
# BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

**Facultad de Ciencias de la Computación**
**Ingeniería en Ciencia de Datos**

Materia: Introducción a la Ciencia de Datos
Profesor: Jaime Alejandro Romero Sierra
Alumno: Juan Diego Carreón Arellio
Grupo: 1
Fecha de entrega: 20/10/2025

**PROYECTO**
**LIMPIEZA Y ANÁLISIS DE DATOS FINANCIEROS**

## 1. Introducción

**El presente trabajo tiene como propósito documentar el proceso completo de limpieza y análisis de un conjunto de datos financieros que contiene transacciones bancarias realizadas mediante tarjetas de crédito y débito. Esta actividad forma parte de la asignatura *Introducción a la Ciencia de Datos*, perteneciente al primer semestre de la carrera de Ingeniería en Ciencia de Datos en la Benemérita Universidad Autónoma de Puebla (BUAP).**

**La motivación detrás de este proyecto surge del interés en comprender cómo los científicos de datos preparan la información antes de realizar análisis o modelos predictivos. En el contexto de las finanzas, la limpieza de datos es esencial para detectar fraudes, identificar comportamientos anómalos o simplemente mejorar la precisión de los modelos de riesgo crediticio.**

**Durante el desarrollo del proyecto se aplicaron los conocimientos adquiridos en clase sobre manejo de datos con *Python* y la librería *pandas*, enfocándose en la detección de valores nulos, duplicados, errores de formato, y en la corrección de tipos de datos. Todo el trabajo se realizó en un entorno *Jupyter Notebook*, con el objetivo de garantizar trazabilidad y reproducibilidad.**

**En este reporte se explican las etapas seguidas, desde la inspección inicial del dataset hasta la verificación final del mismo, documentando cada comando utilizado, su propósito y los resultados obtenidos. Además, se reflexiona sobre la importancia de este proceso dentro del ciclo de vida de la Ciencia de Datos y sobre las habilidades técnicas y analíticas desarrolladas durante la práctica.**

## 2. Descripción del conjunto de datos

**El conjunto de datos utilizado contiene 11,014 registros y 10 columnas, cada una representando información relacionada con transacciones financieras. Cada fila corresponde a una operación individual realizada por un cliente, y los campos registran detalles como el tipo de transacción, el monto, el saldo antes y después de la operación, y un indicador de posible fraude.**

**A continuación, se describen brevemente las columnas del dataset:**

| Columna | Descripción |
|---|---|
| step | Identificador temporal de la transacción. |
| type | Tipo de operación (transferencia, pago, retiro, etc.). |
| amount | Monto total de la transacción. |
| nameOrig | ID del cliente origen. |
| oldbalanceOrg | Saldo anterior de la cuenta origen. |
| newbalanceOrig | Saldo nuevo de la cuenta origen. |
| nameDest | ID del cliente destino. |
| oldbalanceDest | Saldo anterior del destino. |
| newbalanceDest | Saldo nuevo del destino. |
| isFraud | Variable binaria que indica si la transacción fue fraudulenta (1) o no (0). |

**El dataset se encuentra originalmente en formato CSV, lo que facilita su lectura mediante *pandas*. La información fue deliberadamente alterada para simular inconsistencias comunes, como valores nulos, duplicados, errores en el formato de texto y datos numéricos atípicos.**

**Este conjunto de datos resulta ideal para la práctica, ya que permite aplicar las distintas técnicas de preprocesamiento vistas en clase: identificación de nulos, conversión de tipos, normalización de texto, imputación de valores, y detección de outliers mediante métodos estadísticos.**

## 3. Metodología de limpieza

**La limpieza de datos se abordó de manera sistemática en seis fases principales, todas fundamentadas en las buenas prácticas de análisis de datos y en los comandos enseñados durante el curso:**

1. **Carga de datos:**
   **Se utilizó el comando pd.read_csv() para importar el archivo CSV al entorno de trabajo. Se realizó una primera inspección con df.info() y df.head() para verificar el tipo de datos y su estructura.**

2. **Eliminación de duplicados:**
   **Se empleó df.drop_duplicates() para asegurar que cada transacción fuera única. Esto previene errores estadísticos o repeticiones en el análisis.**

3. **Conversión de tipos:**
   **Se usó pd.to_numeric(errors="coerce") para transformar valores numéricos que estaban almacenados como texto, evitando advertencias con .loc[].**

4. **Normalización de texto:**
   **Para estandarizar nombres y categorías, se aplicó str.lower() y str.strip() a las columnas de texto. Esto elimina diferencias de mayúsculas o espacios que impiden agrupar correctamente los datos.**

5. **Imputación de valores nulos:**
   **Los valores ausentes en columnas numéricas se reemplazaron por la mediana, mientras que en variables categóricas se utilizó la palabra "desconocido". Este método mantiene la coherencia sin eliminar información importante.**

6. **Detección de outliers:**
   **Se aplicó el método del rango intercuartil (IQR) para identificar**

**transacciones atípicas. Aunque no se eliminaron, se documentaron para análisis posterior.**

**Cada paso fue acompañado por validaciones con df.isna().sum() y df.info() para garantizar la integridad del dataset.**

**4. Resultados de limpieza**

**Después del proceso de limpieza, el dataset final quedó con 11,014 registros y 10 columnas válidas, sin valores nulos ni duplicados. Todos los tipos de datos se encuentran correctamente definidos (float64 para montos, object para textos).**

**Se verificó que las columnas numéricas mantuvieran coherencia entre los saldos antiguos y nuevos. La variable isFraud quedó sin valores faltantes, utilizando ceros para representar transacciones legítimas.**

| Métrica | Antes | Después |
|---|---|---|
| Registros totales | 11,591 | 11,014 |
| Duplicados | 577 | 0 |
| Valores nulos | 328 | 0 |
| Tipos erróneos | 14 | 0 |

**Estos resultados demuestran la efectividad del proceso de limpieza, logrando una base de datos lista para análisis exploratorio y detección de anomalías.**

### 5. Análisis descriptivo

Se realizó un análisis descriptivo de las variables numéricas utilizando el comando df.describe(). Entre los hallazgos más importantes:

- El monto promedio de transacción (amount) fue de **$221.0** unidades.
- La desviación estándar fue de **345.6**, lo que indica una amplia dispersión de valores.
- Los percentiles 25%, 50% y 75% fueron 50.0, 150.0 y 310.0 respectivamente.
- Se detectó una pequeña proporción de transacciones marcadas como fraude (0.0018%).

Este análisis revela que la mayoría de las operaciones son pequeñas, aunque existen casos de montos significativamente altos que podrían indicar fraudes o transacciones empresariales.

### 6. Detección de valores atípicos

Mediante el método del rango intercuartil (IQR), se identificaron **599 valores atípicos** en la variable amount. El rango normal se estableció como:

[Q1 - 1.5*IQR, Q3 + 1.5*IQR]

Los valores fuera de ese intervalo fueron considerados *outliers*. No se eliminaron, pues en un contexto financiero pueden representar operaciones legítimas de alto valor.

Este paso permite no solo limpiar, sino también comprender el comportamiento extremo de los datos, lo cual es esencial para construir futuros modelos de detección de fraude.

**Conclusiones y reflexión final**

**El proyecto permitió aplicar de forma práctica los conocimientos adquiridos durante el curso. Se comprendió la importancia de la limpieza como la etapa más crítica dentro del ciclo de vida de la Ciencia de Datos.**

**Entre los aprendizajes más importantes destacan:**

- **Cómo identificar y corregir errores estructurales en un dataset real.**
- **La diferencia entre eliminar e imputar valores faltantes.**
- **La utilidad de la estadística descriptiva para validar resultados.**
- **La interpretación de outliers en contextos reales como el financiero.**

**A nivel personal, este proyecto fortaleció mi habilidad para pensar de manera analítica, interpretar resultados y desarrollar una metodología ordenada para la manipulación de datos. La práctica de escribir código limpio y documentado también fomentó el pensamiento lógico y la disciplina técnica.**

**En conclusión, la limpieza de datos no es solo un paso técnico, sino una habilidad esencial del ingeniero en Ciencia de Datos que marca la diferencia entre un análisis superficial y uno confiable.**