Data Scientist Professional Practical Exam Submission

Report: Predicting Popular Recipes for Website Homepage

Introduction

In this notebook, we have been tasked with helping select popular recipes to display on the homepage in order to increase website traffic and subscriptions. The current method of selecting recipes based on personal preference has led to inconsistent results. To address this, we developed a machine learning model in order to standardize results and ensure better recipe selection.

Importing Libraries

Our dataset has 947 rows, with 7 different features or columns.

```
(947, 7)
```

recipe	calories	carbohydrate	sugar	protein	category	servings	high_traffic
1					Pork	6	High
2	35.48	38.56	0.66	0.92	Potato	4	High
3	914.28	42.68	3.09	2.88	Breakfast	1	null
4	97.03	30.56	38.63	0.02	Beverages	4	High
5	27.05	1.85	0.8	0.53	Beverages	4	null
6	691.15	3.46	1.65	53.93	One Dish Meal	2	High
7	183.94	47.95	9.75	46.71	Chicken Breast	4	null
8	299.14	3.17	0.4	32.4	Lunch/Snacks	4	null
9	538.52	3.78	3.37	3.79	Pork	6	High
10	248.28	48.54	3.99	113.85	Chicken	2	null

Rows: 10

Data Validation

Data Type Validation

Adjusting Column Types

When looking at the data types for each column, we see that 'servings' is of type object when it should be numeric. In addition, high_traffic should be of type character.

Converting the appropiate columns to type category.

Filling in missing values with 'Low' since when the data was populated, only high was filled. Also changed column name to be representative of the column itself.

Now all data types correspond to the criteria listed in the instructions and we're ready for the next step!

Handling Duplicate Values

After dropping duplicate values, the rows of our dataset decreased by 23, meaning that there were 23 duplicate values.

Handling Missing Data

Missing values in this dataset make up a small amount of the total percentage of the entire dataset, but it is still significant enough where we shouldn't just drop them out. Instead we will impute values onto them.

In order to decide what statistical measure we will use to impute onto the missing values, we are looking at the variance, mean, and median of the dataset.

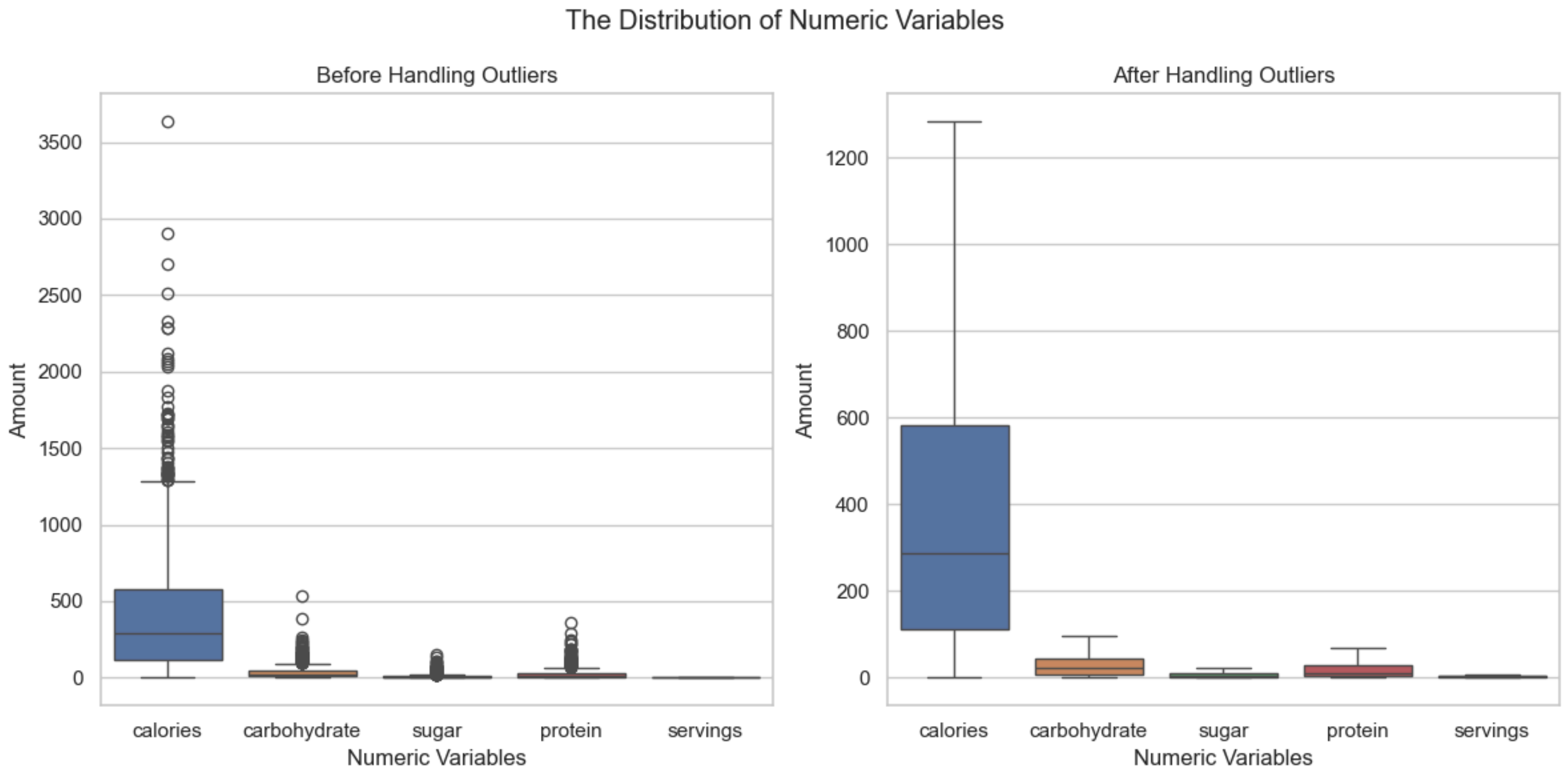
index	calories	carbohydrate	sugar	protein	servings
count	895	895	895	895	
mean	435.9391955307	35.0696759777	9.046547486	24.1492960894	3.4577
std	453.0209971775	43.9490319812	14.6791758036	36.3697385865	1.7390
min	0.14	0.03	0.01	0	
25%	110.43	8.375	1.69	3.195	
50%	288.55	21.48	4.55	10.8	
75%	597.65	44.965	9.8	30.2	
max	3633.16	530.42	148.75	363.36	

Rows: 8

We will impute the median value for each column onto the columns with missing values since there is a huge difference between the lowest value and highest value for each column. This will decrease bias in our dataset.

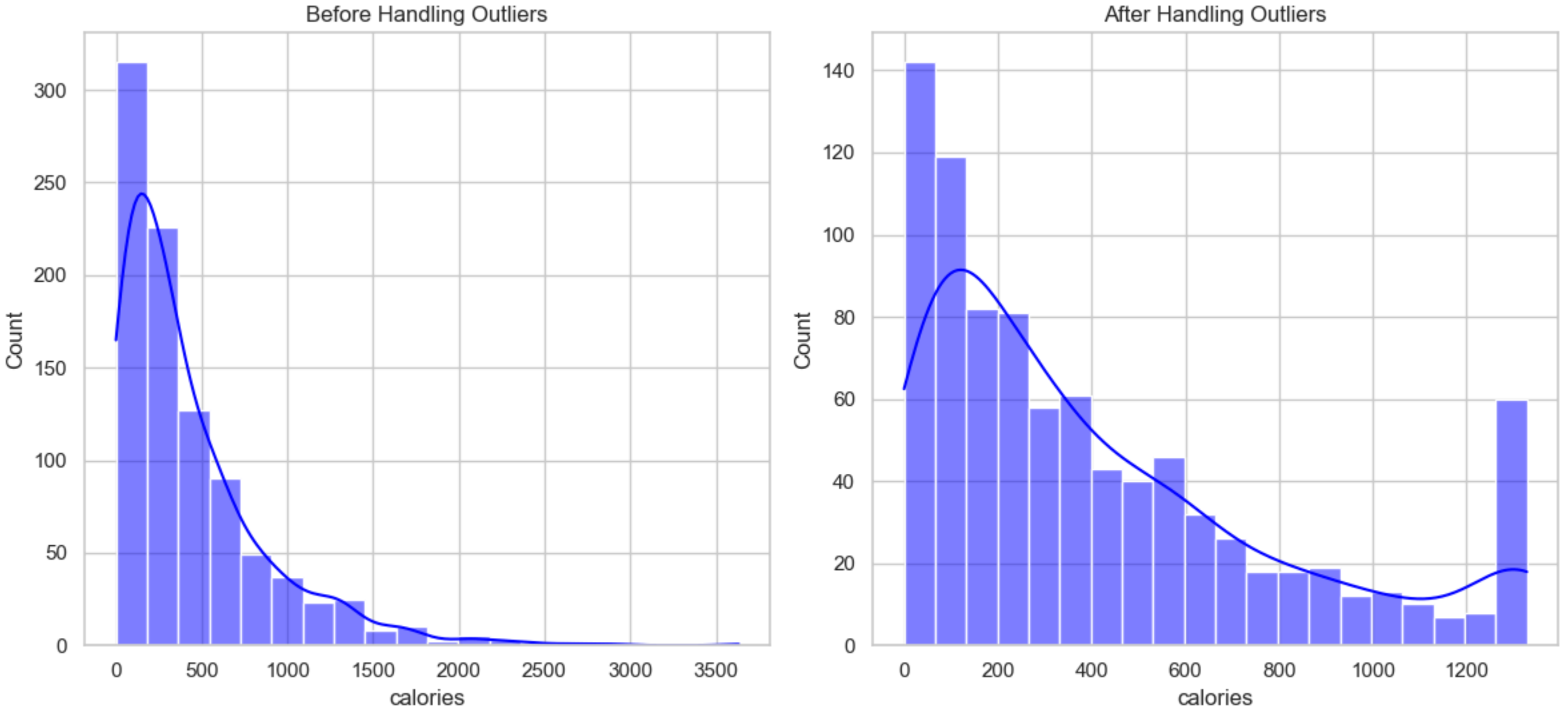
Handling Outliers

We handled outliers using the IQR method and visualized the data distribution for numerical values. It's important to note that there are no outliers in categorical values. Outliers, which represent rare deviations from typical data points, are shown in the graph outside the box:



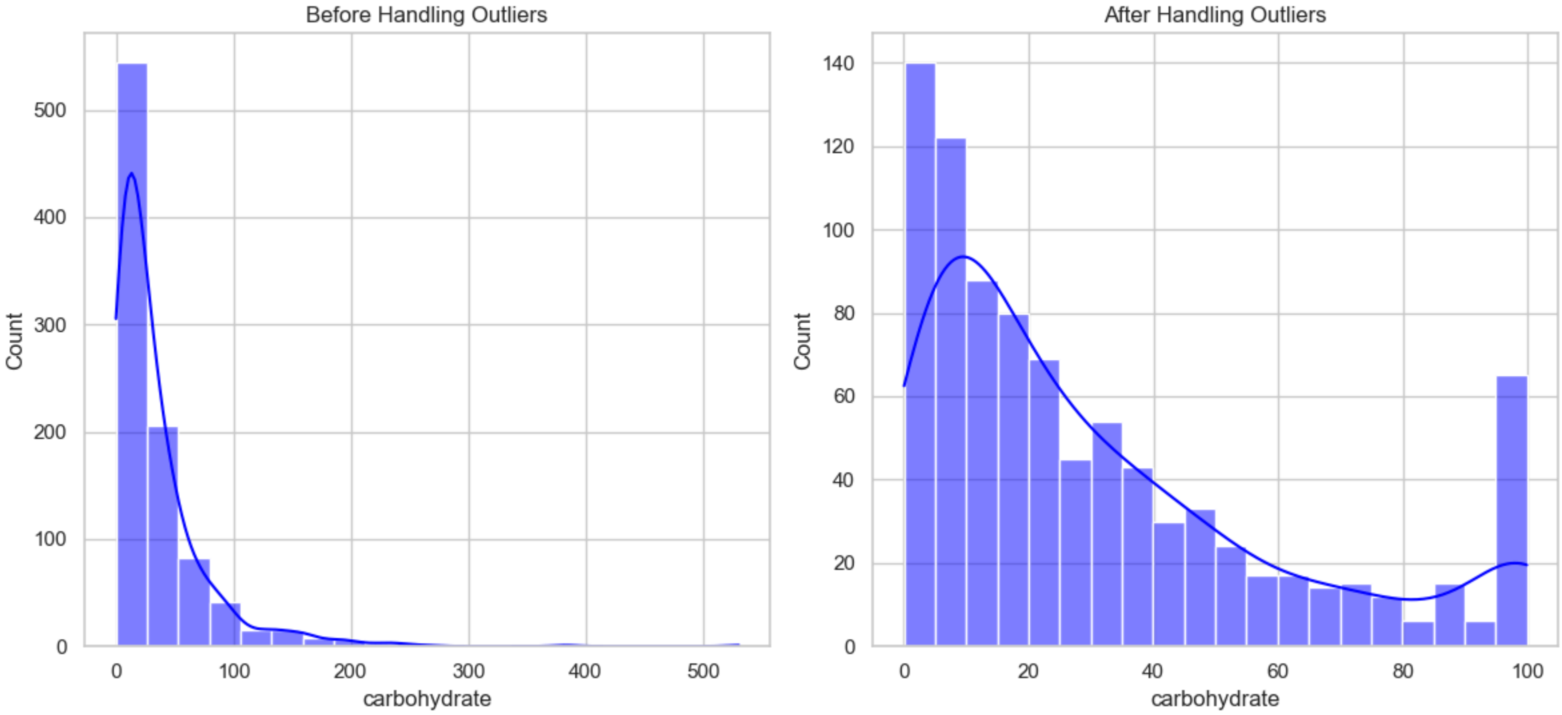
When looking at these graphs, it is clear that our data has a right skew and that handling outliers helps to normalize our dataset.

The Distribution of calories


Before Handling Outliers

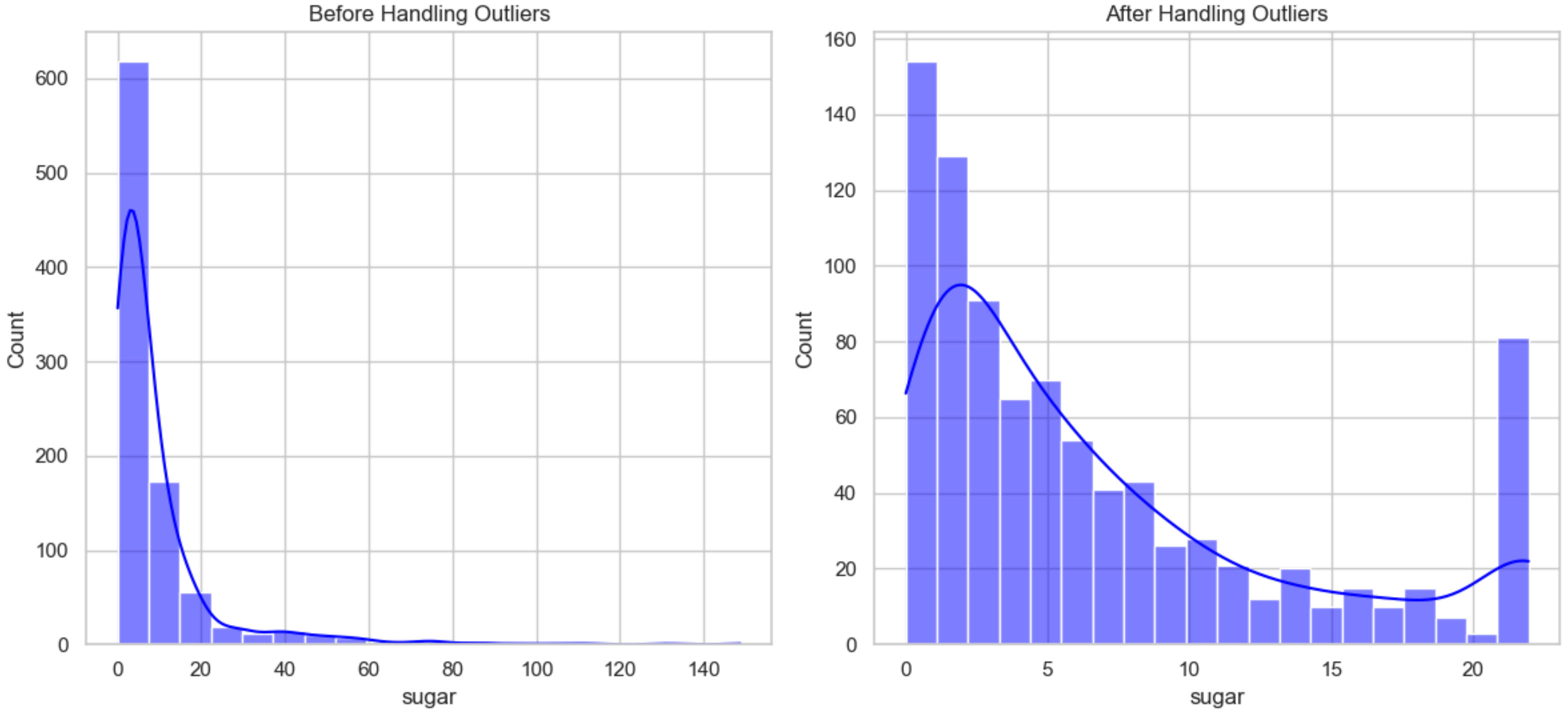
After Handling Outliers

The Distribution of carbohydrate


Before Handling Outliers

After Handling Outliers

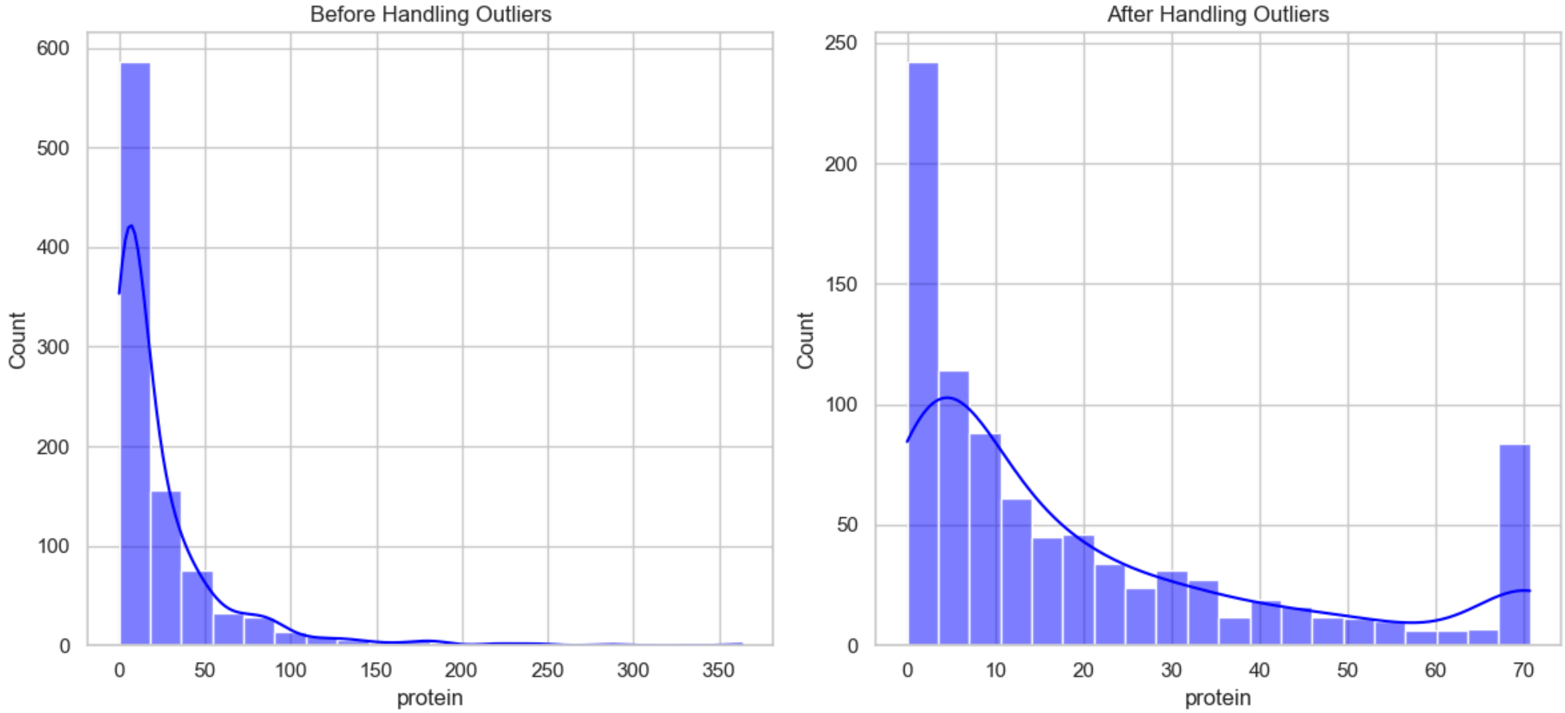
The Distribution of sugar



Before Handling Outliers

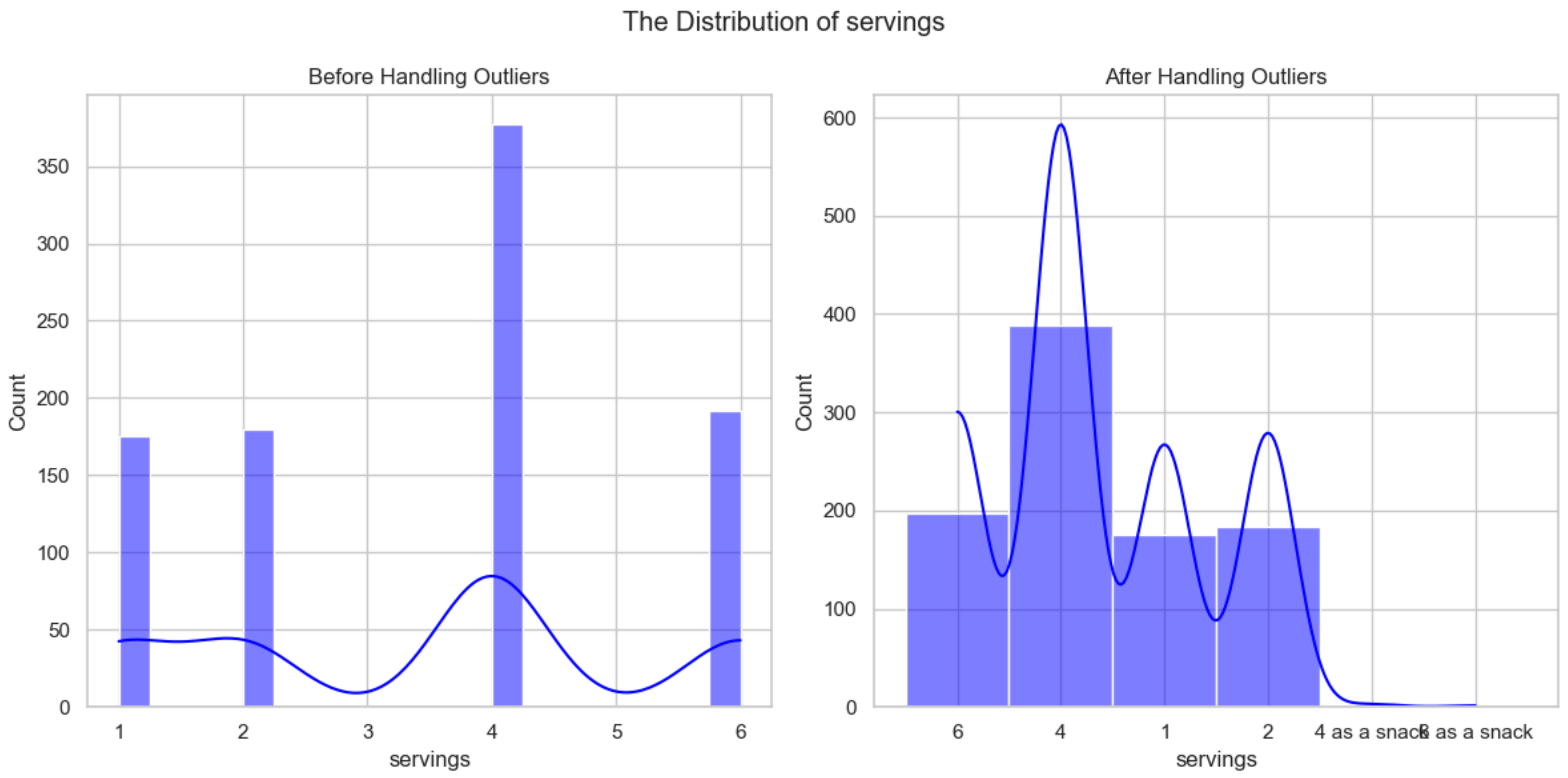
After Handling Outliers

The Distribution of protein



Before Handling Outliers

After Handling Outliers



The last thing we will do in this part is transform our data

Exploratory Analysis

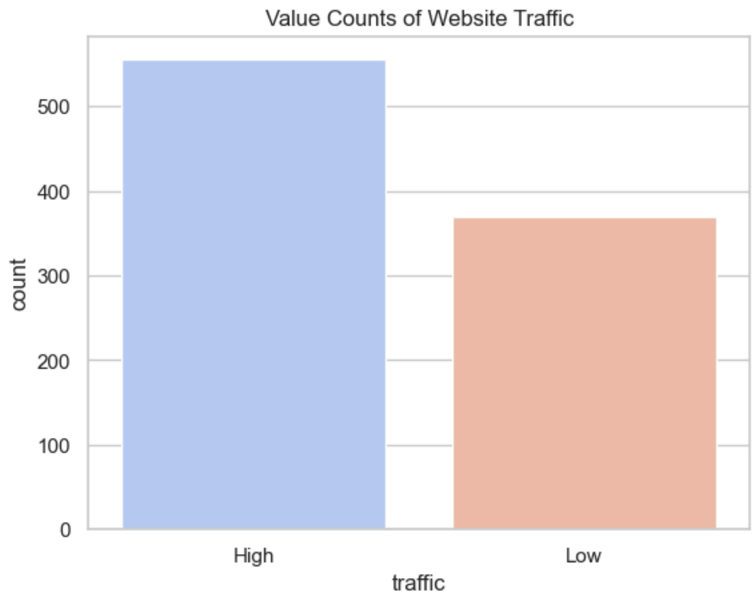
•••	•••	carb... •••	•••	•••	category •••	•••	•••
1	288.55	21.48	4.55	10.8	Pork	6	High
2	35.48	38.56	0.66	0.92	Potato	4	High
3	914.28	42.68	3.09	2.88	Breakfast	1	Low
4	97.03	30.56	21.24	0.02	Beverages	4	High
5	27.05	1.85	0.8	0.53	Beverages	4	Low
6	691.15	3.46	1.65	53.93	One Dish Meal	2	High
7	183.94	47.95	9.75	46.71	Chicken Breast	4	Low

Rows: 7

Since our goal of this notebook is to optimize website traffic, the 'traffic' column is our target variable and will be what we're most interested in.

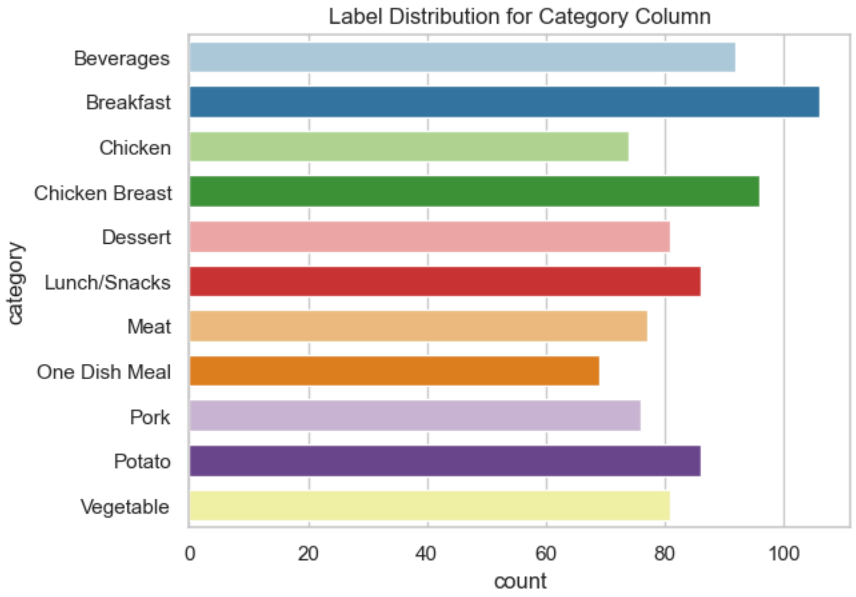
Let's take a look at the class balance in the 'traffic' column!

We see that we have signifcantly more observations with the 'High' label, meaning that we have a class imbalance in our target variable. This could affect the accuracy in our predictions.



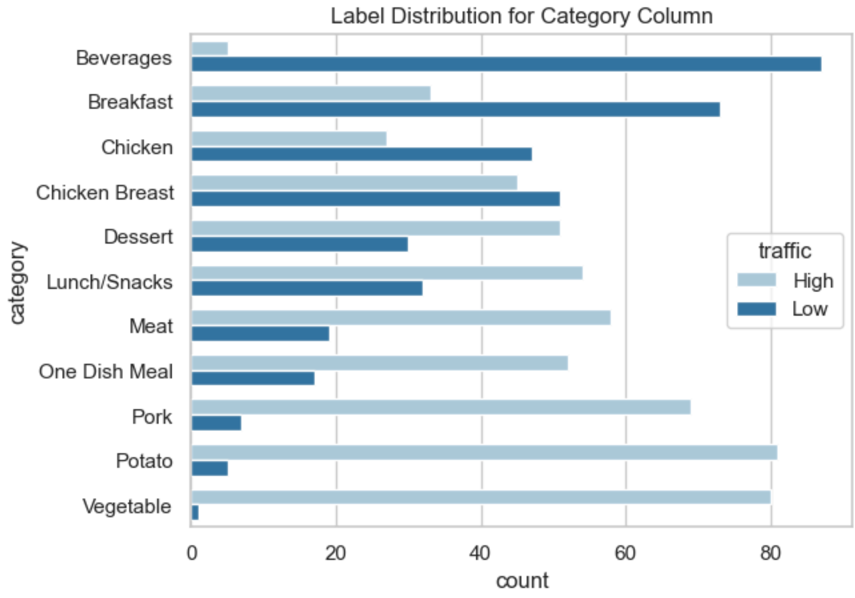
We are also interested how the recipe categories are related and their distribution. Let's check it out!

The class labels are fairly even amoung categories.



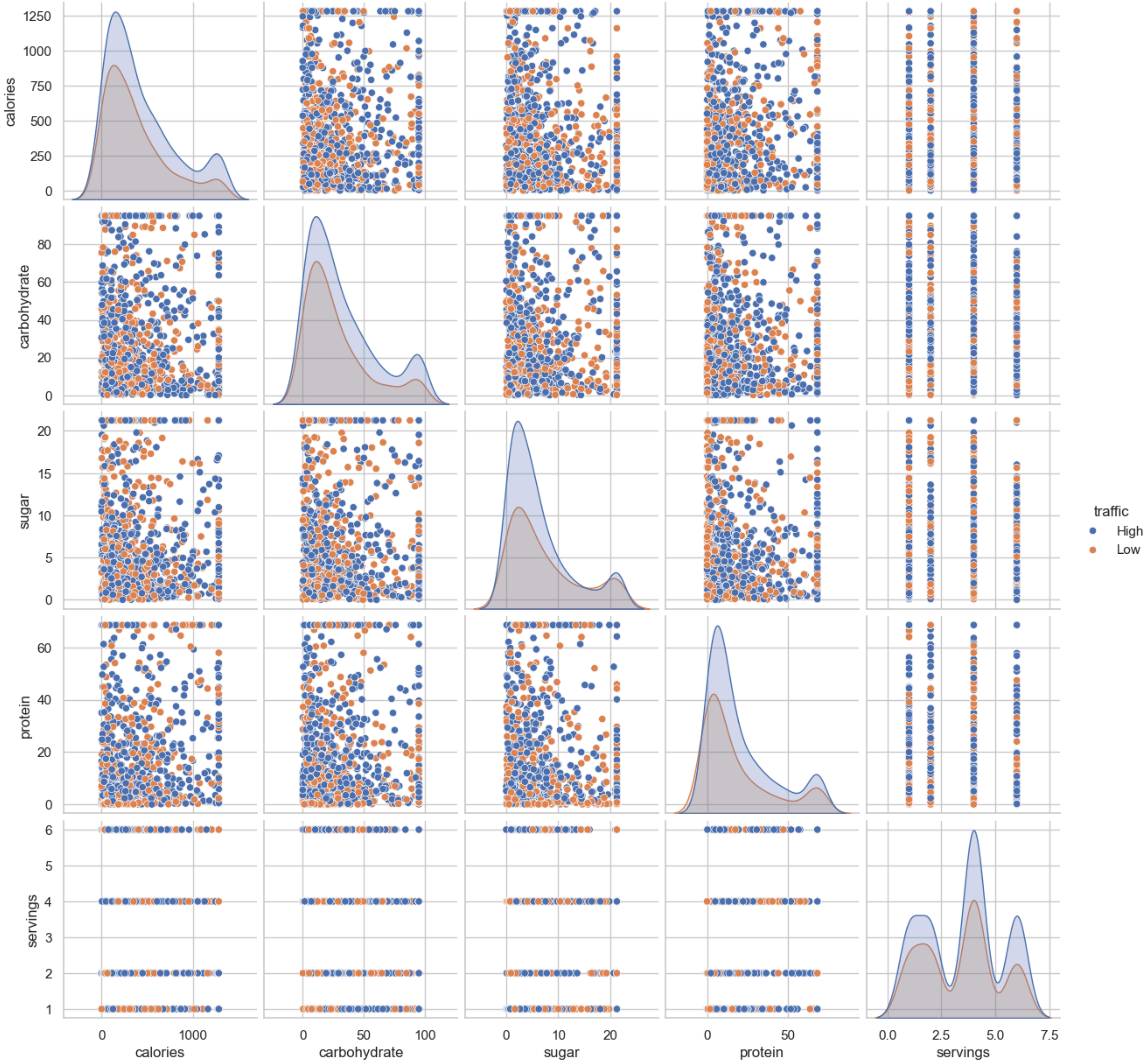
Now let's map traffic column onto this graph!

This clearly shows that some categories, such as Vegetable, Potato, Pork, One Dish Meal, and Meat, are significantly more popular than other categories. This could mean that these if a recipe is in this category, it can indicate if it will be popular!



Finally, let's look at some plots between our numeric variables with traffic quality mapped onto them!

From this pairplot, it is difficult to determine an exact relationship between the target variable and numeric features. This could indicate that the numerical features don't influence the target variable.


calories
carbohydrate
sugar
protein
servings
traffic
● High
● Low
1250
1000
750
500
250
0
80
60
40
20
0
20
15
10
5
0
60
40
20
0
6
5
4
3
2
1
0
1000
0
50
100
0
10
20
0
50
0.0
2.5
5.0
7.5

Model Development

When trying to predict our target variable, we will be trying to do binary classification. In addition, we also have labels for our model to use, meaning that this is a supervised machine learning problem. Therefore, we can use a binary classification supervised model. Some examples of these would be, **Binary Random Forest Classifier** or **Binary Logistic Regression classification**.

We will use the **Logistic Regression model** as the baseline for this task due to it's relative simplicity and fast training. Then we will use the **Random Forest model** as a comparision due to its ensemble training nature.

Modifying Features

We will be encoding our categorical features so that we can use them during our classification. In addition, we will be scaling our features so that all numerical features will be given the same weight, instead of simply larger numbers getting more value. These two measures will decrease bias in our dataset and improve the accuracy of our models.

		carb...				category_Beve...	category_Brea...	category_C...	category_Chicken Breast	category_D...	category_Lunch/
1	288.55	21.48	4.55	10.8	6	0	0	0	0	0	
2	35.48	38.56	0.66	0.92	4	0	0	0	0	0	
3	914.28	42.68	3.09	2.88	1	0	1	0	0	0	
4	97.03	30.56	21.24	0.02	4	1	0	0	0	0	
5	27.05	1.85	0.8	0.53	4	1	0	0	0	0	
6	691.15	3.46	1.65	53.93	2	0	0	0	0	0	
7	183.94	47.95	9.75	46.71	4	0	0	0	1	0	
8	299.14	3.17	0.4	32.4	4	0	0	0	0	0	
9	538.52	3.78	3.37	3.79	6	0	0	0	0	0	
10	248.28	48.54	3.99	68.51125	2	0	0	1	0	0	