**Final report for Machines learning project**

**Wendy Wen**

**08/17/2020**

**Introduction:**

Suppose you are with a family and ready to relocate to California, USA. You have never been to CA before, but attracted by the good weather, beautiful environment and good job opportunities. You will need to make a decision of which city you want to move to in CA. Sometimes, we do not have the option to make a choice due to job limitation or family reason. However, we got a chance to choose this time and we can all be interested to this type of decision making.

In this report, we will use some CA city data to cluster big cities in CA. by doing that, we will rank our top cities for relocation candidates. And each of us can pick a city to live in the future.

**Data:**

In this report, we will be using 3 datasets.

First, the CA location data. the data is from web page: <https://www.w3.org/2003/01/geo/test/ustowns/latlong.htm>

In this dataset, it provides the location: latitude and longitude data for all the big cities in CA. It has 4 columns: ID, Latitude, Longitude and Location. It contains all the cities in the US but we only need the CA cities for this projects. The location data will be used to get the venues from foursquare.com

Second: the CA crime data. When we pick a city to live in, the safety will be my first priority. The crime data by cities will be a good indicator of the safety level. You can find the data here:

<https://en.wikipedia.org/wiki/California_locations_by_crime_rate>

The columns we are interested in are city and violent crime rate. The city will be used to identify the location and the crime rate will be used in the model.

Third: the CA income data. The income level will also be a good indicator of the living conditions. I love to live in a city with average higher income. The data can be found here:

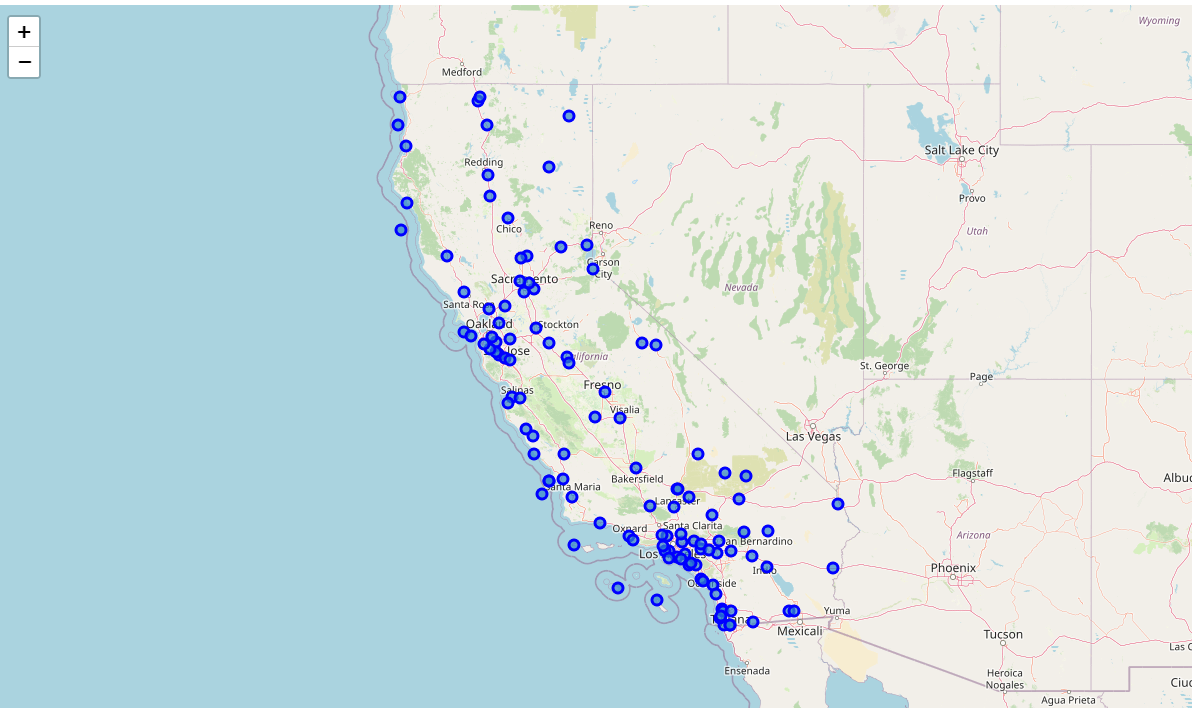
<https://en.wikipedia.org/wiki/List_of_California_locations_by_income>

The columns we are interested in are place and median\_family\_income.

By using these 3 tables, we hope to get a more comprehensive view of the big cities. We can then group them into 3 clusters and look at their average income, crime rate and venue numbers.

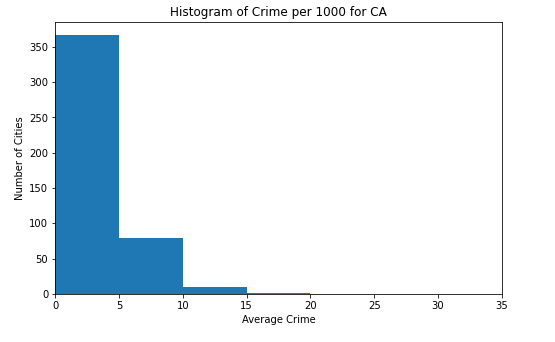
**Methodology** **on data analysis:**

The main methodology for this project is the cluster analysis. we will use K-means cluster approach. Before we do that, we can explore the CA location data, crime data and income data first.

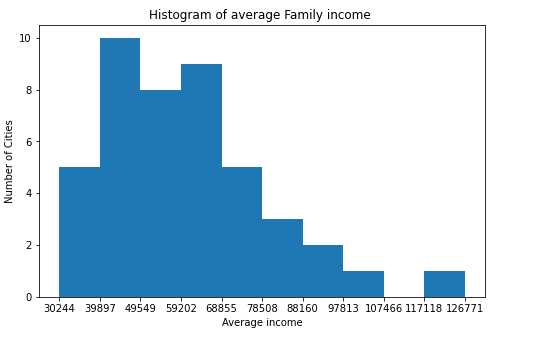


The above is an overview of the cities we start with. We start with 118 big cities in CA. not all of them will make it into the final selection because we may have missing values or not enough information for the clustering process.

Then we did some exploratory analysis on both variables crime and income.



The average crime rate is not normal distributed, which is our expected. Most of the cities have a “crime per 1000” less than 5. But we do see the number can go more than 35 for some cities. Nobody wants to live in a city have relatively high crime rate. This will be a big variable to group the cities.



The average family income is more normally distributed than crime rate. With the income level goes up, the number of cities shrink. We also see one outlier with very high income. That is Palo Alto.

**Methodology on Machine learning:**

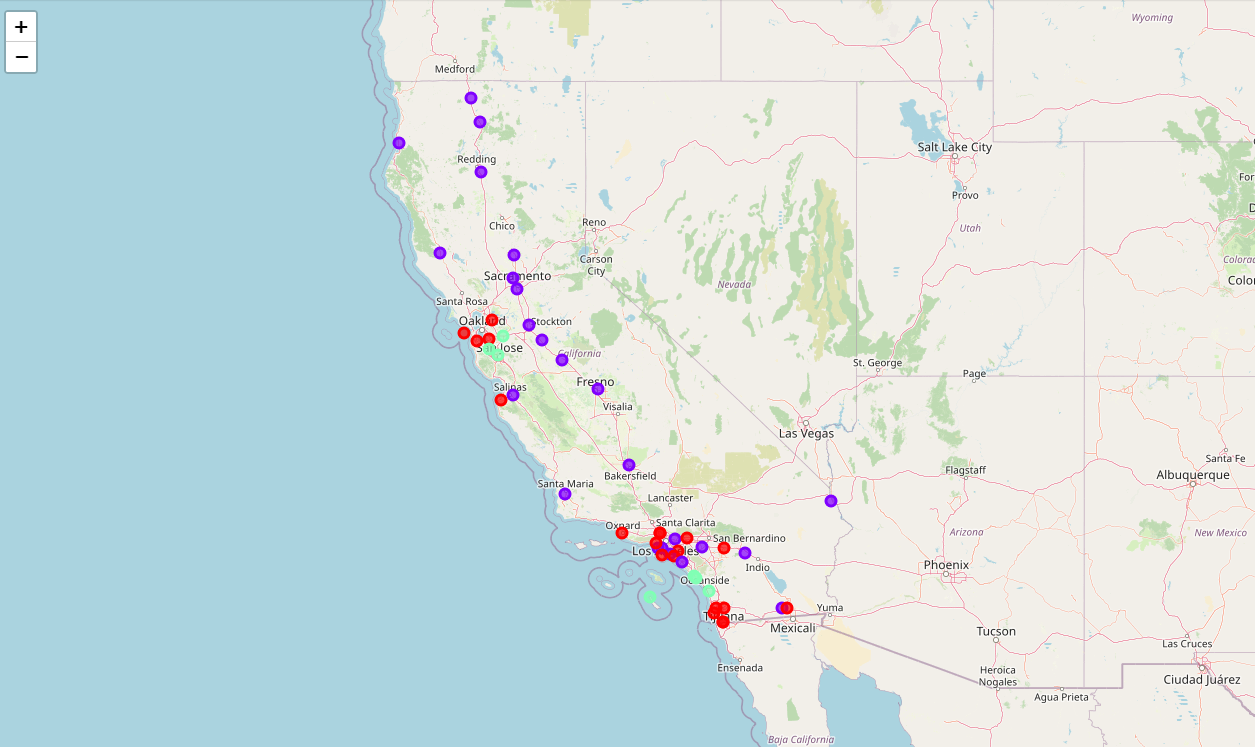
First, the location data is used to get Foursquare venue information through API. The venue information was got through 1000 KM range of the latitude and longitude of the location. I used the “explore” option so it gave me all the surrounding venues of the location. Tustin Mcas has 60 venues, which is the most for all the cities. The data then was got cleaned and sorted out by neighborhood. Top 10 venues were reviewed for each neighborhood.

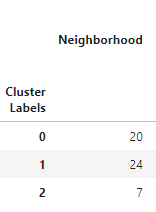
Second, the venue data was joined by the crime data and income data so we have another angle for all the neighborhood.

Last, we use the K-means cluster to do an unsupervised learning. The neighborhood was requested to group into 3 clusters. The purpose is to find the best cities to live in with multiple venues, high income and low crime rate. Please note that this is not a supervised learning, because we really do not know the target variable value in advance. We can not perform a training and learn type of analysis. what we know is all the cities characteristics, we will rely on the cluster algorism to separate those cities based on the venues, crime rate and income level.

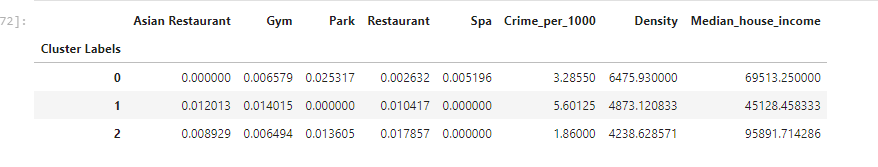
**Result and Discussion:**

Below is a visual result of how the cities got grouped into 3 clusters. In total, we have 51 cities. Remember we start with 118 and about half of them got dropped due to missing values.

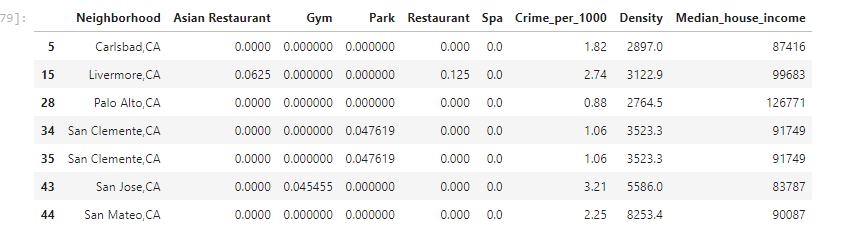




We can see about 20 cities are in group 0 (red), 24 in group 1 (blue) and only 7 in group 2 (green). Let us look at some of the major properties of each cluster:



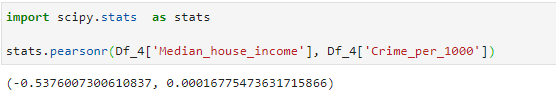
The clusters separately well on the crime rate and family income. You can see the cluster 2 has a highest house income level and a lowest crime level. That makes this cluster outstanding than other clusters. For the key venues, it is kind of in the middle. If we want to look at what is the cities in the cluster 2:



There are 7 cities in this cluster. From the map, we know they are spread out in CA, not only concentrated in one place. That gives us a great list to pick the cities from.

One discussion:

One thing we observe is the Crime rate and family income level seems to have a negative correlation among the clusters. I was wondering if that is the case for all the cities. Here I calculate their Pearson correlation and P value:



It looks like they do have significant negative correlation (-0.5). This makes sense also.

**Conclusion:**

In this report, we use cluster the big cities in CA into 3 groups based on the venues close by, crime rate and family income level. The unsupervised k- means machine learning approach was used. After reviewed the result, we feel comfortable about the approach and the clusters separate quite well. If I need to move to CA, I will likely to pick Palo Alto under the assumption that I do not have budget constraint.