**TAREA: Limpieza de Datos** 

Nombre: Jaime Alexis Hernandez Franco

Materia: Introducción a la Ciencia de Datos

**Profesor:** Jaime Alejandro Romero Sierra

**Fecha:** 20/10/2025

Repositorio GitHub: <u>Alexiiiiss0122/CIENCIA-DE-DATOS-2</u>

## Descripción inicial de la base de datos

La base contiene información sobre clientes bancarios, sus características personales, financieras y de contacto, así como si se suscribieron a una campaña de marketing.

Columna	Descripción
age	Edad del cliente
job	Ocupación
marital	Estado civil
education	Nivel educativo
balance	Saldo promedio en la cuenta
housing	Tiene crédito hipotecario (yes/no)
loan	Tiene otro préstamo (yes/no)
contact	Tipo de contacto (celular o teléfono)
campaign	Número de contactos realizados durante la
	campaña
previous	Número de contactos previos
poutcome	Resultado de la campaña anterior
subscribed	Si el cliente se suscribió (yes/no)

## Proceso de limpieza de datos

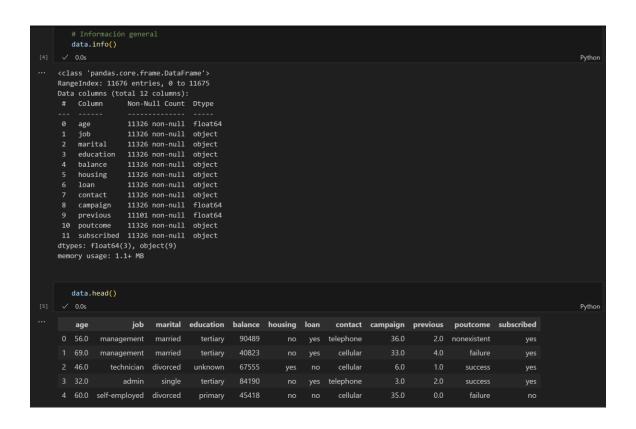
Se realizó un análisis completo para detectar valores nulos, duplicados y errores de texto. Los valores faltantes se rellenaron con el promedio (para columnas numéricas) o con la moda (para columnas categóricas). Los valores anómalos como 'Auto%#' y 'unknown' fueron eliminados y no remplazados. Se eliminaron duplicados también.

## Ejemplos de técnicas aplicadas:

- Relleno de nulos con promedio o moda usando fillna().
- Conversión de columnas numéricas con pd.to\_numeric( no se pudo con balance).

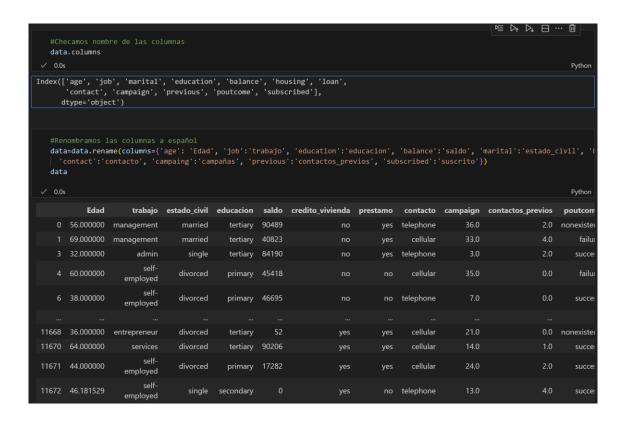
- Eliminacon de palabras inválidas ('Auto%#', 'unknown').
- Eliminación de duplicados con drop\_duplicates().

	rt pa	os la libreria ndas as pd	que vamo	os a ocupar									Pyth
	= pd.	la base sucia read_csv("Base	_Sucia.cs	v")									Pyth
	age	job	marital	education	balance	housing	loan	contact	campaign	previous	poutcome	subscribed	
	56.0	management	married	tertiary	90489	no	yes	telephone	36.0	2.0	nonexistent	yes	
	69.0	management	married	tertiary	40823	no	yes	cellular	33.0	4.0	failure	yes	
	46.0	technician	divorced	unknown	67555	yes	no	cellular	6.0	1.0	success	yes	
	32.0	admin	single	tertiary	84190	no	yes	telephone	3.0	2.0	success	yes	
	60.0	self-employed	divorced	primary	45418	no	no	cellular	35.0	0.0	failure	no	
11671	44.0	self-employed	divorced	primary	17282	yes	NaN	cellular	24.0	2.0	success	no	
11672	NaN	self-employed	single	secondary	NaN	yes	no	telephone	13.0	4.0	success	no	
11673	44.0	retired	married	NaN	13685	yes	no	telephone	46.0	3.0	nonexistent	yes	
11674	66.0	blue-collar	married	secondary	62446	no	yes	telephone	25.0	3.0	nonexistent	no	
11675	64.0	blue-collar	divorced	secondary	-889	no	no	telephone	22.0	2.0	nonexistent	yes	
11676 ro	ws × 1.	2 columns											

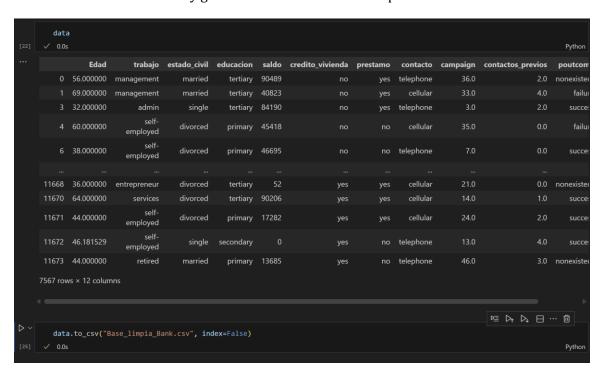








Al final checamos el dataset y guaramos la base de datos limpia.



## **Conclusiones**

Durante la limpieza se identificaron valores nulos, duplicados y errores introducidos intencionalmente para practicar el proceso. Se aplicaron técnicas de imputación y corrección de registros, manteniendo la integridad de la base y traducción de las columnas para facilitar su entendimiento. El resultado final es un conjunto de datos limpio, y listo para su análisis.