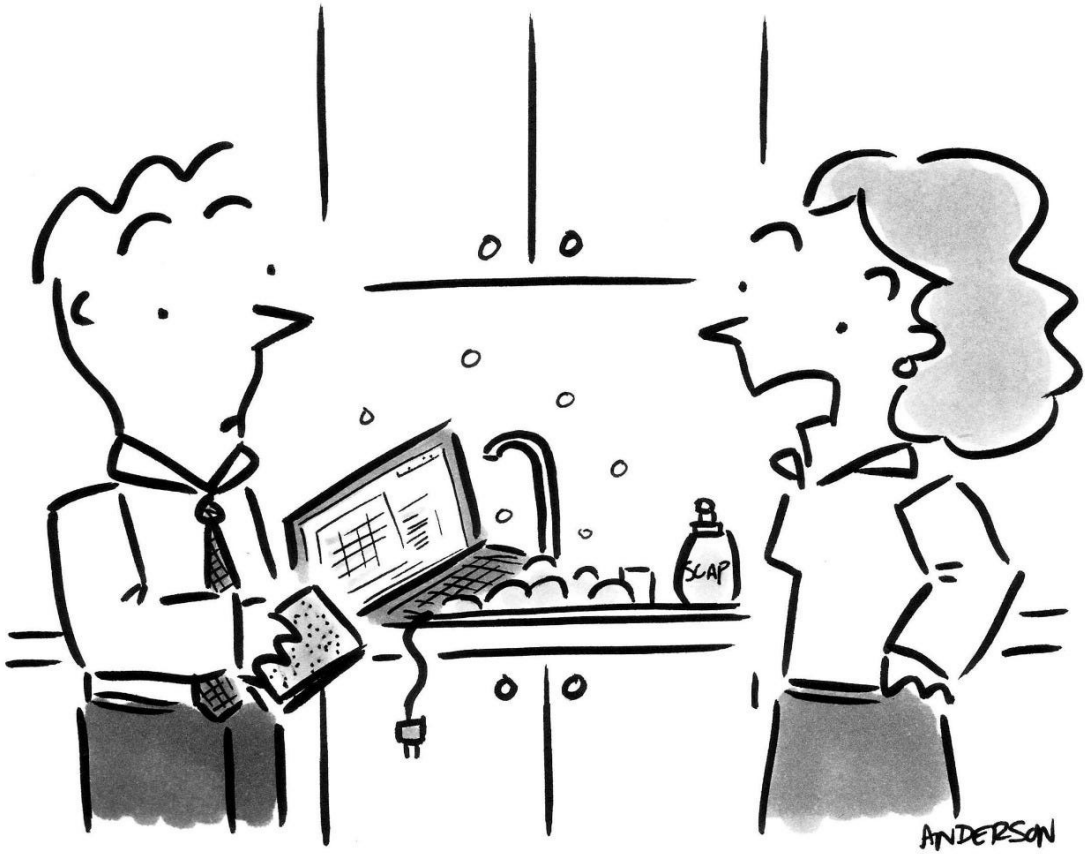
Pontificia Universidad JAVERIANA Bogotá

Educación Continua
Generamos experiencias educativas

Limpieza, Preparación y Curación de Datos

Heriberto Felizzola Jimenez

¿Qué es la limpieza de Datos?



"This is not what I meant when I said 'we need better data cleansing!'"

www.iwaysoftware.com/go/dataquality

Limpieza de Datos

La **limpieza de datos** es el proceso de identificar y corregir problemas dentro de un conjunto de datos, con el objetivo de mejorar su calidad y asegurar que sean adecuados para el análisis o modelado posterior.

Importancia de la limpieza de datos

- Los *“Datos crudos”* suelen contener errores, inconsistencias, valores faltantes y ruidos que afectan la calidad del análisis
- La limpieza asegura que los modelos o análisis trabajen sobre datos representativos y precisos
- Invertir tiempo en la limpieza y curación evita problemas en etapas posteriores del análisis o modelado.
- Los análisis basados en datos sin limpiar pueden llevar a conclusiones imprecisas o erradas

Problemas comunes en los datos

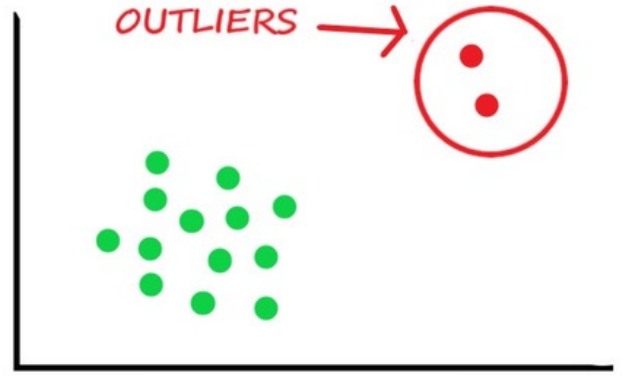
- Valores faltantes (Missing Data)
- Datos duplicados
- Valores atípicos (Outliers)
- Errores tipográficos o de entrada
- Inconsistencias: formatos, categorías, valores numéricos, fechas
- Datos irrelevantes

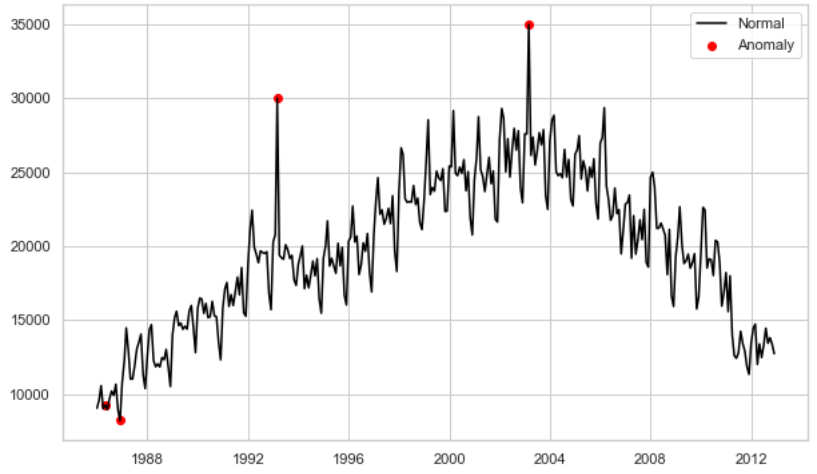
Valores Faltantes

- Son datos ausentes o no registrados en una o más columnas de un conjunto de datos.
- **Tipos de valores faltantes:**
  - **MCAR (Missing Completely at Random)**: Los valores faltan de manera completamente aleatoria.
  - **MAR (Missing at Random)**:
    - Los valores faltan dependiendo de otras variables observadas, pero no de la propia variable con datos faltantes.
    - **Ejemplo**: En una encuesta de salud, la pregunta sobre ingresos anuales está más a menudo sin respuesta entre personas más jóvenes, pero no depende directamente de los ingresos mismos. En este caso, los valores faltantes en la variable de ingresos están relacionados con la edad (otra variable observada), pero no con el propio ingreso.
  - **MNAR (Missing Not at Random)**:
    - La ausencia de datos está relacionada con el valor mismo que falta o con algún factor no observado.
    - **Ejemplo**: En un estudio sobre depresión, las personas con niveles más severos de depresión tienden a no responder preguntas relacionadas con su estado emocional. Los valores faltantes de las respuestas están directamente relacionados con el nivel de depresión, que es precisamente la variable de interés.

Valores Atípicos (Outliers)

- **¿Qué son los outliers?**
  - Son valores que se desvían significativamente del resto de los datos y no siguen el patrón general del dataset.
  - **Ejemplo**: En un análisis de salarios, un valor de $1,000,000 anual en una muestra donde la mayoría de los salarios están entre $30,000 y $80,000.
- **Causas comunes de outliers:**
  - Errores de entrada
  - Fenómenos reales
  - Cambios en el contexto
- **Detección de outliers**:
  - **Métodos gráficos**: Boxplot, histogramas, diagramas de dispersión.
  - **Métodos estadísticos**: cálculo de percentiles (e.g. 1% y 99%), análisis de clúster, modelos predictivos, test de Grubbs.





Error Tipográficos

- **¿Qué son los errores tipográficos?**
  - Son errores en la entrada de datos causados por errores humanos o mecánicos que generan datos incorrectos, mal escritos o inconsistentes.
  - **Ejemplo**: En una base de datos de clientes, "Juan Pérez" puede aparecer como "Juan Perez", "J. Pérez", o "Juaan Pérez".
- **Causas comunes de errores tipográficos:**
  - **Errores humanos**: Entrada manual de datos, donde se cometen faltas de ortografía o errores al teclear.
  - **Falta de estándares**: Falta de una estructura consistente para ingresar información (e.g., diferentes formatos de nombres o fechas).
  - **Integración de fuentes múltiples**: Cuando se combinan datos de diferentes sistemas o bases de datos con formatos y convenciones distintas.

C	D	E	F
site_name	**address**	**zip**	**phone**
Ada S. McKinley Community Services	725 S Wells Street	60607	3852000
Armstrong	2110 W. Greenleaf	60645	5342150
Armstrong , G.	2110 W. Greenleaf		5342150
Chicago Public Schools Mason, Roswell B.	4216 W. 19th St	60623	5341530
Mason	4216 W. 19th St.	60623	5341530
CHICAGO PUBLIC SCHOOLS MASON, ROSWELL B.	4216 W 19TH ST	60623	5341530
Tarkington	3330 W. 71st St.		5354700
HENRY BOOTH HOUSE ALLISON'S INFANT TODDLER	5522 S RACINE	60636	4363193
Henry Booth House - Allison's Infant Toddler	5522 S Racine Ave		4363193
Henry Booth House Allison's Infant Toddler	5522 S Racine	60636	4363193
Allison's Infant and Toddler Center	5522 S. Racine		4363193
Ashburn School	8300 S. St. Louis	60652	5357860
Ashburn School	8300 S. St Louis Ave.		5357860
ERIE NEIGHBORHOOD HOUSE FCCH-TANIA PEREZ	4502 N CHRISTIANA AVE	60625	5041813
Bass	1140 W. 66th St.		5353275
Bass	1140 W. 66th St.	60621	5353275
Catholic Charities-St Mark	1041 N. Campbell	60622	7726606
Catholic Charities Chicago - St. Mark	1041 N Campbell Avenue	60622	7726606
Easter Seals Gilchrist Marchman	1001 West Roosevelt Rd.	60608	4927402
Easter Seals Society of Metropolitan Chicago - Gilchrist-Marchman	1001 W Roosevelt Rd		9395115
Easter Seals Society of Metropolitan Chicago - Gilchrist-Marchman	1001 W Roosevelt Rd		9395115

Inconsistencias en los Datos

- Son errores o discrepancias en cómo los datos están representados o estructurados, lo que dificulta el análisis y la interpretación correcta.
- **Tipos de inconsistencias comunes:**
  - **Formatos de texto y números**:
    - **Problema**: Valores numéricos almacenados como texto o viceversa.
    - **Ejemplo**: "1000" (texto) vs. 1000 (numérico).
  - **Categorías inconsistentes**:
    - **Problema**: Múltiples etiquetas para una misma categoría.
    - **Ejemplo**: "Sí", "si", "SÍ", "Yes" para representar la misma respuesta.
  - **Unidades y escalas mezcladas**:
    - **Problema**: Mezcla de unidades en los datos (e.g., kilogramos y libras en la misma columna).
    - **Ejemplo**: "70 kg" y "154 lbs" en una columna de peso.
  - **Fechas en diferentes formatos**:
    - **Problema**: Fechas registradas con múltiples formatos o como cadenas de texto.
    - **Ejemplo**: "DD/MM/AAAA", "MM-DD-YYYY" o "12/03/2023" vs. "03/12/2023".

Datos Irrelevantes

- **¿Qué son los datos irrelevantes?**
  - Son datos que no aportan valor al análisis o no están relacionados con el objetivo del estudio.
  - **Ejemplo**: En un análisis de ventas, incluir información sobre el clima cuando no tiene impacto directo en las ventas.
- **Causas comunes de datos irrelevantes:**
  - Recolección excesiva de datos
  - Falta de claridad en el objetivo
  - Integración de fuentes diversas

Curación de Datos

- **¿Qué es la curación de datos?**
  - Es el proceso de preparar, transformar y validar los datos para que sean consistentes, precisos y adecuados para el análisis.
  - **Incluye**: Limpieza, transformación, eliminación de errores, estandarización y enriquecimiento de los datos.
- **Objetivos de la curación de datos:**
  - **Mejorar la calidad de los datos**: Asegurar que los datos sean correctos, completos y útiles.
  - **Eliminar ruido e irrelevancias**: Reducir la complejidad eliminando datos innecesarios.
  - **Preparar los datos para análisis o modelado**: Asegurar que los datos estén listos para ser utilizados en análisis estadísticos o modelos de machine learning.

¿Cómo detectar y manejar los outliers?

- **Mantenerlos**: Si reflejan datos reales y son importantes para el análisis.
- **Eliminar o transformar**:
  - **Recorte (trimming)**: Eliminar los valores que están más allá de un umbral razonable.
  - **Winsorización**: Reemplazar outliers por el valor más cercano dentro del rango aceptable.
  - **Transformación**: Aplicar transformaciones logarítmicas o exponenciales para reducir el efecto de los outliers.

Estrategias para el Manejo de Valores Faltantes

- **Eliminar registros o columnas**: Solo si la cantidad de datos faltantes es pequeña y no afecta significativamente el análisis.
- **Imputación de valores**:
  - **Media/Mediana/Moda**: Rellenar con el valor promedio o mediano.
  - **Imputación basada en modelos**: Usar modelos predictivos para estimar los valores faltantes.
- **Relleno con valores anteriores/posteriores** (Series temporales): Usar valores adyacentes para completar los datos o modelo de serie de tiempo.

Cómo manejar los errores tipográficos

- **Normalización de datos**: Establecer formatos estándar para la entrada de datos (e.g., nombres en mayúsculas, fechas en el mismo formato).
- **Corrección automática**: Usar algoritmos de corrección de texto como **Levenshtein distance** o **difflib** en Python para identificar y corregir errores.
- **Revisión manual**: Inspección de registros con errores detectados automáticamente para asegurar una corrección adecuada.
- **Herramientas de deduplicación**: Implementar herramientas que identifiquen registros duplicados o similares para fusionarlos o corregirlos.

Manejo de Inconsistencias en los Datos

- **Tipos de inconsistencias comunes:**
  - **Formatos de texto y números**:
    - **Solución**: Convertir columnas a los tipos de datos correctos (e.g., int, float, string).
  - **Categorías inconsistentes**:
    - **Solución**: Normalizar las categorías con valores únicos y consistentes.
  - **Unidades y escalas mezcladas**:
    - **Solución**: Estandarizar todas las unidades a un mismo sistema (e.g., convertir libras a kilogramos).
  - **Fechas en diferentes formatos**:
    - **Solución**: Convertir todas las fechas a un formato consistente y apropiado (e.g., YYYY-MM-DD).