Universidad de Costa Rica

Facultad de Ciencias

Escuela de Matemática

Departamento de Matemática Pura y Ciencias Actuariales

Herramientas de datos I CA-0204

Bitácoras

Estudiantes:

García Solano Jeremy C33171

Sánchez Mancía Emily Alejandra C27260

Umaña Vega Alessandro C37963

Octubre del 2025

Índice

1. Bitácora 1

1.1. Tema

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.

1.2. Pregunta de Investigación

¿Cómo se correlacionan las variables cuantitativas con las variables categóricas que constituyen un factor de riesgo de muerte mediante el uso deherramientas de análisis de datos en R?

1.3. Objeto de estudio

El 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 utilizando herramientas de análisis de datos en R.

1.4. Conceptos

a) Variable cuantitativa: variable estadística expresada mediante valores numéricos que representan cantidades medibles. Estas se dividen en discretas y continuas.

b) Variable categórica: variable estadística también conocida como cualitativa.

Estas representan características o atributos que no pueden medirse numéricamente. Pueden tomar valores nominales u ordinales.

c) Mortalidad: es la cualidad o el estado de ser mortal (destinado a morir).

1.5. Teorías

- Manríquez, G., & Escudero, C. (2017). Análisis de los factores de riesgo de muerte neonatal en Chile, 2010-2014. Revista chilena de pediatría, 88(4), 458-464.
- 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.
- 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.

2. Bitacora 2

2.1. Descripción de los datos

a) Características de la tabla de datos: El conjunto de datos contiene información de salud y hábitos de 10.000 personas, de las cuales 4.966 son hombres y 5.034 son mujeres. Incluye variables como edad, peso, altura, presión arterial sistólica, número de medicamentos que toma la persona, cantidad de cirugías mayores realizadas, así como la cantidad de bebidas alcohólicas que toma a la semana, el nivel de colesterol, y el sexo del individuo. Además, registra hábitos y condiciones de salud, como consumo de tabaco, cannabis u otras drogas(así como el uso de nicotina en otras formas), uso de opioides, adicción, diabetes, antecedentes de ataques al corazón o derrames, asma, inmunodeficiencia, así como antecedentes familiares de cáncer, enfermedades cardíacas y problemas de colesterol. También se indica si la ocupación o si el estilo de vida de la persona se considera peligrosa.

b) Población de estudio: La población de estudio está compuesta por individuos adultos y adultos mayores registrados en un estudio médico simulado.

c) Muestra Observada: La muestra contiene 10.000 observaciones de personas de edad adulta y adulta mayor, representando una fracción de la población total.

d) Unidad estadística: Cada persona registrada en la base de datos conforma una unidad estadística.

e) Identificar las variables de estudio:

- **Cuantitativas:** Las variables numéricas que se utilizarán en el estudio incluyen edad, peso, altura, presión arterial sistólica, número de medicamentos que toma la persona, cantidad de bebidas alcohólicas consumidas por semana, número de cirugías mayores realizadas y nivel de colesterol.

- **Categóricas:** Las variables categóricas consideradas en el estudio son sexo, consumo de tabaco, consumo de cannabis u otras drogas, adicción, diabetes, antecedentes de ataques al corazón o derrames, asma, inmunodeficiencia, antecedentes familiares de cáncer, enfermedades cardíacas y problemas de colesterol, así como si la ocupación de la persona se considera peligrosa.

2.2. 5 Primeras observaciones

Estas son las primeras 5 lineas de la base de datos

age	weight	sex	height	sys_bp	smoker	nic_other	num_meds	occup_danger	ls_danger	cannabis	opioids	other_drugs	drinks_aweek	addiction	r_surgery	diabetes	hds	cholesterol	asthma	immune_defic	family_cancer	family_heart_disease	family_cholesterol
100	219	m	74	136	n	n	0	1	1	n	n	n	4	n	0	n	y	203	n	n	y	n	y
66	242	m	73	111	n	n	0	1	1	n	n	n	6	y	0	n	n	228	n	n	n	n	n
31	197	f	65	112	n	n	7	1	2	n	n	n	16	y	3	n	y	183	n	n	n	n	n
42	244	f	69	127	n	n	1	2	3	n	n	n	16	n	2	n	y	228	n	n	n	n	n
93	183	f	63	91	y	n	2	3	3	n	n	n	26	y	2	n	n	169	n	n	n	n	n

Figura 1: Elaboración propia, datos recopilados de la base

2.3. Variables cuantitativas

Cuadro 1: Resumen estadístico de variables seleccionadas

	Age	Height	Weight	Major Surgery Num	Cholesterol
Min.	25.00	53.00	97.0	0.000	83.0
1st Qu.	46.00	64.00	190.0	2.000	175.0
Median	64.00	67.00	214.0	4.000	199.0
Mean	64.64	67.24	214.7	4.171	199.7
3rd Qu.	82.00	70.00	238.0	6.000	223.0
Max.	120.00	82.00	366.0	16.000	351.0

Se hará un análisis de las siguientes cinco variables cuantitativas:

- **Edad (age):** Representa la edad de las personas, con valores entre 25 y 120 años. La mediana y la media se aproximan a 64 años, aunque la media es ligeramente mayor (por 0.64), lo que sugiere que una gran proporción de la muestra corresponde a adultos mayores.

- **Peso (weight):** Indica el peso del individuo, con un rango de 97 a 366 lbs. La mediana y la media se acercan a 214 lbs, pero la media es ligeramente mayor (por 0.67), lo que evidencia la presencia de valores extremos y podría indicar casos de obesidad en la muestra.

- **Altura (height):** Corresponde a la altura del individuo, que varía entre 53 y 83 in. La mediana y la media se aproximan a 67 pulgadas, siendo la media ligeramente superior (por 0.24), mostrando una distribución relativamente simétrica con algunos valores extremos.

- **Número de cirugías mayores realizadas:** Refleja la cantidad de cirugías mayores que ha tenido la persona, con un rango de 0 a 16. La mediana y la media se sitúan alrededor de 4 cirugías, con la media ligeramente mayor (por 0.171). Esto indica que algunos individuos han tenido un número significativamente alto de cirugías, ya que el tercer cuartil es 6 y el valor máximo alcanza 16.

- **Colesterol:** Mide la cantidad de colesterol en sangre, con valores entre 83 y 351 mg/dl. La mediana y la media se aproximan a 199 mg/dl, aunque la media es ligeramente mayor (por 0.7). Según la Fundación del Corazón (2024), se considera colesterol alto a partir de 240 mg/dl, mientras que la hipercolesterolemia se define para niveles superiores a 200 mg/dl. Esto indica que la mayoría de la muestra presenta niveles de colesterol elevados.

2.4. Distribución de las variables cuantitativas

A continuación se presenta la distrubición por género de cada una de la variables cuantitativas presentes en la base de datos.

a)-Edad (age)

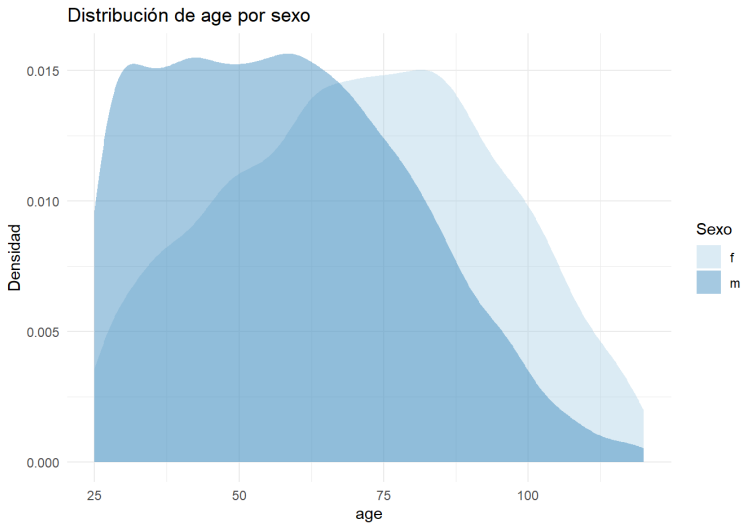

Figura 2: Elaboración propia, datos recopilados de la base

En este gráfico podemos, notar que la mayor concentración de muertes están los primeros años, aproximadamente de los 20 a los 75 años, después de los 75 años menos personas mueren, esto puede ser porque conforme más aumenta la edad, menos personas llegan a estar vivos.

b)-Colesterol (cholesterol)

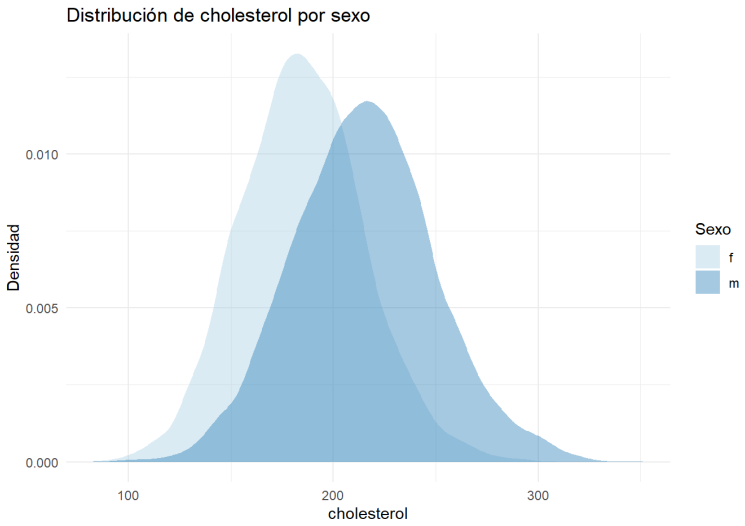

Figura 3: Elaboración propia, datos recopilados de la base

Se puede notar que el grupo de las mujeres, tienden estar más concentrados en valores medios-altos de colesterol, mientras que los hombres tienen una distribución más dispersa, aunque siempre presenta un pico en los valores de colesterol cercanos a 80.

c)-Bebidad por semana (drinks_aweek)

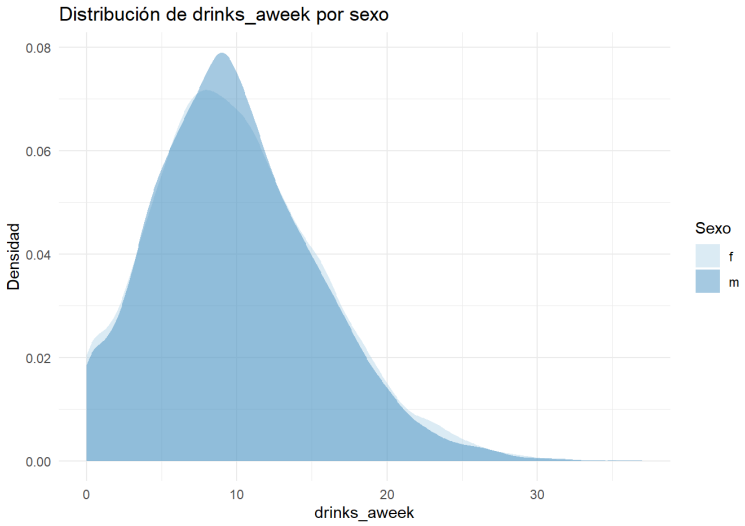

Figura 4: Elaboración propia, datos recopilados de la base

Para ambos grupo, femenino y masculino, se tienen 3 picos principales, los dos primeros picos son alrededor de las 10 bebidas a la semana, y el tercer pico alrededor de las 15 bebidas. También podemos observar que la mayor parte de las personas consumen entre 0 y 15 bebidas y que para valores muy cercanos a las 30 bebidas los que predominan son los hombres, y esto sorprende porque en la base hay más mujeres que hombres.

d)-Altura (height)

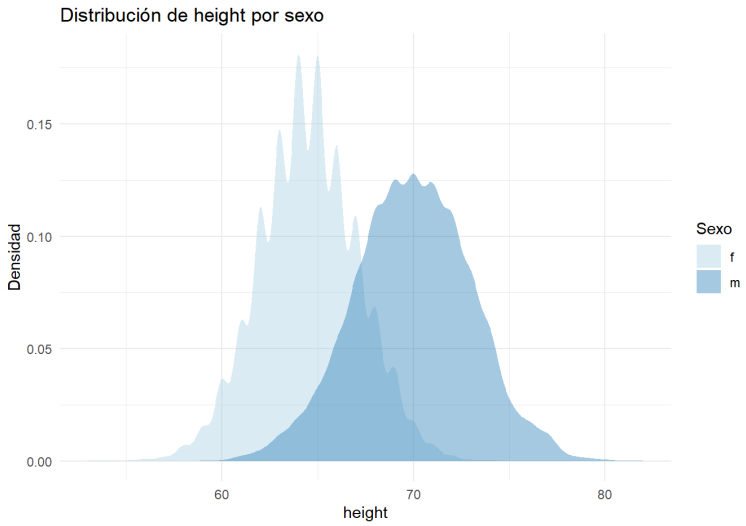

Figura 5: Elaboración propia, datos recopilados de la base

Acá lo que se observa es que la mayoría de las mujeres se concentran dentro de las 65 pulgadas, mientras que los hombres están cercanos a las 70 pulgadas. Después de estos picos, la cantidad de persona baja cada vez más rápido.

e)-Cirugía no ambulatorias (major_surgery_num)

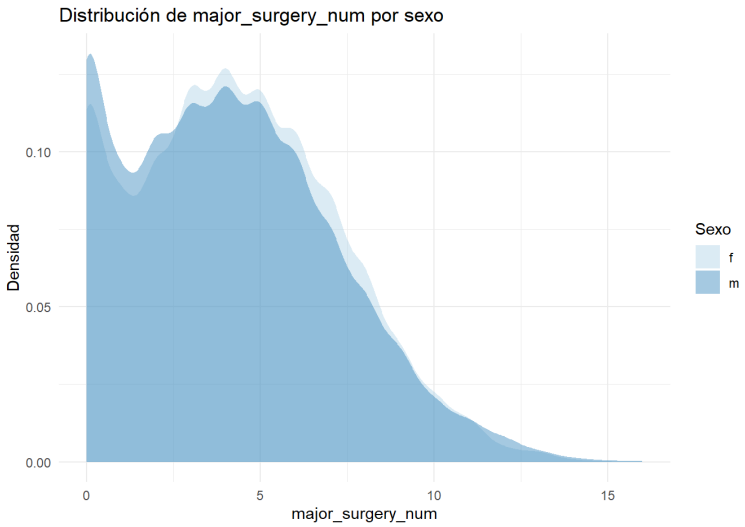

Figura 6: Elaboración propia, datos recopilados de la base

Para ambos grupos se puede observar que el pico principal se sitúa en 0 cirugías. Posterior a este pico, se observa que la mayoría de los individuos se sometieron a entre 4 y 6 cirugías.

e)- Cantidad de medicamentos (num_meds)

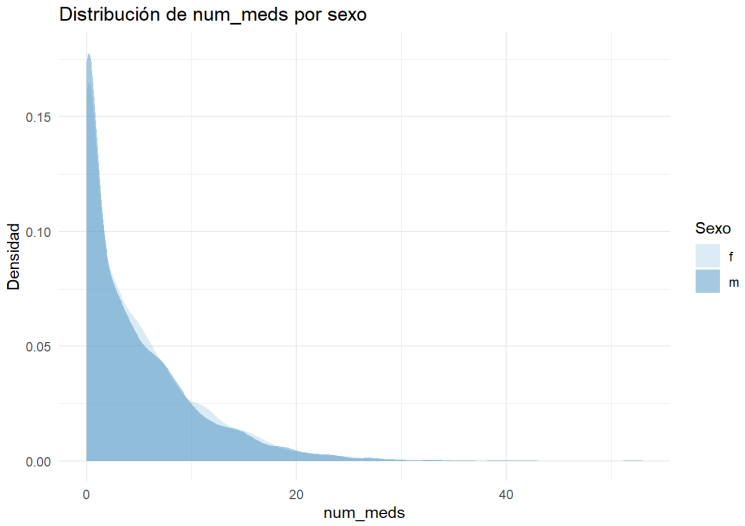

Figura 7: Elaboración propia, datos recopilados de la base

En ambos grupo se observa que la mayor parte de las personas, no tomaban ningún tipo de medicamento,y conforme aumenta el número de medicamentos, menor es la cantidad de personas que toman esa cantidad de medicamentos, es decir tiene una tendencia decreciente respecto cuantas personas toman más pastillas.

f)-Presión Sistólica (sys_bp)

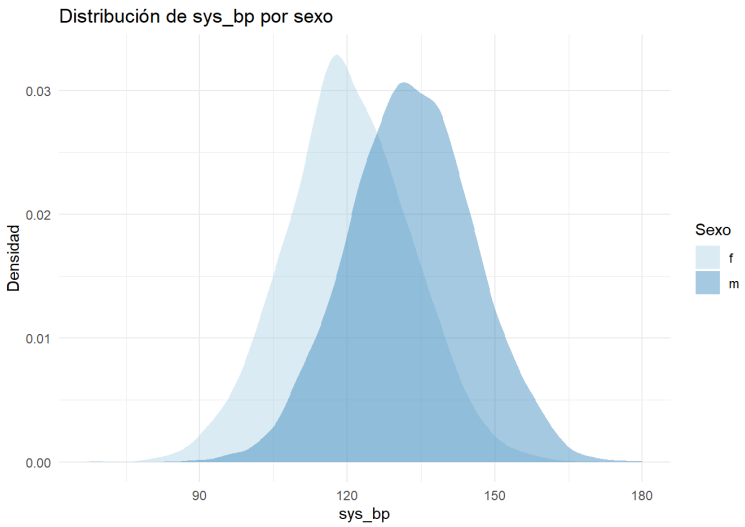

Figura 8: Elaboración propia, datos recopilados de la base

Se puede notar que el en el grupo de las mujeres, la mayor parte de las mujeres, están en un rango medio-alto de presión sistólica, al igual que los hombres, la diferencia es que el pico de los hombres está en valores mas altos, que el pico de las mujeres.

g)- Peso (Weigth)

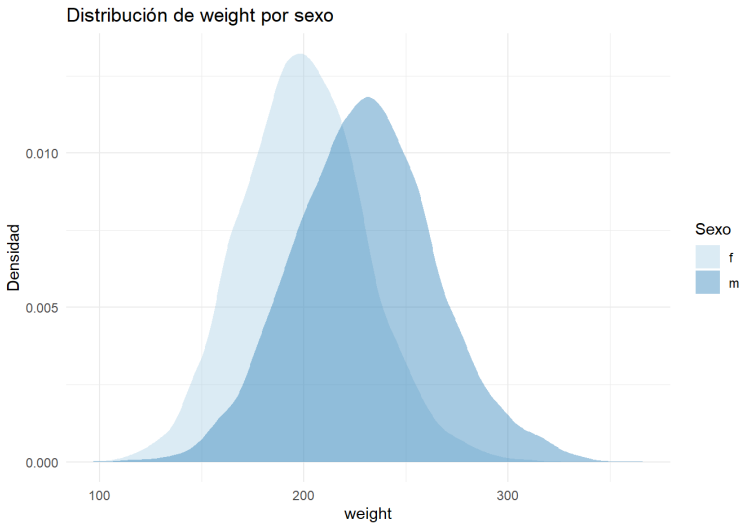

Figura 9: Elaboración propia, datos recopilados de la base

Acá se observa que ambos grupos tienen un pico muy similar, pero en la concentración podemos ver que el grupo de las mujeres, hay más densidad a la izquierda del pico, comparado con el lado derecho, y en el grupo de los hombres pasa al revés, hay más concentración al lado derecho del pico, aunque después de aproximadamente 250 libras, cae bastante.

2.5. Gráficos con variables relacionadas

En esta sección vamos a observar gráficos con variables relacionadas entre si.

a)-Colesterol según antecedentes familiares

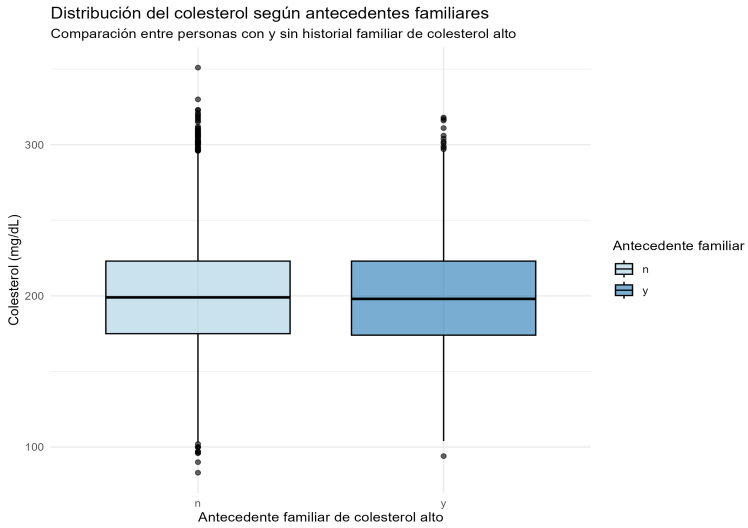

Figura 10: Elaboración propia, datos recopilados de la base

En este gráfico se puede observar que para ambos grupos, la media es prácticamente la misma, con valores cercanos a los 200 mg/dL, lo cual nos puede indicar que el hecho de tener o no tener antecedentes familiares es irrelevante. También se ve que ambos grupos presentan outliers por encima de 300 y por debajo de los 100 mg/dL

b)-Presión arterial según hábito de fumar

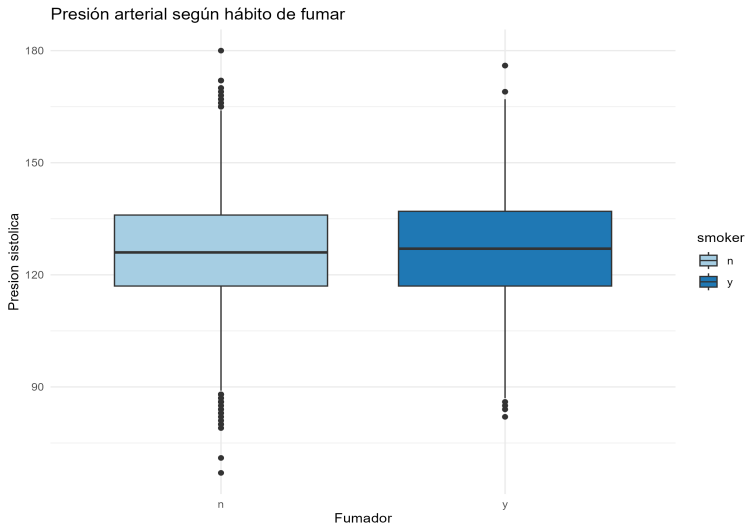

Figura 11: Elaboración propia, datos recopilados de la base

En este gráfico pasa algo similar al gráfico donde se compara el colesterol y los antecedentes familiares; ambos grupos tienen una mediana similar, lo que nos puede indicar que el factor del fumado no influye en la presión arterial. Aún se tiene la presencia de valores atípicos por encima y por debajo del rango intercuartílico.

c)- Edad según el nivel ocupacional

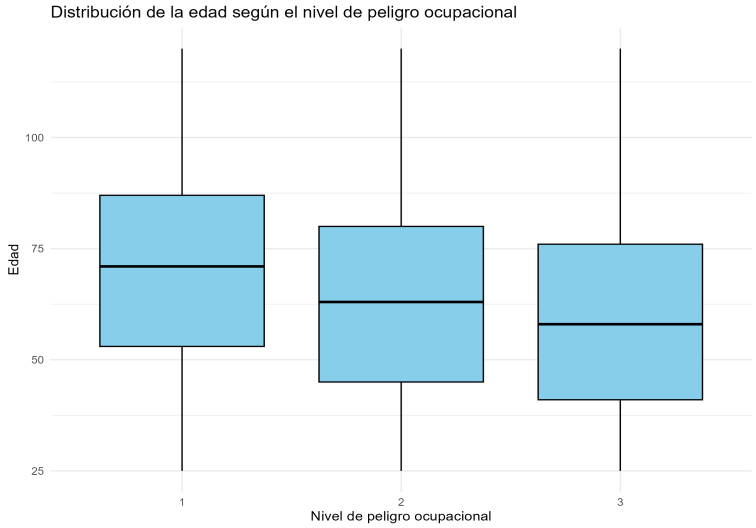

Figura 12: Elaboración propia, datos recopilados de la base

Este es un gráfico interesante, ya que las personas que tienen un trabajo más riesgoso tienden a tener una mediana más baja que las personas que tienen un trabajo de riesgo moderado, y así mismo, las personas que tienen un trabajo con riesgo moderado tienen una mediana menor a la de las personas que tienen los trabajos menos riesgosos. En este gráfico no se muestra la presencia de outliers.

2.6. Distribución de las variables categóricas

Para esta sección se muestran 2 gráficos que representan la distribución de algunas variables categóricas.

a)- Diabetes (diabetes)

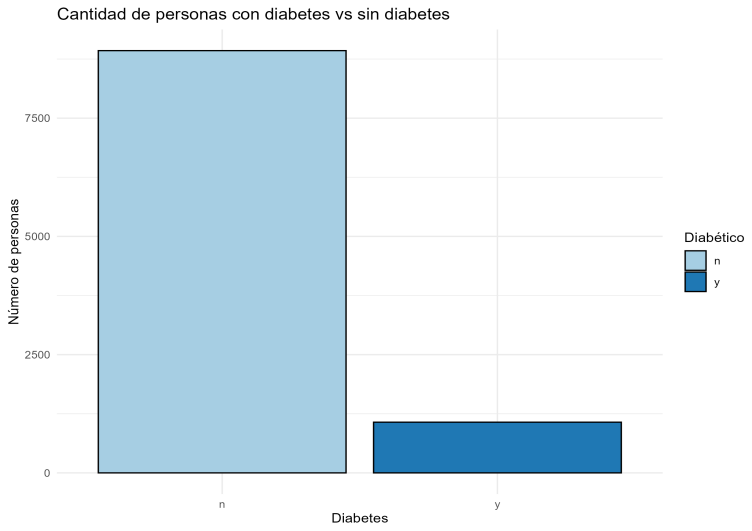

Figura 13: Elaboración propia, datos recopilados de la base

Esta es la distribución de las personas que tienen o no diabetes; se observa que en la población estudiada hay más personas sin diabetes que personas que sí tienen.

a)- Fumador (smoker)

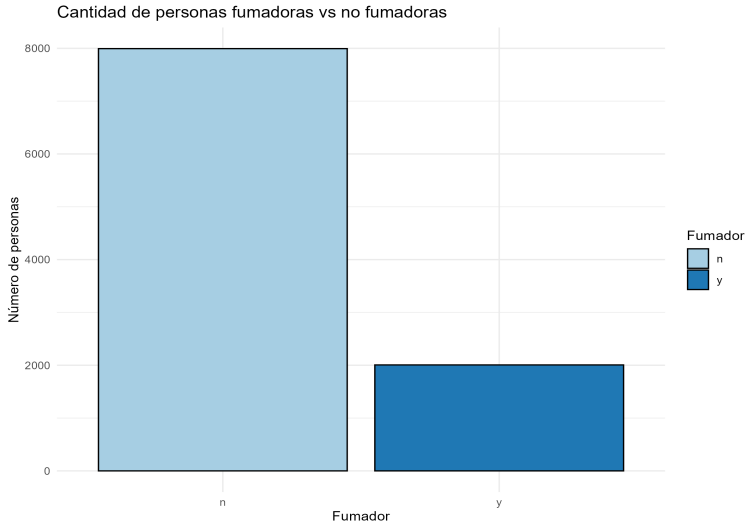

Figura 14: Elaboración propia, datos recopilados de la base

Acá se observa que aproximadamente 2000 personas son fumadores, mientras que unas 8000 personas no fuman.

2.7. Identificación de valores faltantes y posibles outliers

Se llevó a cabo una revisión de la base de datos con el objetivo de identificar posibles valores faltantes o atípicos que pudieran afectar el análisis. Tras este, se concluyó que la base de datos no contiene valores faltantes, lo que garantiza que todas las observaciones estén completas. Sin embargo, hablaremos de como se pueden identificar ambos tipos de valores.

Según Kwak y Kin (2017) dicen que hay varias formas para lograr identifcar outliers en una base de datos. Algunas técnicas que se mencionan son:

- Puntaje Z (Z-score): Se considera outlier cualquier valor que se encuentre a más de 3 desviaciones estándar de la media.

- Rango Intercuartílico (IQR): Se calcula el IQR como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1). Los valores fuera del rango definido por Q1 - 1.5IQR y Q3 + 1.5IQR se consideran outliers.

- Desviación Absoluta Mediana (MAD): Es una medida robusta que calcula la mediana de las desviaciones absolutas respecto a la mediana de los datos. Valores que se desvían significativamente de esta mediana pueden ser considerados outliers.

- Pruebas Formales: Pruebas como el test de Grubbs o el test de Dixon son útiles para detectar outliers en muestras pequeñas.

- Métodos Visuales: Herramientas como diagramas de caja (boxplots), gráficos de dispersión y histogramas ayudan a visualizar y detectar outliers de manera intuitiva.

- Métodos Basados en Aprendizaje Automático: Algoritmos como Isolation Forest y Local Outlier Factor (LOF) identifican outliers al analizar patrones y densidades en los datos.

Por otra parte, según los autores Kwak y Kim (2017), no todos los valores faltantes son iguales. Los valores perdidos pueden clasificarse como:

- MCAR (Missing Completely at Random): los datos faltan de forma completamente aleatoria, sin relación con ninguna otra variable u observación.

- MAR (Missing at Random): la ausencia de datos puede estar correlacionada con otras variables observadas, pero no con el valor que falta en sí mismo.

- MNAR (Missing Not at Random o NMAR): el valor faltante está relacionado con él mismo; es decir, la probabilidad de faltar depende del valor no observado.

Para evaluar la distribución de las variables cuantitativas que se utilizarán en el estudio y detectar posibles anomalías, se usarán métodos visuales, el cual se escogió un diagrama de caja y bigotes. Este gráfico permite identificar valores que se alejan significativamente del resto de la muestra y que podrían considerarse como atípicos, ya sea por errores en la recopilación de datos o por características de algunos individuos.

Como se puede observar en la Figura 15, todas las variables cuantitativas presentan algunos valores atípicos, con la excepción de la variable edad y no muestra desviaciones extremas. Dado que los valores atípicos pueden influir en medidas de tendencia central y dispersión, se aplicarán técnicas de limpieza y ajuste de estos datos a estas variables con el fin de garantizar que los resultados del estudio se basen en información más confiable y representativa.

Figura 15

Identificación de posibles outliers por variable cuantitativa

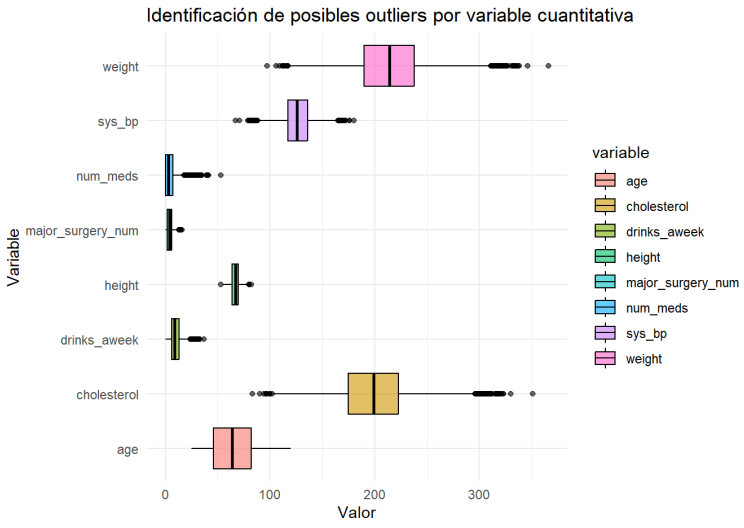

Figura 15: Elaboración propia, datos recopilados de la base

2.8. Técnicas para subsanar valores perdidos y outliers

Al trabajar con bases de datos, es de gran importancia tanto tener información completa como verídica que refleje la realidad.

Entre los mayores obstáculos para conseguir los mejores resultados cuando se aplican modelos sobre bases de datos, están los valores perdidos que consisten en datos faltantes y los outliers que son valores atípicos, es decir registros que se desvían del ”patrón”general que siguen los datos de un conjunto. En numerorsos casos, tener ausencia de datos o tener los datos incorrectos puede afectar, pues muchos modelos no son aceptados y eso puede ocasionar que el modelo falle. Por otro lado, los outliers pueden distorcionar las medidas estadísticas lo cual puede afectar la presición de un modelo. Por ende, es de suma importancia aprender a subsanar estos valores.

Según Kwak y Kim (2017) explican una serie de estrategias que funcionan para trabajar los valores atípicos, asi como con cuales tipos debería de aplicarse cada técnica.

Entre las estrategias que existen para manejar estos valores, la eliminación consiste simplemente en borrar las filas que contienen valores atípicos. A pesar de que es ventajoso porque es fácil de implementar y evita que esos datos afecten cálculos de promedios, desviaciones o modelos, al eliminarlos se pierde información y con ello se pueden perder datos importantes, sesgando así la muestra si los outliers representan casos extremos reales. Esta técnica podría ser muy útil cuando hay certeza de que son errores de registro o datos incorrectos.

Otra estrategia para manejar outliers es la estimación robusta, que consiste en emplear valores menos sensibles a valores extremos. Entre estos se incluyen los estimadores M (M-estimators), el Huber loss, estimadores basados en medianas en lugar de medias y

regresiones robustas. En contextos de regresión, es posible aplicar métodos como Huber Regression, RANSAC o Theil-Sen, los cuales tratan mejor la presencia de valores atípicos que la regresión lineal ordinaria.

Cuando un valor extremo se identifica claramente como un error de registro se puede optar por la imputación del outlier. Esto implica reemplazar el valor por uno más razonable, como la mediana de la variable o un valor estimado por un modelo, suavizando así la influencia de este sin eliminar completamente la observación.

Otra alternativa consiste en aplicar transformaciones a los datos, como logaritmos, raíces cuadradas o la transformación de Box-Cox, con el fin de reducir el impacto de los valores extremos al "compactar" la escala de los datos, haciendo que los outliers resulten menos pronunciados.

También, si los valores atípicos representan un subconjunto distinto dentro de la muestra, es posible tratarlos de manera diferente mediante la segmentación o modelado por subgrupos, ajustando modelos específicos para estos casos y evitando que distorsionen los resultados del conjunto de datos principal. Sin embargo, este método es más complicado pues se requiere de un estudia por cada caso que se considere diferente.

Finalmente, la estrategia de winsorización consiste en reemplazar los outliers por el valor más cercano dentro del rango permitido, es decir, el límite inferior o superior según la regla 1.5 x IQR. Este método tiene la ventaja de mantener todas las filas y reducir el impacto de valores extremos en modelos. Sin embargo, sí existe la desventaja de que se están modificando los valores originales, por lo que sigue habiendo posibilidad de sesgar la muestra. Puede ser muy útil cuando los outliers son muy extremos, pero se desea conservar la totalidad de la muestra.

Ahora para los valores faltantes también estos autores habalan acerca de estrategias de valores faltantes.

La eliminación de casos, al igual que con outliers, también funciona para valores ausentes. Eliminar casos ("complete case analysis") que consiste en eliminar las filas con valores faltantes, de modo que solo se analizan los casos completos. Esto es simple y garantiza que no haya datos faltantes, pero reduce el tamaño de la muestra y puede generar sesgo si los datos faltantes no son MCAR.

También, el análisis con casos disponibles ("available case analysis") consiste en hacer un análisis específico donde se usan los datos disponibles sin requerir que la fila esté completa en todas las variables. Esto permite conservar más datos, aunque el número de observaciones puede variar de variable a variable.

Finalmente, el método de imputar consiste en rellenar o estimar el valor faltante según algún criterio, de modo que se obtenga un conjunto de datos completo. Dentro de la imputación se distinguen:

- Imputación explícita: parte del supuesto de que las variables siguen una determinada distribución probabilística. A partir de esa distribución, se estiman sus parámetros (como la media, mediana, etc.) y se utilizan para reemplazar los valores faltantes. Entre las técnicas más habituales se incluyen la imputación por media, mediana, probabilidad, razón, regresiones (estima el valor faltante usando una ecuación de regresión basada en las demás variables), etc.

- Imputación implícita: se orienta a diseñar o aplicar un algoritmo que permita generar valores imputados de la forma más precisa posible, sin asumir una distribución concreta. Algunos ejemplos son la imputación hot-deck (reemplaza el valor faltante

con el de una observación "similar" dentro del mismo conjunto de datos), la imputación cold-deck (se toma el valor de una fuente externa como por ejemplo, de una base de datos diferente o de registros históricos).

3. Bitácora 3

3.1. Marco 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. Para ello, se utilizarán el coeficiente de correlación de Pearson para evaluar la fuerza y la dirección de la relación lineal entre variables numéricas, y la medida estadística eta cuadrado para determinar la magnitud de la asociación entre variables categóricas y cuantitativas. Esta sección detalla el tipo de estudio, la población y las técnicas de análisis de utilizadas.

El estudio es de carácter correlacional, ya que su propósito central es determinar la magnitud y dirección de la relación entre variables numéricas y cualitativas asociadas a factores de riesgo de muerte. Durante el proyecto se analizarán datos previamente recolectados de un estudio simulado con 10.000 entradas donde cada entrada representa un individuo fallecido entre los 25 y 120 años. Cada registro detalla información sobre los hábitos, condición de salud e información general del sujeto.

Entre las variables numéricas que se utilizarán se encuentran la edad, el peso, la altura, la presión arterial sistólica, el número de medicamentos que toma la persona, la cantidad de bebidas alcohólicas consumidas por semana, el número de cirugías mayores realizadas y el nivel de colesterol.

Por otro lado, las variables categóricas consideradas en el estudio son sexo, consumo de tabaco, consumo de cannabis u otras drogas, adicción, diabetes, antecedentes de ataques al corazón o derrames, asma, inmunodeficiencia, antecedentes familiares de cáncer,

enfermedades cardíacas y problemas de colesterol, así como si la ocupación de la persona se considera peligrosa.

El procedimiento se dividió en 6 etapas:

- **Importación y exploración de los datos en R:** en la que se verificaron la estructura, los tipos de variables y la presencia de valores faltantes o atípicos.
- **Análisis de las distribuciones:** donde se hicieron histogramas para analizar las variables cuantitativas por sexo.
- **Análisis de relaciones:** se hicieron boxplots para analizar las relaciones entre variables cuantitativas y categóricas.
- **Creación de las matrices de correlación:** se calcularon los coeficientes de eta al cuadrado y de Pearson para evaluar el nivel de relación entre las variables.
- **Construcción de los gráficos de correlación:** se elaboraron mapas de calor para observar con facilidad la intensidad y la dirección de dichas relaciones.
- **Elaboración de conclusiones y limitaciones:** después de interpretar los resultados, se elaboraron conclusiones a partir de ellos. Al igual que se detallaron las limitaciones dentro del estudio y observaciones finales.

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) dice que 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. Además 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. Entonces 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.

En este caso 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}.

Por otro lado, presenta algunos requisitos para que se pueda calcular y sea efectivo, la primera es que sea 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 tenemos 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 nos 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, 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.

Por lo que con estas dos técnicas para calcular coeficientes de correlación se 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 mag-

nitud de la relación entre variables categóricas y una variable cuantitativa, encontrando posibles efectos no lineales. Complementando ambas metodologías garantiza un análisis más preciso, asegurando que los resultados obtenidos sean coherentes, válidos y adecuados para responder a las preguntas planteadas en el estudio.

3.2. Resultados

A continuación se van a mostrar los mapas de calor de los dos métodos utilizados para analizar correlaciones.

1. Pearson

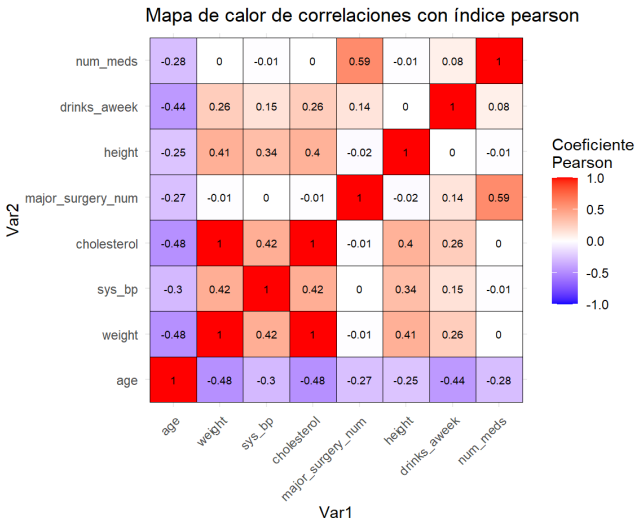

Figura 16: Elaboración propia, datos recopilados de la base

El mapa de calor obtenido muestra que:

- Altura y peso: Presentan una correlación positiva moderada, lo cual es esperado, ya que las personas con mayor estatura tienden a tener mayor peso corporal.
- Edad y número de cirugías mayores: Muestran una correlación positiva débil. Esto sugiere que conforme aumenta la edad, es más probable haber sido sometido a algún procedimiento quirúrgico, aunque la relación no es fuerte.
- Edad y presión sistólica: Presentan una correlación moderada, indicando que la presión arterial tiende a aumentar con la edad.

- Peso y presión sistólica: Muestran una correlación positiva débil a moderada, lo cual coincide con evidencia clínica sobre el impacto del peso en la presión arterial.

Las demás correlaciones cuantitativas resultaron muy bajas, indicando poca o nula dependencia lineal entre esas variables.

Además, se identificó que la relación entre colesterol y peso muestra un valor cercano a 1, indicando una asociación muy fuerte entre estas dos variables en esta base de datos. Este resultado sugiere que, dentro del conjunto simulado, el nivel de colesterol aumenta casi de manera proporcional al peso corporal.

En general, los coeficientes de Pearson sugieren que el conjunto de datos no presenta relaciones lineales fuertes entre la mayoría de las variables cuantitativas.

2. η^2

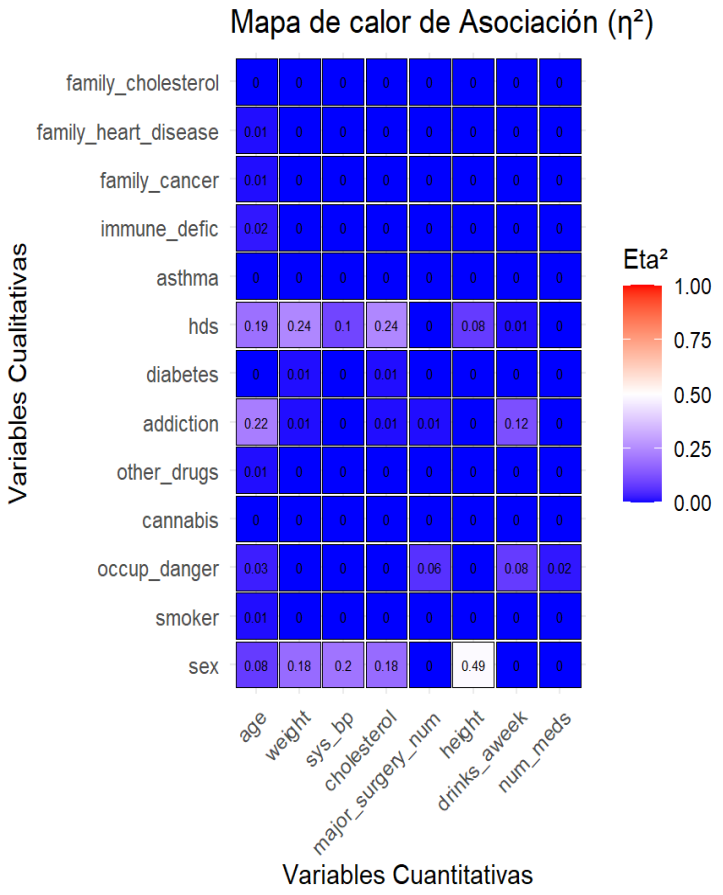

Figura 17: Elaboración propia, datos recopilados de la base

Los resultados del eta cuadrado muestran vínculos importantes entre los datos de un tipo y los de otro:

- El vínculo más marcado se encontró al comparar el género con la estatura, lo cual se alinea con lo que se espera por razones biológicas.
- También se notó una conexión de nivel medio entre el género y la masa corporal, algo que concuerda de nuevo con las variaciones físicas entre las distintas personas.

- Las características de salud que son por categorías (como tener diabetes, adicciones, problemas del corazón, historial familiar, tener un trabajo peligroso, etc.) generalmente tuvieron valores bajos de eta cuadrado. Esto quiere decir que ninguna de estas clasificaciones logra explicar una parte grande de los cambios en datos medibles como el colesterol, la tensión o cuántas medicinas toma alguien.
- El factor ocupación mostró algo de conexión con la edad, aunque fue leve, pareciendo confirmar lo que se vio antes en los gráficos de caja.

Esta evidencia sugiere que, aunque las clasificaciones tienen que ver con los comportamientos típicos, no influyen mucho en el nivel de las medidas numéricas estudiadas, lo que significa que la mayoría de las diferencias en los datos cuantitativos no se deben a estas clasificaciones que revisamos.

4. Limitaciones

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 a contextos clínicos.

Otro aspecto a considerar en relación a 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 a este par el trabajo.

También una limitación con el trabajo es que con el coeficiente de eta cuadrado y de Pearson cuenta con limitaciones para poder usarse de forma correcta, los cuales hay que tomar en cuenta para un buen análisis de los resultados. Por lo que esto implicó revisar que estos requisitos se cumplieran.

Por último, los conocimientos que tiene el grupo limita a la hora de poder realizar análisis más robustos, como una regresión lineal, para realmente poder entender y verificar que estas relaciones entre las variables existen.

5. Conclusiones

El análisis que se realizó permitió entender mejor cómo se relacionan las variables del conjunto de datos y qué tan fuertes son esas relaciones en términos de riesgo de mortalidad. A partir de lo encontrado, se pueden destacar varios puntos importantes:

- Las correlaciones en general fueron bajas, tanto entre variables numéricas como entre variables numéricas y categóricas. Esto significa que, dentro de esta base simulada, la mayoría de las variables no muestran relaciones lineales fuertes entre sí.

- Las asociaciones más evidentes aparecieron entre sexo–altura y sexo–peso, lo cual tiene sentido porque existen diferencias biológicas claras. Sin embargo, estas diferencias no representan por sí solas un mayor riesgo de mortalidad.

- Variables que normalmente se consideran factores de riesgo, como diabetes, problemas cardíacos o consumo de sustancias, no mostraron una relación fuerte con variables como presión arterial, colesterol o número de medicamentos. Esto probablemente se debe a que la base es simulada y no refleja completamente los patrones clínicos de poblaciones reales.

- Tanto Pearson como η^2 ayudan a tener un panorama general de las relaciones entre las variables, pero también queda claro que estos métodos no son suficientes para explicar un fenómeno tan complejo como la mortalidad. Para un análisis más completo 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. Esto 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.

6. Bibliografía

Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: a review. *International journal of psychological research*, 3(1), 58-67.

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.

Diccionario de cáncer del NCI. (s.f.). Cancer.gov. Recuperado de `https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/mortalidad`

Fonnesu, M., & Kuczewski, N. (2019). *Doubts on the efficacy of outliers correction methods*. arXiv preprint arXiv:1907.09864.

Fundación del Corazón. (2024). *Colesterol y riesgo cardiovascular*. Fundación del Corazón. Recuperado de `https://fundaciondelcorazon.com/prevencion/riesgo-cardiovascular/colesterol.html`

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`

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`

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.

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

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.

Sánchez Acero, F. (2024). Variables Cualitativas y Cuantitativas. 2024. Recuperado de https://repositorio.konradlorenz.edu.co/handle/001/6087