Manual de Usuario

RetailIQ Analytics Dashboard

Análisis de Ventas Retail · ETL · EDA · IA Generativa

Autores:

Gia Mariana Calle Higuita
José Santiago Molano Perdomo
Juan José Restrepo Higuita

Universidad EAFIT
SI6001 – Fudamentos en Ciencia de Datos

Febrero 2026
Versión 1.0

Índice

1. Introducción

RetailIQ Analytics es una plataforma web interactiva que integra el ciclo completo de la Ciencia de Datos: desde la ingesta y limpieza de datos (ETL), pasando por el análisis exploratorio visual (EDA), hasta la generación de recomendaciones estratégicas mediante Inteligencia Artificial Generativa.

La plataforma está diseñada para analizar datasets de ventas retail que contienen imperfecciones reales (valores inválidos, nulos, outliers e inconsistencias), y responde tres preguntas estratégicas de negocio:

1. **¿Qué categorías de productos generan mayor ingreso y cuáles dependen más de los descuentos para vender?** — Para optimizar la estrategia de *pricing* y promociones.
2. **¿Existe estacionalidad en las ventas y cómo varía la demanda a lo largo del tiempo?** — Para planificar inventario y campañas de marketing.
3. **¿Cómo difiere el comportamiento de compra entre canales (Online vs In-store) y métodos de pago?** — Para optimizar la estrategia omnicanal.

2. Requisitos del sistema

2.1. Requisitos de software y hardware

Componente	Requisito mínimo
Python	Versión 3.10 o superior
Memoria RAM	4 GB (8 GB recomendado para datasets grandes)
Navegador web	Google Chrome, Mozilla Firefox o Microsoft Edge (actualizado)
Conexión a Internet	Necesaria para la integración con la API de Groq
API Key de Groq	Cuenta gratuita en console.groq.com

Cuadro 1: Requisitos del sistema

2.2. Dependencias de Python

Las siguientes librerías se instalan automáticamente con el archivo `requirements.txt`:

- `streamlit` $\geq$ 1.32.0 — Framework de la aplicación web.
- `pandas` $\geq$ 2.0.0 — Manipulación y análisis de datos.
- `numpy` $\geq$ 1.24.0 — Cómputo numérico.
- `plotly` $\geq$ 5.18.0 — Visualizaciones interactivas.
- `groq` $\geq$ 0.4.0 — Cliente para la API de Groq (LLM).

- `statsmodels` $\geq 0.14.0$ — Líneas de tendencia en gráficos.

3. Instalación y ejecución

3.1. Ejecución local

Siga estos pasos para ejecutar la aplicación en su máquina:

1. **Clone el repositorio:**

```
git clone https://github.com/restreh/
    SI6001-DSc-FinalProject-Retail-IQ-Analytics.git
cd SI6001-DSc-FinalProject-Retail-IQ-Analytics
```

2. **Cree un entorno virtual:**

```
python -m venv venv
```

3. **Active el entorno virtual:**
   - En Linux/Mac: `source venv/bin/activate`
   - En Windows: `venv\Scripts\activate`
4. **Instale las dependencias:**

```
pip install -r requirements.txt
```

5. **Ejecute la aplicación:**

```
streamlit run app.py
```

6. **Acceda desde su navegador:** La aplicación se abrirá automáticamente en `http://localhost:8501`. Si no se abre, copie esa dirección en su navegador.

3.2. Acceso en la nube (Streamlit Cloud)

La aplicación también se encuentra desplegada en Streamlit Cloud. Simplemente acceda a la URL proporcionada en el archivo `README.md` del repositorio. No requiere instalación alguna.

4. Módulo 1: Ingesta y Procesamiento (ETL)

Este módulo permite cargar datos, diagnosticar su calidad, aplicar limpieza interactiva y generar variables derivadas.

4.1. Carga de datos

El sistema soporta tres fuentes de datos:

- **Archivo CSV o JSON:** Utilice el botón «Browse files» para seleccionar un archivo desde su computadora.
- **URL directa:** Ingrese la URL pública de un archivo CSV o JSON.

Nota: El tamaño máximo de archivo admitido es de 500 MB. Para datasets más grandes, se recomienda utilizar una URL directa.

4.2. Auditoría de calidad (datos crudos)

Una vez cargados los datos, el sistema calcula automáticamente el **Health Score**, una métrica ponderada que evalúa tres dimensiones:

Dimensión	Peso	Descripción
Completitud	40 %	Porcentaje de celdas no nulas
Unicidad	30 %	Porcentaje de registros no duplicados
Validez	30 %	Porcentaje de valores válidos (sin ERROR/UNKNOWN)

Cuadro 2: Componentes del Health Score

Además, se muestra la cantidad total de nulos, duplicados y un heatmap de nulidad por columna.

4.3. Limpieza interactiva

El usuario puede activar o desactivar cada paso de limpieza mediante controles interactivos:

- **Eliminar duplicados:** Activa un checkbox para remover filas idénticas. El sistema informa cuántas filas fueron eliminadas.
- **Reemplazar tokens inválidos:** Convierte valores `ERROR`, `UNKNOWN` y `NONE` a `NaN` para su posterior tratamiento.
- **Imputación numérica:** Seleccione entre *Mediana*, *Media* o *Cero* para rellenar valores nulos en columnas numéricas.

- **Tratamiento de outliers:** Aplica el método IQR con multiplicador 3. Los valores extremos se capean a los límites del rango intercuartílico.

4.4. Recálculo inteligente de valores

El sistema aprovecha la relación matemática **Total Spent** = **Quantity** × **Price Per Unit** para recuperar valores faltantes:

- Si falta `Total Spent` pero existen `Quantity` y `Price Per Unit`, se recalcula.
- Si falta `Price Per Unit` pero existen `Total Spent` y `Quantity`, se deduce.
- Si falta `Quantity` pero existen `Total Spent` y `Price Per Unit`, se recupera.

Esto permite conservar la mayor cantidad de registros posible antes de recurrir a la imputación estadística.

4.5. Feature Engineering

Tras la limpieza, el sistema genera automáticamente las siguientes variables derivadas:

Variable	Descripción
`Ticket_Promedio`	Gasto total dividido entre la cantidad comprada (`Total Spent / Quantity`)
`Mes`, `Nombre_Mes`	Mes numérico y nombre del mes, extraídos de `Transaction Date`
`Dia_Semana`	Nombre del día de la semana (Monday, Tuesday, etc.)
`Trimestre`, `Anio`	Trimestre y año de la transacción
`Es_FinDeSemana`	Booleano: `True` si la transacción ocurrió en sábado o domingo
`Rango_Gasto`	Segmentación por cuartiles: Bajo, Medio, Alto, Premium

Cuadro 3: Variables derivadas generadas automáticamente

Importante: Debe presionar el botón **«Ejecutar Limpieza y Feature Engineering»** para que el procesamiento se aplique y los datos queden disponibles para los módulos de EDA e IA.

5. Módulo 2: Visualización Dinámica (EDA)

Este módulo ofrece gráficos interactivos generados con Plotly, organizados en tres pestañas. Todos los gráficos permiten zoom, hover con información detallada y descarga como imagen PNG.

5.1. Pestaña: Análisis Univariado

Permite explorar la distribución de variables individuales:

- **Variables numéricas:** Seleccione una variable y elija entre Histograma o Boxplot. Opcionalmente, puede colorear por una variable categórica para comparar distribuciones entre grupos.
- **Variables categóricas:** Visualice las frecuencias de los valores más comunes mediante gráficos de barras.

5.2. Pestaña: Análisis Bivariado

Explora las relaciones entre dos o más variables:

- **Heatmap de correlación:** Seleccione múltiples columnas numéricas para visualizar la matriz de correlación de Pearson. Los valores se muestran directamente sobre el gráfico.
- **Scatter Plot:** Elija dos variables para los ejes X e Y. Se incluye automáticamente una línea de tendencia OLS. Si el dataset supera los 5 000 registros, se toma una muestra aleatoria para mantener el rendimiento.
- **Evolución temporal:** Seleccione una columna de fecha y una variable numérica para observar su evolución mensual. Puede elegir entre suma, media, conteo o mediana como función de agregación.

5.3. Pestaña: Reporte Estratégico

Contiene gráficos preconfigurados que responden directamente a las tres preguntas de negocio:

5.3.1. Pregunta 1: Ingresos por categoría e impacto de descuentos

- Gráfico de barras con los ingresos totales por categoría de producto.
- Comparativa de ingresos con descuento vs. sin descuento por categoría.
- Tabla con el porcentaje de transacciones con descuento, ticket promedio e ingreso total por categoría.

5.3.2. Pregunta 2: Estacionalidad de las ventas

- Gráfico de barras con los ingresos agrupados por mes del año.
- Gráfico de gasto promedio por día de la semana.
- Evolución trimestral de ingresos desglosada por categoría.

5.3.3. Pregunta 3: Comportamiento por canal y método de pago

- Ingresos y número de transacciones por canal de venta (Online, In-store).
- Transacciones y ticket promedio por método de pago.
- Matriz de calor: ticket promedio cruzando canal con método de pago.

6. Módulo 3: Inteligencia Artificial (Groq)

Este módulo conecta la aplicación con la API de Groq para generar análisis estratégicos automatizados mediante modelos de lenguaje de gran escala (LLM).

6.1. Configuración

1. **Ingrese su API Key de Groq.** Puede obtener una clave gratuita en console.groq.com/keys.
2. **Seleccione el modelo LLM** que desee utilizar.
3. **(Opcional)** Escriba una pregunta específica o contexto adicional en el campo de texto.
4. **Presione «Generar Recomendaciones Estratégicas».**

6.2. ¿Qué genera la IA?

El sistema toma automáticamente el resumen estadístico (`df.describe()`) del dataset filtrado, lo combina con agregados por categoría y canal, y lo envía al modelo LLM junto con el contexto de las tres preguntas de negocio. La respuesta incluye:

- **Tendencias clave:** Los tres patrones más relevantes en los datos.
- **Riesgos detectados:** Alertas tempranas y problemas potenciales.
- **Oportunidades de negocio:** Tres recomendaciones accionables basadas en evidencia.
- **Segmentación sugerida:** Propuesta de segmentos de clientes o productos.

6.3. Modelos disponibles

Recomendación: Para un análisis completo y detallado, utilice `llama-3.3-70b-versatile`. Para iteraciones rápidas durante la exploración, `llama-3.1-8b-instant` ofrece respuestas más ágiles.

Modelo	Característica	Caso de uso recomendado
`llama-3.3-70b-versatile`	Mayor capacidad de razonamiento	Análisis complejos y detallados
`llama-3.1-8b-instant`	Mayor velocidad de respuesta	Consultas rápidas y puntuales
`mixtral-8x7b-32768`	Contexto extenso (32 768 tokens)	Datasets con muchas variables

Cuadro 4: Modelos LLM disponibles en Groq

7. Filtros globales

Los filtros globales se ubican en la barra lateral izquierda y se activan automáticamente una vez que el dataset ha sido procesado en el Módulo 1 (ETL). Estos filtros afectan tanto al Módulo 2 (EDA) como al Módulo 3 (IA Insights), lo que permite analizar segmentos específicos de los datos.

- **Rango de fechas:** Seleccione una fecha de inicio y una fecha de fin para acotar el periodo de análisis.
- **Categorías:** Multiselect que permite filtrar por una o varias categorías de producto.
- **Ubicación:** Filtre por canal de venta (por ejemplo, Online o In-store).
- **Método de pago:** Filtre por método de pago (Cash, Credit Card, etc.).
- **Slider de gasto:** Establezca un rango numérico mínimo y máximo para `Total Spent`.

El contador de registros filtrados se actualiza en la barra lateral cada vez que se modifican los filtros.

8. Preguntas frecuentes

¿Puedo utilizar un dataset diferente al incluido?

Sí. La aplicación acepta cualquier archivo en formato CSV o JSON a través de la interfaz. Las variables derivadas se crearán únicamente si las columnas necesarias (`Quantity`, `Price Per Unit`, `Total Spent`, `Transaction Date`) están presentes en el dataset.

¿La API de Groq es gratuita?

Sí. Groq ofrece un nivel gratuito con límites generosos de uso. Puede registrarse en console.groq.com y generar una API Key sin costo.

¿Puedo descargar los gráficos?

Sí. Cada gráfico generado con Plotly incluye una barra de herramientas en la esquina superior derecha. El ícono de cámara permite descargar el gráfico como imagen PNG.

¿El Health Score se calcula automáticamente?

Sí. Se calcula tanto antes como después de la limpieza, lo que permite cuantificar la mejora en la calidad de los datos.

¿Qué ocurre si mi dataset tiene más de 100 000 filas?

La aplicación puede manejar datasets grandes sin problemas. Los gráficos de tipo scatter toman automáticamente una muestra de 5 000 registros para mantener el rendimiento visual. Los cálculos de ETL e IA procesan la totalidad de los datos.

9. Solución de problemas

Problema	Solución
La aplicación no inicia	Verifique que todas las dependencias estén instaladas ejecutando `pip install -r requirements.txt`. Confirme que está usando Python 3.10 o superior.
Error al subir un archivo	Asegúrese de que el archivo sea un CSV o JSON válido y que no supere los 500 MB de tamaño.
Error de conexión con Groq	Verifique que su API Key sea correcta y que su cuenta tenga créditos disponibles. Compruebe su conexión a Internet.
Los gráficos tardan en cargar	Utilice los filtros globales en la barra lateral para reducir el volumen de datos visualizados.
No se generan las variables derivadas	Las columnas del dataset deben coincidir con los nombres esperados (`Quantity`, `Price Per Unit`, `Total Spent`, `Transaction Date`).
La limpieza no elimina todos los nulos	Verifique que los checkboxes de imputación estén activados. Los nulos en la columna `Transaction Date` que no puedan parsearse como fecha permanecerán como `NaT`.

Cuadro 5: Problemas comunes y sus soluciones

Para soporte adicional o para reportar errores, contacte a los autores a través del repositorio de GitHub del proyecto.