Smart Bites: Nutritional Clustering and Recommender System for Healthier Food Choices

Margvinatta Senesie

Stockton University

DSSA 5302: Final Practicum

August 7, 2025

ABSTRACT

With diet-related chronic diseases on the rise, consumers need practical tools to help them make informed and healthier food choices. This project analyzes 8,789 food items using a publicly available nutrition dataset from Kaggle. The data was cleaned and standardized to ensure consistency across nutrient values. A custom health score was then developed to evaluate and compare the nutritional quality of foods. This score was based on the principle that higher protein and lower fat and sodium levels contribute to better health outcomes.

Using this score, I applied clustering techniques to group foods into three distinct nutrient profiles. Principal Component Analysis (PCA) was used to confirm that the clusters were meaningfully separated based on their nutrient makeup. To further enhance the usefulness of the analysis, I created a recommendation system that suggests healthier alternatives for any selected food item by comparing nutritional similarity and filtering for higher health scores.

The findings are presented through clear visuals, including scatter plots, cluster maps, and nutrient-based word clouds. This project demonstrates how a data-driven approach can simplify the process of identifying better food choices and can serve as a model for future efforts to promote healthier eating habits through personalized and accessible nutrition guidance.

INTRODUCTION

What if we could help consumers identify healthier food options not through calorie counting, but by letting data reveal patterns in what we eat? This project explores that idea by asking: **Can unsupervised clustering, paired with a nutrient-based recommendation system, help consumers discover healthier food alternatives?**

This question matters because poor dietary choices are a major driver of preventable illnesses such as obesity, heart disease, and type 2 diabetes. Yet the average person is often overwhelmed by complex or misleading nutrition labels. With the growing availability of food datasets and tools for analysis, there is an opportunity to simplify how we understand nutrition and promote better choices.

For this project, I used the SmartBites dataset, which contains detailed nutritional information for a wide range of foods. By applying unsupervised clustering, I identified distinct categories of foods based on their nutrient profiles. I then built a similarity-based recommendation system to suggest healthier alternatives within or across these clusters.

To ensure consistency, I cleaned and standardized the data, focusing on key nutrients such as fat, sodium, protein, and sugar. The analysis was conducted in R using packages for clustering (cluster, factoextra), visualization (ggplot2, corrplot), and data transformation (dplyr, tidyr).

This project contributes to the growing field of nutritional informatics. It presents a simple and user-friendly method for identifying better food choices, making nutrition guidance more accessible and data-driven.

METHODS

I conducted this analysis using a dataset downloaded from Kaggle containing nutritional information for 8,789 food items. The dataset was imported into RStudio as an Excel file (nutrition.xlsx) and included 77 variables describing various foods. I used the packages readxl, dplyr, and tidyr in R for data cleaning and transformation.

To prepare the data, I removed units and special characters from nutrient fields and converted all relevant columns into numeric format. I focused on 16 core nutrients for analysis: calories, total fat, saturated fat, cholesterol, sodium, protein, carbohydrate, fiber, sugar, choline, folic acid, calcium, iron, vitamin A, vitamin C, and water. These were selected due to their relevance to chronic health outcomes and overall dietary quality.

I created a custom health score using the formula:

health_score = protein − (total_fat + sodium / 100)

This formula prioritizes foods with high protein and lower levels of fat and sodium.

For clustering, I filtered the dataset to include only complete cases and standardized the nutrient values. I applied k-means clustering with $k = 3$, selected for its clarity and interpretability. Cluster separation was visualized using the factoextra package and Principal Component Analysis (PCA).

To explore trends, I created scatter plots (e.g., calories vs. protein; fat vs. sodium) using **ggplot2**, and used **corrplot** to display nutrient correlations. I also built a simple recommendation system that compares nutrient profiles between foods and suggests alternatives with higher health scores. Nutrient similarity was measured using distance calculations, and recommendations were filtered to prioritize healthier substitutions.

The full code and reproducible analysis are available at:

https://github.com/Margvinatta/Final-DSSA-research-project

RESULTS

Health Score Ranking and Top Foods

Using the custom health score formula (protein − [total_fat + sodium/100]), the system identified the healthiest foods in the dataset. Soy protein isolate received the highest score (87.3), followed by gelatin powder (83.5), whey protein powder, seal meat, and dried egg whites, all of which scored above 65. These foods had high protein content and minimal fat and sodium, supporting the logic of the scoring model. *Table 1* presents the top 10 ranked items.

Table 1. 10 top high-protein nutrients

Name	**Health Score**
Soy protein isolate, potassium type	87.3
Gelatins, unsweetened, dry powder	83.5
Seal, dried (Alaska Native), meat, bearded (Oogruk)	80.3
Beverages, Protein powder whey based	75.0
Soy protein isolate	74.9
Vital wheat gluten	73.0
Egg, glucose reduced, stabilized, dried, white	70.8

Egg, glucose reduced, stabilized, powder, dried, white	70.0
Egg, dried, white	68.3
Egg, glucose reduced, stabilized, flakes, dried, white	65.4

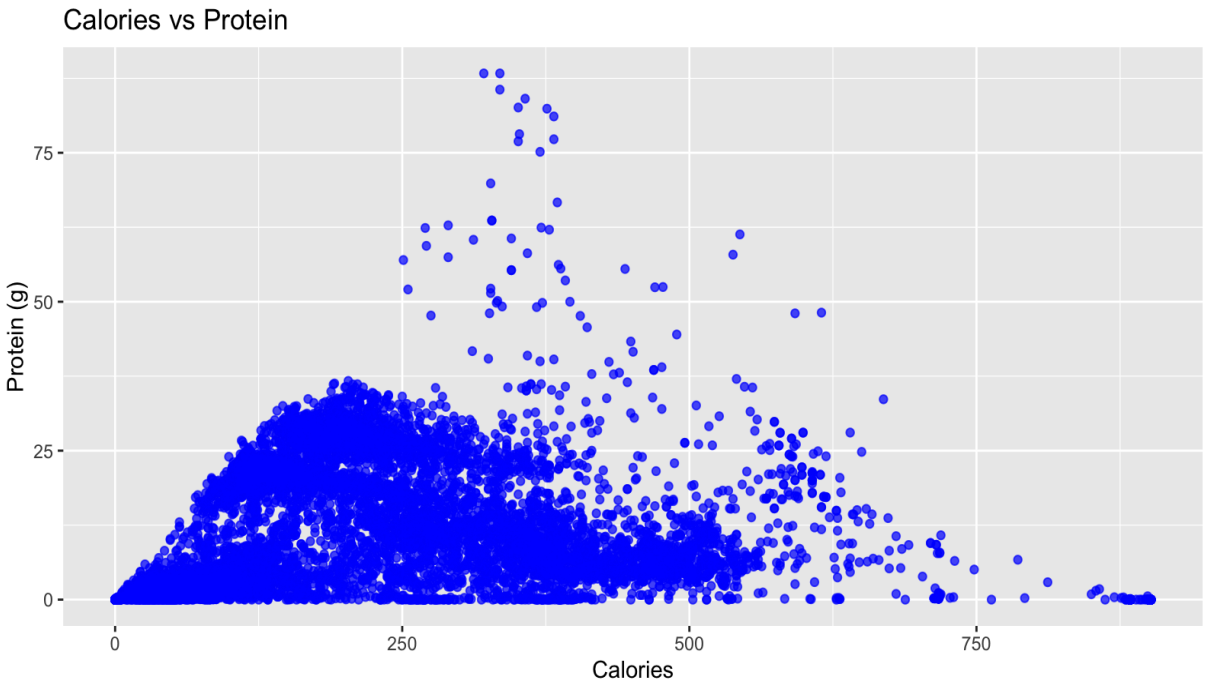

Figure 1. Calories vs Protein

This scatter plot shows the relationship between calorie content and protein levels across food items. We observe that while most items cluster around 100–300 calories and 0–30 grams of protein, some foods have extremely high protein content with relatively moderate calories. These outliers likely include items such as protein powders or lean meats. This pattern supports the

logic behind the health score and helps confirm the importance of protein as a key positive metric in scoring.

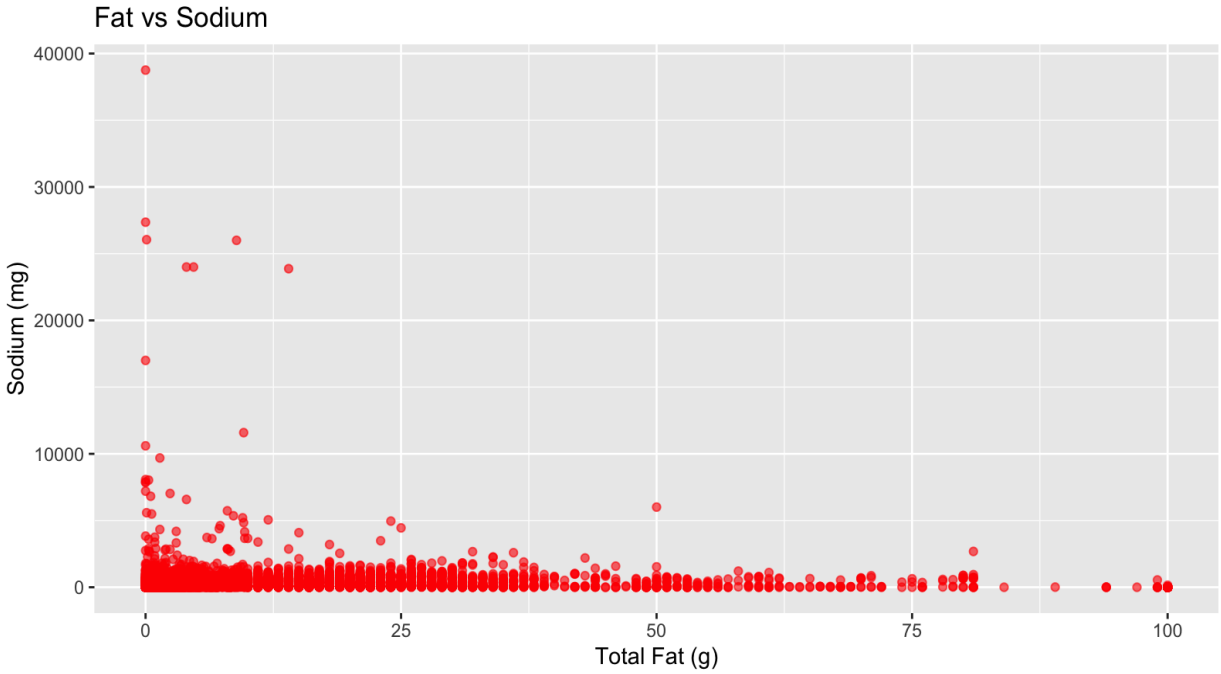

Figure 2: Total Fat vs Sodium

This graph examines the relationship between fat and sodium,two nutrients generally associated with negative health outcomes. The scatter is skewed with many foods having low fat and sodium, but notable outliers show extremely high sodium levels (above 30,000 mg), often linked to processed or preserved items. This visualization reinforced the rationale behind penalizing high sodium in the health score and clustering logic.

Figure 3. PCA Biplot: Nutrient Influence on Cluster Separation

This biplot visualizes the distribution of food items (colored by cluster) along the first two principal components derived from 16 nutrient variables. The arrows represent nutrients, with direction and length indicating their contribution to the variance and cluster separation. For example, *clean_protein* and *clean_water* drive separation toward Cluster 3 (blue), while *clean_calories* and *clean_fat* are more associated with Cluster 1 (red). This plot helps explain which nutrients are most influential in shaping the food groupings generated by k-means clustering.

Cluster Analysis

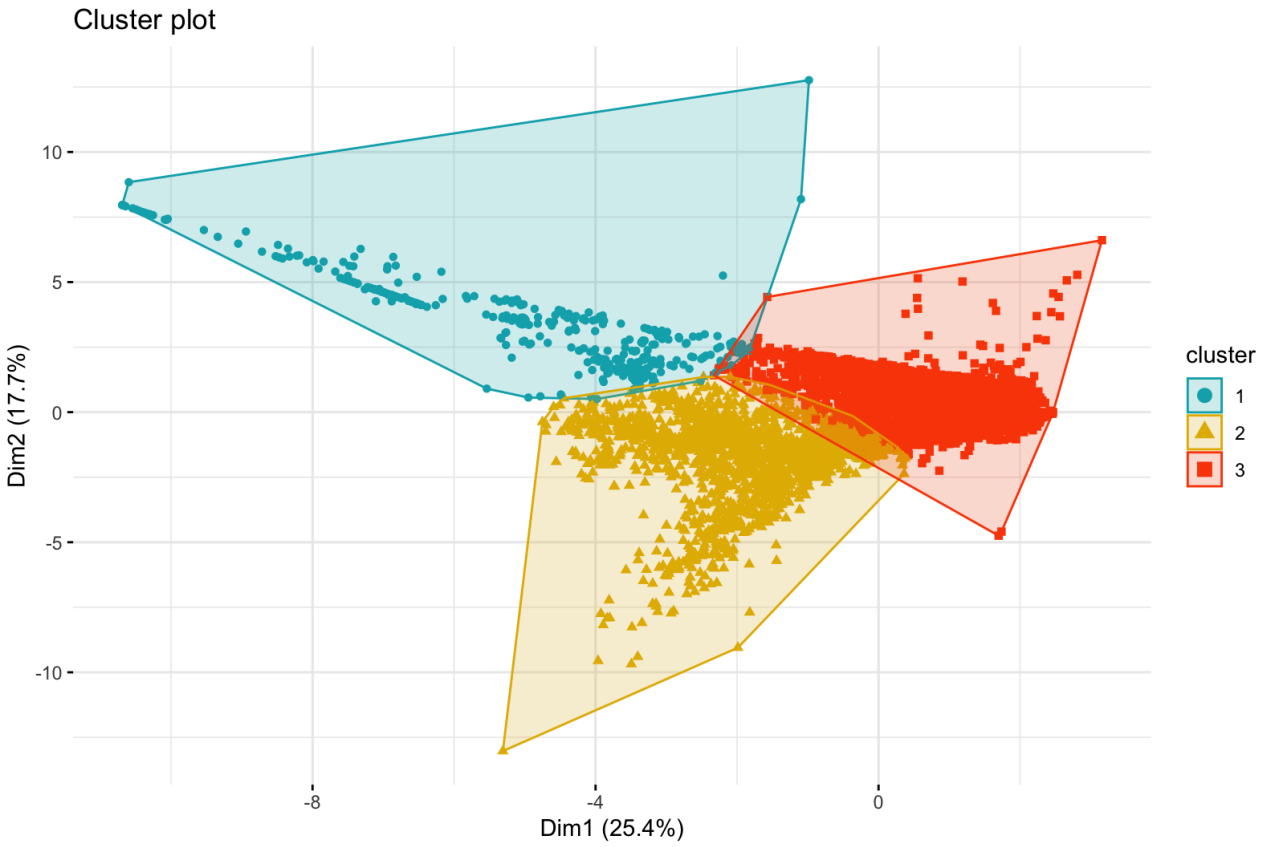

Figure 4. K-means Clustering of Food Items

K-means clustering (k = 3) produced three nutrient-based food clusters, visualized through **PCA plots** and **radar charts**.

- **Cluster 1** contained processed, calorie-dense foods high in fat and sodium (e.g., toaster pastries, cookies).
- **Cluster 2** included mixed-nutrient foods such as cereals and granola bars with moderate health profiles.

- **Cluster 3** grouped lean, high-protein items such as protein powders and eggs.

  These results show clear distinctions between food groups and validate the clustering approach.

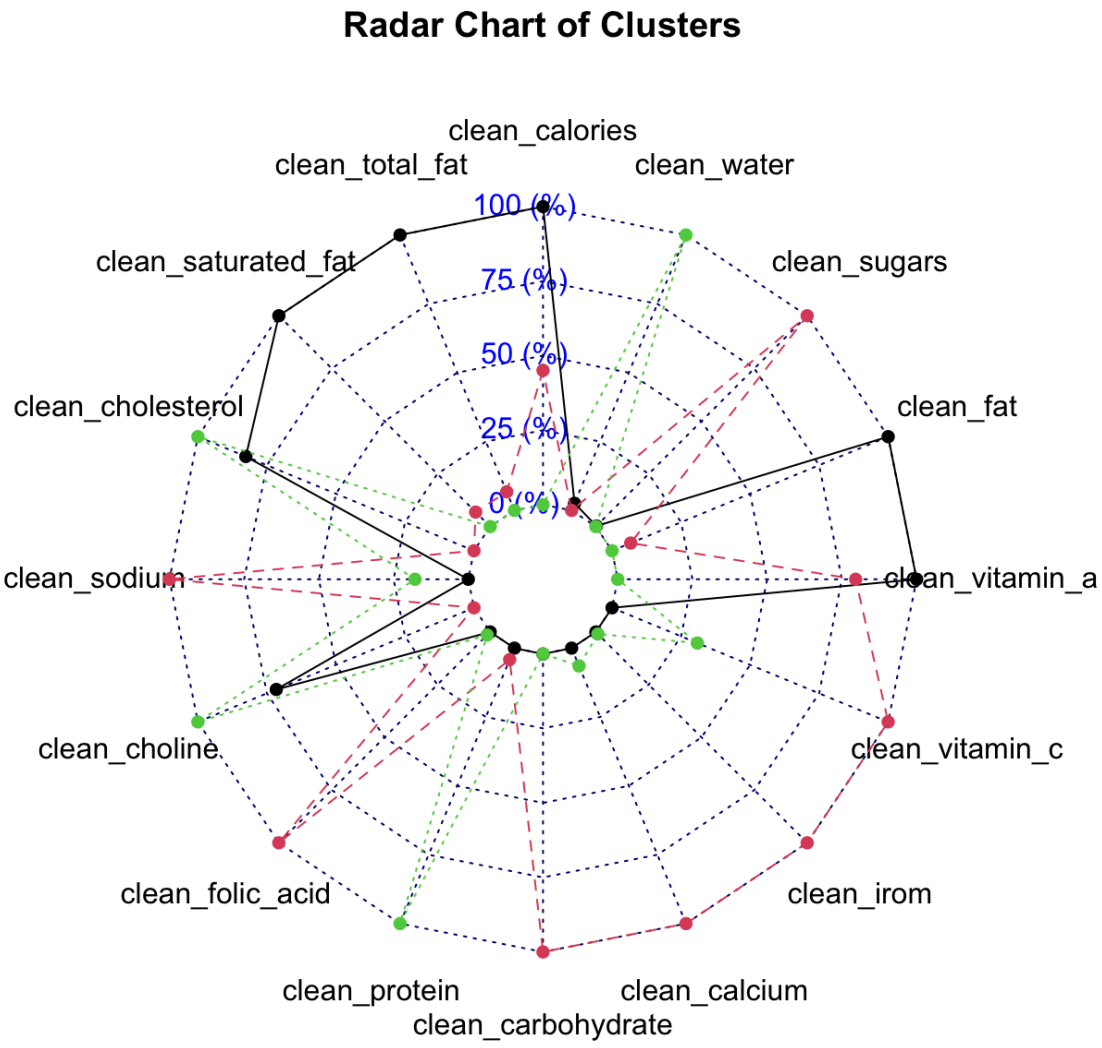

Figure 5. Radar Chart of Cluster Nutrient Profiles

This radar chart displays the **mean normalized values** for 16 nutrients across the three clusters. The visual reveals that:

- **Cluster 1** is highest in fat, calories, and saturated fat.
- **Cluster 2** shows moderate and balanced levels across most nutrients.
- **Cluster 3** stands out for high protein and choline, with low sodium and fat.

This visualization supports the interpretation of the PCA clusters and highlights the distinct **nutritional identities** of each group,making it easier for users or system designers to understand what kind of foods fall into each category.

Word Cloud Visualization

To better understand the composition of each cluster, word clouds were generated (*Figure 4*). These visualizations revealed dominant ingredients and food types within each cluster. Cluster 1 prominently featured desserts and snack items, Cluster 2 included dmixed nutrients with moderate health profiles, while Cluster 3 was dominated by high-protein, minimally processed foods. The clustering themes were consistent with nutrient data trends.

Recommender System Performance

The food recommender system, built using cosine similarity, was tested using a food item with a low health score: *"Frankfurter, meatless"* (health score: 0.9). The system successfully retrieved the top 20 most nutritionally similar foods and filtered them to display only those with higher health scores. The recommended alternatives included lean meats, fish, and high-protein items, with health scores ranging from 18 to 28. These results demonstrate that the system works as intended—preserving nutritional similarity while suggesting healthier substitutions. A screenshot of this test is shown in Figure 6.

```
# A tibble: 20 × 3
   name                                                          health_score cluster
   <chr>                                                                <dbl> <fct>
 1 "Veal, grilled, cooked, boneless, cutlet, cap off, top round, le…     28.4 3
 2 "Veal, grilled, cooked, separable lean only, chop, loin"              24.5 3
 3 "Pork, pan-fried, cooked, separable lean only, bone-in, center l…     20.9 3
 4 "Veal, grilled, cooked, separable lean only, blade chop, shoulde…     20.7 3
 5 "Pork, pan-fried, cooked, separable lean and fat, boneless, top …     20.6 3
 6 "Pork, braised, cooked, separable lean and fat, boneless, top lo…     20.2 3
 7 "Beef, broiled, cooked, all grades, trimmed to 1/8\" fat, separa…     20.0 3
 8 "Fish, dry heat, cooked, sockeye, salmon"                             20.0 3
 9 "Pork, roasted, cooked, separable lean only, bone-in, center loi…     19.7 3
10 "Beef, roasted, cooked, select, trimmed to 0\" fat, separable le…     19.6 3
11 "Beef, grilled, cooked, select, trimmed to 0\" fat, separable le…     19.5 3
12 "Veal, braised, cooked, separable lean and fat, osso buco, fores…     19.2 3
13 "Pork, braised, cooked, separable lean only, bone-in, center rib…     19.0 3
14 "Beef, broiled, cooked, all grades, trimmed to 0\" fat, separabl…     18.9 3
15 "Pork, roasted, cooked, separable lean only, bone-in, center rib…     18.7 3
16 "Fish, dry heat, cooked, pink, salmon"                               18.4 3
17 "Beef, grilled, cooked, choice, trimmed to 0\" fat, separable le…     18.3 3
18 "Pork, pan-fried, cooked, separable lean only, bone-in, center r…     18.3 3
19 "Beef, roasted, cooked, all grades, trimmed to 0\" fat, separabl…     17.9 3
20 "Beef, roasted, cooked, select, trimmed to 0\" fat, separable le…     17.8 3
>
> df_clean[food_index_1, c("name", "health_score")]
# A tibble: 1 × 2
  name                  health_score
  <chr>                        <dbl>
1 Frankfurter, meatless        0.900
>
```

Figure 6. Showing healthier Alternatives to Frankfurter, meatless with higher health scores

DISCUSSION & RESULTS

This project demonstrates that combining unsupervised clustering with a simple health scoring system can provide useful insights into food categorization and healthier substitutions. The results highlight how data-driven approaches can simplify nutrition for consumers by turning complex nutrient profiles into understandable guidance.

The identification of high-protein, low-fat, low-sodium items among the top-scoring foods validates the logic behind the custom health score. Clustering results effectively grouped similar

foods and revealed patterns that can inform dietary decisions, especially for individuals trying to avoid high-fat or highly processed items.

However, the model has limitations. The health score is based on subjective assumptions,namely, that protein is "good" and fat/sodium are "bad",and does not account for other critical nutrients like fiber or sugar. Additionally, all nutrients were weighted equally, which does not reflect how nutritional guidelines or expert dietitians prioritize macronutrients and micronutrients. A more refined scoring system could incorporate Recommended Dietary Allowance (RDA) values or expert-defined weights.

The recommender system shows potential but currently lacks user customization, such as filtering by allergies, dietary preferences, or cultural considerations. Future improvements could integrate these features, making the tool more personalized and practical.

While there is limited published work on cosine-similarity food recommendation systems, the findings here align with the broader movement toward nutrition informatics and consumer-facing dietary tools. Expanding this system into a mobile app or web-based interface could offer everyday users a practical way to discover better food choices based on real data.

A Shiny app version of the system was also developed and successfully tested locally. It includes category filters, food selectors, interactive visualizations, and live recommendations. Unfortunately, due to the large number of selectable food options, deployment to shinyapps.io failed with a server-side selectize error. Despite this, the app demonstrates the project's functionality when run locally using Shiny: : runApp( ). Figure 7 is a screenshot of the recommender system on the shiny app as proof of performance .

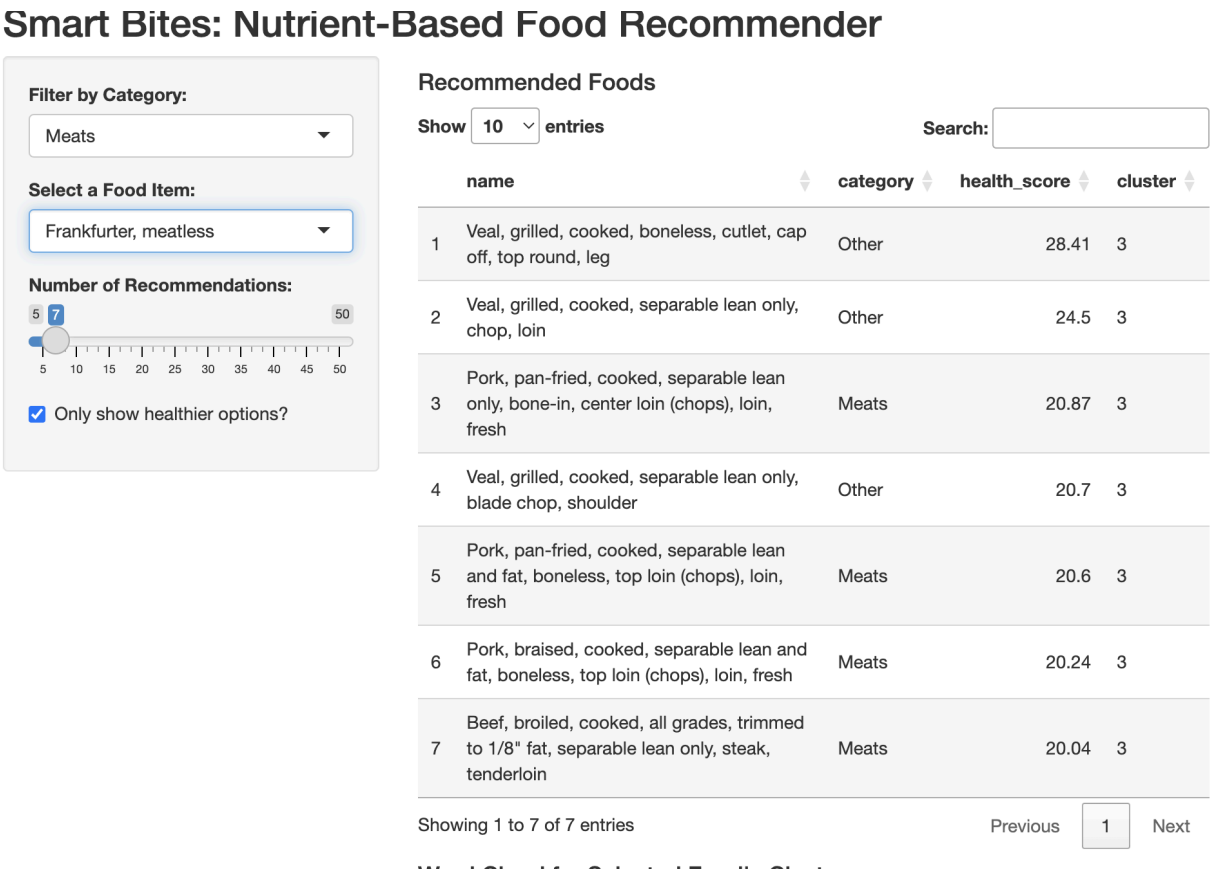

Figure 7. Showing a snippet of locally deployed Shiny App

CONCLUSION

This project demonstrated how unsupervised clustering and cosine similarity can be effectively applied to nutritional data to reveal patterns and suggest healthier food alternatives. By building a custom health score and developing a content-based recommender system, I created a functional prototype that can guide smarter dietary decisions. The clustering process revealed clear distinctions among food types, while the recommender engine provided logical and nutritionally consistent substitutions. These findings support the potential for data science to simplify nutrition and personalize food guidance.

Future work could refine the health score by incorporating additional nutrients such as fiber and sugar, apply expert-weighted nutrient values, and integrate user-specific preferences like

allergies, dietary goals, or cultural context. With further development, this framework could support consumer-facing applications or educational tools that empower individuals to make better food choices based on real data.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to Professors Baldwin and Laurino, and classmates in DSSA 5302 for their valuable guidance, insights, and feedback throughout this project. Their support greatly enhanced the quality of my work.

Special thanks to the R community and the developers of open-source packages such as dplyr, ggplot2, and factoextra, whose tools made this analysis both possible and enjoyable.

REFERENCES

González-Castro, V., & Martínez-Rego, D. (2014). A review of cosine similarity-based recommender systems and their applications. *International Journal of Pattern Recognition and Artificial Intelligence, 28*(4), 1451001. https://doi.org/10.1142/S0218001414510016

Harvard T.H. Chan School of Public Health. (2023). *Protein: Moving the meat off your plate*. https://www.hsph.harvard.edu/nutritionsource/what-should-you-eat/protein/

James, W. P. T. (2008). The epidemiology of obesity: The size of the problem. *Journal of Internal Medicine, 263*(4), 336–352. https://doi.org/10.1111/j.1365-2796.2008.01922.x

Lustig, R. H. (2010). Fructose: It's "alcohol without the buzz." *Advances in Nutrition, 1*(3), 218S–225S. https://doi.org/10.3945/an.110.000073

Monteiro, C. A., Moubarac, J.-C., Cannon, G., Ng, S. W., & Popkin, B. M. (2013). Ultra-processed products are becoming dominant in the global food system. *Obesity Reviews, 14*(S2), 21–28. https://doi.org/10.1111/obr.12107

R Core Team. (2024). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

SmartBites. (2022). *SmartBites – Nutritional information of 8,789 foods* [Data set]. Kaggle. https://www.kaggle.com/datasets/therealsampat/food-nutrition-data

Tan, P.-N., Steinbach, M., & Kumar, V. (2018). *Introduction to data mining* (2nd ed.). Pearson.

U.S. Department of Health and Human Services & U.S. Department of Agriculture. (2020). *Dietary guidelines for Americans, 2020–2025* (9th ed.). https://www.dietaryguidelines.gov/

Wickham, H., & Grolemund, G. (2016). *R for data science*. O'Reilly Media.