Applied Data Science Capstone

Coursera - The Battle of the Neighborhoods
Best Place to Retire in California

SOMA EASWARAMOORTHY

Contents

1. Introduction

1.1 Background:

California is a very nice place to live. It has some of the best beaches along it’s 900 miles of Pacific Coast. From Hollywood to Silicon Valley, from Aerospace companies to giant Social Media companies it has everything an aspiring young person would love to enjoy. It has a diverse population that adds to fun when you are young. It is vibrant and feels like it has more economic ups and downs than the rest of the world. As you get older and start thinking about life after retirement you start to think of stability and familiar faces. As a resident of California, I and many of my friends and family wonder, what is the best County to retire?

1.2 Problem Statement:

In this project we will try to find a good place to **retire in California**. Ideally a place that has diverse population and lots of **parks and cafes**. Also, with **low crime rate**. As there are more attributes to a county that average person’s mind can handle, we need to utilize the tools of data science to figure this out.

1.3 Interest:

My Friends and Family and everyone I know who is in the same place as I am in my life is starting to think of this problem. I hope this project will show them a new way to look at the problem and find data driven solution for more real-life problems.

2. Methodology

I will try to follow the CRISP-DM methodology by John Rollins to solve this problem.

So far, we have tried to ***understand the problem*** space and defined the problem statement and established the ***(business) requirements*** in the form of a problem statement.

The sheer volume of data and the numerous dimensions of multiple dataset makes it harder to solve with traditional approaches. For this problem we will have to use ***machine learning approach*** and data science tools like IBM Watson Studio, Jupyter Notebook, Python Libraries like Numpy, Pandas, Matplotlib, sklearn and folium.

From the above problem statement (Business Requirements) I have searched multiple source of data and understood the multiple source and defined the ***data requirements*** for this project. As part of planning for ***data collection*** I need to research and read about multiple data sources that are in the public domain. The section 3.1 data sources list the available ***data sources***.

After collecting all the data in one place, they will be merged into single data source. One dataset from Census data uses a name for identifying County while Open Justice uses a County Code for identifying a County. I had to find a meta data mapping data set that maps the names of county to the County Code. With all this I have to ***merge data*** the multiple datasets into a single dataset.

In order to ***understand data,*** we need to reference the documentation available for each of the data source. After going through the documentation available for each of the data source I've understood the ways to do ***data cleanup*** and ***normalize data.*** You will see more details of ***data preparation*** steps in the Jupyter Notbook in subsequent phase of this project.

More specifically about modeling, we are going to use the ***K-Means algorithm*** to Cluster data to find out similar Counties and explore each cluster to choose more appropriate County to retire in.

3. Data Acquisition

3.1 Data Requirements:

In order to find optimal solution for this real-life problem we need to get data from multiple sources.

- We need Demographic information for each County in California
- We need Crime Data about each County in California
- We need to figure out popular venues in each County in California

3.2 Data Sources:

I have identified 4 data sets in CSV format, 1 dataset in Excel format and one JSON data from Foursquare API. The following are data sources to get the data mentioned above

- Demographic information can be obtained from US Census data that can be found in https://data.census.gov

- Crime rate can be determined by combining various data in the Open Justice website. Specifically, they have the following data for each County in California https://data-openjustice.doj.ca.gov
    - **Domestic Violence** related Calls for Service https://data-openjustice.doj.ca.gov/sites/default/files/dataset/2019-06/DVRCA_2001-2018.csv
    - Number of Victims of **Hate Crimes** https://data-openjustice.doj.ca.gov/sites/default/files/dataset/2019-06/HATE_2001-2018_0.csv
    - Number of **Arrests** https://data-openjustice.doj.ca.gov/sites/default/files/dataset/2019-06/OnlineArrestData1980-2018.csv
    - Number of **Violent Crimes** Committed Against Senior Citizens (VCASC) https://data-openjustice.doj.ca.gov/sites/default/files/dataset/2019-06/VCASC_2000-2018.csv

- Finally, the popular venues in each County can be derived from the Venues data obtained from the **FourSquareAPI** . Details of the API endpoints can be found in the official FourSquare documentation: https://developer.foursquare.com/docs/

4. Data Wrangling

Exploration, Understanding and Preparation

4.1 Invalid Rows and Columns

Some datasets came with column headers in the 1st row. In those cases we had to remove the first row.

Some dataset came with data that were for entire United States and we were only looking for data for all Counties in California, so we had to filter out the rows that were outside of California.

Some datasets came with rows for multiple years starting from 2001 to 2018. We decided to derive our answers based on the latest data so we filtered out data that was not 2018.

Some columns were sub set and aggregates. For example, the demographic data was available for Native American's as a whole and also divided into multiple columns of sub groups. However, this data was not available in all cases. We decide to just use the aggregate columns and dropped the columns representing sub groups.

4.2 Invalid Data Elements

Some data elements where values were not available the dataset had "N" and we had to replace it with number 0 (zero).

4.3 Normalizing Data

Columns like race were absolute values we had to convert them to percentages to make sure they are normalized. Columns like Population were absolute numbers we normalized those columns using the Simple Scaling formula.

4.4 Merging Multiple Datasets

After filtering out unwanted rows, after dropping invalid columns, cleaning up invalid data and normalizing it we merged the datasets into one based on the County.

In some datasets the County was a string value and we matched the rows based on the County name. In some datasets the County was a code and we had to find the meta data for converting the County code to County name and then matched the rows.

In the end we had a single row representing all the data that belongs to a County.

5. Modeling and Visualization

5.1 Clustering California Counties

After collecting data from all data sources (Census, Open Justice and FourSquare API) and cleaning it up and merging them we used the K-Means algorithm to cluster all California Counties into five clusters. We used folium library to create a map of California and identify each of the county by color coding the clusters using the following notations:

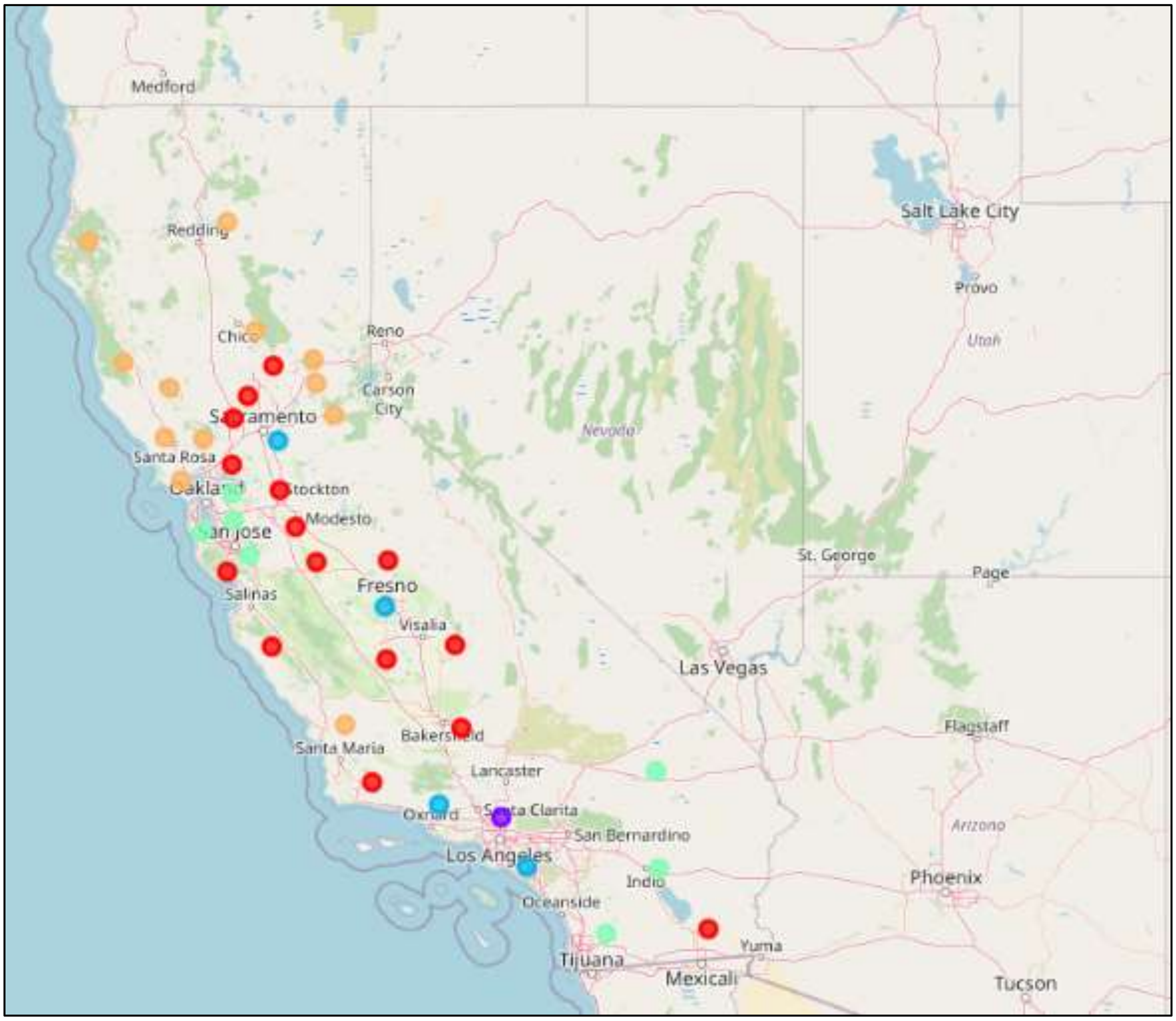
Cluster 0: **(RED DOT)**

Cluster 1: **(BLUE DOT)**

Cluster 2: **(LIGHT BLUE DOT)**

Cluster 3: **(GREEN DOT)**

Cluster 4: **(YELLOW DOT)**

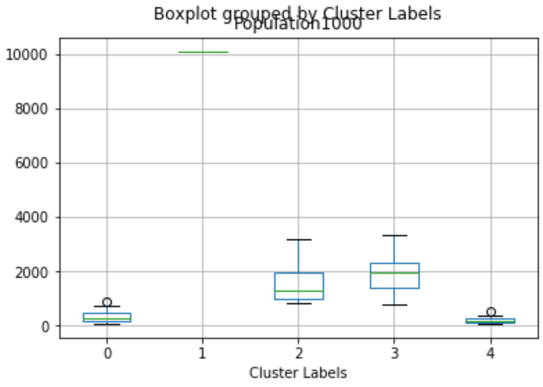


The following table shows all the counties in each Cluster.

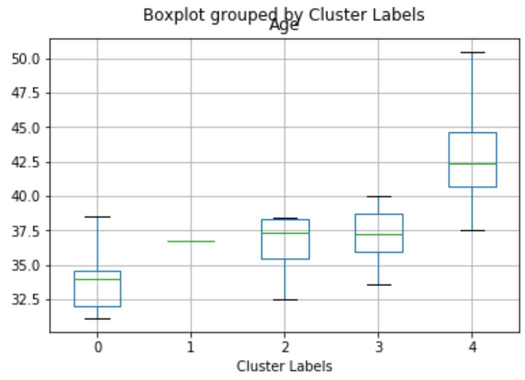
Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Kings County	Los Angeles County	Ventura County	Riverside County	Nevada County
Monterey County		Orange County	Santa Clara County	Shasta County
Yuba County		Sacramento County	Alameda County	Mendocino County
Santa Cruz County		Fresno County	San Bernardino County	Placer County
Stanislaus County			San Diego County	Lake County
Sutter County			Contra Costa County	Sonoma County
Solano County			San Mateo County	Humboldt County
Yolo County				Napa County
San Joaquin County				Butte County
Santa Barbara County				El Dorado County
Tulare County				Marin County
Madera County				San Luis Obispo County
Imperial County				
Kern County				
Merced County				

6. Understanding Results

6.1 How Clusters differ in Population and Median Age

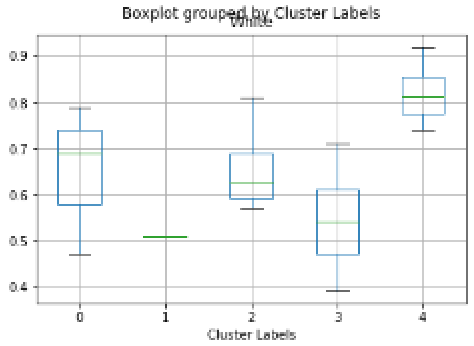


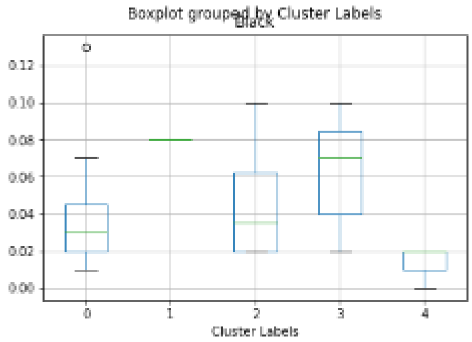
(X-axis: Cluster Labels, Y-axis: Population in 1000s)

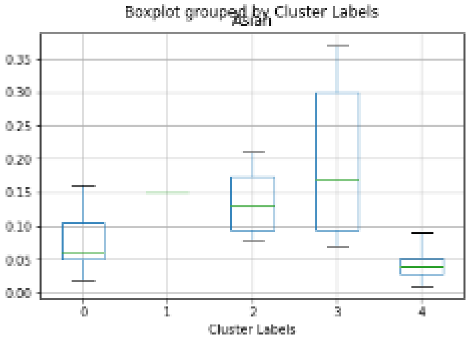


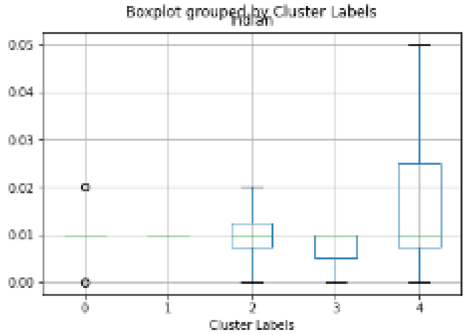
(X-axis: Cluster Labels, Y-axis: Median Age of Population)

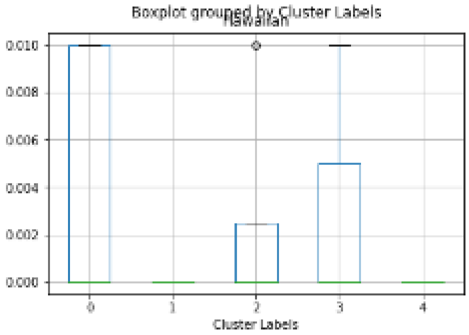
6.2 How Clusters differ in Demographics



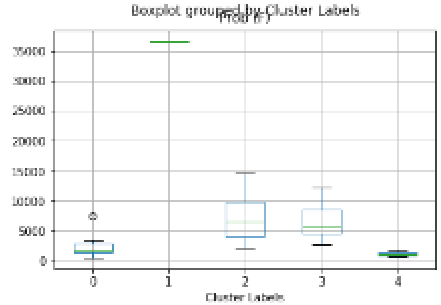


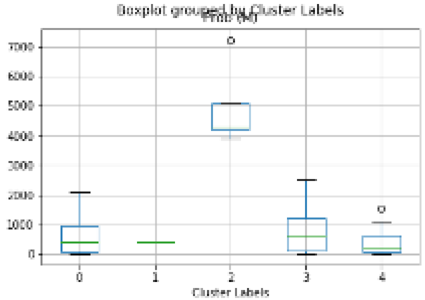


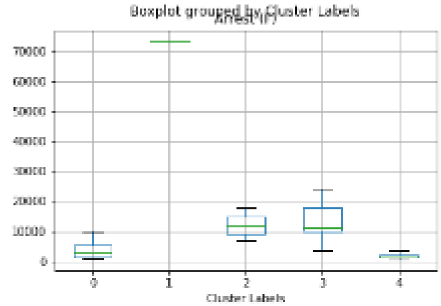


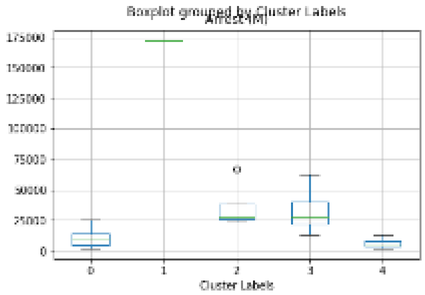


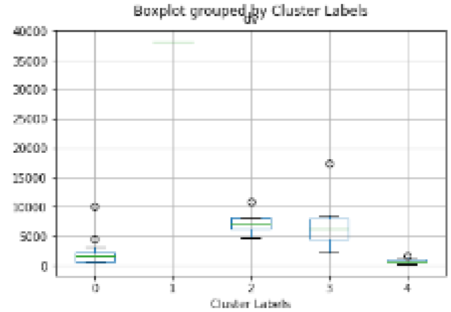
6.3 How Clusters differ in Crime

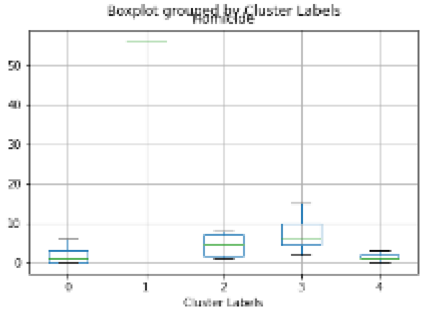


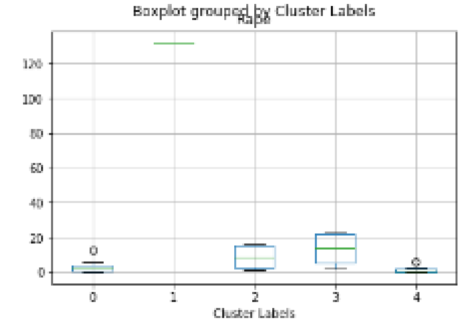


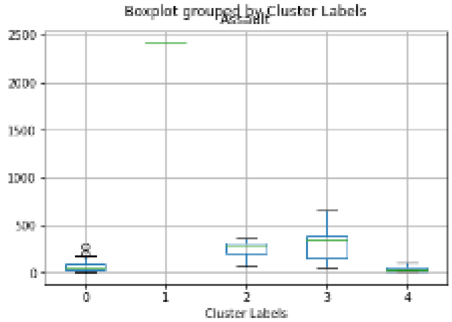








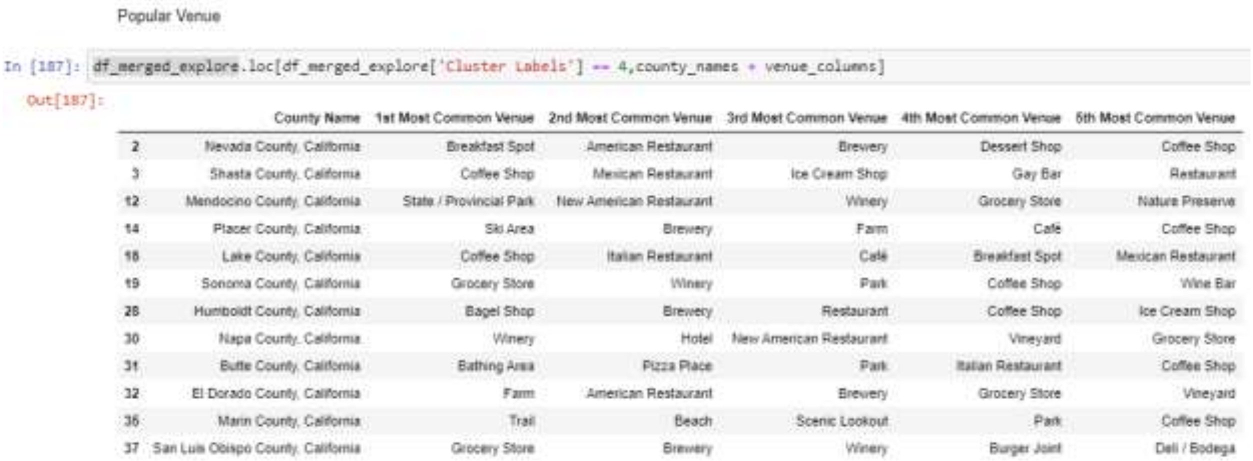




7. Conclusion

Cluster 1 has only one County, that is Los Angeles County. Though Los Angeles has a diverse population, the high crime rate in all kinds of Crime data makes the world-famous Los Angeles as the least preferred place for retiring in California.

Cluster 4 Counties are the ones with reasonable mix of demography. It has the lowest crime rates and that along with the popular places such as Winery, Parks, Trail, Farms and Restaurants makes it the most preferred place to retire. The following are the Counties in Cluster 4 and the popular joints. See the below table and find where you like to retire!



Popular Venue

```
In [187]: df_merged_explore.loc[df_merged_explore['Cluster Labels'] == 4,county_names + venue_columns]
```

Out[187]:

	County Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	Nevada County, California	Breakfast Spot	American Restaurant	Brewery	Dessert Shop	Coffee Shop
3	Shasta County, California	Coffee Shop	Mexican Restaurant	Ice Cream Shop	Gay Bar	Restaurant
12	Mendocino County, California	State / Provincial Park	New American Restaurant	Winery	Grocery Store	Nature Preserve
14	Placer County, California	Ski Area	Brewery	Farm	Café	Coffee Shop
16	Lake County, California	Coffee Shop	Italian Restaurant	Café	Breakfast Spot	Mexican Restaurant
19	Sonoma County, California	Grocery Store	Winery	Park	Coffee Shop	Wine Bar
28	Humboldt County, California	Bagel Shop	Brewery	Restaurant	Coffee Shop	Ice Cream Shop
30	Napa County, California	Winery	Hotel	New American Restaurant	Vineyard	Grocery Store
31	Butte County, California	Bathing Area	Pizza Place	Park	Italian Restaurant	Coffee Shop
32	El Dorado County, California	Farm	American Restaurant	Brewery	Grocery Store	Vineyard
36	Marin County, California	Trail	Beach	Scenic Lookout	Park	Coffee Shop
37	San Luis Obispo County, California	Grocery Store	Brewery	Winery	Burger Joint	Deli / Bodega