Data Cleaning Process

1. Challenges Faced During Data Cleaning

The dataset presented several challenges that required careful handling to ensure a smooth and accurate analysis. Below are the key challenges faced during the cleaning process:

1.1 Inconsistent and Duplicate Entries: The dataset contained duplicate rows for some recipes, which indicated either repeat entries or errors in data entry. These duplicates needed to be removed to ensure each recipe was represented uniquely and to avoid skewing the analysis. Additionally, the data was standardized by converting text to lowercase to maintain consistency across the dataset. The data was also checked for inconsistencies, such as negative values in the nutritional features (calories, protein, fat, sodium).

1.2 Missing Values: Many rows had missing values for important columns, particularly for nutritional information (calories, protein, fat, sodium). These missing values needed to be addressed to prevent skewed results in the analysis. Ignoring them entirely or replacing them with incorrect assumptions could lead to inaccurate conclusions.

1.3 Outliers in Nutritional Information: Outliers were found in columns such as calories, protein, fat, and sodium. Some recipes had extremely high values, which were well outside the normal range for similar recipes. These outliers could distort the analysis if not handled properly.

1.4 Feature Categorization: The dataset has 680 columns, covering various categories like meal types, regions, cooking methods, ingredients, and dietary preferences. It was difficult to understand and organize these columns because of the large number and variety. This step involved grouping similar features to simplify analysis and ensure that key variables such as nutritional information and recipe types were properly categorized for further exploration.

2. Decisions Made During Data Cleaning

To address the challenges mentioned above, the following decisions were made during the cleaning process:

2.1 Handling Inconsistent and Duplicate Entries:

- **Normalization of Categorical Columns:** All text entries were standardized by converting them to lowercase to eliminate discrepancies caused by case sensitivity (e.g., "Vegetable" vs. "vegetable"). This ensured consistent values for categorical columns.
- **Removing Duplicates:** Duplicate rows were removed to ensure that each recipe was represented uniquely without redundant entries.

2.2 Handling Missing Values:

- **Imputation of Missing Nutritional Values:** For columns such as calories, protein, fat, and sodium, where values were missing, we used the median values of the respective columns to fill in the missing data. Median imputation was chosen to avoid the influence of extreme values, making it a more robust option than mean imputation.

2.3 Handling Outliers:

- **Winsorization of Outliers:** Nutritional columns (calories, protein, fat, sodium) with extreme outliers were winsorized. This technique limited the influence of outliers by capping values above the 99th percentile and below the 1st percentile, ensuring that the majority of the data was within a reasonable range.

2.4 Feature Categorization: The dataset, with 680 columns, was organized into several categories to simplify analysis:

- **Recipe Information:** Titles, ratings, and nutritional data (e.g., calories, protein).
- **Ingredients:** List of all ingredients used in the recipes.
- **Meal Types:** Categories like breakfast, lunch, and special occasions.
- **Cooking Methods/Tools:** Cooking techniques and tools (e.g., bake, grill, blender).
- **Tags/Occasions:** Events and dietary tags (e.g., Christmas, low carb).
- **Regions:** Locations associated with the recipes, like specific states or countries.

This categorization made it easier to analyze and extract meaningful insights from the dataset.

3. Assumptions Made During Preprocessing

In some instances, assumptions had to be made to move forward with the analysis. These assumptions are documented below:

3.1 Imputation of Missing Values: It was assumed that the missing values for nutritional information (calories, protein, fat, sodium) were missing at random. This assumption justified the use of median imputation, as it helped maintain the central tendency of the data without making biased predictions about what the missing values might have been.

3.2 Categorization of Binary Flags: When reviewing and adjusting binary flags such as ‘vegetarian’ or ‘christmas’, it was assumed that the title and description of the recipe provided the most accurate information. This assumption led to recategorizing some recipes where the title or description did not match the binary label.

3.3 Winsorization of Outliers: During outlier handling, it was assumed that extremely high or low values in nutritional columns were the result of data entry errors or unusual recipes that would not be representative of most recipes. Thus, winsorization was applied to bring extreme values within a more reasonable range.

4. Summary

In summary, the data cleaning process involved addressing inconsistencies, handling missing values, managing outliers, and ensuring accurate feature categorization. Through this process, careful decisions were made to ensure the dataset was clean, reliable, and ready for analysis. The assumptions made were necessary to maintain the integrity of the dataset and facilitate the exploratory analysis.