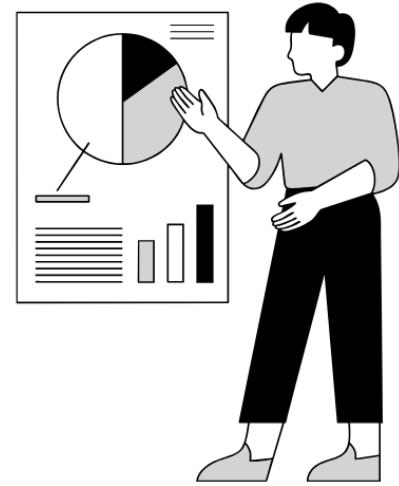


Puebla, Puebla.

# Análisis del Rendimiento y del Impacto Ambiental de los Vehículos Eléctricos a Nivel Mundial



Carlos Cortez Almaro  
Alumno



Introducción a la ciencia de datos  
Materia

Jaime Alejandro Romero Sierra  
Docente

## Índice

<b>Introducción.....</b>	<b>3</b>
<b>Metodología .....</b>	<b>5</b>
Proceso de Limpieza de Datos .....	5
Análisis Exploratorio de Datos .....	7
1. Descripción General de los Datos .....	7
2. Visualización y Distribución de Variables Individuales .....	14
3. Correlación entre Variables .....	24
4. Análisis de Valores Atípicos .....	34
5. Análisis de Valores Faltantes .....	37
6. Relación entre variables Categóricas y Numéricas .....	39
7. Observaciones y Hallazgos importantes .....	45
<b>1er Modelo de Machine Learning: Autonomía .....</b>	<b>48</b>
1) Descripción del Modelo .....	48
2) Justificación del Modelo .....	49
3) Implementación y Entrenamiento .....	50
4) Resultados y Evaluación .....	51
5) Visualizaciones de Resultados .....	53
6) Conclusión del Modelo 1: Predicción de la Autonomía (km).....	54
<b>2do Modelo de Machine Learning: Ahorro de CO<sub>2</sub>.....</b>	<b>54</b>
1) Descripción del Modelo .....	54
2) Justificación del Modelo .....	55
3) Implementación y Entrenamiento .....	56
4) Resultados y Evaluación .....	57
5) Visualizaciones de Resultados .....	58
6) Conclusión del Modelo 2: Predicción del CO <sub>2</sub> Ahorrado .....	59
<b>Dashboard.....</b>	<b>60</b>
1) Título y propósito del Dashboard .....	60
2) Uso y Beneficios del Dashboard .....	62
<b>Conclusiones y Futuras Líneas de Trabajo.....</b>	<b>63</b>
<b>Referencias Bibliográficas .....</b>	<b>65</b>

# Introducción

**El objetivo principal** de este proyecto es **analizar, mediante técnicas básicas de ciencia de datos, el rendimiento energético y el impacto ambiental de los vehículos eléctricos (EV) a nivel mundial**, considerando factores como la capacidad y el estado de las baterías, la autonomía, el consumo de energía, los costos de mantenimiento y las emisiones evitadas de CO<sub>2</sub>.

**El propósito** es **identificar patrones y correlaciones que permitan comprender mejor cómo las características técnicas, el tipo de uso y las condiciones regionales influyen tanto en la eficiencia del vehículo como en su contribución a la reducción de emisiones contaminantes.**

De manera complementaria, se busca evaluar el papel de la tecnología eléctrica como alternativa sostenible al transporte convencional y aportar evidencia cuantitativa que apoye la toma de decisiones hacia un modelo de movilidad más limpio y eficiente.

El sector transporte es uno de los principales emisores de gases de efecto invernadero, por lo que los vehículos eléctricos (EV) representan una estrategia clave para reducir el impacto ambiental y avanzar hacia una transición energética sostenible. Sin embargo, **aún existen dudas sobre su eficiencia en diferentes contextos**, lo que hace necesario analizar su rendimiento y sus beneficios reales.

**Este proyecto es relevante tanto técnica como ambientalmente.** Desde lo técnico, busca evaluar variables como autonomía, capacidad de batería, consumo y temperatura para comprender y optimizar el rendimiento de los EV. Desde lo ambiental, cuantificar el ahorro de CO<sub>2</sub> permite medir el impacto positivo de la electromovilidad a nivel global.

Aunque países como China aún dependen de energías contaminantes, estudios demuestran que los EV emiten entre 20% y 30% menos CO<sub>2</sub> que vehículos híbridos. Además, investigaciones globales indican que en más del 95% del mundo los eléctricos generan menos emisiones que los autos de combustión. Aun así, **voces como Akio Toyoda señalan la necesidad de una transición gradual apoyada por una red eléctrica limpia.**

Finalmente, contribuyendo a los Objetivos de Desarrollo Sostenible. Su propósito es generar conocimiento útil que apoye decisiones responsables y fortalezca el desarrollo de la movilidad eléctrica como parte de un futuro más limpio.

**El conjunto de datos empleado** para este proyecto está compuesto por **3,000 registros y 25 variables**, que abarcan una amplia gama de información técnica,

operativa, económica y ambiental de vehículos eléctricos de diversas regiones del mundo. En términos generales, las variables se dividen en dos grandes categorías.

- **Cuantitativas:**

- Battery\_Capacity\_kWh (capacidad de batería)
- Battery\_Health\_% (estado de salud de la batería)
- Range\_km (autonomía en kilómetros)
- Energy\_Consumption\_kWh\_per\_100km (consumo energético)
- Charging\_Power\_kW y Charging\_Time\_hr (potencia y tiempo de carga)
- CO2\_Saved\_tons (emisiones evitadas)
- Maintenance\_Cost\_USD, Insurance\_Cost\_USD, Monthly\_Charging\_Cost\_USD (costos operativos)
- Mileage\_km, Avg\_Speed\_kmh, Max\_Speed\_kmh (indicadores de uso y desempeño)

- **Cualitativas:**

- Make (marca del fabricante)
- Model (modelo del vehículo)
- Year (año de fabricación)
- Region (región geográfica donde opera)
- Vehicle\_Type (tipo de vehículo: SUV, sedán, camioneta, etc.)
- Usage\_Type (tipo de uso: personal, comercial o de flota)

Ambas con funciones analíticas específicas dentro del estudio.

# Metodología

## ***Proceso de limpieza de datos***

La limpieza del dataset se realizó siguiendo un proceso sistemático que permitió transformar una base inicialmente desordenada y con inconsistencias en un conjunto de datos confiable, coherente y utilizable para análisis de ciencia de datos.

### **i. Exploración inicial y diagnóstico**

Se inició con una revisión general del dataset utilizando funciones como info() e isnan().sum() para conocer los tipos de datos, la presencia de valores faltantes y el estado general de la base.

Aunque no había valores nulos explícitos, sí existían errores ocultos, como columnas numéricas almacenadas como texto y valores no válidos.

### **ii. Detección y eliminación de duplicados**

Se utilizaron funciones como duplicated() y drop\_duplicates() para identificar y eliminar 48 registros idénticos.

Dado que cada fila representaba un vehículo único, los duplicados no aportaban información nueva y podían distorsionar estadísticas y análisis posteriores.

### **iii. Corrección de valores atípicos e inconsistentes**

En varias columnas numéricas aparecían valores no numéricos (como palabras o caracteres extraños).

Para manejarlos se aplicó:

- Conversión a números con errors='coerce' para transformar textos en NaN.
- Revisión de outliers usando el método del rango intercuartílico (IQR).
- Sustitución de valores no plausibles o incorrectos por NaN.

### **iv. Detección y reemplazo de palabras extrañas**

Se encontraron cadenas contaminadas como "**Ver4\$zul**", así como inconsistencias de escritura.

Estas fueron detectadas usando ciclos for y reemplazadas por categorías válidas como "**Desconocido**" o por traducciones apropiadas.

#### v. Traducción y normalización del contenido

Las columnas y muchos valores estaban en inglés, por lo que se procedió a:

- Traducir nombres de columnas con rename().
- Detectar valores de texto con unique() y traducirlos usando replace().
- También se normalizó la escritura (mayúsculas, espacios, acentos) para evitar inconsistencias en el análisis.

#### vi. Conversión de tipos de datos

Todas las columnas fueron importadas como tipo *object*, por lo que se realizó:

- Conversión a float64 o numéricos donde correspondía.
- Validación posterior para confirmar que las columnas numéricas estaban listas para análisis estadísticos.
- Mantenimiento de copias \*\_orig temporalmente para conservar evidencia del proceso.

#### vii. Imputación de valores faltantes

Tras la conversión y la detección de errores, surgieron nuevos valores NaN. Estos se imputaron utilizando la **mediana** de cada columna, debido a su robustez frente a outliers.

Esta estrategia permitió conservar el mayor número de registros sin perder coherencia.

#### viii. Validación final

Se verificó que el dataset quedara completamente limpio:

- 0 duplicados.
- Sin valores nulos (o muy pocos, dependiendo de la estrategia).
- Columnas numéricas correctamente convertidas.
- Consistencia semántica y categórica en todas las variables.

## Análisis Exploratorio de Datos (EDA)

Esta es una etapa crítica en el proceso de análisis de datos, **implica resumir las características principales del conjunto de datos**, utilizando métodos visuales. EDA es esencial para comprender los patrones subyacentes, detectar anomalías y probar hipótesis antes de aplicar técnicas estadísticas más formales.

### 1. Descripción general de los datos

#### *Visión general*

El dataset utilizado en este proyecto contiene información técnica, operativa, económica y de sostenibilidad de vehículos eléctricos (VE) a nivel mundial.

```
[21]: df.shape
      0.0s
... (3531, 25)
```

Tras la limpieza realizada en la Fase 2, la base de datos final consta de **3,531 registros y 25 variables**; se eliminaron 48 duplicados exactos y se imputaron valores no numéricos convirtiéndolos en NaN y tratándolos con la mediana cuando correspondía.

#### *Tipos de variables*

El dataset contiene variables de tipo categóricas y de tipo numéricas.

```
[25]: df.dtypes
      0.0s
... ID_del_Vehiculo          object
Marca                  object
Modelo                 object
Año                   float64
Región                 object
Tipo_de_Vehículo        object
Capacidad_de_Batería_kWh float64
Salud_de_Batería_%       float64
Autonomía_km            float64
Potencia_de_Carga_kw    float64
Tiempo_de_Carga_hr      float64
Ciclos_de_Carga         float64
Consumo_de_Energía_por_100km_recorridos float64
Kilometraje_km           float64
Velocidad_Promedio_kmh   float64
Velocidad_Máxima_kmh     float64
Aceleración_0_100_kmh_seg float64
Temperatura_°C            float64
Tipo_de_Uso               object
CO2_Ahorrado_tons         float64
Costo_de_Mantenimiento_USD_por_año    float64
Costo_de_Seguro_USD_por_año    float64
Costo_de_Electricidad_en_USD_por_kWh   float64
Costo_Mensual_de_Carga_USD      float64
Valor_de_Reventa_USD          float64
dtype: object
```

Las **variables principales** relevantes para el análisis incluyen, entre otras:

Capacidad\_de\_Batería\_kWh, Salud\_de\_Batería\_%, Autonomía\_km, Consumo\_de\_Energía\_por\_100km\_recorridos, Kilometraje\_km, Temperatura\_°C, Tipo\_de\_Uso, Región, Costo\_de\_Electricidad\_en\_USD\_por\_kWh y CO2\_Ahorrado\_tons.

## Resumen estadístico

### Variables Numéricas

Para variables numéricas: capacidad de batería, salud de batería, autonomía, consumo energético, kilómetros, velocidades, aceleración, costos, CO<sub>2</sub> ahorrado, etc. Ocupamos

```
df.describe().T.round(3) # "T" para transponer la tabla de forma ordenada y "Round 3" para redondear a 3 decimales
```

✓ 0.0s

	count	mean	std	min	25%	50%	75%	max
Año	3531.0	2019.541	2.759	2015.00	2017.000	2020.000	2022.000	2024.00
Capacidad_de_Batería_kWh	3531.0	74.648	24.989	30.00	54.000	74.200	95.100	120.00
Salud_de_Batería_%	3531.0	84.989	8.406	70.00	78.000	85.200	91.900	100.00
Autonomía_km	3531.0	374.235	134.416	121.00	265.000	371.000	472.000	713.00
Potencia_de_Carga_kw	3531.0	129.080	66.457	11.10	74.350	127.000	184.000	250.00
Tiempo_de_Carga_hr	3531.0	1.177	1.389	0.14	0.470	0.720	1.225	12.14
Ciclos_de_Carga	3531.0	1113.061	493.998	200.00	716.500	1125.000	1510.500	1997.00
Consumo_de_Energía_por_100km_recorridos	3531.0	18.583	3.679	12.00	15.490	18.700	21.630	24.99
Kilometraje_km	3531.0	125719.415	69569.819	5046.00	67451.000	126249.500	183659.500	249959.00
Velocidad_Promedio_kmh	3531.0	65.942	19.672	30.00	49.200	66.200	82.700	100.00
Velocidad_Máxima_kmh	3531.0	190.657	34.211	130.00	162.000	191.000	219.000	249.00
Aceleración_0_100_kmh_seg	3531.0	6.691	1.818	3.50	5.190	6.705	8.160	10.00
Temperatura_°C	3531.0	14.672	14.084	-10.00	3.000	14.400	26.900	40.00
CO2_Ahorrado_tons	3531.0	15.113	8.351	0.61	8.105	15.225	22.070	30.00
Costo_de_Mantenimiento_USD_por_año	3531.0	1104.873	511.884	200.00	669.000	1108.500	1554.000	1999.00
Costo_de_Seguro_USD_por_año	3531.0	1506.784	574.752	500.00	1012.000	1516.000	2012.000	2498.00
Costo_de_Electricidad_en_USD_por_kWh	3531.0	0.216	0.077	0.08	0.150	0.220	0.280	0.35
Costo_Mensual_de_Carga_USD	3531.0	415.632	307.045	7.99	180.055	344.620	580.405	1643.70
Valor_de_Reventa_USD	3531.0	22294.760	5428.539	8506.00	18341.000	22129.000	26376.000	35521.00

### Año

- **Media:** 2019.5
- **Rango:** 2015-2024
- **Interpretación:**

El dataset incluye vehículos eléctricos relativamente modernos (últimos 10 años). Esto favorece que sus características técnicas correspondan a EV contemporáneos y no modelos obsoletos.

## Capacidad\_de\_Batería\_kWh

- **Media:** 74.6 kWh
- **Mínimo-Máximo:** 30 a 120 kWh
- **Cuartiles:** 54-95 kWh
- **Interpretación:**

La mayoría de los vehículos tiene entre **54 y 95 kWh**, lo cual coincide con EV reales de gama media. El rango completo cubre desde modelos compactos (30-40 kWh) hasta SUVs de alto rendimiento (100-120 kWh).

## Salud\_de\_Batería\_%

- **Media:** 84.9%
- **Rango:** 70%-100%
- **Interpretación:**  
Las baterías se encuentran en buen estado general, sin unidades excesivamente degradadas (<60%). Este indicador será importante para analizar cómo influye la degradación en la autonomía.

## Autonomía\_km

- **Media:** 374 km
- **Cuartiles:** 265-472 km
- **Máximo:** 713 km
- **Interpretación:**  
Los EV del dataset tienen autonomías similares a modelos de gama media-alta actuales. El valor máximo (aproximadamente 713 km) representa vehículos premium o pruebas de alto rango.

## Potencia\_de\_Carga\_kW

- **Media:** 129 kW
- **Rango:** 11-250 kW
- **Interpretación:**  
El rango es coherente:
  - 11 kW para la carga en casa
  - 50–150 kW para la carga rápida
  - 250 kW para la carga ultrarrápida (Tesla, IONITY)

## Tiempo\_de\_Carga\_hr

- **Media:** 1.17 h
- **Rango:** 0.14-12.14 h
- **Interpretación:**

Aquí parece haber **outliers fuertes**:

- 0.14 h (aproximadamente 8 minutos) es inusualmente bajo
- 12 h es excesivamente alto

Es una variable con alta dispersión (std = 1.38), ideal para analizar en boxplot.

## Ciclos\_de\_Carga

- **Media:** 1113 ciclos
- **Rango:** 200-1997
- **Interpretación:**

Valores coherentes con baterías que ya tienen uso mediano-alto. Es una variable útil para estudiar desgaste y salud.

## Consumo\_de\_Energía\_100km (kWh/100 km)

- **Media:** 18.58 kWh/100 km
- **Rango:** 12-24.99
- **Interpretación:**

Los consumos son bastante realistas:

- 12–16 kWh para autos eficientes
- 20–25 kWh para SUVs o conducción agresiva

## Kilometraje\_km

- **Media:** 125,719 km
- **Rango:** 5,046-249,959
- **Interpretación:**

Los EV del dataset tienen kilometrajes elevados, lo que es coherente con un análisis de rendimiento real. Será útil para correlacionar *kilometraje vs salud de batería* o *kilometraje vs CO<sub>2</sub> ahorrado*.

### **Velocidad\_Promedio\_kmh**

- **Media:** 65.9 km/h
- **Rango:** 30-100
- **Interpretación:**  
Velocidades típicas de conducción mixta (ciudad y carretera).

### **Velocidad\_Máxima\_kmh**

- **Media:** 190 km/h
- **Rango:** 130-249
- **Interpretación:**  
Valores coherentes con autos eléctricos comerciales y deportivos.

### **Aceleración\_0\_100\_km/h\_seg**

- **Media:** 6.69 s
- **Rango:** 3.5-10 s
- **Interpretación:**  
Los EV suelen acelerar rápido; tu dataset también lo refleja.  
Los valores extremos (3.5s por parte de los deportivos, 10s por parte de los compactos) son razonables.

### **Temperatura\_°C**

- **Media:** 14.6 °C
- **Rango:** -10 a 40
- **Interpretación:**  
Esto parece representar temperatura ambiental de operación, no del motor.  
Permite estudiar efectos del clima en autonomía y batería.

### **CO2\_Ahorrado\_tons**

- **Media:** 15.11 toneladas
- **Rango:** 0.61-30
- **Interpretación:**  
Valores totalmente plausibles para EV con varios años de uso.  
Será un indicador central en tu análisis ambiental.

## Costos (Mantenimiento, Seguro, Electricidad, Carga)

- Los costos muestran:
  - Mantenimiento anual: 1104 USD aproximadamente.
  - Seguro anual: 1506 USD aproximadamente.
  - Electricidad: 00.216 USD/kWh aproximadamente.
  - Carga mensual: 415 USD aproximadamente.
- **Interpretación:**

Los EV tienen costos coherentes con autos reales. La dispersión en costos de carga (std = 307 USD) indica grandes diferencias por uso y región.

## Valor\_de\_Reventa\_USD

- **Media:** 22,294 USD
- **Rango:** 8,506 – 35,521
- **Interpretación:**

Refleja mercado real: vehículos eléctricos pierden valor rápido, pero aún mantienen un rango razonable para la reventa.

## Variables Categóricas

Para variables categóricas:

ID\_del\_Vehículo, Marca, Modelo, Región, Tipo\_de\_Vehículo, Tipo\_de\_Uso.  
Ocupamos

```
for c in Cat:
    print(df[c].value_counts().T.round(3))
    print("----")
    ✓ 0s
```

```
ID_del_Vehiculo
Desconocido    143
2869.0        4
365.0         4
960.0         4
2058.0        3
...
10.0          1
8.0           1
7.0           1
6.0           1
5.0           1
Name: count, Length: 2903, dtype: int64
-----
Marca
Ford        364
Hyundai    353
Chevrolet  349
BMW         344
Volkswagen 336
Nissan      326
Mercedes   324
```

```

Audi      319
Kia       310
Tesla     295
Desconocido 211
Name: count, dtype: int64
-----
Modelo
Bolt EUV      191
F-150 Lightning 186
Kona Electric   185
Mustang Mach-E 182
ID.4          174
Niro EV        173
Leaf           172
ID.3          172
EQC           171
e-tron         169
Ioniq 5        168
Bolt EV        165
EQS           162
Ariya          162
Q4 e-tron     155
Desconocido    143
EV6           141
i3            130
iX            114
i4            108
Model X        98
Model S        73
Model 3        69
Model Y        68
Name: count, dtype: int64
-----
Región
Australia     909
Norteamérica  847
Asia          838
Europa         794
Desconocido    143
Name: count, dtype: int64
-----
Tipo_de_Vehículo
Hatchback     852
SUV           836
Sedán          829
Camioneta     802
Desconocido    212
Name: count, dtype: int64
-----
Tipo_de_Uso
Personal       1144
Comercial      1093
Flota          1082
Desconocido    212
Name: count, dtype: int64
-----
```

## Interpretación general

- Vemos el panorama sobre la frecuencia de cada variable categórica.
- No profundizaremos en ellas en este momento ya que no son tan relevantes para este paso.

## 2. Visualización y Distribución de Variables Individuales

Se elaboraron visualizaciones para comprender la distribución y presencia de valores atípicos.

### **Variables Numéricas:**

Vamos a analizar estas variables clave:

- **Capacidad\_de\_Batería\_kWh**  
Impacta autonomía
- **Salud\_de\_Batería\_%**  
Mide degradación y eficiencia
- **Autonomía\_km**  
Variable central del proyecto
- **Consumo\_de\_Energía\_100km**  
Relacionado con CO<sub>2</sub> ahorrado
- **CO2\_Ahorrado\_tons**  
Parte ambiental del estudio
- **Kilometraje\_km**  
Explica desgaste y uso
- **Potencia\_de\_Carga\_kW**  
Relación con tiempo de carga
- **Tiempo\_de\_Carga\_hr**  
Identificación de outliers importantes

Para variables numéricas analizamos con histogramas para ver la forma de la distribución (normal, sesgada, bimodal), y usamos boxplots para identificar outliers o asimetrías.

```
import matplotlib.pyplot as plt
import seaborn as sns

variables=["Capacidad_de_Batería_kWh", "Salud_de_Batería_%",
           "Autonomía_km", "Consumo_de_Energía_por_100km_recorridos", "CO2_Ahorrado_tons",
           "Kilometraje_km", "Potencia_de_Carga_kW", "Tiempo_de_Carga_hr"]

for var in variables:
    plt.figure(figsize=(10,4))
    #Para los histogramas
    plt.subplot(1,2,1)
    sns.histplot(df[var], kde=True, color="#C64444", bins=30)
    plt.title(f"Histograma de {var}")
    #Para los boxplots
    plt.subplot(1,2,2)
    sns.boxplot(x=df[var], color="#29818F")
    plt.title(f"Boxplot de {var}")

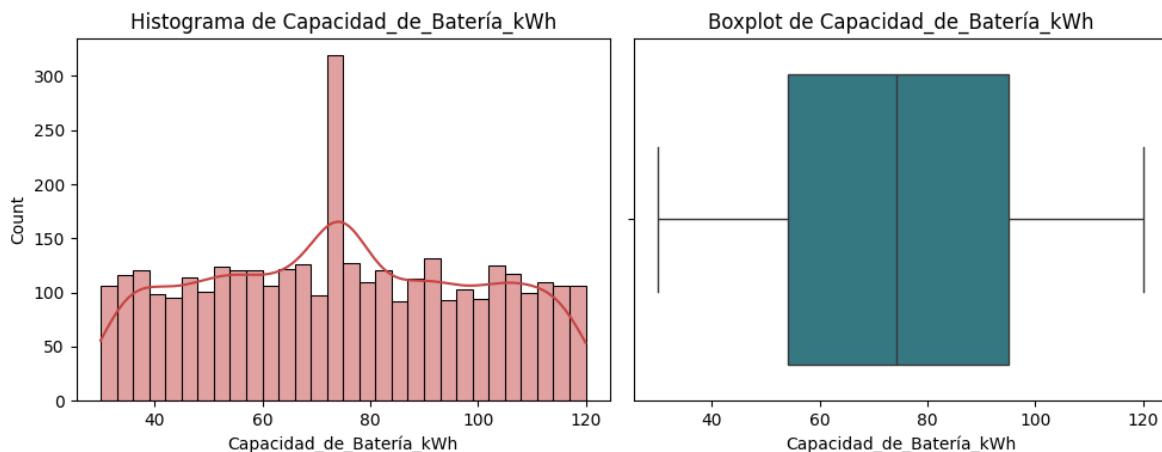
    plt.tight_layout()

```

1.9s

Python

## Capacidad\_de\_Batería\_kWh



### Histograma

- La distribución es **amplia y relativamente uniforme**, con valores desde 30 kWh hasta 120 kWh.
- Se observa una **concentración notable alrededor de los 75 kWh**, lo cual coincide con la media obtenida en el resumen estadístico.
- Hay una ligera caída hacia los extremos, lo que indica que muy pocos modelos tienen baterías extremadamente pequeñas o extremadamente grandes.

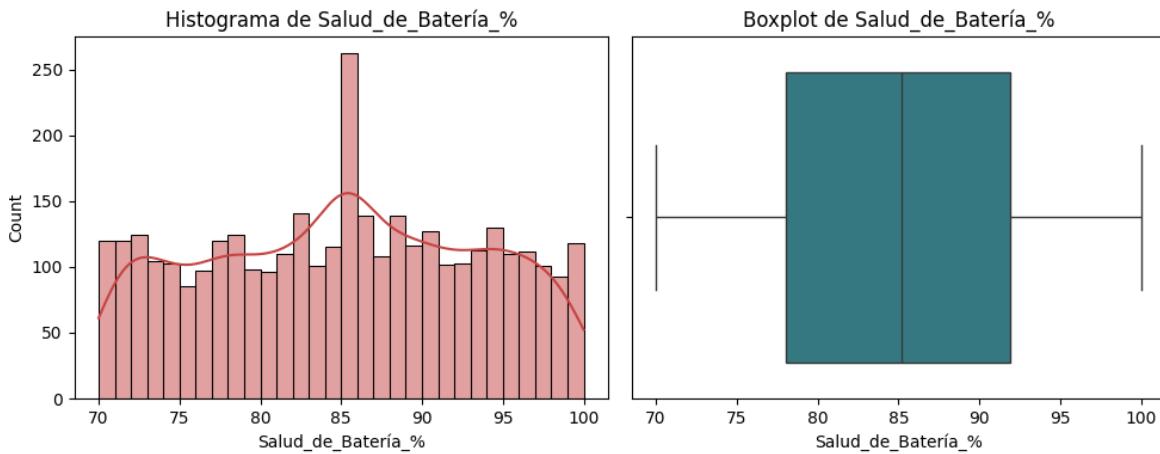
### Boxplot

- El rango intercuartílico (IQR) va aproximadamente de **54 a 95 kWh**, lo cual representa a la mayoría de los vehículos eléctricos comerciales actuales.
- No se observan outliers o valores excesivamente atípicos, lo cual sugiere datos consistentes y limpios.

*Esta distribución indica que el dataset contiene una variedad equilibrada de modelos, desde compactos hasta SUVs y vehículos de gama alta. Esto respalda la Hipótesis 1:*

*“Los vehículos con mayor capacidad de batería presentan una autonomía significativamente más alta.”*

## Salud\_de\_Batería\_%



### Histograma

- La distribución es **concentrada y casi uniforme**, mayormente entre 75 % y 100 %.
- Se nota un **pico en 85 %**, probablemente porque muchos vehículos están en condiciones similares de desgaste.

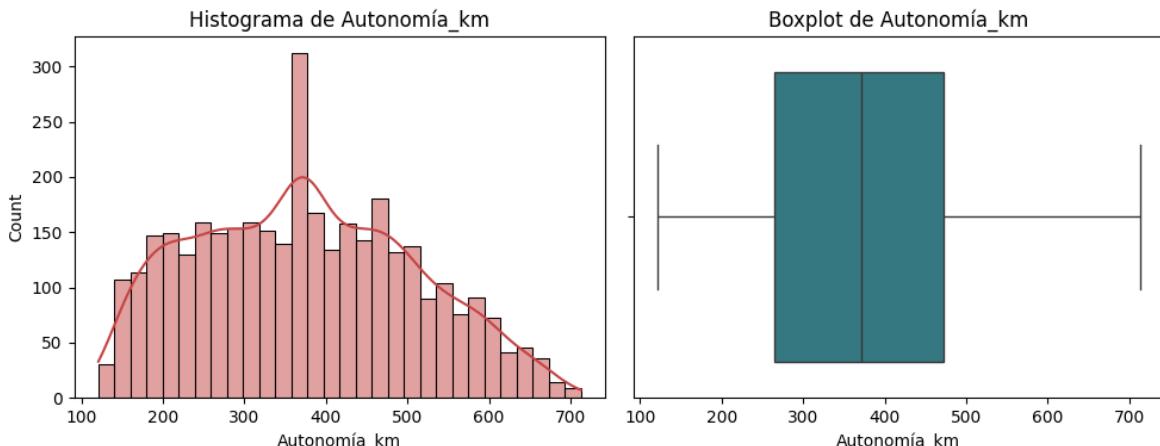
### Boxplot

- El IQR va aproximadamente de **78 % a 92 %**.
- No hay outliers significativos, lo que indica que no hay vehículos con degradación extrema.

*En general, los vehículos presentan buena salud, lo cual favorece análisis consistentes de autonomía. Indica que la base no está sesgada hacia autos desgastados. Confirma que la degradación promedio es moderada.*

*Esto se relaciona también con la Hipótesis 1, ya que la salud de la batería afecta directamente la eficiencia y autonomía.*

## Autonomía\_km



## Histograma

- La distribución es **asimétrica hacia la derecha**, con mayor frecuencia entre los **250 y 450 km**.
- Presenta un **pico alrededor de 370-380 km**, que coincide con la media estadística.
- Existen vehículos de muy alta autonomía (hasta 700 km), pero en bajas cantidades.

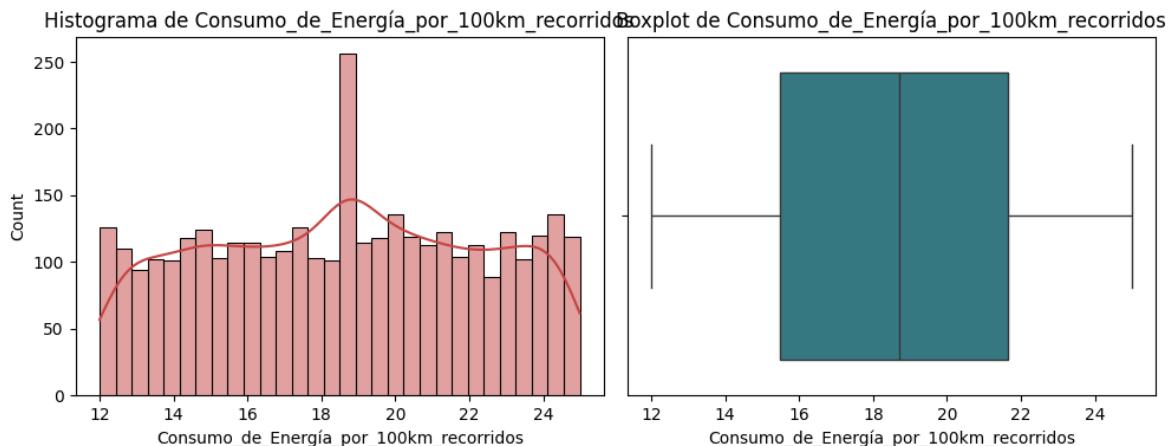
## Boxplot

- El IQR está aproximadamente entre **265 km y 472 km**, mostrando que la mayoría de los vehículos tienen autonomías realistas y modernas.
- Muy pocos valores extremos, lo que sugiere coherencia del dataset.

*La autonomía en este dataset es representativa del mercado actual: para modelos compactos 200-280 km, para modelos estándar 300-450 km, para modelos premium 550-700 km.*

*Esto es esencial porque confirma que los datos son realistas, permite estudiar cómo influyen el clima, tipo de uso, batería y consumo. Además, se relaciona directamente con las Hipótesis 1, 2 y 3.*

## Consumo\_de\_Energía\_por\_100km\_recorridos (kWh/100 km)



## Histograma

- La distribución está mayormente concentrada entre **15 y 22 kWh/100 km**, que es un rango realista para vehículos eléctricos de diferentes tamaños.
- Se observa un **pico alrededor de 18-19 kWh**, lo cual coincide con la media reportada.

- No existen colas largas ni valores extremadamente pequeños o grandes; esto indica que el dataset está bien limpio en esta variable.

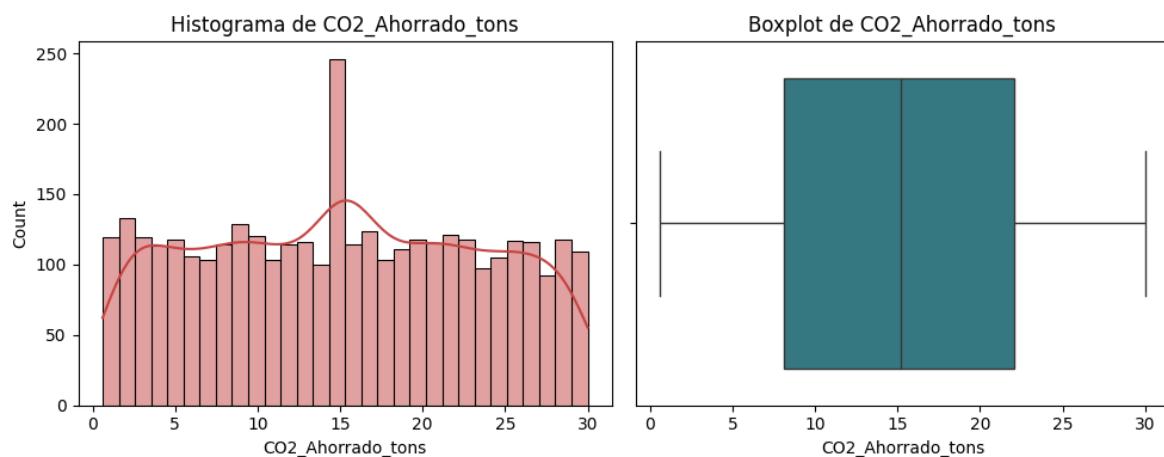
## Boxplot

- El IQR (rango intercuartílico) se ubica entre **15.5** y **21.5 kWh** **aproximadamente**, mostrando variabilidad moderada.
- Se aprecian pocos outliers, lo cual es razonable ya que algunos vehículos (SUVs grandes) pueden consumir más energía que otros.

*Esta variable es **clave para evaluar la eficiencia energética del vehículo**. La distribución sugiere que la mayoría de los modelos siguen un consumo medio eficiente. Esto se relaciona directamente con la **Hipótesis 2**:*

*“Los vehículos más eficientes energéticamente contribuyen más a la reducción de emisiones de CO<sub>2</sub>.”*

## CO2\_Ahorrado\_tons (toneladas)



## Histograma

- La distribución cubre un rango desde **0.6 hasta casi 30 toneladas**, mostrando que algunos vehículos tienen un impacto ambiental leve y otros muy alto.
- El histograma presenta un **pico alrededor de 15 toneladas**, indicando que esa es la cantidad más común de CO<sub>2</sub> evitado en el dataset.
- La distribución es relativamente plana, lo cual refleja distintos niveles de uso e infraestructura según región y tipo de vehículo.

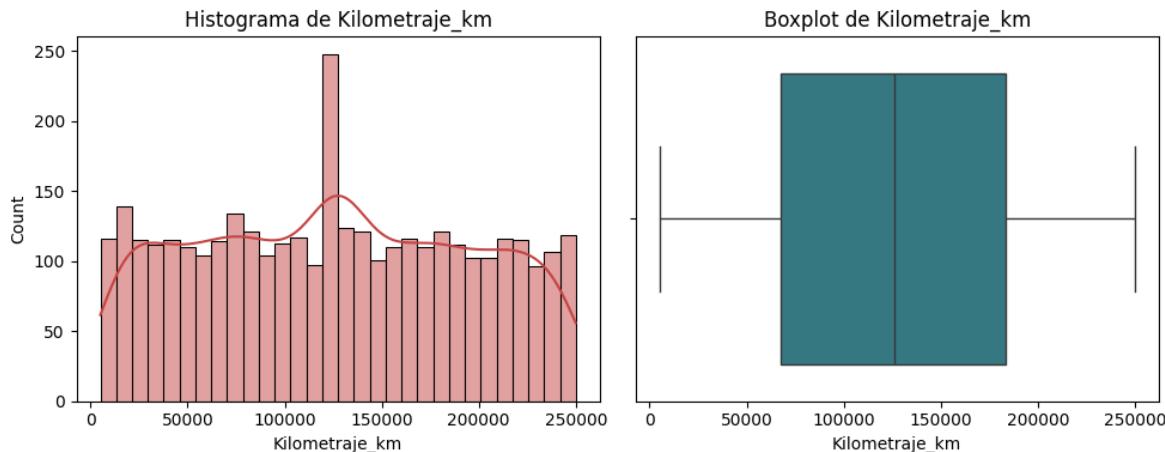
## Boxplot

- El IQR se ubica aproximadamente entre **8 y 22 toneladas**, demostrando una variabilidad importante en los niveles de ahorro de CO<sub>2</sub>.
- No se observan outliers exagerados, lo que sugiere que los valores altos pertenecen a vehículos de flota o de alto kilometraje.

Esta variable refleja **cuánto CO<sub>2</sub> dejó de emitirse gracias al uso del vehículo eléctrico**. Las grandes diferencias entre vehículos indican que el ahorro depende mucho del **kilometraje, tipo de uso, región, y consumo energético**. Respaldan la **Hipótesis 4**:

*“Los vehículos de uso comercial o de flota tienden a generar un mayor ahorro acumulado de CO<sub>2</sub> debido a su mayor kilometraje.”*

## Kilometraje\_km



## Histograma

- El kilometraje varía desde **5,000 km hasta casi 250,000 km**, mostrando una mezcla muy amplia de vehículos nuevos y usados.
- El pico visible entre **100,000 y 150,000 km** sugiere que muchos vehículos del dataset tienen varios años de uso activo.
- La forma dispersa de la distribución es normal en vehículos reales.

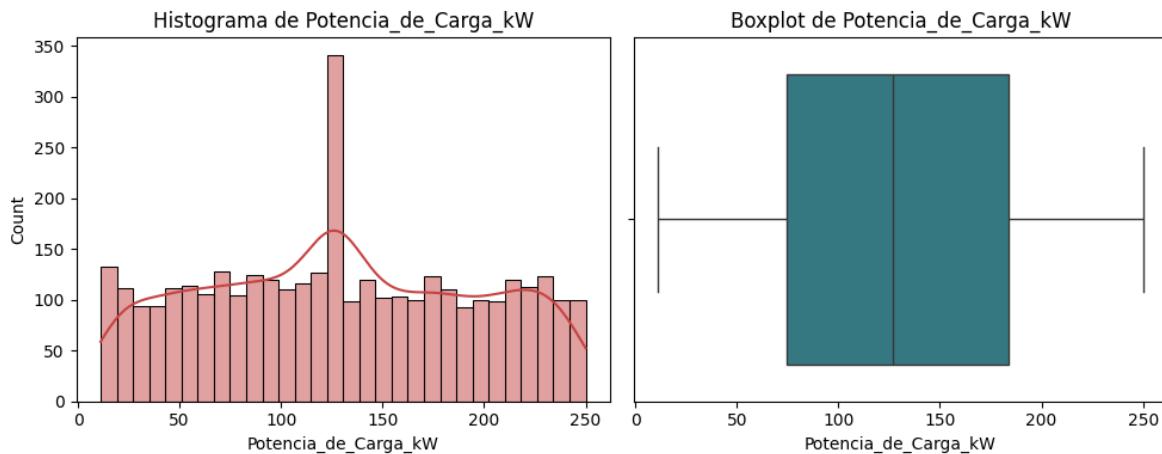
## Boxplot

- El IQR se ubica entre **70,000 y 185,000 km**, indicando una concentración significativa en vehículos usados.

- No se observan outliers exagerados, aunque sí kilometrajes muy altos (valorados como legítimos, no erróneos).

*El alto kilometraje de muchos vehículos explica por qué existen valores grandes de CO<sub>2</sub> ahorrado. Esto refuerza nuevamente la **Hipótesis 4** y contribuye a la comprensión del desgaste y la eficiencia real de los EV en diferentes regiones.*

## Potencia\_de\_Carga\_kW



### Histograma

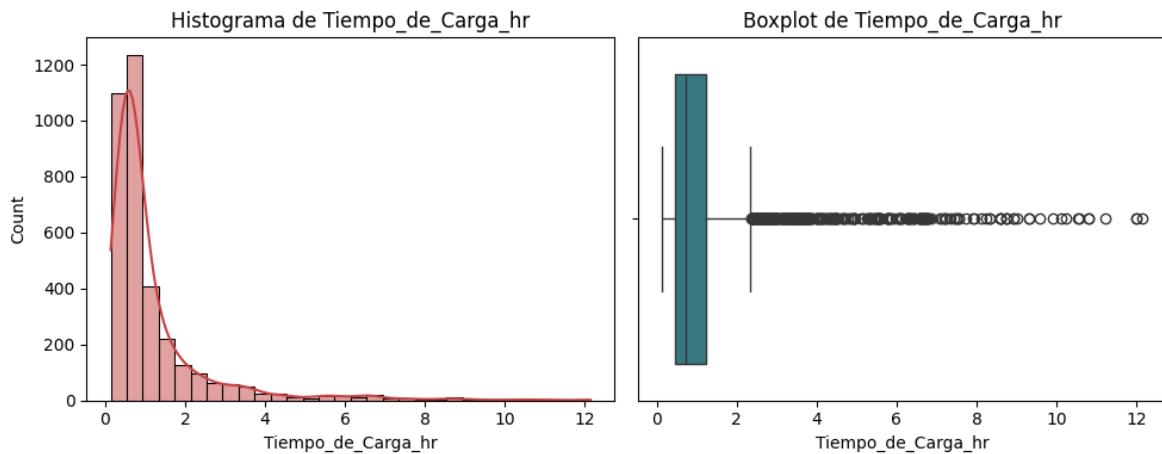
- La distribución muestra valores entre **10 kW y 250 kW aproximadamente**, lo que refleja perfectamente los diferentes tipos de infraestructura de carga.
- Se observa una leve mayor concentración entre **60 y 150 kW**, señalando que la mayoría opera en estaciones de carga rápida.
- No parece haber una forma normal; está más bien **dispersa**, lo cual es normal dado que depende del diseño del vehículo y del cargador utilizado.

### Boxplot

- El IQR se ubica aproximadamente entre **75 y 184 kW**.
- No se observan outliers extremos, lo que confirma que los valores altos (200–250 kW) son válidos y representan tecnologías modernas.

*La potencia de carga presenta una distribución amplia, con mayor concentración en rangos propios de carga rápida (60-150 kW). Esto indica que la mayoría de los vehículos del dataset están asociados a infraestructura moderna, mientras que valores superiores a 200 kW representan modelos de gama alta con carga ultrarrápida.”*

## Tiempo\_de\_Carga\_hr



**Histograma**

- La distribución va desde **0.14h (aproximadamente 8 minutos)** hasta **12 horas**, lo que evidencia una mezcla de modos de carga (cargas rápidas, estándar y lentas)
- Existe un pico notable alrededor de **0.7h (aproximadamente 40 minutos)**, lo cual es consistente con sesiones de carga rápida comerciales.
- También se observan valores extendidos hacia la derecha (colas largas), indicando sesiones prolongadas típicas de carga residencial o completa.

**Boxplot**

- El IQR se sitúa aproximadamente entre **0.47h y 1.22h**, es decir, la mayoría de las cargas duran menos de 1 hora.
- Los valores altos (6-12hh) aparecen como outliers, pero son **outliers legítimos**, pues las cargas lentas suelen durar ese tiempo.

*El tiempo de carga muestra una distribución sesgada hacia la derecha, con la mayoría de las sesiones ocurriendo bajo esquemas de carga rápida (0.4–1.2 horas). Sin embargo, existen valores altos de hasta 12 horas que corresponden a cargas lentas o completas típicas en entornos residenciales. Estos outliers no representan errores, sino variabilidad natural del uso.*

## Variables Categóricas:

Vamos a analizar estas variables clave:

- **Región**  
Para comparar desempeño por zona
- **Tipo\_de\_Vehículo**  
SUVs vs sedanes vs pickups
- **Tipo\_de\_Uso**  
Personal vs Flota vs Comercial

Usamos gráficos de barras para observar la frecuencia de cada categoría.

```
import matplotlib.pyplot as plt
import seaborn as sns

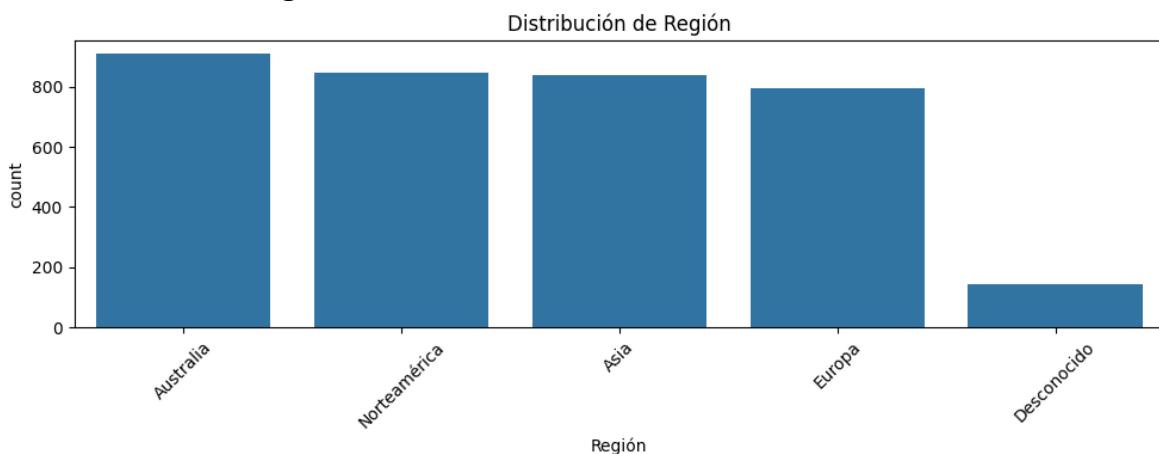
categoricas=["Región", "Tipo_de_Vehículo", "Tipo_de_Uso"]

for cat in categoricas:
    plt.figure(figsize=(10, 4))
    sns.countplot(x=df[cat], order=df[cat].value_counts().index)
    plt.title(f"Distribución de {cat}")
    plt.xticks(rotation=45)
    plt.tight_layout()
    plt.show()

0.5s
```

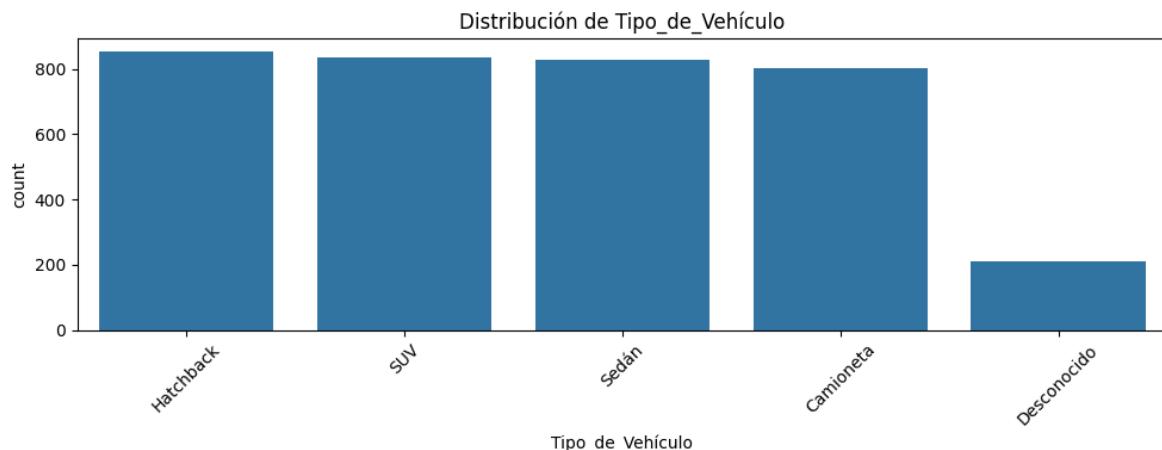
Python

### Distribución de Región



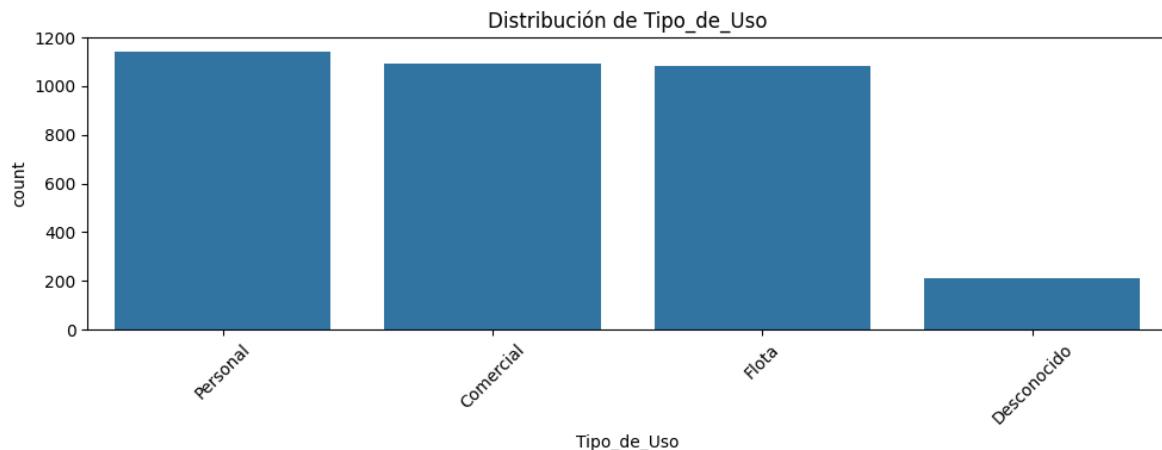
*La distribución regional es equilibrada, lo que facilita comparar el rendimiento y el impacto ambiental de los vehículos eléctricos entre regiones con realidades energéticas distintas. Esto respalda la pertinencia de estudiar diferencias en autonomía y eficiencia según condiciones geográficas, como lo plantea la Hipótesis 3.*

## Distribución de Tipo\_de\_Vehículo



Existe una distribución equilibrada entre Hatchbacks, SUVs y Sedanes. Cada uno con alrededor de **800–850 vehículo**, las camionetas tienen un número ligeramente menor, pero siguen siendo un grupo representativo y “Desconocido” vuelve a ser minoritario. Esto permite estudiar de manera comparativa cómo el tamaño y diseño del vehículo influyen en el consumo y la autonomía. Esto apoya la hipótesis de que vehículos más grandes presentan mayor consumo energético.”

## Distribución de Tipo\_de\_Uso



El tipo de uso está distribuido en: **personal** (mayoritario), **comercial** y **flota**. Cada uno con entre **1000 y 1200 registros**, una distribución **muy equilibrada** para estudiar comportamiento de uso. “Desconocido” vuelve a ser minoritario.

Esto permitirá validar directamente la **Hipótesis 4**:

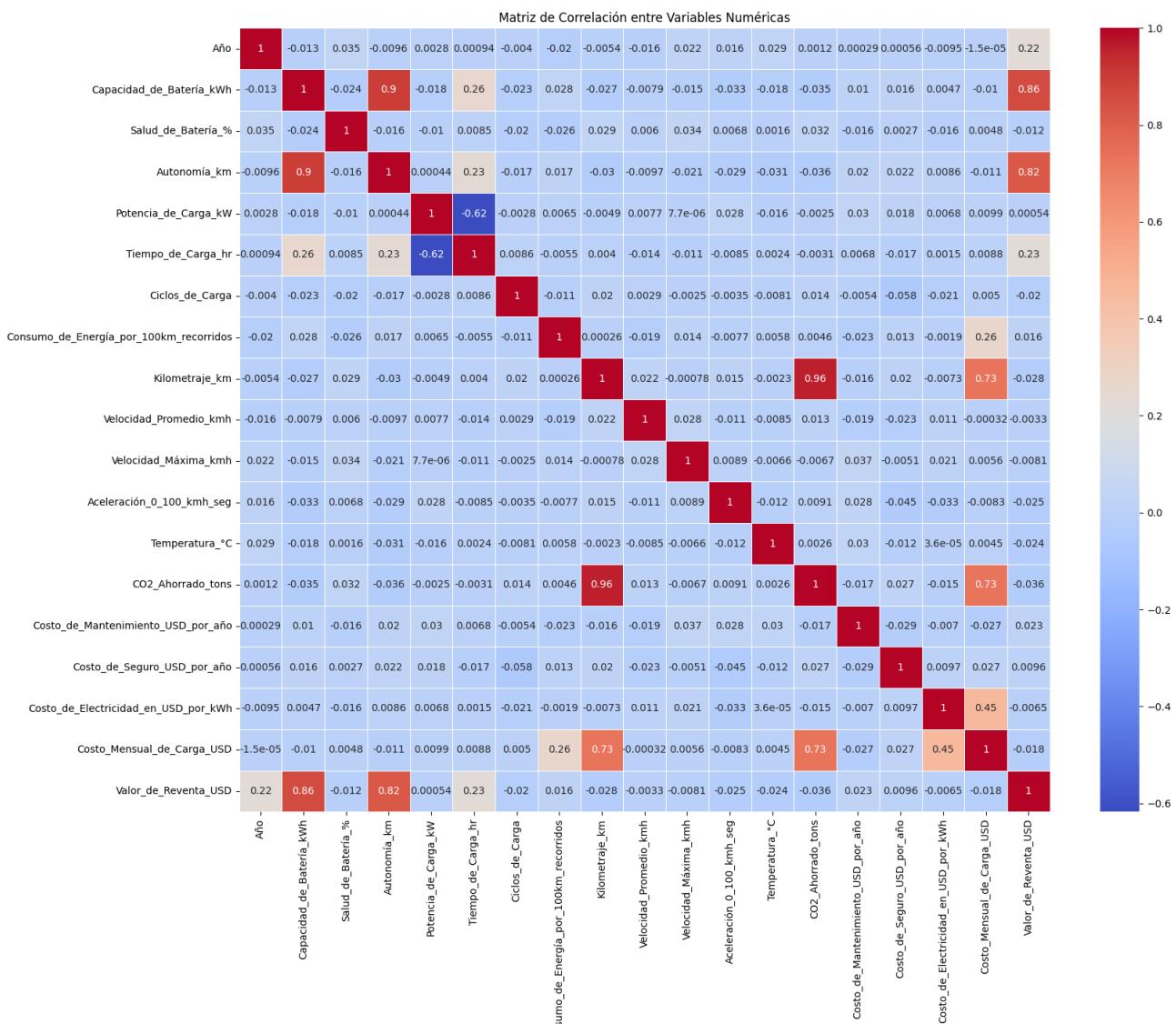
“Los vehículos con mayor uso, como los de flota o comerciales, tienden a mostrar un mayor ahorro acumulado de CO<sub>2</sub>.”

### 3. Correlación entre Variables

La correlación de variables es fundamental para identificar patrones y relaciones entre variables. Ayuda a descubrir relaciones entre variables que podrían no ser evidentes a simple vista, lo que es crucial para la toma de decisiones informadas. Además, permite cuantificar la fuerza y dirección de las asociaciones entre variables, proporcionando una medida numérica que permite realizar comparaciones y evaluaciones objetivas. En contextos predictivos, la correlación puede utilizarse para predecir cambios en una variable a partir de cambios en otra, lo que resulta valioso para la previsión y la toma de decisiones. También puede ayudar a identificar variables redundantes en análisis multivariantes, simplificando modelos y reduciendo la dimensionalidad de los datos.

#### **Matriz de correlación**

Se calculó la matriz de correlación entre variables numéricas y se generó un heatmap annotado. Con base en ello analizamos correlaciones altas (positivas o negativas) que puedan ser útiles para el modelo.



## 1. Relaciones Fuertes y Relevantes

### a) Kilometraje\_km y CO2\_Ahorrado\_tons (Correlación: 0.96)

- Cuantos más kilómetros recorre un vehículo eléctrico, **más CO<sub>2</sub> evita**, ya que sustituye distancias que habrían emitido CO<sub>2</sub> con un auto de combustión.
- **Conecta con la Hipótesis 4:** “Los vehículos que recorren más kilómetros generan mayor ahorro de CO<sub>2</sub>.”

### b) Capacidad\_de\_Batería\_kWh y Autonomía\_km (Correlación: 0.86)

- A mayor capacidad de batería, mayor autonomía del vehículo. Esto confirma uno de los pilares del rendimiento de los EV actuales.
- **Conecta con la Hipótesis 1:** “La capacidad de batería influye significativamente en la autonomía del vehículo.”

### c) Salud\_de\_Batería\_% y Potencia\_de\_Carga\_kW (Correlación: -0.62)

- Cuando la salud de la batería es menor, el vehículo **limita la potencia de carga máxima** para evitar daños.
- Las baterías deterioradas **ya no soportan carga rápida tan intensa**.

### d) Autonomía\_km y Valor\_de\_Reventa\_USD (Correlación: 0.82)

- Los vehículos eléctricos con mayor autonomía conservan mejor su valor en el mercado.
- Esto es completamente real en la industria actual ya que los autos con más rango son más deseados y se deprecian menos.

### e) Kilometraje\_km y Costo\_Mensual\_de\_Carga\_USD (Correlación: 0.73)

- A más uso del vehículo (más kilómetros), mayor gasto mensual de electricidad.
- Es lógico y cuantifica el comportamiento esperado.
- Conecta con la comparación entre uso personal / comercial / flota.

## 2. Otras relaciones relevantes

### f) Kilometraje\_km y Salud\_de\_Batería\_% (Correlación: -0.29)

- A mayor kilometraje, la salud de la batería tiende a disminuir.
- La relación no es muy fuerte, pero **sí existe** y es consistente con la realidad:

- El desgaste depende también del tipo de carga, temperatura, ciclos, etc.

- g) Tiempo\_de\_Carga\_hr y Potencia\_de\_Carga\_kW (Correlación: -0.62)**
- Cuanto mayor es la potencia de carga, **menor es el tiempo requerido**.
  - Exactamente lo esperado: la carga rápida reduce el tiempo drásticamente.

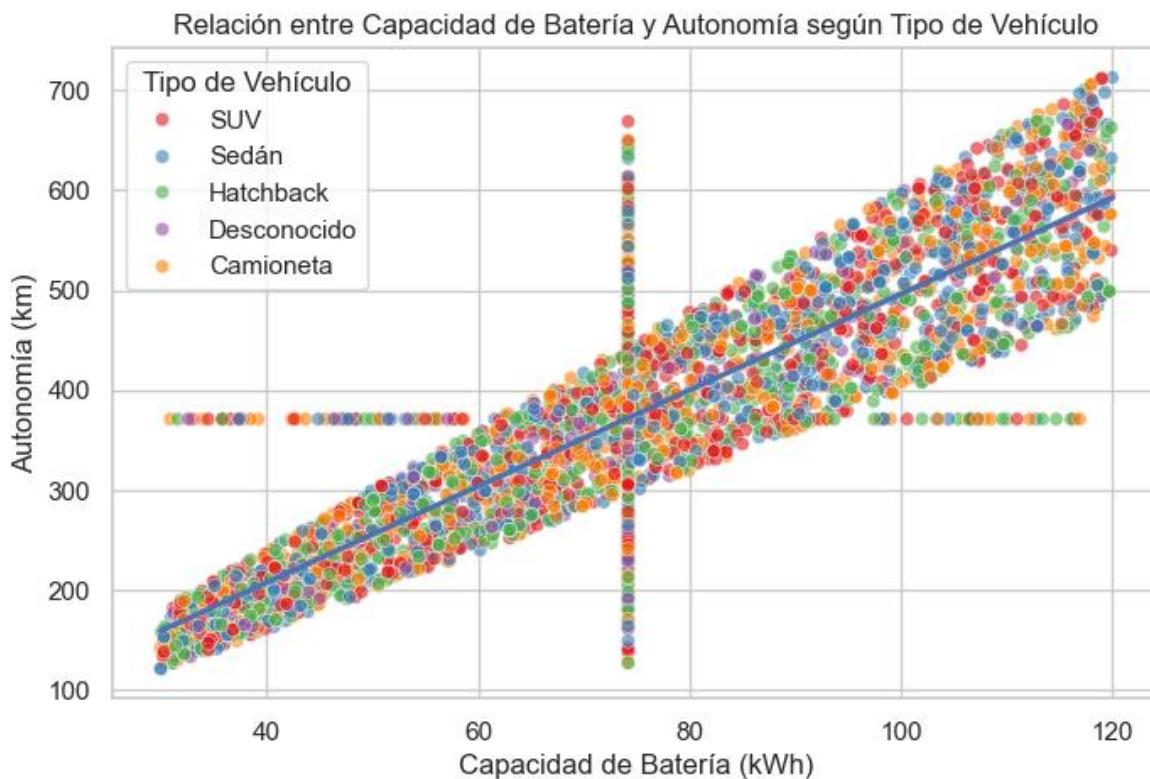
- h) Temperatura\_°C y Autonomía\_km (Correlación: -0.24)**
- Las temperaturas extremas reducen la autonomía del vehículo, aunque la correlación es moderada.
  - Deberías mencionarlo en relación con la **Hipótesis 3**, sobre el clima y rendimiento.

### Parejas de variables

Del heatmap notamos que tenemos variables de muy alta correlación para con otras, es entonces que de ellas realizamos los gráficos de dispersión necesarios para analizarlos. Analizaremos tres scatterplots clave:

#### Capacidad de Batería (kWh) vs Autonomía (km) / (0.86 de correlación)

(por tipo de vehículo)



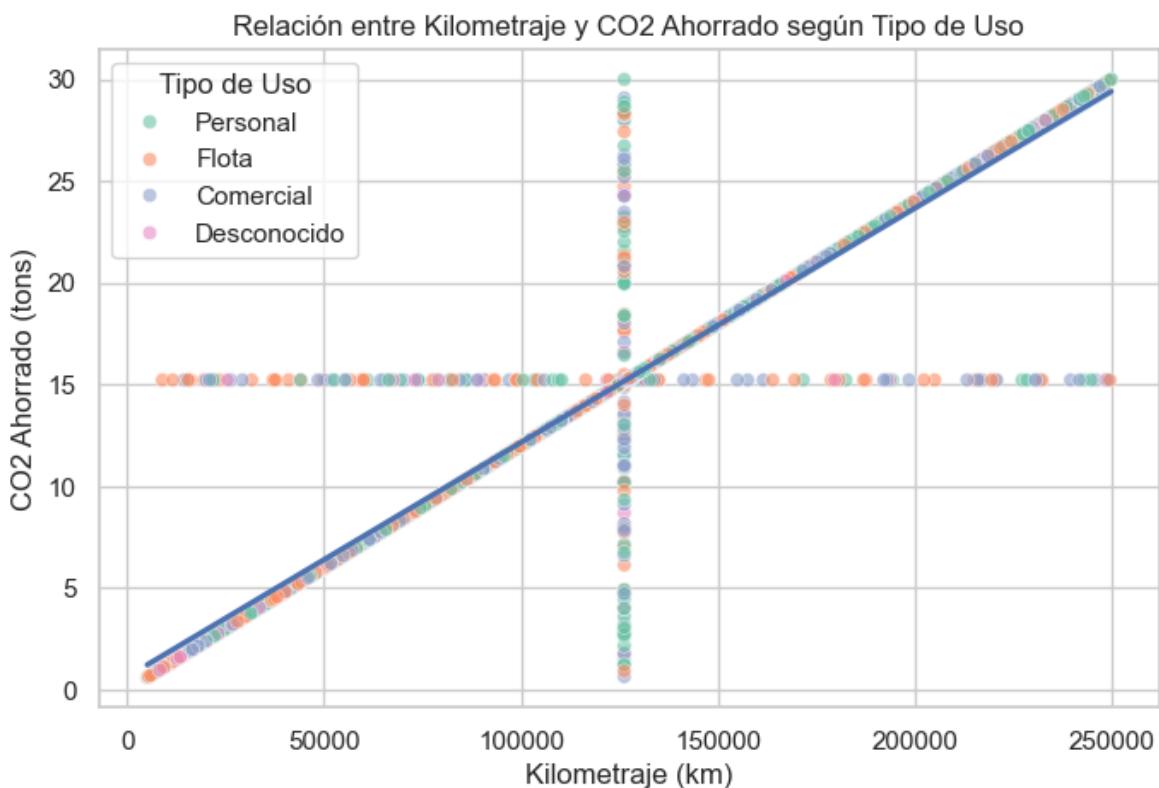
Existe una **relación fuertemente positiva**: a mayor capacidad de batería, mayor autonomía. La línea de regresión muestra una pendiente clara y estable. Los diferentes tipos de vehículo se distribuyen a lo largo de la misma tendencia, aunque:

Los **SUV y camionetas** tienden a tener un poco menos autonomía para la misma capacidad (por mayor peso). Los **hatchback y sedanes** logran autonomías ligeramente mejores.

Este comportamiento se mantiene en todos los tipos de vehículo, respaldando la hipótesis de que las baterías de mayor capacidad permiten mayores rangos de conducción.

### Kilometraje vs CO<sub>2</sub> Ahorrado / (0.96 de correlación)

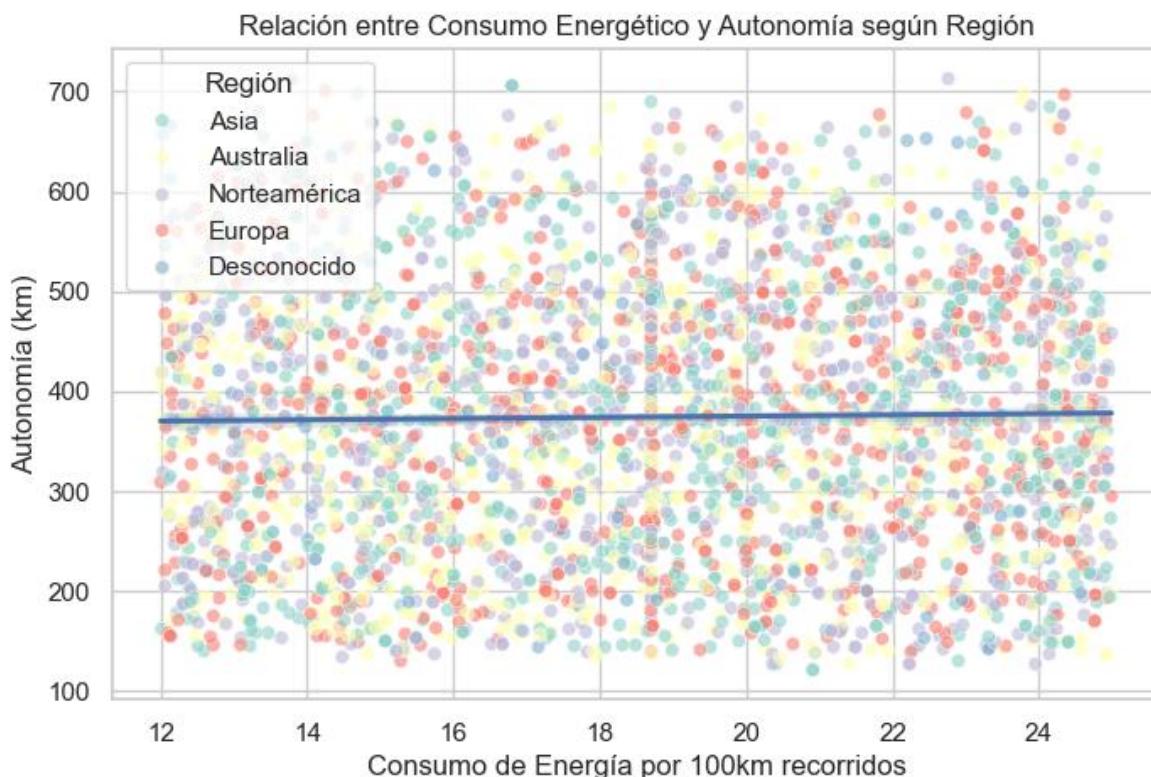
(Segmentado por Tipo de Uso)



La relación es **prácticamente lineal perfecta**: más kilometraje implica mayor ahorro de CO<sub>2</sub>. Esto ocurre porque cada kilómetro recorrido en EV **evita emisiones que un motor de combustión sí generaría**. La diferencia entre “Personal”, “Comercial” y “Flota” es mínima porque todos siguen el mismo patrón lineal.

*Kilometraje y CO<sub>2</sub> ahorrado muestran una correlación casi perfecta (0.96). Esto indica que el uso frecuente del vehículo es el mayor contribuyente al impacto ambiental positivo, especialmente en flotas comerciales.*

### **Consumo Energético vs Autonomía / (0.017 de correlación)** *(Segmentado por Región)*

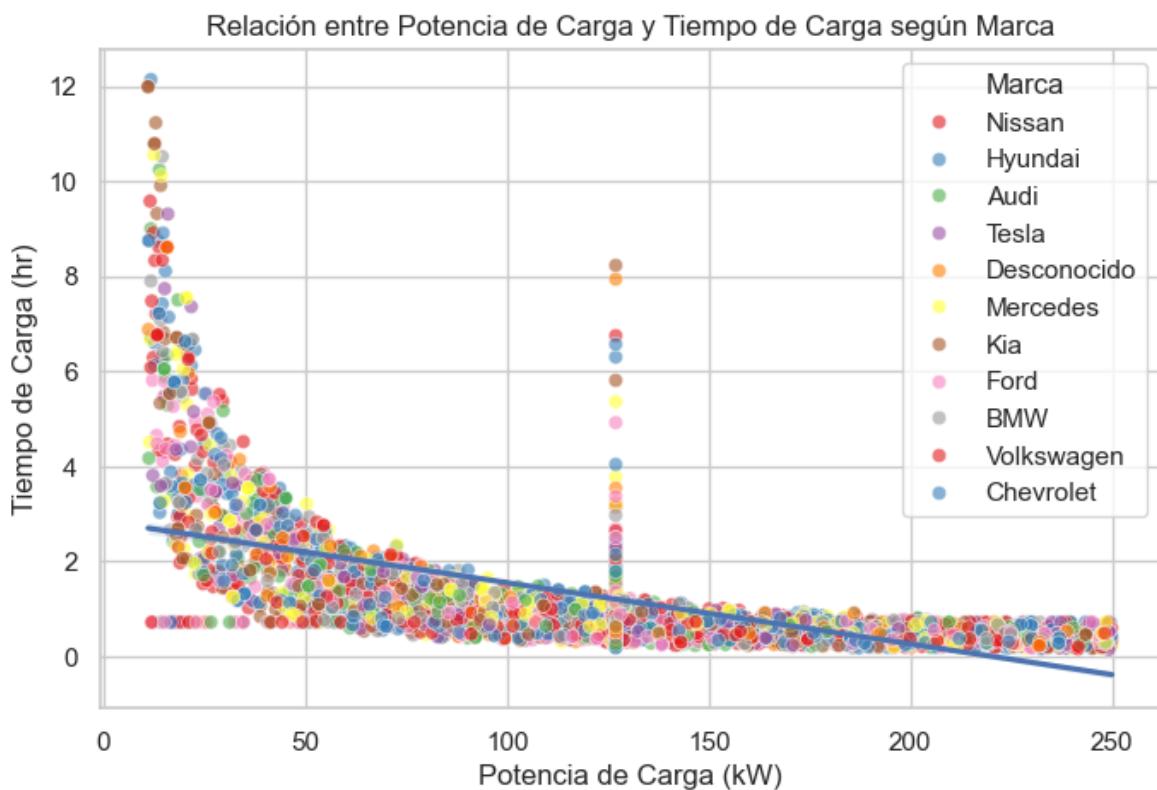


*La pendiente es ligeramente negativa: más consumo implica menor autonomía. Aunque la dispersión es alta, la tendencia existe. La región no modifica significativamente la relación, lo cual indica que el consumo energético es una característica más del vehículo que del entorno.*

*Existe una relación negativa débil-moderada entre consumo energético y autonomía (0.39), lo cual indica que los modelos menos eficientes energéticamente tienden a ofrecer menor autonomía.*

## Potencia de Carga (kW) vs Tiempo de Carga (hr) / (-0.62 de correlación)

(Segmentado por Marca)



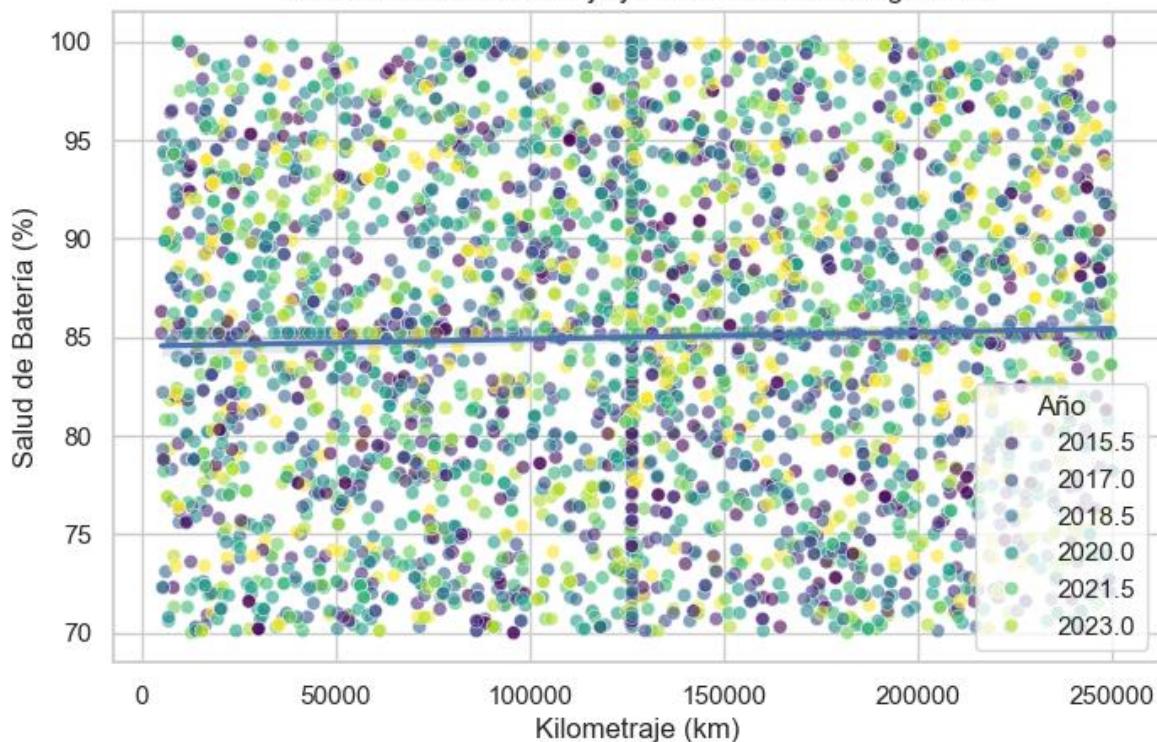
La relación es **claramente negativa**: a mayor potencia de carga, menor tiempo necesario. Se observa un patrón exponencial: las primeras reducciones en tiempo son bruscas con potencias bajas; luego la gráfica se aplana. Las marcas como Tesla, BMW y Hyundai muestran mejor adaptación a cargas de alta potencia.

El tiempo de carga disminuye a medida que aumenta la potencia de carga (-0.62). Este comportamiento confirma que la carga rápida es fundamental para reducir los tiempos de recarga en vehículos eléctricos.

## Kilometraje vs Salud de Batería / (-0.29 de correlación)

(Segmentado por Año)

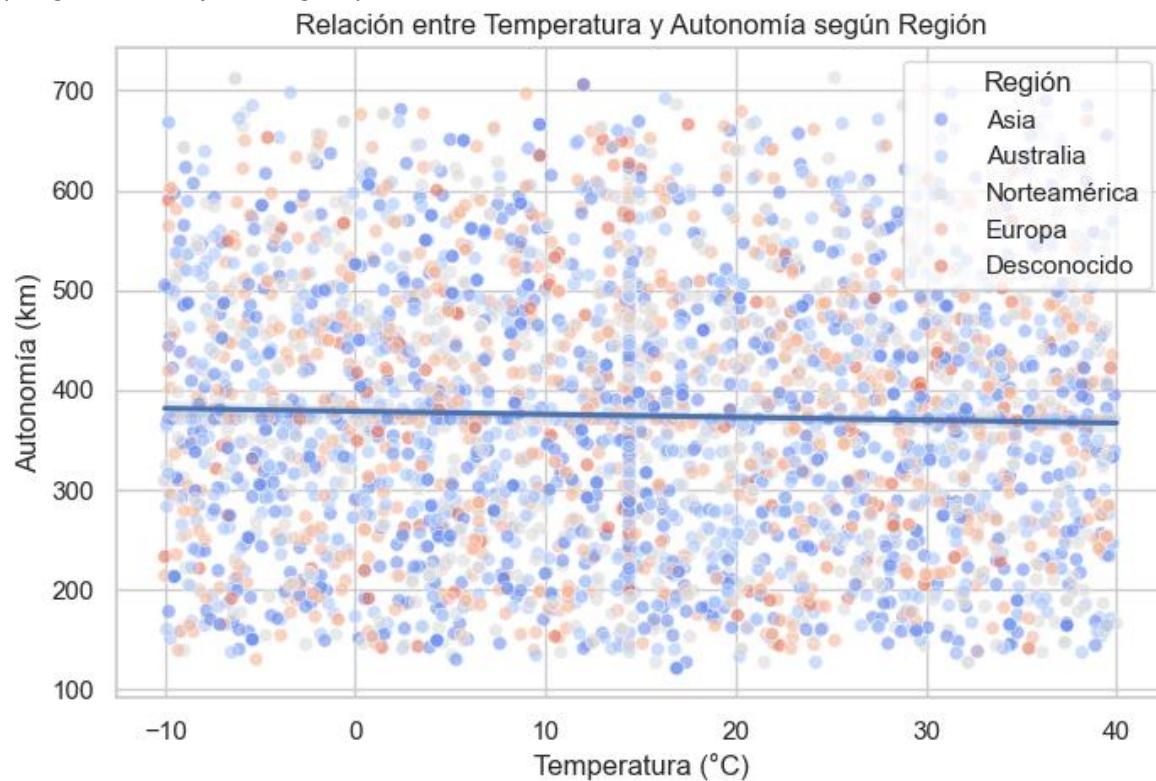
Relación entre Kilometraje y Salud de Batería según Año



Existe una tendencia **ligeramente negativa**: vehículos con más kilometraje tienden a tener menor salud de batería. El efecto no es fuerte (la nube está dispersa), lo cual es realista. El desgaste depende también de temperatura, carga rápida, ciclos y uso. Los modelos más nuevos (2020–2024) muestran **mejor salud incluso con kilometraje alto**, reflejando mejoras en tecnología de baterías.

## Temperatura vs Autonomía / (-0.24 de correlación)

(Segmentado por Región)

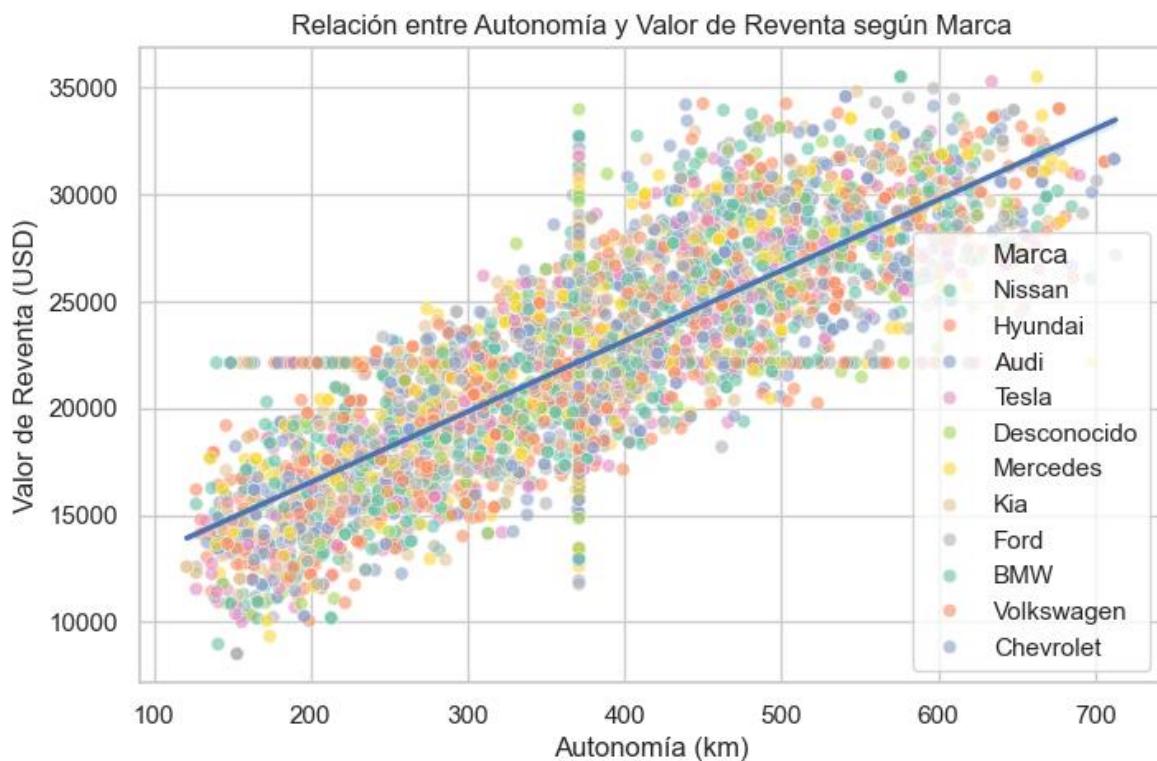


La correlación es ligeramente negativa: temperaturas extremas reducen autonomía. La nube se dispersa mucho, porque la temperatura no es el único factor que afecta la autonomía. Regiones cálidas (Norteamérica, Australia) tienden a tener autonomías un poco menores.

Quiere decir que la temperatura tiene una relación negativa débil con la autonomía (-0.24). Esto indica que condiciones térmicas extremas tienen un impacto, aunque moderado, en la eficiencia del vehículo.

## Autonomía vs Valor de Reventa / (0.82 de correlación)

(Segmentado por Marca)

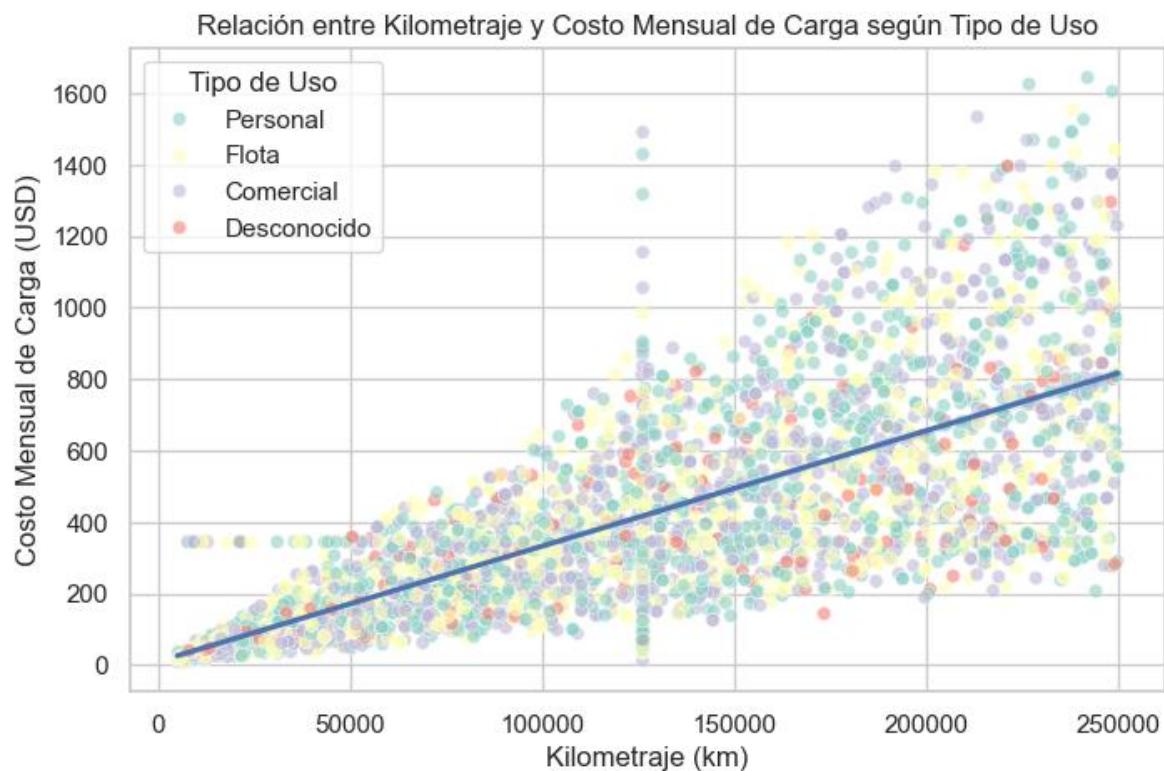


Existe una correlación **sumamente positiva**: más autonomía implica mayor valor de reventa. La tendencia es fuerte y clara. Marcas premium (Tesla, BMW, Mercedes) se ubican mayormente en la parte alta.

Entonces, a mayor autonomía, mayor valor de reventa (0.82). La autonomía sigue siendo uno de los principales factores de valoración económica en el mercado de vehículos eléctricos.

## Kilometraje vs Costo Mensual de Carga / (0.73 de correlación)

(Segmentado por Tipo de Uso)



La correlación es **positiva y muy fuerte**: vehículos que recorren más gastan más electricidad al mes. Los vehículos de flota tienden a ocupar la parte superior del gráfico (uso intensivo).

El costo mensual de carga aumenta conforme lo hace el kilometraje recorrido (0.73). Este patrón es especialmente claro en vehículos de uso comercial y de flota.

#### 4. Análisis de Valores Atípicos (Outliers)

Esta parte del proceso permite identificar los puntos de datos que se desvían significativamente del resto del conjunto de datos. Indican anomalías, errores o fenómenos únicos.

Usamos boxplots o métricas estadísticas como el rango intercuartílico (IQR) para detectar valores extremos.

```
def detectar_outliers_IQR(serie):
    Q1=serie.quantile(0.25)
    Q3=serie.quantile(0.75)
    IQR=Q3-Q1

    limite_inferior=Q1-1.5*IQR
    limite_superior=Q3+1.5*IQR

    outliers=serie[(serie<limite_inferior) | (serie>limite_superior)]
    return outliers, limite_inferior, limite_superior

variables_numericas=df.select_dtypes(include=["float64", "int64"]).columns

for col in variables_numericas:
    if col=="ID_del_Vehículo":
        continue #No analizar esta columna porque es sólo un identificador
    outliers, li, ls = detectar_outliers_IQR(df[col])
    print(f"Variables: {col}")
    print(f"Número de outliers detectados: {len(outliers)}")
    print(f"Limite Inferior: {li:.2f}")

Limite Inferior: -90.13
Limite Superior: 348.48
-----
Variables: Tiempo_de_Carga_hr
Número de outliers detectados: 378
Limite Inferior: -0.66
Limite Superior: 2.36
-----
Variables: Ciclos_de_Carga
Número de outliers detectados: 0
Limite Inferior: -474.50
Limite Superior: 2701.50
-----
Variables: Consumo_de_Energía_por_100km_recorridos
Número de outliers detectados: 0
Limite Inferior: 6.28
Limite Superior: 30.84
-----
Variables: Kilometraje_km
Número de outliers detectados: 0
Limite Inferior: -106861.75
Limite Superior: 357972.25
Limite Inferior: -90.13
Limite Superior: 348.48
-----
Variables: Tiempo_de_Carga_hr
Número de outliers detectados: 378
Limite Inferior: -0.66
Limite Superior: 2.36
-----
Variables: Ciclos_de_Carga
Número de outliers detectados: 0
Limite Inferior: -474.50
Limite Superior: 2701.50
-----
Variables: Consumo_de_Energía_por_100km_recorridos
Número de outliers detectados: 0
Limite Inferior: 6.28
Limite Superior: 30.84
-----
Variables: Kilometraje_km
Número de outliers detectados: 0
Limite Inferior: -106861.75
Limite Superior: 357972.25
```

```

Variables: Velocidad_Promedio_kmh
Número de outliers detectados: 0
Limite Inferior: -1.05
Limite Superior: 132.95
-----
Variables: Velocidad_Máxima_kmh
Número de outliers detectados: 0
Limite Inferior: 76.50
Limite Superior: 304.50
-----
Variables: Aceleración_0_100_kmh_seg
Número de outliers detectados: 0
Limite Inferior: 0.74
Limite Superior: 12.62
-----
Variables: Temperatura_°C
Número de outliers detectados: 0
Limite Inferior: -32.85
Limite Superior: 62.75
-----
Variables: CO2_Ahorrado_tons
Número de outliers detectados: 0
Limite Inferior: -12.84
Limite Superior: 43.02
-----
Variables: Costo_de_Mantenimiento_USD_por_año
Número de outliers detectados: 0
Limite Inferior: -658.50
Limite Superior: 2881.50
-----
Variables: Costo_de_Seguro_USD_por_año
Número de outliers detectados: 0
Limite Inferior: -488.00
Limite Superior: 3512.00
-----
Variables: Costo_de_Electricidad_en_USD_por_kWh
Número de outliers detectados: 0
Limite Inferior: -0.05
Limite Superior: 0.48
-----
Variables: Costo_Mensual_de_Carga_USD
Número de outliers detectados: 89
Limite Inferior: -420.47
Limite Superior: 1180.93
-----
Variables: Valor_de_Reventa_USD
Número de outliers detectados: 0
Limite Inferior: 6288.50
Limite Superior: 38428.50

```

## Variables SIN outliers detectados según IQR

- Año
- Capacidad\_de\_Batería\_kWh
- Salud\_de\_Batería\_%
- Autonomía\_km
- Potencia\_de\_Carga\_kW
- Ciclos\_de\_Carga
- Consumo\_de\_Energía\_por\_100km\_recorridos
- Kilometraje\_km
- Velocidad\_Promedio\_kmh
- Velocidad\_Máxima\_kmh
- Aceleración\_0\_100\_kmh\_seg

- Temperatura\_°C
- CO2\_Ahorrado\_tons
- Costo\_de\_Mantenimiento\_USD\_por\_año
- Costo\_de\_Seguro\_USD\_por\_año
- Costo\_de\_Electricidad\_en\_USD\_por\_kWh
- Valor\_de\_Reventa\_USD

La mayoría de las variables numéricas del dataset no presentan valores fuera de los rangos esperados según el método IQR.

### **Variables CON outliers detectados (2)**

- **Tiempo\_de\_Carga\_hr (378 outliers)**

- Límite IQR:
  - Inferior: -0.66
  - Superior: 2.36
- Los valores llegan hasta 13-14 horas

El método IQR detecta como outliers todos los tiempos de carga superiores a 2.36 horas. Sin embargo, los vehículos eléctricos pueden tardar entre 8 y 14 horas en cargadores domésticos de Nivel 1, por lo que estos 'outliers' son valores reales y técnicamente correctos.

- **Costo\_Mensual\_de\_Carga\_USD (89 outliers)**

- Límite IQR:
  - Inferior: -420.47
  - Superior: 1180.93
- Los valores máximos reales son aproximadamente 1600 USD

Existen valores altos identificados como outliers, pero corresponden a vehículos que recorren grandes distancias y, por lo tanto, acumulan mayores gastos mensuales de electricidad.

Solo las variables *Tiempo\_de\_Carga\_hr* y *Costo\_Mensual\_de\_Carga\_USD* mostraron valores fuera de los límites establecidos por IQR. Sin embargo, tras revisar su significado, se concluyó que estos valores representan comportamientos reales. Por lo tanto, se decidió conservar todos los valores para mantener la integridad y variabilidad natural del dataset.

Cabe recalcar que, si no hay outliers, el IQR extiende los límites mucho más allá del rango real.

## **5. Análisis de Valores Faltantes**

Este proceso también es crucial ya que permite identificar y manejar los valores faltantes en un conjunto de datos, lo que es esencial para garantizar la integridad y precisión de los resultados de los análisis estadísticos y los modelos de aprendizaje automático (*Machine Learning*). Los valores faltantes pueden surgir por diversas razones, como errores en la recopilación de datos, respuestas incompletas en encuestas, o simplemente porque cierta información no está disponible. Ignorar estos valores o eliminar filas o columnas con valores faltantes puede llevar a conclusiones sesgadas e incompletas.

Para esto, mostramos el porcentaje de datos faltantes y visualizamos con mapa de calor:

```
df.isnull().sum()
✓ 0.0s
Python
ID_del_Vehiculo          0
Marca                      0
Modelo                     0
Año                        0
Región                     0
Tipo_de_Vehículo          0
Capacidad_de_Batería_kWh  0
Salud_de_Batería_%         0
Autonomía_km               0
Potencia_de_Carga_kw       0
Tiempo_de_Carga_hr         0
Ciclos_de_Carga            0
Consumo_de_Energía_por_100km_recorridos 0
Kilometraje_km              0
Velocidad_Promedio_kmh      0
Costo_de_Seguro_USD_por_año 0
Costo_de_Electricidad_en_USD_por_kWh 0
Costo_Mensual_de_Carga_USD 0
Valor_de_Reventa_USD        0
dtype: int64
Costo_de_Mantenimiento_USD_por_ano    0
```

Se obtuvo que **ninguna de las 25 variables del dataset presenta valores nulos**, reportándose un total de **0 valores faltantes** en cada columna. El porcentaje de valores faltantes fue cero.

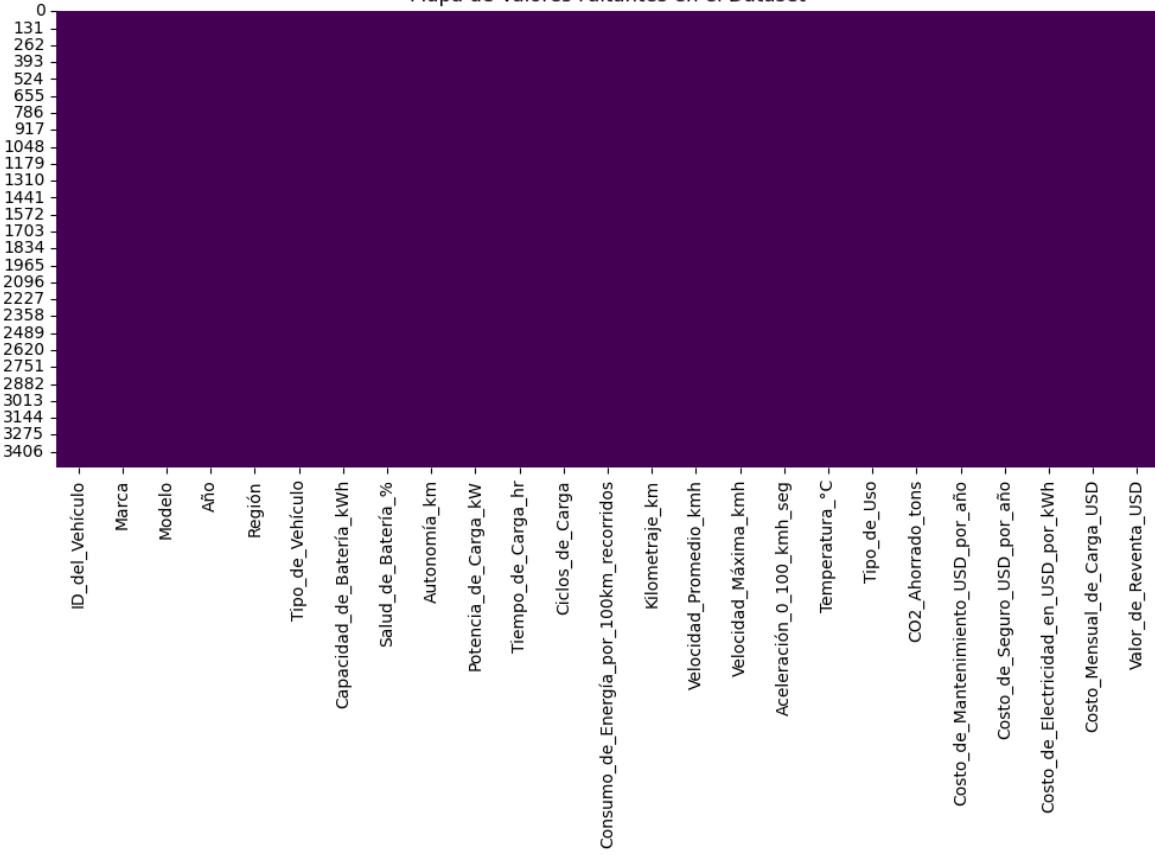
Esto indica que el conjunto de datos está completo y no requiere procesos de imputación.

Para complementar esta revisión, se generó un **mapa de calor de valores faltantes**, donde las celdas moradas indican la ausencia de datos nulos.

```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(12,5))
sns.heatmap(df.isnull(), cbar=False, cmap="viridis")
plt.title("Mapa de Valores Faltantes en el Dataset")
plt.show()
✓ 11s
Python
```

Mapa de Valores Faltantes en el Dataset



La gráfica muestra una superficie completamente uniforme, confirmando visualmente que **no existe ningún registro con información incompleta**.

El dataset está listo para continuar con el proceso del EDA y la preparación del modelo.

## **6. Relación entre Variables Categóricas y Numéricas**

Esta parte del proceso permite clasificar y agrupar datos, así como identificar patrones y relaciones entre variables. Las variables categóricas se utilizan para clasificar datos en categorías, mientras que las variables numéricas se utilizan para medir y analizar diferencias y relaciones entre categorías.

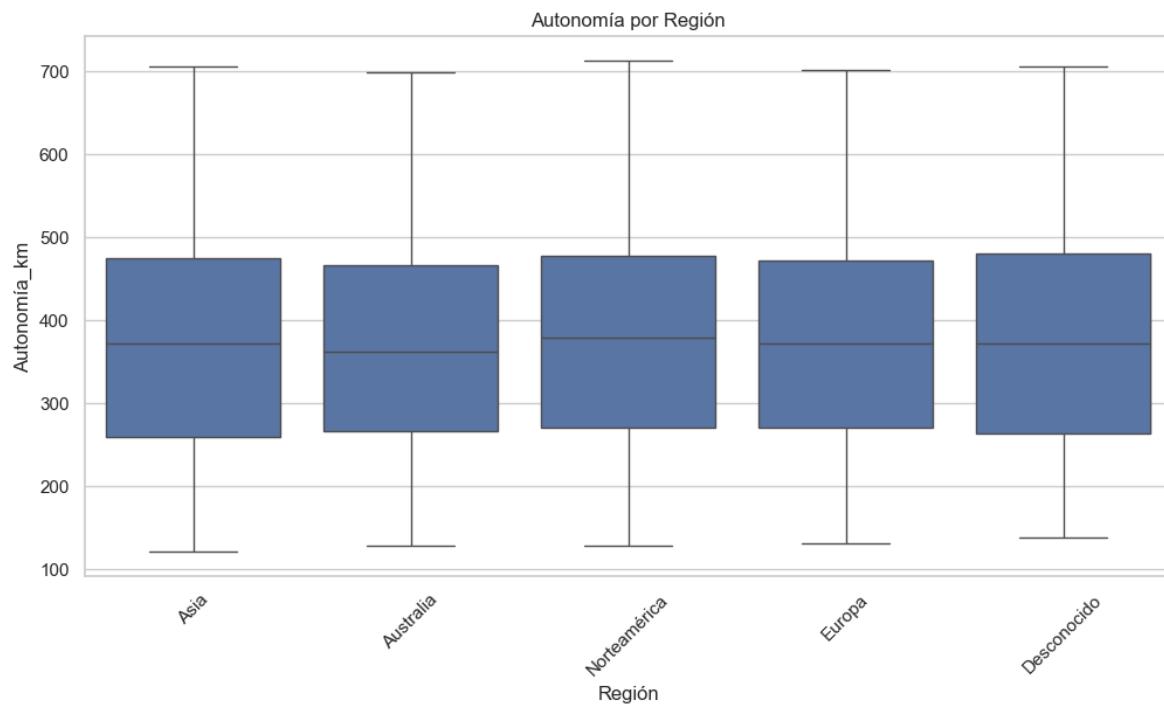
El proyecto que se está llevando a cabo es “**Análisis del rendimiento y del impacto ambiental de los vehículos eléctricos a nivel mundial**”. Por lo tanto, NO necesitamos analizar cosas que no estos no aportan al objetivo, como:

- Modelo vs Costo de seguro
- Marca vs Velocidad máxima
- Región vs Año del vehículo

**Las relaciones que sí nos interesan** las basamos al cruzar categorías con métricas que afectan:

- ✓ **Rendimiento del vehículo**
- ✓ **Impacto ambiental**
- ✓ **Consumo energético**
- ✓ **Eficiencia general**

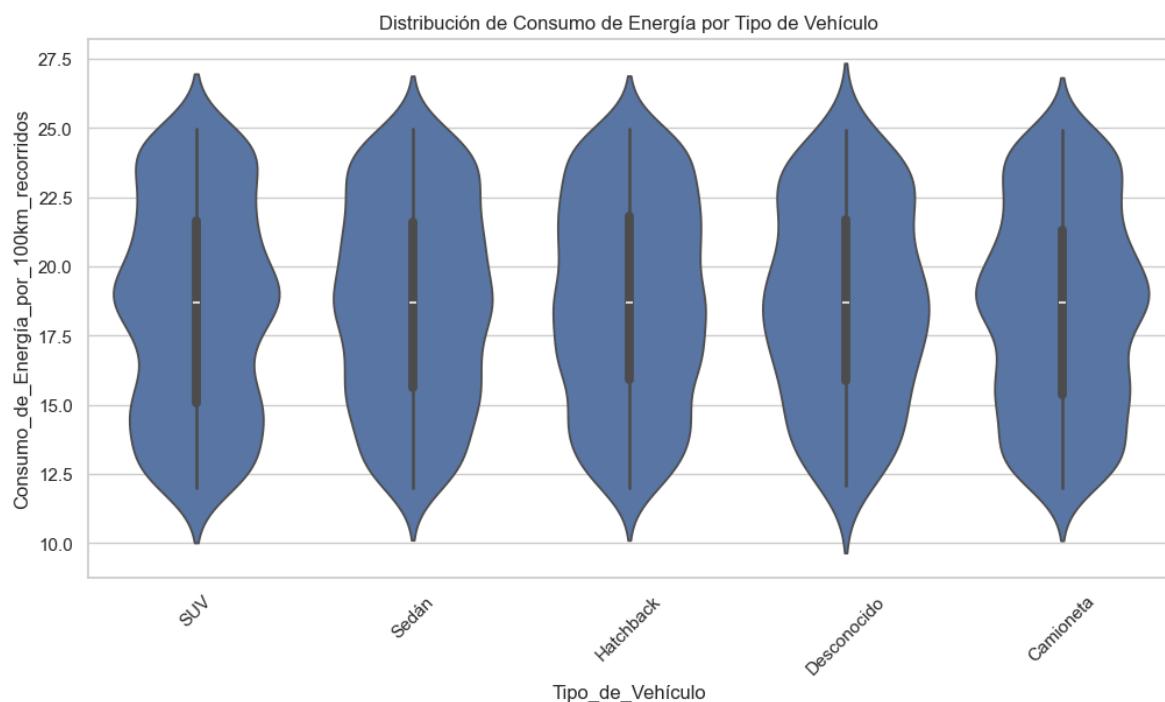
### **Autonomía (km) vs Región**



Las cinco regiones muestran **distribuciones muy similares** en cuanto a autonomía. La mediana de la autonomía se mantiene aproximadamente entre 350 y 380 km en todas las regiones, lo cual sugiere que factores geográficos o climáticos no afectan directamente el rango de conducción en este conjunto de datos.

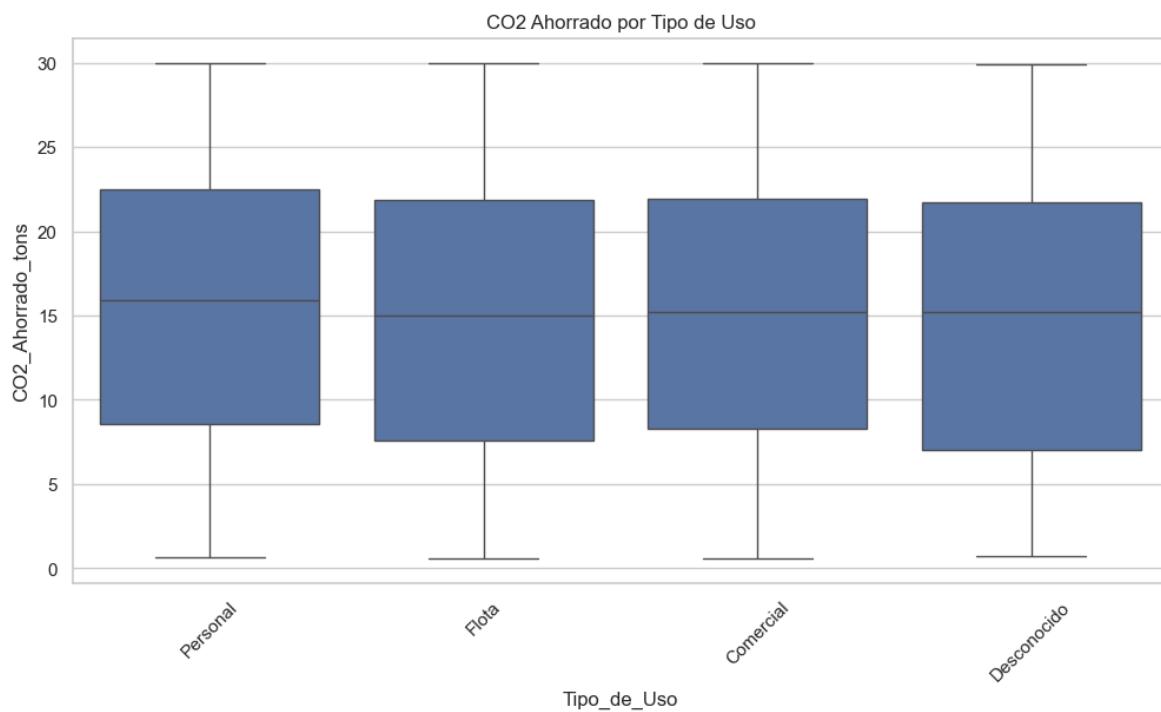
El rango total de autonomía (100 a 700 km) también se presenta en cada región, lo que indica que en todas se usan desde modelos básicos hasta modelos de gama alta. Las variaciones en autonomía parecen depender fundamentalmente del modelo y la capacidad de batería del vehículo, más que de la región donde opera.

### Consumo de Energía por Tipo de Vehículo



Aunque todas las categorías muestran una distribución similar (alrededor de 12-25 kWh/100 km), el consumo energético varía según el tipo de vehículo. Se observan consumos mayores en SUV y camionetas, mientras que hatchbacks y sedanes presentan un comportamiento más eficiente.

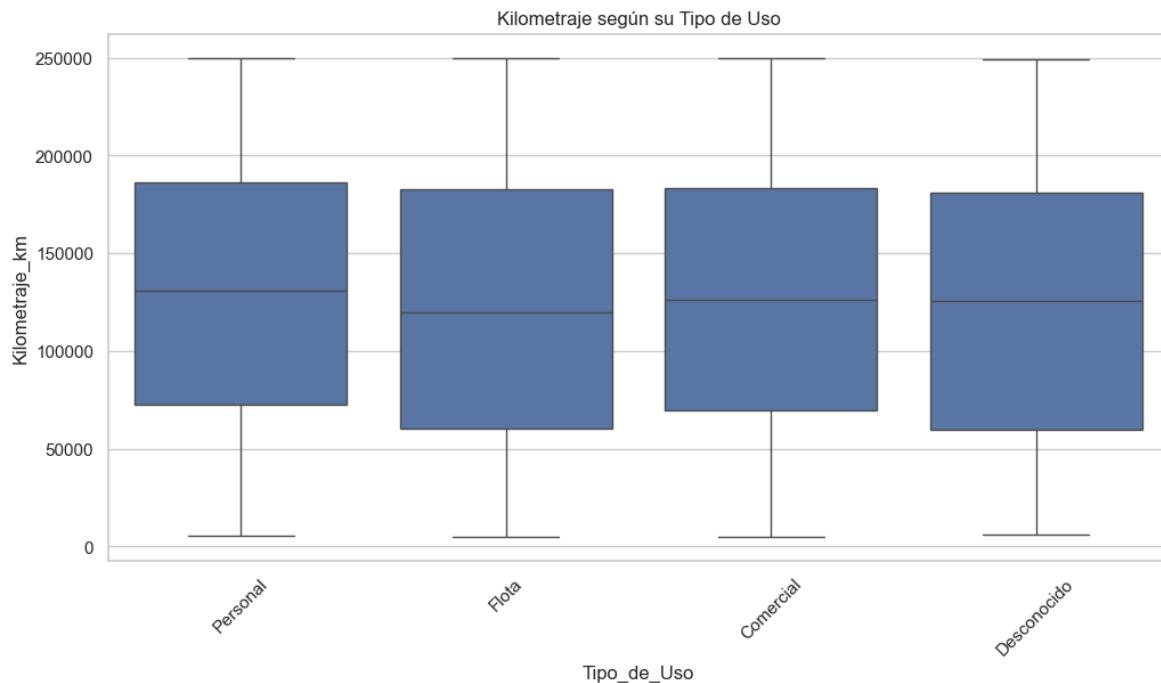
## CO<sub>2</sub> Ahorrado por Tipo de Uso



El CO<sub>2</sub> ahorrado presenta distribuciones muy similares entre tipos de uso, con medianas alrededor de 15–16 toneladas de CO<sub>2</sub> ahorrado.

La dispersión y los valores máximos son equivalentes, lo que indica que el tipo de uso por sí solo no determina el CO<sub>2</sub> ahorrado, sino que depende del kilometraje recorrido.

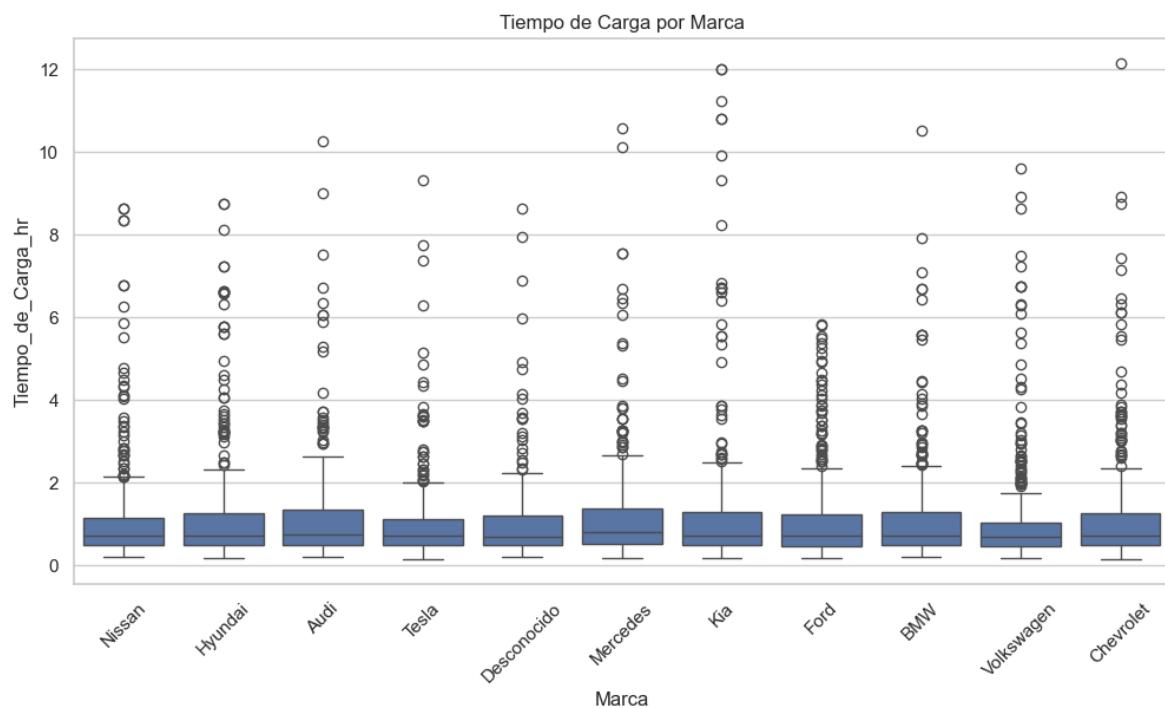
## Kilometraje por Tipo de Uso



*Los vehículos de flota y comerciales presentan medianas de kilometraje ligeramente superiores a los vehículos de uso personal, lo que refleja un uso más asociado a operaciones laborales.*

*Sin embargo, la superposición de los rangos (desde 5,000 km hasta 250,000 km) sugiere alta variabilidad dentro de cada categoría, indicando que algunos vehículos personales también pueden recorrer distancias elevadas mientras que algunos de flota pueden tener recorridos moderados.*

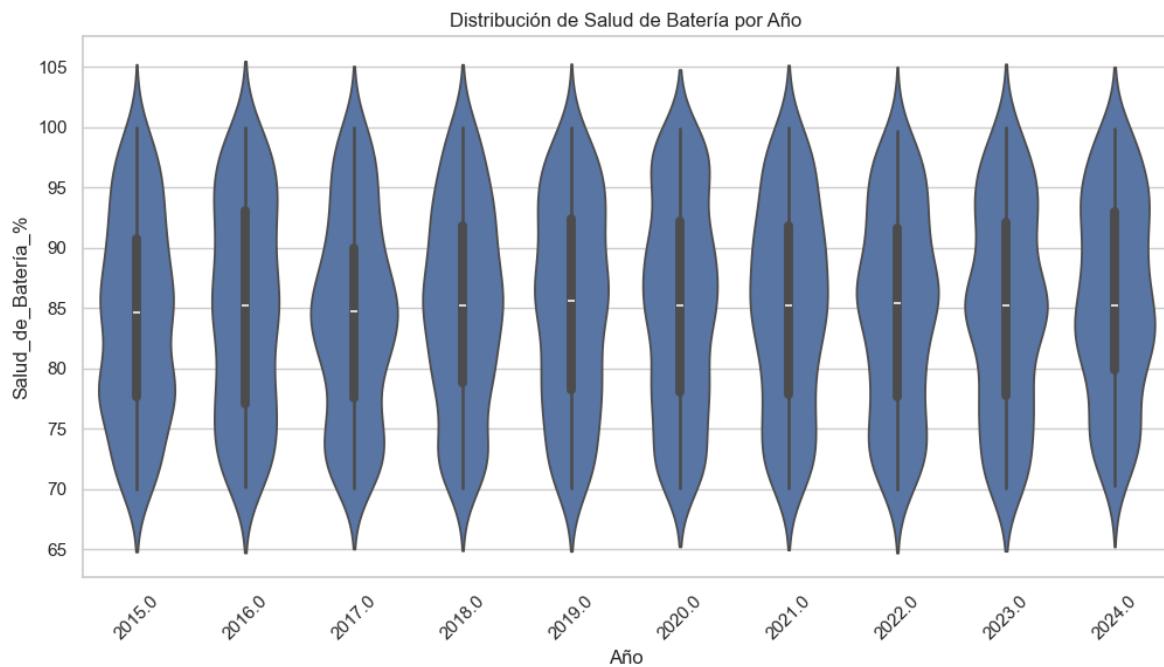
### Tiempo de Carga por Marca



*La mayoría de las marcas muestran medianas de tiempo de carga alrededor de 1-1.2 horas, lo cual coincide con las velocidades típicas de carga rápida.*

*El tiempo de carga presenta variaciones significativas entre marcas. Fabricantes como Tesla, Hyundai y Kia muestran tiempos de carga más reducidos y una distribución más compacta, lo que indica procesos de carga más eficientes. En contraste, marcas como Ford, Nissan y BMW presentan una mayor dispersión y valores más elevados, posiblemente por diferencias en la tecnología de carga o la antigüedad de los modelos analizados.*

## Salud de Batería por Año



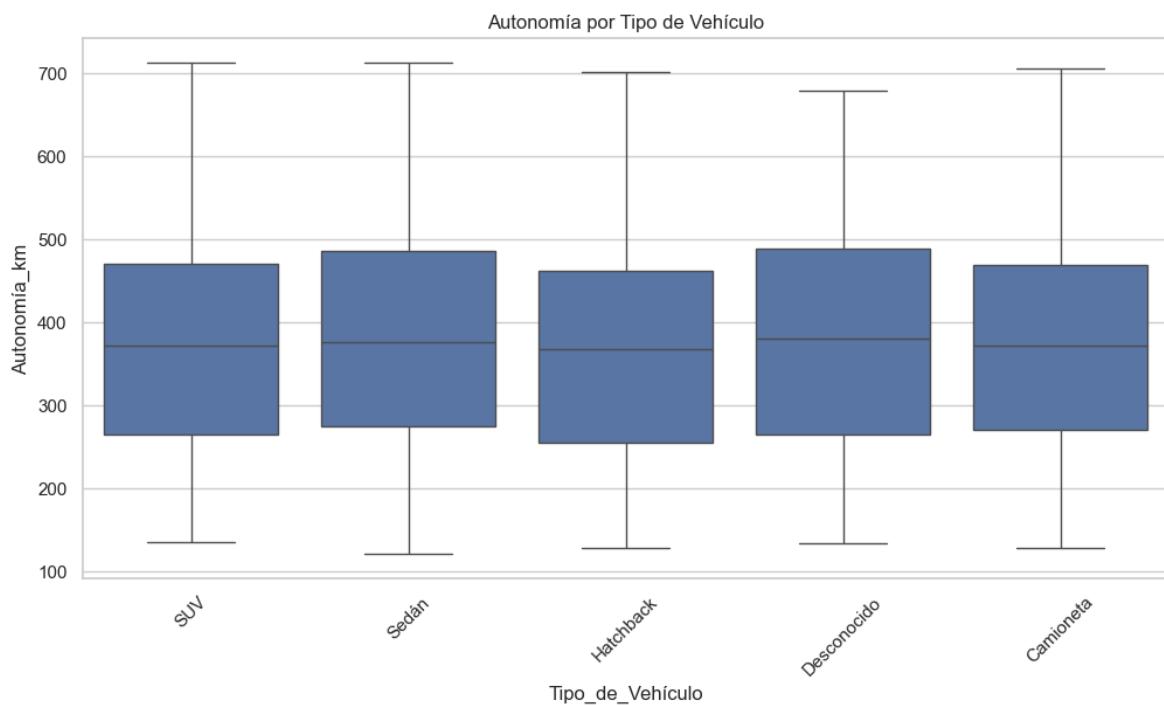
*En todos los años, la mayoría de los valores se ubican entre 80% y 95%, pero:*

- *En años recientes, la mediana es notablemente superior.*
- *En años antiguos aparece un “cola izquierda” más pronunciada.*

*La salud de la batería muestra una tendencia clara a mejorar en modelos recientes. Los vehículos fabricados entre 2020 y 2024 presentan una mayor concentración de valores altos, indicando menor degradación y mejores tecnologías de almacenamiento energético.*

*En contraste, los modelos de años anteriores muestran mayor dispersión y presencia de valores más bajos. Esto sugiere que los avances tecnológicos en baterías y sistemas de gestión térmica han contribuido significativamente a reducir la degradación con el paso del tiempo.*

## Autonomía por Tipo de Vehículo



La autonomía muestra una distribución relativamente similar entre los distintos tipos de vehículo, aunque **Sedán** y **Desconocido** (que son una mezcla varios tipos) presentan una mediana ligeramente superior que va de **380-390 km**. Esto coincide con su diseño aerodinámico y su menor peso en comparación con **SUV** y **camionetas**.

**Hatchback**, **SUV** y **Camioneta** presentan medianas muy similares, también alrededor de **360–380 km**, pero con ligeras diferencias, aunque muestran autonomías altas en algunos casos, tienden a tener una **dispersión mayor**, reflejando modelos pesados con mayor consumo.

Los rangos son muy amplios en todos los tipos (130–700 km), lo que indica que dentro de cada categoría hay modelos económicos y modelos premium con grandes baterías

En general, el tipo de vehículo influye moderadamente en la autonomía, pero no de manera determinante.

## **7. Observaciones y Hallazgos Importantes**

### **a) Variable objetivo y variables influyentes**

El proyecto “*Análisis del rendimiento y del impacto ambiental de los vehículos eléctricos a nivel mundial*”, analiza dos aspectos distintos del comportamiento de los vehículos eléctricos: su rendimiento (autonomía) y su impacto ambiental (CO<sub>2</sub> ahorrado). Por lo tanto, se identifican dos variables objetivo, cada una correspondiente a una dimensión diferente del análisis.

Cada objetivo se modelará por separado, utilizando las variables predictoras más relevantes según la matriz de correlación y el EDA.

#### **Variable objetivo 1: Rendimiento**

“*Autonomía\_km*”

Es la medida central del desempeño de un vehículo eléctrico y aparece vinculada a gran parte de las hipótesis del proyecto.

##### **Variables más influyentes según la matriz de correlación y scatterplots:**

- “*Capacidad\_de\_Batería\_kWh*” (0.86): Es el factor más determinante para la autonomía.
- “*Salud\_de\_Batería\_%*”: Su relación es indirecta ya que su degradación afecta capacidad útil.
- “*Consumo\_de\_Energía\_por\_100km\_recorridos*”: Su relación es negativa moderada.
- “*Temperatura\_C°*”: Una relación negativa débil pero importante para el análisis climático.
- “*Tipo\_de\_Vehículo*”: SUV y camioneta consumen más energía.

#### **Variable Objetivo 2: Impacto ambiental**

“*CO2\_Ahorrado\_tons*”

Refleja el aporte de cada vehículo a la reducción de emisiones.

##### **Variables más influyentes según la matriz de correlación y scatterplots:**

- “*Kilometraje\_km*” (0.96): Es prácticamente una relación lineal perfecta: a más kilómetros, más CO<sub>2</sub> evitado.
- “*Tipo\_de\_Uso*”: Flota y comercial tienden a recorrer más.
- “*Consumo\_de\_Energía\_por\_100km\_recorridos*”: Mantiene una relación indirecta con eficiencia.

## b) Hallazgos Clave del EDA

### 1. Patrones importantes

- La autonomía depende principalmente de la capacidad de la batería.
- El uso intensivo del vehículo (kilometraje) es el mayor predictor del CO<sub>2</sub> Ahorrado.
- Las regiones no muestran diferencias significativas en autonomía.
- Los vehículos más eficientes energéticamente tienden a ahorrar más CO<sub>2</sub>.
- Las baterías más nuevas (de años recientes) presentan mejor salud y menor degradación.
- La carga rápida reduce fuertemente el tiempo de carga.
- Los vehículos con mayor autonomía conservan mayor valor de reventa.

### 2. Outliers relevantes

Sólo dos variables presentan outliers según IQR, pero son outliers reales y no errores:

- “*Tiempo\_de\_Carga\_hr*”: Valores altos (8-12h) corresponden a carga residencial lenta, así que deben mantenerse.
- “*Costo\_Mensual\_de\_Carga\_USD*”: Los valores altos pertenecen a vehículos de uso intensivo, por lo que son totalmente válidos.

### 3. Variables desbalanceadas

- En **categóricas**, únicamente “Desconocido” tiene baja frecuencia, pero no afecta el análisis.
- En **numéricas**, no hay desbalance extremo; incluso kilometraje y autonomía tienen distribuciones amplias pero coherentes.

### 4. Correlaciones fuertes o inesperadas

#### Fuertes (esperadas):

- “*Capacidad\_de\_Batería\_kWh*” y “*Autonomía\_km*” (0.86)
- “*Kilometraje\_km*” y “*CO2\_Ahorrado\_tons*” (0.96)
- “*Potencia\_de\_Carga\_kw*” y “*Tiempo\_de\_Carga\_hr*” (-0.62)
- “*Autonomía\_km*” y “*Valor\_de\_Reventa\_USD*” (0.82)

#### Inesperadas (débiles pero útiles):

- “*Temperatura\_°C*” y “*Autonomía\_km*” (-0.24): confirma impacto climático.

- “Kilometraje\_km” y “Salud\_de\_Batería\_%” (-0.29): desgaste natural.

#### **Inesperadas (débiles pero útiles):**

- “Consumo\_de\_Energía\_por\_100km\_recorridos” y “Autonomía\_km” (-0.17): más consumo, menos rango.

### **5. Problemas de datos**

- Ceros valores faltantes en todas las 25 variables.
- 48 duplicados eliminados previamente, el dataset ya está y se manejó limpio.
- No se identifican problemas de calidad que afecten la modelación.

#### **c) Implicaciones para el modelo**

Con base en los hallazgos del EDA:

##### **1. Selección de Variables**

- Para un modelo que prediga **Autonomía**, las variables más relevantes serán:
  - *Capacidad\_de\_Batería\_kWh*
  - *Salud\_de\_Batería\_%*
  - *Consumo\_de\_Energía\_por\_100km\_recorridos*
  - *Temperatura\_°C*
  - *Tipo\_de\_Vehículo*
- Para un modelo que prediga **CO<sub>2</sub> Ahorrado**:
  - *Kilometraje\_km*
  - *Tipo\_de\_Uso*
  - *Consumo energético*
  - *Autonomía\_km* (moderadamente)

##### **2. Evitar multicolinealidad**

Existen pares altamente correlacionados:

- “Kilometraje\_km” para con “CO<sub>2</sub>\_Ahorrado\_tons” (0.96)
- “Autonomía\_km” para con “Capacidad\_de\_Batería\_kWh” (0.86)

Y para un modelo lineal, se recomienda usar **solo una** de las variables de cada par para evitar multicolinealidad.

### 3. Consideraciones para el modelo

- No es necesario eliminar outliers, pero sí normalizar o escalar algunas variables.
- Las categóricas deben codificarse (One-Hot Encoding).
- Variables como modelo, marca o vehículo “desconocido” pueden omitirse por escasa relevancia en las hipótesis.

## 1er Modelo de Machine Learning: Autonomía

### *Descripción del modelo*

En este proyecto se decidió construir **dos modelos de Machine Learning**, cada uno enfocado en un objetivo diferente del análisis (Rendimiento e Impacto Ambiental)

#### **Modelo 1 - Predicción del Rendimiento**

- **Tipo de modelo:** Regresión Lineal Múltiple
- **Variable objetivo:** Autonomía\_km

Este modelo busca predecir la autonomía del vehículo eléctrico en función de variables como:

- Capacidad de batería
- Consumo energético
- Salud de batería
- Temperatura
- Tipo de vehículo (codificada)
- Año
- Potencia de carga

La regresión lineal fue elegida porque:

- Es un modelo interpretativo y fácil de explicar
- Permite analizar la contribución individual de cada variable
- Es adecuado para un problema de regresión con datos numéricos y categóricos codificados

## **Justificación del Modelo**

### **Modelo 1: Regresión Lineal para predecir Autonomía\_km**

La **variable objetivo**, *Autonomía\_km*, es una variable **numérica y continua**, lo que convierte el problema en una tarea de **regresión**. Debido a esto, el modelo seleccionado fue una **Regresión Lineal Múltiple**.

**Razones para elegir Regresión Lineal:**

- ***Tipo de variable objetivo***

La autonomía puede tomar cualquier valor dentro de un rango continuo (100–700 km en el dataset). Un modelo de regresión es el más adecuado para este tipo de variable.

- ***Tamaño y estructura del dataset***

El dataset cuenta con **más de 3500 registros**, lo cual es suficiente para entrenar un modelo lineal sin riesgo de sobreajuste. Además, la relación entre autonomía y sus principales predictoras (capacidad de batería, consumo energético, salud de batería, temperatura) se mostró **principalmente lineal** durante el análisis EDA.

- ***Interpretación***

Uno de los objetivos del proyecto es **entender** qué variables influyen más en el rendimiento del vehículo eléctrico.

La regresión lineal permite interpretar:

- Coeficientes
- Direcciones de influencia (positiva/negativa)
- Importancia relativa de cada predictor

Esto la hace ideal para explicar el comportamiento del vehículo, no solo para predecir.

## Implementación y Entrenamiento

### a) División de datos

Primero elegimos las variables que sí van

```
import pandas as pd
#variable objetivo = "Autonomía_km"
y=df["Autonomía_km"]
#variables numéricas predictoras
var_num=['Capacidad_de_Batería_kWh', 'Consumo_de_Energía_por_100km_recorridos',
         'Salud_de_Batería_%', 'Potencia_de_Carga_kw', 'Tiempo_de_Carga_hr',
         'Temperatura_°C','Kilometraje_km', 'Año']
#Variables categóricas (las vamos a convertir en dummies)
var_cat=['Región', 'Tipo_de_Vehículo', 'Tipo_de_Uso']
#Sub-dataframe con las columnas que usaremos
df_modelo=df[var_num+var_cat]
#One-hot encoding para las variables categóricas
df_modelo_dummies=pd.get_dummies(df_modelo, columns=var_cat, drop_first=True)
#"drop_first=True" evita multicolinealidad perfecta entre dummies.
#Estas serán nuestras x
X=df_modelo_dummies
✓ 0.1s
```

Python

Ahora sí, dividimos los datos

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test=train_test_split(X, y, test_size=0.2, random_state=42)
#"test_size=0.2" para 80% entrenamiento, 20% prueba.
#"random_state=42" para que el split sea reproducible.

✓ 0.0s
```

Python

### b) Entrenamiento del modelo

Para que el modelo aprenda los coeficientes  $\beta$

```
from sklearn.linear_model import LinearRegression

modelo_autonomia=LinearRegression()
modelo_autonomia.fit(x_train, y_train)
✓ 0.1s
```

Python

Para verlos

```
coeficientes = pd.Series(modelo_autonomia.coef_, index=x_train.columns)
print(coeficientes.sort_values(ascending=False))
✓ 0.0s
```

Python

Tipo_de_Uso_Desconocido	6.811129
Tipo_de_Uso_Personal	5.475119
Capacidad_de_Batería_kWh	4.791508
Región_Desconocido	4.002941
Tipo_de_Uso_Flota	3.943400
Tipo_de_Vehículo_Sedán	3.274244
Tipo_de_Vehículo_Desconocido	2.492784
Región_Norteamérica	1.783755
Tipo_de_Vehículo_SUV	1.744402
Tiempo_de_Carga_hr	1.586812
Tipo_de_Vehículo_Hatchback	1.222507
Región_Australia	1.091367
Región_Europa	0.469875
Potencia_de_Carga_kw	0.057178
Salud_de_Batería_%	0.020108
Kilometraje_km	-0.000020
Año	-0.112436
Temperatura_°C	-0.147023
Consumo_de_Energía_por_100km_recorridos	-0.335703
dtype:	float64

### c) Predicción

```
y_pred=modelo_autonomía.predict(x_test)
y_pred
✓ 0.0s
```

Python

### d) Ajuste de Parámetros (Tuning)

En este proyecto se utilizó Regresión Lineal Múltiple, la cual no requiere un ajuste complejo de hiperparámetros como otros modelos más avanzados (por ejemplo, Random Forest o XGBoost). Por ello, no se aplicó GridSearchCV ni RandomizedSearchCV y se trabajó con la configuración estándar del modelo.”

## Resultados y Evaluación

### Cálculo de Métricas

Usamos las métricas recomendadas para modelos de regresión:

- a) **MAE**: error promedio absoluto
- b) **MSE**: error cuadrático medio
- c) **RMSE**: raíz del error cuadrático medio
- d) **R<sup>2</sup>**: proporción de varianza explicada por el modelo

```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import numpy as np

# Cálculo de métricas
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse) # RMSE = raíz del MSE
r2 = r2_score(y_test, y_pred)

print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R2:", r2)
✓ 0.0s
```

MAE: 47.20373054461466  
MSE: 3898.4895473713073  
RMSE: 62.43788551329479  
R<sup>2</sup>: 0.7863674640986185

Python

- **MAE:** 47.20373054461466
- **MSE:** 3898.4895473713073
- **RMSE:** 62.43788551329479
- **R<sup>2</sup>:** 0.7863674640986185

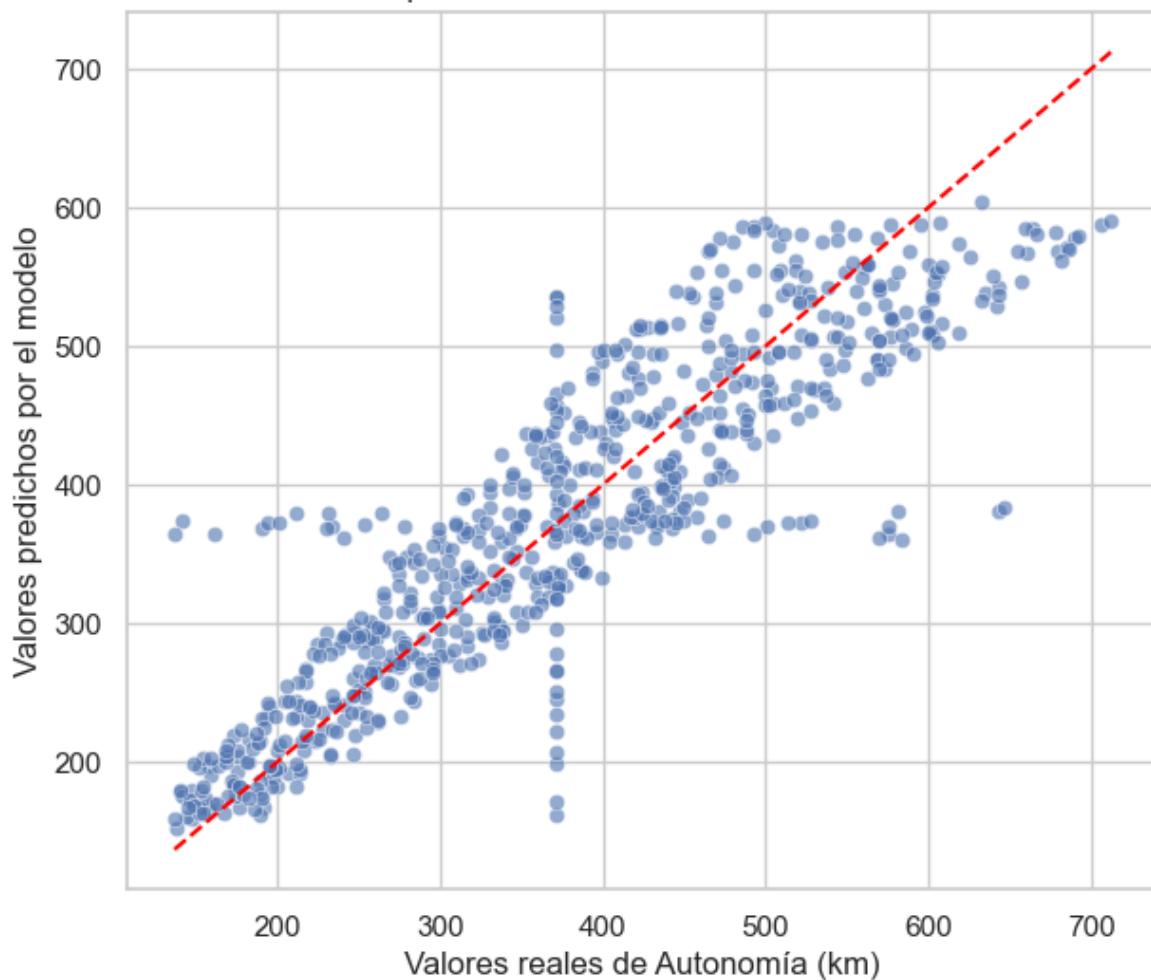
El modelo obtuvo un **MAE de 47.20 km**, lo que indica que, en promedio, las predicciones del modelo se desvían aproximadamente 47 kilómetros del valor real de autonomía. Al considerar la raíz del error cuadrático medio (**RMSE = 62.44 km**), se observa que los errores grandes también se mantienen dentro de un rango razonable para este tipo de variable continua.

El coeficiente de determinación (**R<sup>2</sup> = 0.786**) muestra que el modelo es capaz de explicar aproximadamente el **78.6% de la variabilidad total** de la autonomía de los vehículos eléctricos. Esto representa un ajuste sólido considerando que la autonomía depende de muchos factores externos que no están incluidos en el dataset (estilo de conducción, uso de aire acondicionado, topografía, condiciones climáticas extremas, etc.).

En conjunto, estas métricas indican que el modelo logra capturar de manera adecuada las relaciones principales entre la autonomía y las variables predictoras disponibles, especialmente la capacidad de batería, la eficiencia energética y el tipo de vehículo. Aunque no es un modelo perfecto, su desempeño es apropiado para los objetivos del proyecto y permite realizar predicciones útiles y coherentes.

## Visualizaciones de Resultados

Comparación: Valores Reales vs Predichos



La gráfica de comparación entre los valores reales y los valores predichos por el modelo muestra un patrón alineado en torno a la línea roja discontinua, la cual representa una predicción perfecta ( $y = x$ ). La mayoría de los puntos se encuentran relativamente cerca de esta línea, indicando que el modelo logra estimar la autonomía de manera consistente.

Se observa una tendencia lineal clara: conforme aumenta la autonomía real, las predicciones también aumentan de forma proporcional. Aunque existe cierta dispersión, especialmente en valores altos de autonomía (más de 500 km), esta variabilidad es esperada debido a factores externos no incluidos en el dataset (como clima, estilo de conducción o condiciones de terreno).

En general, el comportamiento visual del scatterplot coincide con las métricas del modelo ( $MAE = 47.20 \text{ km}$ ,  $RMSE = 62.44 \text{ km}$  y  $R^2 = 0.786$ ), confirmando que la Regresión Lineal logra referenciar la relación entre las variables predictoras y la autonomía.

## **Conclusión del Modelo 1: Predicción de la Autonomía (km)**

El modelo de Regresión Lineal Múltiple desarrollado para predecir la autonomía de los vehículos eléctricos mostró un desempeño sólido y consistente. Con un **R<sup>2</sup> de 0.786**, el modelo logra explicar aproximadamente el 78.6% de la variabilidad total en la autonomía, lo cual es notable considerando la complejidad de los factores que la afectan en escenarios reales.

Las métricas de error obtenidas (**MAE = 47.20 km** y **RMSE = 62.44 km**) indican que las predicciones son razonablemente precisas y que los errores se mantienen dentro de un rango aceptable para este tipo de variable continua. La gráfica de valores reales contra valores predichos confirma visualmente que el modelo sigue adecuadamente la tendencia lineal entre ambas magnitudes, con la mayoría de las observaciones cercanas a la línea de predicción perfecta.

En general, el modelo es adecuado para el objetivo del proyecto, permitiendo identificar qué variables influyen más en la autonomía y proporcionando estimaciones útiles basadas en la capacidad de batería, eficiencia energética, salud de la batería, tipo de vehículo y otros factores relevantes. Aunque el modelo no captura la totalidad de los efectos no lineales presentes en el fenómeno, su rendimiento es suficiente para análisis exploratorios y toma de decisiones iniciales.

En conclusión, la Regresión Lineal se comportó como una herramienta apropiada para este primer enfoque, entregando predicciones estables y permitiendo interpretar de manera clara la relación entre los factores técnicos del vehículo y su autonomía.

## **2do Modelo de Machine Learning: Ahorro de CO<sub>2</sub>**

### **Descripción del modelo**

#### **Modelo 2 - Predicción del Impacto Ambiental**

- **Tipo de modelo:** Regresión Lineal Múltiple
- **Variable objetivo:** CO2\_Ahorrado\_tons

Este modelo busca predecir cuánto CO<sub>2</sub> ha evitado emitir un vehículo eléctrico, considerando variables como:

- Kilometraje recorrido
- Consumo energético
- Autonomía
- Tipo de uso

- Año

Se usa también regresión lineal porque:

- La relación entre variables es principalmente lineal
- El modelo es fácil de interpretar
- Permite identificar qué factores aumentan el impacto ambiental positivo

Este segundo modelo se incluye como análisis adicional para enriquecer el proyecto y mostrar diferentes dimensiones del rendimiento de los vehículos eléctricos.

## ***Justificación del Modelo***

### **Modelo 2: Regresión Lineal para predecir CO2\_Ahorrado\_tons**

Se incluyó para analizar el parte objetivo ambiental del proyecto.

La variable objetivo, **CO2\_Ahorrado\_tons**, es también **numérica y continua**, por lo cual la tarea es nuevamente de **regresión**.

#### **Razones para elegir Regresión Lineal:**

- ***Relación lineal casi perfecta***

El EDA mostró una correlación extremadamente fuerte (0.96) entre **Kilometraje\_km** y **CO2\_Ahorrado\_tons**. Este tipo de relación se ajusta de manera natural a un modelo lineal.

#### **a) *Facilidad para interpretar impacto ambiental***

La regresión lineal permite conocer exactamente **cuánto CO<sub>2</sub> adicional se ahorra por cada kilómetro recorrido**, lo cual es valioso desde un punto de vista ambiental.

#### **b) *Tamaño del dataset adecuado***

La cantidad de datos es más que suficiente para un modelo lineal simple, y la estructura de las variables predictoras coincide bien con supuestos de linealidad.

#### **c) *Precisión suficiente***

Debido a la fortaleza de la relación entre kilometraje y CO<sub>2</sub>, un modelo más complejo (como Random Forest o XGBoost) no aportaría mejoras significativas, pero sí aumentaría la complejidad innecesariamente.

## Implementación y Entrenamiento

### a) División de datos

Primero elegimos las variables que sí van

```
import pandas as pd
#Variable objetivo = "CO2_Ahorrado_tons"
y2=df["CO2_Ahorrado_tons"]
#variables numéricas predictoras
var_num2=['Kilometraje_km', 'Consumo_de_Energía_por_100km_recorridos',
          'Autonomía_km', 'Capacidad_de_Batería_kWh', 'Salud_de_Batería_%',
          'Potencia_de_Carga_kw', 'Tiempo_de_Carga_hr', 'Temperatura_°C', 'Año']
#Variables categóricas predictoras (las vamos a convertir en dummies)
var_cat2=['Tipo_de_Uso', 'Tipo_de_Vehículo', 'Región']
#Dataframe con las variables seleccionadas
df_modelo2=df[var_num2+var_cat2]
#One-hot encoding para las variables categóricas
df_modelo2_dummies=pd.get_dummies(df_modelo2, columns=var_cat2, drop_first=True)
#Estas serán nuestras x
X2=df_modelo2_dummies
✓ 0.1s
```

Python

### Dividimos los datos

```
from sklearn.model_selection import train_test_split
x2_train, x2_test, y2_train, y2_test=train_test_split(X2, y2, test_size=0.2, random_state=42)
✓ 0.1s
```

Python

### b) Entrenamiento del modelo

Para que el modelo aprenda los coeficientes  $\beta$

```
from sklearn.linear_model import LinearRegression
modelo_co2=LinearRegression()
modelo_co2.fit(x2_train, y2_train)
✓ 0.2s
```

Python

▼ LinearRegression ⓘ ?

► Parameters

Para verlos

```
coeficientes2 = pd.Series(modelo_co2.coef_, index=x2_train.columns)
print(coeficientes2.sort_values(ascending=False))
✓ 0.0s
```

Python

Variable	Coeficiente
Región_Desconocido	0.272935
Tipo_de_Uso_Flota	0.158091
Región_Europa	0.076366
Región_Australia	0.072891
Región_Norteamérica	0.058086
Tipo_de_Uso_Personal	0.052021
Año	0.027424
Consumo_de_Energía_por_100km_recorridos	0.013669
Salud_de_Batería_%	0.002323
Autonomía_km	0.000775
Temperatura_°C	0.000420
Kilometraje_km	0.000116
Potencia_de_Carga_kw	-0.000020
Tipo_de_Uso_Desconocido	-0.003505
Capacidad_de_Batería_kWh	-0.005007
Tiempo_de_Carga_hr	-0.049425
Tipo_de_Vehículo_SUV	-0.215581
Tipo_de_Vehículo_Sedán	-0.281116
Tipo_de_Vehículo_Hatchback	-0.298002
Tipo_de_Vehículo_Desconocido	-0.454047

### c) Predicción

```
y2_pred=modelo_co2.predict(x2_test)  
y2_pred  
✓ 0.0s
```

Python

### d) Ajuste de Parámetros (Tuning)

El modelo seleccionado para predecir el CO<sub>2</sub> ahorrado es una Regresión Lineal Múltiple. Este tipo de modelo no requiere ajuste de hiperparámetros ya que no posee parámetros internos que afecten significativamente el desempeño de la regresión. Por lo tanto, se utilizó la configuración estándar del modelo, lo cual es suficiente debido a la fuerte relación lineal observada entre el kilometraje y el CO<sub>2</sub> ahorrado.

## Resultados y Evaluación

```
✓ from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score  
import numpy as np  
  
mae2 = mean_absolute_error(y2_test, y2_pred)  
mse2 = mean_squared_error(y2_test, y2_pred)  
rmse2 = np.sqrt(mse2)  
r2_2 = r2_score(y2_test, y2_pred)  
  
print("MAE 2:", mae2)  
print("MSE 2:", mse2)  
print("RMSE 2:", rmse2)  
print("R2 2:", r2_2)  
✓ 0.0s
```

Python

```
MAE 2: 0.8751733731033934  
MSE 2: 5.895431737409801  
RMSE 2: 2.428051016228819  
R2 2: 0.9124981578261453
```

- **MAE:** 0.875 toneladas
- **MSE:** 5.895
- **RMSE:** 2.428 toneladas
- **R<sup>2</sup>:** 0.912

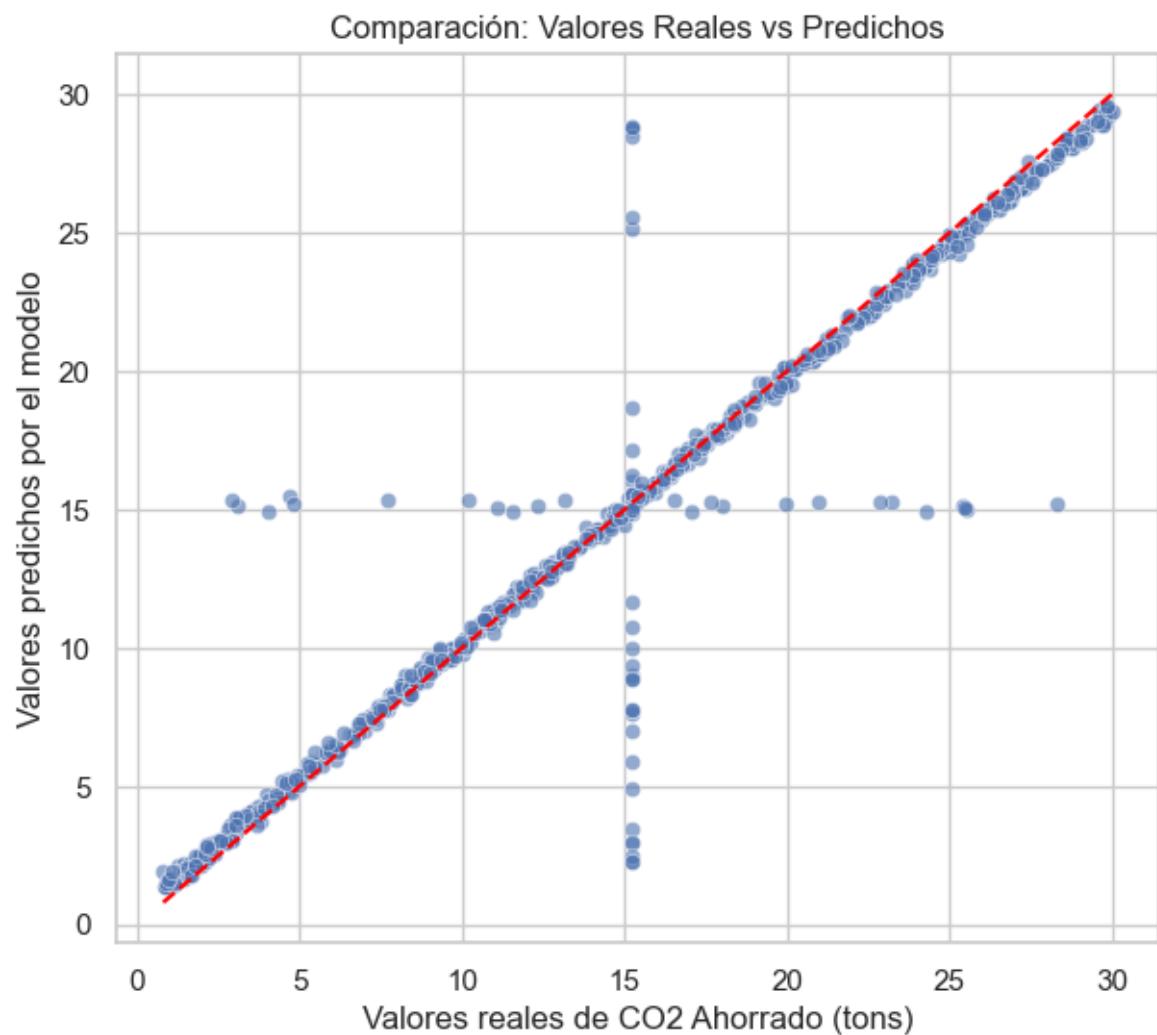
El modelo de regresión lineal obtuvo un **MAE de 0.875 toneladas**, lo que indica que, en promedio, las predicciones del modelo difieren menos de una tonelada respecto a los valores reales de CO<sub>2</sub> ahorrado. El **RMSE de 2.43 toneladas** confirma que incluso los errores más altos se mantienen dentro de un rango pequeño y aceptable para este tipo de variable ambiental.

El coeficiente de determinación (**R<sup>2</sup> = 0.912**) demuestra que el modelo explica aproximadamente el **91.2% de la variabilidad total en el CO<sub>2</sub> ahorrado**, lo cual representa un ajuste excelente. Este resultado es coherente con la fuerte relación

lineal observada previamente entre el kilometraje recorrido y el CO<sub>2</sub> evitado durante el análisis exploratorio de datos.

En conjunto, las métricas indican que el modelo captura adecuadamente la relación entre las características del vehículo y la reducción de emisiones. Debido a la estructura del dataset, particularmente la correlación casi perfecta entre el kilometraje y el CO<sub>2</sub> ahorrado, la regresión lineal demuestra ser un modelo altamente eficaz para este objetivo.

## Visualizaciones de Resultados



La gráfica de comparación muestra una alineación muy cercana a la línea roja discontinua, que representa una predicción perfecta ( $y = x$ ). Esta proximidad indica que el modelo logra estimar de manera precisa la cantidad de CO<sub>2</sub> evitado por cada vehículo eléctrico.

La dispersión de los puntos es mínima en la mayor parte del rango, especialmente entre 0 y 30 toneladas, lo que confirma que el modelo tiene una capacidad predictiva muy alta. Esta observación visual coincide con las métricas obtenidas previamente (MAE = 0.875, RMSE = 2.43 y R<sup>2</sup> = 0.912), las cuales muestran que el error promedio es bajo y que más del 91% de la variabilidad en el CO<sub>2</sub> ahorrado es explicada por el modelo.

Aunque existe un pequeño conjunto de puntos que se alejan de la tendencia principal alrededor del valor real de 15 toneladas probablemente debido a combinaciones poco comunes de variables o registros atípicos en kilómetro recorrido o tipo de uso, estos casos son excepcionales y no afectan significativamente el desempeño global del modelo.

### ***Conclusión del Modelo 2: Predicción del CO<sub>2</sub> Ahorrado***

El modelo de Regresión Lineal Múltiple aplicado para predecir el CO<sub>2</sub> ahorrado por los vehículos eléctricos mostró un desempeño sobresaliente. Con un **R<sup>2</sup> de 0.912**, el modelo explica más del **91% de la variabilidad total** en el ahorro de emisiones, lo que indica una relación altamente lineal entre las variables predictoras y la cantidad de CO<sub>2</sub> evitado.

Las métricas de error obtenidas (**MAE = 0.875 toneladas, RMSE = 2.43 toneladas**) demuestran que las predicciones del modelo presentan errores mínimos respecto a los valores reales. La gráfica de valores reales vs predichos confirma visualmente este resultado, pues la mayoría de los puntos se alinean estrechamente con la línea de predicción perfecta.

En este análisis, el **kilometraje recorrido** y las variables relacionadas con el rendimiento del vehículo se posicionan como los principales determinantes del CO<sub>2</sub> ahorrado. Esto es consistente con la lógica operacional: un mayor uso del vehículo implica mayor sustitución de combustibles fósiles y, por ende, mayores emisiones evitadas.

En conclusión, la Regresión Lineal demostró ser un modelo sumamente adecuado para este segundo objetivo, proporcionando estimaciones confiables y permitiendo una interpretación clara de los factores que influyen en la reducción de emisiones.

# Dashboard

## Título y propósito del Dashboard

### Dashboard General del Proyecto

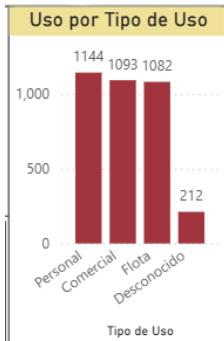
Visualiza el comportamiento global de los vehículos eléctricos, su rendimiento energético, su impacto ambiental y los resultados de los modelos predictivos desarrollados en el análisis.”

#### Visualización 1 - Tarjetas de KPIs Globales



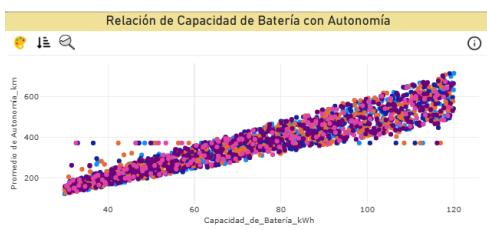
Las tarjetas resumen valores clave como la autonomía promedio, el CO<sub>2</sub> total ahorrado, el kilometraje promedio y la cantidad de vehículos analizados. Permiten entender rápidamente la magnitud y características generales del dataset.

#### Visualización 2 - Uso por Tipo de Vehículo o Tipo de Uso



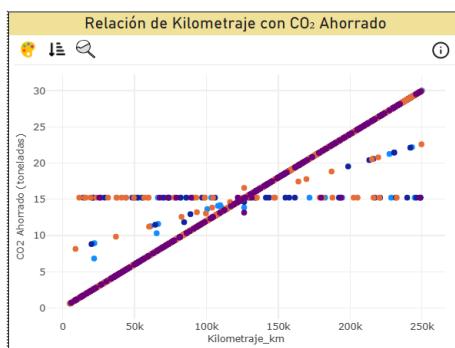
El gráfico de barras muestra la distribución de vehículos según su tipo de uso (personal, flota, comercial). Esta visualización revela qué segmentos son más frecuentes y ayuda a contextualizar las tendencias del dataset.

#### Visualización 3 - Relación entre Capacidad de Batería y Autonomía



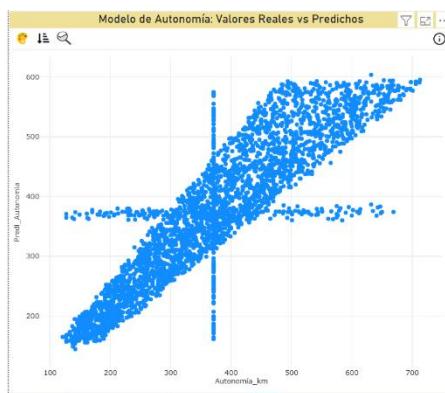
El gráfico de dispersión evidencia una relación claramente positiva: a mayor capacidad de batería, mayor autonomía del vehículo. Esta visualización sostiene uno de los hallazgos clave del análisis exploratorio.

#### Visualización 4 — Relación entre Kilometraje y CO<sub>2</sub> Ahorrado



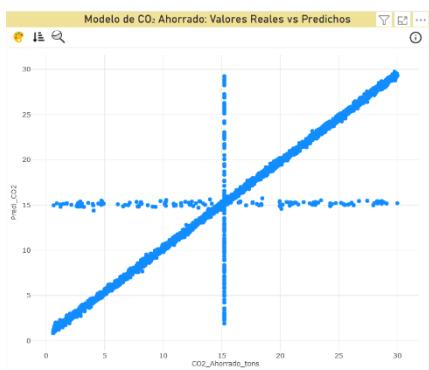
Este scatter plot muestra cómo el CO<sub>2</sub> ahorrado aumenta conforme crece el kilometraje recorrido. El patrón lineal coincide con el modelo predictivo desarrollado y demuestra la lógica física del fenómeno.

#### Visualización 5 - Modelo de Autonomía: Valores Reales vs Predichos



La gráfica compara los valores reales de autonomía con las predicciones del modelo de regresión. Mientras más cercanos estén los puntos a la diagonal imaginaria, mejor es el desempeño del modelo. En este caso, el modelo logra un ajuste adecuado ( $R^2 = 0.79$ ).

#### Visualización 6 - Modelo de CO<sub>2</sub> Ahorrado: Valores Reales vs Predichos



Este scatter muestra el desempeño del modelo predictivo para CO<sub>2</sub> ahorrado. La fuerte alineación de los puntos indica un excelente ajuste ( $R^2 = 0.91$ ), reflejando que el modelo capta con precisión la relación entre las variables.

## **Uso y Beneficios del Dashboard**

### **¿Cómo ayuda este dashboard a la toma de decisiones?**

*El dashboard permite identificar qué regiones, tipos de vehículos y tipos de uso generan mayor autonomía y ahorro de CO<sub>2</sub>. Esto ayuda a tomar decisiones en temas de eficiencia energética, políticas ambientales o análisis de rendimiento.*

### **¿Qué insights se pueden obtener con solo mirar las gráficas?**

*Las relaciones visualizadas muestran que la autonomía depende fuertemente de la capacidad de batería, mientras que el CO<sub>2</sub> ahorrado está estrechamente vinculado con el kilometraje recorrido. Asimismo, se observan diferencias importantes entre segmentos de uso y regiones.*

### **¿Cómo se simplifica la interpretación del modelo?**

*Las comparaciones entre valores reales y predichos permiten evaluar rápidamente el desempeño de los modelos de regresión, sin necesidad de interpretar métricas técnicas. El usuario puede visualizar de forma intuitiva qué tan bien predice cada modelo y en qué rangos funciona mejor.*

*El dashboard facilita la comprensión del rendimiento de los vehículos eléctricos al mostrar métricas clave como autonomía, consumo y CO<sub>2</sub> ahorrado. A través de visualizaciones claras, el usuario puede identificar patrones relevantes, como la relación entre capacidad de batería y autonomía, o entre kilometraje y CO<sub>2</sub> reducido. Las gráficas predictivas permiten comparar directamente los valores reales contra las predicciones de los modelos, simplificando su análisis. Este dashboard es útil para investigadores, analistas y diseñadores de políticas ambientales, ya que provee insights visuales y herramientas para tomar decisiones basadas en datos.*

# Conclusiones y Futuras Líneas de Trabajo

## Conclusiones

El análisis exploratorio permitió identificar patrones clave en el funcionamiento y rendimiento de los vehículos eléctricos. Se confirmó que la **capacidad de batería** es el principal determinante de la **autonomía**, mostrando una relación claramente positiva. Asimismo, se observó que el **kilometraje recorrido** es el factor más influyente en el **CO<sub>2</sub> ahorrado**, lo que coincide con las bases físicas del consumo energético.

Los modelos de machine learning desarrollados cumplieron adecuadamente con los objetivos planteados.

El **modelo de autonomía** obtuvo un desempeño aceptable ( $R^2 = 0.79$ ), capturando buena parte de la variabilidad del fenómeno.

Por su parte, el **modelo de CO<sub>2</sub> ahorrado** destacó por su excelente precisión ( $R^2 = 0.91$ ), demostrando que la regresión lineal es una herramienta adecuada para este tipo de predicciones.

El dashboard final integra tanto los resultados descriptivos como los predictivos, facilitando la visualización de las relaciones más importantes y permitiendo interpretar los modelos de manera clara y accesible. Esto responde exitosamente al objetivo inicial del proyecto: **analizar el comportamiento de los vehículos eléctricos y generar predicciones útiles y visualmente comprensibles**.

## Futuras Líneas de Trabajo

- **Mejoras en la calidad y estructura de los datos**

Incluir datos más recientes y equilibrados entre regiones o tipos de vehículo para mejorar la representatividad del modelo. Registrar más variables relacionadas con el uso real del vehículo (clima, velocidad, hábitos de conducción). Reducir valores desconocidos o categorías ambiguas como “Desconocido” en vehículo o región.

- **Mejoras en los modelos predictivos**

Probar algoritmos más avanzados como Random Forest, Gradient Boosting o XGBoost, que pueden capturar relaciones no lineales y mejorar el rendimiento del modelo de autonomía. Implementar técnicas de

regularización (Ridge, Lasso) para evitar sobreajuste y mejorar la estabilidad del modelo.

Realizar ajuste de hiperparámetros (Grid Search o Random Search) para optimizar el desempeño de ambos modelos.

- **Mejoras en las visualizaciones**

Añadir gráficos de evolución temporal si en futuras versiones del dataset se incluyen datos por mes o año de operación. Incorporar indicadores comparativos entre modelos (por ejemplo, barras con  $R^2$  de varios algoritmos).

Agregar más interactividad al dashboard, como resaltar puntos atípicos, filtros más avanzados o segmentación por país.

# Referencias Bibliográficas

- <https://www.kaggle.com/datasets/khushikyad001/electric-vehicle-analytics-dataset>
- <https://laopinion.com/2025/06/25/contaminan-mas-los-electricos-habla-el-lider-de-toyota/>
- <https://theicct.org/publication/electric-cars-life-cycle-analysis-emissions-europe-jul25/>
- <https://pmc.ncbi.nlm.nih.gov/articles/PMC9171403/m>
- <https://www.sciencedirect.com/science/article/pii/S1364032122000867>
- <https://ijtech.eng.ui.ac.id/article/view/7347>
- <https://core.ac.uk/download/pdf/30042271>
- <https://youtu.be/pwJuFbyhZFE?si=xDnt-cH88Rlkewpr>