Data Scientist Professional Practical Exam Submission

Use this template to write up your summary for submission. Code in Python or R needs to be included.

📝 Task List

Your written report should include both code, output and written text summaries of the following:

- Data Validation:
  - Describe validation and cleaning steps for every column in the data
- Exploratory Analysis:
  - Include two different graphics showing single variables only to demonstrate the characteristics of data
  - Include at least one graphic showing two or more variables to represent the relationship between features
  - Describe your findings
- Model Development
  - Include your reasons for selecting the models you use as well as a statement of the problem type
  - Code to fit the baseline and comparison models
- Model Evaluation
  - Describe the performance of the two models based on an appropriate metric
- Business Metrics
  - Define a way to compare your model performance to the business
  - Describe how your models perform using this approach
- Final summary including recommendations that the business should undertake

Data Card

Importing Libraries

Loading Data

	recipe	calories	carbohydrate	sugar	protein
0	1	null	null	null	nu
1	2	35.48	38.56	0.66	0.9
2	3	914.28	42.68	3.09	2.8
3	4	97.03	30.56	38.63	0.0
4	5	27.05	1.85	0.8	0.5

5 rows

Data Composition

Column Servings is displayed as object (String) data type but as shown in the datacard this column should be a category

Dimensions of the data

Addressing Null values

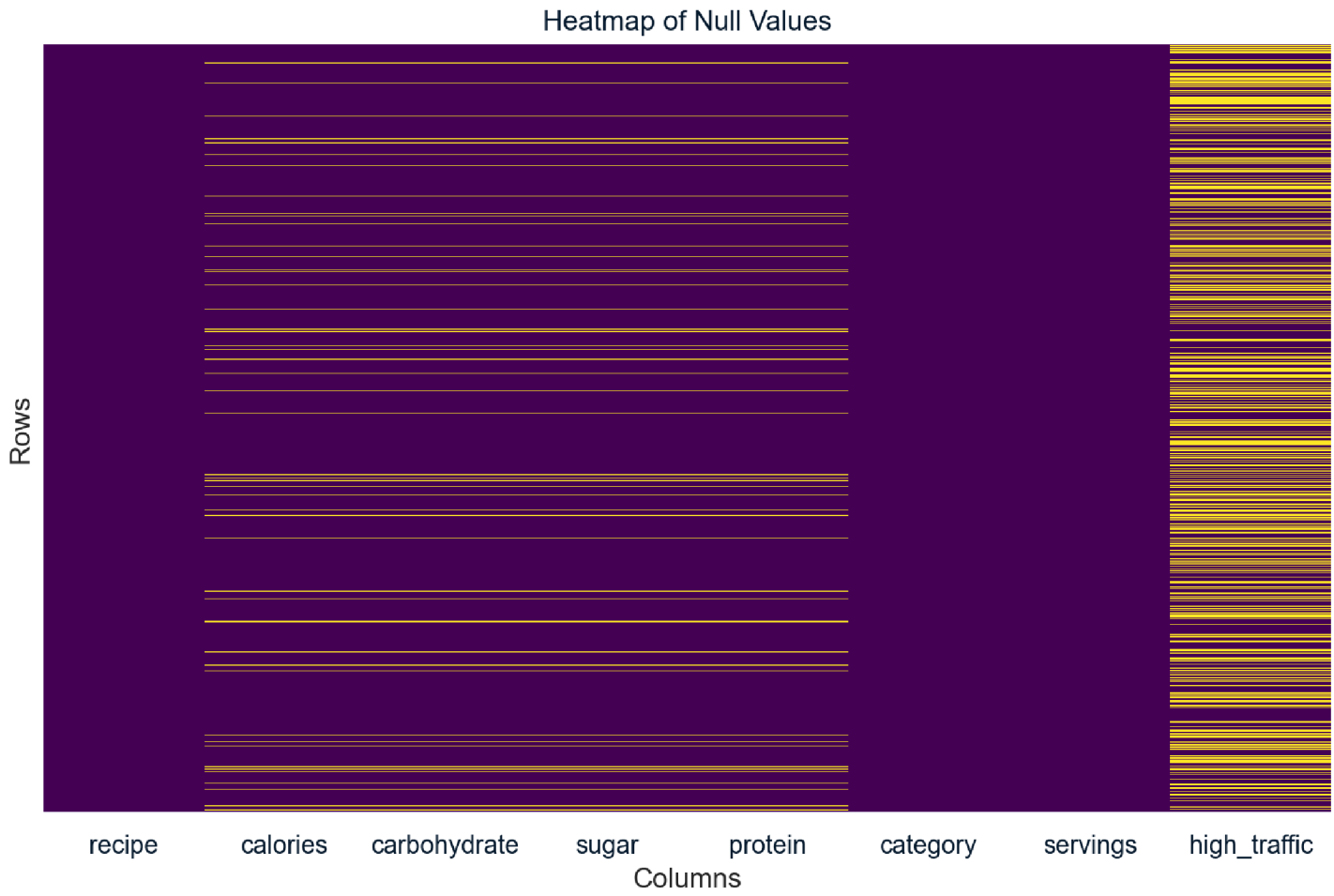
	0
recipe	
calories	
carbohydrate	
sugar	
protein	
category	
servings	
high_traffic	

8 rows

Observation 01:

1. There are 52 recipes where we don't know the calories, carbohydrates, sugar and protein contents of those recipes
2. There are no null values in categories and servings
3. high traffic is the column where we saw 373 null values

Let us visualize them first



Observation 02

1. Alright! so the calories, carbohydrate, sugar and protein value missing in the same row not randomly missing
2. As the missing values are 5.3% and we have a handful amount of data so its not wise to remove those rows
3. I am thinking of imputing these values with category's median for calories, protein, sugar and carbohydrates (Median because the data is left skewed and has alot of high value outliers so mean is not a good option)

11 rows

8 rows

Setting Recipe as Index:

Looking like the recipe column is unique throughout the dataset, if so we can set it as index

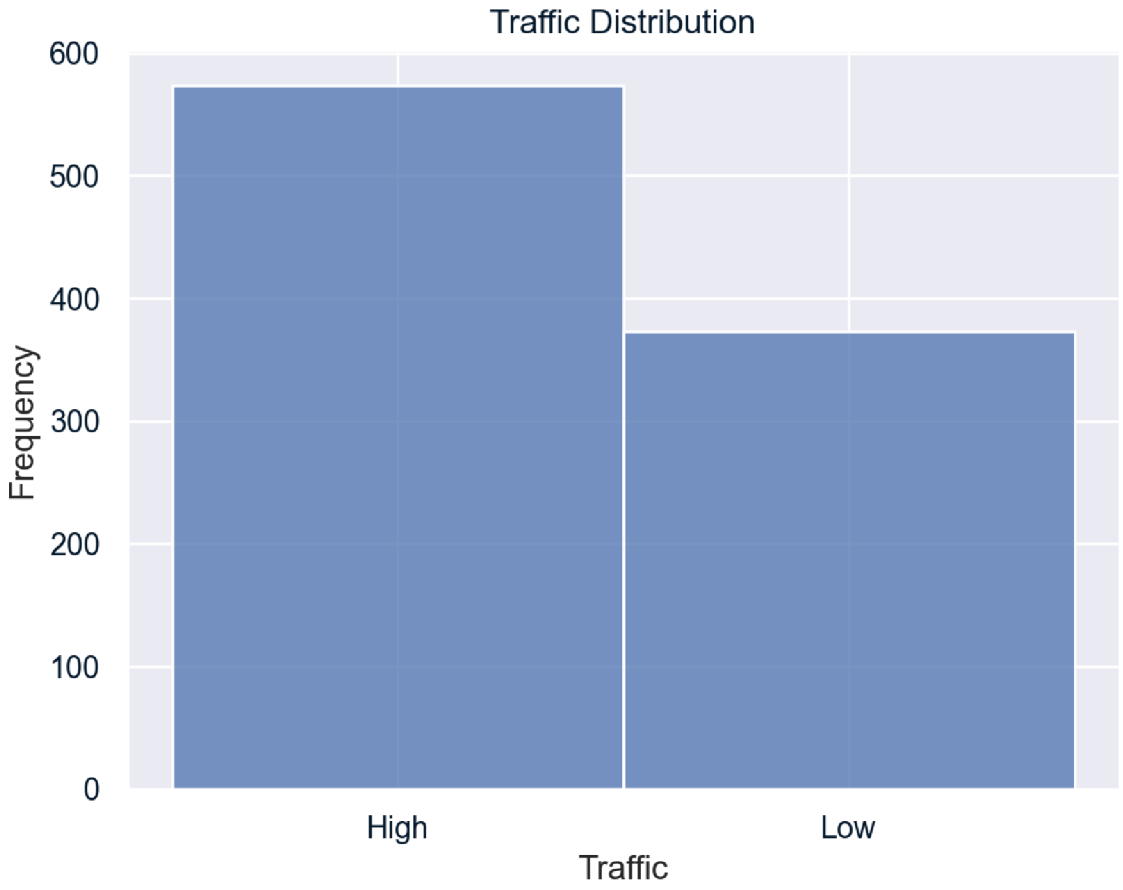
Yes it is unique throughout! We can use it as index

2 rows

Investigate 'high_traffic' column

1 rows

Out of a total of 574 high-traffic recipes, 373 values are not recorded. Given the context in the datacard, these missing values are intentional. Therefore, we can impute these missing entries with the designation 'Low'.



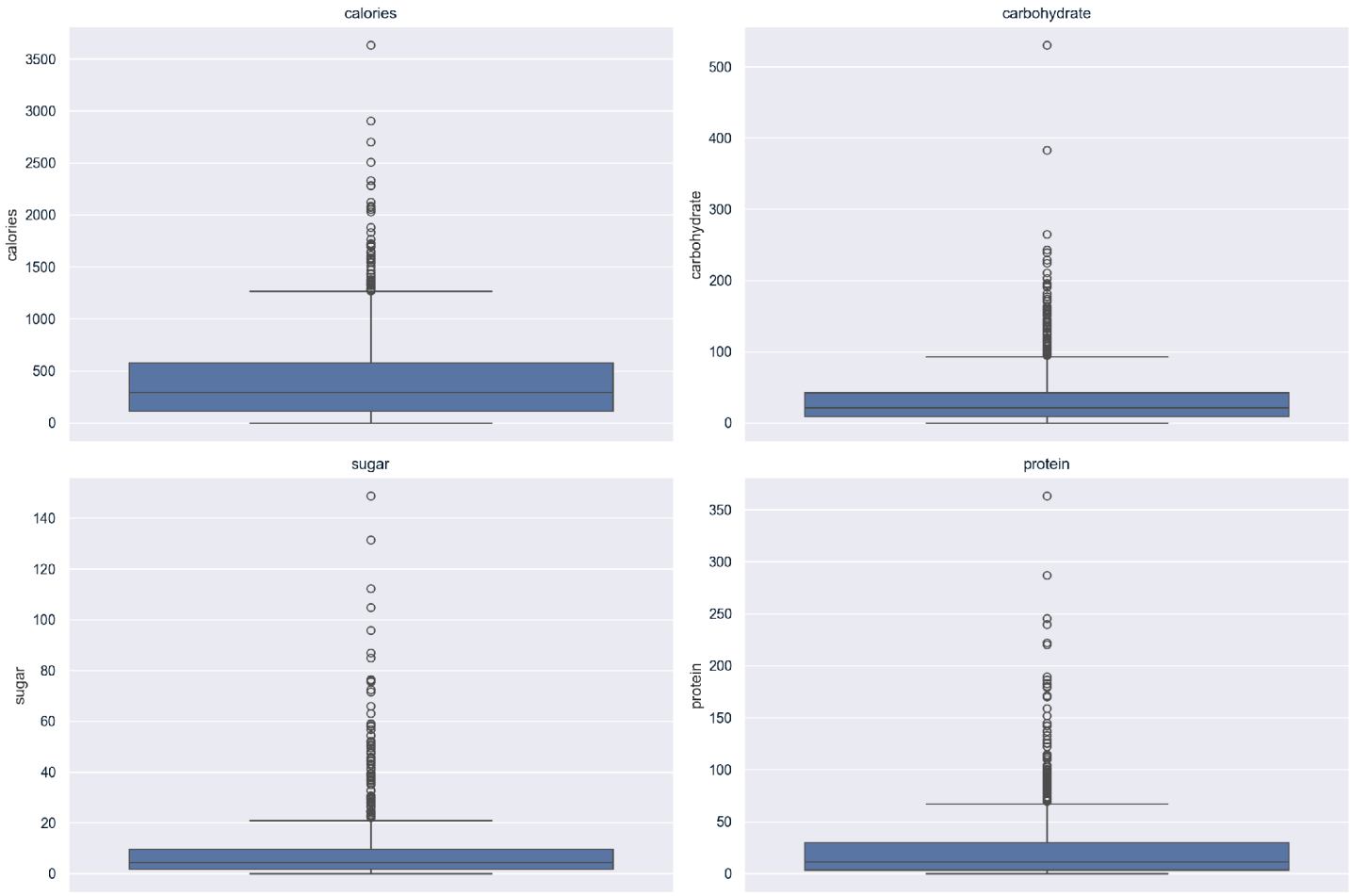
Observation 03:

1. Distribution of the target variable with 574 High traffic recipes and 373 low traffic, slightly skewed but we can balance classes during the modeling stage.

Duplicate values

Outliers in the data

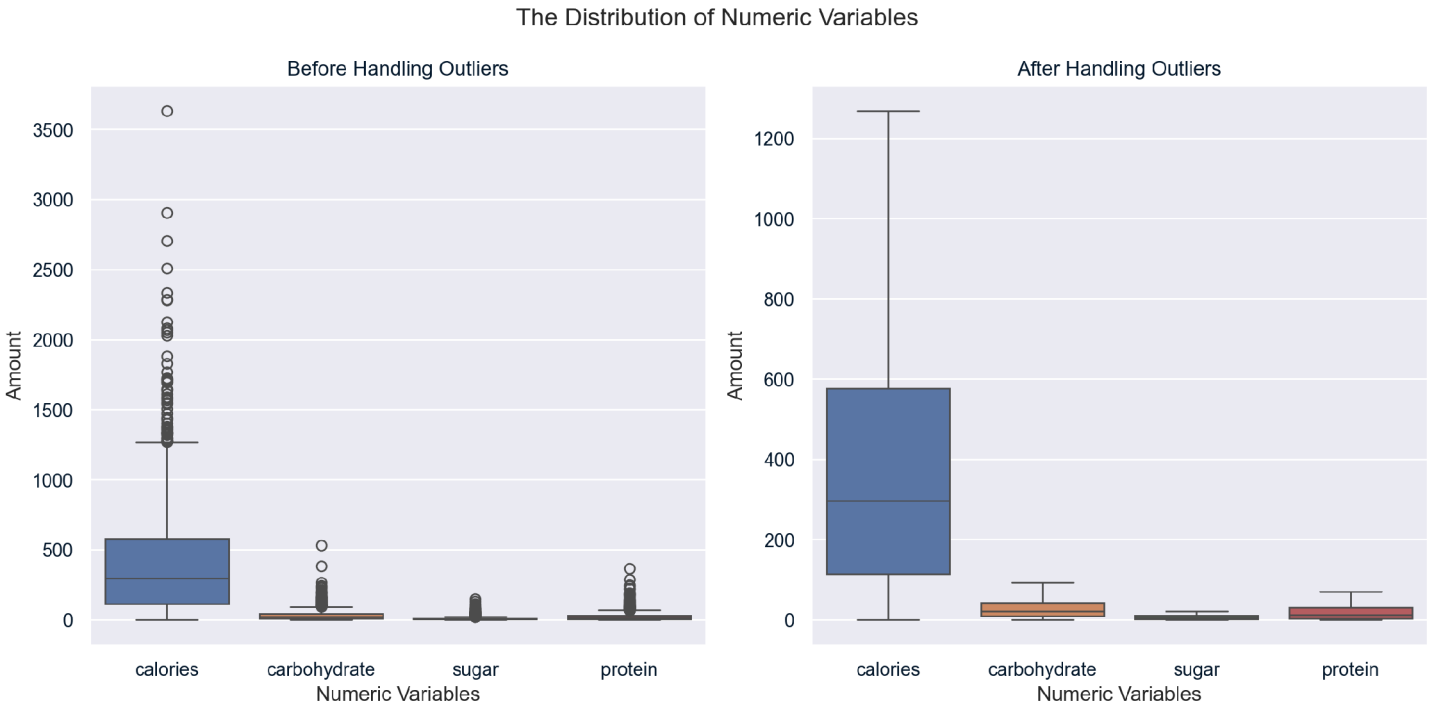
Create boxplots for all numeric columns



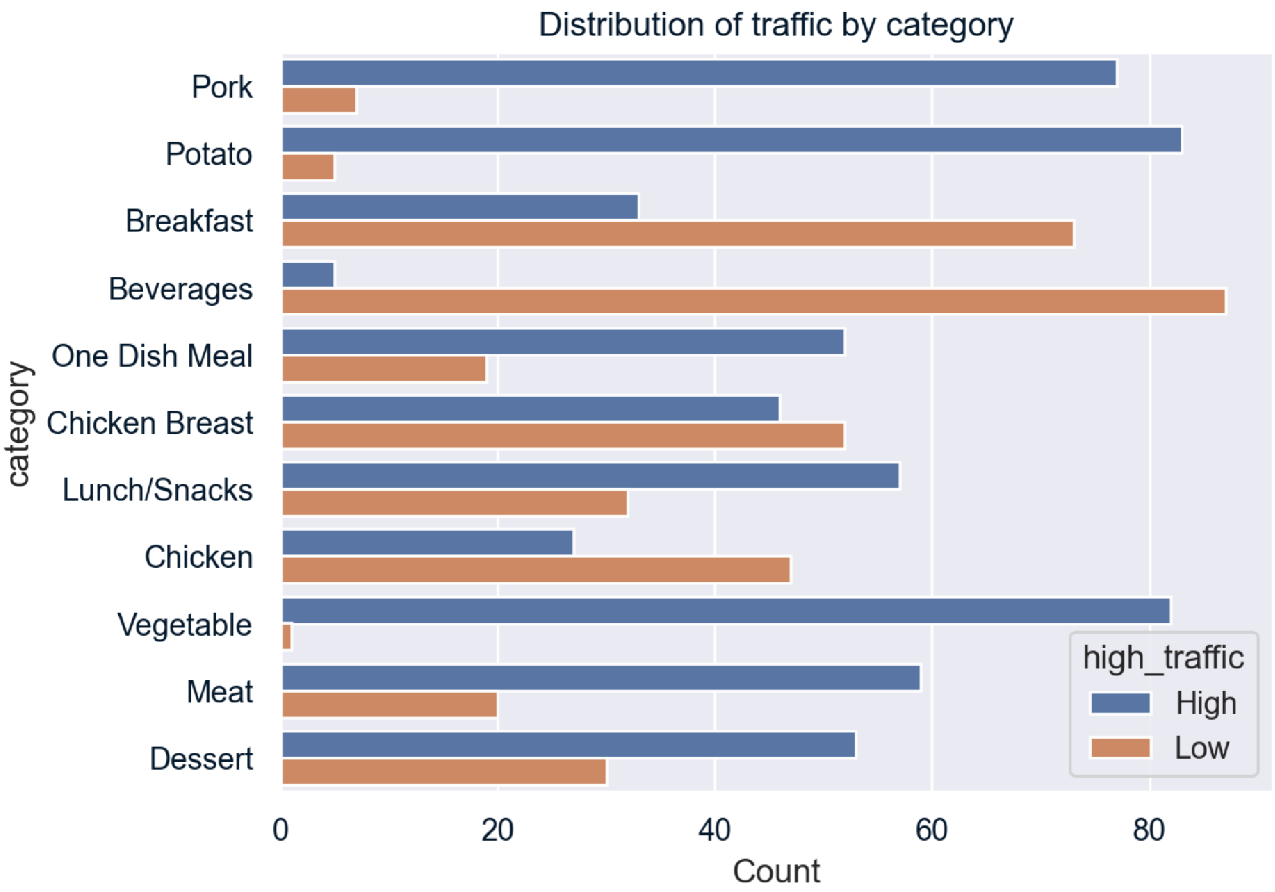
These are some data points that are way above 75th percentile but as these are recipes there can be exceptions but we need to handle these values because they can skew our results and model will underperform.

Handling Outliers with IQR method

Comparison of Numerical variables with outliers and without outliers



Which Dishes had the most amount of traffic, is there any category predominant?



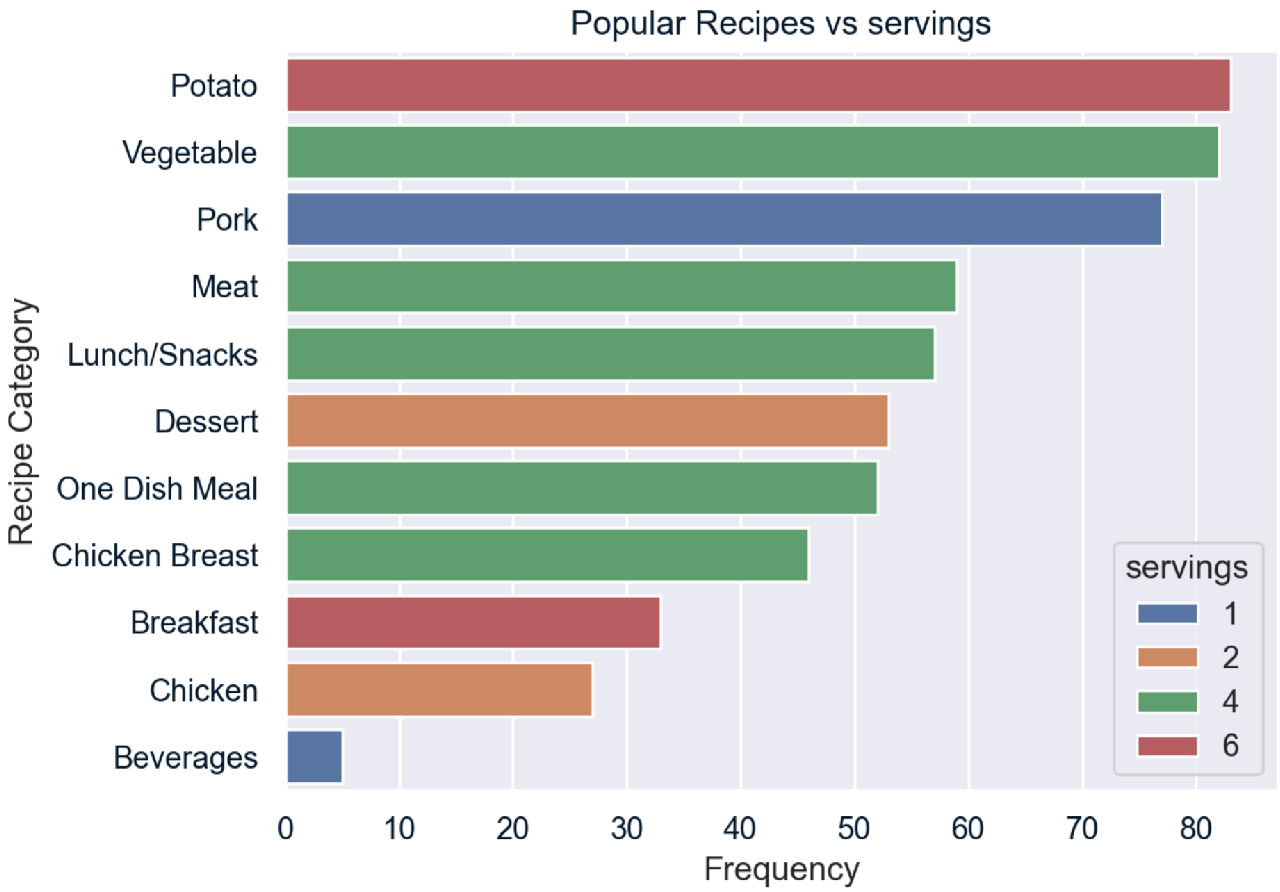
Obersevation 04:

1. The top 4 recipe categories that received the most amount of traffic are Potato, Vegetable, Pork, Meat and Lunch/Snacks.
2. Where as Chicken Breast, Chicken and Beverages are the least popular recipes.
3. Beverages infact are a total disaster, we are not getting any traffic for them.

Business Recomendation 01:

1. Craft more recipe's around Potato, Vegetables and Pork to attract more traffic.

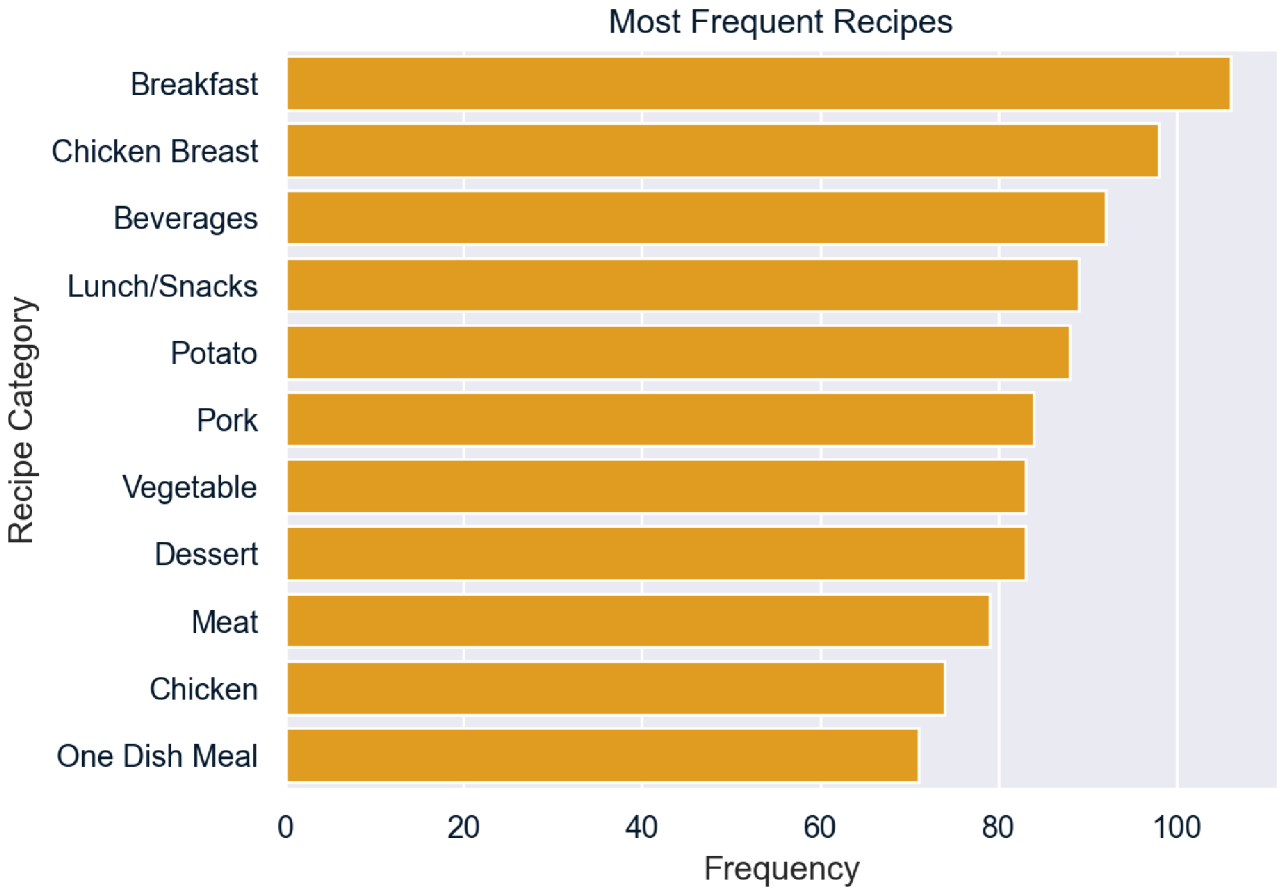
Which recipes are popular when people have to serve 3 or more people?



Observation 04:

1. Visitors usually tend to go for Potato recipe's when serving 4+ people.
2. Pork, chicken and beverages are usually used for 1 serving.
3. However, for 2-4 people Meat and Lunch snacks are common.

Which type of recipes we post most Frequently?



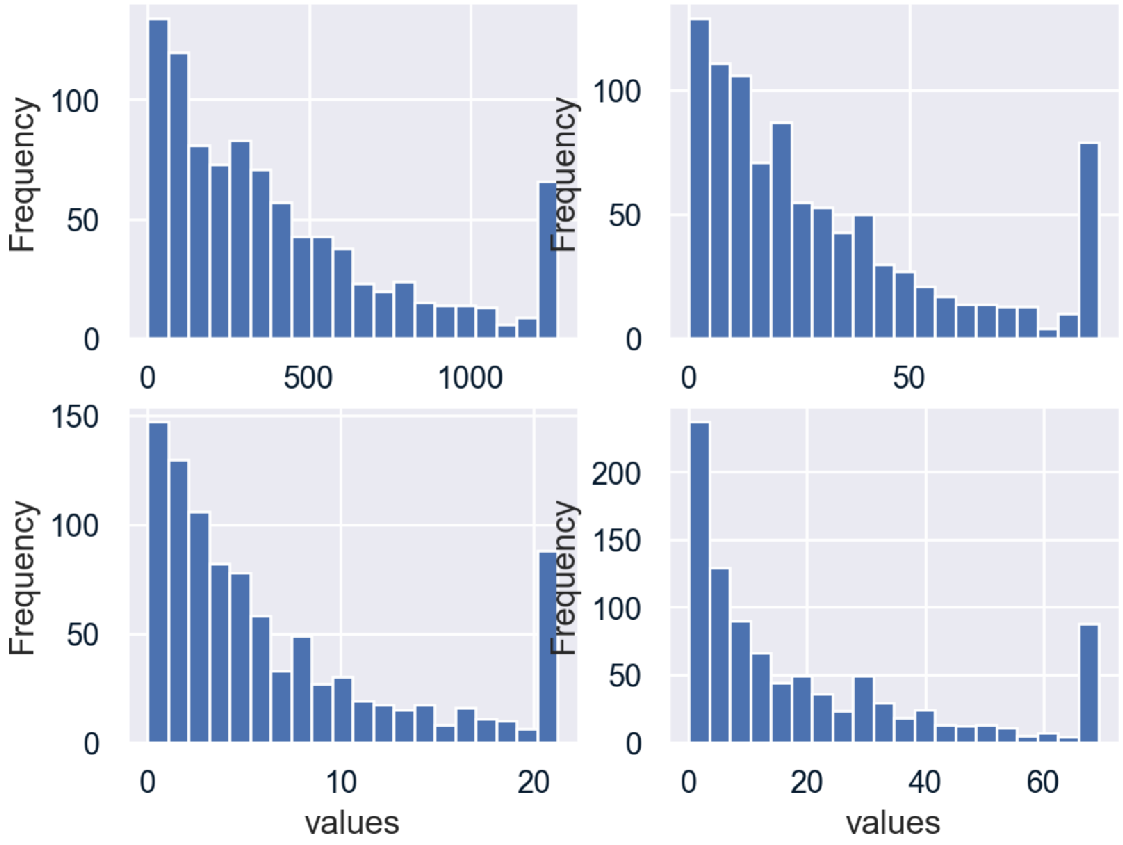
Observation 05:

1. We post Breakfast, Chicken Breast, Beverages and Lunch/ Snack recipes frequently despite the fact these recipes are not performing well in terms of attracting visitors.

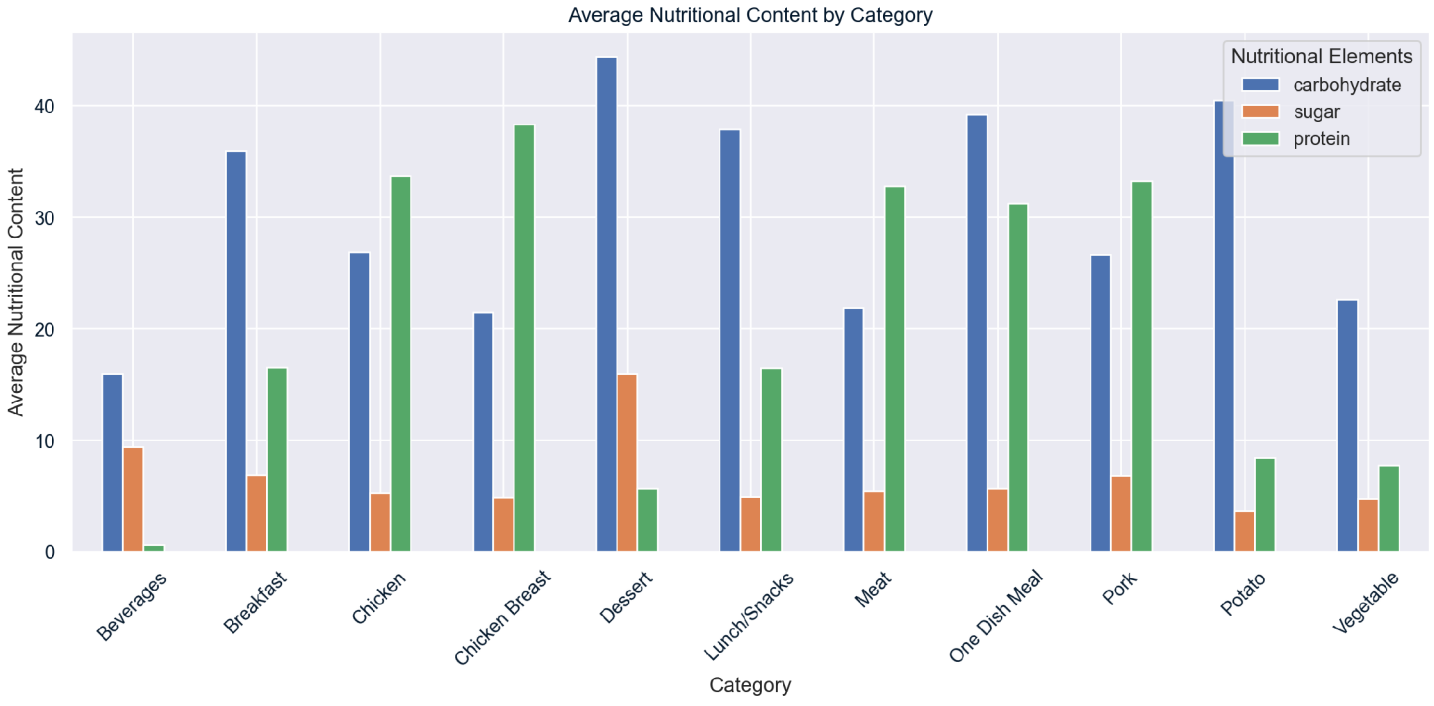
Business Recommendation 02:

1. We need to post more recipe's in Potato, Vegetable, Pork, Meat and Lunch/Snacks categories

What is the distribution of each numeric column? i-e Calories, Carbohydrates, Sugar and Protein



What is the average nutritional content (carbohydrate, sugar, protein) by category?

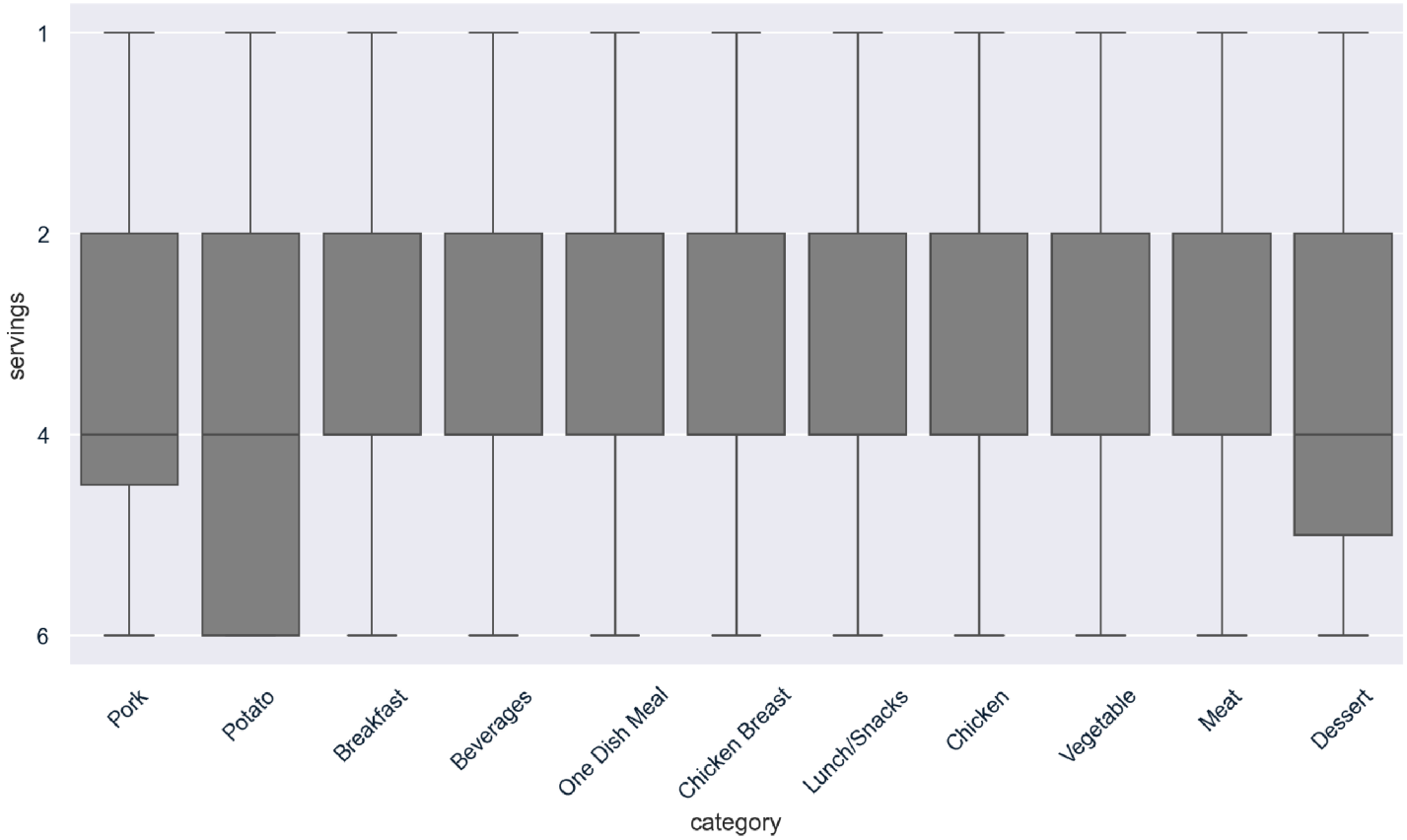


Observation 06:

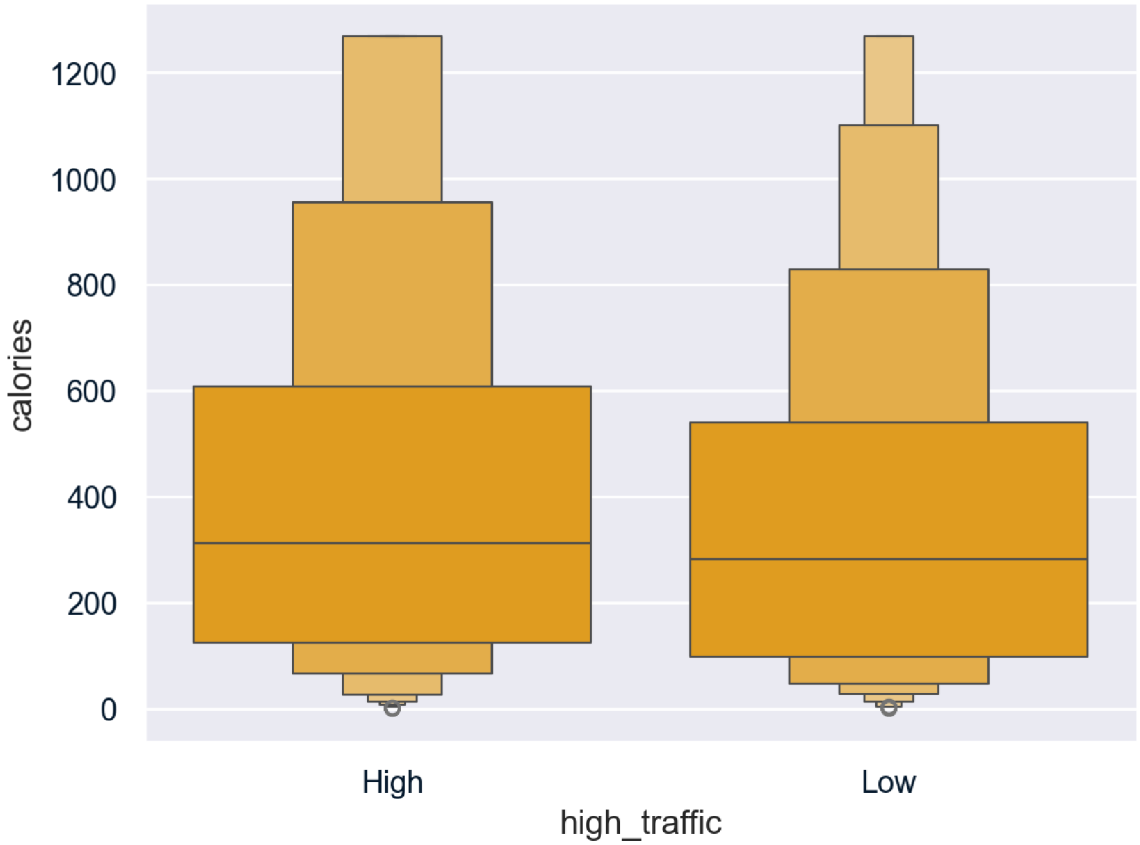
1. Chicken recipe's have the most amount of protein in them while meat and pork are in the middle and least amount of protein is found in vegetables, desserts and beverages.

2. Potato, desserts and one dish meal are mostly comprised of carbohydrates.
3. Desserts and beverages are the sweetest categories ;)

What is the distribution of servings for each category?



Are high-calorie recipes more likely to receive high traffic?

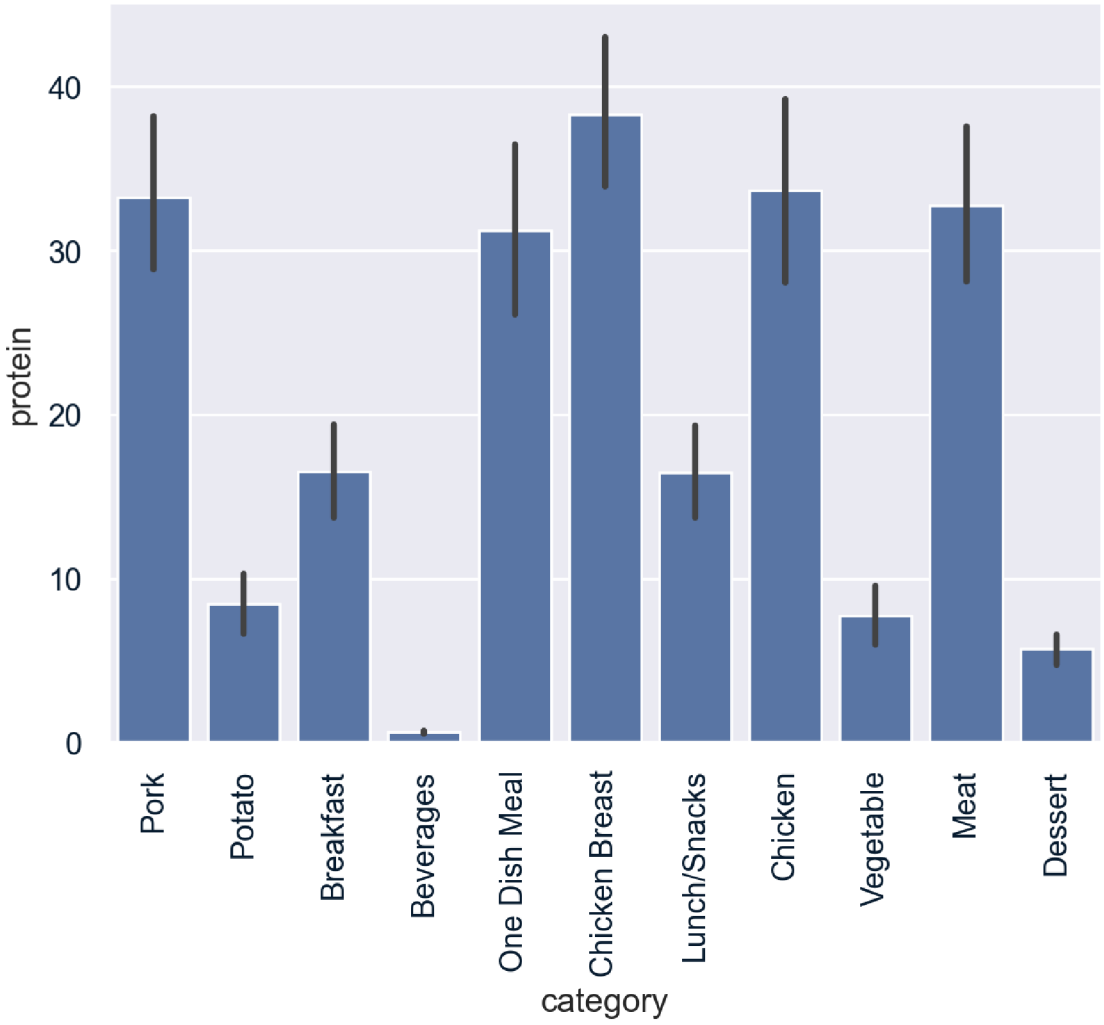


Observation 07:

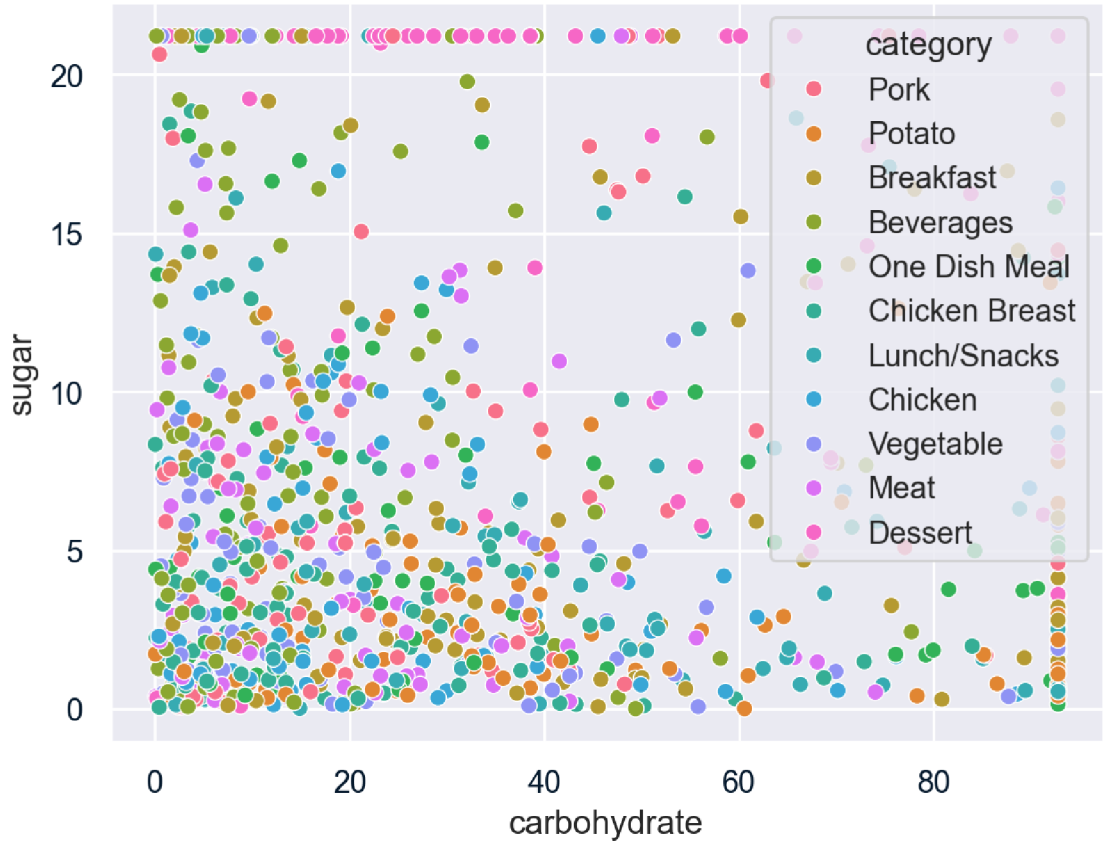
1. There's a slight difference that the high calorie recipes were more likely to receive high traffic that was because we saw that except potato our popular recipes were Pork, meat and lunch/snacks. It will more evident with the visualization given below.

Which categories have the highest calorie content on average?

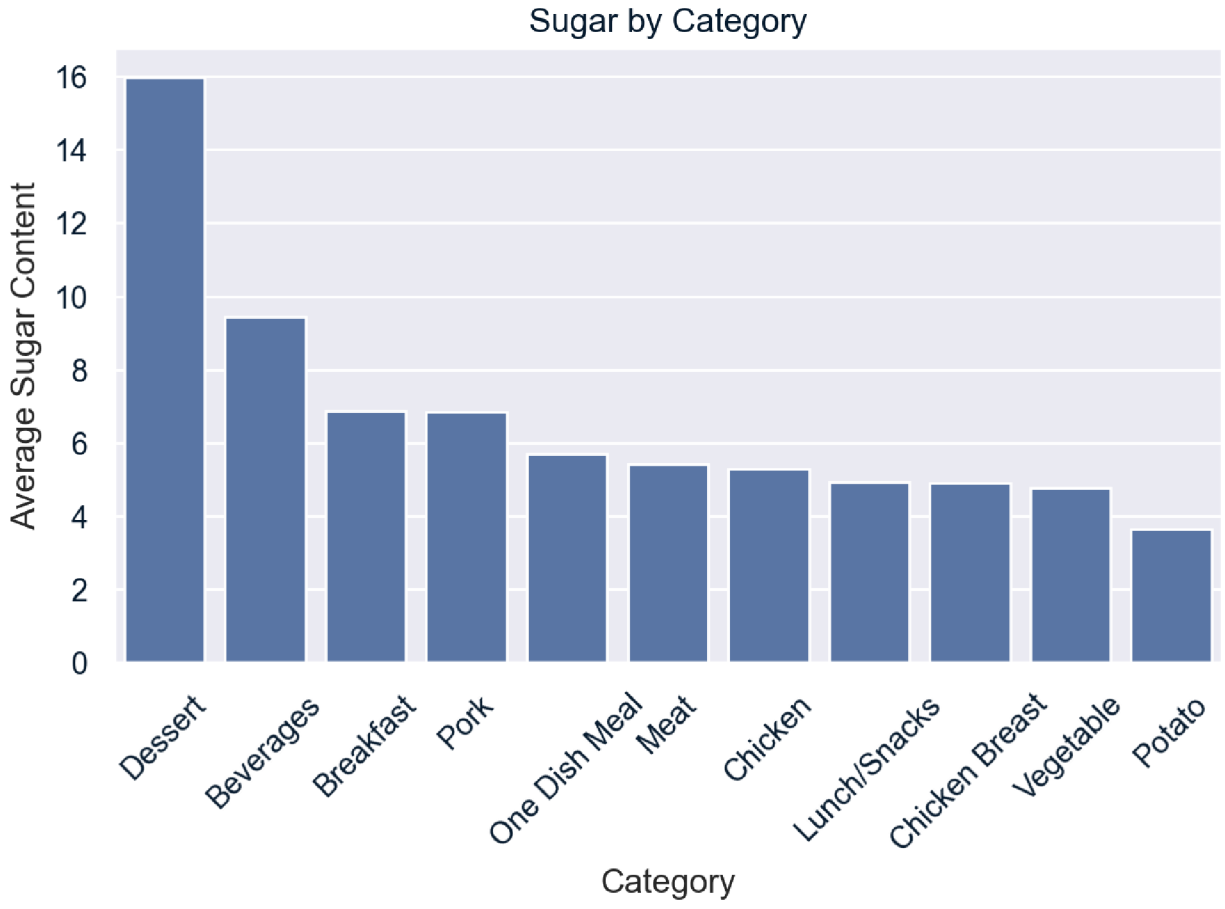
Are recipes with more protein more likely to be in certain categories?



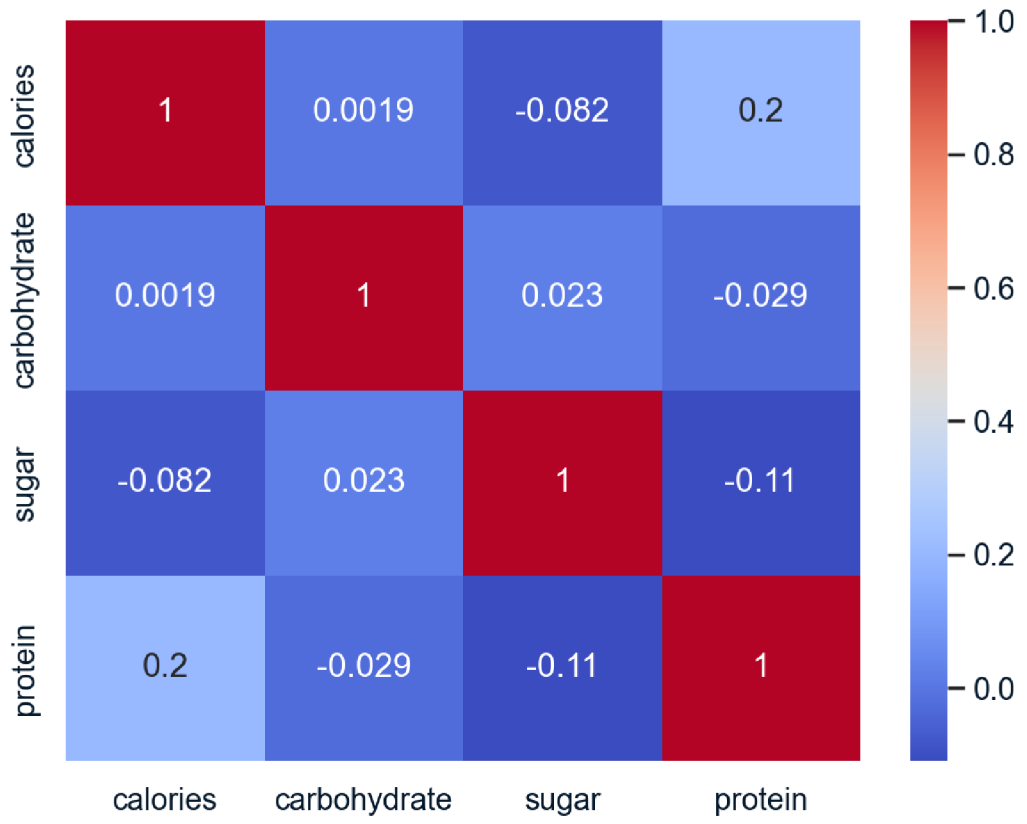
What is the relationship between carbohydrate and sugar content?



Which Categories have the most amount of sugar?



What is the correlation between numeric variables?



Observation 08:

Weak Correlations:

1. Most of the correlations between the numerical features (calories, carbohydrate, sugar, protein) are weak.
2. A weak negative correlation exists between calories and protein. This might indicate that recipes with higher protein content tend to have lower calorie counts.

Positive Correlation:

1. There is a slight positive correlation between carbohydrate and sugar. This suggests that recipes with higher carbohydrate content tend to have higher sugar content.

Model Development

1. Problem Type

** Since the objective of this task is to find out recipe's that are likely to receive high traffic with 80% accuracy, we can call this a ***Binary Classification Problem*** based on the nutritional and categorical features.

2. Baseline Model

Our baseline model will be a logistic regression classifier because:

1. It's a simple, interpretable model for binary classification.
2. It gives us a good initial benchmark without heavy computation.

Separating Target and Input Features

Splitting the data in 80 / 20 Ratio

Fitting the base model (Logistic Regression)

1. Hyperparameter Tuned
2. Scaled Features
3. Balanced Class weights

	precision	recall	f1-score	support
0	0.71	0.71	0.71	77
1	0.80	0.80	0.80	113
accuracy			0.76	190
macro avg	0.75	0.76	0.75	190
weighted avg	0.76	0.76	0.76	190