CHEMENG 765/SEP 767 :Multivariate Statistical Methods for Big Data Analysis and Process Improvement

Title: Nutritional Profiling With Data Analytics

- By Snigdha Pandey

Introduction - Background

- **The Quest for Optimal Health:** At the heart of a healthy lifestyle lies the complex world of nutrition—a realm where vitamins, minerals, and macronutrients play pivotal roles in shaping our well-being.
- **Decoding Dietary Diversity:** Amidst the plethora of food choices and dietary recommendations, individuals face the daunting task of selecting foods that meet their unique nutritional needs. The sheer diversity of nutrients within different food groups necessitates a deeper understanding of what we consume.
- **Empowering Choices Through Data:** Leveraging the power of data analysis transforms the overwhelming nutrition data into intelligible, actionable insights. By dissecting the nutritional content of food with sophisticated analytical tools, we can guide consumers toward more informed and health-conscious decisions.

Introduction - Academic Literature

- **Pioneering Nutritional Data Analysis**

Early studies in nutritional science have established the importance of data-driven approaches to understand the complex interplay of diet and health. Researchers like Willett et al. have emphasized the need for robust analytical methods to decipher the vast nutritional data available.

- **The Emergence of PCA in Nutritional Research**

Principal Component Analysis (PCA) has been a game-changer in identifying nutrient patterns that influence dietary guidelines. The work by Jolliffe and Cadima outlines PCA as a critical tool for reducing dimensionality in nutritional data, allowing for the capture of dietary patterns across diverse populations.

- **Implications for Public Health and Policy**

These studies provide valuable insights for public health initiatives and inform policy-making in nutritional education.This may be methodically incorporated into various phases of the health policy cycle for fact-based and precise health policy decision-making.

References:

- Willett, W. C., et al. "Food Frequency Questionnaires." American Journal of Epidemiology, vol. 147, no. 3, 1998, pp. 283-290.
- Jolliffe, I. T., and Cadima, J. "Principal component analysis: a review and recent developments." Philosophical Transactions of the Royal Society A, vol. 374, no. 2065, 2016.
- Lee, R. H. (2017). "Data Analytics in Public Health: Shaping Policy and Practice." Public Health Reports, 132(4), 472-480.

Database Description

- Source: USDA National Nutrient Database.
- The dataset consists of 8,618 records (food items) and 45 attributes.
- Each record is for 100 grams.
- Attributes include nutritional information like energy (kcal), protein (g), fat (g), vitamins, and minerals.
- Food items are categorized into different groups (e.g., 'Dairy and Egg Products').

Objectives

- **Nutritional Diversity and Diet Planning**

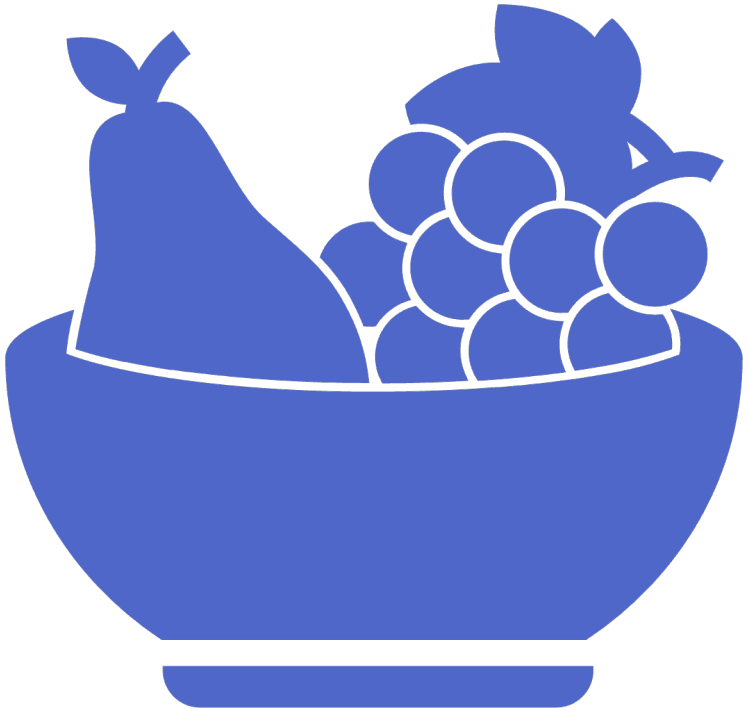
Each food category is a unique tapestry woven with various essential nutrients, critical to a balanced diet. Grasping the intricate nutritional profiles within these groups empowers us to tailor dietary choices that are both informed and beneficial, fostering a diet that is rich in vital nutrients.

- **Insights from PCA Analysis**

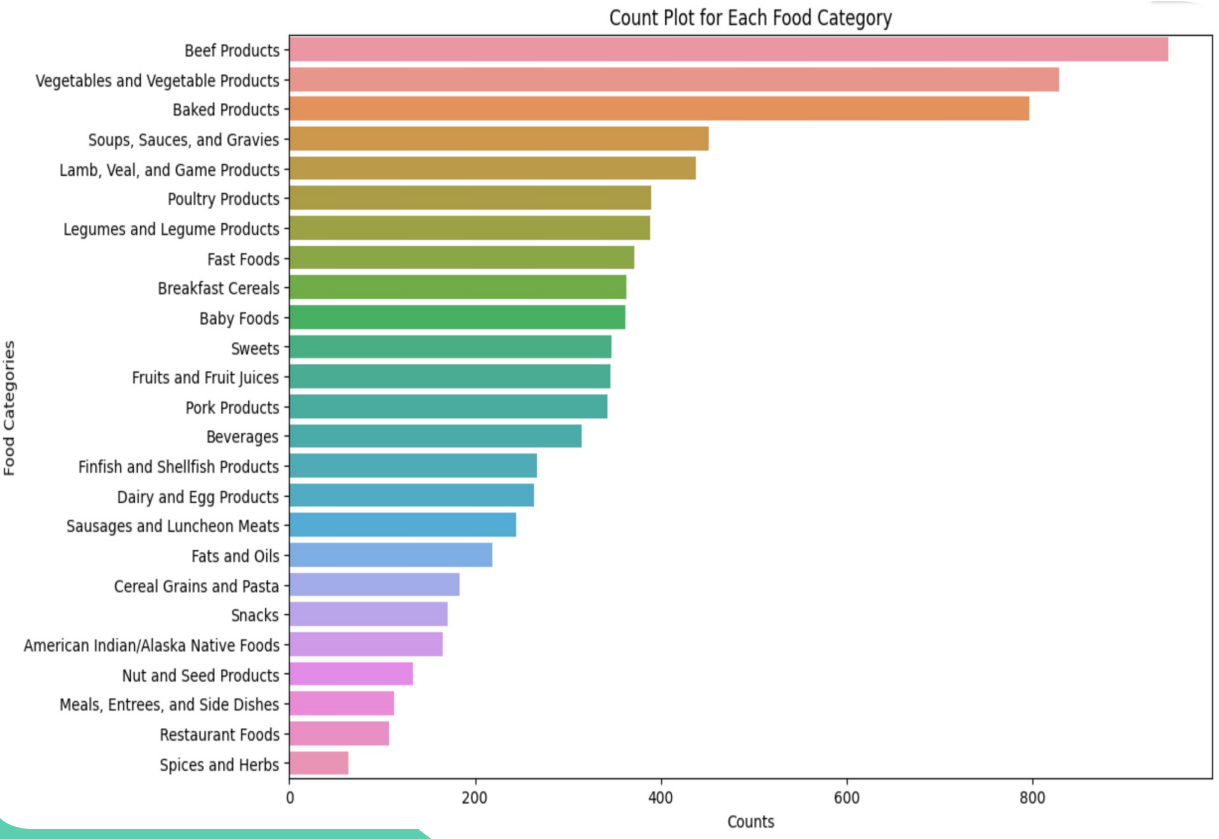
By leveraging the analytical power of Principal Component Analysis (PCA), we unearth the defining nutrients within food groups. This powerful method illuminates the most influential dietary components, serving as a beacon for nutritional planning and education.

- **Advances in Food Classification**

The study progresses to construct a sophisticated model that intelligently categorizes food items. Utilizing a synergistic approach that combines the dimensionality-reducing capabilities of PCA with the robust classification prowess of a Random Forest Classifier, we pave the way for an optimized, data-driven classification system.



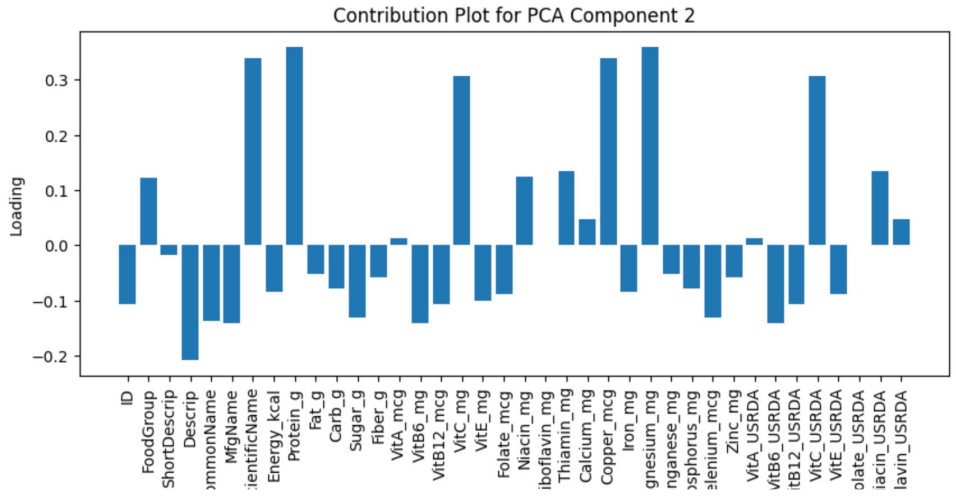
Methodology

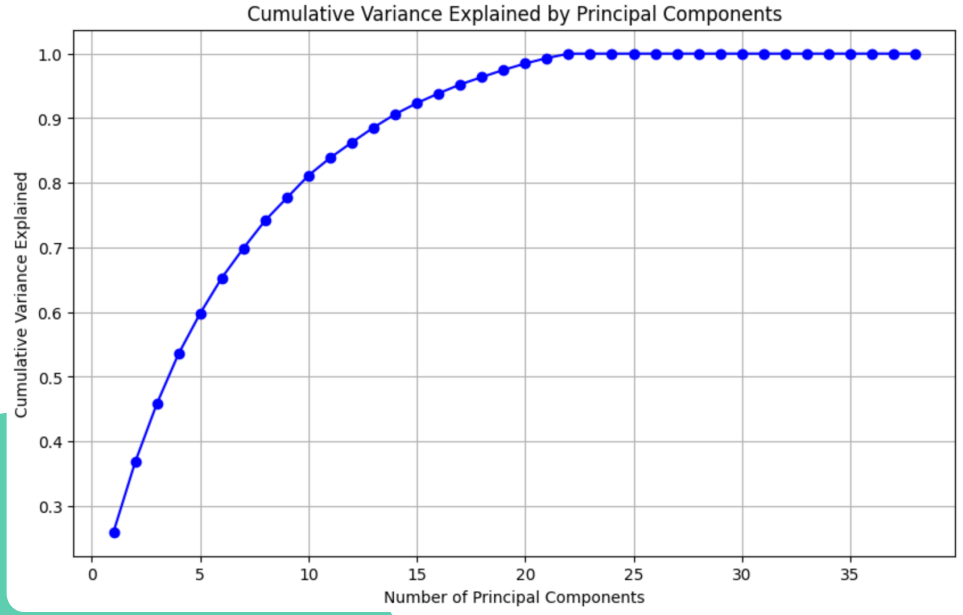


Stage1: Data Preprocessing and Exploration

- Evaluated and confirmed the absence of missing data within key columns, ensuring data integrity.
- Streamlined the dataset by excluding descriptive attributes, focusing solely on quantifiable nutritional values.
- Standardized the numerical features, providing a normalized scale for accurate comparative analysis.

Methodology





Stage 2: Insight Extraction via PCA and Advanced Visualization

- Employed PCA to distill 39 numerical features down to 15, capturing nearly the 90% of the dataset's variance.
- Engaged Plotly, a sophisticated visualization library, to color-code data points by food group within a 3D PCA space.
- Interactive contribution plots were generated, pinpointing essential nutrients for each food category.
- Enhanced the tool's utility for dieticians and interested individuals by enabling in-depth nutritional profiles upon hovering over any food group data point.
- Investigated outlier items visually to discern the nutritional traits that make them distinct.

Methodology

Stage 3: Classifier Development and Dimensionality Reduction

- Undertook the construction of a Random Forest classifier, leveraging both the original and PCA-reduced datasets.
- Assessed the model's efficacy on the full dataset versus the PCA-reduced set to evaluate the trade-offs in performance.
- Despite a marginal decline in accuracy, the PCA-based model demonstrated reduced complexity and increased computational speed.
- The findings indicate a strategic balance between performance and resource utilization, emphasizing efficiency without significantly compromising predictive power.

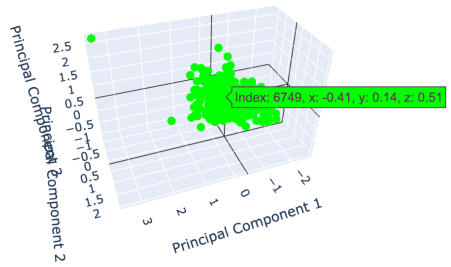
Results - Principal Component Analysis (PCA)

- The visualization tool at the nutritionist's disposal enables an interactive analysis of the 3D graph, allowing for a precise selection and examination of individual data points. This level of detail facilitates an in-depth understanding of the constituent nutrients and their impacts on the overall nutritional profile.
- Through the utilization of contribution plots, we've identified the most influential nutrients within each food group, establishing a foundational nutritional benchmark for these categories. With this data, the nutritionist is equipped to craft personalized diet profiles tailored to the nutritional needs of individuals, taking into account the comprehensive nutritional makeup of each food group. Additionally, the tool empowers the nutritionist to discern and rationalize the presence of outliers, enhancing the precision of dietary recommendations.

```
Important features for Soups, Sauces, and Gravies:
Component 1:
['Phosphorus_USRDA', 'Phosphorus_mg', 'Riboflavin_mg', 'Riboflavin_USRDA', 'Energy_kcal']
Component 2:
['VitA_mcg', 'VitA_USRDA', 'Thiamin_USRDA', 'Thiamin_mg', 'Copper_USRDA']
Component 3:
['Thiamin_USRDA', 'Thiamin_mg', 'VitA_mcg', 'VitA_USRDA', 'VitE_USRDA']
Component 4:
['VitC_mg', 'VitC_USRDA', 'VitB6_USRDA', 'VitB6_mg', 'Sugar_g']
Component 5:
['VitB12_USRDA', 'VitB12_mcg', 'VitC_mg', 'VitC_USRDA', 'Phosphorus_mg']
Component 6:
['VitE_mg', 'VitE_USRDA', 'VitB12_USRDA', 'VitB12_mcg', 'Selenium_mcg']
Component 7:
['Calcium_mg', 'Calcium_USRDA', 'Sugar_g', 'Fat_g', 'VitE_USRDA']
Component 8:
['Magnesium_mg', 'Magnesium_USRDA', 'VitB6_mg', 'VitB6_USRDA', 'VitA_mcg']
Component 9:
['Fat_g', 'VitC_USRDA', 'VitC_mg', 'VitE_USRDA', 'VitE_mg']
Component 10:
['ID', 'Sugar_g', 'Folate_mcg', 'Folate_USRDA', 'Riboflavin_USRDA']
Component 11:
['Selenium_mcg', 'Selenium_USRDA', 'Sugar_g', 'Magnesium_mg', 'Magnesium_USRDA']
Component 12:
['Sugar_g', 'Riboflavin_USRDA', 'Riboflavin_mg', 'Carb_g', 'VitE_mg']
Component 13:
['Zinc_mg', 'Zinc_USRDA', 'Fiber_g', 'VitB6_mg', 'VitB6_USRDA']
Component 14:
['Fat_g', 'Thiamin_mg', 'Thiamin_USRDA', 'Energy_kcal', 'Iron_mg']
Component 15:
['Iron_mg', 'Fat_g', 'Copper_mcg', 'Copper_USRDA', 'VitA_mcg']
```

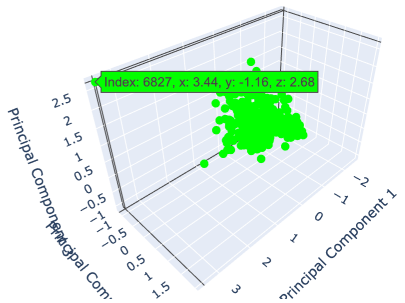
Analyzing Data Points

3D Scatter Plot of PCA Data (Color Coded by Food Groups)

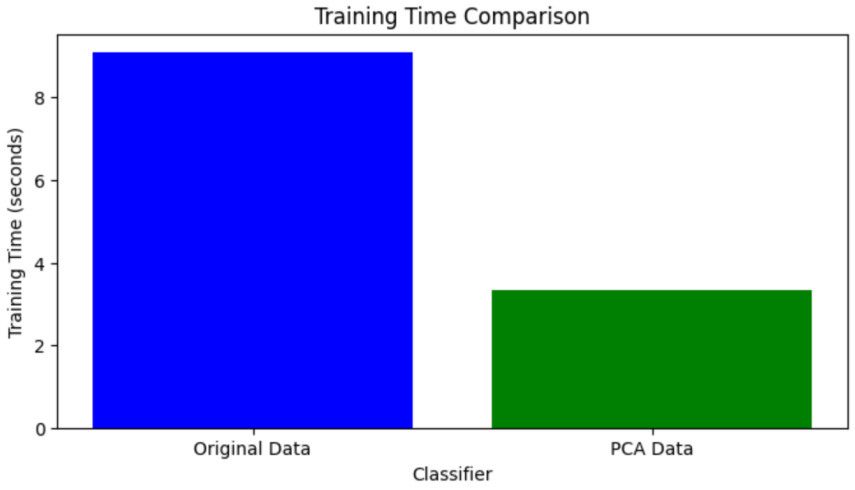


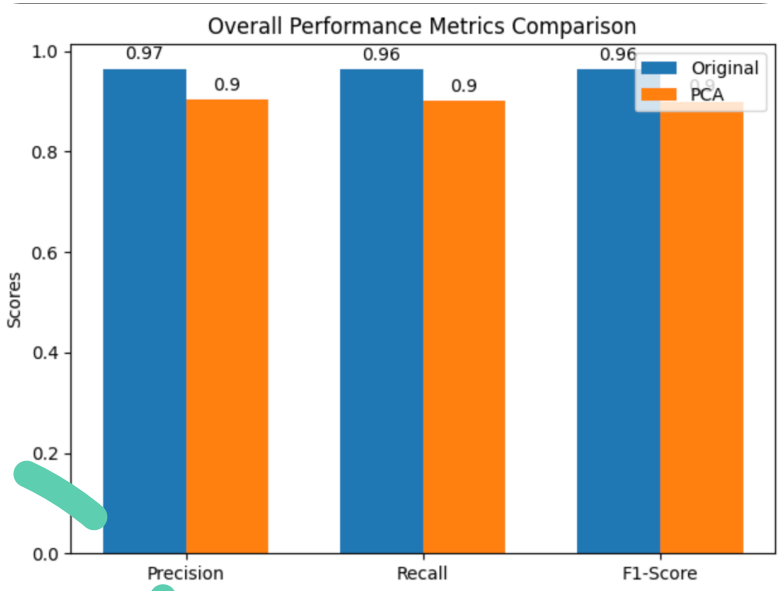

```
Row at index 6749 :
 ID                                21246
FoodGroup                      Fast Foods
ShortDescrip       WENDY'S,CHICK NUGGETS
Descrip          WENDY'S, Chicken Nuggets
CommonName                            NaN
Name: 6749, dtype: object
Top contributing features for datapoint at index 6749 :
 Index(['Protein_g', 'Fat_g', 'Energy_kcal', 'Phosphorus_mg',
       'Phosphorus_USRDA', 'Fiber_g', 'Iron_mg', 'Selenium_mcg',
       'Selenium_USRDA', 'Zinc_mg', 'Zinc_USRDA', 'Sugar_g', 'Carb_g',
       'Copper_USRDA', 'Copper_mcg', 'Calcium_USRDA', 'Calcium_mg',
       'Folate_mcg', 'Folate_USRDA', 'Niacin_USRDA', 'Niacin_mg', 'VitE_mg',
       'VitE_USRDA', 'VitB12_mcg', 'VitB12_USRDA', 'Thiamin_mg',
       'Thiamin_USRDA', 'Riboflavin_mg', 'Riboflavin_USRDA', 'Manganese_mg',
       'VitC_USRDA', 'VitC_mg', 'Magnesium_mg', 'Magnesium_USRDA', 'VitB6_mg',
       'VitB6_USRDA', 'VitA_USRDA', 'VitA_mcg'],
      dtype='object')
```

3D Scatter Plot of PCA Data (Color Coded by Food Groups)




```
Row at index 6827 :
 ID                                        21337
FoodGroup                            Fast Foods
ShortDescrip       MCDONALD'S,PNUTS (FOR SUNDAES)
Descrip          McDONALD'S, Peanuts (for Sundaes)
CommonName                                   NaN
Name: 6827, dtype: object
Top contributing features for datapoint at index 6827 :
 Index(['Energy_kcal', 'Fat_g', 'Magnesium_mg', 'Magnesium_USRDA', 'Protein_g',
       'VitE_mg', 'VitE_USRDA', 'VitB6_USRDA', 'VitB6_mg', 'Phosphorus_mg',
       'Phosphorus_USRDA', 'Niacin_mg', 'Niacin_USRDA', 'Fiber_g', 'Sugar_g',
       'Calcium_USRDA', 'Calcium_mg', 'Folate_USRDA', 'Folate_mcg',
       'VitB12_mcg', 'VitB12_USRDA', 'Zinc_mg', 'Zinc_USRDA', 'Selenium_mcg',
       'Selenium_USRDA', 'Manganese_mg', 'Copper_mcg', 'Copper_USRDA',
       'Thiamin_USRDA', 'Thiamin_mg', 'Iron_mg', 'Riboflavin_mg',
       'Riboflavin_USRDA', 'VitA_USRDA', 'VitA_mcg', 'Carb_g', 'VitC_USRDA',
       'VitC_mg'],
      dtype='object')
```





Results - Classifier

- The PCA-based approach significantly reduces the training time. This is evident from the bar chart, which shows a much shorter bar for the PCA Data compared to the Original Data. This reduction in training time can lead to faster iterations and improvements in model development.
- The PCA approach offers a compromise between efficiency and accuracy. While there might be a slight drop in the performance metrics, this trade-off can be justified by the significant gains in speed and less computational resource usage.
- The PCA reduction likely results in lower memory because of reduced dimensions.
- A PCA-reduced model may generalize better to new data since it's trained on the most relevant features that explain the majority of the variance in the data, potentially avoiding overfitting to noise and less relevant features.

Conclusions & Future work

The integration of data analytics into nutritional profiling has paved the way for more nuanced and personalized dietary recommendations. By harnessing techniques like PCA, we can uncover nutrient patterns that would otherwise be obscured by the complexity of the data. This approach not only assists individuals in making more informed dietary choices but also equips policymakers with the evidence necessary to craft effective public health strategies.

Future Work:

- **Expansion of Datasets:** Future studies should incorporate larger and more diverse datasets to capture a broader spectrum of dietary patterns and preferences.
- **Beyond PCA:** Explore other dimensionality reduction techniques like t-SNE or UMAP before classification to uncover non-linear nutrient patterns.
- **Feature Engineering:** Create interaction features that model the synergy between different nutrients (e.g., Nutrient Interaction Terms such as vitamin D and calcium absorption).