DATA DESCRIPTION, MODELLING, AND ANALYSIS PLAN

Visualise the energy density of food purchases across LSOAs in winter (October-April) and in summer (May-September).

a. Is there a difference between energy density according to season?
b. What about the ratio of sugar to fibre?

DATA DESCRIPTION:

The Tesco food 1.0 dataset is a comprehensive compilation of food purchase records from January through December 2015 from Lower Super Output Areas (LSOAs) in Greater London. Researchers may examine how food consumption patterns vary with the seasons thanks to this dataset, which provides a month-by-month breakdown of purchases. Since Tesco Clubcard transactions are the source of the data, only purchases made with loyalty cards have been documented. Although this guarantees organized and trustworthy data, it does not take into consideration transactions done without a Clubcard, which could lead to some representational gaps.

The dataset offers a wealth of nutritional information for each LSOA, such as the distribution of food categories, total calorie consumption, and the composition of macronutrients (carbohydrates, proteins, fats, sugar, and fibre). This dataset provides a powerful predictor of food buying patterns over time and location, especially considering Tesco's dominance in the UK grocery sector. The monthly availability of the data makes it possible to compare the winter (October–April) and summer (May–September) seasons, which aids in identifying seasonal changes in eating patterns.

The 202 columns in each month's dataset cover important topics such transaction counts, total products bought, calorie density, nutrient composition, and proportions of food categories. Although individual transactions are aggregated at the LSOA level, no personally identifying information is made public because the data has been processed to provide privacy protection. The dataset also has a representativeness score, which aids in determining how well it captures general patterns in food consumption across various regions.

The information does not give a comprehensive picture of food consumption because it only includes Tesco Clubcard purchases; purchases made at other supermarkets, local businesses, and restaurants are not included. Nevertheless, in spite of these drawbacks, it continues to be among the most extensive and extensive datasets accessible for researching London food buying habits. Research on nutrition trends, public health, economic influences, and cultural variables influencing dietary choices can all benefit greatly from it.

DATA PROCESSING PLAN:

The datasets must then be loaded and combined into two consolidated dataframes, df_winter and df_summer. This guarantees that the data is organized to make seasonal analysis easier. The number of rows will increase during the merging process, but the column layout will remain unchanged because each monthly file has 202 columns with a constant structure. The data will be examined for missing values after it has been combined. To preserve data integrity and prevent inconsistencies in the analysis, any rows containing null values will be eliminated.

Not every column is required for the actual analysis. The goals of the study will serve as a guide for choosing the columns. Only the energy_density column will be retained if we are examining energy density; only the sugar and fibre columns will be retained if we are analyzing the sugar-to-fibre ratio. The area_id column, which gives each LSOA a unique identity, will also be kept. To increase computational performance and streamline the dataset, all other columns will be eliminated. A brief validation step will be completed prior to proceeding with the final analysis. This entails examining the dataset's shape to verify the anticipated number of rows and columns, then using descriptive statistics to make sure all required modifications have been carried out accurately. Following verification, we will go on to the following stage, which involves putting the analysis and visualization into practice.

DATA IMPLEMENTATION PLAN

The workflow must then be put into action after the data processing strategy has been established. The read_csv() function in pandas will be used to load the CSV files first. This will be done in a loop to ensure efficiency because we are working with numerous files. To facilitate trend comparison across time periods, all files will be concatenated into two seasonal datasets (df_winter and df_summer) after loading.

df.isnull().sum() will be used to identify missing values after the data has been merged. Depending on the kind and distribution of the data, any null values will either be handled using the proper imputation techniques or eliminated using dropna(). Here, the objective is to prevent errors from being introduced by incomplete records in the final analysis.

Feature transformation and selection follow. Only the essential columns—energy_density, sugar, fibre, and area_id—will be kept, depending on the research topic. At this point, more changes, such figuring out the sugar-to-fibre ratio, will be carried out if necessary. This guarantees that there is no needless overhead and that the dataset is optimized for addressing the research topics.

Finally, a brief validation phase will be carried out before to delving into analysis. Data distribution will be understood through the generation of descriptive statistics , and anomalies will be found using

visualizations like boxplots and histograms. We can move forward with additional investigation, visualization, and insight extraction once everything appears to be in order.

DATA ANALYSIS AND VISUALISATION PLAN:

In this study, we are analysing how energy density and nutritional composition, particularly the sugar-to-fibre ratio, change with seasons. The goal is to see whether there is any significant difference between winter (October-April) and summer (May-September) in terms of food consumption patterns. To get a clear understanding, we will use statistical tests and different types of visualizations to interpret the data effectively.

To begin with, we will first look at descriptive statistics for key variables like energy density, sugar, and fibre. By calculating values like mean, median, standard deviation, and quartiles, we can get an overall idea of the distribution of these variables. This will help us spot patterns, outliers, and variations in the dataset.

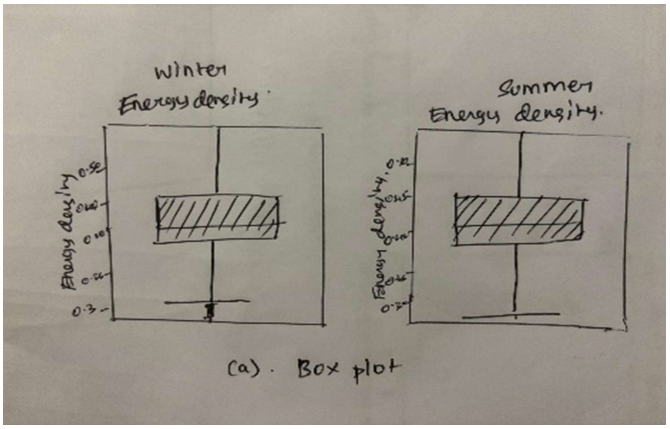
Next, we will perform a T-test to check whether there is a significant difference in energy density between winter and summer. This is important because if energy density is higher in one season compared to another, it could indicate changes in food consumption habits. A two-sample independent T-test will be conducted, and we will check the p-value to interpret the results. If the p-value is less than 0.05, it means that the difference in energy density between the two seasons is statistically significant.

Similarly, another T-test will be performed to compare the sugar-to-fibre ratio between winter and summer. Since fibre is an essential part of a healthy diet, and sugar intake needs to be controlled, checking the balance between these two components is crucial. If the p-value is low (<0.05), it would mean that there is a significant seasonal variation in the sugar-to-fibre ratio. If not, then we can assume that the ratio remains roughly the same throughout the year.

Once the statistical analysis is complete, we will use visualizations to make our findings more understandable. Various types of plots will be created to highlight seasonal trends, relationships, and distributions in the dataset.

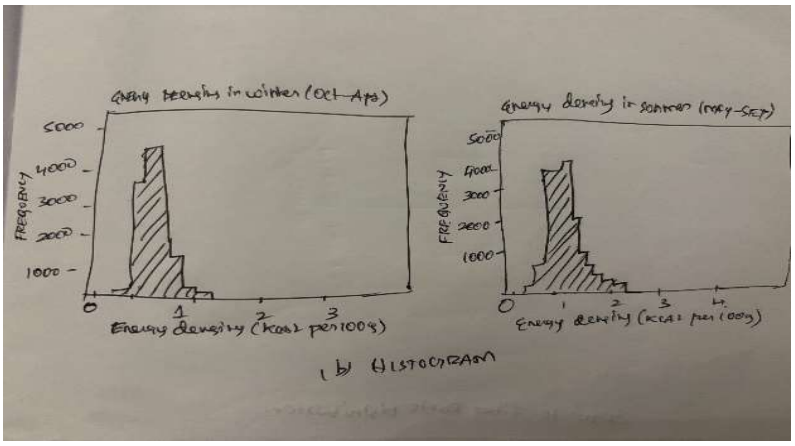
BOX PLOT –

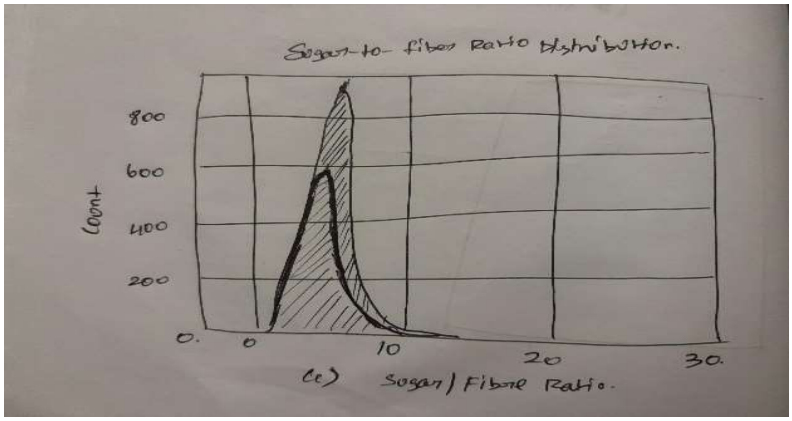
This will compare the distribution of energy density and sugar-to-fibre ratio between winter and summer. Box plots are useful in detecting the outliers and understanding the spectrum of data.



HISTOGRAM –

To check whether energy density follows a normal distribution, a histogram will be plotted. This will help validate the results of the normality test. We are preparing two histograms: one for energy density in winter and summer, and another with a KDE plot overlaid on a histogram for the sugar-to-fibre distribution





With this method, we will conduct an in-depth statistical analysis while also presenting the results using insightful visualizations. By integrating T-tests, correlation analysis, normality checks, and various graphical representations, we can evaluate whether seasonal variations significantly impact dietary patterns. This approach will offer a clear understanding of how food consumption habits fluctuate throughout the year.