ИТМО

# Laboratory 2.


**Authors:**
Jorge Sosa
Elizabetty Guseva
Siraeva Gulnara

**Professors:**
Anna V. Kalyuzhnaya, PhD
Irina Deeva, PhD


**Course:**
Methods and Models for Multivariate Data Analysis

December 11, 2024

# Contents

# 1 Introduction

This analysis focuses on the McDonald's menu dataset, which includes nutritional information such as calories, fats, proteins, carbohydrates, and vitamins. The goal is to explore complex relationships among variables, improve data quality using multivariate techniques, and enhance model performance through feature engineering and synthetic data generation.

## 1.1 Objectives:

- Analyze relationships between nutritional components.
- Identify key patterns and dependencies in menu item compositions.
- Improve data quality through advanced multivariate techniques.
- Generate synthetic data while preserving nutritional relationships.
- Refit models using transformed and optimized data.

# 2 Initial Multivariate Analysis

## 2.1 Correlation Matrix

The correlation matrix identifies linear relationships between independent variables and between independent and dependent variables. Ultimately, the matrix with imputed data was selected for further analysis.

Using imputed data allows for robust and effective multivariate analysis, especially when:

- Missing values are significant and cannot be ignored.
- It is important to maintain the full sample size.
- Biases or information loss in modeling and analysis need to be avoided.

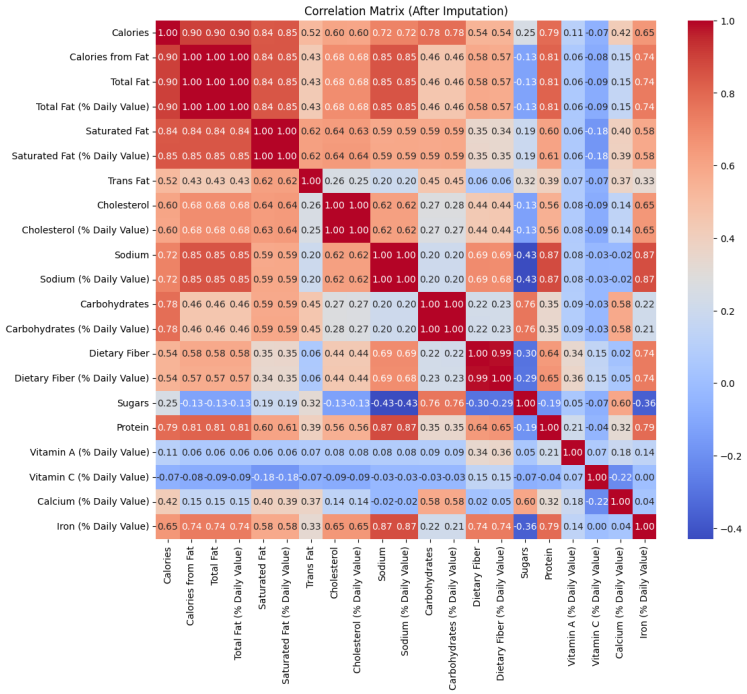


**Figure 1:** *Correlation Matrix*

Key findings:

- Strong correlations were observed between fats, calories, and sodium.
- Weak relationships between vitamins and macronutrients suggest independent contributions.

## 2.2 Multicollinearity Detection

Variance Inflation Factor (VIF) analysis detected severe multicollinearity among predictors in the initial dataset:

- **Severe Multicollinearity (VIF $>$ 1000):**
  - Sodium and its Daily Value percentage (VIF  7900).
  - Cholesterol and its Daily Value percentage (VIF  3770).
  - Total Fat metrics (VIF 2500-3300).
  - Calories and Calories from Fat (VIF 2200-2750).
  - Carbohydrates and its Daily Value percentage (VIF 1400-2100).
- **High Multicollinearity (VIF $>$ 100):**
  - Saturated Fat and its Daily Value percentage (VIF  780).
  - Protein (VIF  118).
- **Moderate Multicollinearity (VIF $>$ 10):**
  - Dietary Fiber and its Daily Value percentage (VIF  45).
  - Sugars (VIF  30).
- **Acceptable Levels (VIF $<$ 10):**
  - Trans Fat.
  - Vitamins and Minerals (Iron, Calcium, Vitamin A, Vitamin C).

After removing variables with the highest VIF values, the multicollinearity situation improved significantly, although some concerns remained:

- **High VIF ($>$ 20):**
  - Sugars (26.71).
  - Sodium (24.86).
  - Carbohydrates (23.07).
  - Total Fat (21.33).
- **Moderate VIF (10-20):**
  - Protein (17.89).
  - Saturated Fat (12.24).
- **Borderline VIF (5-10):**

- Iron ( Daily Value) (8.41).

- **Good VIF ($< 5$):**
  - Calcium ( Daily Value) (4.86).
  - Dietary Fiber (3.91).
  - Trans Fat (2.91).
  - Cholesterol (2.38).
  - Vitamin A ( Daily Value) (1.39).
  - Vitamin C ( Daily Value) (1.21).

While the situation has improved dramatically from the initial VIFs in the thousands, moderate multicollinearity remains among macronutrients (Sugars, Carbohydrates, Total Fat, Protein). This is expected, as these nutrients are often correlated in food items.

# 3 Advanced Multivariate Methods

## 3.1 Principal Component Analysis (PCA)

PCA reduced the dimensionality of the dataset, capturing 95% of the variance with the first eight components:

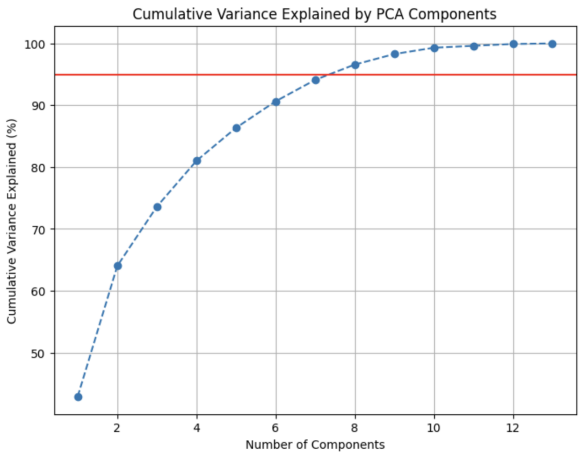


**Figure 2:** *PCA*

- **PC1:** Macronutrients (Total Fat, Protein, Sodium).
- **PC2:** Carbohydrates and Sugars (Sugars, Calcium, Carbohydrates).
- **PC3:** Vitamins and Fiber (Vitamin A, Vitamin C, Dietary Fiber).

- **PC4:** Vitamin-Mineral Relationship (Vitamin C, Vitamin A, Calcium).
- **PC5:** Fat Quality Indicator (Trans Fat, Carbohydrates, Vitamin A).
- **PC6:** Protein-Mineral Relationship (Cholesterol, Protein, Calcium).
- **PC7:** Mineral-Macronutrient Interplay (Calcium, Cholesterol, Carbohydrates).
- **PC8:** Fiber-Mineral Focus (Dietary Fiber, Iron, Total Fat).

## 3.2 Nutritional patterns identified:

- **Fundamental Macronutrient Structure (PC1):**
  - Represents overall food density/portion size.
  - Groups Total Fat, Protein, and Sodium together.
  - Likely captures protein-rich, savory foods.
- **Carbohydrate-Mineral Pattern (PC2):**
  - Links Sugars, Carbohydrates, and Calcium.
  - Possibly represents dairy and sweet foods.
  - Highlights relationships between calcium-rich foods and carbohydrates.
- **Vitamin-Fiber Complex (PC3):**
  - Groups Vitamins A, C, and Dietary Fiber.
  - Likely represents fresh or plant-based foods.
  - Captures nutritional density of produce.
- **Micronutrient Interrelations (PC4):**
  - Correlation between vitamins and calcium.
  - Suggests foods rich in multiple micronutrients.
  - Likely represents fortified foods.
- **Fat Quality Indicator (PC5):**
  - Led by Trans Fat.
  - Shows an interesting connection to carbohydrates and Vitamin A.

- May represent processed foods.

- **Refinement Components (PC6-8):**
  - Show more specific patterns.
  - Focus on particular nutrient relationships, such as fiber-iron and cholesterol-calcium interactions.

## 3.3 t-SNE and UMAP

### 3.3.1 UMAP

This technique visualized high-dimensional data, revealing clusters based on nutritional characteristics. However, it is primarily used for visualization rather than feature reduction.

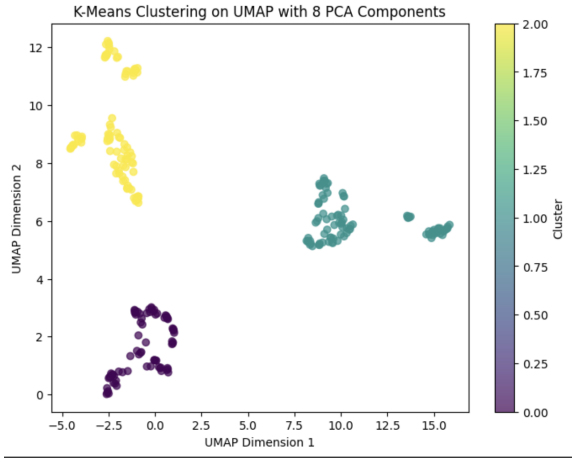


**Figure 3:** *UMAP*

Silhouette Score (UMAP): 0.73

#### Interpreting Clustering with UMAP

The graph shows three well-separated clusters:

- **Yellow cluster (top left):**
  - Located in the negative UMAP Dimension 1 region.
  - Has high values on UMAP Dimension 2.
  - Compact group of points with some spread.

- **Purple cluster (bottom left):**
  - Also in the negative UMAP Dimension 1 region.

  - Low values on UMAP Dimension 2.
  - More sparse structure.

- **Turquoise cluster (right):**
  - Positive values on UMAP Dimension 1.
  - Medium values on UMAP Dimension 2.
  - Contains several point subgroups.

The clusters are well separated spatially, indicating:

- Effectiveness of the chosen number of clusters (k=3).
- Successful application of UMAP for visualization.
- Presence of natural structure in the nutritional data.

Silhouette coefficient of 0.73:

- **Value close to 1 (maximum is 1):**
  - Shows that clusters are well separated.
  - Objects within clusters are similar to each other.
  - Objects from different clusters are substantially different.
- **This is confirmed by visualization:**
  - Clusters are clearly separated in space.
  - Minimal overlap between clusters.
  - Points within clusters are compactly arranged.
- **Practical significance:**
  - 0.7 is considered a high clustering quality indicator.
  - Confirms that choosing 3 clusters was justified.
  - Indicates the presence of natural structure in the nutritional data.

Such a high silhouette coefficient suggests that the clustering successfully identified different groups of products based on their nutritional characteristics.

### 3.3.2 t-SNE

This technique visualized high-dimensional data, revealing clusters based on nutritional characteristics. However, it is primarily used for visualization rather than feature reduction.

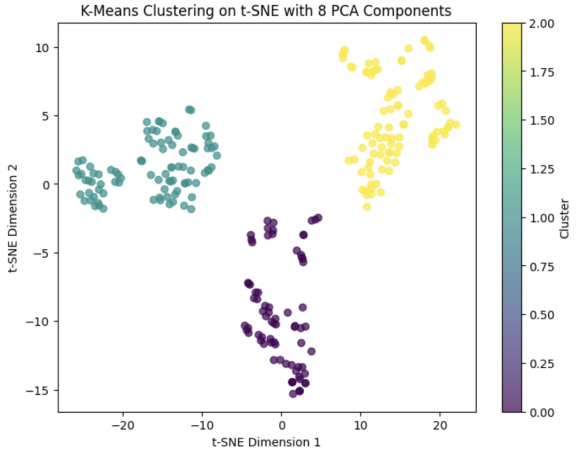


**Figure 4:** *t-SNE*

Silhouette Score (t-SNE): 0.67

**Interpreting Clustering with t-SNE**

1. **Cluster Structure:**
    - **Yellow cluster (right):** Large cluster with spread, possibly representing high-calorie/nutrient-dense foods.
    - **Purple cluster (center):** Vertically elongated cluster, showing internal structure with distinct subgroups.
    - **Teal cluster (left):** Two distinct subgroups, potentially representing foods with similar nutritional profiles.
2. **Characteristics of the Visualization:**
    - Clear separation between clusters.
    - Distinct subgroups within clusters.
    - Non-linear separation boundaries.
    - Good preservation of local structure.
3. **Comparing with UMAP:**
    - t-SNE shows more internal structure.

- Clusters are more spread out.
- Reveals potential subgroups not as visible in UMAP.
- Different spatial arrangement but similar overall grouping.

**Clustering Quality**

- Clusters are well-defined.
- Minimal overlap between clusters.
- Natural-looking separations.
- Clear boundaries between different groups.

The silhouette coefficient of 0.67 for t-SNE is slightly lower than UMAP's 0.73 but still indicates good clustering quality. A comparison:

**UMAP vs t-SNE**

1. **Silhouette Coefficients:**
   - UMAP: 0.73.
   - t-SNE: 0.67.
   - Both values $>0.6$ indicate good cluster separation.
2. **Visual Differences:**
   - t-SNE shows more detailed internal cluster structure.
   - UMAP produces more compact, clearly separated groups.
   - t-SNE reveals potential subgroups within clusters.
3. **Advantages of Each Method:**
   - UMAP better preserves global data structure.
   - t-SNE better shows local relationships.
   - UMAP provides slightly clearer cluster separation.

Both methods confirm the presence of three main groups in the data but show their structure slightly differently. UMAP's higher silhouette coefficient suggests it might be preferable for this dataset.

## 3.4 Factor Analysis

Latent factors were identified:

### Factor Loadings on the 8 Principal Components

**Table 1:** *Factor Loadings for Principal Components*

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 |
|---|---|---|---|---|---|---|---|---|
| Factor 0 | 2.1399 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Factor 1 | -0.0000 | 1.3202 | 0.0000 | -0.0000 | -0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Factor 2 | -0.0000 | 0.0000 | 0.4977 | -0.0000 | -0.0000 | 0.0000 | 0.0000 | 0.0000 |

### Interpreting Factor Loadings

1. **Factor 0 (first row):**
   - Strong correlation with PC1 (2.139928).
   - Almost zero correlations with other PCs.
   - This indicates that the first factor mainly reflects information from PC1 (nutritional density).
2. **Factor 1 (second row):**
   - Strong correlation with PC2 (1.320153).
   - Very small values for other PCs.
   - Corresponds to information from PC2 (sweet/carbohydrate products).
3. **Factor 2 (third row):**
   - Moderate correlation with PC3 (0.497699).
   - Negligible correlations with other PCs.
   - Reflects information from PC3 (vitamin-fiber characteristics).

### Characteristics of the Results

1. Clear factor structure.
2. Each factor is predominantly associated with one PC.
3. Very small cross-loadings (close to zero).

This confirms that the first three PCs indeed capture the main patterns in the nutritional data.

### Cluster Averages for UMAP and t-SNE

**Table 2:** *Average Factors per Cluster (UMAP and t-SNE)*

| Cluster | Factor 1 | Factor 2 |
|---|---|---|
| Cluster 0 | -0.8809 | -0.3186 |
| Cluster 1 | -0.1362 | 0.8139 |
| Cluster 2 | 0.8558 | -0.5630 |

### Cluster Interpretations

**Table 3:** *Cluster Characteristics by Factors*

| Cluster | Factor 1 | Factor 2 | Interpretation |
|---|---|---|---|
| Cluster 0 | -0.8809 | -0.3186 | Represents products with low nutritional density and moderate carbohydrate content. |
| Cluster 1 | -0.1362 | 0.8139 | Represents products high in carbohydrates/-sugars, such as desserts and sweet beverages. |
| Cluster 2 | 0.8558 | -0.5630 | Represents products with high nutritional value, rich in proteins and fats. |

### Distribution of Products by Factor 1 and Factor 2 (UMAP Cluster)

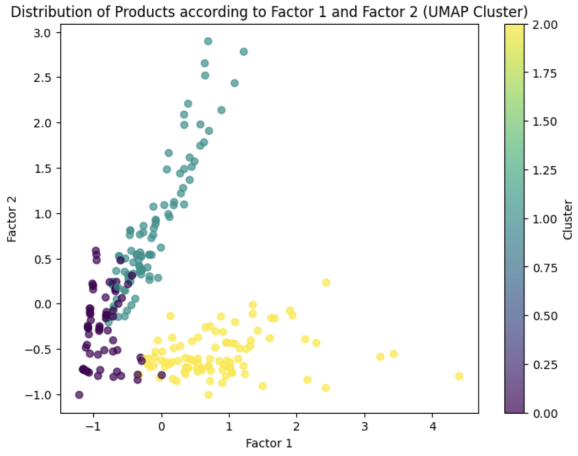


**Figure 5:** *Distribution of Products*

**Yellow Cluster (right/bottom):**

- High Factor 1 values (0.5 to 4.0).
- Negative Factor 2 values (-1.0 to 0).
- Likely represents high-calorie, protein-rich items like burgers and meat dishes.

**Teal Cluster (middle/top):**

- Moderate Factor 1 values (-0.5 to 1.0).
- High Factor 2 values (0.5 to 3.0).
- Represents sweet items and carbohydrate-rich products like desserts and sugary drinks.

**Purple Cluster (left/bottom):**

- Low Factor 1 values (-1.0 to 0).
- Low Factor 2 values (-1.0 to 0.5).
- Likely represents lighter menu items like salads and low-calorie options.

Key Observations:

- UMAP and t-SNE produced identical mean factor values for clusters, demonstrating the stability and robustness of clustering.
- The distribution of factors shows clear separation between clusters, highlighting distinct food categories.
- Cluster interpretations align with nutritional patterns observed in the data.

# 4 Additional Feature Engineering Techniques

## 4.1 Feature Interaction Generation

Interactions between features were created to capture non-linear dependencies. These interactions improved predictive accuracy for calorie estimation.

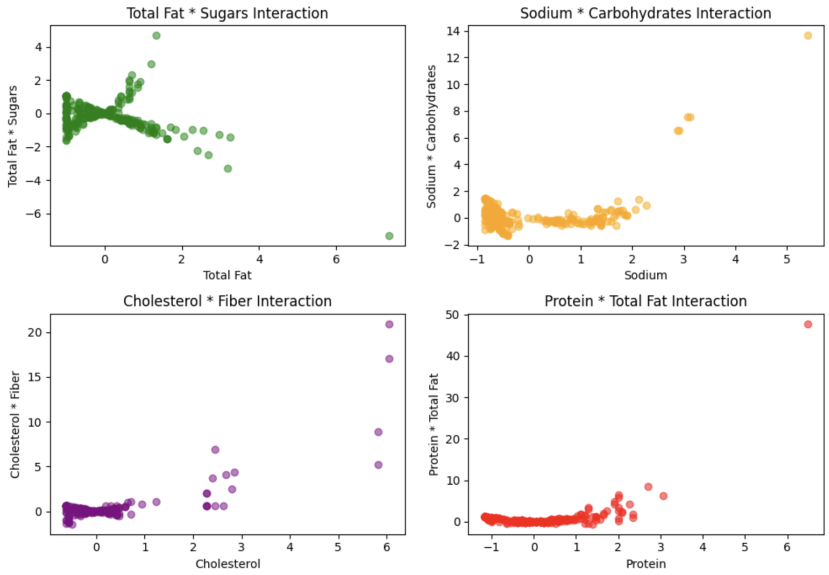


**Figure 6:** *Interaction Generation*

- **Top Left: Total Fat * Sugars Interaction**
  - The interaction between Total Fat and Sugars shows a relatively broad dispersion.
  - Some points lie outside the main range, indicating foods combining high levels of fat and sugar, such as desserts or processed items.
- **Top Right: Sodium * Carbohydrates Interaction**
  - A slight positive relationship between Sodium and Carbohydrates is observed, with most points concentrated near the origin.
  - This may reflect salty foods that are also rich in carbohydrates, such as bread or snacks.
- **Bottom Left: Cholesterol * Fiber Interaction**

- The interaction between Cholesterol and Fiber shows a scattered trend with a few higher values in specific foods.
- This could represent cholesterol-rich foods, such as meats, that also have low fiber levels.

- **Bottom Right: Protein * Total Fat Interaction**
  - This graph highlights the interaction between Protein and Total Fat.
  - A positive trend indicates that higher levels of Protein are often paired with higher levels of Total Fat, common in protein-dense foods such as meats and dairy.

## 4.2 Polynomial Features

Polynomial transformations up to the second degree modeled non-linear relationships, such as quadratic terms for saturated fats and sodium.

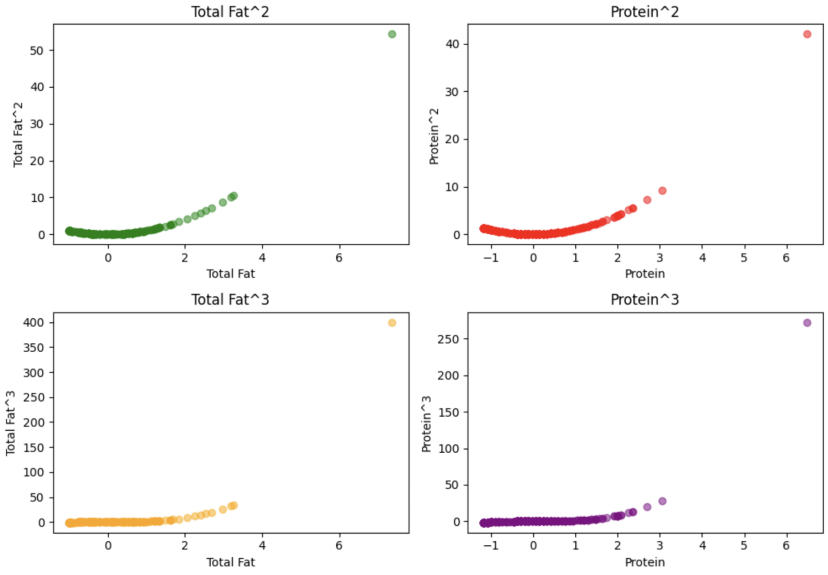


**Figure 7:** *Polynomial*

- **Top Left: Total Fat(2) vs Total Fat**
  - This graph shows a quadratic relationship between Total Fat and its squared version.
  - The smoothly increasing trend indicates that higher levels of Total Fat contribute exponentially to the quadratic sum.

- **Top Right: Protein(2) vs Protein**

  - A similar quadratic trend is observed for Protein.
  - Higher values of Protein show an exponential impact, capturing significant nonlinear relationships.

- **Bottom Left: Total Fat(3) vs Total Fat**
  - This graph represents a cubic relationship between Total Fat and its cubed version.
  - The growth is steeper compared to the quadratic relationship, highlighting stronger nonlinear interactions at higher fat levels.

- **Bottom Right: Protein(3) vs Protein**
  - Similar to the cubic Total Fat graph but for Protein.
  - Higher Protein values exhibit a sharper growth when cubed, suggesting potential nonlinear interactions in complex models.

## 4.3 Feature Selection Techniques

### RFE Results

RFE selected the 5 most important features for calorie prediction:

1. **Primary Nutrients:**
    - Total Fat.
    - Carbohydrates.
    - Protein.
2. **Third-degree Polynomial Features:**
    - Total Fat_poly3 (fat cubed).
    - Protein_poly3 (protein cubed).

**Rationale:**

1. The three main macronutrients directly affect calorie content.
2. Third-degree polynomial features help capture nonlinear relationships.
3. Interestingly, cubic terms were chosen rather than quadratic ones.

**Additional Observations:**

- Simple feature interactions were not selected.
- Micronutrients (vitamins, minerals) were not selected.
- Among polynomial terms, only those for fats and proteins were chosen.

### LASSO Results

LASSO selected many more features than RFE. These features are categorized as follows:

1. **Primary Nutrients:**
    - Total Fat, Saturated Fat, Trans Fat.
    - Carbohydrates, Sugars.
    - Protein.
    - Cholesterol.
    - Dietary Fiber.

2. **Vitamins and Minerals:**
    - Vitamin A (% Daily Value).
    - Vitamin C (% Daily Value).
    - Iron (% Daily Value).
3. **First-order Interactions:**
    - fat_sugars_interaction.
    - sodium_carbs_interaction.
    - cholesterol_fiber_interaction.
4. **Polynomial Features:**
    - Quadratic ($^2$): Total Fat, Sodium, Protein.
    - Cubic ($^3$): Total Fat, Sodium, Protein.
5. **Complex Interactions:**
    - Total Fat * Sodium, Total Fat * Protein.
    - Total Fat$^2$ * Sodium, Total Fat$^2$ * Protein.
    - Total Fat * Sodium$^2$, Total Fat * Sodium * Protein.
    - Total Fat * Protein$^2$.
    - Sodium$^2$ * Protein, Sodium * Protein$^2$.

**Interpretation:**

1. Calorie content has complex nonlinear dependencies.
2. Both primary nutrients and their interactions are important.
3. Some micronutrients also play a role.
4. Cubic terms may be important for accurate prediction.

# 5 Bayesian Networks

## 5.1 Structure Learning

A Bayesian network was constructed using the Hill-Climbing algorithm. Key dependencies included:


**Figure 8:** *Structure Learning*

### Interpretations

The results of learning the Bayesian network structure show interesting relationships between various nutrients and food caloric content. Let's analyze some key observations:

1. **Impact on Caloric Content:**
    - Calories are directly linked to carbohydrates, trans fats, and saturated fats.
    - Total fat also influences caloric content. This makes sense as these components are major sources of energy in food.
2. **Fat Relationships:**
    - Total fat is connected to saturated fats and calories.
    - Cholesterol is linked to total fat, saturated fats, and calories, showing the close relationship between different types of fats in foods.
3. **Role of Carbohydrates:**
    - Carbohydrates are connected to calories and sugars.
    - Sugars directly influence caloric content, reflecting the importance of carbohydrates as an energy source.

4. **Protein:**

    - Protein is linked to cholesterol, sugars, and total fat.
    - Interestingly, protein doesn't show a direct connection to calories in this model.

5. **Micronutrients:**

    - Vitamin A is connected to total fat, cholesterol, and sugars.
    - Vitamin C is linked to protein.
    - Iron is connected to fiber, protein, total fat, and cholesterol, indicating that foods rich in certain nutrients often contain specific micronutrients.

6. **Fiber:**

    - Fiber is connected to protein, vitamin C, and cholesterol, reflecting characteristics of fiber-rich foods.

This model provides interesting insights into the relationships between various nutrients in foods.

## 5.2 Conditional Independence

**Markov Blankets**

- **Markov Blanket for Calories:** Saturated Fat, Sugars, Cholesterol, Trans Fat, Total Fat, Carbohydrates.
- **Markov Blanket for Total Fat:** Saturated Fat, Sugars, Cholesterol, Iron (% Daily Value), Carbohydrates, Calories, Protein, Vitamin A (% Daily Value).
- **Markov Blanket for Protein:** Sugars, Cholesterol, Dietary Fiber, Vitamin C (% Daily Value), Iron (% Daily Value), Total Fat, Vitamin A (% Daily Value).

## Conditional Probability Distributions (CPDs)

### Conditional Probability Distribution for Calories:

Conditional Probability Distribution for Calories:

| Cholesterol | Cholesterol(0.0) | ... | Cholesterol(4.0) | Cholesterol(4.0) |
|---|---|---|---|---|
| Sugars | Sugars(0.0) | ... | Sugars(4.0) | Sugars(4.0) |
| Total Fat | Total Fat(0.0) | ... | Total Fat(3.0) | Total Fat(4.0) |
| Calories(0.0) | 1.0 | ... | 0.0 | 0.0 |
| Calories(1.0) | 0.0 | ... | 0.0 | 0.0 |
| Calories(2.0) | 0.0 | ... | 0.0 | 0.0 |
| Calories(3.0) | 0.0 | ... | 0.0 | 0.0 |
| Calories(4.0) | 0.0 | ... | 1.0 | 1.0 |

**Figure 9:** *Calories*

### Conditional Probability Distribution for Carbohydrates:

Conditional Probability Distribution for Carbohydrates:

| Calories | Calories(0.0) | ... | Calories(4.0) | Calories(4.0) |
|---|---|---|---|---|
| Sugars | Sugars(0.0) | ... | Sugars(4.0) | Sugars(4.0) |
| Total Fat | Total Fat(0.0) | ... | Total Fat(3.0) | Total Fat(4.0) |
| Carbohydrates(0.0) | 1.0 | ... | 0.0 | 0.0 |
| Carbohydrates(1.0) | 0.0 | ... | 0.0 | 0.0 |
| Carbohydrates(2.0) | 0.0 | ... | 0.0 | 0.0 |
| Carbohydrates(3.0) | 0.0 | ... | 0.0 | 0.0 |
| Carbohydrates(4.0) | 0.0 | ... | 1.0 | 1.0 |

**Figure 10:** *Carbohydrates*

### Conditional Probability Distribution for Trans Fat:

Conditional Probability Distribution for Trans Fat:

| Calories | Calories(0.0) | ... | Calories(3.0) | Calories(4.0) |
|---|---|---|---|---|
| Trans Fat(0.0) | 1.0 | ... | 0.75 | 0.2962962962962963 |
| Trans Fat(1.0) | 0.0 | ... | 0.25 | 0.7037037037037037 |

**Figure 11:** *Trans Fat*

### Conditional Probability Distribution for Saturated Fat:

Conditional Probability Distribution for Saturated Fat:

| Calories | Calories(0.0) | ... | Calories(4.0) | Calories(4.0) |
|---|---|---|---|---|
| Cholesterol | Cholesterol(0.0) | ... | Cholesterol(4.0) | Cholesterol(4.0) |
| Total Fat | Total Fat(0.0) | ... | Total Fat(3.0) | Total Fat(4.0) |
| Saturated Fat(0.0) | 1.0 | ... | 0.0 | 0.0 |
| Saturated Fat(1.0) | 0.0 | ... | 0.0 | 0.0 |
| Saturated Fat(2.0) | 0.0 | ... | 0.0 | 0.1111111111111111 |
| Saturated Fat(3.0) | 0.0 | ... | 1.0 | 0.8888888888888888 |

**Figure 12:** *Saturated Fat*

## Conditional Probability Distribution for Total Fat:

Conditional Probability Distribution for Total Fat:

| Cholesterol | ... | Cholesterol(4.0) |
|---|---|---|
| Iron (% Daily Value) | ... | Iron (% Daily Value)(3.0) |
| Protein | ... | Protein(4.0) |
| Total Fat(0.0) | ... | 0.0 |
| Total Fat(1.0) | ... | 0.0 |
| Total Fat(2.0) | ... | 0.03333333333333333 |
| Total Fat(3.0) | ... | 0.3333333333333333 |
| Total Fat(4.0) | ... | 0.6333333333333333 |

**Figure 13:** *Total Fat*

**Conditional Probability Distribution for Vitamin A ( Daily Value):**

Conditional Probability Distribution for Vitamin A (% Daily Value):

| Cholesterol | ... | Cholesterol(4.0) | Cholesterol(4.0) |
|---|---|---|---|
| Protein | ... | Protein(4.0) | Protein(4.0) |
| Total Fat | ... | Total Fat(3.0) | Total Fat(4.0) |
| Vitamin A (% Daily Value)(0.0) | ... | 0.0 | 0.1 |
| Vitamin A (% Daily Value)(1.0) | ... | 0.5 | 0.35 |
| Vitamin A (% Daily Value)(2.0) | ... | 0.1 | 0.2 |
| Vitamin A (% Daily Value)(3.0) | ... | 0.4 | 0.35 |

**Figure 14:** *Vitamin A*

**Conditional Probability Distribution for Protein:**

Conditional Probability Distribution for Protein:

| Dietary Fiber | ... | Dietary Fiber(3.0) |
|---|---|---|
| Iron (% Daily Value) | ... | Iron (% Daily Value)(3.0) |
| Vitamin C (% Daily Value) | ... | Vitamin C (% Daily Value)(1.0) |
| Protein(0.0) | ... | 0.0 |
| Protein(1.0) | ... | 0.0 |
| Protein(2.0) | ... | 0.0 |
| Protein(3.0) | ... | 0.04 |
| Protein(4.0) | ... | 0.96 |

**Figure 15:** *Protein*

**Conditional Probability Distribution for Cholesterol:**

Conditional Probability Distribution for Cholesterol:

| Dietary Fiber | ... | Dietary Fiber(3.0) |
|---|---|---|
| Iron (% Daily Value) | ... | Iron (% Daily Value)(3.0) |
| Protein | ... | Protein(4.0) |
| Cholesterol(0.0) | ... | 0.0 |
| Cholesterol(1.0) | ... | 0.0 |
| Cholesterol(2.0) | ... | 0.05 |
| Cholesterol(3.0) | ... | 0.35 |
| Cholesterol(4.0) | ... | 0.6 |

**Figure 16:** *Cholesterol*

**Conditional Probability Distribution for Sugars:**

Conditional Probability Distribution for Sugars:

| Cholesterol | ... | Cholesterol(4.0) |
|---|---|---|
| Protein | ... | Protein(4.0) |
| Vitamin A (% Daily Value) | ... | Vitamin A (% Daily Value)(3.0) |
| Sugars(0.0) | ... | 0.08333333333333333 |
| Sugars(1.0) | ... | 0.75 |
| Sugars(2.0) | ... | 0.16666666666666666 |
| Sugars(3.0) | ... | 0.0 |
| Sugars(4.0) | ... | 0.0 |

**Figure 17:** *Sugars*

**Conditional Probability Distribution for Dietary Fiber:**

Conditional Probability Distribution for Dietary Fiber:

| Iron (% Daily Value) | ... | Iron (% Daily Value)(3.0) |
|---|---|---|
| Dietary Fiber(0.0) | ... | 0.0 |
| Dietary Fiber(1.0) | ... | 0.028169014084507043 |
| Dietary Fiber(2.0) | ... | 0.28169014084507044 |
| Dietary Fiber(3.0) | ... | 0.6901408450704225 |

**Figure 18:** *Dietary Fibe*

Direct influencers of Calories are Total Fat, Cholesterol and Sugars.
**Probability of Calories given high Total Fat:**

| Calories | phi(Calories) |
|---|---|
| Calories(0.0) | 0.0032 |
| Calories(1.0) | 0.0032 |
| Calories(2.0) | 0.0032 |
| Calories(3.0) | 0.3054 |
| Calories(4.0) | 0.6850 |

**Figure 19:** *Calories given high Total Fat*

### Interpretation of Results

The conditional independence analysis and CPDs provide several important insights into the relationships between nutrients:

- **Caloric Content:**
  - Directly influenced by Total Fat, Sugars, and Cholesterol.
  - Foods high in these components are more likely to have higher calorie content, as observed in the CPD of Calories.
- **Role of Fats:**

  - Saturated Fat and Trans Fat are highly dependent on Total Fat and Calories.
  - Fats, being calorie-dense, significantly contribute to the energy value of foods, emphasizing their critical role in caloric content.

- **Carbohydrates:**
  - Influenced by Calories, Total Fat, and Sugars.
  - Sugars are a crucial component, directly affecting the energy density of carbohydrate-rich foods.

- **Micronutrients and Proteins:**
  - Micronutrients such as Vitamin A and Iron show dependencies with Total Fat and Protein, indicating that nutrient-dense foods typically balance fats and proteins.
  - Protein, while less directly linked to calorie content, interacts significantly with fats and micronutrients.

- **Key Insights:**
  - **Central Role of Macronutrients:** The Markov blankets for Calories, Total Fat, and Protein confirm that macronutrients (fats, proteins, carbohydrates) and specific micronutrients are essential to the nutritional structure.
  - **Simplified Relationships:** Conditional independence among certain variables simplifies the understanding of their interactions, allowing for targeted analysis and optimization in food composition.

This Bayesian network provides a detailed view of the interdependencies in nutritional data, helping to identify key contributors to caloric and nutrient composition.

## 5.3 Application

The Bayesian network was used to generate synthetic samples, maintaining the conditional dependencies observed in the original dataset.

### Comparison of Statistics Between Original and Generated Data

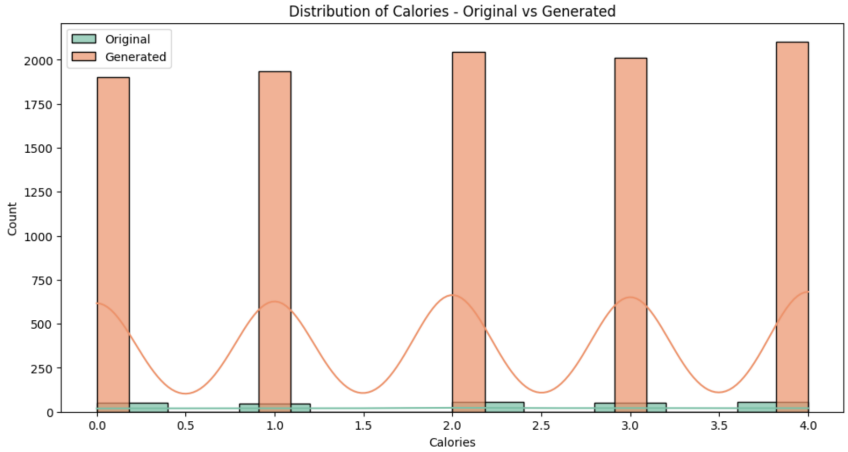


**Figure 20:** *Calories*

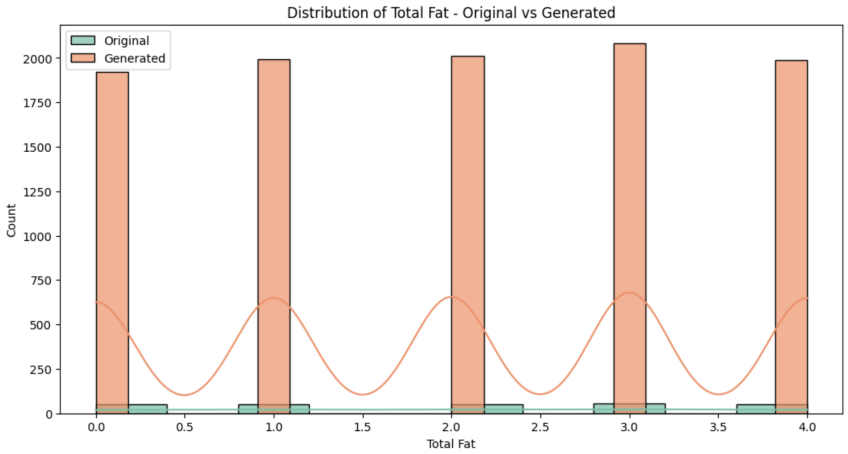


**Figure 21:** *TotalFat*

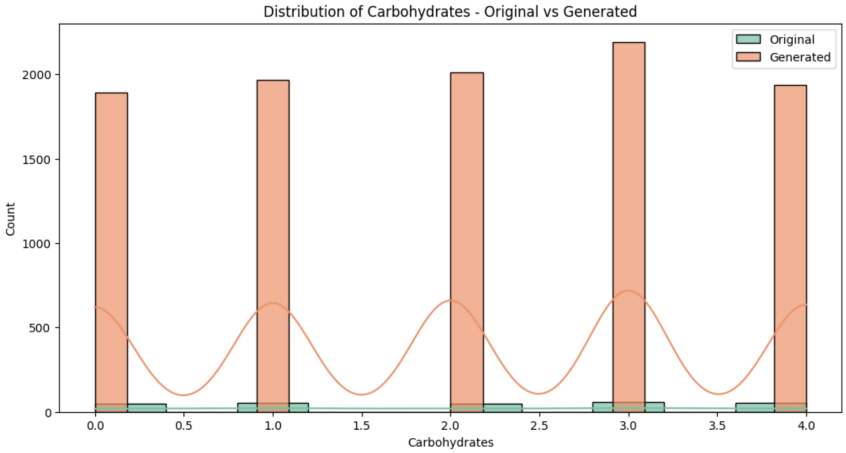


**Figure 22:** *Carbohydrates*

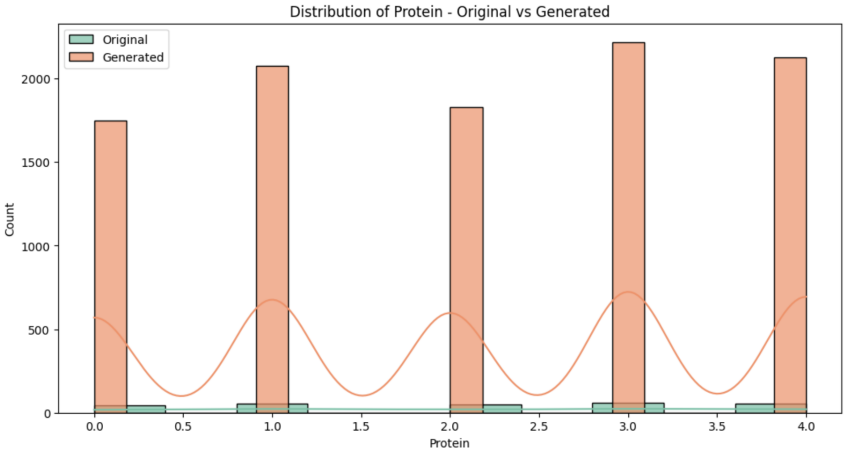


**Figure 23:** *Protein*

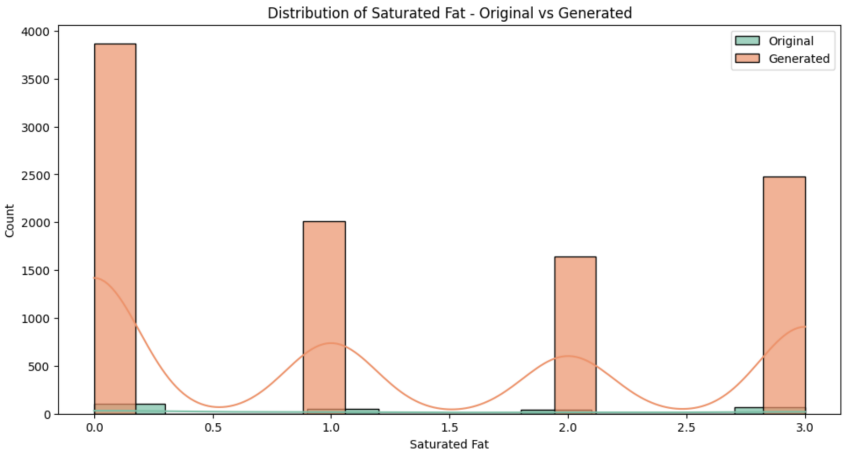


**Figure 24:** *Satured Fat*

- **Data Volume:**
  - The original dataset contained 260 samples, while the generated dataset has 10,000 samples, which aligns with our goal of increasing the data volume.
- **Value Ranges:**
  - The minimum and maximum values for all variables in the generated data match those in the original data, indicating preservation of the overall data structure.
- **Mean Values:**
  - The mean values are very close for all variables. For example:
    - Calories: 2.046 (original) vs 2.050 (generated).
    - Total Fat: 2.035 vs 2.035.
    - Carbohydrates: 2.031 vs 2.005.
  - This suggests that the overall distribution of values has been maintained.
- **Standard Deviations:**
  - The standard deviations are also very close, indicating preservation of data dispersion.
- **Quartiles (25%, 50%, 75%):**
  - The quartiles for most variables are identical or very close, suggesting preservation of the data distribution shape.

- **Specific Observations:**
  - **Trans Fat:** The binary variable (0 or 1) maintained its nature in the generated data.
  - **Vitamin C:** The skewed distribution with a predominance of zero values was preserved.
- **Minor Differences:**
  - There are slight differences in some mean values and standard deviations, which is normal and even desirable for synthetic data, as it prevents exact replication of the original dataset.

# 6 Refitting the Model

## 6.1 Data Transformation and Cleaning

Following the multivariate analysis, the dataset was scaled, reduced, and transformed to optimize predictive performance. Key steps included:

- **Handling Missing Values:** Median imputation was applied to ensure consistency across features.
- **Outlier Treatment:** Winsorization and log transformations were used to address extreme values.
- **Normalization:** Log and square root transformations were employed to reduce skewness in variables such as **Calories**, **Total Fat**, and **Sugars**.
- **Feature Scaling:** Standardization was applied to all numerical variables to ensure a consistent scale.
- **Feature Engineering:** Interaction terms and polynomial features were added to capture non-linear relationships between variables, as identified in the multivariate analysis.

## 6.2 Model Retraining

Two models were evaluated and compared:

1. **Cleaned Model (Baseline):**
   - $R^2$**:** 0.9781.
   - **RMSE:** 0.1219.

- This model significantly improved over the original unprocessed model by addressing data quality issues, reducing overfitting, and enhancing generalizability.

2. **Multivariable Model:**

   - $R^2$**:** 0.9907.
   - **RMSE:** 0.0875.
   - The inclusion of multivariate analysis and engineered features resulted in a substantial improvement in predictive performance.
   - Compared to the cleaned model, this model reduced RMSE by approximately 28.2% and increased $R^2$ by 1.27%.

**Comparison of Model Performance:**

**Table 4:** *Performance Metrics Comparison*

| **Model** | $R^2$ | **RMSE** | **Improvement Over Baseline** |
|---|---|---|---|
| Cleaned Model | 0.9781 | 0.1219 | - |
| Multivariable Model | 0.9907 | 0.0875 | +28.2% RMSE, +1.27% $R^2$ |

# 7 Conclusions and Recommendations

1. **Identified Patterns:**
    - Clear grouping of nutritional variables into macronutrient-dense, sugar-heavy, and protein-rich categories, as revealed by clustering and factor analysis.
    - The analysis identified distinct nutritional patterns, such as the close relationship between fats and calorie content, and the role of carbohydrates and sugars in energy density.
    - Clusters can guide menu optimization, enabling tailored nutritional strategies for different product categories.
2. **Impact of Multivariate Analysis:**
    - Multivariate techniques such as PCA, t-SNE, and Bayesian networks revealed key dependencies among variables, improving data interpretation.
    - Synthetic data generation significantly expanded the dataset from 260 to 10,000 samples while preserving the statistical characteristics of the original data.
    - The synthetic data maintained value ranges, distributions, and specific variable characteristics, enabling robust testing without compromising data integrity.
3. **Model Performance Improvements:**
    - The multivariable model demonstrated superior performance compared to the baseline and cleaned models, achieving an $R^2$ of 0.9907 and an RMSE of 0.0875.
    - Feature engineering, including interaction terms and polynomial features, captured complex nonlinear relationships, enhancing predictive accuracy.
4. **Recommendations:**
    - Leverage identified clusters to design menu items tailored to specific nutritional profiles or consumer preferences.
    - Use the multivariable model for accurate calorie prediction and data-driven decision-making in menu design.
    - Incorporate synthetic data into future analyses to enhance model generalization and support robust testing.

- Continue exploring advanced multivariate techniques to uncover further insights and improve model robustness.

**Conclusion:** The multivariable model demonstrates a substantial improvement over the cleaned baseline model. The reduction in RMSE and the increase in $R^2$ highlight the effectiveness of incorporating multivariate techniques, feature interactions, and polynomial terms. This model is recommended for further predictions due to its accuracy and ability to generalize well to unseen data.

# 8 Annexes

This section includes supplementary visualizations referenced in the analysis. These figures specifically pertain to the section **Comparison of Statistics Between Original and Generated Data** and provide insights into the correlation structures and scatterplots of the original and generated datasets.

## 8.1 Correlation Heatmaps

The following heatmaps illustrate the correlation matrices for the original and generated datasets, highlighting the similarity in statistical relationships between variables.

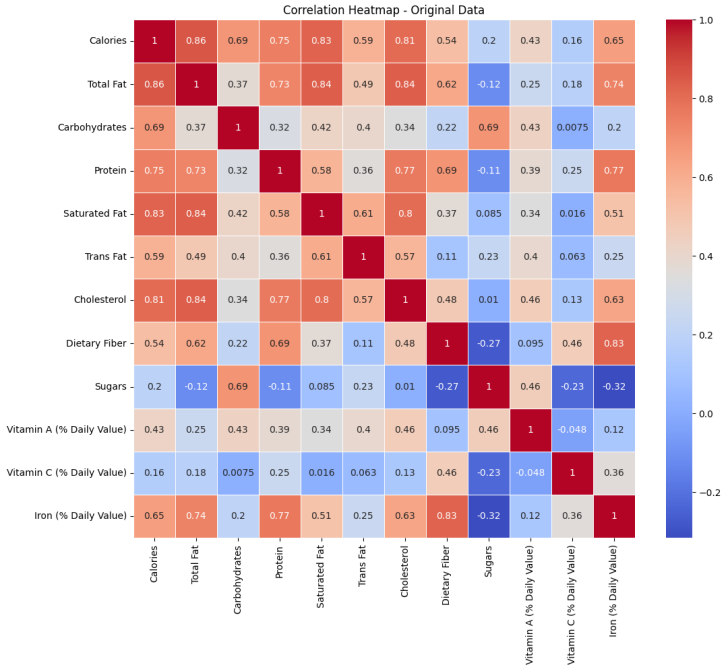


**Figure 25:** *Correlation Heatmap - Original Data*

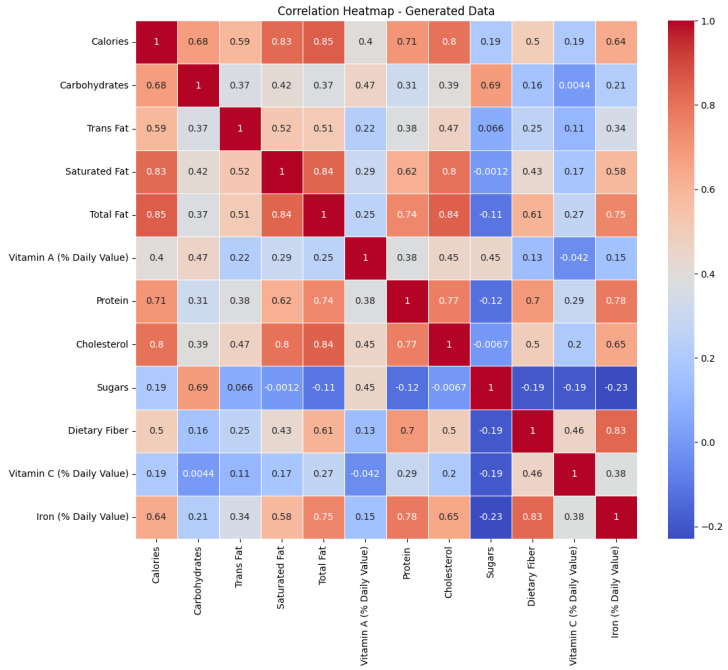


**Figure 26:** *Correlation Heatmap - Generated Data*

## 8.2 Scatter Plots

Scatter plots comparing **Total Fat** vs. **Calories** in the original and generated datasets are provided below. These plots showcase the preservation of patterns and distributions between the datasets.

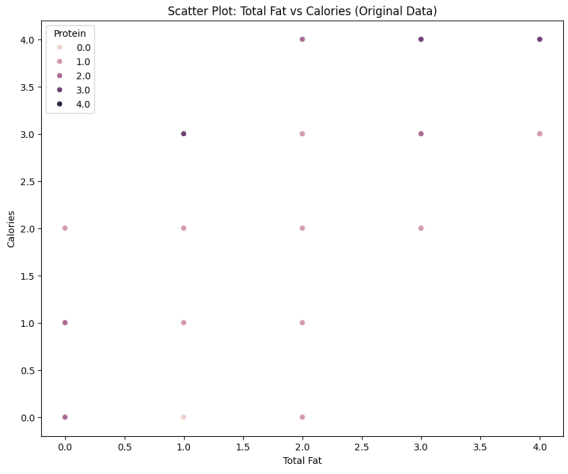


**Figure 27:** *Scatter Plot: Total Fat vs. Calories (Original Data)*

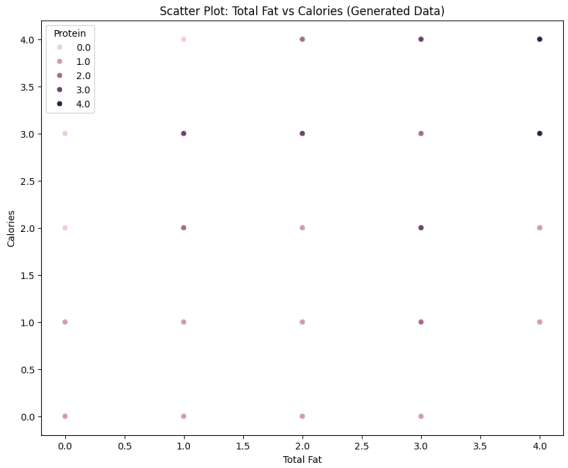


**Figure 28:** *Scatter Plot: Total Fat vs. Calories (Generated Data)*

## Code

The full code used for this analysis is available upon request. The code for this project is implemented in a Google Colab notebook. You can access the notebook using the following link:

Click here to view the Google Colab Notebook

# References


[1] J.H. Friedman. *Greedy Function Approximation: A Gradient Boosting Machine.* The Annals of Statistics, 29(5), 2001.

[2] T. Chen and C. Guestrin. *XGBoost: A Scalable Tree Boosting System.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785 and 794, 2016.

[3] L. Breiman. *Random Forests.* Machine Learning, 45(1), 5-32, 2001.

[4] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. *Learning representations by back-propagating errors.* Nature, 323, 533-536, 1986.

[5] M. Sokolova and G. Lapalme. *A systematic analysis of performance measures for classification tasks.* Information Processing and Management, 45(4), 427-437, 2009.

[6] R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons, 1987.

[7] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning.* MIT Press, 2016. Recuperado de `https://www.deeplearningbook.org`

[8] F. Pedregosa et al. *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, 2825-2830, 2011.

[9] Kaggle. *Titanic - Machine Learning from Disaster.* Recuperado de `https://www.kaggle.com/competitions/titanic`

[10] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009. Recuperado de `https://web.stanford.edu/~hastie/ElemStatLearn/`

[11] K.P. Murphy. *Machine Learning: A Probabilistic Perspective.* MIT Press, 2012.

[12] C.M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[13] K. He, X. Zhang, S. Ren, and J. Sun. *Deep Residual Learning for Image Recognition.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770-778, 2016.

[14] Y. Zhang and Q. Yang. *A Survey on Multi-Task Learning.* IEEE Transactions on Knowledge and Data Engineering, 34(1), 2021.

[15] UCI Machine Learning Repository. *Automobile.* Recuperado de `https://archive.ics.uci.edu/dataset/10/automobile`

[16] R. Kohavi. *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.* International Joint Conference on Artificial Intelligence (IJCAI), 1995.

[17] F. Chollet. *Keras.* GitHub repository, 2015. Recuperado de `https://github.com/keras-team/keras`