

Resumen Profesional del Proceso de Limpieza y Análisis de Datos Odontológicos

1. Recepción y Preparación del Dataset


Se recibió un archivo CSV con registros de consultas odontológicas. Se identificaron inconsistencias en los datos, como errores de codificación (#¡REF!) y etiquetas no estandarizadas en la columna rango etario.

2. Limpieza Inicial de Datos

Se reemplazaron los valores #¡REF! por NaN para facilitar la imputación. Se estandarizaron las etiquetas de edad (rango etario) para unificar criterios como:

- "1 - 4 ANOS", "1 - 4 ANIOS" → "1 - 4 AÑOS"
- "20-39 ANIOS" → "20 - 39 AÑOS"
- "< 1 ANIO" → "< 1 AÑO"

3. Imputación de Valores Faltantes

- **Variables numéricas** (como consulta_cantidad) fueron imputadas con la **media**.
- **Variables categóricas** (como sexo, detalle, saps) fueron imputadas con la **moda** (valor más frecuente).
- Se generó un nuevo archivo limpio:
 consultas_odontologicas_imputado.csv

4. Visualización de Datos con Matplotlib

A) Distribución por Rango Etario

- **Gráfico:** Barras verticales que muestran la cantidad total de consultas por grupo de edad.
- **Resultado principal:**
Grupo más atendido: **20 - 39 AÑOS**
Total de consultas: **26,613**
- **Interpretación:** La mayor demanda odontológica se concentra en adultos jóvenes, posiblemente por mayor autonomía, acceso a servicios o necesidades clínicas específicas.

B) Distribución por Sexo

- **Gráfico:** Barras verticales comparando el total de consultas entre hombres y mujeres.
- **Resultado principal:**
Sexo con más consultas: **Femenino (F)**
Total de consultas: **49,959**

- **Interpretación:** Esto puede reflejar mayor conciencia de salud bucal, acceso a servicios o prevalencia de patologías en mujeres.

C) Top 10 Tipos de Consulta

- **Gráfico:** Barras horizontales mostrando los motivos de consulta más frecuentes.
- **Resultado principal:**
Consulta más frecuente: **URGENCIA**
Total de registros: **16,965**
- **Interpretación:** La alta frecuencia de urgencias sugiere deficiencias en prevención o seguimiento odontológico, lo que podría orientar estrategias de salud pública.

D) Top 10 Profesionales con Más Consultas

- **Gráfico:** Barras horizontales que muestran los odontólogos con mayor volumen de atención.
- **Resultado principal:**
Profesional más activo: **DR. SEMPER**
Total, de consultas realizadas: **8,211**
- **Interpretación:** Este dato permite evaluar la carga laboral, eficiencia y distribución de recursos humanos en el servicio odontológico.

Se detectaron inconsistencias con las fechas, por lo cual Identificaremos convertiremos y reemplazaremos, estas inconsistencias

- Se detectaron 10 fechas inválidas.
- Todas fueron corregidas usando la fecha más frecuente: 2020-02-01.

5. Creación de reglas de coherencia y corrección de Incoherencias entre Edad y Tipo de Consulta.

```
# Definir reglas de coherencia

cond_nino = df['rango_etario'].str.contains("1 - 4|5 - 14|< 1", case=False)
cond_adulto = df['rango_etario'].str.contains("15 - 19|20 - 39|40 - 69|>= 70", case=False)

incoherentes_nino = df[cond_nino & df['detalle'].str.contains("ADULTO", case=False)]
incoherentes_adulto = df[cond_adulto & df['detalle'].str.contains("NIÑOS|NINOS", case=False)]

total_incoherencias = len(incoherentes_nino) + len(incoherentes_adulto)
print(f"Incoherencias detectadas: {total_incoherencias}")

# Corrección
df.loc[cond_nino & df['detalle'].str.contains("ADULTO", case=False), 'detalle'] = \
    df['detalle'].str.replace("ADULTO", "PEDIATRICO", case=False)

df.loc[cond_adulto & df['detalle'].str.contains("NIÑOS|NINOS", case=False), 'detalle'] = \
    df['detalle'].str.replace("NIÑOS|NINOS", "ADULTO", case=False, regex=True)

# Confirmar correcciones
incoherencias_restantes = df[
    (cond_nino & df['detalle'].str.contains("ADULTO", case=False)) |
    (cond_adulto & df['detalle'].str.contains("NIÑOS|NINOS", case=False))
]
print(f"Incoherencias restantes después de corrección: {len(incoherencias_restantes)}")
```

6. Realizaremos una serie de validaciones necesarias, para que el sistema detecte errores antes de ser diligenciada e insertados los pacientes:
 - a) Detección de duplicados
Verificar si hay registros repetidos (mismo profesional, fecha, rango etario, detalle).
 - b) Validación de valores extremos o atípicos
Por ejemplo, si hay consultas con cantidades muy altas o negativas.
 - c) Homogeneización de texto en la columna detalle
Unificar mayúsculas/minúsculas, corregir errores ortográficos comunes.
 - d) Validación de turnos y servicios
Hay que asegurar que los valores en turno y servicio estén dentro de un conjunto permitido (ej. "MAÑANA", "TARDE").
 - e) Generación de indicadores por centro de atención (saps)
Total de consultas, promedio por día, por profesional, etc.
 - f) Clasificación de tipos de consulta
Agrupar los detalles en categorías como "Preventiva", "Urgencia", "Restaurativa", etc.

RESULTADO



Validaciones adicionales realizadas:

1. Registros duplicados detectados: 347
2. Registros con valores negativos en consulta_cantidad: 0
3. Registros con valores extremadamente altos (> percentil 99): 222
4. Turnos inválidos detectados: 5
5. Servicios inválidos detectados: 0

6. Clasificación de tipos de consulta:

```

tipo_consulta
OTROS          8559
RESTAURATIVA   7489
URGENCIA        4836
PREVENTIVA     1811
DERIVACION     284
Name: count, dtype: int64

```

7. Indicadores por centro de atención (Top 5):

	total_consultas	promedio_consultas	cantidad_registros
saps			
DR. SEMPER	8211.000000	5.698126	1441
DR. SUSSINI	8028.000000	3.158143	2542
DR. ALBERTO LIFSCHITZ	7715.000000	5.284247	1460
DR MANUEL A. GONZALEZ	6225.793054	4.156070	1498
ITATI	5026.000000	3.881081	1295

6. Exportar archivos con resultados y correcciones aplicadas

⇒ Archivo Excel 'reporte_validaciones_clinica.xlsx' generado con éxito.

Una vez que ya tenemos nuestro análisis limpio y con las validaciones que se creyeron considerables, pasaremos a exportar diferentes archivos .csv, para generar dashboard con power BI.

⇒ Archivos CSV para Power BI generados:

- powerbi_tipo_consulta.csv
- powerbi_consultas_por_saps.csv
- powerbi_consultas_por_sexo.csv
- powerbi_consultas_por_edad.csv
- powerbi_valores_altos.csv

7. Herramientas Utilizadas

- **Python + Pandas:** para limpieza, imputación y agrupación de datos.
- **Matplotlib:** para visualizaciones estáticas y profesionales.
- **Plotly (previo intento):** se intentó usar para gráficos interactivos, pero se descartó por problemas técnicos.
- **POWER BI:** se generaron distintos archivos CSV con la información relevante para generar nuestro Dashboard, para ayudar con la toma de decisiones y visualización de indicadores.

8. Resultado Final

Se obtuvo un dataset limpio, imputado y visualizado, listo para:

- Presentaciones ejecutivas.
- Análisis estadístico.
- Desarrollo de dashboards o modelos predictivos.

REALIZADO POR:

Leonardo Torres Velilla

Ingeniero de Sistemas – Especialista en BI Analytics.