Coursera Capstone Project- The Battle of Neighborhoods (Week 2)

Introduction

Greater Seattle is home to Amazon and Microsoft, two of the biggest technology companies in the world. There are more than a dozen companies that have headquarters in Seattle including Aerospace manufacturer Boeing. This makes Seattle real estate one of the hottest in the nation. According to Redfin.com, Seattle is rated 91 out of 100 Most Competitive where most homes get multiple offers often with waved contingencies. Homes sell for about 1% to 4% above list price and go pending in about 3 to 7 days. Some homes sell for as high as 22% above list price in that same site.

Audience

This makes it a perfect place for property developers and stakeholders to build and invest in this hot real estate market!

Business Problem

In this project we are going to analyze where is the best location to buy or build a new home. To do so we are going to use Seattle Neighborhoods and location data.

Data (Types of data needed)

-First, we need to get the list of neighborhoods data in order to explore city

-We then get latitude and longitude of the neighborhoods using python package geocoder.

-We also need to utilize the Foursquare Venues data and machine learning tool k-means clustering to group the neighborhoods into clusters.

Where we get the data

We get our data from https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Seattle which has all the 127 neighborhoods in Seattle.

Techniques used

Since our data is not readily available, we will have to use web scraping techniques such as beautiful soup or simply pandas to extract our table from the web page.

We will use geocoders package in python to get the geographical coordinates latitude and longitude of the neighborhoods.

We then use python request package to request venues data from the Foursquare location data and k-means clustering to cluster neighborhoods. The important step of visualizing the data in a map we use the python Folium package.

There are several steps used and that would be in the next section!

Methodology

First of all, we install all the necessary packages needed to work with the data. The most notable ones are like geocoders to get latitude and longitude of our neighborhoods, k-means from sklearn.clusters to cluster our neighborhood to groups and the python package folium to visualize our data.

We use web scraping technique like read html from python.io.html to get the list of neighborhood names and the district they fall in, then make it a clean a dataframe. Since this data does not have the geographical coordinates we use the geocoders with a while loop until we get all the latitude and longitudes of the neighborhoods and add(concat) it to the previous dataframe. By now our data had 127 rows and 4 columns. To reduce the data size and for simplicity we only work with Districts that contained the word Seattle and ended up with 46 rows and 4 columns!

To have a general view of what the map looked like we used Nominatim geolocator to get the coordinates and map Seattle Washington with the neighborhoods using folium. This starts our visualization process like the image below:
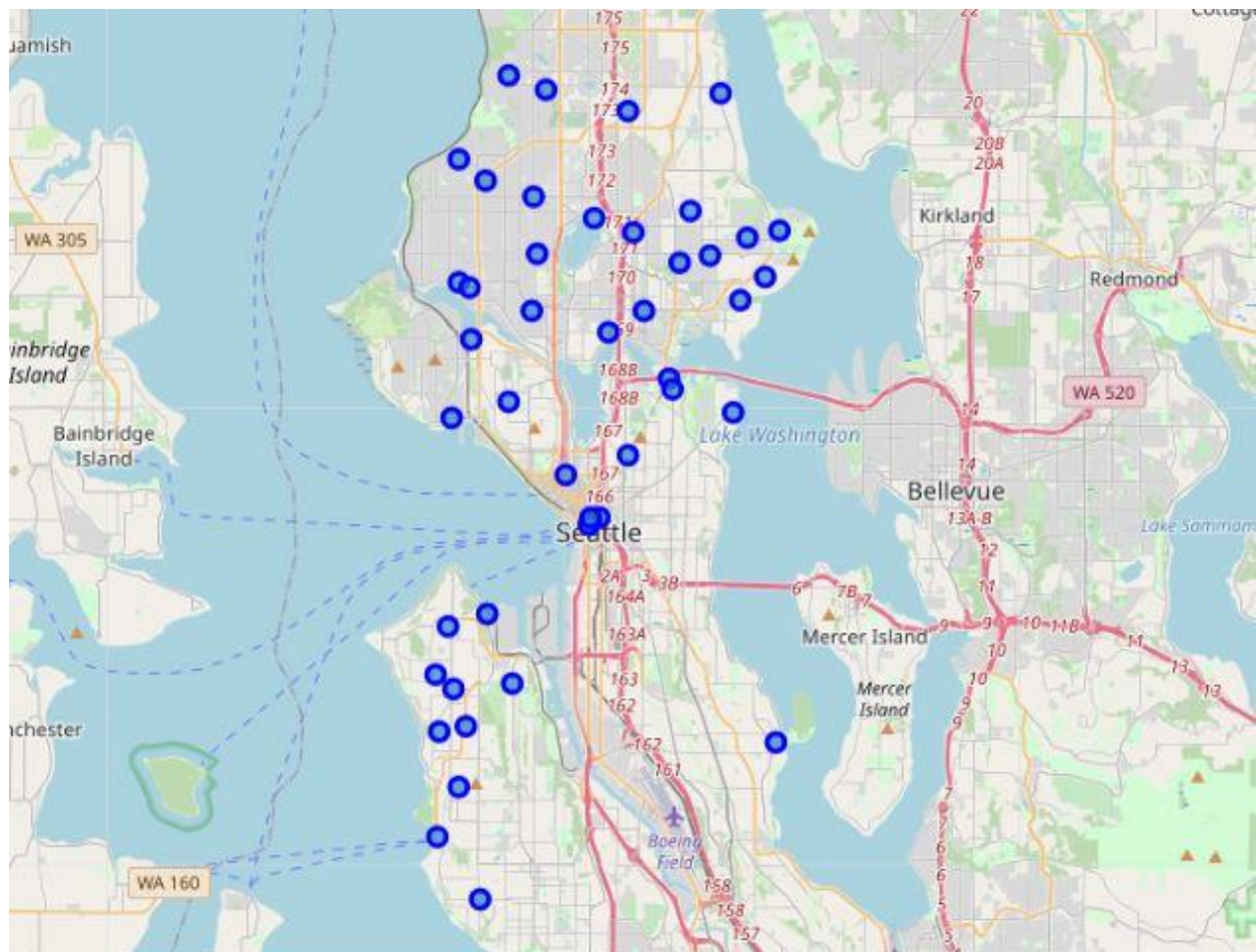
Seattle, Washington with its Neighborhoods.

From here we start exploring the first Neighborhood 'North Seattle' and use foursquare to get venues around (50 venues!) using its coordinates and Get request. Once we get the results we use a for loop to

iterate through all the neighborhoods and create the API url: request and make the Get requests which will return relevant nearby venues.

We analyze that data and use one hot encoding to fill the frequency of the venues in the dataframe. We use this new frequency data that was added to our dataframe to get top 5 and top 10 venues of each neighborhood. Based on this frequency data we use k-means clustering to cluster the Neighborhood into four clusters. We again map this new data using folium but this time with one color for each of the four clusters. The image is in the results section.

I just added an extra step to repeat the analysis process but this time only used the 'Parks' data for example if somebody likes parks instead of the top 10 venues to make the clusters, since building or buying a home is based on one of many factors.

Results

From k-means clustering we can cluster the neighborhoods into for clusters

(Base on Top 10 venues)

Cluster 0 which is red shows it has only two neighborhoods that are very similar

Cluster 1 is the purple cluster which has the most neighborhoods but

Cluster 2 only has one neighborhood is bright green sandwiched in the purple clusters

Cluster 3 this is the most popular of all the clusters and its color is faint olive green.
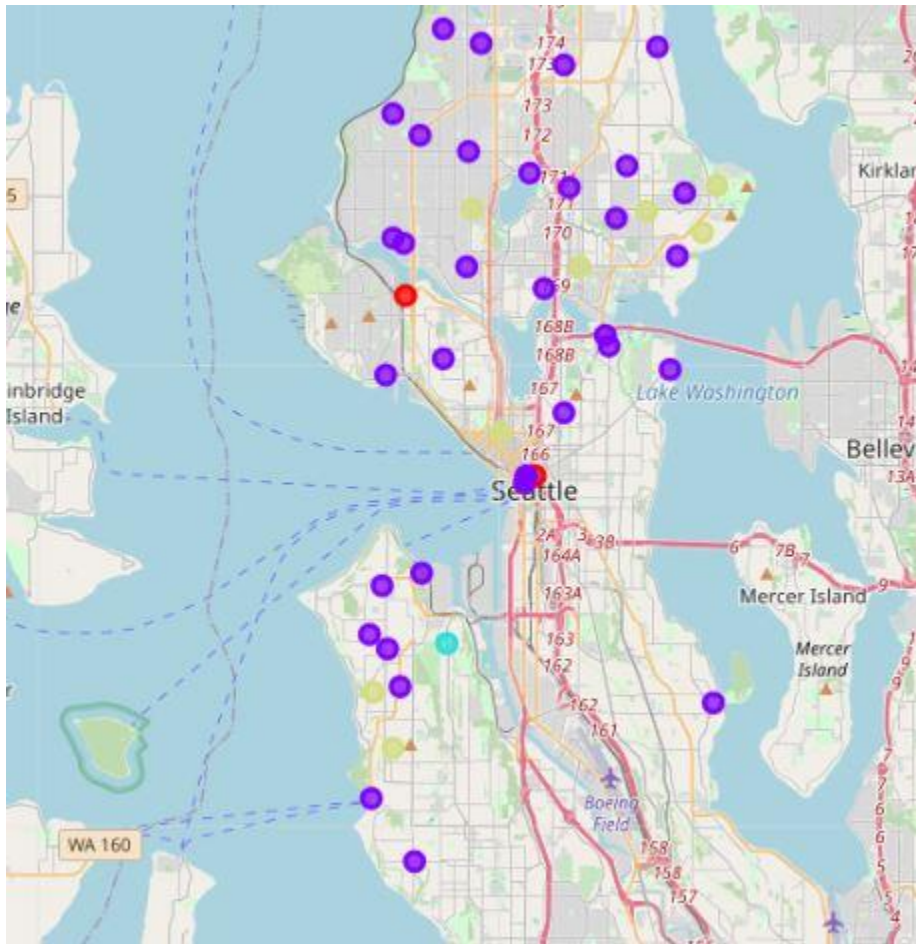
(Clusters Based on Parks)

Cluster 0 is red shows it has only two neighborhoods just like the previous one but in different location

Which had no parks

Cluster 1 is the purple cluster which has the most neighborhoods but very few parks

Cluster 2 only has one neighborhood is bright green like cluster 0 has no parks

Cluster 3 this is the cluster most 'Parks' of all the clusters and its color is faint olive green.

It is also well spread out like cluster one but has more parks!

Observation/Discussion and Conclusion

When we grouped neighborhoods into clusters using k-means clusters we selected the 10 most frequent venues and found that cluster 1 had the most neighborhoods and cluster 2 had the least(only 1)

neighborhood. When Building or Buying a house the are a lot of factors that go into buying a home. One of them is how family friendly the neighborhood is, which depends on the venues that are around it. In this case I went with parks! What we saw is that cluster 0 and 2 had no parks and cluster 1 had very few parks even though it had the most neighborhoods. This leaves us with cluster 3 which had the most parks. What is also interesting about neighborhoods in this cluster is that they are well spread out in the city. We have not taken into consideration other factors such a price which is import but that will require other dataset. But choosing a neighborhood in cluster 3 is the best option here because it is as well spread as cluster 1 and yet had the most parks!

More work and data is needed to come with a comprehensive results to make a better decision as to when and where exactly to build or buy a new home.

Thanks you!

References:

https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Seattle

Redfin.com-seattle housing market

Fourquare.com/docs