




**Curso:** Data Science


**Comisión:** 42310

**Autor:** Leandro Montes Catari

**Tutor:** Ernesto Surijon Frimis


# Índice

 Introducción y contexto.

 Preguntas de Interés.

 Análisis Exploratorio.

 Insights.

 Evaluando Modelos de Machine Learning.

# **Introducción y contexto.**

## **Abstract**

Este conjunto de datos Super Sales Store contiene datos de aproximadamente 10.000 registros de venta de una tienda en Estados Unidos en un periodo determinado. El conjunto de datos incluye datos categóricos de productos, clientes, tipos de envío y datos demográficos de los clientes. Las variables numéricas más importantes son el monto de las ventas, las ganancias, la cantidad de producto y los descuentos.

Este análisis de datos puede resultar de gran utilidad para una variedad de audiencias, incluidos gerentes y propietarios de tiendas minoristas que deseen comprender mejor el rendimiento de sus productos y clientes. También puede ser de interés para analistas de mercado que buscan identificar tendencias de compra y preferencias de los clientes. Además, los especialistas en logística y gestión de inventarios podrían beneficiarse al identificar patrones de demanda y optimizar las estrategias de almacenamiento y distribución.

## **Objetivo**

El objetivo principal de esta investigación es descubrir información procesable a partir de datos de ventas que puedan ayudar a la tienda a tomar decisiones. El análisis se centrará en comprender el comportamiento de los clientes, las tendencias de ventas y el desempeño regional para identificar oportunidades de mejora y maximización de ventas y ganancias.

## **Contexto Empresarial**

La empresa enfrenta el desafío de aumentar las ventas, adquirir nuevos clientes y expandirse a nuevos mercados para seguir siendo competitiva y lograr un crecimiento sostenible.

## **Contexto Comercial**

La empresa necesita identificar áreas de mejora en sus estrategias de ventas, orientación a clientes y expansión regional para maximizar la rentabilidad. Por lo que analizaremos a fondo los datos y con ayuda de las visualizaciones podremos dar respuesta a las preguntas que se nos vayan presentando.

## **Contexto Analítico**

Esta investigación empleará técnicas de análisis de datos para extraer información significativa del conjunto de datos. Los hallazgos se utilizarán para informar decisiones comerciales e impulsar iniciativas de crecimiento estratégico.

# Preguntas de Interés.



## Pregunta Principal.

¿Como aumentar las ventas y ganancias?

## Preguntas Secundarias.

- ¿Cómo han cambiado las ventas con el tiempo?
- ¿Qué regiones están obteniendo buenos resultados y cuáles necesitan mejorar en términos de ventas?
- ¿Qué segmentos de clientes son más propensos a comprar en cada región?
- ¿Qué categoría de productos genera las mayores ingresos y ganancias?

## Conociendo el Dataset

Los datos de nuestro dataset provienen de kaggle <https://www.kaggle.com/datasets/ishanshrivastava28/superstore-sales/data> y en general a si está compuesto:

```
[9] data.shape  
(9994, 24)
```

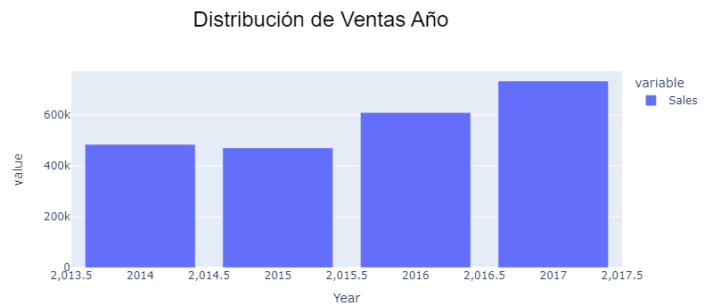
```
[6] data.columns  
  
Index(['Row ID', 'Order ID', 'Order Date', 'Ship Date', 'Ship Mode',  
       'Customer ID', 'Customer Name', 'Segment', 'Country', 'City', 'State',  
       'Postal Code', 'Region', 'Product ID', 'Category', 'Sub-Category',  
       'Product Name', 'Sales', 'Quantity', 'Discount', 'Profit', 'Year',  
       'Month', 'Month Name'],  
      dtype='object')
```

```
[7] data.info()  
  
<class 'pandas.core.frame.DataFrame'  
RangeIndex: 9994 entries, 0 to 9993  
Data columns (total 24 columns):  
#   Column                Non-Null Count  Dtype  
---  -  
0   Row ID                9994 non-null  int64  
1   Order ID              9994 non-null  object  
2   Order Date            9994 non-null  datetime64[ns]  
3   Ship Date             9994 non-null  datetime64[ns]  
4   Ship Mode             9994 non-null  object  
5   Customer ID           9994 non-null  object  
6   Customer Name         9994 non-null  object  
7   Segment               9994 non-null  object  
8   Country               9994 non-null  object  
9   City                  9994 non-null  object  
10  State                 9994 non-null  object  
11  Postal Code           9994 non-null  int64  
12  Region                9994 non-null  object  
13  Product ID            9994 non-null  object  
14  Category              9994 non-null  object  
15  Sub-Category          9994 non-null  object  
16  Product Name          9994 non-null  object  
17  Sales                 9994 non-null  float64  
18  Quantity              9994 non-null  int64  
19  Discount              9994 non-null  float64  
20  Profit                9994 non-null  float64  
21  Year                  9994 non-null  int64  
22  Month                 9994 non-null  int64  
23  Month Name            9994 non-null  object  
dtypes: datetime64[ns](2), float64(3), int64(5), object(14)  
memory usage: 1.8+ MB
```



## 1 ¿Cómo han cambiado las ventas con el tiempo?

Según su distribución las ventas han aumentado con el pasar de los años.



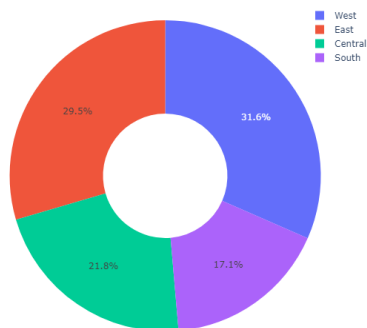
Podemos notar que la media de las ventas más alta fue en el año 2014, luego en 2016 intento repuntar, pero aun así no logro superar el máximo histórico, en 2017 cae, pero hay que considerar que no tenemos los datos completos de este año.



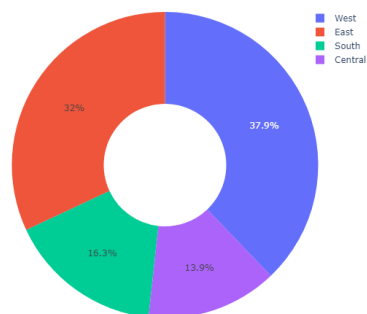
## 2 ¿Qué regiones están obteniendo buenos resultados y cuáles necesitan mejorar en términos de ventas?

En el grafico podemos notar que las regiones con mejores resultados son West y East, la que necesita mejorar sus resultados en ventas es la región de South y en ganancias la región Central.

Distribución de Ventas por Region



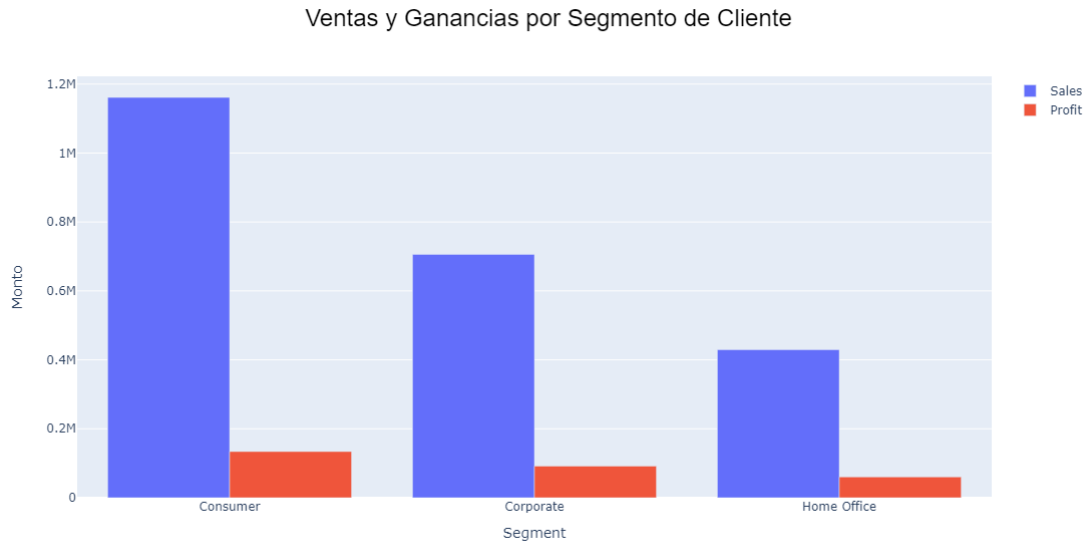
Distribución de Ganancias por Region





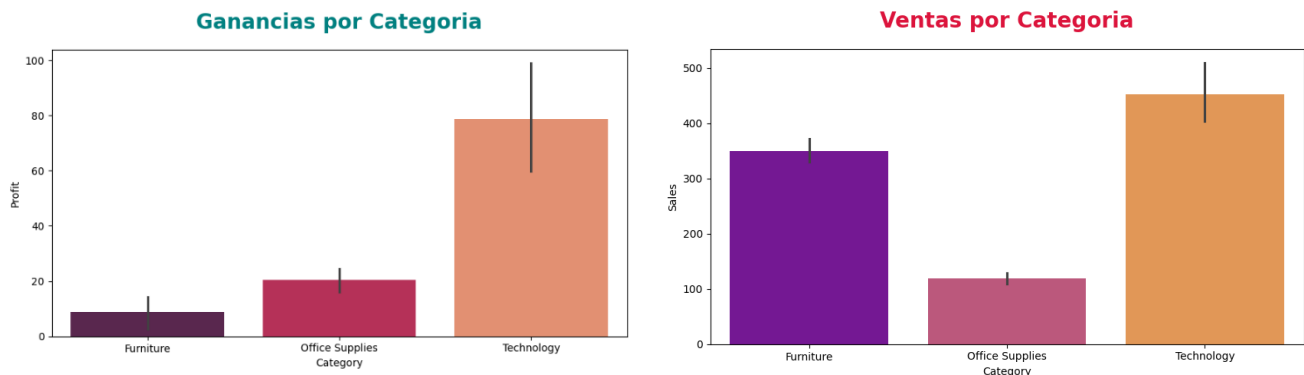
### 3 ¿Qué segmentos de clientes son más propensos a comprar en cada región?

Como se puede apreciar en general los clientes más propensos a comprar son los Consumer por ende la mayor cantidad de ganancias se genera por las compras de este segmento de clientes.



### 4 ¿Qué categoría de productos genera las mayores ingresos y ganancias?

A pesar de que la categoría Technology es la que tiene menos productos es la que mayores ingresos y ganancias genera, podría ser porque los artículos electrónicos son mucho más costosos y comerciales.





## Insights Ventas / Ganancias

- 🛒 Mejores meses de venta son noviembre y diciembre, esto se debe quizás a que es temporada decembrina.
- 🛒 Segmento clientes más propensos a comprar en todas las regiones son los Consumer.
- 🛒 Categoría con mayor cantidad de ventas es la de Office Supplies.
- 🛒 Producto que deja mayores ingresos en la categoría de Office Supplies es Binders.
- 🛒 Segundo producto más rentable son las Chairs pero esta pertenece a la categoría Furniture.
- 🛒 Regiones con más ganancias son las que tienen estados que son prósperos y con alto poder adquisitivo como lo son California y New York.
- 🛒 A pesar de que la categoría Technology es la que tiene menos productos es la que mayores ingresos y ganancias genera, podría ser porque los artículos electrónicos son mucho más costosos y comerciales.
- 🛒 Los productos que más ingresos generan en la categoría technology son los Phones, Machines y Copies estos por ser más costosos



# **Evaluando Modelos de Machine Learning**

## Entrenamiento de modelo

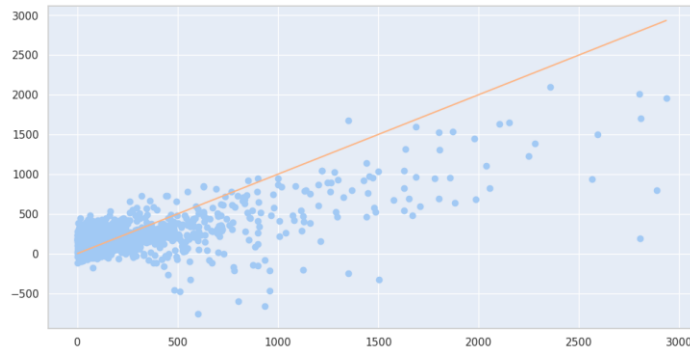
Siendo Sales mi variable objetivo, que es un valor continuo, elegiré un algoritmo de regresión, pero primero realizaremos algunas pruebas para ver cual se adecua mejor. Usaremos un Algoritmo de Regresión Lineal, luego CatBoostRegressor por último LGBMRegressor y obtendremos:

- 🛒 MSE (Mean Squared Error): Error Cuadrático Medio.
- 🛒 RMSE (Root Mean Squared Error): Raíz del Error Cuadrático Medio.
- 🛒 Reg\_score (R-squared): Coeficiente de Determinación.
- 🛒 MAPE (Mean Absolute Percentage Error): Error Absoluto Medio Porcentual.
- 🛒 MAE (Mean Absolute Error): Error Absoluto Medio.

## Resultados del Algoritmo de Regresión Lineal, Interpretación y Conclusión.

### Resultados:

0.38511748943441415  
0.4803313722532023  
MSE: 64499.76654144917  
RMSE: 253.96804236251688  
Reg\_score: 0.4803313722532023  
MAPE: 5.999169238452524  
MAE: 158.2398354331593



### Interpretación:

- 🛒 El MSE y RMSE son elevados, lo que indica que las predicciones del modelo tienen un error relativamente alto.
- 🛒 El valor de Reg\_score es positivo, lo que indica que la regresión lineal no explica una parte significativa de la varianza en la variable objetivo.
- 🛒 El MAPE es moderado, lo que sugiere que hay una desviación promedio del 5.99% de los valores reales.
- 🛒 El MAE indica que el error promedio en las predicciones es de 158.2398 unidades.

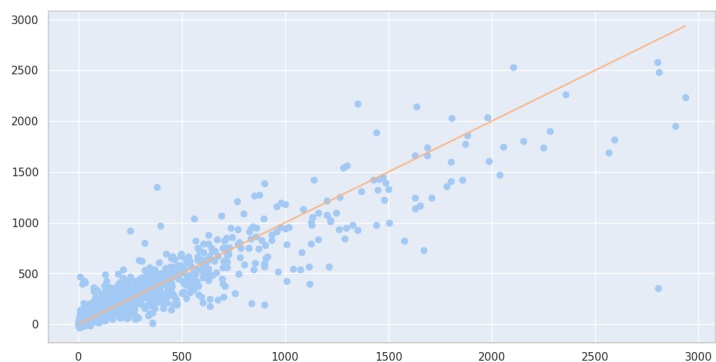
## Conclusiones:

- 🛒 El modelo de regresión lineal ajustado no parece ser un buen predictor de la variable objetivo.
- 🛒 Hay un error considerable en las predicciones del modelo.
- 🛒 Se necesitan más datos o un modelo diferente para obtener resultados más precisos.

## Resultados del Algoritmo CatBoostRegressor, Interpretación y Conclusión.

### Resultados:

0.9465576153491012  
0.8622137054053671  
MSE: 17101.636234033616  
RMSE: 130.77322445376046  
Reg\_score: 0.8622137054053671  
MAPE: 0.9000882339572204  
MAE: 57.817156422685606



### Interpretación:

- 🛒 El MSE y RMSE son inferiores al modelo de Regresión Lineal.
- 🛒 El valor de Reg\_score es positivo, lo que indica que la regresión CAT explica una parte significativa de la varianza en la variable objetivo.
- 🛒 El MAPE es moderado, lo que sugiere que hay una desviación promedio del 0.9% de los valores reales.
- 🛒 El MAE indica que el error promedio en las predicciones es de 57.817 unidades.

## Conclusiones:

- 🛒 El modelo de regresión CAT ajustado parece ser el MEJOR predictor de la variable objetivo.

- 🛒 Las predicciones del modelo tienen un error relativamente bajo.
- 🛒 No se necesitan más datos o un modelo diferente para obtener resultados más precisos.

### Resultados del Algoritmo LGBMRegressor, Interpretación y Conclusión.

#### Resultados:

0.9228893710909888  
0.8417782006079353  
MSE: 19638.03196433973  
RMSE: 140.13576261732666  
Reg\_score: 0.8417782006079353  
MAPE: 1.2220081438035713  
MAE: 63.76185246019997



#### Interpretación:

- 🛒 El MSE y RMSE son inferiores al modelo de Regresión Lineal y mayores al modelo CAT.
- 🛒 El valor de Reg\_score es positivo, lo que indica que la regresión LightGBM explica una parte significativa de la varianza en la variable objetivo.
- 🛒 El MAPE es bajo, lo que sugiere que hay una desviación promedio del 1.22% de los valores reales.
- 🛒 El MAE indica que el error promedio en las predicciones es de 63.761 unidades.

#### Conclusiones:

- 🛒 El modelo de regresión LightGBM ajustado parece ser un buen predictor de la variable objetivo.
- 🛒 Las predicciones del modelo tienen un error relativamente bajo.
- 🛒 No se necesitan más datos o un modelo diferente para obtener resultados más precisos.