MARKET SEGMENTATION | PYTHON | TEAM AWESOME

Where to open a cafe in Toronto?

Toronto boasts a diverse array of neighborhoods, each with its own unique charm and characteristics.

The goal was to identify a neighborhood that provides optimal business opportunities, a robust customer base, and minimal competition.

OBJECTIVE

The objective was to analyze the distribution of markets, assess the existing competition, evaluate the population density and median household income within each neighborhood.

How can I find reliable sources for population and income data?

Which API service best meets the needs?

Which statistical analysis method is best suited for analyzing categorical data?

PROJECT OVERVIEW

Data sources

3

weighed & scrutinized

Wikipedia
Statistics Canada
Google Nearby Places API

Processes

5,000,000+

Lines of data processed

Web scraped
Read, wrote csv
Cleaned, Transformed
API calls
Data models pandas
One Hot encoding
K Means
Plots, Interpretation

Effort

80+

hours of learning

Reading HTML
Evaluating and confirming data
Writing more functions
modeling data
one hot encoding
Silhouette score
k-clusters
Data analysis mindset

THE RECOMMENDATIONS

Rouge Hill/Port Union/Highland Creek

This neighborhood in **Scarborough** experiences high foot traffic due to the presence of numerous transit stations, offices, and gas stations, all of which can significantly boost sales for a cafe.

Population
35,642
Median Household Income
$ 109,785

Lawrence Park

This neighborhood in **Central Toronto** enjoys high foot traffic due to the numerous businesses, offices, clinics, and shopping plazas, all of which is good for cafe sales.

Population
16,058
Median Household Income
$ 137,758

Milliken / Agincourt North / Steeles East

This neighborhood in **Scarborough** experiences high foot traffic due to the presence of numerous transit stations, offices, shopping plazas, churches, all of which these will help the sales of a cafe.

Population
50,825
Median Household Income
$ 64,576

Median Per postal code Population	**- 26128**	census year 2021
Median Per postal code Income	**- $65508**	census year 2021
Median Per postal code Count of cafes	**- 1**	Google Nearby Places

METHODOLOGY

Approach

A commentary, describing the steps enabling the analysis.

WEB SCRAPING AND SECONDARY DATA SOURCES

1. **Scrape Postal Code Data**
   Scrape postal code data of Toronto from Wikipedia.
   Clean and save the postal code data into a data frame.
   Verify if all postal codes are obtained.

2. **Add Geographical Data**
   Add latitude and longitude using pgeocode.
   Inspect, discuss, and clean the geo data for data sanity.

3. **Population Data**
   Download the population csv from Statistics Canada.
   Extract Toronto-specific data from the csv and save it.
   Read the new csv and save the population data frame.

4. **Income Data**
   Download the income csv from Statistics Canada.
   Extract Toronto-specific data from the csv and save it.
   Read the new csv and save the income data frame.

5. **Merge Data**
   Merge the postal-geo data, population data, and income data into a master data frame (**toronto_data**)

6. **Inspect and Clean Master Data**
   Step description

7. **Visual Representation**
   Plot the richest and most populous postal codes for visual representation.

8. **Measure central Tendencies**
   Box Plot Population, Income, Cafe counts
   Discuss to calculate either Mean or Median
   Calculate the Medians

API CALLS AND RE-USEABLE FUNCTIONS

9

API Selection

Analyze Google Places and Foursquare Places.
Set up an API call, read response, read the documentation.

10

Function to get Required Dictionary

Get all Place Type Cafe from the responses.
Use **get_places_cafe** function, append to **toronto_data.**

11

Data Storage

Store the responses as a csv.
Read the csv and create **toronto_data_df**
Store postal codes, geo information, income, population, counts of cafes, and names of cafes.

12

Visualization MAP

Plot toronto_data_df on a map to enable visualization of cafe competition.

13

Function Iteration

Iterate the function into **get_places_all** to get all place types for postal codes.

14

Data Filtering, Function Iteration 2

Iterate **get_places_all** to filter out places with 'point_of_interest' as the first in type list and get the second type if available (this place type is not of business use).

15

Final Data Storage

Store the response as a csv so that no more API calls are needed while working.
Read the csv and create a data frame **toronto_data_places**.

ONE-HOT ENCODING AND K-MEANS

16 **One-Hot Encoding**

Perform **one-hot encoding on the 'Place Type'** column to prepare categorical data for the machine learning model.

17 **Grouping and Analysis**

Group by 'Postal Code' and calculate the mean of each one-hot encoded column to analyze the distribution of different place types across various postal codes in Toronto.
Identify the most common place types for each postal code in Toronto as frequencies.

18 **Function Creation for Common Places**

Create a function to identify the most common places for rows in the coded and grouped Data Frame, make a **lists of top 10 most common places for each postal code.**

19 **Clustering**

Use **Silhouette Score** to establish the number of clusters.
Set up and run the **K-means clustering algorithm.**

20 **Final Data Frame**

Final Data Frames using, dropping columns not necessary for conclusion and discussion. **Display all 4 clusters.**
Plot **Stacked Bar Chart**

21 **Cluster Selection**

Select the most suitable cluster for the cafe business environment using the stacked bar charts.

22 **Prime Location In the cluster**

Score the postal codes for Population and Income and sort the result to get the Top 3.

THE REPORT

Here is the data analysis report resulting from the market segmentation study.

CENTRAL TENDENCIES

Population:

The median is around 25,000.

IQR spans from 15,000 to 40,000.

There is one outlier above 70,000.

Median Household Income:

The median is around 60,000.

IQR spans from 50,000 to 80,000.

There are several outliers above 100,000.

Cafe Counts:

The median is around 2.

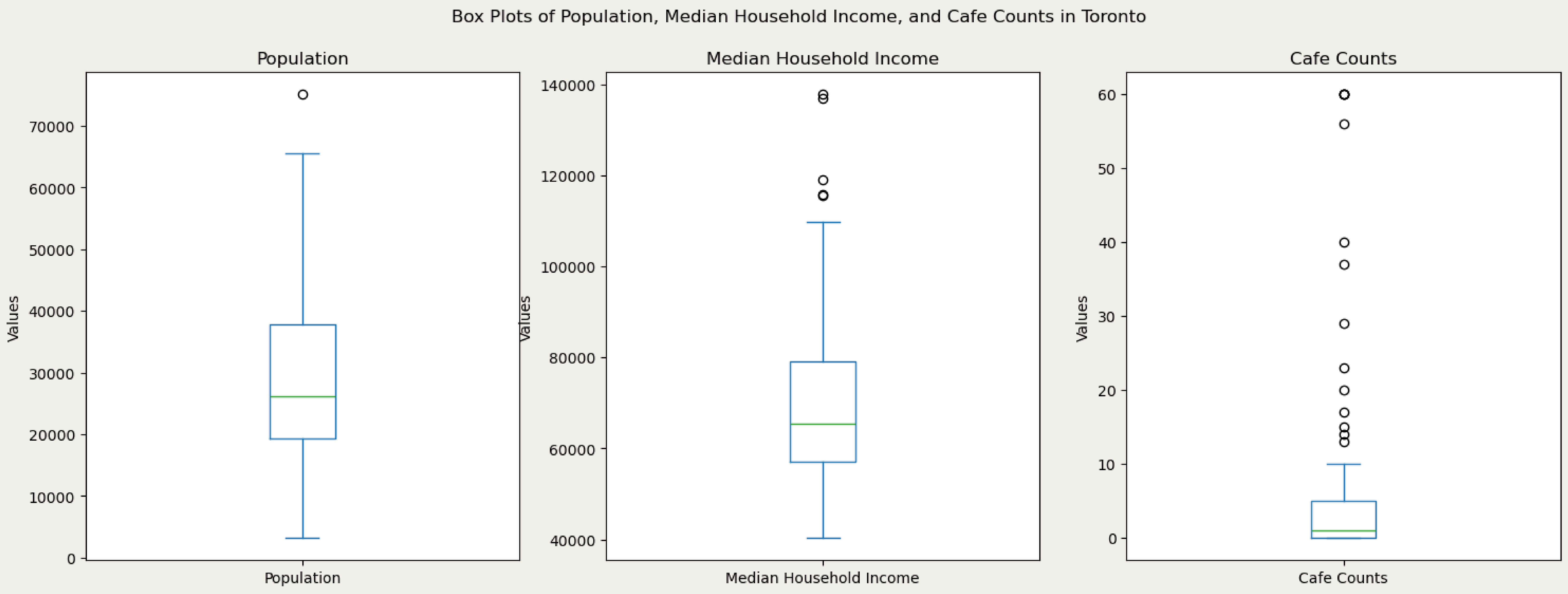
IQR spans from 1 to 6.

There are numerous outliers above 10, with the highest being around 60.

Median Per postal code Population - 26128 census year 2021

Median Per postal code Income - $65508 census year 2021

Median Per postal code Count of cafes - 1 Google Nearby Places



CLUSTERS

Cluster A:

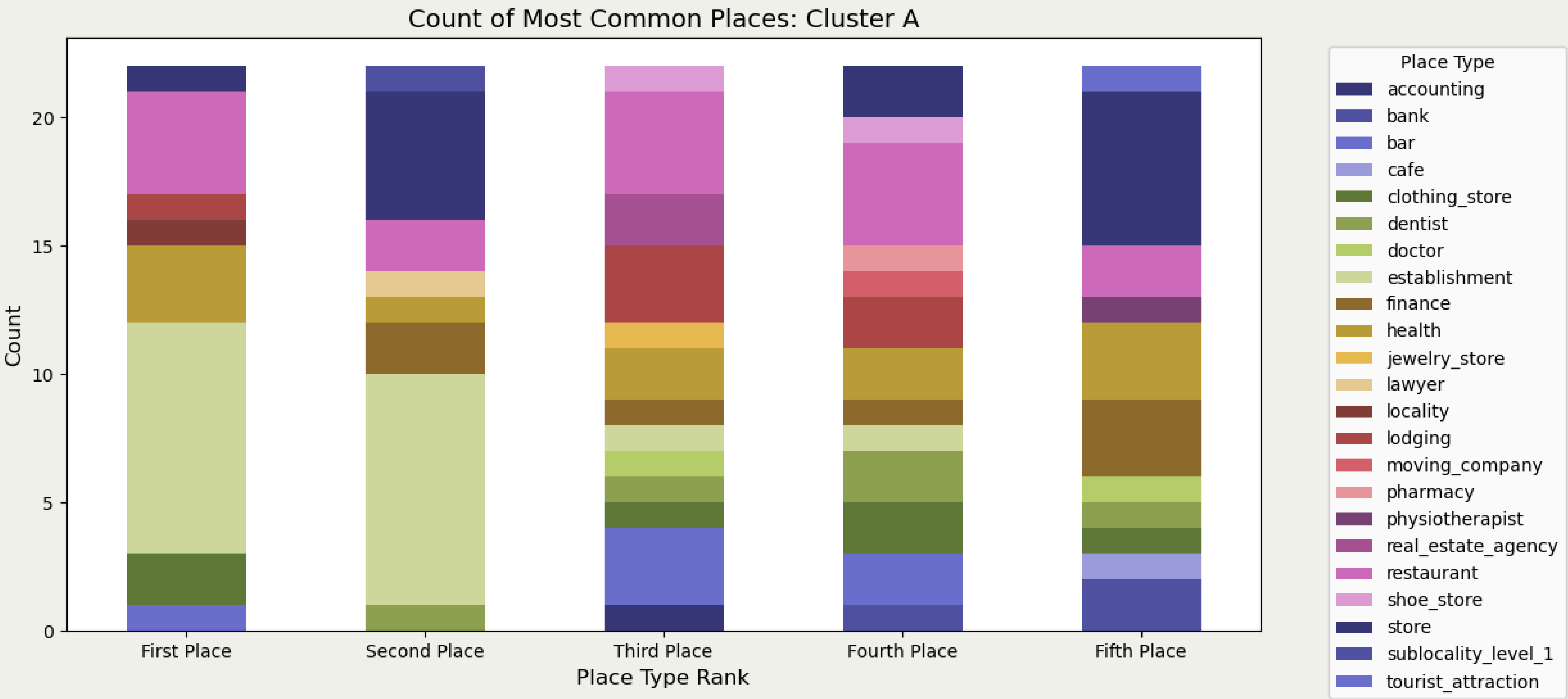
A Business heavy area with restaurants and motels.

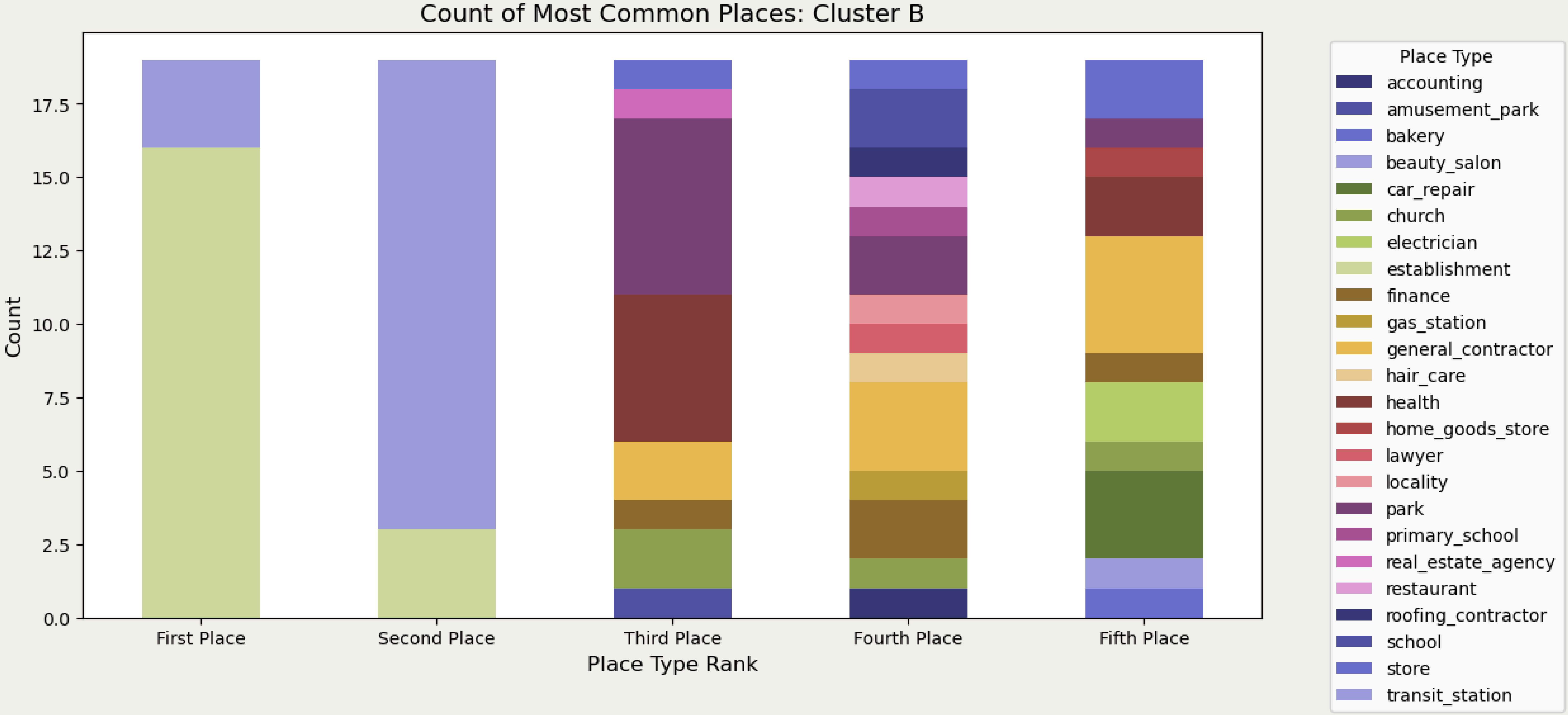
High concentration of cafes in postal codes M5B, M5C, M5E, M5G, and M5H.

Cluster B:

High footfall due to transit points, making it ideal for new cafes.

Few cafes, with small counts in postal codes M1H, M1V, and M9P.





CLUSTERS

Cluster C:

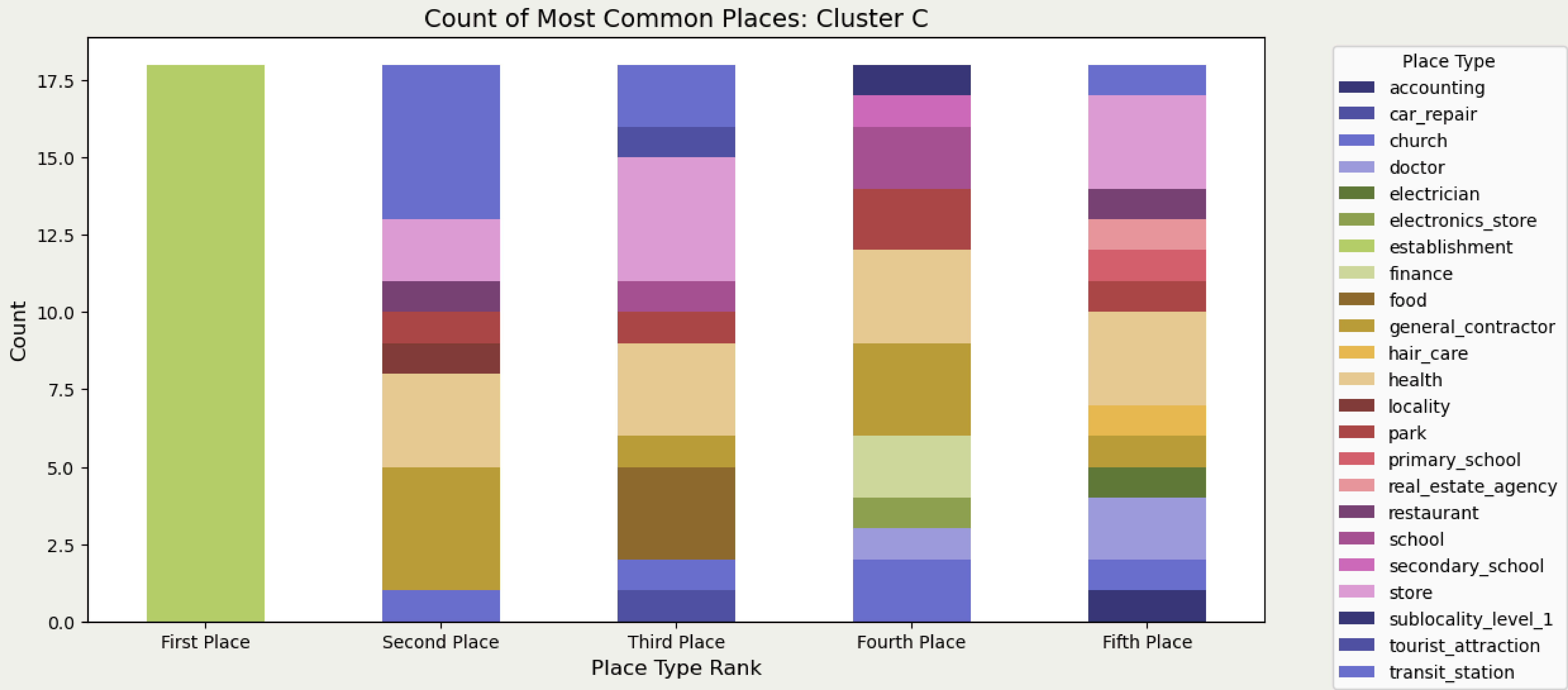
A Business heavy area with shops and doctor clinics.

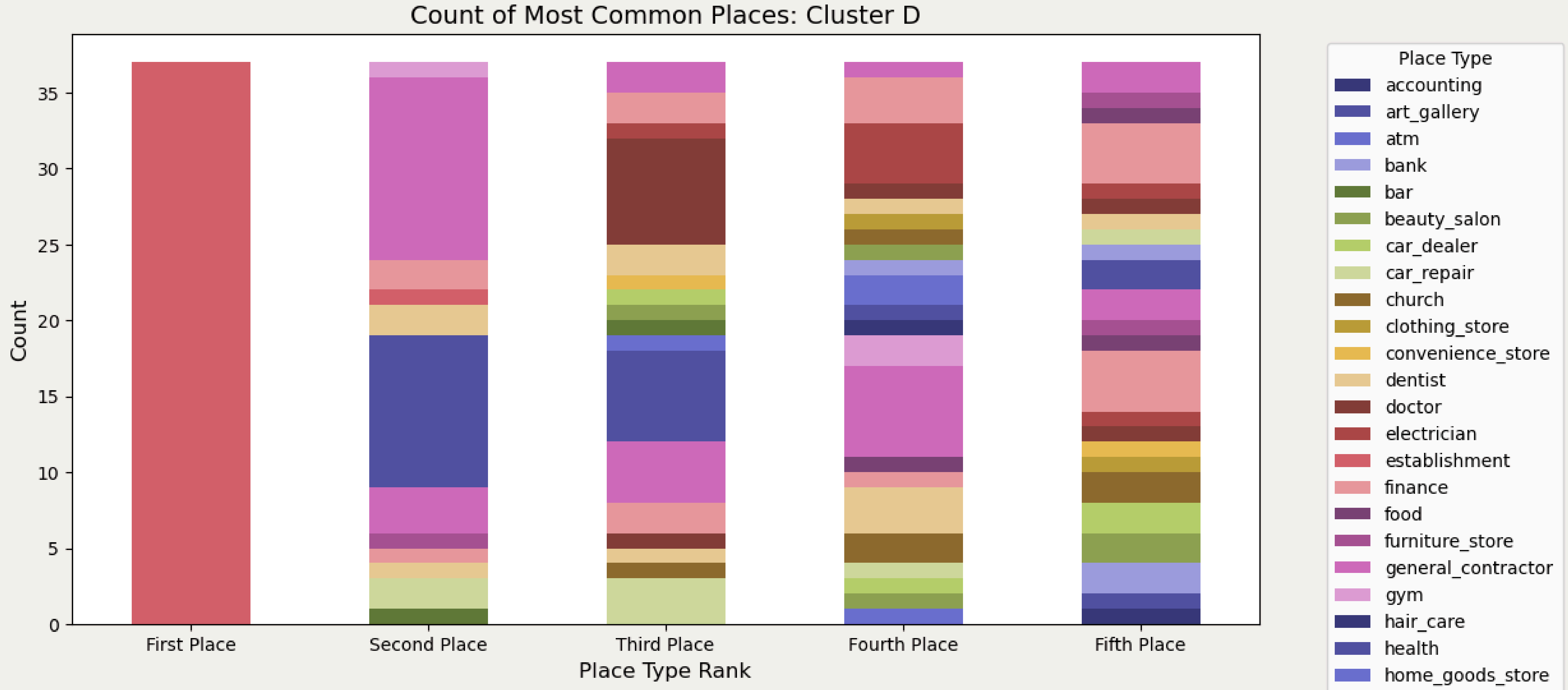
Moderate number of cafes in postal codes M3C, M4E, M5J, and M6K.

Cluster D:

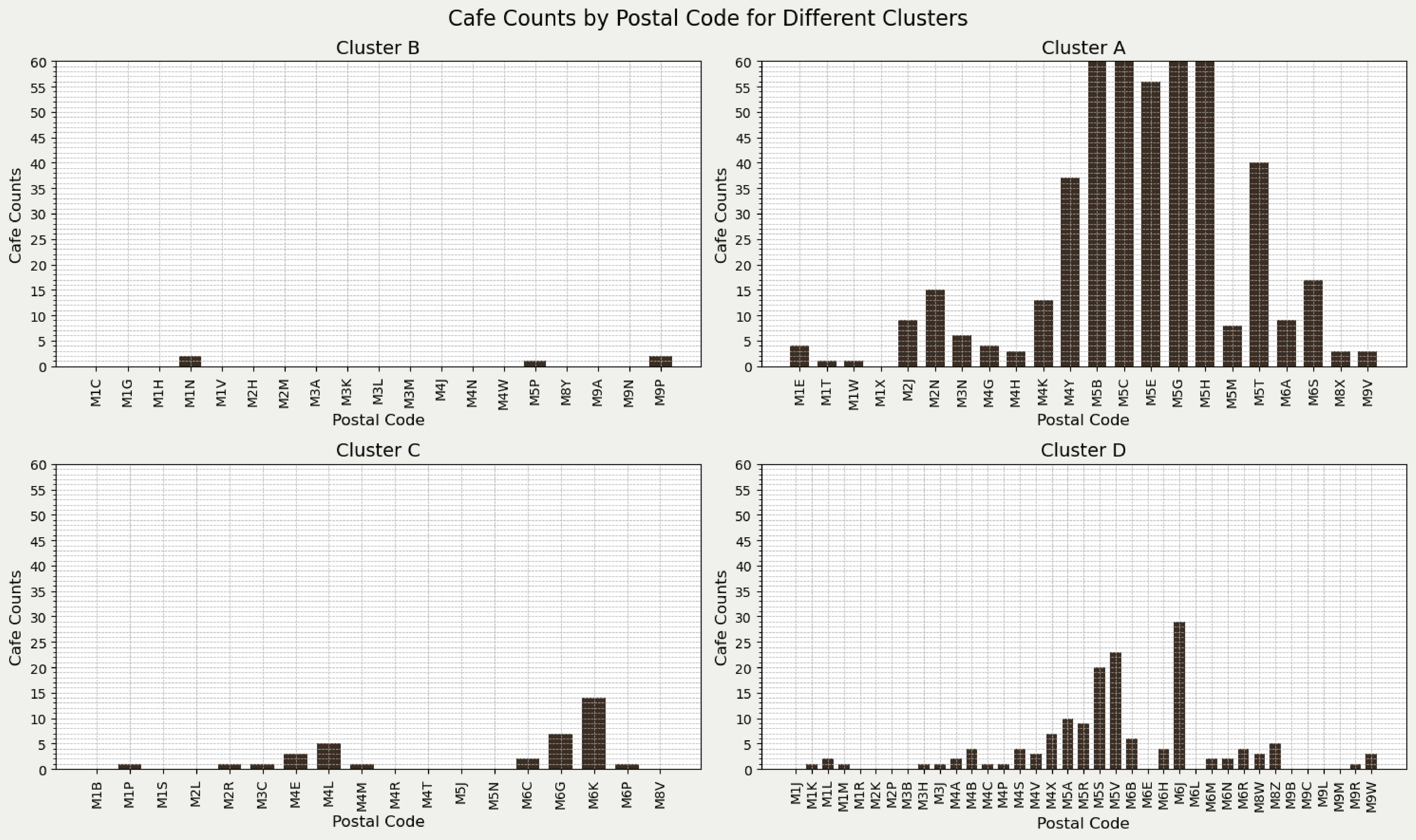
A Industrial heavy area with some shops and clinics.

Notable number of cafes in postal codes M5V, M6G, and M6H.





COMPETITION



Cluster Selection

Cluster B Is the recommended cluster as it has high volume of businesses and offices, with lots of transit points. Such high footfall will yield high cafe sale volume.

RECOMMENDATIONS

Rouge Hill / Port Union / Highland Creek

High household income: $109,785, High population: 35,642

Lawrence Park

Very high household income: $137,758, Sustainable population: 16,058

Milliken / Agincourt North / Steeles East

Comparable household income: $64,576, Very high population: 50,825

	Postal Code	Borough	Neighborhood	Population	Median Household Income	Cafe Counts	1st Most Common Place Type	2nd Most Common Place Type	3rd Most Common Place Type	4th Most Common Place Type	5th Most Common Place Type	6th Most Common Place Type	7th Most Common Place Type	8th Most Common Place Type	9th Most Common Place Type	10th Most Common Place Type	Score
Top	M1C	Scarborough	Rouge Hill / Port Union / Highland Creek	35642.0	109785.0	0	transit_station	establishment	finance	gas_station	car_repair	general_contractor	accounting	car_wash	hair_care	real_estate_agency	50758.1
Second	M4N	Central Toronto	Lawrence Park	16058.0	137758.0	0	establishment	transit_station	health	lawyer	general_contractor	accounting	spa	gym	park	food	49358.0
Third	M1V	Scarborough	Milliken / Agincourt North / Steeles East / L'...	50825.0	64576.0	0	transit_station	establishment	health	roofing_contractor	car_repair	church	finance	physiotherapist	pharmacy	primary_school	44786.9

THE CODE

Here is a showcase of the work that support the commentary of this data analysis.

Data Cleaning for Postal Code

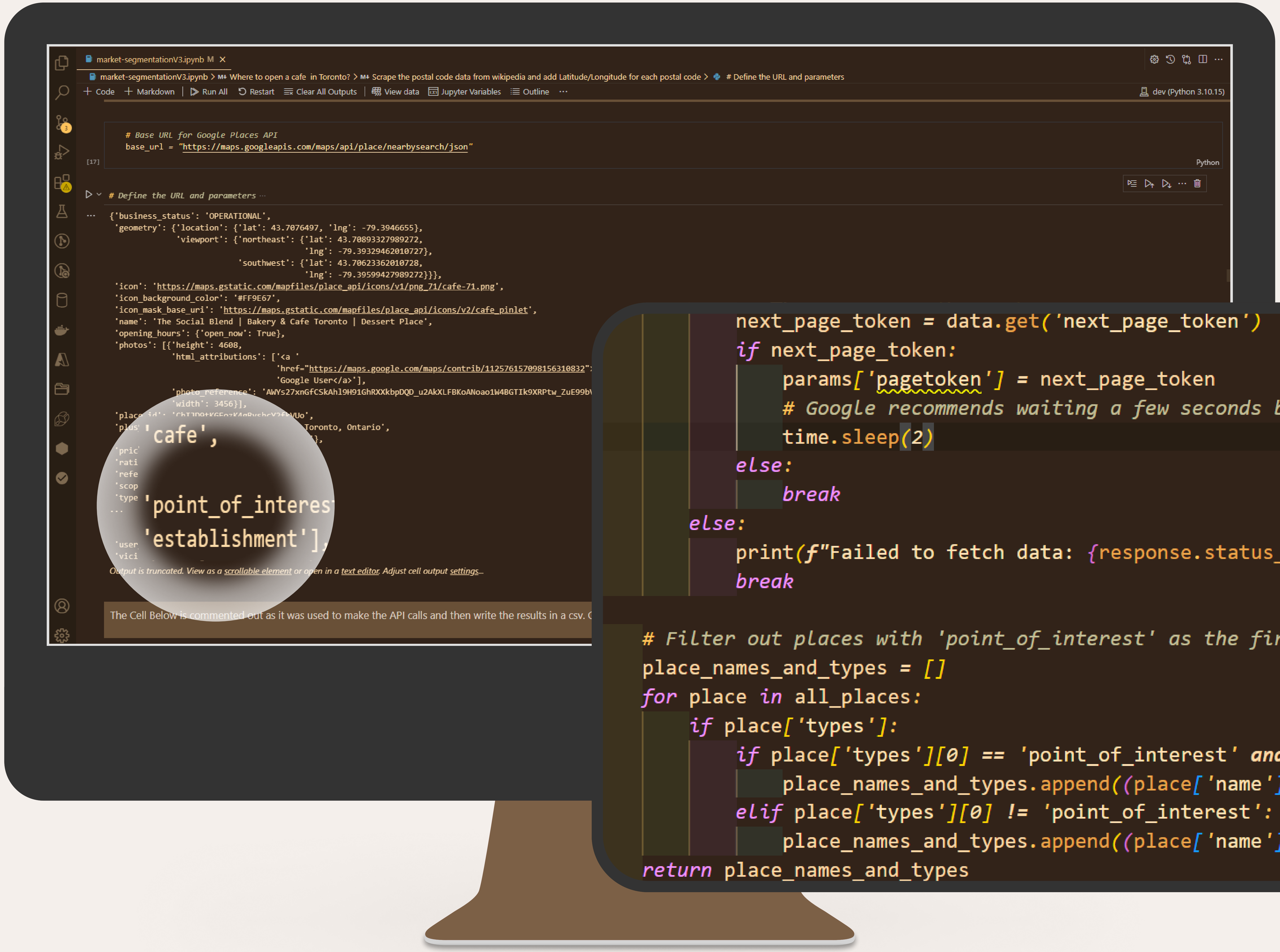
This code shows how the combined data from Wikipedia has been separated in three columns as postal codes, borough and neighborhood.

Use of lambda function to simplify the code and put that value for the designed list.


```python
# Iterate through each column in the DataFrame
for col in wiki:
    # Retrieve the first 3 characters of each string in the column and store in a list
    col_values = wiki[col].astype(str).apply(lambda x: x[:3]).tolist()
    postalcode_list.append(col_values)

    # Retrieve the substring from the 4th place value till the bracket open symbol "("
    substr_values = wiki[col].astype(str).apply(lambda x: x[4:x.find(' (')] if ' (' in x else x[4:]).tolist()
    borough_list.append(substr_values)

    # Retrieve the values within the brackets
    bracket_values = wiki[col].astype(str).apply(lambda x: x[x.find('(')+1:x.find(')')] if '(' in x and ')' in x else '').
    neighborhood_list.append(bracket_values)

return postalcode_list, borough_list, neighborhood_list
```

API call function

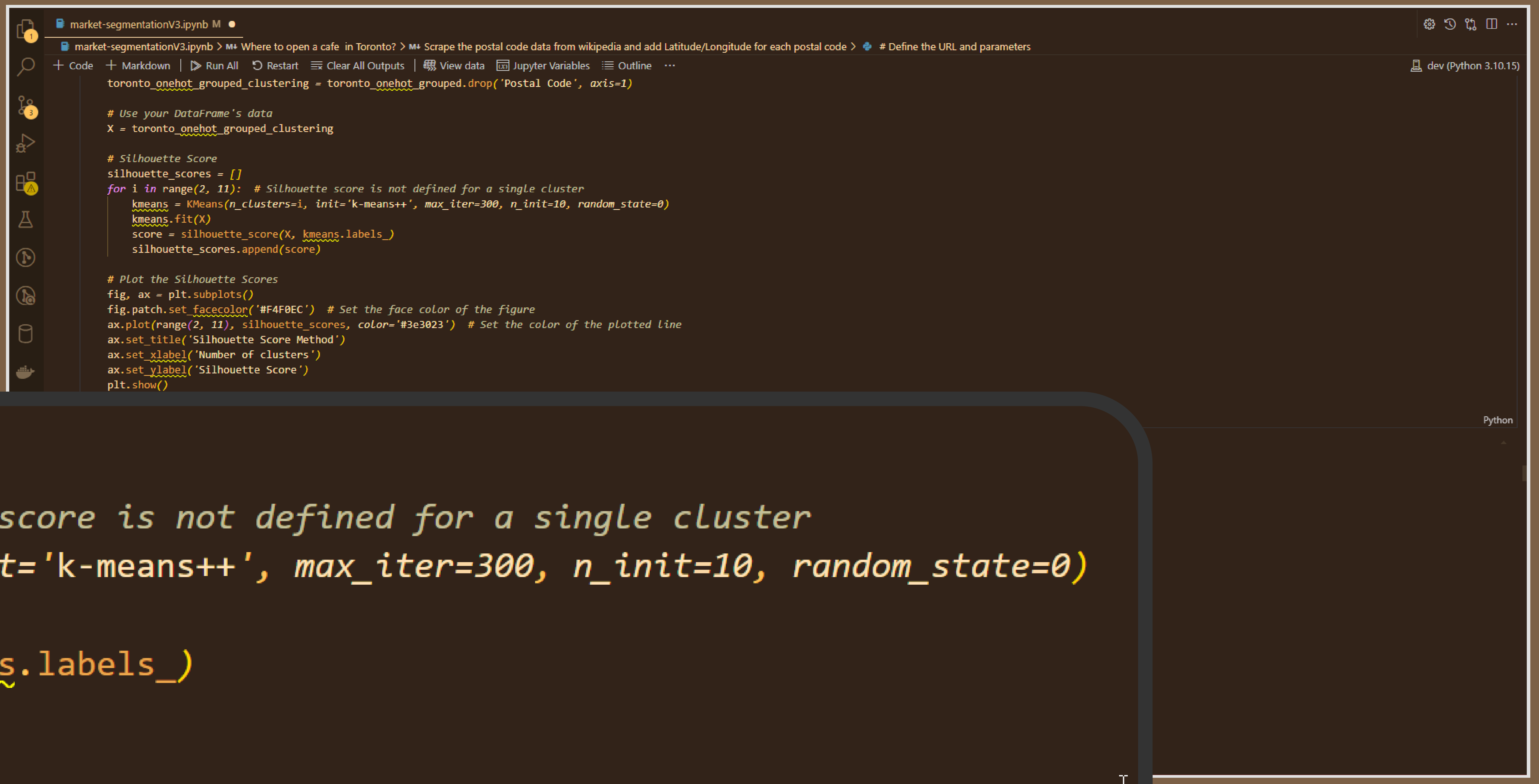
Iterated the API call function to get all "places for all of the postal codes. Filter out the 'point_of_interest' as it is not a place and is a google tag. This makes no sense for the study that we are doing.




```python
            next_page_token = data.get('next_page_token')
            if next_page_token:
                params['pagetoken'] = next_page_token
                # Google recommends waiting a few seconds before making the next request
                time.sleep(2)
            else:
                break
        else:
            print(f"Failed to fetch data: {response.status_code}, {response.json()}")
            break

    # Filter out places with 'point_of_interest' as the first type and get the second type if available
    place_names_and_types = []
    for place in all_places:
        if place['types']:
            if place['types'][0] == 'point_of_interest' and len(place['types']) > 1:
                place_names_and_types.append((place['name'], place['types'][1]))
            elif place['types'][0] != 'point_of_interest':
                place_names_and_types.append((place['name'], place['types'][0]))
    return place_names_and_types
```

Silhouette Score Method

The silhouette score is a measure of how similar an object is to its own cluster compared to other clusters. Higher silhouette scores indicate better-defined clusters.




```python
# Silhouette Score
silhouette_scores = []
for i in range(2, 11):  # Silhouette score is not defined for a single cluster
    kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=0)
    kmeans.fit(X)
    score = silhouette_score(X, kmeans.labels_)
    silhouette_scores.append(score)

# Plot the Silhouette Scores
fig, ax = plt.subplots()
fig.patch.set_facecolor('#F4F0EC')  # Set the face color of the figure
ax.plot(range(2, 11), silhouette_scores, color='#3e3023')  # Set the color of the plotted line
ax.set_title('Silhouette Score Method')
ax.set_xlabel('Number of clusters')
ax.set_ylabel('Silhouette Score')
plt.show()
```

K-mean Cluster

Using the k-mean clusters, divided the dataset into 4 clusters based on the most common places types that comes into the assigned postal codes.

NEXT STEPS

The next iteration of the code will focus on expanding this data model both horizontally and vertically!

Horizontally, the ethnicity data from Statistics Canada's 2021 census can be incorporated, adding another dimension to the market clustering model.

Additionally, more geospatial data can improve postal code zoning. The current radius-based method leads to overlaps, this will enhance accuracy.

Vertically, a decision tree to automatically identify the best market clusters for a cafe, taking into account various dimensions and suggesting optimal markets can be developed .