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.

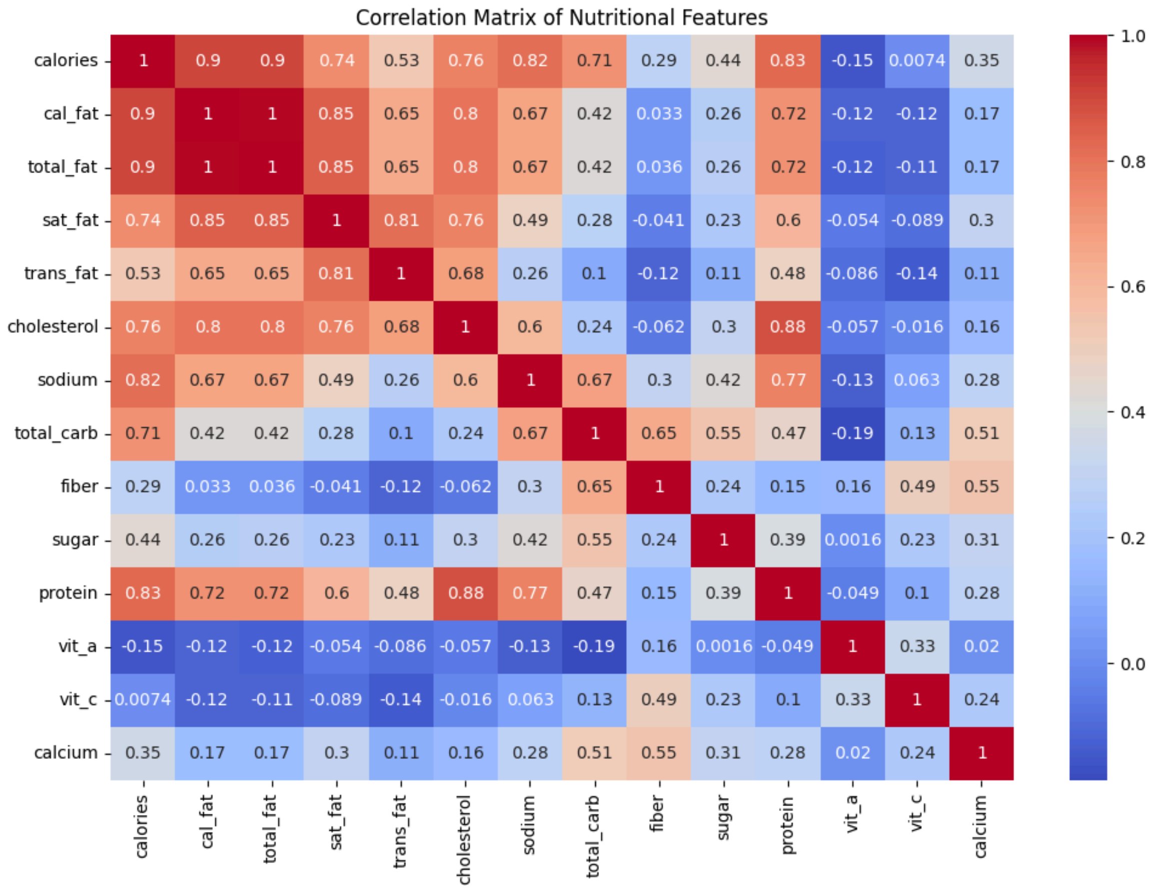
Data Information

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.

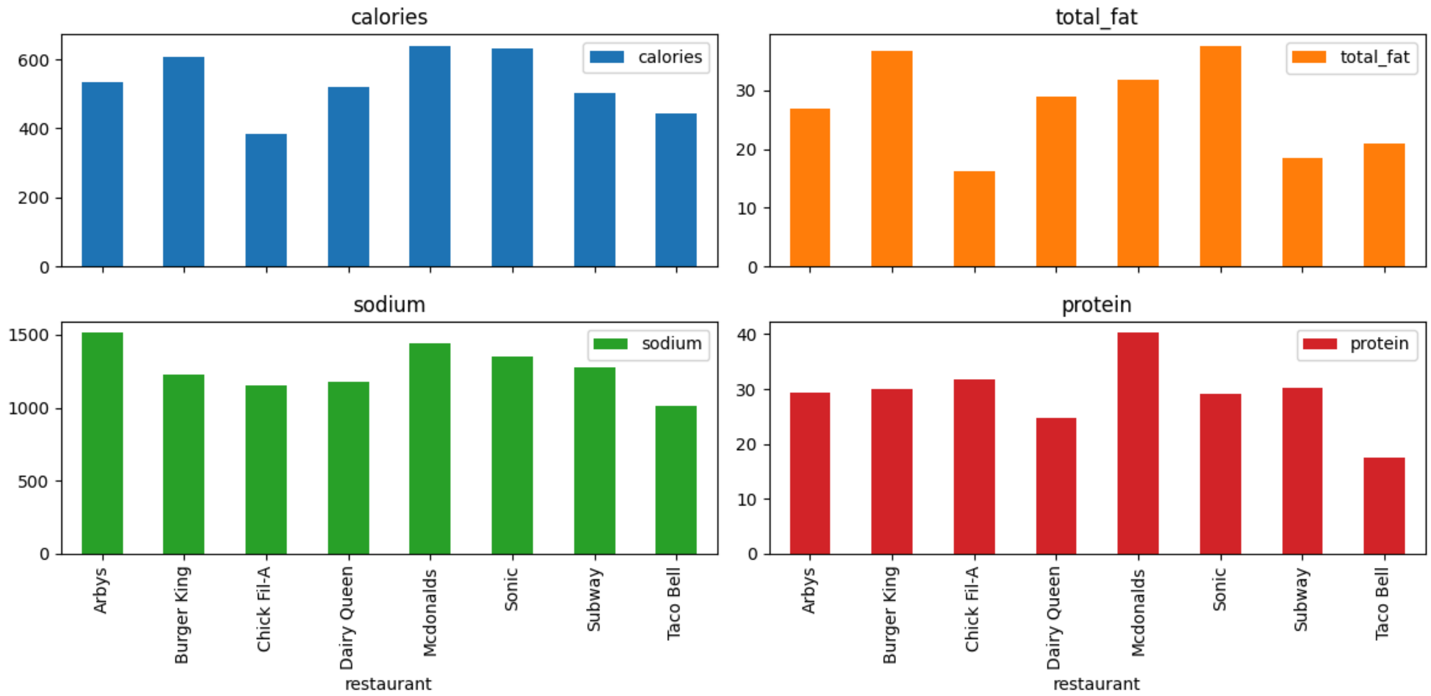


	calories	cal_fat	total_fat	sat_fat	trans_fat	cholesterol	sodium	total_carb	fiber	sugar	protein	vit_a	vit_c	calcium
calories	1	0.9	0.9	0.74	0.53	0.76	0.82	0.71	0.29	0.44	0.83	-0.15	0.0074	0.35
cal_fat	0.9	1	1	0.85	0.65	0.8	0.67	0.42	0.033	0.26	0.72	-0.12	-0.12	0.17
total_fat	0.9	1	1	0.85	0.65	0.8	0.67	0.42	0.036	0.26	0.72	-0.12	-0.11	0.17
sat_fat	0.74	0.85	0.85	1	0.81	0.76	0.49	0.28	-0.041	0.23	0.6	-0.054	-0.089	0.3
trans_fat	0.53	0.65	0.65	0.81	1	0.68	0.26	0.1	-0.12	0.11	0.48	-0.086	-0.14	0.11
cholesterol	0.76	0.8	0.8	0.76	0.68	1	0.6	0.24	-0.062	0.3	0.88	-0.057	-0.016	0.16
sodium	0.82	0.67	0.67	0.49	0.26	0.6	1	0.67	0.3	0.42	0.77	-0.13	0.063	0.28
total_carb	0.71	0.42	0.42	0.28	0.1	0.24	0.67	1	0.65	0.55	0.47	-0.19	0.13	0.51
fiber	0.29	0.033	0.036	-0.041	-0.12	-0.062	0.3	0.65	1	0.24	0.15	0.16	0.49	0.55
sugar	0.44	0.26	0.26	0.23	0.11	0.3	0.42	0.55	0.24	1	0.39	0.0016	0.23	0.31
protein	0.83	0.72	0.72	0.6	0.48	0.88	0.77	0.47	0.15	0.39	1	-0.049	0.1	0.28
vit_a	-0.15	-0.12	-0.12	-0.054	-0.086	-0.057	-0.13	-0.19	0.16	0.0016	-0.049	1	0.33	0.02
vit_c	0.0074	-0.12	-0.11	-0.089	-0.14	-0.016	0.063	0.13	0.49	0.23	0.1	0.33	1	0.24
calcium	0.35	0.17	0.17	0.3	0.11	0.16	0.28	0.51	0.55	0.31	0.28	0.02	0.24	1

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

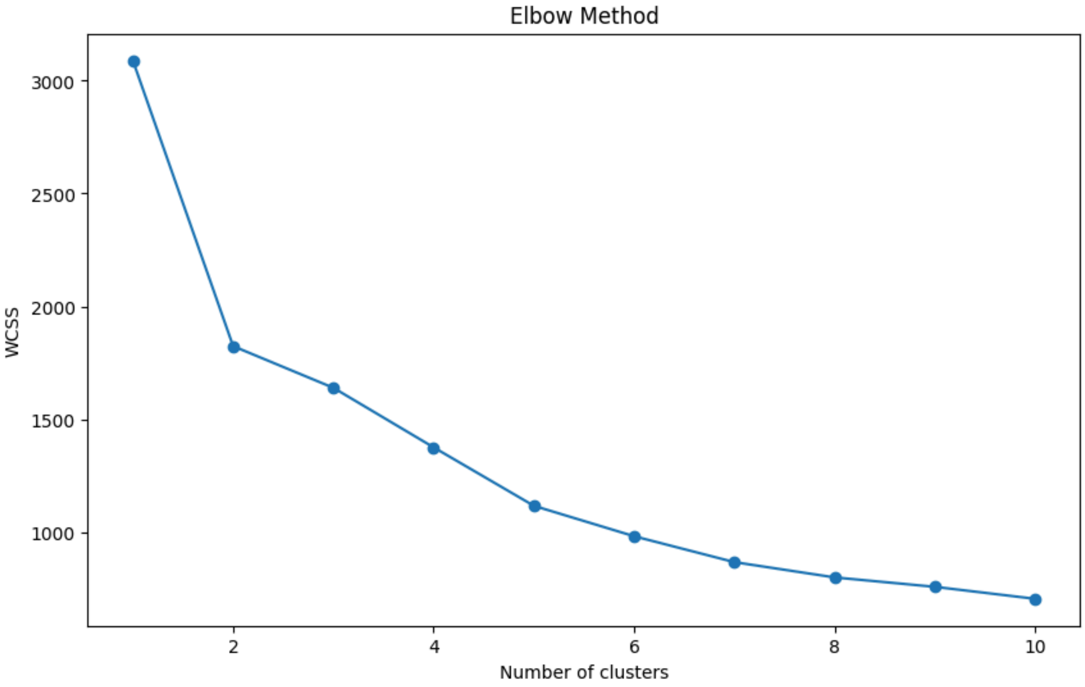
Data Preprocessing

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.

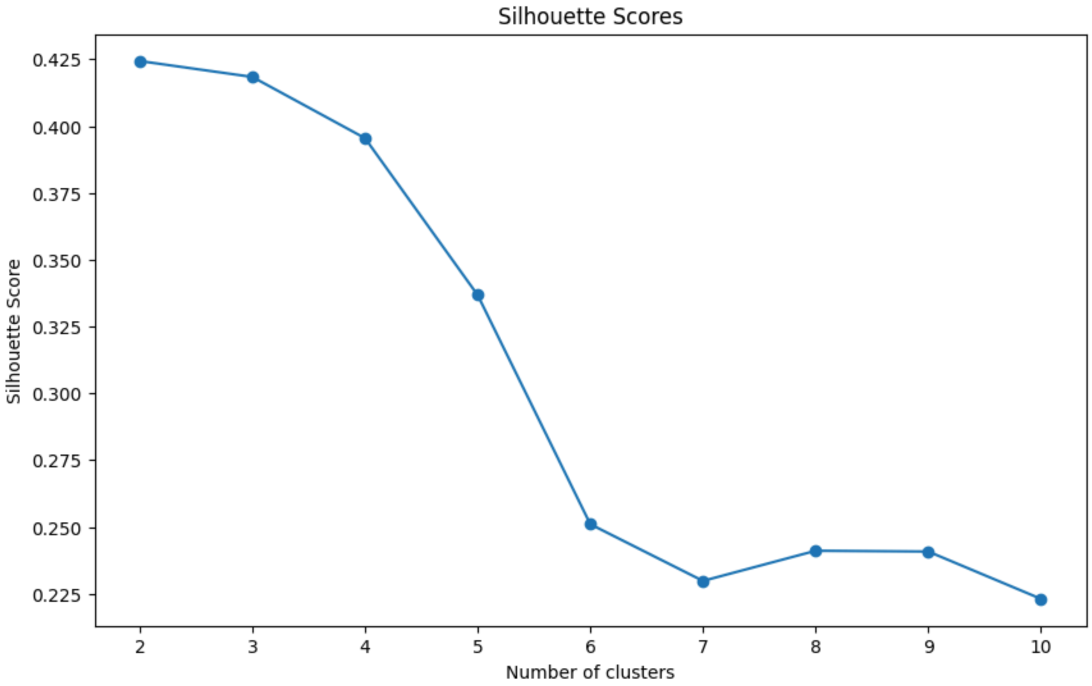
Modeling & Analysis

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.



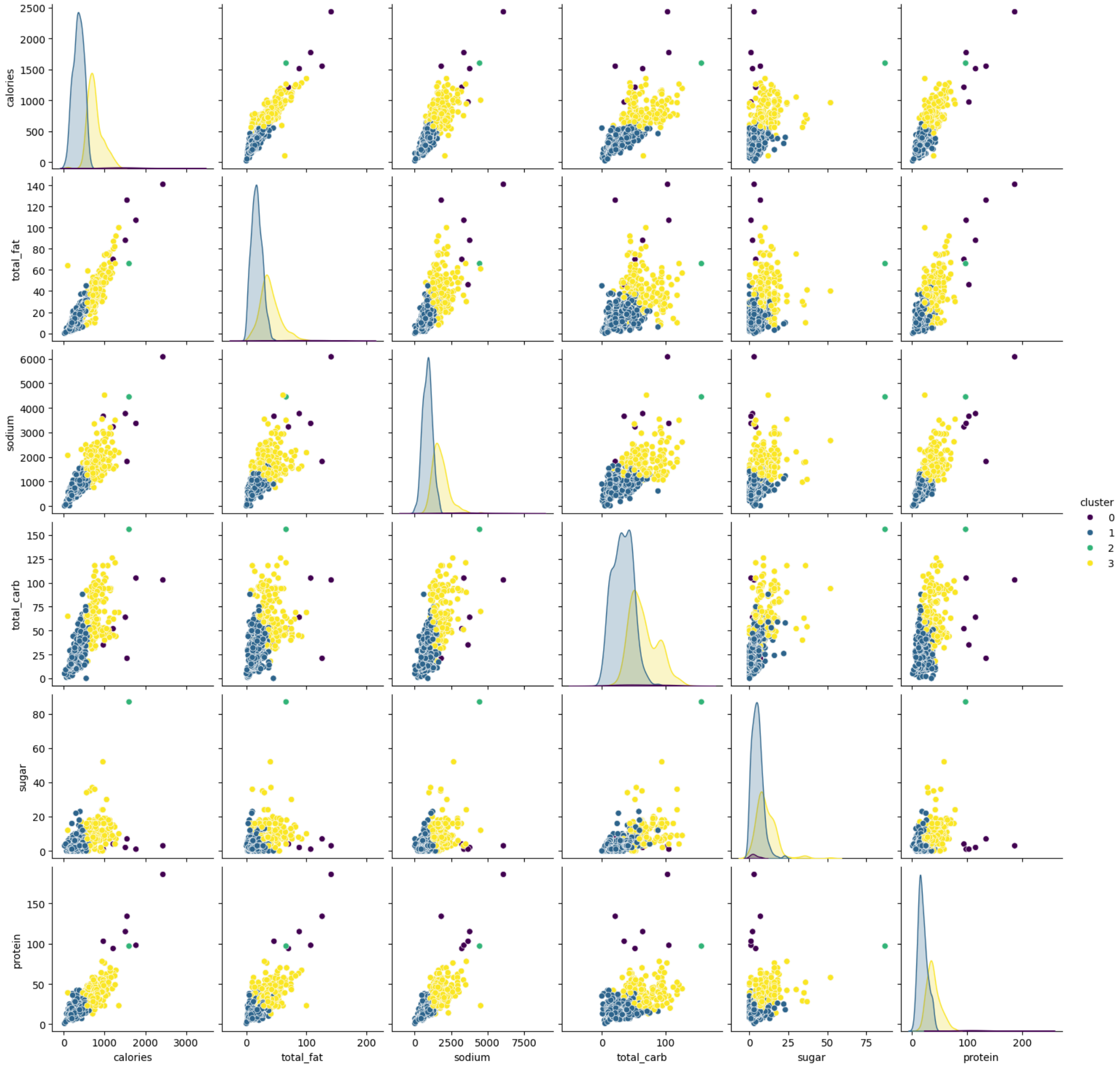
Made on hex with Python

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

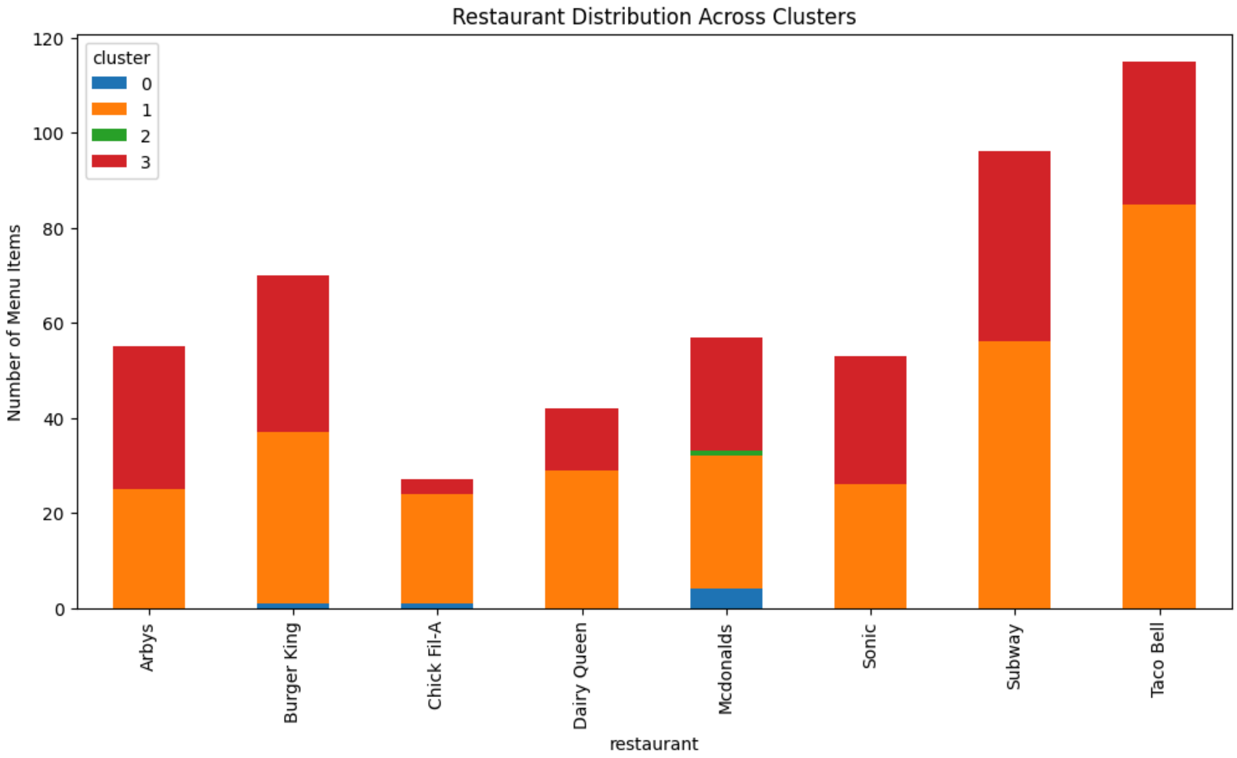
The optimal cluster count is confirmed as 4, with clusters being somewhat different; scores >0.25 indicate significant structure.



Made on hex with Python

The pairplot visualization shows different nutritional profiles across four clusters, clearly separating them by key macronutrients. Cluster 3 includes indulgent items with high calories (over 1000 kcal) and total fat (over 50g). An example is the Bacon King sandwich from Burger King. In contrast, Cluster 0 has healthier options with items under 500 calories and 20g of fat. Protein levels vary interestingly: Cluster 2 has high-protein items (30-70g), like Arby's roast beef sandwiches, which still have moderate calorie counts (500-800 kcal). Meanwhile, Cluster 1 features high-carb and low-protein items, such as McDonald's McFlurries, which form a separate group.

Notable outliers in sodium content appear in all clusters. Several items in Cluster 3 exceed 3000 mg of sodium, getting close to the FDA's daily limit, particularly in Sonic's burger offerings. This visualization shows how different macronutrient combinations form the clusters, even though some overlap exists in mid-range calorie items, where carbohydrate and protein ratios become distinguishing factors.



Made on hex with Python

The restaurant distribution analysis shows important details about nutritional choices. Burger King's menu mainly consists of indulgent items, with 60% of dishes in Cluster 3, such as the Whopper and Chicken Fries. Only 10% of their offerings are healthier options from Cluster 0.

Chick-fil-A takes a more balanced approach, with 40% of their menu featuring healthier items like grilled chicken and salads, and 30% in Cluster 2's high-protein category, which fits their focus on chicken.

Taco Bell's menu is different, with 70% of items in Cluster 1, which includes high-carb tortillas, and almost no high-protein choices. Dairy Queen's menu shows a mix of desserts in Cluster 1 and large burgers in Cluster 3, reflecting their dual role as a treat spot and burger place. This analysis highlights how different restaurant menus lead to various nutritional outcomes, with Chick-fil-A and Subway standing out for their lower-calorie options compared to other chains.

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).