Navegación

Ir a la sección:

- Exploración Inicial
- Limpieza y Validación
- Indicadores y Documentación
- EDA Avanzado & Dashboards
- Modelado de Machine Learning

🏥 Análisis y Calidad de Datos de Pacientes de Hospital

Esta aplicación realiza un análisis exhaustivo de la calidad de los datos de pacientes, seguido de procesos de limpieza, validación, generación de KPIs, EDA avanzado y un modelo de Machine Learning.

1. 📉 Análisis de Calidad de Datos (Exploración)

Identificación de los principales problemas de calidad en la tabla de pacientes.

1.1. Vista Previa de Datos Originales

	id_paciente	nombre	fecha_nacimiento	edad	sexo	email
0	1	Claudia Torres	1954-01-08	None	Female	user1@example.com
1	2	Carlos Gómez	1965-01-01	58	Female	None
2	3	Carlos Gómez	2009-03-08	16	None	user3@example.com
3	4	Andrea López	1951-11-18	47	F	user4@example.com
4	5	Juan Gómez	1961-09-05	81	Female	user5@example.com

1.2. Información General y Tipos de Datos

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5010 entries, 0 to 5009
Data columns (total 8 columns):
 # Column Non-Null Count Dtype
--- ------ -------------- -----
 0 id_paciente 5010 non-null int64
 1 nombre 5010 non-null object
 2 fecha_nacimiento 5010 non-null object
 3 edad 3363 non-null float64
 4 sexo 3987 non-null object
 5 email 2504 non-null object
 6 telefono 3342 non-null object
 7 ciudad 4183 non-null object
dtypes: float64(1), int64(1), object(6)
memory usage: 313.3+ KB

1.3. Valores Faltantes (Nulos)

	Valores Faltantes	Porcentaje (%)
email	2506	
telefono	1668	
edad	1647	
sexo	1023	
ciudad	827	

Observaciones Iniciales sobre Valores Faltantes:

- `edad` : Muestra `null` en el JSON para algunos registros. Esto es un problema, ya que la edad es crucial y puede calcularse a partir de la fecha de nacimiento.
- `fecha_nacimiento` : Aunque no hay nulos directos, es importante verificar el formato y la validez de la fecha.

1.4. Inconsistencias y Formatos

Columna `sexo`

sexo	count
Male	
None	
F	
Female	
M	

Problema: Inconsistencia en la capitalización o variaciones en la columna `sexo` (ej., 'Female' vs 'female', 'F' vs 'f', 'M' vs 'm', u otros valores inesperados).

Columna `fecha_nacimiento`

Problema: Se encontraron **4** fechas de nacimiento con formato inválido.

	id_paciente	nombre	fecha_nacimiento	edad	sexo	email
56	57	Andrea López	02 de nov de 1977	None	M	user57@example.com
64	65	Carlos Pérez	22 de oct de 2002	None	None	user65@example.com
84	85	Juan Torres	1959-06-33	36	M	None
94	95	Claudia Pérez	14 de diciembre de 2007	87	F	None

Columna `email`

Problema: Se encontraron **2506** correos electrónicos con formato potencialmente inválido.

	id_paciente	nombre	fecha_nacimiento	edad	sexo	email	tel
1	2	Carlos Gómez	1965-01-01	58	Female	None	No
9	10	Juan López	1961-04-28	None	M	None	34
21	22	Juan Pérez	1977-03-29	34	None	None	30
22	23	Andrea Pérez	2002-02-14	23	None	None	38
23	24	Andrea Torres	1973-01-29	52	Female	None	30

Columna `telefono`

Problema: Se encontraron **1668** números de teléfono con caracteres no numéricos o que no se convierten a un formato numérico válido.

	id_paciente	nombre	fecha_nacimiento	edad	sexo	email
1	2	Carlos Gómez	1965-01-01	58	Female	None
3	4	Andrea López	1951-11-18	47	F	user4@example.com
4	5	Juan Gómez	1961-09-05	81	Female	user5@example.com
5	6	María López	1966-10-26	59	Male	user6@example.com
13	14	María López	1982-02-07	40	M	user14@example.com

Resumen de Problemas de Calidad (Pacientes):

1. **Valores Nulos:** Principalmente en la columna `edad`.
2. **Inconsistencias de Formato:**
   - `sexo`: Posibles variaciones en la capitalización (`Female` vs `female`), o abreviaciones (`F` vs `f`).
   - `fecha_nacimiento`: Necesita conversión a tipo `datetime` y manejo de posibles formatos incorrectos.
   - `edad`: Debe ser un valor numérico y consistente con `fecha_nacimiento`. Si es `null`, debe ser calculado.
   - `email`, `telefono`: Requieren validación de formato (aunque el ejemplo dado parece limpio, es buena práctica).
3. **Coherencia de Datos:** `edad` debe ser derivable de `fecha_nacimiento` y ser un número positivo.

Nota: Dado que solo tenemos la tabla 'pacientes' de la URL, el análisis se centra en ella.