Fast Food Nutrition Classification System

Larry Johnson

ABSTRACT

Fast food nutrition in the United States (US) is a topic of ongoing debate and concern, reflecting the integral role that fast food plays in American dietary habits. Characterized by high-calorie, high-fat, high-carbohydrates, low-fiber, and low-protein, fast food has been linked to various health issues, including obesity, cardiovascular disease, and diabetes. Despite its convenience and affordability, the nutritional quality of fast food is frequently criticized for its contribution to unhealthy eating patterns and poor health outcomes.

In response, there has been a growing push for transparency and healthier choices within the fast-food industry, leading to menu reforms, the inclusion of more nutritious options, and clearer nutritional labeling to help consumers make informed decisions. Nevertheless, the challenge remains in balancing consumer demand for quick, tasty, and affordable meals with the need for nutritional adequacy and healthier eating habits.

We will build a Fast-Food Nutrition (FFN) classification system that will classify fast-food menu items. Exclusively, utilizing the food nutritional contents of calories, fat, carbohydrates, fiber, and protein, we will build an objective computational model to categorize and then randomly classify future fast-food items based on these nutritional food contents.

1. Overview

The FFN system will be built over a series of milestones. Table 1.1 outlines these mid-level objectives. They start with the collection of data and performing exploratory data analysis (EDA) and ultimately ending with an unsupervised machine learning model.

Table 1.1 – FFN Classification Model Milestones

Milestone	Description
EDA	Process used in data analysis to investigate datasets to discover patterns, anomalies, or relationships without initially having specific hypotheses in mind. It involves using statistical summaries and visualizations to understand the data's underlying structure and characteristics. EDA is a critical step before more formal data analysis, allowing for informed model building and hypothesis testing.
Feature Engineering	Process of transforming raw data into features that better represent the underlying problem to predictive models, resulting in improved model accuracy on unseen data. This involves creating new features from existing data through domain knowledge, data manipulation, and algorithmic techniques to highlight important information for learning. Effective feature engineering can significantly impact the performance of machine learning models by providing them with more relevant information for making predictions.
Model Training	Process of teaching a machine learning algorithm to make predictions or decisions based on data. During training, the model iteratively adjusts its parameters to minimize the difference between its predictions and the actual data outcomes, using a specified learning algorithm and loss function. This process continues until the model achieves a desired level of accuracy on the training data, preparing it to make predictions on new, unseen data.
Model Validation	Practice of assessing a machine learning model's performance on a separate dataset not used during training, to evaluate its generalization ability to new data. This process typically involves using specific metrics, such as accuracy, precision, recall, or mean squared error, depending on the type of model and the problem it aims to solve. Validation helps ensure that the model performs well not just on the training data but also holds its predictive quality in real-world situations, thereby mitigating the risk of overfitting.
Model Prediction	Involves using a trained machine learning model to estimate the output or outcome for new, unseen data based on the patterns it learned during training. This step translates the model's understanding of the data into actionable insights or decisions, often in the form of predicted labels for classification problems or value estimates for regression tasks. The quality and reliability of model predictions depend on the

	model's accuracy, the relevance of the features used, and the model's generalization capability to new data.

2. Goals

The ultimate goal of the project is to build a Fast-Food Nutrition (FFN) classification system that will classify the fast-food menu item. Machine Learning is the best approach to solving this kind of problem due to various reasons including, but are not limited to:

- **Discovering Natural Groupings:** Unsupervised learning algorithms, such as clustering, can help identify natural groupings or patterns in the data without needing predefined labels. This is useful in nutritional science, where the healthiness of food might not be binary or easily categorized without extensive domain knowledge. It allows the exploration of data to uncover relationships and groupings based on nutrient profiles that might not be immediately obvious.
- **Handling Complexity and Nuances:** The healthiness of food is a complex topic influenced by various factors beyond basic nutrients (like the presence of vitamins, minerals, and other bioactive compounds). Unsupervised learning doesn't require a simplified "healthy/unhealthy" label upfront, which means it can handle this complexity better, allowing for the discovery of nuanced relationships between different nutritional components.
- **Adaptability to Broad Definitions of Healthiness:** Different dietary needs and health goals mean that "healthiness" can vary widely among individuals. Unsupervised learning doesn't start with a rigid definition of what makes food healthy or unhealthy but instead can reveal different dimensions of healthiness based on nutrient content, catering to a broader range of dietary approaches.

However, there are also some potential drawbacks and/or things to consider when using this kind of modeling approach:

- **Subjectivity and Interpretation:** The patterns and groupings discovered by unsupervised learning need to be interpreted by medical professionals such as dieticians, nutritionists, or even primary case physicians, which introduces subjectivity into what is considered healthy or unhealthy.
- **Lack of Explicit Labels:** While unsupervised learning can identify patterns and groupings, it doesn't label these groups as "healthy" or "unhealthy" without further analysis and interpretation. This might require additional steps, like consulting a similar list of the aforementioned medical professionals.
- **Complementary Use of Supervised Learning:** In some cases, unsupervised learning might be a preliminary step to reduce dimensionality or to understand the dataset better before applying supervised learning techniques, which can explicitly classify foods based on healthiness using labels generated from the insights of unsupervised learning.

3. Approach

3.1 Dataset

There are a variety of publicly available datasets available concerning fast food nutrition. The website Kaggle, a data science learning and competition website has a plethora of datasets available for public research and consumptions like this project. Table 3.1 lists the initial list of datasets that are available that cover the data needed for the project that cover the food nutritional contents that are in scope to define the model. Data sources could be modified based on the research and analysis necessary to complete the project objectives.

Table 3.1 – Dataset Data Sources

Dataset	Description
Starbucks	Includes the nutritional information for Starbucks' food and drink menu items. All nutritional information for drinks is based on a 12 oz serving size.
McDonalds	Provides a nutrition analysis of every menu item on the US McDonald's menu, including breakfast, beef burgers, chicken and fish sandwiches, fries, salads, soda, coffee and tea, milkshakes, and desserts.
Burger King	Comprehensive collection of nutritional information for all major menu items offered by Burger King. The dataset includes information on the number of calories, total fat, saturated fat, trans fat, cholesterol, sodium, total carbohydrates, and protein found in each menu item.
Wendy's	Comprehensive collection of nutritional information for all major menu items offered by Wendy's. The dataset includes information on the number of calories, total fat, saturated fat, trans fat, cholesterol, sodium, total carbohydrates, and protein found in each menu item.
Chick-Fila	Contains standard recipe-based nutrition and ingredient information, excluding customizations, and notes variability from handcrafting, serving size, and preparation differences.

3.1 EDA

Our dataset encompasses nutritional information on calories, fat, carbohydrates, fiber, and protein across a broad spectrum of food items, featuring five independent numerical variables across approximately 1,100 samples. Again, the variables are calories, fat, carbohydrates, fiber, and protein. Given the diversity of our data sources, normalization is a prerequisite to ensure comparability and reliability in our analyses.

To delve into the intricacies of our dataset and facilitate a comprehensive understanding of its characteristics, we propose an extensive Exploratory Data Analysis (EDA). This analysis will not only inform our subsequent modeling strategies but also provide valuable insights into the underlying patterns and distributions of

nutritional contents. The EDA will encompass the following key components for each nutritional variable:

1. **Scatter Plot:** To visualize the relationships between pairs of variables and identify potential correlations or outliers.

2. **Box Plot:** Both individual and composite plots will be employed to offer a concise summary of the distribution's central tendency and variability, highlighting outliers effectively.

3. **Histogram:** This will facilitate an understanding of the distribution shape of each variable, revealing skewness, kurtosis, and the presence of multiple modes if any.

4. **Normal Assessment:** Through graphical and statistical tests, we will evaluate the adherence of our data to a normal distribution, a critical consideration for many statistical analysis techniques.

5. **Parameters Estimation:** We aim to estimate key statistical parameters, including proportions, mean values, and standard deviations (SD), to succinctly describe each variable's central tendency and dispersion.

3.2 Feature Engineering

In our pursuit to refine and enhance the predictive capabilities of our dataset, we embark on a critical phase of feature engineering, specifically focusing on the application of data transformation techniques to each nutritional variable. The relevance of binarization and quantization, among other transformations, will be meticulously evaluated. These transformations are pivotal for aligning our dataset with the requisites of sophisticated analytical methodologies and for accentuating specific nutritional attributes within the dataset.

The selection of appropriate transformation techniques is a consequential decision, heavily informed by the insights gleaned from our Exploratory Data Analysis (EDA). This ensures that our feature engineering efforts are not only methodologically robust but are also custom fitted to the distinctive attributes of our dataset.

Considering the variance in distribution and scale across the nutritional variables, we may employ the Box-Cox transformation or log transformation to address skewness and normalize the data distribution. These transformations are particularly beneficial for variables that do not follow a normal distribution, thereby making the data more amenable to linear models and other statistical analyses that assume normality.

Additionally, our analysis will explore the utility of normalization and standardization techniques, such as min-max scaling, z-score standardization, and L2-normalization. These methods adjust the scales of our variables, rendering them directly comparable and removing potential bias due to differing units of measurement or scales. Min-max scaling, for instance, rescales the data into a fixed range, usually 0 to 1, which can be advantageous for algorithms that are sensitive to the scale of input data. Z-score standardization transforms the data to have a mean of 0 and a standard deviation of 1, aiding in the identification of outliers and improving the performance of models based on distance measures. L2-normalization, on the other hand, adjusts the values in a way that the sum of the squares is equal to 1, which is useful for comparing the similarity of samples.

Each transformation method will be judiciously considered and applied based on its compatibility with the unique characteristics of each nutritional content variable. Through this deliberate and informed approach to feature engineering, we aim to enhance the analytical readiness of our dataset, paving the way for deeper insights and more robust model performance.

There is not an industry standard on whether a food item is healthy or unhealthy in terms of food science. As a result, we will use an objective food computational model to quantitatively drive the classification of the various food items. This prerequisite step is part of feature engineering as we will be adding additional quantitative computed independent variables for each food nutrient and one composite independent variable based on the following model.

Here's a rough breakdown of how these factors might contribute to a food item, totaling up to 100%:

- High in Calories: 35%
- High in Fat: 30%
- High in Carbohydrates: 20%
- Low in Fiber: 10%
- Low in Protein: 5%

Here, CFIS represents the Composite Food Intake Score, calculated by multiplying the amounts of Calories (C), Fat (F), Carbohydrates (Ca), Fiber (Fi), and Protein (P) in a meal by their respective weights and summing these products. The use of subtraction for Fiber and Protein reflects their negative contribution to the unhealthiness score in this context.

Let's denote the amounts of Calories, Fat, Carbohydrates, Fiber, and Protein in a meal by *C*, *F*, *Ca*, *Fi*, and *P* respectively. Then, the meal score, which we can call the Composite Food Intake Score (*CFIS*), can be mathematically expressed as:

$$CFIS = C * 0.35 + F * 0.30 + Ca * 0.20 - Fi * 0.10 - P * 0.05$$

3.2 Modeling

As previously mentioned, it may be best to approach this problem using an unsupervised learning approach; more specifically clustering to identify the various types of fast-food menu items. Then we will use the new clusters to classify future fast food menu items. Furthermore, again this classification will *not* necessarily tell you what is healthy or unhealthy due to the variability of fast-food menu items.

More specifically, for the clustering, we will use the KMeans algorithm to identify the fast-food group menu items. KMeans clustering can be applied to identify hidden groups within fast food items based on nutritional scores, including caloric_intake_score, fat_intake_score, carbohydrates_intake_score, fiber_intake_score, protein_intake_score, and a composite_intake_score. This process involves several steps, tailored to uncover patterns in how different fast food items can be categorized based on their nutritional content.

Modified Mathematical Algorithm for Fast Food Clustering:

1. **Initialization:** Start by selecting **k** initial centroids randomly from the dataset of fast-food items. These centroids represent the initial guess for the centers of the nutritional score clusters.

2. **Assignment Step:** Each fast-food item is then assigned to the nearest centroid based on its nutritional scores. The distance (d) between a fast food item with its nutritional scores x_i and a centroid c_j could be

calculated using a multidimensional distance metric, like Euclidean distance, considering all nutritional scores:
$d(x_i, c_j) =$
$\sqrt{((caloric_intake_score_i - caloric_intake_score_j)^2 + composite_intake_score_j)^2)}$

3. **Update Step:** Recompute the centroid of each cluster to be the mean of the nutritional scores of all fast food items in that cluster. The new position of each centroid is determined by averaging the scores of all items assigned to the cluster.

$$c_j = \frac{1}{|S_j|} \sum_{x_i \in S_j}^{x_i}$$

4. **Convergence Check:** The assignment and update steps are repeated until the centroids' positions stabilize and no longer change significantly, indicating that the clusters have been successfully identified.

Objective Function for Nutritional Score Clustering:
The aim is to minimize the within-cluster sum of squares (WCSS) across all nutritional scores, effectively grouping fast food items into clusters that minimize the variance within each cluster:

$$J = \sum_{J=1}^{K} \sum_{x_i \in S_j} \|x_i - c_j\|^2$$

Application:

This adapted KMeans algorithm can reveal hidden groups within the fast food dataset based on their nutritional profiles. For instance, it may identify clusters of items that are high in protein but low in carbohydrates, or items that have a balanced nutritional profile according to the composite intake score. These clusters can help consumers make informed choices, nutritionists to recommend healthier options, or fast food companies to identify gaps in their product offerings.

Considerations:

- **Selection of k:** Determining the number of clusters, **k**, is crucial and can be informed by domain knowledge or by using methods like the elbow method, applied to the composite nutritional scores.

- **Nutritional Score Interpretation:** The interpretation of clusters will rely on understanding the nutritional needs and how each score (caloric, fat, carbohydrates, fiber, and protein) contributes to the overall healthiness of fast food items.

This approach provides a structured method to uncover patterns in fast food nutrition that might not be apparent through simple analysis, offering valuable insights into the dietary quality of fast food menu items.

4. EDA

As previously mentioned there are approximately 1,100 samples across the various fast food restaurants. After removing duplicates there are 1,096 menu item samples. Using various types of visualization techniques, we will show the menu item distribution in order to illustrate our EDA activities. Figures 4.1-4.4 show various types of visualizations using box plots, scatterplots, and bar histograms

Figure 4.1 - Fast Food Menu Item Paired Scatterplots

Scatter Plot-Histograms (Diagonal Charts):

Calories: The histogram for calories shows the distribution of calorie counts across the dataset. If the bars are skewed to the left, with a long tail to the right, it indicates that most menu items have a lower calorie count, with a few items having very high calories.

Fat: The fat histogram will show how fat content is distributed among the menu items. A similar left skew could indicate that most items are low in fat, with fewer items being high in fat.

Carbohydrates: The carbohydrates histogram will provide insight into the carb content of the items. If the distribution is more uniform or normal (bell-shaped), it would suggest a more balanced spread of carbohydrate values across the menu items.

Fiber: The fiber histogram can be expected to show that many items have low fiber content, which is common in fast-food items. A peak at the lower end of the scale would confirm this.

Protein: The protein histogram might show a different pattern if there is a variety of menu items, some with low protein content (like drinks or desserts) and others with high protein content (like meat-based items).

Scatterplots (Off-Diagonal Charts):

Calories vs. Other Nutrients: Scatter Plots comparing calories with fat, carbohydrates, fiber, and protein can reveal which of these nutrients contribute more to the calorie content. A positive correlation would show that as one nutrient increases, the calorie content tends to increase as well. For example, a rising trend in the scatterplot of calories vs. fat would indicate that higher-fat items tend to be higher in calories.

Fat vs. Carbohydrates, Fiber, and Protein: These scatterplots would show whether there's a relationship between the fat content and other nutrients. We might not expect a strong correlation between fat and fiber, but there could be a positive relationship between fat and protein if higher-fat items also contain more protein (like cheese or meat).

Carbohydrates vs. Fiber and Protein: These scatterplots can indicate whether higher-carb items tend to have more fiber or protein. For instance, whole-grain items might have both high carbohydrates and fiber.

Fiber vs. Protein: This scatterplot might show a less clear relationship since high-fiber items are not necessarily high in protein and vice versa.
Patterns and Clusters:

If certain scatterplots show clusters, this could indicate that there are distinct categories or groups of menu items, such as desserts (high carb, low protein), salads (low carb, high fiber), or meats (high protein, high fat).

Outliers in these plots can indicate special menu items that are exceptionally high or low in certain nutrients compared to the rest.

Figure 4.2 - Fast Food Menu Item Individual BoxPlot

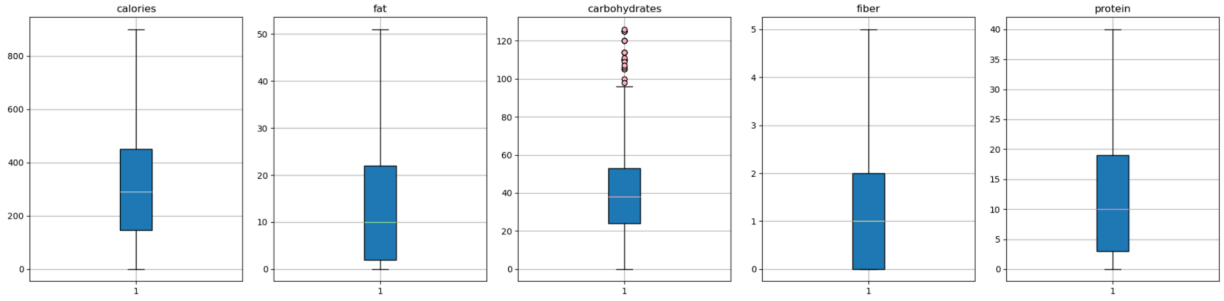


Figure 4.3 - Fast Food Menu Item Combined BoxPlot

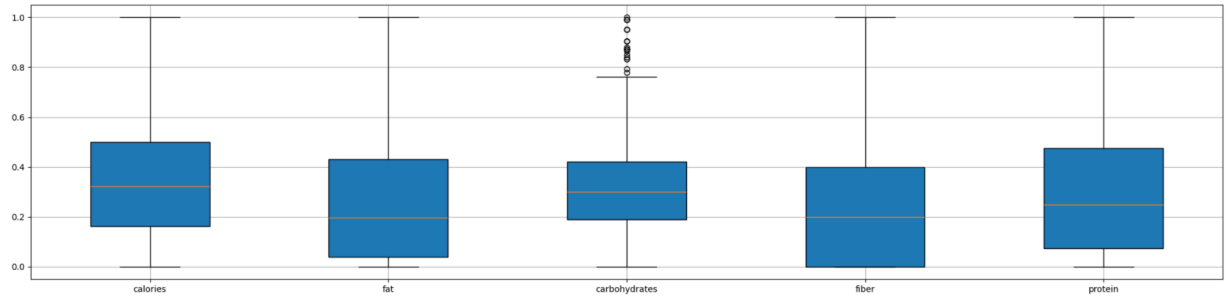
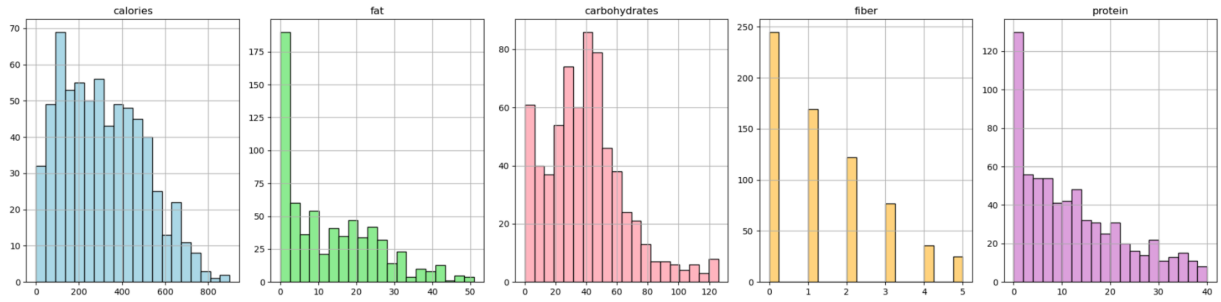


Figure 4.4 - Fast Food Menu Item Individual Histogram



Box Plots

These boxplots help in quickly identifying which nutrients vary the most across menu items and which ones have extreme values. They also provide a snapshot of the nutritional profile of the menu, which can be useful for health analysis, menu planning, or consumer information. Note, for the combined boxplot numerical explanation see section *Normal Assessment* for details in the min-max scaling that was applied.

Calories: The boxplot for calories shows the median, interquartile range (IQR), and potential outliers for calorie content across menu items. The median is indicated by the line within the box, and it seems to be on the lower end of the range, suggesting a right-skewed distribution. Outliers are represented by individual points that fall outside of the 'whiskers'. These points indicate menu items with unusually high calorie counts.

Fat: The boxplot for fat also shows a right-skewed distribution, with a lower median relative to the range of data. There are several outliers, indicating some menu items have a much higher fat content than the majority.

Carbohydrates: The carbohydrates box plot seems less skewed than the calories and fat, but still shows some right skewness. The presence of outliers on the upper range suggests that there are a few menu items with very high carbohydrate content.

Fiber: Fiber content tends to be low across most items, as shown by the concentration of the box near the bottom of the plot.

There are outliers indicating some menu items are particularly high in fiber, which could be items like salads or whole-grain-based products.
Protein:

The protein boxplot has a similar distribution to fiber, with most items having lower protein content and a few outliers that are high in protein. These could be meat or legume-based items.

Histograms:

This set of histograms gives an overview of the nutritional content distribution among the menu items. From a health or dietary perspective, these histograms can help identify how the menu aligns with nutritional guidelines or targets, which could be particularly useful for dietary planning or menu design.

Calories: The histogram for calories shows that the vast majority of menu items have a relatively low calorie count, with the frequency decreasing as the calorie count increases. There are very few items with extremely high calories, which appear as outliers.
The scale for calories is quite large, indicating a wide range of calorie counts among the items.

Fat: The fat content histogram also shows a decrease in frequency as fat content increases, suggesting that most menu items have lower fat content, with fewer high-fat options. Similar to calories, there are a few items with exceptionally high fat content, indicating outliers.
Carbohydrates:

Carbohydrates: follow a pattern similar to fat, with most items being on the lower end of the carbohydrate scale. The frequency of items decreases as carbohydrate content increases. This histogram shows that high-carbohydrate items are less common in the dataset.

Fiber: The fiber content histogram demonstrates that the majority of items contain very little fiber, with a steep drop-off in frequency as fiber content increases. This suggests that fiber-rich menu items are relatively rare in this dataset.

Protein: The histogram for protein indicates that most items have a low to moderate protein content, with very few items containing high levels of protein.

Normal Assessment

In order to perform a normality assessment, we will perform this process in two steps.

1. Performing min-max scaling on the entire menu dataset adjusts all the numerical values of the various nutrients to fall within a new range, typically 0 to 1. This scaling is done for each feature (column) independently and can be particularly useful for machine learning models that are sensitive to the magnitude of input features.

   The formula for min-max scaling of a value x in a feature X is:

$$x_{scaled} = \frac{(x - min(X)}{max(X) - min(X)}$$

   where:
   - x_{scaled} is the scaled value.
   - x is the original value.
   - $min(X)$ is the minimum value in the feature X.
   - $max(X)$ is the maximum value in the feature X.
2. Then we quantize each nutrient in deciles in order to normalize the data to perform a normality assessment.

The results of the normality assessment are shown in Figure 4.3. In Figure 4.3, we can still see that even after scaling the data mostly follows a distribution that is skewed to the right.

Parameters Estimation

Since we have 1,000 samples we can safely assume that these menu items are *not* representative of the entire fast food restaurants industry in the U.S. or the world for that matter. As a result, we will attempt to evaluate the following claims of the calories, fat, carbohydrates, fiber, and protein nutrients of the fast food industry through the following:

1. Make a claim about the population proportion
2. Make a claim about the estimate of the population mean of each nutrient.
3. Make a claim about the two population means of each nutrient

Note, all of the claims below are based on the following criteria:
- A level of significance of 0.05
- N or the number of unique samples are 100.
- H0 = p = 0.50 or 50%
- H1 = p != 0.50 or 50%

After duplicates and data was cleaned were approximately 674 fast food items that make up our population. Table 2 houses the summary statistics for the research.

Table 4.1 - Fast Food Menu Items - Summary Statistics

	calories	fat	carbohydrates	fiber	protein
count	674.000000	674.000000	674.000000	674.000000	674.000000
mean	309.023739	13.223294	39.881306	1.354599	11.944807
std	189.894855	12.271499	25.292068	1.397435	10.540781
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	146.250000	2.000000	24.000000	0.000000	3.000000
50%	290.000000	10.000000	38.000000	1.000000	10.000000
75%	450.000000	22.000000	53.000000	2.000000	19.000000
max	900.000000	51.000000	126.000000	5.000000	40.000000

Population Proportion Claim(s):

Let's make a claims that:
- 50% of the fast foods in the U.S. have exactly 277 calories.
- 50% of the fast foods in the U.S. have exactly 12 grams of fat.
- 50% of the fast foods in the U.S. have exactly 22 carbohydrates.
- 50% of the fast foods in the U.S. have exactly 1.5 grams of fiber.
- 50% of the fast foods in the U.S. have exactly 10 grams of protein.

Table 4.2 – Population Proportion Claims

nutrient	claim	population_proportion_claim	population_proportion_claim_percentage	sample_proportion_claim	sampleproportion_claim_percentage	p0	alpha	z	p_value	critical_value_test	min	max	confidence	confidence_interval_test
calories	277.0	354	52.52	45	45.0	0.5	0.05	-1.0	0.317	do_not_reject_h_0	0.352	0.548	95%	do_not_reject_h_0
fat	12.0	322	47.77	47	47.0	0.5	0.05	-0.6	0.549	do_not_reject_h_0	0.372	0.568	95%	do_not_reject_h_0
carbohydrates	22.0	521	77.3	78	78.0	0.5	0.05	5.6	0.0	reject_h_0	0.699	0.861	95%	reject_h_0
fiber	1.5	260	38.58	38	38.0	0.5	0.05	-2.4	0.016	reject_h_0	0.285	0.475	95%	reject_h_0
protein	10.0	339	50.3	39	39.0	0.5	0.05	-2.2	0.028	reject_h_0	0.294	0.486	95%	reject_h_0

There is sufficient evidence that exactly 50% of the fast food menu items have 277 calories and 12 grams of fat. However, there is insufficient evidence that exactly 50% of fast food items have 22 carbohydrates, 1.5 grams of fiber, and 10 grams of protein.

Population Mean Claim(s):

Let's make a claims that:

- 50% of the fast foods in the U.S. have exactly a mean of 277 calories.
- 50% of the fast foods in the U.S. have exactly a mean of 12 grams of fat.
- 50% of the fast foods in the U.S. have exactly a mean of 22 carbohydrates.
- 50% of the fast foods in the U.S. have exactly a mean of 1.5 grams of fiber.
- 50% of the fast foods in the U.S. have exactly a mean of 10 grams of protein.

Table 4.3 – Population Mean Claims

nutrient	claim	population_mean	sample_mean	p0	alpha	z	p_value	critical_value_test	min	max	confidence	confidence_interval_test
calories	277.0	309.02373887240400	277.29	0.5	0.05	-1.67	0.095	do_not_reject_h_0	240.071	314.509	95%	reject_h_0
fat	12.0	13.223293768546000	11.77	0.5	0.05	-1.18	0.236	do_not_reject_h_0	9.365	14.175	95%	reject_h_0
carbohydrates	22.0	39.881305637982200	37.91	0.5	0.05	-0.78	0.436	do_not_reject_h_0	32.953	42.867	95%	reject_h_0
fiber	1.5	1.3545994065281900	1.4	0.5	0.05	0.32	0.745	do_not_reject_h_0	1.126	1.674	95%	reject_h_0
protein	10.0	11.944807121661700	10.065	0.5	0.05	-1.78	0.075	do_not_reject_h_0	7.999	12.131	95%	reject_h_0

There is sufficient evidence that the average fast food menu item has exactly 277 calories, 12 grams of fat, 22 carbohydrates, 1.5 grams of fiber, and 10 grams of protein.

Two Population Means Claim(s):

Are there differences between the average number of calories, fat, carbohydrates, fiber, and protein between random populations of fast food items?

Table 4.4 – Two Population Means Claims

nutrient	claim	first_sample_mean	first_sample_mean	p0	alpha	z	p_value	critical_value_test	min	max	confidence	confidence_interval_test
calories	7.41	277.29	284.700	0	0.05	-0.29	0.769	do_not_reject_h_0	-57.016	42.196	95%	do_not_reject_h_0
fat	0.01	11.77	11.762	0	0.05	0.0	0.996	do_not_reject_h_0	-3.188	3.204	95%	do_not_reject_h_0
carbohydrates	3.34	37.91	41.250	0	0.05	-1.0	0.319	do_not_reject_h_0	-9.922	3.242	95%	do_not_reject_h_0
fiber	0.04	1.4	1.360	0	0.05	0.2	0.843	do_not_reject_h_0	-0.357	0.437	95%	do_not_reject_h_0
protein	0.46	10.065	10.524	0	0.05	-0.32	0.746	do_not_reject_h_0	-3.255	2.337	95%	do_not_reject_h_0

There is sufficient evidence that two means between random populations of food menu items have the same means for calories, fat, carbohydrates, fiber, and protein.

5. Feature Engineering

In the development of a robust model to classify fast food items based on their nutritional profiles, feature engineering plays a crucial role. Our approach involves transforming raw nutritional data from menu items into insightful and actionable features that help predict dietary impacts effectively. This section details the methodological steps taken to derive these features, leveraging established dietary guidelines and nutritional science principles.

5.1 Data Transformation

The first step in our feature engineering process was to encapsulate the raw nutritional data into a structured format. For each menu item, we extracted key nutritional metrics such as calories, fiber, fat, carbohydrates, and protein. This was achieved using the get_nutrition function, which constructs a Nutrition object for each item. See section 3.2 on the nutritional scoring model calculation.

5.2 Scoring System

To quantify the nutritional impact of each menu item, we developed a scoring system that assesses the caloric intake, fat, carbohydrates, fiber, and protein content. Each nutritional aspect was evaluated against recommended daily values, considering a combined guideline for all sexes, which simplifies the model

without significantly sacrificing the accuracy for initial assessments. The scores were calculated as follows:

- Caloric Intake Score: Reflects the calorie content relative to daily recommended values.
- Fat Intake Score: Assesses the fat content, focusing on saturated fat relative to daily limits.
- Carbohydrate Intake Score: Gauges the carbohydrate content against daily needs.
- Fiber Intake Score: Evaluates fiber content, given its importance in digestion and health.
- Protein Intake Score: Measures protein content, essential for muscle repair and growth.

The weights were determined based on dietary priorities highlighted in health guidelines and the impact of these nutrients on overall health:

Nutritional Guidelines Reference
The scoring thresholds and weights were informed by the 2015-2020 Dietary Guidelines for Americans, which suggest variable caloric intakes based on sex and activity level. Further granularity was incorporated using specific recommendations from sources like the Mayo Clinic for fiber, the NHS for fat, and multiple resources for carbohydrates and protein. These guidelines help ensure that our scoring system is grounded in scientific consensus and practical dietary advice.

According to the 2015-2020 Dietary Guidelines for Americans, [women are likely to need between 1,600 and 2,400 calories a day, and women from 2,000 to 3,000 for men.[1]

According to the Mayo Clinic [women should try to eat at least 21 to 25 grams of fiber a day, while men should aim for 30 to 38 grams a day[2].

According to the National Health System (NHS) men should not eat more than 30g of saturated fat a day and women should not eat more than 20g of saturated fat a day.

According to the Week[3] women consuming 1,600-calorie diets need 180 to 260 grams, women following 2,000-calorie diets need 225 to 325 grams and women consuming 2,400 calories per day require 270 to 390 grams of carbohydrates each day. According to the American Academy of Orthopedic Surgeons, women athletes may require 60 to 70 percent of their calories from carbs, which is equivalent to 360 to 420 grams of carbs for a 2,400-calorie meal plan.

According to Healthline men should eat between 338–488 grams for a diet that consists of 3,000 daily calories[4]. And lastly, Eating Well recommends approximately 88 g to 122 g for women, 105 g to 145 g for men[5]. Remember, this research is a generalization for the worldly population. Please consult a health professional for your personal food intake.

The engineered features thus not only simplify the complex nutritional data but also align it with actionable health standards. This alignment allows our model to effectively classify fast food items based on their potential health impacts, providing valuable insights for consumers aiming to make informed dietary choices. The methodology ensures a balance between scientific accuracy and practical applicability, setting a solid foundation for further enhancements in predictive accuracy and user-specific customization in future iterations.

6. Modeling

6.1 Linear Regression

Collinearity, or multicollinearity, refers to a situation in which two or more predictor variables in a multiple regression model are highly correlated. This means that one predictor variable in the model can be linearly predicted from the others with a substantial degree of accuracy. When collinearity occurs, it can make the model analysis more complicated, potentially leading to unreliable or unstable estimates of regression coefficients. Multicollinearity increases the standard errors of the coefficients, which might lead to failing to identify important relationships between variables due to a lack of statistical significance.

Implications for This Data:
From the correlation analysis, provide in Table 6.1.1, in the provided dataset, several key insights emerge regarding collinearity among different nutritional features:

High Correlation Between Calories and Fat (r = 0.886):

This indicates a strong positive relationship, suggesting that as the calorie content of food items increases, the fat content also tends to be higher. The correlation is significant (p-value = 0.0), and the confidence interval for the correlation coefficient is tightly bound between 0.869 and 0.901, indicating a precise estimation.

Table 6.1.1 - Nutrient Feature Collinearity

	feature_1	feature_2	r	t	p_value	hypothesis	r_confidence_interval
0	calories	fat	0.886	49.533	0.0	reject_h_0	(0.869, 0.901)
1	calories	carbohydrates	0.459	13.393	0.0	reject_h_0	(0.397, 0.516)
2	calories	fiber	0.594	19.141	0.0	reject_h_0	(0.543, 0.641)
3	calories	protein	0.733	27.934	0.0	reject_h_0	(0.696, 0.766)
4	fat	carbohydrates	0.149	3.906	0.0	reject_h_0	(0.075, 0.222)
5	fat	fiber	0.561	17.568	0.0	reject_h_0	(0.507, 0.61)
6	fat	protein	0.689	24.644	0.0	reject_h_0	(0.647, 0.727)
7	carbohydrates	fiber	0.194	5.126	0.0	reject_h_0	(0.121, 0.266)
8	carbohydrates	protein	0.151	3.960	0.0	reject_h_0	(0.076, 0.223)
9	fiber	protein	0.507	15.248	0.0	reject_h_0	(0.449, 0.561)

Moderate to Strong Correlations Involving Calories:
Calories with protein (r = 0.733), fiber (r = 0.594), and carbohydrates (r = 0.459) also show significant positive correlations. These relationships suggest that higher-calorie items also tend to have higher amounts of these nutrients, albeit the strength of these relationships varies.

1 https://health.gov/our-work/nutrition-physical-activity/dietary-guidelines/previous-dietary-guidelines/2015

2 https://www.mayoclinic.org/healthy-lifestyle/nutrition-and-healthy-eating/in-depth/high-fiber-foods/art-20050948#:~:text=Women%20should%20try%20to%20eat,It%20can%20vary%20among%20brands

3 https://www.weekand.com/healthy-living/article/recommended-intake-grams-carbohydrates-per-day-women-18021277.php

4 https://www.healthline.com/nutrition/3000-calorie-meal-plan#method

5 https://www.eatingwell.com/article/290496/this-is-how-much-protein-you-need-to-eat-every-day

Other Noteworthy Correlations:
Fat and protein (r = 0.689), fat and fiber (r = 0.561): These show moderate to strong positive correlations. Carbohydrates with fiber and protein show relatively weaker correlations (r values around 0.15 to 0.2), suggesting less dependency between these variables compared to others.

Meaning and Consequences:
Modeling and Interpretation Challenges: The presence of strong correlations, particularly between calories and fat, might pose challenges in regression models attempting to predict nutritional outcomes based on these features. High collinearity can inflate the variance of the coefficient estimates and make the model sensitive to changes in model specifications.

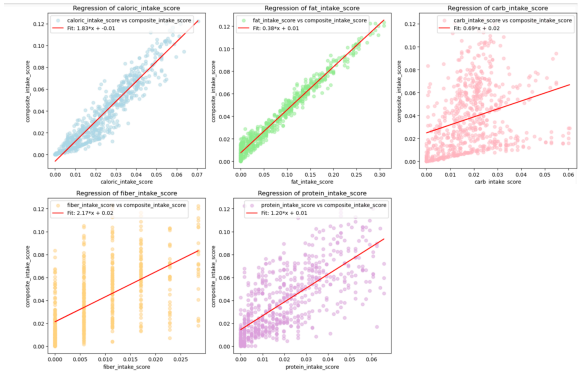
Redundancy:
High collinearity might indicate redundancy in the features, which suggests that some variables could be removed without much loss of information in predictive modeling contexts.
Nutritional Insight: From a nutritional perspective, the strong correlation between calories and fat might be used to inform dietary recommendations or food formulation decisions in a way that manages caloric content without disproportionately increasing fat content.

In summary, the collinearity observed in this dataset helps in understanding the relationships among various nutritional components of fast food items, which is crucial for both developing accurate predictive models and making informed nutritional assessments or recommendations.

There is sufficient evidence that there is no collinearity between the various features of fast food menu item nutrients. Due to the fact that there is no collinearity this will allow us not to apply any feature selection methods thereby we need to eliminate any features in our CFIS model. See more details in the section on the CFIS model.

Figure 6.1.1 shows scatter plots with fitted regression lines, each representing the relationship between different nutritional intake scores and a composite intake score for food items. These charts are a form of regression analysis, a common method used to investigate the relationship between independent variables (or predictors) and a dependent variable.

Figure 6.1.1 - Single Linear Regressions



Individual Chart Analysis:
Caloric Intake Score vs. Composite Intake Score: This chart shows a positive correlation between the caloric intake score and the composite intake score, as indicated by the upward-sloping regression line. The equation provided (Fit: 1.83x + 0.01) suggests that for every unit increase in the caloric intake score, the composite intake score increases by 1.83 units, after controlling for the intercept.

Fat Intake Score vs. Composite Intake Score: The second chart depicts a positive correlation between the fat intake score and the composite intake score. The regression line is less steep compared to the caloric intake score chart (Fit: 0.38x + 0.01), indicating a weaker relationship. Here, for every unit increase in the fat intake score, the composite intake score increases by 0.38 units.

Carbohydrate Intake Score vs. Composite Intake Score: There is a moderate positive relationship between carbohydrate intake score and the composite intake score (Fit: 0.69x + 0.02), suggesting that carbohydrates also contribute significantly to the composite score.

Fiber Intake Score vs. Composite Intake Score: The fiber intake score appears to have a positive correlation with the composite intake score, though the data points are more spread out, indicating more variability. The fit line suggests that the composite score increases by 2.17 units for every unit increase in fiber intake score, which is a strong relationship according to the slope of the line.

Protein Intake Score vs. Composite Intake Score: This chart shows that there is a strong positive relationship between the protein intake score and the composite intake score, with a regression line slope of 1.20 (Fit: 1.20x + 0.01).

Summary:
The composite intake score, which is likely a measure of the overall nutritional value of food items, seems to be positively influenced by all the individual nutritional scores — calories, fat, carbohydrates, fiber, and protein. Calories have the strongest influence, followed by protein and fiber, then carbohydrates, and lastly fat, which has the least impact among the variables shown.

These relationships can be useful in understanding how different nutritional components contribute to the overall dietary quality of food items. The positive slopes across all charts indicate that as the intake scores for individual nutrients increase, so does the composite score, which might be an aggregate measure of nutritional density or quality.

However, the spread of data points around the regression line in each plot indicates the degree of variability in these relationships. A tighter cluster of points around the line would suggest a more consistent relationship across different food items, while a wider spread, as seen in the fiber intake score chart, suggests more variability in how fiber intake scores relate to the overall composite score.

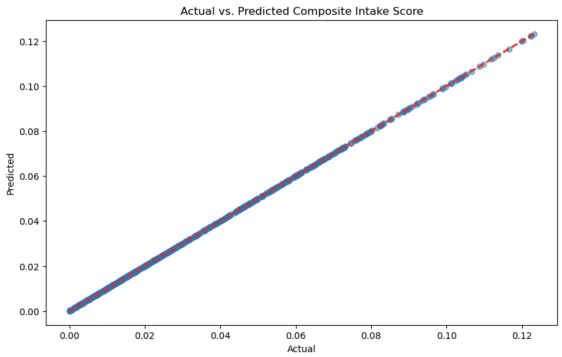
These analyses are important for diet formulation and health-related research, as they help in quantifying and understanding the contribution of various nutrients to overall dietary health.

6.2 Multi-regression

Multi-regression, or multiple regression, models are statistical techniques that explore the relationship between one dependent variable and two or more independent variables. The model assesses the strength of the influence each independent variable has on the dependent variable, while also controlling for the effects of other variables. This is especially useful for understanding complex scenarios where the dependent variable is influenced by multiple factors.

Figure 6.2.1 displays a scatter plot comparing the actual values of a composite intake score against the values predicted by a multiple regression model.

Figure 6.2.1 - Multi-Linear Regressions



The x-axis represents the actual values of the composite intake score. These are the observed values for each item in the dataset.
The y-axis represents the predicted values of the composite intake score. These are the values predicted by the multiple regression model based on the various input features. The dotted line ideally represents where the actual values would perfectly match the predicted values. This is often referred to as the "line of perfect fit" or "identity line". The solid line is the regression line, and it represents the relationship between the actual and predicted scores as determined by the model. In a perfect model, where predictions match reality exactly, all points would lie on the dotted line. The closeness of the points to this dotted line indicates the accuracy of the model's predictions:

If the points are scattered far from the dotted line, the predictions are inaccurate. If the points are close or on the dotted line, it indicates accurate predictions.

Interpretation of This Chart:
The points on the scatter plot are tightly clustered around the dotted line of perfect fit, indicating that the model's predictions are very close to the actual values. This suggests that the multiple regression model has high predictive accuracy for the composite intake score based on the independent variables it includes.

The alignment of the regression line (solid) with the line of perfect fit (dotted) demonstrates that the model is well-calibrated and that the independent variables used in the model are strong predictors of the composite intake score. It also suggests that the model does not consistently overpredict or underpredict the composite intake score, which would have been indicated by a solid regression line above or below the dotted line, respectively.

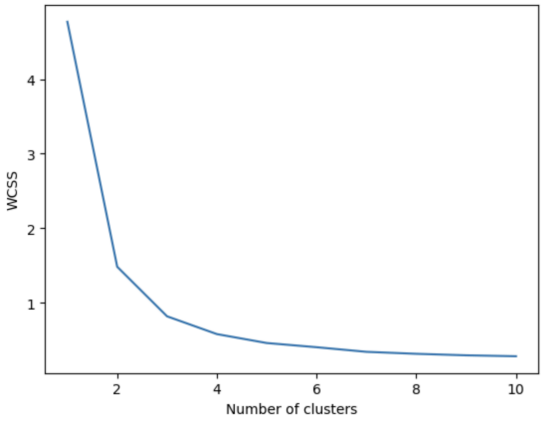
Summary:
Naturally, Figure 6.2.1 shows a strong positive correlation between the actual and predicted composite intake scores, indicating that the multiple regression model is performing well in predicting the composite intake score based on the multiple nutritional intake scores used as predictors in the model.

6.3 K-Means

The elbow method involves running the k-means clustering algorithm on the dataset for a range of values of k (e.g., k=1 to 10), and for each value of k, calculating the WCSS for the model. The values are then plotted, and the "elbow point," where the rate of decrease sharply changes, represents the optimal number of clusters. This point is typically chosen because adding more clusters beyond this number does not provide much better modeling of the data.

Elbow Chart Analysis:
In the provided elbow chart, in Figure 6.3.1, the x-axis represents the number of clusters and the y-axis represents the WCSS. As the number of clusters increases, the WCSS decreases because the data points are closer to the centroids of their respective clusters.

Figure 6.3.1 - Fast Food Classification Elbow Chart

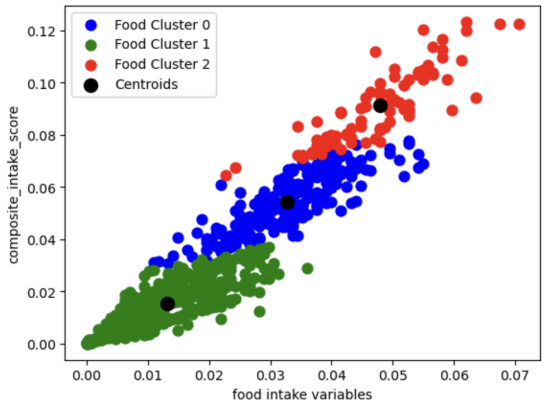


From the chart, we see a rapid decrease in WCSS as the number of clusters goes from 1 to 2, and then from 2 to 3. After 3 clusters, the decrease in WCSS starts to level off, indicating that adding more clusters beyond 3 does not significantly improve the fit of the model.

Choosing the Right Number of Food Clusters:
Based on the chart, the elbow seems to occur at the point where k equals 3. This suggests that three is the optimal number of clusters for this dataset because it is at this point that adding more clusters will result in diminishing returns in terms of decreased WCSS. Hence, we would choose three clusters for this k-means clustering model. This choice is based on the desire to minimize WCSS while also aiming to have a model that is as simple as possible without sacrificing accuracy.

Food Clusters:
This KMeans model scatter plot visualization, in Figure 6.3.2, depicts the clustering of data points based on their features using the KMeans algorithm. Each cluster is represented by a different color:

Figure 6.3.2 - K-Means Scatterplot Food Clusters



Blue dots represent the data points in Food Cluster 0.
Green dots represent the data points in Food Cluster 1.
Red dots represent the data points in Food Cluster 2.
Black dots represent the centroids of each cluster, which are the mean positions of all the data points within each cluster.
From the scatterplot, you can infer the following:

Cluster Formation: There are three distinct clusters, each with a varying degree of composite intake scores and food intake variables. It suggests that the KMeans algorithm has identified three groups within the data that share similar characteristics.

Centroids: The centroids indicate the "center" of each cluster. Their position suggests the average profile of each cluster. For example, the centroid of Food Cluster 1 is lower on both axes than the centroids of Clusters 0 and 2, which suggests that items in Cluster 1 have lower average scores for the food intake variables and composite intake score.

Spread of Clusters:

Food Cluster 1 (Green): appears to have lower scores for both food intake variables and composite intake score, which might indicate a grouping of food items that are lower in these nutritional aspects.

This cluster likely consists of foods with lower values of calories, fat, and carbohydrates because it is associated with lower CFIS values. Given the negative coefficients for fiber and protein in the CFIS formula, it's plausible that the foods in this cluster could have moderate to higher amounts of fiber and protein, which contribute to lowering the CFIS.

Foods in this cluster could be considered as more balanced or possibly healthier options due to their lower calorie, fat, and carbohydrate content relative to other clusters.

Food Cluster 0 (Blue): generally has higher food intake variable scores but similar composite intake scores to Cluster 1, suggesting these items might be richer in specific nutrients but have a similar overall nutritional profile.

Foods in this cluster show moderately high values for calories, fat, and carbohydrates as indicated by moderate CFIS values. These foods might have lower amounts of fiber and protein compared to Cluster 1, as their CFIS is higher, despite fiber and protein's negative influence on the score.

This cluster represents a middle ground, possibly containing foods that are moderately rich in calories, fat, and carbohydrates, but not as nutrient-dense in terms of fiber and protein.

Food Cluster 2 (Red)): has the highest scores on both axes, indicating that these food items might be the most nutrient-dense or have higher nutritional quality based on the scoring system used.

Conclusions:
The visualization clearly shows how the KMeans clustering algorithm has grouped the food items into three categories based on their nutritional content as quantified by the intake scores. These clusters could *potentially* be used to categorize foods into different healthiness levels or to recommend dietary choices based on nutritional content. For instance, items in Cluster 1 could be considered 'basic' or 'standard' foods, Cluster 0 might represent 'moderately nutritious' foods, and Cluster 2 might consist of 'highly nutritious' foods.

Summary:

The success of the clustering would ultimately depend on the specific definitions and purposes of the intake scores, which should be examined in the context of dietary recommendations or nutritional guidelines.

In summary, Cluster 1 might be characterized by food items with lower calories, fat, and carbohydrates, and potentially higher in fiber and protein. Cluster 0 includes foods with moderate levels of unhealthy nutrients (calories, fat, carbohydrates) and possibly lower fiber and protein. Cluster 2 likely consists of high-calorie, high-fat, and high-carbohydrate items with lower fiber and protein content, which could be considered the least healthy options according to the CFIS metric.

7. Conclusion

While this report thoroughly examines the various food nutrients and how they impact fast food menu items from a statistical and modeling perspective there are more experiments to perform that can enhance and/or further this research.

The EDA, feature engineering, and modeling performed on this dataset of fast food nutritional information have yielded insightful results about the nutritional landscape of these foods. The EDA provided a clear overview of the dataset, highlighting the distribution of calories, fat, carbohydrates, fiber, and protein across a variety of fast food items. Through feature engineering, we were able to construct a CFIS that encapsulates the nutritional profile of food items in a single, holistic measure. This facilitated the effective clustering of the food items through the KMeans algorithm, enabling us to categorize them into distinct groups based on their nutritional content.

However, there is always room for improvement and further research that could refine our understanding and application of these findings:

Box Cox Transformation: Applying the Box Cox transformation could help in normalizing the distribution of our features, potentially enhancing the performance of our clustering algorithm and leading to more distinct and interpretable clusters.

Analyzing More Fast Food Restaurants: Including data from a broader range of fast food restaurants would increase the generalizability of our model, providing a more comprehensive view of the fast food industry's nutritional offerings.

Adding Additional Food Nutrients for Analysis and/or Modeling: Incorporating more detailed nutritional information such as sodium content, trans fats, vitamins, and minerals could offer deeper insights into the health implications of fast food items.

Consultation with Food Science Experts and Health Professionals: Engaging with dieticians, food scientists, and medical health professionals can enrich our analysis with expert insights into the implications of our findings and ensure that our modeling approach aligns with industry standards and public health objectives.

Deployment of the Model for Real-Time Classification: Finally, developing a predictive model to classify food items in real time would be a valuable tool for both consumers and food providers, offering immediate nutritional assessments and fostering healthier dietary choices within the fast food landscape.

By addressing these areas, future research can build upon the current work, enhancing the utility and accuracy of nutritional profiling in the fast food industry, and ultimately contributing to better public health outcomes.