Applied Data Science Capstone

The Battle of Neighborhoods – Assignmment

Segmenting and Clustering Neighborhoods in Vancouver

DIEGO NUNES BOTELHO

27 May 2020

Part I – Introduction and Problem Statement

In this project, we will study, analyze, cluster, and compare the neighborhoods of Vancouver which is in British Columbia (Canada). We will investigate on what kinds of businesses are most common in the city, which outdoors and recreation activities and what kinds of restaurants are most common between the neighborhoods.

Doing this project will enable us to get a better understanding of similarities and differences between neighborhoods of Vancouver, which will make it known to business people what types of businesses are more likely to thrive or understand why you should visit this place or, who knows, move there.

Vancouver, city, southwestern British Columbia, Canada. It is the major urban centre of western Canada and the focus of one of the country's most populous metropolitan regions. Vancouver lies between Burrard Inlet (an arm of the Strait of Georgia) to the north and the Fraser River delta to the south, opposite Vancouver Island. The city is just north of the U.S. state of Washington. It has a fine natural harbour on a superb site facing the sea and mountains.

Figure 1 – Vancouver

With its scenic views, mild climate, and friendly people, Vancouver is known around the world as both a popular tourist attraction and one of the best places to live. Vancouver is also one of the most ethnically and linguistically diverse cities in Canada with 52 percent of the population speaking a first language other than English.

Vancouver is made up of a few smaller neighbourhoods and communities. Neighbourhood boundaries provide a way to break up the city's large geographical area for delivering services and resources and identify the distinct culture and character of different areas of our diverse population. However, there is some disagreement on all of the names and boundaries of these areas.

Part II – Data Acquisition and Preparation

In this section, the process of acquiring, cleaning, and preparing the dataset used in this project for the next stages will be specified. To be able to do this project, two types of data are needed:

- **Neighborhood Data:** datasets with the list names of the neighborhoods of Vancouver and their latitude and longitude coordinates. The neighborhoods names were obtained in the website of Vancouver (https://vancouver.ca/news-calendar/areas-of-the-city.aspx) and the latitude and longitude data were obtained using a recursive function that would return the geocode of the address passed into it.
- **Venues Data:** data that describes the top 100 venues (restaurants, cafes, parks, museums, etc) in each neighborhood of Vancouver. The data should list the venues of each neighborhood with their categories. For example:

Table 1 – Example of the Venues Data

Venue	Category
Spartacus Books	Bookstore
Trout Lake Fitness Centre	Gym / Fitness Center

This data will be retrieved from Foursquare which is one of the world largest sources of location and venues data. Foursquare API will be utilized to get and download the data.

1. Neighborhood Data – Vancouver

A dataset was created from the combination of two sources. The following figures show the process for generating the data table.

```
[2]: # Vancouver neighborhoods
van_neighborhoods = ['Arbutus Ridge', 'Cedar Cottage', 'Champlain Heights', 'Chinatown', 'Coal Harbour', 'Collingwood',
                     'Commercial Drive', 'Creekside', 'Downtown', 'Downtown Eastside', 'Dunbar-Southlands', 'Fairview',
                     'False Creek North', 'False Creek South', 'Gastown', 'Grandview-Woodland', 'Granville Island',
                     'Hastings-Sunrise', 'Hastings Crossing', 'Hastings East', 'Kensington-Cedar Cottage', 'Kerrisdale',
                     'Killarney', 'Kitsilano', 'Knight', 'Langara', 'Little Mountain', 'Main', 'Marpole', 'Mole Hill',
                     'Mount Pleasant', 'Musqueam', 'Oakridge', 'Quilchena', 'Renfrew-Collingwood', 'Riley Park',
                     'Shaughnessy', 'South Cambie', 'South Granville', 'South Hill', 'South Vancouver', 'Southlands',
                     'Southwest Marine', 'Sunrise', 'Sunset', 'Victoria-Fraserview', 'West Broadway', 'West End', 'West Point Grey', 'Yaletown']

[3]: df_van = pd.DataFrame(van_neighborhoods)
     df_van.columns = ['Neighborhood']
     df_van.head()
```

[3]:

	Neighborhood
0	Arbutus Ridge
1	Cedar Cottage
2	Champlain Heights
3	Chinatown
4	Coal Harbour

Figure 2 – Neighborhoods of Vancouver

```
[4]: #create a function to handle TimeOuts from Geocoder
from geopy.exc import GeocoderTimedOut
locator = Nominatim(user_agent = "bostonagent")

def do_geocode(address):
    try:
        return locator.geocode(address)
    except GeocoderTimedOut:
        return do_geocode(address)

[5]: neighborhoods = df_van.values.tolist()

latitude = []
longitude = []
for neighborhood in neighborhoods:
    print('-', end='')
    coord = do_geocode('{}, Vancouver'.format(neighborhood))

    #check to make sure all latitude and longitude values are present in the Nominatim API
    #handles the case where Nominatim returns a 'None' object because the neighborhood does not exist in their API

    if (coord == None):
        latitude.append('0')
        longitude.append('0')
    else:
        latitude.append(coord.latitude)
        longitude.append(coord.longitude)

#add coordinates columns to dataframe
df_van['Latitude'] = latitude
df_van['Longitude'] = longitude

df_van.head()
```

Figure 3 – Function to obtain Latitude and Longitude

The Figure 4 shows the resulting dataframe which contains data on 50 neighborhoods.

[5]:

	Neighborhood	Latitude	Longitude
0	Arbutus Ridge	49.240968	-123.167001
1	Cedar Cottage	49.251622	-123.064548
2	Champlain Heights	49.215266	-123.030915
3	Chinatown	49.279981	-123.104089
4	Coal Harbour	49.290375	-123.129281

Figure 4 –Vancouver neighborhood-data dataframe

Having data of the coordinates of Vancouver, it is possible to draw a map using Folium Python package of Vancouver and its neighborhoods. Figure 5 shows this map; each blue circles represents the location of one neighborhood.

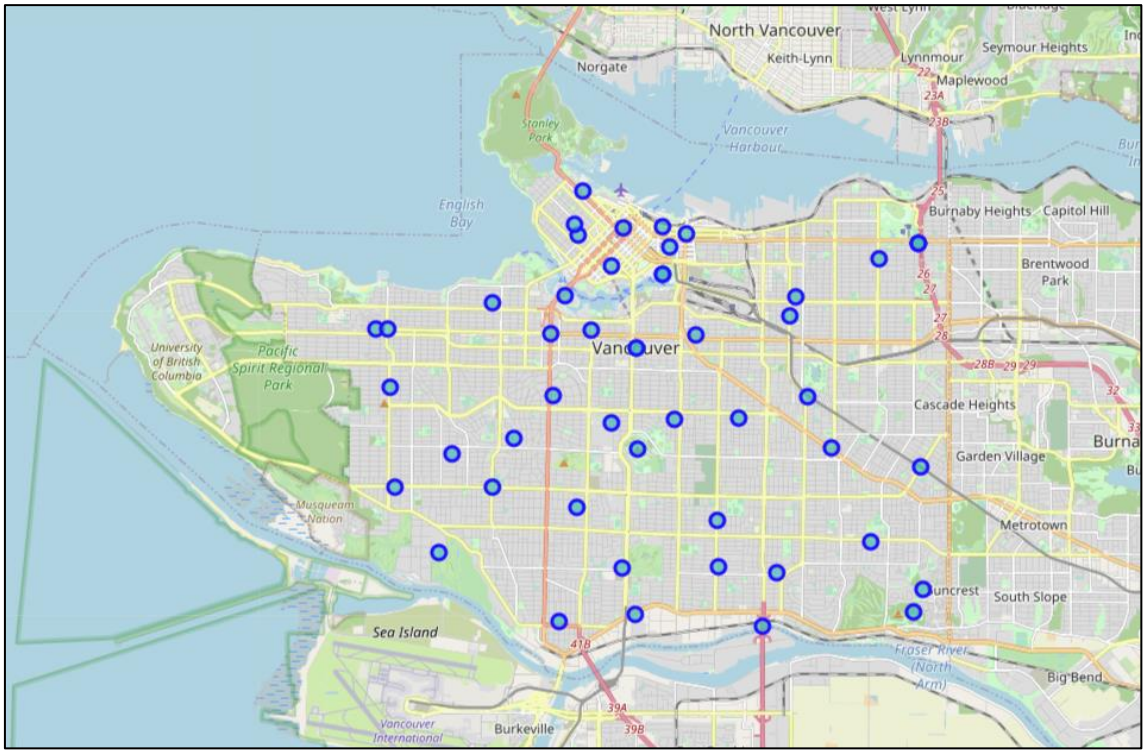


Figure 5 – A map of Vancouver and its neighborhoods.

1. Venues Data – Vancouver

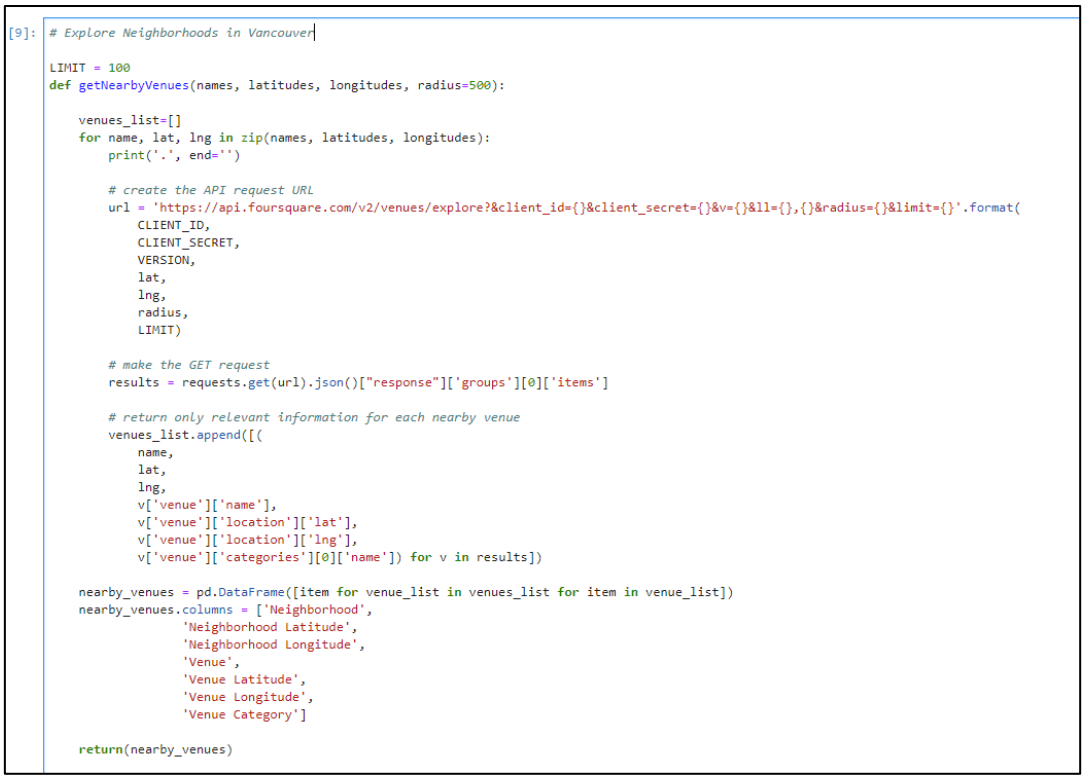
For the city, data that describes the venues of its neighborhoods and the categories of these venues is needed. Venues data will be retrieved from Foursquare which is a popular source of location and venue data. Foursquare API service will be utilized to access and download venues data. To retrieve data from Foursquare using their API, a URL should be prepared and used to request data related a specific location. An example URL is the following:

```
https://api.foursquare.com/v2/venues/search?
&client_id=1234&client_secret=1234&v=20180605&
ll=40.89470517661,-73.84720052054902&radius=500&limit=100
```

Figure 7 – URL example

Where search indicates the API endpoint used, client_id and client_secret are credentials used to access the API service and are obtained when registering a Foursquare developer account, v indicates the API version to use, ll indicates the latitude and longitude of the desired location, radius is the maximum distance in meters between the specified location and the retrieved venues, and limit is used to limit the number of returned results if necessary.

Figure 8 shows the code used to create a function that takes as input the names, latitudes, and longitudes of the neighborhoods, and returns a dataframe with information about each neighborhood and its venues. It creates an API URL for each neighborhood and retrieves data about the venues of that neighborhoods from Foursquare.



```
[9]: # Explore Neighborhoods in Vancouver

LIMIT = 100
def getNearbyVenues(names, latitudes, longitudes, radius=500):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print('.', end='')

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    return(nearby_venues)
```

Figure 8 – Code used to build a venues dataframe for a city neighborhood.

For the study restricting outdoors and recreation activities and restaurants, were necessary do add category ID (Outdoors & Recreation ID- 4d4b7105d754a06377d81259; and restaurants ID - 4d4b7105d754a06374d81259).

Using the function in Figure 8 with Vancouver neighborhood data retrieved data about 1700 venues in Vancouver neighborhoods. For each venue, venue name, category, latitude, and longitude were retrieved. The head of the dataframe returned by the function for Vancouver is shown in Figure 9. We can see that each row in the dataframe contains data about one venue: the venue name, coordinates (latitude and longitude), and category in addition to the neighborhood in which the venue is located and the coordinates of the neighborhood. Different numbers of venues were found in different neighborhoods.

[41]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Arbutus Ridge	49.240968	-123.167001	Butter Baked Goods	49.242209	-123.170381	Bakery
1	Arbutus Ridge	49.240968	-123.167001	The Haven	49.241377	-123.166331	Spa
2	Arbutus Ridge	49.240968	-123.167001	Barktholomews Pet Supplies	49.242746	-123.170193	Pet Store
3	Arbutus Ridge	49.240968	-123.167001	The Dragon's Layer	49.238518	-123.169029	Nightlife Spot
4	Arbutus Ridge	49.240968	-123.167001	The Heights Market	49.237902	-123.170949	Grocery Store
5	Cedar Cottage	49.251622	-123.064548	Commercial Street Cafe	49.252539	-123.068178	Café
6	Cedar Cottage	49.251622	-123.064548	Trout Lake Community Centre	49.255403	-123.065048	Gym
7	Cedar Cottage	49.251622	-123.064548	Trout Lake Fitness Centre	49.255601	-123.065317	Gym / Fitness Center
8	Cedar Cottage	49.251622	-123.064548	The Lower Mainland Childbearing Society	49.252836	-123.068136	Child Care Service
9	Cedar Cottage	49.251622	-123.064548	Flourist Mill & Bakery	49.253881	-123.068209	Bakery

Figure 9 – Venue dataframe for Vancouver.

Part III – Exploratory data Analysis

In this section, the dataset produced in the previous section will be explored via effective visualizations to understand the data better.

1. Most Common Venue Categories

What are the categories that have more venues than the others in Vancouver? To answer this question, the number of occurrences is counted for each venue category in Figure 10. After doing so, a bar plot can be used to visualize the popularity of the most common venue categories in the city.

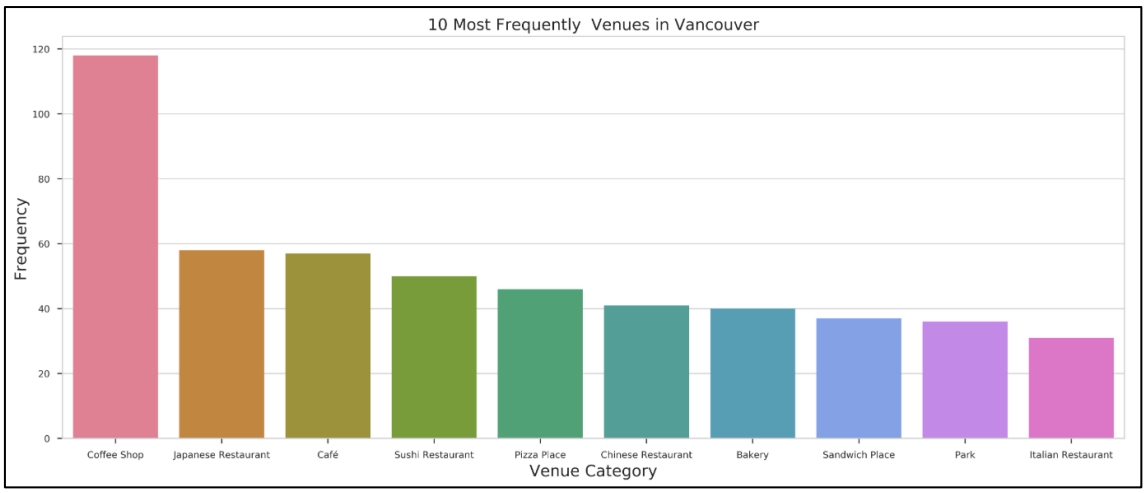


Figure 10 – Ten most frequently venues in Vancouver.

In Figure 11 and 12 we will see the more common outdoors and recreation activities and restaurants. After doing so, a bar plot can be used to visualize the popularity of the most common venue outdoors and recreation activities and restaurants in the city.

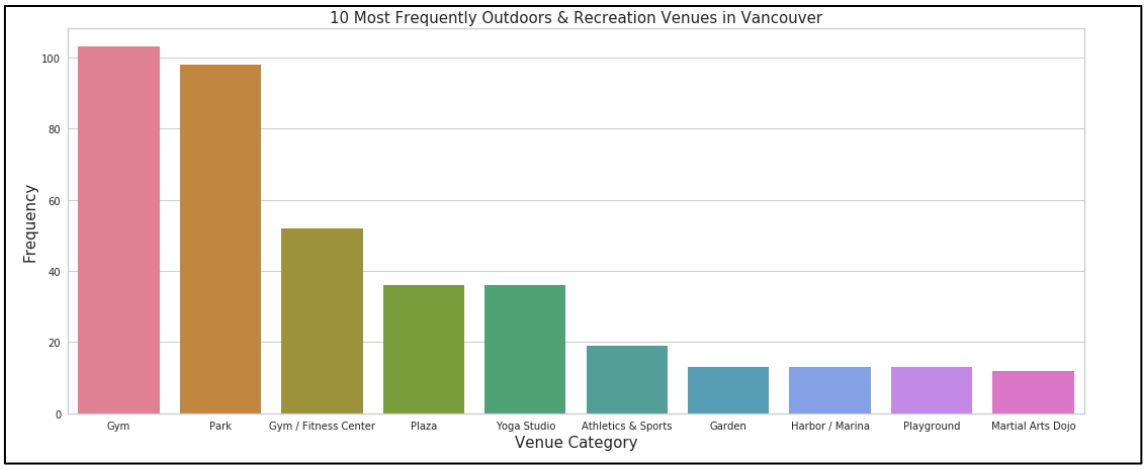


Figure 11 – Ten most frequently outdoors and recreation venues in Vancouver.

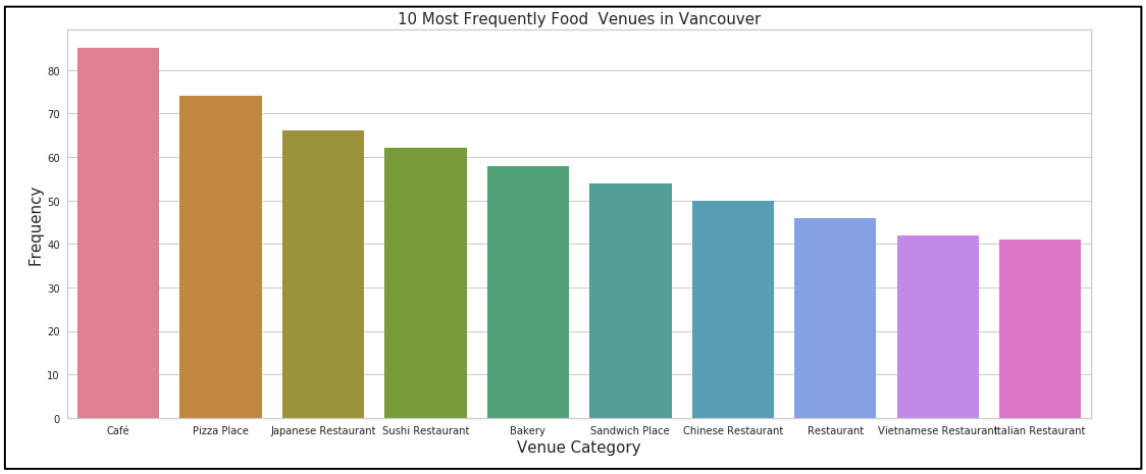


Figure 12 – Ten most frequently food venues in Vancouver.

2. Most Widespread Venue Categories

Now another question is to be answered: What are the venue categories that exist in more neighborhood? This question is different than the one mentioned in 1. To explain the difference with an example, suppose that there are 15 venues with the category “VR Games” and that these venues exist in 7 neighborhoods only out of 80 neighborhoods; also suppose that there are 10 venues with the category “Syrian Restaurant” and that these venues exist in 10 neighborhoods—each one of them in a different neighborhood. Then it can be said that the “VR Games” category is more common than “Syrian Restaurant” category because there are more venues under this category, and it can be said that the

“Syrian Restaurant” category is more widespread than the “VR Games” category because venues under this category exist in more neighborhoods than the other category.

Figure 13 shows the most widespread venue categories in Vancouver. It can be seen that the order of categories this time is different than that of the most common categories (Figure 10). The most widespread category is “Coffee Shop”; Coffee Shop exist in more than 30 of neighborhoods out of the 50 neighborhoods. After that comes the “Pizza Place” category with venues in 30 neighborhoods. In the third place comes the “Café” category with venues in ~25 neighborhoods.

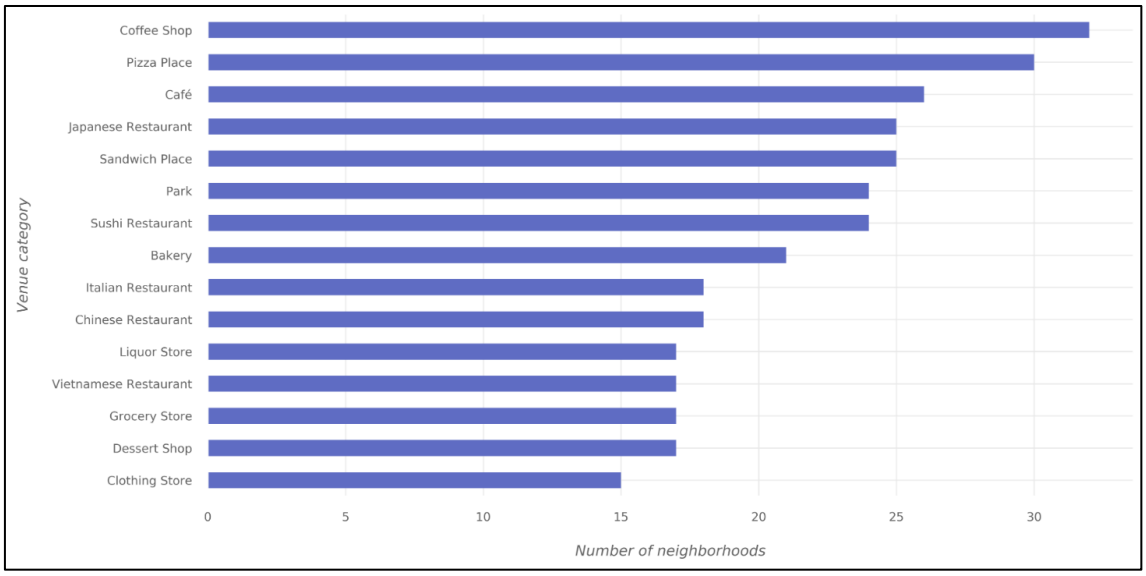


Figure 13 – Most widespread venue categories in Vancouver.

Figure 14 shows the most widespread outdoors and recreation venue in Vancouver. It can be seen that the order of categories this time is different than that of the most common outdoors and recreations categories (Figure 11). The most widespread category is “Park”; Park exist in more than 40 of neighborhoods out of the 50 neighborhoods. After that comes the “Gym” category and in the third place comes the “Gym / Fitness Center” category.

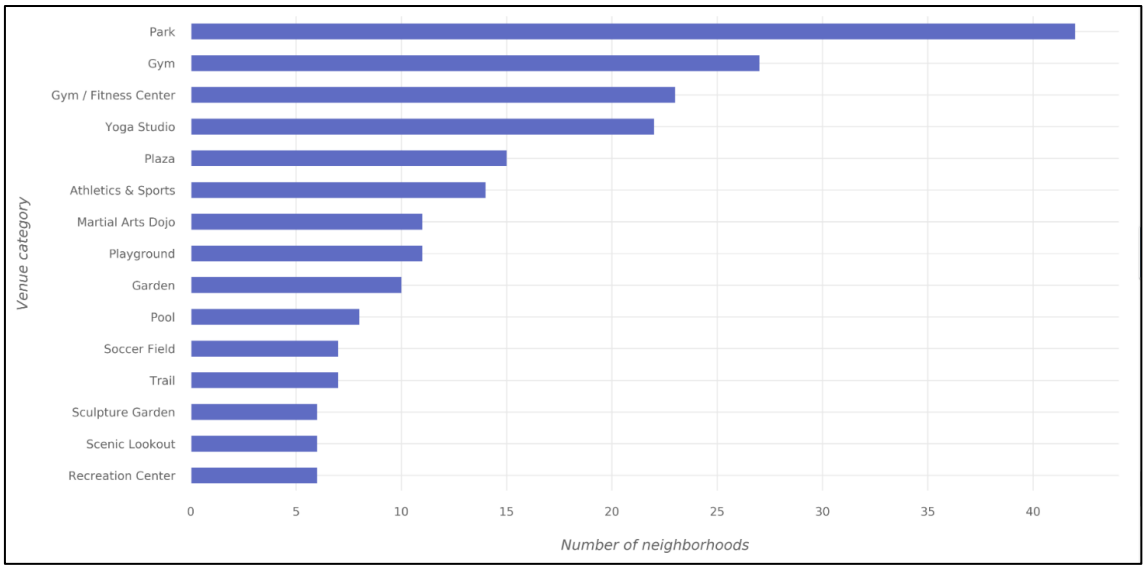


Figure 14 – Most widespread venue outdoors and recreations in Vancouver.

Figure 15 shows the most widespread outdoors and recreation venue in Vancouver. It can be seen that the order of categories this time is different than that of the most common foods place categories (Figure 12). The most widespread category is “Pizza Place”; Pizza Place exist in more than 30 of neighborhoods out of the 50 neighborhoods. After that comes the “Café” category with almost 30 places and in the third place comes the “Bakery” category in 25 neighborhoods.

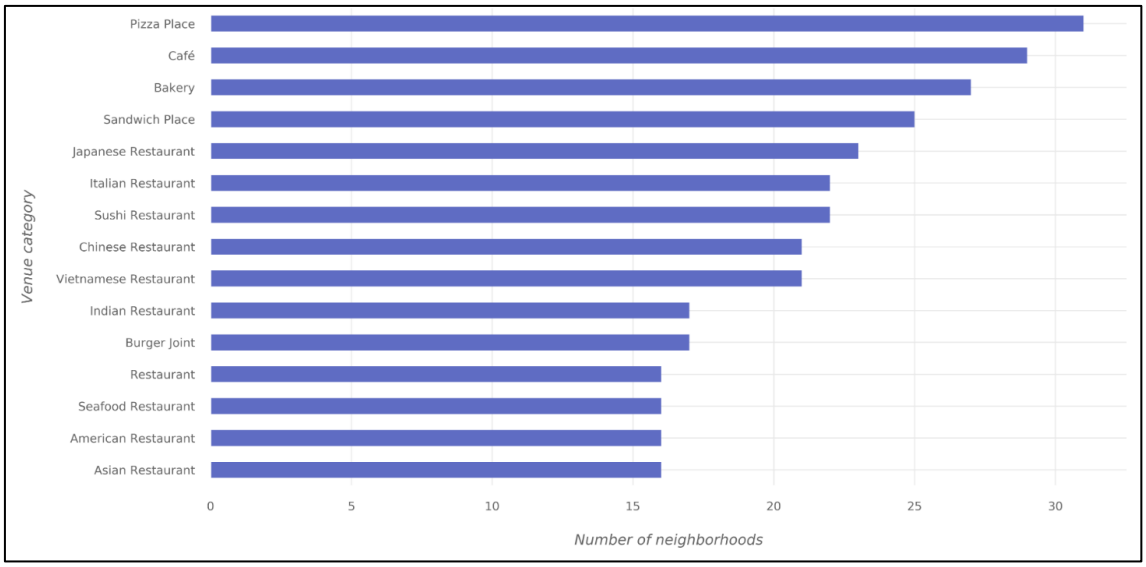


Figure 15 – Most widespread venue food in Vancouver.

Part IV – Clustering of Neighborhoods

In this section, clustering will be applied on Vancouver neighborhoods to find similar neighborhoods in the city. Clustering is the process of finding similar items in a dataset based on the characteristics (features) of items in the dataset. In particular, K-means clustering algorithm of the Scikit-learn Python library will be used. To be able to perform clustering, a dataset suitable for clustering is needed; the datasets described in Figure 9 is not ready to be used with clustering algorithms.

The goal of the clustering is to cluster neighborhoods based on the similarity of venue categories in the neighborhoods. This means that the two things of interest here are the neighborhood and the venue categories in the neighborhood. Thus, the following features will be selected out of the dataframes of Figure 9: “Neighborhood” and “Venue Category”. But still after that, the data is not ready for the clustering algorithm because the algorithm works with numerical features.

For that, one-hot encoding will be applied on the “Venue Category” feature and the result of the encoding will be used for clustering. One-hot encoding will be applied on the Vancouver data. After applying one-hot encoding on Vancouver data, the resulting dataframe looks like the one shown in Figure 16. For example, when looking at the first row in Figure 9, it can be seen that the venue category for that row is “Bakery”, so the column whose value is 1 for the fiest row in Figure 16 is the “Bakery” column; and the same applies for all rows.

[14]:

	Neighborhood	Accessories Store	Airport Terminal	American Restaurant	Amphitheater	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Australian Restaurant	Automotive Shop	BBQ Joint	Bagel Shop	Bakery	Bank	Bar	Baseball Field	Beach
0	Arbutus Ridge	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
1	Arbutus Ridge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Arbutus Ridge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Arbutus Ridge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Arbutus Ridge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 16 – The result of one hot enconding on Vancouver data.

The next step is aggregating the values for each neighborhood so that each neighborhood becomes represented by only one row. The aggregation will be done by

grouping rows by neighborhood and by taking the mean of the frequency of occurrence of each category. Figure 17 shows how the aggregated dataframe looks like for Vancouver.

[16]:

	Neighborhood	Accessories Store	Airport Terminal	American Restaurant	Amphitheater	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Australian Restaurant	Automotive Shop	BBQ Joint	Bagel Shop	Bakery	Bank	Bar	Baseball Field	Beach
0	Arbutus Ridge	0.0	0.000000	0.00	0.0	0.00	0.0	0.000000	0.00	0.0	0.0	0.00	0.0	0.200000	0.0	0.00	0.0	0.0
1	Cedar Cottage	0.0	0.000000	0.00	0.0	0.00	0.0	0.000000	0.00	0.0	0.0	0.00	0.0	0.111111	0.0	0.00	0.0	0.0
2	Champlain Heights	0.0	0.000000	0.00	0.0	0.00	0.0	0.000000	0.00	0.0	0.0	0.00	0.0	0.000000	0.0	0.00	0.0	0.0
3	Chinatown	0.0	0.000000	0.01	0.0	0.01	0.0	0.020000	0.01	0.0	0.0	0.01	0.0	0.030000	0.0	0.01	0.0	0.0
4	Coal Harbour	0.0	0.010204	0.00	0.0	0.00	0.0	0.010204	0.00	0.0	0.0	0.00	0.0	0.000000	0.0	0.00	0.0	0.0

Figure 17 – Part of the aggregated dataframe for Vancouver.

The clustering algorithm grouped neighborhoods of Vancouver in 5 clusters based on the similarity between their venues. The K value was calculated using Elbow method (Figure 18) and Figure 19 shows the resulting of dataset with clusters labels.

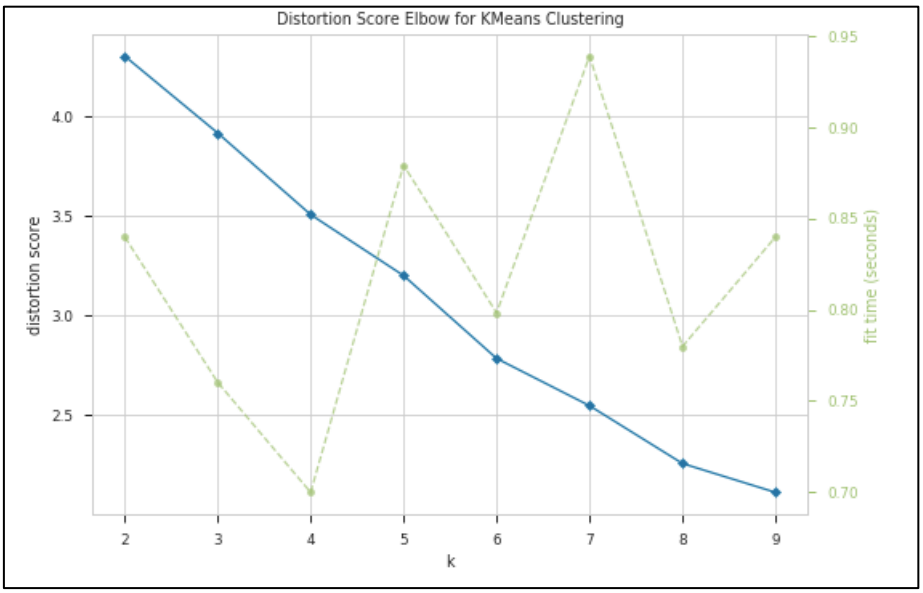


Figure 18 – Using Elbow method for the value of k.

	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Arbutus Ridge	49.240968	-123.167001	1	Spa	Pet Store	Nightlife Spot	Bakery	Grocery Store	Yoga Studio
1	Cedar Cottage	49.251622	-123.064548	1	Lake	Skating Rink	Child Care Service	Bakery	Park	Bookstore
2	Champlain Heights	49.215266	-123.030915	1	Video Store	Recreation Center	Park	Pizza Place	Bus Stop	Flower Shop
3	Chinatown	49.279981	-123.104089	1	Café	Coffee Shop	Sandwich Place	Pizza Place	Chinese Restaurant	Mexican Restaurant
4	Coal Harbour	49.290375	-123.129281	1	Japanese Restaurant	Coffee Shop	Ramen Restaurant	Dessert Shop	Café	Park

Figure 19 – Dataset with cluster labels.

Now, these clusters will be investigated to see the most common categories in each of them. Figures 20, 21, 22, 23 and 24 show each cluster; for each common category dataframe.

```
# Cluster 1
van_c1 = van_merged.loc[van_merged['Cluster Labels'] == 0,
                        van_merged.columns[[0] + list(range(5, van_merged.shape[1]))]]
van_c1.head()
```

	Neighborhood	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
44	Sunset	Dessert Shop	South Indian Restaurant	Home Service	Yoga Studio	Falafel Restaurant	Food Truck	Food Court	Food & Drink Shop	Food

Figure 20 – Cluster 1.

```
# Cluster 2
van_c2 = van_merged.loc[van_merged['Cluster Labels'] == 1,
                        van_merged.columns[[0] + list(range(5, van_merged.shape[1]))]]
van_c2.head()
```

	Neighborhood	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Arbutus Ridge	Pet Store	Nightlife Spot	Bakery	Grocery Store	Yoga Studio	Farmers Market	Food Truck	Food Court	Food & Drink Shop
1	Cedar Cottage	Skating Rink	Child Care Service	Bakery	Park	Bookstore	Gym	Gym / Fitness Center	Café	Fish Market
2	Champlain Heights	Recreation Center	Park	Pizza Place	Bus Stop	Flower Shop	Fast Food Restaurant	Field	Filipino Restaurant	Fish Market
3	Chinatown	Coffee Shop	Sandwich Place	Pizza Place	Chinese Restaurant	Mexican Restaurant	Bakery	Sushi Restaurant	Gastropub	German Restaurant
4	Coal Harbour	Coffee Shop	Ramen Restaurant	Dessert Shop	Café	Park	Restaurant	Breakfast Spot	Sushi Restaurant	Italian Restaurant

Figure 21 – Cluster 2.

```
# Cluster 3
van_c3 = van_merged.loc[van_merged['Cluster Labels'] == 2,
                        van_merged.columns[[0] + list(range(5, van_merged.shape[1]))]]
van_c3.head()
```

	Neighborhood	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
41	Southlands	Bus Stop	Yoga Studio	Fast Food Restaurant	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Food

Figure 22 – Cluster 3.

```
# Cluster 4
van_c4 = van_merged.loc[van_merged['Cluster Labels'] == 3,
                        van_merged.columns[[0] + list(range(5, van_merged.shape[1]))]]
van_c4.head()
```

	Neighborhood	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
18	Hastings Crossing	Theme Park Ride / Attraction	Fast Food Restaurant	Amphitheater	Bus Stop	Event Space	Soccer Field	Coffee Shop	Fish Market	Field
19	Hastings East	Theme Park Ride / Attraction	Fast Food Restaurant	Amphitheater	Bus Stop	Event Space	Soccer Field	Coffee Shop	Fish Market	Field
43	Sunrise	Theme Park Ride / Attraction	Fast Food Restaurant	Amphitheater	Bus Stop	Event Space	Soccer Field	Coffee Shop	Fish Market	Field

Figure 23 – Cluster 4.

```
# Cluster 5
van_c5 = van_merged.loc[van_merged['Cluster Labels'] == 4,
                        van_merged.columns[[0] + list(range(5, van_merged.shape[1]))]]
van_c5.head()
```

	Neighborhood	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
25	Langara	Pool	Yoga Studio	Farmers Market	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Food

Figure 24 – Cluster 5.

Part V – Conclusions

In this project, the neighborhoods of Vancouver were clustered into multiple groups based on the categories (types) of the venues in these neighborhoods. The results showed that there are venue categories that are more common in some cluster than the others; the most common venue categories differ from one cluster to the other. If a deeper analysis—taking more aspects into account—is performed, it might result in discovering different style in each cluster based on the most common categories in the cluster.