IBM/Coursera Capstone Project - The Battle of the Neighborhoods

Explore similar neighborhoods between Manhattan and Seattle

Ray Wong (https://www.linkedin.com/feed/)
2019 Aug.26 @ Kobe, Japan

Table of contents

Introduction

We make decisions everyday, big or small, right or wrong. And every decision we made, makes us who we are.
In year 1994, a 30-year-old young man decided to leave his lucrative job on Wall Street. He drove across the country all the way to an online bookstore. He named it Amazon. And the young man is Jeff Bezos. As he revealed later, he made this decision because o forecasting the rapid growth of Internet, also because of the fear of regret later in life.

Today due to the gold rush of AI, there is an urgent need for qualified IT talents. Many smart minds, including those Wall Street worl be thinking about starting a new career in one of those cities full of tech giants. But transferring to a new career, moving to a new ci very challenging, of course we would hope the new neighborhood we are moving to is similar to the one that we have been living in take us too long and too much to adapt to the new environment.

In this project, we are trying to solve such a problem. We are going to to explore the similarity between neighborhoods in Manhattar to come up with some insights as a reference to which neighborhood in Seattle should one choose if he is moving from a neighborh to Seattle.

To complete this task, we are going to use Foursquare location data to get the most common venue categories in each neighborho on this we will group the neighborhoods in custers with k-means clustering Machine Learning Algorithm, and present the clustering generated with Folium library.

You may wonder why Seattle, why not Silicon Valley. Well, not only just because it is the headquarter of Amazon(one member of the club), it is also the headquarter of Microsoft(another member of the 1 trillion dollar club), Zillow, Tableau, Expedia, etc. Other big tec Google, Facebook, Twitter all have offices in Seattle. What's more, the living cost, especially the housing price is so friendly if comp Silicon Valley. Not to mention the first Starbucks in Pike Place could be a big attraction to the coffee fun. Actually, even Starbucks is especially data science to strengthen its power. So if you are one of those Talents, make sure you take Seattle as an option before y decision.

Data

For this project, we will need to collect dataset below:

- Manhattan neighborhood data with latitude and longitude coordinates
- Seattle neighborhood data with latitude and longitude coordinates
- Foursquare location data: Most common categories

Manhattan neighborhood data

First we need the data of all neighborhoods in Manhattan, the coordinates of each neighborhood is also required since we will neee common venue categories nearby each neighborhood with the coordinates.

Luckily, we can find the the data of New York on link below https://geo.nyu.edu/catalog/nyu_2451_34572 (https://geo.nyu.edu/catalog/nyu_2451_34572)

On the right of the page we could download the file in different format, Shapefile, KMZ, GeoJSON, we will download a file in GeoJS then read the data we need (borough, neighbothood, latitude, longitude) into a pandas dataframe.

Since this data is for the whole New York, we will filter to get a subset of the neighborhoods of borough Manhattan only.

Collecting New York Neighborhood data

Contents

Through a glance at the json data, We notice that the info. we need are stored in the 'featuers' key. We will define a new variable ne include all the values under 'fearures' key.


```
{'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]},
 'geometry_name': 'geom',
 'properties': {'name': 'Wakefield',
  'stacked': 1,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'annoangle': 0.0,
  'borough': 'Bronx',
  'bbox': [-73.84720052054902,
   40.89470517661,
   -73.84720052054902,
   40.89470517661]}}
```


Now let's have a look at the first neighborhood in our neighborhood data.

We can see that the coordinate info is stored in ['geometry']['coordinates'], but the order is [longitude, latitude]. The name of borou ['properties']['borough'].
The name of the neighborhood is stored in ['properties']['name']

Next we will create an empty dataframe called 'ny_df' with columns 'Borough', 'Neighborhood', 'Latitude', 'Longitude'and use a for values of neighborhood.

We will print out the first 5 rows of the ny_df dataframe.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Collecting Manhattan Neighborhood data

The first 5 rows of ny_df dataframe looks good, next we will filter the 'Borough' column in the dataframe to get a subset of Manhatt 'mh_neigh' and print out the first 5 rows.

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

Since we will need to merge all the neighborhoods from Manhattan and Seattle together later, we will change the column name 'Bo 'Borough/City' so that we could use the same column to store both the value of Manhattan Borough and Seattle City.

	Borough/City	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

Double check values of column 'Borough'

```
Manhattan    40
Name: Borough/City, dtype: int64
```

Size of dataframe

```
The mh_neigh dataframe has 40 rows with 4 columns.
```

Check missing values

```
Borough/City    0
Neighborhood    0
Latitude        0
Longitude       0
dtype: int64
```

Check duplicates

```
0
```

Double check basic info. of the dataframe

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 4 columns):
Borough/City    40 non-null object
Neighborhood    40 non-null object
Latitude        40 non-null float64
Longitude       40 non-null float64
dtypes: float64(2), object(2)
memory usage: 1.3+ KB
```

Looks good, there is no missing value, no duplicates, all the data types are correct, let's save it to a csv file and name it 'manhattan_neighborhood.csv'.

Seattle neighborhood data

We did not find any dataset with Seattle neighborhoods as well as coordinates, there is also hardly a webpage from where we could info.
In this case, we will have to divide this task into 3 steps:

1. Get a dataset with Seattle neighborhood and zip code info.

- Get a dataset with Seattle coordinates and zip code info.
- Joint the 2 datasets above together

Get a dataset with Seattle neighborhood and zip code info

We can get such info. from Sub-Regional, City and Neighborhood Designations by Zip Code file (http://www.agingkingcounty.org/w content/uploads/sites/185/2016/09/SubRegZipCityNeighborhood.pdf)

But seems like some neighborhoods are not correct, we will modify the incorrect info. based on resources we get from some other area (http://seattlearea.com/zip-codes/).

Considerring that, Redmond, where The headquarter of Microsoft is located, is also a big attraction to the IT talents, and it is not fa we will also include Redmond in our dataset.

And then we get a working neighbborhood dataset in excel file. We will load the dataset to have a look at it.

	ZIP	City	Neighborhood
0	98052	Redmond	Redmond
1	98101	Seattle	Central Business District, First Hill
2	98102	Seattle	Capitol Hill
3	98103	Seattle	Greennwood, Freemont, Greenlake
4	98104	Seattle	Pioneer Square, International District

Get a dataset with Seattle coordinates and zip code info

We can get this dataset on United States Zip Codes (https://www.unitedstateszipcodes.org/wa/#zips-list).
We will load the data below to explore the dataset.

	zip	type	decommissioned	primary_city	acceptable_cities	unacceptable_cities	state	county	
0	501	UNIQUE	0	Holtsville	NaN	I R S Service Center	NY	Suffolk County	America/
1	544	UNIQUE	0	Holtsville	NaN	Irs Service Center	NY	Suffolk County	America/
2	601	STANDARD	0	Adjuntas	NaN	Colinas Del Gigante, Jard De Adjuntas, Urb San...	PR	Adjuntas Municipio	America/Pu
3	602	STANDARD	0	Aguada	NaN	Alts De Aguada, Bo Guaniquilla, Comunidad Las ...	PR	Aguada Municipio	America/Pu
4	603	STANDARD	0	Aguadilla	Ramey	Bda Caban, Bda Esteves, Bo Borinquen, Bo Ceiba...	PR	Aguadilla Municipio	America/Pu

We can see that there are a lot of columns in the dataset that we don't need, we will just get a subset dataframe.

Let's again print out the first 5 rows.

	ZIP	Latitude	Longitude
41670	98052	47.68	-122.12
41671	98053	47.66	-122.01
41687	98073	47.67	-122.11
41697	98101	47.61	-122.33
41698	98102	47.63	-122.32

Joint the 2 datasets

Joint the 2 datasets above together to get a dataset of Seattle neighborhoods with coordinates info.

	ZIP	City	Neighborhood	Latitude	Longitude
0	98052	Redmond	Redmond	47.68	-122.12
1	98101	Seattle	Central Business District, First Hill	47.61	-122.33
2	98102	Seattle	Capitol Hill	47.63	-122.32
3	98103	Seattle	Greennwood, Freemont, Greenlake	47.67	-122.34
4	98104	Seattle	Pioneer Square, International District	47.60	-122.32

Check value counts of neighborhoods

```
Downtown                                  10
Northeast                                  4
Southwest                                  3
Northwest                                  3
Ballard                                    2
Delridge                                   2
Belltown                                   1
Northgate                                  1
South Lake Union, East Queen Anne          1
Redmond                                    1
University District,  Laurelhurst          1
Rainier Valley, Columbia City              1
Industrial District                        1
Greennwood, Freemont, Greenlake            1
Madison Park, Capitol Hill                 1
Capitol Hill                               1
Central Business District, First Hill      1
Broadview                                  1
Central                                    1
Duwamish                                   1
N. Beacon Hill, Mt. Baker                  1
Beacon Hill                                1
Queen Anne/Magnolia                        1
Magnolia                                   1
Pioneer Square, International District     1
West Seattle, Alki Beach                   1
Name: Neighborhood, dtype: int64
```

We can see that while one zip code may encompass more than one neighborhood, one neighborhood may also contain multiple zip group by the neighborhood, assign the mean value of coordinates of the neighborhood to the latitude and longitude variable, then v dataframe.

	Neighborhood	Latitude	Longitude
0	Ballard	47.675	-122.38
1	Beacon Hill	47.540	-122.31
2	Belltown	47.620	-122.35
3	Broadview	47.730	-122.37
4	Capitol Hill	47.630	-122.32

We will add a column 'Borough/City' to the dataset and assign 'Seattle' to the values of the column.

	Borough/City	Neighborhood	Latitude	Longitude
0	Seattle	Ballard	47.675	-122.38
1	Seattle	Beacon Hill	47.540	-122.31
2	Seattle	Belltown	47.620	-122.35
3	Seattle	Broadview	47.730	-122.37
4	Seattle	Capitol Hill	47.630	-122.32

Size of the dataframe

```
The sea_neigh dataframe has 26 rows with 4 columns.
```

Check missing values

```
Borough/City    0
Neighborhood    0
Latitude        0
Longitude       0
dtype: int64
```

Check duplicates

```
0
```

Check basic info of the dataframe

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26 entries, 0 to 25
Data columns (total 4 columns):
Borough/City    26 non-null object
Neighborhood    26 non-null object
Latitude        26 non-null float64
Longitude       26 non-null float64
dtypes: float64(2), object(2)
memory usage: 912.0+ bytes
```

It looks good, no missing value, no duplicates, all the data types are correct, we will save it to a csv file named 'seattle_neighborhoo

Merged Manhattan and Seattle neighborhood data

Since in this project we are going to do clustering on the neighborhoods of both Manhattan and Seattle, we will need to merge the t
together to get a new working dataset.

Load manhattan neighborhood data and seattle neighborhood data

	Borough/City	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

	Borough/City	Neighborhood	Latitude	Longitude
0	Seattle	Ballard	47.675	-122.38
1	Seattle	Beacon Hill	47.540	-122.31
2	Seattle	Belltown	47.620	-122.35
3	Seattle	Broadview	47.730	-122.37
4	Seattle	Capitol Hill	47.630	-122.32

Merge the Manhattan neighborhood data and Seattle neighborhood data

	Borough/City	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

Check size of the merged dataset

```
The merged mh_sea dataset has 66 rows with 4 columns.
```

check missing values

```
Borough/City    0
Neighborhood    0
Latitude        0
Longitude       0
dtype: int64
```

Check duplicates

```
0
```

Check basic info. of the dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 66 entries, 0 to 65
Data columns (total 4 columns):
Borough/City    66 non-null object
Neighborhood    66 non-null object
Latitude        66 non-null float64
Longitude       66 non-null float64
dtypes: float64(2), object(2)
memory usage: 2.1+ KB
```

The merged working dataset looks good, no missing values, no duplicates, all data types are correct. We will save it to a csv file for

Foursquare location data

After we get the working data of Manhattan neighborhoods and Seattle neighborhoods, we could use Foursquare API to get the mo
categories nearby each neighborhood in Manhattan and Seattle. And then group the neighborhood in custers based on the most co
categories with k-mean clusters Machine Learning Algorithm. For generating the nearby venues, we will choose a radius of 1km an
100 venues.

To start up with Foursquares, we will define Foursquare Credentials and Version first

```
Your credentails:
CLIENT_ID: 1QLDTWGYGVIST1KCL0TY0ADVUHPVTDIUGLAFXPIWOQZEXNWW
CLIENT_SECRET:2QP3AWK3ACZVGBULSKRPG0HS50ZM4MZ2VXVXQDA44FR1H0FZ
```

Now let's get the top 100 venues within 1km for each neighborhood

	City/Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venu Longitud
0	Manhattan	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.90620
1	Manhattan	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.91027
2	Manhattan	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.90893
3	Manhattan	Marble Hill	40.876551	-73.91066	Sam's Pizza	40.879435	-73.90585
4	Manhattan	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.90558

Size of the venues dataset

```
The venues dataset has 5605 rows and 8 columns.
```

Missing values

```
City/Borough              0
Neighborhood              0
Neighborhood Latitude     0
Neighborhood Longitude    0
Venue                     0
Venue Latitude            0
Venue Longitude           0
Venue Category            0
dtype: int64
```

Duplicates

0

Basic Info. of the venues dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5605 entries, 0 to 5604
Data columns (total 8 columns):
City/Borough              5605 non-null object
Neighborhood              5605 non-null object
Neighborhood Latitude     5605 non-null float64
Neighborhood Longitude    5605 non-null float64
Venue                     5605 non-null object
Venue Latitude            5605 non-null float64
Venue Longitude           5605 non-null float64
Venue Category            5605 non-null object
dtypes: float64(4), object(4)
memory usage: 350.4+ KB
```

This dataset looks good, it has the info. we need, with no missing value, no duplicates. And the data types look good. We will save 'nearby_venues.csv'.

Methodology

In this project, we are trying to explore the similarity between neighborhoods in Manhattan and Seattle.

The stakeholders of this project would be talents living in Manhattan who are thinking about starting a new IT career(especially AI) i

To complete this task, we will do the following steps in this project:

- Collecting data: Including Manhattan neighborhood data, Seattle Neighborhood data, venues nearby each neighbothood whicl by Foursquare API.
- Exploratory data analysis to better understand our working dataset
- Creating Map of Manhattan and map of Seattle with Folium library.
- Clustering of the neighborhoods: We will use K means clustering machine learning algorithm since it is fast and especially whe many variables.
- Presenting the clustering result on the map

Analysis

Exploratory Data Analysis

In the section, we will load and the nearby venues data and do an exploratory data analysis on it.

Neighborhoods

First we will check whether the venues dataset has venues for all 66 neighborhoods, 40 in Manhattan and 26 in Seattle.

```
The venues dataset has venues for 66 unique neighborhoods.
```

```
The dataset has 40 unique neighborhoods in Manhattan and 26 unique neighborhoods in Sea
```

Venues

First Let's have a look at how many venues are returned for each neighbohood

Contents

	City/Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Ve Longit
0	Manhattan	Battery Park City	100	100	100	100	
1	Manhattan	Carnegie Hill	100	100	100	100	
2	Manhattan	Central Harlem	100	100	100	100	
3	Manhattan	Chelsea	100	100	100	100	

Contents

	City/Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Ve Longit
4	Manhattan	Chinatown	100	100	100	100	
5	Manhattan	Civic Center	100	100	100	100	
6	Manhattan	Clinton	100	100	100	100	
7	Manhattan	East Harlem	100	100	100	100	
8	Manhattan	East Village	100	100	100	100	
9	Manhattan	Financial District	100	100	100	100	
10	Manhattan	Flatiron	100	100	100	100	
11	Manhattan	Gramercy	100	100	100	100	
12	Manhattan	Greenwich Village	100	100	100	100	
13	Manhattan	Hamilton Heights	100	100	100	100	
14	Manhattan	Hudson Yards	100	100	100	100	
15	Manhattan	Inwood	100	100	100	100	
16	Manhattan	Lenox Hill	100	100	100	100	
17	Manhattan	Lincoln Square	100	100	100	100	
18	Manhattan	Little Italy	100	100	100	100	
19	Manhattan	Lower East Side	100	100	100	100	
20	Manhattan	Manhattan Valley	100	100	100	100	
21	Manhattan	Manhattanville	100	100	100	100	
22	Manhattan	Marble Hill	84	84	84	84	
23	Manhattan	Midtown	100	100	100	100	
24	Manhattan	Midtown South	100	100	100	100	
25	Manhattan	Morningside Heights	100	100	100	100	
26	Manhattan	Murray Hill	100	100	100	100	
27	Manhattan	Noho	100	100	100	100	
28	Manhattan	Roosevelt Island	100	100	100	100	
29	Manhattan	Soho	100	100	100	100	
...	...	...	...	...	...	...	
36	Manhattan	Upper West Side	100	100	100	100	
37	Manhattan	Washington Heights	100	100	100	100	
38	Manhattan	West Village	100	100	100	100	
39	Manhattan	Yorkville	100	100	100	100	
40	Seattle	Ballard	100	100	100	100	
41	Seattle	Beacon Hill	20	20	20	20	
42	Seattle	Belltown	100	100	100	100	
43	Seattle	Broadview	4	4	4	4	
44	Seattle	Capitol Hill	80	80	80	80	
45	Seattle	Central	64	64	64	64	
46	Seattle	Central Business District, First Hill	100	100	100	100	
47	Seattle	Delridge	22	22	22	22	
48	Seattle	Downtown	100	100	100	100	
49	Seattle	Duwamish	56	56	56	56	
50	Seattle	Greennwood, Freemont, Greenlake	86	86	86	86	
51	Seattle	Industrial District	56	56	56	56	
52	Seattle	Madison Park, Capitol Hill	52	52	52	52	
53	Seattle	Magnolia	18	18	18	18	
54	Seattle	N. Beacon Hill, Mt. Baker	68	68	68	68	
55	Seattle	Northeast	95	95	95	95	
56	Seattle	Northgate	21	21	21	21	
57	Seattle	Northwest	71	71	71	71	
58	Seattle	Pioneer Square, International District	100	100	100	100	
59	Seattle	Queen Anne/Magnolia	59	59	59	59	

	City/Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Ve Longit
60	Seattle	Rainier Valley, Columbia City	27	27	27	27	
61	Seattle	Redmond	100	100	100	100	
62	Seattle	South Lake Union, East Queen Anne	100	100	100	100	
63	Seattle	Southwest	8	8	8	8	
64	Seattle	University District, Laurelhurst	100	100	100	100	
65	Seattle	West Seattle, Alki Beach	14	14	14	14	

66 rows × 8 columns

Foursquare API returned 100 venues for most of the neighborhoods, but returned less than 100 venues for some of the neighborho look at which neighborhood does not have 100 venues.

	City/Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Ve Longit
22	Manhattan	Marble Hill	84	84	84	84	
41	Seattle	Beacon Hill	20	20	20	20	
43	Seattle	Broadview	4	4	4	4	
44	Seattle	Capitol Hill	80	80	80	80	
45	Seattle	Central	64	64	64	64	
47	Seattle	Delridge	22	22	22	22	
49	Seattle	Duwamish	56	56	56	56	
50	Seattle	Greennwood, Freemont, Greenlake	86	86	86	86	
51	Seattle	Industrial District	56	56	56	56	
52	Seattle	Madison Park, Capitol Hill	52	52	52	52	
53	Seattle	Magnolia	18	18	18	18	
54	Seattle	N. Beacon Hill, Mt. Baker	68	68	68	68	
55	Seattle	Northeast	95	95	95	95	
56	Seattle	Northgate	21	21	21	21	
57	Seattle	Northwest	71	71	71	71	
59	Seattle	Queen Anne/Magnolia	59	59	59	59	
60	Seattle	Rainier Valley, Columbia City	27	27	27	27	
63	Seattle	Southwest	8	8	8	8	
65	Seattle	West Seattle, Alki Beach	14	14	14	14	

Size of vanue_count dataframe

```
(19, 8)
```

There are 19 neighborhoods have less than 100 venues, some neighorhoods even only have less than 10 venues returned. Among t neighborhoods, one is from Manhattan, the other 18 are from Seattle. This may impact our clustering to some extent, to reduce this only using the top 10 categories for clustering.

Categories

Next we all explore the venue categories returned by Foursquare API.

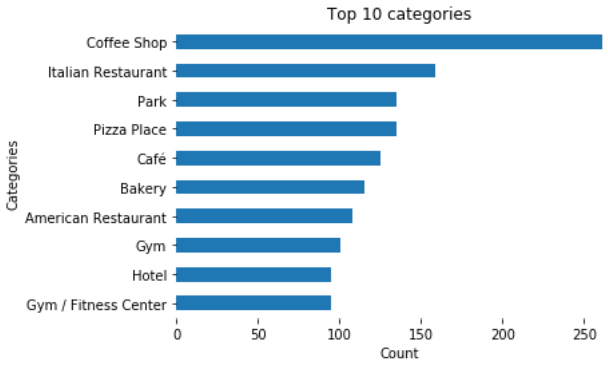
How many unique categories are returned in total?

```
There are 376 uniques categories.
```

How many unique categories for Manhattan and Seattle?

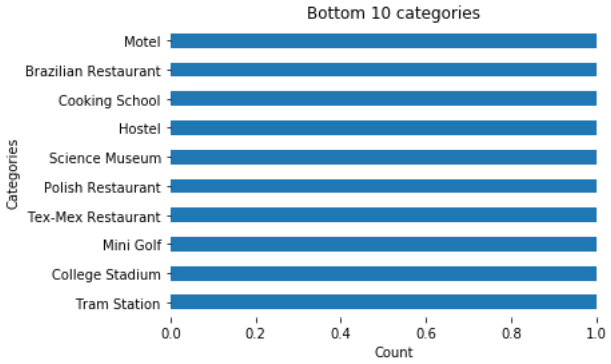
```
Manhattan has 310 unique venue categories
Seattle has 267 unique venue categories
```

Visualize the top 10 categories



From the bar chart above, it is not difficult to figure out that most of the top 10 categories are related to Restaurant, Cafe that kind

Visualize Bottom 10 categories



The bottom 10 categories all only have 1 count. Let's take a look at all those categories whose count is 1.

```
array(['Zoo', 'Photography Studio', 'TV Station', 'Radio Station',
       'Post Office', 'North Indian Restaurant', 'Piano Bar',
       'Kebab Restaurant', 'Ukrainian Restaurant', 'Daycare', 'Locksmith',
       'Costume Shop', 'Storage Facility', 'Herbs & Spices Store',
       'Arcade', 'Irish Pub', 'Insurance Office', 'Metro Station',
       'Lighthouse', 'Pedestrian Plaza', 'Multiplex', 'Nature Preserve',
       'Board Shop', 'Fabric Shop', 'Modern European Restaurant',
       'Tour Provider', 'Volleyball Court', 'Ski Area',
       'Spiritual Center', 'Outdoors & Recreation', 'Church',
       'Veterinarian', 'Bed & Breakfast', 'Gay Bar', 'Synagogue', 'River',
       'Heliport', 'Bowling Alley', 'Construction & Landscaping',
       'Swiss Restaurant', 'Malay Restaurant', 'Child Care Service',
       'Dry Cleaner', 'Indonesian Restaurant', 'Cantonese Restaurant',
       'Laundry Service', 'Basketball Stadium', 'Neighborhood',
       'Bus Line', 'Sports Club', 'Tennis Stadium',
       'Theme Park Ride / Attraction', 'Alternative Healer', 'Reservoir',
       'Casino', 'Student Center', 'Fishing Store', 'Circus', 'Motel',
       'Brazilian Restaurant', 'Cooking School', 'Hostel',
       'Science Museum', 'Polish Restaurant', 'Tex-Mex Restaurant',
       'Mini Golf', 'College Stadium', 'Tram Station'], dtype=object)
```

Ok, that's pretty much it about the exploratory analysis, next let's dive deeper into the dataset.

Top 10 categories for each Neighborhood

In this section, we will first use the one hot encoding to each categories into a seperate column and calculate the frequency of occu
category. Then we will slice the top 10 categories for clustering.

One hot encoding

	Neighborhood	ATM	Accessories Store	Adult Boutique	African Restaurant	Airport	Airport Terminal	Alternative Healer	American Restaurant	Animal Shelter	...	W
0	Marble Hill	0	0	0	0	0	0	0	0	0	...	
1	Marble Hill	0	0	0	0	0	0	0	0	0	...	
2	Marble Hill	0	0	0	0	0	0	0	0	0	...	
3	Marble Hill	0	0	0	0	0	0	0	0	0	...	
4	Marble Hill	0	0	0	0	0	0	0	0	0	...	

5 rows × 376 columns

Frequency of occurrence of each category

	Neighborhood	ATM	Accessories Store	Adult Boutique	African Restaurant	Airport	Airport Terminal	Alternative Healer	American Restaurant	Animal Shelter	...	W
0	Ballard	0.01	0.0	0.0	0.0	0.0	0.00	0.0	0.00	0.0	...	
1	Battery Park City	0.00	0.0	0.0	0.0	0.0	0.00	0.0	0.03	0.0	...	
2	Beacon Hill	0.00	0.0	0.0	0.0	0.1	0.15	0.0	0.00	0.0	...	
3	Belltown	0.00	0.0	0.0	0.0	0.0	0.00	0.0	0.02	0.0	...	
4	Broadview	0.00	0.0	0.0	0.0	0.0	0.00	0.0	0.00	0.0	...	

5 rows × 376 columns

Top 10 categories for each neighborhood

	Neighborhood	1st Most Common Venue	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	9 C
0	Ballard	Coffee Shop	Ice Cream Shop	Mexican Restaurant	Bakery	Burger Joint	Gym	Cocktail Bar	Thai Restaurant	Viet Res
1	Battery Park City	Park	Coffee Shop	Wine Shop	Gym	American Restaurant	Plaza	Memorial Site	Gym / Fitness Center	I
2	Beacon Hill	Airport Terminal	Café	Airport	Fried Chicken Joint	Sandwich Place	French Restaurant	Brewery	Gas Station	Miscel
3	Belltown	Coffee Shop	Exhibit	Pizza Place	Bar	Sculpture Garden	Vietnamese Restaurant	Mexican Restaurant	Movie Theater	
4	Broadview	Trail	Golf Course	Concert Hall	Fish Market	Exhibit	Fabric Shop	Falafel Restaurant	Farmers Market	Fa Res

Use K means to Cluster Neighborhoods into 6 clusters on map

In this section, we will first create maps for Manhattan and Seattle, do the k means clustering and then superimpose the clustering of Manhattan and Seattle to for visualization.

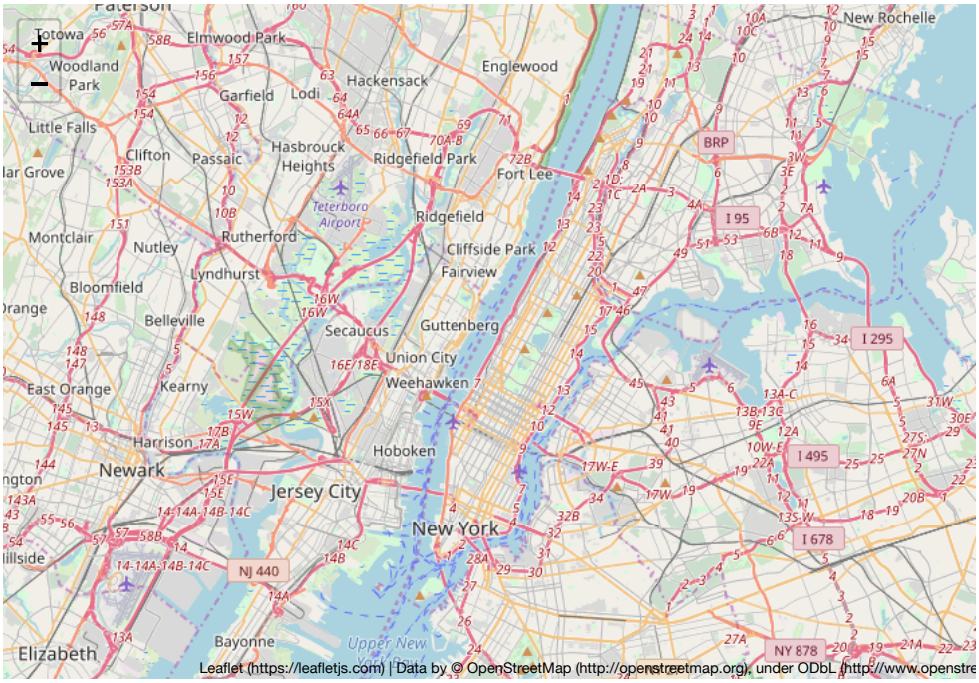
Map of Manhattan

First, let's get the coordinates of Manhattan, using the Nominatim Function.

```
The geograpical coordinate of Manhattan are 40.7900869, -73.9598295.
```

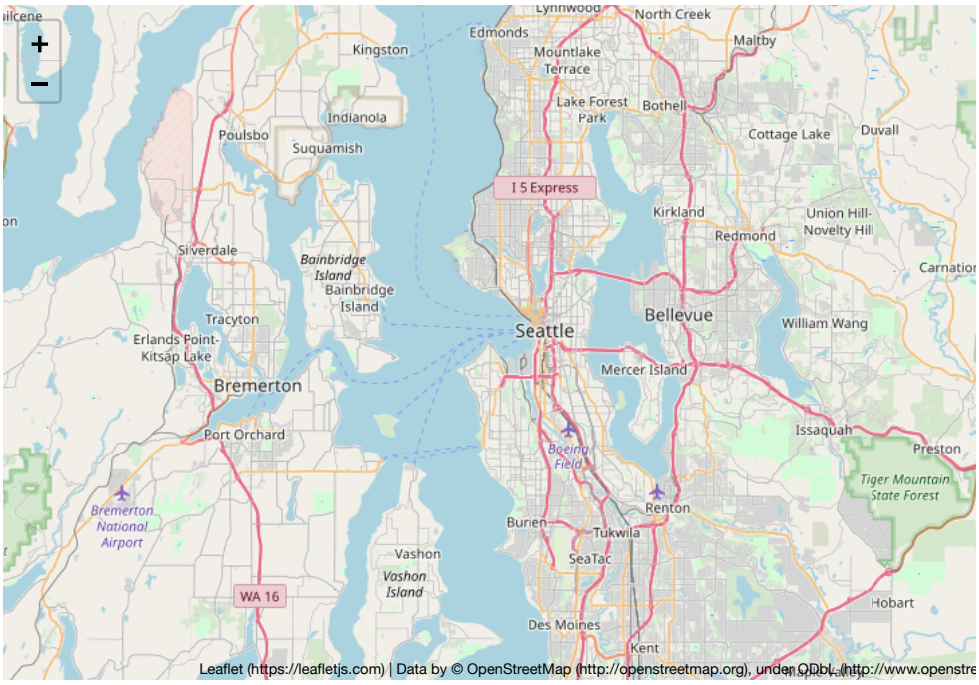
Then we will create a map of Manhattan with folium library, set the zoom_start to be 11 so we have a good view of the whole Manh

Contents



Map of Seattle

With the same steps above, we will create another map for Seattle.
First, let's get the coordinates of Manhattan, using the Nominatim Function.
Then we create a map of Seattle with folium library.



K Means clustering and visualization

Next we will use the k means clustering machine learning algorithm to cluster all the neighborhoods into 5 clusters and examine the

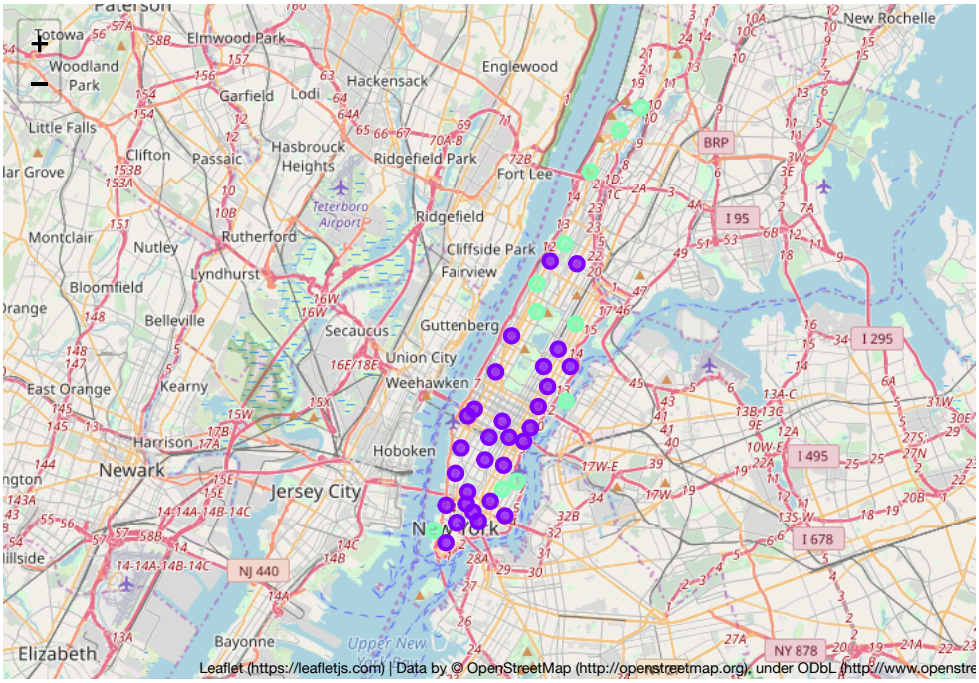
Set up k means clustering and check the first 10 cluster labels

```
array([3, 3, 1, 3, 2, 1, 1, 3, 1, 1], dtype=int32)
```

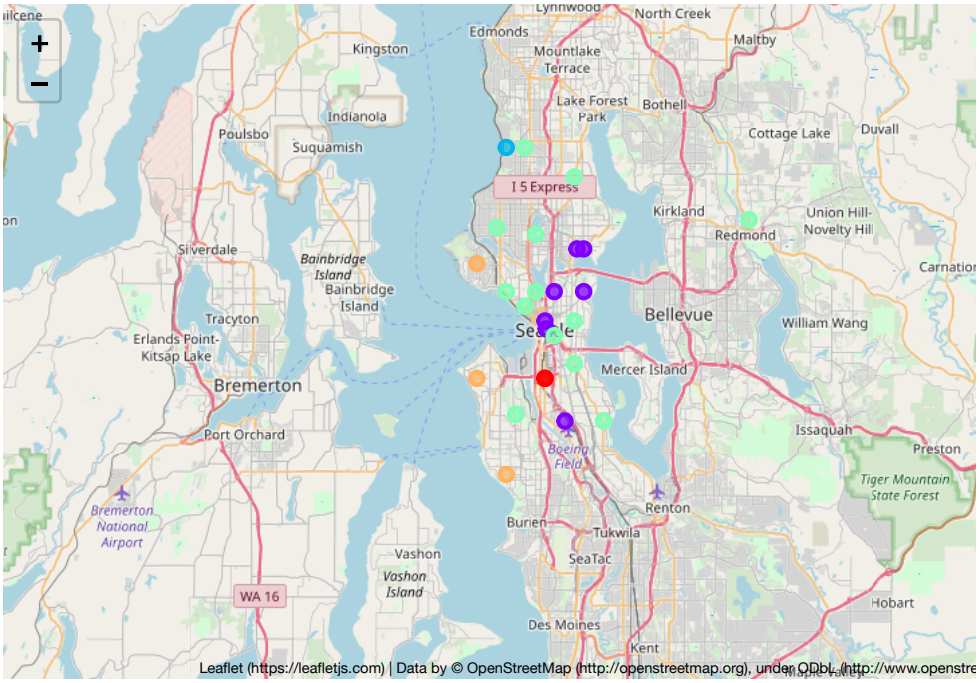
Create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood.

	Borough/City	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Con V
0	Manhattan	Marble Hill	40.876551	-73.910660	3	Park	Pizza Place	Supermarket	Mexican Restaurant	Sp Resta
1	Manhattan	Chinatown	40.715618	-73.994279	1	Sandwich Place	Cocktail Bar	Café	Ice Cream Shop	Win
2	Manhattan	Washington Heights	40.851903	-73.936900	3	Pizza Place	Latin American Restaurant	Bakery	Mexican Restaurant	Sp Resta
3	Manhattan	Inwood	40.867684	-73.921210	3	Latin American Restaurant	Pizza Place	Deli / Bodega	Mexican Restaurant	
4	Manhattan	Hamilton Heights	40.823604	-73.949688	3	Coffee Shop	Bar	Café	Mexican Restaurant	Ethi Resta

Let's show the clusters on the map of Manhattan.



Let's show the clusters on the map of Seattle.



From the map above, we can see that the neighborhoods in Manhattan are all in two clusters, which is cluster 1 and cluser 3.

Next we will print out the list of cluster 1 and cluster 3 to have a look.

Cluster 1

Contents

	Borough/City	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	
1	Manhattan	Chinatown	40.715618	-73.994279	1	Sandwich Place	Cocktail Bar	Café	Ice Cream Shop	
5	Manhattan	Manhattanville	40.816934	-73.957385	1	Park	Mexican Restaurant	Café	American Restaurant	I
6	Manhattan	Central Harlem	40.815976	-73.943211	1	Southern / Soul Food Restaurant	African Restaurant	Café	Gym / Fitness Center	I
8	Manhattan	Upper East Side	40.775639	-73.960508	1	Exhibit	Italian Restaurant	Gym / Fitness Center	Bakery	C
9	Manhattan	Yorkville	40.775930	-73.947118	1	Italian Restaurant	Gym	Coffee Shop	Ice Cream Shop	F
10	Manhattan	Lenox Hill	40.768113	-73.958860	1	Sushi Restaurant	Italian Restaurant	Coffee Shop	Gym / Fitness Center	I
12	Manhattan	Upper West Side	40.787658	-73.977059	1	Italian Restaurant	Wine Bar	Ice Cream Shop	Indian Restaurant	
13	Manhattan	Lincoln Square	40.773529	-73.985338	1	Coffee Shop	Gym / Fitness Center	Gym	Jazz Club	I
14	Manhattan	Clinton	40.759101	-73.996119	1	Theater	Hotel	Coffee Shop	American Restaurant	Gyr
15	Manhattan	Midtown	40.754691	-73.981669	1	Theater	Gym	Hotel	Coffee Shop	I
16	Manhattan	Murray Hill	40.748303	-73.978332	1	Gym / Fitness Center	Korean Restaurant	Japanese Restaurant	Pizza Place	C
17	Manhattan	Chelsea	40.744035	-74.003116	1	Art Gallery	Coffee Shop	American Restaurant	Seafood Restaurant	I
18	Manhattan	Greenwich Village	40.726933	-73.999914	1	Italian Restaurant	Spa	Gym	Seafood Restaurant	I
20	Manhattan	Lower East Side	40.717807	-73.980890	1	Italian Restaurant	Ice Cream Shop	Boutique	Wine Bar	I
21	Manhattan	Tribeca	40.721522	-74.010683	1	Park	American Restaurant	Hotel	Spa	C

Contents

	Borough/City	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	
22	Manhattan	Little Italy	40.719324	-73.997305	1	Clothing Store	Hotel	Coffee Shop	Men's Store	I
23	Manhattan	Soho	40.722184	-74.000657	1	Italian Restaurant	French Restaurant	Men's Store	Women's Store	
24	Manhattan	West Village	40.734434	-74.006180	1	Italian Restaurant	American Restaurant	New American Restaurant	Wine Bar	
27	Manhattan	Gramercy	40.737210	-73.981376	1	American Restaurant	Juice Bar	New American Restaurant	Gym	Mec
29	Manhattan	Financial District	40.707107	-74.010665	1	Coffee Shop	Pizza Place	American Restaurant	Café	C
30	Manhattan	Carnegie Hill	40.782683	-73.953256	1	Coffee Shop	Yoga Studio	Italian Restaurant	Bakery	F
31	Manhattan	Noho	40.723259	-73.988434	1	Italian Restaurant	Japanese Restaurant	Hotel	Cocktail Bar	F
32	Manhattan	Civic Center	40.715229	-74.005415	1	French Restaurant	Coffee Shop	Bakery	Cocktail Bar	I
33	Manhattan	Midtown South	40.748510	-73.988713	1	Korean Restaurant	Gym / Fitness Center	Hotel	Yoga Studio	
34	Manhattan	Sutton Place	40.760280	-73.963556	1	Italian Restaurant	Spa	Salon / Barbershop	Gym	C
35	Manhattan	Turtle Bay	40.752042	-73.967708	1	Coffee Shop	Sushi Restaurant	Seafood Restaurant	Park	I
36	Manhattan	Tudor City	40.746917	-73.971219	1	Coffee Shop	Grocery Store	Japanese Restaurant	Sushi Restaurant	B
38	Manhattan	Flatiron	40.739673	-73.990947	1	Gym / Fitness Center	Gym	Cycle Studio	American Restaurant	
39	Manhattan	Hudson Yards	40.756658	-74.000111	1	Hotel	Gym / Fitness Center	Italian Restaurant	Theater	Dar
41	Seattle	Beacon Hill	47.540000	-122.310000	1	Airport Terminal	Café	Airport	Fried Chicken Joint	
44	Seattle	Capitol Hill	47.630000	-122.320000	1	Coffee Shop	Cocktail Bar	Thai Restaurant	Café	
46	Seattle	Central Business District, First Hill	47.610000	-122.330000	1	Coffee Shop	Hotel	New American Restaurant	Sandwich Place	
48	Seattle	Downtown	47.604000	-122.329000	1	Hotel	Coffee Shop	Café	Italian Restaurant	C
52	Seattle	Madison Park, Capitol Hill	47.630000	-122.290000	1	Café	Garden	Dessert Shop	French Restaurant	
55	Seattle	Northeast	47.660000	-122.297500	1	Pizza Place	Coffee Shop	Arts & Crafts Store	Thai Restaurant	
64	Seattle	University District, Laurelhurst	47.660000	-122.290000	1	Furniture / Home Store	Coffee Shop	Women's Store	Clothing Store	Ar

Cluster 3

Contents

	Borough/City	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Manhattan	Marble Hill	40.876551	-73.910660	3	Park	Pizza Place	Supermarket	Mexican Restaurant
2	Manhattan	Washington Heights	40.851903	-73.936900	3	Pizza Place	Latin American Restaurant	Bakery	Mexican Restaurant
3	Manhattan	Inwood	40.867684	-73.921210	3	Latin American Restaurant	Pizza Place	Deli / Bodega	Mexican Restaurant
4	Manhattan	Hamilton Heights	40.823604	-73.949688	3	Coffee Shop	Bar	Café	Mexican Restaurant
7	Manhattan	East Harlem	40.792249	-73.944182	3	Café	Bakery	Pizza Place	Mexican Restaurant
11	Manhattan	Roosevelt Island	40.762160	-73.949168	3	Park	Coffee Shop	Sushi Restaurant	Pizza Place
19	Manhattan	East Village	40.727847	-73.982226	3	Wine Bar	Bar	Coffee Shop	Ice Cream Shop
25	Manhattan	Manhattan Valley	40.797307	-73.964286	3	Coffee Shop	Park	Indian Restaurant	Ice Cream Shop
26	Manhattan	Morningside Heights	40.808000	-73.963896	3	Italian Restaurant	Coffee Shop	Park	American Restaurant
28	Manhattan	Battery Park City	40.711932	-74.016869	3	Park	Coffee Shop	Wine Shop	Gym
37	Manhattan	Stuyvesant Town	40.731000	-73.974052	3	Pizza Place	Bar	Cocktail Bar	Coffee Shop
40	Seattle	Ballard	47.675000	-122.380000	3	Coffee Shop	Ice Cream Shop	Mexican Restaurant	Bakery
42	Seattle	Belltown	47.620000	-122.350000	3	Coffee Shop	Exhibit	Pizza Place	Bar
45	Seattle	Central	47.610000	-122.300000	3	Coffee Shop	Ethiopian Restaurant	Playground	Thai Restaurant
47	Seattle	Delridge	47.545000	-122.360000	3	Convenience Store	Trail	Food Truck	Gas Station
50	Seattle	Greennwood, Freemont, Greenlake	47.670000	-122.340000	3	Zoo Exhibit	Coffee Shop	Park	Grocery Store
54	Seattle	N. Beacon Hill, Mt. Baker	47.580000	-122.300000	3	Coffee Shop	Vietnamese Restaurant	Pizza Place	Convenience Store
56	Seattle	Northgate	47.710000	-122.300000	3	Park	Marijuana Dispensary	Lake	Tennis Court
57	Seattle	Northwest	47.730000	-122.350000	3	Chinese Restaurant	Coffee Shop	Vietnamese Restaurant	Grocery Store
58	Seattle	Pioneer Square, International District	47.600000	-122.320000	3	Vietnamese Restaurant	Chinese Restaurant	Coffee Shop	Café
59	Seattle	Queen Anne/Magnolia	47.630000	-122.370000	3	Park	Coffee Shop	Cruise	Café
60	Seattle	Rainier Valley, Columbia City	47.540000	-122.270000	3	Mexican Restaurant	Park	Vietnamese Restaurant	Bank
61	Seattle	Redmond	47.680000	-122.120000	3	Coffee Shop	Thai Restaurant	Bakery	Pizza Place
62	Seattle	South Lake Union, East Queen Anne	47.630000	-122.340000	3	Coffee Shop	Hotel	Food Truck	Museum

There are as many as 40 neighborhoods in Manhattan, we are not going to explore each of them, what we wil do is:

On the internet, find out the top 3 popular neighborhoods in Manhattan.

Then check the similar neiborhoods in Seattle for those 3 neighborhoods

After doing some googling, we came up with our top 3 Manhattan Neighborhoods: East Village, West Village and Lower East Side, I webpages below:

- Top 5 Neighborhoods for Wall Street Workers (https://www.careermetis.com/neighborhoods-for-wall-street-workers/)
- Work in the Financial District? 5 great neighborhoods with a quick commute (https://www.brickunderground.com/blog/2015/10/30_minute_commute_financial_district)
- The Top 10 New York City Neighborhoods to Live In, According to the Locals (https://www.forbes.com/sites/trulia/2016/10/04/t york-city-neighborhoods-to-live-in-according-to-the-locals/#5a2079881494)

Let's check out which clusters are the three neighborhoods in.

```
East Village Neighborhood is in cluster 3
West Village Neighborhood is in cluster 1
Lower East Side Neighborhood is in cluster 1
```

OK, so East Village Neighborhood is in cluster 3, Lower East Side Neighborhood and West village Neighborhood are in cluster 1.

You can also double confirm this on the map. East Village Neighborhood is represented by green circles(cluster 2), while Lower Eas Neighborhood and West village Neighborhood are in purple(cluster 1).

Let's check which Seattle Neighborhoods are similar to East Village neighborhood, that is to say, in cluster 3; and which Seattle Nei similar to Lower East Side and West Village neighborhood, which are in cluster 1.

```
If you are moving to Seattle from East Village in Manhattan,
The similar neighborhoods in Seattle that you could choose are:

Ballard
Belltown
Central
Delridge
Greennwood, Freemont, Greenlake
N. Beacon Hill, Mt. Baker
Northgate
Northwest
Pioneer Square, International District
Queen Anne/Magnolia
Rainier Valley, Columbia City
Redmond
South Lake Union, East Queen Anne


If you are moving to Seattle from West Village or Lower East Side in Manhattan,
The similar neighborhoods in Seattle that you could choose are:

Beacon Hill
Capitol Hill
Central Business District, First Hill
Downtown
Madison Park, Capitol Hill
Northeast
University District,  Laurelhurst
```

Results and Discussion

From the exploration above, we find that among the 5 clusters, Manhattan neighborhoods are all in 2 clusters, which is cluster 1 an

Since there are too many neighborhoods in Manhattan, we did not look deep into each one, instead we focused on the top three ne Manhattan, which are:

- East Village Neighborhood
- Lower East Side Neighborhood
- West Village Neighborhood.

And we did find the similar neighborhood in Seattle for each of those 3 neighborhoods in Manhattan.

If you are interested in exploring some other certain neighborhood in Manhattan, you can just feel free to go back to the Manhattan and click on the circle which represent your interested neighborhood to find out the cluster and then go to the Seattle map to figure neighborhood in the same cluster(in the same color).

Or if you are a big fan of charts or lists, you can also check the cluster 1 and cluster 3 neighborhoods list we generated earlier to pr

And please be kindly noted that, although the result in our analysis could throw some insights onto the issue concerned, since our online from multiple sources, we should always be cristical about the confidence of the result and only take as a reference.

Conclusion

If this project, we are trying to find out the similar neighborhoods between Manhattan and Seattle, intend to provide an insight to th to move from Manhattan to Seattle to start a new career.

And luckily, there do exist some neighborhoods in Seattle similar to those in Manhattan as shown above.

But it should also be well acknowledged that the result of this analysis is basically according to the categories of venues nearby the neighborhood, there are some other factor you may also want to take into consideration before you make up your mind, things like commuting, crime rate, and so on.

Hope you had a good time through this journey of exloring similar neighborhoods.