**Handling missing values -**

There were multiple columns that had missing values Ex – calories – 4117 , Protein – 4162 etc

So firstly I checked Threshold value For the dataset which indicated not to drop these columns as it was not more than 70% hence I decided to Impute these Missing values

Imputation was Performed using Median value for each column that had missing value because distribution was Skewed Ex – calories column left skewed and there is a sufficient amount of data available to calculate an accurate median

**Handling Duplicate data –**

Firstly checked for Exact Duplicated and then dropped all duplicates keeping First instance

Duplicates could distort the analysis results, such as in statistical summaries hence removed as the dataset contained numerical columns like Claories , Sodium etc

**Handling Outlier data –**

Remove records with impossible data values. Remove records that have “impossible” data values 1. Ratings of 0. Ratings start with 1 “fork” so a 0 rating is not possible. 2. Calories of 0. Unless its a recipe for water or salt, recipes have to have more than 0 calories. 3. Recipes with fat, protien and sodium all equal to 0.

Assumed that for dietary recommendations, it is critical to focus on a specific range of nutrient values (e.g., sodium, calories).Ex – calories value > 10000 was dropped

used box plots to visually identify outliers in columns then The Interquartile Range (IQR) is calculated as IQR=Q3−Q1IQR = Q3 - Q1IQR=Q3−Q1, where Q is the first quartile (25th percentile) and Q3 is the third quartile (75th percentile). Lower bound and Upper bound was calculated and then filtered data according to the Bounds and Z-score because data deviates from the mean

**Exploratory Data Analysis and Insights -**

Number of Highly Rated Recipes: 13703

Number of LOW Rated Recipes: 3627

Highest Rated - Ginger Spice Cookies (after sorting (ascending))

Lowest Rated - Pesto From the Sea (after sorting (ascending))

Count of Recipes by Meal Type: {'Breakfast': 672, 'Lunch': 1246, 'Dinner': 2303}

Number of Vegetarian Recipes: 6046 i.e 34.9% (according to Pie plot)

Number of Non-Vegetarian Recipes: 11284 i.e 65.1% (according to Pie plot)

Count of Dairy Free dishes – 2774 i.e 16% (according to Pie plot)

Count of Dairy included dishes – 14556 i.e 84% (according to Pie plot)

Number of Dishes by Cooking Method ie – Baked dishes much more as compaired to boil

Dishes Available on Diwali occasion (seasonal) are – 25 by count

More no of dishes and recipies suits the Thanksgiving occasion

The most popular occasion for preparing dishes is Thanksgiving, followed by Christmas and New Year's Day. The least popular occasion is Halloween.

17.9% Dishes are dessert and 82.1% are non-dessert dishes

**Corelation Analysis**

The correlation matrix shows the relationship between the nutritional values (calories, protein, fat, sodium) and the ratings of a dataset. conclusions and insights that can be drawn from the output are:

**Positive Correlations:**

Calories and rating: There is a strong positive correlation between calories and rating. This suggests that foods with higher calorie content tend to have higher ratings.

Protein and rating: There is a moderate positive correlation between protein and rating. This indicates that foods with higher protein content are generally rated higher than those with lower protein content.

Fat and rating: There is a strong positive correlation between fat and rating. This suggests that foods with higher fat content tend to have higher ratings.

**Negative Correlations**:

Sodium and rating: There is a weak negative correlation between sodium and rating. This indicates that foods with higher sodium content are slightly less likely to have high ratings. However, the correlation is weak, so the relationship is not very strong.

**Overall Conclusions**:

The ratings of the foods in the dataset are primarily influenced by the calorie content and fat content.

Protein content also plays a role in determining ratings, but to a lesser extent.

Sodium content has a minimal impact on ratings

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

Recommending Dishes For Lactose Intolerant people that are highly rated

Recommending Dishes for Heart Patients based on nutritional content like max\_sodium ,max\_calories , max\_fat , max\_protein

Meal Recommmendation for Breakfast , Lunch and Dinner based on Ratings and Nutritional content

Recommending dishesh based on cusine Type