Methodology

Initial Dataset and Approach

The core dataset was a mixed dataset combining categorical and numerical variables. Most features were categorical, which strongly influenced our choice of methods. However, the inclusion of numerical data introduced challenges in processing both types together. Thinking through the problem, our approach began with a few key steps: splitting the dataset into numerical and categorical subsets (standardizing numerical features as needed); applying dummy indicators to convert categorical variables into numeric form; applying appropriate clustering algorithms (k-Means for numerical features and k-Modes for categorical features); and exploring ways to combine the results of both models. When clustering categorical data, the most suitable methods require recoding levels—often via dummy indicators—and aggregating by mode, as in k-Modes. Conversely, numerical features typically demand k-Means. We chose **k-Prototypes** for the final clustering step because it unifies these processes and offers native support for mixed-data attributes. :contentReference[oaicite:2]{index=2}:contentReference[oaicite:3]{index=3}

Two-Tier Analysis

Some features were highly dimensional, with 100+ to over 1,000 levels. Considering our computational constraints, we adopted a two-tiered analysis.

First, we addressed the high dimensionality of customer types, which had over 150 distinct levels. Using SAS JMP software, we conducted Multiple Correspondence Analysis (MCA), a technique that reduces high-cardinality categorical variables into principal components. MCA reveals relationships within and between groups of variables and can handle multiple dimensions (not to be confused with the number of levels). By using MCA, we avoided extreme dataset widening and costly granularity, preserving our computational capacity. The resulting MCA biplot revealed at least four distinct customer subgroups, three tightly clustered around a single business group. This finding confirmed strong associations between customer types and business groups and allowed us to introduce an intermediate classification layer between broad business groups and granular customer types. :contentReference[oaicite:4]{index=4}:contentReference[oaicite:5]{index=5}

Next, in a simulated Jupyter Notebook environment (see Appendix), we used the MCA component loadings (correlations with principal components) to visualize the scatterplot of correlations; generate a sample dataset and overlay it on the raw data; apply agglomerative hierarchical clustering (with dendrogram) to assess sample clusters; and use the dendrogram cutoff (k) to create and visualize the MCA-determined customer-type clusters, particularly assessing outliers. Following this initial phase, we applied the elbow method to determine the optimal number of clusters (k) for the full dataset, followed by k-Prototypes using the selected k to generate final clustering profiles. :contentReference[oaicite:6]{index=6}:contentReference[oaicite:7]{index=7}

Optimizations

To optimize our model, we reduced customer-type levels via MCA and increased the association strength (λ) from 1% to 33% by grouping types into observed clusters. We also applied agglomerative hierarchical clustering with an average-link distance measure, which is less sensitive to outliers. :contentReference[oaicite:8]{index=8}:contentReference[oaicite:9]{index=9}