Segmenting and Clustering Neighborhoods in New York

Kurma Jhansi Naveena, Hema Mounika Kotte, Vinay Kumar Karingula
Data Mining and Analysis for Managers (MGMT 635 – 851)
New Jersey Institute of Technology

1. Summary

Information about neighborhoods of New York City are utilized to conduct segmentation and clustering nearby venues. The following analysis converts addresses to their respective latitude and longitude values. Foursquare API is utilized to explore nearby venues in neighborhoods. The **explore** function is used to get the most common (or most visited) venue categories in each neighborhood, and then use this feature to perform clusters. To perform clustering, KMeans clustering algorithm is used to complete the task. Finally, Folium library is used to visualize the neighborhoods and the clusters formed.

2. Introduction

The initial dataset is an unstructured data, namely a JSON file. The primary data is stored in the features section of the JSON file. The data is extracted using json library and converted into a data frame. The dataset consists of borough, neighborhood, latitude and longitude.

Folium library is utilized to visualize geospatial data, with its unique interactive features. One of the main advantages of Folium is that it is completely free, while other geospatial visualization tools keep information on how many API calls are made.

The Foursquare API takes calls from a URL generated by information of geospatial data. A URL is constructed to send a request to the API to search for specific venues, explore the venues, Foursquare users, and to get trending venues around a location.

Since the chosen dataset doesn't have a target labeled, clustering algorithm is implemented, namely KMeans clustering algorithm. The goal of current analysis is to segment and cluster like – venues. Therefore, KMeans is one of the clustering algorithms that is efficient for segmentation.

3. Data mining

a. KMeans clustering [1]

KMeans is has the reputation of being the simplest and popular unsupervised algorithm. The main objective of this algorithm is to group similar data points together and uncover hidden patterns. To achieve this, KMeans looks for a specific number of clusters (k). A target number *k* is defined, which refers to the number of centroids is needed. The centroid is an imaginary location representing center of a cluster. Each data point is then assigned to a cluster that is nearest.

To process the data, the algorithm starts with a group of random, or known, centroids, which are used as initial points for clusters. The process becomes iterative as the calculations optimize the positions of the centroids. The process halts when either

- The centroids are stabilized, i.e. if the positions of the centroids remain unchanged.
- The defined number of iterations has been achieved.

The algorithm is primarily used for data cluster analysis that possess similar characteristics. However, choosing the right value of *k* is always uncertain and unpredictable. As the entire algorithm depends on the number of clusters, it could either lead to overfitting or underfitting; but the delivery of results is quick.

4. CRISP – DM

a. Business understanding

The latitude and longitude of the neighborhoods determine the nearby venues, which eventually are segmented and clustered by their similarity. Such analysis depicts how travel sites, Google Maps, and primarily Foursquare applications work and produce the expected results.

b. Data understanding

The JSON file consists of all the information on New York from borough to their neighborhoods to their geospatial data. As it is unstructured, the data is loaded as a json file initially, and relevant data is extracted that contains necessary information. Once the data is converted into a data frame, the data frame contains 4 features and 306 records, in which it contains 5 boroughs. To look in detail, Manhattan borough is selected to find its neighborhoods' nearby venues, then extract top 10 venues from them using Foursquare API endpoints.

c. Data preparation

Data preparation stage involves exploring the data further and making sure that it is in right format for the chosen machine learning algorithm. The dataset is checked for any missing values. Since the dataset doesn't contain any missing values, handling of missing values isn't necessary.

Once naming conventions are performed, URL that explores top 100 venues within 500-meter radius is generated. The results are stored in a JSON format. Once the results are examined and extracted, the venue category is extracted using a user – defined function. A new data frame is created that contains name of the neighborhood, respective coordinates, and venue details. The data frame then undergoes one – hot encoding and grouped by neighborhood and its average to prepare for clustering.

d. Model building

Data modeling answers two key questions:

- What is the purpose of data modeling?
- What are some characteristics of the process?

Data modeling focuses on developing models that are either descriptive or predictive. Since the Manhattan dataset is a descriptive modeling, it doesn't require training and testing sets.

The selected algorithm for the dataset is KMeans algorithm used for clustering problems. The chosen number of clusters is 5.

e. **Testing and evaluations**

This model isn't built for evaluation; however, it is meant to uncover underlying patterns. In this case, the model is about finding nearby most commonly explored venues within neighborhoods. Once the clusters are formed, a map is constructed using Folium package to visualize the clusters.

f. **Deployment**

Deploying such program tells us not only about trending venues and explored venues, but also shows the behavior of the users – check-ins, tips posted, likes, number of times visited, venues visited, etc. Given the location details to a travel site or a map, can produce us results such as nearby venues based on categories.

g. **Results**

The clusters produce the following results.

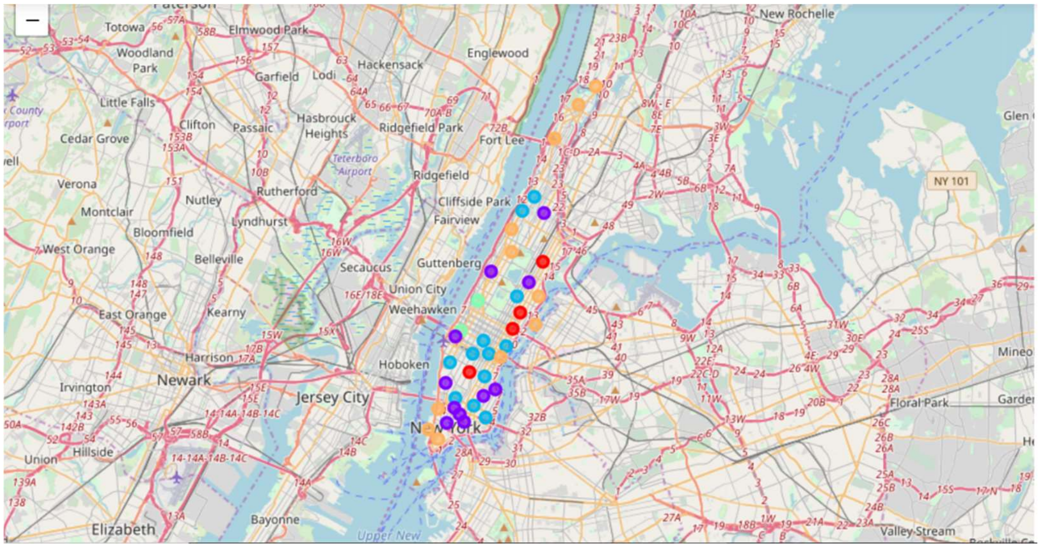

Fig 1. Cluster Map

5. Conclusions

In this report, a KMeans clustering algorithm is implemented on New York's geospatial data, specifically Manhattan data. The implementation and methods are further explained, and CRISP – DM technique is described in correspondence to the business problem. Maps are generated to view the neighborhoods and clusters within neighborhood.

For future reference and applications, such analysis provides us the insights on customer's behavior on traveling, check – ins, and visited venues. Furthermore, a recommendation engine can be added to the existing system to enable content-based recommendation or collaborative filtering.

6. References

[1] “Understanding K-means Clustering in Machine Learning” Dr. Michael J. Garbade, Sep 12, 2018 https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.