Segmenting and Clustering Neighborhoods in San Francisco

Applied Data Science Capstone by IBM/Coursera

photo copy from https://www.flickr.com/photos/canbalci/2435328333

Introduction: Business Problem

In this project we will try to make a recommendation for stakeholders who relocate to San Francisco and want to find an optimal place to live.

Specifically, this project aims to find **safe** neighborhoods in San Francisco, **close to the tech companies** and find the **most common venues** in each neighborhood, so that each area can be clearly expressed and best possible final location can be chosen by stakeholders.

We will use our data science powers to generate a few most promising neighborhoods based on this criteria.

Data

Based on definition of our problem, factors that will influence our decision are:

- number crimes in the neighborhood
- distance between the neighborhoods and tech companies
- venues in the neighborhood

Following data sources will be needed to extract/generate the required information:

- Geographic Locations and Boundaries of Neighborhoods in San Francisco: sfgov
- Crime Incident Report : sfgov
- All Bay Area Tech Companies List: Github open source
- Venues returned for each neighborhood: Foursquare API

Methodology

First Step

We will collected the required data: **geographic locations and boundaries of neighborhoods in San Francisco** and **locations of tech companies in San Francisco**, to accurately calculate distances from the centroid of neighborhood to the centroid of all tech companies, we need to create our grid of locations in **Cartesian 2D coordinate system** which allows us to calculate distances in meters (not in latitude/longitude degrees). Then we'll project those coordinates back to latitude/longitude degrees to be shown on Folium map. So we need to create functions to convert between WGS84 spherical coordinate system (latitude/longitude degrees) and UTM Cartesian coordinate system (X/Y coordinates in meters).

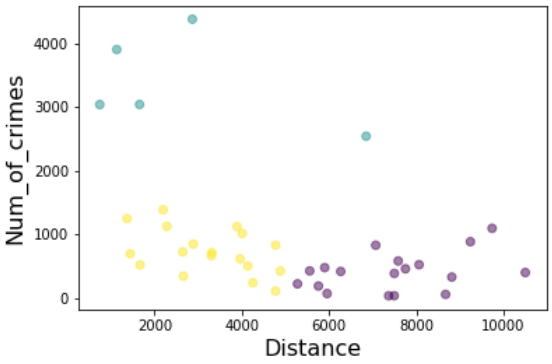
We have also collected the **crime incident data**, keep all data in 2020 and group them by neighborhood to calculate the total number of crimes in each neighborhood. Merge with neighborhoods dataframe and calculate the distance from each neighborhood to the centroid of all tech companies.

Second Step

We will use **k-means clustering** for neighborhood segmentation to find the candidate neighborhoods. We start by normalizing the dataset. Normalization is a statistical method that helps mathematical-based algorithms interpret features with

different magnitudes and distributions equally. We use StandardScaler() to normalize our dataset. Then we run k-means to cluster the neighborhoods into 3 clusters. Note that each row in our dataset represents a neighborhood, and therefore, each row is assigned a label. Then we can easily check the centroid values by averaging the features in each cluster.

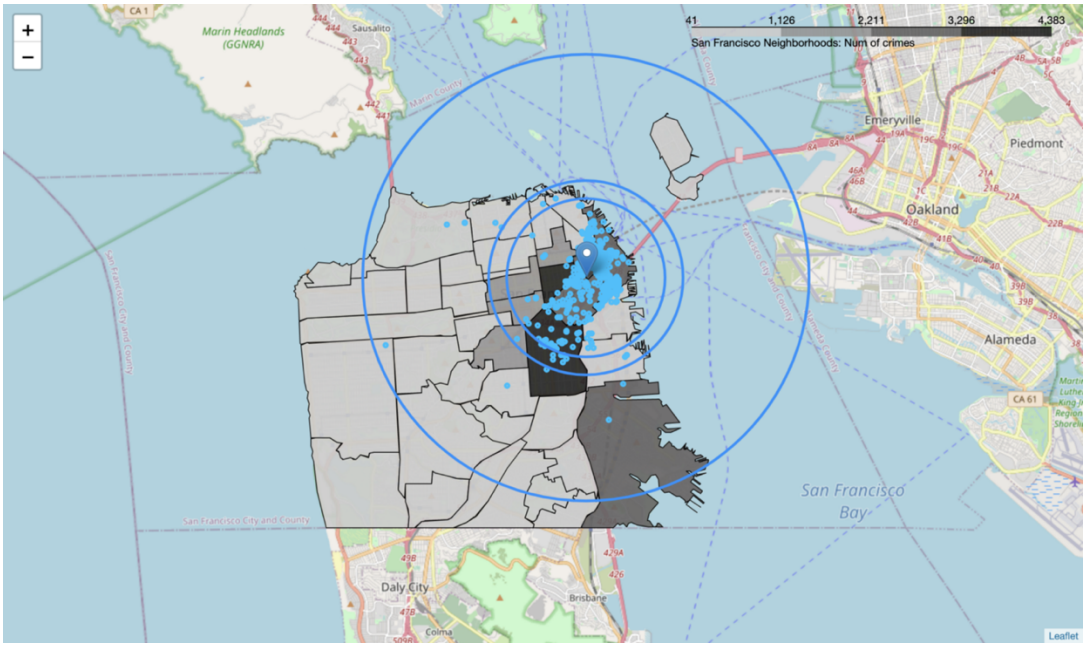
Now, let's look at the distribution of neighborhoods based on their number of crimes and distance to the tech companies:



Therefore we will keep neighborhoods in cluster 2 (yellow). Ok, now we have our candidate neighborhoods ready:

	Neighborhood	Num_of_crimes	Distance	Latitude	Longitude
0	Treasure Island	115	4775.187494	37.815691	-122.367322
1	Presidio Heights	246	4255.033980	37.784668	-122.451945
2	Japantown	349	2662.474054	37.784767	-122.433846
3	Noe Valley	430	4883.753181	37.747638	-122.435390
4	Lone Mountain/USF	507	4140.196872	37.779203	-122.450308
5	Mission Bay	526	1669.362921	37.777850	-122.386186
6	Haight Ashbury	621	3963.476056	37.767229	-122.443559
7	Potrero Hill	672	3301.999091	37.758275	-122.384201
8	Chinatown	699	1440.908662	37.796295	-122.407744
9	Russian Hill	720	3313.224595	37.808877	-122.423921
10	Pacific Heights	731	2648.728079	37.791871	-122.431914
11	Bernal Heights	835	4777.619477	37.741462	-122.414034
12	North Beach	852	2887.511387	37.809381	-122.409134
13	Marina	1019	4012.133119	37.806849	-122.438672
14	Castro/Upper Market	1127	3890.423112	37.761351	-122.437652
15	Hayes Valley	1130	2284.717745	37.774467	-122.426813
16	Nob Hill	1253	1369.714995	37.792162	-122.414994
17	Western Addition	1390	2198.751069	37.783348	-122.428604

We can create a Choropleth map , with lighter areas indicate high levels of safety, maker and circle show where the company gathers. (we need a GeoJSON file that defines the areas/boundaries of San Francisco. We can find the GeoJSON file from sfgov.)



We can also create a map of San Francisco with candidate neighborhoods superimposed on top.

Third Step

We will focus on neighborhoods generated in step 2, use the **Foursquare API** to explore those neighborhoods and to get **the most common venue categories** in each neighborhood, and then use this feature to group the neighborhoods into clusters using **k-means clustering**. We will use the **Folium** library to visualize the neighborhoods in San Francisco and their emerging clusters.

To use Foursquare API, we need to define Foursquare Credentials and Version. Create a function to explore all the candidate neighborhoods in San Francisco. We set LIMIT = 100, radius = 500 to get the top 100 venues within a radius of 500 meters, run the above function on each neighborhood and create a new dataframe called sf_venues.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Treasure Island	37.815691	-122.367322	Woods Island Club	37.818278	-122.367314	Brewery
1	Treasure Island	37.815691	-122.367322	Treasure Island Museum	37.816996	-122.371254	History Museum
2	Treasure Island	37.815691	-122.367322	World Headquarters	37.817392	-122.369802	Art Gallery
3	Treasure Island	37.815691	-122.367322	Treasure Island Yacht Club	37.816616	-122.370483	Harbor / Marina
4	Treasure Island	37.815691	-122.367322	Kite the Bay	37.815853	-122.370839	Harbor / Marina

Then we check how many venues were returned for each neighborhood and how many unique categories can be curated from all the returned venues.

```
          Neighborhood  Num_of_venues
0        Bernal Heights             43
1   Castro/Upper Market             84
2            Chinatown             89
3       Haight Ashbury             31
4         Hayes Valley            100
5            Japantown             93
6    Lone Mountain/USF             33
7               Marina             11
8          Mission Bay             43
9             Nob Hill            100
10          Noe Valley             45
11         North Beach             51
12     Pacific Heights             39
13        Potrero Hill             43
14    Presidio Heights             36
15       Russian Hill             64
16     Treasure Island             11
17    Western Addition             55
There are 227 uniques categories.
```

Then we analyze each neighborhood, group rows by neighborhood and by taking the mean of the frequency of occurrence of each category. Create the new dataframe and display the top 10 venues for each neighborhood.

	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	Bernal Heights	Food Truck	Playground	Bakery	Coffee Shop	Park	Italian Restaurant	Gourmet Shop	Gay Bar	Pet Store	Peruvian Restaurant
1	Castro/Upper Market	Gay Bar	Coffee Shop	Thai Restaurant	Park	Yoga Studio	Clothing Store	Cosmetics Shop	Deli / Bodega	Hill	Playground
2	Chinatown	Coffee Shop	Chinese Restaurant	Café	Pizza Place	Italian Restaurant	Cocktail Bar	Men's Store	New American Restaurant	Dim Sum Restaurant	Bakery
3	Haight Ashbury	Park	Boutique	Breakfast Spot	Thrift / Vintage Store	Scenic Lookout	Coffee Shop	Smoke Shop	Skate Park	Shoe Store	Clothing Store
4	Hayes Valley	Wine Bar	French Restaurant	Sushi Restaurant	Café	Cocktail Bar	Record Shop	Dessert Shop	Clothing Store	Park	Optical Shop

Now let's merge all useful data together.

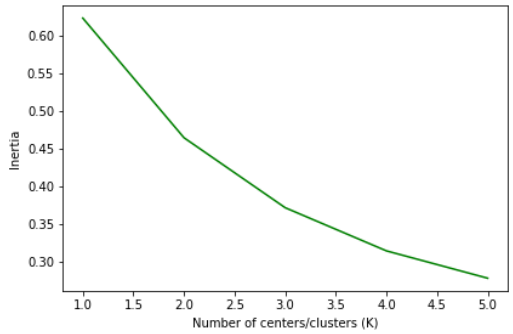
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Num_of_crimes	Distance	Latitude	Longitude
0	Bernal Heights	Food Truck	Playground	Bakery	Coffee Shop	Park	Italian Restaurant	Gourmet Shop	Gay Bar	Pet Store	Peruvian Restaurant	835	4777.619477	37.741462	-122.414034
1	Castro/Upper Market	Gay Bar	Coffee Shop	Thai Restaurant	Park	Yoga Studio	Clothing Store	Cosmetics Shop	Deli / Bodega	Hill	Playground	1127	3890.423112	37.761351	-122.437652
2	Chinatown	Coffee Shop	Chinese Restaurant	Café	Pizza Place	Italian Restaurant	Cocktail Bar	Men's Store	New American Restaurant	Dim Sum Restaurant	Bakery	699	1440.908662	37.796295	-122.407744
3	Haight Ashbury	Park	Boutique	Breakfast Spot	Thrift / Vintage Store	Scenic Lookout	Coffee Shop	Smoke Shop	Skate Park	Shoe Store	Clothing Store	621	3963.476056	37.767229	-122.443559
4	Hayes Valley	Wine Bar	French Restaurant	Sushi Restaurant	Café	Cocktail Bar	Record Shop	Dessert Shop	Clothing Store	Park	Optical Shop	1130	2284.717745	37.774467	-122.426813
5	Japantown	Grocery Store	Gift Shop	Bakery	Tea Room	Boutique	Shopping Mall	Spa	Cosmetics Shop	Café	Japanese Restaurant	349	2662.474054	37.784767	-122.433846
6	Lone Mountain/USF	Coffee Shop	Salon / Barbershop	Wine Bar	Café	Pub	Sushi Restaurant	Supplement Shop	Diner	Liquor Store	Big Box Store	507	4140.196872	37.779203	-122.450308
7	Marina	Harbor / Marina	Coffee Shop	Bank	Lighthouse	Gym	Gym / Fitness Center	Monument / Landmark	Historic Site	Park	Event Space	1019	4012.133119	37.806849	-122.438672
8	Mission Bay	Baseball Stadium	Outdoor Sculpture	Coffee Shop	Park	Harbor / Marina	Sandwich Place	Garden	Café	Pop-Up Shop	Nightclub	526	1669.362921	37.777850	-122.386186
9	Nob Hill	Hotel	Coffee Shop	Italian Restaurant	Grocery Store	Café	Bar	American Restaurant	Spa	French Restaurant	Yoga Studio	1253	1369.714995	37.792162	-122.414994
10	Noe Valley	Gift Shop	Italian Restaurant	American Restaurant	Trail	Bakery	Yoga Studio	Bookstore	Martial Arts Dojo	Burger Joint	Shipping Store	430	4883.753181	37.747638	-122.435390
11	North Beach	Seafood Restaurant	Tour Provider	Coffee Shop	Hotel	Candy Store	Toy / Game Store	Italian Restaurant	Gift Shop	Ice Cream Shop	Cosmetics Shop	852	2887.511387	37.809381	-122.409134
12	Pacific Heights	Cosmetics Shop	Ice Cream Shop	Coffee Shop	Grocery Store	Sandwich Place	French Restaurant	Spa	Bookstore	Boutique	Burger Joint	731	2648.728079	37.791871	-122.431914
13	Potrero Hill	Wine Bar	Brewery	Gift Shop	Cocktail Bar	Bakery	Yoga Studio	Park	Southern / Soul Food Restaurant	Breakfast Spot	Bubble Tea Shop	672	3301.999091	37.758275	-122.384201
14	Presidio Heights	Cosmetics Shop	Italian Restaurant	Wine Bar	American Restaurant	Coffee Shop	Bank	Sporting Goods Shop	Men's Store	Electronics Store	Miscellaneous Shop	246	4255.033980	37.784668	-122.451945
15	Russian Hill	Historic Site	Chocolate Shop	Scenic Lookout	Café	National Park	Park	Gift Shop	Ice Cream Shop	Seafood Restaurant	Bike Rental / Bike Share	720	3313.224595	37.808877	-122.423921
16	Treasure Island	Harbor / Marina	Music Venue	History Museum	Bus Station	Brewery	Fried Chicken Joint	Beach	Event Space	Eastern European Restaurant	Electronics Store	115	4775.187494	37.815691	-122.367322
17	Western Addition	Gift Shop	Grocery Store	Shopping Mall	Park	New American Restaurant	Sushi Restaurant	Japanese Restaurant	Ramen Restaurant	Tea Room	Indian Restaurant	1390	2198.751069	37.783348	-122.428604

Looking good. So now we have all the neighborhoods in San Francisco, and we know which one have less number of crimes, which neighborhoods exactly are in vicinity of the centroid of all tech companies. And we also know the top 10 venues for each neighborhood.

This concludes the data gathering phase—we're now ready to use this data for analysis to produce the report on optimal locations.

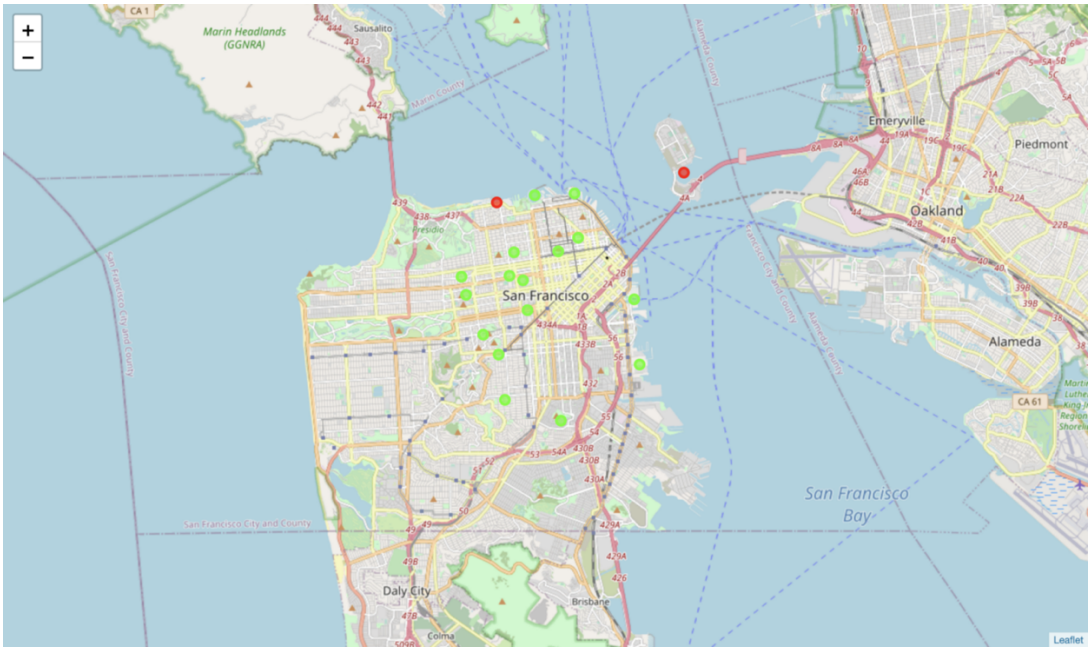
Analysis

K-means is especially useful if you need to quickly discover insights from unlabeled data. To determine the optimal number of clusters, we have to select the value of k at the "elbow":



We can see the elbow point is 2, so we can run k-means to cluster the neighborhood into 2 clusters.

Create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood and create the map:



Let’s examine clusters:

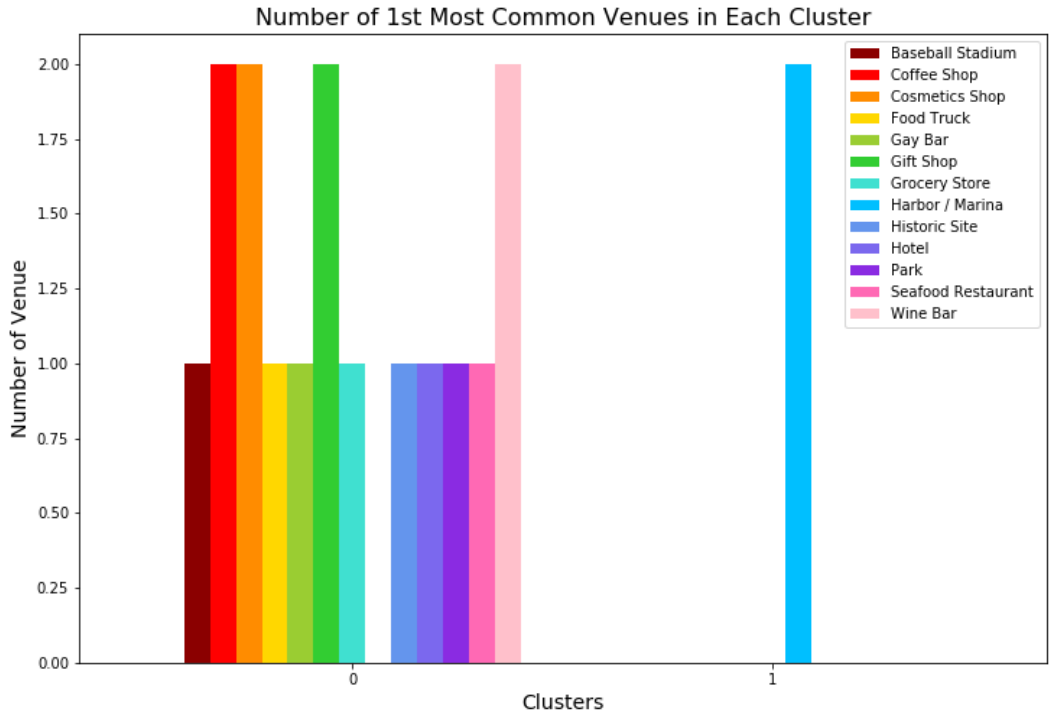
```
# Cluster 1 (green)
```

	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	Bernal Heights	Food Truck	Playground	Bakery	Coffee Shop	Park	Italian Restaurant	Gourmet Shop	Gay Bar	Pet Store	Peruvian Restaurant
1	Castro/Upper Market	Gay Bar	Coffee Shop	Thai Restaurant	Park	Yoga Studio	Clothing Store	Cosmetics Shop	Deli / Bodega	Hill	Playground
2	Chinatown	Coffee Shop	Chinese Restaurant	Café	Pizza Place	Italian Restaurant	Cocktail Bar	Men's Store	New American Restaurant	Dim Sum Restaurant	Bakery
3	Haight Ashbury	Park	Boutique	Breakfast Spot	Thrift / Vintage Store	Scenic Lookout	Coffee Shop	Smoke Shop	Skate Park	Shoe Store	Clothing Store
4	Hayes Valley	Wine Bar	French Restaurant	Sushi Restaurant	Café	Cocktail Bar	Record Shop	Dessert Shop	Clothing Store	Park	Optical Shop
5	Japantown	Grocery Store	Gift Shop	Bakery	Tea Room	Boutique	Shopping Mall	Spa	Cosmetics Shop	Café	Japanese Restaurant
6	Lone Mountain/USF	Coffee Shop	Salon / Barbershop	Wine Bar	Café	Pub	Sushi Restaurant	Supplement Shop	Diner	Liquor Store	Big Box Store
8	Mission Bay	Baseball Stadium	Outdoor Sculpture	Coffee Shop	Park	Harbor / Marina	Sandwich Place	Garden	Café	Pop-Up Shop	Nightclub
9	Nob Hill	Hotel	Coffee Shop	Italian Restaurant	Grocery Store	Café	Bar	American Restaurant	Spa	French Restaurant	Yoga Studio
10	Noe Valley	Gift Shop	Italian Restaurant	American Restaurant	Trail	Bakery	Yoga Studio	Bookstore	Martial Arts Dojo	Burger Joint	Shipping Store
11	North Beach	Seafood Restaurant	Tour Provider	Coffee Shop	Hotel	Candy Store	Toy / Game Store	Italian Restaurant	Gift Shop	Ice Cream Shop	Cosmetics Shop
12	Pacific Heights	Cosmetics Shop	Ice Cream Shop	Coffee Shop	Grocery Store	Sandwich Place	French Restaurant	Spa	Bookstore	Boutique	Burger Joint
13	Potrero Hill	Wine Bar	Brewery	Gift Shop	Cocktail Bar	Bakery	Yoga Studio	Park	Southern / Soul Food Restaurant	Breakfast Spot	Bubble Tea Shop
14	Presidio Heights	Cosmetics Shop	Italian Restaurant	Wine Bar	American Restaurant	Coffee Shop	Bank	Sporting Goods Shop	Men's Store	Electronics Store	Miscellaneous Shop
15	Russian Hill	Historic Site	Chocolate Shop	Scenic Lookout	Café	National Park	Park	Gift Shop	Ice Cream Shop	Seafood Restaurant	Bike Rental / Bike Share
17	Western Addition	Gift Shop	Grocery Store	Shopping Mall	Park	New American Restaurant	Sushi Restaurant	Japanese Restaurant	Ramen Restaurant	Tea Room	Indian Restaurant

```
# Cluster 2 (red)
```

	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
7	Marina	Harbor / Marina	Coffee Shop	Bank	Lighthouse	Gym	Gym / Fitness Center	Monument / Landmark	Historic Site	Park	Event Space
16	Treasure Island	Harbor / Marina	Music Venue	History Museum	Bus Station	Brewery	Fried Chicken Joint	Beach	Event Space	Eastern European Restaurant	Electronics Store

Let’s use bar chart to visualize number of 1st most common venues in each cluster:



After examining each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories we can create a profile for each group, considering the common characteristics of each cluster:

Cluster 0: Multiple Social Venues

Cluster 1: Harbor / Marina Venues

Results and Discussion

Our analysis shows that there are 41 neighborhoods in San Francisco, and we get 18 of them both have less number of crimes, and not far apart to the tech companies. Then by clustering those 18 neighborhoods into two clusters, we can finally make recommendations to stakeholders who relocate to San Francisco and want to find an optimal place to live.

My Project has many place to be improved, for example, we can use population data to calculate the crime rate instead of number of crimes. We can also use commute time instead of distance. Apartment data can be explored further.

Recommended neighborhoods should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only safety and nearby competition but also other factors taken into account and all other relevant conditions met.

Conclusion

By modeling the distribution of neighborhoods based on their number of crimes and distance to the tech companies we have first identified general neighborhoods that justify further analysis. Then explore those neighborhoods and get the top 10 venues for each neighborhood. Clustering of those locations was then performed in order to create major zones of interest.

Because there are only a few neighborhoods in San Francisco, we can't really get insightful clusters.

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.