# **IMFORME EDA**

# **Empresa Farmacéutica**

**PROYECTO**: Análisis de la Esperanza de Vida al Nacer para Factibilidad de lanzamiento de multivitamínico a nivel Global

## **1. Introducción**

En el epicentro de la toma de decisiones informadas se encuentra el análisis exploratorio de datos (EDA), una práctica esencial que arroja luz sobre los matices y patrones ocultos dentro de conjuntos de datos vastos y complejos. Este informe se sumerge en un EDA exhaustivo, basado en ocho conjuntos de datos extraídos del Banco Mundial, cada uno de los cuales abarca factores cruciales que potencialmente inciden en la esperanza de vida.

Los datos, procedentes de una fuente confiable como el Banco Mundial, se han seleccionado meticulosamente para ofrecer una visión holística de los 44 países en estudio, distribuidos a lo largo de los 5 continentes. Estos conjuntos abarcan aspectos que van desde indicadores económicos y sociales hasta datos demográficos, permitiéndonos desentrañar las complejidades que subyacen a la variabilidad en la esperanza de vida.

A través de este EDA, no solo buscamos describir y visualizar la distribución de variables clave, sino también identificar relaciones y tendencias significativas entre los diversos factores. Al comprender la interconexión entre variables como el Producto Interno Bruto, la educación, la nutrición y otros indicadores relevantes, se espera revelar insights valiosos que orientarán la toma de decisiones estratégicas.

## **Data sets**

Los conjuntos de datos fueron descargados directamente desde la página oficial del Banco Mundial, accediendo específicamente al siguiente enlace:

<https://databank.worldbank.org/databases>

Este enlace proporciona acceso a una amplia variedad de series que se pueden filtrar según las necesidades específicas. Para el presente proyecto, se ha optado por descargar los factores más alineados con los objetivos establecidos, los cuales incluyen:

1. Longevidad
2. Población
3. Economía
4. Educación
5. Inversión en Salud

Dibujo en blanco y negro

Descripción generada automáticamente con confianza media Para analizar cada conjunto de datos, se importaron las siguientes librerías:

**DASHBOARD INTERACTIVO**

## 

## **Factor Educación**

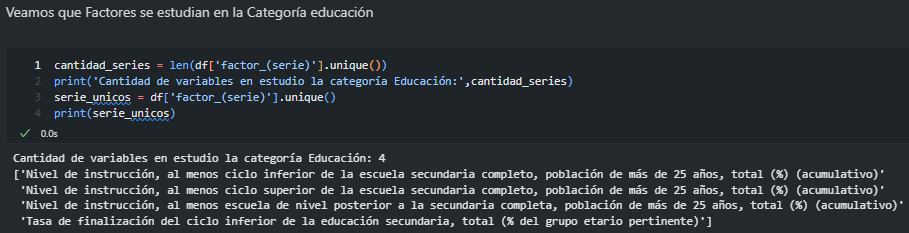
Este Dataset comprende un extenso periodo, desde 1987 hasta 2022, con un rango variable de observaciones, desde un mínimo de 3 a un máximo de 89. Las desviaciones estándar permanecen consistentes, entre 25 y 32, destacando la dispersión de los datos en cada año. La media de los indicadores se mantiene alrededor de 60-70, mostrando cierta estabilidad a lo largo de las décadas. Los valores mínimo y máximo varían notablemente, desde 0 hasta 12.60 como mínimo, y desde 99.86 hasta 129.95 como máximo, evidenciando la amplitud y diversidad de los datos a lo largo del tiempo. En resumen, el conjunto de datos proporciona un amplio espectro para analizar tendencias, cambios y distribución de indicadores específicos para múltiples países a lo largo de un amplio período temporal.

El conjunto de países en estudio refleja una diversidad geográfica significativa, representando un total de 43 naciones. Esta recopilación abarca una amplia distribución global, incluyendo países de distintos continentes, como Asia con naciones como China, India, y Japón; Europa con Alemania, España y Noruega; América con Estados Unidos, Argentina y Venezuela; África con Argelia, Nigeria y Rwanda; Oceanía con Australia, Fiji y Nueva Zelanda, entre otros. Esta diversidad geográfica sugiere una representación global en los datos, proporcionando la oportunidad de estudiar y comparar patrones, tendencias y características distintivas de cada continente en el análisis.

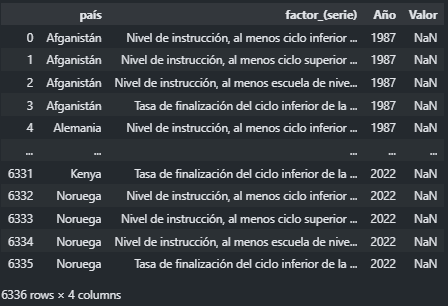
Un primer vistazo al conjunto de datos se obtuvo ejecutando df.head() en Pandas.



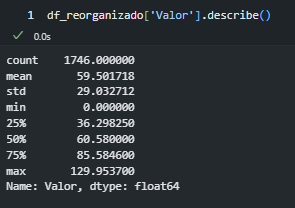
Visualicemos los factores de la serie presentes en este conjunto de datos.

****

Para facilitar la comprensión y análisis de los datos, se ha optado por reorganizar el dataframe de la siguiente manera:

****

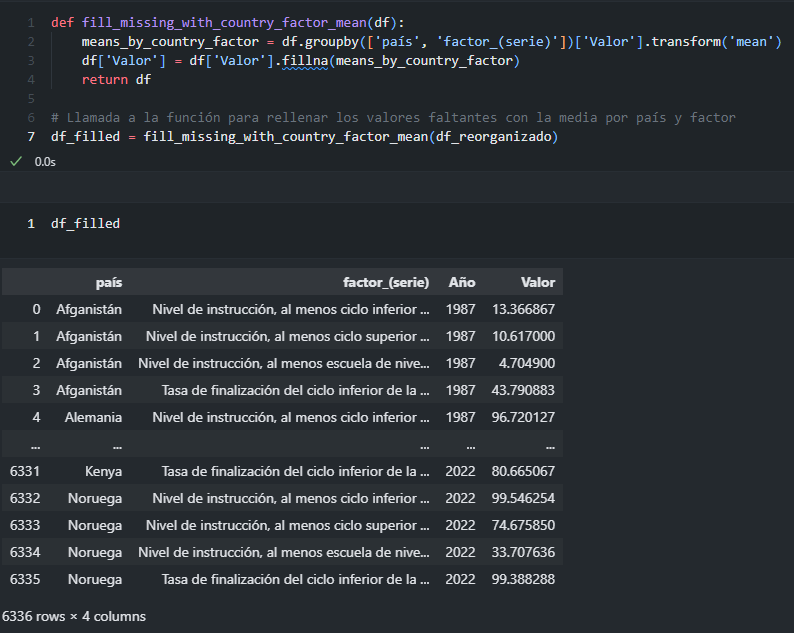
Aplicamos la función .describe() para obtener un resumen estadístico del conjunto de datos.

****

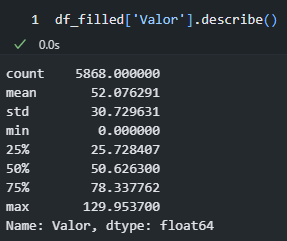
Como se puede observar en el análisis anterior con un total de 1746 registros no nulos, se observa que la media de los valores es aproximadamente 59.50, con una dispersión significativa reflejada por una desviación estándar de alrededor de 29.03 unidades. El rango de valores es considerable, oscilando desde un mínimo de 0 hasta un máximo de 129.95. La mediana, situada en aproximadamente 60.58, indica que el 50% de los datos caen por debajo de este valor, mientras que el otro 50% se encuentra por encima. Los percentiles muestran que el 25% de los datos son inferiores a 36.30 y el 75% son inferiores a 85.58, evidenciando la distribución de los valores en el conjunto de datos. Estos datos resumen la variabilidad, la centralidad y la amplitud de la información presente en la columna 'Valor'.

## **Imputación**

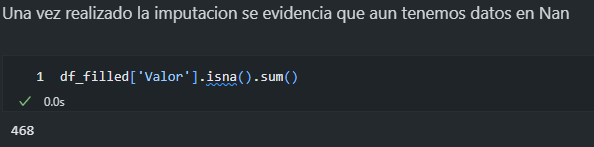
Ante la presencia significativa de valores faltantes (NaN) en nuestro conjunto de datos, se optó por aplicar una estrategia de imputación. Esta estrategia consiste en el reemplazo de los valores faltantes por la media de los datos, tomando en consideración la especificidad de cada país como también la del factor en estudio. Este enfoque nos permite abordar la ausencia de datos, asegurando una mayor integridad en nuestro conjunto de datos al evitar vacíos o valores faltantes, lo que resulta crucial para un análisis robusto y confiable.

****

Volvemos aplicar la función .describe() para obtener un resumen estadístico del conjunto de datos, una vez realizado la imputación

****

Verificamos la presencia de datos Nan, resulta que hay países que no tienen datos en ciertos fatores en ninguno de los años seleccionados.

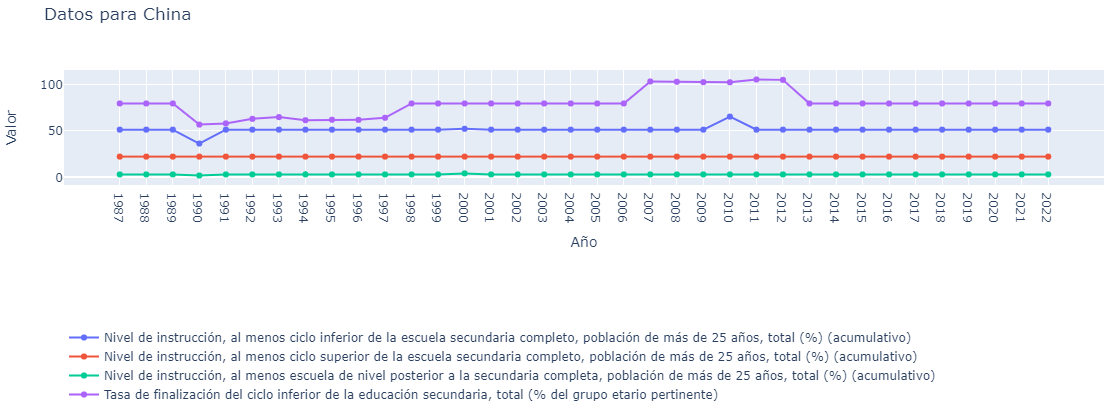
****

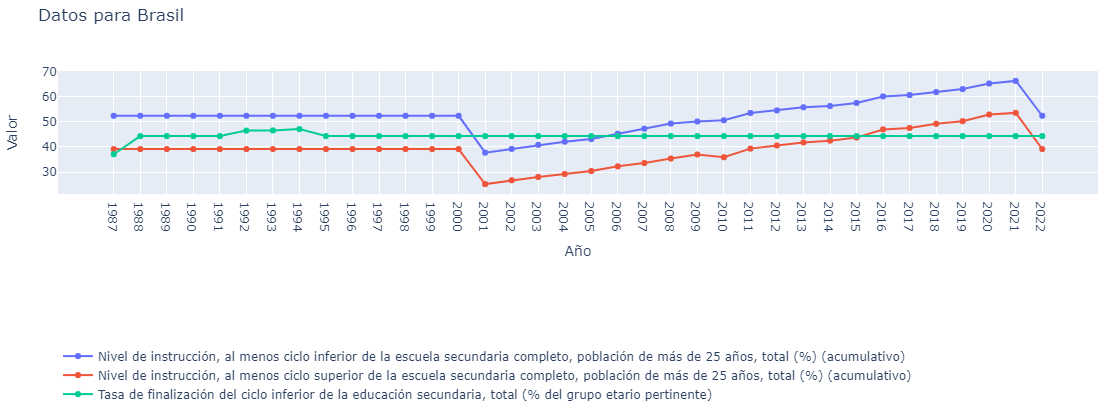
Por lo tanto, procedemos a eliminar los datos correspondientes a aquellos factores para los cuales no existen registros en ninguno de los años. Esta decisión se basa en evitar la generación de datos ficticios que podrían distorsionar la integridad del conjunto de datos, reconociendo las particularidades de cada país.

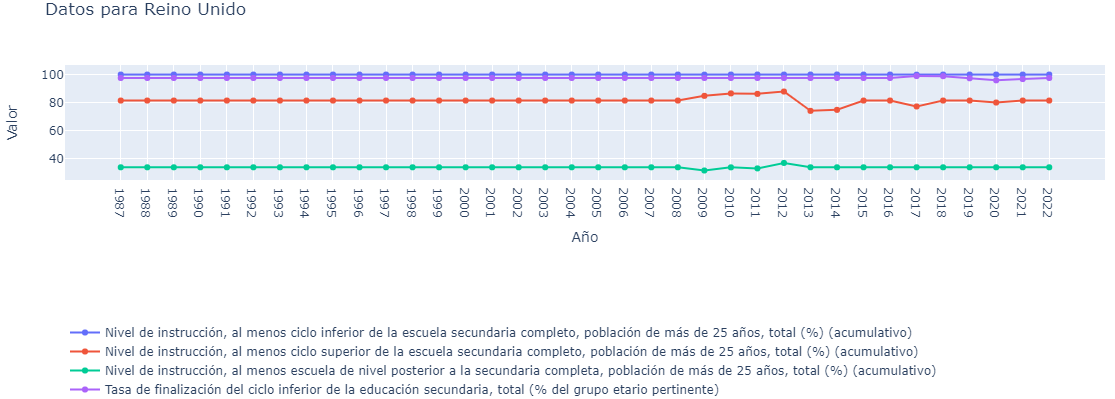
****

## **Distribución**

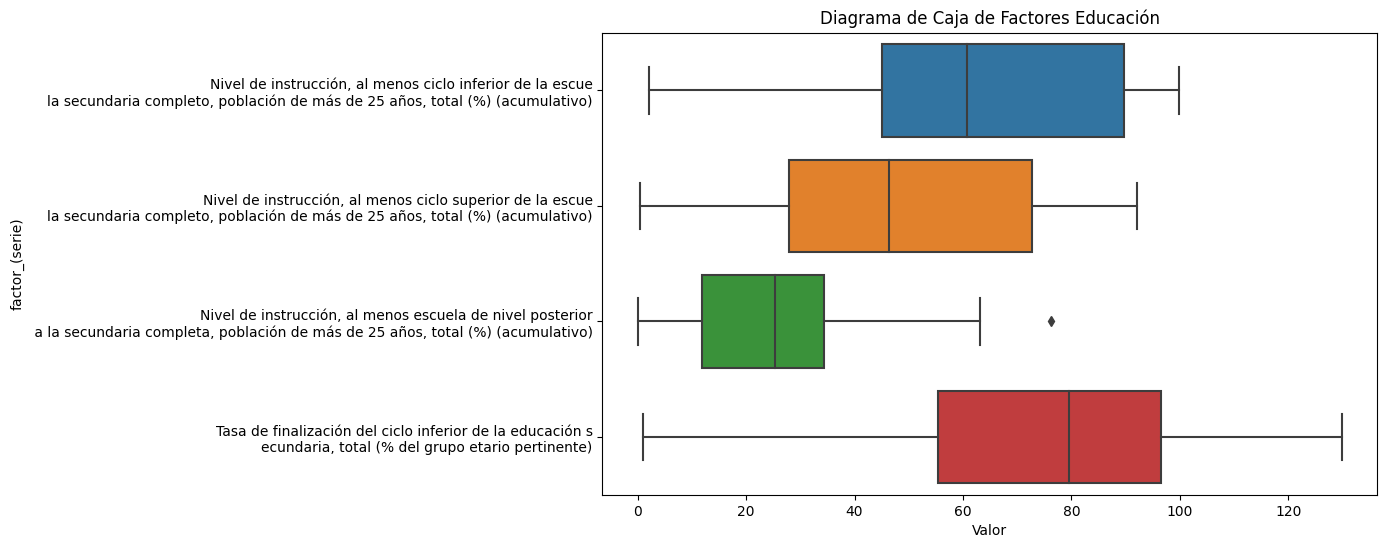
Ahora que contamos con la mayoría de los datos, procedemos a visualizar la distribución mediante gráficos. Se han seleccionado aleatoriamente algunos países para este análisis:







## **Diagrama Boxplot**

****

1.- Nivel de instrucción, al menos ciclo inferior de la escuela secundaria completo:

Este factor presenta un promedio del 60.47%, con una desviación estándar de 28.87%. El rango intercuartílico (IQR) abarca desde el 25% al 75%, con valores que oscilan entre el 45.16% y el 89.69%. El valor mínimo observado es 2.04%, mientras que el máximo alcanza el 99.86%.

2.- Nivel de instrucción, al menos ciclo superior de la escuela secundaria completo:

El nivel de instrucción, al menos ciclo superior de la escuela secundaria completo, tiene un promedio del 47.23%, con una desviación estándar de 26.09%. Su rango intercuartílico (IQR) va desde el 27.96% hasta el 72.74%. Los valores extremos registrados son 0.50% como mínimo y 92.18% como máximo.

3.- Nivel de instrucción, al menos escuela de nivel posterior a la secundaria completa:

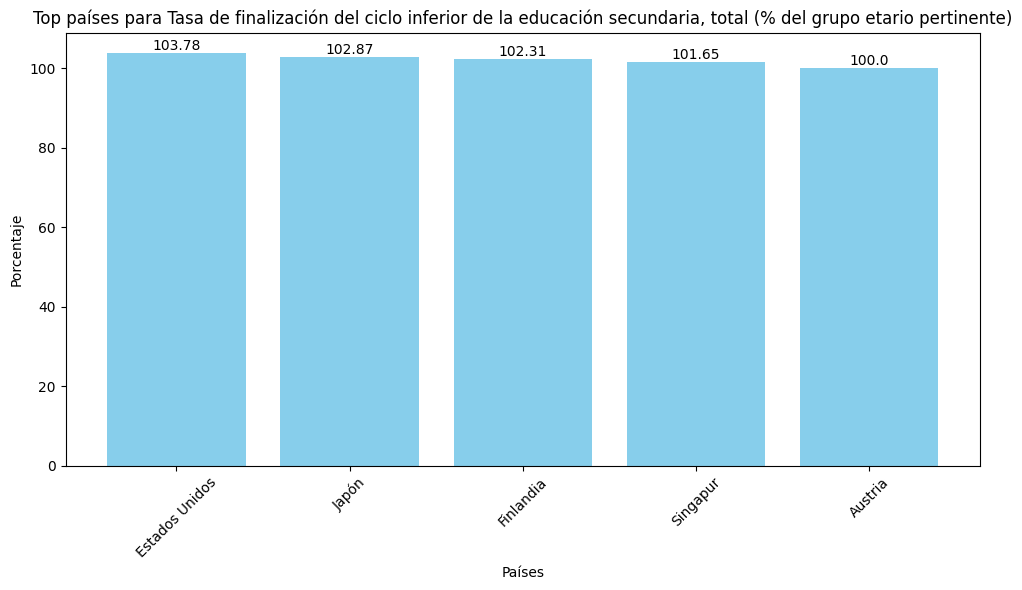
Este factor presenta un promedio del 24.28%, con una desviación estándar de 15.85%. El rango intercuartílico (IQR) abarca desde el 25% al 75%, con valores que oscilan entre el 11.92% y el 34.38%. El valor mínimo registrado es 0%, mientras que el máximo alcanza el 76.29%.

4.- Tasa de finalización del ciclo inferior de la educación secundaria:

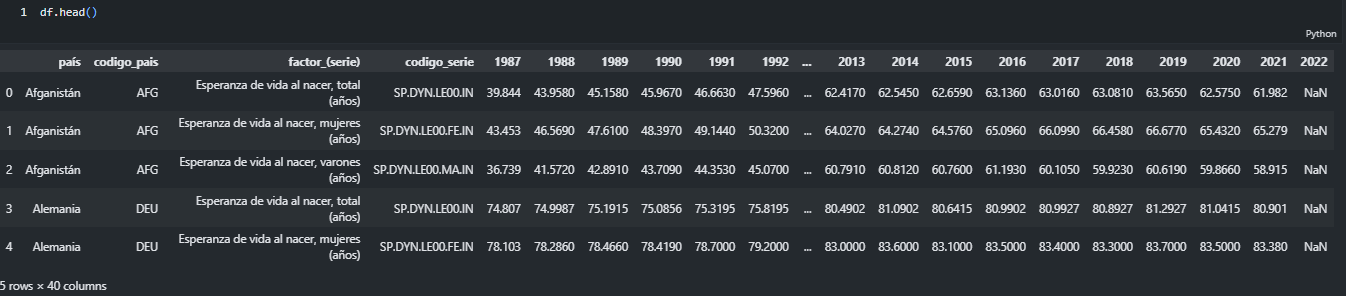
La tasa de finalización del ciclo inferior de la educación secundaria exhibe un promedio del 72.75%, con una desviación estándar de 27.24%. El rango intercuartílico (IQR) comprende desde el 25% al 75%, con valores que varían entre el 55.39% y el 96.55%. La cifra mínima registrada es 0.92%, mientras que la máxima alcanza el 129.95%.

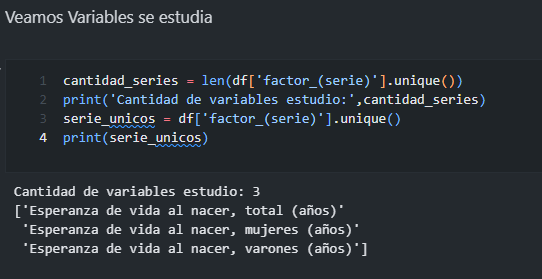
## **El mejor Factor de Educación y su top 5 países**

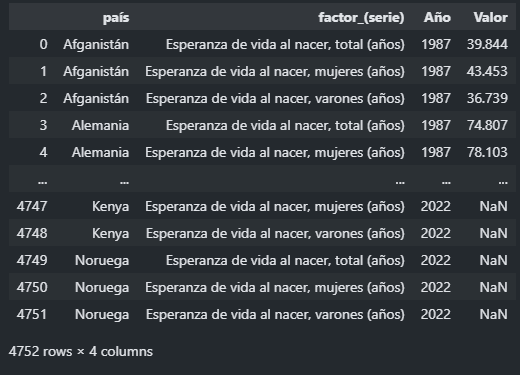
El siguiente gráfico presenta los cinco países líderes en la Tasa de Finalización del Ciclo Inferior de la Educación Secundaria, expresada como un porcentaje del grupo etario correspondiente. Este indicador es crucial para evaluar el éxito del sistema educativo en proporcionar una educación completa a los estudiantes. En la cima de la lista se encuentra Estados Unidos con un impresionante 103.78%, seguido de cerca por Japón con 102.87%. Finlandia, reconocida por su excelencia educativa, ocupa el tercer lugar con un índice del 102.31%, seguida por Singapur con 101.65%. Cerrando la lista, Austria alcanza una tasa de finalización del 100.00%. Estos resultados destacan el compromiso y la eficacia de estos países en garantizar que sus jóvenes completen con éxito la etapa inicial de la educación secundaria.

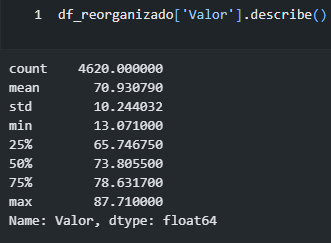


**Factor Esperanza de vida**

****

****

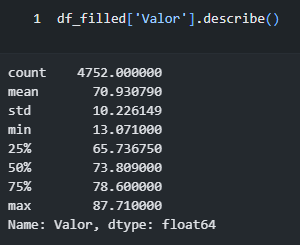
****

****

El resumen estadístico de la serie revela la presencia de 4620 observaciones. La media de los valores es de aproximadamente 70.93, con una desviación estándar de alrededor de 10.24, indicando una dispersión moderada de los datos respecto a la media. Los valores varían desde un mínimo de 13.07 hasta un máximo de 87.71. Los percentiles sugieren que el 25% de los valores son menores que 65.75, el 50% son inferiores a 73.81 y el 75% son menores que 78.63. Esto indica que la mayoría de los datos se concentran entre el percentil 25 y 75, con una dispersión más amplia hacia los valores máximos y mínimos.

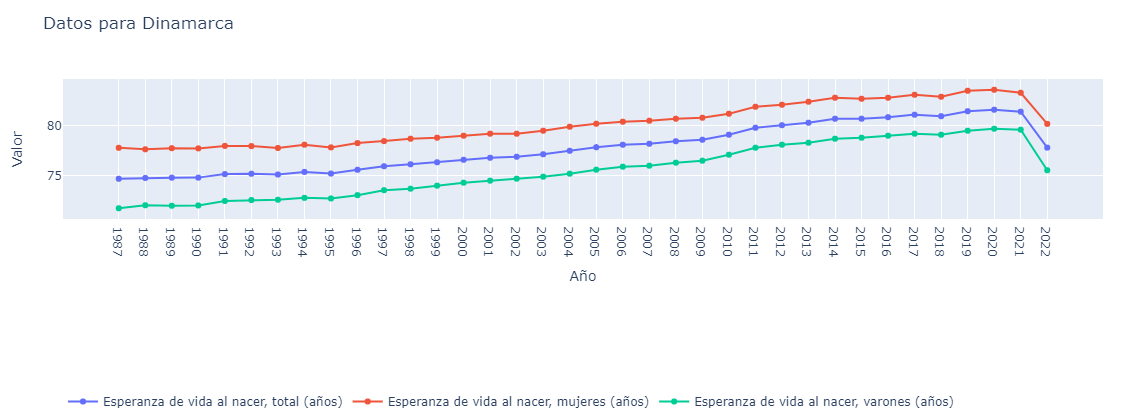
Imputacion

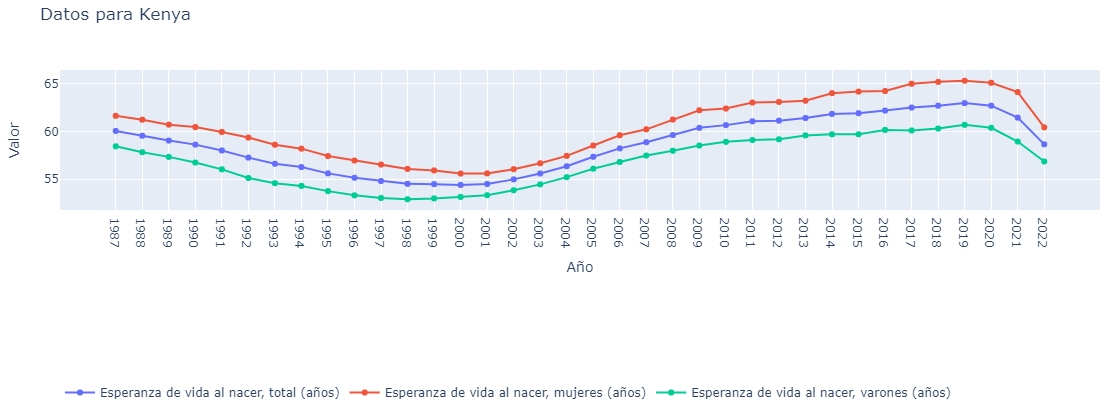
Ante la presencia significativa de valores faltantes (NaN) en nuestro conjunto de datos, se optó por aplicar una estrategia de imputación. Esta estrategia consiste en el reemplazo de los valores faltantes por la media de los datos, tomando en consideración la especificidad de cada país como también la del factor en estudio. Este enfoque nos permite abordar la ausencia de datos, asegurando una mayor integridad en nuestro conjunto de datos al evitar vacíos o valores faltantes, lo que resulta crucial para un análisis robusto y confiable.

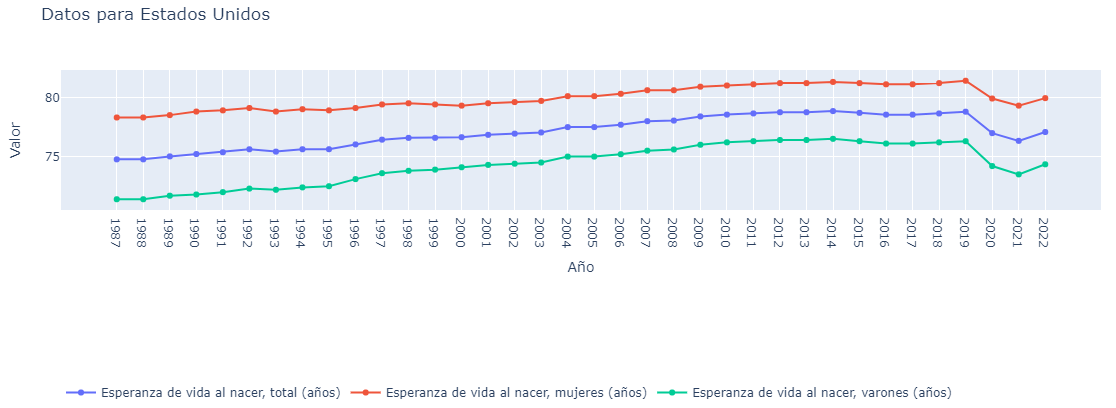
****

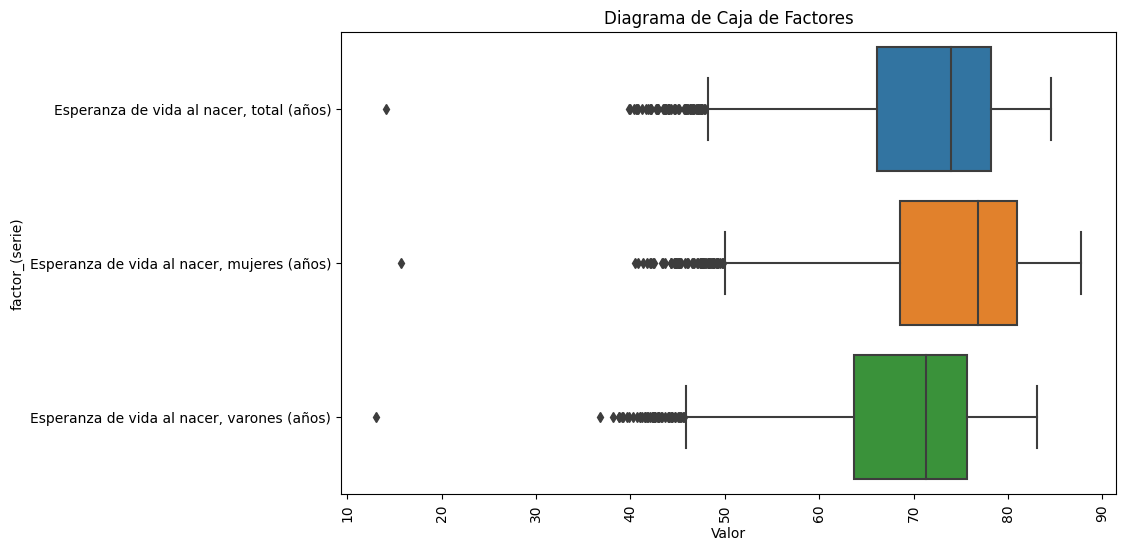
El resumen estadístico de la serie comprende un total de 4752 observaciones. Los valores muestran una media de aproximadamente 70.93 con una desviación estándar de alrededor de 10.22, lo que sugiere una dispersión moderada en relación a la media. Los valores oscilan entre un mínimo de 13.07 y un máximo de 87.71. Los percentiles muestran que el 25% de los datos son menores que 65.75, el 50% son inferiores a 73.8, y el 75% son menores que 78.64. Esto indica que la mayor parte de los datos se concentran entre los percentiles 25 y 75, mientras que hay una dispersión más amplia hacia los valores máximos y mínimos.

**Distribución**

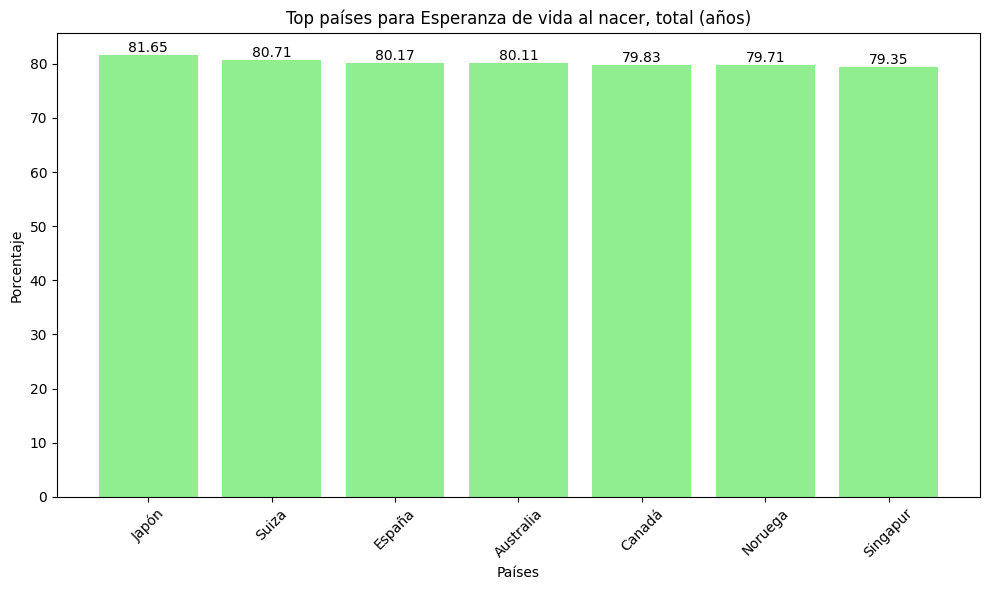
****

****

****

****

Al observar los tres factores representados en un gráfico de caja, cada uno muestra particularidades significativas en la distribución de la esperanza de vida al nacer. 'Esperanza de vida al nacer, total' exhibe un rango intercuartílico (IQR) que va desde 66.18 hasta 78.20 años, con una mediana en 73.98 años y valores extremos desde 14.10 hasta 84.56 años. En el caso de 'Esperanza de vida al nacer, mujeres', el IQR abarca desde 68.57 hasta 81.00 años, con una mediana en 76.83 años y valores extremos entre 15.66 y 87.71 años. Por su parte, 'Esperanza de vida al nacer, varones' muestra un IQR que se extiende desde 63.73 hasta 75.63 años, con una mediana en 71.31 años y valores extremos de 13.07 a 83.10 años. Estas diferencias reflejan variaciones sustanciales en la distribución de la esperanza de vida al nacer entre los grupos representados, con distintos rangos y medias, lo que ilustra las particularidades en la duración de vida entre los distintos conjuntos demográficos.

****