**📊 Should we Standardize?**

**1️⃣ Why is this even a question?**

**Your tutor explains that whether to standardize data before clustering is a debated topic.  
There are arguments for and against, and the decision depends on the situation.  
So let’s explore it with an example.**

**2️⃣ The Apartment Example**

**We have 4 apartments, plotted on a scatter plot:**

* **X-axis: Size (square feet)**
* **Y-axis: Price (USD)**

| **Apartment** | **Size (sqft)** | **Price (USD)** |
| --- | --- | --- |
| **A** | **500** | **50,000** |
| **B** | **500** | **100,000** |
| **C** | **1,200** | **50,000** |
| **D** | **1,200** | **100,000** |

**🔷 Without standardization**

* Looking at the scatter plot, a reasonable clustering is:  
  **AB (small size) and CD (large size)**.  
  Why? Because price dominates the scale (higher numbers) and influences K-means more.

**3️⃣ What happens if we standardize?**

**🔷 Step 1: Standardize only Size (X)**

* Now sizes are standardized to either -1 or 1.
* New clustering appears to make sense as:  
  **AC (low price) and BD (high price)**.

**🔷 Step 2: Standardize both Size (X) and Price (Y)**

* Both variables now range from -1 to 1.
* The plot becomes a **perfect square**, and it’s no longer clear which clustering is better:
  + AB & CD?
  + AC & BD?
  + Both look equally reasonable.

**4️⃣ Why does this happen?**

✅ **Standardization removes the influence of scale.**  
✅ It prevents variables with large ranges (like Price) from dominating clustering.  
✅ It essentially gives all variables equal weight.

But — sometimes equal weighting **doesn’t make sense**, if you already know one variable is more important.

**5️⃣ Why does scale matter?**

In our original (unstandardized) data:

* Price values (50,000–100,000) were much larger than Size values (500–1,200).
* K-means “thinks” Price is more important because of its numerical scale, and clusters accordingly.
* That’s why the clusters leaned towards price.

**6️⃣ So, should we standardize?**

✅ **Yes, standardize if:**

* You want to give all variables equal importance.
* You’re a beginner and don’t yet know which variables matter more.
* You want to avoid letting larger-scale variables dominate.

🚫 **Don’t standardize if:**

* You already know that one variable is more important than others.
* For example, if price is *definitely* more important than size (because people care about price first), you might not want to standardize and give size equal weight.

**7️⃣ Summary Table**

| **Situation** | **Standardize?** |
| --- | --- |
| You want equal weight for all variables | ✅ Yes |
| One variable is inherently more important | 🚫 No |
| You’re not sure yet | ✅ Yes (good practice for beginners) |