Prueba Técnica XpertGroup 2025

Informe Técnico de Calidad y Gobierno de Datos

Entregable: Notebook, Scripts y Dataset Limpio

Autor: Andrés Arroyave Carmona

1. Análisis Exploratorio y Hallazgos de Calidad

1.1. Descripción General del Dataset

El dataset está compuesto por dos tablas principales:

- **pacientes**: Información demográfica y de contacto de los pacientes.
- **citas_medicas**: Registro de citas, especialidades, médicos, costos y estado.

1.2. Principales Problemas de Calidad Detectados

- **Valores nulos y faltantes** en campos críticos (`id_paciente`, `fecha_nacimiento`, `especialidad`).
- **Duplicados** por `id_paciente` y `id_cita`.
- **Inconsistencias de formato** en correos electrónicos, teléfonos y nombres.
- **Especialidades y médicos fuera del conjunto esperado**.
- **Citas sin paciente válido** (integridad referencial).
- **Citas con fechas inválidas** (anteriores a la fecha de nacimiento).
- **Citas de especialidad no coherente con sexo o edad** (ej. ginecología para hombres, pediatría para adultos).
- **Costos negativos o nulos en citas**.
- **Estados de cita inconsistentes con la presencia de fecha**.

---

2. Validaciones Realizadas y Problemas Detectados

2.1. Validaciones de Formato y Consistencia

- **Email**: Expresión regular estándar.
- **Teléfono**: Formato `XXX-XXX-XXXX` o nulo.
- **Nombres**: Solo letras y espacios.
- **Médicos**: Prefijo obligatorio `Dr.` o `Dra.`.
- **Especialidades**: Solo valores válidos predefinidos.
- **Costos**: Deben ser positivos o nulos.
- **Fechas**: Ninguna cita anterior a la fecha de nacimiento.
- **Sexo vs. Especialidad**: Ginecología solo para mujeres.
- **Edad vs. Especialidad**: Pediatría solo menores de 18 años.
- **Estado de cita vs. fecha**: Citas completadas/canceladas deben tener fecha.

2.2. Validaciones Cruzadas

- **Integridad referencial**: Todas las citas deben tener un paciente válido.
- **Pacientes sin citas**: Se reportan pero no se eliminan.

---

3. Estrategia de Limpieza y Supuestos Adoptados

3.1. Supuestos Adoptados

- No se rellenan valores nulos en campos críticos con datos inventados.
- En campos no críticos, se puede usar un valor genérico ("Sin información").
- Correos y teléfonos inválidos se dejan nulos.
- Solo se aceptan especialidades y médicos dentro del conjunto esperado.
- Citas sin paciente válido se eliminan.
- Citas con fechas incoherentes se eliminan.
- Costos negativos se corrigen a 0.
- Se estandarizan nombres y ciudades a formato título.
- Edad se calcula si es posible, si no, se deja nulo.
- Sexo se estandariza a "Femenino" y "Masculino".

3.2. Decisiones de Limpieza

- **Eliminación de duplicados** por `id_paciente` y `id_cita`.
- **Normalización de formatos** en todos los campos relevantes.
- **Eliminación de registros con integridad referencial rota**.
- **Corrección de valores negativos y formatos inválidos**.
- **Reporte de pacientes sin citas y de valores atípicos**.

---

4. Indicadores de Calidad Antes y Después de la Limpieza

Indicador	Antes de la limpieza	Después de la limpieza
Total de pacientes	(ver notebook)	(ver notebook)
Duplicados por id_paciente	(ver notebook)	(ver notebook)
Nulos en campos críticos	(ver notebook)	(ver notebook)
Total de citas médicas	(ver notebook)	(ver notebook)
Duplicados por id_cita	(ver notebook)	(ver notebook)
Citas sin paciente válido	(ver notebook)	0
Citas con fechas inválidas	(ver notebook)	0
Costos negativos	(ver notebook)	0
Especialidades/médicos inválidos	(ver notebook)	0

Nota: Los valores exactos pueden consultarse en las tablas y gráficos del notebook entregado.

---

5. Reglas de Validación Implementadas

- Validación de formato de email, teléfono y nombres.
- Validación de especialidades y médicos.
- Validación de costos y fechas.
- Validación cruzada de integridad referencial.
- Validación de coherencia entre sexo, edad y especialidad.
- Validación de estado de cita vs. presencia de fecha.

---

6. Recomendaciones de Mejora para la Calidad Futura de los Datos

- Implementar validaciones automáticas y monitoreo continuo en futuras cargas.
- Mantener un diccionario de datos y reglas de validación documentadas.
- Usar frameworks de validación como Great Expectations o pytest.
- Realizar auditorías periódicas y establecer responsables de calidad.
- Automatizar la generación de reportes de calidad y alertas ante anomalías.
- Capacitar al personal en buenas prácticas de captura y gestión de datos.
- Versionar los datasets y mantener trazabilidad de los cambios.
- Definir flujos de aprobación para la actualización de datos críticos.
- Simular migraciones y pruebas de integridad en entornos de staging antes de producción.

---

7. Bonus: Pruebas Automáticas y Simulación de Migración

- Se implementaron pruebas automáticas en el script `pruebas_automaticas.py` para validar la integridad y calidad de los datos.
- Se simuló la migración de los datos limpios a una estructura tipo Data Warehouse, validando la consistencia y referencialidad.

Este informe, junto con el notebook, scripts y datasets limpios, conforma la entrega completa de la prueba técnica.