2020

Cluster Analysis Using Geographic Data

Dinesh Ogirala
Grand Valley State University

CLUSTER ANALYSIS USING GEOGRAPHIC DATA: TO DETERMINE THE BEST POSSIBLE LOCATION FOR A COFFEE SHOP IN GRAND RAPIDS

DINESH OGIRALA

A Project Submitted to

GRAND VALLEY STATE UNIVERSITY

In

Partial Fulfillment of the Requirements

For the Degree of

Master of Science in Applied Computer Science

School of Computing and Information Systems

December 2020

The signatures of the individuals below indicate that they have read and approved the project of Dinesh Ogirala in partial fulfillment of the requirements for the degree of Master of Science in Applied Computer Science.

_____

<name of project advisor>, Project Advisor Date

_____

<name of GPD>, Graduate Program Director Date

_____

<name of unit head>, Unit head Date

Abstract

Many businesses suffer from losses after establishing their business due to a lack of proper research before deciding on a new establishment location. The method proposed in this paper can land on the best possible location for a new establishment by web scraping a target list of neighborhoods from [1] using beautifulsoup library [2], and passing this list to geocoder library [3], to retrieve a list of geographical coordinates. API calls are made to Foursquare API [4] with each coordinate as parameter which returns a JSON output consisting all the venues around. After various stages of pre-processing such as data cleaning, normalization and feature engineering are done on the data, this data is fed to a clustering algorithm such as K-means clustering [5]; an unsupervised learning technique which strives to choose its centroids to minimize the inertia in the given data. The number of centroids in K-means clustering is determined by utilizing the two methods namely, Silhouette and Elbow method. The best location is determined by scrutinizing the frequency of coffee shops, hence, the competition/demand of coffee shops in the area and suggest the best possible spot for a new coffee shop. Grand Rapids is chosen as the location for this project. Of course, just like any other business decision, opening a new coffee shop requires various other factors to be considered, such as the audience in that area or any schools around. Nevertheless, determining a location for the new establishment is the primary step that any individual would think of.

Introduction

A coffee shop is an excellent spot for any two individuals to collaborate and work on something, especially for students working on a project. Not only for work, but a coffee shop is also a great place to hang out with friends. For such a place, the location would be a crucial point, to begin with, because establishing a new coffee shop around ten coffee shops would be less profitable. Unless it is a monopoly and has some brand value, it is again a challenging task to achieve. One could search for coffee shops on Google and determine the best location for a new establishment, but that would be a very naïve approach. An individual must be looking at a maximum of 3 to 4 coffee shops on the map. However, a machine learning algorithm takes in 100 coffee shops to determine the best possible location.

The initial project idea was to determine the best location for a new coffee shop using a list of Grand rapids neighborhoods that is web scraped from [1] and passing this list to geocoder library [3], to retrieve a list of geographical coordinates. API calls are made to Foursquare API [4] with each coordinate as parameter which returns a JSON output consisting all the venues around. The JSON result set from the API gave out a lot more information than required, including the street address, Zip code of the venue. After each stage of the project, it was evident that many improvements can be made to predict the best location better. Many of the improvements were already included in this project. Integrating other public datasets such as Census data and a few more improvements have been left out due to its extreme complexity, if included.

Project Management

My estimations regarding timelines were on point during the initiation phase of the project regarding identifying the problem that needs to be solved. However, as the project progressed into the planning and execution stages, there was a slight hindrance due to a lack of resources and unexpected fixes to code.

Organization

To decipher the enigma, we will require the following data:

- Data from web-scraping [1], which is a list of Grand Rapids neighborhoods.
- Getting geographic coordinates of those neighborhoods using the geocoder library.
- Passing the neighborhood coordinates to Foursquare API to retrieve a list of places around each neighborhood for clustering.

Initially, data is extracted (web scaped) from the [1] containing all the neighborhoods, followed by plotting the coordinates on a map using the folium library [6] shown in Figure 2. Now, API calls are made to Foursquare API, passing each neighborhood coordinate, which returns the top 500 popular venues in 1.5 Kilometer radius. All the venues returned are scrutinized and verified if they contain coffee shop data. One-hot encoding is done on the venue data to transform the categorical venue data into numerical data for clustering purposes and mean of occurrence of each category is calculated, shown in Figure 1.

	Neighborhoods	Coffee Shop
1	Auburn Hill	0.021277
2	Baxter	0.060000
3	Belknap Lookout	0.030000
4	Black Hills	0.048780
6	Creston	0.015873

Fig. 1. *Structure of Data before Passing onto K-means*

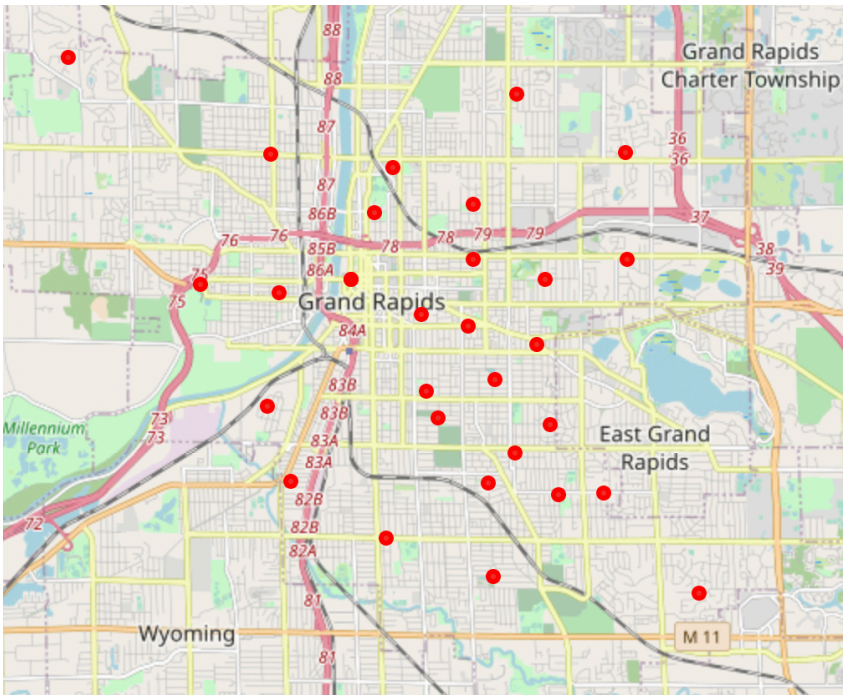

Fig. 2. *Data Locations plotted using Folium Library*

To determine the number of centroids for the k-means algorithm [8], Elbow method [7] and Silhouette method are employed. Both these methods aim to choose the centroids to minimize the inertia or sum-of-squares error in the data. Sum of Squares error formula is given below,

μ=centroid (mean of data points), x=sample space,

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2)$$

Now, the model is trained with the appropriate number of centroids, and clusters are formed. The best cluster is determined by the frequency of coffee shops in it. Hence, based on the mean frequency number of coffee shops at a location in the cluster, the best location is determined.

Reflection

Initially, I had to search for a source to get a list of neighborhoods.After some research, I landed on [1] and was able to extract the data. To make the API calls, I used the Foursquare API, but there are many other alternatives. Moreover, it was a free developer account at Foursquare and not a premium account. We get additional details such as venue ratings, photos, etc for a premium API call, which can be very useful in determining the competition (coffee shops) in an area. There were so many outliers in the data which had to be removed because K-means clustering would not be useful with outliers. An improvement would be using the DBSCAN clustering algorithm, which is very effective in managing outliers. Moreover, DBSCAN has the upper hand in fitting non-spherical clusters. The only reason K-means clustering is employed is to make time for other parts of the project.

When making the API calls, there is a cap on the radius and number of places returned. After some research, I realized that the same venue is returned in more than onc neighborhood because of the large radius value. So, after trimming down both the radius of the search and the number of venues, I managed the data by neighborhoods and verified the data does not contain duplicates. K-means clustering was executed on default parameters.

The model would make a better prediction if there were more data, for example, some more locations in Grand Rapids. The scope of this project can be further scaled out by integrating various other datasets.

Conclusions

The results after the application of the algorithm would generate 5 clusters, shown in Figure 3, based on the density of the coffee shops.

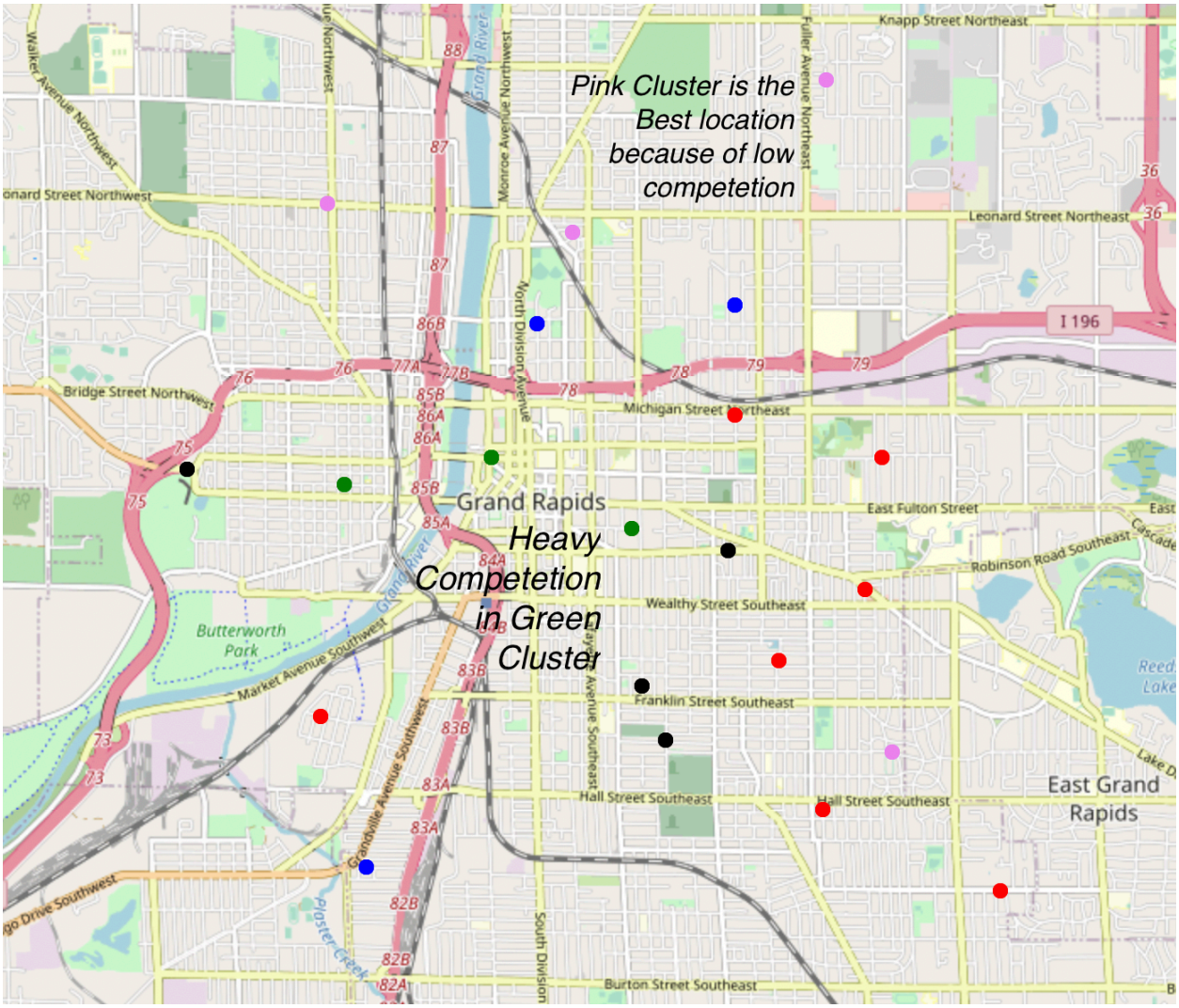

Fig. 3. *Resulting Clusters after K-means*

Cluster 0

	Neighborhood	Coffee Shop	Cluster Labels
1	Auburn Hill	0.021277	0
6	Creston	0.015873	0
23	Ottawa Hills	0.016393	0
30	West Grand	0.015625	0

Fig. 4. *Cluster 0 – Best Cluster because of low frequency of Coffee Shops*

Cluster 2

	Neighborhood	Coffee Shop	Cluster Labels
13	Heartside	0.090000	2
14	Heritage Hill	0.090000	2
29	Swan	0.082474	2
31	Westside Connection	0.090000	2

Fig. 5. *Cluster 2 – Worst Cluster because of high frequency of Coffee Shops*

- Cluster 0 consists of the best possible locations for a new coffee shop
- Cluster 2 represents the worst possible locations because of heavy competition.

However, this project provides the basic initial step for a new establishment, and anyone considering opening a new coffee shop must do further research. I intend to perform further development by adding age data from the Census Data and better predicting the new establishment's profitability. By this developement, the algorithm can better determine its audience rather than solely depend upon the density of coffee shops in an area.

Appendices

[1] Neighborhoods in Grand Rapids. (n.d.). Areavibes. Retrieved November 5, 2020, from https://www.areavibes.com/grand+rapids-mi/neighborhoods/

[2] BeautifulSoup. (n.d.). Retrieved October 14, 2020, from https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[3] Geocoder: Simple, Consistent¶. (n.d.). Retrieved October 14, 2020, from https://matplotlib.org/3.3.3/tutorials

https://scikit-learn.org/0.21/modules/clustering.html#k-means

[4] Places API. (n.d.). Foursquare API. Retrieved November 5, 2020, from

https://developer.foursquare.com/docs/places-api/

[5] K-Means. (n.d.). Scikit-learn. Retrieved November 5, 2020, from

https://scikit-learn.org/0.21/modules/clustering.html#k-means

[6] Folium. (n.d.). Retrieved October 05, 2020, from https://python-visualization.github.io/folium/modules.html

[7] Spatial algorithms and data structures (scipy.spatial)¶. (n.d.). Retrieved November 03, 2020, from https://docs.scipy.org/doc/scipy/reference/spatial.html

[8] Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*, *2*(2), 226–235. doi:10.3390/j2020016

[9] Saxena, E. (2020, April 28). Exploring the Taste of NYC Neighborhoods. Retrieved November 06, 2020, from https://towardsdatascience.com/exploring-the-taste-of-nyc-neighborhoods-1a51394049a4