**Selecting Optimal Location for Pizzeria Restaurant**

**in the Sacramento Region**

by Michael K. Christoff

January 8, 2020

**1. Introduction**

My clients were moving from the east coast to Sacramento, CA in order to be closer to their extended family. They currently ran a successful pizzeria in New Jersey which they are planning to sell, and with the proceeds open a new pizzeria in the Sacramento region. They asked me to help determine, through data science research, which regions would be an ideal spot for their new restaurant.

The cost of retail space was not an important consideration for my clients, but they insisted the new location be a high traffic, growing region that is near event venues, corporate centers and recreational attractions. They also wanted to minimize competition by avoiding areas already inundated with pizzerias.

They made it clear that they really love downtown Sacramento, but they were aware the region surrounding Sacramento had many great opportunities. As long as they are just a few hours away from Sacramento to be able to visit family regularly, they want to explore every opportunity before making their decision on the best location.

1. **Data**

I knew they prized the downtown area, so the first part of my research would focus on the city of Sacramento and neighborhoods. The second part of my research would focus on the cities within a two hour drive of Sacramento.

**2.1 Data for city of Sacramento search**

For my search within the Sacramento city limits, I will use a tourism website planetware, which highlighted top-rated attractions within the city of Sacramento. Based on each attraction, I will explore their immediate surrounding area to identify popular venues and then make comparisons. For this, I will use the Foursquare website. Foursquare, the world’s leader in location-based data, uses crowdsourcing to collect specific information about locations and venues all over the world. I will use Nominatim geolocator location data to covert location name or address to map markers onto a Folium map. I will use K-means clustering to determine cluster labels to add to the Folium map and my results.

**2.2 Data for cities outside Sacramento search**

For the extensive region surrounding Sacramento, I will concentrate on the largest populated cities and those with the highest median family income. I will use U.S. Census Bureau to determine the largest cities in the Sacramento region, as well as specific population and median family income for these cities. I will explore cities surrounding Sacramento based on venue information from Foursquare. I will use Nominatim geolocator location data to map markers onto a Folium map. I will use K-means clustering to determine cluster labels to add to the Folium map and my results.

1. **Methodology**

**3.1 Methodology for city of Sacramento search**

I performed all my analysis using the programming language Python, within a Jupyter Notebook [1].

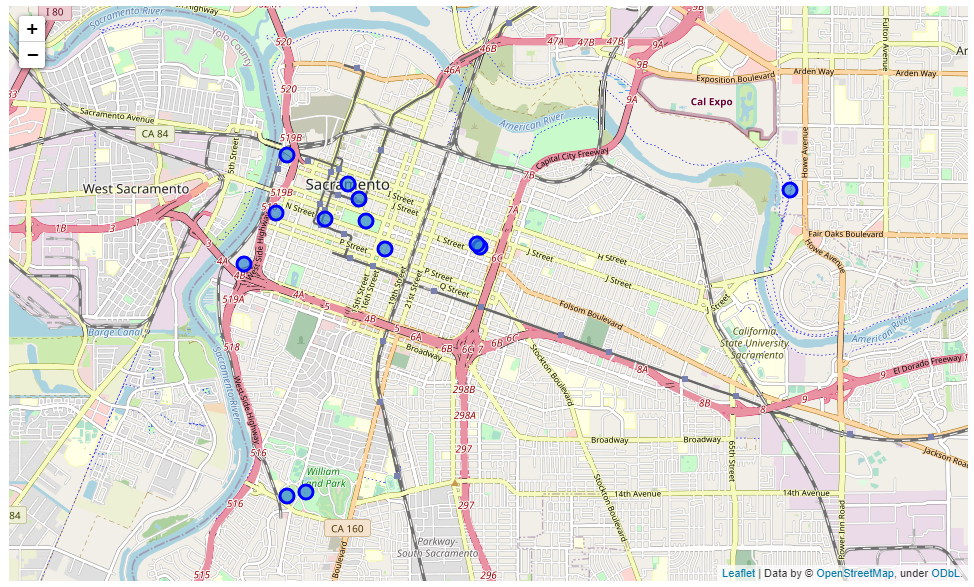
I started with the tourism website planetware article on Sacramento’s top-rated attractions to begin my research [2]. To scrape the list of attractions from the site, I used BeautifulSoup, an HTML parser created by Leonard Richardson. After scraping the attraction names, I stored the list of names into a panda DataFrame ‘Attraction\_Name’.

I used the Nominatim geolocator tool to identify the location of each named attraction. Nominatim is a tool to search OSM (OpenStreetMap) data by name and/or address (geocoding) to generate latitude longitude coordinates for each variable. I only had the attraction name, and when I fed these names into the geolocator, there were instances where the tool did not return the correct coordinates. To adjust for this, I cleaned the “Attraction\_Name” list with a “for loop” statement to remove unnecessary commas and spaces, and then adjusted a few of the attraction name strings. Once clean, I used a “for loop” to send the names through the geolocator and stored the latitude and longitude coordinates into two empty DataFrames I had created for “Latitude” and “Longitude” numerical float variables. Once the three DataFrames were complete and cleaned (‘Attraction\_Name’, ‘Latitude’, ‘Longitude’) I merged the three lists into a single DataFrame named ‘Attractions’ with columns for ‘Attraction’, ‘Latitude’ and ‘Longitude’.

DataFrame ‘Attractions’:

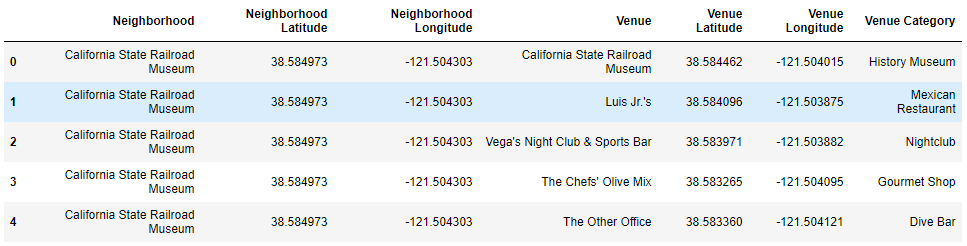


At this point, I wanted to map out each attraction onto an interactive map. I chose Folium, which makes it easy to visualize data that's been manipulated in Python on an interactive leaflet map. I created a base map of Sacramento using its latitude and longitude coordinates using the Nominatim geolocator, and used a ‘for loop’ to add each marker to the base map (see image):

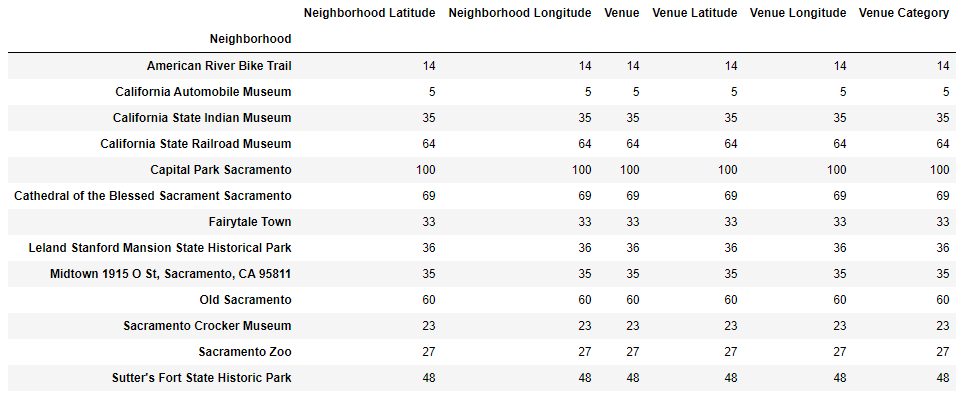


Based on each location on the map, it was time to feed these locations through the Foursquare API to generate a list of venues. I used a function I created named ‘getNearbyVenues’ which fed the location data from my DataFrame ‘Attractions’ to the Foursquare API. I set the search parameters to identify up to 100 venues that were located within a 500 ft radius of each attraction. With the function call, I created a new DataFrame called “sac\_venues” with new columns for ‘Neighborhood’, ‘Neighborhood Latitude’, ‘Neighborhood Longitude’, ‘Venue’, ‘Venue Latitude’, ‘Venue Longitude’ and ‘Venue Category’. Based on a ‘len’ count of the ‘Venue Category” column, the list contains 122 unique venue categories.

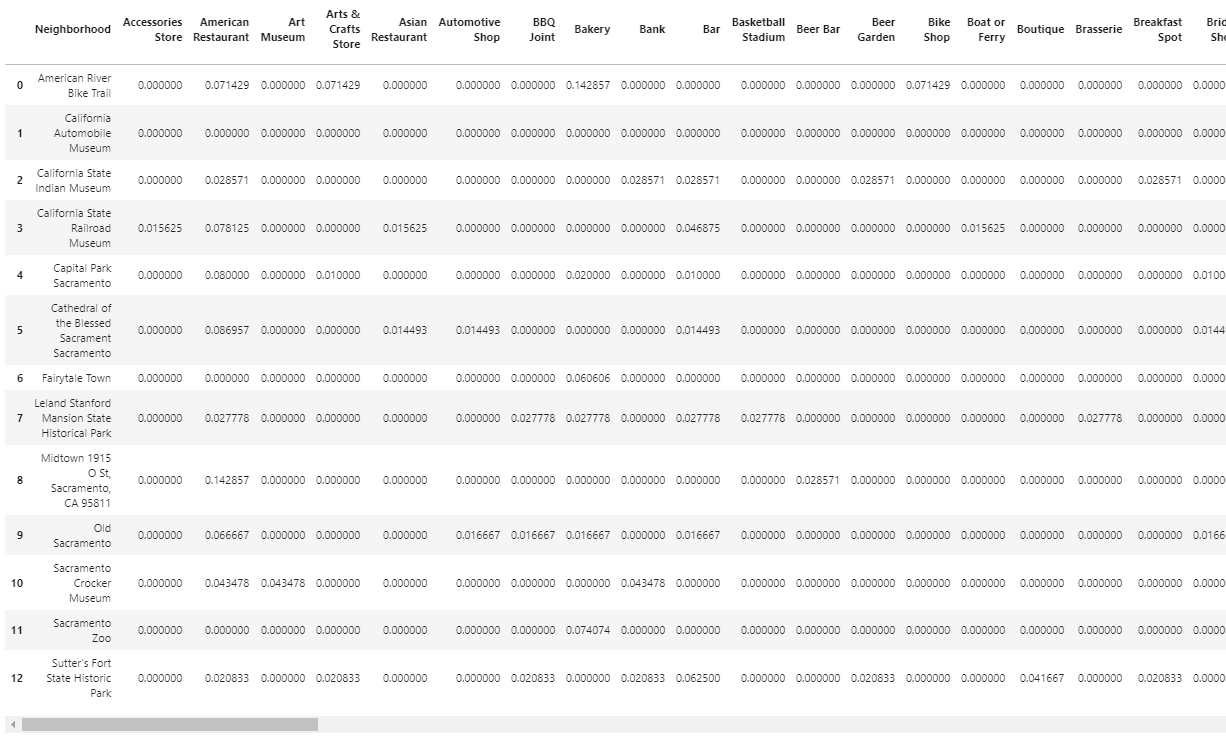
List of first 5 rows of DataFrame ‘sac\_venues’:



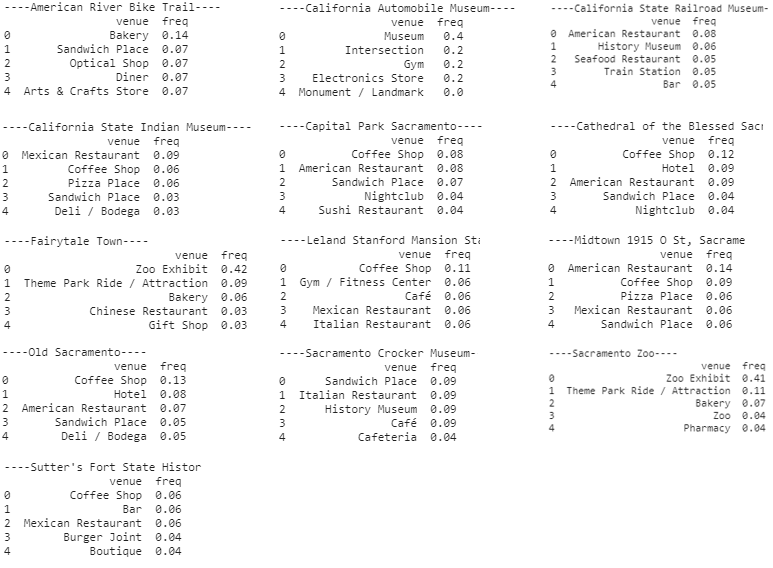
I utilized the ‘groupby’ function to count the ‘sac\_venues” DataFrame, here is the list of total venues located nearby each location:



Next, I used the ‘onehot’ method to determine the top 5 most frequent venue categories around each location. This consists of taking our ‘sac\_venues’ DataFrame and converting the Venue Categories column to binary choices (0, 1) using the pandas ‘get\_dummies’ function. This creates a new DataFrame ‘sac\_onehot’, which we add the ‘Neighbors’ column to from ‘sac\_venues’. We then transform the ‘sac\_onehot’ DataFrame into a new DataFrame named ‘sac\_grouped’ that uses the groupby function to group the Neighborhoods and apply a mean value to each group:



Now that we have this data, I sorted the Neighborhoods in descending order and took the top 5 values for each neighborhood. This gives us the top 5 most common venue category for each Neighborhood:



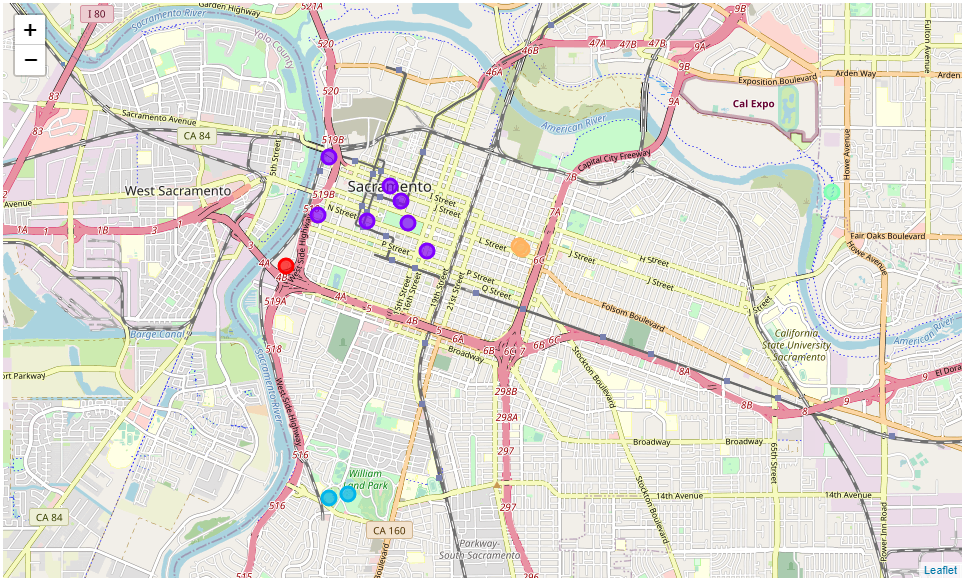
I then created a function that returned the most common venues, and created a new DataFrame ‘neighborhood\_venues\_sorted’ and populated it with the top 10 venues in each neighborhood:



With this data, I was ready to use K-means Clustering algorithm to create correlated clusters, groupings to gain new insight into the data. I utilized a kcluster value of 5. This gave me values [0,1,2,3,4] for each Neighborhood based on the clustering results of the algorithm. I then created a new DataFrame named “sac\_merged’ that combined ‘Attractions’ columns ‘Attraction’(name), ‘Longitude’ and ‘Lattitide’ coordinates with the ‘neighborhood\_venues\_sorted’ data. This prepared me to create an interactive map of the clusters. Here is a peek of the new data frame:

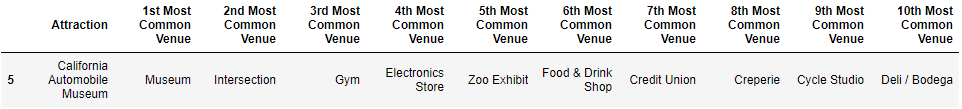


Using Folium, I created an interactive map of this data:

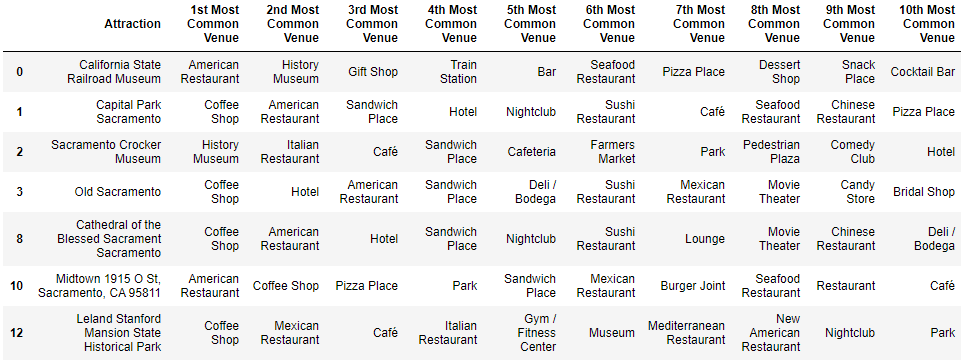


Here are details about each cluster from the map:

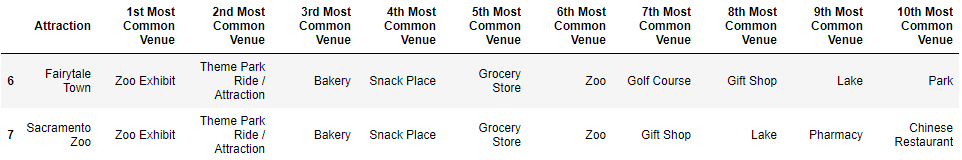
**Cluster 1 Automobile Museum Area:** (**red** marker on map)



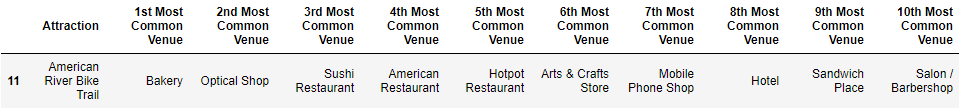
**Cluster 2 Capital Area:** (**purple** markers on map)



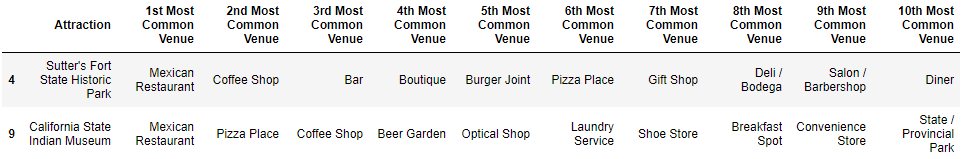
**Cluster 3 Zoo & Theme Park Area:** (**blue** markers on map)



**Cluster 4 American River Area:** (**green** marker on map)

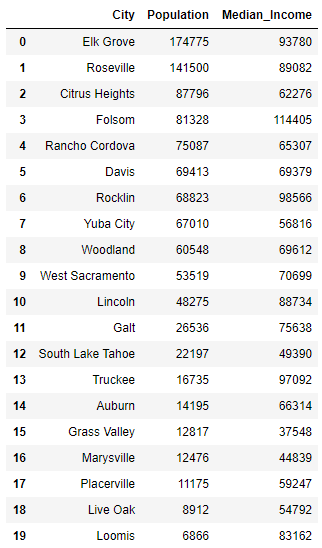


**Cluster 5 Sutter Fort Area:** (orange markers on map)

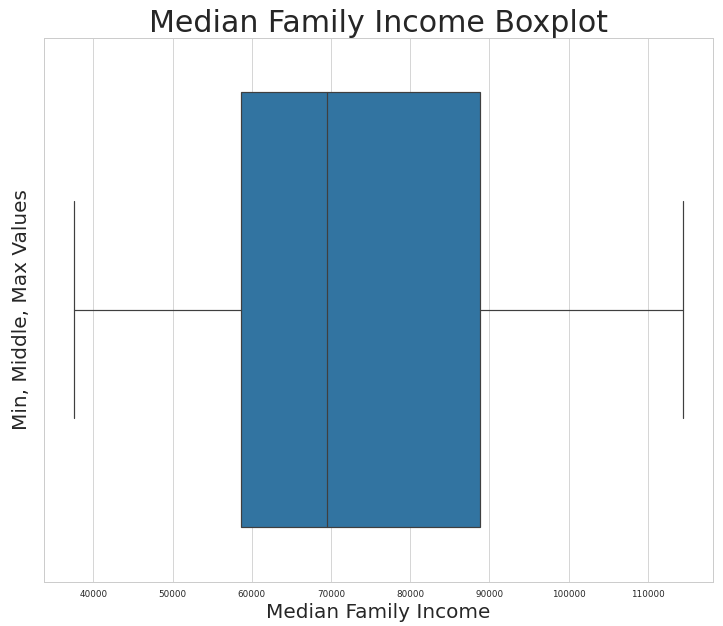


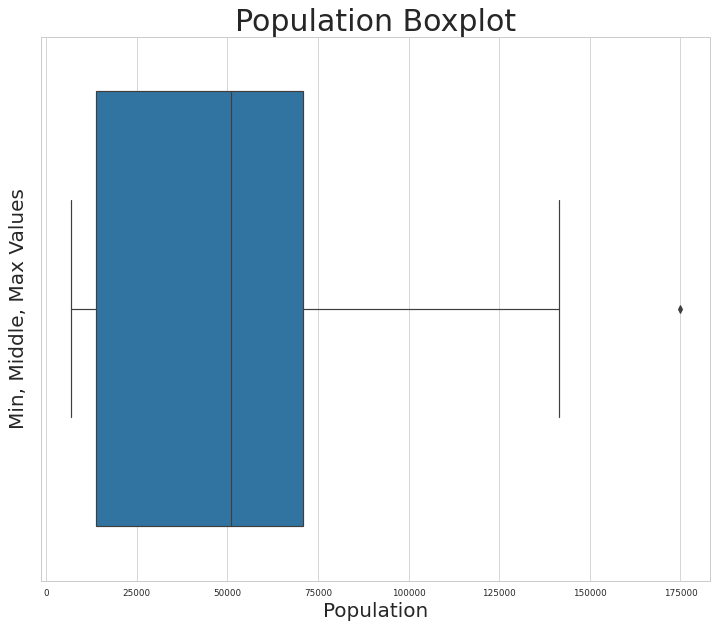
* 1. **Data for cities outside Sacramento search**

I started by creating a CSV file of the top 20 most populated cities in Sacramento, along with their population and median family income[4]. I read the CSV file into my Jupyter Notebook using pandas ‘read\_csv’ function and named the new DataFrame‘df2’:

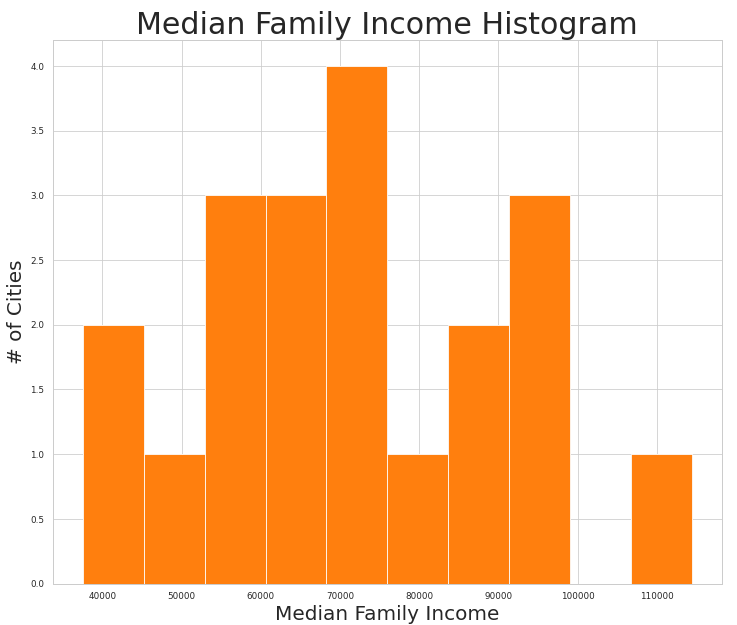


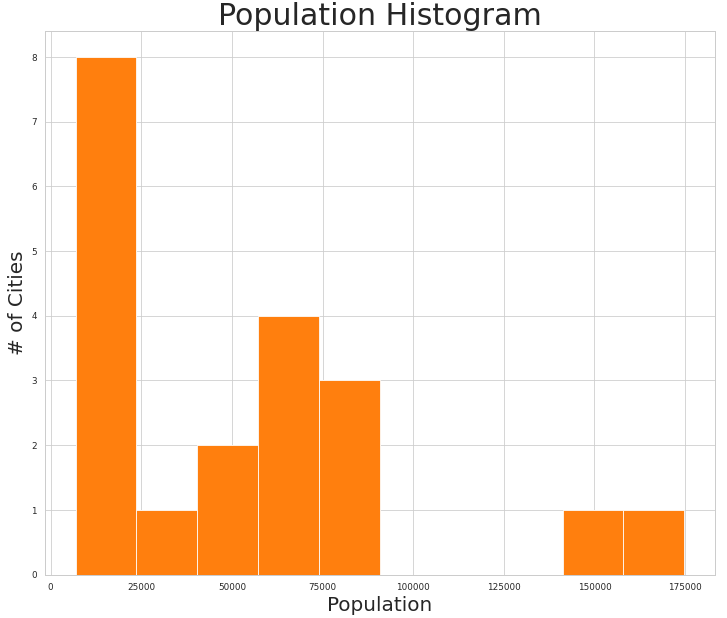
I used pyplot to create **Boxplot** charts for Median Family Income and Population:



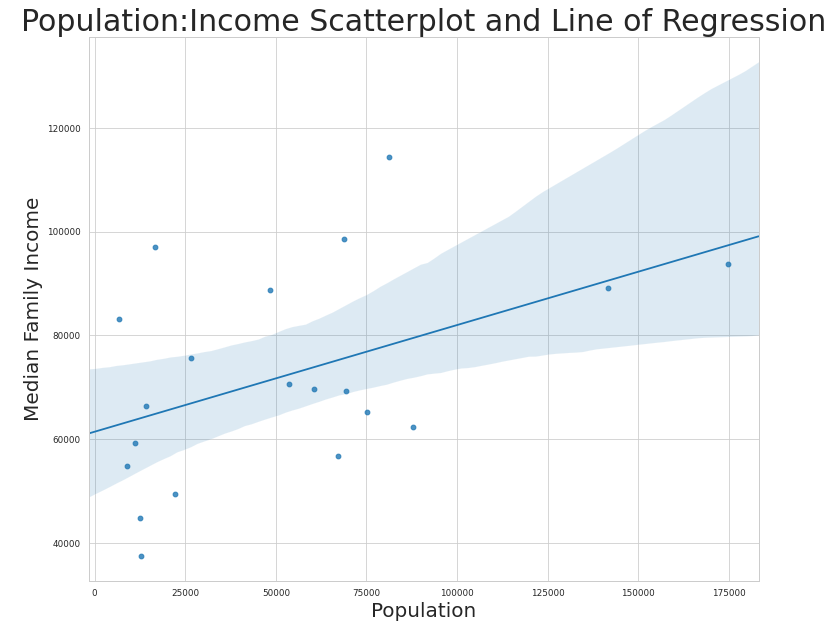


I used pyplot to create **Histogram** charts for Median Family Income and Population:





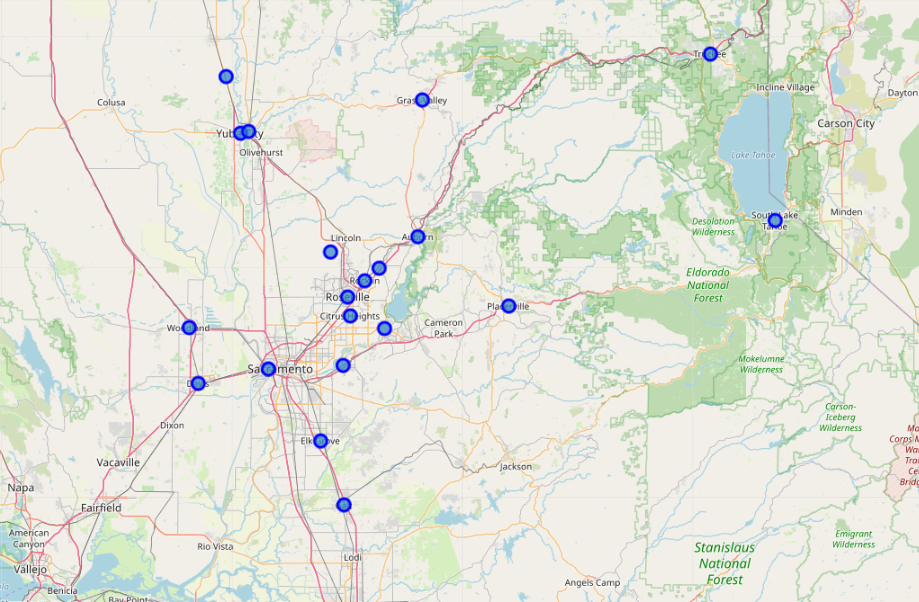
I created a scatterplot with Population on the x-axis, and Median Family Income on the y-axis. Within the scatterplot, I also included a linear line of regression by using ‘regplot’:



I converted the DataFrame into a temporary list using the ‘values.tolist’ function, then ran the list through the Nominatim site to get location data, storing coordinates into new empty DataFrames ‘Latitude2’ and ‘Longitude2’. Then I created a new DataFrame named ‘Cities’ and merged together columns for 'City' (using ‘names’ list), 'Latitude’ (using ‘Latitude2’) and 'Longitude' (using ‘Longitude2’). I cleaned up the ‘City’ column by using the ‘replace’ function to remove unnecessary ‘[’ and ‘]’ special characters:



I created a base map of Sacramento using geolocation coordinates obtained through the Nominatim geolocator tool. I then used the ‘astype’ function to update the ‘Latitude’ and ‘Longitude’ columns to be numerical floats to prepare for the Folium mapping. I utilized Folium to add markers to my base map:

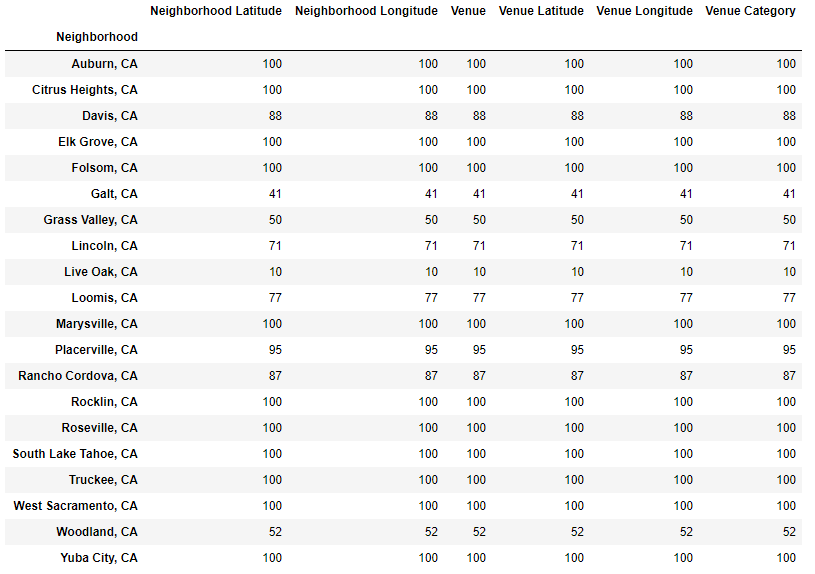


I sent the data from “Cities” to the Foursquare site to obtain venue information about each city. In these surrounding cities, a car or bus is usually how people travel between venues. So my Foursquare search was defined to return a list of 100 top venues within 1 mile (radius=5280 ft) of each city center. I used shape to determine the list of venues (qty 1671 unique venues):

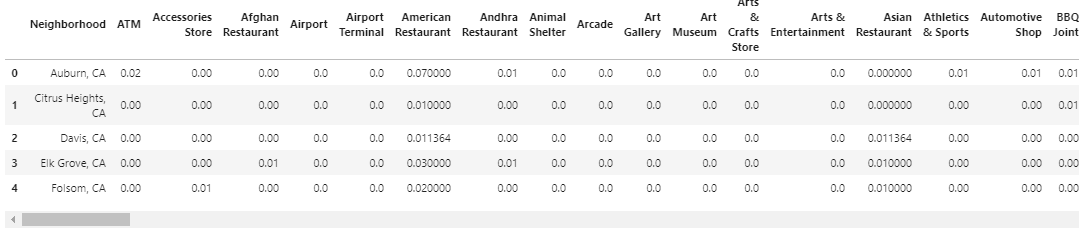
Here is a list of the first 5 Venues:



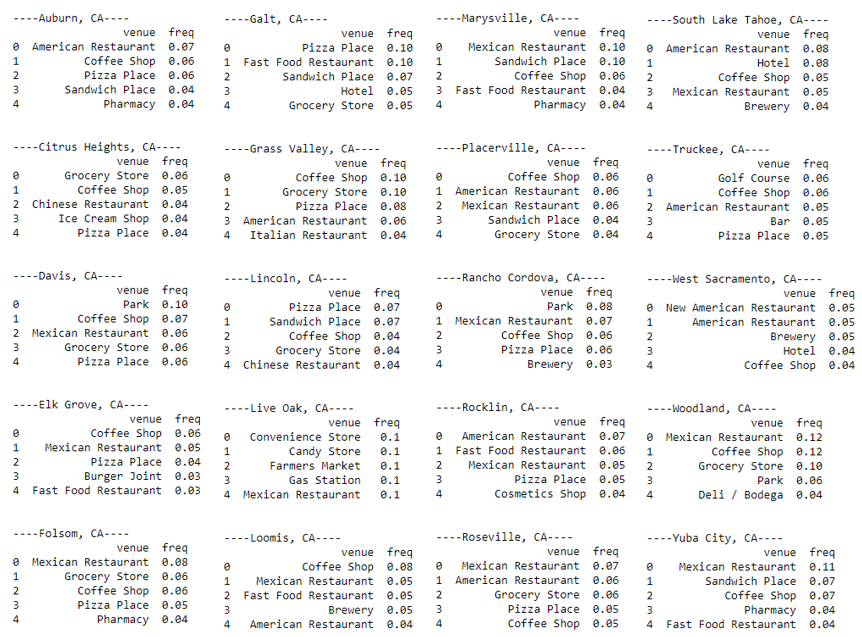
I performed a ‘groupby’ count to determine the number of venues within 1 mile of each city:



I created a new one hot list of this data using the ‘get\_dummies’ function and named it the “region\_onehot’ DataFrame, and moved the Neighborhood column to the first position. I  
 calculated the mean value of each venue category based on the list grouped by “Neighborhood”:



I sorted this DataFrame in descending order, and took the top 5 rows for each Neighborhood in the ‘region\_grouped’ DataFrame. This game me the top 5 common venue for each city:

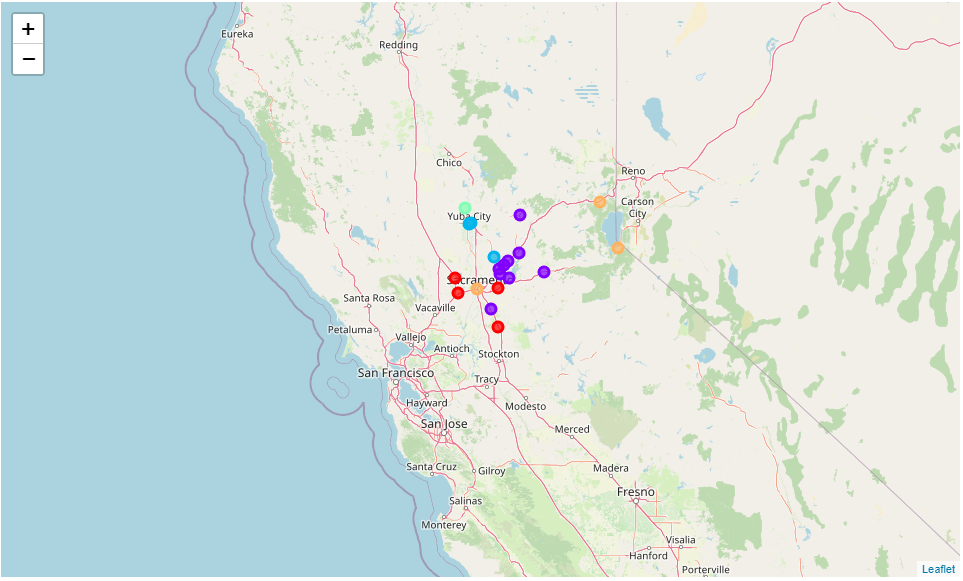


Next, I sorted the ‘region\_grouped’ by ‘Neighborhood’ and created a new DataFrame named ‘neighborhoods1\_venues\_sorted’.

With this data, I was ready to use K-means Clustering algorithm to create correlated clusters, groupings to gain new insight into the data. I utilized a kcluster value of 5. This gave me values [0,1,2,3,4] for each Neighborhood based on the clustering results of the algorithm. I then created a new DataFrame named “sac\_merged’ that combined ‘Attractions’ columns ‘Attraction’(name), ‘Longitude’ and ‘Lattitide’ coordinates with the ‘neighborhood\_venues\_sorted’ data. This prepared me to create an interactive map of the clusters. Then I created a new DataFrame named “region\_merged” and merged ‘neighborhoods1\_venues\_sorted’ and ‘region\_merged’ lists:  
  
 Here are the first 5 rows of this list, sorted with cluster names and location data added:



I took the ‘region\_merged’ DataFrame and fed the location data into Folium and added markers based on clusters to the updated map:



Here are details about each cluster from the map:

**Cluster 1 Near Sacramento Parks Area:** (**red** markers on map)



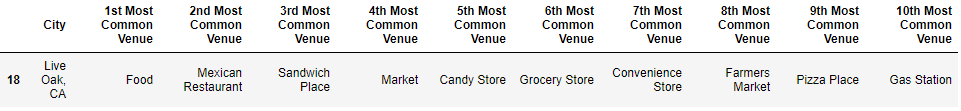
**Cluster 2 HWY80 West Area:** (**purple** markers on map)



**Cluster 3 HWY65 NW Area:** (**blue** markers on map)



**Cluster 4 Farthest Area:** (**green** marker on map)



**Cluster 5 Hotel Area:** (**orange** marker on map)



1. **Results**

**4.1 Results for city of Sacramento search**

Out of the thirteen attractions, the surrounding neighborhood contained 545 venues within 500 feet of the attraction, which represented 120 unique categories of venues.

There were 545 venues found within 500 feet of all 13 attractions.

There were 120 unique venue categories found within 500 feet of all 13 attractions.

Capital Park (97) and the Cathedral (69) represented the most unique venue categories, while the Automobile Museum (4) and Bike Trail (13) represented the least.

Cluster 2 Capital Area seemed to have the bulk of the neighborhoods closest to the city center, while the other clusters were outside the city center in areas with a lower number of inque vendor categories

Three clusters (clusters 1, 3 and 4) did not contain Pizzerias within 500 ft of the neighborhood center.

**4.2 Results for cities outside Sacramento search**

Out of the top 20 most populated cities in Sacramento’s surrounding area, Elk Grove (174,775) and Roseville (141,500) had the highest population while Live Oaks (8,912) and Loomis (6,866) had the least.

Out of the top 20 most populated cities in Sacramento’s surrounding area, Folsom (114,405), Elk Grove (93,780) and Rocklin (98,566) had the highest median family income while Grass Valley (37,548) and Marysville (44,839) had the least.

The majority of median family income was between just under $60,000 and just under $90,000, with the full range from $37,548 and $114,405.

The majority of cities had a population between just under 25,000 and just under 75,000, with the full range from 6,866 and 274,775.

There were eight(8) cities in the results with less than 25,000 population.

Based on the scatterplot and regression analysis comparison, there was a positive correlation between population and median family income. This means as population of the cities were higher, it was more likely those cities had higher median family income as well.

There were 1,671 venues found within 1 mile of all 20 cities.

There were 219 unique venue categories found within 1 mile of all 20 cities.

There were 9 cities that reached the 100 limit of unique venues within 1 mile, while Galt (41) and Live Oak (10) had the fewest.

1. **Discussion**

**5.1 Discussion for city of Sacramento search**

I would avoid the outliers which are a bit far from the city center, such as Cluster 4 American River Area, Cluster 3 Zoo and Theme Park Area. For the Zoo & Theme Park area specifically, there may not be a Pizzeria within 500 feet of this neighborhood, but the nature of the Zoo and Park, which serve their own food within these parks, means that families will probably be full after their day at the park. The American River does not have a Pizzeria either, but traffic would not be as high as a location nearer the city center.

Cluster 5 would be a higher traffic area close to the heart of the city, but there are multple Pizzerias already present which may lead to the competition you wanted to avoid.

Cluster 1 California Automobile Location looks promising. There is no current Pizzaria within 500 ft of this location, and it is situated near Old Sacramento which is a popular destination for both tourists and locals. Also, a highway intersection is near which could benefit a delivery operation to expand profitable opportunities for your restaurant.

Cluster 2 offers the largest traffic opportunity, and is closest to the city center. With this, will come competition from the many other restaurants in the area. This area includes our state capital, and is full of business and government agencies that would lead have a high demand for lunch, delivery and catering opportunities.

**5.2 Discussion for cities outside Sacramento search**

Cluster 4 Live Oak area are low population centers that are far from the city of Sacramento. To add, the Cluster 4 Live Oak area represents a city with a lower median family income when compared to other cities in the region.

Cluster 3 is also a low population area, and Yuba City and Marysville also represent cities with lower median family income levels.

Cluster 1 Parks area great locations if you want to be near parks, but the presence of other Pizzerias may lead to competition which you wanted to avoid.

Cluster 2 are similar in that there would be competition from other Pizzerias, and both contain cities with higher populations and median family income when compared to other Clusters on the map. I would think if the surrounding area is your desire, your choice would come from one of these two clusters.

1. **Conclusion**

**6.1 Conclusion for city of Sacramento search**

If you choose to locate your new restaurant within the Sacramento city limits, the data discovered with Foursquare location data and K-means clustering on the top 13 attractions will provide you with insight that will help you make an informed decision.

If you chose to stay within the city limits, your choice should most likely come from Cluster 1 or 2. If you are competion averse, chose a location in Cluster 1, and if you don’t mind a little competition and can manage the highest traffic area of the city, chose a location in Cluster 2.

**6.2 Conclusion for cities outside Sacramento search**

If you choose to locate your new restaurant outside Sacramento, the data discovered with U.S. Census Bureau data, Foursquare location data and K-means clustering on the top 20 most populated cities within a 2 hour drive of the city will provide you with insight that will help you make an informed decision.

I would think if the surrounding area is your desire, your choice should come from Cluster 1, 2 or 3. Choosing a city with a high population in Clusters 1 or 2 will ensure the highest traffic, with most likely higher median family income. Cluster 3 Hotels is also promising, but two locations, Trukee and South Lake Tahoe, represent areas with harsh winters and seasonal tourism cycles.

1. **References**

[1] Jupyter Notebook on Github, Viewed through nbviewer to enable interactive mapping <https://nbviewer.jupyter.org/github/mkchrist/Coursera_Capstone/blob/f36473776ef7bb7bcde8fecc28c4c20f43941d28/Coursera_Capstone_Notebook_Final.ipynb>

[2] planetware, “13 Top-Rated Tourist Attractions & Things to Do in Sacramento, CA”

Written by Lana Law, Sept 14, 2020:

<https://www.planetware.com/tourist-attractions-/sacramento-us-ca-sc.htm>

[3] U.S. Census Bureau, Regional City Data on Population & Median Family Income

<https://www.census.gov/quickfacts/sacramentocountycalifornia>