

Aprendizaje No Supervisado: Exploración y Reducción de Dimensionalidad

HE2: Consultoría Económica con IA Responsable

Santiago Neira & Catalina Bernal

Universidad de los Andes
Departamento de Economía

Marzo 2026

Agenda de hoy

- 1 El Misterio: ¿Quiénes Son Nuestros Clientes? ¿Cómo caracterizar países / individuos?
- 2 Similitud y Distancia: El Lenguaje del No Supervisado
- 3 Análisis de Componentes Principales (PCA)
- 4 Diseño de Variables para Segmentación

Agenda de hoy

- 1 El Misterio: ¿Quiénes Son Nuestros Clientes? ¿Cómo caracterizar países / individuos?
- 2 Similitud y Distancia: El Lenguaje del No Supervisado
- 3 Análisis de Componentes Principales (PCA)
- 4 Diseño de Variables para Segmentación

Un Caso de Consultoría: El Encargo

Situación: Una cadena de retail con presencia nacional los contrata.

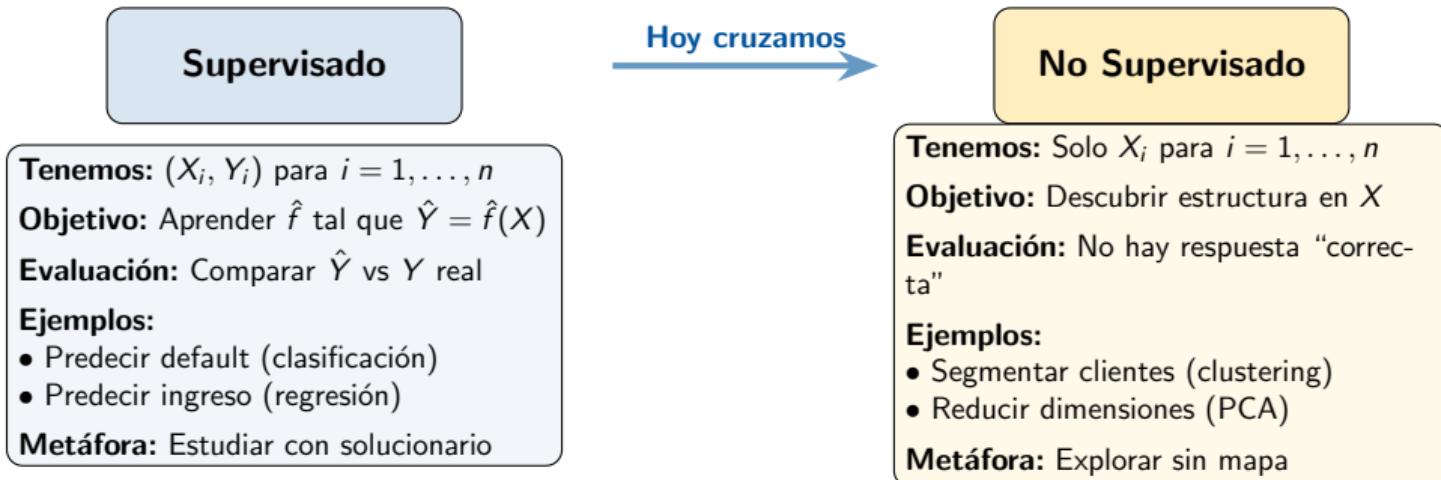
"Tenemos 200,000 clientes activos. Sabemos su edad, ingreso, frecuencia de compra, ticket promedio, canal preferido, categorías que compran... pero no sabemos quiénes son. Necesitamos una estrategia diferenciada, pero no sabemos en cuántos grupos dividirlos ni cuáles son."

¿Qué tiene de diferente este problema?

- No hay variable Y que predecir
- No hay “respuesta correcta” de antemano
- No podemos medir accuracy, precision, ni recall
- El objetivo es **descubrir estructura** en los datos

Bienvenidos al aprendizaje no supervisado.

El Cambio de Paradigma



¿Para Qué Sirve? Dos Grandes Familias

Clustering (Agrupamiento)

Encontrar grupos “naturales” en los datos.

- Segmentación de clientes
- Tipologías de municipios
- Perfiles de riesgo crediticio
- Grupos de países por desarrollo

Próxima clase

Reducción de Dimensionalidad

Resumir muchas variables en pocas, preservando la información esencial.

- Visualizar datos de 50 variables en 2D
- Construir índices compuestos
- Pre-procesamiento antes de clustering
- Eliminar ruido y colinealidad

Clase de hoy

La secuencia del consultor

Antes de agrupar, necesitamos **entender la geometría** de los datos. Reducir dimensiones es el primer paso de exploración.

El Problema Real: Demasiadas Variables

Volvamos al caso de retail. Tenemos 15 variables por cliente:

Demográficas

- Edad
- Género
- Estrato
- Ciudad

Comportamiento de compra

- Frecuencia mensual
- Ticket promedio
- Categorías distintas
- Canal (online/físico)
- Hora pico de compra

Financieras

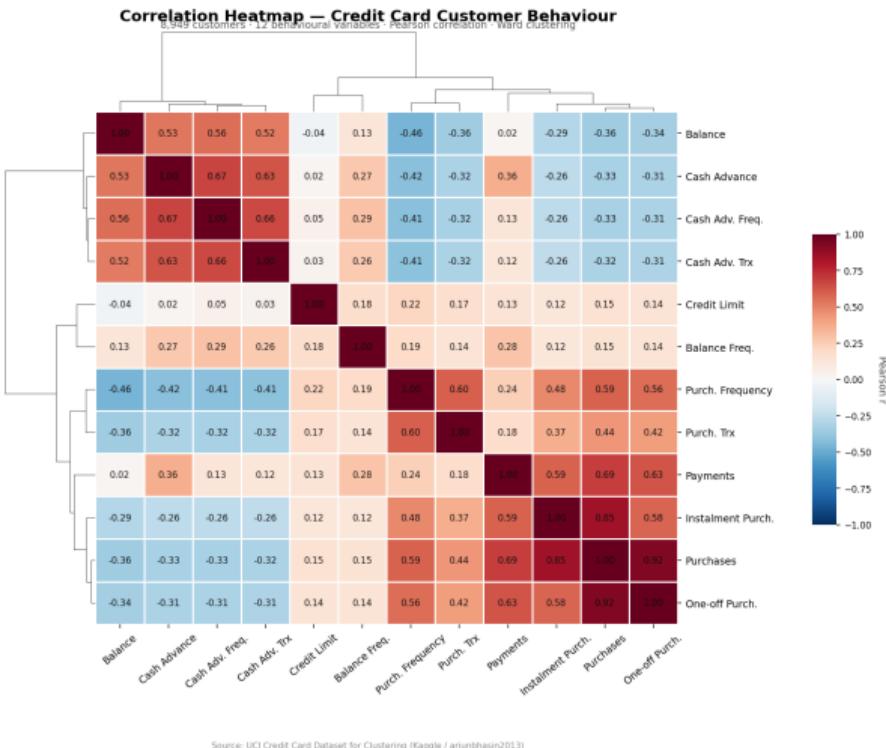
- Gasto total anual
- Uso de crédito
- Devoluciones (%)
- Descuentos usados
- Meses como cliente
- Saldo en cuenta

Pregunta: ¿Cómo visualizar 200,000 puntos en 15 dimensiones?

No podemos. Nuestro cerebro opera en 2D (máximo 3D). Necesitamos una herramienta que **comprima** 15 dimensiones en 2, sin perder lo esencial.

Esa herramienta es PCA.

¿Cómo Se Ve un Dataset de Alta Dimensión?



Observación clave: Muchas variables están correlacionadas entre sí. Existe **redundancia informativa**.

- Bloques de alta correlación \Rightarrow variables redundantes
- Si ingreso y gasto tienen $\rho = 0,92$, ¿necesitamos ambas?
- PCA encuentra dimensiones **no redundantes**

Agenda de hoy

- 1 El Misterio: ¿Quiénes Son Nuestros Clientes? ¿Cómo caracterizar países / individuos?
- 2 Similitud y Distancia: El Lenguaje del No Supervisado
- 3 Análisis de Componentes Principales (PCA)
- 4 Diseño de Variables para Segmentación

¿Qué Significa “Parecido”?

Intuición económica: Dos clientes son “parecidos” si compran de manera similar, tienen ingresos similares, frecuencias similares...

Pero necesitamos **formalizar** esta intuición. Toda la maquinaria de aprendizaje no supervisado descansa sobre una pregunta:

¿Qué tan lejos están dos observaciones en el espacio de características?

Cada observación es un punto en \mathbb{R}^p :

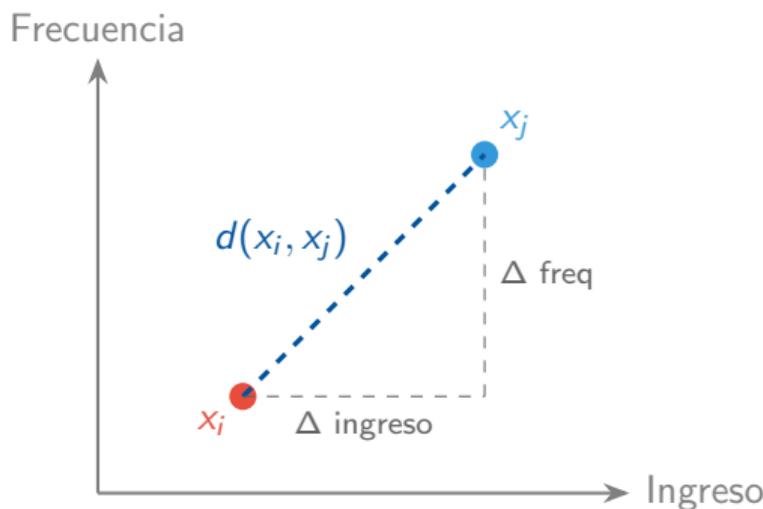
$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathbb{R}^p \quad (1)$$

Un cliente con ingreso \$3M, edad 35 y frecuencia 4 es el punto $(3, 35, 4)$ en \mathbb{R}^3 .

Distancia Euclídea: La Más Natural

Distancia euclídea entre dos puntos x_i y x_j :

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2} = \|x_i - x_j\|_2 \quad (2)$$

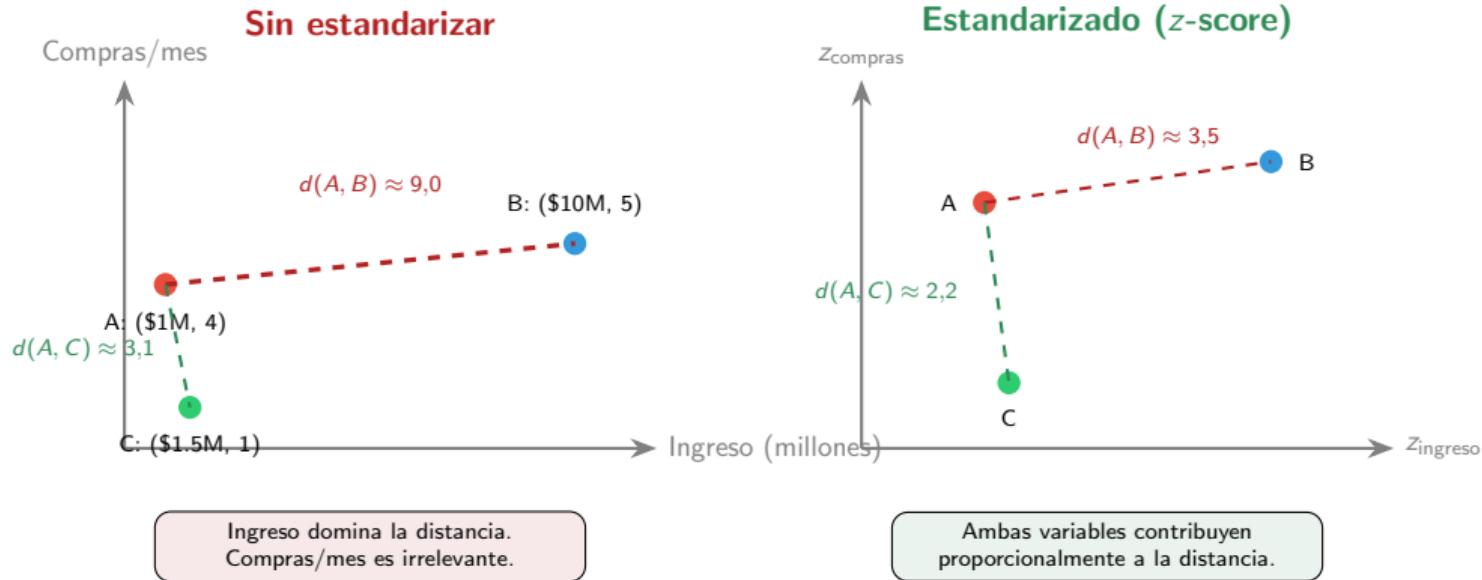


Propiedades:

- $d(x_i, x_j) \geq 0$ (no negativa)
- $d(x_i, x_j) = 0 \iff x_i = x_j$
- $d(x_i, x_j) = d(x_j, x_i)$ (simétrica)
- Desigualdad triangular

Es la extensión del teorema de Pitágoras a p dimensiones.

Cuidado: Las Escalas Importan



Regla de oro

Siempre estandarizar antes de calcular distancias, PCA o clustering. Estandarizar:

$$z_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}.$$

Otras Distancias: ¿Cuándo Usar Cuál?

Distancia	Fórmula	Uso típico
Euclídea (L_2)	$\sqrt{\sum(x_{ik} - x_{jk})^2}$	Default, variables continuas
Manhattan (L_1)	$\sum x_{ik} - x_{jk} $	Robusta a outliers
Coseno	$1 - \frac{x_i \cdot x_j}{\ x_i\ \ x_j\ }$	Texto, dirección
Mahalanobis	$\sqrt{(x_i - x_j)^\top \Sigma^{-1} (x_i - x_j)}$	Correlación entre variables

Para el consultor: La distancia euclídea con datos estandarizados es el punto de partida. Solo cambie si tiene razones de dominio (e.g., datos de texto → coseno; variables correlacionadas → Mahalanobis o PCA primero) **¿Cuándo usar cual?**.

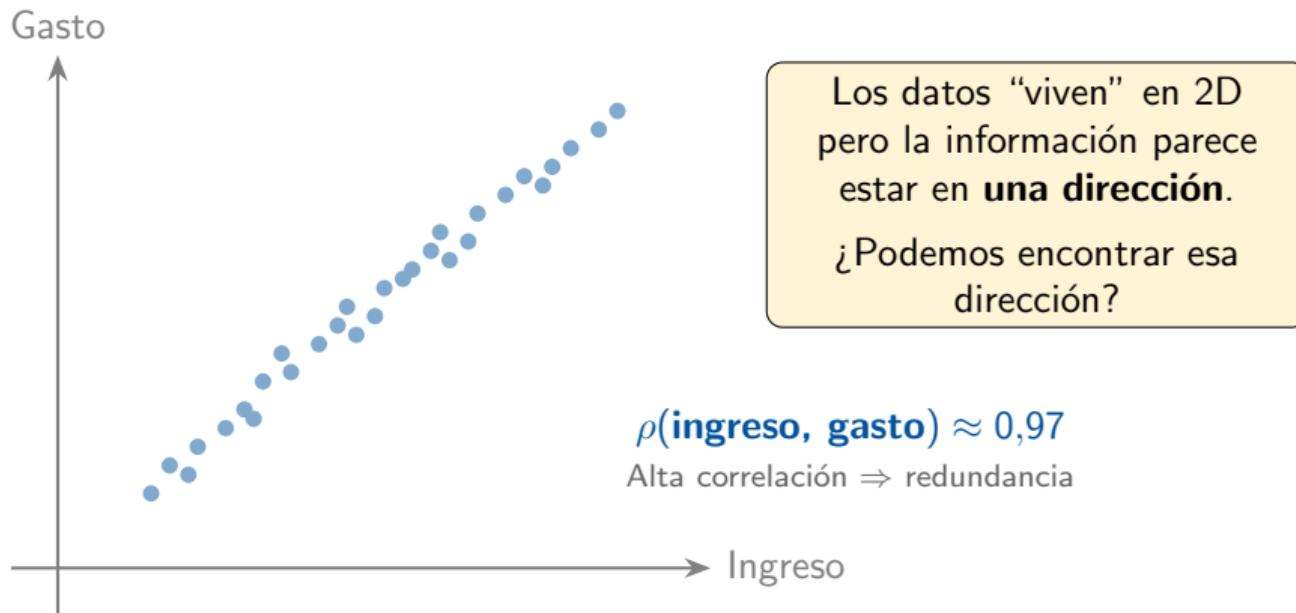
Agenda de hoy

- 1 El Misterio: ¿Quiénes Son Nuestros Clientes? ¿Cómo caracterizar países / individuos?
- 2 Similitud y Distancia: El Lenguaje del No Supervisado
- 3 Análisis de Componentes Principales (PCA)
- 4 Diseño de Variables para Segmentación

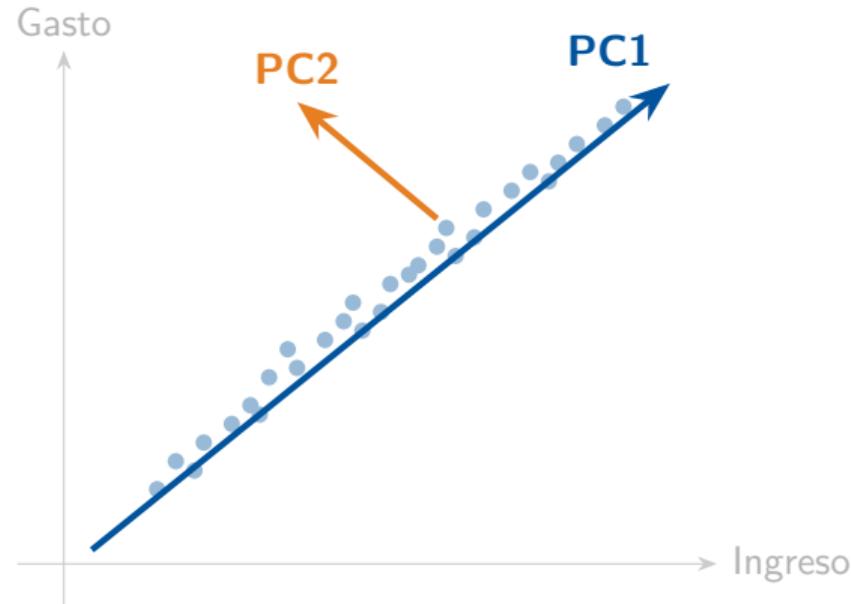
PCA: La Intuición en 2D

Imaginemos que solo tenemos dos variables: ingreso y gasto.

Los datos forman una nube alargada — las variables están correlacionadas.



PCA en 2D: Encontrar la Dirección de Máxima Varianza



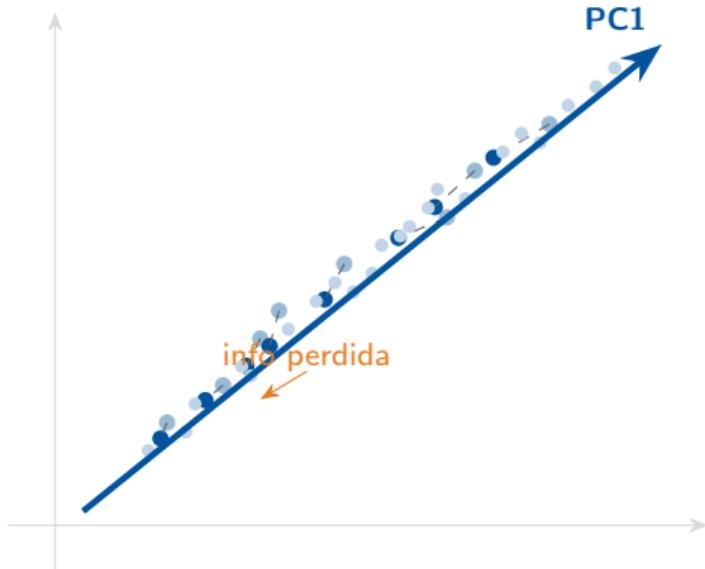
PC1: Dirección de máxima varianza.

Captura la señal principal:
“nivel adquisitivo general”

PC2: Perpendicular a PC1.
Captura la varianza **residual**:
“desviación del patrón”

¿Quién gasta más/menos
de lo esperado dado su in-
greso?

PCA en 2D: La Proyección



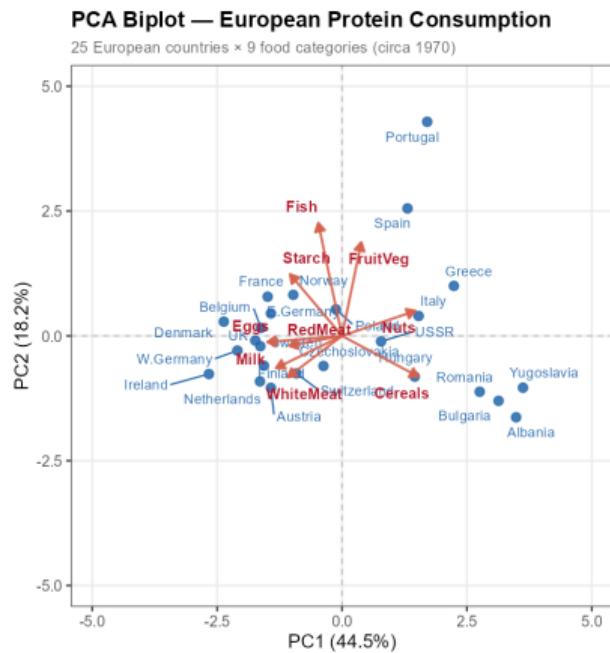
Cada punto se **proyecta** sobre PC1.

La proyección es la “sombra” del dato sobre la dirección principal.

Si PC1 captura el 95 % de la varianza,
la proyección pierde muy poco.

2 variables → 1 componente
(reducción de 2D a 1D)

Ejemplo Real: PCA sobre Datos de Consumo



Dataset clásico: Consumo de proteínas (g/persona/día) por tipo de alimento en 25 países europeos.

Variables: Carne roja, carne blanca, huevos, leche, pescado, cereales, almidones, nueces, frutas/vegetales.

Lo que revela PCA:

- PC1: Dieta dominada en cereales (este) vs dieta dominante en carnes (oeste)
- PC2: Dieta basada en pescado vs dieta basada en almidón / huevos
- Países escandinavos se agrupan, los mediterráneos también

Formalización: ¿Qué Hace PCA Exactamente?

Dato de entrada: Matriz X de $n \times p$ (centrada, idealmente estandarizada).

Paso 1: Calcular la **matriz de covarianza** (o correlación):

$$\Sigma = \frac{1}{n-1} X^\top X \in \mathbb{R}^{p \times p} \quad (3)$$

Paso 2: Encontrar los **eigenvectores** y **eigenvalores** de Σ :

$$\Sigma v_k = \lambda_k v_k, \quad k = 1, \dots, p \quad (4)$$

donde $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$.

Paso 3: Las **componentes principales** son las proyecciones:

$$Z_k = Xv_k \quad (\text{la } k\text{-ésima componente principal}) \quad (5)$$

Interpretación de los eigenvalores

λ_k = varianza capturada por la k -ésima componente.

Proporción de varianza explicada: $\frac{\lambda_k}{\sum_{j=1}^p \lambda_j}$

Los Loadings: ¿Qué Significa Cada Componente?

Cada componente principal es una **combinación lineal** de las variables originales:

$$Z_1 = v_{11}X_1 + v_{12}X_2 + \cdots + v_{1p}X_p \quad (6)$$

Los coeficientes v_{1j} se llaman **loadings**. Nos dicen *cuánto contribuye* cada variable original a la componente.

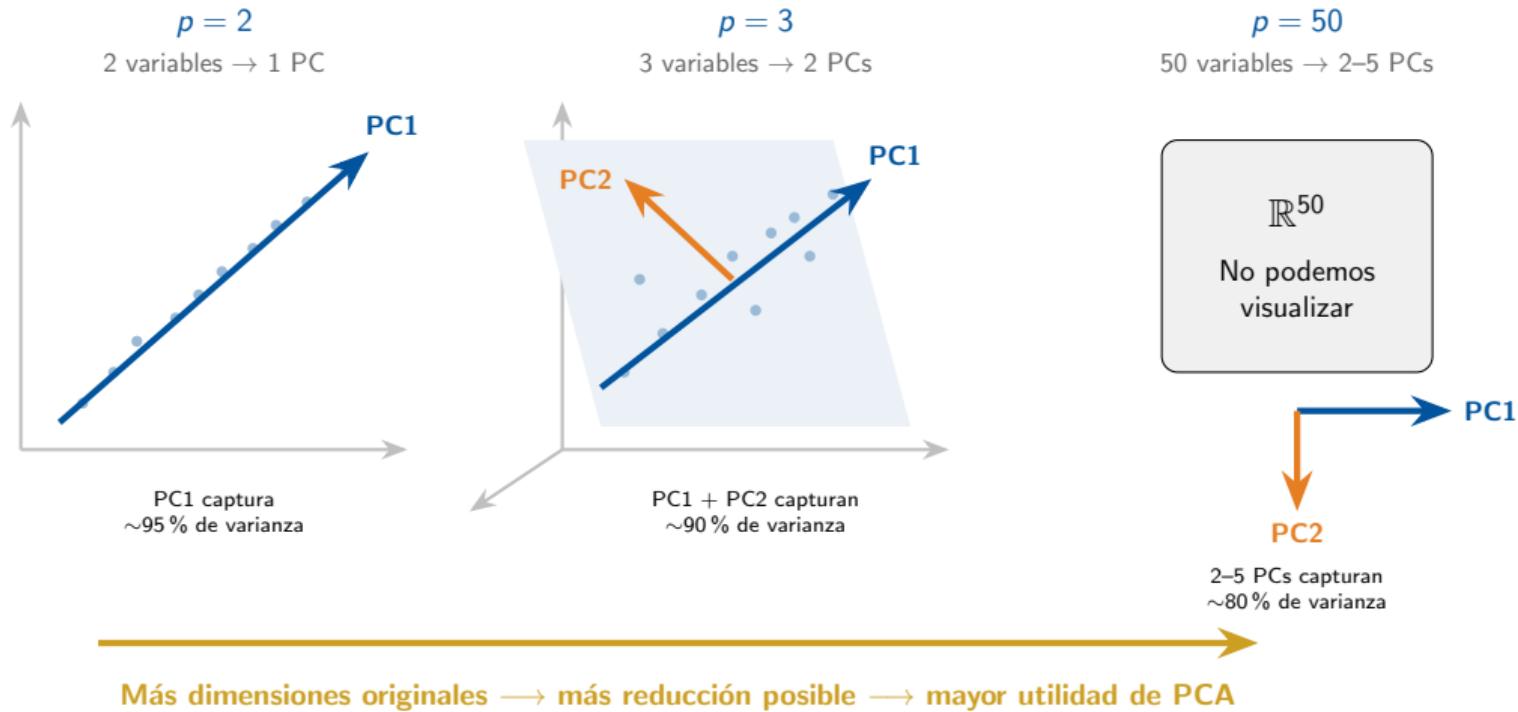
Ejemplo con datos de retail:

Variable	Loading PC1	Loading PC2
Ingreso	0.45	-0,10
Gasto total	0.48	-0,15
Ticket promedio	0.42	0.35
Frecuencia de compra	0.38	-0,55
Uso de descuentos	-0,30	0.60
Devoluciones	-0,15	0.40

Interpretación para el cliente:

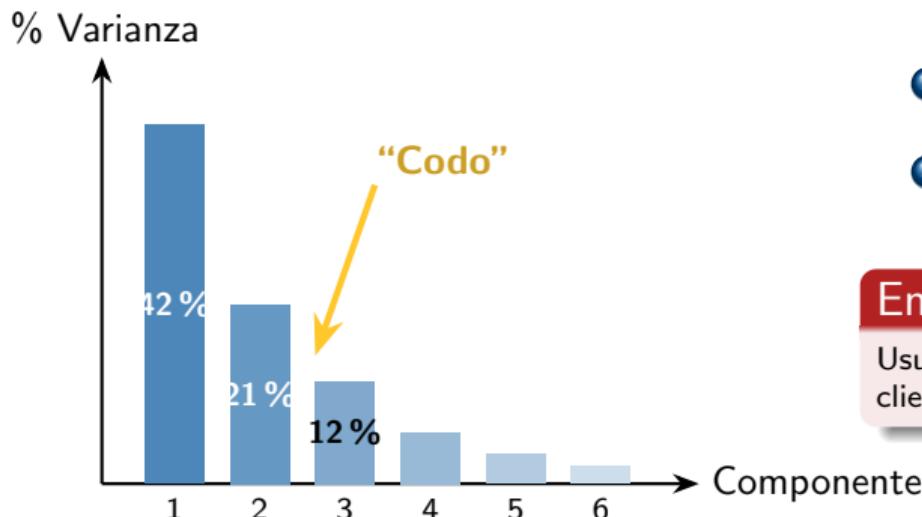
- PC1 ≈ “Poder adquisitivo general” (ingreso, gasto, frecuencia altos)
- PC2 ≈ “Cazador de ofertas” (descuentos altos, devoluciones, frecuencia baja)

PCA: De 2D a p Dimensiones



¿Cuántas Componentes Retener?

Scree Plot (Gráfico de Codo)



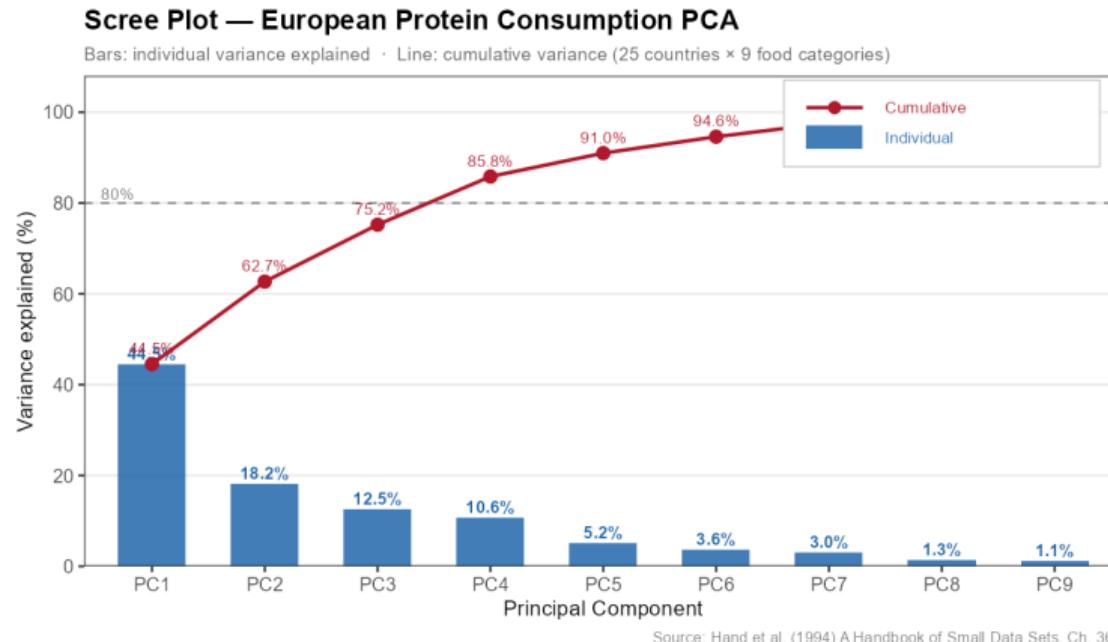
Criterios de selección:

- ① **Regla del codo:** Retener componentes hasta que la curva se “aplana”
- ② **Varianza acumulada:** Retener hasta alcanzar 80–90 % de varianza total
- ③ **Criterio de negocio:** ¿Con cuántos ejes el cliente puede tomar decisiones?

En consultoría

Usualmente 2–3 componentes bastan para visualizar. El cliente necesita un gráfico que quepa en una diapositiva.

Scree Plot en la Práctica



Lectura del scree plot para el cliente:

- “Con solo 3 dimensiones capturamos el 75 % de la información de las 15 variables originales”
- “Aregar una 4ta dimensión solo aporta un 10 % adicional — no vale la complejidad”
- “Esto nos permite hacer un mapa de segmentación en un gráfico de 2 ejes”

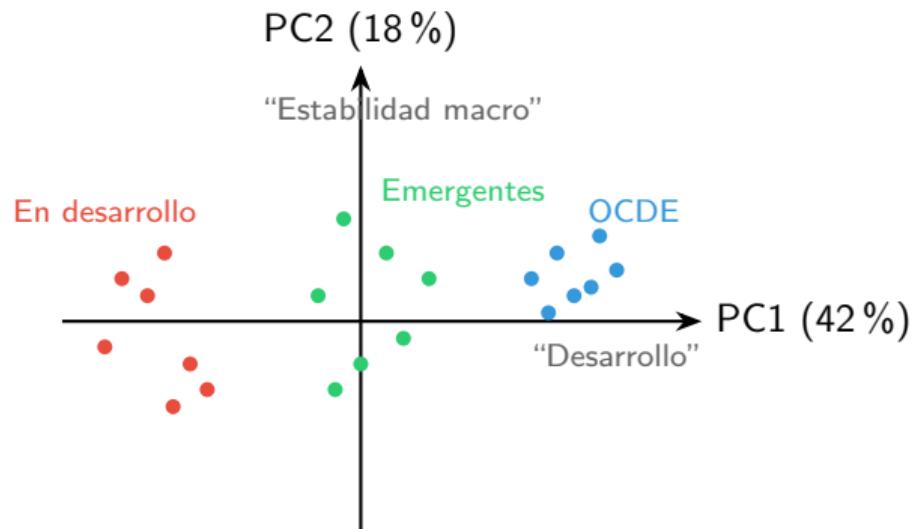
Ejemplo Económico: Indicadores de Desarrollo de Países

Contexto: 150 países, 12 indicadores del Banco Mundial.

Variables originales:

- PIB per cápita
- Esperanza de vida
- Tasa de alfabetización
- Mortalidad infantil
- Acceso a electricidad
- Gasto en salud (%PIB)
- Desempleo
- Inflación
- Gini
- Emisiones CO₂ pc
- Inversión (%PIB)
- Deuda pública (%PIB)

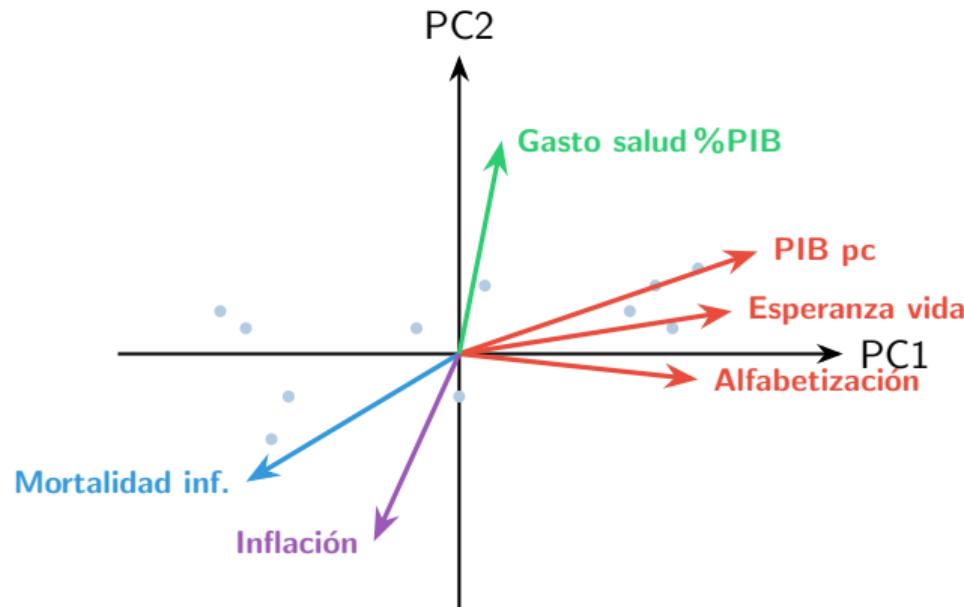
Después de PCA:



12 variables → 2 componentes. Y ya se ven los grupos "naturales". PCA como paso previo a clustering.

Biplot: Visualizar Datos y Variables Simultáneamente

El **biplot** superpone las observaciones (puntos) y las variables (flechas) en el mismo espacio de PCA.



Lectura del biplot:

- Flechas largas = variables bien representadas
- Flechas cercanas = variables correlacionadas
- Flechas opuestas = correlación negativa
- Puntos cercanos a una flecha = valores altos en esa variable

PCA: Propiedades Importantes

- 1 Las componentes son ortogonales:

$$\text{Cor}(Z_i, Z_j) = 0 \quad \text{para } i \neq j \tag{7}$$

Esto elimina la multicolinealidad — útil si PCA es input de otro modelo.

- 2 La varianza total se conserva:

$$\sum_{k=1}^p \lambda_k = \sum_{k=1}^p \text{Var}(X_k) = \text{traza}(\Sigma) \tag{8}$$

PCA redistribuye la varianza, no la inventa.

- 3 PCA es una rotación: Los nuevos ejes son simplemente los ejes originales rotados para alinearse con la dirección de máxima varianza.
- 4 Es lineal: PCA solo captura relaciones lineales. Si la estructura de los datos es no lineal, PCA puede no ser suficiente — existen métodos más avanzados (lo veremos brevemente).

PCA: Ventajas y Limitaciones

Ventajas:

- Determinístico (mismo resultado siempre)
- Preserva estructura global
- Componentes interpretables (loadings)
- Rápido ($O(np^2)$)
- Fundamentación matemática clara (eigendecomposición)

Limitaciones:

- Solo captura relaciones **lineales**
- Sensible a outliers (varianza)
- Asume que varianza = información
- Difícil interpretar con p muy grande
- No preserva distancias locales (vecindarios)

En la práctica del consultor

PCA es la herramienta principal de reducción de dimensionalidad en consultoría. Si los clusters no se ven claros en el biplot, antes de probar métodos más avanzados, revise la selección y transformación de variables.

Nota: Existen Métodos Más Avanzados

PCA es lineal. Si la estructura de los datos es no lineal, existen alternativas:

- **t-SNE, UMAP:** Métodos no lineales para *visualización* en 2D. Útiles para explorar datos de muy alta dimensión, pero los ejes no son interpretables.
- **Kernel PCA:** PCA en un espacio transformado (misma idea del kernel trick de SVM).
- **Autoencoders:** Redes neuronales que comprimen y reconstruyen datos.

Para este curso

PCA es suficiente para la gran mayoría de aplicaciones en consultoría económica. Estos métodos son herramientas complementarias de visualización — si tienen curiosidad, revisen las referencias al final.

Agenda de hoy

- 1 El Misterio: ¿Quiénes Son Nuestros Clientes? ¿Cómo caracterizar países / individuos?
- 2 Similitud y Distancia: El Lenguaje del No Supervisado
- 3 Análisis de Componentes Principales (PCA)
- 4 Diseño de Variables para Segmentación

Antes de Reducir o Agrupar: ¿Qué Variables Usar?

El paso más importante (y el más ignorado): decidir qué variables entran al análisis.

Principio clave

Las variables deben reflejar la **dimensión en la que queremos segmentar**. No todo lo que tenemos es útil.

Ejemplo — segmentación de clientes de retail:

Segmentación por comportamiento:

- Frecuencia de compra
- Ticket promedio
- Categorías compradas
- Canal preferido
- Recencia (días desde última compra)

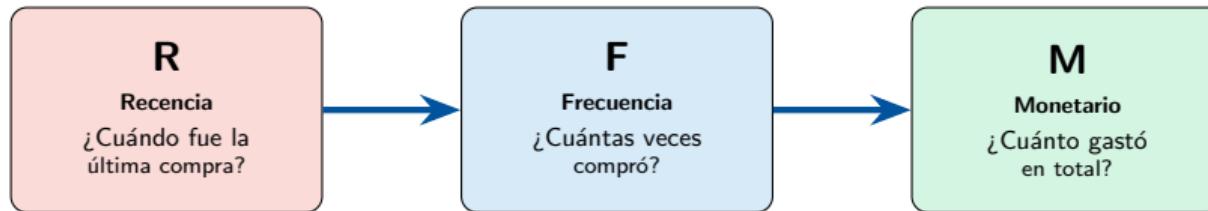
Segmentación demográfica:

- Edad
- Estrato
- Ubicación geográfica
- Género
- Estado civil

Mezclar ambas puede producir clusters difíciles de interpretar y accionables.

RFM: Un Framework Clásico de Segmentación

RFM (Recency, Frequency, Monetary): Un marco simple y poderoso para segmentar clientes por comportamiento transaccional.



Con solo 3 variables bien construidas + PCA o clustering, se obtienen segmentos accionables: "Champions", "At Risk", "Lost", "Promising" ...

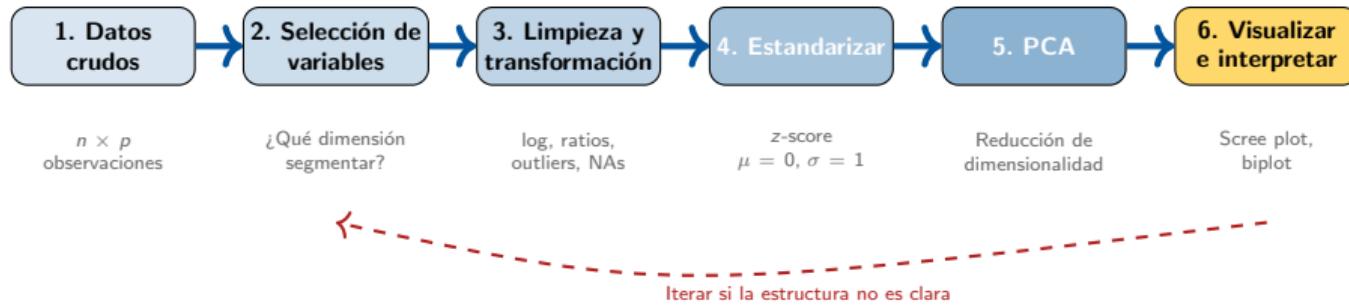
Buenas Prácticas en Selección de Variables

- 1 Alinear con el objetivo de negocio:** ¿Para qué segmentamos? La respuesta define las variables.
- 2 Evitar variables redundantes:** Si ingreso y gasto tienen $\rho > 0,95$, incluir ambas infla la importancia de esa dimensión. PCA ayuda, pero mejor curar antes.
- 3 Cuidado con variables de escala:** Una empresa grande tendrá valores absolutos mayores en todo. Usar ratios o per cápita cuando corresponda.
- 4 No incluir identificadores:** Código de cliente, NIT, o fecha de registro no son features de segmentación (aunque parezca obvio, pasa más de lo que creen).
- 5 Transformar distribuciones sesgadas:** $\log(\text{ingreso})$ suele funcionar mejor que ingreso crudo. Los outliers distorsionan distancias y PCA.

Regla del consultor

Dedica el 60 % del tiempo a entender los datos y construir variables. El 40 % restante es algoritmo + presentación.

Pipeline Completo: De los Datos Crudos a la Exploración



La próxima clase

Con la exploración hecha (PCA), pasamos al paso siguiente: **algoritmos de clustering** (k-means, jerárquico, DBSCAN) para formalizar los grupos que empezamos a intuir hoy.

Resumen: Ideas Clave

- ① **Aprendizaje no supervisado:** No hay Y . El objetivo es descubrir estructura. Explorar antes de agrupar.
- ② **Distancias:** Todo empieza con una noción de similitud. Estandarizar siempre.
- ③ **PCA:** Encuentra las direcciones de máxima varianza. Reduce dimensionalidad preservando estructura global. Loadings dan interpretabilidad. Útil para construir índices y visualizar datos de alta dimensión.
- ④ **Diseño de variables:** La decisión más importante. Alinear con el objetivo de negocio. Ratios, transformaciones y conocimiento del dominio importan más que el algoritmo.

Próxima clase: Algoritmos de clustering (k-means, jerárquico, DBSCAN), elección del número de clusters, y discusión de IA responsable en segmentación.

Referencias y Lecturas Recomendadas

Libros de texto:

- James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (ISLR), 2nd Ed. Cap. 12: Unsupervised Learning. <https://www.statlearning.com>
- Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning* (ESL), 2nd Ed. Cap. 14. <https://hastie.su.domains/ElemStatLearn/>
- Jolliffe, I.T. & Cadima, J. (2016). "Principal component analysis: a review and recent developments." *Phil. Trans. R. Soc. A*, 374(2065).

Aplicaciones económicas de PCA:

- Filmer, D. & Pritchett, L.H. (2001). "Estimating wealth effects without expenditure data — or tears." *Demography*, 38(1), 115–132. (*PCA para índices de riqueza*)
- Vyas, S. & Kumaranayake, L. (2006). "Constructing socio-economic status indices: How to use PCA." *Health Policy and Planning*, 21(6), 459–468.
- Fader, P., Hardie, B. & Lee, K.L. (2005). "RFM and CLV: Using iso-value curves for customer base analysis." *J. Marketing Research*, 42(4), 415–430.

Recursos prácticos: scikit-learn (PCA):

<https://scikit-learn.org/stable/modules/decomposition.html> factoextra (R):

<https://rpkgs.datanovia.com/factoextra/>

¡Gracias!

s.neira10@uniandes.edu.co