

K-Means Clustering



Revised: 9/9/2019



Summary	1
Data Input	
Analysis Options	5
Tables and Graphs	7
Analysis Summary	
Membership Table	
Data Table	
Python Script	
2-D Scatterplot	
3-D Scatterplot	
Save Results	
Calculations	
References	

Summary

The *K-Means Clustering* procedure implements a machine-learning process to create groups or clusters of multivariate quantitative variables. Clusters are created by grouping observations which are close together in the space of the input variables.

The calculations are performed by the "scikit-learn" module in Python. To run the procedure, Python must be installed on your computer together with the *scikit-learn* module. For information on downloading and installing Python, refer to the document titled "Python – Installation and Configuration".

Sample StatFolios: kmeans.sgp

© 2019 by Statgraphics Technologies, Inc.



Sample Data

As an example, data describing characteristics of 188 countries during 2008 is contained in the file named *worldbank2008.sgd*. Some of the interesting information in that file is:

- 1) Population density
- 2) Percentage of population in rural areas
- 3) Percentage of population who are female
- 4) Percentage of population who are of working age (age dependency ratio)
- 5) Life expectancy at birth
- 6) Average number of children born to a woman (fertility rate)
- 7) Infant mortality rate
- 8) Percentage of GDP attributable to trade with other countries
- 9) Average GDP per capita
- 10) Consumer price inflation
- 11) Central government debt
- 12) Gross domestic savings

C:\DocE)ata19\worldba	ank2008.sgd							
	Country Code	Country	Year	Population	Pop. Density	Rural Population	Female Percentage	Age Dependency Ratio	Life Expectancy (Total)
				total	people per sq. km of land area	<pre>% of total population</pre>	<pre>% of total population</pre>	<pre>% of working-age population</pre>	years at birth ;
A	Character	Character	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	ABW	Aruba	2008	105526	586.26	53.22	52.61	41.26	74.68
2	AFG	Afghanistan	2008	32517656	49.86	75.96	48.25	97.18	47.54 ·
3	AGO	Angola	2008	18037964	14.47	43.3	50.51	98.35	49.85
4	ALB	Albania	2008	3181397	116.11	53.28	49.95	49.79	76.62
5	ARE	United Arab Emirates	2008	6206623	74.24	22.12	30.62	22.03	76.22
6	ARG	Argentina	2008	39714298	14.51	8	51.08	55.78	75.29
7	ARM	Armenia	2008	3079087	108.11	36.14	53.43	47.38	73.5
8	AUS	Australia	2008	21498500	2.8	11.26	50.25	47.8	81.4
9	AUT	Austria	2008	8336926	101.11	32.84	51.28	47.55	80.23
10	AZE	Azerbaijan	2008	8763400	106.07	48.08	50.66	39.64	70.07
11	BDI	Burundi	2008	7943385	309.32	89.6	51.02	72.22	48.98
12	BEL	Belgium	2008	10709973	353.7	2.64	51.02	51.82	79.48
13	BEN	Benin	2008	8355980	75.54	58.8	50.79	88.89	54.79
14	BFA	Burkina Faso	2008	15515258	56.71	80.44	50.44	91.39	53.98
15	BGD	Bangladesh	2008	145478300	1117.6	72.86	49.18	58.72	68 (
16	BGR	Bulgaria	2008	7623395	70.19	28.9	51.59	44.56	72.96
$ \bullet \bullet$	🕕 worldbank20	108 B C	4						•

As may be seen in the missing data plot shown below, some of the variables (particularly the economic indices) have a large amount of missing values.





We will use the information in this file to group the countries into clusters with similar characteristics.



Data Input

When the *K-Mean Clustering* procedure is selected from the Statgraphics menu, a data input dialog box is displayed. The columns that will be used to group the countries are entered in this dialog box, together with optional point labels are initial seeds:

K-Means Clustering	×
Country Code Country Year Population Pop. Density Rural Population Female Percentage Age Dependency Ratio Life Expectancy (Total) Life Expectancy (Total) Life Expectancy (Male) Female-Male Fertility Rate Infant Mortality Rate Trade	Data: LOG(Pop. Density) Rural Population Female Percentage Age Dependency Ratio Life Expectancy (Total) Female-Male Fertility Rate Infant Mortality Rate
Central Government Debt Consumer Price Inflation Real Interest Rate Unemployment (Total) Unemployment (female) Unemployment (Male) Gross Domestic Savings	(Point Labels:) Country (Initial Seeds:) (Select:)
Sort column names	
OK Cancel	Delete Transform Help

- **Data:** name of one or more numeric columns containing the data used to group the observations.
- Point labels: name of an optional column of any type used as labels for each observation.
- **Initial seeds:** name of an optional column containing row numbers to be used to form the initial cluster seeds. If not specified, the seeds will be selected randomly.
- **Select:** optional Boolean column or expression identifying the cases (rows of the Databook) to be included in the analysis.



Analysis Options

The Analysis Options dialog box sets various options for grouping the data.

K-Means Clustering Options	×					
Number of clusters:	 Standardize variables Verbose output 					
Algorithm	Estimation					
 Automatic 	Number of runs:					
C Classical EM-style	10					
C Elkan	Maximum iterations:					
- Initial seeds	300					
 Smart selection 	Relative tolerance:					
C Random selection	0.0001					
C. From input dialog	Precompute distances					
- Randomization	Automatic					
✓ Fix random seed:	C Always					
30109	C Never					
1	Center the data					
Missing value treatment						
C Exclude incomplete case	s					
 Assign incomplete cases 	to nearest neighbor					
C Replace missing values with column means						
C Replace missing values with column medians						
C Replace missing values with most frequent value						
ОКС	ancel Help					

- Number of clusters: number of groups into which the observations will be divided.
- Algorithm: algorithm used to create the clusters. *Automatic* will cause the program to use the Elkan method for dense data and the classical EM-style method for sparse data.
- **Initial seeds:** method for selecting the initial seeds. *Smart selection* selects initial cluster centers in a manner that speeds up convergence. *Random selection* selects random rows as

the initial seeds. *From input dialog* uses the column of initial seeds specified on the data input dialog box (if entered).

- **Randomization:** if *fix random seed* is checked, seeds the random number generator used by the algorithm. Enter a number between 0 and 32767. This allows you to generate the same results when repeating the analysis at a later time.
- **Standardize variables:** if checked, each variable is standardized by subtracting its mean and dividing by its standard deviation before distances between clusters are calculated. This is important if the variables are measured in different units,
- Verbose output: if checked, the output will include information about each intermediate step.
- **Number of runs:** number of times the algorithm will be run with different seeds. The final fit will be the run which gives the best results.
- **Maximum iterations:** maximum number of iterations performed during each run of the algorithm.
- **Relative tolerance:** relative tolerance with respect to inertia used by the algorithm to declare convergence.
- **Precompute distances:** whether to calculate and store distances before starting the algorithm. *Automatic* precomputes distances if the number of samples multiplied by the number of clusters does not exceed 12 million. Center the data controls whether the data are centered before precomputing the distance which is more accurate but may introduce small numerical differences which could cause a significant slowdown in the algorithm.
- **Missing value treatment:** how missing values in the data should be handled. The choices are:
 - *Exclude incomplete cases*: any cases with missing values in any column are excluded from the analysis.
 - Assign incomplete cases to nearest neighbor: runs the algorithm with only complete data. Cases with missing data are then assigned to the cluster whose centroid is the closest based on the non-missing data.
 - *Replace missing values with column means*: missing values are imputed by replacing them with its column mean.
 - *Replace missing values with column medians*: missing values are imputed by replacing them with its column median.
 - *Replace missing values with most frequent values*: missing values are imputed by replacing them with the most frequent value in its column.

For more information on the various options, go to <u>https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html</u>. © 2019 by Statgraphics Technologies, Inc.



Tables and Graphs

The following tables and graphs may be created:

Tables and Graphs		×
TABLES Analysis Summary	GRAPHS 2D Scatterplot	ОК
Membership Table	✓ 3D Scatterplot	Cancel
🔽 Data Table		AII
✓ Python Script		Store
		Help



Analysis Summary

The Analysis Summary summarizes the input data and the final results:

K-Means Clustering Data variables: LOG(Pop. Density) Rural Population (% of total population) Female Percentage (% of total population) Age Dependency Ratio (% of working-age population) Life Expectancy (Total) (years at birth) Female-Male (Life Expectancy (Female) - Life Expectancy (Male)) Fertility Rate (births per woman) Infant Mortality Rate (per 1,000 live births)

Number of complete cases: 178 Number of partially missing cases: 10 Partially missing data assigned to nearest cluster centroid.

Number of iterations: 3 Final inertia: 643.533 Variance explained: 54.5528% Calinski-Harabasz index: 105.031

Cluster Summary

Cluster	Members	Percent
1	123	69.10
2	61	34.27
3	4	2.25

Cluster Centroids

Cluster	LOG(Pop. Density)	Rural Population	Female Percentage	Age Dependency Ratio
1	4.25199	36.5002	50.5288	51.1776
2	3.82481	63.8003	50.0257	80.5561
3	5.32367	9.89998	33.9425	28.2225

Cluster	Life Expectancy (Total)	Female-Male	Fertility Rate	Infant Mortality Rate
1	74.5026	5.84195	2.10009	14.8027
2	56.3795	2.17066	4.62836	67.8377
3	75.7825	1.0425	2.2825	8.2

Standardized Cluster Centroids

Cluster	LOG(Pop. Density)	Rural Population	Female Percentage	Age Dependency Ratio
1	0.0887537	-0.383554	0.187263	-0.526274



2	-0.221226	0.81206	1	0.0144	58	1.09234
3	0.866409	-1.54852		-5.510	68	-1.791
Cluster	Life Expectancy (Total)		Female-Ma	le	Fertility Rate	Infant Mortality Rate
1	0.608413		0.527063		-0.588029	-0.598242
2	-1.17522		-0.889492		1.11978	1.16182
3	0.734381		-1.32479		-0.464813	-0.817363

Of particular interest are:

- 1. **Number of complete cases:** number of cases with no missing data for any of the variables specified. Any incomplete cases are removed before the algorithm is performed.
- 2. **Number of iterations:** number of iterations performed by the algorithm before convergence. If this number is less than the maximum number of iterations specified on the *Analysis Options* dialog box, the algorithm has converged to a final solution.
- 3. **Final inertia:** sum of squared distances of the samples to their closest cluster center. This number may be compared with the result of selecting a different number of clusters.
- 4. **Variance explained:** percentage reduction in the sum of squared distances compared to a single cluster.
- 5. Calinski-Harabasz index: comparison of the between cluster sum of squares to the within cluster sum of squares. Larger values of the index are preferable and is often used to help determine the proper number of clusters.
- 6. Cluster summary: number and percentage of cases assigned to each cluster.
- 7. Cluster centroids: the mean value of each variable for the members of each cluster.
- 8. **Standardized cluster centroids:** the mean value of each variable for the members of each cluster in standardized units.

In the sample data, the clusters have 123, 61 and 4 members, respectively.

Membership Table

This table shows the members of each cluster. A portion of the table is shown below:

Memb	ership Table			
Cluste	r 1 (123 members)			
Row	Country	Pop. Density	Rural Population	Female Percentage
1*	Aruba	6.37376	53.22	52.61
4	Albania	4.75454	53.28	49.95
6	Argentina	2.67484	8	51.08
7	Armenia	4.68315	36.14	53.43
8	Australia	1.02962	11.26	50.25
9	Austria	4.61621	32.84	51.28
10	Azerbaijan	4.6641	48.08	50.66
12	Belgium	5.86845	2.64	51.02
16	Bulgaria	4.25121	28.9	51.59
Cluste	r 2 (61 members)			
Row	Country	Pop. Density	Rural Population	Female Percentage
2	Afghanistan	3.90922	75.96	48.25
3	Angola	2.67208	43.3	50.51
11	Burundi	5.73438	89.6	51.02
13	Benin	4.32466	58.8	50.79
14	Burkina Faso	4.03795	80.44	50.44
15	Bangladesh	7.01894	72.86	49.18
22	Bolivia	2.1838	34.42	50.18
27	Botswana	1.23837	40.42	49.7
28	Central African Republic	1.91692	61.42	50.76
Cluste	r 3 (4 members)			
Row	Country	Pop. Density	Rural Population	Female Percentage
5	United Arab Emirates	4.3073	22.12	30.62
17	Bahrain	7.23322	11.48	38.87
93	Kuwait	4.96291	1.64	40.14
4.40	Oatar	1 70123	1 36	26.14

Notice the asterisk next to Aruba in the table for cluster #1. Aruba was not used to create the clusters since it had missing data on at least one variable. However, it was assigned to that cluster after the algorithm finished since it was closer to the centroid for that cluster than to any of the other clusters.



Pane Options

Membership Table Options	×
Include data variables	ОК
	Cancel
	Help

Check the box to include the data as well as the row numbers and identifiers in the table.



Data Table

This table shows the cluster numbers in row order:

Row	Country	Cluster	LOG(Pop. Density)	Rural Population
1	Aruba	1*	6.37376	53.22
2	Afghanistan	2	3.90922	75.96
3	Angola	2	2.67208	43.3
4	Albania	1	4.75454	53.28
5	United Arab Emirates	3	4.3073	22.12
6	Argentina	1	2.67484	8
7	Armenia	1	4.68315	36.14
8	Australia	1	1.02962	11.26
9	Austria	1	4.61621	32.84
10	Azerbaijan	1	4.6641	48.08
11	Burundi	2	5.73438	89.6
12	Belgium	1	5.86845	2.64

*Partially missing data assigned to nearest cluster centroid.

Depending on the method used to treat missing data, some rows may not have been assigned to a cluster.

Pane Options

Data Table Options	×
🔽 Include data variables	ОК
	Cancel
	Help

Check the box to include the data as well as the row numbers and identifiers in the table.



Python Script

This table shows the script that was executed by Python:

from sklearn.cluster import KMeans #Read data X=pandas.read_csv(r'C:\\Users\\NEIL~1.STA\\AppData\\Local\\Temp\\\statgraphics_data.csv') X=X.replace(-32768,numpy.NaN) #Perform cluster analysis kmeans = KMeans(n clusters=3,verbose=0,algorithm='auto',init='kmeans++',random_state=30109,precompute_distances='auto',n_init=10,max_iter=300,tol=0.0001) kmeans.fit(X) #Display cluster centers print(kmeans.cluster centers) [[0.08875374 -0.38355427 0.18726345 -0.52627382 0.60841295 0.5270635 -0.58802943 -0.5982424] [-0.22122597 0.81206076 0.01445796 1.09234284 -1.17521575 -0.88949234 1.1197805 1.16181667] 0.86640925 -1.548515 -5.5106775 -1.790995 0.734381 -1.3247875 -0.46481325 -0.81736275]] numpy.savetxt('C:\Users\NEIL~1.STA\AppData\Local\Temp\\statgraphics centers.csv',kmeans.cluster centers _,delimiter=',') #Display cluster numbers print(kmeans.labels_) [1102000001011102000010000110000111010 000001000010010001100111110000001001 000000000101021010001000001000100001 101011100010010001000200101101101100 001011010100001100000001011111 numpy.savetxt('C:\Users\NEIL~1.STA\AppData\Local\Temp\\statgraphics clusters.csv',kmeans.predict(X)) #Display final inertia print(kmeans.inertia_) 643.5326413615171 #Display number of iterations print(kmeans.n_iter_) 3 **#Save results** output=[kmeans.inertia_,kmeans.n_iter_] numpy.savetxt('C:\Users\NEIL~1.STA\AppData\Local\Temp\\statgraphics_inertia.csv',output,delimiter=',')



2-D Scatterplot

This plot shows the sample data with respect to any 2 variables. Different point symbols are used to identify the clusters:





Clusters 1 and 2 are well separated in the space of *Life Expectancy* and *Fertility Rate*. However, cluster #3 lies totally within cluster #2 for those 2 variables.

Pane Options

This dialog box specifies the variables to be placed on the X and Y axes:

2D Scatterplot Options	×					
X-Axis:						
LOG(Pop. Density) Rural Population Female Percentage						
Life Expectancy (Total)						
Female-Male						
Fertility Rate						
Infant Mortality Hate						
Y-Axis:	_					
LOG(Pop. Density) Rural Population Female Percentage Age Dependency Ratio						
Life Expectancy [1 otal]						
Fertility Bate						
Infant Mortality Rate						
Display centroids						
✓ Draw ellipses around clusters						
OK Cancel Help						

- X-axis: select the variable to plot along the X-axis.
- **Y-axis:** select the variable to plot along the Y-axis.
- **Display centroids:** if checked, the location of the cluster centroids are shown using small plus signs.
- **Draw ellipses around clusters:** if checked, ellipses will be drawn containing the points in each cluster, centered at the cluster centroids with its major axis in the direction of the first principal component.



3-D Scatterplot

This plot shows the sample data with respect to any 3 variables. Different point symbols are used to identify the clusters:



Cluster 3 is separated from the other clusters because of a much lower percentage of females in the population.

Pane Options

This dialog box specifies the variables to be placed on the X, Y and Z axes:

3D Scatterplot Options	×
X-Axis: LOG(Pop. Density) Rural Population Female Percentage Age Dependency Ratio Life Expectancy (Total) Female-Male Fertility Rate	^
Y-Axis: LOG(Pop. Density) Rural Population Female Percentage Age Dependency Ratio Life Expectancy (Total) Female-Male Fertility Rate	*
Z-Axis: LOG(Pop. Density) Rural Population Female Percentage Age Dependency Ratio Life Expectancy (Total) Female-Male Fertility Rate	< >
OK Cancel Help	

- **X-axis:** select the variable to plot along the X-axis.
- **Y-axis:** select the variable to plot along the Y-axis.
- **Z-axis:** select the variable to plot along the Z-axis.
- **Display centroids:** if checked, the location of the cluster centroids are shown using small plus signs.



Save Results

The cluster numbers and cluster centroids may be saved to the Statgraphics DataBook:

Save Results Options		×
Save Save Cluster assignments Cluster centroids	Target Variables CLUSTER CENTROID	Cancel Cancel Help Datasheet C A C N C B C D C C C P C D C Q C E C R C F C S C G C T C H C U C I C V C J C W C K C X C L C Y
Autosave	Save comments	OMOZ

To save results, select:

- **Save:** select the items to be saved.
- **Target Variables:** enter names for the columns to be created.
- **Datasheet:** the datasheet into which the results will be saved.
- Autosave: if checked, the results will be saved automatically each time a saved StatFolio is loaded.
- **Save comments:** if checked, comments for each column will be saved in the second line of the datasheet header.



Calculations

The assignment of observations to clusters are performed by the scikit-learn module in Python. Other calculations are done in Statgraphics including those below.

Let

X = multivariate observation in a single row

 C_i = centroid of cluster *i*

C = centroid of all observations used in the calculations

 n_i = number of observations assigned to cluster i

n = total number of observations used in the calculations

c = number of clusters

 $d^{2}(C_{i},C)$ = squared distance between centroid of cluster *i* and centroid of all the data

 $d^{2}(X,C_{i})$ = squared distance between a single observation X and the centroid of cluster *i*

Calinski-Harabasz index

$$CH = \frac{\sum_{i=1}^{c} n_i d^2(C_i, C) / (c-1)}{\sum_{i=1}^{c} \sum_{X \in C_i} d^2(X, C_i) / (n-c)}$$

References

Liu, Y., Li, Z., Xiong, H., Gao, X. and Wu, J. (2010) "Understanding of Internal Clustering Validation Measures", 2010 IEEE International Conference on Data Mining, 911-916.

Sklearn.cluster.kmeans, scikit-learn.org https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html