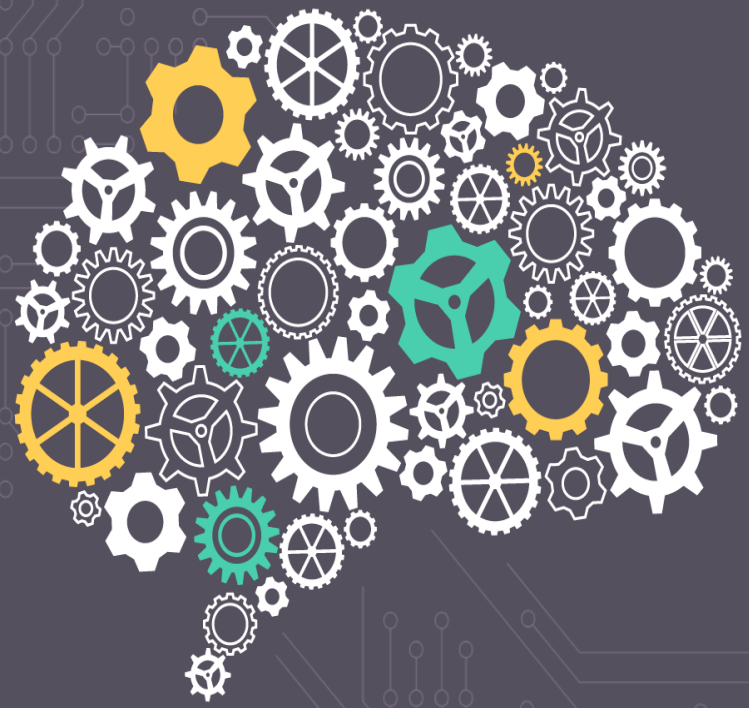
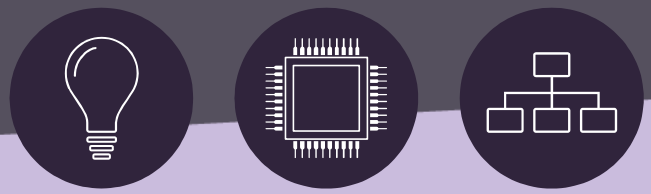
[디지털헬스케어개론]



빅데이터분석을 위한

데이터전처리

창의융합대학 **MSC**교육부 유 현 주



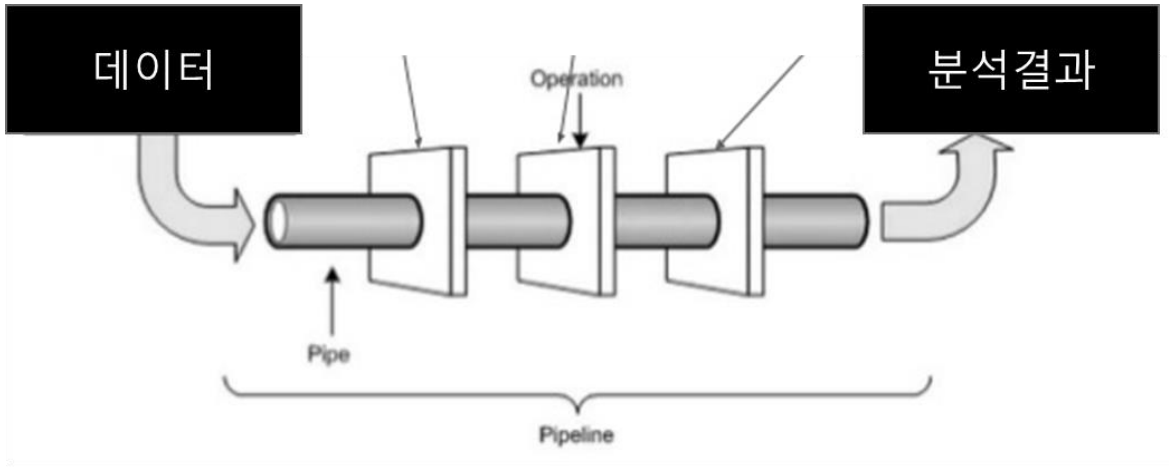
Contents

데이터전처리

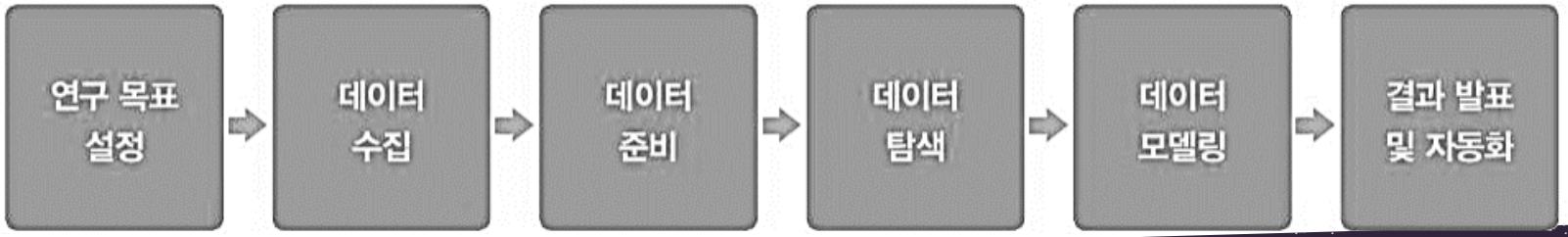
- 정형데이터의 전처리
- 탐색적 데이터 분석(EDA)

데이터 분석 파이프라인

- 데이터 중심 사회에서 데이터를 효과적으로 가져오고 분석하는 일은 매우 중요한 작업
- 데이터를 절차에 따라 분석을 진행 → '빅데이터 분석 파이프라인'

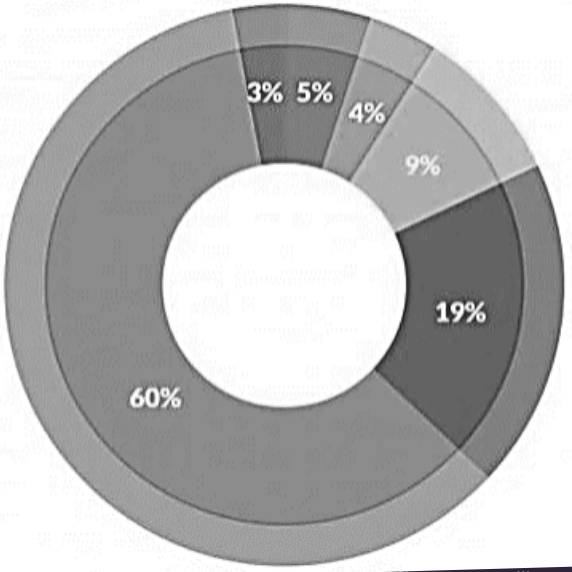


- 데이터 과학 방법론



데이터 전처리

- 데이터를 분석 및 처리에 적합한 형태로 만드는 과정을 총칭하는 개념으로 **데이터 전처리 (Data preprocessing)**, 또는 데이터 조작 (Data Handling, Data Manipulation, Data Wrangling, Data Munging)이라 함
- 수집된 데이터의 형식과 구조는 매우 다양하므로 데이터에 분석 기능을 적용하기 위하여 데이터를 가공, 정리하는 과정
- 데이터 사이언티스트 업무 시간 중 70% 이상을 데이터 수집 및 전처리 과정에 사용
- 아무리 좋은 도구나 분석 기법도 품질이 낮은 데이터로는 좋은 결과를 얻기 힘듦

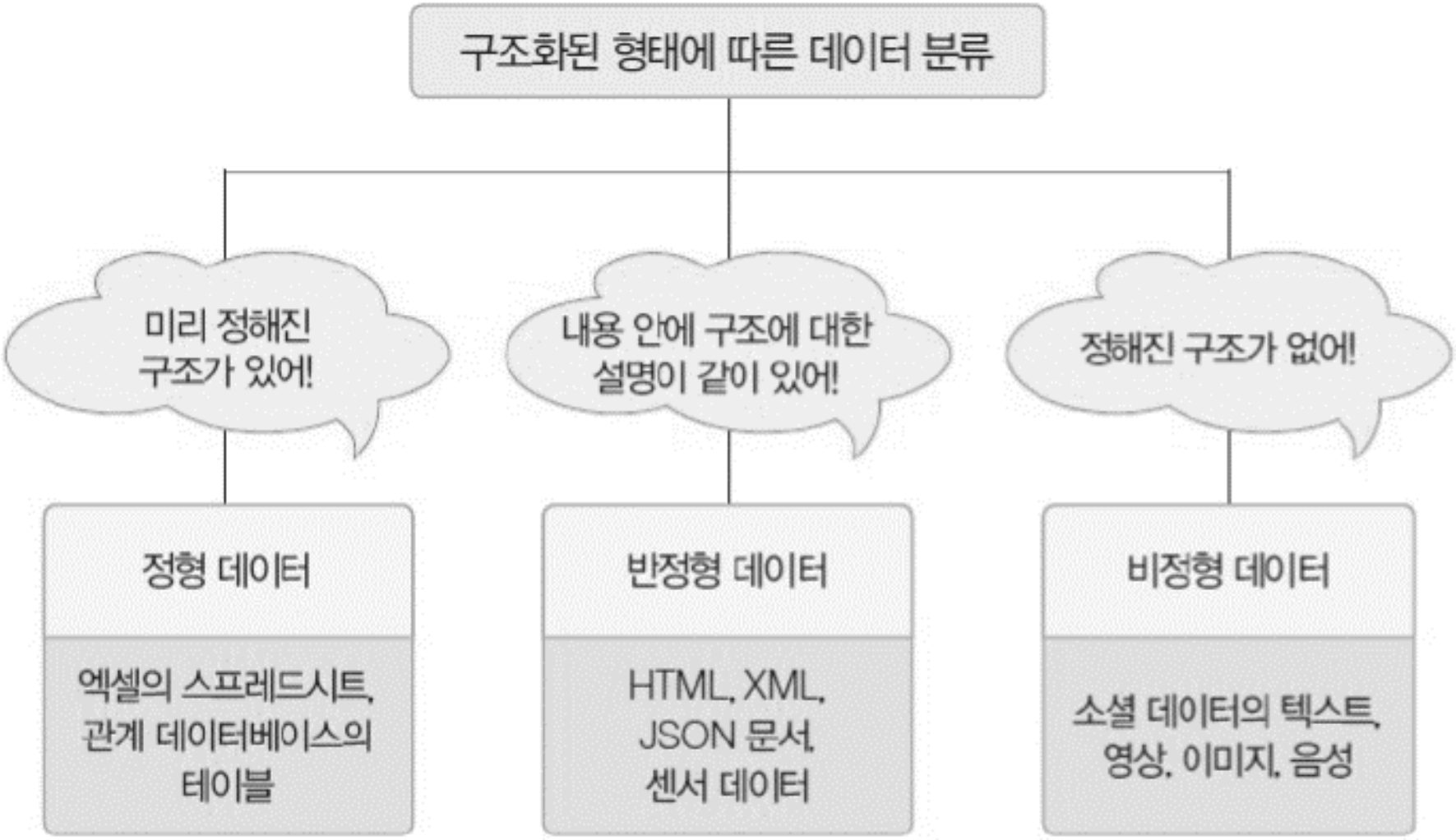


What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

정형데이터의 전처리

데이터 유형


구조화된 형태에 따른 데이터 분류
미리 정해진 구조가 있어!
내용 안에 구조에 대한 설명이 같이 있어!
정해진 구조가 없어!
정형 데이터
엑셀의 스프레드시트, 관계 데이터베이스의 테이블
반정형 데이터
HTML, XML, JSON 문서, 센서 데이터
비정형 데이터
소셜 데이터의 텍스트, 영상, 이미지, 음성

비정형 데이터

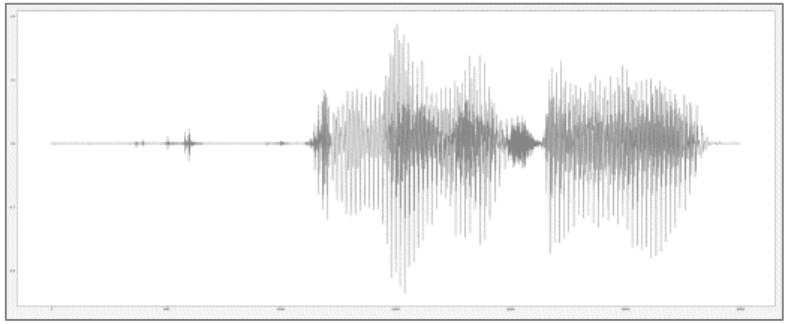
- 비정형데이터(Unstructured data)는 일괄적으로 정해진 구조가 없고 각각 데이터 형식을 가지는 데이터
- 빅데이터의 대부분을 차지하는 텍스트, 이미지, 오디오, 비디오 등이 해당 됨

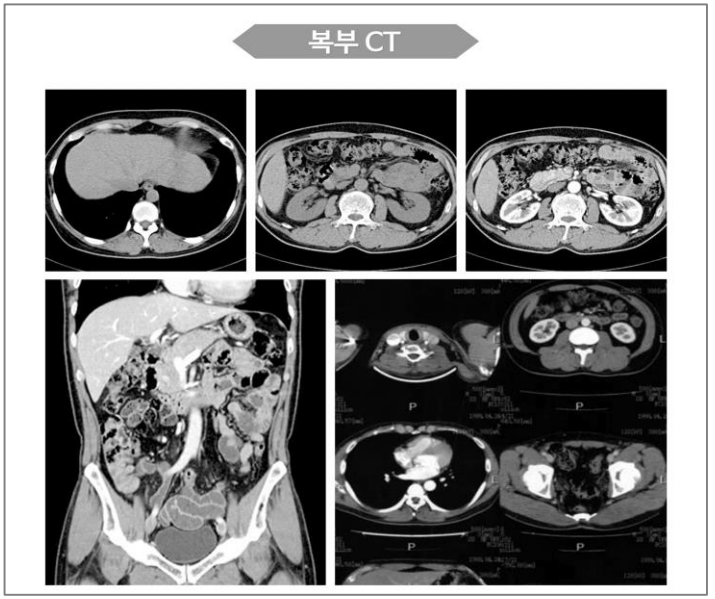
From today's featured article

"X-Cops" is the twelfth episode of the seventh season of the American science fiction television series *The X-Files*. Directed by Michael Watkins and written by Vince Gilligan, the installment originally aired on the Fox network in February 2000. In this episode, Fox Mulder (David Duchovny) and Dana Scully (Gillian Anderson), special agents for the Federal Bureau of Investigation, are interviewed for the Fox network reality television program *Cops* during an X-Files investigation. Mulder, hunting what he believes to be a werewolf, discovers that the monster terrorizing people craves the fear it provokes. While Mulder embraces the publicity of *Cops*, Scully is uncomfortable about appearing on national television. "X-Cops" is one of only two *X-Files* episodes that was shot in real time. The episode has been thematically analyzed for its use of postmodernism and its presentation as reality television. It has been named among the best episodes of *The X-Files* by several reviewers, for its humor and format. **(Full article...)**

Recently featured: Battle of Winterthur • 38th (Welsh) Infantry Division • *New Worlds* (magazine)

Archive • By email • More featured articles





비정형 데이터 전처리 주요 작업

- **자연어 처리를 위한 코퍼스 전처리**
  - 코퍼스(Corpus, 말뭉치): 텍스트 분석을 위한 데이터 셋

정제: Cleaning → 토큰화: Tokenization → 정규화: Normalization → 필터링: Filtering → 품사 태깅: POS Tagger

- **이미지 전처리 ← 컴퓨터 비전(Computer Vision) 활용**

이미지 데이터 인식 Visual Rcognition | 이미지 사이즈조절 Resize | 그레이스케일, 이진화 Grayscalling, Binarization | 이미지 노이즈감소 Denoising | 이미지 증강 Augmentation

- 오디오 데이터 전처리

오디오 데이터 리샘플링 → 데이터셋 필터링 → 오디오 데이터 변환

텍스트 전처리

- 토큰화
  - 주어진 코퍼스내 자연어 문장들을 토큰이라 불리는 최소 단위로 나누는 작업
  - 문장 토큰화 / 단어 토큰화 / 문자 토큰화 / 서브워드 토큰화
- 정제
  - 토큰화 작업에 방해가 되는 부분인 노이즈들을 제거하기 위해 지속적으로 이뤄지는 전처리 과정
  - 어떤 특성이 노이즈인지 판단 기준이나 제거 정도는 언어 특성에 따라 합의로 진행할 수 있음
- 정규화
  - 표현 방법이 다른 단어들을 통합시켜서 같은 단어로 만들어주는 과정
  - 어간추출, 표제어추출 / 대소문자 통합
- 필터링
  - 불용어(stop words) 제거 / 어간 추출(stemming) / 표제어 추출(lemmatization)
- 품사 태깅 (⊂형태소 분석)
  - 형태소 및 어근, 접두사, 접미사, 품사 등 다양한 언어적 속성의 구조 파악
  - 형태소의 뜻과 문맥을 고려하여 단어에 품사를 매핑함

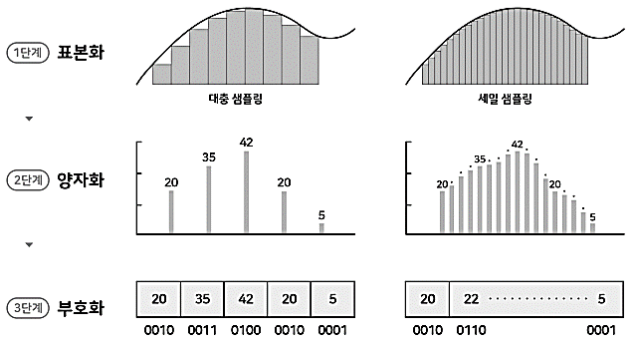
이미지 전처리

- 이미지 인식 (Visual Rcognition)
  - 이미지를 숫자로 변환하여 벡터로 표현하는 과정
- 이미지 사이즈 조절 (Resize)
  - 이미지 색을 숫자로 변환하는 작업인 벡터화를 수행하는 과정에서 이미지를 동일한 크기로 맞춤
- 이미지 그레이스케일 / 이진화 (Grayscailling / Binarization)
  - RGB의 다중 채널 이미지를 그레이스케일 및 이진화로 변환함
  - 데이터 양을 감소시켜 계산의 효율성을 높이는 효과
- 이미지 노이즈 감소 (Denoising)
  - 카메라 성능, 촬영 상황, 저장 및 전송에서 발생된 노이즈 등 분석에 불필요한 부분 제거
  - Blur / Filtering / Morphology 등
- 이미지 증강 (Augmentation)
  - 수집된 이미지 데이터 양을 늘리기 위해 기존 이미지 데이터를 변형하여 다양성을 높임
  - Flipping / Cropping / Rtation / Brightness & Saturation (+other color jittering) 등

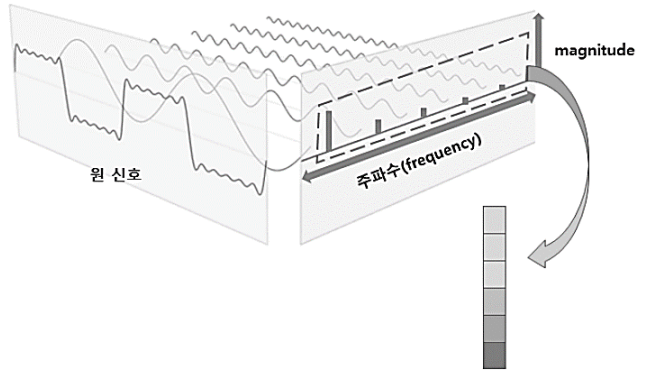
오디오 전처리

- 오디오는 본질적으로 음파(sound wave), 연속적인(시간의 흐름) 신호임
- 연속적인 음파를 이산적인 디지털 표현으로 변환되어야 함 → 샘플링(Sampling)
  - 아날로그 정보를 쪼개 대표값을 사용하여 디지털 정보로 표현, 이를 샘플링이라 함

<소리 정보의 디지털 표현>



- 오디오 데이터 리샘플링
  - 샘플링 속도를 기준에 의해 다시 맞추는 과정
- 오디오 데이터셋 필터링
  - 오디오 데이터를 특정 기준에 맞춰 필터링함
  - 예) 특정 길이에 맞춰 필터링
- 오디오 데이터 모델의 입력에 맞게 변환
  - 모델 학습에 맞는 형식으로 변환함
- 오디오 데이터 특징 추출
  - 디지털화된 음원 신호에서 특징을 추출
  - Waveform 시각화 / 푸리에 변환 / 스펙토그램(Spectogram) / STFT(Short Term Fourier Transformation) / MFCC(Mel-frequency cepstral coefficients)



<푸리에 변환으로 생성된 1차 벡터>

정형데이터 행과 열

- 변수(Variable,열,컬럼,피처,특징): 키, 몸무게, 성별
- 관측치(Observation,행,로우,레코드): 값을 측정한 단위, 각각의 사람
- 값(Value): 152 cm, 80 kg, 여성

1 변수(열,컬럼, 특징,피처)

Name	Sex	Age	Grade	Absence	Blood type	Height	weight
김길동	남자	23	3	유	O	165.3	68.2
이미린	여자	22	2	무	AB	170.1	57
홍길동	남자	24	4	무	B	175	80.1
김철수	남자	23	3	무	AB	182	85.7
손세수	여자	20	1	유	A	168	49.5
박미희	여자	21	2	무	O	162	52
강수진	여자	22	1	무	O	155.2	45.3

5 관측값 변수값

3 관측치

2 행,로우 데이터양

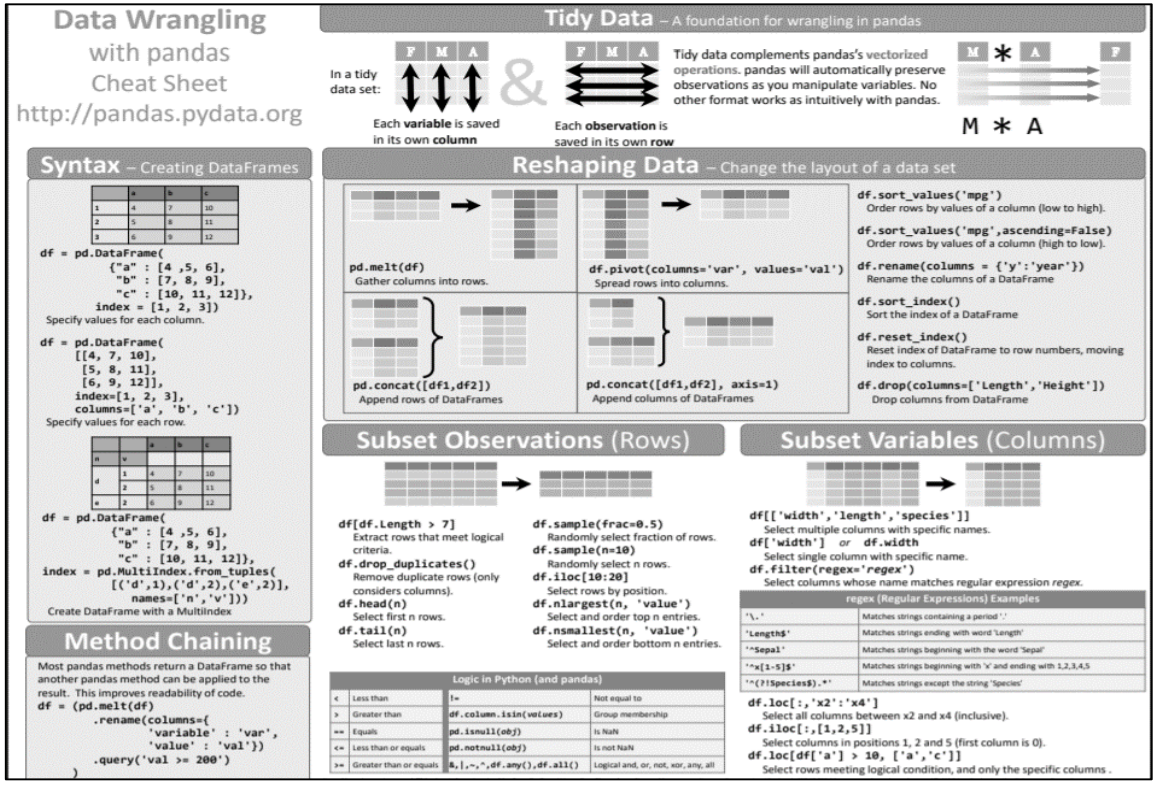
4 변수값

정형 데이터 전처리 주요 작업

- **데이터 정제(Cleansing)**
  - 데이터를 정제하는 기술
  - **결측값 보정, 이상치 검출, 표준화 및 정규화**, 노이즈 제거 등
- **데이터 변환(Transformation)**
  - 데이터 분석을 보다 쉽게 하기 위해 데이터를 변환해 일관성을 확보하고 데이터의 중복을 최소화 해 데이터 분석 시간을 절약하는 기술
- **데이터 필터링(Filtering)**
  - 데이터의 오류를 발견하고 삭제 및 보정을 통해 데이터의 품질 향상시키는 기술
- **데이터 통합(Integration)**
  - 데이터 분석을 수월하게 하기 위해 유사한 성질의 데이터를 연계하는 등 데이터를 통합하는 기술
  - 데이터 구조 변경(차원 변경), 데이터 벡터화 등
- **데이터 축소(Reduction)**
  - 데이터 분석 시간을 단축하기 위해 해당 분석에 사용되지 않는 데이터를 분석 대상에서 제외시키는 기술

정형데이터 분석을 위한 파이썬의 판다스 라이브러리

- 판다스에는 효과적인 데이터 분석을 위한 고수준의 자료구조와 데이터 분석 도구를 제공
  - 1차원 자료구조인 시리즈 (Series)
  - 2차원 자료구조인 데이터프레임 (DataFrame)
  - 3차원 자료구조인 패널(Panel)



데이터프레임

- 데이터프레임은 엑셀의 스프레드시트 형태로 행과 열 자료구조
- 각 열(column)에서는 서로 다른 종류의 값(숫자, 문자열, 불리언 등)을 가질 수 있음
- 시리즈가 복수 개 합쳐진 것
- 파이썬 데이터프레임은 판다스에서 제공하는 기본 구조
- 데이터 분석 시 가장 많이 보게 되는 자료구조

데이터프레임

	동물	나이
0	Dog	7
1	Cat	9
2	Tiger	2
3	Lion	3
4	Monkey	1

인덱스 → 0

나이 → 열

9 → 값

행과 열의 크기

- 데이터프레임에서 행이 많다는 의미는 어떤 의미일까?
- 데이터를 분석하다가 1억건 이상의 데이터를 분석하게 되면 어떻게 될까?
  → 행의 증가는 막대한 컴퓨팅 파워의 증가를 의미
- 행이 늘어나면 컴퓨터의 연산 핵심인 메모리와 CPU를 늘려야 함 → 클라우드 플랫폼
- 열의 감소는 분석 시 다양한 조합이 불가능

* 행의 증가

NO	Name	Sex	Age	Grade	Absence	Blood type	Height	weight
1	김길동	남자	23	3	유	O	165.3	68.2
2	이미린	여자	22	2	무	AB	170.1	57
3	홍길동	남자	24	4	무	B	175	80.1
:								
100,000,000	손세수	여자	20	1	유	A	168	49.5
100,000,001	박미희	여자	21	2	무	O	162	52
100,000,002	강수진	여자	22	1	무	O	155.2	45.3

* 열의 감소

NO	Name	Sex	Blood type	Height
1	김길동	남자	O	165.3
2	이미린	여자	AB	170.1
3	홍길동	남자	B	175
:				
100,000,000	손세수	여자	A	168
100,000,001	박미희	여자	O	162
100,000,002	강수진	여자	O	155.2

정리된 깔끔한 데이터

- 깔끔한 데이터(Tidy data)
  - 제프리크(Jeff Leek)의 데이터분석스타일기초 저서에서 제시한 데이터 분석이 용이한 데이터
  - 깔끔한 데이터는 데이터를 조작하고, 모형화 하고, 시각화가 용이
- 깔끔한 데이터는 특정한 구조를 갖추고 있는데 변수 는 열(column)이고, 관측점은 행(row)이며, 관측 단위에 대한 형태는 테이블(table)로 구성
- 깔끔한 데이터(tidy data) 는 결국 데이터분석을 쉽게 할 수 있는 데이터라 할 수 있음

* Tidy data

NO	변수	값
김길동	A	1
이미린	A	4
홍길동	A	5
김철수	A	3
손세수	A	2
김길동	B	7
이미린	B	3
홍길동	B	6
김철수	B	5
손세수	B	3

* Messy data

NO	A	B
김길동	1	7
이미린	4	3
홍길동	5	6
김철수	3	5
손세수	2	3

데이터 프레임 확인

- 데이터 요약정보의 확인은 데이터분석의 시작
- seabon 패키지에 내장된 'Tips' 데이터
  - 244개 행 인덱스(0~243)와 7개의 열에 관한 정보
- 데이터프레임의 기본 정보 출력 : **df.info( )**
- 데이터프레임의 자료형 확인 : **df.dtypes**

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
...	...	...	...	...	...	...	...
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

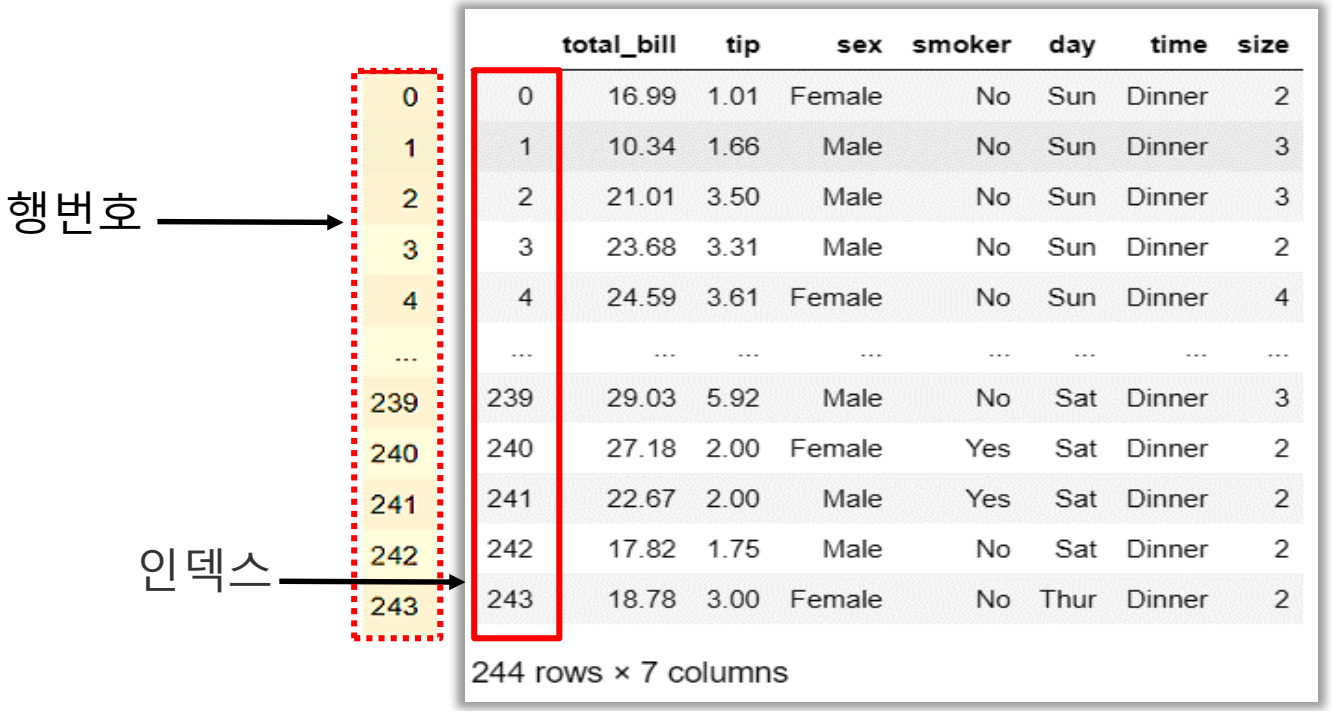
244 rows × 7 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
total_bill    244 non-null float64
tip           244 non-null float64
sex           244 non-null category
smoker        244 non-null category
day           244 non-null category
time          244 non-null category
size          244 non-null int64
dtypes: category(4), float64(2), int64(1)
memory usage: 7.3 KB
```

행 데이터

- 판다스 에서 행 단위로 데이터 보기는 두 가지

속성	설명	탐색 대상
loc	인덱스 기준으로 행 데이터 읽기	인덱스 이름 예) [‘A’: ‘D’] ->’A’, ’B’, ’C’, ‘D’
iloc	행 번호를 기준으로 행 데이터 읽기	정수형 위치 인덱스 예 [1:3] -> 1, 2



	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
...	...	...	...	...	...	...	...
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

244 rows × 7 columns

행 데이터 조작

기능	함수
1) 인덱스 읽기	**loc**로 행 데이터 추출하기 **df.loc[인덱스이름]** **df.loc[인덱스이름1, 인덱스이름2, 인덱스이름n]**
2) 행번호 읽기	**iloc** 속성으로 행 데이터 읽어오기 **df.iloc[행번호]** **iloc**를 통해 마지막 행 데이터 가져오기 **df.iloc[-1]**
3) 특정 행 범위 영역을 선택	**df[시작행:마지막행]**
4) 조건을 이용하여 선택하기	기본 조건식 **: and(&), or(|), not(~),** 비교**(==)**
5) 특정 조건 선택	**df.isin(values)**

열 변수

기능	함수
1) 열 변수 추출	**df.컬럼명 또는 df['컬럼명']**
2) 여러 개 열 변수 한 번에 추출	**df ['컬럼명1', '컬럼명2', '컬럼명n']**
3) 파생변수 만들기	**신규 df.컬럼명 = df.컬럼명 + df.컬럼명**
4) 열 변수 자료형 확인 및 변환	자료형 확인 **: df.dtype , type(df.컬럼명 또는 df['컬럼명'])** 자료형 변환 **: df.astype**

카테고리형(category)과 문자열(object)의 차이

- object 자료형과 category 자료형 모두 문자 표기를 의미 하는 자료형
- category
  - 문자열의 특수한 형태로서, 값의 범위가 제한적인 데이터는 카테고리형으로 지정
  - 일반 문자열형식보다 분석 시 메모리 용량과 속도 면에서 매우 효율적임

*** object**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 8 columns):
total_bill    244 non-null float64
tip           244 non-null float64
sex           244 non-null category
smoker        244 non-null category
day           244 non-null category
time          244 non-null category
size          244 non-null int64
smoker_str    244 non-null object
dtypes: category(4), float64(2), int64(1), object(1)
memory usage: 9.2+ KB
```

*** category**

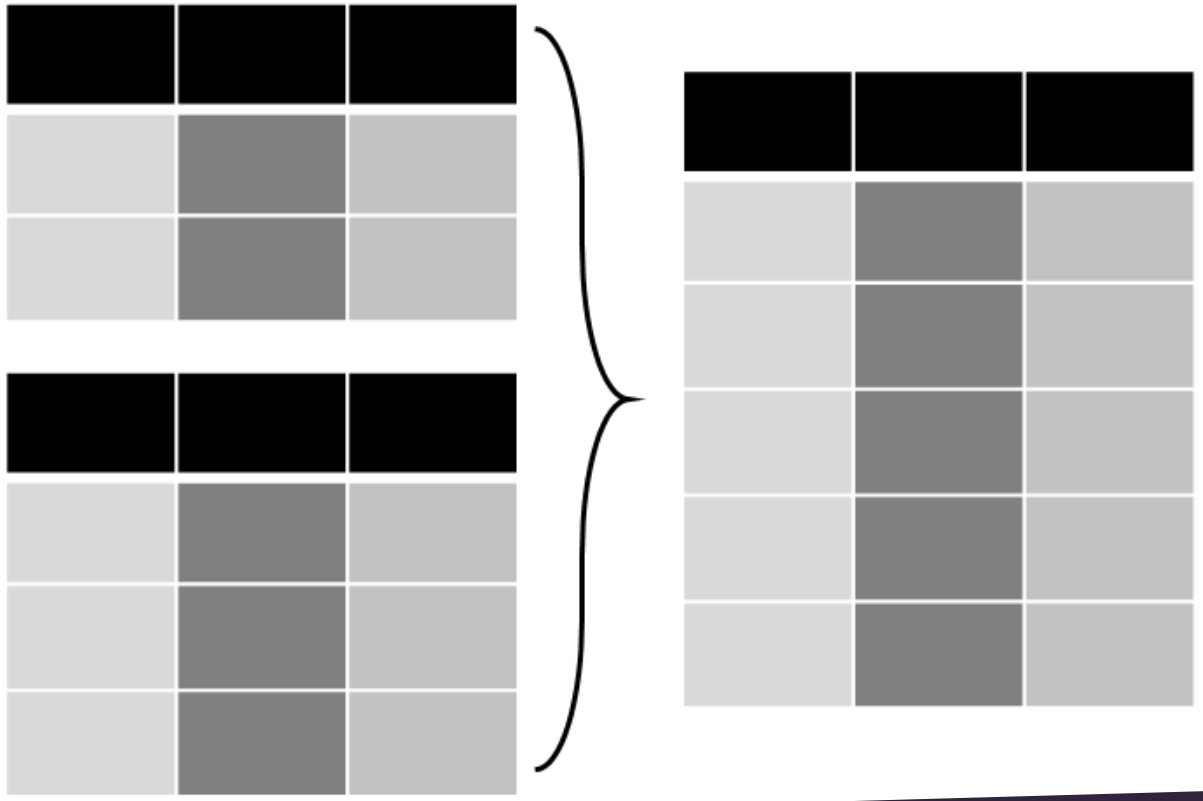
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 8 columns):
total_bill    244 non-null float64
tip           244 non-null float64
sex           244 non-null category
smoker        244 non-null category
day           244 non-null category
time          244 non-null category
size          244 non-null int64
smoker_str    244 non-null category
dtypes: category(5), float64(2), int64(1)
memory usage: 7.6 KB
```

데이터 조작

기능	함수
1) 개수 확인	**count()**
2) 기타보기 인덱스보기 변수보기 데이터보기	**df.index** **df.columns** **df.values**
3) 정렬 오름차순 내림차순	**dataframe.sort_values()** **df.sort_values(ascending=False)**
4) 행/열 합계	**df.sum()**

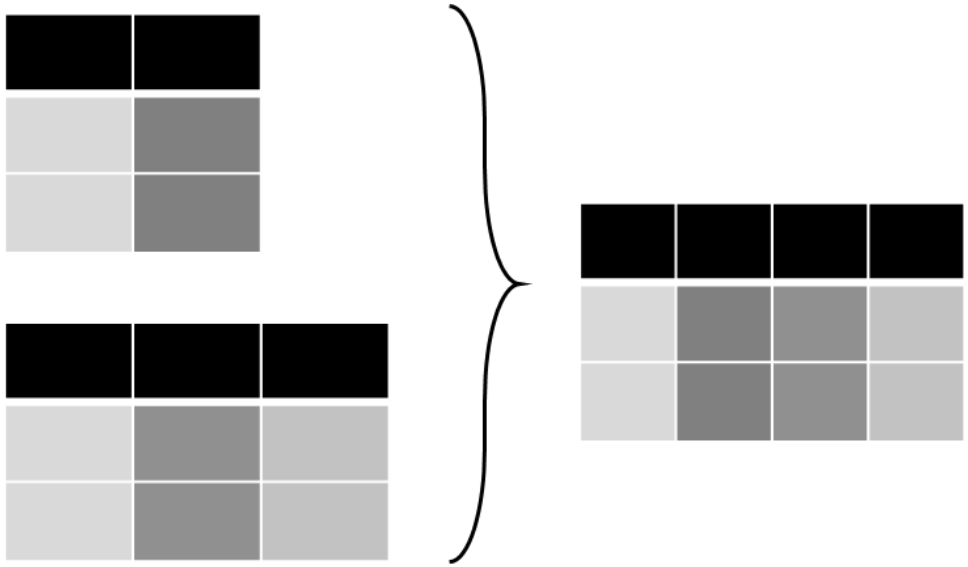
데이터 연결

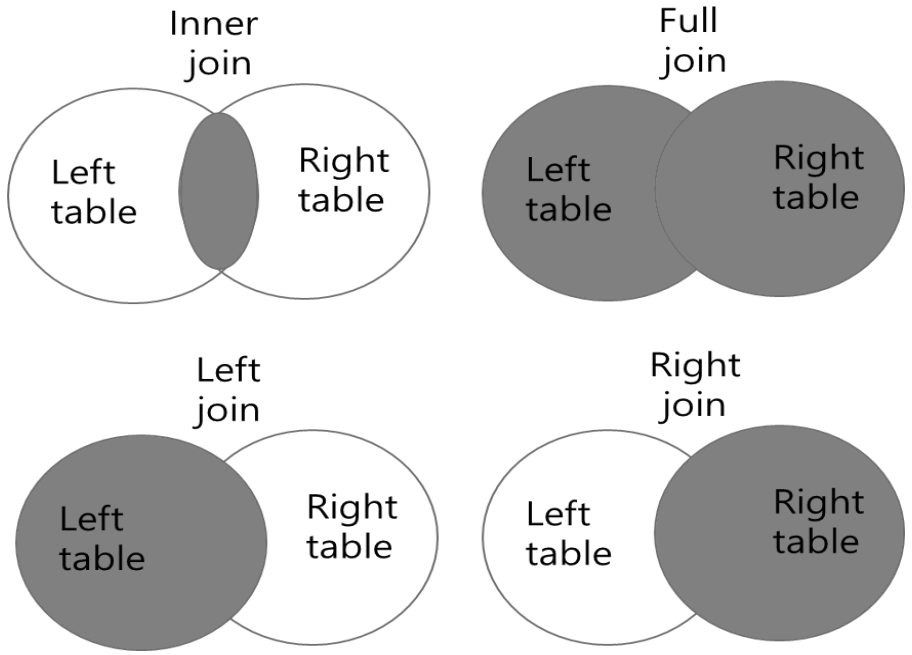
- 판다스의 두 개 이상 데이터프레임을 하나로 결합하는 방식
- 연결(Concatenate)은 데이터를 행과 열로 위/아래로 결합하는 방법
- 위/아래로 데이터 행을 연결하는 과정을 통해 간단히 두 시리즈나 데이터프레임을 연결할 수 있지만 이 경우 인덱스 값이 중복될 수 있음
- **df.concat()**
- **df.append()**



데이터 병합

- 병합(Merge) 명령은 두 데이터 프레임의 공통 열 혹은 인덱스를 기준으로 두 개의 테이블을 병합하는 과정
- 이 때 기준이 되는 열, 행의 데이터를 키(key)라고 함





결측 데이터

- 결측값(Missing data)
  - 데이터클린징 작업으로 데이터 누락값 즉, 결측값(Missing data)이 있는지를 자세히 검토
  - 결측값이 있는 상태로 분석하게 되면 변수간의 관계가 왜곡 될 수 있기 때문에 분석의 정확성이 떨어짐
  - 결측값이 발생하는 유형은 다양한데 결측값이 무작위로 발생하느냐, 아니면 결측값의 발생이 다른 변수와 관계가 있는지 여부에 따라 결측값을 처리하는 방법도 조금씩 달라짐
  - 판다스 에서는 결측값을 'NaN(Not a Number)' 으로 표기하며, 'None'도 결측값을 의미

- 결측값을 처리하는 방법에는 크게 두 가지
  - 제거(Deletion)와 대체(Imputation)방법
  - 제거는 결측값을 포함한 행, 열을 삭제
  - 대체는 특정한 방법 예를 들어 대표값인 평균 등으로 값을 변환 하는 방법

결측 데이터 확인

- isnull() : 결측 데이터이면 True값을 반환하고, 유효한 데이터가 존재하면 False를 반환
- notnull() : 유효한 데이터가 존재하면 True를 반환하고, 누락 데이터면 False를 반환

	Pregnancies	BloodPressure	BMI	Age	Outcome
0	6.0	NaN	33.6	50	1.0
1	NaN	66.0	NaN	31	NaN
2	8.0	64.0	23.3	21	1.0
3	NaN	NaN	28.1	40	NaN
4	0.0	40.0	NaN	33	1.0

df.isnull( )

	Pregnancies	BloodPressure	BMI	Age	Outcome
0	False	True	False	False	False
1	True	False	True	False	True
2	False	False	False	False	False
3	True	True	False	False	True
4	False	False	True	False	False

df.notnull()

	Pregnancies	BloodPressure	BMI	Age	Outcome
0	True	False	True	True	True
1	False	True	False	True	False
2	True	True	True	True	True
3	False	False	True	True	False
4	True	True	False	True	True

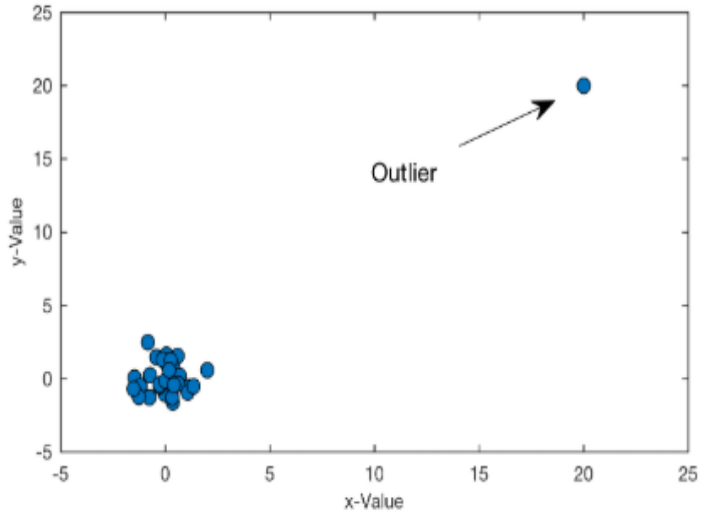
- 칼럼별 결측값 개수 구하기 : df.isnull().sum()
- 행(row) 단위로 결측값 개수 구하기 : df.isnull().sum(1)
- 행(row) 단위로 실제값 개수 구하기 : df.notnull().sum(1)

결측 데이터 처리

- Null의 의미 : 숫자 0과 null 과 같은 결측치는 완전히 다른 개념이니 유의해야 함
- 자료형(실수형, 정수형, 날짜/시간)의 확인
- Null 연산 시 유의할 사항
  - 결측치 제거 전 데이터 백업
- 결측치 제거
  - 행 삭제 데이터프레임.dropna(axis=0)
  - 열 삭제 데이터프레임.dropna(axis=1)
  - 데이터프레임.dropna()
- 결측치 데이터 대체
  - 결측값을 특정 값으로 채우기 : df.fillna(0)
  - 결측값을 특정 문자열로 채우기 : df.fillna(‘ ‘)
  - 결측값을 변수별 평균으로 대체: df.fillna(df.mean())

이상 데이터

- 정상에서 벗어난 데이터를 이상치(Outlier)
- 이상한(비정상적인) 데이터를 검출하는 것을 이상탐지(Anomaly Detection)

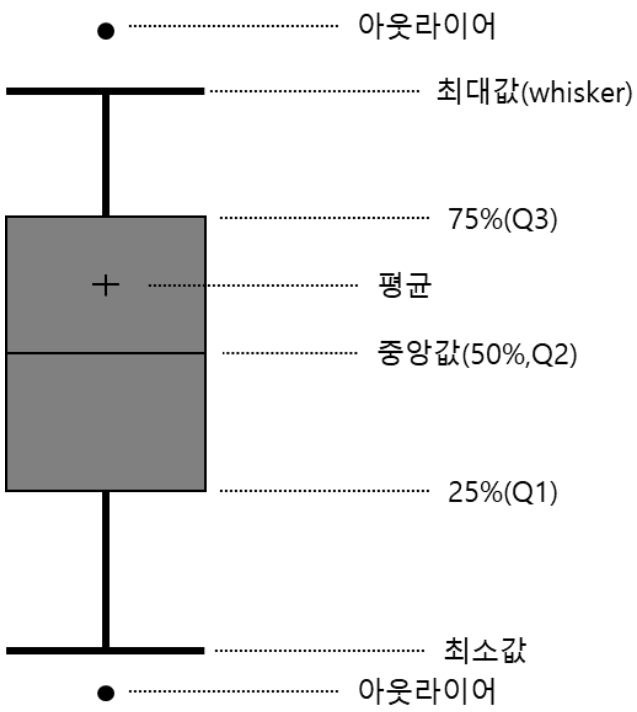


- 이상데이터 처리 방법
- 이상데이터 확인
- 이상데이터 결측 대체/제거
- 이상데이터 반영 확인

확인 → 대체/제거 → 반영확인

이상 데이터의 확인을 위한 시각화

- 중앙값(Median)은 중앙값 50%의 위치한 데이터
- 중앙값은 짝수일 경우 2개가 될 수도 있고, 평균이 중앙값이 될 수도 있음. 홀수일 경우, 중앙값은 1개
- 박스(Box)는 25%(Q1) ~75%(Q3) 까지 값들을 박스로 둘러 쌓게됨.
- 수염 (whiskers)은 박스의 각 모서리(Q1, Q3)로 부터 IQR의 1.5배 내에 있는 가장 멀리 떨어진 데이터 점까지 이어져 있는 것이 수염
- 이상치(Outlier)는 수염(whiskers)보다 바깥쪽에 데이터가 존재하는 데이터로 이것을 이상치로 분류 하면 됨



이상 데이터 처리 방법

- 단순 삭제
  - ✓ 이상값이 논리적 에러에 의해서 발생한 경우 해당 관측치를 삭제
  - ✓ 단순 오타나, 주관식 설문 등의 이상 응답, 데이터 처리 과정에서의 오류 등의 경우에 사용
- 다른 값으로 대체
  - ✓ 데이터의 개수가 작은 경우, 삭제의 방법으로 이상치를 제거하면 데이터 절대량이 작아지는 문제가 발생, 이 경우 관측치를 삭제하는 대신 다른 값(평균 등)으로 대체하거나, 결측값과 유사하게 다른 변수들을 사용해서 예측 모델을 만들고, 이상값을 예측한 후 해당 값으로 대체하는 방법도 사용
- 변수화
  - ✓ 단순 삭제나 대체의 방법을 통해 수립된 모델은 설명/예측하고자 하는 현상을 잘 설명하지 못할 수도 있으므로 자연발생적인 이상값의 경우, 바로 삭제하지 말고 좀 더 이상값에 대해 파악이 중요한데 이럴 경우 다른 변수로 변환을 통해 대체 함
- 리샘플링
  - ✓ 해당 이상값을 분리해서 모델을 만드는 방법
- 케이스분리 분석
  - ✓ 자연 발생한 이상값에 별다른 특이점이 발견되지 않는다면, 단순 제외 보다는 분석 케이스를 분리하여 분석 할 수도 있음

중복 데이터

- 데이터를 수집하는 과정 중 또는 데이터를 병합하는 단계에서 오류 등으로 인해 데이터가 중복이 되는 경우가 생길 수 있음
  - 특히, 유일한 키(key) 값을 관리해야 하는 경우 중복(Duplicates)데이터가 발생하면 분석에 영향을 끼칠 수도 있음
- 데이터 분석 전에 중복 데이터를 확인하고 처리하는 데이터 클린징 작업이 필요
- 데이터 개수가 많으면 육안으로 일일이 확인 하기 어려움
- 판다스에서 중복 데이터를 확인할 때 사용하는 방법으로 **duplicated()** 함수 사용
- 중복값을 처리 할 때 **drop_duplicates()**을 사용

중복 데이터 처리

- 중복 데이터 처리 방법
  - ① 중복 데이터 확인으로 중복이 있으면 무엇을 남길지 확인
  - ② 중복값 처리(유일한 1개 키만 남기고 나머지 중복 제거) **DataFrame.drop_duplicates()**
  - ③ 중복 데이터 결과
- 중복 데이터 확인 : **df.duplicated()**
- 중복 데이터 시작과 끝 확인
  - 중복이 있으면 처음이나 끝에 무엇을 남길지 확인
  - Keep= ' first ' 가 default 이며, 중복값이 있으면 첫번째 값을 duplicated 여부를 False로 반환 나머지 중복값에 대해서는 True를 반환
- 중복데이터 제거
  - 중복데이터에서 단일한 1개 키만 남기고 나머지 제거 하는 방법
  - drop_duplicates()는 중복값을 keep='first', 'last', False argument에 따라 유일한 1개의 key 값만 남기고 나머지는 중복 제거

데이터 재구조화

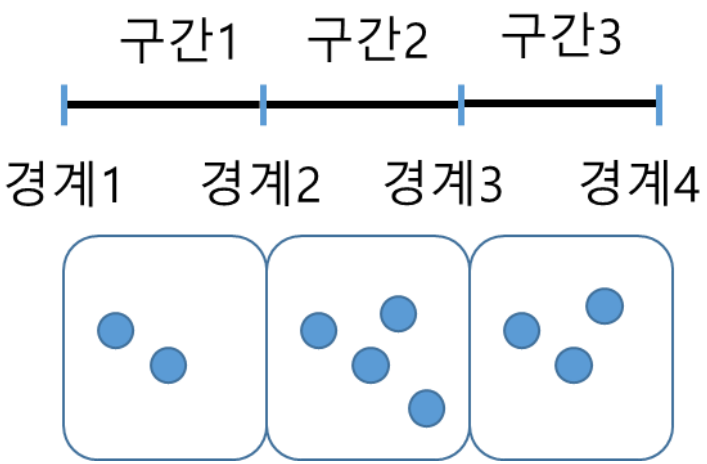
- 분석 과정에서 원본 데이터의 구조가 분석 기법에 맞지 않아 행과 열의 위치를 바꾼다거나, 특정 요인에 따라 집계 해서 구조를 바꿔야 하는 경우가 자주 발생
- 재구조화(Reshaping)라고 하며 판다스는 아래와 같이 재그룹화 기능을 제공

함수	설명
pd.cut(), pd.qcut()	데이터 구간화
pd.get_dummies	원-핫 인코딩
T	데이터 전치
melt()	열,행 전환
stack(), unstack()	행,열 인덱스 전환

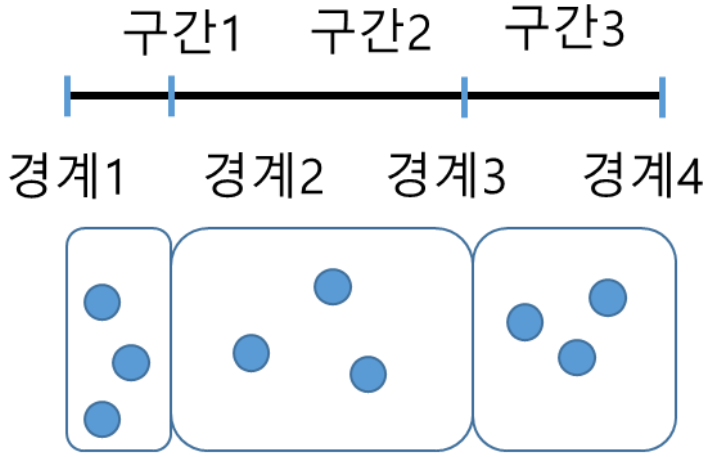
데이터 구간화

- 데이터 분석을 진행할 때 연속형 변수를 바로 사용하는 경우보다 일정 구간(bin)으로 작업하는 경우가 많이 있음
- 연속형 변수를 범주형 변수로 만드는 방법을 데이터구간화(Data Binning)이라 함
- 데이터구간화 에는 특별한 원칙이 있는 것이 아니기 때문에 데이터사이언티스트의 업무 이해도에 따라 창의적인 방법으로 할 수 있음

동일길이로 나누기
pd.cut( )



동일개수로 나누기
pd.qcut( )



원-핫 인코딩

- 딥러닝 알고리즘은 수치 데이터만 이해
- 기계가 이해할 수 있는 형태로 데이터를 변환해야 하는데 이때 범주형 데이터를 원-핫인코딩 형태로 변환
- 원-핫인코딩
  - 해당되는 하나의 데이터만 1로 변경해 주고 나머지는 0(dummy)으로 변경하는 과정
  - 아래 예각각의 동물인 사자, 곰, 여우로 컬럼을 만들고 해당 되는 동물에만 1로 표기 하고 나머지 동물은 0으로 변경

동물
사자
곰
여우
사자

동물	동물_사자	동물_곰	동물_여우
사자	1	0	0
곰	0	1	0
여우	0	0	1
사자	1	0	0

데이터 전치

- 데이터프레임 행과 열의 기준(축)을 바꾸는 방법을 데이터 전치(Transpose)
- 수학 선형대수에서 '전치행렬(transpose matrix)'
  - 행렬의 내적(inner product) 구할 때 a^T*a 처럼 전치 행렬과 원래 행렬을 곱할 때 전치 행렬(a^T)를 사용

	student_no	class	science	english	math	sex
0	1	A	50	98	50	m
1	2	A	60	97	60	w
2	3	A	78	86	45	w
3	4	A	58	98	30	m
4	5	B	65	80	90	w
5	6	B	98	89	50	m
6	7	B	45	90	80	m
7	8	B	25	78	90	w
8	9	C	15	98	20	w
9	10	C	45	93	50	w

df.T

	0	1	2	3	4	5	6	7	8	9
student_no	1	2	3	4	5	6	7	8	9	10
class	A	A	A	A	B	B	B	B	C	C
science	50	60	78	58	65	98	45	25	15	45
english	98	97	86	98	80	89	90	78	98	93
math	50	60	45	30	90	50	80	90	20	50
sex	m	w	w	m	w	m	m	w	w	w

멜트

- 멜트라는 의미 : 녹으면 흘러내린다.
- 멜트(melt)는 열을 행으로 변경하는 재구조화 과정
- 데이터프레임에서는 열이 행으로 흘러 열이 짧아지고 행이 길어 진다고 이해

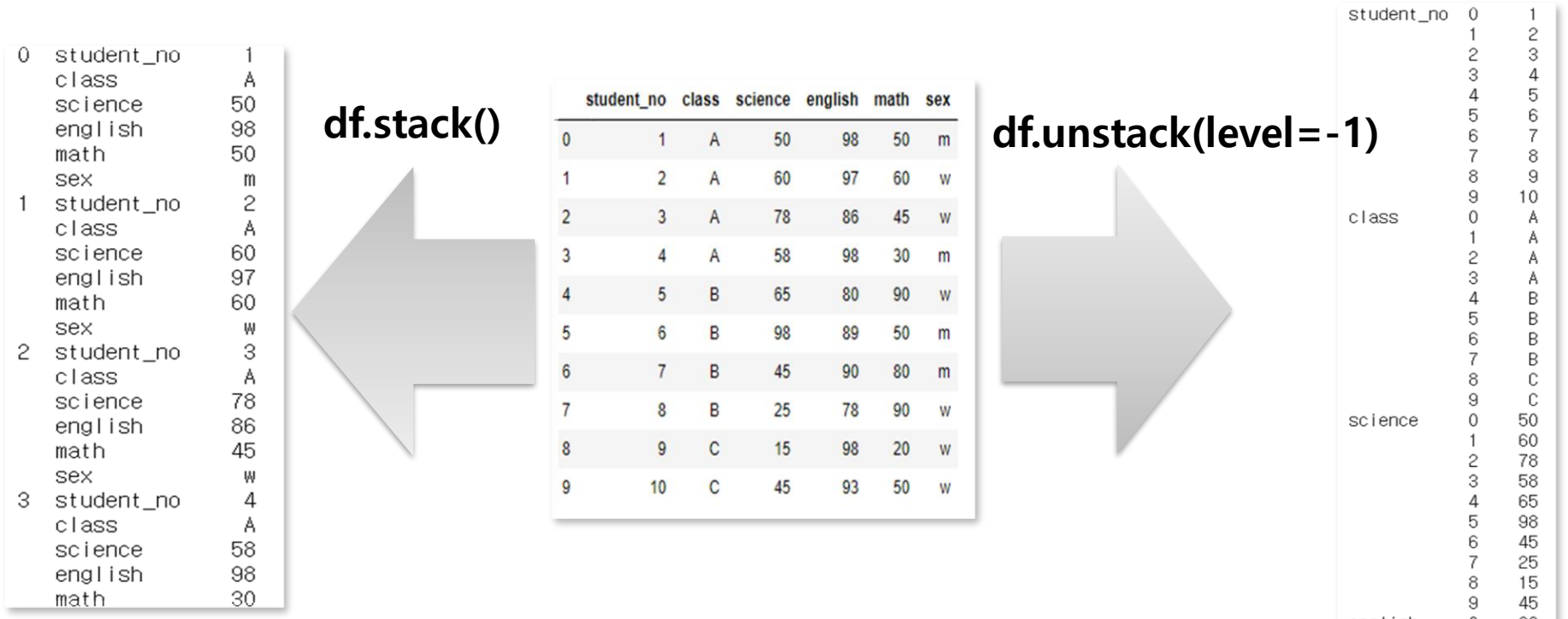
pd.melt(df,id_vars=['student_no', 'class'])

	student_no	class	science	english	math	sex
0	1	A	50	98	50	m
1	2	A	60	97	60	w
2	3	A	78	86	45	w
3	4	A	58	98	30	m
4	5	B	65	80	90	w
5	6	B	98	89	50	m
6	7	B	45	90	80	m
7	8	B	25	78	90	w
8	9	C	15	98	20	w
9	10	C	45	93	50	w

	student_no	class	variable	value
0	1	A	science	50
1	2	A	science	60
2	3	A	science	78
3	4	A	science	58
4	5	B	science	65
5	6	B	science	98
6	7	B	science	45
7	8	B	science	25
8	9	C	science	15
9	10	C	science	45
10	1	A	english	98
11	2	A	english	97
12	3	A	english	86
13	4	A	english	98
14	5	B	english	80
15	6	B	english	89
16	7	B	english	90
17	8	B	english	78
18	9	C	english	98
19	10	C	english	93
20	1	A	math	50
21	2	A	math	60
22	3	A	math	45
23	4	A	math	30
24	5	B	math	90
25	6	B	math	50
26	7	B	math	80
27	8	B	math	90
28	9	C	math	20
29	10	C	math	50
30	1	A	sex	m
31	2	A	sex	w
32	3	A	sex	w
33	4	A	sex	m
34	5	B	sex	w
35	6	B	sex	m
36	7	B	sex	m
37	8	B	sex	w
38	9	C	sex	w
39	10	C	sex	w

스택/언스택

- 행 인덱스와 열 인덱스 교환 시 사용하는 기능
- 스택(Stack) 명령을 사용하면 열을 행으로 변하는데 열 인덱스가 반 시계 방향으로 90도 회전한 것과 같은 모양이 됨
- 언스택(Unstack)은 스택으로 쌓은 것을 옆으로 늘여 놓은 것이라 볼 수 있음
- 둘 다 인덱스를 지정할 때는 문자열 이름과 순서를 표시하는 숫자 인덱스를 모두 사용함



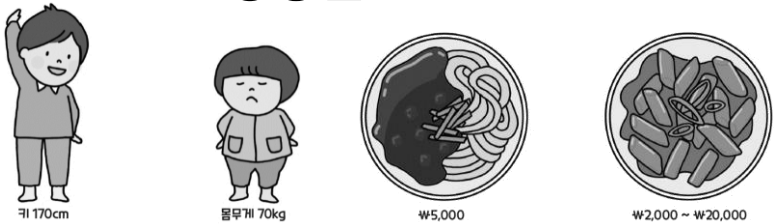
0	student_no	1
	class	A
	science	50
	english	98
	math	50
	sex	m
1	student_no	2
	class	A
	science	60
	english	97
	math	60
	sex	w
2	student_no	3
	class	A
	science	78
	english	86
	math	45
	sex	w
3	student_no	4
	class	A
	science	58
	english	98
	math	30

	student_no	class	science	english	math	sex
0	1	A	50	98	50	m
1	2	A	60	97	60	w
2	3	A	78	86	45	w
3	4	A	58	98	30	m
4	5	B	65	80	90	w
5	6	B	98	89	50	m
6	7	B	45	90	80	m
7	8	B	25	78	90	w
8	9	C	15	98	20	w
9	10	C	45	93	50	w

student_no	0	1
	1	2
	2	3
	3	4
	4	5
	5	6
	6	7
	7	8
	8	9
	9	10
class	0	A
	1	A
	2	A
	3	A
	4	B
	5	B
	6	B
	7	B
	8	C
	9	C
science	0	50
	1	60
	2	78
	3	58
	4	65
	5	98
	6	45
	7	25
	8	15
	9	45
english	0	98
	1	97

표준화와 정규화

- 데이터에서 특성(Feature) 또는 특징이란 분석 대상에 영향을 주는 속성을 말함
- 데이터를 분석할 때는 특성 중 어느 것이 분석 대상에 더 크게 영향을 미치는지 비교하여 선택하게 됨
- 단위가 다르면 비교가 어려움.
  - 키가 170cm인 사람과 몸무게가 70kg인 사람 중 누가 더 큰지 말할 수 없는 것과 마찬가지



- 단위가 같아도 값의 범위가 크게 차이 난다면 비교하기 어려움.
  - A 도시에서 모든 식당의 자장면 값이 5,000원이고 떡볶이 값은 2,000원부터 2만 원까지 분포하여 평균이 5,000원. 이때 A 도시에서 자장면과 떡볶이 중 어느 쪽이 더 비싼지 말할 수 없음.
- 데이터를 분석하려면 특성들의 상대적인 차이를 줄여 특성들이 모두 비슷한 정도로 대상에 영향력을 행사하도록 값을 변환해야 함.
- 표준화와 정규화를 수행하면 특성의 단위에 관계없이 값을 바로 비교할 수 있음.
  - 특성 스케일링(Feature scaling) 또는 데이터 스케일링(Data scaling) 이라고도 함.

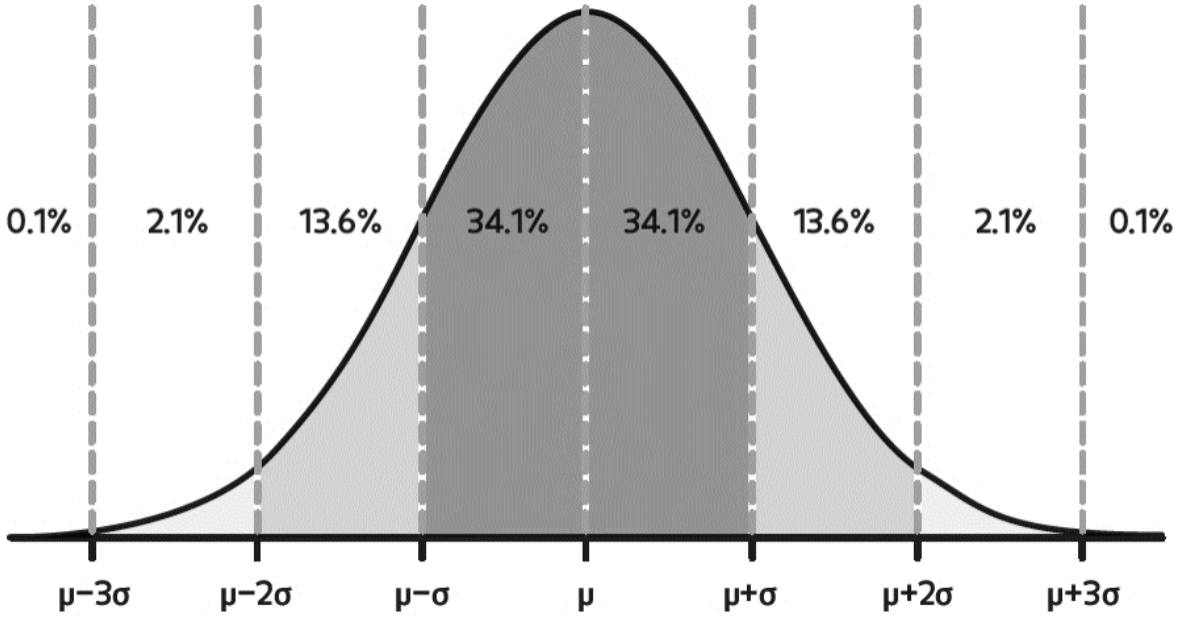
표준화

- 표준화(Standardization)는 데이터 분포를 평균이 0이고 표준편차가 1이 되게 변환하는 데이터 전처리 기법.
- 여러 개의 변수가 있을 때 서로 다른 변수들을 비교하기 편리하게 만듦.
- Z 점수(Z-score)는 어떤 값이 평균에서 얼마나 떨어져 있는지를 나타내는 수치.
  - 확률변수 X가 평균 μ(뮤)로부터 표준편차 σ(시그마)의 몇 배만큼 떨어져 있는지를 Z 점수로 정의.
  - Z 점수를 표준화 점수 또는 표준 점수라고도 함.
  - Z 점수는 표준편차의 배수로 계산됨. 표준편차는 데이터의 분포를 나타낸 값이기 때문에, Z 점수를 알면 서로 다른 분포로부터 나온 데이터를 비교할 수 있음.

$$Z\text{ 점수} = \frac{(X - \text{평균})}{\text{표준편차}}$$

표준화

- 평균을 중심으로 하고 표준편차 단위로 정규 곡선 영역을 나누었을 때 데이터 점 분포.
  - 데이터 점의 68.3%는 평균(μ) ± 1×표준편차(σ) 안에 있음.
  - 데이터 점의 95.4%가 평균(μ) ± 2×표준편차(σ) 안에 있음.
  - 데이터 점의 99.7%가 평균(μ) ± 3×표준편차(σ) 안에 있음.
- 정규분포곡선에서 Z 점수 절댓값이 3 이하인 데이터가 전체의 99.7%, 이 데이터를 이상치로 간주.



표준화

- Z 점수는 응시자 수가 많은 시험에서 개개인의 성적이 전체에서 어떤 위치에 있는지 보여줄 때 사용하기도 함.
  - 정규분포를 따르는 학년 영어점수 평균이 60점이고 표준편차가 10점일 때, 세일이의 점수가 70점이라면 Z 점수는 1.
  - 정규분포를 따르는 학년 국어점수 평균이 60점이고 표준편차가 15점일 때, 세일이의 점수가 75점이라면 Z 점수는 똑같이 1.
  - 세일이의 영어점수와 국어점수는 Z 점수가 같으므로 학년 분포에서 같은 위치에 있다고 할 수 있음.
  - 이 학년 학생 68.2%는 국어점수가 45점(Z 점수가 -1)에서 75점(Z 점수가 1) 사이일 것.

정규화

- 정규화(Normalization)는 데이터를 특정 범위 내의 값으로 조정하는 데이터 전처리 기법.
- 정규화를 하면 다양한 범위와 단위의 데이터를 서로 비교할 수 있음.

$$X' = \frac{(X - Xmin)}{Xmax - Xmin}$$

- 대표적인 정규화 방법으로 최소-최대 정규화(Min-Max scaling).
  - 데이터의 최솟값과 최댓값을 사용하여 데이터를 0과 1 사이의 값으로 변환.
- 정규화를 수행하면 여러 특성을 같은 범위로 맞출 수 있어서, 특성 간 비교가 쉬워져 빅데이터 분석에 유리하고 인공지능 모형의 성능을 높일 수 있음.

표준화와 정규화

- 표준화와 정규화 비교

비교	표준화	정규화
사용하는 값	평균과 표준편차를 사용	최댓값, 최솟값을 사용
목적	평균을 0으로, 표준편차를 1로 만들기	데이터의 범위를 서로 맞추기
전처리 후 데이터	범위가 제한되지 않음	[0, 1] 또는 [−1, 1] 사이로 스케일링
전처리 전 데이터	데이터가 정규분포일 때 유용함	데이터 분포를 모를 때 유용함

데이터 탐색

EDA(Exploratory Data Analysis)

탐색적 데이터 분석(EDA)

- 데이터 탐색(EDA)이란?
  - 데이터 분석의 초기 단계에서 데이터의 특성을 살펴보는 과정
  - 데이터 세트를 분석 및 조사하고 주요 특성을 요약하여 데이터의 패턴 발견, 이상 징후 발견, 가설 테스트, 가정을 확인하는 작업이라 할 수 있음
  - 통계를 기반으로 데이터의 특성을 발견할 수 있음

- **통계(統計, Statistics)**
  - 표준국어대사전: 한데 몰아서 어림잡아 계산, 현상을 통계에 의하여 관찰 · 연구하는 학문
  - 불확실성에 대한 논리를 부여하는 학문, 경험과학의 한 분야이자 대부분 학문의 기초이며, 다양한 정의가 존재하고 축약하면 자료를 연구하는 학문, 데이터를 분석하는 학문
  - 기술통계학(**Descriptive** Statistics) : 데이터를 수집, 정리, 요약하여 데이터 의미를 기술(설명)
  - 추론통계학(**Inferention** Statistics) : 표본 자료에서 얻은 정보를 이용하여 전체 집단(단위)에 대한 정보 및 불확실한 사실에 대해 예측하는 방법과 이론을 제시

빅데이터시대의 기술 통계와 추론 통계

- 기술 통계
  - 기술 통계량으로 데이터를 설명함. 데이터 그룹별 요약 집계, 기술 통계량 확인 등
- 추론 통계
  - 가설 검정: 모집단에서 샘플링한 표본으로 모집단의 특성을 추론하고 그 결과가 신뢰성이 있는지 검정하는 과정

- **빅데이터 시대**
  - 모집단과 표본집단을 구분하기 보다는 **데이터 전체를 표본으로 하고 수집하지 못한 현실 세계 전체 데이터나 미래에 대한 데이터를 모집단으로 볼 수 있음**

예) 통신 회사의 고객 데이터 전체는 데이터 과학자의 샘플링 단계가 없더라도 표본 집단으로 보고, 수집하지 못한 다른 회사의 고객데이터나 다음 달 고객 데이터는 모집단이라고 볼 수 있음

통계 기초 이론: 기술 통계

- 기술통계 기법
  - 수집한 데이터를 대표하는 값이 무엇인지 찾기 → 대표값
  - 수집한 데이터가 어떻게 퍼져 있는지를 설명하기 → 데이터 분포

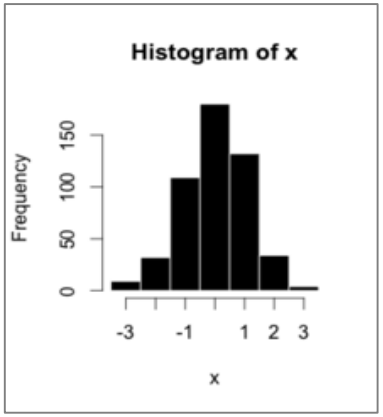
- 대푯값
  - 주어진 자료를 대표하는 특정 값
  - 대표값은 자료의 중심적인 경향이나 자료분포의 중심의 위치를 나타냄
  - 평균 / 중앙값 / 최빈값 등

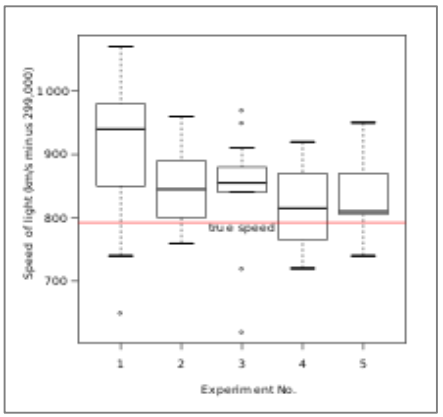
- 데이터 분포
  - 주어진 자료의 퍼짐 정도를 표현하는 값 또는 시각화 차트
  - 분산 / 표준편차 / 사분위값 / 왜도 / 첨도 / 도수분포표 / 히스토그램

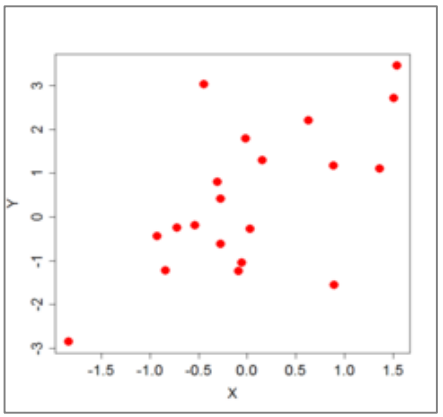
통계 기초 이론: 기술 통계

- 수치적인 탐색 : 기술통계
  - 평균(Mean), 최대값(Max), 최소값(Min), 중앙값(Median), 최빈값(Mode)
  - 표준편차(Standard Deviation), 분산(Variance)
  - 사분위수범위(Interquartile Range)
  - 첨도(Kurtosis), 왜도(Skewness)

- 시각화(그래프) 탐색
  - 히스토그램(Histogram)
  - 상자수염그림(Box plots)
  - 산점도(Scatter plots)

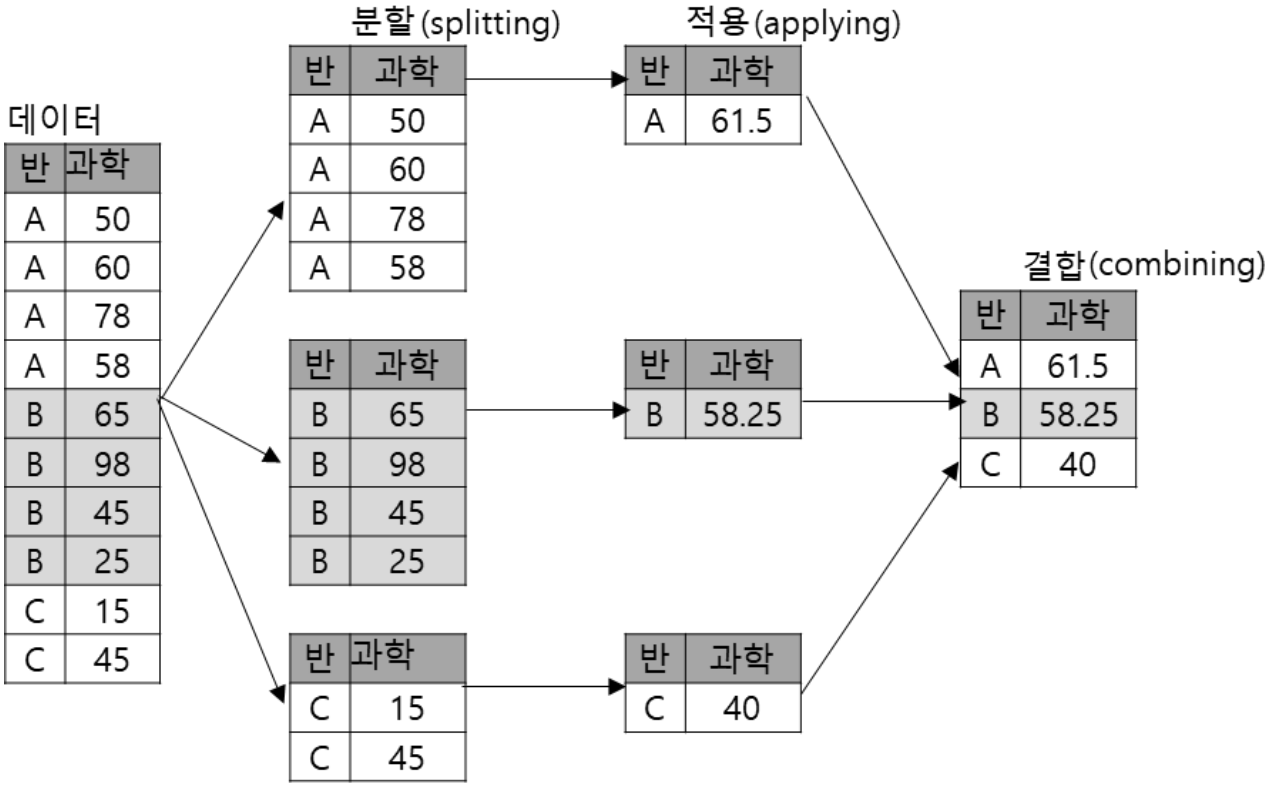






통계 기초 이론: 그룹별 요약

- 특정한 조건에 맞는 데이터가 하나 이상 데이터 그룹을 이루는 경우에는 그룹의 특성을 보여주는 그룹 분석(group analysis)을 자주 사용
- 그룹 분석 은 분석 대상 변수를 그룹별로 데이터를 집계하여 진행하는 분석으로 데이터분석에서 자주 사용하기 때문에 반드시 알아야할 필수적 방법



반	과학
A	50
A	60
A	78
A	58
B	65
B	98
B	45
B	25
C	15
C	45

반	과학
A	50
A	60
A	78
A	58

반	과학
B	65
B	98
B	45
B	25

반	과학
C	15
C	45

반	과학
A	61.5

반	과학
B	58.25

반	과학
C	40

반	과학
A	61.5
B	58.25
C	40

통계 기초 이론: 피봇 테이블

- 피봇(Pivot) 테이블이란 많은 양의 데이터에서 필요한 요약 정보로 재구성하는 기능
- 피봇 테이블을 사용하면 사용자가 원하는 대로 데이터를 정렬하고 필터링 가능
- 판다스는 피봇테이블을 만들기 위한 pivot() 함수를 제공하는데 데이터 열 중에서 두 개의 열을 각각 행 인덱스, 열 인덱스로 사용하여 데이터를 조회하여 펼쳐놓은 것을 의미
- 첫번째 인수는 행 인덱스로 사용할 열 이름, 두번째 인수는 열 인덱스로 사용할 열 이름, 마지막으로 데이터로 사용할 열 이름을 지정

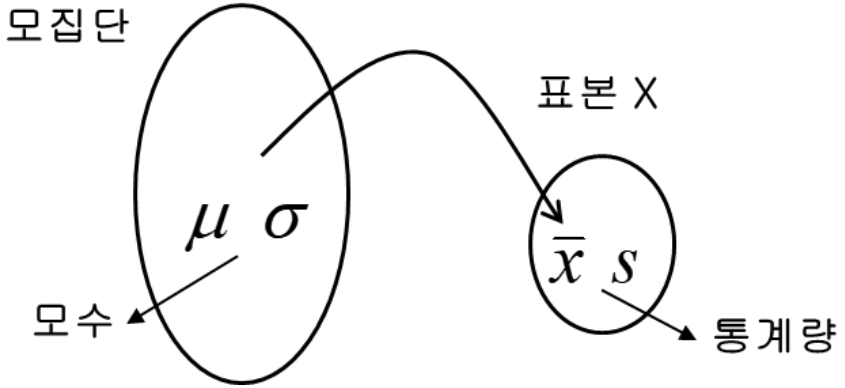
	student_no	class	science	english	math	sex
0	1	A	50	98	50	m
1	2	A	60	97	60	w
2	3	A	78	86	45	w
3	4	A	58	98	30	m
4	5	B	65	80	90	w
5	6	B	98	89	50	m
6	7	B	45	90	80	m
7	8	B	25	78	90	w
8	9	C	15	98	20	w
9	10	C	45	93	50	w

df.pivot_table(data, index, columns, values, aggfunc)

sex	m	w
class		
A	54.0	69.0
B	71.5	45.0
C	NaN	30.0

통계 기초 이론: 모집단과 표본

모집단	표본
관심의 대상이 되는 모든 개체의 관측값이나 측정값의 집합	모집단에서 실제로 추출한 관측값이나 측정값



구분	모수	통계량
대상	모집단의 특성	표본에서 계산한 특성
표시	그리스, 로마자로 표시	알파벳으로 표시
평균	μ	$\bar{x}$
표준편차	σ	s

통계 기초 이론: 가설 수립

- 가설(Hypothesis)은 모수에 대한 예상, 주장, 또는 단순한 추측
  - 예) '외계인은 존재한다, 그 사람은 유죄이다' 처럼 아직은 하나의 추측인 사실을 가설이라 함
- 통계적 가설 검정은 가설에 대해 증거를 수집하고 과학적으로 증명하는 과정
- 가설 검정의 첫 단계는 가설 수립부터 시작
  - 검정하고자 하는 모집단의 모수(조사하고자 하는 자료의 평균, 분산, 표준편차, 상관계수)에 대해 항상 귀무가설과 대립가설 두 가지로 수립

영(귀무)가설(null hypothesis)	대립가설(alternative hypothesis)
H0	H1
기각하기를 희망하여 형식화한 가설, 기존에 받아들이던 가설 모수에 관한 귀무가설은 항상 모수의 정확한 값을 지정하도록 진술 될 것인 반면 대립가설에서는 여러 개의 값의 가능성이 허용	표본을 통해 입증하고자 하는 새로운 가설 모수에 대한 관심의 영역 중에서 귀무가설로 지정되지 않은 모든 경우를 포괄적으로 지정
외계인은 존재하지 않는다. 그 사람은 무죄이다	외계인은 존재한다 그 사람은 유죄이다

통계 기초 이론: 가설 검정 단계

구분	사례1
가설 수립	외계인이 존재할까? 외계인이 존재한다는 확실한 증거 수집 전까지는 외계인은 없다라고 한다. H0 : 외계인 = 0. 외계인은 0 명이다. 외계인은 없다 H1 : 외계인 ≠ 0. 외계인은 0 명이 아니다. 외계인은 있다
가설 검정	외계인이 있다라는 증거가 많이 있다 ⇨ 외계인은 존재한다. 외계인이 있다라는 증거가 조금밖에 없다 ⇨ 외계인은 존재한다는 증거가 부족하다
증거 수집	외계인은 존재하는가? 외계인이 존재한다는 객관적인 증거가 97% 있다. 증거가 95%보다 많으므로 외계인은 존재한다

구분	사례2
가설 수립	그 사람은 무죄일까? 유죄일까? 법정에서는 유죄 판결을 받기 전까지 모든 사람들은 무죄 H0 : 기존에 받아들이던 가설 죄=0. 죄가 없다. 무죄 H1 : H0를 기각하기를 바라는 가설 죄≠0. 죄가 있다. 유죄
가설 검정	무죄가 아니라는 증거가 많이 있다(유죄) 무죄가 아니라는 증거가 조금밖에 없다(증거 불충분으로 무죄)
증거 수집	그 사람은 무죄인가? 유죄인가? 유죄라는 객관적인 증거가 80% 있다 증거가 95%보다 적으므로 증거불충분으로 무죄

통계 기초 이론: 오류의 이해

- 유죄라는 객관적인 증거와 외계인이 존재한다는 객관적인 증거는 얼마나 필요할까? 일반적으로 95% 정도가 필요
- 그럼 객관적인 증거가 95%보다 많으면 실제로 외계인은 존재할까(H1)?
- 아니면 객관적인 증거가 95%보다 적으면 실제로 무죄일까(H0)?
- 반드시 그런 것은 아니며 이것을 오류라 하며, 통계적 오류는 다음과 같이 구분

구분	영가설진실(H0)	대립가설진실(H1)
영가설선택**(H0)**	옳은 판단 신뢰수준(1-α)	2종 오류 유의수준(β)
대립가설선택**(H1)**	1종 오류 유의수준(α)	옳은 판단 신뢰수준(1- β)

통계 기초 이론: 유의수준

비교	결과
P < α = 0.05 **(오류 5% 이하, 95% 이상 진실)**	H1 선택
P >= α = 0.05 **(오류 5% 이상, 95% 이하 진실)**	H0 선택

- 유의수준(α)는 항상 0.05로 고정되어 있을까? 그렇지 않다.
- 유의수준은 일반적으로 0.05를 사용하는데, 연구자의 기준에 따라서 변할 수 있다.
- 그러나, 유의수준은 분석 전에 미리 결정을 하여야 함.
- 유의수준을 0.01로 하면 어떻게 될까? 혹은 유의수준을 0.1로 하면 어떻게 될까?

통계 기초 이론: 통계 결과 해석

- 결과 해석은 유의수준과 유의확률을 비교하여 결정
- 유의수준은 제1종 오류의 최대 허용 한계
  - 유의수준 α 값이 작아지면(오류가 작아진다) 영가설이 틀렸다는 결론을 내리기 어려움
  - 반대로 유의수준 α 값이 커지면 귀무가설이 틀렸다는 결론을 내리기 쉬워짐
- 유의확률(검정통계량)은 P 값 이나 P-value 라고 함
- 영가설(H0)이 맞을 경우, 대립가설 쪽의 값이 나올 확률이 얼마나 되는지를 나타내는 값으로 결론적으로 통계는 유의확률 P와 유의수준 α를 비교하여 영가설과 대립가설을 선택하는 과정
- 대부분통계는 새로운 가설을 선택하는게 목적
- 주로 판정 기준을 이렇게 표현

P < 0.05 이하 기준 이면 새로운 대립가설(H1) 을 선택

통계 기초 이론: 통계 절차 요약

① 통계분석방법 선정

② 분석하고자 하는 목적에 따른 귀무가설(영가설)과 대립가설 설정

③ 분석도구(Excel, SPSS, R, Python programming 등) 검정통계량 실행 및 확인

④ 유의수준(α) 결정 : 0.1, 0.05, 0.01

⑤ 유의확률(P) 확인

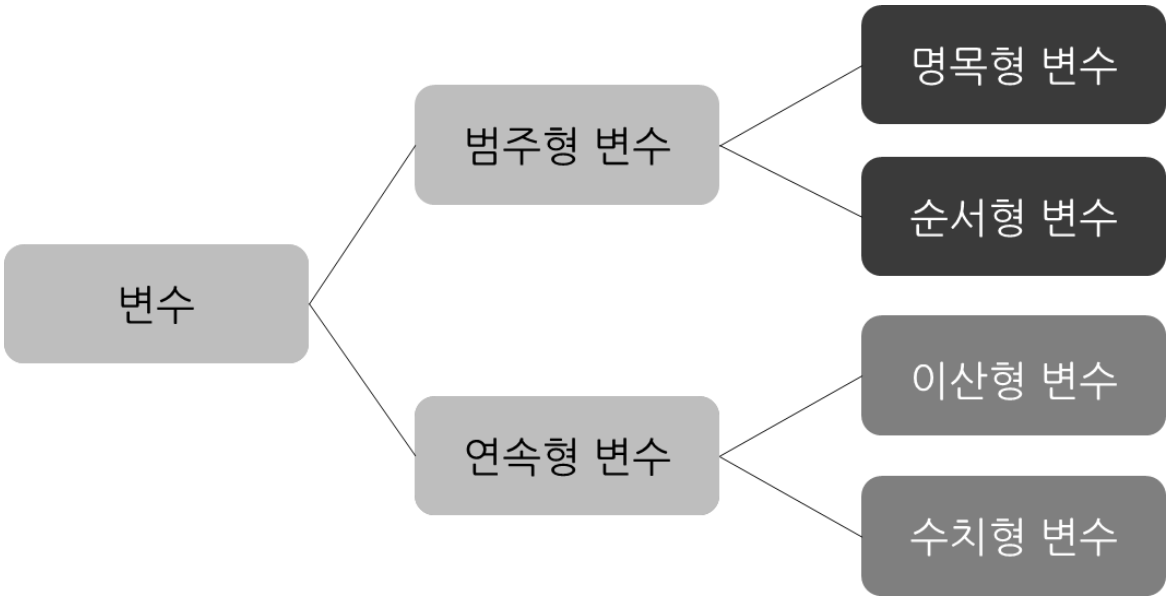
⑥ 유의확률과 유의수준 비교 (< α)

⑦ 귀무가설 과 대립가설 선택

⑧ 분석 결론

통계 기초 이론: 변수 유형

- 범주형 변수
  - 연속형 데이터와 범주형 데이터의 가장 큰 차이는 가감승제가 가능하지만 의미가 없다는 것
    - 예를 들어 성별을 1(남자), 2(여자)로 구분하여 이들의 평균을 구하면 1.5가 됨
  - 명목변수, 서열변수
- 연속형 변수
  - 사칙연산이 가능. 예로부터 자연현상에 대해 과학적 수치를 부여하는 과정을 “수치화” 라고 함
  - 등간, 비율변수

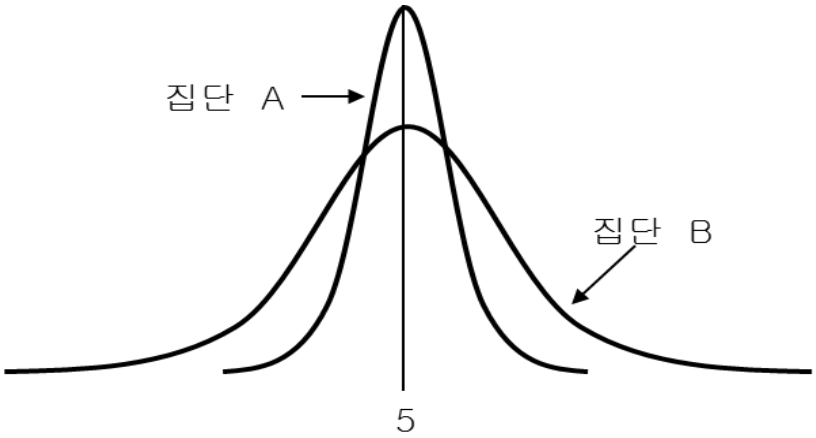


통계 기초 이론: 범주형 변수 분석

- 범주형 변수의 분석의 첫 단계는 해당 변수의 빈도(Frequency)
- 전체데이터 빈도를 계산하는 것부터 시작
- 빈도분석
  - 빈도는 대상 자료에서 반복하는 횟수를 기록하는 것으로 빈도분석을 통해 전반적 분포를 살펴보면서 전체 데이터의 구성을 파악 과정
  - 가장 빈도가 높은 데이터와 가장 빈도가 낮은 데이터를 기록하고 데이터 분석을 시작해야 함
- 교차분석
  - 교차분석은 범주형 변수간 빈도분석을 확장하여 빈도를 교차시킨 분할표(Contingency Table)를 만들어 분석하는 방법
  - 분할표는 교차분석표(Cross table) 라고도 함

통계 기초 이론: 연속형 변수 분석

- 수집된 수치 데이터의 정리, 표현, 요약등을 통해 데이터의 전반적인 특성을 이해하는 분석 진입 단계 → '기술통계' 라고도 함
- 데이터 중심 이해
  - 데이터의 중심은 여러 자료들의 비교하였을 경우 중앙에 위치하는 값으로 자료의 특성을 보여줌
  - 평균, 중위수
- 데이터 퍼짐 정도 이해
  - 분산: 데이터 퍼짐 정도
  - 표준편차: 데이터의 퍼짐 정도를 동일 한 기준을 적용하기 위하여 편차 제곱합의 평균
  - 범위: 데이터의 흩어진 범위
  - 사분위: 데이터의 분포가 좌우대칭이 아니거나 이상치가 있는 경우 평균은 극단적으로 치우진 대표성이 없는 값에 의해 영향을 받음
    - 이 경우 자료를 나열하여(시각화) 전체 데이터를 파악할 수 있음
    - 자료를 순서대로 나열했을 때 50%에 위치하는 수가 중위수(Q2) 이고 25%에 위치하는 수가 Q1

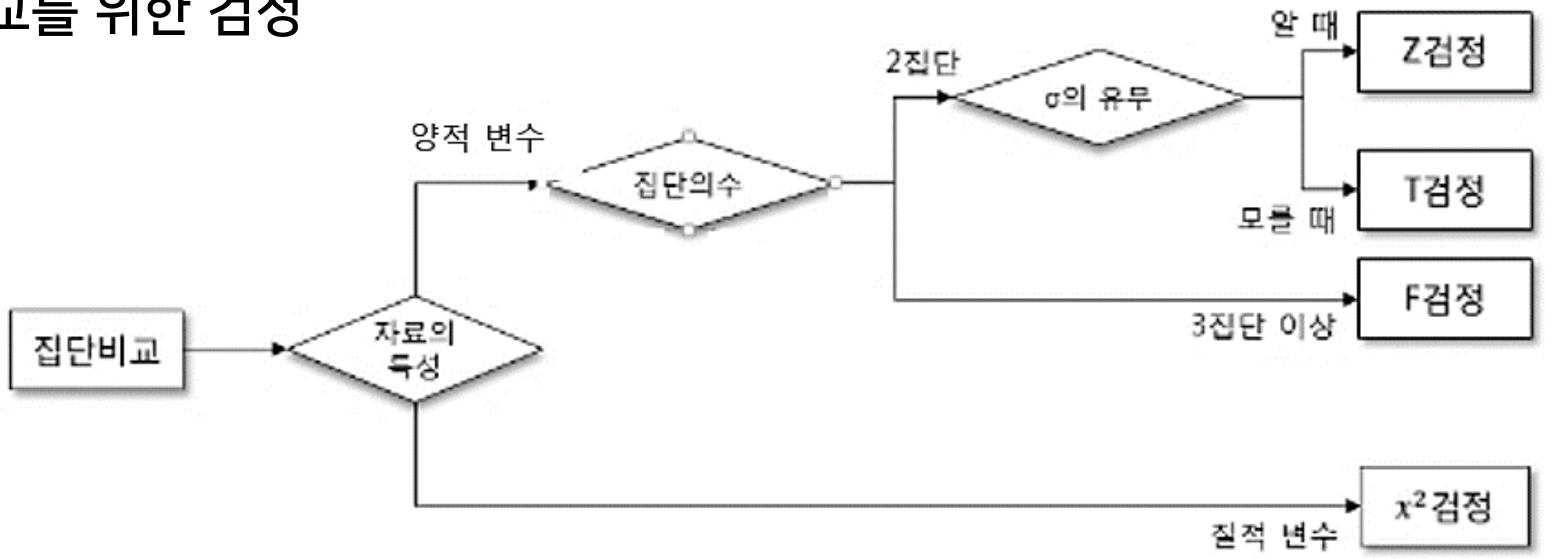


통계 기초 이론: 분석 모형 선택 기준

- 분석하고자 하는 내용(주제) 즉, “가설”을 명확히 규명
- 가설에 맞는 데이터의 변수 척도를 파악 → 변수 척도에 따라 적용할 모형이 정해짐
- 가설 검정의 구분
  - 차이 검정: 집단 간 평균 차이
  - 관계 검정: 함수적 관계 규명
- 종속 변수의 종류가 연속형과 이산형에 따라 분석 모형을 결정
  - 종속변수 Y 가 연속형일 때
    - 차이 : T-test, Anova(분산분석)
    - 관계 : Regression (회귀분석)
  - 종속변수 Y 가 이산형일 때
    - 차이 : Chi-square Independence Test (카이제곱 독립성검정)
    - 관계 : Logistic Regression (로지스틱 회귀분석)

통계 기초 이론: 차이 분석

● 집단 비교를 위한 검정



집단	대상
단일 표본 비교	하나의 집단(단일 표본 - one sample)중 관심 있는 연속형 변수의 모평균이 어떤 특정 값과 같은지 알아보고자 할 때
두 집단 간 평균비교	독립 표본 T : 서로 독립인 두 표본에 의한 모평균 비교 대응 표본 T : 대응하는(Paired) 쌍에 대한 차의 모평균 검정
3 이상 집단간 평균비교	일원배치분산분석 (One-way ANOVA)

통계 기초 이론: T-분석

- T-분석, T-검정은 두 집단의 평균을 비교하는 통계적 검정 방법
- 모집단을 대표하는 표본으로부터 추정된 분산이나 표준편차를 가지고 검정하는 방법으로 "두 모집단의 평균간의 차이는 없다"라는 귀무가설과 "두 모집단의 평균 간에 차이가 있다"라는 대립가설 중에 하나를 선택할 수 있도록 하는 통계적 검정방법
- 단순히 차이의 존재 여부를 떠나 두 집단의 비교가 통계적으로 의미가 있는가를 검정
- 즉, 두 모집단의 차이가 우연에 의해서 인지 아닌지를 검정하는 방법
- T분석의 기본 가설

✓ 가설

- 영 가설 : 집단간의 평균 차이는 없다.
- 대립가설 : 집단간의 평균 차이는 있다.

$$t = \frac{\overline{X} - \mu}{\text{표준오차(SE)}\ \frac{S}{\sqrt{n}}}$$

$\overline{x}$: 두 집단 차이의 평균

μ : 모집단의 평균

S : 두 집단 차이의 표준편차

- 30개 이하의 비교적 적은 수의 표본에 대해 활용 → 30개 이상이면 정규분포의 Z검정
- 모집단의 표준편차를 알 수 없을 때 사용

통계 기초 이론: 상관분석

- 두 연속형 변수 사이 상관관계가 존재하는지를 파악하고, 상관관계의 정도를 확인 하는 것이 상관분석(Correlation analysis)이라 함
- 상관분석에서는 관련성을 파악하는 지표로 상관계수(Correlation coefficient)라는 통계학적 관점에서 선형적 상관도를 확인하여 정도를 파악
- 상관분석은 간단한 분석이지만 머신러닝의 기반이 됨

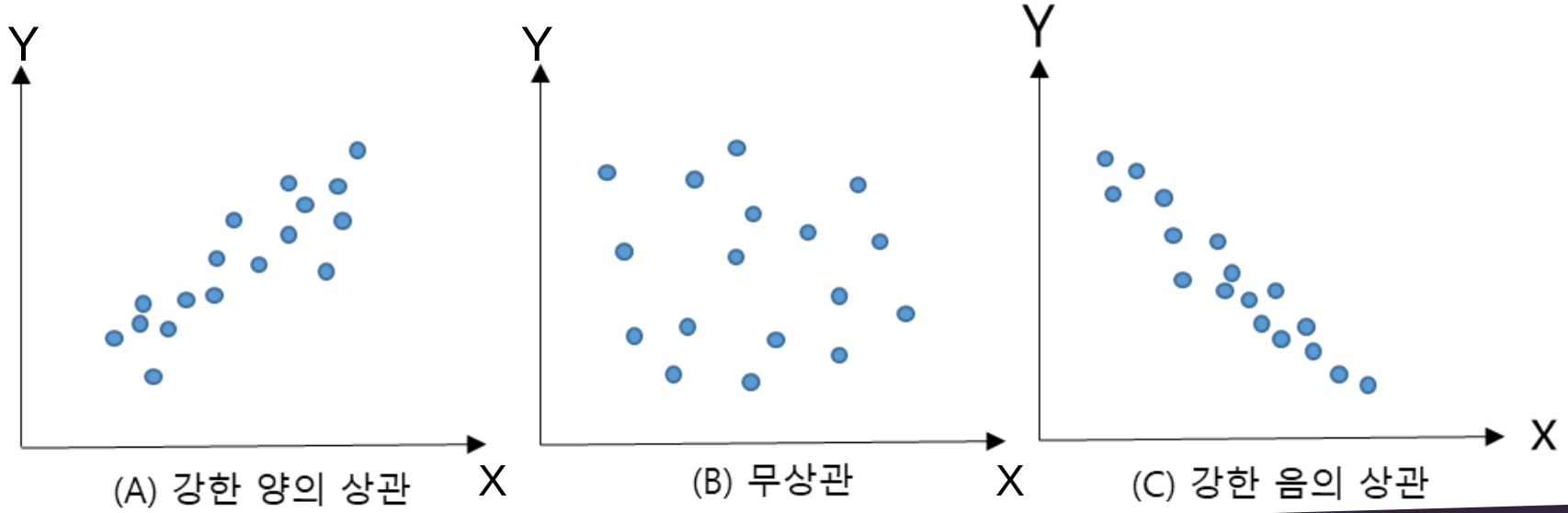
① 산점도(Scatter) 두 변수 상관 파악

② 상관계수 확인

③ 의사결정

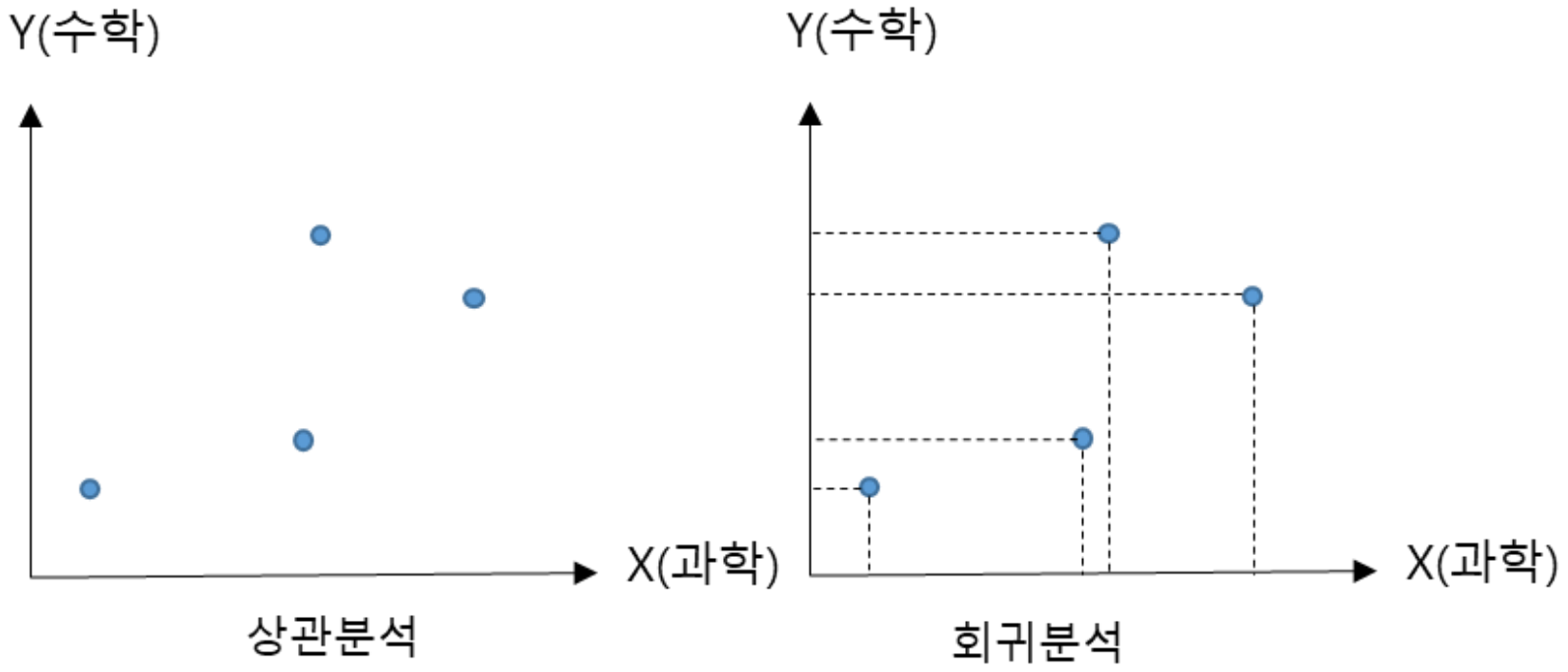
통계 기초 이론: 상관분석

상관	상관계수
양의 상관	+0.1 ~ +0.3 이면, 약한 양의 상관관계 +0.3 ~ +0.7 이면, 뚜렷한 양의 상관관계 +0.7 ~ +1.0 이면, 강한 양의 상관관계 – 그림A
무상관	-0.1 ~ +0.1 이면, 없다고 할 수 있는 상관관계 - 그림 B
음의 상관	-1.0 ~ -0.7 이면, 강한 음의 상관관계 – 그림C -0.7 ~ -0.3 이면, 뚜렷한 음의 상관관계 -0.3 ~ -0.1 이면, 약한 음의 상관관계



통계 기초 이론: 회귀분석

- 상관분석에서는 두 연속형 변수 X(과학)와 Y(수학)의 상관 정도만 알 수 있고 인과관계는 알 수 없었음
- 회귀분석에서는 두 연속형 변수 X와 Y를 독립변수와 종속변수라고 하는 인과관계로 설명
- '과학 점수가 좋으면 수학점수가 좋을까요?' 와 같이 간단 하지만 미래를 예측할 수 있는 머신러닝의 초기 모델이 됨

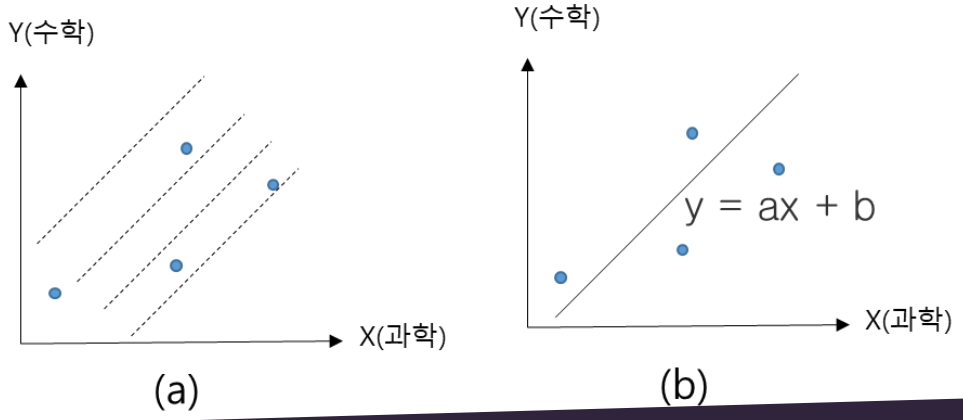


통계 기초 이론: 회귀분석

- 선형회귀분석(Linear Regression Analysis)은 쌍으로 관찰된 연속형 변수들 사이의 관계에 있어서 한 변수를 원인으로 하고 다른 변수들을 결과로 하는 분석
- 독립변수와 종속변수 사이 선형식을 구하고 그 식을 이용하여 변수값 들이 주어 졌을 때 종속변수의 변수 값을 예측하는 분석방법

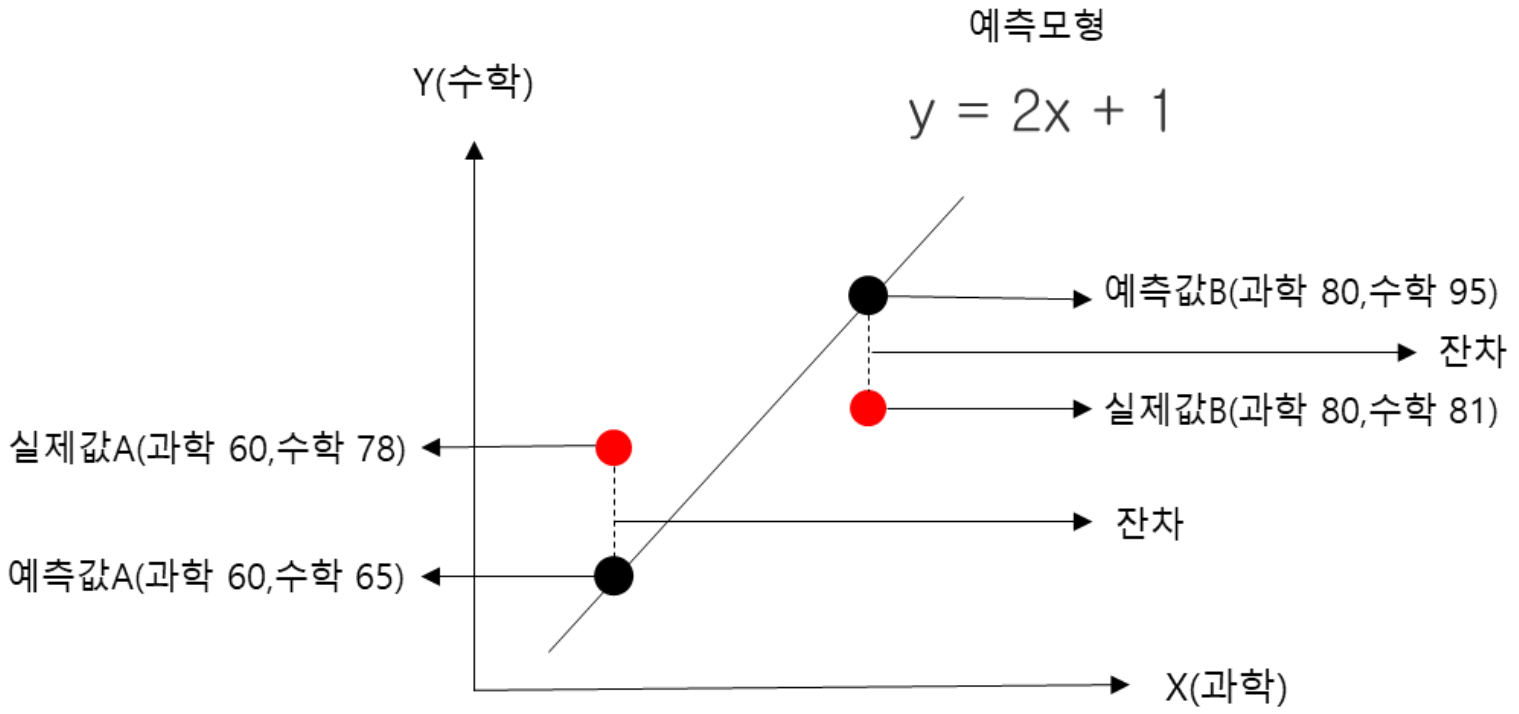
X	Y
독립변수, 설명변수, 원인변수	종속변수, 반응변수, 결과변수 머신러닝(클래스, 라벨)
다른 변수에 영향을 주는 원인	다른 변수에 영향을 받는 결과

- x변수와 y변수 간의 관계를 y = ax + b 와 같은 하나의 선형 관계식으로 표현
- y = ax + b 인 회귀식에서 독립변수 x가 하나인 것이기에 단순선형회귀분석이라함



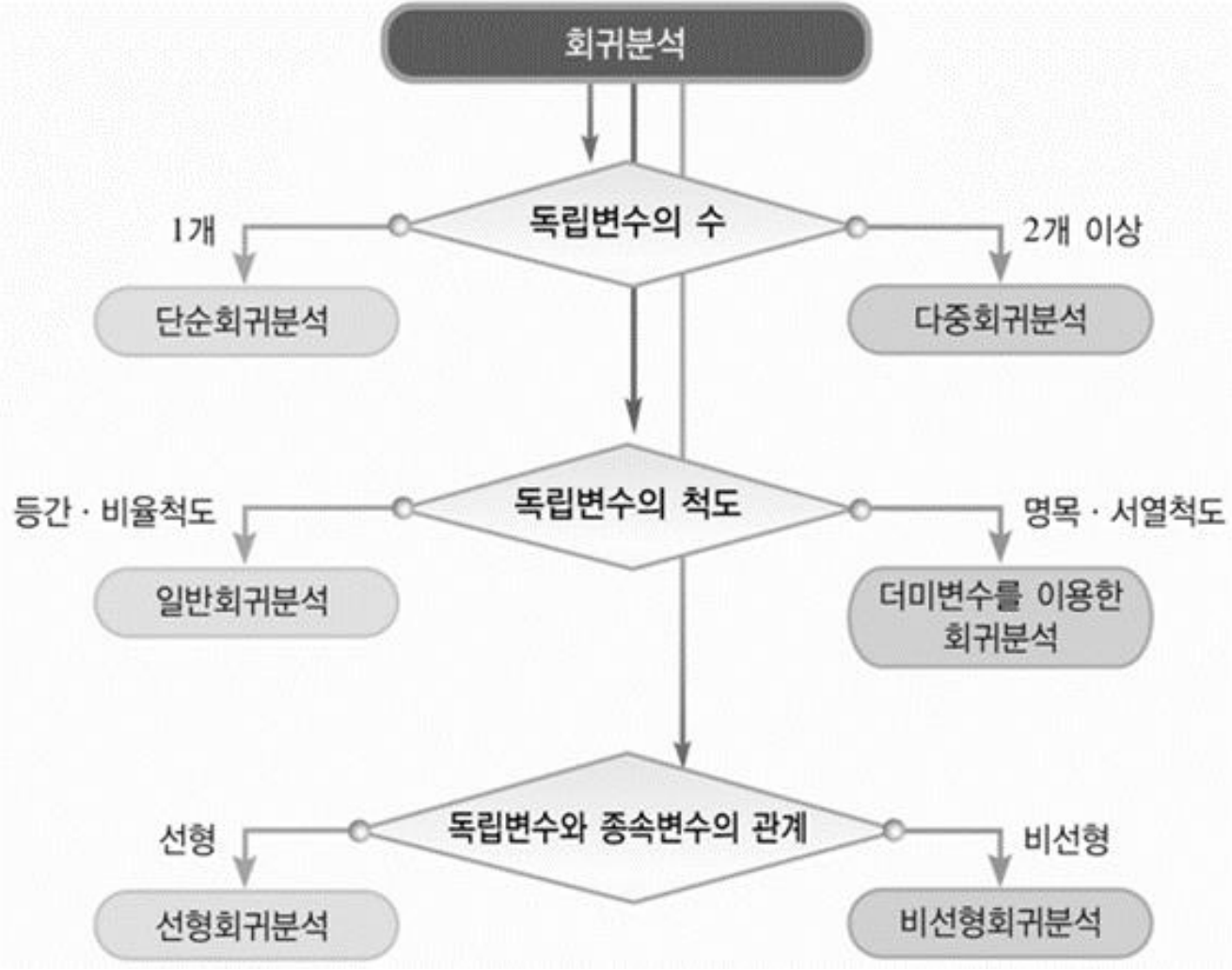
통계 기초 이론: 단순선형회귀분석

- 수집한 데이터를 대표하는 하나의 직선을 찾는 것이 회귀분석의 목적
  - '대표한다'는 기준을 무엇이냐에 따라 회귀선도 달라질 수 있음
- 선형 회귀분석은 최소제곱법 또는 최고자승법이라고도 부르는 것을 기준으로 회귀선을 찾아감
- y = ax + b가 최소제곱법으로 찾은 회귀식이 됨



통계 기초 이론: 선형회귀분석

- 선형회귀분석은 다시 독립변수의 개수에 따라 단순 선형과 다중 선형으로 구분



통계 기초 이론: 변수 종류에 따른 통계 분석 종류

- 종속변수와 독립변수의 종류에 따라 통계 분석법의 종류

Y 종속변수 (반응변수)	X 독립변수(설명변수)	통계분석법	귀무가설
연속형	범주형(2개 범주)	T-검정, paired T-검정	집단 간 평균이 동일
연속형	범주형(3개 이상)	분산분석(ANOVA)	집단 간 평균이 동일
연속형	연속형	회귀분석	회귀 계수 = 0
연속형	혼합(수치형+범주형)	공분산분석(ANCOVA)	
범주형	범주형	$\chi 2$검정/로짓분석	집단 간 연관성이 없음/ 회귀 계수 = 0
범주형	혼합(수치형+범주형)	로짓분석	
생존시간	혼합(수치형+범주형)	생존분석	

EDA 예제

와인 품질 예측

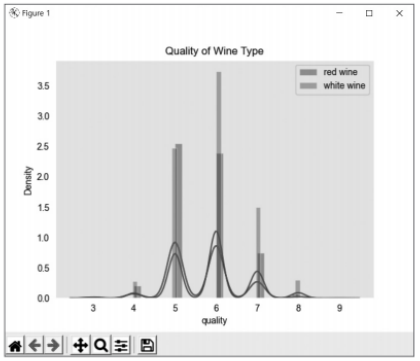
타이타닉호 생존율 분석

01. [기술 통계 분석 + 그래프] 와인 품질 예측하기

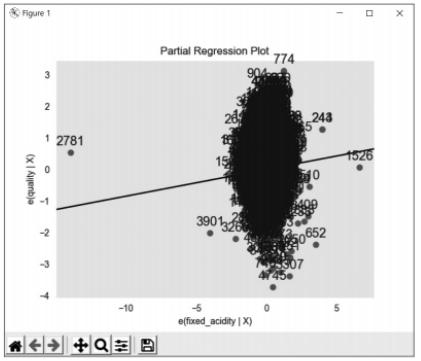
● 분석 미리보기

와인 품질 등급 예측하기	
목표	와인 속성을 분석하여 품질 등급을 예측한다.
핵심 개념	기술 통계, 회귀 분석, t-검정, 히스토그램
데이터 수집	레드 와인/화이트 와인 데이터셋: 캘리포니아 어바인 대학의 머신러닝 저장소에서 다운로드
데이터 준비	수집한 데이터 파일 병합
데이터 탐색	1. 정보 확인: info() 2. 기술 통계 확인: describe(), unique(), value_counts()
데이터 모델링	1. 데이터를 두 그룹으로 비교 분석 • 그룹별 기술 통계 분석: describe() • t-검정: scipy 패키지의 ttest_ind() • 회귀 분석: statsmodels.formula.api 패키지의 ols() 2. 품질 등급 예측 • 샘플을 독립 변수(x)로 지정 → 회귀 분석 모델 적용 → 종속 변수(y)인 품질(quality) 예측
결과 시각화	

1. 히스토그램을 이용한 시각화



2. 부분 회귀 플롯을 이용한 시각화



01. [기술 통계 분석 + 그래프] 와인 품질 예측하기

- 목표 설정
  - 목표: 와인의 속성을 분석한 뒤 품질 등급을 예측하는 것
  - 데이터의 기술 통계를 구함
  - 레드 와인과 화이트 와인 그룹의 품질에 대한 t-검정을 수행
  - 와인 속성을 독립 변수로, 품질 등급을 종속 변수로 선형 회귀 분석을 수행

01. [기술 통계 분석 + 그래프] 와인 품질 예측하기

■ 핵심 개념 이해

- 기술 통계 (요약 통계)
  - 데이터의 특성을 나타내는 수치를 이용해 분석하는 기본적인 통계 방법
  - 평균, 중앙값, 최빈값 등을 구할 수 있음

- 회귀 분석
  - 독립 변수, x와 종속 변수, y 간의 상호 연관성 정도를 파악하기 위한 분석 기법
  - 하나의 변수가 변함에 따라 대응 되는 변수가 어떻게 변하는지를 측정하는 것
  - 변수 간의 인과관계를 분석 할 때 많이 사용
  - 독립 변수가 한 개이면 단순 회귀 분석, 두 개 이상이면 다중 회귀 분석
  - 독립 변수와 종속 변수의 관계에 따라 선형 회귀 분석과 비선형 회귀 분석으로 나눠짐
  - 선형 회귀 분석 식: $y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$

- t-검정
  - 데이터에서 찾은 평균으로 두 그룹에 차이가 있는지 확인하는 방법
  - 예) A와 인의 품질이 1등급인지 2등급인지에 따라 가격에 차이가 있는지를 확인할 때 사용

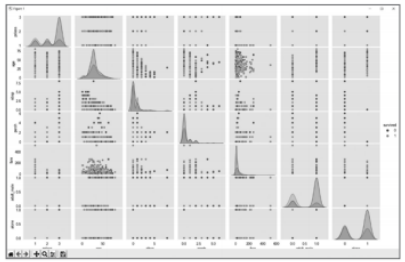
- 히스토그램
  - 데이터 값의 범위를 몇 개 구간으로 나누고 각 구간에 해당하는 값의 숫자나 상대적 빈도 크기를 차트로 나타낸 것

02. [상관 분석 + 히트맵] 타이타닉호 생존율 분석하기

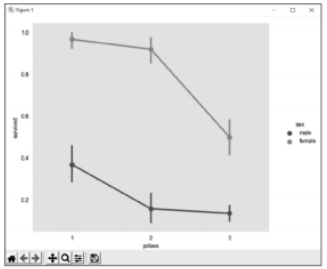
● 분석 미리보기

타이타닉호 생존율 분석하기	
목표	타이타닉호 승객 변수를 분석하여 생존율과의 상관관계를 찾는다.
핵심 개념	상관 분석, 상관 계수, 피어슨 상관 계수, 히트맵
데이터 수집	타이타닉 데이터: seaborn 내장 데이터셋
데이터 준비	결측치 치환: 중앙값 치환, 최빈값 치환
데이터 탐색	1. 정보 확인: info() 2. 차트를 통한 데이터 탐색: pie(), countplot()
데이터 모델링	1. 모든 변수 간 상관 계수 구하기 2. 지정한 두 변수 간 상관계수 구하기
결과 시각화	

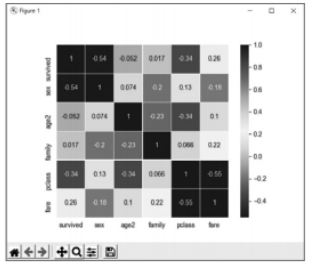
1. 산점도를 이용한 시각화



2. 특정 변수 간 상관관계 시각화



3. 히트맵을 이용한 시각화



02. [상관 분석 + 히트맵] 타이타닉호 생존율 분석하기

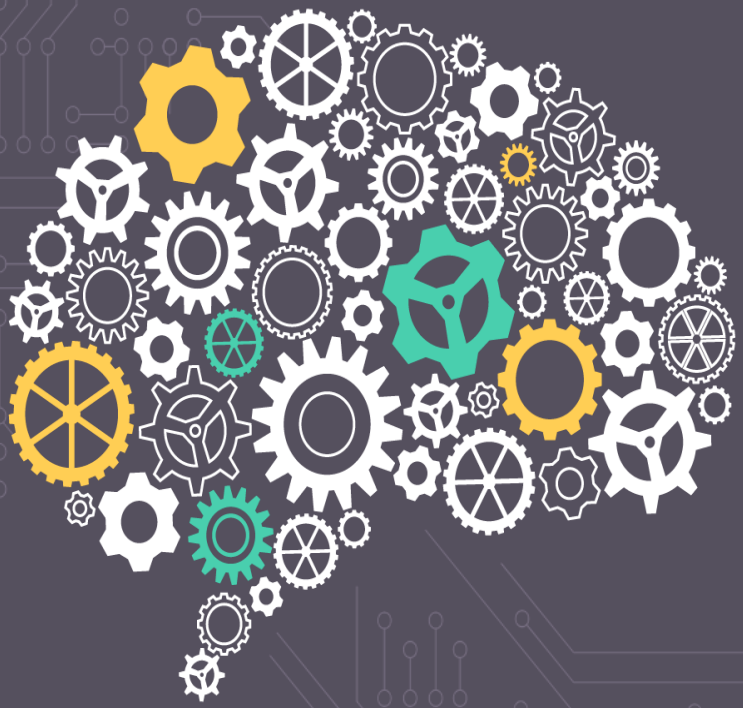
- 분석 미리보기
  - 타이타닉호의 생존자와 관련된 변수의 상관관계를 찾아봄
  - 생존과 가장 상관도가 높은 변수는 무엇인지 분석
  - 상관 분석을 위해 피어슨 상관 계수를 사용
  - 변수 간의 상관관계는 시각화하여 분석

02. [상관 분석 + 히트맵] 타이타닉호 생존율 분석하기

- 핵심 개념 이해
  - 상관 분석
    - 두 변수가 어떤 선형적 관계에 있는지를 분석하는 방법
    - 두 변수는 서로 독립적이거나 상관된 관계일 수 있는데, 두 변수의 관계의 강도를 상관관계 라고함
    - 상관 분석에서는 상관관계의 정도를 나타내는 단위로 모상관 계수 ρ를 사용
    - 상관 계수는 두 변수가 연관된 정도를 나타낼 뿐 인과 관계를 설명하지 않으므로 정확한 예측치를 계산할 수는 없음

    - 단순 상관 분석
      - 두 변수가 어느 정도 강한 관계에 있는지 측정
    - 다중 상관 분석
      - 세 개 이상의 변수 간 관계의 강도를 측정
      - 편상관 분석: 다른 변수와의 관계를 고정하고 두 변수 간 관계의 강도를 나타내는 것

  - 상관 계수 ρ
    - 변수 간 관계의 정도(0~1)와 방향(+, -)을 하나의 수치로 요약해주는 지수로 -1에서 +1 사이의 값을 가짐
    - 상관 계수가 +이면 양의 상관관계이며 한 변수가 증가하면 다른 변수도 증가
    - 상관 계수가 –이면 음의 상관관계이며 한 변수가 증가할 때 다른 변수는 감소
    - 0.0 ~ 0.2: 상관관계가 거의 없음
    - 0.2 ~ 0.4: 약한 상관관계가 있음
    - 0.4 ~ 0.6: 상관관계가 있음
    - 0.6 ~ 0.8: 강한 상관관계가 있음
    - 0.8 ~ 1.0: 매우 강한 상관관계가 있음



Thank you!

Have a good day ☺

창의융합대학 MSC교육부 유 현 주

comjoo@uok.ac.kr