1. Introduction

1.1. Background

Suppose a client Ava got two job offers from two different companies. Company A's office is located in Toronto and is approximately 20 min drive away from Ava's apartment. Ava lives in the central bay street neighborhood in Toronto. Company B's office is in Vancouver, which means Ava has to move to Vancouver in order to work in Company B. However, Company B is offering a higher salary to Ava compared to Company A. One thing that makes Ava hesitate to move to Vancouver is the convenience of living in her current neighborhood, which has various types of venues around and makes her life much comfortable and enjoyable.

1.2. Problem

If Ava were to move over to Vancouver, she would like to find a neighborhood around Company B (maybe 20 min drive away) that's similar to her current one, so as to achieve a good work-life balance as she does now. Thus, this project aims to look for a neighborhood around Company B in Vancouver that's similar to the central bay street neighborhood in Toronto in terms of the venue categories.

2. Data

2.1. Data Sources

Dataset of information about neighborhoods in Vancouver such as their names and geocodes came from the dataset called "Local area boundary" on the website: https://opendata.vancouver.ca/explore/dataset/local-area-boundary/table/?location=12,49.2474,-123.12402.

Geocode of the central bay street neighborhood could be found in this file https://cocl.us/Geospatial_data.

Data of information about all neighborhoods in Toronto are scraped from the Wikipedia page: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Information about venue categories of neighborhoods can be obtained through API calls made to Foursquara.com.

2.2. Data Description

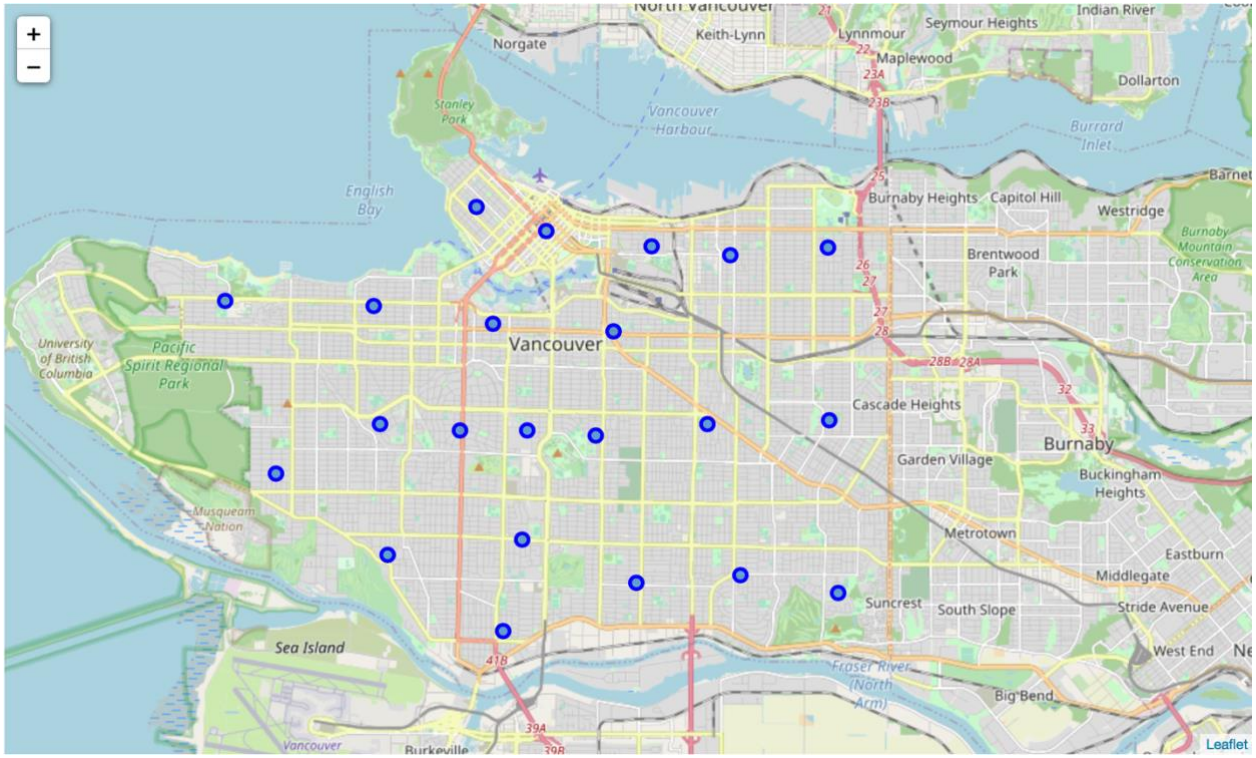
Data scraped from the Wikipedia page contains missing values, for example, some postal codes were not assigned with any borough's or neighborhood's name.

	PostalCode	Borough	2	3	4	5	6	7	8	9	...	21	22	23	24	25
0		None	None	None	None	None	None	None	None	None	None	...	None	None	None	None	None
1	M1A	Not assigned	Not assigned	None	None	None	None	None	None	None	None	...	None	None	None	None	None
2	M2A	Not assigned	Not assigned	None	None	None	None	None	None	None	None	...	None	None	None	None	None
3	M3A	North York	Parkwoods	None	None	None	None	None	None	None	None	...	None	None	None	None	None
4	M4A	North York	Victoria Village	None	None	None	None	None	None	None	None	...	None	None	None	None	None

Those rows with data under 'Borough' not assigned were removed. In order to get the geocodes for the neighborhoods, the csv file https://cocl.us/Geospatial_data was imported and got combined with our first data frame.

	PostalCode	Borough	Neighbourhood	Postal Code	Latitude	Longitude
0	M3A	North York	Parkwoods	M3A	43.753259	-79.329656
1	M4A	North York	Victoria Village	M4A	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	M5A	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	M6A	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	M7A	43.662301	-79.389494

Dataset of neighborhoods in Vancouver was downloaded directly from opendata.vancouver.ca. as a csv file. Only columns of the neighborhoods' names and geocodes (latitude and longitude) are retained. There were 22 neighborhoods of Vancouver in the dataset. Their geocodes would be put into API calls to the Foursquare website to get nearby venues.

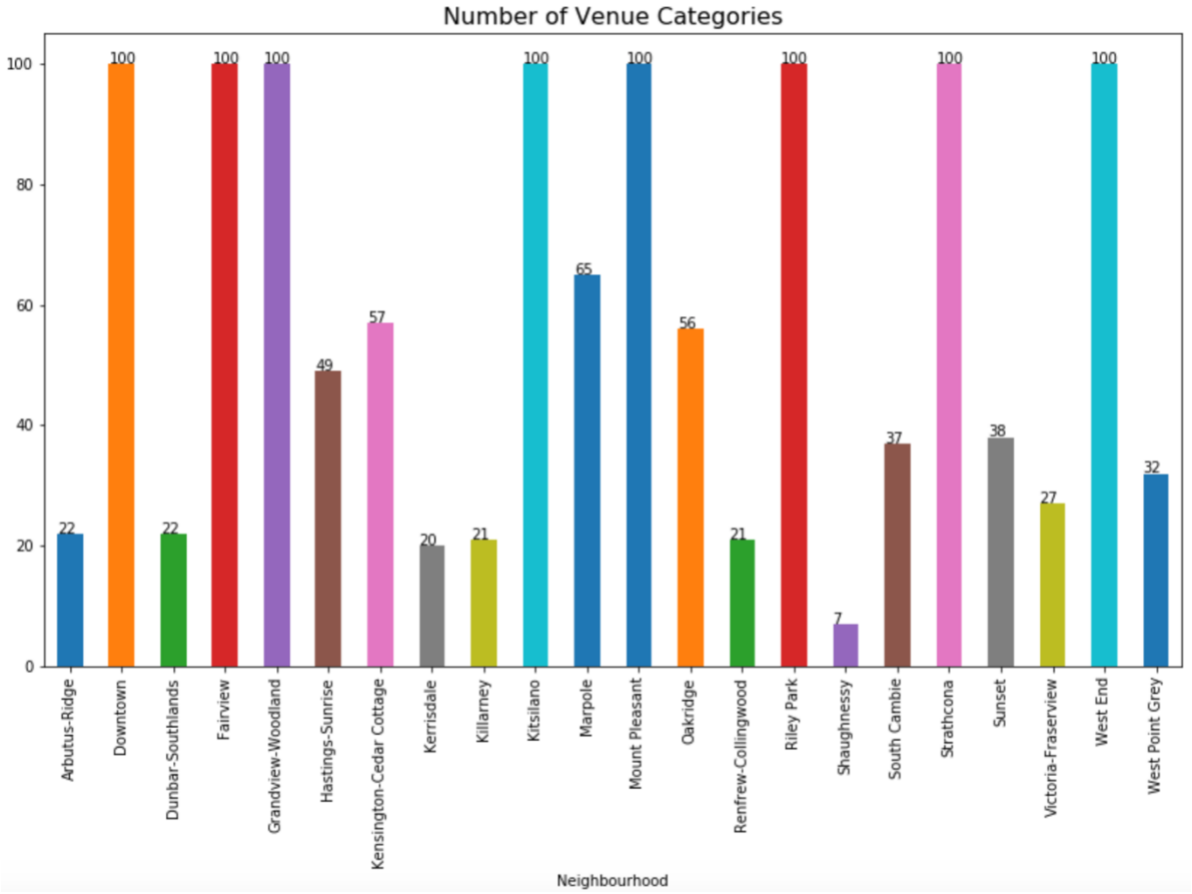


3. Methodology

3.1. Exploratory Data Analysis

In order to get the neighborhoods in Vancouver that are similar to the central bay street neighborhood, I used the pandas get_dummies method first. Venue categories of up to 200 venues (with radius of 1km) within each neighborhood were labeled with 0 or 1 to indicate

whether they belong to a certain category. I noticed that there were 218 unique venue categories among all those neighborhoods. The following graph shows the number of venue categories of each neighborhood.



Then, I used the groupby method in pandas to get a new data frame showing the average existence of each venue category in each neighborhood.

	Neighbourhood	Accessories Store	African Restaurant	American Restaurant	Amphitheater	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Dealership	...	Trade School	Trail	Train Station	Vegetarian / Vegan Restaurant
0	Arbutus-Ridge	0.00	0.0	0.000000	0.000000	0.00	0.00	0.00	0.00	0.0	...	0.00	0.00	0.0	0.000000
1	Downtown	0.01	0.0	0.000000	0.000000	0.01	0.00	0.00	0.00	0.0	...	0.01	0.00	0.0	0.010000
2	Dunbar-Southlands	0.00	0.0	0.000000	0.000000	0.00	0.00	0.00	0.00	0.0	...	0.00	0.00	0.0	0.000000
3	Fairview	0.01	0.0	0.010000	0.000000	0.00	0.02	0.02	0.00	0.0	...	0.00	0.02	0.0	0.010000
4	Grandview-Woodland	0.00	0.0	0.000000	0.000000	0.00	0.00	0.02	0.01	0.0	...	0.00	0.00	0.0	0.020000
5	Hastings-Sunrise	0.00	0.0	0.020408	0.020408	0.00	0.00	0.00	0.00	0.0	...	0.00	0.00	0.0	0.000000
6	Kensington-Cedar Cottage	0.00	0.0	0.035088	0.000000	0.00	0.00	0.00	0.00	0.0	...	0.00	0.00	0.0	0.052632
7	Kerrisdale	0.00	0.0	0.000000	0.000000	0.00	0.00	0.00	0.00	0.0	...	0.00	0.00	0.0	0.000000
8	Killarney	0.00	0.0	0.000000	0.000000	0.00	0.00	0.00	0.00	0.0	...	0.00	0.00	0.0	0.000000
9	Kitsilano	0.00	0.0	0.010000	0.000000	0.00	0.01	0.01	0.00	0.0	...	0.00	0.00	0.0	0.030000

10 rows × 219 columns

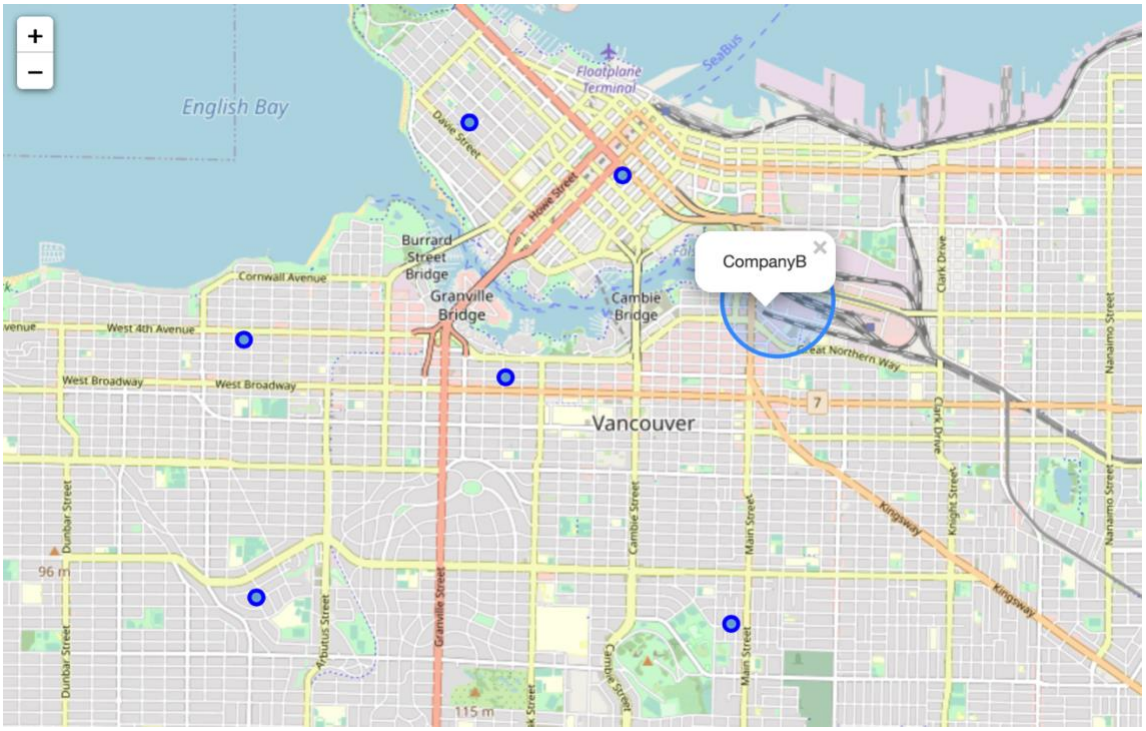
However, after the same method was done to the central bay street neighborhood, I found that there were 64 unique venue categories in there. After I concatenate all unique venue categories of central bay street to the data frame of those in Vancouver, I got 231 unique categories in total.

3.2. Modeling

Now, each row of the combined data frame represents one neighborhood and its venue categories. I then applied the K-Means clustering algorithm with num of clusters set to 7. This would assign each neighborhood a label based on the similarity of their existing venue categories.

4. Results

There were 6 neighborhoods in Vancouver sharing the same label with central bay street, which are: Arbutus-Ridge, Riley Park, Kitsilano, Fairview, West End, Downtown.



5. Discussion

I found that if I set the number of clusters differently, the resulting clusters varied. For, example, I first set the number to 5, there were 11 neighborhoods in the same cluster with the central bay street. Then I decided to increase the number of clusters to 7, and 6 neighborhoods came out this time. I guess reasonably higher number of clusters could do a better job in clustering those neighborhoods (maybe more precise). However, I couldn't test the performance of my model as there were only 22 neighborhoods in the original dataset, and those neighborhoods were not initially classified either.

6. Conclusion

There are 6 neighborhoods Ava can choose to live in Vancouver with similar surroundings as her place in central bay street. However, based on the geocode of CompanyB ([49.27071645, -123.09744954633206]) and as I plotted on the map in my notebook, the Downtown neighborhood and the Fairview neighborhood are closer to CompanyB. Therefore, moving to one of these two neighborhoods gives Ava a shorter commute to work.