Starbucks Data Analysis - Advance Insights

Problem Statement:

The objective of this project is to analyse the **nutritional content** of various beverages and identify key patterns and insights regarding **calories, fat, sugar, protein, and caffeine** across different beverage categories. This analysis will help understand the nutritional profiles of beverages, identify high-calorie or high-sugar drinks, and uncover relationships between nutritional metrics.

Data Set Used:

The dataset consists of multiple beverage types, each with key nutritional components such as:

- **Beverage_category**: The type or category of the beverage (e.g., Coffee, Smoothies, Frappuccino Blended Beverages).
- **Calories**: The number of calories in each beverage.
- **Total Fat (g)**: The fat content in grams.
- **Sugars (g)**: The amount of sugar in grams.
- **Protein (g)**: The protein content in grams.
- **Caffeine (mg)**: The amount of caffeine in milligrams.

Work Done:

1. Data Preprocessing:

Before diving into the analysis, the dataset was carefully pre-processed to ensure all columns were in the correct format and ready for visualization:

- **Data Cleaning**: Converted non-numeric values in critical columns like Caffeine (mg) and Calories into numeric data using pd.to_numeric(), while handling any non-numeric entries (like 'Varies') by converting them to NaN.
- **Handling Missing Data**: Rows containing NaN values were removed from columns where numeric values are essential, such as Calories and Caffeine (mg). This ensured that subsequent analyses and visualizations used complete and valid data.

By cleaning and standardizing the data, the groundwork was laid for accurate and insightful analysis.

2. Visualization of Beverage Categories:

A **count plot** was used to visualize the number of beverages within each category, such as **Coffee**, **Smoothies**, and **Frappuccino Blended Beverages**. This plot provided a quick view of which categories were most and least represented in the dataset.

- **Purpose**: This step helped identify how many drinks belong to each category, which is crucial in understanding the product variety and emphasis (e.g., whether coffee is the primary product or if other drink types are equally represented).

3. Relationship Between Calories and Caffeine:

Using a **regression plot**, we examined the relationship between **calories** and **caffeine content**. This helped assess whether higher caffeine content in a drink correlates with higher calories, and if any notable patterns exist between these two metrics.

- **Purpose**: Caffeine is an important factor for many consumers, and analysing how it relates to calories could provide valuable insights for those looking to balance their energy and calorie intake.

4. Average Calories per Beverage Category:

A **bar plot** was created to show the **average number of calories** for each beverage category, allowing us to compare how caloric different types of drinks are on average.

- **Purpose**: This helped determine which categories are more calorie-dense. For instance, categories like **Smoothies** and **Frappuccino Blended Beverages** were shown to have the highest average calories, indicating that these drinks could be high-calorie indulgences.

5. Pair Plot of Nutritional Metrics:

A **pair plot** was used to visualize relationships between key nutritional metrics, including **calories**, **fat**, **sugar**, and **protein**. This gave an overview of how these metrics interact with each other across the dataset.

- **Purpose**: This step was essential for uncovering correlations between metrics. For example, higher calories were strongly correlated with higher sugar and fat content, suggesting that the more indulgent drinks contain more fat and sugar.

6. Protein Content Distribution:

A **histogram** was used to analyse the **distribution of protein** across the beverages. This showed the frequency of protein content levels across all drink types, revealing that most drinks contain little or no protein.

- **Purpose**: Protein is an important macronutrient, and this step highlighted that most beverages (like coffee and tea) don't contribute significantly to daily protein intake. However, some categories, like smoothies, do contain notable amounts of protein.

7. Relationship Between Protein and Calories:

Another **regression plot** was created to explore the relationship between **protein content** and **calories**. This was used to see if protein-heavy drinks also tend to be high in calories, and whether a strong correlation exists between the two.

- **Purpose**: This visualization showed that drinks with higher protein content (e.g., protein shakes or smoothies) tend to have higher calorie counts. It helped in identifying protein-rich options that might also be higher in calories.

8. Correlation Heatmap of Nutritional Metrics:

A **correlation heatmap** was generated to visualize the **Pearson correlation coefficients** between various nutritional metrics, including **calories**, **fat**, **sugar**, **protein**, and **caffeine**.

- **Purpose**: The heatmap helped identify strong positive correlations, such as those between **calories** and **fat/sugar content**, highlighting the primary contributors to calorie counts in beverages. It also showed weak correlations with caffeine, indicating that caffeine content doesn't significantly impact the other nutritional metrics.

9. Sugar Distribution by Beverage Category:

Finally, a **box plot** was used to visualize the **distribution of sugar content** across various beverage categories. This helped identify categories with the highest sugar content, as well as the presence of outliers (drinks with exceptionally high or low sugar levels).

- **Purpose**: This step was crucial in pinpointing which beverages are most sugar-heavy. Categories like **Frappuccino Blended Beverages** and **Smoothies** had wide sugar distributions, with some drinks containing extremely high sugar levels. It also showed that some categories (like **coffee** and **tea**) generally contain little or no sugar.

Summary of Work Done:

Throughout the project, we employed a combination of **data cleaning**, **visualization techniques**, and **correlation analysis** to uncover key insights about the **nutritional content** of beverages. Each visualization played a critical role in highlighting relationships between calories, fat, sugar, protein, and caffeine, offering insights into which drinks are nutritionally dense and which offer lighter options.

By systematically working through these steps, we were able to:

- Understand the distribution of beverages across categories.
- Explore the relationships between key nutritional metrics.
- Highlight which drinks are calorie-dense or sugar-heavy.
- Identify the primary contributors to caloric content, including fat and sugar.

Each analysis provided specific insights that could guide both consumers and businesses in understanding the nutritional profile of beverages and making informed decisions.

1. Beverage Categories Count:

- **Observation**: The count plot revealed that **Frappuccino Blended Beverages** and **Coffee** dominate the dataset in terms of frequency. This suggests that these two categories are the most represented, which could be due to their popularity or a larger variety of options in these categories compared to others like **Smoothies** or **Tea**.
- **Detailed Insight**: This indicates that coffee-based drinks and blended beverages are a significant part of the beverage offerings. It also implies that the dataset may skew towards these high-caffeine or dessert-like beverages, which is important to consider when analysing the overall nutritional content of the dataset. The dominance of these categories might reflect consumer preferences or business priorities.

2. Calories vs. Caffeine:

- **Observation**: The **regression plot** showed that there is **no strong correlation** between **calories** and **caffeine content** in beverages. Drinks with high caffeine do not necessarily contain more calories, and vice versa.
- **Detailed Insight**: This finding suggests that caffeine content in beverages is more dependent on the type of drink (e.g., brewed coffee, espresso) than on calorie-dense ingredients like sugars, fats, or creams. For example, **high-caffeine drinks** like espresso-based beverages can still be low in calories if consumed without sweeteners or cream, while drinks like **Frappuccino's** can be high in calories but have varying caffeine levels. This dissociation between calories and caffeine highlights that caloric intake in beverages is largely driven by factors other than caffeine, such as sugar and fat content.

3. Average Calories per Beverage Category:

- **Observation**: **Frappuccino® Blended Beverages** and **Smoothies** have the highest average calories per beverage, while categories like **Tazo® Tea Drinks** and **Shaken Iced Beverages** have significantly lower calorie counts.
- **Detailed Insight**: This clearly indicates that **blended beverages** and **smoothies** tend to be more calorie-dense, likely due to added ingredients like **syrups, whipped cream, and full-fat dairy products**. On the other hand, categories such as **tea-based drinks** and **iced beverages** generally contain fewer calorie-rich additives, making them lighter options. This distinction is crucial for those looking to make healthier choices, as beverages like **Frappuccino's** can easily surpass daily calorie recommendations, especially when consumed in larger sizes or with additional toppings.

4. Pair Plot of Nutritional Metrics:

- **Observation**: The pair plot revealed a strong **positive correlation** between **calories, total fat, and sugars**. Drinks that are high in fat also tend to be high in sugar and calories.
- **Detailed Insight**: This observation highlights that beverages with high calorie counts are often loaded with both **sugars and fats**, making them more indulgent options. For instance, drinks that contain syrups, creams, and other rich ingredients tend to elevate both sugar and fat content. This is especially common in categories like **Frappuccino's** and **Smoothies**, where ingredients like whipped cream and flavored syrups are common. Conversely, categories like **black coffee** or **tea** have little to no fat and sugar, and thus, fewer calories. Understanding this correlation helps consumers identify which drinks might pack the most calories due to their sugar and fat content.

5. Protein Content Distribution:

- **Observation**: The **histogram** of protein content revealed that the vast majority of beverages contain **little to no protein**, with only a small subset of drinks (likely smoothies or milk-based beverages) having significant protein content.
- **Detailed Insight**: Protein is an important macronutrient, but the dataset shows that beverages, particularly **coffee-based drinks** and **teas**, are generally not a significant source of protein. Drinks with **high protein content** are likely limited to categories such as **smoothies**, where protein powders or milk products are added. This insight is crucial for consumers looking to boost their protein intake, as most beverages in this dataset are not contributing substantially to daily protein needs. Those who are looking for protein-rich options should focus on smoothies or specific drinks designed with added protein.

6. Relationship Between Protein and Calories:

- **Observation**: The **regression plot** between protein and calories showed a **positive correlation**, meaning that beverages with more protein tend to also have more calories.
- **Detailed Insight**: While protein is a healthy macronutrient, drinks that are **high in protein** (such as **protein shakes** or **milk-based smoothies**) also tend to have **higher calorie counts**. This is because these beverages often contain not just protein, but other calorie-dense ingredients like full-fat milk, yogurt, or added sugars. Therefore, while high-protein drinks can be good for individuals looking for nutritional benefits like muscle recovery or satiety, they also come with higher calorie content, which might not suit individuals looking to cut down on calories.

7. Correlation Heatmap of Nutritional Metrics:

- **Observation**: The **correlation heatmap** showed strong **positive correlations** between **calories, total fat, and sugars** (correlation values around 0.9), while **caffeine** had little to no correlation with these metrics.
- **Detailed Insight**: This finding reinforces the observation that **caloric content** in beverages is primarily driven by **fat** and **sugar**, as these are the main contributors to calorie increases. Caffeine, on the other hand, is mostly independent of these nutritional factors, meaning a drink's caffeine content doesn't influence its fat or sugar levels. For example, a black coffee may be high in caffeine but low in calories, while a high-calorie Frappuccino may contain relatively low caffeine. The strong correlation between calories, fat, and sugar indicates that reducing sugar or fat is key to lowering calorie intake from beverages.

8. Sugar Distribution by Beverage Category:

- **Observation**: The **box plot** revealed that **Frappuccino Blended Beverages** and **Smoothies** had the highest sugar content, with some drinks containing exceptionally high levels of sugar (outliers), while categories like **Coffee** and **Tea** had almost no sugar content.
- **Detailed Insight**: This observation highlights the stark contrast in sugar content between categories. While **blended beverages** and **smoothies** can contain **more than 50 grams of sugar** per serving, drinks like **black coffee** or **tea** often have little to no added sugar. This is a crucial finding for consumers mindful of sugar intake, as regularly consuming high-sugar beverages like **Frappuccinos** can significantly contribute to daily sugar intake, leading to potential health concerns. The outliers in the **Frappuccino** category suggest that certain customized versions of these drinks can be even more sugar-laden, particularly when topped with syrups, whipped cream, or sweeteners.

Conclusion:

The analysis reveals that beverages with higher calorie content generally contain more fat and sugar, with categories like **Frappuccino® Blended Beverages** and **Smoothies** being the most calorically dense. **Caffeine** content, however, is not significantly related to other nutritional metrics. Furthermore, protein is low across most beverages, with only a few exceptions. The visualizations, particularly the **correlation heatmap** and **pair plots**, help in identifying how different nutritional components interact, providing actionable insights for both consumers and businesses aiming to understand the nutritional profiles of beverages.