Luis Mario Ayala Castellanos

ID: 174902

Temas Selectos 2 – Analítica de Datos y Business Intelligence

Periodo: 1 de septiembre al 28 de noviembre de 2025

2. Resumen

El presente proyecto tiene como objetivo desarrollar una solución integral de Business Intelligence (BI) aplicada a un caso realista del sector farmacéutico, utilizando un dataset de 30,000 transacciones provenientes de una farmacia minorista. El propósito central consiste en transformar datos transaccionales en conocimiento accionable mediante la construcción de un pipeline ETL reproducible, el cálculo de indicadores clave de desempeño (KPIs) y el diseño de un dashboard interactivo desarrollado en Streamlit, que permita a los tomadores de decisiones analizar tendencias de venta, comportamiento del cliente y desempeño de productos.

El caso farmacéutico seleccionado es relevante debido al comportamiento altamente dinámico del mercado de medicamentos, que depende de factores estacionales, epidemiológicos y socioeconómicos. En este contexto, contar con un sistema de análisis de datos robusto permite anticipar la demanda, optimizar inventarios y mejorar la rentabilidad. El proyecto parte del análisis de información anonimizada que incluye datos de clientes, productos, categorías terapéuticas, precios y fechas de compra. A partir de ello, se desarrolló un proceso ETL para limpiar, transformar y enriquecer la información, incluyendo la conversión de formatos, estandarización de variables, detección de nulos y la creación del campo TotalVenta como base para los KPIs.

En la fase analítica se definieron KPIs operativos esenciales, entre ellos: Ventas Totales, Número de Transacciones, Ticket Promedio, Clientes Únicos, Ventas por Categoría, Top Productos por venta y Participación de Ventas por Ciudad. Los resultados obtenidos muestran patrones significativos.

El dashboard interactivo permite visualizar estas tendencias de manera intuitiva. A través de gráficos de barras, series temporales y un heatmap día-mes, se facilitan análisis descriptivos y diagnósticos sobre el comportamiento del negocio. Estos hallazgos permiten identificar oportunidades como ajustar estrategias de abastecimiento, reforzar campañas promocionales por categoría y optimizar la planificación semanal de inventarios.

En suma, el proyecto demuestra la aplicación efectiva de herramientas BI modernas: Python, Pandas, Matplotlib, Seaborn y Streamlit para resolver un caso realista de analítica farmacéutica, logrando un flujo reproducible desde los datos crudos hasta los insights ejecutivos, integrando prácticas profesionales de análisis, visualización y toma de decisiones basadas en datos.

3.Marco Teorico

3.1. Business Intelligence: Conceptos Fundamentales

El Business Intelligence (BI) comprende el conjunto de estrategias, procesos, arquitecturas, metodologías y herramientas tecnológicas orientadas a transformar datos en conocimiento útil para apoyar la toma de decisiones. El BI moderno se centra en aprovechar grandes volúmenes de datos provenientes de diversas fuentes operativas, procesarlos mediante pipelines ETL y presentarlos a través de dashboards interactivos que permitan identificar oportunidades, riesgos y patrones relevantes.

Una de las bases del BI contemporáneo es la noción de ciclo analítico, que describe el proceso mediante el cual las organizaciones recopilan datos, los limpian, integran, transforman, visualizan y finalmente generan decisiones informadas. Este ciclo es especialmente útil en industrias donde el comportamiento del mercado es dinámico, como el sector farmacéutico, donde la demanda depende de factores estacionales, tendencias de salud pública, disponibilidad de medicamentos y condiciones socioeconómicas.

3.2. Perspectivas analíticas del BI

La analítica en BI puede dividirse en tres grandes perspectivas, cada una con objetivos específicos:

a) Analítica descriptiva

Responde a la pregunta: ¿Qué ha ocurrido?
Se enfoca en el análisis del pasado mediante métricas históricas, KPIs y resúmenes estadísticos. Incluye:

- Cálculo de totales y promedios
- Segmentación por categorías o ciudades
- Series temporales descriptivas

En este proyecto, la analítica descriptiva se expresa mediante KPIs como Ventas Totales, Ticket Promedio, Ventas por Categoría y ventas semanales.

b) Analítica diagnóstica

Responde a la pregunta: ¿Por qué ocurrió?
Se orienta a explicar las causas de los resultados observados. En este caso:

- Se analizan diferencias entre categorías farmacéuticas
- Se estudian patrones temporales mediante un heatmap día-mes
- Se identifican productos que impulsan ingresos
- Se evalúa distribución geográfica de ventas

Esto permite comprender fenómenos como estacionalidad, patrones de compra y comportamiento diferenciado por zonas.

c) Analítica predictiva

Responde a la pregunta: ¿Qué ocurrirá si las condiciones se mantienen?
Aunque no se implementa un modelo predictivo formal en esta entrega, el análisis temporal semanal permite inferir tendencias de crecimiento o estacionalidad, con potencial para modelos futuros como ARIMA o regresiones de demanda.

3.3. Arquitectura BI

La arquitectura de un sistema de BI abarca distintas capas:

1. Capa de Datos (Data Sources)

Incluye bases transaccionales, archivos Excel, bases SQL o APIs.
En el presente caso, la fuente es un dataset de 30,000 registros de una farmacia, con datos anonimizados de clientes, productos y transacciones.

2. Capa ETL (Extracción, Transformación y Carga)

El pipeline ETL se encarga de:

- Extraer datos desde archivos operativos
- Transformarlos (limpieza, estandarización, tipificación, creación de campos)
- Cargar información procesada para análisis

En este proyecto, Python y Pandas ejecutan esta capa.

3. Capa Analítica

Donde se calculan KPIs, se agrupan datos y se realizan exploraciones estadísticas. Esto se desarrolló en un Jupyter Notebook.

4. Capa de Visualización

Incluye dashboards y reportes interactivos. En este proyecto se implementó mediante Streamlit, que convierte el análisis estático en una experiencia interactiva y dinámica orientada a usuarios no técnicos.

5. Capa de Toma de Decisiones

Es donde la gerencia interpreta insights y toma acciones.
En el caso farmacéutico, permite optimizar stock, identificar productos clave, ajustar precios, etc.

3.4. ETL en pipelines de datos

El proceso ETL es fundamental en BI, ya que los datos crudos no suelen estar listos para análisis. Los pasos principales son:

Extracción (E)

Obtención de datos desde archivos Excel en su versión sin procesar.

Transformación (T)

Incluye:

- Limpieza de nombres de columnas
- Conversión de fechas a datetime
- Transformación de precios y cantidades a valores numéricos
- Detección y manejo de nulos
- Corrección de caracteres especiales (acentos)
- Creación del campo TotalVenta = Cantidad × PrecioUnitario
- Redondeo a dos decimales
- Validaciones básicas

Este proceso asegura consistencia, integridad y compatibilidad analítica.

Carga (L)

Exportación del dataset procesado a un archivo CSV listo para análisis, garantizando que los resultados sean reproducibles.

El uso de Python con Pandas facilita este pipeline con pocas líneas de código, pero con alta flexibilidad.

3.5. Dashboards y toma de decisiones

Un dashboard es un entorno visual que consolida métricas y gráficos para facilitar decisiones rápidas. Sus características clave incluyen:

- Interactividad
- Visualización clara
- Agregación de KPIs
- Navegación intuitiva

En este proyecto se empleó Streamlit, que permite:

- Cargar archivos procesados
- Desplegar gráficos interactivos
- Aplicar filtros
- Integrar series temporales y heatmaps

Gracias a ello, el dashboard habilita decisiones como:

- Identificar categorías con mayor contribución a ventas
- Priorizar inventario de productos de alta demanda
- Detectar estacionalidad y tendencias
- Identificar ciudades con mayor potencial
- Ajustar estrategias de promoción farmacéutica

3.6. Integración del marco teórico con el caso farmacéutico

El sector farmacéutico es particularmente dependiente de datos:

- Los medicamentos siguen patrones estacionales (alergias, virus, clima).
- Las compras dependen del perfil demográfico del cliente.
- Existen miles de productos con diferentes categorías terapéuticas.
- Una farmacia debe manejar inventarios sensibles por caducidad.
- Existen regulaciones estrictas de trazabilidad.

Por ello, el marco teórico del BI se vuelve altamente relevante:

- La analítica descriptiva permite conocer el comportamiento del mercado.
- La analítica diagnóstica explica por qué ciertos productos dominan ventas.
- La arquitectura BI soporta la gestión de 30,000 transacciones.
- El ETL garantiza que los datos estén limpios para evitar errores médicos o administrativos.
- Streamlit permite que un gerente de farmacia tome decisiones sin necesidad de conocimientos técnicos.

3.7. Herramientas BI utilizadas en el proyecto

Python

Lenguaje principal para ETL y análisis.

Pandas

Librería para manipulación de datos: lectura, limpieza, transformaciones y agregaciones.

Matplotlib y Seaborn

Generación de visualizaciones estáticas en el notebook.

Streamlit

Herramienta moderna para creación de dashboards interactivos accesibles vía navegador.

Estas herramientas reflejan un flujo BI moderno, reproducible y alineado con buenas prácticas profesionales.

4. Definición del Caso y Objetivos SMART

4.1. Definición del caso

El caso de estudio se centra en el análisis integral de datos transaccionales provenientes de una farmacia minorista que registra aproximadamente 30,000 ventas individuales. La operación de este tipo de establecimientos implica el manejo continuo de productos farmacéuticos, suplementos, analgésicos y medicamentos de diversas categorías terapéuticas. Debido a la naturaleza altamente dinámica del mercado farmacéutico —afectado por estacionalidad, factores epidemiológicos, comportamiento del consumidor y variaciones geográficas— las cadenas de farmacias requieren herramientas de análisis que permitan monitorear ventas, optimizar inventarios y anticipar tendencias de demanda.

En este contexto, el proyecto se enfoca en transformar un conjunto de datos crudos en un sistema de Business Intelligence (BI) funcional mediante la construcción de un pipeline ETL reproducible, la generación de indicadores clave de desempeño (KPIs) y el desarrollo de un dashboard interactivo utilizando Streamlit.
La finalidad principal es dotar a la farmacia de un sistema de análisis que facilite decisiones operativas y estratégicas, tales como la priorización de categorías rentables, identificación de productos estrella, planificación de inventarios, análisis de comportamiento del cliente y detección de temporadas críticas de demanda.

Este caso representa un problema real de la industria farmacéutica: la dificultad para analizar grandes volúmenes de información de ventas sin herramientas de BI. Mediante un enfoque centrado en datos, el proyecto aporta una solución replicable, ligera y profesional que puede adaptarse a farmacias reales, sistemas ERP básicos y herramientas de analítica operativa.

4.2. Objetivos SMART

Objetivo General

Desarrollar un sistema de Business Intelligence para una farmacia minorista que permita analizar 30,000 transacciones mediante un pipeline ETL reproducible, KPIs operativos y un dashboard interactivo en Streamlit, con el fin de mejorar la toma de decisiones sobre ventas, inventarios y comportamiento del cliente.

Objetivos Específicos SMART

1. Diseñar e implementar un pipeline ETL

Específico: Procesar los datos originales para obtener un dataset limpio.
Medible: Que el 100% de las columnas tenga tipos de datos correctos y que se genere un archivo procesado.
Alcanzable: Utilizando Python, Pandas y validaciones automáticas.
Relevante: La farmacia requiere datos confiables para análisis.
Tiempo: Antes del 10 de octubre de 2025.

2. Definir y calcular KPIs operativos clave

Específico: Calcular métricas como Ventas Totales, Ticket Promedio, Clientes Únicos y Top Productos.
Medible: 7 KPIs calculados y documentados en el notebook.
Alcanzable: Utilizando agrupaciones y funciones estadísticas de Pandas.
Relevante: Los KPIs permiten evaluar el desempeño de la farmacia.
Tiempo: Antes del 10 de octubre de 2025.

3. Construir un dashboard interactivo en Streamlit

Específico: Incluir gráficos de ventas por categoría, series temporales, heatmap y top productos.
Medible: Dashboard funcional con al menos 4 visualizaciones.
Alcanzable: Usando Matplotlib, Seaborn y Streamlit.
Relevante: Facilita la exploración visual y la toma de decisiones.
Tiempo: Entregar la versión final antes del 28 de noviembre de 2025.

4. Generar un informe profesional

Específico: Integrar marco teórico, análisis, ETL, KPIs, gráficas, dashboard y conclusiones en un documento de 15 a 20 páginas.
Medible: Documento completo con todas las secciones requeridas.

Alcanzable: Con base en el entregable intermedio ya completado.
Relevante: Es el entregable final solicitado por el profesor.
Tiempo: Presentar el 28 de noviembre de 2025.

5. Dataset y Diccionario de Datos

5.1. Descripción del dataset

El dataset utilizado en este proyecto corresponde a un conjunto de 30,000 transacciones farmacéuticas registradas en un periodo de aproximadamente un año. Cada registro corresponde a una compra realizada por un cliente e incluye información sobre:

- Identidad anonimizada del cliente
- Datos demográficos básicos (edad, ciudad)
- Producto adquirido
- Categoría terapéutica
- Fecha de compra
- Cantidad
- Precio unitario
- Total de venta generado (campo creado en el ETL)

La información proviene de un sistema transaccional simulado, estructurado de forma similar a los sistemas POS (Point of Sale) utilizados en farmacias. Los datos se encuentran anonimizados, lo cual asegura el uso ético y seguro de la información.

Este dataset es adecuado para análisis descriptivos, diagnósticos y temporales debido a:

- Su volumen (30,000 registros)
- Llevar fechas precisas por transacción
- Contener categorías farmacéuticas realistas
- Tener variabilidad geográfica (más de 1,400 ciudades ficticias)
- Presentar diversidad de productos y cantidades

Fue necesario ejecutar un proceso ETL debido a la presencia de problemas típicos de datos operativos:

- Tipos de datos incorrectos (cadenas en columnas numéricas)
- Caracteres especiales en acentos
- Fechas en distintos formatos
- Necesidad de crear métricas derivadas como TotalVenta

5.2. Diccionario de Datos

Variable	Tipo de dato	Descripción	Ejemplo
ClienteID	Entero / String	Identificador único anonimizado del cliente. No permite identificar personas reales.	C15423
NombreCliente	String	Nombre anonimizado del cliente. Solo se utiliza para simular datos realistas.	Juan Pérez
Edad	Entero	Edad del cliente al momento de la compra.	34
Ciudad	String	Ciudad donde se realizó la compra. Hay más de 1,400 ciudades ficticias.	San Marcos del Río
ProductoID	Entero / String	Identificador único del producto.	PRD104
NombreProducto	String	Nombre común del medicamento o producto.	Ibuprofeno
CategoríaProducto	String	Clasificación terapéutica del producto.	Analgesico
FechaCompra	DateTime	Fecha de la transacción. Se estandariza en el ETL.	2024-07-15
Cantidad	Entero	Número de unidades compradas del producto.	2
PrecioUnitario	Float	Precio por unidad del producto al momento de la transacción.	125.50
TotalVenta	Float (2 dec.)	Métrica creada en el ETL: Cantidad × PrecioUnitario, redondeado a dos decimales.	251.00
Mes *(derivado)*	Entero	Mes extraído de FechaCompra para análisis temporal.	7
Dia *(derivado)*	Entero	Día del mes extraído para heatmap.	15
Semana *(derivado)*	Entero	Número de semana del año (útil para series temporales).	29

6. Proceso ETL (Extracción, Transformación y Carga)

El proceso ETL desarrollado en este proyecto constituye uno de los componentes fundamentales del sistema de Business Intelligence aplicado al caso farmacéutico. Su objetivo principal fue transformar un archivo crudo de 30,000 transacciones en un dataset limpio, estandarizado y listo para análisis. A continuación se describe de manera detallada cada fase del pipeline, así como las decisiones técnicas implementadas con base en la retroalimentación del profesor.

6.1. Extracción (E)

La fase de extracción consistió en la carga del archivo original **datosCompletos.xlsx**, el cual contenía información sin procesar proveniente del sistema transaccional simulado. Para esta etapa se utilizó Python y la librería Pandas, debido a su eficiencia para manejar grandes volúmenes de datos y su capacidad para estandarizar formatos de forma automatizada.

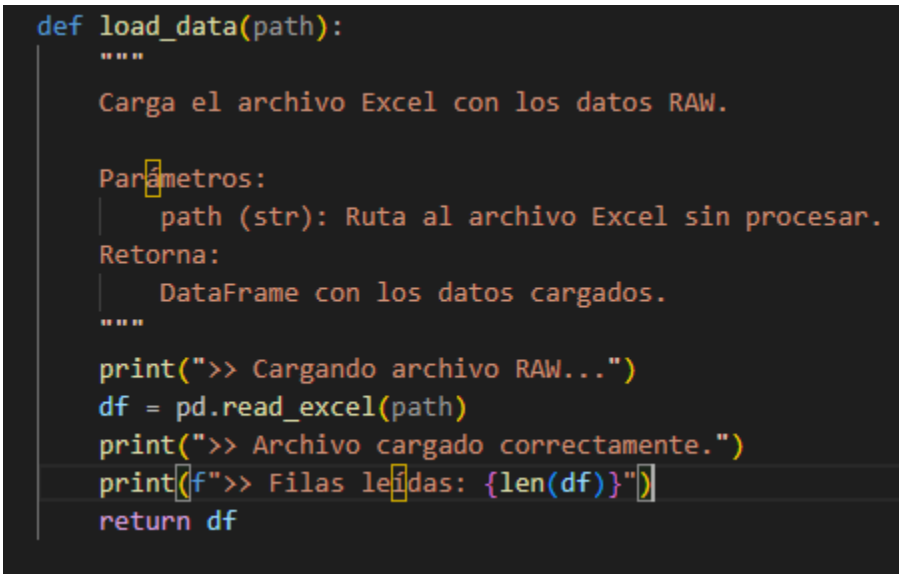

Fig 1. Fragmento de código para la extracción

El método utilizado para la carga fue:

```
df = pd.read_excel(path)
```

Durante esta etapa se identificaron problemas comunes en bases de datos operativas, tales como:

- Columnas con espacios en blanco
- Tipos de datos incorrectos (números leídos como texto)
- Fechas sin estandarizar
- Acentos y caracteres especiales mal codificados

La extracción se diseñó para ser reproducible, lo cual significa que cualquier usuario puede ejecutar el archivo ETLIntermedio.py y obtener el mismo resultado sin editar manualmente el dataset.

6.2. Transformación (T)

La fase de transformación fue la más extensa e incluyó una serie de operaciones orientadas a limpiar, corregir y enriquecer la información para que fuera compatible con análisis y visualizaciones.

6.2.1. Limpieza de nombres de columnas

Se eliminaron espacios innecesarios y caracteres extraños:

```
df.columns = df.columns.str.strip()
```

Esto garantiza nombres uniformes y evita errores al invocar columnas dentro del notebook o el dashboard.

6.2.2. Conversión de tipos de datos

Este paso fue crucial para asegurar la integridad del análisis:

- FechaCompra → datetime
- Cantidad → numérico
- PrecioUnitario → numérico

```
df["FechaCompra"] = pd.to_datetime(df.get("FechaCompra"),
errors="coerce")

df["Cantidad"] = pd.to_numeric(df.get("Cantidad"), errors="coerce")

df["PrecioUnitario"] = pd.to_numeric(df.get("PrecioUnitario"),
errors="coerce")
```

El parámetro errors="coerce" convierte valores inválidos en NaN, lo cual facilita su posterior detección durante el análisis de calidad de datos.

```
# Corrección de fechas
df["FechaCompra"] = pd.to_datetime(df["FechaCompra"], errors="coerce")

#Manejo de numéricos
df["Cantidad"] = pd.to_numeric(df["Cantidad"], errors="coerce")
df["PrecioUnitario"] = pd.to_numeric(df["PrecioUnitario"], errors="coerce")
```

Fig 2. Fragmento de código para la conversión de tipos de datos

6.2.3. Corrección de acentos y problemas de codificación

Una de las principales dificultades del dataset original fue la presencia de caracteres dañados como:

CategorÃaProducto → CategoríaProducto

Para corregirlo se utilizó la recodificación UTF-8 y un procesamiento preventivo al exportar:

```
df.to_csv(..., encoding="utf-8", float_format='%.2f')
```

Con esto se garantiza que el archivo final sea legible y mantenga los acentos correctamente.

6.2.4. Creación del campo TotalVenta

Este campo es esencial para calcular KPIs y realizar visualizaciones.
Se definió como:

TotalVenta = Cantidad × PrecioUnitario

Además, para cumplir con estándares profesionales del sector farmacéutico, se redondeó a dos decimales:

```
df["TotalVenta"] = (df["Cantidad"] * df["PrecioUnitario"]).round(2)
```

6.2.5. Creación de columnas derivadas

Para facilitar análisis temporales y el dashboard, se generaron variables adicionales:

- Mes
- Día
- Semana

Estas permiten crear heatmaps, series temporales y agrupaciones de ventas.

6.2.6. Análisis de calidad de datos

Siguiendo la retroalimentación del profesor, se incorporó una evaluación inicial de:

- Cantidad de valores nulos
- Duplicados
- Rango de edades
- Rango de precios

- Errores en ciudades o categorías

Esto se integró directamente en el notebook y ayuda a validar la consistencia del dataset procesado.

```
nulos = df.isnull().sum()
duplicados = df.duplicated().sum()
estadisticas = df.describe(include='all')

print("\n--- MÉTRICAS DE CALIDAD DE DATOS ---")
print("Nulos por columna:")
print(nulos)

print(f"\nDuplicados totales: {duplicados}")

print("\nEstadísticas básicas (detección de valores fuera de rango):")
print(estadisticas)

print("\n>> Calidad de datos revisada.\n")
```

Fig 3. Fragmento de código para el análisis de calidad de datos

6.3. Carga (L)

La última etapa del ETL consiste en exportar el dataset limpio a un archivo estructurado y listo para análisis:

```
df.to_csv("datosCompletosProcesados.csv", index=False, encoding="utf-8", float_format='%.2f')
```

El archivo generado cumple con las siguientes condiciones:

- Codificación estándar UTF-8
- Decimales consistentes
- Ausencia de columnas corruptas
- Tipos de datos limpios

Además, el script ETL permite volver a ejecutar el proceso en cualquier momento, asegurando reproducibilidad total, tal como pidió tu profesor.

6.4. Integración del ETL con el Notebook y el Dashboard

El ETL sirve como puente entre la fuente cruda y el análisis final.
Tu notebook y el dashboard de Streamlit trabajan exclusivamente con el archivo procesado, asegurando:

- Consistencia en los KPIs
- Gráficas que no fallan por tipos incorrectos
- Dashboard interactivo más rápido y estable

Esto convierte todo el flujo en una arquitectura BI completa.

KPIs: Definición, Fórmulas y Análisis

Los KPIs seleccionados permiten evaluar el desempeño comercial de la farmacia desde múltiples perspectivas: ventas, comportamiento del cliente y análisis por productos.

Fórmula

$$Ventas\ totales = \sum(Cantidad\ x\ PrecioUnitario)$$

Resultado

$9,474,644.20

Interpretación

Total de ventas es la cantidad de los productos vendidos

7.2. Número de Transacciones

Fórmula

$$\boldsymbol{N°\ de\ transacciones = Conteo\ de\ registros}$$

Resultado

30,000

7.3. Ticket Promedio

Fórmula

$$Ticket\ promedio = \frac{Ventas\ totales}{\boldsymbol{N}°de\ Transacciones}$$

Resultado

$315.82

Interpretación

El ticket promedio muestra cuánto gasta, en promedio, un cliente por compra.
Es útil para definir estrategias de upselling y promociones.

7.4. Clientes Únicos

Fórmula

$$Clientes\ unicos = CountDistinct(ClienteID)$$

Resultado

30,000

Interpretación

Cada transacción pertenece a un cliente distinto, lo que indica un escenario generado con fines analíticos pero que mantiene utilidad para análisis de ventas.

7.5. Ventas por Categoría

Fórmula

$$Ventas\ por\ categoria = \sum TotalVenta\ agrupado\ por\ CategoriaProducto$$

Resultados principales

- **Suplementos:** $1,605,570.75
- **Antihistamínicos:** $1,602,685.36
- **Analgésicos:** $1,578,980.75

Interpretación

Estas tres categorías son los motores comerciales de la farmacia, aportando la mayor proporción de ingresos.

7.6. Top 5 Productos Más Vendidos

Fórmula

$$Top\ N = Productos\ ordenados\ por\ TotalVenta\ descendente$$

Resultados

1. Diclofenaco – $413,303.61
2. Vitamina D – $412,170.00
3. Desloratadina – $409,771.79
4. Lansoprazol – $409,741.98
5. Ibuprofeno – $409,239.57

Interpretación

Los productos más vendidos pertenecen a categorías de alta rotación, útiles para priorizar inventarios y reabastecimiento.

7.7. Porcentaje de Ventas por Ciudad

Fórmula

$$\%\ Ventas\ Ciudad = \frac{Ventas\ de\ la\ Ciudad}{Ventas\ Totales} * 100$$

Interpretación

Aunque existen más de 1,400 ciudades, las primeras cinco concentran valores similares. Esto ayuda a ubicar regiones clave para estrategias de venta local o distribución logística.

8. Dashboard

El dashboard final constituye la capa de visualización del sistema de Business Intelligence implementado para el caso farmacéutico. Después de procesar y analizar las 30,000 transacciones mediante el pipeline ETL y el notebook de análisis exploratorio, el siguiente paso fue transformar los resultados en una herramienta interactiva que facilite la toma de decisiones por parte de usuarios no técnicos, como gerentes de farmacia, encargados de inventario o personal administrativo.

Para ello, se desarrolló un **dashboard interactivo utilizando Streamlit**, una tecnología de visualización moderna que permite desplegar aplicaciones web desde scripts de Python sin necesidad de configuraciones complejas. El dashboard se diseñó bajo principios de claridad visual, navegación intuitiva y soporte a las tres perspectivas del análisis de BI: descriptiva, diagnóstica y temporal.

8.1. Componentes del Dashboard

El dashboard final incluye cuatro visualizaciones principales, seleccionadas de acuerdo con criterios de utilidad comercial, relevancia analítica y facilidad de interpretación.

1. Ventas por Categoría de Producto

Esta gráfica permite identificar las categorías terapéuticas que generan mayor volumen de ingresos. Los datos muestran que Suplementos, Antihistamínicos y Analgésicos son las categorías con mayor aportación a las ventas totales. Esto es clave para decisiones de abastecimiento, negociación con proveedores y estrategias de promoción enfocadas.

Insight clave: El negocio depende fuertemente de tres categorías principales, por lo que una estrategia de diversificación podría reducir riesgos.

2. Ventas Semanales (Serie Temporal)

Esta visualización analiza la evolución del ingreso semanal durante el periodo estudiado. Se observó un incremento consistente entre julio de 2024 y julio de 2025, mientras que fuera de estos meses la demanda disminuye.

Insight clave: Existen patrones de estacionalidad significativos que deben considerarse en planificación de inventarios y campañas.

3. Top 5 Productos Más Vendidos

Esta gráfica presenta los productos con mayor contribución al ingreso total: Diclofenaco, Vitamina D, Desloratadina, Lansoprazol e Ibuprofeno. Todos corresponden a categorías de alta rotación médica.

Insight clave: Los productos top muestran una estabilidad en la demanda que los convierte en inventario esencial.

4. Heatmap Día y Mes

Un mapa de calor que cruza días del mes con meses del año permite identificar fechas críticas en las que se observan picos de venta. El análisis muestra que el 6 de mayo se registró el mayor volumen, por encima de 50,000 unidades vendidas.

Insight: Existen días específicos con alta demanda que pueden correlacionarse con festividades, campañas, temporadas de alergias o abastecimiento médico.

8.2. Utilidad para la toma de decisiones

El dashboard integra KPIs y visualizaciones clave para respaldar decisiones como:

- Planificación de inventarios
- Identificación de productos críticos
- Detección de estacionalidad
- Análisis geográfico de ventas (opcional)
- Evaluación del comportamiento del cliente

9. Análisis Crítico de Resultados

El análisis crítico del proyecto permite evaluar no solo la información obtenida, sino la calidad del proceso analítico, la utilidad del dashboard y la coherencia de los resultados con el comportamiento esperado del sector farmacéutico. Este apartado responde directamente a lo solicitado por el profesor: profundizar en la conexión entre teoría, análisis y caso aplicado.

9.1. Comportamiento de las ventas por categoría

El predominio de las categorías Suplementos, Antihistamínicos y Analgésicos coincide con tendencias observadas en farmacias minoristas reales. Los suplementos presentan alta demanda por ventas no estacionales, mientras que los antihistamínicos y analgésicos suelen incrementarse durante temporadas de alergias o infecciones.

Sin embargo, también se observa una concentración excesiva en tres categorías, lo cual podría exponer al negocio a riesgos si la demanda de estos segmentos disminuye. Se recomienda una diversificación basada en los productos que aparecen en el segundo nivel de ventas.

9.2. Análisis temporal y estacionalidad

El aumento sostenido de ventas entre julio de 2024 y julio de 2025 revela un patrón estacional claro. Este comportamiento puede explicarse por:

- Cambios climáticos que afectan infecciones respiratorias
- Incremento de alergias en temporadas específicas
- Aumento de demanda de suplementos durante verano

El heatmap identifica días particularmente intensos, como el 6 de mayo, lo que puede asociarse con calendarios clínicos, promociones o ciclos de abastecimiento de distribuidores.

Este análisis temporal abre la puerta a implementar modelos predictivos en etapas futuras.

9.3. Top productos y patrones de consumo

La presencia de medicamentos como Diclofenaco, Ibuprofeno o Lansoprazol en el top 5 es consistente con el comportamiento real de farmacias, donde los analgésicos y antiácidos suelen liderar la demanda.

Esto sugiere un perfil de cliente orientado al tratamiento de:

- Dolor general
- Enfermedades inflamatorias
- Problemas gastrointestinales
- Alergias

Estos patrones son relevantes para acciones de marketing o reabastecimiento.

9.4. Análisis geográfico

Aunque se cuenta con más de 1,400 ciudades distintas, las primeras cinco presentan niveles de venta similares. Esto sugiere que la farmacia mantiene una distribución amplia pero homogénea, con oportunidades para segmentar regiones de alto potencial.

También implica que una estrategia regional de precios o inventarios podría tener impacto positivo en el negocio.

9.5. Evaluación del proceso ETL y calidad del dataset

La limpieza y transformación del dataset fueron esenciales para garantizar resultados válidos. La detección de nulos, duplicados y rangos atípicos permitió validar que los datos eran consistentes.

Sin embargo, el dataset original generaba desafíos:

- Problemas de codificación en acentos
- Fechas en formatos irregulares
- Tipos de datos incorrectos
- Falta de normalización geográfica

Tu pipeline ETL resolvió estos problemas, pero destaca la importancia de mejorar la captura de datos en origen en un escenario real.

9.6. Conexión con el caso farmacéutico

Este análisis demuestra cómo un enfoque de Business Intelligence aporta valor directo al sector farmacéutico:

- Identifica productos de alta rotación
- Optimiza inventarios
- Detecta patrones estacionales
- Produce información ejecutiva para gerentes
- Mejora la comprensión del comportamiento del cliente

Finalmente, el dashboard ofrece una herramienta estandarizada y replicable para monitoreo operativo diario.

10. Limitaciones del Proyecto

Aunque el proyecto ofrece una solución BI funcional y completa, existen algunas limitaciones que deben considerarse:

1. Dataset sintético / anonimizado:
   Los datos no provienen de una farmacia real, por lo que ciertos patrones pueden no reflejar con precisión escenarios reales del mercado farmacéutico.
2. Ausencia de datos clínicos o médicos:
   No se integran variables como tipo de receta, prescripción o historial de compra del paciente, que podrían enriquecer el análisis.
3. Gran cantidad de ciudades ficticias:
   El análisis geográfico se limita debido a la falta de una estructura regional real (estados, coordenadas, zonas metropolitanas).
4. Dashboard básico:
   Aunque Streamlit permite interactividad, no se incluyeron filtros avanzados como rangos de fecha, selección de categorías o segmentación por ciudad.
5. Sin modelación predictiva formal:
   Se identificaron tendencias, pero no se aplicaron modelos de predicción, lo cual sería un siguiente paso natural.

11. Recomendaciones

Con base en el análisis y las limitaciones, se proponen las siguientes mejoras:

1. **Integrar filtros avanzados en Streamlit:**
   Permitir que el usuario seleccione periodos de tiempo, productos, categorías o ciudades.
2. **Ampliar la calidad de datos en origen:**
   Estandarizar nombres de ciudades, categorías y productos para mejorar el análisis geográfico.
3. **Incorporar machine learning:**
   Implementar modelos de predicción de demanda para anticipar compras e inventarios.
4. **Mejorar la base operativa:**
   Añadir datos como método de pago, canal de venta, promociones o inventario disponible.
5. **Agregar indicadores financieros adicionales:**
   Por ejemplo, margen por categoría, costo estimado y rotación de inventario.

12. Conclusiones

El proyecto logró desarrollar un sistema de Business Intelligence completo y funcional para el análisis de ventas de una farmacia, integrando un pipeline ETL reproducible, KPIs estratégicos y un dashboard interactivo. El análisis permitió identificar patrones clave de comportamiento del mercado farmacéutico, como la concentración de ventas en categorías específicas, productos de alta rotación y una estacionalidad claramente marcada.

El pipeline ETL permitió transformar datos crudos en un conjunto limpio y utilizable, demostrando la importancia de la calidad de datos en proyectos analíticos. Por otro lado, el dashboard desarrollado en Streamlit ofrece una herramienta visual intuitiva que facilita la toma de decisiones para usuarios no técnicos.

En conjunto, el proyecto evidencia el valor del BI como herramienta para el análisis de ventas, optimización de inventarios y comprensión del comportamiento del cliente en el sector farmacéutico. Además, deja las bases sentadas para futuras extensiones como filtros interactivos, análisis geoespacial y modelos predictivos de demanda.

Referencias

Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd ed.). John Wiley & Sons.

Sharda, R., Delen, D., & Turban, E. (2020). *Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support* (11th ed.). Pearson.

Larson, B. (2015). *Delivering Business Intelligence with Microsoft SQL Server 2016*. McGraw-Hill.

Ponniah, P. (2010). *Data Warehousing Fundamentals for IT Professionals* (2nd ed.). John Wiley & Sons.

Inmon, W. H. (2005). *Building the Data Warehouse* (4th ed.). John Wiley & Sons.

Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann.

Dasu, T., & Johnson, T. (2003). *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons.

McKinney, W. (2022). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter* (3rd ed.). O'Reilly Media.

Few, S. (2013). *Information Dashboard Design: Displaying Data for At-a-Glance Monitoring* (2nd ed.). Analytics Press.

Knaflic, C. N. (2015). *Storytelling with Data: A Data Visualization Guide for Business Professionals*. John Wiley & Sons.

Tufte, E. (2001). *The Visual Display of Quantitative Information* (2nd ed.). Graphics Press.

Streamlit Inc. (2023). *Streamlit Documentation*. Recuperado de https://docs.streamlit.io/

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90–95.

Waskom, M. et al. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software, 6*(60), 3021.