# Análisis de Ventas de Autos Eléctricos

Marco Salvatierra

Octubre 2024

## Introducción

En la actualidad, la exploración y el modelado de datos juegan un rol fundamental en la ciencia de datos, ya que permiten a las organizaciones comprender mejor su entorno y tomar decisiones bien fundamentadas. En esta actividad, me centraré en la limpieza y preparación de un conjunto de datos relacionado con las ventas globales de autos eléctricos, utilizando técnicas avanzadas de análisis de datos.

La predicción precisa de estas ventas resulta crucial tanto para los fabricantes de automóviles como para la formulación de políticas de sostenibilidad. Esto facilita la planificación en áreas como la producción, distribución y desarrollo de estrategias de mercado.

## Propósito

El objetivo de este proyecto es desarrollar un proceso de limpieza de datos utilizando técnicas avanzadas para depurar y estructurar la información sobre las ventas de autos eléctricos. Para ello, se emplearán librerías de Python como *pandas* y *NumPy*, entre otras especializadas en la gestión de datos. Además, se evaluará la precisión de los resultados obtenidos y se presentarán los hallazgos mediante un panel interactivo desarrollado en *Streamlit*, lo que permitirá una visualización clara y accesible del análisis de los datos.

## Instrucciones

### a) Investigación

La investigación consistirá en obtener y analizar un conjunto de datos que refleje las ventas de autos eléctricos a nivel global. Aplicaré técnicas avanzadas de limpieza y preparación de datos, tales como el manejo de datos faltantes, la detección de valores atípicos y la normalización, utilizando librerías de Python. También exploraré la creación de visualizaciones interactivas con *Streamlit* para facilitar la interpretación y el análisis de los datos preparados.

### b) Análisis de Caso: Ventas de Autos Eléctricos

En los últimos años, las ventas de autos eléctricos han mostrado un crecimiento constante a nivel global, impulsado por factores como la expansión de la infraestructura de carga, incentivos gubernamentales, y la creciente diversidad de modelos disponibles. Este proyecto tiene como objetivo preparar un conjunto de datos históricos de ventas de autos eléctricos, aplicando un proceso riguroso de limpieza para asegurar la calidad de la información.

El análisis inicial del conjunto de datos incluirá la revisión de la cantidad de registros y columnas, la identificación de los tipos de datos en cada columna, y la detección de valores nulos o inconsistentes. Esta fase es fundamental para entender la estructura de los datos y resolver posibles problemas antes de realizar un análisis más detallado.

Posteriormente, realizaré un análisis descriptivo calculando estadísticas como la media, mediana, moda, desviación estándar y las distribuciones de las variables. Estas estadísticas permitirán detectar posibles anomalías y obtener una visión cuantitativa clara de los datos.

## Visualización de Datos

La visualización de datos será clave en este análisis. Utilizando *Streamlit*, crearé un panel interactivo que facilitará la exploración de los resultados. Además, emplearé herramientas como *Plotly Exprés*, *Altair* y *Matplotlib* para generar gráficos dinámicos (líneas, barras y dispersión) que ilustren la evolución de las ventas de autos eléctricos a lo largo del tiempo. Esto permitirá comparar las ventas entre diferentes regiones y analizar la relación entre las características de los autos y sus ventas.

## Descripción de las Columnas

El conjunto de datos incluye varias columnas que proporcionan información detallada sobre las ventas de vehículos eléctricos. A continuación, se describen las principales columnas:

- **region**: Región geográfica donde se registran las ventas.
- **category**: Tipo de vehículo eléctrico, como "sedán", "SUV", "camioneta", entre otros.
- **parameter**: Parámetros específicos que describen características del vehículo o métricas de venta, como "ventas totales" o "ventas por modelo".
- **mode**: Modo de operación o uso del vehículo.
- **powertrain**: Tipo de tren motriz del vehículo, como "eléctrico", "híbrido", o "híbrido enchufable".
- **year**: Año en el que se registraron las ventas.
- **unit**: Unidad de medida utilizada para el valor en la columna *value*, como "número de unidades" o "miles de dólares".
- **value**: Representa el número de unidades vendidas o el valor monetario asociado a las ventas de vehículos eléctricos.
- **price**: Precio de venta promedio de los vehículos eléctricos en la categoría y región especificadas.
- **range_km**: Distancia máxima que un vehículo eléctrico puede recorrer con una sola carga, medida en kilómetros.
- **charging_time**: Tiempo necesario para cargar completamente el vehículo eléctrico, medido en horas.
- **sales_volume**: Número total de unidades vendidas de vehículos eléctricos en un periodo específico.
- **co2_saved**: Cantidad de emisiones de CO2 evitadas gracias a las ventas de vehículos eléctricos, medida en toneladas.
- **battery_capacity**: Capacidad de la batería del vehículo eléctrico, medida en kilovatios-hora (kWh).
- **energy_efficiency**: Eficiencia energética del vehículo, medida en vatios-hora por kilómetro (Wh/km).
- **weight_kg**: Peso total del vehículo eléctrico, medido en kilogramos.
- **number_of_seats**: Número total de asientos disponibles en el vehículo.
- **motor_power**: Potencia del motor del vehículo eléctrico, medida en kilovatios (kW).
- **distance_traveled**: Distancia total recorrida por los vehículos eléctricos, medida en kilómetros.

Este proyecto me permitirá obtener una visión clara sobre el comportamiento del mercado de los autos eléctricos a través del análisis de sus ventas. Además, la creación de un panel interactivo facilitará la visualización y el entendimiento de los datos, apoyando la toma de decisiones en este sector.

# Exploración inicial de los datos

Para la exploración de los datos se utilizará un código en Python `"5_analisisbasico.py"` que estará disponible en el enlace de GitHub proporcionado al final del documento.

## Cantidad de registros

| **Tamaño del Dataset** | (12654, 19) |
|---|---|

Table 1: Tamaño del dataset

El dataset cuenta con 12654 registros y 19 columnas.

## Tipo de datos

| **Columna** | **Tipo de Dato** |
|---|---|
| region | objeto |
| category | objeto |
| parameter | objeto |
| mode | objeto |
| powertrain | objeto |
| year | entero |
| unit | objeto |
| value | flotante |
| price | flotante |
| range_km | flotante |
| charging_time | flotante |
| sales_volume | flotante |
| co2_saved | flotante |
| battery_capacity | flotante |
| energy_efficiency | flotante |
| weight_kg | flotante |
| number_of_seats | flotante |
| motor_power | flotante |
| distance_traveled | flotante |

Table 2: Tipos de datos de las columnas del dataset

Se puede observar que la mayor parte de las columnas son flotantes (numericos), seguido de columnas de tipo objeto (atributos) y con una sola columna de tipo de dato entero.

### Valores de datos nulos y inconsistentes

| Columna | Valores Nulos |
|---|---|
| region | 0 |
| category | 0 |
| parameter | 0 |
| mode | 0 |
| powertrain | 0 |
| year | 0 |
| unit | 0 |
| value | 0 |
| price | ”2,158” |
| range_km | ”2,058” |
| charging_time | ”2,652” |
| sales_volume | ”2,532” |
| co2_saved | ”2,243” |
| battery_capacity | ”2,107” |
| energy_efficiency | ”2,224” |
| weight_kg | ”2,596” |
| number_of_seats | ”2,072” |
| motor_power | ”2,116” |
| distance_traveled | ”2,741” |

Table 3: Valores nulos encontrados en el dataset

Los valores proporcionados son los valores nulos encontrados en cada columna del dataset.

## Limpieza de datos

Para la limpieza de los datos se utilizará un código en Python `"5_analisisbasico.py"` que estará disponible en el enlace de GitHub proporcionado al final del documento.

| Descripción | Cantidad |
|---|---|
| Filas duplicadas | 0 |

Table 4: Resumen de filas duplicadas en el dataset

Se puede observar que el data set no cuenta con datos duplicados, por ende no se eliminara ninguna fila, sobre los valores inconsistentes se los solucionara más adelante con la normalización.

# Imputación de datos faltantes

Para la imputación de datos faltantes se utilizará un código en Python `"7_maeciimputacion.py"` que estará disponible en el enlace de GitHub proporcionado al final del documento.

| Columna | Promedio Original | Promedio Imputado |
|---|---|---|
| price | 55081.99 | 55012.44 |
| charging_time | 6.46 | 6.47 |
| range_km | 345.75 | 345.97 |
| motor_power | 186.75 | 186.54 |
| sales_volume | 2532.22 | 2531.40 |
| co2_saved | 349.07 | 349.93 |
| battery_capacity | 59.38 | 59.52 |
| energy_efficiency | 217.37 | 217.13 |
| number_of_seats | 4.50 | 4.49 |
| distance_traveled | 50168.32 | 49945.37 |
| weight_kg | 1748.65 | 1747.91 |

Table 5: Promedios de las columnas antes y después de la imputación

En este caso se realizó la imputación de valores faltantes utilizando el algoritmo IterativeImputer (MICE). En la tabla se muestran los valores antes de imputar y luego de imputar. El archivo Excel con los valores imputados estará disponible en GitHub bajo el nombre `"datos_modificados_maeci"`.

# Manejo de valores atípicos

Para la deteccion y manejo de valor atipicos se utilizará un código en Python `"detector_ati.py"` que estará disponible en el enlace de GitHub proporcionado al final del documento.

El objetivo general del código es preparar el conjunto de datos `"datos_modificados_maeci"`, limpio y procesado sobre vehículos eléctricos, eliminando valores atípicos y transformando datos clave para facilitar el análisis y la visualización en futuras investigaciones o informes.

## Transformación Logarítmica

Las columnas a las que se les aplica la transformación logarítmica son:

- value
- price
- range_km
- charging_time
- sales_volume
- co2_saved
- battery_capacity
- energy_efficiency
- weight_kg
- number_of_seats
- motor_power
- distance_traveled

**Resumen**

El logaritmo se aplica a todas las columnas que se encuentran desde la columna 9 en adelante del DataFrame original, lo que incluye varios parámetros numéricos relacionados con los vehículos eléctricos, con el objetivo de normalizar la distribución de los datos y facilitar el análisis. El excel con los valores atipicos eliminados estara disponible en el github con el nombre de `"IEA_Global_EV_Data_Sin_Outliers"`.

# Normalización y estandarización de datos

Para la normalizacion y estandarizacion de datos se utilizará un código en Python `"11_normalizacionStandarizacion.py"` que estará disponible en el enlace de GitHub proporcionado al final del documento.

## Contexto

La normalización y estandarización son técnicas de preprocesamiento de datos utilizadas en análisis estadístico y aprendizaje automático para asegurar que las variables tengan una escala comparable. Estas transformaciones son especialmente útiles cuando las variables tienen diferentes unidades o rangos, lo que puede afectar el rendimiento de los algoritmos de aprendizaje automático.

## Normalización (Min-Max Scaling)

**Descripción:** La normalización transforma los datos para que estén en un rango específico, generalmente entre 0 y 1. Esto se logra mediante la fórmula:

$$X_{\text{normalizado}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

**Uso:** Es útil cuando se desea mantener la distribución de los datos originales y se necesita que todos los atributos estén en la misma escala. Esto es especialmente importante en métodos que utilizan distancias, como k-NN o redes neuronales.

## Estandarización (Z-Score Standardization)

**Descripción:** La estandarización transforma los datos de tal manera que tengan una media de 0 y una desviación estándar de 1. La fórmula es:

$$X_{\text{estandarizado}} = \frac{X - \mu}{\sigma}$$

donde $\mu$ es la media y $\sigma$ es la desviación estándar de la variable.

**Uso:** Es particularmente útil para algoritmos que asumen que los datos están distribuidos normalmente (por ejemplo, regresión lineal, SVM, etc.). La estandarización es menos afectada por los outliers en comparación con la normalización.

## Variables Estandarizadas y Normalizadas

En tu caso, se están aplicando estas técnicas a las siguientes variables logarítmicas transformadas:

- Log_price: Precio de un producto o servicio en su forma logarítmica.
- Log_range_km: Rango de distancia (en kilómetros) que un vehículo eléctrico puede recorrer con una carga.
- Log_charging_time: Tiempo necesario para cargar el vehículo (en horas).
- Log_sales_volume: Volumen de ventas del producto (puede ser en unidades vendidas).
- Log_co2_saved: Cantidad de CO2 ahorrado en comparación con los vehículos convencionales.

- Log_battery_capacity: Capacidad de la batería (en kWh) en su forma logarítmica.
- Log_energy_efficiency: Eficiencia energética del vehículo (puede ser en km/kWh).
- Log_weight_kg: Peso del vehículo (en kilogramos).
- Log_number_of_seats: Número de asientos en el vehículo.
- Log_motor_power: Potencia del motor (en kW) en su forma logarítmica.
- Log_distance_traveled: Distancia total recorrida (en km) por el vehículo.

### Importancia

La normalización y estandarización de estas variables son cruciales para la comparación y el análisis de datos relacionados con vehículos eléctricos. Al transformar estas variables, se pueden aplicar modelos de aprendizaje automático de manera más efectiva y se mejora la interpretabilidad de los resultados. Los excel proporcionados por el codigo estara disponible en github con los nombres de `"Datos_Normalizados.csv"` y `"Datos_Estandarizados.csv"`.

## Análisis de Componentes Principales (PCA)

Para Análisis de Componentes Principales se utilizará un código en Python `"13_PCA.py"` que estará disponible en el enlace de GitHub proporcionado al final del documento.

### Aplicacion

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica de reducción de dimensionalidad que transforma un conjunto de variables posiblemente correlacionadas en un conjunto de variables no correlacionadas, llamadas *componentes principales*. El objetivo es reducir la dimensionalidad del conjunto de datos mientras se retiene la mayor cantidad de varianza posible.

### Selección de Componentes

En este caso, se han seleccionado 3 componentes principales para el análisis. La decisión de usar 3 componentes se basa en el balance entre la cantidad de varianza explicada y la simplicidad del modelo. A continuación se presenta el porcentaje de la varianza total explicada por los primeros tres componentes principales:

- Componente Principal 1: 20.52%
- Componente Principal 2: 14.13%
- Componente Principal 3: 9.82%

### Varianza Total Explicada

La varianza total explicada por los 3 componentes principales es del 44%. Esto significa que estos 3 componentes capturan el 44% de la información (varianza) presente en el conjunto de datos original. Aunque el PCA ha reducido el número de variables de manera significativa, sigue reteniendo una porción importante de la estructura original de los datos.

$$\text{Varianza Total Explicada} = 0.44$$

El análisis muestra que, a pesar de la reducción en el número de variables, los tres componentes principales logran capturar una cantidad suficiente de información para representar eficazmente el comportamiento de los datos. Aunque no se alcanza el 50% de varianza explicada, los componentes principales aún pueden proporcionar una visión clara y simplificada de la variabilidad en los datos, lo que facilita el análisis y la interpretación posterior.

### Importancia de los Componentes Principales

El uso de estos 3 componentes principales permite simplificar el análisis sin sacrificar demasiada información. Esta reducción es particularmente útil en aplicaciones de aprendizaje automático, ya que ayuda a disminuir el ruido en los datos, acelera el procesamiento y puede mejorar el rendimiento de los modelos predictivos.

## Preguntas reflexivas

### ¿Cuál es la importancia de la limpieza de datos en el proceso de análisis y modelado?

Al momento de aplicar todos los codigos a la base de datos, lo que me falto es entender cada columna, porque me salieron cosas que no tenian mucho sentido, la limpieza es muy importante ya que gracias a ello, podremos realizar un analisis a detalle.

### ¿Cómo influye la visualización de datos en la toma de decisiones en el contexto de ventas de autos eléctricos?

Al ver los datos en bruto, números, etc, es muy estresante ver muchos números, la verdad ni el jefe tomaria atención si le paso un reporte con puros números, en cambio preparan graficos y que se pueda seleccionar las variables que deseas es muy distinto, puedes ver como esta el precio de un producto a lo largo del tiempo, por país, etc, si se logra hacerlo a detalle hasta un niño de 5 años podria leer los graficos.

### ¿Qué desafíos podrían surgir al trabajar con conjuntos de datos grandes y complejos, y cómo se pueden abordar?

Cuando se trabaja con un conjunto de datos grandes, los datos faltantes son el problema, se registra el tiempo de cada venta pero si en una fecha no se registro una venta ya sea porque no hubo la venta o si en caso alguien lo registro mal y no lo elimino, si queremos realizar predicciones no sera muy sertera, se puede reemplazar utilizando la tecnica MAECI pero agregariamos algo que no se dio.