DSC 478 FINAL PROJECT

What Makes Food Healthy: An Analysis

The Flying Squirrels

Executive Summary:

Project Goals:

Our goal was to generate insights into what makes foods healthy or unhealthy using various methods like clustering analyses and a classification model. Our chosen dataset has a few features that could be used to classify a food as healthy or unhealthy, and we set our to examine what features contribute to that classification.

Data:

Our data set has 1722 different meals (food items) and 57 features. The data schema is single comprehensive tabular dataset. Features include categories like: very healthy, ingredinents, percent fat, etc.

Methods:

In order to determine what makes a food healthy, we first needed to preprocess our dataset. We performed a correlation analysis to determine if there were any features that should be dropped. Then we trained and tested a classification model (KNN) on the dataset to review our initial results. Afterwards we began running unsupervised clustering algorithms to identify any natural groupings of data, and to identify the most important features in the data. Lastly, we re-ran our classifier to determine whether our clustering analysis proved useful in cleaning the data and identifying the most influential features.

Conclusions:

In conclusion, we have determined that there is no clear way to predict a food's health. Gluten Free foods or Whole30 diet foods, for example, are not indicative of whether a food is healthy or not. We believe this is likely because our target label (in this case health score) might be a subjective one. Different foods and their components are good or bad for different people at different times.

Team Member Contributions:

Alexandria- KNN/Classification Models, Class Presentation
Rudhvish: Agglomerative Clustering, Class Presentation
Nikki: K-Means, Class Presentation
John : DBSCAN and Presentation, Class Presentation

Introduction:

The hunting and foraging days are behind us. Today, there are a plethora of food options available to everyone, everywhere in much of today's global society. All you need is a stomach and a credit card. We've become more disconnected with the food we eat as a result; we don't stop to think as much about what our body needs. Lifestyle changes, slowing down and spending time preparing meals and eating them more slowly, can help you reconnect with your food. However, for many a busy life is a reality that they must face. So, to help us all understand how food impacts our bodies, the flying squirrels' team has decided to use data science and machine learning to re-discover what makes a food healthy.

Often, we think of a food or meal as one simple thing, but a food is often a complicated mix of nutrients, ingredients, and drugs. When foods have lots of nutritional features in common, they can be grouped into vegetarian food items or gluten free food items. These food groups are something our data group will examine. Like many people we are interested in simple solutions. For instance, It would be really nice to say vegetarian food is always more healthy than non-vegetarian food. We don't think it's likely this will be a finding, but it's worth exploring. It's more likely, based on current food science research, that we will see negative health impacts associated with sugars and fats. According to Gillespie KM, Kemps E, White MJ, and Bartlett SE in their paper called The Impact of Free Sugar on Human Health-A Narrative Review, there are studies that "suggest a negative effect of excessive added sugar consumption on human health and wellbeing." This finding specifically calls out added sugars. The difference between added sugar and sugar is not well defined in our data set for food items, nonetheless, we will be interested in finding the correlation, or lack thereof, between sugar and food healthiness.

Our plan of action in this endeavor is to first, preprocess the data by dropping unnecessary or highly cross correlated features. Then run a classification model on our preliminary set to set a baseline target for the feature we decide to use as our measure of healthiness. This could be a feature like health score or very health (T/F). Then, we will perform various unsupervised methods to look for patters in our dataset. This will help us see if the different types of food form any groups naturally. In addition, we will run test scenarios on our unsupervised clustering algorithms with different sets of features. This will help us see which features may contribute the most to how foods are naturally separated from one another. Finally, we will use what we have learned so far to run a classification model, compare the results to our baseline, and determine if we found what makes a food healthy.

Methods:

Pre-Processing:

We initially reviewed the dataset for features that would not serve our data. Features like url links should be dropped and cleaned out. In addition, we looked at finding a correlation with health Score to see if it makes sense to run a KNN classifier against it. When we did run a classifier with all the features that we decided to keep we determined that targets like healthScore may yield results but need improvement.

K-Means:

In this section of the project, we tried running many different data frames through K-Means. Generally speaking, we would run a data set with select features and determine whether the data could cluster in meaningful ways. If it could, we would look into it further for insight. The way we evaluated each clustering strategy, was by plotting silhouette scores and SSE over k values. A high silhouette score would indicate high separation and cohesion for each individual food item within each cluster. SSE scores indicate how tightly the items in a cluster are packed around the centroid. Generally, low SSE scores and high silhouette scores are good, however, when they are too high, it's likely you have stumbled upon an outlier. This is why size was always examined in addition to the aforementioned scores. If your clusters aren't sized fairly evenly, or you have a cluster with just a few points, it's likely that cluster isn't very meaningful when trying to find insights for the entire dataset. It's only clustering around a specific value or outlier.

So, in our first trial we dropped id, title, ingredients, alcohol, caffeine, image, spoonacularsourceurl, dishtypes, cusines, gaps, spoonacularscore, and healthscore. We dropped these variables because they were either so sparse they we deemed unhelpful in preprocessing, or they contain data that is to unlike the other features. For instance, SourceUrl is a web address, which would have no bearing on how healthy a food it or not. Health Score, in this case, was dropped because we were experimenting with the idea that this could be our label. This run also included many categorical variables: whole30, ketogenic, lowfoodmap, verypopular, veryhealthy, sustainable, dairyfree, glutenfree, vegan, vegetarian. These variables were true false, and they were converted to 1s and 0s.

Next, we normalized the data using the Min-Max method. This brought all the data points to a standardized value between 0 and 1. This was chosen instead of the Standard Scaler method in this case because we have categorical (1/0) values in this run. To retain the true false information for those variables we must use min-max scaling.

Lastly, we clustered in K-Means over different values of K (clusters) and evaluated our results. We repeated this process many times to experiment with features to drop or include and standardization methods. In subsequent trials we tried:

1. Run a baseline K-Means experiment with features identified in pre-processing.
2. Dropping all categorical features.

3. Conducting a PCA analysis to find the most important numerical features and running K-Means on the features that accounted for 99% of all variance. At the same time normalizing the data with the standard scaler method because there were no categorical features.
4. Dropping health score, very popular, and very healthy.

For each of the above experiments graphs and insights were gathered. See results section for those insights.

DBSCAN Clustering Analysis:

Applied (Density-Based Spatial Clustering of Applications with Noise) to detect food clusters based on nutrient content and health-related features. Unlike K-Means and Agglomerative Clustering, this allowed us to try and identify natural groupings without assuming a fixed number of clusters.

Hyperparameter Selection was used for DBSCAN implementation. Used the K-Distance Graph to determine optimal eps value. The elbow graph suggested eps = 6 as an appropriate choice. Min samples was set to 5 meaning a minimum of 5 points on the pre-processed dataset. Used PCA to visualize clustering results. Much tuning was done to see if there are separate clusters with distinction.

Agglomerative Clustering:

Agglomerative clustering is a hierarchical clustering technique that groups data points into clusters based on their similarity. It begins with each data point treated as an individual cluster, then progressively merges the closest clusters until a specified number of clusters or a stopping criterion is reached. This approach is particularly useful when the number of clusters is not known in advance and allows intuitive interpretation through dendrograms.

In our experiments, we explored various aspects of agglomerative clustering to effectively analyze the healthiness of food recipes. Specifically, we experimented with different linkage methods—including Ward, complete, average, and single linkage—to determine the best way to measure similarity between clusters. Additionally, we tested different numbers of clusters to identify meaningful groupings within our recipe dataset.

To ensure accurate and meaningful results, we standardized the selected key nutritional, dietary, and economic features using StandardScaler. This step guaranteed equal contribution of all features to the clustering process. Ultimately, by refining our feature set and systematically varying the clustering parameters, we aimed to uncover clear, interpretable clusters that reflect meaningful patterns in food healthiness and dietary profiles.

Classification:

In this section, we wanted to see if the classification models (KNN, Random Forest, and Decision Tree) could accurately determine if a food item was healthy or not. We focused on the *healthscore* feature and changed it to healthy (1) or not healthy (0) target, with 30 being the splitting score. We decided to use 30 as the splitting score because it was the mean of the *healthscore* feature. For KNN, we tuned the model using PCA and feature selection. We used the clusters from K-Means in a Decision Tree to see if it could predict whether a food item was healthy or not.

Results

K-Means Experiment 1:

The first of our K-Means experiments was dropping Health Score and other features identified as unnecessary or unhelpful from preprocessing as defined in the methods section. The results of the clustering analysis are as follows:

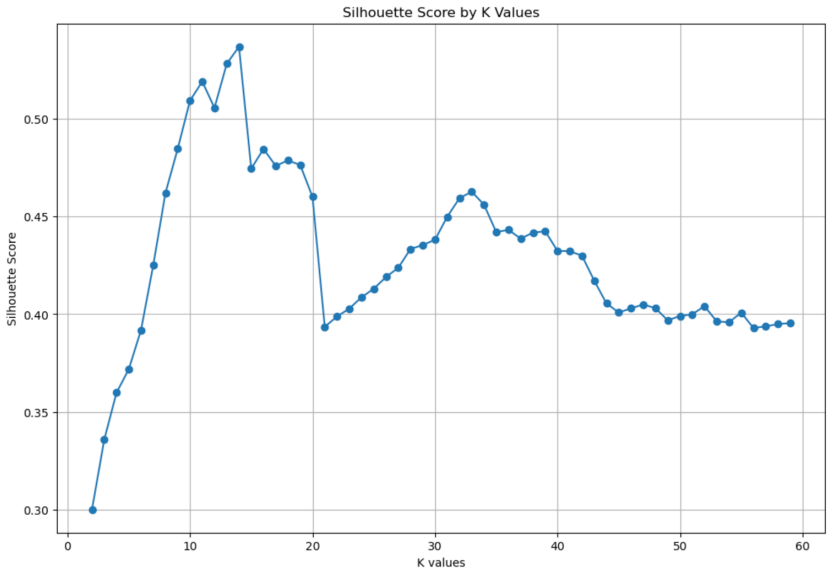


Figure *Experiment 1 - Silhouette Scores*

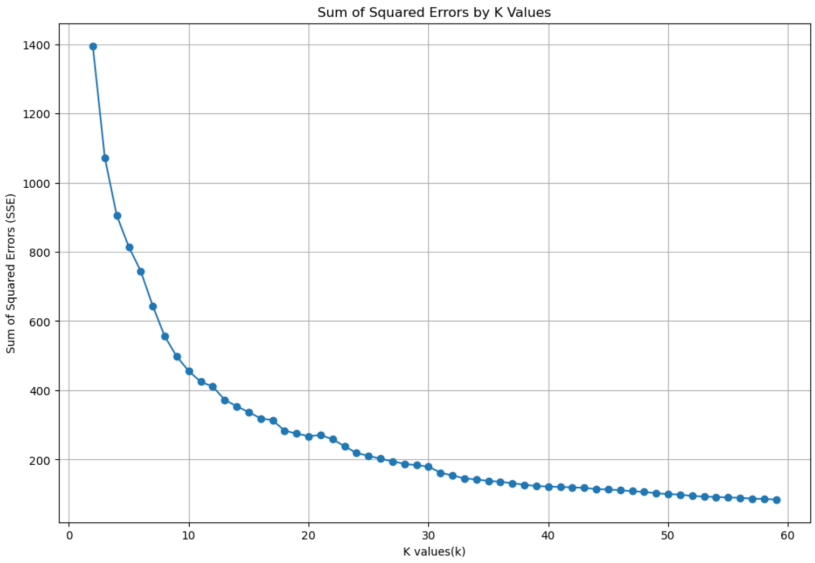


Figure *Experiment 1 – SSE Scores*

Table Experiment 1 – Cluster Sizes when k = 11

```
Size of Cluster 0 =  415
Size of Cluster 1 =  131
Size of Cluster 2 =  344
Size of Cluster 3 =  38
Size of Cluster 4 =  145
Size of Cluster 5 =  111
Size of Cluster 6 =  78
Size of Cluster 7 =  72
Size of Cluster 8 =  85
Size of Cluster 9 =  152
Size of Cluster 10 =  151
```

The key takeaway from this experiment was simply that we could cluster our data successfully. We didn't see a strong indication of where a kick in our elbow plot (SSE figure) would be, but that was ok because we were going to try more experiments. We chose 11 because in addition to having what resembled an elbow in our SSE plot, it had a relatively high silhouette score. Our hope was to increase our silhouette score in future experiments.

K-Means Experiment 2:

In this experiment we dropped all categorical values from the data frame and applied the standard scaling method instead of min max scaling. The results were as follows:

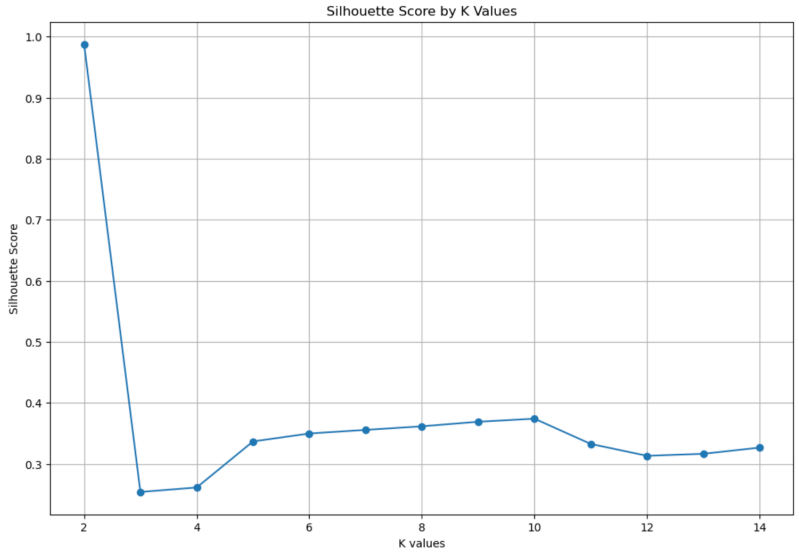


Figure *Experiment 2 - Silhouette Scores*

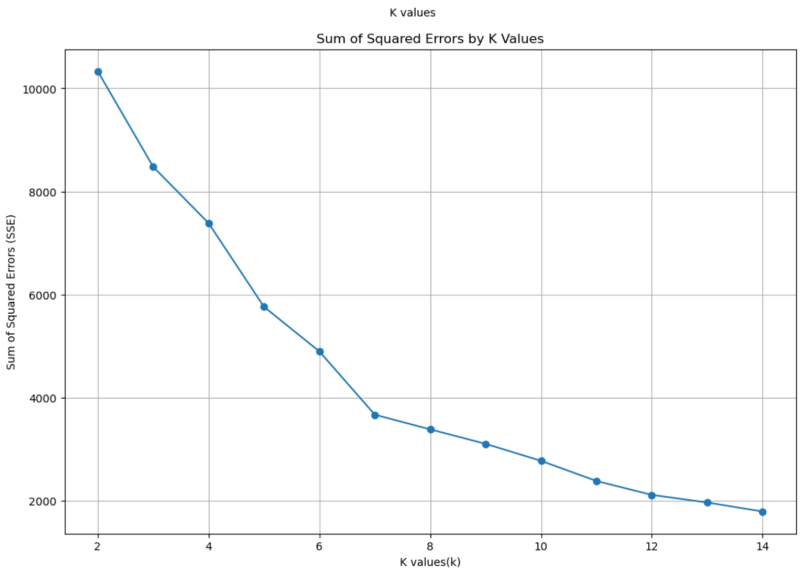


Figure *Experiment 2 – SSE Scores*

Table *Experiment 2 – Cluster Sizes*

```
Size of Cluster 0 =  1721
Size of Cluster 1 =  1
```

Key takeaways from this experiment were that these features alone seemed to all follow each other. With a change in K from 2 to 3 we saw silhouette scores and SSE scores go from suspiciously high to suspiciously low. Relying on these features alone may not be the best path forward.

K-Means Experiment 3:

In this experiment we wanted to ensure that noise in the data generated by some of our non-categorical features wasn't inhibiting our ability to cluster. So, ranked features by feature importance and selected the features contributing to the top 99% of the variance in our target.

Table *Experiment 3 – Feature Rankings*

Feature importance

	feature	importance value
9	Saturated Fat/g	4.23
8	Fat/g	4.08
25	Zinc/mg	4.01
15	Vitamin B3/mg	3.85
29	Vitamin B5/mg	3.85
19	Vitamin B2/mg	3.82
21	Vitamin B1/mg	3.70
12	Cholesterol/mg	3.67
30	Vitamin B6/mg	3.63
23	Potassium/mg	3.61
27	Magnesium/mg	3.58
16	Selenium/µg	3.56

Of the features in the Table above, the top 6 were determined to cause 99% of the variance. So, we ran K-Means only including those 6 features and got the following results.

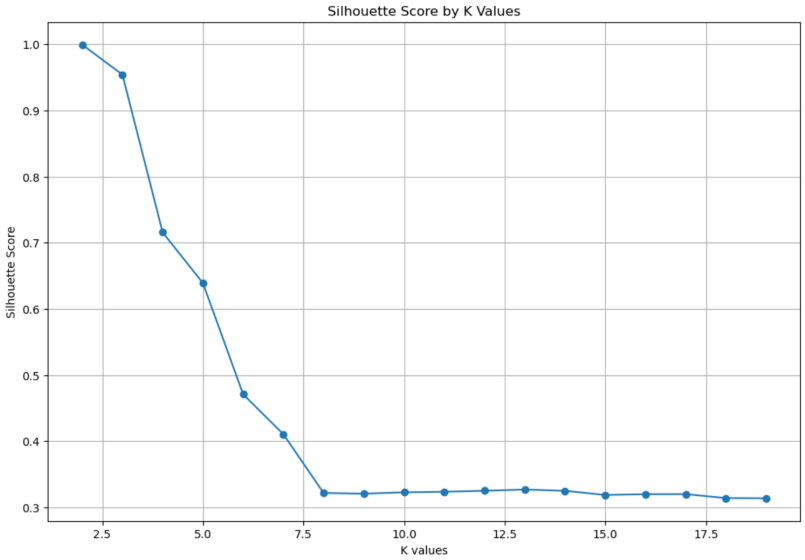


Figure *Experiment 3 - Silhouette Scores*

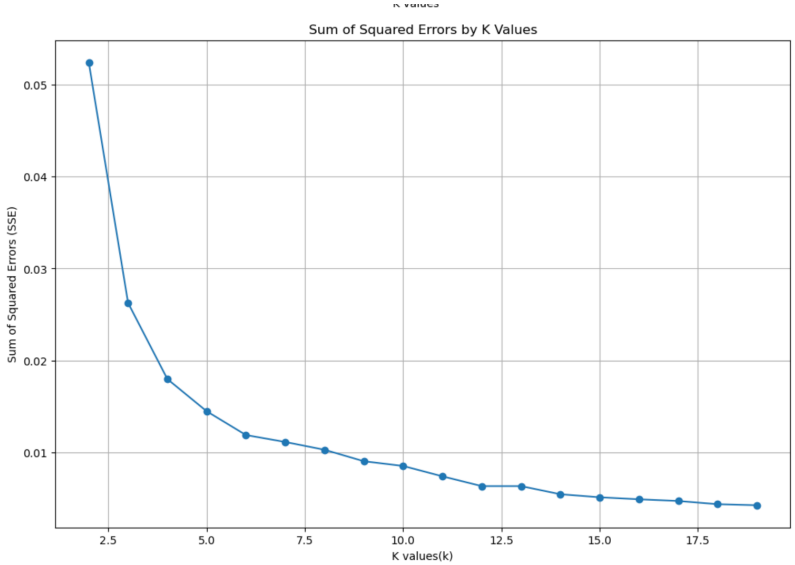


Figure *Experiment 3 – SSE Scores*

Table *Experiment 3 – Cluster Sizes*

```
Size of Cluster 0 =  1610
Size of Cluster 1 =  1
Size of Cluster 2 =  3
Size of Cluster 3 =  108
```

It looked like this might be the best clustering test so far, until we reviewed with cluster sizes. It looks like we are centralized around one or two features in the case of cluster 1 and 3. While we have impressive scores at a k of 3 or 4, these clusters likely wouldn't generalize well the data set in classification. There are just too many data points in cluster 0.

K-Means Experiment 4:

In our last experiment, we examined K-Means clustering by recreating experiment one, this time, we also dropped features: very healthy, and very popular; in addition to health score. We decided that these could all be reasonable categorical targets we examine later on. The results are as follows:

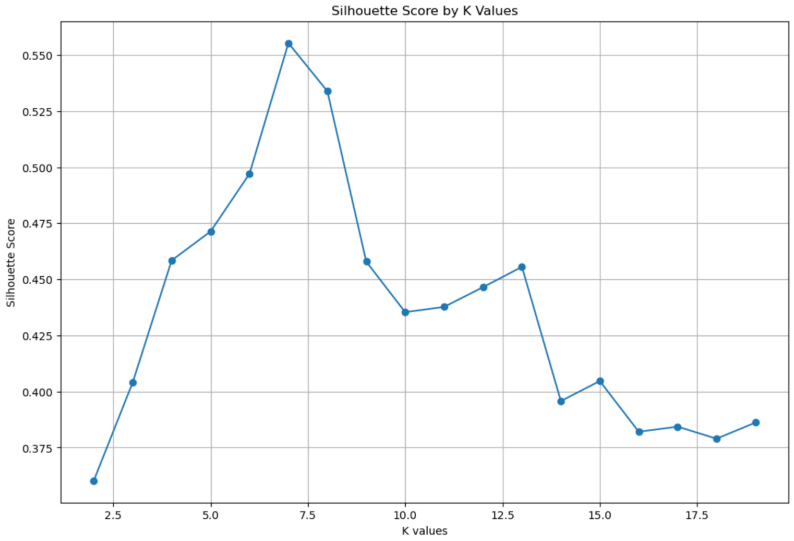


Figure *Experiment 4 - Silhouette Scores*

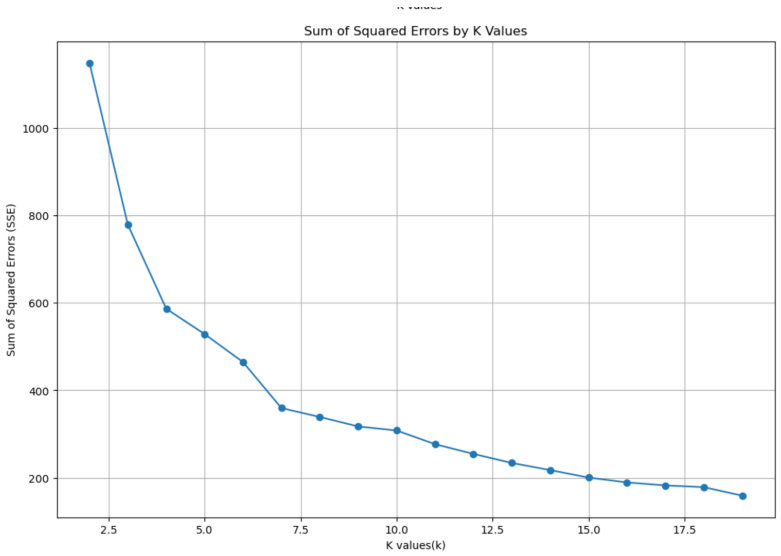


Figure *Experiment 4 – SSE Scores*

Table *Experiment 4 – Cluster Sizes*

```
Size of Cluster 0 =  572
Size of Cluster 1 =  245
Size of Cluster 2 =  404
Size of Cluster 3 =  422
Size of Cluster 4 =  79
```

It looks like at K = 5 we have a kick in our elbow, and we have a silhouette score near 0.5. We ran K-Means again this time just on K = 5 in order to plot the silhouette coefficient for each cluster and to verify our results.

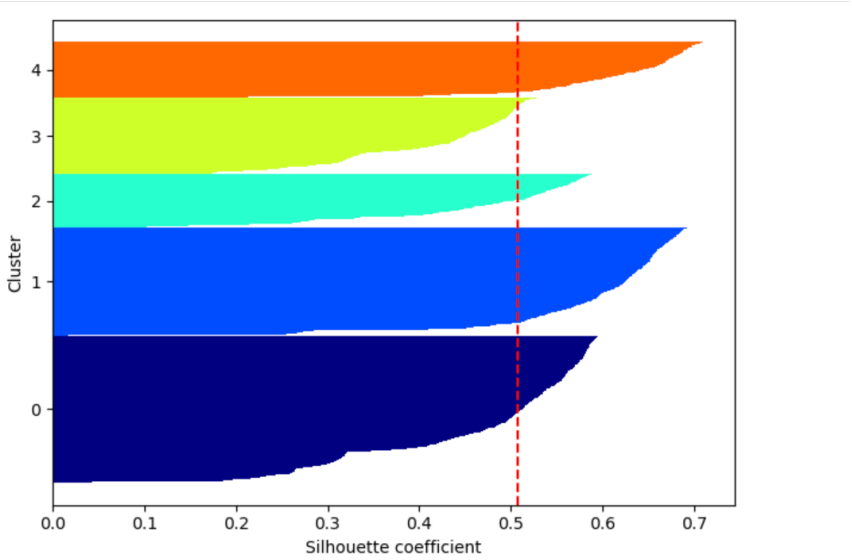


Figure *Experiment 4 – Cluster Silhouette Coefficients*

On this verifying run, we also saw a slightly improved average silhouette score of 0.51, as shown in the chat above. Because the results were so promising, we took a look at the centroids to see what might be contributing to the clustering assignments.

Table *Experiment 4 – Centroid Analysis*

	0	1	2	3	4
pricePerServing	0.00	0.00	0.00	0.00	0.00
weightPerServing	0.00	0.00	0.00	0.00	0.00
vegetarian	0.04	0.03	0.03	0.04	0.01
vegan	0.00	-0.00	0.02	0.03	0.01
glutenFree	1.00	-0.00	1.00	1.00	-0.00
dairyFree	-0.00	0.00	1.00	1.00	1.00
sustainable	0.00	0.00	0.00	0.00	0.00
lowFodmap	0.12	-0.00	0.10	0.19	0.00
ketogenic	0.04	0.01	0.01	0.02	0.00
whole30	0.01	-0.00	-0.00	1.00	-0.00
readyInMinutes	0.01	0.01	0.02	0.02	0.01
aggregateLikes	0.01	0.02	0.02	0.01	0.01
percentProtein	0.35	0.28	0.44	0.34	0.25
percentFat	0.55	0.49	0.39	0.57	0.39
percentCarbs	0.18	0.31	0.27	0.17	0.43

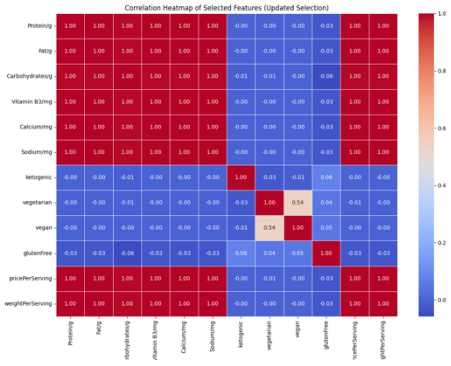
Based on the centroid analysis, we could see that the clusters relied heavily on whole30, glutenFree, and dairyFree. Cluster 4 for instance has a high average for the dairyFree feature. While Cluster 0 is gluten free. Clusters 2 and 3 are different combinations of gluten free, dairyfree, and whole 30. Cluster 1 is full of data points not associated with those features. Additionally, we can see insights like how cluster 1 (not associated with whole 30 gluten free or dairy free) and cluster 4 (dairyFree) are generally higher in carbs than the other clusters. This insight led us to store the cluster labels from this analysis for future analysis.

DBSCAN :

We conclude that DBSCAN was ineffective for this dataset, as it primarily labeled points as noise. The next steps would include to use other clustering techniques such as Agglomerative Clustering or K-Means to see if they provide better defined clusters. After the extensive tuning that was attempted the only other future improvement would include dimensionality reduction to try and improve the performance of DBSCAN.

Agglomerative Clustering:

The correlation analysis provided insights into the relationships between key features of food recipes. Calories strongly correlated with fat content, highlighting fats as a primary calorie contributor. Protein showed a notable correlation with Vitamin B3, while calcium had a mild positive association with fats. Dietary flags such as vegetarian and vegan were closely related, and gluten-free and dairy-free often co-occurred. Additionally, price per serving exhibited a slight positive relationship with serving weight. Understanding these correlations helped streamline the selection of relevant attributes, ensuring a meaningful clustering analysis of recipe healthiness.

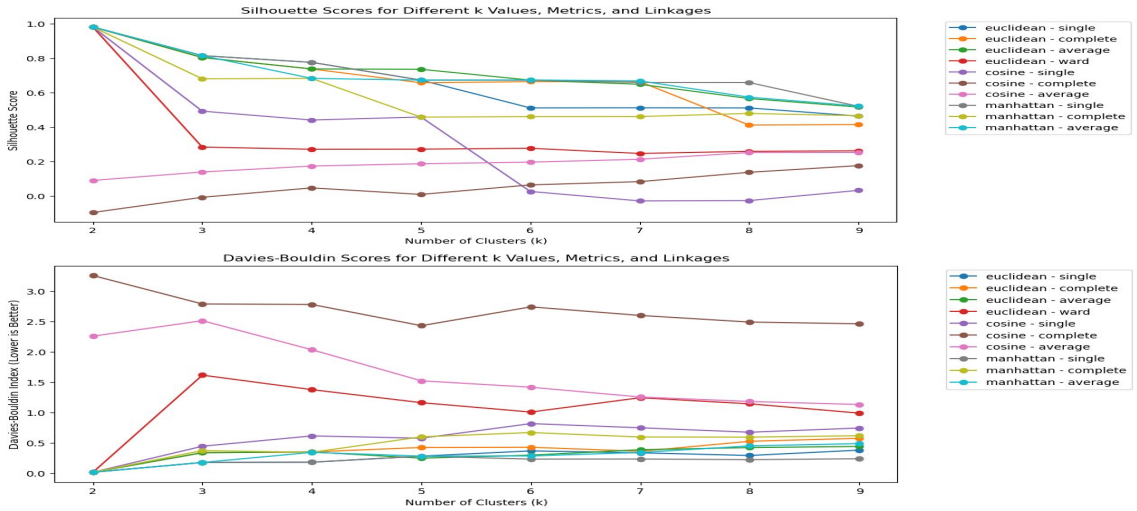


Initially, we explored various combinations of distance metrics (Euclidean, Cosine, and Manhattan) and linkage methods (Single, Complete, Average, and Ward) to evaluate their influence on clustering quality. We aimed to identify combinations that produced clearly defined and meaningful clusters.

Our experiments showed that the combination of Single Linkage with Euclidean distance consistently delivered superior performance, evident from higher silhouette scores (indicating better cohesion and separation among clusters) and lower Davies-Bouldin indices (signifying more compact and distinct clusters).

Building on these findings, we further refined our analysis by systematically varying the number of clusters (k) from 2 to 10, evaluating each using silhouette scores and Davies-Bouldin indices. Our results confirmed that Single Linkage with Euclidean distance remained the most effective, highlighting clear and meaningful cluster structures despite the method's known tendency for chaining.

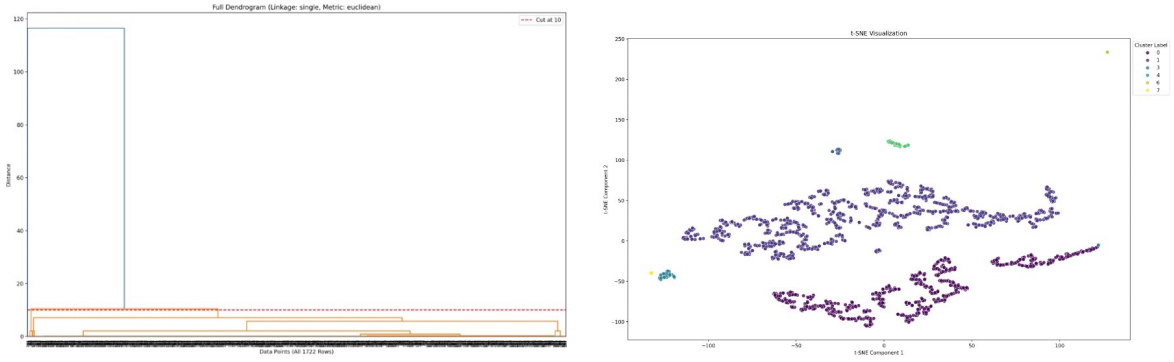
Consequently, we selected Single Linkage–Euclidean distance with an optimal number of clusters (determined through the silhouette and Davies-Bouldin evaluations) as the final configuration for our clustering analysis.



In our analysis, we initially visualized truncated dendrograms (showing only the top 5 hierarchical levels) across different linkage methods to quickly compare their clustering structures. Single Linkage demonstrated a notable chaining effect, where most data points merged early at low distances, leaving a few outliers merging later at higher distances. In contrast, Complete, Average, and Ward linkage methods showed more balanced merging patterns, although they still indicated high similarity among most recipes.

To further investigate these findings, we produced a full dendrogram using the optimal Single Linkage–Euclidean distance configuration and selected a cut distance of 10 through experimentation. This threshold provided a balanced clustering, ensuring that only recipes with very similar nutritional profiles merged together, thus avoiding excessive fragmentation or overly broad groupings. This dendrogram analysis validated our choice of clustering parameters and visually confirmed meaningful natural groupings within the dataset.

Full Dendrogram and Cut Analysis and t-SNE Visualization



In our analysis, we applied t-SNE to the scaled dataset to visualize the high-dimensional nutritional data in two dimensions. The resulting plot showed a prominent central cluster representing the majority of recipes with similar nutritional profiles, several smaller distinct clusters capturing unique recipe characteristics (such as extreme macronutrient profiles or specific dietary patterns), and isolated points identifying unique outliers.

To gain deeper insights, we assigned cluster labels derived from our optimal clustering configuration back to the original dataset and calculated average values for nutritional predictors and healthScore within each cluster. This revealed clusters with high protein and low fat generally exhibited higher healthScores, aligning with our hypothesis regarding healthiness criteria.

Our experiments provided essential insights: Single Linkage with Euclidean distance emerged as the optimal configuration, offering the clearest cluster separation based on silhouette scores and Davies-Bouldin indices. Dendrogram analyses confirmed natural hierarchical groupings and guided our selection of the cluster cutoff distance. Cluster profiling underscored the importance of Protein/g, Fat/g, Carbohydrates/g, Vitamin B3/mg, and Calcium/mg in distinguishing healthy recipes, and t-SNE visualization reinforced our quantitative findings, visually validating meaningful cluster formations. Overall, this comprehensive analysis using agglomerative clustering effectively grouped recipes into meaningful nutritional profiles, providing valuable insights into the key nutritional features that define food healthiness. These results are valuable for targeted nutritional recommendations and further research into dietary health.

Classification:

KNN :

We split the data into testing and training data and set *healthscore* as the target. We focused on the healthscore feature and changed it to healthy (1) or not healthy (0) target, with 30 being the splitting score. We decided to use 30 as the splitting score because it was the mean of the healthscore feature. The first KNN model we ran had very little tunning and had an accuracy score of .78.

	precision	recall	f1-score	support
0.0	0.77	0.83	0.80	179
1.0	0.80	0.73	0.76	166
accuracy			0.78	345
macro avg	0.78	0.78	0.78	345
weighted avg	0.78	0.78	0.78	345

Then, we want to see if PCA components would be better at finding the main features in indicating if a food item was healthy or not. We tried out different numbers of components and found that 2 components gave the best results. The accuracy was less than the not tunned model with a score of .69 vs.78.

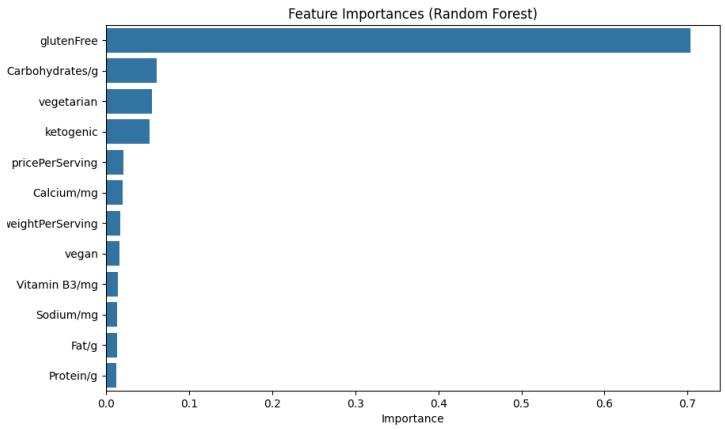
	precision	recall	f1-score	support
0.0	0.78	0.69	0.73	211
1.0	0.59	0.69	0.63	134
accuracy			0.69	345
macro avg	0.68	0.69	0.68	345
weighted avg	0.70	0.69	0.69	345

Lastly, we used linear regression forward selection to determine which features to use. We found 3 features have the best results. The feature the model found the most important were Manganese, Potassium, and Vitamin K. Using these three features gave the best accuracy, with a score of .82.

	precision	recall	f1-score	support
0.0	0.89	0.81	0.85	211
1.0	0.74	0.84	0.78	134
accuracy			0.82	345
macro avg	0.81	0.82	0.81	345
weighted avg	0.83	0.82	0.82	345

Random Forest:

The Random Forest feature importance analysis revealed that the dietary flag ‘glutenFree’ was the most influential predictor by a significant margin, highlighting dietary restrictions as a key factor differentiating recipes. Other dietary attributes, such as ‘ketogenic’ and ‘vegetarian’, along with macronutrients like ‘Carbohydrates/g’, also showed moderate importance. In contrast, economic factors (‘pricePerServing’ and ‘weightPerServing’) and certain micronutrients (‘Calcium/mg’, ‘Vitamin B3/mg’, ‘Sodium/mg’) had comparatively lower impacts on clustering outcomes.



Decision Tree:

We used the clusters from the K-Means model as features in a decision tree to determine if a food item was healthy or not. I labeled the clusters: 'Cluster_0': 'glutenFree', 'Cluster_1': 'control_group', 'Cluster_2': 'glutenFree_dairyFree', 'Cluster_3': 'glutenFree_dairyFree_whole30', 'Cluster_4': 'dairyFree'. The model was not very accurate with an accuracy score of .56 and the tree gave no clear way to determine if a food item was healthy or not.

```
Accuracy: 0.5681159420289855
|--- control_group <= 0.50
|   |--- glutenFree <= 0.50
|   |   |--- dairyFree <= 0.50
|   |   |   |--- glutenFree_dairyFree_whole30 <= 0.50
|   |   |   |   |--- class: 1
|   |   |   |--- glutenFree_dairyFree_whole30 >  0.50
|   |   |   |   |--- class: 1
|   |   |--- dairyFree >  0.50
|   |   |   |--- class: 1
|   |--- glutenFree >  0.50
|   |   |--- class: 0
|--- control_group >  0.50
|   |--- class: 0
```

Conclusion and Discussion

Working with this dataset we tried many things. First, we looked to see if there were any strong correlations with health score as this was the first target we sought to evaluate. Second, we examined many different unsupervised methods looking for patterns in the data, both with and without categorical features. In our unsupervised analysis we confirmed that the natural groupings were K = 5. This was confirmed in both Agglomerative and K-Means analyses. In K-means we saw that the groups formed about different categories and combinations of dairyFree, GlutenFree, and Whole30. Later, in our classification analysis we determined that those categories were not good predictors of whether a food item is healthy. We also performed an analysis by means of PCA and feature reduction that showed Manganese, Potassium, and Vitamin K, were fairly good predictors of food health, however, that data still contained a bit of noise, those features are highly correlated with each other, and it looks like the model we ran that determined this was not generalizable.

In conclusion, we have determined that there is no clear way to predict a food's health. Gluten Free foods or Whole30 diet foods, for example, are not indicative of whether a food is healthy or not. We believe this is likely because our target label (in this case health score) might be a subjective one. Different foods and their components are good or bad for different people at different times.

References

Gillespie, K. M., Kemps, E., White, M. J., & Bartlett, S. E. (2023). The Impact of Free Sugar on Human Health-A Narrative Review. Nutrients, 15(4), 889. https://doi.org/10.3390/nu15040889