INFORME

Docente: Jesús Ariel Bonilla

Estudiante: Paula Andrea Terrios Ossa

Corporación Universitaria Del Huila (CORHUILA)

Ing. Ambiental

Ciencia de Datos

Neiva – Huila

2025

**Introducción**

**Tema:** Ciencia de datos

**Autor:** Paula Andrea Terrios Ossa

**Año:** 2025

El presente informe desarrolla un flujo completo de análisis de datos aplicando técnicas de **ciencia de datos** sobre el conjunto Online Retail del repositorio UCI.  
El propósito principal es implementar un proceso **ETL (Extracción, Transformación y Carga)** complementado con consultas SQL, análisis temporal, métricas por cliente e integración con una API de divisas.

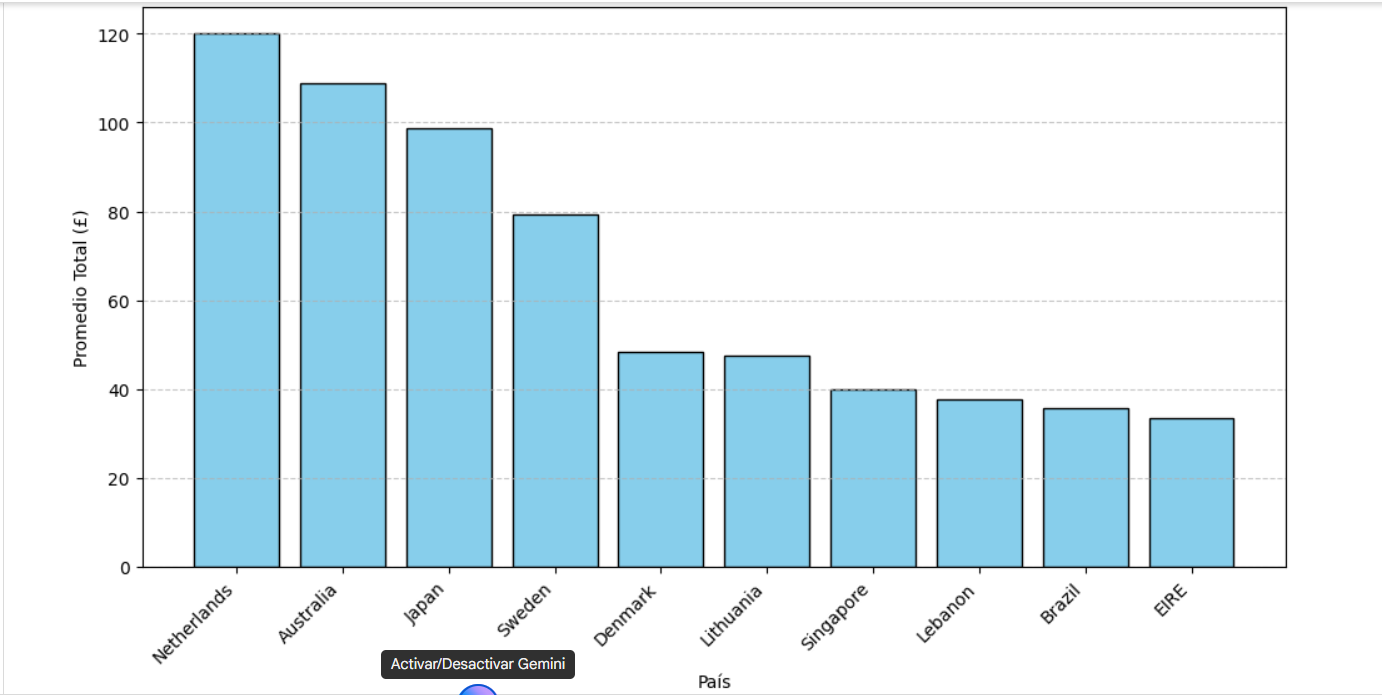
El dataset contiene transacciones de ventas en línea realizadas por una empresa minorista del **Reino Unido durante el año 2011**, y permite identificar patrones de comportamiento comercial, evolución de ventas y segmentación de clientes.

La aplicación de herramientas como **Python, Pandas, SQLite, Matplotlib, Seaborn y Requests** facilitó la manipulación, análisis y visualización de los datos para obtener información útil que respalde la toma de decisiones empresariales.

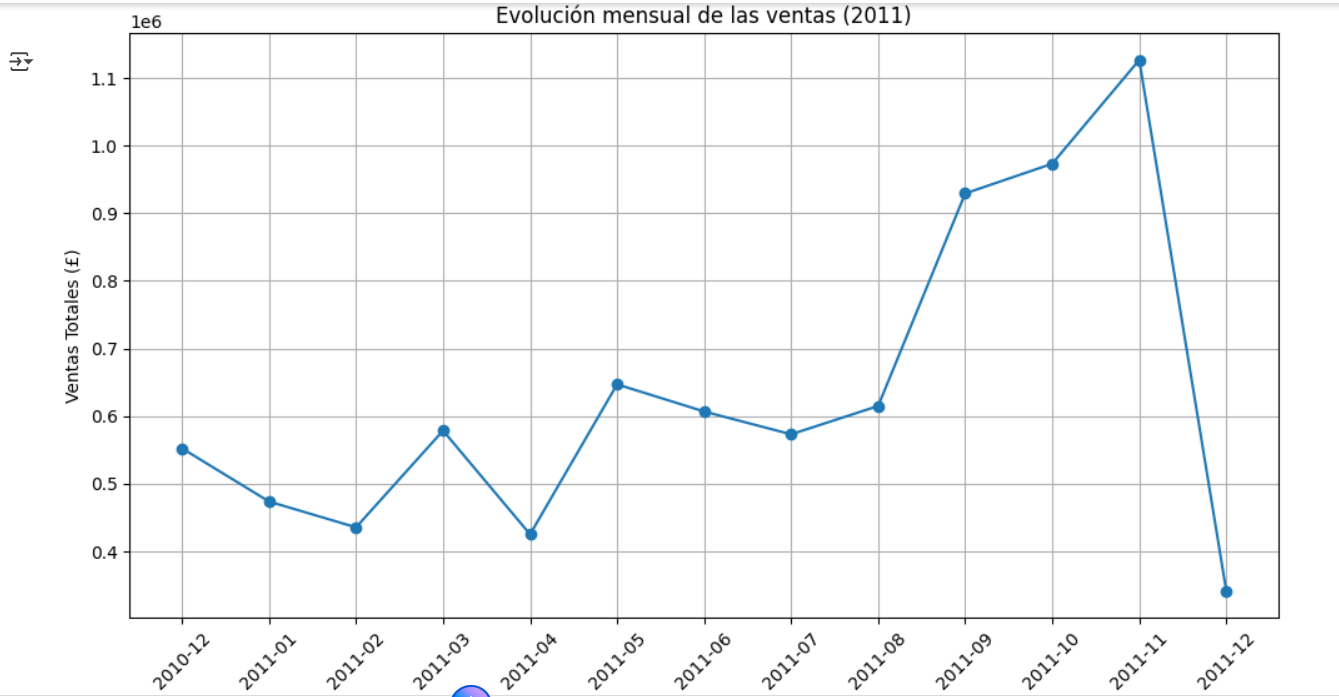
## **Metodología**

El desarrollo del trabajo se dividió en siete fases principales:

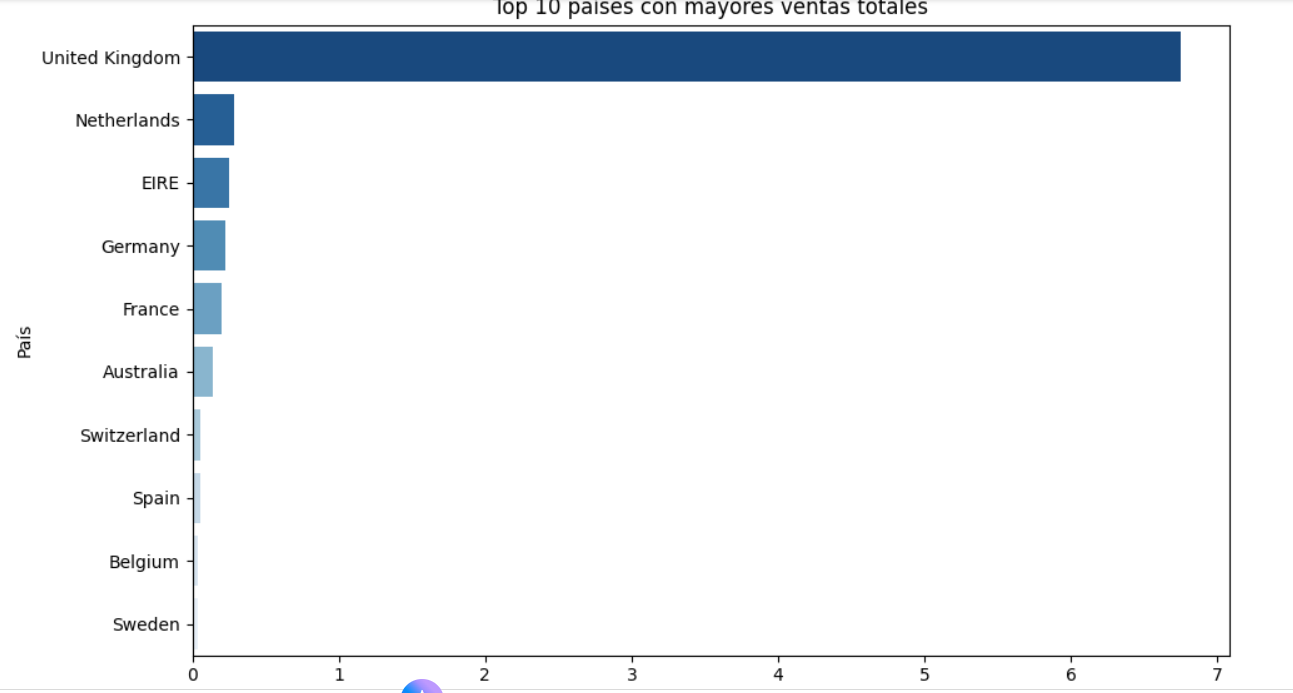
1. **Carga y exploración inicial:**  
   Se importó el dataset desde el repositorio UCI y se verificaron tipos de datos, valores nulos y duplicados.
2. **Limpieza y transformación:**  
   Se eliminaron registros duplicados y valores faltantes, se creó la columna Total = Quantity × UnitPrice y se normalizó el texto de las descripciones.
3. **Conexión y consulta SQL:**  
   Se creó una base de datos local ventas.db en SQLite y se cargaron los datos en una tabla llamada ventas.  
   Luego, se ejecutó una consulta SQL para calcular el **promedio del valor de transacción por país**.

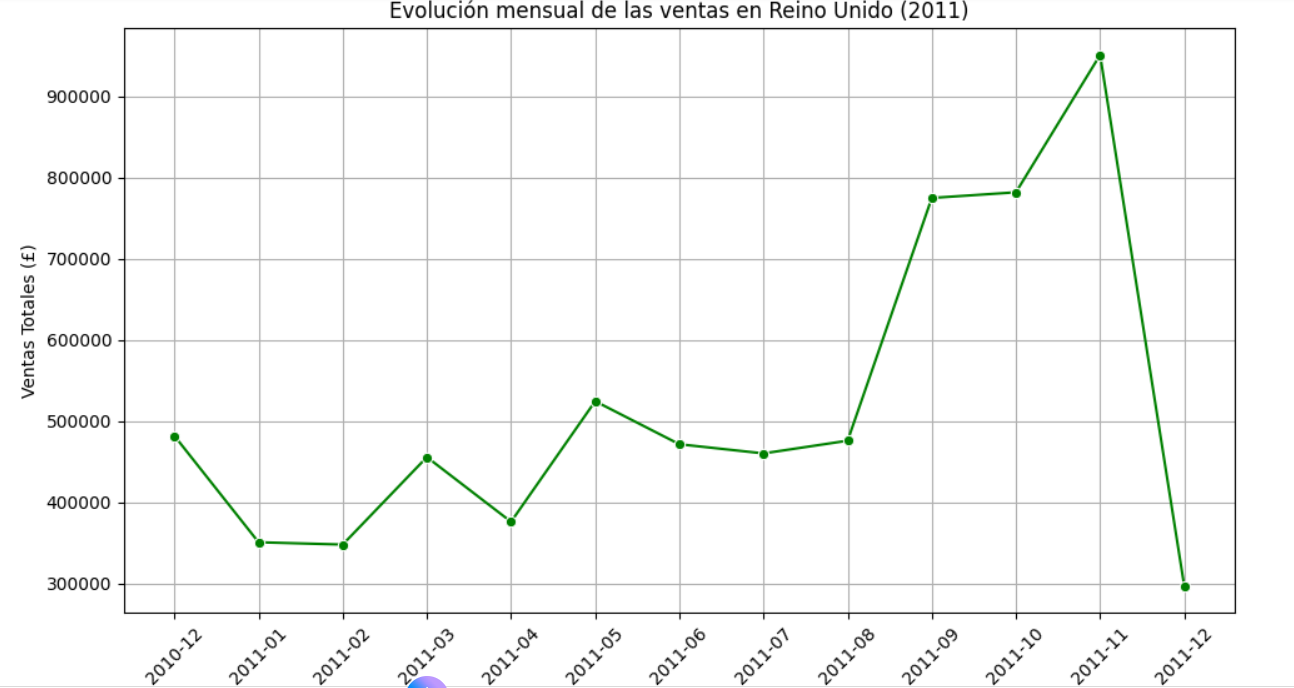


1. **Análisis temporal:**  
   Se agruparon las ventas por mes para observar la evolución de ingresos durante el año.



1. **Visualización comparativa:**  
   Se elaboraron gráficos que comparan los **países con mayores ventas totales** y la **evolución mensual del Reino Unido**.



****

1. **Métricas de clientes:**  
   Se calcularon indicadores por cliente en el Reino Unido: número de compras, gasto total y promedio por compra.
2. **Integración con API de divisas:**  
   Se conectó con la **API Frankfurter** para obtener tasas de cambio y convertir las ventas a **USD, EUR y COP**.

## **Resultados**

Los principales resultados obtenidos durante el proceso fueron los siguientes:

* Se **eliminaron registros nulos y duplicados**, conservando únicamente datos válidos para análisis.
* El **rango temporal** del dataset abarca desde **1 de diciembre de 2010** hasta **9 de diciembre de 2011**.
* El país con mayor número y valor promedio de ventas fue el **Reino Unido**, confirmando su papel como mercado principal.
* El análisis temporal mostró fluctuaciones mensuales con picos significativos en los meses intermedios del año.
* Las métricas de clientes revelaron diferencias notables en gasto y frecuencia de compra, permitiendo identificar clientes de alto valor.
* La conversión de divisas mostró cómo el valor promedio por transacción varía al expresarse en USD, EUR o COP.

## **Análisis**

El análisis de los resultados permitió extraer varias conclusiones clave:

* El flujo ETL fue fundamental para **garantizar la calidad de los datos** y la consistencia en las métricas calculadas.
* Las consultas SQL facilitaron la obtención de **resúmenes estadísticos precisos**, como el gasto promedio por país.
* El análisis temporal reveló patrones de **estacionalidad** en las ventas, información útil para planificar estrategias de marketing.
* Las métricas de clientes permiten segmentar y priorizar esfuerzos de **fidelización** hacia los compradores más rentables.
* La integración con la API de divisas demostró la **interoperabilidad** de sistemas, extendiendo el análisis económico a contextos internacionales.

## **Conclusión**

El taller permitió integrar distintas herramientas del ecosistema de ciencia de datos para desarrollar un flujo analítico completo.  
La aplicación de procesos **ETL, SQL, análisis temporal, visualización y consumo de APIs** permitió transformar un conjunto de datos sin procesar en información estructurada, visual y útil para la toma de decisiones empresariales.

El proyecto evidenció la importancia de la limpieza y validación de los datos, así como la capacidad de Python y sus bibliotecas para conectar diferentes etapas del análisis, desde la preparación hasta la presentación de resultados.  
Finalmente, el uso de visualizaciones y métricas por cliente demostró cómo el análisis de datos puede orientar estrategias comerciales más efectivas y personalizadas.