

# Agrupamiento con K-Means

## Ventajas y Desventajas

# K-Means Ventajas y Desventajas

## Ventajas

- Fácil de comprender
- Rápido en agrupar
- Ampliamente disponible (Python, R y otros)
- Fácil de implementar
- Siempre da un resultado\*\*

\*\* esto también puede ser un problema, ya que puede ser engañoso)

# K-Means Ventajas y Desventajas

## Desventajas

1. Hay que escoger el valor de K
2. Es sensitivo a la inicialización

## Remedio

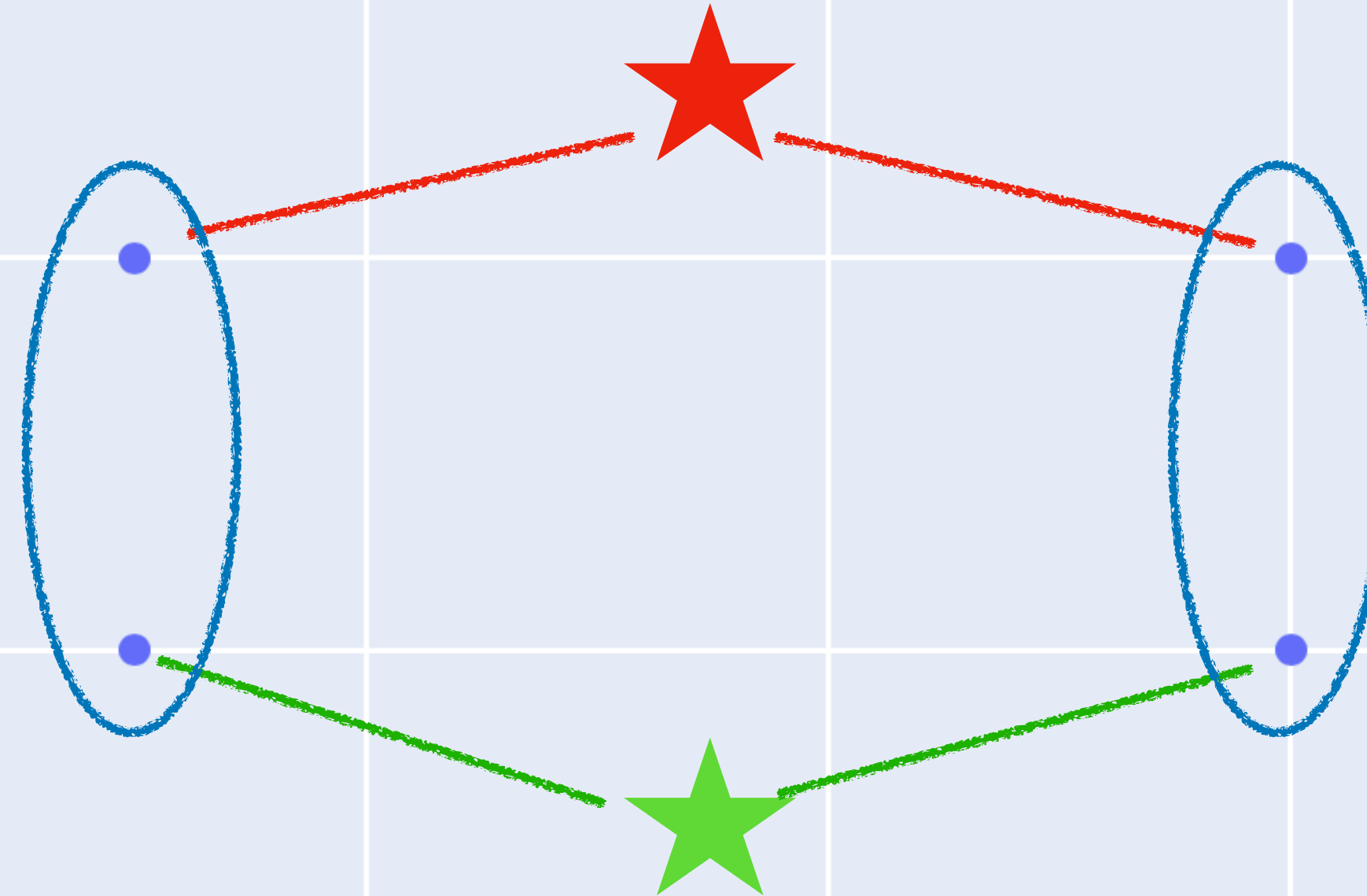
1. Método del “codo” (elbow)

# Desventajas

## Sensitivo a la inicialización

Pero pareciera que una solución más adecuada es:

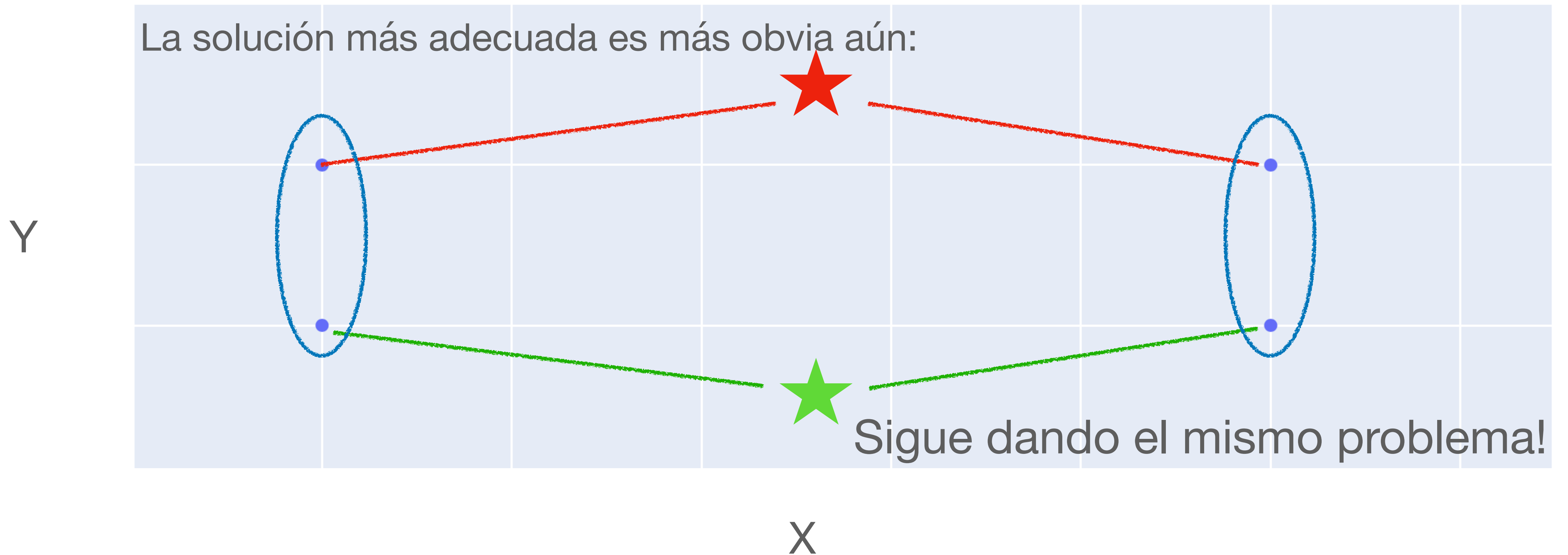
Y



X

# Desventajas

Es más, si separamos los puntos



# K-Means Ventajas y Desventajas

## Desventajas

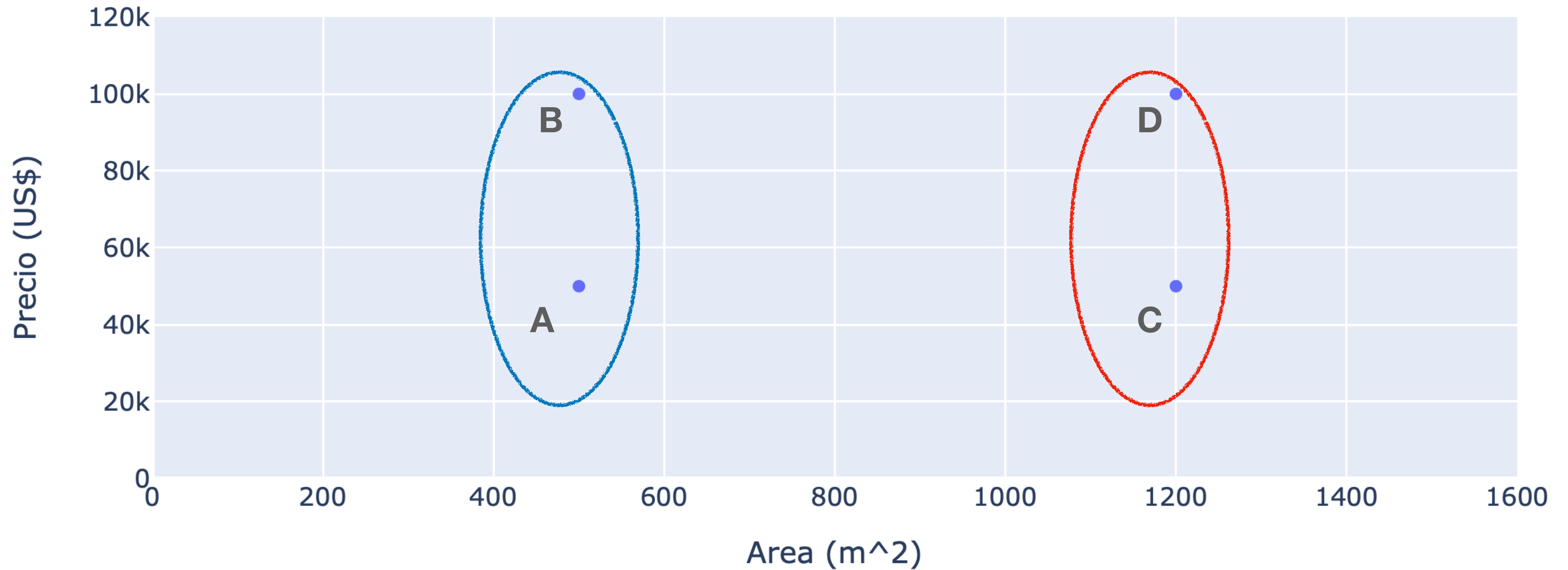
1. Hay que escoger el valor de K
2. Es sensitivo a la inicialización
3. Es sensitivo a valores atípicos
4. Produce soluciones esféricas
5. Estandarización

## Remedio

1. Método del “codo” (elbow)
2. k-means++
3. Eliminar los valores atípicos

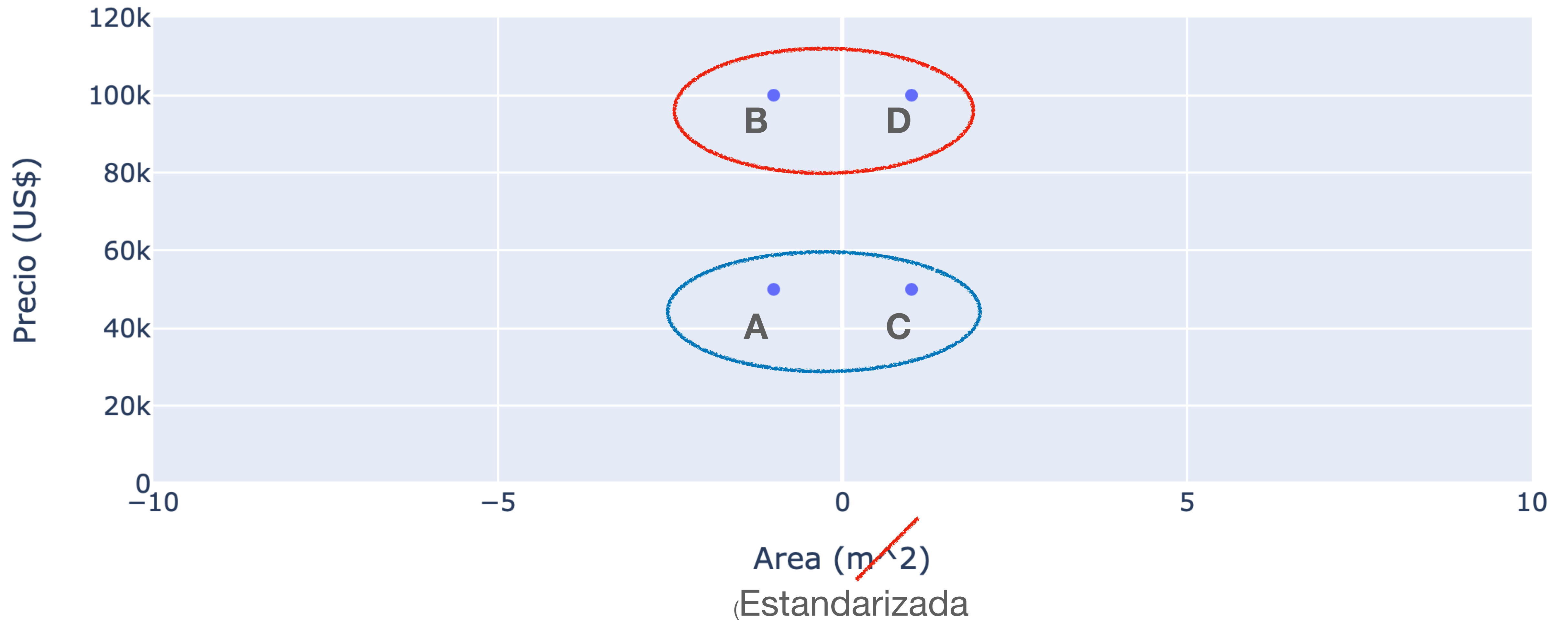
**Estandarizar o no estandarizar**

¿Cómo podemos agrupar las 4 observaciones?  
A simple vista:





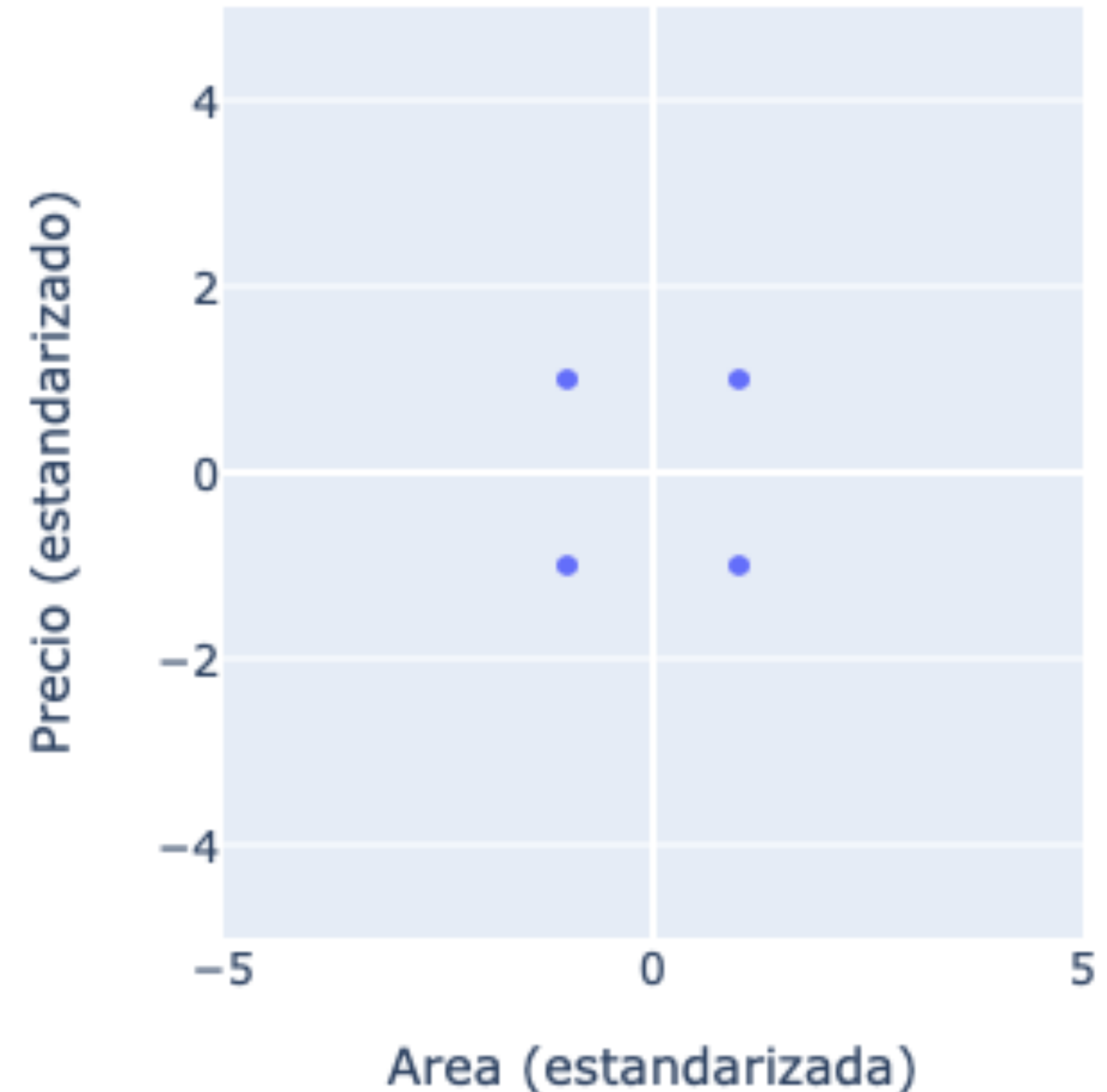
¿Y si estandarizamos el Area?



# ¿Y si estandarizamos las dos variables?

Ahora no tenemos cómo decidir, **es un cuadrado perfecto**.

Nos fuimos de una solución, a otra totalmente diferente, a ninguna solución.



**La meta final de la estandarización es reducir el peso de los número más grandes, e incrementar el de los más bajos.**

**El rango de las variables actúa como un peso y el agrupamiento favorece a los más grandes**

**La estandarización intenta poner a todas las variables en términos equitativos.**

**Es buena práctica hacerlo,  
especialmente si nos estamos iniciando  
en este campo.**

**Sin embargo, si nosotros queremos que ciertas variables tengan más peso, entonces no se debe utilizar la estandarización.**

**En estos ejemplos, sabemos que el precio, en general, tiene más valor que el área. Si no podemos pagarlo, de nada sirve qué área tiene el terreno!**

**¿Cómo sabemos si debemos, o no, estandarizar?  
Una vez más entra en juego la experiencia...el experto en el dominio!**

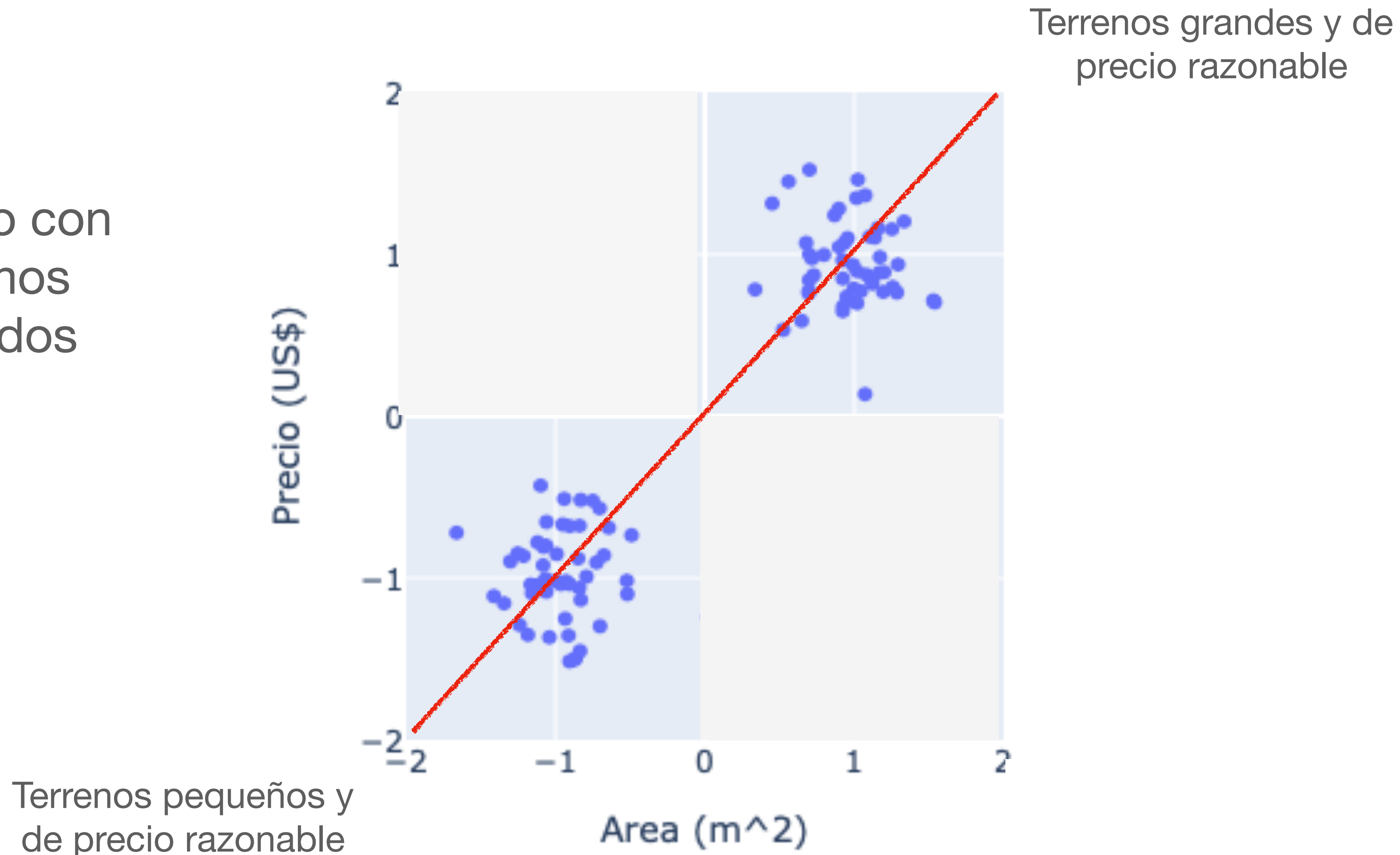
¿Cuál es la relación entre  
Regresión y Clustering?

# Hemos estado trabajando con una gráfica simplificada

La realidad probablemente es más como esta

Nos queda algo con lo que ya estamos muy familiarizados

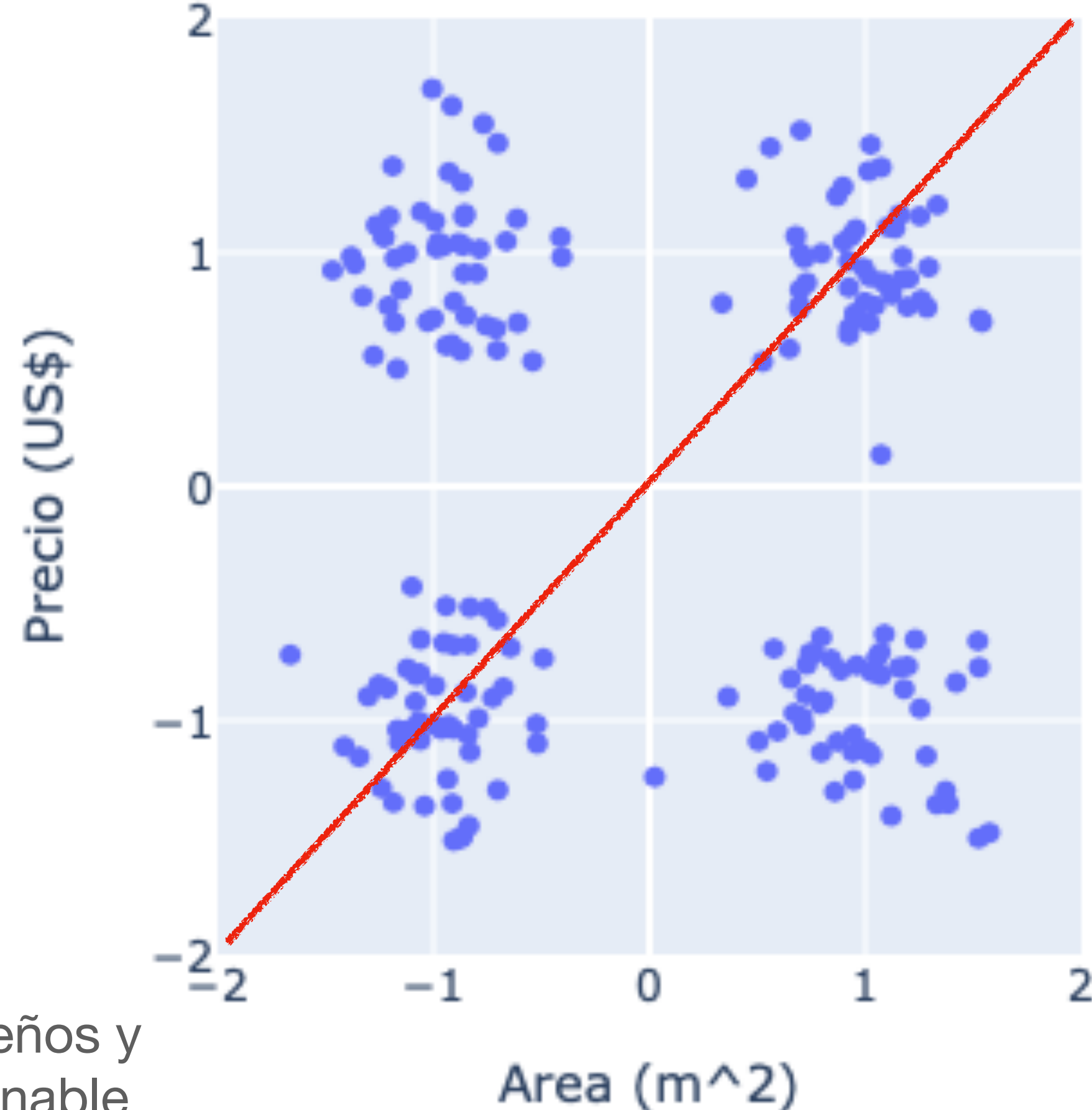
Una Regresión





# El Clustering si tiene relación con la regresión!

Puede que estos no sean estafas, sino terrenos en el centro de la ciudad



Terrenos grandes y de precio razonable

Quizás debiéramos considerar una variable de ubicación.

Clustering nos puede ayudar a detectar el **Sesgo de Variables Omitidas (OVB)** por sus siglas en inglés).

Terrenos pequeños y de precio razonable

Puede que estos terrenos estén en suburbios