

# Statistics and Data Analysis

## Unit 05 – Lecture 07: Case Study: PCA + Clustering

Tofik Ali

School of Computer Science, UPES Dehradun

February 14, 2026

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

Overview  
○

Pipeline  
○

Interpretation  
○

Exercises  
○○○○○○○

Demo  
○○

Summary  
○○

# Quick Links

Overview

Pipeline

Interpretation

Exercises

Demo

Summary

# Agenda

1 Overview

2 Pipeline

3 Interpretation

4 Exercises

5 Demo

6 Summary

# Learning Outcomes

- Run PCA before clustering for visualization/stability

# Learning Outcomes

- Run PCA before clustering for visualization/stability
- Explain why scaling matters for clustering

# Learning Outcomes

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

# 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

# Pipeline: Key Points

- Scale features
- Run PCA (2D for visualization)

# Pipeline: Key Points

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

# Interpretation: Key Points

- Clusters are patterns, not truth

# Interpretation: Key Points

- Clusters are patterns, not truth
- Check stability across seeds/k

# 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)

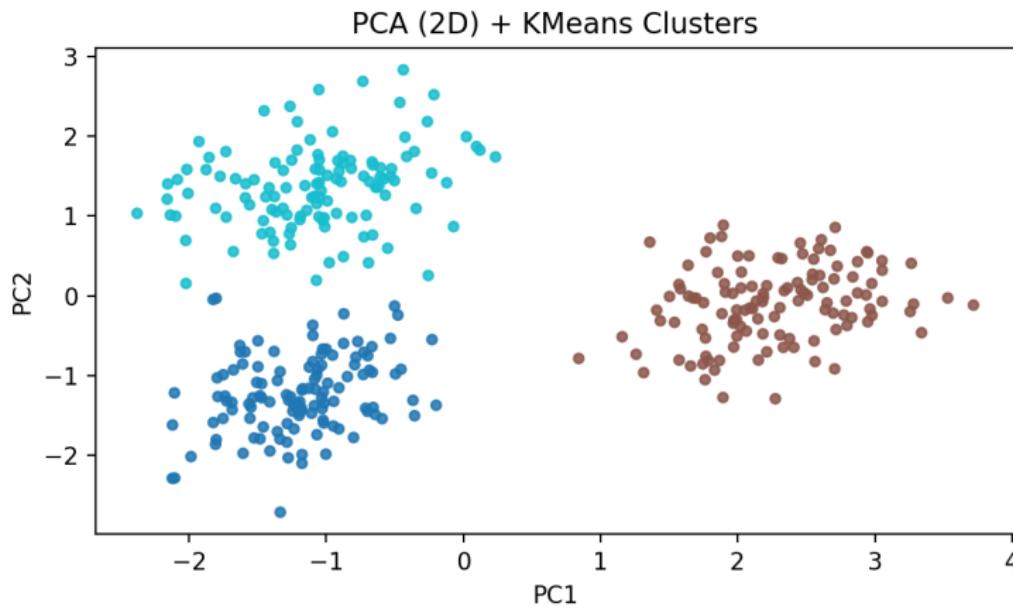
Run from the lecture folder:

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

Outputs:

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

# Demo Output (Example)



# Summary

- Key definitions and the main formula.

# Summary

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

# 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?