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 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 – 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 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 2 lists the initial list of datasets that are available that seems to 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 2 – 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 six independent numerical variables across approximately 1,100 samples. 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, we feel it is 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. Feature Engineering

5. Modeling

6. Report

6.1 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 1-4 show various types of visualizations using box plots, scatterplots, and bar histograms

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 1 - Fast Food Menu Item Paired Scatterplots

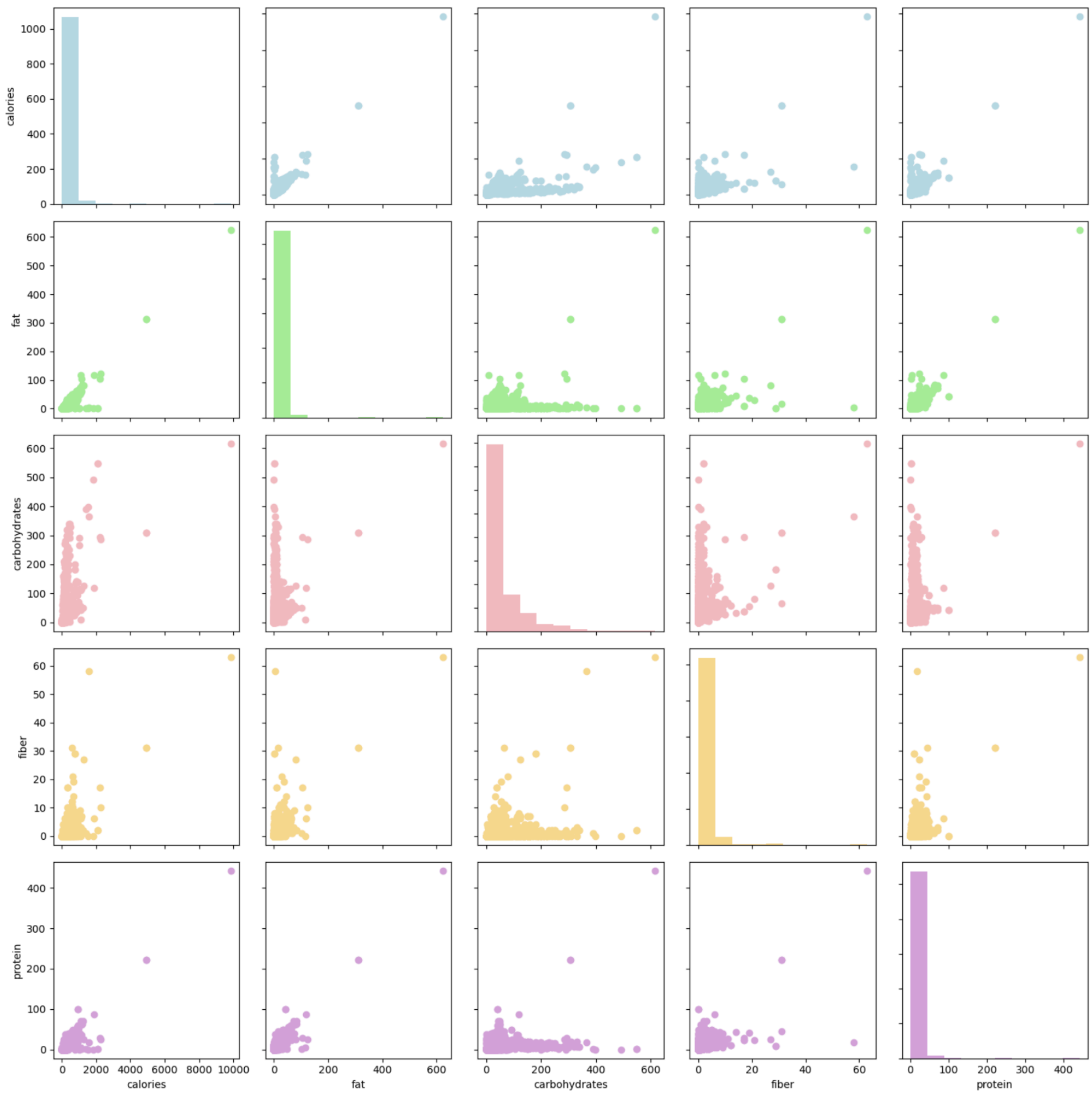


Figure 2 - Fast Food Menu Item Individual BoxPlot

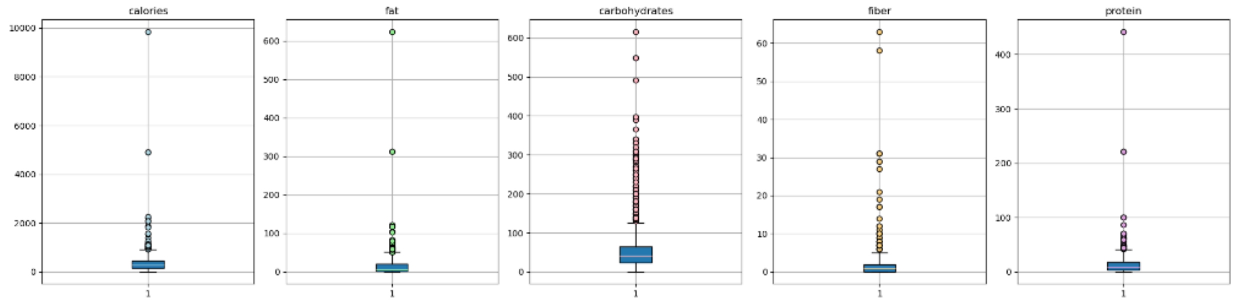


Figure 3 - Fast Food Menu Item Combined BoxPlot

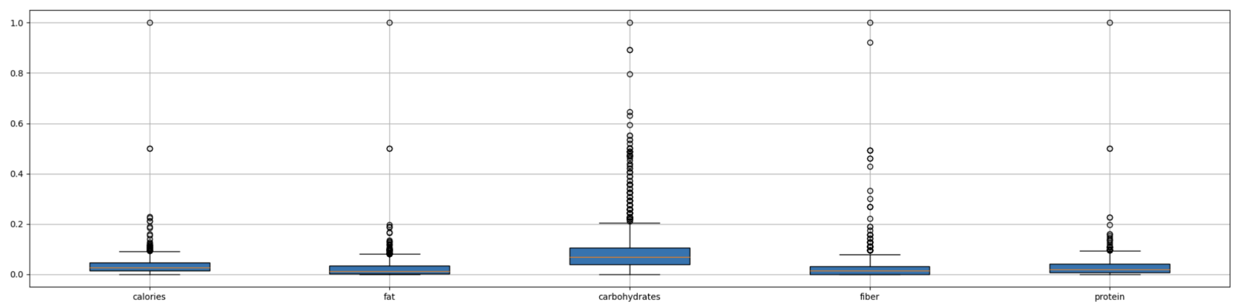
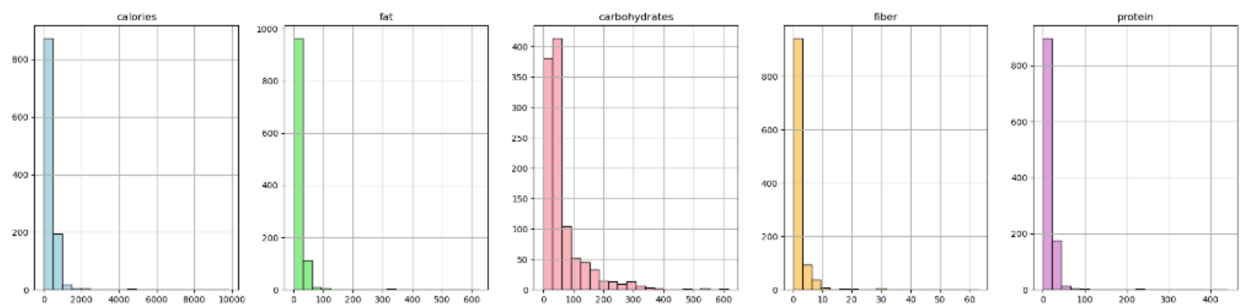


Figure 4 - Fast Food Menu Item Individual Histogram



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.

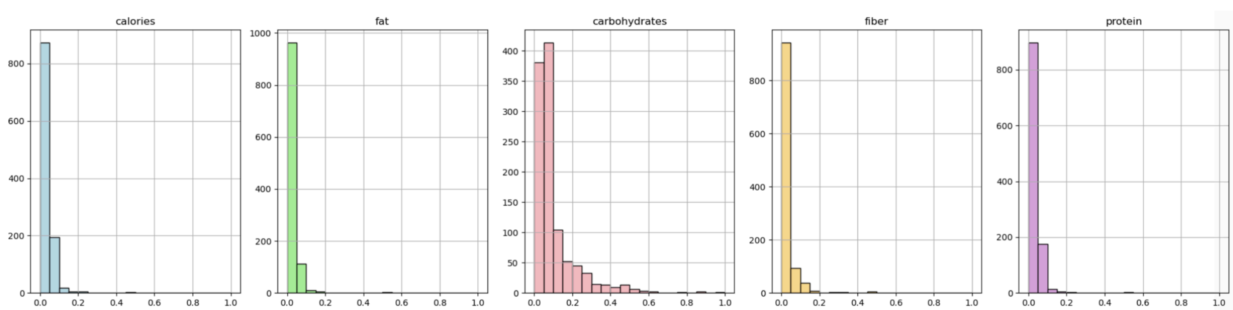
   The formula for min-max scaling of a value x in a feature X is:

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

   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 5. In Figure 5, we can still see that even after scaling the data mostly follows a distribution that is skewed to the right.

Figure 5 - Min-Max Scaling - Fast Food Menu Item Individual Histogram



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 standard deviation 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 1117.
- H0 = p = 0.50 or 50%
- H1 = p > 0.50 or 50%

Population Proportion Claim(s):

Table 3 – Population Proportion Claims

Claim	z	p-value	Conclusion
The total number of unique fast food items on the menu are 1117 and the total number of food items that have more than 270 calories are 547 out of 1117 for a percentage of 48.97%.	73.7868	0.00	There is sufficient evidence that more than 50% of the fast foods in the U.S. have more than 270 calories.
The total number of unique food items on the menu are 1117 and the total number of fast food items that have more than 1 gram are 545 out of 1117 for a percentage of 48.79%.	73.5128	0.00	There is sufficient evidence that more than 50% of the fast foods in the U.S. have more than 1 gram of fat.
The total number of unique fast food items on the menu are 1117 and the total number of food items that have more than 41 carbs are 558 out of 1117 for a percentage of 49.96%.	75.2939	0.00	There is sufficient evidence that more than 50% of the fast foods in the U.S. have more than 41 carbs.
The total number of unique fast food items on the menu are 1117 and the total number of food items that have more than 1 gram of fiber are 414 out of 1117 for a percentage of 37.06%.	55.6560	0.00	There is sufficient evidence that more than 50% of the fast foods in the U.S. have more than 1 gram of fiber.
The total number of unique fast food items on the menu are 1117 and the total number of food items that have more than 1 gram of fiber are 576 out of 1117 for a percentage of 51.57%.	77.7448	0.00	There is sufficient evidence that more than 50% of the fast foods in the U.S. have more than 8 grams of protein.

While we correctly performed the calculations z-statistic, the numbers are going to be incorrect due to the fact that each of the nutrients does *not* follow a normalized distribution.

Population Mean Claim(s):

Table 4 – Population Mean Claims

Claim	t	p-value	Conclusion
The mean number of calories for fast foods in the U.S is greater than 341.26.	26.0738	0.00	There is sufficient evidence that the mean number of calories for fast foods in the U.S is greater than 341.26.
The mean number of calories for fast foods in the U.S is greater than 14.35.	17.3788	0.00	There is sufficient evidence that the mean number of grams of fat for fast foods in the U.S is greater than 14.35.
The mean number of carbs for fast foods in the U.S is greater than 61.12.	29.2954	0.00	There is sufficient evidence that the mean number of carbs for fast foods in the U.S is greater than 61.12.
The mean number of grams of fat for fast foods in the U.S is greater than 1.84.	15.5585	0.00	There is sufficient evidence that the mean number of grams of fiber for fast foods in the U.S is greater than 1.84.
The mean number of grams of protein for fast foods in the U.S is greater than 13.0.	21.6382	0.00	There is sufficient evidence that the mean number of grams of protein for fast foods in the U.S is greater than 13.0

While we correctly performed the calculations for the t-statistic, the numbers are going to be incorrect due to the fact that each of the nutrients does *not* follow a normalized distribution.

Standard Deviation Claim(s):

Table 5 – Population Standard Deviation Claims

Claim	X^2	p-value	Conclusion
The standard deviation for the number of calories for fast foods in the U.S is greater than 400.	1334.6610	0.00	There is sufficient evidence that standard deviation for the number of calories for fast foods in the U.S is greater than 400.
The standard deviation for the number of grams of fat for fast foods in the U.S is greater than 251.	1360.8296	0.00	There is sufficient evidence that standard deviation for the number of grams of fat for fast foods in the U.S is greater than 25.
The standard deviation for the number of carbs for fast foods in the U.S is greater than 653.	1284.3060	0.00	There is sufficient evidence that the standard deviation for the number of carbs for fast foods in the U.S is greater than 65.
The standard deviation for the number of grams of fiber for fast foods in the U.S is greater than 3.	946.0995	0.00	There is sufficient evidence that the standard deviation for the number of grams of fiber for fast foods in the U.S is greater than 3.
The standard deviation for the number of grams of protein for fast foods in the U.S is greater than 20.	1125.1480	0.00	There is sufficient evidence that the standard deviation for the number of grams of protein for fast foods in the U.S is greater than 20.