Statistics and Data Analysis
Unit 05 – Lecture 07 Notes
Case Study: PCA + Clustering

Tofik Ali

February 17, 2026

Topic

Case study: PCA + KMeans clustering; visualize and interpret clusters.

How to Use These Notes

These notes are written for students who are seeing the topic for the first time. They follow the slide order, but add the missing 'why', interpretation, and common mistakes. If you get stuck, look at the worked exercises and then run the Python demo.

Course repository (slides, demos, datasets): `https://github.com/tali7c/Statistics-and-Data-Analysis`

Time Plan (55 minutes)

- 0–10 min: Attendance + recap of previous lecture
- 10–35 min: Core concepts (this lecture's sections)
- 35–45 min: Exercises (solve 1–2 in class, rest as practice)
- 45–50 min: Mini demo + interpretation of output
- 50–55 min: Buffer / wrap-up (leave 5 minutes early)

Slide-by-slide Notes

Title Slide

State the lecture title clearly and connect it to what students already know. Tell students what they will be able to do by the end (not just what you will cover).

Quick Links / Agenda

Explain the structure of the lecture and where the exercises and demo appear.

- Overview

- Pipeline
- Interpretation
- Exercises
- Demo
- Summary

Learning Outcomes

- Run PCA before clustering for visualization/stability
- Explain why scaling matters for clustering
- Use KMeans and interpret clusters cautiously
- Visualize clusters in PCA space

Why these outcomes matter. **PCA** finds new axes (principal components) that capture maximum variance. It is a rotation of the feature space. Because PCA is variance-based, it is sensitive to scaling: standardize features first unless all features are already comparable.

Pipeline: Key Points

- Scale features
- Run PCA (2D for visualization)
- Cluster (KMeans) and visualize

Explanation. **PCA** finds new axes (principal components) that capture maximum variance. It is a rotation of the feature space. Because PCA is variance-based, it is sensitive to scaling: standardize features first unless all features are already comparable.

Interpretation: Key Points

- Clusters are patterns, not truth
- Check stability across seeds/k
- Explain clusters using original variables

Exercises (with Solutions)

Attempt the exercise first, then compare with the solution. Focus on interpretation, not only arithmetic.

Exercise 1: Scaling

Why scale before KMeans?

Solution

- Distance-based; scale dominates otherwise.

Exercise 2: Choose k

Name one heuristic to choose k.

Solution

- Elbow, silhouette, domain knowledge.

Exercise 3: Explain cluster

How to explain cluster to non-technical audience?

Solution

- Describe in original variables (high spend, frequent visits, etc.).

Mini Demo (Python)

Run from the lecture folder:

```
python demo/demo.py
```

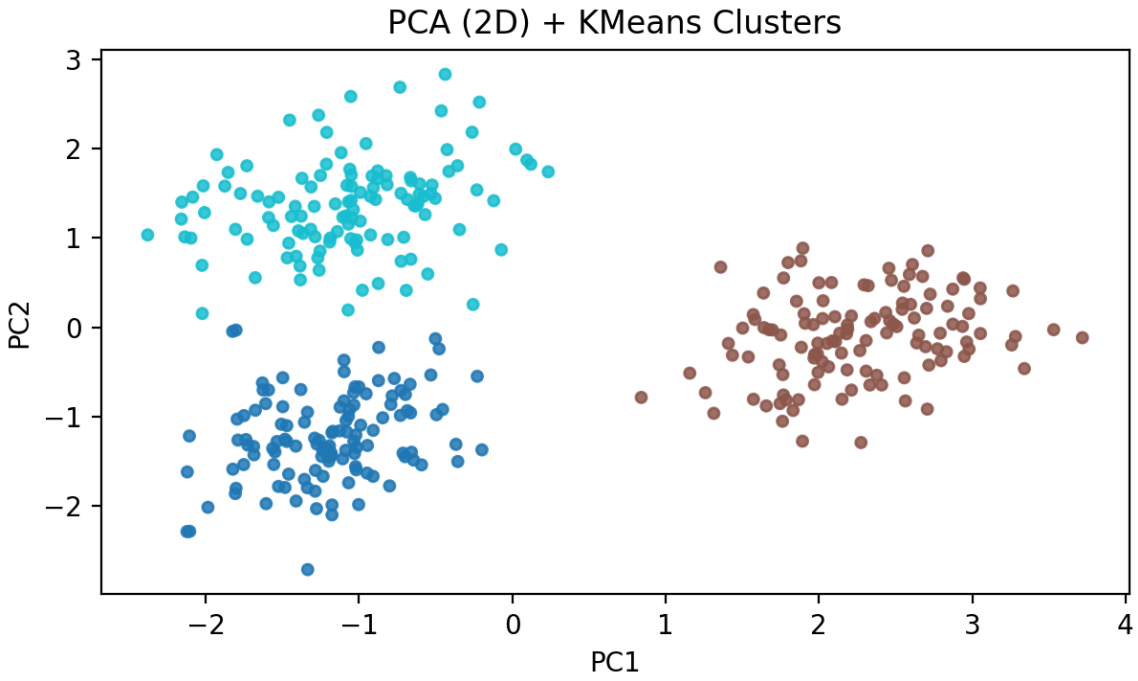
Output files:

- `images/demo.png`
- `data/results.txt`

What to show and say.

- Scales features, runs PCA for 2D visualization, then clusters with KMeans.
- Plots clusters in PCA space to discuss patterns vs ground truth.
- Use it to talk about choosing k (elbow/silhouette) and stability checks.

Demo Output (Example)



Summary

- Key definitions and the main formula.
- How to interpret results in context.
- How the demo connects to the theory.

Exit Question

Why should you validate cluster stability before using clusters for decisions?

Suggested answer (for revision). Clusters can change with seed/k; stability checks prevent making decisions based on accidental patterns in one run.

References

- Montgomery, D. C., & Runger, G. C. *Applied Statistics and Probability for Engineers*, Wiley.
- Devore, J. L. *Probability and Statistics for Engineering and the Sciences*, Cengage.
- McKinney, W. *Python for Data Analysis*, O'Reilly.

Appendix: Slide Deck Content (Reference)

The material below is a reference copy of the slide deck content. Exercise solutions are explained in the main notes where applicable.

Title Slide

https://github.com/tali7c/Statistics-and-Data-Analysis

Quick Links

Overview Pipeline Interpretation Exercises Demo Summary

Agenda

- Overview
- Pipeline
- Interpretation
- Exercises
- Demo
- Summary

Learning Outcomes

- Run PCA before clustering for visualization/stability
- Explain why scaling matters for clustering
- Use KMeans and interpret clusters cautiously
- Visualize clusters in PCA space

Pipeline: Key Points

- Scale features
- Run PCA (2D for visualization)
- Cluster (KMeans) and visualize

Interpretation: Key Points

- Clusters are patterns, not truth
- Check stability across seeds/k
- Explain clusters using original variables

Exercise 1: Scaling

Why scale before KMeans?

Solution 1

- Distance-based; scale dominates otherwise.

Exercise 2: Choose k

Name one heuristic to choose k.

Solution 2

- Elbow, silhouette, domain knowledge.

Exercise 3: Explain cluster

How to explain cluster to non-technical audience?

Solution 3

- Describe in original variables (high spend, frequent visits, etc.).

Mini Demo (Python)

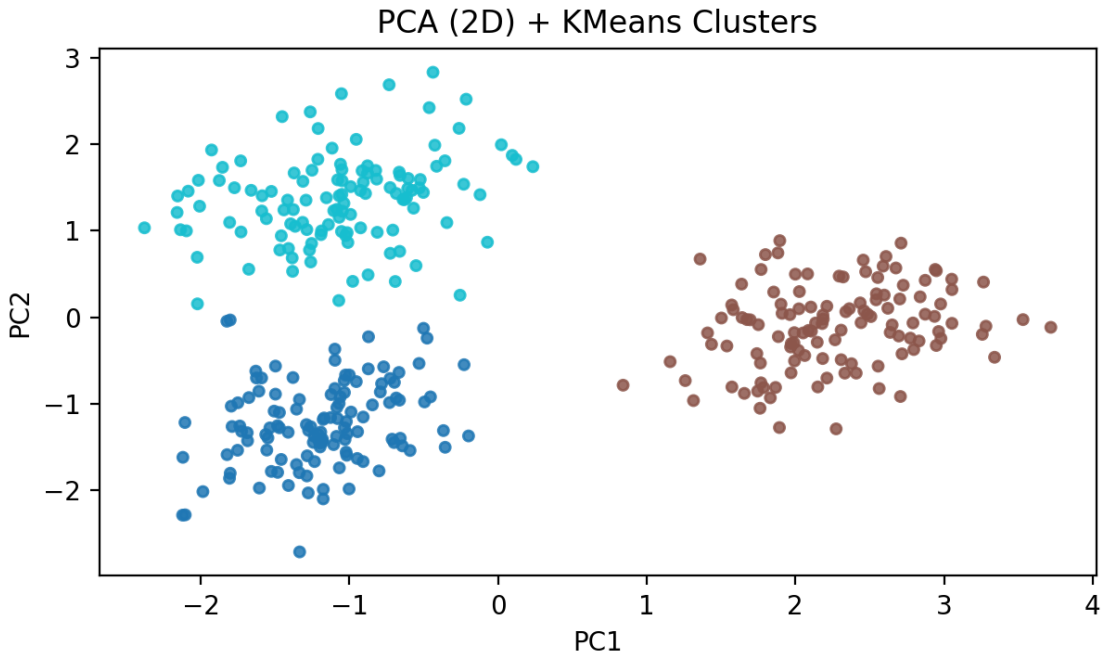
Run from the lecture folder:

```
python demo/demo.py
```

Outputs:

- `images/demo.png`
- `data/results.txt`

Demo Output (Example)



Summary

- Key definitions and the main formula.
- How to interpret results in context.
- How the demo connects to the theory.

Exit Question

Why should you validate cluster stability before using clusters for decisions?