6-1. 기초 통계

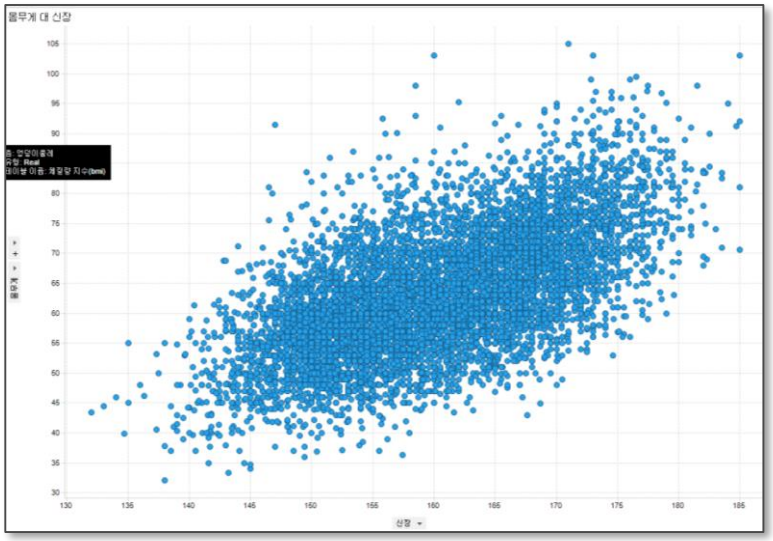
라인 및 곡선/데이터 상관성 분석/K평균 군집분석

김 성 기

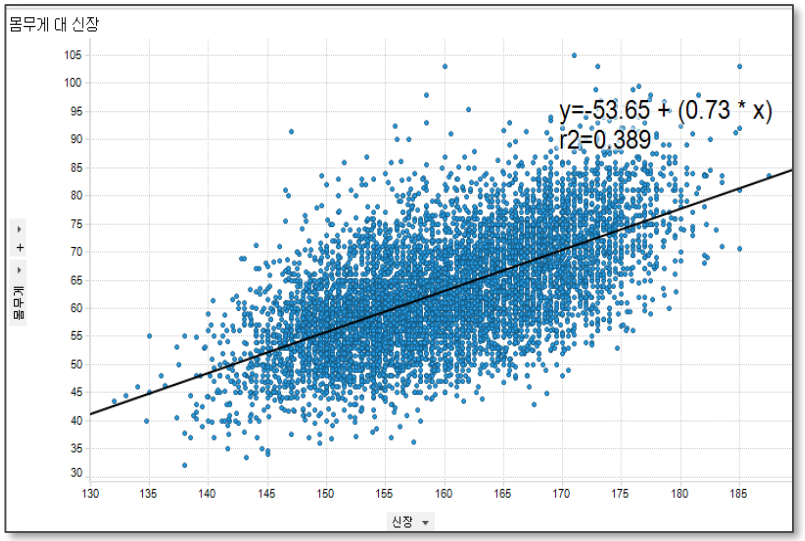
목 차

1. 라인 & 곡선(Lines & Curves)

- 필요에 따라서 사용자가 만들어 놓은 시각화(차트)에 참조 선이나 기본적인 통계수치를 추가로 표시해야 하는 경우가 있다. Spotfire에서 이러한 추가 정보들은 해당 시각화 유형의 속성 창에서 설정할 수 있다.
- 예를 들어 데이터 포인트가 특정 다항식 곡선 맞춤 또는 로지스틱 회귀 곡선 맞춤에 얼마나 잘 들어 맞는지를 확인할 수 있다.



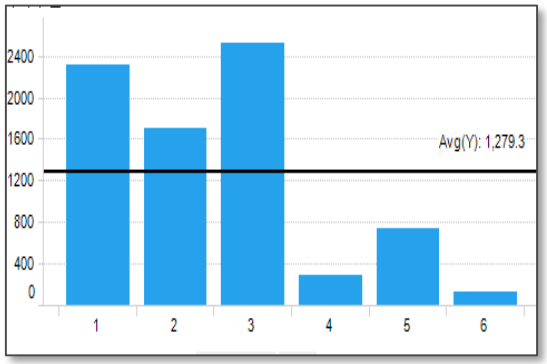
아무런 정보를 표시하지 않은 경우

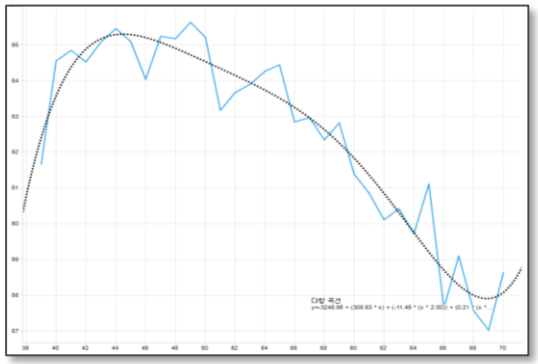


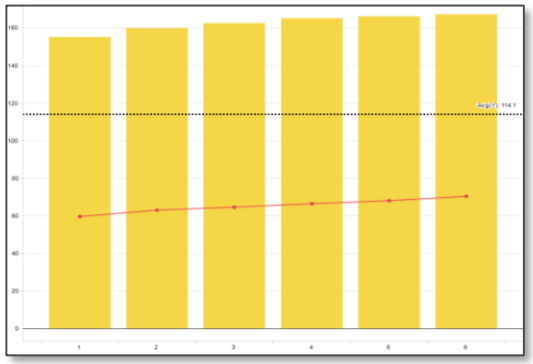
'라인 **&**곡선' 을 이용하여 직선 맞춤 **(Straight Line Fit)** 선과 관련 정보를 표시하여 놓은 경우

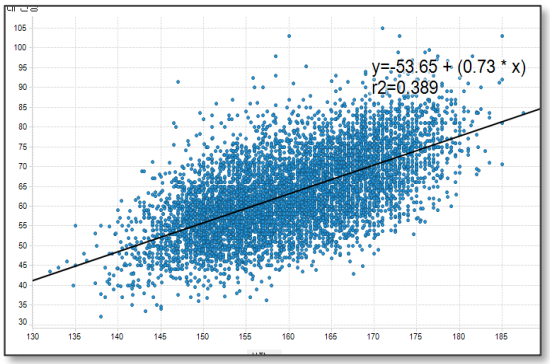
1. 라인 & 곡선(Lines & Curves)

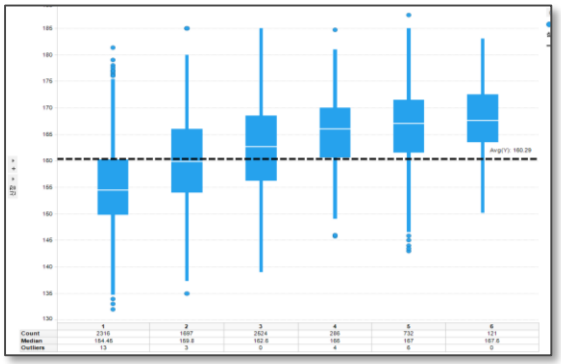
- 시각화의 종류와 데이터의 타입에 따라서 라인 및 곡선 정보를 표시하는 경우가 제한될 수 있다.
- 현재 막대 그래프, 선 그래프, 콤비네이션 차트, 산점도 및 상자 그래프 시각화에서 라인을 그릴 수 있다.





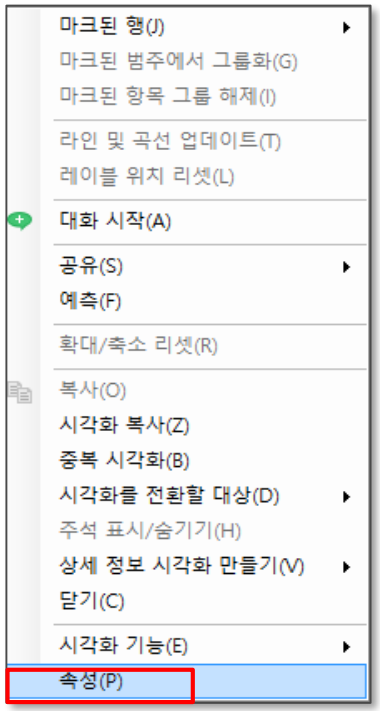


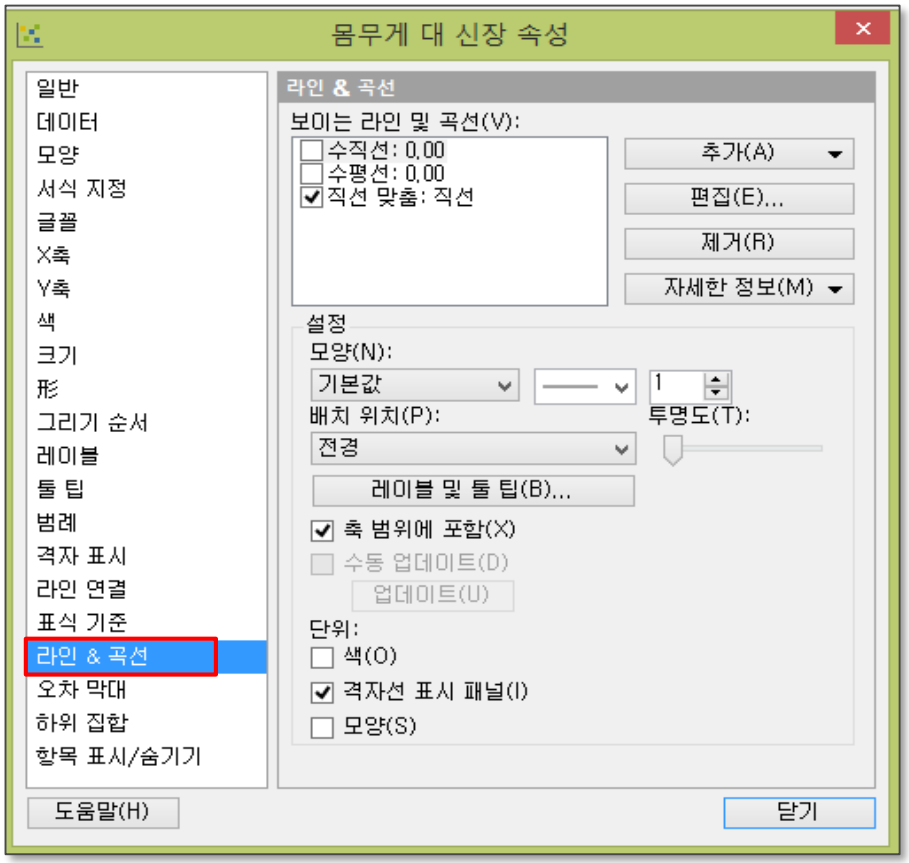




1. 라인 & 곡선(Lines & Curves)

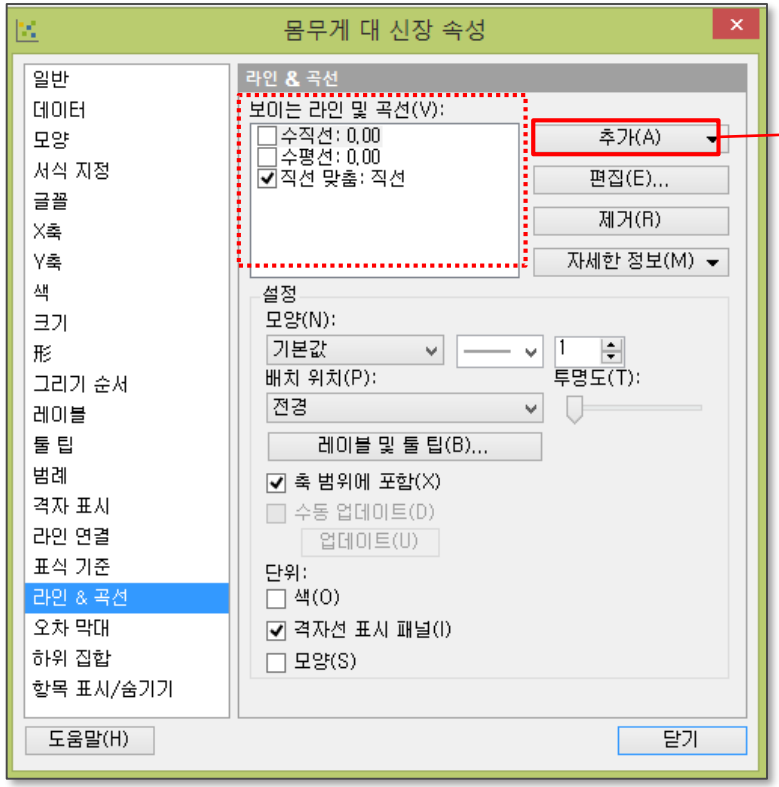
- Spotfire의 시각화에서 마우스 우클릭 > '속성' > '라인 & 곡선' 메뉴를 통해서 이용할 수 있다.

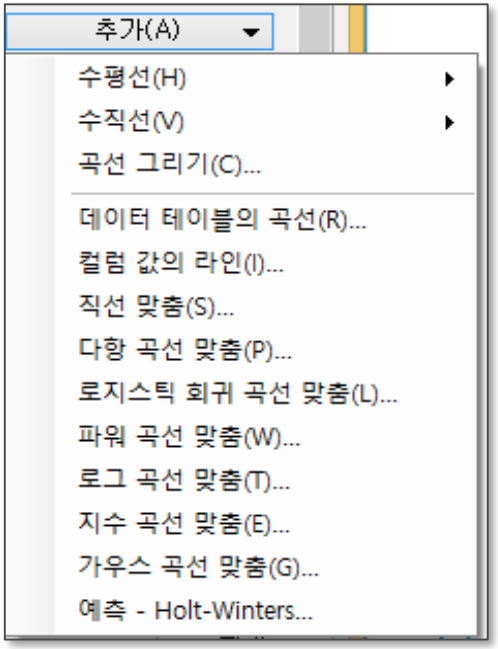




1. 라인 & 곡선(Lines & Curves) - 추가

- Spotfire에서 제공되는 Line & Curve 메뉴는 시각화 종류마다 기본적으로 거의 유사하지만, 기본적으로 첫 화면에서 보여지는 메뉴는 조금씩 다르다.

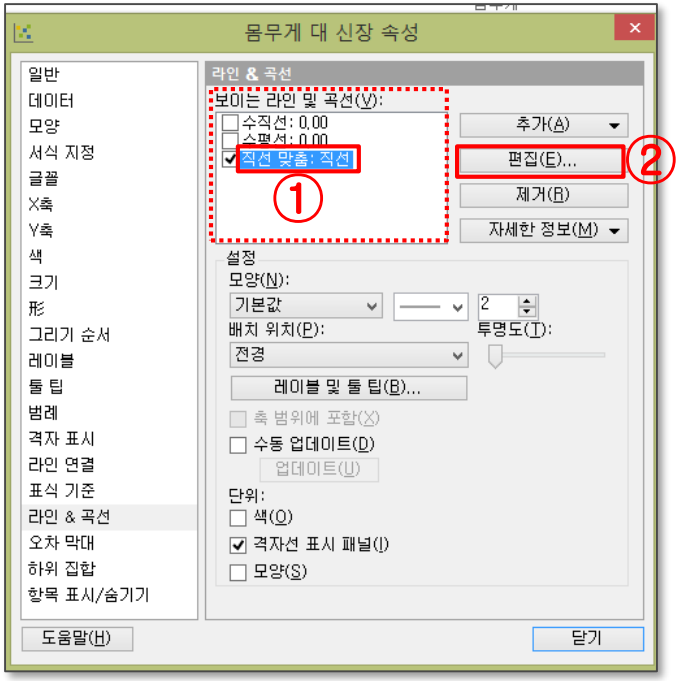


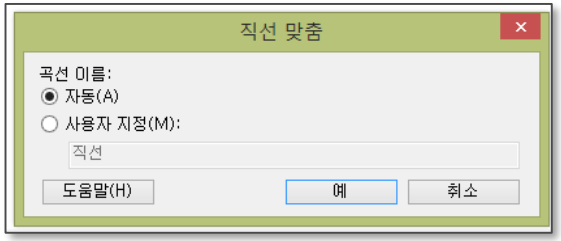


'보이는 라인 및 곡선' 창에 원하는 라인이 없으면 여기에서 원하는 라인을 선택하여 추가할 수 있다.

1. 라인 & 곡선(Lines & Curves) – 편집

- '보이는 라인 및 곡선' 의 편집(수정) 방법
  - 여기서 꼭 명심해야 할 부분은 먼저 '보이는 라인 및 곡선' 창의 여러 곡선들 이름 중에서 편집을 원하는 '라인 및 곡선'을 선택하고 나머지 작업을 시작해야 한다는 것이다. 즉, '라인 및 곡선'을 선택하게 되면 아래 그림과 같이 파란색으로 표시(①)가 된다.
  - 이제 '편집' 버튼을 누르게 되면(②) 각 '라인 및 곡선'이 특성에 따라 다양한 편집 화면이 표시되게 된다.

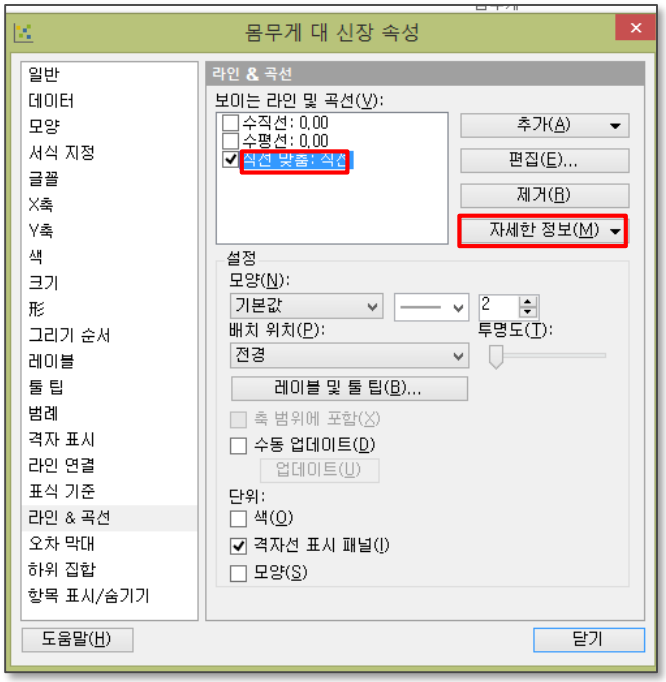


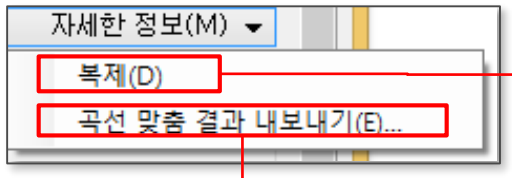


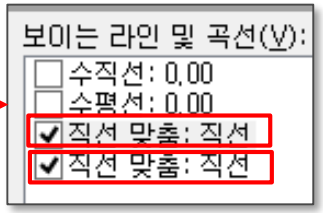
- '직선 맞춤' 의 경우에는 '직선'의 표시 이름을 편집할 수 있는 설정 화면만 나타난다.
- '사용자 지정' 버튼 선택시 사용자가 원하는 이름으로 직선의 이름을 변경할 수 있다.

1. 라인 & 곡선(Lines & Curves) – 레이블 정보 파일로 저장

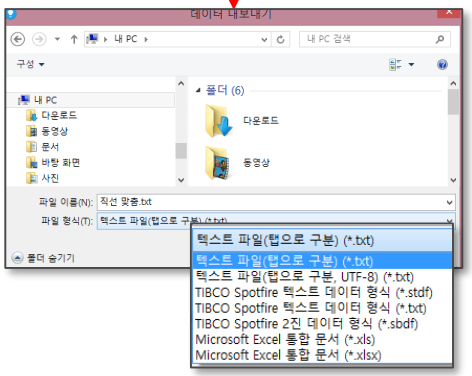
- 이미 작성한 '라인 & 곡선'을 복제하여 쉽게 또 다른 유사한 '라인 & 곡선' 을 만들 수 있다.
- '라인 & 곡선' 을 통해서 얻은 관련 정보들[예, 직선 맞춤(Straight Line Fit)의 경우 : 직선의 기울기, Y절편값, 상관계수(R^2)]을 .txt나 .xls등의 파일로 저장할 수 있다.



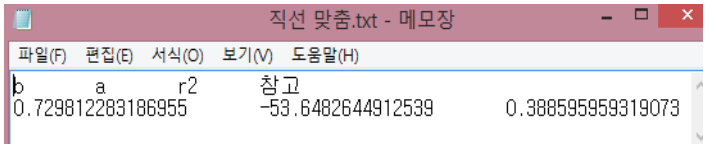




동일한 '라인 & 곡선' 을 복사하여 추가한다.

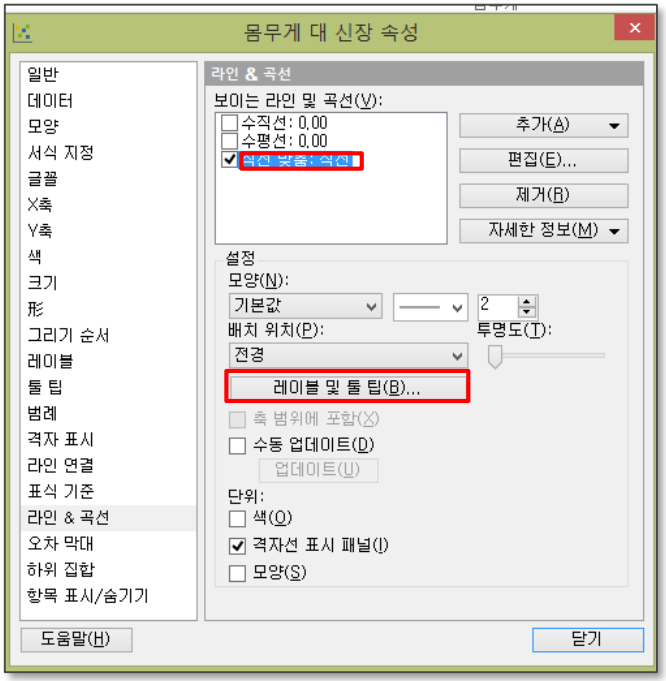


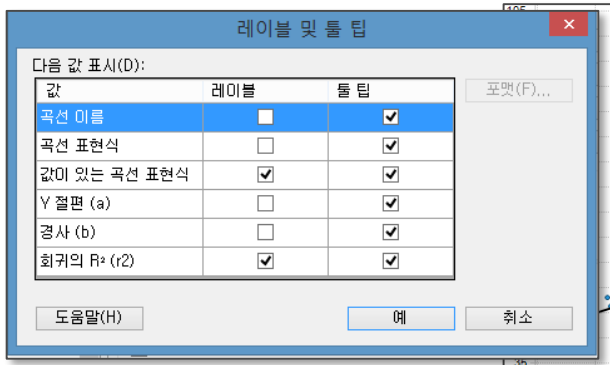
기울기(b), 절편값(a), 상관계수(r2) 등 통계적 수치가 포함된 정보들을 별도의 문서 파일로 저장할 수 있게 한다.

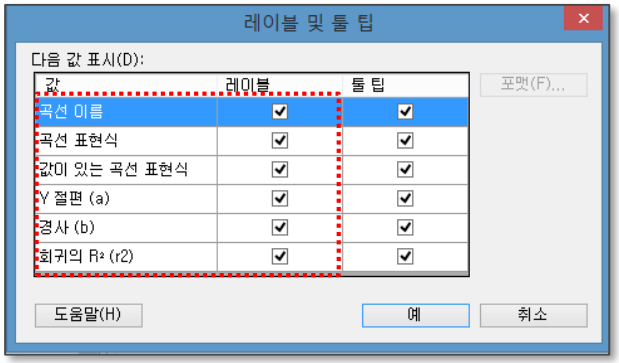


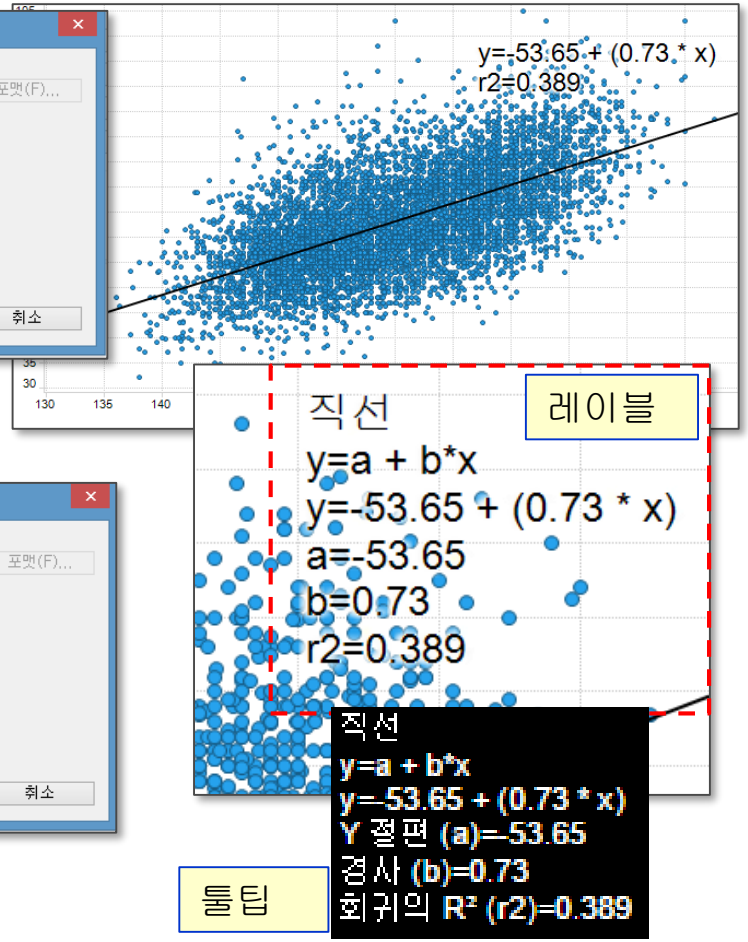
1. 라인 & 곡선(Lines & Curves) – 레이블/툴팁의 표시 정보 추가/변경

- 작성한 '라인 & 곡선'에 대하여 시각화에 나타내는 레이블(화면의 직선 근처에 표시될) 정보들을 편집, 수정할 수 있다.



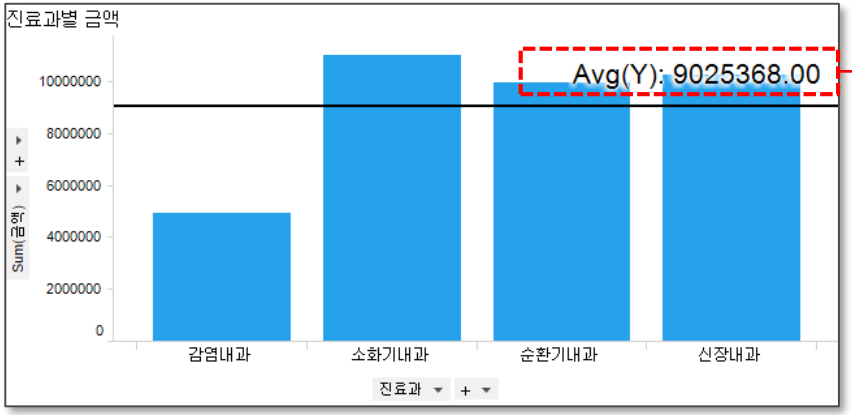




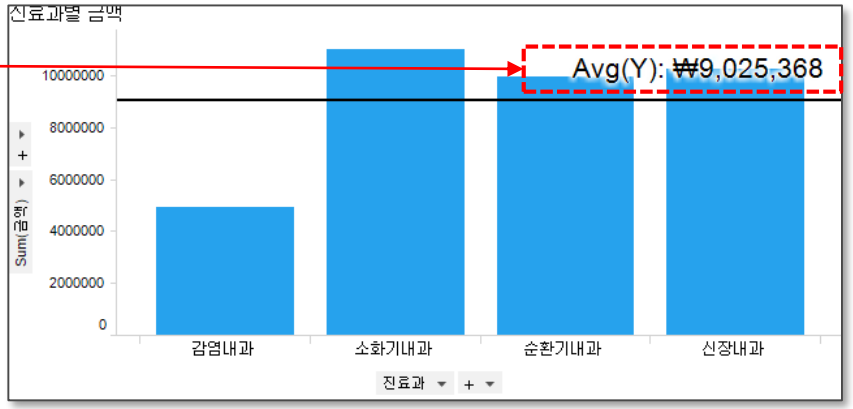


1. 라인 & 곡선(Lines & Curves) – 레이블의 서식 변경

- 표시된 레이블의 형식(예, 소수점 아래 자리수 등)을 변경할 수 있다.



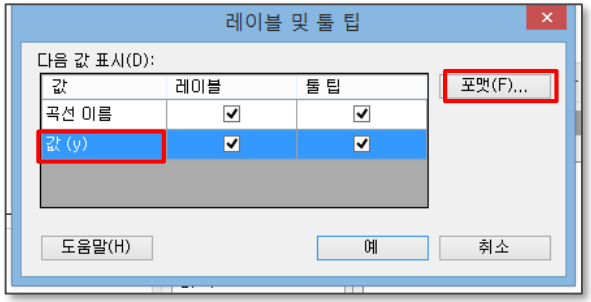
기본 설정으로 표시된 레이블

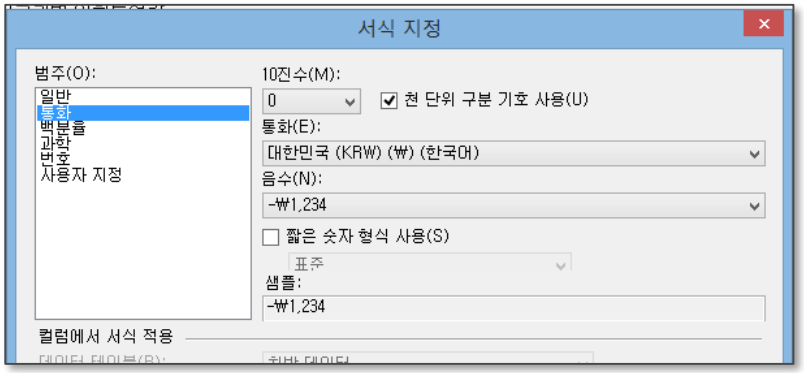


설정 변경 후 표시된 레이블

서식 변경 방법 :

마우스 우클릭 > '속성' > '라인 & 곡선' > '보이는 라인&곡선'에서 라인 선택 > '레이블 및 툴팁' > 원하는 '값(y)' 선택 > '포맷' 클릭

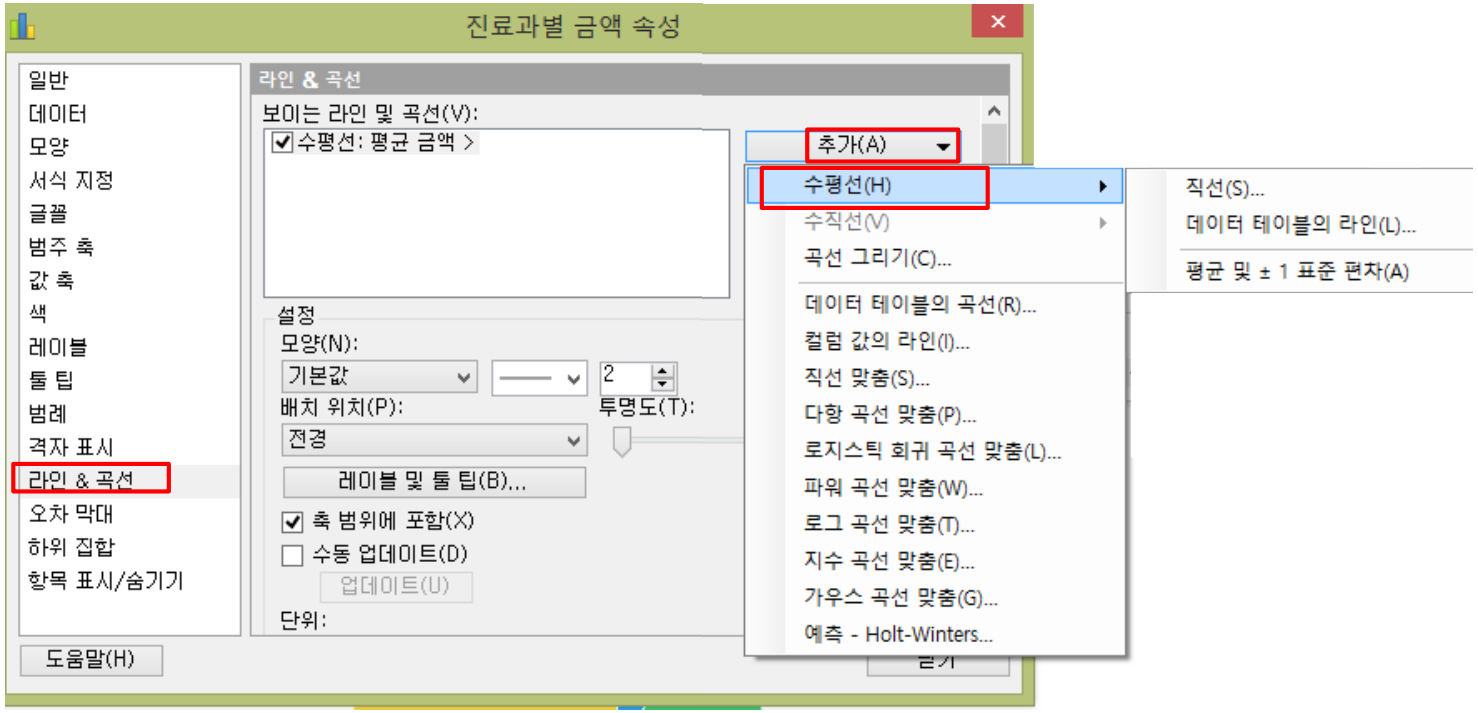




1. 라인 & 곡선(Lines & Curves) – 수평선 설정 추가

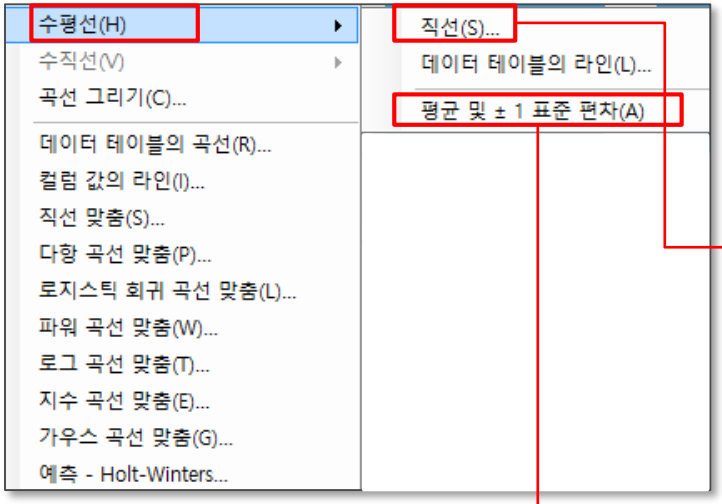
- 시각화에 새로 수평선(Horizontal Line)을 추가할 수 있다.

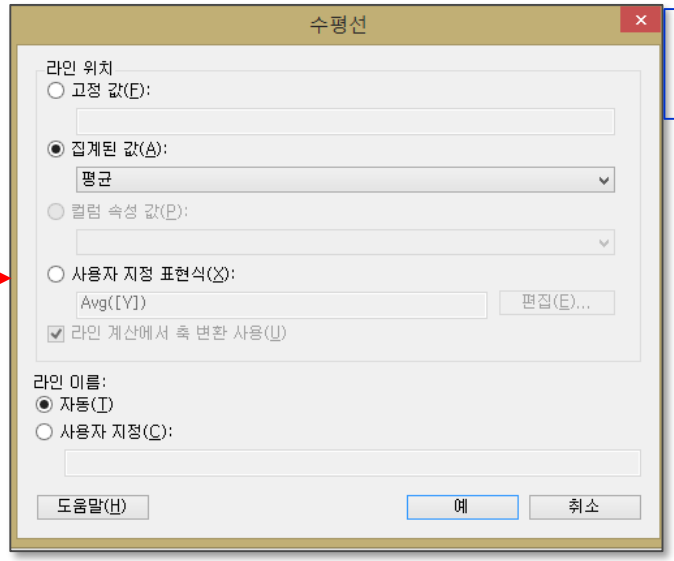
수평선 추가 방법 : 마우스 우클릭 > '속성' > '라인 & 곡선' > '추가' > '수평선' 선택



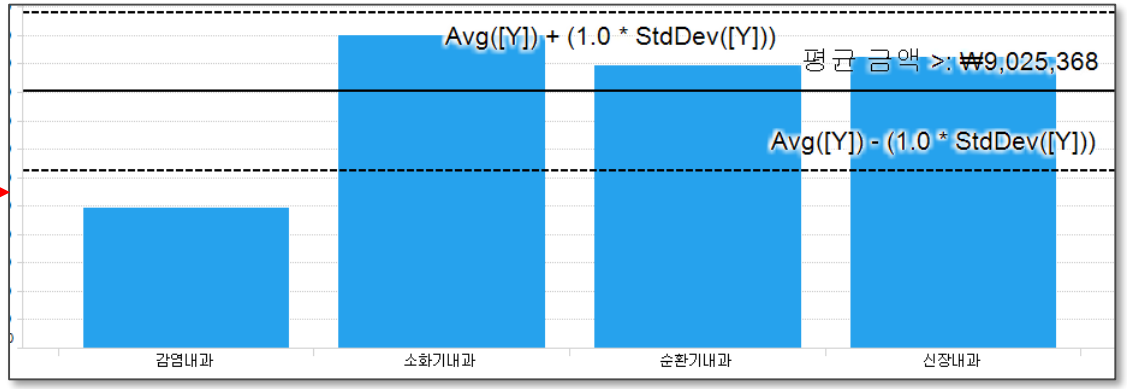
1. 라인 & 곡선(Lines & Curves) – 수평선 설정 추가

- 수평선(Horizontal Line)을 추가하는 몇 가지 방법이 있다.



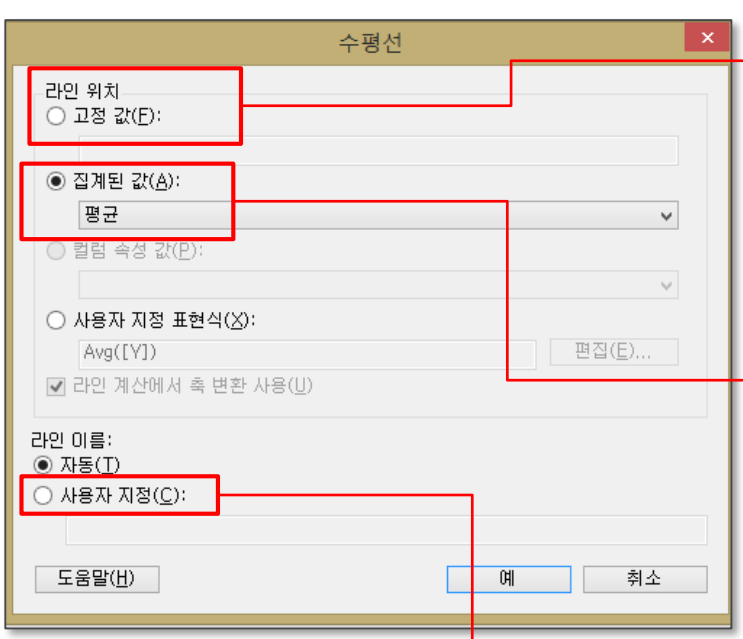


다음 페이지에 설정관련 설명이 나와 있다.

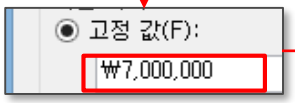


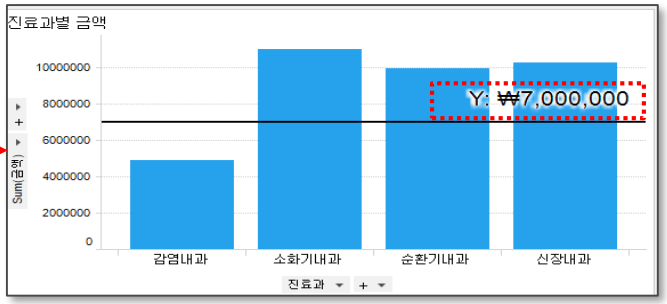
1. 라인 & 곡선(Lines & Curves) – 수평선 설정 방법

- '수평선'에 대하여 특성을 설정할 수 있다.

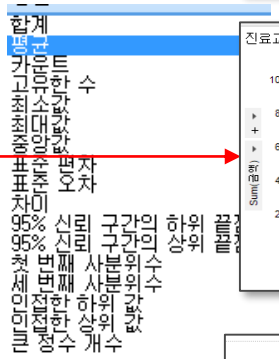


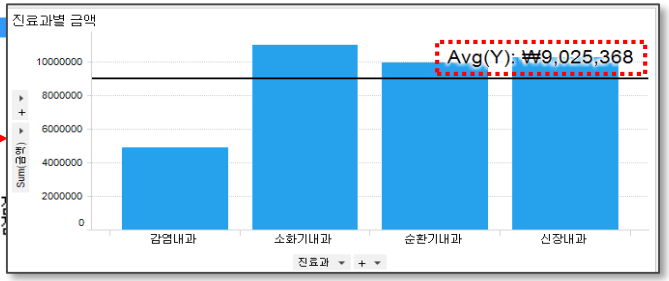
사용자가 지정하는 값의 위치에 선을 표시한다.



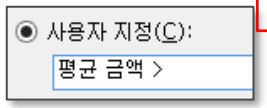


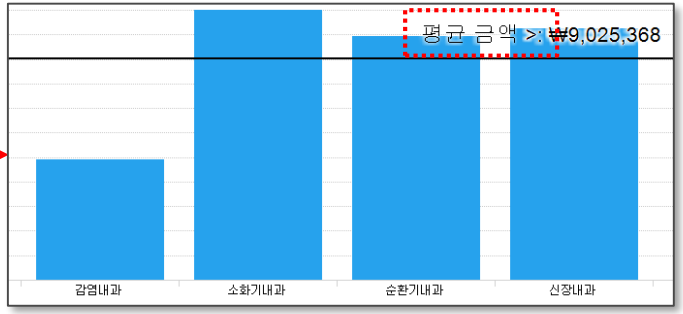
수평선의 집계 방법을 선택한다.





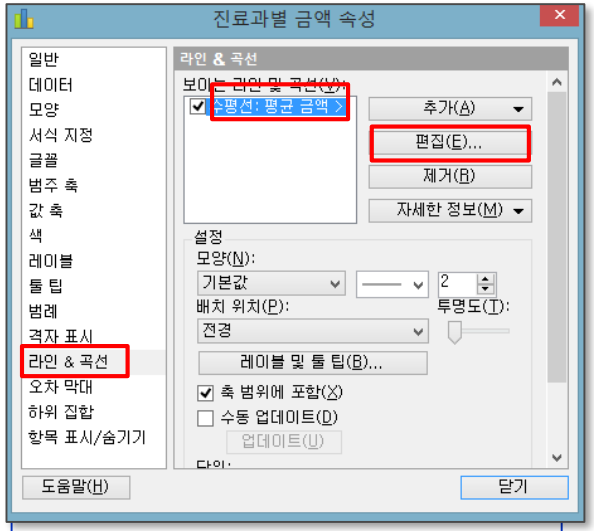
사용자가 원하는 이름으로 수평선의이름을 지정한다.

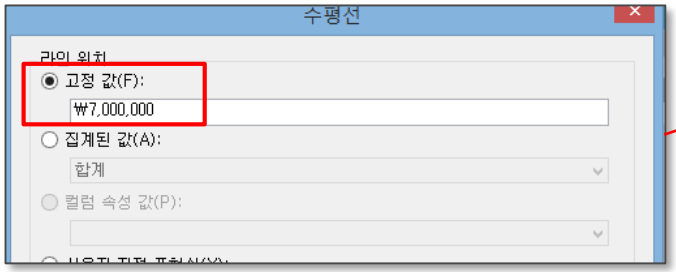


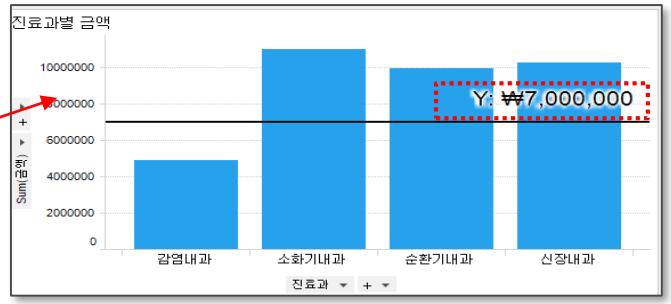


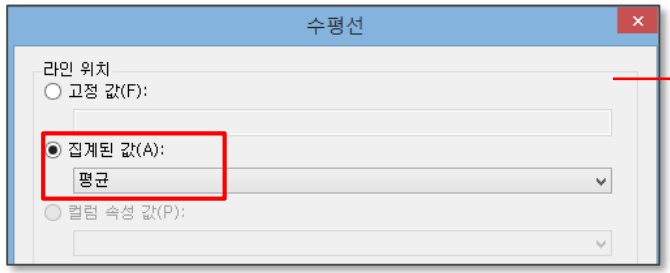
1. 라인 & 곡선(Lines & Curves) – 수평선 설정 변경 방법

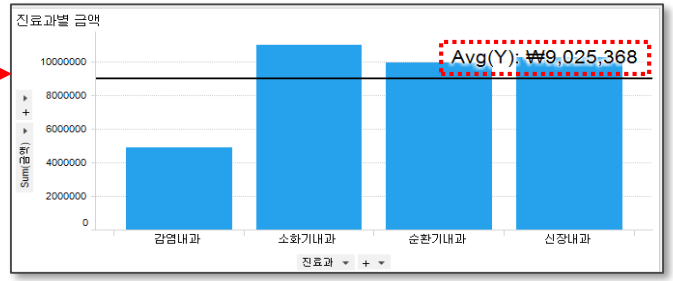
- 이미 작성해 놓은 레이블의 표시 형식(예, 소수점 아래 자리수 등)을 변경할 수 있다.











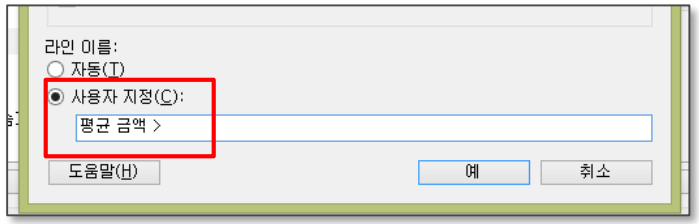
수평선 서식 변경 방법 :

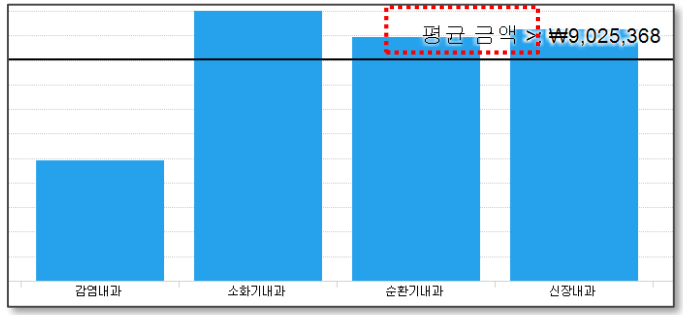
마우스 우클릭 > '속성'

> '라인 & 곡선'

> '보이는 라인&곡선'에서 '수평선: ...' 라인 선택

> '편집' 클릭

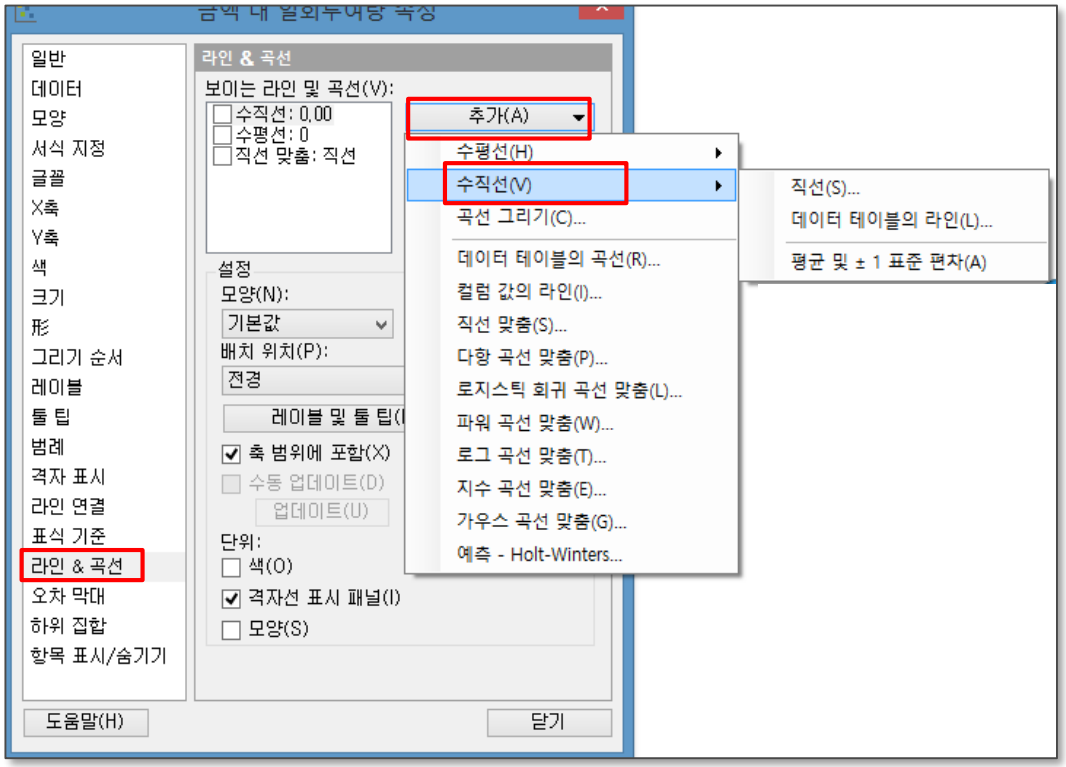




1. 라인 & 곡선(Lines & Curves) – 수직선 설정

- 시각화에 새로 수직선(Horizontal Line)을 추가할 수 있다.

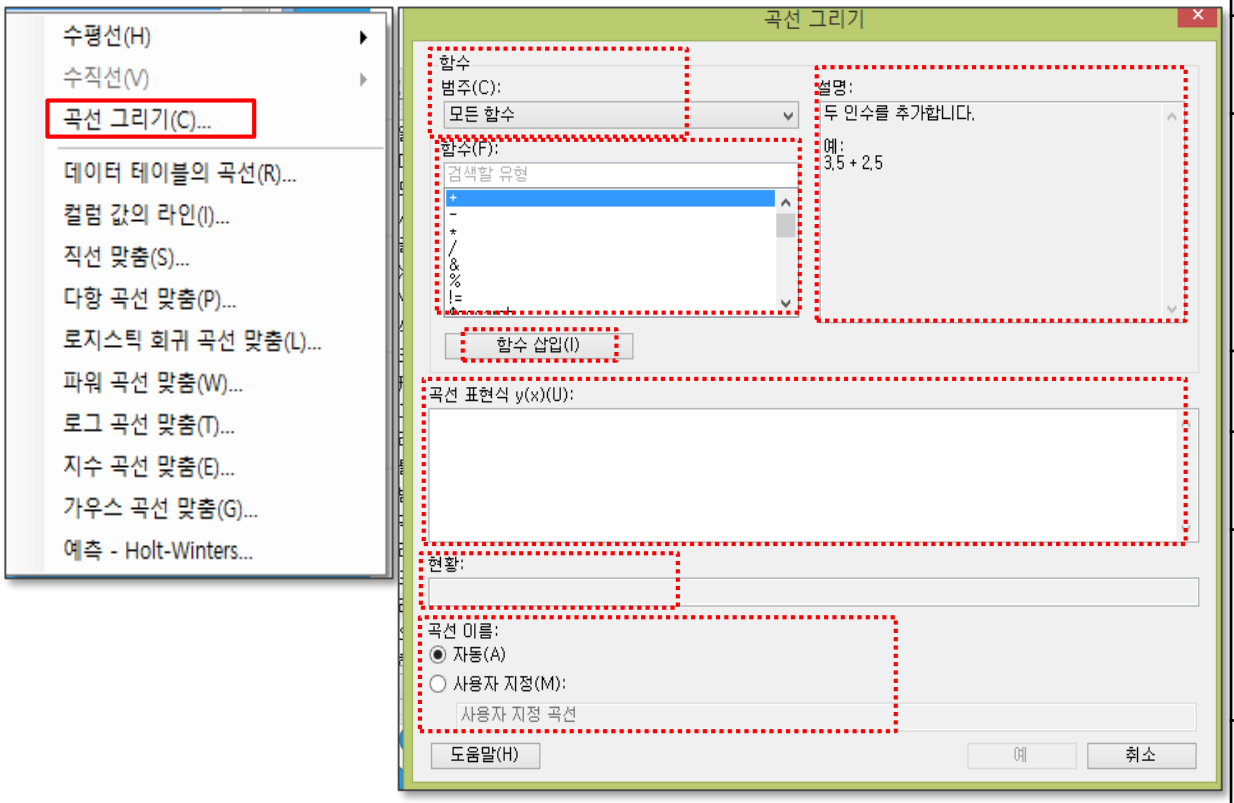
수직선 추가 방법 : 마우스 우클릭 > '속성' > '라인 & 곡선' > '추가' > '수직선' 선택



* 수직선의 관련 설정 방법은 '수평선'의 설정방법과 완전히 동일하므로 '수평선'의 설정방법을 참조한다.

1. 라인 & 곡선(Lines & Curves) – 곡선 그리기 설정

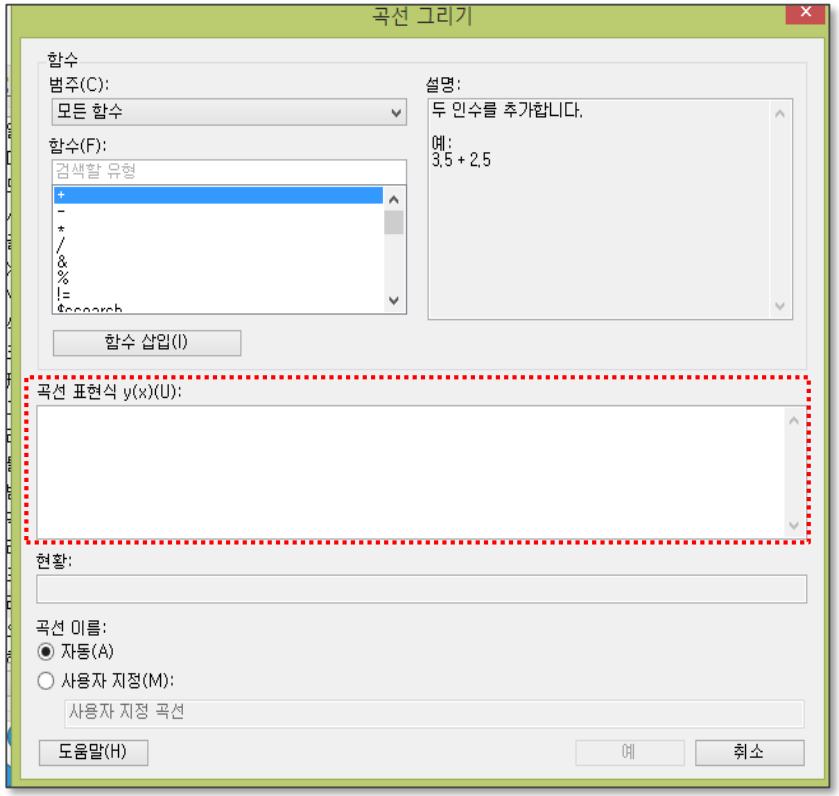
- 시각화에 새로운 곡선을 추가(Curve Draw) 할 수 있다.

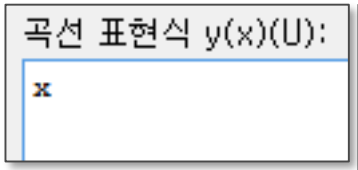


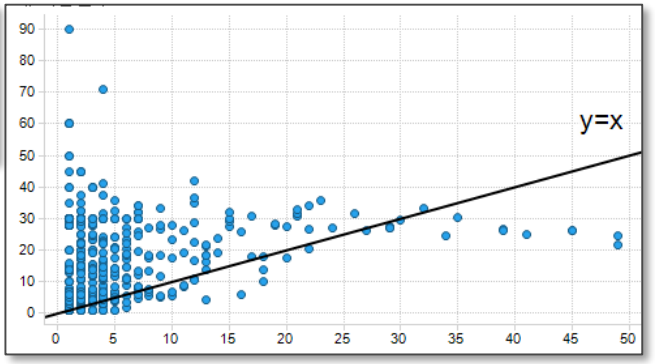
옵션	설명
범주 (Category)	함수의 범주를 선택하여 함수 목록의 선택 옵션을 제한한다.
함수 (Function)	텍스트 필드에 검색 문자열을 입력하여 함수 목록의 항목 수를 제한할 수 있다. 함수를 클릭하고 원하는 함수 이름의 첫 글자를 입력하여 목록의 특정 위치로 바로 이동할 수 있다. 여기에서 원하는 함수를 선택한다.
설명 (Description)	선택한 함수에 대한 간략한 설명이 표시.
함수 삽입	곡선 표현식 필드의 현재 커서 위치에 선택된 함수를 삽입.
곡선 표현식 y(x)	표현식 필드는 표현식을 작성하는 텍스트 필드. 목록에서 함수를 삽입하거나 표준 텍스트 편집기에서와 같은 방법으로 텍스트를 입력한다.
현황 (Status)	현재 표현식의 상태를 표시한다. 이 필드에 오류가 표시되는 경우 표현식에 문제가 있는 것이다.
곡선 이름	자동 곡선 이름을 만들지 또는 **사용자 지정** 곡선 이름을 입력할지 여부를 지정.

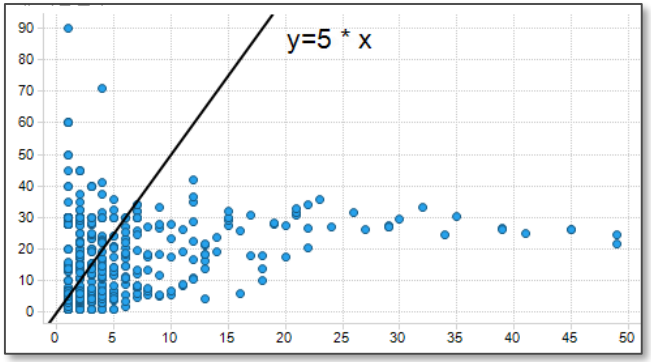
1. 라인 & 곡선(Lines & Curves) – 곡선 그리기 설정

- 곡선 그리기(Curve Draw)의 '**곡선 표현식**' 사용 예 :



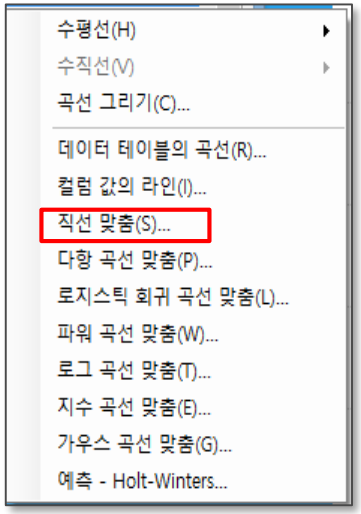


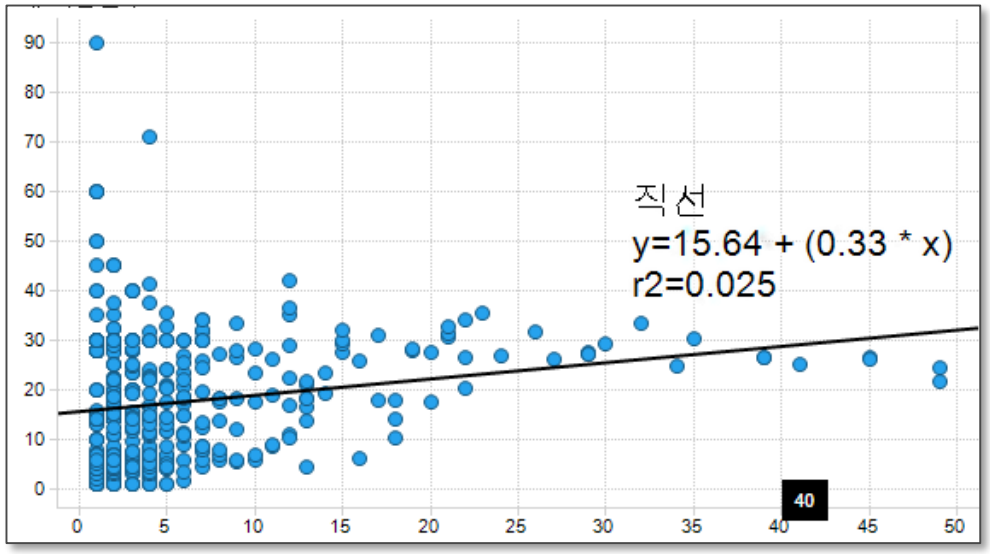




1. 라인 & 곡선(Lines & Curves) – 직선 맞춤(straight Line Fit)

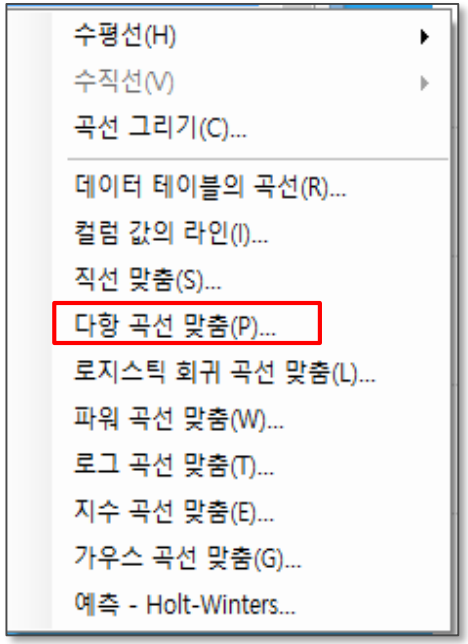
- '직선 맞춤'은 일련의 데이터 요소에 가장 잘 맞는 직선을 구성하는 프로세스이다.
- **Spotfire**에서는 데이터를 가장 잘 표현할 수 있는 직선과 관련 정보들을 제공한다.
  - **Straight line**의 정보 중에서 가장 중요한 정보는 '회귀의 **R^2(상관계수)**' 이다.
  - 적합도는 **R^2** 값으로 표시되는데, 값이 **R^2=1.0**이면 완벽한 맞춤을 나타내고, **R^2=0.0**이면 회귀 모델이 이 데이터 형식에 적합하지 않음을 나타낸다. 일반적으로 **0.5** 이상이면 비교적 상관성이 존재하는 것으로 유추할 수 있다.
- 방정식 : $y = a + bx$

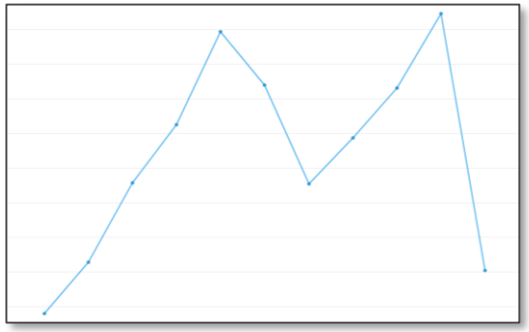




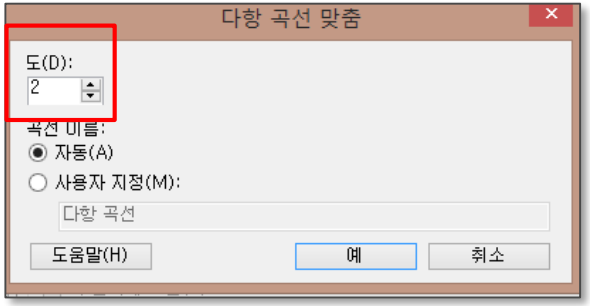
1. 라인 & 곡선(Lines & Curves) – 다항 곡선 맞춤 설정

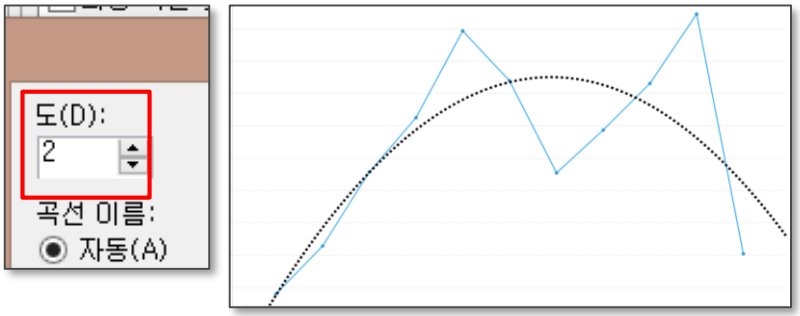
- 시각화에 다항 곡선 맞춤(Polynomial Curve Fit)를 추가로 표시할 수 있다.
- 방정식 : $y = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5$
  (최대 5차까지 표현 가능)

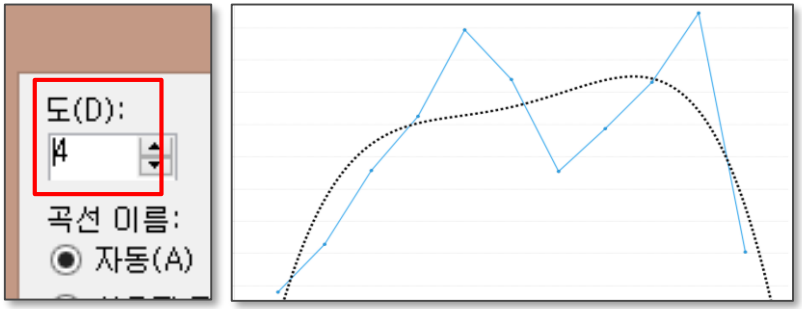


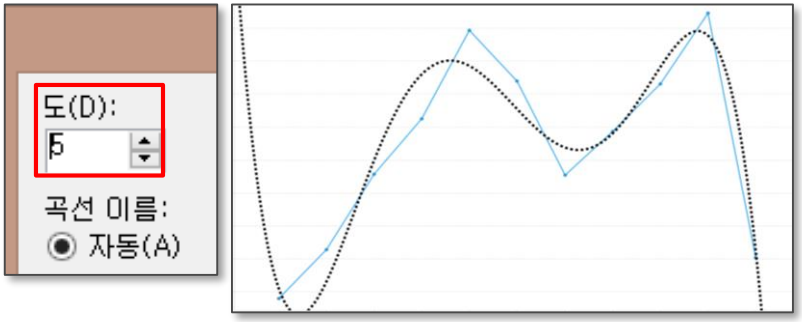


초기 선 그래프



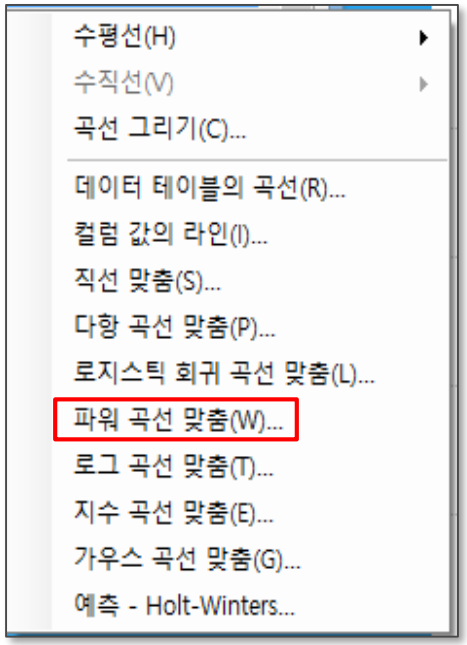


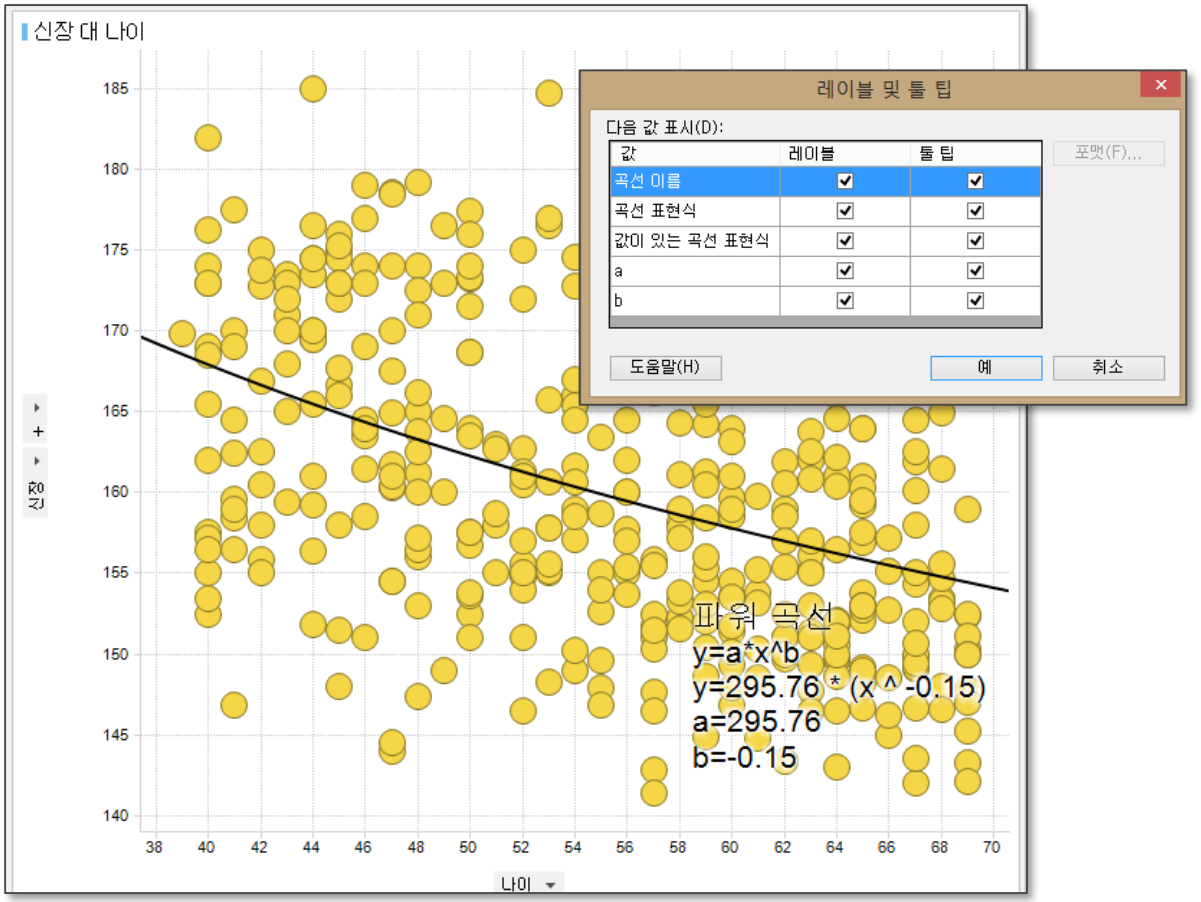




1. 라인 & 곡선(Lines & Curves) – 파워 곡선 맞춤 설정

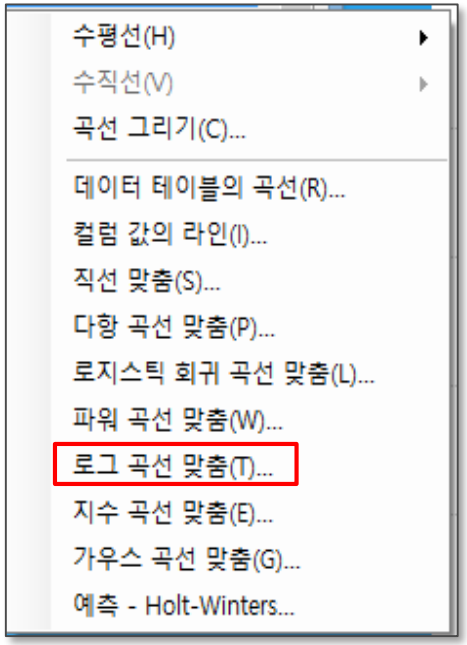
- 시각화에 파워 곡선 맞춤(Power Curve Fit)를 추가로 표시할 수 있다.
- 방정식 : $y = ax^b$

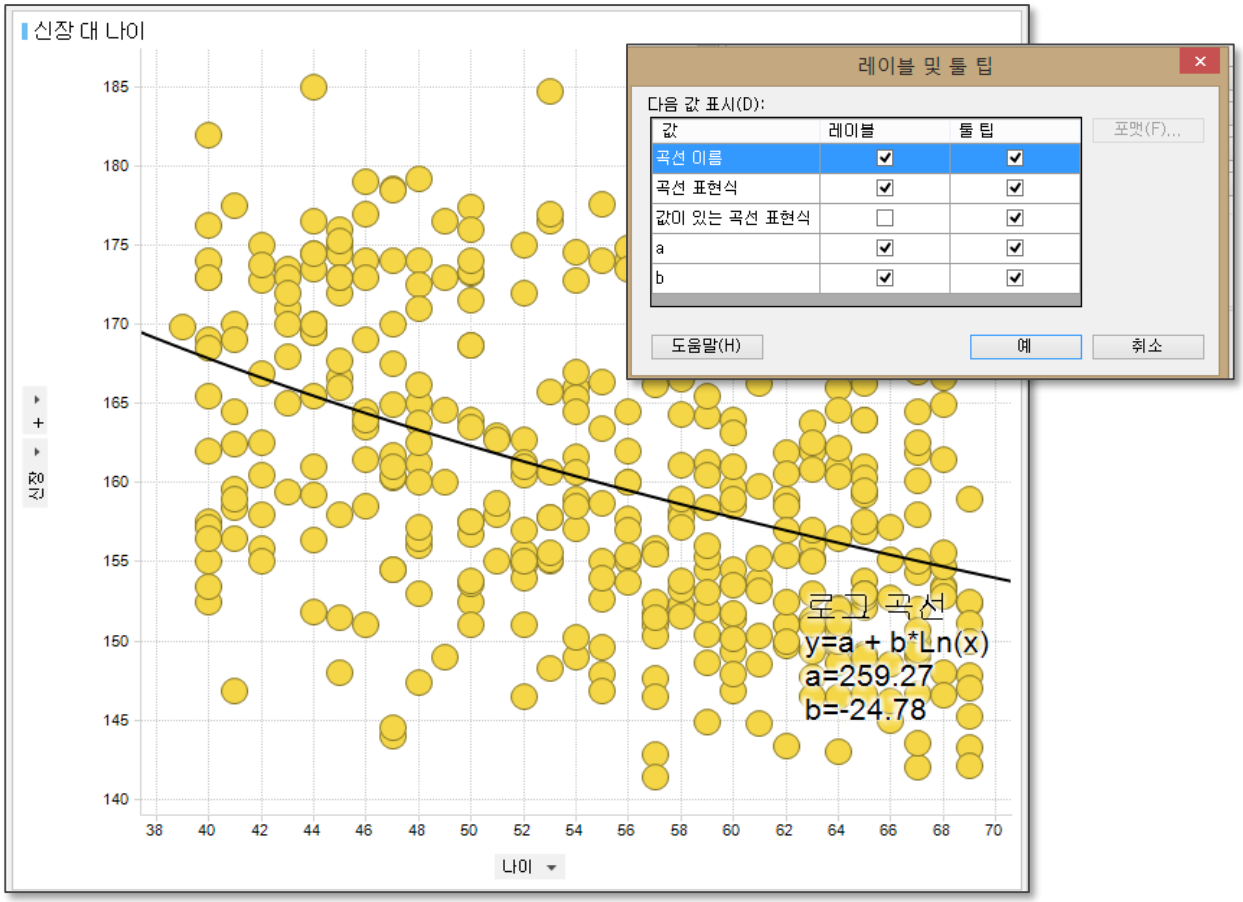




1. 라인 & 곡선(Lines & Curves) – 로그 곡선 맞춤 설정

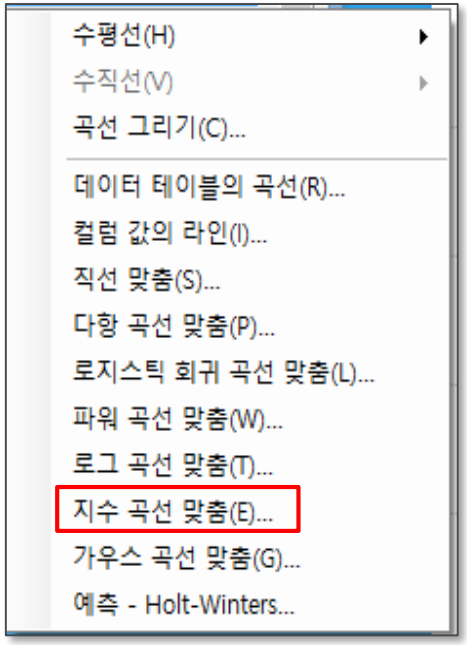
- 시각화에 로그 곡선 맞춤(Logarithmic Curve Fit)를 추가로 표시할 수 있다.
- 방정식 : $y = a + b \ln x$

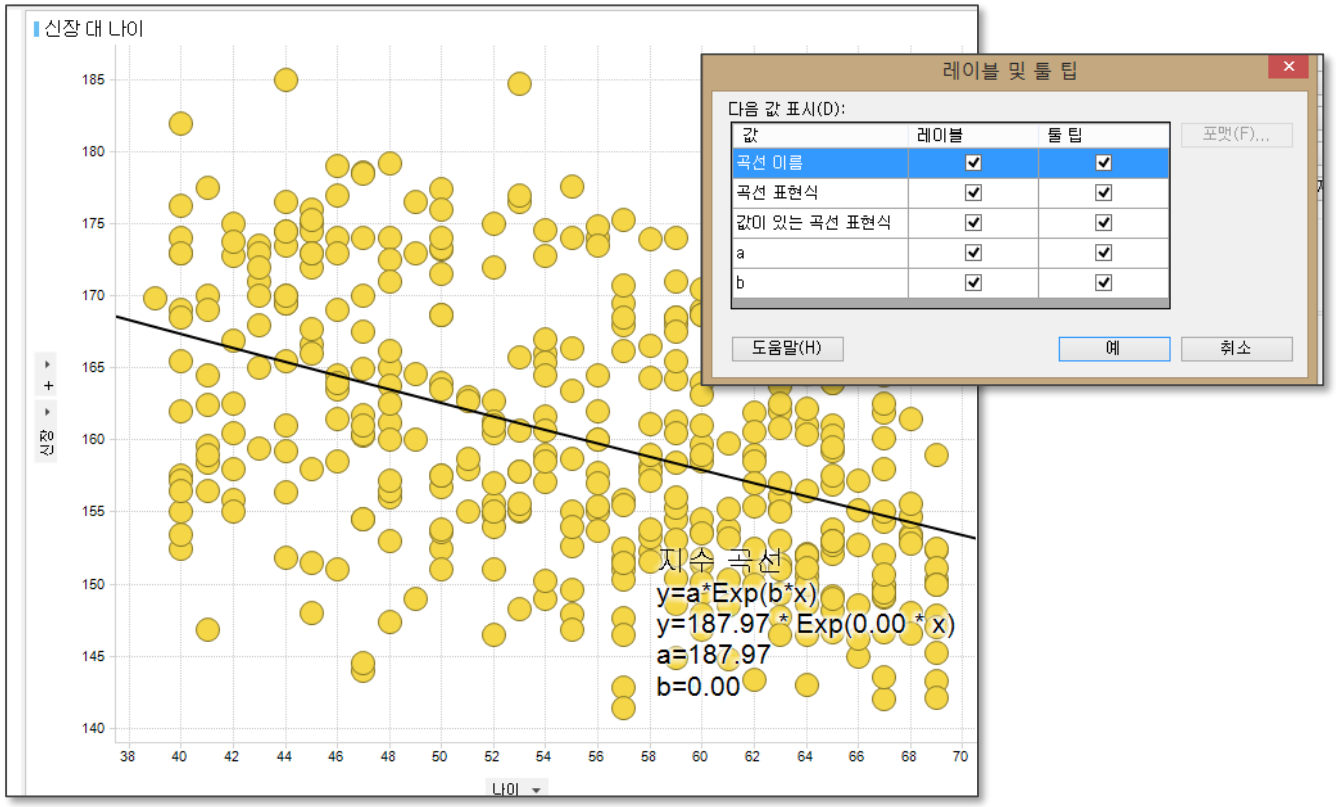




1. 라인 & 곡선(Lines & Curves) – 지수 곡선 맞춤 설정

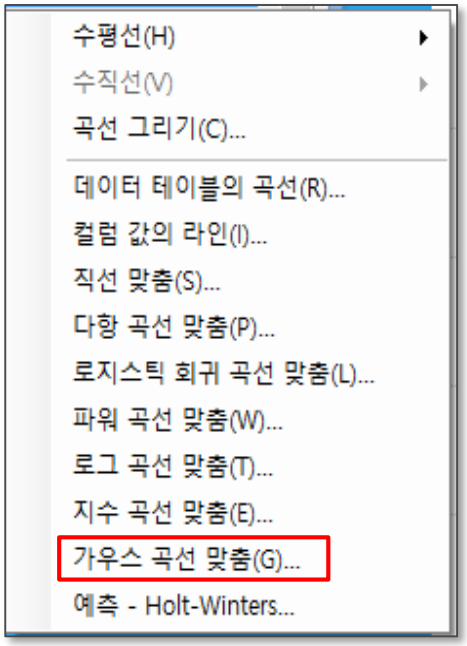
- 시각화에 지수 곡선 맞춤(Exponential Curve Fit)를 추가로 표시할 수 있다.
- 방정식 : $y = ae^{bx}$
- 지수 곡선은 예를 들어 박테리아의 기하급수적 성장에 대한 생물학적 응용에 일반적으로 사용된다.

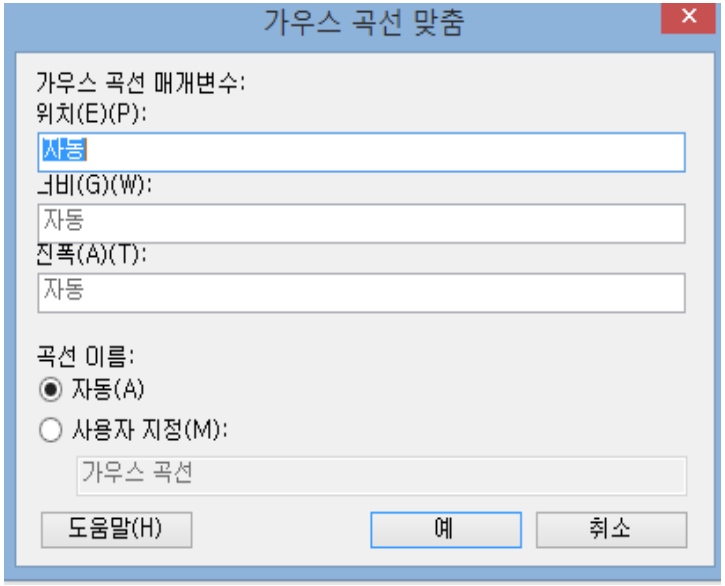




1. 라인 & 곡선(Lines & Curves) – 가우스 곡선 맞춤 설정

- 시각화에 가우스 곡선 맞춤(Gaussian Curve Fit)를 추가로 표시할 수 있다.
- 가우스 곡선 맞춤에서는 다음 방정식을 사용하여 정규 분포를 설명하는 데 적합한 종형 곡선을 계산한다.
- 방정식 : $y = A \cdot e^{-\left(\frac{x-E}{G}\right)^2}$



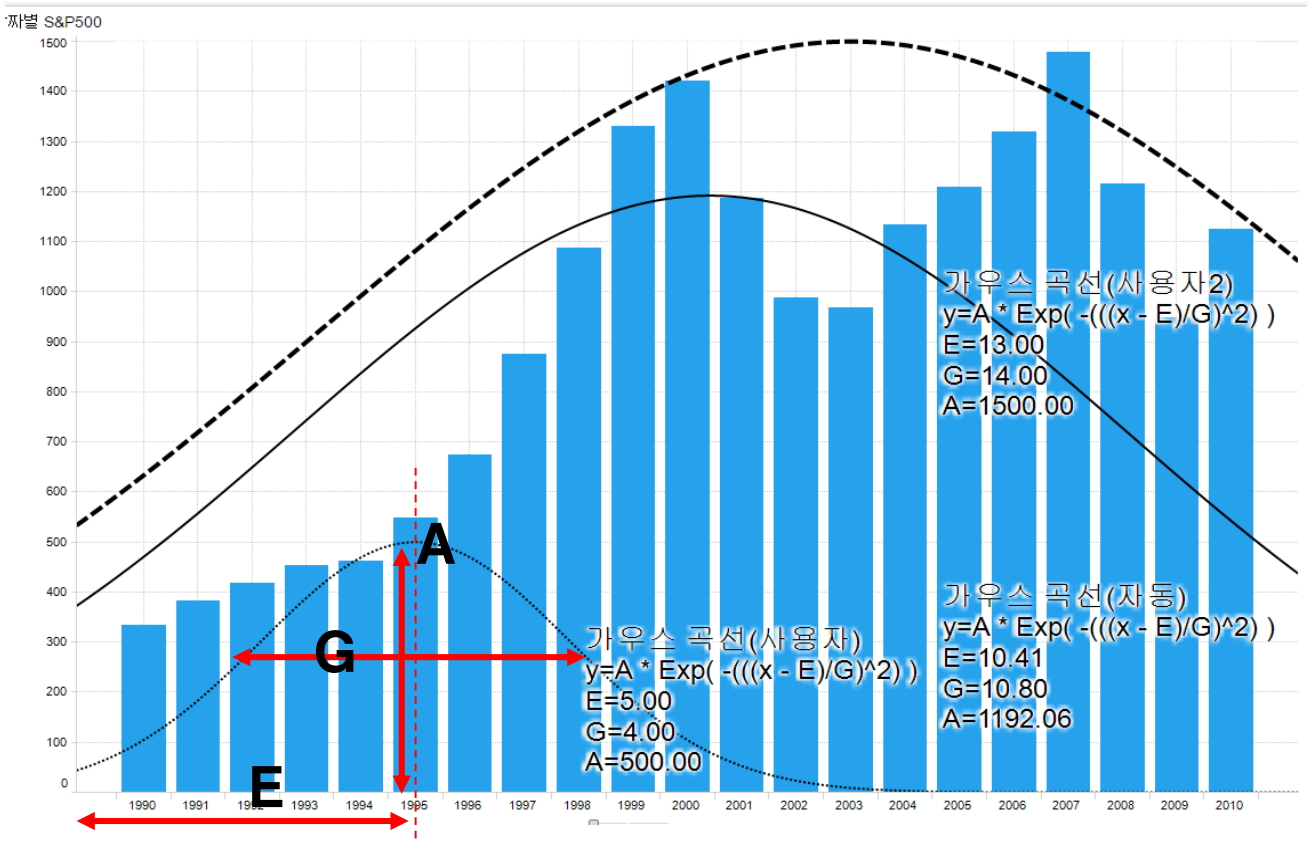


옵션	설명
위치(E) (Position (E))	가우스 분포 곡선에 대한 중심점의 위치를 지정할 수 있다.
너비(G) (Width (G))	가우스 분포 곡선의 너비를 지정할 수 있다.
진폭(A) (Amplitude (A))	가우스 분포 곡선의 고도(높이)를 지정할 수 있다.

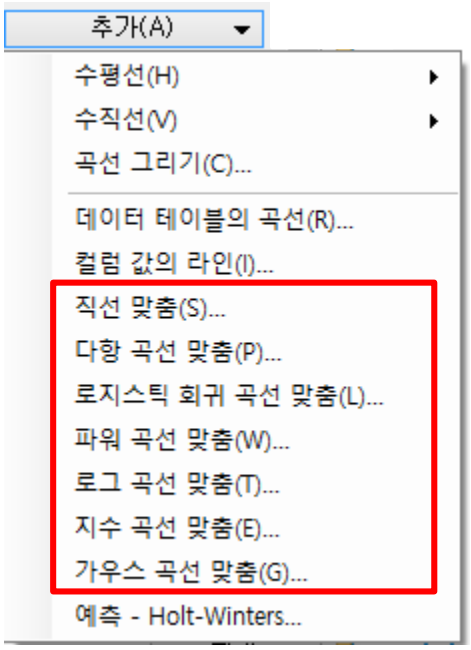
1. 라인 & 곡선(Lines & Curves) – 가우스 곡선 맞춤 설정

- Spotfire에서는 시각화에 사용자가 원하는 값으로 가우스 곡선을 그리거나, 아니면 자동으로 현재 표시되어 있는 데이터를 기반으로 가우스 곡선을 표시해 주는 기능이 있다.

위치(E) (Position (E))
너비(G) (Width (G))
진폭(A) (Amplitude (A))



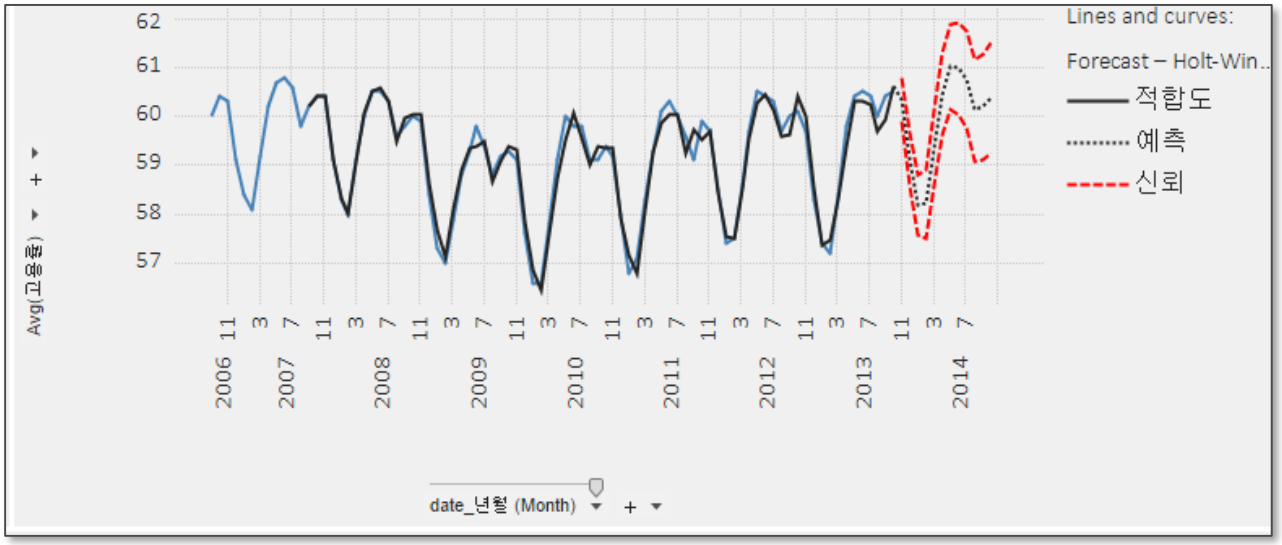
1. 라인 & 곡선(Lines & Curves) – 방정식



곡선 맞춤 모델	방정식
직선	$y = a + bx$
다항	$y = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5$
파워	$y = ax^b$
로그	$y = a + b\ln x$
지수	$y = ae^{bx}$
가우스	$y = A \cdot e^{-\left(\frac{x-E}{G}\right)^2}$

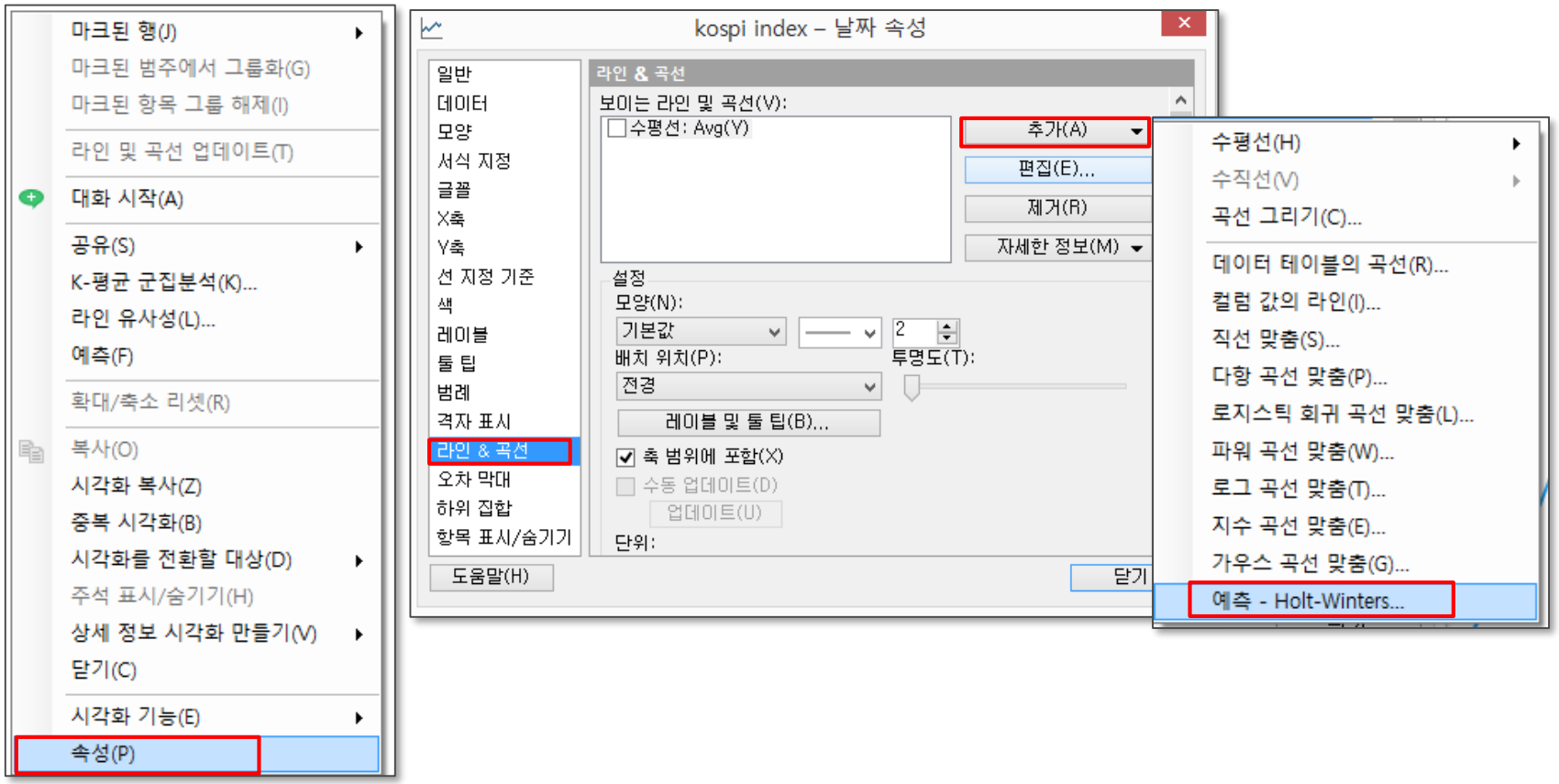
1. 라인 & 곡선(Lines & Curves) – 예측 (Holt-Winters) 설정

- Spotfire의 시각화에 시계열 예측을 할 수 있는 Holt-Winters Fitting Line을 사용할 수 있다.
- 추세와 계절적 변동이 모두 존재하는 시계열의 경우에 '홀트 윈터스 지수평활법 (Exponential Smoothing)'을 사용한다.
- 홀트 윈터스 지수평활법은 수준, 추세, 계절 요인을 반영할 수 있는 예측 방법이다.
- Holt-Winters 에서는 Level, Trend, Seasonal의 경우 별도로 값을 부여하지 않을 경우 자동으로 산출하여 그 결과를 출력한다.
- Spotfire 예측 기능은 대략적인 예측을 할 수 있도록 보조선 형태로 Chart에 추가할 수 있다.

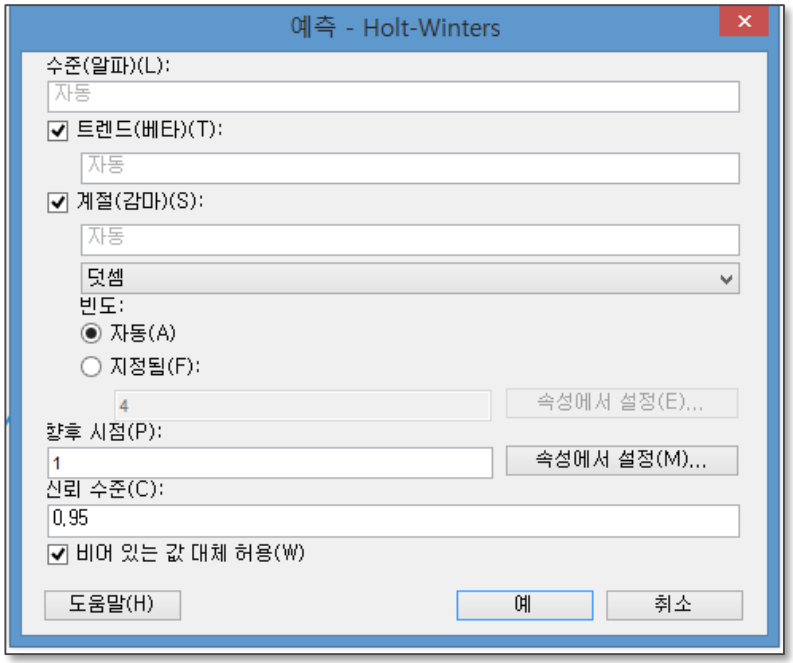


1. 라인 & 곡선(Lines & Curves) – 예측 (Holt-Winters) 설정

- 실행 방법 : 예측을 원하는 시각화에 마우스를 이동한 후,
  마우스 우클릭 > 속성 > 라인 & 곡선 > 추가 > '예측 - Holt-Winters'



1. 라인 & 곡선(Lines & Curves) – 예측 (Holt-Winters) 설정



- **향후 시점(Times points ahead)**
  - 예측하고자 하는 개수
- **신뢰 수준(Confidence level)**
  - 신뢰구간에 대한 유의확률(0<x<≦1)
- **비어있는 값 대체 허용**
  - 인접한 값을 보간(interpolating)하여 비어 있는 값을 대체.

- **수준[Level(alpha)]**
  - Level에 대한 조정 계수(0<x<≦1)
  - 0에 가까울수록 과거 값에, 1에 가까울수록 최근 값에 weight를 더 부여함
  - 0~1에 대한 조정 값은 트랜드, 계절 모두 동일함
- **트랜드[Trend(beta)]**
  - Trend에 대한 조정 계수(0<x<≦1)
- **계절[Seasonal(gamma)]**
  - Season에 대한 조정 계수(0<x<≦1)
  - 덧셈 : * (수준 + 트렌드 +계절)로 모델링됨
    * 계절적인 변동이 시계열 전반에 걸쳐 거의 일정할 때 사용
  - 곱셈 : * [(수준 + 트렌드) * 계절]로 모델링됨
    * 계절적인 변동이 시계열의 수준에 비례하게 변할 때 사용.

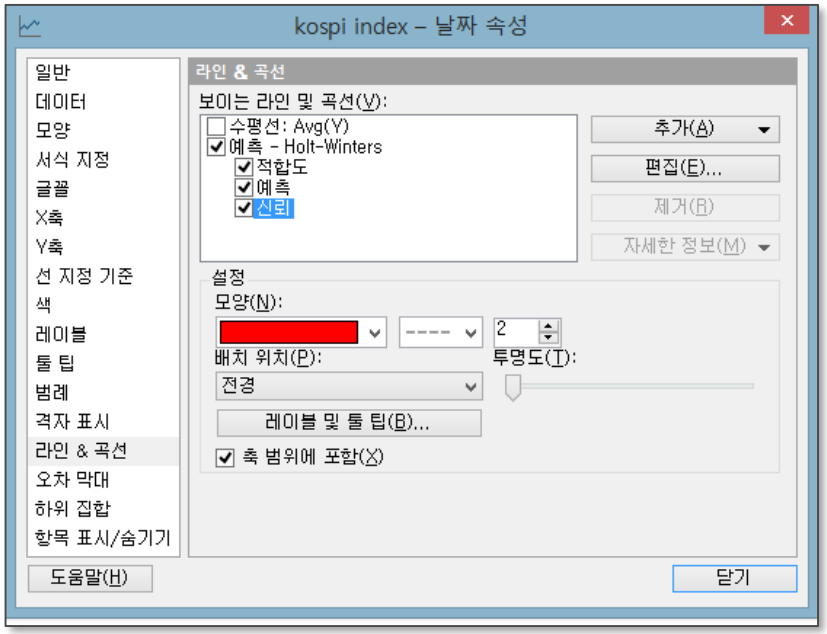
빈도(Frequency)

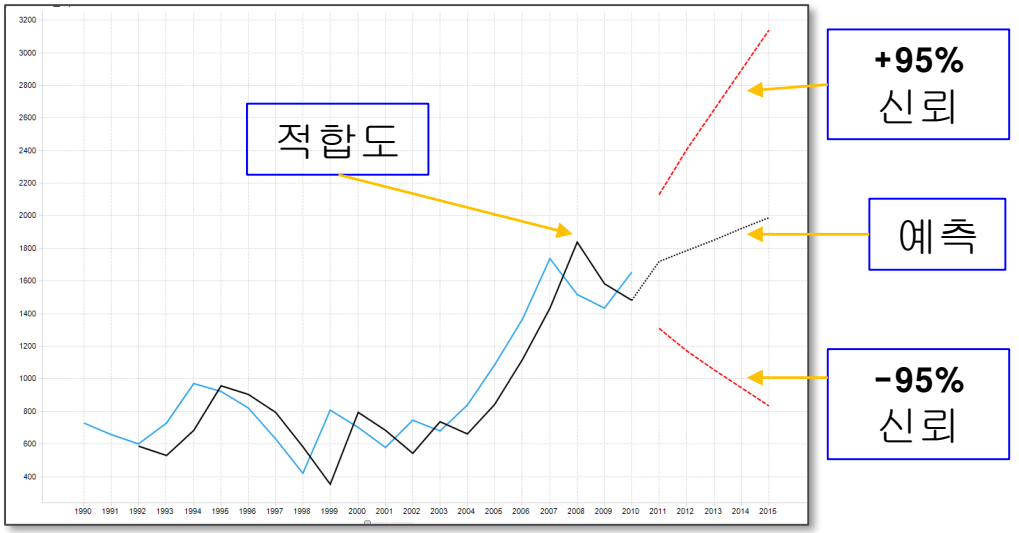
- 계절(Seasonal)에 대한 주기
  - 시작 값을 계산하는 데 사용할 계절 기간 수 즉, 샘플링 기간당 관찰 수. 예를 들어, 월별 데이터의 빈도는 12이다.

출처 : **CNFCJ https://danielykim.gitbooks.io/fpp-kr/content/doc/fpp-kr-7-5.html**

1. 라인 & 곡선(Lines & Curves) – 예측 (Holt-Winters) 설정

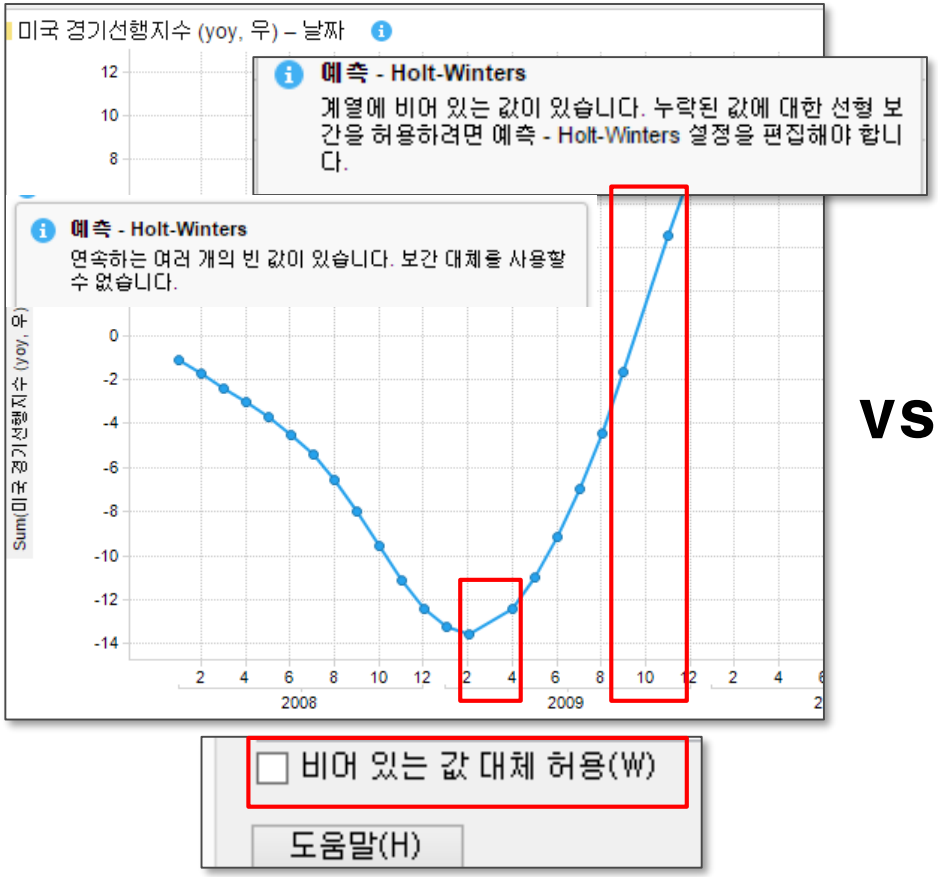
- Spotfire는 다른 제품의 Holt-Winters와 달리, 별도의 속성값들을 일일이 설정하지 않아도 자동으로 최적의 예측 모델을 찾아서 수행하는 특징을 갖고 있다.
- 일반적으로 시계열 예측에 많이 사용되는 것은 ARIMA 모델 등이 있지만, Holt-Winters는 사용이 간편하다는 장점이 있는 대신에, point 가 누락되어 있으면 수행이 불가하다는 단점이 있다.
- 실행 후에는 예측 결과가 신뢰 구간(± **95%)**과 함께 표시된다.



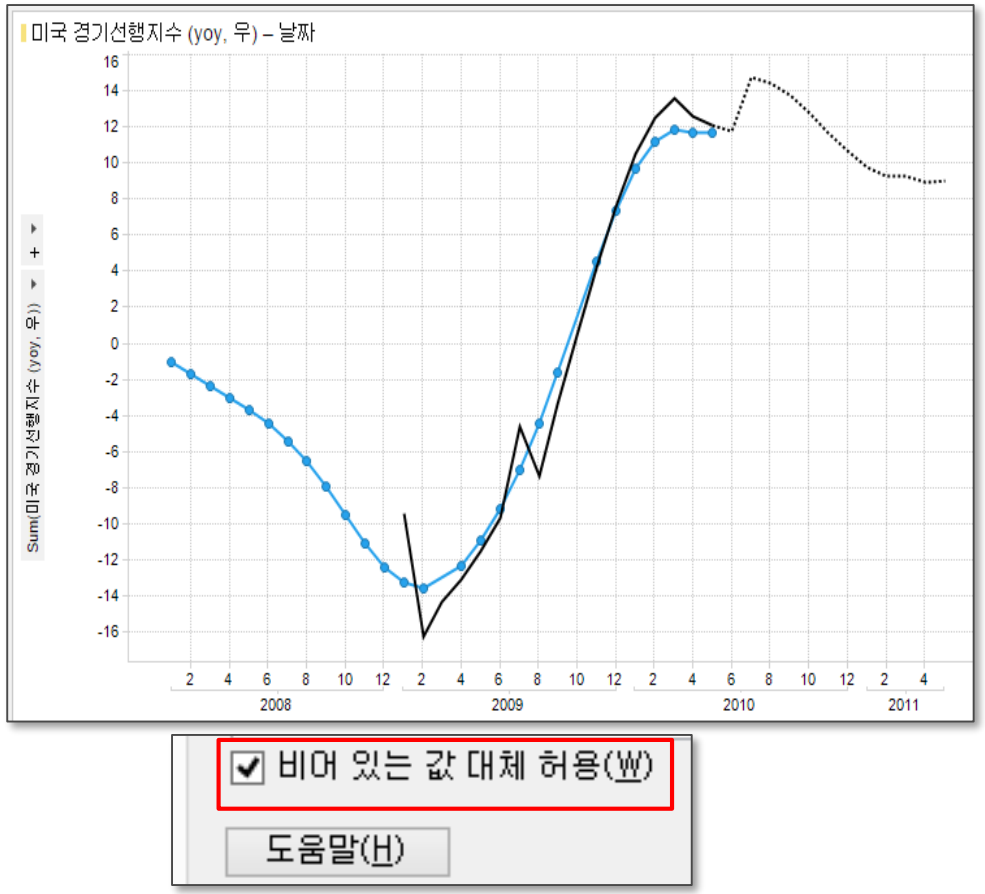


1. 라인 & 곡선(Lines & Curves) – 예측 (Holt-Winters) 설정

- Holt-Winters는 X축의 값이 연속되는 값이어야 한다. 한 point라도 빠지는 날이 있으면 안되며 이는 월, 년인 경우도 동일하다.
- Spotfire에는 비어있는 값을 보정해 주는 옵션이 있어서 이를 사용할 수 있다.



vs



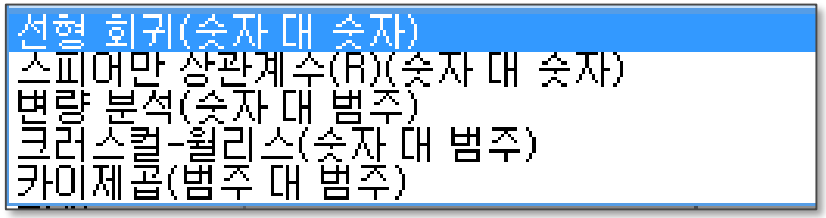
2. 데이터 상관성(Data Relationship) 분석

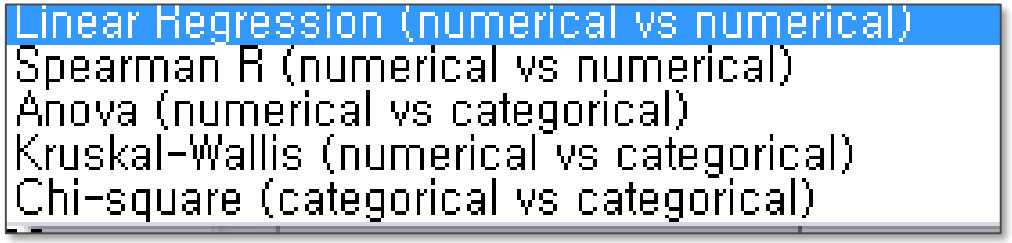
상관분석(Correlation Analysis)

- **상관분석은** 확률론과 통계학에서 두 변수간에 어떤 선형적 관계를 갖고 있는 지를 분석하는 방법이다.
- 두변수는 서로 독립적인 관계로부터 서로 상관된 관계일 수 있으며 이때 두 변수간의 관계의 강도를 상관관계**(Correlation, Correlation coefficient)**라 한다. 상관분석에서는 상관관계의 정도를 나타내는 단위로 모상관계수 ρ를 사용한다.
- 상관관계의 정도를 파악하는 상관계수**(Correlation coefficient)**는 두 변수간의 연관된 정도를 나타낼 뿐 인과관계를 설명하는 것은 아니다. 두 변수간에 원인과 결과의 인과관계가 있는지에 대한 것은 회귀분석을 통해 인과관계의 방향, 정도와 수학적 모델을 확인해 볼 수 있다.

2. 데이터 상관성(Data Relationship) 분석

- 데이터 상관성 분석 도구는 서로 다른 컬럼들간에 서로 상관 관계가 존재하는지를 조사하는 데 사용된다.
- 이 도구는 항상 현재 필터링된 데이터에 대해서만 동작한다.
- **Spotfire**에서는 데이터의 타입에 따라서 총 **5**가지의 상관성 분석 도구를 제공한다.





	판매 금액	처방건수	진료과
2	594	1	신장내과
3	999	1	소화기내과
4	1,256	1	소화기내과
5	1,260	1	소화기내과
6	1,332	1	신장내과
7	1,422	2	신장내과
8	1,710	1	신장내과
9	2,100	1	신장내과
10	2,212	1	소화기내과
11	2,360	1	소화기내과
12	2,370	1	소화기내과
13	2,660	1	신장내과
14	2,664	2	소화기내과
15	2,930	1	소화기내과
16	2,970	1	신장내과
17	3,495	1	소화기내과
18	2,536	1	신장내과
19	2,643	1	신장내과
20	4,257	5	신장내과
21	4,266	2	신장내과
22	4,455	2	소화기내과

2. 데이터 상관성(Data Relationship) 분석

- **Column** 간의 상관성을 분석하기 위한 기능으로 지정된 **Column**의 쌍별**(Pair-wise)**로 비교하여 결과를 출력한다.
- 각 컬럼 조합에서는 이 도구로 첫 번째 컬럼이 두 번째 컬럼의 값을 예측하는 정도를 나타내는 **p-value**를 계산한다. **낮은 p값은 두 컬럼 간의 강력한 연결 가능성을 나타낸다.**
- 결과 테이블은 **Y** 및 **X** 컬럼의 각 조합에 대한 **p-value**를 표시한다. 테이블은 **p**값을 기준으로 정렬된다. 컬럼 머리글을 클릭하면 그 컬럼에 따라 행이 다시 정렬된다.

구분	데이터 형	내용	예시
Linear Regression	수치형 vs 수치형	두 수치형 변수 간의 선형적 관계를 도출 회귀식으로 산출함	아버지와 아들 간의 키 (아버지의 키 vs 아들의 키)
Spearman R	수치형 vs 수치형	두 수치형 변수간의 상관관계를 검증하고 상관성 정도를 산출함	아파트 평수와 가격의 관계 (아파트 평수 vs 아파트 가격)
Anova	수치형 vs 명목형	명목형에 따라 수치형의 변화를 확인하는 방법 세개 이상의 집단에 대한 평균을 비교하는 방법 (모수적 방법* : Anova, 비모수적 방법 : Kruskal-wallis)	학급별 성적의 차이 (성적 vs 학급명)
Kruskal-Wallis	수치형 vs 명목형		학급별 성적의 차이 (성적 vs 학급명)
Chi-square	명목형 vs 명목형	두 명목형 변수간의 상관관계를 확인	지역에 따른 지지정당의 차이 (지역 vs 지지정당)

* 모수적**(parametric)** 방법 : 모집단의 분포에 대한 가정을 포함하는 통계적 방법 **(**예, 모집단의 분포는 정규분포를 따른다.**)**

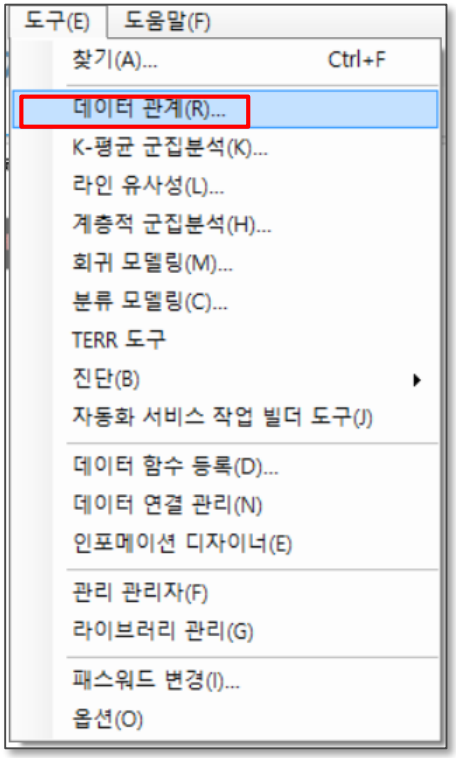
* 데이터테이블 : 처방 데이터.xls

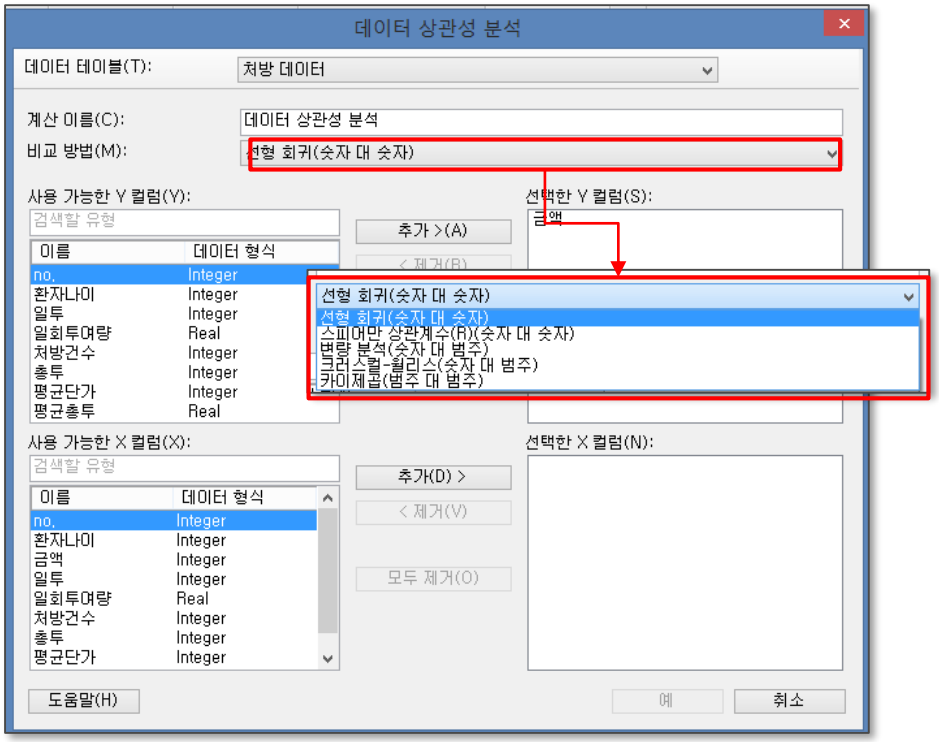
2. 데이터 상관성(Data Relationship) 분석

- **Spotfire**에서 데이터 상관성 분석을 수행해 보자. 여기서는 '처방 데이터**.xls**' 를 이용하여 '금액' 에 가장 상관성이 높은 컬럼을 찾아내 보기로 한다.

  1. 상단의 메인 메뉴 : '도구' > '데이터 관계' 클릭

  2. 비교 방법' 에서 원하는 알고리즘 방법을 선택**[**여기서는 선형회귀**(**숫자 대 숫자**)**' 선택**]**

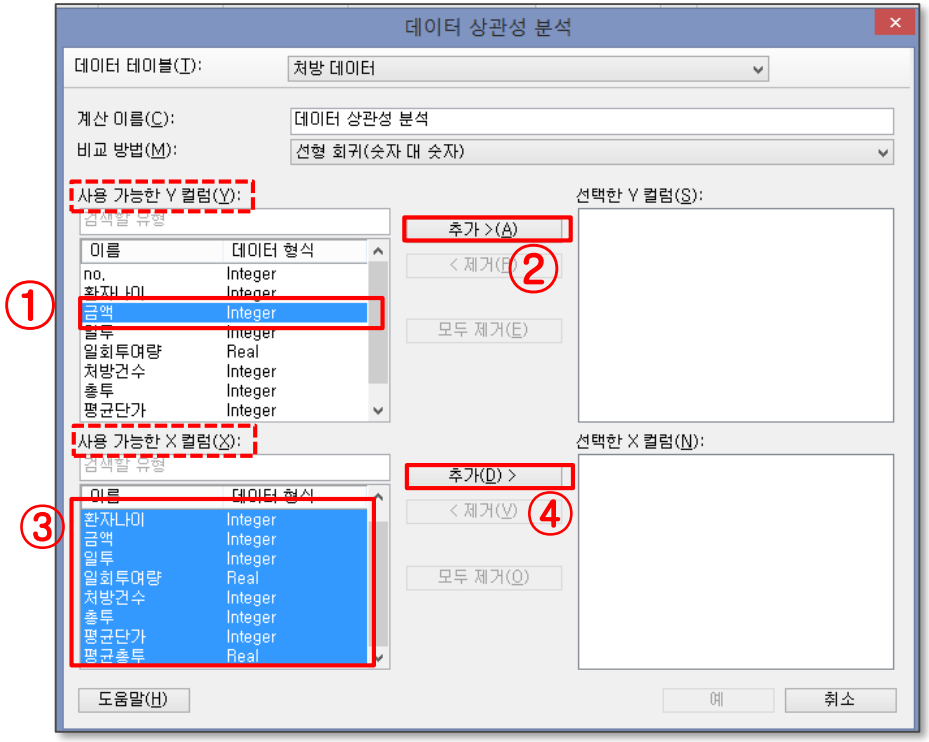




* 데이터테이블 : 처방 데이터.xls

2. 데이터 상관성(Data Relationship) 분석

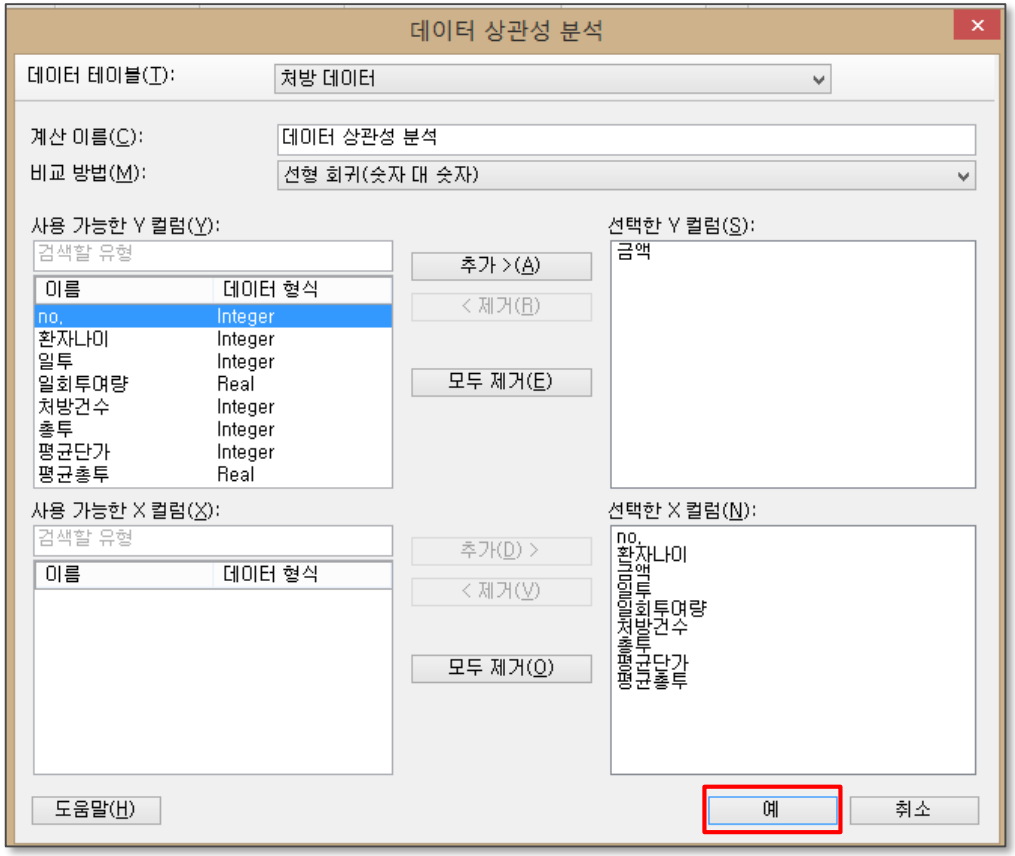
3. ‘사용 가능함 **Y** 컬럼**(Y)**’ 에서 원하는 컬럼을 클릭하고 ‘추가’ 버튼을 누른다.
   (여기서는 ‘금액’ 을 선택한다.**)**
4. ‘사용 가능함 **X** 컬럼**(X)**’ 에서 원하는 컬럼들을 클릭하고 추가 버튼을 누른다.
   (여기서는 리스트에 나와있는 모든 컬럼들을 선택한다. **(**이때는 첫번째 컬럼을 선택한 후 스크롤 박스를 움직여 맨 아래로 이동 후에 **shift +** 마지막 컬럼’ 을 선택**)** 그리고 ‘추가’ 버튼을 누른다.



* 데이터테이블 : 처방 데이터.xls

2. 데이터 상관성(Data Relationship) 분석

5. 이제 아래와 같은 화면으로 선택이 완료되었으면, 마지막으로 '예' 버튼을 누른다.



* 데이터테이블 : 처방 데이터.xls

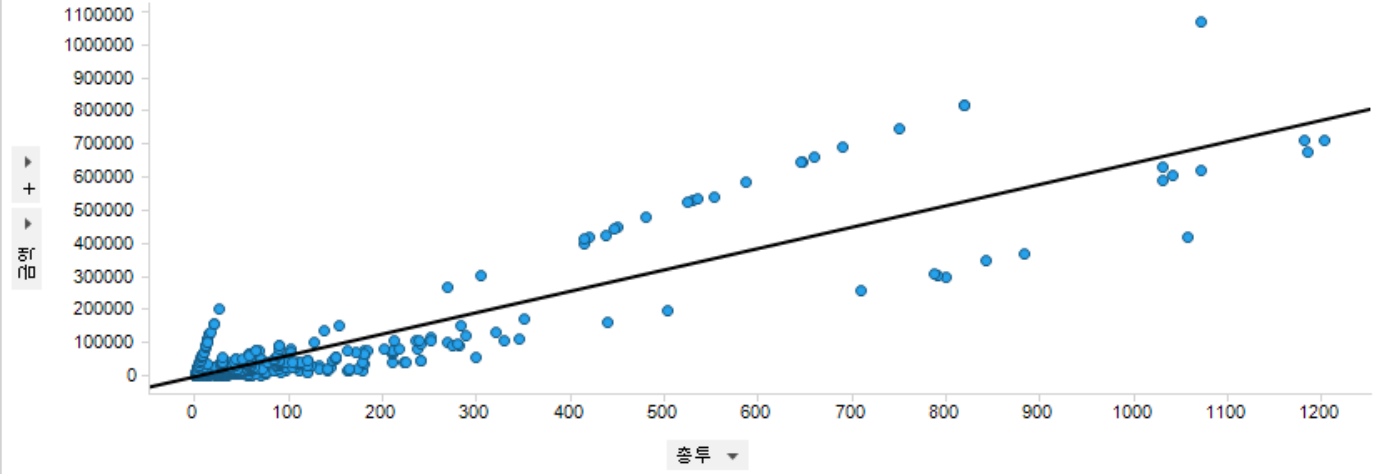
2. 데이터 상관성(Data Relationship) 분석

6. 자동으로 다음 페이지에 '데이터 상관성 분석' 이라는 페이지가 생성되면서 아래와 같은 시각화와 분석 결과가 만들어 진다.

데이터 상관성 분석 (3) (선형 회귀)

Y (numerical)	X (numerical)	p-value	FStat	RSq	R	Df	n
금액	총투	6.38E-284	3362.58	0.81	0.90	774	776
금액	처방건수	3.49E-211	1909.88	0.71	0.84	774	776
금액	일회투여량	1.59E-082	475.29	0.38	0.62	774	776
금액	일투	3.75E-076	428.90	0.36	0.60	774	776
금액	평균총투	9.76E-012	47.82	0.06	0.24	774	776
금액	평균단가	1.23E-005	19.36	0.02	0.16	774	776
금액	환자나이	2.05E-003	9.57	0.01	0.11	772	774
금액	no.	3.05E-001	1.05	0.00	0.04	774	776

데이터 상관성 분석 (3)(상세 정보)



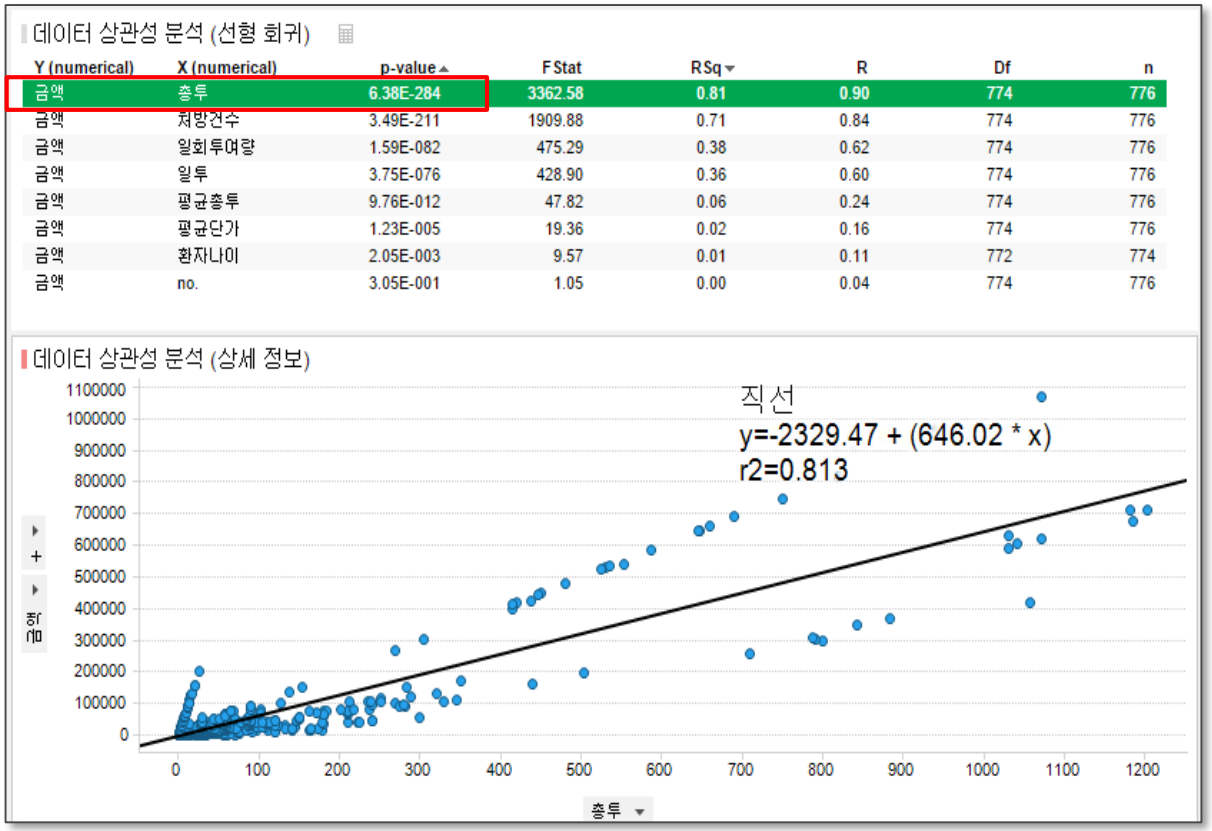
* 데이터테이블 : 처방 데이터.xls

2. 데이터 상관성(Data Relationship) 분석

7. '데이터 상관성 분석(선형회귀)' 결과는 자동으로 **p-value**가 가장 작은 값부터 정렬되어 표시된다. 시각화의 가장 첫번째 행이 **P-value**가 가장 작으며, 이는 '금액' 과 '총투' 가 가장 상관성이 높음을 의미한다.

* **P-value**는 일반적으로 **0.05** 이하일때 **X**와 **Y**간에 상관성이 의미가 있으며, **Rsq**값(**-1.0~1.0**사이)은 절대값이 **1**에 가까울수록 의미가 있으며, **-1**인 경우에는 **X**와 **Y**간에 음의 상관성을 갖는다.

* 마우스 우클릭 > 속성 > '라인&곡선' > '레이블 및 툴팁' 을 이용하여 수식과**R2**값을 나타낼 수 있다.



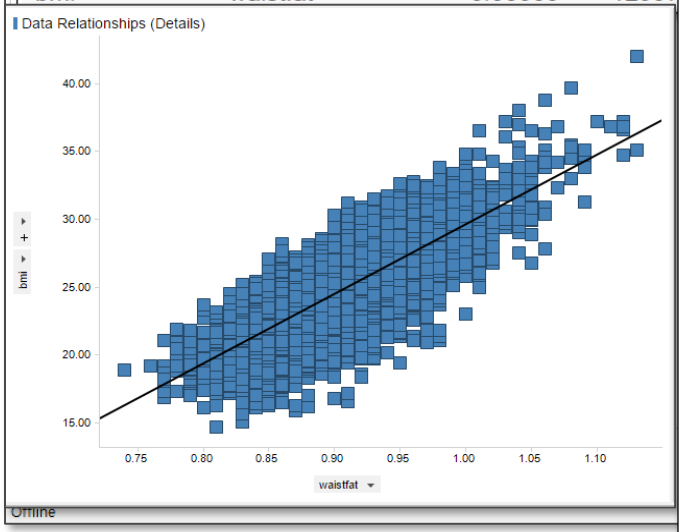
* 데이터테이블 : 체질량 지수(bmi).xls

2. 데이터 상관성 분석 – 선형 회귀(Linear Regression :숫자 대 숫자)

- 독립변수(원인변수)가 종속변수(관심변수)에 미치는 영향력의 크기를 측정하여 독립변수의 일정한 값에 대응되는 종속변수의 값을 예측하는 통계분석 방법이다.
- 독립변수가 종속변수에 영향을 미치는 지의 여부와 그 영향력의 크기를 검증하는데 목적

Data Relationships (Linear Regression)

Y (numerical)	X (numerical)	p-value	FStat	RSq	R	Df	n
bmi	age	0.07473	3.18	0.0004	-0.02	7372	7374
bmi	bodyfat	0.00000	4466.16	0.3773	0.61	7372	7374
bmi	hip	0.00000	9194.33	0.5550	0.74	7372	7374
bmi	waist	0.00000	8669.77	0.5404	0.74	7372	7374
bmi	waistfat	0.00000	12997				

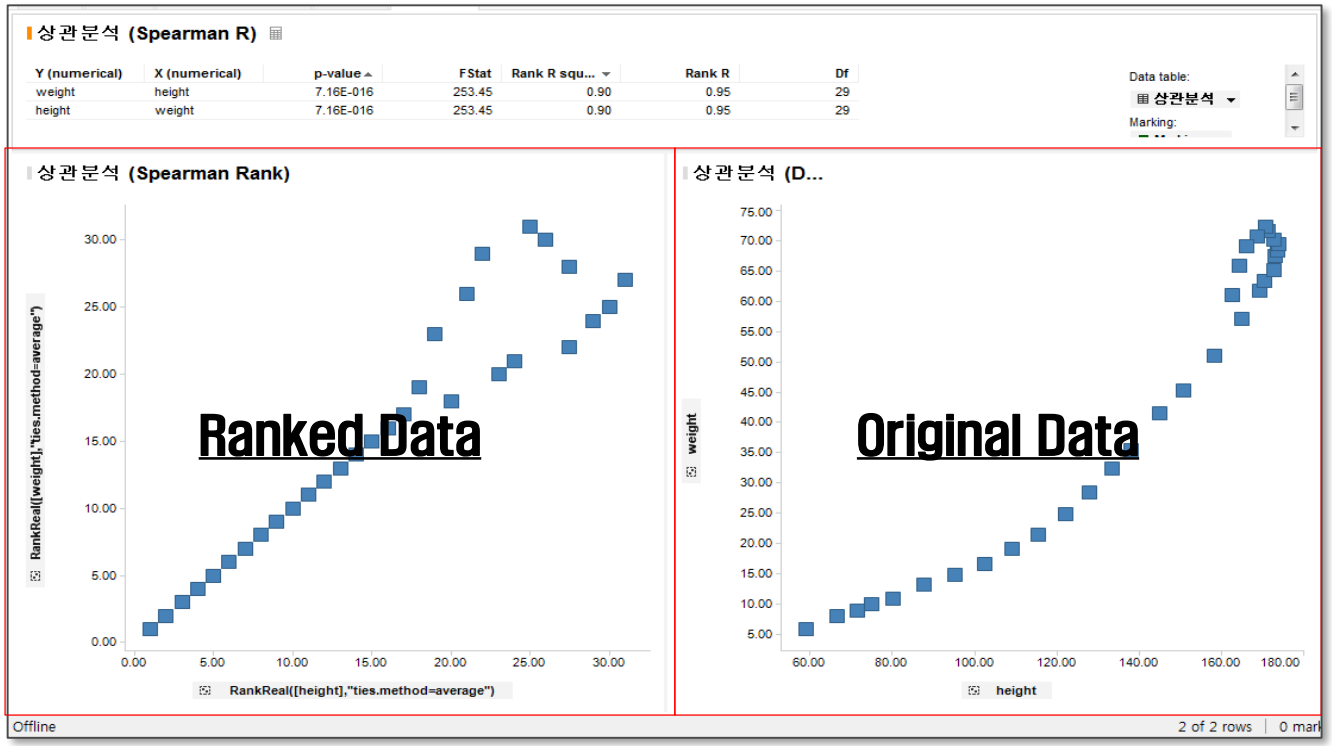


1) 모형이 통계적으로 유의한지 확인하고 위해 p-value가 유의수준 이하인지 확인한다.
   → 보통 0.05를 유의수준으로 하며 0.05 이하일 경우 모형이 통계적으로 유의하다고 이야기하며 이 모형에 대해서만 아래와 같은 과정으로 거쳐 분석을 진행한다.
2) RSq 값을 확인하여 적합된 회귀식이 자료에 대해 잘 설명하고 있는 지를 확인한다.
   → 이론적으로 0.8 이상, 분야에 따라서는 0.6이상일 경우 적합된 모형이 자료를 잘 설명하고 있다고 판단하며, 0인 경우에는 서로 완전히 독립적인 관계이다.

* 데이터테이블 : 체질량 지수(bmi).xls

2. 데이터 상관성 분석 – 스피어만 (Spearman-R : 숫자 대 숫자)

- 두 숫자형 컬럼(변수) 간의 선형적 관계에 초점을 둔 분석으로 단순히 음 또는 양의 상관관계와 상관 정도를 알 수 있다.
- 데이터의 실제값 대신 두 값의 순위를 사용해서 분석. 비선형관계의 연관성을 파악할 수 있다는 장점을 가지고 있다. 이산형, 순서형 데이터에 적용 가능하다.



산술평균보다 중앙값 사용이 더 적절한 자료의 경우 **spearman** 상관계수 사용.

* 데이터테이블 : 체질량 지수(bmi).xls

2. 데이터 상관성 분석 – Spearman R(숫자 대 숫자)

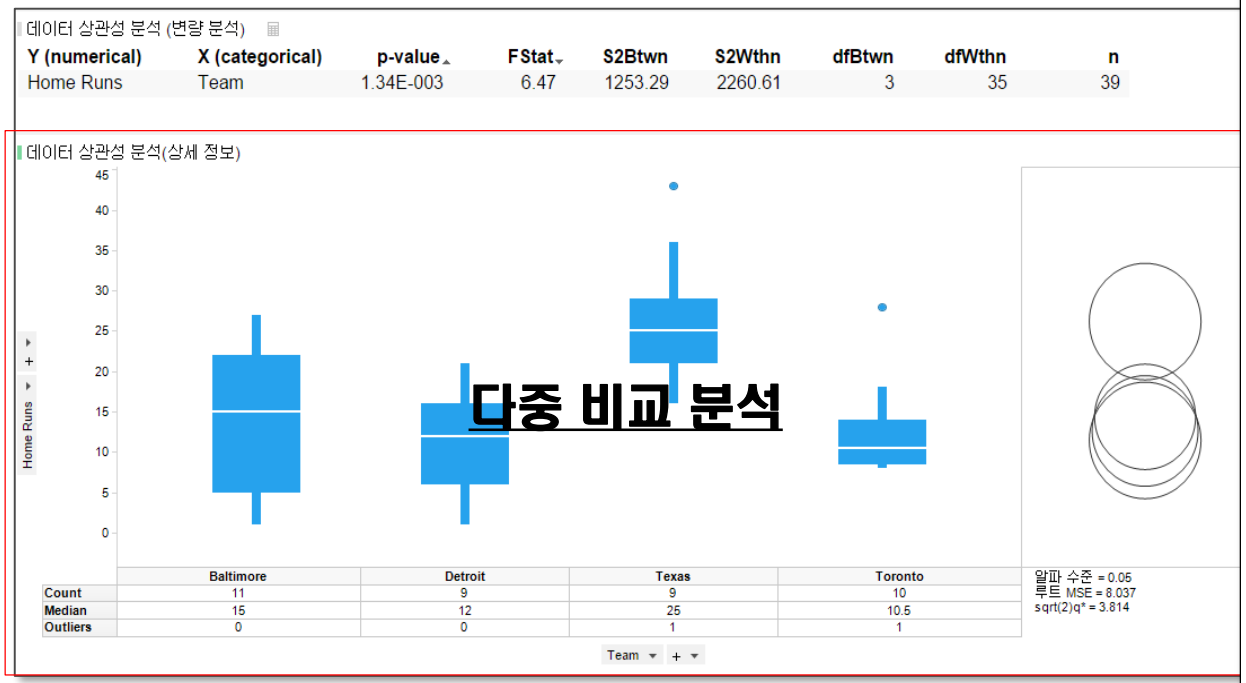
스피어만 상관계수 (Spearman correlation coefficient)

- 스피어만 상관계수는 데이터가 서열 척도인 경우, 즉 자료의 값 대신 순위를 이용하는 경우의 상관계수로서, 데이터를 작은 것부터 차례로 순위를 매겨 서열 순서로 바꾼 뒤 순위를 이용해 상관계수를 구한다.
- 두 변수 간의 연관 관계가 있는지 없는지를 밝혀주며 **자료에 이상점이 있거나 표본크기가 작을 때 유용하다.**
- 스피어만 상관계수는 **-1**과 **1** 사이의 값을 가지는데 두 변수 안의 순위가 완전히 일치하면 **+1**이고, 두 변수의 순위가 완전히 반대이면 **-1**이 된다. 예를 들어 수학 잘하는 학생이 영어를 잘하는 것과 상관이 있는지 없는지를 알아보는데 쓰일 수 있다.

출처 : **https://ko.wikipedia.org/wiki/%EC%83%81%EA%B4%80%EB%B6%84%EC%84%9D**

2. 데이터 상관성 분석 – 변량 분석(Anova : 숫자 대 범주)

- 범주(명목)형 변수에 따른 숫자형 변수에 대한 상관성을 분석 하는 방법으로 세 가지 이상의 집단에 대한 수치를 비교할 때 주로 사용하는 방법이다.
- 자료의 정규성과 등분산성을 만족할 경우에는 **Anova**를 사용한다.
- 변량 분석 옵션은 각 그룹의 데이터 평균 값을 비교하여 그룹 간 차이를 계산한다



1) 모형이 통계적으로 유의한지 확인하고 위해 p-value가 유의수준 이하인지 확인한다.
   → 보통 0.05를 유의수준으로 하며 0.05 이하일 경우 모형이 통계적으로 유의하다고 이야기하며 이 모형에 대해서만 아래와 같은 과정으로 거쳐 분석을 진행한다.
2) 일반적으로 통계에서는 Anova 분석 후에 어느 집단에 차이가 있는 지 확인 하기 위해 다중비교 분석을 실시하는데 Spotfire에는 그런 기능이 없다. 대신 BoxPlot을 활용하여 분석하도록 하고 있는데 이때 Box-Plot의 Comprarison Circle 옵션을 사용하면 유용하다.

2. 데이터 상관성 분석 – 변량 분석(Anova : 숫자 대 범주)

변량 분석(Analysis of variance, ANOVA, 분산 분석)

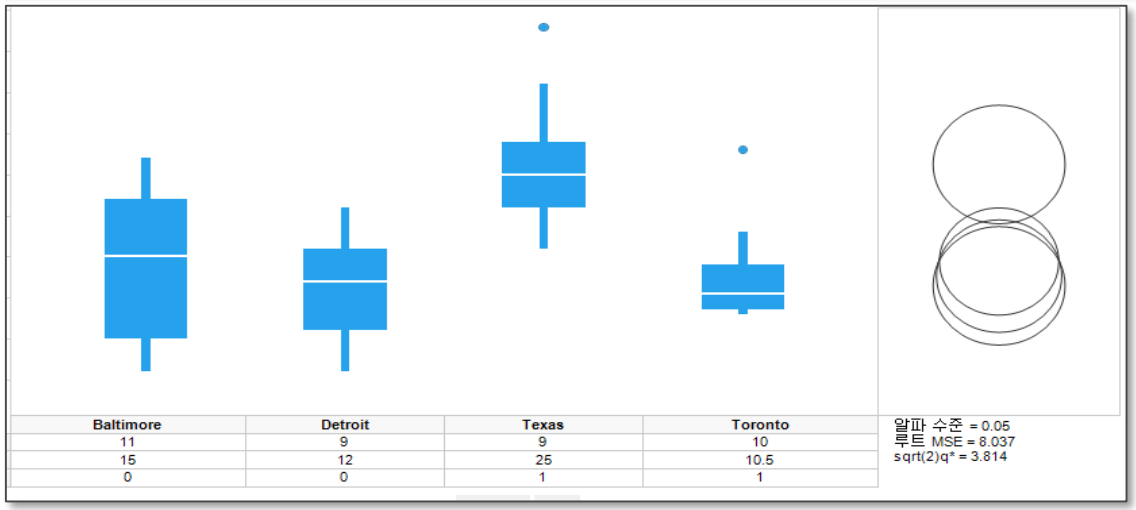
- **변량 분석(ANOVA ;** 변량분석**)**은 통계학에서 두 개 이상 다수의 집단을 비교하고자 할 때 집단 내의 분산**,** 총평균과 각 집단의 평균의 차이에 의해 생긴 집단 간 분산의 비교를 통해 만들어진 **F**분포를 이용하여 가설검정을 하는 방법이다**.**
- **F**분포는 분산의 비교를 통해 얻어진 분포비율이다**.** 이 비율을 이용하여 각 집단의 모집단분산이 차이가 있는지에 대한 검정과 모집단평균이 차이가 있는지 검정하는 방법으로 사용한다**.** 즉 **F = (**군간변동**)/(**군내변동**)**이다**.** 만약 군내변동이 크다면 집단간 평균차이를 확인하는 것이 어렵다**.**
- **분산분석에서는 집단간의 분산의 동질성을 가정하고 하기 때문에** 만약 분산의 차이가 크다면 그 차이를 유발한 변인을 찾아 제거해야 한다**.** 그렇지 못하면 분산분석의 신뢰도는 나빠지게 된다**.**
- 통계학자이자 유전학자인 로날드 피셔**(R.A. Fisher)**에 의해 **1920**년대에서 **1930**년대에 걸쳐 만들어졌다**.**

2. 데이터 상관성 분석 – 크러스컬-월리스(Kruskal-Wallis:숫자 대 범주)

- 범주(명목)형 변수에 따른 숫자형 변수에 대한 상관성을 분석 하는 방법으로 세 가지 이상의 집단에 대한 수치를 비교할 때 주로 사용하는 방법이다.
- 자료의 정규성과 등분산성을 만족하지 못할 경우에 비모수적 방법인 **Kruskal-Wallis**를 사용한다.

데이터 상관성 분석 (2) (크러스컬-윌리스)

Y (numerical)	X (categorical)	p-value	H-stat	Df	n
Home Runs	Team	8.75E-003	11.63	3	39



1) 모형이 통계적으로 유의한지 확인하고 위해 p-value가 유의수준 이하인지 확인한다.
→ 보통 0.05를 유의수준으로 하며 0.05 이하일 경우 모형이 통계적으로 유의하다고 이야기하며 이 모형에 대해서만 분석을 진행한다.

2. 데이터 상관성 분석 – 비모수 통계

비모수 통계(Nonparametric Statistics)

- **비모수 통계** 는 확률 분포**(Pprobability Distributions)**의 매개 변수화 된 계열을 기반으로 하지 않는 통계이다**.** 확률분포의 전형적인 매개 변수**(parameters)**는 평균**(mean)**, 분산**(variance)** 등 이다**.**
- 예를 들어 랭킹 순서 **(**예 **: 1 ~ 4** 개의 별을받는 영화 리뷰**)**를 연구하는데 적용된다**.**
- 여기에는 설명**(descriptive)** 및 추론**(inferential)** 통계가 모두 포함된다**.**
- 비모수 **(**또는 분포가 없는**)** 추론 통계 방법 은 통계적 가설 검정을 위한 수학적 절차로서**,** 매개 변수**(Parametric Statistics)** 통계 와 달리 평가되는 변수 의 확률 분포 **(Pprobability Distributions)**에 대한 가정을 하지 않는다**.**
- 비 파라 메 트릭 방법은 가정이 적기 때문에 적용 가능성이 해당 파라 메 트릭 방법보다 훨씬 넓다**.** 특히 문제의 응용 프로그램**(application)**에 대해 알려진 내용이 적은 상황에 적용될 수 있다**.**
- 또한**,** 적은 가정**(assumptions)**에 의존하기 때문에**,** 비 파라 메 트릭 방법이 더 강력하다**.**

2. 데이터 상관성 분석 – 카이 제곱(chi-square : 범주 대 범주)

- 두 범주(명목)형 변수간의 상관성을 분석하는 방법이다

Chi-Square (Chi-square)

Y (categorical)	X (categorical)	p-value	Chi2-stat	Df	n
성별	지지정당	0.5464	1.21	2	20
지지정당	성별	0.5464	1.21	2	20

비율에 대한 교차표

Chi-Square (Details)

성별 \ 지지정당	공화당	민족당	신민당
남성	2	2	5
여성	5	2	4

Chi-Square (Details)

성별 \ 지지정당	공화당	민족당	신민당
남성	0.22	0.22	0.56
여성	0.45	0.18	0.36

1) 모형이 통계적으로 유의한지 확인하고 위해 p-value가 유의수준 이하인지 확인한다.
→ 보통 0.05를 유의수준으로 하며 0.05 이하일 경우 모형이 통계적으로 유의하다고 이야기하며 이 모형에 대해서만 분석을 진행한다.

2) Spotfire에서는 교차표만 결과로 출력하지만 열 또는 행에 대한 비율을 계산하면 결과를 해석하는데 도움을 받을 수 있다 [Count() / Count() over (All([Axis.Columns]))]

2. 데이터 상관성 분석 – 카이 제곱(chi-square : 범주 대 범주)

카이 제곱(Chi-square)

- Chi-squared 검정은 흔히 '제곱 오차 의 합' 또는 '표본 분산' 을 통해 구성된다.
- 카이 제곱 분포를 따르는 테스트 통계는 독립적으로 정규 분포 된 데이터의 가정에서 발생한다.
- **Chi Square** 통계는 두 개 **(**또는 그 이상**)**의 독립적 인 그룹 사이의 범주 형 응답의 집계 또는 계수를 비교한다**.**
- 카이 제곱 시험은 백분율**,** 비율**,** 평균 등이 아닌 실제 숫자**(actual numbers)**에서만 사용할 수 있다**.**
- 카이 제곱 통계는 두 가지 범주 형 변수 사이의 관계를 보여주는 방법이다**.** 통계에는 두 가지 유형의 변수**[** 숫자 형 **(**계수 형 **; countable)** 변수와 비 수치 형 **(**범주 형**; categorical)** 변수**]**가 있다**.**

2. 데이터 상관성 분석 – 카이 제곱(chi-square : 범주 대 범주)

카이 제곱(Chi-square)

- **카이 제곱 테스트** 에는 **두 가지 유형이** 있다. 둘 다 다른 목적으로 카이 제곱 통계 및 분포를 사용한다.
- **카이 제곱 적합도 검정(goodness of fit test)** 은 표본(sample) 데이터가 모집단(population)과 일치하는지 여부를 결정한다.
- 독립성에 **대한 카이 제곱 검정 (chi-square test for independence)** 은 우연성(contingency) 표의 두 변수를 비교하여 관련성을 확인한다. 보다 일반적인 의미에서, 범주 형 변수의 분포가 서로 다른지 여부를 확인한다.
  - **아주 작은 카이 제곱 검정 통계** 는 관찰 된 데이터가 예상 데이터를 매우 잘 **맞춘다** 는 것을 의미한다. 즉, 관계가 있다.
  - **매우 큰 카이 제곱 검정 통계** 는 데이터가 잘 맞지 **않는다는** 것을 의미한다. 즉, 관계가 없다.

2. 데이터 상관성 분석

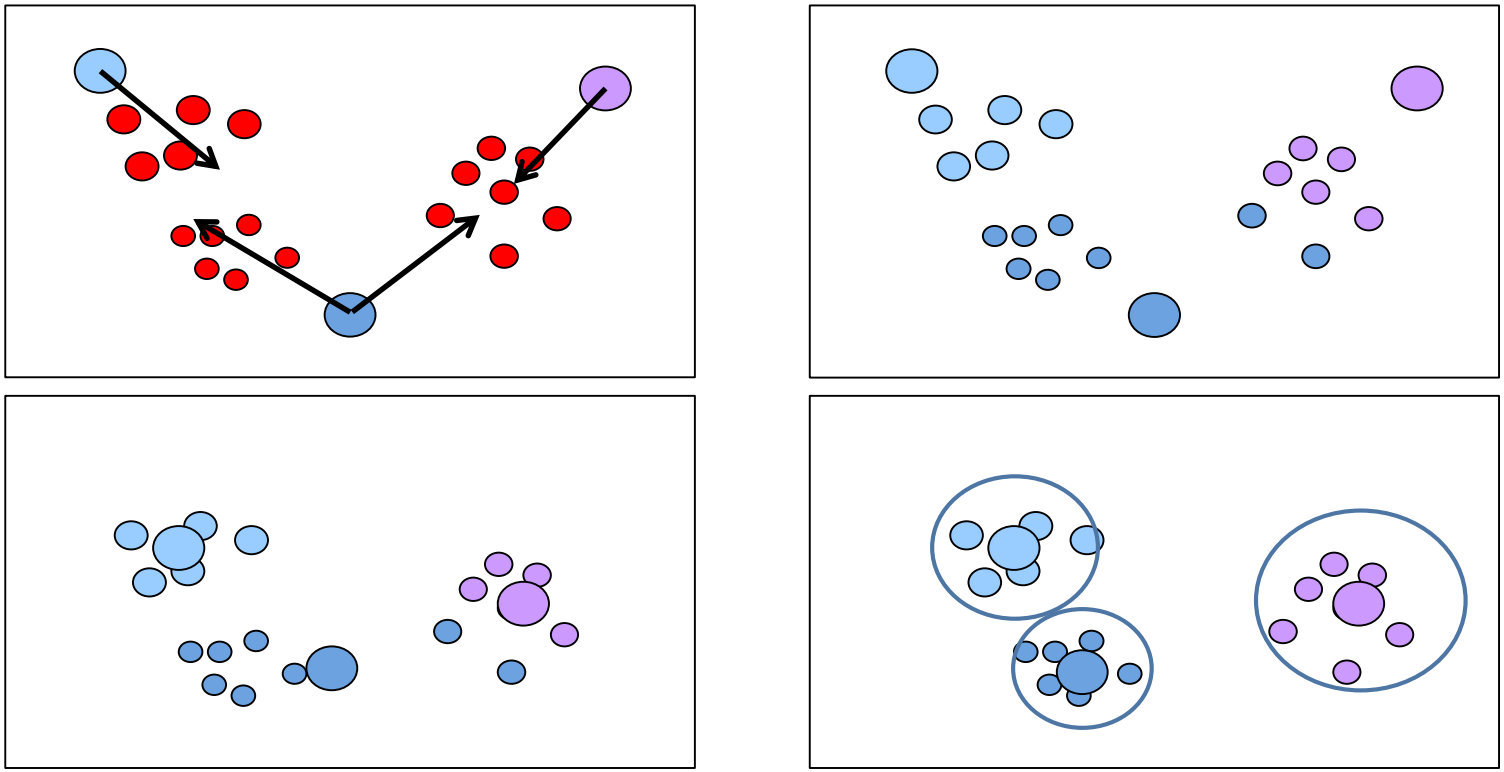
데이터 상관성 분석에 대한 입력 데이터의 데이터 분포 요구 사항

- 변량 분석**(ANOVA)** 및 선형 회귀**(Linear Regression)** 비교는 다음을 전제로 한다.
  - 데이터는 대량 정상적으로 분포된다.
  - 개별 그룹의 변형**(**선형 회귀의 경우에는 오류의 변형**)**이 거의 동등하다.

- 데이터가 이 조건을 충족하지 않는 경우에는 변량 분석 및 선형 회귀 비교가 믿을 수 없는 결과를 낳을 수 있다. 이 경우에는 크러스컬-월리스**(Kruscal Wallis)** 또는 스피어만 상관계수**(Spearman-R)** 비교를 사용하는 것이 더 유효할 수 있다.

3. K-평균 군집분석

K-평균 군집분석은 군집화 **(Clustering)** 문제를 해결하는 가장 간단한 자율학습 **(Unsupervised Learning)** 알고리즘 중 하나이다**.** 사전에 정해진 어떤 수의 클러스터를 통해서 주어진 데이터 집합을 분류하는 간단하고 쉬운 방법이다**.**



3. K-평균 군집분석

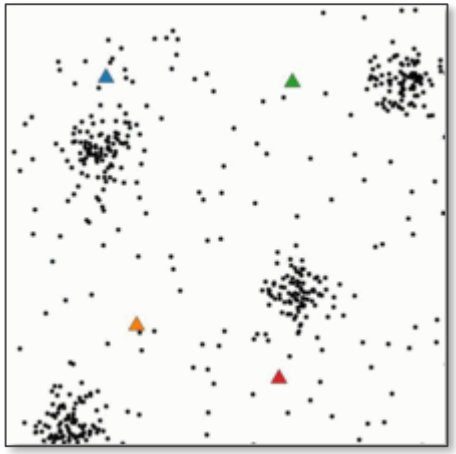
- **K-**평균 알고리즘**(K-means algorithm)**은 주어진 데이터를 **k**개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작한다.
- **K-**평균 클러스터링 알고리즘은 클러스터링 방법 중 분할법에 속한다. 분할법은 주어진 데이터를 여러 파티션 **(**그룹**)** 으로 나누는 방법이다. 예를 들어 **n**개의 데이터 오브젝트를 입력받았다고 가정하자. 이 때 분할법은 입력 데이터를 **n**보다 작거나 같은 **K**개의 그룹으로 나누는데, 이 때 각 군집은 클러스터를 형성하게 된다. 다시 말해, 데이터를 한 개 이상의 데이터 오브젝트로 구성된 **K**개의 그룹으로 나누는 것이다. 이 때 그룹을 나누는 과정은 거리 기반의 그룹간 비유사도 **(dissimilarity)** 와 같은 비용 함수 **(cost function)** 을 최소화하는 방식으로 이루어지며, 이 과정에서 같은 그룹 내 데이터 오브젝트 끼리의 유사도는 증가하고, 다른 그룹에 있는 데이터 오브젝트와의 유사도는 감소하게 된다. **K-**평균 알고리즘은 각 그룹의 중심 **(centroid)**과 그룹 내의 데이터 오브젝트와의 거리의 제곱합을 비용 함수로 정하고, 이 함수 값을 최소화하는 방향으로 각 데이터 오브젝트의 소속 그룹을 업데이트 해 줌으로써 클러스터링을 수행하게 된다.

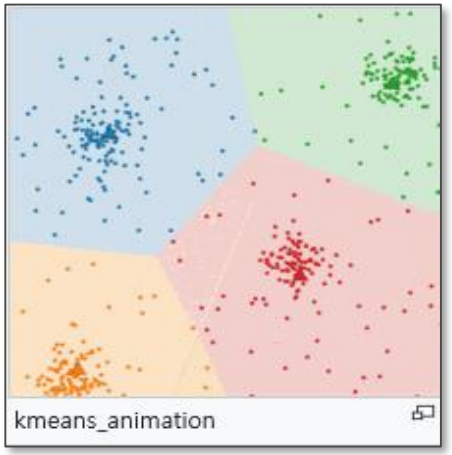
출처 : **https://ko.wikipedia.org/wiki/K-%ED%8F%89%EA%B7%A0_%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98**

3. K-평균 군집분석 - 한계점

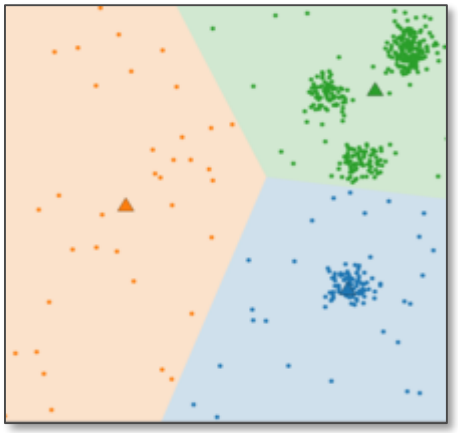
K-평균 알고리즘은 몇 가지 한계점을 가지고 있다**.**

- **클러스터 개수 K값을 입력 파라미터로 지정해주어야 한다.**
  - 이 알고리즘은 **K**값에 따라 결과값이 완전히 달라진다**.** 예를 들어**,** 실제 데이터가 **4**개의 클러스터를 가지고 있는데**, K=3**으로 입력했다고 가정하자**.** 그러면 아래와 같은 결과가 나올 수 있다**(**아래 맨 우측**).** 이는 실제 클러스터의 수 보다 **K**값이 작을 때 발생하는 현상이다**.**





K=4

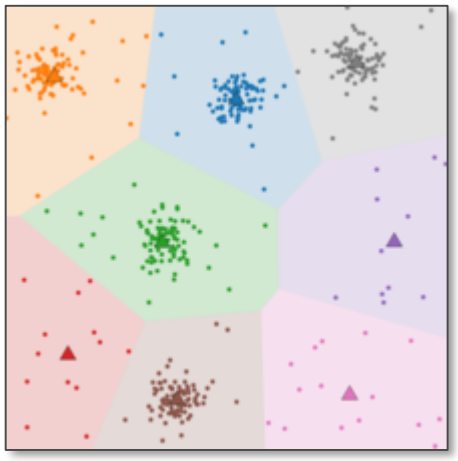


K=3

출처 **: https://ko.wikipedia.org/wiki/K-%ED%8F%89%EA%B7%A0_%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98**

3. K-평균 군집분석 - 한계점

- 반대로, 실제 클러스터가 **5**개인데 **K=8**을 입력했다고 가정하자. 결과는 아래 맨 좌측과 같다. 따라서, **K**값을 어떻게 주느냐에 따라 클러스터링의 결과가 극명하게 달라지며, 좋지 못한 결과를 보여줄 가능성이 있다

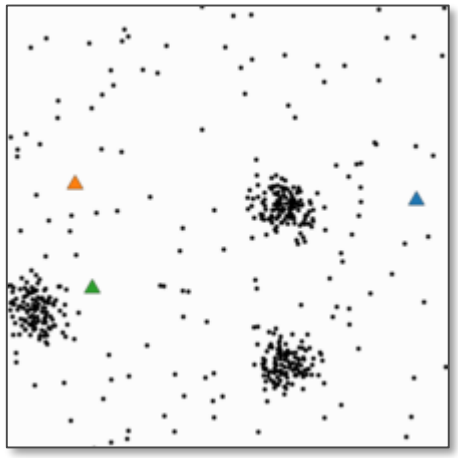


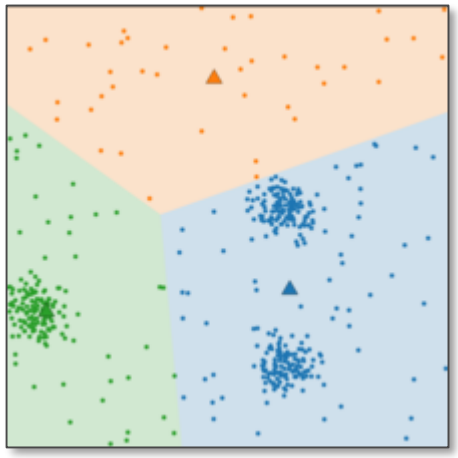
K=8

- **알고리즘의 에러 수렴이 전역 최솟값이 아닌 지역 최솟값으로 수렴할 가능성이 있다.**
  - 이 알고리즘은 초기값을 어떻게 주느냐에 따라 최적화의 결과가 전역 최적값 **(global optimum)** 이 아닌 지역 최적값 **(local optimum)** 으로 빠질 가능성이 있다

3. K-평균 군집분석 - 한계점

- 예를 들어, 실제 데이터는 **3**개의 클러스터를 가지고 있고, **K**도 **3**이라고 하자. 데이터와 초기 중심값의 분포는 아래 좌측과 같다.



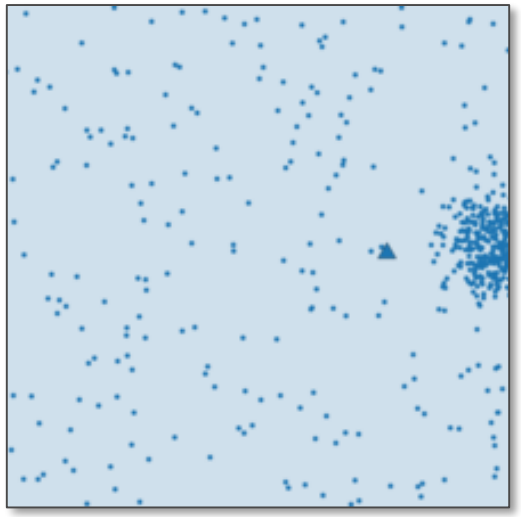


K=3

- 그러나 알고리즘을 수행하면서 비용 함수 최소화를 진행하면 위 우측과 같이 지역 최솟값에 빠져 그대로 수렴하게 된다. 즉, 비용 함수의 함수 공간에서 최적화를 시행할 때, 에러가 줄어드는 방향으로 최적해를 찾아가게 되는데, 전역이 아닌 지역 최솟값에 도달해도 알고리즘의 수렴 조건을 만족하게 되므로 더 이상 최적화를 진행하지 않게 된다. 따라서, 사용자가 기대한 결과를 얻지 못하게 되는 것이다.

3. K-평균 군집분석 - 한계점

- 이상값 **(outlier)** 에 민감하다.
  - **K-**평균 알고리즘은 이상값**(outlier)** 에 민감하게 반응한다. 이상값이란 다른 대부분의 데이터와 비교했을 때 멀리 떨어져 있는 데이터를 의미한다. 이러한 이상값은 알고리즘 내에서 중심점을 갱신하는 과정에서 클러스터 내의 전체 평균 값을 크게 왜곡시킬 수 있다. 따라서 클러스터의 중심점이 클러스터의 실제 중심에 있지 않고 이상값 방향으로 치우치게 위치할 수 있다. 이를 방지하기 위해 **K-**평균 알고리즘을 실시하기 전에 이상값을 제거하는 프로세스를 먼저 실행하거나, 분할법의 일종인 **K-**대표값 알고리즘 **(K-medoids algorithm)** 을 이용하면 이상값의 영향을 줄일 수 있다.



클러스터에서 크게 멀어진 중심점

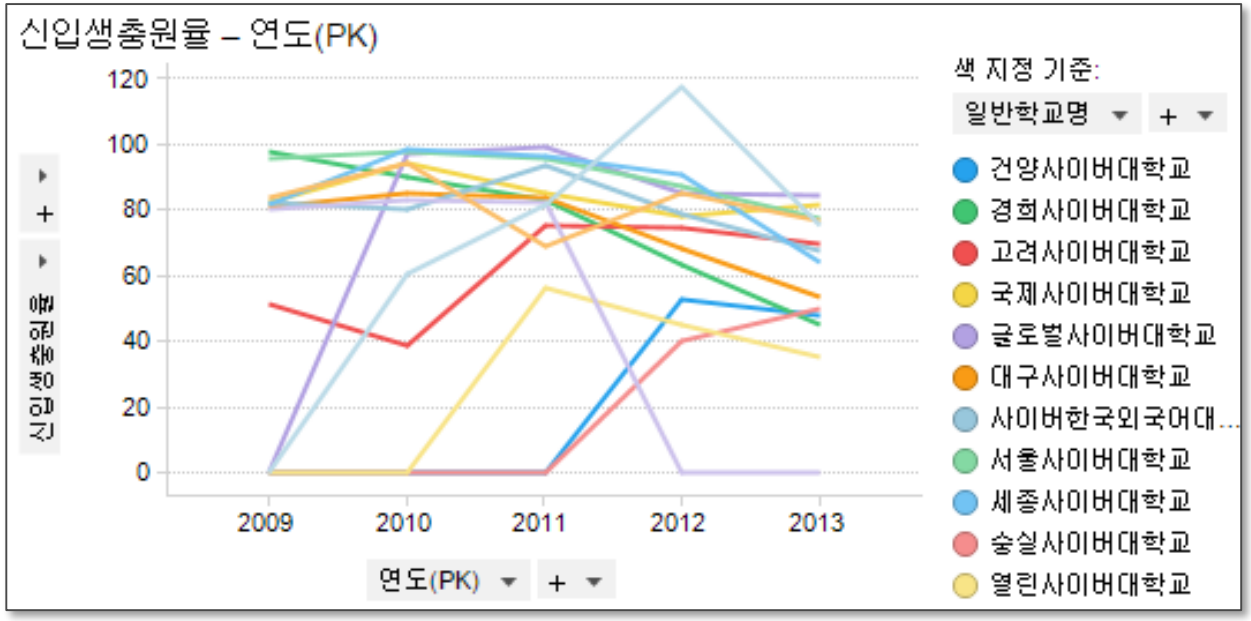
* 데이터테이블 : 대학정보공시(사이버대).txt

3. K-평균 군집분석

- **Spotfire**에서 **K**군집 평균 분석을 쉽게 수행할 수 있다.

  1. Spotfire에서 데이터를 가져온다.

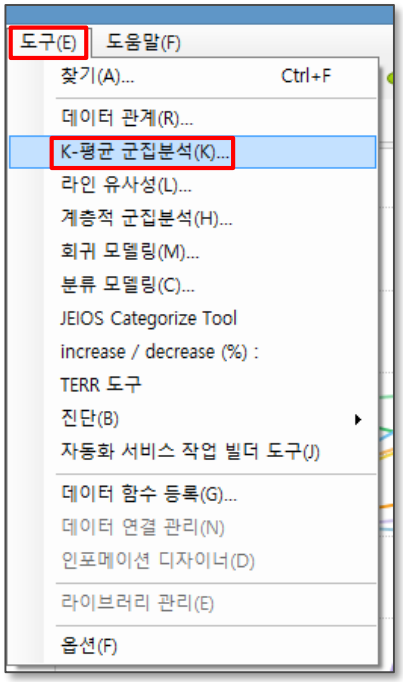
  2. K군집 평균 분석을 실행하기 위해서는 먼저 라인차트나 평형좌표 그래프를 생성한다.
  이 예제에서는 라인차트를 생성하고 아래와 같이 설정해 본다.



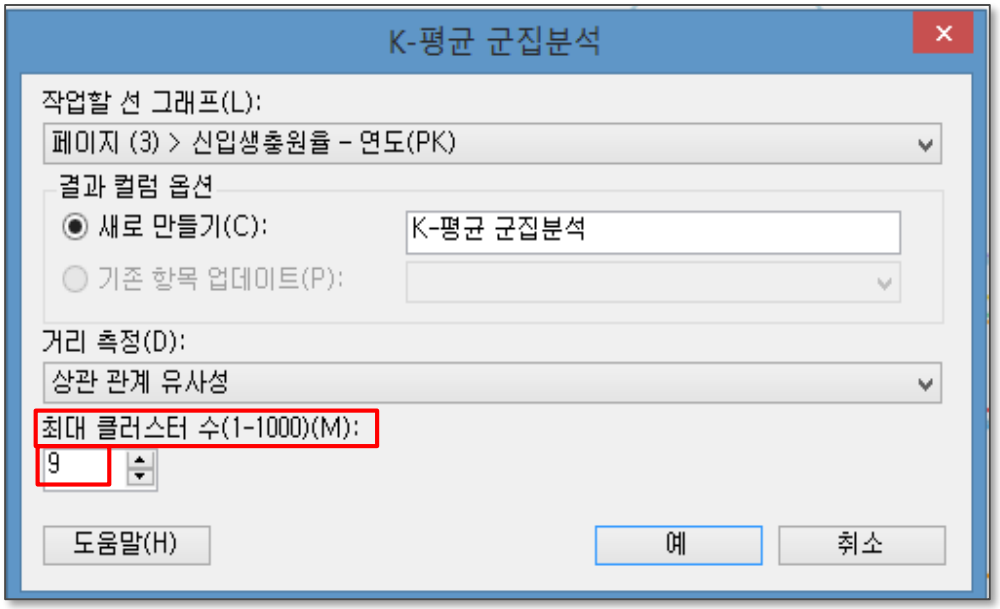
* 데이터테이블 : 대학정보공시(사이버대).txt

3. K-평균 군집분석

3. **Spotfire**의 메인 메뉴에서 ‘도구’ > ‘**K-**평균 군집분석’을 실행한다.



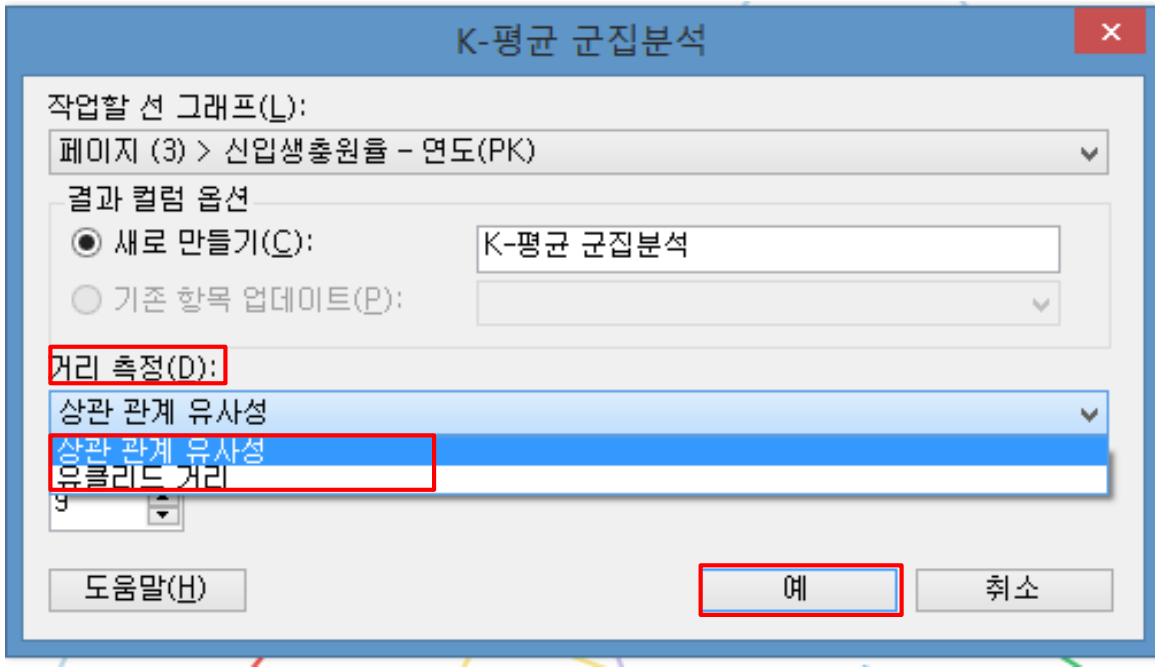
4. **Spotfire**의 메인 메뉴에서 원하는 ‘최대 클러스터 수’를 입력한다.



* 데이터테이블 : 대학정보공시(사이버대).txt

3. K-평균 군집분석

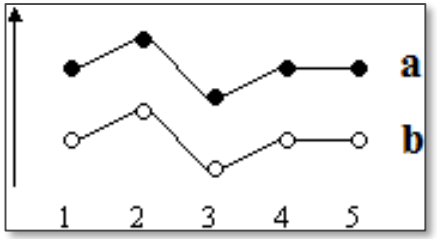
5. ‘거리 측정’에서 원하는 옵션을 선택하고(**default**는 ‘상관 관계 유사성’ 옵션이다.) 예를 누른다.



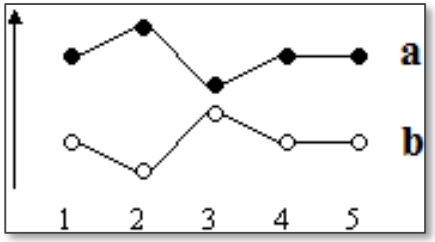
3. K-평균 군집분석 – 거리측정 옵션

상관 관계(Corelation relation)

이 상관 관계는 ***Pearson Product Momentum Correlation***라고 하며, 줄여서 ***'Pearson의 상관 관계'*** 또는 ***'Pearson의 r'*** 이라고도 한다. 범위는 **+1 ~ -1**이고, **+1**이 가장 높은 상관 관계이다. 정반대인 지점은 상관 관계가 **-1**이다.



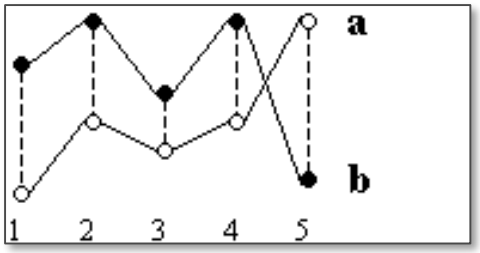
a 및 ***b***가 동일하면 최고의 상관 관계이다.



a 및 ***b***가 완벽하게 미러링되면 최고의 음수 상관 관계를 나타낸다.

유클리드 거리(Euclidean Distance)

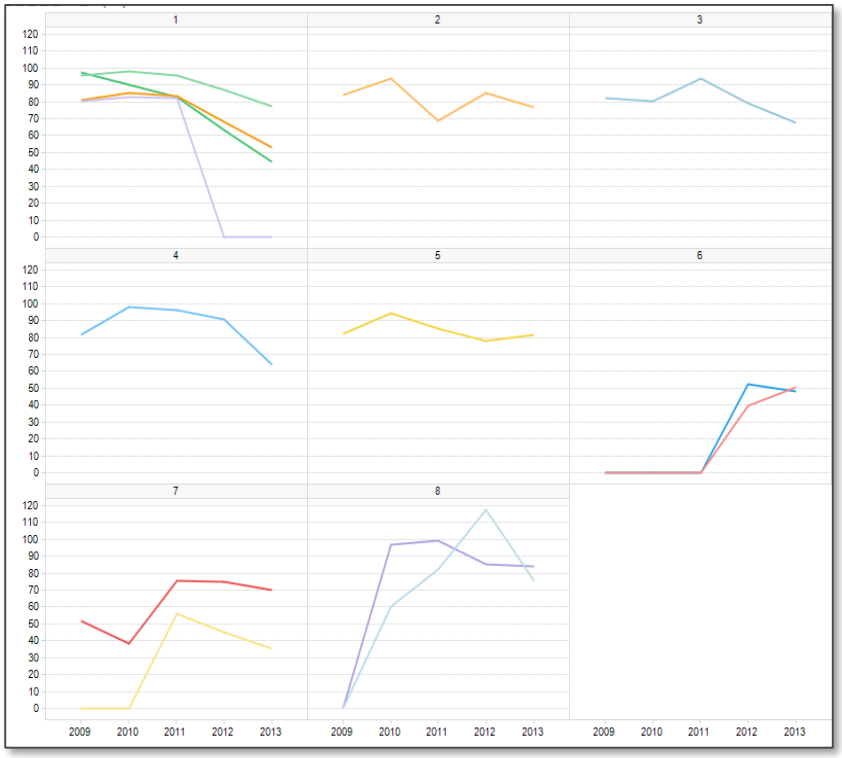
유클리드 거리는 항상 **0**보다 크거나 같아야 한다. 동일한 지점에 대해서는 측정치가 **0**이 되고 유사성이 거의 없는 지점에 대해서는 측정치가 높게 나타난다.
아래 그림은 ***a*** 및 ***b*** 지점의 예를 보여준다. 각 지점은 **5**개의 값으로 설명된다. 그림에서 점선은 거리 (a_1-b_1), (a_2-b_2), (a_3-b_3), (a_4-b_4) 및 (a_5-b_5) 이다.



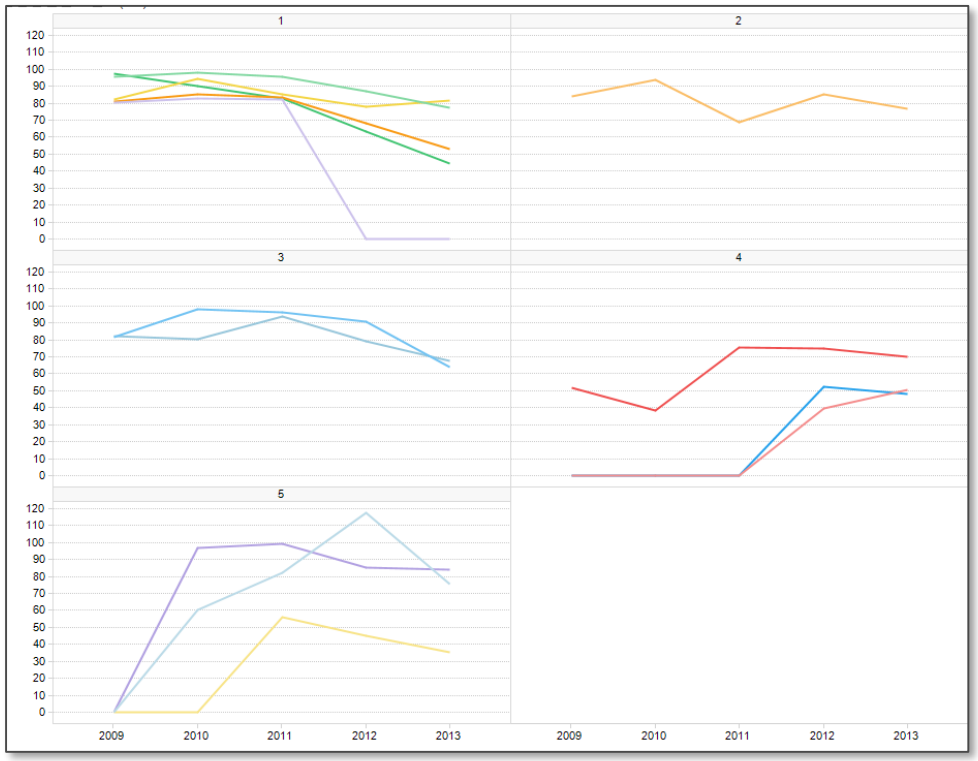
* 데이터테이블 : 대학정보공시(사이버대).txt

3. K-평균 군집분석

6. 최대 클러스터 수를 **default**인 **'9'** 로 수행해 본다.



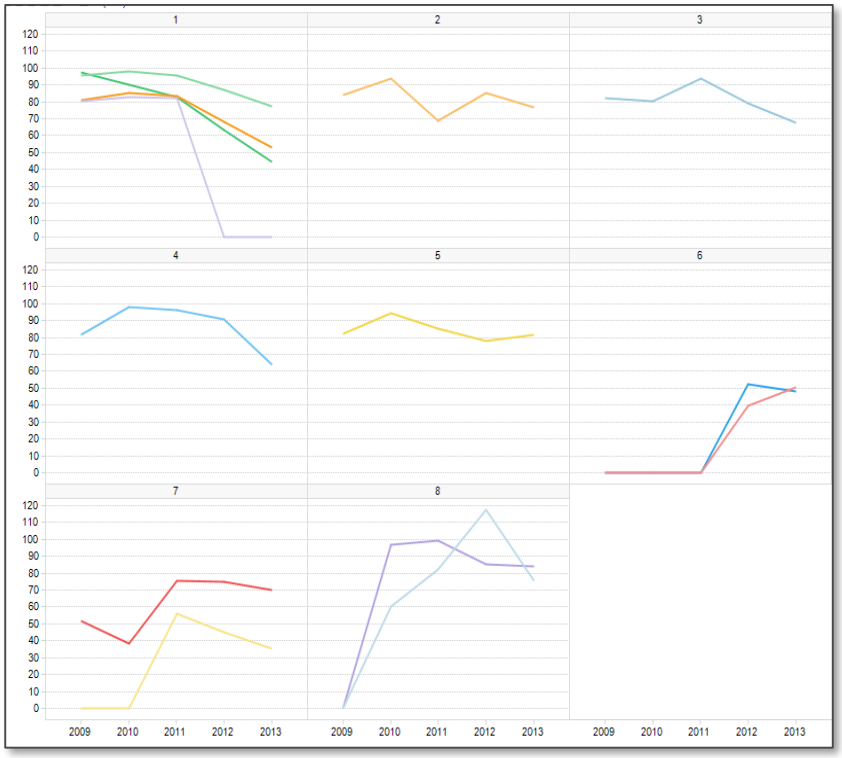
7. **6**번의 결과를 보고 유사한 추세를 감안하여 최대 클러스터 수를 **'5'** 로 수행해 본다.



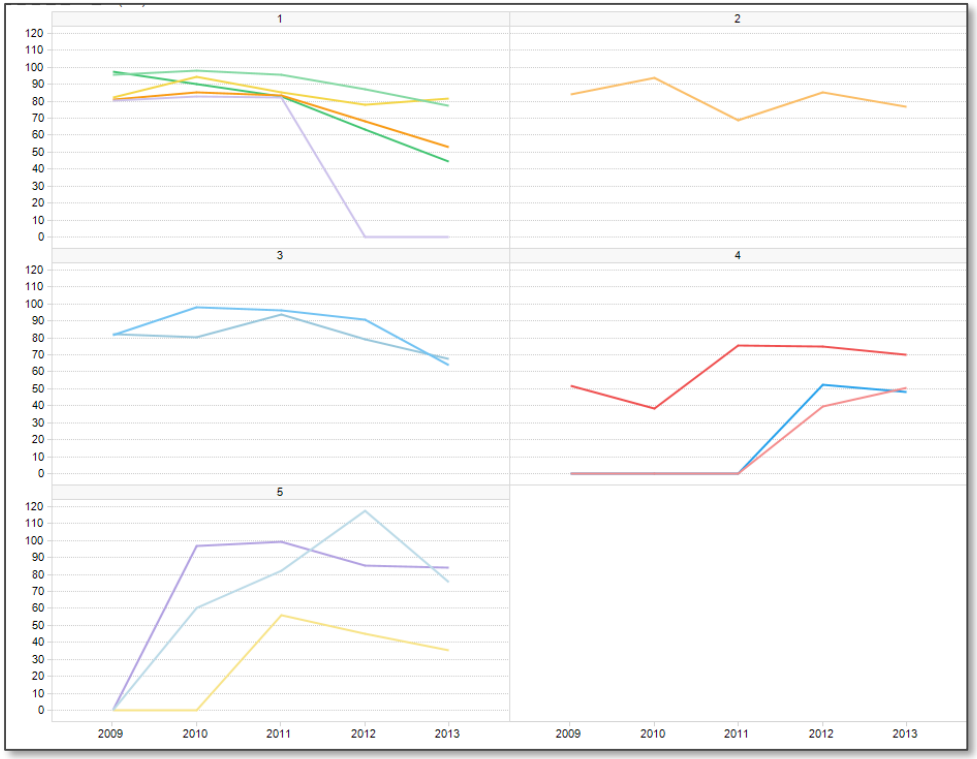
* 데이터테이블 : 대학정보공시(사이버대).txt

3. K-평균 군집분석

6. 최대 클러스터 수를 **default**인 **'9'** 로 수행해 본다.



7. **6**번의 결과를 보고 유사한 추세를 감안하여 최대 클러스터 수를 **'5'** 로 수행해 본다.

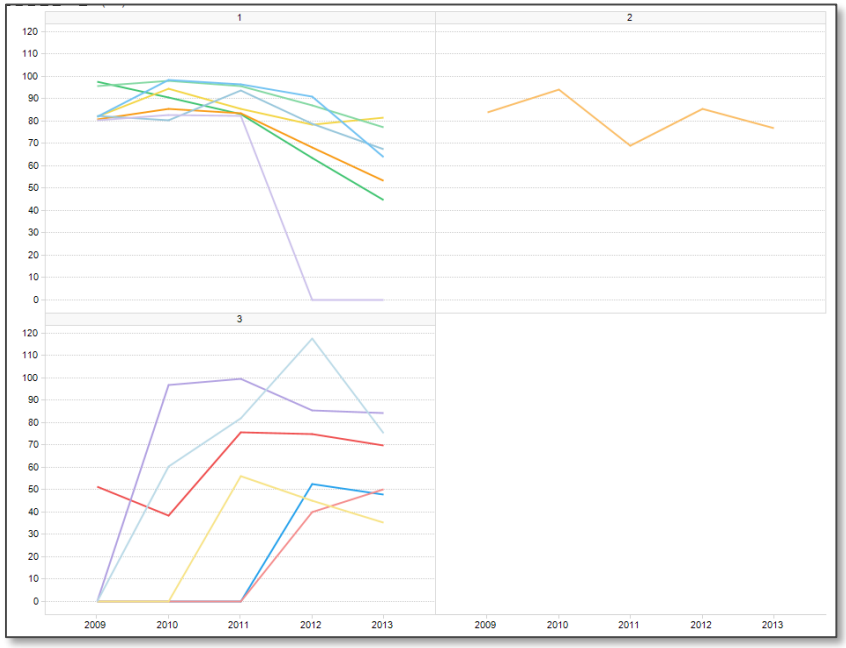


* 데이터테이블 : 대학정보공시(사이버대).txt

3. K-평균 군집분석

8. **7**번의 결과를 보고 다시 유사한 추세를 감안하여 최대 클러스터 수를 **‘4’** 로 수행해 본다.

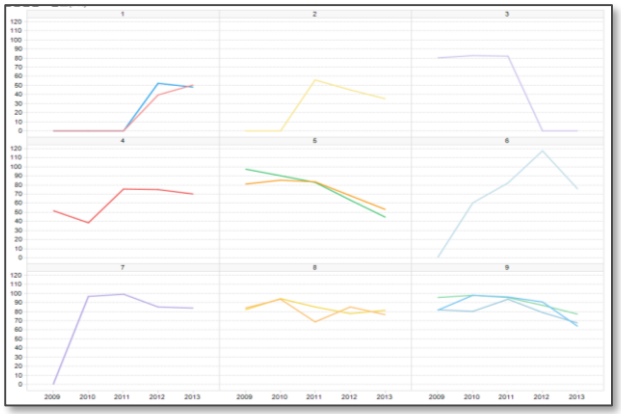
9. 아래는 최대 클러스터 수를 **‘3’** 으로 수행한 결과이다. **8**번의 결과보다 만족스럽지 못하다.



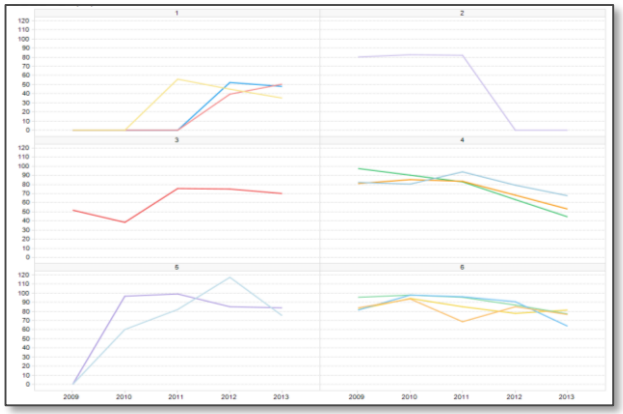
* 데이터테이블 : 대학정보공시(사이버대).txt

3. K-평균 군집분석

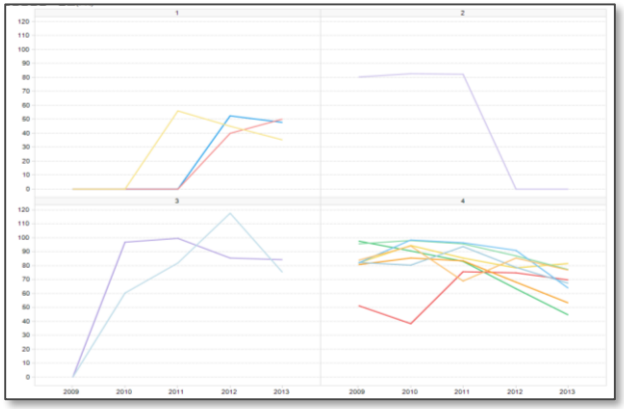
10. 아래는 ‘거리 측정’ 옵션을 ‘유클리드 거리’로 하고 ‘최대 클러스터 수’를 변경해 가며 수행해 본 결과이다.



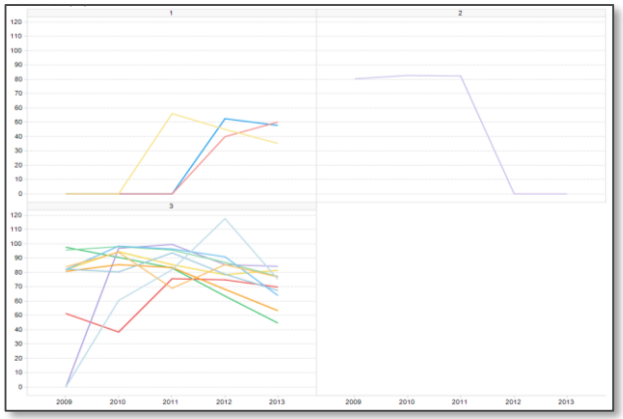
최대 클러수터 수 : 9



최대 클러수터 수 : 6



최대 클러수터 수 : 4



최대 클러수터 수 : 3