STAT 8230 – Applied Multivariate Data Analysis

Final Project – Comparison of Clustering Analysis Methods for Identifying Students with Disabilities

May 3rd, 2021

Connor Armstrong

Description of Data

The dataset for this analysis was published to data.gov on November 10th, 2020 and contains a variety of information related to math test results for New York City schools, grades 3 through 8, disaggregated by borough and disability status.

The scores are divided into 4 categories, Levels 1 through 4. Level 1 means the student did not meet learning standards. Level 2 means the student partially met learning standards. Level 3 means the student met learning standards. Level 4 means the student met learning standards with distinction.

Each observation of the dataset contains Borough (Location), Grade Level, Year, Demographic (With/Without disabilities), Number of students tested, Mean score, Number of students in each "Level", and Percentage of students in each "Level".

The following table contains descriptions of each column in the provided dataset:

Column	**Description**
Borough	City Borough from which the data was collected
Grade	Denotes Grade Level
Year	Year test was administered
Demographic	Students with/without disabilities
Number Tested	Total number of students tested in that Borough, in that Grade, in that Year, in that demographic
Mean Scale Score	Average score of students tested
Num Level 1	Number of students who scored in Level 1 range
Pct Level 1	Percentage of students who scored in Level 1 range
Num Level 2	Number of students who scored in Level 2 range
Pct Level 2	Percentage of students who scored in Level 2 range
Num Level 3	Number of students who scored in Level 3 range
Pct Level 3	Percentage of students who scored in Level 3 range
Num Level 4	Number of students who scored in Level 4 range
Pct Level 4	Percentage of students who scored in Level 4 range
Num Level 3 and 4	Number of students who scored in both Level 3 and 4 range
Pct Level 3 and 4	Percentage of students who scored in both Level 3 and 4 range

Description of Research Question

It is assumed that students with disabilities will perform differently on math tests than those without disabilities. Evaluating the extent of this difference could be performed in a variety of ways, and could be performed without multivariate analysis techniques. This analysis will be performed to evaluate a variety of clustering techniques for their ability to identify populations within this dataset of students either with or without disabilities. Hotelling's T^2 will be used to compare the mean vectors of the number of students in each Level and to evaluate the hypothesis that the populations from which the data were taken have similar characteristics with respect to the numbers of students in each Level. The accuracy of the clustering techniques used will depend largely on the extent to which students with disabilities perform differently on math tests than those without disabilities.

The clustering algorithms used in this analysis are classified as hierarchical, agglomerative clustering algorithms. Hierarchical clustering algorithms operate by clustering groups of similar objects into groups called clusters, and these algorithms can either be agglomerative or divisive. Divisive algorithms start with a single cluster that is divided into smaller ones, while agglomerative clustering algorithms start with many small clusters and merge them into bigger clusters.

Agglomerative clustering, as mentioned, operates by merging smaller clusters into larger ones. The decision to merge clusters together is made by grouping the clusters which are the closest together at each step. This distance can be defined in a number of ways, and depending on how distance is defined are the various "linkage" criteria defining the next classification of clustering algorithms. Generally, distance is defined as the Euclidean distance between points (ie the square root of the sum of the squares of their respective values) but can also be defined as the Manhattan distance (the minimum distance between points not allowing for travel at anything but 90 degree angles, similar to travelling in a city with parallel and perpendicular streets).

Complete linkage clustering calculates the distance between clusters as the longest distance between any two points in each cluster. Average linkage clustering calculates the distance between clusters by taking the average distance between each point in one cluster and every other point in the other cluster. Ward linkage clustering defines distance as the sum of squared differences within the clusters.

The specific clustering algorithms which will be evaluated for this paper are:

1. Complete linkage hierarchical agglomerative clustering with Euclidean distance
2. Average linkage hierarchical agglomerative clustering with Euclidean distance
3. Ward linkage hierarchical agglomerative clustering with Euclidean distance

Python will be used for the clustering analysis specifically, while R will be used for data manipulation, evaluation, plots, and figures.

The remainder of this paper will be dedicated to answering the following question: *Which clustering algorithm of the 3 defined above will perform best in determining which populations of students from the New York City schools either had or did not have disabilities based on their distribution of math scores in the four performance Levels 1, 2, 3, and 4.*

Presentation of Statistical Analysis of Data

The dataset contains data for grades 3 through 8, but also contains the totals for all grades for each Borough, Year, and Demographic. These were removed so that each observation contained distinct results not depending on the others.

Similarly, the columns which combine the scores for Levels 3 and 4 (“Num Level 3 and 4” and “Pct Level 3 and 4”) were removed so that the columns used for analysis would contain unique students for each combination. Percentage columns are dependent on the scores in the others, and were therefore removed. The decision to only evaluate the number of students in each category was made on this basis, and so that the dimensions of the clustering analysis can be interpreted as representing a similar concept. Therefore, the four columns used as input to the selected clustering analysis algorithms are for the number of students in each “Level” (“Num Level 1”, “Num Level 2”, “Num Level 3”, and “Num Level 4”). The summary statistics of these variables are described in the following:

Statistics by Level for All Students

Level	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
1	7	253	525.5	664.6	919	4024
2	75	696	1320	1890	2311	7929
3	64	595.8	1576	3077.2	5247	12165
4	0	91.75	431	1511.66	2217	7593

Statistics by Level for Students with Disabilities

Level	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
1	42	282.5	574.5	660.5	929.5	1813
2	183	519.5	856.5	946.2	1438.8	1882
3	64	356.8	595.5	712.6	968.8	2295
4	0	42.25	91.5	118.67	169.75	449

Statistics by Level for Students without Disabilities

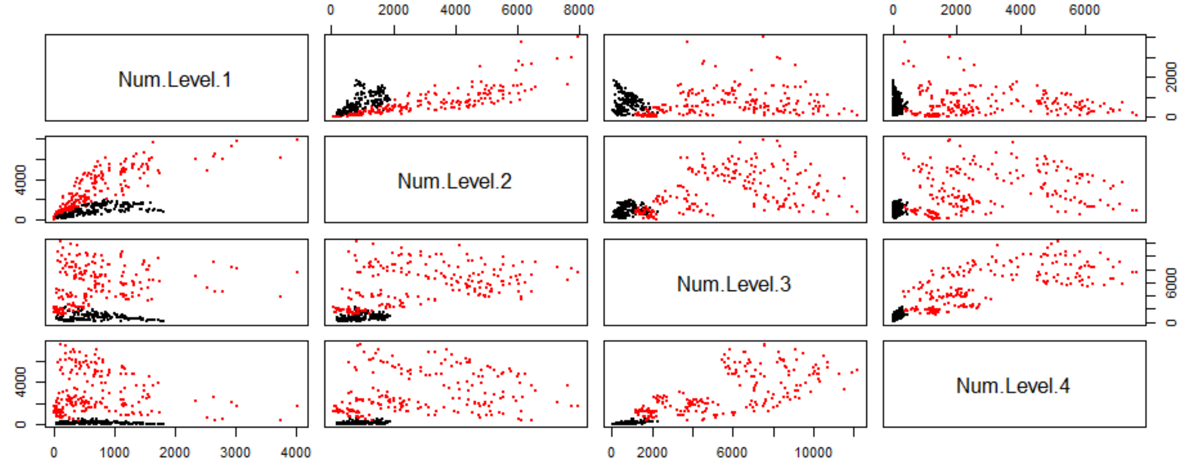
Level	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
1	7	191.5	485	668.7	917.8	4024
2	75	1034	2314	2834	4522	7929
3	1183	2608	5248	5442	7725	12165
4	358	1278	2221	2905	4726	7593

These tables do not immediately indicate an obvious relationship between math scores and demographic, but visualization of the relationships between the levels in the scatterplot matrix on the next page shows a clearer delineation between students with and without disabilities.

A Hotelling’s two sample T^2-test was performed to evaluate the hypothesis that the mean vectors of the two populations are not equivalent. The resulting test indicated that the probability that the data for the two groups came from the same population is less than $2.2*10^{-16}$, or very small.

Scatterplot Matrix of Number of Students which scored in each Level colored by Demographic

Red dots are students without Disabilities **Black dots** are students with disabilities



The clustering analysis algorithms were implemented and the resulting clusterings are described below. SWD indicates students with disabilities and SWOD indicates students without disabilities.

Linkage Technique	# Classified SWD	# Classified SWOD	# Classified Correctly	# Classified Incorrectly	# Classified as SWD but SWOD	# Classified as SWOD but SWD
Complete	327	93	303	117	117	0
Average	294	126	336	84	84	0
Ward	320	100	310	110	110	0
Target	*210*	*210*	*420*	*0*	*0*	*0*

The resulting clusterings, in general, performed better at identifying the students without disabilities than those without. This is likely due to the greater variation (and range) in numbers for the students without disabilities in each of the levels. The scatterplots in the appendix demonstrate the selected clusterings for each linkage technique described above. As before, red dots indicate students without disabilities and black dots indicate students with disabilities.

Average linkage performed better than Complete and Ward linkage, but Complete and Ward linkage performed similarly. All 3 linkage techniques used clustered all students with disabilities together.

Conclusion

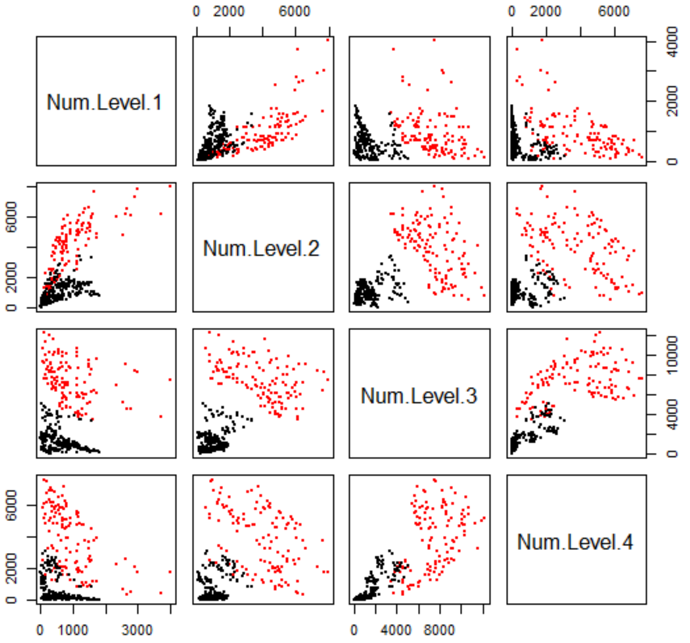
Clustering is a useful method for identifying populations with different characteristics, but is not without its limitations. The selection of linkage technique for hierarchical agglomerative clustering algorithms can result in different clusterings, and these clusterings should we evaluated against known information as applicable in the analysis.

REFERENCES

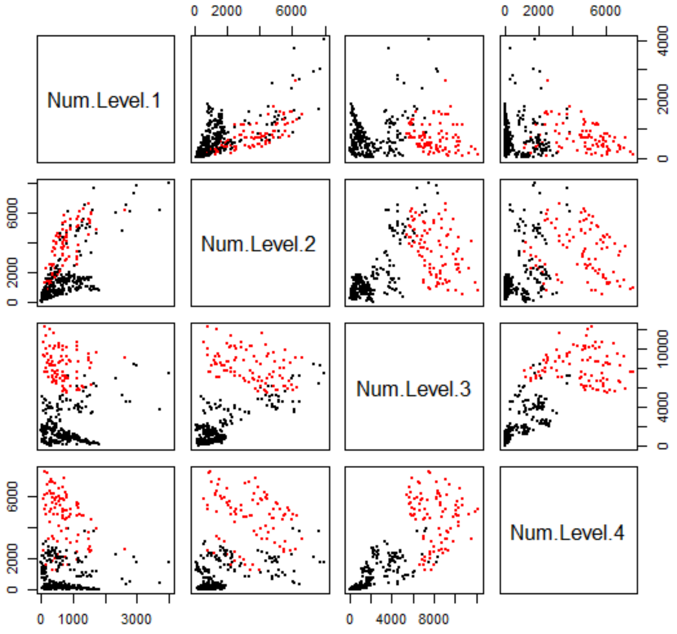
https://towardsdatascience.com/machine-learning-algorithms-part-12-hierarchical-agglomerative-clustering-example-in-python-1e18e0075019

APPENDIX – ADDITIONAL FIGURES

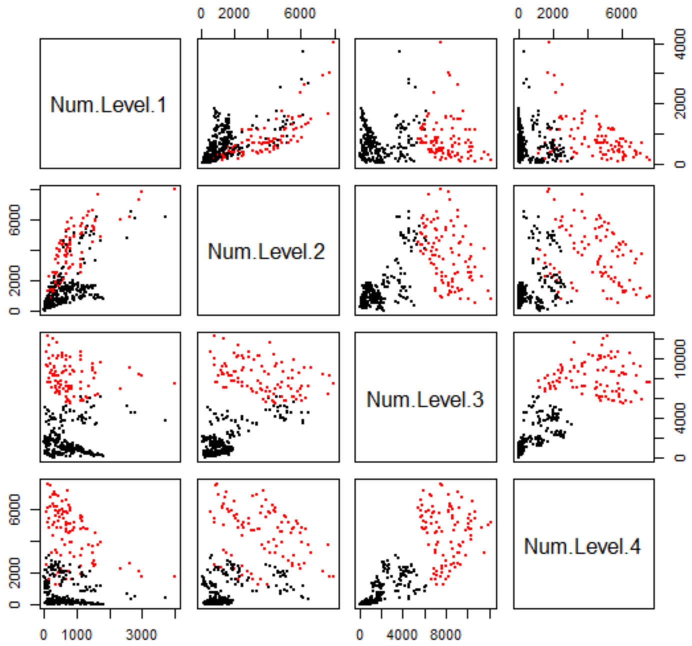
Scatterplot of Number of Students in each Level Colored by Demographic Clustering – Average Linkage



Scatterplot of Number of Students in each Level Colored by Demographic Clustering – Complete Linkage



Scatterplot of Number of Students in each Level Colored by Demographic Clustering – Ward Linkage


Num.Level.1
Num.Level.2
Num.Level.3
Num.Level.4