Proyecto Resuelto (Nivel Básico) – Calidad y Visualización de Datos Clínicos

Formación: IFCD104 – IA y Big Data aplicado al ámbito biosanitario

Fecha de generación: 2025-09-29 12:17

# 1. Objetivo

Cargar, perfilar y limpiar un dataset clínico simulado; generar visualizaciones y un informe de calidad reproducible.

# 2. Dataset y perfilado inicial

Filas (crudo): 205 | Columnas: 10 | Valores perdidos totales: 33 | Duplicados: 5

Valores perdidos por columna:

• patient\_id: 0

• sex: 0

• age: 0

• sbp\_mmHg: 0

• dbp\_mmHg: 0

• heart\_rate\_bpm: 0

• temperature\_c: 17

• spo2\_pct: 16

• glucose\_mgdl: 0

• admission\_date: 0

# 3. Reglas de validación (rangos plausibles)

• age: 0–120

• sbp\_mmHg: 70–220

• dbp\_mmHg: 40–130

• heart\_rate\_bpm: 30–200

• temperature\_c: 34.0–41.0

• spo2\_pct: 80.0–100.0

• glucose\_mgdl: 50–500

# 4. Estrategia de limpieza

• Imputación mediana en temperature\_c y spo2\_pct

• Normalización mediante recorte (clip) a rangos plausibles

• Corrección de edades negativas → imputación por mediana

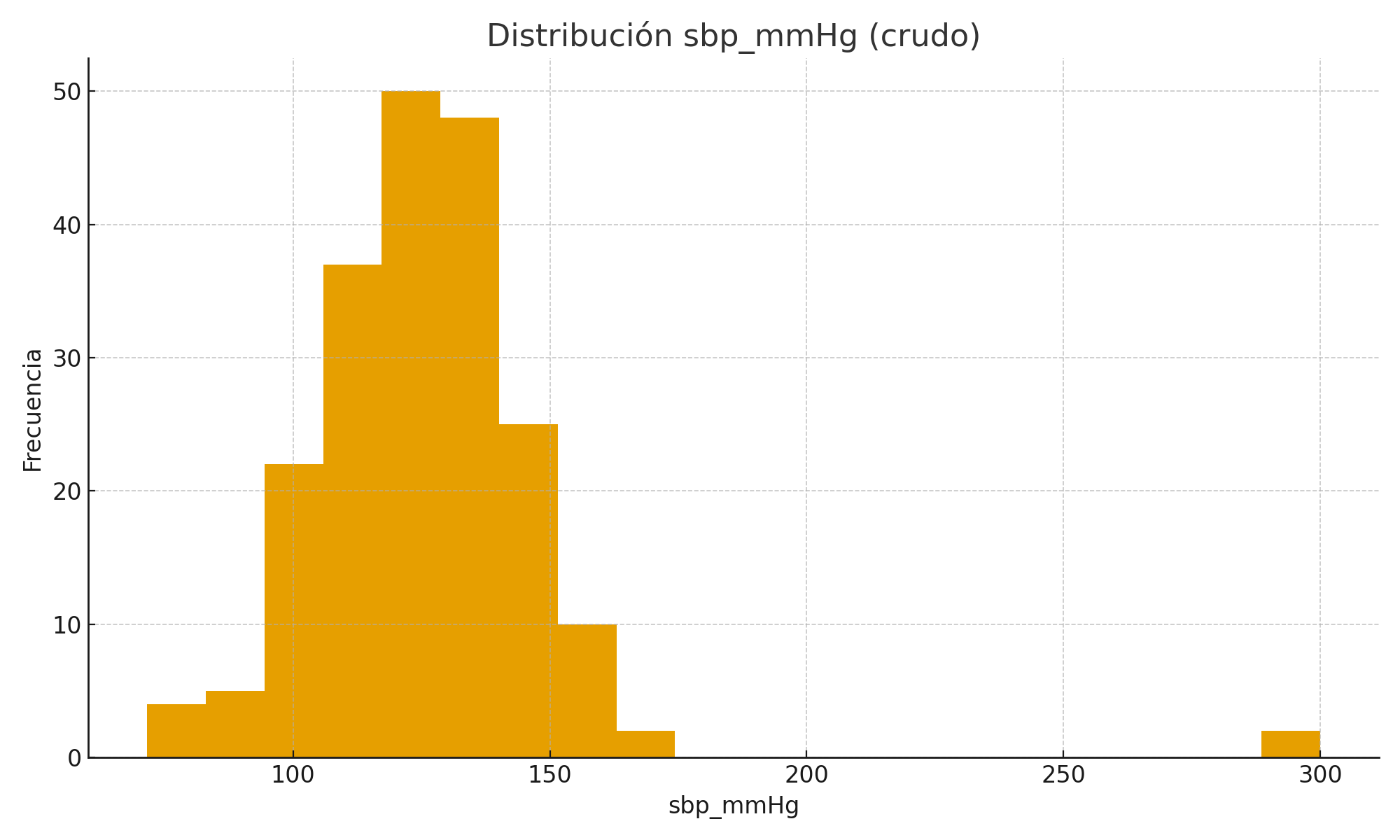
• Eliminación de duplicados exactos

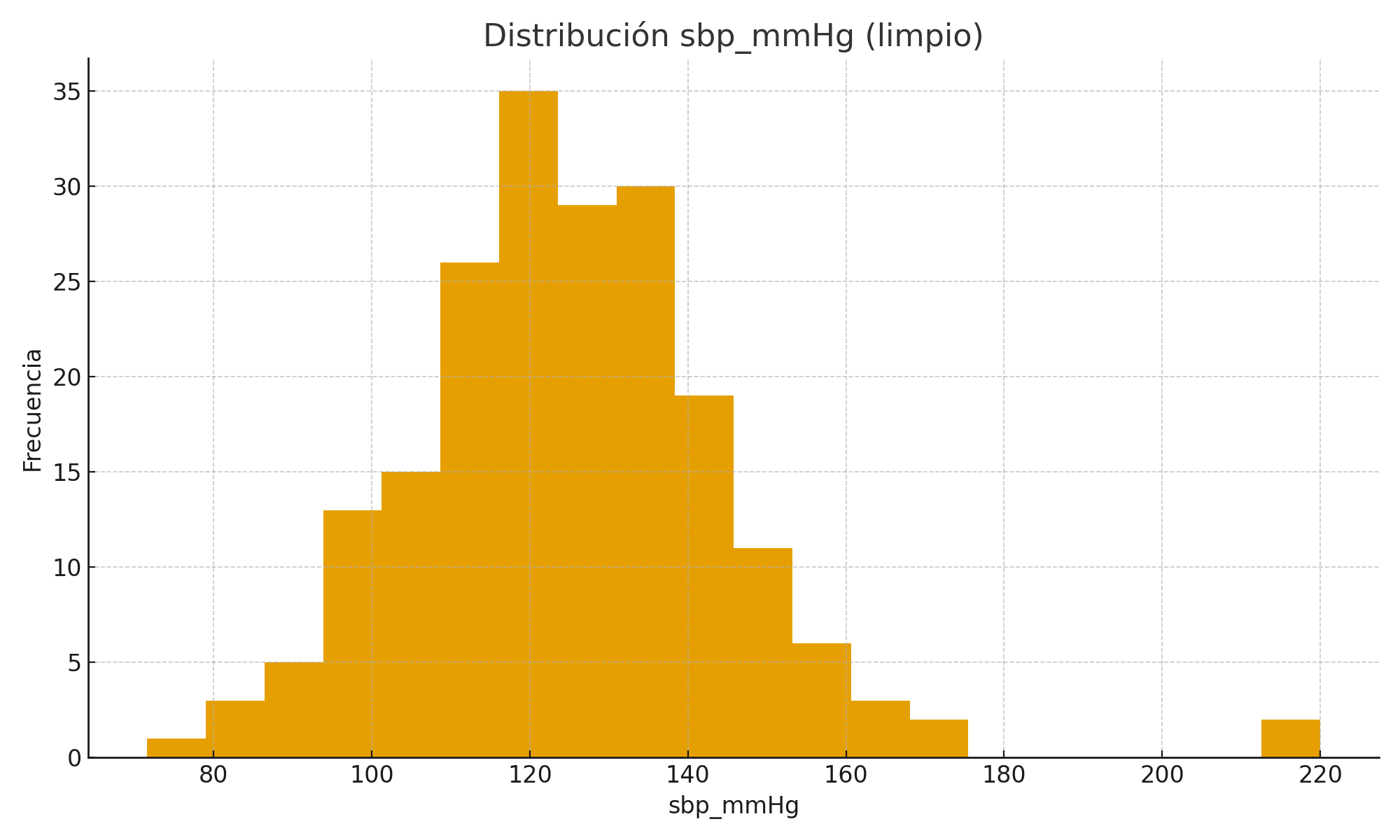
# 5. Resultados tras limpieza

Filas (limpio): 200 | Valores perdidos totales: 0 | Duplicados: 0

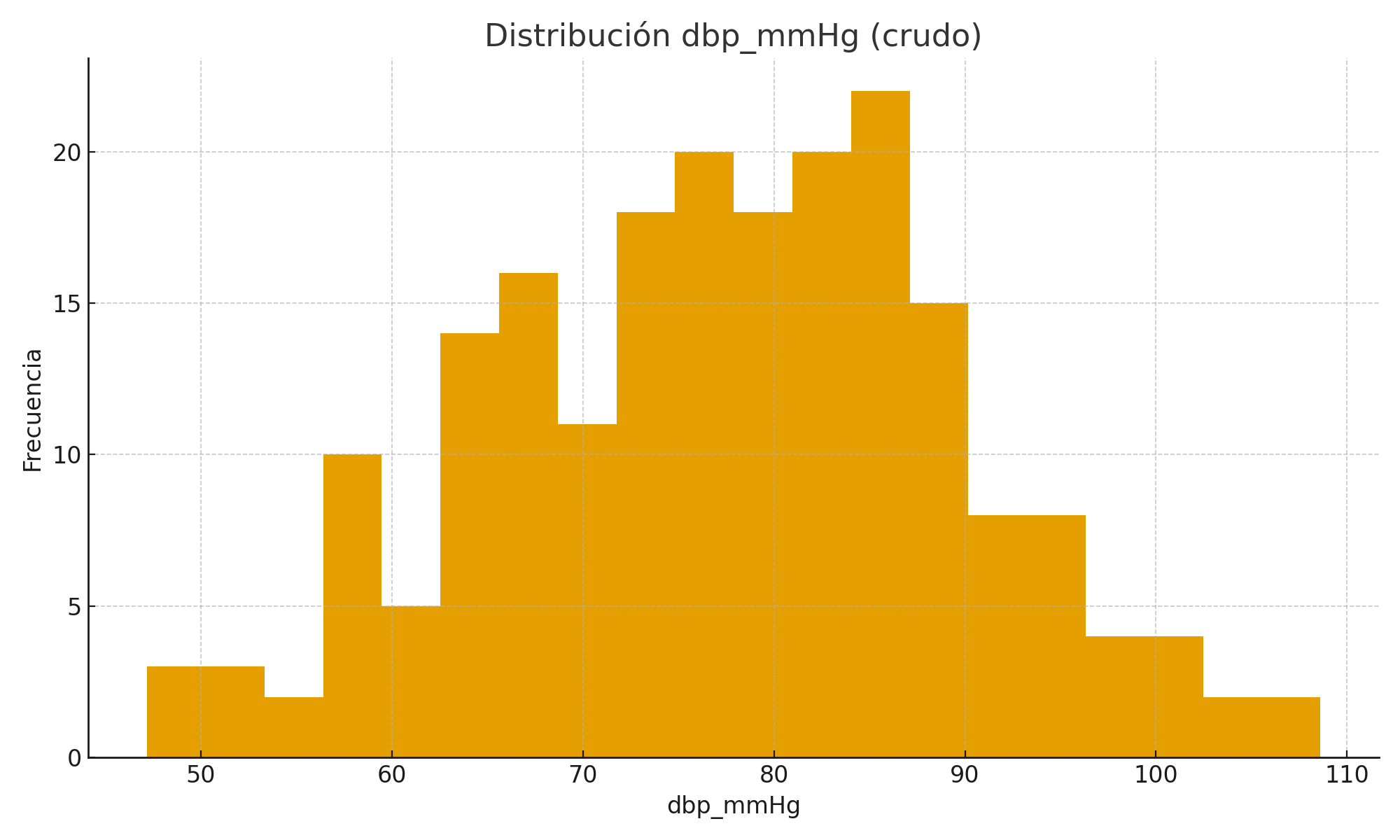
# 6. Visualizaciones (antes / después)

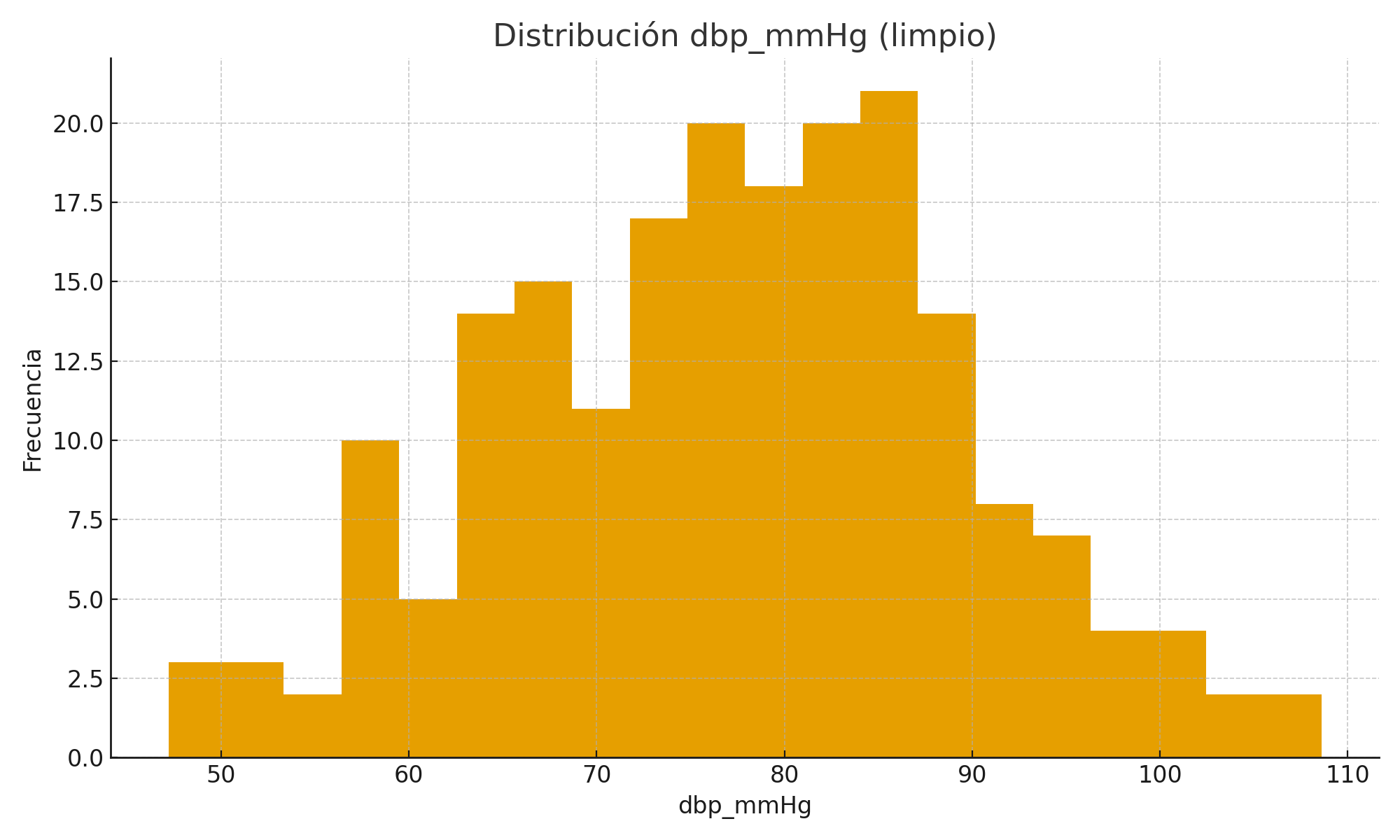
sbp\_mmHg



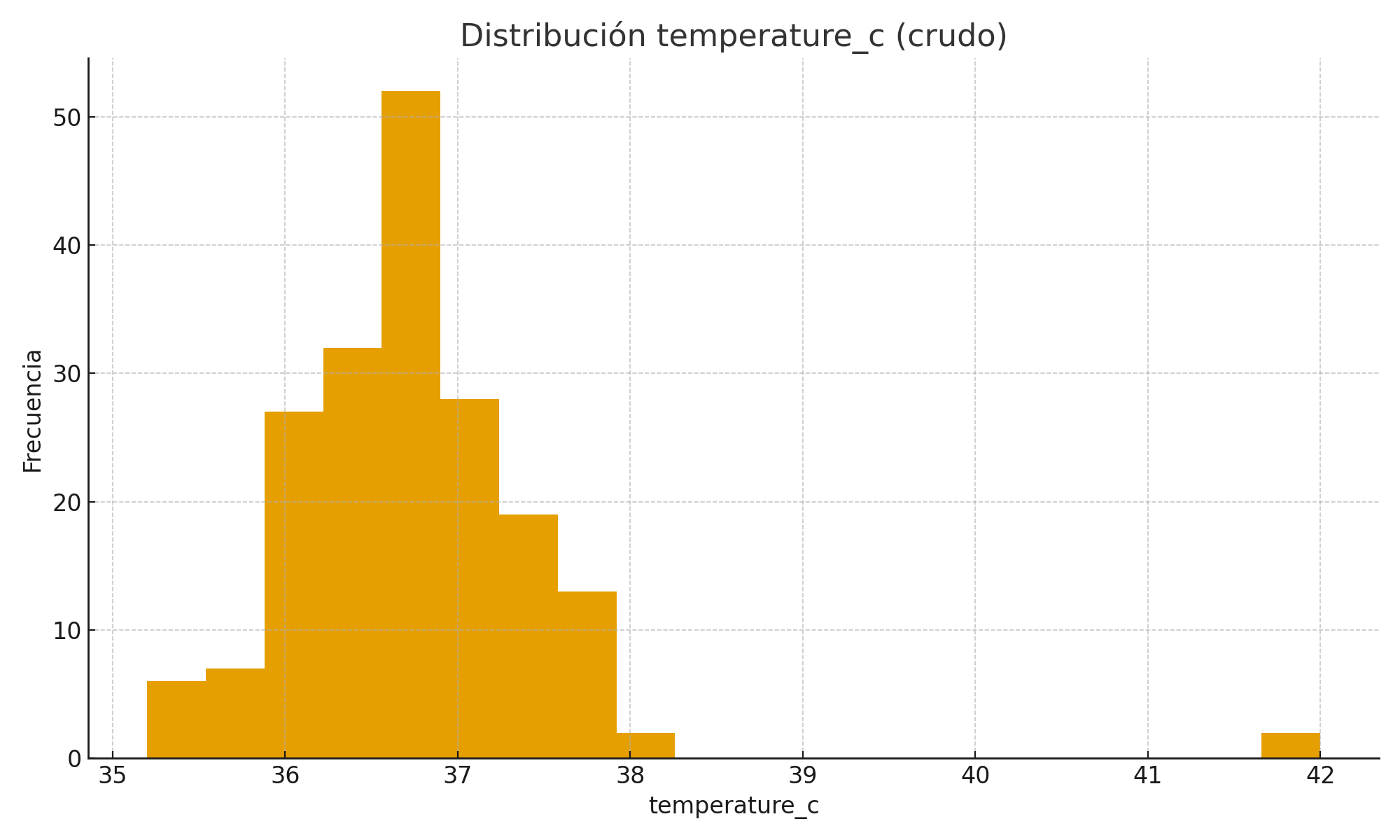


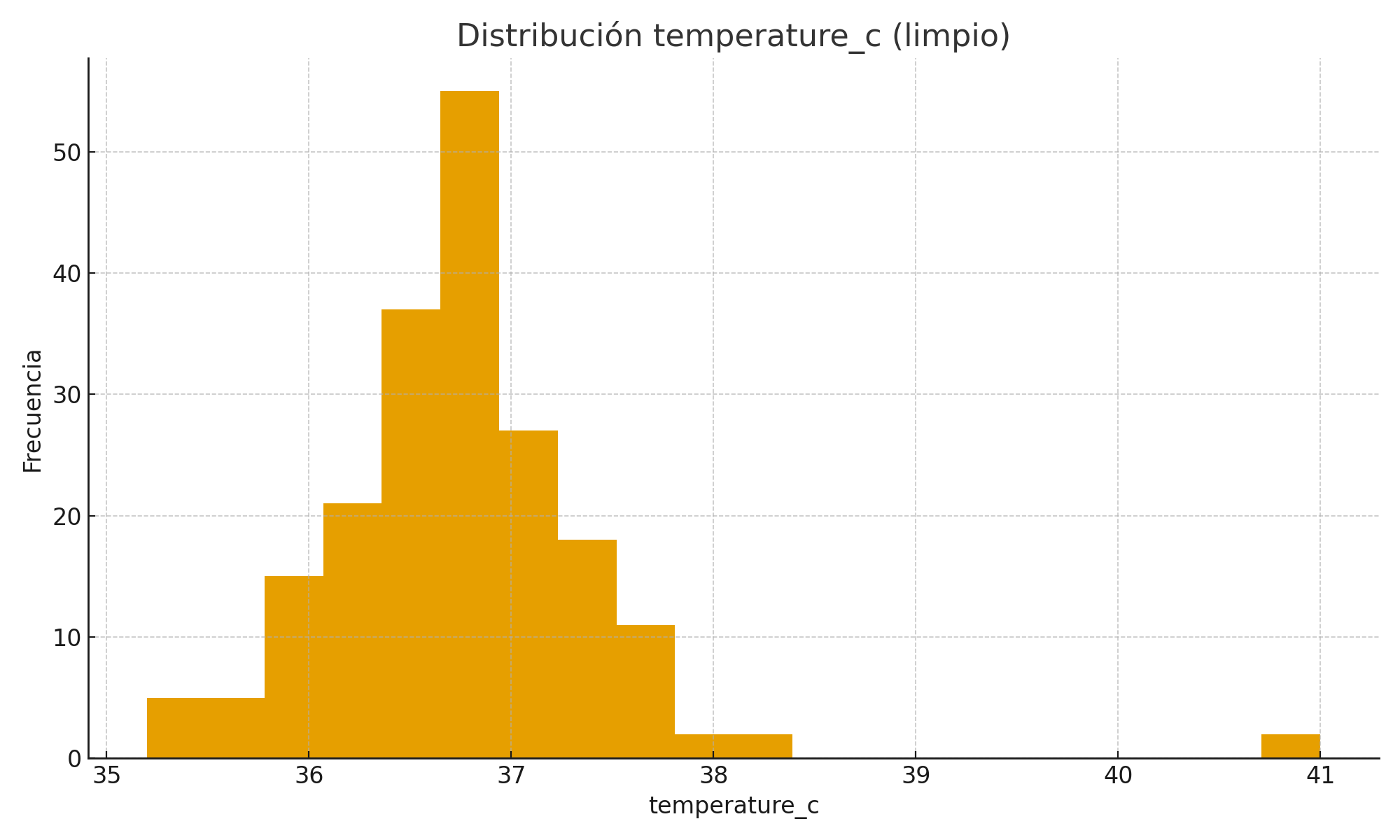
dbp\_mmHg



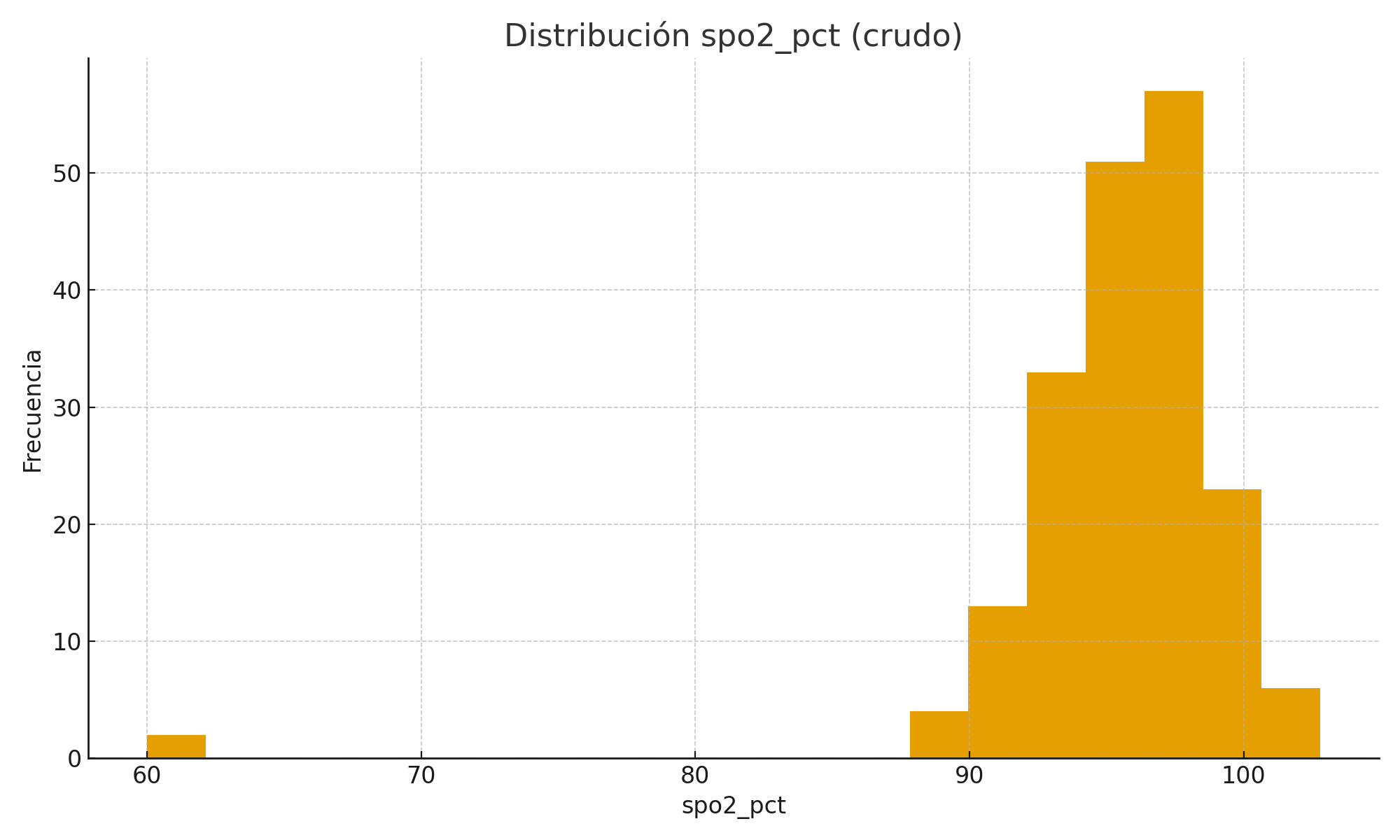


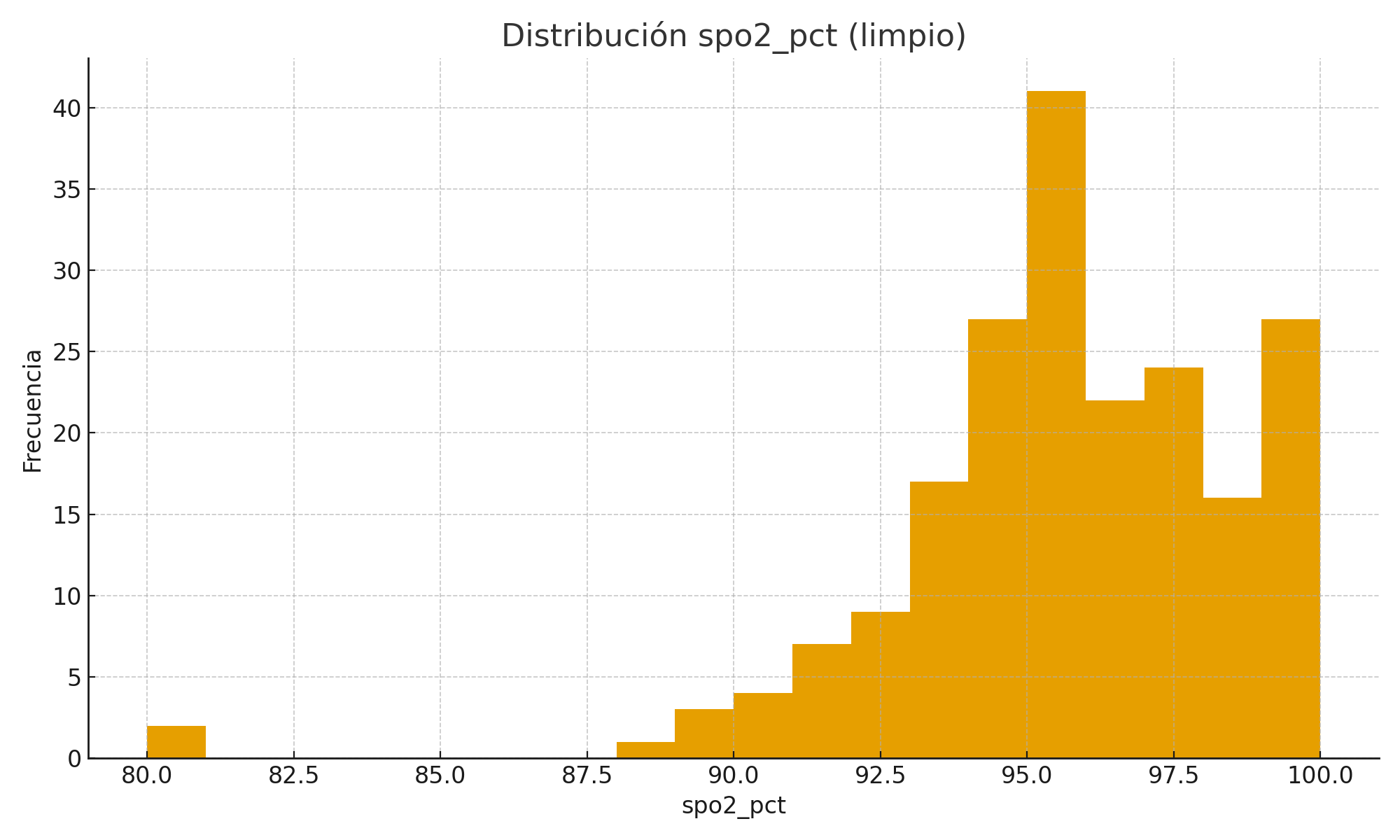
temperature\_c



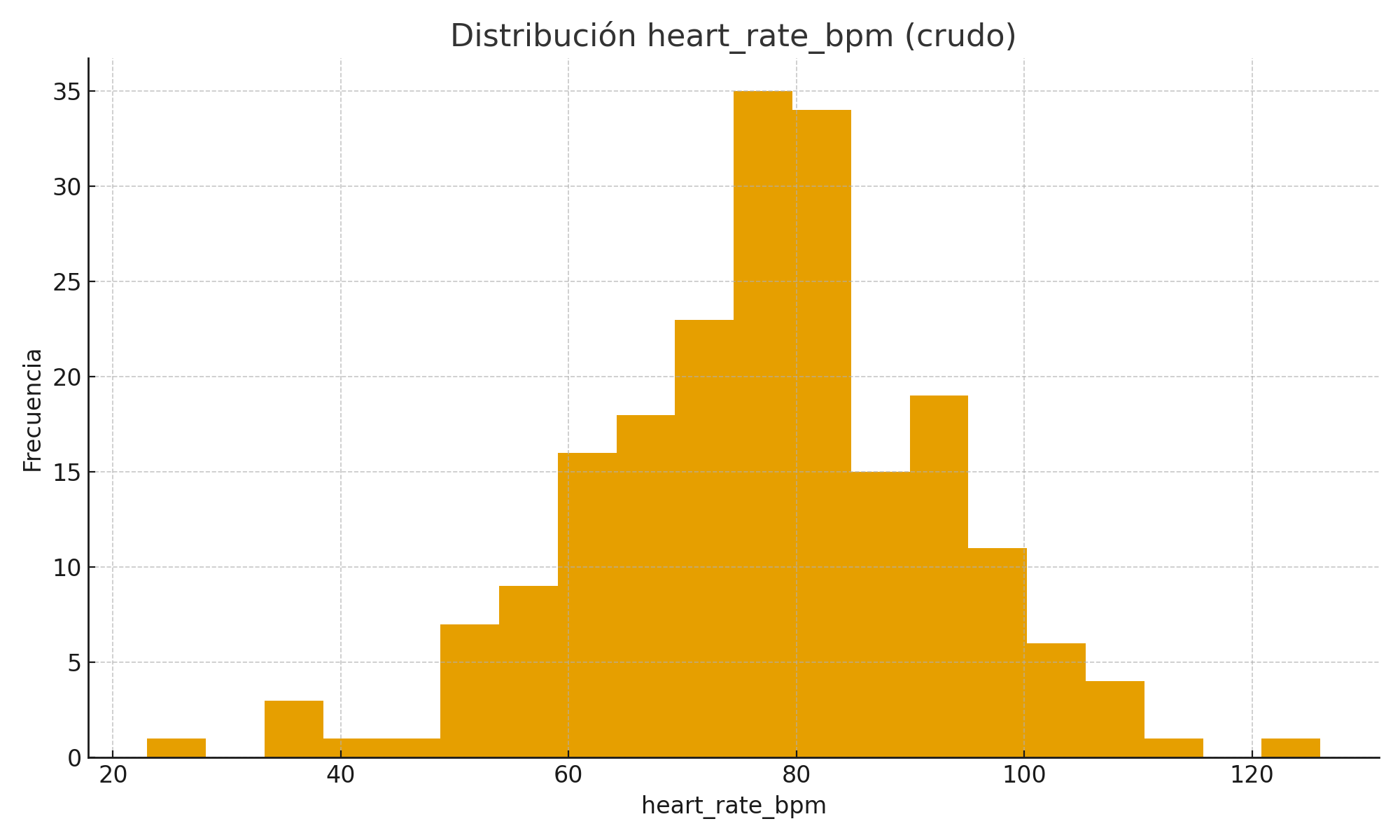


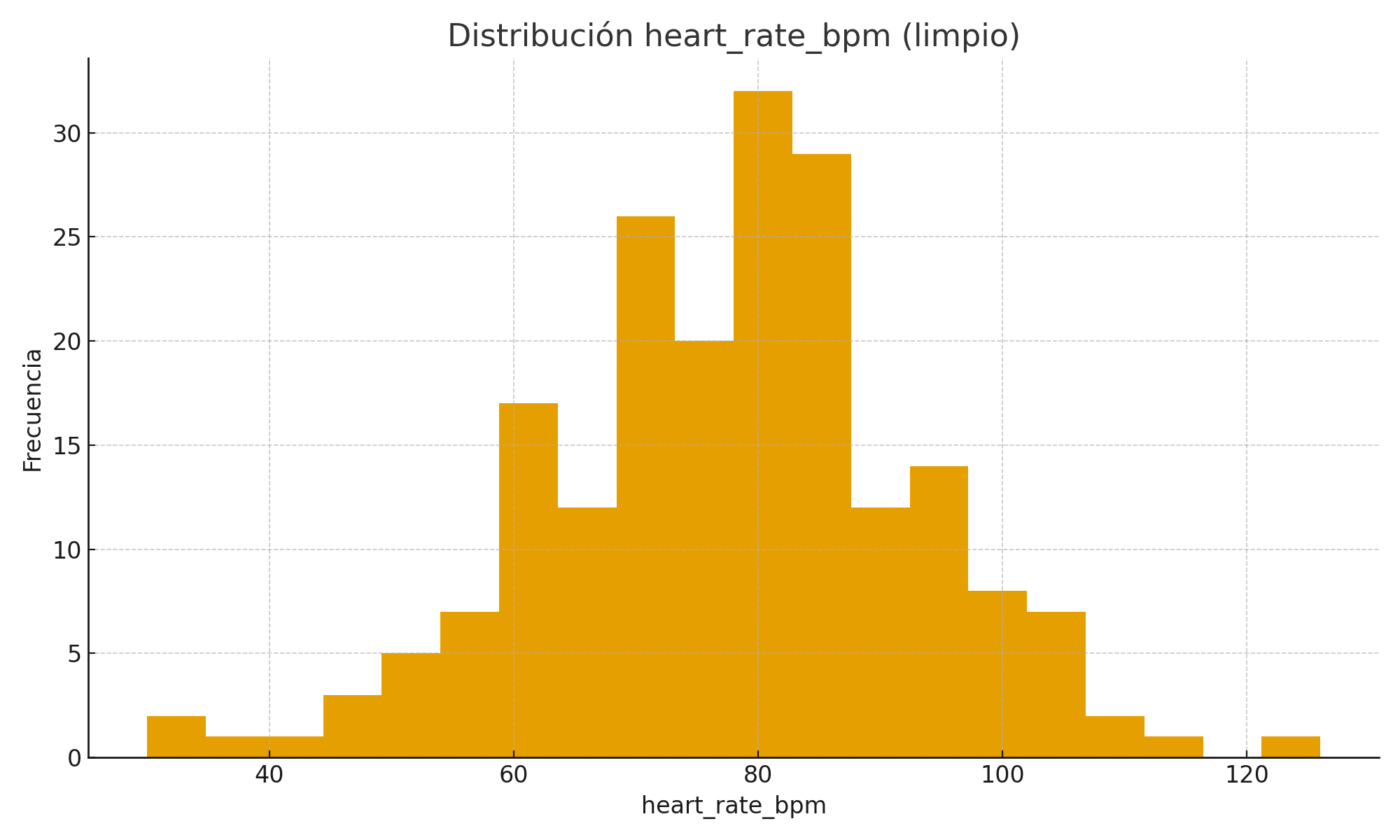
spo2\_pct





heart\_rate\_bpm





# 7. Conclusiones y recomendaciones

• La limpieza ligera reduce valores extremos no plausibles y elimina duplicados.

• La imputación simple permite análisis descriptivo, pero se recomienda evaluar su impacto en modelos clínicos.

• Mantener un diccionario de datos y registrar transformaciones mejora la trazabilidad.

# 8. Entregables

• clinical\_data\_raw.csv (datos crudos)

• clinical\_data\_clean.csv (datos limpios)

• Carpeta figures/ con histogramas antes/después

• Este informe DOCX