Exploratory Analysis of Educational Proficiency Data using K-Means and Hierarchical Clustering

Andres Cambronero

Before conducting any sort of inferential statistical test, researchers need to perform extensive exploratory analysis to understand patterns in their data. In social science, histograms, bivariate plots and summary tables are often used to examine differences in distributions, identify outliers and determine correlations. While such an approach is perfectly valid and necessary, clustering methods are powerful tools that can enhance researchers exploratory analysis. unfortunately, such methods are rarely used in social science. Incorporating clustering methods into the initial stages of data exploration can lead to greater understanding of the data at hand and ultimately enrich the final analysis conducted.

Clustering methods refer to a broad range of unsupervised learning approaches intended to uncover intrinsic patterns in data. Using different algorithms, these methods attempt to find relatively homogenous subgroups among observations in the data. Two popular clustering methods are K-Means and Hierarchical Agglomerative Clustering. K-Means splits the observations in the data into K non-overlapping groups, where K is a pre-specified number of clusters. On the other hand, hierarchical clustering creates subgroups in the data, but does not require a pre-specified number of clusters. In different ways, both methods strive to establish subgroups of observations that have high intra-cluster similarity and low inter-cluster similarity. Such grouping can help researchers better understand their data.

This report uses both K-means and hierarchical clustering to perform exploratory analysis of data about school proficiency in New York. The report conducts clustering using different number of clusters and assesses the clustering quality of the results. The clustering analysis is performed on the dataset's predictors only and excludes the dataset's response level, proficiency level. Once the analysis is conducted, the response label is re-assigned to observations to determine if the subgroups formed by the clustering methods have some overlap with the response variable. Additionally, once the clusters have been determined summary statistics of each group are presented to understand the clusters' characteristics.

Methodology

This section describes the two clustering algorithms used in the report.

K-Means: This algorithm randomly assigns each observation to one of K clusters that the researcher has previously specified. For each of the clusters, it computes the cluster centroid, which is defined as the vector of p feature means for the observations in the cluster. Then, it re-assigns each observation to the cluster with the closest centroid. This process is repeated until the cluster assignments stabilize.

K-means finds a local optimum that minimizes the sum-of-squared deviations from its centroid within each cluster. The squared Euclidean distance is often used as a measure of distance. Because the algorithm only provides a local optimum, which varies depending on the initial assignment of observations to clusters, the algorithm is often repeated multiple times and the best results are reported. In this report, the algorithm was repeated 10 times and it was allowed to iterate a maximum of 1000 times to converge.

Hierarchical Clustering: Unlike K-means, hierarchical clustering does not require the user to pre-specify the number of clusters to create. Instead, the algorithm starts with each observation as its own cluster. Then, it groups the two clusters that are least dissimilar to each other and proceeds iteratively until all observations have been grouped. The squared Euclidean distance between observations is often used as a dissimilarity measure. The user can then decide how many clusters are reflective of the data's structure.

Before grouping observations into clusters, the algorithm requires the user to determine a notion of linkage between observations, which is a measure of dissimilarity between groups of observations. The type of linkage used often results in very different cluster structure and quality. The linkages used in this report are the following:

- Single Linkage: Compute all pairwise distances between clusters, measuring the distance between the clusters from the observations that is closest to the other cluster. Group the clusters that are least dissimilar to each other.

- Complete Linkage: Compute all pairwise distances between clusters, measuring the distance between clusters from the observation that is furthest from the other cluster. Group the clusters that are least dissimilar to each other.
- Ward Linkage: Compute the increase in total sum of squares around the mean for the merge of any two clusters. Group clusters whose merge would lead to the smallest increase in total sum of squares.

In order to visually examine the clustering results, the data points had to be projected from their original multidimensional space to lower 2 dimensional space. To do so, we use classical multidimensional scaling. Classical MDS finds a matrix $Z = [z_1, ..., z_n]^T \in R^{n \times p}$ that minimizes the function:

$$S(Z) = ||ZZ^T - XX^T||_F$$

where $X = [x_1, ..., x_n]^T \in R^{n \times q}$ for our original data and $p < q$. Minimizing this function will provide a low rank approximation to the matrix $G = XX^T$.

Data Processing

The data used in this report contains information about the proficiency of elementary and middle schools in New York. The dataset contained 666 observations. Out of these observations, 56 cases were of class "Proficient" and the rest were "Not Proficient" in the response variable "Proficiency level". As mentioned earlier, the dataset's response variable was excluded from the clustering analysis and only used after the analysis to determine if the clusters created corresponded to the proficiency levels in the data.

The following procedure was followed to merge and clean the dataset into the final stage.

First, the outcome variable "PROF_LEVEL" was created following guidelines of the Every Student Succeeds Act (ESSA) and New York's standard of proficiency. Under ESSA, a school is in need of improvement if English language learners, students receiving special education, racial minorities, or students in poverty consistently underperform. Using the Department of Education's 2015-16 3-8 Assessment Database, each subpopulation of interest within a school received a score of 1 (proficient) or -1 (not proficient) if it's assessment mean achieved the state's standards. Then, each school received a weighted score based on the school's demographics. If the final score was negative, the school was classified as "Not Proficient" and "Proficient" otherwise.

To have a completed dataset, values had to be artificially generated to supplement the information available on the New York State's Department of Education databases, the National Center for Educational Statistics and ACS. These databases contained the following incomplete variables: Suspensions, Average Class Size, Number of Teachers, Percent of Not highly Qualified Teachers, Mean Teacher Score, Enrollment, Expenditure per student, Violent crimes and Property crimes. For these observations, random values were generated from a bounded normal distribution with the mean and variance of the observed values of a school's district. Any remaining missing observations were completed drawing numbers from a normal distribution with the mean and variance of the values observed in the column. With this process, each observation had complete information for each column.

The final dataset contained 26 variables. A full list of variables in the final dataset is provided in the index.

Analysis

Before clustering the observations using the methods proposed, Figure 1 shows a 2-D representation of the data. The plot used the Euclidean distance between observations. Without labels and other identifiers, the plot suggests that there are at least two possible clusters in the data. One cluster is tightly gathered on the left hand side of the plot, while the second cluster is loosely scatted across the right hand side of the figure. Without identifiers, V1 and V2 cannot be associated with any of the variables in the data.

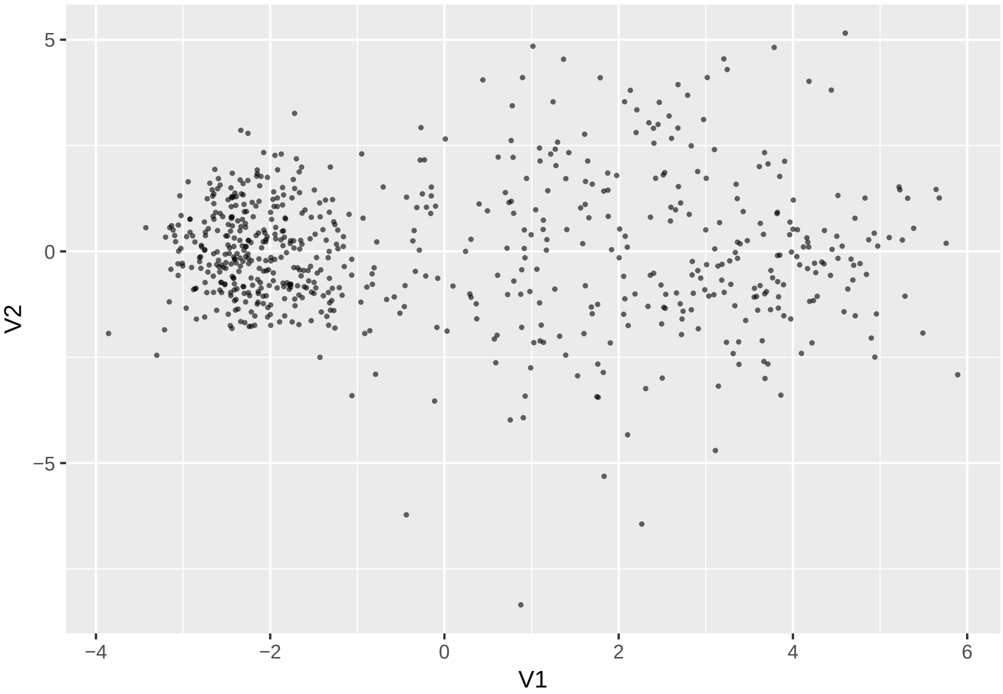

Figure 1: 2-D Representation of Data

Hierarchical Clustering

Below are the results of using Hierarchical clustering with single, complete and ward linkage.

One measure of clustering structure is the agglomerative coefficient. According to Holland, this coefficient measures the dissimilarity of an object to the first cluster it joins, divided by the dissimilarity of the final merger in the cluster analysis, averaged across all samples. In our case, coefficients close to 1 indicate strong clustering structure. The agglomerative coefficients obtained for the linkages single, complete and ward respectively are 0.75, 0.87 and 0.96. This result indicates that a stronger clustering structure is achieved using Ward linkage compared to single and complete.

Figure 2 shows the full dendrogram obtained using the three linkage types. As expected, the initial clusters formed were relatively close as suggested by the low height at which the observations were grouped together. These clusters formed at low height are likely to have observations that are more similar to each other than clusters formed at higher heights.

The single linkage clustering created a lopsided dendrogram. The distance between the clusters at each merge seems small as indicated by the short difference in heights at each merge. Compared to the complete and ward distances, the single linkage formed a single cluster at a much lower height.

The complete linkage clustering also created a lopsided dendrogram, but not as imbalanced as the single linkage clustering. Unlike single linkage where height difference between clusters merges were small throughout the entire process, the complete linkage plot suggests that, as fewer clusters remained, the clusters being merged were quite dissimilar as indicated by the large difference in height difference at each merge towards the end of the process.

Unlike the other two linkage types, the ward linkage created a balanced dendrogram. This linkage created a single cluster at a much higher height than the other two types shown in Figure 2. This results suggest that the clusters merged at the end of the process were more dissimilar to each other than they were for the other two linkage types.

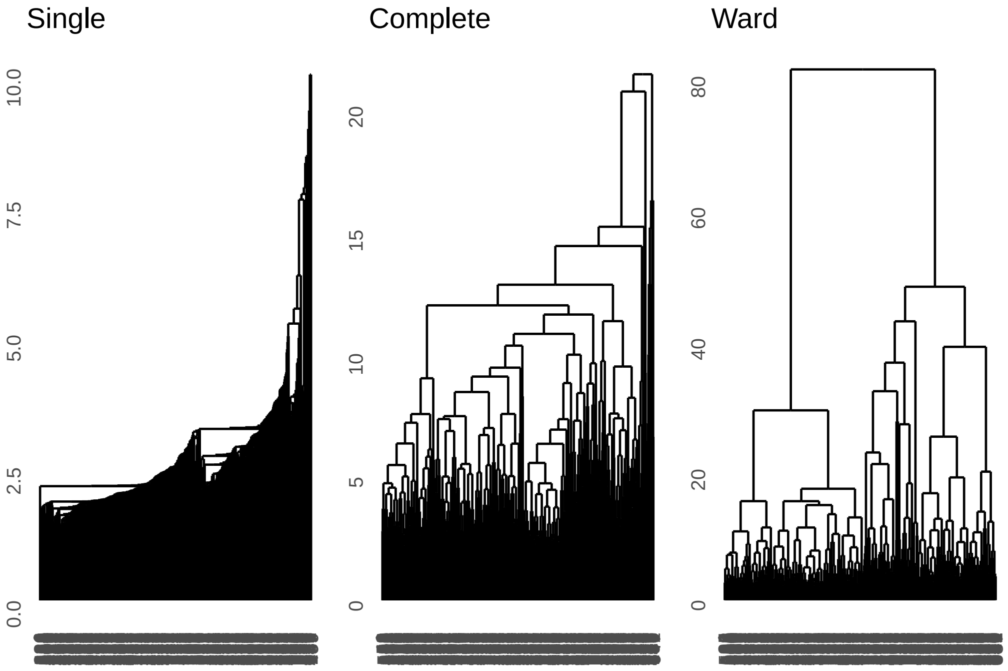

Figure 2: Full Dendrograms for Single, Complete and Ward Linkages

To compare the clustering quality for these three distances, we can analyze the silhouette plots for different number of clusters. The silhouette coefficient compares how close an observation is to other objects in its own cluster with how close it is to observations in other clusters. A silhouette coefficient close to 1 indicates that most observations within a cluster are quite close to each other relative to their position to observations in other clusters.

Figure 3 shows the silhouette plot for the single linkage clustering with two clusters. Cluster 1 contains 665 observations and Cluster 2 only contains 1 observation. Despite this imbalance, Cluster 1 has an silhouette coefficient of 0.48, which indicates that the observations within the cluster are still relatively similar to one another. Using single linkage, the results do not provide much insight into the data's structure.

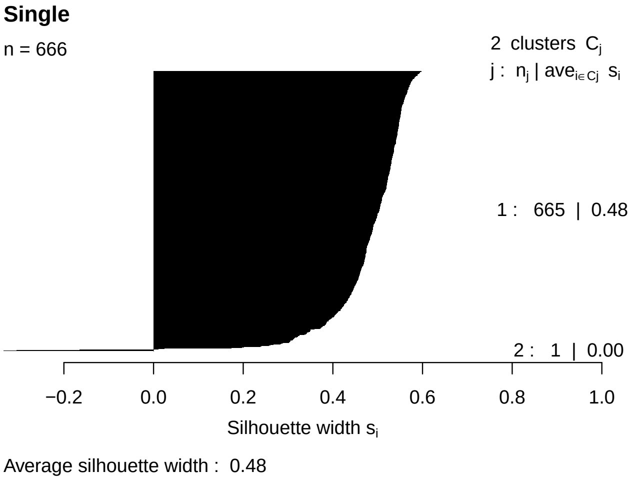

Figure 3: Two Cluster Silhouette Coeficient Plot for Single Linkage

Figure 4 shows the silhouette plot for the complete linkage clustering with two clusters. Cluster 1 contains

647 observations and Cluster 2 contains 19 observations. The silhouette coefficient is 0.36 within Cluster 1 and -0.02 within cluster 2. The average coefficient for this linkage type is 0.35. These results suggest that using complete linkage with two clusters does not improve the clustering quality over single linkage with two clusters. Again, these results do not provide much insight into the data's structure.

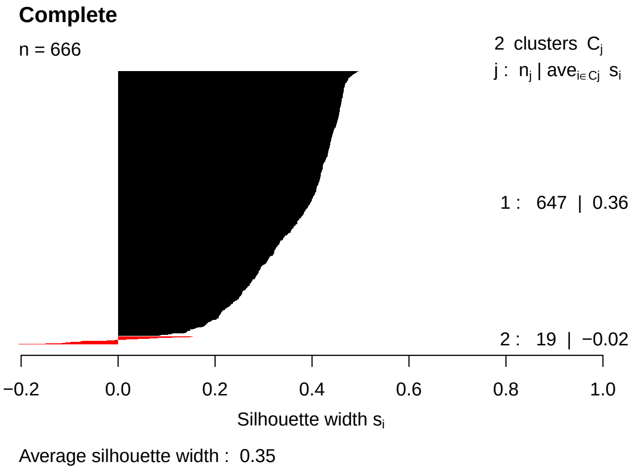

Figure 4: Two Cluster Silhouette Coeficient Plot for Complete Linkage

Figure 5 shows the silhouette plot for the Ward linkage clustering with two clusters. Ward managed to assign relatively similar number of observations into Cluster 1 (340 obs) and Cluster 2 (326 obs). While the silhouette coefficient for Cluster 2 is low, the coefficient is relatively high for Cluster 1. This result indicates that, while the observations in Cluster 2 might be quite dissimilar amongst themselves, the observations in Cluster 1 are similar to one another.

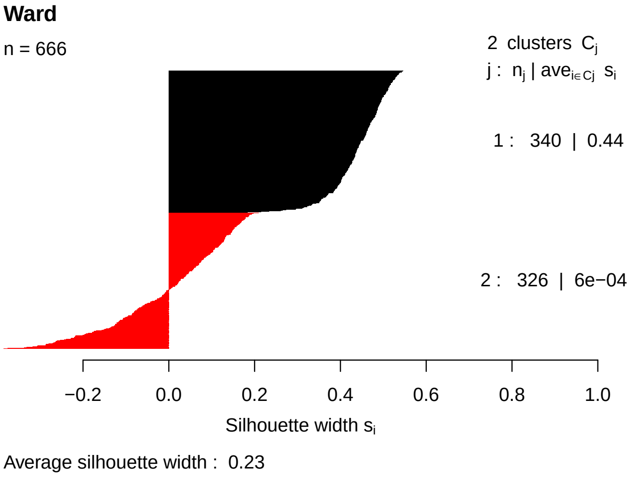

Figure 5: Two Cluster Silhouette Coeficient Plot for Ward Linkage

Since there is no objective way to determine the ideal number of clusters to use, it is important to examine the clustering quality achieved using a different number of clusters. In this case, the hierarchical clustering at 3 and 4 clusters did not suggest higher silhouette coefficients and thus did not indicate higher quality clustering. As a result, only the results for the two cluster hierarchical clustering are presented.

Figure 6 shows 2-D Projection of Hierarchical Clustering with 2 clusters using Ward Linkage. This method identified the two clusters that were highlighted at the beginning of the report. Cluster 1 is composed observations that are tightly located on the left hand corner of the plot, while Cluster 2 is composed of observations scattered across the right hand side of the plot. This scattered pattern clearly explains why Cluster 1 had a much higher silhouette coefficient than Cluster 2; the observations in Cluster 2 are further apart from each other and therefore less similar to one another.

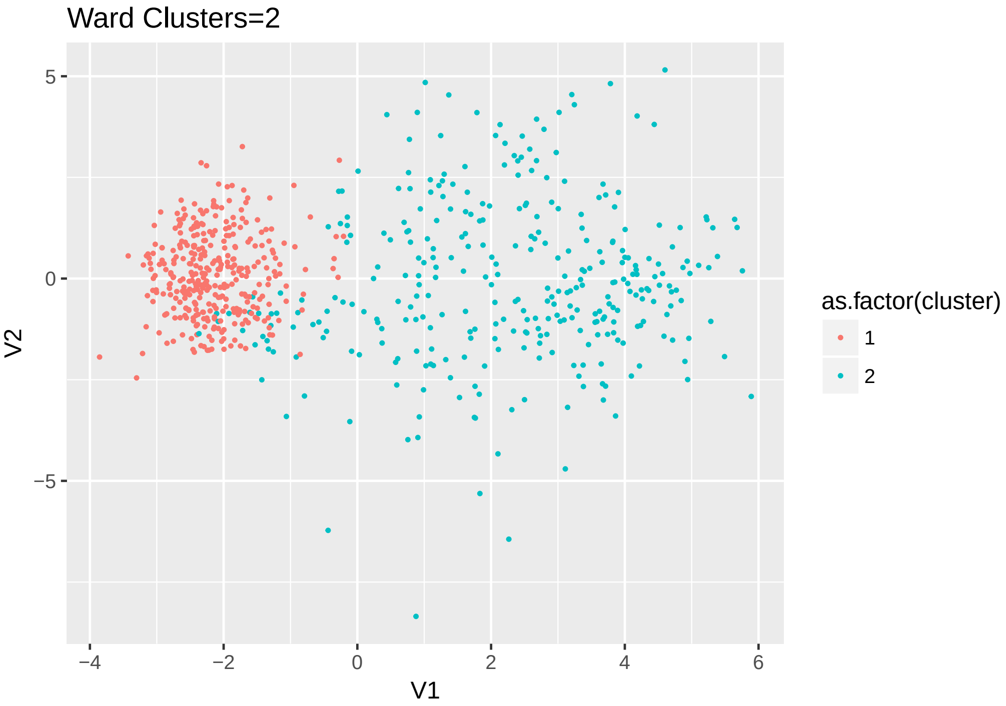

Figure 6: 2-D Projection of Hierarchical Clustering with 2 clusters using Ward Linkage

Table 1 shows the mean value for observations in Cluster 1 and Cluster 2 for certain variables that are worth mentioning. The table shows that the clusters are very different along some dimensions and not that different in others. First, the clusters do not vary along the response variable “Proficiency Level.” Of the observations in Cluster 1, about 9.4% are considered proficient, while about 7.3% of observations in Cluster 2 are of that class.

However, the clusters do seem to vary along economic and racial variables. The mean percentage of economically disadvantaged students in Cluster 2 is much higher than in Cluster 1, 76.75 compared to 32.75 rspectively. Similarly, the mean percentage of white students in schools in Cluster 1 is 81.27, while it is only 16.91 for schools in Cluster 2. Additionally, the mean percentage of Black, Hispanic and Asian students is substantially higher for Cluster 2 than for Cluster 1. As a result, Cluster 1 can be categorized as the cluster of schools composed of White, economically privileged students and Cluster 2 can be considered the cluster of minority economically struggling schools.

Table 1: Mean Percentage Values of Certain Variables for Cluster 1 and Cluster 2 Hierarhical Clustering

Variable_Name	Cluster_1	Cluster_2
Proficiecy	0.0941	0.0736
PER_ECDIS	32.7598	76.7546
PER_WHITE	81.2779	16.9121
PER_BLACK	3.2583	34.0358
PER_HISP	7.8431	37.0041
PER_ASIAN	5.1294	9.6779

K-Means

We can also cluster observations using k-means. In this section, we discuss the results of using clustering with k=2 and k=3.

Below are plots of the silhouette coefficients for k=2 and k=3. With **k=2** clusters, the algorithm groups a large number of observations in each of the two clusters. Cluster 1 has 392 observations and Cluster 2 has 274 observations. As with hierarchical clustering, the silhouette coefficient of Cluster 1 is substantially higher than that of Cluster 2, suggesting higher overall clustering quality. With k=2, the silhouette coefficient of the clusters created by K-Means is roughly equivalent to that produced using hierarchical clustering with Ward linkage and creating 2 clusters.

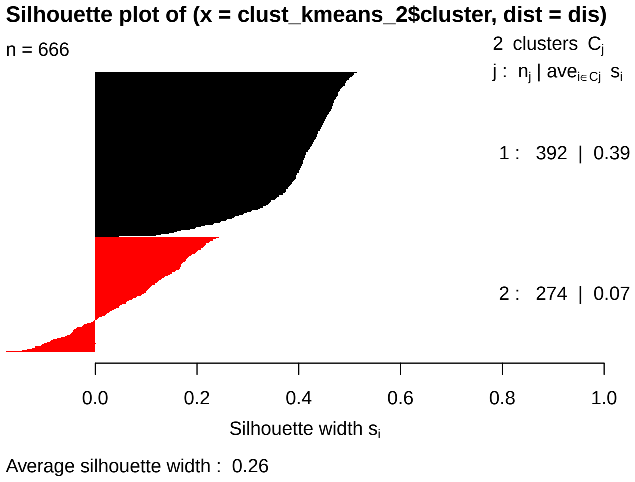

Figure 7: Silhouette Coeficient Plot K=2

Increasing the number of clusters to k=3 does not meaningfully increase the clustering quality. The algorithm still groups a large number of observations into each of the three clusters, 195, 98 and 373 into each of the three respective clusters. While the silhouette coefficient in Cluster 3 is higher than the silhouette coefficient of either cluster when k=2, the silhouette coefficient of Cluster 1 and Cluster 2 is quite low. The slight increase in the silhouette coefficient of one cluster does not merit the substantial decrease in silhouette coefficient of the other two clusters. As with hierarchical clustering, it seems that the best clustering quality is achieved with two clusters.

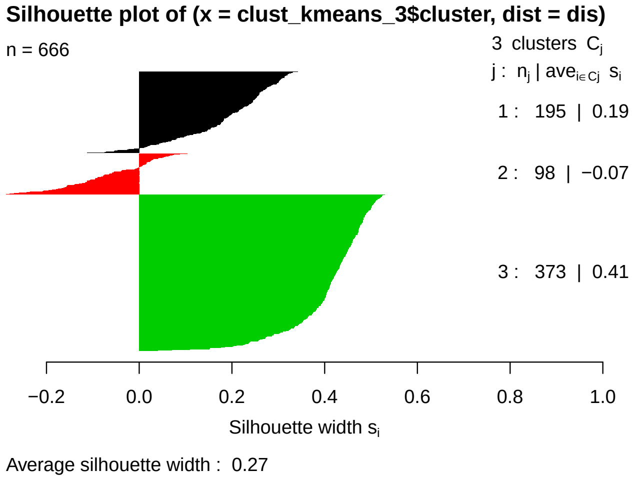

Figure 8: Silhouette Coeficient Plot K=3

The quality of clustering can be further investigated by assessing the total within-cluster sum of squares and total between sum of squares for k=2 and k=3. If clear and distinct clusters exist in the data, we would expect a small within-cluster sum of squares and large total between sum of squares. Table 2 shows the total within-cluster sum of squares and total between sum of squares for k=2 and k= 3. Based in this metric of clustering quality, K-means with K=3 seems to provide a better result than k=2. While the difference in total within-cluster SS between k=2 and k=3 is relatively small, there is a substantial increase in total between-cluster SS when k increases from 2 to 3.

Table 2: Total Within and Between Sum of Squares for K=2 and K=3

Clusters	Total Within-Cluster SS	Total Between-Cluster SS
k=2	12959.261	3665.739
k=3	11996.15	4628.85

Deciding whether the results of K-means using k=2 are a better representation of the data's structure than the results of k=3 is imprecise and, in this case, visual inspection of the 2-D projection might be the best approach. These projections are shown in Figure 9 below. As with hierarchical clustering with ward distance using 2 clusters, K-Means with k=2 identifies the two clusters, one group is whoen tightly on the left hand side of the plot and one loosely scattered on the right hand side of the plot. With k=3, the algorithm still identifies the tightly grouped observations as one cluster and groups the loosely scattered data points into two clusters. The projection shows two clusters more evidently than it does three. Therefore, despite the improved clustering quality achieved by using k=3 based on within and between SS, k-Means with k=2 seems to reflect the data's structure more accurately than k=3.

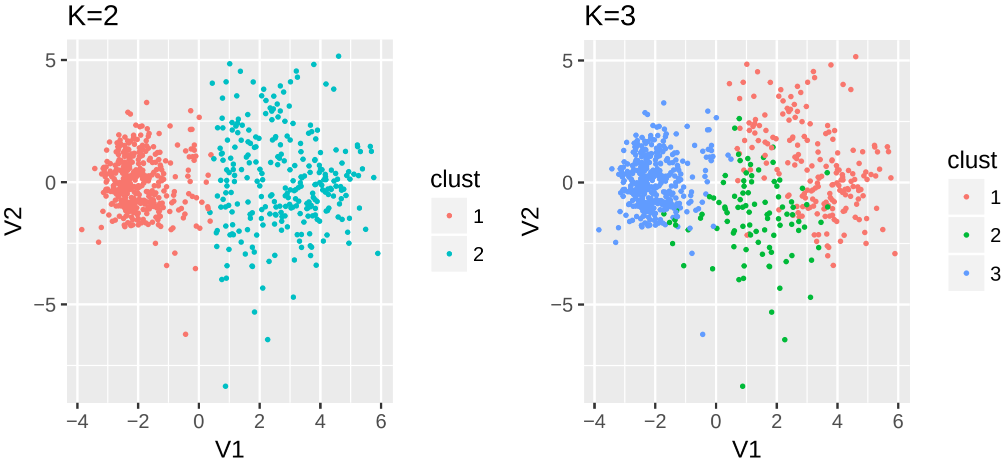

Figure 9: 2-D projection

Table 3 shows the mean value for observations in Cluster 1 and Cluster 2 obtained using K-means for certain variables that are worth mentioning. The summary statistics for these variables is very similar to that obtained using hierarchical clustering. Cluster 1 shows schools with significantly higher mean percentage of economically disadvantaged and minority students than schools in Cluster 2. In fact, it seems that the schools in the "minority" cluster produced with K-means shows a higher concentration of such students than the schools in the "minority" cluster produced using hierarchical clustering. Based on this result, it seems that K-means produces clusters that are more differentiated along these economic and racial dimensions.

Table 3: Mean Percentage Values of Certain Variables for Cluster 1 and Cluster 2 K-Means

Variable_Name	Cluster_1	Cluster_2
Proficiecy	0	0.0839
PER_ECDIS	36.0323	80.4221
PER_WHITE	77.3822	10.2701
PER_BLACK	4.8967	37.5328
PER_HISP	9.767	39.7859
PER_ASIAN	5.0264	10.6886

Using 2-D the projection where observations are differentiated by their percentage of white students, economically disadvantaged and proficiency level, it is clear that the two clusters identified using K-means and Hierarchical clustering truly represent a cluster of "poor, minority" schools and one of "privileged, white schools." Figure 10 shows the observations differentiated into two groups. Group 1 represents schools that have a student body that is more than 50% white students and Group 2 are the other schools. The pattern in Figure 10 resembles that found using the two clustering methods. This result suggests that in the projection V1 could represent, among other things, a measure of a school's student body racial composition.

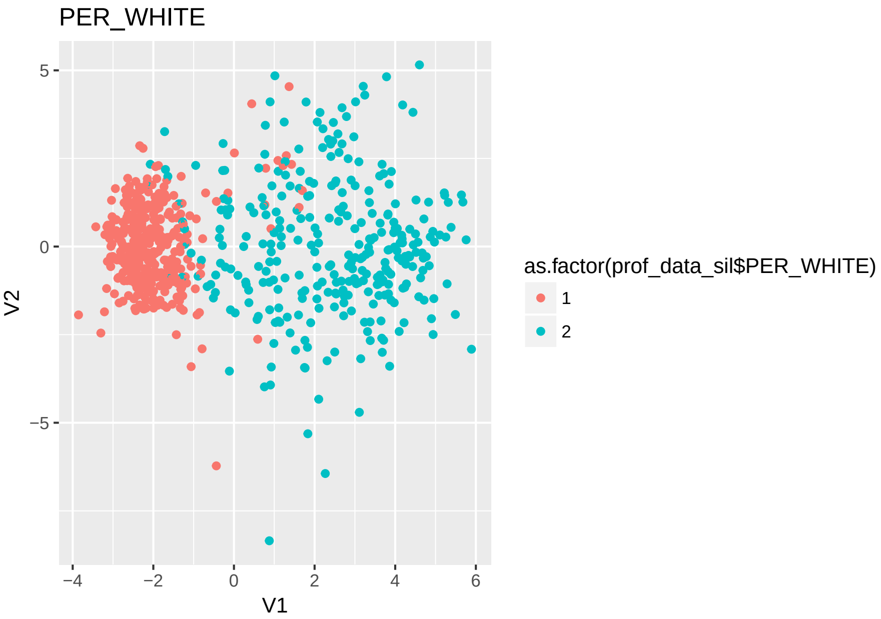

Figure 10: 2-D Projection Differentiated by Pct. White

Figure 11 shows a similar pattern to that of Figure 10. Here, schools with a student body with more than 50% economically disadvantaged are in Group 1 and other schools are in Group 2. Again, this pattern resembles that found using both clustering methods. While the separation along economic lines is not as distinct as along racial lines, the separation is present along the V1 dimension of the projection. Along with a measure of race, V1 seems to represent a measure of the economic position of a school's student body.

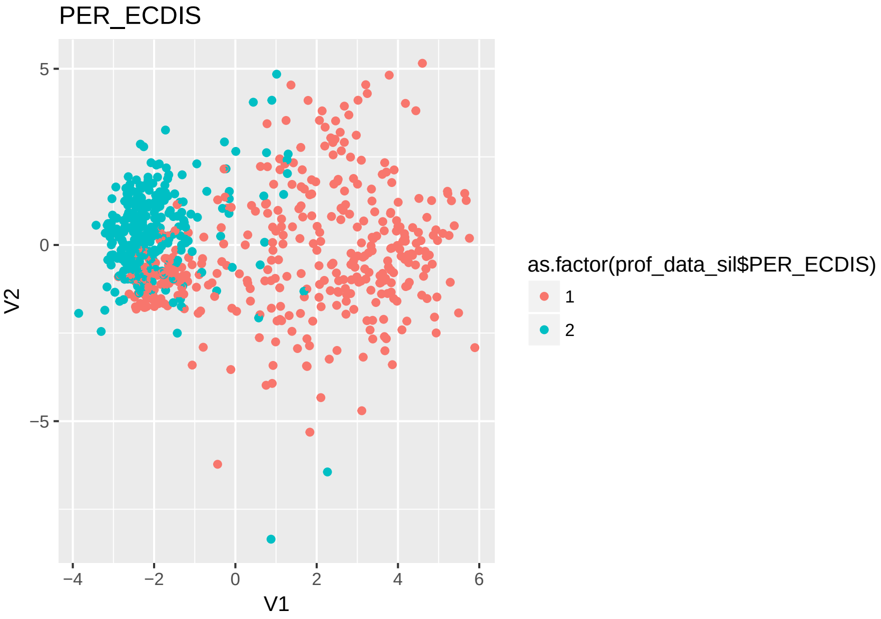

Figure 11: 2-D Projection Differentiated by Pct. ECDIS

While not as distinct as in the previous two figures, it seems that V2 represents some measure school proficiency. Proficient schools slightly greater concentration on the upper half of the Figure 12.

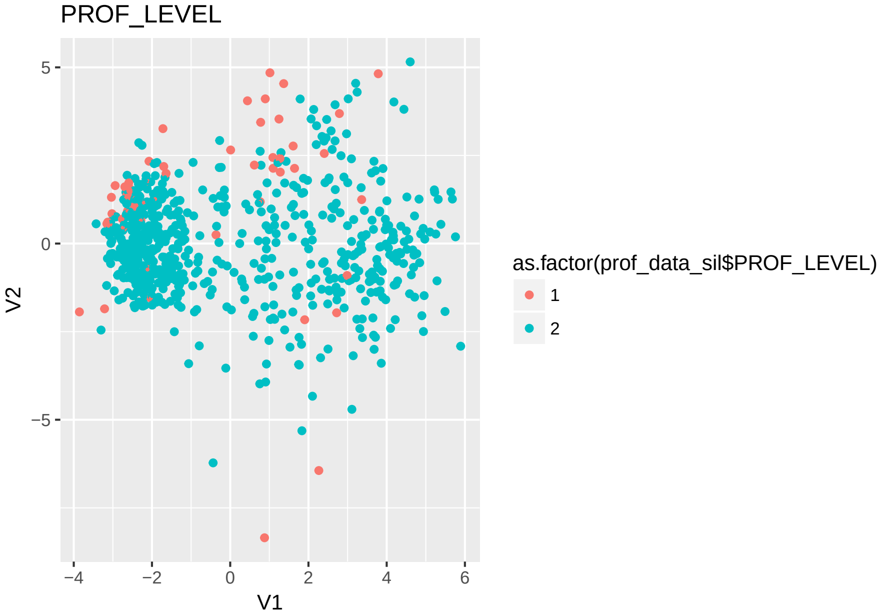

Figure 12: 2-D Projection Differentiated by Proficiency

Discussion

Clustering can be a powerful tool that helps researchers better understand their data in the exploratory phase (and beyond) of their data analysis. As an example of such applications, this report used K-means and hierarchical clustering to identify subgroups in data about school proficiency in New York. The differences between clusters identified in the report could be incorporated into further analysis or help gain a better understanding of the results of inferential methods.

The results demonstrate that the quality of clustering often depends on the specifications of the methods used. With the dataset used in this report, hierarchical clustering with single and complete linkage did not yield clusters that provided any insight into the data's structure. On the other hand, using ward linkage created clusters that differentiated observations based on racial and economic characteristics; variables which are often central to social science projects. Similarly, the number of clusters a researcher determines accurately reflect the data is a holistic process without a definitive answer. This issue was evidenced in the report when having to decide whether k-means with k=2 or k=3 yielded the best result. Ultimately, clustering requires some amount of exploration and trial and error. Such a process will researchers better understand their data.

References

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibsharani. An Introduction to Statistical Learning with Applications in R. Vol. 1. 2013.

Holland, Steven. Data Analysis in the Geosciences: Cluster Analysis. Retrieved from : http://strata.uga.edu/8370/lecturenotes/clusterAnalysis.html

Index: List of Variables and Abreviations

The following is a list of variable abreviations:

```
-PROF_LEVEL: Proficiency level
-PER_FREE_LUNCH: Percent of student body in a school on free lunch
-PER_REDUCED_LUNCH: Percent of student body on reduced lunch
-PER_LEP: Percent of student body that is English Language Learner
-PER_AM_IND: Percent of student body that is Native American
-PER_BLACK: Percent of student body that is Black
-PER_HISP: Percent of student body that is Hispanic
-PER_ASIAN: Percent of student body that is Asian
-PER_WHITE:Percent of student body that White
-PER_Multi: Percent of student body that Multiracial
-PER_SWD: Percent of student body that has a disabiliy
-PER_FEMALE: Percent of student body is female
-PER_ECDIS: Percent of student body that is economically disadvantaged
-PER_SUSPENSIONS: Percent of student body that has received a suspension
-MEAN_CLASS_SIZE: Average class size in grades 3-8
-MEAN_ENROLL: School's average enrollment in 2014-2015
-PER_NOT_HQ: Percent of teachers that are not highly qualified
-STUD_TEACH_RATIO: Ratio of students to teacher
-PER_HS_DEGREE: Percent of people in school district that has a high school degree
-MEAN_INC: Mean income at the school district  level
-EXP_PER_ST: Expenditure per student
-MEAN_TEACH_SCORE: Average teacher assessment score
-CHARTER: Whether school has charter status
-VIOL_CRIME_PER_CITIZEN: Violent crimes per citizen at the county level
-PROP_CRIME_PER_CITIZEN: Property crimes per citizen at the county level
-TOTAL_DIST_POP: Total population at the school district level
```