

Capstone Project

Machine learning for identification of commercial
opportunities in Lincoln, Nebraska

Kevin Cheek

June 2019

Table of contents

Section	Page
Introduction	3
Business Problem	3
Target Audience	3
Data	3
Methodology	4
Results	7
Discussion	8
Conclusion	8
References	8

Introduction

Lincoln is the capitol city of Nebraska. It is the home of the University of Nebraska at Lincoln. It has a population of 287,000 and the city limits encompass an area of approximately 96 square miles. The ability to pinpoint areas of the city and surrounding area which are ripe for economic development could aide entrepreneurs, developers, and city planners to focus efforts and resources where they will provide an optimum return.

For this project, I will split the city of Lincoln, Nebraska into a 26x26 grid of $\frac{1}{4}$ square mile regions and explore the makeup of these regions by applying machine learning principles to FourSquare places data. Using insights gained from this analysis, I will categorize each region on the grid by the types of venues they contain.

Business Problem

Can FourSquare places data be used to identify suitable locations for commercial development or expansion and minimize search time and costs? Based on this data, could a recommendation be made for a suitable location within the city of Lincoln for an entrepreneur to open a new Chinese restaurant?

Target Audience

The target audience of this project could include entrepreneurs seeking opportunity for new business development, commercial real estate brokers, and city planners and other government agencies.

Data

- Latitude and Longitude coordinates of a suitable city center for Lincoln, Nebraska
 - Coordinates of the city center will be obtained by converting an address to coordinates using the Nominatim geocoding tool
- Latitude and Longitude of all points in a grid overlay of the city
 - The grid overlay and associated Latitude and Longitude coordinates will be calculated using Python.
- Venue data from FourSquare for each region in the grid
 - Venue data will be retrieved from the FourSquare places api and will leverage the explore endpoint

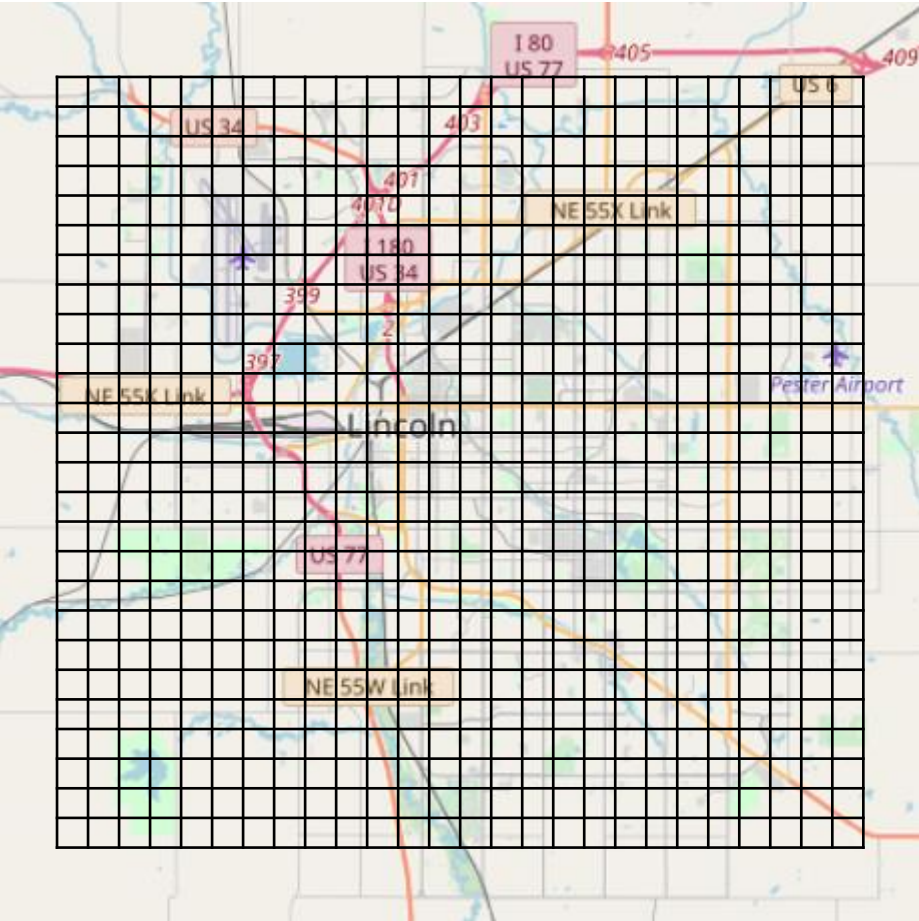
Methodology

After finding an address in the approximate center of the city, I generated a grid of 26x26 0.25sq mile regions centered on that point.

Using the grid overlay coordinate points, I retrieved FourSquare data including up to 100 total venues found within each of the grid regions. Venue category counts were then calculated for each region in the grid and the values stored in a dictionary for easy retrieval and converted to a pandas dataframe for further analysis.

Using the dataframe of venue counts by region, I found that Chinese was the 6th highest of the food-related venue categories. This meant that Chinese restaurants are relatively popular in Lincoln but are not as abundant as some of the standard fare including Sandwiches, Pizza, and Mexican.

Fast Food	169.0
Sandwiches	140.0
Pizza	128.0
Convenience Store	110.0
Mexican	87.0
Coffee Shop	87.0
Park	86.0
Bar	83.0
American	77.0
Hotel	70.0
Pharmacy	67.0
Gym / Fitness	64.0
Grocery Store	58.0
Chinese	57.0
Gas Station	54.0
Construction	48.0
Spa	44.0
Gym	42.0
Burgers	40.0
Ice Cream	39.0



Methodology cont.

Once I confirmed the viability of introducing additional Chinese venues in Lincoln, I used KMeans clustering to group the grid regions into similar categories based on the mean of the venue counts for each category. I chose 6 for the number of clusters to optimize the differentiation between the clusters but still be able to easily identify different usage patterns within each cluster.

cluster	0	cluster	2	cluster	4
Fast Food	1.524590	Bar	7.0	Sandwiches	2.500000
Sandwiches	0.819672	Sandwiches	6.5	Coffee Shop	2.250000
Pizza	0.721311	Pizza	3.5	Fast Food	1.666667
American	0.606557	Brewery	2.5	Mexican	1.500000
Pharmacy	0.573770	Burgers	2.5	Pizza	1.083333
Hotel	0.557377	Coffee Shop	2.5	American	1.000000
Convenience Store	0.540984	Mexican	2.5	Bar	1.000000
Chinese	0.491803	Cocktail	2.0	Hotel	1.000000
Mexican	0.491803	Concert Hall	2.0	Mobile Phones	1.000000
Grocery Store	0.442623	Hotel	2.0	Grocery Store	0.666667

cluster	1	cluster	3	cluster	5
Park	0.068	Apparel	6.5	Convenience Store	0.484848
Bar	0.066	Women's Store	3.0	Pizza	0.373737
Fast Food	0.056	Department Store	2.0	Sandwiches	0.262626
Baseball Field	0.054	Lingerie	2.0	Park	0.252525
Pizza	0.054	Mexican	2.0	Coffee Shop	0.242424
Gym / Fitness	0.048	Shoes	2.0	Fast Food	0.242424
Convenience Store	0.046	Accessories	1.5	Mexican	0.202020
Golf Course	0.046	Gift Shop	1.5	American	0.151515
Lake	0.046	Pharmacy	1.5	Spa	0.151515
Sandwiches	0.038	Supplement Shop	1.5	Gas Station	0.141414

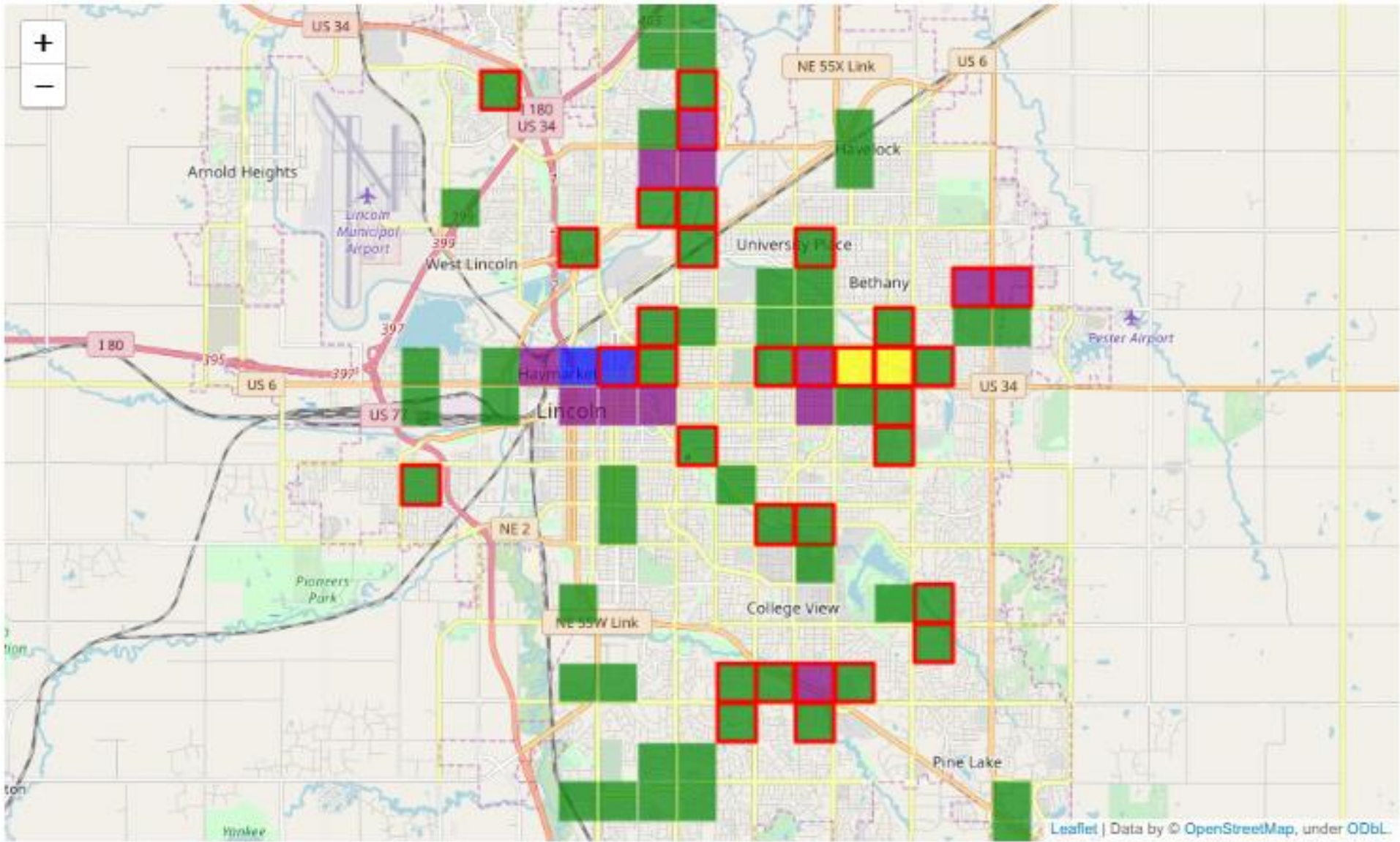
Analysis of the cluster data made it clear that two of the clusters (1 and 5) were regions with primarily residential usage. These clusters included venues like parks, baseball fields, golf courses, and lakes in the top 10 venues. I excluded these two regions from consideration for the new Chinese restaurant. The rest of the clusters were mostly commercial regions with higher food, entertainment, and shopping venue scores. I was also able to further define the remaining clusters based on their relative mean score values (high/med/low) and the primary venue types (food/shopping/entertainment).

Cluster 0:	Cluster 2:	Cluster 3:	Cluster 4:
Tier 3 commercial	Tier 1 commercial	Tier 1 commercial	Tier 2 commercial
Low-Medium venue density	High venue density	High venue density	Medium venue density
Primarily Food and Service venues	Primarily Bars/Food/Entertainment	Primarily Shopping/Food	Primarily Food

Methodology cont.

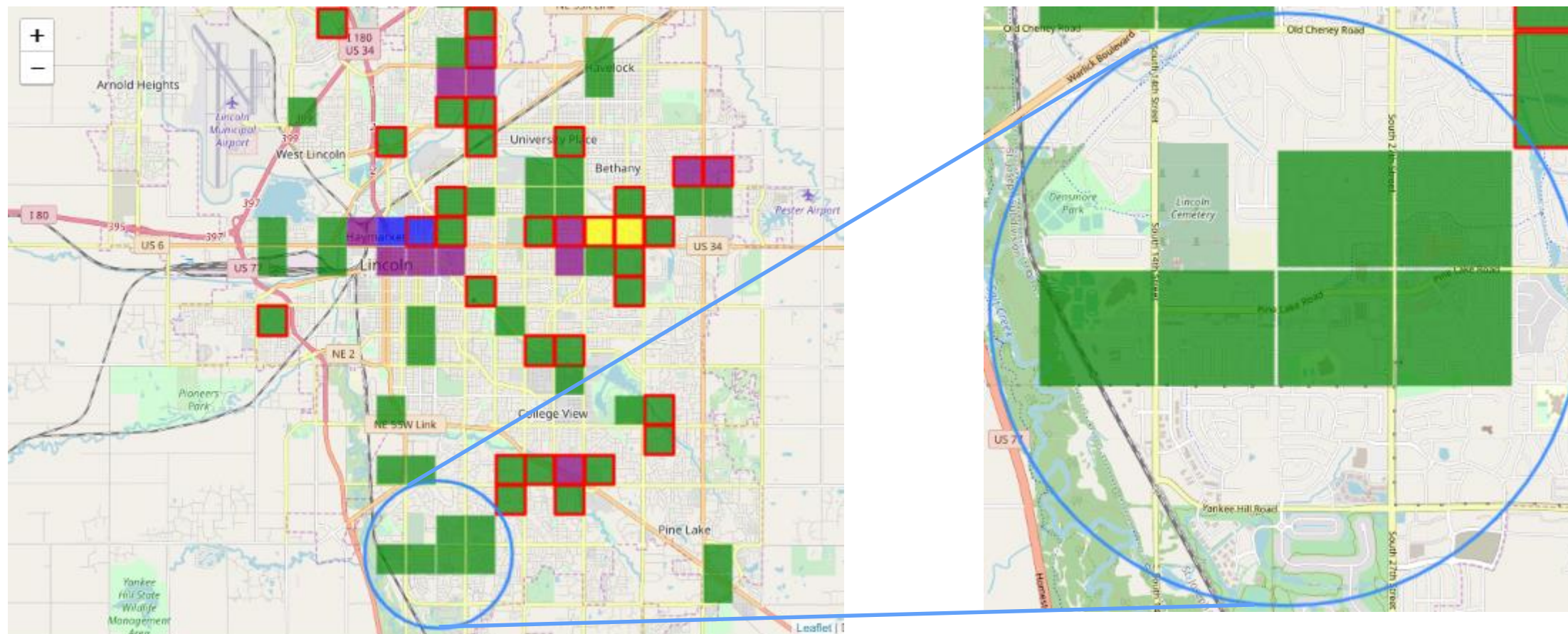
Having defined the clusters, I generated a map showing the plot regions to be considered and color coded them according to the cluster to which they belonged. I also added a red border if the region already included a Chinese venue.

- Green Cluster 0
- Blue Cluster 2
- Yellow Cluster 3
- Purple Cluster 4



Results

There are several areas of Lincoln which should be able to support a new Chinese restaurant. From looking at the map it becomes clear that the majority of the viable regions either already have a Chinese venue or are located next to another region which already has one. In the interest of keeping competition to minimum, it would appear that the southwest corner of the city would be the best choice for the new location. This area has a large cluster of Tier 3 commercial areas without any existing Chinese venues. The area is surrounded by dense residential areas, includes a mid-sized shopping mall, grocery, and several fast food venues. It is also located at a crossroads of 2 major north-south roads and 1 major east-west road.



Discussion

I was able to use the data available from FourSquare to identify and differentiate the commercial regions of the city of Lincoln. I was also able to identify areas that were underserved in the Chinese venue category and recommend a specific area of town for a new Chinese restaurant. However, this project was very limited in data scope and there are several key pieces of information that would still need to be considered before selecting a location for a new restaurant.

Conclusion

This project was successful in using machine learning to classify geographic regions based on the venues located within them. I was able to identify the key commercial districts within a city and pinpoint areas that could be potential sites for a new venture. Stakeholders would be able to leverage this information as a starting point to focus their search for economic expansion opportunities within a target city.

References

Lincoln, NE statistics

https://en.wikipedia.org/wiki/Lincoln,_Nebraska

FourSquare api

<https://developer.foursquare.com/docs/api>

Nominatim geocoder

<https://wiki.openstreetmap.org/wiki/Nominatim>