Step01. 작업환경 셋팅

1. R 다운로드 및 설치

https://cran.r-project.org/

r for the first time 클릭하여 다운로드

2. RStudio 다운로드 및 설치

http://www.rstudio.com

Products 탭을 선택

RStudio => RStudio Desktop을 선택

R 패키지의 업데이트 - 콘솔창에 입력

update.packages()

3. Anaconda3 다운로드

C드라이브에 Anaconda3 폴더를 미리 생성한 후 설치 작업을 수행

https://www.anaconda.com/download/

archive 파일을 이용하여 Anaconda3 2021.05 버전을 다운받아 설치 진행

Anaconda Prompt를 관리자 권한으로 실행하여 R 패키지 설치 진행

cmd에서 실행(설치 시간이 오래 걸림)

conda install -c r r-essentials

ipython notebook을 실행한 후

우측 상단의 New 메뉴를 누르면 R이 보임

4. Jupyter Notebook 커스텀마이징

C드라이브의 사용자(User) 폴더안에 해당 .jupyter 폴더가 보이지 않을 경우

Anaconda Prompt를 관리자 권한으로 실행한 후

jupyter notebook —generate-config 입력 후 Enter

c드라이브의 사용자 폴더안에 .jupyter 폴더가 생성된 것을 확인

custom이라는 하위 폴더를 생성한 후 폴더 안에 custom.css 파일을 생성하여 커스텀마이징 수행

Anaconda에서 제공하는 커스텀마이징을 수행하여 진행

pip install jupyterthemes 설치하거나 기존에 설치된 것을 업데이트

pip install —upgrade jupyterthemes

jt -l(list) // 현재 사용 가능한 모든 테마 목록 출력

Step02. 자료 구조(Data Structure)

- R에서 일반적인 프로그래밍 언어에서 흔히 사용되는 정수, 부동소수, 문자열이 기본적으로 지원
- 수치형(numeric), 문자형(character), 논리형(logical), 복소수형(complex)
- 자료구조 : 변수, 벡터(vector), 행렬(matrix), 데이터 프레임(data frame), 리스트(list)등

1. 변수(Variable)

- 하나 이상의 값을 담기위한 기억공간을 의미
- 변수명 사용자 정의 규칙 : 알파벳, 숫자, '-', '_', '.'로 구성되며 **첫 글자는 문자** 또는 '.'으로 시작
- 변수명의 첫 글자를 '.'으로 시작한다면 뒤에는 숫자가 올 수 없다.
- 할당 연산자(or 대입 연산자) : 〈-, ->, =

2. 팩터(Factor)

- R에서 범주형 데이터(Categorical Data, 질적 자료)를 표현
- 서열형 데이터 : 항목 간의 서열이 존재하는 범주형 데이터
- 명목형 데이터 : 단순 분류인 범주형 데이터
- factor(x, levels, ordered)

  x : 팩터로 변환할 벡터

  levels : 입력한 벡터 x의 범주를 정의한 벡터

  ordered : 서열형 데이터인 경우 TRUE(기본값 FALSE, 명목형 데이터)

3. 벡터(Vector)

- 물리학 용어

  scalar : 크기만 가지고 있는 물리량, Vector : 크기와 방향을 모두 가지고 있는 물리량

- R에서의 Vector : 값들의 집합

  R에서 Scalar인 하나의 숫자도 하나의 원소를 가진 Vector로 취급

- Vector 안의 각 수치를 요소라고 함

- Vector의 특징

  벡터는 R에서 가장 기본적인 자료 구조로 1차원 배열 형태를 가짐

  벡터의 길이는 별도의 선언 없이 요소를 추가한 만큼 늘어남

  하나의 벡터에는 하나의 자료형만 사용

  벡터는 c(combine) 함수로 생성

  일반 벡터 생성 c(), 순열 벡터 생성 seq(from=1, to=10, by=2), 반복 벡터 생성 rep(벡터, times=2)

  벡터의 인덱스는 1부터 시작

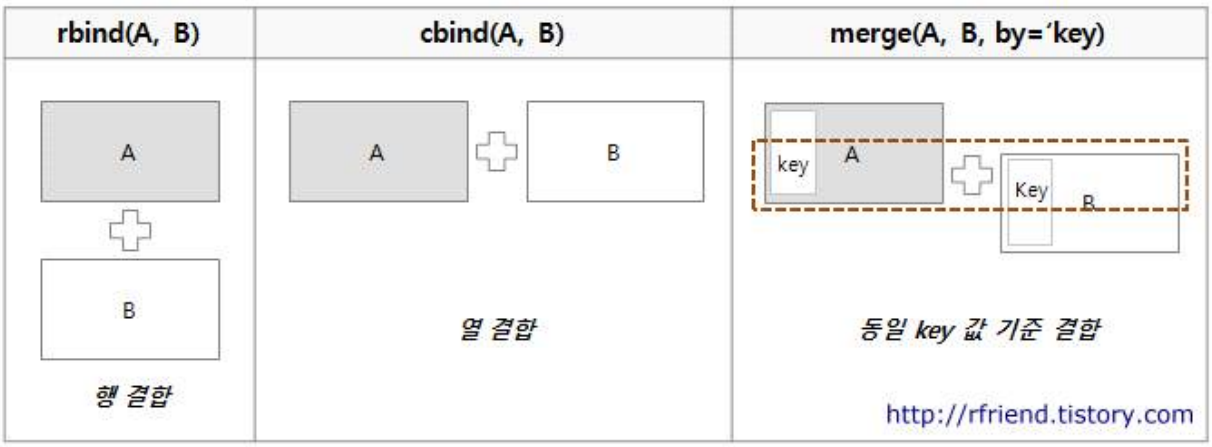
  벡터에서 결측값(누락된 값)은 NA(Not Available) 사용

- Vector 연산의 주요함수

  cor(), cumsum(), length(), max(), mean(), min(), range(), rank(), rev(), sd(), sort(), sum(), summary()

4. 배열(행렬, Matrix)

- 2차원의 자료(행[row]과 열[column]로 구성된)를 저장하는 자료 구조
- 배열 생성 함수 : array, matrix, cbind, rbind 등
- array는 N차원 배열을 생성, array(1:5, c(2, 4))
- matrix는 2차원 배열 생성, matrix(1:12, nrow = 3, bylow = T)
- 기존 벡터를 묶어 배열을 생성할 때 cbind는 열 단위로 묶고, rbind는 행 단위로 묶음



- 유용한 함수

  apply() : 배열의 행 또는 열 단위로 함수 적용

  dim() : 배열의 크기(차원의 수) 반환

  sample() : 벡터나 배열에서 샘플 추출

5. 데이터 프레임(Data Frame)

- R뿐만 아니라, 데이터 분석에서 가장 많이 사용되는 자료 구조
- 표(table, 행과 열로 구성된 틀) 형태로 데이터를 적재하여 활용하는 자료 구조
- 여러 개의 벡터의 조합, 각 벡터의 자료형은 서로 달라도 가능
- data.frame(vector1, vector2, ..., stringsAsFactors)
  vector1..N 데이터 프레임에서 열이 될 벡터
  stringsAsFactors=T/F 문자 데이터형일 경우 팩터로 변환할지 여부, 기본값(생략시, TRUE, 팩터)
- 데이터 접근 : 데이터 프레임명[행 위치 벡터, 열 위치 벡터] or 데이터 프레임명$열이름

6. 리스트(List)

- 데이터 프레임 보다 더 넓은 개념의 자료 구조
- 데이터 프레임과 달리 모든 속성의 크기가 같을 필요가 없음
- 리스트 요소 접근 : $와 [[]]
- 리스트에 유용한 함수 : lapply(), sapply(), is.list(), as.list()

Step03. 제어문과 함수

1. 제어문

- 프로그램의 흐름이나 순서를 조건에 따라 처리할 때 사용

2. 조건문 or 선택문 or 비교/판단문

- [ ] 조건식 : 변수명[행 조건식, 열 조건식]
- if ~ else 문

```
if(조건식){
   조건식이 참일 경우 수행할 문장
}else{
   조건식이 거짓일 때 수행할 문장
}
```

- ifelse(조건식, 조건이 참일 때 실행할 문장, 조건이 거짓일 때 실행할 문장)

3. 반복문

- 주어진 조건 또는 횟수만큼 문장을 반복해서 처리할 때 사용
- repeat{
  반복 수행할 문장
  }
- while(조건식){
  조건식이 참일 때 수행할 문장
  }
- for(개별변수 in 집합변수){
  주어진 횟수만큼 반복해서 수행할 문장
  }

4. 함수(Function)

- 반복적으로 수행되는 기능이나 명령을 하나의 이름으로 묶어서 관리
- 코드의 간결성을 위해 사용
- R 프로그램 안에 내장된 내장함수와 사용자 정의함수로 구성
- 사용자 정의 함수 구조

```
function(매개변수들){
    함수의 내용정의부
    return(반환할 객체)
}
```

Step04. 데이터 분석

1. 데이터 분석이란?

데이터 분석이란 데이터를 정리, 변환, 모델링하는 과정을 통해 유용한 정보를 발굴하여 의사결정을 지원하는 것을 말한다.

데이터 분석이란 산재된 퍼즐조각을 맞추어 의미 있는 인사이트나 가치를 이끌어내는 것

2. 데이터 분석 절차

- 문제 정의 : 분석을 하기 전 분석을 통해 알고 싶은 것이 무엇인지 명확하게 정의하는 것
- 데이터 수집 : 다양한 유형을 데이터를 수집하는 과정
- 데이터 정제 : 수집한 데이터를 분석이 가능한 형태로 처리하는 과정
  변수 선택, 이상치, 결측치 처리 등
- 데이터 분석 : 실제로 데이터를 분석하는 단계, 즉 탐색적 데이터 분석(EDA)라고 부른다.
  기술 통계분석 : 데이터의 빈도, 비율, 평균, 분산 등 일반적인 수치 특성을 분석하는 과정
  추론 통계분석 : 평균 차이, 독립성, 적합도 분석, 상관분석, 회귀분석, 시계열 분석등 표본을 통해 모집단의 특성을 추론하는 분석방법
  데이터마이닝 분석 : 군집분석, 연관분석, 분류분석, 텍스트마이닝 등 대규모 데이터의 패턴이나 규칙을 분석하는 것
- 해석 및 활용 : 데이터 활용이란, 데이터 분석을 통해 추출한 가치 있는 정보 및 지식을 활용하여 문제 또는 환경에 능동적으로 대응하거나 변화 예측에이용하는 것을 뜻한다.

3. 빅데이터 분석 단계

- 문제 정의 단계 : 분석하고자 하는 분야를 이해하고, 해결해야할 문제를 객관적이고 구체적으로 정의하는 단계
- 데이터 수집 단계 : 분석에 필요한 데이터 요건을 정의하고, 데이터를 확보하는 단계
- 데이터 전처리 단계 : 수집한 데이터에 존재하는 결측값이나 오류를 수정/보완하거나 경우에 따라 데이터의 구조나 특성을 변경하는 단계
- 데이터 모델링 단계 : 다양한 관점을 반영한 데이터 설계 단계
- 시각화 및 탐색 단계 : 다양한 유형의 시각화 도구를 이용해서 인사이트를 도출하는 단계

4. 데이터 정제 : 결측값 처리

- 결측값missing value이란 통계에서 누락된 데이터 또는 데이터 수집 단계에서 변수에 값이 저장되지 않아 발생하는 값을 뜻한다.
- 결측값이 포함된 경우 해당 데이터를 삭제하고 사용하거나 보간법interpolation을 써서 다른 값으로 대체하는 작업을 수행한다.
- 결측값 처리 방법

  is.na() 함수 이용 : NA인 데이터가 있으면 T, 없으면 F 반환

  na.omit() 함수 이용 : NA인 데이터를 제거한다. 즉 NA가 포함된 행을 지운다.

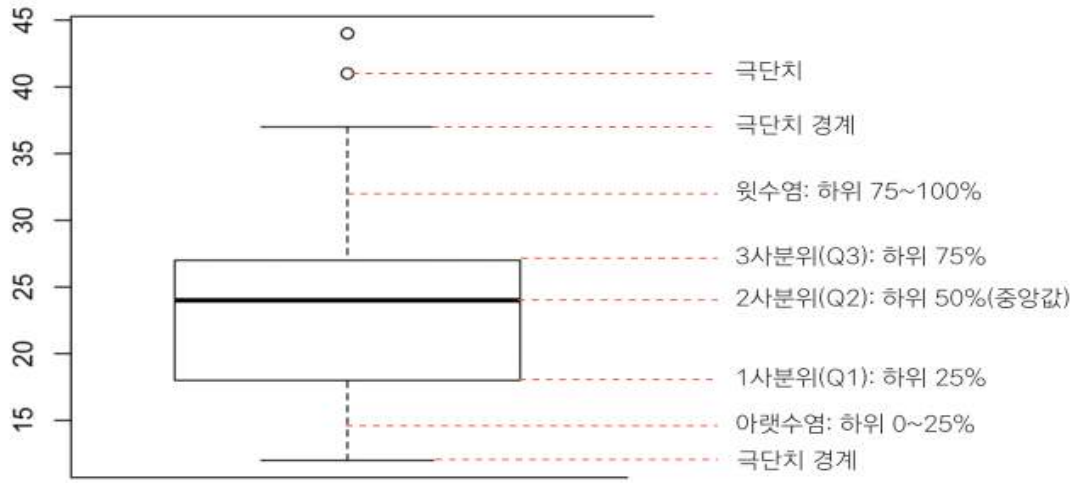
  함수의 속성 이용 : na.rm=T 옵션을 추가해 함수 수행시 NA를 제외한다.

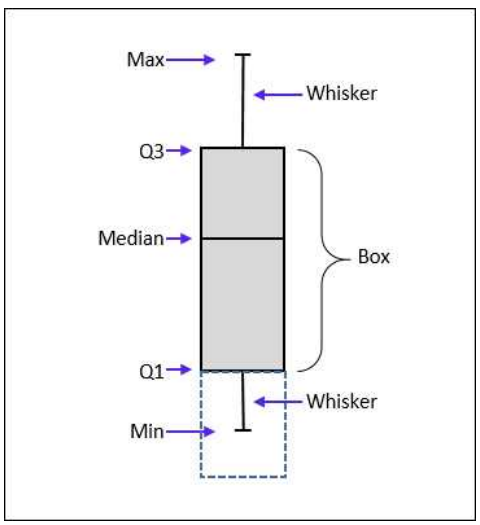
5. 데이터 정제 : 이상값 처리

- 이상값[outlier]는 데이터에 논리적 혹은 통계적으로 맞지 않는 값, 즉 다른 관측값과 멀리 떨어진 관측값을 의미한다.
- 이상값은 측정값의 변동이 원래 큰 경우나 실험적 오류등으로 발생하며 이런 이상값은 통계 분석에서 심각한 문제를 일으킬 수 있는 요소가 된다.
- 정상 범주에서 크게 벗어난 값, 논리적으로 존재할 수 없는 값은 결측 처리후 분석에서 제외
- 통계적 판단 : boxplot(상자수염 그래프)을 이용하여 극단치 기준 정해서 제거

6. Boxplot

- 상자 그림 또는 상자수염 그림
- 이상값(극단치), 최솟값(아래 극단치 경계), 최댓값(위 극단치 경계), 위/아래 수염, 1~3사분위
- IQR[Inter Quartile Range] = 3사분위수 - 1사분위수
- 최솟값 : '중앙값 - 1.5 x IQR' 보다 큰 데이터 중 가장 작은 값
- 최댓값 : '중앙값 + 1.5 x IQR'보다 작은 데이터 중 가장 큰 값
- 이상치(극단치, outlier) = 특이점 : 최솟값보다 작은 데이터 또는 최대값보다 큰 데이터





Step05. 기술 통계분석

- 통계분석은 기술 통계와 추론 통계로 나눌 수 있다.
- 기술 통계(Descriptive statistics)는 데이터를 요약해 설명하는 통계 기법
- 추론 통계(Inferential statistics)는 단순히 숫자를 요약하는 것을 넘어 어떤 값이 발생할 확률을 계산하는 통계 기법

1. 기술 통계 or 기초 통계량 분석

- 수치형 데이터의 통계

  숫자를 이용한 요약 : 기술통계량(평균, 중앙값, 4분위수, 표준편차, 분산, …)

  시각화 도구를 이용한 요약 : 줄기-잎그림, 상자수염그래프, 히스토그램...

- 범주형 데이터의 통계

  숫자를 이용한 요약 : 표(table)

  시각화 도구를 이용한 요약 : 원(파이 차트)그래프, 막대그래프

2. 기술 통계 실습

- 데이터 분석의 시작은 데이터를 “있는 그대로 보는 것” 이다.

2.1 예제 데이터 - 3개의 식당의 배달시간을 측정한 데이터(단위, 분)

- A식당 : 20,21,23,22,26,28,35,35,41,42,43,45,44,45,46,47,47,46,47,58,58,59,60,56,57,57,80
- B식당 : 5,6,11,13,15,16,20,20,21,23,22,27,27,30,30,32,36,37,37,40,40,43,44,45,51,54,70,600
- C식당 : 5,5,5,12,10,11,20,20,20,20,20,21,20,30,32,31,31,31,36,40,40,51,61,51,61,61,70

2.1 평균(mean)

- 데이터의 집합이 어떤 값을 중심으로 분포되어 있는지를 알기 위해 사용
- 일반적으로 자료 전체의 합을 자료 개수로 나눈 산술평균을 의미
- 직관적이고 이해가 쉬워 집합을 대표하는 지표로 많이 사용
- 평균은 이상치에 영향을 많이 받는다는 단점이 있다.

2.2 중앙값(median)

- 데이터 집합을 크기에 따라 차례로 나열했을 때 가운데에 놓이는 값
- 중앙값은 이상치에 의한 영향을 덜 받음(평균의 단점 보안)
- 데이터 분포가 비대칭이면 평균보다 더 의미 있는 지표가 된다.

2.3 사분위수(quantile)

- 데이터 집합을 더 세분화한 것으로 크기순으로 정렬한 후 1/4(25%, 1사분위수), 2/4(중앙값, 50%, 2사분위수), 3/4(75%, 3사분위수), 4/4(100%, 4분위수) 위치에 해당하는 지점의 숫자를 의미

2.4 상자그림(boxplot)

- 사분위수와 이상치(Outlier)를 시각화해 데이터의 중심위치와 분포를 파악하는데 유용한 그래프

2.5 히스토그램(Histogram, hist)

- 연속된 수를 구간별로 나누고 그 구간에 해당하는 빈도수를 표현한 그래프
- hist(x, main, xlab, ylab)

  x : 데이터 벡터　　main : 그래프 제목

  xlab : x축 제목　　ylab : y축 제목

2.6 분산(Variance, var)

- 평균값으로부터 떨어져 있는 정도를 나타내는 지표
- 각 데이터 요소들과 평균 간의 차에 대한 평균

2.7 표준편차(Standard Deviation, sd)

- 분산을 통해 집단 간의 차이는 비교할 수 있으나 어느 정도 차이가 나는지 확인하기 위해 분산을 원래 데이터의 단위로 보정하는 것

2.8 범주형 데이터

- 일반적으로 범주형 데이터들은 범주별 건수를 추출해 범주 간 차지하는 비율을 확인
- 파이차트(pie chart, pie)와 막대차트(bar chart, barplot)를 사용
- table() 함수를 활용하여 범주별 개수를 산출
- pie(x, labels, col, lty, main)

  x : 데이터 벡터　　labels : 범주명을 지정한 벡터

  col : 각 범주별 색상을 지정한 벡터　　main : 그래프 제목

  lty : 선종류

  (0=blank, 1=solid(기본값), 2=dashed, 3=dotted, 4=dotdash, 5=longdash, 6=twodash)

- 실습 데이터

  blood.type <- c('A','B','A','AB','O','A','O','B','A','O','O','B','O','A','AB','B','O','A','A','B')

Step06. 추론 통계분석

1. 모집단과 표본

- 모집단 : 우리가 알고자 하는 대상 전체, 조사 대상의 범위
- 표본 : 모집단으로부터 조사하기 위해 선택된 조사 대상

2. 전수조사와 표본조사

- 전수조사 : 모집단을 구성하는 대상 전부를 조사하는 것
  가장 정확하지만, 비용과 시간이 많이 들게 됨
  전수조사가 불가능한 경우도 있음(예를 들어 감기약의 경우 모두 복용을 해야만 효과를 알 수 있음)
- 표본조사 : 표본을 대상으로 조사

3. 통계분석 기법

- 어떤 그룹, 집단, 형태 등의 **차이를 검정**
  1개, 2개 또는 그 이상의 데이터 차이가 있다고 볼 수 있는지를 검정하는 것
  독립표본 t검정, 대응표본 t검정, ANOVA 등

- 요소와 요소간의 인과관계(상관관계)를 파악
  상관분석 - 변수와 변수 사이의 직선 관계를 상관계수를 이용해서 분석
  회귀분석 - 종속변수와 독립변수간의 관계를 모형화하여 분석

4. 가설검정

- 가설 : 모집단의 특성, 특히 모수에 대한 가정 혹은 잠정적인 결론
  분석의 목적이 정해지면 분석을 통해 확인하고자 하는 명제

- 과학 분야에서의 증명 : **반증법에 의거해 증명**
  "모든 사람은 정직하다"라는 명제가 있을 때 이 명제가 참인지 거짓인지를 확인하는 접근법에는 2가지가 존재
  ✔ 모든 사람을 일일이 조사해서 정직한지 확인하는 방법
  ✔ 정직하지 못한 사람(사례)을 하나 찾아내 명제가 거짓임을 입증하는 방법

  위 2가지 방법 중 어느 것이 효율적일까? (현실적으로 가능한 방법일까?)
  1000명이 정직함을 확인했을 때, 1001명째 사람이 정직하다고 확언할 수 있을까?

5. 귀무가설(H_0)과 대립가설(H_1)

- 귀무가설(Null Hypothesis) : 가설검정에서 연구자의 주장에 대한 부정 진술
  현재까지 주장되어 온 것이거나 기존과 비교하여 변화 혹은 차이가 없음을 나타내는 가설
- 대립가설(Alternative Hypothesis) : 가설검정에서 연구자의 주장이 담긴 진술, 연구가설

- 예를 들어 한 제약회사가 기존 약 A와 비교하여 효과가 향상된 두통약 B를 개발했다고 가정해보자. 제약회사는 신약 B가 효과가 있음을 입증하고자 할 것이다. 이러한 상황에서
귀무가설은 "두통약 A와 B 간의 효과 차이가 없다"와 같이 나타낼 수 있지만,
대립가설은 "두통약 A와 B 간의 효과 차이가 존재한다"와 같이 나타낼 수 있다.

- 이렇게 도출된 가설들은 분석을 통해 통계적으로 의미가 있는지를 검증하게 되는데 두 가설을 모두 검증하는 것이 아니라 기존의 가설(귀무가설)이 잘못됐다는 것을 증명함으로써 새로운 가설(대립가설)을 채택하는 방식을 주로 사용

6. p-value(유의확률)

- 통계의 유의성을 대표하는 지표
- 귀무가설이 참이라는 가정 아래 얻은 통계량이 귀무가설을 얼마나 지지하는지를 나타내는 확률
- 유의수준 : 가설 채택 또는 기각의 기준
- 사회과학 분야의 임계 값(알파 = 0.05), 의생명분야의 임계 값(알파 = 0.01)
  p-value 〉= 0.05 : 대립가설 기각, 귀무가설 채택
  p-value 〈 0.05 : 대립가설 채택, 귀무가설 기각

7. 독립표본 t-검정

- 서로 독립된 두 집단 간의 평균의 차이가 통계적으로 유의미한지 비교하고자 할 때 사용 즉,
  서로 독립된 두 집단에 대해 각 집단별 특정 연속형 변수 평균값이 서로 차이가 있는지 없는지를 통계적으로 검정할 때 사용되는 기법
  예〉 전체 응답자 중 남자와 여자 사이의 연령은 차이가 있는가?

8. 대응표본 t-검정(Paired-sample t-test)

- 서로 동일한 모집단에서 추출된 두 표본에 대해 특정 연속성 변수 평균값이 서로 차이가 있는지, 없는지를 통계적으로 검정할 때 사용되는 기법
  예〉 한 회사에서 자사가 개발한 한 달간의 식이요법 프로그램이 효과가 있는지를 분석하고자 함

9. 일원배치 분산분석(One-way ANOVA)

- 세 개 이상의 집단간의 평균의 차이가 통계적으로 유의미한지 비교하고자 할 때 사용
  예〉 학력수준에 따라 직무만족도의 수준은 차이가 있는가?

Step07. 데이터 분석

1. 데이터 분석 기법

- 지도학습 : 학습데이터(Training Data)를 기반으로 분석을 통해 모델을 도출하고 새로 입력받은 데이터를 모델에 적용해 예측하는 데이터 분석 기법
  회귀분석(regression analysis), 분류분석(classification) 등이 있다.
- 비지도학습 : 특정한 답 없이 주어진 데이터를 분석해 자율적으로 답을 찾아가는 분석 기법
  군집분석(clustering)이나 연관성 분석(association analysis) 등이 있다.

2. 회귀분석

- 어떤 현상을 발생시키는 원인들(독립변수)이 결과(종속변수)에 얼마나 영향을 미치는지를 간략화된 회귀모델 방정식으로 표현하고 이를 통해 분석/예측하는 방법
- 예측에 사용되는 독립변수의 값이 1개면 단순 회귀분석(Simple Linear Regression), 2개 이상이면 다중 회귀분석(Multiple Linear Regression)이 된다.
- lm(fomula, data) 함수
- formula : 회귀분석을 하기 위한 표현식으로 '~', '+', '.', '-', '*'을 이용해 표현
  ~ : 종속변수 ~ 독립변수
  + : 독립변수가 여러 개인 경우, 종속변수 ~ 독립변수1 + 독립변수2 + 독립변수3 ...
  . : 전체 항목
  - : 선택된 독립변수중 제외하고 싶은 독립변수명 앞에 붙임
  * : 독립변수뿐 아니라 상호관계항까지 고려할 때, 종속변수 ~ 독립변수A*독립변수B