Use geo-data to cluster neighborhoods and generate recommendation for visitors to the Greater Seattle area

1. Introduction

1.1. Background

Seattle is a seaport city on the West Coast of the United States. It is the seat of King County, Washington. Seattle is the largest city in both the state of Washington and the Pacific Northwest region of North America. According to U.S. Census data released in 2019, the Seattle metropolitan area's population stands at 3.98 million, and ranks as the 15th-largest in the United States.

1.2. Interest

In this study, by utilizing the Foursquare API and the geo-locator tools, information of more than 5000 venues around 66 neighborhoods of Seattle was retrieved. Through unsupervised learning, it was attempted to categorize the neighborhoods into different clusters, based on the composition and percentage of nearby venues. In addition, from the resulting clusters, the top popular venues were listed to perform a supervised classification, and the model was used to generate travel route recommendation for a new visitor to the Seattle area based on his/her travel preference.

2. Data Collection and Cleaning

2.1. Data Mining

A PDF file containing the postal information of 122 neighborhoods was downloaded. Then Tabula was used to extract the tables incorporated in the file. By combining the tables and deleting the unrelated information, the final table of information of 84 neighborhoods was created. The neighborhoods mainly came from the major cities in the Seattle area, including Seattle, Bellevue, Kirkland, Redmond, Kent, Renton, SeaTac, Medina, Mercer Island, Federal Way and Auburn.

With the postal codes and sub-region names obtained, the map locations of the neighborhoods were retrieved through geo-locator, which returned the latitudes and longitudes. Neighborhoods whose coordinates were not available were discarded. Then, using the explore function of Foursquare API, venues within a radius of 2000m of each neighborhood were queried, resulting a total number of 5283 venues.

2.2. Data Preprocessing

Because of the presence of most common venue types, such as coffeeshops, bus stop, ATMs and so on, the unique characteristics of each neighborhood were hard to be observed. Therefore, the categories with counts greater than 100 were removed from the dataset, including “Coffee Shop”, “Pizza Place”, “Sandwich Place”, etc. Similarly, due to potential difference in category naming, 53 venue categories with only 1 count were also removed. In the final data frame, there were 3946 venues in total, belonging to 273 categories.

3. Modeling

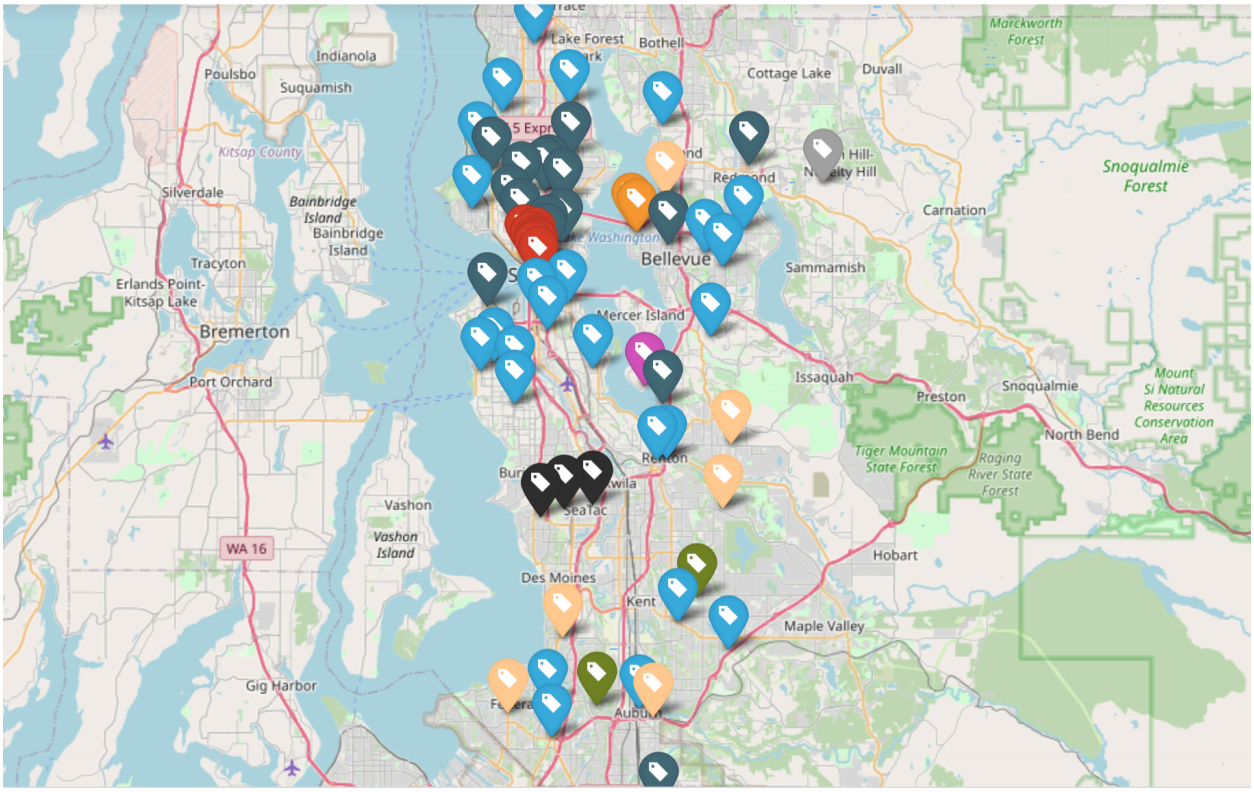
3.1. K-means method for neighborhood clustering

With the processed data frame containing venues of 273 categories, additional 273 dummy columns were created, each representing one category, in order to convert the textual variables into numerical. By computing the percentage of each category' appearance among all for every neighborhood, the data frame was converted to the one having dimension of 66 X 274: 66 rows for the neighborhoods, 1 column identifying the neighborhood postal code, and 273 columns, each identifying the percentage of one venue category.

Before implementing the K-means algorithm, an elbow graph was plotted to determine the best choice of clusters number for this study. From the resulting graph, K = 9 is the turning point, or "elbow", which indicates that 9 clusters would be sufficient to stratify the neighborhoods.

3.2. Displaying the clusters on a folium map

Adopting the choice of 9 for number of clusters, the results of this unsupervised learning were displayed on a folium map. As shown by the map, downtown Seattle forms a major cluster, and the rest parts of Seattle city form another major cluster. Denoted by the light blue labels, the rural area of most cities, including Seattle, Bellevue, Kent, Auburn and Federal Way falls into the third major cluster. In addition, each of the rest clusters represents an area with a unique theme. For example, SeaTac is featured by car rental companies and airport lounges, as the Seattle-Tacoma International Airport is located there, while Medina is featured by outdoor activities such as trails and golf courses.



3.3. KNN for traveller recommendation

After the neighborhoods had been divided into clusters, a list of top 7 popular venues was created by selecting 7 categories that had the largest percentages among all categories for each neighborhood. Based on this popularity list, the popular themes of a certain neighborhood could be derived. For example, a person who would like to experience the Seattle food, especially the asian food, should probably consider visiting the neighborhood located at 98104, as among the top 7 popular venues, 6 of them are restaurant, and 4 of those 6 are asian restaurants.

	index	Zip	City_Name	Latitude	Longitude	Cluster	Number 1 popular local venue type	Number 2 popular local venue type	Number 3 popular local venue type	Number 4 popular local venue type	Number 5 popular local venue type	Number 6 popular local venue type	Number 7 popular local venue type
0	0	98101	Seattle	47.610763	-122.336182	1	Breakfast Spot	Sushi Restaurant	Yoga Studio	Bar	Market	Brewery	Seafood Restaurant
3	3	98104	Seattle	47.600708	-122.331334	1	Vietnamese Restaurant	Seafood Restaurant	Dumpling Restaurant	Sushi Restaurant	Italian Restaurant	Deli / Bodega	Noodle House
15	15	98121	Seattle	47.615494	-122.344680	1	Breakfast Spot	Sushi Restaurant	Yoga Studio	Marijuana Dispensary	Seafood Restaurant	Sculpture Garden	Scenic Lookout
22	22	98138	Seattle	47.606038	-122.331993	1	Seafood Restaurant	Breakfast Spot	Italian Restaurant	Dumpling Restaurant	Deli / Bodega	Sushi Restaurant	New American Restaurant
25	25	98154	Seattle	47.606189	-122.333591	1	Breakfast Spot	Seafood Restaurant	Italian Restaurant	Dumpling Restaurant	New American Restaurant	Brewery	Fish Market

Following this intuition, a data frame was created, featuring themes such as Asian Food, Western Food, Go Seattle, Casual, Seaside, Leisure, etc. Then, for each neighborhood, one score is computed for each theme, based on if and at which position a category of this theme is present in the popularity list. As shown by the table below, most neighborhoods of Cluster 1 (downtown Seattle) contain a mixed amount of Asian and Western restaurants, a rich environment for Seattle style venues, a fair amount of lifestyle venues (bookstores, art craft shops, etc.) and lots of beach activities. Apparently, this area would be a good choice for traveller to Seattle or for someone who has a mood to enjoy the seaside charm.

Cluster	Asian_food	Western_restaurant	Middle_east_restaurant	Go_Seattle	Local_life	Casual	Leisure	Lifestyle	Arrival_departure	Seaside	formal_food
1	8.888889	0.0	0.000000	3.333333	1.538462	0.0	4.615385	10.000000	0.0	3.333333	10.000000
1	10.000000	2.0	0.000000	6.666667	0.000000	1.0	0.000000	0.000000	0.0	6.666667	7.777778
1	4.444444	0.0	0.000000	5.000000	0.000000	2.0	0.000000	5.000000	0.0	1.666667	5.000000
1	3.333333	6.0	0.000000	6.666667	0.000000	1.0	0.000000	3.333333	0.0	6.666667	7.777778
1	2.222222	6.0	0.000000	8.333333	0.000000	0.0	0.769231	3.333333	0.0	8.333333	7.222222
1	3.333333	6.0	0.000000	6.666667	0.000000	0.0	0.769231	3.333333	0.0	6.666667	7.777778
1	3.333333	6.0	0.000000	6.666667	0.000000	1.0	0.000000	3.333333	0.0	6.666667	7.777778
1	2.222222	6.0	0.000000	6.666667	0.000000	2.0	0.000000	3.333333	0.0	6.666667	7.222222
1	4.444444	2.0	0.000000	5.000000	0.000000	1.0	0.000000	5.000000	0.0	5.000000	6.111111
1	1.111111	0.0	3.333333	5.000000	0.000000	0.0	4.615385	3.333333	0.0	5.000000	1.666667

After the construction of the theme score table, a KNN classifier was used to fit those 66 neighborhoods and to learn the theme compositions.

4. Prediction

With the KNN model, recommendations for travel destinations could be generated based on a user's input. For example, if the person wants to do a casual tour to experience the local life of Seattle, the rural area (cluster 2) would be a good choice. On the other hand, if he/she plans to leave in 3 hours, traveling to SeaTac (cluster 4) is recommended, as most of the car rental and airport services are located there.

```
Asian_food_score = 5
Western_food_score = 0
middle_east_food_score = 3
Go_Seattle_score = 1
local_life_score = 3
casual_score = 3
leisure_score = 2
lifestyle_score = 5
airport_score = 0
seaside_score = 0
formal_food_score = 2
x_pred = pd.DataFrame(columns = store_categories[1:])
x_pred.loc[0] = [Asian_food_score,Western_food_score,middle_east_food_score,Go_Seattle_score,local_life_score,casual_sc
x_pred
```

	Asian_food	Western_restaurant	Middle_east_restaurant	Go_Seattle	Local_life	Casual	Leisure	Lifestyle	Arrival_departure	Seaside	formal_food
0	5	0	3	1	3	3	2	5	0	0	2

```
y_pred = knn.predict(x_pred)
```

```
y_pred[0]
```

```
2
```