Data Science

Shahid Beheshti University

Spring 2025

User segmentation using clustering

Assignment 3

1 Theoretical

1. Explain how Gaussian Mixture Models (GMM), K-means++, and Spectral Clustering work for clustering data. Highlight scenarios where each method is most suitable.
2. What strategies exist for clustering datasets containing both numerical and categorical variables?
3. Compare and contrast soft clustering (e.g., GMM) with hard clustering methods (e.g., K-Means). In what scenarios is soft clustering more appropriate?
4. Can clustering be used for anomaly detection? Explain how DBSCAN or GMM can detect outliers.
5. What challenges arise when clustering imbalanced datasets, and how can we ensure meaningful cluster formation in such cases?
6. You are given a dataset of customer transactions with the following features:
   - Annual income
   - Age
   - Total spending in the last year

   You apply hierarchical clustering using the Ward linkage method.

   (a) Explain how hierarchical clustering builds the cluster hierarchy.

   (b) How would you determine the optimal number of clusters from the dendrogram?

   (c) Suppose you cut the dendrogram at 3 clusters. How can you interpret these clusters in terms of customer behavior?

2 Practical

In this assignment you will **discover and profile data-driven customer segments** for the Brazilian marketplace Olist. Your task is to design a clustering workflow, defend every design choice, and translate the resulting groups into concrete business recommendations.

2.1 Dataset Overview

The dataset consists of nine CSV tables (100 000 orders, 2016–2018) that capture payments, products, delivery performance, customer geography, reviews, and more. An entity–relationship diagram with join keys is provided. You *do not* need to use every table—select only those that help you build meaningful *customer-level* features.

2.2 Assignment Tasks

2.2.1 Data Pre-processing & Exploratory Analysis

- Clean and merge the selected tables, handling missing values, outliers and data-type issues.
- Aggregate records so that each row represents a single customer.
- Perform visual and statistical EDA to uncover patterns (e.g. spend distribution, review scores, geography, payment mix).
- Justify every feature you keep, transform or create.

2.2.2 Baseline Clustering

- Fit an initial clustering model of your choice (e.g. K-means with a plausible k, hierarchical, DBSCAN).
- Report internal validation metrics (silhouette, Davies–Bouldin, etc.) and diagnostic plots (elbow curve, dendrogram, cluster overlays on PCA/t-SNE).

2.2.3 Model Variants & Dimensionality Reduction

- Experiment with *at least one* alternative approach or parameter set:
  - different distance metrics or linkage methods,
  - density-based versus partitioning algorithms,
  - PCA, UMAP, or alternative scaling schemes.
- Discuss trade-offs such as stability, interpretability and business usefulness.

2.2.4 Cluster Profiling & Business Insight

- For each final segment, provide:
  - key statistics (average order value, favourite categories, sentiment, region, etc.),
  - a concise marketing label (e.g. "Premium gadget lovers", "Rural bargain hunters"),
  - *at least two* actionable recommendations (targeted promotion, logistics tweak, loyalty perk, . . .).

2.2.5 Model Enhancement through Feature Engineering (optional but recommended)

- Test interaction terms, temporal behaviour (order frequency) or review-text sentiment scores.
- Explain how each change affects cluster cohesion or interpretability.

2.3 Deliverables

1. **Notebook** (Jupyter or Colab) — fully reproducible and commented.
2. **PDF report** that includes:
   - a concise summary of pre-processing and EDA (with figures),
   - validation results for all clustering variants,

- clear cluster portraits and recommended actions,
- reflections on limitations and future work.

2.4 Bonus Ideas (Optional)

- Combine with the separate *Marketing Funnel* dataset to explore alignment between segments and acquisition channels.
- Test temporal stability by training clusters on 2017 data and evaluating them on 2018.(use GMM)
- Prototype an uplift model that predicts which segment is most responsive to free-shipping offers.