# 데이터 과학자

인용도서 – 데이터분석:R
한빛미디어

pykwon@hanmail.net

비즈니스에 관한 경험, 지식
(기획, 영업)
고객의 마음을 알아내는 능력

설계기법에 관한 경험, 지식
(학자, 연구원)
통계해석 능력

그래프화
크로스집계
바스켓분석
클러스터링

통계 가설 검정
회귀분석
분류
머신러닝
딥러닝

Python, R, etc

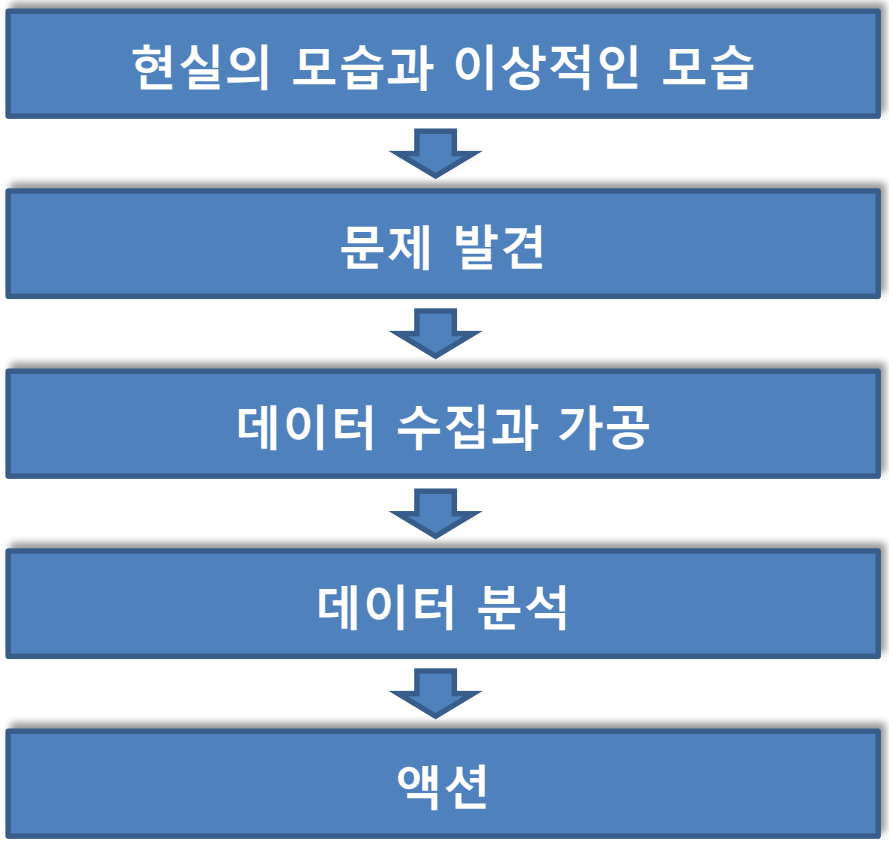
기계학습 활용
수량화 이론
SQL
강화학습
Echo system
저장소
(File,
RDBMS,
Hadoop)


3가지 필요조건

데이터 처리에 관한 경험, 지식
(IT 엔지니어)
데이터 처리 프로그래밍 작성 능력

# Business에서 데이터 분석의 흐름

데이터 분석이란 현재 상황으로부터 이상적인 모습에 빨리 도달하기 위해 문제를 추출하는 것에 주안점을 두고 아래와 같은 순서대로 실시하면 효과적이다.

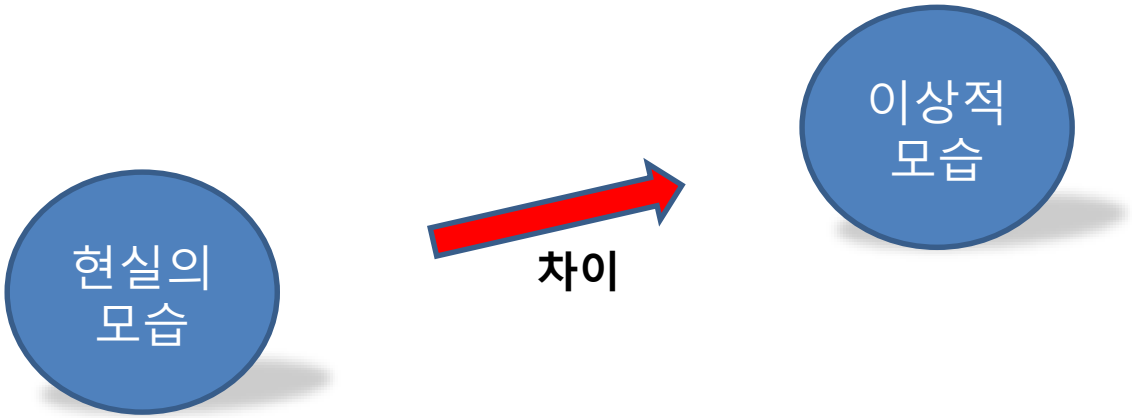



비즈니스에서 데이터 분석은 비즈니스에서 발생한 여러 가지 질문을 통계해석이나 기계학습, 데이터 마이닝 등의 각종 방법론을 구사하여 해결하는 것이 목적이다.
이 때에 고도의 복잡한 모델을 통해 분석한 결과가 반드시 높은 가치가 있다고 볼 수는 없다. 간단하게 분할표를 만들어 단시간에 분석결과를 얻어 내는 것이 더 큰 가치가 있을 수도 있다.
주어진 문제에 대해 맞는 분석기법을 설계하고 실행하는 것이 가장 중요하며 이 것을 간과하면 데이터 분석결과는 투자에 비해 가치가 떨어질 가능성이 있다는 점을 잊지 말자!

# 현실의 모습과 이상적인 모습

**문제**는 어떤 시점에 어떠한 비즈니스가 처한 환경의 "이상적인 모습"에 의해 바뀔 수 있다. 즉, "이상적인 모습"과 "현실의 모습" 사이에 차이가 있어야 "문제가 있다."라고 할 수 있다.

예를 들어 '어떤 상품의 매상이 오르고 있다.' 라는 현상을 보자. 보통 매상이 오르는 상황이라면 별 문제가 없을 것 같지만 '실제로 그 상품에 들이고 있는 광고비에 맞지 않는 매상' 이라면 문제가 될 수 있다.



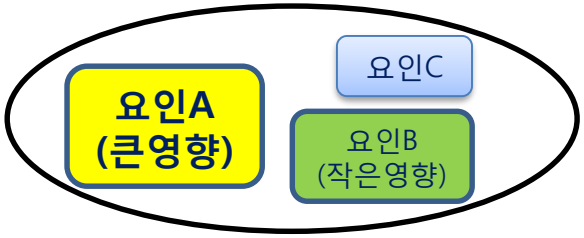
# 문제 발견

**데이터 분석에서 '현상'과 '문제'는 명확히 구분할 필요가 있다.** '현상'이 공유되고 '이상적인 모습'이 공유되었을 때 데이터 분석을 수행할 토대가 마련되었다고 할 수 있으며, 데이터 분석으로 문제를 해결해 나갈 수 있다.

| 현상 | 전체 | 이상적 모습 | 문제인가? |
|---|---|---|---|
| 매상이 오름 | 광고비용이 높다 | 광고비용 낮춤 | 문제 |
| | 광고비용이 적당 | 지금 상태 유지 | 문제 아님 |

*** 데이터 분석을 실시(문제발견)할 때의 3가지 관점**

1) **크기 보기**
- 요인의 영향도




2) **분해해서 보기**
- 일어난 현상의 구성요소를 다양한 주제로 분해하여 현상의 원인 요소를 찾기
- 상호배제, 전체포괄


**어느 요인을 조절하는 것이 매상을 늘리기 쉬운가?**

3) **비교해서 보기**
- 문제가 있는 데이터와 없는 데이터를 비교해서 차이에 대한 발생원인을 찾기

# 데이터 수집과 가공

*** 데이터 수집 시 검토사항**

- 문제를 검증하기 위해 어떤 데이터가 필요한가?
- 분석자가 사용할 수 있는 곳에 필요한 데이터가 보존되어 있고 사용이 가능한가?
- 필요한 데이터가 없는 경우 새로 데이터를 취득할 수 있는가?
- 보존된 데이터가 없는 경우 대체 데이터가 있는가?

* 필요한 데이터가 보존되어 있다면 대개 '파일', '데이터베이스', '하둡' 등에 저장되어 있다.

* 데이터 가공은 데이터 분석의 목적, 보존상태, 형태 등에 따라 달라질 수 있기 때문에 기본적으로는 개별적으로 다루어야 한다. 실무에서 분석을 수행할 때는 데이터 가공을 적절히 수행할 수 있는지 여부가 가장 중요한 포인트가 된다.

- 데이터 결합
- 판정용 변수 작성
- 이산화 변수 작성


데이터를 수집한 후에는 데이터를 분석에 맞는 형식으로 바꿔 주는 **전처리 과정**을 거쳐야 한다.  예를 들어 데이터를 수집할 때는 남/녀라는 대답을 수집했다고 합시다. 그런데 우리가 실질적으로 데이터를 분석할 때는 수식에 입력해야 하는데 이 때 남/녀를 넣을 수는 없다. 그래서 이런 데이터를 수치화해서 바꿔줘야 한다.
남자는 1 여자는 0 뭐 이런 식으로. 이래야 분석을 할 수 있는 명목형 데이터가 된다.

더불어서 해줘야 하는 일은 **결측값 처리**다. 설문지조사나 직접 데이터를 얻을 때 주로 발생하는 상황으로 데이터가 비어있는 경우에는 이를 처리해 줘야 하겠다. 이를 처리하는 방법도 다양한데 그 중에 결측치를 버리는 방법도 있고 데이터 수가 적다면 평균값으로 대체하거나 다른 수로 대체하는 방법 등도 있다.

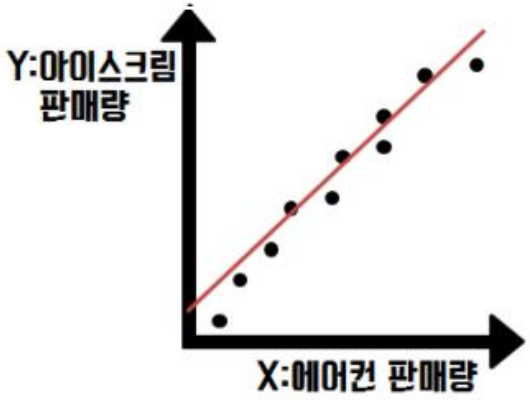
# 데이터 분석

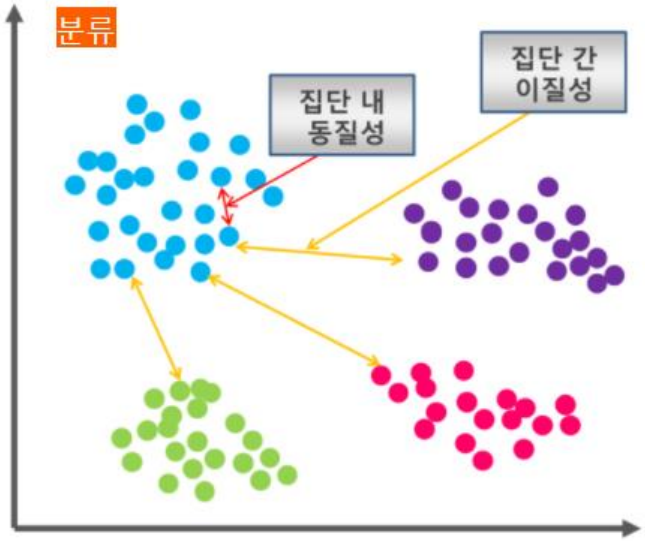
데이터 분석은 문제의 종류에 따라 '의사결정지원'과 '자동화-최적화'로 나눌 수 있다.

| | 의사결정지원 | 자동화-최적화 |
|---|---|---|
| 목 적 | 사람이 행동결정을 지원 | 컴퓨터가 스스로 행동결정 |
| 목 표 | 의사소통 비용을 절감 | 추정 정확도 향상, 계산량 삭감 |
| 사용기법 | 단순집계, 크로스 집계<br>(예측모델 구축) | 기계학습, 알고리즘 구축 |

**회귀분석** X와 Y는 어떤 관계일까?

:변수들의 관련성을 규명하기 위하여 어떤 수학적 모형을 가정하고, 이 모형을 측정된 변수들의 데이터로부터 추정하는 통계적인 방법
독립변수의 값에 의하여 종속변수의 값을 예측하기 위함



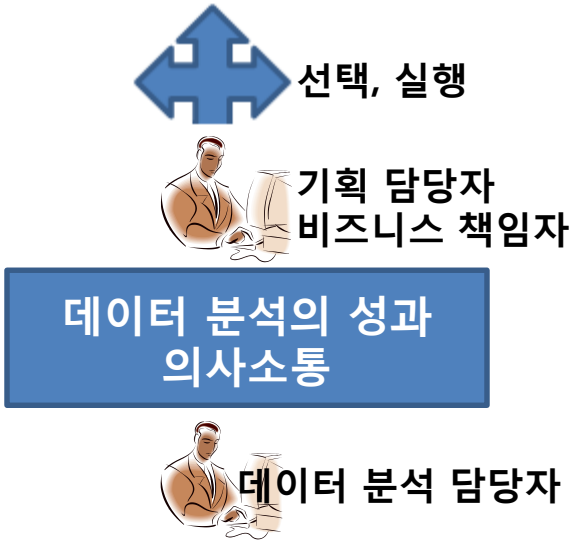




데이터가 완성되었다면 우리가 어떤 분석을 해야 할 지 결정을 해야 한다. 분석방법에는 그룹 비교나 자료 추이, 자료 예측 등 매우 다양하다.
그렇다면 어떤 기준으로 방법을 정해야 할까? 그건 자신이 분석하고자 하는 주제에 따라 결정해야 한다.
어떤 그룹과 어떤 그룹의 데이터 비교를 하기 위해서는 **그룹비교**,
데이터의 전체적인 변화 모습을 알아보고 싶다면 **자료 추이**,
그리고 미래의 데이터를 구하고 싶다면 **자료 예측**의 방법을 선택하고
각 카테고리 별로 적절한 모델이나 방법에 맞춰 분석을 진행해야 한다.

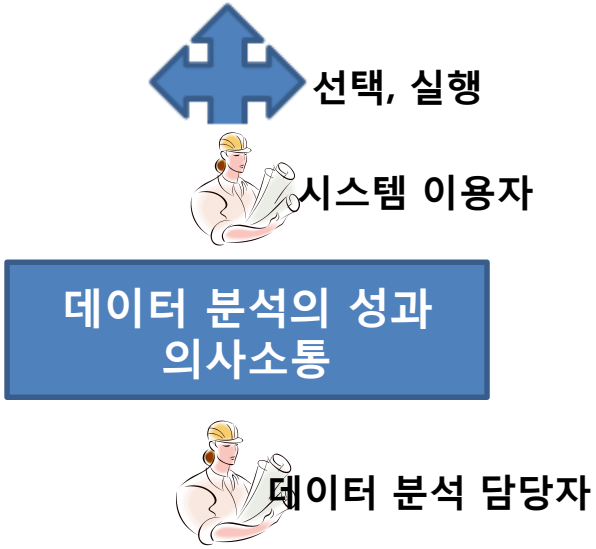
# 액션

데이터 분석의 최종 단계에서는 분석결과를 가지고 실제로 행동으로 옮길지 여부를 검토한다. 액션에는 '사람이 의사결정을 해서 새로이 뭔가를 시작 또는 관두는 것(**의사결정 지원**)'과 '액션을 실행하기 위한 알고리즘을 구축해서 컴퓨터에 실행시키는 것(**자동화-최적화**)'의 두 가지 종류가 있다.

종류마다 실시하기까지 필요한 의사소통 비용이 다르다. 대부분의 조직에서 '의사결정지원'은 상사나 기획팀에 대한 설득비용이 크며, '자동화-최적화'는 개발팀 혹은 서비스 운영팀에 대한 설득비용이 큰 경우가 많다.




**의사결정 지원의 경우**




**자동화-최적화의 경우**