JMP053: Cluster Analysis in the Public Sector

Exploratory Data Analysis (EDA), Principal Component Analysis (PCA), and Cluster Analysis

Produced by

Robert H. Carver, Professor Emeritus
Stonehill College and Brandeis University International Business School
robert.carver@comcast.net

Cluster Analysis in the Public Sector

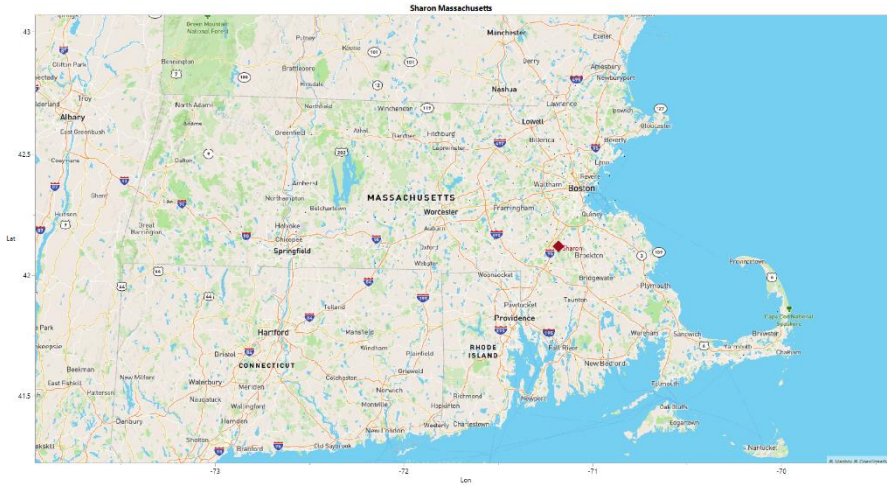
Exploratory Data Analysis (EDA), Principal Component Analysis (PCA), Cluster Analysis

Key ideas:

A Massachusetts (US) town has a research team that wants to exchange information with similar communities about current best practices in local governance. There are 350 other communities in the state, and the all-volunteer team has limited time to interview local officials in other towns. The case demonstrates how multivariate clustering methods can produce a useful list of peer communities. Clustering is a form of unsupervised machine learning in which there is no target, or dependent, variable. We start with a list of informative variables and identify meaningful groupings, or clusters, of individuals who share similar characteristics or attributes.

Background

Like many communities in the northeastern United States, the town of Sharon, MA (2020 pop. 18,575), governs itself with an Open Town Meeting (OTM), an elected Select Board of three members, and several other officials, boards, and departments. The annual Open Town Meeting is the legislative body of the town. During an OTM, any adult citizen of the town can attend, speak, and vote on policies and expenditures. Sharon's Select Board of three members and annual OTM have been the basic governance model since 1765.



Recently, citizen participation has been declining and the Select Board appointed a Governance Study Committee (GSC), charged with reviewing the structures and procedures of town government. The Select Board has asked the GSC to recommend modifications aimed at stimulating citizen engagement in the coming years. The GSC has reviewed state statutes, interviewed and surveyed citizens and local officials, and researched the practices in other states.

Under state law, the municipal legislative function can be performed by

- an Open Town Meeting,
- a Representative Town Meeting (RTM), in which citizens elect a group of neighbors to represent their interests, or
- a smaller Town or City Council.

The Select Board has identified two top-priority issues: 1) should the town adopt an alternative legislative form, and 2) should the Select Board grow from three to five members? The committee's mandate includes other areas, but these are its highest priorities.

Committee members want to interview officials from towns like Sharon to exchange ideas about improving communication with the increasingly diverse citizenry to stimulate involvement in town committees, streamline Town Meetings, use technology more effectively, and so on. Some current town boards have a list of 22 peer towns that they use as a comparison group when setting competitive salaries for municipal employees. Many of these towns are geographically close to Sharon and hence in the same labor market.

The GSC has concluded that "similarity" goes well beyond physical geography and that the factors that might influence civic participation are different from those that drive competition in hiring. Rob is the lone statistician on the GSC, and he has volunteered to locate publicly available data and develop a more suitable list of comparative towns. Imagine that Rob has invited you to assist in this task, and along the way learn about cluster analysis techniques.

Thus far, Rob has gathered data from several sources. The committee has brainstormed a list of variables that might influence voter engagement in local political affairs, and Rob has already done some analysis to narrow down the list of variables and is ready to move forward with clustering.

The Task

At this stage of the project, the following tasks remain:

- **Prepare statewide summary statistics:** The GSC needs a baseline starting point for comparing Sharon to other communities.
- **Data reduction:** The number of candidate explanatory factors is large, with several redundant variables available. We need to reduce the number of columns.
- **Clustering:** Use hierarchical and K Means methods to develop clusters.
- **Evaluate and interpret clusters:** Examine resulting clusters from the perspective of municipal governance.
- **Summarize the governance structures used by similar towns:** Compare Sharon to the most relevant other communities.
- **Prioritize a list of towns to interview:** The GSC has limited time and resources to conduct interviews and must select a small number of communities from a ranked list.

The Data MATowns.jmp

Rob previously compiled data from several public sources including the US Census, the Massachusetts Department of Revenue, the Donahue Institute of the University of Massachusetts, the Massachusetts Taxpayers Foundation, The Massachusetts Municipal Association, the Metropolitan Area Planning Council, and WBUR (a National Public Radio station in Boston).

Rob explains that you'll be treating the data as cross-sectional, but variables were reported at slightly different times. Much of the data comes from 2018, but some columns were gathered at other times as noted below. With limited resources for data gathering, this will have to suffice. In this study, we're not looking for parameter estimates but just intend to improve the list of comparable communities. The columns of the table are as follows:

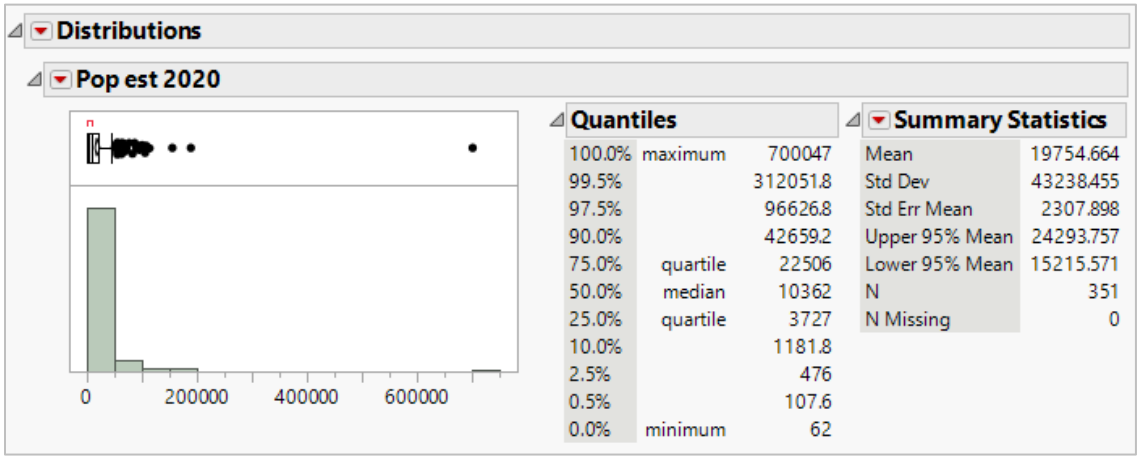
Column	Description
DOR ID Code	MA Dept of Revenue ID code
Legislative	Type of legislative body: Council (number of members), Open Town Meeting, Representative Town Meeting
Name	Community name
SelectSize	Number of Select Board members; cities with Councils have no Select Board
Pop est 2020	Estimated 2020 population (prior to US 2020 Census)
Pop Density 2018	Population per square mile of land area
Pop 18+_p	Percent of population 18 years and older
IncomePerCap	Annual income per capita, 2018
LowIncStudents_p	2020 percent of low-income students enrolled in public schools
RegVoters2018	Number of registered voters living in the community
RegVoters_p 2018	Percent of eligible residents who are registered voters
HouseUnits2020	Total number of housing units
Household Pop 2020_p	Percent of residents living within households (as opposed to university housing, nursing facilities, prisons, etc.)
EQV Per Cap	Equalized property valuations per capita. MA communities report standardized estimates of the market value of all properties. The values are used to assess property taxes.
assist_p	Percentage of households receiving public assistance, 2015-19
nonCit_18o_p	Percentage of noncitizens 18 and over, 2010-14
hugrow2010_2020	Percent growth in housing units, 2010-20
labforce_p201418	Labor force as a percent of the population, 2014-18
lab16ovr201418	Number of people 16 years and older in the labor force, 2014-2018
Lab16over_p	Percentage of individuals 16 years and older in the labor force, 2014-2018
Commute_60_p	Percentage of residents with a daily one-way commute of 60 minutes or longer
bachplus_p	Percentage of residents with at least a bachelor's degree
noncitz_p	Percentage of residents who are not US citizens
White%	Percentage of residents who identify as white
lat	Latitude of city or town hall
lon	Longitude of city or town hall

Analysis

Descriptive Statistics

Because the Committee’s top priorities involve citizen engagement in the legislative and executive functions, let’s begin with descriptive statistics of town populations and basic governance structures.

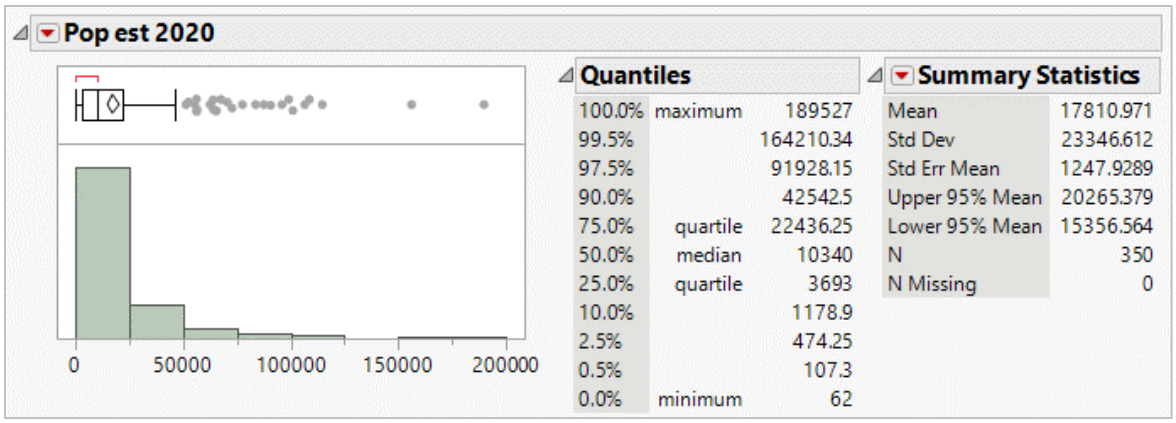
Exhibit 1 Summary Statistics of Population



To create Exhibits 1 through 4, Analyze>Distribution. Drag Pop est 2020, Legislative, and SelectSize into the Y drop zone > OK. Under the red triangle next to Distributions, select Stack to align the output horizontally.

We see a strongly right-skewed distribution with a single outlier, which is the city of Boston. Boston is the state capital and economic hub of the state. Given its unique characteristics, the GSC decided to omit it from further analysis. Exhibit 2 shows the variation in local populations for the 350 remaining communities, after hiding and excluding Boston's row (Note: Boston is row number 36, where the estimated population is 700,047).

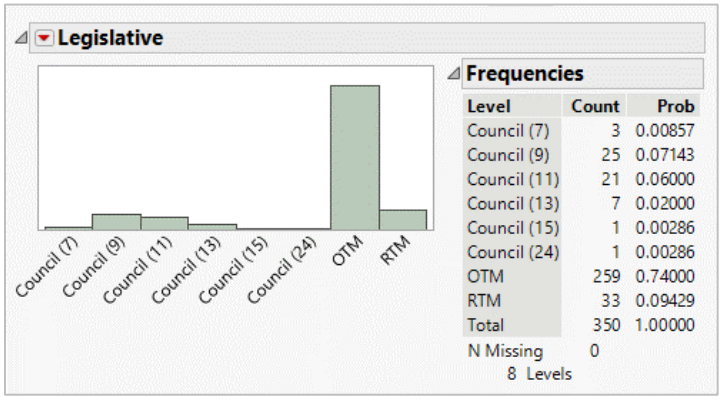
Exhibit 2 Summary Statistics of Population, Hiding and Excluding Boston



To create, highlight the rightmost outlier in the Exhibit 1 boxplot. Right-click and choose Row Hide and Exclude. Then under the red triangle next to Distributions, select Redo>Automatic Recalc.

At the time the data table was constructed, Sharon's estimated 2020 population was 17,656, which is very close to the statewide mean.

Exhibit 3 Distribution of Legislative Bodies



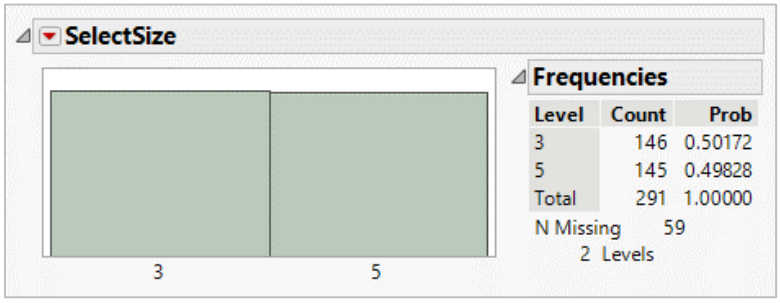
Level	Count	Prob
Council (7)	3	0.00857
Council (9)	25	0.07143
Council (11)	21	0.06000
Council (13)	7	0.02000
Council (15)	1	0.00286
Council (24)	1	0.00286
OTM	259	0.74000
RTM	33	0.09429
Total	350	1.00000

N Missing 0
8 Levels

This exhibit was generated at the same time as Exhibit 1 and modified when we removed Boston from the analysis. Note that there are only 350 communities in the distribution.

OTM is by far the most common legislative form, with 74% of the communities using it. Fifty-eight communities (17%) have Councils of different sizes, and just 9% use RTMs.

Exhibit 4 Distribution of Select Board Size



Level	Count	Prob
3	146	0.50172
5	145	0.49828
Total	291	1.00000

N Missing 59
2 Levels

This exhibit was generated at the same time as Exhibit 1.

Under Massachusetts law, the 59 municipalities with Councils are classified as cities, and the executive/administrative function is performed by a mayor or city manager. The remaining towns are split almost evenly between three- and five-member Select Boards.

In short, the traditional combination of a three-member board and an Open Town Meeting remains the dominant model among Massachusetts communities. Within the framework of these basic structures, there are procedural options that might address the future needs of the busy lives of citizens, the growing complexity of legislative issues, and the increasing diversity of the communities. Local self-governance is unfamiliar to many new residents, especially those arriving from other nations or US states.

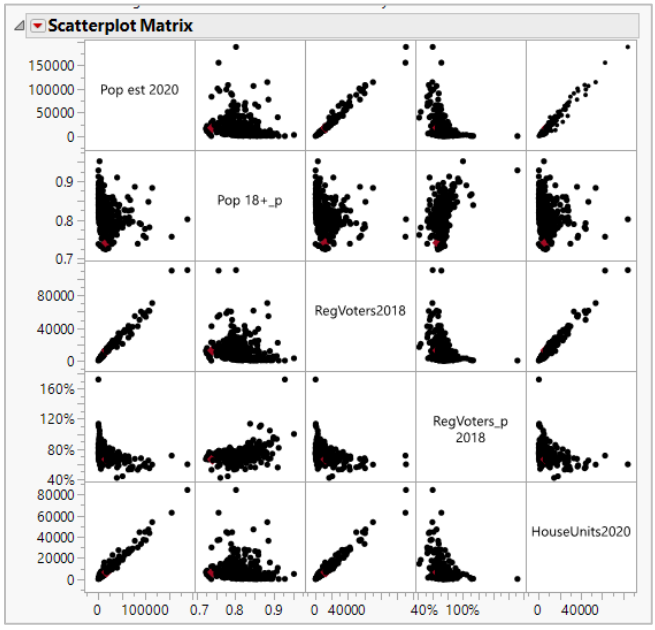
Before proceeding, you ask Rob if there are outliers in addition to Boston or whether missing data will be an issue. Fortunately, the data table is reasonably complete. Nearly all the candidate clustering variables have 350 or 351 observations. There are just two communities with missing data, but that should not impede the rest of the analysis. In preliminary work, Rob also confirmed that apparent outliers are accurate, if unusual.

Data Reduction and Variable Selection: Principal Component Analysis (PCA)

Unfortunately, there is no central repository of Town Meeting attendance data in Massachusetts. In a prior analysis that modeled voter participation in the 2018 statewide election for governor as a proxy for citizen engagement, Rob identified several constructs as predictors of voter turnout, including community size, affluence, levels of education, commuting habits, diversity, and tax base. In our data table, these are the 20 variables represented in columns 5 through 24. Several of the columns appear to measure a single construct, and Rob is concerned about the eventual need to explain an unnecessarily complex model to the GSC and Select Board.

In general, we want to maximize the information value of the available data to model town similarity, without including extraneous or redundant variables. For example, we have several highly correlated columns that reflect the number of residents in a town. While town size is surely relevant, once we have accounted for the population size, adding the number of registered voters may not tell us much more.

Exhibit 5 Relationships Among Several Columns

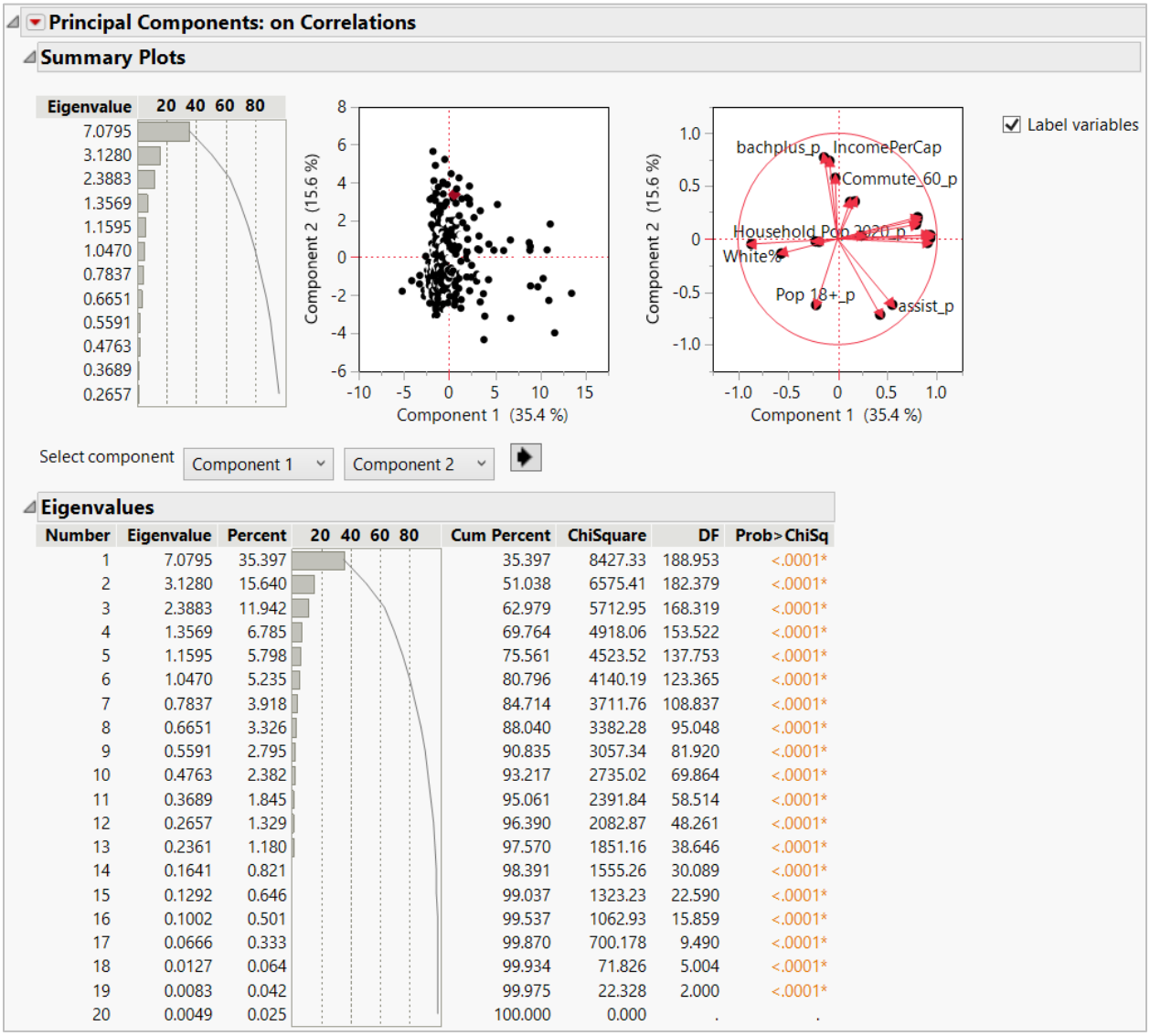


To create, Analyze>Multivariate Methods>Multivariate. Select the columns shown as Y, Columns. Click OK.

One common method for reducing the number of columns is principal component analysis (PCA). As a simple conceptual introduction to PCA, consider the scatterplot matrix in Exhibit 5. Three of the columns have strong, positive, bivariate relationships. PCA would essentially construct an artificial column that captures the common information embedded within the three columns and allow us to condense the three into one.

After consulting with some academic colleagues and the JMP Academic Program team, Rob proposes that we first run a PCA using all 20 columns and capture a modest number of principal components. For now, look at the PCA report in Exhibit 6.

Exhibit 6 The Principal Component Analysis Report



Eigenvalues

Number	Eigenvalue	Percent	Cum Percent	ChiSquare	DF	Prob>ChiSq
1	7.0795	35.397	35.397	8427.33	188.953	<.0001*
2	3.1280	15.640	51.038	6575.41	182.379	<.0001*
3	2.3883	11.942	62.979	5712.95	168.319	<.0001*
4	1.3569	6.785	69.764	4918.06	153.522	<.0001*
5	1.1595	5.798	75.561	4523.52	137.753	<.0001*
6	1.0470	5.235	80.796	4140.19	123.365	<.0001*
7	0.7837	3.918	84.714	3711.76	108.837	<.0001*
8	0.6651	3.326	88.040	3382.28	95.048	<.0001*
9	0.5591	2.795	90.835	3057.34	81.920	<.0001*
10	0.4763	2.382	93.217	2735.02	69.864	<.0001*
11	0.3689	1.845	95.061	2391.84	58.514	<.0001*
12	0.2657	1.329	96.390	2082.87	48.261	<.0001*
13	0.2361	1.180	97.570	1851.16	38.646	<.0001*
14	0.1641	0.821	98.391	1555.26	30.089	<.0001*
15	0.1292	0.646	99.037	1323.23	22.590	<.0001*
16	0.1002	0.501	99.537	1062.93	15.859	<.0001*
17	0.0666	0.333	99.870	700.178	9.490	<.0001*
18	0.0127	0.064	99.934	71.826	5.004	<.0001*
19	0.0083	0.042	99.975	22.328	2.000	<.0001*
20	0.0049	0.025	100.000	0.000	.	.

To create, Analyze>Multivariate Methods>Principal Components. Select the columns from Pop est 2020 through White% as Y, Columns. Click OK. Select Eigenvalues and Bartlett Test from the top red triangle.

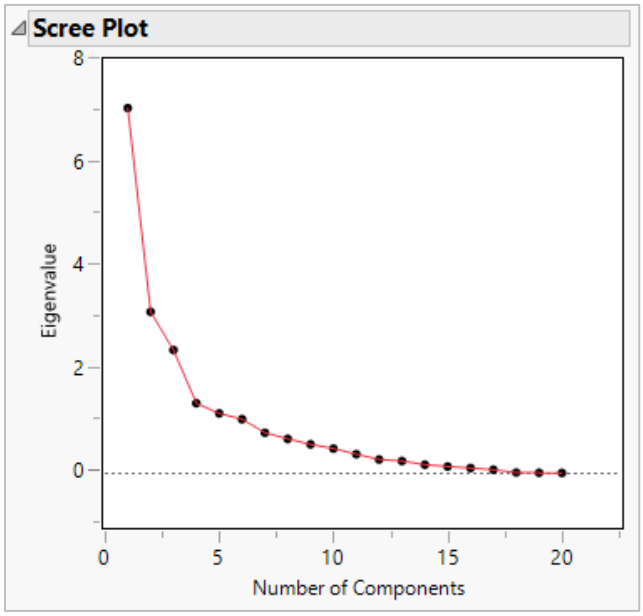
The rightmost column in the report refers to Bartlett's test, which checks if the observed correlation matrix diverges significantly from the identity matrix. That is, it tests against the null hypothesis that the variables are orthogonal. One can perform PCA only if we reject the null hypothesis. In this case, the Prob>Chisq values are less than 0.05 and thus we can proceed with PCA.

The PCA always generates as many potential components as the number of columns. As we see in the left plot of eigenvalues, though, the first few components account for most of the variation in the data. The score plot in the center shows the Component 1 and 2 scores for each town. We note that these two components are orthogonal and embody different information about the towns.

The rightmost loading plot displays the unrotated loading matrix for the first two components, with each variable labeled, showing how components and the original variables are related. The red circle has a radius of 1, and points closer to 1 are variables that are heavily loaded in the component. We see several variables, including lab16ovr20142018 and noncitz_p, point in the Component 1 direction, while bachplus_p and IncomePerCap point in the Component 2 direction. These components and variables are also orthogonal.

The first two components capture much of the variability across the 20 columns. Capturing all of the variability requires using 20 components, which defeats the goals of dimension reduction. This prompts the question, how many components should we use to reduce the number of columns from 20 to a more manageable number? For that judgment, we must look at a scree plot.

Exhibit 7 A Scree Plot



To create, click the red triangle next to Principal Components: on Correlations. Select Scree Plot.

A scree plot places eigenvalues on the y-axis and the number of components on the x-axis. Scree plots are always concave and slope downward. We look for the kink in the curve as a guide when deciding how many components to rely on for further use. In Exhibit 7, it appears that four components capture most of the information value. After six components, the eigenvalues drop off substantially. It seems that the wise course is to use between four and six components, and for that choice we want to think about what the components represent in the context of this use case.

To relate principal components to their real-world meanings, a loading matrix helps. We can look at which variables load most heavily into each component and then rely on the GSC's domain knowledge to describe the different components. JMP facilitates the process by dimming small loading values, and the user can further control the dimming and highlighting.

Exhibit 8 A Formatted Loading Matrix

Formatted Loading Matrix

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11
lab16ovr201418	0.933425	0.036650	0.084837	0.074029	0.301685	0.076674	-0.085165	0.034249	-0.029467	0.006804	-0.009574
Pop est 2020	0.931853	0.006751	0.100151	0.034523	0.329576	0.037027	-0.032088	0.023995	-0.031328	-0.014184	-0.027513
HouseUnits2020	0.904993	-0.038891	0.145311	0.074776	0.362635	0.053352	-0.034449	0.015728	-0.021590	0.039405	-0.017582
RegVoters2018	0.894061	0.033661	0.106966	0.047694	0.415757	0.050519	-0.043561	0.005511	-0.037477	-0.005206	-0.017749
nonCit_18o_p	0.804969	0.204638	0.159609	0.061440	-0.461194	-0.028344	-0.007398	-0.003463	0.041110	0.005203	-0.078584
noncitz_p	0.799331	0.177190	0.208845	0.059455	-0.464745	-0.024765	-0.033587	-0.004218	0.011970	0.021696	-0.060157
Pop Density 2018	0.788861	0.136365	0.109867	0.173276	-0.269485	0.033626	-0.172836	0.084520	0.085846	0.222613	-0.028113
assist_p	0.550532	-0.622748	-0.166924	-0.092065	0.023932	-0.265425	0.274875	-0.017650	-0.030341	-0.050045	-0.016205
bachplus_p	-0.135567	0.770389	0.478788	-0.002964	0.030175	0.018310	-0.115682	-0.081568	-0.130303	0.083529	0.062303
IncomePerCap	-0.081728	0.737168	0.295668	-0.143655	0.112137	-0.295442	0.062552	-0.306179	0.028191	0.172146	0.255273
EQV Per Cap	-0.220290	-0.024078	0.714485	0.446096	-0.018440	0.006125	0.182588	-0.030205	0.259259	-0.281793	0.034167
RegVoters_p 2018	-0.564181	-0.137775	0.506176	0.498558	0.180500	0.046691	0.046687	0.068381	0.033722	-0.054295	0.019042
Pop 18+_p	-0.215135	-0.624754	0.418827	0.063200	-0.111568	0.342021	-0.308167	0.179563	0.053875	0.241524	0.101175
Household Pop 2020_p	-0.186111	-0.027730	-0.339118	0.749498	0.025219	-0.294872	0.186635	-0.095122	-0.104726	0.325170	-0.150900
labforce_p201418	0.126122	0.350234	-0.543342	0.472942	-0.122869	0.117625	-0.241497	0.097857	-0.337059	-0.268436	0.220872
White%	-0.865383	-0.049338	-0.130907	-0.004529	0.230002	0.129025	-0.174610	0.048875	-0.017205	0.199163	-0.006503
hugrow2010_2020	0.174121	0.354277	-0.157448	-0.006620	-0.044972	0.646288	0.597429	0.115257	-0.031455	0.131876	0.103852
Commute_60_p	-0.023074	0.575919	-0.308728	0.003285	0.112312	-0.315817	0.009875	0.581618	0.333646	0.024787	0.066785
Lab16over_p	0.240211	0.028971	-0.650582	0.193021	0.047405	0.259384	-0.170871	-0.380184	0.467933	-0.019921	0.068046
LowIncStudents_p	0.427858	-0.716666	0.002338	0.040237	-0.096242	-0.228791	0.089675	0.016624	-0.006222	0.093572	0.429317

Suppress Absolute Loading Value Less Than 0.3

Dim Text 0.4

To create, click the red triangle next to Principal Components: on Correlations. Select Formatted Loading Matrix. For later purposes, save these six components by clicking the red triangle once again and choosing Save Columns>Save Principal Components. Request six principal components. This will update the data table with six principal components, which are nothing but the linear combination of the variables.

In the data table, assign the names shown in the list below.

In this case, some meanings emerge quite naturally but others are ambiguous. After consulting with others on the GSC, Rob tentatively names the first six components as:

- **Size**: various measures of the number of people in the locale
- **Affluence**: measures of financial well-being and need
- **PropertyValue**: market value of real estate and factors contributing to value
- **PermanentPopulation**: portion of the population that ordinarily resides in the locale
- **PolitySize**: the voting population relative to the total population
- **Growth**: recent increases in housing stock

Exhibit 9 First 10 Rows of the Principal Components

Size	Affluence	PropertyValue	PermanentPopulation	PolitySize	Growth
0.2250601763	0.4577789106	-1.717751369	0.8134618806	0.1359012933	-0.116582339
1.9274542813	3.1830396763	0.1031732954	0.1233951295	-1.433148166	-0.511472003
-0.850191696	-1.108161135	-1.206394345	0.0432245066	0.1053824467	0.2416221166
-0.76405107	-2.714119813	-0.78103626	-0.240565015	0.0901752109	-0.798944945
0.6612271307	-1.50596554	-0.287983242	0.1785296662	0.8229328425	0.6870087396
-2.234220575	-1.200755303	1.8103716198	-0.010325375	-0.349582517	-0.683955762
-0.214908061	0.0380101296	-0.945138955	0.3599631097	0.3279537975	0.7692341734
3.0156328501	-0.868398208	4.9515188262	-7.77321794	-1.790550505	3.7484316764
1.3533836405	3.1241492124	0.7581646834	-0.532442185	0.6027724664	0.0480533054

With these six principal components, we can derive clusters of towns that share these attributes. The clustering analysis will use the six components rather than the original 20 variables.

Clustering: Two Algorithms

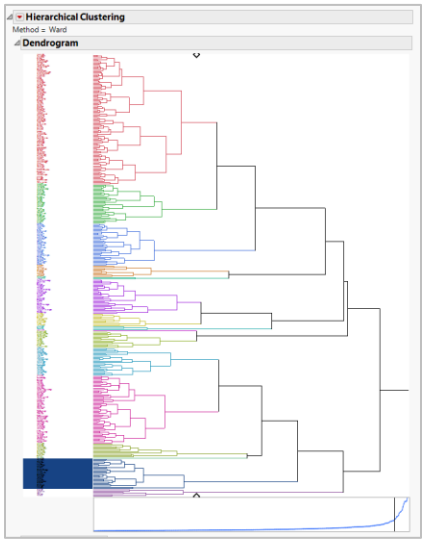
Rob explains that we'll apply two clustering algorithms to the data and compare the results. Both methods compute multivariate Euclidean distances between each pair of observations (municipalities) based on the standardized values of the principal components. Those distances are then stored in a 350 x 350 distance matrix containing all pairwise distances. The algorithms are iterative, developing a set of clusters based on the multivariate distances between towns.

In hierarchical clustering, each of the 350 communities is initially treated as a unique cluster. Working with the distance matrix, JMP finds the two communities that are most similar. They become a new cluster, and JMP calculates the mean of their combined distances. That mean becomes the cluster centroid, and subsequent iterations combine the next two closest clusters. Step by step, the number of clusters is reduced until there is one that contains all 350 observations. JMP then recommends an optimal number of clusters.

With the K Means approach, the user initially specifies a desired number of clusters, K. The algorithm initially creates K clusters, assigning each observation to a cluster and then computing the centroid (mean) for each cluster in the multivariate space. In each iteration, it compares the distance of each observation to other nearby centroids. If observation *i* is closer to a different cluster's mean than it is to its current cluster's mean, that observation is reassigned to the new cluster. Then all centroids are recomputed, and the process continues until distances are minimized or fall within a prespecified limit.

Exhibit 10 shows the results of the hierarchical method.

Exhibit 10: Hierarchical Cluster Dendrogram on Six PCs: 16 Clusters

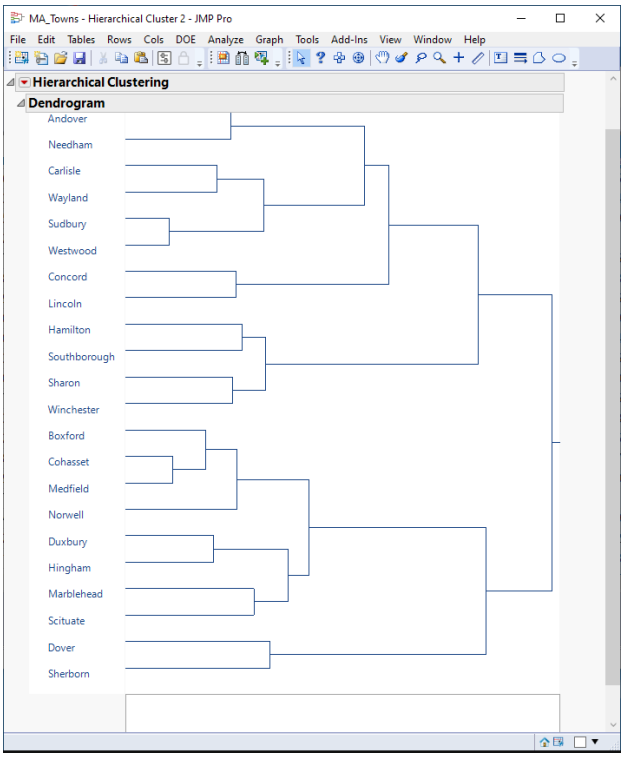


To create, select Analyze>Clustering>Hierarchical Cluster. To color the clusters, click the red triangle next to Hierarchical Clustering and select Color Clusters. To select one cluster, click on the branch that the cluster elements share. Save the cluster assignments to the data table by clicking the red triangle next to Hierarchical Clustering and choosing Save Clusters.

The most distinctive feature of the cluster report is the dendrogram, a tree diagram that illustrates the agglomerative process of gathering observations into clusters. Reading from left to right, notice the initial list of 350 unique town clusters. Moving rightward, each branch indicates the addition of towns and clusters. Also, note the diamond-shaped handles at the top and bottom. Initially, they are positioned to indicate the optimal number of clusters, but the user can slide them left or right to adjust the number of clusters. In this example, JMP recommended 16 clusters, and Rob has highlighted the cluster containing

Sharon. In Exhibit 10, the names of individual communities in the cluster are difficult to discern; Exhibit 11 shows an enlarged image.

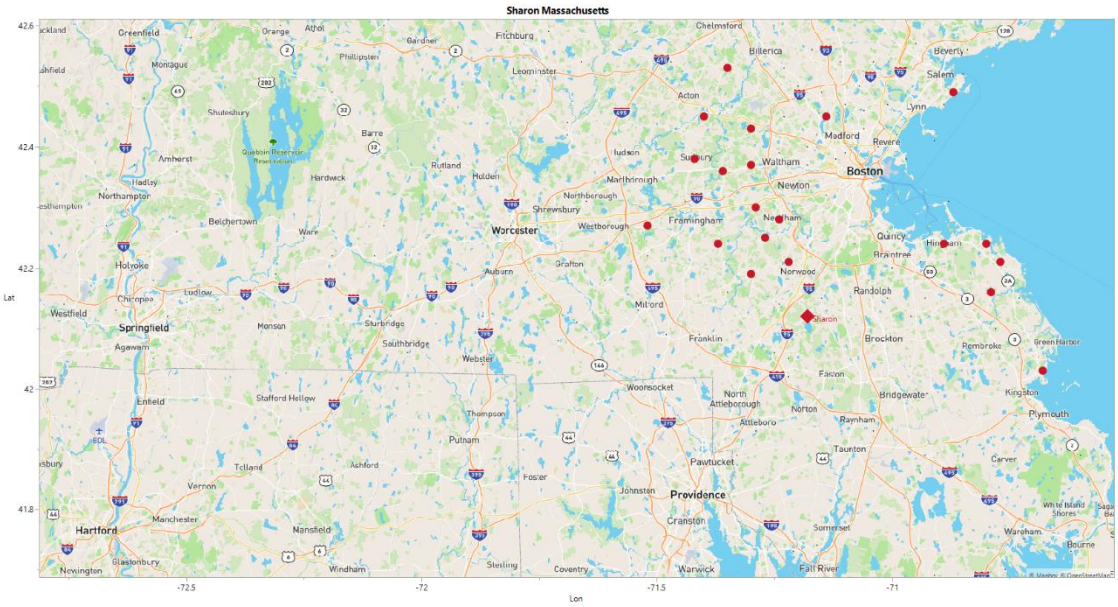
Exhibit 11: Zoom In on Cluster Including Sharon



To create, click the red triangle next to Hierarchical Cluster and choose Zoom to Selected Rows.

In the enlarged portion of the tree, we can see that Sharon first paired with the town of Winchester. In turn, the Sharon-Winchester dyad is more similar to the pair of Hamilton and Southborough than to any other pair of towns.

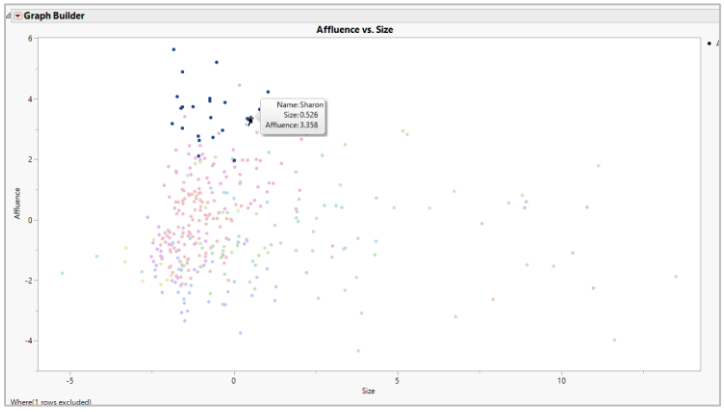
Exhibit 12: Geographic Locations of Towns in Cluster 15



To create, first subset the data table to include only Cluster 15 towns. Then open Graph>Graph Builder and drag the lat and lon columns into the central area. Select the Mapbox Outdoors as the background map, and Color by Cluster.

Because hierarchical clustering is unfamiliar to the GSC members and other town officials, Rob suggests you construct a scatterplot spreading the towns out across the first two principal components, Size and Affluence. From Exhibit 6, we know that these two components together account for approximately 50% of the variation among the towns and that the two components are orthogonal. Hence, this scatterplot will spread the towns out widely. If we then color the points by component, this graph (see Exhibit 13) can be a useful way of communicating what we mean by choosing towns like Sharon across several dimensions.

Exhibit 13 Scatterplot of First Two Principal Components

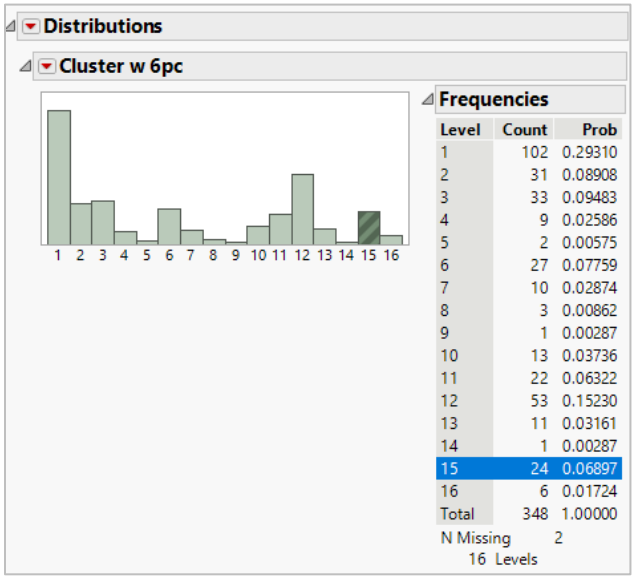


To create, Graph>Graph Builder. Place Affluence on the y-axis and Size on the x-axis and click Done. Remove the default Smoother.

We see the dark blue dots from Sharon's cluster and note that they are all in the same region of the scatterplot. Many towns may be approximately the same size as Sharon but quite different along the Affluence dimension. A viewer might wonder why there are some pink and pale-blue dots that are close to Sharon in the scatterplot but in different clusters. At that point, it should be easy to explain that they are farther apart when other components are taken into account.

Thus far, we might say that towns in Sharon's cluster are of above-average size and among the most affluent in the state. What else can we say descriptively about the clusters?

Exhibit 14 Tally of Clusters

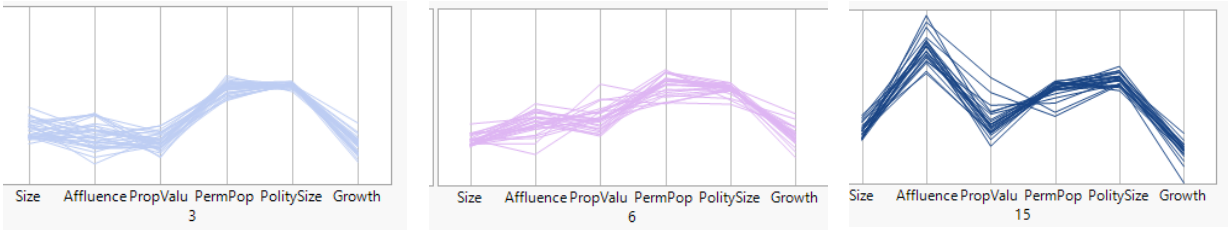


Level	Count	Prob
1	102	0.29310
2	31	0.08908
3	33	0.09483
4	9	0.02586
5	2	0.00575
6	27	0.07759
7	10	0.02874
8	3	0.00862
9	1	0.00287
10	13	0.03736
11	22	0.06322
12	53	0.15230
13	11	0.03161
14	1	0.00287
15	24	0.06897
16	6	0.01724
Total	348	1.00000
N Missing	2	
16 Levels		

To create, Analyze>Distribution. Select the Cluster column as Y, Column.

Hierarchical clustering produced 16 clusters, and Sharon is one of 24 towns in Cluster 15. This represents slightly under 7% of the towns. How do towns in the is cluster compare to one another and differ from other clusters? To answer this question, we consult a parallel plot.

Exhibit 15 Comparing the Attributes of Several Clusters



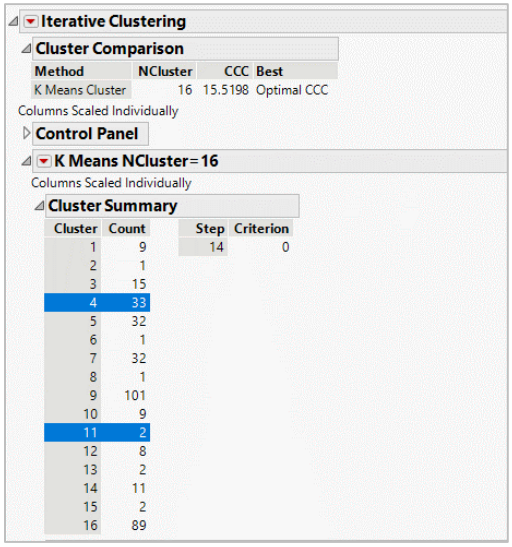
To create, within the Hierarchical Clustering report, click the red triangle at the top and choose Parallel Coord Plots.

Exhibit 15 displays the plots for three of the 16 clusters to illustrate how to interpret parallel plots. Each town is represented by a colored line. The height of each line represents the relative value for each of the six components.

We see, for example, that compared to Cluster 3, Cluster 15 towns are a bit larger and far more affluent, with slightly higher property values, similar permanent population proportions, slightly larger polity sizes, and similar but more varied growth profiles. In contrast, towns in Cluster 6 are faster growing, poorer, and so on. Assuming that these six principal components are those most relevant to governance, it makes sense for the GSC to look to the towns in Cluster 15 for peer-to-peer comparisons and information sharing.

Hierarchical clustering is not the final word. Before accepting these findings as a basis for action, it makes sense to run a K Means analysis and compare the results. Because the hierarchical approach yielded 16 clusters, let's use 16 as the desired number of clusters.

Exhibit 16 Results of K Means Method



Iterative Clustering

Cluster Comparison

Method	NCluster	CCC	Best
K Means Cluster	16	15.5198	Optimal CCC

Columns Scaled Individually

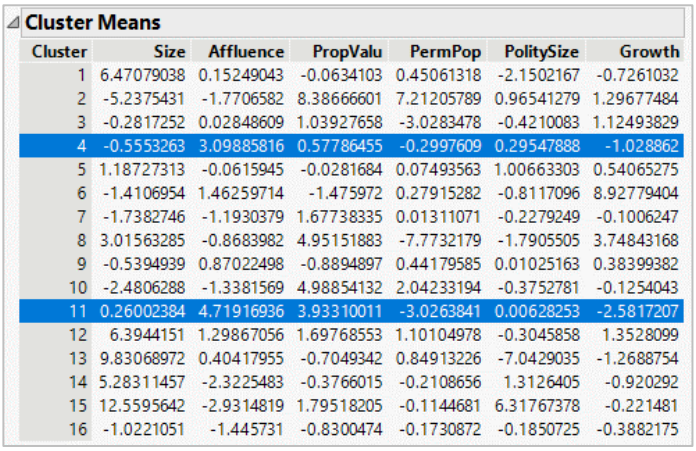
Control Panel

K Means NCluster=16

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	9	14	0
2	1		
3	15		
4	33		
5	32		
6	1		
7	32		
8	1		
9	101		
10	9		
11	2		
12	8		
13	2		
14	11		
15	2		
16	89		



Cluster Means

Cluster	Size	Affluence	PropValu	PermPop	PolitySize	Growth
1	6.47079038	0.15249043	-0.0634103	0.45061318	-2.1502167	-0.7261032
2	-5.2375431	-1.7706582	8.38666601	7.21205789	0.96541279	1.29677484
3	-0.2817252	0.02848609	1.03927658	-3.0283478	-0.4210083	1.12493829
4	-0.5553263	3.09885816	0.57786455	-0.2997609	0.29547888	-1.028862
5	1.18727313	-0.0615945	-0.0281684	0.07493563	1.00663303	0.54065275
6	-1.4106954	1.46259714	-1.475972	0.27915282	-0.8117096	8.92779404
7	-1.7382746	-1.1930379	1.67738335	0.01311071	-0.2279249	-0.1006247
8	3.01563285	-0.8683982	4.95151883	-7.7732179	-1.7905505	3.74843168
9	-0.5394939	0.87022498	-0.8894897	0.44179585	0.01025163	0.38399382
10	-2.4806288	-1.3381569	4.98854132	2.04233194	-0.3752781	-0.1254043
11	0.26002384	4.71916936	3.93310011	-3.0263841	0.00628253	-2.5817207
12	6.3944151	1.29867056	1.69768553	1.10104978	-0.3045858	1.3528099
13	9.83068972	0.40417955	-0.7049342	0.84913226	-7.0429035	-1.2688754
14	5.28311457	-2.3225483	-0.3766015	-0.2108656	1.3126405	-0.920292
15	12.5595642	-2.9314819	1.79518205	-0.1144681	6.31767378	-0.221481
16	-1.0221051	-1.445731	-0.8300474	-0.1730872	-0.1850725	-0.3882175

To create, Analyze>Clustering>K Means Cluster. Cast the six principal components as Y, Columns. Specify 16 clusters and click Go.

Exhibit 16 displays the results. By this method, our previously selected rows now fall into two different clusters. Sharon is in Cluster 4 by this method, with 33 towns in all, and two of the prior Cluster 16 towns move to Cluster 11. Exhibit 15 lists all 33 towns in Cluster 4.

Exhibit 17 Cluster 4 Towns by K Means Method

DOR ID Code	Legislative	Name	SelectSize
335	OTM	Westwood	3
288	OTM	Sudbury	5
119	OTM	Hamilton	5
315	OTM	Wayland	5
038	OTM	Boxford	5
344	RTM	Winchester	5
175	OTM	Medfield	3
065	OTM	Cohasset	5
034	OTM	Bolton	3
115	OTM	Groton	5
168	OTM	Marblehead	5
298	OTM	Topsfield	5
189	RTM	Milton	5
051	OTM	Carlisle	5
199	RTM	Needham	5
082	OTM	Duxbury	3
092	OTM	Essex	3
266	OTM	Sharon	3
157	OTM	Lincoln	3
324	OTM	West Newbury	3
264	OTM	Scituate	5
159	OTM	Longmeadow	5
277	OTM	Southborough	5
009	OTM	Andover	5
219	OTM	Norwell	5
023	OTM	Bedford	5
131	OTM	Hingham	3
155	RTM	Lexington	5
067	OTM	Concord	5
078	OTM	Dover	3
002	OTM	Acton	5
269	OTM	Sherborn	5
026	RTM	Belmont	3

To create, click the red triangle next to K Means NCluster = 16 and choose Save Clusters. Also save Cluster Distances. This creates two new columns in the data table. Select the rows where this cluster = 4 and select Tables>Subset to extract a new data table of Cluster 4 towns. Within the subset, Tables>Sort by Cluster Distances.

This cluster is larger than Cluster 15 in the hierarchical analysis. When we compare the list to Exhibit 11 and recall that just two towns fall into K Means Cluster 11, we note that most of the towns from Cluster 15 also are grouped with Sharon in this approach.

What Can the Clusters Teach Us?

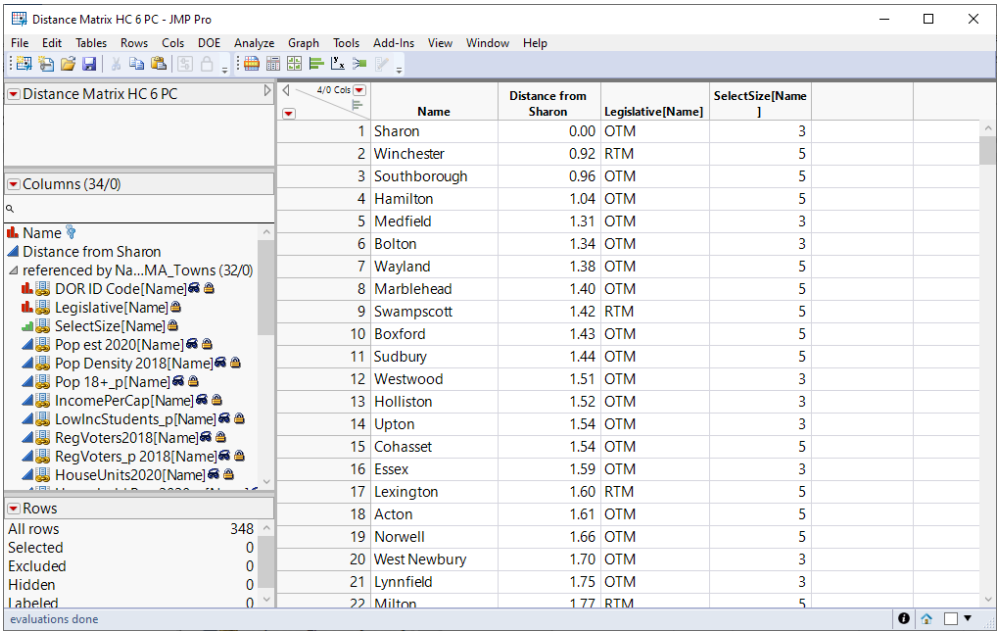
Exhibit 18 Governance Structures in HC Cluster 15, K Means Cluster 4, and Statewide

HC Cluster 15 (n = 24) **KMeans Cluster 15 (n = 33)** **Statewide (n=350)**

Regardless of the clustering method, we find that the basic governance structures among towns like Sharon follow very similar patterns: Open Town Meetings and five-member Select Boards. Approximately 85-88% of towns like Sharon use OTM, in contrast to 74% statewide. Notably, in both clustering methods, only one-third of towns have a three-member Select Board like Sharon has. Statewide, both board sizes are equally common. The prevalence of three-member boards certainly looks different when we focus on the communities most similar to one another, rather than the all-state base rate.

The GSC also wants to interview officials from peer towns. The GSC is an appointed volunteer committee that is subject to the Massachusetts Open Meeting Law, which requires all meetings to be conducted publicly in person or via publicly accessible video conference. With no budget and limited time, it is unrealistic to hold repeated public meetings with numerous peer officials testifying. The GSC wants to speak with people from OTM, RTM, and Council municipalities, and from three- and five-member Select Board towns. As such, the committee wants to prioritize its contact efforts. Because both clustering methods rely on multivariate Euclidean distances between towns, we can use the distance matrix to sort towns from most to least similar.

Exhibit 19 Distance Matrix Sorted Closest to Farthest



To create, open the Distance Matrix from the Hierarchical Clustering report. Delete all columns except for Name and Sharon. Change the Name Column Properties to make it a Link Reference to the original data table. Unhide Legislative and SelectSize.

In contrast to the lists shown in Exhibits 11 and 17, this list contains all towns regardless of cluster. It is helpful to the GSC in several ways. It identifies more communities that are similar to Sharon while highlighting that the distance scores may be quite similar for two closely ranked towns. Also, since the GSC wants to consult with RTM and Council communities that might be similar to Sharon but in a different cluster, the distance matrix makes short work of selecting likely prospects.

Of course, the GSC cannot compel another community's assistance in this study, so it is it helps to prioritize a roster of potential contacts. The customary impulse is to confer with our nearest geographic neighbors with whom we often cooperate, but this analysis underscores the fact that neighboring towns may not be most comparable. The 22 towns visible in Exhibit 19 vary in physical distance from Sharon, and none of them shares a border with Sharon.

Summary

Statistical Insights

To summarize, this case called upon analysts to use JMP for the following:

- Generating summary statistics
- Exploring variables and possible relationships with Distribution and scatterplot matrices
- Performing and interpreting a principal component analysis
- Using hierarchical clustering to identify groups
- Using parallel plots to characterize clusters
- Using Graph Builder to prepare presentation materials
- Using K Means clustering to identify groups
- Interpreting a distance matrix

Implications

Rob and the Governance Study Committee can draw the following conclusions from the analysis:

- The towns most like Sharon tend to use Open Town Meetings more often than towns statewide.
- The towns most like Sharon are far more likely to have a five-member Select Board than a three-member board, as Sharon does.
- As the GSC reaches out to peer communities to exchange experience and ideas, this study provides a way to prioritize their contact efforts.

JMP Features

This study made use of several JMP platforms. Early on, it used the Distribution platform to display histograms and summary statistics; it also used Graph Builder to visualize data on geographic maps and scatterplots. The Multivariate platform displayed correlations.

Essential to addressing the committee's questions were principal component analysis for data reduction and clustering analysis to identify groups of similar communities.

Exercises

In the case study, the research team had numerous options along the way, and at each stage the case reports one alternative. As an exercise, you can now explore the sensitivity of conclusions by repeating the process with the following modifications:

- Principal components are intended to be uncorrelated. Use the Multivariate platform to examine the first six PCs and comment on their correlations and scatterplots.
- In the PCA, the two columns Labforce_p201418 and Lab16over_p did not load very heavily into any component. Pop Density 2018 was one of several columns in PC1, but it did not load into any other PC. Rerun the PCA omitting those three columns and use the first six PCs in a hierarchical clustering analysis. Describe how the list of communities like Sharon compares to the list in the case.
- PCA is complicated to explain. Looking at the original six components, identify the six columns that load most heavily into each PC. Rerun the cluster analysis with those six columns, rather than the PCs, and comment on the resulting peer communities cluster.
- In the study, the team used the first six principal components. How would the results change if the clustering used only four or five of the PCs?
- In a public meeting, a citizen asks, "In Sharon, we often share public safety services with our neighboring towns like Foxboro, Stoughton, and Canton. Why have you excluded those towns from your list?" How might you respond? [Hint: look at the Distribution Matrix.]