


## 1) 정의

- 수집한 데이터가 들어왔을때, 이를 다양한 각도에서 관찰하고 이해하는 과정
- 데이터를 분석학 전에 그래프나 통계적인 방법으로 자료를 직관적으로 바라보는 과정임



- ✓ 데이터의 분포 및 값을 검토
- ✓ 하여 데이터가 표현하는 현상 이해
- ✓ 데이터에 대한 잠재적 문제 발견
- ✓ 문제정의 단계에서 발견하지 못했던 다양한 패턴 발견

- ✓ 분석전 데이터 수집 결정
- ✓ 기존의 가설을 수정하거나 새로운 가설을 세울수 있음.

## 2) 필수요소

구분	요인분석 수행목적
결측치 유무:	<ul style="list-style-type: none"><li>제거 / 대체(평균, 중앙값, 추정(시계열데이터에 또는 관계성이 있다면 결측치를 제외하고 값이 있는 것들끼리 모델링한후 결측치에 값 추정하여 넣음)등)</li><li>새로운 값으로 코딩(예: 미응답은 NoAnswer로 새값 작성)</li></ul>
변수의 분포	<ul style="list-style-type: none"><li>정규분포, 표준편차, 분산, 왜도, 첨도 값의 범위가 극단적이면: 정규화 / 분포가 극단적이면: 로그변환 또는 box-cox 변환</li><li>sparse data(0행열)</li></ul>
연속형 변수중 이상치	<ul style="list-style-type: none"><li>이상치(outlier)와 레버레이지(leverage) 구별 / 상관관계</li></ul>
빈도수	<ul style="list-style-type: none"><li>빈도수가 유독 적은 카테고리는 없는지</li></ul>
feature 들끼리의 관계성	<ul style="list-style-type: none"><li>설명변수와 종속변수간의 관계성: 다중공선성</li></ul>

## 2. 차원 축소(Dimensionality Reduction)

[들어가기](#)[▲ 학습하기](#)[정리하기](#)

### 1) 차원의 정의

- 수학에서 공간내에 있는 점 등의 위치를 나타내기 위해 필요한 축의 개수
- 특징공간: 관측값들이 있는 공간이며 여러 차원으로 구성될수 있으며 n개의 특징에 따라 n개의 차원이 형성됨. 즉 y값을 구성하는 x변수가 여러개일수 있음.

1차원

X
1
1
1

3건

2차원

X1	X2
1	2
1	3
1	4

9건

3차원

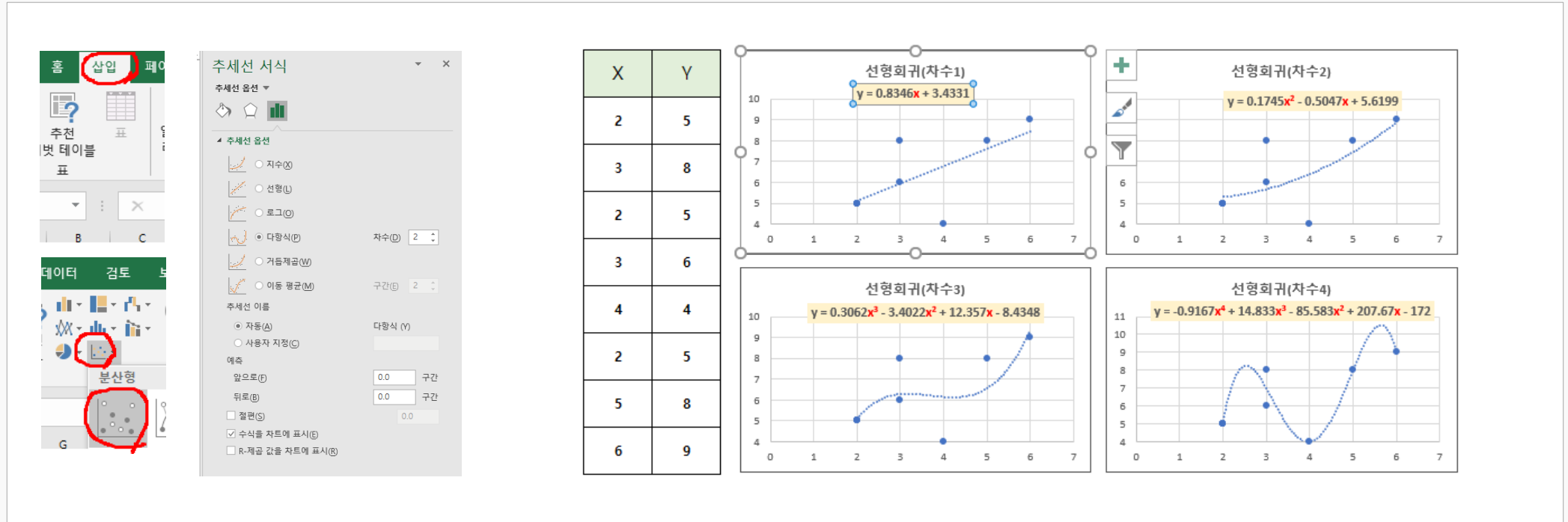
X1	X2	X2
1	2	3
1	3	4
1	4	5

27건

## 2. 차원 축소(Dimensionality Reduction)

[들어가기](#)[▲](#)  
[학습하기](#)[정리하기](#)

### 2) 차원 선형 & 다항식 예시



### 3) 차원의 축소

- 차원의 저주(curse of Dimensionality)

- 특징의 수가 늘어나 차원이 커지면서 발생하는 문제를 차원의 저주라함.
- 다항식( $y = \omega_0 + \omega_1 x + \omega_2 x^2 + \dots + \omega_N x^N$ ) 에서 차원이 많은 경우
- 차원을 더 늘릴 경우 필요한 데이터 양은 기하급수적으로 늘어나는 현상

- 고차원은 새로운 샘플로 훈련 샘플과 멀리 떨어져 있을 가능성이 높다는 뜻.
- 훈련 샘플의 밀도가 충분히 높아질때까지 훈련 데이터의 크기를 키우는 것임.

- 차원 축소

- 차원 축소: 고차원의 데이터를 저차원의 데이터로 변환하는데 사용됨
- 기대효과: 머신러닝에서는 모델의 성능을 강화시켜주는 것이고, 통계적으로는 적은수의 특징만으로 특정 현상을 설명하려고 할때 사용함.

### 1) 요인분석

- 요인분석: 여러개의 특징들을 몇개의 잠재된 특징으로 찾아내는 것을 요인분석이라함
  - 특징들간의 상관관계를 고려하여 서로 유사한 변수들끼리 묶어줌
  - 많은 특징으로 구성된 데이터가 몇개의 요인에 의해 영향을 받은가를 알아볼수 있음.

구분	요인분석 수행목적
입력변수들의 특성파악	• 데이터 이해 과정(여러 특징들간의 상관)
새 특징값의 생성효과 (전처리)	• 원래특징값보다 더 적절한 값을 생성해줌. (즉 잠재변수 추가가능)
데이터축소	• 특징값의 개수를 줄여줌. 결정트리나 회귀분석에서 단순하게 모델가능
다중공선성 문제해결	• 강한상관관계시 분석 결과 모델의 오류 또는 결과 해석의 오류생김방지

### 2) 주성분분석(PCA:Principal component analysis)

- 차원축소의 두가지 방법중(Feature Selection과 Feature extraction)중 후자에 해당함
  - 여러개의 양적변수들 사이의 분산-공분산 관계를 이용하여 변수들의 선형결합으로 표시되는 주성분을 찾고, 2-3개의 주성분으로 전체변동의 대부분을 설명하고자 하는 **다변량 분석법**
  - 분산의 최대축 feature를 찾아야함.

#### 【 첫째 주성분 】

데이터 프레임의 총 변동을 대부분 설명할 수 있는 변수 선형 조합을 찾아내야함.

#### 【 둘째 주성분 】

첫번째 주성분과는 상관이 낮아서 첫번째 주성분이 설명하지 못하는 나머지 변동을 정보 손실없이 가장 많이 설명할 수 있도록 변수의 선형조합을 만듦

### 3) 차이점과 공통점

#### 공통점

- 데이터를 축소함.
- 새로운 데이터를 만들수 있음.

#### 차이점

- 생성되는 변수의 수(요인분석:정의되지않음, PCA:보통 2개찾음)
- 생성되는 변수의 의미(요인분석: 분석가가이름지정, PCA:제1,제2..)
- 생성된 변수들의 관계(요인분석: 대등함, PCA:제1이 가장중요함)
- 분석방법의 의미(요인분석: 변수를 비슷한것끼리 묶어서 새로운 변수생성

PCA:종속변수를 고려한 주성분을 찾아냄)



### 1) 정의

- 다중공선성이 존재하는 데이터로 의사결정트리를 생성할 경우 분석결과에 문제 생김
  - 문제1: 중요한 변수를 발견하지 못함(상관성이 높은 변수를 트리에서 표시하지 않음)
  - 문제2: 결과 모델의 성능(분류율)이 떨어짐
    - ✓ 결정트리는 나무 모델을 생성하는 과정에서 한 노드에서 한 변수만을 선택하여 데이터를 분류함.
    - ✓ 즉 결정트리는 하나의 특징값에 수직적으로 데이터가 잘 분리될때 좋은 성능을 지님
    - ✓ 다중공선성이 있는 자료는 두 개 이상의 특징값에 자료가 연결되어 있어 잘 구별되지 않음.
- 다중공선성이 존재할 경우 상관도가 높은 변수들을 하나의 주성분 혹은 요인으로 축소

## 2) R에서 PCA함수

```
##### 주성분분석 PCA #####
```

```
pca_iris = prcomp(tmp, center=T, scale.=T)
pca_iris
print("----- PCA 요소 값 확인 -----")
```

```
### ==> PC1(제1요인)이 가장 편차(Standard deviation)가 크고, PC2(제2요인)가 다음으로 크다.
summary(pca_iris) # Cumulative Proportion(누적비율값 확인)
```

### PCA 1~4의 편차와 누적비율

[1] "----- PCA 요소 값 확인 -----"

Importance of components:

	PC1	PC2	PC3	PC4
Standard deviation	1.7084	0.9560	0.38309	0.14393
Proportion of Variance	0.7296	0.2285	0.03669	0.00518
Cumulative Proportion	0.7296	0.9581	0.99482	1.00000

### 고유벡터의 계수(특징값과 같은개수나옴)

Rotation (n x k) = (4 x 4):

	PC1	PC2	PC3	PC4
Sepal.Length	0.5210659	-0.37741762	0.7195664	0.2612863
Sepal.Width	-0.2693474	-0.92329566	-0.2443818	-0.1235096
Petal.Length	0.5804131	-0.02449161	-0.1421264	-0.8014492
Petal.Width	0.5648566	-0.06694199	-0.6342727	0.5235971

꽃잎 또는 길이

꽃받침 또는 너비

