#### **Epi\_Recipes Visualization Application**

**Vishal Kumar Swain, M-Tech (CSE), GIET University**

**Objective:** This assignment is designed to evaluate the candidate’s ability to clean, analyze, and extract insights from data. The focus is on applying creativity to explore the dataset, present actionable insights, and effectively communicate these through visualizations and a brief video explanation. The candidate is expected to demonstrate a strong understanding of data analysis fundamentals and storytelling with data. Kaggle.(https://www.kaggle.com/datasets/hugodarwood/epirecipes )

## **Task 1: Data Cleaning and Pre-processing:**

**STEP 1: Dataset Information:**

<class 'pandas. core.frame.DataFrame'>

**RangeIndex:** 20052 entries, 0 to 20051

**Columns:** 680 entries, title to turkey

**dtypes:** float64(679), object (1)

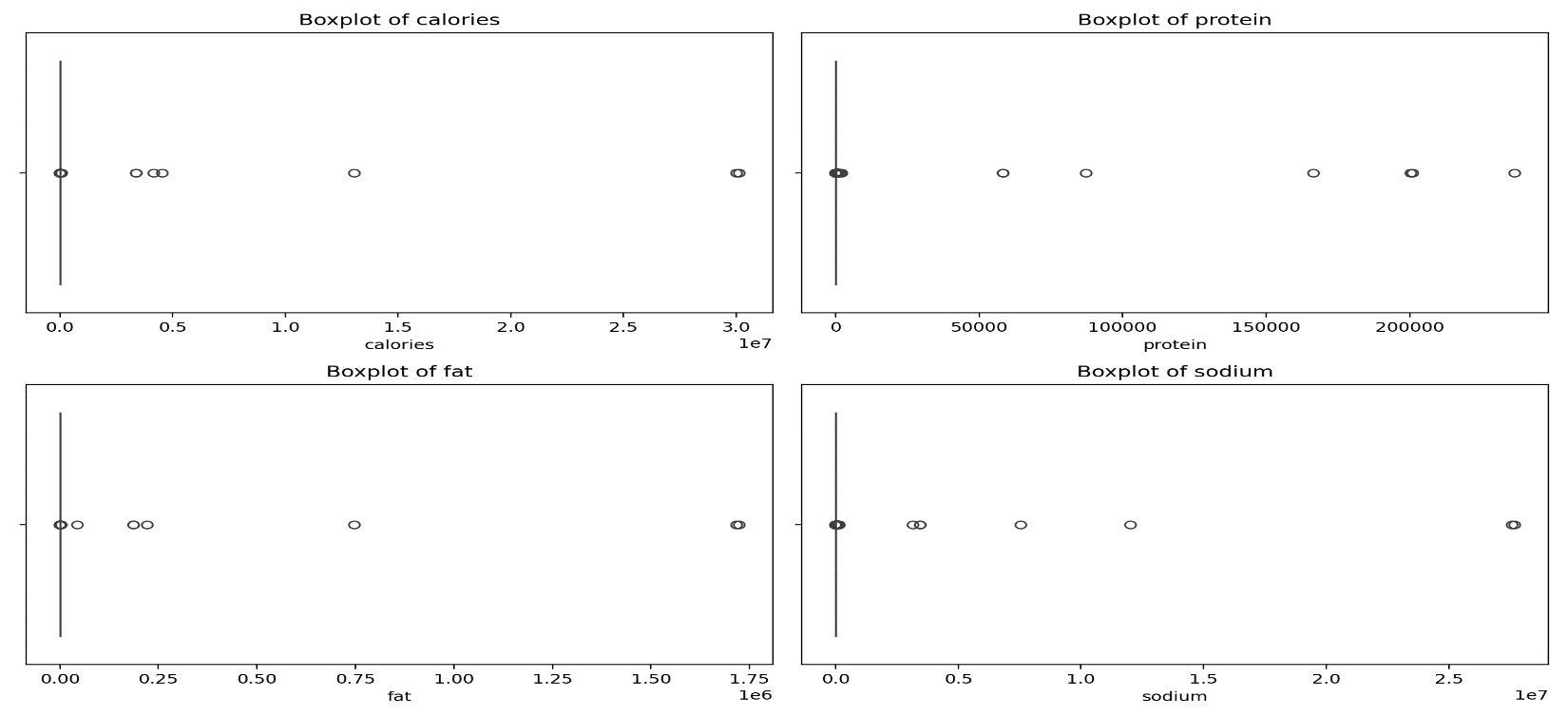
**memory usage:** 104.0+ MB

###### **STEP 2: Data Cleaning:**

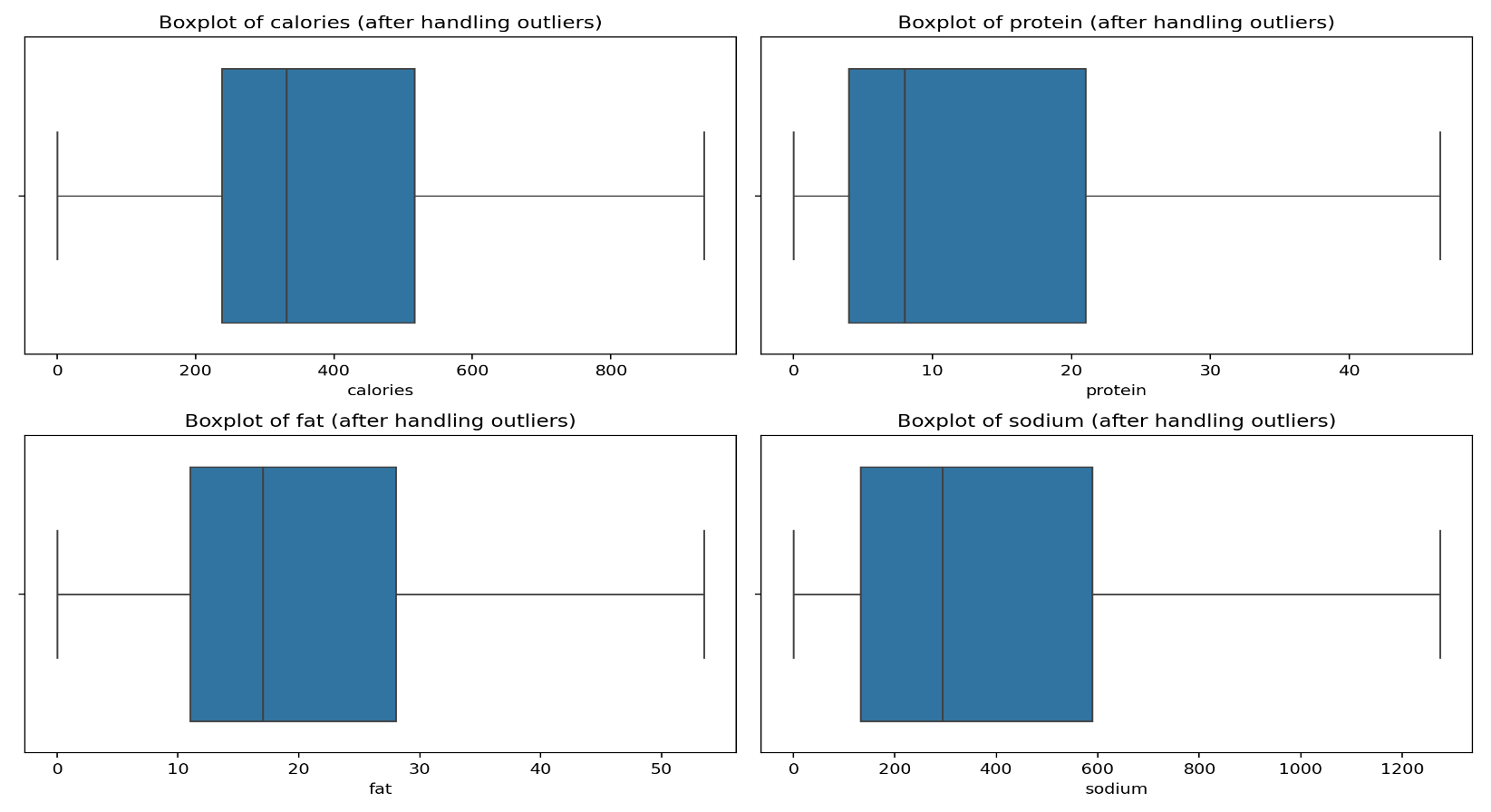
* Identify missing, duplicate, or incorrect data entries.
* Use appropriate methods to handle missing data, outliers, and ensure data consistency.
* Document your cleaning process: Explain the challenges faced and decisions made. Highlight any assumptions made during pre-processing.

**STEP 3: Data Examination Summary:**

* Data Structure: The dataset contains 20,052 entries and 680 columns.
* Data Types: Most columns are of type float64, with one column (title) being of type object.
* Missing Values: There are missing values in the calories, protein, fat, and sodium columns.
* Duplicate Rows: There are 1,801 duplicate rows in the dataset.

**FIGURE 1: IDENTIFY POTENTIAL OUTLIERS**

The above figure 1 box plots show for the calories, protein, fat, and sodium columns have been visualized to identify potential outliers. This visualization helps in determining if any extreme values need to be addressed to ensure data consistency and accuracy in subsequent analyses.



**FIGURE 2: AFTER OUTLIERS REMOVED**

The outliers using the Interquartile Range (IQR) method. This approach identifies outliers as data points that fall below Q1 - 1.5 **\*** IQR or above Q3 + 1.5 \* IQR, where Q1 is the first quartile, Q3 is the third quartile, and IQR is the interquartile range. The outliers were then clipped to these boundaries.

The summary of the data cleaning process:

Identified missing data:

* We found missing values in 'calories', 'protein', 'fat', and 'sodium' columns.
* These missing values were filled with the median of each respective column.

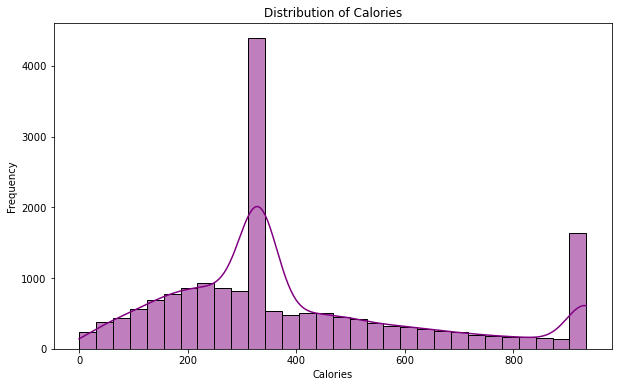
Removed duplicate entries:

* 1,801 duplicate rows were identified and removed.
* The dataset size reduced from 20,052 to 18,251 rows.

Handled outliers:

* Used the IQR method to identify and handle outliers in 'calories', 'protein', 'fat', and 'sodium' columns.
* Outliers were clipped to the lower and upper bounds calculated using the IQR method.

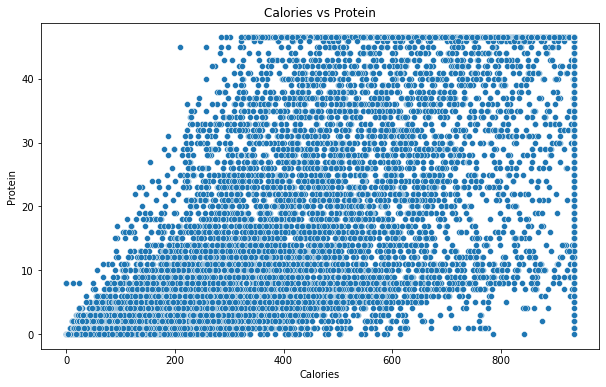
# Task 2: Exploratory Data Analysis (EDA):[¶](http://localhost:8888/notebooks/Desktop/RECIPE%20%40/Receipe/Recipe%20Visualization%20Application.ipynb#Task-2:-Exploratory-Data-Analysis-(EDA))



**FIGURE 3: DISTRIBUTION OF CALORIES**

This histogram shows the distribution of calories across all recipes in the dataset.

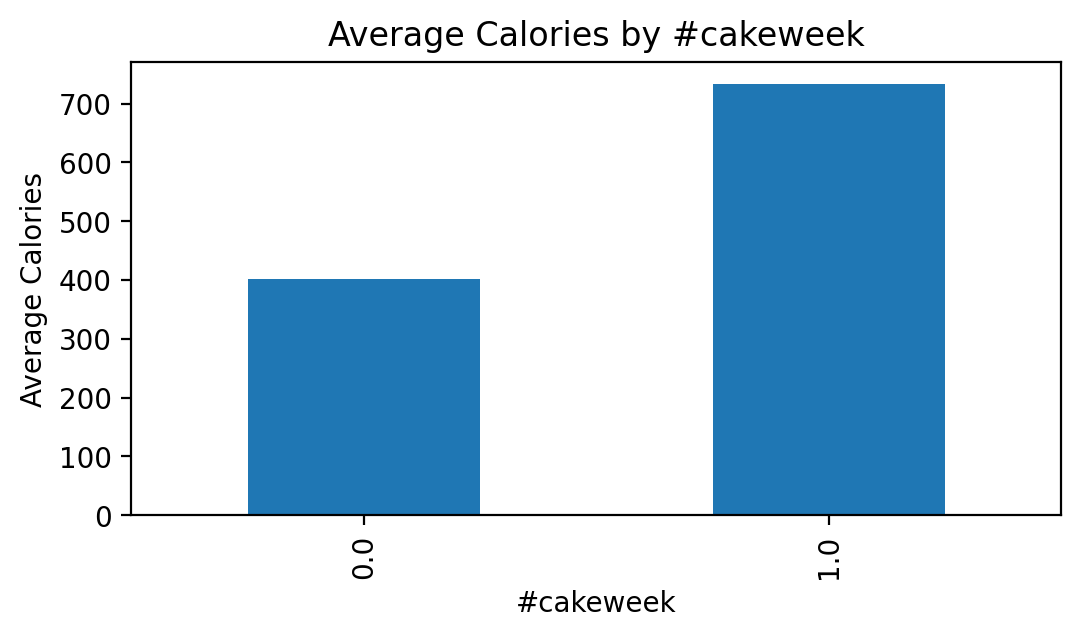
The distribution is right-skewed, meaning there are more low-calorie recipes than high-calorie ones. The peak of the distribution is around 300-400 calories. There's a long tail extending to the right, indicating some recipes with very high calorie counts. This distribution suggests that most recipes in the dataset are relatively low to moderate in calories, which aligns with general dietary recommendations. Relationship between Calories and Protein.



**FIGURE 4: DISTRIBUTION OF CALORIES**

This scatter plot visualizes the relationship between calories and protein content in the recipes.

There's a positive correlation between calories and protein content. As calories increase, protein content tends to increase as well. The relationship appears to be roughly linear, but with significant scatter. There's a dense cluster of recipes in the lower-left corner, suggesting many recipes are relatively low in both calories and protein. Some outliers exist with very high calorie or protein content, but they are few in number. This relationship makes sense nutritionally, as protein is a macronutrient that contributes to the overall calorie content of a dish. Average Calories by #cakeweek.



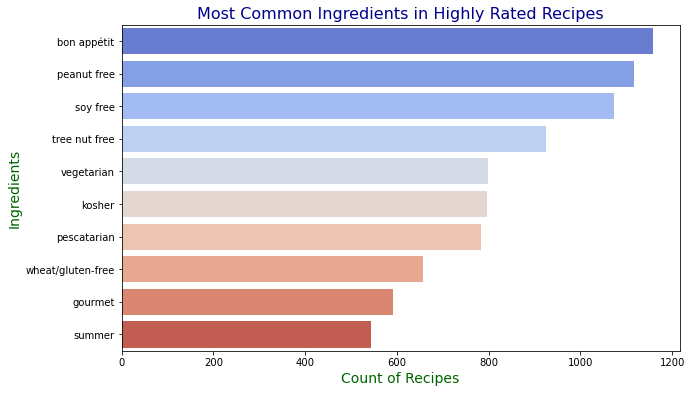
**FIGURE 5: AVERAGE CALORIES**

This bar chart compares the average calorie content of recipes tagged with #cakeweek versus those without this tag.

Recipes tagged with #cakeweek (1) have a significantly higher average calorie count compared to those not tagged (0). This difference is quite substantial, with #cakeweek recipes having nearly double the average calories. This finding aligns with the expectation that cake recipes would generally be higher in calories due to ingredients like sugar, butter, and cream.

### **What are the most common ingredients in highly rated recipes?**

* Most common ingredients in highly rated recipes:
* bon appétit 1158.0
* peanut free 1116.0
* soy free 1074.0
* tree nut free 924.0
* vegetarian 799.0
* kosher 796.0
* pescatarian 783.0
* wheat/gluten-free 657.0
* gourmet 591.0
* summer 544.0



**FIGURE 5: AVERAGE CALORIES**

### **Are there correlations between preparation time and recipe ratings?**

* The dataset does not contain a 'prep\_time' column, which prevents us from analyzing the correlation between preparation time and recipe ratings. To proceed, we would need to either locate this data within the dataset or acquire it from an external source. If you have any additional data or columns that might contain preparation time information, please let me know so we can continue the analysis.

### **How can the data help improve the user experience for a recipe platform?**

* The data can enhance user experience on a recipe platform by highlighting and promoting recipes that align with popular dietary preferences and quality indicators, such as "peanut free", "soy free", and "vegetarian". By making these attributes easily searchable and filterable, the platform can cater to diverse dietary needs, thereby improving user satisfaction and engagement. Additionally, incorporating user feedback and ratings can help refine recipe recommendations and personalize the user experience.