Luis Alfredo García Oquendo

Análisis de ventas y predicciones de ingresos

Máster en Big Data, Data Scientics y Analítica de Datos

2023-2024

Sinopsis

Este estudio aborda la predicción de las ventas semanales de 45 tiendas en Estados Unidos, mediante la implementación de modelos de regresión lineal, árboles de decisión, vecinos más cercanos y máquinas de soporte vectorial en Python, utilizando datos del periodo 2010 al 2012. A través de validación cruzada y optimización de hiperparámetros, se identificó que los modelos basados en árboles explicaban más del 94 % de la variabilidad de los datos. Además, se determinó que cinco variables clave (Store, CPI, Unemployment, Holiday_Events y Week) tenían una mayor influencia en las predicciones de los modelos en comparación con el conjunto total de características. Asimismo, se observó que el modelo con mejor rendimiento (XGBRegressor) presentaba heterocedasticidad en los residuos, lo que limitaba su capacidad para capturar las relaciones entre las variables predictoras y la variable objetivo. Estos hallazgos sugieren la necesidad de un análisis más detallado de los valores atípicos y las variables desbalanceadas (semanas festivas) con el fin de mejorar la precisión de las predicciones.

Tabla de contenido

Índice de figuras

Índice de tablas

Objetivos

Objetivo general:

- Entrenar y evaluar múltiples modelos de regresión lineal en Python, con datos del periodo 2010 y 2012, para predecir las ventas semanales de 45 tiendas en Estados Unidos, determinando cuál ofrece el mejor ajuste a los datos analizados.

Objetivos específicos:

- Diseñar una metodología en SQL para la limpieza, transformación y normalización de datos en la base de datos.
- Identificar valores atípicos mediante el análisis del incremento porcentual y procesarlos para garantizar la integridad y calidad de las series de tiempo.
- Analizar las fechas de las ventas para determinar los meses, trimestres y años con mayor actividad, con el fin de identificar patrones y tendencias temporales.
- Evaluar las ventas regionales mediante el análisis de ventas promedio, variación de ventas y ventas totales, con el objetivo de detectar las ciudades y tiendas líderes en ventas.
- Visualizar gráficamente el impacto de variables predictoras (temperatura, tasa de desempleo, índice de precios al consumidor y precio del combustible) sobre las ventas récord por ciudad y tienda, con el fin de identificar patrones en su comportamiento.
- Calcular el grado de sesgo y las correlaciones entre las variables predictoras y la variable objetivo, para determinar cuáles tienen mayor influencia en el desempeño de los modelos.
- Seleccionar las variables más influyentes en la toma de decisiones de los modelos mediante la técnica de eliminación recursiva de características.
- Validar el modelo con mejor desempeño mediante pruebas estadísticas (Residual Plot, Breusch-Pagan test, Durbin-Watson test, Q-Q plot y Shapiro-Wilk test) para verificar el cumplimiento de los supuestos de la regresión lineal.
- Diseñar un informe interactivo en Power BI que cargue y procese datos desde la base de datos, integre modelos predictivos entrenados en Python para la predicción de ventas, y permita la visualización dinámica de estadísticas que apoyen la toma de decisiones.

Introducción

En la mayoría de los proyectos de Machine Learning, es común recurrir a metodologías estructuradas como Knowledge Discovery in Databases (KDD), SEMMA, CRISP-DM y CRISP-ML para garantizar resultados efectivos, reproducibles y alineados con los objetivos del proyecto, como lo señala Big Data International Campus (1). Entre estas metodologías, CRISP-DM se ha destacado por su flexibilidad y amplio uso en diferentes dominios, gracias a su enfoque claro y sistemático para guiar el desarrollo de proyectos de análisis de datos.

Muchos de estos proyectos se centran en la predicción de variables económicas, atmosféricas, ambientales, de salud, entre otras, utilizando técnicas de aprendizaje supervisado y no supervisado. Estas metodologías han demostrado ser efectivas en distintos ámbitos, como la predicción de ventas en tiendas minoristas (2), la estimación de ventas en restaurantes (3), el pronóstico de la progresión y tasa de infección del COVID-19 (4) y predicciones razonablemente precisas de la tendencia del nivel del mar en regiones específicas para periodos de uno a tres años (5). Sin embargo, para garantizar la reproducibilidad y validez de estos proyectos, es fundamental aplicar una metodología estructurada.

Dado lo anterior, este trabajo tiene como objetivo entrenar un modelo predictivo de regresión lineal con datos del período comprendido entre 2010 y 2012 para estimar las ventas semanales de tiendas ubicadas en distintas ciudades de Estados Unidos, utilizando la metodología CRISP-DM como marco de referencia. Para ello, se emplea un conjunto de datos compuesto por series temporales de ventas semanales, junto con variables predictoras como temperatura, precio de combustible, índice de precios al consumidor (CPI), tasa de desempleo, días festivos y fechas especiales.

La preparación de los datos se lleva a cabo mediante SQL , realizando tareas de limpieza, transformación y normalización de los datos. Posteriormente, se entrenan y evalúan distintos modelos de regresión de Machine Learning, incluidos XGBRegressor, RandomForestRegressor, GradientBoostRegressor, LinearRegression y SVR. La selección del modelo con mejor desempeño se realiza mediante validación cruzada y optimización de hiperparámetros con Optuna, con el objetivo de maximizar el coeficiente de determinación R^2.

Finalmente, los resultados del modelo y los análisis descriptivos se integran en Power BI, proporcionando una herramienta interactiva para visualizar los datos y facilitar la toma de decisiones estratégicas en la gestión de ventas.

Cronología

Los proyectos de Machine Learning requieren de una metodología que garantice resultados efectivos, reproducibles y alineados con los objetivos del proyecto. CRIPS-DM es una metodología estructurada y orientada a extraer el valor de los datos (7). Esta se conceptualiza en seis fases como se evidencia en la Figura 1.

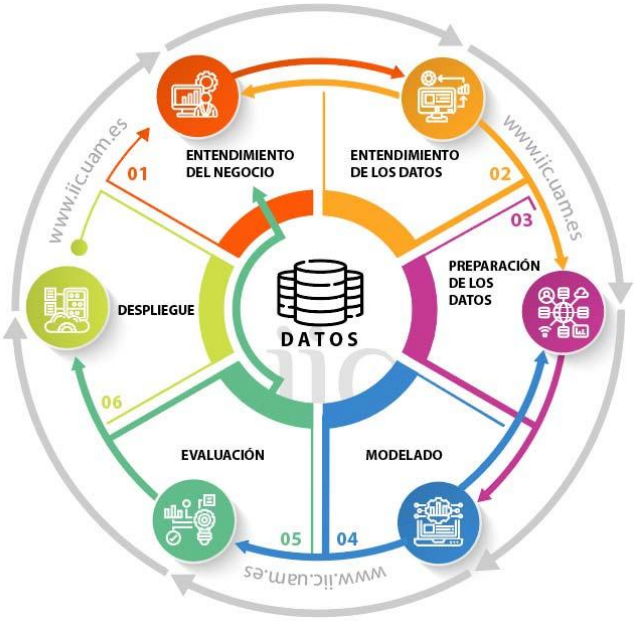


Figura 1. Metodología CRIPS-DM.

Entendimiento del negocio

En la fase de entendimiento del negocio, se identificaron los objetivos y requisitos del trabajo de fin de máster titulado “Análisis de ventas y predicciones de ingresos”. El propósito general fue desarrollar un informe interactivo en Power BI que permita analizar el comportamiento de las ventas y proyectar posibles ingresos futuros. Para ello, se aplicaron modelos de regresión lineal previamente entrenados que generan predicciones facilitando la toma de decisiones estratégicas en la gestión comercial. Esto permitirá a los usuarios explorar los datos de manera interactiva y tomar decisiones informadas sobre las estrategias comerciales a seguir.

Entendimiento de los datos

En la fase de entendimiento de los datos, se llevaron a cabo consultas en la base de datos para identificar las tablas y variables que necesitaran ser limpiadas, transformadas o estandarizadas. Con base a estas consultas se limpiaron y/o transformaron las tablas garantizando la limpieza e integración de los datos mediante código SQL.

Posteriormente, en Python al analizar las ventas semanales mediante un diagrama de cajas y bigotes, se observaron numerosos valores atípicos, y en el gráfico de series de tiempo, se detectó un exceso de ruido. Estos valores atípicos se eliminaron y se dio paso a la exploración de los datos.

Exploración de datos

La exploración de datos tuvo como objetivo obtener una visión general del estado de los datos. Dado que estos rara vez son perfectos, es imprescindible organizarlos, comprender su contenido, identificar las variables más relevantes y analizar las relaciones entre ellas. Además, es necesario detectar patrones, gestionar valores faltantes y tratar datos atípicos para finalmente extraer conclusiones significativas.

Estas actividades conforman lo que se denomina análisis exploratorio de datos, el cual puede definirse como el proceso de comprender, visualizar y extraer información clave de un conjunto de datos con el fin de determinar la estrategia o técnica más apropiada para su procesamiento posterior. Este análisis es fundamental y siempre constituye el primer paso en cualquier proyecto de machine learning o ciencia de datos (8).

En este proceso, se plantearon las siguientes preguntas claves:

- ¿Qué características permiten realizar predicciones con una mayor probabilidad de éxito?
- ¿Existen patrones que puedan ayudar a predecir los ingresos futuros de cada una de las tiendas?
- ¿Es viable implementar un modelo de regresión lineal para realizar estas predicciones?

En consecuencia, se utilizaron estadísticos descriptivos como el promedio, medidas de dispersión (mínimo y máximo), desviación estándar y coeficiente de variación, los cuales permitieron analizar:

- La distribución de las ventas por mes, trimestre y año.
- Las ventas a nivel regional.
- El comportamiento general de las variables, incluyendo sesgo, valores atípicos y correlaciones.

El *promedio* es la suma de todos los valores dividida por el número de valores (9). La fórmula para calcular la media de un conjunto de n valores x_1, x_2, … x_n se define por:

$$Media = \bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$$

El *mínimo* es el valor más pequeño o menor dentro de un conjunto de datos, representando así el límite inferior de las ventas registradas(10). Matemáticamente, si tienes un conjunto de datos X= {x_1, x_2, ..., x_n}, el mínimo se denota como:

$$\min(X) = \min(x_1, x_2, \ldots, x_n)$$

El valor *máximo* es el valor más grande o mayor dentro de un conjunto de datos y representa el límite superior de los valores observados (10). El máximo está denotado como:

$$\max(X) = \max(x_1, x_2, \ldots, x_n).$$

La *desviación* es una medida estadística que indica que tan dispersos están los datos de un conjunto con respecto a la media(9). En otras palabras, refleja cuánto se alejan, en promedio, los datos individuales de la media del conjunto. La fórmula para calcular la desviación estándar de un conjunto de n valores $x_1, x_2, \ldots x_n$, está definida por:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{x})^2}{n}}$$

El *coeficiente de variación,* también conocido como coeficiente de variación de Spearman, es una herramienta estadística que mide la dispersión relativa de un conjunto de datos. Esta métrica se emplea ampliamente en estadística, ya que relaciona la desviación estándar con la media aritmética de los datos. En esencia, el coeficiente de variación permitió medir el grado de la variabilidad relativa del conjunto de datos en comparación con su media (11). La fórmula para calcular el coeficiente de variación está dada por:

$$CV = \frac{\sigma}{\bar{x}} * 100\%$$

Preparación de los datos

Con la visión general obtenida en las etapas anteriores, se generó un nuevo conjunto de datos limpio y depurado de valores atípicos. Además, se incorporó la fecha como variable predictora, lo que implicó la creación de cinco nuevas características: año, mes, día, día de la semana y semana del año. La fecha, por su parte, se estableció como el índice del dataframe y se usó en la etapa del modelado de datos.

Modelado de datos

En esta etapa se entrenaron varios modelos de regresión lineal en su configuración predeterminada, y fueron elegidos los modelos que alcanzaron un coeficiente de determinación mayor al 60%. El coeficiente de determinación R^2 “es una importante medida de la regresión lineal que representa la parte de la variación de la variable objetivo que viene explicada por un conjunto de variables independientes; por lo tanto, mide la capacidad predictiva del modelo estimado(12).”

Varios modelos superaron el umbral del 60%; no obstante, los modelos elegidos para la siguiente fase fueron XGBRegressor, RandomForestRegressor, GradientBoostingRegressor, KNeighborsRegressor. Estos modelos fueron sometidos a optimización de hiperparámetros y validación cruzada con el objetivo de identificar el modelo más adecuado para el conjunto de datos.

Evaluación

Una vez identificado el modelo óptimo en la etapa de modelado, se procedió a evaluar sus predicciones. Esto incluyó analizar las variables que el modelo consideraba más relevantes para las predicciones, utilizando herramientas como SHapley Additive exPlanations (SHAP) y eliminación recursiva de características (RFE).

El método **RFE** elimina iterativamente características y construye modelos con las variables restantes, permitiendo identificar aquellas que tienen mayor impacto en la predicción de la variable objetivo, según la precisión del modelo(13). Por otro lado, **SHAP** asigna un valor de importancia a cada característica del modelo, proporcionando una interpretación detallada de su impacto en las predicciones (14).

De esta forma, se determinaron las características que el modelo consideraba más importantes al momento de hacer las predicciones. Posteriormente, utilizando únicamente estas variables relevantes, se realizaron validaciones para verificar si el modelo cumplía con los supuestos de la regresión lineal.

Esto implicó evaluar varios aspectos clave: la distribución normal de los residuales, la existencia de una relación lineal entre las variables dependientes e independientes, la ausencia de multicolinealidad asegurando que las variables independientes no estuvieran altamente correlacionadas entre sí, y finalmente, la validación de la homocedasticidad, asumiendo que la varianza de los errores residuales se mantuviera constante a lo largo de los valores de las variables independientes (6).

De este modo, se evaluó la capacidad predictiva del modelo y se determinó si era adecuado para realizar inferencias estadísticas.

Despliegue

Este modelo fue implementado en Power BI, donde se desarrolló un informe interactivo capaz de cargar y procesar datos directamente desde una base de datos en SQL Server. Además, se integraron tres modelos predictivos para generar pronósticos de ventas y presentar, de manera dinámica, estadísticas claves que ayudarán a tomar decisiones informadas basadas en los resultados de los tres modelos predictivos.

Detalles y justificación

En las fases previamente mencionadas, se presentó un resumen general de los procesos implementados para la ejecución del proyecto. A continuación, se detalla cada una de las etapas descritas, justificando las decisiones tomadas desde un punto de vista técnico y práctico. Además, se incluyen los resultados gráficos obtenidos y las explicaciones correspondientes, con el objetivo de brindar una visión completa y fundamentada del desarrollo del proyecto.

Entendimiento de los datos

El entendimiento de los datos se dividió en cuatro etapas fundamentales: reconocimiento de los datos mediante consultas en SQL, limpieza y transformación de datos en SQL, detección y eliminación de valores atípicos utilizando Python, y, finalmente, la exploración de los datos.

Consultas en SQL

Las consultas SQL se realizaron con el objetivo de comprender el estado de la información en cada una de las tablas y obtener una visión clara de los procesos necesarios para garantizar la integridad y calidad de los datos. Estas consultas se enfocaron en identificar columnas con datos duplicados, incumplimiento de formatos, caracteres especiales no permitidos, direcciones de correo electrónico con formatos incorrectos y la verificación de variables numéricas.

Para los campos de nombres, fechas, correos electrónicos y números, se implementaron diversas validaciones. En el caso de los nombres, se identificaron y corrigieron espacios consecutivos al inicio, en medio y al final de cada registro, además de detectar símbolos ajenos al alfabeto español. Para las fechas, se garantizó el cumplimiento de un formato estándar (YYYY-MM-DD HH:mm:ss). En los correos electrónicos, la validación se realizó conforme a las especificaciones del estándar RFC 3696 (15), asegurando que cada registro no excediera los 320 caracteres, incluyera un "@" seguido de un punto, y que la longitud del dominio (parte posterior al "@") no superara los 255 caracteres. Finalmente, para las variables numéricas, se verificó que no contuvieran caracteres alfabéticos.

Este análisis permitió identificar que, en la tabla "**Store**", varias columnas presentaban inconsistencias en nombres de ciudades, fechas, códigos postales, correos electrónicos y números de teléfono, los cuales requerían corrección (ver Tabla 1).

Tabla 1. Registros por corregir en la tabla Store de la base de datos TFM.

Columna	# de registros	Observaciones
City	32	Requieren ser transformados y capitalizados
Open_Date	3	Requieren convertirse a un formato válido de fecha
Zip_Code	9	Contienen caracteres no numéricos
Phone	35	Contienen caracteres no numéricos
Email	13	No cumplen con el formato de un correo

Por otro lado, en la tabla **"Walmart_Sales"** se constató que todas las columnas estaban libres de registros nulos. Sin embargo, se identificó que era necesario transformar las columnas a sus respectivos formatos, con excepción de aquella destinada a las fechas, ya que estaba correctamente configurada para almacenar valores de tipo fecha.

Limpieza y transformación en SQL

Con este panorama general sobre el estado de los datos, se llevaron a cabo todas las correcciones necesarias, diseñando algoritmos específicos para cada tabla. En el caso de la tabla **"Store"**, se desarrolló un algoritmo que incluye funciones especializadas, entre las cuales destaca la función **CleanToText**. Esta función, inspirada en la propuesta de Gavin Clayton(16), procesa una columna de texto identificando la posición del primer carácter que no sea una letra, un espacio o un guion bajo. A partir de este punto, se ejecuta un bucle *while* que elimina iterativamente los caracteres especiales detectados. Posteriormente, se reemplazan los guiones bajos por espacios y se ejecuta un segundo bucle *while* para eliminar espacios consecutivos, garantizando un formato uniforme y limpio en los datos de la columna procesada.

También se desarrolló una función llamada **"InitCap"**, diseñada para capitalizar los nombres de las ciudades, basada en el comportamiento de su homónima en Oracle (17). Esta función recibe como parámetro una columna y transforma cada registro de la siguiente manera: convierte la primera letra de cada palabra en mayúscula y, cada vez que se encuentra un carácter no alfanumérico (como un espacio, un signo de puntuación, etc.), se asegura de que la letra siguiente también se convierta en mayúscula. El resto de las letras se transforma a minúsculas, garantizando un formato uniforme y adecuado en los datos.

Para el manejo de fechas, se utilizó la función ISDATE, la cual devuelve un valor de 0 cuando una fecha no cumple con un formato estándar. En los casos donde se identificaron registros inválidos, se extrajeron los componentes de la fecha (año, mes, día, hora, minutos y segundos) y se transformaron al formato estándar “YYYY-MM-DD HH:mm:ss”, asegurando la integridad y consistencia de los datos temporales.

En el caso de los datos numéricos, se diseñó la función “**CleanToNum**”, la cual procesa una columna numérica para garantizar que solo contenga valores válidos. La función emplea un bucle *while* para eliminar de manera iterativa los caracteres no numéricos presentes en los registros. Posteriormente, se eliminan los espacios al inicio, en medio y al final de cada valor, asegurando que el resultado final sea un dato estrictamente numérico y libre de inconsistencias.

Finalmente, para la corrección de direcciones de correo electrónico, se analizaron los registros de la columna Email mediante consultas SQL, identificando múltiples casos que no cumplían con el formato estándar. En algunos registros, el carácter @ era reemplazado por **at** o el símbolo **#**, mientras que, en otros, el punto (.) era sustituido por una coma (,). Para corregir estos problemas, se implementó un proceso de sustitución que reemplaza estos caracteres incorrectos por los correspondientes, asegurando que todas las direcciones de correo electrónico cumplan con el formato adecuado.

Para la tabla **Walmart_Sales**, se desarrolló un algoritmo independiente que, en primer lugar, verifica el formato de cada columna y la transforma a un tipo de dato adecuado. En el caso de las columnas **Temperature**, **Fuel_Price** y **Unemployment**, los datos se transformaron de texto a VARCHAR y posteriormente a FLOAT. Por otro lado, las columnas **Store** y **Holiday_Flag** se convirtieron al formato INT. Finalmente, la columna **Date** se transformó a VARCHAR.

Con las columnas transformadas a estos formatos, procedimos a realizar las correcciones necesarias. Empezamos por la fecha, la cual se transformó del formato AAAA-MM-DD al formato DD-MM-AAAA.

Posteriormente, se creó la columna Holiday_Events, la cual clasifica los eventos especiales en las siguientes categorías: Super Bowl, Labour Day, Thanksgiving y Christmas, asignándolos a las fechas correspondientes:

- Super Bowl: 12-Feb-2010, 11-Feb-2011, 10-Feb-2012, 8-Feb-2013.
- Labour Day: 10-Sep-2010, 9-Sep-2011, 7-Sep-2012, 6-Sep-2013.
- Thanksgiving: 26-Nov-2010, 25-Nov-2011, 23-Nov-2012, 29-Nov-2013.
- Christmas: 31-Dec-2010, 30-Dec-2011, 28-Dec-2012, 27-Dec-2013

Finalmente, se procesó la columna CPI para cumplir con las especificaciones requeridas. Dado que este campo debía representarse como un valor numérico con tres cifras enteras y el resto en decimales, se insertó un punto decimal después del tercer dígito, garantizando así el formato correcto.

Detección y eliminación de valores atípicos

Con las tablas completamente depuradas y transformadas, se inició el proceso de identificación y eliminación de valores atípicos. En esta etapa, se generaron diagramas de cajas y bigotes, los cuales permitieron identificar que las variables Weekly_Sales, Temperature y Unemployment presentaban valores atípicos (ver Figura 2)

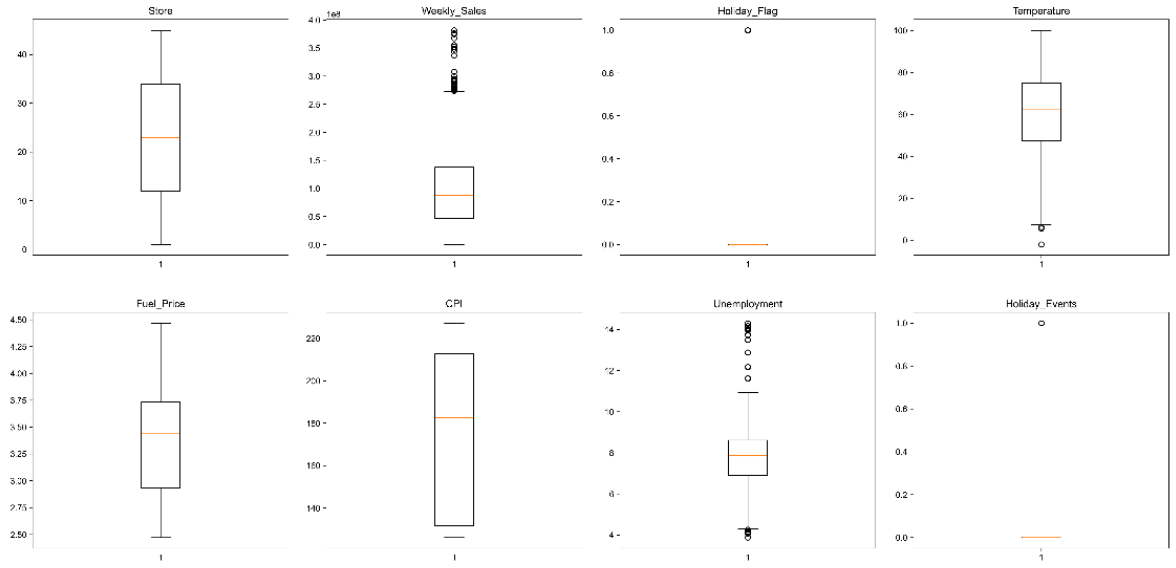


Figura 2. Valores extremos y rango de datos.

Posteriormente, las variables identificadas se graficaron para analizar su comportamiento a lo largo del tiempo. A través de una inspección visual, se determinó que la variable con mayor nivel de ruido era Weekly_Sales (ver Figura 3), lo que permitió centrar los esfuerzos en la limpieza específica de esta serie temporal.

Figura 3. Comportamiento general de las variables sin depuración de valores atípicos.

Dado que el conjunto de datos incluía un total de 45 tiendas, se implementó una función para identificar valores atípicos mediante el análisis del incremento porcentual. Para ello, se calculó una media móvil de tres pasos, la cual sirvió como referencia para evaluar las variaciones en los datos. A partir de esta media móvil, se determinó el incremento porcentual en cada instante de tiempo, midiendo así la variación de la serie respecto a la media calculada. Los valores atípicos, identificados a partir de los extremos mostrados en el diagrama de cajas y bigotes del incremento porcentual, se resaltaron en color rojo en las series temporales (ver Figura 4).

Adicionalmente, se graficó la media móvil calculada y el porcentaje de variación con respecto a la media.

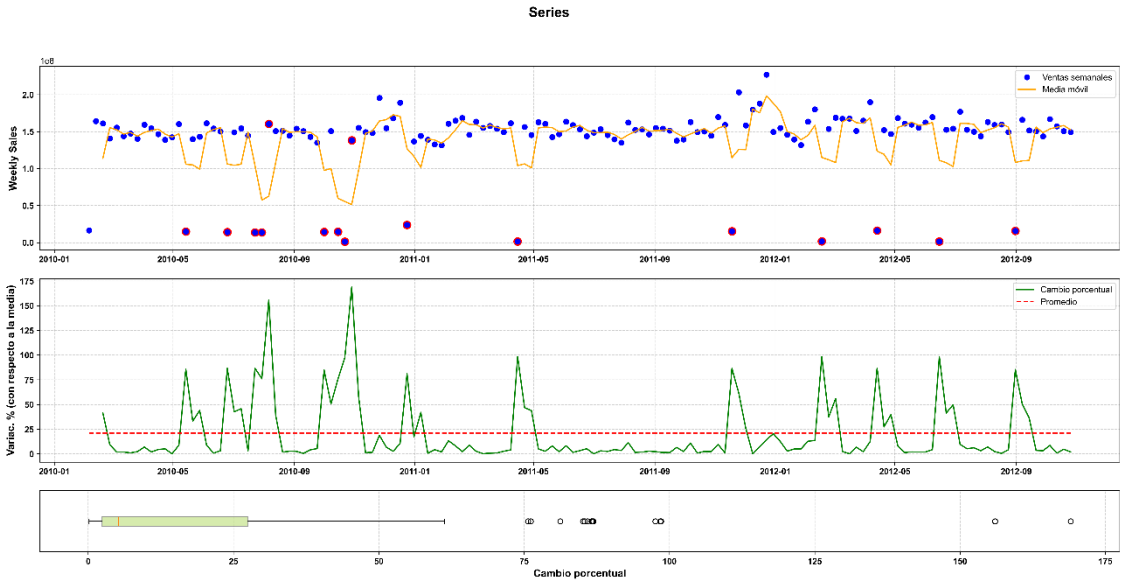


Figura 4. Incremento porcentual y valores atípicos en las ventas semanales.

En este caso, se observó que la serie presentó valores atípicos inferiores a 0.5×10^8, los cuales destacaban significativamente en comparación con el comportamiento general de la serie. Por otro lado, la media móvil reflejó una tendencia estable a lo largo del tiempo; no obstante, se identificaron picos y valles que se desviaban del patrón general de la serie. Estos puntos coincidían con los valores atípicos mencionados anteriormente, es decir, aquellos menores a 0.5×10^8.

Por otro lado, se observó que el porcentaje de variación oscilaba entre 0% y 175%, con un promedio cercano al 24%, lo que indicaba presencia de fluctuaciones significativas en los datos. Además, el diagrama de cajas y bigotes reveló que los valores atípicos se encontraban por encima del 60% del incremento porcentual, correspondiendo a ventas con comportamientos anómalos.

Dado que el incremento porcentual identificó una gran cantidad de valores atípicos, se decidió eliminar aquellos que se desviaban del comportamiento general de la serie. Este proceso se realizó de manera manual, tienda por tienda, hasta lograr que todas las series quedaran depuradas. Como resultado, las series limpias mostraron un comportamiento general consistente con el patrón presentado en la Figura 5.

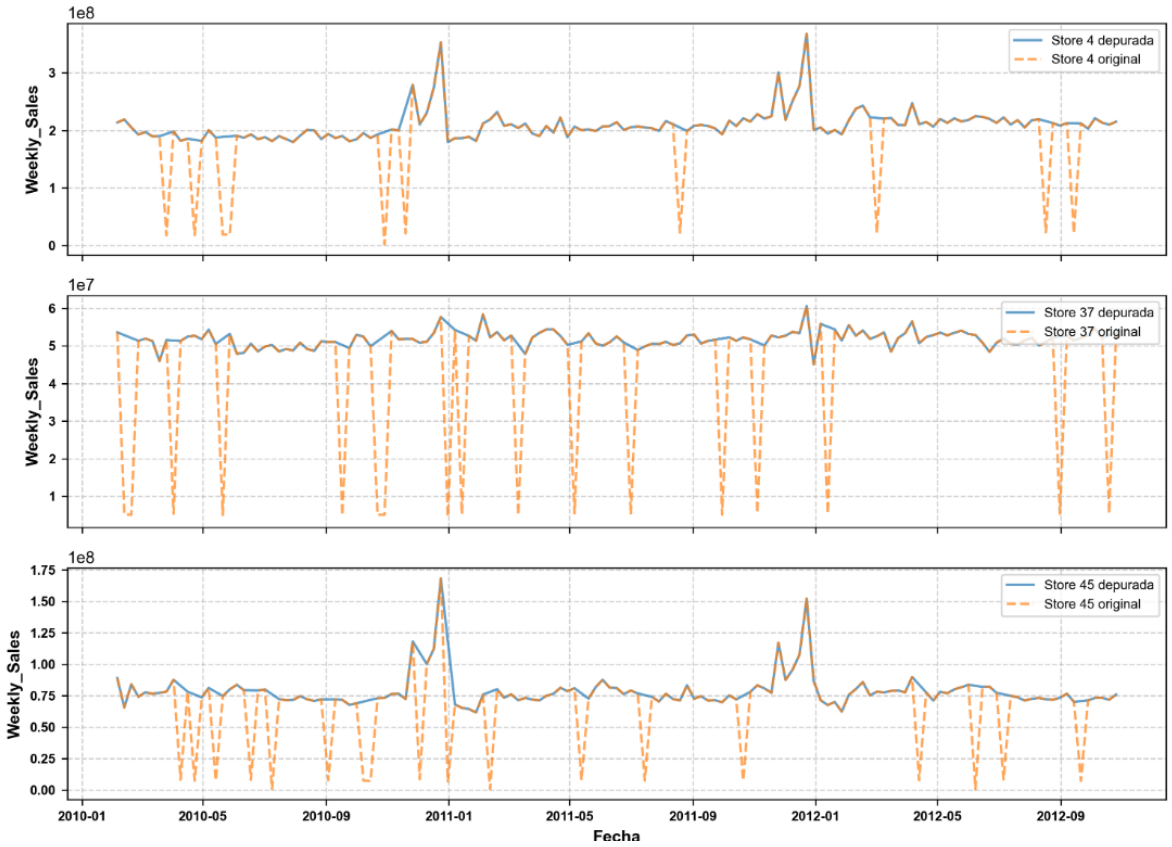


Figura 5. Serie de ventas semanales depuradas vs originales (sin depurar).

Durante el proceso de limpieza, se eliminó aproximadamente el 10.4% del total de los datos (ver Figura 6), una proporción considerable que sugiere un impacto significativo de los valores atípicos. Aunque este fenómeno merece un análisis detallado, no será abordado en este estudio. En su lugar, se utilizará el 89.6% restante de los datos para realizar el análisis exploratorio y el entrenamiento de los modelos.

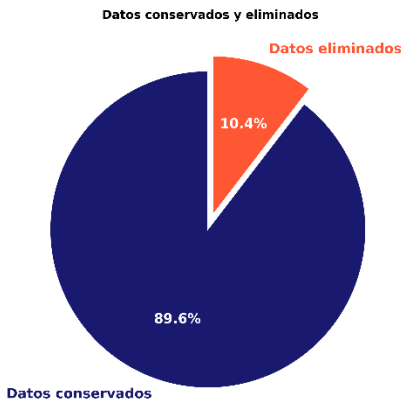


Figura 6. Porcentaje de datos conservados y eliminados.

Análisis exploratorio

Con los datos depurados de valores atípicos, procedimos a realizar el análisis exploratorio. En primer lugar, nos enfocamos en verificar la integridad de los datos, lo que incluyó la revisión de datos faltantes, registros duplicados y tipos de datos. Como resultado, se identificaron cuatro columnas de tipo **float64**, cuatro columnas de tipo **int64** y una columna de tipo **object**, tal como se detalla en la Tabla 2.

Tabla 2. Columnas y tipos de datos del dataset.

#	Column	Non-Null Count	Dtype
0	Date	5763 non-null	object
1	Store	5763 non-null	int64
2	Weekly_Sales	5763 non-null	int64
3	Holiday_Flag	5763 non-null	int64
4	Temperature	5763 non-null	float64
5	Fuel_Price	5763 non-null	float64
6	CPI	5763 non-null	float64
7	Unemployment	5763 non-null	float64
8	Holiday_Events	5763 non-null	int64

No se identificaron valores faltantes ni registros duplicados. Sin embargo, la columna Date no estaba en el formato adecuado. Por ello, se convirtió al formato datetime64[ns] utilizando la biblioteca Pandas.

Con la certeza de que los datos estaban depurados, se llevaron a cabo diversos análisis, entre ellos: la distribución de las ventas por fechas, el comportamiento de las ventas totales por ciudad y su variabilidad, la identificación de las tiendas con mayores ingresos semanales en cada ciudad y la evaluación del impacto de factores como el precio del combustible, la temperatura y la tasa de desempleo durante los períodos de ventas récord.

Además, se examinaron métricas clave como las ventas promedio, el precio promedio del combustible, el índice de precios al consumidor y la tasa de desempleo, junto con su variabilidad. Por último, se analizaron la distribución de las variables, el sesgo, la presencia

de valores atípicos y las correlaciones entre los datos. A continuación, se presentan en detalle los resultados obtenidos en estos análisis.

Distribución de las ventas por fechas

La Figura 7 presenta tres gráficos que muestran la cantidad de ventas realizadas a nivel mensual, trimestral y anual. En el análisis mensual, los meses con los porcentajes más altos de ventas fueron julio (9.98%), abril (9.70%), septiembre (9.28%), agosto (9.20%), junio (9.06%) y marzo (9.01%). Por otro lado, los meses con porcentajes intermedios, que oscilaban entre el 8% y el 8.95%, incluyeron octubre (8.95%), mayo (8.31%) y febrero (8.28%). Finalmente, los meses con los porcentajes más bajos de ventas fueron diciembre (7.08%), noviembre (5.59%) y enero (5.57%).

En el análisis trimestral, se observó que los trimestres con los mayores porcentajes de ventas fueron el segundo y el tercero, coincidiendo con las estaciones de primavera y verano.

Por último, a nivel anual, el año con mayor volumen de ventas fue 2011.

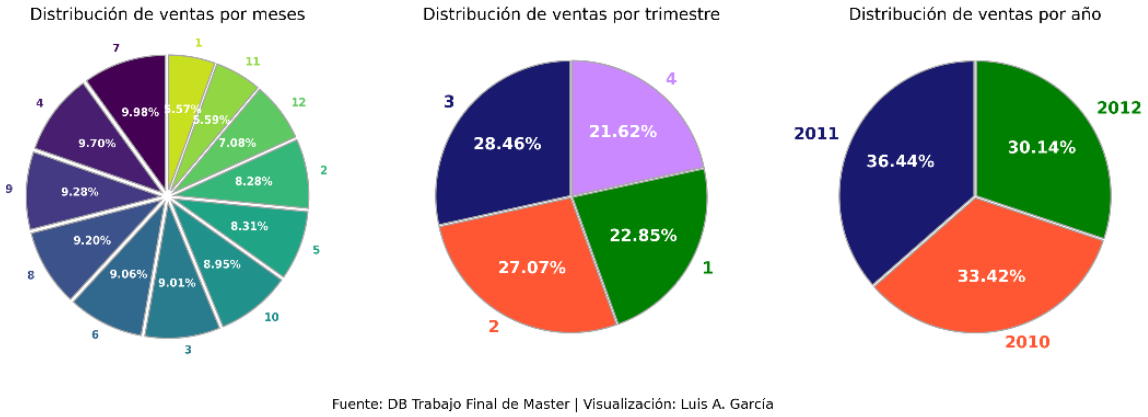


Figura 7. Distribución de ventas en meses, trimestres y años.

Análisis de ventas totales por ciudad y su coeficiente de variación (2010-2012)

La Figura 8 se compone de tres elementos principales. En primer lugar, se presenta un diagrama circular de barras que refleja las ventas totales por ciudad, utilizando colores para indicar la variabilidad de dichas ventas. En segundo lugar, se incluye un gráfico de barras horizontales que muestra los coeficientes de variación de las ventas para cada ciudad. Por último, se encuentra un gráfico tipo Lollipop que representa el promedio de ventas por

ciudad.

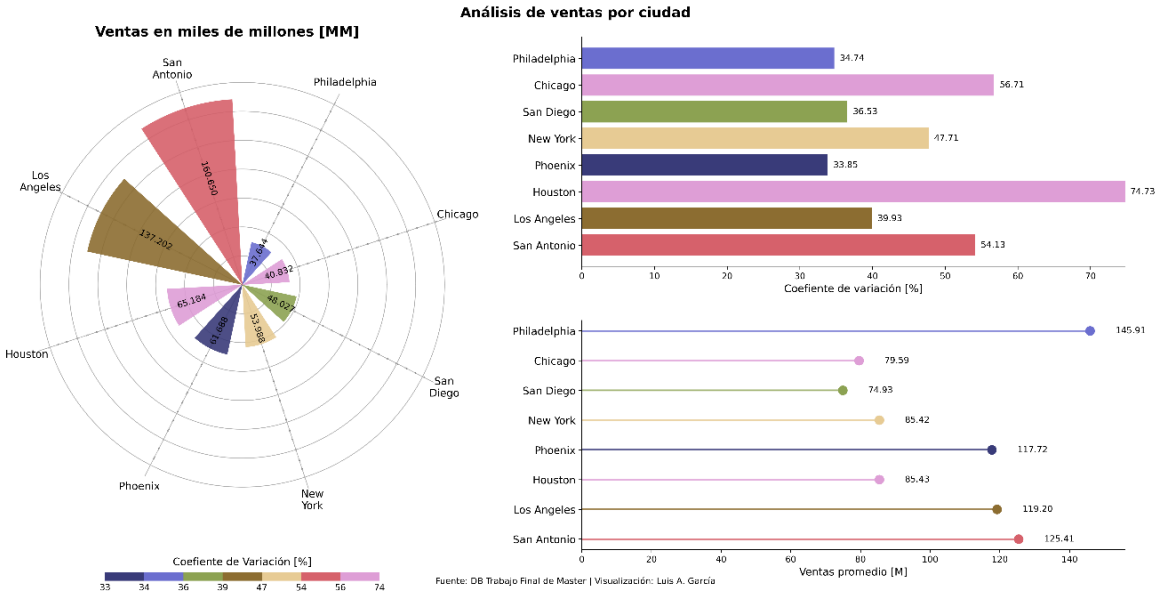


Figura 8. Análisis de ventas totales por ciudad.

En el diagrama circular de barras se observa que, entre los años 2010 y 2012, las ciudades con mayores ventas fueron San Antonio y Los Ángeles, con cifras de 160.65 MM y 137.20 MM, respectivamente. Sin embargo, San Antonio presenta un coeficiente de variación significativamente más alto (54.13%) en comparación con Los Ángeles (39.93%), lo que indica que las ventas en San Antonio están más dispersas en relación con su promedio. Por el contrario, en Los Ángeles, las ventas muestran una menor dispersión respecto a su promedio, lo que sugiere una mayor estabilidad en comparación con la alta variabilidad observada en San Antonio.

En el gráfico de barras horizontales, se evidencia que todas las ciudades presentan un coeficiente de variación superior al 33%, lo que indica que, en general, las ventas son poco estables. No obstante, Houston destaca como la ciudad con la mayor variabilidad (CV = 74.73%), mientras que Phoenix registra la menor dispersión (CV = 33.85%).

Finalmente, el gráfico Lollipop revela que, aunque Philadelphia registró las ventas totales más bajas (37.64 MM), presentó el promedio de ventas más alto (145.91 M). Esto sugiere que, a pesar de su bajo volumen total, mantuvo ventas consistentemente cercanas a este valor promedio en todo momento, lo que se confirma con su bajo coeficiente de variación (CV = 34.74%). Este comportamiento contrasta con ciudades como San Antonio, que, a pesar de tener las ventas totales más altas (160.65 MM), ocupa el segundo lugar en ventas promedio (125.41 M). Esto se explica por su alto coeficiente de variación (54.13%), lo que indica que sus grandes volúmenes de ventas totales se deben a altas variaciones, en lugar de una consistencia en las ventas. Así, el análisis sugiere que las ciudades con mayores ventas totales pueden atribuir este resultado tanto a la cantidad de tiendas como a sus niveles de ventas.

Para comprender a fondo la dinámica comercial de estas ciudades, fue necesario analizar el desempeño de las ventas individuales que impulsaron estos resultados.

La Figura 9 muestra que la cantidad de tiendas varía significativamente entre las ciudades, lo que genera desigualdades notables en el desempeño comercial de cada localidad. Se observó que las ciudades con un mayor número de tiendas, como San Antonio (10 tiendas) y Los Ángeles (9 tiendas), presentaron coeficientes de variación muy distintos (54.13% y 39.93%, respectivamente), lo que sugiere una mayor diversidad en los rangos de ventas.

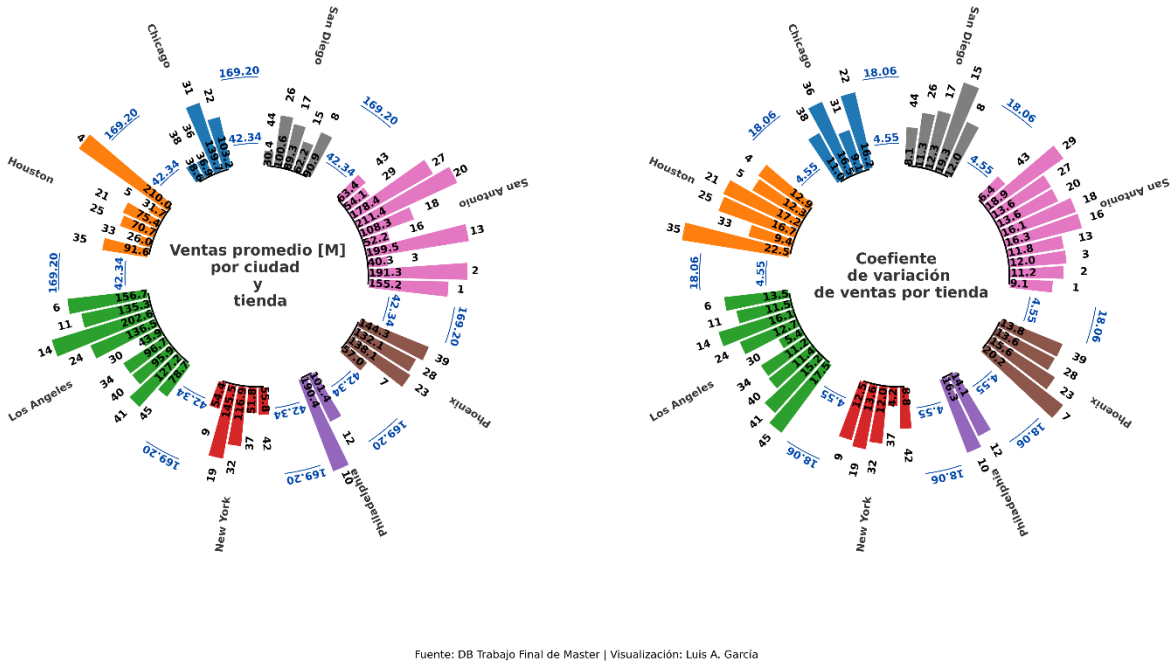


Figura 9. Análisis de ventas promedio y su variabilidad por tienda y ciudad

Al examinar la ciudad de Philadelphia, que se destacó por registrar la mayor venta promedio (145.91 M), se observó que contaba únicamente con dos tiendas: una alcanzó 190.4 M de ventas promedio y la otra registró 101.4 M. Asimismo, los coeficientes de variación de estas dos tiendas (16.3 % y 14.1 %) resultaron muy similares, lo que confirmó el bajo coeficiente de variación de la ciudad (34.74 %).

En el caso de San Antonio, que es la ciudad con más tiendas (10), cinco de sus tiendas superaron el promedio de ventas (125M), mientras que las restantes se situaron muy por debajo de este valor. Al analizar sus coeficientes de variación, se evidenció que las ventas no eran estables en todas las tiendas, lo que se correspondió con el elevado coeficiente de variación total de la ciudad (54.13 %).

Por su parte, Los Ángeles, que ocupó la tercera posición en cuanto a ventas promedio (119.20 M), contó con nueve tiendas, de las cuales cinco superaron el valor promedio de ventas. Los coeficientes de variación revelaron la existencia de dos grupos de tiendas con valores similares: el primero osciló entre 11.2 % y 13.5 %, y el segundo entre 15.2 % y 17.5 %. No obstante, se detectó una tienda con un coeficiente de variación particularmente bajo (5.4 %). Esta dispersión explicó el porqué, a pesar de tener numerosas tiendas con distintos niveles de ventas, la ciudad presentó un coeficiente de variación relativamente bajo (39.93 %).

Finalmente, el caso de Houston, con seis tiendas, permitió comprender su elevado coeficiente de variación. Solo dos de sus establecimientos superaron el promedio de ventas de la ciudad (85.43 M), destacándose uno que alcanzó un valor muy alto (210 M), mientras que los demás se situaron por debajo de 95 M. Al revisar los coeficientes de variación individuales, se encontró que la tienda con mayor venta promedio, mostró uno de los coeficientes de variación más bajos, en contraste con el resto de las tiendas, cuyos valores de variación fueron más elevados. De este modo, se confirmó que la cantidad de tiendas no tenía una relación directa con la variabilidad de las ventas y se atribuyó el alto coeficiente de variación de la ciudad a la significativa dispersión en las ventas de sus tiendas.

Análisis de ventas récord semanales por tienda en cada ciudad (2010-2012)

Una vez comprendidas las relaciones entre las ventas promedio y los coeficientes de variación de las tiendas, se procedió a examinar las ventas máximas alcanzadas en cada establecimiento por ciudad. Este análisis, complementado con la información sobre coeficientes de variación (CV) y ventas totales por ciudad presentada en la Figura 8, proporcionó una visión más profunda de las dinámicas comerciales.

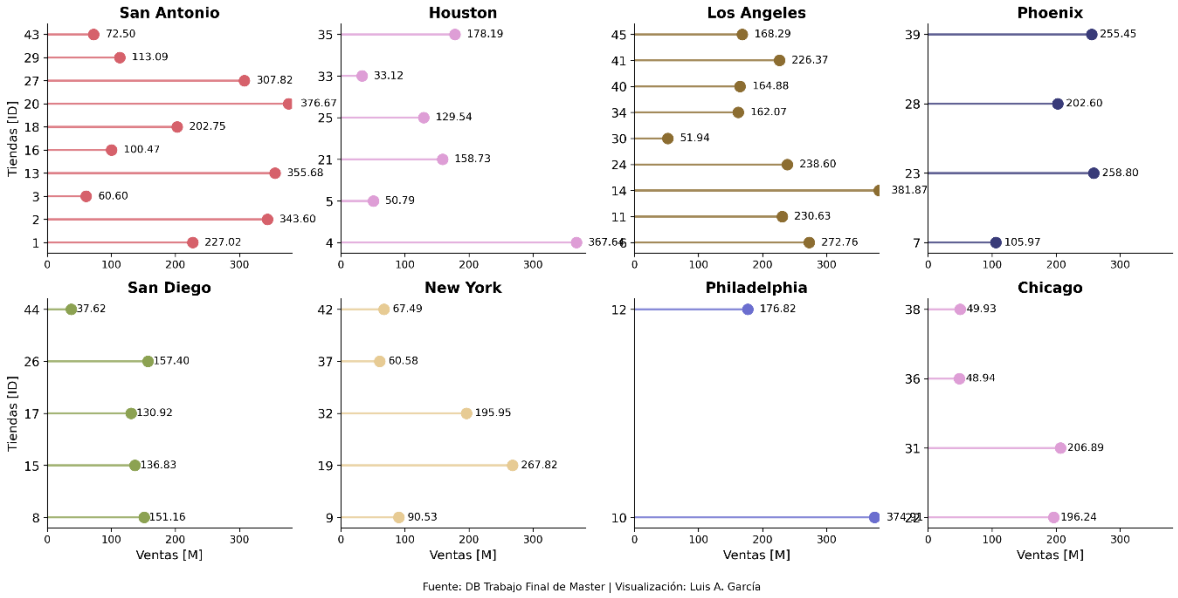


Figura 10. Ventas récord semanales registradas en cada tienda y ciudad.

En San Antonio, que lideró en ventas totales (160.65 MM) pero con un alto coeficiente de variación (54.13%), se observó que las tiendas individuales presentaron una amplia dispersión en sus ventas máximas. Solo las tiendas 2, 13, 20 y 27, registraron ventas récord entre 307 y 376 millones, lo que contribuyó significativamente a la alta variabilidad de la ciudad. Este comportamiento sugiere que las ventas totales de San Antonio están impulsadas por el desempeño excepcional de ciertas tiendas, en lugar de una consistencia generalizada.

Por otro lado, en Los Ángeles, que también tuvo un volumen de ventas totales alto (137.20 MM) pero con un coeficiente de variación más moderado (39.93%), se observó una disparidad significativa entre las tiendas. Por ejemplo, la tienda 14 lideró con 381.87 millones, mientras que la tienda 30 reportó solo 51.94 millones. A pesar de esta diferencia, la mayoría de las tiendas en Los Ángeles mantuvieron un desempeño más estable en

comparación con San Antonio, lo que explica su menor coeficiente de variación. Esto sugiere que, aunque existe variabilidad, esta no es tan extrema como en San Antonio.

En las ciudades con menos tiendas, como Philadelphia (2 tiendas) y Phoenix (4 tiendas), los coeficientes de variación son más moderados (34.74% y 33.85%, respectivamente). En Philadelphia, las dos tiendas registraron ventas máximas de 176.82 y 374.9 millones, lo que refleja una marcada diferencia a pesar del reducido número de establecimientos. Sin embargo, esta ciudad se destaca por su consistencia, como se observa en la Figura 8, donde Philadelphia tuvo el promedio de ventas más alto (145.91 M) a pesar de sus bajas ventas totales (37.64 MM). Esto indica que, aunque hay disparidad entre las tiendas, su desempeño tiende a ser más estable en comparación con ciudades como San Antonio.

En el caso de Phoenix, con solo 4 tiendas, las ventas máximas oscilaron entre 105.97 y 300 millones, lo que refleja una menor dispersión en comparación con ciudades como San Antonio y Los Ángeles. Este comportamiento se alinea con su bajo coeficiente de variación (33.85%), que sugiere un desempeño más uniforme entre sus tiendas.

Por su parte, Chicago, que también tiene 4 tiendas, presenta un coeficiente de variación significativamente más alto que Phoenix (56.71%), lo que indica una mayor dispersión en el desempeño de sus tiendas. En la Figura 10, se observa que las ventas máximas en Chicago oscilaron entre 48.94 y 206.89 millones, lo que refleja una marcada diferencia entre las tiendas. Esta alta variabilidad sugiere que, a pesar de tener un número reducido de tiendas, Chicago enfrenta desafíos en la consistencia de su desempeño comercial, similar a lo observado en ciudades con más tiendas, como San Antonio.

Estos resultados indican que el número de tiendas no está necesariamente relacionado con el nivel de disparidad observado: una mayor cantidad de tiendas no implica automáticamente una mayor variabilidad en los rangos de ventas máximas. Sin embargo, las diferencias significativas entre los rangos de ventas récord sí incrementan el coeficiente de variación, lo que sugiere que la desigualdad en el desempeño individual de las tiendas es un factor clave en la variabilidad general.

Además, al considerar los datos de la Figura 8, se refuerza la idea de que ciudades con ventas totales altas, como San Antonio y Los Ángeles, pueden atribuir sus resultados a momentos de ventas excepcionales en ciertas tiendas, en lugar de una consistencia generalizada. A continuación, se analiza cómo influyeron las variables locales en estas ventas récord.

Análisis de factores influyentes en las ventas récord

La Figura 11 presenta un contraste entre el precio del combustible y los registros de ventas récord semanales por tienda y ciudad, con el propósito de determinar la existencia de una correlación entre ambas variables. El análisis evidenció la ausencia de una relación significativa entre estos factores, lo que sugiere que el precio del combustible no influyó de manera determinante en las ventas récord semanales durante el período estudiado. Esta falta de correlación podría atribuirse a la homogeneidad en los precios del combustible en la mayoría de las ubicaciones analizadas, con algunas excepciones que no modificaron sustancialmente la tendencia general. En consecuencia, los hallazgos indican que el precio del combustible no constituyó un elemento clave en la variación de las ventas récord semanales.

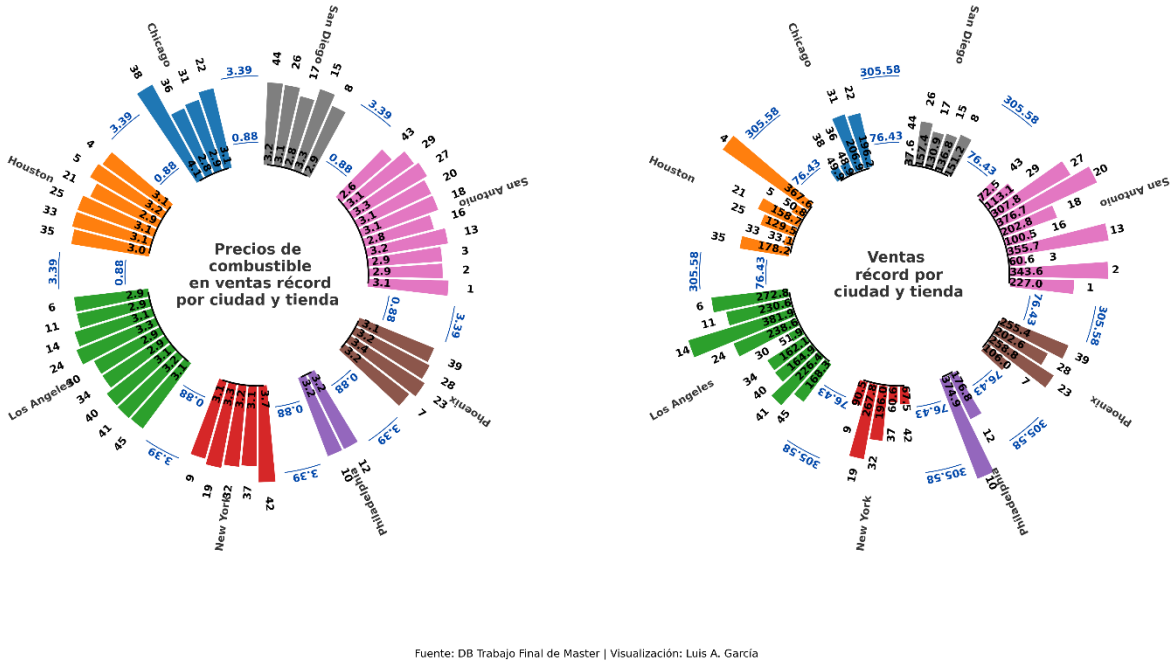


Figura 11. Relación visual entre el precio del combustible y las ventas récord en cada tienda y ciudad.

La Figura 12 presenta un análisis visual de la relación entre las ventas récord y la temperatura en diversas ciudades y tiendas, medida en grados Fahrenheit. Los datos muestran una amplia gama de temperaturas, desde 1.5°F hasta 449°F, lo que sugiere una variabilidad significativa en las condiciones climáticas que podrían influir en los patrones de ventas. Esta variabilidad podría indicar una relación más definida entre las ventas y la temperatura, sugiriendo un patrón estacional en el comportamiento del consumidor.

No obstante, es importante considerar que, aunque la temperatura parece tener un impacto notable en las ventas, otros factores externos también podrían estar influyendo en estos resultados. Elementos como festividades, eventos locales o características geográficas específicas de cada ciudad podrían condicionar y modificar la relación observada entre temperatura y ventas.

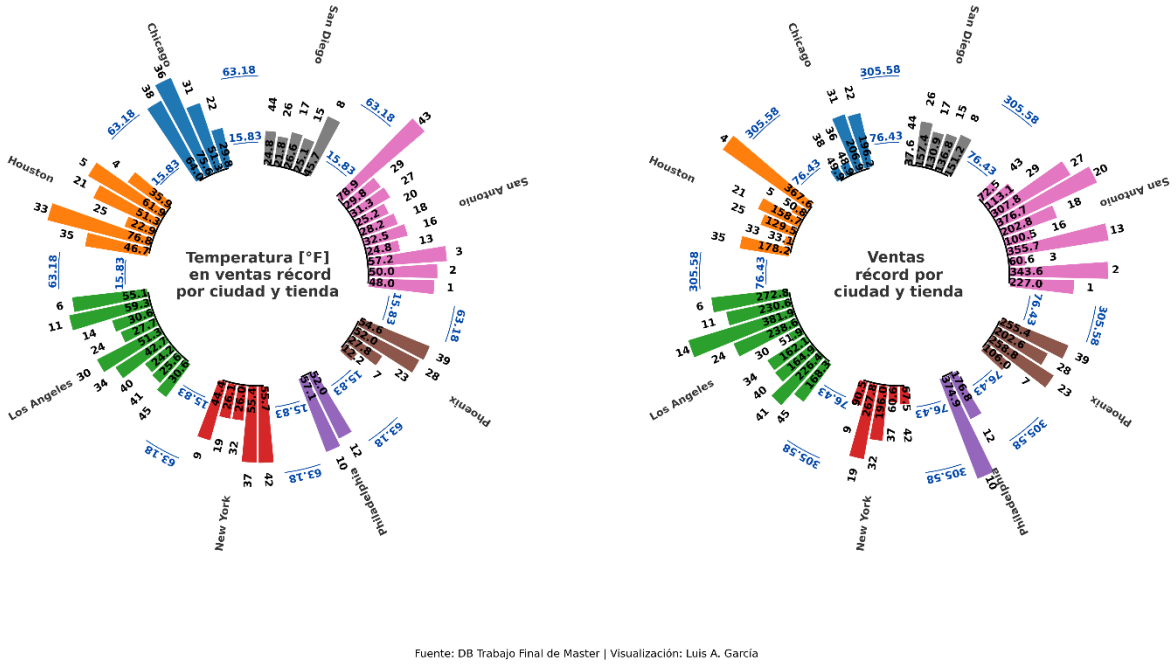


Figura 12. Relación visual entre la temperatura y las ventas récord en cada tienda y ciudad.

En relación con el desempleo, los datos presentados en la Figura 13 revelan efectos divergentes según la localidad analizada:

- **Philadelphia y Chicago**: Se observa que las ventas récord tienden a coincidir con periodos de bajas tasas de desempleo, lo que podría sugerir una relación positiva entre la estabilidad laboral y el consumo.
- **Nueva York**: Contrariamente, en esta ciudad se registraron picos de ventas durante períodos de desempleo relativamente alto, lo que indica una dinámica distinta y posiblemente influenciada por factores específicos del contexto local.
- **Otras ciudades**: En el resto de las localidades analizadas, no se identificó una tendencia clara o consistente entre los niveles de desempleo y las ventas, lo que da la idea de que esta relación no es uniforme.

Esta disparidad en los resultados sugiere que, aunque el desempleo podría tener cierto grado de influencia en las ventas, no existe una correlación directa ni generalizable a nivel global. Más que establecer una relación causal, los hallazgos subrayan la complejidad multifactorial que caracteriza el comportamiento de las ventas. Variables contextuales, sociodemográficas y económicas interactúan de manera no lineal, lo que dificulta la identificación de patrones universales. Por lo tanto, es fundamental considerar estas interacciones al interpretar los datos y diseñar estrategias comerciales.

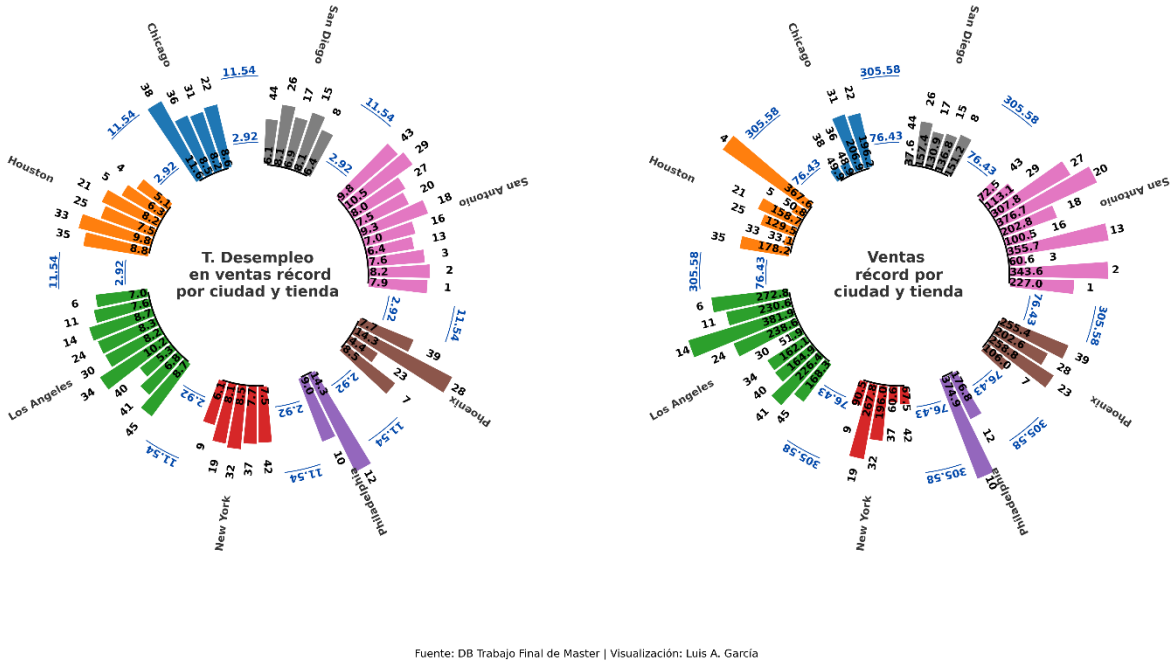
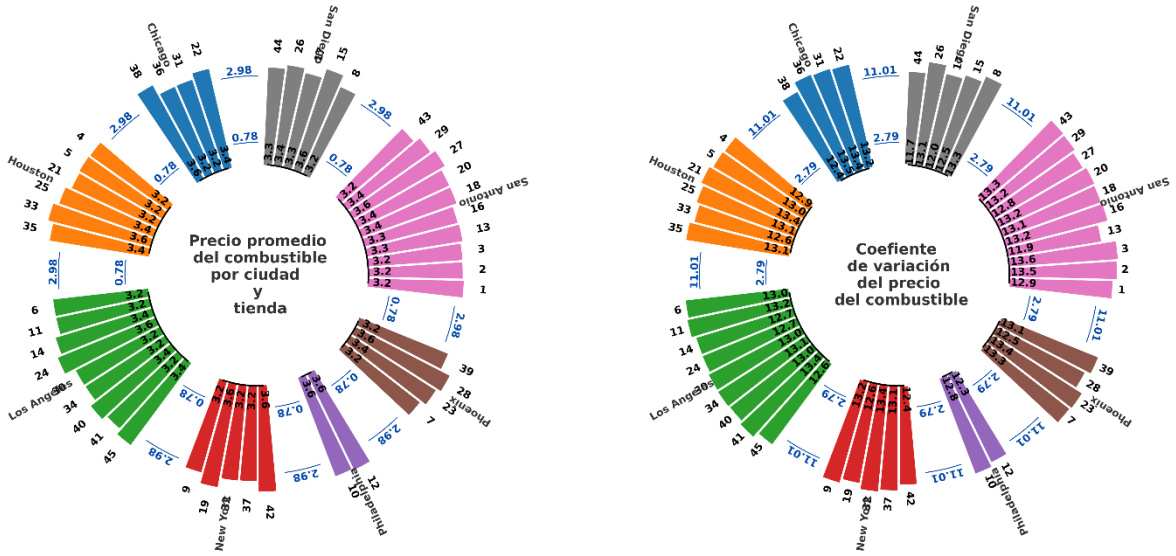


Figura 13. Relación visual entre la tasa de desempleo y las ventas récord en cada tienda y ciudad.

Análisis de promedios y variaciones

Tras analizar las influencias que las variables pudieron haber tenido sobre las ventas récord, y al no observarse patrones claros de correlación, se procedió a examinar los promedios del precio del combustible, el Índice de Precios al Consumidor (CPI) y la tasa de desempleo, tanto por tienda como por ciudad. El objetivo de este nuevo análisis fue obtener una apreciación general del comportamiento promedio de estas variables en cada ciudad y tienda.

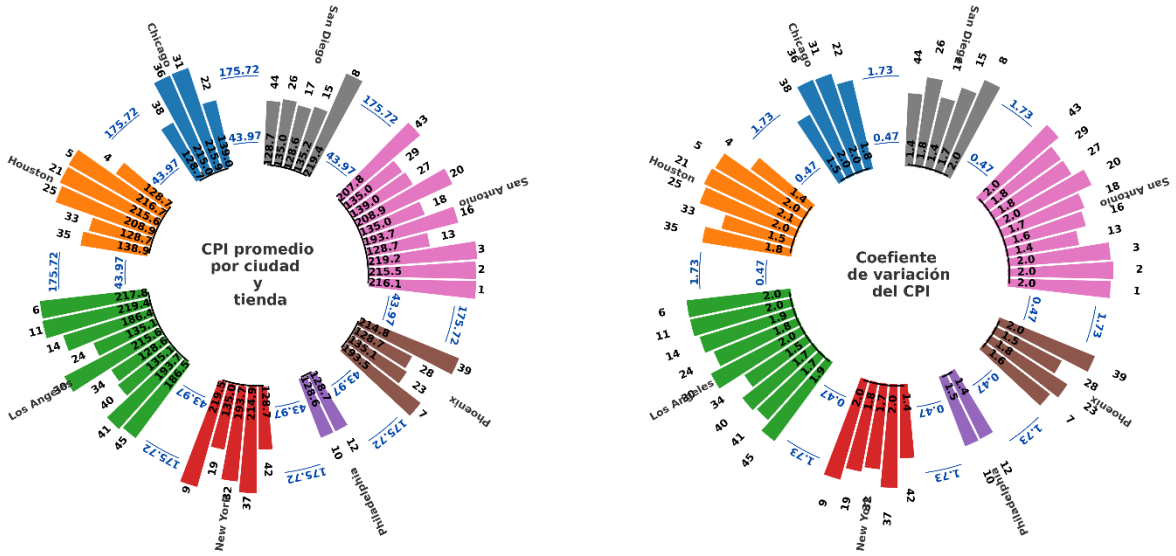
Los promedios de los precios del combustible para cada tienda (Figura 14) muestran una notable uniformidad, con leves variaciones. Esto se refleja también en los coeficientes de variación, que mantienen valores similares para cada tienda en cada ciudad.



Fuente: DB Trabajo Final de Master | Visualización: Luis A. García

Figura 14. Variabilidad del costo promedio del combustible por ciudad y tienda.

En cuanto al índice de precios al consumidor (Figura 15), se evidenció que en todas las ciudades las tiendas presentaron variaciones comprendidas entre el 1.4% y 2%. La única excepción fue Philadelphia, donde el CPI promedio no superó un coeficiente de variación del 1.5%.



Fuente: DB Trabajo Final de Master | Visualización: Luis A. García

Figura 15. Variabilidad del índice de precios al consumidor.

En relación con la tasa de desempleo (Figura 16), se observó que las ciudades de San Diego, San Antonio, Nueva York, Los Ángeles y Houston exhibieron una estabilidad relativa, con valores que oscilaron entre 5.9 y 9.9. En contraste, Chicago, Philadelphia y Phoenix registraron picos atípicos, alcanzando 13.1 en las tiendas 38, 12 y 28, respectivamente. Asimismo, se evidenció que la mayor variabilidad se presentó en Houston, donde la tienda 4 mostró un coeficiente de variación del 23.5%, seguida por San Diego (14.1%) y San Antonio (12.7%).

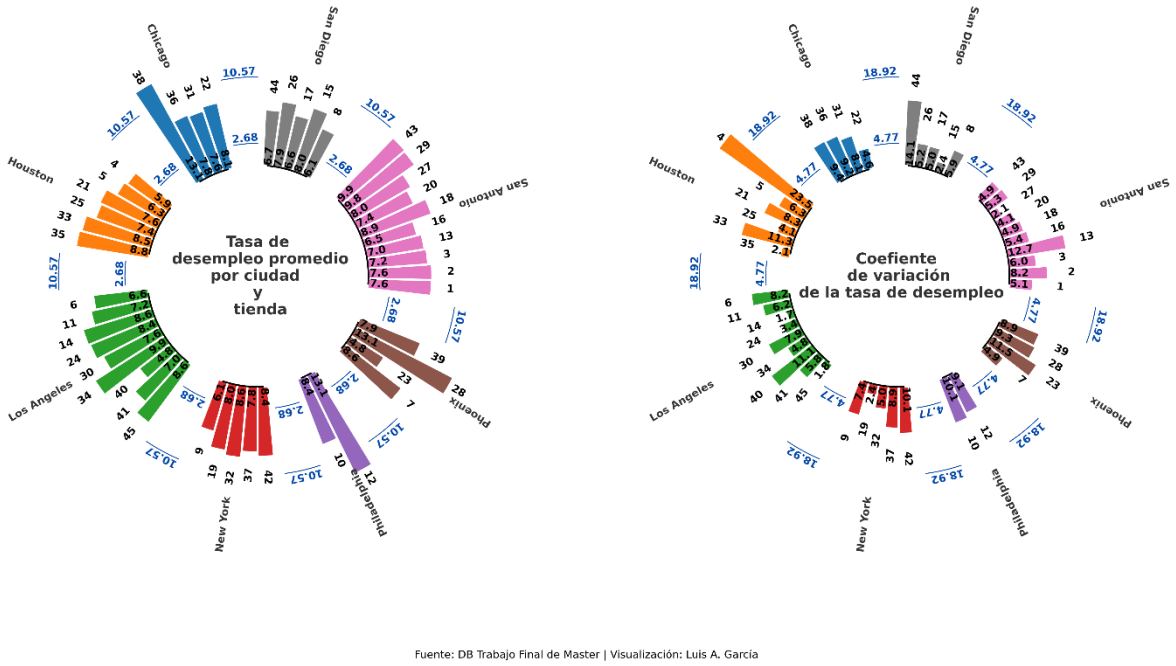


Figura 16. Variabilidad de la tasa de la tasa de desempleo.

Estos resultados subrayan que, aunque ciertas variables como el combustible mantienen patrones homogéneos, otras como ventas y desempleo exhiben dinámicas influenciadas por factores locales. Veamos ahora a nivel global las distribuciones estadísticas de algunas de estas variables.

Distribución de las variables

Al examinar el comportamiento de las variables (Figura 17), se identificaron patrones diferenciados. En el caso de la variable objetivo (Weekly_Sales), esta presentó un sesgo pronunciado hacia la derecha, señalando la presencia de valores atípicos y una alta dispersión. Esto sugiere que aplicar una transformación logarítmica podría mejorar la estabilidad de la varianza y mitigar el sesgo en análisis posteriores.

La variable Holiday_Flag evidencio un desequilibrio notable: la mayoría de las semanas registradas no eran festivas, y solo una fracción mínima correspondía a períodos festivos. Este desbalance podría afectar la capacidad predictiva de los modelos si no se ajusta adecuadamente. Por su parte, los datos de temperatura mostraron un mejor ajuste a una distribución normal, con una concentración marcada de valores entre 60 y 80 °F alrededor de la media.

En contraste, el precio del combustible (Fuel_Price) registró una distribución bimodal, con dos rangos de precios dominantes, aunque sin dispersiones extremas.

En variables como el Índice de Precios al Consumidor (CPI), la distribución fue multimodal, lo que sugiere la existencia de subgrupos diferenciados, posiblemente vinculados a factores geográficos o socioeconómicos locales. Finalmente, la tasa de desempleo presentó un ligero sesgo a la derecha, con valores frecuentes concentrados entre 6 y 8 por ciento. No obstante, su inclusión en modelos requerirá validar si cumple con los supuestos de normalidad en los residuos.

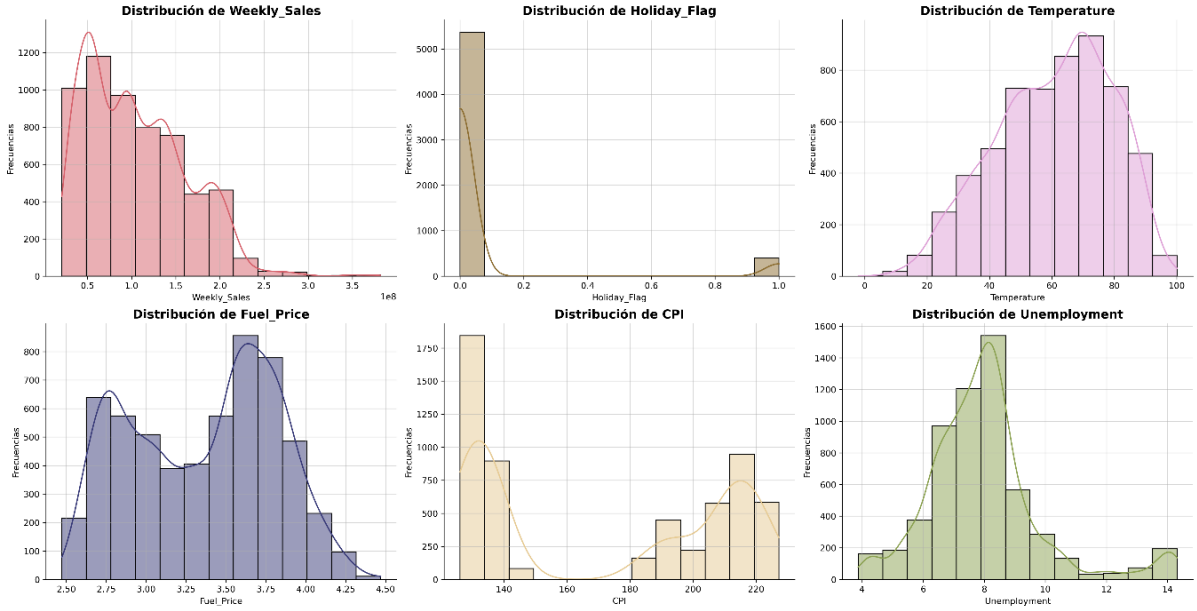


Figura 17.Distribución de variables.

Skewness (Sesgo) de las variables numéricas

La Tabla 3 confirma los patrones visuales previamente identificados en las distribuciones (Figura 17), destacando que la mayoría de las variables no siguieron un comportamiento normal, con sesgos hacia la izquierda o derecha.

Tabla 3. Valores de Skewness (sesgo).

Columna	Sesgo
Weekly_Sales	0.657795
Holiday_Flag	3.421459
Temperature	-0.330335
Fuel_Price	-0.101333
CPI	0.06789
Unemployment	1.179785
Holiday_Events	3.421459

Valores atípicos

Los diagramas de cajas y bigotes (Figura 18) revelaron que, pese a la eliminación del 10.4 % (Figura 6) de datos sospechosos, persistió una variabilidad significativa en las variables Weekly_Sales, Unemployment y Temperature. Ante este escenario, se optó por priorizar la integridad del dataset, evitando recortes adicionales y evaluando el impacto directo de esta variabilidad en el rendimiento de los modelos. Esta decisión permitirá contrastar si la heterogeneidad restante, lejos de ser ruido, contiene información relevante al momento de entrenar los modelos.

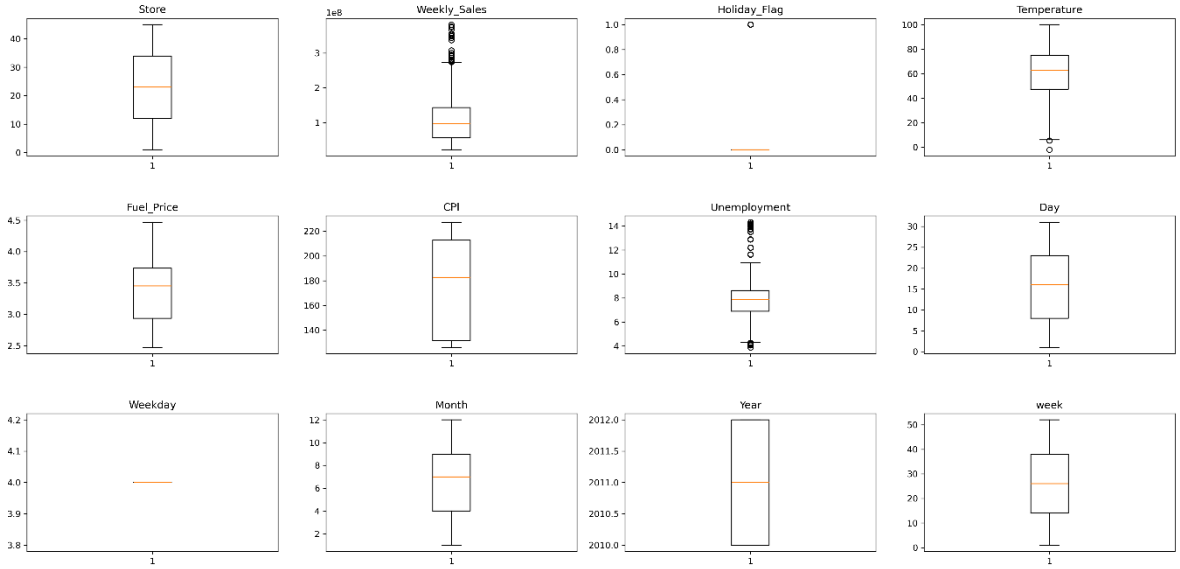


Figura 18. Dispersión de los datos.

Correlaciones

Como se evidenció en esta exploración de datos, la complejidad del dataset continuo incluso tras la limpieza de los datos. En la Figura 19 se profundizó en este análisis al mostrar las correlaciones entre las variables. En este caso, se destacó la variable Store, que registró la mayor correlación con la variable objetivo (Weekly_Sales). No obstante, se identificaron correlaciones significativas entre variables predictoras, un fenómeno que podría vulnerar el supuesto de independencia en modelos de regresión lineal.

No obstante, en línea con la decisión de preservar datos pese a la gran dispersión que se observó en algunas variables, se priorizará un **análisis de importancia de características**. Este enfoque permitirá discriminar entre correlaciones estadísticamente problemáticas y aquellas con valor predictivo real, equilibrando los requisitos técnicos del modelo con la relevancia operativa de los datos.

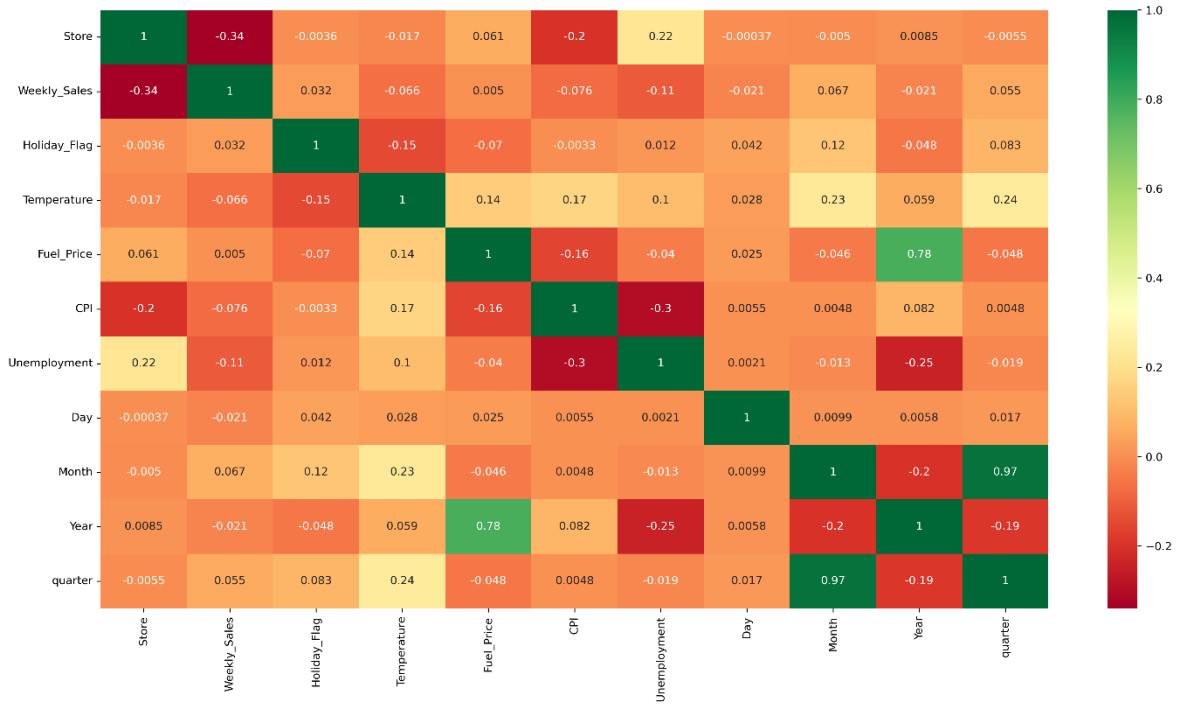


Figura 19. Correlación entre variables numéricas.

En base a lo anterior, este análisis exploratorio realizado permitió concluir que, si bien es viable implementar un modelo de regresión lineal múltiple debido a la existencia de correlaciones significativas entre variables, resulta fundamental verificar el cumplimiento de los supuestos inherentes a esta técnica. Esto se justifica particularmente ante la evidencia de 1) multicolinealidad entre predictores y 2) presencia de valores atípicos significativos, factores que podrían comprometer la validez del modelo. Para garantizar resultados robustos, se evaluarán aspectos como la normalidad, homocedasticidad e independencia de los residuos para entender si las predicciones del modelo sirven para inferencia estadística.

Preparación de los datos

La exploración de datos dio a entender que es posible entrenar modelos regresión lineal, no obstante, evidencio la presencia de valores atípicos en las variables predictoras, aun después de la depuración inicial. Dado que modelos como Random Forest son sensibles a valores atípicos, se procedió a realizar un escalado robusto.

RobustScaler utiliza estadísticas que son robustas a los valores atípicos. “El método emplea métricas estadísticas resistentes a la influencia de valores extremos. Esta técnica suprime la mediana de los datos y realiza el escalamiento en función del rango entre cuantiles (por defecto se utiliza el IQR: rango intercuartílico). El IQR corresponde a la diferencia entre el tercer cuartil (percentil 75) y el primer cuartil (percentil 25). La estandarización se aplica de forma individual a cada variable, calculando los parámetros necesarios exclusivamente a partir de los datos de entrenamiento. Estos valores (mediana e IQR) se almacenan para ser implementados posteriormente en nuevas observaciones a través del método transform” (18).

Como complemento al proceso, se incorporó la fecha como variable predictiva mediante la descomposición temporal. Esta transformación implicó la generación de nuevas características numéricas correspondientes al año, mes, día, día de la semana, semana del año y trimestre.

Sobre el conjunto de datos preprocesado (con todas las variables escaladas excepto la columna objetivo), se realizó una división estratificada: el 80% de los registros se destinó al entrenamiento del modelo y el 20% restante se reservó para la fase de evaluación predictiva.

Modelado de datos

Utilizando el 80% del conjunto de datos asignado para entrenamiento, se implementaron nueve algoritmos de aprendizaje supervisado en su configuración predeterminada:

- **Modelos lineales**: LinearRegression, Lasso, Ridge, ElasticNet.
- **Métodos basados en árboles**: RandomForestRegressor, DecisionTreeRegressor, GradientBoostingRegressor, XGBRegressor
- **Otros enfoques**: KNeighborsRegressor (vecinos cercanos) y SVR (máquinas de vectores de soporte vectorial)

La evaluación se realizó conservando los hiperparámetros predeterminados de cada modelo, lo que permitió obtener los siguientes resultados:

Tabla 4. Comparación de Modelos de Regresión: Desempeño según R^2, MSE y MAE.

Modelo	R^2	Error cuadrático medio (MSE).	Error absoluto medio (MAE)
LinearRegression	15.19 %	2831310798748848.5	43746750.178
Lasso	15.25 %	2829311221517238.5	43732088.348
Ridge	15.25 %	2829399417076720.5	43732811.409
ElasticNet	9.57 %	3018814964079387.5	45702475.715
RandomForestRegressor	97.0 %	100246849805178.56	5403749.772
DecisionTreeRegressor	95.55 %	148706645843901.0	6720447.785
GradientBoostingRegressor	89.47 %	351497109811468.44	13806501.478
XGBRegressor	97.64 %	78853724689358.83	5065163.02
KNeighborsRegressor	37.62 %	2082500040850201.5	35213431.204
SVR	-3.0 %	3438462285323254.0	47437409.815

Del análisis de la Tabla 4 se destacaron los siguientes modelos con mejor rendimiento, ordenados por su coeficiente R^2: XGBRegressor (97.64%), RandomForestRegressor (97.0%), DecisionTreeRegressor (95.55%), GradientBoostingRegressor (89.47%) y KNeighborsRegressor (37.62%). Para el estudio, se optó por excluir únicamente al modelo DecisionTreeRegressor, decisión basada en dos motivos principales: en primer lugar, se buscó evitar redundancias metodológicas, ya que otros modelos seleccionados (XGBRegressor, RandomForestRegressor y GradientBoostingRegressor) emplean enfoques basados en árboles de decisión; y en segundo lugar, aunque KNeighborsRegressor mostró un desempeño inicial bajo, se priorizó su mejora mediante

ajustes de hiperparámetros, aprovechando que su algoritmo tenia un enfoque diferente al de árboles de decisión.

Optimización de hiperparámetros y validación cruzada

Previo al proceso de optimización de hiperparámetros, se realizó un análisis de las configuraciones predeterminadas de cada modelo. Este paso permitió definir rangos de valores inferiores a los establecidos por defecto en los algoritmos. La elección de esta metodología respondió a un enfoque sistemático: al restringir los intervalos de los parámetros, se buscó explorar combinaciones más eficientes que las configuraciones base, evitando así el riesgo de sobreajuste y reduciendo simultáneamente la complejidad computacional del proceso.

La optimización de hiperparámetros se llevó a cabo utilizando Optuna, un framework especializado en la optimización automática de funciones objetivo. En este proceso, se eligió el coeficiente de determinación (R^2) como métrica a maximizar y se evaluó mediante validación cruzada para garantizar la robustez de los modelos. Optuna exploró iterativamente el espacio de búsqueda definido, proponiendo configuraciones de hiperparámetros y validando su eficacia a través de 50 ciclos de entrenamiento y evaluación. En cada iteración, el modelo se ajustaba con los hiperparámetros propuestos, se calculaba el R^2 en los subconjuntos de validación y este valor servía como guía para optimizar las siguientes combinaciones, generando así los siguientes resultados.

Tabla 5. Mejor puntaje para los parámetros optimizados.

Modelo	R^2
XGBRegressor	0.9736
RandomForestRegressor	0.9653
GradientBoostingRegressor	0.9773
KNeighborsRegressor	0.4167

Los resultados de la Tabla 5, evidenciaron que la optimización con hiperparámetros reducidos, mejoró la eficiencia de los modelos. Por un lado, XGBRegressor conservó su alto desempeño (R^2 = 0.976) utilizando solo 89 árboles y una profundidad máxima de 5, frente a los 100 árboles predeterminados. En concordancia, RandomForestRegressor mostró una leve disminución en su puntuación (de R^2 = 0.97 a 0.96) al ajustarse a 23 árboles con profundidad máxima de 4, en contraste con los 100 árboles iniciales. Finalmente, GradientBoostingRegressor destacó al lograr un R^2 de 0.92 con 99 árboles y profundidad máxima de 12, superando claramente su resultado original de 0.89 con 100 árboles. Mientras que KNeighborsRegressor solo mejoró de 0.37 a 0.41.

Se sometieron estos modelos a competencia utilizando los hiperparámetros óptimos previamente identificados. La evaluación integral abarcó la eficiencia computacional (tiempo de entrenamiento), la agilidad operativa (tiempo de predicción) y la precisión estadística (coeficiente de determinación R^2) (ver Figura 20).

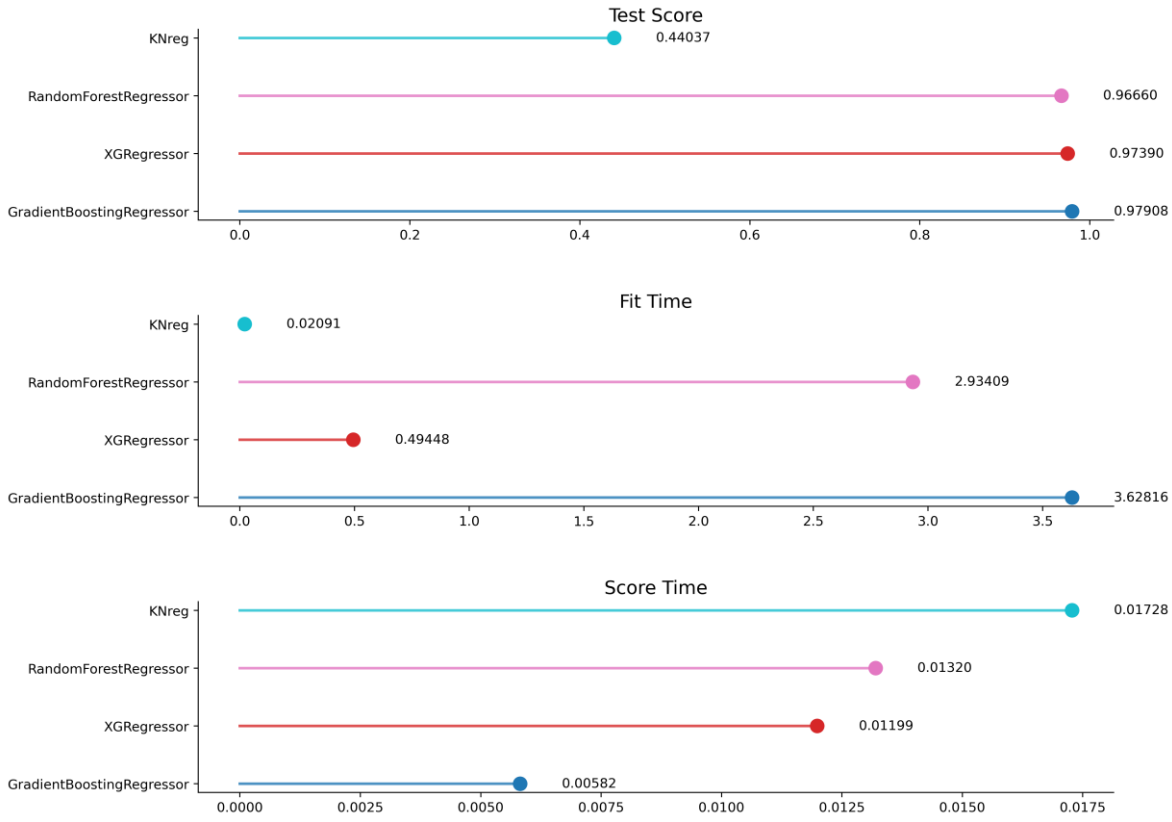


Figura 20. Comparación de modelos de regresión: Rendimiento (R^2), Tiempo de entrenamiento y Evaluación

Tras el análisis comparativo de rendimiento y eficiencia, los resultados obtenidos bajo las configuraciones óptimas revelaron diferencias significativas entre los modelos evaluados. GradientBoostingRegressor destacó al lograr el R^2 más alto (0.979), consolidándolo como el modelo con mejor ajuste a los datos. No obstante, este alto rendimiento tuvo un costo computacional: requirió el tiempo de entrenamiento más prolongado debido a su complejidad algorítmica (utiliza una mayor cantidad de árboles). Sin embargo, compensó esta desventaja al demostrar ser el más rápido en predicciones, lo que lo hace ideal para entornos que priorizan respuestas inmediatas tras el despliegue.

Por otro lado, XGBRegressor alcanzó un R^2 competitivo (0.973), ubicándose en segundo lugar en precisión, y mostró un equilibrio notable entre eficiencia y rendimiento. Fue el segundo más rápido tanto en entrenamiento como en predicción, superando a la mayoría de los modelos en velocidad operativa.

RandomForestRegressor, aunque obtuvo un R^2 sólido (0.96), quedó relegado al tercer puesto. A pesar de emplear una cantidad reducida de árboles en comparación con su configuración por defecto, resultó ser el tercer modelo más lento en entrenamiento y el segundo más lento en predicción.

Finalmente, KNeighborsRegressor confirmó su inadecuación para el problema analizado, registrando el R^2 más bajo del grupo. Su enfoque basado en proximidad, útil en contextos específicos, no logró capturar la complejidad de los datos, lo que subraya la importancia de seleccionar algoritmos alineados con la naturaleza del conjunto de datos.

A pesar de la ligera ventaja en R^2 del GradientBoostingRegressor, se seleccionó al XGBRegressor como modelo óptimo. Esta elección se basó en su balance entre precisión y eficiencia: no solo mantiene un rendimiento estadístico cercano al máximo, sino que también minimiza costes computacionales en entrenamiento (ideal para actualizaciones frecuentes) y predicción (clave para implementaciones en producción). Así, XGBRegressor fue seleccionado como la solución más robusta y sostenible para este problema.

Evaluación

Identificado el modelo más óptimo, se procedió a mejorarse. Estas mejoras consistieron en determinar cuáles eran las características que el modelo consideraba más importantes al momento de realizar las predicciones. Para esto, se utilizó SHapley Additive exPlanations (SHAP) donde se observó que las variables con mayor importancia fueron Store y CPI (ver Figura 21).

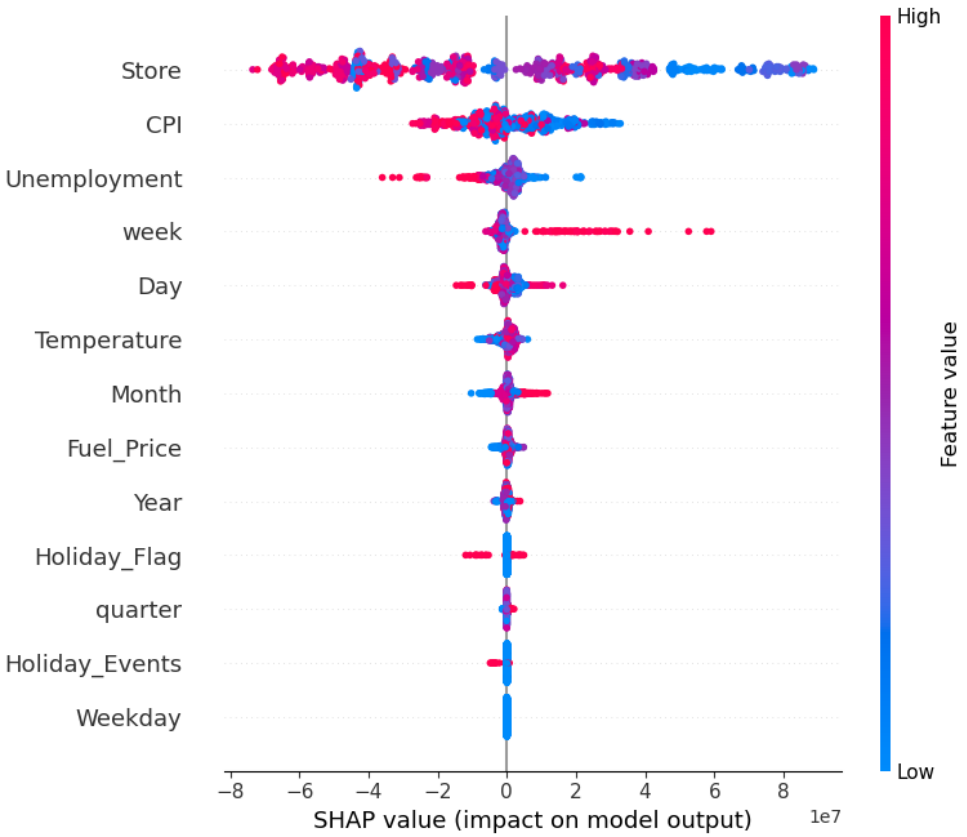


Figura 21. Importancia de Características mediante valores SHAP.

No obstante, se optó por implementar un segundo método de selección de características, la Eliminación Recursiva de Características (RFE), con el propósito de obtener una perspectiva más detallada. La aplicación de este método permitió identificar las cinco variables con mayor influencia en las decisiones del modelo, obteniendo un coeficiente de determinación (R^2) de 0.97 al considerar las siguientes características: Store, CPI, Unemployment, Holiday_Events y week (semanas del año).

Posteriormente, se evaluó la posibilidad de conservar un desempeño similar utilizando únicamente las tres características más influyentes. Al emplear Store, CPI y Unemployment, el coeficiente de determinación se redujo a 0.93, lo que sigue siendo un valor elevado y sugiere que el modelo puede realizar predicciones aceptables con estas tres variables.

No obstante, se optó por conservar las cinco características más influyentes, dado que variables como Holiday_Events (de naturaleza desbalanceada) y week probablemente aportan información relevante para la toma de decisiones del modelo. A partir de esta decisión, se llevó a cabo el entrenamiento del modelo utilizando distintos tamaños de muestra. Los resultados obtenidos indicaron que el modelo no presentaba signos de sobreajuste, ya que las curvas de aprendizaje eran relativamente cercanas entre sí y la curva de validación no se estabilizaba en un valor significativamente inferior a la de entrenamiento (ver Figura 22).

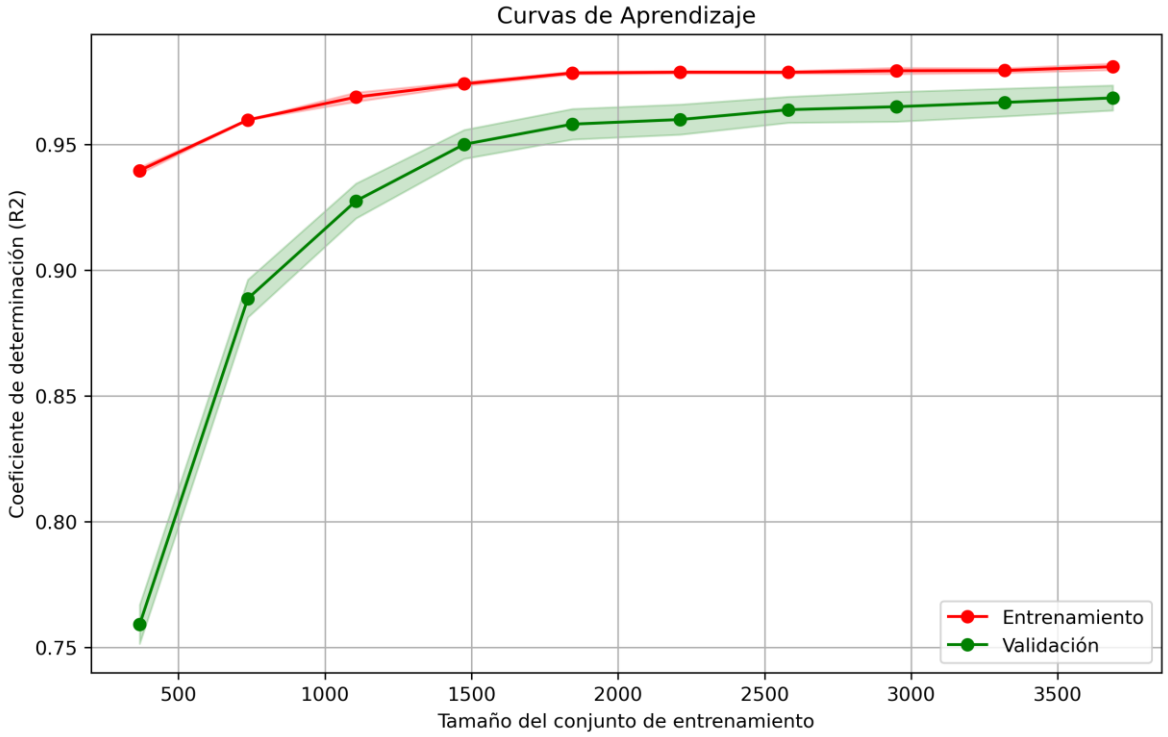


Figura 22. Curvas de aprendizaje del modelo: Relación entre el tamaño del conjunto de entrenamiento y el desempeño (R^2).

Esta prueba de entrenamiento y validación permitió determinar que el modelo alcanzó un coeficiente de determinación R^2 de 0.9645 y un R^2 ajustado de 0.9643. Al evaluar el modelo con un conjunto de datos no visto previamente, se obtuvo un error general en las predicciones del 10.24%

Cumplimiento de supuestos de la regresión lineal

En vista de que el modelo presentaba resultados satisfactorios, se procedió a verificar si este cumplía con los supuestos de la regresión lineal, con el fin de saber las limitaciones del modelo.

La prueba de residuos (Residual Plot y Breusch- pagan) expuso que la varianza de los residuos no era constante, indicando que el modelo no estaba capturando adecuadamente algunas relaciones, señalando la presencia de posibles valores atípicos que generaban heterocedasticidad en los residuos, siendo muy probable que fuera por usar variables desbalanceadas.

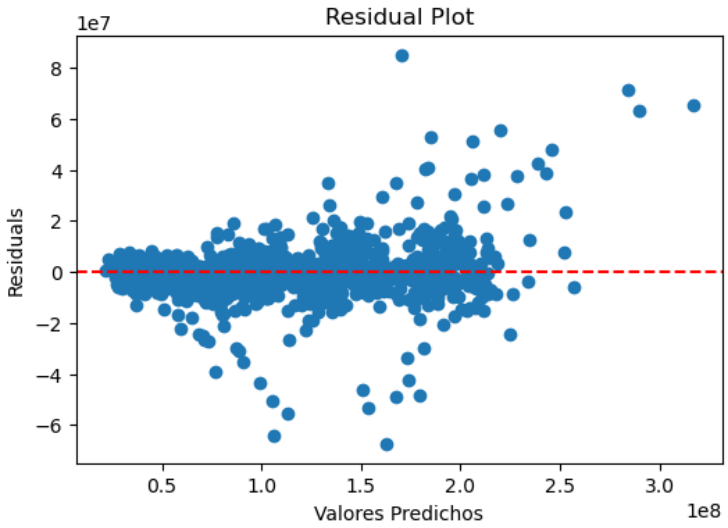


Figura 23. Gráfico de residuos: Evaluación de heterocedasticidad del modelo.

Las pruebas de normalidad de los residuos (Q-Q Plot y Shapiro -Wilk) indicaron que los residuos no siguieron una distribución normal. Adicionalmente, no se encontraron variables altamente correlacionadas con las demás.

Despliegue

Teniendo en cuenta las limitaciones del modelo con mejor desempeño (XGBRegressor), se optó por utilizar los tres mejores modelos (XGBRegressor, GradientBoostingRegressor y RandomForestRegressor) y desplegarlos en Power BI. De esta forma, se pueden realizar predicciones futuras y comparar sus resultados (Figura 24). Además, el informe en Power BI permite consultar de forma interactiva las estadísticas tanto regionales como de cada tienda.

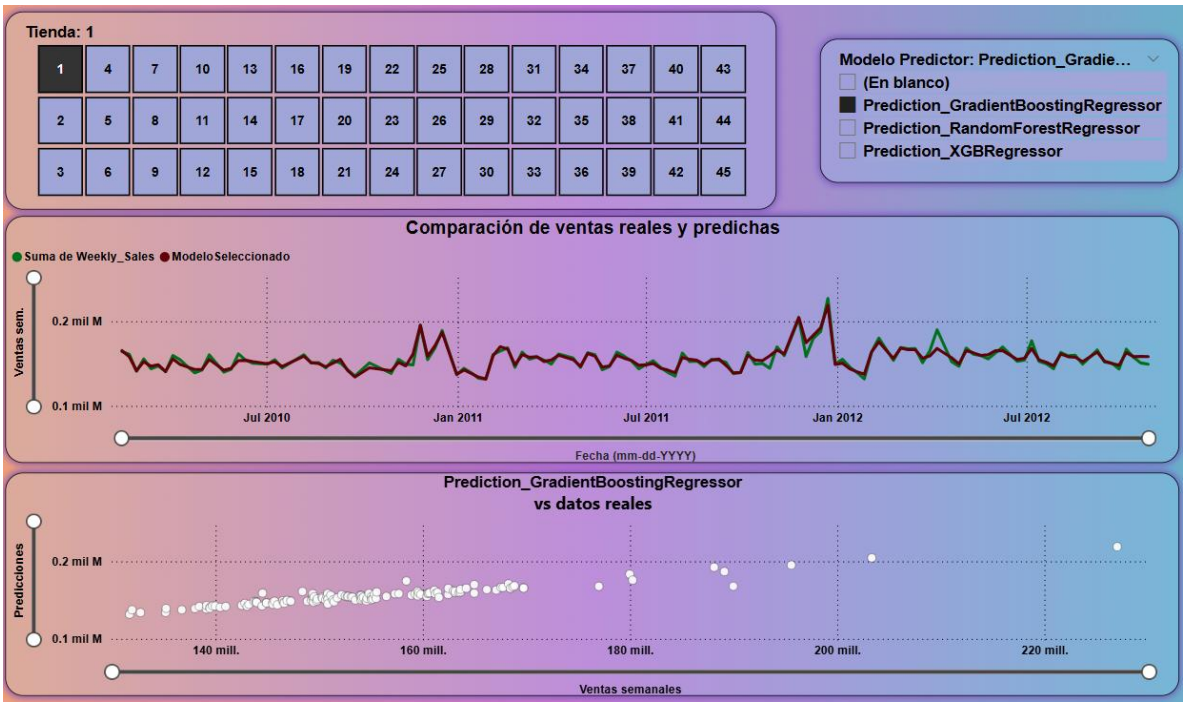


Figura 24. Predicciones en Power Bi.

Conclusiones

La implementación estructurada de las fases de la metodología CRISP-DM permitió obtener resultados satisfactorios y reproducibles, lo que dio lugar a las siguientes conclusiones.

- Luego del entrenamiento de diversos modelos de regresión, tales como regresión lineal, árboles de decisión, k-vecinos más cercanos y máquinas de soporte vectorial, los basados en árboles de decisión demostraron el mejor ajuste a los datos, alcanzando una explicación de más del 90% de la variabilidad en las ventas semanales.

- Se logró eficazmente la implementación de metodologías efectivas para la limpieza de datos, eliminación de valores atípicos, cálculo de estadísticas descriptivas y generación de representaciones gráficas correspondientes, garantizando un análisis integral y reproducible.

- Las variables más influyentes identificadas fueron Store, CPI, Unemployment, Holiday_Events y Week, logrando un R^2 de 0.96 con el modelo XGBRegressor. No obstante, las pruebas estadísticas aplicadas evidenciaron que este no capturó plenamente las relaciones complejas en los datos, lo que sugirió la persistencia de valores atípicos o variables desbalanceadas (como las semanas festivas), lo que podría explicar la heterocedasticidad observada en los residuos.

- Se logró el desarrollo de un informe interactivo en Power BI, donde se integran y depuran los datos, se emplean los modelos entrenados para generar predicciones y se facilita la visualización dinámica de las estadísticas regionales y locales, incluyendo un desglose por establecimiento.

- A pesar de que los modelos entrenados lograron explicar más del 90% de la variabilidad en las ventas es importante abordar la heterocedasticidad detectada en los residuos, lo cual implica estudiar en detalle los valores atípicos y las variables desbalanceadas (como las semanas festivas). Adicionalmente, se recomienda explorar transformaciones logarítmicas en los datos, reentrenar los modelos usando otros parámetros y comparar su desempeño con los resultados obtenidos en este estudio.

Bibliografía

1. Big Data International Campus [Internet]. 2024 [citado 31 de diciembre de 2024]. proceso de Análisis de Datos. Disponible en: https://www.campusbigdata.com/blog/procesos-de-analisis-de-datos/

2. Raizada S, Saini JR. Comparative Analysis of Supervised Machine Learning Techniques for Sales Forecasting. Int J Adv Comput Sci Appl [Internet]. 2021 [citado 25 de enero de 2025];12(11). Disponible en: http://thesai.org/Publications/ViewPaper?Volume=12&Issue=11&Code=IJACSA&SerialNo=12

3. Schmidt A, Kabir MWU, Hoque MT. Machine Learning Based Restaurant Sales Forecasting. Mach Learn Knowl Extr. 30 de enero de 2022;4(1):105-30.

4. Thai-Nichi International College, Thai-Nichi Institute of Technology, Bangkok, Thailand, Joseph FJJ. Time series forecast of Covid 19 Pandemic Using Auto Recurrent Linear Regression. J Eng Res [Internet]. 22 de mayo de 2022 [citado 25 de enero de 2025]; Disponible en: https://kuwaitjournals.org/jer/index.php/JER/article/view/15425

5. Nieves V, Radin C, Camps-Valls G. Predicting regional coastal sea level changes with machine learning. Sci Rep [Internet]. 7 de abril de 2021 [citado 25 de enero de 2025];11(1). Disponible en: https://www.nature.com/articles/s41598-021-87460-z

6. Sosa E, Linares D. RPubs - Modelo de regresión lineal múltiple [Internet]. 2023 [citado 31 de diciembre de 2024]. Disponible en: https://rpubs.com/eric_jsosa/1036059

7. Haya P. La metodología CRISP-DM en ciencia de datos - IIC [Internet]. Instituto de Ingeniería del Conocimiento. 2021 [citado 25 de enero de 2025]. Disponible en: https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/

8. Sotaquirá M. Codificando Bits. 2021 [citado 4 de enero de 2025]. ¿Cómo hacer el Análisis Exploratorio de Datos? - Guía paso a paso. Disponible en: https://codificandobits.com/blog/analisis-exploratorio-de-datos/

9. Bruce P. Estadística Práctica para Ciencia de Datos con R y Python. 2da ed. Barcelona: Marcombo, S.A; 2022. 1 p.

10. Espinosa JLS. Statistics and health at work Descriptive statistics (III): Measures in grouped data. 2021;4(6).

11. DELSOL S. Coeficiente de variación ¿Qué es? [Internet]. 2022 [citado 4 de enero de 2025]. Disponible en: https://www.sdelsol.com/glosario/coeficiente-de-variacion/?srsltid=AfmBOopQ5h1GMhYm6L1FSqFlGYUk00k907f8TK2d7ZV45yfjbT_C7wet/

12. Rodríguez Sánchez A, Salmerón Gómez R, García C. The coefficient of determination in the ridge regression. Commun Stat - Simul Comput. (1):12.

13. Sayak P. Tutorial de selección de características en Python Sklearn [Internet]. 2024 [citado 27 de enero de 2025]. Disponible en: https://www.datacamp.com/tutorial/feature-selection-python

14. Awan AA. Una introducción a los valores SHAP y a la interpretabilidad del machine learning [Internet]. 2024 [citado 27 de enero de 2025]. Disponible en: https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability

15. Klensin JC. Application Techniques for Checking and Transformation of Names [Internet]. Internet Engineering Task Force; 2004 feb [citado 3 de enero de 2025]. Report No.: RFC 3696. Disponible en: https://datatracker.ietf.org/doc/rfc3696

16. Clayton G. Limpieza de cadenas de texto en SQL Server manteniendo letras y números en SQL Server [Internet]. [citado 3 de enero de 2025]. Disponible en: https://es.claytabase.com/Academia/Diseño-de-base-de-datos-de-aprendizaje/Uso-de-las-funciones-de-SQL-Server/Cuerdas-Limpieza-de-texto-en-SQL-Mantener-Letras-y-números

17. Oracle C. INITCAP [Internet]. [citado 3 de enero de 2025]. Disponible en: https://docs.oracle.com/en/database/oracle/oracle-database/21/sqlrf/INITCAP.html

18. scikit learn. sklearn.preprocessing.RobustScaler — documentación de scikit-learn - 0.24.1 [Internet]. 2020 [citado 29 de enero de 2025]. Disponible en: https://qu4nt.github.io/sklearn-doc-es/modules/generated/sklearn.preprocessing.RobustScaler.html