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

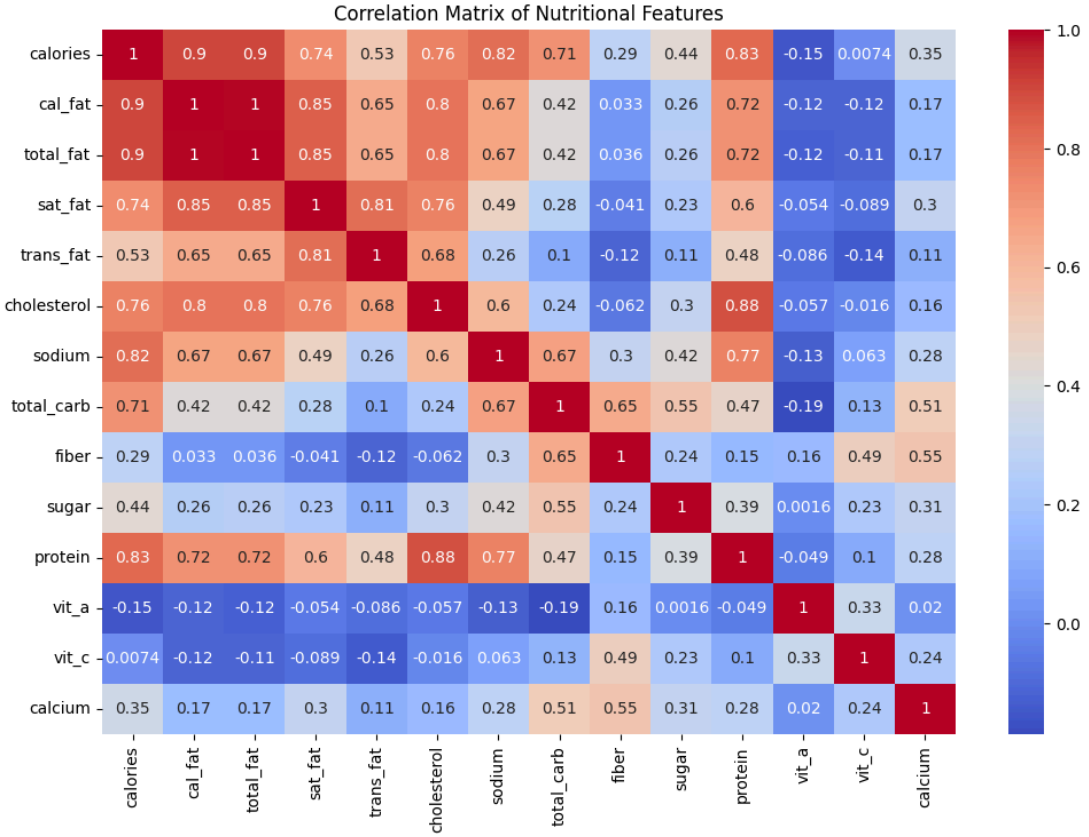
The dataset comes from Kaggle and contains nutritional information for menu items from 8 popular fast food restaurants in the US. The features include:

- ★ restaurant: *the fast food chain*
- ★ item: *the menu item name*

- ★ calories: *total energy content*
- ★ cal_fat: *calories from fat*
- ★ total_fat: *total fat in grams*
- ★ sat_fat: *saturated fat in grams*
- ★ trans_fat: *trans fat in grams*
- ★ cholesterol: *cholesterol in mg*
- ★ sodium: *sodium in mg*
- ★ total_carb: *total carbohydrates in grams*
- ★ fiber: *dietary fiber in grams*
- ★ sugar: *sugar in grams*
- ★ protein: *protein in grams*
- ★ vit_a, vit_c, calcium: *vitamin and mineral content*

Data Visualization

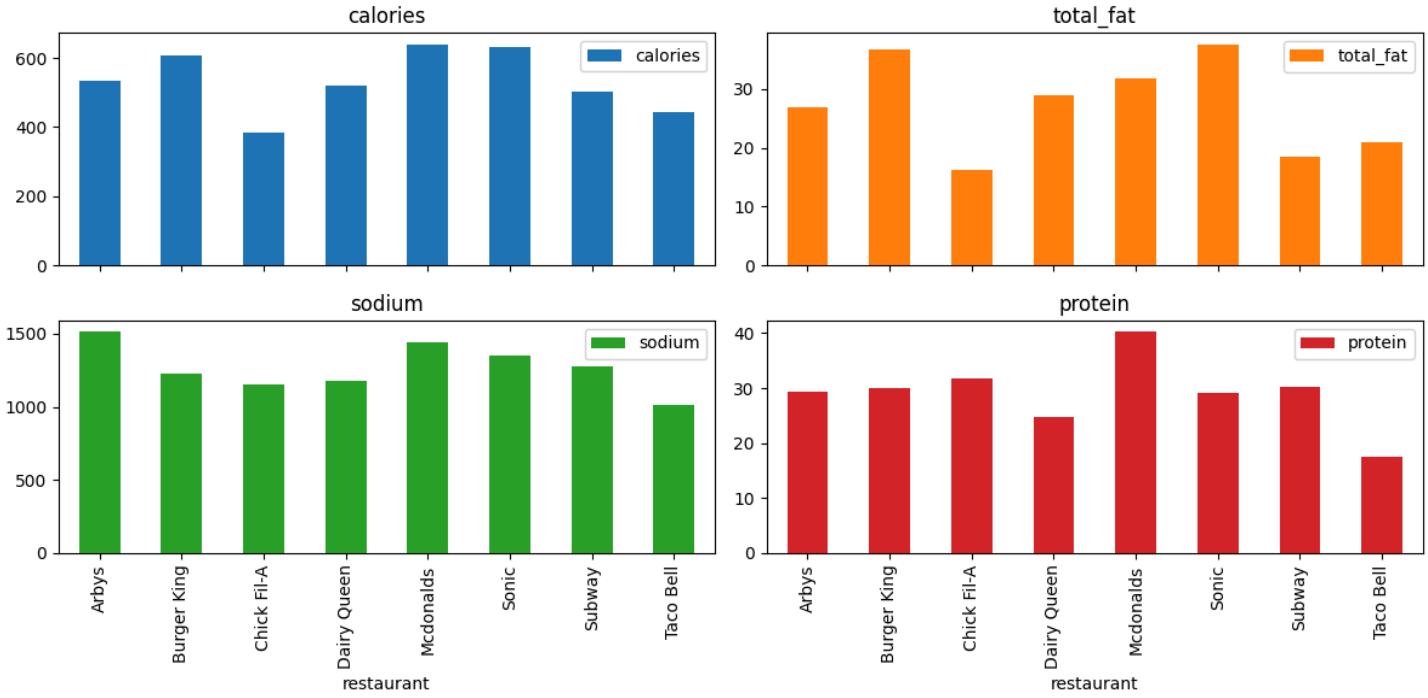
Let's explore the data to understand its characteristics and identify any patterns.



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)



Made on hex with Python

The bar graphs show:

- ★ Sonic and Burger King tend to have higher calorie and fat content
- ★ Subway and Chick-fil-A have more moderate nutritional profiles
- ★ Taco Bell has lower protein content on average

Data Preprocessing

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

- ★ Handling missing values: the dataset has a large number of missing values in the vit_a, vit_c, and calcium columns, so I removed them from the dataset
- ★ Removing irrelevant columns: I removed the “salad” column
- ★ Scaling: standardizing nutritional features because they are on different scales