Finding the best place to live in Seattle

By: Ahmed Sherief

November， 2019

Introduction

The Seattle area is home to some of the world's most famous companies, Microsoft, Starbucks and Amazon are just three. Seattle is located on the west coast of the United States on Puget Sound, which is connected to the Pacific Ocean. It has one of the largest ports in the country and is the closest US port to Asia It also has good connections to Alaska and Canada as well as the rest of the US. Seattle has a busy and growing international airport. Seattle's on the west side of Washington State with good connections to the major agricultural areas in the eastern part of the state. Because of its location, Seattle has a long history of international trade. In fact, 40% of jobs in Seattle are related to international trade.

It is truly a global trade hub. Seattle has a highly skilled and educated workforce. There are over 175,000 people working in the tech industry in the Seattle area and it has the biggest groups of aerospace workers in the U.S. It's also an active promoter of start-ups to develop new products and technologies. This has led to the Seattle area being one of the main recipients of investment from venture capitalists outside of California. **The targeted audience of this project is the people who seeks to find the best place to live in Seattle**

Business problem

The objective of this project is to analyze data and find the best place of living in the city to live near to all services such as hospitals, restraints, bus station, metro café, parking and gardens……etc. using data science methodology and machine learning technique like clustering. We have to find the best cluster with the largest number of services.

Data

1. List of neighborhoods in Seattle. We have to scrape it from (Category:Neighborhoods in Seattle - Wikipedia)
2. Latitude and longitude coordinates of this neighborhoods. We will get it using Python geocoder package.
3. Venue data, and we will use it to perform clustering on the neighborhoods. And we will use foursquare API to get the venue.

Methodology

Firstly, we need to get the list of neighborhoods in the city of Seattle. Fortunately, the list is available in the Wikipedia page (Category:Neighborhoods in Seattle - Wikipedia). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data.

It is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Kuala Lumpur.
Next, we will use Foursquare API to get the top 100 venues that are within a radius of 200 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop.

Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the data to get the largest number of venues.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for venues. The results will allow us to identify which neighborhoods have higher concentration of venues while which neighborhoods have larger number of venues.

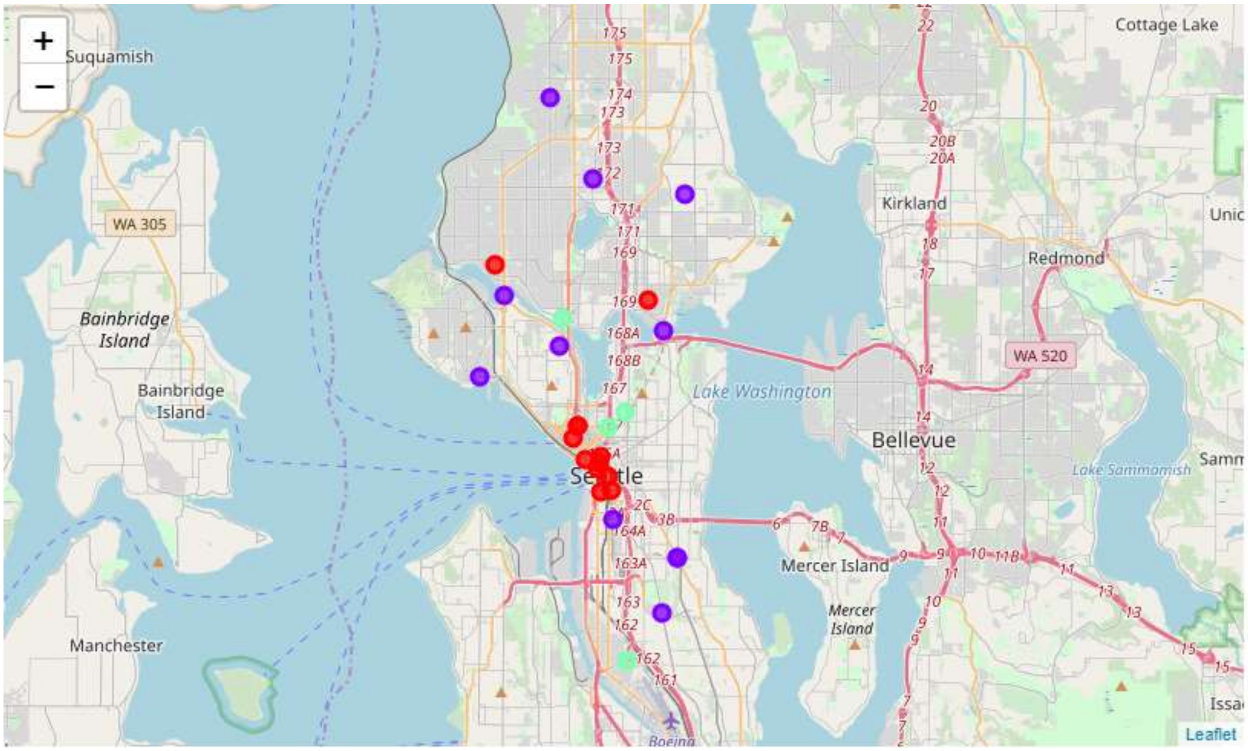
Based on the number of venues it well answers the question as to which neighborhoods are most suitable for living.

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the number of venues:

- Cluster 0: Neighborhoods with large number of Venues.
- Cluster 1: Neighborhoods with the lowest number of existence of venues.
- Cluster 2: Neighborhoods with moderate concentration of venues.

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color.



Discussion

From the Results section, most of the neighborhoods with the large number of venues are in the central area of Seattle, with the highest number in cluster o and moderate number in cluster 2. On the other hand, cluster 1 has very low number of venues in the neighborhoods.

This represent the best place of living in the city to live near to all services such as hospitals, restraints, bus station, metro café, parking and gardens……etc.

The best cluster is (cluster 0) based on the high number of venues. then (cluster 2), and be careful living in (cluster 1) will be bad for you and your family because there well be few number of services compared to the other two clusters.

Conclusion

As a result, people are turning to big cities to start a business or work. For this reason, people can achieve better outcomes through their access to the platforms where such information is provided.

Not only for investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.

Ahmed Sherief