The Big Fish

The Battle of the Neighborhoods:

The search for the largest and warmest US coastal City with the most suited neighborhood, with alternatives, to start a chain of Seafood Restaurants!

Applied Data Science: kMeans Clustering

(Unsupervised Machine Learning Algorithm)

Author: Armand van der Merwe

arri.vdm@gmail.com

+27 72 285 5596

South Africa

July 2020

Table of Contents

1. Introductory Section

1.1. Discuss the business problem

Background

The client is the proprietor of a thriving chain of high-end SEAFOOD RESTAURANTS in the city of **Johannesburg, Gauteng**, South Africa. The company also has its **head office situated in Johannesburg**, the most populous city in the Gauteng province. The restaurants are clustered around the city, with its 12, 270, 000 residents.

The **CENTRALIZATION** of the chain around Johannesburg has made it more manageable to operate them from a logistical perspective. The fact that the chain is in a **LARGE city** also has many advantages. The client would have preferred **a COASTAL city**; for fast access to fresh seafood produce.

In recent years the client has become increasingly concerned with South Africa's socio-economic environment:
- The **high crime** rates,
- **High unemployment** rates and,
- Low incomes

Some of the factors mentioned above, in time, have led to a **struggling economic climate**, which has **lead to thefts and violent crimes**. The client is gravely concerned, as it has affected his family, as well as his business. Although he cares for his country enormously, he has ACCEPTED an offer to buy his chain of seafood restaurants. **THE CLIENT HAS DECIDED TO IMMIGRATE TO THE UNITED STATES OF AMERICA AND START A NEW CHAIN OF SEAFOOD RESTAURANTS.**

Business Problem

THE CITY: The client would like to find the **best possible coastal city and then neighborhood to start his new chain of Seafood Restaurants** based on factual and scientific data analysis:

The client has requested a coastal city in the US, for fast access to fresh seafood produce. He then wants a city with the highest possible population to give his new Seafood Restaurant chain the best possible chance. As the client is from Johannesburg South Africa, he is very mindful of climate, as he is used to comfortable temperatures. It has to be expressly noted that he requested that the cities be compared, in terms of the highest-high and highest-low temperatures throughout the year.

THE NEIGHBORHOOD: Further to the above, the client then wants to know the best cluster of neighborhoods in the city to start the chain of Seafood restaurants. He could decide to place the chain of restaurants in the clusters of neighborhoods with high counts of seafood restaurants. He would thus effectively be taking on the competition head-on. Alternatively, he could place the restaurants in strategic positions away from the clusters of neighborhoods with high counts of seafood restaurants, effectively to avoid the competition, yet be in the correct general location. It is vital for him to be able to visualize the neighborhoods on an interactive map with the clusters to make sound business decisions.

Stakeholders

The **client [1]** has the resources and sees the entire process as a business investigation, and he has shrewdly chosen to use a scientific approach to ascertain the most advantageous neighborhoods.

Upon a recommendation from a trusted business associate, the client has elected to enlist the help of a **Data Scientist, Armand van der Merwe [2]**, to help him solve this critical and complex business problem.

Goal

The client would like to find the largest and warmest US coastal City with the most suitedneighborhood, with alternatives, to start a chain of Seafood Restaurants.

The Measure of the Result

The measure of the result would be if the outcome presented to the client by myself, the appointed Data Scientist, meets all his requirements, as set out above in the Business Problem Section.

1.2. Discuss who would be interested in this project

THE CITY: A private or business individual(s) who would like to know which US coastal city has the largest population, the best average highest-high temperatures, and the best average highest-low temperatures per year. The result is Los Angeles, California; it is necessary at this stage to name the city in order to continue.

THE NEIGHBORHOOD: A private or business individual(s) who is interested in Seafood Restaurants in Los Angeles, California. The concerned party would like to know where the most prominent clusters of Seafood Restaurants are. Also, within these clusters, the neighborhoods associated with the groups. The regions contain further information on the names and locations of individual Seafood Restaurant venues.

This could be very useful in terms of deciding where to open a Seafood Restaurant, with varying business strategies possible. One of these strategies could include being close to "The competition," and another could be to be in the strategic locations "away from the competition".

2. Data Section

2.1. Data Sources

THE CITY:

A1. Description: As part of the brief, I needed to determine the largest coastal cities in the US. In my research, I came upon a Wikipedia webpage with the information I needed. Now, there was no need to do anything other than view the image to identify the largest COASTAL CITIES.
Method: Visually, it is possible to narrow them down with data on Wikipedia:

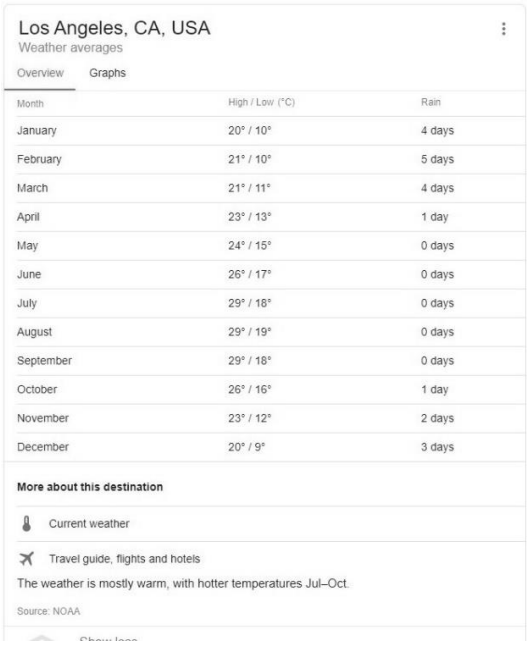
- New York (1st largest)
- Los Angeles (2nd largest)
- San Diego (8th largest)
- San Jose (10th largest)

Source: https://simple.wikipedia.org/wiki/List_of_United_States_cities_by_population
Example:

A2. Description: I needed a table with data on how US cities rank in size. I also needed the actual city names with their populations in millions. It was helpful to find this table as it was very well prepared and would not need a massive amount of Data Wrangling. Having visually established the largest US cities above, it made the process easier. This table is an example of structured data as it is in the form of rows and columns.
Method: Website Scraping
Source: https://simple.wikipedia.org/wiki/List_of_United_States_cities_by_population
Example:

2017 rank	City	State	2017 estimate	2010 Census	Change	2016 land area		2016 population density		Location
1	*New York*[3]	New York	8,622,698	8,175,133	+5.47%	301.5 sq mi	780.9 km²	28,317/sq mi	10,933/km²	40.6635°N 73.9387°W
2	*Los Angeles*	California	3,999,759	3,792,621	+5.46%	468.7 sq mi	1,213.9 km²	8,484/sq mi	3,276/km²	34.0194°N 118.4108°W
3	*Chicago*	Illinois	2,716,450	2,695,598	+0.77%	227.3 sq mi	588.7 km²	11,900/sq mi	4,600/km²	41.8376°N 87.6818°W
4	*Houston*[4]	Texas	2,312,717	2,100,263	+10.12%	637.5 sq mi	1,651.1 km²	3,613/sq mi	1,395/km²	29.7866°N 95.3909°W
5	***Phoenix***	Arizona	1,626,078	1,445,632	+12.48%	517.6 sq mi	1,340.6 km²	3,120/sq mi	1,200/km²	33.5722°N 112.0901°W
6	*Philadelphia*[5]	Pennsylvania	1,580,863	1,526,006	+3.59%	134.2 sq mi	347.6 km²	11,683/sq mi	4,511/km²	40.0094°N 75.1333°W
7	San Antonio	Texas	1,511,946	1,327,407	+13.90%	461.0 sq mi	1,194.0 km²	3,238/sq mi	1,250/km²	29.4724°N 98.5251°W
8	San Diego	California	1,419,516	1,307,402	+8.58%	325.2 sq mi	842.3 km²	4,325/sq mi	1,670/km²	32.8153°N 117.1350°W
9	Dallas	Texas	1,341,075	1,197,816	+11.96%	340.9 sq mi	882.9 km²	3,866/sq mi	1,493/km²	32.7933°N 96.7665°W
10	San Jose	California	1,035,317	945,942	+9.45%	177.5 sq mi	459.7 km²	5,777/sq mi	2,231/km²	37.2967°N 121.8189°W

A3. Description: I need to determine the average monthly **highest-high** and **highest-low temperatures** throughout the year and later, average it to per year. I had to do the same search for all the US coastal cities and Johannesburg in South Africa. Because I was unable to apply website scraping, I copied the data into an excel sheet and created my structured data (data in rows and columns). After having completed the process, I just had to save the excel sheet as a .CSV file and upload the data. I would have preferred a structured source, but none was available to have the data in the amalgamation and format that I needed; therefore, it was necessary to do it manually.
Method: Create .CSV files with temperature information, as no structured source data exists
Source: https://www.noaa.gov/
Example:



Los Angeles, CA, USA
Weather averages

Overview | Graphs

Month	High / Low (°C)	Rain
January	20° / 10°	4 days
February	21° / 10°	5 days
March	21° / 11°	4 days
April	23° / 13°	1 day
May	24° / 15°	0 days
June	26° / 17°	0 days
July	29° / 18°	0 days
August	29° / 19°	0 days
September	29° / 18°	0 days
October	26° / 16°	1 day
November	23° / 12°	2 days
December	20° / 9°	3 days

More about this destination

Current weather

Travel guide, flights and hotels

The weather is mostly warm, with hotter temperatures Jul–Oct.

Source: NOAA

THE NEIGHBORHOOD:

B1. Description: Further on in the project, I was going to have to do kMeans clustering, which is an unsupervised machine learning algorithm. I was also going to need to illustrate the results on a folium map. Therefore, I needed all the neighborhoods in Los Angeles with their latitudes and longitudes. The data from the .json file was unstructured, not in tables and rows, and would need a significant amount of data wrangling.
Method: Downloading .Json data
Source: https://usc.data.socrata.com/
Example:


```
['row-nhgs-3gdi~sq5y',
 '00000000-0000-0000-47CF-3583B50548BD',
 0,
 1503434412,
 None,
 1503434412,
 None,
 '{ }',
 'L.A. County Neighborhoods (Current)',
 'acton',
 'MULTIPOLYGON (((-118.20261747920541 34.53898972076929, -118.18946958918568 34.5385546636616, -118.18950400422953 34.5349457732411, -118.185124836341 34.53482956044709, -118.18516440876348 34.5
312465197055 3, -118.17601577983017 34.5313547024300 15, -118.1761893084381 34.523803185624594, -118.167025613659 65 34.523512278232 81, -118.162940265952 81 34.523716853632315, -118.162988882794 76 3
4.527586918918 19, -118.15426797766412 34.527789861082844, -118.154027259229 34.527320039393956, -118.15365520720269 34.527429379780635, -118.150635306375 65 34.524590832337 48, -118.150644182314 82
34.524313334147, -118.15033440972929 34.524307408942 22, -118.14850508010 56 34.522586027103 74, -118.14786405424 21 34.521983357228 365, -118.14301598797277 34.
51742182339587, -118.14300391540229 34.516887411009 35, -118.142436353702 82 34.516875548731 65, -118.132920427128 95 34.507919655412 6, -118.131514988540 2 34.507971058742 97, -118.123428576294 47 34.5
00365923139725, -118.12284845735702 34.498783702161 31, -118.122420538616 23 34.498784381407 44, -118.122779404886 77 34.498399398663 054, -118.122464904342 79 34.497537010677 036, -118.122743942493 22
34.497220917234 6, -118.122098284476 88 34.496830223965 37, -118.122096349408 34 34.496529557217 2, -118.120995885319 6 34.495083096980 31, -118.079704339167 39 34.495368133098 62, -118.079850853160 25 34.
480840427405 88, -118.079701778533 52 34.473579517385 176, -118.096728941806 15 34.473520016606 62, -118.096796561183 62 34.480676701266 776, -118.118628566047 56 34.480677896362 835, -118.118624111287 83
34.477046436738 72, -118.118235622968 34 34.477046815790 5, -118.117850370948 09 34.476425875279 766, -118.118660520234 63 34.476424611898 611, -118.118729707499 88 34.466169238143 47, -118.131644186933 23
34.466174605450 966, -118.140485411072 92 34.466422355307 934, -118.149620169869 48 34.466439080812 93, -118.149606781455 03 34.452009871133 23, -118.184708587521 02 34.451958343862 283, -118.184743510552
89 34.455701009850 02, -118.197194722989 42 34.455802461932 74, -118.197524559575 34 34.455524583551 88, -118.197261455379 18 34.455802353469 466, -118.197530946458 66 34.455671583806 37, -118.197563077 5
758 34.456309087600 43, -118.198102394624 34 34.455844288512 68, -118.206302312083 95 34.456135718729 94, -118.206608963338 21 34.462886925156 575, -118.207846364840 69 34.462864851995 25, -118.208004468
63117 34.463656896194 394, -118.207898670891 94 34.464038648775 73, -118.207992612127 16 34.466057465210 52, -118.219682261905 79 34.465710534571, -118.219666468632 8 34.462667538203 51, -118.237027101 9
7352 34.463398142499 884, -118.237020886515 05 34.467402818446 28, -118.245336415259 87 34.467322489050 14, -118.245031506698 34 34.478108005466 01, -118.246078309983 34.478112895286 55, -118.245978342 7
1934 34.481709432066 75, -118.257737606507 73 34.481836792646 89, -118.259918606053 43 34.495903101058 374, -118.255544672136 34.495883892046 82, -118.255672457437 07 34.510049171616 565, -118.255505428
81398 34.539292874673 706, -118.237882969511 273 34.538947952229 54, -118.22024775908258 34.539088978245 836, -118.220182089695 67 34.542752039858 854, -118.206986468049 1 34.542696626173 37, -118.207021
92802952 34.539012185669 77, -118.20261747920541 34.538989720769 29)))',
 'L.A. County Neighborhood (Current)',
 'acton',
 'Acton',
 'Acton L.A. County Neighborhood (Current)',
 '39.3391089485',
 'unincorporated-area',
 None,
 None,
 '-118.16981019229348',
 '34.497355239240846',
 'POINT(34.497355239240846 -118.16981019229348)']
```

B2. Description: I would need to find the main categories of venues, then after that the sub-categories and finally all the types of 'Food' venues. Once I had the food venues, I would need the Seafood, Sushi, Japanese, and Fish and Chips restaurants. Foursquare API calls would return unstructured data, not in rows and columns, and would need quite a bit of filtering and data wrangling.
Method: Foursquare API calls
Source: https://developer.foursquare.com/
Example:

```
4d4b7104d754a06370d81259 Arts & Entertainment
4d4b7105d754a06372d81259 College & University
4d4b7105d754a06373d81259 Event
4d4b7105d754a06374d81259 Food
4d4b7105d754a06376d81259 Nightlife Spot
4d4b7105d754a06377d81259 Outdoors & Recreation
4d4b7105d754a06375d81259 Professional & Other Places
4e67e38e036454776db1fb3a Residence
4d4b7105d754a06378d81259 Shop & Service
4d4b7105d754a06379d81259 Travel & Transport
```

2.2. Data Wrangling (cleaning)

THE CITY:

A1. Description: I needed to determine the largest coastal cities in the US. As described above, this was a visual process, and there was no need to create any documents, nor any data wrangling. It might have been necessary if there were more than just four coastal cities in the top ten largest cities in the US.
Method: Visually it is possible to narrow them down with data on Wikipedia
Data Wrangling (cleaning):
None was needed.

A2. Description: I needed to find the cities with their population sizes in millions.
Method: Website Scraping
Data Wrangling (cleaning):
A Wikipedia page could have more than one table; in this case, it had four. It was necessary to extract the correct table from Wikipedia. Once the table had been extracted, it was saved as .CSV file and converted to a pandas dataframe. I had to drop the unnecessary columns: '2019rank', 'State[c]', '2010Census', 'Change', '2016 land area', '2016 population density', 'Location', 'Unnamed: 9', 'Unnamed: 10', rename the heading '2019estimate' to 'Population', sort the 'Population' column values by descending order, rename row 'New York[d] ' to 'New York', and delete all rows which did not contain the cities .

A3. Description: I need to determine the average monthly highest-high and highest-low temperatures and then convert these to yearly averages. Although no data wrangling was necessary, there was quite a bit of work involved in creating my own structured data to work with. I would say there was a degree of 'data wrangling' necessary, but not within the Jupyter Notebook, but rather in the excel sheet that would eventually be saved as a .CVS file.
Method: Create .CSV files with temperature information, as no source data exists
Data Wrangling (cleaning):
Before I could compare the temperatures, it was necessary to understand how the seasons align between the US and South Africa. I had to manually create a .CSV file to compare seasons and convert the data to a dataframe. With the seasons aligned, I would know where to, for example, place the first temperature for the first month of spring. It would otherwise not have given me accurate, comparable data. I had to also manually create a .CSV file with the highest-highs per month for the cities I was comparing and convert it to a dataframe. In order for the graphs to work, I needed to set the index to the column Seasons. Finally, I had to repeat the process for the highest-lows.

THE NEIGHBORHOOD:

B1. Description: I need all the neighborhoods with their latitudes and longitudes.
Method: Downloading .Json data
Data Wrangling (cleaning):
I needed to download the .json file with !wget, open the file and save the contents in a dataframe – it was in dictionary format. Then I had to find the data keys and their value lengths from the .Json file to establish where the data was. I had to create an empty dataframe and populate it with information from the .Json file. I needed to convert the latitudes and longitudes to float64 in order for them to work on a folium map.

B2. Description: I need venue categories, then the 'food' sub-category, and lastly, Seafood venues per neighborhood with their latitudes and longitudes.
Method: Foursquare API calls
Data Wrangling (cleaning):
There was a substantial amount of data wrangling involved in this section. I had to create a Foursquare API and a URL: 'https://api.foursquare.com/v2/venues/categories?&client_id={}&client_secret={}&v={}'. I saved my secret client credential in client_id, client_secret, and version for calling when needed. I requested structure and the keys to find categories. I then had to find out what the main categories' names were. I had to get the SUB-categories of: 'Food' as a dictionary. Get the Food Venues for all neighborhoods in Los Angeles **within a radius of 1km**. Check how many venues were returned for each neighborhood. Find out how many unique food categories could be found from all the returned venues. I only wanted: Seafood Restaurants, Sushi Restaurants, Japanese Restaurants, Fish & Chips Shops. I got a list of all the food categories and used it to remove any venues which were not Seafood Restaurants, Sushi, Japanese or Fish & Chips Shops (i.e., remove all the generalized categories, like 'Churrascaria,' 'Breakfast Spot,' 'Café,' 'Cupcake Shop,' 'Pizza Place' and 'Burger Joint'). I had to delete the column called 'index' from the dataframe. Got the number of Seafood restaurants per neighborhood with One-Hot Encoding ("One-hot Encoding is a type of vector representation in which all of the elements in a vector are 0, except for one, which has 1 as its value, where 1 represents a boolean specifying a category of the element." Source: https://stackabuse.com/one-hot-encoding-in-python-with-pandas-and-scikit-learn/). I counted the venues of each category in each neighborhood. Grouped rows by neighborhood and calculated the average of the frequency of occurrence of each category. Printed each neighborhood along with the most common venues. Wrote a function to sort the venues in descending order. Created the new dataframe and displayed the top 4 venues for each neighborhood. Finally, I counted the venues sorted by the Seafood Restaurant column.

2.3. How will the data be used to solve the problem

THE CITY:

A1. Description: I need to determine the largest coastal cities in the US.
Method: Visually, it is possible to narrow them down with data on Wikipedia
How will the data be used to solve the problem:
Visually I ascertained that there were only four coastal cities in the top ten largest cities in the US. This knowledge would be used in narrowing down the data in the next section.

A2. Description: I need the population sizes of the cities.
Method: Website Scraping
How will the data be used to solve the problem:
I used this data to get the largest coastal city names and population sizes in millions, already narrowed down with the above process. I would use this data in a bar chart to visualize the data for better understanding. I created a point system to compare various features to assist in selecting a city that meets all the client's requirements. Points would be assigned for the largest population in descending order.

A3. Description: I need to determine the average monthly highest-high and highest-low temperatures (later, the averages would be converted to yearly).
Method: Create .CSV files with temperature information, as no source data exists
How will the data be used to solve the problem:
I used this data to get the average monthly highest-high and highest-low temperatures per city; there would be a table for each. I would then calculate the average highest-high and highest-low temperatures per year separately per city, again in individual tables. I would use this data to populate an area chart and bar chart, to better understand the data visually. These two tables would be part of the point system to compare various features in order to select a city, that meets all the client's requirements. Points would be assigned for the average highest-high and highest-low temperatures per year per city. Each of the two features would get assigned points.

THE NEIGHBORHOOD:

B1. Description: I need all the neighborhoods with their latitudes and longitudes.
Method: Downloading .Json data
How will the data be used to solve the problem:
I used the .Json data to get a list of all 272 neighborhoods with their latitudes and longitudes in Los Angeles. This data would be used in a folium map to visualize the neighborhoods around the city. The dataframe would later be joined with the Foursquare table with the Seafood, Sushi, Japanese and Fish & Chips Restaurant venues, and their latitudes and longitudes. This information is crucial and will be used in part to determine the chosen neighborhood. It is important to remember that the client's primary area of focus is Seafood Restaurants.

B2. Description: I need venue categories, then the 'food' sub-category, and lastly venues per neighborhood with their latitudes and longitudes.
Method: Foursquare API calls
How will the data be used to solve the problem
I used Foursquare API calls to get all the Seafood Restaurants, Sushi, Japanese, and Fish & Chips Shops venues with eaches latitudes and longitudes in Los Angeles. The latitudes and longitudes are also crucial for calculating the geometry points, which would be used to create a folium cluster map with information in each cluster on the venue information. The dataframe will later be joined with the neighborhood table above to include this data. This will be used to run k-Means Clustering (Unsupervised Machine Learning Algorithm) to find clusters where each of the following is the most common venue: Seafood Restaurants, Sushi Restaurants, Japanese Restaurants, and Fish & Chips Shops. Ultimately, I need the cluster with the most significant number of Seafood Restaurants and to then derive the neighborhood that wins the Battle of the neighborhoods from that.

3. Methodology section

3.1. Exploratory data analysis

<u>THE CITY</u>:

Once I determined the **largest cities with their populations**, the cities were sorted in descending order by population from high to low. As there is a point system, the bar chart helped me explore/visualize the data, see Figure 1.

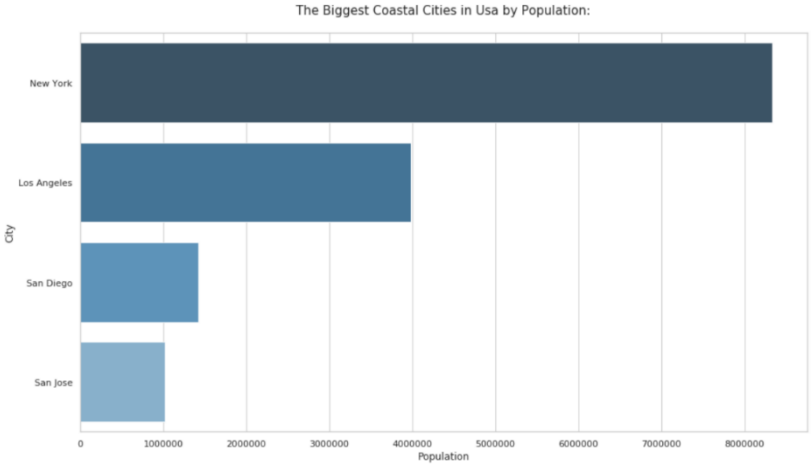


Figure 1

Then I needed the **Highest-HIGH temperatures throughout the year** per city compared with each other in order to determine which city was the best in this regard. First, however, I needed to align the seasons to compare the temperatures appropriately. After that, an area plot was generated to help explore/visualize the data, see figure 2. Although the area plot was interesting and some understanding could be derived from it, it still did not answer the question of how the cities compared on the basis of Highest-HIGH temperatures throughout the year per city. I had to calculate the yearly average and then compare the cities. I used a bar chart to explore/visualize the data, see figure 3.

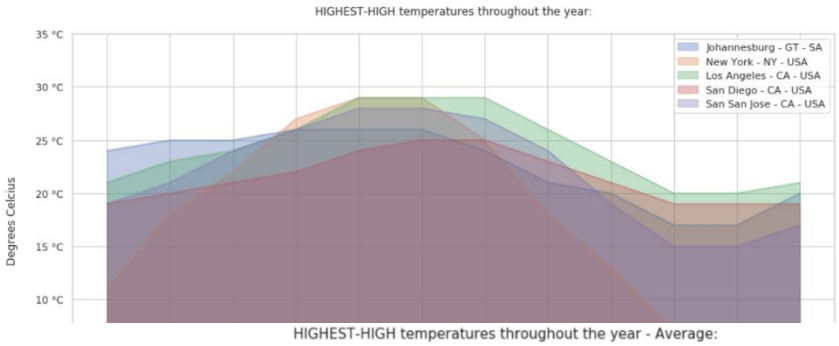


Figure 2

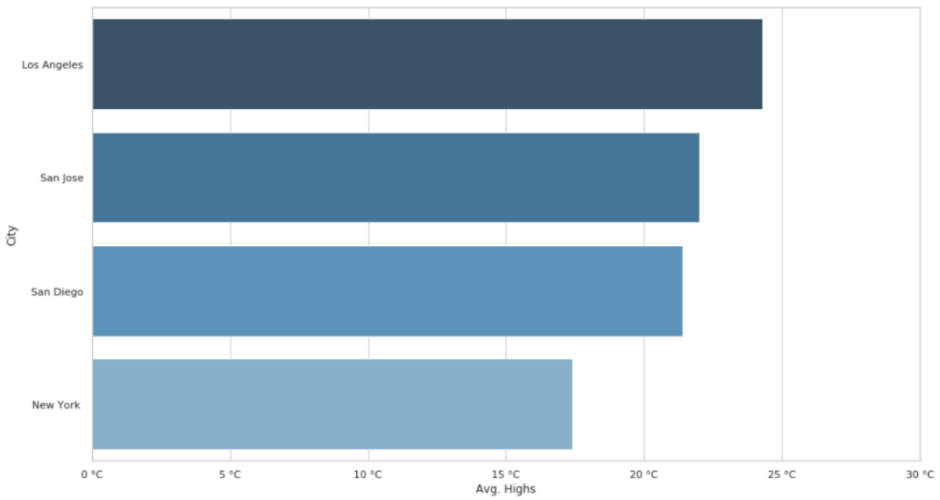


Figure 3

Then I needed the **Highest-Low temperatures throughout the year** per city compared with each other in order to determine which city was the best in this regard. First, however, I needed to align the seasons in order to compare the temperatures correctly. After that, an area plot was generated to help explore/visualize the data, see figure 4. Although the area plot was thought-provoking and some understanding could be derived from it, it still did not answer the question of how the cities compared on the basis of Highest-LOW temperatures throughout the year per city. I had to calculate the yearly average and then compare the cities. I used a bar chart to explore/visualize the data, see figure 5.

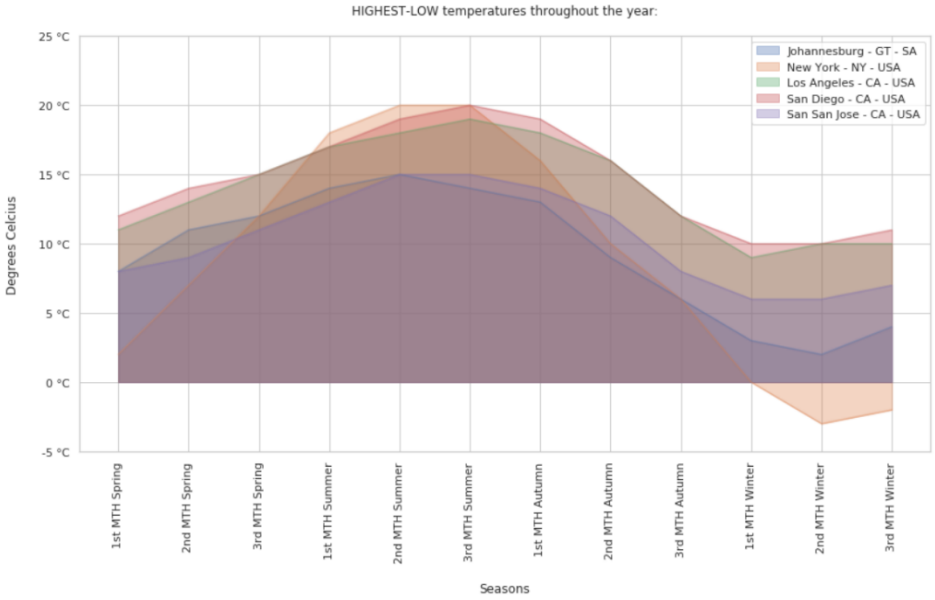


Figure 4

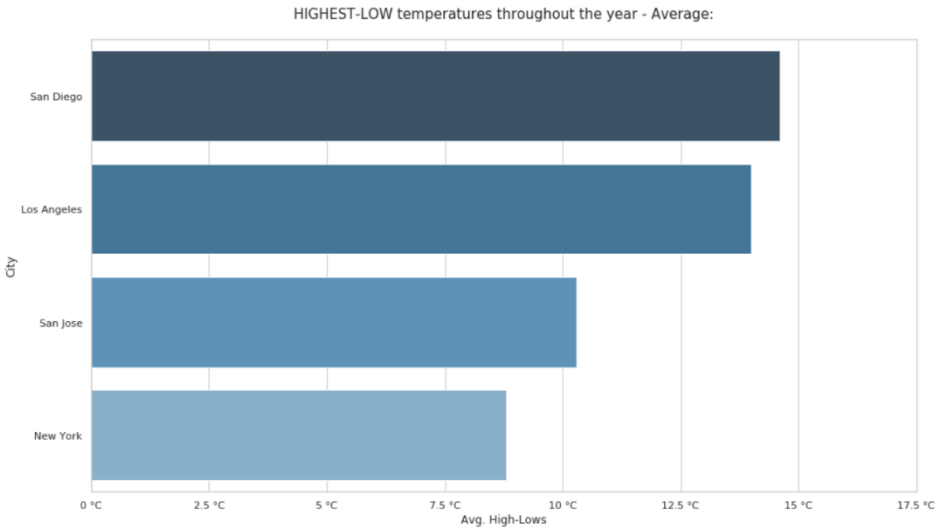
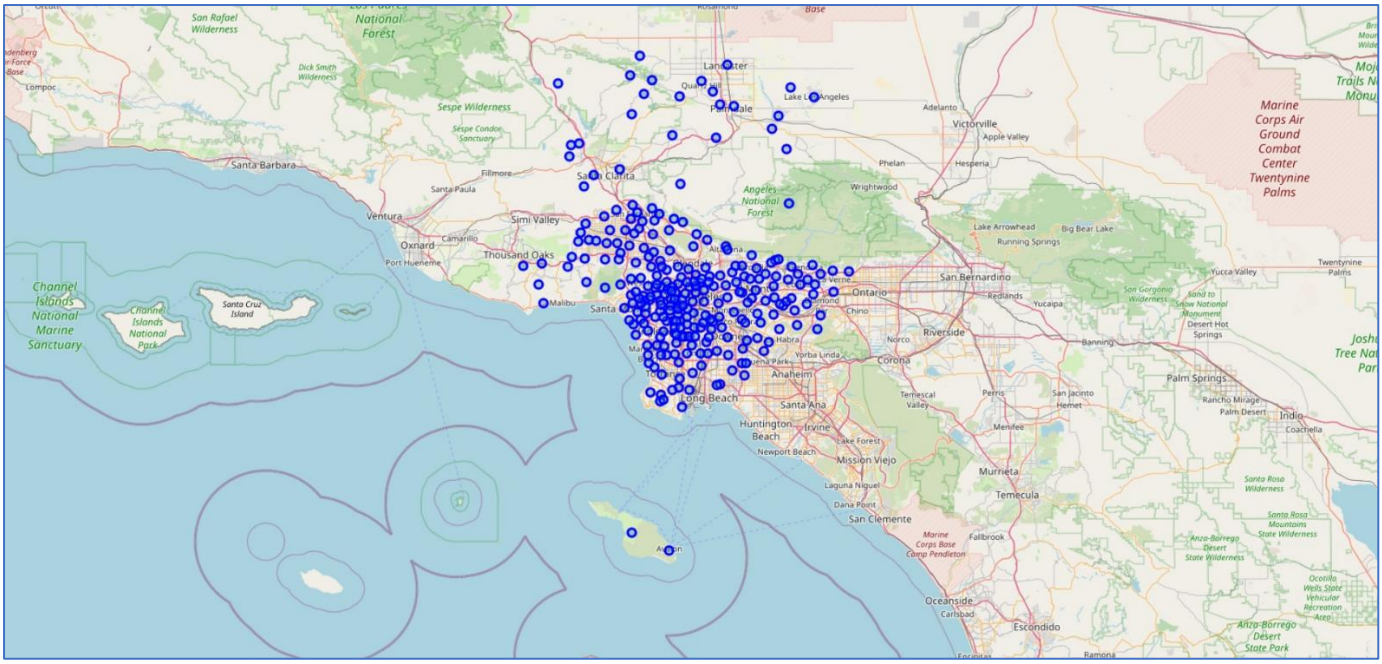


Figure 5

THE NEIGHBORHOOD:

First, I wanted to get an idea of the neighborhoods. I created an interactive folium map of Los Angeles, using the original . Json file (converted to a pandas dataframe), with its latitude and longitude values to visualize the Neighborhoods, see Map 1.



Map 1

With the data obtained with the Foursquare API, I could now visualize the four types of restaurants, and numbers of them, sorted by neighborhood, see figure 6. Although one could argue that this has already answered the business problem, this would not entirely true. It is essential to visualize clusters of Seafood Restaurants in order to make observations and recommendations. The client also wished to see cluster sizes on an interactive map.

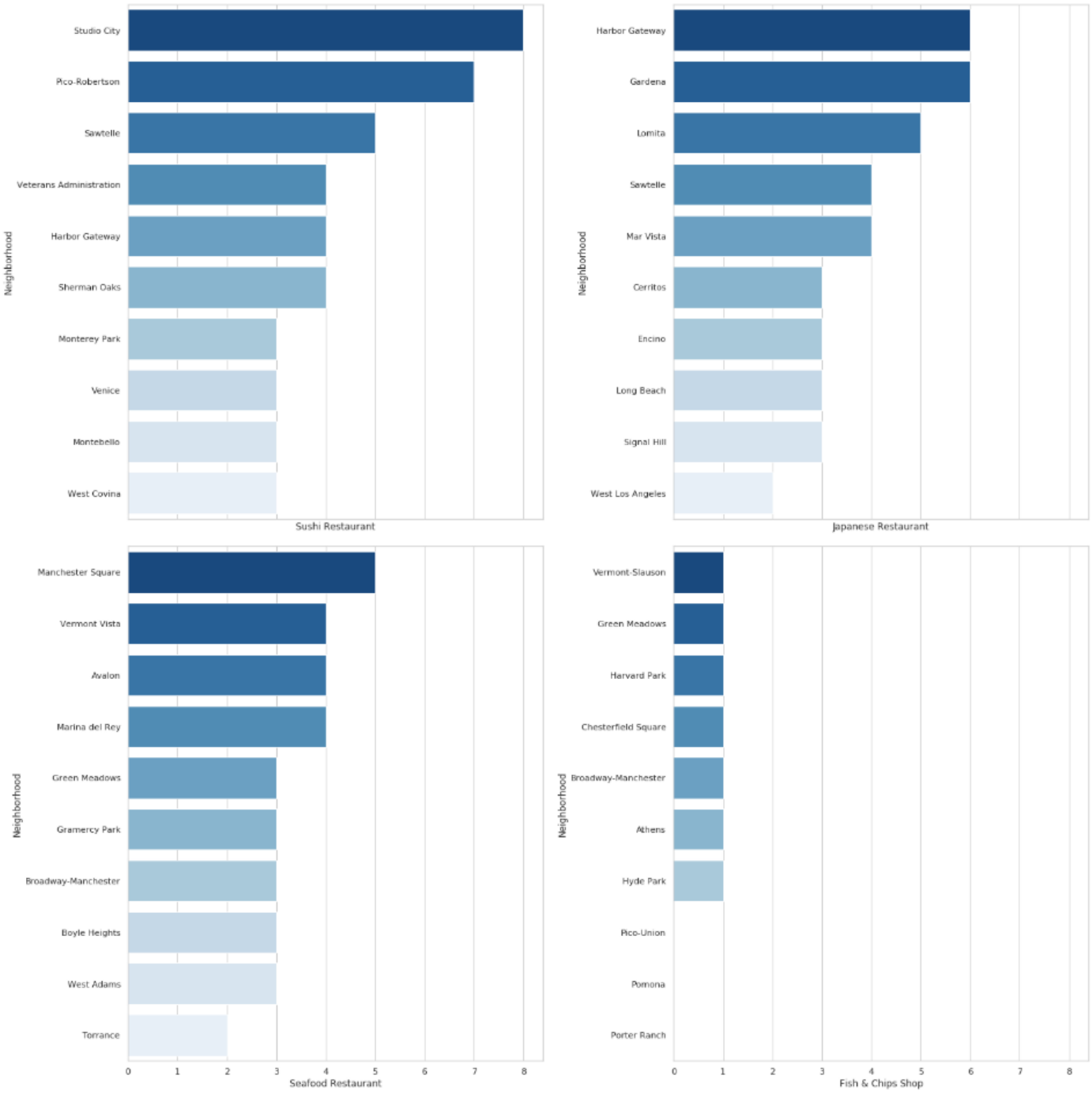


Figure 6, above

3.2. Inferential statistical testing

THE NEIGHBORHOOD:

I would be using kMeans Clustering (Unsupervised Machine Learning Algorithm) to create clusters and ultimately decide on the neighborhood that triumphs. However, before that, I had to find the optimal number of clusters (K's). I ran three statistical testing methods to select the optimum number of K's positively. I used The Elbow method, The Silhouette Method, and lastly, Gap Statistics. A description of each will follow:

The Elbow Method:
The Elbow method is an analytical approach used in determining the number of clusters in a data set in cluster analysis. The method consists of plotting the variation against the number of clusters. Picking the elbow of the curve as the number of clusters to use can be tricky. Basically, it is the area where the angle levels off.
The relationship is graphed between the number of clusters and WCSS (the sum of the centroid and its squared distance between each member of the cluster). It seemed to me that the curve started to level off between 3 and 11. Therefore, I was not sure which K to choose, see figure 7. How can I be sure that the number of clusters I have chosen is CORRECT? In the following section, I aim to answer the question.

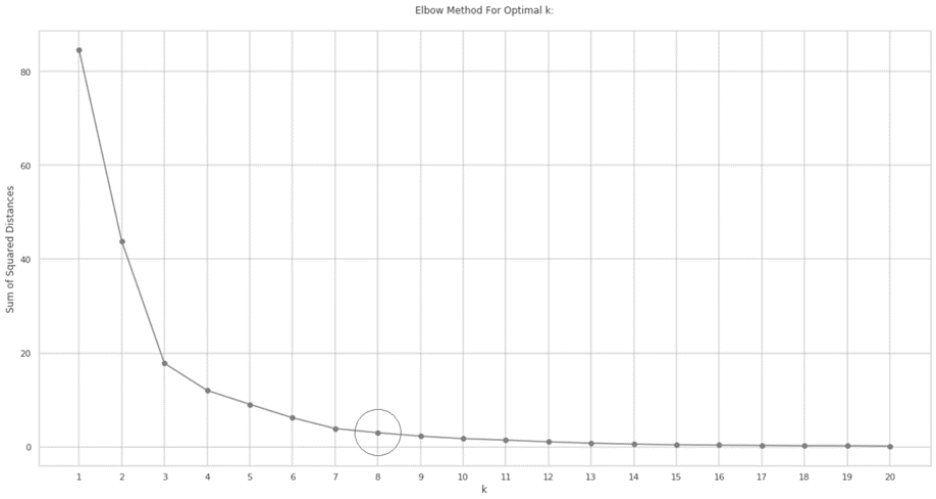


Figure 7

The Silhouette Method:
This method measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation) - The Silhouette Method is secondary to the Elbow method.
"Briefly, it measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering. Thee Silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximizes the average silhouette over a range of possible values for k (Kaufman and Rousseeuw 1990)", source: https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/
In the graph, figure 8, it peaks at 3, but this will give me very broad clustering. The next peak is at k = 8, which can be seen on the graph. At this stage, I am relatively confident that the optimum number of clusters is 8! That said, I felt I still needed further confirmation.

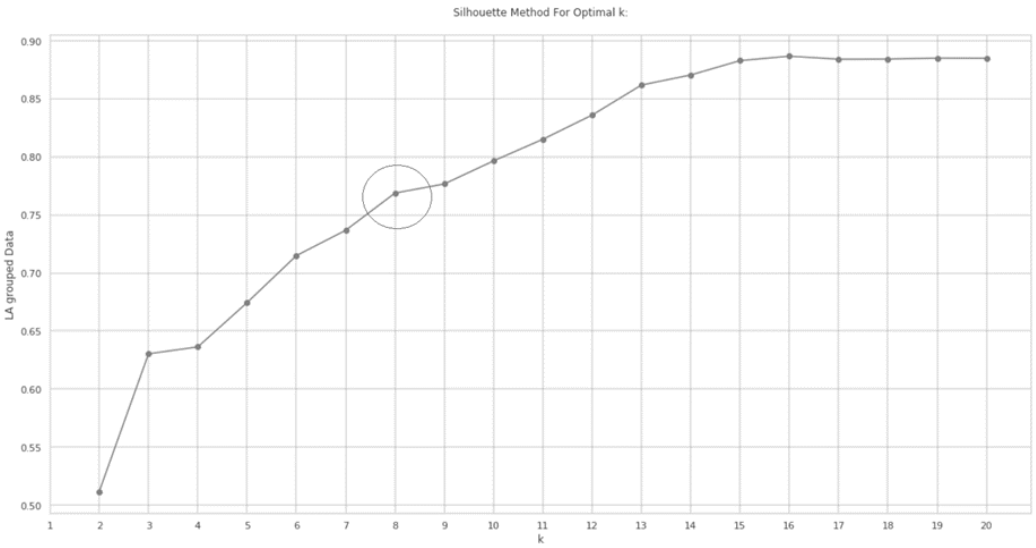


Figure 8

Gap Statistics:
Gap Statistics compares the clusters inertia (how far away points are from their centroid). The optimal choice of K is at a peak in the graph, see figure 9. A large gap illustrates that the clustering structure is very far away from the random unvarying distribution of points. The GAP (distance from each) is maximum for K=8! **Therefore, another confirmation that k=8 is optimal!**

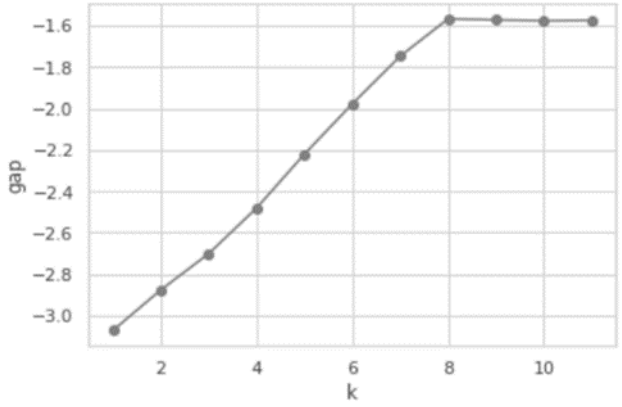


Figure 9

I was still concerned that the Elbow Method's K=3 and the Silhouette Method peak at K=3. It seemed that perhaps, I was missing something. I thought it best to run the kMeans Clustering algorithm on k=3 for further investigative purposes. I was surprised at how random the clustering was and how inconclusive the results were. I had a fear that I might be overfitting the data, but I was pleasantly surprised by this trial and its confirmation. I now had full confidence to continue with the choice of K=8.

I decided to graph the clusters using k=8 to understand the clusters, see figure 10 visually representing each K. Having scrutinized figure 10, it was clear to me that at k=8 I had the best results.

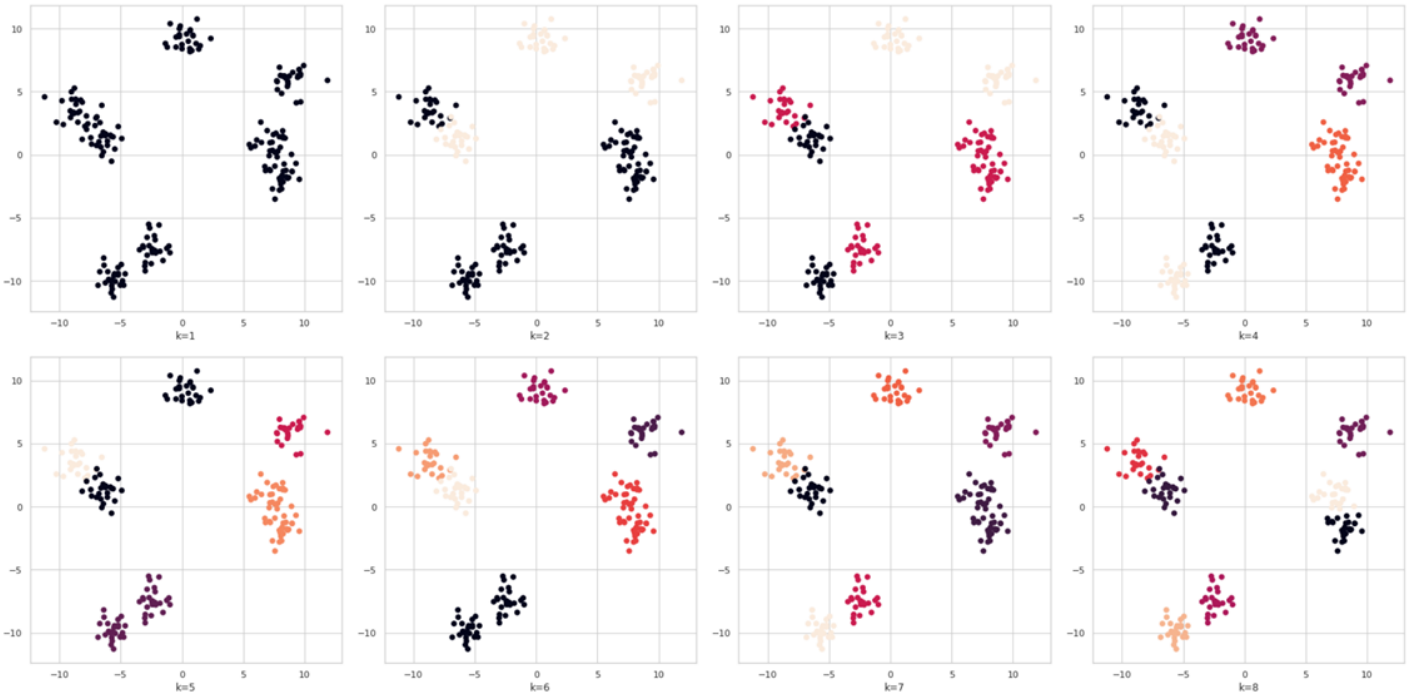
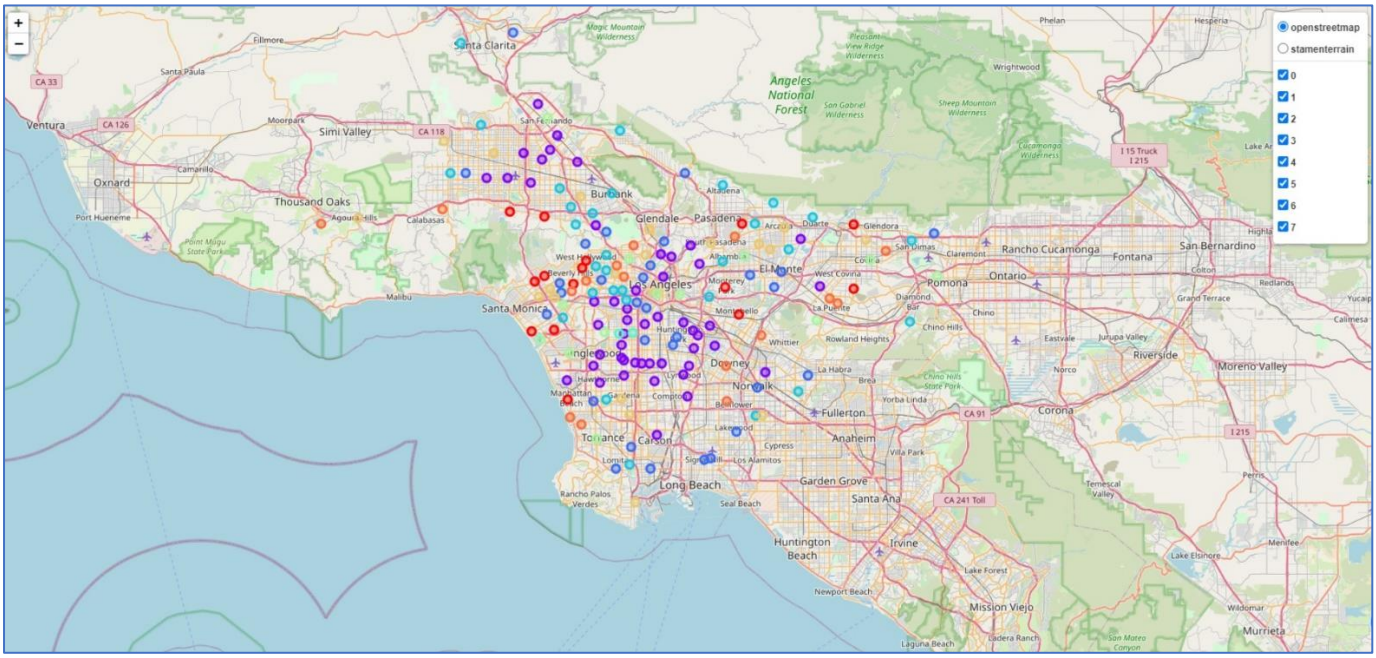


Figure 10

3.3. Machine learnings - kMeans Clustering

(Unsupervised Machine Learning Algorithm)

I ran the K-means Clustering Algorithm K=8 with a 1km radius, established in the Inferential Statistical Testing section above. The first thing I wanted to do was to plot the cluster on a map, see Map 2, to observe the distribution of the clusters and how they compared to one another. This could be very useful to make observations and make recommendations. I created a new dataframe after Data Wrangling (cleaning) was done to include the clusters:



Map 2

Next, I scrutinized the clusters (1km radius), below is a description of the cluster results:

Cluster 0:

1st Most Common Venue: Japanese Restaurant = 30

2nd Most Common Venue: Sushi Restaurant = 26

Cluster 1:

1st Most Common Venue: Sushi Restaurant = 24

2nd Most Common Venue: Seafood Restaurant = 24

Cluster 2:

1st Most Common Venue: Seafood Restaurant = 47

2nd Most Common Venue: Sushi Restaurant = 47

Cluster 3:

1st Most Common Venue: Sushi Restaurant = 34

2nd Most Common Venue: Seafood Restaurant = 30

Cluster 4:

1st Most Common Venue: Sushi Restaurant = 14

2nd Most Common Venue: Japanese Restaurant = 14

Cluster 5:
1st Most Common Venue: Fish & Chips Shop = 2

2nd Most Common Venue: Sushi Restaurant = 2

Cluster 6:
1st Most Common Venue: Seafood Restaurant = 5

2nd Most Common Venue: Fish & Chips Shop = 5

Cluster 7:
1st Most Common Venue: Seafood Restaurant = 12

2nd Most Common Venue: Japanese Restaurant = 11

Immediately Cluster 2 had my attention, as it had as its first most common venue Seafood Restaurants, with a number of 47. This was the highest in all the clusters (all neighborhoods within a 1km radius). Interestingly enough, it also had as the second most common venue, Sushi Restaurants with a similarity of 47. Could this mean something? I think so! I believe that kMeans Clustering was the correct approach to take as it classified data precisely, without having to first be trained with labeled data.

4. Results Section

THE CITY:

The client has requested a coastal city in the US, for fast access to fresh seafood produce. He then wants a city with the highest possible population to give his new Seafood Restaurant chain the best possible chance. As the client is from Johannesburg South Africa, he is very mindful of climate, as he is used to comfortable temperatures. It has to be expressly noted that he requested that the cities be compared, in terms of the highest-high and highest-low temperatures throughout the year.

As described above, I created a point system to equally compare the features in a city; this was exactly what the client wanted. Points would be assigned on the highest population, average highest-high temperatures per year and average highest-low temperatures per year. I was pleased to have found my own system, though very simple in design, to achieve a result the client would be pleased with, see figure 11. The CITY, which received the most points was: **Los Angeles, California!**

	City	Population - Points	Avg. Highs - Points	Avg. Highs-Lows - Points	TOTAL - POINTS
0	Los Angeles	3	4	3	10
2	San Diego	2	2	4	8
1	New York	4	1	1	6
3	San Jose	1	3	2	6

Figure 11

THE NEIGHBORHOOD:

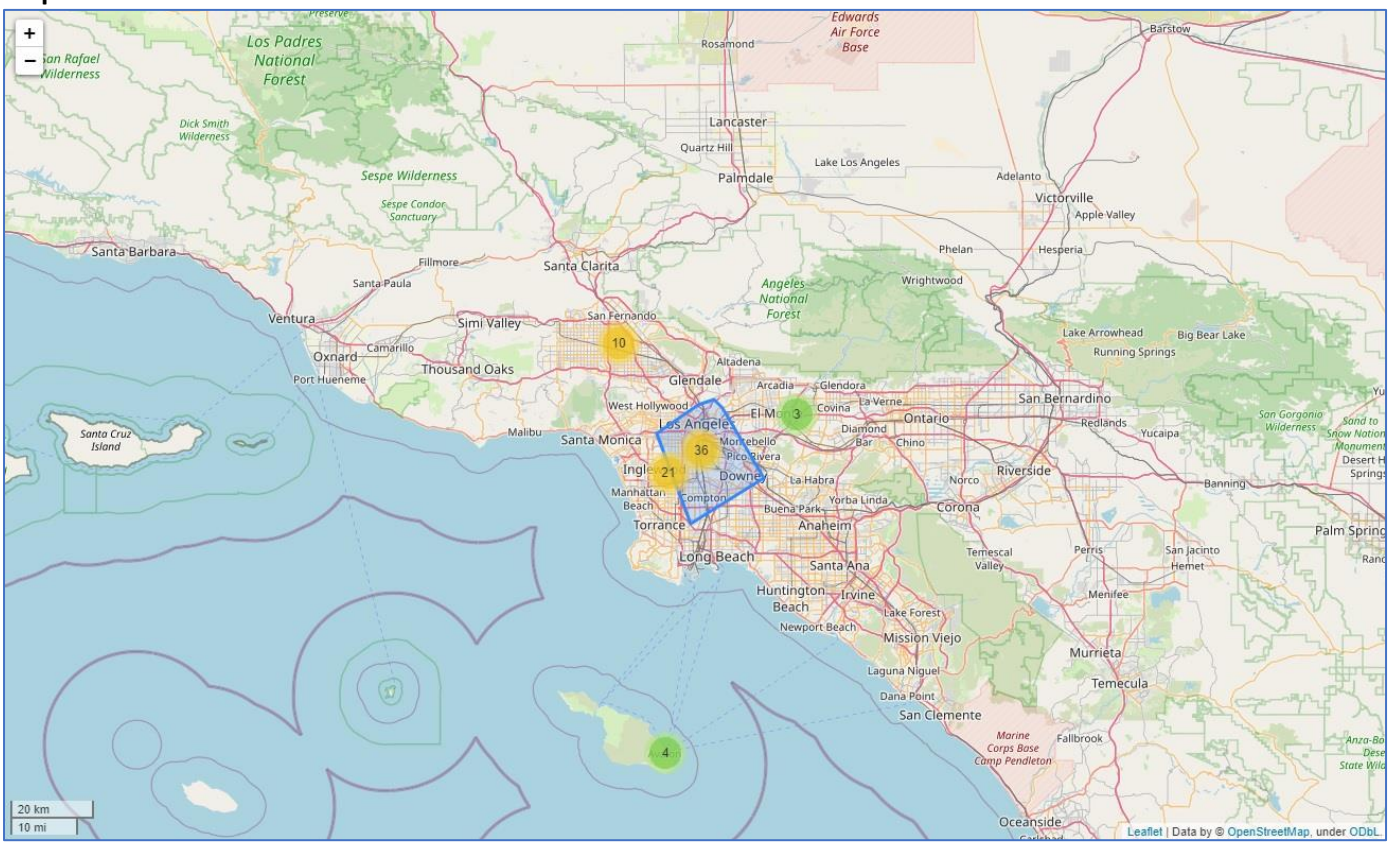
As mentioned, the client then wants to know the best cluster of neighborhoods in the city to start the chain of Seafood restaurants. He could decide to place the chain of restaurants in the clusters of neighborhoods with high counts of seafood restaurants. He would thus effectively be taking on the competition head-on. Alternatively, he could place the restaurants in strategic positions away from the clusters of neighborhoods with high counts of seafood restaurants, effectively to avoid the competition, yet be in the correct general location. It is vital for him to be able to visualize the neighborhoods on an interactive map with the clusters to make sound business decisions.

With the client's requirements in mind, Cluster 2 immediately had my attention, as it had as its first most common venue, Seafood Restaurants with a number of 47, the highest in all the clusters. Interestingly enough, it also had as the second most common venue, Sushi Restaurant with a number of also 47. Could this mean something? I think so!

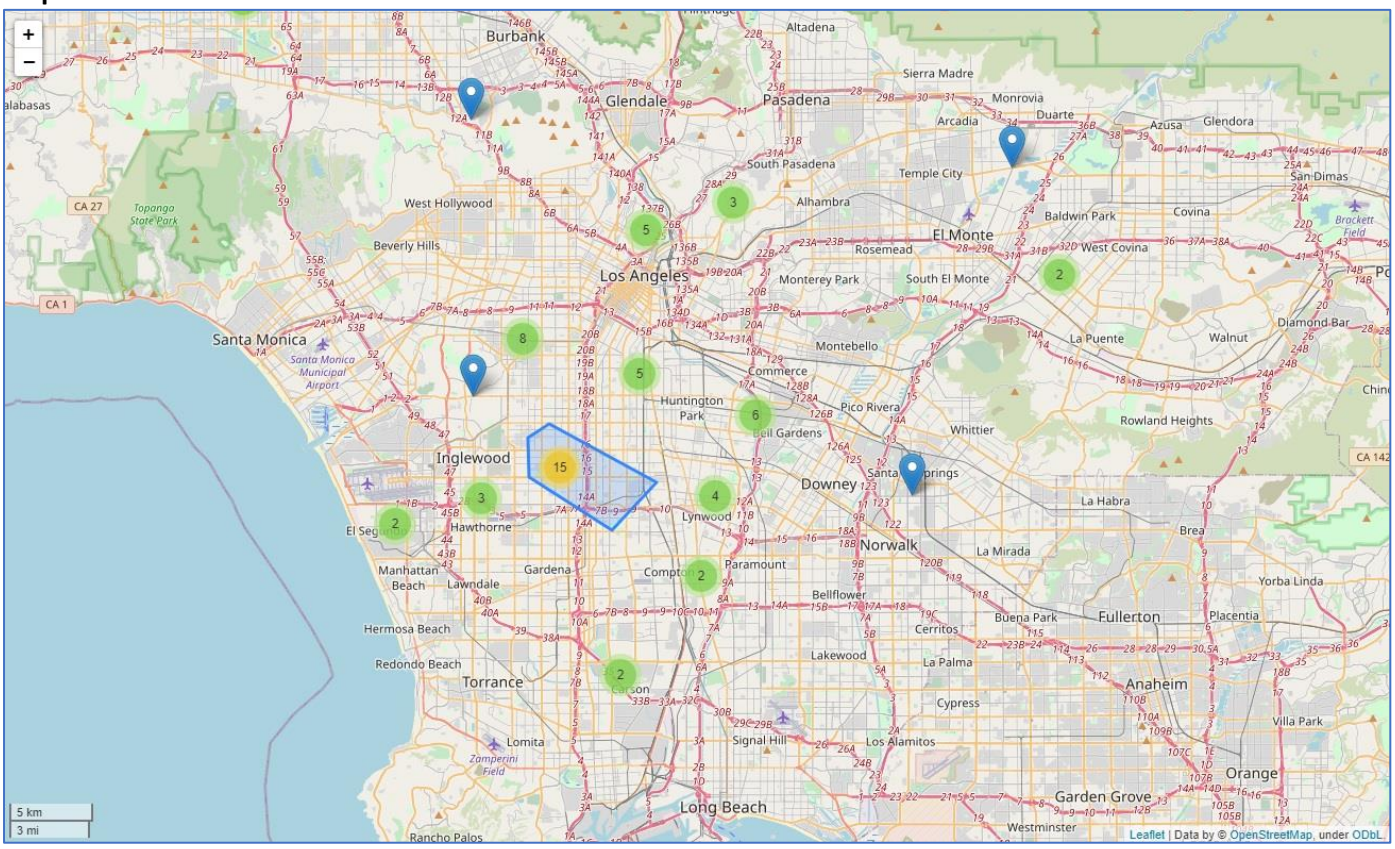
I used geopandas to transform longitudes and latitudes into a list of shapely point objects and set it as a geometry. I then created a GeoDataFrame to view the clusters. It would be necessary to start with the most significant cluster, then zoom in to the next and so on, to eventually derive at the best neighborhood see Map 3 to 5. The results were based on a 1km radius. After having done this, the neighborhood that won Battle of the Neighborhoods, with 5 Seafood Restaurants, was: **Manchester Square, Los Angeles**!

Vermont Vista, Avalon, and Marina del Rey had 4 restaurants each within a 1km radius. I believe this is noteworthy and that the client should be aware of this. If Manchester Square is not to be a viable option for the client, then these 3 neighborhoods offer him alternative options.

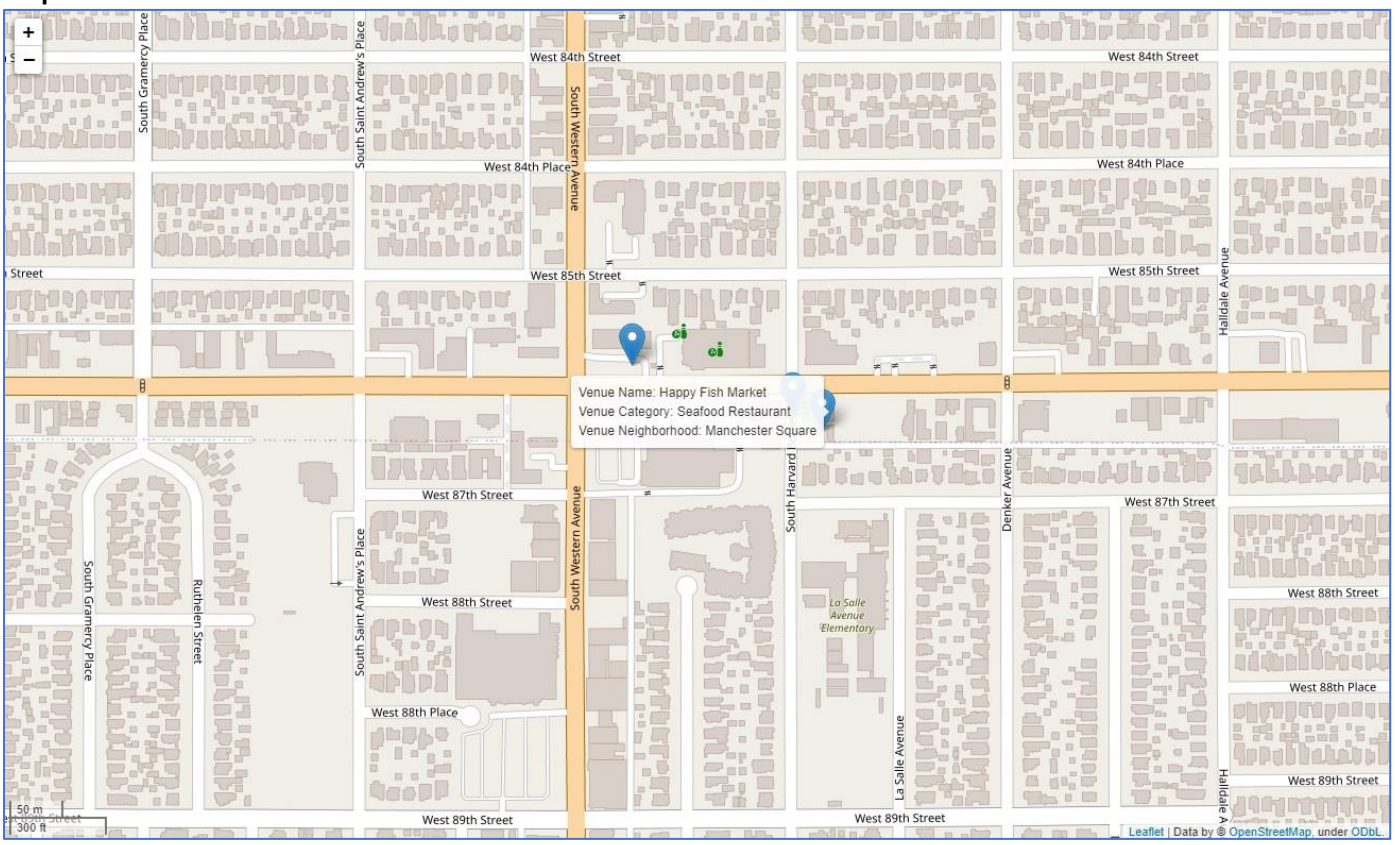
Map 3


Los Padres National Forest
Edwards Air Force Base
Rosamond
Barstow
Lancaster
Quartz Hill
Palmdale
Lake Los Angeles
Adelanto
Victorville
Apple Valley
Hesperia
Phelan
Wrightwood
Angeles National Forest
Sespe Wilderness
Dick Smith Wilderness
Santa Barbara
Fillmore
Santa Paula
Santa Clarita
Ventura
Simi Valley
San Fernando
Oxnard
Camarillo
Thousand Oaks
Port Hueneme
Santa Cruz Island
Malibu
Santa Monica
Altadena
Glendale
Arcadia
Glendora
West Hollywood
El Monte
La Verne
Covina
Diamond Bar
Ontario
Chino
San Bernardino
Redlands
Yucaipa
Lake Arrowhead
Big Bear Lake
Running Springs
Los Angeles
Montebello
Pico Rivera
Inglewood
Downey
La Habra
Norco
Riverside
Manhattan Beach
Compton
Yorba Linda
Buena Park
Corona
Torrance
Anaheim
Long Beach
Santa Ana
Temescal Valley
Perris
San Jacinto
Hemet
Huntington Beach
Irvine
Menifee
Lake Forest
Newport Beach
Mission Viejo
Murrieta
Laguna Niguel
Temecula
Dana Point
San Clemente
Marine Corps Base Camp Pendleton
Fallbrook
Oceanside
10
3
36
21
4
20 km
10 mi
Leaflet | Data by © OpenStreetMap, under ODbL

Map 4


Burbank
Altadena
Sierra Madre
Glendale
Pasadena
Monrovia
Arcadia
Duarte
Azusa
Glendora
South Pasadena
Temple City
San Dimas
CA 27
Topanga State Park
West Hollywood
Alhambra
Baldwin Park
Covina
Beverly Hills
El Monte
Rosemead
West Covina
Los Angeles
Monterey Park
South El Monte
CA 1
Santa Monica
Santa Monica Municipal Airport
Montebello
La Puente
Walnut
Diamond Bar
Commerce
Huntington Park
Pico Rivera
Rowland Heights
Whittier
Inglewood
Santa Fe Springs
Downey
La Habra
Lynwood
Norwalk
La Mirada
El Segundo
Hawthorne
Manhattan Beach
Gardena
Paramount
Bellflower
Yorba Linda
Lawndale
Buena Park
Fullerton
Placentia
Hermosa Beach
Cerritos
Redondo Beach
Lakewood
La Palma
Torrance
Anaheim
Cypress
Villa Park
Lomita
Signal Hill
Los Alamitos
Orange
Garden Grove
Long Beach
Westminster
Rancho Palos
15
8
6
5
4
3
2
5 km
3 mi
Leaflet | Data by © OpenStreetMap, under ODbL

Map 5


Venue Name: Happy Fish Market
Venue Category: Seafood Restaurant
Venue Neighborhood: Manchester Square
West 84th Street
West 84th Place
West 85th Street
West 87th Street
West 88th Street
West 88th Place
West 89th Street
South Gramercy Place
South Saint Andrew's Place
South Western Avenue
South Harvard
Denker Avenue
Halldale Avenue
La Salle Avenue
Ruthelen Street
La Salle Avenue Elementary
50 m
300 ft
Leaflet | Data by © OpenStreetMap, under ODbL.

5. Discussion Section

5.1. Observations

I observed that the most extensive grouping of clusters is mostly south/southwest of the Los Angeles city center. These appeared to be roughly in the middle of 5 airports, Manchester Square; the winning neighborhood is a 10-minute drive from LAX.
I also noticed that the clusters are surrounded by a few of the most expensive neighborhoods in Los Angeles, e.g., Beverley Hills, Long Beach, Palo Verdes, Redondo Beach, and Hermosa Beach. I had noted that Cluster 2 had as its first most common venue Seafood Restaurants totaling 47, and the second most common venue was Sushi Restaurants, also totaling 47. I asked the question, "Could this mean something? I think so!". When looking at the observations above, knowing that the southern neighborhoods have a few of the most expensive neighborhoods surrounding them, it made sense that Sushi restaurants would be grouped here as well (as sushi is generally costly). Vermont Vista, Avalon, and Marina del Rey had 4 restaurants each within a 1km radius. I believe this is noteworthy too.

5.2. Recommendations

I would recommend making sure that the Seafood restaurants or chain of restaurants are located to the south/southwest of the Los Angeles city center. I noted that at least 5 airports, including LAX, surrounded the clusters; therefore, I would endorse staying within this "airport outer border" area. The clusters are surrounded by some of the most expensive neighborhoods (especially to the north of the Los Angeles city center), and I observed that generally, more expensive food types were present in these cluster surrounds. It would thus be recommended that the restaurants could benefit from being designed to have an exclusive ambiance. I would also suggest that if Manchester Square was for any reason not viable, that Vermont Vista, Avalon, or Marina del Rey offer feasible alternatives.

6. Conclusion Section

FIRST SECTION - THE CITY:

In conclusion, the client then wants to know the best cluster of neighborhoods in the city to start the chain of Seafood restaurants. He could decide to place the chain of restaurants in the clusters of neighborhoods with high counts of seafood restaurants. He would thus effectively be taking on the competition head-on. Alternatively, he could place the restaurants in strategic positions away from the clusters of neighborhoods with high counts of seafood restaurants, effectively to avoid the competition, yet be in the correct general location. It is vital for him to be able to visualize the neighborhoods on an interactive map with the clusters to make sound business decisions.

A point system was used to derive the city, **Los Angeles, California**, by equally comparing the features: highest population, average highest-high temperatures per year, and average highest-low temperatures per year.

I believe that the business problem was wholly addressed with the results in this, the first, of two sections. The business problem **answer**: **Los Angeles, California!**

SECOND SECTION - THE NEIGHBORHOOD:

Further concluding, the client wanted to know the best cluster of neighborhoods and, ultimately neighborhood, in the city to start the chain of Seafood restaurants.

I used kMeans Clustering, Unsupervised Machine Learning Algorithm, to answer this question. The most significant clusters were to the south of Los Angeles, and the neighborhood with 5 Seafood Restaurants, is: **Manchester Square, Los Angeles –** it was also noteworthy that Sushi restaurants were equally represented in the cluster. This could potentially add to the client's business strategy!

There were also additional feasible options with Vermont Vista, Avalon, and Marina del Rey (4 restaurants each within a 1km radius).

Additional information was made available; observations and recommendations. I am confident that the client's business problem was addressed with the results, and the supplementary information derived would be most beneficial to him. The business problem **answer**: **Manchester Square, Los Angeles!**

COMMENTS FROM THE DATA SCIENTIST: Armand van der Merwe

This was challenging, yet at the same time, a rewarding project for myself. In the process of doing the project, much was learned, and exciting information gathered. It was incredibly satisfying to be able to confidently answers the business problem and present the results that the client requested accurately. I look forward to future projects.

Thank you for perusing my report!

7. References

Source: https://simple.wikipedia.org/wiki/List_of_United_States_cities_by_population
Reason: A table with data on how US city names and populations in millions

Source: https://www.noaa.gov/
Reason: Temperature information

Source: https://usc.data.socrata.com/
Reason: Location of .Json file

Source: https://developer.foursquare.com/
Reason: Foursquare Developers page for registrations, to access their databases with API calls

Source: https://stackabuse.com/one-hot-encoding-in-python-with-pandas-and-scikit-learn/
Reason: An explanation of One-Hot-Encoding

Source: https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/
Reason: An explanation of the Silhouette Method

8. Acknowledgments

None

9. Appendix

None