Coursera Capstone Project

IBM Applied Data Science Capstone

Opening a new coffee place in Seattle, Washington

By:

Kiranpreet Chawla

Date: July 2019

Introduction

Seattle is regarded as a world center for coffee roasting and coffee supply chain management. Related to this, many of the city's inhabitants are coffee enthusiasts; the city is known for its prominent coffee culture and numerous coffeehouses. People in Seattle consume more coffee than in any other American city; one study stated that there are 35 coffee shops per 100,000 residents and that Seattle people spend an average of $36 a month on coffee. It is nearly impossible to walk past a single block in a commercial area in Seattle without walking past at least one coffee shop. Coffee drinkers can get coffee at a local sidewalk stand, parking lot, tiny coffee houses, big coffee houses, drive-through, and even delivery. Seattle is home to first Starbucks coffee shop along with several coffee roasters like Starbucks roaster, Tully's coffee, Espresso Vivace and others.

The aim is to open a new coffee shop in Seattle and for that, collect and analyze data on coffee places in various neighborhoods in Seattle. With such a competitive market, it is important to conduct a research, collect facts and then, choose a neighborhood to open a new coffee shop.

Business Problem

The objective is to analyze and select the best neighborhood to open a new coffee shop in Seattle. Using data science methodologies and machine learning techniques like clustering and segmentation, find out facts about coffee places in different neighborhoods. The question is- If you've to open a new coffee shop in Seattle, where would you recommend opening it?

Data

To solve the problem, we'll need following data-

- List of neighborhoods in Seattle
- Latitude and Longitude of those neighborhoods. This will help us make a map and visualize the data
- Coffee places data in different neighborhoods. This will help us cluster the data

Data Source and utilizing the data

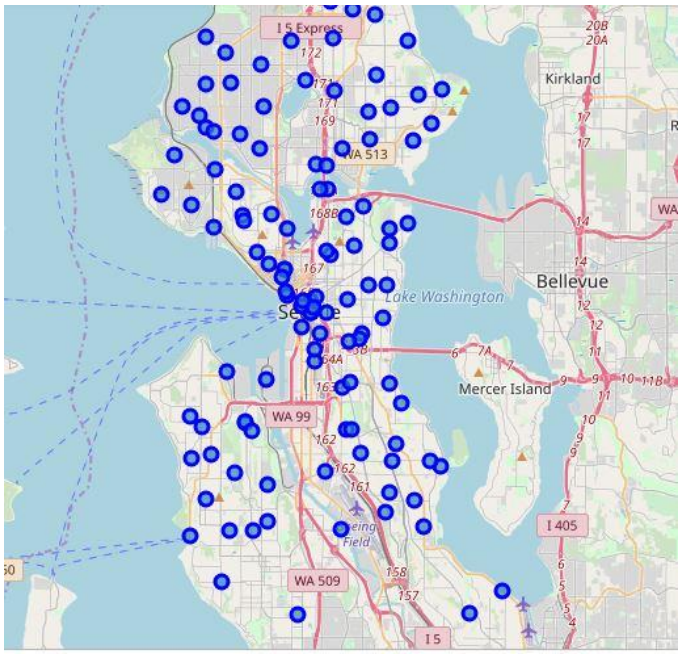
The list of neighborhoods in Seattle can be found on this url-
https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Seattle

We'll further use BeautifulSoap library to parse the data and get neighborhood names. Then, create a python data frame for further processing. We'll use Python Geocoder library to get the geographical coordinates of the location and then, use FourSquare API to find out coffee places venues. We'll also use Folium library for maps and visualization along with machine learning techniques to analyze the data.

Methodology

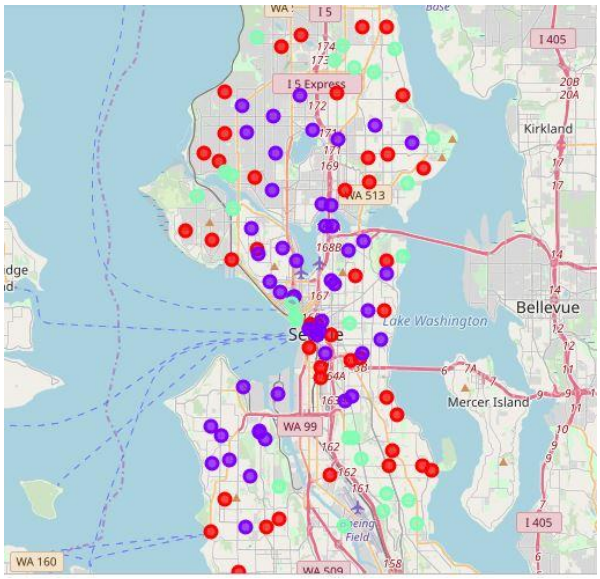
To start with, we'll get the neighborhood data from Wikipedia page- https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Seattle.

Next, we'll scrape only the required data (here, it is neighborhood's name) from this page using BeautifulSoap package. Next, we'll create a data frame from this data and get the geographical coordinates of the neighborhoods. We'll visualize the data using Folium library and create a map to get a picture of data. We need to create a Foursquare account to get the venues data for a given location. Then, using Foursquare API, we'll get the data for venues, it in a data frame and run k-means clustering. K-means clustering algorithm identifies k number of centroids, and allocates every data point to the nearest cluster, while keeping the centroids as small as possible. Create a map for better visualization and in the end, analyze the data.



Results

We cluster the neighborhoods into three clusters based on the frequency of occurrence of coffee shops. The results can be visualized as shown in the map below- Cluster 0 is in red color, cluster 1 in purple color, cluster 2 in mint green color.



Discussion

As observed in the results, there are more coffee places concentration in the neighborhoods mentioned in Cluster 1 and cluster 0, and the least in Cluster 2. Hence, the neighborhoods in Cluster 2 have good potential of opening a new profitable coffee place, given less competition.

Conclusion

In this project, we identified the business problem, find out the data that we'll required, extract the data, perform machine learning methodologies, form clusters and then analyze the results. The findings will help the stakeholders to identify the best neighborhood to open a new coffee place.