# Restaurant Food Data

Arissa Trombley, Chaz Gillette, Decker Mecham

## Introduction

Our data originally revolved around grocery items with nutritional data, but with access restrictions, we had to settle for restaurant sample data, which included items from restaurants with nutritional data. The variables available include 'brand_name', 'item_name', 'brand_id', 'item_id', 'upc', 'item_type', 'item_description', 'nf_ingredient_statement', 'nf_calories', 'nf_calories_from_fat', 'nf_total_fat', 'nf_saturated_fat', 'nf_trans_fatty_acid', 'nf_polyunsaturated_fat', 'nf_monounsaturated_fat', 'nf_cholesterol', 'nf_sodium', 'nf_total_carbohydrate', 'nf_dietary_fiber', 'nf_sugars', 'nf_protein', 'nf_vitamin_a_dv', 'nf_vitamin_c_dv', 'nf_calcium_dv', 'nf_iron_dv', 'nf_potassium', 'nf_servings_per_container', 'nf_serving_size_qty', 'nf_serving_size_unit', 'nf_serving_weight_grams', 'images_front_full_url', 'updated_at', and 'section_id'.
With the data, we would be able to cluster meals to look at what high-calorie meals have in common. With predictive models, we would also be able to narrow down specific variables that contribute to a high calorie count. This informations can be applied to make food recommendations and inform consumers about what to look out for when ordering out.
With our data analysis, we aimed to define characteristics of the data that can help food distributors sell products to the right customers and have the customers be able to choose the diet they want.

## Methods

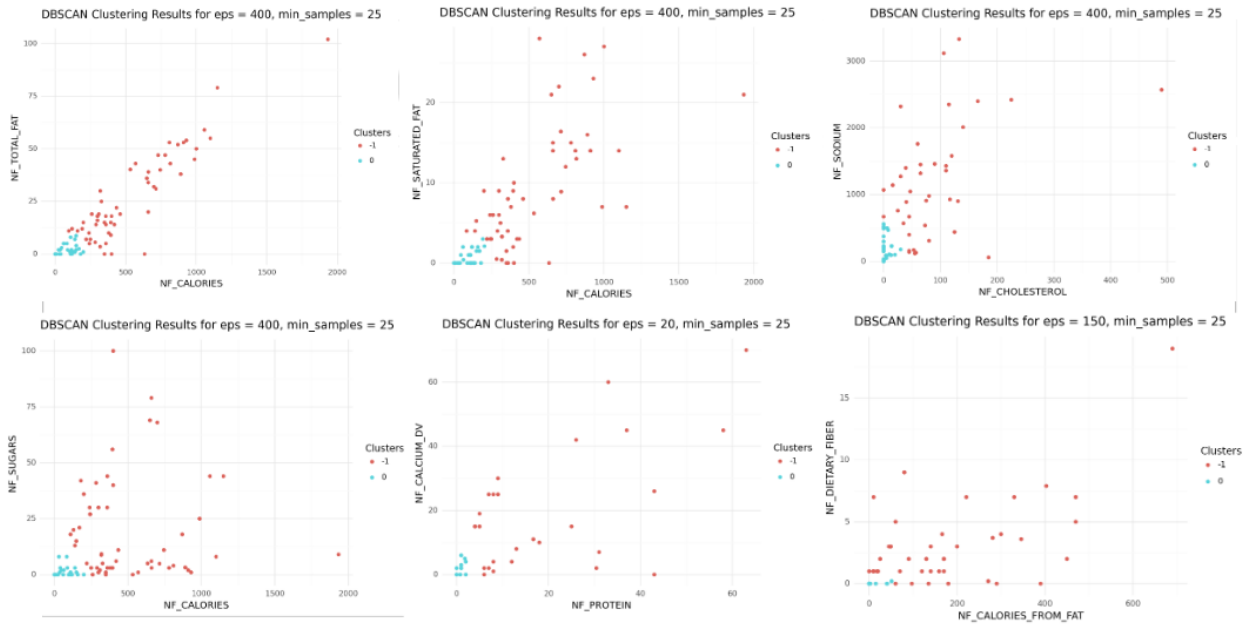
There are many null entries in our data set, so the rows with null values for the respective attributes used in each model were removed, and the index was reset to clean the data. The continuous variables were preprocessed through z-scoring to essentially standardize the values for the models. The clustering analysis applied to the data includes Gaussian Mixture and DBSCAN. These clustering algorithms offer insight into how different types of food items are related to each other when put up against chosen nutritional attributes. To determine certain model parameters, both of these clustering techniques benefit from the elbow method in which the value at the ‘elbow’ of a curve is grabbed for our parameter.

To create a recommendation system, a supervised version of the K Nearest Neighbors model was used to grab the closest items in the restaurant data for a new data set of customers who typically purchase high-sugar foods. We train the model with the restaurant data and have it output the most similar items from the customer data items depending on 'nf_protein',

'nf_calories', 'nf_total_fat', 'nf_sugars', and 'nf_cholesterol'. For the visualization, we group the dataset by brand and calculate the average sodium content for each brand. Also, we identify the brand with the highest average nf_sodium content. This is done through multiple gg-plots. For the PCA, I used a cumulative variance plot to decide on 4 principle components to achieve 90% explained variance. MAE is performed on a linear regression model for the PCA and for all nutrient variables.

In order to find out which variables contributed the most to calorie count we decided to use a supervised model. The variables included were: 'nf_calories', 'nf_total_fat', 'nf_saturated_fat', 'nf_trans_fatty_acid', 'nf_cholesterol', 'nf_sodium', 'nf_total_carbohydrate', 'nf_dietary_fiber', 'nf_sugars', 'nf_protein'. Values were dropped and a regression model was fit. After calculating the MSE and R^2 to determine the accuracy of the model, the data was plotted to look at the influence of each coefficient.

## Results

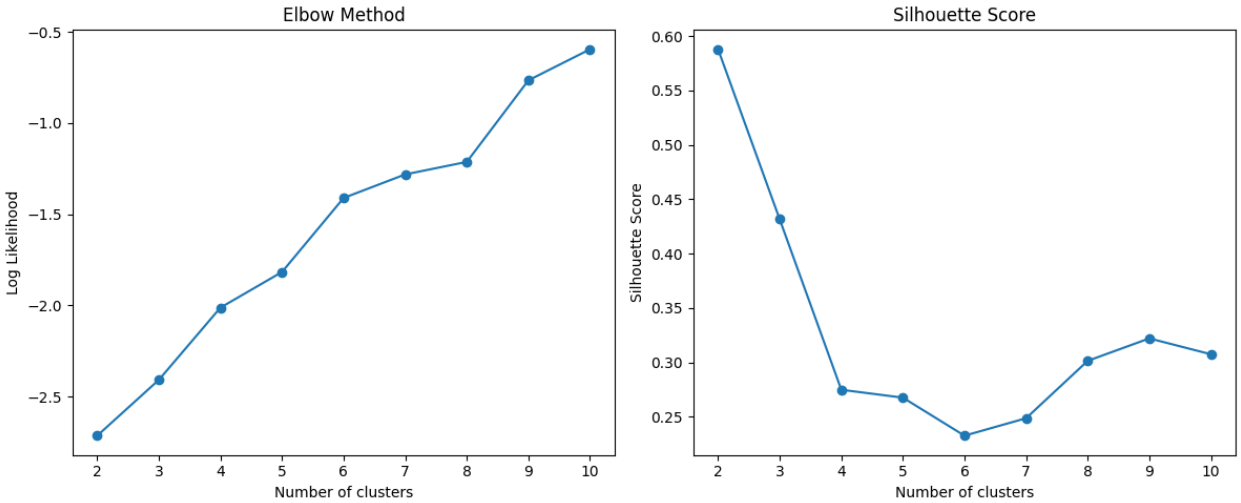



Graphs DBSCAN applied data between the variables "nf_calories" and "nf_total_fat", "nf_calories" and "nf_saturated_fat", "nf_cholesterol" and "nf_sodium", nf_calories" and "nf_sugars", "nf_protein" and "nf_calcium_dv, nf_calories_from_fat" and "nf_dietary_fiber"

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **2** | 1 | 209 | 2 | John Doe | Brownie | 38 | 250 | 15 | 4 | 6 | [22, 34, 24, 52, 1, 31, 25, 12, 8, 73] |
| **3** | 1 | 204 | 1 | John Doe | Soda | 40 | 120 | 1 | 0 | 0 | [8, 34, 22, 12, 25, 31, 41, 78, 77, 1] |
| **4** | 1 | 208 | 4 | John Doe | Milkshake | 35 | 400 | 20 | 8 | 12 | [24, 52, 1, 73, 22, 31, 12, 25, 34, 77] |

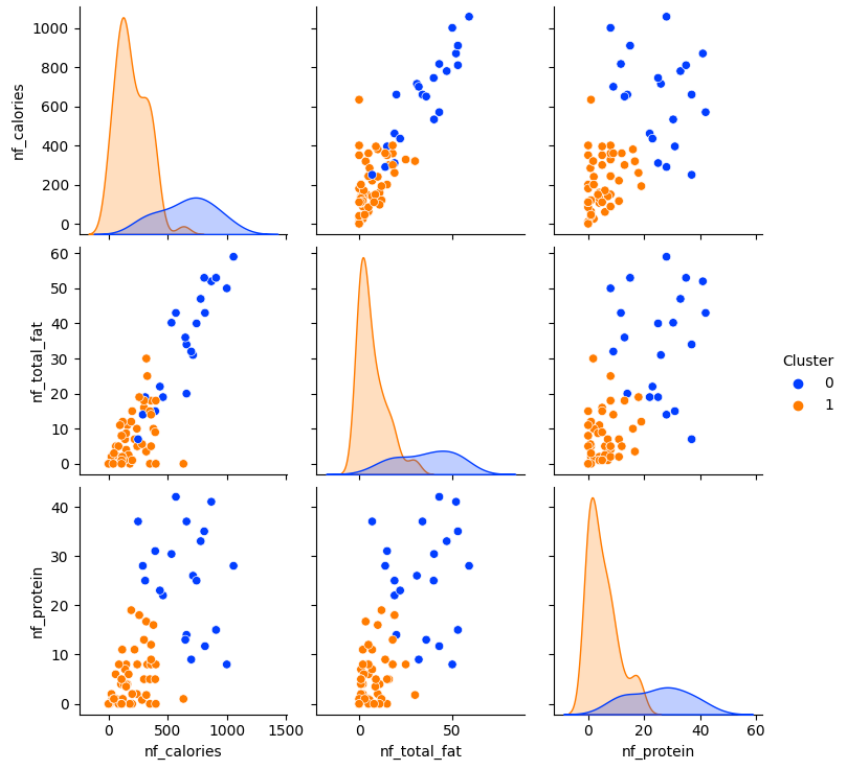
This is an example of the outputs for the recommendation system. It shows the ten nearest neighbors to each data set. The model accurately recommends a ‘Kamikaze Brownie’ for the Brownie data points.

## GMM



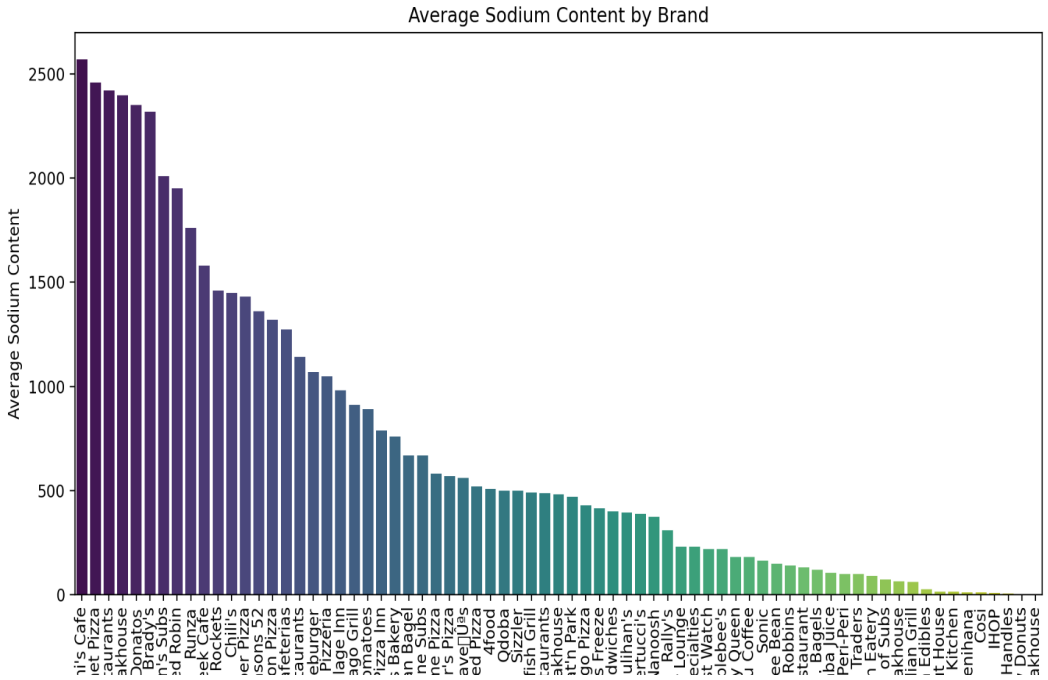

Elbow/Silhouette Scores: The Elbow method here shows that the data is likely not appropriate for GMM as the log-likelihood increases in the elbow method but the Silhouette Score decreases with the number of clusters

.




GMM Clusters: The Linear relationship shows the clustering between healthy and unhealthy foods


Linear Model: Total Fat, Carbs, and Protein show their contribution to calories.




The bar chart shows the average sodium content in the products of different brands. Each bar represents a brand, and the height of the bar shows the average amount of sodium in their products. Brands on the left have higher average sodium levels in their products, while those on the right have lower levels.

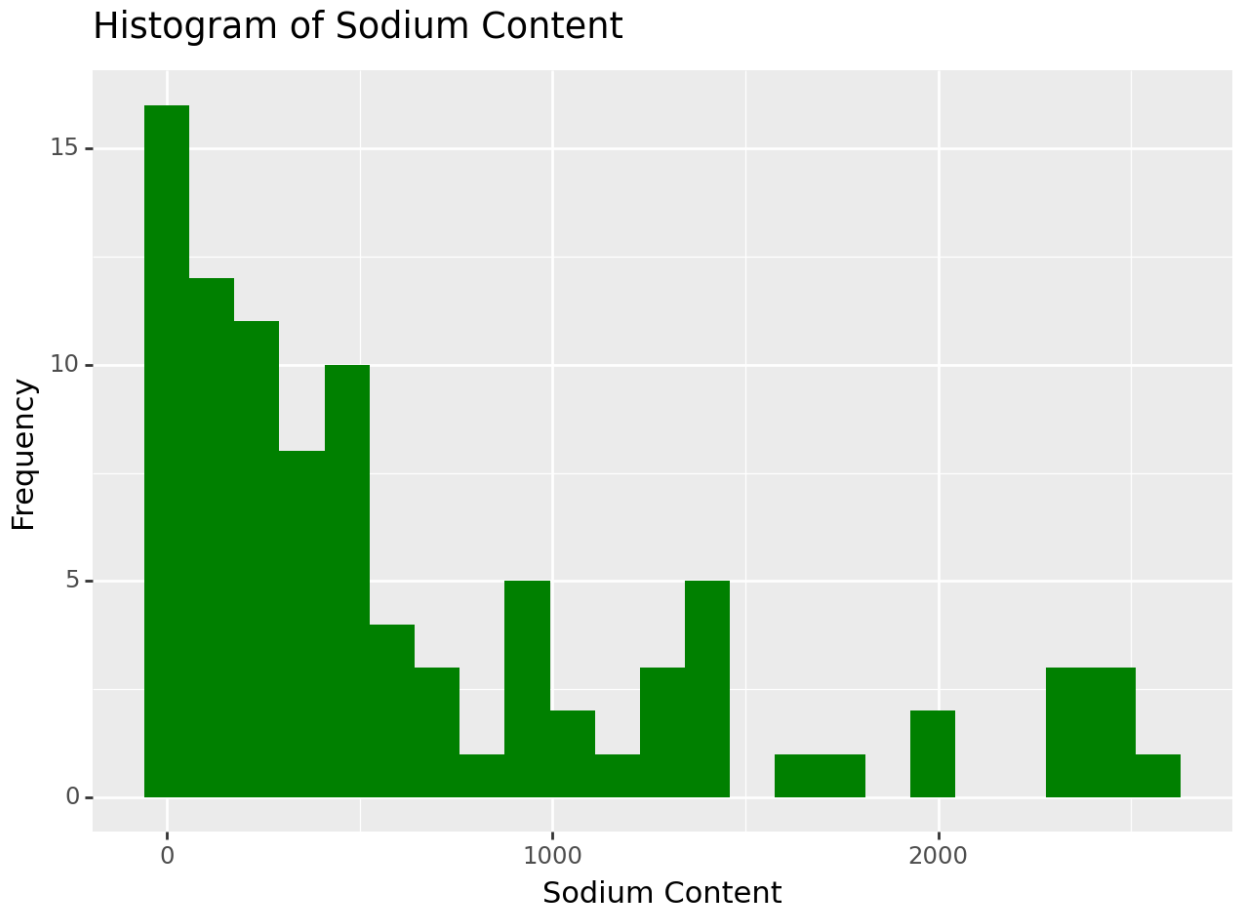
Median Sodium Content Across All Brands: 397.0 mg

Average Sodium Content Across All Brands: 648.1194565217392 mg

Brand with the highest average sodium content:

brand_name Mimi's Cafe

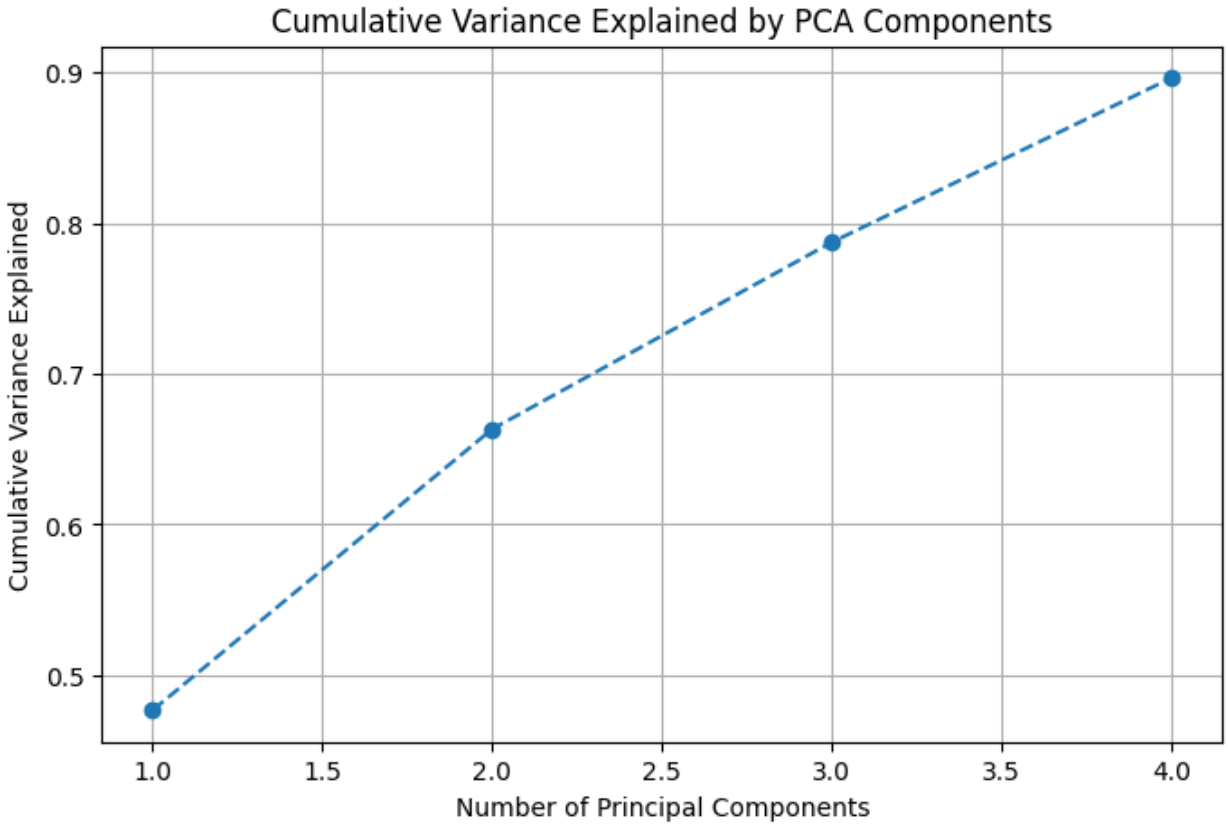
nf_sodium 2570.0 mg




This histogram shows the distribution of sodium content in different foods or samples. Each bar represents a range of sodium values, and the height of the bar shows how many foods fall into that range. For example, a lot of foods have a sodium content between 0 and, say, 500 milligrams because the bars are tall. Fewer foods have very high sodium content, as shown by the shorter bars at the end. This kind of graph helps us see which sodium levels are most common in the foods tested. It is important to be aware of sodium consumption as a consumer because the side effects of too much sodium intake are below.

- Enlarged heart muscle
- Headaches
- Kidney disease
- Osteoporosis
- Stroke
- Heart failure
- High blood pressure
- Kidney stones
- Stomach cancer

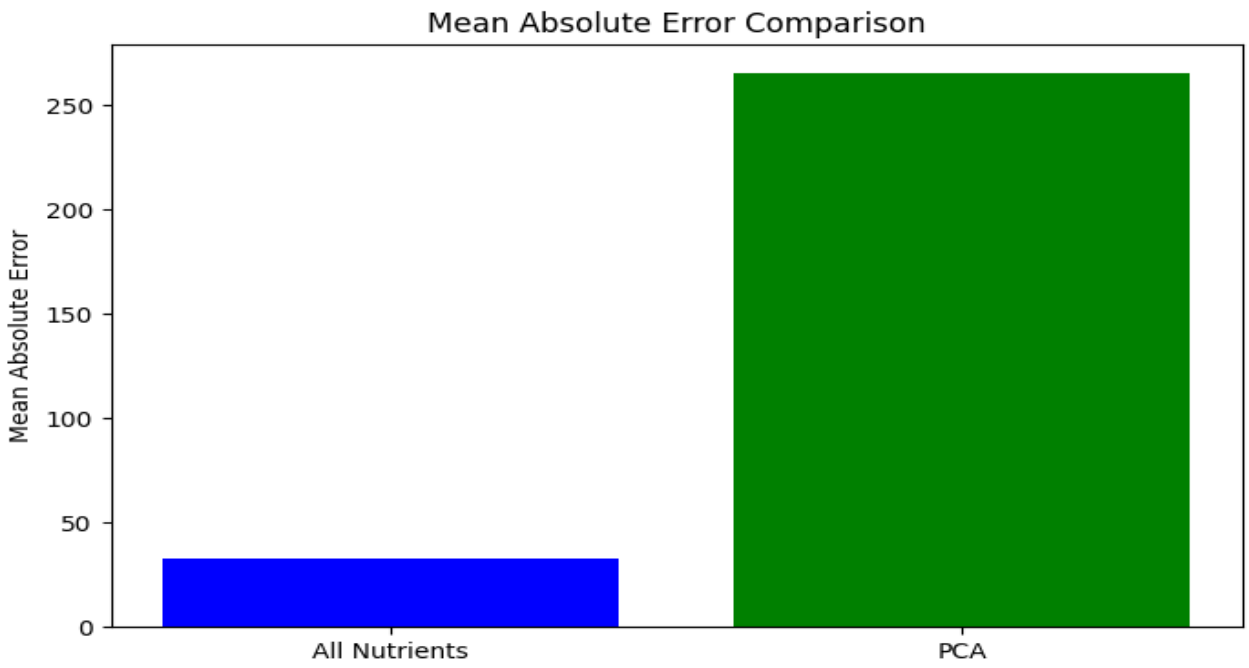
When comparing a model using PCA on all the nutrient variables (total fat, saturated fat, trans fat, cholesterol, sodium, total carbohydrate, dietary fiber, sugars, protein) in the dataset and retaining enough components to keep 90% of the variance, to a model using all the nutrient variables, how much of a difference is there in mean absolute error when predicting calorie count?




This plot shows the cumulative variance explained by the first five principal components of a PCA (Principal Component Analysis).

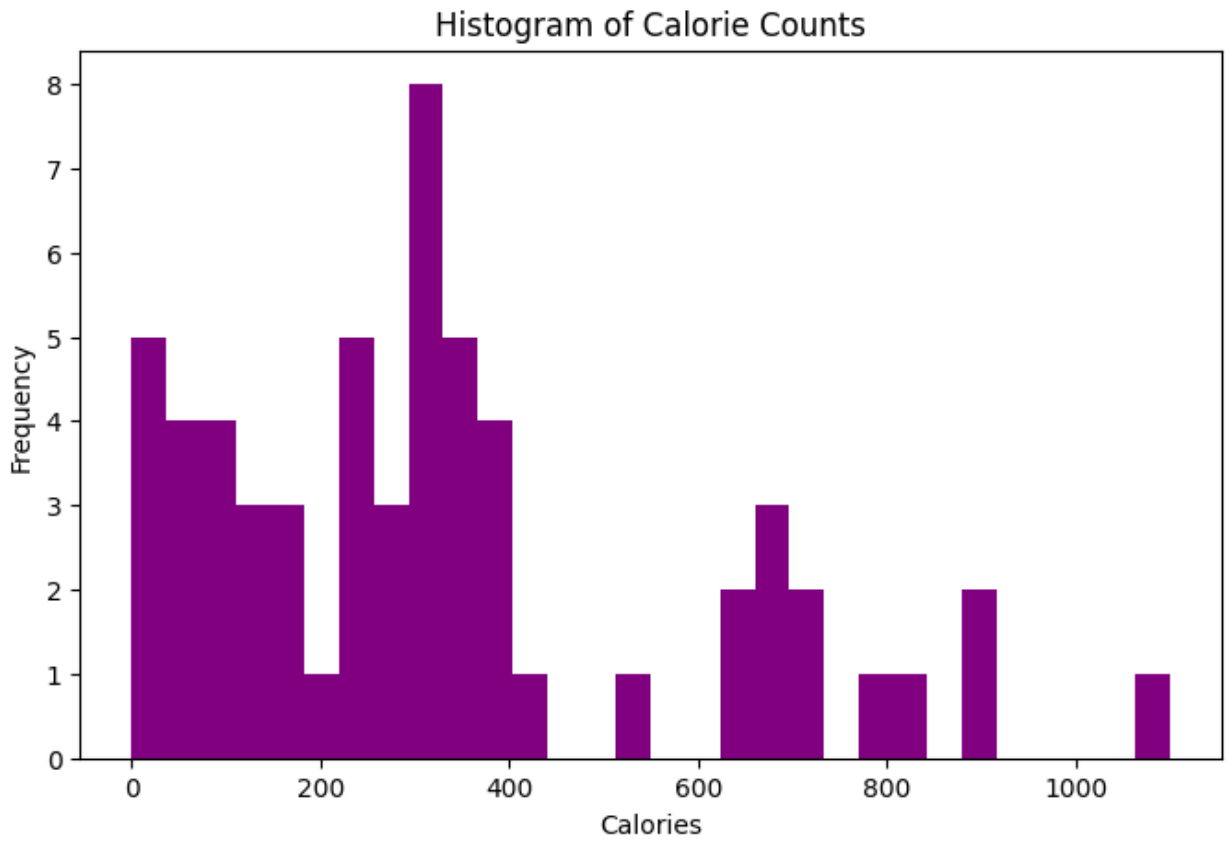
For example, the first principal component (PC1) explains a little over 40% of the variance. When you add the second principal component (PC2), together they explain around 65% of the variance. This trend continues, and by the time you include the fourth principal component, you are explaining about 90% of the variance in the dataset.

The plot is useful for deciding how many principal components to keep. You want to capture as much information (variance) as possible, with as few components as possible. In this case, using four components gets you very close to explaining 90% of the variance, allowing us to reduce the dimensionality of our data significantly.

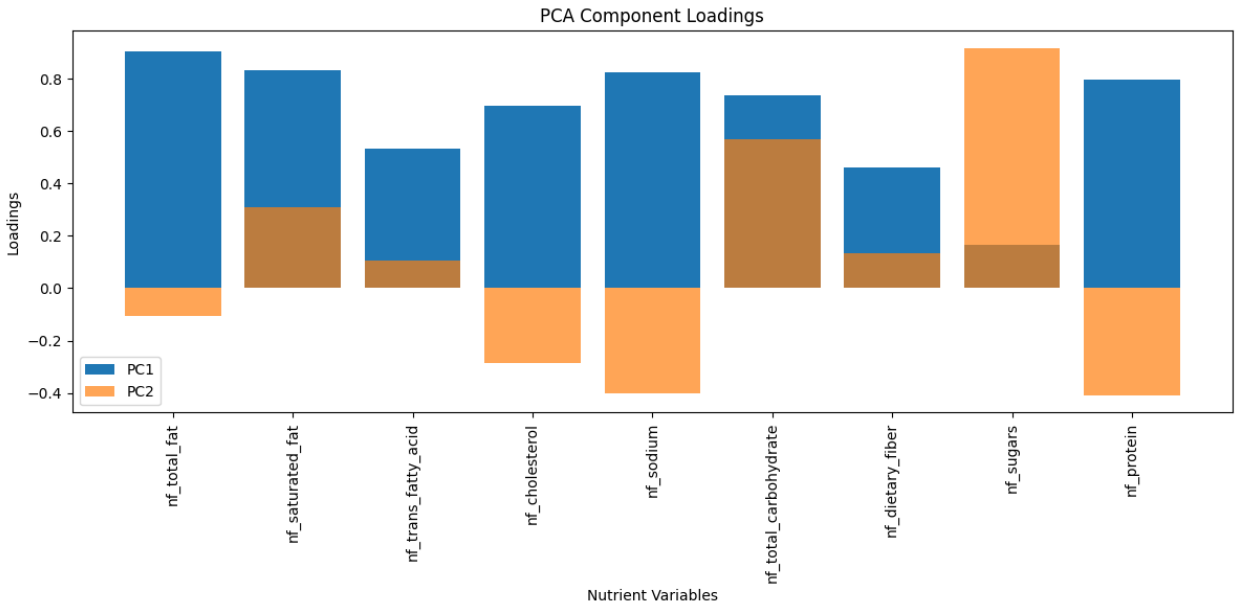



Mean Absolute Error (All Nutrients): 32.93801126874494
Mean Absolute Error (PCA): 265.805624985228
p-value for the t-test: 0.0014184784125880316

The data might be such that calories can be predicted with pretty good accuracy from the given features. This makes sense, because carbs, protein, and fat all have the same amount of calories every time. The things that make it hard to be a perfect 0 mae for all nutrients is most likely fiber. We see the difference between MAE is pretty big, nearly 10 fold that is. Maybe if we had more data, we might see the PCA MAE decrease a little bit, as the predictions will get better. But with our small data set, we were bound to have skewed results.




Histogram showing how many foods had a certain number of calories in their product.



Imagine you're part of a team, and each member has different skills. Some are great at math, some are superstars in art, and others are sports aces. You want to put together smaller groups (teams) where each group has a mix of these skills. But instead of picking just one skill for a team, you mix a bit of each skill to create a unique blend for each team.

This graph is doing something similar but with food nutrients instead of skills. The graph is a way of showing how each nutrient (like total fat, protein, etc.) contributes to creating two new 'teams' named PC1 and PC2.

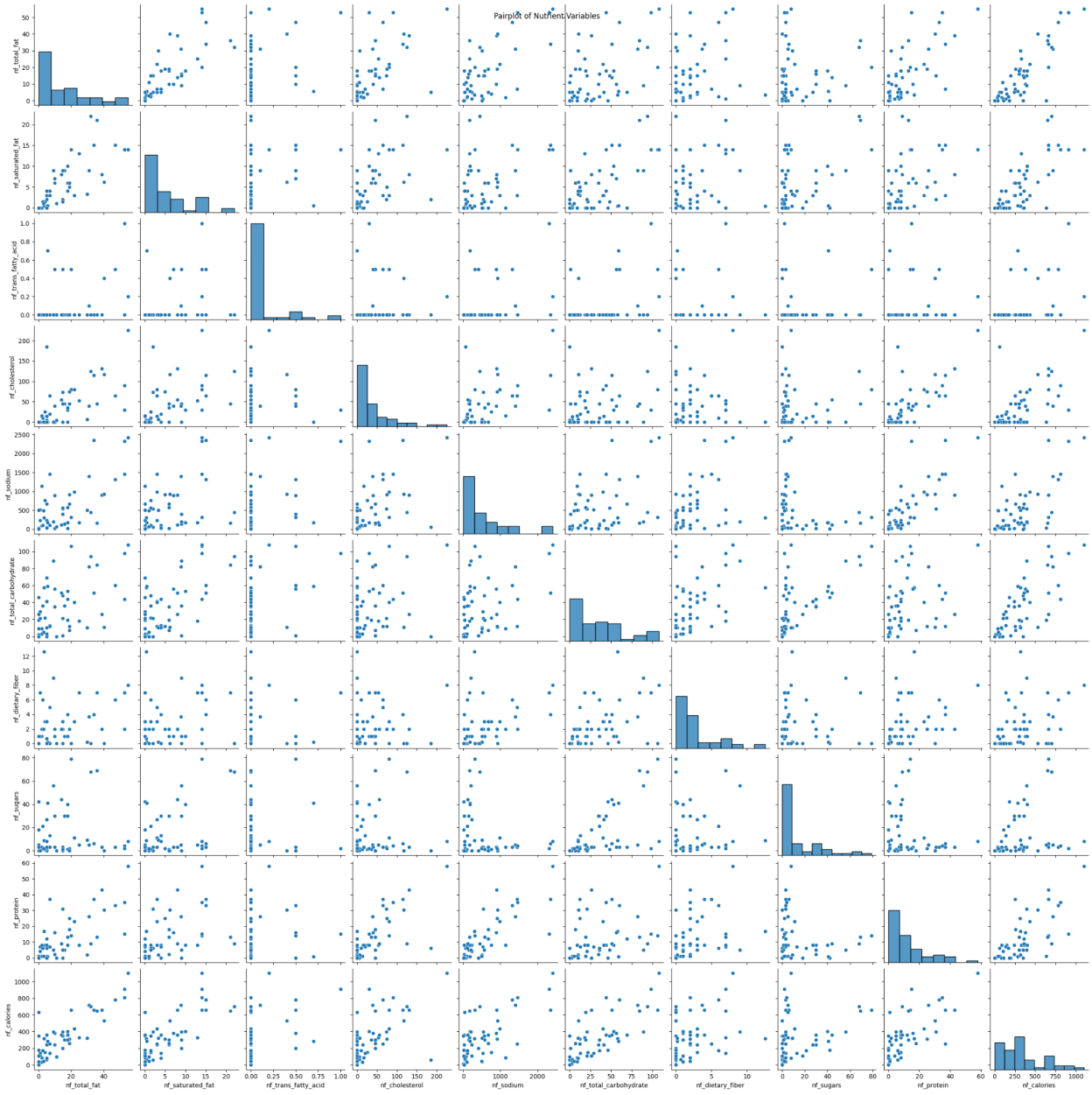
- The blue and brown bars represent how much each nutrient contributes to each team.
- The blue bars show the contribution of each nutrient to Team PC1.
- The brown bars show the contribution of each nutrient to Team PC2.

If a bar is tall, that nutrient has a big say in that team. If it's short, it's more like a quiet member. And some bars can go below zero (into the negatives), which means they kind of have the opposite effect on that team.

For example, you can see that 'nf_protein' has a tall blue bar in Team PC1, which means it's a major player there. But in Team PC2, 'nf_protein' doesn't contribute as much (the brown bar is shorter).

So, when scientists want to make things simpler and not look at every single nutrient individually, they use this method to create 'teams' (PC1 and PC2) that summarize the data in a

way that's easier to manage and understand.




The graph above shows the piece wise relationship between all the variables in our data set.

We see as total fat goes up, so does calories in a linear fashion. We see as sodium and cholesterol increase, so do calories as well, more so than calories increase with sugar or protein.

With the bar graphs, we can see carbohydrates remain steady as the bars are similar height. This shows that most products have carbohydrates. While on the other hand, the trans fatty acid bar chart shows that almost all foods do not have trans fat. This is good to know as trans fat is really bad for your health.

We also see a decent amount of foods have no sugar at all via the bar graphs, and we see a lot of items are low in protein because of the tall boxes to the left in the bar chart of protein. Calories tend to cluster toward the lower end as shown in the scatter plot of dense points to the lower calory end as well as the bar chart showing that all tall (frequency) boxes are under 500 calories.

A key relationship is that we see protein levels rise with cholesterol levels, this makes sense as meat contains cholesterol. And meat is a form of protein.

## Discussion/Reflection

**Question 1:** (Clustering) Using DBSCAN clustering, what are the distinct groups of food items based on their nutrient content, and how do these groups differ from each other?

Due to the limited data sample, the DBSCAN clustering algorithm did not produce informational results. In the future with a larger data sample with grocery items, the clustering should produce more distinct groups between nutrient content. The nutrient content that define calories consist of carbohydrates, protein, and fat, so the graphing shows this relationship despite the lack of distinct clusters. We can see that the fats and protein values are in an almost linear relationship to calories as it is proportional to each other.

**Question 2:** (Recomendation System) Using a K nearest neighbors model, what are the top 5 food items that a user who frequently consumes high-sugar snacks should try based on their past purchases?

The recommendation system produced ten food items from the restaurant sample data that were similar to the snacks that the customer purchased. It was interesting to see the system be able to accurately output similar nutrient content food items like milkshakes and brownies from the sample data. The most interesting recommendations were of the heavy sandwiches and lunch/dinner food items that were suggested with the food items of a sweet variety. It recommends heavy food items to people who consume high-sugar snacks.

**Question 3:** ( GMM Clustering) When clustering food items using a Gaussian Mixture Model, which distinct clusters emerge based on nutritional factors such as calories, total fat, and protein, and how are these clusters characterized?

Clustering with a GMM was intended to find groups of foods that had common nutrition, hoping to group vegetables, meats, and other groups together. Because the foods were only from restaurants, the blending of ingredients to make meals removed the ability for independent

groups to form. High-calorie meals were high in fat and protein, etc. The elbow method and silhouette score are used to determine the number of clusters or groups the foods should be categorized into. The groups were not distinct enough, so the two graphs contradicted each other. The elbow score recommended using many clusters, and the silhouette score recommended using very few. Ultimately, two clusters were chosen to show "healthy" and "Unhealthy" foods. These results are less than ideal, and if I were to run the tests over again, I would have used a data set that would have had less of a positive linear correlation between the variables.

**Question 4:** In a supervised learning model, which variables among brand name, item type, ingredient statement, total fat, saturated fat, trans fat, cholesterol, sodium, total carbohydrate, dietary fiber, sugars, protein, vitamin A, vitamin C, calcium, iron, and potassium have the strongest relationship with calorie count?

Individual nutrition facts were used in a regression analysis to determine which of them had the strongest correlation with calorie count. After dropping the brand name and item type because all of the values were unique, the final variable swerve was put to the test. The resulting graph showed that fat, carbs, and protein were the leaders, which makes perfect sense because the addition of these is what makes up a calorie. That said, the fat was weighted the heaviest of all of these variables, and a variable that ended up having an inverse relationship was sugar. This could be purely because it's not used in the calculation of a calorie, so it's being thrown out to make the model work, but it could also be because some of the high-sugar foods like sweet iced tea or dressing aren't high in calories but very high in sugar. If I were to do this analysis again I would have left out the variables used in calorie calculation so I could see if other variables can make good predictors in order to learn something new.

**Question 5:** Which brand has food items with the highest average nf_sodium content and how does it compare to the average nf_sodium content of other brands?

The analysis is effective because it allows for a direct comparison of the average nf_sodium content across different brands. By identifying the brand with the highest average nf_sodium content, and visually representing the data through graphs, it becomes easier to understand how that brand compares to others. The data made sense as a pizza place ranked a top the sodium charts. In comparison to the average sodium content of other brands, the pizza place contained about 4 times the average sodium as the average foods.

**Question:6** When comparing a model using PCA on all the nutrient variables (calories, total fat, saturated fat, trans fat, cholesterol, sodium, total carbohydrate, dietary fiber, sugars, protein) in

the dataset and retaining enough components to keep 90% of the variance, to a model using all the nutrient variables, how much of a difference is there in mean absolute error when predicting calorie count?

The analysis is effective because it addresses the trade-off between using all nutrient variables and reducing dimensionality through PCA. The use of PCA allows for a more concise set of features while retaining a substantial amount of the variance. By comparing the MAE of the models, one can understand the impact of dimensionality reduction on the predictive performance for calorie count. Although the MAE was about 10 fold higher with the pca versus all of the nutrients, we can conclude that the PCA MAE would of been much better with a larger data set.