

## 7-2. 계층적 군집분석(실습)

김 성 기

# 목 차

## 1. 계층적 군집 분석(Hierarchical Clustering)

: KOSPI 지수에 영향을 끼치는 여러 경제 지표들에 대하여  
월별로 유사한 군집으로 분석

# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

- **Spotfire** 에서 기본적으로 제공하는 분석 도구를 활용하여 **Kopsi** 관련 경제지표 데이터로부터, **KOSPI**에 영향을 끼치는 중요한 인자들을 찾고, 그 인자들을 중심으로 유사한 특성을 갖는 그룹(기간들)을 찾는 분석을 수행 해 보자.

## 1. Spotfire에 KOSPI data.xls 파일을 load한다.

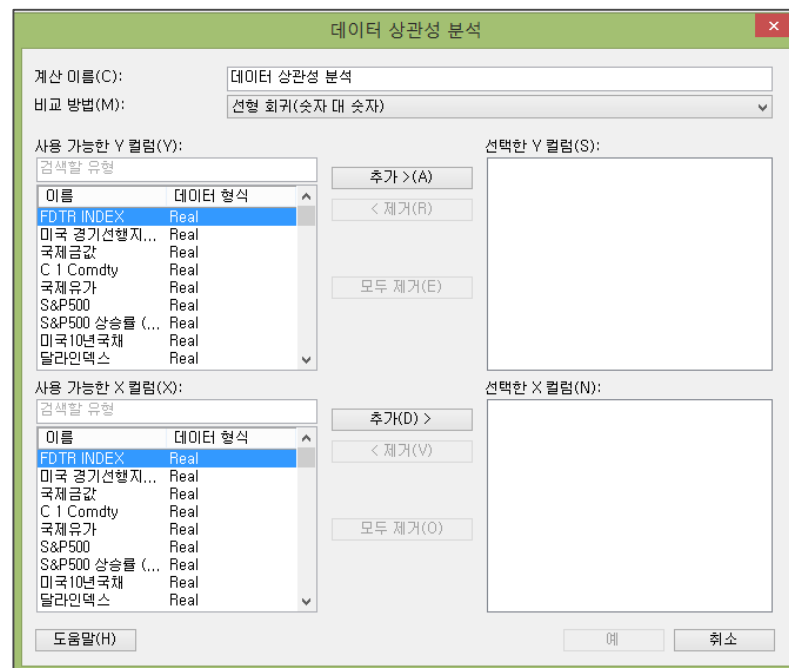
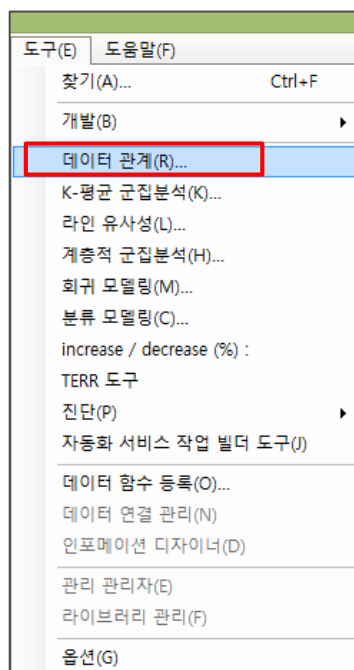
The screenshot shows the Spotfire software interface with the 'KOSPI Data - TIBCO Spotfire' window. The data table is displayed with the following columns:

날짜	FOTR INDEX	한국 경제지표...	국제금융...	C 1 Comodity	국제통화...	S&P500	S&P500 상승...	한국10년국채	달러환율...	Kospi index	KOSPI 상승...	한국부동산...
1990-01-31	8.25	8.98	412.44	238.25	22.68	329.88	8.42	92.55	898.16			
1990-02-28	8.25	8.93	408.99	247.00	21.54	331.89	8.52	93.39	891.59			
1990-03-30	8.25	1.15	379.28	261.25	28.28	339.94	8.63	93.77	849.89	5.0		
1990-04-30	8.25	1.23	371.48	281.25	18.54	338.80	8.92	93.68	688.86	6.0		
1990-05-31	8.25	1.28	364.30	278.00	17.40	361.23	8.80	92.93	797.95	5.0		
1990-06-29	8.25	8.90	359.90	295.50	17.87	358.82	8.41	91.15	708.79	3.7		
1990-07-31	8.00	8.18	370.08	288.25	29.69	356.15	8.34	87.20	678.38	3.7		
1990-08-31	8.00	-8.81	384.60	243.25	27.32	322.56	8.85	86.58	606.87	3.7		
1990-09-28	8.00	-1.91	402.50	228.00		308.05	8.80	85.91	602.88	1.8		
1990-10-31	7.75	-2.93	381.30	229.25		304.80	8.62	83.27	699.16	1.8		
1990-11-30	7.50	-3.71	383.30	227.75	29.85	322.22	8.25	82.84	687.83	1.8		
1990-12-31	7.50	-4.15	386.20	231.75	28.44	330.22		83.12	696.11	0.6		
1991-01-31	8.75	-4.23	386.80	244.25	21.54	343.93	8.81	82.86	835.40	-29.10	0.6	
1991-02-28	6.25	-3.87	369.00	241.25	19.16	367.87	10.60	8.93	84.07	675.57	-21.59	0.6
1991-03-29	6.00	-3.16	357.18	252.75	19.83	375.22	10.36	8.96	92.91	699.85	-21.53	1.0
1991-04-30	5.75	-2.17	357.79	246.50	28.96	375.34	13.46	8.81	91.54	645.81	-4.25	1.0
1991-05-31	5.75	-9.96	361.40	245.75	21.13	389.83	7.82	8.86	93.15	611.35	-23.38	1.0
1991-06-28	5.75	8.42	370.08	233.00	29.56	371.16	3.87	8.23	98.63	605.27	-14.38	1.5
1991-07-31	5.75	1.87	363.60	258.00	21.68	387.81	8.89	8.15	93.49	717.83	5.70	1.5
1991-08-30	5.50	3.25	349.80	249.00	22.28	395.43	22.59	7.82	93.49	683.11	12.56	1.5
1991-09-28	5.25	4.47	364.60	249.25	22.23	387.86	26.73	7.48	89.80	705.88	16.95	1.1
1991-10-31	5.00	5.42	359.50	251.00	23.37	382.45	29.10	7.46	89.94	695.94	0.64	1.1
1991-11-29	4.75	6.07	367.80	239.75	21.48	375.22	18.45	7.38	88.19	682.11	-6.44	1.1
1991-12-31	4.50	6.45	355.20	251.50	19.12	417.89	26.31	8.70	83.53	610.82	-12.24	2.0
1992-01-31	4.00	6.60	357.00	264.25	18.90	408.79	18.86	7.27	87.49	680.51	-7.10	2.0
1992-02-28	4.00	8.48	354.10	264.75	18.68	412.79	12.43	7.25	88.89	612.59	-8.34	2.0
1992-03-31	4.00	8.14	343.70	264.25	18.44	403.89	7.59	7.53	88.65	606.32	-8.11	2.0
1992-04-30	3.75	1.61	337.80	244.00	29.85	414.95	10.55	7.58	89.62	615.87	-4.59	2.0
1992-05-29	3.75	4.65	336.40	259.50	22.11	418.56	6.85	7.32	87.22	574.28	-6.88	2.0
1992-06-30	3.75	4.24	344.40	248.25	21.60	408.14	9.96	7.12	83.72	552.03	-8.80	2.1
1992-07-31	3.25	3.67	357.40	229.25	21.87	424.22	9.39	6.71	81.91	589.85	-28.88	2.1
1992-08-31	3.25	3.39	343.80	218.75	21.48	414.83	4.70	6.80	78.68	562.80	-17.61	2.1
1992-09-30	3.00	3.45	347.80	215.25	21.71	417.80	7.72	6.35	81.10	513.82	-27.13	3.1
1992-10-30	3.00	3.80	340.10	207.25	20.82	418.68	6.60	6.79	87.63	615.58	-11.55	3.1
1992-11-30	3.00	4.27	334.30	212.50	19.89	421.35	14.90	6.84	90.83	603.36	-1.73	3.1
1992-12-31	3.00	4.59	333.10	216.50	19.50	435.71	4.48	6.89	92.36	678.44	11.05	2.2
1993-01-29	3.00	4.60	330.20	214.50	20.26	438.78	7.34	6.36	92.48	670.56	-1.46	2.2
1993-02-28	3.00	4.30	339.80	211.25	20.80	443.38	7.43	6.82	94.81	642.96	-4.97	2.2
1993-03-31	3.00	3.82	337.60	230.25	20.44	451.67	11.89	6.82	91.87	668.75	9.97	1.4
1993-04-30	3.00	3.18	357.20	228.00	20.53	448.09	6.88	6.80	97.25	721.53	88.50	1.4
1993-05-31	3.00	3.00	378.30	224.50	20.82	458.19	8.39	6.15	88.08	762.31	31.62	1.4

# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

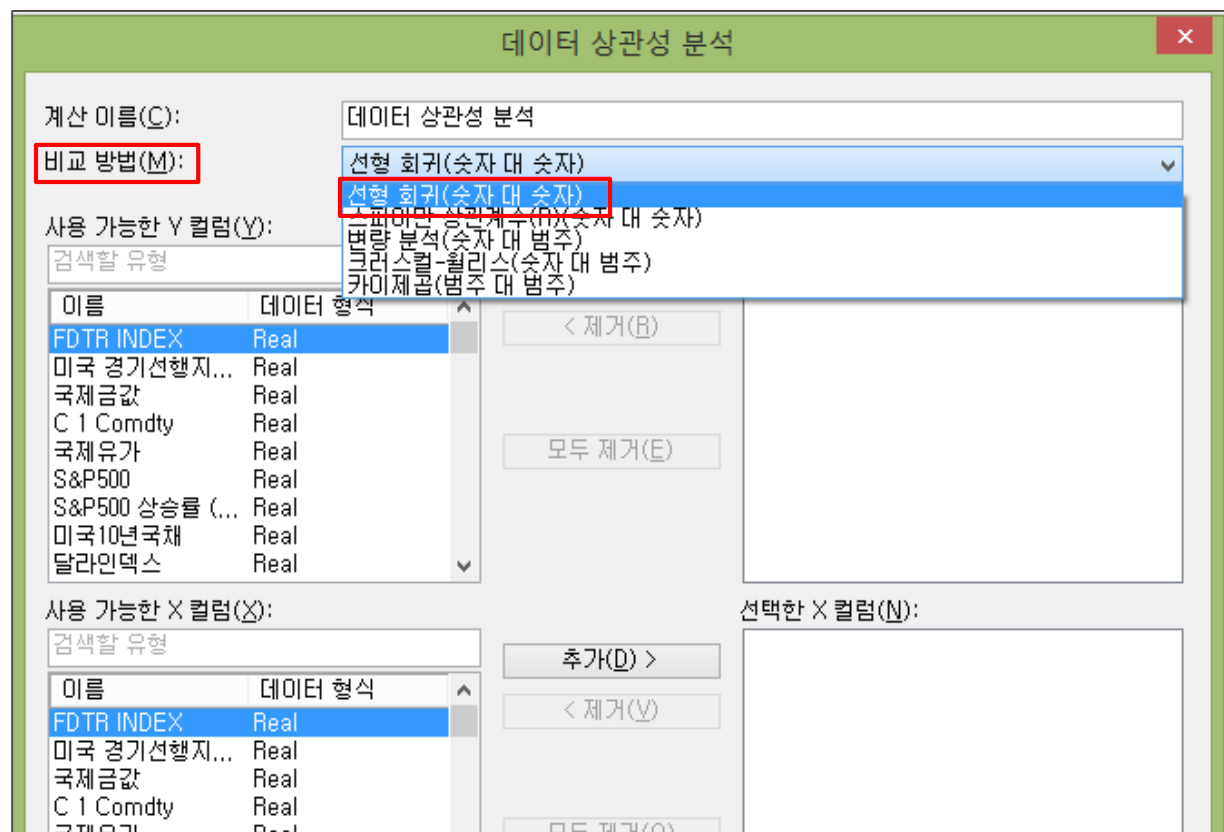
데이터 상관성 분석을 수행하여 **KOSPI**에 가장 영향을 미치는 인자들을 찾아 낸다.

2. 메인 메뉴 > 도구 > ‘데이터 관계’ 를 선택한다.



# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

3. ‘비교 방법’ 에서 원하는 상관성 분석 방법을 선택한다. **KOSPI data.xls**에 있는 컬럼의 값들은 종속 변수인 **kospi index**를 포함하여 거의 모두 숫자형 데이터 들이므로 여기서는 ‘선형 회귀(숫자 대 숫자)’ 방법을 선택한다.



# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

4. ‘사용 가능한 Y 컬럼’ 리스트 중에서 종속 변수를 하나 선택(여기서는 **kospi index**)하여 ‘추가’ 버튼을 눌러서 ‘선택한 Y 컬럼’ 란으로 이동 시킨다.

데이터 상관성 분석

계산 이름(C): 데이터 상관성 분석

비교 방법(M): 선형 회귀(숫자 대 숫자)

사용 가능한 Y 컬럼(Y):

검색할 유형

이름	데이터 형식
<b>kospi index</b>	Real
KOSPI 상승률 (yoy, 우)	Real
미국부동산가격지수 (yoy, 우)	Real
MTIByoy Index	Real
retail sales	Real
미시건	Real
미국통화량증가 (yoy, 좌)	Real
미국통화량증가 (yoy, 우)	Real
미국1개월국채	Real

추가 >(A)

< 제거(R)

모두 제거(E)

선택한 Y 컬럼(S):

사용 가능한 X 컬럼(X):

검색할 유형

이름	데이터 형식
<b>FDTR INDEX</b>	Real
미국 경기선행지수 (yoy, 우)	Real
국제금값	Real
C 1 Comdty	Real
국제유가	Real
S&P500	Real
S&P500 상승률 (yoy, 좌)	Real
미국10년국채	Real
달라인덱스	Real

추가(D) >

< 제거(V)

모두 제거(Q)



데이터 상관성 분석

계산 이름(C): 데이터 상관성 분석

비교 방법(M): 선형 회귀(숫자 대 숫자)

사용 가능한 Y 컬럼(Y):

검색할 유형

이름	데이터 형식
<b>FDTR INDEX</b>	Real
미국 경기선행지수 (yoy, 우)	Real
국제금값	Real
C 1 Comdty	Real
국제유가	Real
S&P500	Real
S&P500 상승률 (yoy, 좌)	Real
미국10년국채	Real
달라인덱스	Real

추가 >(A)

< 제거(R)

모두 제거(E)

선택한 Y 컬럼(S):

**kospi index**

사용 가능한 X 컬럼(X):

검색할 유형

이름	데이터 형식
<b>FDTR INDEX</b>	Real
미국 경기선행지수 (yoy, 우)	Real
국제금값	Real
C 1 Comdty	Real
국제유가	Real
S&P500	Real
S&P500 상승률 (yoy, 좌)	Real
미국10년국채	Real
달라인덱스	Real

추가(D) >

< 제거(V)

모두 제거(Q)

선택한 X 컬럼(N):

FDTR INDEX

미국 경기선행지수 (yoy, 우)

국제금값

C 1 Comdty

국제유가

S&P500

S&P500 상승률 (yoy, 좌)

미국10년국채

# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

5. ‘사용 가능한 X 컬럼’ 리스트에 있는 모든 독립 변수들을 선택(**shift** 키 이용)하고 ‘추가’ 버튼을 눌러서 ‘선택한 X 컬럼’ 란으로 이동 시키고 ‘예’ 를 클릭한다.

사용 가능한 Y 컬럼(Y):

검색할 유형

이름	데이터 형식
KOSPI 상승률 (y...	Real
미국부동산가격...	Real
MTIByoy Index	Real
retail sales	Real
미시건	Real
미국통화량증가...	Real
미국통화량증가...	Real
미국1개월국채	Real
미국3개월국채	Real

추가 >(A)

< 제거(R)

모두 제거(E)

선택한 Y 컬럼(S):

kospi index

사용 가능한 X 컬럼(X):

검색할 유형

이름	데이터 형식
한국 CPI	Real
엔화	Real
유로화	Real
위안화 선물	Real
mscina (3)	Real
goldman carry	Real
원/엔 환율	Real
실질금리	Real
한국 3년물 국채	Real

추가(D) >

< 제거(V)

모두 제거(O)

선택한 X 컬럼(N):

예

취소



사용 가능한 Y 컬럼(Y):

검색할 유형

이름	데이터 형식
FDTR INDEX	Real
미국 경기선행지...	Real
국제금값	Real
C 1 Comdty	Real
국제유가	Real
S&P500	Real
S&P500 상승률 (...)	Real
미국10년국채	Real
달러인덱스	Real

추가 >(A)

< 제거(R)

모두 제거(E)

선택한 Y 컬럼(S):

kospi index

사용 가능한 X 컬럼(X):

검색할 유형

이름	데이터 형식
한국 CPI	Real
엔화	Real
유로화	Real
위안화 선물	Real
mscina (3)	Real
goldman carry	Real
원/엔 환율	Real
실질금리	Real
한국 3년물 국채	Real

추가(D) >

< 제거(V)

모두 제거(O)

선택한 X 컬럼(N):

FDTR INDEX  
미국 경기선행지수 (yoy, 우)  
국제금값  
C 1 Comdty  
국제유가  
S&P500  
S&P500 상승률 (yoy, 좌)  
미국10년국채  
달러인덱스  
kospi index  
KOSPI 상승률 (yoy, 좌)  
미국부동산가격증가율  
MTIByoy Index  
retail sales  
미시건

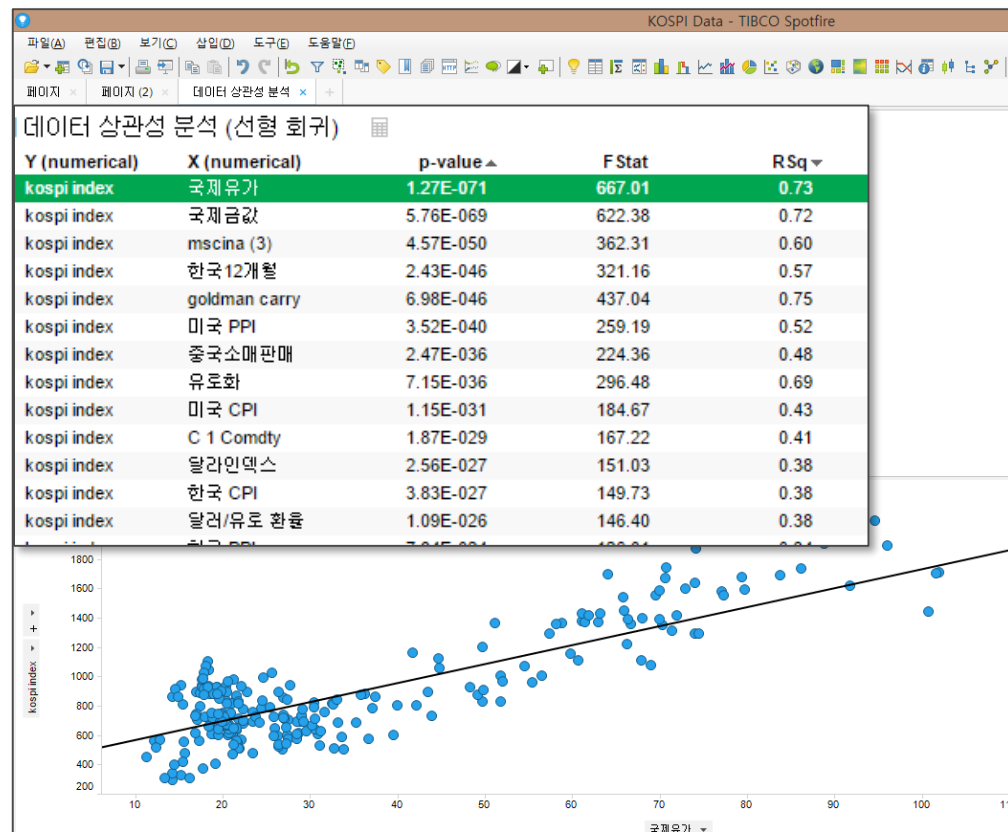
예

취소

# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

6. ‘데이터 상관성 분석’이라는 새로운 페이지가 생성되면서 상관 분석 결과가 표시된다.

- 자동으로 **p-value**가 작고 **Rsq**값이 큰 순서대로 정렬되어 있다.
- **P-value**의 값이 **0.05**이하인 경우 **Y**와 **X**는 서로 상관성이 존재한다고 말할 수 있다.
- 국제유가 > 국제 금값 > **mscina(3)** ... 등의 순으로 **kospi index**와 상관성이 강하게 있다는 것을 알 수 있다.
- 여기서는 상위 **10**개 정도의 인자들만 고려해 보기로 한다.

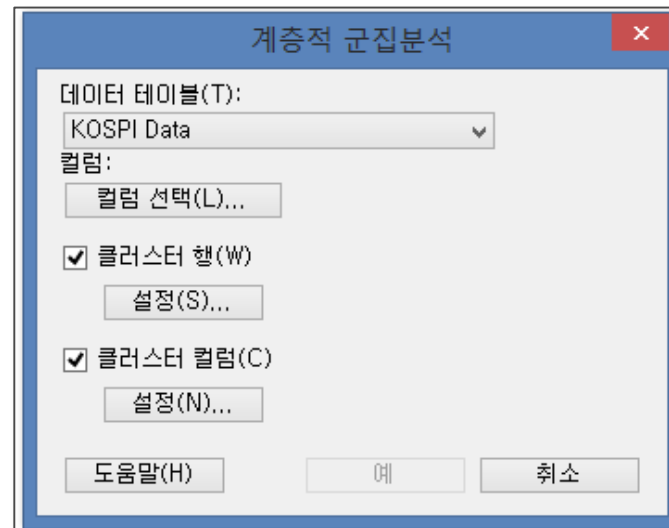
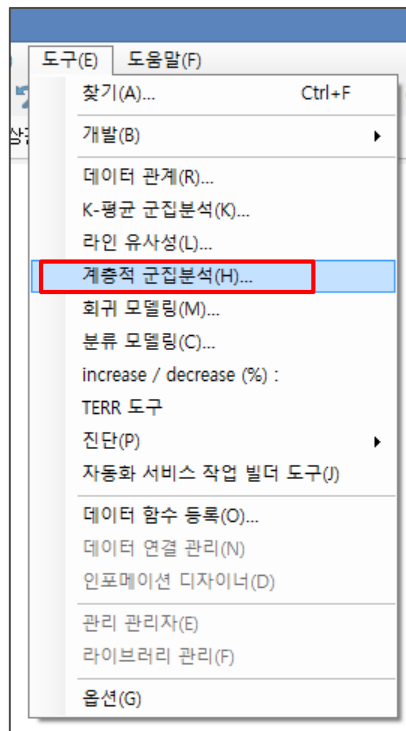




# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

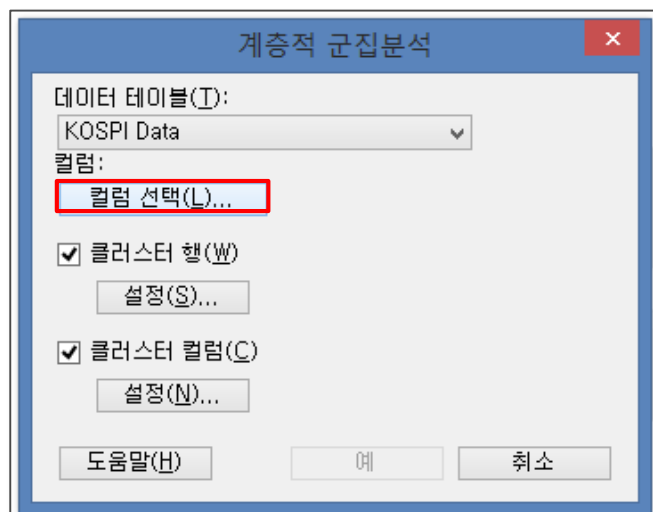
계층적 군집 분석을 수행한다.

7. 새 페이지를 만들고, 메인 메뉴 > 도구 > ‘계층적 군집 분석’ 을 선택한다.

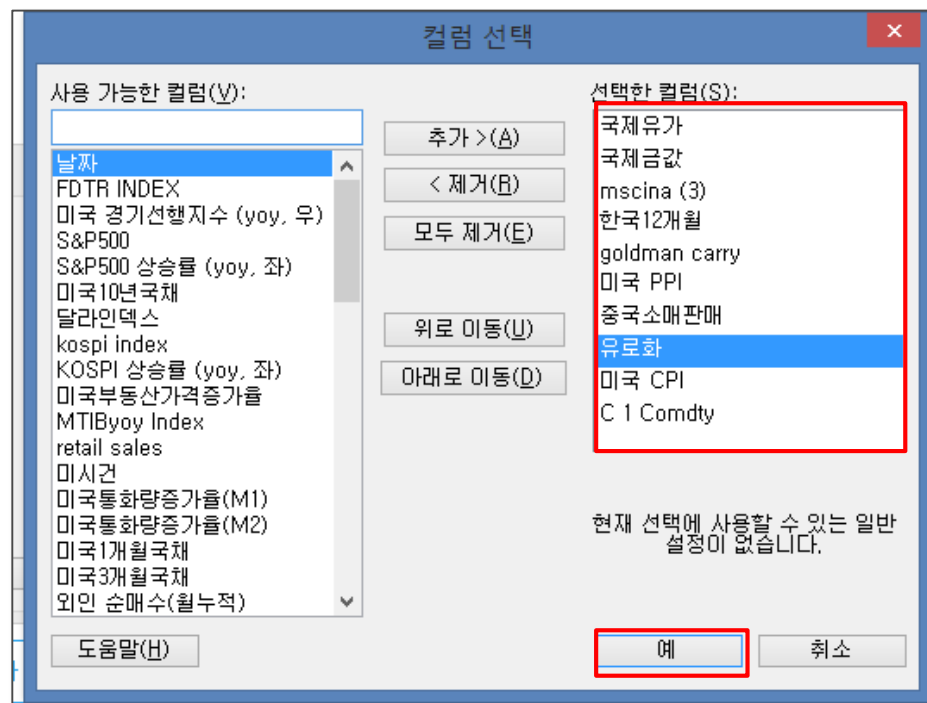


# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

8. ‘계층적 군집 분석’ 에서 ‘컬럼 선택’ 을 선택한다.

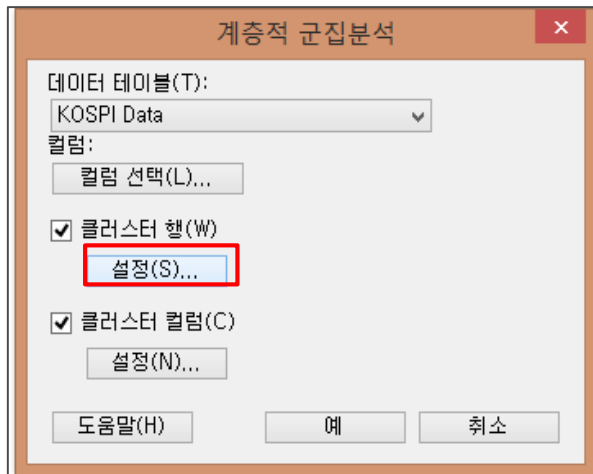


9. ‘6. 데이터 상관성 분석’ 수행 결과로 얻은 **p-value**가 가장 낮은 **10**개의 컬럼들을 ‘사용 가능한 컬럼’ 에서 선택하여 ‘선택한 컬럼’ 으로 ‘추가’ 하고 ‘예’ 를 클릭한다.

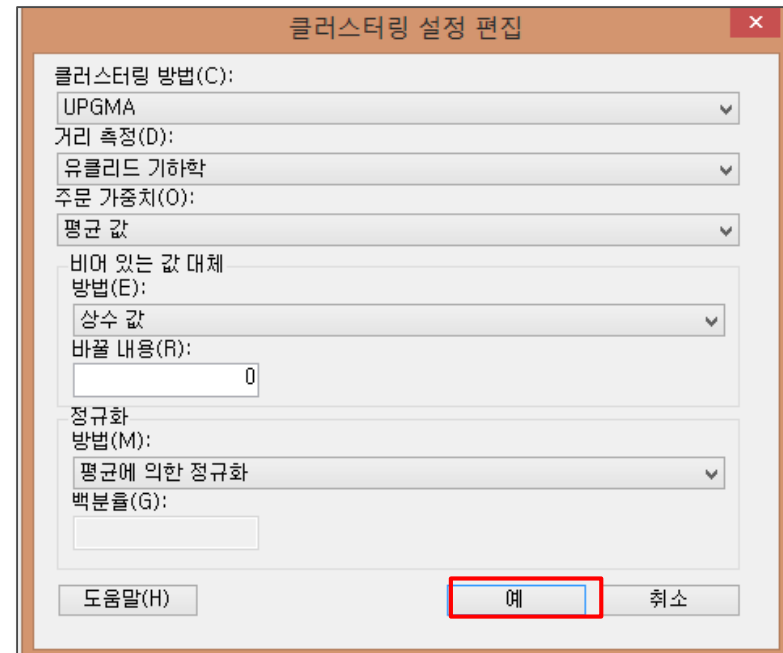


# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

10. ‘계층적 군집 분석’에서 ‘클러스터 행’을 선택한다.

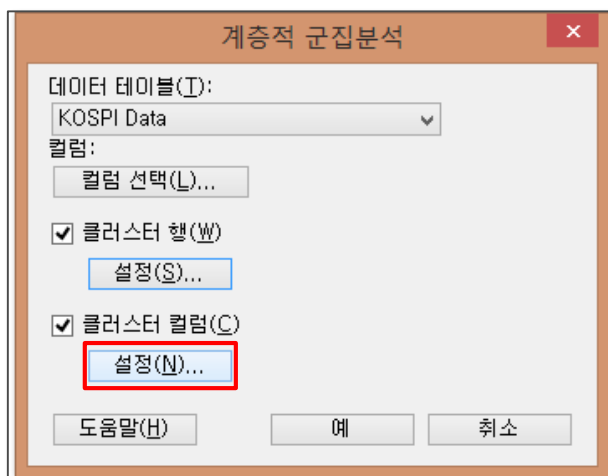


11. ‘클러스터링 설정 편집’에서 원하는 설정을 변경하고 ‘예’를 클릭한다.  
(여기서는 별도의 설정을 변경하지 않고 **default** 값들을 이용한다.)

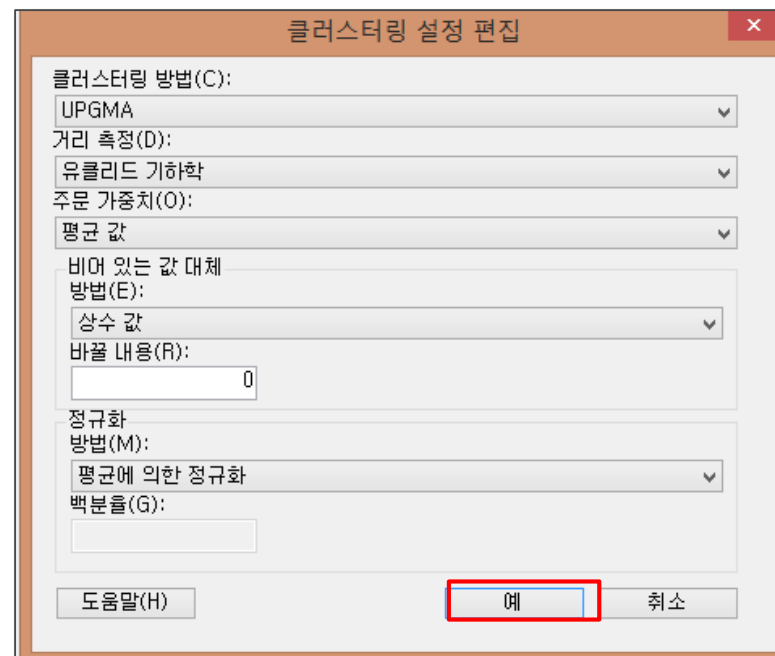


# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

12. ‘계층적 군집 분석’에서 ‘클러스터 컬럼’을 선택한다.

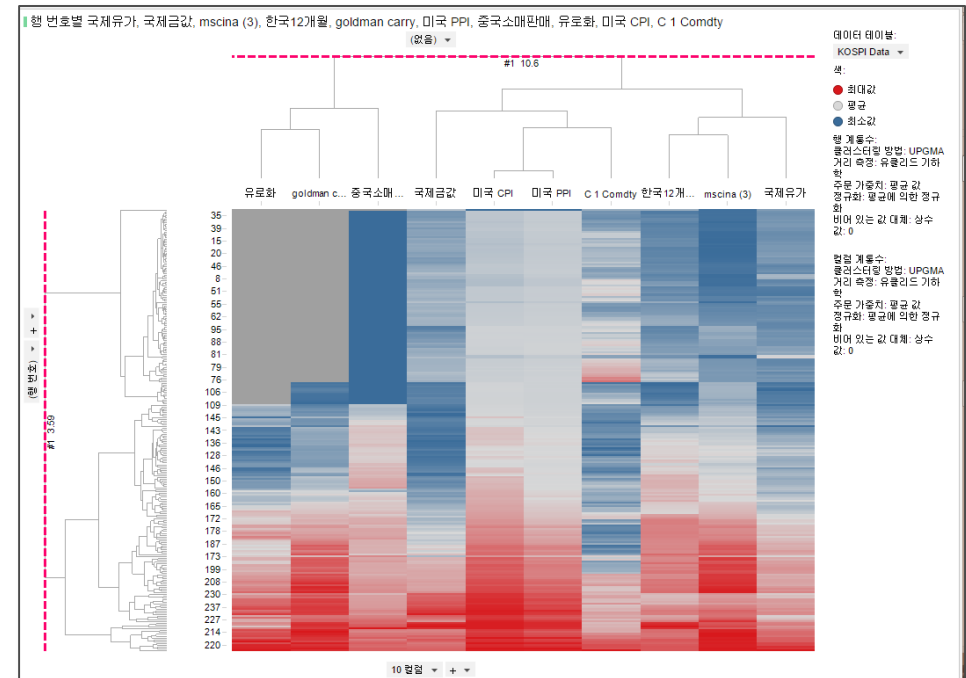
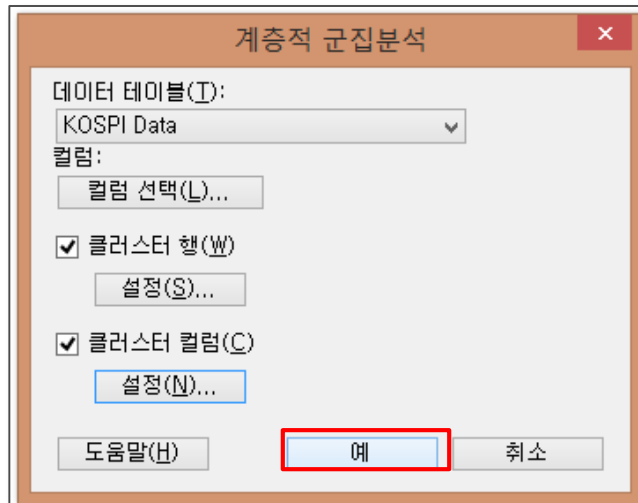


13. ‘클러스터링 설정 편집’에서 원하는 설정을 변경하고 ‘예’를 클릭한다.  
(여기서는 별도의 설정을 변경하지 않고 **default** 값들을 이용한다.)



# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

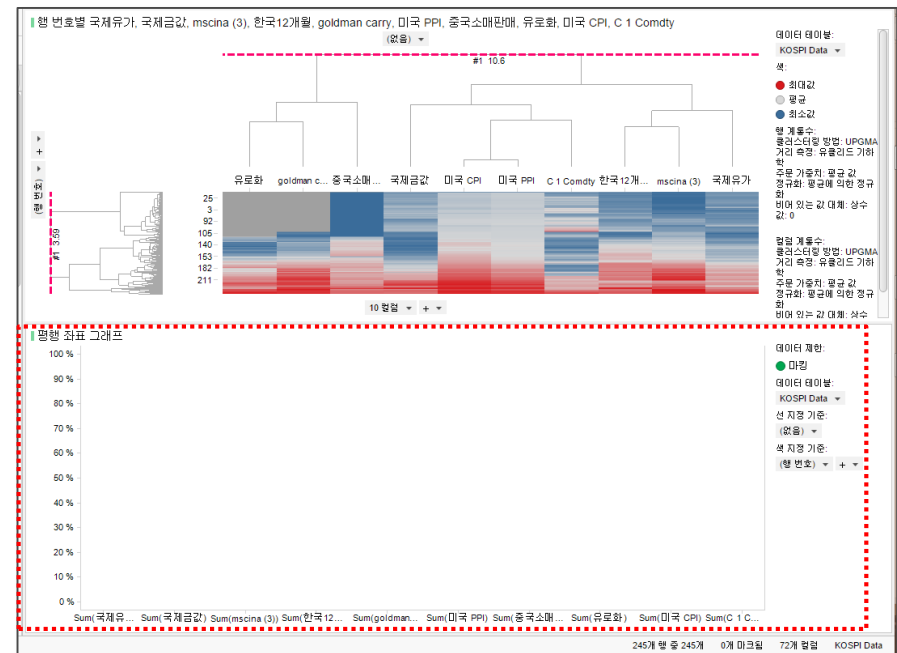
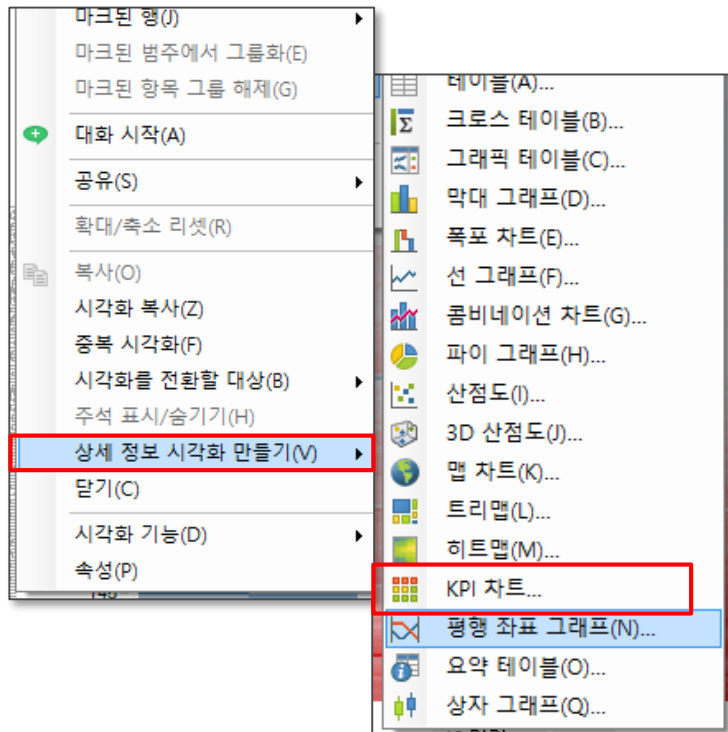
14. ‘계층적 군집 분석’에서 ‘예’를 클릭하여 실제 수행을 실시한다.  
수행 결과로 계통수가 표시된 히트맵이 생성된다.



# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

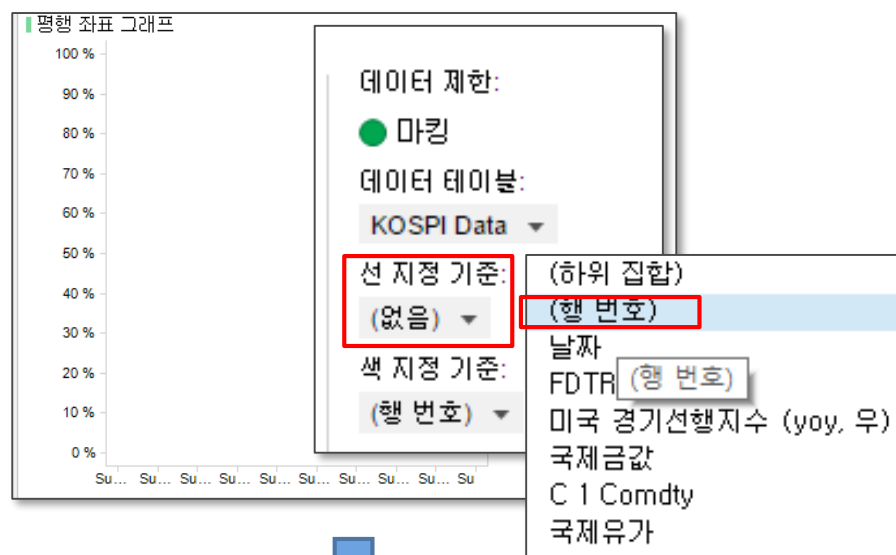
계층적 군집 분석으로부터 **drill down** 기능을 이용, 클러스터링을 수행한다.

15. 생성된 히트맵에서 마우스 우클릭 하여 ‘상세 정보 시각화 만들기’ > ‘평형 좌표 그래프’ 를 클릭한다. 수행 결과 아래에 ‘평형좌표 그래프’ 시각화가 생성된다.

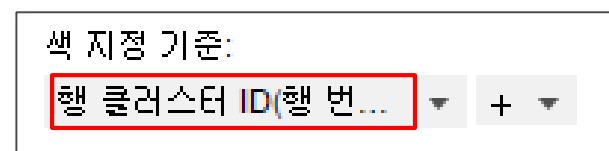
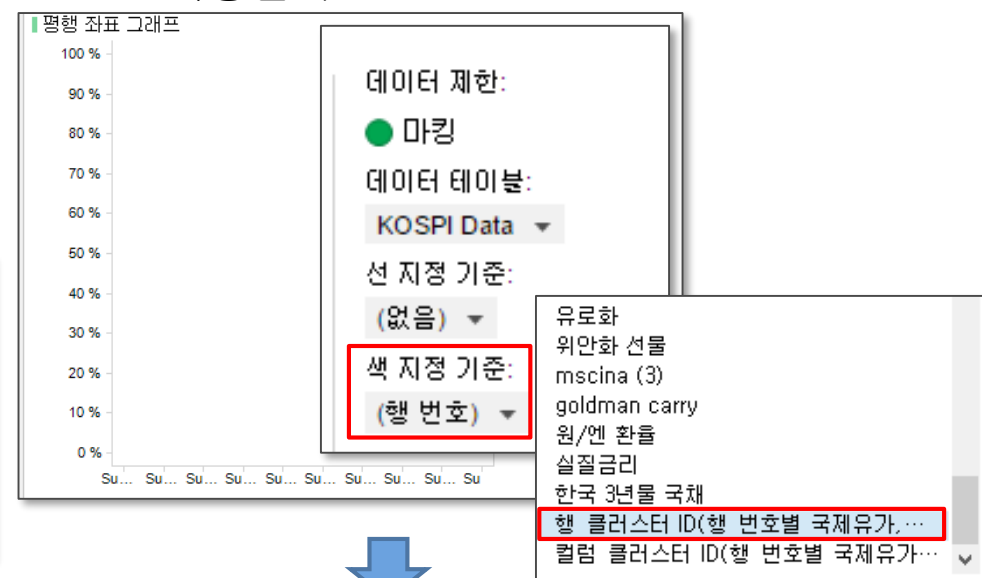


# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

16. '평형 좌표 그래프'의 범례에서 '선 지정 기준'을 '행 번호'로 지정한다.

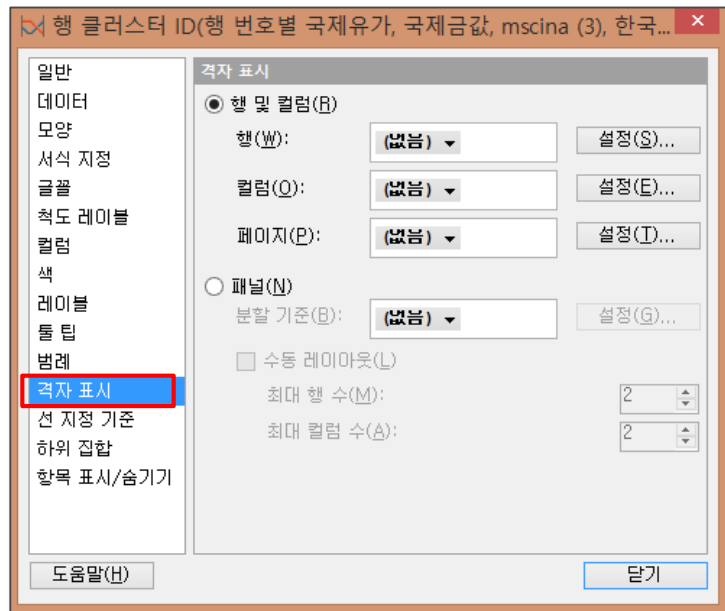


17. '평형 좌표 그래프'의 범례에서 '색 지정 기준'을 '행 클러스터 ID(행번호별 국제유가,...)'로 지정한다.

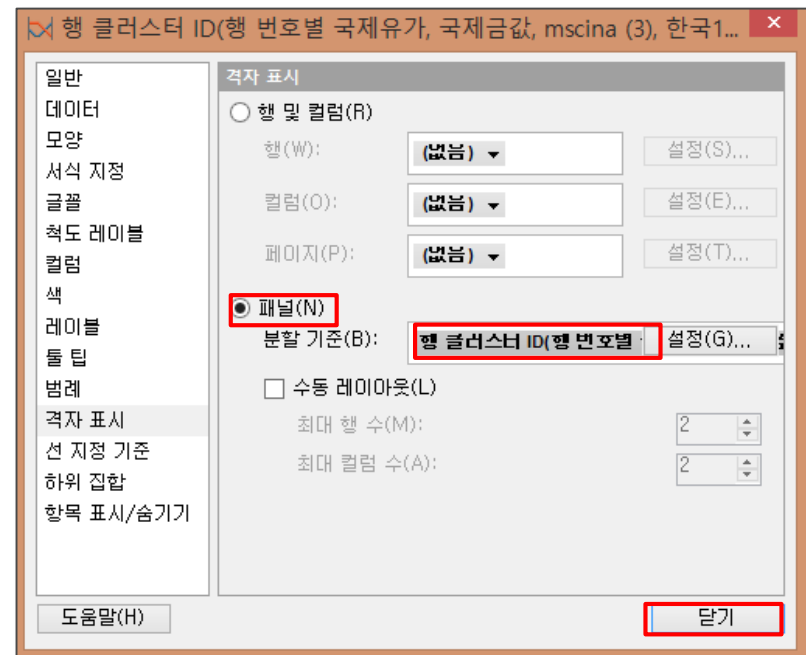


# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

18. 평형 좌표 그래프에서 마우스 우 클릭 하여 ‘속성’ > ‘격자 표시’를 클릭한다.



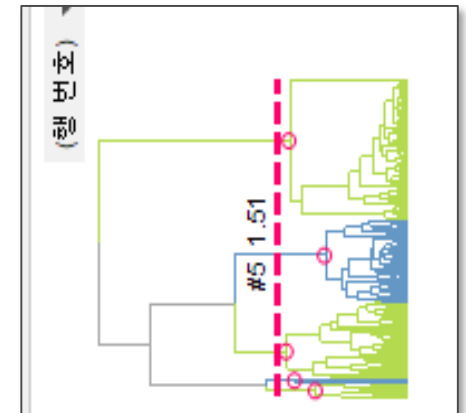
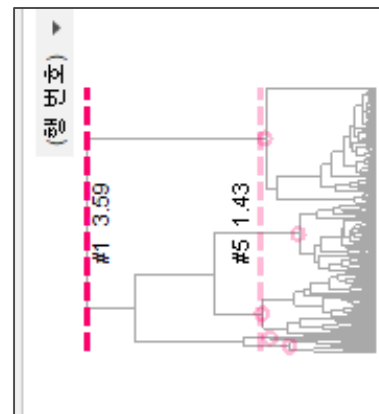
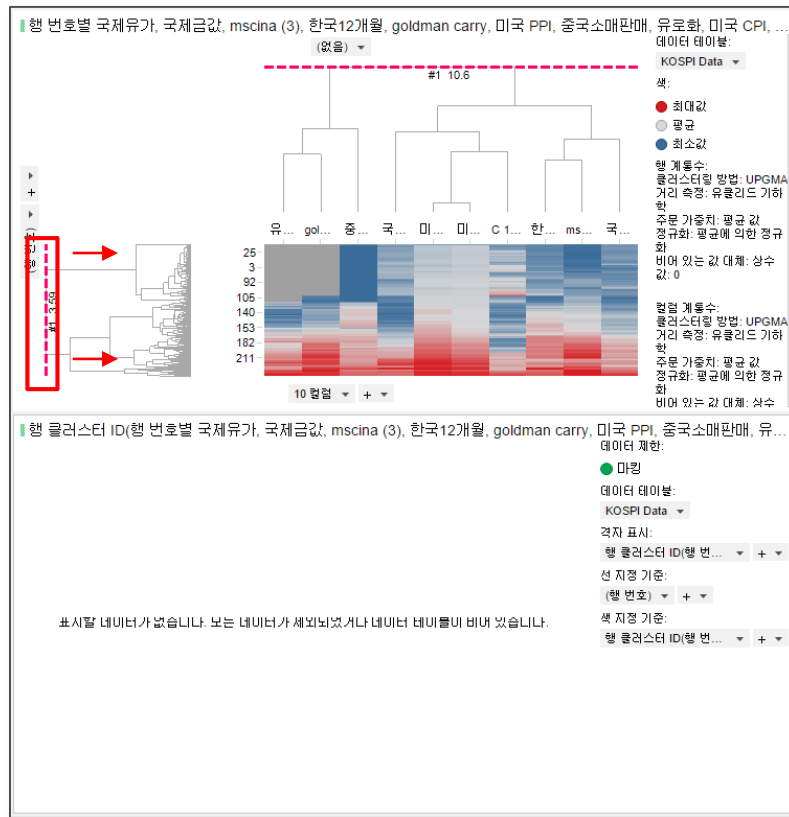
19. ‘패널’ 버튼을 선택하고 ‘분할 기준’의 선택 화살표를 눌러서 ‘행 클러스터 ID...’을 선택하고 ‘닫기’를 클릭한다.





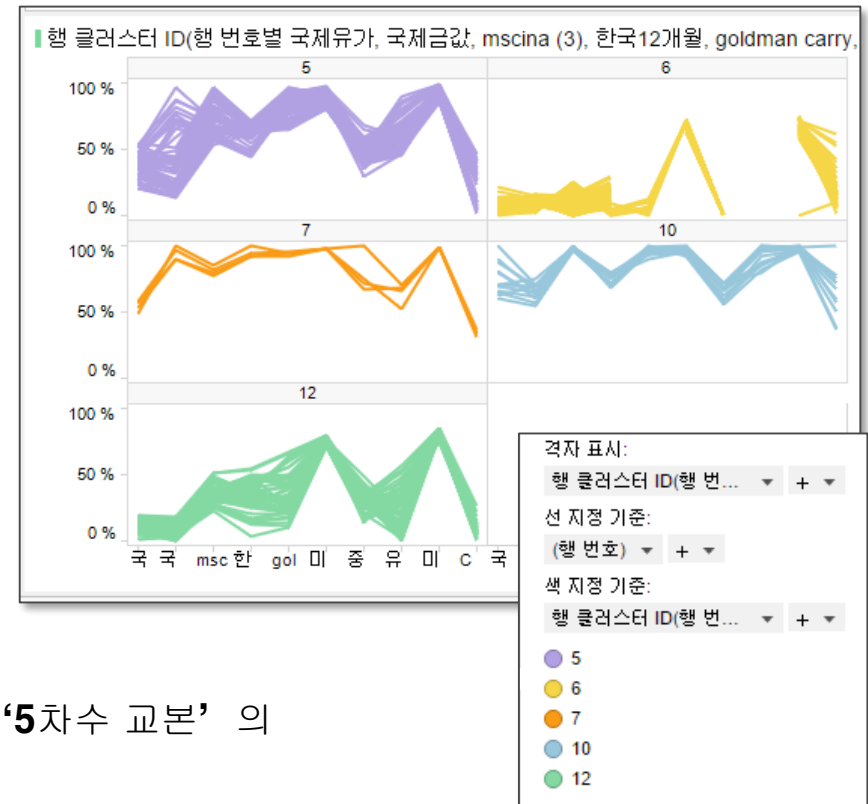
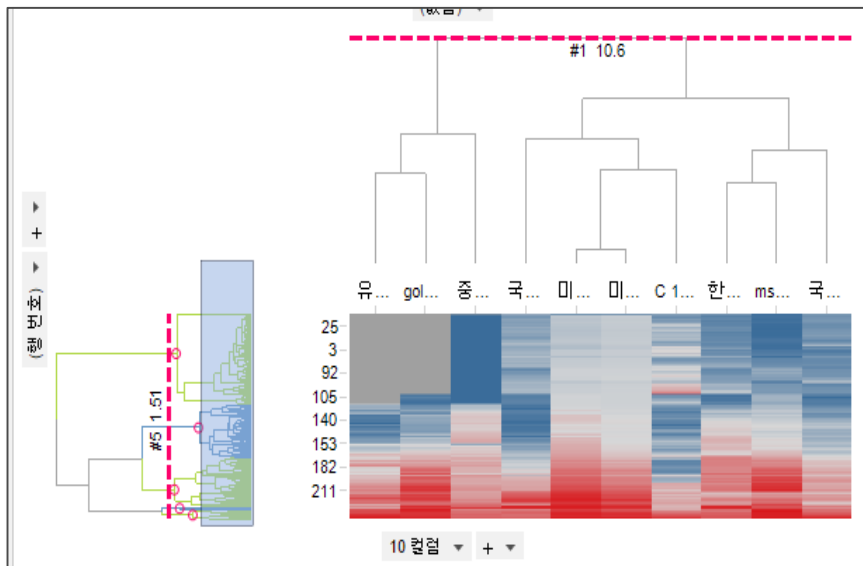
# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

**20.** 히트맵의 좌측에 빨간색 점선으로 표시되어 있는 선(자르기 라인 ; **pruning**)을 좌/우측으로 이동하면 해당되는 라인들간의 거리가 표시되고, 분홍색 원의 위치가 변경되면서 클러스터링의 개수를 알 수 있다. 아래 그림에서는 총 **5**개의 그룹이 존재하는 것을 알 수 있다.(총 **5**개의 분홍색 원)



# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

**21.** 이제 히트맵에서 자르기 라인의 우측 부분에 해당되는 부분을 아래 좌측 그림과 같이 마킹해 보자. 그 아래에 **drill down**으로 미리 설정해 놓은 평형좌표 그래프에 **5**개의 클러스터가 격자형태로 나뉘어져서 표시된다.(아래 우측 그림)

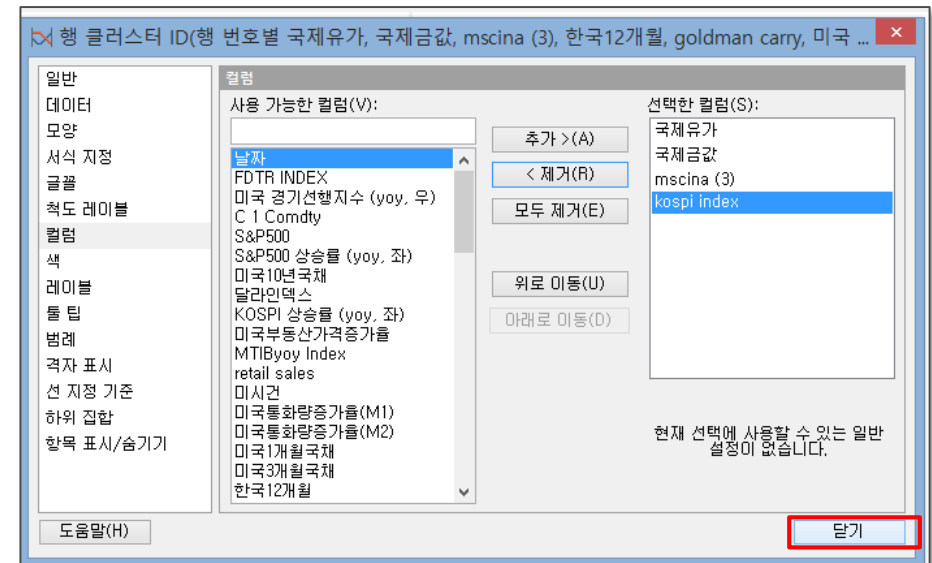
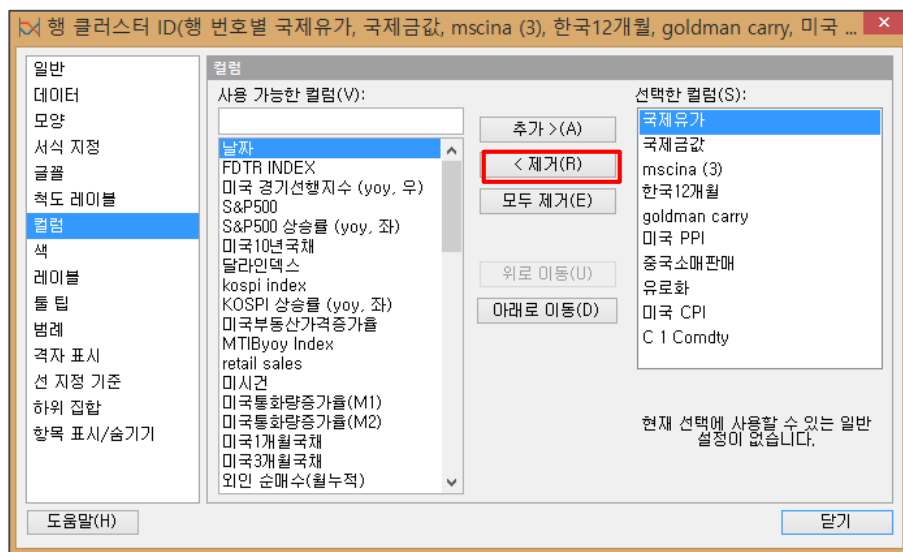


\* **Node ID** 번호, **pruning level**, 계통수에 대한 설명은 ‘5차수 교본’의 10page를 참조 바람.

# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

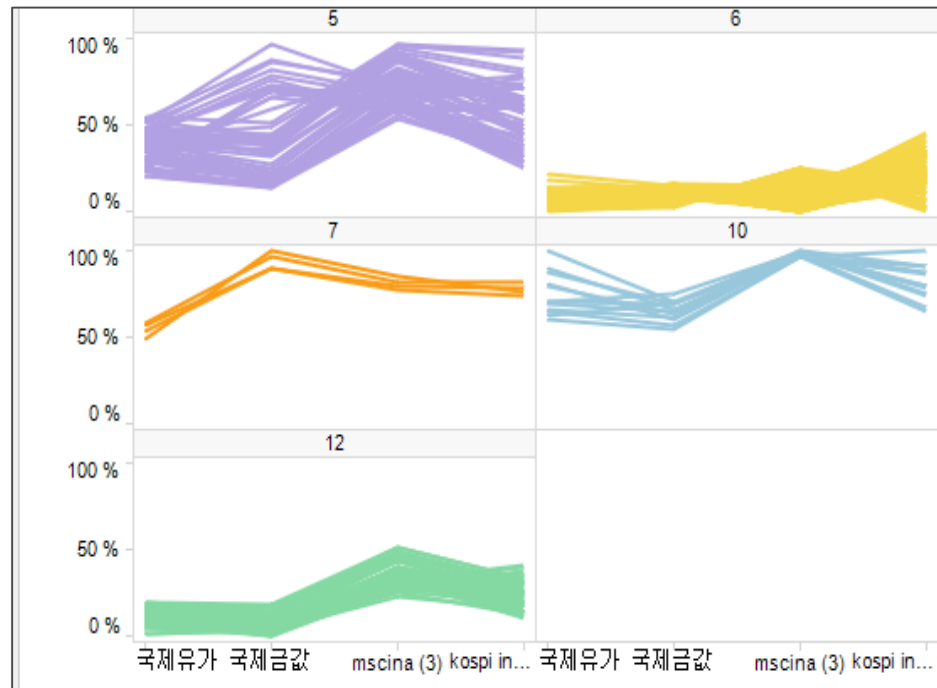
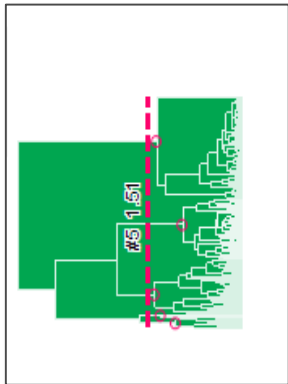
**22.** 평형좌표 그래프의 **X**축에 해당되는 부분에는 여러 개의 컬럼명들이 존재하고 있다. 여기에 너무 많은 컬럼들이 있으면 전체적인 패턴을 알기가 쉽지 않다. 따라서 사용자가 이해할 수 있는 정도로 컬럼 수를 줄이는 것이 좋다.

평형좌표 그래프에 마우스 커서를 놓고 마우스 우클릭 > 속성 > 컬럼 을 선택하면 컬럼을 선택할 수 있다. 여기에서 우측의 ‘선택한 컬럼’ 리스트 중에서 핵심적인 몇 개만 남겨두고 ‘사용 가능함 컬럼’ 으로 ‘제거’ 시키면 된다. (여기서는 추가로 ‘koshi index’를 추가하였다.)



# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

- 23.** 변경된 평형좌표 그래프를 확인하기 위하여, **‘21번 단계’**를 반복해 보자. 히트맵에서 자르기 라인의 우측 부분에 해당되는 부분을 마킹해 본다. 그 아래에 **drill down**으로 미리 설정해 놓은 평형좌표 그래프가 이제 **4개의 컬럼**만으로 표시되므로 **5개의 클러스터링 그룹**에 대하여 직관적으로 쉽게 비교할 수 있다.



# 1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

**24.** 마지막으로 사용자가 임의로 자르기 라인을 이동해 가면서 원하는 클러스터링 그룹이 적당해질 때 까지 실행하면 된다.

