Report

Project Report: Zomato Restaurant Clustering

Project Type: Unsupervised Machine Learning **Contribution:** Individual

1. Executive Summary

In this project, I analyzed the Zomato dataset to identify distinct segments of restaurants based on cost, popularity, and cuisine offerings. By leveraging **K-Means Clustering** and **Principal Component Analysis (PCA)**, I successfully grouped restaurants into strategic clusters (Budget, Premium, Casual Dining). These insights enable targeted marketing and improved recommendation systems for the platform.

2. Problem Statement

With thousands of restaurants on Zomato, users struggle to find places that match their specific preferences beyond simple filters. The goal was to build an unsupervised model that automatically groups restaurants into "Clusters" based on shared characteristics like price point, rating, and menu diversity without pre-existing labels.

3. Methodology

3.1 Data Preprocessing & Feature Engineering

- **Handling Unstructured Data:** The Cuisines column was stored as complex lists (e.g., ['North Indian', 'Chinese']). I built a custom pipeline to "explode" and normalize these lists, allowing for granular analysis of cuisine types.
- **Feature Creation:** I engineered a new feature, Total_Cuisine_Count, to quantify menu diversity and differentiate between specialty shops and multi-cuisine restaurants.
- **Log Transformation:** To handle the highly skewed Cost data (where a few luxury places distorted the scale), I applied a Log Transformation (np.log1p), ensuring the model wasn't biased toward outliers.

3.2 Dimensionality Reduction (PCA)

To manage the high dimensionality created by encoding multiple cuisines, I implemented **Principal Component Analysis (PCA)**. This compressed the features into **3 Principal**

Components, retaining the essential variance while reducing noise and enabling 3D visualization.

3.3 Model Implementation (K-Means)

- **Optimal K:** I utilized the **Elbow Method** to plot WCSS against cluster counts, identifying the optimal number of clusters where the error rate leveled off.
- **Validation:** I performed **Silhouette Analysis** to validate the quality of the clusters. The resulting Silhouette Plots confirmed that the clusters were well-separated and distinct.
- **Results:** The final K-Means model successfully segmented the restaurants into clear groups:
    1. **Budget / Quick Bites:** Low cost, limited menu.
    2. **Premium Dining:** High cost, high ratings.
    3. **Casual / Family Dining:** Moderate cost, high cuisine variety.

4. Conclusion & Impact

I successfully transformed raw metadata into a strategic decision-making tool.

- **For Zomato:** The clusters allow for segment-specific marketing (e.g., promoting "Gold" memberships to the Premium cluster).
- **For Users:** The model powers a "Smart Recommendation" system, suggesting new restaurants that belong to the same cluster as a user's favorites, ensuring a consistent dining experience.

This project demonstrates the power of unsupervised learning to uncover hidden patterns in complex, real-world data.