**Introduction**

This report aims to analyze the business landscape within various postal codes in Toronto, focusing on identifying the best locations for opening a new cafe. The analysis leverages data from the Google Nearby Places API, demographic information, and clustering techniques to provide actionable insights.

* The analysis is based on the following data sources.
  + Wikipedia: To obtain postal code data: [Wikipedia](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
  + Population data CSV: [Statistics Canada](https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=9810001901)
  + Income data CSV: [Statistics Canada](https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/download-telecharger.cfm?Lang=E) ["Forward sortation areas (FSAs)]
  + Google Places API to obtain nearby places data: [Google Nearby Places](https://developers.google.com/maps/documentation/places/web-service/supported_types?_gl=1*gb02v5*_up*MQ..*_ga*MTY4MjAzNTA0MS4xNzMzNjg2NTMz*_ga_NRWSTWS78N*MTczMzY4NjUzMi4xLjEuMTczMzY4Nzk5Ni4wLjAuMA..)
* Approach on a high level
  + Get the postal code data and add latitude and longitudes.
  + Get income and population data for each postal code/neighbourhood.
  + Use the API to get cafes for each neighbourhood area.
  + Count the number of cafes for each neighbourhood area.
  + Use the API to get the top 100 Place Types and their Place Type categories within a 500-meter radius of the postal code latitude and longitudes.
  + Check the frequency of each Place Type category, group them by postal codes, and look at the top 10 most frequent Place Type types to understand the most common business types in each neighbourhood.
  + Finally, come up with the most suitable neighbourhoods where cafe competition is low, population is adequate, income is on a higher spectrum, and the other common businesses are not adverse for cafe.

**Data Collection and Preparation**

1. Scaped the data from [Wikipedia](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) and then reconfirmed it.
   1. Read html!
   2. Create a data frame from the Wikipedia html table. A screen shot of a black and white screen

      Description automatically generated
   3. Read through the Table row by row and store the substrings in separate lists:
      1. The first 3 Letter are the Postal codes.
      2. The text after the postal code and before the bracket starts is the Borough Name.
      3. The text within the bracket and after the bracket is the Neighbourhood Name.
      4. Then use the 3 lists to make a data frame.
   4. Also verified the count of postal codes form [Geonames](https://download.geonames.org/export/zip/CA_full.csv.zip), presented as a separate notebook.
   5. Add Latitude and Longitude data using “pgeocode” library.
   6. It was found that for one postal code the Latitude and Longitude did not populate as it was an enclave for another postal code and reset the index.
   7. Drop the Nan and where “Neighbourhood” is an “Enclave.”
2. Download the Population data from [Statistics Canada](https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=9810001901) and save a population with postal code data frame.
   1. Read the csv obtained from stats Canada website.
   2. Collect the row where the "Geographic name" column starts with "M", as all postal codes of Toronto start with "M".
3. Download the Income data from [Statistics Canada](https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=9810001901) and save a Income with postal code data frame.
   1. Read the original large csv line by line and for line where GEO\_NAME and GEO\_CODE (POR) starts with "M" retain those and writing them into a new csv as output, as this file is very large.
   2. Read the output CSV row by row and collect the rows where column named "CHARACTERISTIC\_NAME" has the value "Median total income of household in 2020 ($)"
   3. For the collected rows get the value from column "GEO\_NAME', 'Dim: Sex (3): Member ID: [1]: Total - Sex."
   4. Create a data frame with GEO\_NAME as postal code and "C1\_COUNT\_TOTAL" as Median Household Income.
4. Merge the Postal codes, Lat, Lon, Population, Income in one clean data frame.
   1. Merge the 3 data frames and look for the rows that missing in among the intersection using "indicator.”
   2. It was found that postal codes - M5K, M5L, M5X are office blocks and M7A is Canadian Parliament building, these do not have income or population data. Drop these in our Postal, income and population data frames and merge them.
   3. Drop these postal codes.
   4. Look for duplicates and finally have a clean data frame with postal code Household Income and population.
5. Plotted a Pie Chart to explore the Top 10 Populous Neighbourhoods in Toronto. A diagram of neighborhoods in toronto

   Description automatically generated with medium confidence
6. Plotted a bar graph to explore the Top 10 Rich Neighbourhoods in Toronto. A graph of a number of neighborhood

   Description automatically generated
7. Google Nearby Places API
   1. Used the API to gather data on different types of businesses within 500 meters of each postal code's latitude and longitude.
   2. The data was processed to exclude irrelevant types like 'point\_of\_interest' and stored in a Data Frame.
   3. This was done using functions that was iterated making it workable for the final gets.
   4. First, used the function to get only the cafes, and counted the cafes can appended it to our master data frame. A screenshot of a black screen

      Description automatically generated
   5. Inspected the Population, Income, Café count data spread and decided to calcualte means so that it can be use this as a becnhmark for final selections of postal codes. A screenshot of a computer

      Description automatically generated
   6. Analysis of Central Tendency
      1. Population:
         1. The median is around 25,000.
         2. The interquartile range (IQR) spans from approximately 15,000 to 40,000.
         3. There is one outlier above 70,000.
         4. The median is a better measure of central tendency due to the presence of an outlier, which can skew the mean.
      2. Median Household Income:
         1. The median is around 60,000.
         2. The IQR spans from approximately 50,000 to 80,000.
         3. There are several outliers above 100,000.
         4. The median is a better measure of central tendency due to the presence of multiple outliers, which can skew the mean.
      3. Cafe Counts:
         1. The median is around 2.
         2. The IQR spans from approximately 1 to 6.
         3. There are numerous outliers above 10, with the highest being around 60.
         4. The median is a better measure of central tendency due to the presence of numerous outliers, which can skew the mean.
      4. Median of Population per postal code – 26128
      5. Median of Income per postal code – 65508
      6. Median of café per postal code – 1
8. Plotted the Café counts on a map that can be zoomed in using Nomiatim, (Green shows Below median)

A map of a city with many dots

Description automatically generated

1. Google Nearby Places API - getting all business for all postal codes.
   1. Using the Google Nearby Places API to look for other types of businesses that dominate the market with in 500 meters of each post code Latitude/longitude and count them.
   2. Also compared the result with foursquare API places, It was found that google gets us better count and quality of data when randomly selected results were verified. The comparison of the two is presented in a separate notebook.
   3. Alter the get cafe function to get all places within the set parameter, while testing it was also found that Types in Google also has a dimension that they call "point\_of\_interest", which is irrelevant to our business problem, hence it is being filtered out.
   4. Use the get all places function to get from one example postal code.
   5. Further build the code to get all places for all postal codesA computer screen shot of a program

      Description automatically generated
   6. Create a data frame to store the Postal codes, Place type, Place name in a new data frame totonto\_data\_places
   7. Write toronto\_data\_places as a csv and then read it back again as otonto\_dataplaces\_df so that there is no need to make the API calls when the notebook run again. A computer screen shot of a black screen

      Description automatically generated
2. One-Hot Encoding
   1. Categorical variables in the 'Place Type' column were converted into binary columns using one-hot encoding, resulting in a DataFrame, `toronto\_onehot`.
      1. One-hot encoding will convert categorical variables into a form that can be provided to ML algorithms to do a better job in prediction. Each unique value in the 'Place Type' column becomes a separate column with binary/bool values (0 or 1).
      2. This code performs one-hot encoding on the 'Place Type' column, reintroduces the 'Postal Code' column, rearranges the columns to place 'Postal Code' first, and then prints and displays the resulting DataFrame. This process is useful for preparing categorical data for machine learning models while maintaining important contextual information
   2. one-hot encoded data was grouped by 'Postal Code' and calculated the mean of each one-hot encoded column for each postal code, to enable Clustering Analysis.
      1. In the resulting DataFrame each row represents a postal code, and the columns contain the average values of the one-hot encoded place types.
      2. This aggregated data becomes useful for analyzing the distribution of different place types across various postal codes in Toronto.
      3. Create a new DataFrame that lists the top 10 most common Place Types for each postal code in Toronto.
3. K-Means Clustering
   1. Calculate the silhouette score to decide how many clusters to form. A graph with a line going up

      Description automatically generated
   2. Performed K-means clustering on 4 clusters for the one-hot encoded data to group postal codes into clusters based on the types of businesses present.
      1. Retrieve the cluster labels for the first 10 rows in the DataFrame.
      2. The cluster labels indicate which cluster each row (postal code) belongs to, based on the similarity of their Place Type frequencies.
      3. Created the final Data frame using the toronto\_onehot\_merged, drop the columns not necessary for conclusion and discussion.

**Data Analysis**

1. Cluster Characteristics - Each cluster was analyzed to identify its characteristics:
   1. Cluster A: Business-heavy area with restaurants and motels.
   2. Cluster B: High-traffic business and office area with numerous transit points.
   3. Cluster C: Business-heavy area with shops and doctor clinics.
   4. Cluster D: Industrial-heavy area with some shops and clinics.
2. Detailed Cluster Analysis
   1. Cluster A:
      1. High concentration of cafes in postal codes M5B, M5C, M5E, M5G, and M5H.
      2. Suitable for businesses targeting office workers and tourists.

A screenshot of a graph

Description automatically generated

* 1. Cluster B:
     1. Few cafes, with small counts in postal codes M1H, M1V, and M9P.
     2. High footfall due to transit points, making it ideal for new cafes.

A screenshot of a graph

Description automatically generated

* 1. Cluster C:
     1. Moderate number of cafes in postal codes M3C, M4E, M5J, and M6K.
     2. Suitable for businesses targeting local residents and shoppers.

A screenshot of a graph

Description automatically generated

* 1. Cluster D:
     1. Notable number of cafes in postal codes M5V, M6G, and M6H.
     2. Suitable for businesses targeting industrial workers and nearby residents.

A screenshot of a computer

Description automatically generated

1. The bar chart below shows the cafe counts for these Clusters.

A graph of different numbers

Description automatically generated with medium confidence

**Recommendations**

1. 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.
2. Prioritized postal codes with higher populations while still considering income and place types.
   1. Adjust Weights as coffee is not that costly and wider customer base matters over income: Increased the weight for population to 0.5 and decreased the weight for income to 0.3, while keeping the weight for place types at 0.2.
   2. Calculate Score: Updated the score calculation to reflect the new weights.
   3. Sort Data: Sorted the DataFrame by the updated score in descending order.
   4. Select Top 3: Selected the top 3 postal codes based on the updated score.
3. Based on the analysis, the following postal codes are recommended for opening a new cafe:
   1. Top Recommendation: Rouge Hill / Port Union / Highland Creek (Postal Code: M1C)
      1. High household income: $109,785
      2. High population: 35,642
      3. No competition
   2. Second Best Recommendation: Lawrence Park (Postal Code: M4N)
      1. Very high household income: $137,758
      2. Sustainable population: 16,058
      3. No competition
   3. Third Best Recommendation\*\*: Milliken / Agincourt North / Steeles East (Postal Code: M1V)
      1. Comparable household income: $64,576
      2. Very high population: 50,825
      3. No competition