

**Dirección**

**Empresa Autonoma**  
Distrito Nacional  
Republica Dominicana

**WhatsApp**

Movil : + (829) 263-8429

**Online**

Email: teofilocorrea@gmail.com  
GitHub: <https://github.com/teofilocorrea>  
Linkedin: teofilocorrea



**CORPORATE**

# FORMULARIO DE INGRESO

Project ID #01

Registro de Recepción y Validación de Dataset Externo

---

**Preparado por:**

Teofilo Correa Rojas

**GitHub: teofilocorrea**

Distrito Nacional

**P : +(829) 263-8429**

**E : teofilocorrea@gmail.com**

---

**Emitido y válido**

Propuesta emitida :

**10.11.2025**

Propuesta válida hasta:

**21.11.2025**

---

**Declaración de confidencialidad**

La información contenida en este documento es de carácter confidencial y está destinada únicamente a fines de análisis, validación y documentación de procesos de datos. Su uso, reproducción o divulgación total o parcial sin autorización del autor está estrictamente prohibido.



# PROPUESTA DE PROYECTO

## TABLA DE CONTENIDO

CASO DE ESTUDIO	<b>04</b>
INFORMACIÓN GENERAL	<b>05</b>
REVISIÓN TÉCNICA INICIAL	<b>06</b>
ACCIONES REALIZADA	<b>07</b>
APROBACIÓN Y FIRMA	<b>08</b>
TERMINO Y CONDICIONES	<b>09</b>

# Caso de Estudio

## DESCRIPCIÓN GENERAL

---

El conjunto de datos **Dirty Cafe Sales** contiene 10 000 filas de datos sintéticos que representan transacciones de ventas en una cafetería. Este conjunto de datos es intencionalmente «sucio», con valores que faltan, datos incoherentes y errores introducidos para proporcionar un escenario realista para la limpieza de datos y el análisis exploratorio de datos (EDA). Se puede utilizar para practicar técnicas de limpieza, manipulación de datos e ingeniería de características.

## SOLUCIÓN

---

Practicar técnicas de limpieza de datos, incluyendo el manejo de valores faltantes, la eliminación de duplicados y la corrección de entradas no válidas, complementándolo con la exploración de técnicas EDA mediante visualizaciones y estadísticas resumidas, así como la realización de ingeniería de características orientada a flujos de trabajo de aprendizaje automático.

## ESPECIFICACIONES DEL PROYECTO

---

01. Manejo de valores faltantes.

---

02. Gestionar valores no válidos.

---

03. Consistencia de fechas.

---

04. Ingeniería de características.

# Información General

## Campo y Descripción

### NOMBRE DEL DATASET

dirty\_cafe\_sales.csv

### FUENTE

Proveedor externo / socio comercial

### REMITENTE

Ahmed Mohamed – Finanzas

### FECHA DE RECEPCIÓN

10/11/2025

### MEDIO DE RECEPCIÓN

Enlace de descargar

### RESPONSABLE DE CARGA

Teofilo Correa Rojas

### CLASIFICACIÓN

Kaggle

# Revisión técnica inicial

Campo y Descripción
<b>FORMATO</b> Código Seguro de Verificación (.csv)
<b>CODIFICACIÓN</b> UTF-8
<b>CANTIDAD DE REGISTROS</b> 10000
<b>COLUMNAS CON NULOS</b> 08
<b>DUPPLICADOS DETECTADOS</b> En verificación
<b>OBSERVACIONES</b> Columnas con espacios

# Acciones realizadas

Durante el procesamiento del dataset se llevaron a cabo acciones orientadas a garantizar la integridad, consistencia y trazabilidad de los datos antes de su integración al entorno analítico.

---

## BLOQUE 01

---

Este bloque corresponde al conjunto de pasos ejecutados en la etapa de limpieza y estructuración inicial del dataset Dirty\_Cafe\_Sales.csv, proveniente de Kaggle. El objetivo principal fue garantizar la integridad y coherencia de los registros antes de su incorporación a la capa RAW del entorno QA, siguiendo los estándares de calidad definidos para la transición hacia STAGING.

- **Preparar Entorno local:** DuckDB
  - **EDA Preliminar:** Identificación de valores nulos, duplicados y errores de formato.
  - **EDA RAW:** Validación post-carga en QA confirmando 10,000 registros válidos.
  - **Documento asociado:** dirty\_cafe\_sales.csv
-

# Aprobaciones/Firma

## Dirty Cafe Sales

Con esta firma se valida que la información documentada ha sido revisada y aprobada por las partes responsables. El firmante confirma haber verificado la consistencia técnica, la calidad de los datos y la correcta ejecución del proceso descrito.

Asimismo, se deja constancia de que los resultados reflejados en este documento cumplen con los estándares internos definidos para la gestión, limpieza y documentación de datasets. Cualquier modificación o uso posterior deberá ser comunicada al responsable técnico o al área encargada del control de versiones.

### Firma del cliente

---

Nombre / Titulo / Posición

---

Fecha

---

Firma del Cliente

---

Nombre de la compañía

---

Email Institucional

---

Número de Contacto

**GRACIAS POR CONFIAR EN NOSOTROS.**

# Términos y condiciones.

Este documento ha sido elaborado con fines técnicos y educativos. Su contenido está orientado a la gestión, validación y documentación de procesos de ingeniería de datos. Cualquier uso distinto deberá realizarse bajo responsabilidad del usuario o equipo que lo implemente.

## Términos y condiciones



La información contenida en este documento es propiedad intelectual de su autor. Se prohíbe su reproducción o distribución total o parcial sin autorización previa. Los datos utilizados con fines de práctica o demostración no representan información real ni pertenecen a ninguna entidad corporativa.

## MATERIAL



El material técnico presentado fue desarrollado con propósitos de aprendizaje, simulación y documentación de flujos de trabajo en ingeniería de datos. Se recomienda su uso exclusivo en entornos de práctica o proyectos personales, evitando su aplicación sobre datos sensibles o productivos sin la debida revisión de seguridad.



CORPORATE

# GRACIAS POR SU APOYO

---

Dirty Cafe Sales

---

## Dirección

**Empresa Autonoma**  
Distrito Nacional  
Republica Dominicana

## WhatsApp

Movil : + (829) 263-8429

## Online

Email: teofilocorrea@gmail.com  
GitHub: <https://github.com/teofilocorrea>  
Linkedin: teofilocorrea

---

## Un breve resumen

Otatempor ma quam facepudit vendit, quunte sitis anis ipsam et quas autem verferum ipient min cum eati vide volupid et, officiu ntibeatur solut aut am fuga. Cuptaest rerchil lendaes at erum nobis dunt quo oftem id maios ab idis aut ut plant ipsus maio te volendae. Tur simagnatur, atat ese rem rem quatur.

---