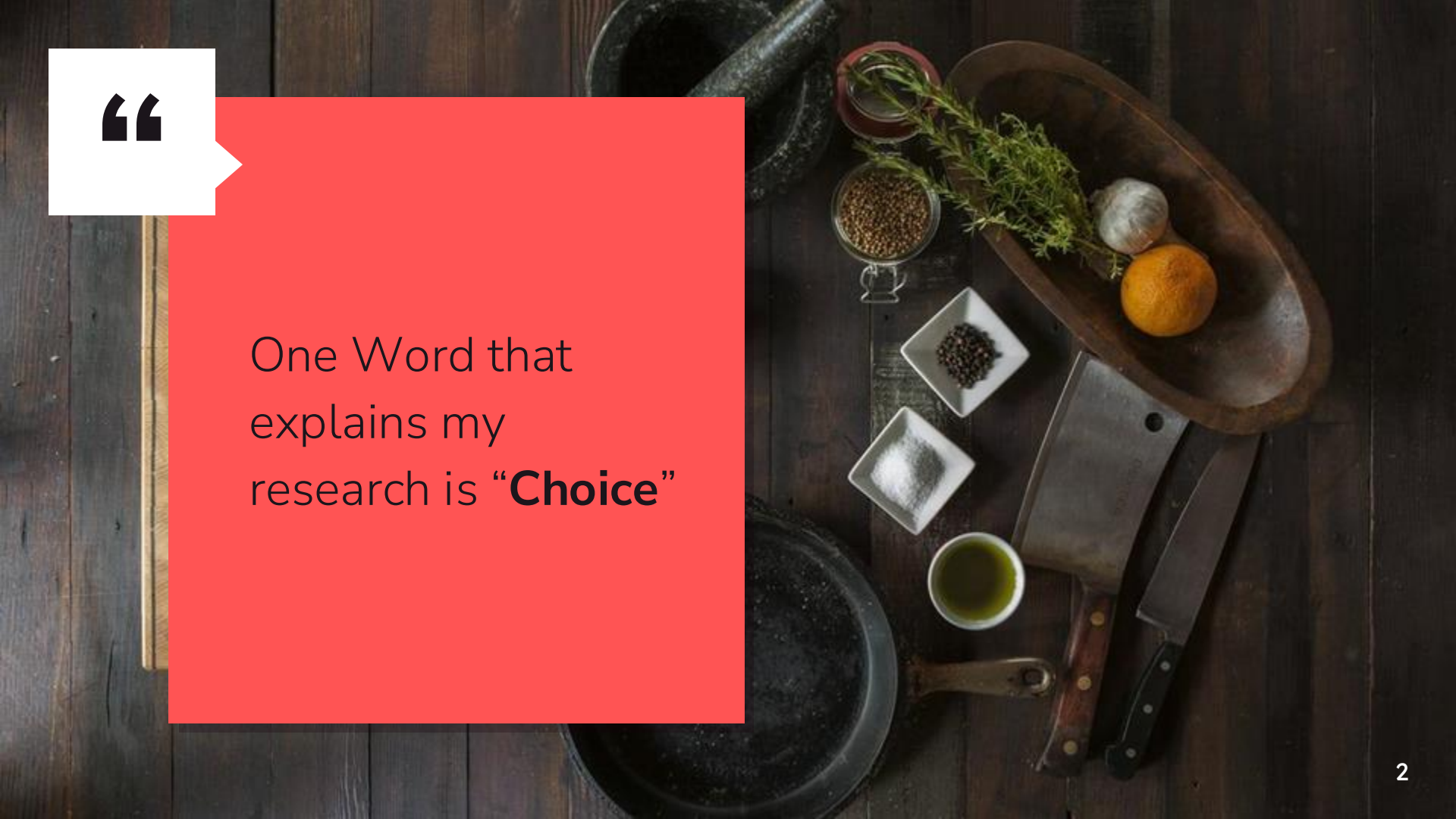
SmartBites:

Nutritional Clustering and Recommender System for Healthier Food Choices

By Margvinatta Senesie



“

One Word that explains my research is “**Choice**”

1 The Big Picture (AND..)

- We are surrounded by food data and bad choices
- 1 in 2 U.S. adults have diet-related chronic disease
- Nutrition labels are confusing or misleading
- Food apps rarely recommend based on nutrient composition

The Problem (BUT..)

- Data alone doesn’t translate to smarter eating
- Data is Overwhelming
- No Standard scoring system
- Nutrient balance isn’t obvious

The Approach (THEREFORE..)

- I built a Data-Driven Food Recommender
- Cleaned & standardized nutrient data
- Created a health score
- Applied unsupervised clustering
- Used cosine similarity to suggest alternatives

Health Score Model

- Health Score = Protein – (Total Fat + Sodium / 100)
- Prioritizes high-protein, low-fat, low-sodium foods
- Top ranked foods: soy protein isolate, egg whites, gelatin powder

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

Visual Insight 1

- Show Most items cluster between 100–300 calories and 0–30g protein.
- But outliers like protein powder stand out with high protein and moderate calories. This supported protein as a key part of my scoring logic.

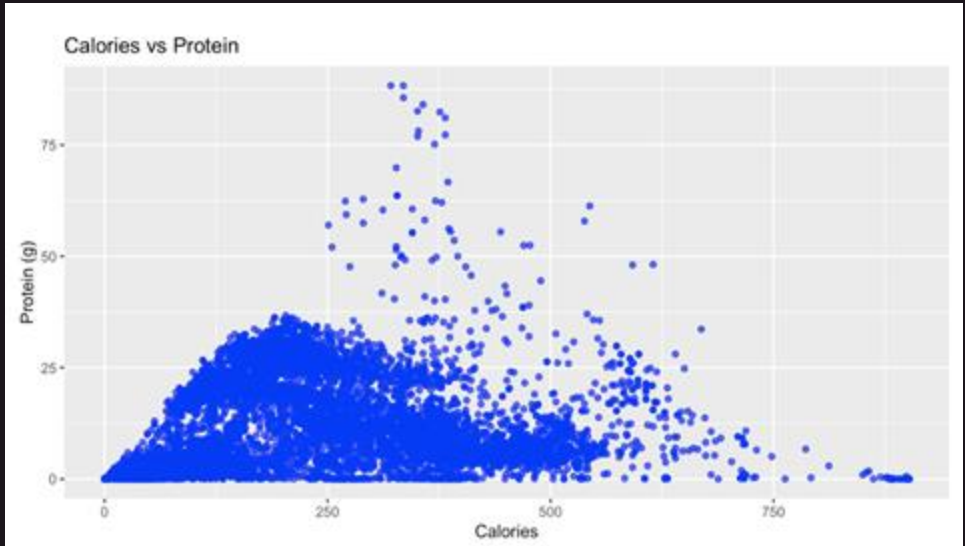


Figure 1: Calories vs Protein

Visual Insight 2

- Foods with extreme sodium levels (30,000+ mg) are visible here.
- These tend to be processed foods. This justified penalizing high sodium in the score.

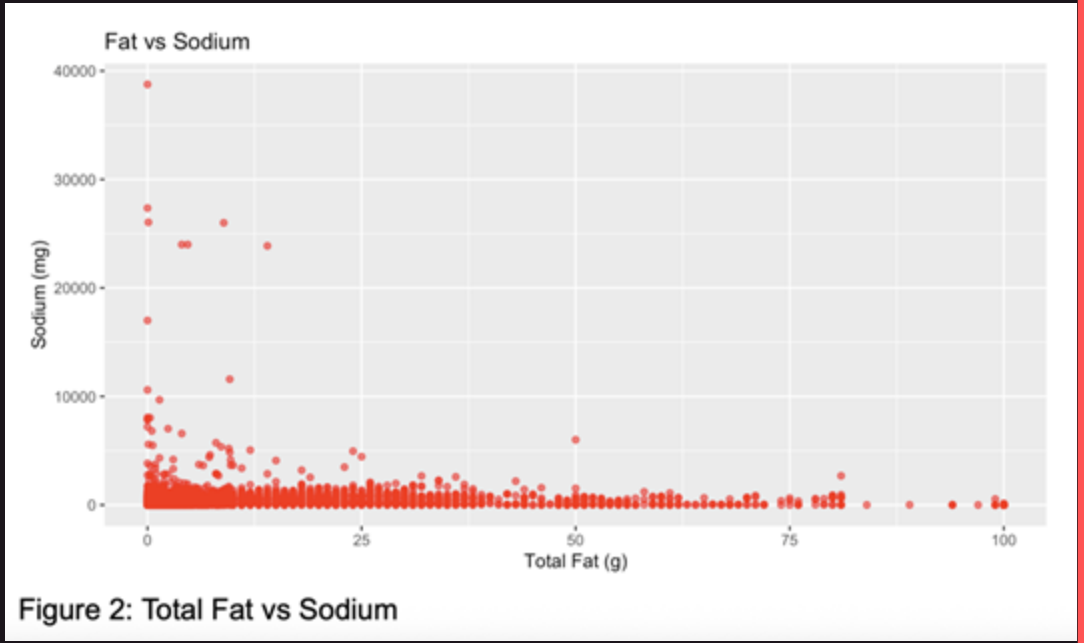
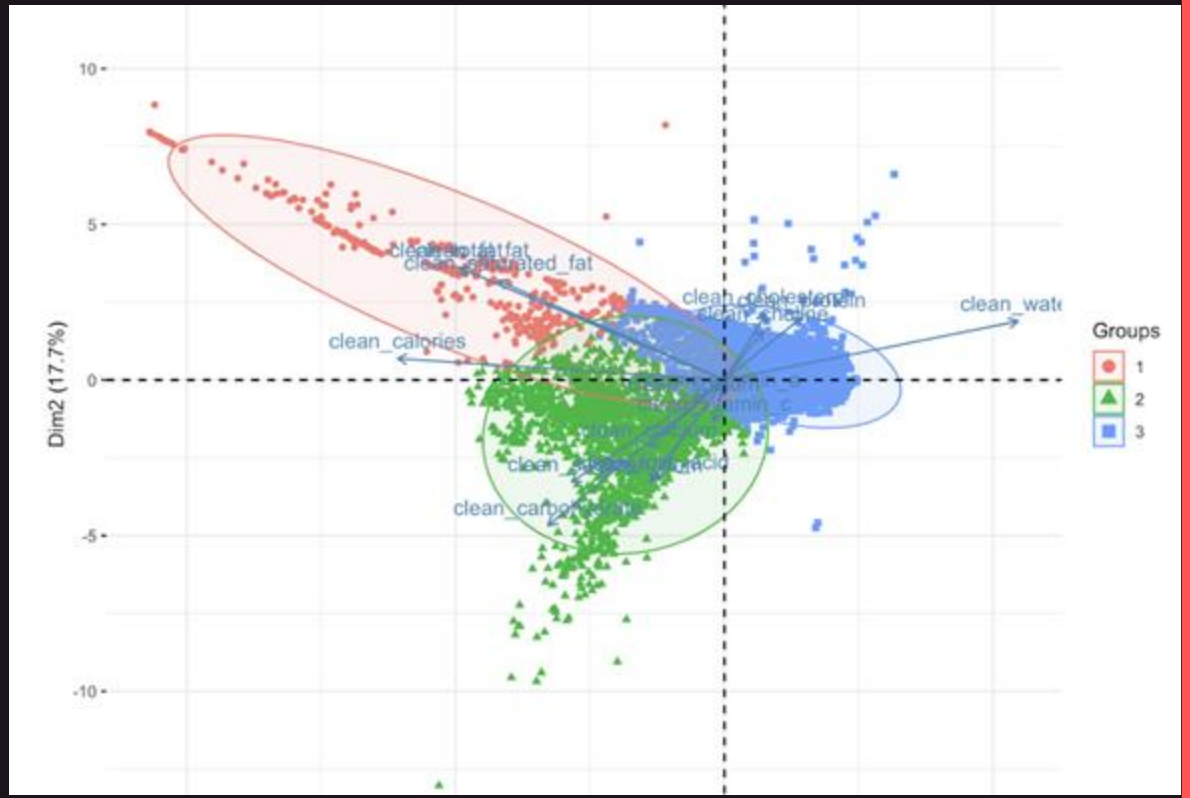


Figure 2: Total Fat vs Sodium

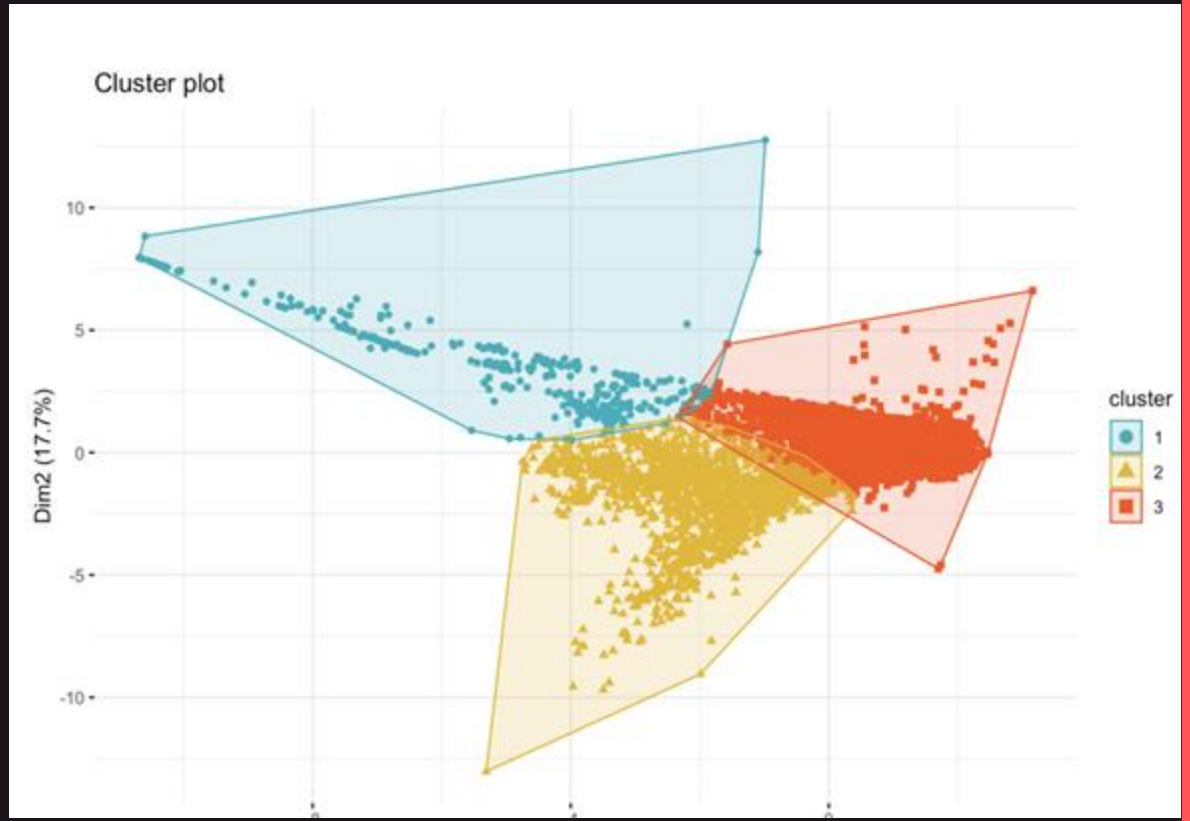
PCA Biplot

- This biplot shows how each nutrient contributed to the clustering.
- For example, protein and water pushed items toward Cluster 3. Fat and calories pointed to Cluster 1.



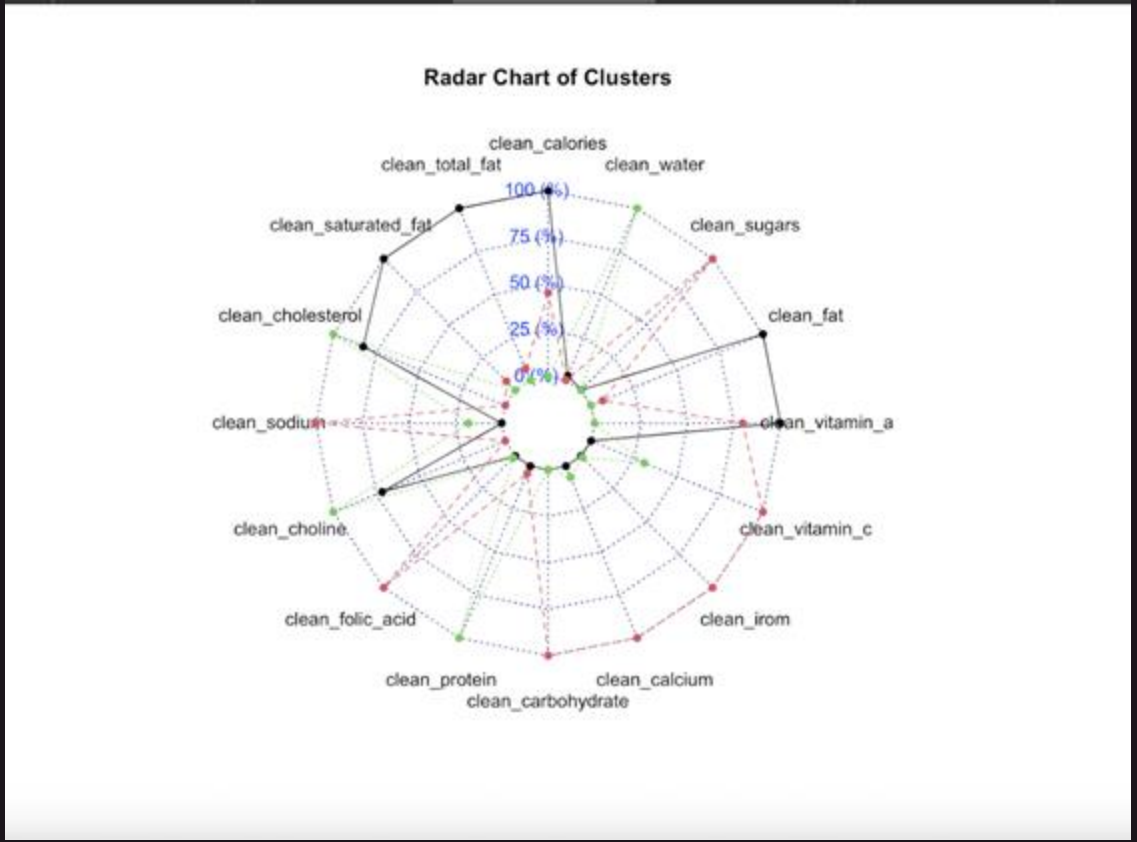
K-means Cluster Plot

- This clearly shows three distinct clusters.
- Blue is high-protein and clean. Yellow is mixed. Red is high-calorie, high-fat. It validated my clustering approach.



Radar Chart

- This is a summary of each cluster's average nutrient profile.
- Cluster 1 is highest in fat, calories, saturated fat.
- Cluster 3 is highest in protein, lowest in sodium.



.

Recommender System

- Suggesting Healthier Substitutes with Cosine Similarity
- Compares nutrient profiles to suggest better alternatives
- Example: healthy granola recommendations

```r
# PHASE 3: Cosine Similarity Recommender
# Compute cosine distance matrix and convert to square matrix
nutrient_matrix <- df %>%
  select(starts_with("clean_")) %>%
  drop_na() %>%
  scale() %>%
  as.matrix()

# Calculate cosine distances
similarity_dist <- proxy::dist(nutrient_matrix, method = "cosine")

# Convert to full matrix
similarity_matrix <- as.matrix(similarity_dist)


# Recommend similar foods
# Pick any food row index (e.g., 100)
food_index <- 100
```

Recommender System

```
  name                                                                health_score cluster
  <chr>                                                                      <dbl> <fct>
1 Snacks, mixed flavors, chewy, KASHI TLC Bar, granola bar                  -0.360 2
2 Cereals ready-to-eat, Wheat and Honey, Oats, 100% Natural Granola…        -1.95  2
3 Snacks, wasabi-flavored, roasted, peas                                    -2.89  2
4 Snacks, mixed flavors, crunchy, KASHI TLC Bar, granola bar                -4     2
5 Babyfood, cookies                                                         -4.2   2
6 Cereals ready-to-eat, Pumpkin Granola, Organic FLAX PLUS, NATURE'…        -7.51  2
7 KEEBLER, KEEBLER Chocolate Graham SELECTS                                 -9.9   2
>
> df_clean[food_index, c("name", "health_score")]
# A tibble: 1 x 2
  name                                    health_score
  <chr>                                          <dbl>
1 Cereals ready-to-eat, homemade, granola        -10.6
>
```

- Food Index = 1590
- Name of Food Item: Cereals, ready-to-eat, homemade, granola (Health Score = -10.6)
- Suggested 7 healthy options ranging from [-9.9, -0.36] in Cluster 2

Recommender System

- Food Index = 55
- Name of Food Item: Frankfurter, meatless (Health Score = 0.9)
- Suggested 20 healthy options ranging from [17.8, 28.4] in Cluster 3

```
# 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
>
```

From Data to Decisions

- Personalized nutrition
- Opportunity for mobile app integration
- Limits: subjective scoring, no fiber/sugar weighting

References

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

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

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

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

Thanks!

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

Any questions?