Segmentation of Restaurants by Median Household Income in Metro Vancouver

Coursera Capstone Final Report

Zhongjie Shen
November 30, 2020

Preface

This report is the final milestone of Coursera's IBM Data Science Certificate and will be a great reflection of my study in the past few months.

First of everything, I give my special thanks to Krystal who encourages and supports me unconditionally along the entire path.

I also want to appreciate all the peers and group mates who reviewed my previous work as well as this report. It was an amazing experience in studying with people all around the globe and see all the fantastic work been done in this field.

Table of Contents

Introduction

Background

Metro Vancouver is a concept of a specific region near and include Vancouver, British Columbia, Canada. Metro Vancouver has a wide range of restaurants selections as well as a very diverse culture and races compositions. Like similar metropolis across the world, Metro Vancouver's residents work in a large number of different sectors and industries and have a wide band of income levels.

Problem

There is always interest in finding out the impact of income levels on different consumer sectors especially like restaurants.
Finding out the distribution and relationship between their income levels and restaurants categories can be greatly helpful in determining the profile of restaurants industries and providing an overview of the income level's impact on this industry as well.

Interest

Since the author has been lived in Metro Vancouver area for more than 7 years, it is in great interest of the author to be able to find and visualize the pattern of restaurants in different sub-areas in Metro Vancouver. Also, Coursera provided author essential knowledge and tools to make this happen.

Data

Data Source

This report will discuss the results of restaurants segmentation work produced by using a wide range of data sources including 2016 Canada Census Profile, Canada Post, Foursquare Places API etc.

To be specific, the data is retrieved from the following places in Public Domain:

1. 2016 Canada Census Profile: Statistics Canada Website
2. Canada Post: There is no direct data interaction. However, its Forward Sorting Area code is used to separate regions in the 2016 Canada Census Profile. Forward Sorting Area (FSA) is a trademark controlled by Canada Post
3. Foursquare Places API: The restaurants categories and relevant profiles will be retrieved from Foursquare Places API.
4. Geolocation Services Canada API: The API will be used to find the centroid coordinates of each FSA

Data Cleaning

Most data we use in this project is in decent shapes and formats. As a result, there is no extensive data cleaning necessary in this project.
Some minor filtering and joining of data will be handled by the Jupyter Notebook discussed later in this report.

Feature Selection

The report will use the following fields in respective data sources detailed in the following table:

Data Source	Fields	Notes
2016 Canada Census Profile	Median Household Income	Income information
	Forward Sorting Area	FSA used as a way to determine elements in segmentation process
Foursquare Places API	Category Name	Name of Category
Geolocation Services Canada API	Coordinates	Coordinates of each FSA's centroid

Methodology

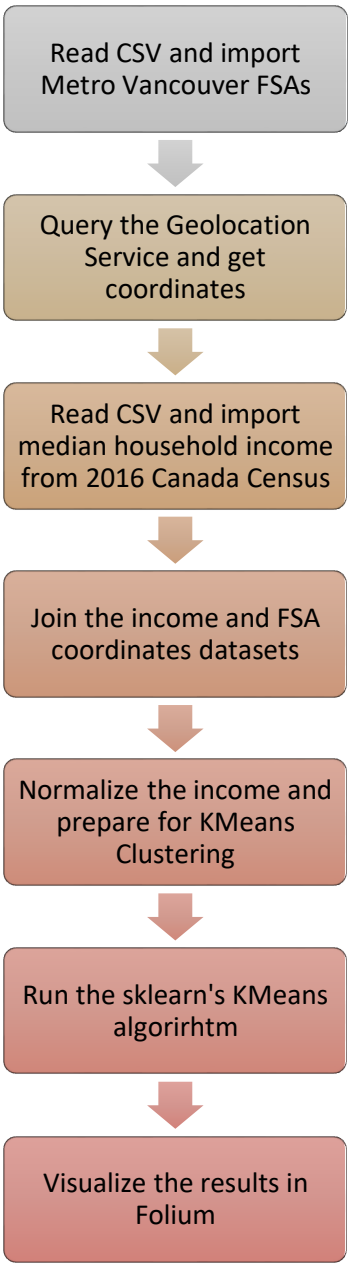
The report will use the output from a Python Jupyter Notebook that was developed during this process. Unless stated otherwise, all the data using in the Notebook is from Public Domain and are restricted by their copyright agreement.

Several Python packages will used:

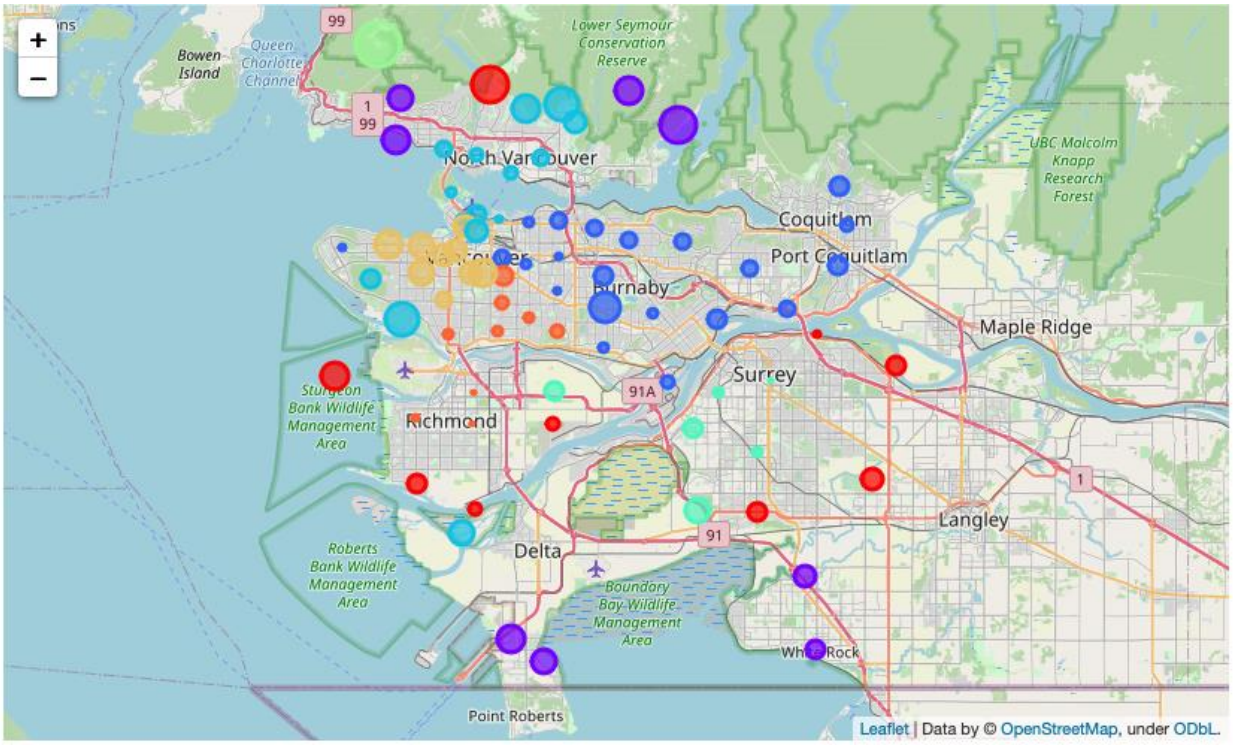
- Pandas
- Requests
- Sklearn
- Folium

Basically, the Notebook is a data processing pipeline that involves multiple steps and stages. The Notebook calls a number of API services followed by the ETL process, and then use the sklearn's KMeans clustering method in grouping the neighbourhoods (or FSA, more specifically).

For a high-level illustration, the flowchart on next page can be a great reference:


Read CSV and import Metro Vancouver FSAs
Query the Geolocation Service and get coordinates
Read CSV and import median household income from 2016 Canada Census
Join the income and FSA coordinates datasets
Normalize the income and prepare for KMeans Clustering
Run the sklearn's KMeans algorirhtm
Visualize the results in Folium

Results



As you could see from the result above, the popular restaurants categories show an obvious pattern in Metro Vancouver.

The colour of the bubble is the Cluster of neighbourhoods that have similar categories of restaurants in popularity. The size of the bubble represents the normalized scale of income, the bigger the bubble, the higher median household income present.

Discussion

When we examine each cluster with different label (colour), there are quite different in distribution of restaurants categories.

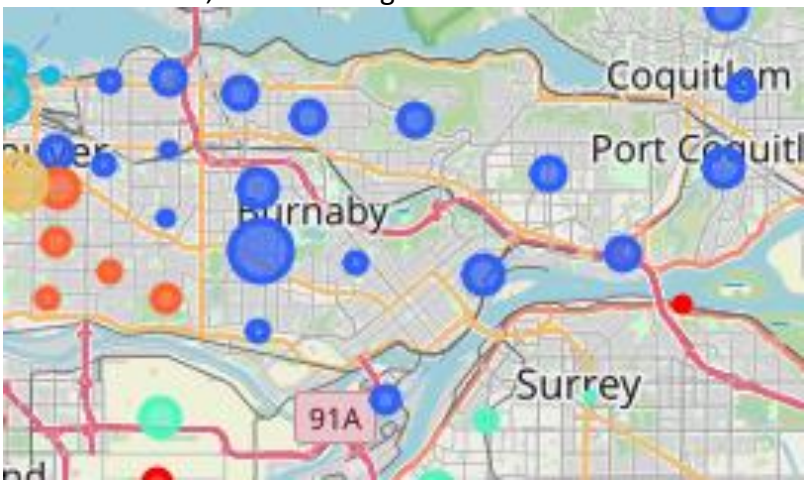
For example, the most common restaurants in the yellow cluster which is at the centre-right of the map above are:

- Bakery Restaurant
- Seafood
- Japanese
- French

And the average median household income in that cluster is approximately $107562 CAD which is much higher than the national average ($88306).

As you can see, the overall average of yellow bubble size is significantly larger than the other parts which means the households in that area have higher income. This shows a clear impact on their restaurant choices. The Bakery, Seafood, French restaurants usually set a relatively high price points that are more approachable by higher income households.

On the contrast, when looking into the Blue Cluster on the east part of Metro Vancouver:



The most common restaurants in this region are quite different from the Yellow Cluster we discussed above, the common restaurants are now:

- Sushi
- Japanese
- Burgers
- Vietnamese
- Pizza

These are relatively more affordable places compare with the restaurants in Yellow Cluster. This is also supported by the average median household income of $86775 which is about 20% lower than the Yellow Cluster and also slightly below Canadian average.

Conclusion

To sum up, this project shows an interesting outcome that the popular restaurants in an area is very closely connected with its income level. The higher average household income area generally contains a more expensive or luxury set of restaurants like Seafood, French etc.

One important thing needs to be mentioned is that the household size can also be a noise factor in the income data. For instance, the Blue Cluster above in east part of Metro Vancouver may have more large families with 5-6 people while the Yellow Cluster may have 3-4 instead. This can pose a huge effect on the available funds to dine out and eventually result in different food choices.

However, the effect on popular restaurants is not just from the household income. It may consist multiple factors like culture, race etc. Due to the time and computation restraints, it is not a highly comprehensive analysis and there is a long way to go for a more solid conclusion.

As a future direction, segmentation of restaurants like this needs analysis on more independent variables and should be based on more data. This obviously requires deeper understanding of data analytics field and expertise in relevant fields.