

Taller de Consultoría (Nivel Senior)

El Desafío de los Datos Erróneos e Interconectados

Curso: Fundamentos en Ciencia de Datos (Maestría)

Docente: Jorge Iván Padilla-Buriticá

Universidad EAFIT - Periodo 2026-1

Resumen Ejecutivo del Proyecto (Valor: 25 %)

TechLogistics S.A.S.(Ficticio), un gigante del retail tecnológico, ha detectado una erosión en su margen de beneficios y una caída drástica en la lealtad de sus clientes. La junta directiva sospecha que la causa raíz es la *invisibilidad operativa*: sus tres sistemas principales (ERP de Inventarios, Logística y Feedback) no hablan el mismo idioma. Usted ha sido contratado como **Consultor Senior** para realizar una curaduría profunda y diseñar un Sistema de Soporte a la Decisión (DSS) en un Dashboard que transforme este caos en una estrategia de recuperación rentable.

1 El Ecosistema de Datos (The Raw Reality)

El consultor recibirá tres datasets en formato CSV, los cuales han sido deliberadamente extraídos con fallos de sistema para evaluar su capacidad de juicio técnico:

1. **inventario_central_v2.csv (2,500 registros)**: Contiene el maestro de productos. Presenta inconsistencias de tipo (fechas y lead times mezclados), costos atípicos (desde \$0.01 hasta \$850k) y existencias negativas que desafían la lógica contable.
2. **transacciones_logistica_v2.csv (10,000 registros)**: El histórico de ventas. Su principal desafío es la **integridad referencial**; miles de ventas corresponden a SKUs que no figuran en el inventario oficial. Además, los formatos de fecha varían y los tiempos de entrega presentan *outliers* de hasta 999 días.
3. **feedback_clientes_v2.csv (4,500 registros)**: La voz del cliente. Incluye registros duplicados intencionales, edades imposibles (195 años) y una escala de satisfacción (NPS) que requiere normalización para ser interpretada.

2 Pipeline de Trabajo: Del Caos al Storytelling

2.1 Fase 1: Auditoría de Calidad y Transparencia

Un consultor senior no limpia datos sin dejar rastro. Su aplicación debe calcular y visualizar un **Health Score** por cada dataset antes y después del procesamiento.

- **Métricas de Calidad:** Reportar porcentaje de nulidad por columna, número de registros duplicados eliminados y la magnitud de los *outliers* detectados.
- **Decisión Ética:** ¿Qué registros decidió eliminar y cuáles decidió imputar? Justifique el uso de media, mediana o moda basado en la distribución de los datos.

2.2 Fase 2: Integración y Feature Engineering

Debe realizar una unión estratégica (*Merge/Join*) para crear una Sola Fuente de Verdad.

- **Dilema del SKU Fantasma:** Debe decidir cómo tratar las ventas sin SKU en inventario. ¿Son ventas de productos nuevos no catalogados o errores de digitación? Su decisión afectará el cálculo del margen total.
- **Variables Derivadas:** Cree al menos tres métricas nuevas (ej. *Margen de Utilidad*, *Brecha de Entrega vs Prometido*, *Ratio de Soporte por Categoría*).

2.3 Fase 3: Inteligencia Artificial con Groq

Integre un módulo de IA que, mediante el modelo **Llama-3**, analice el resumen estadístico de los datos filtrados por el usuario y genere tres párrafos de recomendación estratégica en tiempo real.

3 El Reto Analítico: 5 Preguntas de Alta Gerencia

Su Dashboard y el documento de hallazgos deben responder con evidencia visual y estadística:

Interrogantes Estratégicos Obligatorios

1. **Fuga de Capital y Rentabilidad:** Localice los SKUs que se están vendiendo con margen negativo. ¿Representan una pérdida aceptable por volumen o es una falla crítica de precios en el canal Online?
2. **Crisis Logística y Cuellos de Botella:** ¿En qué ciudades y bodegas la correlación entre Tiempo de Entrega y NPS bajo es más fuerte? Identifique la zona que requiere un cambio inmediato de operador.
3. **Análisis de la Venta Invisible:** Cuantifique el impacto financiero (en USD) de las ventas cuyos SKUs no están en el maestro de inventario. ¿Qué porcentaje del ingreso total está en riesgo por falta de control de inventario?
4. **Diagnóstico de Fidelidad:** ¿Existen categorías de productos con alta disponibilidad (stock alto) pero con un sentimiento de cliente negativo? Explique la paradoja: ¿Es mala calidad de producto o sobrecosto?
5. **Storytelling de Riesgo Operativo:** Visualice la relación entre la antigüedad de la *Última Revisión* del stock y la tasa de *Tickets de Soporte*. ¿Qué bodegas están operando a ciegas y cómo impacta esto en la satisfacción final?

4 Entregables y Normas de Calidad

4.1 Repositorio y Dashboard (Fecha: 31 de Enero de 2026)

- **Streamlit:** Debe ser intuitivo, usar `st.sidebar`, `st.tabs` y permitir al usuario descargar el reporte de limpieza generado.
- **Documento de Hallazgos:** Un PDF (dentro del repo) que actúe como informe de consultoría. Debe contener gráficas extraídas de la app y una narrativa dirigida a la junta directiva.

- **Código:** Uso de buenas prácticas (PEP8), manejo de excepciones y un README.md que explique cómo un usuario externo puede replicar el análisis.

Criterio de Evaluación	Peso en la Nota (Total 25 %)
Calidad del Preproceso y Transparencia	30 % (Rigor técnico y limpieza)
Análisis Cruzado y EDA Dinámico	40 % (Resolución de las 5 preguntas)
Integración de IA y Storytelling	30 % (Capacidad de recomendación y UX)

Los datos limpios son útiles, pero los datos sucios bien interpretados salvan empresas.