10K Challenge: Explorer

Reporte

Índice

Contenido

Introducción: Estrategias para aumentar la esperanza de vida

La base de datos seleccionada para este reporte se titula “Life expectancy”, la cual contiene datos registrados por la Organización Mundial de la Salud entre los años 2000 y 2015 referente a la esperanza de vida y diversos factores sociales, económicos y de salud que impactan en ella. La esperanza de vida depende de la situación demográfica en la que se encuentre cierta población en un tiempo determinado y se refiere al número de años promedio que se espera viva una persona. Siendo que, mientras mayor sea su esperanza de vida, indica un mejor desarrollo económico y social en la población (INEGI, 2020).

A lo largo de este reporte se buscará encontrar una relación entre la esperanza de vida y las demás variables que vienen registradas a través de análisis estadísticos y visualizaciones de datos según sea el caso. La base registrada corresponde a 183 países y tiene un panorama de hasta 22 columnas, variables que pueden ser dispuestas a análisis. Por otro lado, se encontró relevante encontrar estrategias que sirvan a los países a incrementar la esperanza de vida media de sus

ciudadanos. Esto de acuerdo con los factores que generen mayor impacto en la misma. Si bien la situación económica que un país o una sociedad tiene influye en cuánto dinero estos pueden invertir en salud, escolaridad, hogar digno y alimentación; esto no necesariamente es igual en todos los casos ni representa la única manera de conseguir una vida longeva.

Créditos

Nombre y puesto de los integrantes. (Analista, programador, data scientist…)
Sofia Alvarez Sandoval (data engineer, programadora principal y documentadora)
David Alejandro Matamoros Alvarado (programador secundario y data scientist)
Erandi Abigail Ramírez Muñoz (data scientist y analista)
Cynthia Cristal Quijas Flores (analista de datos y investigadora)

Etapa 1: Exploración

Descripción de los datos

La base de datos se compone de 22 columnas que describen, de acuerdo con 183 países, características que generan un impacto en la esperanza de vida de la población de dichas regiones.

Entre las características previamente mencionadas, las cuales igual representan las variables a tratar, se pueden encontrar:

- Country: El país sobre el que se registran los datos
- Year: Rango de años que fueron registrados (2000 a 2015)
- Status: Estatus en el que se encuentra cada país en el año estudiado (desarrollado o en desarrollo)
- Life expectancy: Esperanza de vida
- Adult mortality: Tasa de mortalidad, la cual incluye ambos sexos en un rango de 15 a 60 años por cada 1,000 habitantes
- Infant deaths: Tasa de mortalidad en infantes (menores de un año).
- Under-five deaths: Tasa de mortalidad de niños que se encuentran por debajo de 5 años de edad.
- Polio: Número de personas inmunizadas a lo largo de su primer año de vida contra la poliomielitis (Pol3).
- Diphtheria: Número de personas inmunizadas a lo largo de su primer año de vida contra la difteria (DTP3).
- Hepatitis B: Número de personas inmunizadas a lo largo de su primer año de vida contra la Hepatitis B (Hep3).
- Alcohol: Litros de alcohol consumidos per cápita por personas de 15 años o más
- Percentage expenditure: Porcentaje que se gasta del Producto Interno Bruto per cápita en salud
- Measles: Número de casos registrados de sarampión por cada 1,000 habitantes.
- BMI: Promedio del índice de masa corporal de la población
- Total expenditure: Porcentaje que invierte el gobierno en salud del total de sus gastos
- HIV/AIDS: Muertes por cada 1,000 nacimientos entre los 0 y 4 años a causa del virus de inmunodeficiencia humana
- GDP: Producto Interno Bruto per cápita en dólares
- Thinness 1-19 years: Porcentaje de delgadez (índice de masa corporal menor a la media por el equivalente a dos desviaciones estándar) entre 10-19 años
- Thinness 5-9 years: Porcentaje de delgadez (índice de masa corporal menor a la media por el equivalente a dos desviaciones estándar) entre 5-9 años

- Income composition of resources: Índice de desarrollo humano en términos de la composición de ingresos. Corresponde al porcentaje total de ingresos que le corresponde a cada persona en un país.
- Schooling: Años de escolaridad en la población

Las variables están principalmente relacionadas con factores que influyen a la salud, como lo es la inmunización en edad temprana contra enfermedades o el IMC; así como factores económicos, tanto a nivel gubernamental como a nivel social.

Esta base de datos fue escogida en razón de lo relevante que resulta tener datos como ayuda para medir el desarrollo de un país, en especial en el sector salud. Con esta base se consideró que se podrían encontrar los mayores beneficios para garantizar una vida sana en países menos desarrollados, al realizar un análisis de los datos reflejados, teniendo en cuenta las variables proporcionadas y comparando con países más desarrollados.

Objetivo

Objetivo general: Encontrar las estrategias más efectivas con el fin de aumentar la esperanza de vida, contribuyendo al cumplimiento del Objetivo de Desarrollo Sostenible 3 (garantizar una vida sana y promover el bienestar de todos a todas las edades) mediante la identificación de relaciones entre factores socioeconómicos y de salud con la longevidad en diferentes países.

Preguntas guía

- ¿Qué enfermedades impactan más la longevidad?
- ¿Qué rol juega el cuidado de la salud en edades tempranas en la esperanza de vida?
- ¿Qué factores afectan más a la salud de los niños?
- ¿Qué factores facilitan alcanzar una mayor esperanza de vida?
- ¿Qué papel tienen los factores socioeconómicos (escolaridad, población, etc.) en la esperanza de vida?
- ¿Qué tendencias sigue el valor de la esperanza de vida alrededor del mundo?
- ¿Cómo se relaciona la mortalidad en adultos con la mortalidad en niños? ¿Qué factores se ven involucrados en esta interacción?

Selección

La base seleccionada, a simple vista, se mostró llamativa para el equipo, pues se encontró compuesta de datos ordenados y con nombres que permiten darse una buena idea de aquello que representan los números. No obstante, al comenzar el análisis, se volvió imposible no notar algunos aspectos en los que la base podría no cumplir con las necesidades del equipo según los objetivos propuestos, desde aspectos tan básicos como que los nombres de las variables contengan espacios o que algunas variables no contuviera ningún dato para algunos años específicos.

Para comenzar, se decidió pasar la base por una serie de filtros para obtener los países más representativos a analizar según el objetivo. Se comenzó seleccionando todos los países pertenecientes al año 2015 para poder hacer un análisis de los últimos datos registrados en la base. Lo que se esperaría conseguir de este filtrado es una base con menos países que permitiera analizar el desarrollo de algunos de los más representativos de estos, así como visualizar los cambios más fácilmente. Se dividió la base en cuatro sub-categorías con respecto al valor de la variable *GDP*, referente al Producto Interno Bruto per cápita de cada país. Estas se definieron como: high (valores mayores a 12235), low (valores menores a 1005), lower middle (valores entre 1005 y 3955), y upper middle (valores entre 3955 y 12235). Posteriormente, tomando en cuenta la columnas de *Adult mortality* y *Life expectancy*, se agregaron más filtros con el objetivo de definir los países que tienen una vida más longeva. Esto se decidió de este modo porque, por un lado, solo analizar la variable de

la esperanza de vida no representaría en su totalidad lo que se buscaba; además, un solo filtro resultaba insuficiente y hubiera llevado a encontrar países con iguales valores de *GDP per capita.*

Con esto último en mente, habiendo dividido ya en cuatro rangos de *GDP,* se buscaron aquellos países que tuvieran un valor máximo, valor mínimo o el valor más cercano al promedio de su rango. Con esto se obtuvieron en total 12 países representativos de la base de datos que pueden dar una idea más general de su comportamiento. *(Nota: resultan 12 países al dividir en tres categorías cada uno de los cuatro rangos de PIB -valor máximo, mínimo y promedio)*.

Al empezar a trabajar con los datos se encontraron muchas inconsistencias con los valores, en especial con aquellos que se usaron para el filtrado principal, con la variable de *GDP*. Debido a grandes incoherencias de los resultados, se decidió complementar la base con datos obtenidos directamente del Banco Mundial para el PIB de los países en la primera base. De esta manera, se tuvo que iniciar un proceso de filtros ligeramente diferentes que permitieran agregar estos nuevos datos al data frame original. Es aquí donde se presentó, tal vez, el reto más grande de filtrado, pues los nombres, por razones geopolíticas, estaban escritos de diferente manera en ambas bases de datos. Si solo se intentaban igualar las bases sin tener esto en cuenta, el análisis se hubiera visto desprovisto de varios datos relevantes. No obstante, el equipo se decidió por la alternativa de simplificar un poco los nombres de los países en ambas bases de datos de manera que no presentaran un problema y el sistema pudiera encontrar similitudes al instante. Con esto, se logró anexar los nuevos datos del Producto Interno Bruto per cápita y realizar el filtrado como se explicó con anterioridad. Esto resultó muy fácil gracias al uso de funciones que permiten automatizar todo el proceso y regresar una base de datos “simplificada” rápidamente.

Estas medidas también dieron solución a otras problemáticas que hubieran hecho el trabajo más tedioso con la base anterior. Entre ellas la falta de datos de países que pudieron haber resultado relevantes al filtrar, que obligaba a quedarse sin forma de saber en realidad qué tanto pudieron haber afectado los resultados. Ejemplos de estos son países sin datos para el PIB son: Estados Unidos, Reino Unido, Venezuela y la República de Corea. De igual manera, después de cierta investigación más a fondo sobre los países, esta posible pérdida de datos importantes se volvió irrelevante.

Exploración

Para comenzar con la exploración, se realizó un filtrado que permitiera encontrar los países que, dentro de las categorías para el Producto Interno Bruto per cápita presentadas por Semak (2019) para el 2015, correspondiera a los valores de esperanza de vida máximo, mínimo y promedio (o el más cercano a él) para cada estrato en dicho año. Estos países (cuya esperanza de vida y PIB per cápita se muestra en la Figura 1) se usarían para conocer la evolución de los diferentes factores que componen la base de datos.

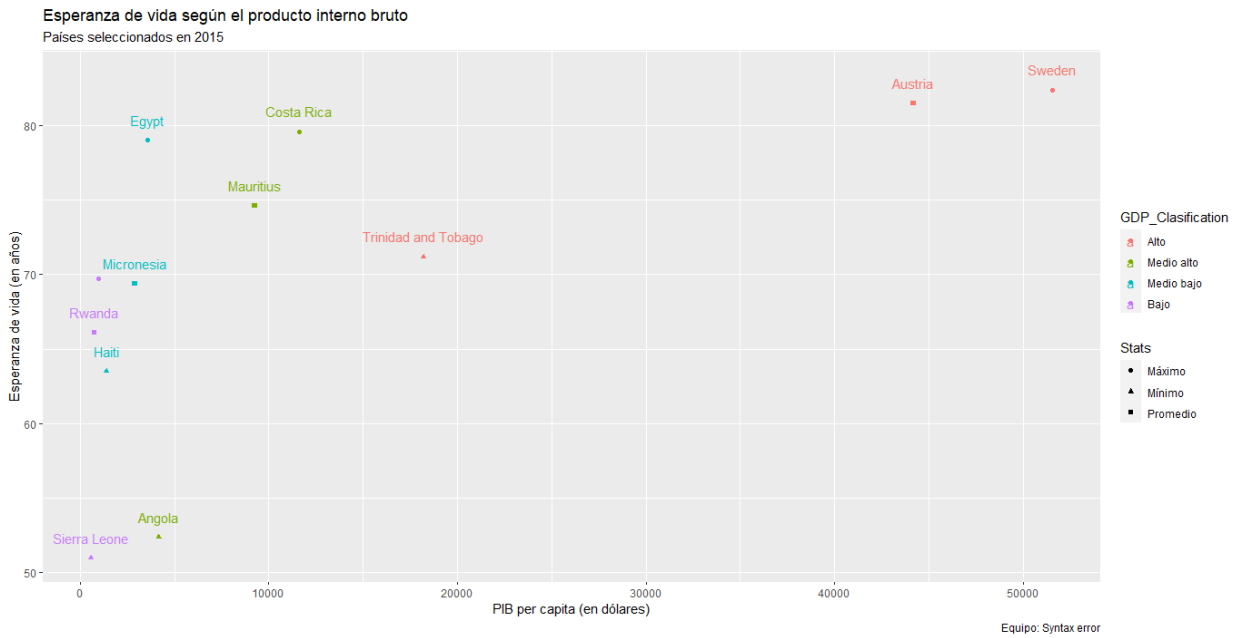

Fig 1. Esperanza de vida filtrando países por rangos de PIB

A primera impresión, podría ser muy evidente que la longevidad es más fácilmente alcanzable en un país rico. Es por esto que se decide encontrar la relación entre el dinero que tienen los países seleccionados y buscar una relación con una mayor esperanza de vida. Es fácil notar que la mayoría de los países se encuentran aglomerados a la izquierda, siendo que los únicos con mayor PIB en realidad no sobrepasan más que por dos años al país inmediatamente detrás de estos en esperanza de vida: Costa Rica. Puede tener relevancia fijarse en qué porcentaje del PIB dedican estos países a cada uno de sus ciudadanos, como se mostrará a continuación:

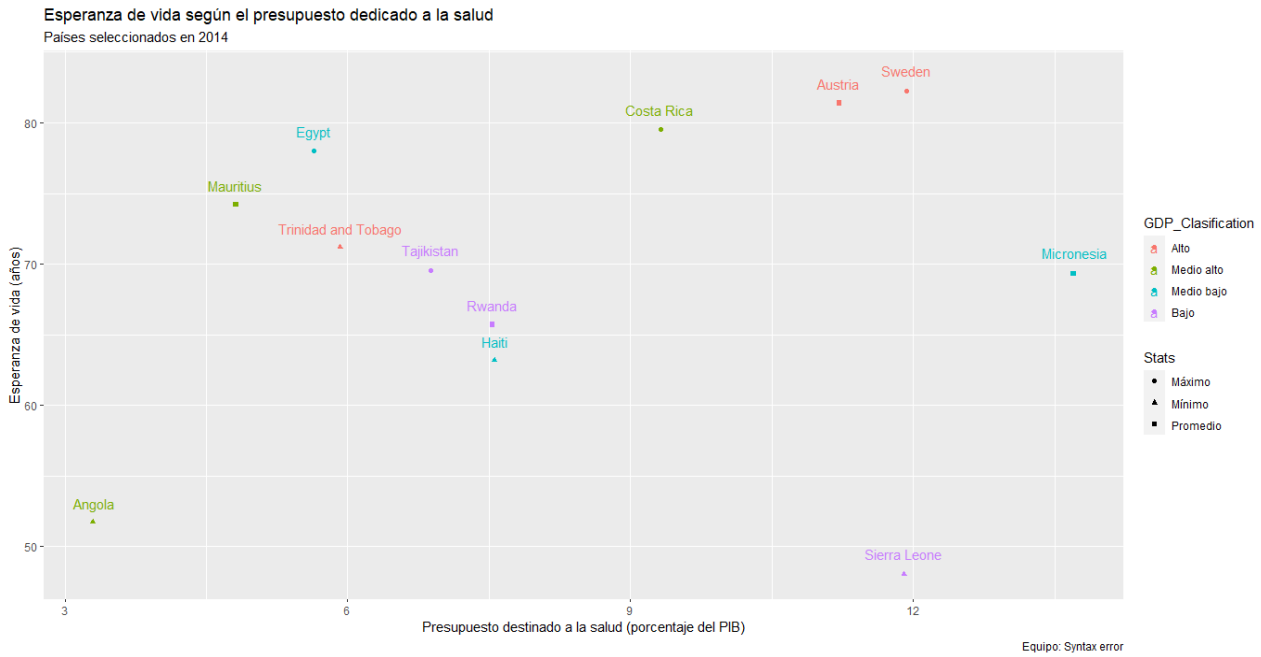

Fig 2. Esperanza de vida según el presupuesto dedicado a la salud

Aquí es evidente cómo los dos países de la gráfica anterior siguen liderando la estadística; sin embargo, aquí los países ya no parecen aglomerarse en una esquina, sino que se desplazan un poco hacia la derecha, en especial países como Micronesia y Sierra Leone que parecen dedicar un buen

porcentaje de su Producto Interno Bruto a la salud. Aunque es claro que Angola y Sierra Leona son los países que menor esperanza de vida alcanzan, existe una enorme diferencia entre el porcentaje que cada país dedica a este servicio. Tal vez algo que valga la pena mirar más a fondo.

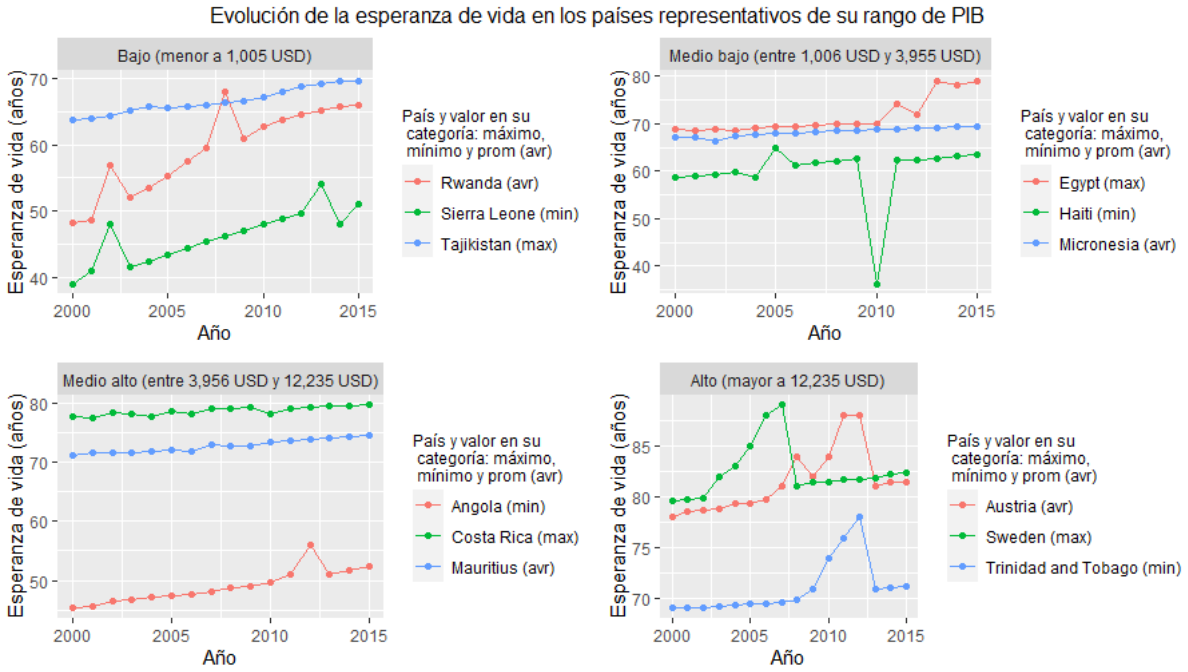

Fig 3. Evolución de la esperanza de vida en los países representativos para cada rango de PIB

Para tener una mejor perspectiva de los cambios en la esperanza de vida de los países elegidos dentro de cada rango, se graficó su valor a lo largo de los años presentes en la base de datos. En general, se observa que la esperanza de vida sigue una tendencia positiva; esto es, que aumenta con el paso de los años. Los países con un Producto Interno Bruto en el rango medio alto y medio bajo presentaron un crecimiento más constante, sin picos ni caídas destacables a excepción del pico en 2012 en Angola y la caída en Haití en 2010 (este último fenómeno puede atribuirse al terremoto que aconteció en ese año y devastó al país). En cambio, los países clasificados por su PIB como alto y bajo presentaron comportamientos más diferenciados, como se observa en las gráficas. En los países con un PIB alto, dichos picos en la esperanza de vida se vieron seguidos por un decremento hasta la tendencia normal, lo que puede interpretarse como errores en el registro de los datos, puesto que dichos valores no se relacionan con el desempeño de otras variables, como se observa en figuras posteriores.

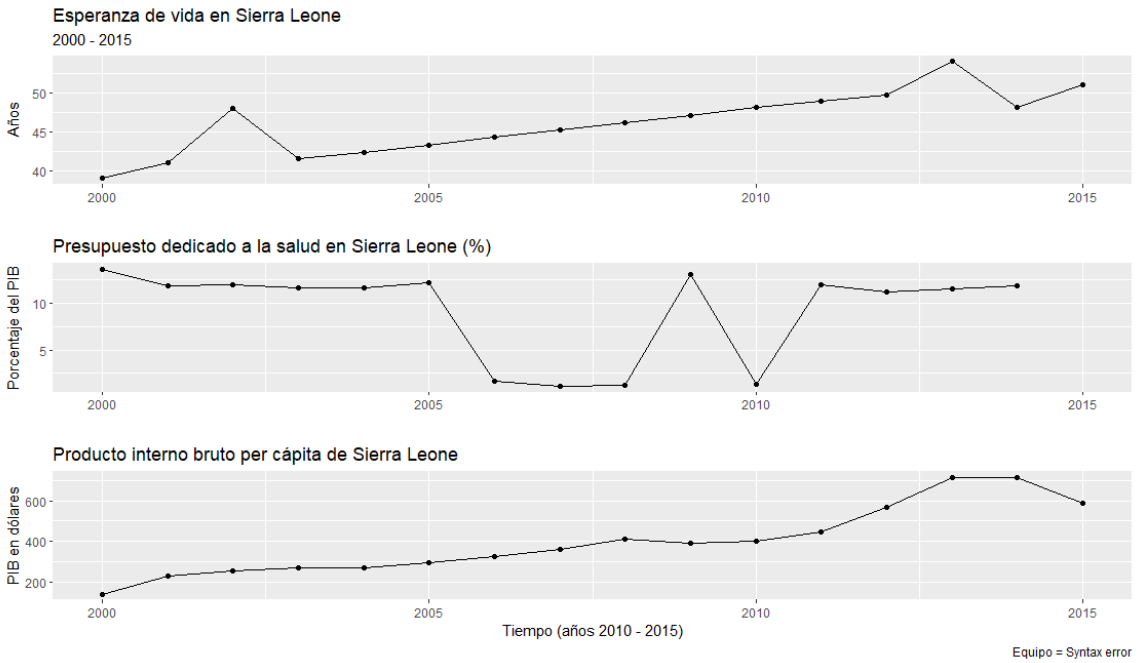

Fig 4. Comparación de variables para Sierra Leone

La esperanza de vida de este país comenzó baja y subió repentinamente de 2001 a 2002, datos que coinciden con el final de una guerra civil este mismo año. Además, este parámetro también ascendió bastante en 2012, ascenso que se encuentra también en su PIB desde un año antes. Asimismo, se señala cómo, a lo largo de los años (desde el 2003 y hasta el 2013) se muestra un crecimiento constante en la esperanza de vida, sin importar el decaimiento del porcentaje que se le dedicaba a la salud durante un gran período de ese intervalo. Quizá otro de los repentinos cambios que presenta la gráfica es el declive del dinero destinado a la salud en 2005. Dado que este dato es derivado del PIB, su decremento podría deberse al incremento del otro.

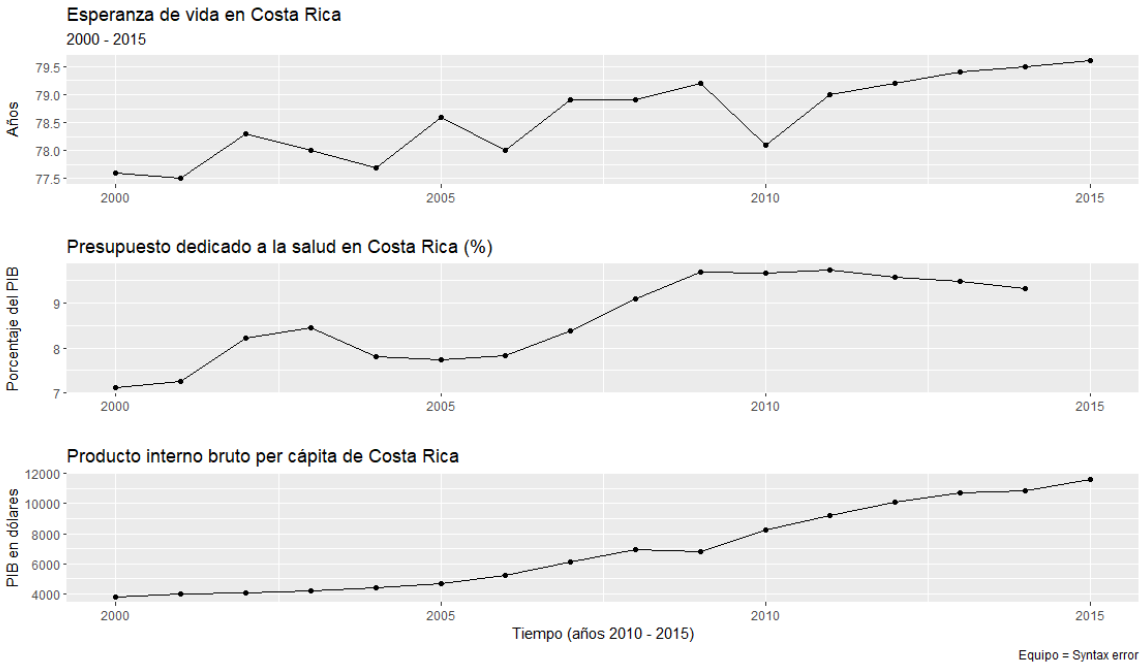

Fig 5. Comparación de variables para Costa Rica.

Este país se mostró de interés al observar la *Fig. 2.* Es curioso cómo, en este caso, para llegar a la alta esperanza de vida que se había observado antes, hubo un camino bastante más turbulento que con el país previo. En casi todos los casos, el aumento en la esperanza de vida en Costa Rica se relaciona con un aumento en el presupuesto dedicado al sector salud del país, tal vez solo a excepción del 2001 y el 2008. Aunado a esto, es interesante notar que un pequeño cambio en el porcentaje puede generar un gran cambio en la esperanza de vida, como es el caso para el 2005 o, con un cambio negativo, en el 2006. El PIB se mantiene en constante ascenso durante casi el período de 15 años completo, lo que quizá le da al país una cierta estabilidad.

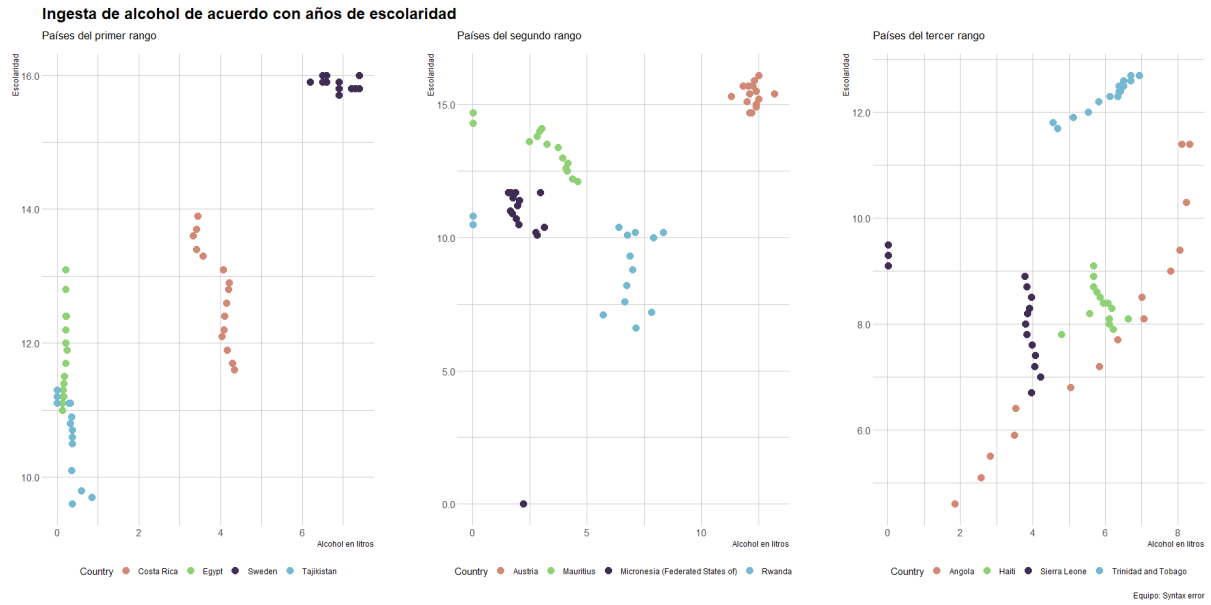

Fig 6. Gráfica de relación años de escolaridad y litros de alcohol consumidos

Otro dato interesante a notar es el consumo de alcohol, que se encuentra, en mayor parte, sin datos, en especial para el 2015, el año que se busca analizar. Por esto, se podría buscar relacionar la variable con alguna otra para los años en los que sí se presentan datos. Los países que aparecen en las gráficas 2 y 3 en color rosa corresponden al país con mayor PIB, los que aparecen en verde al medio-alto, los que aparecen en negro al medio-bajo, y los que aparecen en azul al menor.

En la primera gráfica, se hizo una relación de dos variables (alcohol y escolaridad) para poder comprobar si existe algún patrón entre los países. Como se puede observar en las tres gráficas, el país que tiene mayor PIB per cápita es de igual forma el que tiene mayor escolaridad y, a su vez, el mayor consumo de alcohol en litros. Esto indica que no necesariamente el tener menor escolaridad está ligado a un consumo de estas bebidas adictivas, conocidas por la población porque tienden a reducir la longevidad de vida de una persona.

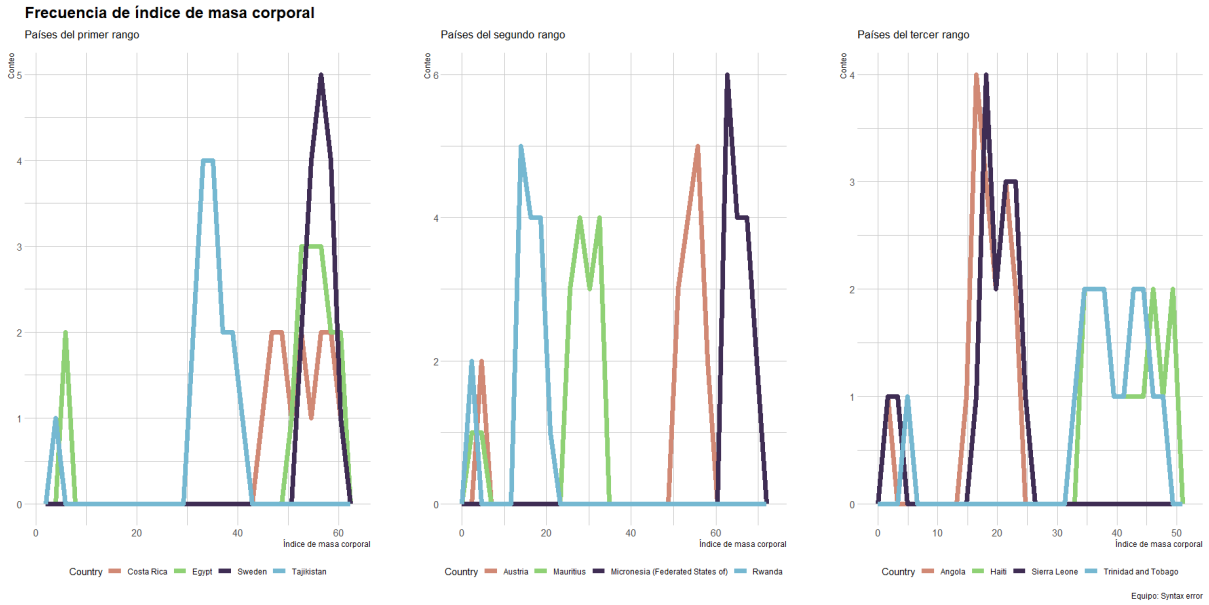

Fig 7. Gráfica de frecuencia de índice de masa corporal en la población

En el caso de esta gráfica, se buscó un enfoque más la población. En el primer rango de países se puede observar que el índice de masa corporal está principalmente entre 40 y 60, lo cual indica que las personas en esos países tienen una buena relación entre su masa y la estatura que tienen, mientras que en los países del segundo y tercer rango se encuentran valores por debajo de 20, lo que indica un índice de desnutrición ayor en una proporción significativa con relación a su masa y estatura.

Para este caso, el PIB juega un papel importante debido al tener recursos económicos per cápita superiores en los países del primer rango, tienden a tener mayor posibilidad de acceder a servicios de salud o tener una dieta balanceada.

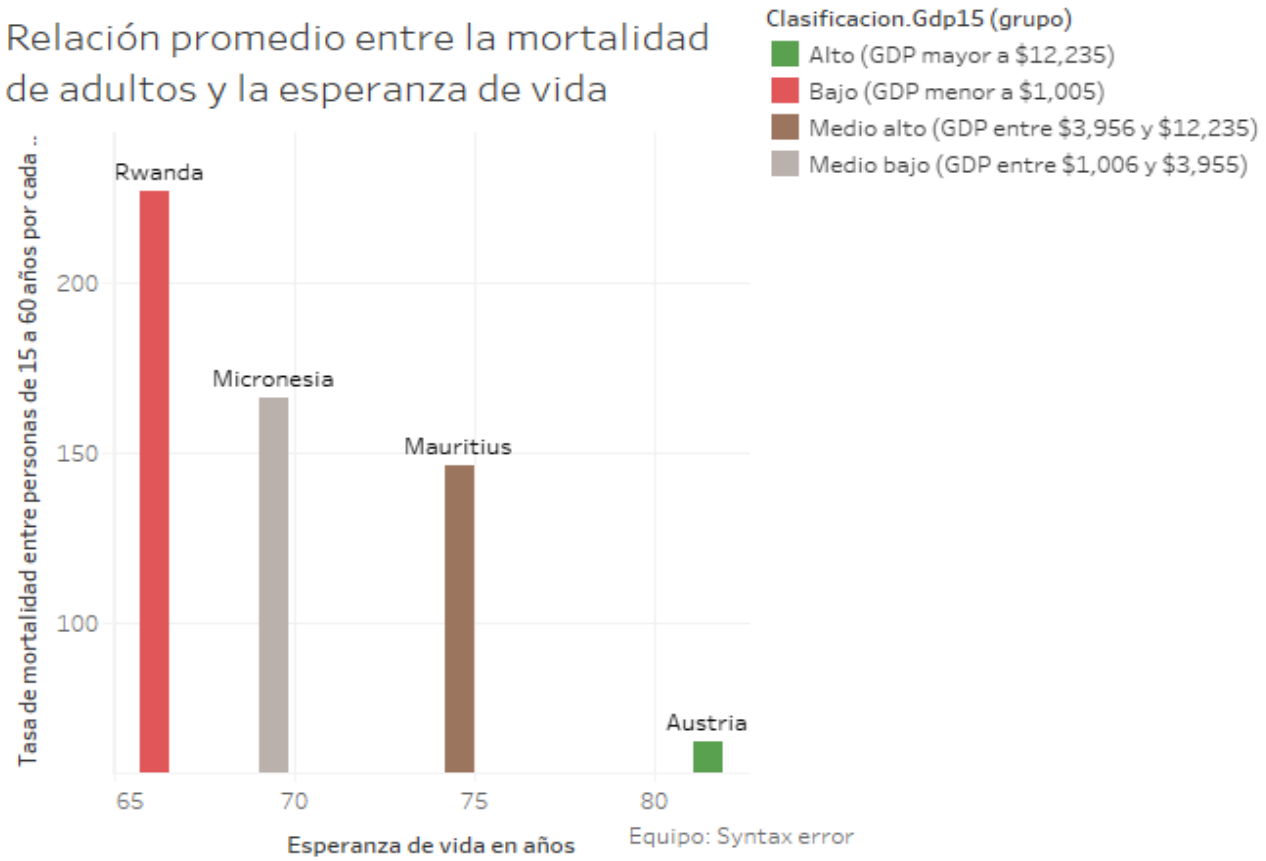

Fig 8. Gráfica de relación promedio entre la mortalidad de adultos y la esperanza de vida

Buscando un impacto más directo con la esperanza de vida y, como su nombre lo indica, esta gráfica muestra la relación existente entre los países con un valor cercano al promedio en la esperanza de vida dentro de su respectiva categoría de GDP y la mortalidad de adultos. Como se puede apreciar, entre menor esperanza de vida en años tenga un país, mayor es el valor de la tasa de mortalidad en adultos. Además, también se muestra que los países que tienen menor esperanza de vida son aquellas que se encuentran en las categorías de 'Bajo' y 'Medio bajo' en cuanto a GDP. Se demuestra que estas dos variables se relacionan de manera inversa, ya que mientras una aumenta, la otra disminuye.

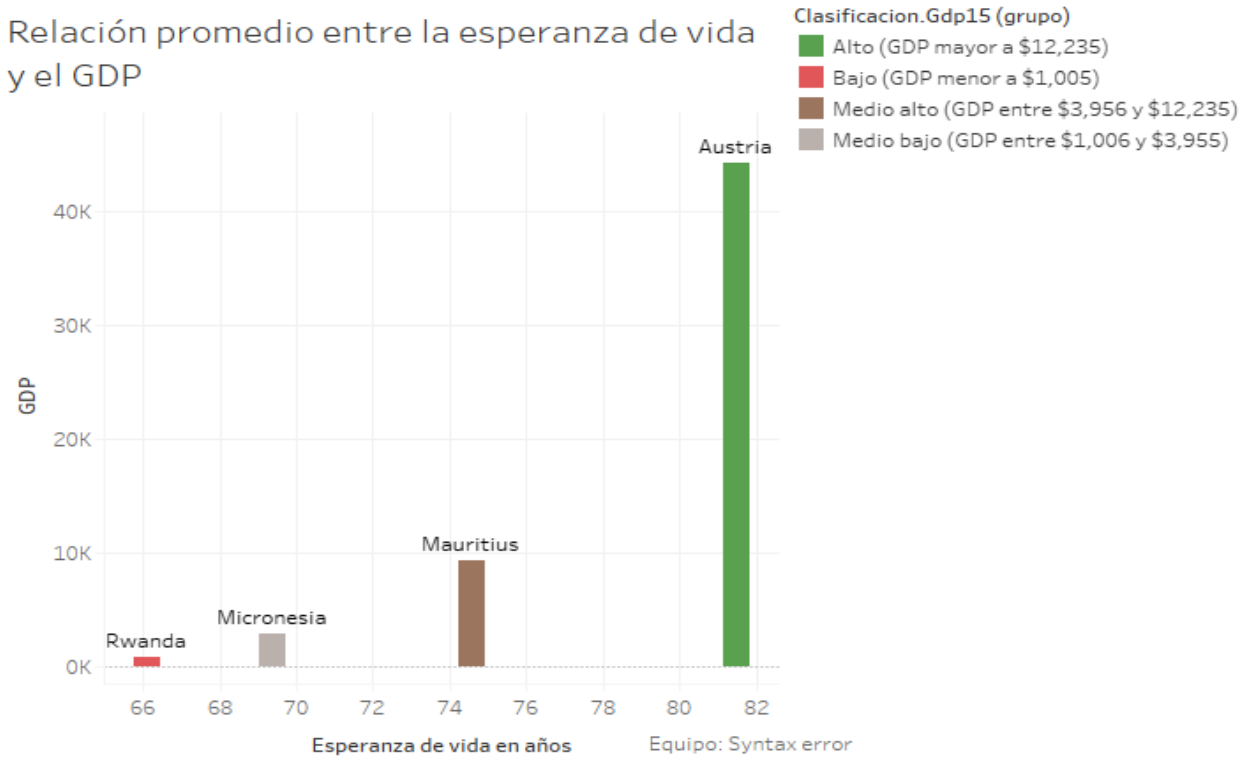

Fig 9. Relación promedio entre la esperanza de vida y el GDP

Por último, fijando la mirada en el ámbito económico, esta gráfica muestra la relación existente entre los países con un valor cercano al promedio en la esperanza de vida dentro de su respectiva categoría de GDP y precisamente el GDP. Como se puede apreciar, entre mayor esperanza de vida en años tenga un país, mayor es el valor del GDP. Además, también se muestra que los países que tienen mayor esperanza de vida son aquellas que se encuentran en las categorías de 'Alto' y 'Medio alto' en cuanto a GDP. Por lo que estas dos variables se relacionan de manera directa, ya que mientras una aumenta, la otra también. De aquí, es posible pasar a considerar las correlaciones entre las variables.

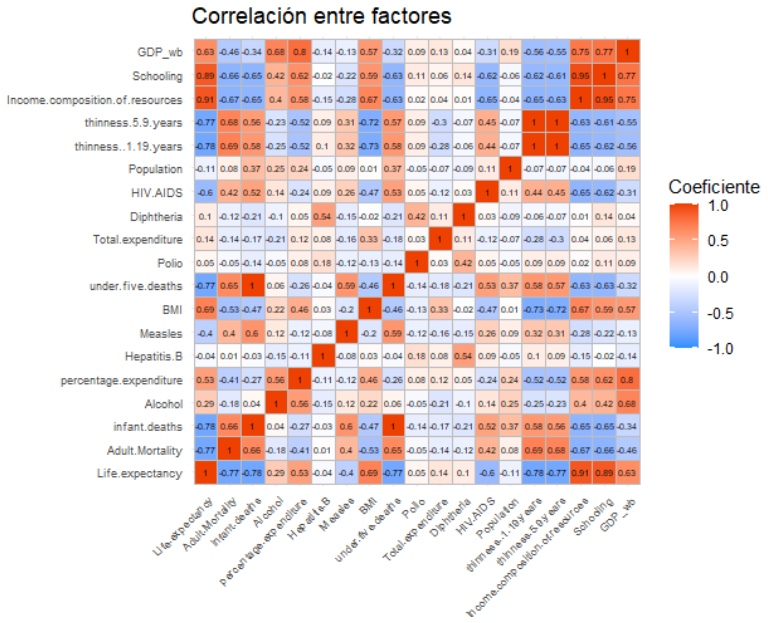

Fig 10. Representación gráfica de las correlaciones entre las variables numéricas de la base de datos

Mediante el desarrollo de las correlaciones mostradas en la figura anterior, fue posible comenzar a identificar algunas relaciones entre variables que se impactan entre sí: primeramente, según el signo del valor en cada recuadro (positivo o negativo) es posible saber si la relación entre ambas es directa o inversa, respectivamente, y el valor absoluto de dicha magnitud (representado gráficamente también por la intensidad del color, ya sea rojo para valores positivos o azul para negativos) cuán relacionadas están dos variables entre sí. Esta información es de gran relevancia en el caso de la esperanza de vida, ya que permite identificar con certeza las variables que más cercanamente se relacionan a dicho valor, lo que contribuyó significativamente a enfocar los siguientes gráficos y análisis estadísticos al objetivo del reporte.

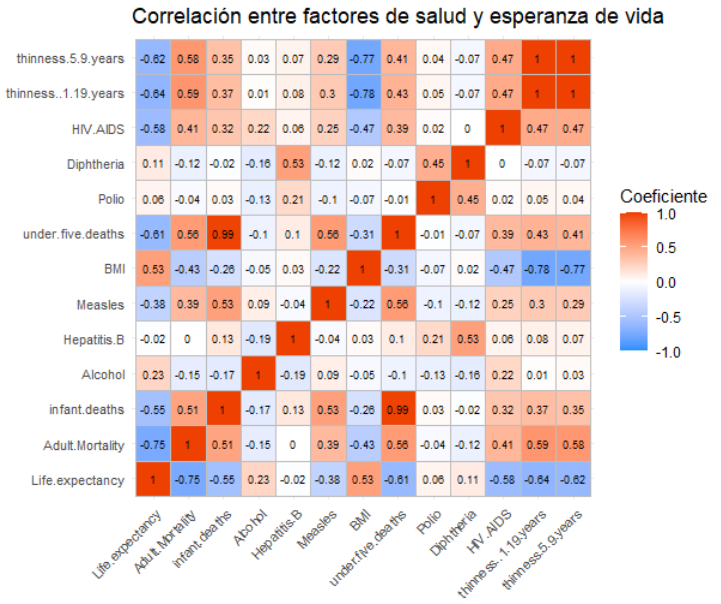

Fig 11. Correlaciones entre factores de salud y esperanza de vida

Con el fin de identificar mejor las interacciones entre los diferentes factores evaluados, se optó por clasificarlos en dos categorías: de salud y socioeconómicos. Dentro de la primera, se encuentran aspectos como prevalencia de enfermedades, vacunación, delgadez, índice de masa corporal, y mortalidad. Esto permitió identificar con mayor facilidad la forma en que los valores se relacionan entre sí, destacando relaciones interesantes como el sarampión (Measles) y la mortalidad en infantes y niños menores de 5 años, la vacunación contra la hepatitis B y la difteria, y la prevalencia del VIH SIDA como uno de los mayores factores que contribuyen a aumentar la mortalidad de niños y adultos por igual. Asimismo, cabe destacar que, dado que la variable referente a muertes en niños menores a 5 años abarca las muertes de infantes (menores a un año de vida), si se analiza su impacto en la esperanza de vida general, se puede inferir que los fallecimientos entre el nacimiento y los 12 meses de edad representan alrededor del 90% de la mortalidad infantil.

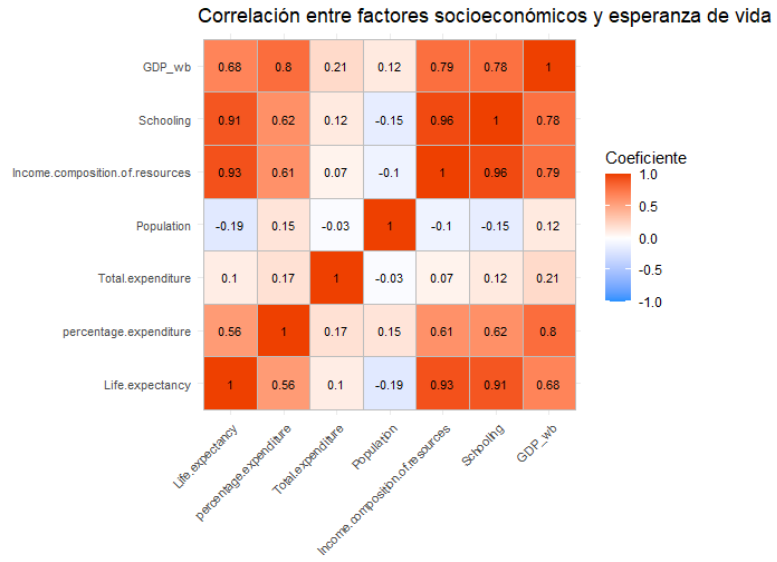

Fig. 12 Correlciones entre factores socioeconómicos y esperanza de vida

Por otro lado, en los factores socioeconómicos se consideraron la distribución promedio de la riqueza (PIB per cápita e índice de desarrollo humano), escolaridad, población y gasto en salud pública. Resultó importante la relación inversa entre el tamaño de la población y su esperanza de vida, así como el poco impacto del gasto del presupuesto gubernamental en salud, mientras que la escolaridad y la distribución equitativa de recursos impactó positivamente de manera importante la esperanza de vida. Cabe destacar que dichas variables están fuertemente relacionadas entre sí, por lo que impulsar una o ambas las beneficiaría enormemente, con lo que se estaría contribuyendo a mejorar la esperanza de vida.

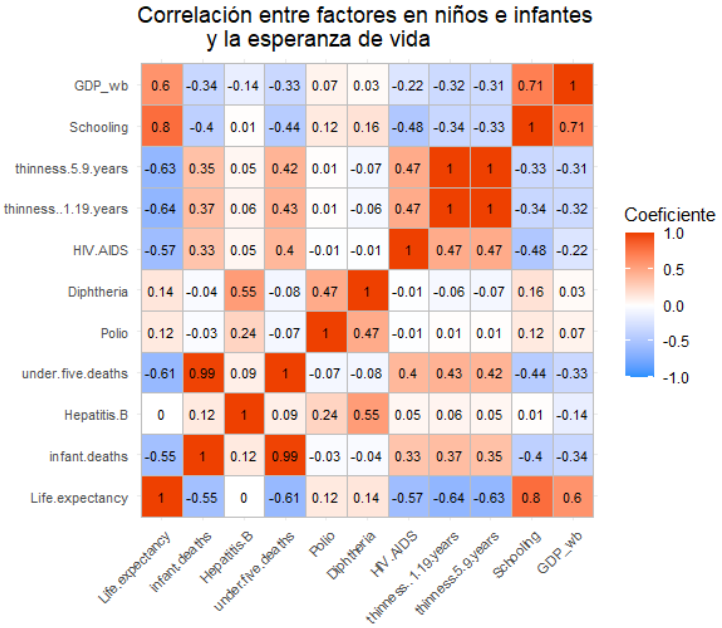

Fig. 13 Correlaciones entre factores relacionados a niños e infantes con esperanza de vida

Una de las observaciones más importantes obtenidas de las correlaciones presentadas en la figura 6 es la relevancia de las muertes en infantes dentro de la esperanza de vida, siendo este factor junto con la delgadez entre 1 y 19 años los más influyentes, según las relaciones presentadas. Por

ello, se creó una cuarta correlación, identificando las variables relativas a la salud de este grupo de edad (vacunación contra hepatitis B, polio y difteria), e incluyendo los datos para niños menores a 5 años. Gracias a esto, se logró identificar a la delgadez (en ambos rangos de edad presentados) y el VIH SIDA como los mayores factores responsables de incrementar la mortalidad a edades tempranas, mientras que la escolaridad y un nivel socioeconómico más elevado resultaron los principales factores para evitar esto. Si bien en esta correlación la vacunación contra las enfermedades aparenta contribuir poco a disminuir la mortalidad en infantes, en la primera se logra apreciar que podrían jugar un papel importante en reducir este indicador.

Etapa 2: Análisis Estadístico

Análisis estadístico descriptivo

Con ayuda de un software de programación estadística, en este caso con el lenguaje de programación R, se obtuvieron los siguientes datos para cada una de las variables que se consideraron:

- Histograma
- Media
- Varianza
- Desviación estándar
- Simetría
- Curtosis

De esta manera, se encuentran los valores requeridos dentro de las visualizaciones de la distribución de cada variable, acompañada de una descripción de esta como se muestra a continuación:

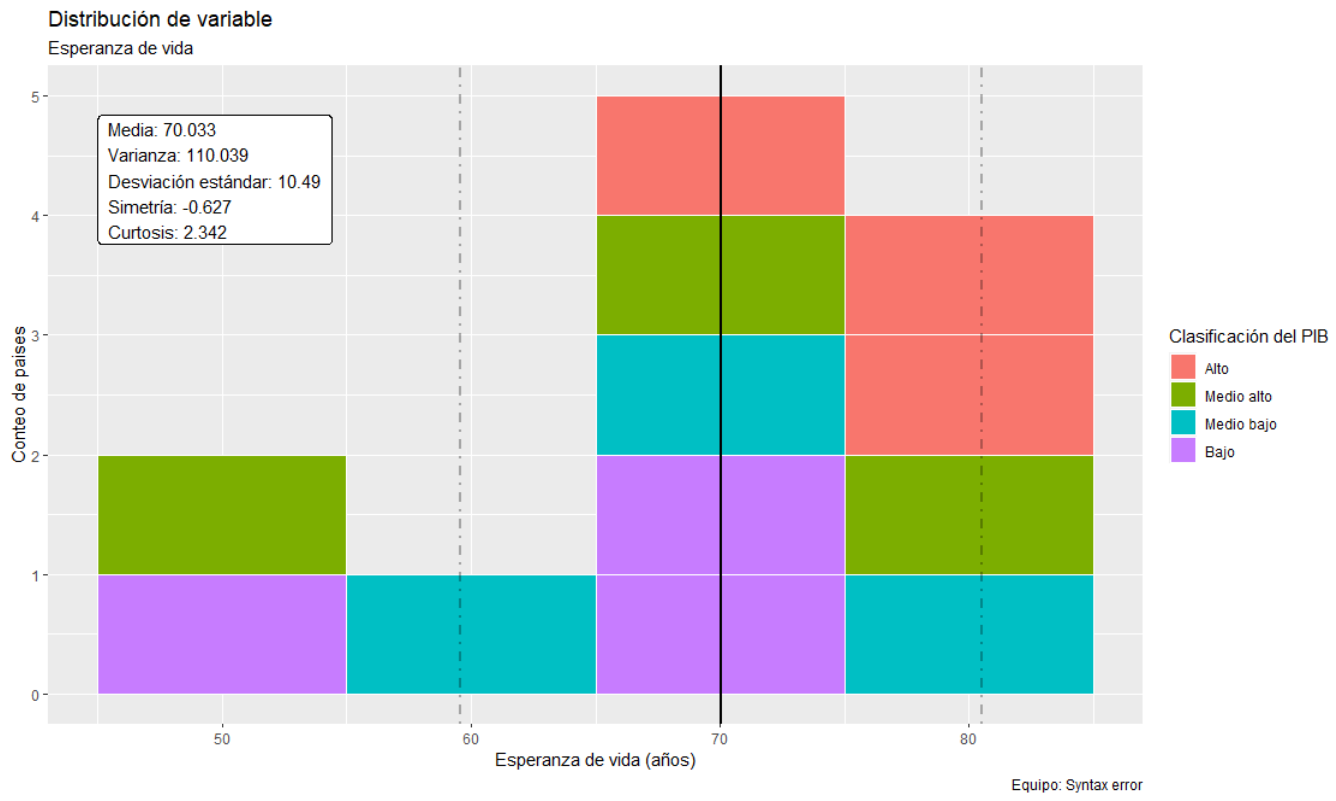

Fig 14. Análisis estadístico de la variable "esperanza de vida".

Hipótesis/descripción: Basándose en la forma del histograma, esta variable podría tener una distribución leptocúrtica, ya que los datos están muy concentrados en la media y se encuentra sesgada hacia la izquierda. Con esto se puede asumir que la variable tiene una mayor cantidad de información en el rango de números después de la media. Además, la gran varianza y las líneas de desviación estándar tan separadas de la media indican una gran dispersión en los datos de esta variable para el 2015.

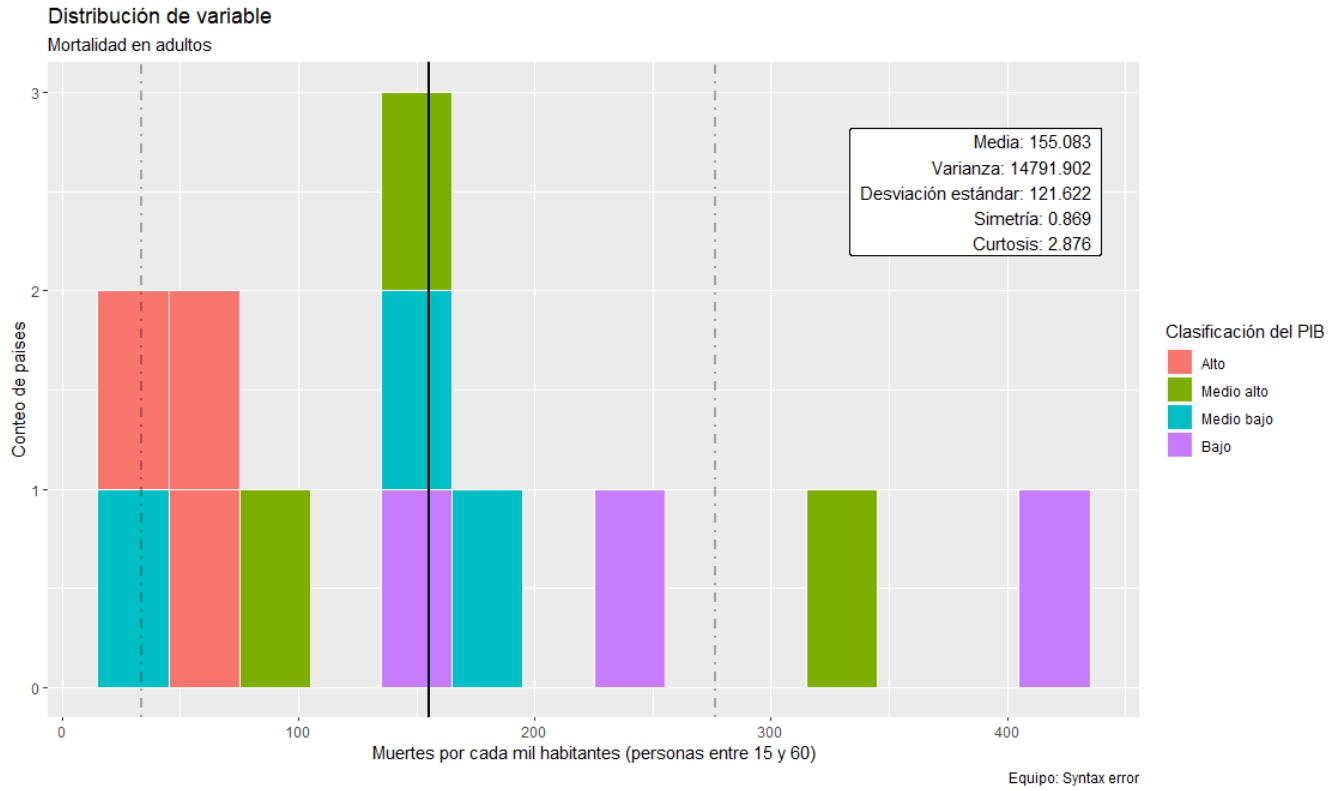

Fig 15. Análisis estadístico de la variable "muertes por cada mil habitantes".

Hipótesis/descripción: Basándose en la forma del histograma, esta variable podría tener una distribución leptocúrtica, ya que los datos están muy concentrados en la media y se encuentra sesgada hacia la derecha por lo que se puede asumir que la variable tiene una mayor cantidad de información en este rango de números (antes de la media). Además, el valor de la varianza y la desviación estándar indican una gran diferencia en los valores de mortalidad en adultos de estos 12 países.

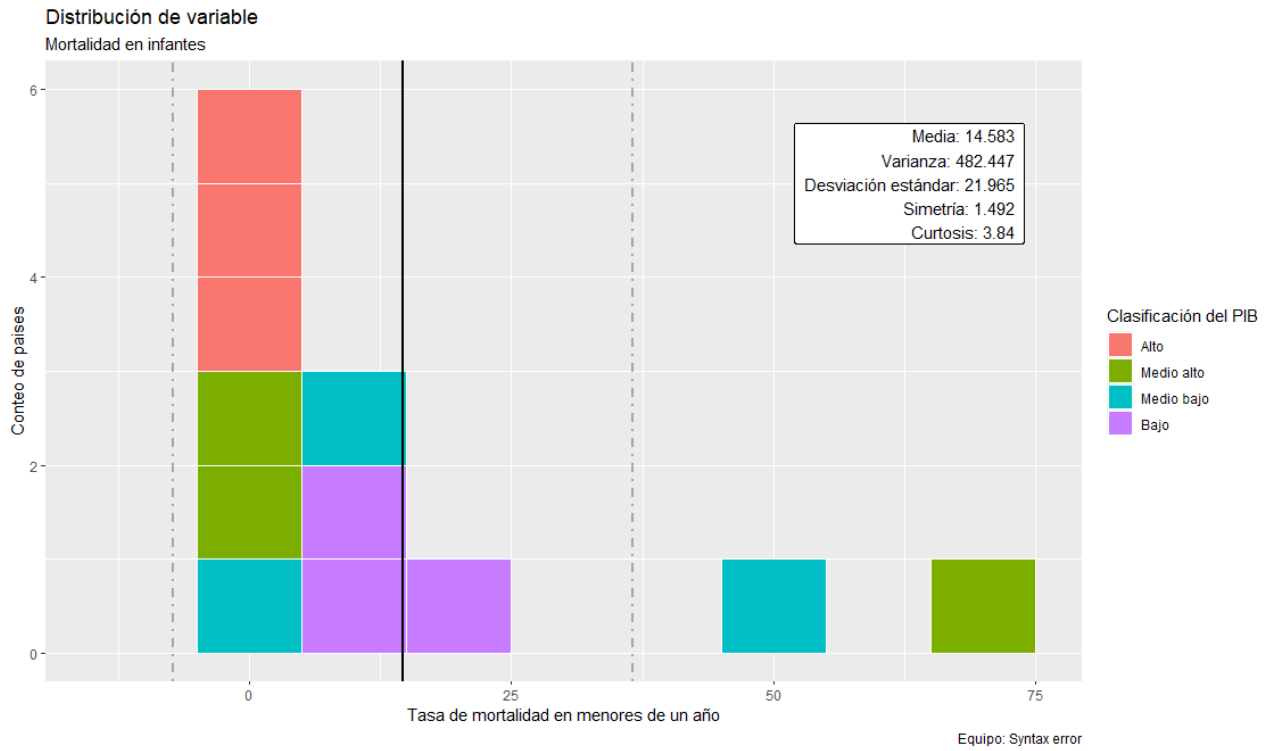

Fig 16. Análisis estadístico de la variable "tasa de mortalidad en menores de un año".

Hipótesis/descripción: Basándose en la forma del histograma, esta variable podría tener una distribución platicúrtica, ya que hay menos concentración de datos en la media y se encuentra sesgada hacia la derecha, por lo que se puede asumir que la variable tiene una mayor cantidad de información en el rango de números anterior a la media. Además, se identifica una gran dispersión de los datos, pues no se concentran cerca de la media.

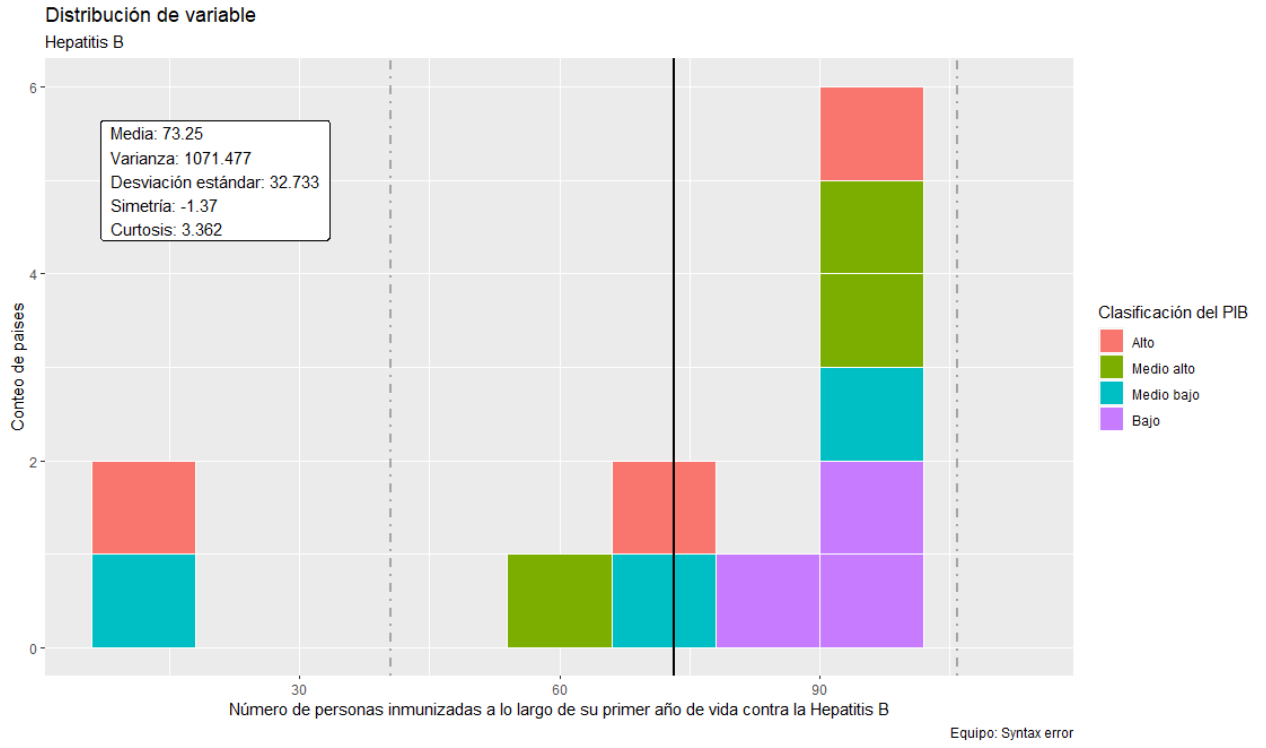

Fig 17. Análisis estadístico de la variable "número de personas inmunizadas a lo largo de su primer año de vida contra la Hepatitis B".

Hipótesis/descripción: Basándose en la forma del histograma, esta variable podría tener una distribución platicúrtica, ya que hay muy poca concentración de datos en la media y se encuentra sesgada hacia la izquierda, por lo que se asume que la variable tiene una mayor cantidad de información en este rango de números. Además, se encuentra una gran dispersión en qué tantas personas están inmunizadas contra este virus en los diferentes países.

a

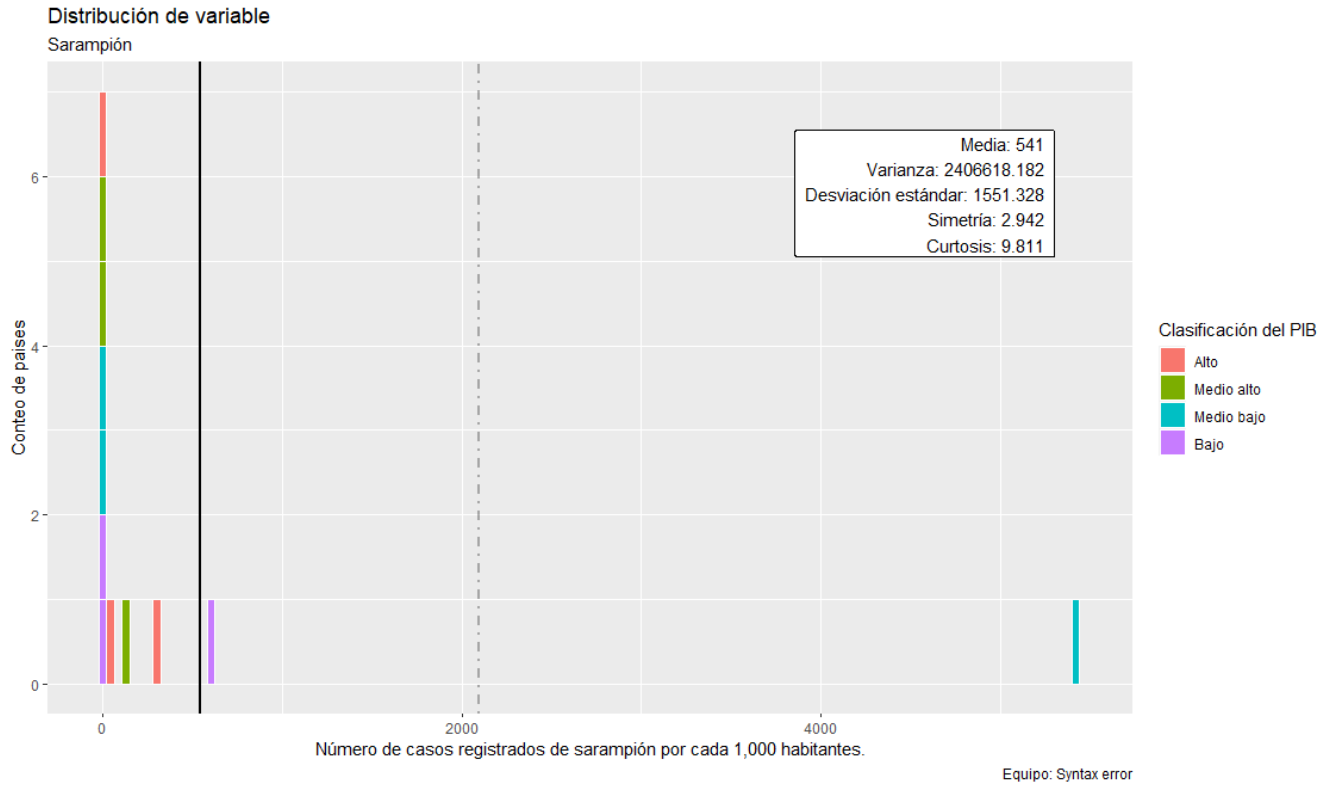

Fig 18. Análisis estadístico de la variable "tasa de mortalidad en menores de un año".

Hipótesis/descripción: Basándose en la forma del histograma, esta variable podría tener una distribución platicúrtica, con un valor muy por encima de lo que se esperaría de cualquier variable. Hay muy poca concentración de datos en la media y se encuentra sesgado hacia la derecha, por lo que se podría asumir que la variable tiene una mayor cantidad de información en el rango de números antes de la media. Se identifica incluso que el ancho de los bins es muy pequeño, lo que llama la atención hacia el hecho que hay un país que se encuentra con muchísimos más casos que todos los demás, llegando hasta los miles.

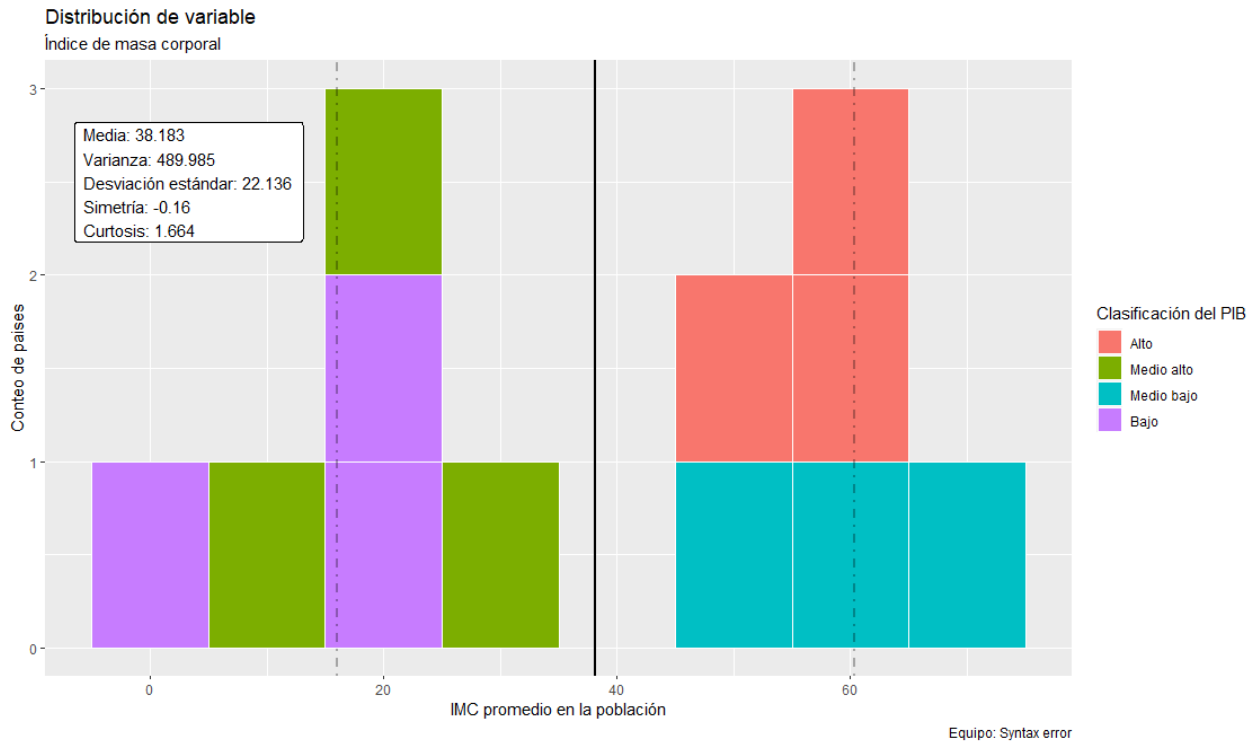

Fig 19. Análisis estadístico de la variable "IMC promedio en la población".

Hipótesis/descripción: como se aprecia en la gráfica, la distribución es de doble pico, lo que quiere decir que es posible que la mayor parte de los datos se encuentre fuera de la desviación estándar o cerca de las colas. Con esta distribución bimodal, se representa cómo se dividen los países de rango alto y medio bajo y los de rango medio alto y bajo en índice de masa corporal promedio.

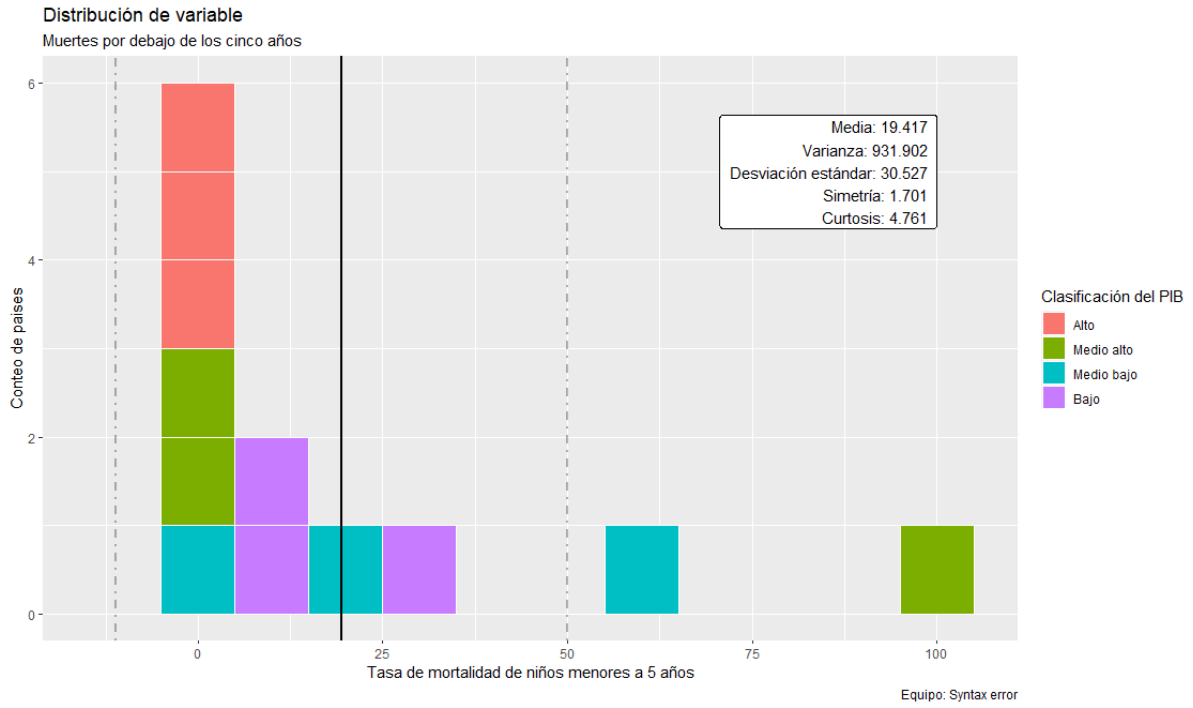

Fig 20. Análisis estadístico de la variable tasa de mortalidad de niños menores a 5 años.

Hipótesis/descripción: Basándose en la forma del histograma, esta variable podría tener una distribución platicúrtica, ya que hay muy poca concentración de datos en la media y se encuentra

sesgada hacia la derecha, por lo que se podría asumir que la variable tiene una mayor cantidad de información en los valores menores a la media.

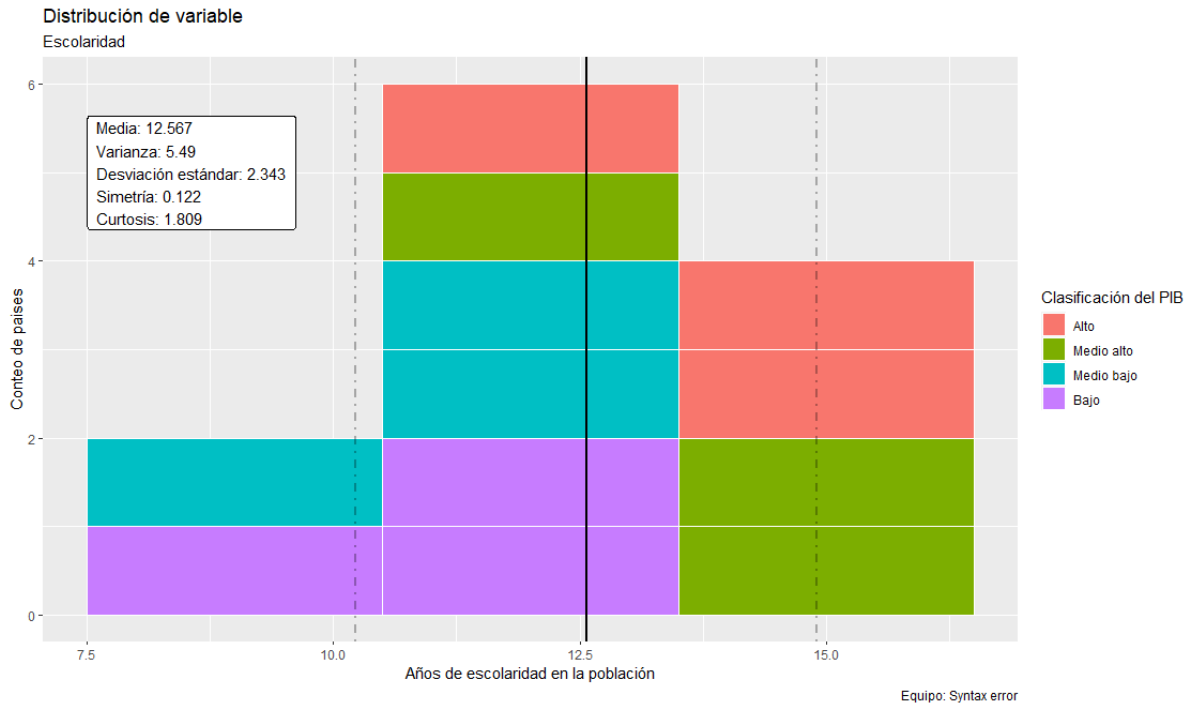

Fig 21. Análisis estadístico de la variable "años de escolaridad en la población".

Hipótesis/descripción: Basándose en la forma del histograma, esta variable podría tener una distribución leptocúrtica, ya que la mayor concentración de datos es en la media y se encuentra sesgada hacia la izquierda por lo que podríamos asumir que la variable tiene una mayor cantidad de información en este rango de números. En realidad es un poco más difícil notarlo en este caso, debido a la simetría tan cercana a 0 de la gráfica.

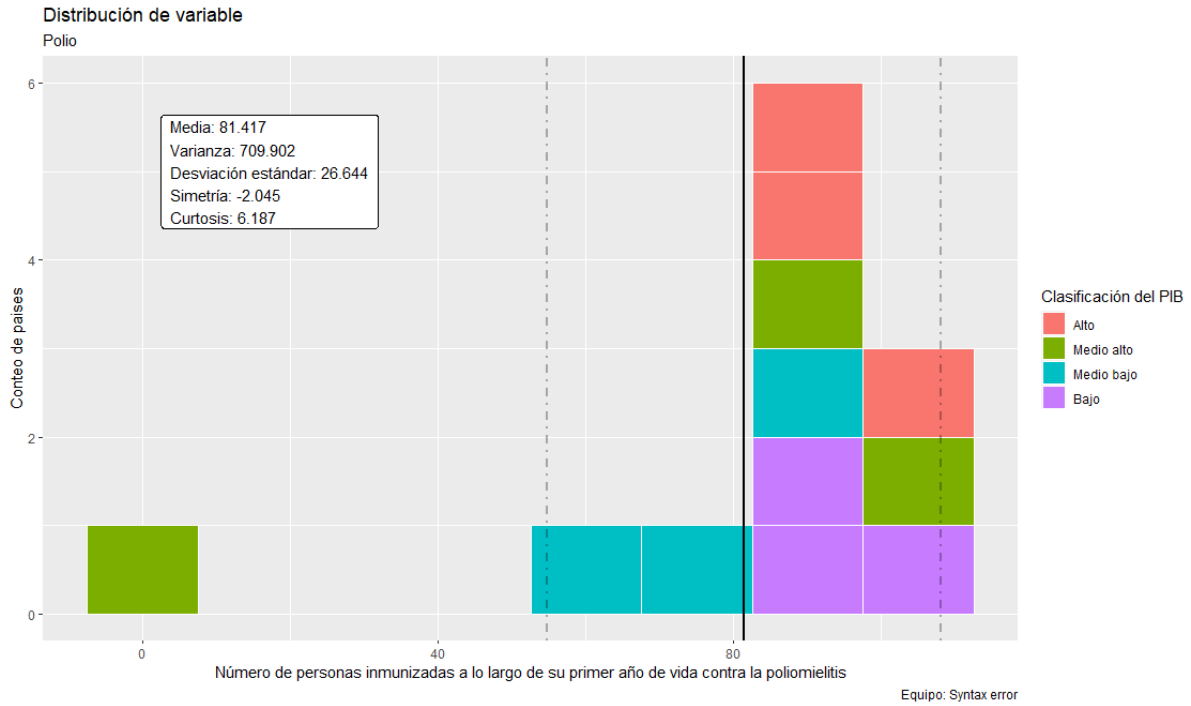

Fig 22. Análisis estadístico de la variable "número de personas inmunizadas a lo largo de su primer año de vida contra la poliomielitis".

Hipótesis/descripción: Basándose en la forma del histograma, esta variable podría tener una distribución platicúrtica, con la mayor concentración de datos cerca de la media, casi todos solo a una desviación estándar. La distribución se encuentra sesgada a la izquierda por lo que podríamos asumir que la variable tiene una mayor cantidad de datos después de la media. Para este caso esto significa que la mayoría de los infantes se inmunizan contra el polio.

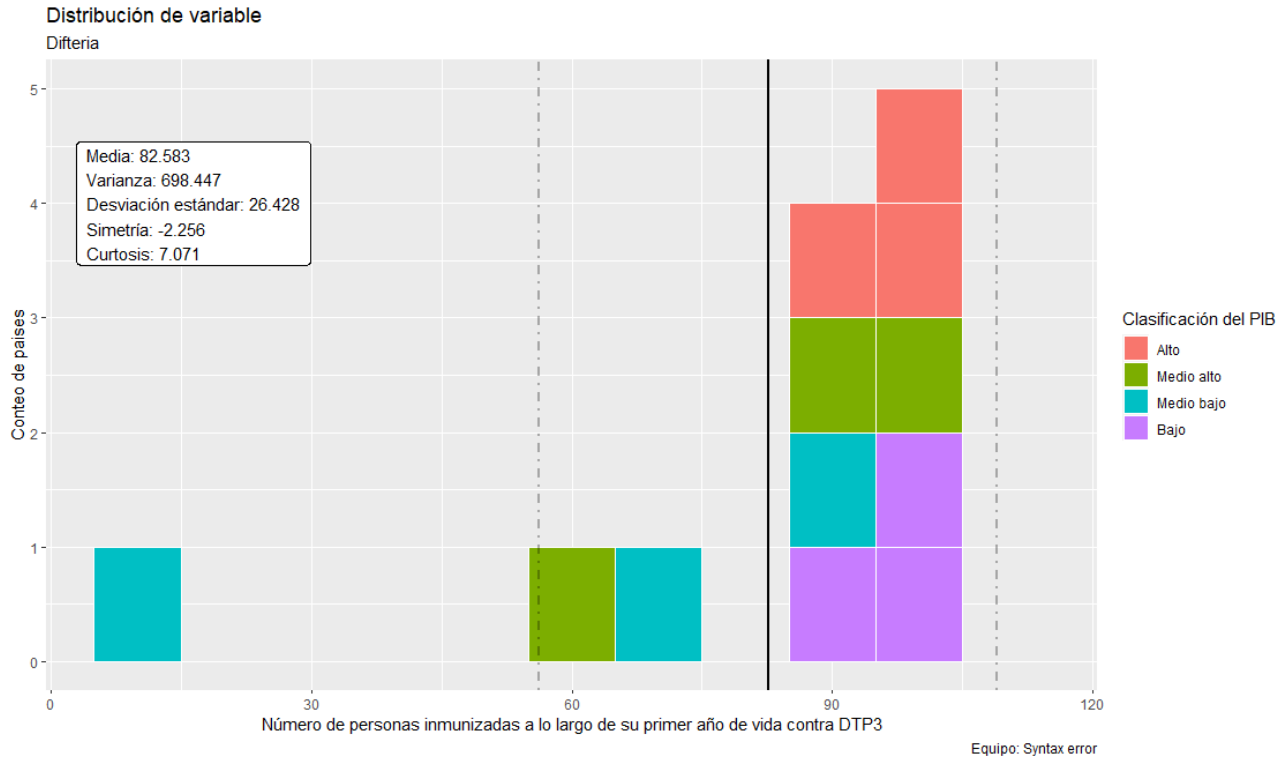

Fig 23. Análisis estadístico de la variable"número de personas inmunizadas a lo largo de su primer año de vida contra DTP3".

Hipótesis/descripción: Basándose en la forma del histograma, esta variable podría tener una distribución platicúrtica, con la mayor concentración de datos cerca de la media; además presentando un sesgo hacia la izquierda, con la moda de personas inmunizadas contra difteria siendo más alta que el promedio para estos países.

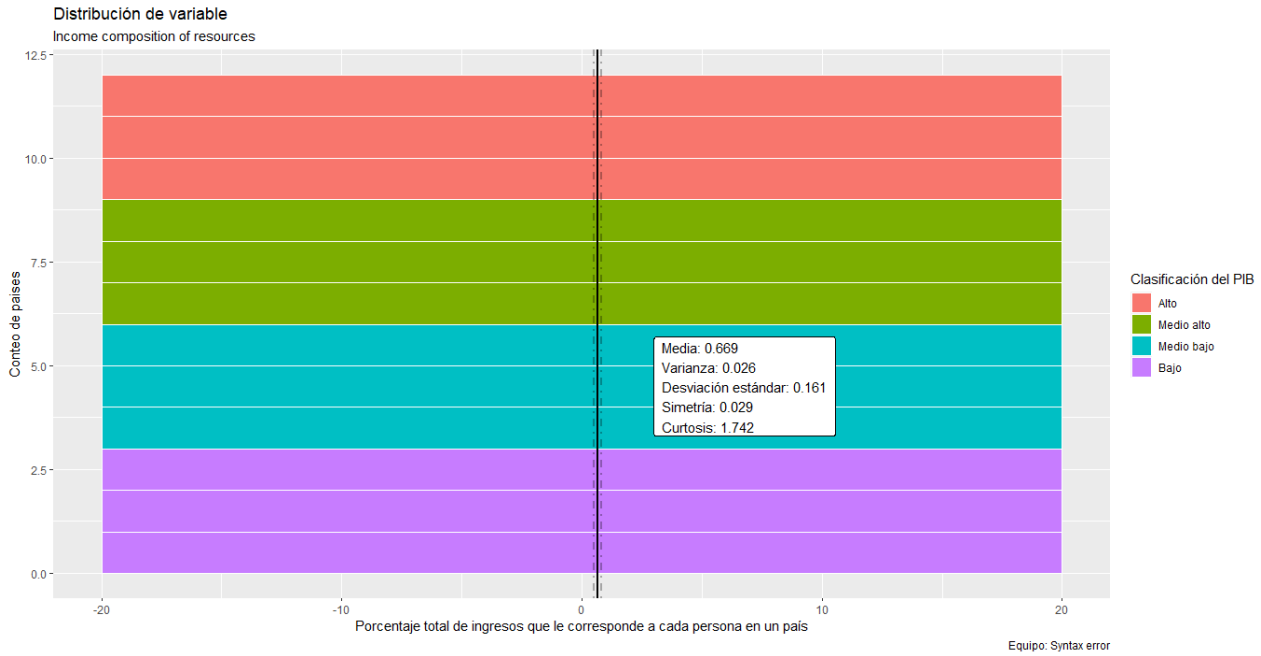

Fig 24. Análisis estadístico de la variable "Porcentaje total de ingresos que le corresponde a cada persona en un país".

Hipótesis/descripción: Basándose en la forma del histograma, esta variable tiene datos muy similares entre sí, ya que la varianza y desviación estándar son mínimos. Además, la simetría resulta casi nula, menor a 0.03, resultando en una distribución prácticamente sin sesgo. Esto quiere decir que, en estos países, el porcentaje de ingresos que le corresponde a cada persona con respecto al dinero del país, es muy similar.

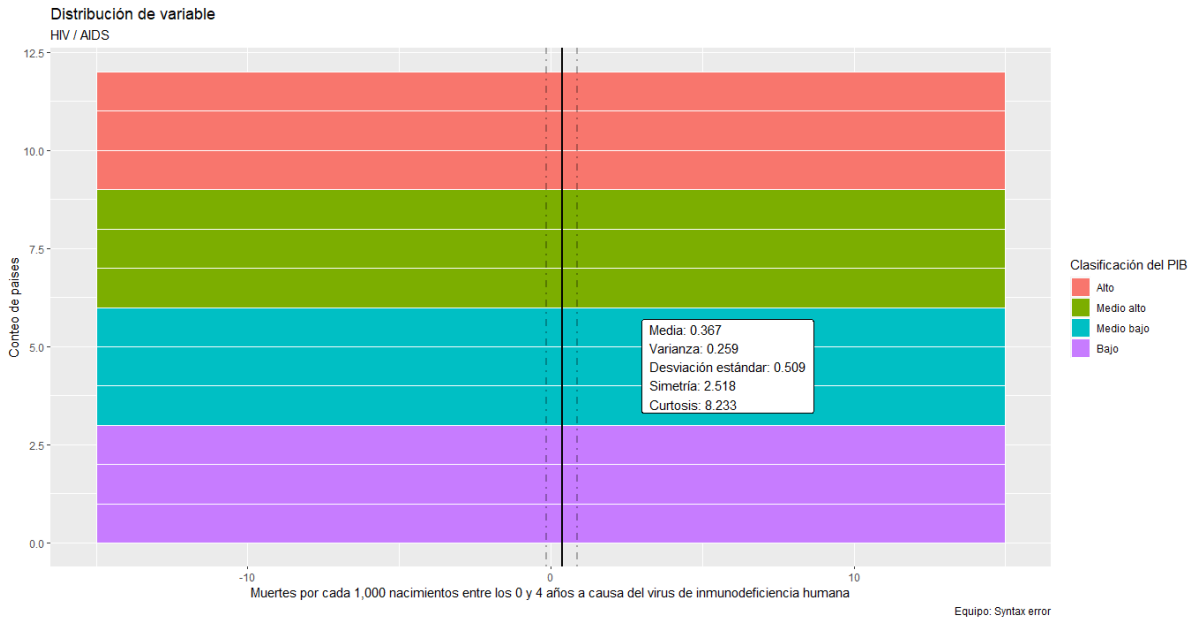

Fig 25. Análisis estadístico de la variable "Muertes por cada 1000 nacimientos entre los 0 y 4 años a causa del virus de inmunodeficiencia humana".

Hipótesis/descripción: Basándose en la forma del histograma, esta variable tiene datos muy similares entre sí para este año y países seleccionados, siendo que todos los valores menos uno tienen un valor de 0.5 o menor. Ya que la varianza y desviación estándar son mínimas, pero aún mayores a 0, se sabe que sí existe un valor un poco más alejado de los demás.

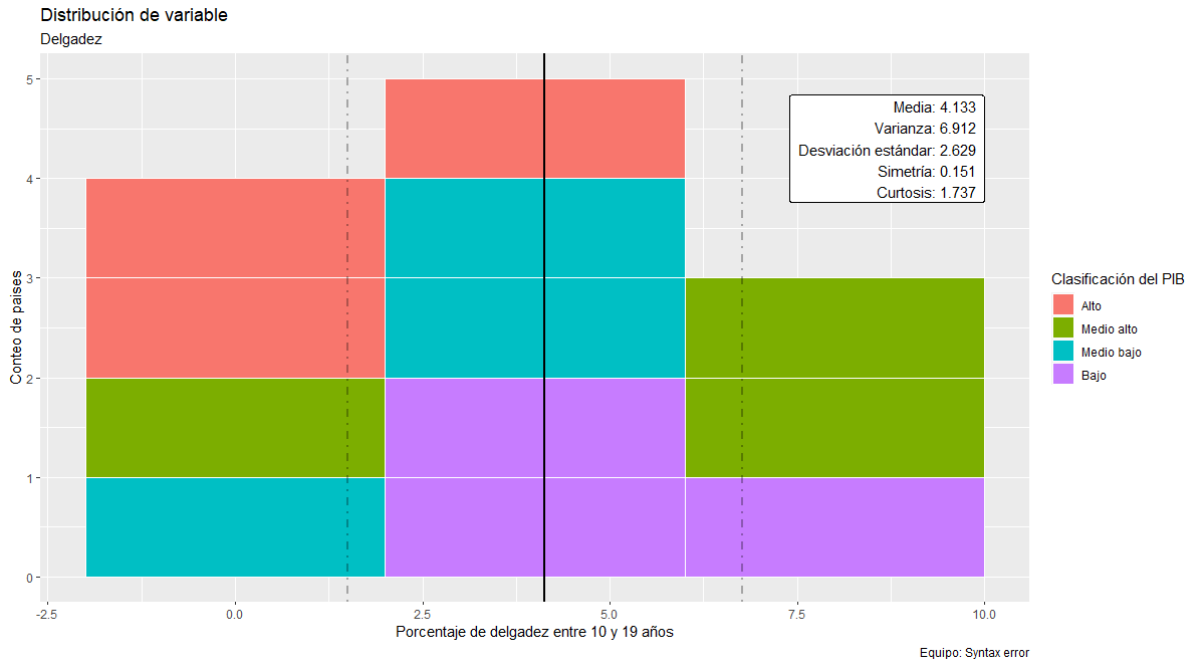

Fig 26. Análisis estadístico de la variable "Porcentaje de delgadez en personas de 10 a 19 años".

Hipótesis: Basándose en la forma del histograma, esta variable podría tener una distribución leptocúrtica, con la mayor concentración de datos en la media. Además, la distribución se encuentra con un valor de simetría muy cercano a 0, aunque aún sesgada ligeramente hacia la derecha. Por esto se entiende que existe mayor aglomeración de datos antes de la media.

De las variables de la base de datos. Incluye en esta sección la interpretación de los estadísticos descriptivos sobre dichas variables, es decir, ofrece una explicación sobre la varianza, media, desviación estándar. Estas son hipótesis sobre los datos. Recuerda redactar esto precisamente como hipótesis, y basarte en los números que obtuviste en el análisis. Por último, para cada una de las variables incluye también tu inferencia sobre qué distribución de probabilidad y explica el comportamiento de la variable. Por ejemplo: Basándonos en la forma del histograma, pensamos que esta variable podría tener una distribución normal de la información y se encuentra sesgado hacia la derecha por lo que podríamos asumir que la variable tiene una mayor cantidad de información en este rango de números.

Etapa 3: Modelación (opcional)

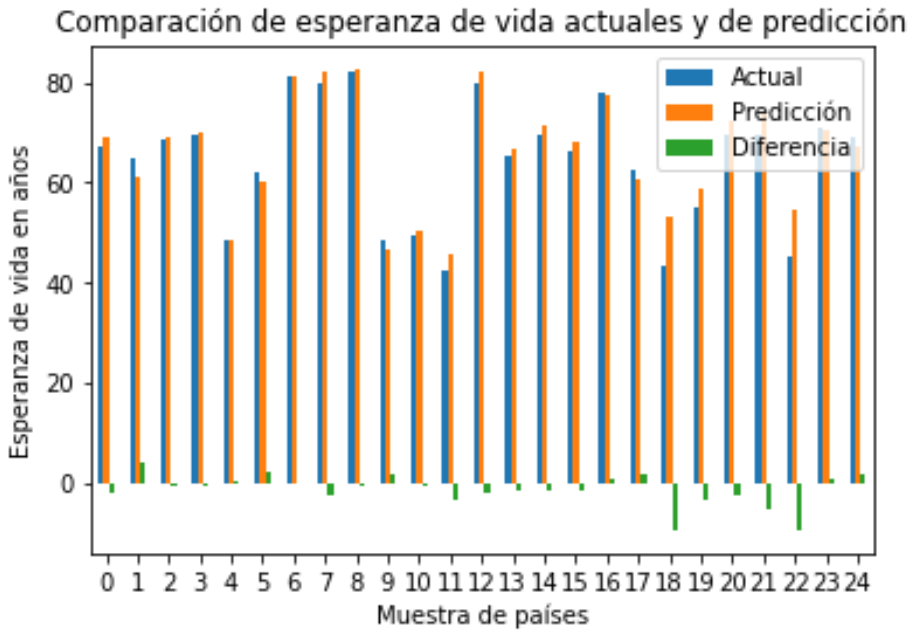

Fig. 27 Gráfica que compara las predicciones y las esperanzas de vida reales con datos de entrenamiento y prueba.

	Coeficientes
Adult.Mortality	-0.014954
infant.deaths	0.539447
percentage.expenditure	0.000250
Measles	-0.000031
BMI	-0.031840
under.five.deaths	-0.419335
Polio	0.009918
Diphtheria	0.026458
HIV.AIDS	0.176259
thinness..1.19.years	-0.266075
thinness.5.9.years	-0.327868
Income.composition.of.resources	43.744727
Schooling	-0.374415
GDP_wb	0.000002

Fig. 28 Coeficientes de regresión lineal por cada variable

	Actual	Predicción	Diferencia
0	67.2	69.162031	-1.962031
1	65.0	61.040330	3.959670
2	68.8	69.263578	-0.463578
3	69.4	70.211907	-0.811907
4	48.6	48.387814	0.212186
5	62.1	60.102576	1.997424
6	81.1	81.165737	-0.065737
7	79.9	82.264887	-2.364887
8	82.3	82.793860	-0.493860
9	48.7	46.789028	1.910972
10	49.6	50.296338	-0.696338
11	42.3	45.865842	-3.565842
12	79.8	82.002911	-2.202911
13	65.5	66.810108	-1.310108

Fig. 29 Matriz que compara datos reales, con datos de predicción que generó el modelo y calcula la diferencia entre ambos.

```
r2_score(y_test, y_pred)
```

```
0.9251789465723237
```

Fig. 30 Valor de la métrica R cuadrada o coeficiente de determinación que da la precisión del modelo.

La modelación que se realizó con los datos fue una regresión en el lenguaje de programación Python. Primeramente, se volvió a limpiar la base de datos, descartando aquellas columnas que contuvieran datos nulos o datos no numéricos, ya que, para realizar este modelo, se necesitan estrictamente datos numéricos. Posteriormente, con la librería sklearn, se importó la herramienta train_test_split, con la que se dividieron los datos en dos conjuntos: xdr de entrenamiento, tomando el 80% de datos, y datos de prueba, tomando el 20%. Después, se calcularon los coeficientes de cada variable relacionado al impacto que tienen sobre la esperanza de vida.

Con estos coeficientes, se construyó el modelo de regresión y se probó con datos de predicción, comparándolo con los valores reales y obteniendo la diferencia de años entre ambos valores. Se calculó la precisión del modelo, que en este caso es de 0.92. Esto significa que el modelo es bastante preciso ya que se acerca al valor 1. Finalmente, se graficaron estos valores, donde se puede ver gráficamente que la diferencia es mínima. Como guía para la realización de este código, se utilizó como inspiración un código perteneciente a la optativa de ciencia de datos y matemáticas para la toma de decisiones (Domínguez & Aguilar, 2021).

Etapa 4: Conclusiones

Para concluir, después de haber realizado la exploración de datos, el análisis estadístico y el modelo de regresión, hay varios factores que resaltan como aquellos que generan más impacto sobre la esperanza de vida. Es por esto por lo que, la creación y desarrollo de iniciativas enfocadas en algunos de estos factores podría ser de gran ayuda para incrementar la esperanza de vida en distintos Estados.

El análisis exploratorio arrojó cuatro variables principales que afectan la esperanza de vida: escolaridad, Producto Interno Bruto per cápita, distribución de la riqueza, muertes en la infancia y niñez, y la prevalencia de enfermedades infecciosas (en especial, el virus de inmunodeficiencia humana y la enfermedad que provoca, el síndrome de inmunodeficiencia adquirida). Mediante el análisis estadístico, pudimos conocer la distribución de los países (agrupados según su rango en el PIB per cápita) en las diferentes variables incluidas en la base de datos obtenida de la Organización Mundial de la Salud. A partir de ello, encontramos los siguientes datos de relevancia:

- La esperanza de vida y la mortalidad (tanto en adultos como niños) están relacionadas directamente con el Producto Interno Bruto per cápita de cada país, de manera general.
    - Parte significativa de países clasificados con un PIB medio alto presentan una elevada mortalidad infantil (tanto en menores de 1 año como en menores de 5).
- Los países con un PIB bajo cuentan con más niños inmunizados contra enfermedades infecciosas (poliomelitis, hepatitis B y difteria) a comparación de países clasificados como PIB alto y medio alto.
    - Dado que la mortalidad en niños es menor en estos países, se puede señalar a las vacunas como uno de los factores a considerar para reducir los decesos en niños.
- Los países con un producto interno bruto alto y medio bajo registran valores de IMC mayores que aquellos clasificados como bajo y medio alto.
- Los países clasificados como PIB bajo cuentan con los menores valores de índice de masa corporal y registran los mayores porcentajes de población con delgadez.
    - Se puede inferir que dar un enfoque hacia la nutrición sería beneficioso para dichos territorios en busca de reducir la mortalidad.
- La escolaridad está directamente relacionada con el valor del PIB per cápita.
- Los países con un PIB per cápita bajo y medio bajo presentan los menores niveles de gasto público (tanto con respecto al presupuesto gubernamental como el PIB).

A partir de estas y otras consideraciones, se definieron algunas propuestas de solución ante el objetivo planteado de encontrar relaciones entre las variables disponibles en la base de datos para mejorar la esperanza de vida. En primer lugar, se propone que todos los países inviertan en salud pública, lo cual incluye seguir fomentando la inmunización contra enfermedades infecciosas, una cultura de prevención de enfermedades como el síndrome de inmunodeficiencia adquirida (siendo que este es la mayor causa de reducción en la esperanza de vida) y mejores hábitos alimenticios, de modo que se busque reducir el porcentaje de delgadez en países clasificados con un PIB bajo o medio bajo, o bien, reducir enfermedades como la obesidad (que se puede inferir padecen aquellos países con valores de IMC altamente elevados, presentes en aquellos dentro de las categorías de PIB alto y medio alto). En segundo lugar, para aquellos Estados que fueron clasificados con un Producto Interno Bruto medio-bajo y bajo, incrementar la inversión en el sector educativo, puesto que mejorar la escolaridad está directamente relacionada con una mayor supervivencia en infantes y niños, lo que se traduce en adultos más longevos.

Referencias

Domínguez, G. & Aguilar, J. (2021). *Guía para: Implementación y visualización de una regresión lineal múltiple en Python*. Noviembre 11, 2021, de ITESM Sitio web:

https://experiencia21.tec.mx/courses/198531/pages/guia-para-implementacion-y-visualizacion-de-una-regresion-lineal-multiple-en-python?module_item_id=9163170
INEGI. (2020). *Población. Esperanza de vida.* Recuperado de: http://cuentame.inegi.org.mx/poblacion/esperanza.aspx?tema=P
Life Expectancy (WHO) (2015). *Statistical Analysis on factors influencing Life Expectancy*. Kaggle. Recuperado de: https://www.kaggle.com/kumarajarshi/life-expectancy-who
Naciones Unidas (2021). *Garantizar una vida sana y promover el bienestar de todos a todas las edades.* Objetivos de Desarrollo Sostenible. Departamento de Asuntos Económicos y Sociales. Recuperado de: https://sdgs.un.org/es/goals/goal3