Introduction

California is a big state of America. The population of California is about 39.56 million (2018). California is divided into 58 counties. In this capstone project, we suppose that one restaurant investor who want to know where he/she should invest for new restaurant.

There are many factors effect on the decision to open a new restaurant at one place. These factors may be population of area, median of personal income, median of Family income, house price and the number of existing restaurants in the area. If we have this information of each county, we can cluster counties in the clusters different. Analyzing these clusters could help the investor to make decision.

Therefore, in this capstone project, we will collect data of counties from many sources to get enough necessary information. Then, we will cluster the counties of California based on this information (features).

Data

At the beginning, we collect data about population, Income, house price and using Foursquare to get information about venues of each county. For the population and income, we found out one Wikipedia contains this information (https://en.wikipedia.org/wiki/List_of_California_locations_by_income). This information in form a table of HTML. In order to get information about house price, we downloaded one csv file from the web site https://www.zillow.com/research/data/. For venues information, we set data through out the Foursquare API. This data returns in form of json file.

Method

1. **Preparing Data**
   - **1.1 Population and Income data**
     Using function read_html of dataframe, we can get the table from the Wikipedia, the column names are changed, data type of monetary columns must be changed. We filter data to keep only information of 58 counties of California
   - **1.2 House Price Data**
     The house price data is read from csv file by read_cvs function, we convert data type and filter to get the house price of 58 counties of California
   - **1.3 Venues Information**
     We merge two data frames obtained in two above steps. Using Geopy library to get geo coordinate of the counties. And then, with name and coordinate of county, we call Foursquare API to get 100 venues in the radius 2000m of the coordinate.
     We filter to keep the venues which is categorized 'restaurant' (with supposition category of venue contains string "restaurant").
     We calculate the number of restaurants for each county as a feature for county

   At the end of the preparation data, we have a data frame contains 58 counties. Each county has properties : population, per_capita_income,family_income, house price, number of restaurants in county. We will base on the properties (features) to cluster the counties.

2. **Clustering**
   1. **Select k**
      We use eblow method to estimate k for K means. We get the best for k=6
   2. **Clustering**

Using scikit-learn, we cluster the counties

3. **Visualize**
   We used matplotlib, seaborn, folium to visualize data.

Analyze the result

1. Per_Captipol_Income is correlated with the House Price
2. Cluster 3 includes the counties which has many restaurants
3. Los Angeles is one cluster seperated with other county
4. The others clusters need anlyze more. Almost counties in this cluster has zero restaurant

Conclusion

In this capstone project, we have collect date from many sources, pre-process data to get the given fearues of each county. Then, The K-Mean Clustering is used to cluster the counties of california into 4 clusters. At the end, we visualize the clusets on the map using Folium library. Given features for clustering need to be evaluated