





Programa de especialización

DATA DRIVEN DECISIONS 2026

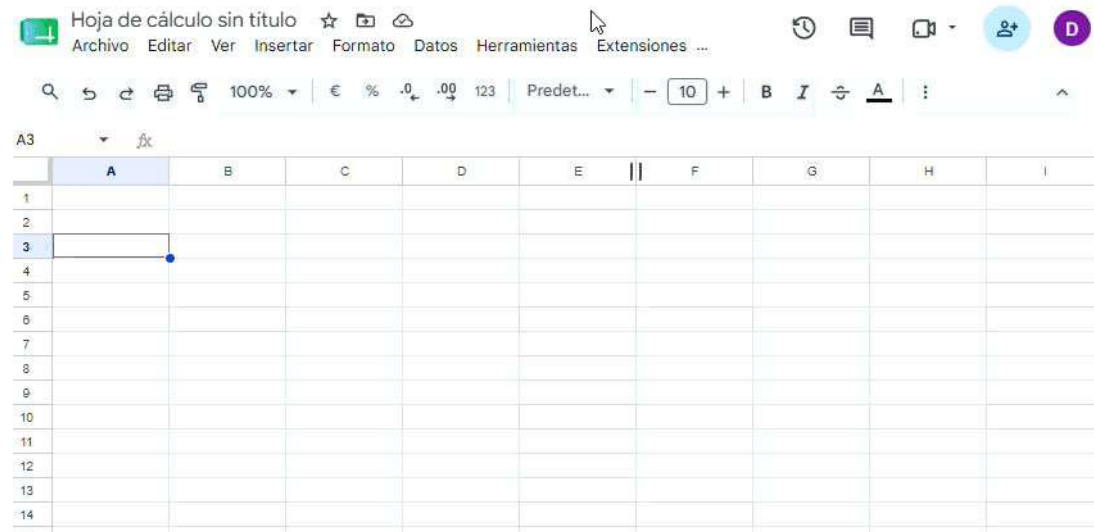
Nivel Intermedio

SESIÓN 3: Principios de limpieza de datos

TRANSFORMAR AHORA,
trascender siempre

Antes de comenzar: ¡verifiquemos tu instalación!

Antes de iniciar la sesión, asegúrate de tener instalado y activo el complemento AI for Sheets en Google Sheets.



Sesión 3: Principios de limpieza de datos

Logro de aprendizaje

Al finalizar la sesión, lograrás:

Justificar la importancia de la limpieza de datos en la obtención de resultados precisos.

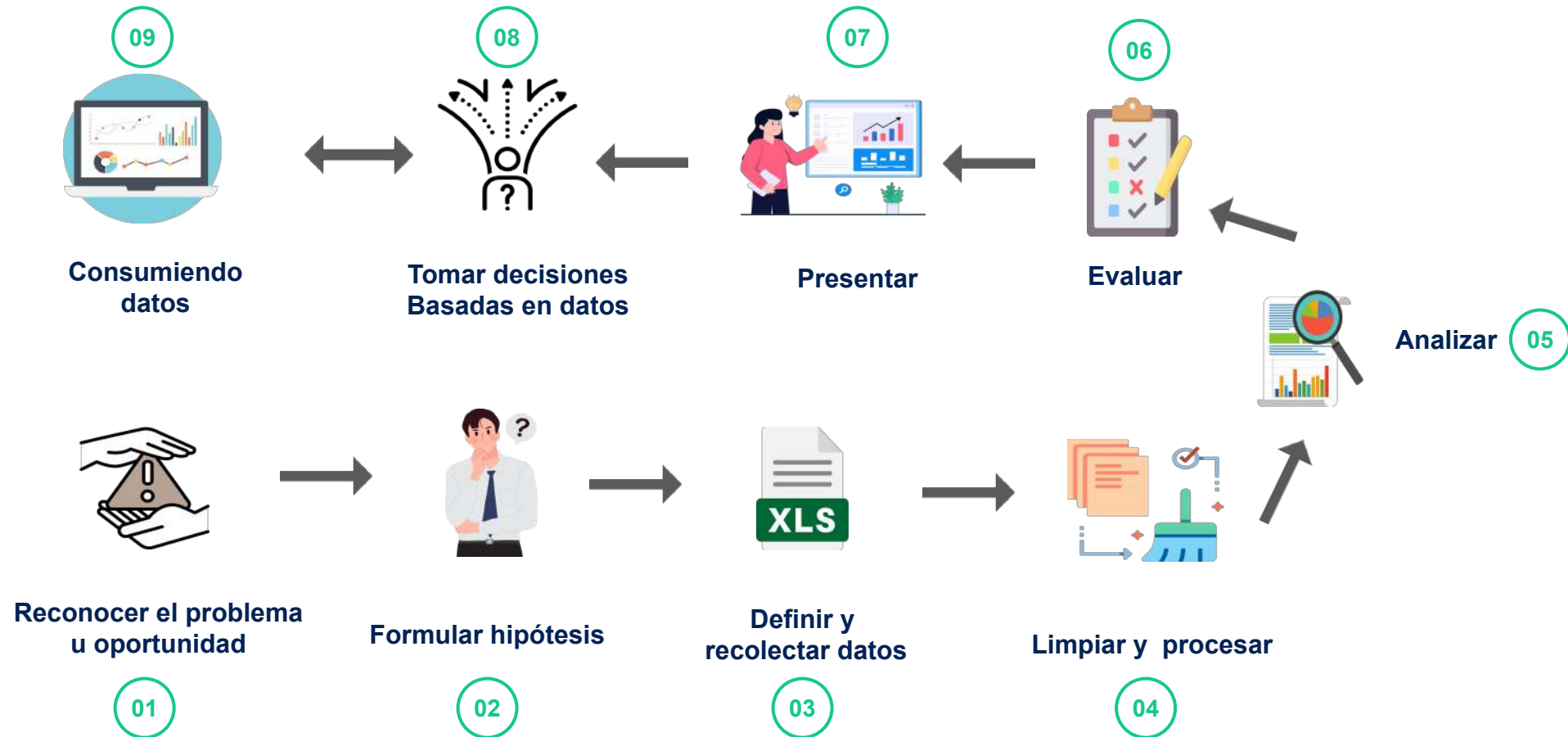
Conocer un framework para la limpieza de datos.

Contenido:

- 1. Proceso de análisis de datos
- 2. Definición limpieza de datos
- 3. Framework de limpieza de datos
- 4. Ejercicios con IA

1.Proceso de análisis de datos

En el día a día



Reconocer el problema o cuál es la oportunidad

Al iniciar el proceso de análisis, es fundamental definir las bases:

01

Identificar el problema

Comprender con claridad cuál es el problema central o la oportunidad de mejora, determinando con precisión por qué es importante resolverlo.

02

Partes interesadas

Reconocer e involucrar a todos los actores clave del problema en la organización, abarcando desde los líderes estratégicos hasta los analistas.

03

Tomar decisiones

Definir con anticipación qué tipo de decisiones y acciones concretas se podrán implementar a partir de los resultados obtenidos en el análisis.

Formular hipótesis

Para iniciar el proceso de análisis de datos, es fundamental:

Definir el problema

Definir muy bien el problema central de negocio o investigación que vamos a abordar mediante el análisis de datos.

Identificar variables

Determinar con precisión cuáles son las principales variables que vamos a incluir en el análisis para validar la hipótesis.



Formular hipótesis

Guiarse con los resultados de análisis previos

- **Qué datos utilizaron:** Identificar las fuentes de información y variables críticas consideradas.
- **Qué tipo de análisis:** Evaluar las metodologías y enfoques analíticos que se aplicaron.
- **Qué resultados obtuvieron:** Comprender las conclusiones para construir sobre terreno firme.

Otros Recursos

- Internet
- Libros
- Conversar con analista
- Gestión de conocimiento de la universidad



Encuadre del problema

¿Has definido claramente un problema u oportunidad que sea importante para tu área/ organización?

¿Has identificado a los interesados para el problema?

¿Estás seguro de que la manera en que planeas resolver el problema resonará con los interesados y que utilizarán los resultados para tomar una decisión?

¿Has considerado múltiples formas/ alternativas de resolver el problema?



Preparación de datos

Fases esenciales para el procesamiento de información:

01

Encontrar datos

Identificar y recopilar las fuentes de información necesarias para iniciar el proceso de análisis.

02

Limpiar datos

Detectar y corregir errores, valores nulos o duplicados para garantizar la máxima calidad de la información.

03

Reestructurar

Organizar y adaptar el formato de los conjuntos de datos para facilitar su posterior procesamiento.

04

Combinar fuentes

Consolidar y fusionar múltiples bases de datos para obtener una visión unificada y enriquecida.

Preparación de datos

30%-80%

del tiempo total

de un proyecto de analítica se destina a:

01 Reorganizar y reestructurar

02 Limpiar nulos y duplicados

03 Corregir errores tipográficos

Analizar los datos

Descriptiva

¿Qué está pasando o qué pasó?

Diagnóstico

¿Por qué está pasando o pasó?

Prescriptiva

¿Qué debemos hacer?

Inferencial

¿Qué pasó con el resto?

Predictiva

¿Qué es lo más probable que pase?

Presentar datos y tomar decisiones basadas en datos

- Comprender la audiencia
- Identificar los objetivos
- Seleccionar las visualizaciones adecuadas
- Contextualizar los datos
- Destacar puntos clave
- Fomentar la participación y el debate
- Tomar decisiones basadas en datos



Repaso

Es hora de poner a prueba lo aprendido. Únete a la actividad interactiva ingresando al enlace desde tu computadora o dispositivo móvil.



Respuestas a repaso

01. ¿Existe una fórmula mágica única para limpiar los datos?

R/ No

02. Este no es un problema de limpieza de datos

R/ El tamaño de la muestra

03. Puedo realizar un análisis de datos sin realizar limpieza y tengo la seguridad de que los resultados no van a presentar errores

R/ Falso

04. Conocer el número de estudiantes matriculados el año pasado hace parte de

R/ Analítica descriptiva

05. Responder ¿Por qué es más probable que algunos estudiantes no caigan en bajo rendimiento? hace parte de la analítica

R/ Diagnóstico

06. ¿Qué familias tienen mayor probabilidad de no pagar a tiempo la matrícula? Esta pregunta la puedo resolver con la analítica

R/ Predictiva

2. Definición limpieza de datos

¿Qué es limpiar los datos?

Definición:

Proceso de corregir o eliminar datos incorrectos, corruptos, formateados incorrectamente, duplicados o incompletos dentro de un conjunto de datos.

Cuándo es común que ocurra:

- En una sola fuente de datos, especialmente en variables de texto y digitadas manualmente
- Al combinar varias fuentes de datos



¿Por qué es importante limpiar los datos?

- Si los datos son incorrectos, los resultados (desde las visualizaciones hasta los resultados de los algoritmos) no van a ser confiables, aunque parezcan correctos.
- El riesgo es alto si se piensa tomar decisiones con datos sin limpieza.

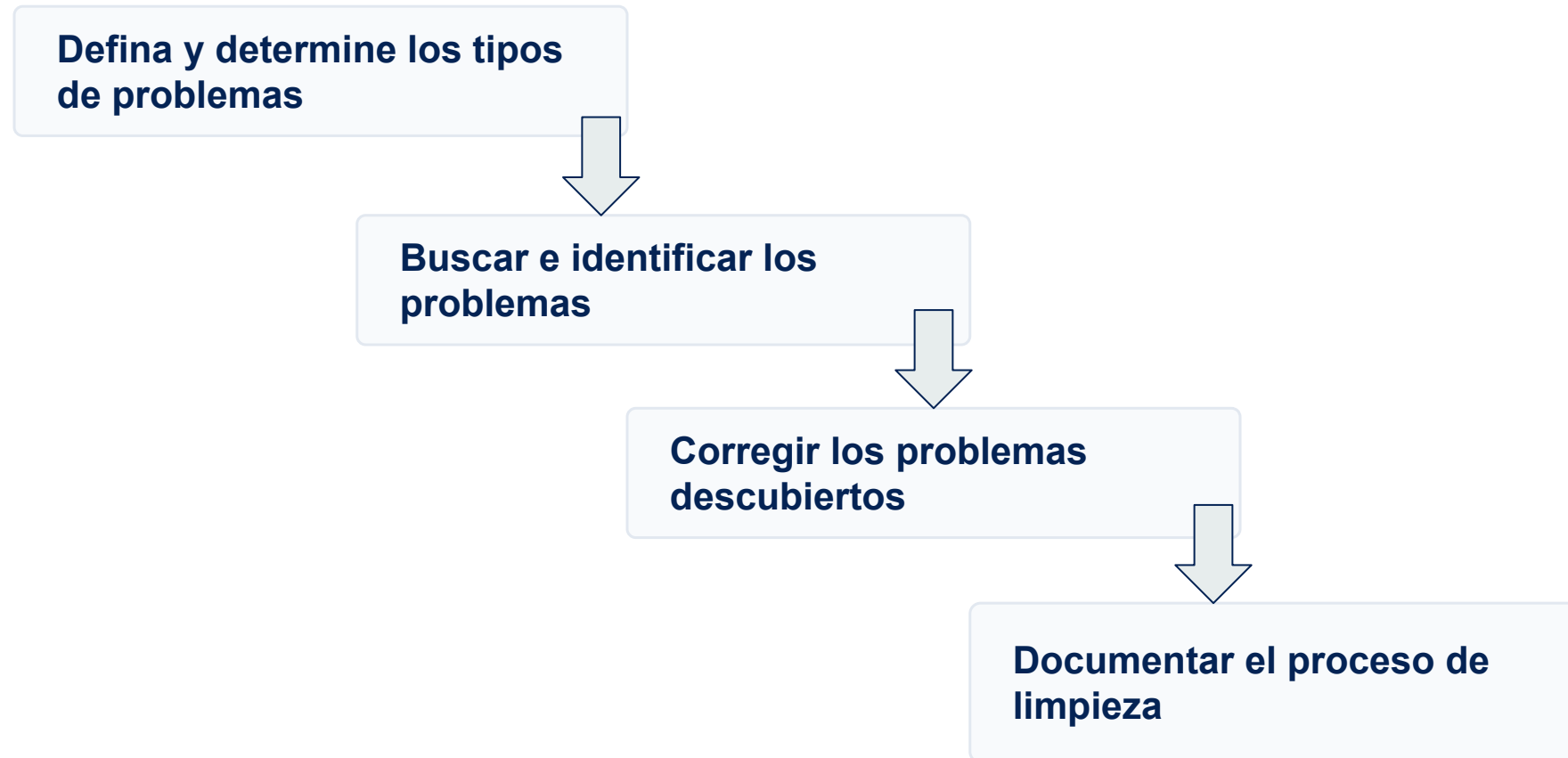


¿Existe una fórmula para limpiar los datos?

No existe una forma única que indique cuáles son los pasos exactos del proceso de limpieza de datos y cómo hacerlos porque los procesos variarán de un conjunto de datos a otro.

3. Framework limpieza de datos

Framework de limpieza de datos



Principales problemas en la limpieza de datos

Observaciones
duplicadas

Valores perdidos

Errores de digitación

Valores
inconsistentes

Valores fuera del
diccionario de
variables

Formato de la
variable no coincide
con su tipo

Observaciones duplicadas

Observaciones duplicadas

Flujo de De duplicación

ID	VALOR
101	Activo
102	Inactivo
102	Inactivo
103	Activo

Set de Datos Original



ID	VALOR
101	Activo
102	Inactivo
103	Activo
[Eliminado]	

Limpieza Aplicada

Observaciones duplicadas

ID estudiante	Fecha	Actividad
123	15-may-2024	Clase de piano
123	15-may-2024	Psicología
123	15-may-2024	Fútbol

¿Está o no duplicada?

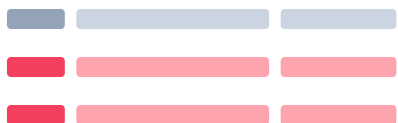
Depende de lo que queramos responder

Observaciones duplicadas

01

Duplicados Exactos

Filas idénticas en todas las columnas. Representan una redundancia absoluta de la información registrada.



02

Duplicados Parciales

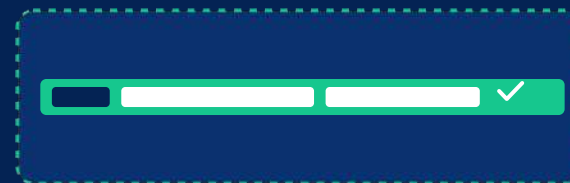
Filas que comparten el mismo identificador clave pero difieren en el contenido de otros campos o columnas.



03

Solución

Definir una regla de negocio para priorizar y conservar el registro más reciente o con la información más completa.



Valores perdidos

Valores perdidos

- Identifique dónde se encuentran los valores perdidos y si es necesario recodifíquelos correctamente.
- Intente entender la naturaleza de los datos perdidos.
- Determine cuál es el mejor método para tratarlos

ID	Edad	Sexo	Ciudad
123	19	F	Lima
456		F	NA
789	24	999	Lima

Valores perdidos

Atrición por un proceso natural o social

Pérdida aleatoria de información

Patrón de salto en la encuesta

**Falta intencional/sistemática como parte del
proceso de recopilación de datos**

Valores perdidos

Determine cuál es el mejor método para tratarlos.

Reconocer su existencia y no efectuar tratamiento alguno. Pero tenerlo en cuenta en la aplicación de los modelos y el análisis.

Eliminar observaciones con valores perdidos

Imputación de datos: sustituir valores perdidos en un campo por otros empleando:

- Media
- Mediana
- Moda
- Modelos estadísticos



Errores de digitación

Ejemplos

Convenciones de denominación

Lima
Ciudad de Lima

Representaciones diferentes

Sí
Si
si
sí

13-03-2024
03-13-2024
3 mar 2024

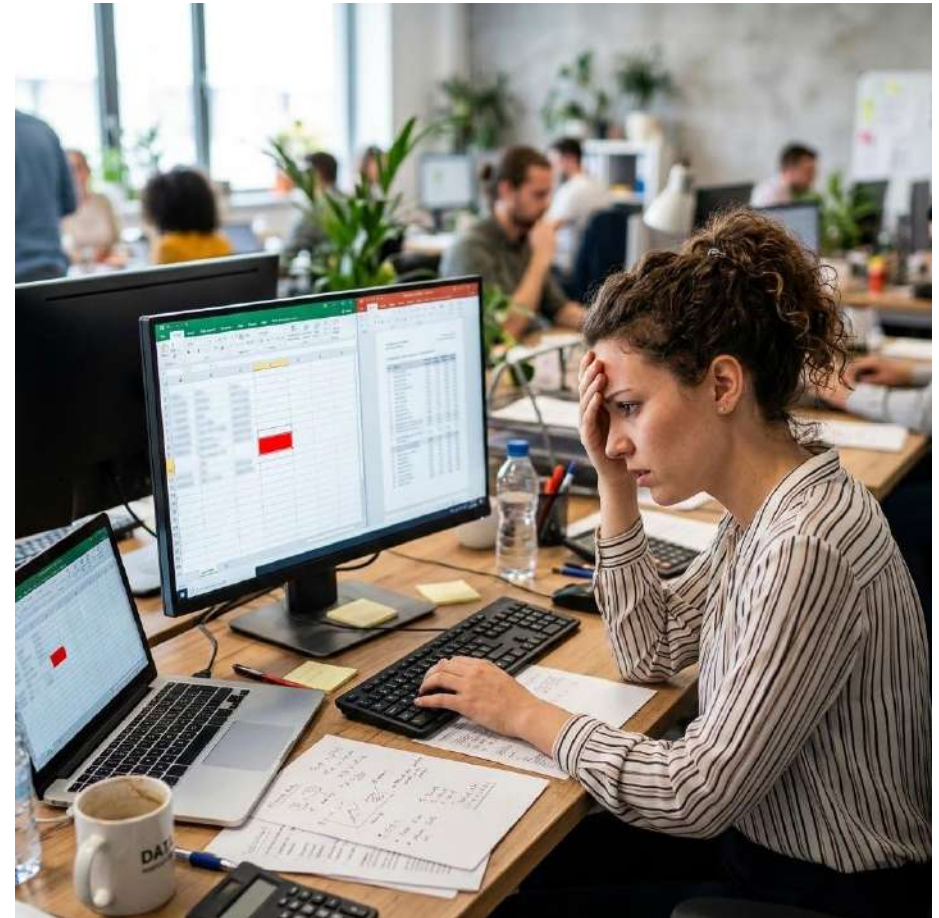
Espacios vacíos

“Mujer”
“ Mujer ”
“Mujer “
“ Mujer”

Errores de digitación

Soluciones

- Estandarice las opciones de respuesta desde la fuente de origen
- Intente recolectar los datos usando medios digitales
- Estandarice la digitación:
 - Uso de tildes
 - Uso de mayúsculas
 - Digitación de fechas
- Con los datos:
 - Utilice la función reemplazar para homogenizar respuestas y eliminar espacios vacíos



Valores inconsistentes

Valores inconsistentes

¿Qué es un dato inconsistente?

Es información que, a pesar de cumplir con el formato técnico, carece de coherencia lógica o viola las reglas fundamentales del negocio.

Edad	Fecha de nacimiento	Edad calculada
27	1990/01/17	27
60	11/11/1956	60
47	7/04/1940	77

Datos inconsistentes

Principales inconsistencias de lógica y registro en las variables:

Edad	Número de hijos
10	9
15	0
25	1

Nombre	Estatura
Andrés	-1.5
Daniela	1.65
María	1.77

Valores inconsistentes

Posibles soluciones

- Convertir los valores inconsistentes en valores perdidos.
- Sin embargo, si tiene acceso a la fuente de datos original (encuesta), puede revisar las respuestas y corregir si es necesario.

Edad	Fecha de nacimiento	Edad calculada
27	1990/01/17	27
60	11/11/1956	60
47	7/04/1940	77

**Valores por fuera del diccionario de
datos**

Diccionario de datos

Un diccionario de datos es documento de referencia que nos sirve para identificar las variables que componen la base de datos. Nos da información de nombres, definiciones y atributos.

Por ejemplo:

- Nombre de la variable : Sexo
- Descripción: Hace referencia al sexo de la persona
- Valores que puede tomar
- 1: Mujer
- 0: Hombre



Valores por fuera del diccionario de datos

Nombre	Estado civil
Paola	1
Esteban	4
Cesar	5
Carlos	8
Mayra	2

Código Estado Civil	
1	Soltero
2	Casado
3	Separado o divorciado
4	Unión libre
5	viudo

Diccionario de Datos

La unificación conceptual resuelve la discrepancia de nombres entre sistemas de origen.

El Desafío de Múltiples Fuentes

Diferentes sistemas de origen suelen registrar la misma entidad bajo nombres inconsistentes. Esto genera silos de información y dificulta los análisis cruzados.

COD_ALUMNO

ID_ESTUDIANTE

EALUMNO_ID

La Solución: Tabla de Mapeo

La unificación conceptual se logra implementando reglas de homologación claras en el modelo final:

- **Mapeo Centralizado**
- **Capa Semántica Única**

Esquema de Integración y Unificación

Académico (S1)
COD_ALUMNO

Matrícula (S2)
ID_ESTUDIANTE

Biblioteca (S3)
ALUMNO_ID

TABLA DE MAPEO

S1.COD_ALUMNO
→ id_estudiante

S2.ID_ESTUDIANTE
→ id_estudiante

S3.ALUMNO_ID
→ id_estudiante

MODELO FINAL
id_estudiante
✓ Unificado

**Formato no coincide con el tipo de
dato**

Formato no coincide

- La edad es una variable cuantitativa y debería tener un formato numérico, si tiene un formato de texto puede ocasionar problemas para realizar cálculos y estimar modelos.
- Todo depende de la herramienta empleada.
- Solución: verifique el formato y haga las correcciones pertinentes

Edad
18
41
23
50
17

Formatos no adecuados

Errores comunes de formato que dificultan el procesamiento y análisis de datos

01

Inconsistencia de Fechas

Fechas registradas como texto o con múltiples formatos mezclados en la misma columna.

15/04/2026 ✖

04-15-26 ✖

2026-04-15 ✔

02

Números como Texto

Presencia de símbolos de moneda o comas que fuerzan al sistema a interpretar valores numéricos como texto.

S/ 1,250.50 ✖

1250.50 ✔

03

Pérdida de Ceros

Identificadores (DNI, RUC) que pierden el cero a la izquierda al ser importados incorrectamente como números.

9876543 ✖

"09876543" ✔

04

Estandarización ISO

Conversión y validación obligatoria de todas las fechas al estándar internacional ISO 8601 (YYYY-MM-DD).

ESTÁNDAR ISO 8601

YYYY-MM-DD

✔ RECOMENDADO

Formatos de Fecha

Guía de normalización y control de calidad en procesos de preparación de datos.

01

Origen de Datos

Importación de reportes externos con fechas configuradas en formato americano.

Formato Recibido:

MM/DD/YYYY: 12/31/2025

Riesgo de ambigüedad:

Confundir días con meses (ej: 05/06 puede ser 5 de junio o 6 de mayo).

02

Transformación

Conversión activa de la estructura de la fecha al estándar local o de base de datos.

Formatos Destino:

Local DD/MM/YYYY: 31/12/2025

ISO YYYY-MM-DD: 2025-12-31

Beneficio clave:

Permite ordenamiento cronológico correcto y consultas directas en SQL/BI.

03

Validación de Tipos

Asegurar que el software de análisis reconozca correctamente el tipo de cada variable.

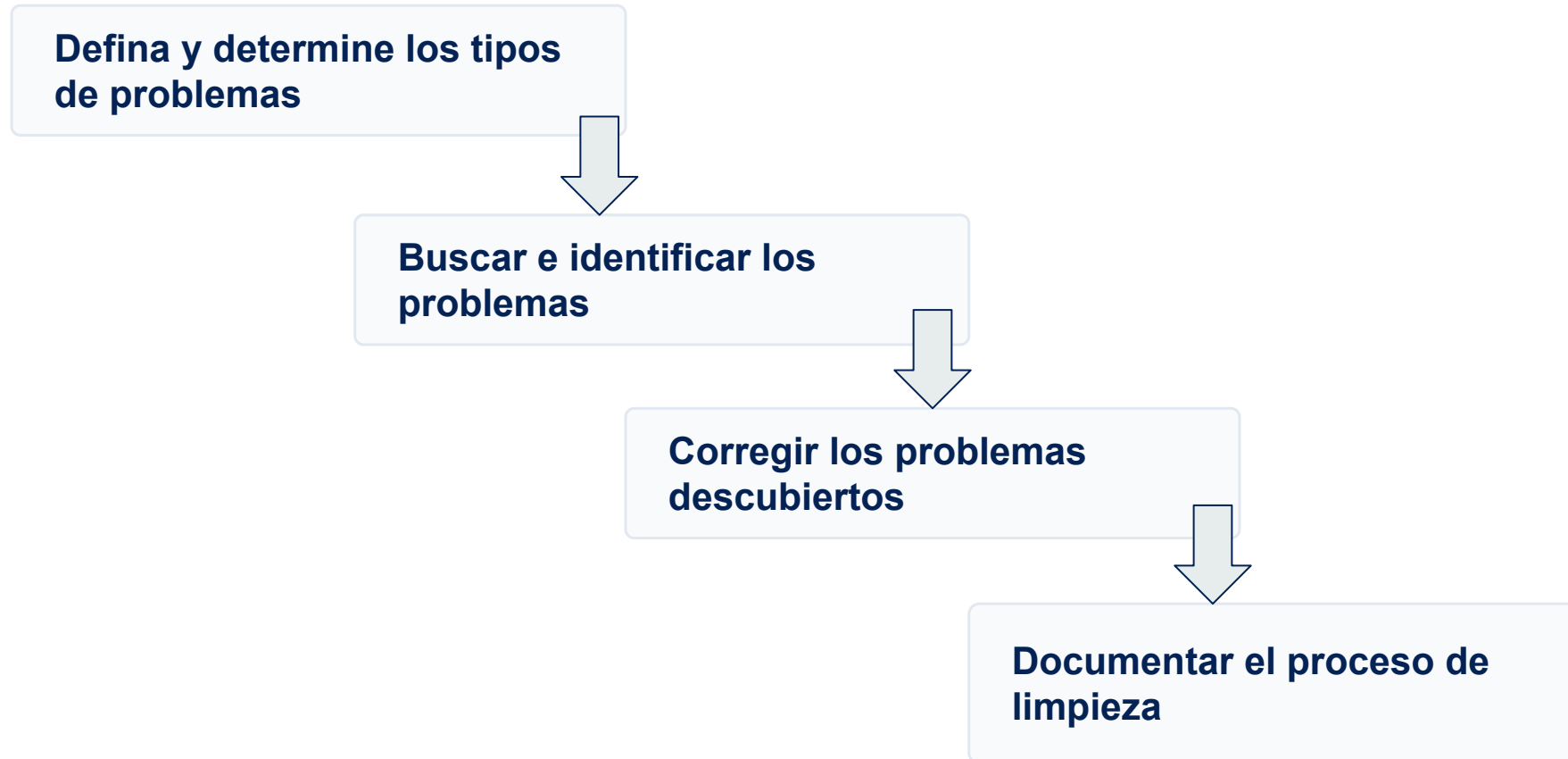
Puntos de Control:

✓ **Tipo Date:** Habilita filtros temporales y gráficos de series de tiempo sin errores.

✓ **Tipo Numérico:** Verifica campos de cálculo numérico.

✓ **Tipo Texto:** Mantiene formatos de códigos con ceros a la izquierda (ej: IDs).

Framework de limpieza de datos



Resumen del Framework de Limpieza

Principios de diseño para asegurar la calidad de los datos

Proceso Secuencial

La limpieza es un proceso secuencial e iterativo que evoluciona de manera constante.

Preparación de Datos

Cada paso del flujo de trabajo prepara los datos refinándolos para la fase siguiente.

Resultados Confiables

Un dataset limpio garantiza análisis reproducibles y la construcción de modelos confiables.

Buenas prácticas de limpieza de datos

01

Copia de seguridad (Backup)

Guardar siempre una copia del dataset original (Backup).

02

Documentar transformaciones

Documentar las reglas de limpieza y transformaciones aplicadas.

03

Automatizar scripts

Automatizar scripts de limpieza para procesos recurrentes.

04

Validar resultados

Validar resultados con expertos del dominio de negocio.

Herramientas para limpieza de datos

01

Hojas de cálculo (Excel, Google Sheets)

Limpieza básica mediante filtros, ordenamiento y funciones/fórmulas esenciales.

02

Power Query (Power BI / Excel)

Procesamiento ETL visual e intuitivo sin necesidad de escribir código complejo.

03

Lenguajes de programación (Python / R)

Automatización avanzada con librerías potentes como Pandas para flujos reproducibles.

04

Herramientas de IA generativa

Asistencia inteligente en la creación de expresiones complejas y detección automatizada de patrones erróneos.

Checklist final de limpieza

☒ ¿Sin filas ni identificadores duplicados?

☒ ¿Datos dentro de límites lógicos y consistentes?

☒ ¿Tratamiento adecuado de valores nulos realizado?

☒ ¿Variables homologadas según diccionario?

☒ ¿Texto sin espacios extra ni errores de digitación?

☒ ¿Formatos de fecha y números validados?

4.Actividad

Actividad

En el archivo de excel de ejemplo vamos a revisar cómo podemos realizar un diagnóstico de la calidad de los datos hacerlo con gemini y con el complemento

<https://app.tango.us/app/workflow/Diagn-stico-de-datos-con-Gemini-901fc2ebe55b4ecda728b91d6315b00e>

<https://app.tango.us/app/workflow/Analizar-archivo-con-Gemini-en-Google-Sheets-y-Gemini-e25451d7216e4aafa0e75c71709b9a01>

Asignación

Leer guía de limpieza de datos con Power Query para la próxima sesión



Asignación opcional

Aplica lo aprendido revisando con Gemini y/o el complemento de IA a un conjunto de datos que tengas



Evaluación de la Sesión



<https://docs.google.com/forms/d/e/1FAIpQLSefDWVWtRfqMuFD1mJMU3xsePBLH3zG-sga1nLQpqBigGBwyA/viewform?usp=dialog>

Fin Sesión 3



La *transformación digital*
es parte de nuestro

Plan
Estratégico
Institucional

23
27



Conoce más ↗