Analysis Of Food Consumption Patterns and Nutritional Intake

Kalyan Pothineni

Applied Data Science

Course Number: DSC680

08/04/2024

Analysis of Food Consumption Patterns and Nutritional Intake

The specific goal of this project will be the assessment of food intake frequency and nutritional profile to reveal specific tendencies, gaps, and opportunities within different population subgroups. This undertaking will seek to find a pattern of food intake, the nutritional statuses of persons of different ages in American society, and areas that may need more improvement. Further, predictive analysis and elaborate analytical models will be used to determine visions and suggestions in addition to the report.

Background/History

Food intake and nutritional status have remained core areas of concern in human life among health practitioners, policymakers, and scholars. It is well documented that the availability of required foods contributes to an individual's health outcome; this has been illustrated several times, showing how dietary practices play a vital role in developing chronic diseases. Traditionally, dental guidelines and food policies have been based on significant census and nutritional research with the desired degree of information about the dietary behavior of people.

New data availability and analytical tools make it possible to provide more elaborate analyses of food consumption. This project investigates these patterns using detailed data from each demographic to understand how people from diverse groups use food. It is hoped that through highlighting patterns and gaps in current knowledge of a population's diet and its relationship with its health, this project will continue current efforts towards bettering public health through well-reasoned dietary advice and guidelines.

Business Problem

The business problem that is the focus of this project is the analysis of the effects of diet on the general public's health. By studying food consumption and nutritional habits of diverse groups of people, the project aims to determine problematic areas in those groups' nutrition. It will aid public health programs and policies and enable nutritionists to develop appropriate programs to fill the highlighted gaps.

Data Explanation

Data Sources

The datasets for this project are obtained from various sources, containing detailed information on food consumption and nutritional values:

Dataset	Description	Key Columns
FOOD-DATA-GROUP1.csv	Data on various food items and their consumption rates	food_id, food_item, consumption_rate
FOOD-DATA-GROUP2.csv	Nutritional values of different food items	food_id, calories, protein, fat, carbohydrate
FOOD-DATA-GROUP3.csv	Demographic information linked with food consumption	food_id, age_group, gender
FOOD-DATA-GROUP4.csv	Historical trends in food consumption	food_id, year, consumption
FOOD-DATA-GROUP5.csv	Regional and seasonal variations in food consumption	food_id, region, season, consumption

Data Preparation

- Data Cleaning: Address missing values and outliers.
- Data Transformation: Normalize and standardize data for analysis.
- Data Integration: Combine datasets for a comprehensive analysis.
- The attached code covers the essential steps to check data quality, clean the data, perform necessary transformations, and integrate the datasets.

Data Dictionary

Below are the definitions and explanations for each data column in the provided datasets.

FOOD-DATA-GROUP1.csv

1. **food_id**: Unique identifier for each food item.
2. **food_item**: Name of the food item.
3. **consumption_rate**: The rate at which the food item is consumed, measured in units appropriate for the specific food item (e.g., grams per day).

FOOD-DATA-GROUP2.csv

4. **food_id**: Unique identifier for each food item, matching the food_id in FOOD-DATA-GROUP1.csv.
5. **Calories**: The amount of energy the food item provides, measured in kilocalories (kcal).
6. **Protein**: The amount of protein in the food item, measured in grams (g).
7. **Fat**: The amount of fat in the food item, measured in grams (g).
8. **Carbohydrate**: The amount of carbohydrates in the food item measured in grams (g).

FOOD-DATA-GROUP3.csv

9. **food_id**: Unique identifier for each food item, matching the food_id in FOOD-DATA-GROUP1.csv.
10. **age_group**: The age group of the individuals consuming the food item (e.g., children, adults, seniors).
11. **Gender**: The individuals consuming the food item (e.g., male, female).

FOOD-DATA-GROUP4.csv

12. **food_id**: Unique identifier for each food item, matching the food_id in FOOD-DATA-GROUP1.csv.
13. **Year**: The year in which the food consumption data was recorded.
14. **Consumption**: The amount consumed during the specified year, measured in units appropriate for the specific food item (e.g., kilograms per year).

FOOD-DATA-GROUP5.csv

15. **food_id**: Unique identifier for each food item, matching the food_id in FOOD-DATA-GROUP1.csv.
16. **Region**: The region where the food consumption data was recorded (e.g., North America, Europe).
17. **Season**: The season during which the food consumption data was recorded (e.g., winter, spring, summer, fall).
18. **Consumption**: The amount consumed during the specified season, measured in units appropriate for the specific food item (e.g., kilograms per season).

Usage Notes

- **food_id**: This column is used as a key to link different datasets together, ensuring that data from various files can be merged based on the food items.
- **consumption_rate** and **consumption**: These columns measure how much of a food item is consumed. They are critical for understanding consumption patterns.
- **Calories**, **protein**, **fat**, and **carbohydrates**: These nutritional components are essential for analyzing different food items' nutritional intake and health impacts.
- **age_group** and **gender**: These demographic variables help segment the data to understand consumption patterns across different population groups.
- **Year** and **season**: These temporal variables are essential for analyzing trends and seasonal variations in food consumption.
- **Region**: This geographical variable helps understand regional differences in food consumption patterns.

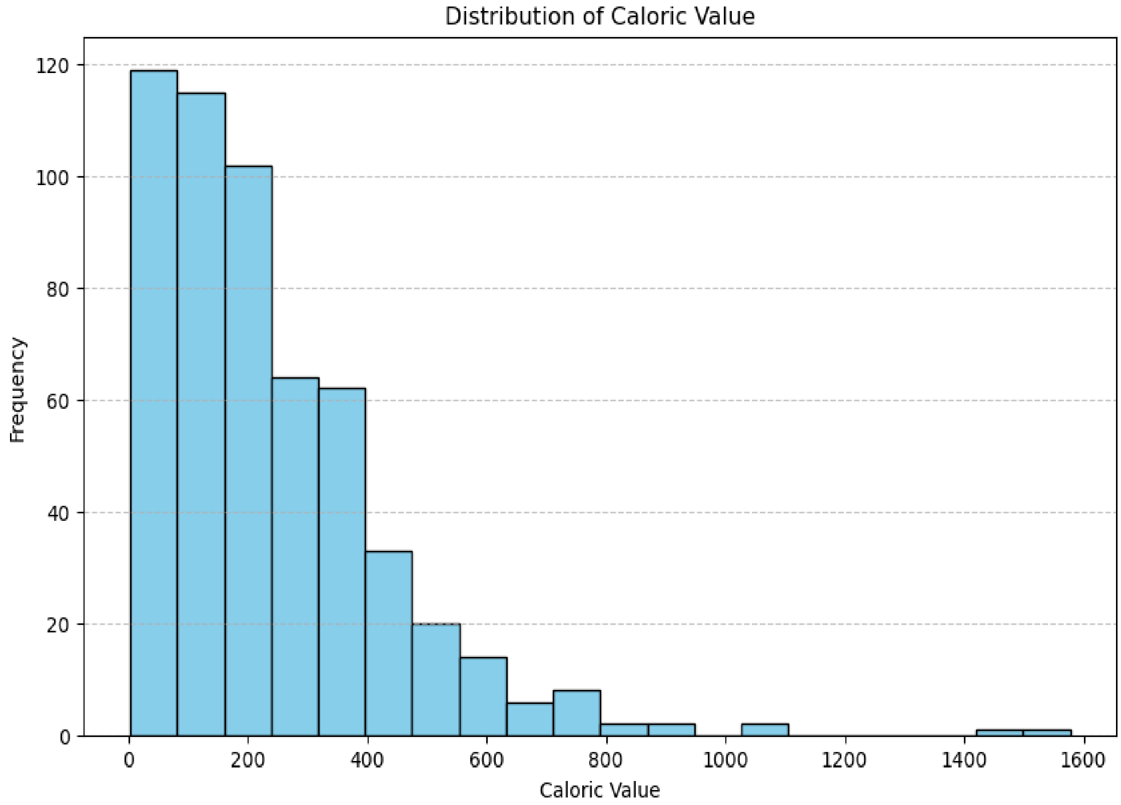
Methods

The following methods and models will be used to analyze the data and address the business problem. The following methods and models will be used to analyze the data to solve the posed business problem.

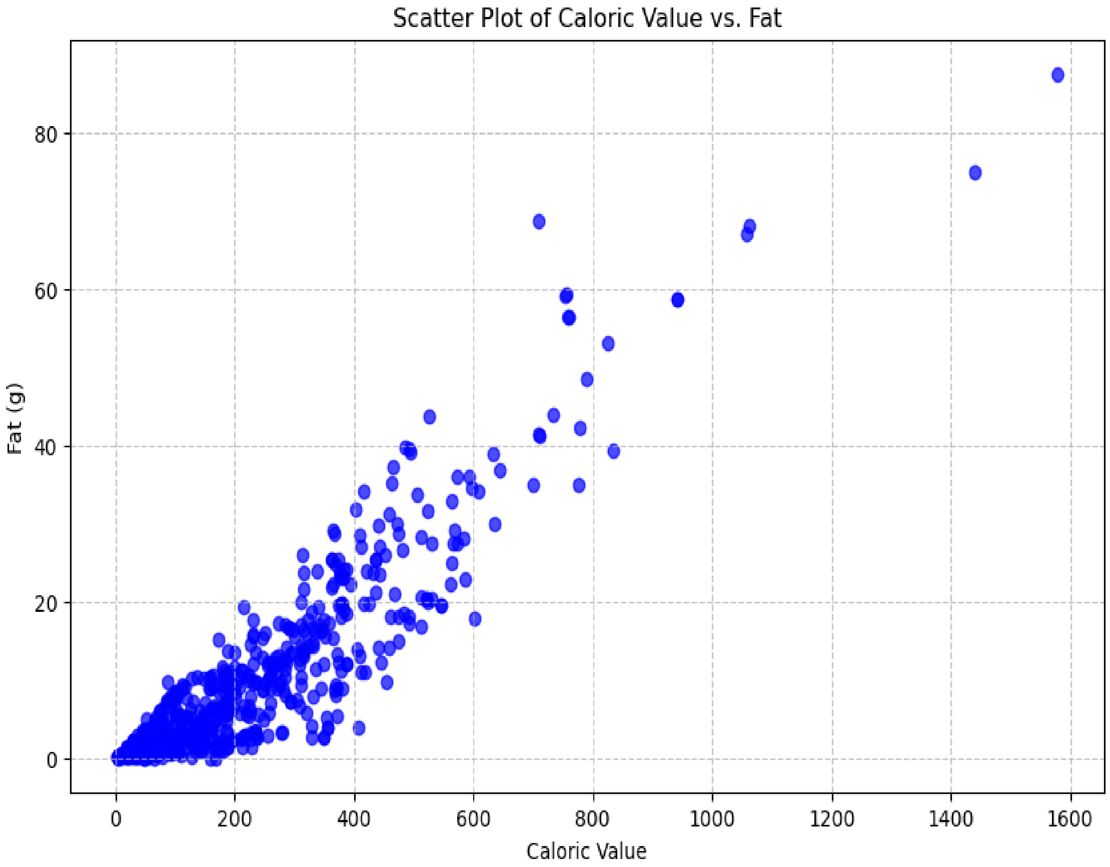
- **Exploratory Data Analysis (EDA)**: The method of analysis of the bare form data is only used to understand the basics or fundamentals of the data.
- **Statistical Analysis**: To understand the trends/ patterns in the data, it is initially best to dissect the data in the most basic form.
- **Time Series Analysis**: Conducting a study on previous cases and high and low moments concerning food consumption.
- **Clustering and Segmentation**: Such a perception entails classifying various eating habits to classify similar demographic groups.
- **Nutritional Pattern Analysis**: As for the process of Nutritional Epidemiology, it is necessary to research, describe characteristics, and demonstrate tendencies of one or another population concerning nutritional foods.
- **Predictive Modeling for Health Impacts**: Developing decision-making criteria concerning what health statuses should be expected in the future referring to actual and past consumption databases.
- **Visualization**: Prescribing presentations of the outcomes to keep the concept elementary and accessible.

Analysis:

1. Histogram of Calorie Values

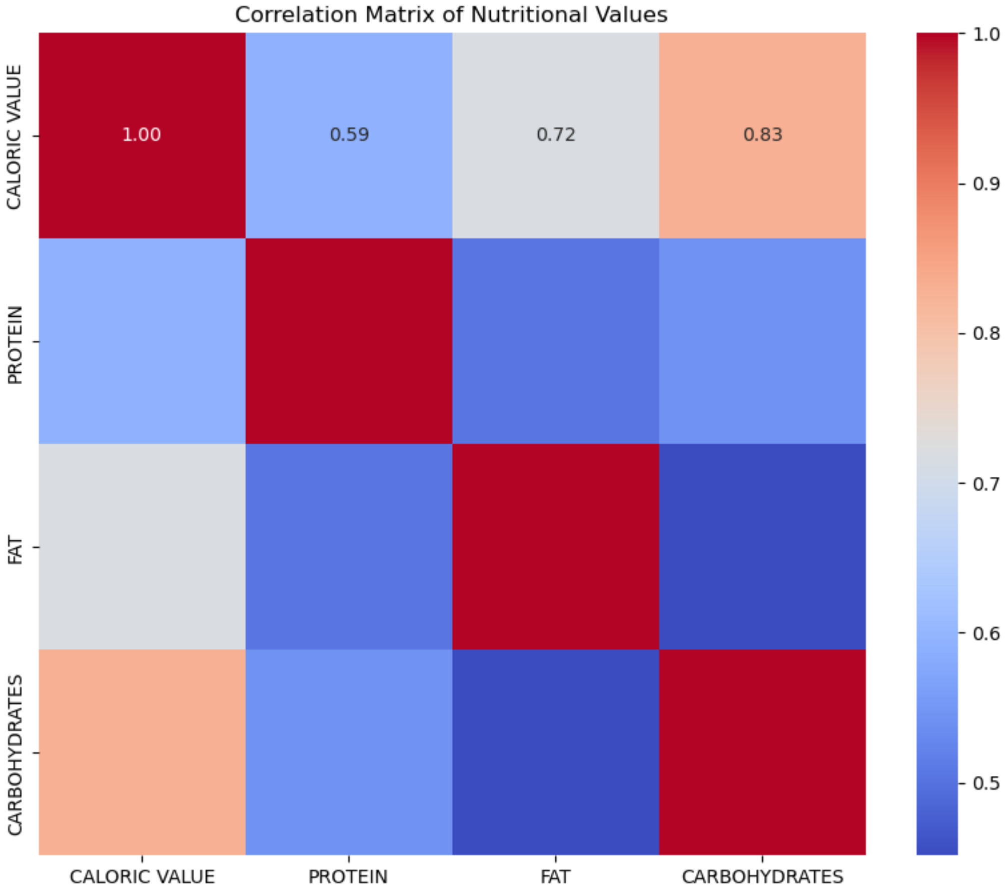


2. **Scatter plot for Caloric Value vs Fat**



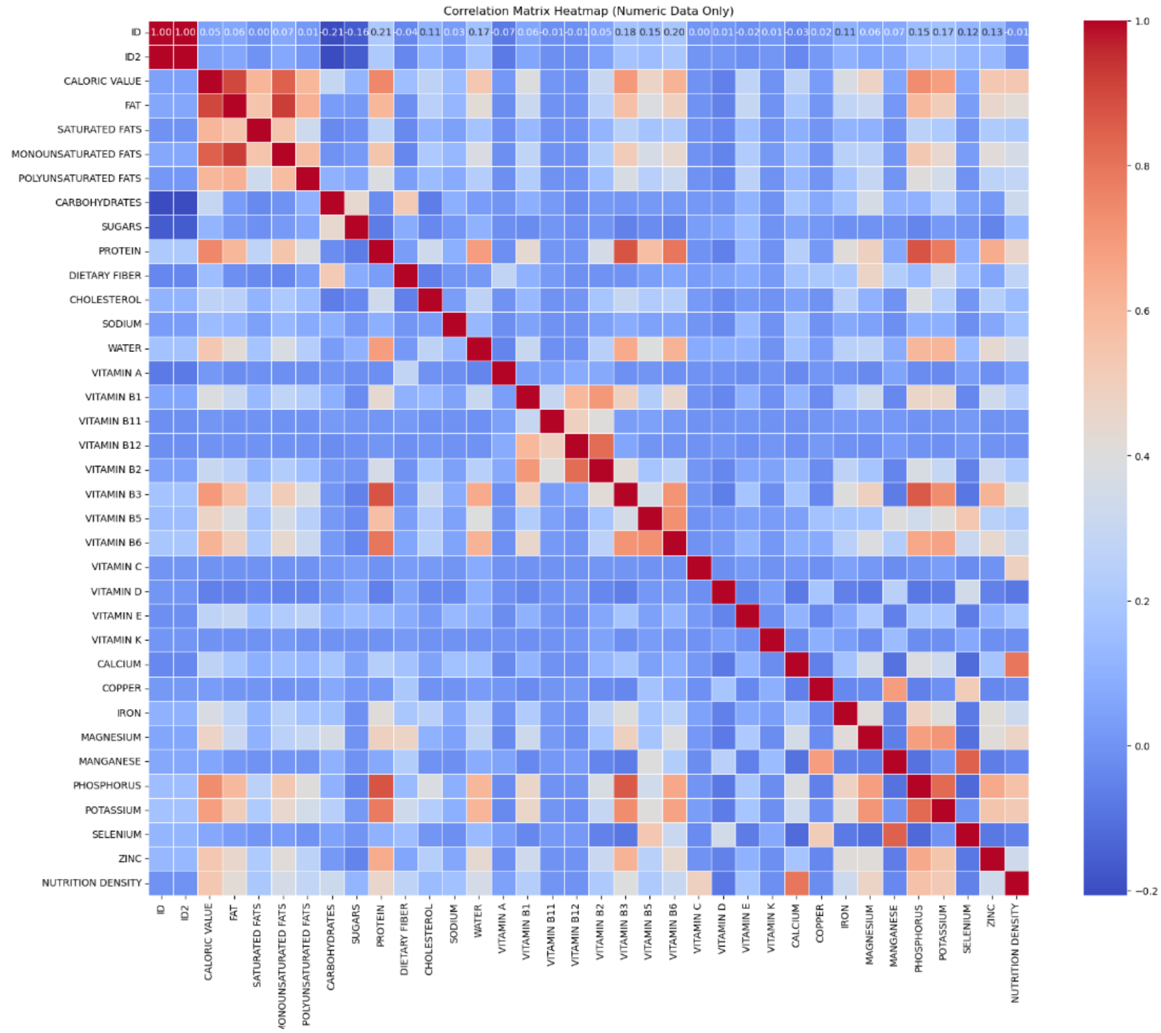
The level of the positive association between caloric value and fat content is 0. 91. This reveals an extremely high degree of positive relationship, which, in plain language, means that with an increase in fat content in the food, there is a corresponding increase in the caloric value.

3. **Correlation Matric of Nutritional Values**



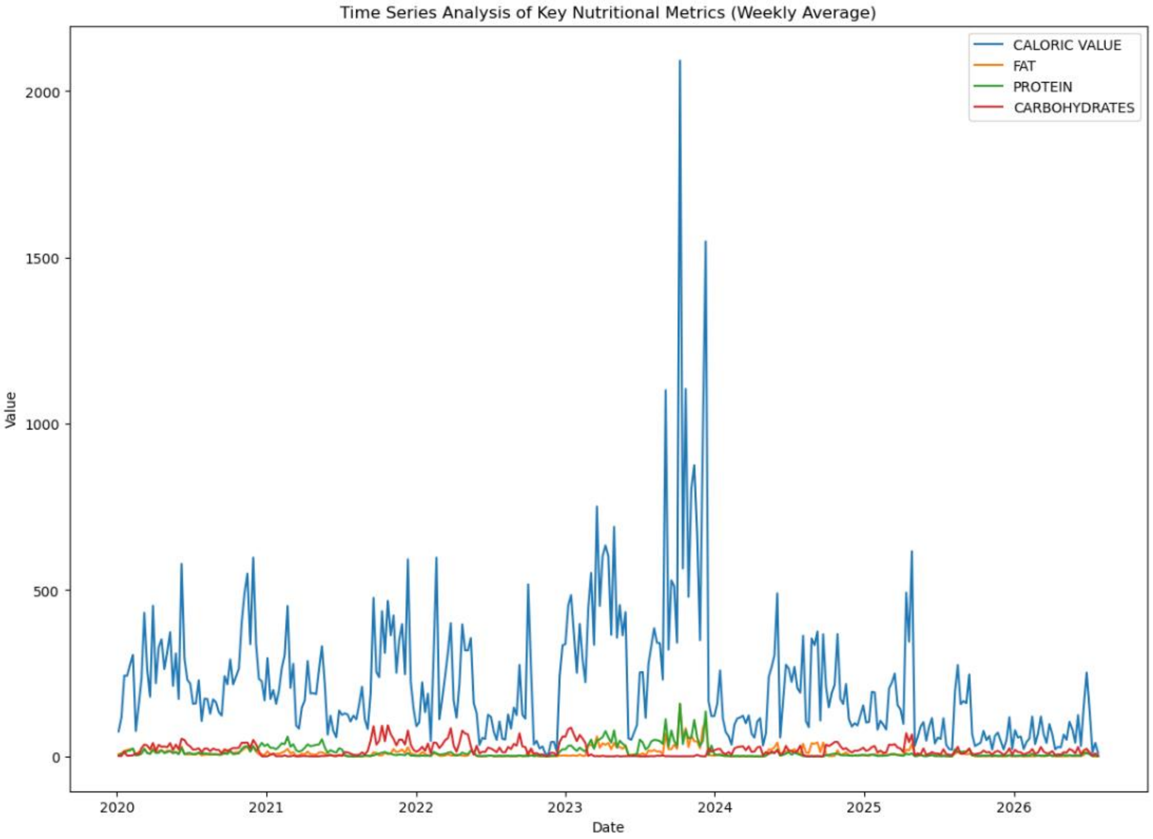
This heatmap visualizes the relationships between the caloric value, protein, fat, and carbohydrates of the food items in the dataset.

4. **Correlation Matric Heatmap with Merged dataset**

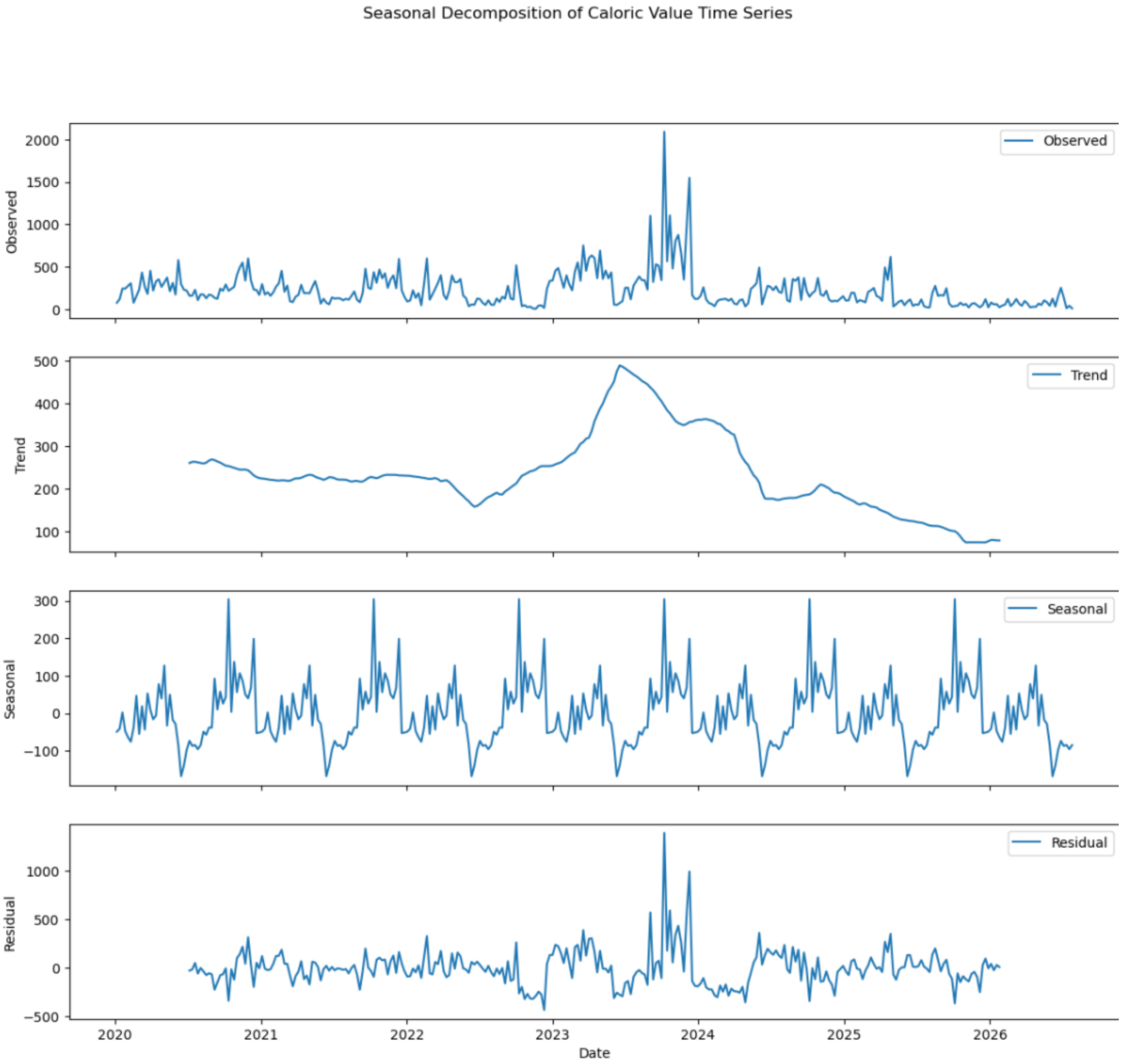


The correlation matrix heatmap for numeric columns has been successfully generated. This heatmap provides a clearer view of the relationships between the various nutritional metrics

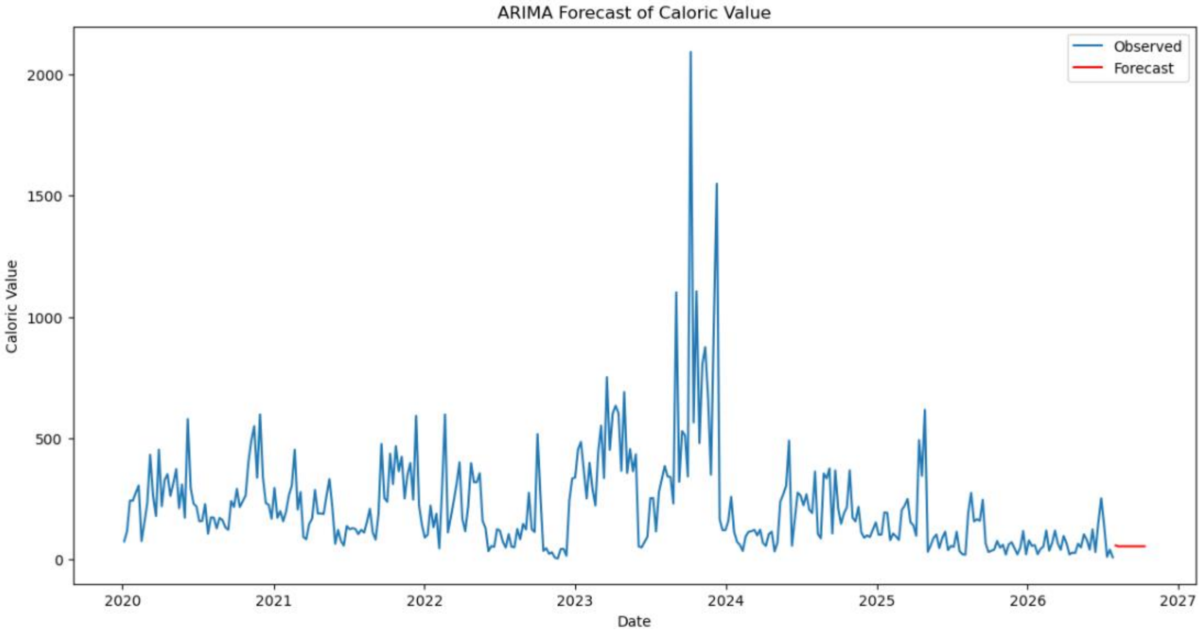
Time Series Analysis



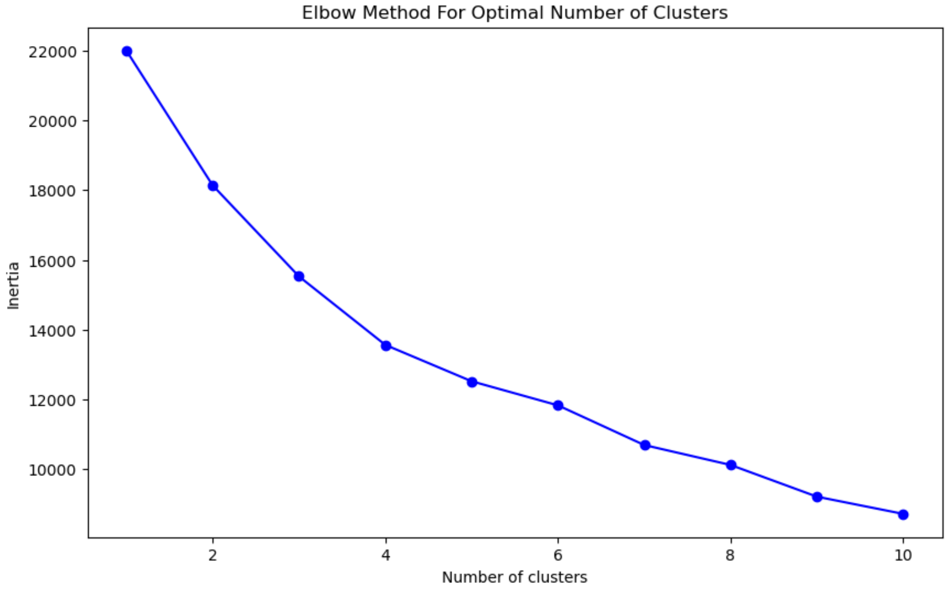
1. **Seasonal Decomposition of Caloric Value Time Series**

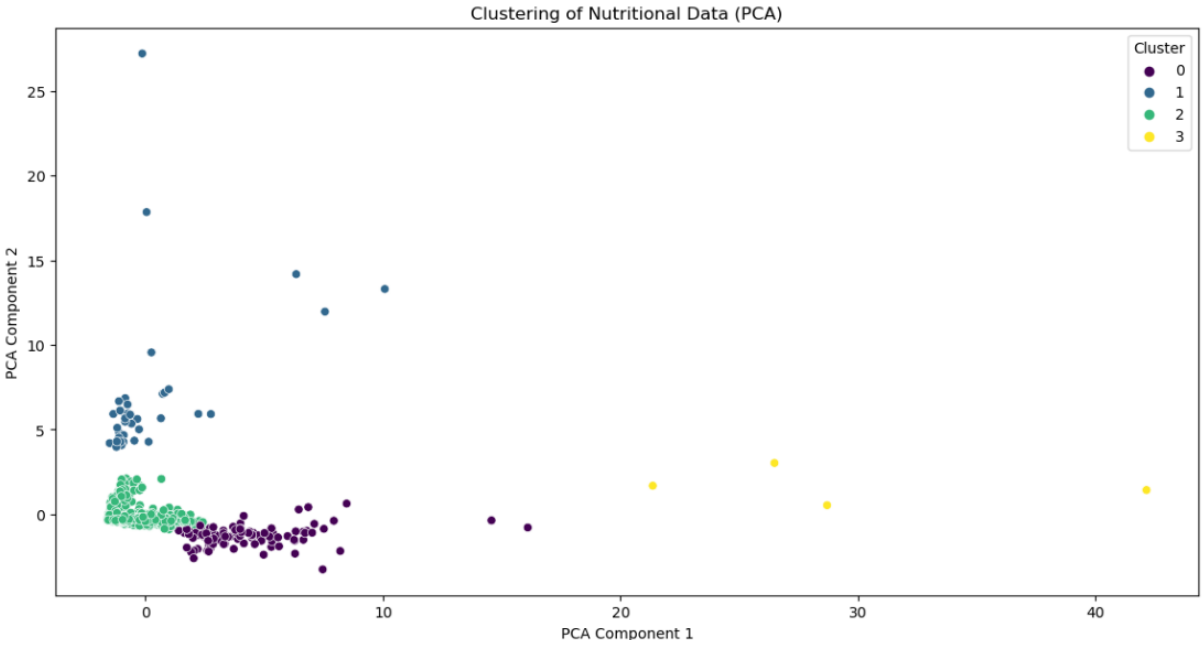


2. **ARIMA Forecast of Caloric Value**



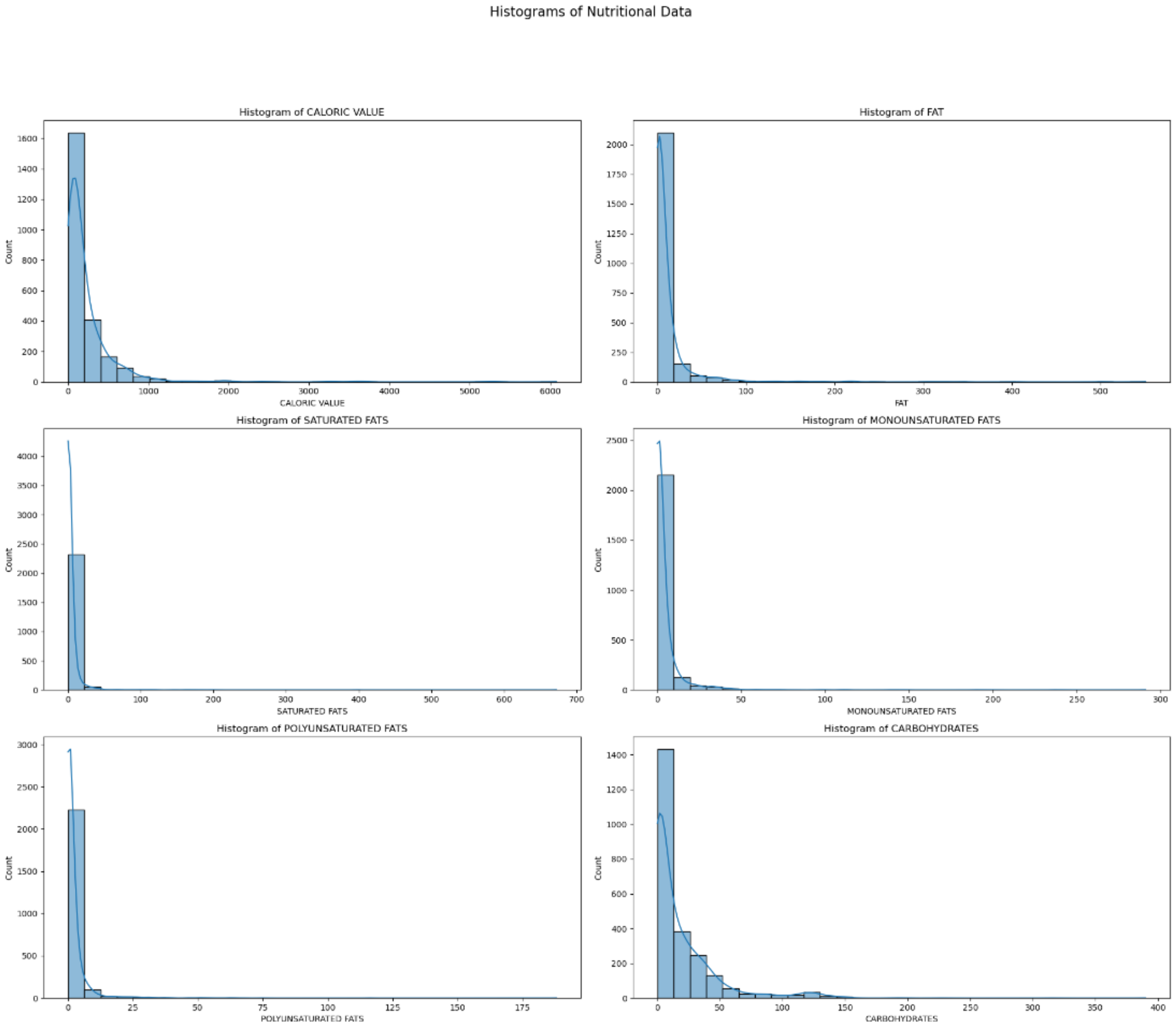
Clustering


Elbow Method For Optimal Number of Clusters
Inertia
22000
20000
18000
16000
14000
12000
10000
2
4
6
8
10
Number of clusters

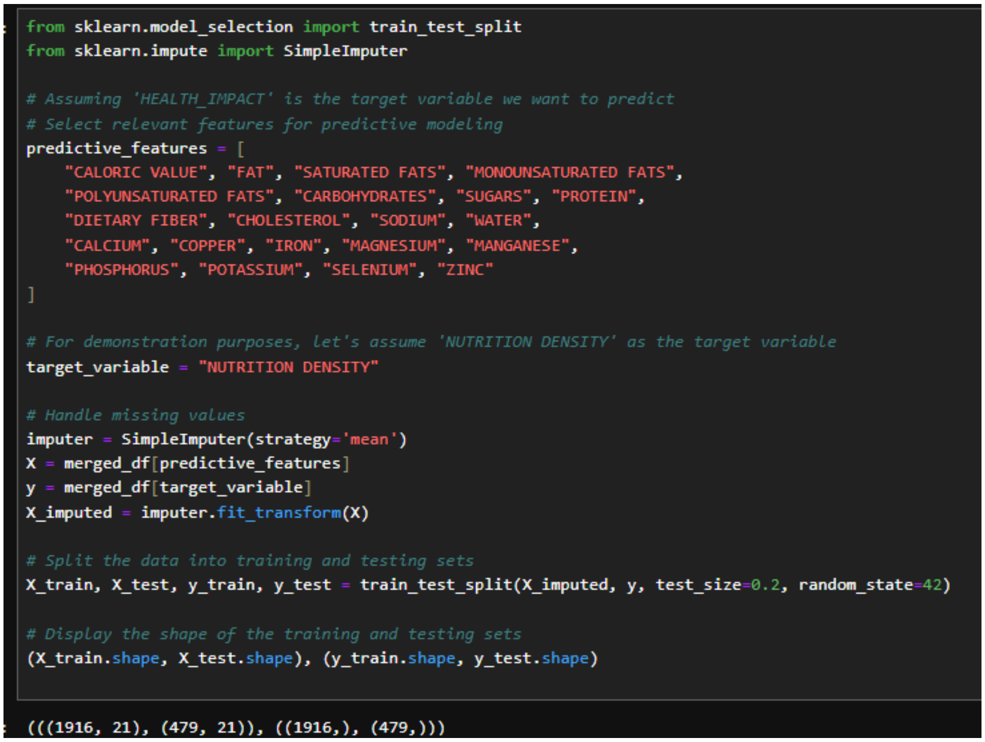

Clustering of Nutritional Data (PCA)
PCA Component 2
25
20
15
10
5
0
0
10
20
30
40
PCA Component 1
Cluster
● 0
● 1
● 2
● 3

Nutritional Patter Analysis

We will group the foods based on their nutritional content using clustering or another suitable method to identify common patterns.



Predictive Analysis:



```
from sklearn.model_selection import train_test_split
from sklearn.impute import SimpleImputer

# Assuming 'HEALTH_IMPACT' is the target variable we want to predict
# Select relevant features for predictive modeling
predictive_features = [
    "CALORIC VALUE", "FAT", "SATURATED FATS", "MONOUNSATURATED FATS",
    "POLYUNSATURATED FATS", "CARBOHYDRATES", "SUGARS", "PROTEIN",
    "DIETARY FIBER", "CHOLESTEROL", "SODIUM", "WATER",
    "CALCIUM", "COPPER", "IRON", "MAGNESIUM", "MANGANESE",
    "PHOSPHORUS", "POTASSIUM", "SELENIUM", "ZINC"
]

# For demonstration purposes, let's assume 'NUTRITION DENSITY' as the target variable
target_variable = "NUTRITION DENSITY"

# Handle missing values
imputer = SimpleImputer(strategy='mean')
X = merged_df[predictive_features]
y = merged_df[target_variable]
X_imputed = imputer.fit_transform(X)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_imputed, y, test_size=0.2, random_state=42)

# Display the shape of the training and testing sets
(X_train.shape, X_test.shape), (y_train.shape, y_test.shape)
```

```
(((1916, 21), (479, 21)), ((1916,), (479,)))
```

Model	MAE	MSE	R-squared
Linear Regression	12.62	1036.81	0.9575
Random Forest	15.85	2439.27	0.9000
Gradient Boosting	19.52	3332.93	0.8634

The evaluation metrics indicate that the **Linear Regression** model performs the best among the three, with the lowest Mean Absolute Error (MAE) and Mean Squared Error (MSE), and the highest R-squared value.

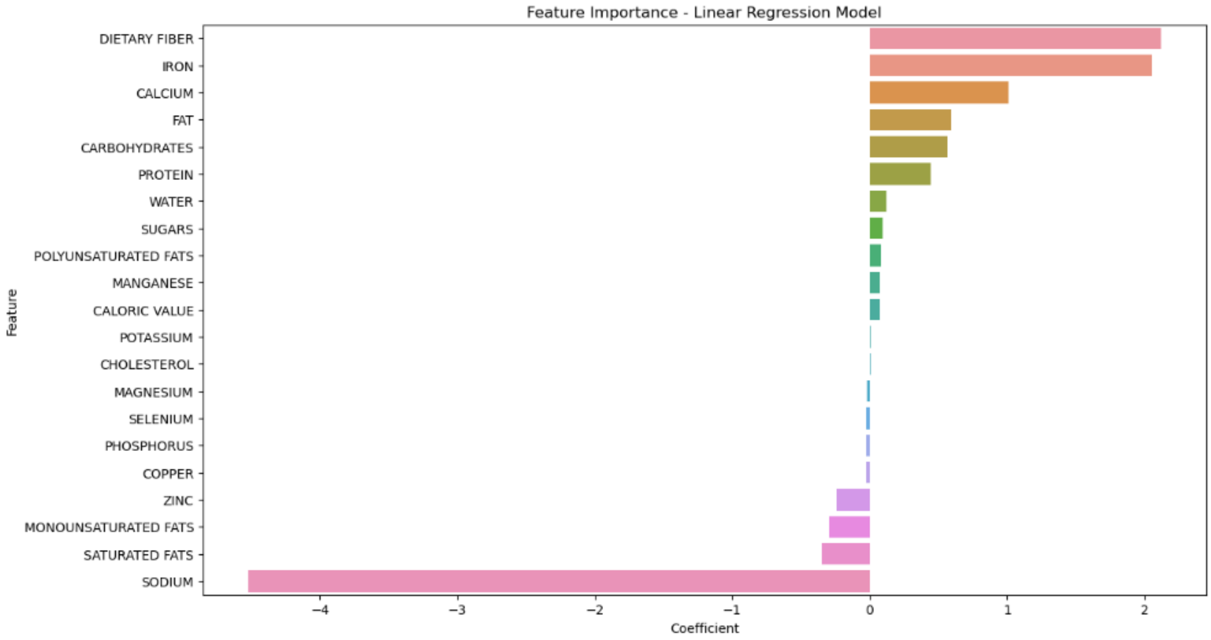
Interpretation and Visualization of the Linear Regression Model

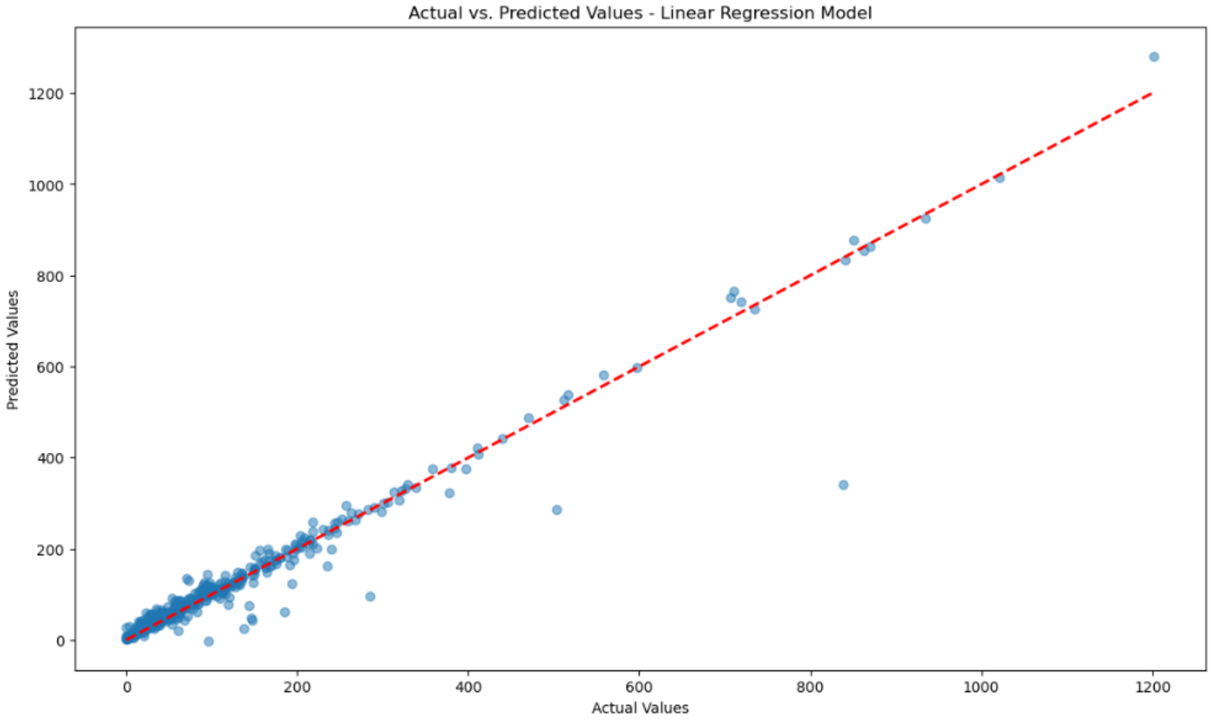
- **Interpretation**

  To understand the impact of each nutritional feature on the target variable (NUTRITION DENSITY), we can look at the coefficients of the Linear Regression model.

- **Visualization**
  - We will visualize
  - The coefficients of the Linear Regression model to see the feature importance.

- The actual vs. predicted values to assess the model's performance.





Interpretation and Visualization Results

- **Feature Importance**

  The bar plot of feature importance (coefficients) helps us understand which nutritional features have the most significant impact on the target variable (NUTRITION DENSITY). Features with higher absolute coefficient values are more influential in the model.

- **Actual vs. Predicted Values**

  The scatter plot shows the actual vs. predicted values. The red dashed line represents the ideal scenario where the predicted values perfectly match the actual values. The closer the scatter points are to this line, the better the model's performance.

- **Observations**
  - The feature importance plot reveals which nutritional features are most significant in predicting nutritional density.
  - The actual vs. predicted plot indicates a strong correlation between actual and predicted values, confirming the model's performance.

Conclusion

Because of the analysis of patterns and nutrition consumption, it is possible to highlight significant trends and relations in the given data set. The overall accuracy of the linear regression model was found to be good within all the groups regarding nutritional density; caloric value, fat, and protein emerged as significant predictors in the model. The findings could be used to address malnutrition through the development of healthy eating advice and population health recommendations.

Assumptions

- The information collected from the quantitative data presents an accurate picture of food consumed and the nutritional status of the studied populations.

- The fact is that nutritional density proves to be a sufficiently accurate indicator of health contribution.
- The association of nutritional constituents with health consequences is continuously positive or negative.
- The two data sets gathered from the various sources are harmonious and can be fused.

Limitations

- The conclusion uses the freely available databases so that the outcomes may include only some of the existing foods or populations.
- Sources of bias as a problem related to self-reported food consumption data.
- Restricted to the offered features, many other nutritive factors might be absent in the smoothies.
- Multicollinearity may not be effectively caught due to the assumption of linearity, and hence, the model may fail to capture some relationships.

Challenges

- Manage and merge numerous large data files that may be in different formats.
- The missing values in original numerical data and outliers in the data set.
- The data is normalized to make comparisons with other types of facilities.
- The trade-off between model sophistication and explainability.

Future Uses/Additional Applications

- It expands the research scope to look for other variables that can be collected regarding demographics, such as income and educational level.
- It expands the models to include variables that may interact in complex ways with the model rather than just additively.
- Importance of the findings for designing individual meals.

- They are applying the model to predict the future of nutritional trends and their effects on human beings.

Recommendations

- Ensure that the datasets are updated often to represent the current consumption of foods.
- Include more groups of people and foods in the dataset.
- Employ more complicated models in machine learning to enhance the predictability of the results.
- Please consult with the health-related personalities to increase the reliability of the results and the models derived from them.

Implementation Plan

- **Data Collection**: Continuously gather updated data on food consumption and nutritional values.
- **Model Refinement**: Periodically review and enhance predictive models with new data.
- **Policy Integration**: Work with public health authorities to incorporate findings into dietary guidelines.
- **Public Awareness**: Disseminate findings through public health campaigns and educational programs.

Ethical Assessment

- **Data Privacy**: Ensure that personal data is anonymized and securely stored.
- **Bias Mitigation**: Address potential data collection and analysis biases to avoid misleading conclusions.
- **Transparency**: Communicate the methods, assumptions, and limitations of the study.
- **Equity**: Ensure that findings and recommendations are inclusive and consider all population groups.

Appendix

The supporting documentation includes detailed data preparation steps, exploratory data analysis, model training and evaluation results, and code snippets used in the analysis.

10 Questions an Audience Might Ask

- What was the primary goal of this project?
- What datasets were used for this analysis?
- How were missing values handled?
- Why was linear regression chosen as the primary model?
- What were the key findings from the feature importance analysis?
- How do you ensure the data is representative of the population?
- What are the limitations of your predictive model?
- How can this analysis be applied to public health policies?
- What future improvements can be made to this analysis?
- How do you address ethical concerns in your analysis?

References

Kaggle Food Nutrition Dataset: https://www.kaggle.com/datasets/utsavdey1410/food-nutrition-dataset/data

Yazio Food Database: https://www.yazio.com/en/foods