Starbucks Beverage Nutritional Data: A Comprehensive Analysis

Name : Sahasa Yadavalli **Student ID :** 23079491
Email ID : sahas96669@gmail.com **Git hub Link :** https://github.com/Yadavallisahasakumar/Clustering-and-Fitting

Introduction

The Starbucks dataset of 242 beverages has `Calories` as one of the main nutritional metrics, with an average of 193.87 kcal and a range of 0 to 510 kcal, which is a huge variation in energy carried by each beverage. Similarly, `Trans Fat` is 1.31 g on average, with a standard deviation of 1.64 g, reflecting variability in fat content. The average values are `Sodium` is 6.36 mg, maximum 40 mg, and `Total Carbohydrates` is 128.88 g, ranging from 0 to 340 g.

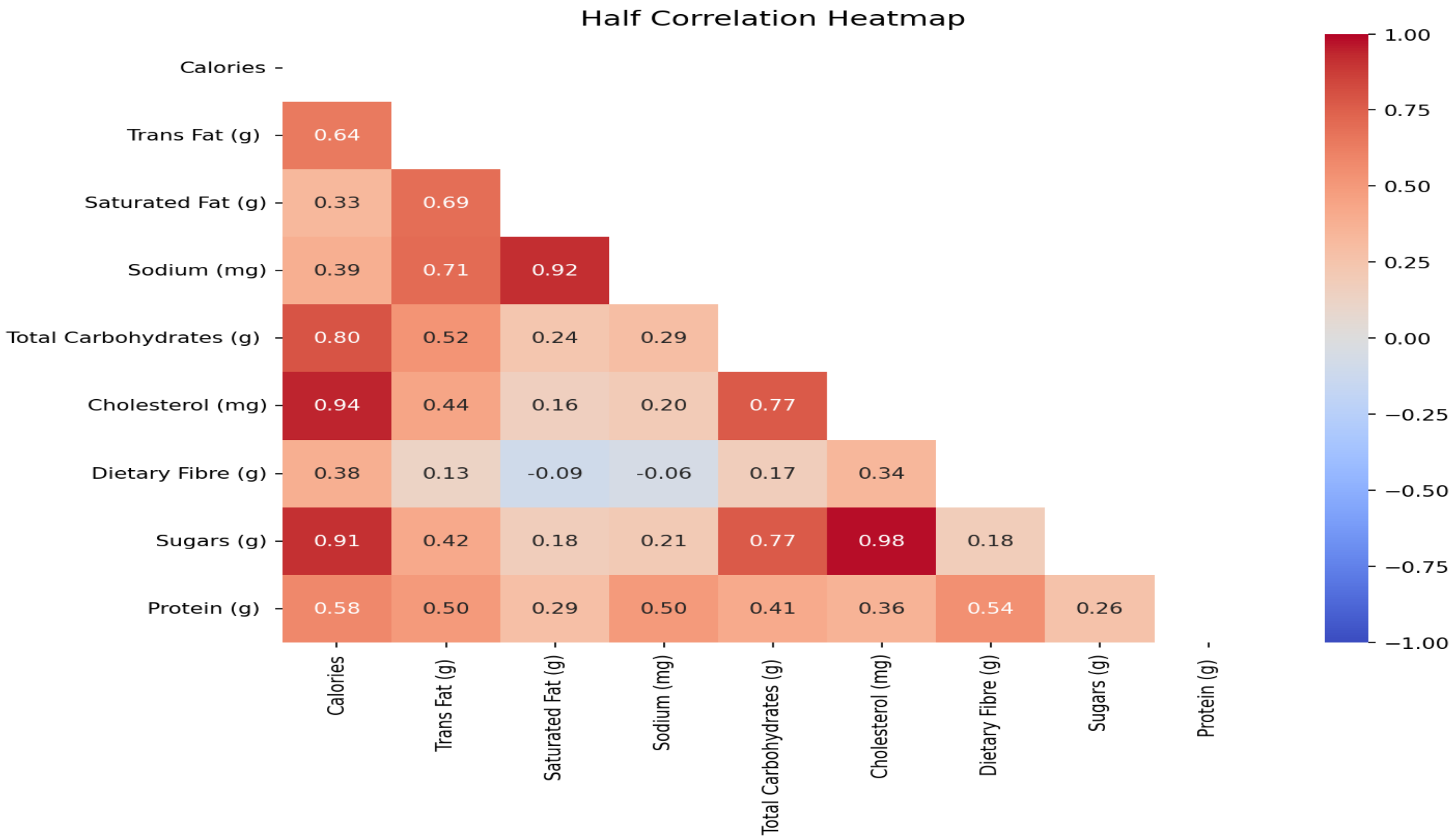
Other nutrients are `Cholesterol`, averaging 35.99 mg, and `Sugars` at an average of 32.96 g. Dietary fibre has an average of 0.81 g, while the protein content averages 6.98 g, with some beverages containing as high as 20 g of protein. The `Caffeine` content has some missing data; the dataset also includes the percentage of key vitamins and minerals (A, C, Calcium, and Iron), but those are stored as categorical variables, which would need further analysis for detailed insight.

This dataset provides an overview of the composition of beverages, which is useful in analysing their nutritional value and comparing different beverage options.

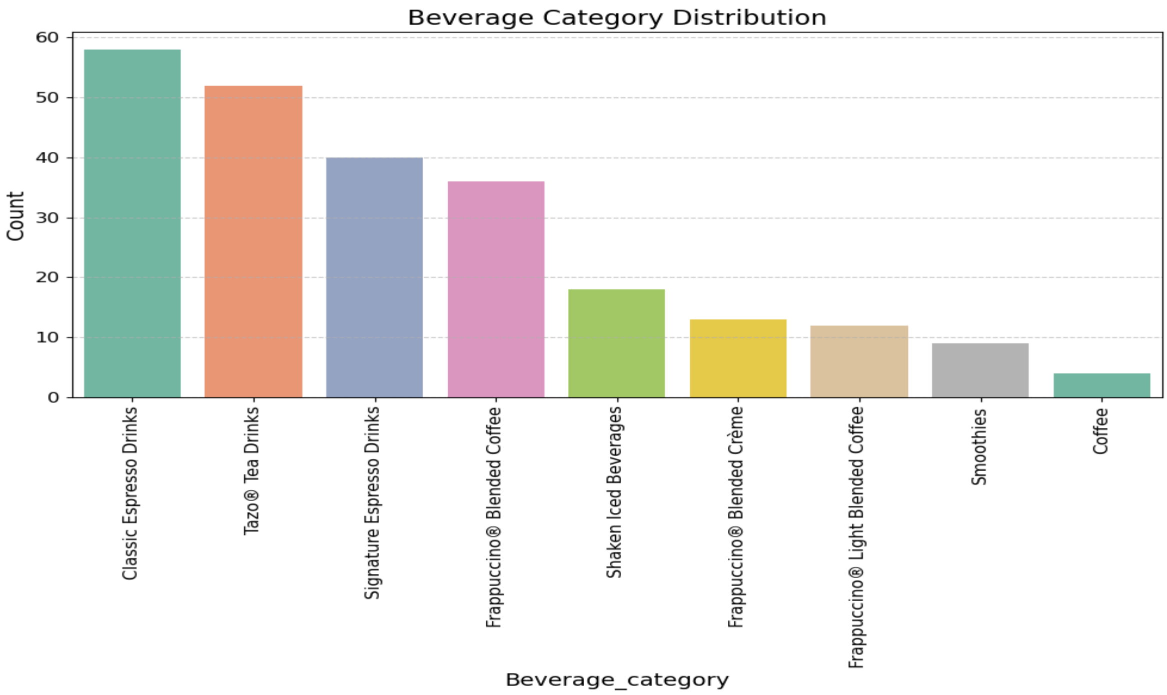
Visualizations

1. **Correlation Heatmap**

The correlation heatmap portrays the relationships of Calories, Sugars, Cholesterol, Sodium, and Saturated Fat nutritional components. Some significant correlations: Calories vs Sugars-0.91, Calories vs Cholesterol-0.94, and very strong relations between Sugars and Cholesterol with a correlation score of 0.98, thus providing a better way of understanding dietary patterns and giving informed decisions about food choice and nutritional intake.

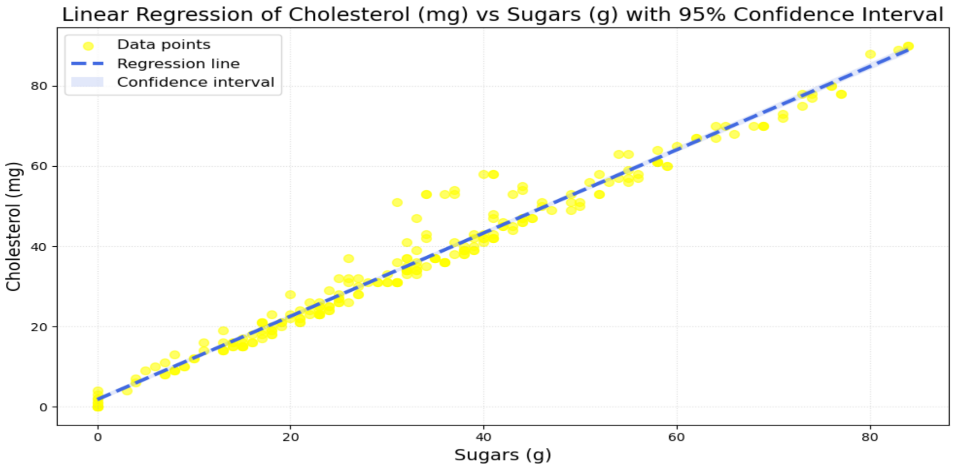


2 . Bar Chart



The bar chart, "Beverage Category Distribution," shows the count of beverages across many categories. The x-axis represents the categories of beverages, while the y-axis represents the number of beverages in each category. Classic Espresso Drinks has the highest count, at about 60, followed by Tazo Tea Drinks at ~55 and Signature Espresso Drinks at ~45. Frappuccino Blended Coffee is at ~40, and Shaken Iced Beverages are at ~30. Smoothies at ~20 and Coffee at ~15 are the least represented, indicating not as many offerings or they are less popular. The chart will give a proper overview of beverage distribution, useful in understanding consumer preferences or guiding product management decisions.

Linear Fit Analysis



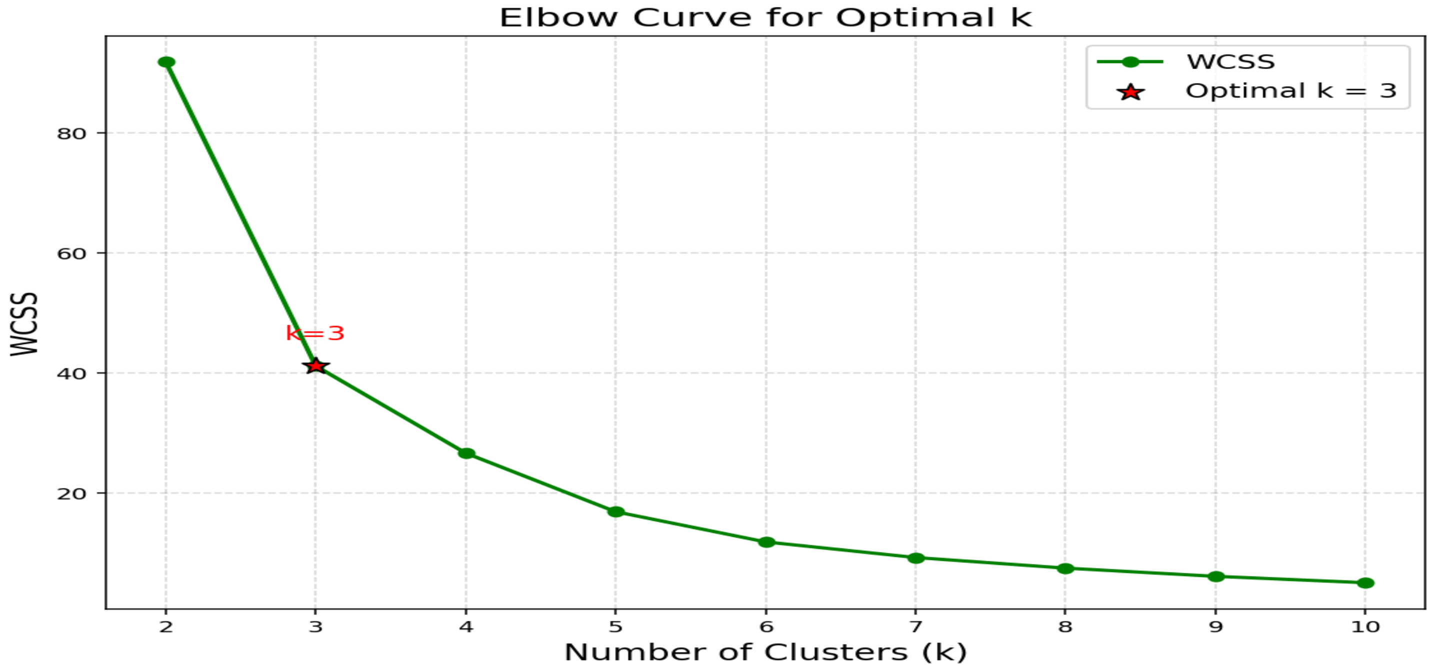
This scatter plot, "Linear Regression of Cholesterol (mg) vs Sugars (g) with 95% Confidence Interval," displays the relationship between sugar intake, x-axis, and cholesterol levels, y-axis. The yellow dots reflect individual data points, while the blue dashed line reflects the linear regression trend of the plot. From this plot, it can be seen that there is a positive linear relation between the two variables: the higher the intake of sugar, the higher the cholesterol levels. The 95% confidence interval of the regression is the shaded area around the regression line. It reflects the uncertainty of the regression estimate. This graph allows for the examination of the correlation between sugar intake and cholesterol levels, as well as the scatter of the data.

Clustering Analysis - Elbow Method and Silhouette Scores

The clustering evaluation combines insights from the **Elbow Method** and **Silhouette Scores** to determine the optimal number of clusters.

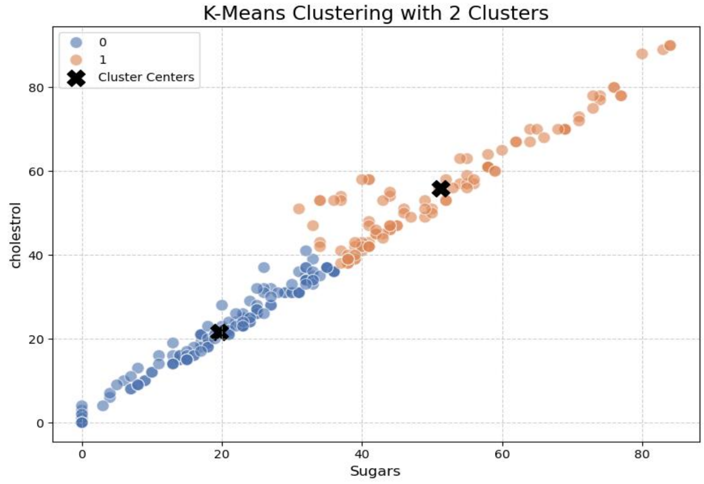
Elbow Method Insights:

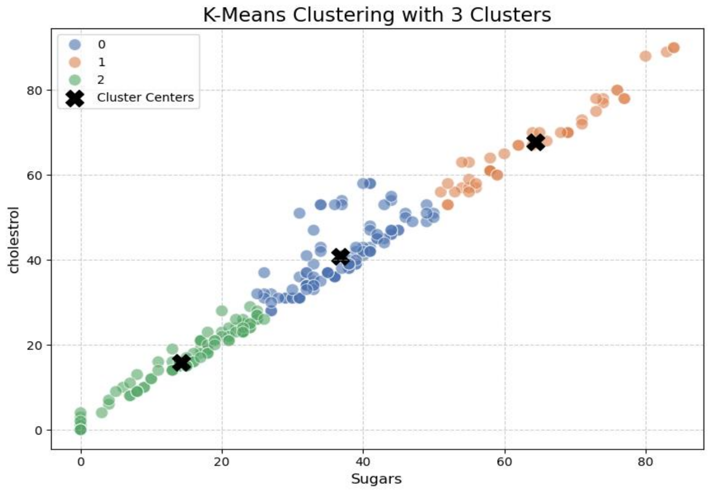
This picture shows an **Elbow Curve** for choosing the best number k in k-means clustering. On the **x-axis** is the number of clusters (k), from 2 to 10, and on the **y-axis** is the WCSS. The **curve** is plotted out of **green points and line**, representing the WCSS value for each k. A **red star** marks the point at k=3, labeled "k=3," showing the optimum number of clusters. Based on the curve, the significant drop in WCSS is from k=2 to k=3, and then a gradual drop beyond this k value, suggesting that at k=3, it is pretty optimal because further clustering beyond this gives very poor improvements. The legend shows that the **green line** represents WCSS values, while the **star in red** represents Optimal k. **Silhouette Score Insights**:


Elbow Curve for Optimal k

Best Cluster Configuration : The highest **3 clusters silhouette score = 0.56**, which indicates excellent cluster separation and cohesion.

K-Means Clustering Analysis


K-Means Clustering with 2 Clusters


K-Means Clustering with 3 Clusters

The scatter plots depict the K-Means clustering results for Sugars (x-axis) vs. Cholesterol (y-axis):

1. For (k=2): Data falls into two clusters—blue, for lower values of sugars and cholesterol, and orange for higher values, with black crosses at the centre of the clusters.

2. For (k=3): Data is divided into three clusters—green, blue, and orange. The trend of the data is diagonally upward, which shows that as sugars increase, so does cholesterol.

These plots illustrate how K-Means clustering separates data into classes based on these two features.

Conclusion:

The Starbucks beverage dataset shows strong relationships among the nutritional components, especially the perfect correlations among calories, sugars, and cholesterol. Clustering analysis based on sugar and cholesterol content identifies distinct groupings that provide insights into beverage profiles. This dataset offers valuable information for understanding nutritional trends and making informed decisions about beverage options.