CLUSTER ANALYSIS IN CA CITIES

WENDY WEN

INTRODUCTION

- Scenario: pick a city in CA to move in assumed you do not have financial constraint.
- Approach: unsupervised learning with cluster analysis

DATA

- 3 data source:
  - Big city venue data from foursquare.com by locations:

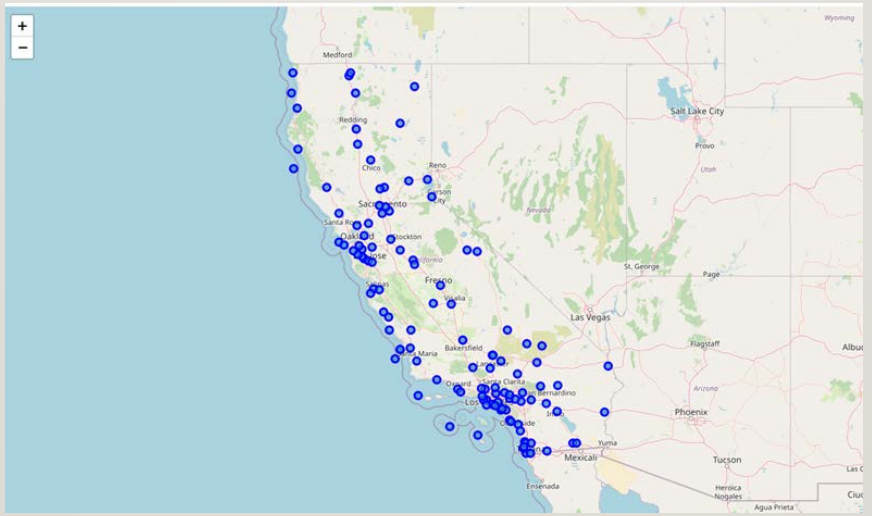
    https://www.w3.org/2003/01/geo/test/ustowns/latlong.htm
  - Crime data:

    https://en.wikipedia.org/wiki/California_locations_by_crime_rate
  - Income data:

    https://en.wikipedia.org/wiki/List_of_California_locations_by_income

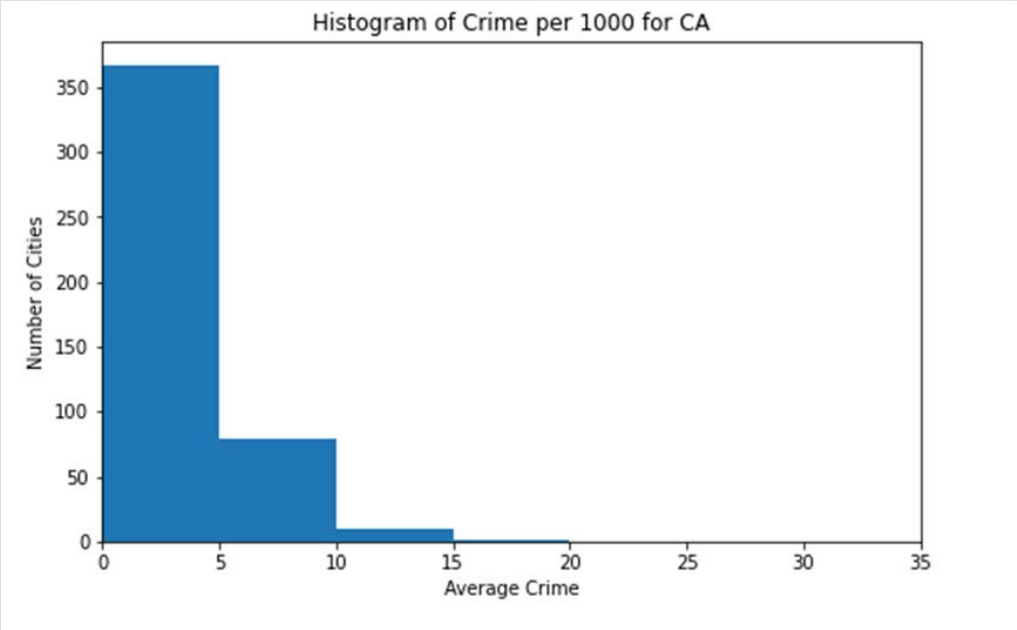
EXPLORATORY ANALYSIS

- Location distribution: all the big cities are included in the analysis
- The location latitude and longitude are used to get the venue data from Foursquare.



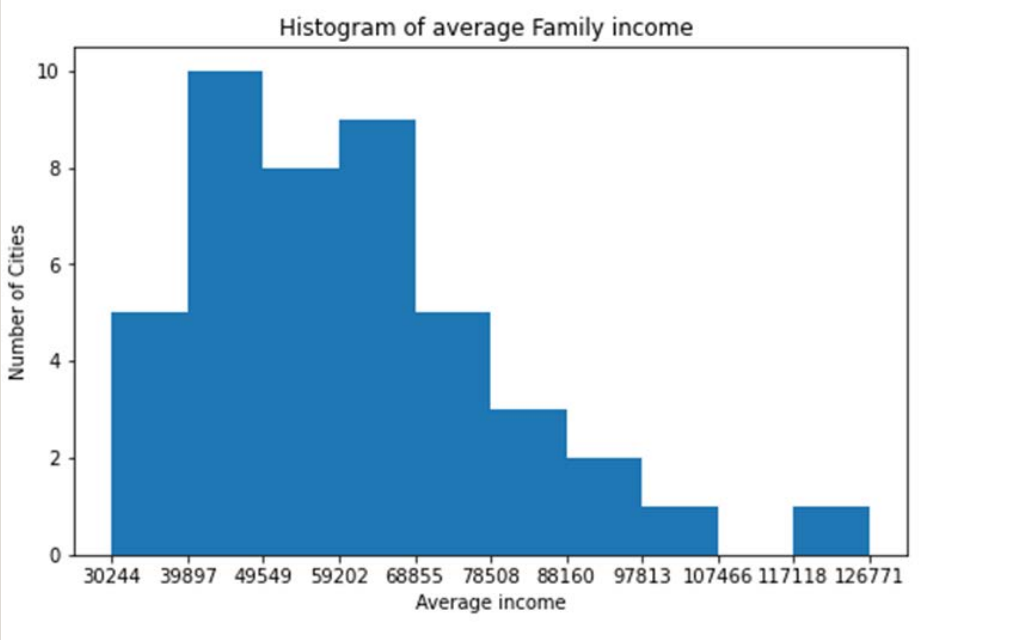
EXPLORATORY ANALYSIS

- Crime rate: The crime rate is not normally distributed, with most cities have low crime rate and a few outliers.



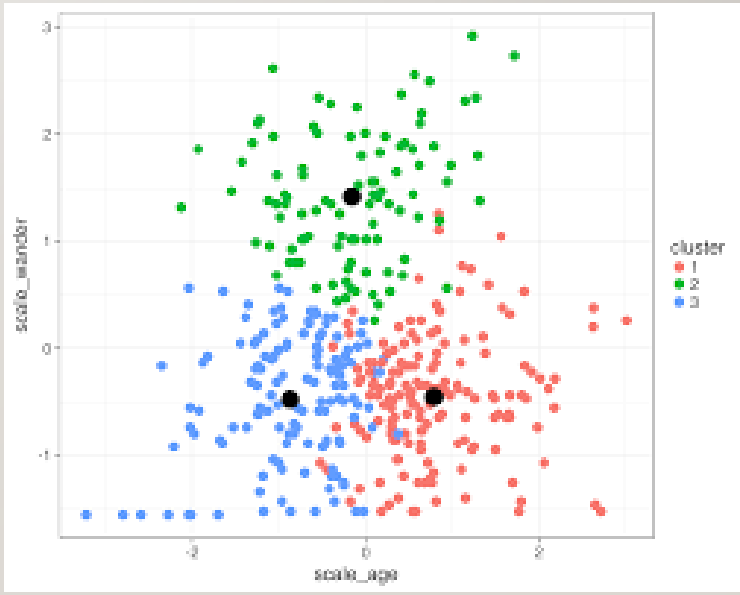
EXPLORATORY ANALYSIS

- Family Income: The family income looks more normally distributed.



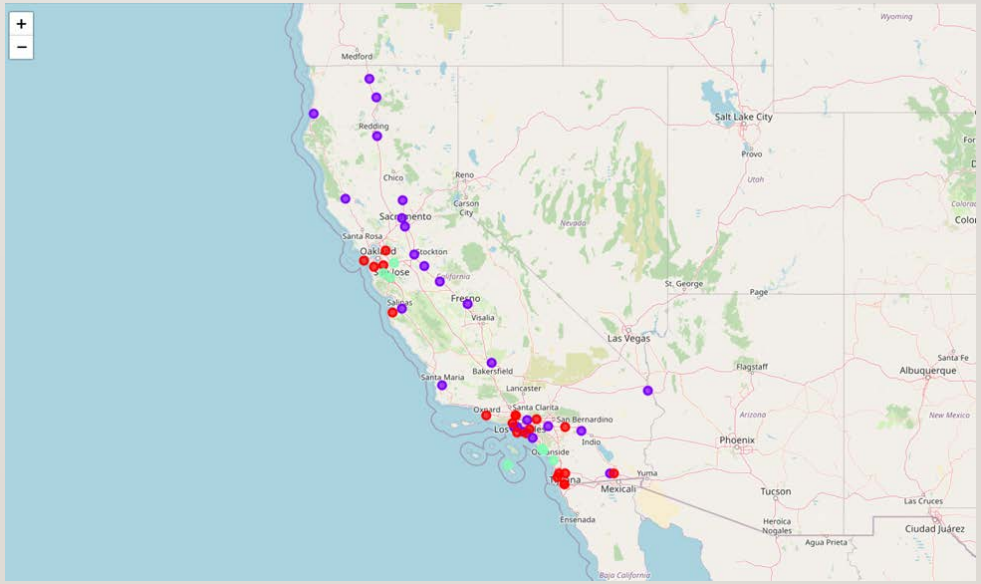
MACHINE LEARNING

- Unsupervised learning algorism.
- K-means cluster analysis was used for all the data.



RESULT AND DISCUSSION

- 3 clusters are formed based on the 3 metrics.



RESULT AND DISCUSSION

- The metrics separate 3 groups well.
- Cluster 2 has the highest income and lowest crime rate.

72]:

Cluster Labels	Asian Restaurant	Gym	Park	Restaurant	Spa	Crime_per_1000	Density	Median_house_income
0	0.000000	0.006579	0.025317	0.002632	0.005196	3.28550	6475.930000	69513.250000
1	0.012013	0.014015	0.000000	0.010417	0.000000	5.60125	4873.120833	45128.458333
2	0.008929	0.006494	0.013605	0.017857	0.000000	1.86000	4238.628571	95891.714286