Análisis de la correlación entre variables cuantitativas y categóricas que actúan como factores de riesgo de mortalidad, con el fin de identificar qué características aumentan la probabilidad de mortalidad

Jeremy García, Emily Sánchez, Alessandro Umaña

Abstract

Este estudio examina la relación entre variables cuantitativas y categóricas asociadas a factores de riesgo de mortalidad en una base de 10 000 individuos adultos y adultos mayores. El análisis descriptivo permitió caracterizar la distribución de variables clínicas, hábitos de consumo y antecedentes de salud, así como identificar valores atípicos que fueron conservados por su posible relevancia. Para evaluar las asociaciones entre variables, se emplearon dos métodos estadísticos: el coeficiente de Pearson para medir relaciones lineales entre variables numéricas y el coeficiente eta cuadrado para cuantificar la magnitud de la relación entre variables categóricas y cuantitativas. Los resultados mostraron que la mayoría de correlaciones fueron débiles o nulas, con excepciones esperadas como las asociaciones entre peso y altura, peso y colesterol, sexo y estatura, y el vínculo moderado entre número de cirugías y medicamentos. Las variables categóricas presentaron muy poca capacidad para explicar la variabilidad de las medidas numéricas. En conjunto, los hallazgos indican que las relaciones observadas no son suficientemente fuertes para explicar patrones de mortalidad.

Keywords: Análisis descriptivo, mortalidad, correlación

Jeremy García
Escuela de Matemática, Universidad de Costa Rica, San José, Costa Rica, e-mail: jeremy.garciasolano@ucr.ac.cr

Emily Sánchez
Escuela de Matemática, Universidad de Costa Rica, San José, Costa Rica, e-mail: emily.sanchezmancias@ucr.ac.cr

Alessandro Umaña
Escuela de Matemática, Universidad de Costa Rica, San José, Costa Rica, e-mail: alessandro.umana@ucr.ac.cr

1 Introducción

La presente investigación se centra en el análisis de la correlación de variables cuantitativas y categóricas de una base de datos, las cuales muestran factores de riesgo de mortalidad. Esto con el propósito de identificar cuáles características incrementan la probabilidad de fallecimiento de una población determinada. Este estudio parte de la pregunta de cómo se relaciónan ambos tipos de variables cuando se aplican herramientas estadísticos y de análsis de datos en R, lo que permite reconocer y analizar de mejor manera los patrones que existen dentro de las variables que no se pueden observar a simple vista.

El objetivo del estudio es examinar la interacción entre variables numéricas con variables numéricas y variables cualitativas con variables numéricas, esto también para poder analizar dos tipo de correlaciones, una lineal y una no lineal para lograr un alcance mayor con los patrones que puedan haber detras de las variables y tambien para respetar la naturaleza de los tipos de datos que se proporcionan. Para ello, se consideran los conceptos fundamentales como variable cuantitativa, la cual es entendida como valores numéricos los cuales son medibles, variables categóricas, que describen características o atributos que no son medibles numéricamente y el concepto de mortalidad, el cual es la condición inherente al ser humano de estar sujeto al fin de la vida, donde se busca encontrar esta probabilidad mediante el análisis a través del trabajo.

Mediante esta aproximación el estudio busca aportar herramientas que puedan facilitar la interpretación de diferentes factores que influyen a la mortalidad, ofreciendo una base teórica y metodológica sólida para el análisis de los datos en el contexto de salud y estadística.

2 Marco teórico conceptual

El presente marco teórico conceptual establece las bases necesarias para comprender los elementos fundamentales usados en el análisis de los factores de riesgo de mortalidad. Para ello, se abordan los conceptos clave relacionados con los tipos de variables utilizadas en el estudio, así como las condiciones médicas y características personales que pueden influir en la probabilidad de muerte. Este apartado proporciona una guía conceptual que permite interpretar de manera adecuada los resultados del análisis estadístico, asegurando una comprensión mas integra de los fenómenos y de las herramientas empleadas en el trabajo.

Primero se explicará la base de datos, las variables cuantitativas del estudio representan medidas numéricas directamente asociadas al estado de salud de las personas. La edad indica los años cumplidos y permite evaluar el riesgo asociado al envejecimiento, el peso y la altura aportan información antropométrica relevante para estimar condiciones como sobrepeso u obesidad, la presión arterial sistólica refleja la fuerza con la que la sangre circula por las arterias y constituye un indicador fundamental de salud cardiovascular. Además, el número de medicamentos que

una persona consume y la cantidad de bebidas alcohólicas ingeridas por semana permiten conocer hábitos que pueden influir en su estado de salud. El número de cirugías mayores realizadas da una visión del historial clínico del individuo, mientras que el nivel de colesterol funciona como un marcador del riesgo cardíaco.

Por su parte, las variables categóricas describen características o condiciones que no se expresan numéricamente, pero que pueden influir en la mortalidad. El sexo permite diferenciar riesgos biológicos asociados al género, el consumo de tabaco, cannabis u otras drogas evidencia hábitos que pueden afectar la salud, la presencia de adicción refleja una condición crónica con posibles consecuencias negativas. Además, se consideran enfermedades o antecedentes como diabetes, ataques al corazón, derrames, asma, inmunodeficiencia, antecedentes familiares de cáncer, enfermedades cardíacas y problemas de colesterol, todos ellos factores con potencial impacto en la probabilidad de mortalidad. Finalmente, se incluye si la ocupación de la persona es peligrosa, ya que ciertos trabajos incrementan la exposición a riesgos que pueden comprometer la vida.

coePor otro lado, se trabajan la correlación, la cual es una medida que indica el grado y dirección entre dos variables, esta puede ser lineal o no lineal, donde si es lineal nos indica la fuerza y la dirección entre dos variables, mientras si es no lineal se da cuando el aumento y disminución entre dos variables no se da con la misma intensidad, por eso es importante realizar ambos tipos de correlaciones, para observar los diferentes comportamientos que tienen las variables respecto a otras variables.

3 Marco teórico metodológico

El presente marco metodológico expone los procedimientos empleados para analizar la relación entre variables cuantitativas y categóricas que actúan como factores de riesgo de mortalidad. Entre los principales métodos y técnicas que se utilizaron para el análisis estuvieron los coeficientes de correlación eta cuadrado y Pearson.

Según Lalinde, et al., (2018), el coeficiente de Pearson es una medida de coeficiente correlación donde muestra la asociación lineal y la fuerza de dirección que existe entre las variables, donde esta se presenta con un rango de -1 a 1, donde entre más cercana a 1 significa una correlación positiva en el mismo sentido, una cercanía a -1 es una correlación negativa donde indica que ambas variables se relacionan de manera inversa, pero en direcciones opuestas y entre más se aproxime a 0, significa que no existe una correlación lineal. De igual forma, existen categorías para determinar que tan fuerte es la correlación entre variables, entonces se considera una correlación nula si se encuentra entre 0.00 y 0.1, correlación débil entre 0.1 a 0.3, correlación moderada entre 0.3 a 0.5 y 0.5 a 1 se considera que tiene una correlación fuerte. La fórmula matemática está dada por:

$$\gamma_{XY} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{[\sum(X-\overline{X})^2 \sum(Y-\overline{Y})^2]^{1/2}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

Donde $\overline{X}$ y $\overline{Y}$ son las medias muestrales de X y Y y S_{XX}, S_{YY} y S_{XY} son las sumas de los cuadrados corregidas de X, Y y el producto cruzado de XY.

Para hacer uso de la Pearson es necesario que ambas variables sean de intervalo o de razón, pero no es necesario que ambas tengan el mismo nivel de medición para que este funcione. También se menciona que es importante que no hayan casos faltantes, ya que estos se descartarán del todo a la hora de realizar el estudio.

Además, se menciona el cuidado de que no presentar efectos de enmascaramiento o empantanamiento. Donde se dice que el enmascaramiento es según Lalinte et al. (2020) "sobreviene cuando un dato aberrante no es descubierto debido a la presencia de otros valores atípicos adyacentes." y el empantanamiento es "ocurre cuando una observación no extrema es clasificada como outlier producto de la existencia de otros datos normales".

Por otro lado, para eta cuadrado se tiene que DeCaires et al., (2018) define este como un índice general de correlación el cual presenta una regresión curvilínea, donde deben existir dos variables: la variable independiente, en este caso será X, y la variable dependiente, en este caso será Y.

Para la práctica, la variable X será la variable categórica y la variable Y una variable cuantitativa. Hay que notar que el valor para cada regresión difiere con como se calcule, ya que es diferente calcular γ_{XY}, que calcular γ_{YX}.

Eta cuadrado presenta algunos requisitos para que se pueda calcular y sea efectivo, la primera es que haya una variable cuantitativa (variable dependiente) y una variable categórica (variable independiente), la población debe estar normalmente distribuida y esta se puede aplicar a muestras grandes. Lo cual se cumple para las variables que vamos a tratar.

El cálculo de esta es la división de la suma de los cuadrados de los grupos por la suma total de los cuadrados, así mostrando la "proporcionalidad" que existe entre las variables. Donde su fórmula es la siguiente:

$$\eta^2 = \frac{\sum_x \eta_x (\overline{y}_x - \overline{y})^2}{\sum_{x,i} (y_{xi} - \overline{y})^2}$$

Para la interpretación de los resultados se tiene que si el resultado es menor o igual a 0.3 se considera que tiene un valor significativo pero débil, si es menor o igual a 0.6, pero mayor a 0.3 tiene un valor moderado y si es mayor a 0.6 tiene un valor fuerte.

Sin embargo, se dice que tiene limitaciones, tales como que el coeficiente siempre dará un valor positivo, por lo que no se toma en cuenta la dirección, pero en el momento de realizar la interpretación se puede aclarar la dirección que pueden tomar las diferentes variables usando un poco el contexto de los datos, el coeficiente se realiza para muestras grandes, por lo que su cálculo puede ser muy extenso y este coeficiente es equivalente al coeficiente de determinación que calcula correlación lineal.

En síntesis, estas dos técnicas para calcular coeficientes de correlación permitirá abordar y analizar factores de riesgo de muerte. Mientras Pearson cuantifica la fuerza y dirección de la asociación lineal entre variables cuantitativas, η^2 permite evaluar la magnitud de la relación entre variables categóricas y una variable cuantitativa, en-

contrando posibles efectos no lineales. Complementar ambas metodologías garantiza un análisis más preciso, garantizando que los resultados obtenidos sean coherentes, válidos y adecuados para responder a las preguntas planteadas en el estudio.

4 Datos y análisis descriptivo

El conjunto de datos utilizado en este estudio está conformado por 10 000 observaciones, cada una correspondiente a una persona adulta o adulta mayor fallecida, con edades entre 25 y 120 años. Para cada individuo se registró información clínica, hábitos de consumo y antecedentes familiares, permitiendo un análisis profundo sobre posibles factores asociados con la mortalidad. La población incluye 5 034 mujeres y 4 966 hombres, lo que garantiza una distribución relativamente balanceada por sexo.

Las variables cuantitativas analizadas incluyen edad, peso, altura, presión arterial sistólica, nivel de colesterol, número de medicamentos consumidos, cantidad de bebidas alcohólicas por semana y número de cirugías mayores realizadas. De forma general, estas variables presentan distribuciones con ligeras asimetrías, especialmente en peso, colesterol y número de cirugías, donde se encuentran valores extremos característicos de condiciones clínicas graves.

Resultados Descriptivos Cuantitativos:

Estos son algunos resultados obtenidos después de realizar el análisis de las variables cuantitativas: La edad con una mediana de 64 años y una media de 64.64, nos dice que la mayor parte de la población estudiada pertenece a personas con una edad avanzada. El colésterol varía ente valores de 83 mg/dL y 352 mg/dL, con una media de 199.7mg/dL y una mediana de 199 mg/dL. Sin embargo, como se observa en el gráfico (fig 1), existe mayor densidad de mujeres en niveles medios-altos, lo cual contrasta con la concentración masculina en valores bajos-medios.

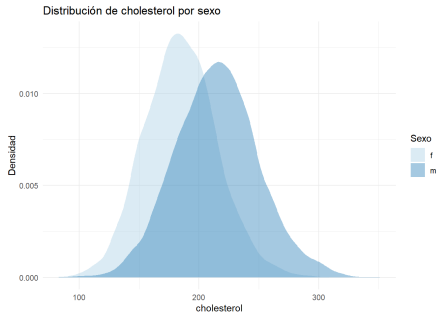

Fig. 1 Elaboración propia, datos recopilados de la base

Resultados Descriptivos Cualitativos: Este análisis nos permite entender a la población estudiada y su como está compuesta según algunas características.

La Diabetes es una de las variables categóricas presente en la base de datos, la mayoría de las personas estudiadas no presentan diabetes, haciendo que este grupo sea minoritario. Esta diferencia de tamaños entre poblaciones es relevante ya que afecta a los resultados que se obtener de esta variable. Además, como más adelante se va a mencionar, tener o no tener diabetes no tiene una relación muy fuerte con variables como presión arterial o el colesterol.

Otra de la variables que se estudian es el Fumado, la cual tambien tiene un gran desequilibrio, aproximadamente 8000 personas no fuman, mientras que solamente 2000 si lo hacen. Gracias a este desequilibrio, las relaciones que se quieran estudiar, se van a ver sesgadas con esta limitante de población. Esto se puede ver reflejado en el gráfico de presión arterial según fumado, donde las medianas son casi idénticas, A pesar de que biológicamente el fumado si influye en la presión arterial.

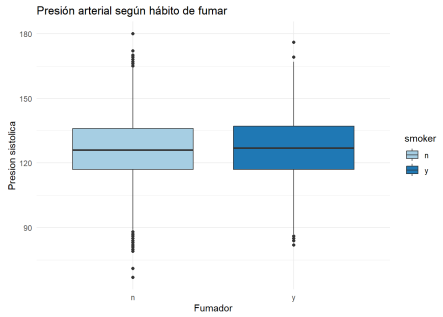

Fig. 2 Elaboración propia, datos recopilados de la base

Durante la exploración preliminar no se identificaron valores faltantes, lo que facilita el uso de técnicas estadísticas que requieren bases completas como Pearson. No obstante, se detectaron valores atípicos (outliers) en la mayoría de las variables numéricas, especialmente en colesterol, peso y consumo semanal de bebidas alcohólicas. Estos valores fueron identificados mediante gráficos de caja y métodos basados en el rango intercuartílico (IQR).

Aunque son observaciones extremas, se decidió no eliminarlas porque pueden corresponder a condiciones clínicas reales y aportar información relevante al análisis.

El análisis descriptivo también incluyó la inspección gráfica por género, donde se observaron patrones esperados: las mujeres se concentran en valores menores de peso y altura, mientras que los hombres tienden a presentar más casos en rangos altos de bebidas por semana y cirugías mayores. Asimismo, variables como presión sistólica y colesterol mostraron dispersión considerable, lo que refuerza la necesidad de métodos estadísticos que permitan evaluar asociaciones sin asumir relaciones estrictamente lineales.

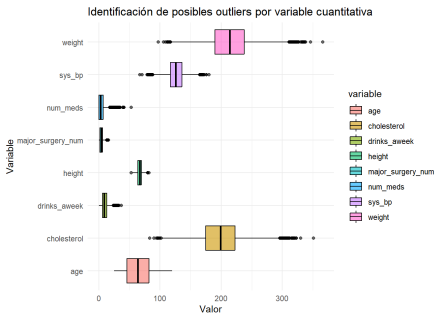

Fig. 3 Elaboración propia, datos recopilados de la base

5 Métodos y Resultados

A continuación se presentan los resultados obtenidos a partir de los dos métodos utilizados:

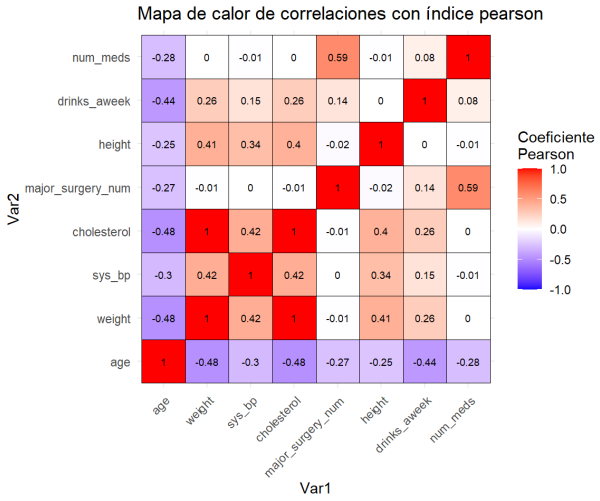

Fig. 4 Elaboración propia, datos recopilados de la base

En el caso del coeficiente de Pearson, el mapa de calor permitió identificar algunas relaciones relevantes. La relación entre la altura y el peso mostró una correlación positiva moderada, coherente con lo que suele observarse en poblaciones reales, en donde una mayor estatura tiende a asociarse con un mayor peso corporal. Por otro lado, la relación entre la edad y el número de cirugías mayores evidenció una correlación negativa débil. Esto indica que, conforme aumenta la edad, el número de cirugías presenta una ligera disminución, aunque esta relación es tan baja que no

constituye una asociación significativa. De forma similar, se observó una correlación moderada entre la edad y la presión sistólica, señalando que la presión arterial tiende a disminuir con el incremento de la edad, mientras que el peso y la presión sistólica mostraron una correlación moderada y positiva, consistente con evidencia clínica que vincula el aumento de peso con incrementos en la presión arterial. De igual forma, el análisis evidenció que el número de medicamentos y el número de cirugías presentan una relación lineal de magnitud moderada. Esto implica que, en general, las personas que han pasado por más procedimientos quirúrgicos tienden también a utilizar una mayor cantidad de medicamentos. Aunque la relación no es perfecta, sí sugiere un patrón consistente en el que ambos factores se incrementan de manera conjunta, reflejando posiblemente una mayor complejidad en el estado de salud de estas personas.

En términos generales, la mayoría de las demás correlaciones cuantitativas fueron muy bajas, lo que sugiere que no existe relación lineal entre esas variables dentro del conjunto de datos. Sin embargo, llamó la atención la relación entre colesterol y peso, que presentó un coeficiente de 1, reflejando una asociación extremadamente fuerte en el conjunto de datos. Este resultado indica que, en estos individuos, el nivel de colesterol aumenta o disminuye junto al peso corporal.

En conjunto, los coeficientes de Pearson señalan que la base de datos analizada no contiene relaciones lineales fuertes entre la mayoría de las variables numéricas estudiadas.

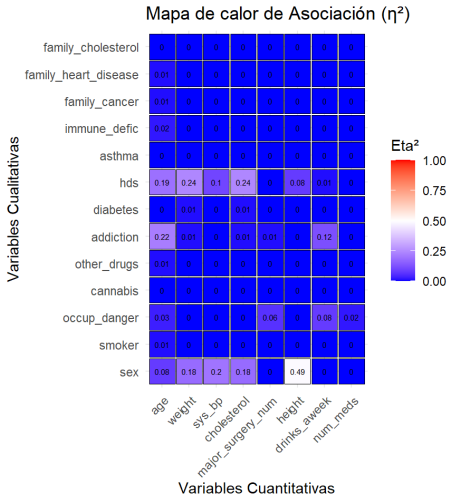

Fig. 5 Elaboración propia, datos recopilados de la base

En cuanto a la Figura 4, análisis mediante eta cuadrado, los resultados no mostraron relaciones relevantes entre variables categóricas y cuantitativas. El vínculo más destacado se observó entre el género y la estatura, un resultado esperado por diferencias biológicas existentes. Asimismo, se identificó una relación débil entre el

género y el peso, lo cual coincide con variaciones físicas que suelen presentarse entre los grupos. Por otra parte, las variables categóricas relacionadas con condiciones de salud como presencia de diabetes, adicciones, problemas cardíacos, antecedentes familiares o riesgos ocupacionales no mostraron relaciones significativas, con la excepción de las relaciones débiles que se encontraron de enfermedades cardíacas con edad, peso, presión sistólica y colesterol. Esto sugiere que estas clasificaciones explican muy poco la variación de medidas cuantitativas como colesterol, presión arterial o cantidad de medicamentos utilizados.

En conclusión, si bien las categorías analizadas tienen cierta relación con patrones esperados, su capacidad para explicar las diferencias en las variables numéricas es muy débil. Esto sugiere que la mayor parte de la variabilidad en los datos cuantitativos no depende de estas clasificaciones, sino de otros factores no incluidos o no captados en este estudio.

6 Conclusiones

Si bien el análisis realizado permitió identificar asociaciones relevantes entre variables cuantitativas y categóricas que actúan como factores de riesgo de mortalidad, es necesario reconocer ciertas limitaciones inherentes al conjunto de datos y a los métodos estadísticos empleados. En primer lugar, dado que se desconoce el origen de la información del conjunto de datos, este difícilmente puede reflejar la complejidad de la población real, lo cual limita la posibilidad de generalizar los resultados. Otro aspecto a considerar en relación con las técnicas utilizadas radica en que el uso del coeficiente eta cuadrado requirió un proceso previo de investigación y aprendizaje, ya que no formaba parte de los conocimientos iniciales del equipo. Esto implicó tiempo adicional para su comprensión y desarrollo en RStudio, lo que pudo restringir la exploración de métodos complementarios durante el trabajo.

El análisis realizado permitió entender mejor cómo se relacionan las variables que aumentan el riesgo de mortalidad del conjunto de datos y qué tan fuertes son estas mismas. Se observó que las correlaciones en general fueron bajas tanto entre variables numéricas como entre variables numéricas y categóricas, lo que indica que en este conjunto de datos la mayoría de las variables no muestran relaciones lineales fuertes, a excepción de algunas cuantas como peso-colesterol, colesterol-presión arterial sistólica, numero de medicamentos-número de cirugías mayores, peso-altura y sexo-altura. Se puede concluir que las relaiciones más fuertes que se evidenciaron fueron entre variables numéricas. Tanto Pearson como eta cuadrado ayudaron a tener un panorama general de las relaciones entre las variables, pero también fue evidente que estos métodos no fueron suficientes para explicar un fenómeno tan complejo como la mortalidad, por lo que sería necesario aplicar modelos más avanzados, como regresiones o técnicas predictivas. Todo el proceso refuerza la importancia de revisar valores atípicos, explorar los datos y asegurarse de entender su estructura antes de aplicar métodos estadísticos más complejos. En resumen, aunque se identificaron algunos patrones, los resultados no muestran relaciones realmente fuertes entre las

variables estudiadas y la mortalidad, lo que deja abierta la puerta a futuros análisis más profundos y al uso de bases de datos más realistas o específicas para obtener conclusiones más sólidas.

Como posibles mejoras para futuros trabajos, se considera necesario aplicar regresiones logísticas o modelos predictivos que permitan analizar las relaciones con mayor profundidad. También sería útil aplicar técnicas más robustas para el manejo de outliers y distribuciones no normales, así como crear nuevas variables que permitan que el análisis sea más completo.

References

1. *Correlaciones lineales*. (s.f.). ChReinvent. https://www.chreinvent.com/recursos/correlaciones-lineales
2. Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: a review. *International journal of psychological research*, 3(1), 58-67.
3. DeCaires, A., Fernández, C., García, G., Olivo, M., & Vera, L. (2018). *COEFICIENTE ETA CUADRADO O RAZÓN DE CORRELACIÓN* (Documento no publicado). Universidad Central de Venezuela.
4. Diccionario de cáncer del NCI. (s.f.). Cancer.gov. Recuperado de https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/mortalidad
5. Fonnesu, M., & Kuczewski, N. (2019). *Doubts on the efficacy of outliers correction methods*. arXiv preprint arXiv:1907.09864.
6. Fundación del Corazón. (2024). *Colesterol y riesgo cardiovascular*. Fundación del Corazón. Recuperado de https://fundaciondelcorazon.com/prevencion/riesgo-cardiovascular/colesterol.html
7. Hyndman, R. J., & Athanasopoulos, G. (s. f.). *13.9 Tratamiento de valores atípicos y ausentes*. En Forecasting: Principles and Practice (ed. 3). Recuperado de https://otexts.com/fppsp/missing-outliers.html
8. Kwak, S. K., Kim, J. H. (2017). *Statistical data preparation: management of missing values and outliers*. Korean Journal of Anesthesiology, 70(4), 407-411. Recuperado de https://doi.org/10.4097/kjae.2017.70.4.407
9. Lalinde, J. D. H., Castro, F. E., Rodríguez, J. E., Rangel, J. G. C., Sierra, C. A. T., Torrado, M. K. A., ... & Pirela, V. J. B. (2018). Sobre el uso adecuado del coeficiente de correlación de Pearson: definición, propiedades y suposiciones. Archivos venezolanos de Farmacología y Terapéutica, 37(5), 587-595.
10. Ocaña Peinado, F. M. (s. f.). *Tratamiento estadístico de outliers y datos faltantes*. Universidad de Granada. Recuperado de https://www.ugr.es/~fmocan/MATERIALES%20DOCTORADO/Tratamiento%20de%20outliers%20y%20missing.pdf
11. Palmer, A., Jiménez, R., & Montaño, J. J. (2000). Tutorial sobre coeficientes de correlación con una o dos variables categóricas. Revista Electrónica de Psicología, 4(2), 1-19.
12. Sánchez Acero, F. (2024). Variables Cualitativas y Cuantitativas. 2024. Recuperado de https://repositorio.konradlorenz.edu.co/handle/001/6087

    Sanjuán, F. J. M. (2022, 24 noviembre). *Relación no lineal - Definición, qué es y concepto — Economipedia*. Economipedia. https://economipedia.com/definiciones/relacion-no-lineal.html

7 Anexos

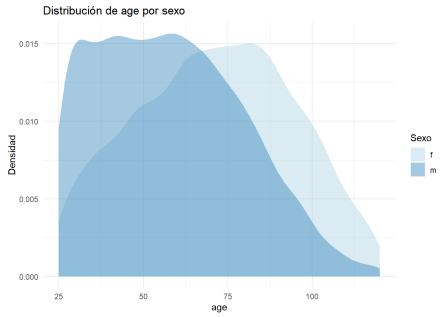

Fig. 6 Elaboración propia, datos recopilados de la base

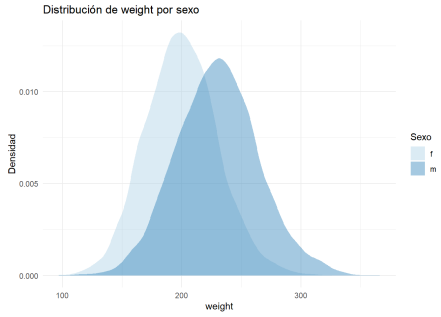

Fig. 7 Elaboración propia, datos recopilados de la base

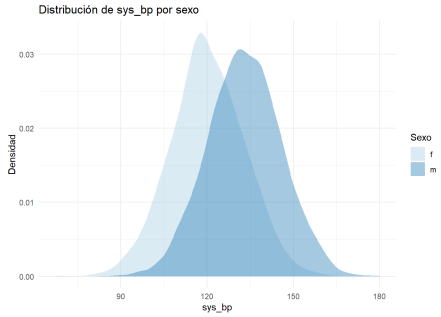

Fig. 8 Elaboración propia, datos recopilados de la base

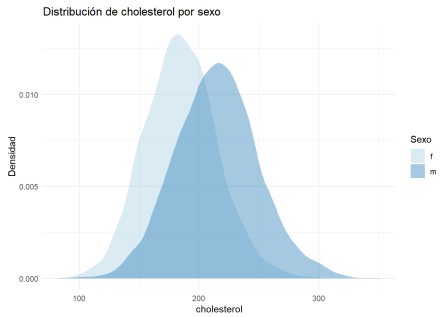

Fig. 9 Elaboración propia, datos recopilados de la base

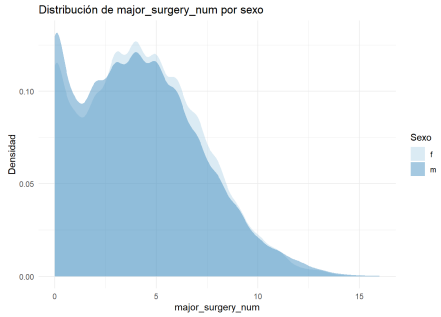

Fig. 10 Elaboración propia, datos recopilados de la base

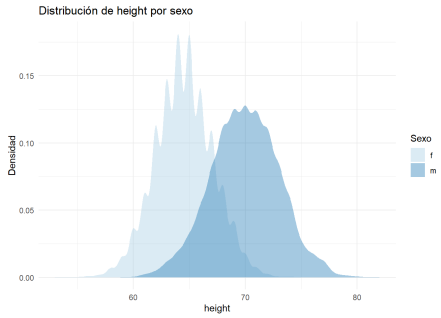

Fig. 11 Elaboración propia, datos recopilados de la base

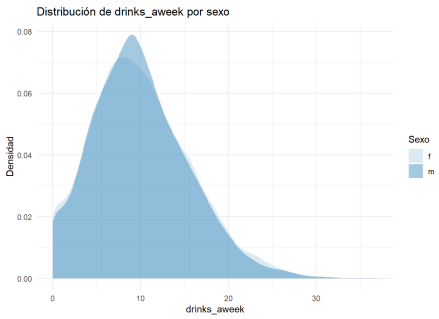

Fig. 12 Elaboración propia, datos recopilados de la base

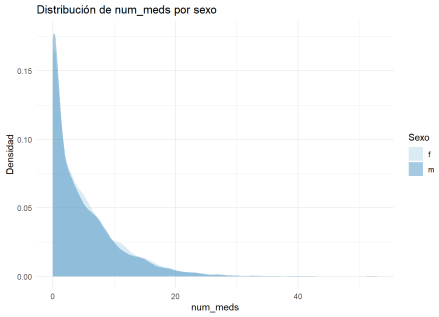

Fig. 13 Elaboración propia, datos recopilados de la base

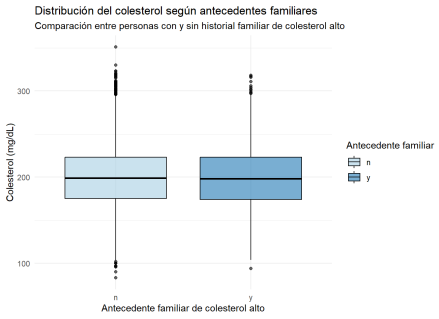

Fig. 14 Elaboración propia, datos recopilados de la base

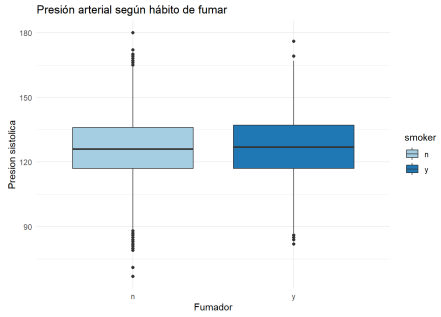

Fig. 15 Elaboración propia, datos recopilados de la base

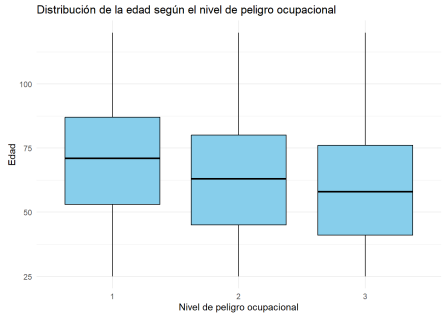

Fig. 16 Elaboración propia, datos recopilados de la base

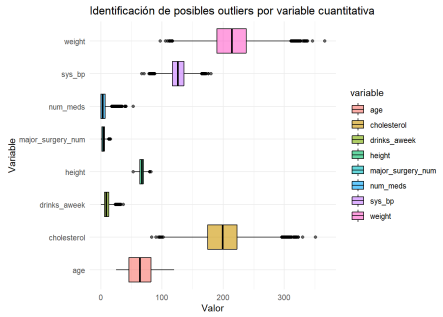

Fig. 17 Elaboración propia, datos recopilados de la base

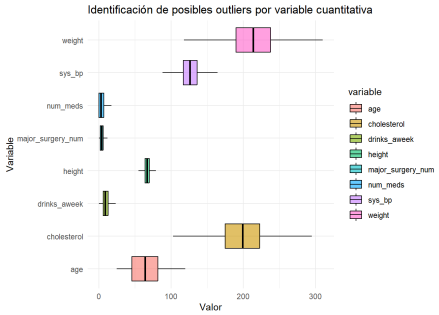

Fig. 18 Elaboración propia, datos recopilados de la base

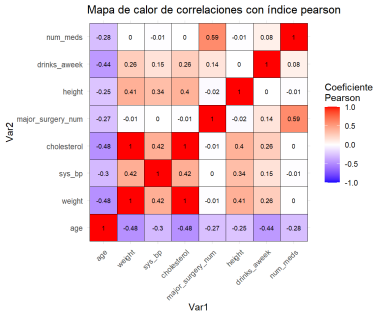

Fig. 19 Elaboración propia, datos recopilados de la base

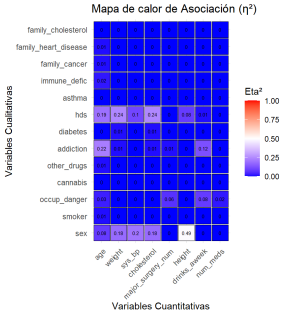

Fig. 20 Elaboración propia, datos recopilados de la base