**CAPSTONE PROJECT: THE BATTLE OF THE NEIGHBORHOODS [SEATTLE]**

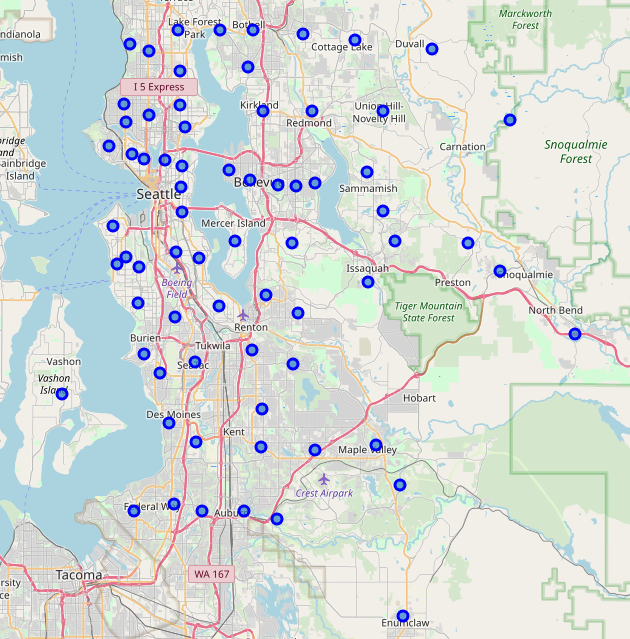
Eric Frezell 07/15/2019

**Defined Problem (Introduction):** A close friend has planned to open a Japanese/Asian fusion restaurant somewhere in Seattle. To be successful, they're searching for a when it comes to this type of food. What is served will be high quality and priced appropriately. The target customer will have a discerning taste when it comes to Japanese food and are likely searching for a differentiated offering. Since our close friend uses high quality ingredients and has priced the menu in accordance, the target customer will have no issue paying a premium for dishes on the menu. There are several neighborhoods to choose from that contain growing demand for new, unique and quality food options. Which neighborhood is the best fit for our friend's new restaurant?

**Data Description:** The data collection will involve using the Foursquare API to pull venue information for the Seattle area. The assumption made is that opening the restaurant in an area that does not already contain Asian food will draw in customers due to its novelty. API will deliver venue and location data that can be used to describe a neighborhoods unique food offering by exploring the types of restaurant operating in that area. The second assumption is that the restaurant will be more profitable in a moderately high to very high wealth area. There is a housing price data set for Seattle from the mid-2010s available from Kaggle that will be used to segment areas into degrees of wealth. This dataset includes home price, size of lots and other residential related details.

**Methodology:**  Working first with the housing data, I established a link on the transfer.sh website to host the Kaggle table. I can pull from the dataset anywhere if I have the transfer.sh link. After reading the data into a pandas data frame I searched the columns headers for notable information. Using the shape method I found the dataset houses over 20,000 home prices. There were several columns I was not interested since they did not tell me very much about the cost of living. I removed columns such as floors, waterfront, view, condition and grade. In the same cell I aggregated much of the pricing related data and stored it into a new data frame. By grouping by the zip code, I can tell what sort of neighborhood was likely wealthier than others. The mean price of a home in areas such as 98039 – the highest in the dataset – alluded to their high cost of living.

Next, I wanted to visualize the dataset to gather an understanding of the area of Seattle this dataset encompasses. Using folium, I was able to add markers for each zip to a map centered around the mean latitude and longitude of the housing dataset.



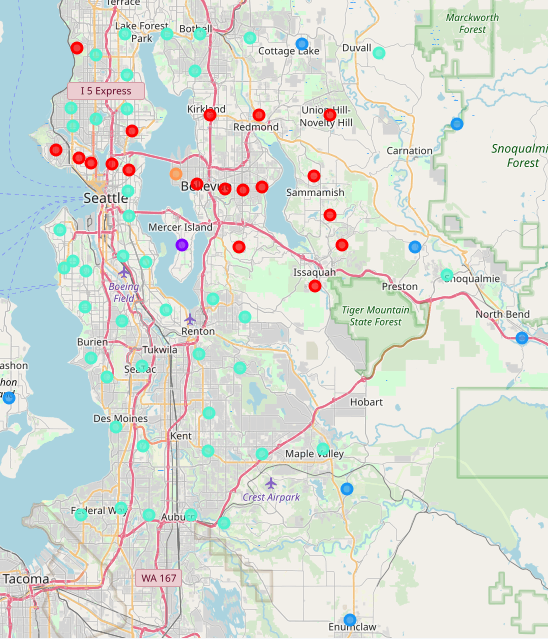
This neatly laid out all the zip codes within the Seattle area that the housing dataset encompassed. Examining that zip code I found with the highest mean home price. I zoomed on top of the marker and found the city to be Medina; one of the most expensive zip codes in the United States.

Now that I had my home price info firmly settle into its data frame, I began to code the inputs necessary to pull API data from Foursquare. The nearby venue data available through Foursquare’s API will compare well with the housing data I’ve already pulled. I initiated the query and formatted the information into a data frame. Unfortunately, only 121 venues were available nearby when select the mean latitude and longitude with a 20 miles radius. I drop 42 rows that are missing either a zip code or category data. That leaves me with only 79 nearby venues. The API appears to be limited in its ability to provide a full accounting of all nearby venue data. Despite this, I encode a new dataset with the venue category names as column headers and a frequency calculation of each category compared to every zip code. I am truly only interested in two categories, “Japanese Restaurant” and “Asian Restaurant” as they relate to my problem statement. Building a new data frame, I merge the housing data with venue data only keeping such relevant data and am ready to start using the kmeans algorithm to cluster after normalizing each feature and filling any missing feature data with ‘0’ to still utilize all of the housing data.

I had aligned on 6 different clusters after experimenting with several different outcomes. Kmeans was chosen to provide an efficient way to preform the segmentation while also providing some control around the number of groups that will a part of the outcome.

**Results:**

What was raised from the algorithms output was 6 separate and distinct clusters. The first cluster – highlighted red in the map – is defined for its high cost of living. The northern Seattle and Bellevue area contain some of the most expensive homes in the data set collectively. The second cluster is heavily impacted by the venue data that was available and centered around mean location within Mercer Island. Thus, zip code 98040 was placed into its own unique cluster rightly so. The third cluster represents lower cost of living location grouped into Seattle’s southside and remote far northern corridor. My fourth cluster is missing in the map due to insufficient housing and venue data. Lastly, we have the most expensive and wealthiest zip code 98039 in its own cluster as the sixth group. I used the counter function to list the totals per each cluster: Counter({0: 19, 1: 1, 2: 7, 3: 42, 4: 2, 5: 1}).



**Discussion:** Based on the available data and the core objective developed from the problem statement, zip code 98039 is the prime location for our friend’s new Japanese/Asian fusion restaurant. This area will serve to launch the restaurant into a pool of largely wealthy customers.

**Conclusion:** Had the Foursquare API better serve a more proportioned number of venues to compare against the housing dataset, the cluster would likely appear different.