Introducción al Business Analytics

Week 07: Sobre los datos, su limpieza y exploración

Eduard F. Martinez Gonzalez, Ph.D.

Departamento de Economía, Universidad Icesi

September 8, 2025

Objetivos de aprendizaje

Al finalizar la unidad, el estudiante será capaz de:

- Identificar y clasificar fuentes de datos relevantes para BA.
- Explicar *base de datos*, *data warehouse*, *data lake*, *ETL* y *metadata*.
- Ejecutar una limpieza inicial en R y documentar decisiones.
- Reconocer y tratar problemas comunes en datos (NA, duplicados, inconsistencias, tipado).

Roadmap

1. Datos
   - Fuentes de datos
   - Tipos de datos
   - Estructura de los datos
2. Almacenamiento y Arquitecturas
   - Bases de datos, DW y DL
   - ETL / ELT
3. Diccionario y Metadata

Panorama de fuentes (qué, para qué y limitaciones)

- **Datos abiertos/públicos** (3P/OGD): DANE, SECOP, Geoportal, World Bank.
  - *Usos:* enriquecimiento exógeno (macro, clima, geografía).
  - *Limitaciones:* rezagos de publicación, cambios metodológicos, licencias.
- **Eventos web/app**: pageviews, rutas, CTR, embudos, sesiones.
  - *Usos:* atribución, CRO, segmentación comportamiento.
  - *Limitaciones:* muestreo, bloqueo cookies, *tracking* inconsistente.
- **Solicitados** (encuestas, grupos focales):
  - *Usos:* NPS, CSAT, preferencia declarada, test de concepto.
  - *Limitaciones:* sesgos (no-respuesta, deseabilidad), diseño de ítems.
- **APIs** (Twitter/X, Wikipedia, Yahoo! Finance, Google Maps, etc.):
  - *Usos:* señales externas (tendencias, precios, POIs, tráfico).
  - *Limitaciones:* límites de tasa, cuotas, términos de uso, estabilidad.

Cómo seleccionar la fuente adecuada

Preguntas guía

1. ¿Cuál es el objetivo de negocio/hipótesis? (métrica de éxito)
2. ¿Qué *granularidad* y *frecuencia* mínimos necesito?
3. ¿Qué sesgos esperados hay (cobertura, medición, supervivencia)?
4. ¿Licencia, privacidad y gobernanza (PII, retención, acceso)?

Criterios comparativos

- *Calidad*: completitud, consistencia, exactitud, puntualidad.
- *Costos*: adquisición, limpieza, mantenimiento, *egress*.
- *Trazabilidad*: metadata, diccionario, linaje, controles.

Tipos de datos en BA (más allá de Q vs. Cualitativos)

Texto

- *Tareas*: limpieza, tokenización, TF–IDF, tópicos, *sentiment*.
- *R*: `tidytext`, `quanteda`, `udpipe`.

Imágenes

- *Tareas*: clasificación, detección, OCR, QA visual.
- *R*: `magick`, `torch`, `tesseract`.

Geoespacial

- *Tareas*: joins espaciales, buffers, rutas, heatmaps.
- *R*: `sf`, `terra`, `osmdata`.

Redes

- *Tareas*: centralidad, comunidades, difusión.
- *R*: `igraph`, `tidygraph`, `ggraph`.

Tiempo/series

- *Tareas*: descomposición, forecast, anomalías.
- *R*: `tsibble`, `fable`, `anomalize`.

Tabular clásico

- *Tareas*: EDA, imputación, *feature engineering*.
- *R*: `tidyverse`, `recipes`, `skimr`.

Riesgos y preprocesamiento por tipo

- **Texto:**
  - *Riesgos:* ambigüedad semántica, sarcasmo, múltiples idiomas, emojis y caracteres especiales.
  - *Requieren:* normalización (*lowercase*, quitar tildes), eliminación de *stopwords* y lematización.
- **Imagen:**
  - *Riesgos:* problemas de iluminación, ruido y resolución.
  - *Requieren:* normalización de escala y técnicas de aumentación (rotar, brillo, contraste).
- **Geoespacial:**
  - *Riesgos:* diferencias de proyección o precisión GPS limitada.
  - *Requieren:* transformar a un CRS común, validar topologías y corregir *outliers* espaciales.
- **Redes:**
  - *Riesgos:* nodos aislados o duplicados.
  - *Requieren:* limpiar identificadores, decidir si la red es dirigida/no dirigida y filtrar aristas redundantes.
- **Series de tiempo:** inconsistencias de *timezone*, huecos y estacionalidad; calendarizar, imputar valores faltantes e incorporar ajustes estacionales.

Estructura de los datos

Estructurados

- Datos organizados en filas y columnas con tipos fijos.
- Ejemplos: tabla de clientes en SQL (ID, nombre, edad, ciudad).
- Ventaja: consultas rápidas y consistencia.
- Desafío: poca flexibilidad ante cambios de esquema.

Semi-estructurados

- Datos con cierta organización pero sin un esquema rígido.
- Ejemplos: un archivo JSON con información de compras anidadas por cliente.
- Ventaja: flexibles y fáciles de compartir vía APIs.
- Desafío: requieren normalización y validación antes del análisis.

No estructurados

- Datos sin formato predefinido ni modelo explícito.
- Ejemplos: correos electrónicos, fotos de productos, audios de llamadas.
- Ventaja: riqueza informativa y variedad de fuentes.
- Desafío: necesitan técnicas avanzadas (NLP, visión por computador) para analizarlos.

Roadmap

1. Datos
   - Fuentes de datos
   - Tipos de datos
   - Estructura de los datos
2. Almacenamiento y Arquitecturas
   - Bases de datos, DW y DL
   - ETL / ELT
3. Diccionario y Metadata

Bases de datos, Data Warehouse y Data Lake

Base de datos

- Colección organizada de datos (filas y columnas, tipos definidos).
- Ejemplo: sistema transaccional de un banco (clientes, cuentas, movimientos).
- Ventaja: velocidad de acceso y soporte a operaciones diarias (OLTP).

Data Warehouse (Schema-on-Write)

- Integra y limpia datos de múltiples fuentes antes de cargarlos.
- Optimizado para reportes, KPIs y analítica descriptiva (OLAP).
- Ejemplo: dashboard de ventas con información consolidada de todas las sucursales.

Data Lake (Schema-on-Read)

- Almacena datos en bruto, sin transformación inicial (texto, JSON, imágenes).
- Flexibilidad para analítica avanzada, ML e investigación exploratoria.
- Ejemplo: repositorio en la nube con logs de sensores IoT y archivos multimedia.

ETL / ELT

ETL (Extract–Transform–Load)

- **E**xtract: lectura desde orígenes (BD transaccionales, APIs, archivos).
- **T**ransform: limpieza, estandarización, validaciones de calidad.
- **L**oad: carga en destino estructurado (normalmente un DW).
- Ventaja: datos consistentes y listos para BI.

ELT (Extract–Load–Transform)

- Primero se cargan los datos crudos en el Data Lake.
- Las transformaciones se aplican bajo demanda, en el momento de análisis.
- Ventaja: mayor flexibilidad y escalabilidad para ciencia de datos.

Idea clave: ETL asegura orden y calidad antes de cargar, ELT da más agilidad y potencia para exploración.

Roadmap

1. Datos
   - Fuentes de datos
   - Tipos de datos
   - Estructura de los datos
2. Almacenamiento y Arquitecturas
   - Bases de datos, DW y DL
   - ETL / ELT
3. Diccionario y Metadata

Metadata y Codebook

¿Qué es la metadata?

- "Datos sobre los datos": describe significado, formato, reglas y calidad.
- Permite entender de dónde viene la información y cómo debe usarse.
- Ejemplo: variable *fecha_nacimiento* → tipo: `date`, formato: `YYYY-MM-DD`.

Diccionario de variables (Codebook)

- Documento técnico que explica cada variable de un dataset.
- Contiene: nombre, descripción, tipo de dato, valores posibles y codificación.
- Ejemplo: *estado_civil* → 1=Soltero, 2=Casado, 3=Divorciado, 4=Unión libre.

¿Por qué importa?

- Facilita la trazabilidad y la auditoría de los datos.
- Evita errores de interpretación y duplicación de esfuerzos.
- Clave para el trabajo colaborativo en proyectos de analítica.

Buenas prácticas con Metadata

- Mantener el diccionario actualizado cada vez que cambian las variables.
- Incluir información sobre población, muestreo, tasas de respuesta y estructura del archivo.
- Usar formatos abiertos (CSV, JSON, YAML) para compartir metadata.
- Integrar la metadata en pipelines de ETL/ELT $\rightarrow$ asegura calidad y consistencia.
- Recordar: sin metadata confiable, los datos pierden gran parte de su valor analítico.