg	0.0187927	0.0054361	3.4570	0.0005462	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Figura 23. Coeficientes para el ajuste 5 con festivos y outliers.

Observamos que todos los coeficientes son estadísticamente significativos, ya que no hay ningún p-valor > 0.05, por lo que esta será nuestra propuesta de modelo final.

4.4.3. Fase de predicción.

En esta fase lo que buscamos es que una vez se haya ajustado el modelo a los datos con los que hemos trabajado, el modelo sea capaz de generar datos a futuro lo más parecidos a lo que en la realidad ocurrirá, es decir, con el menor error de predicción. Para ello, se utiliza la fórmula de error MAPE (Means Absolute Percentage Error):

$$Error = \frac{100 * |real - predicc\acute{o}n|}{real}$$

Ésta nos ayudará a poder decir si nuestro modelo es óptimo o no. El error global sobre histórico MAPE obtenido en el último modelo ajustado es de 11.7894, y el error global a futuro MAPE es 11.5386. Estos valores pueden deberse a la gran variabilidad de datos del modelo.

Para finalizar el análisis se realiza un gráfico con los datos utilizados en la muestra de entrenamiento, y con los estimados mediante el modelo SARIMA:

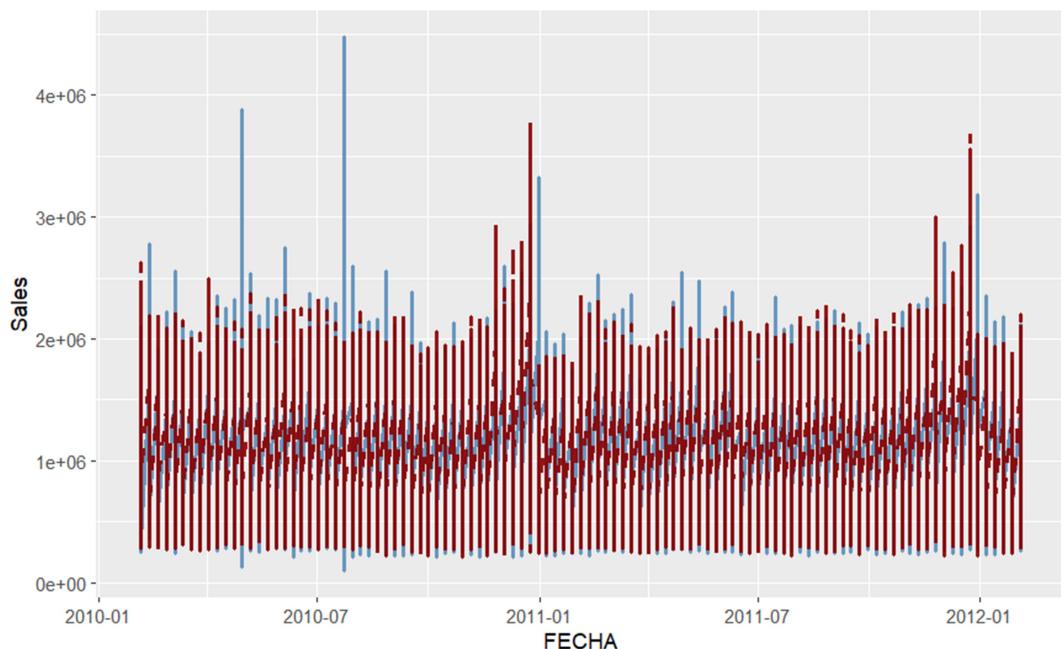


Figura 24. Datos de entrenamiento vs datos predichos.

Como podemos ver, el modelo seleccionado realiza una muy buena estimación sobre los datos con los que se han ido trabajando.

Para poder saber si este modelo se podría generalizar con otros datos y no existe un sobreajuste, se aplica con la muestra de validación:

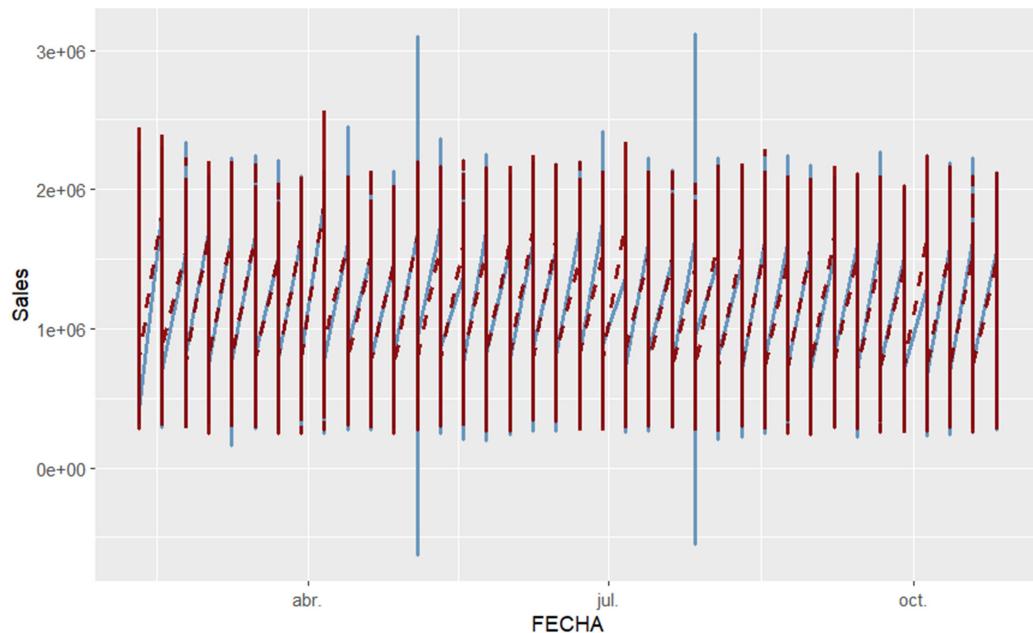


Figura 25. Datos de validación vs datos predichos.

Concluimos que el modelo obtenido funciona también con los datos de validación, por lo que se podría aplicar este ajuste sobre los datos del dataset Walmart para poder realizar una predicción a futuro.

5. Discusión de los resultados

Durante este apartado se van a analizar y discutir los resultados obtenidos mediante la implementación de las técnicas utilizadas en el apartado 4.

5.1. Clusterización

Con la implementación de los algoritmos de clusterización que se han aplicado sobre el dataset, se ha obtenido la siguiente información:

- A través del método del codo, y del método de la silueta se ha podido establecer el número óptimo de clústeres en 4.
- Se han podido aplicar diferentes métodos de Clustering como K-Means, Agglomerative Clustering y DBSCAN, obteniendo resultados muy similares en la segmentación de los datos.
- Se han clasificado los locales de Walmart en 4 clústeres, pudiendo determinar en qué lugares se produce un menor volumen de ventas.

5.2. Regresión lineal

Con las técnicas de modelado estadístico basado en la regresión lineal múltiple, se ha podido establecer:

- La implementación de un modelo basado en todas las variables regresoras.
- La eliminación de las variables que no son estadísticamente significativas sobre el modelo.
- El uso de la función StepAIC para estudiar si existe algún modelo mejor que el ya propuesto.
- Se cumple la no multicolinealidad entre las variables regresores.
- Se verifican las hipótesis de linealidad del modelo, analizando la normalidad, homocedasticidad e independencia de los residuos.

5.3. Series Temporales

La implementación de la metodología Box-Jenkins, basada en el empleo de modelos SARIMA que se ha aplicado en el apartado 4.4, ha proporcionado la siguiente información:

- Mediante el test de Dickey-Fuller se ha comprobado que la serie es estacionaria en media.
- Mediante la metodología Box-Cox se ha propuesto una transformación logarítmica para asegurarnos que la serie es estacionaria en varianza.
- Se han implementado los siguientes ajustes sobre la serie:
 - SARIMA (0, 0, 1)x(0, 0, 0)₁₂
 - SARIMA (0, 1, 1)x(0, 0, 0)₁₂
 - SARIMA (4, 1, 1)x(0, 0, 0)₁₂
 - SARIMA (0, 1, 5)x(0, 0, 0)₁₂
 - SARIMA (0, 1, 5)x(1, 1, 0)₁₂
 - SARIMA (0, 1, 5)x(1, 1, 0)₁₂ + wIt^{festivos}
- Se ha obtenido un error global sobre histórico MAPE de 11.7894, y un error global a futuro MAPE de 11.5386
- Se ha realizado una predicción muy precisa y muy parecida a los datos del dataset.

6. Conclusiones

Mediante este trabajo de fin de máster, se han podido alcanzar los objetivos propuestos inicialmente, y cuya resolución y explicación se va a detallar a continuación:

- A. Segmentar las tiendas y determinar en qué locales habría que realizar más campañas con descuentos.

Gracias a la implementación de las técnicas de clusterización, se han podido segmentar las tiendas en 4 grupos. También se ha podido determinar el valor medio de ventas por cada clúster, siendo el clúster 1 el que menor volumen medio de ventas tiene.

Es por esto que sería óptimo realizar campañas de ofertas y descuentos en las siguientes tiendas, con el objetivo de mejorar las próximas ventas en esos lugares:

Clúster 1 - Stores: 3, 5, 7, 9, 15, 16, 21, 25, 29, 30, 33, 36, 37, 38, 42, 43, 44

- B. Analizar el impacto de la temperatura y del precio del combustible sobre las ventas.

Mediante el ajuste del modelo de regresión lineal múltiple, recordemos que se obtuvo que la variable Temperature sí era estadísticamente significativa, mientras que el precio del combustible no lo era:

Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.961e+06	5.346e+05	3.669	0.000245	***
Store	-1.539e+04	5.225e+02	-29.457	< 2e-16	***
Date	2.555e+00	3.934e+01	0.065	0.948214	
IsHoliday	7.288e+04	2.606e+04	2.796	0.005187	**
Temperature	-9.767e+02	3.765e+02	-2.594	0.009504	**
Fuel_Price	8.325e+03	2.454e+04	0.339	0.734485	
CPI	-2.322e+03	1.903e+02	-12.206	< 2e-16	***
Unemployment	-2.181e+04	3.948e+03	-5.524	3.45e-08	***

Figura 26. Coeficientes del modelo de regresión lineal

Es decir, el precio del combustible no influye en el volumen de ventas que pueda tener el supermercado. Esto puede deberse a que son variables independientes, y que es irrelevante este precio, ya que los clientes seguirán realizando sus compras en el supermercado.

Por su parte, la variable temperatura sí que influye en las ventas de Walmart.

Si analizamos la estimación del modelo de regresión lineal, que recordemos que era:

$$\hat{Y} = 2.031.946 - 15.374X_1 - 2.346X_2 - 22.196.X_3 + 72.222.X_4 - 929.X_5$$

Y X_5 la variable Temperatura, observamos que la relación lineal entre el número de ventas y la temperatura es negativa. Esto quiere decir que a mayor temperatura, los clientes comprarán menos.

C. Analizar el impacto que tiene la tasa de desempleo sobre el número de ventas.

Volviendo a la ecuación de regresión lineal anterior, y recordando que la tasa de desempleo era la variable X_3 , observamos que la relación lineal entre el número de ventas y esta tasa también es negativa, es decir, a mayor tasa de desempleo se obtendrán menos ventas. Este resultado puede resultar muy lógico, y es que los ciudadanos que se encuentren en esta situación, tendrán que comprar los productos más básicos, o realizar compras cada mucho más tiempo.

D. Analizar el impacto que tienen las semanas con festivos sobre el número de ventas.

Una vez más, utilizando la expresión de la estimación del modelo de regresión, y recordando que la variable IsHoliday es X_4 , observamos que en este caso, la relación lineal es positiva. Esto quiere decir que las semanas con días festivos se producen mayores ventas.

Además, esta variable resulta significativa en la implementación de forecasting, por lo que resulta muy relevante almacenar este tipo de variable en un dataset para el análisis de datos que se quiera realizar, ya que como hemos podido observar, es muy influyente sobre las ventas de un comercio.

E. Predecir el volumen de ventas que tendrá la cadena de supermercados en unos meses.

Para predecir el volumen de ventas se ha implementado un modelo:

$$\text{SARIMA } (0, 1, 5) \times (1, 1, 0)_{12} + wIt^{\text{festivos}}$$

Si quisieramos predecir las ventas futuras, bastaría con saber los festivos que habrá en una semana, y aplicar las diferenciaciones junto con las autorregresiones y medias móviles con los datos ya conocidos, para poder determinar las ventas futuras.

Con este trabajo de fin de máster he podido comprobar la importancia que tiene almacenar cualquier tipo de variable que a priori podría parecer que no resulta relevante sobre alguna predicción que se quisiera realizar sobre la variable principal del análisis.

He podido observar además, el gran valor que supone para las empresas poder almacenar datos, y realizar análisis de datos con técnicas de Big Data para poder determinar patrones desconocidos, estimaciones o predicciones que son de gran utilidad para mejorar los ingresos de la compañía.

7. Referencias Bibliográficas

- Casino, F.** (6 de noviembre de 2020). Trabajaba de mesero y le pagaban con un plato de comida: conocé la increíble historia del genio que creó Walmart. IproUp.
<https://www.iproup.com/innovacion/18085-como-nacio-walmart-la-historia-de-su-creador-sam-walton>
- Garbade, M.** (12 de septiembre de 2018). Understanding K-means Clustering in Machine Learning. [Comentario en la página web Towards Data Science].
<https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>
- González, A.** (8 marzo de 2019). Segmentación utilizando K-means en Python. Machine Learning para todos. <https://machinelearningparatodos.com/segmentacion-utilizando-k-means-en-python/>
- Güemes, L.** (27 de agosto de 2020) Evolución de los supermercados, ¿cómo serán en el futuro? Chavsa. <https://www.chavsa.com/evolucion-los-supermercados-seran-futuro/>
- Hernández, O.** (4 de julio de 2022). Repositorio TFM Big Data & Business Analytics.
https://github.com/OlgaHernMartinez/TFM_BigData-BusinessAnalytics_OlgaHernandez
- InStoreView** (2 de octubre, 2020). Big Data en supermercados: ¿cuánto se puede saber de un consumidor? <https://www.instoreview.com/blog/big-data-en-supermercados-cuanto-se-puede-saber-de-un-consumidor#:~:text=Gracias%20al%20Big%20Data%2C%20los,de%20acuerdo%20a%20la%20demanda>
- Moreno, I.** (s.f.) Clustering DBSCAN. Stat Developer. <https://www.statdeveloper.com/clustering-dbscan/>
- Moya, R.** (12 de septiembre de 2016). Selección del número óptimo de Clústers. Jarroba.com. <https://jarroba.com/seleccion-del-numero-optimo-clusters/>
- PyShark** (5 de marzo de 2022) Dunn Index for K-Means Clustering Evaluation. Python-Bloggers.<https://python-bloggers.com/2022/03/dunn-index-for-k-means-clustering-evaluation>
- Rodrigues, A.** (2018). Repositorio de Github. Dunn-sklearn.py.
<https://gist.github.com/keizerzilla/a72056c0905e57a036e57b03b557c896>
- Rodríguez, D.** (22 de abril de 2020). Solucionar la multicolinealidad con VIF. Analytics Lane. <https://www.analyticslane.com/2020/04/22/solucionar-la-multicolinealidad-con-vif/>
- Soage, J.C.** (s.f.). Transformación de Box Cox en R. R CODER. <https://r-coder.com/transformacion-box-cox-r/>

Solares, C. (s.f.) Cómo afecta el clima en el comportamiento del consumidor y su decisión de compra. Neuromarketing. <https://neuromarketing.la/2017/10/afecta-el-clima-en-el-comportamiento-del-consumidor/#:~:text=Seg%C3%BAn%20Weather%20Unlocked%2C%20el%20clima,usamos%2C%20qu%C3%A9%20escuchamos%2C%20etc>.

Walmart Analysis Dataset. Kaggle. <https://www.kaggle.com/datasets/tanujdhiman/walmart-analysis-dataset?select=walmart.csv>