Applied Data Science Capstone Project, IBM, Coursera

Investing in a New Eatery in California: A Data Driven Approach

Ritobrata Ghosh (ritobrata-ghosh@outlook.com)

July 21, 2020

Abstract

This report aims to outline an example of data driven decision making through providing advice for investing in new eateries in the California area aided by Machine Learning. Publicly available data has been collected from government agencies, and Foursquare API has been leveraged for information about eateries. After choosing essential features, KMeans Clustering algorithm has been applied and meaningful clusters of California Counties were formed. Advice was provided based on clusters formed by Machine Learning Algorithm.

Keywords
investment, eatery, clustering, analysis, data driven decision

1 Introduction

When an investment firm is looking into investment options in new eateries in a city- California, the option and location to invest in is not an easy decision. A new eatery might find itself in a high competition environment or might face lack of sales- rendering the investment hard or impossible to garner profit.

In this situation, the investors want to take a data-driven approach in which they invest in specific eateries in specific locations which are much more likely to be financially successful. They want to find counties and eatery types which are found by rigorous data analysis, in which investments are more likely to be successful.

This project intends to advise in investment to new eateries in California. There are numerous different kinds of eateries in which investment is possible. such as- Indian restaurant, deli/bodega, dessert shop etc. There are 58 different counties in California. If an investment firm is looking to invest in a new eatery in the California area, it is the objective of the project to deliver them a list of counties and a list of eatery types to invest so these investment becomes quickly profitable.

Any investment firm looking to invest in new eateries in California will be interested in this project. And investors interested in investing in venues which have multiple kinds who want to take data-driven decisions will be benefited from this project.

2 Data Sources

For solving this problem, data from four sources will be leveraged.

2.1 Location Data

Location data titled "California Counties" provided in **California Open Data Portal** provided by **Government of California** for the geographical location data. This data is in .csv format.

2.2 Venues Data

The **Foursquare** **API** for information about established restaurants based on location.

2.3 Population Data

County-wise population data from **US Government Census site.**(File Link). This file is in .xlsx format. Only the latest data (year 2019) has been kept, and it has be turned into a CSV file for further cleaning.[1]

2.4 Economic Data

County-wise Real GDP data provided by **Bureau of Economic Analysis, U.S. Department of Commerce**. (File Link). This data is also in .xlsx format. Irrelevant data has been truncated and the file has been converted to CSV format for further cleaning.

3 Methodology

3.1 Exploratory Data Analysis

3.1.1 Location Data

The location data has 58 rows for California's 58 counties. And it has 3 columns- one for the counties' names and one each for the latitudes and longitudes. The centers of California's 58 counties were visualized.

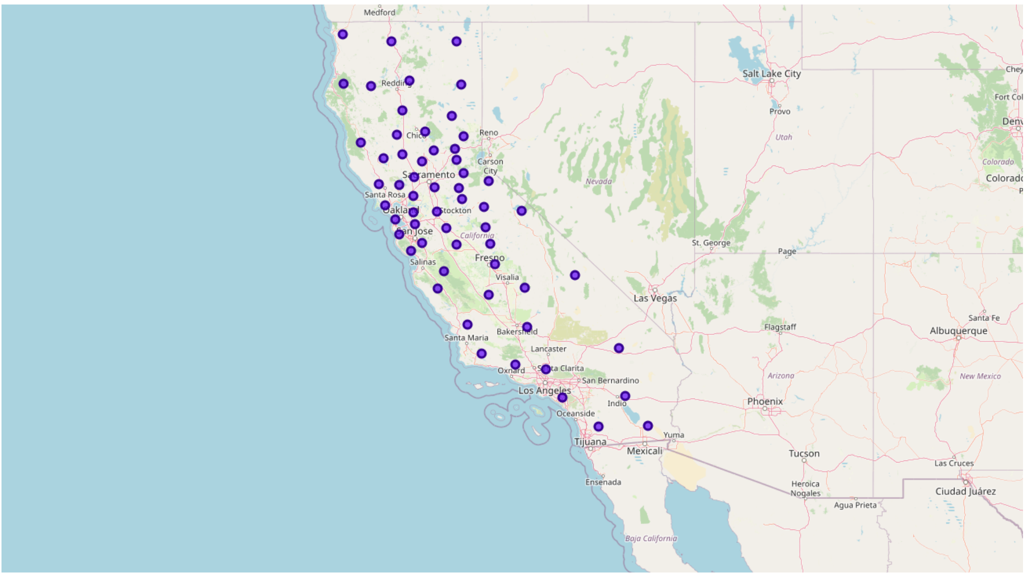

Figure-1: Location of County Centers on California

3.1.2 Population Data

The population dataframe has 58 rows and two columns for names of the counties and their respective population. Here Los Angeles County has, by far, the highest population.

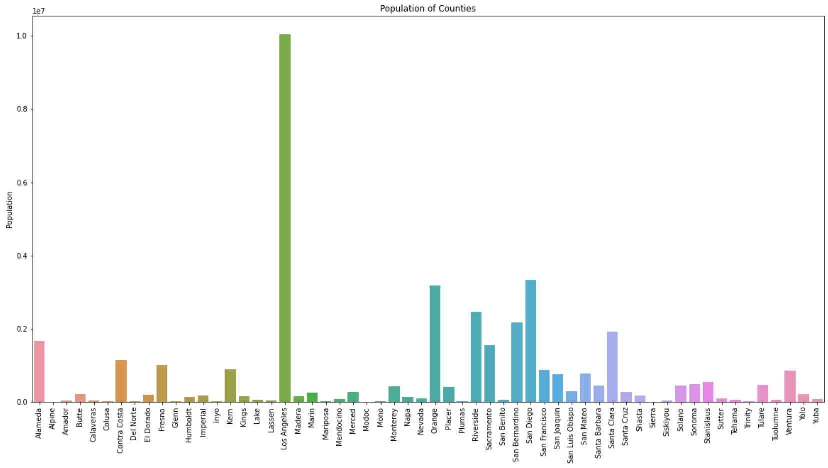

Figure-2: Distribution of Population in California Counties

3.1.3 GDP Data

The GDP dataframe has 58 rows and two columns for names of the counties and their respective GDPs. Here, too, it is clearly visible that the Los Angeles county has the highest GDP.

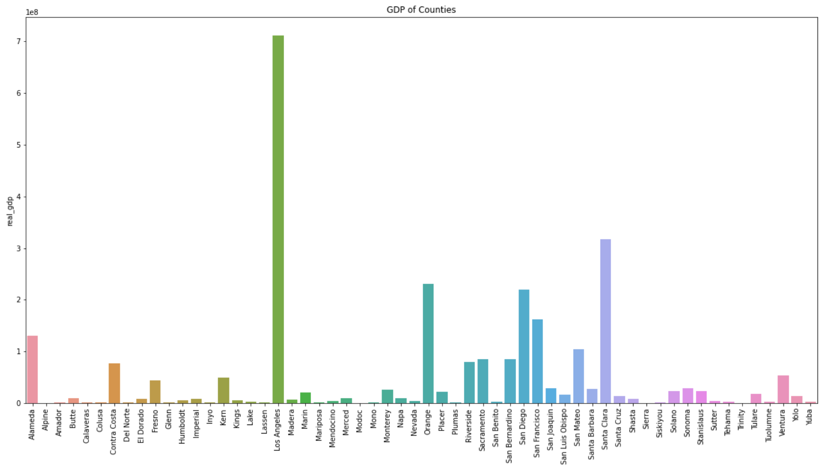

Figure-3: Distribution of GDP in California Counties

3.1.4 Foursquare Data

With an API call to Foursquare for all venues within 10 km of the county center categorised as 'Food' retrieved a list of 1329 venues in all over California. But number of venues was capped at 50 for all venues.

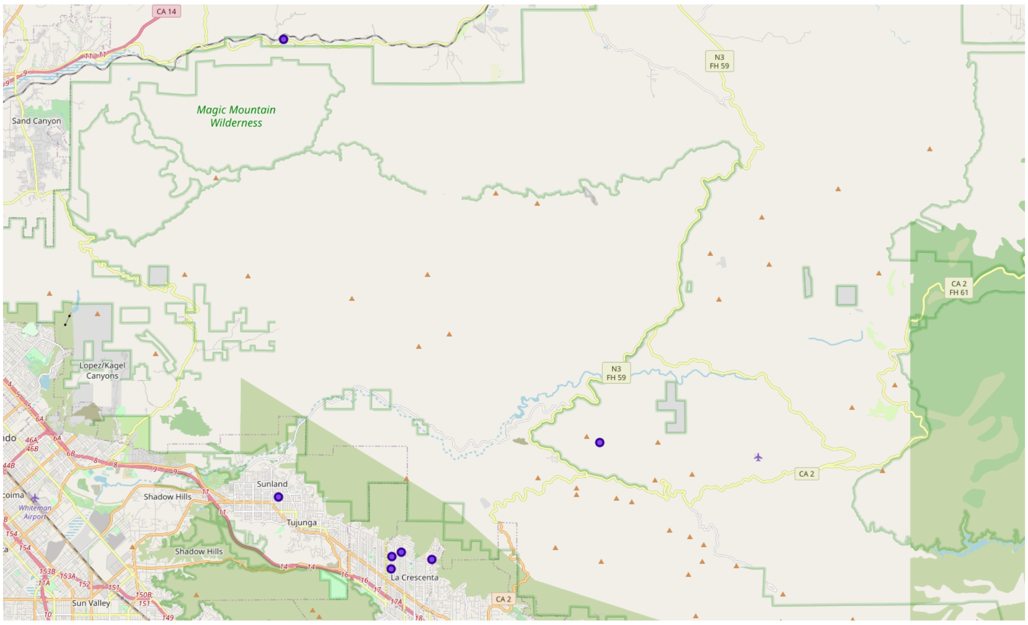

Figure-4: Location of Eateries Retrieved by Foursquare API in Los Angeles County

Counties in California have different number of venues.

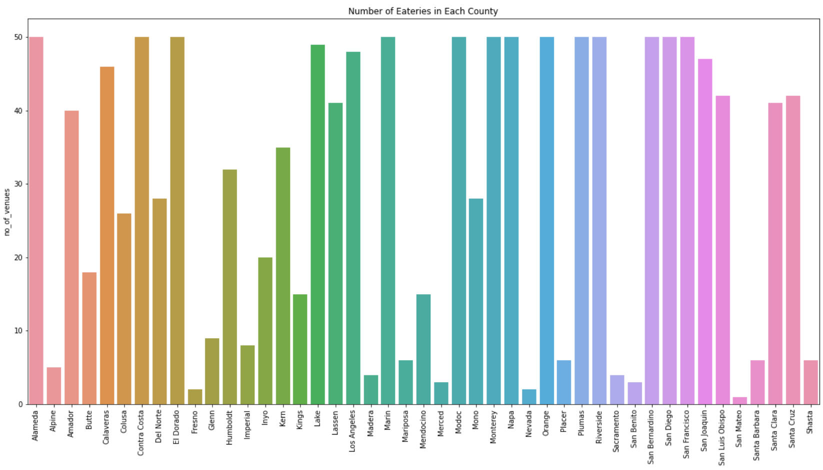

Figure-5: Distribution of Number of Eateries in California Counties (Capped at 50)

No information of any eatery was returned by Foursquare API for 13 counties.

A list of 10 most common eatery types was formed by data manipulation through calculating frequencies for each county. Here is the table.

	county	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Alameda	Coffee Shop	Fast Food Restaurant	Mexican Restaurant	Bakery	Bubble Tea Shop	Ice Cream Shop	Donut Shop	Indian Restaurant	American Restaurant	New American Restaurant
1	Alpine	American Restaurant	Sandwich Place	Café	Diner	Wings Joint	Donut Shop	Convenience Store	Creperie	Cupcake Shop	Czech Restaurant
2	Amador	Pizza Place	American Restaurant	Bakery	Café	Buffet	Asian Restaurant	Coffee Shop	Burger Joint	Ice Cream Shop	Market
3	Butte	Pizza Place	American Restaurant	Fast Food Restaurant	Sandwich Place	Supermarket	Coffee Shop	Food Truck	Comfort Food Restaurant	Food Court	Snack Place
4	Calaveras	Mexican Restaurant	Pizza Place	Café	Coffee Shop	New American Restaurant	Bakery	Restaurant	Food	Sandwich Place	Ice Cream Shop
5	Colusa	Fast Food Restaurant	Mexican Restaurant	American Restaurant	Sandwich Place	Ice Cream Shop	Pizza Place	Burrito Place	Italian Restaurant	Coffee Shop	Taco Place
6	Contra Costa	Coffee Shop	Pizza Place	Burger Joint	Fast Food Restaurant	Café	American Restaurant	Chinese Restaurant	Sandwich Place	Seafood Restaurant	Donut Shop
7	El Dorado	Food	American Restaurant	Pizza Place	Sandwich Place	Breakfast Spot	Daycare	Deli / Bodega	Café	Burger Joint	Fast Food Restaurant
8	Fresno	Fast Food Restaurant	Coffee Shop	Mexican Restaurant	American Restaurant	Chinese Restaurant	Café	Sports Bar	Market	Burrito Place	Sandwich Place
9	Glenn	Cupcake Shop	Snack Place	Wings Joint	Chocolate Shop	Comfort Food Restaurant	Convenience Store	Creperie	Czech Restaurant	Daycare	Deli / Bodega

Figure-6: Top 10 Venues of Counties (10 displayed)

3.2 Statistical Test

3.2.1 Correlation Between Population and GDP of California Counties: Pearson Coefficient

To measure the correlation between GDP and Population of California Counties, Pearson Coefficient was calculated. The Pearson Coefficient came out to be ≈ 0.952, which is very close to 1. And the p-Value was 1.32×10^{-30} which is $<< 0.05$.

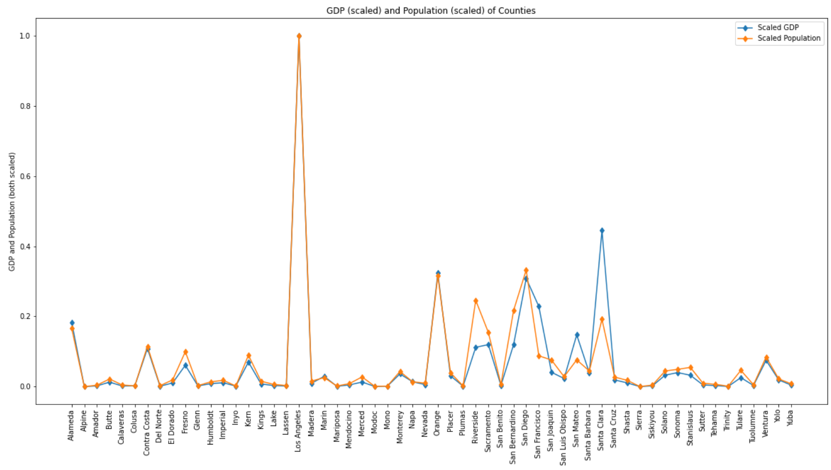

Figure-7: A plot of GDP and Population of of California Counties

These results suggests a very strong correlation between the GDP and Population of a county.

3.3 Choosing Machine Learning Model

3.3.1 Choosing KMeans Clustering

The business problem is to look for eatery types and locations to invest in. The data is not labelled. This renders the problem to be solved a classical application of unsupervised learning.

The aim is not to look for a value or look for a class. The aim is not suggesting someone only one advice for investment. To suggest the stakeholders a list of likely venues is the goal.

And this can be achieved by clustering the counties based on GDP and Population. And KMeans Clustering is the best Statistical Learning algorithm to achieve this.

Scikit-learn[2] library's implementation for the KMeans Clustering algorithm was used.

3.3.2 Choosing the Best k

To choose the best k for applying KMeans Clustering, the elbow method was applied i.e. Distortion and Inertia for each k was plotted against values of k, and the k that was chosen was where the *elbow* appeared.

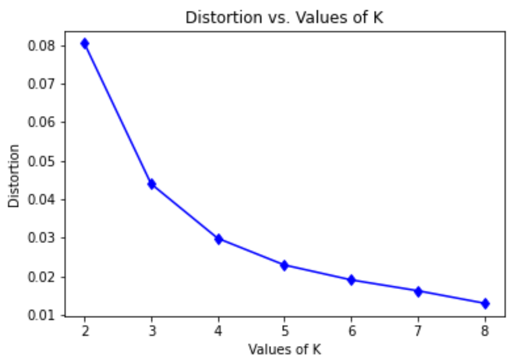

Figure-8: Distortions vs. Values of k

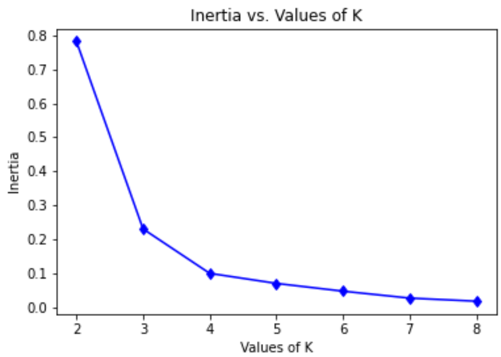

Figure-9: Inertias vs. Values of k

Evident from the figures, the best value of k is 4. And KMeans Clustering algorithm will be applied with k set to 4.

4 Results

After applying KMeans Clustering algorithm, 4 clusters were formed.

These clusters can be visualized in map.

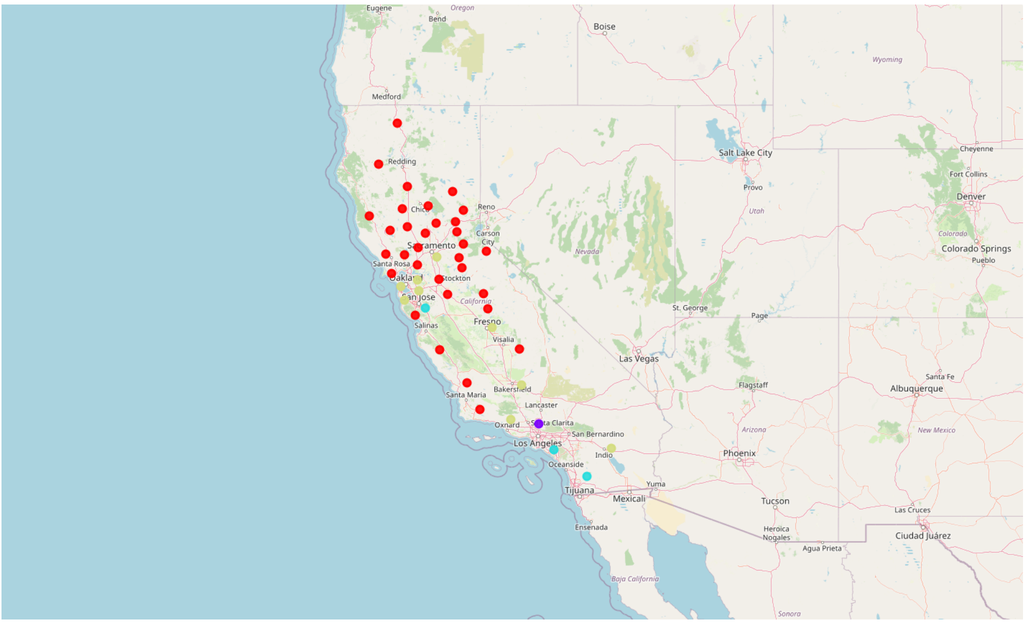

Figure-10: California Counties as Clustered by KMeans Clustering

5 Discussion

5.1 Investment Advice

5.1.1 General Advice

It can be seen in the Results section, where 4 clusters were formed based on GDP and Population of Counties, that there is one cluster which contains only one city- Los Angeles. And there is one cluster with cities with high population and GDP. It would be profitable to invest in uncommon eateries. Investing in common eateries is preferred after investing in uncommon eateries. Then, there are some counties with high population and GDP which are not on par with counties with high GDP and high populations. Only uncommon eateries in these counties are advised investment options. Then there are some counties with low GDP and low population. Investment in these eateries is not advised. If it is chosen to invest in these eateries, investment should be poured in uncommon eateries.

5.1.2 Detailed Recommendation

In clusters 2, 3 we have counties with high population and high GDP. In these counties, it will be profitable to invest in any eatery while it is advisable to invest in a eatery which is not in top 3 venues.

	Cluster Labels	county	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
12	1	Los Angeles	Café	Food Court	Deli / Bodega	Dessert Shop	Burger Joint	Donut Shop	Creperie	Cupcake Shop

Figure-11: Recommendations for Counties in Cluster 2

	Cluster Labels	county	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
20	2	Orange	Bubble Tea Shop	Bakery	Grocery Store	Dessert Shop	Donut Shop	American Restaurant	Korean Restaurant	Burger Joint
25	2	San Diego	American Restaurant	Bakery	Pizza Place	Restaurant	Diner	Snack Place	Dessert Shop	Deli / Bodega
31	2	Santa Clara	Pizza Place	Bagel Shop	Bakery	Fried Chicken Joint	Vegetarian / Vegan Restaurant	Thai Restaurant	Donut Shop	American Restaurant

Figure-12: Recommendations for Counties in Cluster 3

In cluster 4, population and GDP of counties are higher than those of the counties in cluster 1, but lower than those of counties in 2 or 3. Investment in these counties is preferred after counties in cluster 2 and cluster 3, in that order. Investment should be done in uncommon eateries so that they face lesser competition.

	Cluster Labels	county	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	3	Alameda	Mexican Restaurant	Bakery	Bubble Tea Shop	Ice Cream Shop	Donut Shop	Indian Restaurant	American Restaurant	New American Restaurant
6	3	Contra Costa	Burger Joint	Fast Food Restaurant	Café	American Restaurant	Chinese Restaurant	Sandwich Place	Seafood Restaurant	Donut Shop
8	3	Fresno	Mexican Restaurant	American Restaurant	Chinese Restaurant	Café	Sports Bar	Market	Burrito Place	Sandwich Place
10	3	Kern	Bakery	Sandwich Place	Steakhouse	Indian Restaurant	Asian Restaurant	Food Truck	Diner	Convenience Store
23	3	Riverside	Pizza Place	Wings Joint	Diner	Comfort Food Restaurant	Convenience Store	Creperie	Cupcake Shop	Czech Restaurant
24	3	Sacramento	Bubble Tea Shop	Fried Chicken Joint	Asian Restaurant	Burger Joint	Bakery	Dessert Shop	Dim Sum Restaurant	Sandwich Place
26	3	San Francisco	Food Court	Donut Shop	Fast Food Restaurant	Grocery Store	Burger Joint	Filipino Restaurant	Sandwich Place	Italian Restaurant
29	3	San Mateo	Fast Food Restaurant	Café	Pizza Place	BBQ Joint	Gastropub	New American Restaurant	American Restaurant	Burger Joint
42	3	Ventura	Italian Restaurant	Pizza Place	Food	Fast Food Restaurant	Sandwich Place	Ice Cream Shop	Snack Place	Café

Figure-13: Recommendations for Counties in Cluster 4

Cluster 1:
Cluster 1 is dominated by lower population and lower GDP counties. Investment in these counties, should be preferred after investments in counties in clusters 2, 3, and 4. Investment in most common eateries is not advised at all. Investment in these counties is least advised.

	Cluster Labels	county	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	0	Alpine	Café	Diner	Wings Joint	Donut Shop	Convenience Store	Creperie	Cupcake Shop	Czech Restaurant
2	0	Amador	Bakery	Café	Buffet	Asian Restaurant	Coffee Shop	Burger Joint	Ice Cream Shop	Market
3	0	Butte	Fast Food Restaurant	Sandwich Place	Supermarket	Coffee Shop	Food Truck	Comfort Food Restaurant	Food Court	Snack Place
4	0	Calaveras	Café	Coffee Shop	New American Restaurant	Bakery	Restaurant	Food	Sandwich Place	Ice Cream Shop
5	0	Colusa	American Restaurant	Sandwich Place	Ice Cream Shop	Pizza Place	Burrito Place	Italian Restaurant	Coffee Shop	Taco Place
7	0	El Dorado	Pizza Place	Sandwich Place	Breakfast Spot	Daycare	Deli / Bodega	Café	Burger Joint	Fast Food Restaurant
9	0	Glenn	Wings Joint	Chocolate Shop	Comfort Food Restaurant	Convenience Store	Creperie	Czech Restaurant	Daycare	Deli / Bodega
11	0	Lake	Diner	Burger Joint	Pizza Place	Food	Mexican Restaurant	Chinese Restaurant	Peruvian Restaurant	Deli / Bodega
13	0	Madera	Steakhouse	Mexican Restaurant	Dessert Shop	Sandwich Place	Italian Restaurant	Burger Joint	Buffet	Fast Food Restaurant
14	0	Marin	Coffee Shop	Indian Restaurant	Bakery	Deli / Bodega	Restaurant	Chinese Restaurant	New American Restaurant	Cheese Shop
15	0	Mariposa	Food Truck	Restaurant	Resort	Sandwich Place	Fast Food Restaurant	Burger Joint	Dessert Shop	Comfort Food Restaurant
16	0	Mendocino	American Restaurant	Fast Food Restaurant	Coffee Shop	Burger Joint	Donut Shop	Café	Food	Chinese Restaurant
17	0	Monterey	Pizza Place	Fast Food Restaurant	Bakery	Tea Room	Sandwich Place	Burger Joint	Breakfast Spot	Ice Cream Shop
18	0	Napa	Mexican Restaurant	Pizza Place	Restaurant	French Restaurant	Bakery	Deli / Bodega	Italian Restaurant	Food
19	0	Nevada	Deli / Bodega	Fast Food Restaurant	Convenience Store	Creperie	Cupcake Shop	Czech Restaurant	Daycare	Dessert Shop
21	0	Placer	Chinese Restaurant	Bar	Café	Breakfast Spot	Wings Joint	Donut Shop	Creperie	Cupcake Shop
22	0	Plumas	Chinese Restaurant	Bagel Shop	Ice Cream Shop	Diner	Bakery	Food Truck	Asian Restaurant	Coffee Shop
27	0	San Joaquin	Coffee Shop	Chinese Restaurant	Indian Restaurant	Bakery	Breakfast Spot	Sushi Restaurant	Sandwich Place	Fried Chicken Joint
28	0	San Luis Obispo	Wings Joint	Fast Food Restaurant	Convenience Store	Creperie	Cupcake Shop	Czech Restaurant	Daycare	Deli / Bodega
30	0	Santa Barbara	Steakhouse	Food	Pizza Place	Food Truck	Diner	Comfort Food Restaurant	Convenience Store	Creperie
32	0	Santa Cruz	Seafood Restaurant	Ice Cream Shop	American Restaurant	Food Court	Burrito Place	Café	Bakery	Pizza Place
33	0	Sierra	Diner	Coffee Shop	Wings Joint	Donut Shop	Convenience Store	Creperie	Cupcake Shop	Czech Restaurant
34	0	Siskiyou	Japanese Restaurant	Wings Joint	Diner	Comfort Food Restaurant	Convenience Store	Creperie	Cupcake Shop	Czech Restaurant
35	0	Solano	Sandwich Place	Bakery	Chinese Restaurant	Mexican Restaurant	Burger Joint	Burrito Place	Pizza Place	Wings Joint
36	0	Sonoma	American Restaurant	New American Restaurant	Café	Dessert Shop	Pizza Place	Bakery	Ice Cream Shop	German Restaurant
37	0	Stanislaus	Food Truck	Mexican Restaurant	Pizza Place	Donut Shop	Restaurant	Mediterranean Restaurant	American Restaurant	Fried Chicken Joint
38	0	Sutter	Coffee Shop	Pizza Place	Breakfast Spot	Indian Restaurant	American Restaurant	Sandwich Place	Burger Joint	Chinese Restaurant
39	0	Tehama	Fast Food Restaurant	Pizza Place	Mexican Restaurant	Asian Restaurant	BBQ Joint	Café	Sandwich Place	Thai Restaurant
40	0	Trinity	Coffee Shop	Convenience Store	Creperie	Cupcake Shop	Czech Restaurant	Daycare	Deli / Bodega	Dessert Shop
41	0	Tulare	Restaurant	Mexican Restaurant	Steakhouse	Diner	Convenience Store	Creperie	Cupcake Shop	Czech Restaurant
43	0	Yolo	Food Truck	Bakery	Coffee Shop	Burger Joint	Fast Food Restaurant	Diner	Chinese Restaurant	Latin American Restaurant
44	0	Yuba	Restaurant	Bar	Mexican Restaurant	Dim Sum Restaurant	Coffee Shop	Comfort Food Restaurant	Convenience Store	Creperie

Figure-14: Recommendations for Counties in Cluster 1

5.2 Limitation of This Project

The project treats GDP and Population as indicators of a new eatery's success. But GDP and population does not capture the whole picture. Other information like consumer spending habits, financial information from already established eateries etc. could prove valuable additions to considerations undertaken to make the decision.

The Foursquare API limits the number of venues returned in an API call to 50. While for some counties, this limit is not a hindrance, for some counties, this number is too low. This is evident from Figure-5. Counties like Los Angeles, El Dorado, Alameda, San Diego etc have many venues. We cannot get a complete picture without considering those missing venues.

Foursquare API returned less than 10 eateries for some counties. It might provide a bar in deciding in investing in those counties.

5.3 Possible Sources of Error

In other scenarios, an outlier like Los Angeles would have been discarded from consideration. But in real life scenario, such as this, where investment options are being considered, an outlier cannot be simply discarded. Although, the presence of this outlier provides some inconvenience in aesthetic visualization, as clustering algorithm has been applied, it does not create any major problem which could have been possible if classification or regression algorithms were to be applied.

6 Conclusion

I would like to emphasize on the generality of the approach. When investing in an eatery (or multiple eateries), relevant data should be taken into serious consideration. And clustering algorithm can be applied to break the investment options into meaningful clusters which will play crucial role in decision making. This project provides an outline on how to use publicly available data and apply clustering algorithm to form data driven decision.

This project can be further improved by including more relevant data, such as consumer spending habit, availability of talent and skilled labor etc. And including those information and applying machine learning algorithm to that data can aid in better data driven decision making.

The machine learning pipeline can be further improved by including more machine learning algorithms. There are countless possibilities of such applications.

Acknowledgements

I would like to acknowledge the role of the instructors of all the courses in "IBM Data Science Professional Certificate" Specialization on Coursera. I would like to thank the Data Science community of Stack Exchange. Help of my family was instrumental in the completion of this project.

References

[1] *Annual Estimates of the Resident Population for Counties in California: April 1, 2010 to July 1, 2019 (CO-EST2019-ANNRES-06) Source: U.S. Census Bureau, Population Division Release Date: March 2020*

[2] *Scikit-learn: Machine Learning in Python by Pedregosa, Varoquaux, Gramfort et al.*