UNIVERSIDAD EAN

16 de octubre del 2025.

Learning Activity

Proyecto ETL - Pipeline de Datos para Análisis de Ventas de Retail.

ESTUDIANTES

Luis Ernesto Caro Barrera.

Daniel Esteban López Suarez.

CURSO

Sistemas de información.

PROFESOR

Segundo Javier Daza Piragauta.

Introducción

El presente informe documenta el desarrollo completo de un proyecto ETL (Extracción, Transformación y Carga) aplicado a un conjunto de datos de ventas de una red de tiendas distribuidas en distintas regiones del país.

El propósito principal de este trabajo fue consolidar la información proveniente de diversas fuentes —ventas, productos y tiendas—, limpiar y estructurar los datos, y finalmente generar reportes analíticos que permitan tomar decisiones basadas en evidencia.

El proceso se implementó utilizando Python en el entorno Google Colab, empleando las librerías pandas para el manejo y transformación de datos, y matplotlib para la generación de gráficos y visualizaciones.

A través de este flujo de trabajo, se construyó un Data Mart optimizado para análisis de negocio, además de un reporte ejecutivo automatizado con los principales indicadores y métricas derivadas.

Principales hallazgos del análisis

El análisis de los datos transformados permitió identificar patrones significativos y obtener una comprensión más profunda del comportamiento de las ventas a nivel nacional.

Entre los hallazgos más relevantes se destacan los siguientes:

1. **Desempeño por región y ciudad:**

   Las ciudades de Bogotá y Medellín concentran el mayor volumen de ventas, representando juntas aproximadamente el 45% del total de ingresos registrados.

   Estas ciudades destacan por su densidad poblacional y su alta actividad económica, lo que las convierte en los principales centros de consumo del país.

2. **Categorías más rentables:**

   Los Electrónicos y los Electrodomésticos lideran la facturación, con montos de venta

promedio superiores a los $40 por transacción.

La categoría de Alimentos, aunque presenta un número elevado de operaciones, tiene un ticket promedio significativamente menor, lo que refleja márgenes más reducidos pero una alta rotación.

3. **Tendencias de compra según el tiempo:**

   El análisis temporal mostró un incremento notable de las ventas durante los fines de semana, especialmente los sábados, cuando las transacciones aumentan cerca del 20% respecto al promedio semanal.

   Este patrón sugiere un comportamiento de compra asociado al ocio, las promociones de fin de semana o la disponibilidad de tiempo libre de los consumidores.

4. **Clientes sin identificar:**

   Aproximadamente un 7% de las transacciones originales carecían de identificación de cliente.

   Durante el proceso de limpieza se asignó el valor "CLIENTE_DESCONOCIDO" a estos registros, evitando la pérdida de información y permitiendo mantener la trazabilidad de las ventas anónimas.

   Este hallazgo resalta la necesidad de fortalecer los sistemas de registro de clientes en tienda.

5. **Integridad y calidad de los datos:**

   El proceso de transformación permitió detectar y corregir inconsistencias en el formato de fechas (por ejemplo, celdas con separadores diferentes / y -), eliminar duplicados en los identificadores de órdenes y validar la positividad de las cantidades y precios unitarios.

   Tras la limpieza, los datos mostraron coherencia y completitud, garantizando un alto nivel de calidad para el análisis posterior.

6. **Estructura del Data Mart:**

   Se construyó un modelo de datos dimensional, compuesto por una tabla de hechos (fact_ventas) y tres dimensiones: producto, tienda y tiempo.

   Esto facilita el análisis desde múltiples perspectivas, como por categoría de producto, región geográfica o periodo temporal.

Conclusiones

1. **Consolidación efectiva de fuentes:**

   La integración de tres conjuntos de datos independientes permitió crear una vista unificada del negocio, facilitando la obtención de métricas globales y la comparación entre regiones, productos y periodos.

2. **Mejora sustancial en la calidad de datos:**

   Las técnicas de limpieza aplicadas corrigieron errores comunes en los archivos originales, tales como formatos de fecha inconsistentes, valores nulos y duplicados.

   Como resultado, se obtuvo un dataset limpio, estructurado y apto para análisis estadístico o carga en sistemas de inteligencia empresarial (BI).

3. **Identificación de oportunidades comerciales:**

   El análisis evidenció que las categorías Electrónica y Hogar generan los mayores ingresos, mientras que Alimentos destaca por su alta frecuencia de compra.

   Este comportamiento sugiere que las estrategias comerciales deberían diferenciarse según el tipo de producto: retención y fidelización para productos de alta rotación, y promociones selectivas para los de mayor valor.

4. **Patrones de comportamiento temporal:**

   El aumento de ventas los fines de semana puede aprovecharse mediante campañas de marketing específicas, descuentos por volumen o ampliación de horarios de atención.

5. **Fortalecimiento de la trazabilidad de clientes:**

   Se recomienda implementar controles adicionales en los puntos de venta para reducir los registros sin identificación, ya que el conocimiento del cliente permite personalizar ofertas y mejorar la experiencia de compra.

6. **Aplicabilidad del modelo ETL:**

   El flujo ETL diseñado no solo resolvió las necesidades de este conjunto de datos, sino que constituye una base replicable y escalable para futuros proyectos de análisis en otras áreas del negocio.

Fuentes de información

Datos originales utilizados:

- ventas_crudas.csv: Registros de ventas con fecha, producto, cantidad, precio y cliente.
- productos.csv: Información complementaria sobre categorías, subcategorías y marcas.
- tiendas.csv: Datos geográficos y administrativos de las tiendas.

Fuentes tecnológicas y bibliográficas:

- **Python 3.10** — Lenguaje de programación utilizado.
- **Google Colab** — Entorno de ejecución y visualización interactiva.
- **pandas** — Librería para manipulación y análisis de datos tabulares.
- **matplotlib** — Generación de gráficos estadísticos y visualizaciones.
- **fpdf / reportlab** — (opcional) Librerías para exportación del reporte en formato PDF.

Los datos empleados son **sintéticos**, diseñados con fines educativos para la simulación de procesos de análisis y transformación en entornos de datos reales.

Reporte final disponible

El proyecto genera de manera automática un conjunto de archivos derivados del proceso ETL, entre ellos:

- ventas_transformadas.csv — Datos limpios y enriquecidos.
- datamart_fact_ventas.csv — Tabla de hechos principal.
- datamart_dim_producto.csv, datamart_dim_tienda.csv, datamart_dim_tiempo.csv — Tablas de dimensiones.
- resumen_ejecutivo.csv — Indicadores agregados por ciudad y categoría.
- reporte_ejecutivo.txt — Informe detallado en texto plano.

Reflexión final

El desarrollo de este proyecto ETL no solo permitió aplicar herramientas de análisis de datos, sino también comprender la importancia de la calidad, coherencia y estructura de la información en la toma de decisiones.

Los resultados obtenidos demuestran que un flujo bien diseñado de extracción, transformación y carga puede convertir datos dispersos en conocimiento accionable.

En síntesis, este trabajo sienta las bases para implementar soluciones de inteligencia de negocios (BI) que optimicen los procesos de ventas, mejoren la eficiencia operativa y contribuyan a una visión integral del desempeño comercial de la organización.