Opening new Yoga Studio in Seattle,WA

IBM Applied Data Science Capstone Project

Jesmeen Fatema

March 2020

Introduction:

Seattle is situated in Pacific Northwest, surrounded by water, mountains and evergreen forests in Washington State. Seattle has population close to 3.4 Million and is home to major Tech Corporations like Microsoft, Amazon and many others. People in the city are health conscious and willing to spend money to take care of their wellbeing. I visited Seattle multiple times and fell in love with the city and hope to move there one day. I decided to select Seattle for my project to open a new Yoga Studio.

Business Problem

The objective of this project is to analyze and identify locations in Seattle that have good potentials to open a new Yoga Studio and it is important to choose a location where there is less or no Yoga Studio. Using Data Science Methodology and Machine Learning techniques like clustering we'll be able to determine that.

Target Audience

Anybody looking to open a new Yoga Studio is a target audience. Whether single location for individual entrepreneur or multiple locations for big business, it is a good investment to fulfill health and wellbeing needs of modern, health conscious population. It can also help people to choose what options they have and where the Yoga Studios are located.

Data:

Data required

- List of Neighborhoods in Seattle, WA. This defines the scope of the project which is confined to city of Seattle, WA.
- Latitude and the Longitude of the Neighborhoods. This is required to plot the map and get the venues.
- Venue data, specifically related to Yoga Studio. This data will be used to perform Clustering of the neighborhoods.

Source and method to extract Data

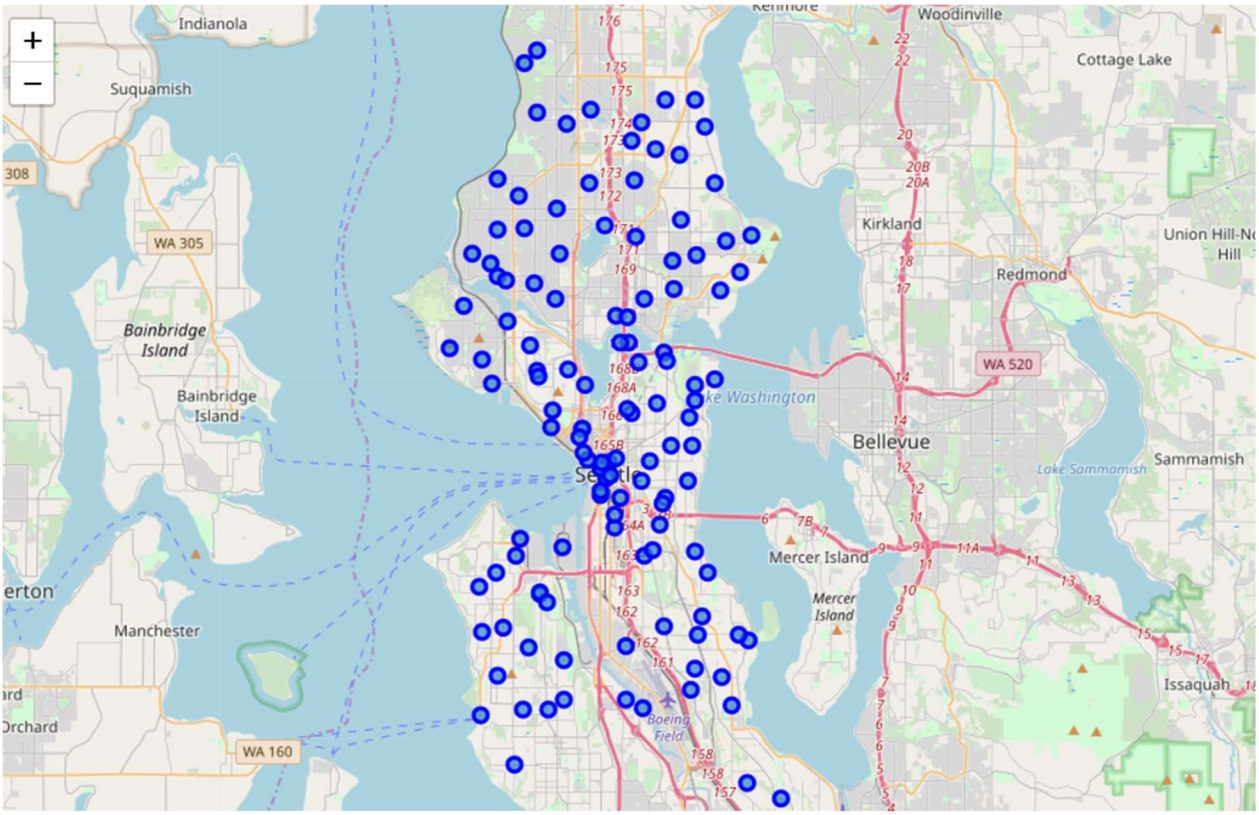
From Wikipedia (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Seattle) we extract and scrape Seattle Neighborhood data using various Python commands. Next, we get the geographical coordinates, data for the Latitude and the Longitude of the Neighborhoods by using Geocoder library. With list of Neighborhoods and their Latitude and Longitude we use Foursquare API to get venue information and we select the Yoga Studio category for further analysis. We are using K-mean Clustering (Machine Learning Technique) to determine suitable locations for our new business as well as Folium library to locate them in the Map. The processing of data help us identify which neighborhoods has less concentration of Yoga Studio, therefore indicating suitable location to open a new one.

Methodology:

The first data we need is the list of neighborhoods and it is available at Wikipedia (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Seattle). We extract and clean the data by web scraping method and using various python commands to get neighborhood data. Next, we need the geographical coordinates, data for the Latitude and the Longitude of the Neighborhoods to utilize Four square API to get venues and detail analysis. We get data for Latitude and the Longitude using Geocoder library. After gathering data for List of neighborhoods and the Latitude and the Longitude we create a table using pandas data frame.

	Neighborhood	Latitude	Longitude
0	North Seattle	47.643727	-122.302939
1	Broadview	47.722380	-122.364980
2	Bitter Lake	47.718680	-122.350300
3	North Beach / Blue Ridge	47.700460	-122.384170
4	Crown Hill	47.695200	-122.374100

We used python folium library to visualize geographic details of Seattle and its neighborhoods and created a map of Seattle with neighborhoods superimposed on top.

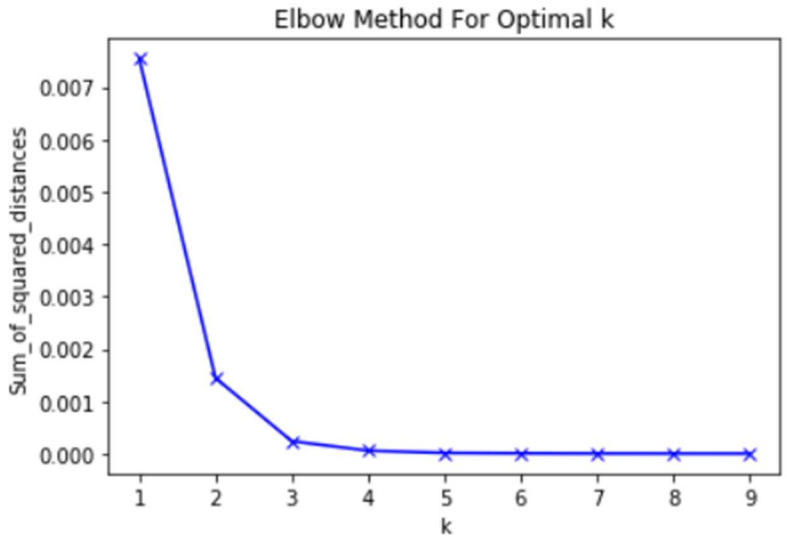


Next, we use Four square API to get the top 100 venues within a radius of 1500 meters. Making a call to Foursquare API we receive venue name, venue category, venue Latitude and Longitude.
9,930 venues were returned by Foursquare. Here is a merged table of neighborhoods and venues.

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	North Seattle	47.643727	-122.302939	Washington Park Arboretum	47.639740	-122.294721	Botanical Garden
1	North Seattle	47.643727	-122.302939	Seattle Public Library - Montlake	47.640520	-122.302413	Library
2	North Seattle	47.643727	-122.302939	Cafe Lago	47.639698	-122.302256	Italian Restaurant
3	North Seattle	47.643727	-122.302939	Montlake Cut	47.647094	-122.304686	Canal
4	North Seattle	47.643727	-122.302939	Arboretum Waterfront Trail	47.642934	-122.291802	Trail

We check how many venues were returned and how many unique categories can be extracted from all the returned venues. We analyze each neighborhood by grouping the rows by each neighborhood and taking the mean frequency of occurrence in each venue category. This also helps preparing the data for use in clustering. We filter the data for yoga studio as venue category for the neighborhood.

We used unsupervised learning **K-means algorithm** to cluster the neighborhoods. K-means clustering algorithm identifies k number of centroids and then allocates every data point to the nearest cluster, while keeping the centroid as small as possible. It is one of the simplest and popular unsupervised algorithms

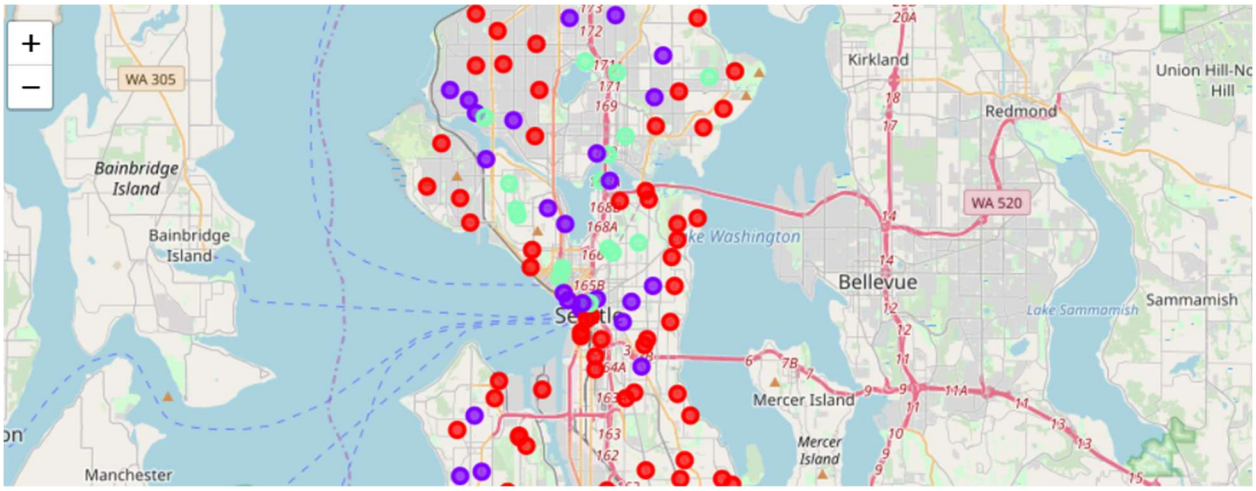


and suited for our project. We determine the optimal K for K-means algorithm by elbow method and it indicated 3 degree for optimal K and we cluster the neighborhoods into 3 clusters for yoga studio.

Results and Discussion:

The result from K-mean clustering shows that we can catagorize the neighborhood into 3 clusters based on the occurance for yoga studio.

- Cluster 0: Neighborhoods with low number to no yoga studio (red circles in map)
- Cluster 1: Neighborhoods with moderate number of yoga studio (purple circles in map)
- Cluster 2: Neighborhoods with high concentration of yoga studio (mint circles in map)



From the result it shows that most of the yoga studios are located in cluster2 (total 17 neighborhoods) which is mostly located near Seattle downtown and north of the city. There are moderate number of yoga studios in cluster1 (total 29 neighborhoods) which is mostly centrally located around the city. Very few yoga studios are located in cluster 0 (total 81 neighborhoods) and mostly located on east and west side of the city. This represents a great opportunity to open new Yoga Studio in neighborhoods located in cluster 0 where there is very little to no competition.

Conclusion:

Purpose of this project is to identify Seattle neighborhoods with low number of yoga studio in order to aid stakeholders in narrowing down the search for optimal location for a new yoga studio. Clustering of those locations is performed to identify major areas of interest to be used as starting points for final exploration by stakeholders. Final decision on optimal yoga studio location will be made by stakeholders based on specific characteristics of neighborhoods and locations, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.

References:

[1] Seattle-Wikipedia: https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Seattle

[2] Foursquare API

[3] Google Map