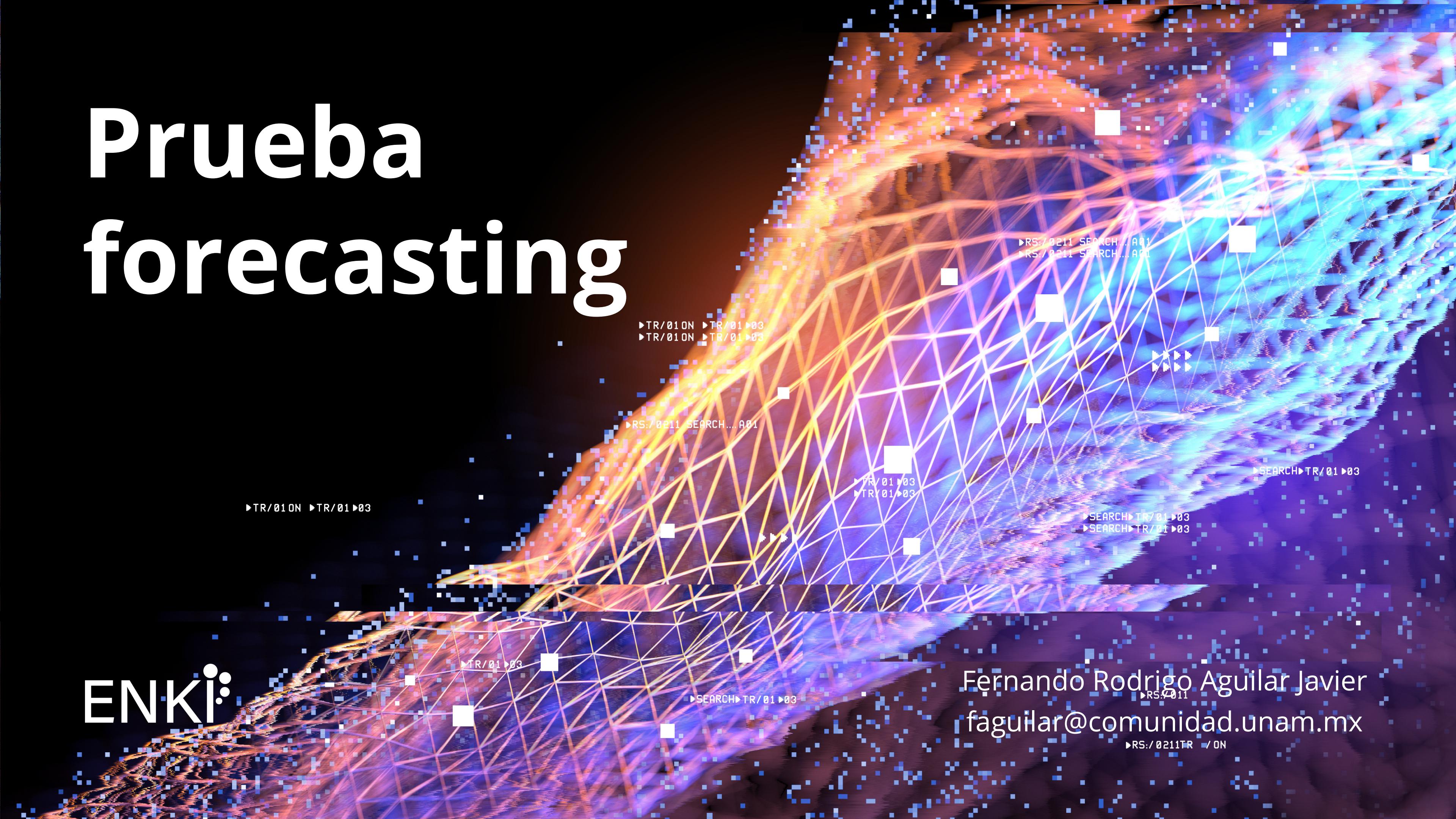


Prueba forecasting

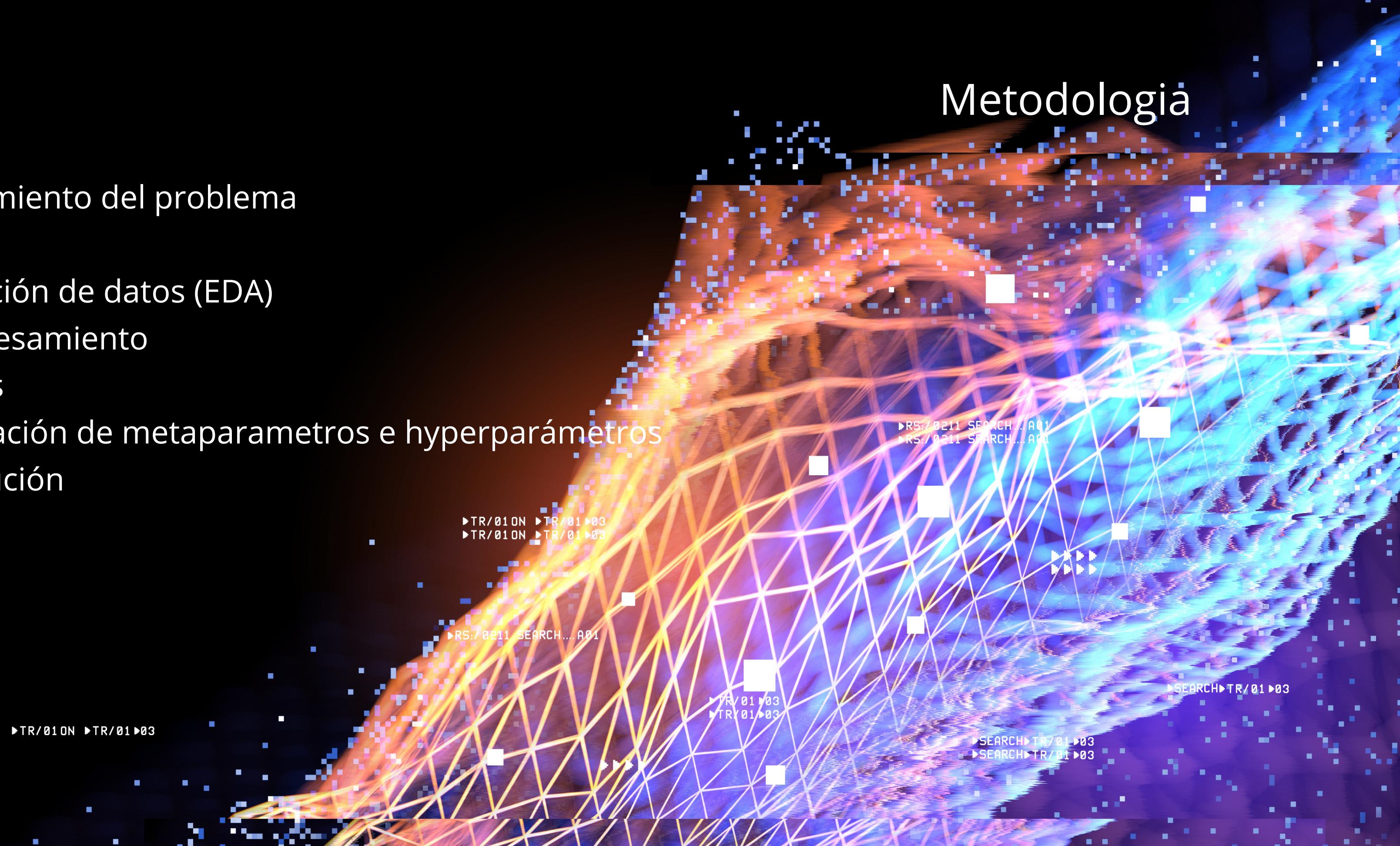
ENKI

Fernando Rodrigo Aguilar Javier
faguilar@comunidad.unam.mx



Metodología

1. Planteamiento del problema
2. Datos
3. Exploración de datos (EDA)
4. Preprocesamiento
5. Modelos
 - a. Afinación de metaparametros e hyperparámetros
 - b. Solución



1. Problemática y descripción de objetivo

Se requiere mejorar el pronóstico de ventas en retail, el objetivo es modelar el comportamiento de manera autorregresiva utilizando técnicas de ML.

Para ello se dispone de un conjunto de ventas tomadas durante un periodo de 3 meses, Agosto-Octubre de 2020-2021.



2. Datos

426377 ventas

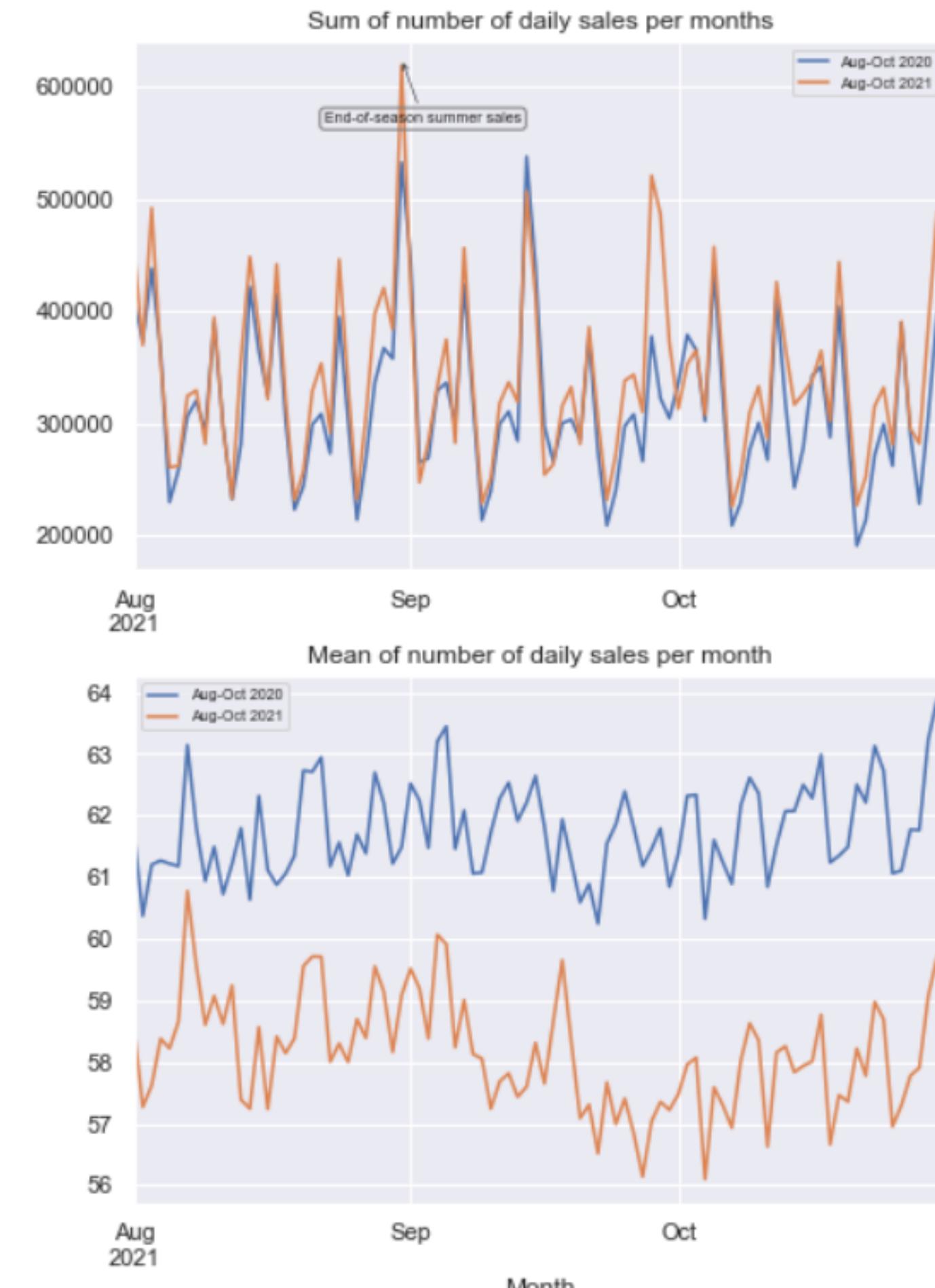
Los datos son números de ventas por producto y su respectivo precio que se hizo durante el día, en un periodo de 3 meses agosto-octubre 2020 y 2021, teniendo un total de 6 meses de registro de ventas con un total de 426377 ventas.

Insights

En la 1a fila, 1a columna, se observan dos picos durante la quincena, que es de esperarse dado que la mayoría de las personas gastan más dinero durante esas fechas.

En la 2a fila, 1a columna, se ve la mediana del precio de los productos diario durante esos 6 meses. Podemos decir que los precios son estables durante los 6 meses, a excepción de los días de quincena, donde se ve un ligero aumento.

En la 1a fila 2a columna son las ventas por día durante los 6 meses, en él se observa un máximo global que con base en las fechas, se cree podría ser provocado por las rebajas de la temporada de primavera-verano.

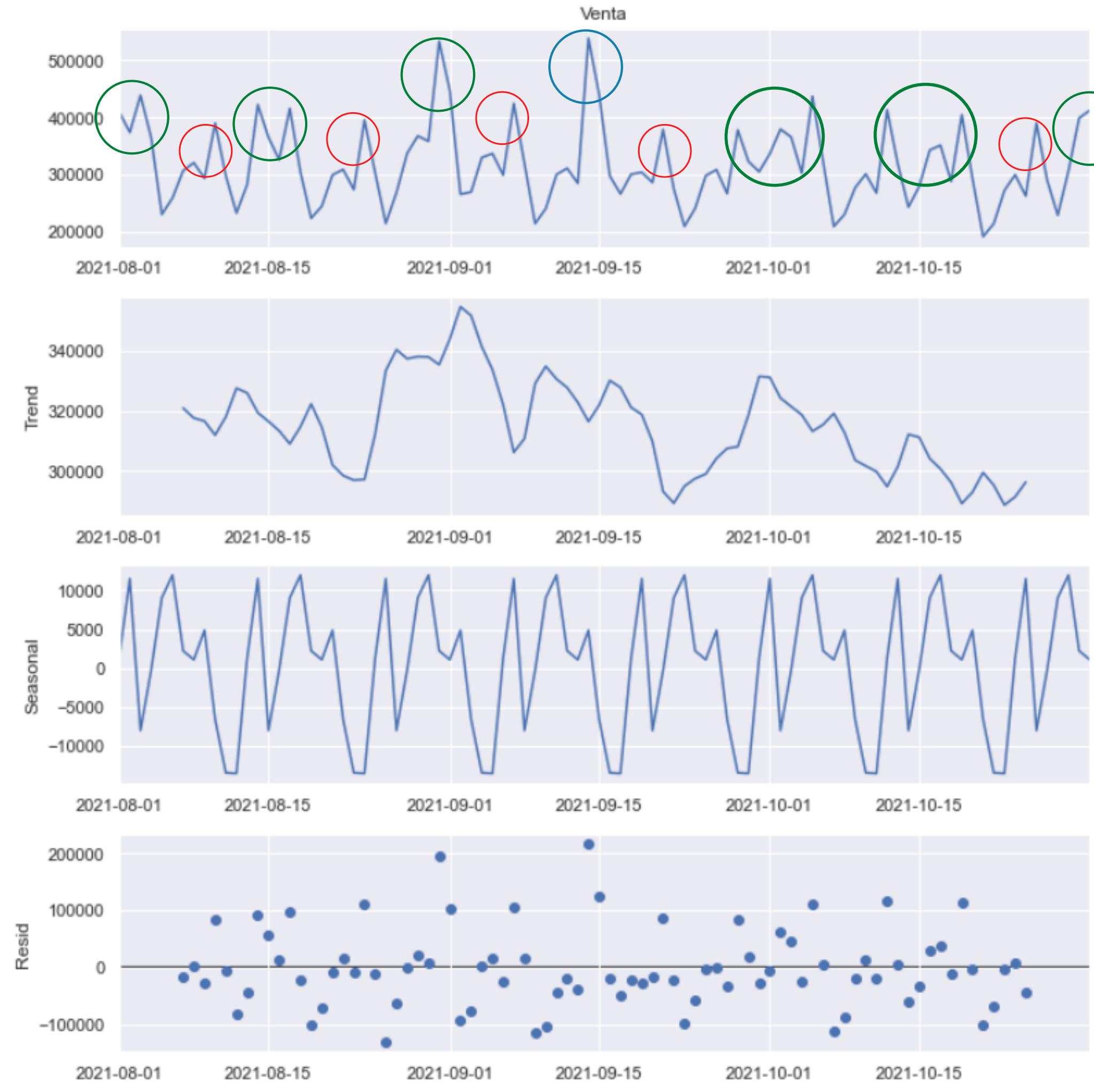


Insights

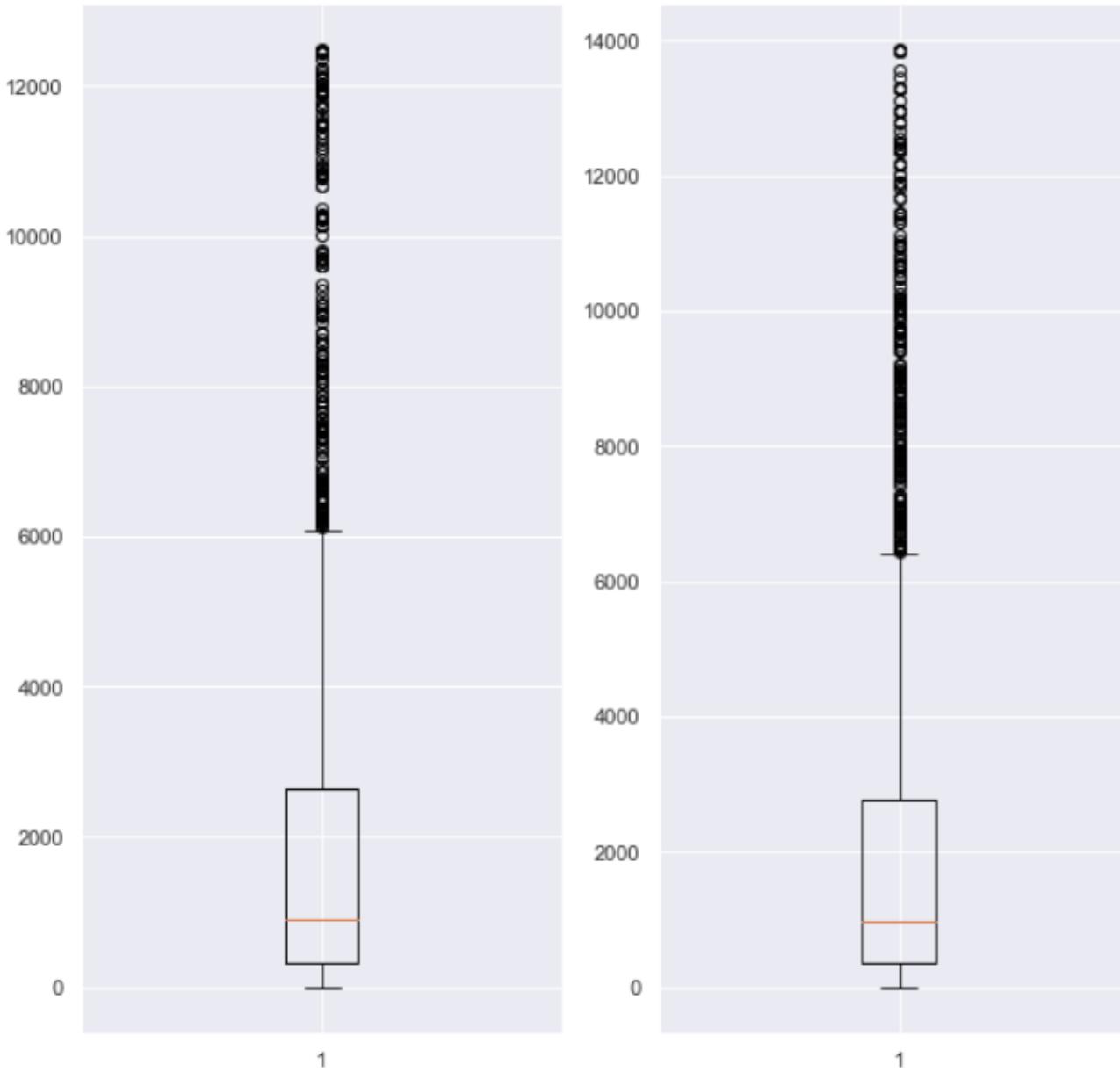
Cómo se puede observar en la gráfica, existe un **aumento en las ventas en cada fin de semana**, en especial en los días de pago de quincenas, principalmente durante agosto y septiembre. En octubre ocurre la misma tendencia, pero en menor medida, esto atribuido al cambio de precios por el inicio de la nueva temporada otoño-invierno.

- Fin de Semana
- Quincena

La descomposición puede ser un modelo aditivo o multiplicativo, para nuestro caso, dado que la tendencia es lineal y la variación estacional.



What is the most sold period? 31511504.28 sales in 2020



Insights

Hubo más ventas durante el periodo de 2020 que en 2021.

Los productos **456** y **480** fueron los más vendidos durante 2020-2021.

11527912 es el producto con 1 sola venta

480 es el producto con más ventas 1513788.1

4. Preprocesamiento

Data Enginnering

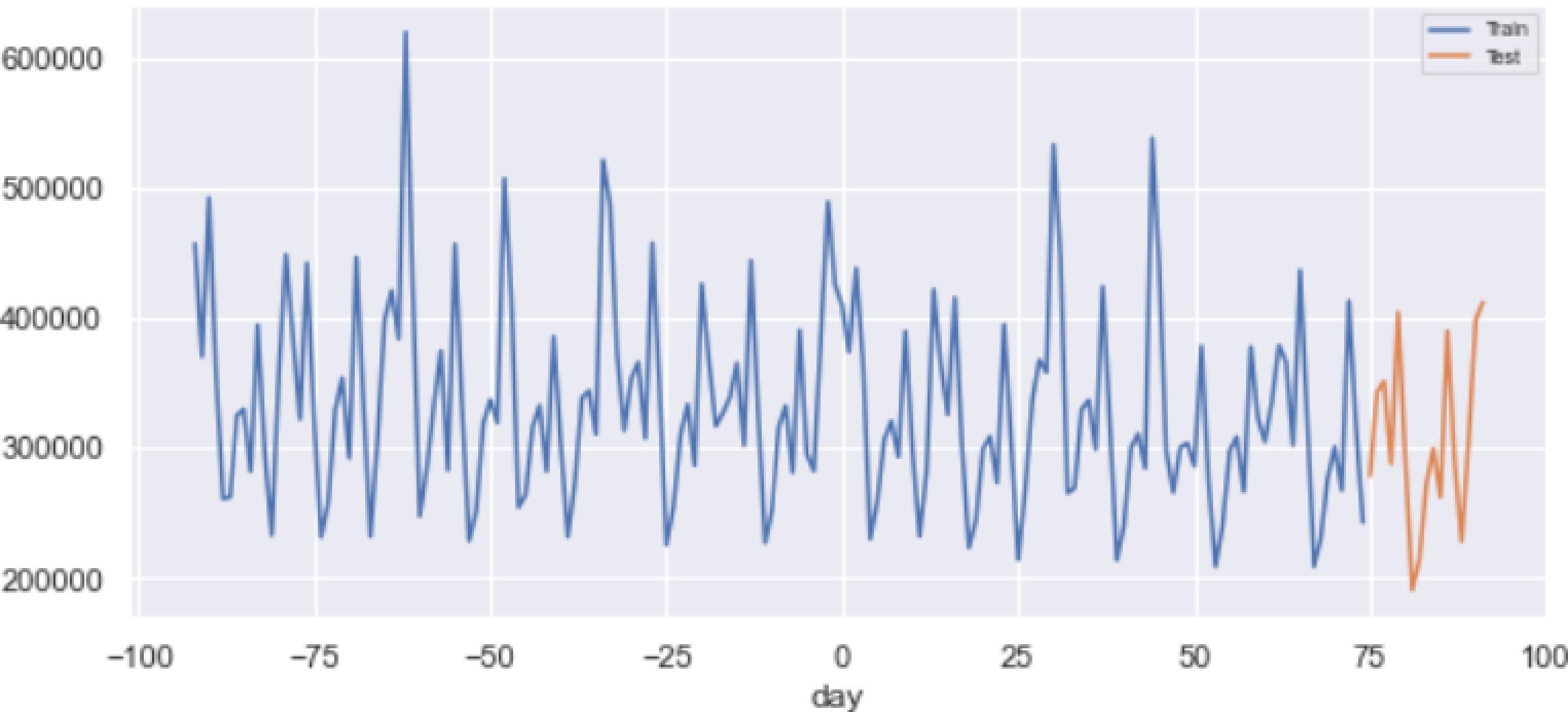
Los atributos Tienda, Formato, Subformato, producto y Unnamed 0 se eliminaron debido a que no aportan nada para entrenar al modelo

Se creo un nuevo atributo llamado **total** a partir de la suma de todos los productos vendidos durante el día, multiplicando el número de ventas por su precio.

5. Modelos

Se proponen los siguientes modelos:

1. Lineal
2. Polinomial
3. Regresión con Bosques Aleatorios
4. Regresión Ridge
5. Regresión Ridge bayesiana



Modelos, variables y métricas consideradas

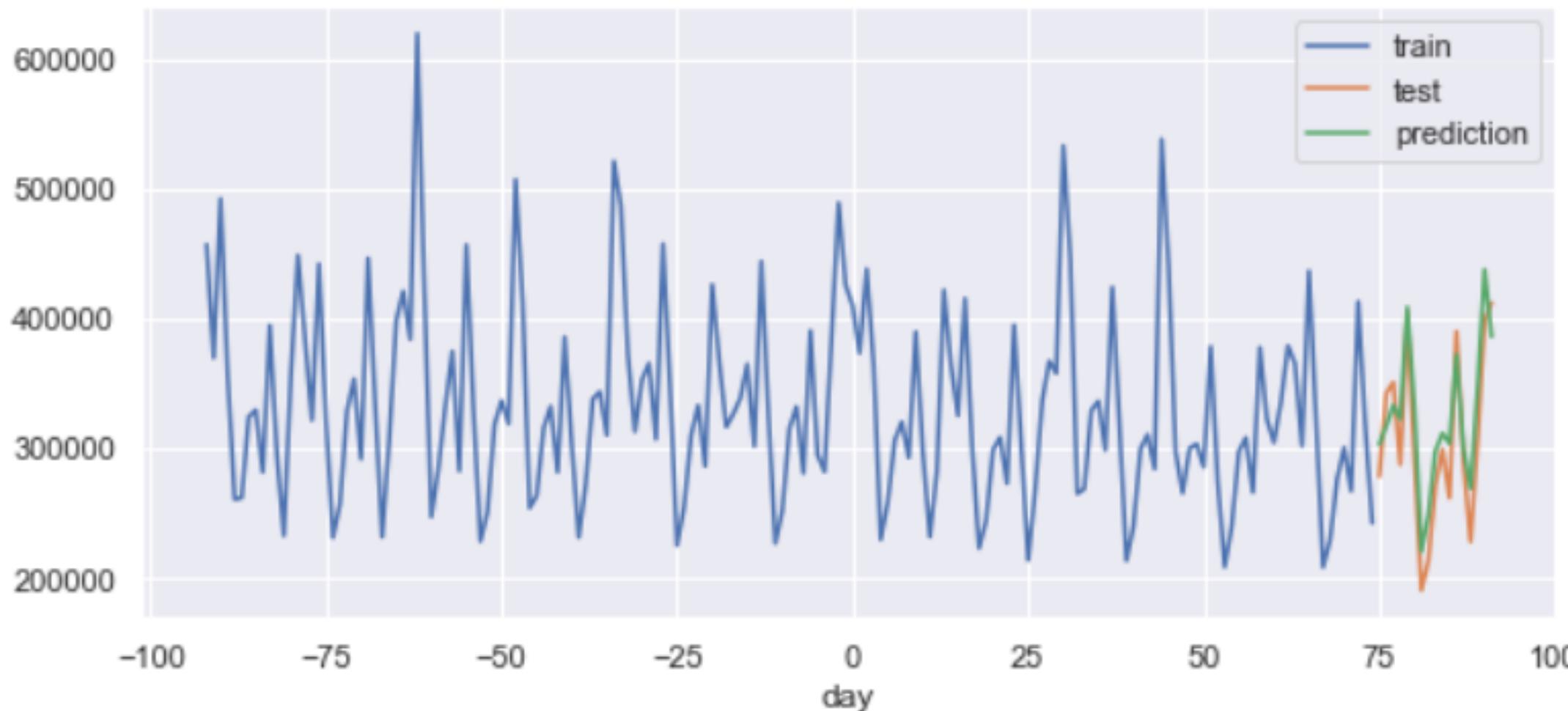
En la primera fase de entrenamiento de los modelos, se utilizó únicamente la variable venta, desde el primer paso se observó un error bastante alto, por lo que se continuó con el tuning de parámetros, pero pese a eso, el error seguía siendo muy alto, por lo que se decidió crear un nuevo atributo combinando el precio y la venta por producto sumado por día, de tal manera que obtendríamos el monto total de las ventas por día.

Los modelos presentaron una disminución del error notable (véase tabla 1), por lo que después de optimizar los parámetros para disminuir el error medido a través de las métricas MAE, MSE, RMSE, el modelo Lineal Ridge y el simple fueron los que mejor se ajustaron a los datos dado que son los que minimizan más el error.

Tabla 1

Modelo	MAE	MAE
Lineal	61520	27385.153
Polinomial	61572	28989
Regresión ridge	61521	27385.2
Regresión con bosques aleatorios	64924	29472
Ridge bayesiana	61128	27790

En la gráfica de abajo se observa el modelo ajustado en color verde que se logra ajustar de muy buena manera a los datos



Siguientes pasos

Yo creo que el modelo lineal ridge se justa de muy buena manera a los datos y predice valores bastante cercanos a los esperados, por lo que podría pensarse en colocar el modelo en producción. No sin antes aplicar una técnica de validación conocida como **Backtesting**. Una vez validado es preciso importar los pesos del modelo ya entrenad utilizando las bibliotecas pickle y joblib, colocarlo en un servidor, ya sea en físico o con algún servicio en la nube o a través de un container, una vez puesto en producción, vigilar su comportamiento y continuar alimentándolo con más datos.

Otro paso antes de la puesta en producción es validar si solo se busca hacer forecasting durante esos meses donde se tomaron los datos o es preciso reentrenar el modelo con otro dataset que contenga información de todos los meses del año, al cual se le podría añadir otras variables como meses pandémicos, índice de contagios, semáforo de covid por mes, inflación mensual, leads y datos de funnel de ventas, membresías, información del cliente, como poder adquisitivo, nivel de estudios, tipo de cliente, entre otros. También otros modelos que se pueden entrenar son las redes neuronales, máquinas de soporte vectorial o inclusive métodos más tradicionales en estadística y análisis de series de tiempo como un ARIMA.

Para este problema, ¿Qué es mejor, un modelo general con toda la información o un modelo particular para cada producto - tienda?

Esto depende de los objetivos, pero lo ideal es hacer ambos casos, un modelo general que podría ayudar a determinar los objetivos de venta para los próximos meses, mejorar el manejo de presupuesto en campañas de marketing, mientras que un modelo por producto ayudaría a priorizar canales de venta para ciertos productos, como los más vendidos, aquellos que no se venden, mejorar la cadena de suministro, entre otros.