Calidad y Limpieza de Datos en Ciencia de Datos

Sergio M. Nava Muñoz
s3rgio.nava@gmail.com
CIMAT/INFOTEC

March 5, 2025

...os de la Presentación

1. Comprender la importancia de la limpieza de datos.
2. Identificar los errores más comunes en datasets.
3. Aprender técnicas prácticas de exploración y validación de datos.
4. Profundizar en:
    - **Validación de supuestos estadísticos**.
    - **Manejo de datos faltantes**.
    - **Identificación y tratamiento de valores atípicos (outliers)**.

I: ¿Por qué es importante la limpieza de datos?

- **Garbage In, Garbage Out (GIGO)**
  - La calidad de la salida de un sistema depende directamente de la calidad de los datos de entrada.
  - Si se alimentan modelos, algoritmos o análisis con datos incorrectos, incompletos o sesgados, los resultados serán igualmente defectuosos, sin importar cuán sofisticados sean los métodos utilizados.
- **Mito de la Robustez:** No todos los métodos estadísticos son robustos a violaciones de supuestos.
- **Errores comunes en la literatura científica:**
  - Falta de validación de supuestos estadísticos.
  - Falta de reporte de pruebas de normalidad y homogeneidad de varianza.
  - Uso de métodos no adecuados para la distribución de los datos.
- **Impacto en la replicabilidad de estudios científicos.**

I.1: Validación de Supuestos Estadísticos

La mayoría de los métodos estadísticos están desarrollados para hacer inferencia y esta dependen de ciertos supuestos sobre los datos. Antes de realizar análisis, debemos validar los supuestos:

lidad

mayoría de los modelos estadísticos tradicionales asumen datos normalmente distribuidos.

todos de validación:

- Histogramas y gráficos Q-Q.
- Pruebas estadísticas: Shapiro-Wilk, Kolmogorov-Smirnov.

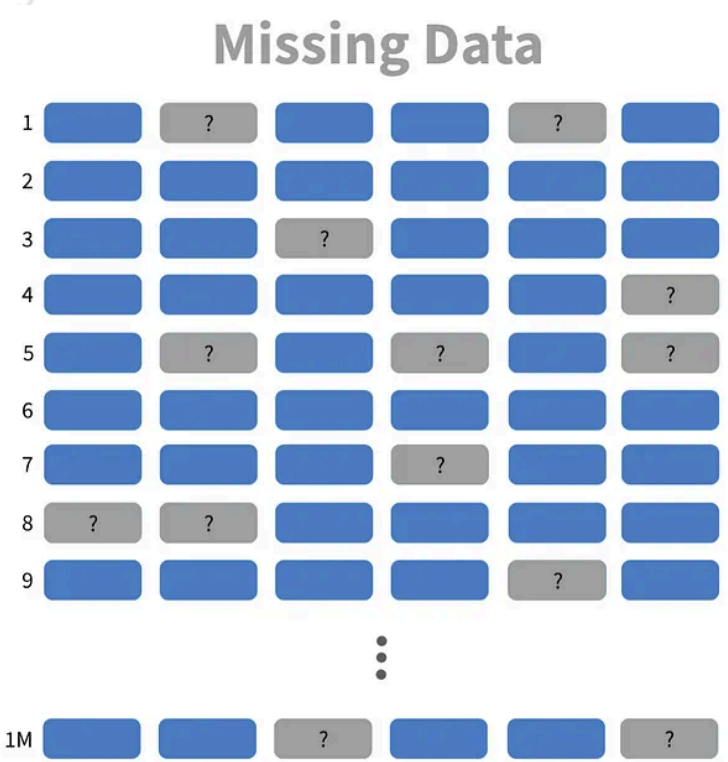
2. **Homogeneidad de varianza**
   - Métodos como ANOVA requieren que las varianzas de los grupos sean similares.
   - Se puede evaluar con pruebas como Levene o Bartlett.
3. **Independencia de los datos**
   - En modelos de regresión y ANOVA, las observaciones deben ser independientes.
   - Se puede evaluar con la prueba de Durbin-Watson en series temporales.
4. **Ausencia de multicolinealidad**
   - En regresión múltiple, la alta correlación entre variables predictoras puede distorsionar los coeficientes.
   - Se evalúa con el Factor de Inflación de Varianza (VIF).
5. **Linealidad**
   - En regresión, la relación entre variables debe ser lineal.
   - Se puede verificar visualmente con gráficos de dispersión.

2: Evaluación de la Estructura de Datos

- Tipos de formatos comunes:
  - **CSV (Comma-Separated Values)**
  - **JSON (JavaScript Object Notation)**
  - **XML (Extensible Markup Language)**
  - **Excel (XLS, XLSX)**
- Técnicas para interpretar la estructura:
  - Inspeccionar cabeceras y delimitadores.
  - Verificar la presencia de nombres de columnas y etiquetas.

3: Validación de Campos y Valores

- **Errores comunes en los datos:**
  - Valores faltantes (`NULL`, `N/A`, `NaN`).
  - Datos fuera de rango (ejemplo: valores negativos en precios).
  - Errores de formato (ejemplo: fechas en formatos inconsistentes).
  - Desviaciones inesperadas en distribuciones.



3.1: Tipos de Datos Faltantes

faltantes pueden distorsionar análisis y modelos. Existen distintos enfoques para manejarlos.

MCAR (Missing Completely at Random)

La falta de datos no está relacionada con ninguna variable del estudio.

✅ **Ejemplos:**

- **Respuestas omitidas en una encuesta**: En un cuestionario de satisfacción, algunos participantes no responden debido a un problema técnico en la plataforma de recolección de datos, afectando aleatoriamente a cualquier encuestado.
- **Sensor defectuoso**: Un sensor ambiental deja de registrar temperatura en momentos aleatorios debido a fallas esporádicas en la transmisión de datos, sin relación con las condiciones climáticas.

3.2: Datos Faltantes Condicionados a Otras Variables

MAR (Missing at Random)

La falta de datos depende de otras variables observadas.

✅ **Ejemplos:**

- **Datos de ingresos en una encuesta**: En un estudio de mercado, las personas más jóvenes tienen más probabilidades de omitir su ingreso mensual en el cuestionario, pero dentro de cada grupo de edad, la falta de datos es aleatoria.
- **Exámenes médicos faltantes**: En un estudio de salud, los pacientes con menor edad pueden tener más registros faltantes de presión arterial porque los médicos tienden a medirla con menos frecuencia en personas jóvenes.

3.3: Datos Faltantes No Aleatorios

MNAR (Missing Not at Random)

La ausencia de datos está relacionada con el valor en sí mismo.

✅ **Ejemplos:**

- **Ingresos altos no reportados**: En una encuesta sobre salarios, las personas con ingresos muy altos pueden optar por no revelar su sueldo, lo que genera un sesgo en la distribución de datos.
- **Pacientes con enfermedades graves**: En un estudio médico, los pacientes con enfermedades graves pueden no asistir a sus citas de seguimiento, lo que provoca datos faltantes que están relacionados con su estado de salud.

3.4: Estrategias de Manejo de Datos Faltantes

Diversas estrategias para tratar los datos faltantes, dependiendo de la cantidad y el patrón de ausencia.

Eliminación de Filas o Columnas

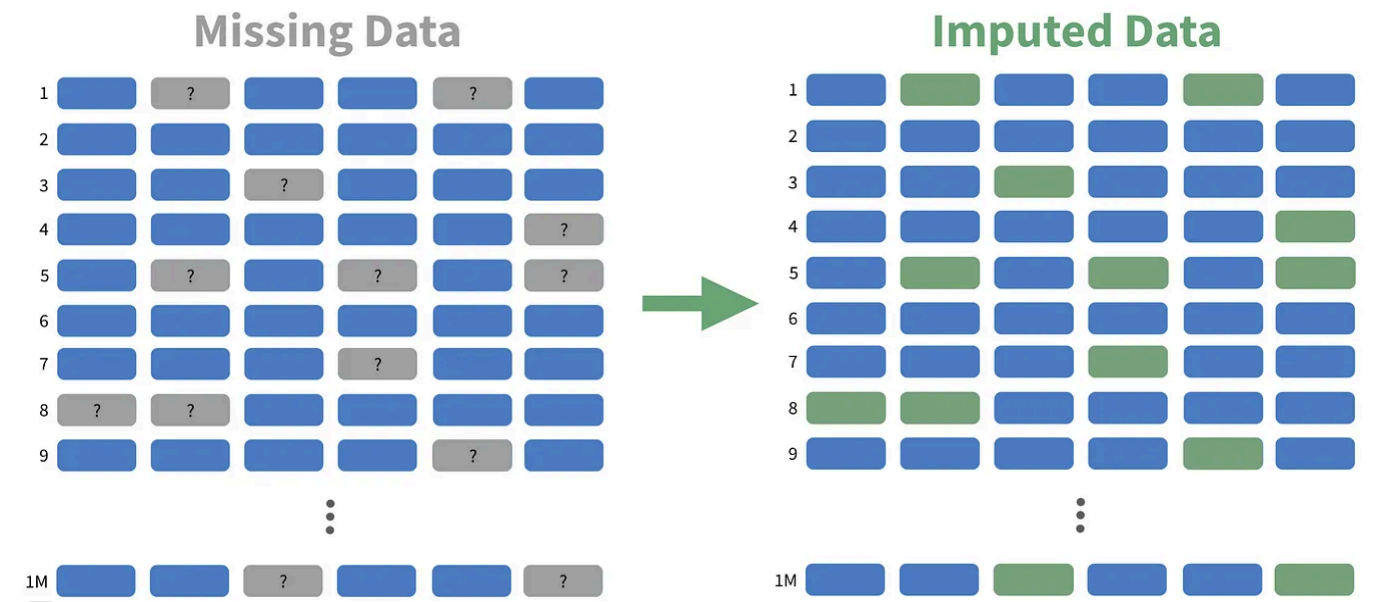
Características:

- Útil si la cantidad de datos faltantes es pequeña.
- No recomendable si se pierde demasiada información.

Ejemplo: Si una encuesta tiene solo un 2% de valores faltantes en una columna específica, eliminarlos puede ser una solución viable. Sin embargo, si el 40% de los datos están ausentes, la eliminación puede comprometer la validez del análisis.

3.5: Imputación de Valores

Eliminar datos no es una opción, se pueden utilizar técnicas de imputación para completar los valores faltantes.



Métodos de Imputación:

Media/Mediana

- Rellena los valores con la media o mediana del conjunto de datos.
- Útil para variables continuas con pocos valores faltantes.

Interpolación

- Usa valores adyacentes para estimar los datos faltantes.
- Aplicable en series temporales donde los valores cercanos tienen una relación lógica.

Modelos Predictivos

en usar **regresiones** o **algoritmos de machine learning** (KNN, árboles de decisión) para
alores faltantes.

e un conjunto de datos suficientemente grande y con correlaciones fuertes entre variables.

3.6: Métodos de Imputación Avanzados e Impacto en los Modelos

Métodos Basados en Múltiples Imputaciones

Características:
- Genera varias imputaciones para reflejar la incertidumbre sobre los valores reales.
- Métodos como **MICE (Multivariate Imputation by Chained Equations)** son ampliamente usados en investigación.

Impacto de los Datos Faltantes en Modelos

Riesgos potenciales:
- La eliminación de datos puede **sesgar los resultados** si los datos faltantes no son aleatorios.
- La imputación inadecuada puede **reducir la varianza artificialmente**, afectando la precisión de los modelos.
- Siempre se deben **reportar las estrategias usadas** para manejar datos faltantes en estudios y análisis.

4: Estadísticas Descriptivas y Validación Numérica

- **Medidas clave para detectar anomalías:**
  - **Mínimo y máximo:** Detecta valores fuera de rango.
  - **Media y mediana:** Evalúa el centro de la distribución.
  - **Desviación estándar:** Identifica dispersión anormal.
  - **Histogramas y gráficos de caja:** Muestran patrones inusuales.

5: Visualización para la Calidad de Datos

- **Histogramas:** Para entender distribuciones y valores **atípicos**.
- **Gráficos de series temporales:** Para detectar patrones estacionales o datos faltantes.
- **Ejemplo de outliers en datos financieros:**
  - En datos de publicidad digital (PPC), una brecha en valores puede indicar errores en la recopilación.

5.1: Identificación y Manejo de Outliers

Outliers son valores que se alejan significativamente de la distribución general de los datos y pueden influir negativamente en los análisis.

¿Cómo identificar outliers?

1. **Métodos basados en estadística descriptiva**
   - Rango intercuartil (IQR):
     - Outliers = valores fuera del rango `[Q1 - 1.5*IQR, Q3 + 1.5*IQR]`.
   - Desviación estándar:
     - Un dato puede considerarse un outlier si está a más de 3σ de la media.
2. **Métodos gráficos**
   - **Boxplots**: Detectan valores atípicos visualmente.
   - **Histogramas**: Pueden revelar datos extremos en la distribución.
   - **Gráficos de dispersión**: Útiles para detectar outliers en relaciones entre variables.
3. **Modelos estadísticos y aprendizaje automático**
   - **Isolation Forest**: Algoritmo basado en la capacidad de aislar outliers con árboles de decisión.
   - **Local Outlier Factor (LOF)**: Método basado en densidad para detectar outliers.
   - **DBSCAN**: Clustering que permite identificar puntos atípicos.

5.2: Estrategias para Manejar Outliers

1. Eliminación de Outliers

- Se recomienda solo si hay evidencia de que los valores son errores de medición.
- Puede distorsionar la distribución de datos si se eliminan valores legítimos.

2. Transformación de Datos

- **Escalado Logarítmico**: Reduce la influencia de outliers en distribuciones sesgadas.
- **Winsorización**: Sustituye outliers por valores límite.

3. Modelado Robusto

- Métodos de regresión robusta pueden mitigar la influencia de valores extremos.
- Algoritmos de machine learning como Random Forest tienden a ser menos sensibles a outliers.

5: Series Temporales y Detección de Errores

- **Ejemplo: Datos de Wikipedia**
  - Detectar días con valores inusuales en tráfico web.
  - Identificar días con datos duplicados o faltantes.

Conclusiones

- **La limpieza de datos NO es opcional.**
- Validar datos con inspección manual y análisis estadístico.
- Usar herramientas adecuadas para detectar errores en datos estructurados y series temporales.
- Reportar siempre los procesos de limpieza y validación en los estudios científicos.
- **La validación de supuestos estadísticos es clave** para evitar errores en el análisis de datos.
- **Los outliers NO deben eliminarse automáticamente**. Se debe analizar su origen y posible impacto.
- **La imputación de datos faltantes debe realizarse con métodos adecuados** para evitar sesgos.
- **Usar herramientas estadísticas y de machine learning** para detectar y manejar outliers de forma efectiva.

ıtas y Discusión