UNIVERSIDAD NACIONAL DE SAN AGUSTÍN DE AREQUIPA

ESCUELA PROFESIONAL DE CIENCIA DE LA COMPUTACIÓN

CIENCIA DE DATOS

INFORME DEL ANÁLISIS EXPLORATORIO DE DATOS

Alumno :

Roy Angel Choquehuanca Anconeyra

AREQUIPA – PERÚ

2024

Análisis Exploratorio de datos sobre personas con enfermedades cardiovasculares internados en UCI

Plan de Análisis

1. Motivación y Contexto

En los últimos años, el análisis de registros electrónicos de salud (EHR) se ha vuelto una herramienta clave en la investigación y monitoreo de pacientes críticos, especialmente en Unidades de Cuidados Intensivos (UCI) cardiovasculares. Estos registros contienen grandes volúmenes de datos heterogéneos que, si se analizan correctamente, pueden ofrecer una mejor comprensión del estado de los pacientes y apoyar la toma de decisiones clínicas. Sin embargo, uno de los desafíos principales es presentar esta información de forma comprensible y útil para los médicos, quienes requieren visualizaciones claras, rápidas e interactivas para extraer conclusiones relevantes.

Con la llegada de entornos computacionales interactivos como los notebooks (por ejemplo, Jupyter), se ha abierto la posibilidad de integrar código, resultados y visualizaciones en un mismo espacio de trabajo [1]. Estas herramientas han transformado el análisis exploratorio de datos clínicos, permitiendo mayor flexibilidad, trazabilidad y replicabilidad. Sin embargo, siguen existiendo barreras, como la necesidad de conocimientos técnicos avanzados para generar visualizaciones útiles, lo cual puede limitar su adopción por parte del personal médico o de investigadores sin formación en ciencia de datos.

Tareas críticas como la creación de visualizaciones, la identificación de anomalías en los datos clínicos o la exploración de tendencias en signos vitales y resultados de laboratorio continúan siendo procesos manuales, que interrumpen el flujo de trabajo y pueden generar errores o pérdida de información relevante [1][3]. Para atender estos retos, se han desarrollado herramientas como Lux, que genera automáticamente visualizaciones sugeridas al mostrar un dataframe, facilitando así la detección de patrones sin necesidad de escribir comandos complejos [1]. Asimismo, herramientas como Mage integran manipulaciones gráficas de datos y modelos con código, reduciendo la fricción cognitiva [2], y AutoProfiler (Dead or Alive) proporciona un perfilado de datos continuo con resúmenes visuales interactivos que permiten identificar errores y comportamientos inesperados en tiempo real [3].

Estas tecnologías surgen de la necesidad de asistir a usuarios durante la fase exploratoria de los datos clínicos, etapa crucial en la comprensión del estado de salud de los pacientes. Aplicadas a UCI cardiovasculares, estas herramientas pueden representar de manera automática tendencias de presión arterial, frecuencia

cardíaca, oxigenación, historial de procedimientos o medicación, permitiendo a los médicos visualizar la evolución de los pacientes y tomar decisiones más informadas con rapidez.

El impacto de estas herramientas recae en su capacidad para aumentar la productividad, mejorar la calidad del análisis de datos clínicos y facilitar la accesibilidad para usuarios con diferentes niveles técnicos. Automatizar parcialmente tareas como la generación de gráficos, la validación de calidad de los datos y el resumen continuo se convierte en una estrategia crucial para fortalecer el monitoreo, predicción y prevención de reingresos a la UCI en pacientes cardiovasculares [1][2][3].

2. Preguntas de Hipótesis

- ¿Existe una relación significativa entre la edad del paciente y su frecuencia cardíaca (heart_rate) por lo cual puede variar su estadía en UCI o su reingreso consecutivo?
- ¿La relación entre la edad y la frecuencia cardíaca varía según el género del paciente (gender)?
- Niveles anormales de creatinina (creatine) están asociados con alteraciones en la presión arterial media (abp_mean)
- ¿La relación entre creatinina y presión arterial media es moderada por la edad del paciente?

3. Objetivos del Análisis

La revisión de donde es que se recolectaron los datos es esencial para saber el contexto de los mismos, entender a detalle sus valores y lo que representan, y entender a detalle los datos para saber cómo realizar la manipulación de los mismos. Por lo que, para investigar las Hipótesis propuestas, se tuvieron en consideración algunos pasos que seguiremos en el informe.

3.1.Análisis de Valores

- Revisar los valores usando en los conjuntos de datos, sus tipos, formato, y analizar sus valores a partir de medidas estadísticas
- Identificar y examinar los valores que estan fuera de la norma o que no son iguales a los demás

3.2.Evaluación de la calidad de los datos

- Revisar la consistencia de los datos, verificando si hay errores de entrada o inconsistencias en la estructura de los registros
- Examinar la integridad de los datos en términos de valores faltantes o inconsistencias temporales.
- Considerar la precisión de los instrumentos de medición y la metodología de recolección de datos para evaluar la fiabilidad de los registros.

3.3.Análisis de correlación de los datos

- Calcular coeficientes de correlación para explorar las relaciones entre las variables de interés, como la presión arterial, ph, etc que guarden relación con la edad y el género
- Realizar análisis gráficos, como diagramas de dispersión, para visualizar las relaciones entre las variables y determinar la fuerza y dirección de la correlación.

3.4.Interpretación de resultados

- Evaluar los hallazgos obtenidos en cada paso del análisis en función de las hipótesis planteadas.
- Identificar patrones significativos, relaciones causales o tendencias emergentes que respalden o refuten las hipótesis formuladas.

2. Fuente de Datos

En este trabajo se analiza una base de datos que fue ya tratada previamente y fusionada de varias bases de datos.

El objetivo de esta data es poder hacer un análisis con respecto a enfermedades cardiacas en UCI, del cual se puede obtener varios resultados, como preveer si una persona saldra de UCI, hacer un estudio para determinar a que grupo esta afectando más y visualizar de forma clara los datos médicos.

Descripción de los datos

Columna	Descripción	Tipo	Naturaleza	Límites	Unidad de medida	% de datos faltantes
subject_id	Identificador único y anonimizado del paciente. Permite rastrear registros individuales a lo largo del tiempo sin revelar su identidad.	int / str	Discreto, categórico	1000 – 1499	*	0%
date	Fecha en la que se registraron los signos vitales o exámenes clínicos. Formato: YYYY-MM-DD.	fecha	Discreto temporal (2,191 fechas únicas)	2001-03-31 – 2007-03-29	*	0%
time	Hora del día en que se tomó la muestra o se midió el dato clínico.	hora	Discreto temporal (1,440 valores únicos)	00:00:00 – 23:59:00	*	0%
age	Edad del paciente al momento del registro.	int	Discreto (puede tratarse como continuo)	19 – 89	años	0%
gender	Sexo biológico del paciente.	str / categórico	Nominal (2 valores únicos: 'M' y 'F')	M, F	*	0%
temperature	Temperatura corporal del paciente. Indicador de infecciones o respuesta inflamatoria.	float	Continuo	36.0 – 40.0	°C	0%
abp_systolic	Presión arterial sistólica	float	Continuo	70.0 – 170.0	mmHg	0%
abp_diastolic	Presión arterial diastólica	float	Continuo	30.0 – 80.0	mmHg	0%
abp_mean	Presión arterial media, muy importante en UCI para evaluar perfusión.	float	Continuo	43.4 – 110.0	mmHg	0%

heart_rate	Frecuencia cardíaca en latidos por minuto.	float	Continuo	50.0 – 157.0	bpm (latidos por minuto)	0%
oxygen_saturation	Saturación de oxígeno en sangre	float	Continuo	90.0 – 100.0	%	0%
weight	Peso del paciente	float	Continuo (con posible error de signos)	-329.0 – 157.0	kg	0%
creatine	Nivel de creatinina en sangre	float	Continuo	0.40 – 2.60	mg/dL o µmol/L	0%
ph	Medida del pH sanguíneo. El valor normal está entre 7.35 y 7.45	float	Continuo	6.8 – 7.7	adimensional	0%
sodium	Concentración de sodio en sangre	float	Continuo	117.0 – 166.0	mEq/L	0%
potassium	Nivel de potasio en sangre	float	Continuo	2.0 – 8.8	mEq/L	0%
hematocrit	Porcentaje de volumen de glóbulos rojos en la sangre	float	Continuo	8.9 – 53.3	%	0%
bilirubin	Nivel de bilirrubina en sangre	float	Continuo	0.1 – 45.0	mg/dL o µmol/L	0%

Analisis Exploratorio de Datos

Análisis del comportamiento de los datos

Manipulación y transformación de datos del archivo
Se procedera con el análisis del archivo que contiene los datos a utilizar sobre los casos de enfermedades cardiacas

- **Formato de los archivos:**

  data_clinic: El archivo se encuentra en formato CSV, es un formato ligero de intercambio de datos.

- **Encoding del archivo**
  Se revisa el encoding del archivo para saber en que formato se leerá posteriormente el archivo, en este caso el resultado del encoding es ascii, por lo cual enconding como UTF-8 también pueden ser usados como encoding para este archivo.

- **Tamaño del archivo**
  El archivo sobre casos tiene un peso de 51.6MB, con lo cual se puede decir que es un archivo de tamaño mediano, y por lo tanto se trabajará con el de manera directa, es decir no hay necesidad de trabajarlo por partes.

- **Limpieza de datos**
  La data estaba previamente limpia pero se hizo un análisis para ver si algún dato era irregular. Se encontró datos con valores negativos en la columna (weight), de los cuales representaban más o menos el 60% de toda la data.

Preguntas:

¿Cuántos registros hay?
Las dimensiones de la data son 516 413 x 18

¿Son demasiado pocos?
No, son una cantidad adecuada para poder hacer el análisis de los datos de forma correcta

¿Son muchos y no tenemos Capacidad (CPU+RAM) suficiente para procesarlo?

No, esta tabla ocupa un total de 51.6MB de memoria RAM por lo cual esta dentro de las capacidades de Google Colab.

¿Hay datos duplicados?

No, se hizo una comprobación en Colab buscando si filas enteras eran lo mismo y se encontró que ninguna era una copia de la otra.

¿Cuales son los tipos de datos de cada columnas?

Tabla data_clinic:

- **subject_id**: Número entero positivo o texto (Identificador único anonimizado del paciente
- **date**: Fecha (Formato AAAA-MM-DD)
- **time**: Hora (Formato HH:MM:SS)
- **age**: Número entero positivo (Edad del paciente en años)
- **gender**: Texto (Valores posibles: 'M' o 'F')
- **temperature**: Número decimal (Temperatura corporal en °C)
- **abp_systolic**: Número decimal (Presión arterial sistólica en mmHg)
- **abp_diastolic**: Número decimal (Presión arterial diastólica en mmHg)
- **abp_mean**: Número decimal (Presión arterial media en mmHg)
- **heart_rate**: Número decimal (Frecuencia cardíaca en latidos por minuto - bpm)
- **oxygen_saturation**: Número decimal (Saturación de oxígeno en sangre en %)
- **weight**: Número decimal (Peso del paciente en kg)
- **creatine**: Número decimal (Nivel de creatinina en sangre en mg/dL o µmol/L)
- **ph**: Número decimal (Nivel de pH sanguíneo, adimensional)
- **sodium**: Número decimal (Concentración de sodio en sangre en mEq/L)
- **potassium**: Número decimal (Nivel de potasio en sangre en mEq/L)
- **hematocrit**: Número decimal (Porcentaje de volumen de glóbulos rojos en la sangre en %)
- **bilirubin**: Número decimal (Nivel de bilirrubina en sangre en mg/dL o µmol/L)

¿Entre qué rangos están los datos de cada columna?, valores únicos, min, max?

- subject_id: 1000 – 1499
- date: 2001-03-31 – 2007-03-29
- time: 00:00:00 – 23:59:00
- age: 19 – 89
- temperature: 36.0 – 40.0 °C
- abp_systolic: 70.0 – 170.0 mmHg
- abp_diastolic: 30.0 – 80.0 mmHg
- abp_mean: 43.4 – 110.0 mmHg
- heart_rate: 50.0 – 157.0 bpm
- oxygen_saturation: 90.0 – 100.0 %
- weight: -329.0 – 157.0 kg
- creatine: 0.40 – 2.60 mg/dL o µmol/L
- ph: 6.8 – 7.7
- sodium: 117.0 – 166.0 mEq/L
- potassium: 2.0 – 8.8 mEq/L

- hematocrit: 8.9 – 53.3 %
-
- bilirubin: 0.1 – 45.0 mg/dL o µmol/L

¿Todos los datos están en su formato adecuado?
No todos, el date y time estan en formato str. Y hay datos que son enteros pero estan como flotantes.

subject_id int64
date object
time object
age int64
gender object
temperature float64
abp_systolic float64
abp_diastolic float64
abp_mean float64
heart_rate float64
oxygen_saturation float64
weight int64
creatine float64
ph float64
sodium float64
potassium float64
hematocrit float64
bilirubin float64

¿Los datos tienen diferentes unidades de medida?
Si, en algunos casos 2 o 3 columnas tiene misma unidad de medida; las que encontramos son: años, °C, mmHg, bpm, kg, mg/dL, mEq/L y %

¿Cuáles son los datos categóricos, ¿hay necesidad de convertirlos en numéricos?
El único es el genero, y no hay necesidad de convertirlos a números ya que puede haber una confusión al colocar números o también el aumento en las categorías para género.

3. Limpieza

¿Están todas las filas completas o tenemos campos con valores nulos? Sí, con ayuda de la herramienta AutoProfiler nos ayudo a visualizar si teníamos algún dato vacio

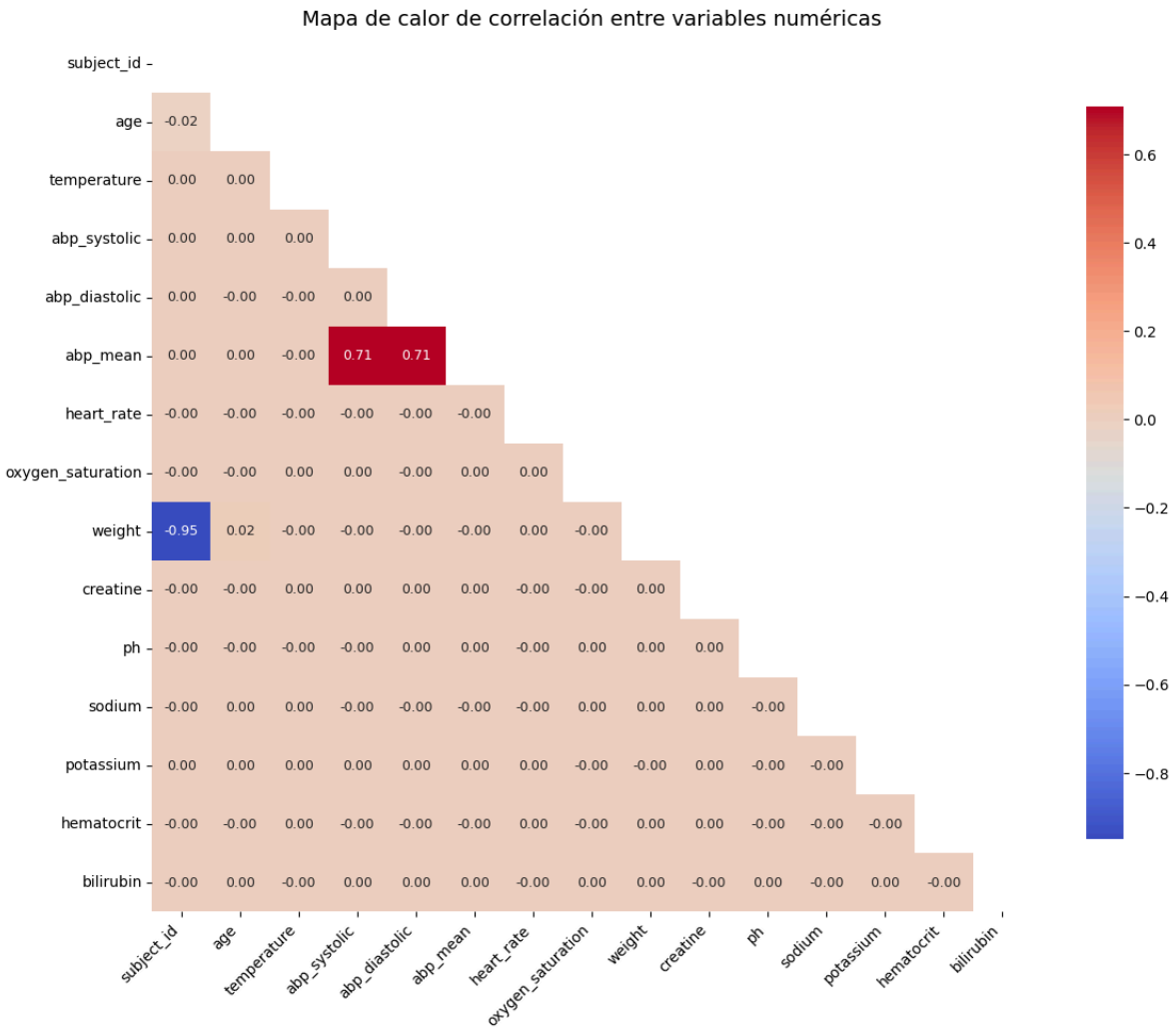
4. Análisis de medidas de tendencia

¿Siguen alguna distribución?

Al aplicar describe() y analizar la columna de fechas en la tabla, se observa una distribución temporal. Que al pasar lo años las personas internadas va en aumento.

.

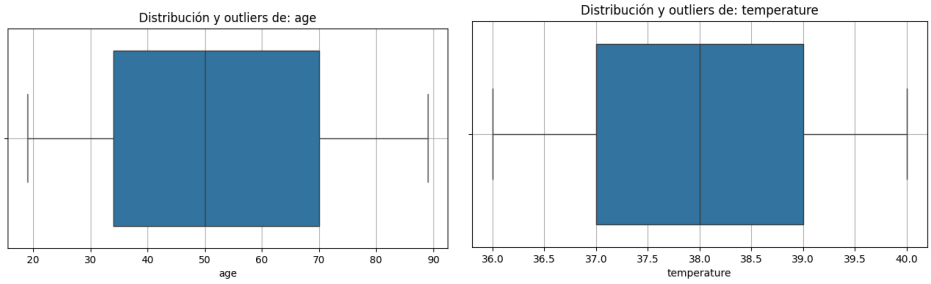
Medidas de tendencia central: media aritmética, geométrica, armónica, mediana, moda, desviación estándar.

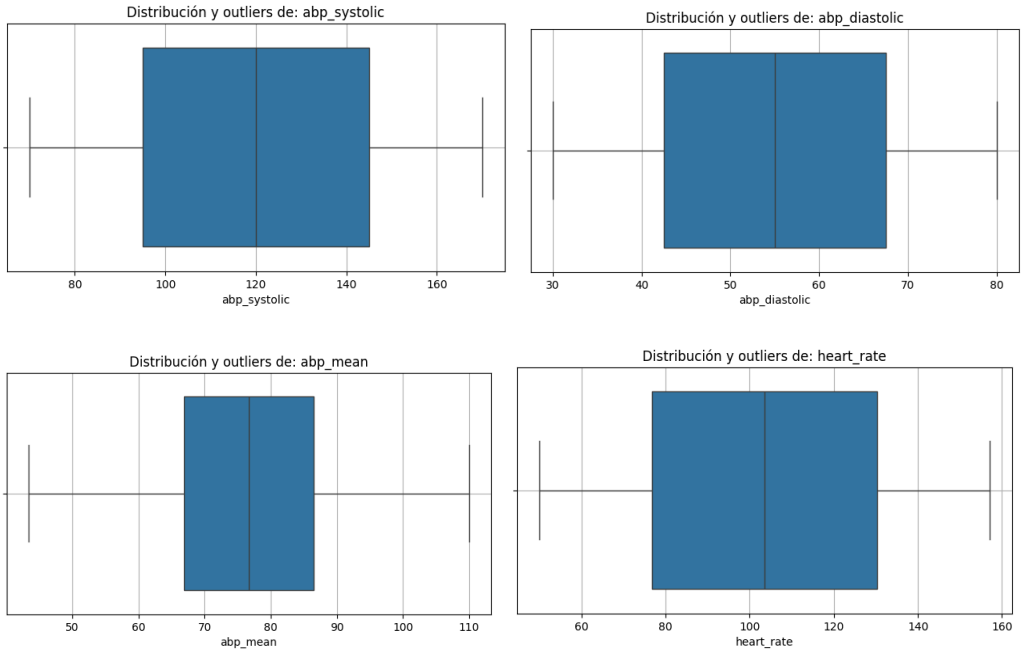


Se observa una fuerta relación entre abp_mean y abp_systolic y abp_dialostic

¿Cuáles son los Outliers? (unos pocos datos aislados que difieren drásticamente del resto y "contaminan" ó desvían las distribuciones)

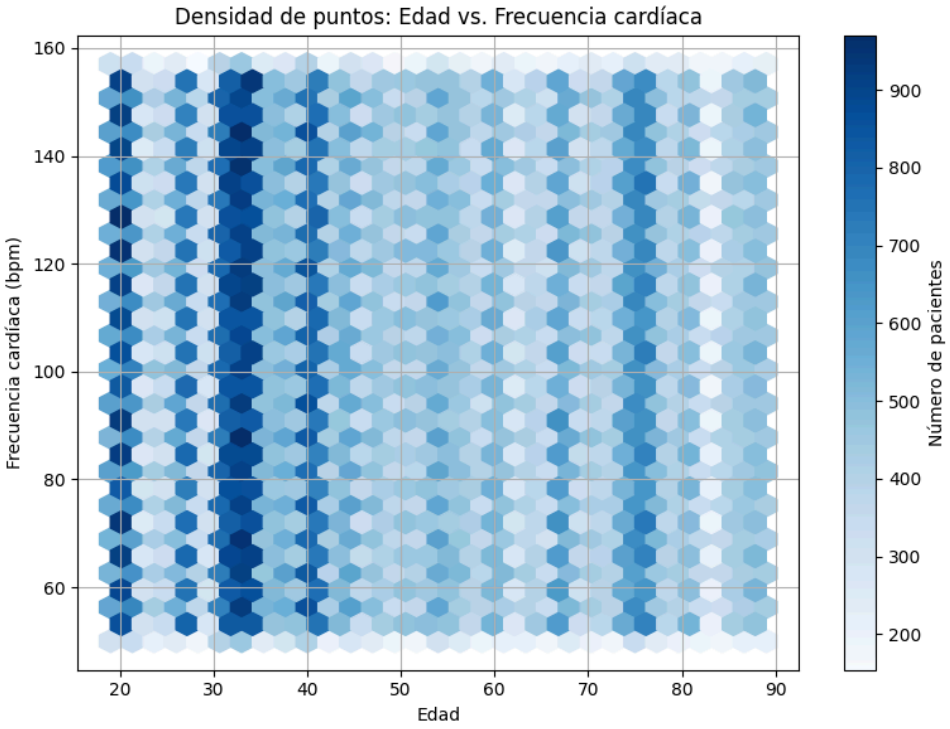
No se encontro Outliers



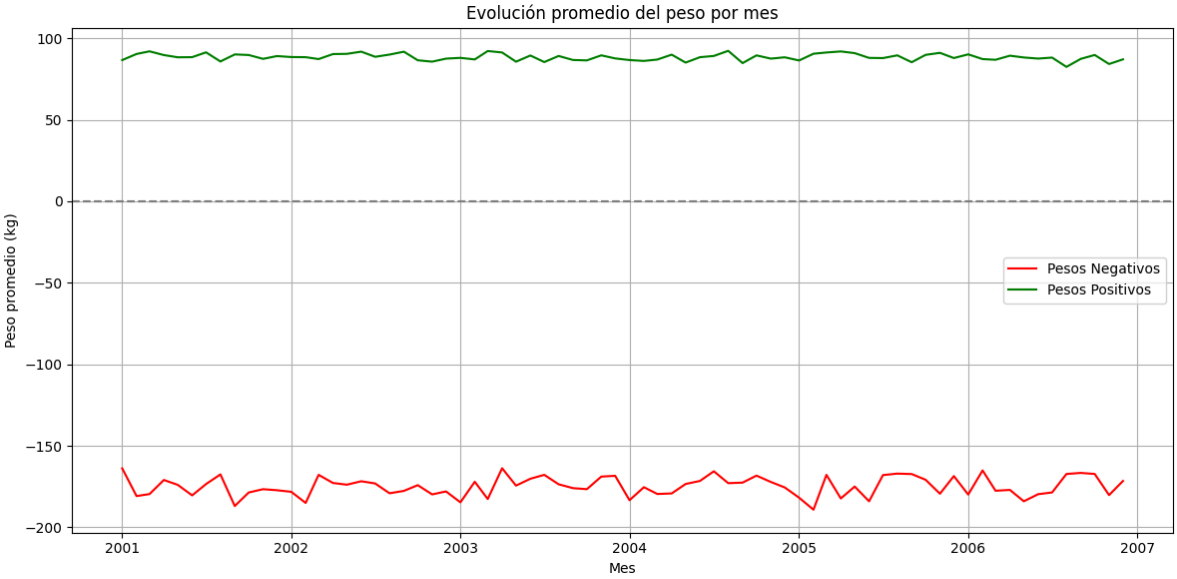


Exploración y Visualización

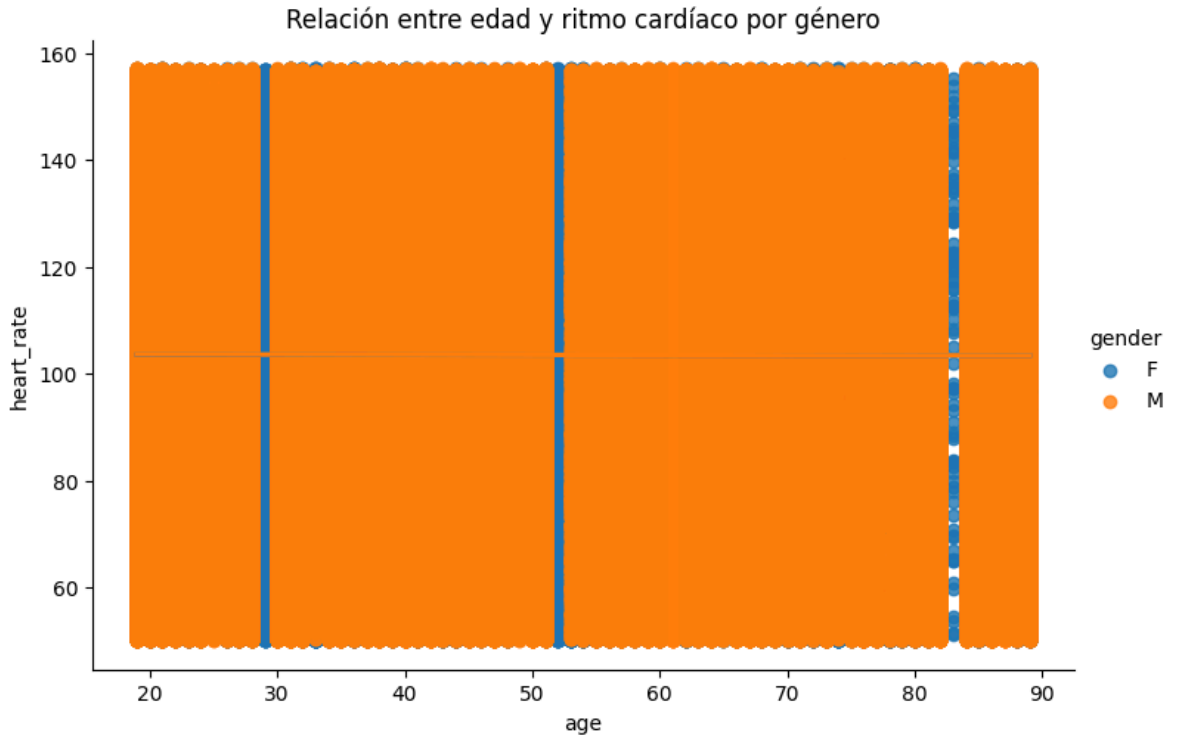
1. Densidad de Edad y frecuencia cardíaca



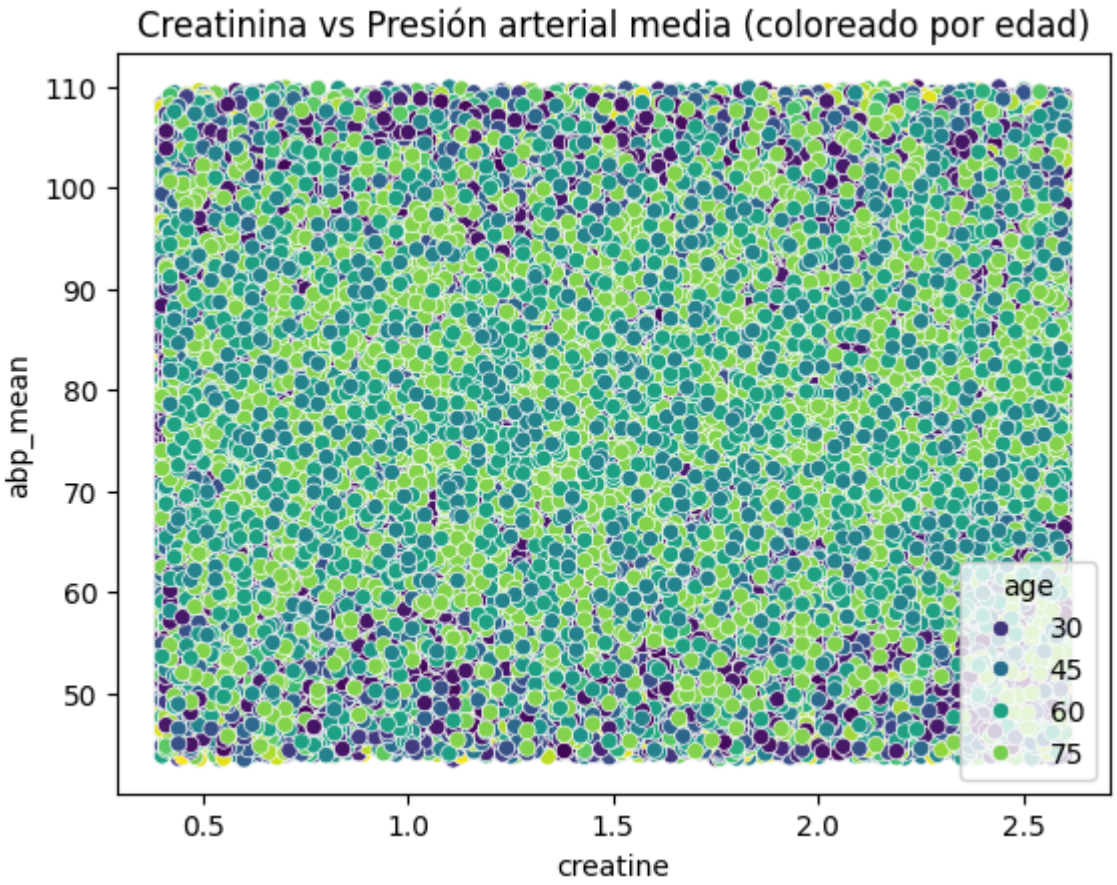
2. Como es la comparación entre los pesos positivos y negativos



3. Existe una relación significativa entre la edad del paciente y su frecuencia cardíaca (heart_rate)



4. Niveles anormales de creatinina (creatine) están asociados con alteraciones en la presión arterial media (abp_mean)



Conclusiones

El análisis y visualización de datos clínicos en UCI cardiovasculares revela la importancia de herramientas intuitivas y automatizadas para mejorar la comprensión del estado de los pacientes. A lo largo del proyecto, se evidenció que tecnologías como AutoProfiler, al integrarse con notebooks computacionales, permiten generar visualizaciones automáticas y perfilados continuos que facilitan la detección de patrones críticos sin requerir conocimientos técnicos avanzados. Esto representa una ventaja significativa para el personal médico, quienes requieren retroalimentación visual inmediata para tomar decisiones en entornos de alta presión. A pesar de que algunas correlaciones como las de edad-frecuencia cardíaca o creatinina-presión arterial no mostraron una relación lineal fuerte, los resultados estadísticamente significativos en ciertos casos sugieren que el análisis visual asistido puede ser clave para identificar relaciones sutiles o no lineales en los datos. En resumen, automatizar y democratizar el análisis exploratorio de datos clínicos representa un paso crucial hacia una medicina intensiva más proactiva, precisa y accesible.

Anexos

Enlace de Código de Data Wrangling: https://colab.research.google.com/drive/1dTylY-YeU-gl5cA2ivWVn9NVcBhhEmJy?usp=sharing

Enlace del dashboard: http://royangelanconeyra.github.io/

Código fuente del dashboard: https://github.com/RoyAngelAnconeyra/Dashboard.git