TRABAJO PRÁCTICO
DE
ESTADÍSTICA DESCRIPTIVA

Integrantes: Cristian Damián Fortunesky Barrios, Matias De Vivo Reynot, Juan Pablo Catalano y Joaquín Albeneda.

Asignatura: Probabilidad y Estadística para la gestión de datos

Profesora: Natalia Estigarribia

Fecha de entrega: 12/05/2024

Institución: IFTS24

Consignas:

En grupo, presentar un informe de un conjunto de datos a elección. Este último debe contener almenos dos variables cuantitativas y una variable cualitativa. No hay restricción de cantidad de registros pero que tenga al menos 200/300 casos.

En el informe deben describir de qué trata el conjunto de datos seleccionado. Describir cada una de las variables indicando el tipo.

Deben presentarse medidas, gráficos y tablas.

Cada uno de los puntos anteriores debe ir acompañado de una explicación o interpretación (NO deben poner cuadros y gráficos sueltos)

Los gráficos y cuadros deben tener título y número de cuadro.

Documentos a presentar:

1) PDF con el informe

2) Script con las consultas empleadas.

3) Conjunto de datos (en xls, txt, csv)

La base de datos elegida es de tipo epidemiológico a nivel mundial y está constituida por las siguientes variables:

Tipos de Variables Estadísticas:

- Variables Cualitativas Nominales: País y Estado
- Variables Cuantitativas Discretas: Año, Muertes infantiles, Sarampión y Muertes menores de cinco años.
- Variables Cuantitativas Continuas: Esperanza de vida, Mortalidad Adulta, Alcohol, Gasto de Porcentaje, Hepatitis B, IMC, Polio, Gasto total, Difteria, VIH SIDA, PIB, Población, Delgadez, Composición de ingresos de recursos y Escolarización.

Las variables a utilizar en este estudio son las siguientes:

- Cualitativas Nominales: Estado.
- Cuantitativas Continuas: Esperanza de vida, Mortalidad Adulta, hepatitis B, Polio, Difteria y VIH SIDA.

A lo largo de este trabajo, se enfocará particular atención al análisis univariado de las variables de Esperanza de Vida, Mortalidad Adulta y en el caso del análisis bivariado (crossplots) las correlaciones se hacen entre enfermedades.

GRÁFICOS:

HISTOGRAMA ESPERANZA DE VIDA (Figura 1):

- La mayor densidad del histograma se encuentra en el intervalo Q_1 (25%) y Q_3 (75%). Dentro de la distancia intercuartil Q_3-Q_1 puede decirse que la mayor concentración de datos está entre Q_2 y Q_3, ya que la distancia entre cuartiles es menor que en el caso Q_1-Q_2 y además Q_2-Q_3 contiene la moda. El 50% de los datos se encuentra entre 63 y 76 años (Q_1 y Q_3, respectivamente) mientras que el rango está comprendido entre 0 a 89 años.
- La distribución es unimodal.
- La distribución es asimétrica negativa porque la moda>mediana>media.
- La distribución es Platicúrtica.

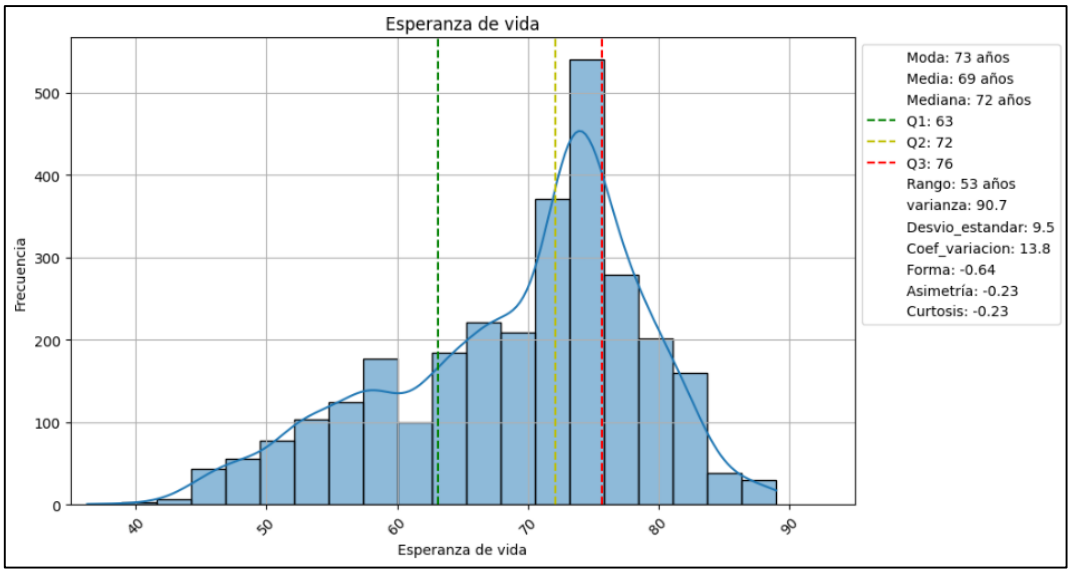

Figura 1: Histograma Eperanza de Vida

BOXPLOT ESPERANZA DE VIDA (Figura 2)**:**

Dentro del boxplot la caja contiene el 50% de los datos. Entre Q_1 y la mediana (Q_2) se encuentran el 25% de los datos y entre la mediana (Q_2) y Q_3 el restante 25% de los datos. Se observa una caja compacta y asimétrica, donde la mayor densidad de datos se encuentra entre los percentiles Q_2 y Q_3 que están relativamente cerca (donde, Q_2(72) y Q_3(76)) que contiene la moda coincidente con el intervalo de clase [73,76] del histograma “Esperanza de Vida”.

Entre los percentiles Q_1 y Q_2 la distancia es más grande que aquella entre Q_2-Q_3. A través del análisis del boxplot se confirma que la distribución es asimétrica negativa, como fue se justificado en el análisis del histograma (moda>mediana>media), porque los datos se concentran entre Q_2-Q_3 y el bigote Ls es de menor tamaño que Li. Los *outliers* se concentran por debajo del extremo inferior del bigote Li.

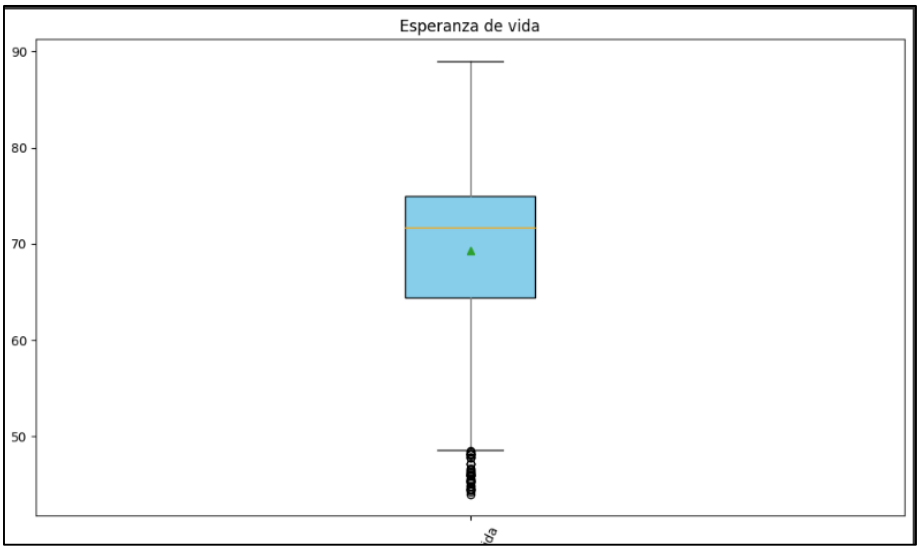

Figura 2: Boxplot Eperanza de Vida

El conjunto de datos seleccionados representa el rango de esperanza de vida humana, abarcando desde los 0 hasta los 89 años. La esperanza de vida promedio, aproximadamente de 69 años, sugiere que en las poblaciones representadas de todos los continentes, las personas tienen una expectativa de vida en torno a esa edad. Sin embargo, la desviación estándar de aproximadamente 9.52 años indica una variabilidad significativa en la esperanza de vida entre las diferentes poblaciones o regiones incluidas en el conjunto de datos.

La mediana de 72.1 años sugiere que el 50% de las observaciones tienen una esperanza de vida igual o superior a esa edad, mientras que el otro 50% tiene una esperanza de vida inferior a ese valor. Esta medida proporciona una comprensión más equilibrada de la distribución de la esperanza de vida en comparación con el promedio, ya que no se ve influenciada por valores extremos.

Por otro lado, la esperanza de vida mínima se encuentra en un rango de 33 años y la máxima de 89 años, la mayor frecuencia de esperanza de vida se encuentra entre los 63 y los 76 años siendo que la moda está comprendida entre el valor 72 y 73, q2 y q3 respectivamente (Figura 1).

Este amplio espectro refleja la diversidad de condiciones de vida, atención médica y otros factores que influyen en la longevidad de las poblaciones estudiadas.

En conjunto, estas estadísticas indican que existe una amplia variabilidad en la esperanza de vida entre las poblaciones representadas en los datos, con un promedio de alrededor de 69 años y una dispersión considerable alrededor de este valor.

MORTALIDAD ADULTA (Figura 3):

- La distribución está concentrada o tiene mayor densidad de datos, entre el mínimo (valor 0) y Q_3. Entre el mínimo=0 y Q3=228 están comprendidas las dos modas, dentro de una distribución con un rango entre 0 y 725.

- La distribución es bimodal. La primera moda tiene el intervalo de clase [0, 37] y la segunda moda [107, 144]. La primera moda se encuentra entre el mínimo y el Q1 y la segunda moda se encuentra entre Q1 y Q2.

- La distribución es Asimétrica Positiva porque la moda<mediana<media.

- La distribución es Platicúrtica.

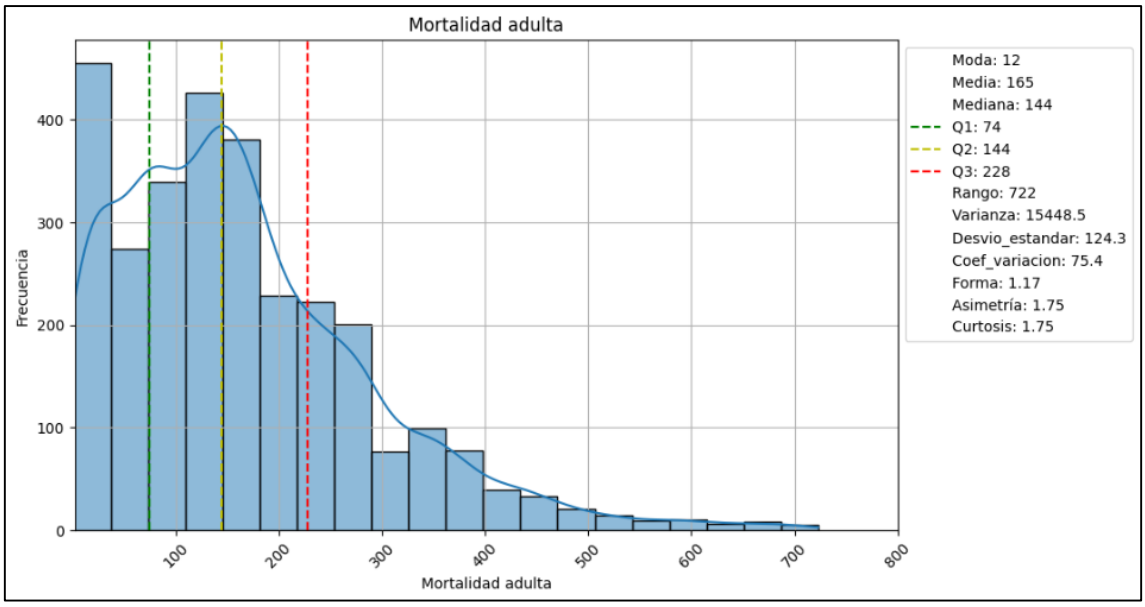

Figura 3: Histograma Mortalidad Adulta

BOXPLOT MORTALIDAD ADULTA (Figura 4):

La caja contiene el 50% de los datos. Dentro de esta, entre Q1 y la mediana (Q2) se encuentran el 25% de los datos y entre la mediana (Q2) y Q3 el restante 25% de los datos. Se observa una caja compacta y levemente asimétrica, donde la mayor densidad de datos se encuentra entre los percentiles Q1 y Q2 que están relativamente cerca (donde, Q1 (74) y Q2 (144)) que contiene la moda coincidente con el intervalo de clase [109, 144].

Entre los percentiles Q2 y Q3 la distancia es levemente más grande que entre Q1-Q2. A través del análisis del boxplot se confirma que la distribución es asimétrica positiva, como fue se justificado en el análisis del histograma (moda<mediana<media), porque los datos se concentran entre Q1-Q2 y el bigote Li es de menor tamaño que Ls. Los outliers se concentran por arriba del extremo superior del bigote Ls.

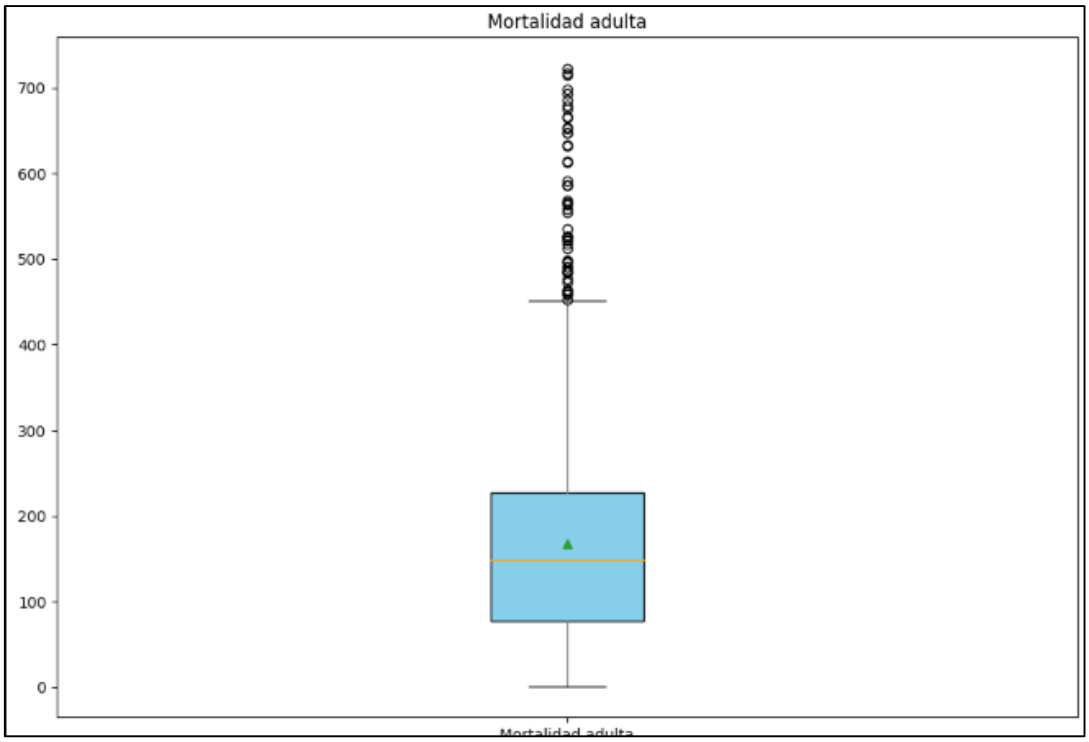

Figura 4: Boxplot Mortalidad Adulta

GRÁFICO DE TORTA - CLASIFICACIÓN DE ESTADOS:

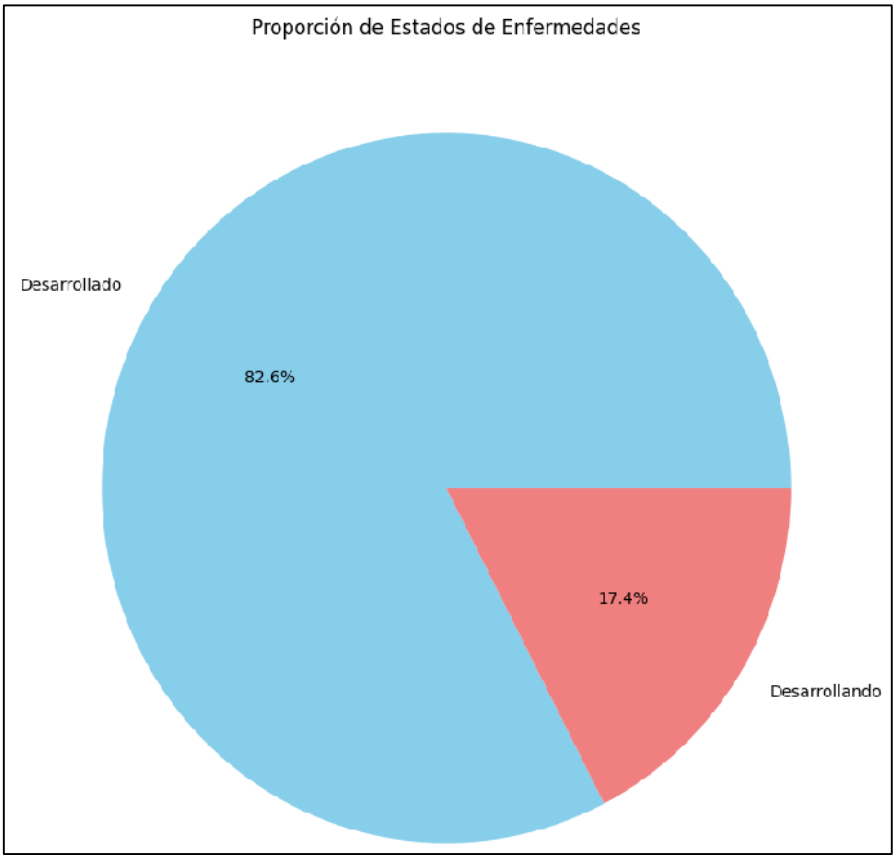

Figura 5: Gráfico de torta, proporción de Estados con Enfermedades

El gráfico de torta muestra la distribución de los estados de enfermedades en dos categorías: "Desarrollado" y "Desarrollando". A continuación, se presentan los resultados:

- Desarrollado: 82.6% de los casos se encuentran en este estado, lo que indica que la mayoría de las enfermedades se encuentran en una fase avanzada de desarrollo.

- Desarrollando: 17.4% de los casos se encuentran en este estado, lo que sugiere que una minoría de las enfermedades se encuentran en una fase inicial de desarrollo.

GRÁFICO DE BARRAS - ESPERANZA DE VIDA POR CONTINENTE (Fig. 6):

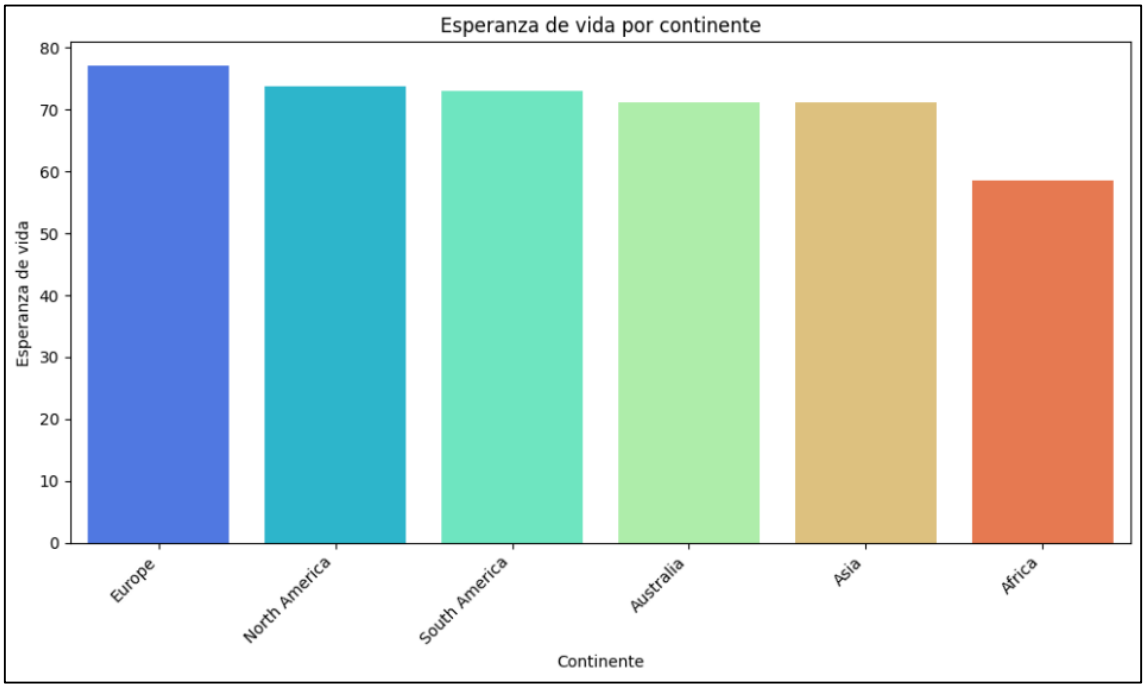

Figura 6: Gráfico de barras **-** Esperanza de Vida por Continente

GRÁFICO DE TORTA - ESPERANZA DE VIDA POR CONTINENTE (Fig. 7):

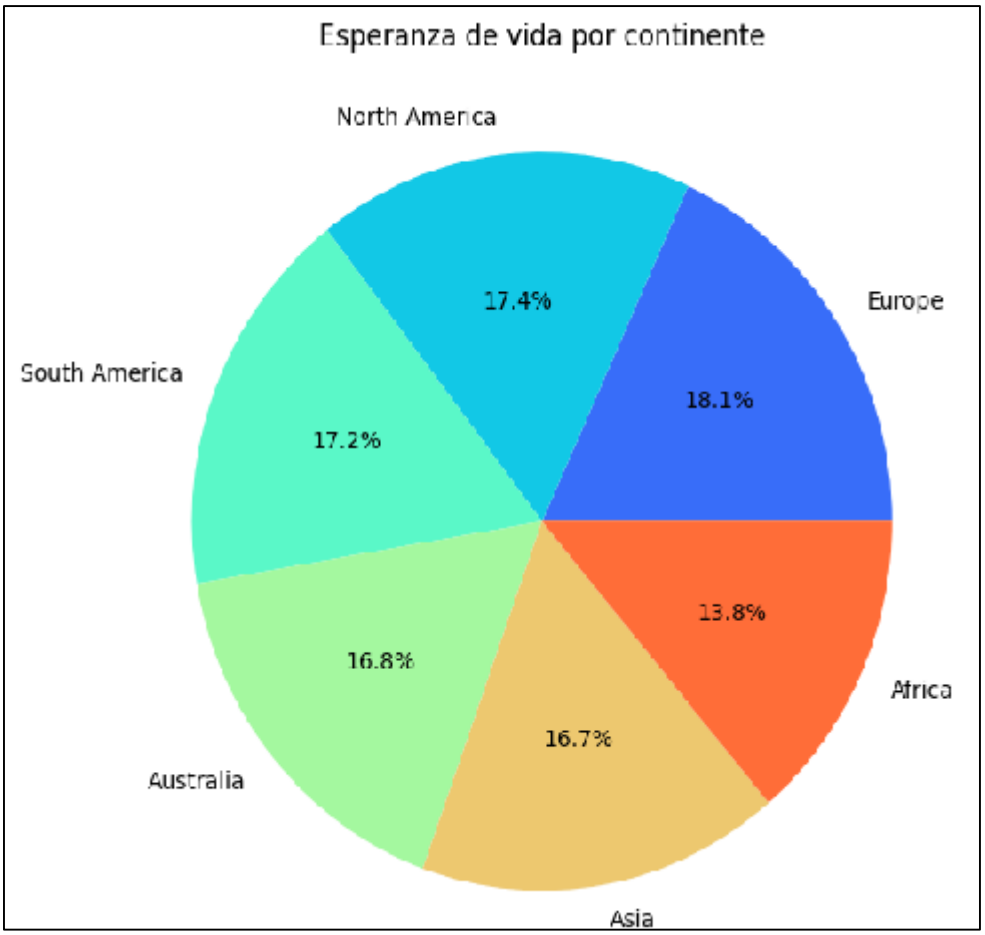

Figura 7: Gráfico de torta - Esperanza de Vida por Continente

El análisis de los gráficos que representan la distribución porcentual de la esperanza de vida por continente revela varias interpretaciones significativas:

En primer lugar, se observan notables disparidades en la esperanza de vida entre los diferentes continentes. Europa exhibe la mayor esperanza de vida con un porcentaje del 18.1%, seguido de Norteamérica con un 17.4%, mientras que África registra el porcentaje más bajo con un 13.8%.

Estas diferencias subrayan la importancia de las políticas de salud pública y los sistemas de atención médica en cada región, así como las disparidades socioeconómicas y ambientales que pueden influir en la salud y el bienestar de la población.

Además, el análisis del gráfico sugiere posibles tendencias demográficas, como una población envejecida en Europa y Norteamérica en contraste con poblaciones más jóvenes en África y Asia.

En última instancia, el gráfico de torta proporciona una valiosa perspectiva para explorar temas relacionados con la equidad global, la salud pública y el impacto de factores socioeconómicos y ambientales en la esperanza de vida a nivel mundial.

ANALISIS DE CORRELACIÓN:

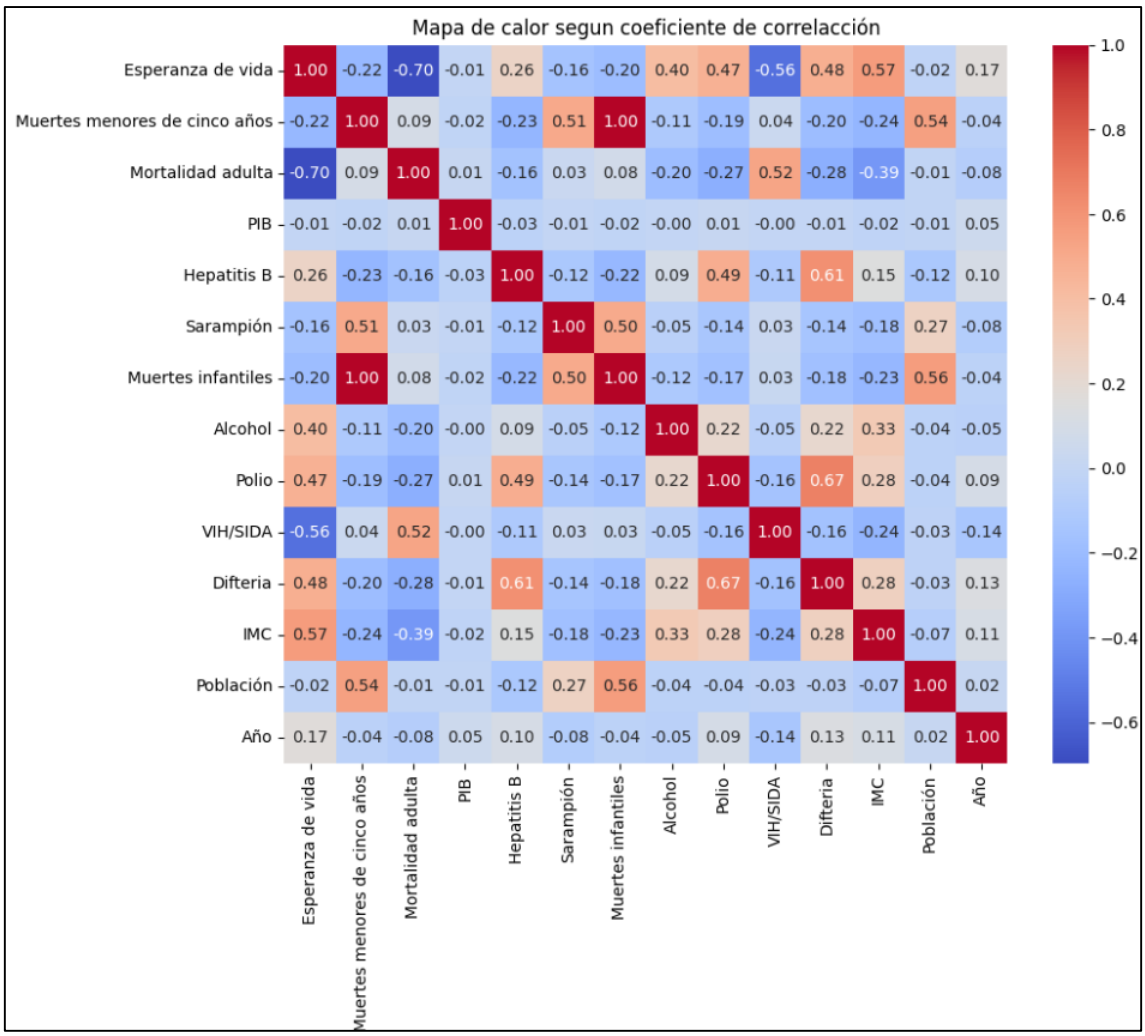

Figura 8: Mapa de calor. Los colores azules representan las correlaciones más bajas entre variables y los rojos las más altos. El grado de correlación entre variables además está representado por el coeficiente de Pearson

Utilizando los coeficientes de Pearson, se consideran los siguientes *cutt offs*: [0-4] (bajo), moderado (4-7] y alto >7.

Utilizando el coeficiente de correlación de Pearson para determinar la dependencia lineal entre nuestras variables cuantitativas, podemos observar:

Correlación entre Esperanza de vida y IMC:

- Existe una correlación positiva moderada de 0.542 entre la esperanza de vida y el índice de masa corporal (IMC). Esto sugiere que hay una relación positiva entre la esperanza de vida y el IMC, aunque moderada.

Correlación entre Esperanza de vida y Alcohol:

- Hay una correlación positiva moderada de 0.402 entre la esperanza de vida y el consumo de alcohol. Esto indica que la esperanza de vida tiende a aumentar ligeramente con el consumo moderado de alcohol.

Correlación entre Esperanza de vida y Polio:

- Se observa una correlación positiva baja de 0.327 entre la esperanza de vida y la incidencia de polio. Esto sugiere que la presencia de casos de polio puede influir modestamente en la esperanza de vida en una población.

Correlación entre Esperanza de vida y Difteria:

- También se evidencia una correlación positiva baja de 0.341 entre la esperanza de vida y la incidencia de difteria. Esto indica que la presencia de casos de difteria puede tener un impacto similar en la esperanza de vida.

Correlación entre Muertes menores de cinco años y Población:

- Existe una correlación positiva moderada de 0.658 entre las muertes de niños menores de cinco años y la población. Esto sugiere que en áreas con una población más grande, es probable que haya más muertes en niños menores de cinco años.

Correlación entre Sarampión y Muertes menores de cinco años:

- Se observa una correlación positiva moderada de 0.517 entre la incidencia de sarampión y las muertes de niños menores de cinco años. Esto indica que el sarampión puede ser un factor contribuyente a las muertes en este grupo de edad.

Correlación entre Sarampión y Muertes infantiles:

- También se evidencia una correlación positiva moderada de 0.532 entre la incidencia de sarampión y las muertes infantiles en general. Esto sugiere que el sarampión puede tener un impacto significativo en las tasas de mortalidad infantil.

Correlación entre Mortalidad adulta y VIH/SIDA:

- Hay una correlación positiva moderada de 0.550 entre la mortalidad adulta y la prevalencia de VIH/SIDA. Esto indica que la presencia de casos de VIH/SIDA puede influir en las tasas de mortalidad en adultos.

Correlación entre Hepatitis B y Polio:

- Se observa una correlación positiva moderada de 0.463 entre la incidencia de hepatitis B y la de polio. Esto sugiere cierta relación entre estas enfermedades infecciosas.

Correlación entre Hepatitis B y Difteria:

- También se evidencia una correlación positiva moderada de 0.588 entre la incidencia de hepatitis B y la de difteria. Esto sugiere una relación más robusta entre estas dos enfermedades.

Correlación entre IMC y Alcohol:

- Existe una correlación positiva baja de 0.353 entre el índice de masa corporal (IMC) y el consumo de alcohol. Esto indica que puede haber alguna influencia del consumo de alcohol en el IMC, pero la relación es débil.

Correlación entre IMC y Polio:

- Hay una correlación positiva baja de 0.186 entre el índice de masa corporal (IMC) y la incidencia de polio. Esto sugiere que puede haber una ligera asociación entre el IMC y la incidencia de polio, pero la relación es débil.

Correlación entre Hepatitis B y Esperanza de vida:

- Se observa una correlación positiva baja de 0.199 entre la incidencia de hepatitis B y la esperanza de vida. Esto sugiere que la presencia de casos de hepatitis B puede tener un impacto modesto en la esperanza de vida de una población.

ANALISIS DE DISPERSIÓN (Figura 9):

Para el análisis de los gráficos de dispersión se utilizan los coeficientes de correlación de Pearson calculados para la Figura 8.

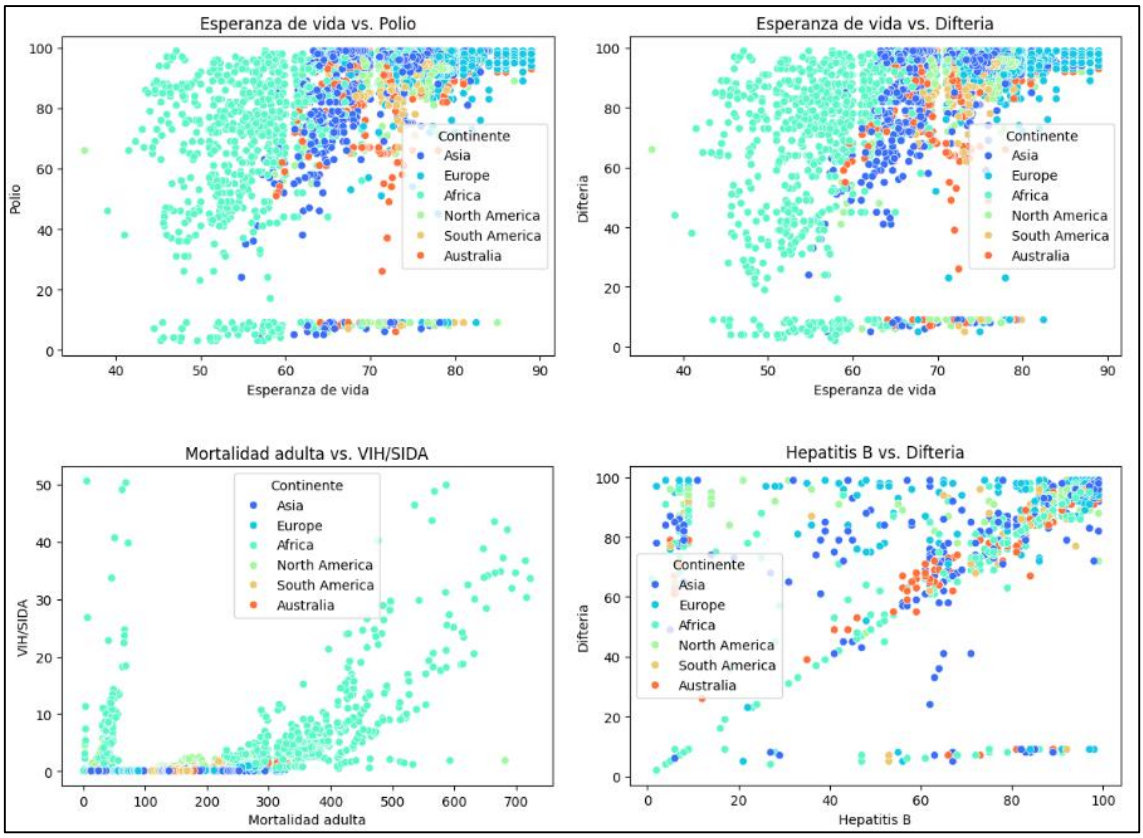

Figura 9: Correlación entre distintas enfermedades filtrada por continentes (esta tercer variable está en color)

Esperanza de vida vs. Polio / Difteria:

- Ambas correlaciones con la esperanza de vida (0.33 para Polio y 0.34 para Difteria) son positivas pero relativamente bajas. Esto sugiere que la incidencia de estas enfermedades puede estar relacionada con la esperanza de vida, pero otros factores también pueden influir significativamente.

Mortalidad adulta vs. VIH/SIDA:

- Hay una correlación positiva moderada (0.55) entre la mortalidad adulta y el VIH/SIDA. Esto indica que los países con una mayor tasa de incidencia de VIH/SIDA tienden a tener también una mayor mortalidad adulta.

Hepatitis B vs. Difteria:

- La correlación Hepatitis B y Difteria (0.588) son positivas y moderadas. Esto sugiere que la incidencia de Hepatitis B puede estar asociada con la incidencia de estas enfermedades, pero nuevamente, otros factores pueden influir.

Los análisis de dispersión basados en el coeficiente de correlación de Pearson revelan patrones significativos en la relación entre diversas variables y la esperanza de vida, dividida por continentes.

En primer lugar, al comparar la esperanza de vida con la incidencia de la polio y la difteria, se observa que África y Asia tienden a tener las menores esperanzas de vida, lo que sugiere una posible correlación entre la incidencia de estas enfermedades y la salud general de la población en estas regiones. Por otro lado, Europa muestra consistentemente una esperanza de vida más alta en relación con la incidencia de estas enfermedades, lo que podría indicar mejores sistemas de salud y prevención en este continente.

En el tercer análisis, que compara la mortalidad adulta con la prevalencia de VIH/SIDA, se destaca que África exhibe una mortalidad adulta notablemente más alta en comparación con otros continentes, lo que sugiere una relación entre la carga de enfermedades transmisibles, como el VIH/SIDA, y la mortalidad en esta región. Este hallazgo subraya la urgencia de intervenciones médicas y programas de salud pública en África para abordar esta preocupante disparidad.

Finalmente, al examinar la relación entre la incidencia de hepatitis B y la difteria, se observa que en África existe una correlación más lineal entre ambas variables, mientras que en Asia la relación es más dispersa en diferentes edades. Esto podría indicar diferentes patrones de exposición a estas enfermedades o variaciones en la eficacia de las políticas de vacunación entre los continentes.

En conjunto, estos análisis resaltan la importancia de comprender las interrelaciones entre la salud, las enfermedades infecciosas y los factores socioeconómicos en diferentes regiones del mundo, así como la necesidad de implementar estrategias efectivas para mejorar la salud y la calidad de vida en áreas donde las tasas de enfermedad y mortalidad son más altas.