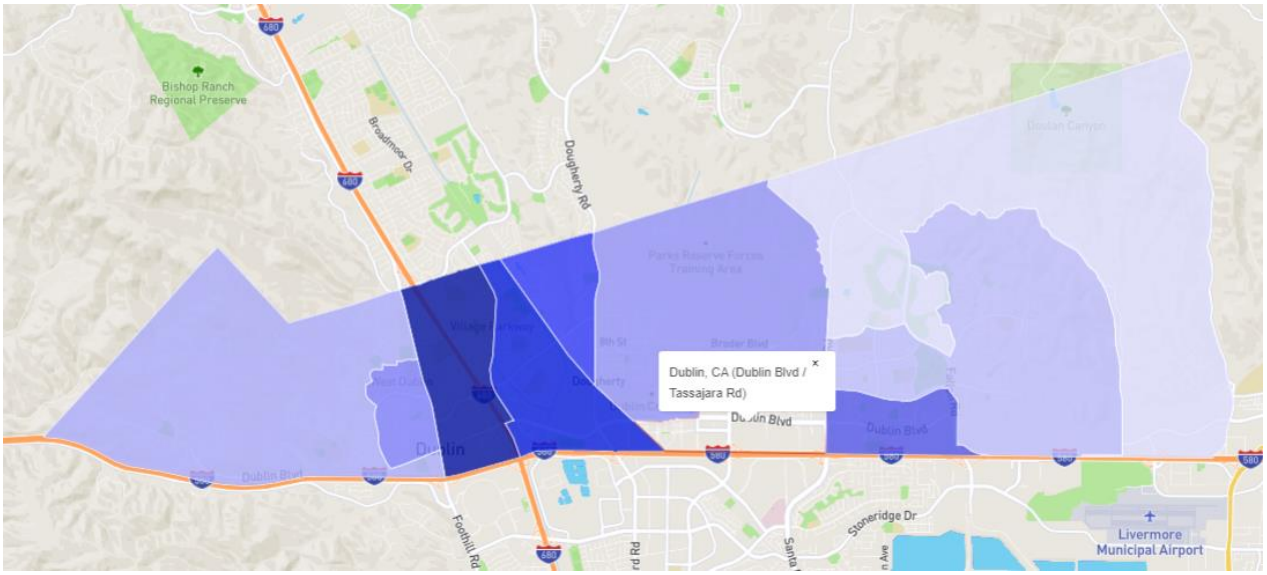
IBM Applied Data Science Certificate Capstone

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

Xiaoxiong Ma **October 2019**



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 10 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 trending 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 or eager to accommodate some great corporations, including Workday, SAP, Oracle, Ross stores, Safeway, etc.

According to the 2018 released report by the Bay Area Council Economic Institute, 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.

There are 3 main reasons that will endorse Mayor'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.

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 employers, will have high motivation to facilitate the business/working sites shifting to Dublin area.
Obviously, employers will benefit from low housing 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 short commute in the future.

Target audience

For my project, target audience would be employers' executives who are considering to tag Dublin area as their future working sites. I would like to leverage my data science skills to help those potential employers to locate their business address, through an inter-neighborhood evaluation within city of Dublin and help shed some light upon employers about the most suitable neighborhood from the business location 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
- Venue data, particularly related to employment

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 Sources and ETL methodology
Unfortunately, demographic data from Wikipedia is not narrowed down by neighborhoods. So we need to leverage other social media sites that are specializing in neighborhoods. I cross checked 3 websites:

- 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. 6 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 69homes.com is the best option. 29 neighborhoods are defined and average housing

price information is provided as well. We will 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 captured by Python Geocoder package. Foursquare API, crowdsourcing more than 100 million venues and 125,000 developing, will help access the venue data for all 29 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 & life necessity
  - Food & Beverage (café, bakery, fast food, etc.)
  - Commute (bus station, gas station, ATM, etc.)
  - Shopping (shopping mall, grocery stores, convenience stores, etc.)
- ✓ **Productivity**: rejuvenate work morale, focus on work life balance
  - Lifestyle (health & beauty service, pharmacy, bookstore, etc.)
  - Workout (gym/fitness, parks, trails, etc.)
  - Entertainment (bar, movie & music, recreation center, etc.)

To wrap up, 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, and next section will be discussing about the methodology and results portion of the project.