Data

Chamdia, Raunaq

Coursera IBM Applied Data Science Capstone

September 2019

Table of Contents

Introduction

Restaurant food is increasingly becoming popular in the United States as working young adults shift for home cooking to eating outside. This creates a growing demand for restaurant variety in different locations. Hence new restaurants looking to open up in an area should have a strong understanding of the varieties of restaurants in an area and the shortages of restaurant types in different areas. Hence for each type of restaurant that one is looking to open up it is important to understand the fundamentals in terms of competition and supply.

Business Problem

The objective of this capstone project is to analyze the neighborhoods of Sacramento, consider distribution of popular venues and determine the best place to open a Chinese Restraint. Using the data science methodology and analyzing the data through concepts learnt in this course such as machine learning clustering, one hot encoding etc. we can answer the question:

Where would you recommend one open a Chinese restaurant in Sacramento, CA?

Target Audience

This project is particularly useful to project developers and investors that are seeking to open a new Chinese restaurant in the Sacramento Area. Since Sacramento is a small city it is still developing in terms of infrastructure and would benefit from the opening of small businesses and restaurants. A Chinese restaurant was chosen as the type of small business based on personal preference, but the analysis can be applied to a multitude of business types. This small developing town was named by Thrillist as a ‘city that about to blow up as a food destination’. Given this steady development of the town many new investors and business owners would be looking to expand into the area.

Data

Description

To adequately analyze the business problem, one would need the following data:

- Information about the neighborhoods in Sacramento. Sacramento is a small city in California that can be a useful demonstration of machine learning techniques on relatively small data sets.
- Coordinates in terms of latitude and longitude of each neighborhood to visualize the neighborhoods and further analyze venue data for each neighborhood.
- Data about Chinese restaurant venues in each neighborhood, it would be easier to get general venue data in each category and then filter that to specifically analyze what the project requires in terms of classification and clustering.

Source

The list of neighborhoods in Sacramento can be found on Wikipedia (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Sacramento,_California) with a total of 24 neighborhoods. This data can be extracted from this database using web scraping through packages such as requests and Beautiful Soup. Additionally, packages such as Geocoder allows one to determine the coordinates of each neighborhood and further refine our dataset. Once this preliminary dataset is defined any outliers without definite data can be filtered out and the data can be further processed.

To determine popular venue area in each neighborhood one can, deploy the Foursquare API. Foursquare has one of the largest databases of 105+ million places and is used by 125,000 developers. The API is able to provide venue data and venue category which would help us identify business types and further refine our analysis.

Methodology

Initially, we need the rundown of neighborhoods in Sacramento. This is accessible in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur). Utilizing python packages requests and Beautiful Soup we can easily obtain this by identifying the type of HTML class the data is stored in and using the find all function to retrieve the necessary data. Next, we have to get the coordinates of each of the neighborhoods we have to use with the Foursquare API. The geocoder package is helpful for this task as we can use the Nominatim package to determine latitude and longitude based on the location name. With the names and coordinates for each neighborhood we can filter out any non-existing values and create a data frame of the information. The folium package can be used with this data now to visualize the map of the city and the location of each neighborhood on it.

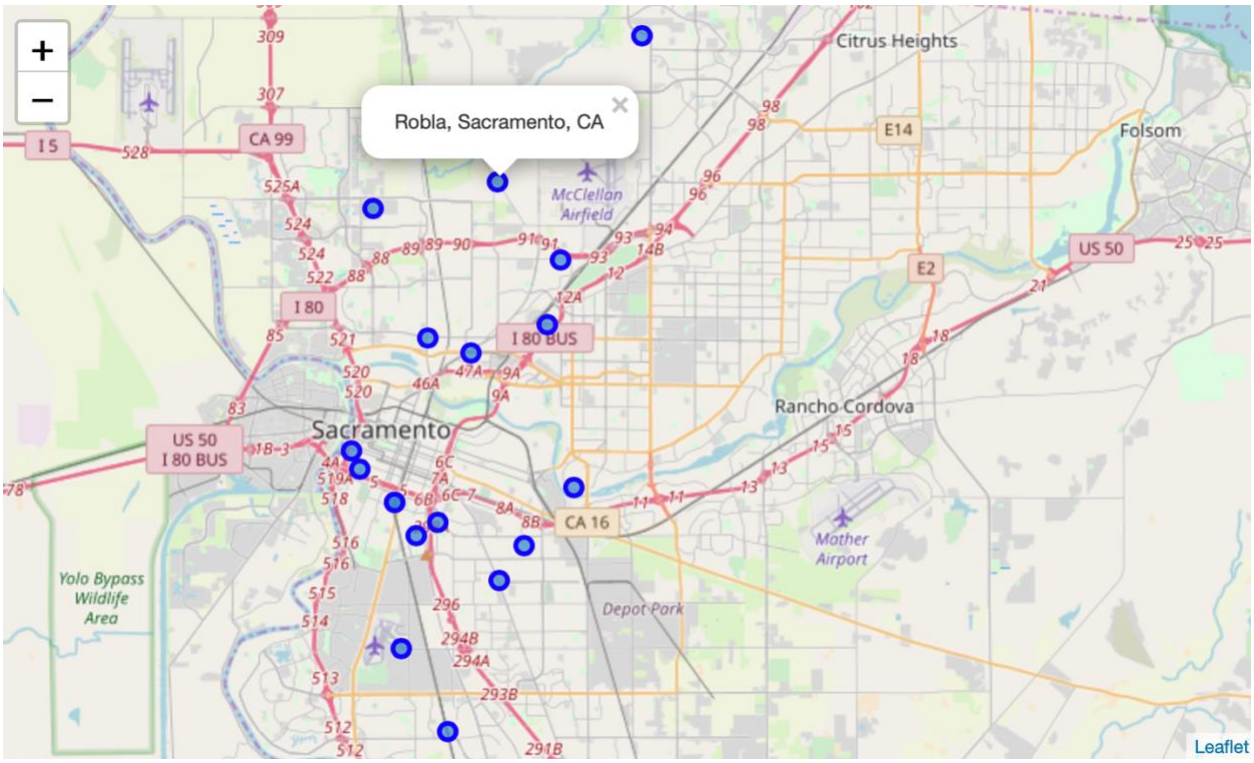


Figure 1. Sacramento neighborhoods seen using Folium

Now with access (credentials) to the Foursquare API one can develop a function to obtain data for the top 100 venues within a 1000-meter radius for each coordinate neighborhood. The API can return data for each neighborhood in a JSON format from which one can determine the venues name, category, and coordinates. It is essential to have the json package to access the returned data. Now we can conduct all sorts of machine learning analysis to determine different information trends about different aspects of the data.

```
The API found 632 venues in the Sacramento Neighborhoods
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ben Ali	38.615511	-121.42604	River City Phoenix	38.610096	-121.432634	Marijuana Dispensary
1	Ben Ali	38.615511	-121.42604	Dolphin Scuba Center	38.610841	-121.431171	Sporting Goods Shop
2	Ben Ali	38.615511	-121.42604	Sarom's Southern Kitchen	38.611439	-121.422809	Southern / Soul Food Restaurant
3	Ben Ali	38.615511	-121.42604	RT Metro	38.616119	-121.431389	Train Station
4	Ben Ali	38.615511	-121.42604	7-Eleven	38.620045	-121.419002	Convenience Store

Figure 2. First 5 Results of Foursquare API

For the purposes of this study we will use one hot encoding and standardization to process our categorical variable (venue category) into a form that machine learning algorithms can further classify the data. The K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. In this project k-means clustering was used to cluster the Chinese restaurant frequency data obtained into 3 categories, neighborhoods with no Chinese restaurants, neighborhoods with a moderate density of Chinese restaurants and neighborhoods with a high density of Chinese restaurants. Hence clustering would allow us to determine which areas are suitable to open up a new Chinese restaurant based on the competition and shortages of the category in each neighborhood in Sacramento.

Results

The results from the k-means clustering illustrate that we can classify the density of Chinese restaurants in Sacramento neighborhoods into 3 categories:

Cluster 0:

Neighborhoods with no Chinese restaurants

	Neighborhood	No. of Chinese Restraunts	Cluster Labels
0	Meadowview	0.0	0
1	Robla	0.0	0
2	North Sacramento	0.0	0
3	Natomas	0.0	0
4	East Sacramento	0.0	0
5	Valley View Acres	0.0	0

Cluster 1:

Neighborhoods with a moderate density of Chinese restaurants

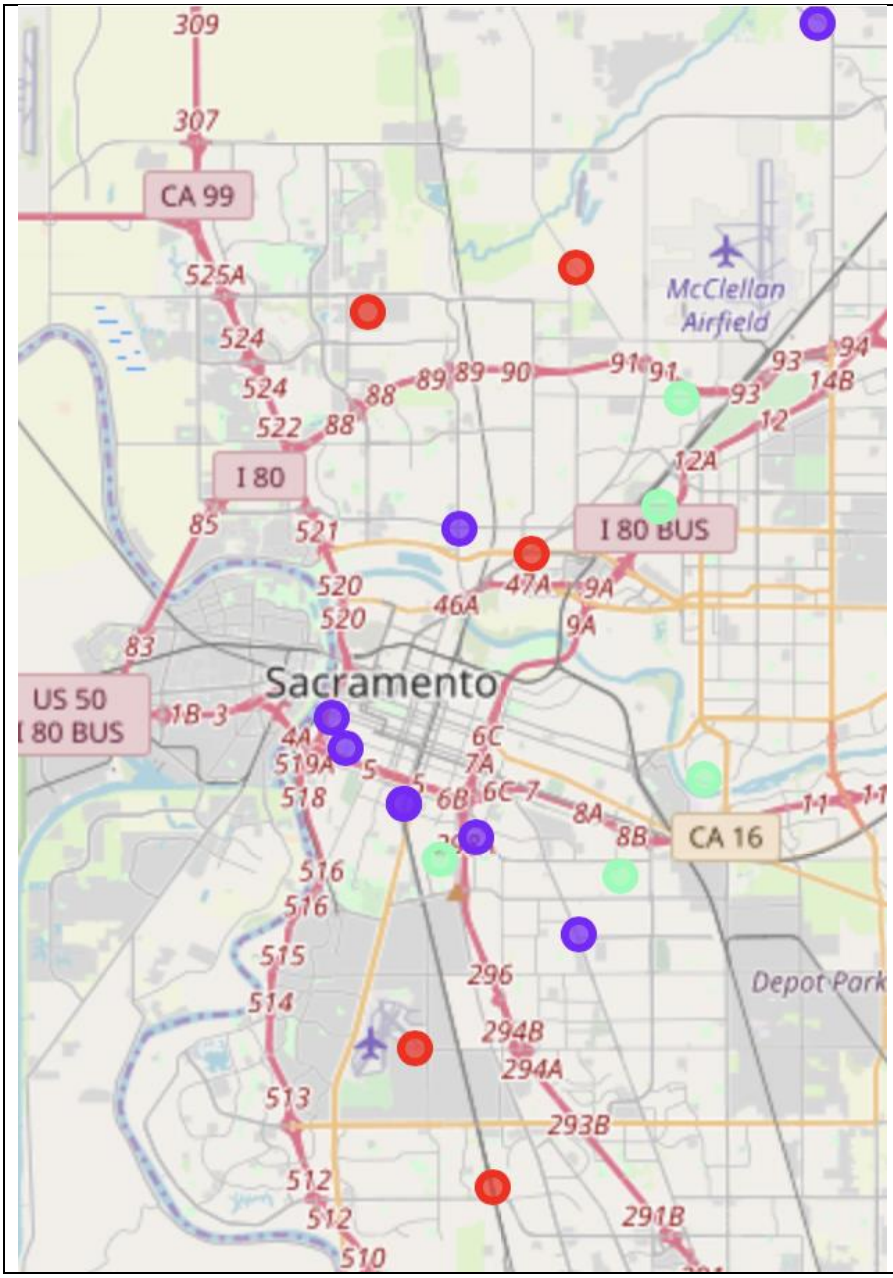
	Neighborhood	No. of Chinese Restraunts	Cluster Labels
6	Gardenland	0.038462	1
7	Land Park	0.039474	1
8	Upper Land Park	0.039474	1
9	Midtown Sacramento	0.040000	1
10	Oak Park	0.048780	1
11	Colonial Heights	0.029412	1
12	Southside Park	0.042254	1
13	Downtown Sacramento	0.020000	1

Cluster 2:

Neighborhoods with a high density of Chinese restaurants

	Neighborhood	No. of Chinese Restraunts	Cluster Labels
14	Elmhurst	0.062500	2
15	Del Paso Heights	0.100000	2
16	Curtis Park	0.065217	2
17	Tahoe Park	0.076923	2
18	Ben Ali	0.058824	2

We can also visualize this clusters on a map using Foluim:



Legend:

- Red: Cluster 0
- Purple: Cluster 1
- Green: Cluster 2

Discussion

It is evident that the majority of the Chinese restaurants are centrally concentrated in Sacramento. The highest density of Chinese restaurants is seen in Cluster 2, followed by a moderate density in cluster 1 and no density of Chinese restaurants in cluster 0.

Hence while choosing a neighborhood to open a Chinese restaurant it would be wise to consider locations in cluster 0 and 1 due to lack of competition and ability to conquer the market in that area. It is also important to remain central in the city to cover the maximum radius of potential customers. Hence given these criteria the location options for a new Chinese restaurant are narrowed down to North Sacramento, Southside Park or Upper Land Park. This list can further be narrowed depending on population density and property costs. This analysis would allow the developer to make an informed decision about the more profitable place of business to set up shop where the scarcity for their service is the highest. Minimizing competition and distinguishing your business from your immediate surrounding would allow the business to prosper and better develop strategies to grow.

Limitations

It is important to realize that this study only uses the density of one category to cluster its areas. In reality location could also depend on population density, income and resident preferences. It would be important to further consider these aspects given that the necessary data can be obtained and can be processed to be used in parallel to location data. Additionally, one should also consider property costs so as to not exceed the investors budget and provide an informed decision to all the stakeholders. This study serves as a preliminary analysis of the business problem but a through approach would require access to resources beyond the scope of this study.

Conclusion

This project was concerned with the data science methodology where one identified a business problem, collected the data required, processed the data, used relevant methods to prepare data for analysis by the appropriate machine learning techniques and lastly used the results to recommend solutions to relevant stakeholders.

Here we analyzed the best place to open a Chinese restaurant within the neighborhoods of Sacramento, California. The analysis was successful able to shortlist neighborhoods in Sacramento where there is a shortage of Chinese restaurants and supply for such a service is scarce.

References

- https://developer.foursquare.com/docs/api/endpoints
- https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Sacramento,_California
- https://www.thrillist.com/travel/nation/sacramento-up-and-coming-food-city-us
- Coursera IBM Applied Data Science Capstone Course

Acknowledgements

I would like to thank Alex Aklson and his team and Coursera/ IBM for creating this opportunity to learn about data science and get real experience in developing projects, analyzing data to identify problems and making creative solutions.