Desirable Neighborhoods in San Francisco

Vien Nguyen
March 15, 2019

Introduction

- A couple is moving into San Francisco, and they would like to pick their neighborhood based on several factors. They want to be close to:
  - nearby parks
  - Restaurants
  - public transit
- They would also like the adjacent neighborhoods to be desirable as well.

Problem

- Can we build a model to recommend a neighborhood for the couple to live in, based on their specified criteria?
- Target audience: couples that value living near parks, food, and transit in the city of San Francisco

Data Acquisition

- GeoJSON file available on DataSF
  - San Francisco Neighborhoods as designated by the San Francisco Association of Realtors (SFAR) in August 2010
- FourSquare API
  - Venues

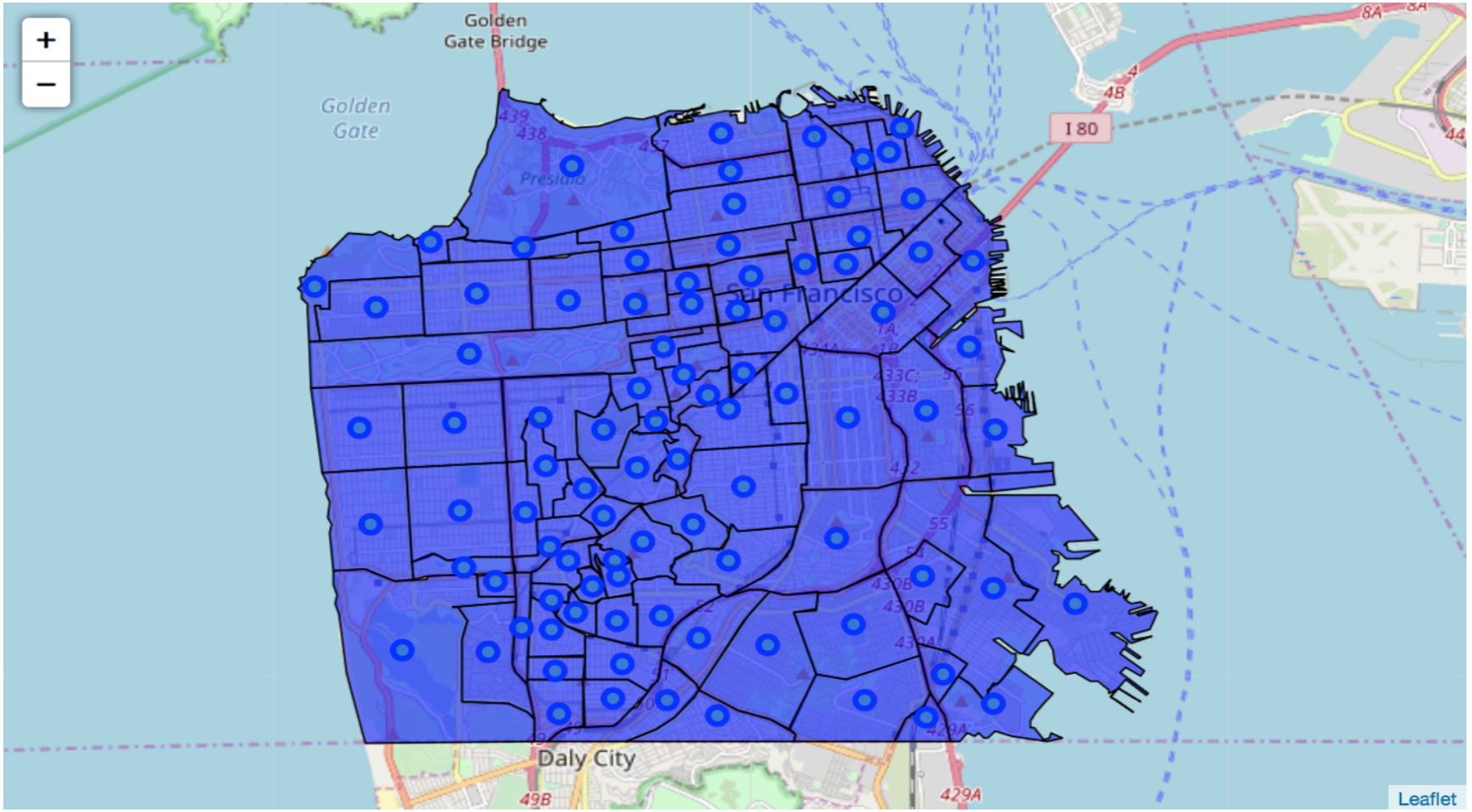
Data Cleaning

- GeoJSON file
  - Parsed into Pandas dataframe
  - Shapely used to determine center
- Neighborhood Name, Latitude, Longitude


```
{
  "type": "FeatureCollection",
  "features": [
    {"type":"Feature","properties":{"nbrhood":"Alamo Square","nid":"6e","sfar_distr":"District 6 -
        Central North"},"geometry":{"type":"MultiPolygon","coordinates":
        [[[[-122.42948394891741,37.775096230704314],[-122.43101153840207,37.77490283132814],
        [-122.43264862289246,37.77469266061221],[-122.43429428444675,37.774475816629221],
        [-122.43594032806337,37.77427667844604],[-122.4376106240872,37.774061646405066],
        [-122.43774490824637,37.774044328330085],[-122.43867122231956,37.778615269308666],
        [-122.43345910855257,37.779277718636891],[-122.4331005636341,37.77750542996215],
        [-122.43157464229915,37.77769992520567],[-122.43005093549058,37.77789364870233],
        [-122.43003059443478,37.77780251540478],[-122.43003083164434,37.777802485238761],
        [-122.42987991643541,37.77705573292902],[-122.42948394891741,37.775096230704314]]]]}},
        {"type":"Feature","properties":{"nbrhood":"Anza Vista","nid":"6a","sfar_distr":"District 4
        Central North"},"geometry":{"type":"MultiPolygon","coordinates":
        [[[[-122.44746439135872,37.77986335309237],[-122.44735192713006,37.780152060669641],
        [-122.44751798255064,37.781069511309816],[-122.44753905927386,37.781510294135381],
        [-122.44730151303307,37.782391642037541],[-122.44279310007408,37.782885336383714],
        [-122.4396173770431,37.783290990343251],[-122.43890760984469,37.779782258762644],
        [-122.44723434672437,37.778717244904861],[-122.44746439135872,37.779863353092371]]]]}},
        {"type":"Feature","properties":{"nbrhood":"Balboa Terrace","nid":"4a","sfar_distr":"Distri
        4 - Twin Peaks West"},"geometry":{"type":"MultiPolygon","coordinates":
        [[[[-122.46450886214802,37.732208495544021],[-122.4650443773425,37.731756190765941],
        [-122.46522807013629,37.731514843545381],[-122.46524536126304,37.731020385476971],
        [-122.46542006656442,37.730168255203161],[-122.46543848657501,37.730094926110674],
```


	Neighborhood	Latitude	Longitude
0	Alamo Square	37.776076	-122.433919
1	Anza Vista	37.780611	-122.443255
2	Balboa Terrace	37.730649	-122.468267
3	Bayview	37.732391	-122.387170
4	Bernal Heights	37.740230	-122.415885

San Francisco Neighborhoods



Foursquare API

- Venue data
  - Name
  - Category
  - Latitude
  - Longitude


```
{'meta': {'code': 200, 'requestId': '5c81f1b51ed2196e48bf5806'},
 'response': {'venues': [{'categories': [{'icon': {'prefix': 'https://ss3.4sqi.net/img/categ
ories_v2/travel/busstation_',
       'suffix': '.png'},
      'id': '52f2ab2ebcbc57f1066b8b4f',
      'name': 'Bus Stop',
      'pluralName': 'Bus Stops',
      'primary': True,
      'shortName': 'Bus Stop'}],
    'hasPerk': False,
    'id': '4f6d0ddae4b0725b60d290e3',
    'location': {'address': 'Van Ness Ave',
     'cc': 'US',
     'city': 'San Francisco',
     'country': 'United States',
     'crossStreet': 'at McAllister St',
     'distance': 103,
     'formattedAddress': ['Van Ness Ave (at McAllister St)',
      'San Francisco, CA 94102',
      'United States'],
     'labeledLatLngs': [{'label': 'display',
       'lat': 37.779925625368705,
       'lng': -122.4200782149065}],
     'lat': 37.779925625368705,
     'lng': -122.4200782149065,
     'postalCode': '94102',
     'state': 'CA'},
    'name': 'MUNI Bus Stop - Van Ness & McAllister',
    'referralId': 'v-1552019893'},
   {'categories': [{'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/travel/bussta
tion_',
       'suffix': '.png'},
      'id': '52f2ab2ebcbc57f1066b8b4f',
```

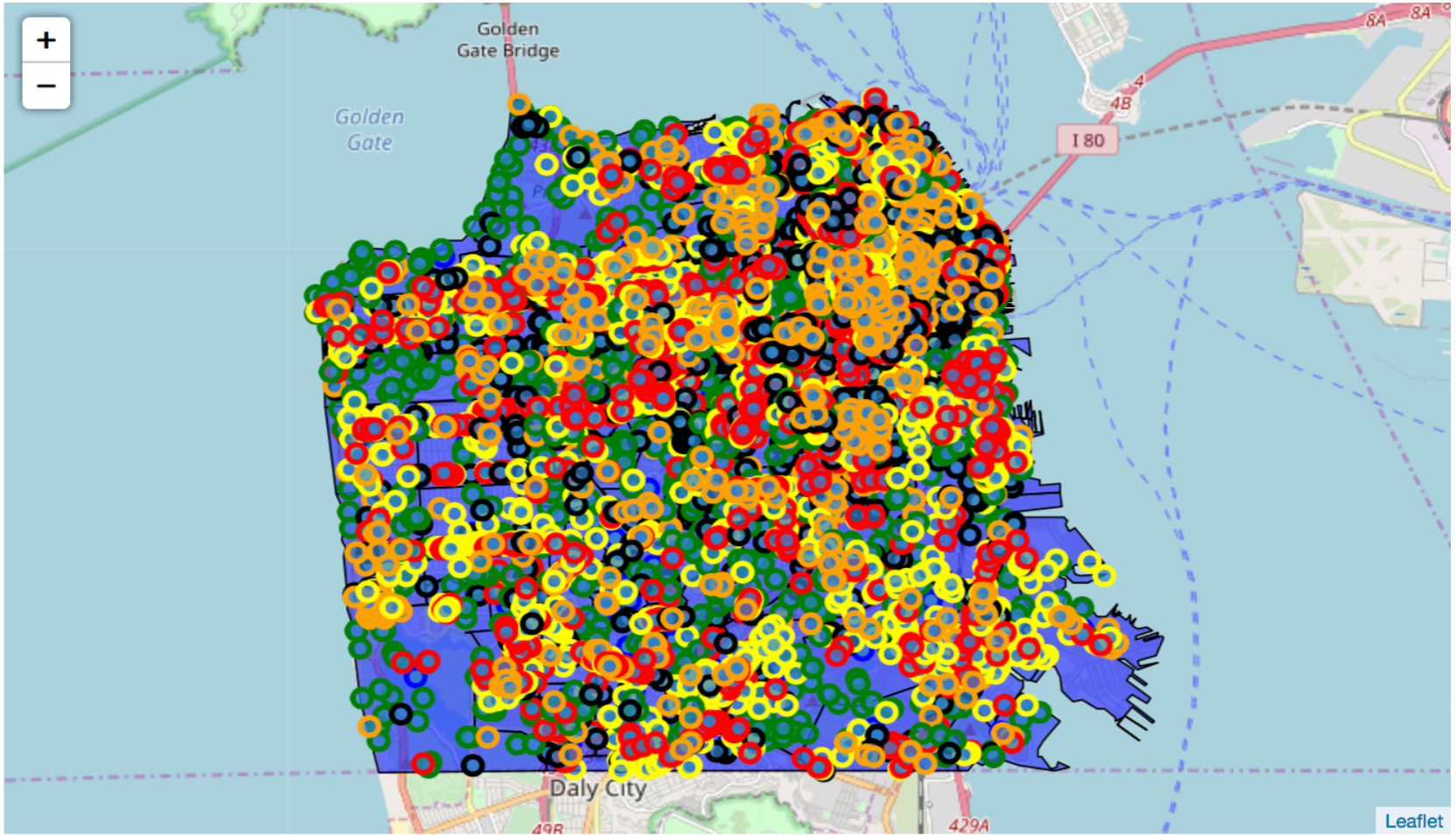

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Master Category
0	Alamo Square	37.776076	-122.433919	Alamo Square	37.776062	-122.433622	Park	Outdoors & Recreation
1	Alamo Square	37.776076	-122.433919	Alamo Square Dog Park	37.775878	-122.435740	Dog Run	Outdoors & Recreation
2	Hayes Valley	37.776076	-122.433919	Yoga Garden	37.771982	-122.437107	Yoga Studio	Outdoors & Recreation
3	Hayes Valley	37.776076	-122.433919	Duboce Park	37.769458	-122.433013	Park	Outdoors & Recreation
4	Hayes Valley	37.776076	-122.433919	Yoga Tree Hayes	37.776507	-122.425014	Yoga Studio	Outdoors & Recreation

Foursquare API Limits

- 50 results maximum per call
- To get around this, I made API calls for specific categories:
  - **Arts & Entertainment** 4d4b7104d754a06370d81259
  - **Food** 4d4b7105d754a06374d81259
  - **Outdoors & Recreation** 4d4b7105d754a06377d81259
  - **Shop & Service** 4d4b7105d754a06378d81259
  - **Transit**
    - **Bus Station** 4bf58dd8d48988d1fe931735
    - **Bus Stop** 52f2ab2ebcbc57f1066b8b4f
    - **Cable Car** 52f2ab2ebcbc57f1066b8b50
    - **Light Rail Station** 4bf58dd8d48988d1fc931735
    - **Metro Station** 4bf58dd8d48988d1fd931735
    - **Train Station** 4bf58dd8d48988d129951735
    - **Tram Station** 52f2ab2ebcbc57f1066b8b51

Venues around SF

- Results combined for 500m and 2000m radii



Data Processing

- After removal of duplicates:
  - 8,923 unique venues
    - Outdoors & Recreation: 1523
    - Food: 2063
    - Transit: 1823
    - Arts & Entertainment: 940
    - Shop & Service: 2574

Feature Selection

- One-hot encoding method

	Neighborhood	Arts & Entertainment	Food	Outdoors & Recreation	Shop & Service	Transit
0	Alamo Square	0	0	1	0	0
1	Alamo Square	0	0	1	0	0
2	Hayes Valley	0	0	1	0	0
3	Hayes Valley	0	0	1	0	0
4	Hayes Valley	0	0	1	0	0

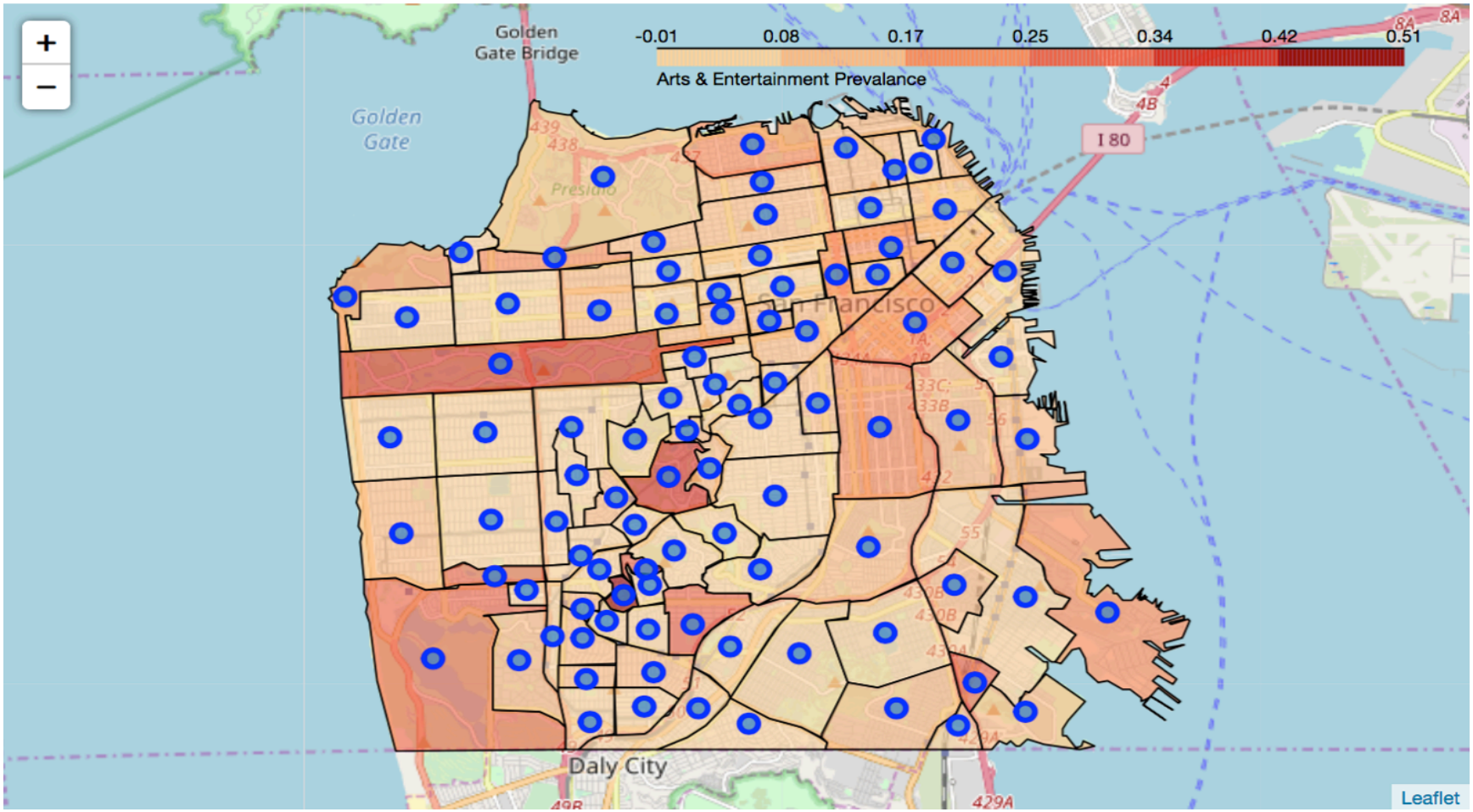
- Determines proportion of each category of total venues

	Neighborhood	Arts & Entertainment	Food	Outdoors & Recreation	Shop & Service	Transit
0	Alamo Square	0.090000	0.260000	0.100000	0.360000	0.190000
1	Anza Vista	0.000000	0.107143	0.000000	0.714286	0.178571
2	Balboa Terrace	0.125000	0.125000	0.000000	0.625000	0.125000
3	Bayview	0.061111	0.244444	0.100000	0.472222	0.122222
4	Bayview Heights	0.272727	0.181818	0.272727	0.090909	0.181818

Exploratory Data Analysis

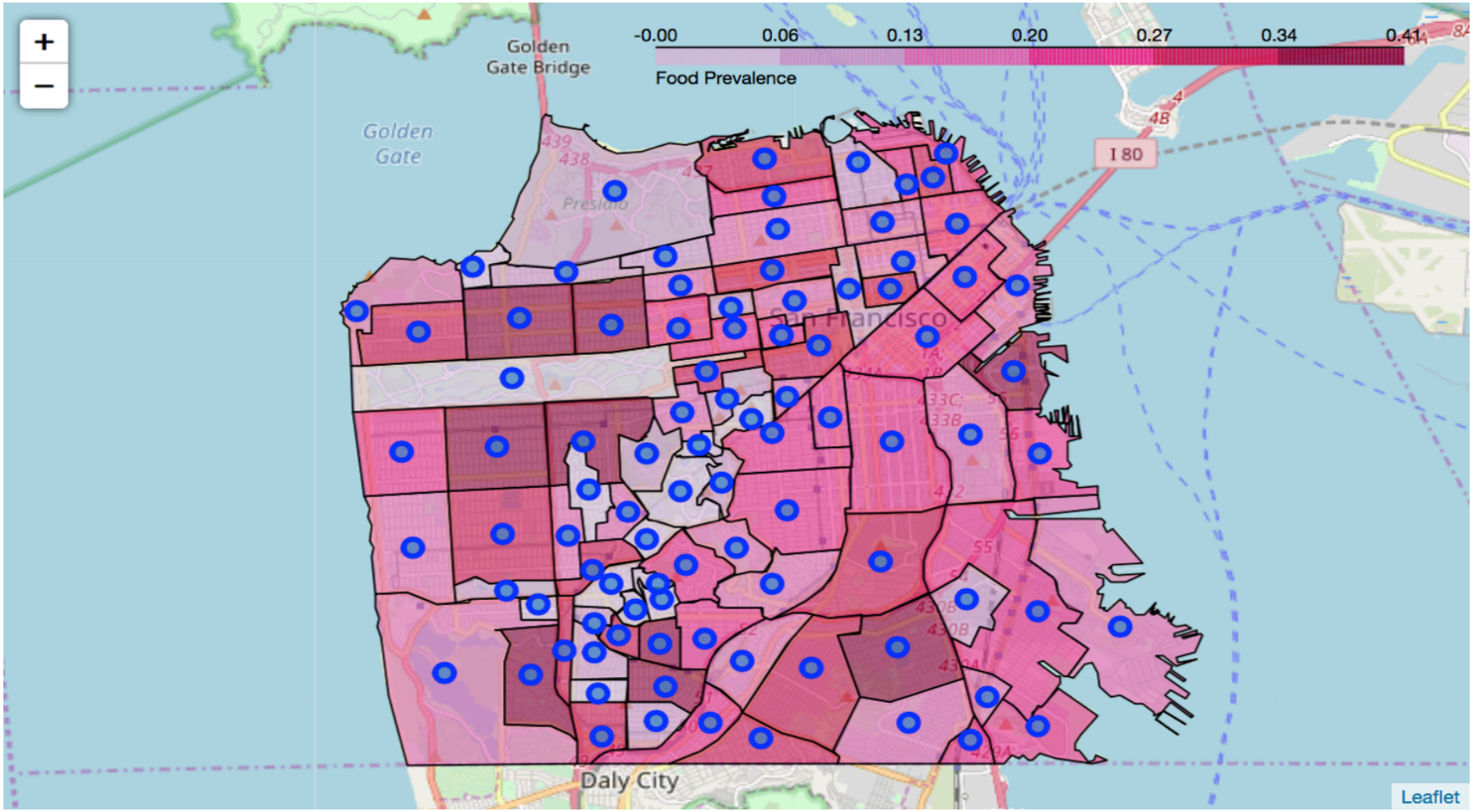
- Choropleths of each category
- K-means clustering
  - Groups similar neighborhoods based on features
- Ranked score
  - Each category is weighted
- Average Adjacent score
  - Mean of Ranked Score for all surrounding neighborhoods

Arts & Entertainment Prevalence



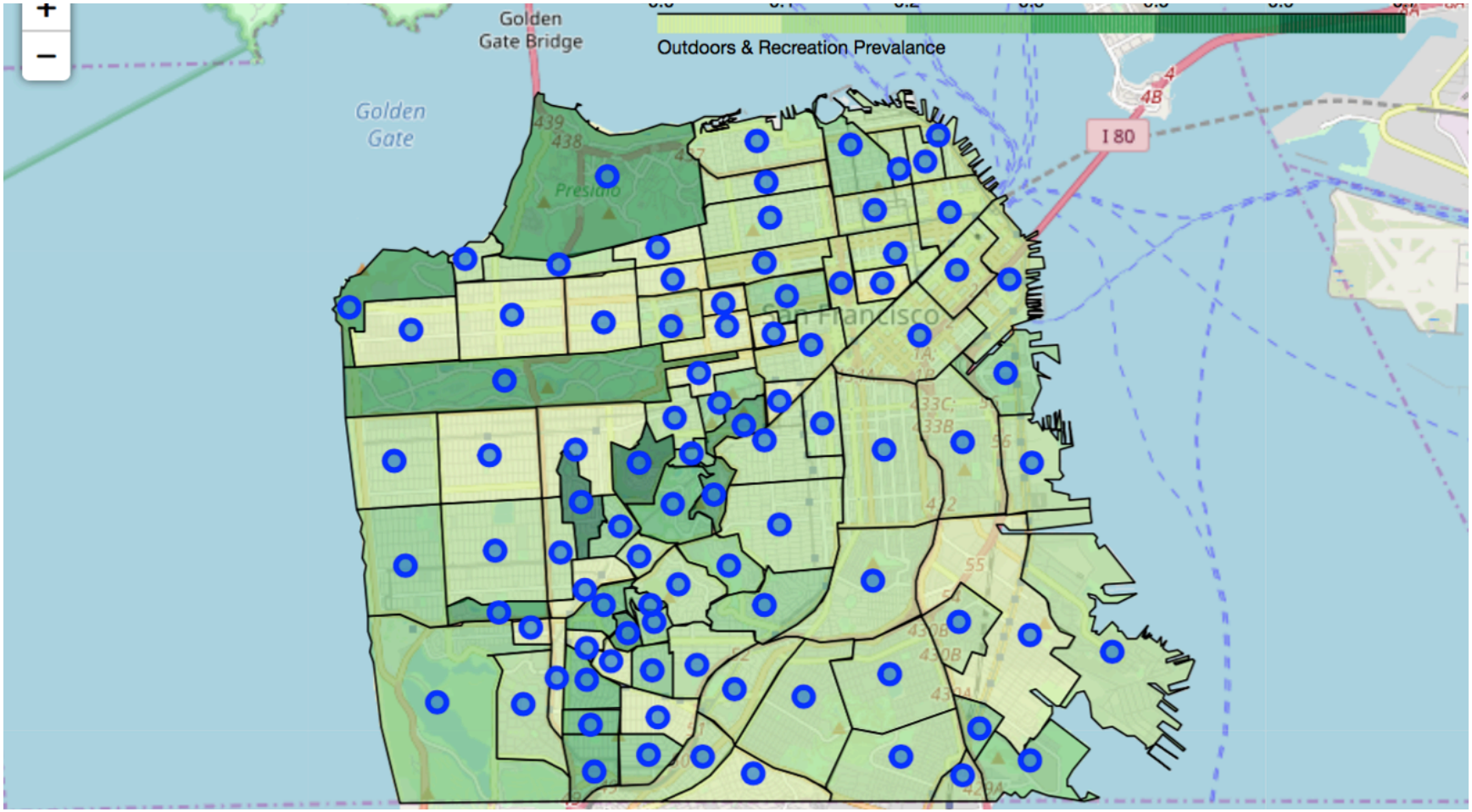
- Monterey Heights, Midtown Terrace, and Golden Gate Park

Food Prevalence



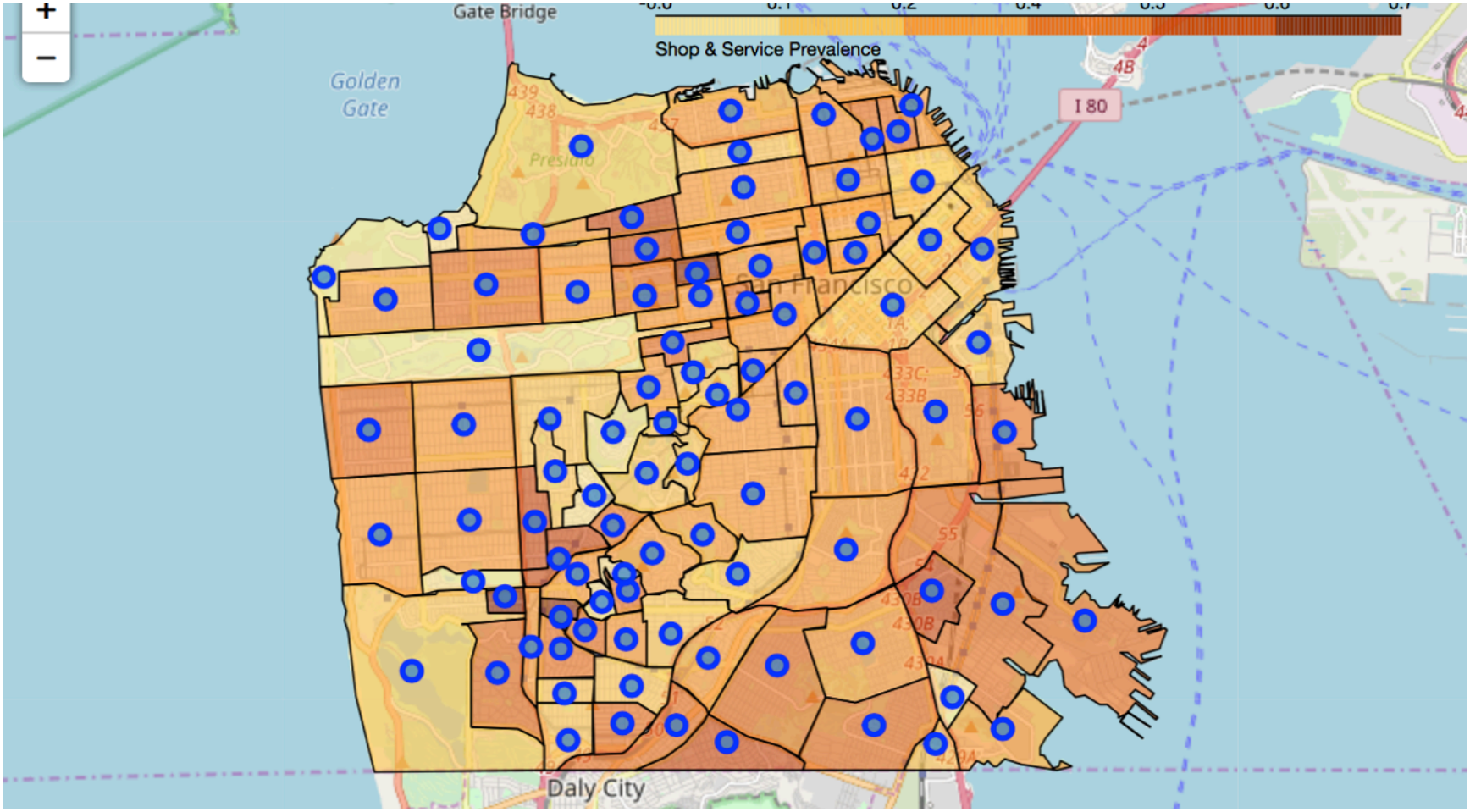
- Central and Inner Richmond, Central and Inner Sunset, Portola, Mission Bay, Stonestown, Westwood Park, and Ingleside

Outdoors & Recreation Prevalence



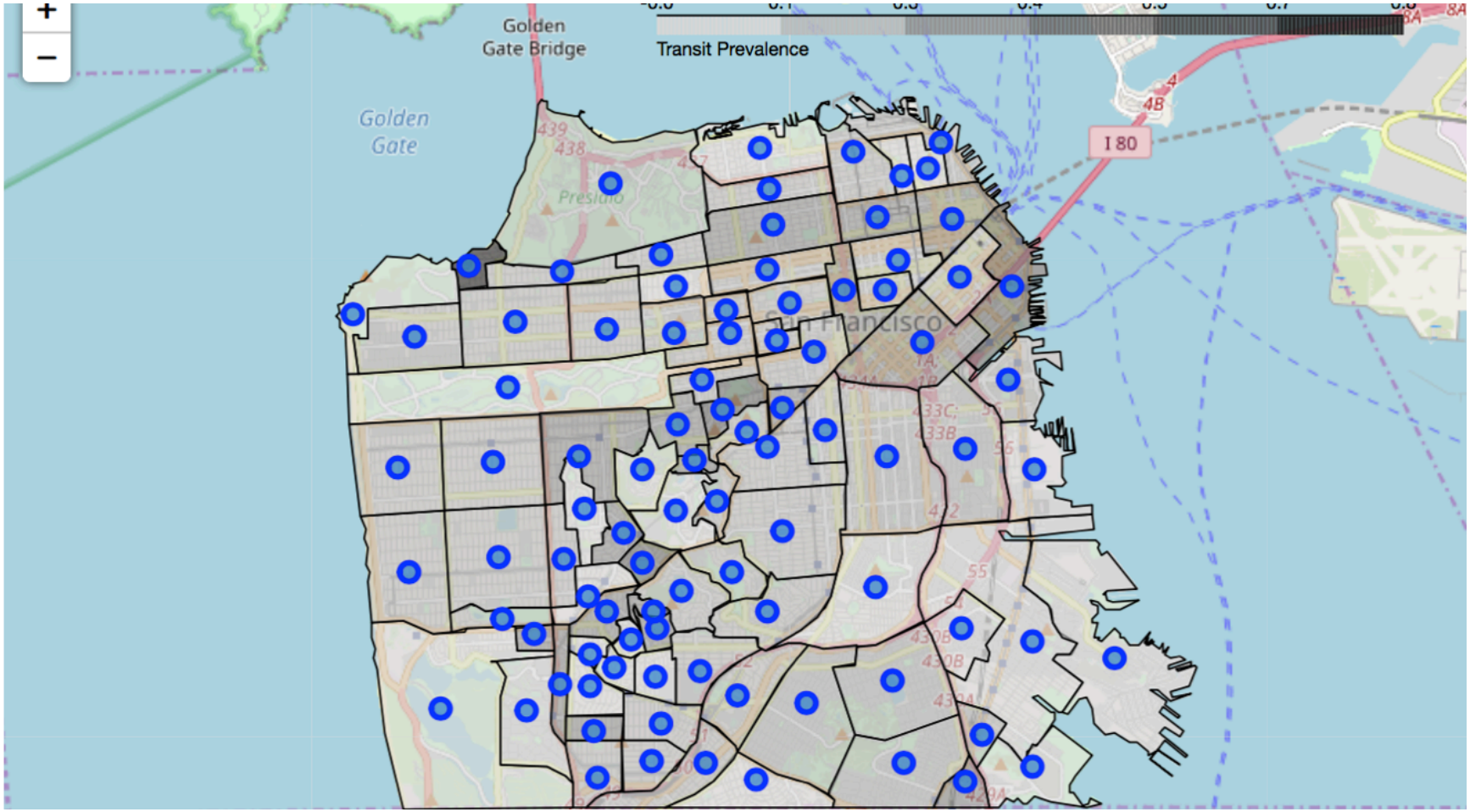
- Forest Knolls and Golden Gate Heights

Shop & Service Prevalence



- Anza Vista, Balboa Terrace, and Merced Manor

Transit Prevalence



- Sea Cliff

k-means Clustering

- This method works by initially randomly placing cluster centroids within the points of the feature set. It then calculates the distance between the centroid and all other points. After each iteration, it moves the centroid closer to the features and recalculates the distances. After enough iterations, the distances between the feature points and the centroids has been minimized enough to consider all the points within a cluster as related.

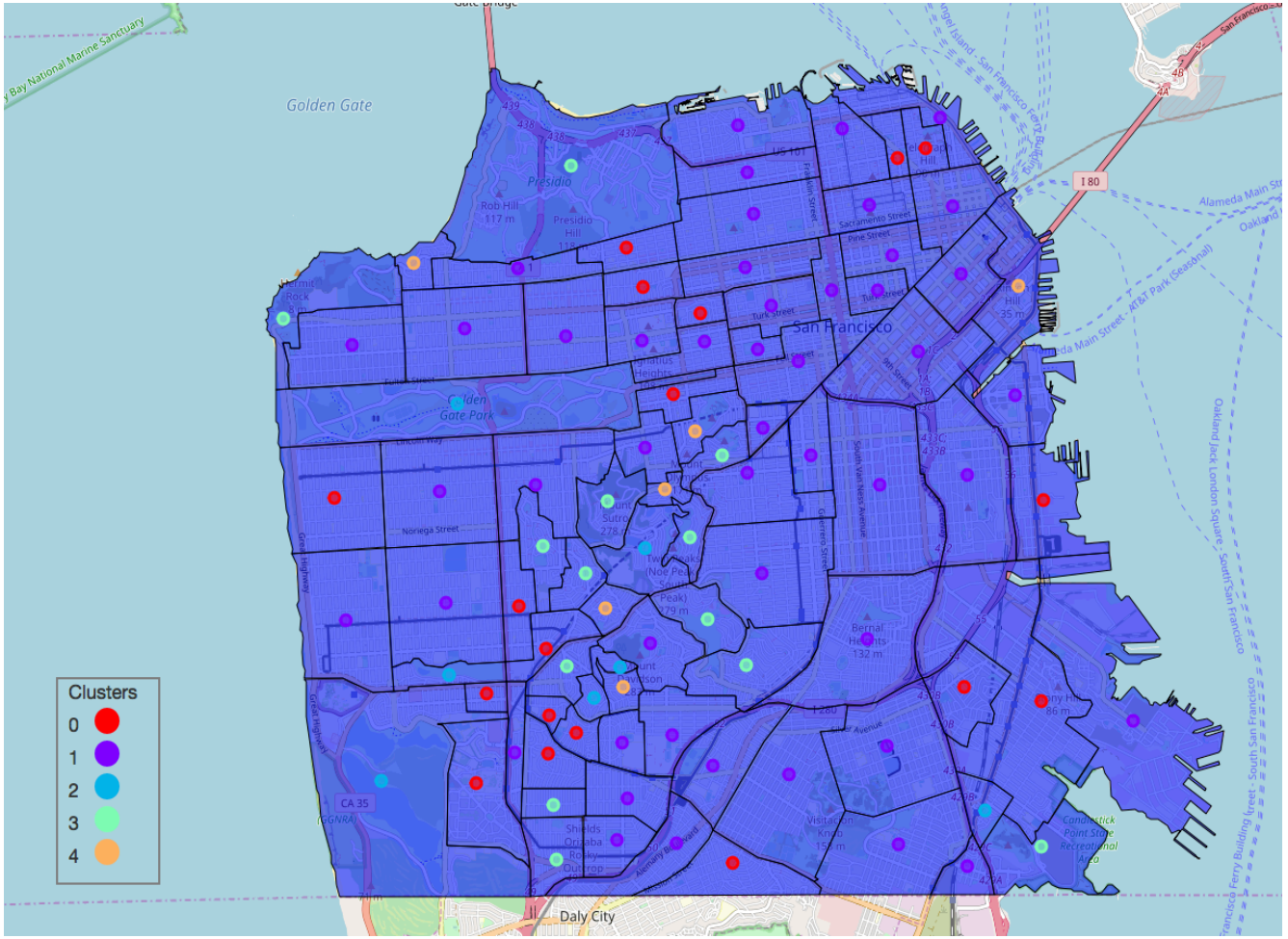
k-means Clustering

- By using k-means clustering, we group like neighborhoods together based on the values of the five categories.

	Arts & Entertainment	Food	Outdoors & Recreation	Shop & Service	Transit
0	0.090000	0.260000	0.100000	0.360000	0.190000
1	0.000000	0.137931	0.000000	0.689655	0.172414
2	0.125000	0.125000	0.000000	0.625000	0.125000
3	0.059783	0.244565	0.103261	0.483696	0.108696
4	0.272727	0.181818	0.272727	0.090909	0.181818

We can eliminate clusters 1 and 2 due to low Outdoors & Recreation

SF Neighborhoods Clustered

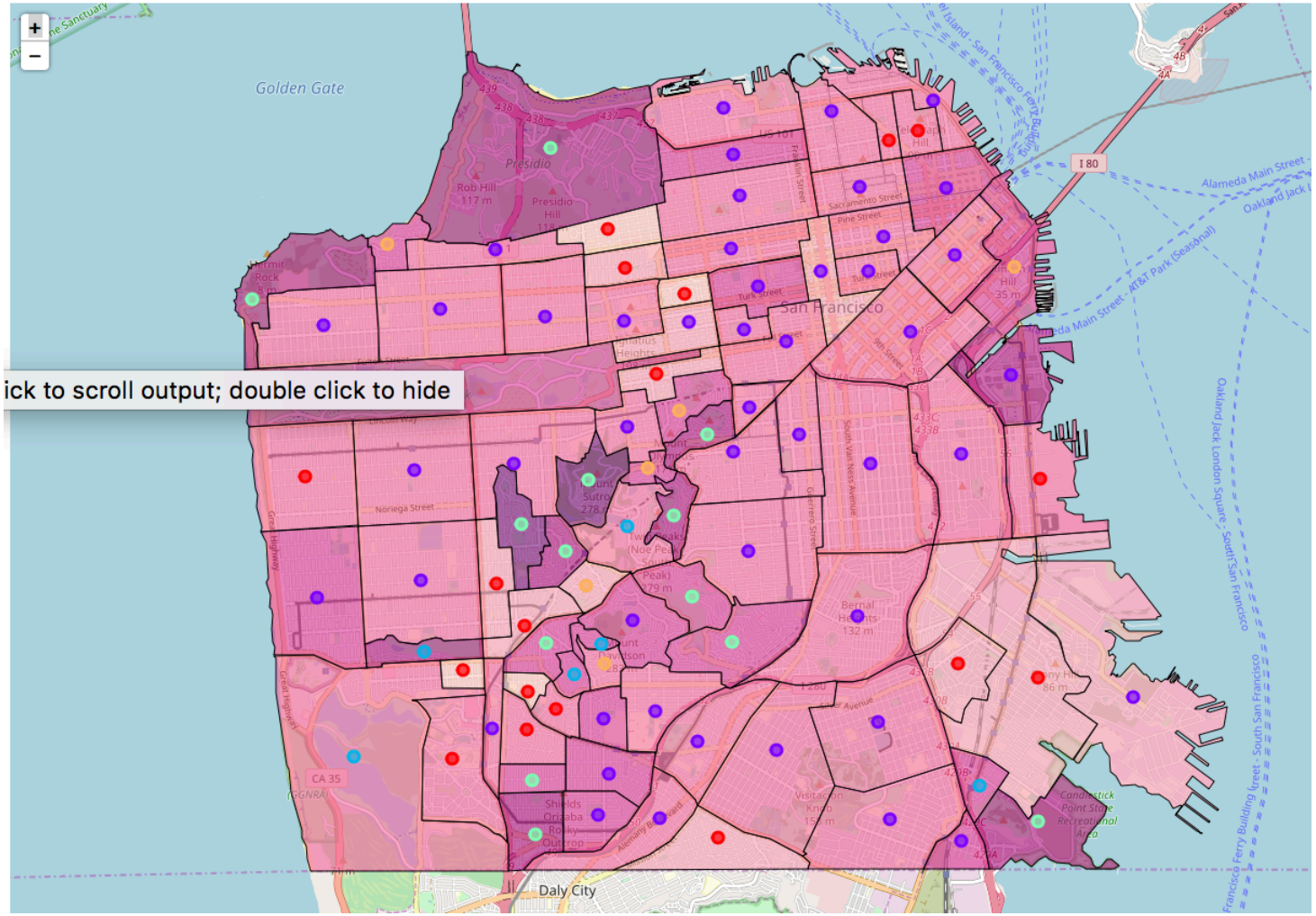


Primary Target Variable

- Ranked Score
  - 40% Outdoors & Recreation
  - 30% Food
  - 20% Transit
  - 5% Arts & Entertainment

	Neighborhood	Arts & Entertainment	Food	Outdoors & Recreation	Shop & Service	Transit	Ranked Score
26	Golden Gate Heights	0.000000	0.117647	0.647059	0.117647	0.117647	0.323529
24	Forest Knolls	0.052632	0.105263	0.631579	0.105263	0.105263	0.313158
32	Ingleside Heights	0.051282	0.307692	0.384615	0.128205	0.128205	0.280769
42	Lincoln Park	0.183673	0.163265	0.489796	0.081633	0.081633	0.274490
84	Twin Peaks	0.086957	0.086957	0.478261	0.130435	0.217391	0.271739

Choropleth of Ranked Score



Note: Clusters are from previous map

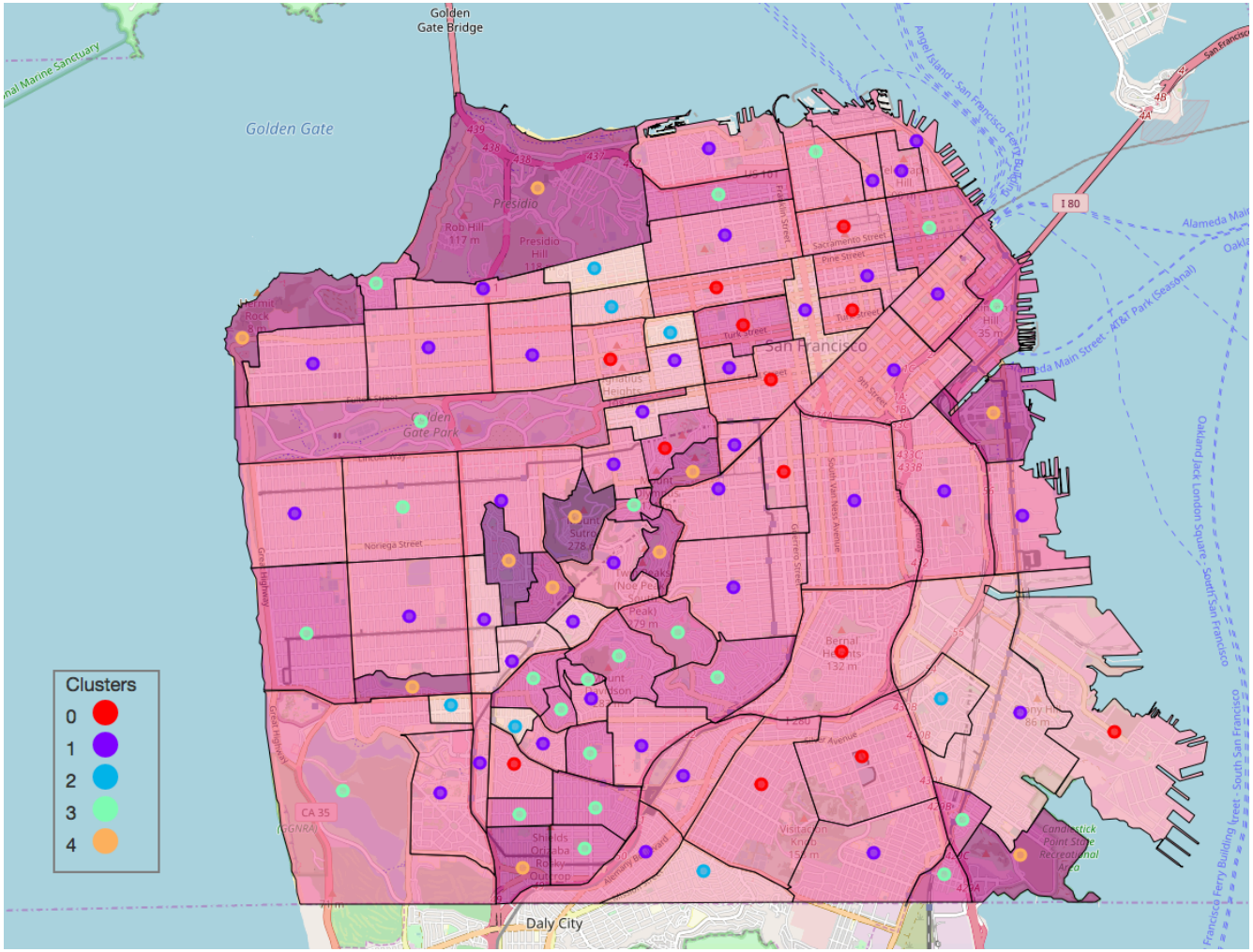
Secondary Target Variable

- Average Adjacent Score
  - Average of Ranked Score for all neighborhoods touching that neighborhood
  - One-hot encoding of whether neighborhoods touch

Neighborhood	Alamo Square	Anza Vista	Balboa Terrace	Bayview	Bayview Heights	Bernal Heights	Buena Vista Park/Ashbury Heights	Candlesti Poi
Alamo Square	1	0	0	0	0	0	0	
Anza Vista	0	1	0	0	0	0	0	
Balboa Terrace	0	0	1	0	0	0	0	
Bayview	0	0	0	1	0	0	0	
Bayview Heights	0	0	0	0	1	0	0	

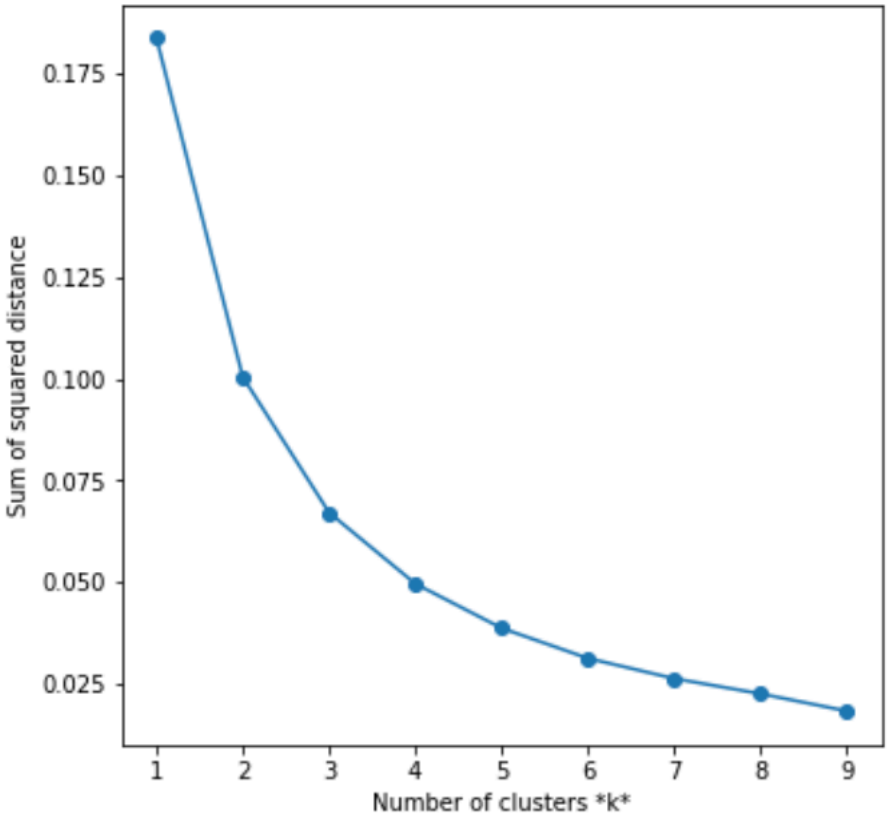
Neighborhood	Average Adjacent Score
Alamo Square	0.187764
Anza Vista	0.175835
Balboa Terrace	0.190930
Bayview	0.197360
Bayview Heights	0.210203

Clustered by Ranked Score



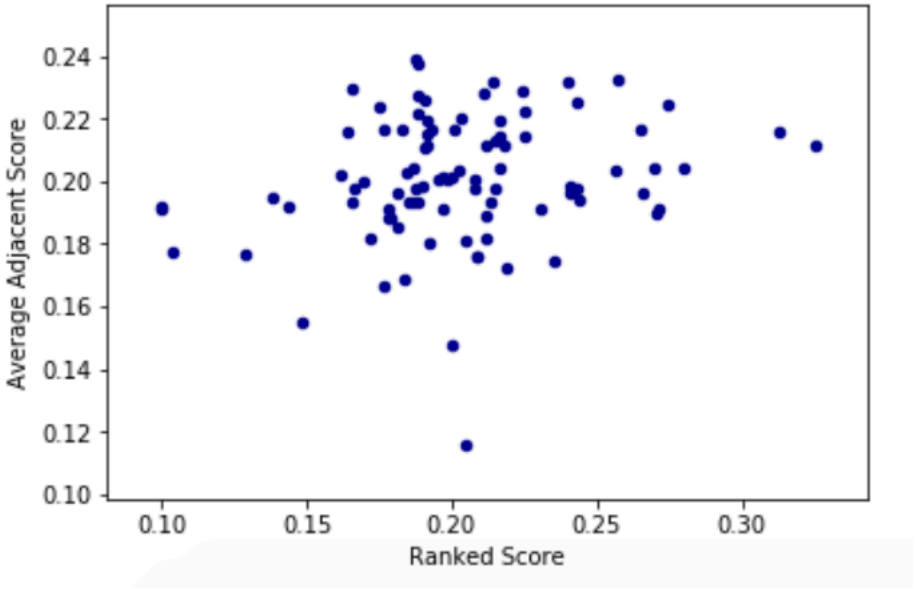
Evaluation of k

- Elbow method
- No clear elbow
- k=5 is fine



Results

Neighborhood	Ranked Score	Average Adjacent Score
Golden Gate Heights	0.325000	0.211571
Forest Knolls	0.313158	0.216164
Ingleside Heights	0.279730	0.204443
Lincoln Park	0.274490	0.224153
Twin Peaks	0.271739	0.191271



- Golden Gate Heights is a strongly ranked candidate

Discussion

- Clear recommendation based on the weights given by the target audience.
- Grouping of similar neighborhoods by features
- Limitations:
  - Foursquare API may exclude some venues
  - Search of each area could be more extensive
  - Other venues such as government, hotels could be added to the analysis
  - Number of venues can be used in a future analysis

Conclusion

- Using GIS data from DataSF and the venue data from Foursquare, I was able to determine which neighborhood would be desirable for a target audience that has a preference for Outdoors & Recreation, Food, and Transit.
- By using k-means clustering, I could group similar neighborhoods, which allows me to eliminate less desirable neighborhoods from the final recommendation.