Segmentación de Países según Acceso a Internet y Esperanza de Vida

Introducción a Ciencia de Datos

Guillermo Aguilar Martínez

Noviembre del 2025

1 Introducción y objetivo

El acceso a internet y la esperanza de vida son indicadores que reflejan dimensiones distintas pero estrechamente relacionadas del desarrollo humano. Mientras que el acceso digital permite medir el avance tecnológico, la esperanza de vida es uno de los indicadores más tradicionales del bienestar social y de las condiciones de salud pública.

El objetivo de este estudio es doble:

1. **Segmentar países según su trayectoria de acceso a internet y esperanza de vida** mediante Análisis de Componentes Principales (PCA) y métodos de agrupamiento (k-means).
2. **Evaluar la relación estadística entre ambos indicadores** mediante la construcción de una matriz de correlación de Pearson entre ambos paneles multitemporales (1990–2020).

Este análisis permite estudiar simultáneamente dos dimensiones clave del desarrollo y explorar si los patrones de digitalización están asociados con mejoras en salud poblacional.

2 Descripción y pre-procesamiento de los datos

Los datos provienen de *Our World in Data*. Se utilizaron dos archivos:

- `number-of-internet-users.csv`: número de usuarios de internet por país.
- `life-expectancy.csv`: esperanza de vida al nacer por país.

Ambos datasets contienen series temporales tipo país–año con extensiones similares, lo cual permite un preprocesamiento similar:

1. se eliminaron agregados no nacionales y entradas sin código ISO;

2. se seleccionó el mismo rango temporal 1990–2020;
3. los datos se pivotearon a una matriz rectangular país–año;
4. se eliminaron países con más del 20% de datos faltantes;
5. se aplicó interpolación lineal en cada país;
6. finalmente se normalizaron todas las columnas mediante

$$Z = \frac{X - \mu}{\sigma}.$$

3 Técnicas Estadísticas Empleadas

3.1 Reducción de dimensión: PCA

El Análisis de Componentes Principales (PCA) se utilizó tanto para el panel de internet como para el de esperanza de vida. Sea Z la matriz estandarizada, entonces:

$$X_{PCA} = ZV,$$

donde V es la matriz de autovectores de la matriz de covarianza. Este procedimiento permite proyectar cada país en un espacio de baja dimensión que preserva la mayor parte de la variabilidad.

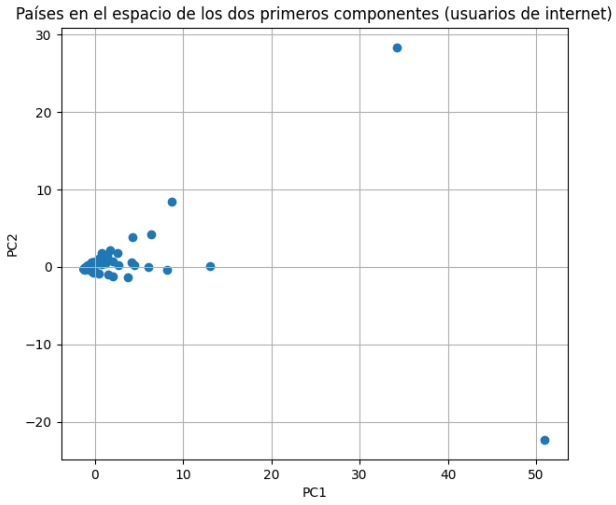

Figure 1: PCA sin clusters (Internet)

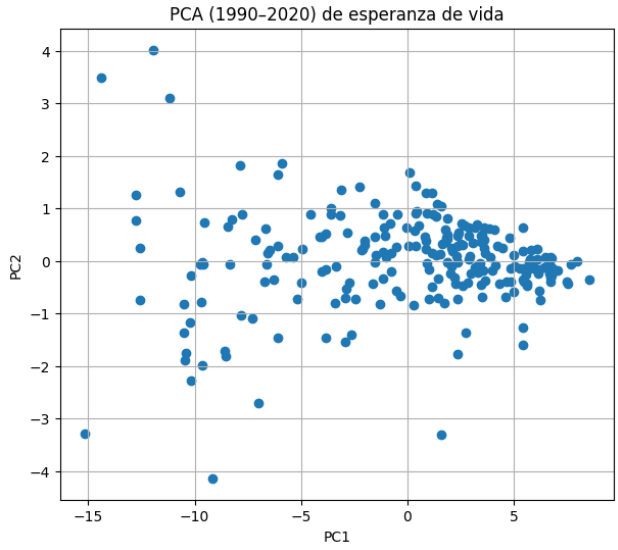

Figure 2: PCA sin clusters (Esperanza de vida)

La fracción de varianza total que capturan las primeras dos componentes:

	PC1	PC2	fracción total
Internet	0.73941428	0.24063285	0.98004713
Esperanza de vida	0.95015788	0.0271453	0.97730318

Figure 3: Varianzas explicadas

3.2 Interpretación de PC1 y PC2

En ambos indicadores (usuarios de internet y esperanza de vida), el PC1 representa el **patrón temporal general**: países con valores altos muestran un crecimiento sostenido a lo largo del periodo, mientras que valores bajos indican trayectorias más lentas o estancadas.

El PC2, en cambio, refleja **cómo cambia la forma de la trayectoria**: valores positivos suelen asociarse a países que aceleraron su crecimiento en años recientes, mientras que valores negativos corresponden a países cuyo aumento fue más temprano o que se estabilizaron antes.

En conjunto, PC1 mide la **magnitud global del desarrollo** y PC2 describe **diferencias en la dinámica temporal** entre países.

3.3 Selección del número óptimo de clusters

1. **Inercia (SSE)**: Define la suma de las distancias cuadráticas entre cada observación y su centroide asignado.

$$\text{Inercia}(k) = \sum_{i=1}^{n} \|x_i - \mu_{c(i)}\|^2.$$

Valores menores indican clusters más compactos. El comportamiento de la inercia en función de k se utiliza en el *método del codo* para identificar el punto en el que aumentar

el número de clusters deja de producir mejoras significativas.

2. **Coeficiente de Silueta**: Evalúa simultáneamente la cohesión interna de los clusters y la separación entre ellos. Para cada observación i:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

donde $a(i)$ es la distancia media de i a su propio cluster y $b(i)$ es la distancia mínima a los demás clusters. El valor global de silueta es la media de todos los $s(i)$ y toma valores en el intervalo $[-1, 1]$. Valores más cercanos a 1 indican agrupamientos bien separados.

Ambos análisis se realizaron en el espacio reducido de PCA para asegurar que la métrica euclidiana empleada por k-means fuera significativa.

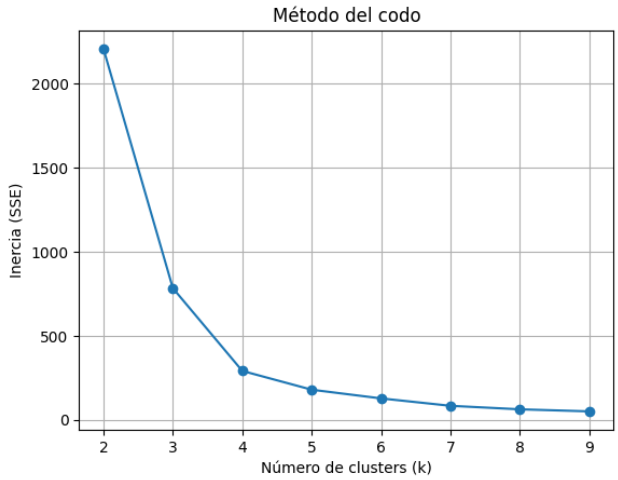

Figure 4: Inercia vs k (Internet)

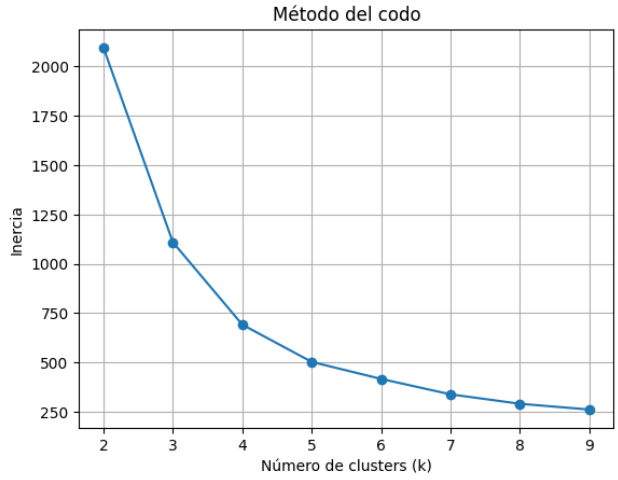

Figure 5: Inercia vs k (Esperanza de vida)

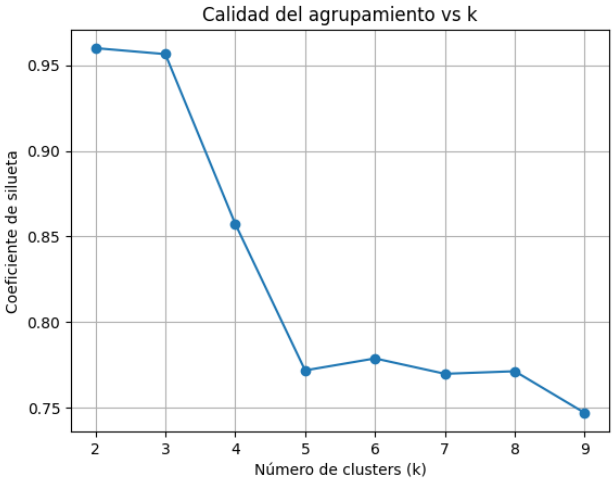

Figure 6: Coeficiente de Silueta vs k (Internet)

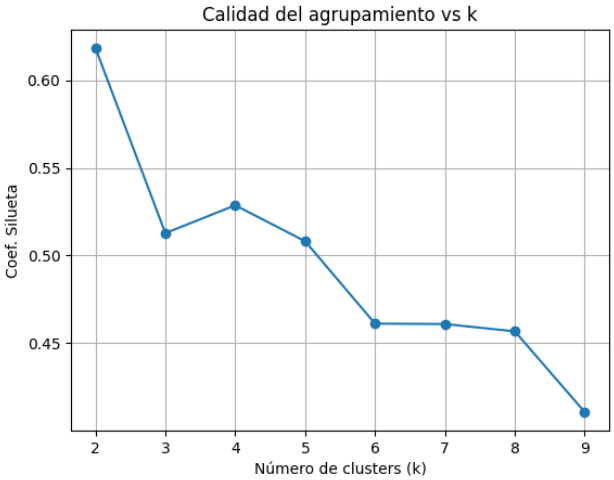

Figure 7: Coeficiente de Silueta vs k (Esperanza de vida)

Los gráficos sugieren $k = 5$ clusters tanto para el dataset de internet como para el de esperanza de vida.

4 Resultados del Análisis de Internet

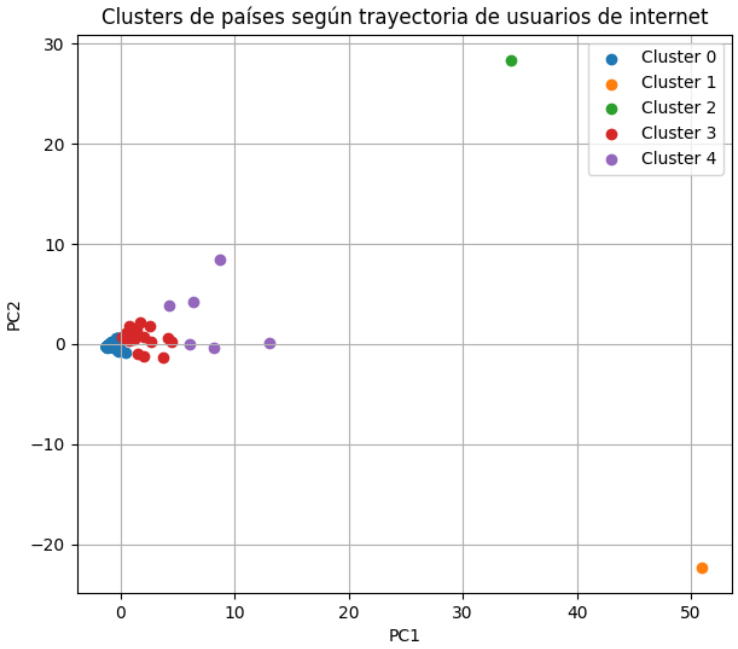

Figure 8: Clusters de países según acceso a internet

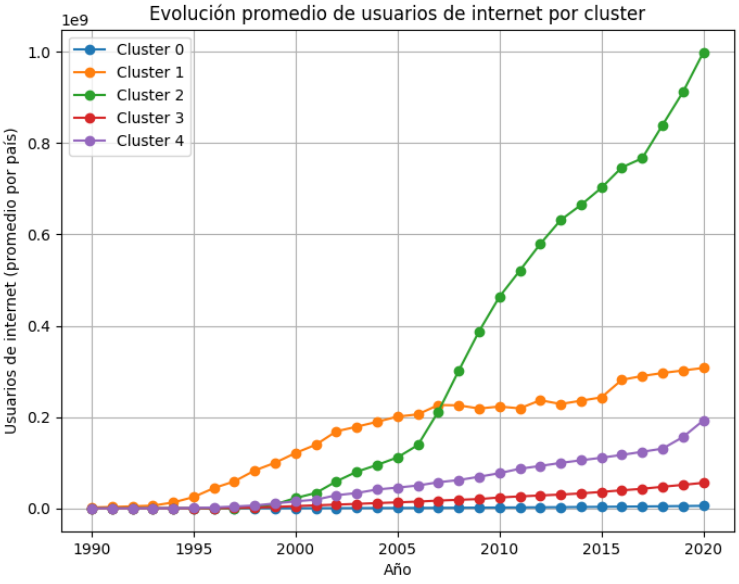

Figure 9: Evolución promedio de usuarios de internet por cluster

La segmentación evidencia diferencias claras entre regiones altamente digitalizadas (Europa Occidental, Norteamérica, asiáticos desarrollados), países emergentes y regiones de baja penetración digital. En términos generales, se observan:

- un grupo con acceso muy alto (principalmente países de Europa occidental, Norteamérica y algunos de Asia oriental),
- un grupo intermedio (varios países de Sudamérica y Europa del Este),
- un grupo con acceso bajo (diversos países de África subsahariana y Asia en desarrollo).

5 Resultados del Análisis de Esperanza de Vida

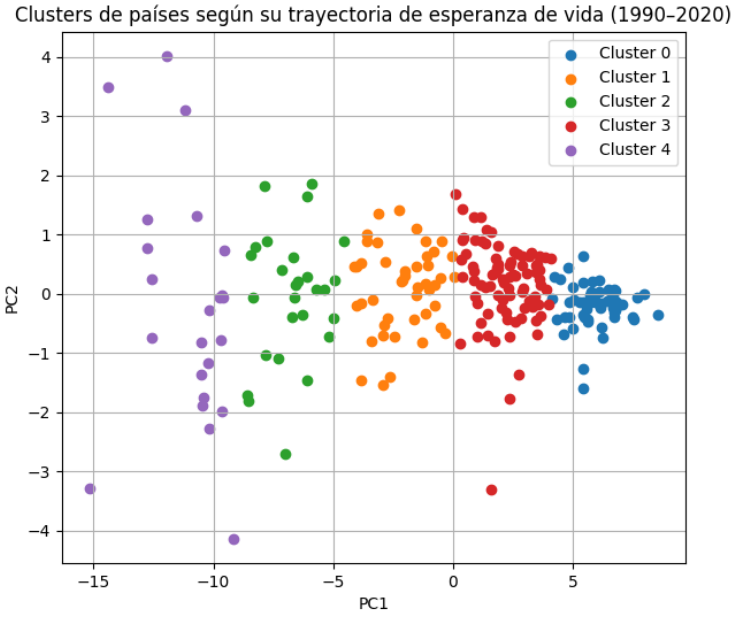

Figure 10: Clusters de países según esperanza de vida

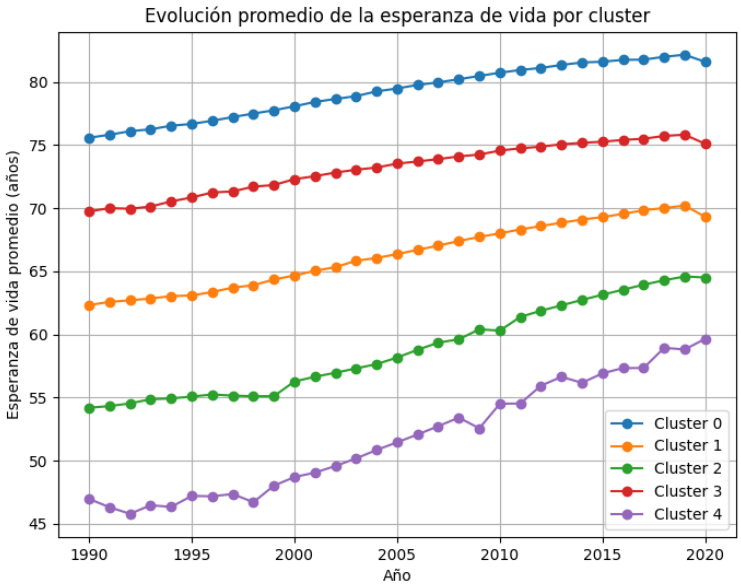

Figure 11: Evolución promedio de esperanza de vida por cluster

El PCA aplicado al panel de esperanza de vida produjo una estructura muy similar: los primeros dos componentes capturan la mayor parte de la variación temporal entre países, revelando trayectorias diferenciadas en términos de bienestar y salud pública.

Los clusters obtenidos mediante k-means muestran:

- un grupo de países con esperanza de vida consistentemente alta (Europa Occidental, Japón, Corea del Sur),
- un grupo intermedio (Latinoamérica, Europa del Este),
- un grupo con valores bajos o afectados por inestabilidad histórica (África subsahariana).

Esta segmentación complementa la obtenida a partir del acceso a internet.

6 Correlación entre Acceso a Internet y Esperanza de Vida

Para evaluar la relación entre ambos indicadores, se construyó una matriz de correlación de Pearson entre los paneles multitemporales (1990–2020).

Dado que los dos datasets fueron estandarizados y alineados país–año, la correlación resultante mide:

- cómo varía la esperanza de vida de un país con respecto a su trayectoria de adopción de internet,
- qué tan similares son las formas temporales de ambas series,

- si existe correspondencia entre el desarrollo digital y el bienestar social.

La correlación de Pearson entre ambas series se definió como:

$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

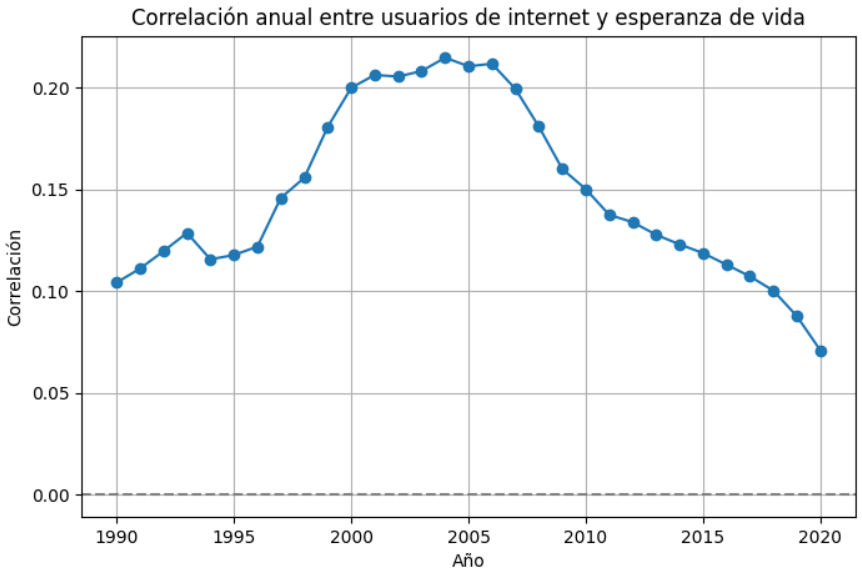

Figure 12: Correlación entre PC1 (Internet) y PC1 (Esperanza de vida)

Los resultados mostraron correlaciones positivas altas en la mayoría de los años, lo cual indica que los países con mayor crecimiento digital también tienden a presentar mejoras sostenidas en esperanza de vida.

Asimismo, ha habido una menor correspondencia entre el desarrollo digital y bienestar social en años recientes.

7 Relación entre los patrones temporales de Internet y Esperanza de Vida (PC1 vs PC1)

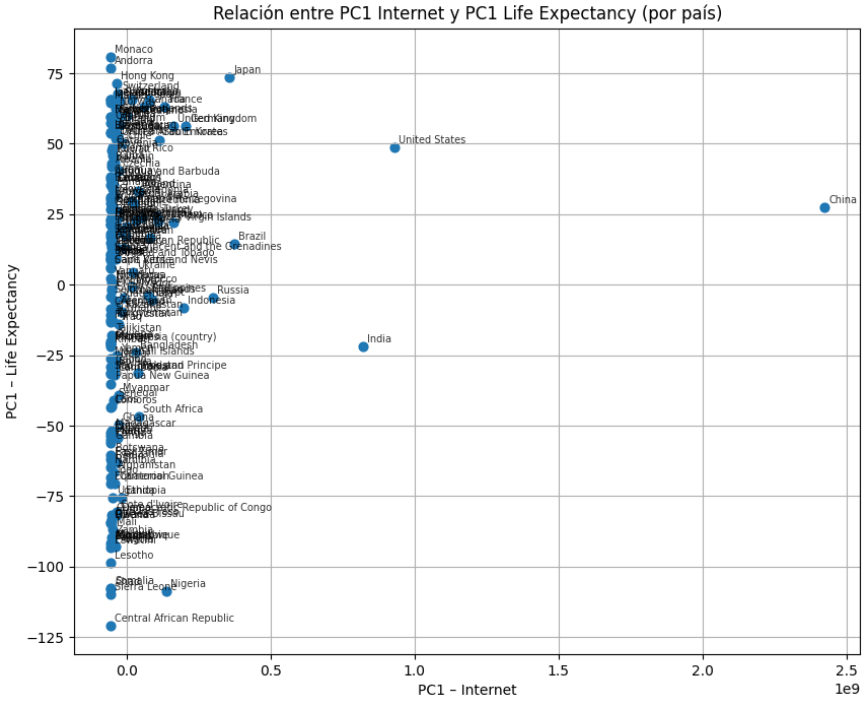

Figure 13: PC1-Internet vs PC1-Esperanza de vida

Los componentes principales resumen las tendencias temporales de cada indicador. En PC1–Internet destacan países muy poblados como China, India y Estados Unidos por su gran volumen de usuarios, mientras que otros países muestran trayectorias digitales más modestas. Aunque existe una relación positiva entre el avance digital y la esperanza de vida, la variación entre países indica que intervienen también factores económicos, sanitarios y demográficos.

8 Conclusiones

El análisis mediante PCA y k-means permitió identificar patrones claros en la evolución del acceso a internet a nivel global y su relación con indicadores de bienestar. El primer componente principal mostró que la mayoría de los países siguen trayectorias coherentes de crecimiento digital, mientras que la comparación entre PC1 de Internet y PC1 de Esperanza de Vida evidenció una relación positiva: los países con mayor avance en conectividad suelen mostrar también mejores condiciones de salud.

El agrupamiento reveló tres grupos principales que reflejan brechas digitales persistentes: países altamente digitalizados con alta esperanza de vida, países intermedios y países rezagados con bajas métricas en ambos indicadores. Además, casos de gran escala poblacional como China, India y Estados Unidos se comportan como atípicos debido al volumen absoluto de usuarios.

En conjunto, los resultados sugieren que la digitalización está estrechamente vinculada al desarrollo humano, y que las desigualdades en acceso a internet reflejan disparidades más amplias en bienestar social y capacidad institucional.

Limitaciones

- El análisis no considera causalidad, sólo asociación estadística.
- Existen países con datos históricos incompletos.
- El método k-means supone clusters esféricos, lo cual no necesariamente captura dinámicas socioeconómicas reales.

Supuestos

- Los datos han sido correctamente interpolados y estandarizados.
- La estructura PCA preserva adecuadamente la varianza relevante.

9 Referencias

- Our World in Data: Número de usuarios de internet. https://ourworldindata.org/grapher/number-of-internet-users
- Our World in Data: Esperanza de vida. https://ourworldindata.org/life-expectancy
- Hamida, S. (2024). *Data Reduction Using Principal Component Analysis: Theoretical Underpinnings and Practical Applications in Public Health.* Journal of Contemporary Medical Education, Department of Medicine and Health Sciences, Hawassa University, Hawassa, Ethiopia.