CAPSTONE PROJECT - THE BATTLE OF NEIGHBORHOODS

OBJECTIVE

- Irvine is the largest city in area in Orange County with 65.61 sq. miles, and 3rd largest in population with 212,375 people according to census in 2010.

- Several corporations, particularly in the technology and semiconductor sectors, have their national or international headquarters in Irvine. A lot of them also have office in Bay Area (this analysis will focus on Santa Clara county). During the last couple of years, I had quite a few friends who relocated to Santa Clara county. One big task in the relocation is to find housing, and neighborhood is an important consideration.

- The purpose of this project is to conduct clustering analysis on the 15 cities in Santa Clara county and find out which one is the most like Irvine.

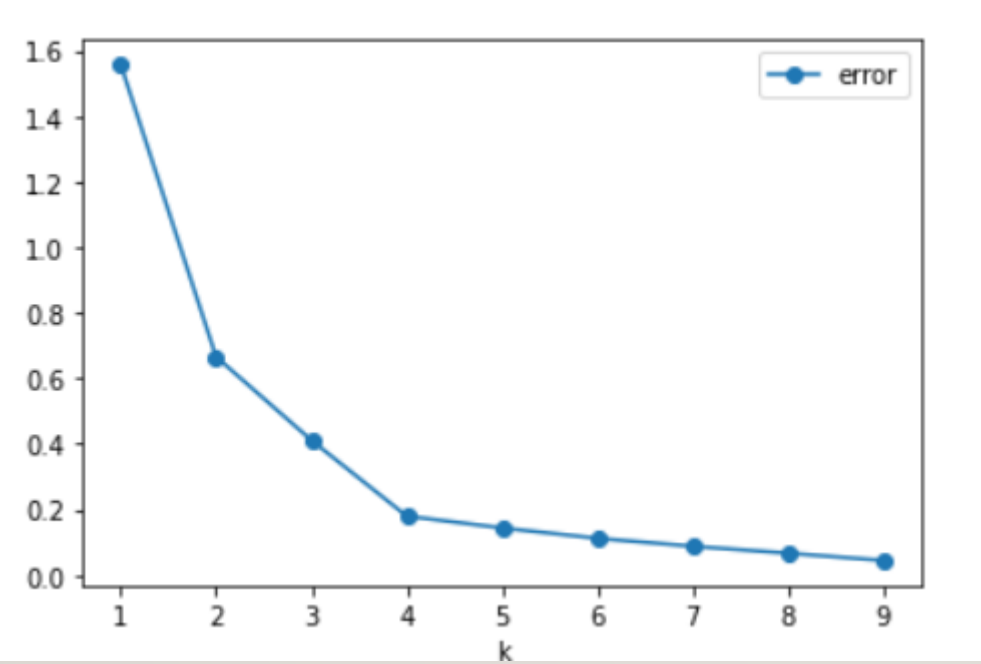
DATA

- Data used in this analysis mainly include:
- Population and demographic related data from Wikipedia pages extracted using web scaping approach;
  - https://en.wikipedia.org/wiki/Santa_Clara_County,_California and pages linked
  - https://en.wikipedia.org/wiki/Irvine,_California
- venue information from foursquare API.

METHODOLOGY

- Extract city names from Wikipedia pages.
- Use Nominatim to get longitude and latitude.
- Pull venue information using Foursquare API.
- Run K-means clusters analysis on the cities and identify groups.

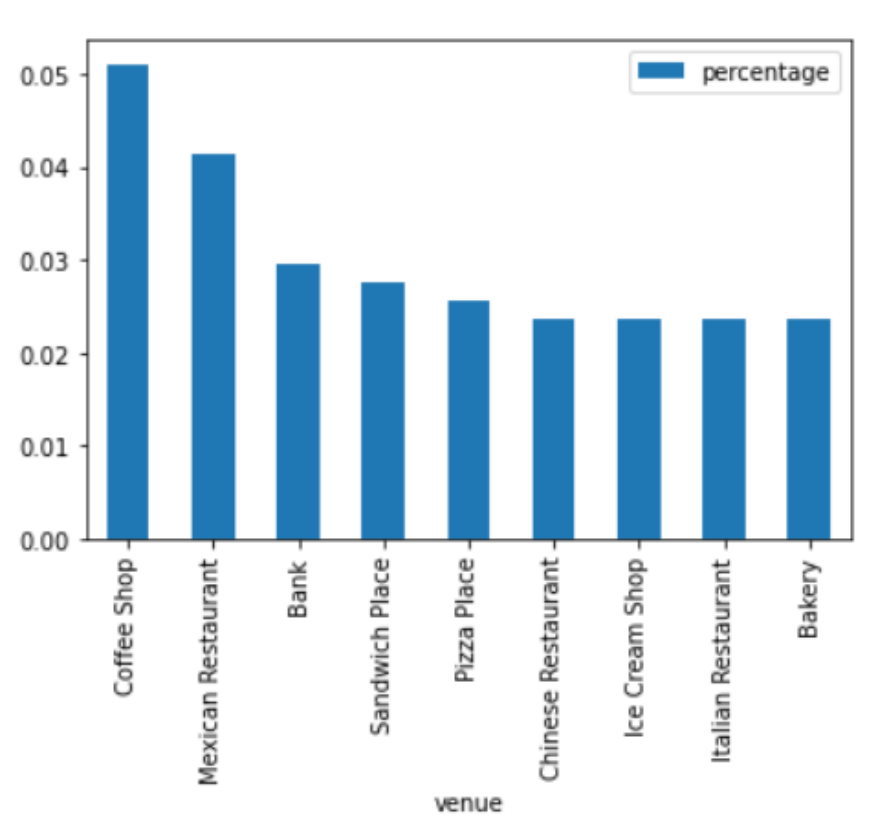
K-MEANS CLUSTERING




```
In [22]: kmeans.labels_
Out[22]: array([1, 1, 3, 0, 1, 1, 2, 1, 1, 1, 1, 1], dtype=int32)
```

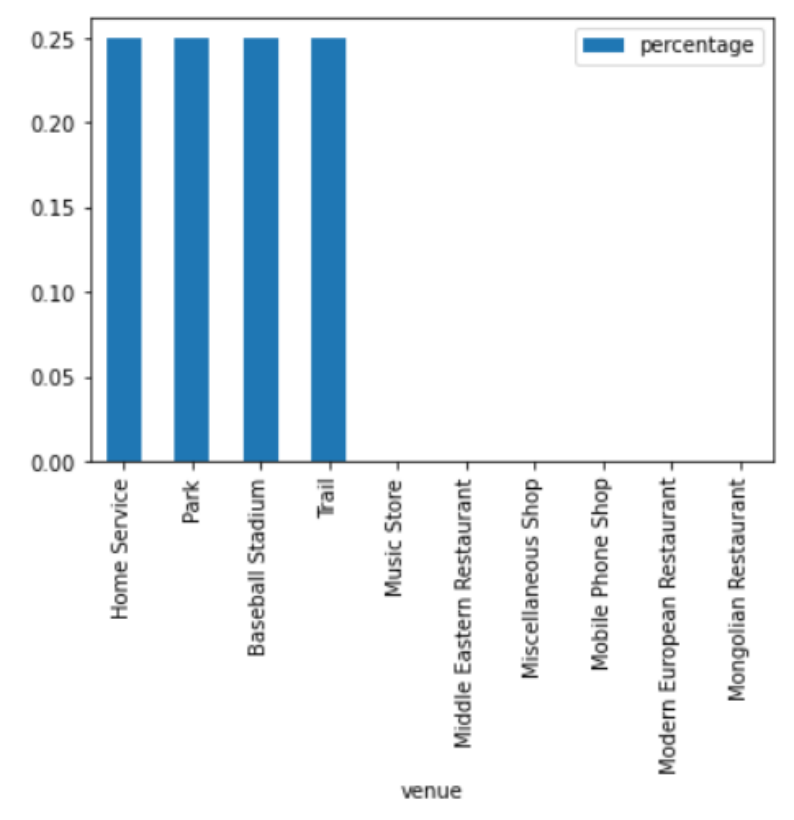
- 1 to 9 clusters are tested, and the slope for error (total distance) flattened significantly after 4.
- 4 clusters will be optimal for this project.
- It appeared that most cities in Santa Clara are similar (group 1), while there are 3 cities significantly different from others and form different clusters by themselves.

COMPARISON



- Unfortunately, Irvine is one of the 3 cities that formed their own clusters.
- The chart on the left shows top venues in cluster 1 (with most Santa Clara cities).
  - Focus more on food and drink shops.

COMPARISON



- Venues in Irvine, on the other hand, are sparsely distributed (only 4 captured in the radius).
  - Focus more on family activities.

CONCLUSION

- Unfortunately, Irvine is one of the 3 cities that formed their own clusters.
- However, the data accuracy is a question, as there are not a lot of venue tracked by Foursquare API in Irvine.
- Further studies will be helpful using different APIs and extract more venue information.