Project Proposal
Restaraunt Segmentation Analysis

MDS Students: Chen Lin, Eric Tsai, Morris Zhao, Xinru Lu
Capstone Partner: Sitewise Analytics
Mentor: Gittu George

2023-05-12

Contents

Executive Summary

The Restaurant Segmentation Analysis project, in collaboration with Sitewise Analytics, a SaaS company specializing in building a strategic real estate road map for restaurant owners, aims to use machine learning to identify key factors driving traffic to a specific store location and recognize patterns among similar locations. In the following proposal, we will first highlight the problem and fundamental goals of the project, introduce the data and discuss data science techniques that we will employ to tackle the problem, and conclude with a rough timeline of the project.

Introduction

To plan for future expansions or market the new store strategically, restaurant franchise owners need to know the main factors that drive traffic to a location as well as the major customer group, whether it is office workers in downtown or students in a neighbourhood. This involves analyzing the surrounding population demographic, consumer behaviour in the trade area and nearby competitor/sister store information. Given that Sitewise does not currently have these factors, the Restaurant Segmentation Analysis project will address this problem by using data from Smoothie King locations in the United States and Subway locations in Canada and the United States to build machine learning data pipelines for Sitewise to incorporate into their consulting service for those respective clients.

Given that Smoothie King, Subway US, and Subway Canada are all different clients of Sitewise and have different marketing strategies, it is necessary to build three separate models for each respective client. Ultimately, the factors that drive traffic to each of the three restaurants may differ and depend on many things such as a difference in customer demographic between the US and Canada. Thus, the solution will be based on the data gathered for each restaurant client. At the end of the project, we expect to have the following three machine learning pipelines:

1. A supervised machine learning pipeline using data from Smoothie King US locations to predict a store's category from one of five pre-labeled categories:
    - Home
    - Shopping
    - Work
    - Travel
    - Other

The prediction will be human-interpretable in that the key features that determine the category for a store location will be outlined for the users.

2. An unsupervised machine learning pipeline based on data of US Subway locations that cluster locations by similar features.

3. An unsupervised machine learning pipeline based on data of Canadian Subway locations that cluster locations by similar features.

The two unsupervised machine learning pipelines will also have human-interpretable results, including ways to identify similar features that caused different locations to be clustered together.

These machine learning data pipelines are expected to be incorporated into our partner's consulting services for these clients. The final product will be integrated into a GitHub repository, including the scripts for the machine learning data pipelines, reproducible results and reports, and documentation.

Data Summary

The data for this project contains three datasets for each respective customer: Smoothie King, Subway Canada, and Subway US. Each dataset consists of five CSV files for demographic, point of interest, store-specific data, competition sister store data, and trade area, where each row represents a single store location and the columns represent the variables/features of that store. All features in the demographic, point of

interest, competition sister store, and trade area files are numeric, whereas the store-specific data files contain categorical features such as state and market size.

For Smoothie King, there are over 1000 features combined for 796 stores.

For Subway US, there are over 1000 features combined for approximately 14,000 stores.

For Subway Canada, there are around 100 features combined for around 1,800 stores.

Understanding the Data

To help understand the data, we started with a selection of important categorical features and created some bar charts below to visually represent the distribution of these features in three datasets. The complete datasets can be accessed in the `data` directory of the repository.

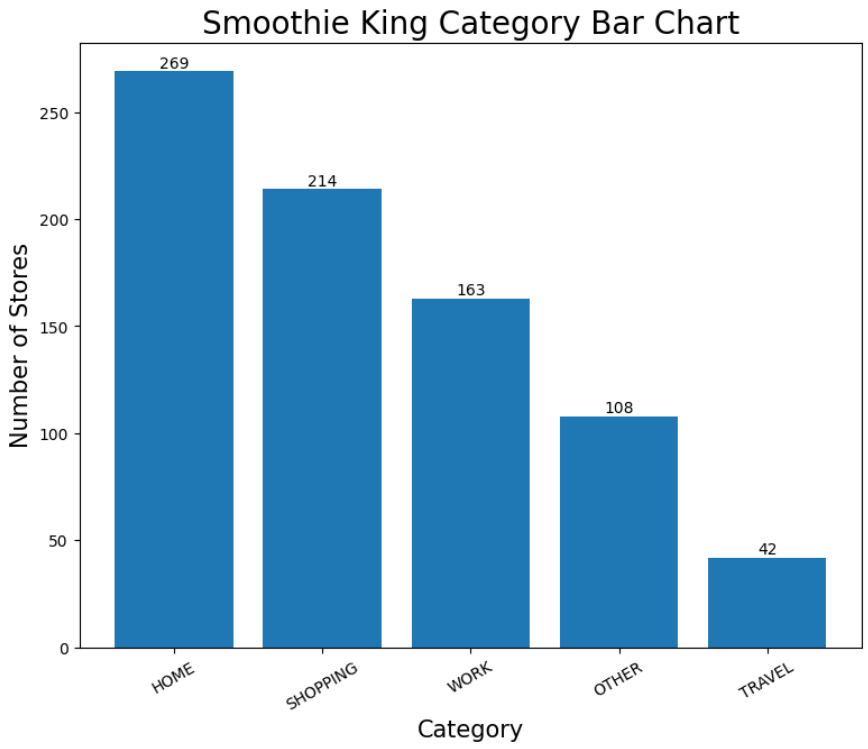

Figure 1: Number of stores in each category for Smoothie King US locations. The distribution is unbalanced.

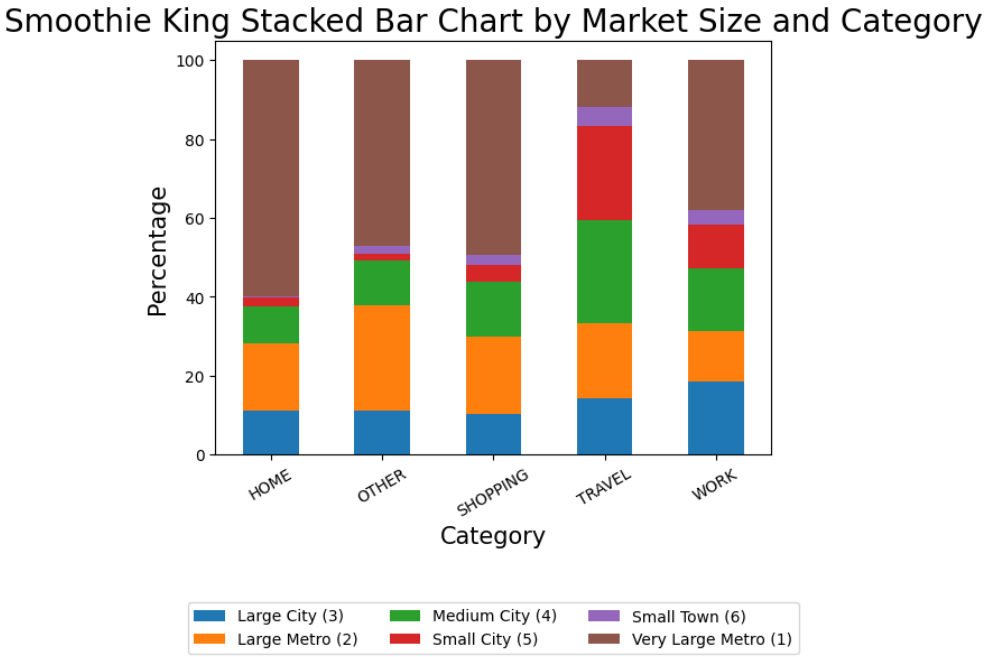

Figure 2: The highest portion among categories are "Very Large Metro". "Home" has a high percentage of "Very Large Metro". "Travel" has a high percentage of "Small City".

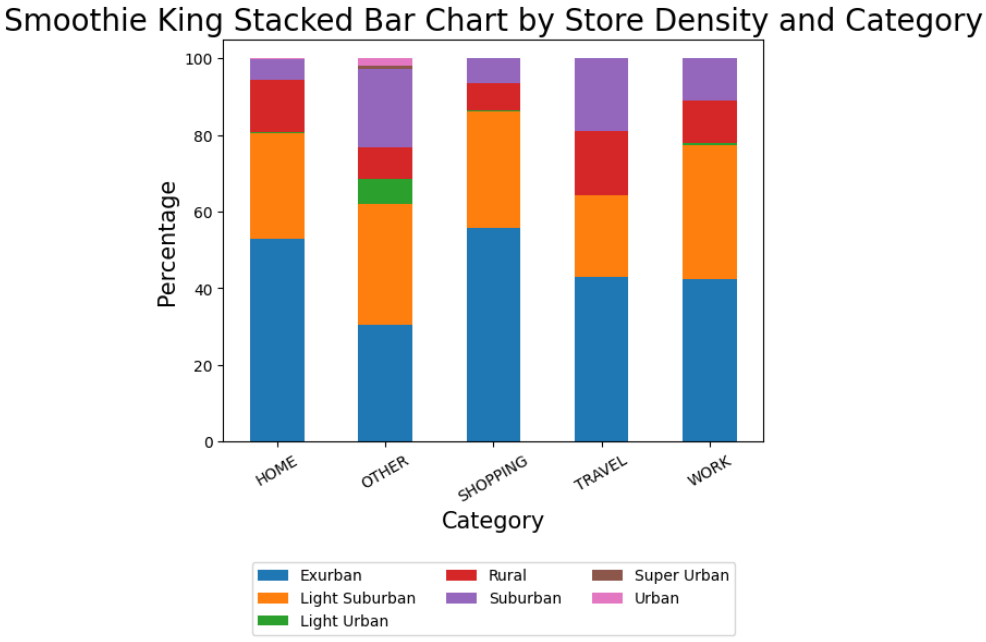

Figure 3: Most of the stores are located in "Exurban" and "Light Suburban".

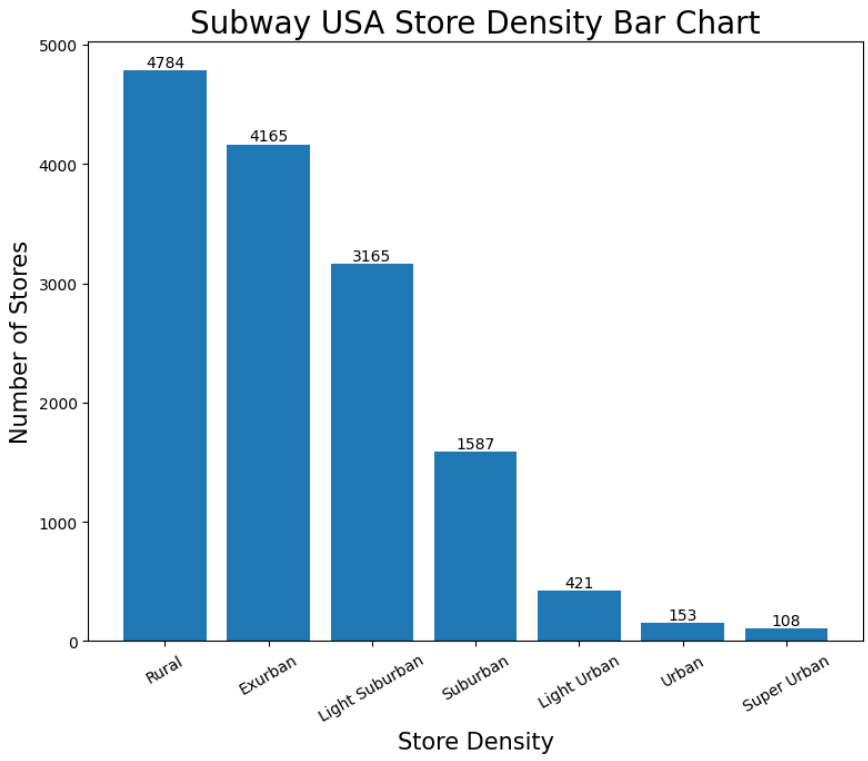

Figure 4: Rural has the highest count.

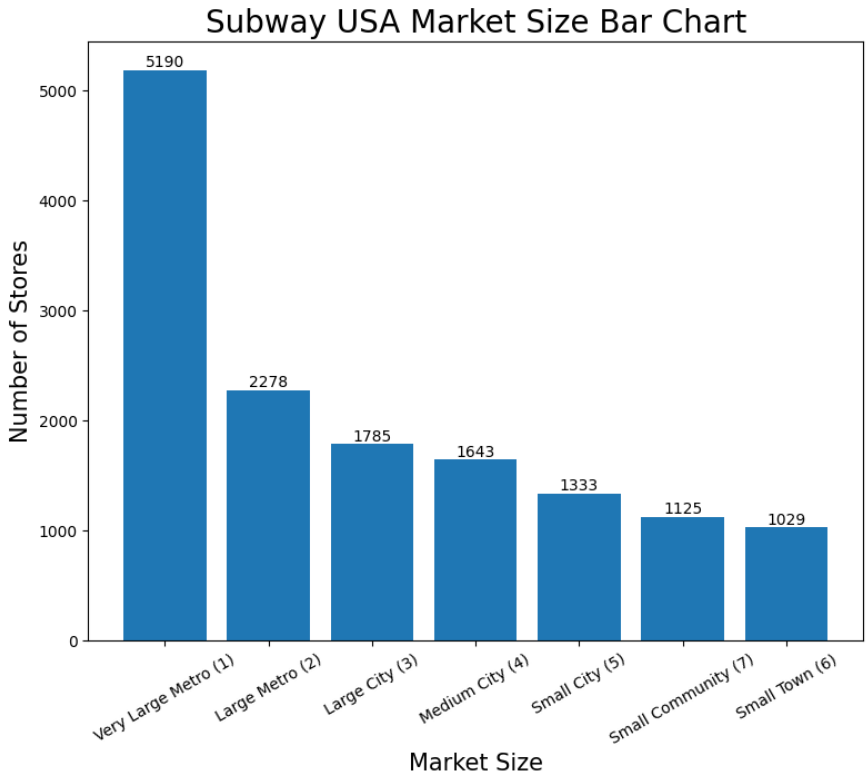

Figure 5: Very Large Metro has the highest count with 5190 stores, and other categories have similar counts.

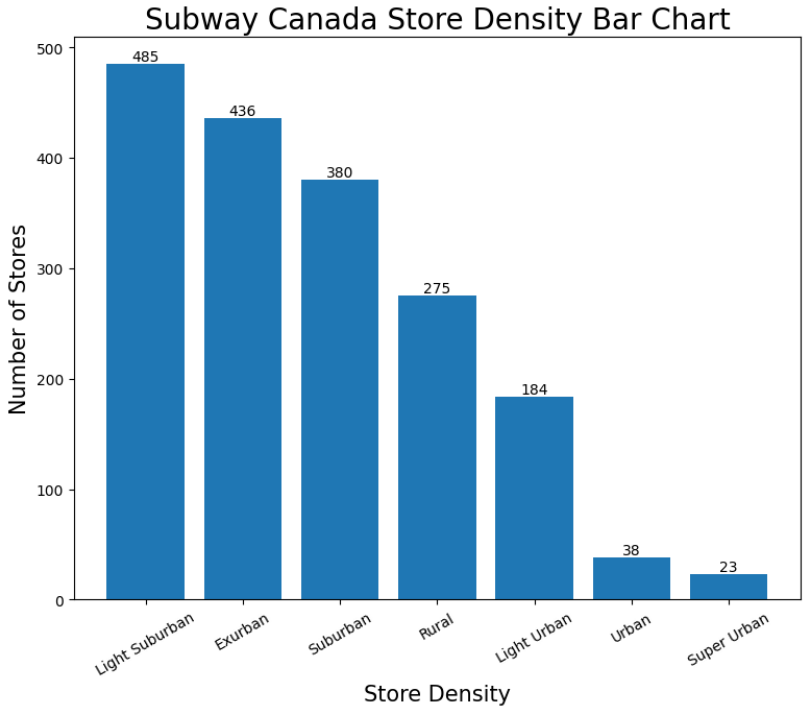

Figure 6: The most common market size is Very Large Metro with 577 stores.

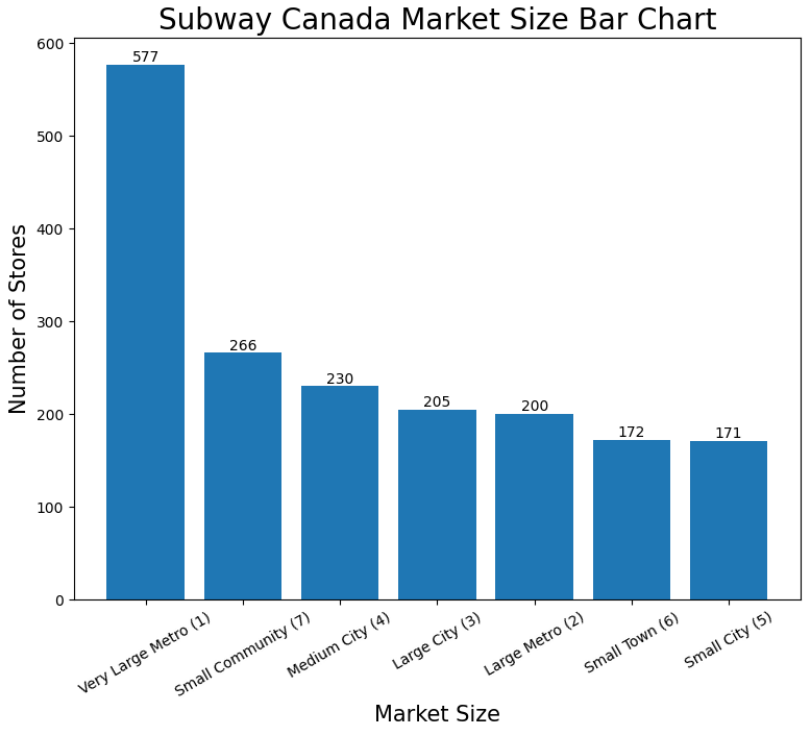

Figure 7: Very Large Metro has the highest count with 577 stores, and the other categories have similar counts.

Data Science Techniques

With three distinct datasets and the challenge of building tailored solutions for each restaurant chain client, we propose the following procedures to build a solution to the problem:

1. For Smoothie King, many features are highly correlated to one another. Therefore, it is important to start with feature selection and dimensionality reduction. Other than manually analyzing the data columns with correlation scores, we are planning to use Principal Component Analysis (PCA) to reduce the dimension and Recursive Feature Elimination to select the most important features.

   With the data and the supervised classification objective, we propose starting with a Logistic Regression model as our baseline since it is easy to interpret. We will also explore other models such as a simple Random Forest Tree regression model, as well as an LGBMClassifier as it effectively trains high-dimensional datasets and usually leads to relatively high accuracy for the multi-class problems.

   With the labeled data, the models will be evaluated by its accuracy score, with the target being 80% accuracy as outlined by Sitewise. The score would indicate if the model is capable of detecting the most important traffic driving factor to a store. The result can be interpreted with SHAP (SHapley Additive exPlanations) plots where we can visualize which features drive the decision of a particular category assignment for a certain store.

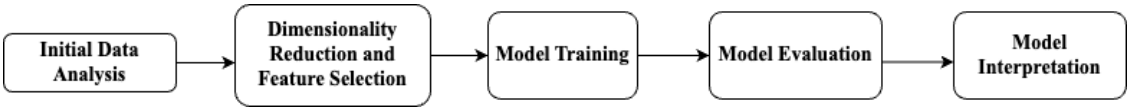

Figure 8: Flow chart of Smoothie King

2. For Subway US, the main goal is to cluster the stores into different clusters where stores that share similar features are in the same cluster. We propose starting with a DBSCAN model as a baseline and evaluate the clusters empirically. Afterwards, we will perform hierarchical clustering and apply PCA for dimensionality reduction and compare the results. The rationale for using hierarchical clustering is that we do not have to pre-define the number of clusters and it also allows unbalanced cluster sizes.

   Since both Smoothie King and Subway US are in the US and share similar features in their datasets, they could share similar segmentations to a certain extent. To perform PCA, the most driven features from the supervised model for Smoothie King can be used as a reference to evaluate this PCA step where more matched important features could indicate a more reasonable PCA result.

   For clustering, all possible linkage criteria will be tested to evaluate how to find similarities between clusters. Finally, since ground truth labels are not known, evaluation can be done through the Silhouette Coefficient where a Silhouette Coefficient closer to 1 suggests that a model defines well separated clusters.

   To interpret the results and empirically validate the clusters, we will randomly select a sample of stores in the same cluster and visualize them on Google Maps or Sitewise's internal application to check if they share similar geographic attributes (near highway exits, dense residential area, etc.).

3. For Subway Canada, we will take a similar approach to Subway US and evaluate the result.

   If the performance is not ideal, consider removing the PCA step and run clustering again to evaluate the result again with the same procedure performed on the Subway US model.

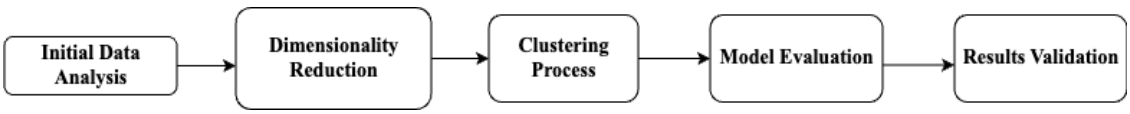

Figure 9: Flow chart of Subway US and Canada

Timeline

Schedule		Objective
Week 1 (5/1 - 5/7)	EDA & Proposal	Understand the problem, perform initial EDA on the dataset, and propose potential models and approaches to each objective.
Week 2 (5/8 - 5/14)	Feature Selection	Explore a variety of methods to determine the most important features for each of the three datasets.
Week 3 (5/15 - 5/21)	Supervised Model (Smoothie King)	Train and test a supervised classification model on the labeled dataset, as well as a list of the most important features as major indicators.
Week 4 (5/22 - 5/28)	Unsupervised Model (Subway US)	Train an unsupervised clustering model on one of the unlabeled datasets and apply several potential evaluation metrics.
Week 5 (5/29 - 6/4)	Unsupervised Model (Subway Canada)	Train an unsupervised clustering model on one of the unlabeled datasets and apply several potential evaluation metrics.
Week 6 (6/5 - 6/11)	Models Tuning	Perform parameter tuning and optimization on the three models.
Week 7 (6/12 - 6/18)	Final Presentation	Present the final models as well as the list of important indicators from the list of features. Address potential directions and approaches for further study.
Week 8 (6/19 - 6/25)	Final Product	Draft submission for the final product and report and iterate on feedback before final submission.