Starbucks Nutritional Analysis Report

Andrici Sabina, Ardelean Alexandru,

Boncu Ana Maria, Anghelscu Adelina

Executive Summary

This project leverages the Starbucks nutritional dataset to analyze menu items, identify patterns in their nutritional profiles, and group similar beverages using machine learning techniques. Principal Component Analysis (PCA) reduced data complexity while retaining significant variance. Hierarchical Cluster Analysis (HCA) then segmented beverages into distinct groups based on their nutritional features. Key findings include clear differentiation among low-calorie, high-sugar, and protein-rich clusters, providing valuable insights for health-conscious consumers and menu optimizations.

1. Data Description

Source: Kaggle Dataset Starbucks menu

(https://www.kaggle.com/datasets/henryshan/starbucks?resource=download)

Overview:

- The dataset includes nutritional information for Starbucks menu items.

- Key columns include:

- Beverage category: Classifies beverages into categories like coffee, tea, or smoothies.

- Beverage: Specific names of drinks (e.g., Caramel Macchiato, Green Tea Latte).

- Beverage prep: Preparation method (e.g., hot or cold, whipped cream, syrup).

- Calories, Total Fat (g), Sugars (g), Protein (g): Key nutritional values.

- Contains both categorical and numerical data.

Observations:

- Data required preprocessing to handle missing values (e.g., NaNs in Caffeine (mg)).

- Key features for analysis were primarily nutritional metrics.

2. Data Preprocessing

Objective: Prepare the raw dataset for analysis by cleaning data, handling missing values, and standardizing numerical features.

Steps Taken:

- Removed extra spaces from column names.

- Converted non-numeric columns to numeric where necessary (e.g., Caffeine (mg)).

- Filled missing values with column means (e.g., Caffeine (mg) NaNs).

- Standardized numerical columns (e.g., Calories, Sugars (g)) for PCA compatibility.

- Saved cleaned data to `cleaned\_starbucks\_menu.csv` for further analysis.

3. Exploratory Data Analysis (EDA)

Objective: Gain insights into the dataset through statistical summaries and visualizations.

Key Findings:

- Distributions:

- Most beverages had calorie values concentrated between 100 and 300.

- High variability observed in sugar and caffeine content.

- Correlations:

- Strong positive correlation between Calories, Sugars (g), and Total Fat (g).

- Weak correlations with Caffeine (mg).

- Outliers:

- Boxplots identified outliers in Calories and Sugars (g) for certain beverages like Frappuccinos.

- Visualizations:

- Histograms, correlation heatmaps, and pair plots provided a clearer view of feature relationships.

4. Principal Component Analysis (PCA)

Objective: Reduce data dimensionality while retaining significant variance.

Steps Taken:

- PCA was applied to standardized numerical features.

- Retained components explaining ~90% of the variance.

- Scree plot indicated optimal retention of three principal components.

Findings:

- First three components captured most of the variability:

- PC1: High loadings on Calories, Sugars (g), and Total Fat (g).

- PC2: Moderate loadings on Protein (g) and Sodium (mg)

- PC3: Primarily driven by Caffeine (mg).

Outputs:

- PCA results saved to `pca\_results.csv`.

5. Clustering Analysis

Objective: Group beverages into clusters based on nutritional profiles.

Steps Taken:

- Hierarchical Cluster Analysis (HCA) was performed using Ward's method.

- Clusters were determined based on dendrogram visualizations, cutting the tree at an appropriate threshold to form distinct groups.

Findings:

- Cluster 1: Beverages with low calorie and sugar content, typically brewed teas.

- Cluster 2: High-calorie and high-sugar beverages, often blended drinks.

- Cluster 3: Moderate-calorie, protein-rich beverages, such as lattes.

Outputs:

- Clustered data saved to `hierarchical\_clustered\_results.csv`.

- Cluster summaries saved to `hierarchical\_cluster\_summary.csv`.

6. Results and Visualizations

Cluster Nutritional Insights:

- Generated bar plots comparing mean values for Calories, Sugars (g), and Protein (g) across clusters.

- Scatter plots visualized cluster separations on principal component axes.

Key Takeaways:

- Clusters revealed distinct beverage groups:

- Health-conscious options.

- High-calorie indulgent drinks.

- Protein-focused items for dietary needs.

Outputs:

- Visualizations confirmed meaningful grouping and differentiation.

7. Discussion and Conclusion

Insights:

- PCA and HCA effectively categorized Starbucks beverages into meaningful groups.

- Low-calorie beverages are predominantly teas, while high-calorie options include blended drinks with significant sugar content.

- Protein-rich beverages offer distinct dietary benefits.

Challenges:

- Limited dataset size restricted the depth of analysis.

- Exclusion of categorical data from PCA may have reduced context.

Opportunities:

- Future analysis could integrate categorical data (e.g., Beverage\_category).

- Insights could inform Starbucks’ product marketing and health-conscious menu design.