



Taller 1. Sistemas Avanzados de Producción

Julieth Dayanna Cifuentes Melo

Michelle Renata Cuadrado Suárez

Sistemas Avanzados De Producción

Universidad Ecci

2026-01

Febrero, 28 De 2026

Índice

Fase 1: Estadística Descriptiva y Análisis Exploratorio	3
Caracterización Numérica	3
Análisis de Distribución y Atípicos	5
Análisis según los valores	5
Simetría	5
Exploración de la Nube de Puntos	8
Evaluación de Asociación	11
Interpretación de los coeficientes de Pearson y Spearman	11
Matriz de Correlación de Pearson	11
Matriz de Correlación de Spearman	12
Fase 2: Regresión Lineal y Diagnóstico	13
Modelamiento Múltiple	13
Interpretación de Parámetros	13
Intercepto ($\beta_0 = 2.9389$)	13
Pendiente TV ($\beta_1 = 0.0458$)	13
Pendiente Radio ($\beta_2 = 0.1885$)	13
Pendiente Newspaper ($\beta_3 = -0.0010$)	14
Estimación Matricial	14
Matrices iniciales	14
Transpuesta de X	15
Producto $X'X$	15
Inversa de $X'X$	15
Producto $X'Y$	15
Estimadores β^{\wedge}	16
Resultado Final	16
Fase 3: Árboles de Decisión y Comparación de Modelos	16

Fase 1: Estadística Descriptiva y Análisis Exploratorio

Se cargó el archivo Advertising.csv directamente desde la URL oficial para garantizar la integridad y trazabilidad de los datos. Posteriormente, se verificó la estructura del conjunto de datos, confirmando que contiene las variables TV, Radio, Newspaper (inversión publicitaria en distintos medios) y Sales (ventas), todas de tipo numérico. Finalmente, se realizó una revisión de valores faltantes, comprobando que no existen datos ausentes que puedan sesgar el análisis estadístico o los modelos posteriores.

Tabla 1. Carga y Preparación de Datos

#	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

Caracterización Numérica

A continuación, se presentan las medidas de tendencia central y dispersión para las variables analizadas (TV, radio, newspaper y sales).

Tabla 2. Medidas de Tendencia Central y Dispersión

Variable	Media	Mediana	Desv. Estándar	Mínimo	Máximo	Asimetría	Curtosis
TV	147.042	149.75	85.854236	0.7	296.4	-0.06985	-1.22649

Radio	23.2640	22.90	14.846809	0.0	49.6	0.094175	-1.26040 1
Newspaper	30.5540	25.75	21.778621	0.3	114.0	0.894720	0.649502
Sales	14.0225	12.90	5.217457	1.6	27.0	0.407571	-0.40886 9

En términos generales, la inversión en TV muestra la media más alta (147.04) y también la mayor variabilidad (DE = 85.85), lo que indica una amplia dispersión en los montos destinados a este medio. La mediana (149.75) es muy cercana a la media y la asimetría es prácticamente nula (-0.07), lo que sugiere una distribución aproximadamente simétrica. Además, la curtosis negativa (-1.23) indica una distribución ligeramente platicúrtica, es decir, con menor concentración de valores en torno a la media en comparación con una distribución normal.

En el caso de radio, se observa una media de 23.26 y una desviación estándar de 14.85, lo que refleja una dispersión moderada. La asimetría también es cercana a cero (0.09), lo que indica una distribución relativamente simétrica. La curtosis negativa (-1.26) sugiere una forma más aplanada que la normal.

Por su parte, newspaper presenta una media de 30.55 y una desviación estándar de 21.78, evidenciando una variabilidad considerable. La asimetría positiva (0.89) indica que la distribución está sesgada hacia la derecha, es decir, existen algunos valores altos que elevan la media. La curtosis positiva (0.65) sugiere una mayor concentración de datos alrededor de la media en comparación con una distribución normal.

Finalmente, la variable sales tiene una media de 14.02 y una desviación estándar de 5.22, mostrando una dispersión menor en comparación con las variables de inversión publicitaria. La asimetría positiva moderada (0.41) indica una ligera inclinación hacia valores altos, mientras que la curtosis negativa (-0.41) señala una distribución ligeramente más aplanada que la normal.

En conjunto, los resultados permiten observar diferencias importantes en los niveles de dispersión y forma de las distribuciones entre los distintos medios publicitarios y las ventas, lo cual resulta relevante para el análisis posterior de relaciones y modelos explicativos.

Análisis de Distribución y Atípicos

Para determinar la simetría se analiza el coeficiente de asimetría (skewness):

- Si es ≈ 0 → distribución simétrica
- Si es > 0 → asimetría positiva (cola hacia la derecha)
- Si es < 0 → asimetría negativa (cola hacia la izquierda)
- Valores entre -0.5 y 0.5 suelen considerarse aproximadamente simétricos.

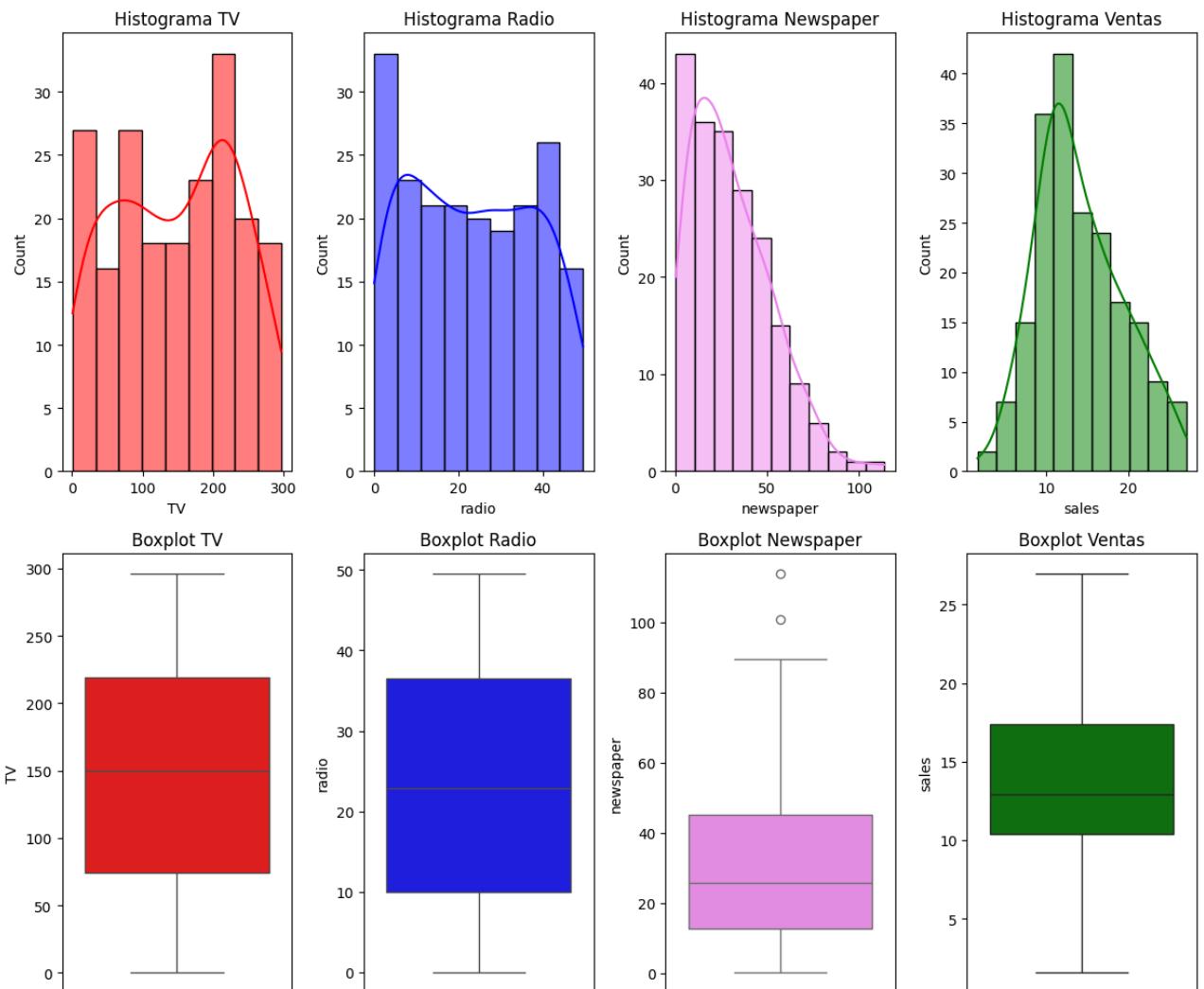
Análisis según los valores

Simetría

En el ámbito de la simetría, para la variable de TV (-0.069853), el valor es muy cercano a 0. Indica una distribución prácticamente simétrica, con una ligera inclinación hacia la izquierda, pero no significativa. Por otra parte, para radio (0.094175), también es muy cercano a 0. Se considera aproximadamente simétrica, con leve tendencia positiva, pero sin sesgo relevante. En el caso de Newspaper (0.894720), se presenta asimetría positiva moderada. La distribución está sesgada hacia la derecha, lo que implica que existen valores altos que alargan la cola derecha y elevan la media. Y por último, Sales (0.407571) muestra asimetría positiva leve, aunque tiene una ligera inclinación hacia la derecha, aún puede considerarse relativamente cercana a la simetría.

A continuación se presentan los gráficos, en la parte superior de la figura se observan los histogramas correspondientes a las variables TV, Radio, Newspaper y Ventas, cada uno acompañado por su curva de densidad suavizada y en la parte inferior se presentan los diagramas de caja de las mismas variables, lo que permite complementar el análisis visual de la distribución, la dispersión y la presencia de posibles valores atípicos.

Gráfico 1. Histogramas y Box Plot variables TV, Radio, Newspaper y Ventas.



En la variable TV, el histograma muestra una distribución amplia que abarca prácticamente todo el rango de valores, desde niveles muy bajos hasta cercanos a 300. La frecuencia de observaciones se distribuye de manera relativamente homogénea, sin una concentración marcada en un intervalo específico. La curva de densidad sugiere una forma cercana a la simetría, aunque con ligeras variaciones en algunos tramos. El diagrama de caja evidencia un rango intercuartílico amplio, lo que confirma una alta variabilidad. La mediana se encuentra aproximadamente centrada dentro de la caja y los extremos se extienden de forma equilibrada, sin observarse valores atípicos destacados.

En el caso de Radio, el histograma presenta una distribución extendida dentro de un rango menor en comparación con TV, aproximadamente entre 0 y 50. Las observaciones se concentran principalmente en valores intermedios, aunque se aprecia una distribución relativamente uniforme en el rango completo. La forma general no muestra sesgos pronunciados. El diagrama de caja indica una dispersión moderada, con un rango intercuartílico considerable y una mediana ubicada cerca del centro de la caja. Los extremos presentan una extensión similar y no se identifican valores atípicos relevantes.

Para Newspaper, el histograma evidencia una mayor concentración de observaciones en los valores bajos y medios, con una disminución progresiva de frecuencias a medida que aumentan los valores. La curva de densidad muestra claramente una asimetría positiva, con una cola más prolongada hacia la derecha. El diagrama de caja confirma esta característica, mostrando una mayor extensión en el extremo superior y la presencia de valores atípicos en la parte alta, lo que indica la existencia de inversiones considerablemente superiores al resto de los datos. El rango intercuartílico es amplio, reflejando una variabilidad importante.

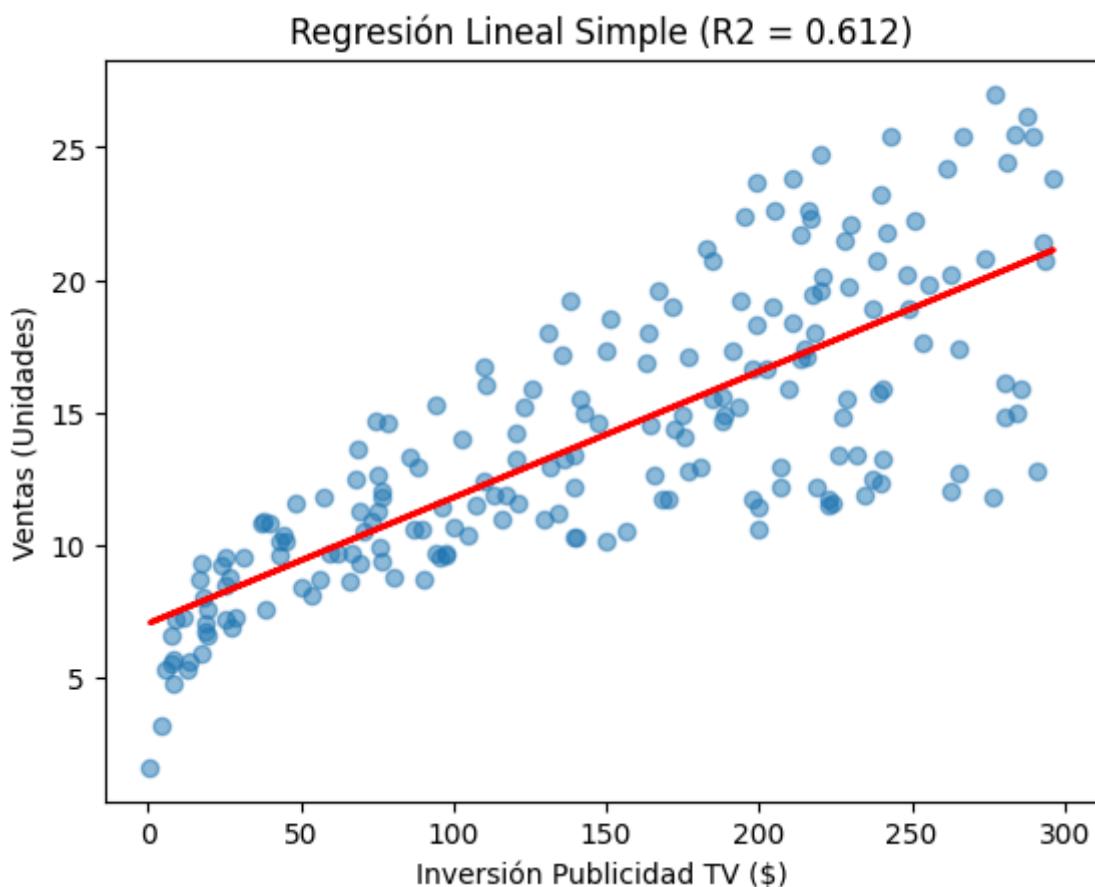
En cuanto a la variable Ventas, el histograma presenta una distribución unimodal con concentración de observaciones en torno a los valores centrales, especialmente entre aproximadamente 10 y 15 unidades. La forma general se aproxima a una distribución relativamente simétrica, aunque con ligera inclinación hacia la derecha. El diagrama de caja muestra un rango intercuartílico más reducido en comparación con las variables de inversión publicitaria, lo que indica menor dispersión. La mediana se encuentra centrada dentro de la caja y no se observan valores atípicos extremos claramente marcados.

En conjunto, la representación gráfica permite apreciar diferencias en la forma y variabilidad de las distribuciones. Las variables de inversión publicitaria presentan mayor amplitud y dispersión, especialmente TV y Newspaper, mientras que la variable Ventas muestra una distribución más concentrada y estable dentro de un rango más acotado.

Exploración de la Nube de Puntos

A partir de los dispersogramas presentados, se realiza el siguiente análisis visual:

Gráfica 2. Sales vs. Inversión en Publicidad en TV

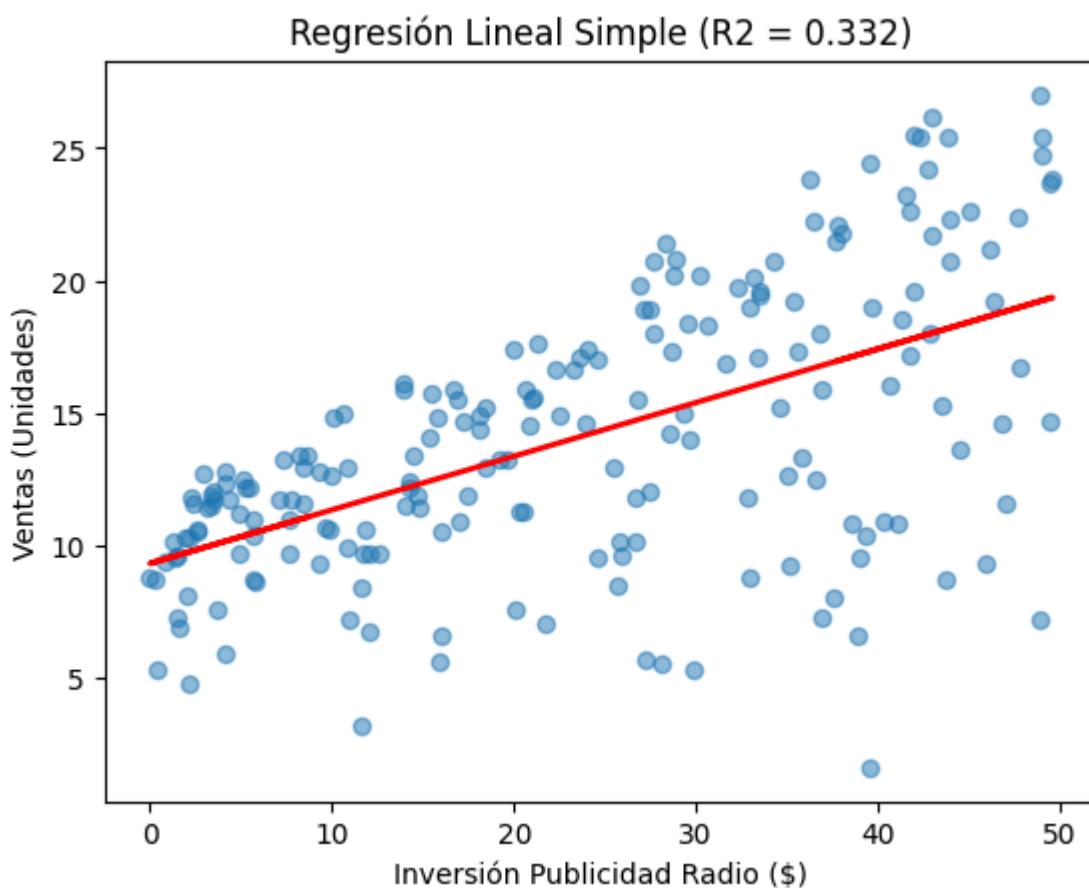


La nube de puntos muestra una tendencia creciente claramente definida. A medida que aumenta la inversión en publicidad en televisión, las ventas tienden a incrementarse. La distribución de los puntos alrededor de la recta de regresión es relativamente compacta, lo que indica una asociación lineal moderadamente fuerte.

No se observan patrones curvilíneos evidentes, como comportamientos parabólicos, ni cambios de pendiente que sugieran relaciones segmentadas. Tampoco se identifican valores atípicos extremos que aparenten ejercer una influencia desproporcionada sobre el ajuste. El coeficiente de determinación $R^2 = 0.612$ respalda visualmente que el modelo lineal simple explica una proporción considerable de la variabilidad en las ventas.

En consecuencia, el gráfico sugiere adecuadamente una relación lineal simple entre la inversión en TV y las ventas.

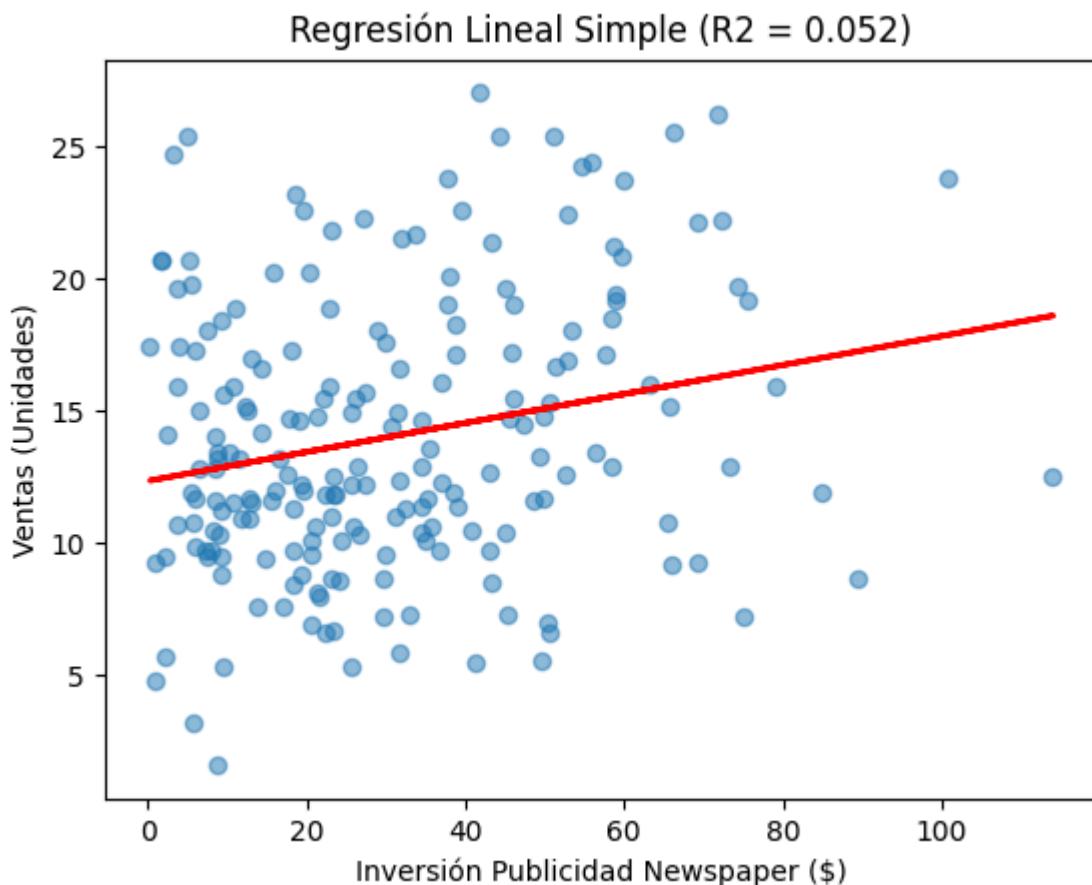
Gráfica 3. Sales vs. Inversión en Publicidad en Radio



En este caso también se observa una tendencia positiva: mayores niveles de inversión en radio tienden a asociarse con mayores ventas. Sin embargo, la dispersión de los puntos alrededor de la recta es mayor que en el caso de la televisión, lo que indica una relación lineal más débil.

No se aprecian patrones no lineales claros, como formas parabólicas o estructuras por tramos. Aunque existen algunos puntos relativamente alejados, no se evidencian valores extremos claramente influyentes a simple vista. El coeficiente de determinación $R^2 = 0.332$ confirma que la relación lineal es más débil y que la inversión en radio explica una proporción menor de la variabilidad en las ventas.

Gráfica 4. Sales vs. Inversión en Publicidad en Newspaper



El coeficiente de determinación $R^2 = 0.052$ indica que el modelo lineal simple explica únicamente el 5.2 % de la variabilidad de las ventas. Este valor es muy bajo, lo que sugiere una relación lineal prácticamente inexistente o muy débil entre la inversión en prensa escrita y las ventas.

Visualmente, un R^2 de esta magnitud suele corresponder a una nube de puntos altamente dispersa, sin una tendencia lineal clara y con gran variabilidad alrededor de la recta de regresión. Es probable que los puntos no sigan un patrón definido y que la pendiente estimada sea pequeña.

Por lo tanto, se puede concluir que:

- La inversión en televisión presenta una relación lineal positiva moderadamente fuerte con las ventas ($R^2 = 0.612$).
- La inversión en radio muestra una relación lineal positiva, pero más débil ($R^2 = 0.332$).

- La inversión en newspaper presenta una relación lineal muy débil o prácticamente nula con las ventas ($R^2 = 0.052$).

En términos comparativos, la publicidad en televisión es la covariable que mejor explica las ventas dentro de un modelo de regresión lineal simple, seguida por radio, mientras que newspaper aporta muy poca capacidad explicativa en forma individual.

Evaluación de Asociación

Interpretación de los coeficientes de Pearson y Spearman

Con el propósito de analizar la relación entre las variables TV, radio, newspaper y sales, se calcularon los coeficientes de correlación de Pearson y Spearman. Ambos permiten medir la fuerza y dirección de la asociación entre variables, aunque desde enfoques distintos: Pearson evalúa relaciones lineales, mientras que Spearman analiza relaciones monótonas basadas en el orden o rango de los datos.

Matriz de Correlación de Pearson

El análisis de Pearson muestra que:

- Existe una **correlación positiva fuerte** entre **TV y sales ($r = 0.78$)**. Esto indica que, a mayor inversión en publicidad en televisión, mayores son las ventas, siguiendo una relación lineal considerablemente marcada.
- La variable **radio presenta una correlación positiva moderada con sales ($r = 0.58$)**. Esto sugiere que la inversión en radio también influye en las ventas, aunque en menor medida que la televisión.
- En el caso de **newspaper y sales ($r = 0.23$)**, la correlación es positiva pero débil. Esto indica que la publicidad en prensa tiene una relación poco significativa con el incremento de ventas.
- Entre las variables independientes, se observa una correlación moderada entre **radio y newspaper ($r = 0.35$)**, lo que podría indicar cierta relación en la asignación del presupuesto publicitario entre estos dos medios. Sin embargo,

TV muestra correlaciones muy bajas con radio (0.05) y newspaper (0.06), lo que sugiere que la inversión en televisión se comporta de manera relativamente independiente respecto a los otros medios.

Matriz de Correlación de Spearman

El análisis mediante el coeficiente de Spearman arroja resultados similares:

- La relación entre **TV y sales ($\rho = 0.80$)** es fuerte y positiva, confirmando que, incluso al considerar el orden de los datos, la televisión mantiene una asociación significativa con el aumento de ventas.
- La variable **radio y sales ($\rho = 0.55$)** muestra una correlación positiva moderada, coherente con el resultado obtenido en Pearson.
- La relación entre **newspaper y sales ($\rho = 0.19$)** es débil, lo que reafirma que la publicidad en prensa tiene una influencia limitada sobre las ventas.
- Entre los medios publicitarios, nuevamente se observa una correlación moderada entre **radio y newspaper ($\rho = 0.32$)**, mientras que las demás asociaciones son bajas.

Al comparar ambas matrices, se observa que los coeficientes de Pearson y Spearman son muy similares en magnitud y dirección. Esto indica que las relaciones entre las variables no solo son lineales, sino también consistentemente monótonas. No se evidencian distorsiones importantes por valores atípicos ni comportamientos no lineales marcados.

En conclusión:

- La televisión es el medio con mayor impacto en las ventas.
- La radio tiene un efecto moderado.
- La prensa escrita presenta una influencia baja.

Estos resultados sugieren que, dentro de la estrategia publicitaria analizada, la asignación de recursos en televisión resulta más determinante para el comportamiento de las ventas.

Fase 2: Regresión Lineal y Diagnóstico

Modelamiento Múltiple

El modelo estimado puede expresarse de la siguiente forma:

$$[Sales = 2.9389 + 0.0458(TV) + 0.1885(Radio) - 0.0010(Newspaper)]$$

A continuación, se interpreta cada coeficiente en el contexto del modelo:

Interpretación de Parámetros

Intercepto ($\beta_0 = 2.9389$)

El intercepto representa el valor esperado de las ventas cuando la inversión en TV, Radio y Newspaper es igual a cero.

En este caso, si no se invierte en ningún medio publicitario, el modelo estima que las ventas serían aproximadamente 2.94 unidades.

Aunque en la práctica puede no ser realista tener inversión cero, este valor funciona como punto de partida del modelo.

Pendiente TV ($\beta_1 = 0.0458$)

Este coeficiente indica que, manteniendo constantes Radio y Newspaper, por cada unidad adicional invertida en televisión, las ventas aumentan en promedio 0.0458 unidades.

Es una relación positiva, lo que confirma que la publicidad en TV tiene un efecto favorable sobre las ventas. Sin embargo, su impacto individual es menor que el de la radio cuando se analizan conjuntamente en el modelo.

Pendiente Radio ($\beta_2 = 0.1885$)

El coeficiente de Radio indica que, manteniendo constantes TV y Newspaper, por cada unidad adicional invertida en radio, las ventas aumentan en promedio 0.1885 unidades.

Este es el coeficiente más alto del modelo, lo que sugiere que, dentro del análisis multivariable, la radio tiene el mayor efecto marginal sobre las ventas.

Pendiente Newspaper ($\beta_3 = -0.0010$)

Este coeficiente es negativo y muy cercano a cero. Indica que, manteniendo constantes las demás variables, un aumento de una unidad en la inversión en prensa se asocia con una disminución promedio de 0.0010 unidades en las ventas.

Sin embargo, debido a su magnitud tan pequeña, este efecto es prácticamente nulo. Esto sugiere que la publicidad en prensa no tiene un impacto relevante en las ventas dentro del modelo.

El modelo muestra que:

- Radio tiene el mayor impacto marginal sobre las ventas.
- La televisión también influye positivamente, aunque en menor medida.
- Newspaper no presenta un efecto significativo en términos prácticos.
- El intercepto establece el nivel base estimado de ventas sin inversión publicitaria.

En conjunto, la regresión confirma que los medios electrónicos (TV y Radio) son los principales impulsores de las ventas en comparación con la prensa escrita.

Estimación Matricial

Matrices iniciales

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{pmatrix}, \quad Y = \begin{pmatrix} 2 \\ 4 \\ 6 \\ 7 \\ 9 \end{pmatrix}$$

Transpuesta de X

$$X' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

Producto X'X

$$X'X = X' \cdot X = \begin{pmatrix} 5 & 15 \\ 15 & 55 \end{pmatrix}$$

$$\text{Sumatorias: } \begin{cases} n = 5 \\ \sum x = 15 \\ \sum x^2 = 55 \end{cases}$$

Inversa de X'X

$$(X'X)^{-1} = \frac{1}{\text{Det}} \begin{pmatrix} 55 & -15 \\ -15 & 5 \end{pmatrix} = \begin{pmatrix} 1.1 & -0.3 \\ -0.3 & 0.1 \end{pmatrix}$$

$$\text{Determinante: } \text{Det} = 5 \cdot 55 - 15 \cdot 15 = 50$$

Producto X'Y

$$X'Y = \begin{pmatrix} 28 \\ 100 \end{pmatrix}$$

$$\text{Sumatorias: } \begin{cases} \sum y = 28 \\ \sum xy = 100 \end{cases}$$

Estimadores β^{\wedge}

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} 1.1 & -0.3 \\ -0.3 & 0.1 \end{pmatrix} \begin{pmatrix} 28 \\ 100 \end{pmatrix} = \begin{pmatrix} 0.80 \\ 1.60 \end{pmatrix}$$

$$\begin{cases} \hat{\beta}_0 = 0.80 \\ \hat{\beta}_1 = 1.60 \end{cases}$$

Resultado Final

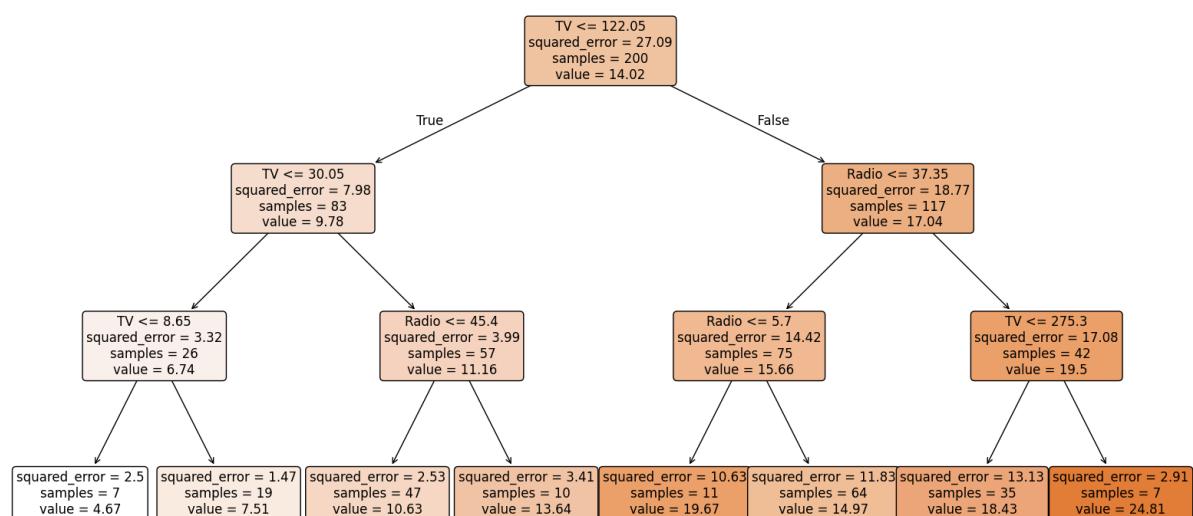
$$\hat{y} = 0.80 + 1.60x$$

Interpretación: $\begin{cases} \text{Intercepto } \beta_0 = 0.80 \text{ (valor de } y \text{ cuando } x = 0) \\ \text{Pendiente } \beta_1 = 1.60 \text{ (incremento promedio de } y \text{ por unidad de } x) \end{cases}$

Fase 3: Árboles de Decisión y Comparación de Modelos

Gráfico 5. Árbol de Decisión

Estructura del Árbol de Regresión: Ventas vs (TV y Radio)



Las particiones en los nodos se realizan bajo los criterios de reducción de varianza en donde el algoritmo analiza cada paso de la variable para que los nodos resultantes sean lo más homogéneos posibles .

Mecánica de selección: Bajo este criterio se toma la división que produce mayor disminución en la suma total de los errores a comparación de con los nodos que los anteceden .

Tabla 3. Cálculo Manual de Incertidumbre (Clasificación)

Mercado	Inversión TV	Ventas (Clase)
1	Alta	Altas
2	Alta	Altas
3	Baja	Bajas
4	Baja	Altas

- Entropía del Nodo Padre

$$H(S) = \sum pi \log_2(pi) = - [3/4 \log_2(3/4) + 1/4 \log_2(1/4)]$$

$$H(S) = - [0.311 - 0.5] = 0.811$$

- Índice de Gini

$$H(S) = \sum p^2 i = 1 - [(3/4)^2 + (1/4)^2]$$

$$1 - [0.5625 + 0.0625] = 1 - 0.625 = 0.375$$

- Ganancia de información

Hijo 1 = Alta en los mercados 1 y 2 ambos son altas con la entropía en $H(H1)=0$

Hijo 2 = Baja en los mercados 3 y 4, uno es baja y el otro es alta en donde $p=0.5$ para cada uno

$$H(H2) = - [0.5 \log_2(0.5) + 0.5 \log_2(0.5)] = 1 \text{ bit}$$

Entropía ponderada

$$H(S, TV) = 2/4(0) + 2/4(1) = 0.5$$

$$IG=H(s)-H(S,TV)=0.811-0.5=0.311$$

De acuerdo a los resultados obtenidos la entropía del nodo padre es de 0.811, con un índice de gini del 0.375 y con un resultado en la ganancia de la información por cada variable del 0.311

Finalmente la variable que se identifica con mayor peso en el árbol es la variable TV puesto que es dominante en ambos modelos por su mayor presencia con las ventas; esta variable en el árbol de decisión toma el liderazgo por su capacidad de reducir el error, por otro lado también se puede deducir que la regresión lineal presenta el coeficiente más alto. Basándonos en dichos resultados e interpretaciones se deduce que es preferible usar la regresión lineal cuando el objetivo es tener como resultado una relación matemática coherente y estable, sin embargo los árboles de decisión permiten tomar segmentos de mercado específicos y relaciones no lineales más complejas entre las variables.