LOCATION-BASED BUSINESS RECCOMENDATION ENGINE

BAN620 Data Mining

Rishabh Rao

CSU East Bay | Business Analytics

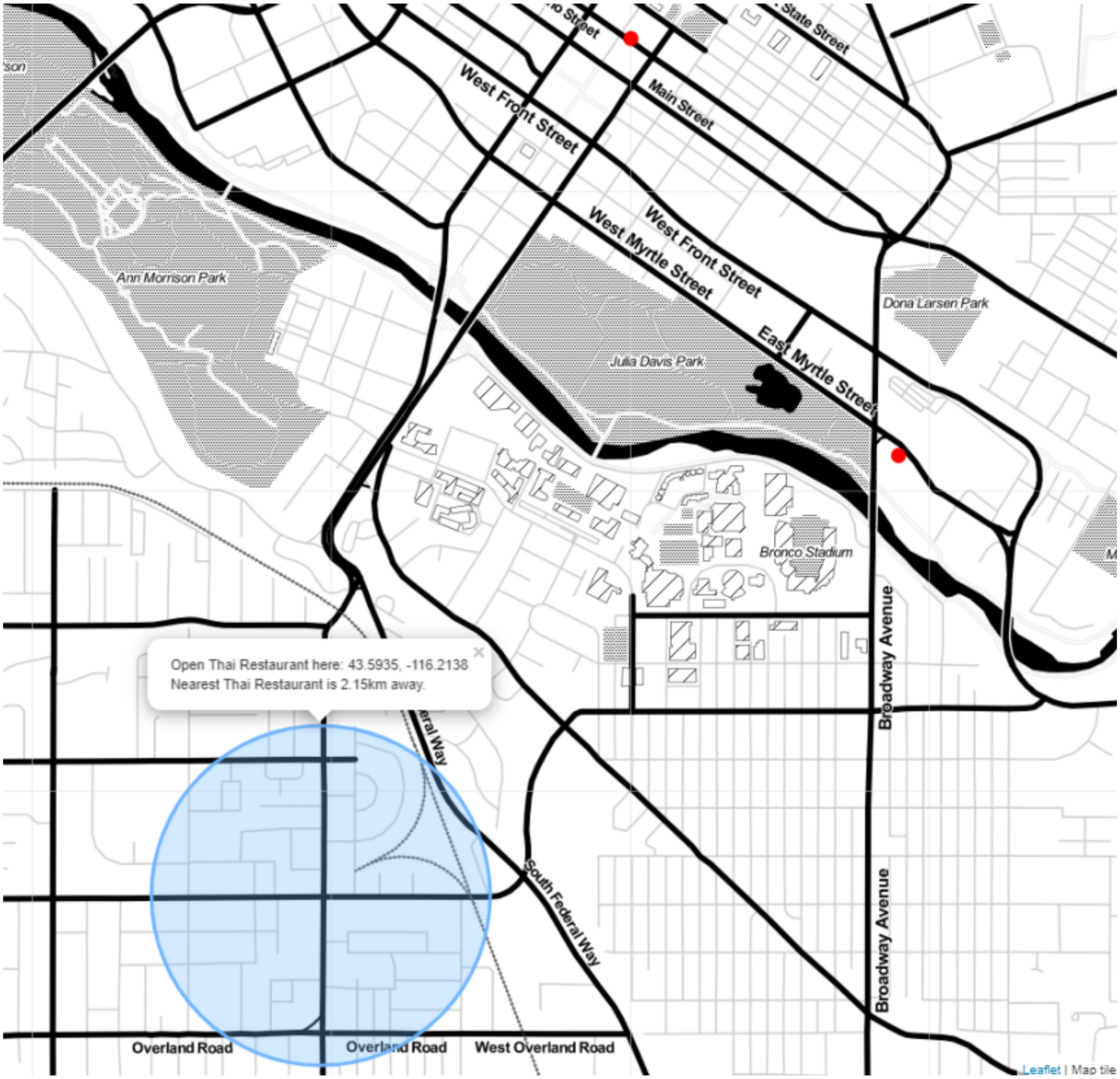
Summary

Are you an entrepreneur? Considering your next business venture? How can you determine exactly what the market around you needs based on real data? Where do you even begin?

With over 33,000,000 small business in America, these are some challenges every budding entrepreneur faces at some point or another. Our project will seek to answer these questions by developing a recommendation method that can be applied to a variety of businesses.

Introduction

The goal of our project is to determine the best location for a business using several factors. We will employ a general method that can be applied to a variety of business types and different data sets, though we will explore restaurants here. The data mining approach we will mainly focus on is K-Means Clustering and K Nearest Neighbors with geospatial data.



Open Thai Restaurant here: 43.5935, -116.2138
Nearest Thai Restaurant is 2.15km away.

Main Chapter

Step 1: Gather and Prep

Before we begin our recommendation method, we must collect data to analyze both the regions and the possibility of brick-and-mortar establishments. We will establish metropolitan regions (metros) using Census Bureau data regarding demographics, socioeconomic data, and more. We can encode the geospatial data of our metro regions using the geopy library and ArcGIS as the geocoder. ArcGIS does not require credentials to access, currently and is more accurate than geopy's default geocoder.

We will leverage this data alongside the Foursquare Places API to gather geospatial data of restaurant types, latitude and longitude; there is also a Google Maps API it is less robust. An expansion of our method could potentially leverage the API services of both Foursquare, Google Maps, possibly Yelp and other vendors and validating the data against each other. Different services also employ and measure different parameters that could all be of value.

Most of the data regarding American businesses can be found on the Census Bureau Website (https://www.census.gov/). Most of the data is not user-friendly and has many errors that is not conducive to bulk scrubbing and manipulation via Python. In this case, it will be easier to manually perform the bulk of the initial data scrubbing in Microsoft Excel or Google Sheets.

To ensure we are not just searching the immediate center (latitude and longitude) of our metropolitan areas, we will conduct a 25km circular sweep around our city geolocations, and repeat for each of Foursquare's 350+ restaurant and venue types. Each circular sweep can only return a maximum of 50 businesses at a time, so it is crucial we break the search area into several smaller sectors and search each. We utilize the recursive function getRestaurantCoordsAndType() to achieve this.

Foursquare and other similar services do not only categorize restaurants and have a variety of business types available on their platforms. We can repeat and tweak our recommendation method to work with other business types as well.

The main focus of our analysis will be to find underrepresented restaurant types in each region that would be a good market. We will determine underrepresentation of restaurant categories and also find populous regions that would benefit the most from having a restaurant of.

This specific use-case models a real-world example that may be employed in a business setting; we will focus on using K-Means Clustering and explore the geospatial data instead of expanding any other methods that are not capable of compensating for geolocation.

Step 2: Grouping

In our case, we will focus on the Pacific Northwest Region (PNW hereon) of America as it is easily available on the Census Bureau website via CSV download or API. The API is available for a pull however, since the PNW region has many metropolitan regions with non-user-friendly data it is best we utilize manual scrubbing as aforementioned. We have gathered the geospatial data of our regions and now we must apply the geolocation (latitude and longitude) to each region so we can conduct our K-Means Clustering. The process is tedious and took many steps inside our Jupyter notebook as we had to initially convert all of our null values returned from the API and then also add the coordinates to the metropolitan data.

We collect our data in batches as we experienced many timeouts when trying to collect data all at once. Since we have about 519 records, we collect them in batches of approximately 100. The geospatial data is also saved to a new CSV so that we can utilize this at a later stage.

We will conduct further analysis by choosing a specific region of interest to zone in on. The Pacific Northwest has many metropolitan and micropolitan areas that are often unincorporated towns and cities that lack exact data. To gather reliable results for our restaurant data we have to utilize Foursquare's API, which only allows 5000 calls per hour at the free level. This delays our method significantly, however we were able to find a method (Caine 2) that would allow us to time our API calls and maximize efficiency at the lowest possible spend. This is an important metric as many startups face surmounting costs from cloud computing firms that charge great deals for API usage (Twilio, Stripe, AWS, etc).

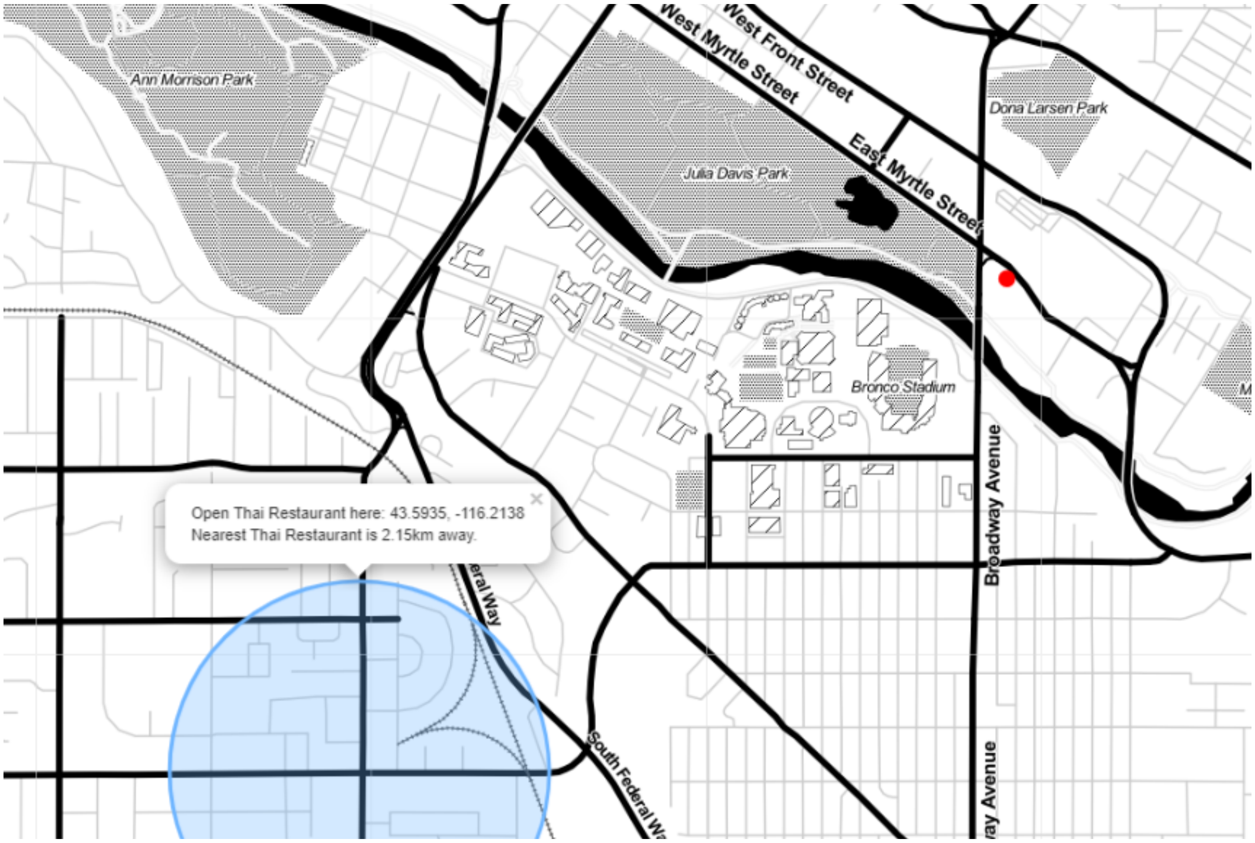
Step 3: Finding Regions For Restaurants

The next step is to utilize the demographic data to determine what populous regions have the most demand for restaurants. An important cluster would be the "population vs total restaurants" ratio for any given metropolitan region. We can assume that the demand for restaurants in higher if our ratio is higher. The opportunity to start a successful restaurant would be the greatest, given the population would be in most demand for restaurants. In other words, we are targeting the neighborhoods with the least amount of restaurants per person. We are able to extend our analysis to determine which are the most underrepresented restaurant types in the area, because the Foursquare API will allow us to categorize based on restaurant type. This extension will be how we make our case to determine the best location for the most underrepresented restaurant type.

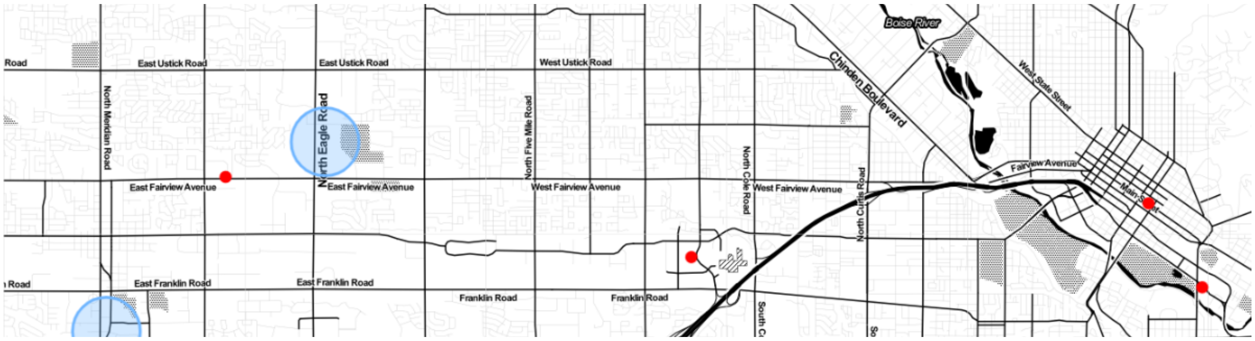
Our K-Means Clustering allows us to draw impactful insights from the data and utilize the statistics to find the most underrepresented restaurant types in each region. The next step is to find the specific location we should open each restaurant and how far they would be from competitors. We utilize the scikit-learn’s DBSCAN algorithm to handle clustering across city streets as restaurant distribution is uneven and non-random. We are able to generate these maps using the Folium library and procedurally generate them using our recommendation methodology.

Conclusion

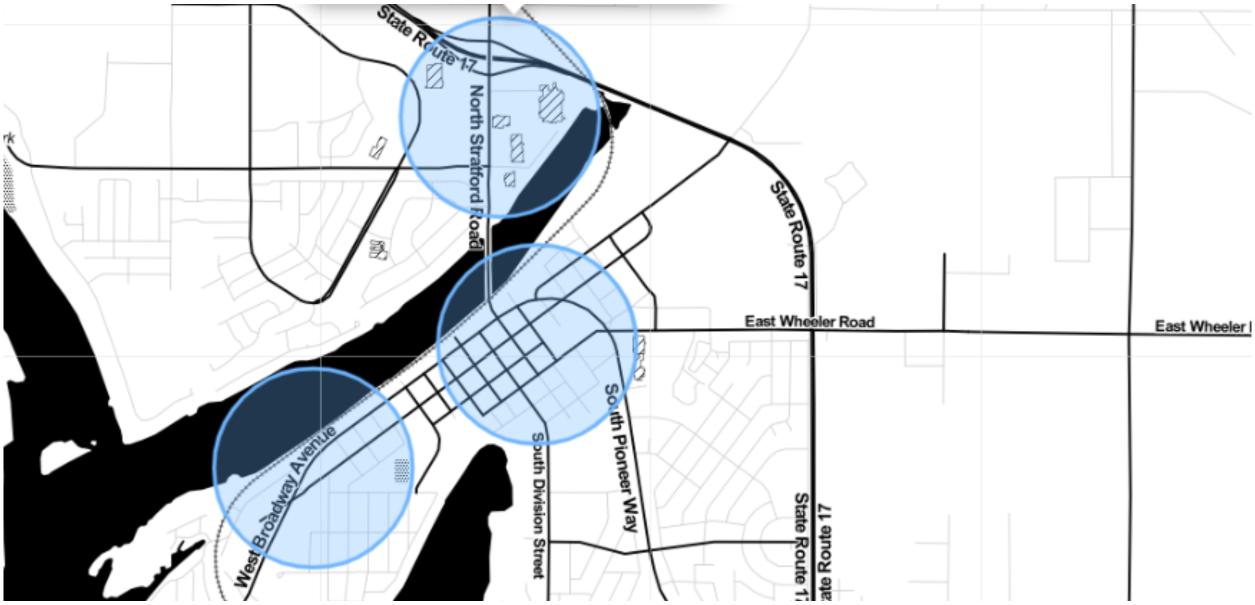
We can consider our “recommendation engine” to be a preliminary tool to be used for analysis purposes only. With considerable tweaks and taking advantage of data cross validation from various sources, we could have eventually built a real-world business ready application. The goal of our analysis yielded “The Best Restaurant Type” to open by region and population. We can consider a few of these examples below as underrepresented restaurants that may fare well in the Pacific Northwest Region. There are many open gaps in the Foursquare database that fall short of canonical standards but are good enough.



Best area to open a Thai Restaurant in Boise, Idaho. The maps are generated by Folium.



All of the recommended locations to open Thai restaurants in Boise Idaho, above. Our recommendation engine is driven primarily by the data we feed it; our API usage and Census data is considered canonical for these purposes but not necessarily.



Best area to open an American restaurant in the Moses Lake, WA area. Foursquare's API is not perfect and there is overlap between Breakfast food diners and American food diner. Further exploration could involve cross validation of the data from different sources to best accomplish segmentation of business type.

The best industries this could fare for are repeatable chains and franchises, that have parameters that they would be able to control for each new location. Additional limitations include that our methodology only works for America currently given that Foursquare and

Census data is only readily available for USA. Multinational chains that open new branches abroad often face even greater challenges in foreign markets and can often fail. The launch of a new product or a expanding into new markets is always a tricky question. The fields of corporate development and business development have traditionally been financial equations, however with the burgeoning field of data mining we are able to find novel methods to answer age-old questions.

Bibliography

Braun, Jeffrey. “Chipotle Locations.” *Kaggle*, 28 July 2020, www.kaggle.com/datasets/jeffreybraun/chipotle-locations/data.

Bureau, US Census. *Census.Gov*, 7 May 2024, www.census.gov/.

Deshpande, Rutvik. “How to Choose the Ideal Site for Designing Your Restaurant Using Data Science.” *Medium*, The Startup, 10 Aug. 2020, medium.com/swlh/how-to-choose-the-ideal-site-for-designing-your-restaurant-using-data-science-2cbfb9853f93.

“Get Started.” *Developer*, docs.foursquare.com/developer/reference/places-api-get-started. Accessed 12 May 2024.

“How to Choose the Best Business Location Analysis Platform?” *Mapchise*, 10 Sept. 2021, mapchise.com/blog/how-to-choose-business-location-analysis-platform/.

“Recommender Systems for Business - A Gentle Introduction.” *Width.Ai*, www.width.ai/post/recommender-systems-recommendation-systems. Accessed 12 May 2024.