

IBM Software

Clustering Course

Lab 4: Getting familiar with K-means clustering method

Contents

Getting familiar with K-means clustering method

1. Getting familiar with dataset
2. Data preparation
3. Applying K-means clustering

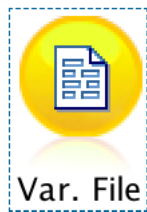
Summary

Hello everybody! In this lab, we are going to get familiar with K-means clustering algorithm which is one of the most popular algorithms in machine learning.

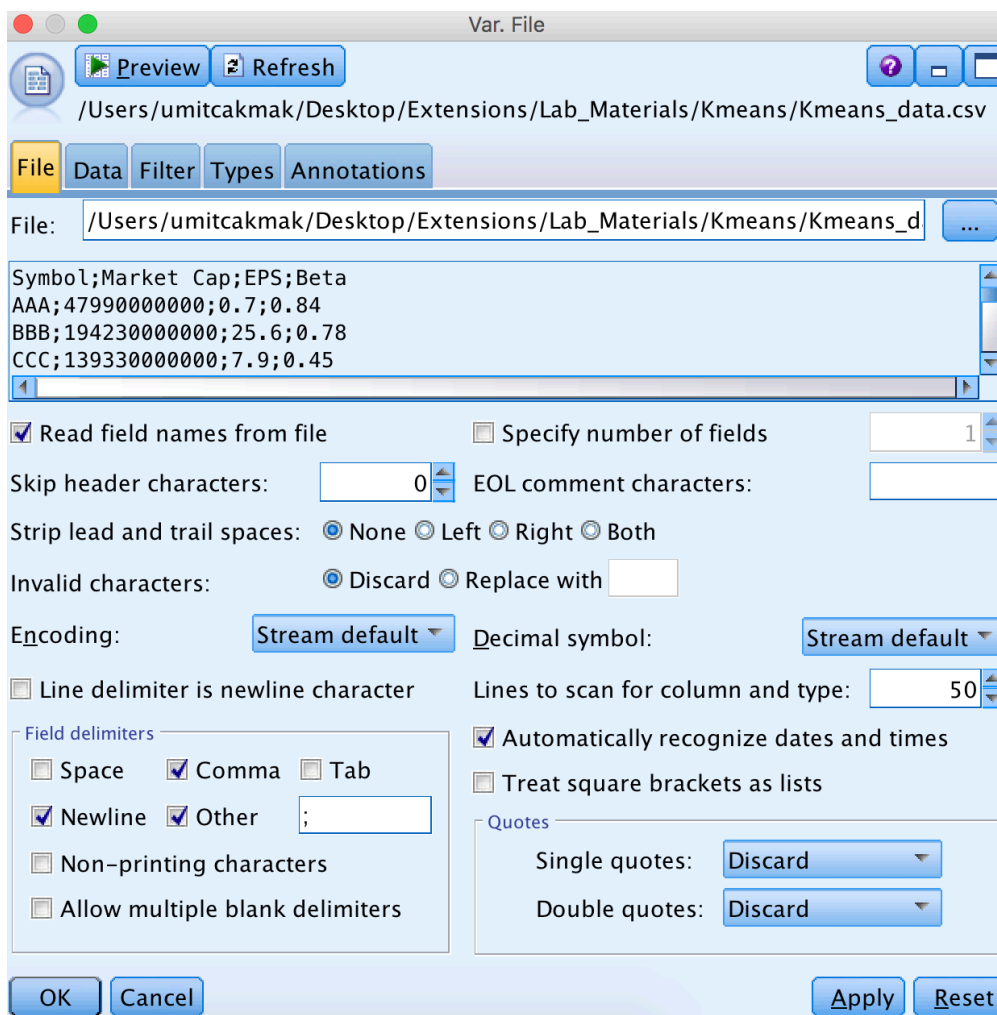
Let's get started.

1. Getting familiar with dataset

We need to import "Kmeans_data.csv". Add a "Var. File" node from "Sources" palette.



Browse to "Kmeans_data.csv" file, check the box "Others" and type semicolon as shown below and click OK.



Var. File

Preview Refresh

/Users/umitcakmak/Desktop/Extensions/Lab_Materials/Kmeans/Kmeans_data.csv

File Data Filter Types Annotations

File: /Users/umitcakmak/Desktop/Extensions/Lab_Materials/Kmeans/Kmeans_d ...

Symbol;Market Cap;EPS;Beta
AAA;4799000000;0.7;0.84
BBB;19423000000;25.6;0.78
CCC;13933000000;7.9;0.45

☒ Read field names from file ☐ Specify number of fields 1

Skip header characters: 0 EOL comment characters:

Strip lead and trail spaces: ☒ None ☐ Left ☐ Right ☐ Both

Invalid characters: ☒ Discard ☐ Replace with

Encoding: Stream default Decimal symbol: Stream default

☐ Line delimiter is newline character Lines to scan for column and type: 50

Field delimiters

☐ Space ☒ Comma ☐ Tab

☒ Newline ☒ Other ;

☐ Non-printing characters

☐ Allow multiple blank delimiters

☒ Automatically recognize dates and times

☐ Treat square brackets as lists

Quotes

Single quotes: Discard

Double quotes: Discard

OK Cancel Apply Reset

You can add “Table” node from “Output” palette to see all records.

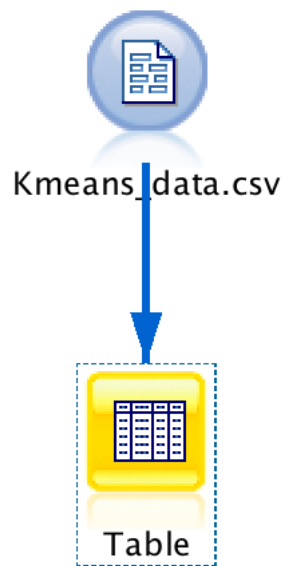


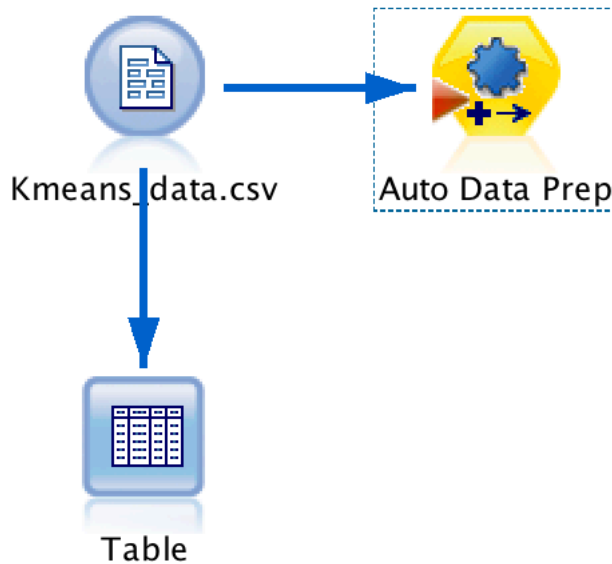
Table (4 fields, 25 records)

	Symbol	Market Cap	EPS	Beta
1	AAA	47990000000	0.700	0.840
2	BBB	194230000000	25.600	0.780
3	CCC	139330000000	7.900	0.450
4	DDD	50270000000	3.200	0.820
5	EEE	99800000000	1.100	0.860
6	FFF	198540000000	22.300	0.690
7	GGG	150560000000	11.300	0.520
8	HHH	212300000000	22.100	0.760
9	III	124990000000	7.600	0.550
10	JJJ	48570000000	3.700	0.780
11	KKK	97200000000	2.300	0.820
12	LLL	196250000000	26.020	0.810
13	MMM	45110000000	1.300	0.770
14	NNN	189120000000	24.800	0.720
15	OOO	46550000000	1.000	0.800
16	PPP	138750000000	7.500	0.620
17	QQQ	145980000000	9.000	0.480
18	RRR	89000000000	4.100	0.790
19	SSS	143540000000	8.700	0.490
20	TTT	187145000000	23.400	0.800

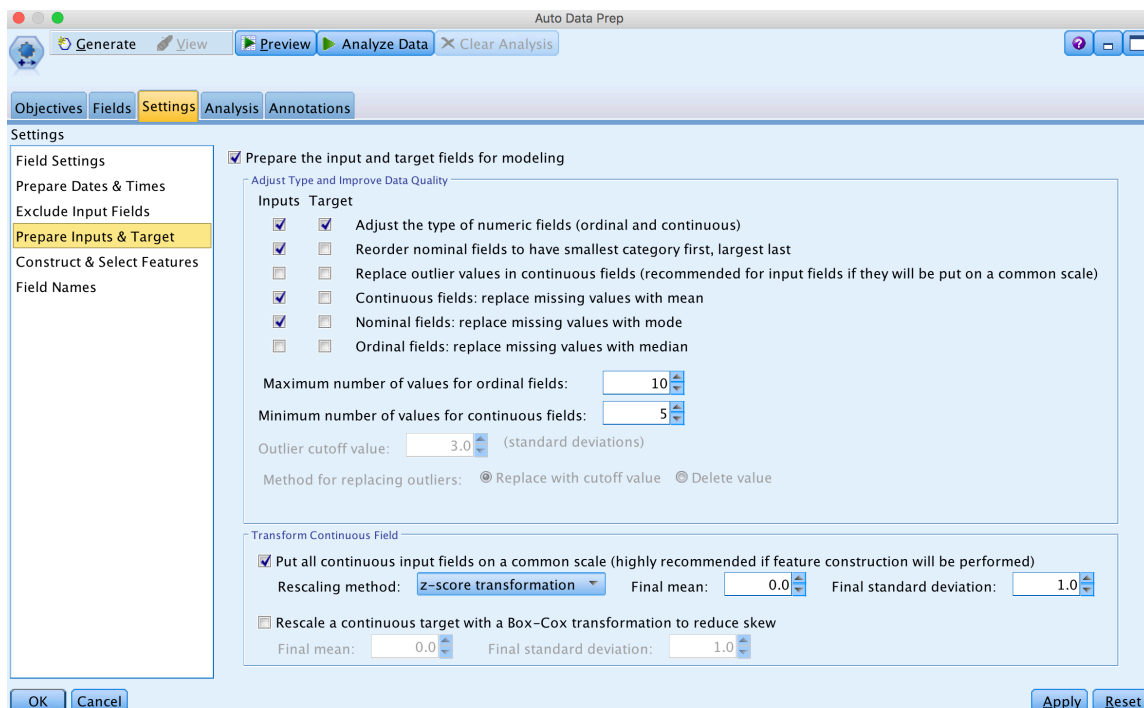
OK

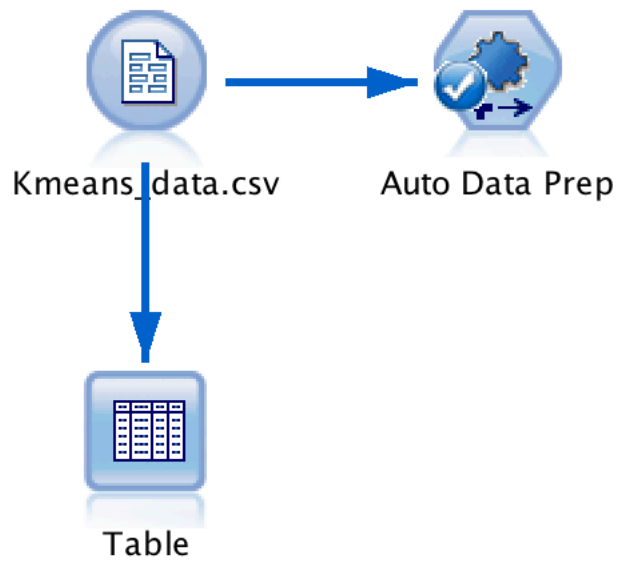
2. Data preparation

We will add “Auto Data Prep” node to normalize our data set so that large values in scale compared to other will not distort the results.

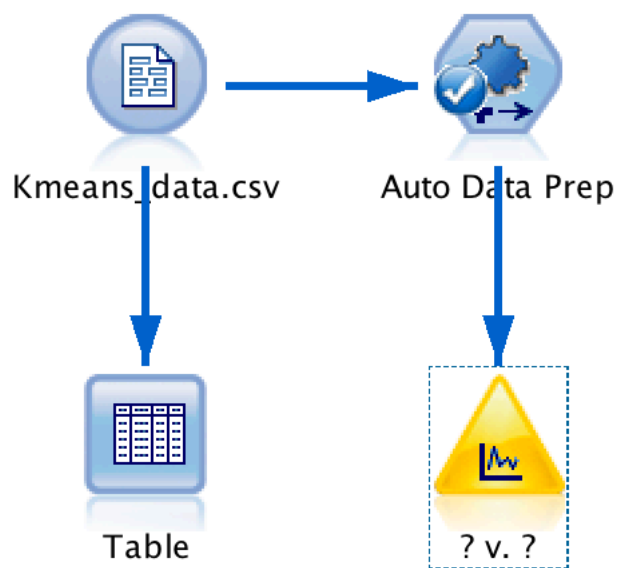


Open “Auto Data Prep” node and in “Settings” tab, “Prepare Inputs & Target” section will allow you to transform all records. Default settings are already set properly for normalization, we can click “Analyze Data” on top to perform normalization.

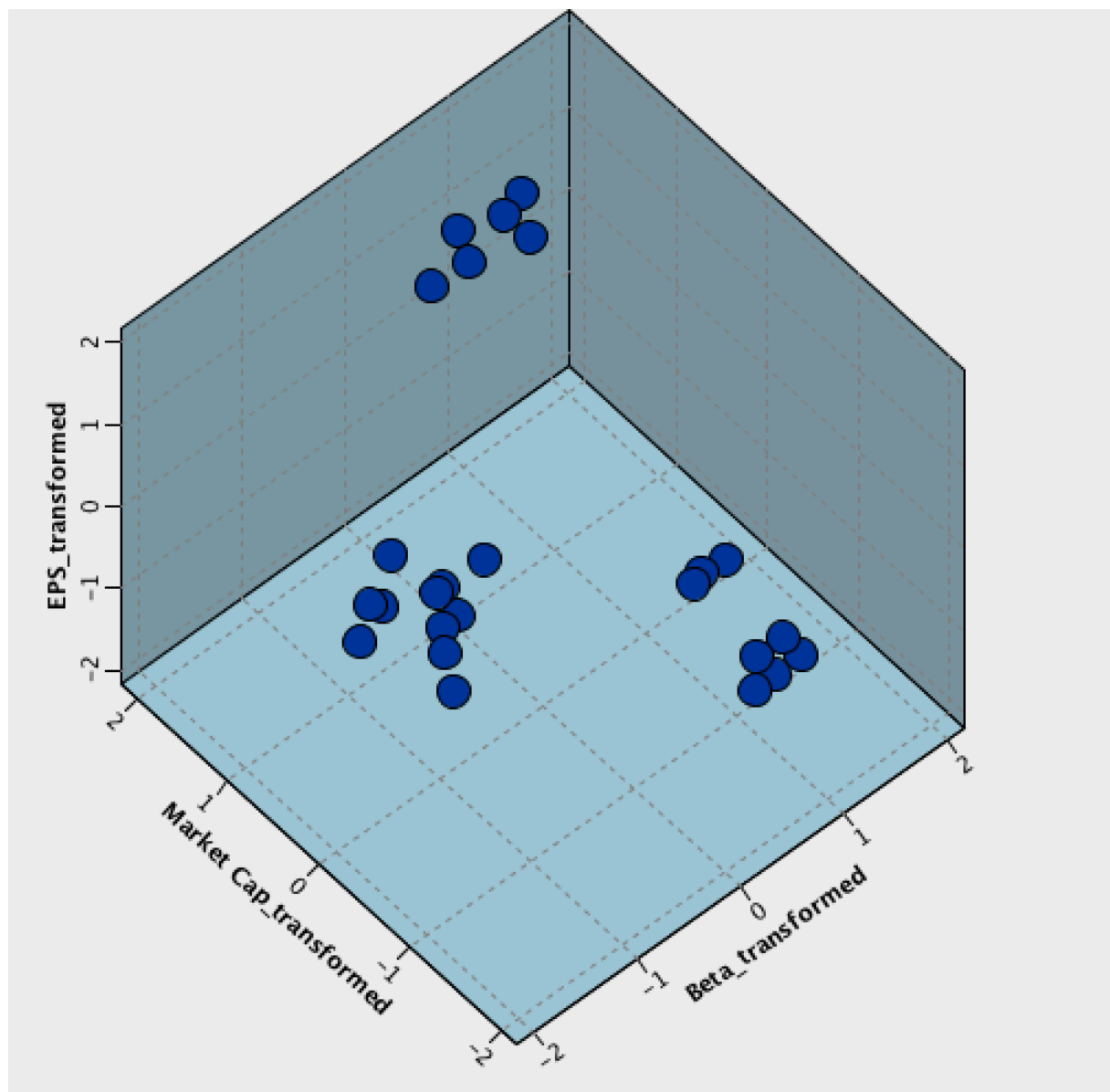
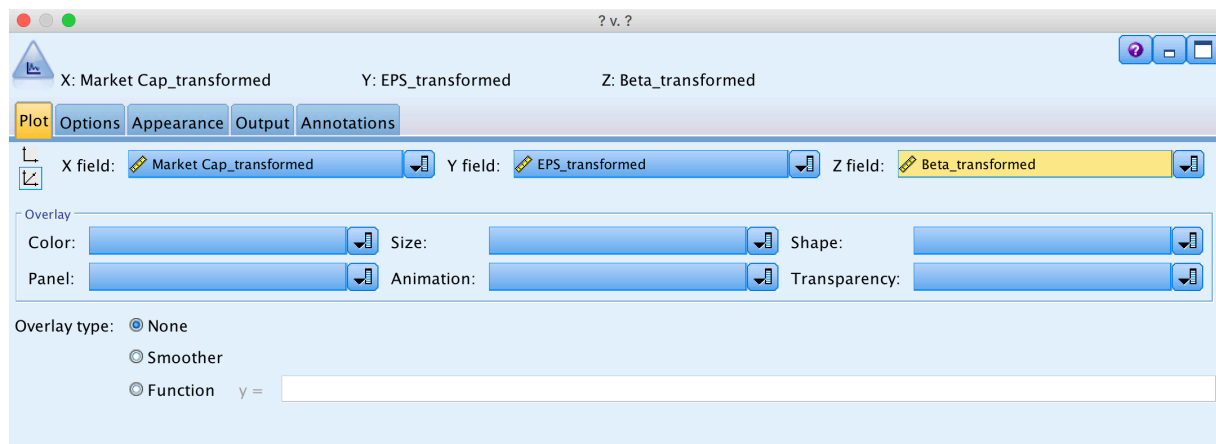




Let's add a "Plot" node from "Graphs" palette to see what our data looks like



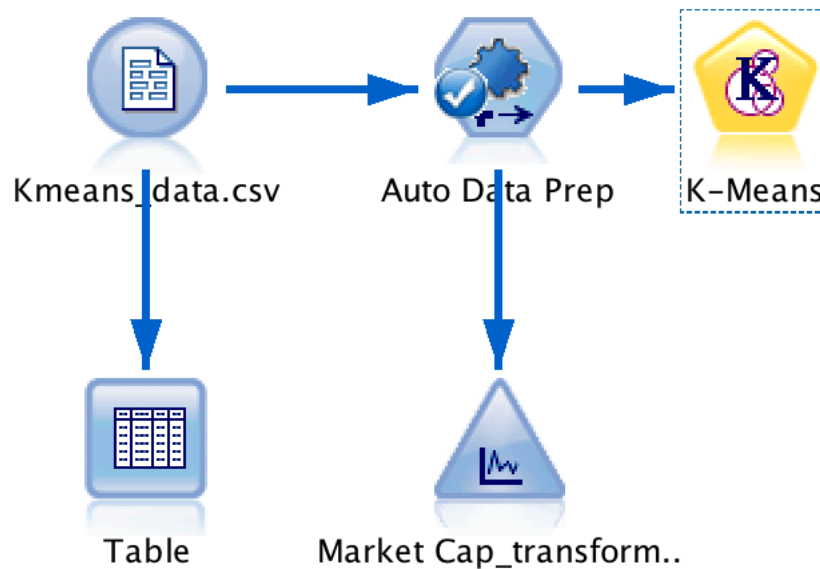
Adjust settings as shown below and click “Run” to see 3D-plot



We can see 3 different groupings on this chart, and next step is to apply K-means to group these records.

3. Applying K-means clustering

We will add “Kmeans” node from “Modeling” palette.



We will set “Number of clusters” to 3 and “Run” the model.

K-Means

Fields Model Expert Annotations

Model name: ☒ Auto ☐ Custom

☒ Use partitioned data

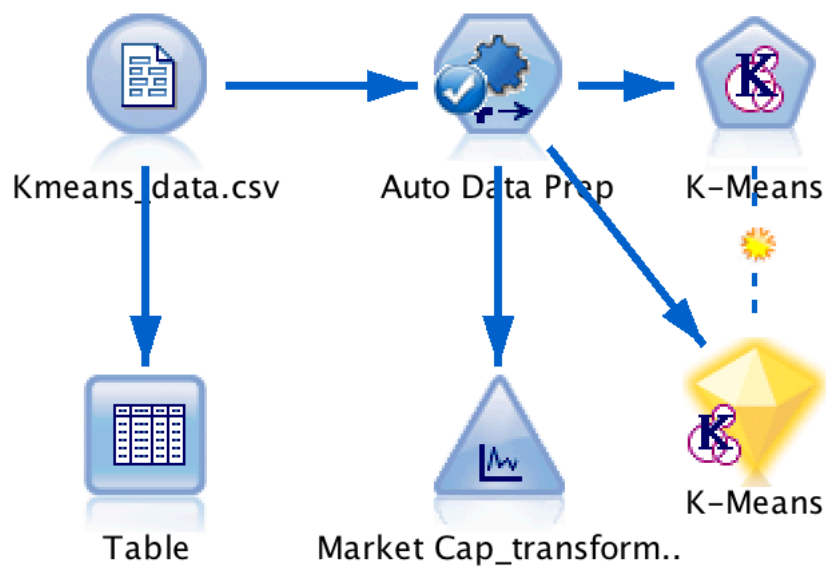
Number of clusters:

☐ Generate distance field

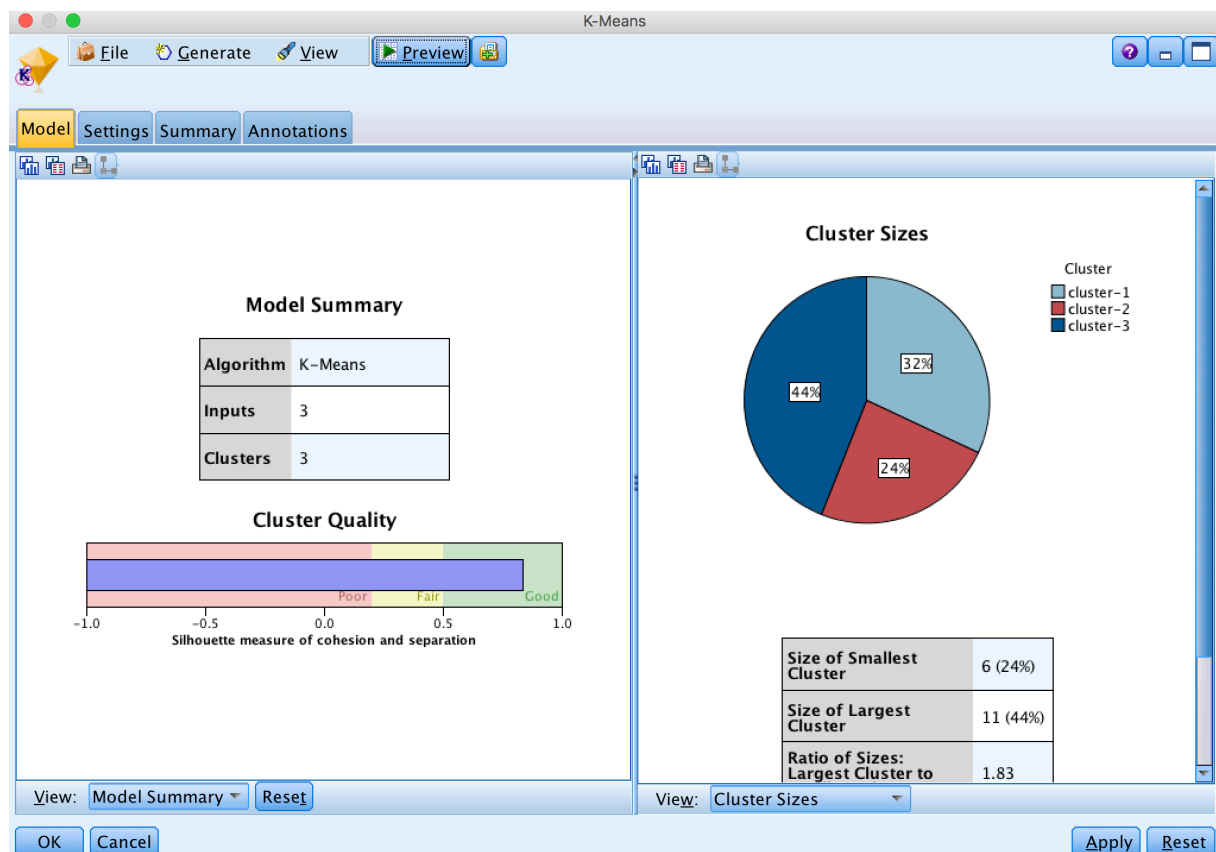
Cluster label: ☒ String ☐ Number

Label prefix:

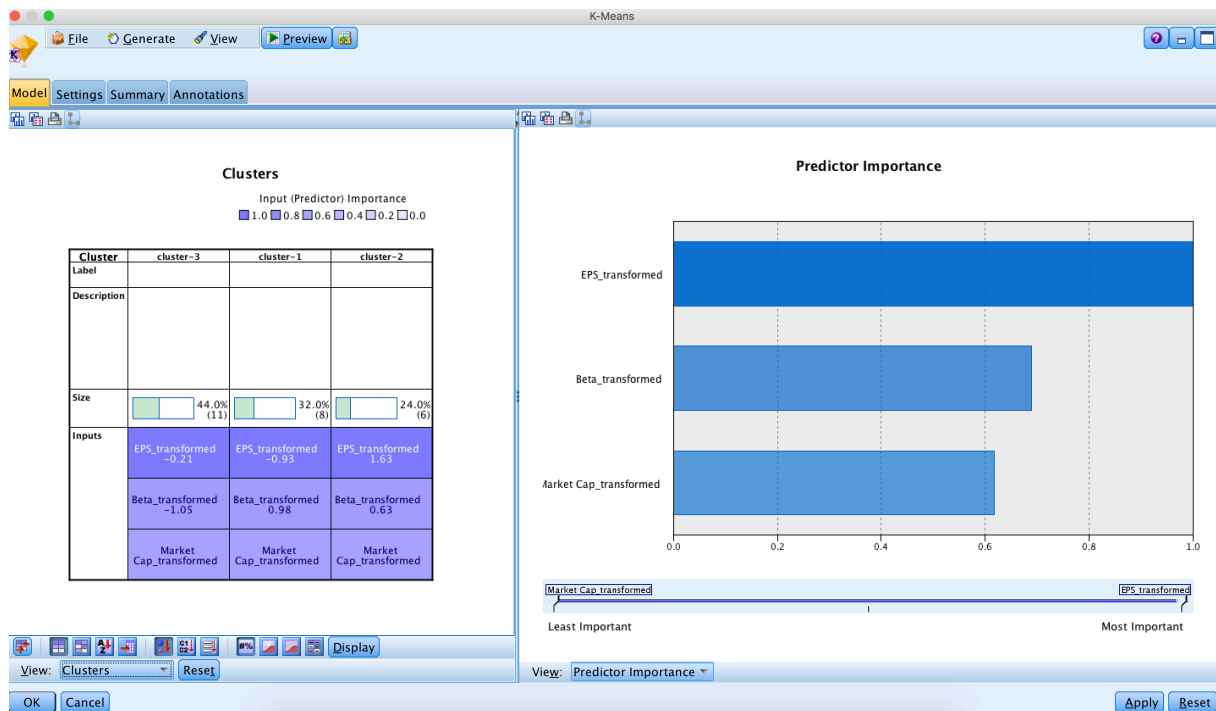
Optimize: ☐ Speed ☒ Memory



Once we have model nugget, we can double-click on it to open model summary



We can see model summary and various measures to assess the performance of clustering method.



We can also add “Table” from “Output” node to see group memberships of individual records.

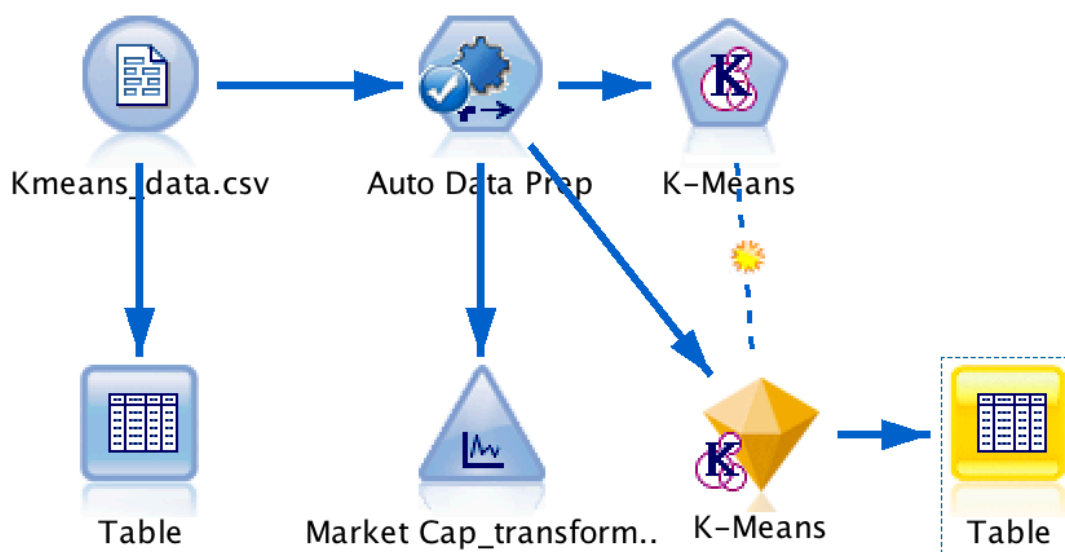
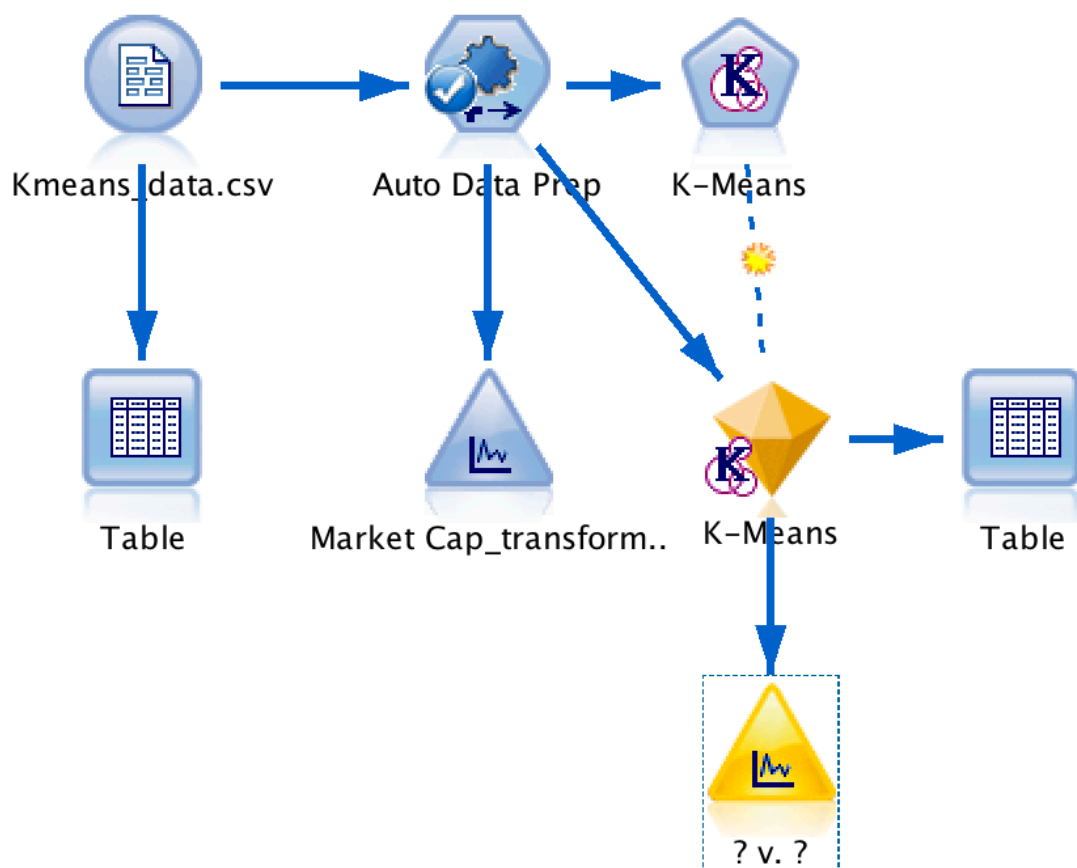
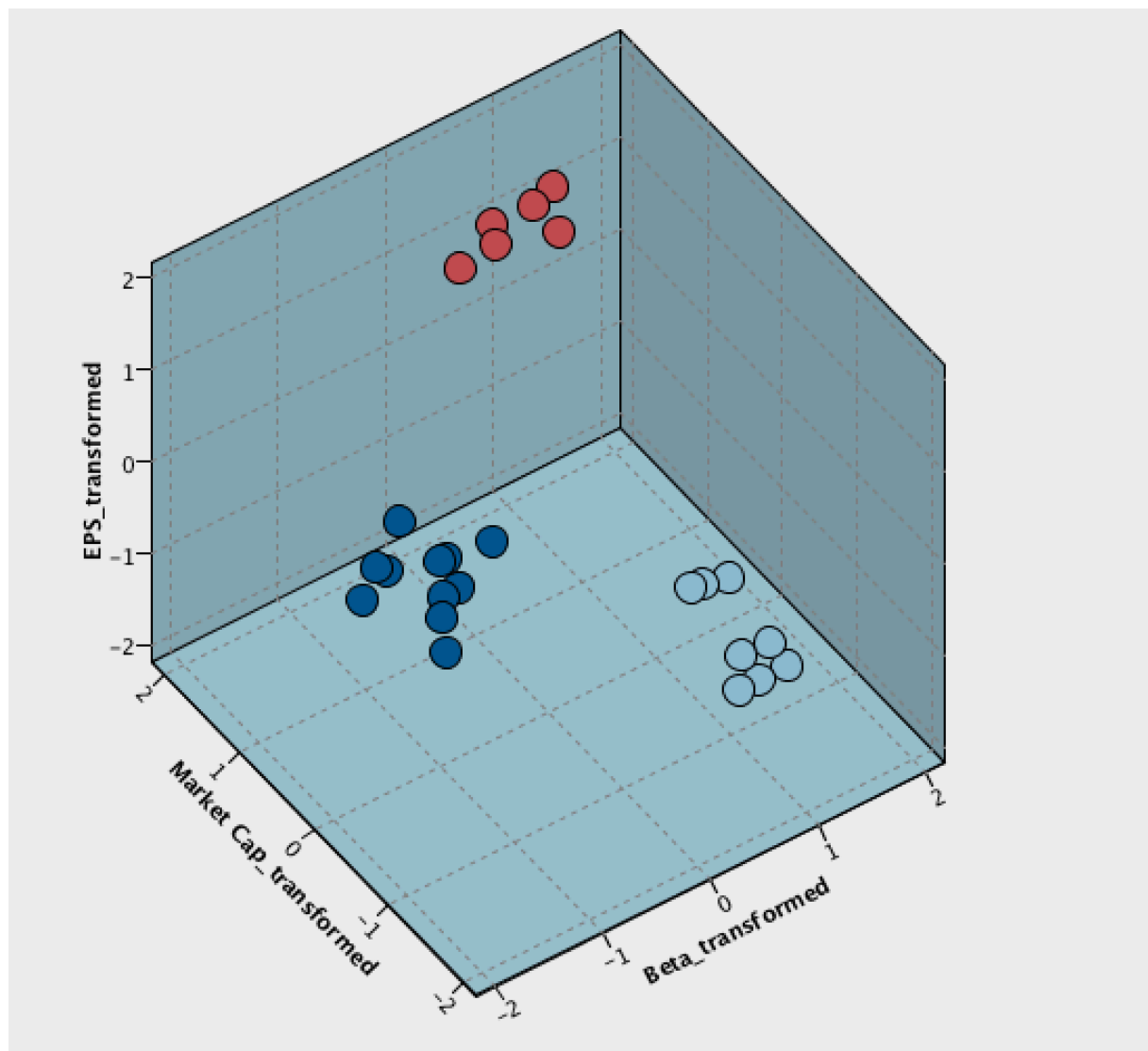
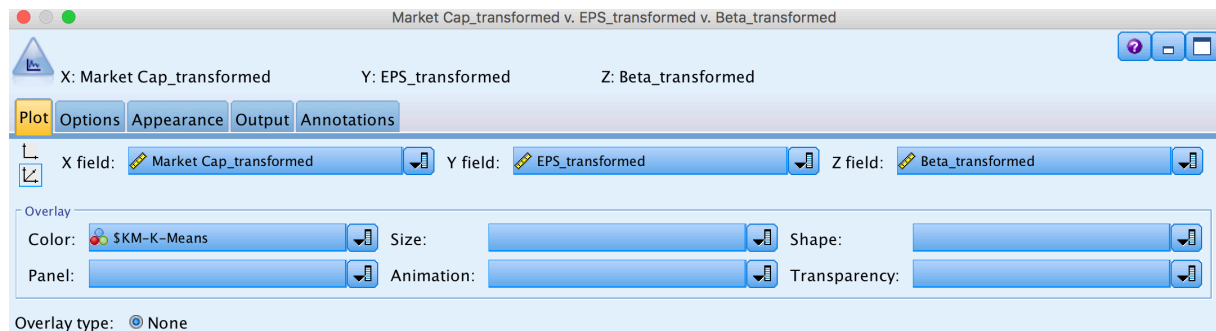


Table	Annotations				
	Market Cap_transformed	EPS_transformed	Beta_transformed	Symbol_transformed	\$KM-K-Means
1	-1.492	-1.107	1.186 00		cluster-1
2	1.297	1.813	0.766 01		cluster-2
3	0.250	-0.262	-1.548 02		cluster-3
4	-1.448	-0.813	1.046 03		cluster-1
5	-0.504	-1.060	1.327 04		cluster-1
6	1.379	1.426	0.135 05		cluster-2
7	0.464	0.136	-1.057 06		cluster-3
8	1.642	1.403	0.625 07		cluster-2
9	-0.023	-0.297	-0.847 08		cluster-3
10	-1.480	-0.755	0.766 09		cluster-1
11	-0.553	-0.919	1.046 10		cluster-1
12	1.336	1.863	0.976 11		cluster-2
13	-1.546	-1.036	0.696 12		cluster-1
14	1.200	1.719	0.345 13		cluster-2
15	-1.519	-1.071	0.906 14		cluster-1
16	0.239	-0.309	-0.356 15		cluster-3
17	0.377	-0.133	-1.338 16		cluster-3
18	-0.710	-0.708	0.836 17		cluster-1
19	0.330	-0.168	-1.268 18		cluster-3
20	1.162	1.555	0.906 19		cluster-2
21	0.144	-0.122	-0.847 20		cluster-3
22	0.127	-0.098	-0.917 21		cluster-3
23	-0.015	-0.333	-0.987 22		cluster-3
24	-0.456	-0.403	-1.268 23		cluster-3
25	-0.201	-0.321	-1.127 24		cluster-3

You can also visualize memberships by add another “Plot” node to model nugget.



Adjust settings as shown below and click “Run” to see chart



Summary

In this lab, you have learned how you can apply Kmeans clustering methods and how to check model summary to see model performance.

Thank you and see you in next lab.