IBM Applied Data Science Capstone

Opening a New Restaurant in Minneapolis, Minnesota, US
By: Shaoqing Liu
Dec 2019

Introduction

Restaurants have always played an essential role in the business, social, intellectual and artistic life of a thriving society. The major events of life, personal and professional, are celebrated in restaurants. Right now, restaurants are more important than ever. Restaurants today lie at the heart of 21st-century American life. And for the foreseeable future, millions of Americans will wait tables, cook food, or wash dishes for their livelihoods. Currently, there are many restaurants in the city of Minneapolis and many more are being built. Of course, as with any business decision, opening a new restaurant requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the restaurant is one of the most important decisions that will determine whether the location will be a success or a failure. For investors, choosing a restaurant location is one of the more permanent choices a restaurant owner makes. Restaurant location influences the success or failure of a restaurant in a host of ways, from attracting enough initial customer interest to being convenient to visit. But the restaurant's location is also interrelated to other factors, some of which are changeable, while others are not. So making a snap decision without doing any research may leave an owner with a location he or she may later regret.

Business Problem

The objective of the project is to analyze and select the best location in the city of Minneapolis to open a new restaurant. Using data science methodology and machine learning techniques, this project aims to provide solutions to answer the business question: In the city of Minneapolis, Minnesota, if a property developer is looking to open a new restaurant, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful for investors looking to open a new restaurant in Minneapolis, Minnesota. This project is timely as the city has more and more restaurants to be built in the following year. Data from the 2018 Minnesota Quarterly Census of Employment and Wages showed that an additional 15 per cent will be added to existing plazas. In addition, the Minnesota Employment and Economic Development agency shows that the current restaurant shows different spatial distribution properties, highly depending on population demographics, leasing rate and other important factors. So the project here will further identify the spatial clustering characteristics of restaurant and help investors for their decision-making.

Data

To solve the problem, we need the following data:

- List of suburbs in Minneapolis. This defines the scope of this project which is confined to the city of Minneapolis, Minnesota
- Latitude and longitude coordinates of suburbs. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to restaurants. We will use this data to perform clustering on the suburbs.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Hennepin_County,_Minnesota) contains a list of suburbs in Minneapolis, with a total of 45 suburbs. We will use web scraping to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the suburbs using Python Geocoder package which will give us the coordinates of the suburbs.

After that, we will use Foursquare API to get the venue data for those suburbs. Foursquare has one of the largest database of 100 million places and is used by over 100 thousands developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the restaurant category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping, working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we obtain the list of suburbs in the city of Minneapolis from the following Wikipedia page (https://en.wikipedia.org/wiki/Hennepin_County,_Minnesota). We use web scraping with the requests and Beautifulsoup Python packages to extract the list of suburbs data. Then we use Geocoder package to get the geographical coordinates in the form of latitude and longitude which are required by Foursquare API. After gathering the geographic coordinates, we convert the data into pandas DataFrame and then visualize the suburbs in a map using Folium package. This allows us to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Minneapolis.

Next, we use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. By making API calls to Foursquare passing in the geographical coordinates of the suburbs, Foursquare returns the venue data in JSON format and we extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each suburb and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each suburb by grouping the rows by suburb and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Restaurant” data, we will filter the “Restaurant” as venue category for the suburbs.

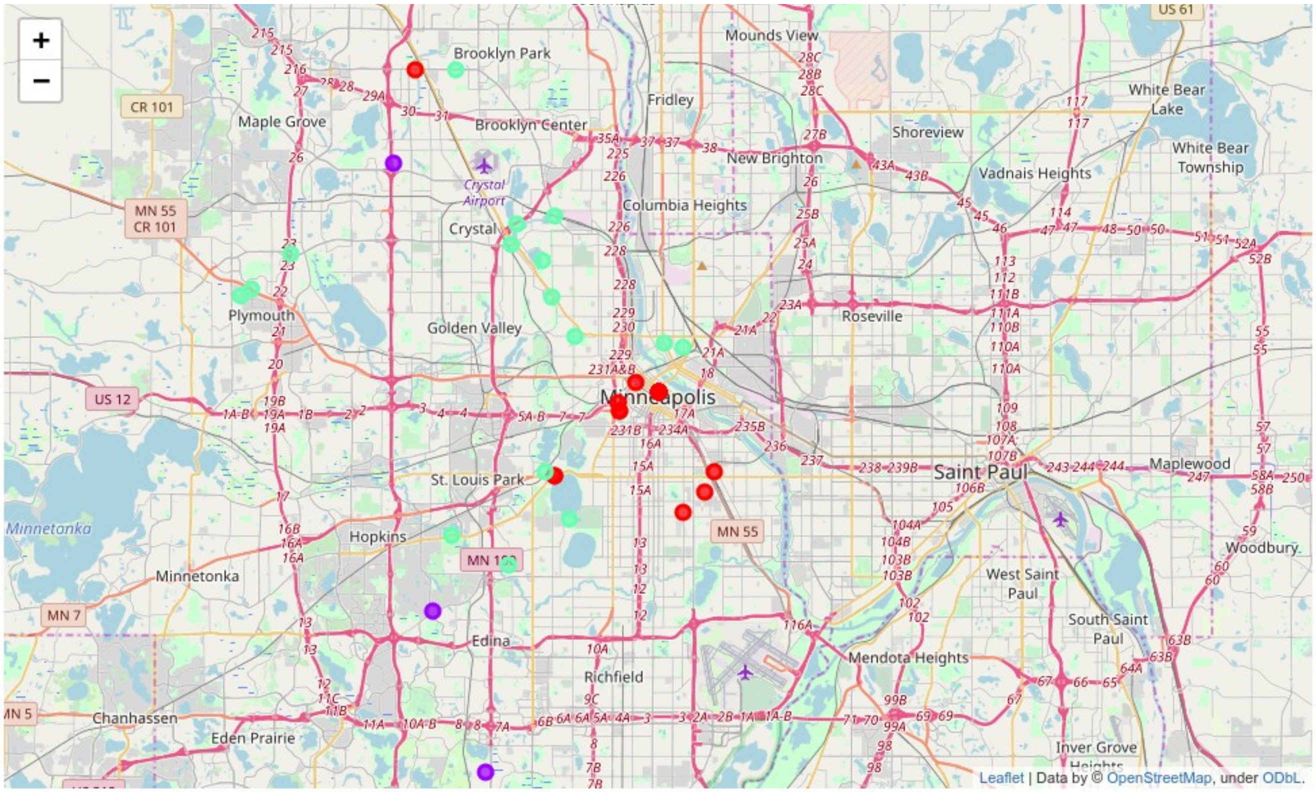
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the suburbs into 3 clusters based on their frequency of occurrence for “restaurant”. The results will allow us to identify which suburbs have higher concentration of restaurants while which suburbs have fewer number of restaurants. Based on the occurrence of restaurants in different suburbs, it will help us to answer the question as to which suburbs are most suitable to open new restaurants.

Results

The results from the k-means clustering show that we can categorize the suburbs into 3 clusters based on the frequency of occurrence for “restaurant”:

- Cluster 0: Suburbs with high number of restaurants
- Cluster 1: Suburbs with low number to no existence of restaurants
- Cluster 2: Suburbs with moderate number of restaurants

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.



Discussion

As observations noted from the map in the Results section, most of the restaurants are located in the central area of Minneapolis, with the largest number in cluster 0 and moderate number in cluster 2. In contrast, cluster 1 has very low number restaurants in suburbs. This indicates a great opportunity and high potential areas to open news restaurants because there is very little to no competition from existing restaurants. Meanwhile, restaurants in cluster 0 are likely suffering from intense competition due to the large number of restaurants. The results also show that the oversupply of restaurants mostly in the central area of Minneapolis, with the suburb area still have very few restaurants. Therefore, this project recommends investors to open new restaurants in suburbs within cluster 1. Property developers can also open new restaurants in suburbs within cluster 2 with moderate competition, and avoid suburbs in cluster 0 which already have large amount of restaurants.

Limitations and Suggestions for Future Research

In this project, we only considered the frequency of occurrence of restaurants from Foursquare API, there are other factors such as population demographics and residents income which could also affects the location decision of a new restaurant. Future research should incorporate more relevant data to be used in the clustering algorithm to make a better decision. In addition, we didn’t differentiate the restaurant type here. What we did is to consider all restaurant types. There are could be bias in the analysis, since some suburbs may have more specific restaurant type (e.g. American restaurant) but in fact these suburbs has lower total restaurant numbers. Finally, this project used of the free Account of Foursquare API that came with limitations as to the number of API calls. Future research could make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have identified the business problem, specified the data required, extracted and processed the data, performed machine learning by clustering the data into 3 clusters based on their frequency of occurence, and lastly provided recommendations to investors regarding the best locations to open a new restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: the suburbs in cluster 1 are the most preferred locations to open a new restaurant. The findings of this project will provide investors the insights about high potential locations while avoiding intense competition areas.

References

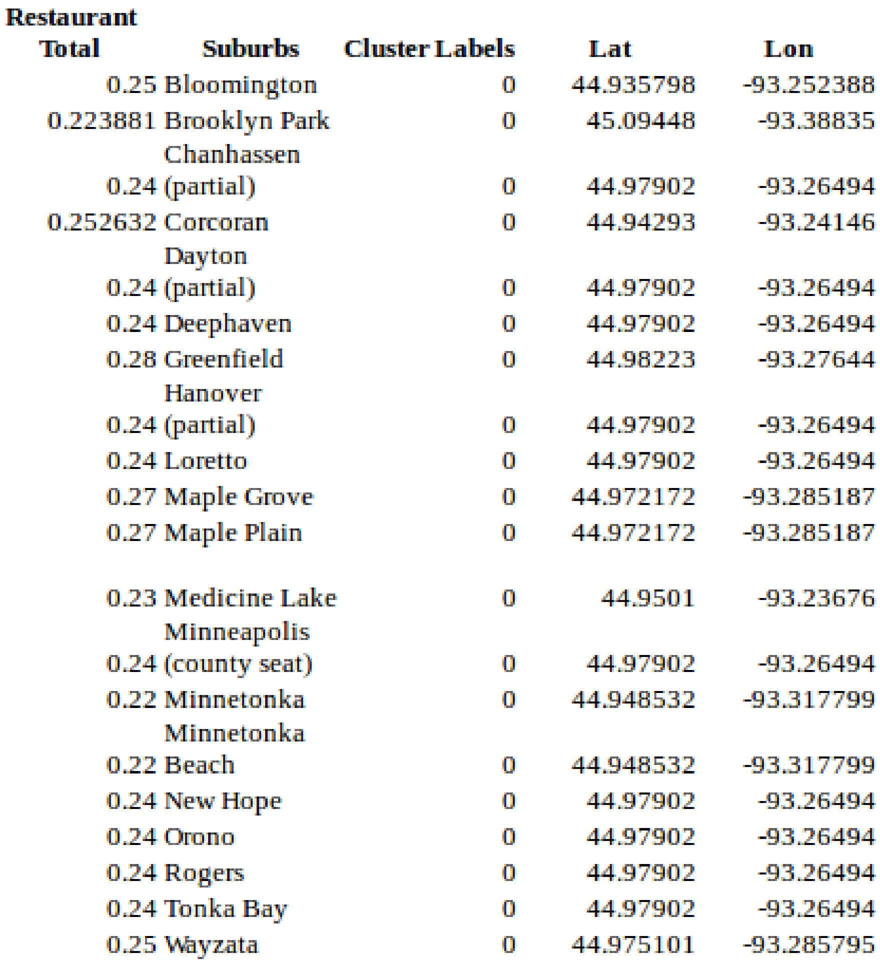
Hennepin County, Minnesota
 https://en.wikipedia.org/wiki/Hennepin_County,_Minnesota

Foursquare Developers Documentation. Foursquare. Retrieved from https://developer.foursquare.com/docs

2018 Minnesota Quarterly Census of Employment and Wages
https://mn.gov/deed/data/data-tools/qcew/

Appendix

Restaurants in Cluster 0



Restaurant Total	Suburbs	Cluster Labels	Lat	Lon
0.25	Bloomington	0	44.935798	-93.252388
0.223881	Brooklyn Park	0	45.09448	-93.38835
0.24	Chanhassen (partial)	0	44.97902	-93.26494
0.252632	Corcoran	0	44.94293	-93.24146
0.24	Dayton (partial)	0	44.97902	-93.26494
0.24	Deephaven	0	44.97902	-93.26494
0.28	Greenfield	0	44.98223	-93.27644
0.24	Hanover (partial)	0	44.97902	-93.26494
0.24	Loretto	0	44.97902	-93.26494
0.27	Maple Grove	0	44.972172	-93.285187
0.27	Maple Plain	0	44.972172	-93.285187
0.23	Medicine Lake	0	44.9501	-93.23676
0.24	Minneapolis (county seat)	0	44.97902	-93.26494
0.22	Minnetonka	0	44.948532	-93.317799
0.22	Minnetonka Beach	0	44.948532	-93.317799
0.24	New Hope	0	44.97902	-93.26494
0.24	Orono	0	44.97902	-93.26494
0.24	Rogers	0	44.97902	-93.26494
0.24	Tonka Bay	0	44.97902	-93.26494
0.25	Wayzata	0	44.975101	-93.285795

Restaurants in Cluster 1

Restaurant				
Total	Suburbs	Cluster Labels	Lat	Lon
0.055556	Eden Prairie	1	44.900383	-93.379409
0.108696	Hopkins	1	44.811134	-93.286478
0.033333	Independence	1	45.060778	-93.399329
0	Minnetrista	1	44.957607	-93.650264
0	Mound	1	44.957607	-93.650264
0.076923	Shorewood	1	44.801397	-93.309462
0.0625	St. Bonifacius	1	44.90409	-93.74419
0.097561	Woodland	1	44.842149	-93.353159

Restaurants in Cluster 2

Restaurant				
Total	Suburbs	Cluster Labels	Lat	Lon
0.205882	Brooklyn Center	2	45.094393	-93.367998
0.157895	Champlin	2	45.17053	-93.392277
0.173913	Crystal	2	45.026042	-93.323711
0.16	Edina	2	44.91645	-93.341182
0.125	Excelsior	2	44.927425	-93.370291
0.169492	Golden Valley	2	44.998758	-93.307808
0.157895	Greenwood	2	45.012919	-93.318908
0.206897	Long Lake	2	45.039022	-93.336887
0.164835	Medina	2	45.013158	-93.476999
0.2	Osseo	2	45.041994	-93.317944
0.16	Plymouth	2	45.015579	-93.471502
0.19	Richfield	2	44.933421	-93.310531
0.19403	Robbinsdale	2	45.03178	-93.33998
0.169492	Rockford (partial)	2	45.028237	-93.451479
0.19	Spring Park	2	44.995077	-93.252461
0.2	St. Anthony (partial)	2	44.99617	-93.26199
0.17	St. Louis Park	2	44.950067	-93.322706