**PART 1**

Principal Component Analysis (PCA) For Exploring Nutritional Patterns in Foods, part1

**Introduction**

In this report I chose to perform principal component analysis (PCA) on the nutritional profiles of foods. In today’s world where dietary awareness is growing, understanding the nutritional structure of food is crucial not only for individual health decisions but also for public health and food product development, nutrition datasets often involve numerous variables including macronutrients, micronutrients, and energy values which can make pattern recognition difficult due to the complex interrelationships among variables. In this report applies Principal Component Analysis (PCA) to a curated dataset of food items derived from the 2017–2018 USDA Food and Nutrient Database for Dietary Studies (FNDDS), which includes 13 standardised nutrient variables across 7084 rows of food items.

PCA is used to simplify the dataset by reducing the dimensionality while preserving the core structure of the data, enabling the identification of dominant nutritional patterns and food groupings. This approach can highlight latent features such as high-fat, high-calorie or low-fat, high-protein foods, helping to reveal relationships that are not obvious in the raw data.

The mean objective of this analysis is to identify a small set of principal components that explain the majority of variance in the dataset and to interpret how these components relate to the nutritional characteristics of foods.

This report addresses the following key questions:  
• Can PCA effectively reduce the complexity of nutrition data while preserving key information?  
• What underlying nutritional patterns or groupings exist among food items?  
• Which nutrients contribute most significantly to the identified components

**Background**

Nutritional data is inherently multivariate and complex. Each food item typically includes values for calories, fats, carbohydrates, proteins, sugars, fibre, and a wide range of micronutrients. These variables are often intercorrelated for example, high-fat foods often contain higher calories which can make analysis challenging using traditional univariate or bivariate methods.

Principal Component Analysis (PCA) offers a solution to this complexity by reducing the number of variables into a lower set of independent components that explain most of the variation in the dataset. PCA is widely used in fields such as genomics, image processing, and marketing analytics, and has proven useful in nutritional studies for identifying dietary patterns and clustering foods based on nutrient profiles  
By applying PCA to this nutritional dataset, we can simplify the data structure, improve interpretability, and uncover dominant nutritional dimensions that are not immediately obvious in the raw format.

**Data Description**

The dataset I used for this analysis is based on the 2017–2018 USDA Food and Nutrient Database for Dietary Studies (FNDDS), the original dataset contained extensive nutritional profiles for a wide range of food items. I used python to prepare and clean data for this analysis,

in (Figure 1):

* import the necessary libraries and loaded csv file.
* Fixed header and dropped unnecessary rows.
* cleaned and shortened the columns names for clarity.

In (Figure 2):

* convert all nutrient columns to numeric format.
* dropped rows with missing values.
* standardise the nutrient values using z-scores.
* create the final PCA ready dataFrame.
* saved the final dataset to a new csv file.

After cleaning I selected a subset of 13 nutrient variables, including calories, protein, fat, carbohydrates, sugars, fibre, saturated fat, cholesterol, sodium, calcium, iron, and water. These nutrients were chosen for their relevance to nutritional quality and public health interest.

The final dataset used in this analysis '2017\_2018\_Food\_Descriptions.csv' contains standardised values for each of these nutrients. The preparation process ensured that all variables were on the same scale to contribute equally to the PrincipalComponent Analysis regardless of their original units.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 1: Initial Data Loading and Preprocessing

A screenshot of a computer

AI-generated content may be incorrect.

Figure 2: Data Cleaning, Standardisation, and Export

**Results**

The figure 3 scree plot shows that the first three principal components capture the most variance in the dataset, with PC1 alone accounting for approximately 34.6% of the total variance. The cumulative variance reaches over 80% by the fifth component, indicating that a reduced set of components mostly explained the variability in the data without significant information loss.

The figure 4 output confirms that the first five components cumulatively explain about 87% of the total variance. This aligns with the scree plot findings and supports selecting around 4–5 components for interpretation.

In figure 5 loading matrix shows how each nutrient contributes to the principal components. For example:

* PC1 is heavily influenced by Calories, Fat, Carbs, and Water (negatively) suggesting it captures overall energy density.
* PC2 is influenced by Protein, Cholesterol, and Fat, potentially representing high-protein food patterns.
* PC3 loads highly on Iron, Calcium, and Fiber, possibly related to nutrient-rich plant-based or fortified foods.

These loadings provide insight into the dominant nutritional patterns underlying the dataset.

The figure 6 scatter plot of PC1 vs PC2 shows clear variation among food items. Foods on the far right of PC1 may be high in calories and fat, while those with high PC2 scores may be higher in protein.

A graph with a line

AI-generated content may be incorrect.

Figure 3: Scree plot show the proportion of variance explained by each principal component.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 4: Cumulative variance explained by principal components.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 5: Loadings of each nutrient on the first 12 principal components.

**A blue dot diagram on a grid

AI-generated content may be incorrect.**

Figure 6: Distribution of food items in PCA space (PC1 vs PC2). (scatter plot)

**Discussion and Analysis of Results**

The results of the Principal Component Analysis (PCA) show meaningful nutritional patterns that help simplify the complex structure of the dataset, the first three principal components alone explained a large share of the total variance, with PC1 capturing over 34%, and the first five components together explaining more than 80%. This confirms that a small number of derived dimensions can effectively represent the most important variation in food composition.

The nutrient loadings offer further insight into these components, PC1 represents energy density as it is strongly associated with calories, fat, and carbohydrates, and inversely related to water content. This suggests that PC1 identifies differences between high-calorie, energy-dense foods and low-calorie, high-moisture foods. PC2 is driven primarily by protein andcholesterol, likely shows a pattern related to foods high in protein such as meat and dairy.

PC3 highlights a strong contribution from iron, calcium, and fibre, potentially reflecting whole grains, legumes, or fortified products.

The PCA scatter plot (PC1 vs PC2) shows differences between food items, suggesting that there are natural groups or types based on nutrition.

PCA has proven effective in this analysis, is a linear technique, which means it may not capture complex, non-linear relationships between nutrients. Additionally, PCA is unsupervised and does not consider any external outcome variables example like health outcomes, which limits its ability to assess nutritional quality directly. Despite these constraints, PCA remains a valuable exploratory tool for uncovering structure in large, multidimensional datasets like this one.

**Conclusion**

This report applied Principal Component Analysis (PCA) to a standardised USDA nutritional dataset to reduce its complexity and identify dominant patterns, the PCA effectively transformed 13 interrelated nutrient variables into a smaller number of uncorrelated components that explain the larger part of variance in the data.  
Key findings include the identification of components represent energy density, protein content, and micronutrient richness. These components allowed the visualisation of food items in a reduced dimensional space revealing clear differences in nutritional composition across the dataset. The results support the use of PCA as a method for analysing complex nutrition data and for guiding decisions in dietary research, food labelling, or public health planning.

**PART 2**

Two-way ANOVA For the Effect of Cooking Method and Sauce Type on Flavour Intensity, Part 2

**Introduction**

In this report it’s based on theme of food, I test the effects of cooking methods and sauce types on the perceived flavour intensity of food, flavour is a multidimensional sensory experience influenced by preparation techniques, ingredients, and seasoning. In cookery and food science background, understanding how different factors interact to shape flavour perception is critical for improving recipe development, product design, and consumer satisfaction. This analysis uses a recipe dataset sourced from Food.com, where key variables cooking method like grilled, fried, boiled, steamed, baked and sauce type like spicy, sweet, savory, sour were extracted from unstructured recipe texts, a Two-Way ANOVA is applied to explore how these two factors, individually and in combination, influence a simulated dependent variable: flavour intensity score on a 1-10 scale .  
  
This report addresses the following key question:  
Does cooking method and sauce type significantly affect flavour intensity, and is there an interaction between these two factors?

**Background**

Two-Way ANOVA (Analysis of Variance) is used to analyse the effects of two independent categorical variables on a continuous outcome variable, it assesses both the main effects of each factor and their interaction effect, given a deeper understanding of how the two factors influence the response variable individually and together. This method is hugely used in behavioural science, agriculture, marketing, and increasingly, food studies, where sensory evaluation is influenced by multiple factors. In this report, I apply Two-Way ANOVA to food preparation data the dataset, sourced from Food.com Recipes and Interactions dataset on Kaggle, includes over 180,000 user-contributed recipes. From this, a subset was processed by identifying cooking methods through recipe instructions for example grill, boil and sauce types from ingredients and tags for example sweet, spicy. Although actual consumer taste ratings were unavailable, a simulated flavour intensity score was used for demonstration.

The use of ANOVA in food science is widely recognized, particularly for its effectiveness in sensory evaluation. It is instrumental in detecting statistically significant differences among food products prepared under different conditions. For example, Ares et al. (2014) highlighted the application of ANOVA alongside methods like CATA (Check-All-That-Apply) to analyses consumer preferences in food studies. By applying these statistical techniques, researchers can uncover key factors influencing sensory perceptions. This report uses a Two-Way ANOVA to investigate how cooking methods and sauce type interact to affect perceived flavour intensity. These methods support the goal of this report to statistically test the influence of cooking technique and sauce style on flavour intensity in a factorial design framework.

Before running the Two-Way ANOVA, it was important to consider a few key assumptions that make this type of analysis valid:

1. **Independence,** each observation is assumed to be collected independently of the others.
2. **Normality,** ANOVA assumes that the residuals, which represent the deviation between observed and model-predicted values, follow a normal distribution, while the dependent variable in this project was simulated using a uniform scale from 1 to 10, the balanced design and moderate sample size to make this assumption reasonably and acceptable.
3. **Equal Variance,** the spread of scores (variance) across all groups is assumed to be exactly the same.
4. **Measurement Scale,** the outcome being measured (Flavour Intensity Score) should be continuous and measured on an interval scale. Although the scores are integers, the 1-10 range is enough for analysis.

Based on this setup, the following hypotheses were tested in the analysis:

**Main Effect – Cooking Method**

* **Null hypothesis (H₀)**: All cooking methods result in similar average flavour intensity scores.
* **Alternative hypothesis (H₁)**: At least one cooking method made a significantly different flavour intensity score.

**Main Effect – Sauce Type**

* **H₀**: There is no difference in average flavour intensity between sauce types.
* **H₁**: At least one sauce type leads to a different flavour intensity score.

**Interaction Effect – Cooking Method × Sauce Type**

* **H₀**: There is no interaction the effect of sauce type does not depend on the cooking method.
* **H₁**: There is an interaction the effect of one factor changes depending on the level of the other.

**Data description**

The dataset I used in this analysis is retrieved from the RAW\_recipes.csv file, which contains thousands of user-submitted recipes to prepare this data for a Two-Way ANOVA, the following steps were performed using Python:  
  
**Figure 1: Loading the Dataset**  
The dataset was loaded using the panda’s library it contains text based fields such as recipe steps, tags, and ingredients extracted from the RAW\_recipes.csv file.

**Figures 1 and 2: Feature Extraction**

* **Cooking Method**: A custom Python function parsed the 'steps' column to identify keywords such as grill, fry, bake, boil, and steam based on these, each recipe was used to assign cooking method.
* **Sauce Type**: A second function analysed the tags and ingredients columns to classify recipes by sauce or flavour profile keywords such as sweet, spicy, savory, and sour were used to assign a sauce type.

**Figure 3: Data Cleaning and Variable Creation**  
To make sure that dataset was suitable for Two-Way ANOVA, all rows missing values for either factor were removed using dropna function.  
A new column, Flavour Intensity Score, was then created using np.random.randint function to assign random scores between 1 and 10, simulating sensory evaluation data.

**Figure 4: Balancing the Design**  
The cleaned data was grouped by Cooking Method and Sauce Type, and exactly 3 observations were randomly sampled from each of the 20 unique combinations this resulted in a 5×4 factorialdesign with 3 replications per cell, totalling 60 rows.

The final dataset includes the following columns:

* Cooking Method (Factor 1: Grilled, Fried, Baked, Boiled, Steamed)
* Sauce Type (Factor 2: Sweet, Spicy, Savory, Sour)
* Flavour Intensity Score (Dependent variable: 1–10 scale)

This cleaned dataset was saved as FlavorIntensity\_ANOVA\_Dataset.csv and used for the subsequent Two-Way ANOVA analysis.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 1: Loading the Dataset and Extracting Cooking Method from Recipe Steps

A screenshot of a computer

AI-generated content may be incorrect.

Figure 2: Extracting Sauce Type from Tags and Ingredients

A screenshot of a computer

AI-generated content may be incorrect.

Figure 3: Cleaning the Data and Generating Simulated Flavour Scores (1-10)

A screenshot of a computer

AI-generated content may be incorrect.

Figure 4: Sampling for a Balanced 5×4 Factorial Design and Export

**Results**

The figure 5 shows the results from the Two-Way ANOVA revealed the following effects on Flavour Intensity Score:

* The main effect of Cooking Method was statistically significant p = 0.047. Therefore, we reject the null hypothesis, indicating that different cooking methods result in significantly different flavour intensity ratings.
* The main effect of Sauce Type was not significant p = 0.55. Therefore, we fail to reject the null hypothesis. Show no significant differences in flavour intensity across different sauce types.
* The interaction between Cooking Method and Sauce Type was also not statistically significant p = 0.18. Therefore, we fail to reject the null hypothesis. Show that the combined effect of these two factors does not significantly influence flavour intensity.

The figure 6 Shows the bar plot representing the mean flavour intensity scores for each cooking method. Figure 7 shows the interaction plot illustrating how sauce types interact with cooking methods.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 5: Importing Libraries and Two-Way ANOVA Output Table with Sum of Squares, Degrees of Freedom, F-Values, and p-values

A graph of different colored squares

AI-generated content may be incorrect.

Figure 6: Main Effect of Cooking Method on Flavour Intensity. (Bar plot)

A diagram of different colored lines

AI-generated content may be incorrect.

Figure 7: Interaction Plot of Cooking Method × Sauce Type on Flavour Intensity. (pointplot)

**Discussion and Analysis of Results – Main Effects**

The main effect of Cooking Method was statistically significant (p = 0.047), indicating that flavour intensity ratings varied depending on how the food was prepared, we reject the null hypothesis for cooking method. Among the methods, fried dishes received the highest mean flavour scores, followed by grilled and baked items. In other hands, steamed foods received the lowest ratings, suggesting that certain cooking techniques enhance perceived flavour intensity more than others.  
For the main effect of Sauce Type, the result was not statistically significant (p = 0.55), and therefore, we fail to reject the null hypothesis, this implies that was no strong evidence to decide that the type of sauce (sweet, spicy, savory, or sour) significantly impacted the overall flavour intensity score when averaged across cooking methods.

**Discussion and Analysis of Results – Interaction Effect**

The interaction between Cooking Method and Sauce Type was not statistically significant. However, visual examination of the interaction plot reveals some non-parallel lines, indicating potential practical (though not statistically confirmed) interactions between certain combinations. For example, sweet sauces paired with fried dishes scored higher than when paired with steamed items. These insights may be useful for culinary product development.

The interaction between Cooking Method and Sauce Type was also not statistically significant (p = 0.18), so we fail to reject the null hypothesis for the interaction effect. This suggests that the effect of one factor cooking method on flavour intensity does not significantly depend on the level of the other factor sauce type.

However, visual inspection of the interaction plot (Figure 7) reveals some non-parallel lines and crossing trends, which may suggest practical interactions even if they are not statistically significant. For instance, sweet sauces paired with fried foods appeared to produce higher flavour intensity scores than when the same sauces were used with steamed items. While these differences weren’t strong enough to be statistically confirmed, they may provide useful insights for future product development and culinary testing.

**Conclusion**

This report aimed to explore how cooking methods and sauce types influence perceived flavour intensity of food. The result of two-way ANOVA shows that cooking method has a statistically significant impact, while sauce type did not show a significant effect and the interaction between the two factors was also not statistically significant, visual analysis of the interaction plot revealed patterns that may hold practical relevance. These findings provide a deeper understanding of how preparation methods impact sensory experience and highlight opportunities for future research in food science and product development.