Capstone Project 1 - Final Project Report

Nutritional Analysis of the Products in the Supermarket for popular Keto Diet Meal Planning

Asha shravanthi Pidathala
Springboard Capstone Project 1
Jan 6th 2020 Cohort
Problem Statement

The ketogenic diet (or keto diet, for short) is a low-carb, high-fat diet that offers many health benefits. It involves drastically reducing carbohydrate intake and replacing it with fats. This reduction in carbohydrates puts your body into a metabolic state called ketosis. Low-carb diets have been controversial for decades. Some people assert that these diets raise cholesterol and cause heart disease due to their high fat content. However, in most scientific studies, low-carb diets prove their worth as healthy and beneficial. There are many studies ongoing to determine the long term health effects of Low-carb/Keto diets and we will know only in future how that information will unfold the potential uses/risks of these diets. The bottom line is to not cut back on nutritious food for the purpose of weight loss. With this project I hope to distribute the various products available in the supermarkets based on its nutritive values and help the customers make an informed and smart decision when picking the meals for Keto dieting.

Dataset

The dataset was obtained from Open Food Facts downloaded as a tsv file named ‘en.openfoodfacts.org.products.tsv’. Open Food Facts is a free, open, collaborative database of food products from around the world, with ingredients, allergens, nutrition facts and all the tidbits of information we can find on product labels.

Data Exploration and Cleaning

1. **Data Cleaning**

   All the essential libraries were imported and the raw data file read into a dataframe for further analysis. Exploring the shape and contents of the dataset revealed that it has 356027 rows and 163 columns. Using the head(), tail(), and info() commands, I explored the contents of the dataset and noticed there are a lot of Nan entries. Using isnull() and dropna() on the entire dataset would have not worked in this case as almost all columns had some Nan entries. I first started by visualizing the missing values in the dataset. Any column with more than 70 percent of missing data will be dropped to avoid misleading results. Two ways were demonstrated to remove the columns with more than 70 percent missing values, one was **df.dropna(thresh=105620, axis='columns')** and the other was **df.loc[:, df.isnull().mean() < 0.7]**. With thresh you can mention the cut off number of Nan entries you want i.e; anything more than 105620 Nan entries columns will be removed. The second way is much easier where you are using the mean of isnull() entries and anything less than 0.7 which is 70% you keep it and the rest are removed.

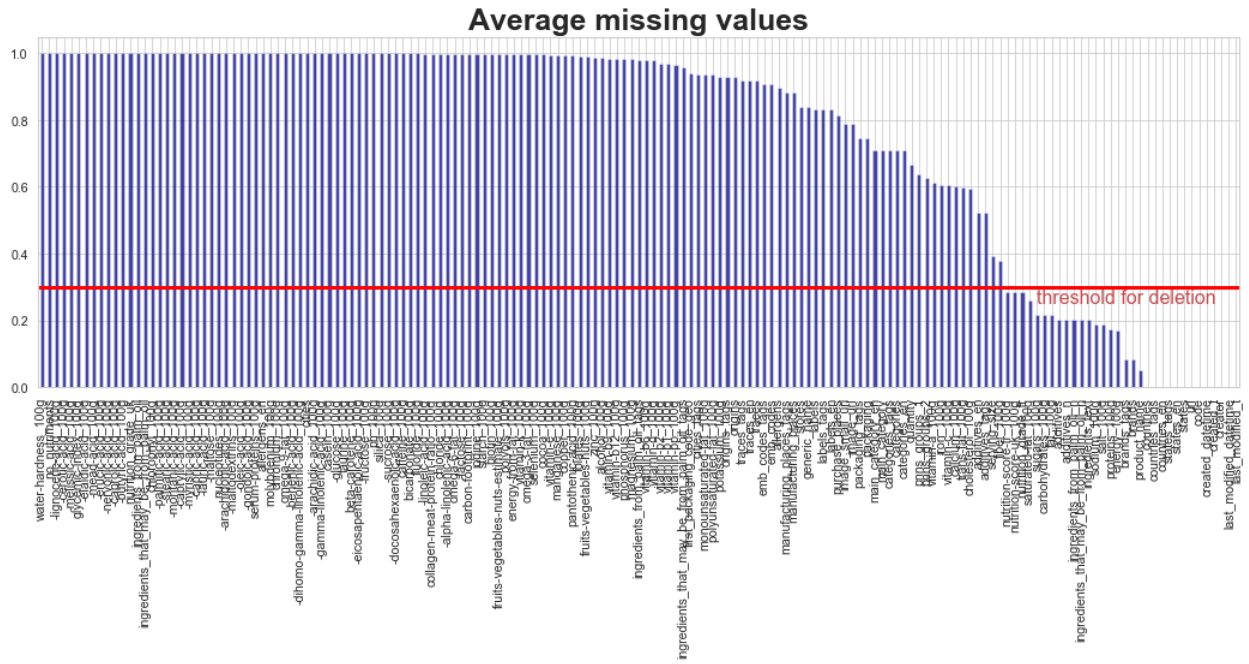

Figure 1: Mean null values from all columns in the dataset

2. Data Exploration

Removing the columns with more than 70 percent missing values dropped the columns from 163 to 45 i.e; we dropped 118 columns. Using info(), head(), and describe(), the new dataset was further explored. This showed that some columns had extremely high mean as compared to the min value, which indicated that there were outliers. Ideally if we do not know the dataset very well, it is not recommended to remove any values as outliers. However, in this dataset there are some very obvious outliers which correspond to bad entries. Using boxplot visualization, we see there are four bad entries with really high values. These four entries along with one more entry from energy (87000 kcal still very high for 100g of product) were removed as bad entries. We further explored the dataset by checking the countries present. USA and France are the countries with most products in this dataset, the next thing to do is check the brands represented in this dataset. Since the USA and France are the most represented countries in this dataset, the most frequent brands are French and American supermarkets. This capstone project will be focusing on the brands available in the USA. Upon filtering and sorting the brands available in the US, you see that Meijer has the greatest number of products, hence Meijer will be used for further analysis in this project.

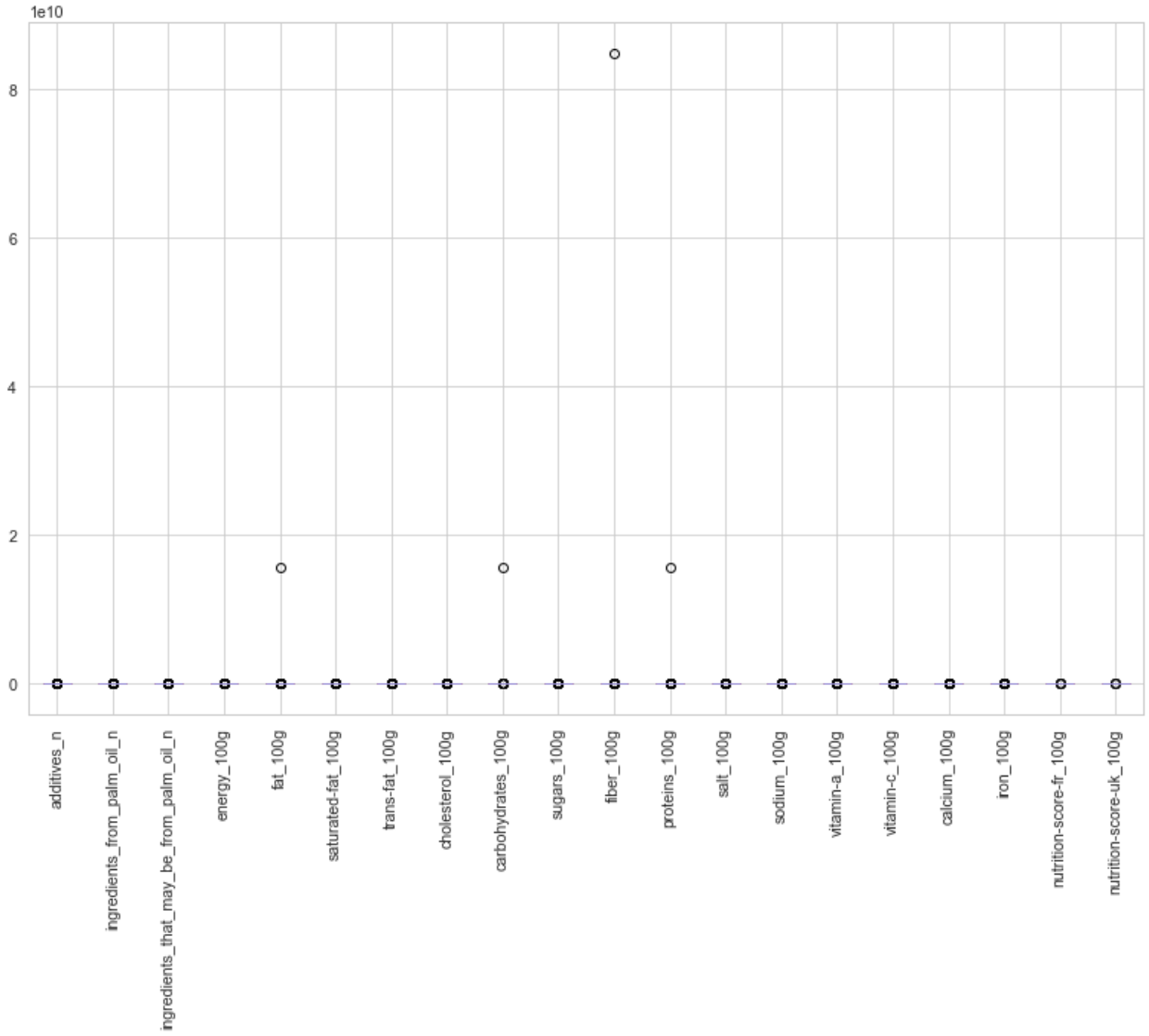

Figure 2: The bad entries from the dataset represented by the boxplot

Data Storytelling

A new dataset is prepared containing only Meijer products from the USA. Data storytelling will focus on preparing the nutritious Keto diet plan with Meijer brand only. Exploring this new dataset reveals some missing values, which can be filled with fillna() so that the dataset is ready to be used for some visual storytelling. By preparing a heatmap of Pearson's correlation coefficients between all the columns from the meijer dataset, one can visually see the columns that are strongly correlated.

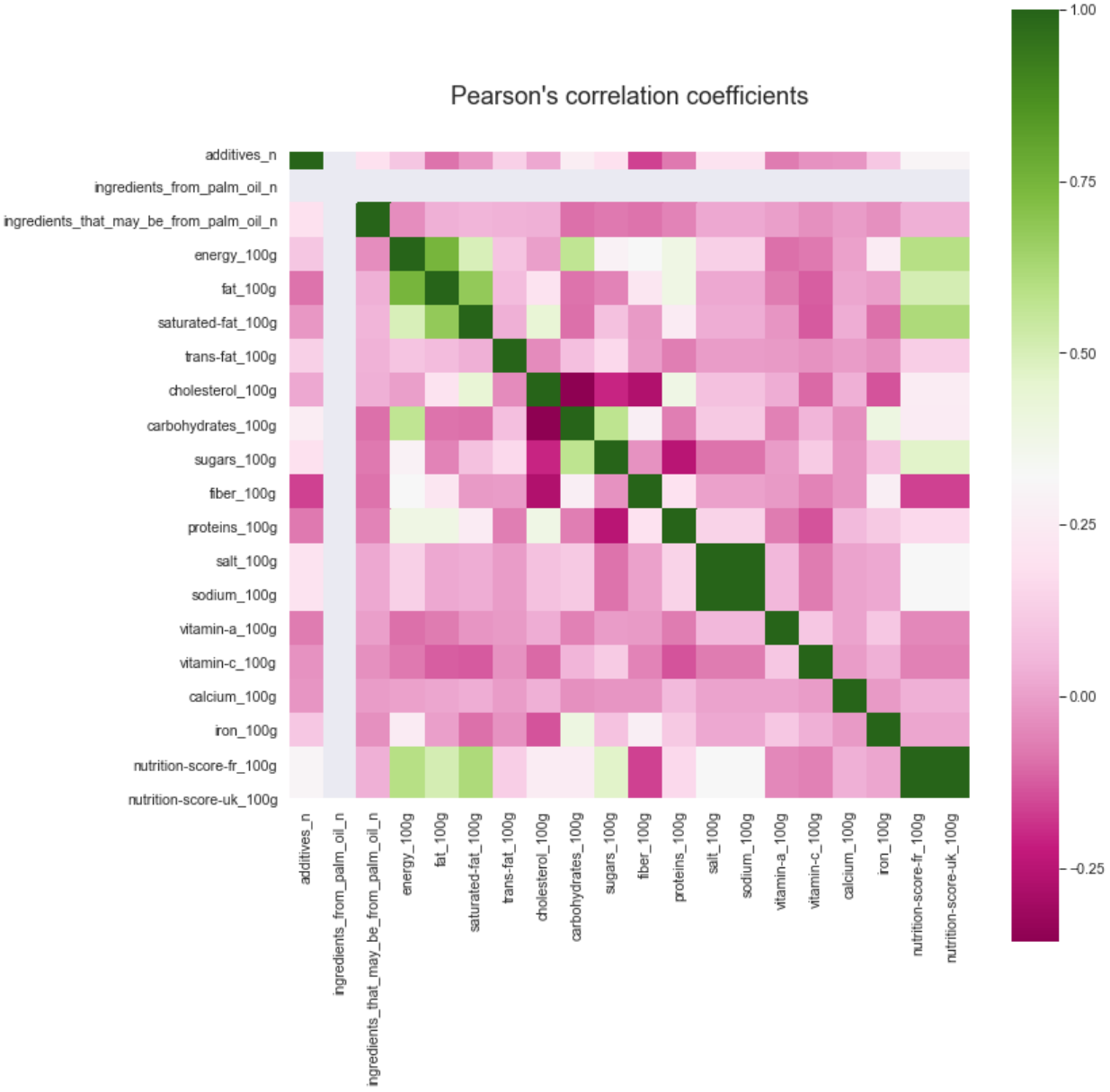

Figure 3: Pearson's correlation coefficients of meijer brand products

From the correlation heatmap, we can see a strong correlation between many features:

- ❖ Sugars and carbohydrates
- ❖ Carbohydrates and energy
- ❖ Fat, saturated fat and energy

And a strong correlation between the nutrition score (FR and UK) and energy, saturated fat, fat, and sugars. This shows that the scores are given based on the amounts of fat, calories and carbohydrates in the product. Let's check the type of relation between the strongly correlated features.

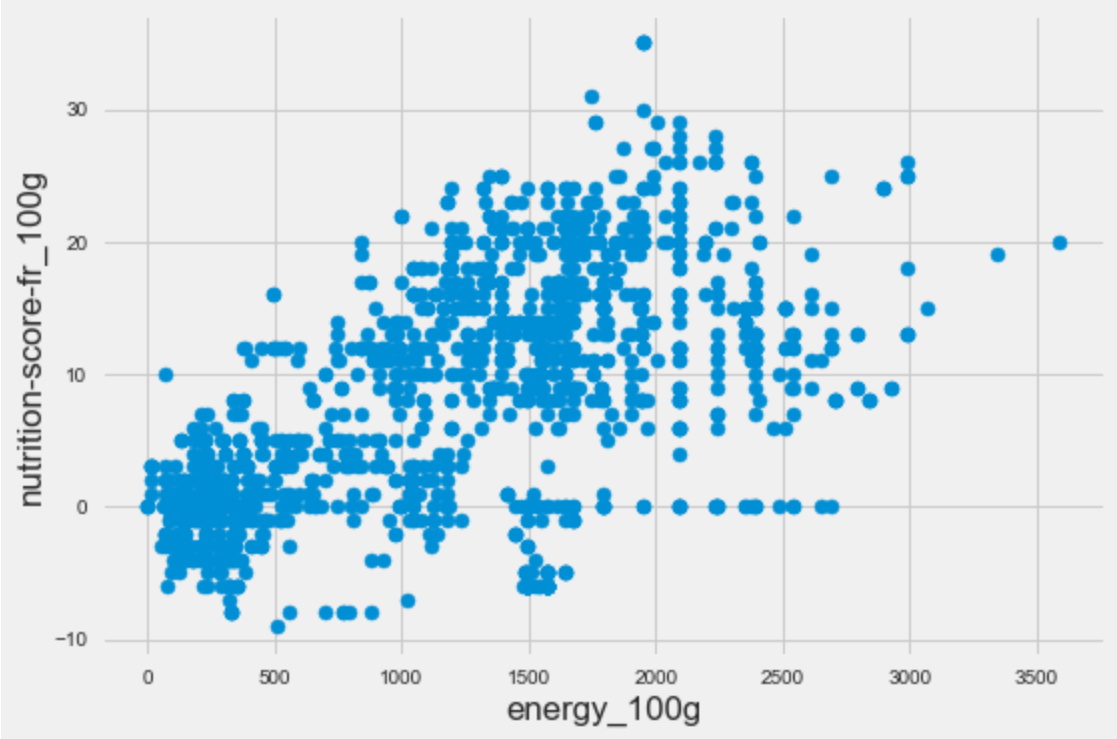

Figure 4: Scatter plot of energy and nutrition score

We can see a pattern in this scatter plot, the more energy the product has, the higher score it gets. Also, many high calorie products have a 0-nutrition score.

Statistical Methods for Data Analysis

As part of the hypothesis testing, we will be analyzing the top three brands. We will be comparing the nutritional scores of the products from top three brands to see if the analysis using just the Meijer products can be easily applied to other popular brands. Separate datasets were prepared for Kroger and Great value products and missing values filled to be used for hypothesis testing. Both frequentist inference and bootstrap inference methods were used to compare the other two datasets with Meijer dataset.

The hypothesis testing between Meijer and Kroger brands proved that there is no difference between the nutritional scores between the two. However, the frequentist method on the Meijer and Great Value brands showed that there is a difference between the two brands in terms of their nutritional scores (small single dataset). The bootstrapping method with 10000 replicates proved otherwise, and hence it can be concluded that the analysis I will be doing using Meijer brand dataset can be easily applied to other popular brands. We still need to consider that the original dataset is very large and unstructured, and having so many Nan values did not help for an easy and clear analysis.

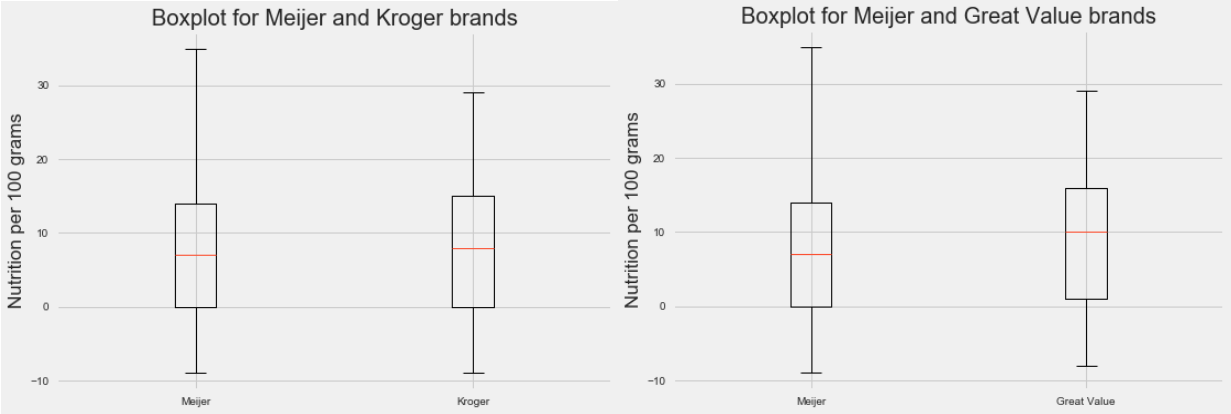

Figure 5: Boxplots comparing nutritional scores between Meijer, Kroger, and Great Value brands

The histograms of the bootstrapped replicates of Meijer and Kroger brands showed that the observed difference in mean value is well within the 95% confidence interval. This is not the same with Meijer and Great Value brand products. Even though the observed difference in the mean value is within the range of values from the replicates, it is outside the confidence interval which shows the uncertainty with comparison between Meijer and Great Value products.

Keto shopping guide in the Meijer Supermarket

While everyone's body and needs are slightly different, the diet typically translates to:

- 60 to 75 percent of your calories from fat
- 15 to 30 percent of your calories from protein
- 5 to 10 percent of your calories from carbs.

For a 2000-calorie diet, this translates to about 165 grams fat, 40 grams carbohydrate, and 75 grams protein. Because it lacks carbohydrates, a ketogenic diet is rich in proteins and fats. This is just an overview of the products that might be suitable for a Keto diet meal plan.

Let’s further shrink the Meijer dataset by keeping only 7 features that we are interested in preparing a Keto diet plan. We are keeping the columns 'product_name', 'energy_100g', 'fat_100g', 'saturated-fat_100g', 'carbohydrates_100g', 'sugars_100g', 'proteins_100g'. The new dataset now has 1631 products and 7 columns. We made a new column ‘total_fat’ by adding the ‘fat_100g’ and ‘saturated-fat_100g’ columns which will be the true depiction of fat in the product.

We will now filter the products based on the percent of calories we need from each category which will be energy/calories < 2000g, carbs < 40g, fat < 165g, and protein < 75g. This will discard the elements that will have high calorie content. We have 832 keto

products in Meijer supermarkets in the United States. Visual representation of the filtered keto products shows the good distribution of the components in accordance to our keto diet meal plan. Most products have calories in between 0-500, carbohydrates around 0-20 g, protein around 0-25 g, and fat around 0-20 g, 40g, and 50g. All these are in accordance with the Keto dietary requirements.

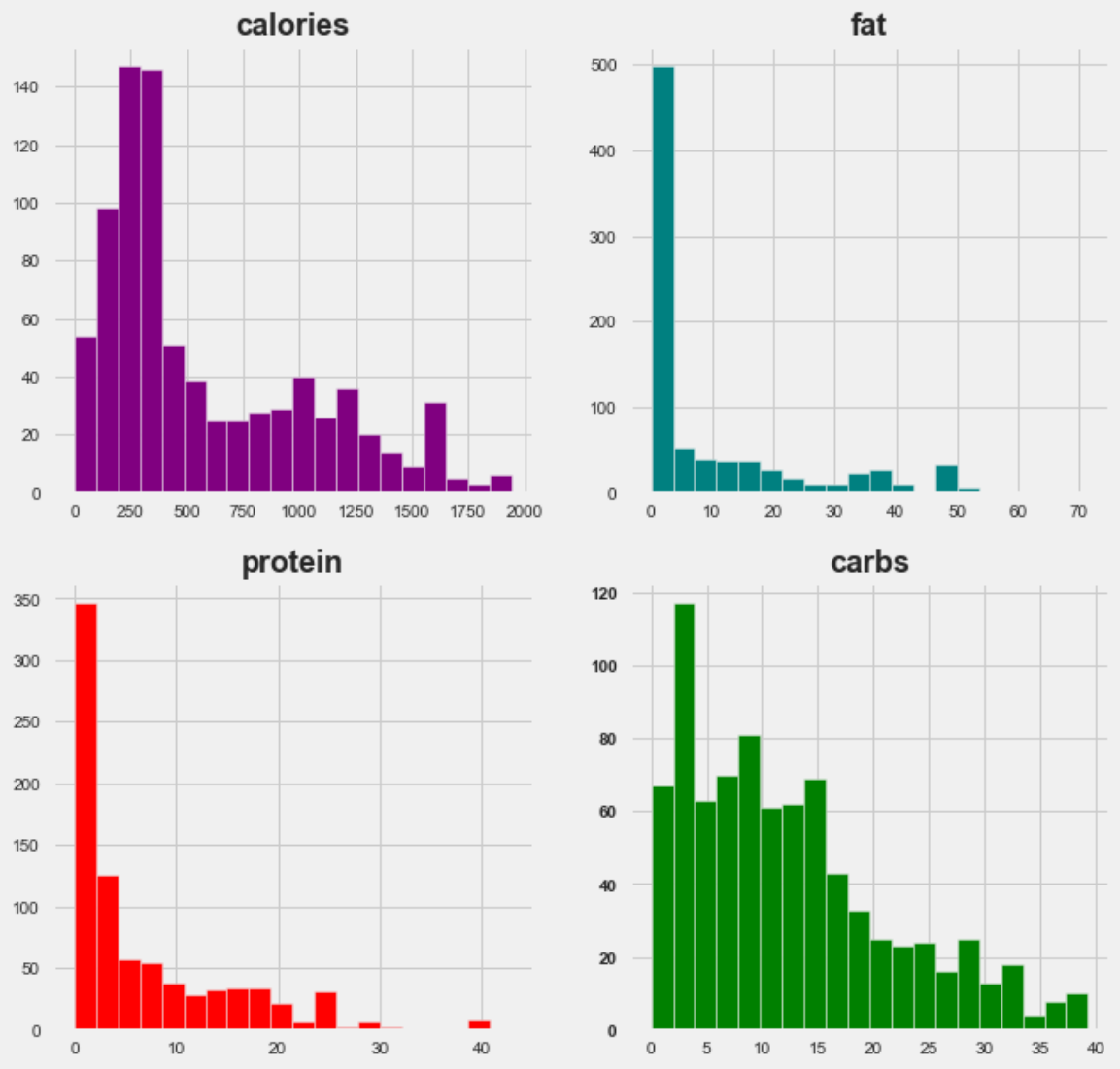

Figure 6: Histogram of components from filtered keto products

The filtered data frame was further separated into Low, medium, and high products based on energy, fat, protein, and carbohydrates content and made a WordCloud of these products based on calorie content.

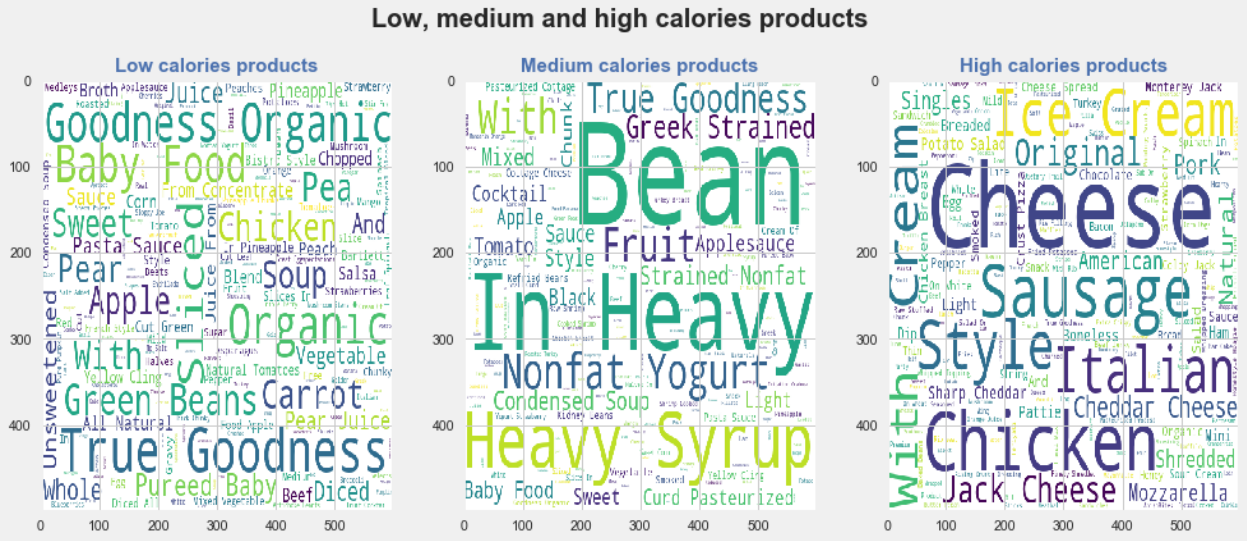

Figure 7: WordCloud of Low, Medium, and High calorie products

From the WordCloud figure, it is very evident that products like fruits and vegetables are low in calories whereas products like cheese and meats are high in calories. To further visualize each macronutrient and what percentage of products are useful for Keto diet planning, a visual representation using pie chart was generated.

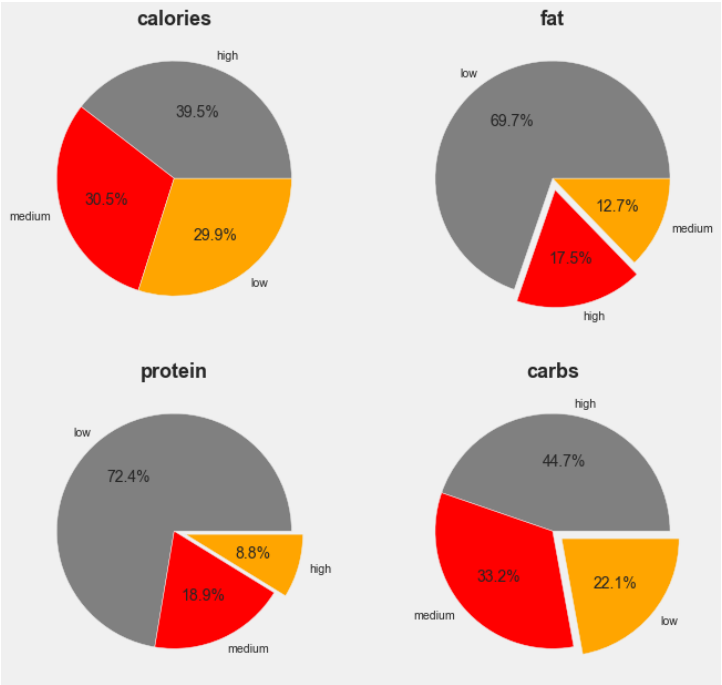

Figure 8: Pie chart representing % of Low, Medium, and High products for different micronutrients

Machine Learning - Clustering:

Preprocessing

We start preparing the data for Machine Learning. We have separated the columns of interest containing numeric values required for fitting and predicting the clusters. First thing to do is scale and standardize the data using StandardScaler(). Standardization is useful for data which has negative values which is the case here. It arranges the data in a standard normal distribution without changing the original shape of the distribution of the data. We store the scaled data, and further use this for fitting the machine learning model.

KMeans Clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes. The 'means' in the K-means refers to averaging of the data; that is, finding the centroid. The idea of KMeans clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. We calculated the inertias of different k_clusters (1 - 15 clusters) to determine which number of clusters divides the data appropriately so that similar products are grouped together. Inertia is the sum of squared error for each cluster. Therefore the smaller the inertia the denser the cluster (closer together all the points are). This graph is popularly known as an elbow shaped graph, where the number of clusters at which the inertias makes an elbow shape is considered the right number of n_clusters.

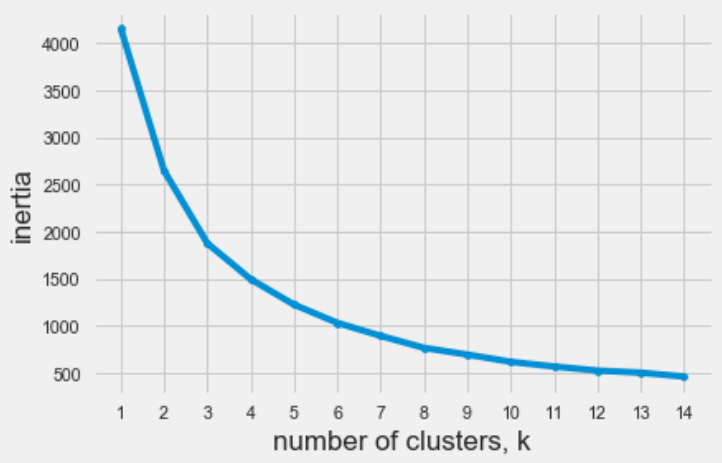

Figure 9: Elbow graph of inertia vs clusters

From the above figure, the elbow shape is not that clear. It is more of a curved graph which makes it difficult to pick the right number of clusters for fitting the model. I chose 6 clusters to begin with a random_state=10 to keep them consistent. The predicted values from fit_predict() of the scaled data was stored as a new column in the 'df_keto_filtered'

data frame. Plotting the clusters as a bar chart shows that we got two big, two medium, and one small cluster.

Since the elbow graph from inertia values did not give a clear understanding of the ideal number of clusters, we determined silhouette scores for different n_clusters. The silhouette score is from -1 to 1 and shows how close or far away the clusters are from each other and how dense the clusters are. The closer your silhouette score is to 1 the more distinct your clusters are. If your score is 1 think of your clusters as perfect little balls that are far away from each other with no miss classification. We calculated and plotted the silhouette score for n_clusters = 2 to 10.

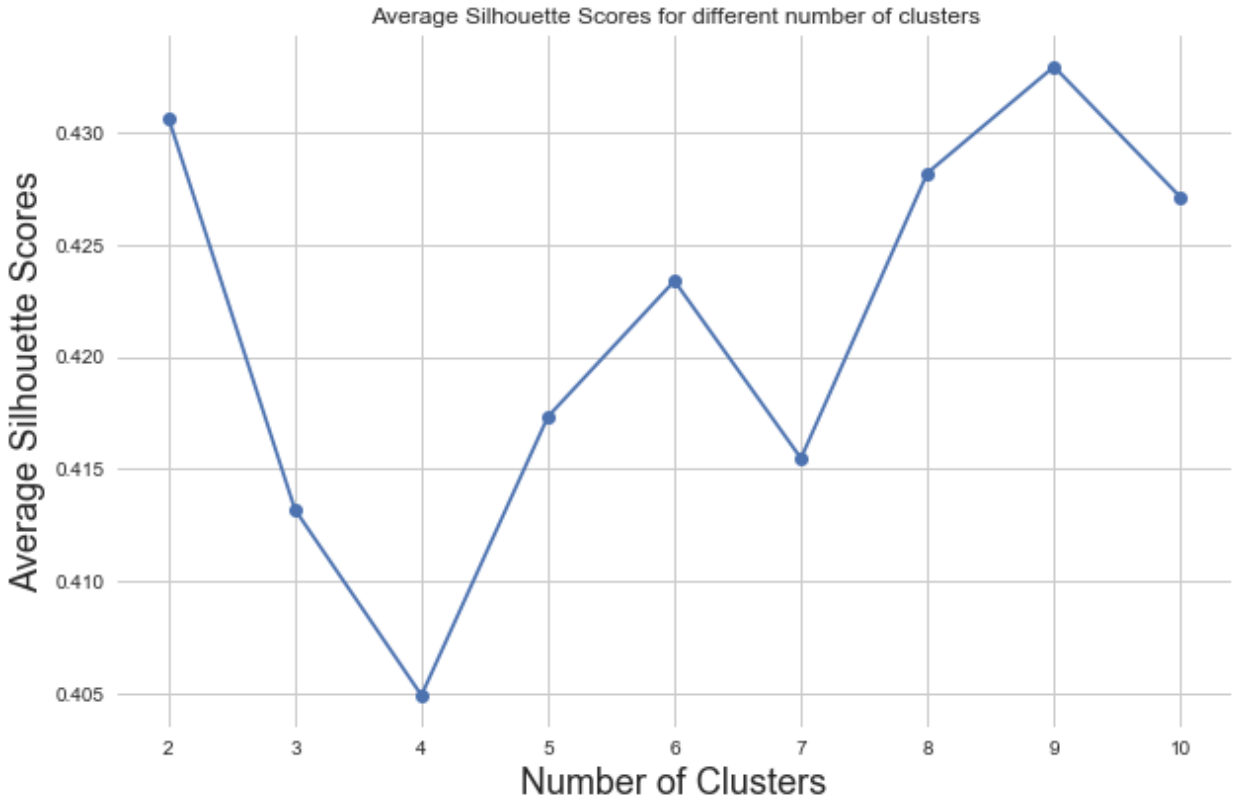

Figure 10: Plot showing silhouette score vs cluster

From the above figure, it looks like n_clusters = 2 and n_clusters = 9 have the highest silhouette score. Using PCA, we tried to plot and visualize the distribution of the points in the clusters for n_clusters = 2 to 10.

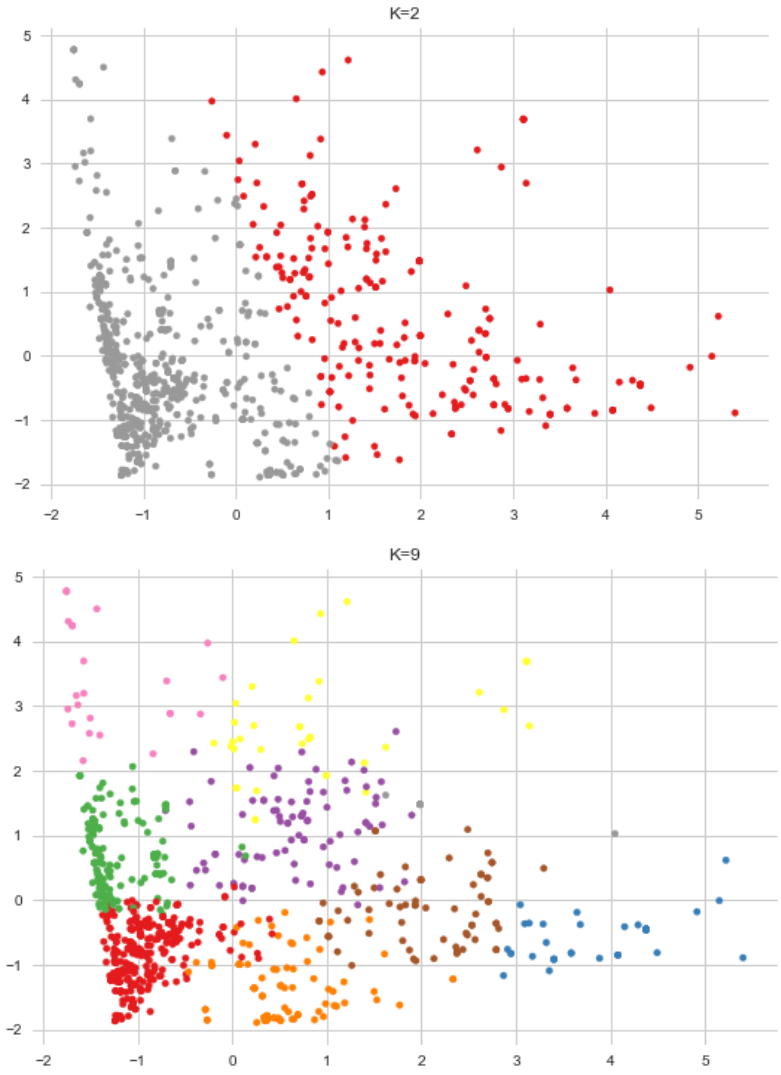

Figure 11: 2-D Scatter plot for n_clusters 2 and 9 using PCA

Looking at the silhouette_score, clusters 2 and 9 show the best score. However, looking at the 2-D scatter plot of the points using PCA shows that the values are too spread out, so choosing 2 clusters to analyze the data will not be a good choice and 9 clusters will be a lot to analyze and draw good conclusions out of each cluster. So the previous choice of 6 clusters also has a decent silhouette score and with fewer clusters it will be easy to analyze and draw conclusions, so we will stick with 6 clusters and use the predicted values of n_clusters = 6 which is stored in column k_clusters.

We make the final data frame taking only the required columns, here we include the predictions from fitting the model to 6 clusters and also the nutritional scores to make conclusions and suggest which cluster will give you the nutritious Keto diet products. We plotted a few columns against each other to see the clustering distribution.

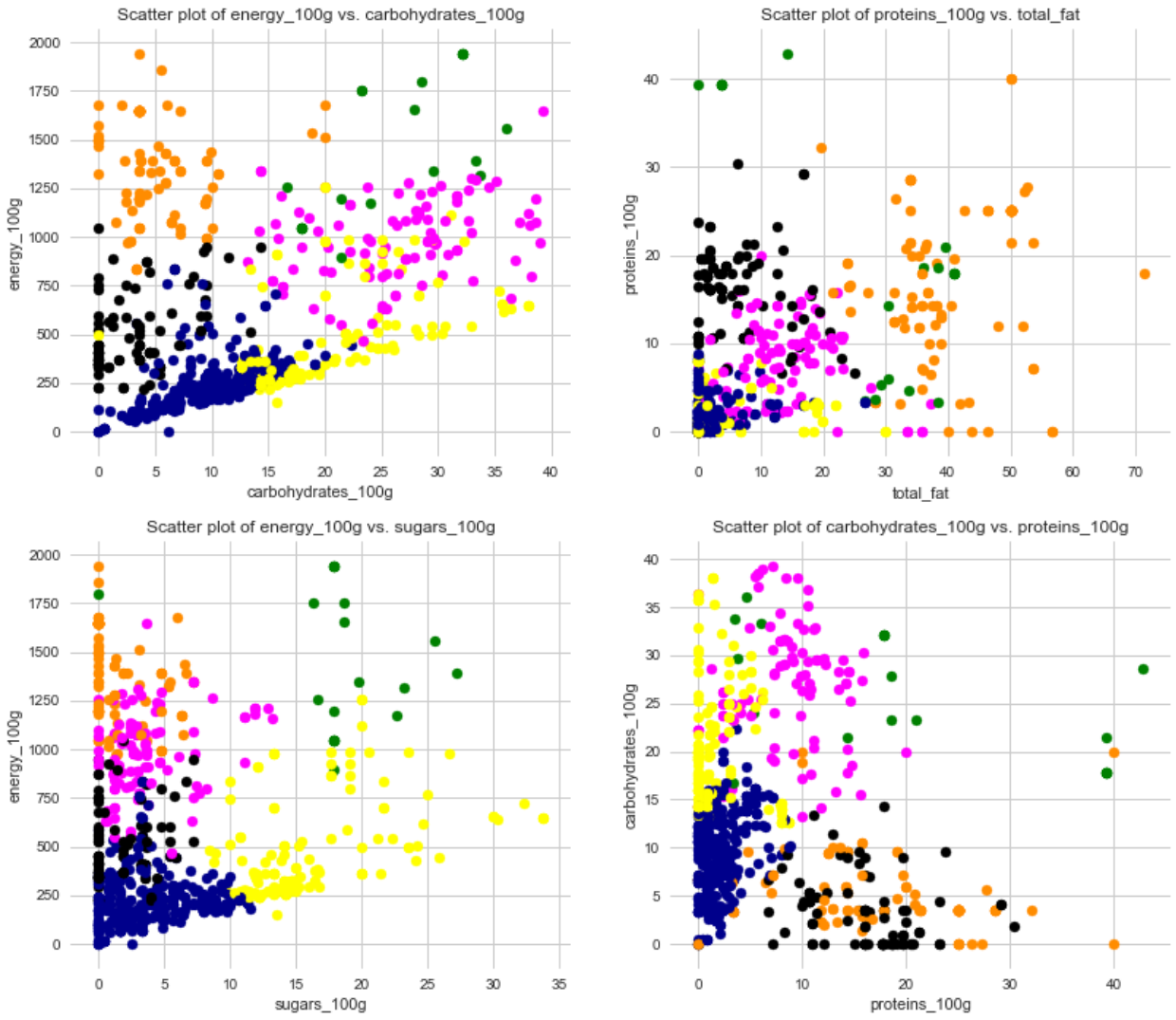

Figure 12: 2-D Scatter plot for columns energy, carbohydrates, protein, fat, and sugars.

We see clear clusters for the plots with ‘energy_100g’ as calories depend on all other macronutrient molecules like protein, carbohydrates, fat, sugars etc. Not seeing clear clusters when plotting the fat, protein, and carbs against each other is something we expect to see as the clustering we did is not based on calories, but rather on the individual macronutrient content.

The main food categories in a supermarket are: Meat, fruits/vegetables, sweets, bread/pasta, beverages, milk/cheese/yogurt. We are going to further analyze the clusters to see if we can find clear food categories in the clusters. We made a WordCloud for each of the 6 clusters with product names.

Figure 12: WordCloud showing products in Clusters 0 to 5

From the above figure, we see that some clusters are mixed while some are clear. Cluster 0 is the biggest cluster and hence has most of the mixed items. From WordCloud we can make assumptions of the food types in each cluster.

Cluster 0: Mixed items (Salads/Vegetables/Soups)
Cluster 1: Snacks
Cluster 2: Cheese/Cream
Cluster 3: Fruits/Juices

Cluster 4: Salad ingredients/some unhealthy options like pizza and pasta
Cluster 5: Meats

Using descriptive statistics, we will analyze each cluster for its individual macromolecules vs. the nutritional score to get a nutritious keto meal. We calculated the mean and median for each cluster and most clusters were good to be picked for Keto diet shopping, cluster 2 was perfect with high fat and protein, low carb and sugar, and good nutritional scores. Cluster 3 is not favourable for Keto diet and hence can be eliminated when considering shopping for Keto.

Mean of all macronutrients from Cluster 2

	Cluster 2 Mean
energy_100g	1374.81
total_fat	40.5786
carbohydrates_100g	4.96991
sugars_100g	1.53296
proteins_100g	17.8239
nutrition-score-fr_100g	19.2222
k_clusters	2

Median of all macronutrients from Cluster 2

	Cluster 2 Median
energy_100g	1393
total_fat	38.1
carbohydrates_100g	3.57
sugars_100g	0
proteins_100g	18.455
nutrition-score-fr_100g	20
k_clusters	2

Figure 13: Descriptive statistics for cluster 2

Conclusions

Cluster analysis based on WordCloud and Descriptive Statistics:

Generally, popular ketogenic resources suggest an average of 70-80% fat from total daily calories, 5-10% carbohydrate, and 10-20% protein. For a 2000-calorie diet, this translates to about 165 grams fat, 40 grams carbohydrate, and 75 grams protein.

If you're on a keto diet, you know that staying and getting into ketosis (the whole goal of going keto), is achieved by eating a higher fat, moderate protein, and low-carb diet. The perfect amount of daily carbs is different for each person; some people can easily get into ketosis and stay there on 50 grams of total carbs per day while others need to stay at around 20 grams of total carbs per day.

The Mean and Median of all clusters are pretty similar indicating good classification of products into each cluster based on its macronutrients.

Cluster 0:
Looks like this cluster has low energy products (200 cal), no significant high fat or high protein products required for keto diet specifically and also has very low nutritional score. However since this group also had low sugars and carbs it is not all bad. This cluster can be used as snacks between the meals.

Cluster 1:
From the WordCloud for this cluster, it looks like this cluster has a mix of good snacks like Beef Jerky, Trail Mix, Cheese, Almonds and also some sweets like Cranberries, Cake, Eclairs, Tiramisu, Custard which makes this cluster a little high on carbohydrate and sugar levels which leads to high energy. But this cluster also has relatively high fat and high protein products which have decent nutritional scores. In this cluster there are some products which are a good pick for Keto diet.

Cluster 2:
This cluster is one of the prefect clusters for Keto diet products. It has all the dietary requirements for a Keto diet. This cluster has high fat, high protein, low carb, and low sugar products with good nutritional scores.

Cluster 3:
This cluster has low fat, low protein products with relatively high carbs and sugars. Considering this cluster has fruits and juices according to WordCloud, this is the right prediction of macronutrients. This cluster is not so good for Keto diet planning and shopping.

Cluster 4:
This cluster has a decent amount of fats and proteins, even though it is relatively high on carbs (still <40g) and has decent nutritional scores. According to WordCloud, some products like pizzas and pasta are making this cluster a little unfavourable.

Cluster 5:
This cluster is high in protein and low in carbs and sugars, considering this has mostly meats according to the WordCloud. This is also one of the good clusters to consider for the Keto because of the high protein content.

In conclusion, if I had to suggest which clusters to pick for a nutritional Keto diet planning, I would mix products from the clusters 1,2,4, and 5. Can consider cluster 0 for some unharmful snacking.
Effectively, k=6 clusters was not the optimum value since we got some clusters that are not clear with mixed products, which means there is still significant room for improvement.

However, in the other clusters, we could see a clear dominance of certain types of food, for example, cluster 5 has mainly: boneless meat, chicken, turkey, ham, bacon, sausages, and shrimp. The clear clusters tend to be the smallest, with fewer products in comparison to other clusters. This invites us to increase the number of clusters in order to decrease the variation between clusters. Maybe going with the silhouette score (n_clusters=9) would not have been a bad option.