Informe Técnico – Prueba de Ingeniero de Datos
Generado: 2025-08-13 / Por: Sebastian Jaimes G

1. Análisis Exploratorio y Hallazgos de Calidad
Tabla: pacientes
- **Cantidad inicial:** 5,010 registros.
- **Principales problemas detectados:**
- 10 duplicados exactos por “id_paciente”.
- Campos con alta proporción de valores nulos:
- “edad” → 1,647 nulos.
- “sexo” → 1,023 nulos.
- “email” → 2,506 nulos.
- “telefono” → 1,668 nulos.
- “ciudad” → 827 nulos.
- Inconsistencias entre “edad” y “fecha_nacimiento” en varios registros.
- Variabilidad de formatos en “sexo” (M, F, masculino, femenino, etc.).

Tabla: citas_medicas
- **Cantidad inicial:** 9,961 registros.
- **Principales problemas detectados:**
- 3,278 nulos en “fecha_cita”.
- 1,673 nulos en “especialidad”.
- 2,033 nulos en “medico”.
- 1,724 nulos en “costo”.
- 2,542 nulos en “estado_cita”.
- 4,807 registros con el mismo “id_paciente” (varias citas por paciente).
- 190 registros huérfanos → “id_paciente” no existe en la tabla pacientes.

2. Estrategia de Limpieza y Supuestos
- **Estandarización de fechas** a formato “YYYY-MM-DD”.
- **Recalculo de edad** en base a “fecha_nacimiento”, corrigiendo inconsistencias.
- **Normalización de “sexo”** a “"M"“ o “"F"“.
- **Eliminación de duplicados**:
- Pacientes: 10 registros eliminados por “id_paciente” repetido.
- Citas: no se detectaron duplicados por “id_cita”.
- **Validación de integridad**:
- Se detectaron 190 citas huérfanas, listadas en “reports/orphan_citas.csv”.
- No se eliminaron de la tabla final de citas para mantener trazabilidad, pero se recomienda revisión.

Supuestos adoptados:
1. Si “edad” estaba vacía pero “fecha_nacimiento” era válida, se recalculó.
2. Si “sexo” no estaba en formato M/F, se intentó mapear según equivalencias conocidas.
3. No se imputaron valores para “email”, “telefono”, “ciudad”, “especialidad”, “medico”, “costo” ni “estado_cita” por falta de reglas de negocio.

3. Indicadores de Calidad Antes y Después

Tabla	Filas iniciales	Filas finales	Nulos totales iniciales	Nulos totales finales	Duplicados iniciales	Duplicados finales
pacientes	5,010	5,000	7,671	6,021	10	0
citas_medicas	9,961	9,961	11,250	11,250	0	0

4. Validaciones Cruzadas

- **Integridad referencial**:
- "id_paciente" en citas debe existir en pacientes.
- Resultado: 190 registros huérfanos.
- **Duplicados**:
- Eliminados duplicados por "id_paciente" en pacientes.
- Confirmado "id_cita" único en citas.

5. Recomendaciones de Mejora

1. **Completar campos clave** ("sexo", "email", "telefono", "ciudad") desde otras fuentes.
2. **Revisar citas huérfanas** para:
- Corregir "id_paciente" erróneos.
- O asignar un paciente válido.
3. **Estandarizar catálogo de especialidades y médicos** para evitar variantes de texto.
4. **Definir reglas de negocio** para imputar datos faltantes en costos y estado de cita.

6. Archivos Entregados

- **Datos**:
- "data/raw/dataset_hospital.json" → Fuente original.
- "data/interim/" → Datos intermedios.
- "data/clean/" → Datos limpios.
- "data/hospital_dataset_clean.xlsx" → Dataset limpio consolidado.
- **Reportes**:
- "reports/exploration_report.md"
- "reports/cleaning_summary.md"
- "reports/orphan_citas.csv"
- "reports/final_report.md" (este documento)
- **Scripts**:
- "src/explore.py"
- "src/clean.py"
- "src/export_excel.py"
- "src/compare.py"

Extras:

Validación de calidad de datos

Se ejecutaron las pruebas incluidas en "tests/test_data_quality.py" usando "pytest".

Todas las pruebas pasaron correctamente, confirmando que:

1. No hay duplicados en los identificadores de pacientes ni citas.
2. Todos los pacientes cuentan con un ID válido.
3. La integridad referencial entre citas y pacientes se cumple.
4. Las columnas requeridas están presentes en ambos datasets.

Resultado: Todas las pruebas se ejecutaron exitosamente.
El detalle se encuentra en “reports/test_results.txt”.

Simulación de migración a Data Warehouse
Se implementó un script “load_to_dw.py” que simula la carga de los datos limpios a una estructura destino tipo Data Warehouse, utilizando SQLite como motor.

Estructura destino:
- **dim_pacientes**: Tabla con la información de pacientes.
- **fact_citas**: Tabla con la información de las citas médicas.