Tasty Bytes

A K-Means Clustering Approach

Jalal Kaddoura

27 June, 2024

Project Overview & Business Goal

Tasty Bytes aims to understand its vast recipe catalog to better serve users and optimize content strategy.

Original Goal: Correctly predict high traffic recipes 80% of the time.

The Business Challenge: Predicting Popular Recipes

So, how can we identify which recipes will lead to high traffic to display on the homepage?

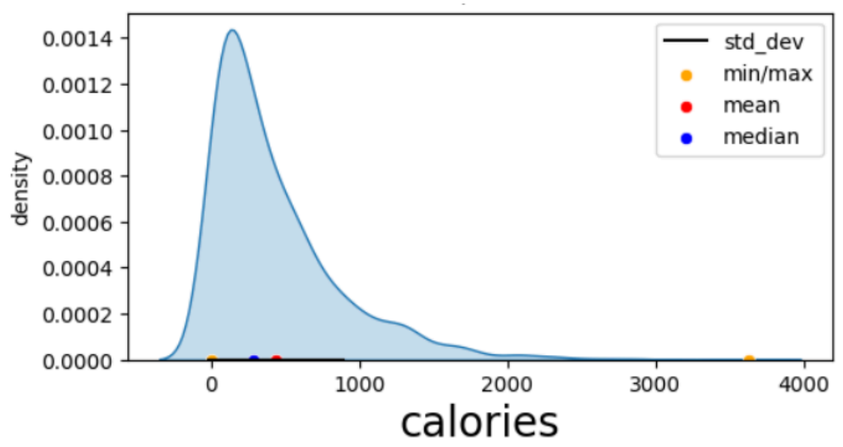
While direct 80% prediction is challenging , I focused on segmenting recipes to understand the characteristics of high-traffic recipes, providing foundational insights for future predictive efforts and immediate content strategy.

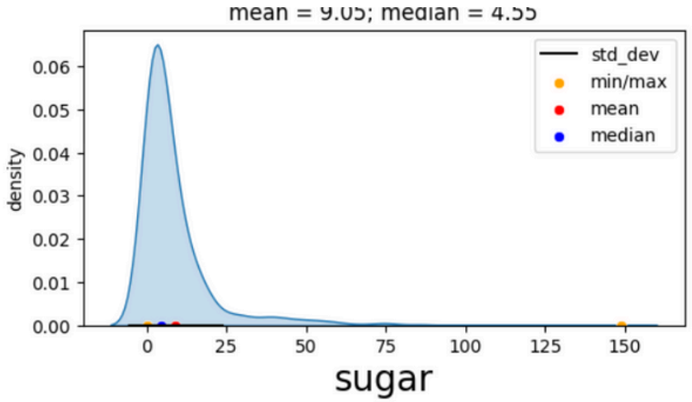
Data & Preparation Highlights

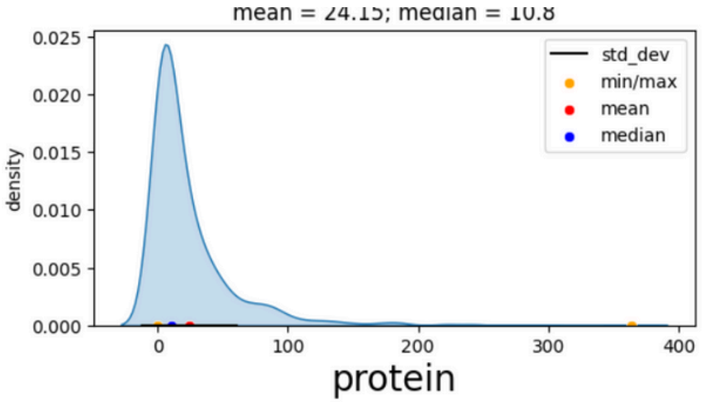
The Porvided Datset: recipe_site_traffic_2212.csv – containing nutritional info, category, servings, and high_traffic status for the amount of recipes.

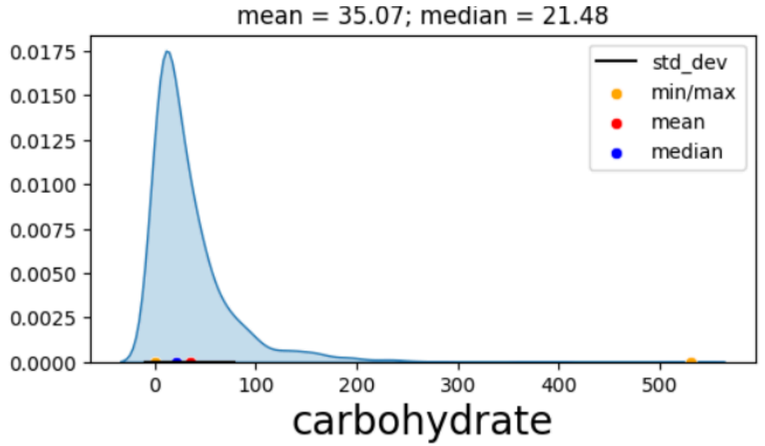
Key Features Used: Calories, Carbohydrates, Sugar, Protein, Category, Servings, High Traffic.

Initial Observations (Univariate Analysis Highlights):





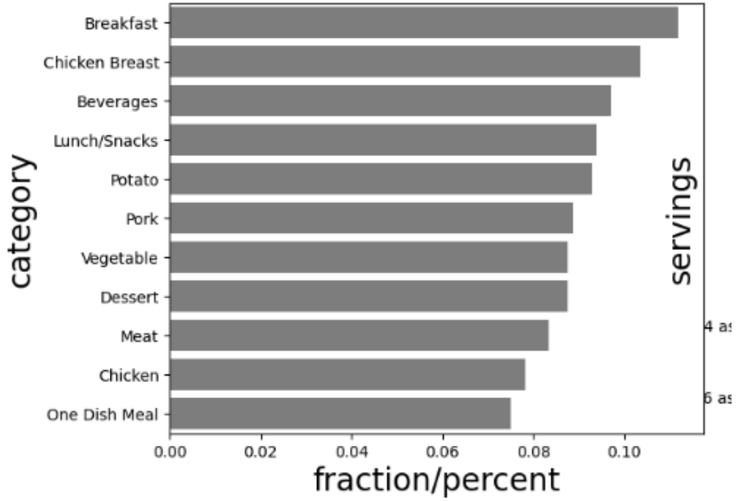




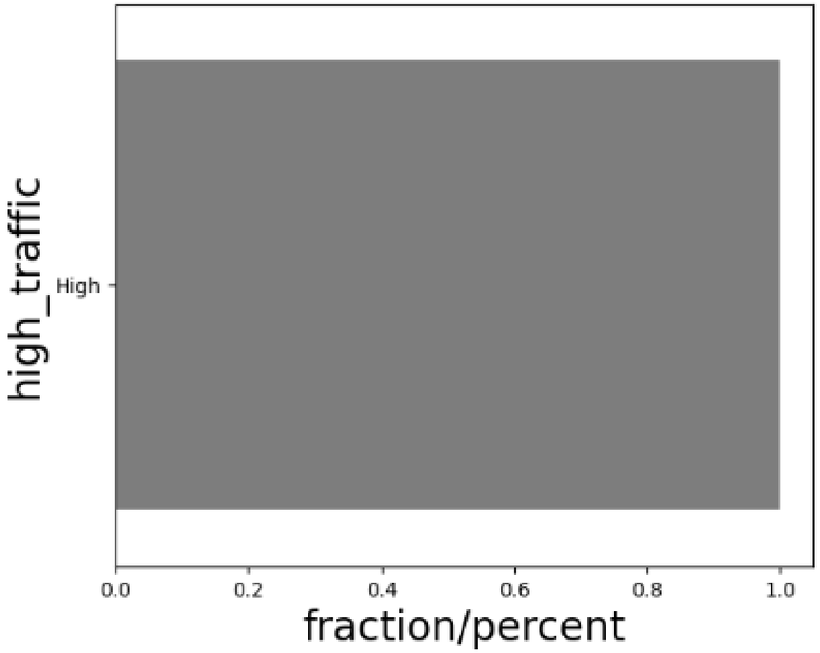
Data & Preparation Highlights

Initial Observations (Univariate Analysis Highlights):

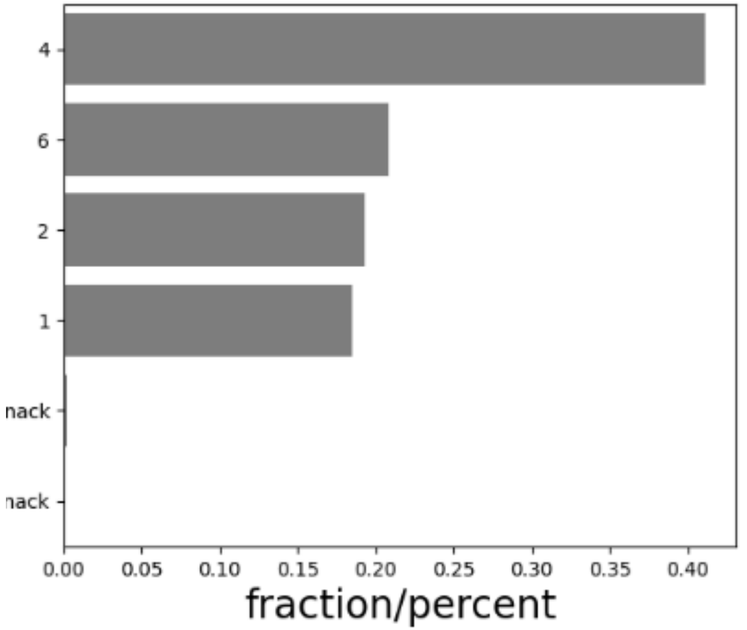
n_uniques = 11
value counts
category
Breakfast 0.111932
Chicken Breast 0.103485
Beverages 0.097149
Lunch/Snacks 0.093981
Potato 0.092925
Pork 0.088701
Vegetable 0.087645
Dessert 0.087645
Meat 0.083421
Chicken 0.078141
One Dish Meal 0.074974
Name: proportion, dtype: float64;



n_uniques = 1
value counts
high_traffic
High 1.0
Name: proportion, dtype: float64;



n_uniques = 6
value counts
servings
4 0.410771
6 0.208025
2 0.193242
1 0.184794
4 as a snack 0.002112
6 as a snack 0.001056
Name: proportion, dtype: float64;



Data & Preparation Highlights

What Was Concluded and Applied:

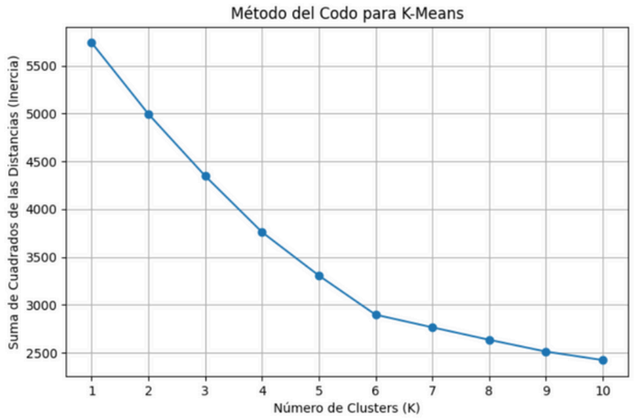
- Imputed missing nutritional values with the median and replaced missing high_traffic values with 'No_Traffic_Info' (assuming they are not high traffic).
- Converted servings into a numerical format for analysis.
- Applied StandardScaler to numerical features and One-Hot Encoding to categorical features to ensure fair contribution to clustering.

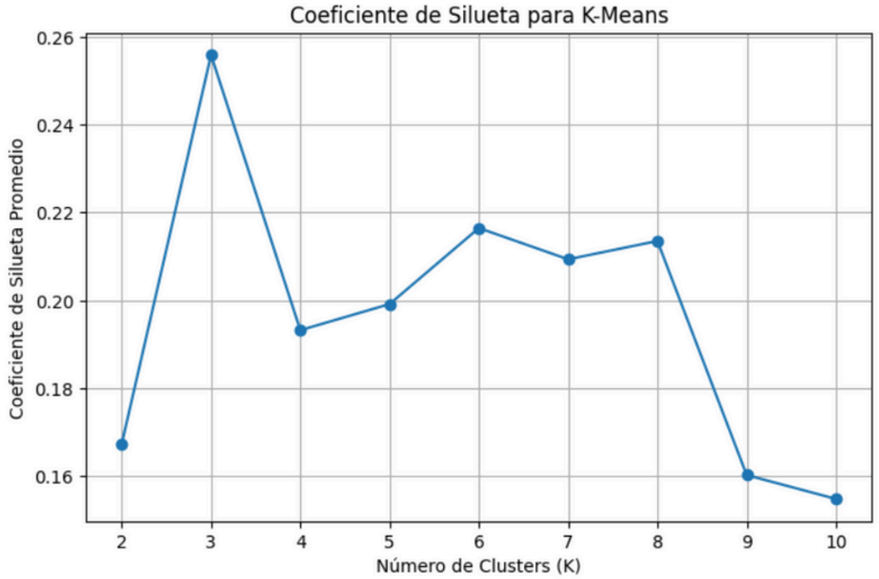
Methodology: Uncovering Recipe Segments with K-Means

K-Means clustering was employed to group recipes based on their inherent similarities across various features.

- Determining Optimal Segments (K=4):
- Elbow Method: Showed a clear "bend" or "elbow" at K=4, indicating diminishing returns beyond this point.
- Silhouette Analysis: Supported K=4 with a good average silhouette score, suggesting well-defined and separated clusters.

- **Outcome:** 4 distinct and interpretable recipe segments were identified.





Overview of the 4 Recipe Segments

Recipes naturally fall into 4 distinct groups, each with unique nutritional, categorical, and traffic profiles.

Cluster	Recipes
0	493
1	50
2	303
3	101

Overview of the 4 Recipe Segments

Recipes naturally fall into 4 distinct groups, each with unique nutritional, categorical, and traffic profiles.

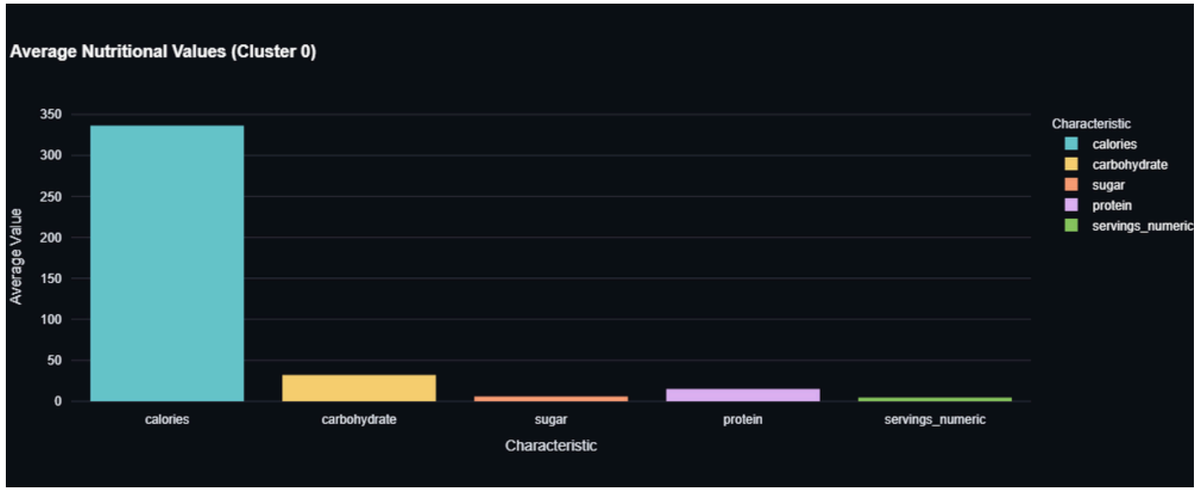
Cluster	Recipes
0	493
1	50
2	303
3	101

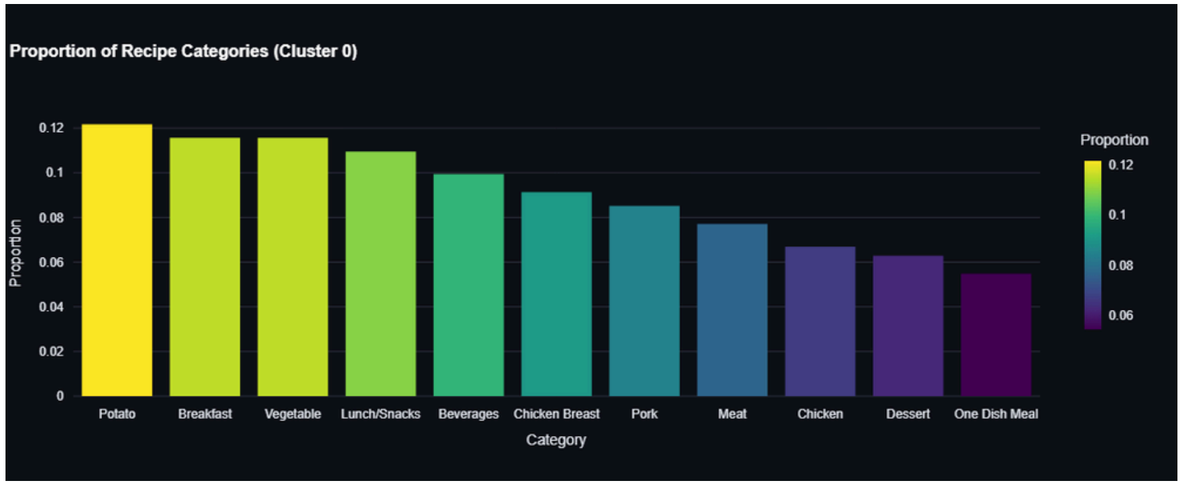
The PCA plot demonstrates good visual separation among the identified clusters in a reduced 2D space.

The clusters vary in size, reflecting the natural distribution of recipe types within our catalog.

Segment 0

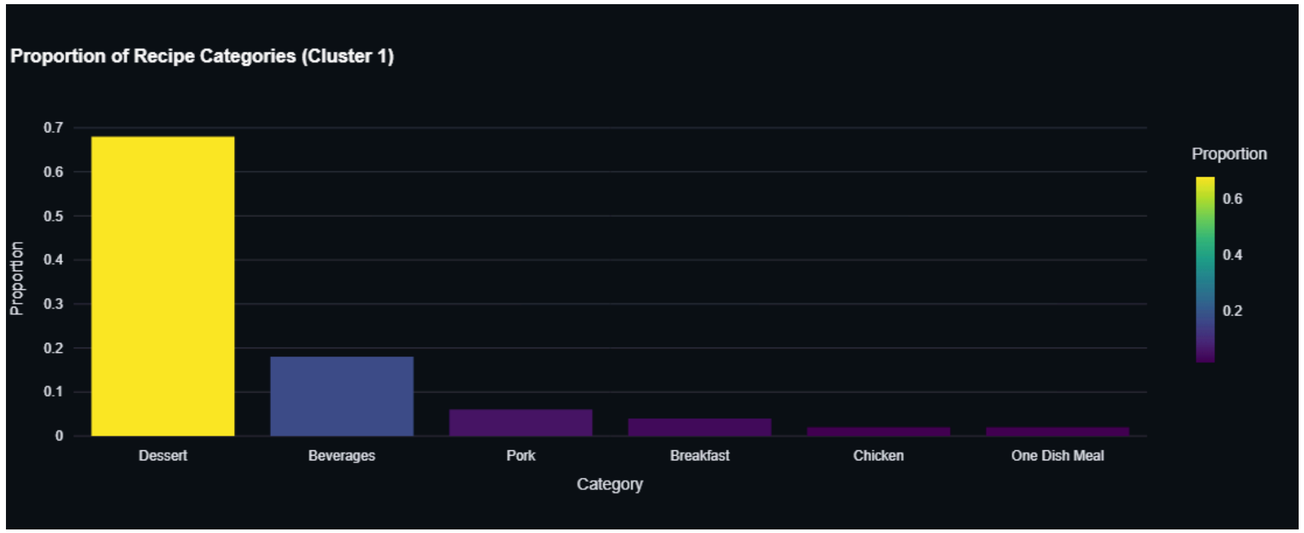
Recipes naturally fall into 4 distinct groups, each with unique nutritional, categorical, and traffic profiles.

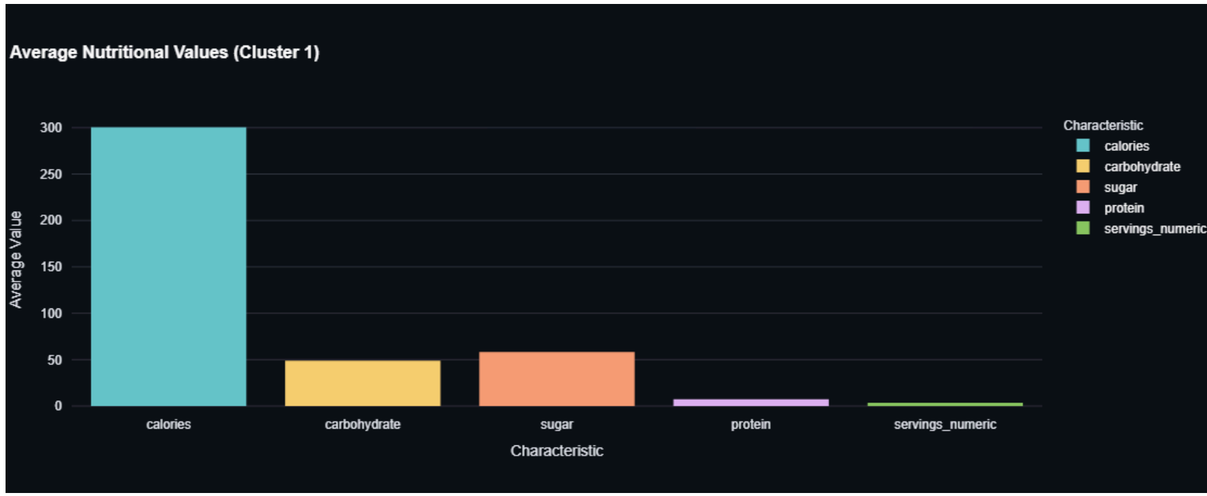




This segment represents the largest group of recipes, characterized by a balanced nutritional profile and suitability for standard family meals.

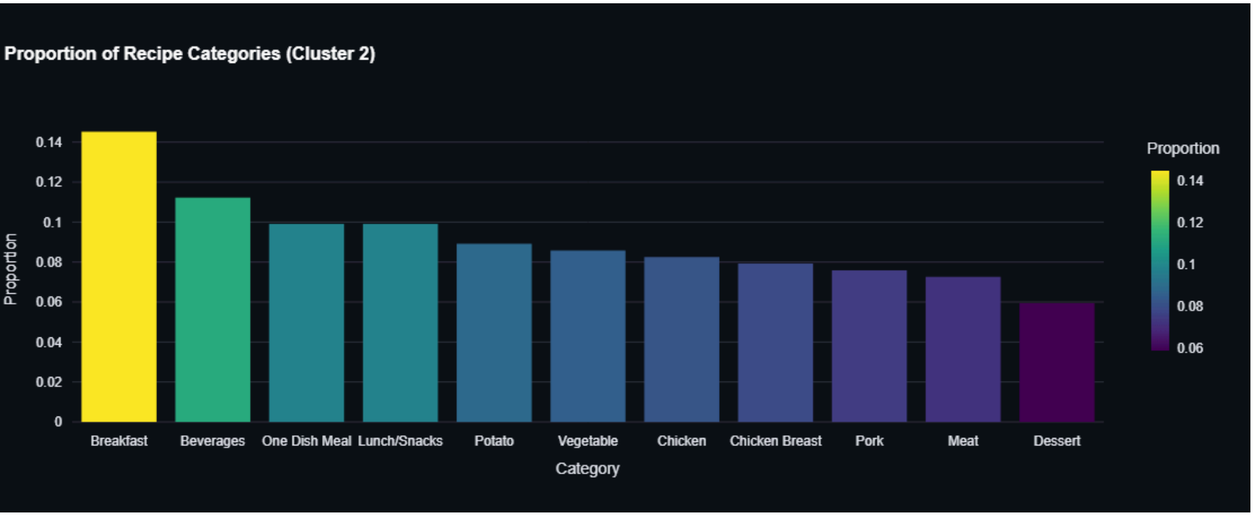
Segment 1

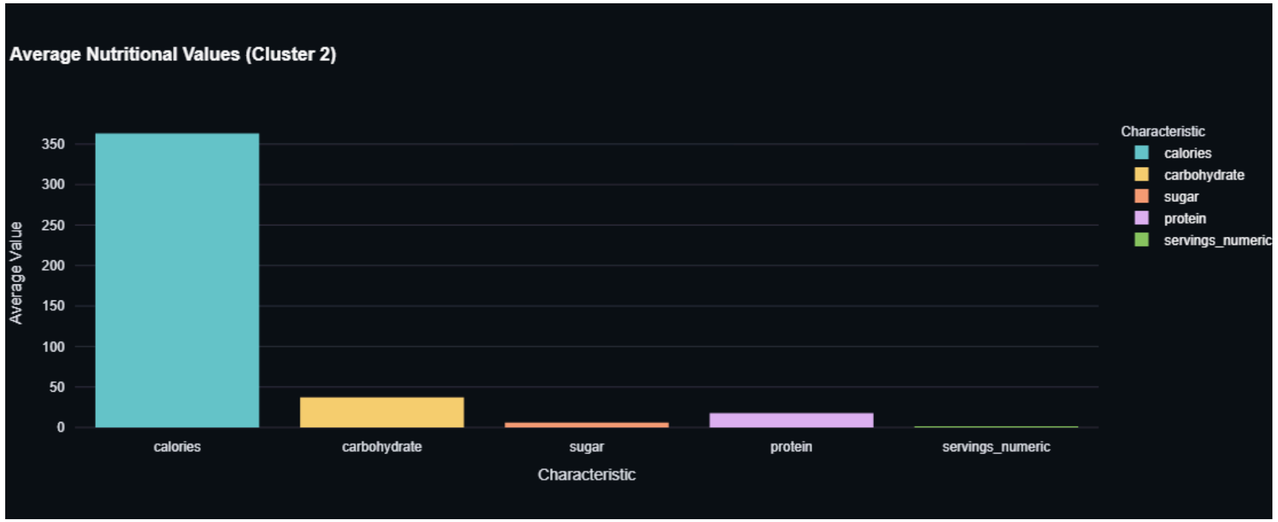




This smaller segment is distinctly defined by its high sugar content, primarily comprising desserts and sweet beverages.

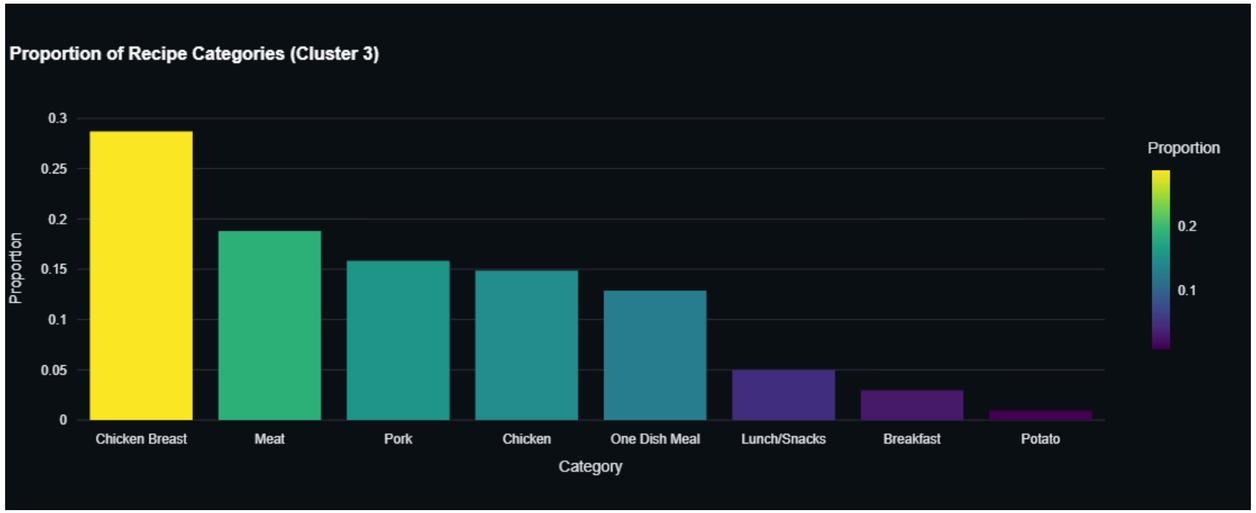
Segment 2

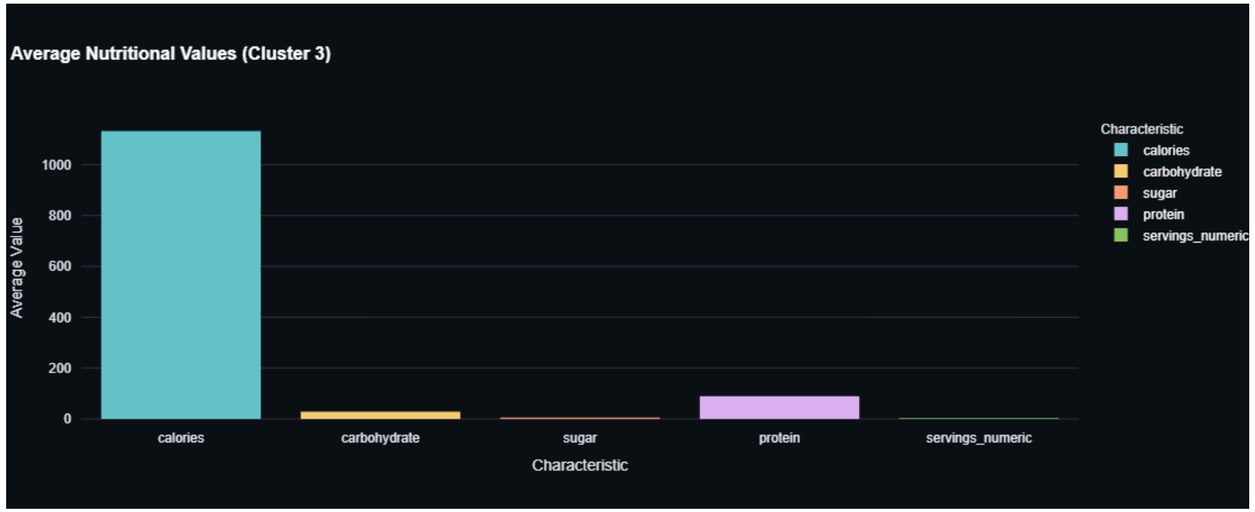




This segment groups recipes designed for single or small portions, often suitable for quick meals or snacks.

Segment 3





This segment features recipes with the highest protein and calorie content, typically robust main dishes centered around meat and poultry.

Traffic Analyisis

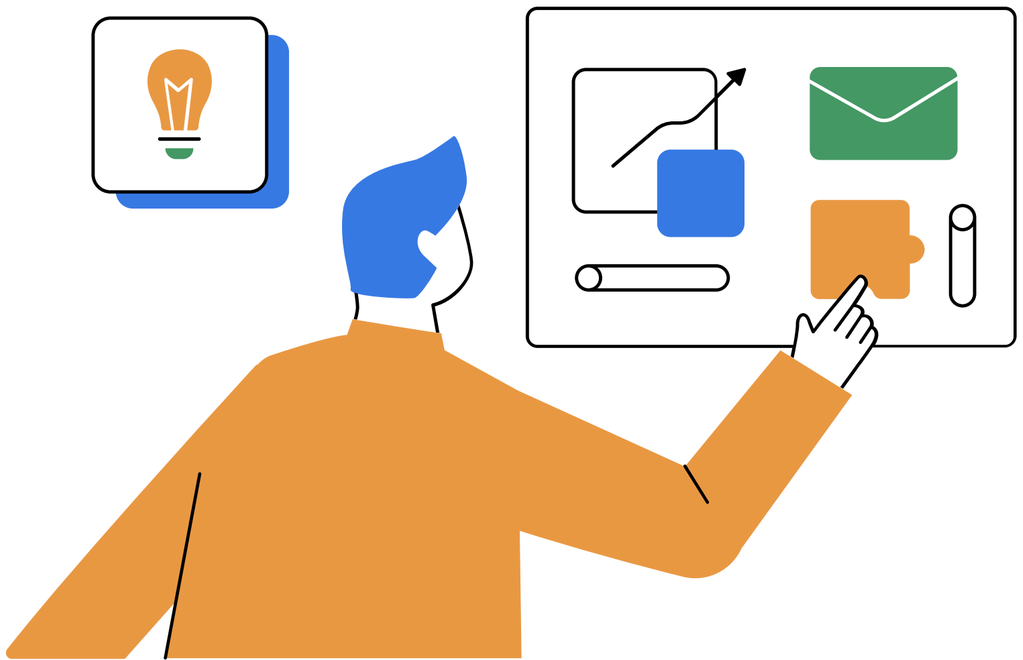
Analyzing the high_traffic variable, which is crucial for our business objective, we found that the original dataset only provided explicit data for 'High' traffic. A significant number of records simply lacked this information (NaN). There was no explicit 'low traffic' label.

Given the business's focus on identifying high-performing recipes and the absence of a distinct 'low traffic' category, we decided to interpret and treat these NaN values as 'No_Traffic_Info'. This approach created a binary benchmark, allowing us to clearly differentiate recipes proven to be popular from those for which we have no high-performance evidence, which was essential for our segmentation analysis.

But...

Analyzing the high_traffic variable within this segment, we observe that the distribution between 'High Traffic' recipes and those with no traffic information (NaN) is consistently similar to the overall dataset's proportion. Approximately [60% - 63%] of the recipes in this cluster show 'High Traffic', while the remaining [37% - 48%] corresponds to values without information. This suggests that the clustering process has grouped recipes by other distinctive characteristics (such as ingredients, nutrition, or servings), but it has not isolated segments that are inherently more or less prone to high traffic than what is observed in the general dataset.

Results and Conclusion



The project successfully delivered a robust solution for identifying high-traffic recipes. While the clustering analysis provided valuable segmentation based on recipe characteristics, it was the SVC prediction model that directly achieved the key business metric: predicting high-traffic recipes with over 80% precision. This empowers the Product Manager to make data-driven decisions on which recipes to highlight on the homepage, minimizing the chance of featuring unpopular content.