7-2. 계층적 군집분석(실습)

김 성 기

목 차

1. 계층적 군집 분석(Hierarchical Clustering)

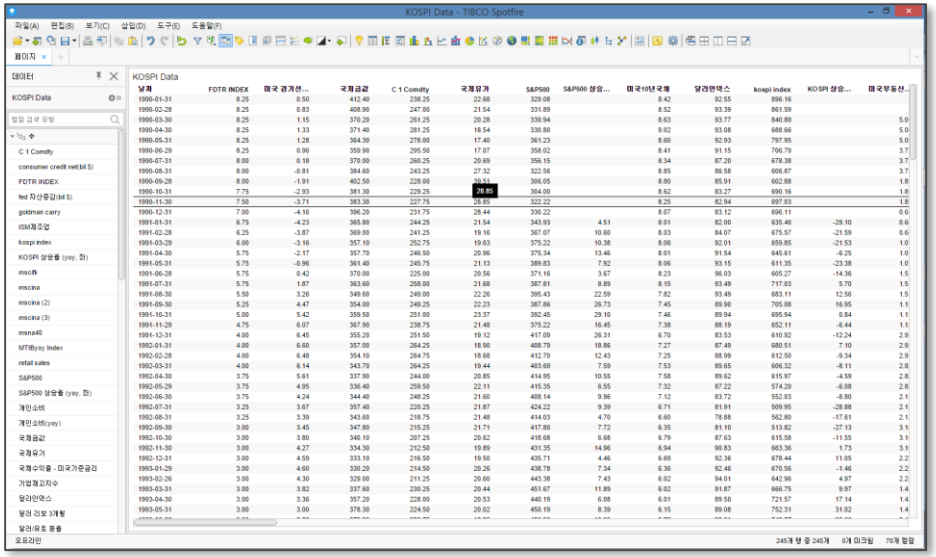
   : KOSPI 지수에 영향을 끼치는 여러 경제 지표들에 대하여 월별로 유사한 군집으로 분석

KOSPI Data.xls

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 **load**한다**.**

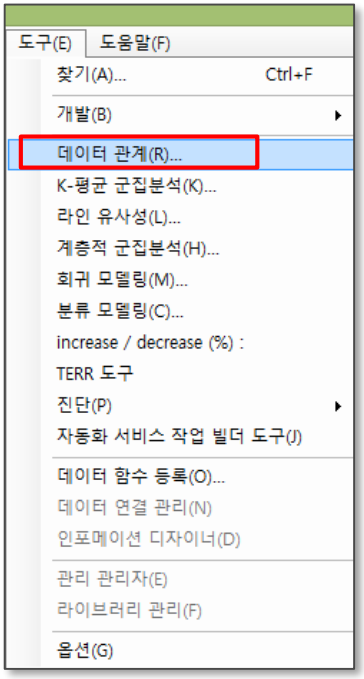


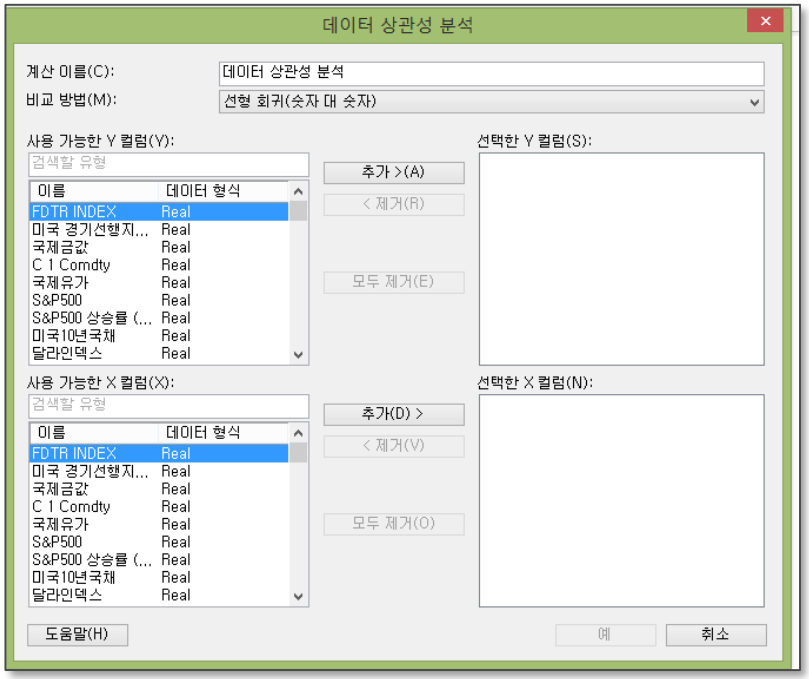
KOSPI Data.xls

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.

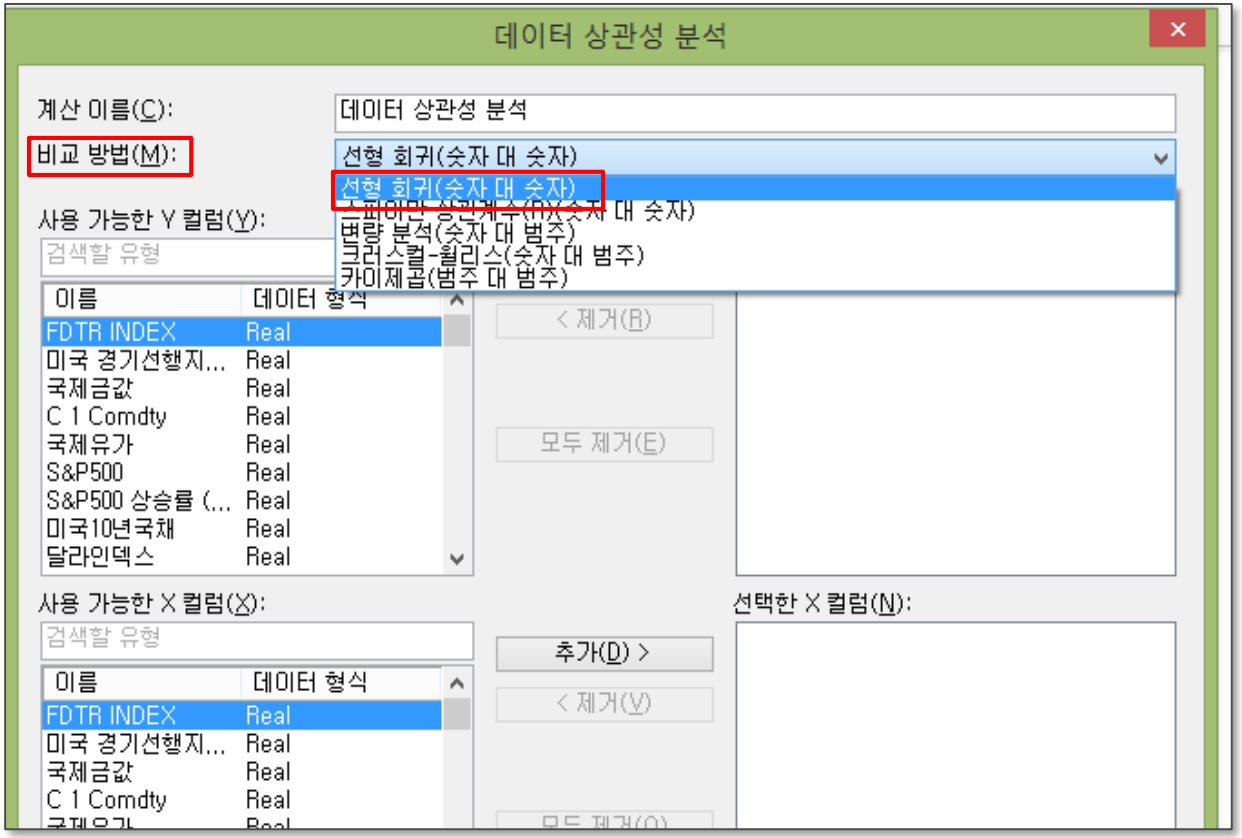




KOSPI Data.xls

1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

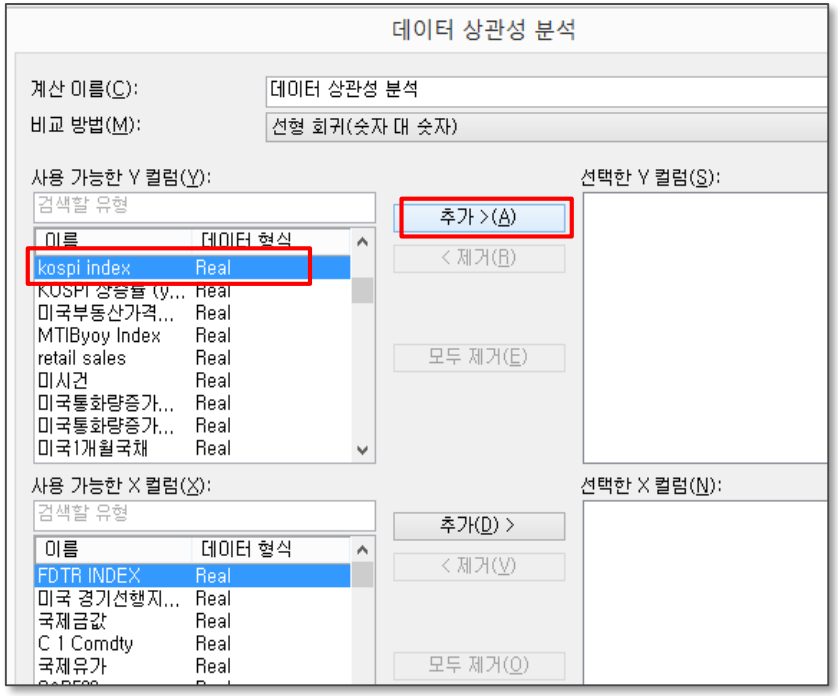
3. ‘비교 방법’에서 원하는 상관성 분석 방법을 선택한다. **KOSPI data.xls**에 있는 컬럼의 값들은 종속 변수인 **kospi index**를 포함하여 거의 모두 숫자형 데이터 들이므로 여기서는 ‘선형 회귀(숫자 대 숫자)’ 방법을 선택한다.

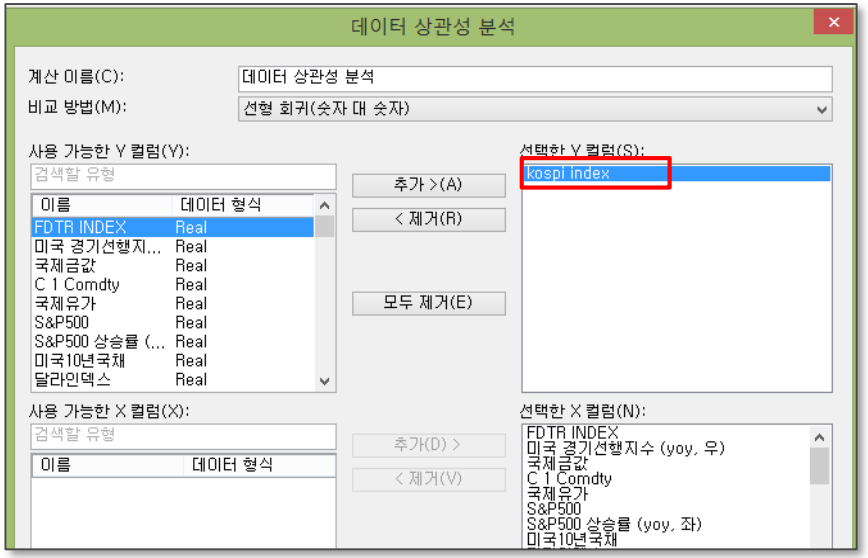


KOSPI Data.xls

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**.**

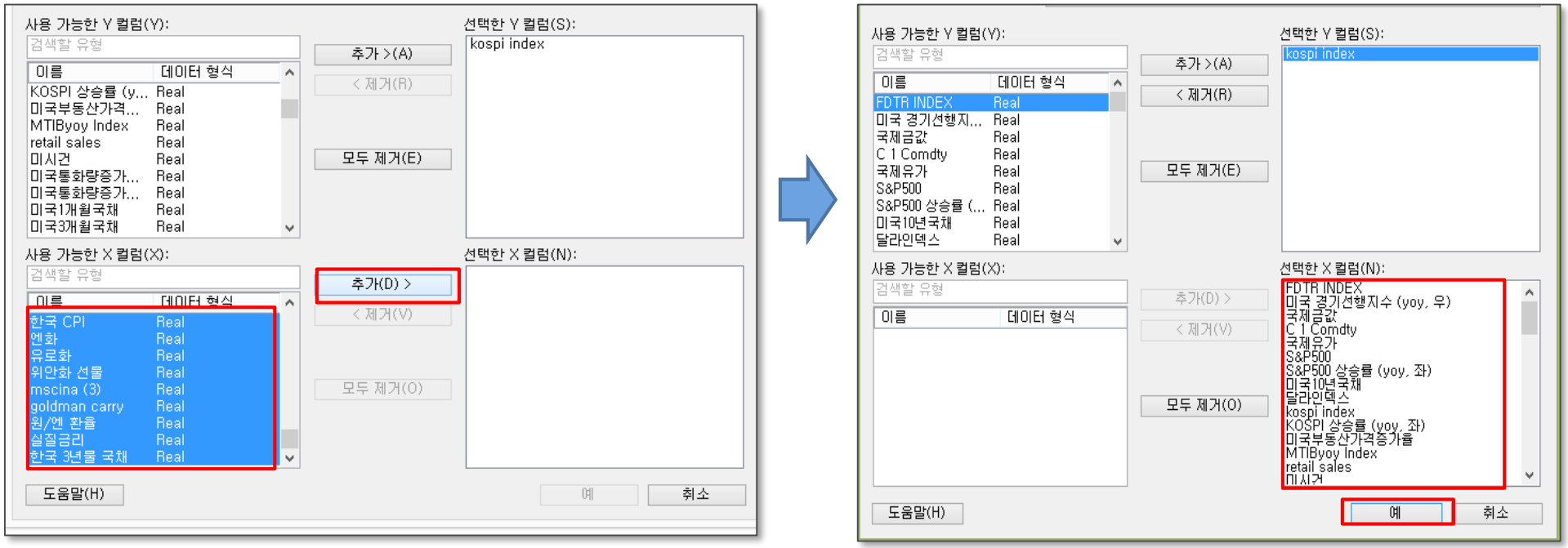




KOSPI Data.xls

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**.**

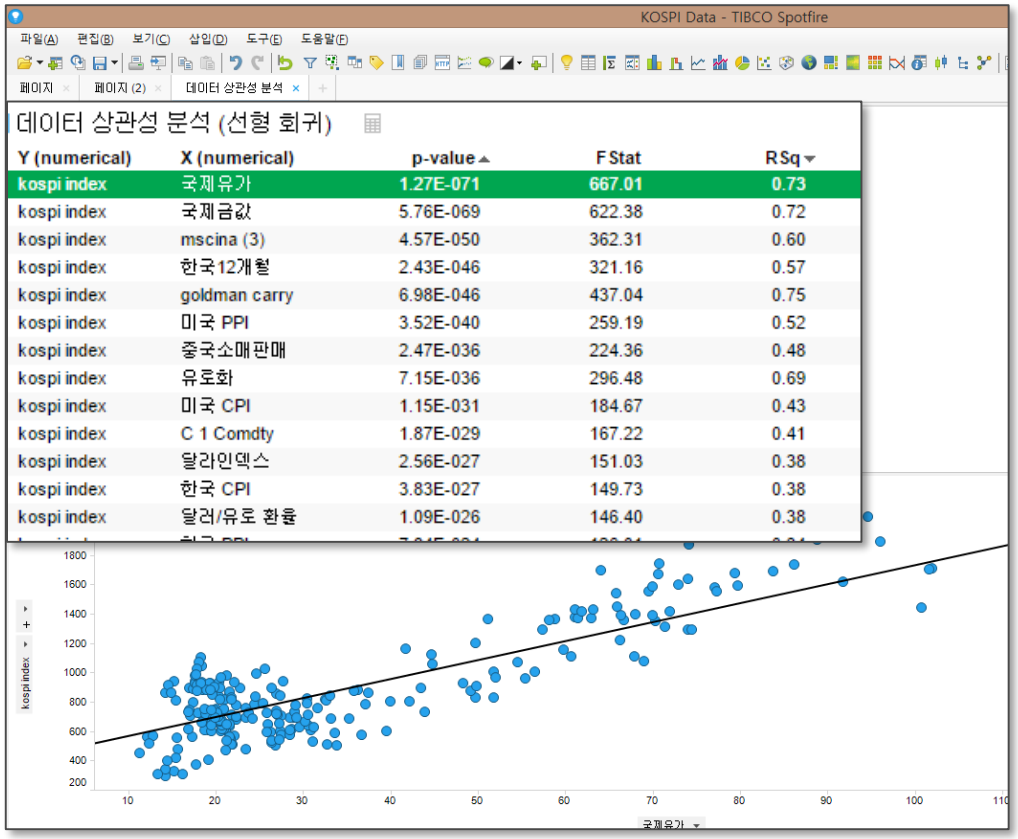


KOSPI Data.xls

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.

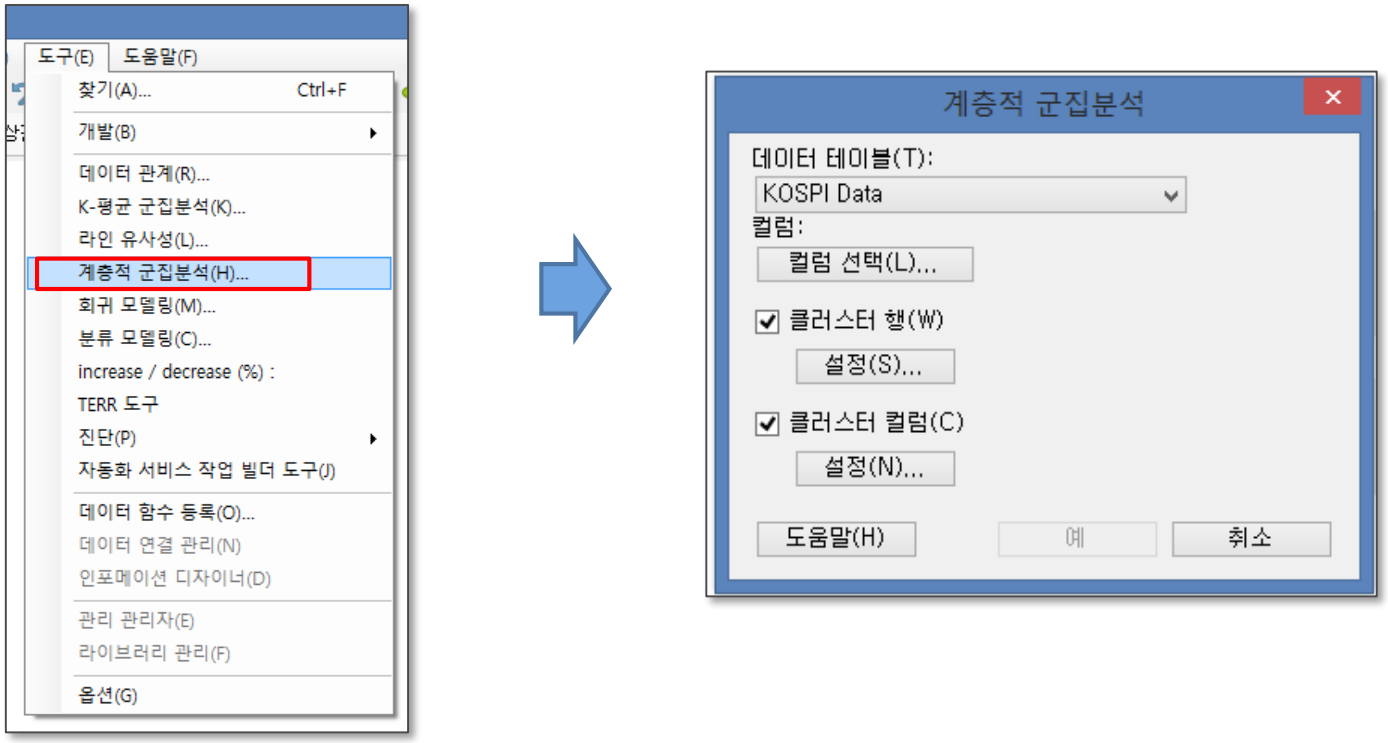


KOSPI Data.xls

1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

계층적 군집 분석을 수행한다.

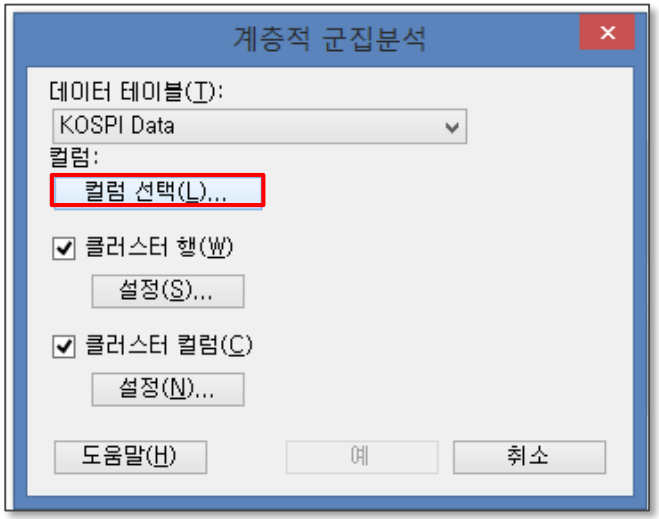
7. 새 페이지를 만들고, 메인 메뉴 > 도구 > '계층적 군집 분석'을 선택한다.



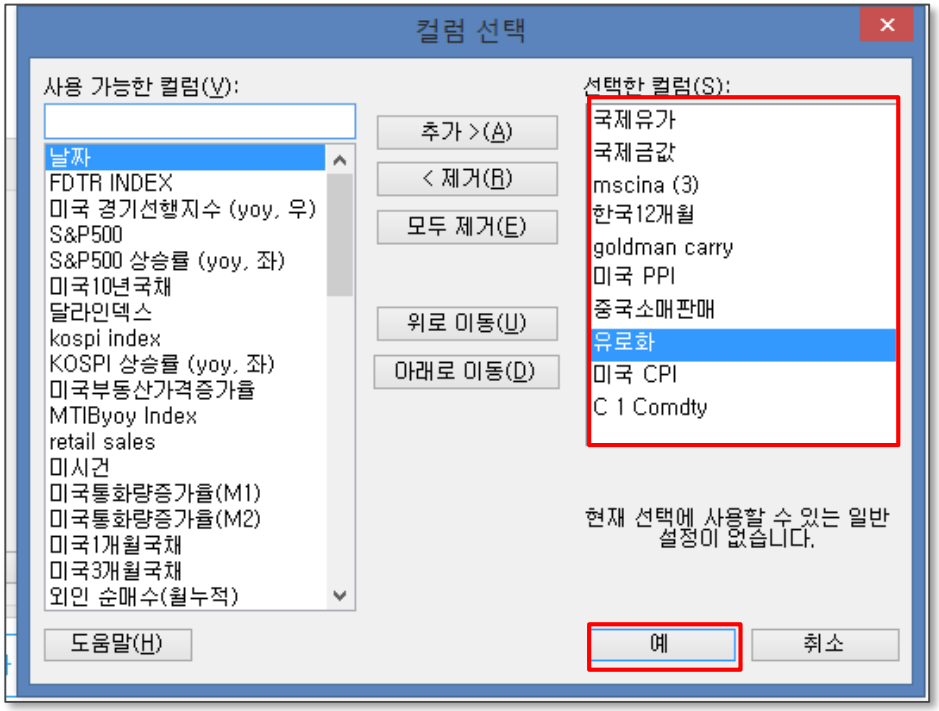
KOSPI Data.xls

1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

8. ‘계층적 군집 분석’에서 ‘컬럼 선택’을 선택한다.



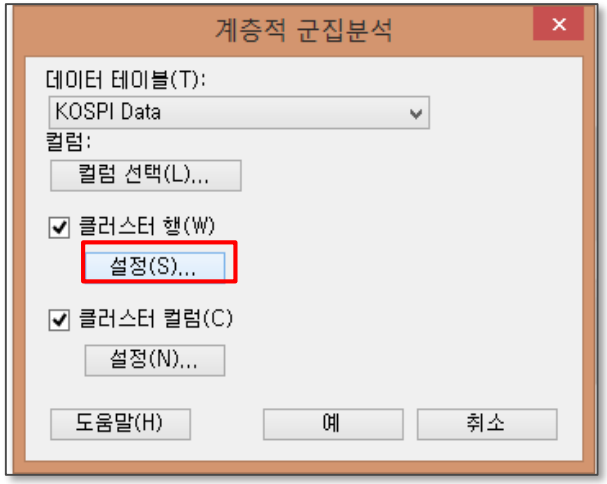
9. ‘**6.** 데이터 상관성 분석’ 수행 결과로 얻은 **p-value**가 가장 낮은 **10**개의 컬럼들을 ‘사용 가능한 컬럼’에서 선택하여 ‘선택한 컬럼’으로 ‘추가’하고 ‘예’를 클릭한다.



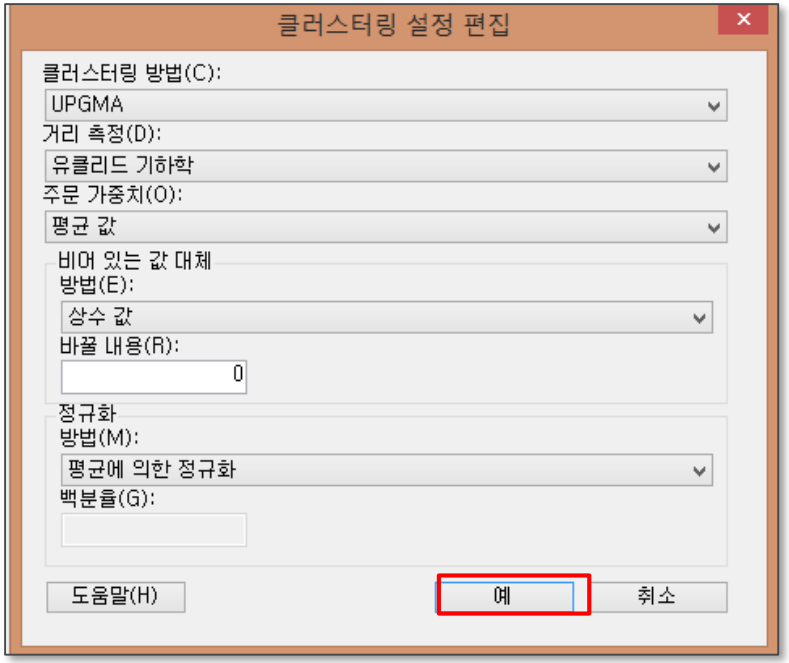
KOSPI Data.xls

1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

10. ‘계층적 군집 분석’ 에서 ‘클러스터 행’ 을 선택한다.



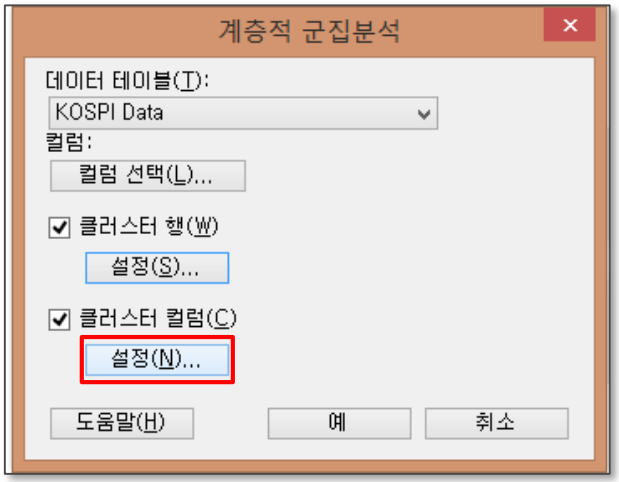
11. ‘클러스터링 설정 편집’ 에서 원하는 설정을 변경하고 ‘예’ 를 클릭한다.
(여기서는 별도의 설정을 변경하지 않고 **default** 값들을 이용한다**.)**



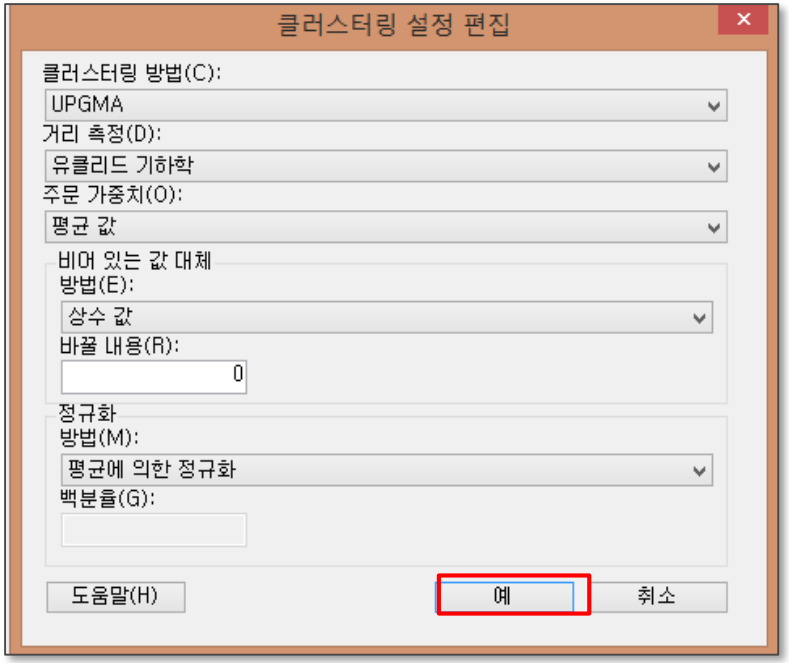
KOSPI Data.xls

1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

12. ‘계층적 군집 분석’ 에서 ‘클러스터 컬럼’ 을 선택한다.



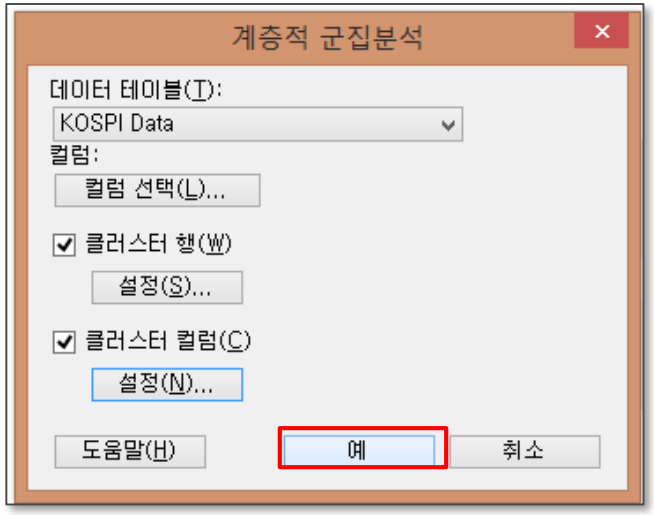
13. ‘클러스터링 설정 편집’ 에서 원하는 설정을 변경하고 ‘예’ 를 클릭한다.
(여기서는 별도의 설정을 변경하지 않고 **default** 값들을 이용한다**.)**

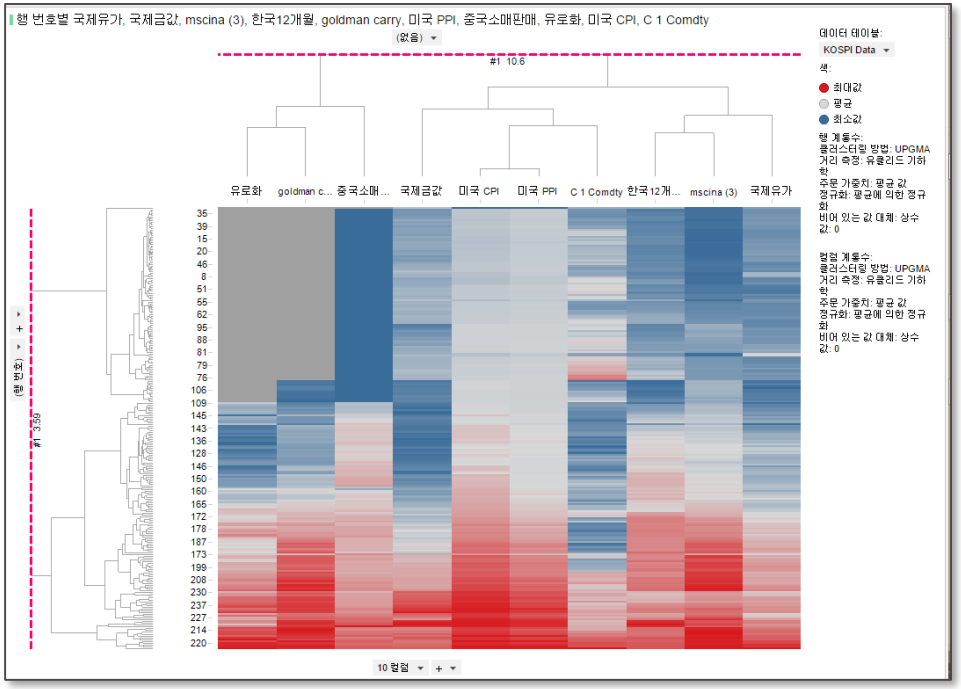


KOSPI Data.xls

1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

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



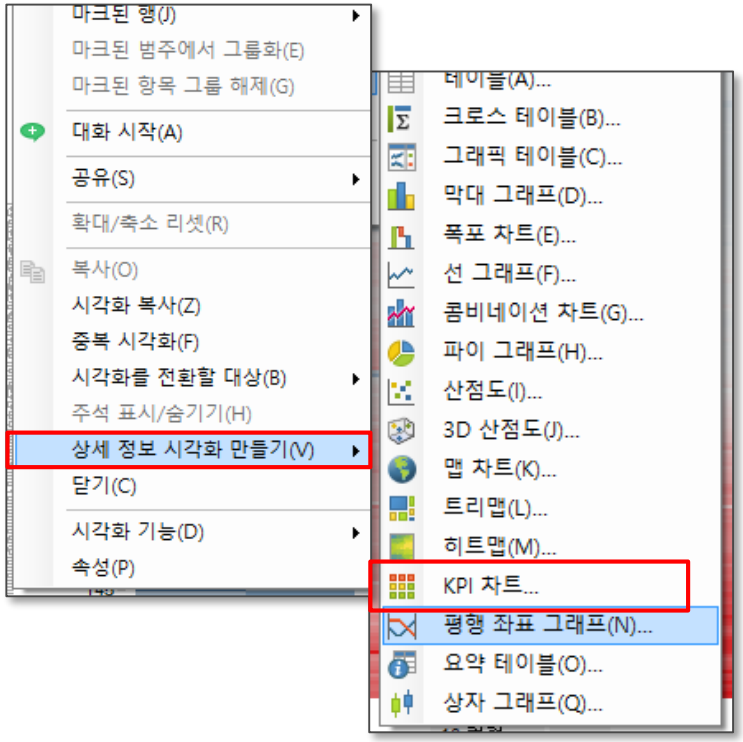


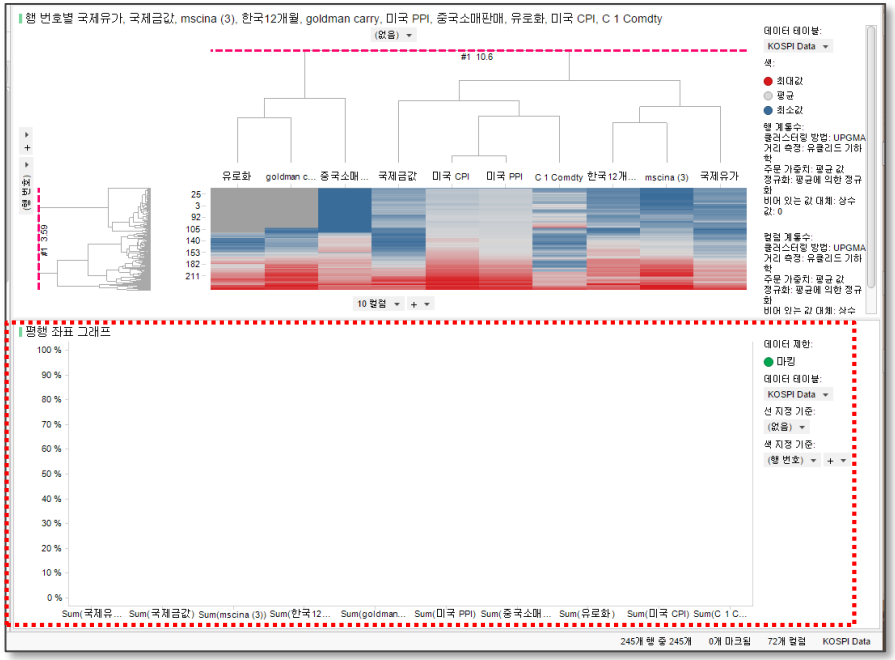
KOSPI Data.xls

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.

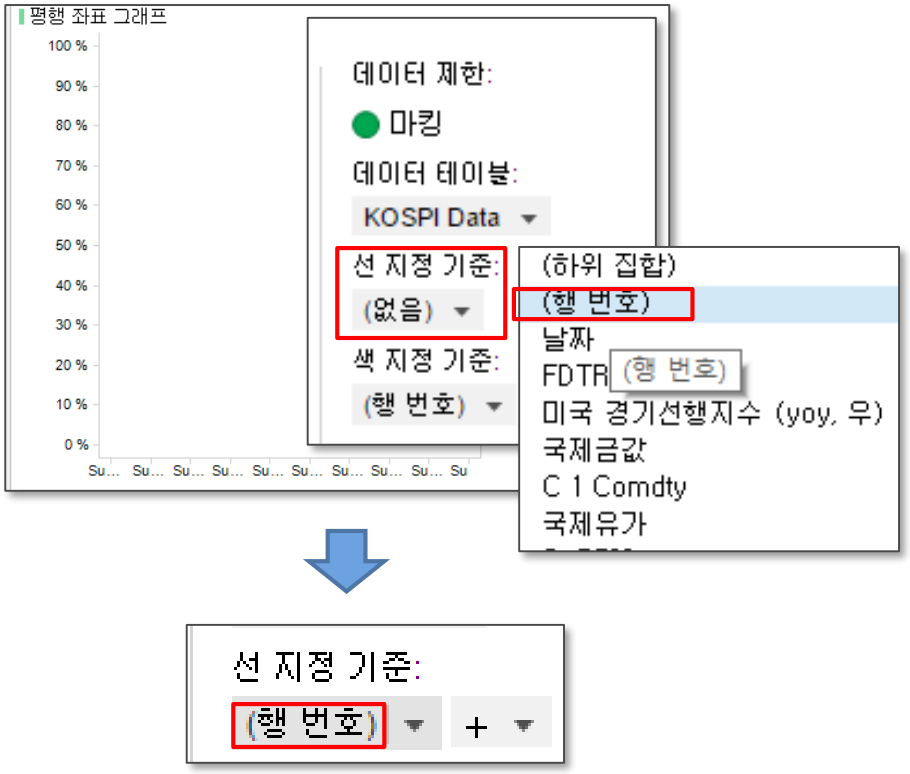




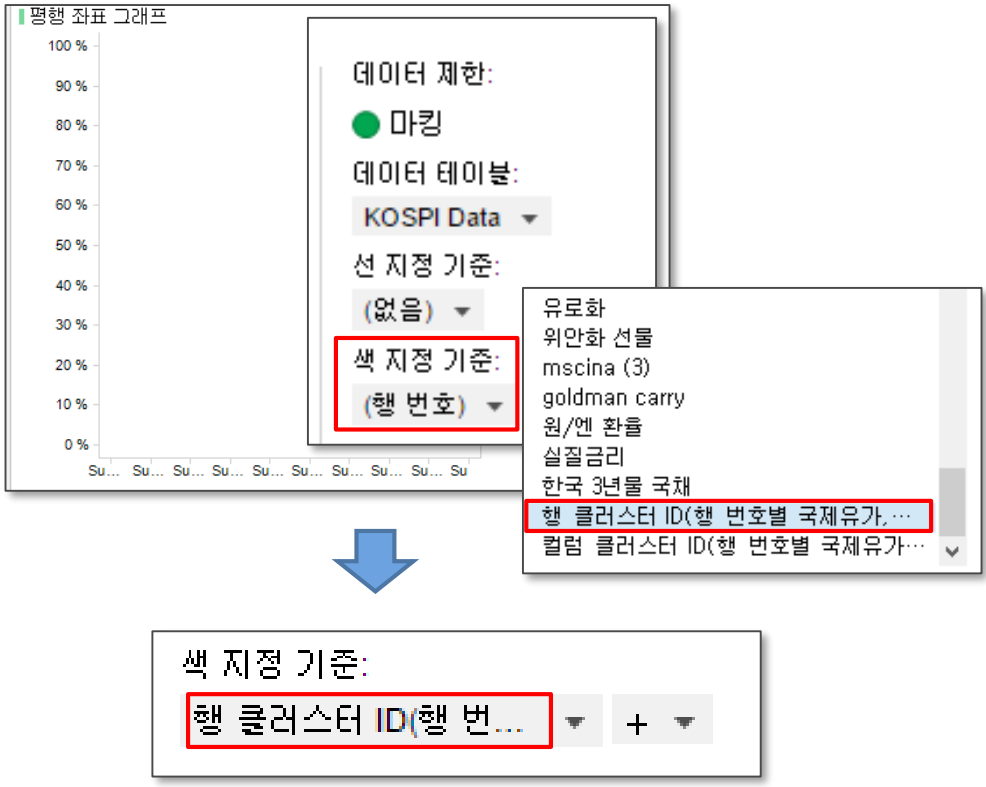
KOSPI Data.xls

1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

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



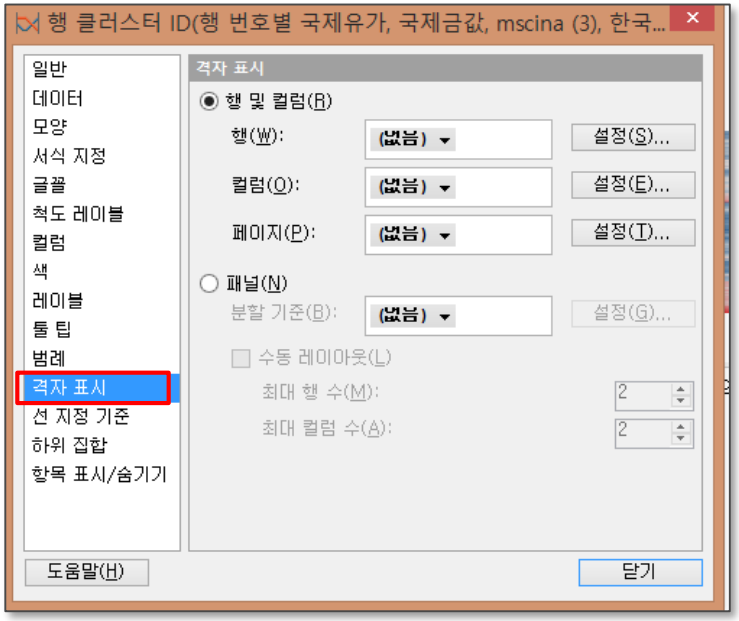
17. ‘평형 좌표 그래프’ 의 범례에서 ‘색 지정 기준’ 을 ‘행 클러스터 **ID(**행번호별 국제유가,… ’ 로 지정한다.



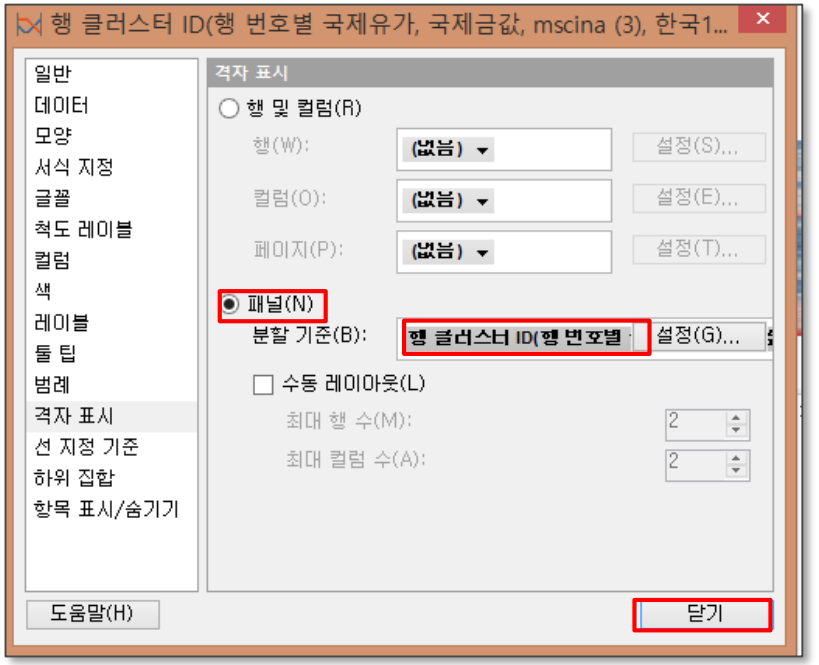
KOSPI Data.xls

1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

18. 평형 좌표 그래프에서 마우스 우클릭 하여 '속성' > '격자 표시'를 클릭한다.



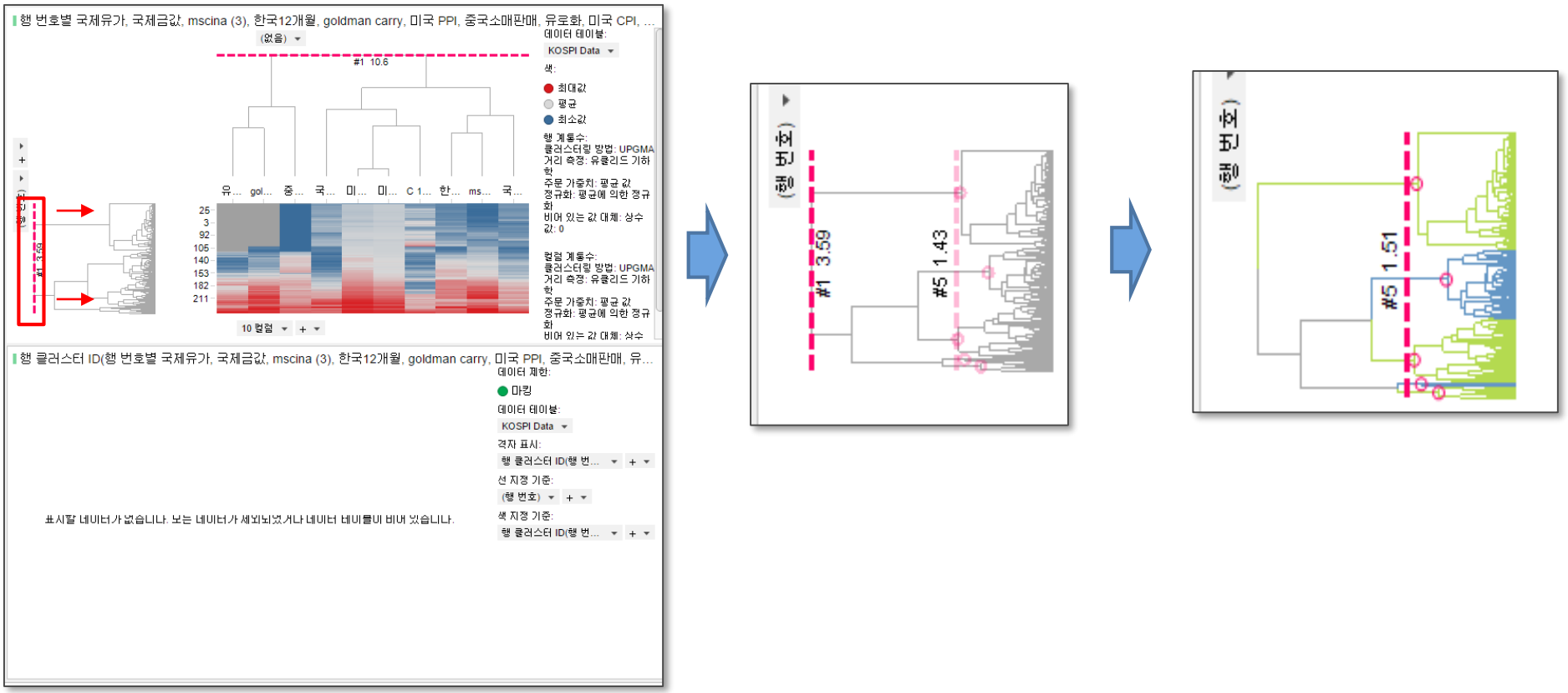
19. '패널' 버튼을 선택하고 '분할 기준'의 선택 화살표를 눌러서 '행 클러스터 **ID...**'을 선택하고 '닫기'를 클릭한다.



KOSPI Data.xls

1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

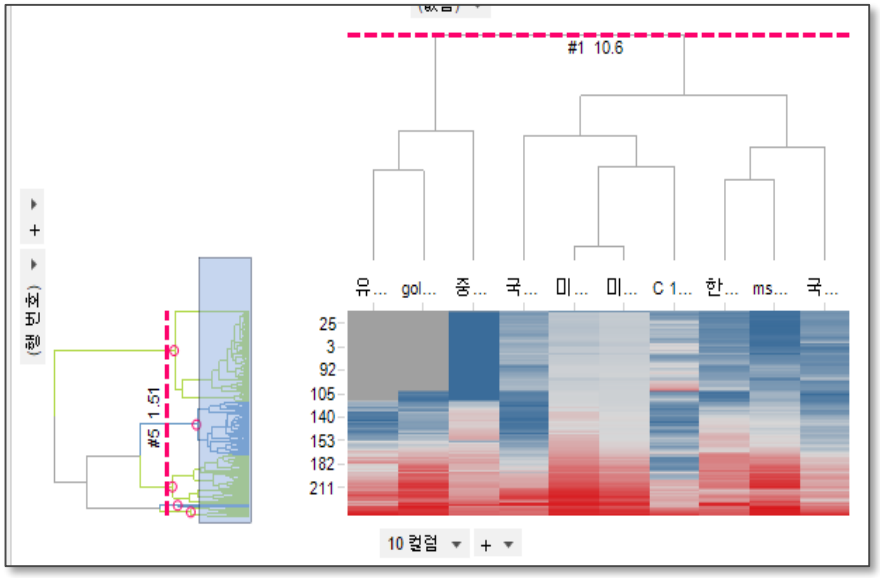
20. 히트맵의 좌측에 빨간색 점선으로 표시되어 있는 선**(**자르기 라인 **; pruning)**을 좌/우측으로 이동하면 해당되는 라인들간의 거리가 표시되고**,** 분홍색 원의 위치가 변경되면서 클러스터링의 개수를 알 수 있다**.** 아래 그림에서는 총 **5**개의 그룹이 존재하는 것을 알 수 있다**.(**총 **5**개의 분홍색 원**)**

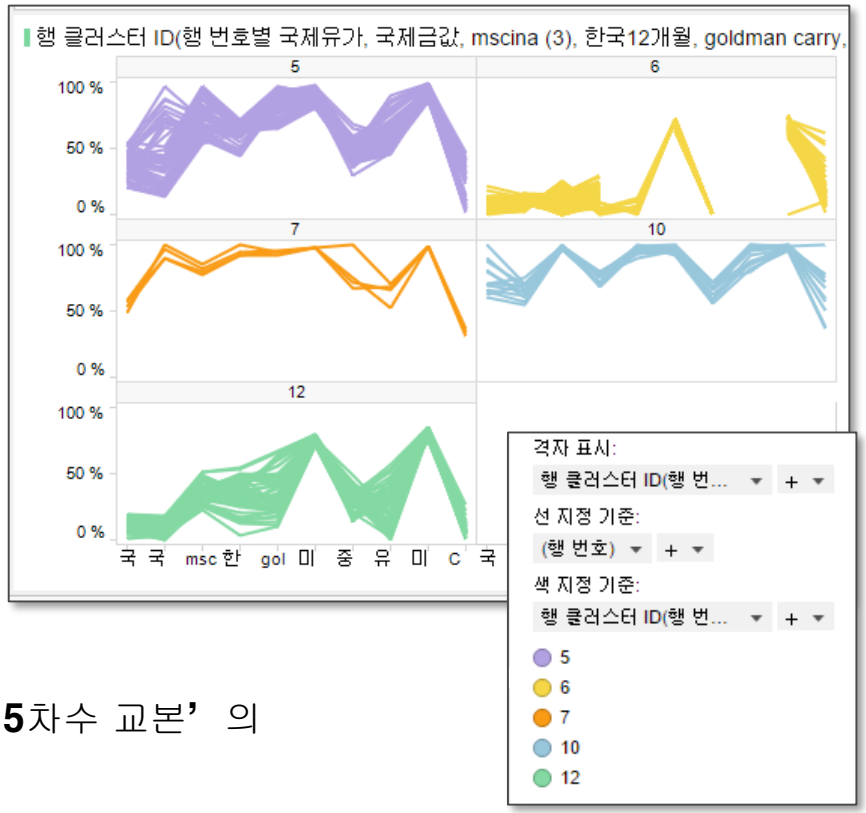


KOSPI Data.xls

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 나눠져서 표시된다**.(**아래 우측 그림**)**





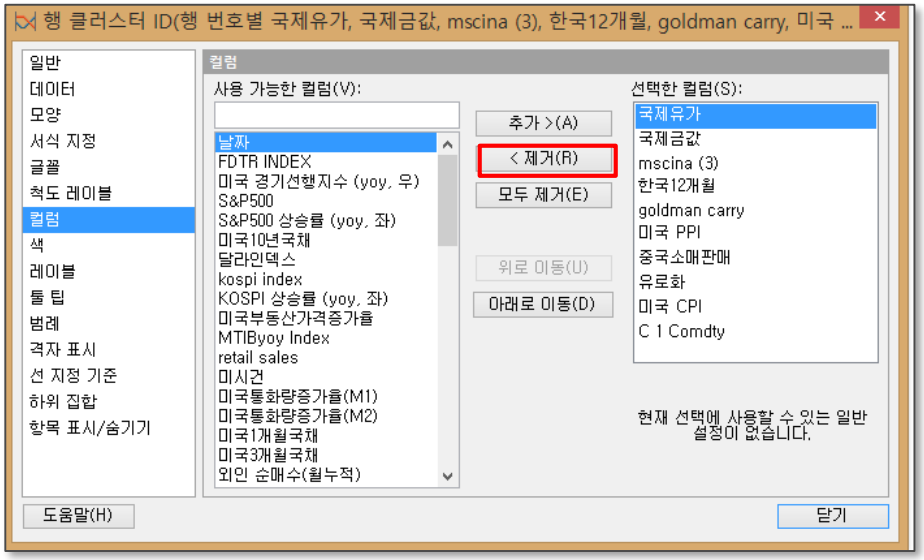
* **Node ID** 번호**, pruning level,** 계통수에 대한 설명은 **‘5**차수 교본**’** 의 **10page**를 참조 바람**.**

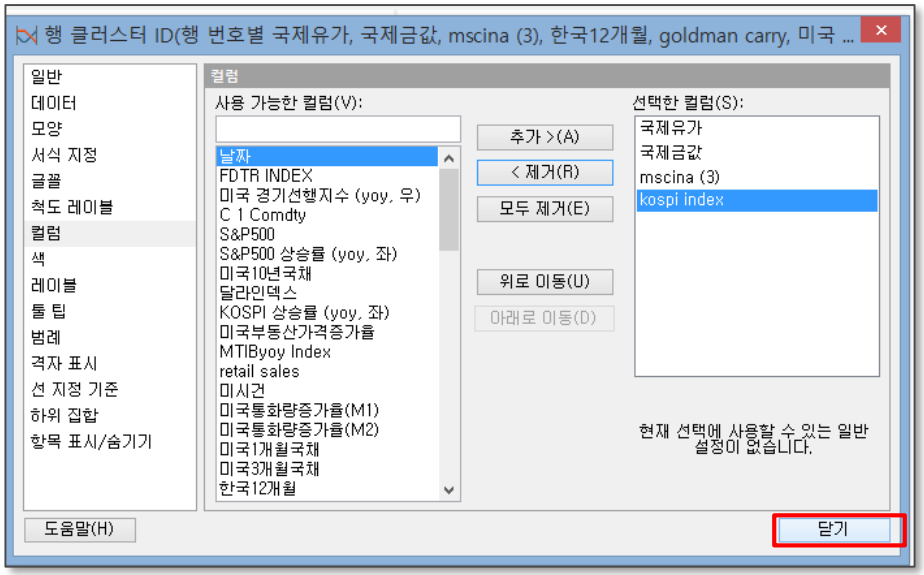
KOSPI Data.xls

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 **'kospi index** 를 추가하였다.**)**

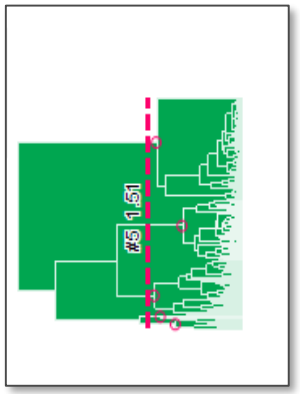


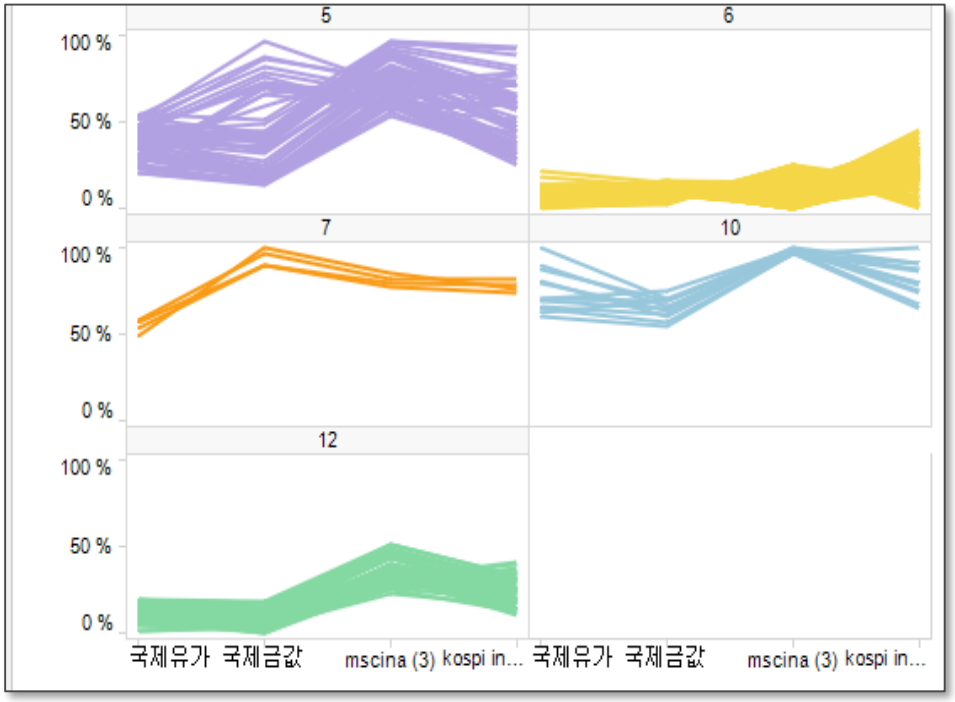


KOSPI Data.xls

1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

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





KOSPI Data.xls

1. 계층적 군집 분석(Hierarchical Clustering) 따라하기

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