**Introduction and Description of the Problem**

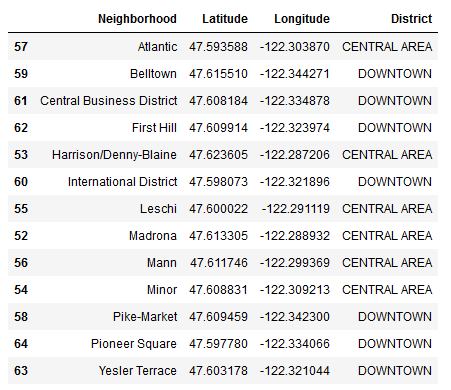
Seattle is a city that has grown rapidly over the last decade. From 2010-2018, the population of Seattle grew 22.4%, whereas the population of the United States grew 6.0%. As the city continues to grow, it is important for potential business owners to examine the types of businesses in the area. For example, a restaurant owner may not want to open a restaurant in a neighborhood that has many similar types of restaurants. To examine the types of venues in the area, Foursquare location data can be leveraged in order to select a neighborhood that is not saturated with a particular type of restaurant. Analyzing median income can also be useful in determining price points for the menu. To explore different neighborhoods in Seattle, I will create a map of the neighborhoods showing median income and clustered by venue density.

**Data**

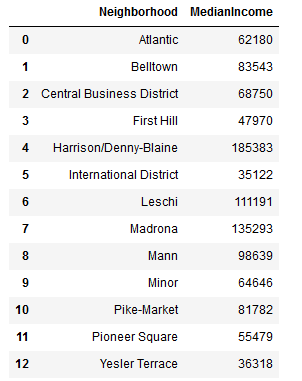
Seattle neighborhoods are not official, and many of the boundaries are informal. To examine neighborhoods and districts, I will be pulling in a GeoJSON map of the Seattle neighborhoods from the City of Seattle. This file includes the smaller neighborhoods, larger neighborhoods, and latitude and longitude. For this project, I will be focusing on the neighborhoods that are located near central Seattle. In order to examine the neighborhoods, I will leverage Foursquare API data to retrieve the 50 most common venues of each of the smaller neighborhoods in Seattle. I will then use a k-means algorithm to cluster the neighborhoods. By examining the most common venues in the clusters, I can then identify appropriate neighborhoods for a restaurant. I will also pull in data of median incomes by neighborhood. By examining median incomes, this will help business owners decide reasonable prices.

**Method**

From the GeoJSON map of the Seattle neighborhoods, I created a dataframe including Neighborhood, District, and Latitude and Longitude of the center of the neighborhood.

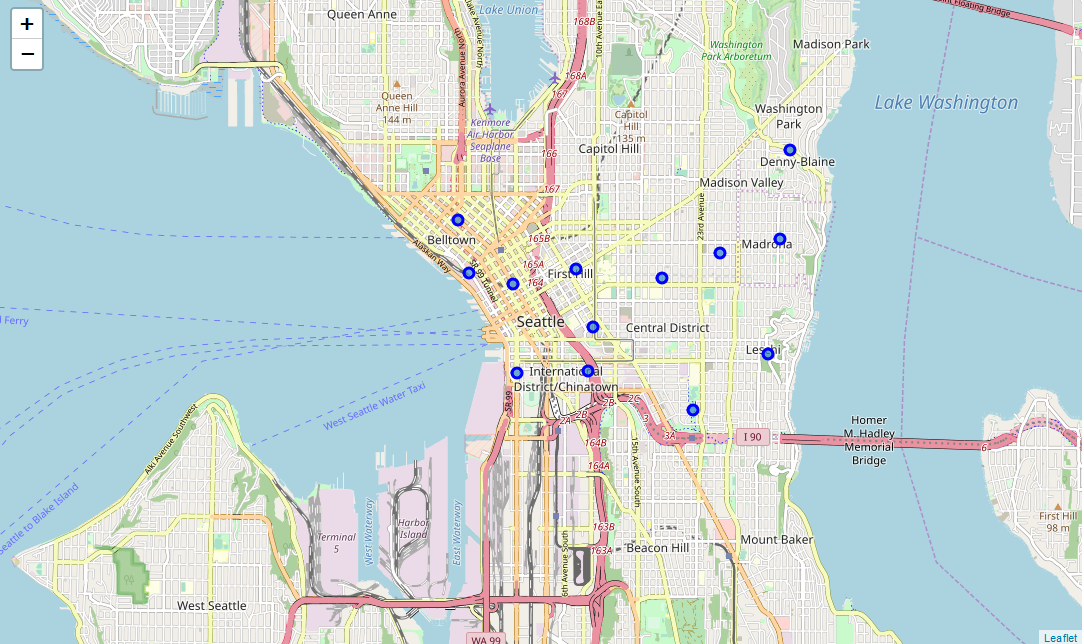


Because the neighborhoods are informal, the median incomes of the neighborhoods are not readily available from a single source. I created a csv file of median income from multiple sources (<https://statisticalatlas.com/>, <https://www.city-data.com/>, and <https://www.point2homes.com/>). This csv file was then saved to my GitHub repository. The csv file was then imported as a dataframe and merged with the dataframe of neighborhoods and coordinates.

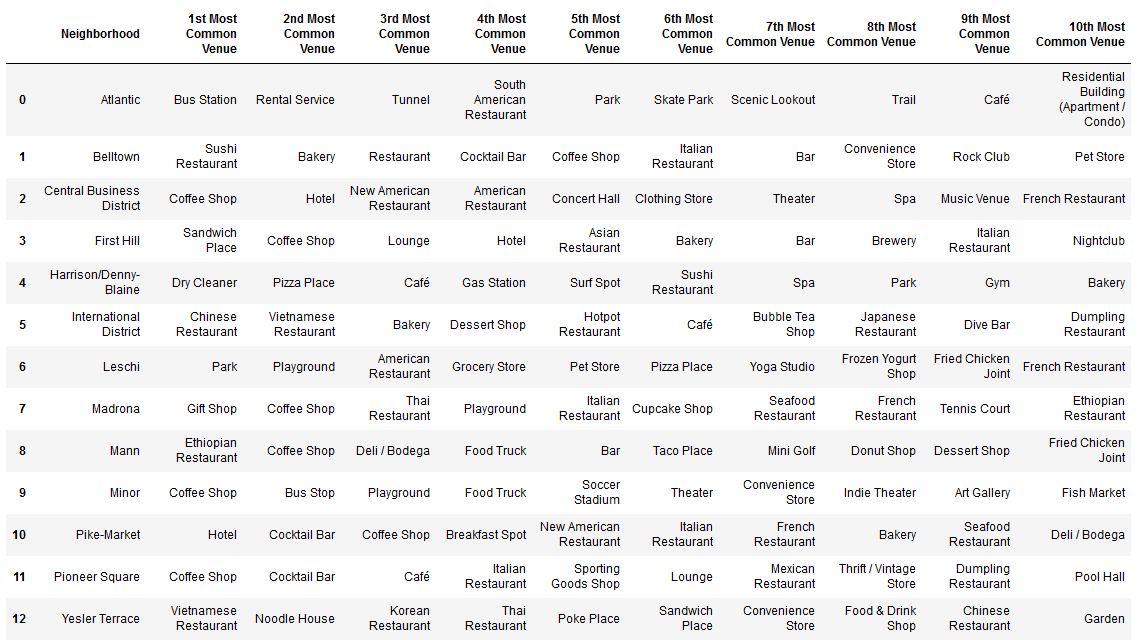




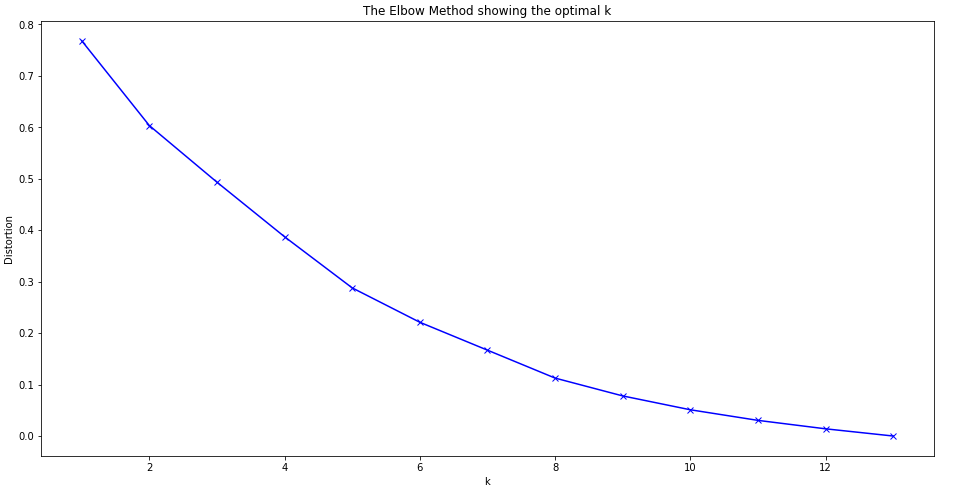
I then used folium to create a map of Seattle with the neighborhoods with markers to indicate the neighborhoods.



To examine venues in the area, I used the Foursquare API to generate a list of the top 50 venues within a 500 m radius. I created a dataframe of the 10 most common venues in each neighborhood.



To further examine the venues, I used k-means clustering, an unsupervised learning method. This is a commonly used method that groups observations together based on distance from the center of the cluster. Using the list of venues created from the Foursquare API, I first created a plot of the elbow method to determine the optimal k clusters.



When examining the graph, there is not a distinct bend to indicate the optimum number of clusters. However, there is a slight bend around 5, so I ran the k-means cluster analysis on the Foursquare data using 5 as the number of clusters. The cluster numbers were then merged into a dataframe with the different types of venues.

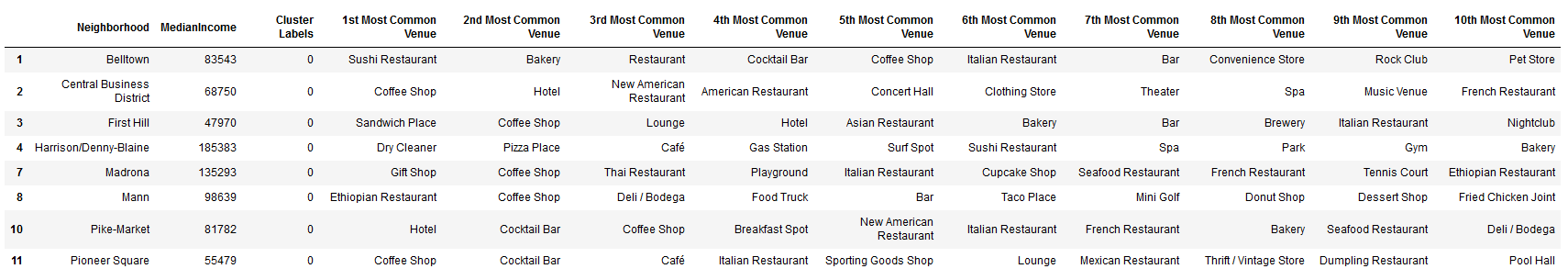
After conducting the k-means analysis, I then generated a choropleth map including median income of each neighborhood and markers to indicate the different clusters.

**Results**

***Neighborhood Clusters***

I identified 5 clusters of neighborhoods. We can further examine the characteristics of the neighborhoods.

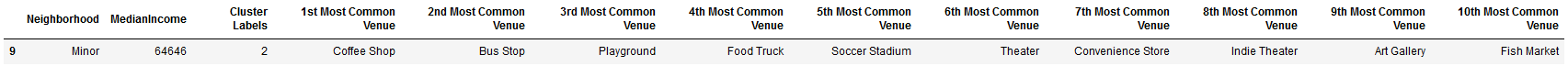
First cluster:

This cluster was the largest. The top venues in the above cluster are mainly coffee shops (first and second most common venues), but there are also a large number of restaurants of different varieties.

Second cluster:

The second cluster only consisted of one neighborhood. Venues were mostly related to travel and nature.

Third cluster:

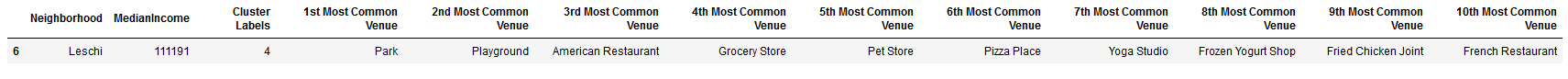


There was only one neighborhood in this cluster. The cluster consisted of many entertainment venues (e.g., soccer stadium, theater, indie theater, and art gallery).

Fourth cluster:

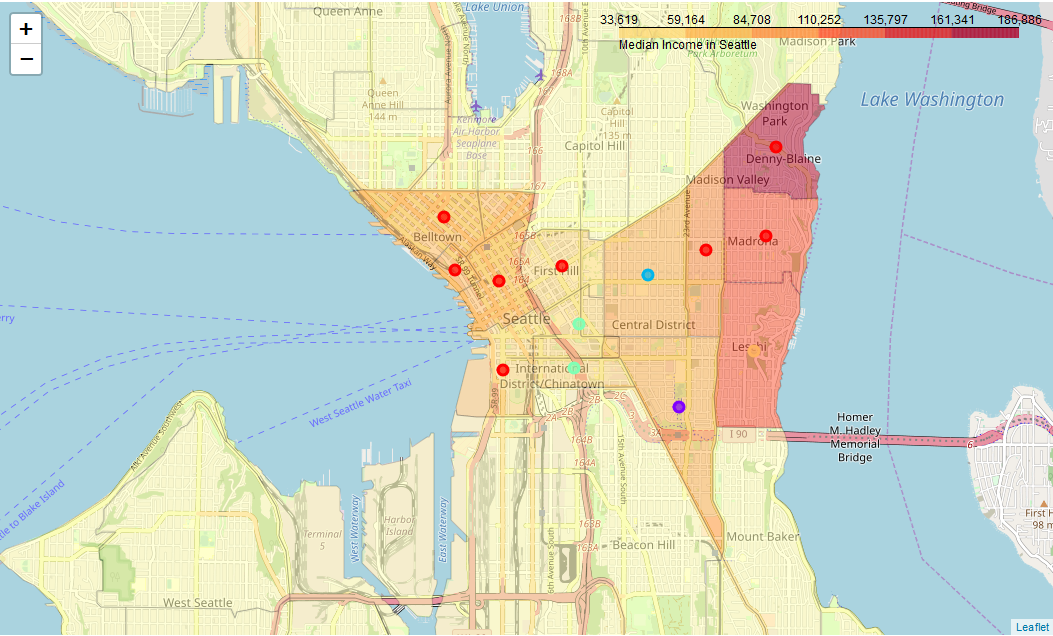
The fourth cluster had two neighborhoods and consisted mainly of Asian (Chinese, Vietnamese, Korean, Thai, and Japanese) restaurants.

Fifth cluster:

This cluster included only one neighborhood. Top venues in this neighborhood were related to recreation (park, playground, yoga studio) and a few restaurants.

***Clusters and Median Income***

To explore the median incomes of the neighborhoods, I generated a choropleth map of the median incomes of the neighborhoods of Seattle. I then added markers with different colors to indicate the different clusters. This map helps identify the neighborhood clusters in order to identify the most common venues in each neighborhood and the median income.



**Discussion**

For this project, I focused on the Downtown and Central Area districts. Using the elbow method to test the k-means algorithm did not result in an obvious optimal number of clusters, and for the k-means cluster analyses I set the value to 5 clusters. These clusters can be further explored by using the choropleth map to take into account the median income in the neighborhood.

The largest cluster consisted of eight neighborhoods. The highest median incomes for this cluster were on the east side, in the Madrona and Denny-Blaine neighborhoods. A restaurant owner could potentially set higher menu prices in these neighborhoods than in First Hill and Pioneer Square, the neighborhoods with the lowest median incomes. This cluster may consist of optimal neighborhoods since it seems to be a location where restaurants are common. Potential restaurant owners and investors will want to carefully assess the types of restaurants in order to determine how much competition in each neighborhood.

The second largest cluster consisted of two neighborhoods, International District and Yesler Terrace. If opening a restaurant in this area, a business owner or investor should take into account that there are many Asian (Chinese, Vietnamese, Korean, and Japanese) restaurants in the neighborhood, and thus there may be a lot of competition if opening up a restaurant. In addition, the median income of neighborhoods in this cluster is relatively low compared to other neighborhoods. Thus, it would likely be best to set lower menu price points.

There were three clusters that consisted of only one neighborhood each. One cluster consisted of Leschi, where the two most common venues were parks and playgrounds. Leschi also had a relatively high median cost of living, and a restaurant owner could set higher menu prices in this neighborhood. Another cluster consisted of the Atlantic neighborhood, and the top three venues were related to travel (bus station, rental service, and tunnel) and outdoors venues (park, skate park, scenic lookout, and trail). There are fewer restaurants in this neighborhood, and this may not be the optimal place to open a restaurant because there may be less demand for a sit-down restaurant. The final cluster consisted of only the Minor neighborhood. Some of the most common venues in this neighborhood were coffee shops and food trucks. This indicates that this may not be the best place to open up a restaurant either. Other common venues were attractions, including soccer stadium, theater, indie theater, and art gallery. This seems to be a neighborhood where people are most likely looking for quick food options.

**Conclusion**

As Seattle continues to grow, using data to determine optimal locations can be especially useful for future business owners and investors. Leveraging Foursquare API data to examine the most common venues in neighborhoods provides extremely valuable information for potential business owners and investors. By combining this data with information on median incomes, business owners can determine optimal neighborhoods and potential costs of menus. In the present analyses, there were some challenges since the boundaries are somewhat informal and median incomes were not available from a single source. In addition, there wasn’t a distinct optimal k identified using the elbow method. Further analyses can be done selecting different numbers of clusters to run the k-means analysis. Nevertheless, I was able to identify some neighborhoods as less than optimal due to type of venues in the neighborhood. When making a final decision, business owners may want to take into account other factors including population and real estate availability.