**Clustering and Grouping Data**

David Yedlin Broom

Northcentral University

TIM\_8501: Exploratory Data Analysis

Dr. Marty Crossland

October 2, 2022

**Clustering and Grouping Data**

Some useful ways to find patterns in a data set are clustering methods.  Na et al. (2010) states there is an increasing demand for clustering methods in various areas of business, government, science, and technology. Clustering is also helpful for finding hidden or complex patterns in large data sets.  A well known method for clustering data is called the k-means algorithm which involves unsupervised machine learning.

Clustering is a way to group data by its similar properties.  Several questions should be answered before beginning a cluster analysis.  The data should be explored and understood before starting a clustering analysis.  The shape and size of the data is relevant for planning a cluster analysis.

It would be important to understand how the various forms of data are related or if data is not comparable in some way.  Next it would be important to see if there is a need to standardize the data.  The standardization of data can help cluster methods run more efficiently.  Finally it is important to find if any variables are correlated.  If there is a correlation then that correspondence should be investigated to see how the relational properties of that data could be utilized.

The first step in cluster analysis is to identify and label a problem.  Then a similarity measure is chosen.  After a similarity method is chosen it is best to choose the grouping method.  The number of clusters should be determined to implement in the grouping method.  Finally the clusters should be analyzed and described.

Sometimes with large amounts of data it is difficult to see if two variables will be suitable for clustering.  The principle component analysis is useful for determining if two variables are appropriate for clustering analysis.  There are four steps in principal component analysis.  Each step will help reduce a number of variables to only one most related pair.

Principal component analysis should be performed before starting a cluster analysis.  The steps of principal component analysis in order are standardization, covariance matrix, eigenvalues and vectors, and last is feature.  The main purpose of principal component analysis is to reduce the dimension.

A principal component analysis was not performed on this data set.  There are a limited number of variables with discrete or continuous numerical data.  The only two variables that have continuous data are longitude and latitude.

The k-means algorithm first randomly selects a point in the data.   It begins an iterative process to converge to a local minimum. The algorithm uses the Euclidean distance as a measure.  The result is a desired number of tight clusters that can be studied further.  Alsabti et al. (1997) states the k-means method is shown to be effective in making good clusters for many different applications.

The data set contains 1.01 million rows and thirty-five columns related to forestry planting spaces for New York City parks.  The last update for the data set was in September of 2022.  The data was created and made for public use in March of 2017.  The data is collected by the Forestry Management System.

Several of the columns describe facts about the location of a planting space.  There is location information in the form zip code, address, and geographical coordinates.  The only location information in the set that is numerical and continuous is geographical coordinates.  The other fields contain only nominal data.

The geographical data is listed in several forms.  The is a separate longitudinal and latitudinal field.  These two fields would be good for doing a cluster analysis since the data is quantitative and continuous.

There are two additional fields with discrete numerical data involving length and plot width of planting spaces.  Many entries contain a plot width and length of zero for both fields.  Since there is not a very strong relation to any particular length or width these two fields would not be useful for a clustering analysis.

The other fields in the data set contain information related to the name of the park, political jurisdiction, and identification number.  Each space contains only one tree to one planting space.  The data set could be useful for improving the city’s aesthetics and managing all trees that belong to the city.

The data set provides a lot of useful information about trees that are maintained by the New York City government.  For management purposes it would be useful to divide the tree points up into groups containing a similar geography.  What would be the number of groups?  What characteristics would be important for each group?

The K Means clustering method was used to form different clusters of the different latitude and longitude methods.  The difference between the points is a Euclidean Distance similarity measure.  An elbow method can be used to find the number of clusters that should be used.

             Python was used to do all calculations on the data set.  First the data set was curated.  When performing the k means clustering method it is important to remove all null values from the set.  The DROPNA() function was used to remove all null values.  A separate data frame was created with only columns with only relevant data.

The longitude and longitude coordinates were standardized with the STANDARDSCALER() function.  That transformed standardized data was placed into two columns named Latitude\_T and Longitude\_T.  Next the elbow method was used to determine what the appropriate number of clusters should be.  A function named OPTIMISE\_K\_MEANS() was created. The OPTIMISE\_K\_MEANS() function took in two parameters DATA and MAX\_K.  In the function a for loop ran various k-means methods for different amounts of cluster.

The k means algorithm was implemented with the KMEANS() function.  The KMEANS() function is part of the SKLEARN.CLUSTER library.  A new variable name KMEANS\_3 was populated with assigned cluster numbers.  The clusters were then plotted with the PLT.SCATTER() function with the original latitude and longitude values.

A for loop was used to create a small table of counts and percentages for each cluster.  In the for loop the .LOC() function was used to isolate each cluster.  The LEN() function was used to find the number of entries for each cluster.   Another for loop along with the VALUE\_COUNTS() function was used to find the counts of all Boroughs, Parks, and Street values of each cluster of tree spaces.

The data set row count before any curation was 1014386.  After all null values were removed from the latitude and longitude variables there were 792605 rows.  Twenty one percent of the data set had null values for latitude and longitude.  Before and curation took place the number of Boroughs with tree spaces with labels Queens, Brooklyn, Bronx, Staten Island, and Manhattan was 343996, 267109, 148337, 144720, and 109711 respectively.  The number of tree spaces before any curation labeled inside a park or street was 191403 and 822954.

             After curation the most noticeable difference was in tree space with the PSSite variable labeled park with only 403 entries.  99.789% of all entries of the PSSite labeled park had null values for the latitude and longitude field.  After curation the labels for different Boroughs labeled Queens, Brooklyn, Staten Island, Bronx, and Manhattan were 282753, 210582, 124803, 101737, and 72730 respectively.

Chart, line chart

Description automatically generated

             It would be best to use five for the amount of clusters which was found by using the elbow method. The curve in the graph of mean straightens greatly after the number five.   The clusters had counts of entries from cluster labeled zero to four at 150832, 124804, 203979, 183266, and 129755 respectively.  The percent of data for clusters zero to four was 19.0%, 15.7%, 25.7% 23.1% and 16.4% respectively.

The graph of clusters was well defined and all groups were easy to see.  The range of values for latitude and longitude was about [40.4, 41.0] and [-73.6,-74.4] respectively.  The clusters were displayed in different colors.

Icon, map

Description automatically generated

Calculations were performed with data in each of the five clusters.  The clusters were well formed and showed a systematic way to cluster the geography of tree spaces.  Each cluster contained almost a similar count of data.  Without a clustering algorithm it would have been very difficult to organize over 750,000 location points into five separate neatly organized groups.

**References**

Alsabti, K., Ranka, S., & Singh, V. (1997). An efficient k-means clustering algorithm. *Electrical Engineering and Computer Science - All Scholarship*, 1. https://surface.syr.edu/eecs/43

Na, S., Xumin, L., & Yong, G. (2010). *Research on k-means clustering algorithm: an improved k-means clustering algorithm* [Paper]. 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Jian, China.