Traffic Of Recipe Recommend

Data Scientist Practical Exam DS601P

by Alice Lyu

Report Overview

Data Validation

Validation and cleaning steps for each column in the dataset.

Exploratory Analysis

Graphics showing single and multiple variables to demonstrate data characteristics and relationships.

Model Development

Reasons for selecting models and a statement of the problem type, including code to fit baseline and comparison models.

Model Evaluation

Description of model performance based on an appropriate metric.

Data Validation and Cleaning

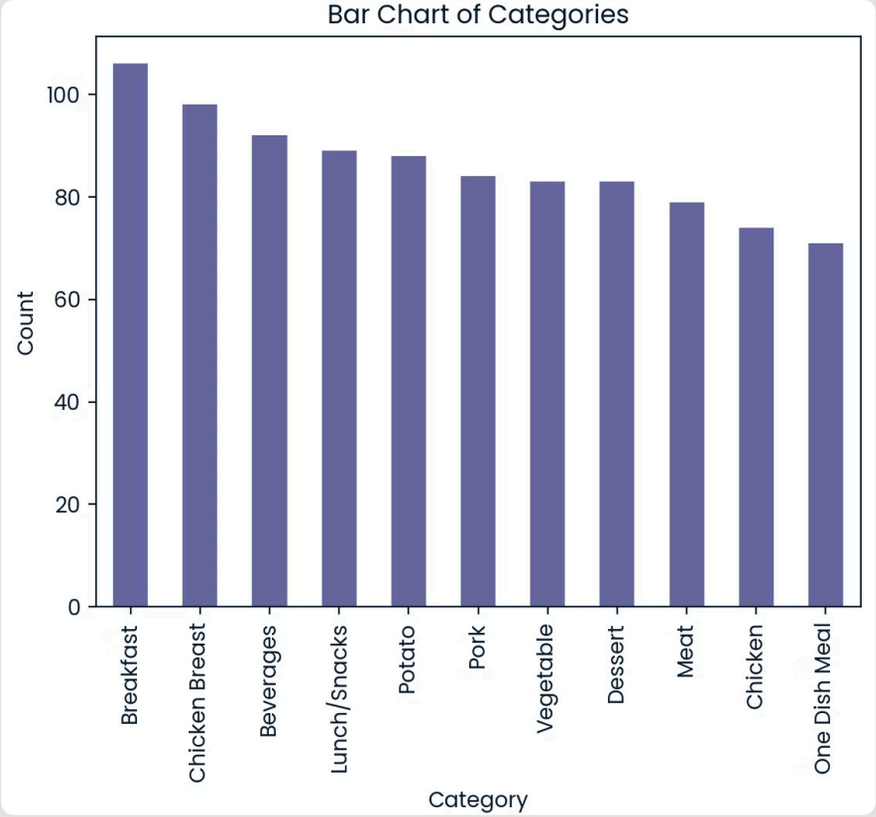
- Dataset: **947** rows, **8** columns.
- Cleaning:
  - missing values in columns ('calories', 'carbohydrate', 'sugar', and 'protein') replaced with multi-index mean values.
  - 'high_traffic' missing values filled with 'Low'.
  - 'servings' standardized to integers.

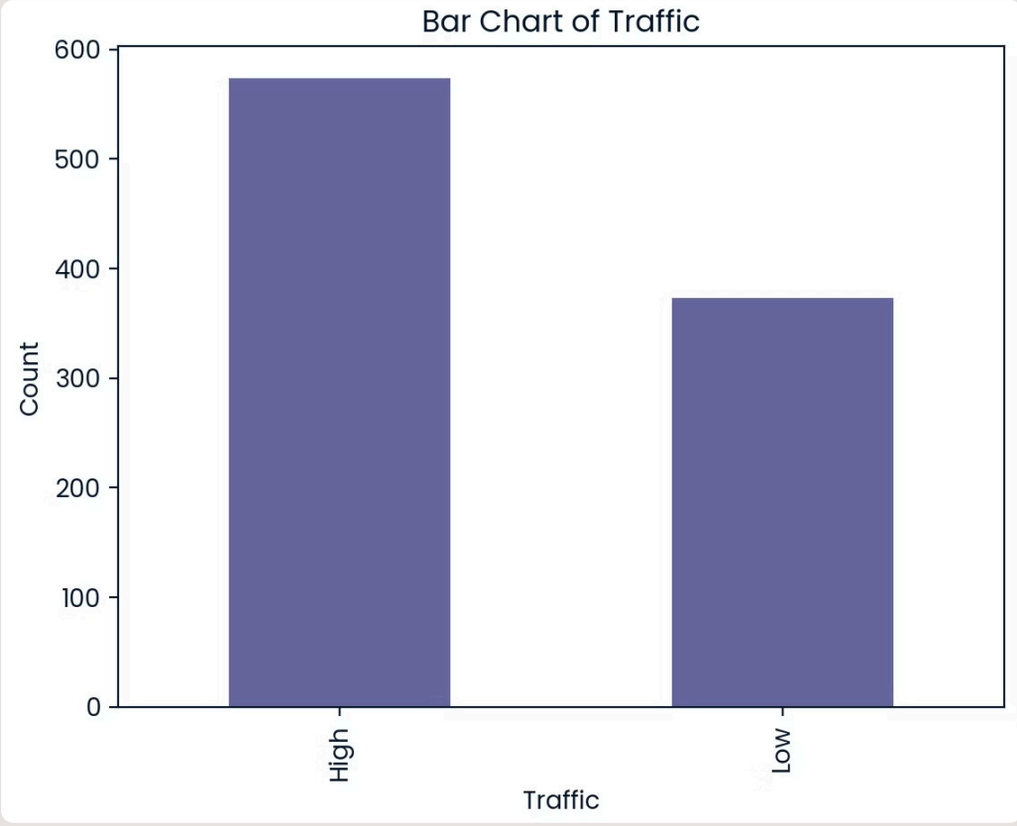
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 947 entries, 0 to 946
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   recipe        947 non-null    int64
 1   calories      947 non-null    float64
 2   carbohydrate  947 non-null    float64
 3   sugar         947 non-null    float64
 4   protein       947 non-null    float64
 5   category      947 non-null    object
 6   servings      947 non-null    int64
 7   high_traffic  947 non-null    object
dtypes: float64(4), int64(2), object(2)
memory usage: 59.3+ KB
```

EDA

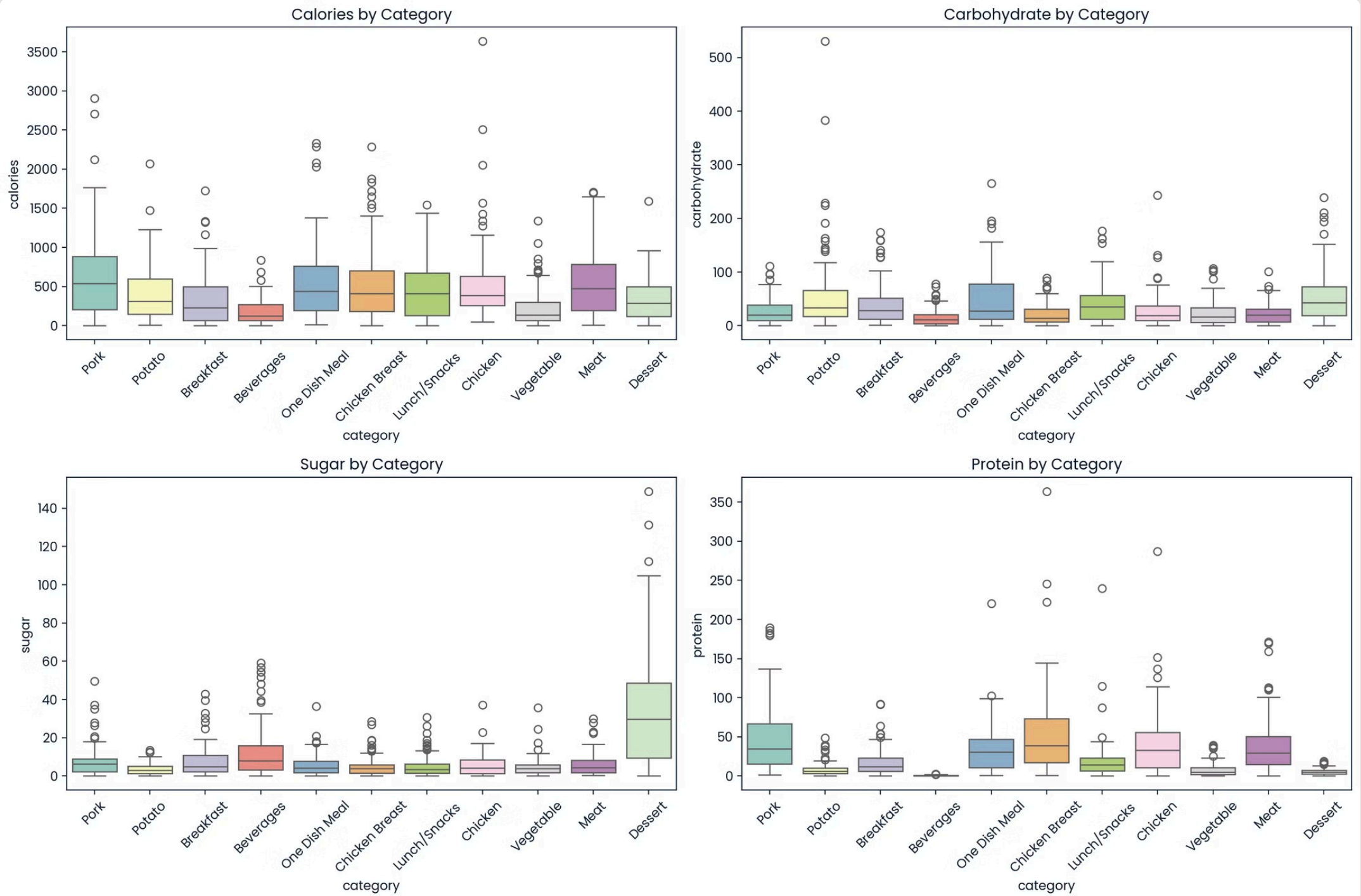
- Category Distribution: Breakfast and chicken breast recipes are the most frequent.
- 'High' traffic recipes are more than 'low' ones.





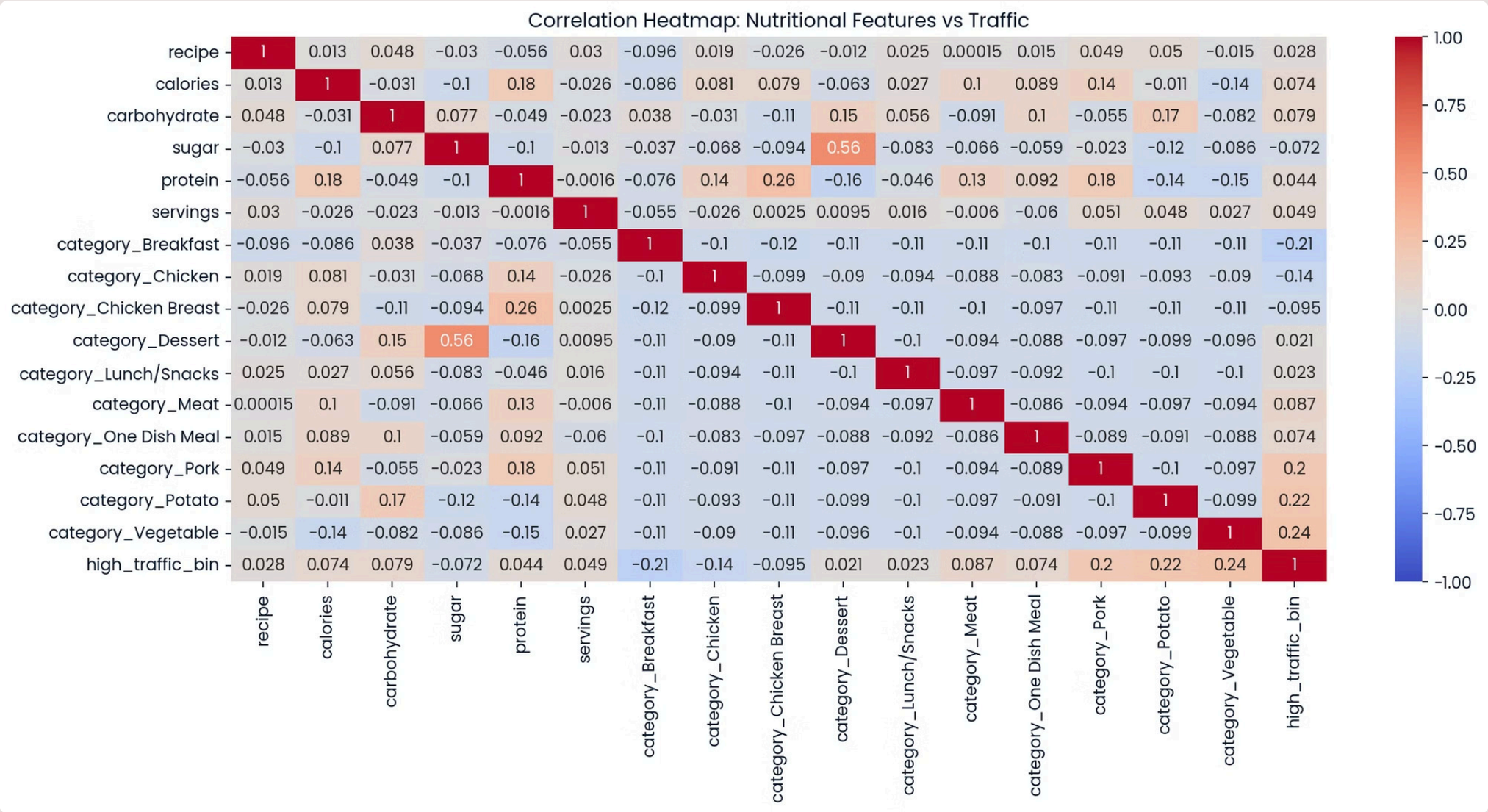
EDA

- Certain nutrients are associated with specific dish categories



EDA

- There's no single variable showing a great correlation to the high traffic.
- Thus, we need to look into the nutrients as a combination and its relations with category and high traffic.



EDA — nutrients & traffic

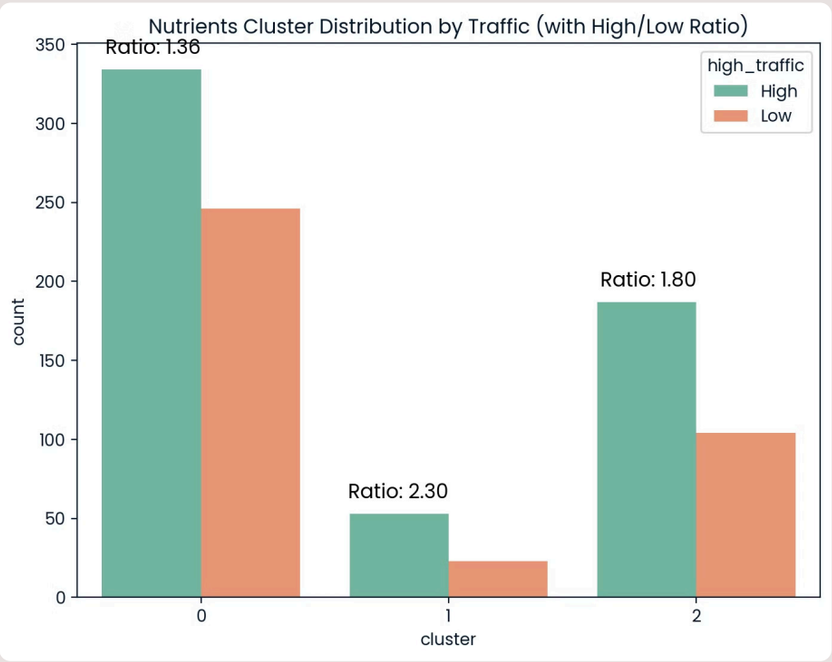
- Cluster using the nutrients:

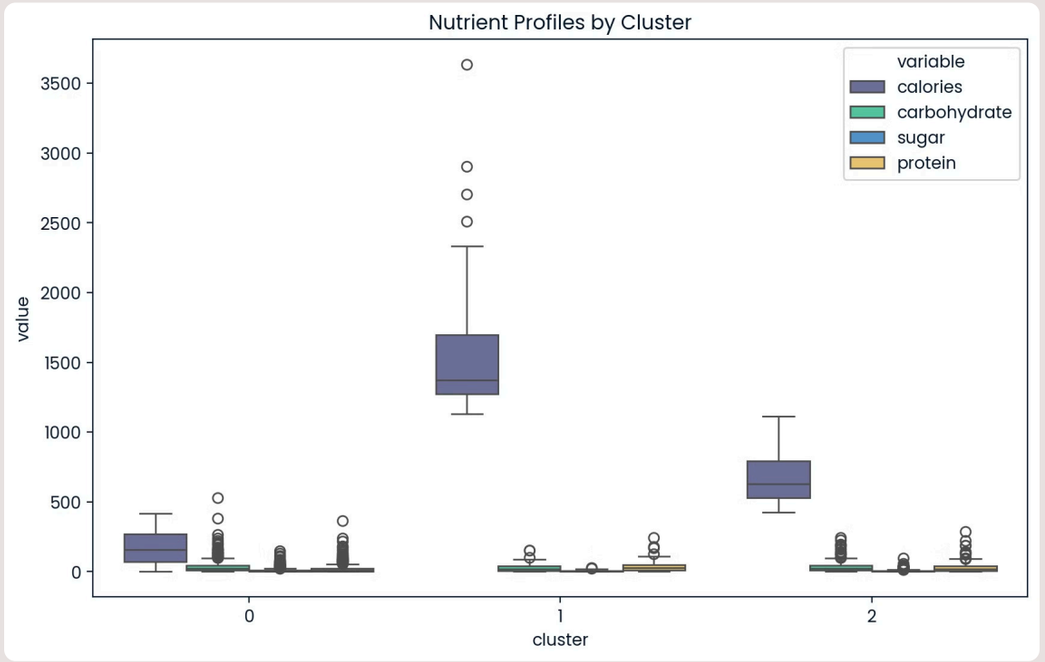
  Cluster 0: Likely represents everyday, balanced recipes; could target health-conscious users.

  Cluster 1: Represents rich, indulgent recipes; might attract users seeking treats or special occasions.

  Cluster 2: Suggests protein-focused or main meals; could appeal to those seeking substantial or protein-heavy options.

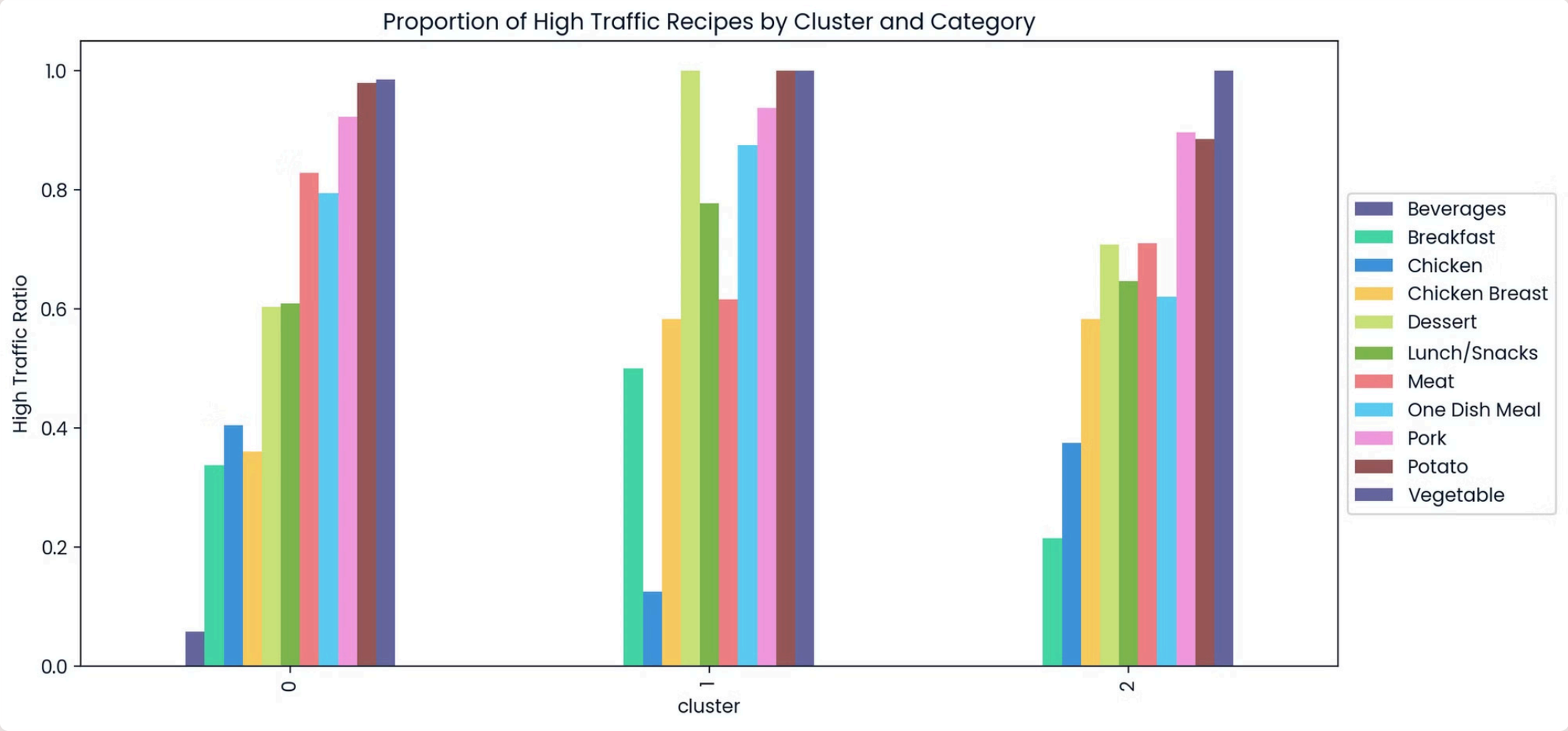
- Cluster 1 and cluster 2 are more likely to related to high_traffic.
- Cluster 1 and cluster 2 are calory-heavy and protein-heavy foods.





EDA — nutrients & category & traffic

- The Chi-squared test shows that most categories are statistically significant with traffic, such as vegetable, potato, breakfast, pork, chicken, meat, etc. The p-value of these categories are way less than 0.05.



Model Development: Binary Classification

1

Problem Type

Binary Classification

2

Models

Logistic Regression: a simple and linear model estimates the probability of an outcome.

Decision Tree: captures non-linear relationships and handles feature interactions

Random Forest: more accurate & robust than single tree

3

Goal

Recall > 80% for High Traffic (Recall for High Traffic Recipes = Correctly predicting a high-traffic recipe / Total actual high-traffic recipes)

Model Evaluation and Comparison

1

Metrics

Recall, Precision,

F1-Score, Accuracy

2

Models

Logistic Regression

Decision Tree

Random Forest

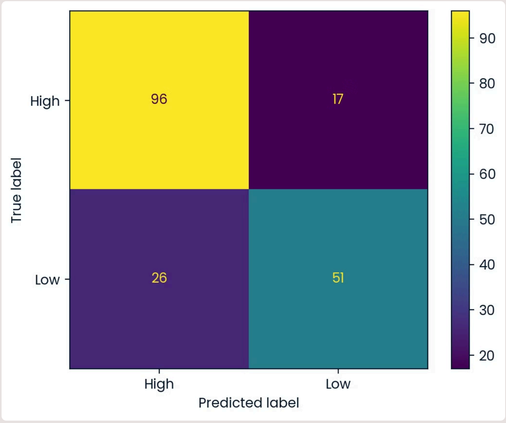
3

Conclusion

LR provides the best balance

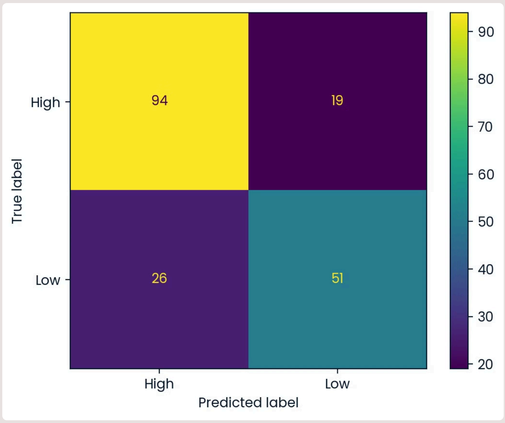
Metrics of LR

	precision	recall	f1-score	support
High	0.79	0.85	0.82	113
Low	0.75	0.66	0.70	77
accuracy			0.77	190
macro avg	0.77	0.76	0.76	190
weighted avg	0.77	0.77	0.77	190



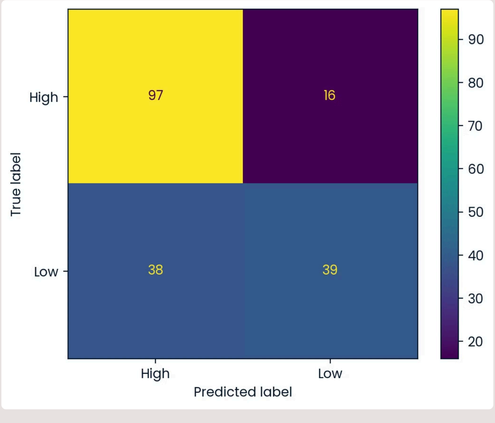
Metrics of DT

	precision	recall	f1-score	support
High	0.78	0.83	0.81	113
Low	0.73	0.66	0.69	77
accuracy			0.76	190
macro avg	0.76	0.75	0.75	190
weighted avg	0.76	0.76	0.76	190



Metrics of RF

	precision	recall	f1-score	support
High	0.72	0.86	0.78	113
Low	0.71	0.51	0.59	77
accuracy			0.72	190
macro avg	0.71	0.68	0.69	190
weighted avg	0.71	0.72	0.70	190



Business Metric

Business Goal

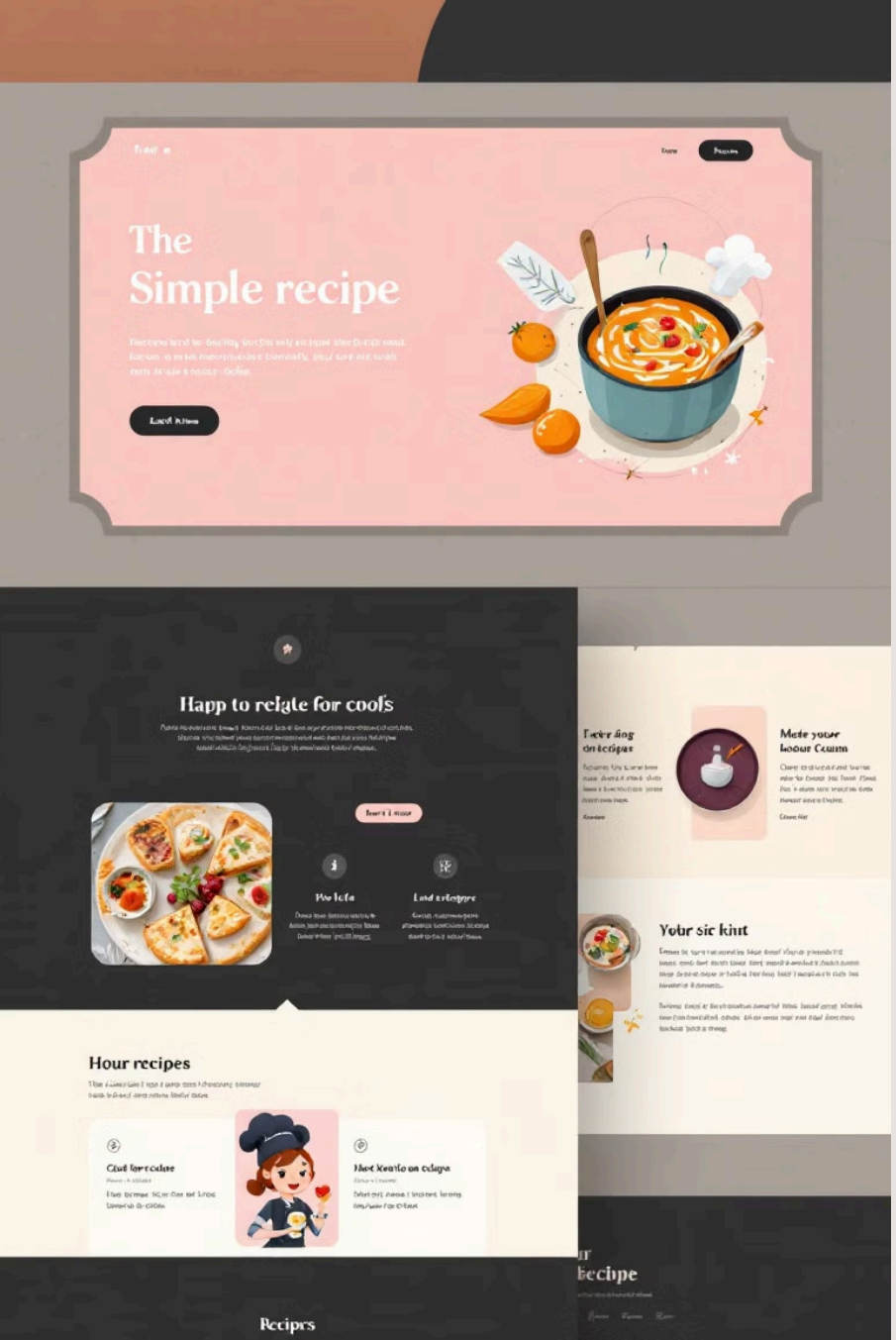
Display popular recipes

Metric

Recall for High Traffic Recipes

Value

85% (exceeds 80% target)



Final Summary and Recommendations

- **Logistic regression** is recommended for deployment, achieving 85% recall while maintaining reasonable precision.
- Prioritize high-traffic recipe types like **Meat, Chicken, and high calory-heavy and high protein** foods.
- Prefer recipes with **four or more servings** (family-sized).
- Monitor performance weekly and enhance model features with user ratings.
- Periodically retraining the model