Chicago Metropolitan Area High School Data Exploration

Wenxuan Gu and Ryan Brim

Introduction

This data project is an exploration of Chicago Public Schools - Progress Report Cards (2011-2012). This original dataset is retrieved from University of Chicago repository which contains a total of 556 Chicago metropolitan elementary, middle, and high schools' information from 2011 – 2022 school year.

Data Cleaning

The step of cleaning our data presented a few challenges. Because it is a school data set, there were a lot of variables that wouldn't help an analysis, like phone number, state, and city, since all the schools are in the Chicago, IL area and the phone numbers to the schools are not relevant to analysis. Other variables had over 25% NDAs, meaning there was a lot of missing data, and were removed. They include Family Involvement Icon, Family Involvement Score, Leaders Score, Leaders Icon, Teachers Icon, Teachers Score, Pk-2 Math %, Students Taking Algebra %, Students Passing Algebra %, 9th Grade EXPLORE (2009), 9th Grade EXPLORE (2010), 10th Grade PLAN (2009), 10th Grade PLAN (2010), Net Change EXPLORE and PLAN, 11th Grade Average ACT (2011), Net Change PLAN and ACT, College Eligibility %, Graduation Rate %, College Enrollment Rate %, and Freshman on Track Rate %. As you can tell from the names, many of these variables describe some sort of attribute of middle or high schoolers, which less than ⅕ of the data was comprised of. Because of this, all middle and high schools were also removed from the dataset. Things like zip code, ward, police district, etc. all describe location, yet we kept them in the dataset because they describe different identifiers for an area, and they aren't always describing the same area.

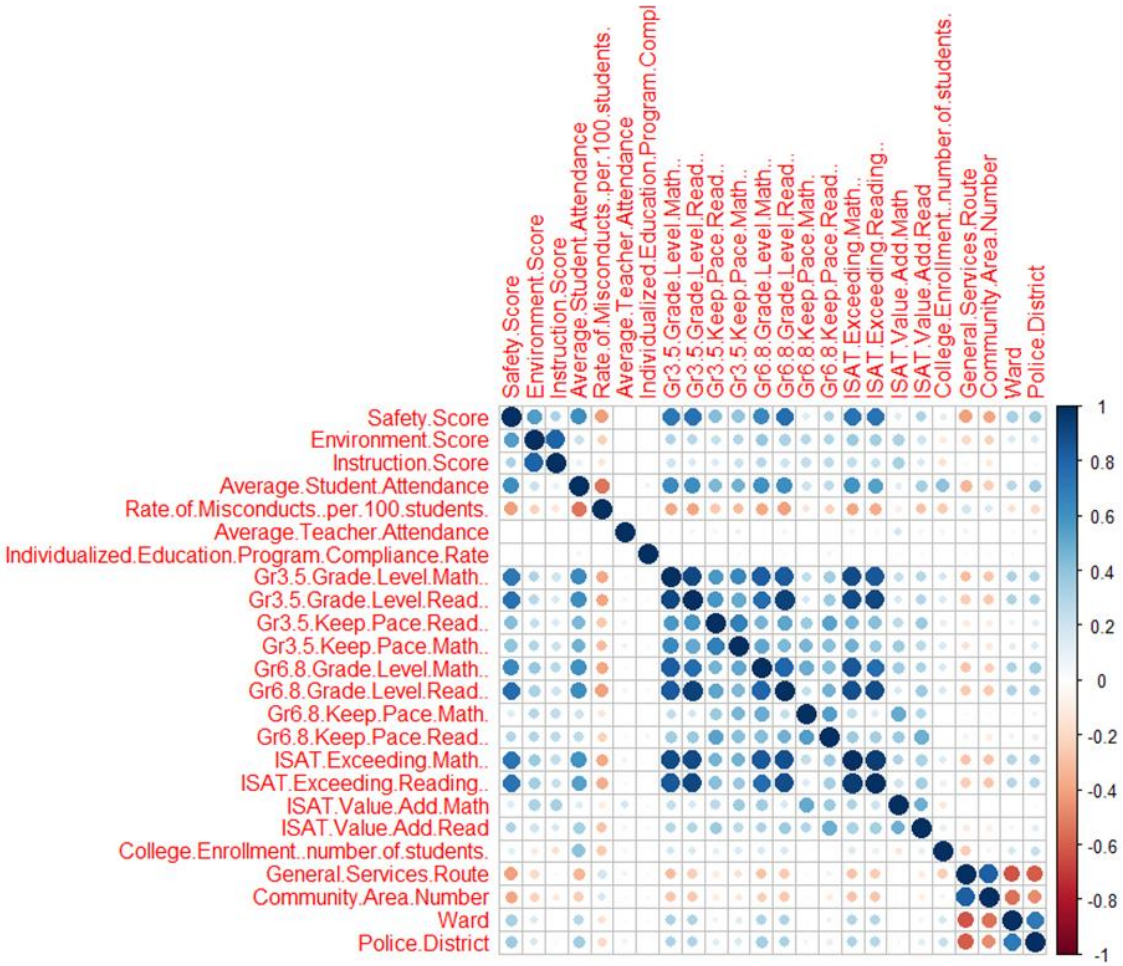
We went from 567 to 407 rows, and from 79 to 39 variables. Column values with over 25% NDAs were removed as stated previously. Column values with duplicate or close meanings were removed, such as X_COORDINATE, Y_COORDINATE, Latitude, Longitude, and Location (a tuple of latitude and longitude). Column values that are categorical, but contain useless information were removed. This 3-step cleaning process was performed to lower our NDA threshold from 30% to 0% to ensure that we minimize the noise in the data as much as possible. A small number of NDAs should not affect our analysis on variables, but we were left with 0 instances of NDAs once variables or rows with over 25% of NDAs were removed. Based on the 19-variable categorical dataset, it presents three different classes of categorical variables which are: categorical, binary, and ordinal. We have a total of 4 binary variables, 10 ordinal variables, and the rest are categorical. By splitting the cleaned dataset into two parts, the numeric and the categorical data sets, we were able to explore potential overfitting, multicollinearity, and outlier issues on all numeric variables and classes of categorical variables individually.

After cleaning, we moved on to more exploration of the data. We began by looking at potential outliers in this data by using the libraries "outliers" and "EnvStats" with Rosner's Test, and we were able to spot less than two potential outliers for each variable. No observations appeared more than twice for any variable, so it is not likely there is much influence on the data by any singular data point.

OLS Regression

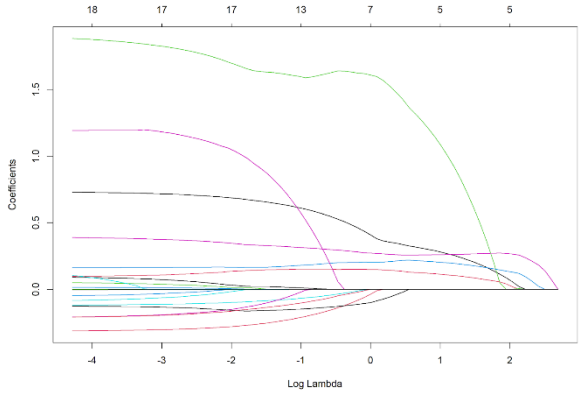
We ran an OLS model on just the numerical variables with Safety Score as the response variable. This was chosen by looking at the Correlation matrix. It was highly correlated with many variables and gave good results. There were several obvious signs of overfitting. The R^2 value was 79.11%, which is good, but there were beta coefficient values in an unexpected (large) range. Also, the Training RMSE was 9.21 and the Testing RMSE was 10.24, and a much larger Testing than Training RMSE is another indication of overfitting. It likely needs some sort of regularized regression.

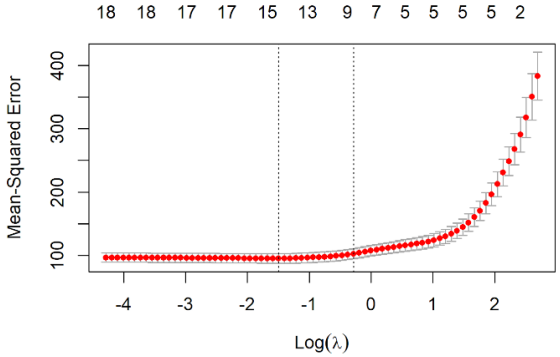
Finally, we looked into the relationships between the variables. A correlation heatmap of the numeric variables was created to look for groupings of variables that had high correlation. It can be seen that there are a couple groups along the diagonal of the heatmap. These are variables strongly correlated with one another. PCA and/or PFA would be good to investigate these relationships and trends in the variables.



Regularized Regression

As there was some indication of overfitting (unexpected beta coefficients and a higher Test than Training RMSE), several regularized regression models were performed. With ridge regression, there was a much better RMSE ratio than before. The new Test RMSE is now 9.40, making a ratio of 1.02, compared to the ratio of 1.11 with OLS. With lasso, we found even better results. The new Test RMSE is 9.23, making a ratio of 1.00. The lasso regression used a lambda of 0.7647 and an R^2 value of 74.62%, only losing 2.64% of the variance. It removed 8 variables: Rate of Misconducts per 100 students, Average Teacher Attendance, Gr 3-5 Keep Pace Read, Gr 3-5 Keep Pace Math, Gr 6-8 Grade Level Math, ISAT Exceeding Math, ISAT Value Add Read, and College Enrollment number of students.



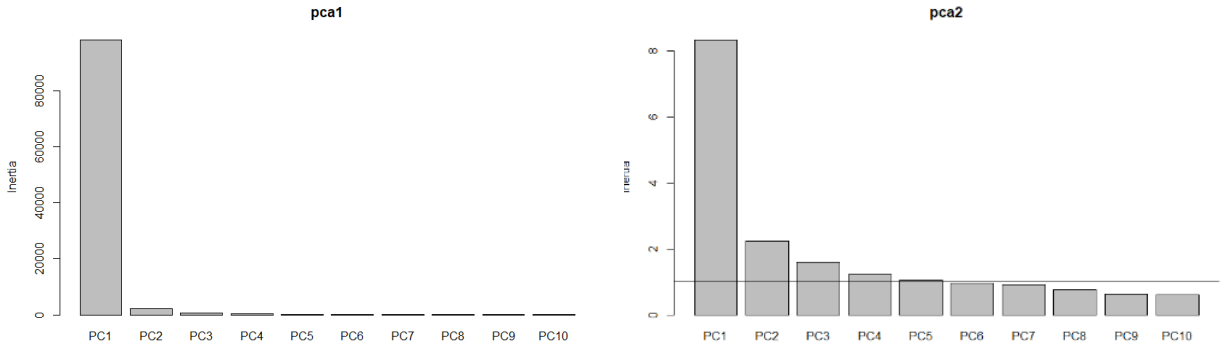


PCA & PFA Analysis

For our PCA analysis, we start it off by checking the scaling issues of our numeric dataset. By using "prcomp" to generate our loadings, it is obvious to see that the unscaled version appears to have scaling issues since the PC1 has a variation of 96.01%, dominating all other principal components. In comparison, in the scaled version, the first 6 PCs present a cumulative variation of 72.09%. The second, scaled version is the one we expect to use for future analyses.

To visualize and determine the how many PCs we expect to use, two versions of scree plots are presented as below. As we can see in the unscaled version, PC1 totally dominates the overall variation, and the rest seems like noise instead of useful information. The situation turns out to be much better by simply scaling the dataset as the right figure presents. By drawing the ab line, a total of 5 principal components are above the 1.0 variance line and are expected to be kept as PCs. The overall scaling issue testing turns out to be reasonable and better than we expected since there exists 5 PCs out of a total of 24 variables. By running scaled version of PCA, we see that it does a good job on dimensionality

reduction and the following analysis would be using PFA to reveal the latent factors behind our PCs. With principal function and varimax rotation, we can generate the loading outputs as below:



A standard metric of setting our cutoff as 0.4 and scores equal to true is applied to our PFA analysis. The first loading shows 5 factors/components and each of the RCs can be labeled with a certain trend. By looking at RC1, it mainly includes a series of variables which are the "course programs." Variables such as "Gr3.5.Grade.Level.Math.." are specific course programs that elementary schools students take. "Gr3.5.Grade.Level.Read.." and "ISAT.Exceeding.Reading.." have the highest contribution to RC1. By combining these courses with variable "Safety.Score", "Average.Student.Attendence" and misconducts rate, we can expect RC1 to be labeled as "blue ribbon" elementary schools since this category not only presents safe locations, decent student attendance, and negative misconducts rate but also highly focuses on reading and math programs.

```
Loadings:
                                                    RC1    RC2    RC3    RC4    RC5
Safety.Score                                         0.778
Environment.Score                                           0.882
Instruction.Score                                           0.885
Average.Student.Attendance                           0.591                0.561
Rate.of.Misconducts..per.100.students.                                   -0.562
Average.Teacher.Attendance                                               -0.472  0.436
Individualized.Education.Program.Compliance.Rate                                 0.821
Gr3.5.Grade.Level.Math..                             0.900
Gr3.5.Grade.Level.Read..                             0.943
Gr3.5.Keep.Pace.Read..                               0.448  0.602
Gr3.5.Keep.Pace.Math..                               0.412  0.602
Gr6.8.Grade.Level.Math..                             0.773
Gr6.8.Grade.Level.Read..                             0.909
Gr6.8.Keep.Pace.Math.                                       0.792
Gr6.8.Keep.Pace.Read..                                      0.740
ISAT.Exceeding.Math..                                0.927
ISAT.Exceeding.Reading..                             0.937
ISAT.Value.Add.Math                                         0.690
ISAT.Value.Add.Read                                         0.589
College.Enrollment..number.of.students.                                  0.724

                 RC1   RC2   RC3   RC4   RC5
SS loadings    6.515 3.230 1.975 1.626 1.071
Proportion Var 0.326 0.162 0.099 0.081 0.054
Cumulative Var 0.326 0.487 0.586 0.667 0.721
```

RC2 presents an interesting observation which forms pairs of course programs. It matches three different levels of math & read programs together as three pairs and it can be expected as a trend of ""important pathway programs."

RC3 can be interpreted as “fundamental score for high schools” and the example would be schools in Lincoln Park area could be considered as the best neighborhood in Chicago and it tend to have relatively high instruction scores and environmental scores.

RC4 could be labeled as “participation status” because it combines both teachers and students attendance plus the rate of misconducts. RC5 here is bit different since it is a combination of teachers’ attendance and compliance rate. It might be labeled as “instruction strategy.” In order to have a holistic view of all our RCs, the proportion of each variables shows the RC4 and RC5 generate a total of 13.5% variance which are not as high as RC1 and RC2. Thus, we have a sense to know that RC4 and RC5 can be considered as additional information for an overall review of this data. RC1 and RC2 are the main trends we derived from this dataset. They could be used as criteria for evaluating Chicago high schools.

By comparing loadings with 8 RCs, we see that the additional three RCs do not generate as much information as the original five RCs in our previous loading outputs. RC6 and RC5 are basically just single variables. RC7 are a bit redundant since it is technically a simplified version of RC4 and RC8.

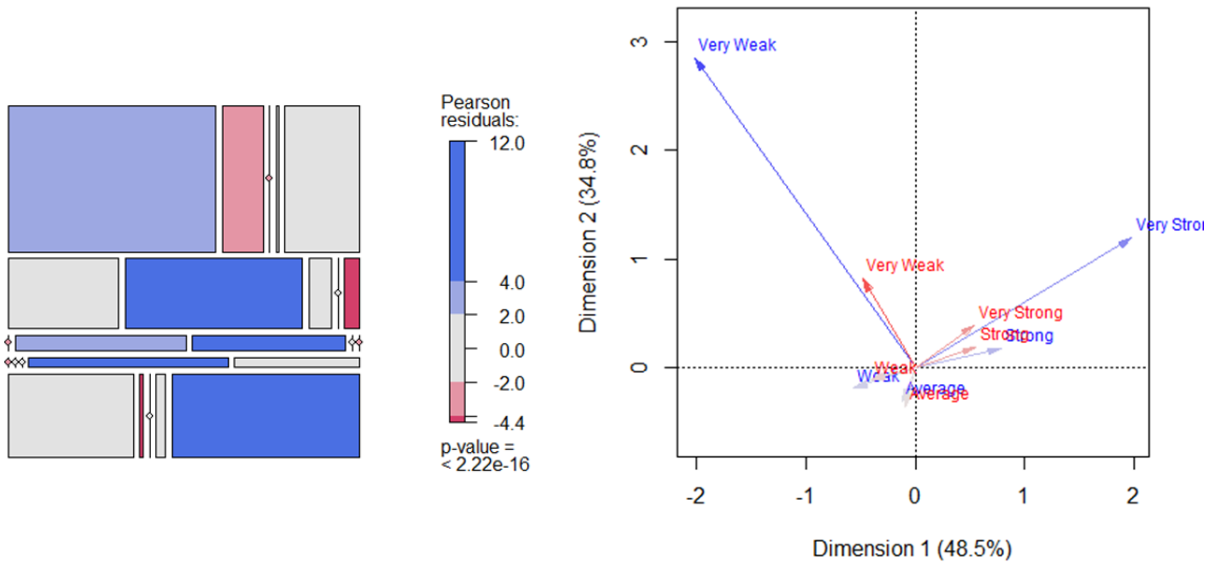
```
Loadings:
                                                      RC1    RC3    RC8    RC2    RC4    RC7    RC6    RC5
Safety.Score                                           0.752
Environment.Score                                             0.909
Instruction.Score                                             0.900
Average.Student.Attendance                             0.557                         0.592
Rate.of.Misconducts..per.100.students.                                              -0.611
Average.Teacher.Attendance                                                                        0.974
Individualized.Education.Program.Compliance.Rate                                                         0.978
Gr3.5.Grade.Level.Math..                               0.883
Gr3.5.Grade.Level.Read..                               0.922
Gr3.5.Keep.Pace.Read..                                        0.824
Gr3.5.Keep.Pace.Math..                                        0.735
Gr6.8.Grade.Level.Math..                               0.803         0.413
Gr6.8.Grade.Level.Read..                               0.899
Gr6.8.Keep.Pace.Math.                                         0.839
Gr6.8.Keep.Pace.Read..                                        0.478                 0.510
ISAT.Exceeding.Math..                                  0.938
ISAT.Exceeding.Reading..                               0.949
ISAT.Value.Add.Math                                           0.700
ISAT.Value.Add.Read                                                                 0.864
College.Enrollment..number.of.students.                                      0.837

                 RC1   RC3   RC8   RC2   RC4   RC7   RC6   RC5
SS loadings    6.261 2.032 1.862 1.760 1.622 1.424 1.043 1.037
Proportion Var 0.313 0.102 0.093 0.088 0.081 0.071 0.052 0.052
Cumulative Var 0.313 0.415 0.508 0.596 0.677 0.748 0.800 0.852
```

Correspondence Analysis

Besides the PCA and PFA analyses, a correspondence analysis is also conducted in this milestone. By paring up our categorical variable instruction icon and environment icon, a contingency table is generated as the following:

	Average	Strong	Very Strong	Very Weak	Weak
Average	117	23	0	1	42
Strong	30	48	6	0	4
Very Strong	0	10	9	0	0
Very Weak	0	0	0	8	5
Weak	40	1	0	3	60

The mosaic plot shows the observations displayed as proportions. The darker blue or red a box is, the higher or lower observed proportion than expected the occurrence is, respectively. For example, the bottom right box is very dark blue, meaning that instance of a “weak” instruction icon and a “weak” environment icon are higher than expected. In fact, each instruction icon mapped to each environment icon has a higher-than-expected proportion, though the “average”-to-“average” instance is not as high as the other instances. This means that there is a strong positive correlation between instruction icon and environment icon. In general, as a school is rated higher (or lower) for instruction capabilities, it is highly likely it will also be rated higher (or lower) for the environment of the school. For the two dark red boxes, there is a lower-than-expected number of schools with a “strong” environment icon and a “weak” instruction icon, and vice versa, which makes sense.




```
Principal inertias (eigenvalues):
           1        2        3        4
Value      0.535063 0.383918 0.139015 0.044891
Percentage 48.51%   34.81%   12.6%    4.07%
```

Above are the principal inertias, or variances for each eigenvalue. The graph on the right only shows the first two eigenvectors to display the results of the contingency table in a more readable form. It shows the direction and magnitude of the pairs of each environment and instruction icon. The angle between each instance is very small/acute. Each arrow labeled with the same name are pointed in nearly the same direction, meaning the correlation between each icon for an instance is high. The “strong” and “very strong” are pointed to the positive side of the first dimension, while “weak” and “very weak” are pointed in the negative direction. Since the first dimension makes up the most variance, it makes sense that it has the largest span (-2:2). The second dimension is still large, and therefore it also has a large span (-0.5:3).

LDA Exploration

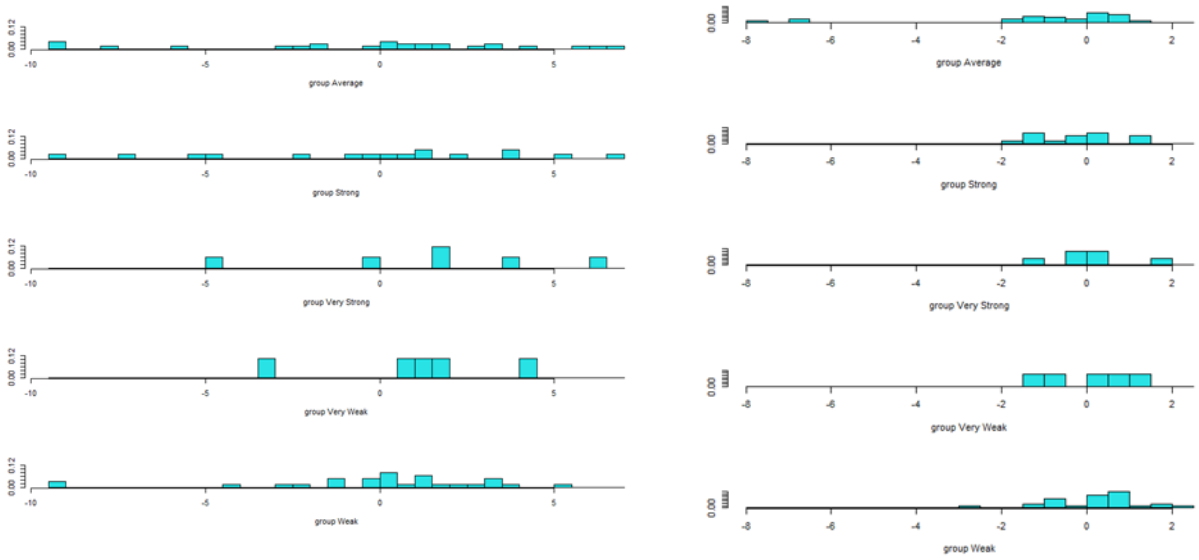
As an expletory, a try-out of LDA analysis is implemented in this milestone. By using categorical variable “Safety Icon” as our response and the rest as exploratory variables, LDA is generated to observe if safety icon separates classes well. By setting our training group with 80% observations, we get 4 LDs as our output based on training set since there are five levels of “safety icon” which are weak, very weak, strong, very strong, and average. The loadings are not included since it is not useful for us to make valuable interpretations.

As a summary of LDs, we can tell that the LD1 is explaining 95% variations overall which is not ideal since the rest LDs are becoming noise. This is an indicator telling us that our response might not be a good categorical variable that classifies groups well. Ideally, we would expect to have traces behave as 0.5, 0.3, 0.2, 0.1 instead of one dominating all others in this case.

```
Proportion of trace:
   LD1    LD2    LD3    LD4
0.9506 0.0360 0.0071 0.0063
```

By generating two LDA histograms, we are able visualize how well each group separates from each other. With dimensionality equal to 1, it is very hard for us to tell the separations since they are all behaving as histograms spreading evenly from range -10 to 5. Ideally, we would expect at least two groups of them are on the opposite side of each other which indicates that such groups are well separated by our LDs. A similar result with two-dimensional space presents something similar. They are all grouping on the right side on the scale, and it is not helpful for us to pick out certain distinct groups overall. This explains that using “safety icon” as our response might not be helpful for grouping.

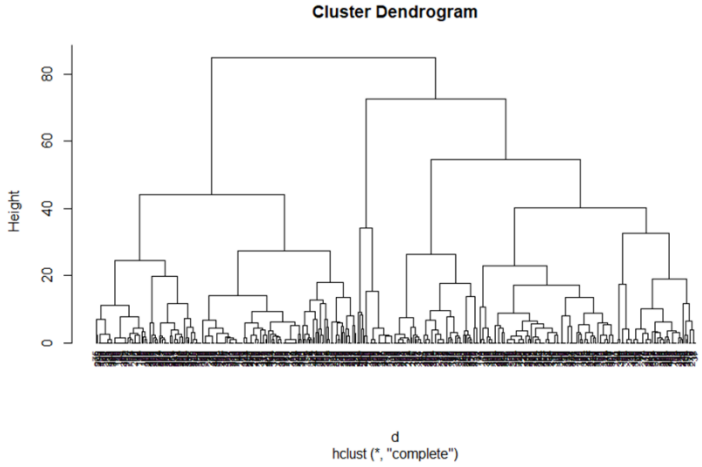
In this case, even if LDA is not working well for our dataset, it doesn’t mean that this try-out phase is useless. This might be a foresee for us to apply a different technique like clustering on this dataset with other categorical variables.



Clustering

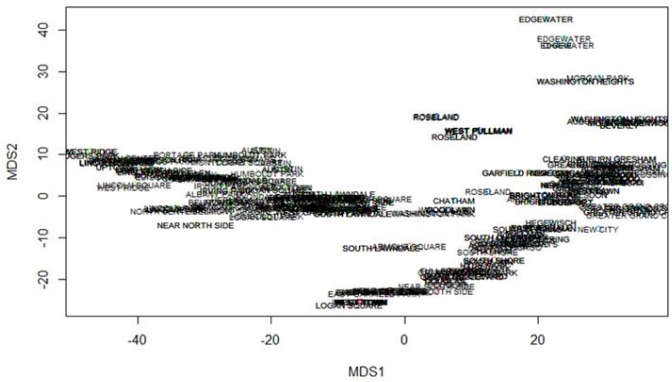
As another route of exploration, a clustering analysis is performed. These results are based on these four variables to explore if there exists a pattern and meaning behind areas.

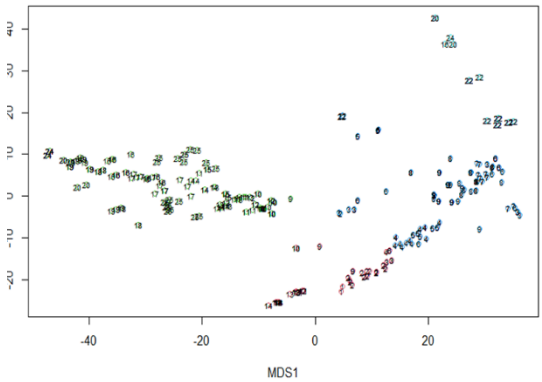
By using Euclidian distance and package “vegan”, we can calculate out the stress which is 0.0278. By applying our calculated Euclidian distance, we can generate a dendrogram:



This dendrogram does provide a preliminary view of our clustering as we can see that they were separated to 4 branches at a height of 50 (three on the right and one on the left). Therefore, we might choose the 4 clusters for our k-means clustering.

By setting our new variable “clustCut”, we form a new column that has 4 different groups, and they are used in our data frame for clustering. The clustering results are generated as below: the two graphs show the similar data, the difference is that one includes police districts and the other includes specific community area names.



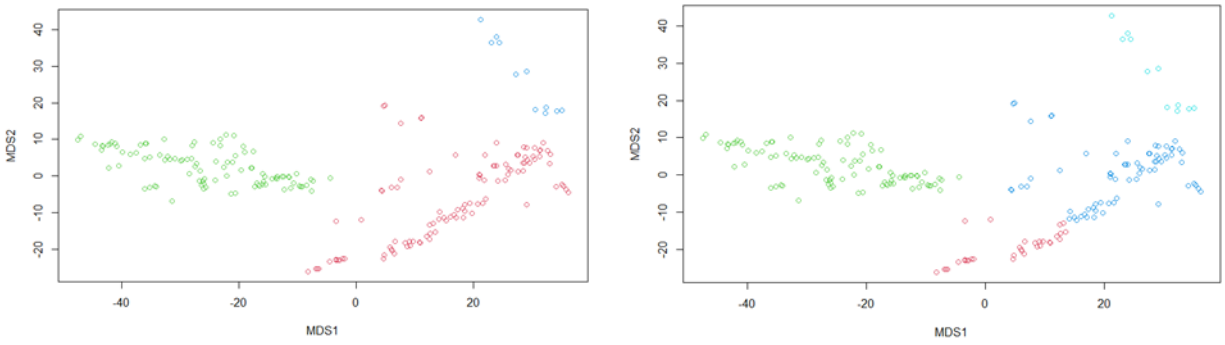


As we relate to the Chicago district and area map shown below, we see that areas in green are considered as safe neighborhoods such as Lake View, Old Town, Lincoln Park, Belmont, Lincoln Square. The sky-blue dots represent relatively remote areas such as Edge Water, Logan Park, and Washington Heights. The dark blue dots represent the danger zone. The most typical examples can be found such as Chatham, Garfield Park, Greater Grand Crossing, Auburn Gresham, Englewood. According to the website of the most dangerous neighborhoods in Chicago area, areas mentioned in the article are also all included in the dark blue zone and red zone.

The red area represent risky zone that are not as dangerous as areas in dark blue but might also considered as neighborhoods to be avoided such as Hyde Park, Fuller Park, South Shore, Near South Side. Dots that are at the boundary of the danger-zone might also considered as dangerous places such as South Deering and Riverdale. By also combining the police districts and neighborhood areas, it is even more obvious to see that that police district such as 2,3,5,6,7,8,9 are all centered at the dark blue and red dots which represents police districts in the south side of Chicago.

The rest green and blue dots have corresponding police districts with two digits, and they represent the safer police districts in comparison. Based on the clustering outputs, high schools can be grouped with four levels of zones which are: “dangerous”, “risky”, “safe” and “remote.”

By setting our cluster with k equal to 3, we generate 3 distinct groups. It is more obvious for us to see that dark bule dots and red dots can also be grouped together as red dots which indicates that it can be well separated from the rest two groups and considered as the new “danger-zone.” This new “danger-zone” represents all risky and dangerous areas and makes our total clustering more succinct and interpretable.



Conclusion

Our models can be used for many applications. It can be used by parents with schoolchildren trying to make an informed decision where to move so their kids can be in a good school. They can use our clustering analysis to find schools in neighborhoods that are both safe and have good schools. They can also use our correspondence analysis and regularized regression models to determine the best school in each neighborhood based on a number of variables.

On another side, the City of Chicago could potentially use our models to determine which schools deserve more funding. Schools with poorer performance could benefit from the extra resources from an increased budget.

Reference

[1] https://usaestaonline.com/most-dangerous-neighborhoods-in-chicago

[2] https://news.wttw.com/sites/default/files/Chicago%20Police%20Districts.pdf