# Waitoa Fusion: Custom Segmentation & Opportunity Scoring Methodology

## Overview

This document outlines the methodology and technical implementation of the Waitoa Fusion project, a data science pipeline developed to identify high-opportunity supermarkets for the Waitoa Chicken brand in New Zealand.

The goal was to create a data-driven targeting framework using store-level sales data, demographic overlays, and geospatial clustering to prioritize digital media spend and campaign focus areas.

## 1. Data Sources

The pipeline integrates multiple sources including:

- Historical category sales data (Butchery/Fresh, Frozen, Deli) via Dunnhumby (2024)

- Monthly sales updates from Excel files (2025)

- New Zealand supermarket master list with geocoordinates

- Stats NZ population data and SA2-level geographic shapefiles (2023-2025)

All files are ingested from Google Cloud Storage (GCS) buckets using the `google.cloud.storage` library.

## 2. Data Preparation

Each product category is loaded separately and cleaned by:

- Renaming ambiguous columns

- Filling missing values in sales/volume with 0

- Aggregating over all time-specific columns to produce `TOTAL\_SALES` and `TOTAL\_VOLUME`

## 3. Product Segmentation

Products are split into two major groups based on their names:

- Waitoa products (identified by substring match on 'WAITOA')

- Competitor products (everything else)

This allows sales and volume to be calculated separately for both groups per store.

## 4. Store Matching

The supermarket names from Waitoa sales data are not always cleanly aligned with the master supermarket list.

To resolve this, fuzzy string matching is applied using `fuzzywuzzy`, scoring matches and selecting the closest match with a threshold.

Successfully matched records are then merged with store-level attributes such as geocoordinates, addresses, and town.

## 5. Market Share Calculation

For each product category (Fresh, Frozen, Deli), we calculate:

- Waitoa sales and volume per store

- Competitor sales and volume per store

- Market share = Waitoa / (Waitoa + Competitors)

These share metrics are then merged with SA2-level population data to provide local catchment context.

## 6. Opportunity Scoring

A custom scoring mechanism was implemented based on:

- `opp\_sales`: Competitor sales minus Waitoa sales (floored at 0)

- `opp\_volume`: Competitor volume minus Waitoa volume

- Local SA2 `population`

Each factor is min-max normalized and combined using weighted coefficients:

- 50% weight on normalized sales gap

- 30% weight on normalized volume gap

- 20% weight on local population

Each store receives a final `opportunity\_score` and is ranked accordingly.

## 7. Clustering

Using the top 100 Waitoa stores by opportunity score in each category, K-Means clustering is applied on their geographic coordinates.

Key parameters:

- Cluster count (55–75) per category

- Dynamic radius computation based on max geodesic distance from centroid

- Radius constraints: minimum 4km, maximum up to 25km

Folium is used to plot interactive maps with cluster centroids, coverage circles, and store pins.

## 8. Cluster Expansion

To identify white-space or growth zones, the system also:

- Plots all supermarkets that are **not currently Waitoa** within cluster radii

- These “non-Fresh”, “non-Frozen”, or “non-Deli” stores are tagged for possible acquisition

This mechanism helps identify adjacent stores with similar geographic or demographic characteristics.

## 9. Deliverables

The pipeline outputs:

- Top 100 cluster summaries by product category (`waitoa\_top\_100\_stores\_analysis.xlsx`)

- Clusters for all other stores (`waitoa\_others\_stores\_analysis.xlsx`)

- Interactive HTML maps for top and other stores (Fresh, Frozen, Deli)

Each map displays clusters with hoverable tooltips, popup store details, and colored pin markers.

## 10. Extensibility

The code is modular and can be extended in several ways:

- Add new sales years or product categories

- Adjust opportunity scoring weights

- Modify clustering logic (e.g., use KMeans)

- Overlay media spend or campaign performance metrics for feedback loops

- Extend SA2 or meshblock geospatial enrichment

This project is intended for internal campaign planning and performance optimization.