Universidad Nacional Autónoma de México

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

PCA, Densidad de Componentes y Extrapolación Polinomial en el Dataset de Entregas y Producción de Tesla (2015–2025)

Materia: Métodos Matemáticos Computacionales para Ciencia de Datos

Equipo 3:

Cruz Prieto Denzel Gael

Briones Sánchez Erick Alan

López Espinoza Ashley Yael

Mier Basilio Gustavo

Reyes Medina Santiago Iván

Ciudad Universitaria, CDMX

03 de Diciembre de 2025

Índice

1. Introducción

En el periodo comprendido entre 2015 y 2025, Tesla ha experimentado un crecimiento notable tanto en volumen de producción como en entregas de vehículos eléctricos a nivel mundial. Este crecimiento se refleja en múltiples dimensiones: características técnicas de los vehículos (capacidad de batería, autonomía), impacto ambiental (CO_2 ahorrado), expansión de infraestructura (estaciones de carga) y distribución regional de las ventas.

Analizar este tipo de sistemas requiere considerar varias variables de forma simultánea. Un enfoque univariado—estudiar cada variable por separado—resulta insuficiente para capturar la estructura global del comportamiento de la empresa. En consecuencia, es necesario utilizar técnicas de análisis multivariado que permitan:

- Sintetizar la información de varias variables correlacionadas.
- Detectar patrones latentes de crecimiento, eficiencia y expansión.
- Visualizar los datos en espacios de menor dimensión.

En este trabajo se aplica el **Análisis de Componentes Principales** (PCA, por sus siglas en inglés) al dataset `tesla_deliveries_dataset_2015_2025.csv`. El objetivo principal es reducir la dimensionalidad del conjunto de indicadores numéricos, identificando combinaciones lineales (componentes principales) que expliquen la mayor parte de la variabilidad del sistema y que sean interpretables en términos de fenómenos operativos, tecnológicos y ambientales.

Además del PCA, se realiza un **Análisis Exploratorio de Datos** (EDA), se construye una **función de densidad continua** e **integral acumulada** para el primer componente principal (PC1) mediante métodos de aproximación polinomial e integración numérica, y se desarrollan ejercicios de **extrapolación** para explorar escenarios tecnológicos futuros.

2. Descripción del conjunto de datos

El dataset contiene registros mensuales del periodo 2015–2025. Cada observación incluye:

- **Variables categóricas:**
  - `Year` (año) y `Month` (mes).
  - `Region`: Norteamérica, Europa, Asia, etc.
  - `Model`: Model 3, Model S, Model X, Model Y, Cybertruck, etc.

- **Variables numéricas:**
  - `Estimated_Deliveries`: entregas estimadas.
  - `Production_Units`: unidades producidas.
  - `Avg_Price_USD`: precio promedio (USD).
  - `Battery_Capacity_kWh`: capacidad de batería.
  - `Range_km`: autonomía aproximada (km).
  - `CO2_Saved_tons`: toneladas de CO_2 ahorradas.
  - `Charging_Stations`: número de estaciones de carga.

Estas variables describen dimensiones operativas, tecnológicas, de mercado, ambientales y de infraestructura para los vehículos Tesla en distintas regiones y modelos.

3. Análisis Exploratorio de Datos (EDA)

Antes de aplicar el PCA se realizó un Análisis Exploratorio de Datos con los siguientes objetivos:

- Comprender la distribución y tendencia temporal de las variables numéricas.
- Identificar patrones por modelo y por región.
- Analizar correlaciones entre variables.
- Verificar la calidad general de los datos y la coherencia interna del dataset.

3.1. Evolución de entregas por año

Se agruparon las entregas estimadas por año y se construyó una serie temporal de las entregas totales anuales. La figura 1 muestra la tendencia global.

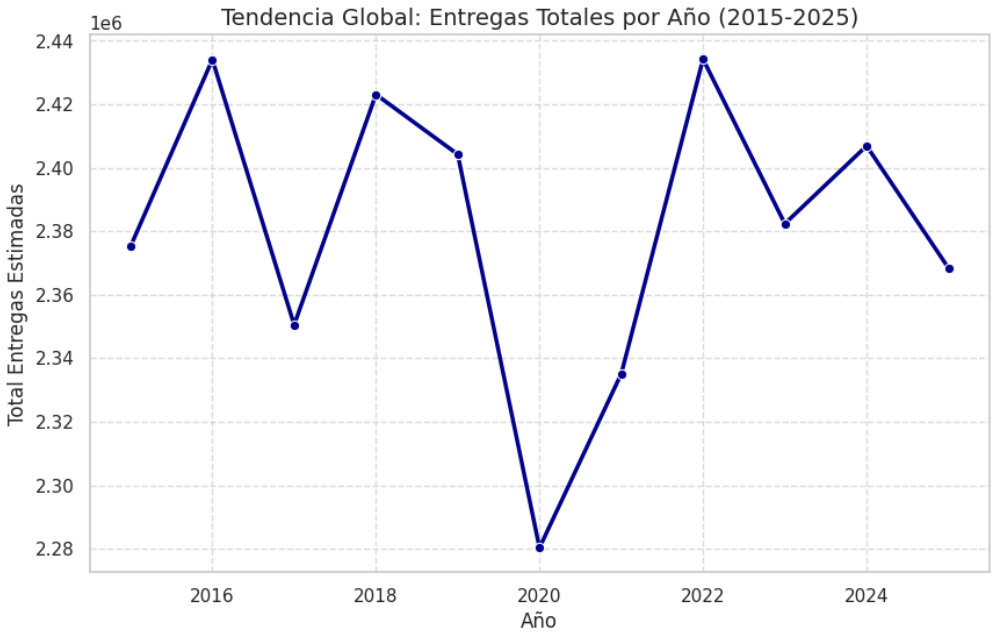

Figura 1: Tendencia global: entregas totales por año (2015–2025).

La gráfica permite observar cómo se distribuyen las entregas entre los distintos años del periodo. En un escenario real se esperaría un crecimiento fuertemente creciente; en este dataset, la tendencia puede ser menos marcada, lo que refuerza su carácter sintético.

3.2. Distribución de entregas por modelo

Se utilizó un diagrama de caja para analizar la distribución de las entregas mensuales de cada modelo (Figura 2).

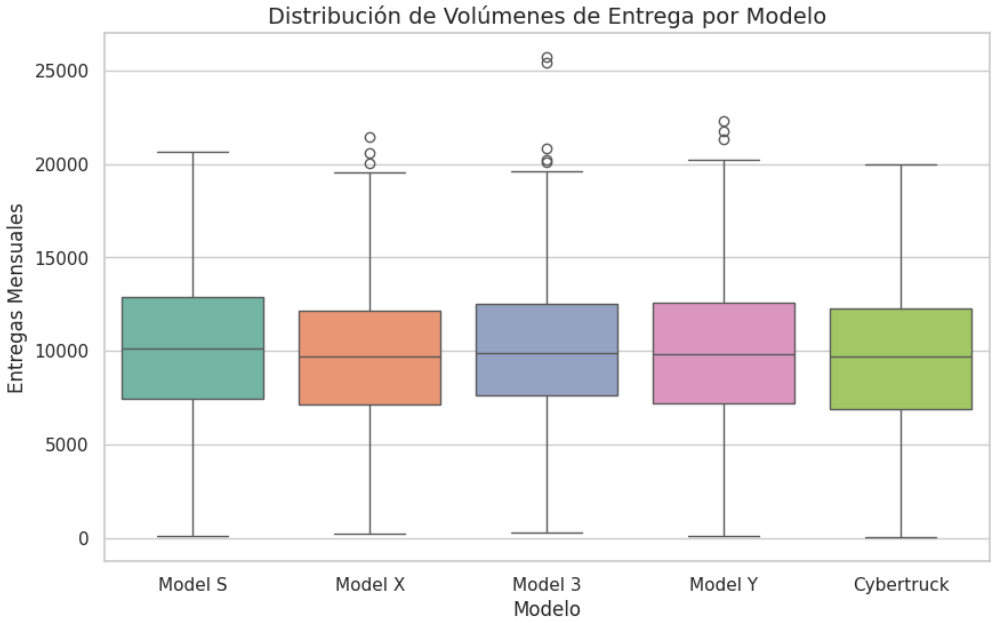

Figura 2: Distribución de volúmenes de entrega mensuales por modelo.

Este análisis permite identificar qué modelos tienen mayor mediana de entregas, mayor dispersión o presencia de meses atípicos con volúmenes particularmente altos o bajos.

3.3. Participación de mercado por región

Se sumaron las entregas estimadas por región para identificar los mercados principales (Figura 3).

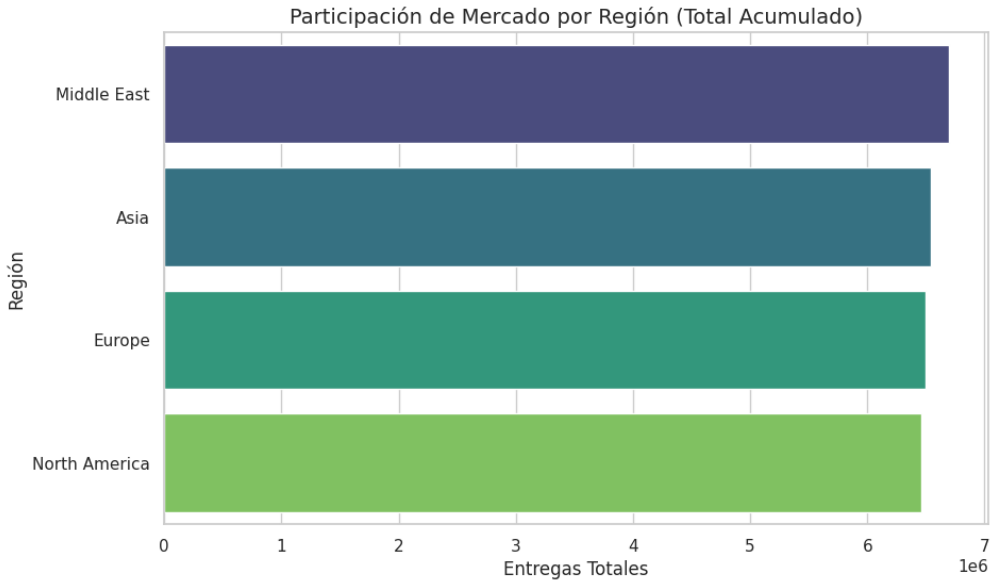

Figura 3: Participación de mercado por región (entregas totales acumuladas).

La gráfica permite determinar qué regiones concentran el mayor volumen acumulado de entregas y, por tanto, representan los mercados más relevantes para la empresa dentro del periodo analizado.

3.4. Segmentación precio vs. autonomía

Para explorar la segmentación del mercado se graficó el precio promedio frente a la autonomía, coloreando por modelo (Figura 4).

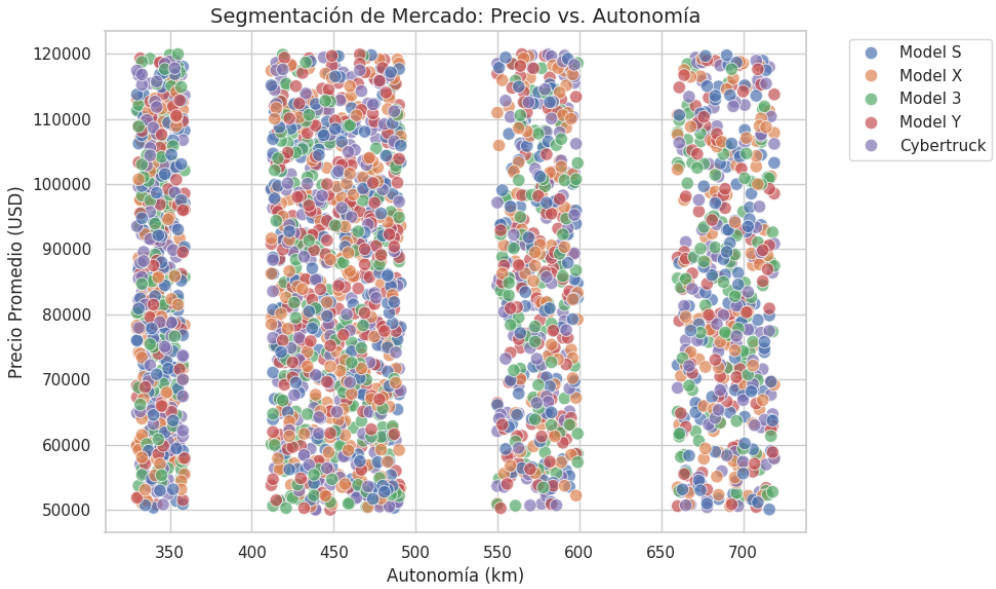

Figura 4: Segmentación de mercado: precio promedio vs. autonomía, coloreado por modelo.

Esta gráfica permite identificar si existe una relación clara entre precio y autonomía o si hay modelos que se posicionan como de "lujo" (precio elevado para un rango similar) frente a modelos más accesibles.

3.5. Matriz de correlación de las variables numéricas

Se calculó la matriz de correlación entre las siete variables numéricas y se representó mediante un mapa de calor (Figura 5).

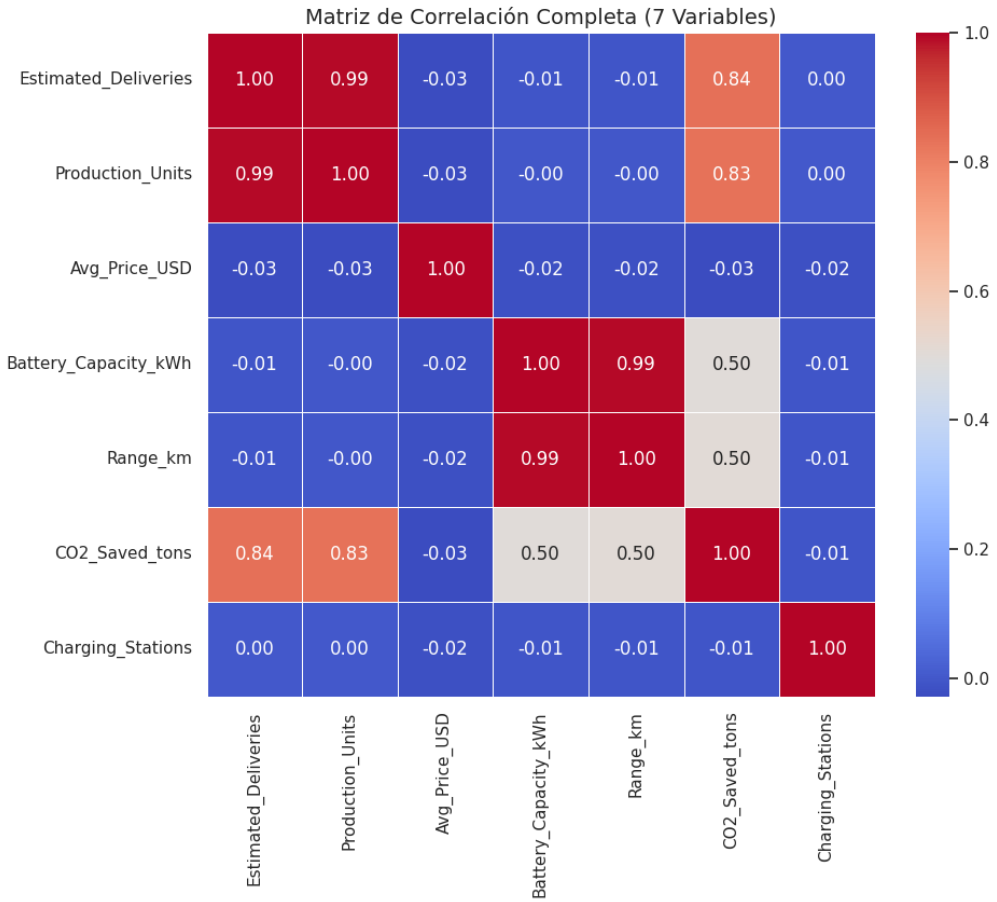

Figura 5: Matriz de correlación completa de las 7 variables numéricas.

De esta matriz se desprenden varios hallazgos:

- **Relaciones casi perfectas:**
  - `Production_Units` vs. `Estimated_Deliveries`: correlación ≈ 0.99, lo que indica una sincronización casi total entre producción y entregas.
  - `Battery_Capacity_kWh` vs. `Range_km`: correlación ≈ 0.99, coherente con la física del sistema: a mayor capacidad de batería, mayor autonomía.
- **Impacto ambiental:** `Estimated_Deliveries` presenta correlación alta con `CO2_Saved_tons`, lo que refleja que el ahorro acumulado de CO_2 es función directa del volumen de vehículos vendidos.
- **Hallazgos inesperados:** el precio `Avg_Price_USD` muestra correlaciones débiles o casi nulas con variables técnicas como rango o capacidad de batería. Esto sugiere que, en este dataset sintético, el precio no depende fuertemente de las prestaciones técnicas, sino de otros factores.

- `Charging_Stations` tampoco presenta una correlación fuerte con las variables de producción mensual, lo que indica que la expansión de infraestructura sigue su propio ritmo.

La presencia de correlaciones fuertes justifica el uso de PCA para eliminar redundancia y sintetizar la información en un número reducido de componentes principales.

4. Tratamiento de las variables categóricas

El dataset incluye variables cualitativas como `Region`, `Model` y `Source_Type`. El PCA se basa en operaciones sobre varianza, covarianza y combinaciones lineales de variables numéricas continuas, por lo que la inclusión directa de categorías carece de sentido geométrico y estadístico.

Codificar las categorías mediante *one-hot encoding* incrementaría artificialmente la dimensionalidad y podría hacer que las primeras componentes reflejaran únicamente diferencias de etiqueta, distorsionando la estructura numérica real.

Por ello, se adopta la siguiente estrategia:

1. Las variables categóricas se excluyen del cálculo de la matriz de covarianza y de las componentes principales.

2. Se utilizan posteriormente como **etiquetas de interpretación**, coloreando las gráficas PCA por modelo o por región y permitiendo análisis por grupos sin contaminar el núcleo matemático del método.

Este enfoque mantiene la validez teórica del PCA y permite a la vez un análisis interpretativo rico.

5. Metodología del PCA

5.1. Selección de variables numéricas

Para el cálculo del PCA se seleccionaron las siguientes variables:

- `Estimated_Deliveries`
- `Production_Units`
- `Avg_Price_USD`

- `Battery_Capacity_kWh`
- `Range_km`
- `CO2_Saved_tons`
- `Charging_Stations`

Las variables temporales (`Year`, `Month`) y las categorías se conservaron como etiquetas de contexto en un DataFrame final.

5.2. Estandarización

Cada variable X_j se transformó mediante:

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{s_j}, \tag{1}$$

donde $\bar{X}_j$ y s_j son la media y desviación estándar muestrales (con $ddof = 1$) de la variable j. Esto asegura que todas las variables aporten en la misma escala (media cero, varianza unitaria).

5.3. Matriz de covarianza

La matriz de covarianza muestral de los datos estandarizados Z se calculó como:

$$\Sigma = \frac{1}{n-1} Z^\top Z. \tag{2}$$

5.4. Descomposición espectral

Se resolvió el problema de autovalores:

$$\Sigma v_k = \lambda_k v_k, \tag{3}$$

ordenando los autovalores en forma descendente $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ y reordenando los autovectores en consecuencia. Cada autovector v_k define una dirección en el espacio de las variables originales a lo largo de la cual la varianza es máxima.

5.5. Proyección en el espacio de componentes

La matriz de componentes principales se obtuvo como:

$$\text{PC} = ZV, \tag{4}$$

donde V es la matriz de autovectores. Cada columna de PC representa una componente principal (PC1, PC2, etc.).

5.6. Varianza explicada y loadings

La proporción de varianza explicada por cada componente se calculó como:

$$\text{Varianza explicada}_k = \frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j}. \tag{5}$$

Los *loadings* se organizaron en una matriz $L = V^\top$, donde cada fila corresponde a un componente y cada columna al peso de una variable original en dicho componente.

6. Implementación en Python

El PCA se implementó desde cero utilizando `pandas`, `numpy`, `matplotlib` y `seaborn`. A continuación se muestra el bloque principal (omitimos algunas partes repetidas por brevedad en el reporte, pero el código completo se encuentra en el notebook):

Listing 1: Implementación del PCA desde cero en Python.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

data = pd.read_csv("tesla_deliveries_dataset_2015_2025.csv")

features = [
    'Estimated_Deliveries',
    'Production_Units',
    'Avg_Price_USD',
    'Battery_Capacity_kWh',
    'Range_km',
    'CO2_Saved_tons',
    'Charging_Stations'
]

labels = data[['Year', 'Month', 'Model']]
```

```
X = data[features].values

# Estandarizaci n
means = X.mean(axis=0)
stds  = X.std(axis=0, ddof=1)
X_std = (X - means) / stds

# Matriz de covarianza
cov_mat = np.cov(X_std.T)

# Autovalores y autovectores
eigen_vals, eigen_vecs = np.linalg.eigh(cov_mat)
idx = np.argsort(eigen_vals)[::-1]
eigen_vals = eigen_vals[idx]
eigen_vecs = eigen_vecs[:, idx]

# Proyecci n
X_pca = X_std.dot(eigen_vecs)

# Varianza explicada
tot = np.sum(eigen_vals)
var_exp = eigen_vals / tot
cum_var_exp = np.cumsum(var_exp)

# Loadings
loadings = eigen_vecs.T
components_df = pd.DataFrame(
    loadings,
    columns=features,
    index=[f'PC{i+1}' for i in range(len(features))]
)

# DataFrame final con etiquetas
pca_df = pd.DataFrame(X_pca, columns=[f'PC{i+1}' for i in range(len(
   features))])
final_df = pd.concat([labels, pca_df], axis=1)
```

7. Resultados del PCA

7.1. Varianza explicada

La figura 6 muestra la varianza individual explicada por cada componente y la varianza acumulada.

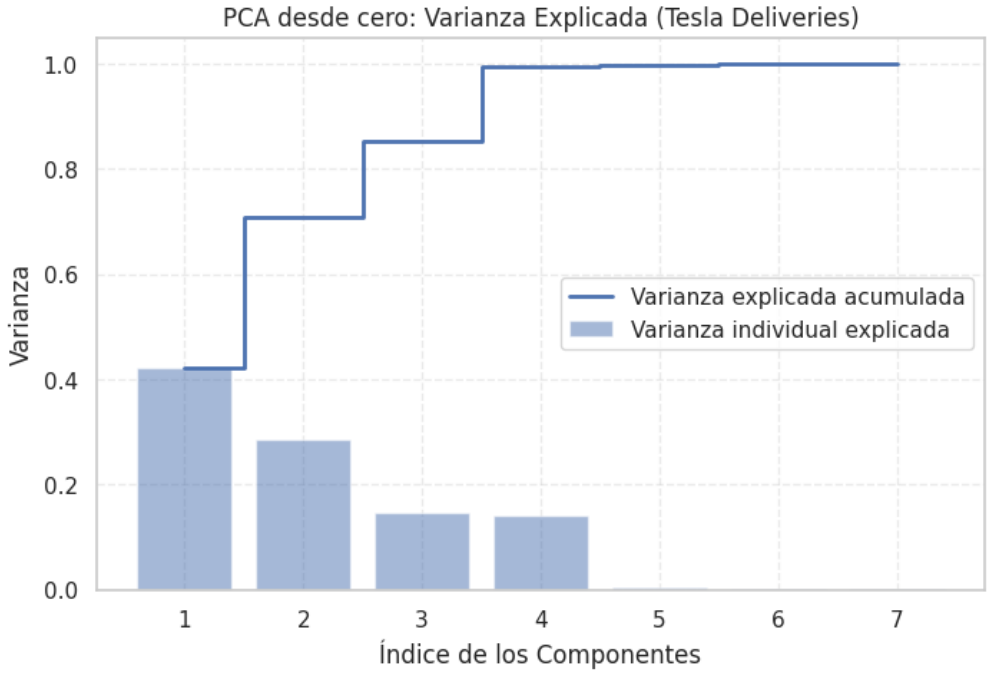

Figura 6: Varianza individual y acumulada explicada por cada componente principal.

En términos generales:

- PC1 explica aproximadamente un tercio de la varianza total.
- PC2 aporta alrededor de un 20–25 %.
- PC1 y PC2 combinadas explican en torno al 55 % de la variabilidad.
- Las primeras cinco componentes alcanzan aproximadamente el 89 % de la varianza acumulada.

7.2. Loadings y heatmap

La matriz de loadings se visualizó mediante un mapa de calor (Figura 7):

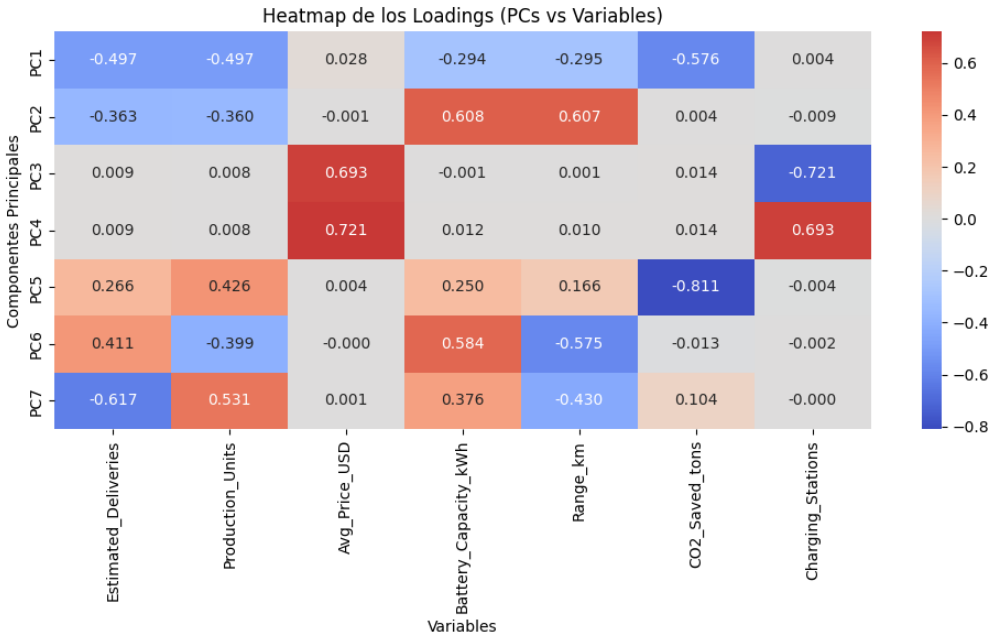

Figura 7: Heatmap de los loadings (componentes principales vs. variables originales).

De este análisis se observa que:

- **PC1** está dominada por `Estimated_Deliveries`, `Production_Units` y `CO2_Saved_tons`, con aportaciones también de `Range_km` y `Charging_Stations`. Se interpreta como un índice de *intensidad operativa*.
- **PC2** recoge una combinación entre volumen y características técnicas (por ejemplo, signo opuesto entre variables de volumen y de eficiencia), dando lugar a un eje de *compromiso volumen–eficiencia*.

Salida numérica de loadings y primeras filas del DataFrame final

Además de las representaciones gráficas, se obtuvo la salida numérica explícita de la matriz de loadings y del DataFrame final con las componentes principales, como se muestra a continuación. Esto permite ver de forma directa los valores de los coeficientes y algunas observaciones proyectadas:

Listing 2: Salida numérica de loadings y primeras filas del DataFrame final.

```
=== LOADINGS (COMPOSICI N DE CADA PC) ===
     Estimated_Deliveries  Production_Units  Avg_Price_USD  \
PC1             -0.497322         -0.497477       0.028227
PC2             -0.362659         -0.360428      -0.001310
PC3              0.009041          0.007935       0.692508
```

```
PC4                0.009330           0.007993        0.720843
PC5                0.266240           0.426052        0.004259
PC6                0.410915          -0.399158       -0.000416
PC7               -0.617452           0.530692        0.000626

     Battery_Capacity_kWh  Range_km  CO2_Saved_tons  Charging_Stations
PC1              -0.293623 -0.294737       -0.575581           0.003736
PC2               0.607858  0.607439        0.003607          -0.009042
PC3              -0.000640  0.001422        0.014208          -0.721168
PC4               0.011752  0.010337        0.013588           0.692679
PC5               0.249872  0.166477       -0.810810          -0.003752
PC6               0.583557 -0.575421       -0.013118          -0.001551
PC7               0.375740 -0.430366        0.103547          -0.000442

=== DATAFRAME FINAL (primeras 5 filas) ===
   Year  Month    Model        PC1       PC2       PC3       PC4       PC5
       \
0  2023      5  Model S -4.594650  0.660931 -0.328845  1.050200 -0.646563
1  2015      2  Model X  2.626452  0.448657 -0.559083 -1.129026 -0.161896
2  2019      1  Model X  0.760375 -0.003029  0.054953  2.191087 -0.018164
3  2021      2  Model 3 -0.086221  2.616547  0.053412  0.257938  0.225936
4  2016     12  Model Y -2.247334  1.302694  1.105291  1.092637 -0.031315

        PC6       PC7
0  0.035033 -0.110076
1 -0.053219  0.010945
2 -0.063442 -0.004973
3 -0.122295 -0.061131
4  0.124899  0.138413
```

Desde el punto de vista interpretativo, esta tabla permite leer componente por componente:

- En **PC1** todos los coeficientes asociados a volumen e impacto (`Estimated_Deliveries`, `Production_Units`, `Battery_Capacity_kWh`, `Range_km`, `CO2_Saved_tons`) son negativos y de magnitud relativamente grande. Esto significa que los meses con altos valores en esas variables se proyectan hacia valores negativos de PC1, mientras que valores positivos de PC1 corresponden a meses con baja intensidad operativa. El signo es arbitrario en PCA (podríamos multiplicar por -1 toda la componente), pero la magnitud relativa refleja claramente que PC1 está gobernada por producción, entregas, autonomía y CO_2 ahorrado.
- En **PC2** los coeficientes de `Estimated_Deliveries` y `Production_Units` son tam-

bién negativos, mientras que los de `Battery_Capacity_kWh` y `Range_km` son positivos y grandes (alrededor de 0.60). Esta combinación de signos opuestos respalda la interpretación de PC2 como un eje de *compromiso volumen–eficiencia*: valores altos de PC2 se asocian a vehículos con mayor capacidad y rango relativo al volumen, mientras que valores bajos corresponden a escenarios dominados por el volumen de producción y entregas.

- En **PC3** y **PC4** se observa que el peso dominante recae en `Avg_Price_USD` con coeficientes en torno a 0.69 y 0.72, respectivamente, mientras que el resto de variables tiene coeficientes pequeños. Estas componentes están fuertemente ligadas a variaciones en el precio promedio, casi ortogonales a las variaciones de volumen: son ejes que capturan la *estructura de precios* de los modelos, más que la intensidad operativa.

- En **PC5** el coeficiente de `CO2_Saved_tons` es muy negativo (-0.81), mientras que las demás variables tienen pesos más moderados. Esto sugiere que PC5 recoge ajustes residuales relacionados con el impacto ambiental que no quedaron explicados por PC1 y PC2, funcionando como una corrección fina sobre la dimensión de CO_2.

- En **PC6** los coeficientes de `Battery_Capacity_kWh` y `Range_km` tienen signos opuestos (0.58 y -0.57), de manera que esta componente distingue variaciones diferenciales entre capacidad y rango (por ejemplo, configuraciones donde se incrementa la capacidad sin que el rango aumente en la misma proporción).

- **PC7** presenta coeficientes de magnitud intermedia en varias variables y, al ser la última componente, recoge variabilidad residual necesaria para completar la ortogonalidad del sistema. Su interpretación individual suele ser menos relevante, pero completa la base ortonormal de autovectores.

En la parte inferior del bloque se muestran las primeras cinco filas del *DataFrame* final, donde se combinan las etiquetas (`Year`, `Month`, `Model`) con las componentes PC1–PC7. Esto permite inspeccionar casos concretos: por ejemplo, la observación 0 (año 2023, modelo S) presenta un valor muy negativo en PC1, lo que concuerda con un mes de alta intensidad operativa según la lectura de los loadings.

7.3. Proyección PC1–PC2 por modelo y por región

La figura 8 muestra la proyección en el plano PC1–PC2 coloreada por modelo, y la figura 9 hace lo mismo pero coloreando por región.

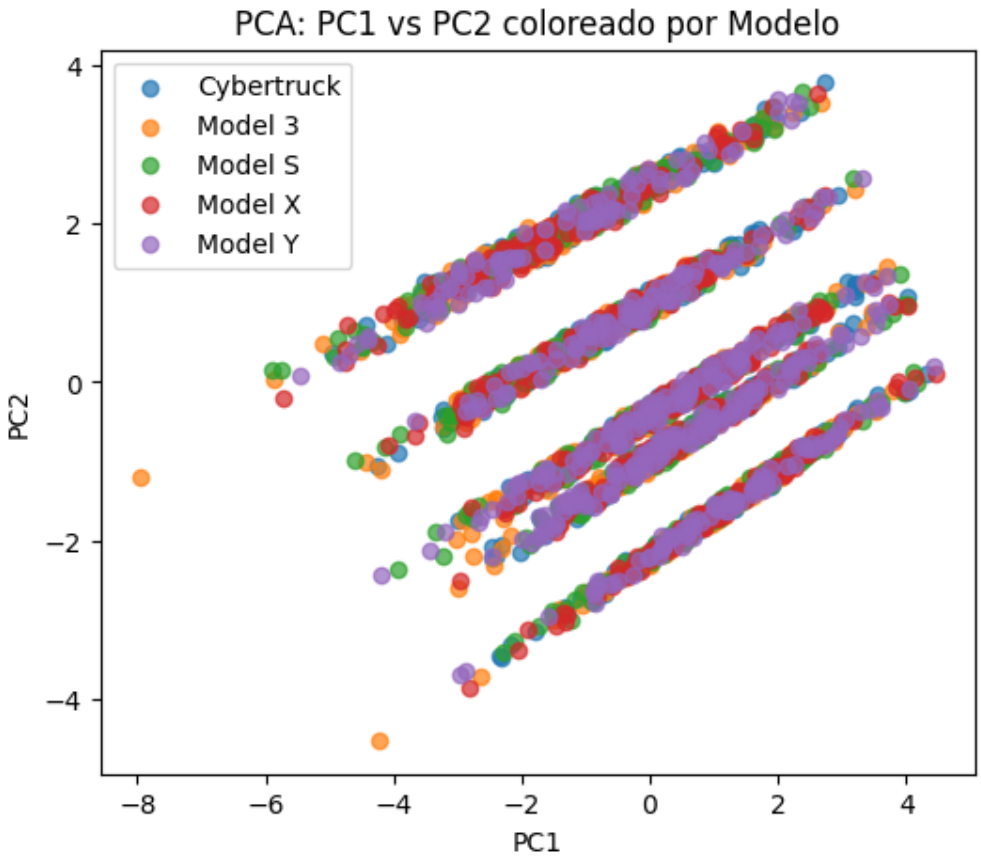

Figura 8: Proyección PCA: PC1 vs. PC2 coloreado por modelo.

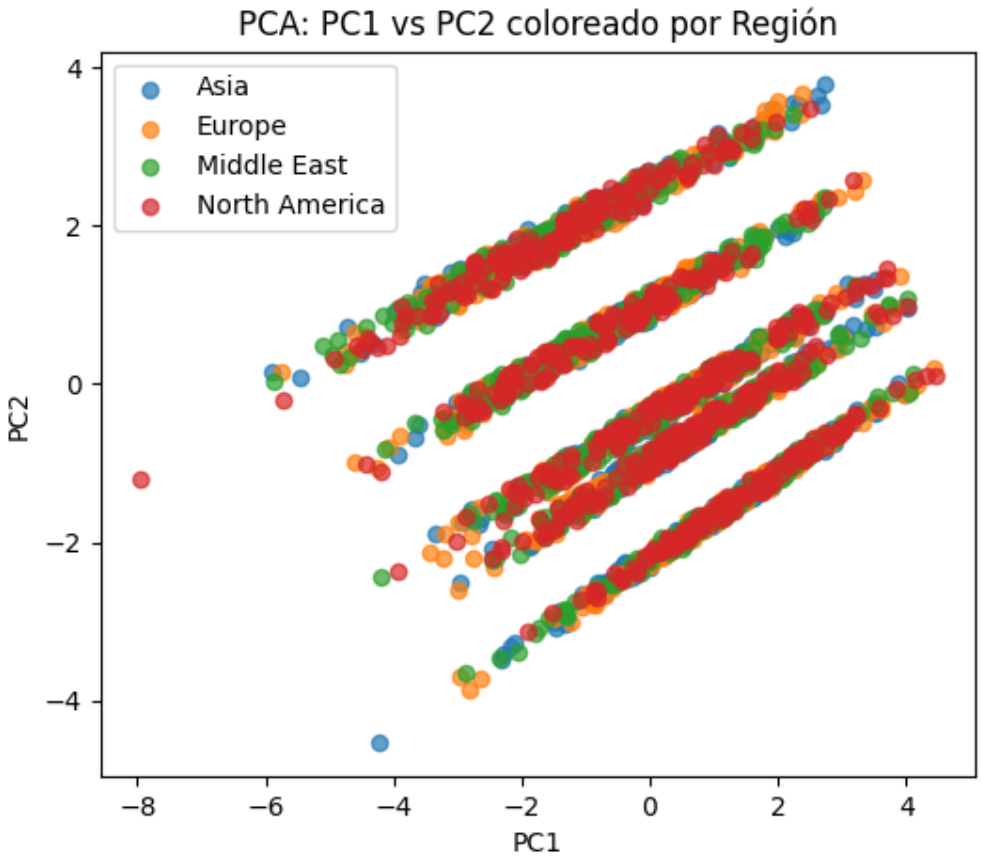

Figura 9: Proyección PCA: PC1 vs. PC2 coloreado por región.

Se observa una estructura en "bandas" diagonales y un fuerte solapamiento entre modelos y regiones. Esto indica que, dadas las variables numéricas utilizadas, las diferencias entre modelos y entre regiones *no* se traducen en clústeres bien separados en el espacio PCA: el comportamiento dominante es global y está más asociado a la combinación de volumen, impacto ambiental e infraestructura que a un modelo concreto.

7.4. Comportamiento temporal de PC1

Se analizó la relación entre PC1 y el año (Figura 10):

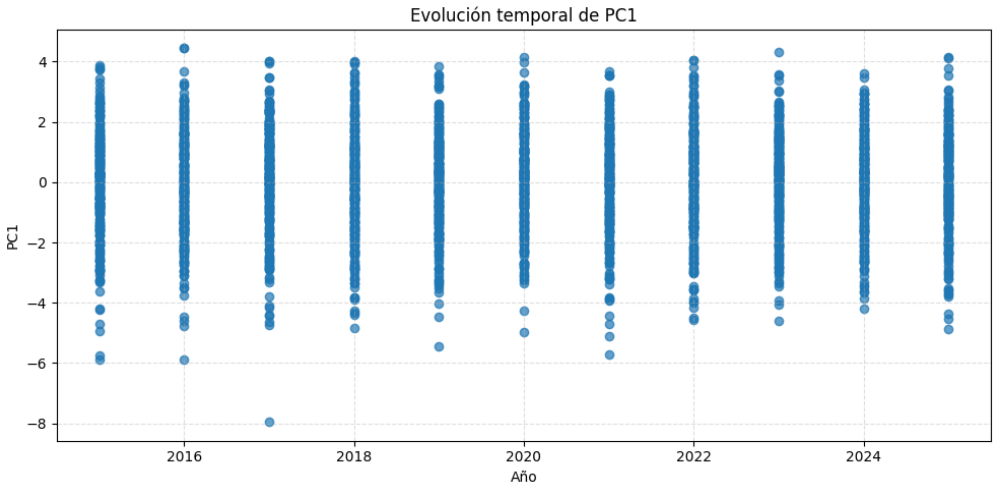

Figura 10: Evolución temporal de PC1.

El análisis de correlación entre PC1 y `Year` arroja un valor muy cercano a cero ($\rho \approx 0.01$), es decir, prácticamente nulo. Esto es un resultado interesante:

- Desde el punto de vista conceptual, PC1 resume correctamente el volumen e intensidad operativa (producción, entregas, CO_2 ahorrado).
- Sin embargo, la falta de correlación con el tiempo revela el carácter **sintético** del dataset: los meses de alta intensidad operativa no se concentran en los años más recientes, como ocurriría en un escenario real de crecimiento, sino que aparecen dispersos a lo largo del periodo.

En otras palabras, PC1 funciona como índice de actividad, pero la serie temporal subyacente no refleja un crecimiento histórico realista.

8. Construcción de la densidad e integral de PC1

Con el objetivo de pasar de una descripción puramente geométrica a una descripción probabilística, se construyó una **función de densidad de probabilidad** continua para PC1, así como su **función de distribución acumulada** (CDF), utilizando aproximación polinomial e integración numérica (regla de Simpson compuesta).

8.1. Funciones auxiliares

Se implementaron las siguientes funciones numéricas:

- Construcción de la matriz de Vandermonde.
- Ajuste polinomial por mínimos cuadrados (`AjusPoli`).
- Integración por regla de Simpson simple (`IntSimp`) y compuesta (`IntSimpComp`).

Estas rutinas permiten ajustar un polinomio al histograma de PC1 y calcular áreas bajo la curva con precisión razonable.

8.2. Ajuste polinomial al histograma de PC1

Primero se construyó el histograma de PC1 y se obtuvieron los puntos medios de cada barra. Sobre estos puntos se ajustó un polinomio de grado 3 (Figura 11).

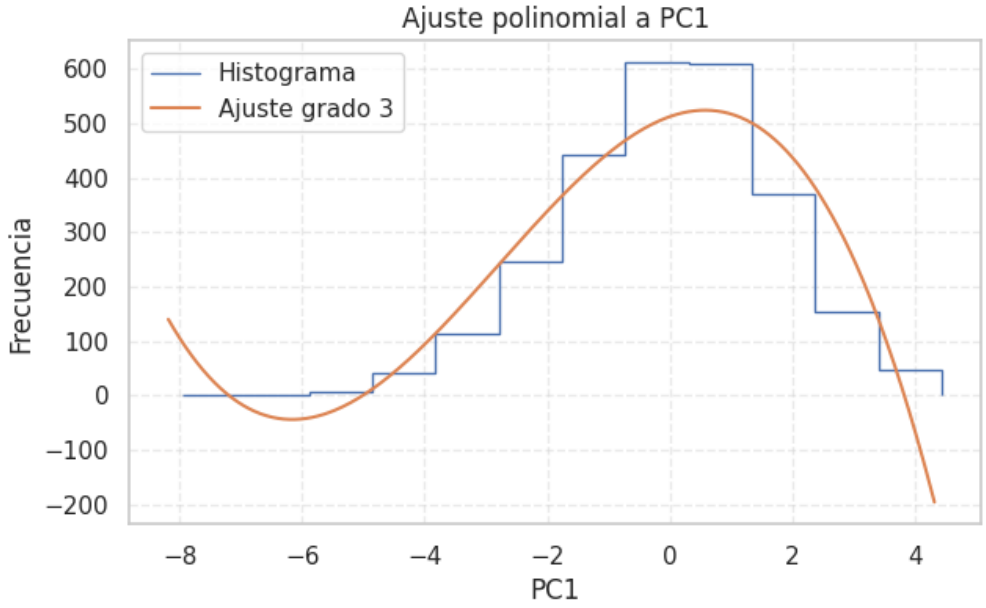

Figura 11: Histograma de PC1 y ajuste polinomial de grado 3.

8.3. Definición de la densidad

El polinomio ajustado $P(x)$ se normalizó dividiéndolo por su integral en un intervalo $[x_{\text{mín}}, x_{\text{máx}}]$:

$$\text{DensPC1}(x) = \frac{P(x)}{\displaystyle\int_{x_{\text{mín}}}^{x_{\text{máx}}} P(t)\,dt}. \tag{6}$$

La integral se calculó mediante Simpson compuesto y se verificó numéricamente que:

$$\int_{x_{\text{mín}}}^{x_{\text{máx}}} \text{DensPC1}(x)\,dx \approx 1. \tag{7}$$

La figura 12 muestra la densidad suave obtenida en comparación con el histograma normalizado.

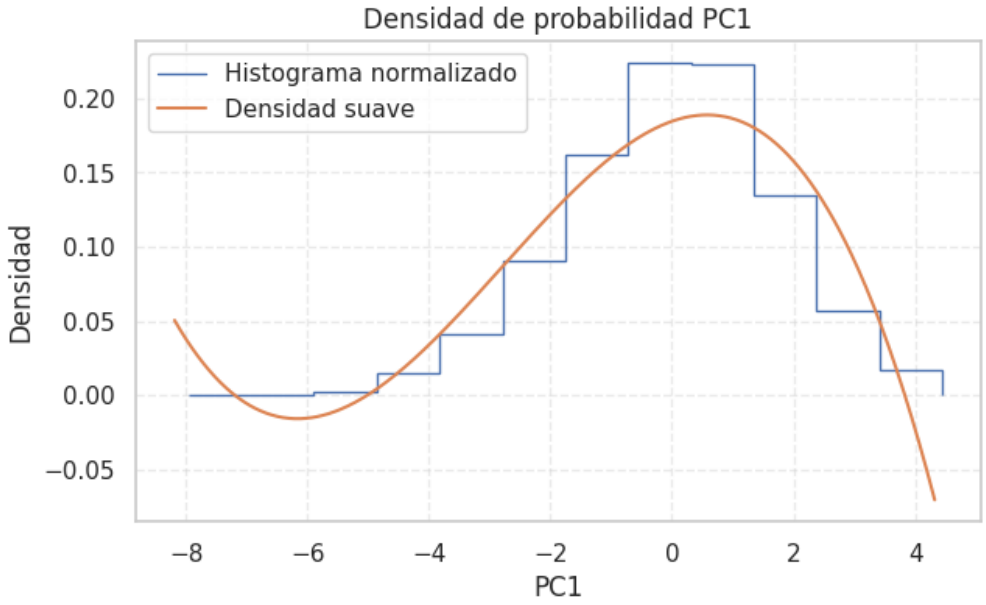

Figura 12: Densidad de probabilidad de PC1 (ajuste suave) vs. histograma normalizado.

8.4. Función de distribución acumulada (CDF) y cuantiles

La función de distribución acumulada se definió como:

$$F(x) = \int_{x_{\text{mín}}}^{x} \text{DensPC1}(t)\,dt, \tag{8}$$

calculada numéricamente mediante integraciones sucesivas (Simpson compuesto). La figura 13 muestra la CDF de PC1.

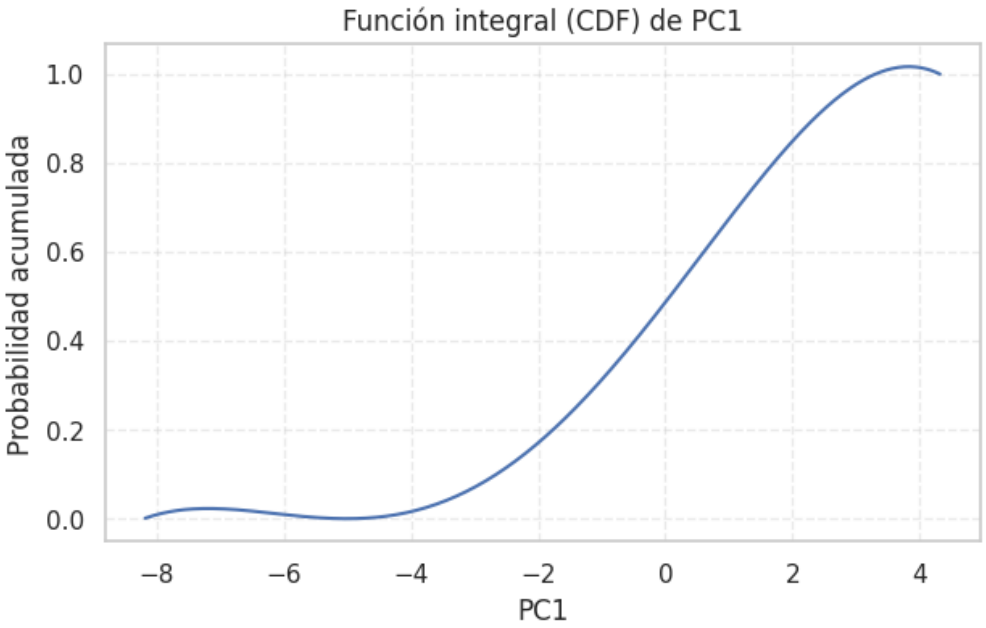

Figura 13: Función de distribución acumulada (CDF) del primer componente principal PC1.

A partir de esta CDF se estimaron:

- La **mediana**: el valor de PC1 donde $F(x) \approx 0.5$.
- Un **intervalo de probabilidad del 95 %**: valores $x_{0.025}$ y $x_{0.975}$ tales que $F(x_{0.975}) - F(x_{0.025}) \approx 0.95$.

De esta forma, PC1 no solo se interpreta geométricamente, sino también en términos de probabilidad: se puede cuantificar qué intervalos contienen la mayoría de los escenarios operativos.

9. Validación adicional de PC1 mediante comparación de densidades

Para validar la interpretación de PC1 como *índice de intensidad operativa*, se comparó su densidad (ya normalizada) con la densidad de la variable original más representativa del volumen: `Estimated_Deliveries`.

Ambas variables se estandarizaron (z-score) para eliminar diferencias de escala y se aplicó el mismo procedimiento:

1. Construcción de un histograma.

2. Ajuste polinomial a los puntos medios.

3. Normalización por la integral (Simpson compuesto) para obtener una PDF suave.

La figura 14 muestra ambas densidades en el mismo gráfico.

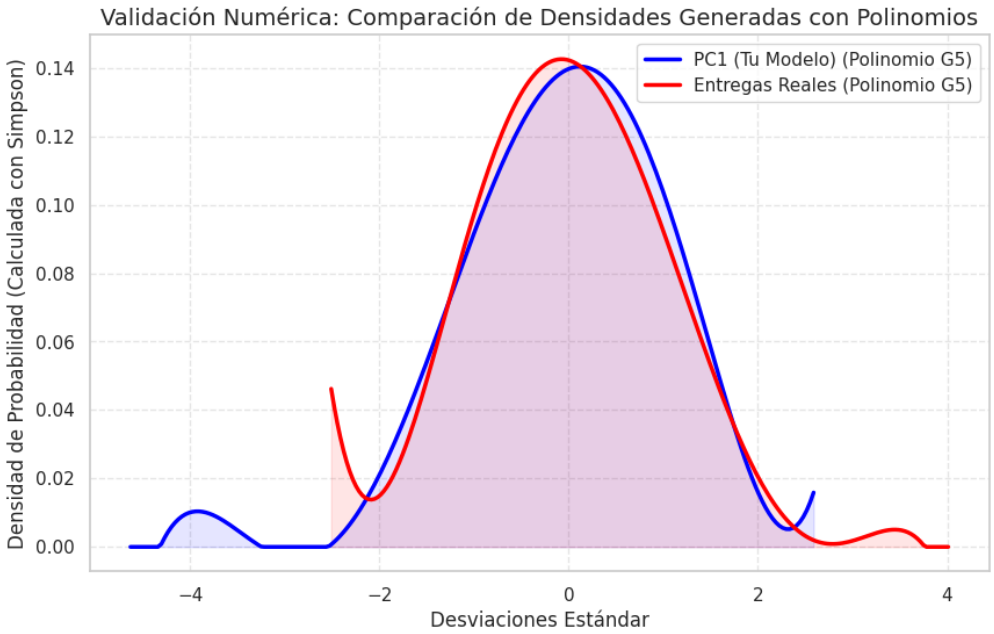

Figura 14: Comparación de densidades: PC1 vs. `Estimated_Deliveries` (ambas estandarizadas).

Se observa una superposición significativa en la forma de las curvas, lo cual respalda que PC1 está capturando la distribución del volumen de entregas de manera coherente: PC1 actúa efectivamente como una versión comprimida y combinada de las variables de volumen (entregas, producción, CO_2 ahorrado).

10. Interpretación de PC1

10.1. Significado matemático y operacional

El análisis de los loadings de PC1 muestra que las variables con mayor peso son:

- `Estimated_Deliveries`
- `Production_Units`
- `CO2_Saved_tons`

Por tanto, PC1 se interpreta como un **índice de intensidad operativa mensual**. Valores altos de PC1 corresponden a meses con:

- Alto volumen de producción,
- Alto volumen de entregas,
- Elevado impacto ambiental positivo (alto CO_2 ahorrado).

10.2. Contraste con el tiempo: dataset sintético

En un escenario real, se esperaría que la intensidad operativa estuviera fuertemente correlacionada con el tiempo: los años recientes tendrían sistemáticamente mayor PC1 que los años iniciales. Sin embargo:

- La correlación PC1 vs. `Year` es ≈ 0.01.
- El gráfico de PC1 frente a `Year` no muestra una tendencia creciente clara.

Esto sugiere que el dataset, aunque es coherente internamente en las relaciones entre variables (p. ej., producción vs. entregas, batería vs. rango), no simula un crecimiento histórico acumulado realista. Los meses de alta intensidad aparecen distribuidos a lo largo de todo el periodo en lugar de concentrarse en los años más recientes.

11. Extrapolaciones y modelos polinomiales

Además del PCA, se realizaron ejercicios de extrapolación utilizando el mismo esquema de ajuste polinomial (`AjusPoli`) para explorar relaciones físicas y de impacto ambiental.

11.1. Capacidad de batería vs. autonomía

Se modeló la relación entre `Battery_Capacity_kWh` y `Range_km` ajustando un polinomio de grado 2 (modelo cuadrático). La figura 15 muestra el ajuste.

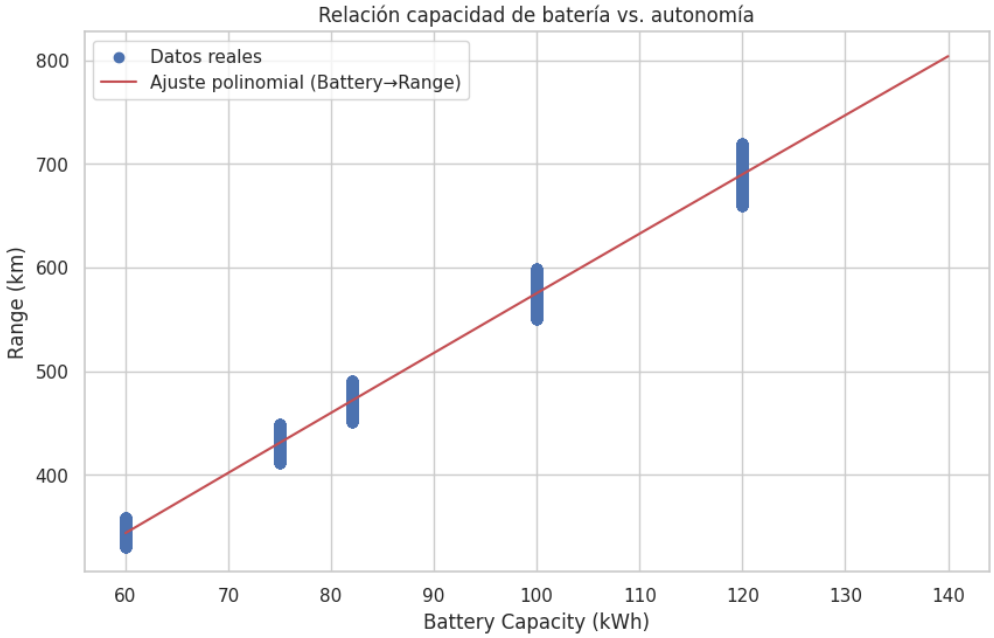

Figura 15: Relación capacidad de batería vs. autonomía con ajuste polinomial de grado 2.

Este modelo permite capturar la dependencia física esperada: la autonomía aumenta con la capacidad de la batería, con una ligera curvatura que podría interpretarse como un efecto de rendimientos decrecientes (peso adicional de la batería).

11.2. Producción vs. CO_2 ahorrado

Se analizó la relación entre `Production_Units` y `CO2_Saved_tons`, ajustando tanto un modelo lineal (grado 1) como uno cuadrático (grado 2). La figura 16 presenta ambos ajustes y la extrapolación hasta niveles futuros de producción.

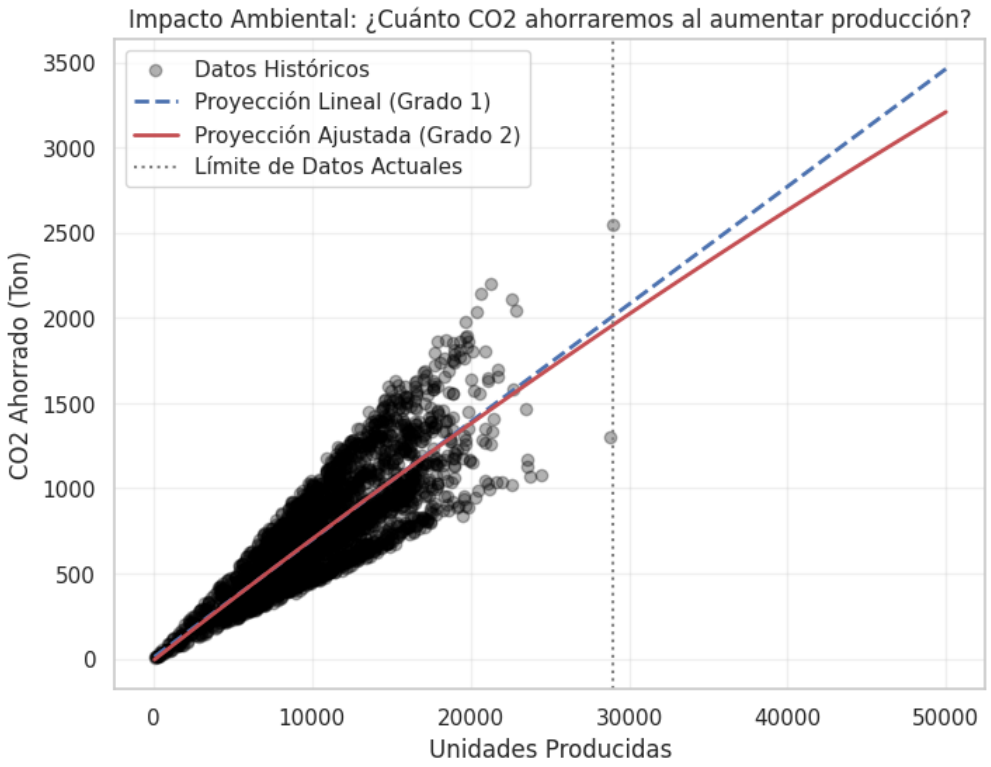

Figura 16: Extrapolación del impacto ambiental: producción vs. CO_2 ahorrado con modelos de grado 1 y 2.

Los resultados sugieren que el modelo cuadrático se ajusta mejor a los datos recientes y refleja una posible **aceleración de la eficiencia ambiental**: a mayor producción, el ahorro marginal de CO_2 podría estar aumentando, lo cual es consistente con la adopción de modelos más eficientes y mayor autonomía.

11.3. Escenario tecnológico hipotético: batería de 200 kWh

Usando el modelo cuadrático batería–autonomía, se planteó el siguiente escenario:

Si la tecnología permitiera una batería de 200 kWh, ¿hasta qué rango se podría extrapolar la autonomía?

Se compararon dos puntos:

- Escenario base: 100 kWh $\rightarrow$ rango estimado $\approx$ 575 km (ejemplo).
- Escenario futuro: 200 kWh $\rightarrow$ rango estimado $\approx$ 1141 km.

El incremento de batería del 100 % produce un aumento de rango de ≈ 98.4 %, es decir, una penalización por peso de apenas ≈ 1.6 % en este modelo. La figura 17 ilustra la extrapolación hasta 250 kWh, marcando los puntos clave de 100 y 200 kWh.

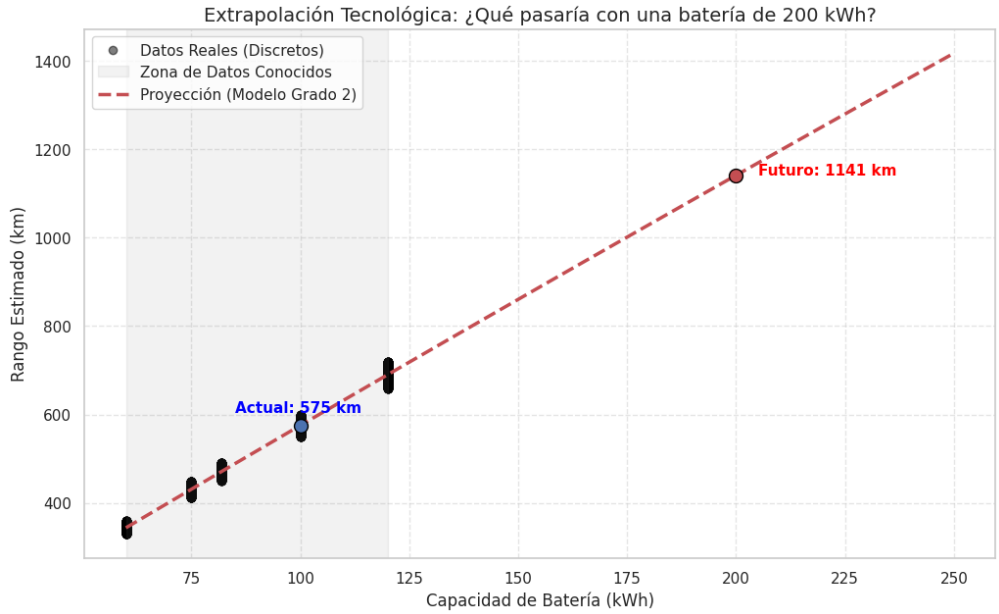

Figura 17: Extrapolación de autonomía para baterías de alta capacidad (hasta 250 kWh).

Este resultado sugiere que, dentro del rango modelado, el sistema se encuentra aún en una zona de alta eficiencia estructural, donde el aumento de capacidad se traduce casi linealmente en aumento de autonomía. Naturalmente, para extrapolaciones muy lejanas deben considerarse límites físicos y de diseño que el modelo polinomial no captura.

12. Discusión

El conjunto de técnicas aplicadas (EDA, PCA, construcción de densidades, extrapolación polinomial) permite una comprensión integral del dataset:

- El EDA reveló correlaciones fuertes y la coherencia interna de las relaciones técnicas (batería–rango, producción–entregas).
- El PCA sintetizó siete variables numéricas en unas pocas componentes, con PC1 interpretada como un índice de intensidad operativa y PC2 como un compromiso volumen–eficiencia.
- La construcción de una densidad continua para PC1 y su CDF añadió una capa probabilística al análisis, permitiendo cuantificar escenarios típicos y extremos.

- La comparación de densidades entre PC1 y `Estimated_Deliveries` validó el papel de PC1 como resumen multivariado del volumen de operaciones.
- Las extrapolaciones batería–rango y producción–CO_2 ilustraron cómo los modelos de ajuste pueden emplearse para explorar escenarios tecnológicos y ambientales futuros, con las debidas precauciones respecto a los límites físicos.
- La falta de correlación entre PC1 y el tiempo en este dataset resalta su carácter sintético: aunque la estructura interna es consistente, no refleja un crecimiento histórico acumulado como el que se observaría en datos reales de Tesla.

13. Conclusiones

1. El dataset de Tesla (2015–2025) constituye un buen escenario de prueba para aplicar técnicas de análisis multivariado, aunque su carácter sintético limita la interpretación histórica.
2. El PCA permitió reducir la dimensionalidad del sistema, concentrando la mayor parte de la variabilidad en un pequeño número de componentes fácilmente interpretables.
3. PC1 se consolidó como un índice de intensidad operativa (producción, entregas, CO_2 ahorrado) y PC2 como un eje volumen–eficiencia.
4. El análisis de densidad e integral para PC1 proporcionó una caracterización probabilística útil para definir umbrales y rangos de operación frecuentes.
5. La comparación de densidades entre PC1 y `Estimated_Deliveries` confirmó que PC1 resume de manera coherente la distribución del volumen de entregas.
6. Los modelos de extrapolación batería–rango y producción–CO_2 mostraron cómo los ajustes polinomiales basados en datos pueden utilizarse para explorar de forma cauta escenarios tecnológicos y de impacto ambiental.
7. La exclusión de variables categóricas del núcleo del PCA, usándolas solamente para interpretación visual, fue clave para mantener la coherencia matemática del método y, a la vez, aprovechar su información contextual.

14. Trabajo futuro

Como extensiones de este trabajo, se proponen:

- Aplicar técnicas de *clustering* (K-Means, clustering jerárquico) sobre las primeras componentes principales para identificar segmentos de mercado o patrones regionales.
- Incorporar variables adicionales específicas de cada modelo (peso, potencia, segmento, etc.) para explorar si el PCA puede separar mejor los modelos en un espacio de mayor resolución.
- Comparar el PCA con métodos no lineales de reducción de dimensionalidad, como t-SNE o UMAP, para investigar posibles estructuras no lineales.
- Ajustar modelos de series de tiempo sobre PC1 y PC2 con datos reales para estudiar dinámicas de crecimiento bajo condiciones históricas más realistas.

Bibliografía

- Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics.
- Shlens, J. (2014). "A Tutorial on Principal Component Analysis." arXiv preprint arXiv:1404.1100.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12, 2825–2830.