Battle of the Bakers

Finding the best neighborhood in the Bay Area to open a bakery using Data Science.

Kristen Vrionis

2020

**Table of Contents**

[**1. Introduction** 1](#_Toc48309565)

[**2. The Data** 2](#_Toc48309566)

[**3. The Analysis** 6](#_Toc48309567)

[**4. Discussion:** 11](#_Toc48309568)

[**5. Conclusion** 11](#_Toc48309569)

**1. Introduction**

#### The Business Problem

The Bay Area is the one the most diverse and populous cities in the United States. It is diverse, wealthy, a global hub, and has many new enterprises being started every day. Because of these factors, and more, The Bay Area has attracted many individuals and businesses.

The Bay Area is also very competitive and expensive. With all the new enterprises being started, one needs to do a fair amount of due diligence and research before starting a venture of their own.  If a location is not picked strategically, there is a higher chance of failure due to factors such as increased competition, lack of customers, and high operating costs. My client, Cleo’s Cookies, is looking for a place to start a bakery in The Bay Area. I was hired to research the best location to open a new bakery within The Bay Area. The insights derived from analysis will give good understanding of the business environment which help in strategically targeting the market. This will help in reduction of risk and the increase the potential return on investment.

The Bay Area is famous for its excellent and diverse cuisine. It's food culture includes an array of international cuisines influenced by the city's immigrant history. Specifically, bakeries are popping up and offer a variety of baked goods. To survive in this competitive market, a business must plan and be strategic about its location. A few of the factors that will be considered are:

1. Bay Area Counties
2. Cities Population
3. Cities Demographics
4. Are there any venues (offices, gyms, schools) nearby where “floating population” is high?
5. Who are the competitors in that location?
6. What classifies as a bakery (café, dessert, ice cream, etc.)?

#### b. Target Audience

Cleo’s Cookies would like to know which neighborhood or area of The Bay Area would be the most optimal to start a new bakery in. Beyond my client, this project would be of interest to anyone in The Bay Area who is wanting to start their own bakery.

**2. The Data**

#### Data Sources

The data that will be required will be a combination of CSV files that have been prepared for the purposes of the analysis from multiple sources. County information, including which cities are part of each county and neighborhood demographics were found via Wikipedia ([here](https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_the_San_Francisco_Bay_Area)). The Geographical location of each town/neighborhood was determined via the Geolocator package and Venue data pertaining to bakery competitors was obtained via Foursquare. The Venue data will help find which neighborhood is best suitable to open a bakery in.

#### Geographical Location Data Using the Geocoder Package

The second source of data provided us with the Geographical coordinates of the neighborhoods with the respective longitudes and latitudes. However, in order to get an accurate result, data cleansing had to be performed first.

##### Data Cleansing

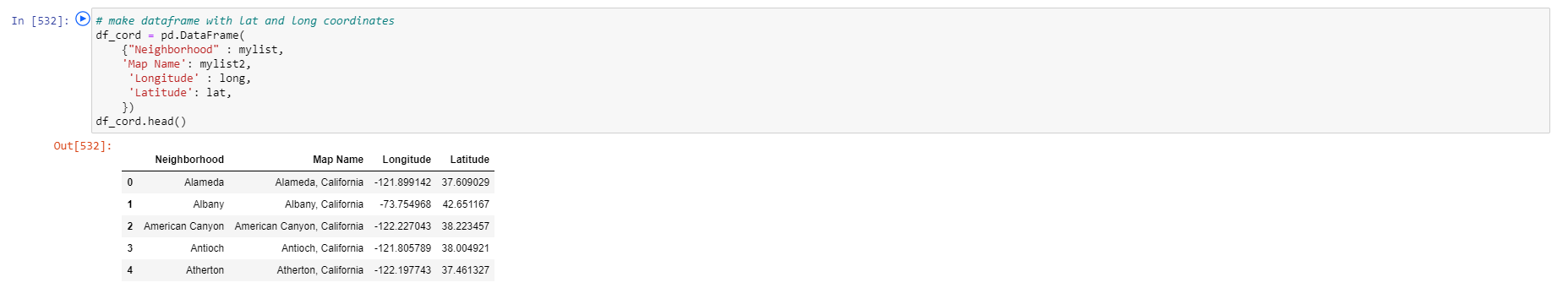
##### The data from the Wikipedia page only contained neighborhood names, which was a problem considering there are multiple cities throughout the world that contain the same name. In order to prevent inaccuracies in determine the longitude and latitude of the cities, I transformed the data and created a new column that would format the neighborhood name as “City, California”.

##### Geolocator

Now that the data was in the City, State format, I could use geolocator to determine the longitudes and latitudes of each city.

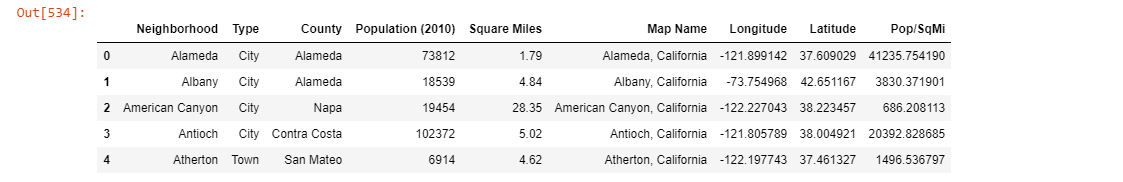


After all the data was collected and put into data frames, merging of the data was required to start the process of analysis. The coordinates information that was determined in the step above was converted into a data frame (df\_cord), shown below.



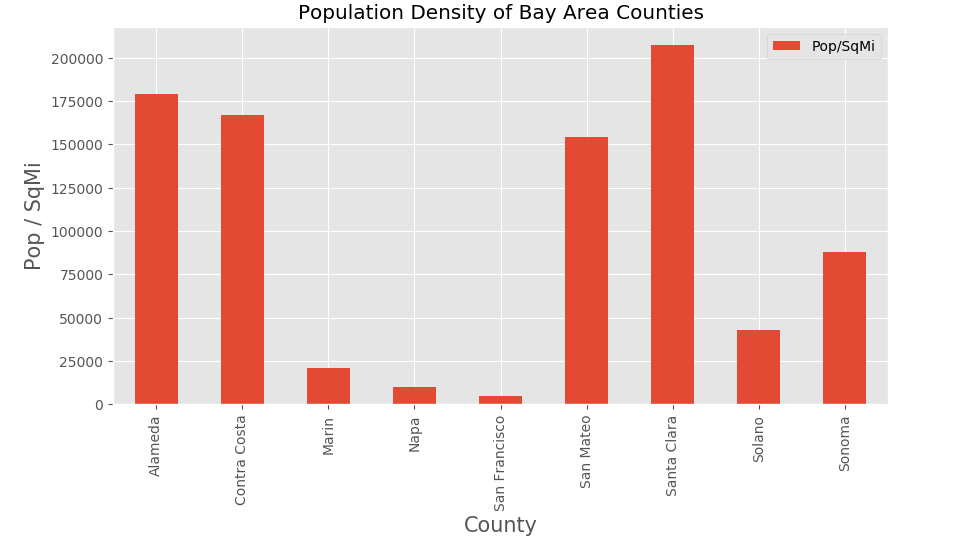
Next, df\_cord was merged with the original data set from Wikipedia. Now, all cities were listed along with their corresponding county, population, square footage, population, longitude, and latitude.





#### County Selection

After loading in the data, I realized that there was a total of 9 counties and 101 cities within the Bay Area. In order to narrow down my analysis, I decided to narrow down my analysis and pick three counties that have the some of the highest population density. In order to determine population density for each county, I first determined the population density on the neighborhood level. To do this, for each neighborhood I used the available data fields Population (as of 2010) and Square Miles. After the population density was calculated for each city, I created a bar chart (grouped by county) to help visualize the Population per Square Mile in each of the counties.



From the bar chart, I determined that the top five most densely populated counties were: Santa Clara, Alameda, Contra Costa, San Mateo, and Sonoma. Using knowledge about the area, I decided that the best counties to select for a new bakery, out of the top five, were Santa Clara, San Mateo, and Alameda. These three counties were selected to be used for the remainder of the analysis, all other counties were filtered out.

#### Venue Data from FourSquare API

##### Finding the Venues

After determining the counties in scope for the analysis, I obtained the venue data for each of the three counties using Foursquare API. Shown below is an example of obtaining the venue data for Santa Clara county as well as the Santa Clara county venue data frame.

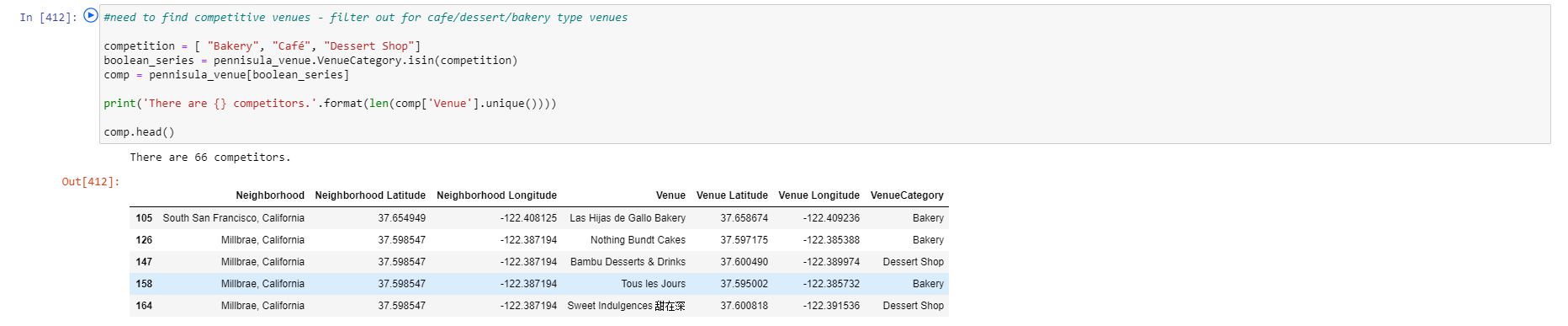




Each data frame had a total of 100 entries. I took the data frames for each county (three total) and merged them into one. After all the data frames were combined, I removed duplicates values to verify that all 300 data points were unique.

##### Identifying the Competition

There was a total of 245 unique venue categories that were returned. In order to identify who the competition was, the venue categories needed to be narrowed down. After listing the categories, I determined that the three categories that were related to Cleo’s Cookies were: Bakery, Dessert Shop, and Café.

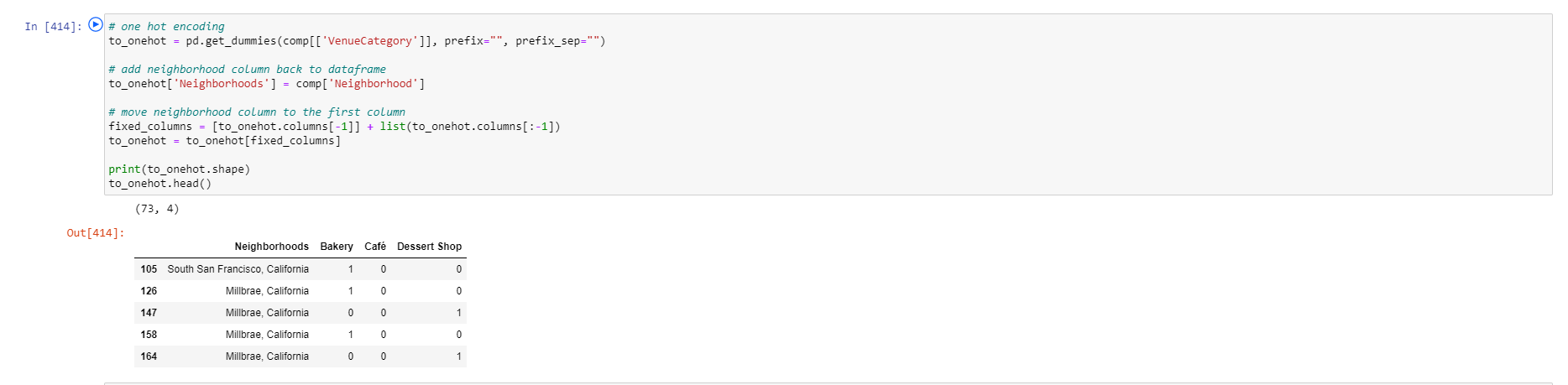


I filtered the for the desired venue categories and was left with 66 potential competitors for Cleo’s Cookies within the three selected counties. With the competitors identified, it was now time to explore the data and perform an analysis.

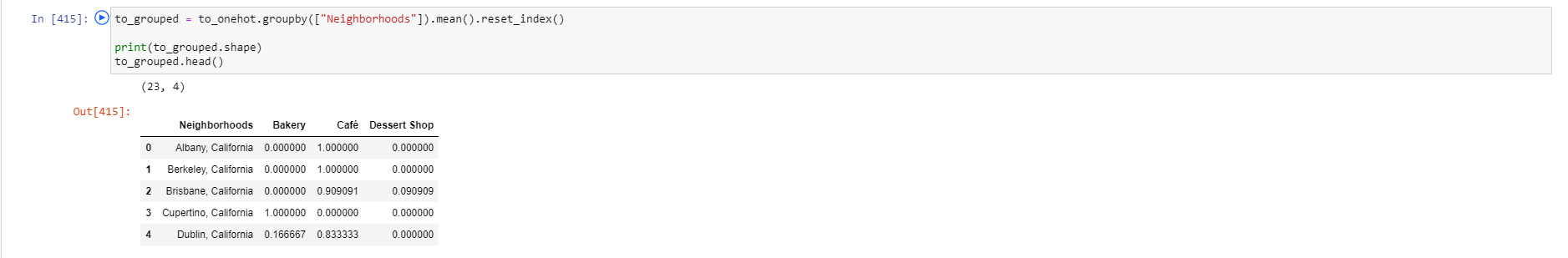
**3. The Analysis**

#### One Hot Encoding

To analyze the data, I first performed a technique called One Hot Encoding in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. For each of the cities, the venues were used to determine a “frequency” for each category that the venue belonged to.

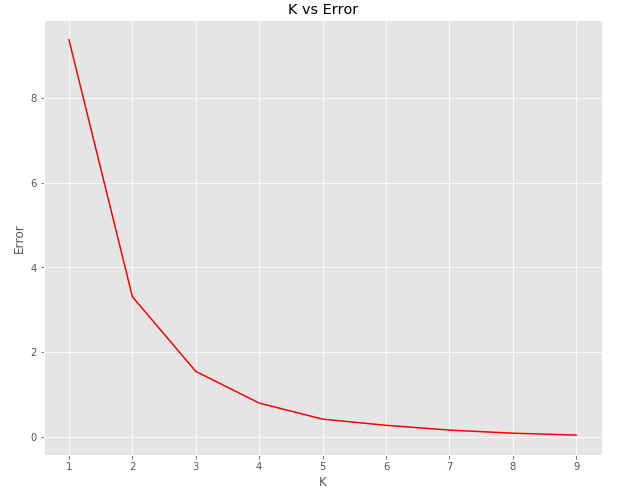


All venues were then grouped by neighborhood and the average frequency of occurrence of each Venue Category was calculated.

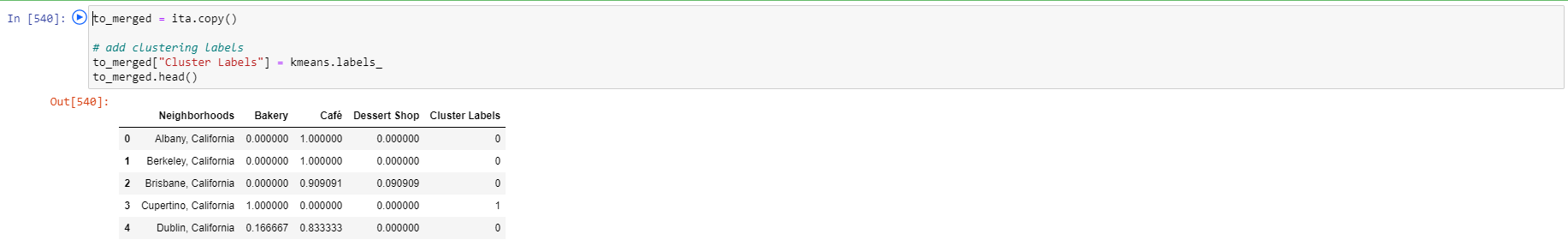


#### K-Means Clustering

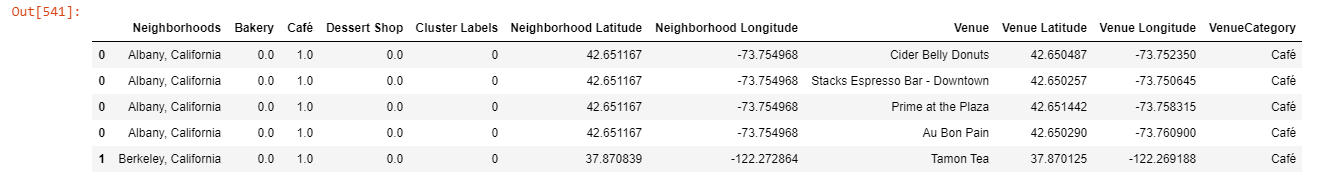
To make the analysis more interesting, we wanted to cluster the cities based on the cities that had similar averages of bakeries, cafes, and dessert shops. To do this we used K-Means clustering. To get the optimum K value, I used the Elbow Point Technique, which is performed by running a test with different number of K values. The best K value is where the line has the sharpest turn. K was determined to equal two, meaning that there will be two total clusters.



Cities that had a similar mean frequency of competitive venues were grouped into two clusters (note that cluster indexing starts at 0).

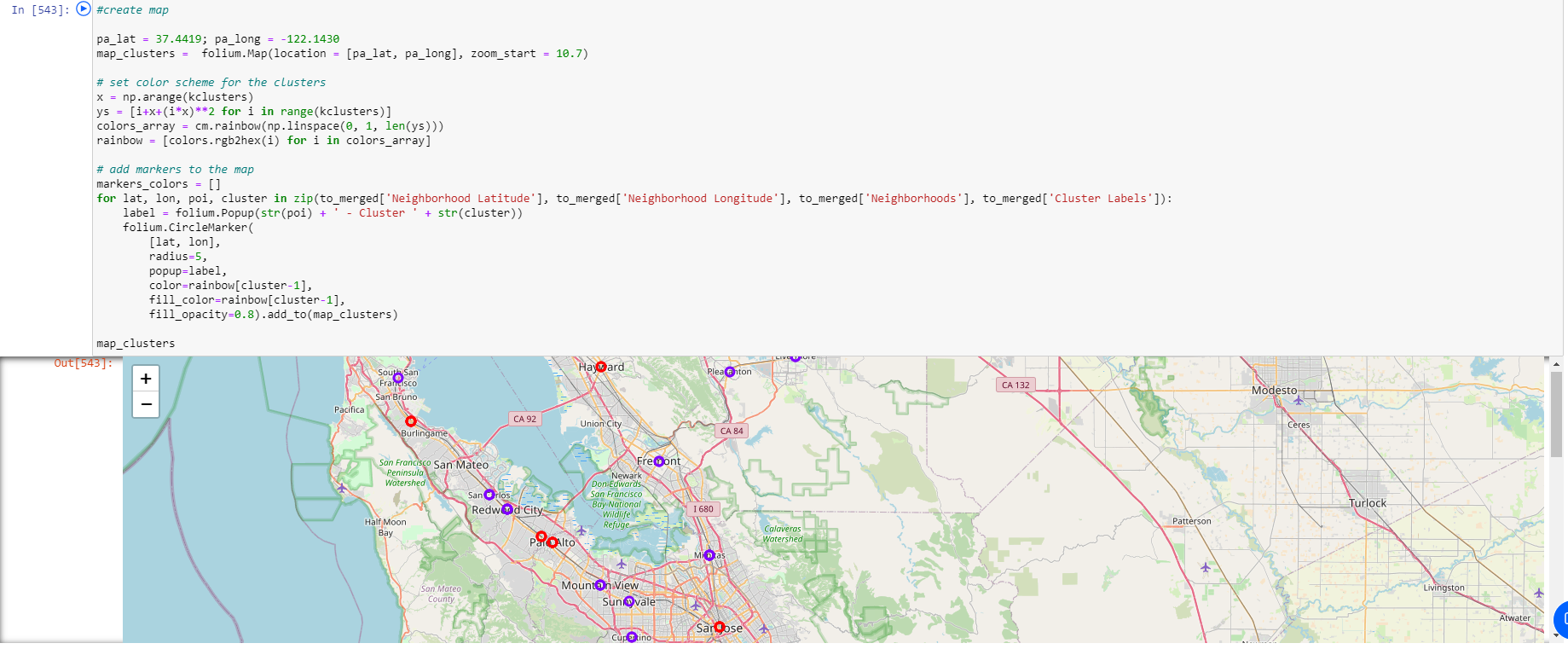


I then merged the venue data with the table above, creating a new table which would be the basis for analyzing new opportunities for opening a bakery in the Bay Area.



This table was used to create a map using the Folium package in Python, where each neighborhood was colored based on the cluster label.

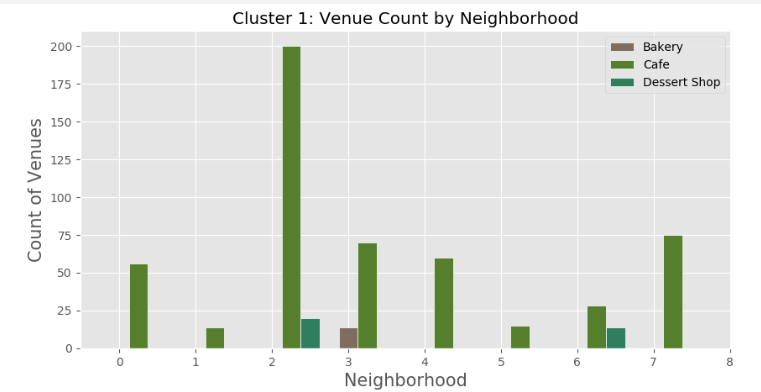
* Cluster 1 — Red
* Cluster 2 — Purple

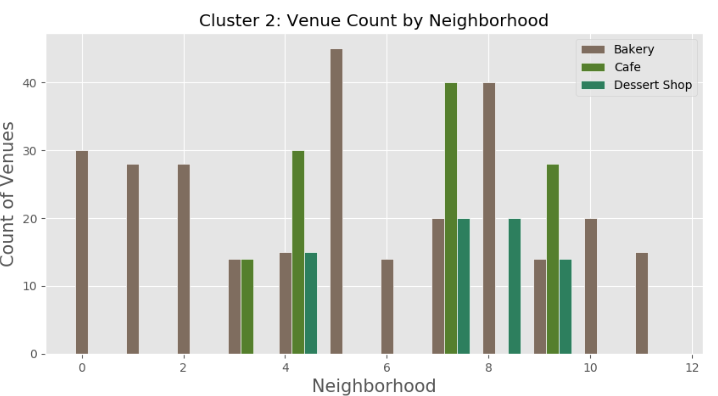


The map above shows the different clusters that had a similar mean frequency of bakeries, cafes, and dessert shops.

#### Cluster Analysis

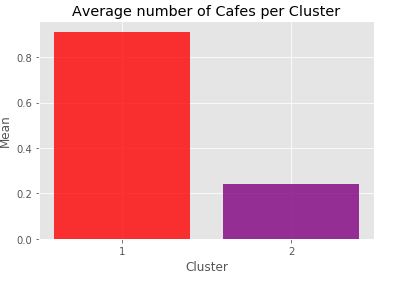
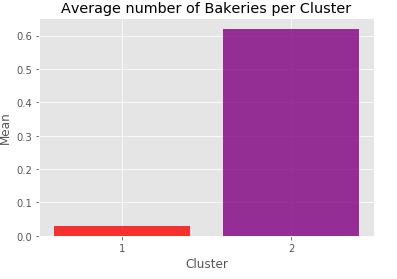
Before analyzing the clusters by venue, we first analyzed the frequency of venues in cities per cluster. From the graphs below, we can see that cities in Cluster 1 have less bakeries and dessert shops than in Cluster 2.

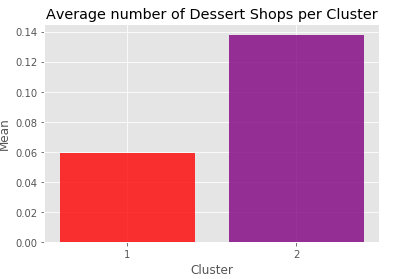




This information is crucial as we can see that even though one neighborhood in neighborhood in Cluster 1, that has a high count of cafes, dessert shops and bakeries seem to be quite scarce among the rest of the neighborhoods. However, this is not the case with Cluster 2, as there are higher bakery counts among the neighborhoods.

In addition to determining the number of venues per neighborhood, a cluster analysis was performed for each venue category: Bakery, Café, and Dessert Shop. A bar chart was generated to aid in the visualization of the cluster analysis for each of these categories. The charts aid in supporting the earlier conclusion that Cluster 1 has less competition in the Bakery and Dessert Shop categories.





**4. Discussion:**

From our analysis, it is determined that cities within Cluster 1 would be optimal choices to open a bakery in. The cities that have both the highest population and density within Cluster 1 are Palo Alto, Dublin, and Brisbane. Based on the performed analysis, I would recommend Cleo’s cookies to open a bakery either in Palo Alto or Dublin.

Some of the drawbacks of this analysis are — the clustering is completely based on data obtained from Foursquare API. Also, the analysis does not take into consideration rent prices or additional demographic information such as age, proximity to schools or businesses, and income. This concludes the optimal findings for this project and recommends the entrepreneur to open a bakery in either Palo Alto or Brisbane to have little to no competition.

**5. Conclusion**

In conclusion, to end off this project, I had an opportunity on a business problem, and it was tackled in a way like how a genuine data scientist would do. I utilized numerous Python libraries to fetch the information, controlled the content, and analyzed and visualized those datasets. Also, I utilized the Foursquare API to investigate the settings in neighborhoods within the San Francisco Bay Area and utilized different plots present in seaborn and Matplotlib libraries. Finally, I applied AI strategies that were taught through the data science course to come up with a final proposal on where a new bakery should be built in the Bay Area.

If I had to do this project over, I wouldn’t filter out the counties in the beginning. It would be interesting to see how the additional data could be leveraged and if it would lead to a different or even better result. Additionally, it would be interesting to perform additional analyses on data points such as neighborhood incomes and facility rent prices.