Project 4: Predicting House Prices Using Regression

Introduction

Fast food has become popular in today's fast-paced world due to its convenience and low cost. However, as more people focus on health, there is a growing interest in understanding the nutritional value of these menu items. This project intends to analyze fast food nutrition data by using clustering techniques to group menu items by their dietary profiles.

The main problem I'm trying to solve is how to categorize fast food menu items based on their nutritional content to help consumers make more informed choices. Specifically, I want to answer:

- ★ Can we find natural groupings of foods like "low-calorie," "high-protein," or "high-fat/high-sodium"?
- ★ Do certain restaurants offer more items in specific nutritional categories?
- ★ Are there surprising items (foods marketed as healthy) that cluster with unhealthy options?

What is Clustering?

Clustering is an unsupervised machine learning technique that groups similar data points based on their characteristics. One of the most popular clustering algorithms is K-means. This algorithm minimizes the within-cluster sum of squares, which measures how tightly the clusters are packed together. Another clustering algorithm is agglomerative hierarchical clustering, which begins with each item as an individual cluster and gradually merges pairs of clusters based on their similarities until all items are combined into one large cluster.

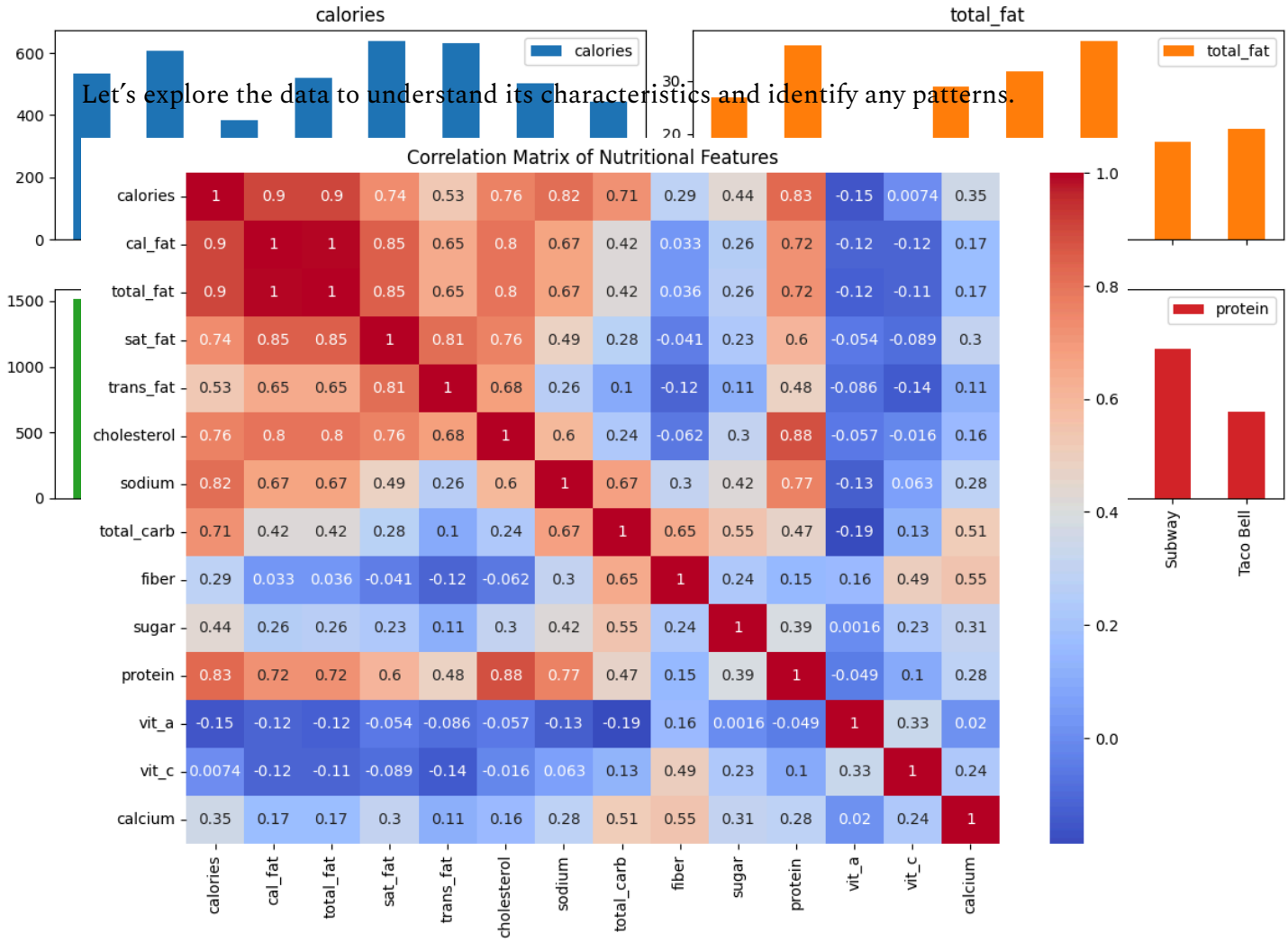
Data Information

The dataset comes from Kaggle and contains nutritional information for menu items from 8 popular fast food restaurants in the US. The key features include:

- ★ Macronutrients (calories, fat, protein, carbs)
- ★ Micronutrients (vitamins, calcium) – excluded due to 40% missing data

Data Visualization

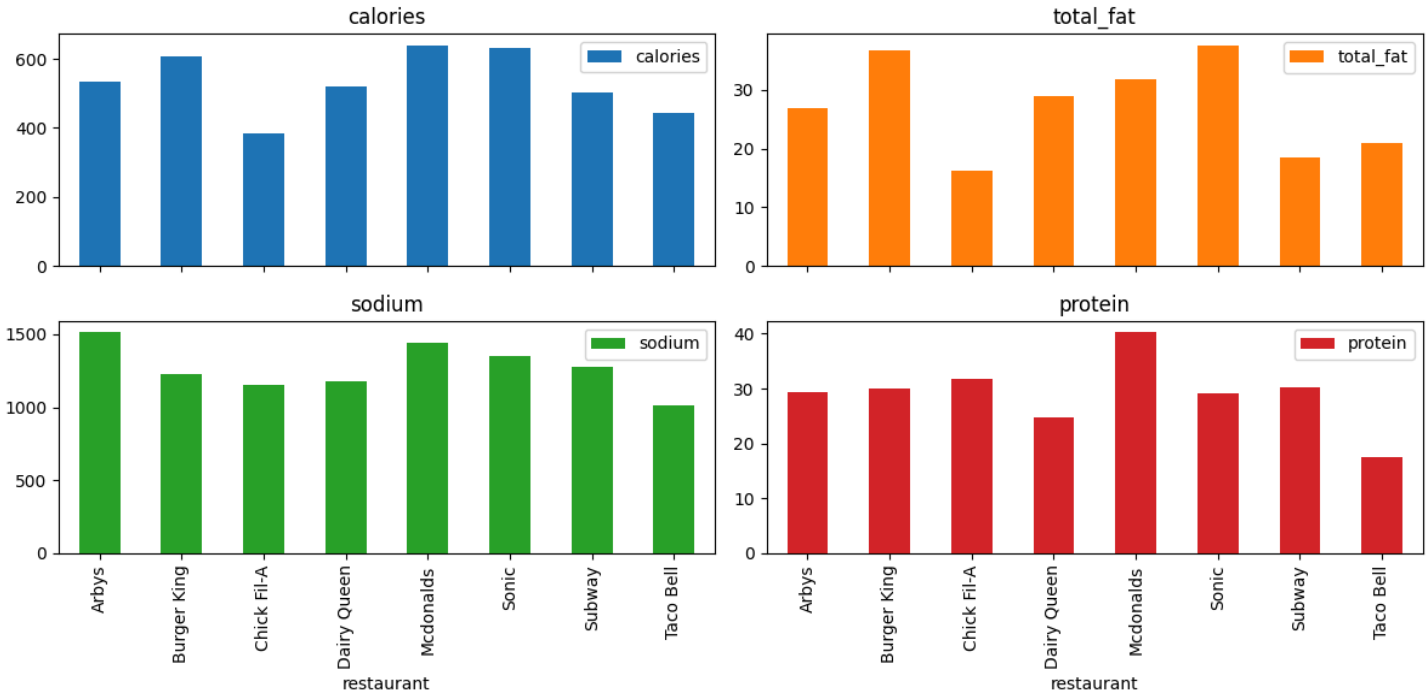
Let's explore the data to understand its characteristics and identify any patterns.



Made on hex with Python

I created a correlation matrix to reveal important relationships between nutritional features:

- ★ Strong positive correlations (0.7-0.9):
  - Calories show very strong correlations with fat-related metrics (cal_fat, total_fat, sat_fat)
  - Protein correlates strongly with calories (0.83) and cholesterol (0.88)
  - Sodium correlates strongly with calories (0.82)
- ★ Weak correlations
  - Fiber and vitamins show minimal ties to calories, justifying their exclusion.



Made on hex with Python

The bar graphs show:

- ★ Burger King, McDonald's, and Sonic tend to have higher calorie content
- ★ Subway and Chick-fil-A have more moderate nutritional profiles
- ★ Chick-fil-a and Subway have lower protein content on average, which is surprising because they are known for using fresh ingredients

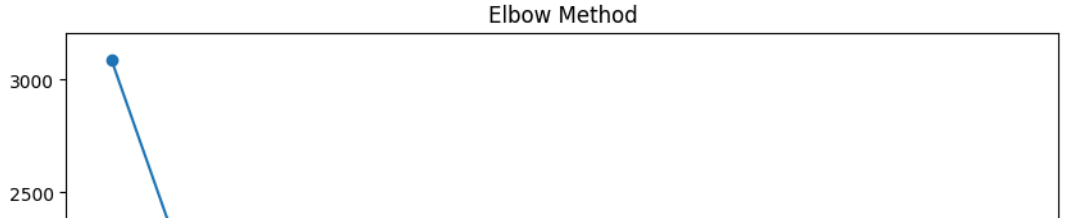
Data Preprocessing

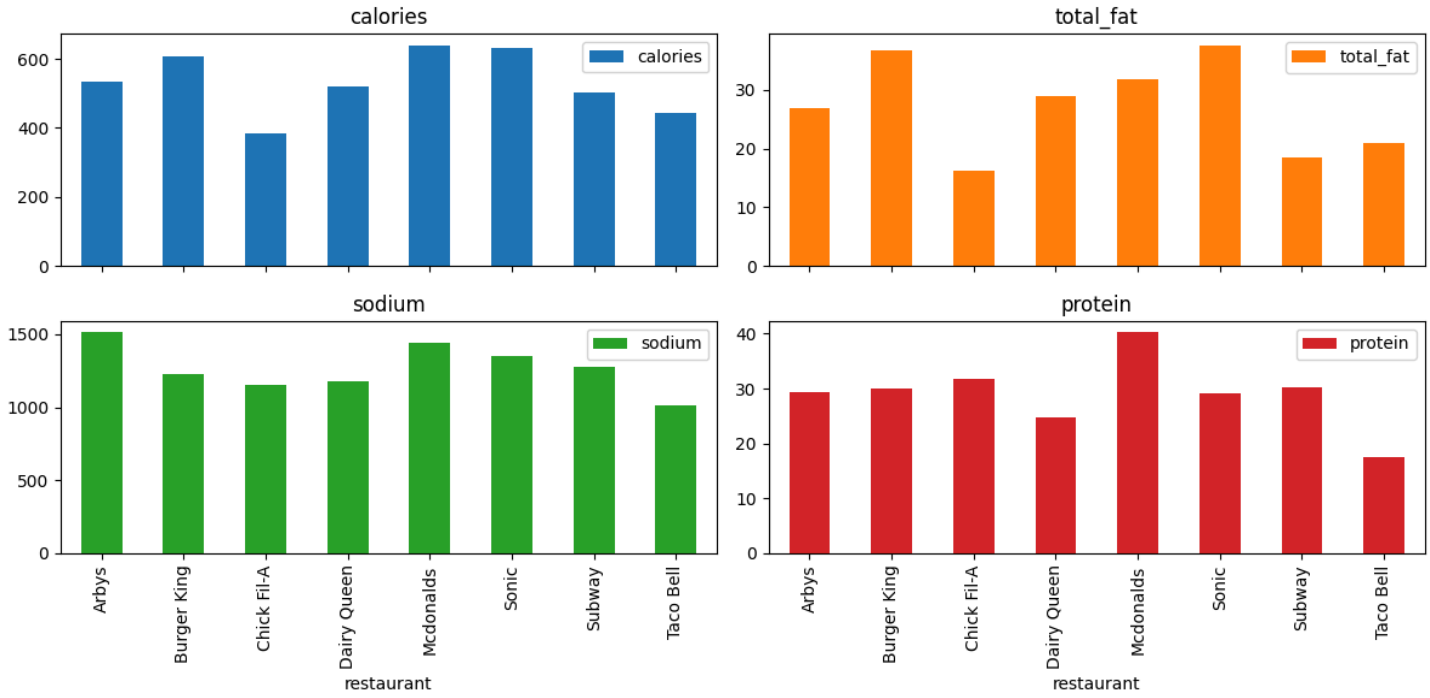
Before clustering, I took these pre-processing steps to prepare the data:

- ★ Dropped vit_a, vit_c, calcium (too many missing values)
- ★ Scaled features (e.g., calories range 50–3000; fat 0–150g)
- ★ Encoded restaurant for analysis, but not clustering.

Modeling & Analysis

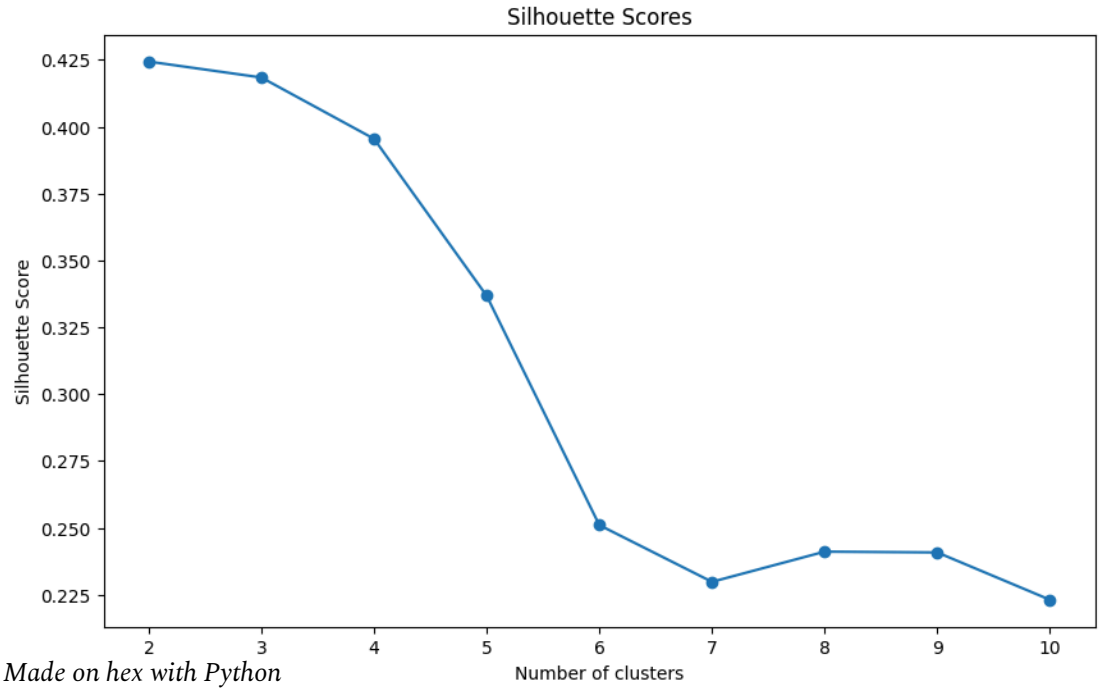
In this experiment, I used k-means and agglomerative clustering. First, I used the elbow method and silhouette score to determine the best k for k-means.



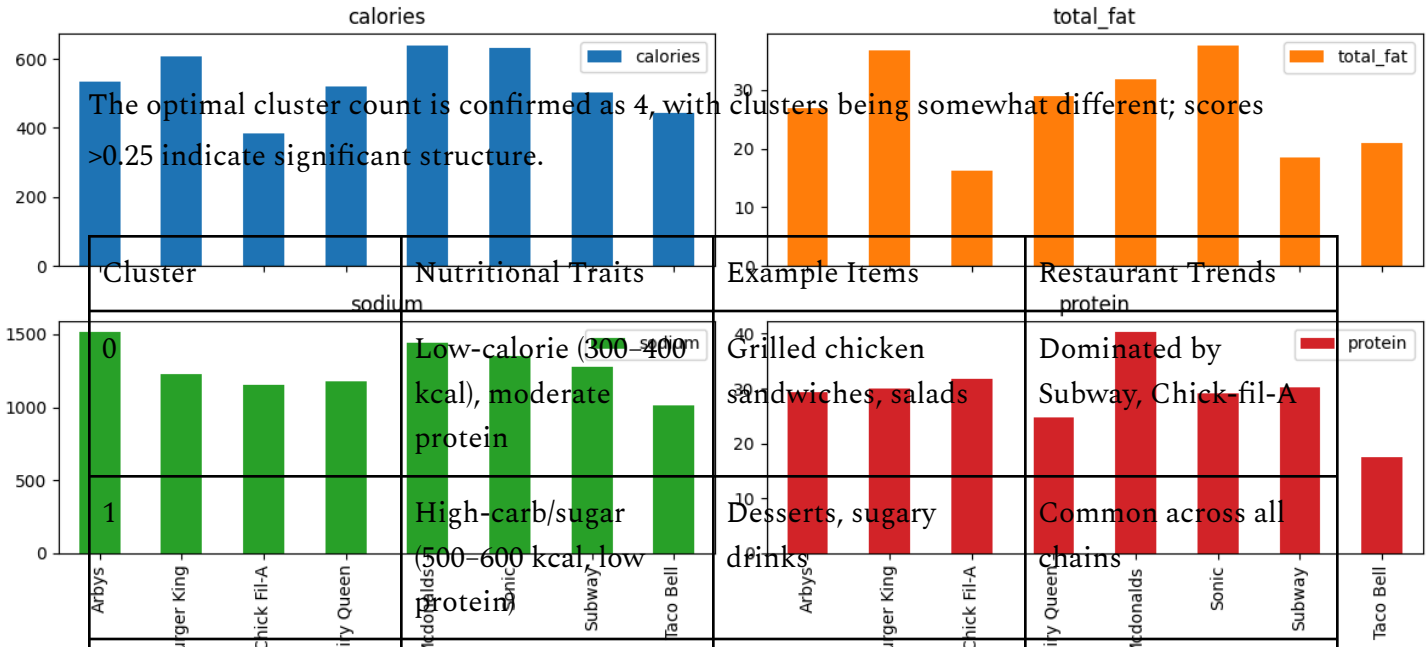


Made on hex with Python

We see a sharp decrease in WCSS (within-cluster sum of squares) from k=1 to k=4, gradual flattening after k=4, and the "elbow" at k=3 or k=4. This indicates 3 or 4 clusters would optimally balance model complexity with explanatory power. I chose k=4 to capture more nuanced groupings.



Made on hex with Python



The optimal cluster count is confirmed as 4, with clusters being somewhat different; scores >0.25 indicate significant structure.

Cluster	Nutritional Traits	Example Items	Restaurant Trends
0	Low-calorie (300–400 kcal), moderate protein	Grilled chicken sandwiches, salads	Dominated by Subway, Chick-fil-A
1	High-carb/sugar (500–600 kcal, low protein)	Desserts, sugary drinks	Common across all chains
2	High-protein (600–800 kcal), moderate fat	Burgers, meat-centric meals	Arby's, Sonic
3	Extreme indulgence (>1000 kcal, high fat/sodium)	Double cheeseburgers, large combos	Burger King, Dairy Queen

Cluster table insights:

- ★ Clear separation in calorie-fat-protein space.
- ★ Overlap between Clusters 1 (high-carb) and 2 (high-protein) for moderate-calorie items.

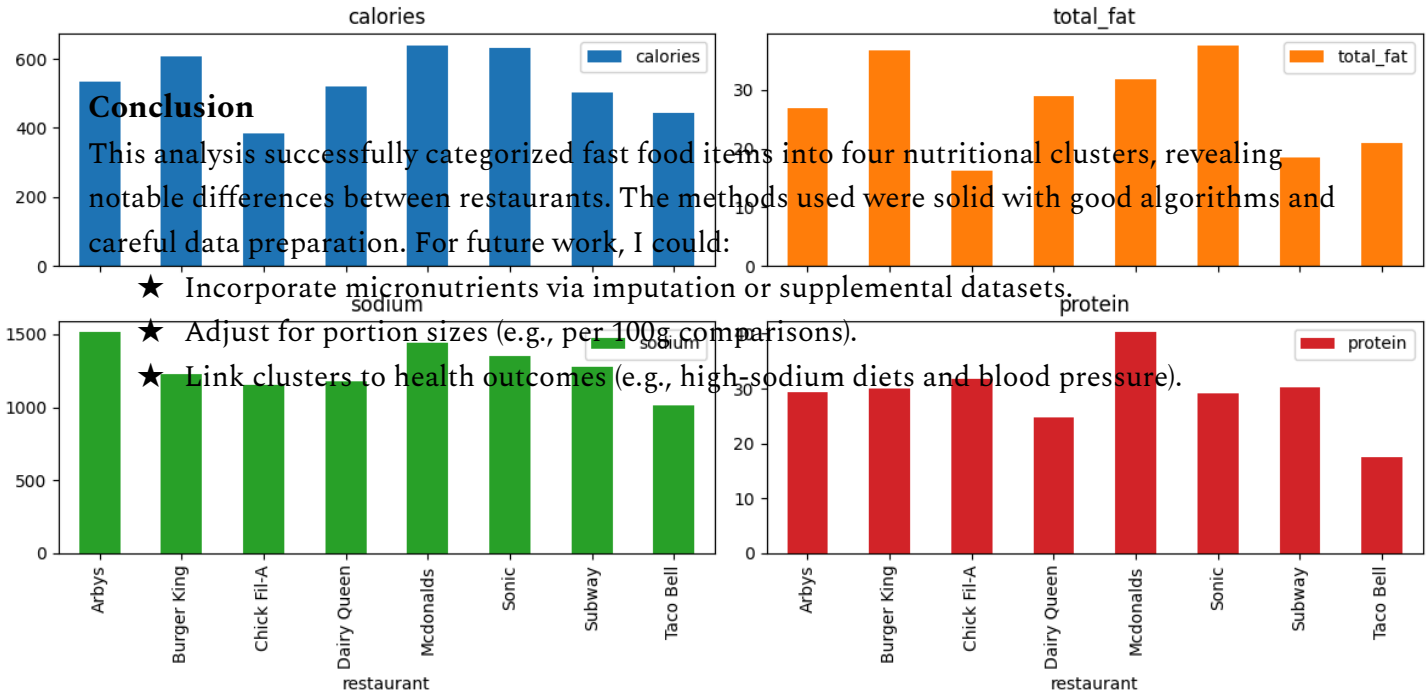
Impact

Key Insights

- ★ Subway and Chick-fil-A offer 3x more "healthy" (Cluster 0) items than Burger King.
- ★ 80% of Burger King's menu falls into Clusters 2–3 (high-calorie/indulgent).
- ★ Some salads (e.g., crispy chicken Caesar) cluster with high-carb meals due to dressings/croutons.
- ★ "Grilled" items sometimes rival burgers in sodium (e.g., grilled chicken sandwiches at Sonic).

The project has both positive and negative potential impacts:

- ★ Positive: Empowers consumers with data-driven comparisons.
- ★ Negative: Oversimplifies "health" (e.g., high-protein ≠ nutritious if loaded with sodium), could unfairly stigmatize certain restaurants without context (e.g., BK also offers salads).



Conclusion

This analysis successfully categorized fast food items into four nutritional clusters, revealing notable differences between restaurants. The methods used were solid with good algorithms and careful data preparation. For future work, I could:

- ★ Incorporate micronutrients via imputation or supplemental datasets.
- ★ Adjust for portion sizes (e.g., per 100g comparisons).
- ★ Link clusters to health outcomes (e.g., high-sodium diets and blood pressure).