CODERHOUSE

**DATA SCIENCE II**

**Profesor: Jorge Ruiz**

**Autor: Lic. Alejandro Meschini**

# ***Informe de Análisis Exploratorio y Preprocesamiento de Datos: Mercado de Autoelevadores***

## 1. Introducción y Objetivo del Proyecto

El presente proyecto tiene como objetivo desarrollar un modelo predictivo para estimar el precio de venta de Autoelevadores en el mercado industrial. En esta primera fase, nos enfocamos en garantizar la calidad del dato (*Data Cleaning*), entender las relaciones entre variables (*EDA*) y preparar el dataset para algoritmos de Machine Learning.

## 2. Tratamiento de Datos y Calidad de la Información

### 2.1 Limpieza e Imputación

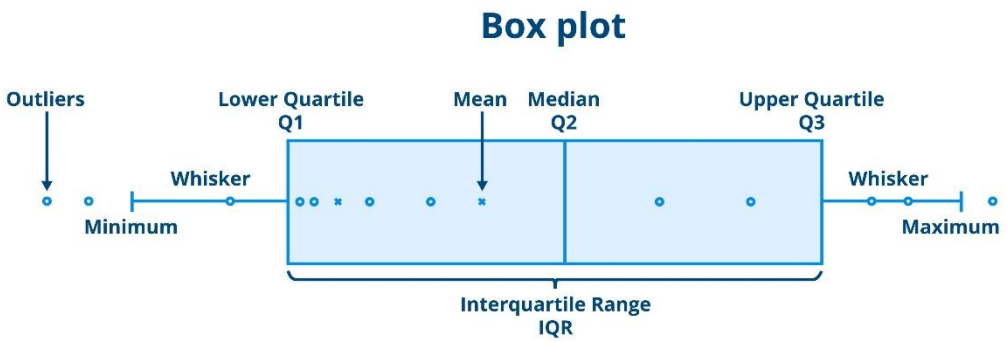
Se identificó una tasa de valores nulos cercana al **8%** en variables críticas.

- **Variables Numéricas:** Se utilizó la **mediana** para imputar Año_fabricación y Horas_uso, evitando el sesgo de valores extremos.
- **Variables Categóricas:** Se aplicó la **moda** para Provincia y Método_pago, manteniendo la integridad de la distribución original.
- **Lógica de Negocio:** Para las horas de uso, se validó que los equipos con estado "nuevo" tuvieran valor 0 antes de proceder con la imputación general.

### 2.2 Análisis de Outliers (Valores Atípicos)

Utilizando el método del **Rango Intercuartílico (IQR)**, se detectaron y trataron registros que distorsionaban la varianza:

- Se eliminaron registros con combinaciones ilógicas (ej. equipos nuevos con exceso de horas).
- El resultado fue un dataset con un **Skewness de 0.45**, lo que indica una distribución aproximadamente simétrica, ideal para la convergencia de modelos de regresión.




### 3. Análisis Estadístico y Visual (EDA)

#### 3.1 Correlaciones y Significancia (p-valores)

Mediante pruebas de hipótesis y coeficientes de **Pearson**, se determinaron los "drivers" del precio:

- **Capacidad de Carga ($p < 0.05$):** Es la variable con mayor correlación positiva. El mercado paga una prima lineal por el tonelaje.
- **Depreciación ($p < 0.05$):** Las Horas_uso muestran una correlación negativa moderada. No obstante, en marcas premium, la depreciación es menos agresiva, un fenómeno detectado mediante el análisis de **Boxplots**.

#### 3.2 Análisis de Distribución (QQ-Plot)

El gráfico **QQ-Plot** reveló que, si bien el precio tiene una ligera cola hacia la derecha, la distribución es lo suficientemente robusta para ser modelada sin transformaciones agresivas en esta etapa inicial.

### 4. Ingeniería de Características (Feature Engineering)

Para mejorar la capacidad de generalización del modelo, se realizaron las siguientes transformaciones:

- **Agrupación Regional:** Se redujo la cardinalidad de la variable Provincia creando la variable Zona (Norte, Centro, Sur), eliminando el ruido geográfico menor.
- **Codificación Jerárquica:** Se aplicó **Ordinal Encoding** a las variables Estado y Método_pago. Esto permite que el modelo entienda que un equipo "Nuevo" (valor 3) es intrínsecamente superior a uno "Usado" (valor 1).

- **Codificación Nominal:** Se utilizó **One-Hot Encoding** para las variables Marca y Tipo, evitando la trampa de la multicolinealidad mediante el parámetro drop_first=True.

**5. Conclusiones y Diagnóstico**

1. **Validación de Hipótesis:** El precio no es aleatorio; responde a una estructura lógica dominada por la capacidad técnica y el estado del equipo.
2. **Estado del Dataset:** Los datos han sido normalizados y están libres de inconsistencias graves.
3. **Recomendación Técnica:** Dada la baja asimetría (0.45) y la alta correlación de variables numéricas, el próximo paso lógico es la implementación de un modelo de **Regresión Lineal Múltiple** como base, comparándolo con un **Random Forest Regressor** para capturar posibles relaciones no lineales entre la marca y el precio.