**📊 Lecture: Clustering, Regression, and Omitted Variables**

**🎯 1️⃣ Real-World Context of the Apartment Example**

**In the previous example, we looked at 4 points (A, B, C, D), representing apartments with different sizes and prices.**

**But in real life:**

**✅ You’d have many data points, and they’d likely form 4 clusters — not just 4 isolated points.**

**🏠 2️⃣ What do the clusters mean?**

**The instructor gave an interpretation of the 4 groups:**

| **Cluster** | **Description** |
| --- | --- |
| **A** | **Small, reasonably priced apartments** |
| **B** | **Small, expensive apartments (rip-offs)** |
| **C** | **Big, cheap apartments (bargains)** |
| **D** | **Big, reasonably priced apartments** |

**Real-life insight:**

* Small & expensive = likely city center (high demand → high price).
* Big & cheap = likely suburbs (low demand → low price).

So clustering helped us uncover these **patterns in the data.**

**🔷 3️⃣ Connection to Regression**

If you remove the “rip-offs” and “bargains” (outliers) and only keep the usual pattern:

* Small apartments → cheaper.
* Big apartments → more expensive.

That’s just a **classic regression relationship**:

Price = f(Size)  
where price increases as size increases.

So **clustering and regression can complement each other.**

**🧩 4️⃣ Clustering Helps Identify Omitted Variable Bias**

**What does that mean?**

When you run a regression with only **Size → Price**,  
you might miss another important variable — like **Location**.

How does clustering help?

* You notice the data splits into groups that size alone can’t explain.
* This suggests you’re missing a variable (e.g., apartments in city center vs. suburbs).
* Including **Location** in your regression would improve your predictions.

So clustering is a way to explore the data and discover hidden patterns — and sometimes realize your regression model is missing something important.

**📌 5️⃣ Key Takeaways**

✅ Clustering and regression can work **together**:

* Clustering: find patterns, detect groups, discover missing variables.
* Regression: quantify relationships, make predictions.

✅ Use clustering to **explore the data first**, and identify if any key variables are missing.

✅ Then build a better regression model that includes all important variables.

**🔷 Final note:**

This lecture isn’t just about clustering — it’s about **thinking like a data scientist:**

* Explore the data.
* Understand the patterns.
* Build models that reflect the real world.