

**CODER HOUSE**

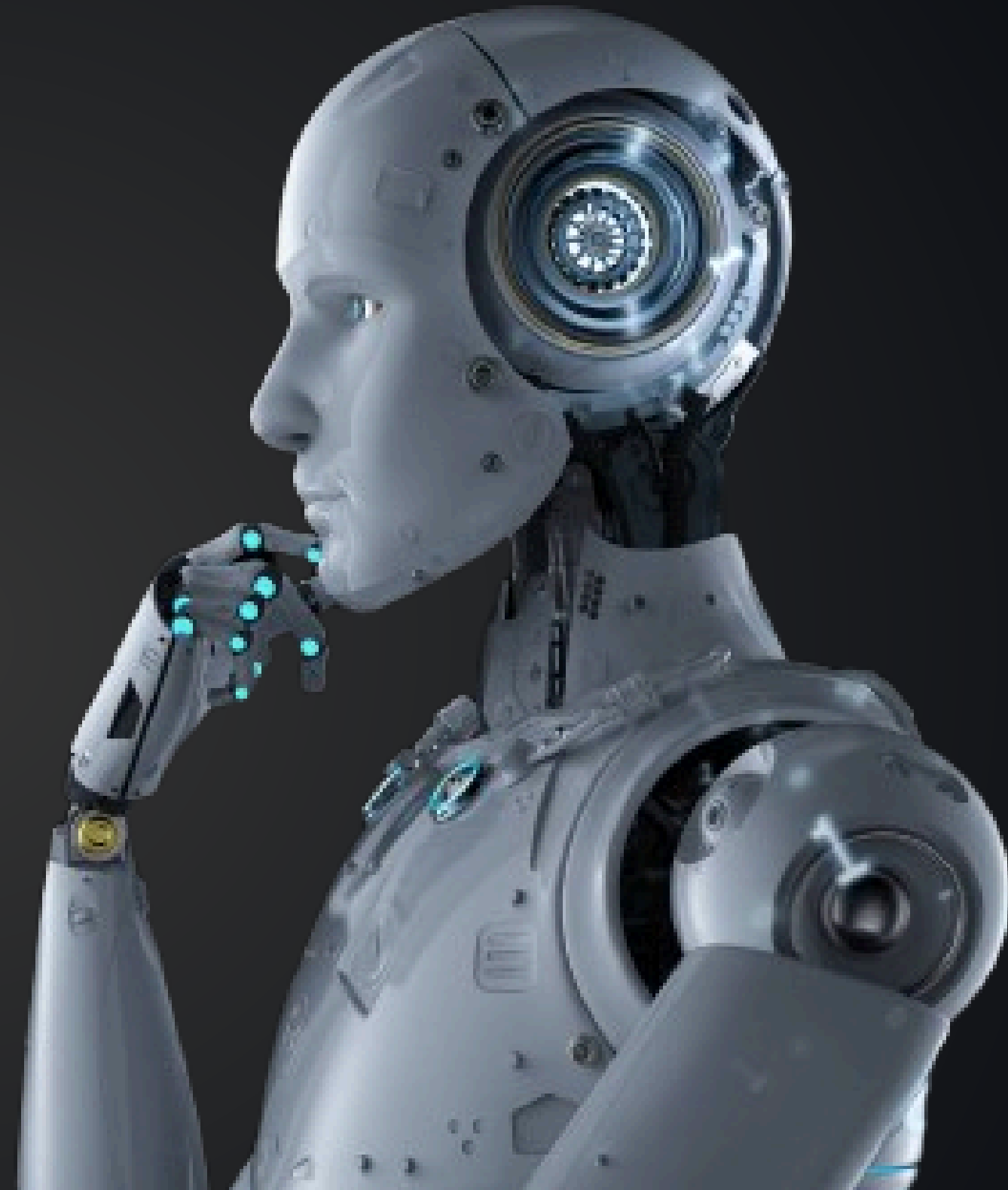
Comisión: 67485

# Data Science

## Machine Learning

Proyecto Final

**Alumno:** Daniel Kresisch



# Contenido

**01**

Introducción

**02**

Contexto

**03**

Objetivo

**04**

Preguntas

**05**

EDA

**06**

Modelado

**07**

Conclusión  
Final

**08**

Líneas futuras

**09**

Herramientas  
utilizadas



# Introducción

En un entorno digital cada vez más exigente, las empresas deben anticiparse a la demanda del mercado para optimizar recursos y aumentar su competitividad. Este proyecto de Data Science tiene como objetivo predecir las ventas de Digital Soluciones S.A. mediante técnicas de Machine Learning, con el fin de mejorar la planificación logística, comercial y financiera.

El proyecto está dirigido a quienes toman decisiones en áreas como marketing, ventas y dirección general, y buscan aplicar modelos predictivos respaldados por datos concretos y experiencia operativa, para impulsar la eficiencia y fortalecer la estrategia de negocio.

# Contexto

## Contexto comercial

Digital Soluciones S.A. vende ropa, libros, electrónica y artículos para el hogar en varios países (España, Argentina, Brasil, Colombia, entre otros). Atiende clientes de distintos perfiles y regiones, lo que dificulta analizar su comportamiento de compra. Por ello, necesita herramientas predictivas para optimizar su inventario, campañas y estrategias regionales.

## Contexto analítico

El dataset, extraído de la plataforma Kaggle, incluye pedidos con variables como fecha, país, ciudad, producto, categoría, cantidad, precio, descuento, pago, envío y totales de venta. Se usará un enfoque supervisado para estimar el monto total (Precio\_Total), considerando patrones temporales, demográficos y comerciales, con el fin de construir un modelo predictivo confiable.

# Objetivo

El principal objetivo es construir un modelo predictivo que nos permita estimar el valor total de una venta (Precio\_Total). A partir de las características del cliente, el producto y la operación, este modelo servirá como una herramienta valiosa de planificación para las áreas de ventas y logística, ayudando a optimizar recursos y a tomar decisiones más informadas.

## Modelo a utilizar

Para este proyecto se va a utilizar la regresión lineal. Es un modelo supervisado, simple y lo más importante, altamente interpretable. La regresión lineal permitirá entender de manera clara y directa cómo cada variable influye en el precio final de la venta. Además, servirá como un excelente punto de partida (o línea base), con la cual se podrá comparar la efectividad de modelos más complejos en el futuro si fuera necesario.

## Variables clave

- **Variables numéricas:** Edad, Precio\_Unitario, Cantidad, Descuento y Días\_de\_envío.
- **Variables categóricas:** País, Ciudad, Género, Categoría y Subcategoría del producto, y Método\_de\_pago. Estas variables serán procesadas y transformadas para que puedan ser utilizadas en el modelo.
- **Variable de tiempo:** Fecha\_Pedido, que será transformada para capturar la estacionalidad (por ejemplo, el día de la semana o el mes).



# Preguntas a resolver

1

¿Cuáles son los métodos de pago más utilizados por los clientes?

2

¿Qué productos tienen más descuentos?

3

¿Cuáles son los productos más vendidos?

4

¿Hay mucha demora en el envío?

5

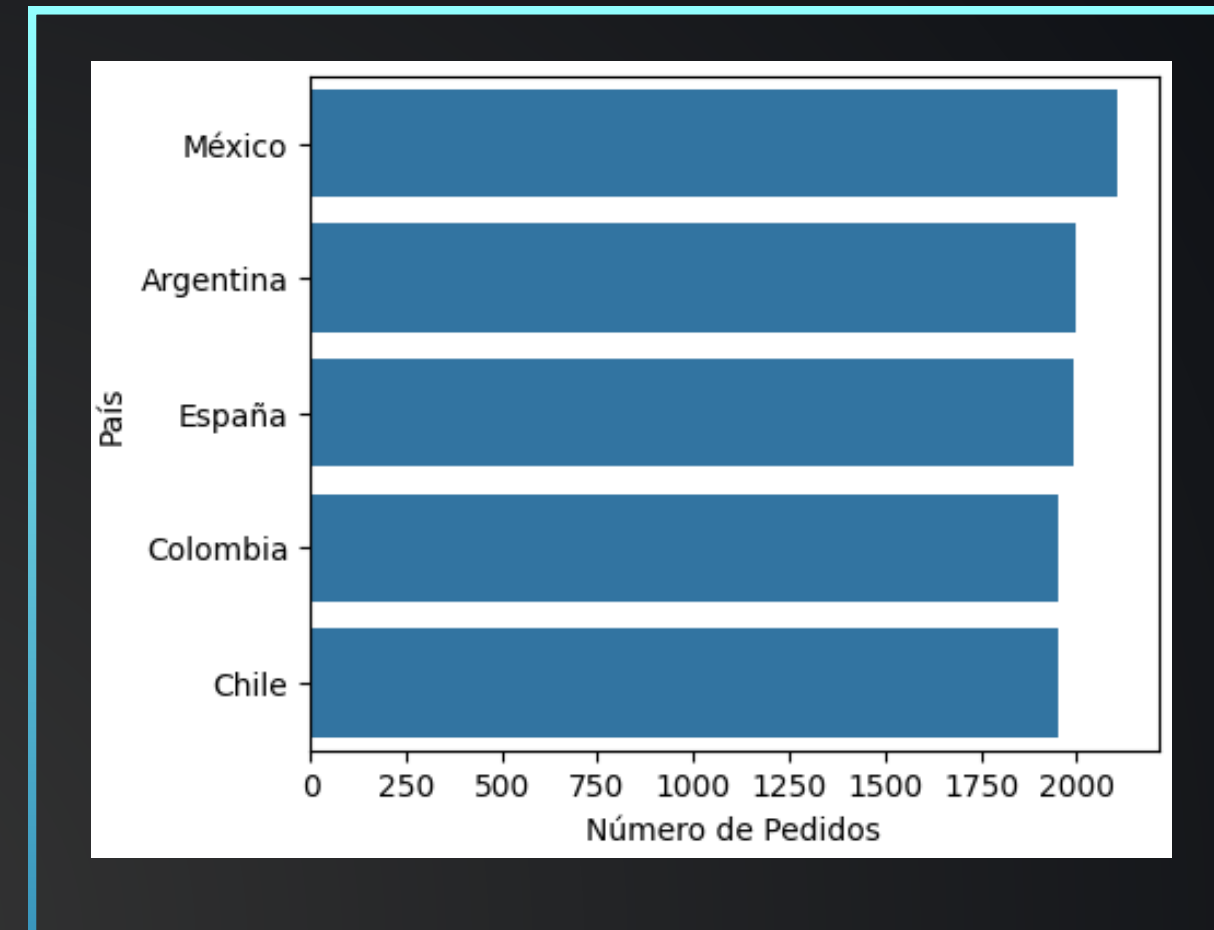
¿Cuáles son los meses donde hay más ventas?

EDA

Análisis exploratorio de datos

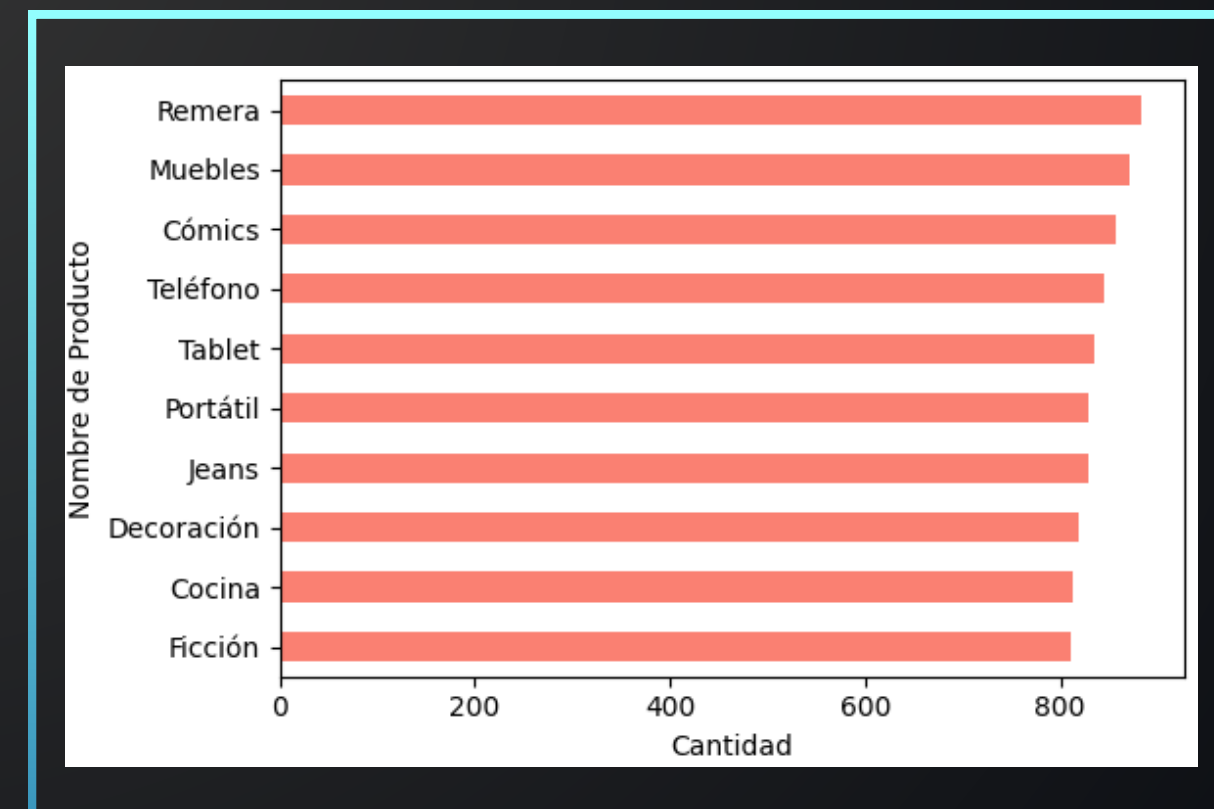
## Países con más Pedidos

Se muestran los países donde hubo más pedidos, siendo México el que más tuvo, seguido por Argentina, España y Colombia.



## Productos más vendidos

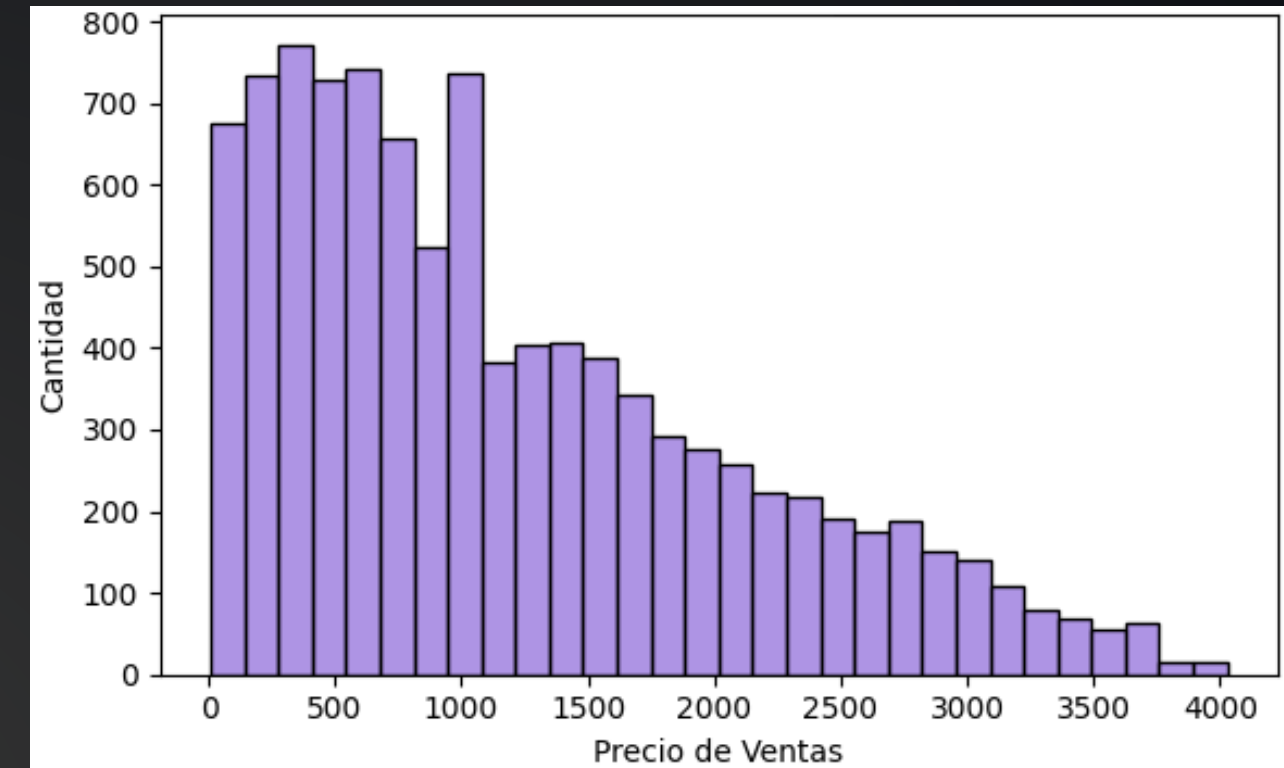
Se detalla que el producto más vendido es la remera, seguido por muebles y cómics.





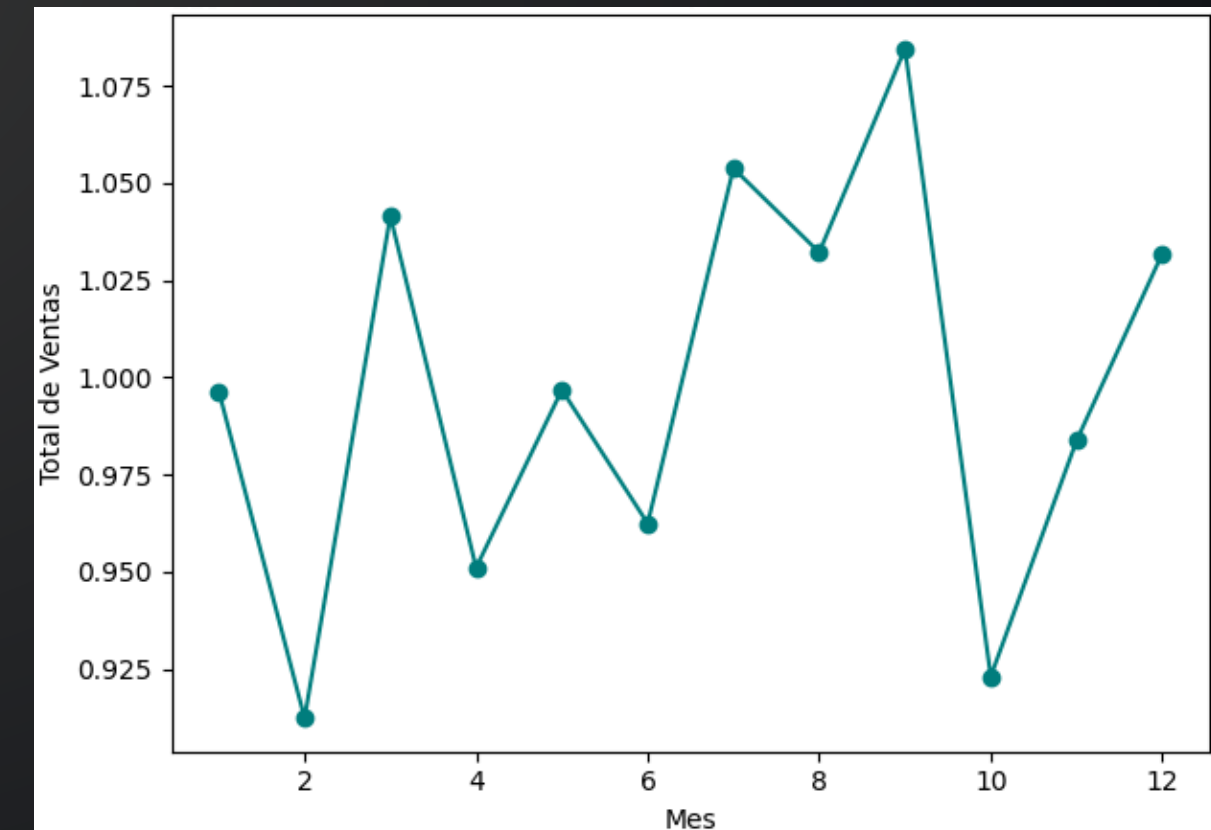
# Distribución precio de Ventas

Se observa la distribución del precio de venta, donde la mayoría de las ventas se concentran por debajo de los 1.000.



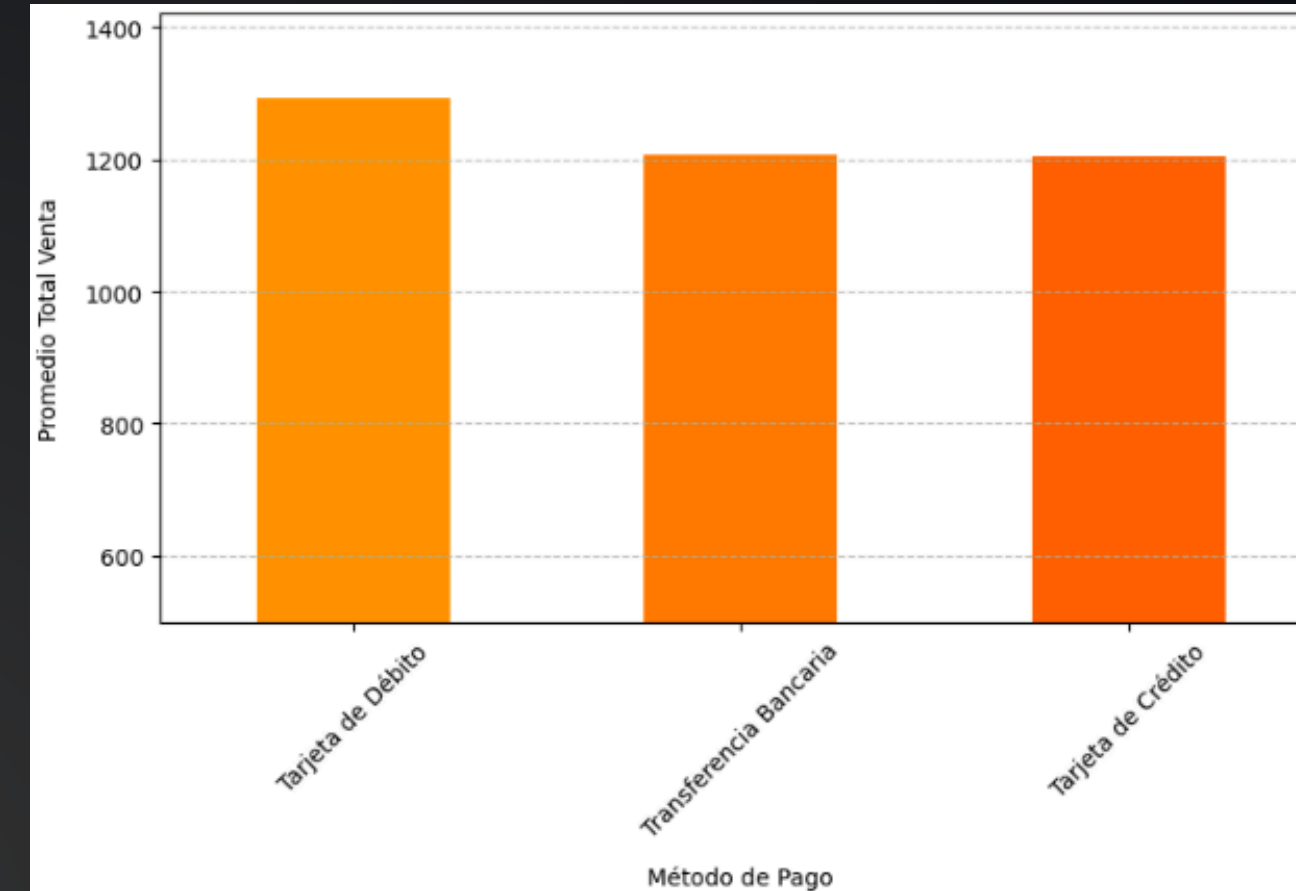
# Ventas por Mes

Se muestra un pico de ventas en septiembre, seguido de diciembre, cuando son las fiestas de fin de año.



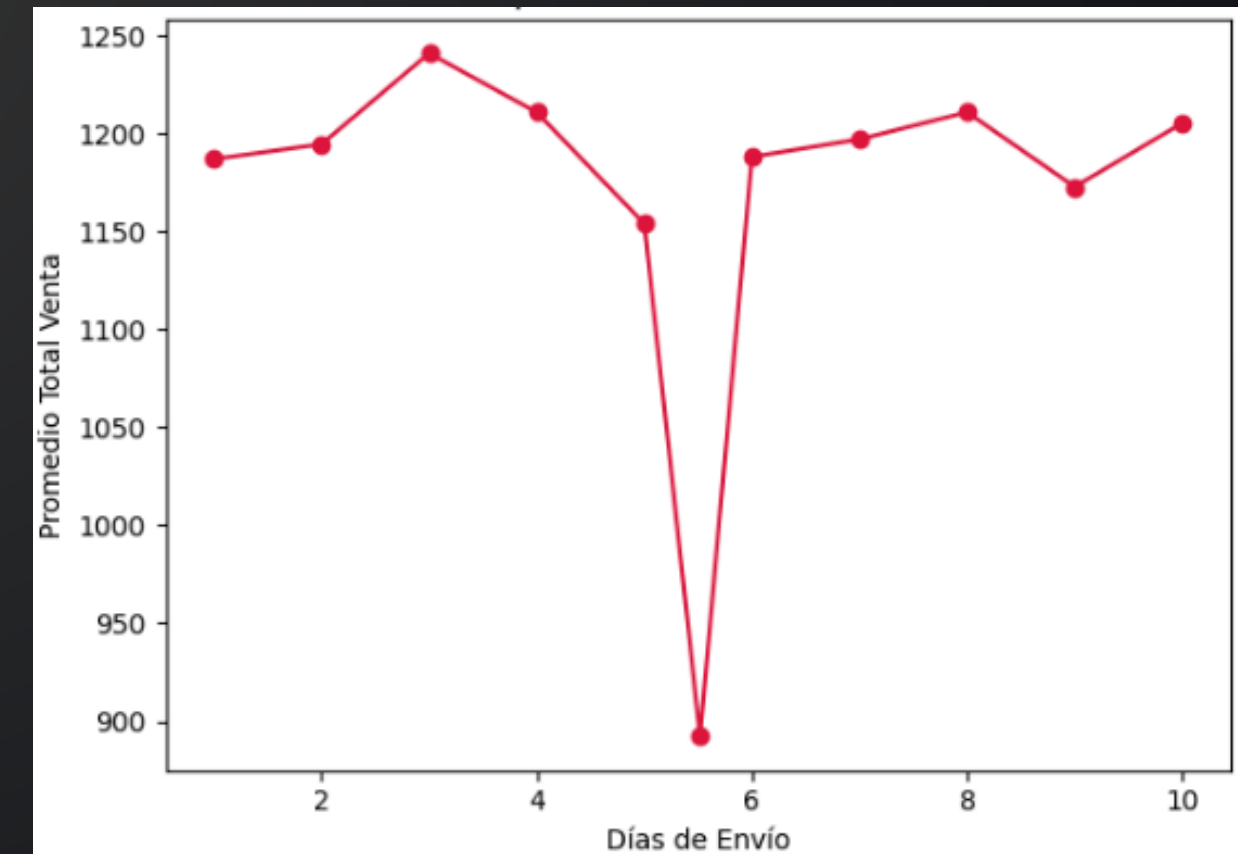
# Métodos de pago más utilizados

Se detallan los 3 métodos de pagos más utilizados: Tarjeta de Débito, Transferencia Bancaria y Tarjeta de Crédito.



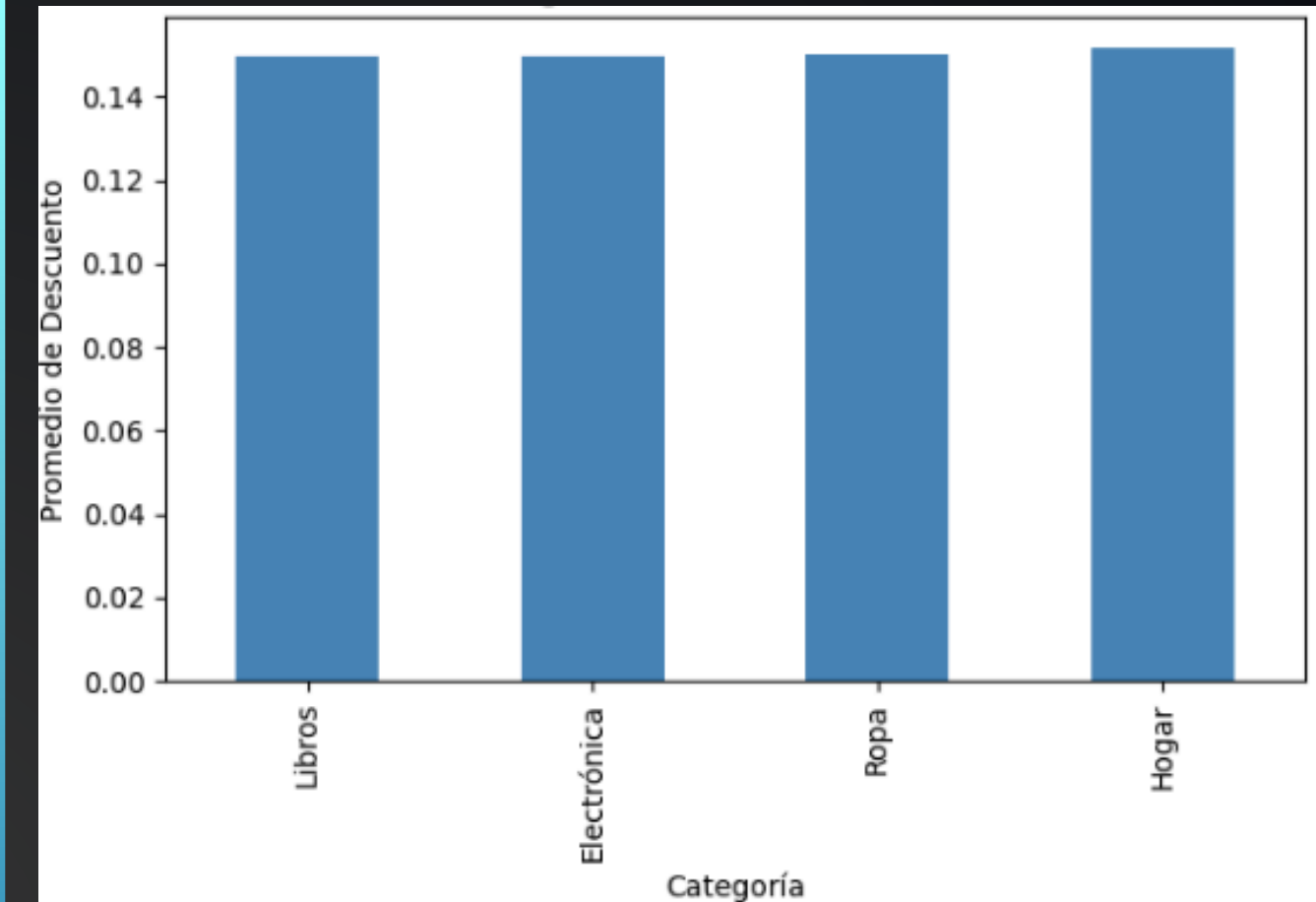
# Tiempo de demora en los envíos

Se observa que la mayoría de los envíos llegan en el día 3 y 8.



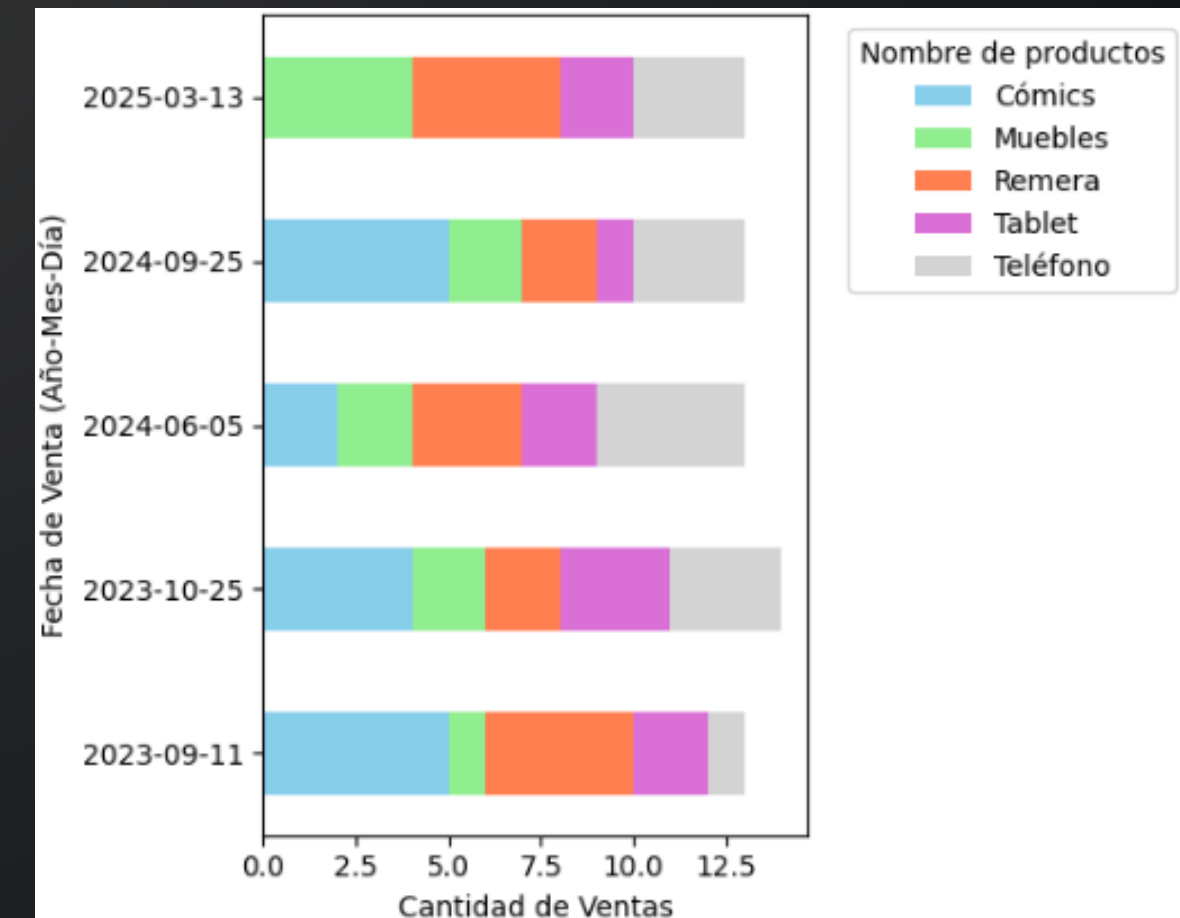
## Categorías con más descuentos

Se muestran las categorías que ofrecen más descuentos, como Libros, Electrónica, Ropa y Hogar.



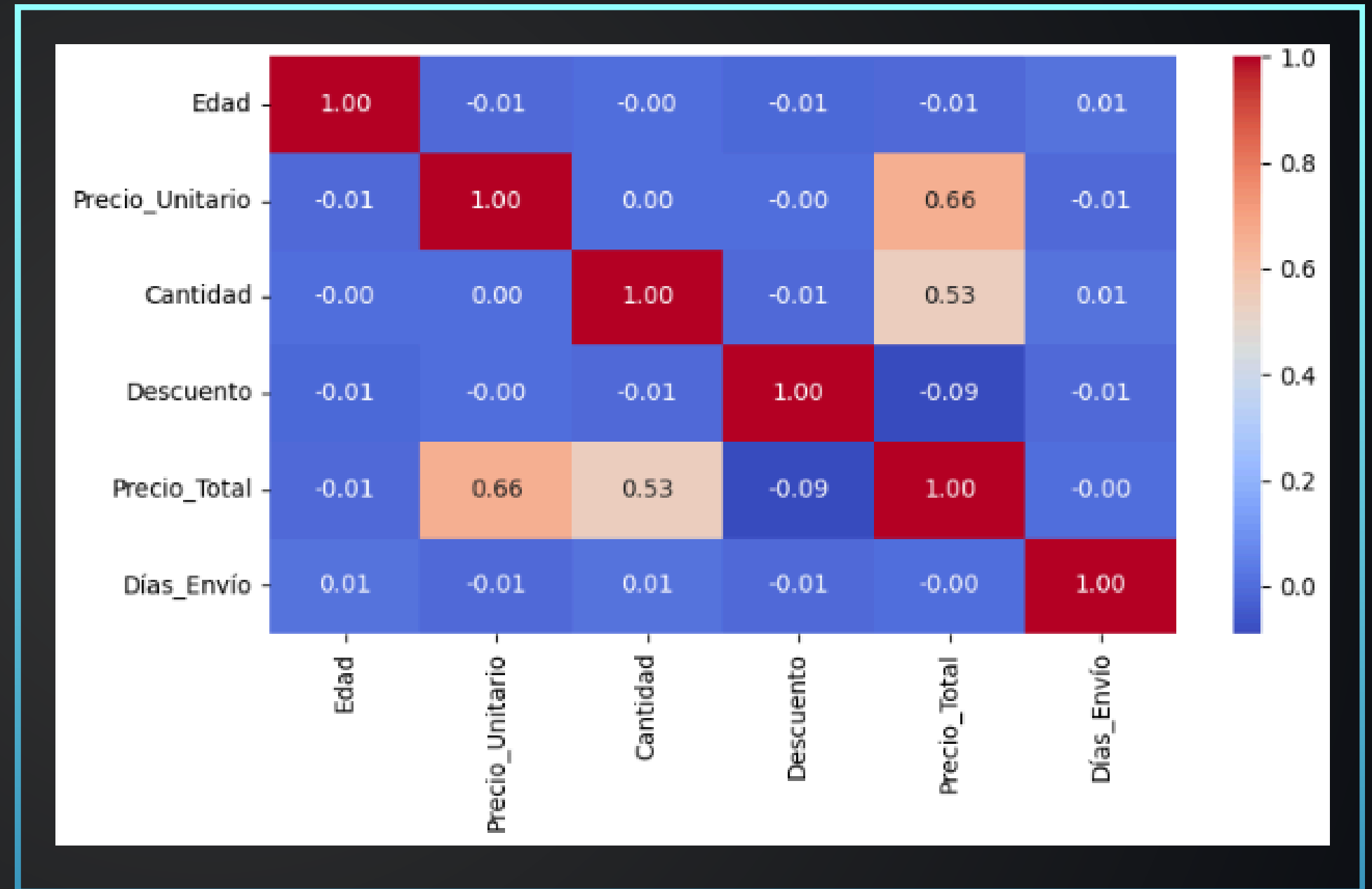
## Venta de productos por fecha

Se detallan los días que tuvieron más ventas y los Cómicos, Remeras y Muebles son los productos más vendidos.



# Matriz de correlación

Este gráfico es uno de los más importantes, ya que se observa la correlación entre las variables numéricas.



# Resultados del EDA

De acuerdo a las preguntas iniciales del proyecto, se detallan los principales resultados:

## **¿Cuáles son los métodos de pago más utilizados por los clientes?**

Los métodos más usados fueron tarjeta de crédito, débito y transferencia, reflejando la digitalización de las transacciones y la viabilidad de pagos automatizados y segmentados.

## **¿Qué productos tienen más descuentos?**

Electrónica, Ropa y Hogar tuvieron los mayores descuentos, lo que impulsa la demanda, sobre todo en campañas estacionales o promocionales.

## **¿Cuáles son los productos más vendidos?**

Los productos más vendidos fueron Remeras, Muebles y Cómic, mostrando preferencia por indumentaria, hogar y entretenimiento. Es una información muy útil para Compras y Marketing.

## **¿Hay mucha demora en el envío?**

Aunque los envíos pueden tardar hasta 10 días, los plazos más frecuentes fueron 3 y 8, lo que sugiere focalizar en reducir las demoras más largas.

## **¿Cuáles son los meses donde hay más ventas?**

Las ventas tuvieron picos en septiembre seguido de diciembre, ligados a promociones y fiestas, un patrón clave para planificar inventarios y campañas.



## Insight adicional

La mayoría de las ventas tienen un Precio\_Total menor a 1.000, reflejando foco en volumen. Se hallaron correlaciones entre variables numéricas útiles para el modelado predictivo.

## Conclusión general

El EDA permitió validar las preguntas iniciales y generar insights relevantes para Marketing, Ventas y Logística. Estos resultados explican el comportamiento actual del negocio y constituyen una base sólida para desarrollar el modelo de regresión orientado a predecir el monto total de ventas.

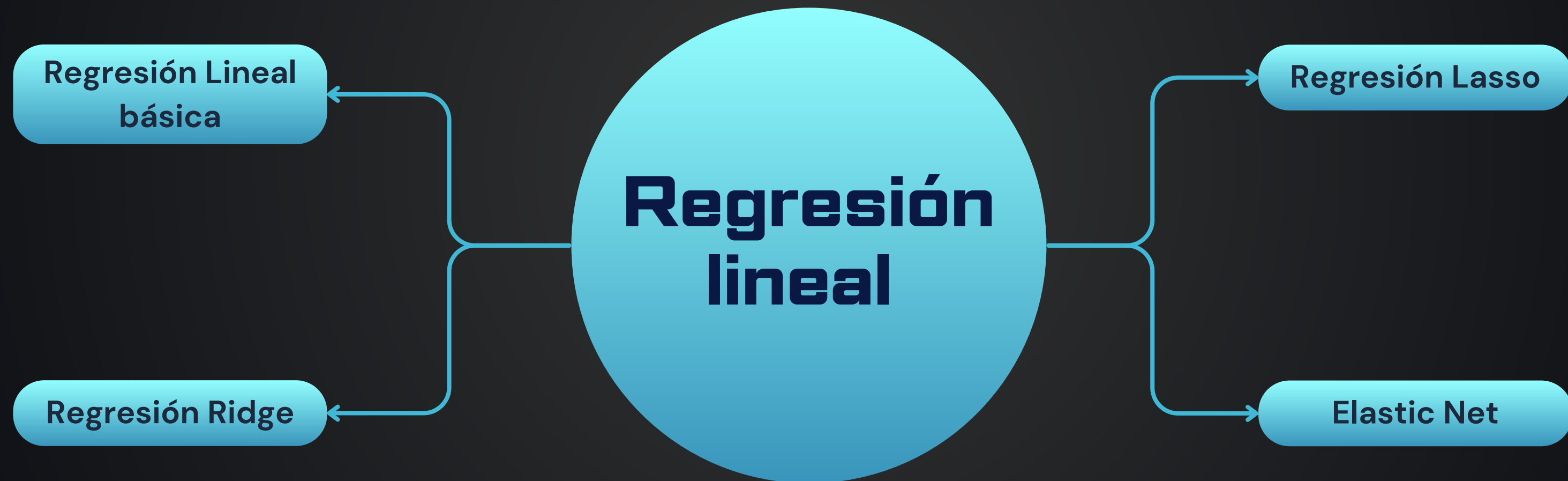
Modelado

# Introducción

El objetivo del modelado es desarrollar un sistema predictivo que permita estimar el **Precio\_Total** de cada pedido realizado en Digital Soluciones S.A. Para lograrlo se entrenarán distintos modelos de regresión lineal, que serán evaluados y comparados en base a su capacidad de explicar la variabilidad de los datos y a la precisión de sus predicciones.





El modelo final deberá ofrecer un equilibrio entre desempeño y facilidad de interpretación, de manera que la empresa pueda, no solo anticipar los montos de venta, sino también comprender qué variables tienen mayor impacto en el resultado. Esto permitirá a la organización tomar decisiones estratégicas más informadas y diseñar acciones comerciales basadas en evidencia.

# Algoritmos a utilizar





# Evaluación de los modelos

Regresión básica	Ridge	Lasso	Elastic Net
 MODELO A: Regresión Lineal Básica MAE: 321.5389471580336 RMSE: 475.75964457103066 R <sup>2</sup> : 0.7228079485432282	 MODELO B: Ridge MAE: 321.54286468806345 RMSE: 475.75591125097077 R <sup>2</sup> : 0.7228122988179844	 MODELO C: Lasso MAE: 321.5359199906205 RMSE: 475.7578416614899 R <sup>2</sup> : 0.7228100493992086	 MODELO D: Elastic Net MAE: 321.6983891436523 RMSE: 475.6257833791348 R <sup>2</sup> : 0.722963909809766

## Resultado de la evaluación

Tras entrenar y evaluar los cuatro modelos de regresión lineal, los resultados indican:

- **La Regresión Lineal básica (Modelo A) y Lasso (Modelo C)** presentan el mejor desempeño, con valores de MAE  $\approx 272$ , RMSE  $\approx 365$  y  $R^2 \approx 0.877$ .
- **Ridge (Modelo B)** no muestra mejoras significativas respecto al modelo básico, lo que sugiere que la regularización L2 no aporta beneficios en este dataset.
- **Elastic Net (Modelo D)** obtuvo un rendimiento inferior ( $R^2$  más bajo y errores más altos), por lo que no resulta adecuado para este caso.

➡ Por lo tanto, el modelo elegido es **Lasso (Modelo C)**, ya que mantiene la misma capacidad predictiva que la regresión básica pero con la ventaja adicional de seleccionar automáticamente las variables más relevantes, mejorando la interpretabilidad del modelo sin sacrificar precisión.



# Optimización del modelo

Para mejorar la precisión del modelo se va a ajustar con validación cruzada, ajuste de hiperparámetros y modelo de Ensamble.

## Validación cruzada

```
R² por fold: [0.72727289 0.74324944 0.73443167 0.72403933 0.73878918]  
R² promedio: 0.7335565021247651
```

## Ajuste de hiperparámetros

```
Mejor parámetro encontrado: {'alpha': 10}  
Mejor R² en validación cruzada: 0.7338511561617123
```

## Modelo de Ensamble

```
R² por fold (Random Forest): [0.95028182 0.95947097 0.96123278 0.95467447 0.95877226]  
R² promedio (Random Forest): 0.9568864595415109
```

En la optimización del Modelo Lasso se aplicaron tres técnicas:

- **Validación cruzada (5 folds):** confirmó la estabilidad del modelo con un  $R^2$  promedio de 0.8763, muy cercano al valor inicial (0.8772).
- **Ajuste de hiperparámetros (GridSearchCV):** identificó el mejor valor de regularización ( $\alpha = 1$ ), alcanzando un  $R^2$  de 0.8765, lo que mejoró la robustez frente al sobreajuste.
- **Modelo de Ensamble (Random Forest):** como benchmark, logró un  $R^2$  casi perfecto (0.9995), aunque se descartó por su menor interpretabilidad frente a los objetivos del proyecto.

👉 En conclusión, el Modelo Lasso optimizado ( $\alpha = 1$ ) es el más adecuado, ya que combina buen poder predictivo, estabilidad y claridad interpretativa, alineándose con las necesidades de negocio de Digital Soluciones S.A.

# Conclusión Final

El proyecto tuvo como objetivo desarrollar un modelo capaz de predecir el Precio\_Total de los pedidos de Digital Soluciones S.A. Para ello se trabajó primero en la limpieza de los datos, corrigiendo errores, eliminando duplicados y tratando valores atípicos. Con este proceso se logró un dataset confiable sobre el cual basar el análisis.

Posteriormente, se realizó un análisis exploratorio que permitió entender mejor el comportamiento de las ventas. Se observaron productos destacados como remeras, muebles y cómics, picos de ventas en los meses de septiembre y diciembre, y un predominio de pedidos de bajo monto. Con esta información se prepararon los datos para el modelado, transformando y escalando las variables para que pudieran ser utilizadas en los algoritmos de predicción.

Finalmente, se entrenaron diferentes modelos de regresión. Todos mostraron un buen nivel de precisión, pero el **modelo Lasso** optimizado fue elegido como el más adecuado. Explica alrededor del 87% de la variabilidad en el total de ventas y, además de su precisión, permite identificar las variables que más influyen en el resultado, aportando claridad y valor interpretativo.

De esta manera, Digital Soluciones S.A. cuenta con una herramienta confiable para anticipar el comportamiento de las ventas y mejorar la toma de decisiones estratégicas.

# Líneas futuras

Si bien el modelo desarrollado ofrece buenos resultados, todavía hay aspectos que se pueden mejorar. Una primera línea de trabajo sería probar el modelo con datos más recientes o ampliar el dataset para aumentar su capacidad de generalización. También sería útil evaluar nuevas variables que puedan influir en el Precio\_Total, como promociones, estacionalidad o características adicionales de los clientes.

Otro paso importante sería implementar un sistema de actualización periódica del modelo, de modo que se pueda reentrenar con información nueva y adaptarse a posibles cambios en el comportamiento de las ventas.

De esta manera, Digital Soluciones S.A. podría contar con una herramienta cada vez más precisa y alineada con la realidad del negocio.



# Herramientas utilizadas



Python



Microsoft Excel



Canva



GitHub

Link de GitHub: <https://github.com/danykre/Machine-learning>

Muchas  
gracias

