Epicurious Recipes

By Kamini Sharma

Challenges deep-dive

Task 1

Data Cleaning and Preprocessing

Handling Missing Values: Identify missing data and handle it by either removing rows/columns or imputing values using methods like mean, median, or mode.

Removing Duplicates: Detect and remove duplicate entries to avoid redundant data that can skew analysis.

Task 2

Exploratory Data Analysis (EDA)

Descriptive Statistics: Use functions like df.describe() to get an overview of the dataset, including metrics like mean, median, min, max, and standard deviation for numerical columns (e.g., calories, fat, protein, etc.).

Missing Value Analysis: Identify missing values using df.isnull().sum() to assess the completeness of the data and determine the appropriate handling method.

Task 1 : Data Cleaning and Preprocessing

Problem Statement

I am tasked with analyzing a recipe dataset to understand the nutritional content (such as calories, sodium, fat, and protein) and how it varies across different recipes. However, the dataset contains missing values in key columns like calories, protein, fat, and sodium, which may hinder accurate analysis and insights.

My approach is focusing on the calories, sodium, fat, and dessert columns to explore the nutritional content of various recipes. The dataset has missing values in several key columns like calories, protein, fat, and sodium, which could skew the analysis if not handled properly.

Solution

- **Inspect Data Shape**: First, check the shape of both datasets to understand the size of the data, ensuring that they are structured correctly and confirming the number of entries and columns available.
- **Identify Missing Values**: Use `isnull().sum()` to check for missing values in the selected columns. Missing values could potentially affect the accuracy of your analysis.
- **Drop Rows with Missing Values**: If necessary, you can drop rows with missing values to create a cleaner dataset for specific analyses that cannot handle missing data.
- **Impute Missing Values**: For certain columns like `calories`, `protein`, `fat`, and `sodium`, instead of dropping rows, you can fill missing values with the mean value of that column. This ensures that you don't lose too many data points while making the dataset usable for analysis.

Task 2: Exploratory Data Analysis

Solution

Data Overview:

Inspect Data Types and Structure: Use df.info() to check the types of each column (e.g., numerical, categorical) and identify the presence of missing values.
Basic Statistics: Use df.describe() to generate summary statistics (mean, median, standard deviation) for numerical columns like calories, sodium, fat, and protein.

Missing Value Analysis:

Visualizing Missing Data: Identify the extent of missing values across the dataset using heatmaps or simple summaries (df.isnull().sum()).
Handling Missing Values: Decide on the strategy to handle missing data (e.g., dropping rows with missing values or imputing them using mean/median).

Multivariate Analysis:

Correlation Matrix: Compute and visualize correlations between key variables like `calories`, `protein`, `fat`, and `sodium` to see if there are strong relationships between these nutritional factors.
Pairplots: Use pair plots to visualize pairwise relationships between multiple variables.

Categorical Data Analysis:

Bar Plots for Categorical Variables: Visualize the distribution of categorical columns like `dessert` to see the breakdown of recipes by category.

Outlier Detection:

Box Plots: Use box plots to detect outliers in variables like `calories`, `fat`, or `sodium`. This helps in understanding any extreme values that might affect the analysis.

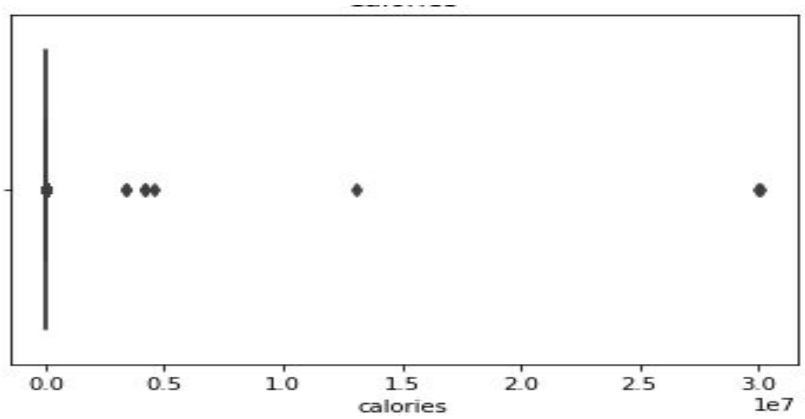
Analysis Based on the Boxplot of Calories and Sodium

Calories:

- The boxplot for calories shows significant outliers. Most of the data points are concentrated near the lower end of the scale, with a few extreme outliers extending to very high values.
- The outliers represent recipes with extremely high-calorie counts, far beyond the typical range. These outliers could skew the overall analysis if not handled properly.

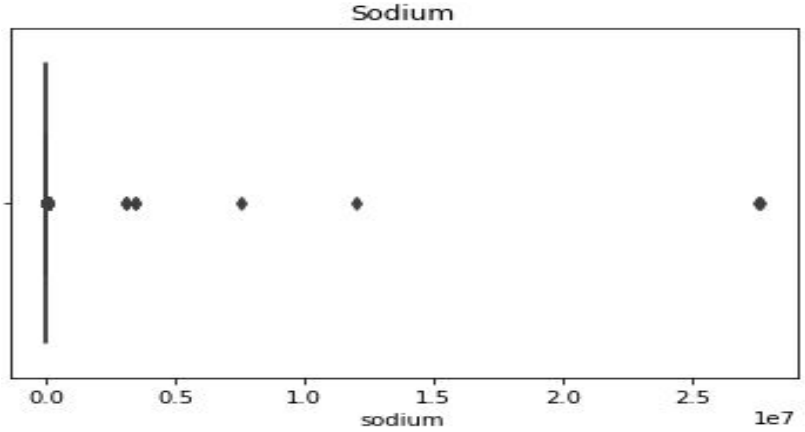
Sodium:

- Similar to the calorie plot, the sodium boxplot reveals a concentration of data near the lower end, with a few extreme outliers representing recipes with very high sodium levels.
- The presence of outliers could indicate the need for further investigation into the recipes with unusually high sodium content to ensure these values are correct or appropriately handled in the analysis.




```
sns.boxplot(x=epicurious_calories_sodium_df.sodium)
```

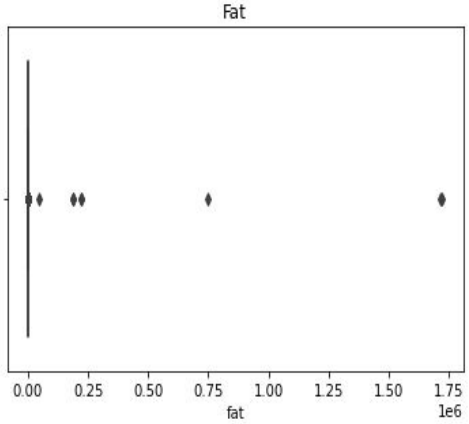
```
Text(0.5, 1.0, 'Sodium')
```



Analysis Based on the Boxplot of Fat and Dessert

Fat:

- The boxplot for fat content shows a similar pattern to the previous plots, where most of the data points are concentrated near the lower end of the scale, with several extreme outliers.
- These high-fat outliers could represent recipes that include large amounts of oils, butter, or other fat-rich ingredients, which may require special attention in health-related analyses.

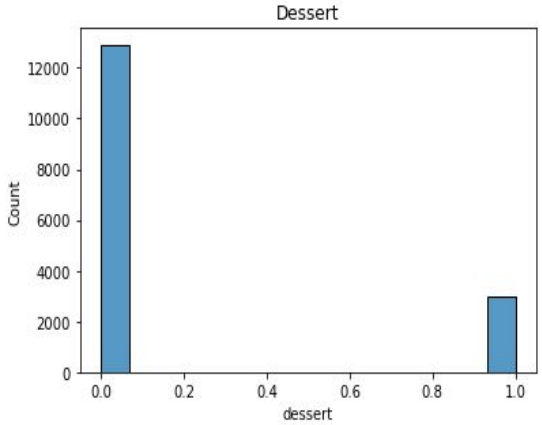


Dessert:

- The majority of recipes (about 12,000) are non-dessert items (`dessert = 0`), while a smaller portion (around 2,000 recipes) are classified as desserts (`dessert = 1`).
- This indicates that the dataset contains a wide range of recipe types, but non-dessert recipes make up the bulk of the data.

```
[220…  sns.histplot(x=epicurious_calories_sodium_df.dessert).set_title("Dessert")
```

```
[220…  Text(0.5, 1.0, 'Dessert')
```



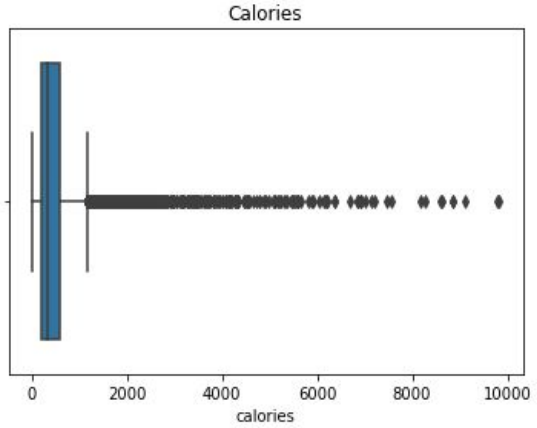
Analysis Based on the Boxplot of Calories and Sodium

Calories:

- The boxplot for calories shows a high concentration of data points towards the lower end, with many outliers extending up to approximately 10,000 calories.
- The majority of recipes have a calorie count below 2,000, while a small number of extreme outliers are significantly higher. These outliers may represent calorie-dense meals, possibly desserts or large serving meals.

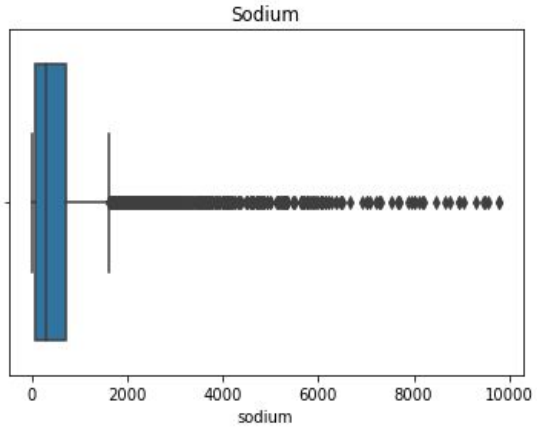
Sodium:

- The sodium boxplot similarly shows a large number of outliers, with sodium levels extending up to nearly 10,000 mg.
- Most of the data points are clustered around the lower end (under 2,000 mg), indicating that while some recipes contain moderate amounts of sodium, there are quite a few recipes that contain extremely high sodium levels, likely preserved foods or heavily salted dishes.




```
18…  sns.boxplot(x=epicurious_calories_sodium_df.sodium).set_title("Sodium")
```

```
18…  Text(0.5, 1.0, 'Sodium')
```



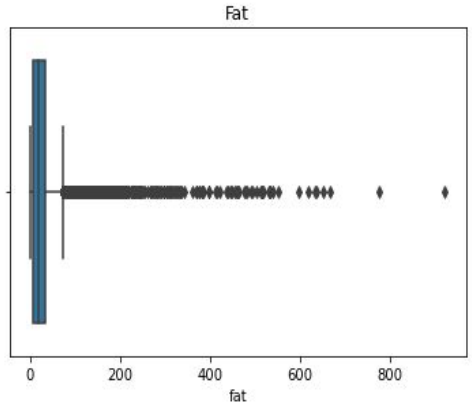
Analysis Based on the Boxplot for Fat and Histogram for Dessert:

Fat:

- The boxplot for fat content shows that the majority of recipes have low to moderate fat content, with values concentrated below 100 grams of fat.
- However, there are several extreme outliers, with fat content exceeding 600 grams in some cases. These outliers likely represent recipes with high amounts of oils, butter, or other fat-heavy ingredients.

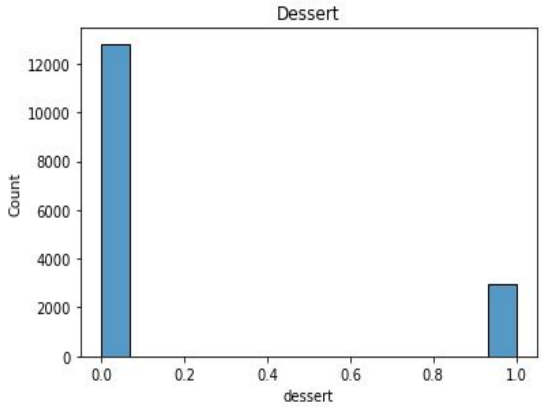
Dessert:

- The histogram for the `dessert` variable shows that most recipes in the dataset are non-desserts (`dessert = 0`), while a smaller number of recipes (around 2,000) are categorized as desserts (`dessert = 1`).
- This suggests that the dataset is skewed towards savory or non-dessert recipes, with only a limited number of dessert recipes included.




```
[420…  sns.histplot(x=epicurious_calories_sodium_df.dessert).set_title("Dessert")
```

```
[420…  Text(0.5, 1.0, 'Dessert')
```



Analysis Based on the Line Graph of the Number of Recipes by Year:

Spike in 2005:

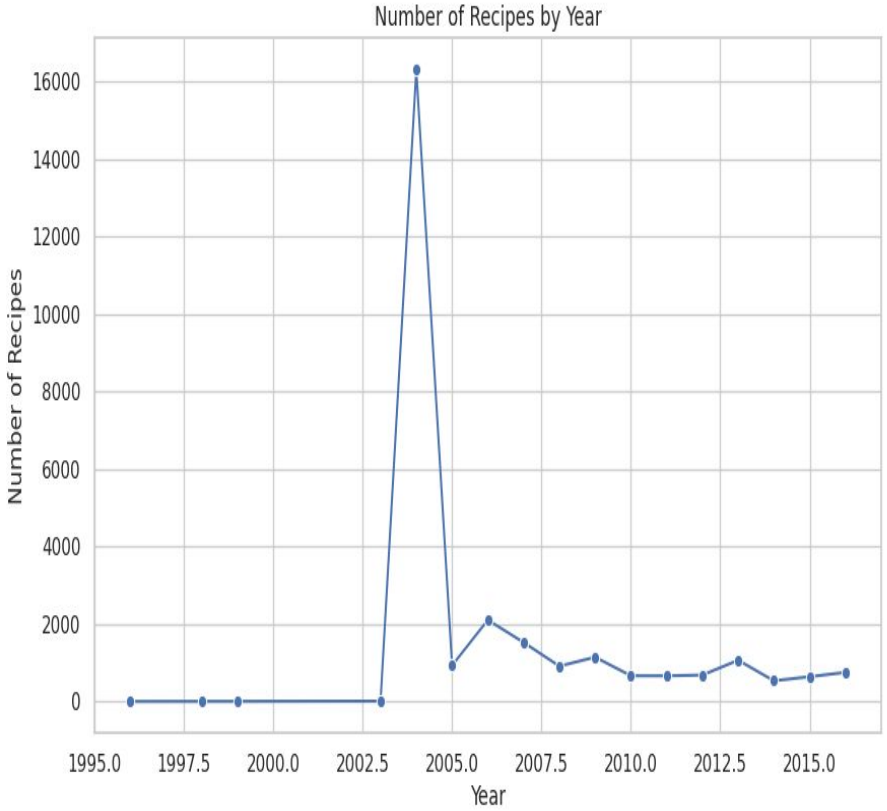
- There is a dramatic spike in the number of recipes published around 2005, with over 16,000 recipes added in that year alone,indicates a special event or campaign, such as a major content push, recipe collection, that led to a significant surge in recipe additions.

Decline After 2005:

- From 2006 onwards, the number of recipes per year is consistently below 2,000, showing a steady but minimal level of recipe entries.

Stable Activity After 2006:

- After the decline in 2005, the number of recipes stabilizes around 1,000 to 2,000 per year. This suggests a more consistent and moderate level of recipe publishing activity in subsequent years.



Analysis Based on the Bar Chart of Ratings:

Highly Rated Recipes:

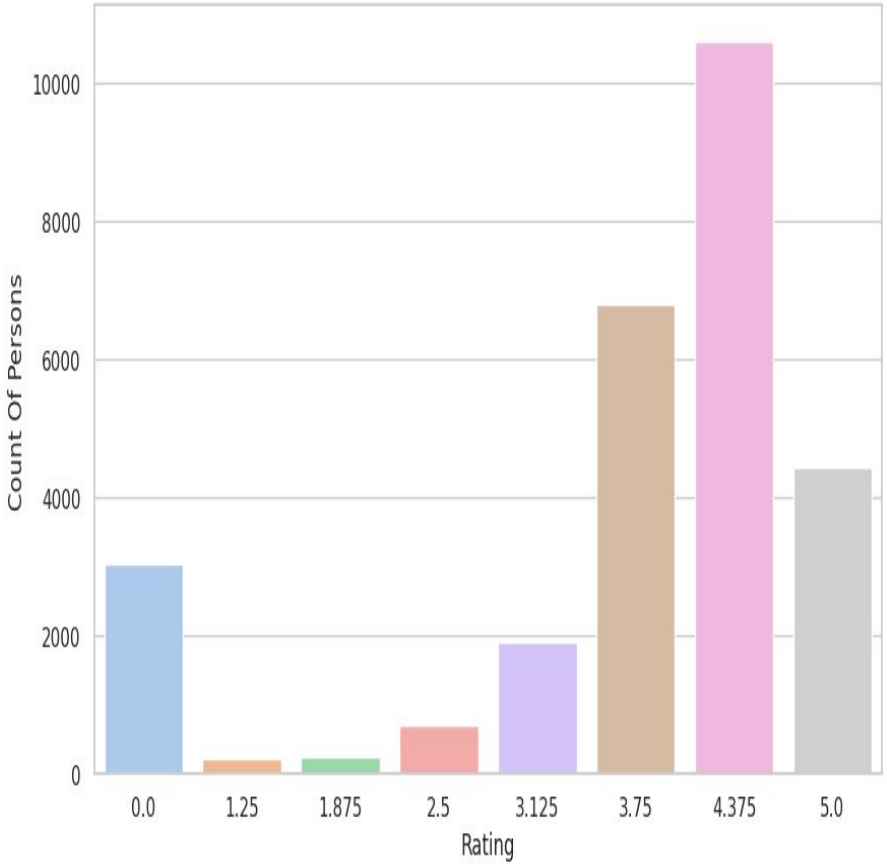
- The majority of recipes have high ratings, with the most frequent rating being **4.375**. Over 10,000 recipes fall into this category, indicating that a significant portion of the dataset is well-received by users.

Low and No Ratings:

- A considerable number of recipes (around 4,000) have a rating of **0.0**, indicating that these recipes might not have received any ratings or feedback from users.

Rating Distribution:

- The distribution of ratings is heavily skewed towards the higher end, with the majority of recipes rated **3.75** and above. This suggests that either users tend to rate recipes favorably, or that the recipes themselves are generally of good quality.



Analysis Based on the Bar Chart of Recipes by Diet Types:

Healthy Recipes Dominate:

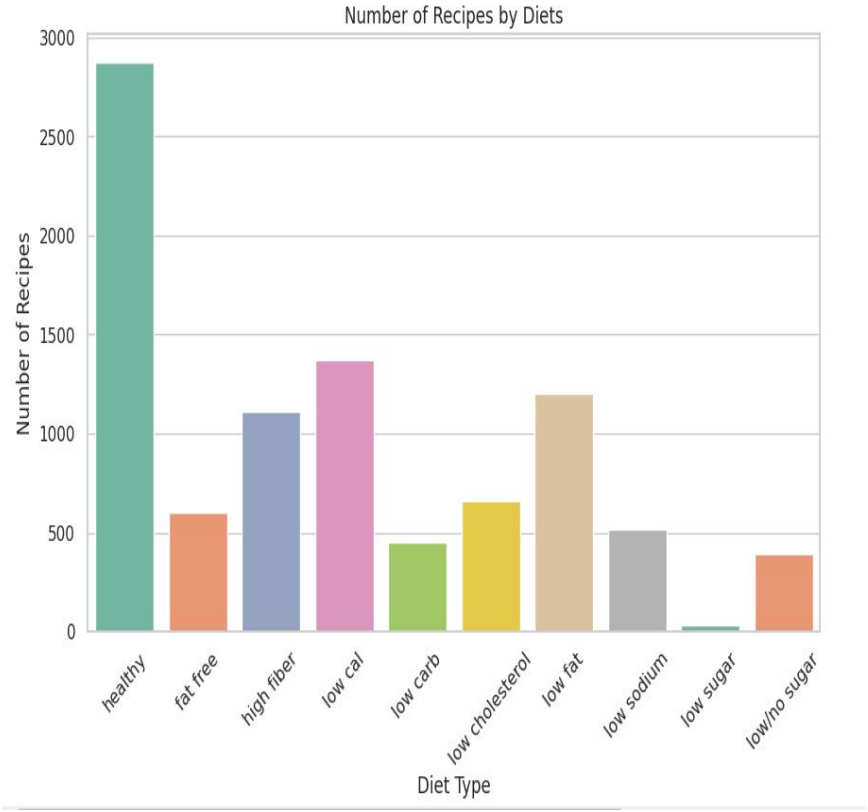
- The largest category by far is **healthy** recipes, with nearly 3,000 recipes. This suggests that there is a strong focus on recipes designed for general health-conscious consumers.

Low Calorie and Low Fat:

- **Low calorie** and **low fat** recipes are also popular, with over 1,500 and 1,800 recipes respectively. These diet types are likely targeting users who are focused on weight management or seeking to reduce their calorie and fat intake for health reasons.

Lower Presence of Low Sodium and Low/No Sugar Recipes:

- **Low sodium** and **low/no sugar** recipes are much less common, with fewer than 500 recipes in each category. This could represent an opportunity to expand the number of recipes catering to users with specific dietary needs, such as those managing hypertension or diabetes.



Analysis Based on the Bar Chart of Recipes by Diet Types:

Healthy Recipes Dominate:

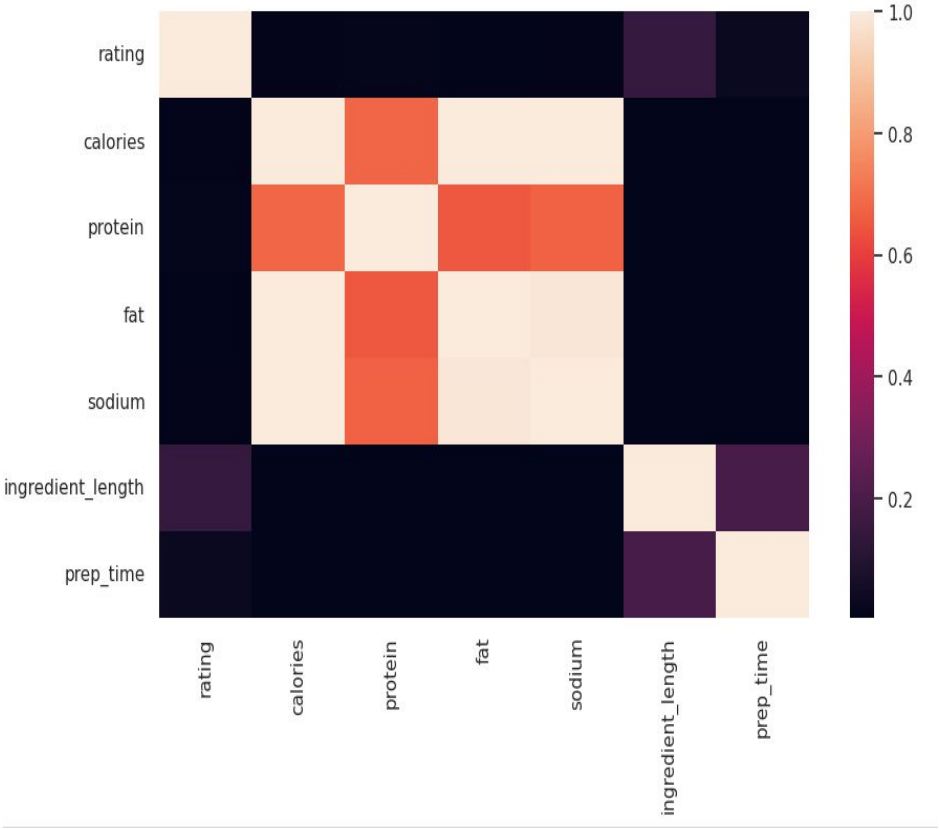
- The largest category by far is **healthy** recipes, with nearly 3,000 recipes. This suggests that there is a strong focus on recipes designed for general health-conscious consumers.

Low Calorie and Low Fat:

- **Low calorie** and **low fat** recipes are also popular, with over 1,500 and 1,800 recipes respectively. These diet types are likely targeting users who are focused on weight management or seeking to reduce their calorie and fat intake for health reasons.

Lower Presence of Low Sodium and Low/No Sugar Recipes:

- **Low sodium** and **low/no sugar** recipes are much less common, with fewer than 500 recipes in each category. This could represent an opportunity to expand the number of recipes catering to users with specific dietary needs, such as those managing hypertension or diabetes.



Recommendations

Handle Outliers:

- Consider capping extreme outliers in calorie, fat, and sodium content to avoid skewing analysis and to provide a clearer picture of the average recipe's nutritional value.

Expand Specific Diets:

- Increase the number of recipes catering to **low sodium** and **low/no sugar** diets to appeal to users with dietary restrictions (e.g., hypertension, diabetes).

Focus on Highly Rated Recipes:

- Promote and highlight recipes with ratings above 4.0 to enhance user satisfaction and engagement.

Investigate the 2005 Surge:

- Look into the factors that led to the massive spike in recipe additions in 2005 to identify opportunities for future growth and engagement.

Conclusion

Healthy and Popular Recipes: The dataset is largely focused on healthy, low-calorie, and low-fat recipes that cater to general wellness, with high user satisfaction.

Opportunities for Growth: Expanding recipe categories for **low sodium** and **low/no sugar** diets presents a key opportunity to meet growing health-conscious trends.

Future Strategy: Capitalize on high-rated recipes for promotional purposes, while addressing the gaps in specific dietary needs to maintain a broad and engaged user base.