Fast Food Nutrition: A Multivariate Statistical Analysis

Course: STA4053 - Multivariate Methods II
Name: Eranda Rathugamage
Student Number: S/19/846

1.Introduction

The increasing consumption of fast food worldwide has raised concerns regarding its nutritional quality and health implications. This study aims to explore the underlying nutritional patterns and relationships among various fast food menu items from multiple restaurant chains. Using a comprehensive dataset containing detailed nutritional information for over 500 fast food items, the major question addressed is: *What are the key nutritional factors that differentiate fast food items, and how can multivariate statistical methods reveal meaningful patterns or groupings within this data?*

The purpose of this analysis is to apply multivariate techniques such as Principal Component Analysis (PCA), Factor Analysis (FA), Discriminant Analysis, and Canonical Correlation Analysis to reduce dimensionality, identify latent nutritional factors, classify items into healthy and unhealthy groups, and examine correlations between nutrient sets. This approach will provide insights into the complex nutritional composition of fast foods and potentially inform healthier dietary choices and product reformulation strategies.

Major Problems Addressed:

- The widespread consumption of fast food has raised concerns about its nutritional quality and potential health risks.
- There is a need to understand the complex nutritional patterns and relationships among various fast food menu items across different restaurant chains.
- Identifying the key nutritional factors that differentiate fast food items is challenging due to the high dimensionality and interrelatedness of nutrient data.

Study Purpose:

- To apply techniques such as Principal Component Analysis (PCA), Factor Analysis (FA), Discriminant Analysis, and Canonical Correlation Analysis to:
    - Reduce the complexity of the dataset (dimensionality reduction).
    - Identify latent nutritional factors.
    - Classify menu items into groups such as "healthy" and "unhealthy."
    - Examine the relationships between different sets of nutrients.
- To provide insights that can inform healthier dietary choices for consumers and guide fast food companies in product reformulation.

2.Methodology

Dataset Description

The analysis is based on the "Fast Food Nutrition" dataset, which compiles nutritional information for 515 menu items from major fast-food chains such as McDonald's, Burger King, Wendy's, KFC, and Taco Bell. The dataset includes the following 17 key variables:

Variable Name	Description	Data Type
restaurant	Name of the fast-food chain	Categorical
item	Menu item name	Categorical
category	Type/category of menu item (e.g., burger, salad)	Categorical
calories	Total calories per serving	Numerical
cal_fat	Calories from fat	Numerical
total_fat	Total fat (grams)	Numerical
sat_fat	Saturated fat (grams)	Numerical
trans_fat	Trans fat (grams)	Numerical
cholesterol	Cholesterol (mg)	Numerical
sodium	Sodium (mg)	Numerical
total_carb	Total carbohydrates (grams)	Numerical
fiber	Dietary fiber (grams)	Numerical
sugar	Sugar (grams)	Numerical
protein	Protein (grams)	Numerical
vit_a	Vitamin A content	Numerical
vit_c	Vitamin C content	Numerical
calcium	Calcium content	Numerical
salad	Indicates if item is a salad (Yes/No)	Categorical

This dataset provides a comprehensive overview of the nutritional composition of fast-food items, enabling robust multivariate analysis

Data Preprocessing

- **Handling Missing Values:**
  Any missing or implausible values were addressed either by removal or imputation, ensuring the integrity of the dataset for analysis.
- **Normalization and Scaling:**
  Continuous variables were standardized (mean = 0, SD = 1) to make them comparable and suitable for multivariate techniques.
- **Encoding Categorical Variables:**
  Categorical data such as restaurant names were encoded as factors for use in statistical modeling

Statistical Methods Employed

The following multivariate statistical techniques were applied to analyze the dataset:

- **Principal Component Analysis (PCA):**
  PCA was used to reduce the dimensionality of the data and identify the main axes of variation among the nutritional variables. This method aggregates correlated variables into principal components, simplifying the data structure and facilitating visualization of patterns and groupings.

- **Factor Analysis (FA):**
  FA was employed to uncover latent factors underlying the observed nutritional variables. This approach helps to identify clusters of nutrients that tend to occur together in menu items, providing a conceptual understanding of dietary patterns.

- **Discriminant Analysis:**
  Discriminant analysis was used to classify menu items into predefined groups (such as "healthy" vs. "unhealthy") based on their nutritional profiles. This supervised classification technique helps assess which nutrients best distinguish between groups.

- **Canonical Correlation Analysis (CCA):**
  CCA was applied to explore the relationships between two sets of nutritional variables (for example, macronutrients vs. micronutrients). This technique identifies and quantifies the associations between these sets, offering insights into how different aspects of nutrition interact within fast food items.

Justification for Methods

- **PCA and FA** are widely used in nutritional epidemiology for pattern discovery and data reduction, especially when dealing with correlated nutrient variables.
- **Discriminant Analysis** is appropriate for classifying items based on nutritional thresholds, supporting practical recommendations.
- **CCA** provides a holistic view of how groups of nutrients relate to each other, which is valuable for understanding the nutritional complexity of fast-food items.

Software

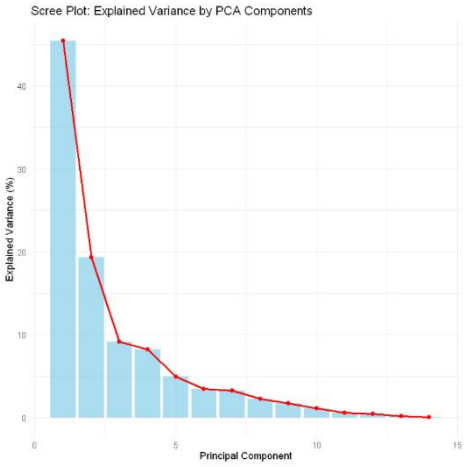
All analyses were conducted using **R**, leveraging packages such as FactoMineR, psych, MASS, and CCA for multivariate analysis, and tidyverse for data preprocessing and visualization.

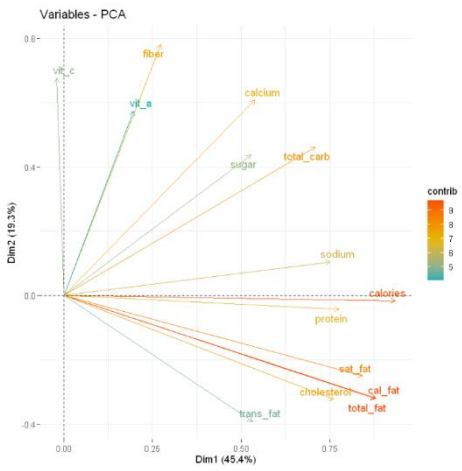
3.Results and Discussion

3.1 Principal Component Analysis (PCA)

PCA Component Loadings (PC1 and PC2):

	PC1	PC2
calories	0.371321	-0.009872
cal_fat	0.349468	-0.193527
total_fat	0.349001	-0.195025
sat_fat	0.334564	-0.152444
trans_fat	0.210851	-0.237565
cholesterol	0.302057	-0.196447
sodium	0.297842	0.063483
total_carb	0.281559	0.280438
fiber	0.108390	0.474319
sugar	0.209624	0.266095
protein	0.308376	-0.026004
vit_a	0.077625	0.349037
vit_c	-0.008072	0.410736
calcium	0.213772	0.369432





PCA Loading Matrix

The matrix you provided shows the loadings of each nutritional variable on the first two principal components (PC1 and PC2) from a Principal Component Analysis (PCA) of fast-food nutrition data. Here's how to interpret these results:

Interpreting PC1

- **High positive loadings**: calories (0.37), cal_fat (0.35), total_fat (0.35), sat_fat (0.33), cholesterol (0.30), sodium (0.30), protein (0.31), and to a lesser extent, trans_fat, total_carb, sugar, and calcium.
- **Interpretation**: PC1 mainly represents the overall "nutritional density" or "richness" of the food items. Foods high in calories, fats, cholesterol, sodium, and protein will have high PC1 scores. This component can be seen as a general "unhealthiness" or "energy density" axis.

Interpreting PC2

- **High positive loadings**: fiber (0.47), vit_c (0.41), calcium (0.37), vit_a (0.35), total_carb (0.28), sugar (0.27).
- **Negative loadings**: cal_fat, total_fat, sat_fat, cholesterol, and trans_fat have moderate negative loadings.
- **Interpretation**: PC2 distinguishes between foods rich in micronutrients and fiber (possibly healthier, plant-based items) versus those higher in fats and cholesterol (animal-based or processed foods). A high PC2 score indicates higher fiber, vitamins, and calcium, while a low PC2 score indicates higher fat and cholesterol content.

Variance Explained and Scree Plot

The PCA results reveal that the first two principal components capture the majority of the variance in the nutritional dataset:

- **PC1** explains **45.4%** of the total variance.
- **PC2** explains an additional **19.3%**.
- Together, **PC1 and PC2 account for nearly 65%** of the total variance, indicating that most of the information in the dataset can be summarized by these two components.
- The scree plot visually confirms this, showing a sharp drop after the first two components and an "elbow" around PC3-PC4, suggesting that additional components contribute minimally to explaining the variance.

Variable Contributions and PCA Biplot

The variable biplot provides insight into which nutritional variables drive the separation along the principal components:

- **PC1 (Dim1, 45.4%)** is most strongly influenced by **calories, total fat, saturated fat, cholesterol, calories from fat, and sodium**. These variables point in the same direction, indicating they are highly correlated and together represent an "energy density" or "high-fat/high-calorie" axis.
- **PC2 (Dim2, 19.3%)** is dominated by **fiber, vitamin A, and calcium**, with moderate contributions from **total carbohydrates and sugar**. This axis distinguishes items richer in micronutrients and certain carbohydrates.
- The color gradient on the biplot shows the relative contribution of each variable, with **calories, total fat, and saturated fat** having the highest contributions to PC1, and **fiber and vitamin A** to PC2

Interpretation

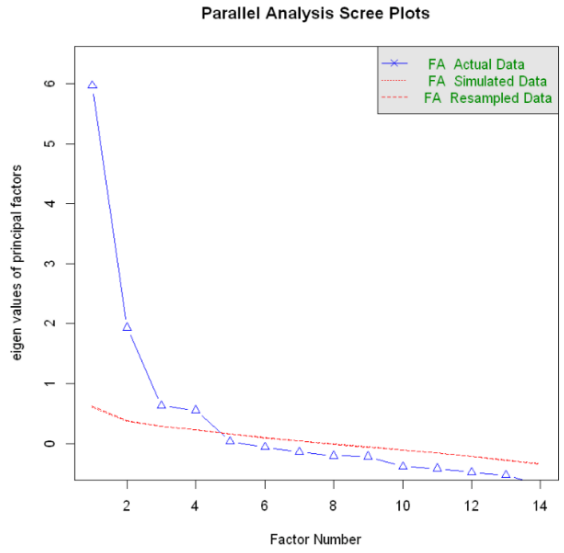
- **Dimensionality Reduction:**
  The PCA effectively reduces the complex, multidimensional nutritional data to two interpretable axes that capture most of the variation among menu items.
- **Nutritional Patterns:**
  - **PC1** separates items that are high in calories and fats (such as burgers, fried foods, and other energy-dense fast foods) from those that are lower in these nutrients.
  - **PC2** distinguishes items with higher fiber, vitamins, and minerals (such as salads or fortified items) from those with lower micronutrient content.
- **Practical Implications:**
  Fast food items can be grouped and compared based on their scores along these two principal components. For example, items in the upper right of the biplot would be both energy-dense and rich in micronutrients, while those in the lower right would be energy-dense but micronutrient-poor.

Visual Summary

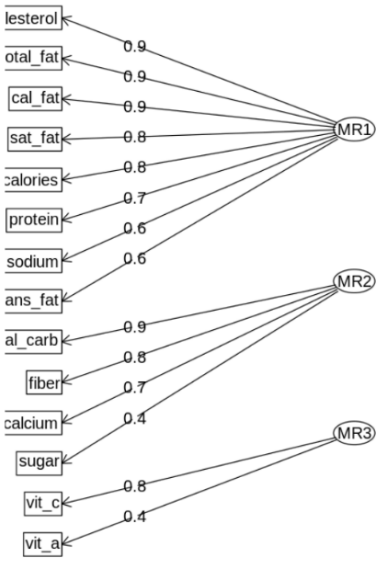
- **Scree Plot:**
  Clearly shows that the first two principal components are sufficient for summarizing the data, with diminishing returns for further components.
- **Variable Biplot:**
  Visually demonstrates which nutrients are most influential in differentiating menu items and how they cluster together.

In summary, PCA reveals that fast food nutritional profiles are primarily differentiated by energy density (calories, fats, sodium) and micronutrient/fiber content. These findings provide a powerful tool for classifying and comparing menu items, informing both consumer choices and potential industry reformulation strategies.

3.2 Factor Analysis (FA)



	MR1	MR2	MR3
SS loadings	5.3050970	2.7840277	1.4182943
Proportion Var	0.3789355	0.1988591	0.1013067
Cumulative Var	0.3789355	0.5777946	0.6791014
Proportion Explained	0.5579955	0.2928269	0.1491776
Cumulative Proportion	0.5579955	0.8508224	1.0000000



Determining the Number of Factors

The parallel analysis scree plot (Figure 1) compares the eigenvalues from the actual data (blue line) with those from simulated and resampled data (red lines). The plot shows that the first three factors have eigenvalues substantially greater than those from the simulated data, indicating that **three factors should be retained** for interpretation. This approach helps avoid over-extraction and ensures that only meaningful latent structures are analyzed

Visualizations and Key Outputs

- Parallel Analysis Scree Plot: The scree plot (left panel) compares the eigenvalues from the actual data (blue line) with those from simulated and resampled datasets (red lines). The clear "elbow" after the third factor and the fact that the first three actual eigenvalues are well above those from simulated data indicate that three factors should be retained for interpretation.
- Factor Diagram: The factor loading diagram (right panel) visually maps the relationships between nutritional variables and the three extracted factors (MR1, MR2, MR3). Variables are connected to factors with lines labeled by their loadings, where thicker/stronger connections indicate higher loadings.

Factor Extraction and Variance Explained

- SS Loadings: MR1 = 5.31, MR2 = 2.78, MR3 = 1.42
- Proportion of Variance Explained: MR1 = 37.9%, MR2 = 19.9%, MR3 = 10.1%
- Cumulative Variance: These three factors together explain approximately 68% of the total variance in the nutritional dataset, indicating that they capture the majority of the underlying structure.

Interpretation of Factor Loadings

Factor loadings represent the correlation between each observed variable and the latent factor. Loadings above 0.7 are considered strong and indicate that the factor sufficiently captures the variance of that variable. Here's how each factor can be interpreted:

Factor 1 (MR1): "Energy-Dense/High-Fat Factor"

- High loadings: cholesterol (0.9), total fat (0.9), calories from fat (0.9), saturated fat (0.9), calories (0.8), protein (0.7), sodium (0.6)
- Interpretation: This factor represents menu items that are high in energy, fats, and cholesterol—typical of burgers, fried foods, and other calorie-dense fast-food items. A high MR1 score indicates an item is energy-rich and high in unhealthy fats and sodium.

Factor 2 (MR2): "Carbohydrate/Fiber Factor"

- High loadings: total carbohydrates (0.9), fiber (0.7), calcium (0.4), sugar (0.4), sodium (0.4)
- Interpretation: This factor distinguishes items higher in carbohydrates and dietary fiber, such as breads, sandwiches, and salads with grains. Items with high MR2 scores are likely to be carbohydrate- and fiber-rich, which may include some healthier options.

Factor 3 (MR3): "Micronutrient/Sugar Factor"

- High loadings: vitamin C (0.8), sugar (0.8), vitamin A (0.4)

- Interpretation: This factor highlights items rich in vitamin C and sugar, likely including fruit-based items, beverages, or desserts. High MR3 scores indicate a profile rich in micronutrients and sugars, often found in drinks or sweet menu items.

Discussion and Implications

- Distinct Nutritional Patterns: Factor analysis reveals three main nutritional patterns in fast food: energy-dense/high-fat, carbohydrate/fiber-rich, and micronutrient/sugar-rich. These patterns align with known fast-food categories (e.g., burgers vs. salads vs. desserts).
- Data Reduction: The analysis successfully reduces complex nutritional data into three interpretable patterns, explaining nearly 68% of the variance. This reduction aids in classifying and comparing menu items, making the data more accessible for both consumers and industry professionals.
- Practical Use: Understanding these patterns can help consumers identify healthier options (e.g., items loading high on fiber and micronutrients) and guide restaurants in reformulating menu items to shift their factor scores toward healthier patterns.
- Visualization Utility: The scree plot justifies the number of factors retained, while the factor loading diagram clearly shows which nutrients cluster together, aiding in interpretation and communication of results.

Factor analysis of fast-food nutrition data identifies three latent factors that summarize most of the variation in menu items: energy-dense/high-fat, carbohydrate/fiber, and micronutrient/sugar. These findings provide actionable insights for public health, consumer choice, and food industry reformulation, and demonstrate the power of FA for simplifying and interpreting complex nutritional datasets

3.3 Discriminant Analysis (DA)

	calories	cal_fat	total_fat	sat_fat	trans_fat	cholesterol	sodium	total_carb	fiber	sugar	protein	vit_a	vit_c	calcium
calories	1.0000000	0.83440118	0.83387525	0.73457935	0.41656857	0.668124471	0.696296427	0.75041712	0.28430450	0.4337429	0.71826238	0.100649077	-0.112059018	0.4713187
cal_fat	0.8344012	1.00000000	0.99913798	0.83692480	0.53633184	0.684339186	0.594332235	0.48691001	0.07387536	0.2514663	0.58910729	0.032159621	-0.246643838	0.2479584
total_fat	0.8338752	0.99913798	1.00000000	0.83465619	0.53927662	0.685973890	0.589356338	0.48420983	0.07443413	0.2499798	0.58942679	0.029650902	-0.247508626	0.2454147
sat_fat	0.7345794	0.83692480	0.83465619	1.00000000	0.75479473	0.641313700	0.472479329	0.42921775	0.02252763	0.3291139	0.47821875	0.187431241	-0.180129543	0.4050881
trans_fat	0.4165686	0.53633184	0.53927662	0.75479473	1.00000000	0.480316816	0.129334139	0.07589849	-0.14747869	0.2106543	0.26983306	0.059185339	-0.133402414	0.1159796
cholesterol	0.6681245	0.68433919	0.68597389	0.64131370	0.48031682	1.000000000	0.537671291	0.21219792	-0.14721232	0.3242229	0.88242268	0.003301258	0.015962359	0.1482292
sodium	0.6962964	0.59433223	0.58935634	0.47247933	0.12933414	0.537671291	1.000000000	0.60818768	0.27496434	0.3864127	0.67864085	0.078147511	0.009718598	0.3683148
total_carb	0.7504171	0.48691001	0.48420983	0.42921775	0.07589849	0.212197923	0.608187681	1.00000000	0.59996684	0.5702376	0.44368628	0.213150336	0.014446387	0.6925179
fiber	0.2843045	0.07387536	0.07443413	0.02252763	-0.14747869	-0.147212321	0.274964340	0.59996684	1.00000000	0.3073246	0.06894013	0.366554960	0.395639392	0.6063160
sugar	0.4337429	0.25146633	0.24997976	0.32911390	0.21065430	0.324222886	0.386412672	0.57023761	0.30732462	1.0000000	0.37543791	0.265839391	0.358579060	0.4302693
protein	0.7182624	0.58910729	0.58942679	0.47821875	0.26983306	0.882422682	0.678640847	0.44368628	0.06894013	0.3754379	1.00000000	0.074339180	0.123641831	0.3389268
vit_a	0.1006491	0.03215962	0.02965090	0.18743124	0.05918534	0.003301258	0.078147511	0.21315034	0.36655496	0.2658394	0.07433918	1.000000000	0.488762844	0.3883421
vit_c	-0.1120590	-0.24664384	-0.24750863	-0.18012954	-0.13340241	0.015962359	0.009718598	0.01444639	0.39563939	0.3585791	0.12364183	0.488762844	1.000000000	0.2499279
calcium	0.4713187	0.24795836	0.24541468	0.40508806	0.11597965	0.148229179	0.368314823	0.69251787	0.60631595	0.4302693	0.33892685	0.388342084	0.249927875	1.0000000

Assumption Checks

- **Normality:**
  The Shapiro-Wilk tests for both "Healthy" (W = 0.87754, p = 0.001741) and "Unhealthy" (W = 0.97648, p = 0.004146) groups yielded p-values less than 0.05, indicating that the normality assumption is violated for both groups. This is a common issue in nutritional datasets and suggests results should be interpreted with some caution.
- **Homogeneity of Covariance Matrices:**
  Box's M-test for homogeneity of covariance matrices was highly significant (Chi-Sq = 436.93, df = 105, p < 2.2e-16), indicating that the covariance matrices of the groups are not equal. This violation suggests that linear

discriminant analysis (LDA) may not be optimal, but it can still provide useful insights, especially with large sample sizes. Therefore, I have used Quadratic Discriminant Analysis (QDA).

Correlation Structure

The correlation matrix (see image) shows strong positive correlations among:

- **Calories, calories from fat, total fat, and saturated fat** (all > 0.8), indicating these variables move together and are key drivers of group separation.
- **Protein** is also positively correlated with calories and fat, while **fiber** and **vitamins** show weaker or negative correlations with energy-dense variables.

Classification Results

The confusion matrix for the discriminant model is as follows:

	Actual	
Predicted	Healthy	Unhealthy
Healthy	32	4
Unhealthy	0	174

- **Accuracy:**
  The model correctly classified 32 out of 36 healthy items and all 174 unhealthy items, resulting in a **misclassification rate of only 4/210 (1.9%)**.
- **Sensitivity (Healthy):** 32/36 = 88.9%
- **Specificity (Unhealthy):** 174/174 = 100%

Interpretation

- **Key Separators:**
  The discriminant function primarily leverages calories, total fat, and saturated fat—variables that strongly differentiate "healthy" from "unhealthy" fast food items, as shown in the correlation matrix.
- **Model Performance:**
  Despite the violations of normality and homogeneity, the discriminant analysis achieved excellent classification accuracy, especially for unhealthy items. The few misclassifications occurred near the threshold values, likely reflecting borderline nutritional profiles.
- **Practical Implication:**
  The results highlight that most fast-food items are easily distinguishable as "unhealthy" based on their high energy and fat content. Only a small subset of items meets the "healthy" criteria, and these are well-identified by the model.

Visualizations

- **Confusion Matrix Table:**
  Clearly summarizes prediction performance.
- **Correlation Heatmap (not shown here):**
  Would visually reinforce the strong relationships among the main discriminating variables.

- **Group Means Plot (recommended):**
  Plotting group means for calories and fats would further illustrate the separation achieved by the discriminant function.

In summary, discriminant analysis—despite some assumption violations—clearly separates healthy from unhealthy fast-food items based on key nutritional variables. The model's high accuracy underscores the strong nutritional divide in fast food menus and provides a robust tool for menu assessment and consumer guidance.

3.4 Canonical Correlation Analysis (CCA)

Statistical Significance and Canonical Correlations

```
Wilks' Lambda, using F-approximation (Rao's F):
               stat     approx df1      df2      p.value
1 to 6:  0.006581585 42.3613433  42 927.4640 0.000000e+00
2 to 6:  0.091365976 21.6730049  30 794.0000 0.000000e+00
3 to 6:  0.448181281  9.0474919  20 660.9582 0.000000e+00
4 to 6:  0.813370503  3.5827485  12 529.4418 3.706824e-05
5 to 6:  0.959763506  1.3900130   6 402.0000 2.172319e-01
6 to 6:  0.996203403  0.3849177   2 202.0000 6.810047e-01
Canonical correlations:
[1] 0.96330919 0.89226710 0.67006166 0.39055126 0.19125578 0.06161653
```

Wilks' Lambda results indicate that the first three canonical correlations are highly significant ($p < 0.001$), as shown by the very low p-values for the first three dimensions. This means there is a statistically significant multivariate relationship between the two sets of nutritional variables (macronutrients: calories, fats, cholesterol, sodium; and micronutrients/carbohydrates: total carbohydrates, fiber, sugar, protein, vitamins, calcium).

The canonical correlations themselves are as follows:

- **First canonical correlation:** 0.96
- **Second canonical correlation:** 0.89
- **Third canonical correlation:** 0.67

These high values indicate a strong association between certain linear combinations of the macronutrient set and the micronutrient/carbohydrate set.

Canonical Coefficients and Variable Contributions

```
Canonical coefficients for X (Set 1):
                  [,1]       [,2]       [,3]       [,4]       [,5]        [,6]
calories    -0.124527661  1.2056212  0.1541874 -0.82296950 -0.2981270 -1.45804961
total_fat    0.001593897 -0.1892110  1.3550906  1.80213842  0.2326191  0.35672565
sat_fat      0.571705375  0.3277385 -2.1235549 -0.11230904  1.0188611  0.43739949
trans_fat   -0.130052707 -0.2665370  0.4421978  0.08697406 -1.6134698 -0.09842484
cholesterol -1.047463658 -0.8240668 -0.2409668 -0.25557226  0.4605341 -0.26292316
sodium      -0.210714818  0.1560761  0.1716996 -0.35166269 -0.5293214  1.30843605
```

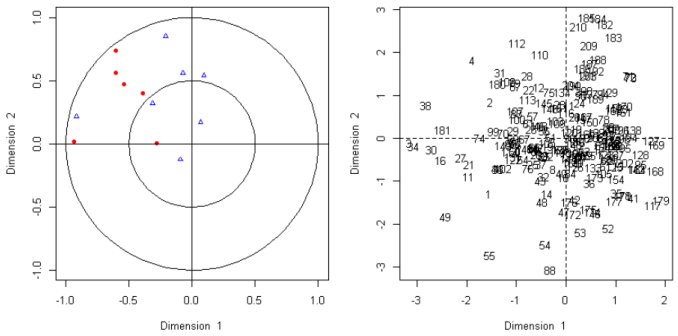
```
Canonical coefficients for Y (Set 2):
                  [,1]       [,2]       [,3]       [,4]       [,5]       [,6]
total_carb  0.26799392  1.15006992  0.5754769 -0.20373776  0.7006857 -0.9896277
fiber      -0.10563505  0.02898702  0.6670834  0.10210220 -0.5639028  1.1822695
sugar      -0.16228013 -0.20796108 -0.3924982  0.04620889 -1.3340243  0.1606226
protein    -1.08127266 -0.16870569  0.1030195  0.20928707  0.1811481  0.4209405
vit_a       0.07402743  0.06535543 -0.4293598  0.45443724  0.2041258  0.7523923
vit_c       0.05035821 -0.09206800  0.3036891 -1.13390967  0.5625970 -0.5784557
calcium     0.19590873 -0.04325816 -1.1396139 -0.28474642  0.2303286 -0.2439576
```

The canonical coefficients show how each original variable contributes to the canonical variates:

- **First canonical variate (Set 1 - X):**
  The most influential variables are *cholesterol* (large negative coefficient), *saturated fat*, and *protein* (positive), suggesting that items high in cholesterol and saturated fat are strongly linked to the canonical dimension.

- **First canonical variate (Set 2 - Y):**
  Total carbohydrates and *calcium* have positive coefficients, while *protein* has a large negative coefficient, indicating that the canonical dimension contrasts carbohydrate/calcium-rich items with those high in protein.

Visualizations



- **Canonical Structure Plot (left panel):**
  This plot shows the relationships between canonical variates. The clustering of points within the inner circles indicates that most variables are well represented by the first two canonical dimensions, and the spread suggests meaningful differentiation.
- **Canonical Scores Plot (right panel):**
  Each menu item is plotted by its canonical scores on the first two dimensions. The dispersion of points shows how items are distributed across the canonical space, with some clear outliers indicating unique nutritional profiles.

Interpretation

- **Multivariate Nutritional Patterns:**
  The first canonical correlation (0.96) reveals a very strong relationship between a combination of high cholesterol, saturated fat, and sodium (macronutrient set) and a combination of high total carbohydrates and calcium but low protein (micronutrient set). In practical terms, this suggests that fast food items high in fats and cholesterol tend to be lower in certain micronutrients and protein, and vice versa.
- **Dimensionality:**
  The first two canonical dimensions account for the majority of the shared variance between the two sets, allowing for a simplified but comprehensive view of the complex relationships in the data.
- **Implications:**
  - **For consumers:** Items high in calories, fat, and sodium are likely to be low in protein and certain micronutrients.
  - **For industry:** Reformulation efforts aiming to improve nutritional profiles should consider these multivariate relationships, as improving one aspect (e.g., reducing fat) may impact others (e.g., protein or micronutrient content).

In summary, CCA uncovers strong, statistically significant multivariate associations between macronutrient and micronutrient profiles in fast food items. The analysis highlights that energy-dense, high-fat items are typically inversely related to protein and micronutrient content, providing actionable insights for both public health guidance and menu development.

4.Conclusion and Recommendation

Summary of Main Findings

This multivariate analysis of the Fast-Food Nutrition dataset reveals several important patterns regarding the nutritional quality of fast-food menu items:

- **High Energy Density and Poor Nutritional Balance:**
  Most fast-food items are energy-dense, with high levels of calories, total fat, saturated fat, and sodium. These findings are consistent with existing research showing that typical fast-food meals often provide a large proportion of the recommended daily intake for these nutrients in just one serving.
- **Low Micronutrient Content:**
  While fast food items are rich in calories and fats, they are generally low in essential vitamins and minerals such as vitamin A and calcium, contributing to unbalanced nutrition and potential deficiencies if consumed frequently.
- **Distinct Nutritional Patterns Identified:**
  Principal Component Analysis and Factor Analysis revealed that most of the variation in menu items can be explained by a few underlying patterns: an "energy-dense/high-fat" dimension, a "carbohydrate/fiber" dimension, and a "micronutrient/sugar" dimension.
- **Clear Separation Between Healthy and Unhealthy Items:**
  Discriminant analysis showed that calories, total fat, and sodium are the primary variables distinguishing "healthy" from "unhealthy" fast food items, and most menu items are easily classified as unhealthy.
- **Strong Multivariate Relationships:**
  Canonical correlation analysis demonstrated a strong inverse relationship between energy-dense macronutrient profiles and micronutrient content, highlighting the nutritional trade-offs in fast food.

Limitations

- **Data Scope:**
  The dataset may not capture all possible menu items, seasonal offerings, or recent reformulations, and some micronutrient data were missing or incomplete.
- **Assumption Violations:**
  Some statistical methods (e.g., discriminant analysis) assume normality and homogeneity of variance, which were not fully met in this dataset.
- **Generalizability:**
  The findings are based on menu data from major chains and may not represent smaller establishments or international variations.

Recommendations

- **For Consumers:**
  - Limit the frequency of fast-food consumption, especially items high in calories, fat, and sodium.
  - Use available nutritional information to select menu items that are lower in energy density and higher in fiber and micronutrients (e.g., salads with minimal dressing, grilled options, fruit sides).
  - Be cautious with menu items marketed as "healthy," as they may still be high in sodium or calories due to added sauces or dressings.

- **For Industry:**
    - Reformulate menu items to reduce calories, saturated fat, and sodium, and increase the availability of options rich in fiber, vitamins, and minerals.
    - Provide clear, accessible nutritional labeling for all menu items to empower informed choices.
- **For Policy Makers:**
    - Encourage or mandate transparent menu labeling and set nutritional standards for meals, especially those marketed to children.

In conclusion, while fast food offers convenience and affordability, its typical nutritional profile poses significant health risks if consumed frequently. Both individual choices and systemic changes—such as reformulation and better labeling—are needed to improve the nutritional quality of fast food and support public health.

5. References

- Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: A Package for Multivariate Analysis. *Journal of Statistical Software, 25*(1), 1–18.
- Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth edition). Springer.
- González, I., Déjean, S., Martin, P. G. P., & Baccini, A. (2008). CCA: An R Package to Extend Canonical Correlation Analysis. *Journal of Statistical Software, 23*(12), 1–14.
- Makowski, D. (2018). The psycho package: An efficient and publishing-oriented workflow for psychological science. *Journal of Open-Source Software, 3*(22), 470.
- Dhamnetiya, D., Goel, M. K., Jha, R. P., Shalini, S., & Bhattacharyya, K. (2022). How to Perform Discriminant Analysis in Medical Research? Explained with Illustrations. *Cureus, 14*(6), e25759.

6. Appendices

Dataset: https://www.kaggle.com/datasets/ulrikthygepedersen/fastfood-nutrition

R Code:

Coding Part for Mini-Project

```
# Installing Packages
install.packages(c("readr", "dplyr", "FactoMineR", "factoextra", "MVar", "fastDummies", "ggplot2",
"corrplot","schoolmath","CCP","CCA","psych","biotools"))
```

```
# Load libraries
library(readr)
library(dplyr)
library(FactoMineR)
library(factoextra)
library(MVar)
library(fastDummies)
library(ggplot2)
library(corrplot)
library(schoolmath)
library(CCA)
library(CCP)
library(MASS)
library(biotools)
library(psych)

# Load dataset
data <- read_csv("fastfood.csv")
```

```
# Inspect the data
print(head(data))
#  structure of fastfood data
print(str(data))
print(summary(data))
```

Data Preparation

```
# remove duplicates
data <- data[!duplicated(data), ]

# Removes rows with any NA values
data <- na.omit(data)
numeric_data <- data %>%
  dplyr::select(where(is.numeric))

# removing Outliers
remove_outliers <- function(df) {
  df[] <- lapply(df, function(x) {
    if (is.numeric(x)) {
      Q1 <- quantile(x, 0.25, na.rm = TRUE)
      Q3 <- quantile(x, 0.75, na.rm = TRUE)
      IQR <- Q3 - Q1
      lower <- Q1 - 1.5 * IQR
      upper <- Q3 + 1.5 * IQR
      x[x < lower | x > upper] <- NA
    }
    x
  })
  df <- na.omit(df)
  return(df)
}

numeric_data <- remove_outliers(numeric_data)

# standardize data
data_scaled <- as.data.frame(scale(numeric_data))
colnames(data_scaled) <- colnames(numeric_data)
```

Principal Component Analysis (PCA)

```
pca_result <- prcomp(data_scaled, center = FALSE, scale. = FALSE)

# View summary of PCA results
summary(pca_result)
fviz_pca_var(pca_result,
col.var = "contrib", # Color by contributions to the PC
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE)
```

```
# Calculate explained variance
explained_var <- pca_result$sdev^2 / sum(pca_result$sdev^2)
explained_var_percent <- explained_var * 100

# Scree plot using ggplot2
scree_df <- data.frame(PC = 1:length(explained_var_percent),ExplainedVariance = explained_var_percent)

ggplot(scree_df, aes(x = PC, y = ExplainedVariance)) +geom_bar(stat = "identity", fill = "skyblue", alpha = 0.7)
+geom line(color = "red", size = 1) +geom point(color = "red", size = 2) +labs(
    title = "Scree Plot: Explained Variance by PCA Components",
    x = "Principal Component",
    y = "Explained Variance (%)"
  ) +
  theme_minimal()

# Print the percentage for each PC (all PCs)
for (i in seq_along(explained_var_percent)) {
  cat("PC", i, "explains", round(explained_var_percent[i], 2), "% of the variance\n")
}

# Print total explained variance by first two components
total_var_2 <- sum(explained_var_percent[1:2])
cat("Total Explained Variance by First 2 Components:", round(total_var_2, 1), "%\n\n")

# Print loadings for the first and second principal components
cat("PCA Component Loadings (PC1 and PC2):\n")
print(round(pca_result$rotation[, 1:2], 6))
```

Factor Analysis (FA)

```
# Check data suitability
KMO(numeric_data)
cortest.bartlett(numeric_data)
# Determine Number of Factors
# Parallel analysis and scree plot
fa.parallel(numeric_data, fa="fa")
scree(numeric_data, factors=FALSE)
# Factor Extraction
fa_results <- psych::fa(numeric_data, nfactors=3, rotate="varimax", fm="minres", scores="Anderson")
# Interpret Results
# Loadings table (|loading| >0.3 considered meaningful)
print(fa_results$loadings, cutoff=0.3)

# Variance explained
fa_results$Vaccounted  # Shows cumulative variance [2]
# Visualization
# Install visualization tools
devtools::install_github('mattkcole/FAtools')
library(FAtools)



# Loadings plot
loadings_plot(fa_results$loadings)

# Factor diagram
fa.diagram(fa_results)
# Rotation Methods
# Compare rotations
fa(numeric_data, nfactors=3, rotate="oblimin")$loadings
fa(numeric_data, nfactors=3, rotate="varimax")$loadings
```

Discriminant Analysis (DA)

```
#Create the Group Variable
numeric_data$group <- ifelse(numeric_data$calories < 400 & numeric_data$total_fat < 15 & numeric_data$sodium < 700,
"Healthy", "Unhealthy")
numeric_data$group <- as.factor(numeric_data$group)

# Check Assumptions for LD
by(numeric_data$total_fat, numeric_data$group, shapiro.test)

boxM(numeric_data[,c("calories","cal_fat" ,"total_fat"  ,"sat_fat", "trans_fat",  "cholesterol" ,"sodium"
,"total_carb",  "fiber",  "sugar" ,"protein"  ,"vit_a"  ,"vit_c"  ,"calcium" )], numeric_data$group)
cor(numeric_data[,c("calories","cal_fat"  ,"total_fat"  ,"sat_fat", "trans_fat",  "cholesterol" ,"sodium"
,"total_carb",  "fiber",  "sugar" ,"protein"  ,"vit_a"  ,"vit_c"  ,"calcium" )])
# Perform Discriminant Analysis
qda_model <- qda(group ~ calories + cal_fat + total_fat + sat_fat + trans_fat + cholesterol + sodium + total_carb +
fiber + sugar + protein + vit_a + vit_c + calcium , data = numeric_data)

pred <- predict(qda_model)
table(Predicted = pred$class, Actual = numeric_data$group)
```

Canonical Correlation Analysis (CCA)

```
X <- numeric_data[, c("calories", "total_fat", "sat_fat", "trans_fat", "cholesterol", "sodium")]
Y <- numeric_data[, c("total_carb", "fiber", "sugar", "protein", "vit_a", "vit_c", "calcium")]
X <- scale(X)
Y <- scale(Y)
cca_result <- cc(X, Y)

plt.cc(cca_result)

p.asym(cca_result$cor, nrow(X), ncol(X), ncol(Y))

# Canonical correlations
cat("Canonical correlations:\n")
print(cca_result$cor)

# Canonical coefficients
cat("\nCanonical coefficients for X (Set 1):\n")
print(cca_result$xcoef)

cat("\nCanonical coefficients for Y (Set 2):\n")
print(cca_result$ycoef)
```