Section 1: Business Need and Importance

The prevalence of obesity in the United States has reached alarming levels, with about 42% of adults considered obese in 2023 according to the Trust For American Health (Faberman 2023). This epidemic not only poses significant health risks but also places a substantial economic burden on the healthcare system, with obesity-related medical costs estimated to range from $147 to $210 billion annually (STOP 2024). Addressing this issue requires innovative solutions that promote healthier dietary habits and lifestyle choices. By leveraging unsupervised learning techniques, we can create a robust recommendation system for foods tailored to individual needs and preferences. This system has the potential to empower individuals to make healthier food choices by providing personalized recommendations based on nutritional content and food preferences, contributing to the reduction of obesity rates, and improving public health outcomes. The scalability and accessibility of such a recommendation system make it a viable solution for addressing the obesity epidemic on a large scale, reaching diverse populations across different demographic groups and socioeconomic backgrounds. Through this analysis, we aim to not only combat the obesity crisis but also drive positive social and economic impact by promoting healthier lifestyles and reducing healthcare costs associated with obesity-related diseases.

Section 2: Statistical Methodology

I have selected data from the USDA's Nutrition Dataset, which contains nutritional information on almost 8000 different foods. This information includes food categories and values for calories, macronutrients, vitamins, and minerals. For my analysis, I will be using two unsupervised learning techniques. The first technique is Agglomerative Clustering. I decided to use hierarchical clustering because it's a helpful method when you don't know the ideal number of clusters, giving

you a starting point for further analysis. The agglomerative coefficient shows the strength of clustering, and the dendrogram gives a better visual of the possible clusters for better interpretation. This method also allows you to summarize the values for each variable within each cluster. As a certified nutrition coach, having these values helps me classify the clusters in terms that my clients can understand, such as high protein or low fat, making it easier for them to make decisions and act.

After determining the optimal number of clusters produced by Agglomerative Clustering, I then applied K-Means clustering to the same data. Choosing the right clustering algorithm is crucial to ensure accurate results. While Agglomerative clustering may seem like a good option, it's unstable at the beginning as each observation is considered as its own cluster. This means that adding in a new variable, such as Sodium, would possibly produce different clusters. K-means clustering, on the other hand, establishes the number of random centroids upfront and assigns observations to the nearest centroid based on their distance to form k number of clusters. With the centroids being recalculated until optimal, this method is better suited for situations where new data is added. After determining the optimal number of clusters using Agglomerative clustering, K-means was tested with K-1, K, and K+1 clusters to evaluate the results. This approach can help reveal patterns in nutritional composition and aligning these clusters with specific food categories can also assist in accounting for dietary restrictions in our recommendations.

Section 3: Results and Interpretation

Beginning with Agglomerative Clustering, the goal was to determine if there was a strong clustering connection between the ingredients. This is determined by the agglomerative coefficient, but the methods for calculating distance between points can vary. To make this analysis more well-rounded, both the Euclidean & Manhattan distances were calculated and used to build the agglomerative clustering model. While both were strong, the model that utilized Manhattan distances

proved slightly better, with an Agglomerative coefficient of 0.9985 compared to 0.9983 with the Euclidean distance. This produced the following dendrogram with 4 distinct clusters (Figure 1)

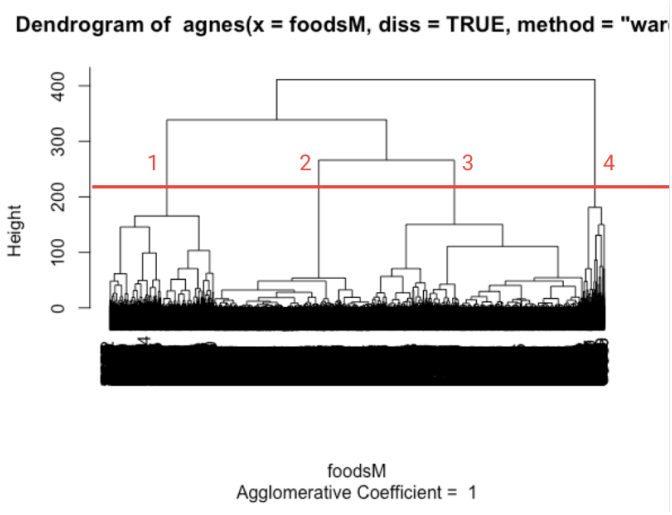

Figure 1: Dendrogram produced by Agglomerative Clustering with Manhattan distances

From here the next step was to analyze the summary of values for each cluster, the table below shows the mean values of each variable within each cluster (Table 1)

Cluster (# of foods)	Calories	Carbs (g)	Protein (g)	Fat (g)	Trans Fat (g)	Saturated Fat (g)	Monounsaturated Fat (g)	Polyunsaturated Fat (g)	Fiber (g)	Sugar (g)
Cluster 1 (1632)	1106	65	8	11	0	3.5	3	2	5	20
Cluster 2 (2491)	191	12	2.5	1	0	0	0	0	1.5	4
Cluster 3 (327)	620	4	21	12	0	4	4.5	1.5	0	1
Cluster 4 (343)	1903	10	8	70	2	19	27	19	3	2

Table 1: Mean values of each variable within each cluster

These values allow for the categorization of foods in more common terms. For example, cluster 3 can be described as high protein, low carb foods. Verification was then performed through K-Means Clustering with 4 clusters chosen, for a more reliable method of clustering. Figure 2 below shows the visualization for K = 4, the visuals for K = 3 and K = 5 can be found in the Appendix. The plot

shows the clusters graphed against the first 2 principal components, which explain 54.42% of the variability in the data. One way to evaluate the effectiveness of a K Means clustering method is by looking at the Average Silhouette Width of the total dataset. This metric provides an overall assessment of how well the clusters are separated from each other and how cohesive they are. When comparing the results of K = 5 and K = 4, the average silhouette width is only slightly higher for K = 5, with values of 0.35 and 0.33 respectively. Therefore, the difference between the two is not significant enough to justify using 5 clusters over 4. Moreover, K = 5 exhibits more overlap between clusters, which reduces the uniqueness of each cluster. For these reasons, it is reasonable to settle on an optimal value of K = 4. It should also be noted that these two principal components only explain about 55% of the variability in the data. Therefore, incorporating more or different data in the clustering algorithm could potentially lead to better results.

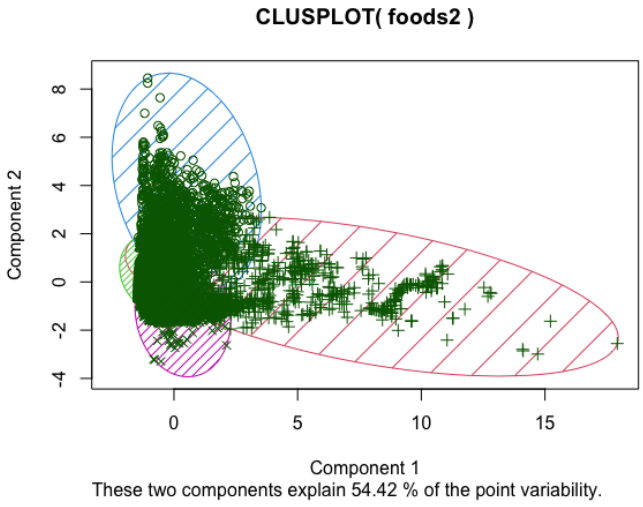

Figure 2: Cluster Plot of K-Means Clusters with K = 4

Section 4: Alternative Approaches

When selecting unsupervised learning techniques for the analysis, some alternative methods were considered but not chosen. One of the methods considered was Association Analysis, which

could have been useful in analyzing food recipes to identify which ingredients are commonly used together. However, this method does not allow a detailed examination of the nutritional values of ingredients to determine the most nutritious foods for combating obesity. Another alternative method was Monte Carlo simulation, which would have enabled me to simulate different scenarios of calorie intake and macronutrient composition to provide personalized nutrition recommendations and informed decision-making. However, this method was excluded as I do not currently have biometric data or data on my clientele's eating patterns necessary to begin formulating simulations.

Section 5: Conclusions

Obesity affects almost half of the adult population, costing billions of dollars each year, and therefore requires a solution. As a nutrition coach, I have observed that people tend to avoid healthy eating because they believe it involves bland foods and repetitive meal prep. However, that's not true. Even in a limited dataset, we can find a variety of options for high-protein, low-carb foods across 20 different categories (view the recommendations dataframe in my R script). By analyzing this dataset, we can create a robust food recommendation system that aligns with an individual's dietary, cultural, religious, and health goals. If someone doesn't like a recommended food, we can suggest another food from the same category and cluster until they find one that suits their taste. The goal is to simplify healthy eating, fight obesity, and reduce food waste. We could even establish a system between this recommendation tool and local food banks to take advantage of their resources and help those who can't afford groceries regularly. There is enormous potential for this analysis to be used, particularly as more foods are added and algorithms are refined.

APPENDIX

Farberman, Rhea. “State of Obesity 2023: Better Policies for a Healthier America.” *TFAH*, 22 Nov. 2023, www.tfah.org/report-details/state-of-obesity-2023/#:~:text=Nationally%2C%2041.9%20percent%20of%20adults,percent%20and%2045.6%20percent%20respectively.

Strategies to Overcome & Prevent (STOP) Obesity Alliance. *Costs of Obesity*, stop.publichealth.gwu.edu/sites/g/files/zaxdzs4356/files/2022-06/fast-facts-costs-of-obesity.pdf. Accessed 3 May 2024.

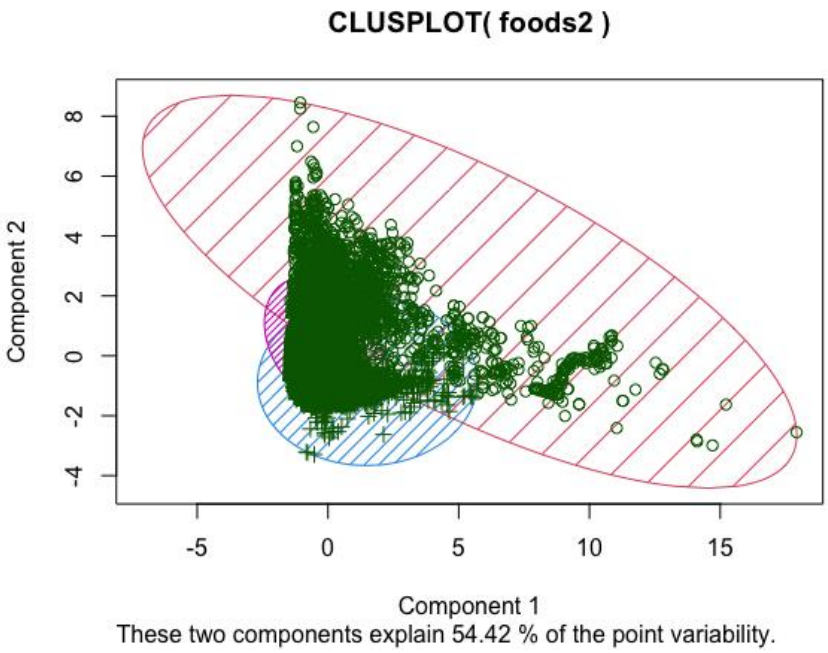

Figure 3: Cluster Plot of K-Means Clusters with K = 3

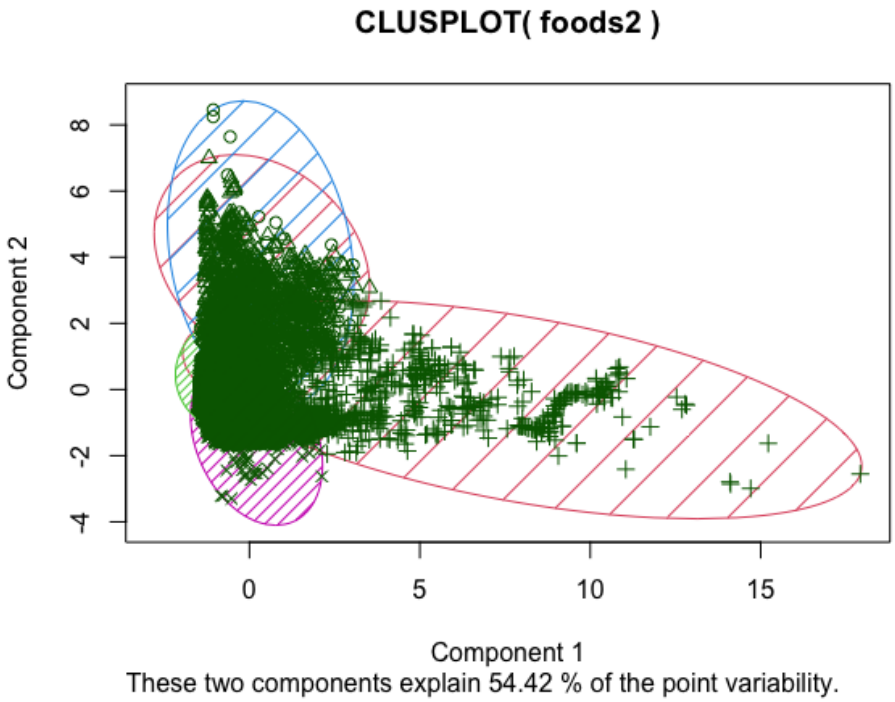

Figure 3: Cluster Plot of K-Means Clusters with K = 5

aClusters	Food.Category	Total
1	Baked Products	490
1	Sweets	251
1	Breakfast Cereals	165
1	Snacks	160
1	Cereal Grains and Pasta	121
2	Vegetables and Vegetable Products	726
2	Fruits and Fruit Juices	305
2	Beverages	304
2	Baby Foods	260
2	Soups, Sauces, and Gravies	195
3	Beef Products	938
3	Lamb, Veal, and Game Products	443
3	Poultry Products	381
3	Pork Products	327
3	Fast Foods	271
4	Fats and Oils	165
4	Nut and Seed Products	83
4	Legumes and Legume Products	25
4	Lamb, Veal, and Game Products	18
4	Beef Products	13

Table 2: Top 5 Food Categories for each Cluster after Agglomerative Clustering