**Applied Data Science Capstone**

**Clustering Analysis For Restaurant Location in San Francisco**

By Jatin Selmokar

January 2020

# Contents

[Contents 2](#_Toc29151078)

[Introduction/Business Problem 3](#_Toc29151079)

[Data 4](#_Toc29151080)

[Methodology 4](#_Toc29151081)

[Results 5](#_Toc29151082)

[Discussion 6](#_Toc29151083)

[Conclusion 6](#_Toc29151084)

[References 7](#_Toc29151085)

[Appendix 8](#_Toc29151086)

# Introduction/Business Problem

Opening a restaurant in a densely populated city is always challenging and often requires understanding of the current food preferences, location, competition, and the capital investment associated with it.

San Francisco being one of the most populated cities in the US has plethora of restaurants offering different cuisines. Although, offering great food at a lower cost is one of the success metrics, there are external factors that define restaurant’s success. Thus, in order to establish a prosperous business model, it is imperative for a business owner to understand and gauge the restaurant business in and around San Francisco.

Our main objective in this capstone project is to guide the business client in choosing the perfect location to open a restaurant. This project aims to analyze and provide insights on restaurant businesses around San Francisco neighborhoods using unsupervised learning techniques (KMeans Clustering) so that the business owner can make an informed decision.

# Data

For the data analysis in this project, we will need the following data

* San Francisco neighborhood locations
* Geo-coordinates for the SF neighborhoods.
* Surrounding restaurant venue data

# Methodology

The neighborhood data for SF is gathered from the Wikipedia page (<https://en.wikipedia.org/wiki/List_of_neighborhoods_in_San_Francisco>). The data is extracted using Beautiful Soup and Requests packages available in python. We then use Google Maps API to extract geo-coordinates for the neighborhoods. This will be our primary data that will be used in getting surrounding venues in each neighborhood.

We will use Foursquare API to get nearby venues for all the neighborhoods. Since our business client is interested in restaurants, we will filter out all the other categories from the venues data.

Once data is clean and formatted, we will then use unsupervised learning method (K-Means clustering) to form groups of clusters that are similar in nature.

To get the optimal value of cluster size, the data is fitted in K-Means model for different values of k. The inertia metric i.e., within-cluster sum-of-squares criterion is observed and plotted against all values of K. The elbow point is found and its associated K-value is chosen for further analysis.

# Results

The clustering algorithm forms 5 clusters which can be further used to understand the restaurant outlook in San Francisco. They also give information on the popular restaurants in each of the neighborhood areas.

Cluster Details -

Cluster 1 (Red) – Japanese, Sushi, and Chinese restaurants

Cluster 2 (Purple) – Mexican & Southern/Soul restaurants

Cluster 3 (Blue) – Asian restaurants

Cluster 4 (Green) – American restaurants

Cluster 5 (Orange) – Vietnamese, Italian, and New American restaurants

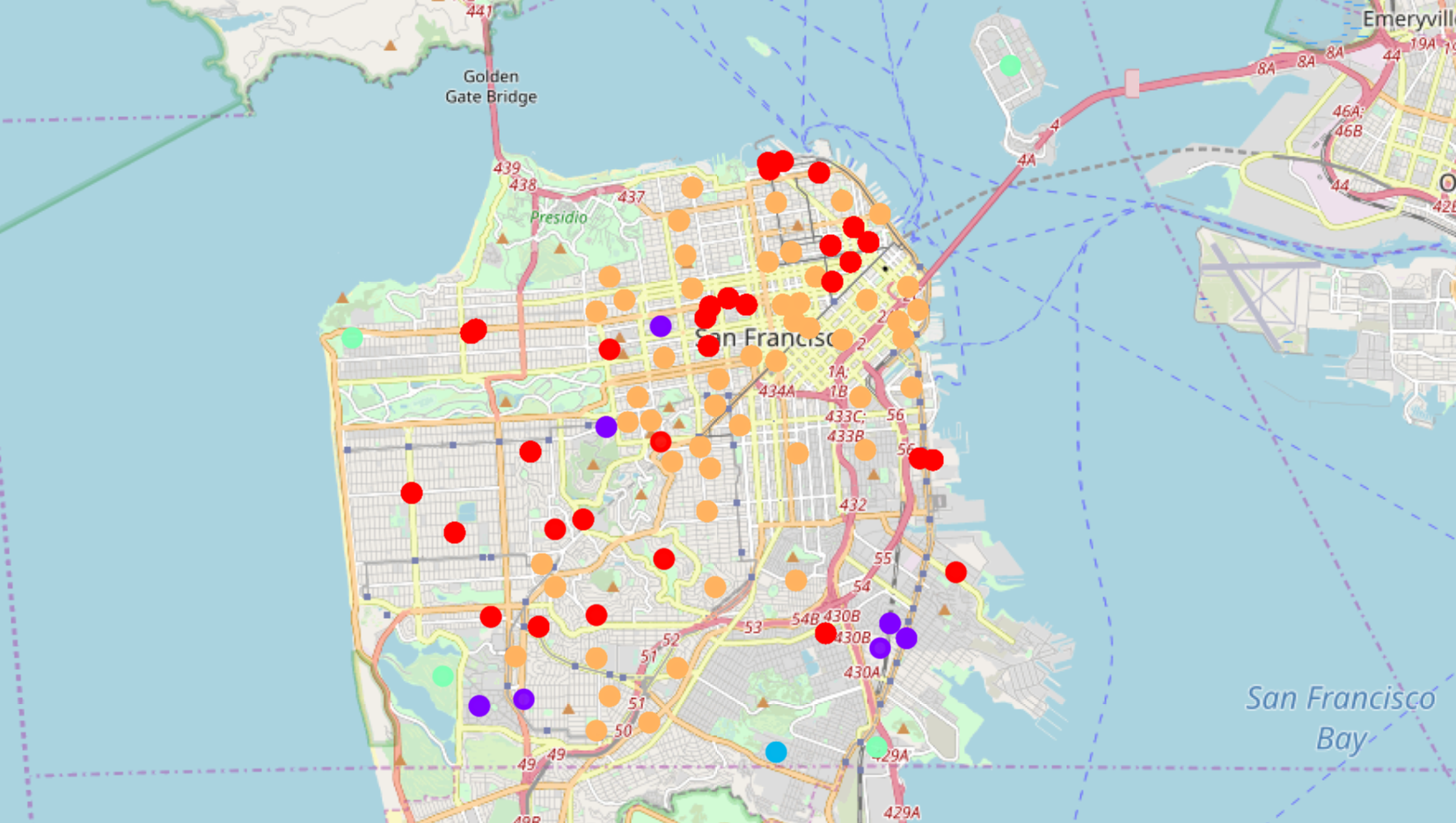


Figure 1 San Francisco Neighborhoods After Clustering

# Discussion

From the clusters, it is evident that there are wide variety of cuisines available in San Francisco. It is also interesting to note that majority of the neighborhoods fall in the orange cluster (Vietnamese, Italian, and New American) which suggests that there is huge demand for these popular cuisines.

The business owner can leverage this information to open restaurants in locations where there is demand and little to no competition. For instance, if the business client had to open an Indian restaurant, neighborhoods under the cluster 2,3, and 4 seem to be good options.

# Conclusion

The clusters give valuable insights on the restaurant businesses in San Francisco area. Business client can use this information to open the restaurant of his choice.

Although the clustering was done on the frequency of restaurants around a neighborhood, considering the average household income and population will provide a detailed perspective and will assist in better decision making.

# References

1. <https://en.wikipedia.org/wiki/List_of_neighborhoods_in_San_Francisco>
2. <https://scikit-learn.org/stable/modules/clustering.html>
3. <https://www.geeksforgeeks.org/ml-determine-the-optimal-value-of-k-in-k-means-clustering/>
4. [https://github.com/limchiahooi/Coursera\_Capstone/](https://github.com/limchiahooi/Coursera_Capstone/blob/master/week5_final_report.pdf)

# Appendix

**Cluster 1 Neighborhoods**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Alamo Square | Corona Heights | Fishermans Wharf | Japantown | Outer Sunset | Union Square |
| Balboa Terrace | Diamond Heights | Forest Hill | Laguna Honda | Parkside | Western Addition |
| Belden Place | Dogpatch | India Basin | Little Russia | Portola | Westwood Highlands |
| Buena Vista | Fillmore | Inner Sunset | Lone Mountain | Richmond District |  |
| Cathedral Hill | Financial District | Irish Hill | Merced Manor | South End |  |
| Chinatown | Financial District South | Jackson Square | North Beach | Sunset District |  |

**Cluster 2 Neighborhoods**

|  |  |
| --- | --- |
| Anza Vista | Parnassus |
| Butchertown (Old and New) | Portola Place |
| Merced Heights | Silver Terrace |
| Parkmerced |  |

**Cluster 3 Neighborhoods**

Sunnydale

**Cluster 4 Neighborhoods**

|  |  |  |  |
| --- | --- | --- | --- |
| Lakeshore | Little Hollywood | Treasure Island | Vista del Mar |

**Cluster 5 Neighborhoods**

|  |  |  |
| --- | --- | --- |
| Ashbury Heights | Hayes Valley | Mission Dolores |
| Bernal Heights | Ingleside | Nob Hill |
| Castro | Ingleside Terraces | Noe Valley |
| Cayuga Terrace | Jordan Park | North of Panhandle |
| China Basin | Lakeside | Oceanview |
| Civic Center | Laurel Heights | Outer Mission |
| Cole Valley | Lincoln Manor | Pacific Heights |
| Cow Hollow | Little Saigon | Polk Gulch |
| Design District | Lower Haight | Potrero Hill |
| Dolores Heights | Lower Pacific Heights | Presidio Heights |
| Duboce Triangle | Lower Nob Hill | Rincon Hill |
| Embarcadero | Marina District | Russian Hill |
| Eureka Valley | Mid-Market | Saint Francis Wood |
| Glen Park | Mission Bay | South Beach |
| Haight-Ashbury | Mission District | South of Market |
| West Portal | Telegraph Hill | South Park |
| Westwood Park | Tenderloin | Yerba Buena |