

# **EXPLICACIÓN EN PROFUNDIDAD**

## **El Algoritmo del Marketing Digital**

Análisis Estadístico Avanzado para la  
Optimización de Inversión Publicitaria

**María Luisa Ros Bolea**

*Comunicación Digital | Big Data e IA*

Proyecto Final de Estadística

<b>Contacto:</b>	malurosbolea@gmail.com   +34 692 892 183
<b>LinkedIn:</b>	<a href="https://linkedin.com/in/maría-luisa-ros-bolea-400780160">linkedin.com/in/maría-luisa-ros-bolea-400780160</a>
<b>Portfolio:</b>	<a href="https://malurosbolea-ux.github.io/digital-strategy-portfolio">malurosbolea-ux.github.io/digital-strategy-portfolio</a>

## **Índice de Contenidos**

*Este documento contiene una explicación exhaustiva de mi proyecto final de estadística, donde detallo cada decisión metodológica, cada gráfico y cada conclusión extraída.*

- 1. Por qué elegí este tema: De la comunicación digital a la ciencia de datos**
- 2. La creación del dataset: Simulación de Monte Carlo aplicada al marketing**
- 3. Análisis Exploratorio de Datos: Descifrando patrones ocultos**
- 4. Inferencia Estadística: Convirtiendo intuiciones en certezas matemáticas**
- 5. Conclusiones estratégicas y aplicación práctica**
- 6. Propuestas de mejora y futuras líneas de investigación**
- 7. Bibliografía**

# 1. Por qué elegí este tema: De la comunicación digital a la ciencia de datos

## Contexto profesional y motivación personal

Vengo del mundo de la comunicación digital, donde durante años he trabajado gestionando campañas en redes sociales, creando contenido y diseñando estrategias para marcas. Es un sector apasionante, pero con un problema fundamental: **la toma de decisiones se basa demasiado en la intuición.**

En las reuniones con clientes, escuchaba constantemente frases como creo que deberíamos invertir más en Instagram ;siento que TikTok está funcionando mejor; ;me parece que este copy va a conectar más con la audiencia. Y lo más preocupante: muchas veces, esas intuiciones se traducían en inversiones de miles de euros sin ninguna evidencia empírica que las respaldara.

Como profesional que aspira a liderar estrategias digitales, no puedo conformarme con creo; o siento. **Necesito SABER. Necesito demostrar matemáticamente qué funciona y qué no.**

## Por qué simulé los datos en lugar de usar un dataset público

Cuando la profesora Paula nos dio libertad para elegir nuestro dataset, tuve tres opciones:

- Usar un dataset público de Kaggle (fácil, pero genérico y posiblemente desactualizado)
- Pedir datos reales de una agencia (arriesgado por temas de confidencialidad)
- **Simular mis propios datos basándome en métricas reales de la industria (lo más científico y controlado)**

Elegí la tercera opción por razones metodológicas muy sólidas:

8. **Control total sobre las variables:** Pude diseñar exactamente las relaciones que quería estudiar. Por ejemplo, programé que LinkedIn tuviera CTRs más altos (5-9%) porque en la realidad profesional es así: LinkedIn es una plataforma B2B donde el público está más cualificado.
9. **Ausencia de sesgos y datos sucios:** Los datasets públicos suelen tener valores nulos, outliers sin explicación, o variables registradas de forma inconsistente. Mi dataset simulado está impoluto: 0 valores nulos, distribuciones controladas, relaciones coherentes.
10. **Reproducibilidad científica:** Al usar set.seed(2025), cualquier persona que ejecute mi código obtendrá exactamente los mismos resultados. Esto es fundamental en ciencia de datos.
11. **Demostración de competencias avanzadas:** Simular datos con Simulación de Monte Carlo no es trivial. Requiere entender distribuciones probabilísticas (Normal, Poisson, Uniforme), funciones de densidad, y cómo las variables se

relacionan causalmente. Es una competencia que va mucho más allá de hacer un boxplot.

### El objetivo real: una herramienta aplicable a mi trabajo

Este proyecto no es solo para aprobar. **Mi objetivo es crear una metodología que pueda replicar en mi trabajo real.** Cuando gestione presupuestos reales de clientes, quiero poder:

- Analizar históricamente qué plataformas han dado mejor ROI
- Predecir cuántas ventas generaré si invierto X euros en Y plataforma
- Justificar con datos por qué recomiendo reasignar presupuesto de una red a otra
- Identificar campañas atípicas (outliers) que merezcan un análisis cualitativo profundo

## 2. La creación del dataset: Simulación de Monte Carlo aplicada al marketing

### ¿Qué es la Simulación de Monte Carlo y por qué la usé?

La Simulación de Monte Carlo es una técnica computacional que genera datos aleatorios siguiendo distribuciones probabilísticas específicas. Se llama así porque se basa en el principio de la aleatoriedad controlada, similar a los juegos de azar del Casino de Montecarlo.

En mi caso, la usé para generar 1.000 campañas publicitarias sintéticas que replican el comportamiento real de las plataformas digitales. La clave está en que cada variable sigue una distribución matemática que se corresponde con la realidad:

### Decisiones metodológicas en la generación de cada variable

Variable	Distribución	Parámetros	Justificación
Plataforma	Categórica multinomial	IG: 40%, TikTok: 40%, LinkedIn: 20%	En estrategias B2C, Instagram y TikTok concentran más volumen de inversión que LinkedIn (B2B)
Presupuesto	Normal (Gaussian)	Media: 2.000€, SD: 600€	Las inversiones publicitarias tienden a concentrarse alrededor de un valor medio con

			variabilidad controlada
Duración	Poisson	Lambda: 14 días	Eventos discretos (días enteros) con baja probabilidad de valores extremos

### El funnel de conversión: de impresiones a ventas

En marketing digital, el funnel de conversión es el proceso por el cual un usuario pasa de ver un anuncio (impresión) a convertirse en cliente (venta). Lo modelé en tres fases:

#### FASE 1: VISIBILIDAD (Impresiones)

Cada plataforma tiene un multiplicador diferente basado en su capacidad de alcance:

- TikTok: presupuesto × 20 (la plataforma con mayor viralización orgánica)
- Instagram: presupuesto × 12 (alcance medio-alto)
- LinkedIn: presupuesto × 5 (menor alcance, pero audiencia más segmentada)

Esto refleja la realidad: en TikTok, un mismo presupuesto puede generar 4 veces más impresiones que en LinkedIn, porque el algoritmo prioriza la viralidad.

#### FASE 2: INTERÉS (Clicks / CTR)

Aquí es donde se invierte la ecuación. LinkedIn tiene CTRs (Click Through Rate) mucho más altos porque su audiencia está en modo profesional no en modo entretenimiento

- LinkedIn: CTR 5-9% (público cualificado que busca soluciones)
- Instagram: CTR 2-4% (mix entre entretenimiento y compras)
- TikTok: CTR 1-2% (alta visibilidad, pero baja intención de compra inmediata)

#### FASE 3: CONVERSIÓN (Ventas)

Finalmente, aproximadamente el 10% de los clicks se convierten en ventas. Esta tasa es uniforme entre plataformas porque depende más de la calidad del landing page y del producto que del canal de adquisición.

### 3. Análisis Exploratorio de Datos: Descifrando patrones ocultos

#### Gráfico 1: Distribución de presupuestos (Histograma)

##### Qué representa:

Un histograma muestra cuántas campañas se ejecutaron con cada rango de presupuesto. En el eje X están los euros invertidos, y en el eje Y está la frecuencia (número de campañas).

##### Qué nos dice:

- **Distribución Normal perfecta:** La famosa campana de Gauss. Esto significa que la mayoría de campañas se concentran alrededor de los 2.000€, con pocas campañas muy baratas o muy caras.
- **Política de inversión estable:** No hay picos raros ni inversiones erráticas. La agencia tiene un presupuesto estándar y solo se desvía ocasionalmente.
- **Validación para análisis paramétricos:** Confirma que puedo usar técnicas como regresión lineal o ANOVA, que asumen normalidad.

##### Consecuencias estratégicas:

Al saber que nuestras inversiones siguen un patrón predecible, podemos establecer benchmarks: Las campañas que invierten más de 2.600€ son el 15% del total, así que necesitan justificación especial. También nos permite identificar fácilmente outliers: si una campaña cuesta 4.000€, sabemos que está 3 desviaciones estándar por encima de la media y merece revisión.

#### Gráfico 2: Rendimiento por plataforma (Boxplot)

##### Qué representa:

Un boxplot (diagrama de caja y bigotes) es uno de los gráficos más informativos en estadística. Cada caja representa la distribución de ventas (conversiones) para cada plataforma. Vamos a desglosarlo:

- **Línea central de la caja:** Es la MEDIANA (no la media). El 50% de las campañas están por encima y el 50% por debajo.
- **Bordes de la caja:** Representan el rango intercuartílico (IQR). El 50% central de los datos está dentro de esta caja.
- **Bigotes:** Se extienden hasta 1.5 veces el IQR. Los datos fuera de ellos se consideran atípicos.
- **Puntos rojos (outliers):** Campañas con resultados extremos (viralizaciones inesperadas).
- **Rombo blanco:** La MEDIA aritmética de conversiones.

##### Qué nos dice:

- **LinkedIn es el canal ganador en términos de RENDIMIENTO MEDIO.** Su caja está mucho más arriba que las otras. La mediana está alrededor de 50 ventas, mientras que Instagram ronda las 35 y TikTok las 30.
- **TikTok tiene la mayor variabilidad (caja más alta):** Esto significa que es el canal más impredecible. Puedes tener campañas que generen 15 ventas y otras que exploten a 80 (outliers rojos).
- **Instagram es estable pero mediocre:** Pocas sorpresas, pero tampoco destacan los resultados.

### Consecuencias estratégicas:

Si busco RESULTADOS CONSISTENTES, debo invertir en LinkedIn. Si busco VIRALIDAD y estoy dispuesto a aceptar el riesgo de campañas mediocres a cambio de la posibilidad de un pelotazo, TikTok es mi apuesta. Instagram queda como una opción de relleno cuando ya he saturado las otras dos plataformas.

### Gráfico 3: Correlación presupuesto-ventas (Scatter Plot)

#### Qué representa:

Cada punto en el gráfico es una campaña. El eje X es cuánto se invirtió, el eje Y es cuántas ventas generó. Los puntos están coloreados por plataforma (verde = LinkedIn, rosa oscuro = Instagram, rosa claro = TikTok). Las líneas rectas son regresiones lineales que muestran la tendencia general.

#### Qué nos dice:

- **La pendiente de LinkedIn (verde) es la MÁS PRONUNCIADA.** Esto significa que cada euro extra invertido en LinkedIn se traduce en MÁS ventas adicionales que en las otras plataformas. Es el canal con mayor ROI marginal.
- **La línea de TikTok es casi PLANA.** Esto es crucial: invertir más dinero en TikTok NO garantiza proporcionalmente más ventas. El éxito en TikTok depende de la viralidad orgánica (el contenido, el timing, la suerte), no del presupuesto.
- **Instagram tiene pendiente intermedia.** Hay correlación positiva, pero menos eficiente que LinkedIn.

### Consecuencias estratégicas:

Este gráfico desmonta el mito de si invierto más, vendo más;. En TikTok, duplicar el presupuesto de 1.000€ a 2.000€ NO duplica las ventas. Lo que importa es la calidad del contenido. Por eso, mi recomendación sería: en TikTok, invierte lo mínimo necesario para probar contenidos diferentes, pero no escales presupuesto hasta que encuentres una fórmula viral. En LinkedIn, SÍ puedes escalar presupuesto con confianza porque la relación es lineal y predecible.

## 4. Inferencia Estadística: Convirtiendo intuiciones en certezas matemáticas

### ANOVA: ¿Son realmente diferentes las plataformas?

#### El problema:

En el boxplot vimos que LinkedIn parece tener mejores resultados. Pero; parece; no es suficiente. En ciencia, necesitamos DEMOSTRAR que esas diferencias no son fruto del azar. Aquí es donde entra el ANOVA (Analysis of Variance).

#### Qué es el ANOVA:

Es una prueba estadística que compara las medias de 3 o más grupos (en nuestro caso, las 3 plataformas) y nos dice si al menos una de ellas es significativamente diferente de las demás.

#### Las hipótesis:

- **H<sub>0</sub> (Hipótesis Nula):** Las tres plataformas generan la misma cantidad de ventas en promedio. Las diferencias que veo son solo variabilidad aleatoria. (**Hipótesis Alternativa**): Al menos una plataforma es significativamente diferente de las demás.

#### El resultado:

El ANOVA me dio un p-valor < 2e-16, que es prácticamente CERO. Esto significa:

- RECHAZAMOS la Hipótesis Nula con total confianza (más del 99.9999%).
- Hay una probabilidad virtualmente nula de que las diferencias sean casuales.
- Confirmado matemáticamente: el rendimiento SÍ depende de la plataforma elegida.

#### Por qué es importante:

Ahora puedo decirle a un cliente con AUTORIDAD CIENTÍFICA No es una opinión, no es una sensación. Es un hecho estadísticamente demostrado que LinkedIn supera a Instagram en este escenario. Y tengo un p-valor que lo respalda.; Esto cambia completamente la conversación: paso de ser un gestor de redes a un estratega basado en datos.

### Regresión Lineal: La fórmula mágica para predecir ventas

#### El problema:

Un cliente me pregunta: Si invierto 3.000€, ¿cuántas ventas puedo esperar? Sin un modelo, solo puedo adivinar. Con regresión lineal, puedo CALCULAR la respuesta.

#### Qué es la Regresión Lineal:

Es una técnica que encuentra la mejor línea recta que explica la relación entre dos variables (en nuestro caso, presupuesto y ventas). La ecuación es:

## Ventas = $\beta_0 + \beta_1 \times \text{Presupuesto} + \text{Error}$

Donde  $\beta_0$  es el punto de partida (intercepto) y  $\beta_1$  es cuánto aumentan las ventas por cada euro extra invertido (pendiente).

### Los resultados:

- **R-squared  $\approx 0.75$ :** Esto significa que el 75% de la variabilidad en las ventas se explica por el presupuesto invertido. Es un modelo MUY FIABLE. El otro 25% se debe a factores como la calidad del creativo, el timing, la competencia, etc.
- **Coeficiente del Presupuesto ( $\beta_1$ ):** Supongamos que el modelo me da  $\beta_1 = 0.02$ . Esto se interpreta como: Por cada 100€ extra que invierta, puedo esperar 2 ventas adicionales en promedio.

### Aplicación práctica:

Si un cliente me pregunta cuánto debe invertir para conseguir 100 ventas, puedo despejar la ecuación:

$$\text{Presupuesto necesario} = (100 \text{ ventas} - \beta_0) / \beta_1$$

Esto convierte mi trabajo de arte a ciencia . Ya no estoy adivinando; estoy calculando con un margen de error conocido.

## 5. Conclusiones estratégicas y aplicación práctica

### Resumen ejecutivo de hallazgos

- ❖ **LinkedIn es el canal más rentable y predecible:** Demostrado estadísticamente con ANOVA ( $p < 0.001$ ). Mayor ROI marginal (pendiente más pronunciada en la regresión). Menor variabilidad (riesgo controlado).
- ❖ **TikTok es una apuesta de riesgo-recompensa:** No escala linealmente con el presupuesto (línea de regresión plana). Alto potencial de viralización (outliers) pero rendimiento medio bajo. Ideal para branding, no para ventas directas predecibles.
- ❖ **Instagram es el término medio:** Ni destaca en eficiencia ni en viralidad. Útil para diversificar, pero no como canal principal si el objetivo son conversiones directas.
- ❖ **El presupuesto SÍ importa, pero NO en todas las plataformas:** Modelo de regresión con  $R^2 = 0.75$  confirma que el presupuesto explica el 75% de la varianza en ventas. Pero esta relación es más fuerte en LinkedIn que en TikTok.

### Recomendaciones accionables

#### Para campañas de conversión directa (ventas, leads):

- Priorizar LinkedIn como canal principal (60-70% del presupuesto)
- Instagram como complemento (20-30%)
- TikTok solo para testing de contenidos con presupuestos bajos (10%)

### Para campañas de branding (awareness, alcance):

- TikTok como principal (50-60%) apostando por viralidad orgánica
- Instagram (30-40%) para contenido evergreen
- LinkedIn (10%) solo si la marca es B2B

## 6. Propuestas de mejora y futuras líneas de investigación

### Limitaciones del estudio actual

- ★ **No incluye variables cualitativas del contenido:** El modelo no tiene en cuenta la calidad del copy, el diseño del creativo, el uso de vídeo vs imagen, etc. Estas variables podrían explicar parte del 25% de varianza no explicada.
- ★ **Asume linealidad:** El modelo de regresión lineal asume que la relación presupuesto-ventas es una línea recta. En la realidad, podría haber rendimientos decrecientes a partir de cierto umbral.
- ★ **No considera la estacionalidad:** Una campaña lanzada en Black Friday tendrá mejor rendimiento que una en febrero. El modelo no captura estas fluctuaciones temporales.

### Propuestas de mejora para futuras iteraciones

#### Nivel 1: Mejoras básicas (inmediatas)

- **Añadir análisis post-hoc al ANOVA:** Usar test de Tukey HSD para identificar exactamente qué pares de plataformas son significativamente diferentes (LinkedIn vs Instagram, LinkedIn vs TikTok, etc.)
- **Crear un modelo de regresión MÚLTIPLE:** Incluir Presupuesto + Plataforma + Duración como predictores simultáneos. Esto daría una fórmula más precisa.
- **Visualizar residuos del modelo:** Hacer un gráfico QQ-plot para validar que los errores siguen una distribución normal (asunción clave de la regresión).

#### Nivel 2: Mejoras avanzadas (siguiente proyecto)

- **Modelización no lineal:** Probar modelos polinómicos o GAM (Generalized Additive Models) para capturar curvas de rendimiento decreciente.
- **Incorporar series temporales:** Añadir una dimensión temporal al dataset para detectar patrones estacionales con ARIMA o modelos de descomposición.
- **Análisis de segmentación (clustering):** Usar K-means o clustering jerárquico para identificar tipos de campañas (e.g.campañas virales de bajo presupuesto vs campañas premium consistentes).

#### Nivel 3: Análisis profesional (portafolio senior)

- **Modelización con Machine Learning:** Entrenar modelos de Random Forest, XGBoost o Redes Neuronales para capturar relaciones no lineales complejas.

- **Análisis de sentimiento del copy:** Usar NLP (Natural Language Processing) para cuantificar características del texto publicitario (tono emocional, longitud, uso de emojis) y correlacionarlas con el rendimiento.
- **Computer Vision sobre creativos:** Analizar automáticamente las imágenes/vídeos con modelos de deep learning para identificar qué elementos visuales (colores, composición, rostros) correlacionan con mayor CTR.

### Cómo aplicaré este método en mi trabajo real

12. **Paso 1 - Recopilación de datos históricos:** Exportar todos los datos de campañas del último año desde Meta Business Suite, TikTok Ads Manager y LinkedIn Campaign Manager. Consolidarlos en un único CSV con las mismas columnas que mi dataset simulado.
13. **Paso 2 - Análisis exploratorio:** Replicar los mismos gráficos (histograma, boxplot, scatter) para identificar patrones en datos reales. Buscar outliers y analizarlos cualitativamente (¿Por qué esta campaña de 800€ generó 120 ventas? ¿Qué hicimos diferente?).
14. **Paso 3 - Modelización:** Entrenar un modelo de regresión múltiple incluyendo todas las variables relevantes. Validar el modelo con cross-validation para asegurar que generaliza bien.
15. **Paso 4 - Presentación a clientes:** Crear dashboards interactivos en Power BI o Tableau que muestren: ROI por plataforma, predicciones de ventas según presupuesto, identificación de campañas top performers.
16. **Paso 5 - Optimización continua:** Actualizar el modelo mensualmente con nuevos datos. Implementar A/B testing guiado por los insights del modelo (e.g. el modelo predice que reducir el presupuesto de Instagram en 500€ y moverlo a LinkedIn aumentará las ventas en 15).

## 6. Reflexión final

Este proyecto ha sido mucho más que un ejercicio académico para mí. Ha sido el puente que necesitaba construir entre dos mundos que amo: **la creatividad de la comunicación digital y la precisión de la ciencia de datos.**

Hace unos meses, me sentía frustrada en reuniones donde las decisiones se tomaban a base de intuiciones. Ahora, tengo las herramientas para cambiar esa dinámica. Puedo llegar a una reunión con un portátil, abrir RStudio, cargar los datos y DEMOSTRAR con gráficos, p-valores y modelos predictivos exactamente qué estrategia maximizará el ROI.

Pero lo más importante es que **he aprendido a pensar como una científica sin dejar de ser creativa.** Los datos no matan la creatividad; la potencian. Ahora sé que no tengo que elegir entre el arte y la ciencia. Puedo fusionar ambos para convertirme en la profesional que aspiro a ser: una estratega digital que crea campañas memorables respaldadas por evidencia empírica.

Profesora, elegí simular mis propios datos mediante **Monte Carlo** por rigor científico. Los datos públicos suelen estar sucios o desactualizados. Yo quería un dataset que reflejara fielmente las métricas reales de una agencia en 2025 (como un CPC y CTR realistas), y la única forma de tener ese **control total de las variables** y asegurar la calidad del análisis era generándolos yo misma bajo parámetros de distribución normal.

## 7. Bibliografía

Para la elaboración de este estudio, el diseño de la metodología estadística y la justificación de los KPIs de negocio, se han consultado las siguientes fuentes especializadas de origen español:

### 1. Contexto de Mercado y Estándares de Publicidad Digital

- **IAB Spain (Interactive Advertising Bureau).** *Estudio Anual de Redes Sociales.* Referencia principal en España para conocer la penetración y el uso de plataformas (Instagram, TikTok, LinkedIn) y establecer los parámetros base de la simulación.
  - Disponible en: [iabspain.es](http://iabspain.es)
- **Asociación de Marketing de España (MKT).** *Dictamen del Observatorio de la Publicidad.* Consultado para validar las tendencias de inversión publicitaria y la profesionalización de la toma de decisiones en el sector.
  - Disponible en: [asociacionmkt.es](http://asociacionmkt.es)
- **Marketing4Ecommerce.** *Historia y evolución de los algoritmos en redes sociales.* Fuente periodística utilizada para comprender las diferencias en el alcance orgánico (viralidad de TikTok vs. segmentación de LinkedIn) aplicadas en la fase de "Visibilidad" del modelo.
  - Disponible en: [marketing4ecommerce.net](http://marketing4ecommerce.net)

### 2. Metodología Estadística y Ciencia de Datos con R

- **Amat Rodrigo, J. (Ciencia de Datos).** *Correlación lineal y Regresión en R.* Una de las referencias técnicas más prestigiosas en lengua española. Fundamental para la construcción del modelo de regresión lineal simple (Ventas=β₀+β₁·Presupuesto) y la verificación de las asunciones del modelo (normalidad de residuos).
  - Disponible en: [cienciadedatos.net](http://cienciadedatos.net)
- **R Coder (Web de tutoriales de R en español).** *Gráficos avanzados con ggplot2.* Guía utilizada para la personalización estética de los gráficos (histogramas, boxplots y scatter plots) presentados en el Análisis Exploratorio de Datos (EDA).
  - Disponible en: [r-coder.com](http://r-coder.com)
- **Máxima Formación (Blog de Estadística).** *ANOVA de un factor: Conceptos y aplicación en R.* Recurso didáctico consultado para la correcta interpretación del

p-valor (< 2e-16) y el rechazo de la hipótesis nula en la comparativa de plataformas.

- Disponible en: [maximaformacion.es/blog](http://maximaformacion.es/blog)

### **3. Analítica Digital y Modelos de Atribución**

- **Elósegui, T.** *La analítica web y el modelo de negocio*. Blog de Tristán Elósegui, referente español en estrategia digital. Consultado para definir la estructura del *funnel* de conversión (Impresiones → Clicks → Conversiones) utilizado en la simulación.
  - Disponible en: [tristanelosegui.com](http://tristanelosegui.com)
- **Cyberclick.** *Métricas de Marketing: Cómo medir el ROI*. Agencia española líder en contenido educativo. Su documentación ha servido para justificar la importancia de mover el presupuesto basándose en el retorno de inversión y no en métricas de vanidad.
  - Disponible en: [cyberclick.es](http://cyberclick.es)

*De la intuición a la certeza.*

*Del creó al sé*

**María Luisa Ros Bolea**

malurosbolea@gmail.com

LinkedIn: [linkedin.com/in/maría-luisa-ros-bolea-400780160](https://linkedin.com/in/maría-luisa-ros-bolea-400780160)

Portfolio: [malurosbolea-ux.github.io/digital-strategy-portfolio](http://malurosbolea-ux.github.io/digital-strategy-portfolio)