2. Limpieza de datos y preprocesamiento

2.1 Comprensión de los datos limpios

Los datos limpios se refieren a conjuntos de datos bien organizados, fáciles de analizar y estructurados de acuerdo a reglas específicas que facilitan un análisis preciso y evita errores.

Los pilares de los datos limpios

- **Cada variable forma una columna**
    - Una variable representa una característica o atributo específico, como el color de un auto o la edad de una persona, y cada una debe ocupar su propia columna dentro de un conjunto de datos.

Name	Age	Sex	Height
Ocean	11	F	4'3"
Donald	14	M	4'5"
Charles	31	M	6'1"
Betsy	31	F	5'5"

- **Cada observación forma una fila**
    - Una observación es un único registro o instancia de datos que combina diferentes variables. En un conjunto de datos, cada fila debe representar una única observación.

Name	Join Date	Order Date
Raheem	01/20/2020	9/14/2021
Fatima	05/30/2019	10/6/2021
Carla	12/11/2022	1/29/2022
Denis	04/05/2021	3/4/2020

- **Cada tipo de unidad de observación forma una tabla**
  - Diferentes tipos de datos (p. ej. "orders" (pedidos) y "customers" (clientes)) deben segregarse en sus propias tablas para mantener claridad y evitar la duplicidad.

Order ID	Order Date	Customer ID
1	9/14/2021	1
2	10/6/2021	2
3	1/29/2022	3
4	3/4/2020	4

2.2 Cómo limpiar datos importados

Cómo dividir columnas mediante un separador

La función "Dividir texto en columnas" de Google Sheets te permite dividir una columna en varias, utilizando un carácter específico como el separador. Esto es especialmente útil para separar strings de datos complejos en componentes más manejables y analizables.

Antes de dividir:

User_ID	Contact_info	Registration_date
U001	"JohnDoe|john.doe@example.com"	2022-01-15
U002	"JaneSmith|jane.smith@example.com"	2022-03-22
U003	"AlexJohnson|alex.johnson@example.net"	2022-07-05

Después de dividir:

User_ID	Name	Email	Registration_date
U001	JohnDoe	john.doe@example.com	2022-01-15
U002	JaneSmith	jane.smith@example.com	2022-03-22
U003	AlexJohnson	alex.johnson@example.net	2022-07-05

Cómo eliminar filas duplicadas

Las filas duplicadas pueden distorsionar el análisis de datos. "Eliminar duplicados" se usa para ayudarte a identificar y eliminar estas entradas redundantes con base en columnas seleccionadas que identifican cada renglón de forma única.

Antes de eliminar duplicados:

TransactionID	Date and Time	Amount
TX123	2023-04-01T10:00:00.000Z	$150
TX123	2023-04-01T10:00:00.000Z	$150
TX456	2023-04-02T10:00:00.000Z	$200
TX456	2023-04-02T10:00:00.000Z	$200

Después de eliminar duplicados:

TransactionID	Date and Time	Amount
TX123	2023-04-01T10:00:00.000Z	$150
TX456	2023-04-02T10:00:00.000Z	$200

Cómo buscar y reemplazar texto

"Buscar y reemplazar es una herramienta para hacer cantidades grandes de modificaciones en tu conjunto de datos. Ya sea corregir errores ortográficos comunes o actualizar información en varias entradas, esta función agiliza el proceso.

Antes de Buscar y reemplazar:

Product ID	Product Name	Price
P001	Grahpical Card	$299
P002	Grahpical Card	$299
P003	Grahpical Card	$299
P004	Grahpical Card	$299

Después de Buscar y reemplazar:

Product ID	Product Name	Price
P001	Graphical Card	$299

P002	Graphical Card	$299
P003	Graphical Card	$299
P004	Graphical Card	$299

2.3 Funciones de fecha y hora

En Google Sheets, los valores fecha/hora son numéricos y comienzan desde el 12/30/1899. Esta representación numérica permite manipular fechas y horas fácilmente, así como sumar, restar y extraer componentes específicos como el mes o el año.

Funciones de fecha y hora

- **Cómo extraer los componentes de fecha/hora**:
  - `YEAR(date)`: devuelve el componente de año de una fecha. Para `1/25/2021`, devuelve `2021`.
  - `MONTH(date)`: recupera el componente de mes de una fecha. Para `1/25/2021`, arroja `1`.
  - `DAY(date)`: extrae el componente de día de una fecha. Para `1/25/2021`, el resultado es `25`.
  - `HOUR(date)`, `MINUTE(date)`, `SECOND(date)`: devuelve la hora, minuto y segundo de un valor de tiempo. Para `9:35:32 AM`, la función devuelve `9`, `35` y `32`, respectivamente.
  - `WEEKNUM(date)`: proporciona el número de semanas del año para una fecha dada. Por ejemplo, `WEEKNUM("1/25/2021")` podría devolver `4`, indicando la cuarta semana del año.
  - `WEEKDAY(date)`: devuelve el día de la semana como un número (Domingo = 1, Sábado = 7) para una fecha especificada.
- **Cómo convertir entre texto y fecha**:
  - `DATEVALUE("date_string")`: convierte un string que representa una fecha en un valor de datos numérico que Google Sheets reconoce.

- `DATE(year, month, day)`: crea un valor de fecha a partir de los componentes año, mes y día. Es útil para volver a armar fechas a partir de campos de datos independientes.

- **Rangos de fechas y horas**:
  - `DATEDIF(start_date, end_date, "unit")`: calcula la diferencia entre dos fechas. `"unit"` puede ser `"Y"`, `"M"`, `"D"`, `"MD"`, `"YM"`, o `"YD"`, dedependiendo de si estás midiendo en años, meses, días, etc.
  - `TODAY()`: devuelve la fecha actual.
  - `NOW()`: proporciona la fecha y hora actuales.

2.4 Validación de datos

La validación de datos es un paso importante en la preparación de datos, asegurándonos que todas las entradas en un conjunto de datos satisfacen criterios específicos y, por lo tanto, conservando la integridad y precisión de los datos.

Cómo personalizar los validadores

- **Selección del rango de celdas**
  - Especifica en qué celdas de la hoja de cálculo aplican las reglas de validación.
  - Normalmente, involucra seleccionar columnas completas, excluyendo los encabezados (p. ej., `A2:A` para la primera columna) para aplicar las reglas de manera uniforme.
- **Cómo establecer criterios de validación**
  - Al especificar las condiciones que cada valor de celda debe satisfacer, estos criterios determinan que datos se consideran válidos dentro del rango seleccionado.
  - **Tipos de criterio**:
    - **Listas desplegables**: restringe los valores de las celdas a una lista predefinida de opciones válidas.

- **Condiciones de texto**: valida las entradas de texto con base en el contenido, como el comprobar correos electrónicos o URL válidas.
- **Condiciones de fechas**: asegura que las entradas de fecha caigan dentro de un rango específico o relativo a la fecha actual.
- **Condiciones numéricas**: restringe a los números para que caigan dentro de ciertos límites o coincidan con criterios numéricos específicos.
- **Fórmulas personalizadas**: utiliza las fórmulas de Google Sheets para crear condiciones de validación personalizadas y complejas que devuelvan `TRUE` o `FALSE`.

Cómo implementar validadores personalizados

1. **Navega hacia Validación de datos**: selecciona "Datos > Validación de datos" desde el menú.
2. **Escoge el rango de celdas**: ingresa el rango al que quieres aplicarle la validación en la casilla "Aplicar a un rango".
3. **Selecciona el tipo de criterio**: elige los criterios de validación adecuados del menú desplegable o ingresa una fórmula personalizada.
4. **Configura ajustes adicionales**: de forma opcional, ajusta configuraciones avanzadas como los mensajes de entrada o el tratamiento de datos no válidos (p. ej., mostrar advertencia o rechazar entrada).