CLUSTERING ANALYSIS

1. OVERVIEW

This project applies unsupervised machine learning techniques to cluster restaurants in Bangalore based on cost, cuisine type, and service attributes. Using K-Means clustering, we aim to uncover hidden patterns and segment restaurants into meaningful groups that reflect pricing strategy, service offerings, and cuisine preferences.

2. KEY FEATURES USED FOR CLUSTERING

The following variables were selected as input features:

- averagecost: Average cost for two people.
- ishomedelivery: Binary indicator for availability of home delivery.
- istakeaway: Binary indicator for availability of takeaway services.
- isvegonly: Binary indicator for vegetarian-only restaurants.
- isindoorseating: Binary indicator for indoor seating availability.
- area: Locality or neighborhood in Bangalore.
- cuisines: Cuisine(s) offered by the restaurant.

3. DATA PREPROCESSING

3.1 Handling Missing Values

- Missing averagecost values are filled with the median cost to avoid skewing cluster centers with extreme values.
- Missing area entries are replaced with "Unknown" to preserve data and handle categorical encoding.
- Missing cuisines are filled with "Other" to maintain consistency.

3.2 Simplification Of Cuisine

- To reduce complexity, only the first cuisine listed is retained for each restaurant.
- This approach simplifies multi-label cuisine data into a single categorical variable, making encoding and clustering easier.

Note: While this reduces complexity, it may lose information about multi-cuisine offerings.

3.3 Feature Transformation

- Numeric features are standardized to mean 0 and variance 1 to equalize their influence.
- Categorical features are one-hot encoded, converting each category into a binary feature.
- Binary features are retained without transformation.
- This pipeline produces a numeric matrix suitable for K-Means clustering.

4. CLUSTERING METHODOLOGY

4.1 Clustering Approach

We used the K-Means clustering algorithm with k = 5, which segments the data into 5 distinct clusters. The choice of k=5 was based on an assumed balance between detail and interpretability, though this could be fine-tuned using methods like the elbow method or silhouette analysis in future iterations.

4.2 Model Training

- The preprocessed feature matrix (including standardized numeric, one-hot encoded categorical, and binary features) was input to K-Means.
- The algorithm assigned each restaurant to one of 5 clusters.

5. PCA VISUALISATION OF CLUSTERS

To better understand the clustering results, we applied Principal Component Analysis (PCA) to the preprocessed feature matrix. This dimensionality reduction technique helps project the high-dimensional data into two dimensions, allowing visual inspection of cluster separation.

The scatter plot below represents each restaurant as a point in this reduced space, with colors indicating their respective cluster assignments. This visualization highlights the relative distinction between clusters and suggests that the selected features contribute effectively to differentiating restaurant profiles.

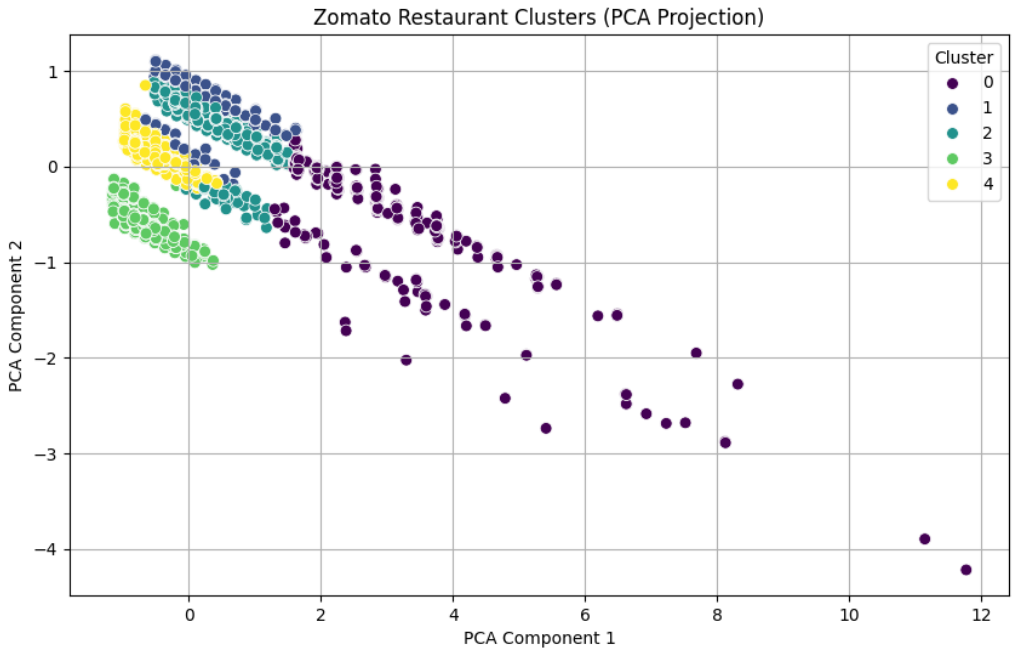

Figure 1: PCA projection of restaurant clusters using KMeans (k=5). Each point represents a restaurant projected onto two principal components, colored by cluster assignment.

6. RESULTS

6.1 Cluster summary statistic

cluster	averagecost	ishomedelivery	istakeaway	isvegonly	isindoorseating	Restaurant_count
0	1159.483961	1.000000	0.920502	0.061367	0.972106	717
1	323.637821	1.000000	0.982372	0.248397	0.939103	624
2	396.755240	1.000000	0.982784	0.085704	0.993263	2672
3	197.961631	0.993491	0.000000	0.024666	0.004111	2919
4	183.098945	1.000000	1.000000	0.072828	0.000502	1991

Interpretation:

- **Cluster 4** consists of premium restaurants with the highest average cost and high indoor seating.
- **Cluster 3** also represents high-cost restaurants with nearly universal indoor seating and delivery.
- **Cluster 0** is mid-range with high home delivery and takeaway but moderate indoor seating.
- **Cluster 1** is mostly delivery only (0% takeaway) and very low indoor seating.
- **Cluster 2** offers a balance of home delivery and takeaway with moderate indoor seating.

6.2 Top Cuisines Per Cluster

Interpretation:

- **Cluster 0:** Pizza, Fast Food, Desserts, Beverages — suggests casual dining and quick service.
- **Cluster 1:** North Indian, South Indian, Ice Cream, Bakery — traditional Indian fast casual.
- **Cluster 2:** South Indian, Bakery, Biryani, Street Food — popular local and snack foods.
- **Cluster 3:** North Indian, Chinese, Mughlai, Thai, Seafood — diverse cuisines with emphasis on dine-in.
- **Cluster 4:** BBQ, Mughlai, Biryani, North Indian — premium and specialty cuisines.

7. CONCLUSION

This clustering analysis using KMeans has revealed five distinct groups based on pricing, service options, location, and primary cuisine type. Each cluster represents a unique restaurant profile — from budget-friendly, delivery-only outlets to premium dine-in experiences with diverse, high-end cuisines. The clustering effectively distinguishes operational models and culinary focus, which can inform strategic decisions for restaurant owners, food delivery platforms, and marketers. For instance, Cluster 1 identifies low-cost, delivery-centric places that could benefit from expanded takeaway options, while Cluster 4 represents upscale restaurants with opportunities for exclusive marketing and loyalty programs. Overall, the clustering successfully uncovers hidden structure in the restaurant ecosystem, making it easier to understand market segments and customer targeting.

8. LIMITATIONS

Despite its insights, this analysis has several limitations. Simplifying multiple cuisines to just the first listed one removes valuable information about diverse offerings, which can distort cluster interpretations. The number of clusters (k=5) was chosen heuristically without validation through techniques like the elbow method or silhouette score, which may affect clustering quality. Additionally, important features like customer ratings, review sentiments, and geographical coordinates were not included, although they likely influence restaurant popularity and type. Finally, KMeans assumes spherical clusters and equal variance, which may not reflect the true structure of complex real-world data, potentially oversimplifying some relationships.