Team Blue - Anna Grigoryan, Sean Jeffry, Andrew Fletcher, Jayden Murata

Dataset: https://www.key2stats.com/data-set/view/1604
We will be using the Nutrition_facts_for_Starbucks_Menu_1604_26.csv file. It has 242 rows and 18 columns of data. There are three non-numerical variables (names and categories of drinks) and fifteen numerical variables, each a different nutritional value.

Business Question: How can Blue Café navigate the evolving consumer preferences for healthier and more nutritionally conscious beverage options?

Why: As consumer preferences continue to evolve into a more health-conscious nature, having healthier options available will be beneficial to our business to provide our customers with better drink options and increase revenue growth for our brand.

Questions:

1. **(Supervised Model)** Looking at the coefficients, which independent variables [milk_choice, beverage_size, Total Fat (g), Trans Fat (g), Saturated Fat (g), Sodium (mg), Total Carbohydrates (g), Cholesterol (mg), Dietary Fibre (g), Sugars (g), Protein (g), Caffeine (mg)] have the strongest relationship on the dependent variable, calories?

2. **(Supervised Model)** Looking at the coefficients, which variables (Vitamin A (% DV) , Vitamin C (% DV), 'Calcium (% DV) , Iron (% DV)) have the strongest relationship on sugars(g)?

3. **(Clustering)** When considering the percentage of sugar (g), sodium (mg), protein (g), trans fat (g), what clusters emerge and what characterizes those clusters?

4. **(Dimensionality Reduction)** When comparing a model using Principal Component Analysis (PCA) on all the continuous variables in the dataset and retaining enough PCs to keep 90% of the variance, to a model using all the continuous variables, how much of a difference is there in mean absolute error when predicting calories?