Unidad: Manipulación y Preparación de Datos

Fundamentos teóricos para Ciencia de Datos

Nicolás Sidicaro
Abril 2025

Agenda para Hoy

1. **Limpieza de datos**
    - Tratamiento de valores faltantes
    - Detección y manejo de outliers
2. **Transformación de variables**
    - Tipos de transformaciones
    - Creación de variables derivadas
3. **Normalización y estandarización**
    - Escalado de variables
    - Métodos principales
4. **Joins y combinación de fuentes de datos**
    - Tipos de joins
    - Consideraciones prácticas

1. Limpieza de datos: Introducción

¿Por qué es importante la limpieza de datos?

- Los datos del mundo real rara vez vienen "limpios" y listos para análisis
- El proceso de limpieza puede consumir hasta 80% del tiempo en un proyecto de datos
- "Garbage in, garbage out" - los resultados del análisis serán tan buenos como los datos de entrada
- La limpieza de datos ayuda a evitar sesgos y conclusiones erróneas

Principales problemas a abordar:

- Valores faltantes (NA, NULL, espacios en blanco, etc.)
- Valores atípicos (outliers)
- Inconsistencias y errores de formato
- Duplicados
- Problemas de codificación y tipo de datos

Tratamiento de valores faltantes

Tipos de datos faltantes:

1. **MCAR (Missing Completely At Random)**
    - La probabilidad de que falte un valor es independiente de todas las variables
    - Ejemplo: Un sensor que falla aleatoriamente
2. **MAR (Missing At Random)**
    - La probabilidad de que falte un valor depende de valores observados de otras variables
    - Ejemplo: Personas con ingresos altos tienden a no reportar su patrimonio
3. **MNAR (Missing Not At Random)**
    - La probabilidad de que falte un valor depende del valor que falta
    - Ejemplo: Personas con ingresos muy altos tienden a no reportar su ingreso

Tratamiento de valores faltantes

Detección de valores faltantes:

- Identificación de valores faltantes explícitos (NA, NULL)
- Identificación de valores faltantes implícitos (códigos como -999, espacios en blanco)
- Análisis de patrones de faltantes (¿hay alguna estructura?)
- Visualización de la matriz de valores faltantes

Métricas importantes:

- Porcentaje de valores faltantes por variable
- Porcentaje de valores faltantes por observación
- Correlación entre patrones de valores faltantes

Tratamiento de valores faltantes

Estrategias de manejo:

1. **Eliminación**
    - Eliminar filas (listwise deletion)
    - Eliminar columnas (variables con demasiados faltantes)
    - **Ventajas**: Simple, mantiene la distribución de los datos
    - **Desventajas**: Pérdida de información, posible sesgo si no es MCAR
2. **Imputación simple**
    - Media, mediana, moda
    - Último valor observado (LOCF - Last Observation Carried Forward)
    - Interpolación lineal
    - **Ventajas**: Preserva el tamaño de la muestra
    - **Desventajas**: Subestima la varianza, distorsiona relaciones entre variables

Tratamiento de valores faltantes

Estrategias de manejo (continuación):

1. **Imputación múltiple**
    - Genera múltiples conjuntos de datos completos con diferentes valores imputados
    - Analiza cada conjunto y combina resultados
    - **Ventajas**: Preserva incertidumbre, menor sesgo
    - **Desventajas**: Mayor complejidad computacional

2. **Imputación basada en modelos**
    - Regresión
    - k-Nearest Neighbors (k-NN)
    - Algoritmos de machine learning (random forests, etc.)
    - **Ventajas**: Aprovecha relaciones entre variables
    - **Desventajas**: Riesgo de sobreajuste

Tratamiento de valores faltantes

Estrategias de manejo (continuación):

1. **Variables indicadoras**
    - Crear variables dummy que indiquen si un valor estaba originalmente faltante
    - **Ventajas**: Mantiene información sobre el patrón de datos faltantes
    - **Desventajas**: Aumenta dimensionalidad

Detección y manejo de outliers

¿Qué son los outliers?

Valores atípicos que se desvían significativamente del resto de observaciones.

Tipos de outliers:

1. **Univariados**: Extremos en una sola variable
2. **Multivariados**: Combinaciones inusuales de valores en múltiples variables
3. **Contextual/condicional**: Valores anómalos en un contexto específico. Por ejemplo, dentro de un grupo

Causas comunes:

- Errores de medición o recolección
- Errores de procesamiento o entrada de datos
- Variabilidad natural extrema
- Fraude o comportamiento anómalo intencional
- Cambios en el proceso generador de datos

Detección y manejo de outliers

Métodos de detección:

1. **Métodos estadísticos**

   - Rango intercuartílico (IQR)
     - Outliers leves: < Q_1 - 1.5×IQR o > Q_3 + 1.5×IQR
     - Outliers extremos: < Q_1 - 3×IQR o > Q_3 + 3×IQR
   - Z-score: |z| > 3 (asumiendo normalidad)
   - Test de Grubbs $G = \frac{|x_{max} - \overline{x}|}{s}$ (o con mínimo)

2. **Métodos basados en densidad y distancia (ML) - Menos usuales**

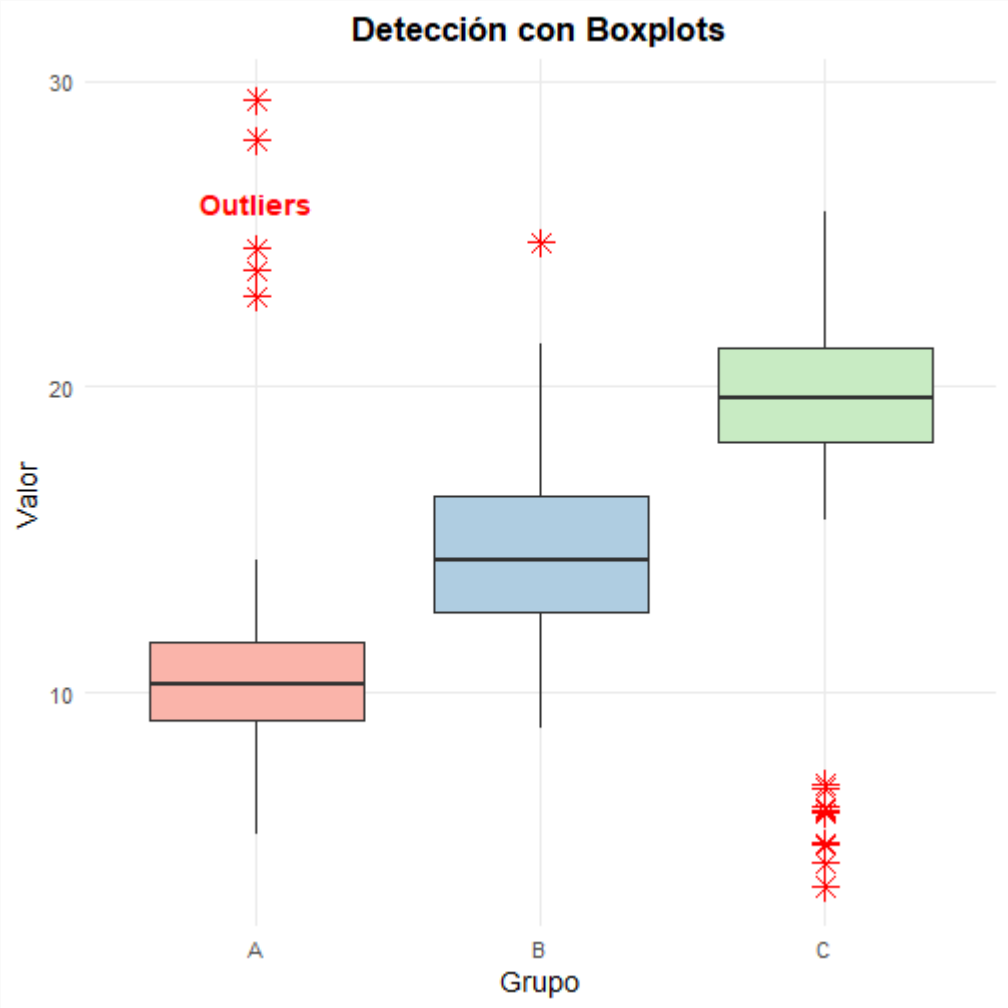
   - DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
   - LOF (Local Outlier Factor)
   - Distancia de Mahalanobis
   - k-Nearest Neighbors (k-NN)

Detección y manejo de outliers

Métodos de detección (continuación):

1. **Métodos de visualización**
    - Boxplots
    - Histogramas y densidades
    - Scatterplots
    - Gráficos Q-Q

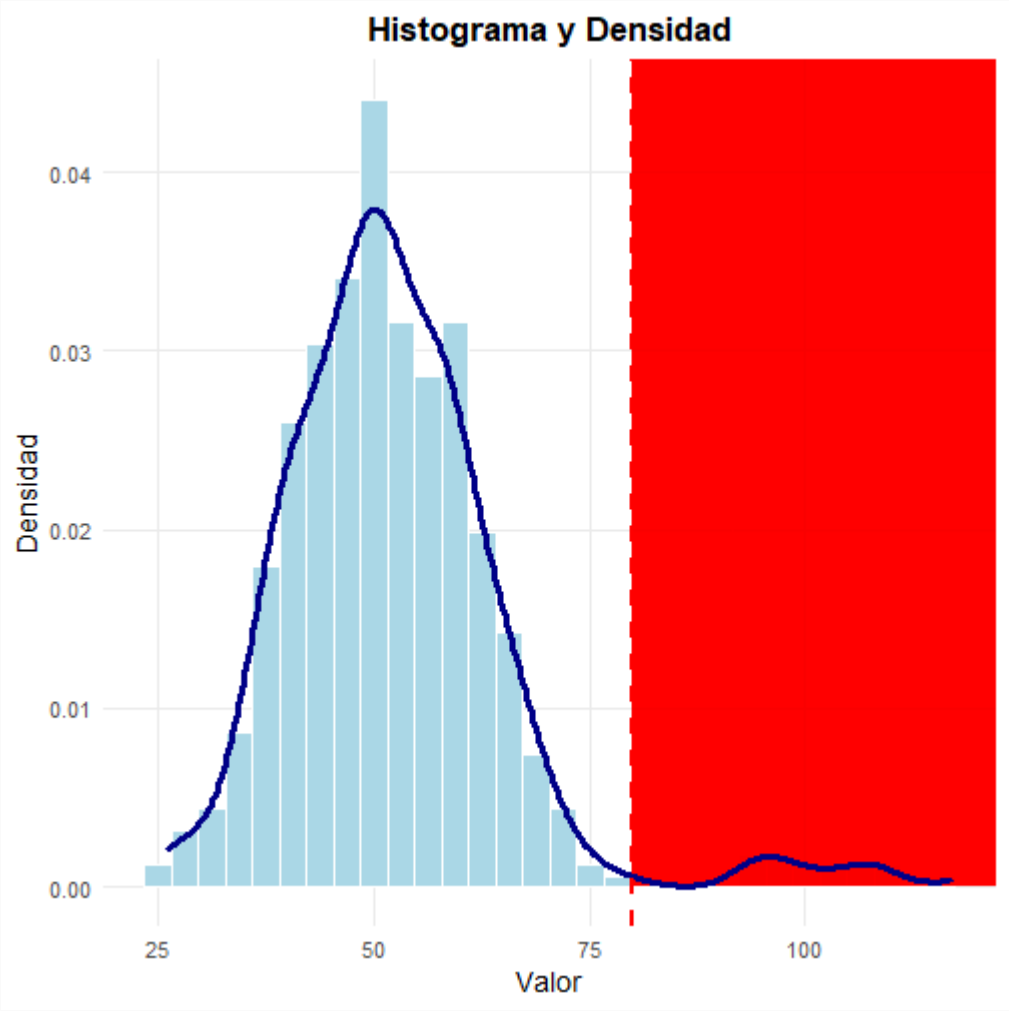
Boxplots



Boxplot

- Muestran la **distribución** de los datos
- Identifican valores atípicos (puntos rojos)
- Se basan en el **rango intercuartílico (IQR)**
- Formula: Valores > Q3 + 1.5×IQR o < Q1 - 1.5×IQR
- Son robustos a distribuciones no normales
- Permiten comparar entre grupos

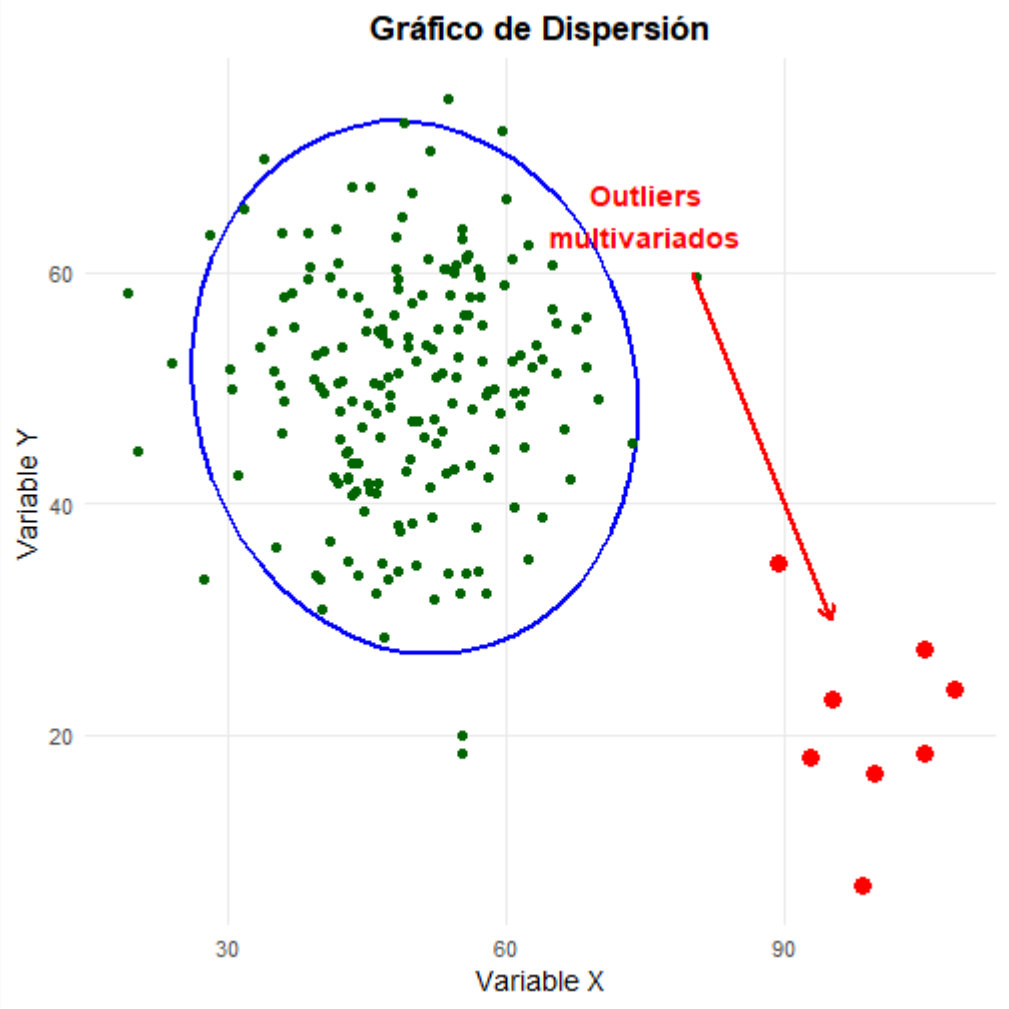
Histogramas y Densidades para Outliers



Histogramas y Densidades para Outliers

- Muestran la **forma de la distribución**
- Outliers aparecen como:
  - Valores aislados
  - Colas largas
  - Separados del cuerpo principal
- El umbral (línea roja) indica:
  - Q3 + 1.5×IQR para asimetría positiva
  - Q1 - 1.5×IQR para asimetría negativa
- Área sombreada contiene outliers

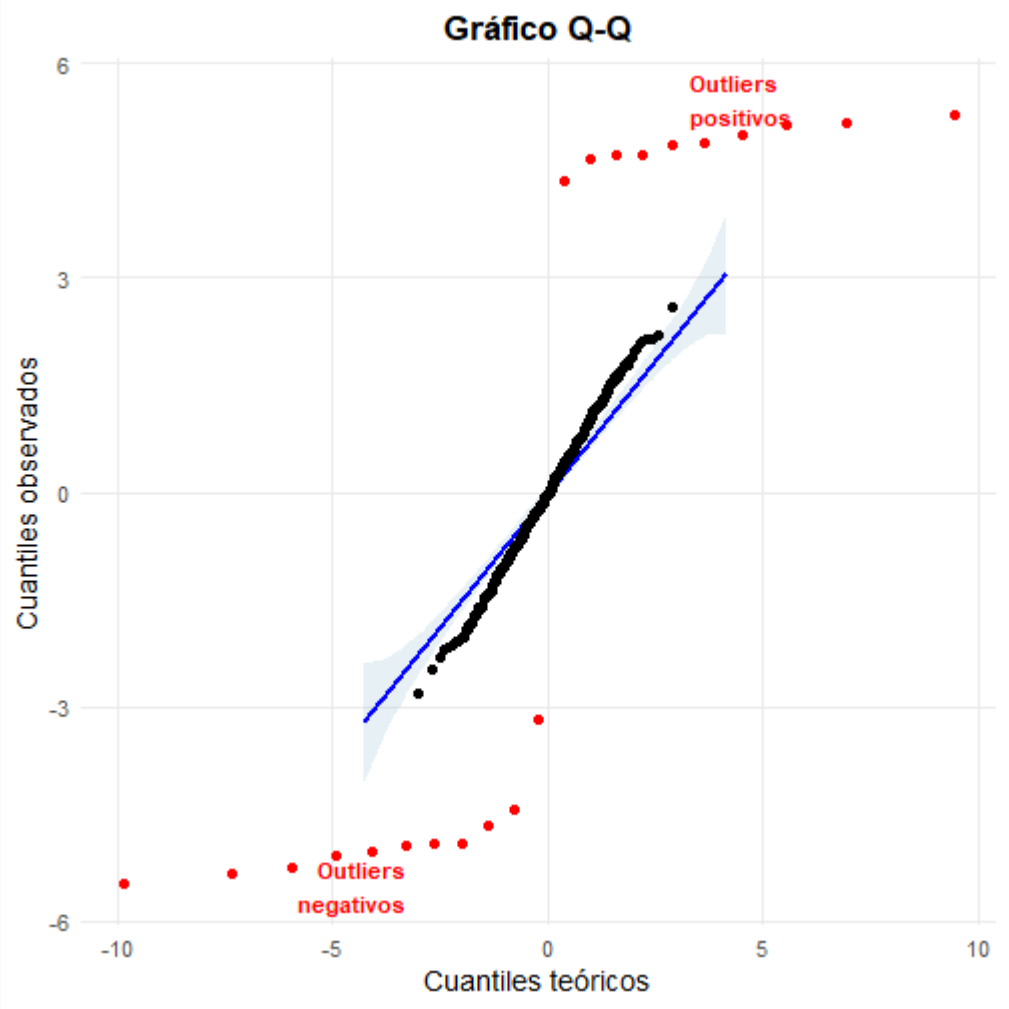
Gráficos de Dispersión para Outliers



Gráficos de Dispersión para Outliers

- Detectan **outliers multivariados**
- Algunos puntos pueden ser:
  - Normales en cada variable individual
  - Atípicos en su combinación
- **Métodos de detección:**
  - Distancia de Mahalanobis
  - Elipses de confianza (95%)
  - Técnicas de clustering
- Los puntos fuera de la elipse azul son **estadísticamente atípicos**
- Permiten visualizar relaciones anómalas entre variables

Gráficos Q-Q para Evaluar Normalidad y



Gráficos Q-Q para Evaluar Normalidad y

- Comparan los **cuantiles empíricos** con los teóricos de una distribución normal
- Outliers aparecen como:
  - Puntos que se **desvían** de la línea diagonal
  - Puntos **fuera** de las bandas de confianza
- Bandas azules muestran el rango de variación normal (95%)
- Útiles para:
  - Verificar supuestos de normalidad
  - Identificar asimetrías
  - Detectar anomalías en ambos extremos

Detección y manejo de outliers

Estrategias de manejo:

1. **Retención**
    - Mantener outliers si representan fenómenos reales de interés
    - Importante en detección de fraude, análisis de riesgos, etc.
2. **Eliminación**
    - Remover observaciones con outliers
    - Adecuado cuando representan errores o no son parte de la población objetivo

Detección y manejo de outliers

Estrategias de manejo:

1. **Tratamiento**
    - Winsorización: reemplazar outliers por valores en percentiles extremos (ej. 5% y 95%)
    - Transformación: aplicar transformaciones que reduzcan el impacto (log, raíz cuadrada)
    - Discretización: convertir la variable continua en categorías
2. **Modelado robusto**
    - Usar métodos estadísticos robustos menos sensibles a outliers
    - Ejemplo: regresión robusta, mediana en lugar de media

2. Transformación de variables

¿Por qué transformar variables?

- Cumplir supuestos de modelos estadísticos (normalidad, linealidad, homocedasticidad)
- Manejar diferentes escalas y unidades
- Reducir el impacto de outliers
- Mejorar la interpretabilidad
- Crear características más informativas para el modelado

Tipos de transformaciones:

1. **Transformaciones matemáticas simples**
2. **Transformaciones no paramétricas**
3. **Creación de variables derivadas**

Transformaciones matemáticas simples

- **Logarítmica**: $\log(x)$ o $\ln(x)$
  - Útil para: datos con asimetría positiva, relaciones multiplicativas, estabilizar varianza
  - Aplicaciones: ingresos, precios, poblaciones
  - Consideración: requiere valores positivos (usar $\log(x + c)$ para valores no positivos)
- **Raíz cuadrada**: $\sqrt{x}$
  - Útil para: datos de conteo, datos con asimetría positiva moderada
  - Menos agresiva que la transformación logarítmica
  - Consideración: requiere valores no negativos
- **Inversa**: $1/x$ o $-1/x$
  - Útil para: datos con asimetría positiva severa
  - Consideración: sensible a valores cercanos a cero

Transformaciones no paramétricas

Transformación cuantílica:

- Mapea valores a sus cuantiles teóricos en una distribución objetivo
- Preserva el orden pero modifica las distancias entre observaciones
- Útil para: forzar una distribución específica sin conocer la transformación paramétrica

Discretización:

- Convierte variables continuas en categóricas
- Métodos: cortes por cuantiles, intervalos iguales, clustering, etc.
- Útil para: variables con relaciones no lineales, reducción de ruido
- Desventaja: pérdida de información y poder estadístico

Creación de variables derivadas

Técnicas comunes:

1. **Interacciones**
   - Productos entre variables: $x_1 \times x_2$
   - Capturan efectos conjuntos no aditivos
   - Ejemplo: edad × educación puede predecir ingresos mejor que ambas por separado
2. **Polinomios**
   - Términos cuadráticos, cúbicos: x^2, x^3
   - Capturan relaciones no lineales
   - Ejemplo: la relación entre edad y productividad puede ser curvilínea
3. **Ratios y proporciones**
   - División entre variables: x_1/x_2
   - Útiles en análisis financiero, demografía, etc.
   - Ejemplos: ROI, velocidad (distancia/tiempo), densidad poblacional

Creación de variables derivadas

Técnicas comunes (continuación):

1. **Variables temporales**
    - Extraer componentes de fechas: año, mes, día, día de la semana
    - Diferencias entre fechas
    - Indicadores de eventos especiales (feriados, fin de semana)
    - Rezagos y diferencias: x_{t-1}, $x_t - x_{t-1}$

2. **Agregaciones**
    - Sumas, promedios, máximos por grupos
    - Ejemplo: ventas totales del último trimestre, temperatura media por región
    - Ventanas móviles (rolling window)
    - Ejemplo: suma acumulada, media movil de los últimos tres períodos

3. Normalización y estandarización

¿Por qué escalar variables?

- Evitar que variables con magnitudes mayores dominen el análisis
- Requisito para muchos algoritmos de machine learning (k-means, SVM, redes neuronales)
- Mejora la convergencia de algoritmos de optimización
- Facilita la comparación e interpretación de coeficientes

Diferencias conceptuales:

- **Normalización**: generalmente se refiere a escalar a un rango específico [0,1] o [-1,1]
- **Estandarización**: transformar para tener media 0 y desviación estándar 1

Nota: Estos términos a veces se usan indistintamente o con definiciones diferentes según el contexto

Métodos principales de escalado

Normalización Min-Max:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Escala los datos al rango [0,1]
- Para escalar a [a,b]: $x_{norm} = a + \frac{(x - x_{min})(b - a)}{x_{max} - x_{min}}$
- **Ventajas**: Preserva relaciones exactamente, rango conocido
- **Desventajas**: Sensible a outliers

Métodos principales de escalado

Estandarización (Z-score):

$$x_{std} = \frac{x - \mu}{\sigma}$$

- Transforma a distribución con media 0 y desviación estándar 1
- **Ventajas**: Menos afectada por outliers que min-max, útil para métodos que asumen normalidad
- **Desventajas**: No garantiza un rango específico, valores pueden ser muy grandes si hay outliers extremos

Métodos principales de escalado

Escalado robusto:

$$x_{rob} = \frac{x - \text{mediana}(x)}{\text{IQR}(x)}$$

- Similar a z-score pero usa estadísticos robustos
- **Ventajas**: Menos sensible a outliers
- **Desventajas**: Puede ser menos eficiente con datos realmente normales

Métodos principales de escalado

Normalización L1 (Manhattan):

$$x_{L1} = \frac{x}{\sum_{i=1}^{n} |x_i|}$$

- Escala para que la suma de valores absolutos sea 1
- Útil en procesamiento de texto, problemas de optimización L1

Normalización L2 (Euclidiana):

$$x_{L2} = \frac{x}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

- Escala para que la norma euclidiana sea 1
- Útil en métodos basados en distancias, SVM, PCA

Consideraciones para el escalado

¿Cuándo usar cada método?

- **Min-Max**: cuando se requiere un rango específico
- **Z-score**: métodos que asumen normalidad, PCA, regresión
- **Escalado robusto**: cuando hay outliers y no se quiere que afecten el escalado
- **Normalización L1/L2**: en problemas de optimización, SVM, regularización

4. Combinación de fuentes de datos

Conceptos fundamentales:

- **Clave primaria**: Identifica de manera única cada registro en una tabla
- **Clave foránea**: Referencia a una clave primaria en otra tabla
- **Relaciones**:
  - Uno a uno (1:1)
  - Uno a muchos (1:N)
  - Muchos a muchos (N:M)

Desafíos comunes:

- Diferentes formatos y estructuras (character vs numeric)
- Inconsistencias en claves (codigos de países)
- Duplicados
- Granularidad diferente (niveles de agregación)
- Períodos temporales no alineados (N° de semana vs fecha del lunes)
- Conflictos de nombres de columnas

Tipos de joins

Inner Join:

- Retorna filas donde hay coincidencia en ambas tablas
- Descarta filas sin coincidencia
- Ejemplo: Combinar ventas con información de clientes, solo para clientes con ventas

```
A     B          A ⋈ B
1     3    →     1,3
2     5          2,5
4
```

Left Join:

- Retorna todas las filas de la tabla izquierda y las coincidencias de la derecha
- Rellena con NA/NULL donde no hay coincidencia
- Ejemplo: Mantener todos los clientes, incluso los que no tienen ventas

```
A     B          A ⟕ B
1     3    →     1,3
2     5          2,5
4                4,NA
```

Tipos de joins

Right Join:

- Retorna todas las filas de la tabla derecha y las coincidencias de la izquierda
- Rellena con NA/NULL donde no hay coincidencia
- Ejemplo: Mantener todos los productos, incluso los que no tienen ventas

```
A     B          A ⋈ B
1     3    →     1,3
2     5          2,5
      6          NA,6
```

Tipos de joins

Full Join (Outer Join):

- Retorna todas las filas de ambas tablas
- Rellena con NA/NULL donde no hay coincidencia
- Ejemplo: Ver todos los clientes y todos los productos, con o sin ventas

```
A     B          A ⟗ B
1     3    →     1,3
2     5          2,5
4                4,NA
      6          NA,6
```

Tipos de joins

Cross Join (Producto Cartesiano):

- Combina cada fila de la primera tabla con cada fila de la segunda
- Resultado tiene n×m filas (donde n y m son los tamaños de las tablas)
- Útil para generar todas las combinaciones posibles
- Ejemplo: Todas las combinaciones posibles de productos y tiendas

```
A     B          A × B
1     a     →    1,a
2     b          1,b
                 2,a
                 2,b
```

Técnicas especializadas de combinación

Fuzzy matching:

- Une registros basados en similitud aproximada, no coincidencia exacta
- Útiles cuando hay errores tipográficos, formatos inconsistentes, etc.
- Técnicas: distancia de edición, similitud fonética, n-gramas

> Nota: no necesariamente va a estar bien, es más complicado de utilizar y es necesario ir probando con distintos parámetros de distancia máxima, así como también es necesario aceptar cierto margen de error

Consideraciones prácticas para joins

Optimización:

- Indexar columnas de unión
- Filtrar antes de unir
- Seleccionar solo columnas necesarias
- Considerar el orden de las operaciones

Resolución de conflictos:

- Estrategias para nombres de columnas duplicados:
  - Renombrar (prefijos/sufijos)
  - Seleccionar una
  - Crear una nueva combinando ambas
- Estrategias para valores conflictivos:
  - Priorizar una fuente
  - Usar el valor más reciente

Consideraciones prácticas para joins

Validación post-join:

- Verificar número de filas resultante
- Comprobar integridad de claves
- Buscar valores faltantes inesperados

Ejemplos de validación:

- Si A tiene n filas y B tiene m filas:
  - Inner join: $\leq \min(n,m)$ filas
  - Left join: exactamente n filas
  - Right join: exactamente m filas
  - Full join: entre $\max(n,m)$ y $n+m$ filas
  - Cross join: exactamente $n \times m$ filas