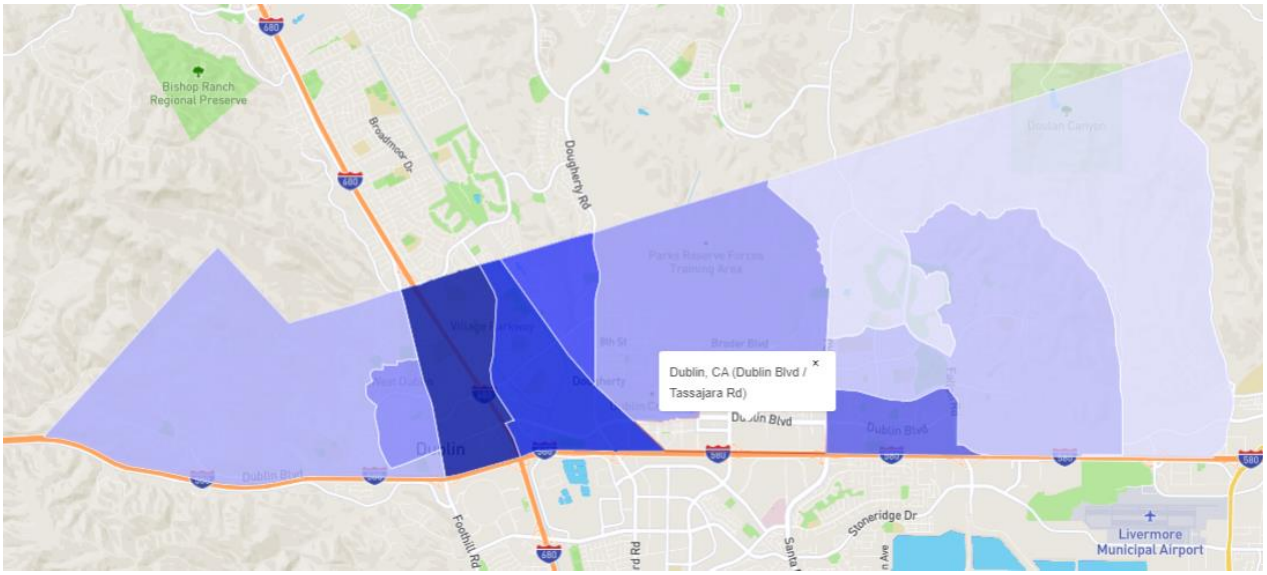
IBM Applied Data Science Certificate Capstone

Relocation to Dublin, Tri-valley to ramp up your business

Xiaoxiong Ma **October 2019**


680
Bishop Ranch Regional Preserve
Broadmoor Dr
Dougherty Rd
Dublin Canyon
Parks Reserve Forces Training Area
Village Parkway
8th St
Dougherty
Dublin, CA (Dublin Blvd / Tassajara Rd)
Dublin Blvd
Dublin
Dublin Blvd
580
Foothill Rd
Stoneridge Dr
Livermore Municipal Airport

Introduction

Dublin, CA is the suburban city of San Francisco East bay. It is part of Tri-valley region of Alameda County, California. As of 2019, city of Dublin consists of 19 constituent neighborhoods, and is the home of a population of 60,939 people, including me and my family.

We have been living in Dublin, CA for almost 5 years and are consistently seeing significant upside trends of domestic population. More and more highly educated people are relocating to Dublin area from various cities across the bay. Real estate developers and investors have been and are still diving into this area, especially east part of Dublin, to construct business and residential properties.

Dublin, together with its sister city Pleasanton, which is adjacent to Dublin right across the highway 580, are currently accommodating some great global corporations, including Workday, SAP, Oracle, Ross stores, Safeway, etc.

According to the 2018 released report by the Bay Area Council Economic Institute[1], *Tri-valley area, where Dublin & Pleasanton are located, are now home to over 450 tech companies, economically boosted by more than $4 billion investment over last decade.*

Regionally, 12 percent job growth for last 4 years is only the start, "*We want the next Google and the next Facebook and the next LinkedIn to be created and housed here in the Tri-valley*," said Dublin Mayor David Haubert.[2]

3 main reasons are endorsing Mayor Haubert's statement:

- Relatively low rent (Dublin $36/sqft, compared with Silicon Valley $52/sqft, & SF $72/sqft)
- Increased density of population (population has been boosted for last several decades)
- No shortage of talents (61.48% of its adults holding 4-year or even advanced degree)

DUBLIN
CALIFORNIA
THE NEW
AMERICAN
BACKYARD

Census	Population	Rate
1970	13,641	N/A
1980	13,496	-1.1%
1990	23,229	72.1%
2000	29,973	29.0%
2010	46,036	53.6%
2019	60,939	32.4%

At the same time, Dublin residents are suffering longer than average commuting time, specifically, people here spend average 40 minutes each day getting to work, which is significantly higher than the national average.

[1] http://www.bayareaeconomy.org/report/tri-valley-rising-2018/

[2] https://sanfrancisco.cbslocal.com/2018/07/18/tri-valley-area-becoming-an-attractive-destination-for-tech-companies/

Business Problem

Keeping the background and stakeholders' interests in mind, I am making reasonable assumption that business relocation will drive out a win-win scenario, both parties, local residents and future employers, will have high motivation to facilitate the business/working sites shifting to Dublin area.
Obviously, employers will benefit from decreased rental price, and expanded talent pool by renting and hiring locally in Dublin. Dublin residents will gain increasing exposure to local job opportunities and take advantage of the shorter commuting time in the future.

Target audience

For my project, target audience will be employers' executives who are considering to tag Dublin area as their potential future headquarter or subordinate working locations. I would like to leverage my data science skills to help those employers to locate their business address, through an inter-neighborhood evaluation within city of Dublin and help shed some light about the most suitable neighborhood from business relocation point of view.

Data

I am gathering and preparing the neighborhood information to facilitate the data exploration.

- List of neighborhoods in Dublin, CA.
- Neighborhood housing prices information
- Neighborhood Latitude and longitude coordinates
- High density venue data within each neighborhood

List of neighborhoods defines the scope of this project which will be focusing on the city of Dublin, CA. Housing Price is a key parameter that will differentiate neighborhoods when employers are selecting the business locations. Latitude and longitude coordinates data is required to visualize the map and facilitate the extraction of venue data. Venue data can be grouped and break down into several features to finalize the clustering.

Data Source & Description
Unfortunately, demographic data from Wikipedia is not narrowed down by neighborhoods. Therefore, we need to leverage other social media sites that are specializing in neighborhoods:

- Neighborhood scout https://www.neighborhoodscout.com/ca/dublin
- Nextdoor.com https://nextdoor.com/city/dublin--ca/
- 680homes.com https://www.680homes.com/Search-By-Neighborhood/Dublin/

10 neighborhoods are tagged at Neighborhood scout, visualized information is most likely qualitative and hard to be extracted through web scraping. Nextdoor.com identified 20 local neighborhoods, with each one followed with designated population, average age and neighborhood favorites. It seems to be attractive at the beginning, however, I found a lot of data missing after looking into it, i.e. more than 5 neighborhoods are missing the population data. Meanwhile, since new employers will definitely share

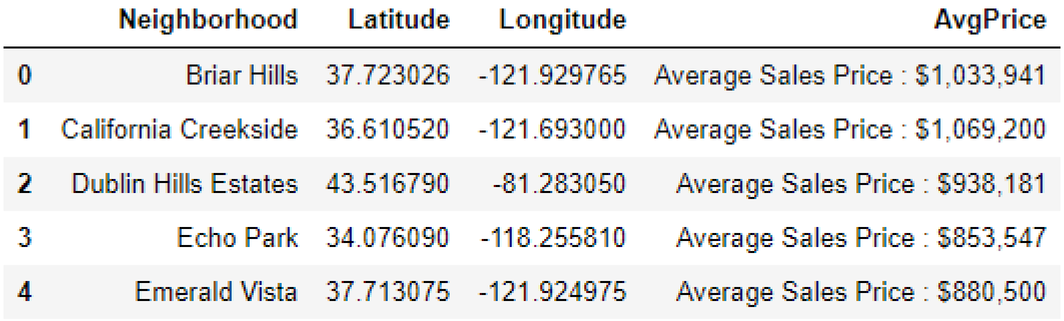
the talent pool within all neighborhoods, taking population as a factor might lead to overfitting. Eventually, we find 680homes.com is the best option. 19 neighborhoods are clearly defined and average Housing price information is provided as well. I am able to use web scraping techniques to extract neighborhood and housing price data from 680homes.com, by utilizing Python requests and beautifulsoup packages. Geographical coordinates of the neighborhoods can be obtained by Python Geocoder package. Foursquare API will help access the venue data for all 19 neighborhoods. Out of all the venue categories, we will select the most representative feature categories and narrow down into 2 dimensions as follows,

- **Convenience**: provide employee work accessibility & life necessity
- **Productivity**: rejuvenate work morale, focus on work life balance & potential growth

This project will focus on the topic of business location selection. Data science skills will be made full use of to drive out deliverables, from web scraping to gather the data, working with API (Foursquare) to get venue coordinates, data cleaning, data wrangling, to machine learning (K-means clustering) and data visualization (matplotlib and folium). Housing price, Convenience, and productivity are 3 main factors that I designed to build up clustering model.

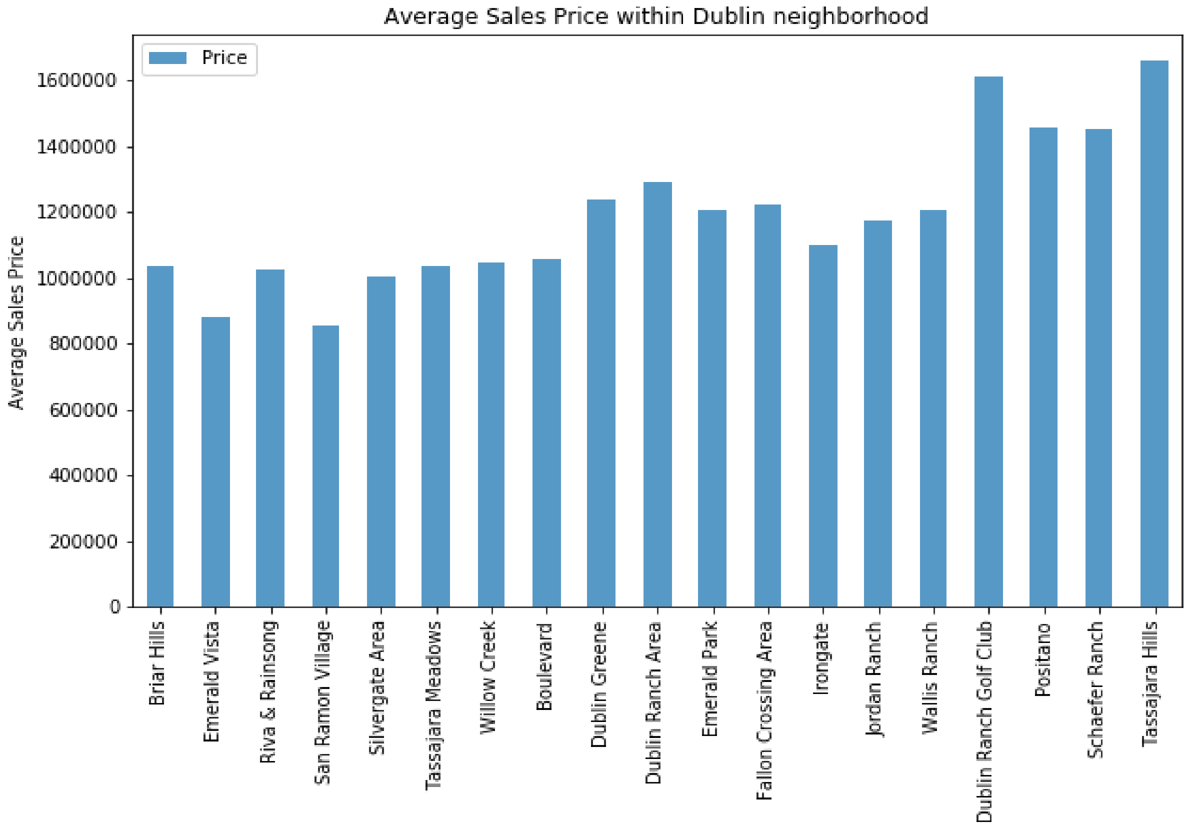
Methodology

Since neither Wikipedia nor US census has clearly classified neighborhoods of city of Dublin, I have to go through 3 other neighborhood specialized websites to locate the data. 680Home.com won my attention after the comparison. Geocoder package helps extract the coordinates of each neighborhood and a portion of consolidated neighborhood dataframe is listed below.

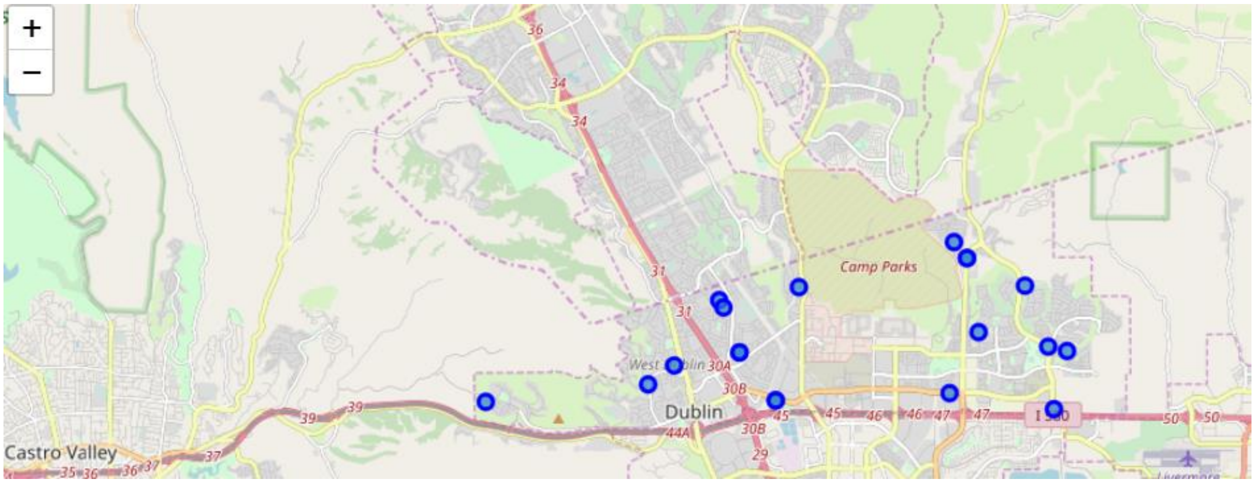


	Neighborhood	Latitude	Longitude	AvgPrice
0	Briar Hills	37.723026	-121.929765	Average Sales Price : $1,033,941
1	California Creekside	36.610520	-121.693000	Average Sales Price : $1,069,200
2	Dublin Hills Estates	43.516790	-81.283050	Average Sales Price : $938,181
3	Echo Park	34.076090	-118.255810	Average Sales Price : $853,547
4	Emerald Vista	37.713075	-121.924975	Average Sales Price : $880,500

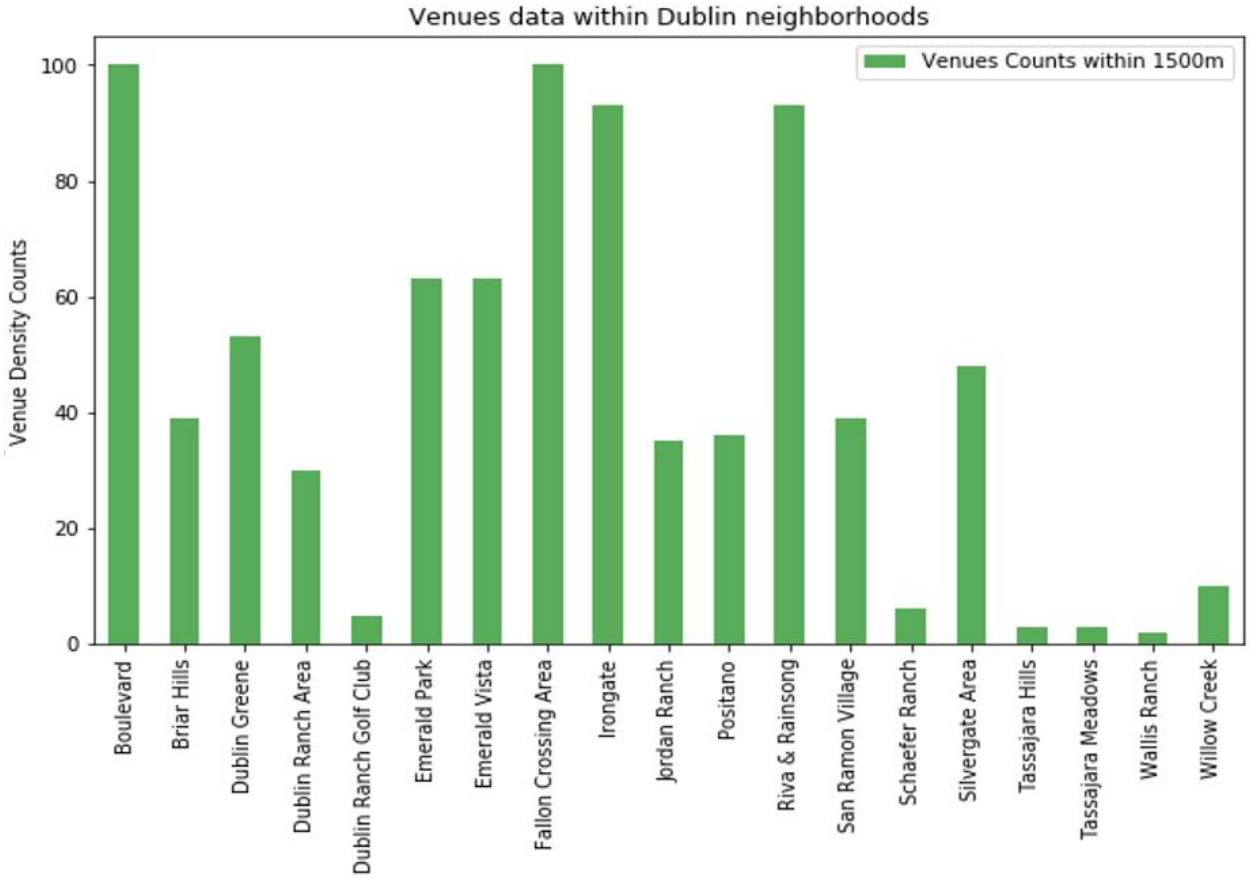
Matplotlib package is also used to visualize the average price of each neighborhood. For 19 local neighborhoods, prices range from $850,000 to $1,600,000 as below. Neighborhoods, prices are plotted as X axis and Y axis as below:



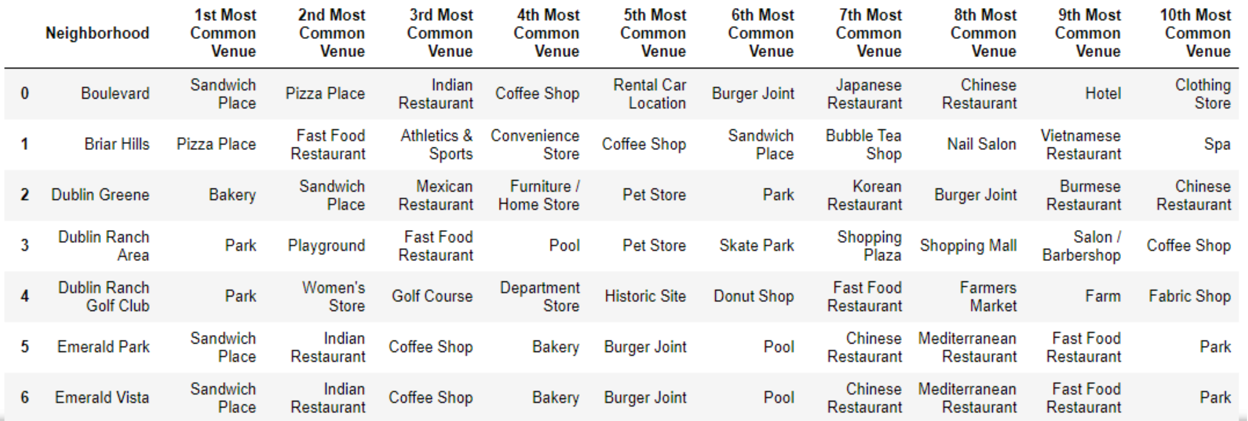
Python folium library is utilized to visualize the geographical details of all neighborhoods. Neighbor data points are superimposed on top of the Dublin map. I also use latitudes and longitudes values to get the visual as below:



Foursquare API is utilized to explore the venue data points around each neighborhood. Foursquare will return JSON format venue data. Requests library can be used to parse JSON into data frame. Radius has been set as 1500m and total venue limit is 100. 7 out of 19 neighborhoods have more than 50 venue within the radius area, while less than 10 venues are found in 5 neighborhoods.



Totally, 835 venues and 163 unique venue categories are returned for all 19 neighborhoods by Foursquare API. Top 10 venue categories for several neighborhoods are listed as below:



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Boulevard	Sandwich Place	Pizza Place	Indian Restaurant	Coffee Shop	Rental Car Location	Burger Joint	Japanese Restaurant	Chinese Restaurant	Hotel	Clothing Store
1	Briar Hills	Pizza Place	Fast Food Restaurant	Athletics & Sports	Convenience Store	Coffee Shop	Sandwich Place	Bubble Tea Shop	Nail Salon	Vietnamese Restaurant	Spa
2	Dublin Greene	Bakery	Sandwich Place	Mexican Restaurant	Furniture / Home Store	Pet Store	Park	Korean Restaurant	Burger Joint	Burmese Restaurant	Chinese Restaurant
3	Dublin Ranch Area	Park	Playground	Fast Food Restaurant	Pool	Pet Store	Skate Park	Shopping Plaza	Shopping Mall	Salon / Barbershop	Coffee Shop
4	Dublin Ranch Golf Club	Park	Women's Store	Golf Course	Department Store	Historic Site	Donut Shop	Fast Food Restaurant	Farmers Market	Farm	Fabric Shop
5	Emerald Park	Sandwich Place	Indian Restaurant	Coffee Shop	Bakery	Burger Joint	Pool	Chinese Restaurant	Mediterranean Restaurant	Fast Food Restaurant	Park
6	Emerald Vista	Sandwich Place	Indian Restaurant	Coffee Shop	Bakery	Burger Joint	Pool	Chinese Restaurant	Mediterranean Restaurant	Fast Food Restaurant	Park

After looking into the top 10 common venue categories, we can easily find that common venues for each neighborhood are quite versatile and different. However, most of common venues will fall under certain groups, such as restaurants, outdoor places, and convenience stores. I picked 17 representative venue categories out of 163 (12%) to effectively eliminate the homogeneous factors and comply with our pre-designed features: ***convenience and productivity.*** These categories will facilitate the clustering model.

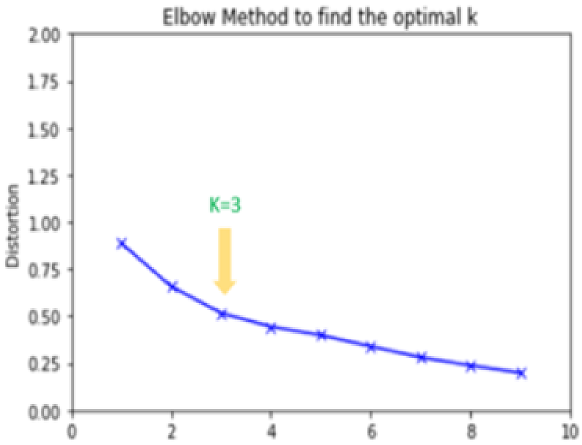
Convenience

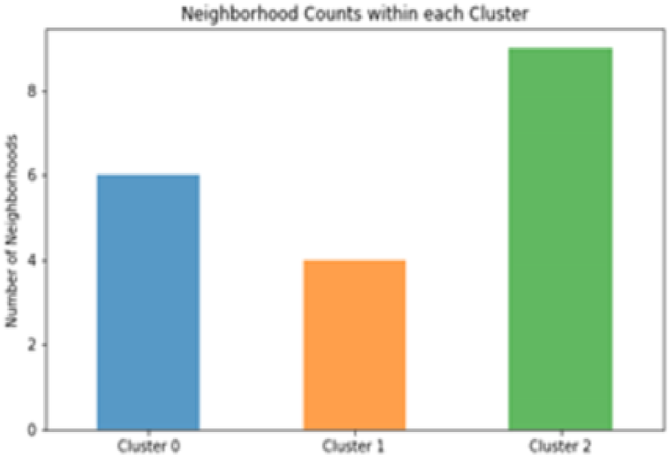
1. **Food**: café & coffee shop, bakery, fast food, diverse style restaurants
2. **Shopping**: shopping mall, grocery stores
3. **Commute**: bus station, gas station, ATM.

Productivity

1. **Workout**: gym/fitness center, parks, trails
2. **Entertainment**: bar, movie & music, recreation center
3. **Lifestyle**: pharmacy, bookstore, pet store

After the features selection and grouping, it is time to practice the clustering modeling. I chose K-means clustering, which is an unsupervised machine learning algorithm, to analyze my on-hand data. 3 most important factors, **Housing price, Convenience, and Productivity,** are factored into the model to locate data point of each neighborhood. In order to determine what is the K number of centroids to initiate the labelling, I implemented the elbow method. We can see that there are 2 slight elbow shape points created, which are 3 and 5. In this case, I will be choosing K=3 to avoid the overfitting situation since the graph below is indicating differences between neighborhoods are not as significant as I first thought.

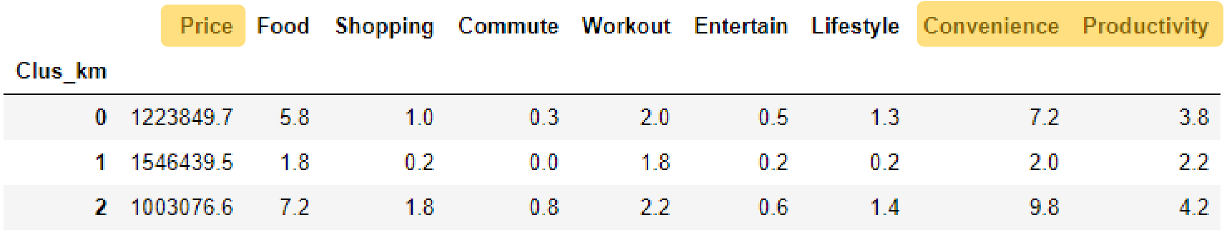




Results

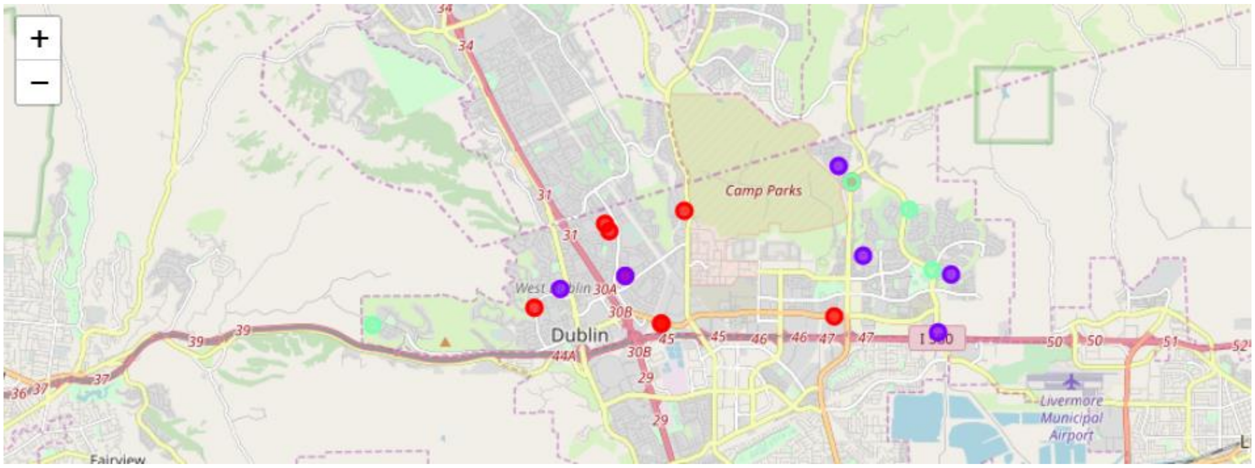
Sklearn package is mainly leveraged to normalize the data and implement K means algorithm. Folium can be used again to visualize the neighborhoods by each cluster. Results of clusters and their corresponding features are plotted above and explained below,

- ***Cluster 0: Medium rent, medium high convenience, and high productivity (6 neighborhoods)***
- ***Cluster 1: High rent, low convenience, and medium productivity (4 neighborhoods)***
- ***Cluster 2: Low rent, high convenience, and medium high productivity (9 neighborhoods)***



Clus_km	Price	Food	Shopping	Commute	Workout	Entertain	Lifestyle	Convenience	Productivity
0	1223849.7	5.8	1.0	0.3	2.0	0.5	1.3	7.2	3.8
1	1546439.5	1.8	0.2	0.0	1.8	0.2	0.2	2.0	2.2
2	1003076.6	7.2	1.8	0.8	2.2	0.6	1.4	9.8	4.2

From the folium mapping, purple data points stand for cluster 0, green mint data points stand for cluster 1, red data points stand for cluster 2.



Discussion

When scraping online data, we found that neighborhood definitions of Dublin, CA are conflicting within different websites. I chose the most detailed one to gather both neighborhood name and average housing prices. I disregarded some other pre-designed factors, such as neighborhood population, education level, and crime rates, etc. Population feature is dropped to avoid overfitting since future employers will be sharing the talent pool across all neighborhoods in Dublin. Similar case with education level, since business should not only hire high-educated people from certain neighborhoods, whole Dublin, or even the whole Try-valley area will be highly exposed. Local rent must be considered as it is directly affecting business operational cost. Local crime rate matters, however, this factor is also excluded from my analysis due to the hardship to obtain quantitative crime data for each neighborhood of Dublin.

Foursquare API delivers 163 unique venue categories and some of them are highly homogenous, such as, Japanese restaurants v.s. sushi restaurants, coffee shops v.s. café, etc. I decided to strip out the redundant categories and reclassify them by 2 criteria, convenience and productivity[3]. Convenience takes employees' daily necessities into account. It is critical for employees to have high accessibility to their point of interests, such as food, gas station, and convenience stores. Productivity pays more

[3] https://www.virgin.com/entrepreneur/six-factors-consider-when-choosing-location-your-business

attention on employees' growth potential, either personally or professionally. That is why gym/fitness center, bookstores, and pharmacies are baked into this feature.

Selection of K is compromising. We technically did not see an obvious elbow shape when we run the optimal-k visualization. It implies that distinction between each neighborhood is not statistically significant. Dublin, CA is a suburban and fast-developing city, it embraces great opportunities and foresees the promising future, but is after all not a metropolitan city, like San Francisco. Another hypothesis might be, filtered local venue categories and renting prices are not sufficient enough to distinguish them and group them into unique clusters, and cannot trigger the significant elbow shape. Future study might focus on involving more diversified features to help differentiate neighborhoods from multiple angles.

Overall, **K=3 is still acceptable**, since distortion rate firstly drops down to less than 1 when K=3, and when K keeps increasing, centroids are closer to the clusters centroid, improvements is declining, so K=3 is the best choice at this point.

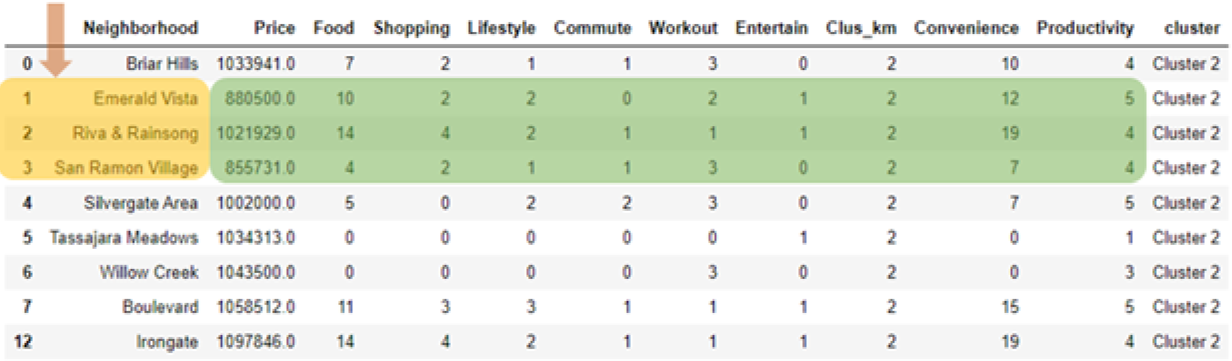
Within the scope of my project, neighborhoods falling under cluster 1 are the least attractive to future employers as they will be paying higher rent than other local areas, and vicinity areas are also not convenient or productive enough to cater to their employees.

Without cluster 2, cluster 0 can be an option since those 6 neighborhoods are averagely providing a commuter friendly working environment. However, after cross checking the other features, especially the rent, cluster 2 eventually stands out. The average rent price of its 9 neighborhoods is 20% lower than cluster 0. Meanwhile, infrastructure and other facilities are well spread out or under construction in this cluster to provide great point of interests for the future employees.

Conclusion

Neighborhoods falling under **Cluster 2** is the best choice to relocate working sites.

Relative low rent, fully equipped infrastructure and diverse community facilities all help establish attractive working environment. Within cluster 2, I highlighted 3 neighborhoods as the priority business locations, which are **Emerald Vista, Riva & Rainsong,** and **San Ramon Village.**



	Neighborhood	Price	Food	Shopping	Lifestyle	Commute	Workout	Entertain	Clus_km	Convenience	Productivity	cluster
0	Briar Hills	1033941.0	7	2	1	1	3	0	2	10	4	Cluster 2
1	Emerald Vista	880500.0	10	2	2	0	2	1	2	12	5	Cluster 2
2	Riva & Rainsong	1021929.0	14	4	2	1	1	1	2	19	4	Cluster 2
3	San Ramon Village	855731.0	4	2	1	1	3	0	2	7	4	Cluster 2
4	Silvergate Area	1002000.0	5	0	2	2	3	0	2	7	5	Cluster 2
5	Tassajara Meadows	1034313.0	0	0	0	0	0	1	2	0	1	Cluster 2
6	Willow Creek	1043500.0	0	0	0	0	3	0	2	0	3	Cluster 2
7	Boulevard	1058512.0	11	3	3	1	1	1	2	15	5	Cluster 2
12	Irongate	1097846.0	14	4	2	1	1	1	2	19	4	Cluster 2

Reference

1. Tri-Valley Area Becoming An Attractive Destination for Tech Companies
2. Six factors to consider when choosing a location for your business
3. Tr-Valley Rising 2018 Bay Area Council Economic Institute
4. Foursquare API https://developer.foursquare.com
5. 680homes.com https://www.680homes.com/Search-By-Neighborhood/Dublin/
6. Neighborhood scout https://www.neighborhoodscout.com/ca/dublin
7. Nextdoor.com https://nextdoor.com/city/dublin--ca/