Statistics and Data Analysis
Unit 05 – Lecture 07 Notes

Tofik Ali

February 14, 2026

Topic

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

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

Detailed Notes

These notes are designed to be read alongside the slides. They expand each slide bullet into plain-language explanations, small worked examples, and common pitfalls. When a formula appears, emphasize (1) what each symbol means, (2) the assumptions needed to use it, and (3) how to interpret the final number in the problem context.

Pipeline

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

Interpretation

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

Exercises (with Solutions)

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.).

Exit Question

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

Demo (Python)

Run from the lecture folder:

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

Output files:

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

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.