

Segmentación de países por uso de Internet y Esperanza de vida

INTRODUCCIÓN A CIENCIA DE DATOS

Guillermo Aguilar Martínez

Introducción y Objetivos

- ▶ El acceso a internet se ha convertido en un indicador clave del desarrollo moderno, mientras que la esperanza de vida refleja el bienestar y las condiciones sanitarias de una población. Analizar su evolución conjunta permite identificar patrones globales de desarrollo.
- ▶ Mediante técnicas como PCA y k-means se pueden extraer tendencias temporales dominantes y agrupar países según su nivel de conectividad digital y salud, revelando brechas claras entre naciones altamente desarrolladas, países en transición y regiones con rezago persistente.

Datos y Preparación

- ▶ Dataset: number-of-internet-users.csv y life-expectancy.csv (Our World in Data).
- ▶ Se eliminaron agregados como 'World', 'Europe', 'Asia'.
- ▶ Período analizado: 1990–2020.
- ▶ Matriz de datos: País × Años
- ▶ Estandarización de datos: $Z = (X - \mu) / \sigma$.

Manejo de Datos Faltantes

- ▶ Eliminar países con $>20\%$ de años faltantes
- ▶ Interpolación temporal lineal para huecos pequeños
- ▶ Forward-fill y backward-fill para completar extremos

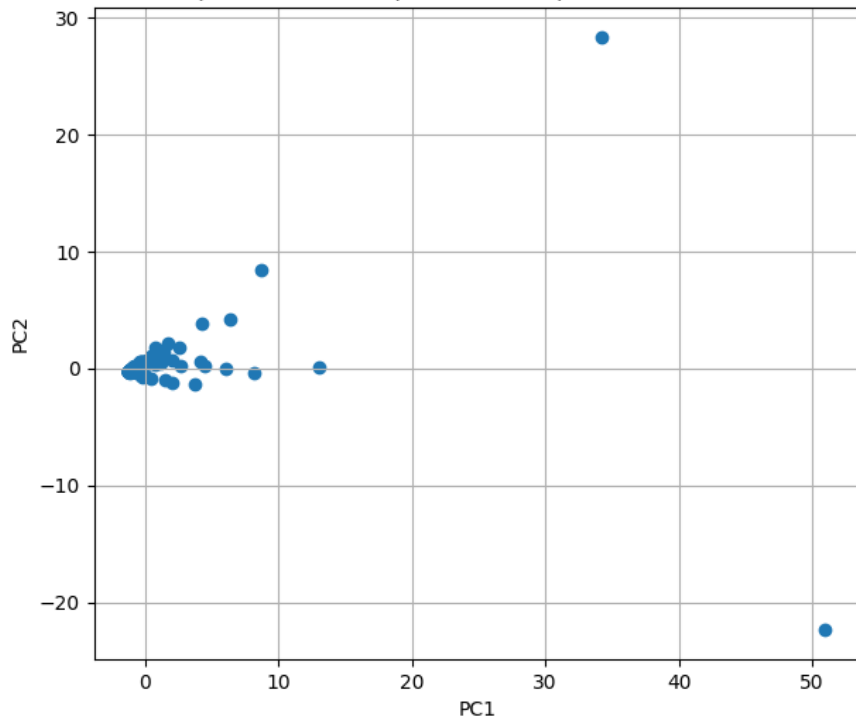
Metodología: PCA

- ▶ Reducción de 31 años a 2 componentes principales.
- ▶ PC1: Magnitud global del desarrollo.
- ▶ PC2: Diferencias en la dinámica temporal.

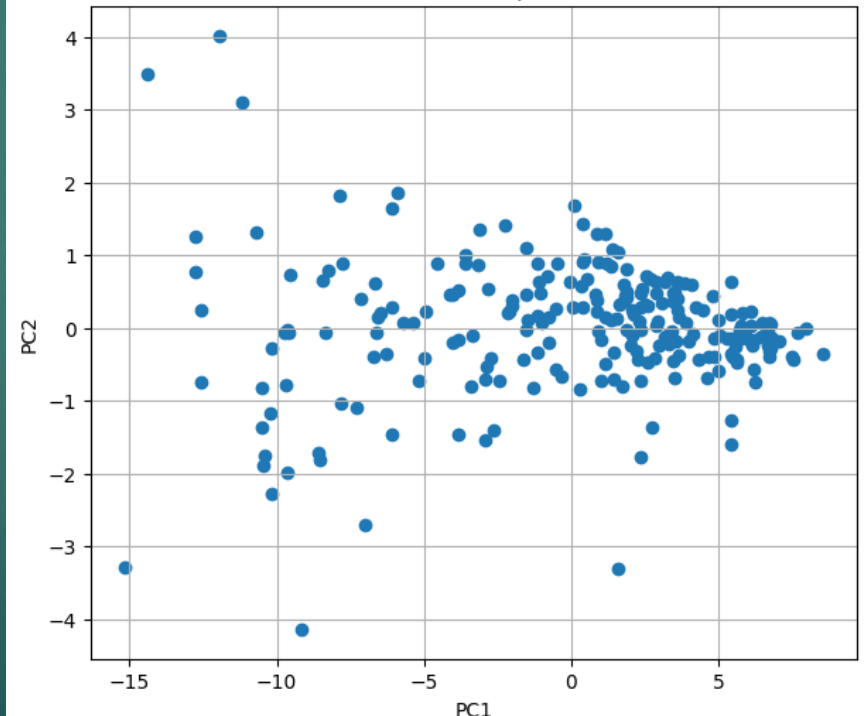
PCA sin Clusters

- ▶ Diferencias claras entre países con adopción alta vs lenta.
- ▶ La estructura visual motiva aplicar clustering.

Países en el espacio de los dos primeros componentes (usuarios de internet)



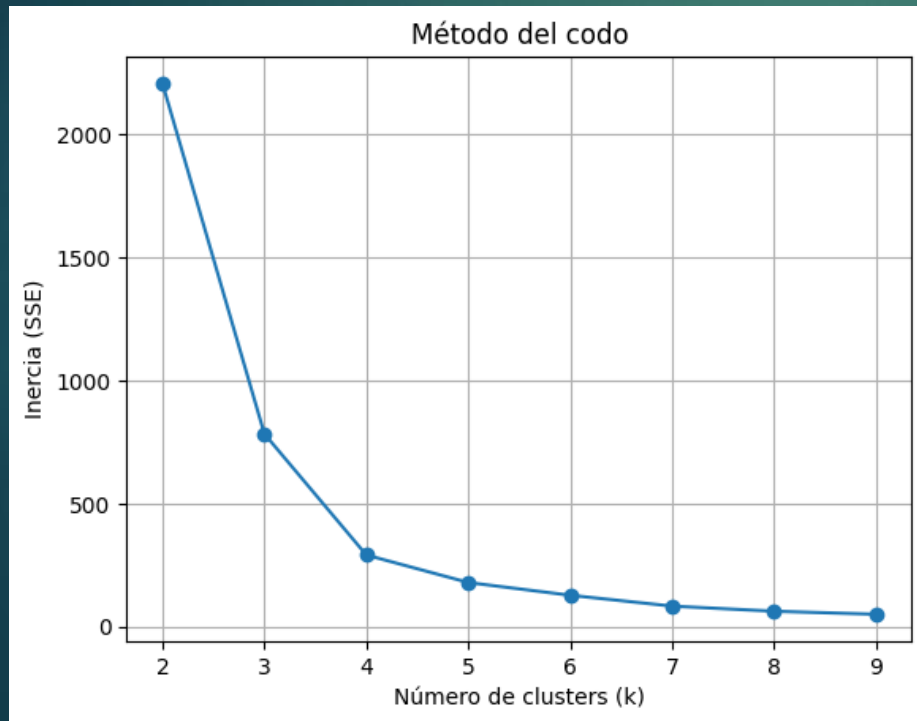
PCA (1990-2020) de esperanza de vida



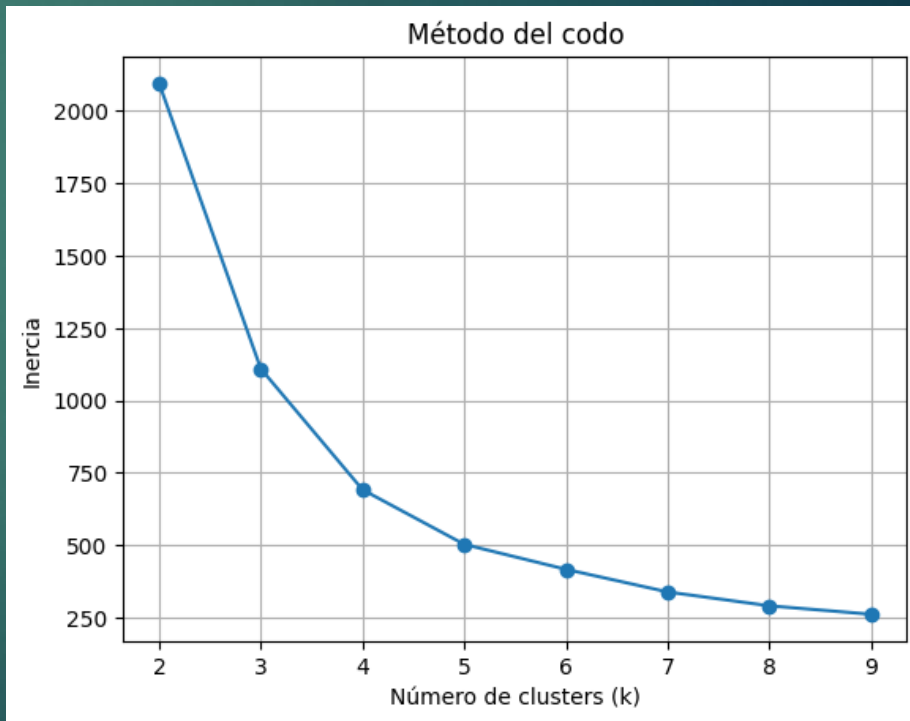
K- Means: Método del codo

- ▶ Se probaron valores $k = 2 \dots 9$.
- ▶ Después de $k=5$ la mejora marginal es menor.

Curva de inercia:



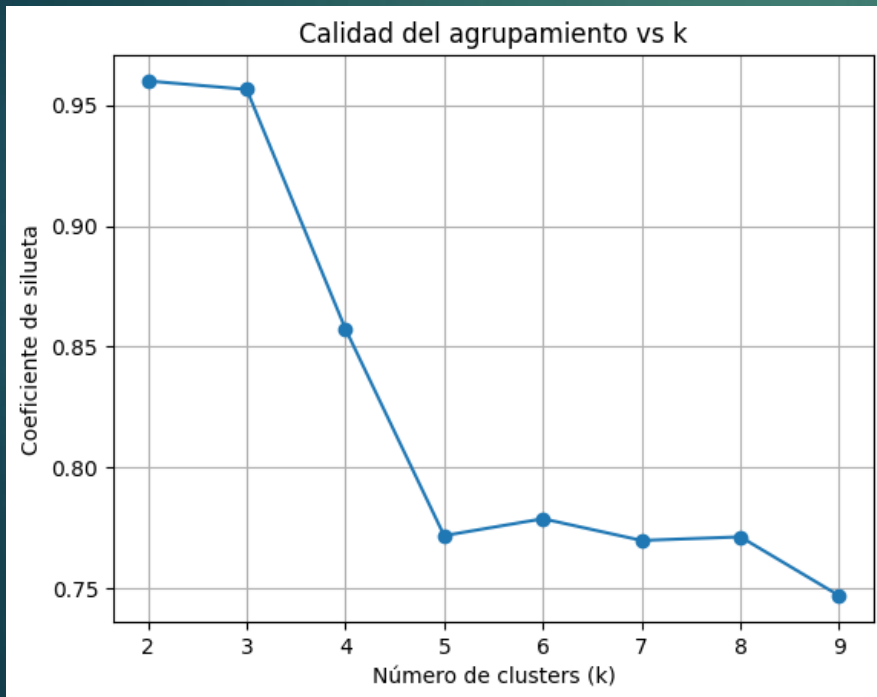
Internet



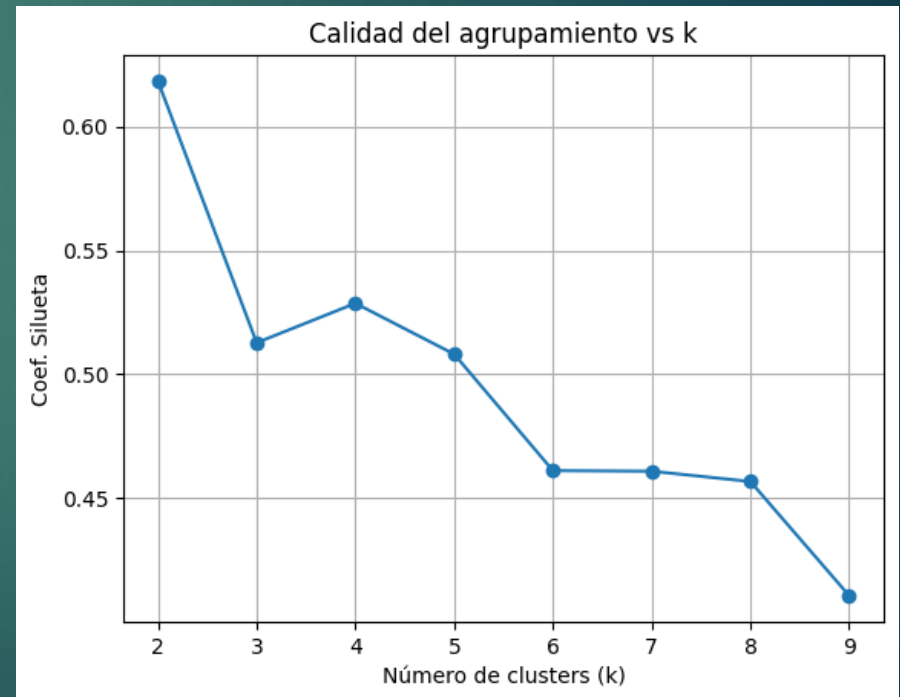
Esperanza de vida

K-Means: Método de la silueta

- ▶ Silueta máxima para k entre 3 y 5.
- ▶ Valores mayores reducen separación.
- ▶ Se eligió $k=5$.



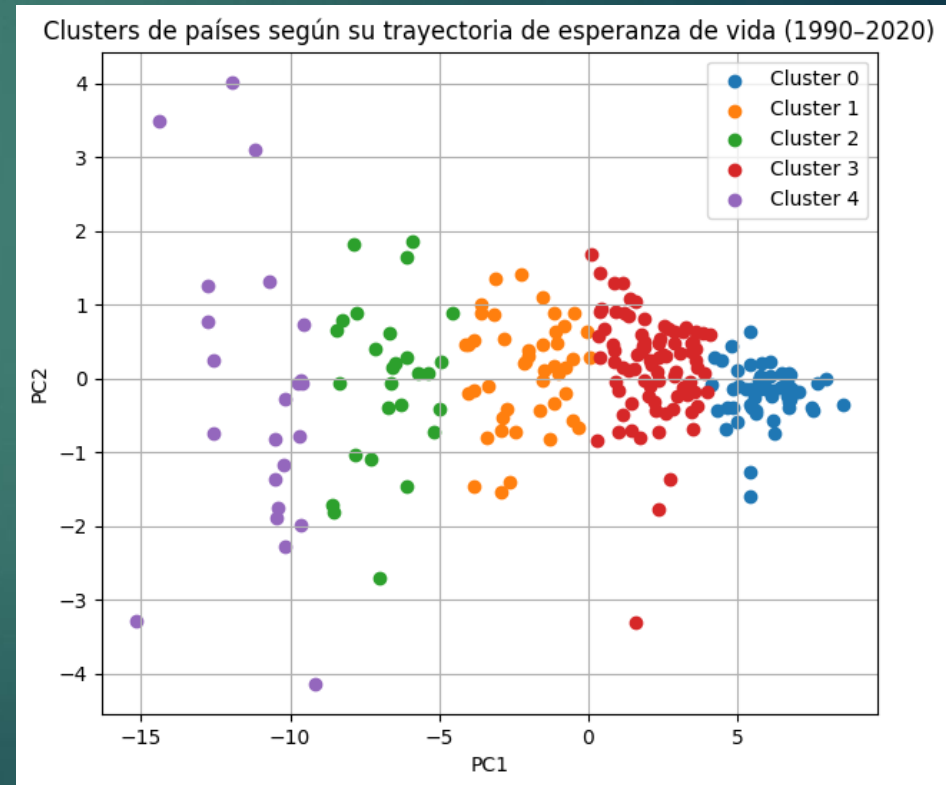
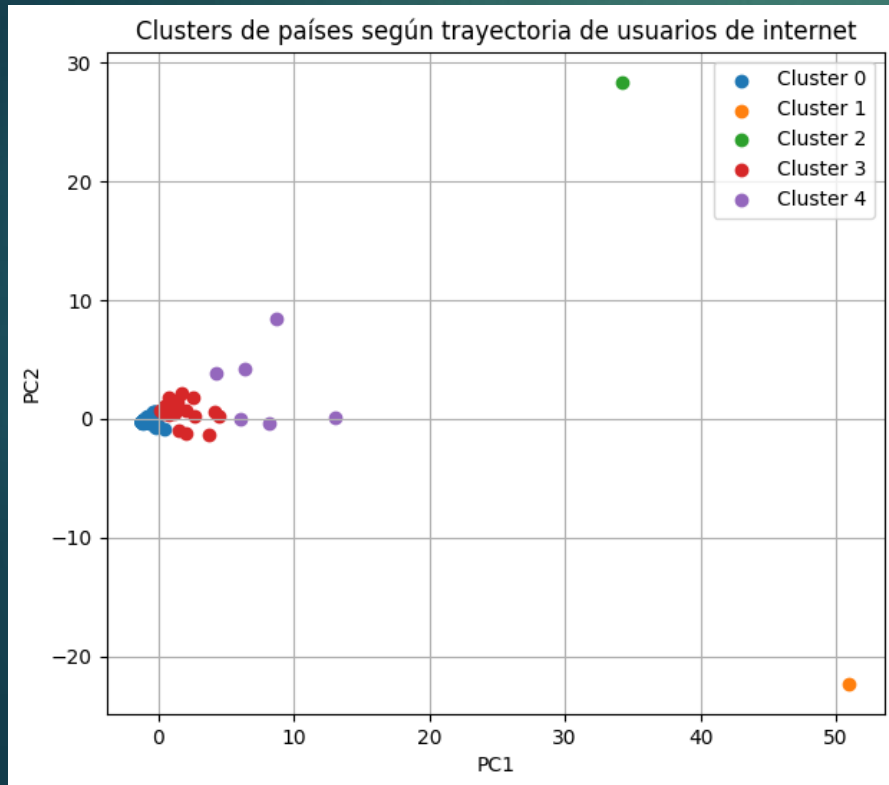
Internet



Esperanza de vida

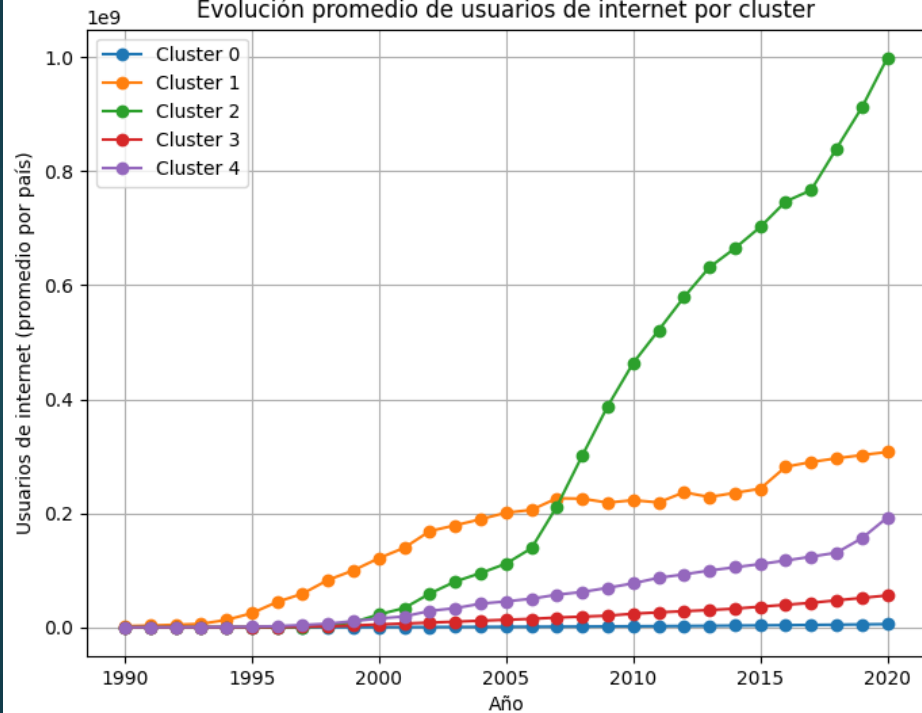
Clusters en PC1–PC2

- ▶ Visualización PC1–PC2 muestra patrones claros.
- ▶ Clusters reflejan brecha digital y de sanidad global.

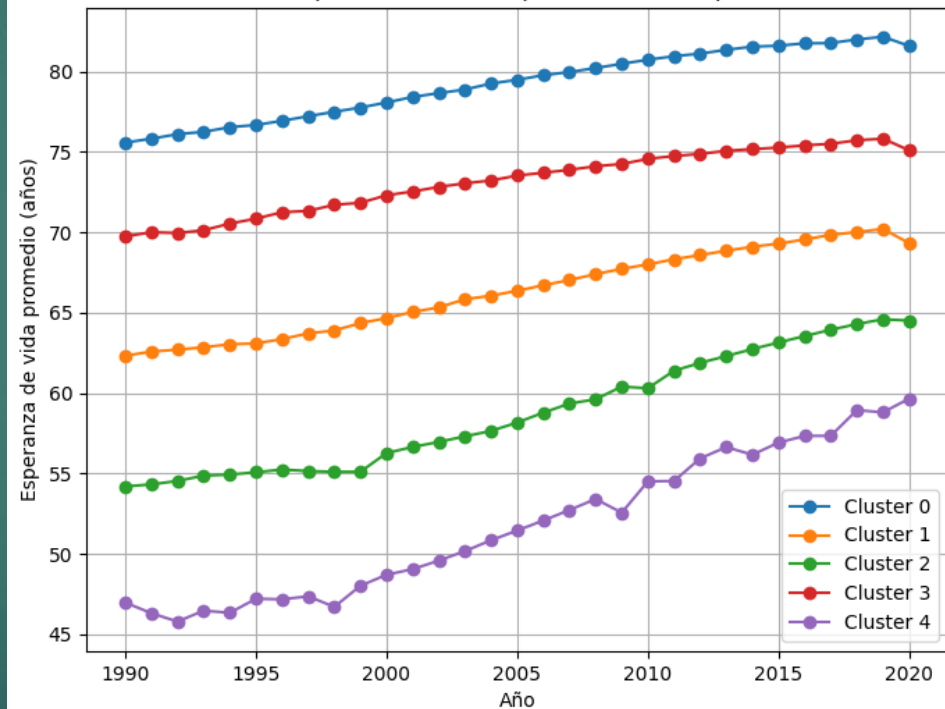


Evolución Promedio por Cluster

Evolución promedio de usuarios de internet por cluster



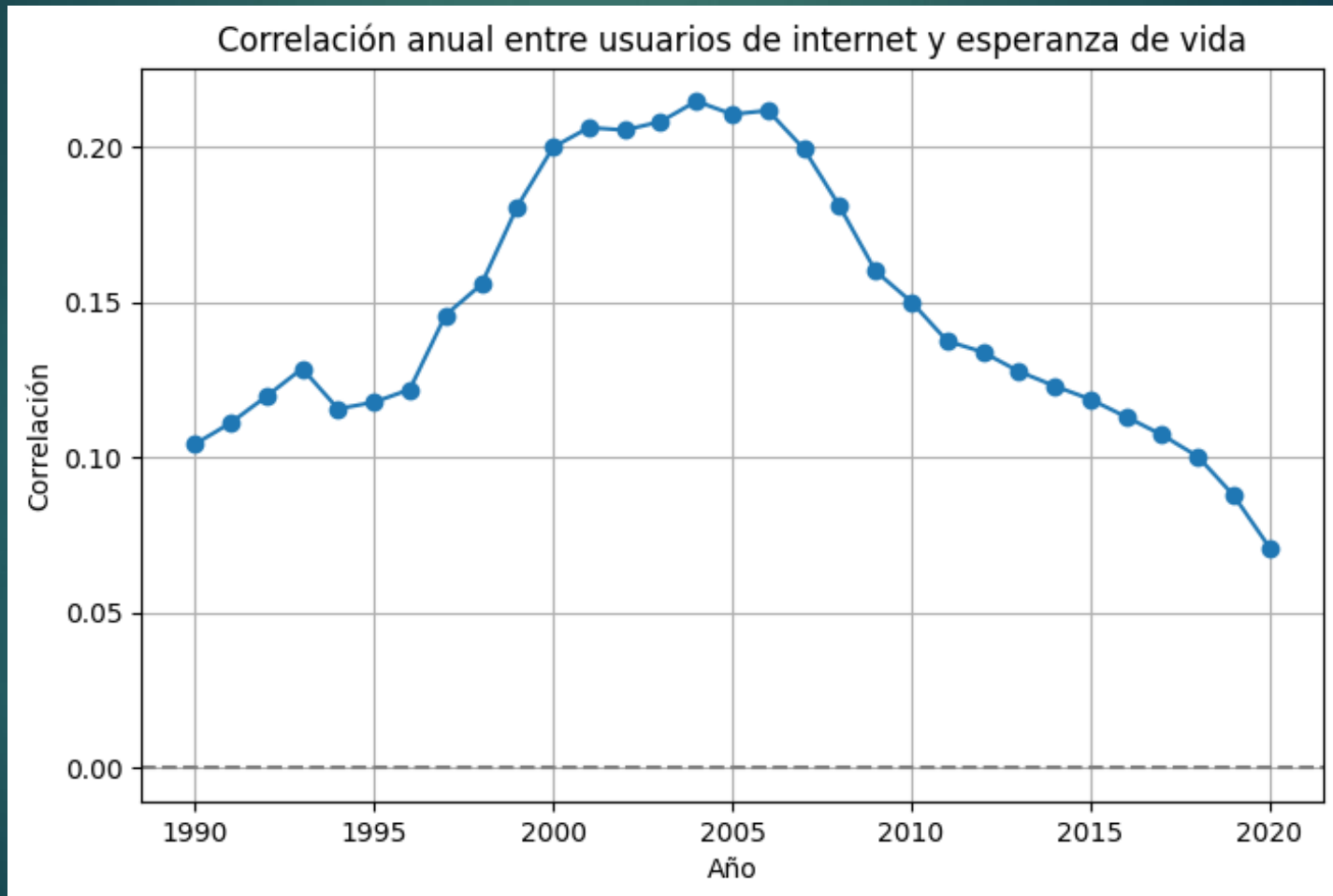
Evolución promedio de la esperanza de vida por cluster



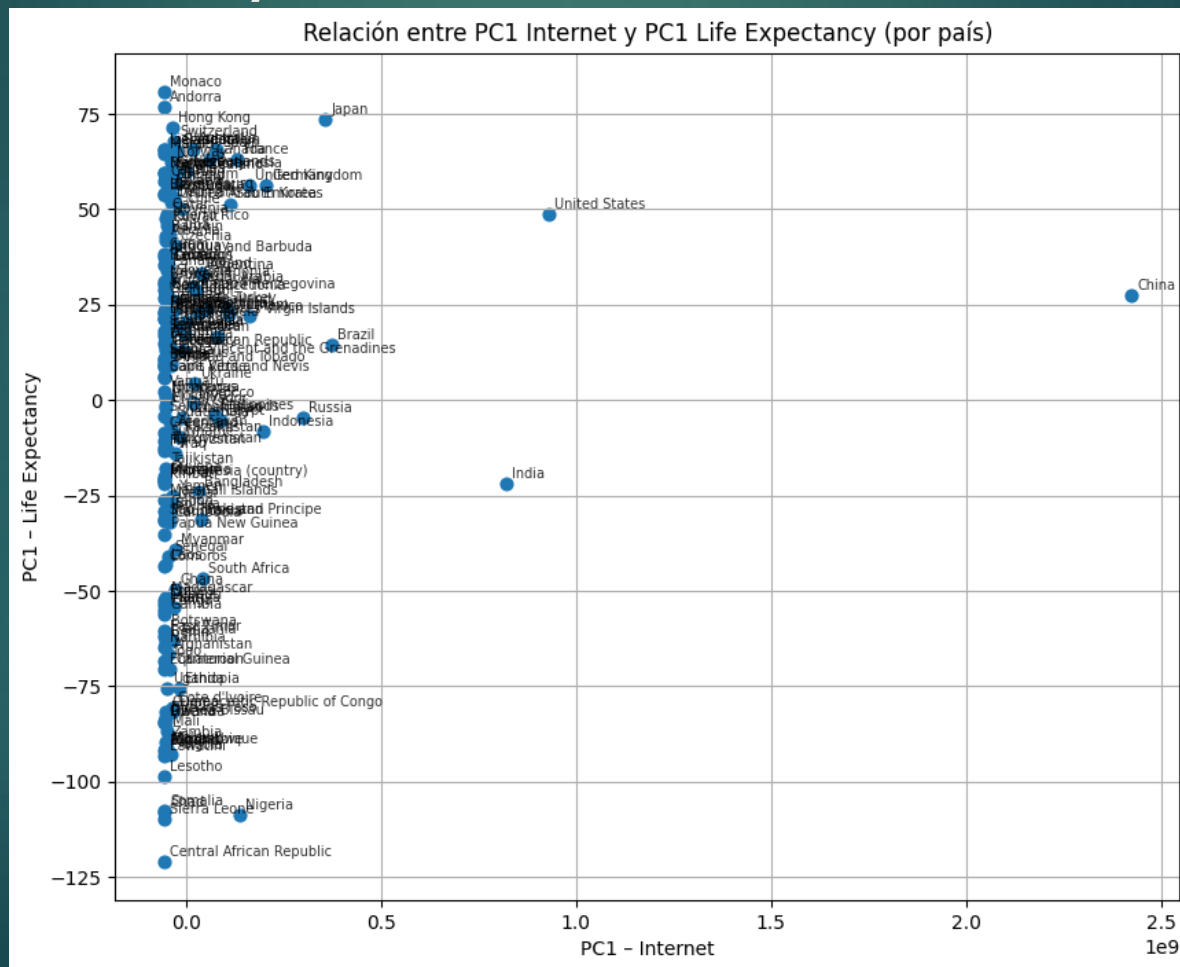
Interpretación

- ▶ PC1–PC2 visualiza estructura clara entre países.
- ▶ PC1 de Internet explica la mayor parte de la variabilidad ($\approx 73\%$).
- ▶ PC1 de Esperanza de vida también explica la mayor parte de la varianza ($\approx 95\%$).

Correlación



Patrones globales de digitalización y bienestar (PC1–PC1)



Comparando

- ▶ La correlación anual entre ambos indicadores aumenta hasta mediados de los 2000 y luego disminuye, mostrando que la relación existe pero depende también de factores económicos y sanitarios.
- ▶ Países con PC1 de Internet cercano a cero muestran trayectorias digitales más lentas y gran variabilidad en esperanza de vida.

Conclusiones

- ▶ La evolución del acceso a internet muestra patrones globales que se asocian con variaciones en la esperanza de vida, permitiendo identificar países con trayectorias digitales y sanitarias similares.
- ▶ El PCA muestra que los primeros dos componentes explican cerca del 98% de la variación total (Internet y Esperanza de vida).
- ▶ Los clusters revelan tres grupos: alta conectividad y alta esperanza de vida, transición intermedia, y rezago digital con menores niveles de salud.

Referencias

- ▶ <https://ourworldindata.org/grapher/number-of-internet-users?time=earliest..2021&country=MEX~CHN~IND~POL~MAR~MYS~UKR~BEL~AGO~CMR~CIV~ECU~ROU~NLD>
- ▶ <https://ourworldindata.org/life-expectancy>
- ▶ Hamida, S. (2024). Data Reduction Using Principal Component Analysis: Theoretical Underpinnings and Practical Applications in Public Health