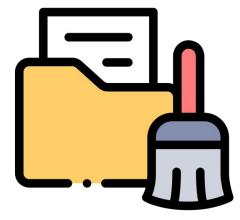
# Módulo 7 La importancia de la Integridad. Datos impecables.





#### **AGENDA DE LA CLASE**

- Integridad de los Datos
- O Datos Insuficientes
- C Limpieza de datos
- Herramientas para la limpieza de datos en hojas de cálculos



# 01 - Integridad de los Datos

```
La integridad de los datos consiste en la:
exactitud,
integridad,
consistencia y
confiabilidad
```

de los datos a lo largo de su **ciclo de vida**.





Si la integridad de los datos está comprometida, un análisis entero puede estar mal hecho porque los datos con los que trabajabas eran erróneos desde el principio.

# 01 - Integridad de los Datos

Formas de comprometer la integridad:

**Al replicar los datos**: Al archivar en distintas locaciones pueden perder **sincronización**.





Al transferir los datos: Si el proceso se interrumpe, los datos quedan incompletos.





Al manipular los datos: En el proceso de hacer los datos más eficientes, un error puede comprometer la integridad.



**Otros**: Errores humanos, Virus, Malware, Hackeos.









#### 02 - Datos insuficientes

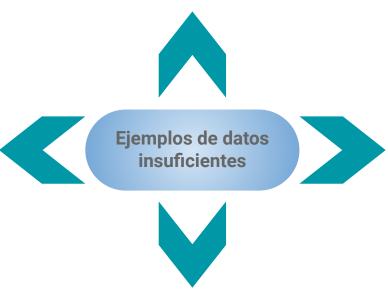
#### Datos limitados geográficamente

Solución: buscar set de datos globales

# Datos se actualizan constantemente.

#### Solución:

- Esperar por más datos.
- Realizar estudios preliminares para luego ser ampliados.



Datos desactualizados

Solución: buscar set de datos más actualizados

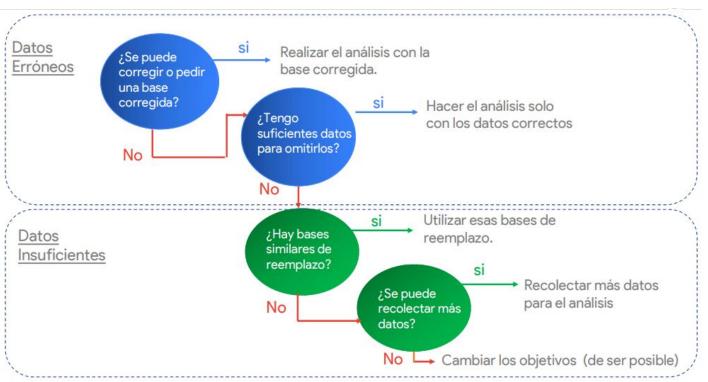
# Datos de **un solo origen o fuente**.

#### Solución:

- Buscar set de datos externos para complementar.
- Identificar tendencias generales con lo disponible.
- Buscar datos de reemplazo

#### 02 - Datos insuficientes

¿Qué hacemos cuando los datos son erróneos o insuficientes?



## 03 - Limpieza de datos

Tipos de datos

sucios

Los **datos sucios** son datos **incompletos**, **incorrectos** o **irrelevantes** respecto del problema que intentas resolver.

Duplicados: Registros de datos que aparecen más de una vez.

Por ejemplo: Por ingreso manual o mala migración de datos.

Desactualizados: Datos viejos que deberían reemplazarse por información más actualizada y exacta.

Por ejemplo: cuando las personas cambian de función o empresa, o software y sistemas que quedan obsoletos.

**Incompletos**: **Datos que carecen de campos importantes**.

Por ejemplo: ocurre con la recopilación de datos inadecuados o ingreso de datos incorrectos.

**Incorrectos o inexactos**: **Datos que son completos pero inexactos**.

Por ejemplo: ocurre por error humano durante el ingreso de datos, información falsa o datos ficticios.

**Inconsistentes**: Datos que usan diferentes formatos para representar la misma cosa.

Por ejemplo: ocurre por errores generados durante la transferencia de datos.

## 03 - Limpieza de datos

Pautas a tener en cuenta:



**6**. Remover formatos para unificar

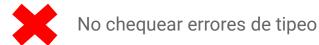
**5**. Corregir errores de tipeo y puntuación

- 4. Remover espacios extra
- 3. Remover datos irrelevantes
- 2. Buscar duplicados y eliminarlos
- 1. Hacer una copia del conjunto de datos



## 03 - Limpieza de datos

## Errores comunes de la limpieza de datos:





Perder el seguimiento de los objetivos de negocio



No documentar errores



No tener en cuenta los tiempos de limpieza en el cronograma de trabajo



Mirar solo un conjunto de datos y no la película completa



No solucionar la fuente del error



No hacer backup previo antes de la limpieza



Pasar por alto valores faltantes o nulos





# 04 - Herramientas para la limpieza de datos en hojas de cálculos

#### **FORMATO CONDICIONAL**

Se trata de una herramienta que cambia la forma en que se ven las celdas cuando los valores cumplen ciertas condiciones.

#### **QUITAR DUPLICADOS**

Se trata de una herramienta que permite quitar duplicados de una tabla o conjunto de celdas.

#### **FORMATO FECHA**

Hay distintos formatos de fechas para seleccionar.

#### **MANIPULEO DE TEXTO**

Una **cadena de texto** es un grupo de caracteres dentro de una celda.

Una característica importante de una cadena de texto es su **longitud**, que es el número de caracteres en ella.

Una **subcadena** es un subconjunto más pequeño de una cadena de texto.



# ¿PREGUNTAS?



# ¡Muchas gracias!



