Probability & Statistics

통계 기초

강사 양석환

통계학 개요

통계학 개요

- 통계학(Statistics)이란?
  - 수학의 한 분야로서 데이터를 수집, 정리, 분석, 해석하고 이해하는 학문
  - 모집단, 변동, 데이터 축소 방법에 대한 연구하는 학문

  - 다양한 분야에서 활용
    - 비즈니스, 과학, 의학, 사회과학, 정치학 등 다양한 분야에서 중요한 역할을 수행
    - 특히 현대 데이터 과학과 머신러닝 분야에서 데이터 분석과 통계적 기법은 예측, 패턴 인식, 의사 결정 등에 핵심적으로 사용되고 있음

통계학의 목적

- 데이터 수집
  - 통계학은 관심 있는 현상이나 문제에 대한 데이터를 수집하는 과정을 다루며
  - 이러한 데이터는 실험, 조사, 측정, 관찰 등 다양한 방법으로 얻을 수 있음
- 데이터 분석
  - 수집한 데이터를 정리하고 요약하여 패턴이나 추세를 파악하는 과정
  - 이를 통해 데이터의 특성을 파악하고 인사이트를 도출할 수 있음
- 통계 모델링
  - 데이터로부터 확률적인 모델을 구축하고 이를 사용하여 미래의 예측을 수행하는 과정
  - 경제 예측, 의학 연구, 기업 분석 등 다양한 분야에서 활용됨

통계학의 목적

- 불확실성 처리
  - 데이터의 불확실성을 고려하고 처리하는 방법을 제공
  - 표본 오차, 신뢰 구간, 가설 검정 등을 사용하여 결과의 신뢰성을 평가
- 의사 결정 지원
  - 통계학은 의사 결정 과정을 지원하는데 사용됨
  - 데이터를 분석하고 모델링하여 최선의 결정을 내릴 수 있도록 지원

데이터의 기술

기술 통계 분석

- 기술 통계 분석(Descriptive Statistics)
  - 데이터의 주요 특성을 요약하고 설명하는 통계적 기술을 다루는 분석 방법
  - 데이터 집합을 이해하고 요약하여 데이터의 패턴을 파악할 수 있음
  - 데이터 탐색 및 이해의 첫 단계로서 중요함
  - 데이터의 특성을 요약하고 시각화하여 패턴을 파악하는 데 도움을 줌
  - 추론 통계 분석을 통해 데이터에 대한 통계적 가설 검정 및 예측 모델링을 수행할 수 있음

기술 통계 분석

- 기술 통계 분석의 주요 요소
  - 중심 경향성(Measures of Central Tendency)
    - 데이터 집합의 대표값을 계산하며, 이는 데이터의 중심 위치를 나타냄
    - 주요 중심 경향성 측정값: 평균(average), 중앙값(median), 최빈값(mode) 등
  - 분산성(Measures of Variability)
    - 데이터의 퍼짐 정도를 나타냄
    - 주요 분산성 측정값: 분산(variance), 표준 편차(standard deviation)
  - 분포 형태(Distribution Shape)
    - 데이터 분포의 형태를 이해하고 설명함
    - 히스토그램, 박스 플롯 등의 시각적 도구를 사용하여 데이터 분포를 확인할 수 있음

기술 통계 분석

- **기술 통계 분석의 주요 요소**
  - **이상치(Outliers) 탐지**
    - 이상치: 일반적인 데이터 값과 동떨어진 값
    - 기술 통계 분석을 통해 이상치를 식별하고 처리할 수 있음
  - **데이터 요약**
    - 데이터 집합의 합계, 최소값, 최대값, 범위 등의 요약 통계량을 계산하여 데이터의 전반적인 특성을 파악함

탐색적 데이터 분석(EDA)

탐색적 데이터 분석 개요

- **데이터 분석의 접근 방법**
  - **확증적 데이터 분석(CDA: Confirmatory Data Analysis)**
    - 가설을 설정한 후, 수집한 데이터로 가설을 평가하고 추정하는 전통적인 분석
    - 관측된 형태나 효과의 재현성 평가, 유의성 검정, 신뢰구간 추정 등의 통계적 추론을 하는 분석 방법
    - 설문조사나 논문에 관한 내용을 입증하는 데 사용

탐색적 데이터 분석 개요

- **탐색적 데이터 분석(EDA, Exploratory Data Analysis)**
  - 원 데이터(Raw data)를 가지고 유연하게 데이터를 탐색하고, 데이터의 특징과 구조로부터 얻은 정보를 바탕으로 통계모형을 만드는 분석방법
  - 주로 빅데이터 분석에 사용됨

탐색적 데이터 분석 개요

- 확증적 데이터 분석은 *추론통계로, 탐색적 데이터 분석은 *기술통계로 구분할 수 있음
  - 추론통계
    - 수집한 데이터를 이용하여 추론 예측하는 통계 기법
    - 신뢰구간 추정, 유의성 검정 기법 등을 이용함
  - 기술통계
    - 수집한 데이터를 요약 묘사 설명하는 통계 기법
    - 데이터의 대푯값, 분포 등을 이용함

탐색적 데이터 분석 개요

- 탐색적 데이터 분석
  - 벨 연구소의 수학자 존 튜키가 제안한 데이터 분석 방법
  - 통계적 가설 검정 등에 의존한 기존 통계학으로는 새롭게 나오는 많은 양의 데이터의 핵심 의미를 파악하는 데 어려움이 있다고 생각하여 이를 보완한 탐색적 데이터 분석을 도입
  - 데이터를 분석하고 결과를 내는 과정에서 원 데이터에 대한 탐색과 이해를 기본으로 가지는 것이 가장 중요

탐색적 데이터 분석 개요

- 탐색적 데이터 분석의 분석 방향
  - 데이터의 분포와 값을 다양한 각도에서 관찰하며
  - 데이터가 표현하는 현상을 더 잘 이해할 수 있도록 도와주고
  - 데이터를 다양한 기준에서 살펴보는 과정을 통해
  - 문제 정의 단계에서 미처 발견하지 못한 다양한 패턴을 발견하고
  - 이를 바탕으로 기존의 가설을 수정하거나 새로운 가설을 추가할 수 있도록 함
  - 데이터에 대한 관찰과 지식이 이후에 통계적 추론이나 예측 모델 구축 시에도 사용되므로 데이터 분석 단계 중 중요한 단계라고 볼 수 있음

탐색적 데이터 분석 개요

- 탐색적 데이터 분석의 목표
  - 관측된 현상의 원인에 대한 가설 제시
  - 가설은 적절한 통계 도구 및 기법의 선택을 위한 가이드 역할
  - 통계 분석의 기초가 될 가정을 평가
  - 추가 자료수집을 위한 기반 제공

탐색적 데이터 분석 방법

- 탐색적 데이터 분석은...
  - 한 번에 완벽한 결론에 도달하는 것이 아니라
  - 아래와 같은 방법을 반복하여 데이터를 이해하고 탐구하는 과정
    1) 데이터에 대한 질문 & 문제 만들기
    2) 데이터를 시각화하고, 변환하고, 모델링하여 그 질문 & 문제에 대한 답을 찾아보기
    3) 찾는 과정에서 배운 것들을 토대로 다시 질문을 다듬고 또 다른 질문 & 문제 만들기

탐색적 데이터 분석 방법

- 이러한 과정을 기반으로...
- 데이터에서 흥미 있는 패턴이 발견될 때까지, 더 찾는 것이 불가능하다고 판단될 때까지
- 도표, 그래프 등의 시각화, 요약 통계를 이용하여 전체적인 데이터를 살펴보고 개별 속성의 값을 관찰하여
- 데이터에서 발견되는 이상치를 찾아내어
- 전체 데이터 패턴에 끼치는 영향을 관찰하고,
- 속성 간의 관계에서 패턴을 발견함

탐색적 데이터 분석 단계

탐색적 데이터 분석 단계

탐색적 데이터 분석 단계

- 1 단계: 전체적인 데이터 살펴보기
  - 데이터 항목의 개수, 속성 목록, NAN 값, 각 속성이 가지는 데이터형 등 확인
  - 데이터 가공 과정에서 데이터의 오류나 누락이 없는지 확인
  - 데이터의 head와 tail을 확인
  - 데이터를 구성하는 각 속성값이 예측한 범위와 분포를 갖는지 확인

탐색적 데이터 분석 방법

- **2 단계: 이상치(Outlier) 분석**

  1) 개별 데이터를 관찰하여 전체적인 추세와 특이사항을 관찰

  - 데이터가 많다고 특정 부분만 보게 되면 이상치가 다른 부분에서 나타날 수도 있으므로 앞, 뒤, 무작위로 표본을 추출해서 관찰
    - 이상치들은 작은 크기의 표본에서는 나타나지 않을 수도 있음

  2) 적절한 요약 통계 지표를 사용

  - 데이터의 중심을 알기 위해서 평균, 중앙값, 최빈값을 사용
  - 데이터의 분산도를 알기 위해서는 범위, 분산 등을 이용
  - 통계 지표를 이용 시, 평균과 중앙값의 차이처럼 데이터의 특성에 주의해서 이용해야 함

탐색적 데이터 분석 방법

3) 시각화 활용

- 시각화를 통해 데이터의 개별 속성에 어떤 통계 지표가 적절한지를 결정
- 시각화 방법: Histogram, Scatterplot, Boxplot, 시계열 차트 등

- **그 외에도**
  - 기계학습의 K-means 기법
  - Static based detection 기법
  - Deviation based method 기법
  - Distance based Detection 기법 등

  을 이용하여 이상치를 발견

탐색적 데이터 분석 방법

- 3 단계: 속성 간의 관계 분석
  - 속성 간의 관계 분석을 통해 서로 의미 있는 상관관계를 갖는 속성의 조합 도출
  - 분석의 대상이 되는 속성의 종류에 따라서 분석 방법 선택
    - 범주형 (Categorical) 변수: 명목형 데이터, 순서형 데이터 등
    - 수치형 (Numeric) 변수: 연속형 데이터, 이산형 데이터 등

탐색적 데이터 분석 방법

- **이산형 변수-이산형 변수의 경우**
  - 상관계수를 통해 두 속성 간의 연관성 확인
  - Heatmap이나 Scatterplot을 이용하여 시각화
- **이산형 변수-범주형 변수의 경우**
  - 카테고리별 통계치를 범주형으로 나누어서 관찰
  - Box plot, PCA plot 등으로 시각화
- **범주형 변수- 범주형 변수의 경우**
  - 각 속성값의 쌍에 해당하는 값의 개수, 분포를 관찰
  - Piechart, Mosaicplot 등을 이용하여 시각화

시각화

데이터 시각화(Data Visualization)

- 데이터 시각화란?
  - 정보와 데이터를 그래프로 나타내는 것
  - 차트, 그래프, 맵과 같은 시각적 요소를 사용하여
  - 데이터에서 추세, 이상 값 및 패턴을 보고 이해할 수 있도록 해 주며
  - 데이터 분석에 쉽게 접근할 수 있도록 하는 방법
  - 특히 빅 데이터의 세계에서, 데이터 시각화 도구와 기술은 막대한 양의 정보를 분석하고 데이터 기반 의사 결정을 내리는 데에 필수적

데이터 시각화(Data Visualization)

- 데이터 시각화의 필요성
  - 인간은 시력을 통해 얻는 정보양은 다른 기관의 정보보다 훨씬 많음
  - 지나치게 많은 데이터로 인해 이를 관리하고 이해하는 어려움이 계속해서 증가
  - 대부분의 사람들은 통계 데이터에 대해 잘 알지 못하며, 기본적인 통계 방법(평균, 중위수, 범위 등)은 인간의 인지적 성격과 맞지 않음
  - 통계 방법에 따라 규칙을 보는 것은 어렵지만, 데이터가 시각화되면 규칙은 매우 명확히 인지 가능(예: 안스콤비의 4중주)

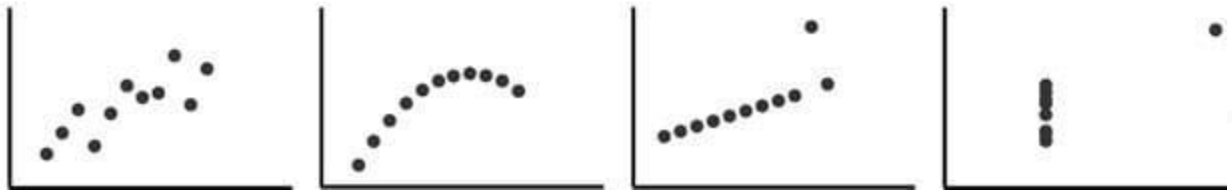
데이터 시각화(Data Visualization)

- 안스콤비의 4중주(Anscombe's quartet)

a

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

b



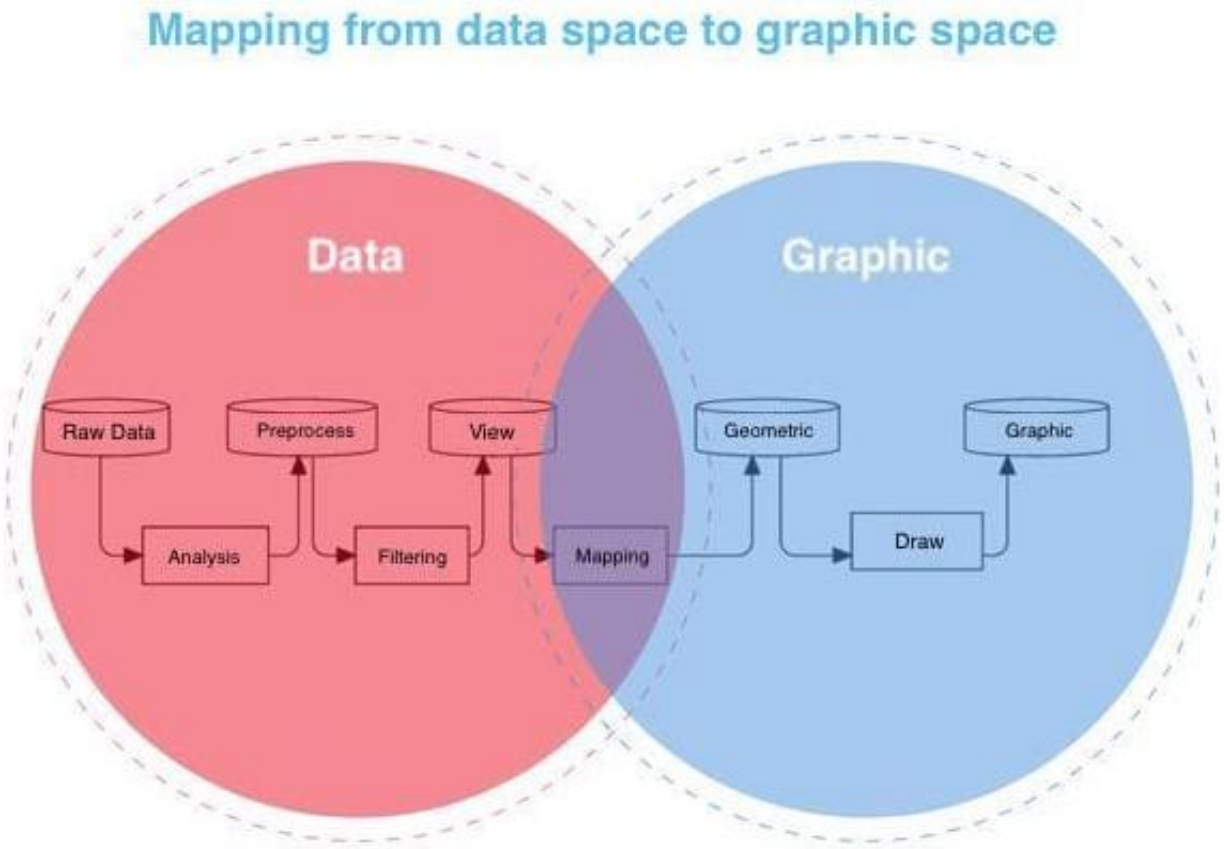
데이터 시각화의 방법

- 데이터 시각화는 데이터 공간에서 그래픽 공간으로의 매핑이다
- 기본적인 시각적 구현 절차

  1.데이터를 처리하고 필터링

  2.표현 가능한 시각적 형태로 변환

  3.사용자가 볼 수 있는 보기로 렌더링

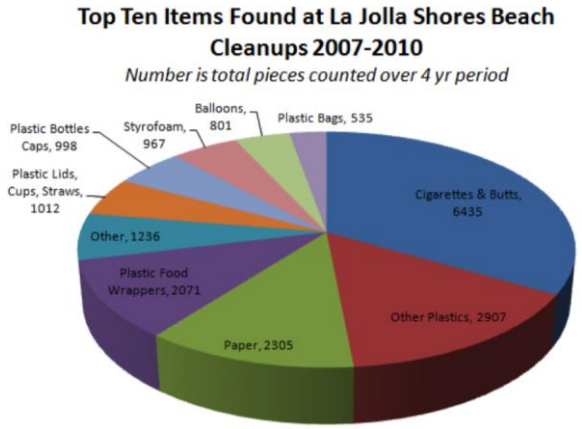


데이터 시각화의 방법

- 데이터 시각화에서 요구되는 기술
  - 기초수학: 삼각함수, 선형대수, 기하 알고리즘
  - 그래픽: 캔버스, SVG, WebGL, 연산 그래픽, 그래프 이론
  - 엔지니어링 알고리즘: 기본 알고리즘, 통계 알고리즘, 공통 레이아웃 알고리즘
  - 데이터 분석 : 데이터 정리, 통계, 데이터 모델링
  - 디자인 미학: 디자인 원리, 미적 판단, 색상, 상호작용, 인지
  - 시각화 기반 : 시각 부호화, 시각 분석, 그래픽 상호 작용
  - 시각화 솔루션: 차트의 올바른 사용, 공통 비즈니스 시나리오의 시각화

데이터 시각화의 유형

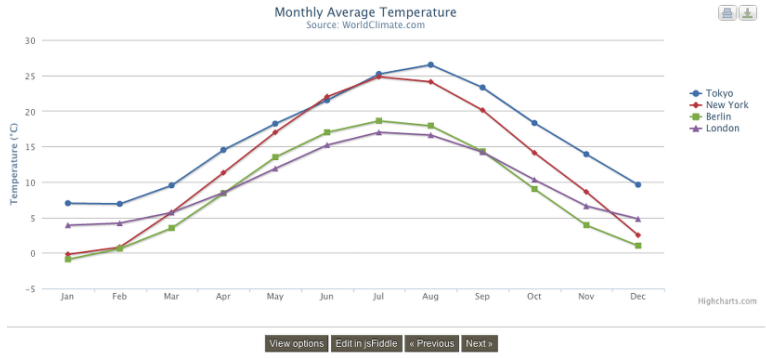
- 널리 사용되는 데이터 시각화의 일반적인 유형
  - 차트
  - 테이블
  - 그래프
  - 맵
  - 인포그래픽
  - 대시보드

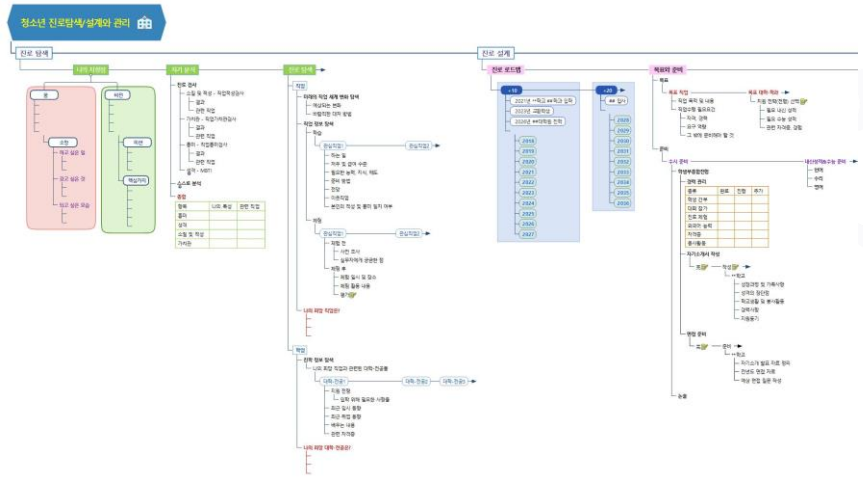


휴대폰 제조업체 매출현황

단위:백만원

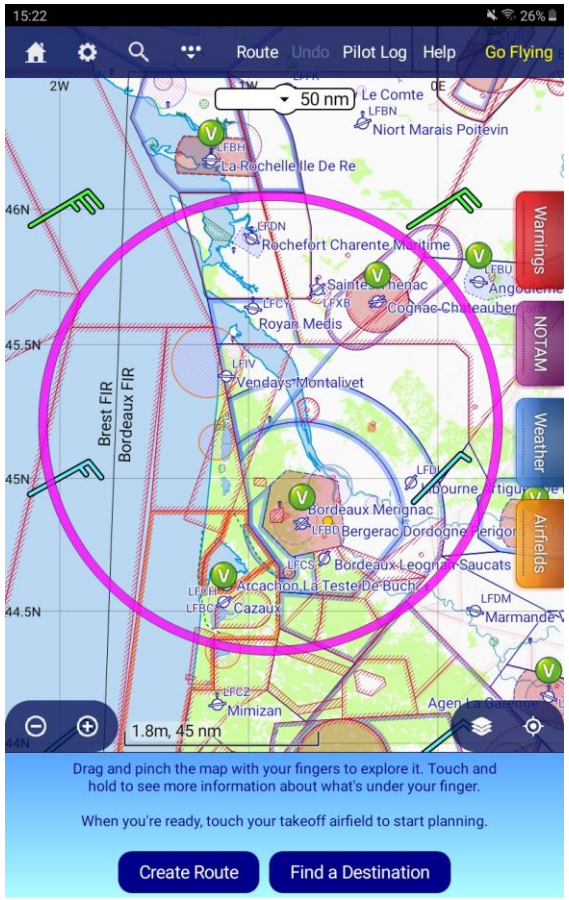
브랜드명	제조업체	생산지	최신모델	2007년	2008년	2009년	2010년	평균매출액
스카이	KS	미국	IM-7100	20,000	24,000	30,000	34,600	27,150
흐르미	NIC	일본	MS-150	16,000	18,600	20,000	24,000	19,650
멀티규	NIC	일본	SCP-A011	12,000	16,000	19,000	18,500	16,375
큐텔	GLS	한국	S2	18,600	23,500	26,400	28,800	24,325
레디안	GLS	한국	SD2100	21,000	30,000	32,000	41,000	31,000
애드콜	SAMS	한국	E-170	35,000	42,000	56,000	66,400	49,850
콜맨	SAMS	한국	E-2500	-	52,000	26,000	28,400	26,600
스카이	KS	한국	IM-8100	-	24,000	26,000	34,000	21,000
흐르미	NIC	일본	SCP-A012	12,000	16,000	19,000	18,500	16,375
큐텔	GLS	한국	S3	23,500	32,400	26,400	23,400	26,425
레디안	GLS	한국	SD2101	32,100	21,000	32,000	32,000	29,275
애드콜	SAMS	한국	E-171	43,000	45,200	32,100	25,000	36,325
콜맨	SAMS	미국	E-2600	-	32,600	26,000	15,000	18,400

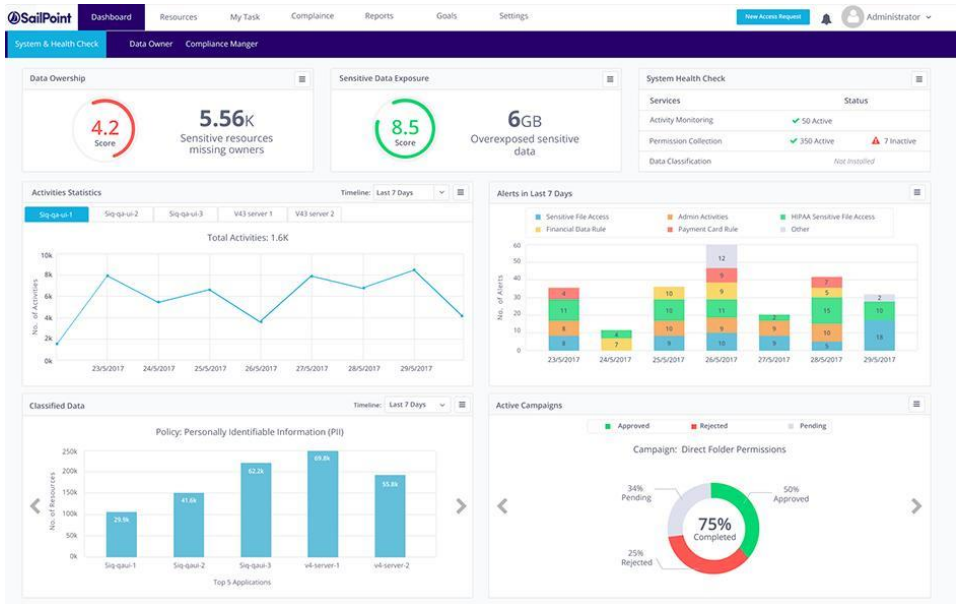




데이터 시각화의 유형

- 널리 사용되는 데이터 시각화의 일반적인 유형





데이터 시각화의 유형

- **데이터 시각화의 구체적인 예**

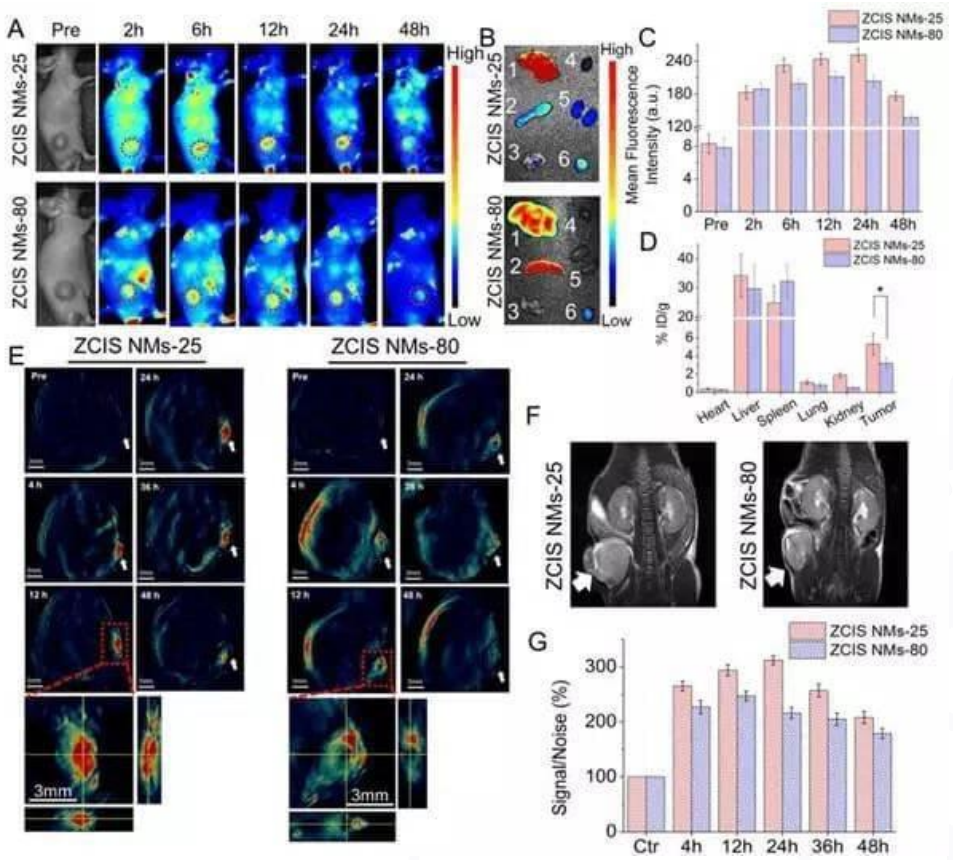
- 영역 차트
- 막대 차트
- 상자-수염 차트
- 버블 클라우드
- 불릿 그래프
- 카토그램
- 원 뷰
- 점 분포 맵
- 간트 차트

- 히트 맵
- 하이라이트 테이블
- 히스토그램
- 행렬
- 네트워크
- 극좌표형 영역(Polar Area)
- 방사형 트리
- 분산형 차트(2D / 3D)
- 스트림 그래프

- 텍스트 테이블
- 타임라인
- 트리 맵
- 쐐기형 누적 그래프 (Wedge Stack Graph)
- 워드 클라우드
- 대시보드를 통한 모든 유형의 조합 등

데이터 시각화의 분류

- 과학적 시각화 (Scientific Visualization)
  - 과학 분야의 학제적 연구 및 응용 분야
  - 건축, 기상학, 의학, 생물학적 시스템과 같은 3차원 현상의 시각화에 초점
  - 과학적 시각화의 목적은 과학자들이 데이터에서 패턴(pattern)을 이해하고, 설명하고, 수집할 수 있도록 과학 데이터를 그래픽으로 설명하는 것



데이터 시각화의 분류

- 정보 시각화 (Information Visualization)
  - 인간의 인식을 향상시키기 위한 추상 데이터의 대화형 시각적 표현에 대한 연구
  - 추상적인 데이터에는 지리적 정보 및 텍스트와 같은 디지털 데이터와 비디지털 데이터가 모두 포함
  - 히스토그램, 추세 그래프, 흐름도 및 트리 다이어그램과 같은 그래픽은 모두 정보 시각화에 속함
  - 이러한 그래픽의 설계는 추상적 개념을 시각 정보로 변환

데이터 시각화의 분류

- 시각적 분석 (Visual Analytics)
  - 과학적 시각화와 정보 시각화의 발전과 함께 진화한 새로운 분야
  - 대화형 시각화 인터페이스를 통한 분석 추론을 강조

Matplotlib 활용

Matplotlib

- **Matplotlib**
  - 파이썬에서 플롯(그래프)을 그릴 때 주로 쓰이는 2D, 3D 플롯팅 패키지(모듈)
  - 저명한 파이썬 라이브러리 개발자인 John Hunter에 의해 개발됨
  - 2003년 version 0.1이 발표된 이후 현재까지 꾸준히 발전해온 약 20년의 역사를 가진 패키지
  - 산업, 교육계에서 널리 쓰이는 수치해석 소프트웨어인 MATLAB과 유사한 사용자 인터페이스를 가지고 있어 각 업계에서 쉽게 접근 가능

Matplotlib

- **Matplotlib의 장점**
  - 동작하는 OS를 가리지 않음
  - 다양한 그래프와 그 구성요소에 대하여 상세한 서식을 설정 가능
  - 다양한 출력형식(PNG, SVG, JPG 등) 지원
  - MATLAB과 유사한 사용자 인터페이스

Matplotlib 활용 기초

- 선 그래프 (Line Plot)
  - 연속하는 데이터 값들을 직선 또는 곡선으로 연결하여 데이터 값 사이의 관계를 나타냄
  - 기본 사용법

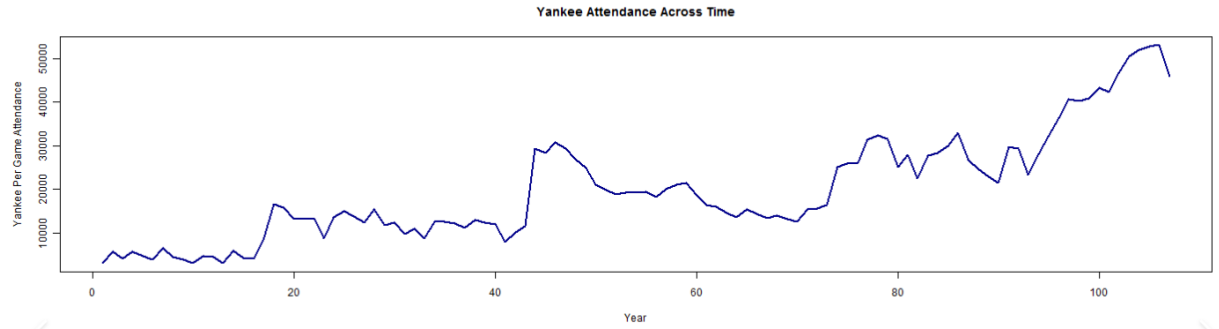
```
import matplotlib.pyplot as plt
plt.plot(x축, y축)
```

```
제목: plt.title('제목')
X축 이름 설정: plt.xlabel('x축이름')
Y축 이름 설정: plt.ylabel('y축이름')
범례 표시: plt.legend()
그래프 표시: plt.show()
```

Matplotlib 활용 기초

- 선 그래프 (Line Plot)
  - Style

옵션	설명
'o'	점 그래프로 표현
marker=마커모양	마커 모양 (예: 'o', '+', '*', '.')
markerfacecolor=색	마커 배경색
markersize=숫자	마커 크기
color=색	선의 색
Linewidth=숫자	선의 두께
label=label이름	label 지정



Color

character	color
'b'	blue
'g'	green
'r'	red
'c'	cyan
'm'	magenta
'y'	yellow
'k'	black
'w'	white

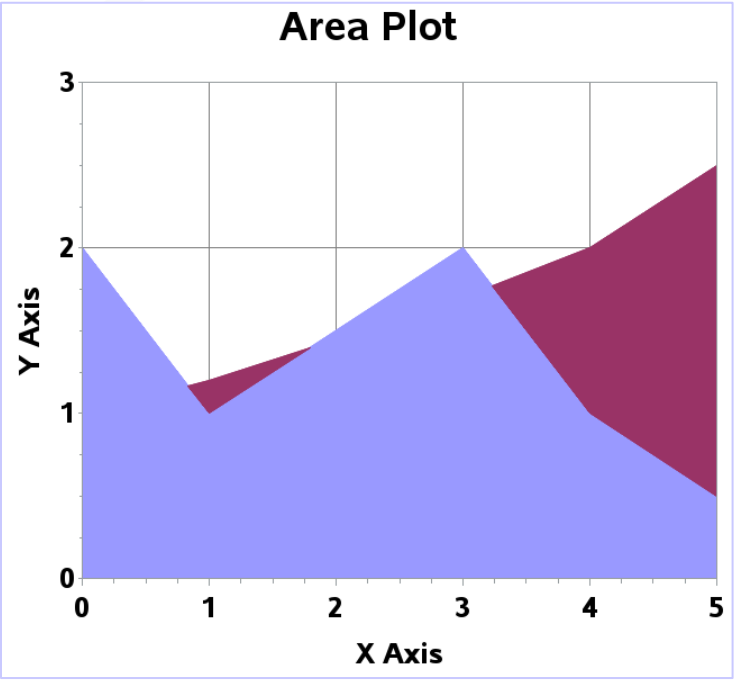
LineStyle

character	description
'-'	solid line style
'--'	dashed line style
'-.'	dash-dot line style
':'	dotted line style

Matplotlib 활용 기초

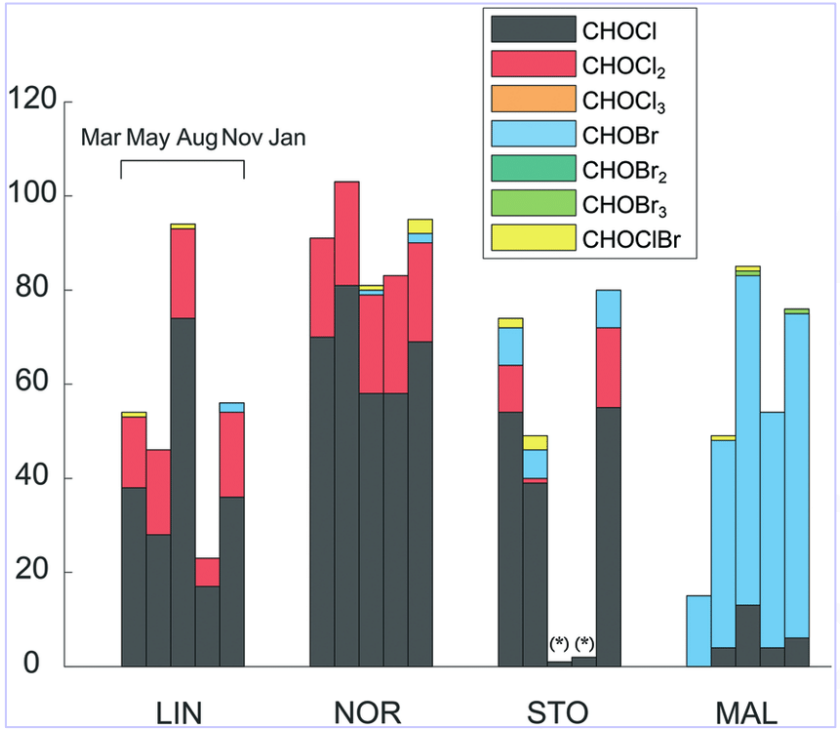
- **면적 그래프(Area Plot)**
  - **선 그래프를 확장한 개념**
  - **각 열의 패턴과 함께 열 전체의 합계가 어떻게 변하는지 파악할 수 있음**
  - **기본 사용법**

- DataFrame객체.plot() 함수에 kind = 'area' 옵션 추가
- 누적 여부 설정: stacked=True/False (기본값: True)
- 색의 투명도 설정: alpha=값(0~1범위, 기본값: 0.5)



Matplotlib 활용 기초

- 막대 그래프 (Bar Plot)
  - 데이터 값의 크기에 비례하여 높이를 가지는 직사각형 막대로 표현
  - 세로형 막대 그래프는 시계열 데이터를 표현하는데 적합
  - 가로형 막대 그래프는 각 변수 사이의 값의 크기 차이를 설명하는데 적합

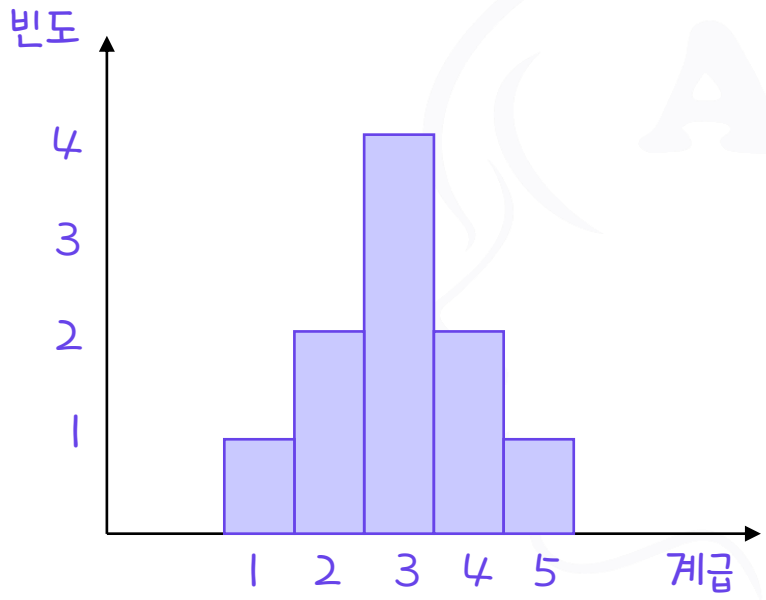


히스토그램

- **히스토그램(Histogram)이란?**
  - **표 형태로 되어 있는 빈도표(Frequency Table)를 그래프 형태로 나타낸 것**
  - **빈도표에서 계급 값은 '값'이 될 수도 있고 '구간'이 될 수도 있음**

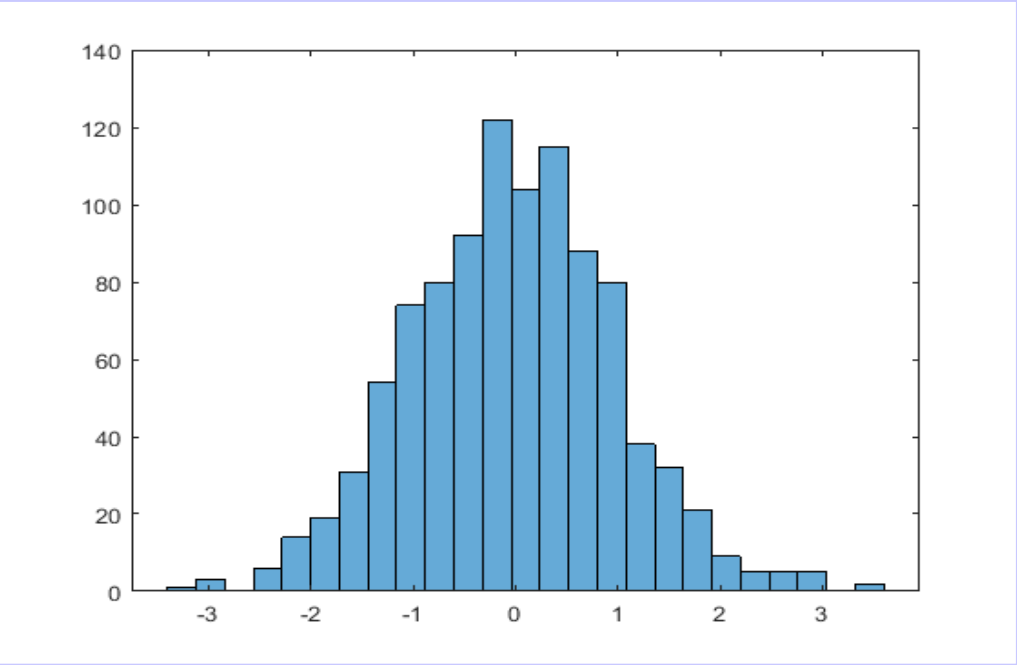
빈도표 = 도수분포표

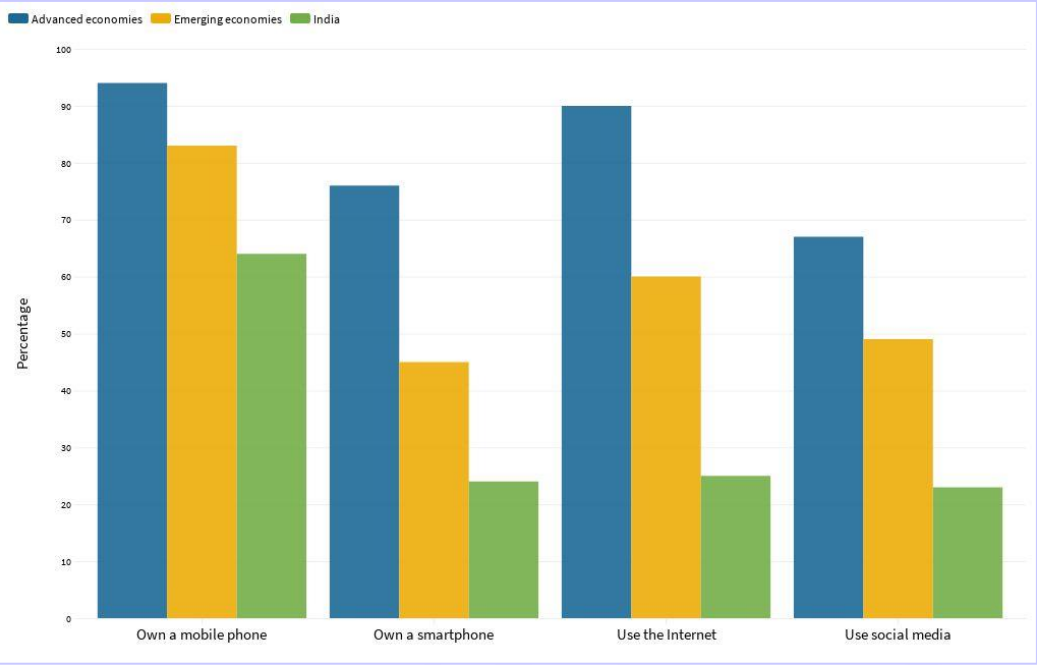
계급	빈도
1	1
2	2
3	4
4	2
5	1



Matplotlib 활용 기초

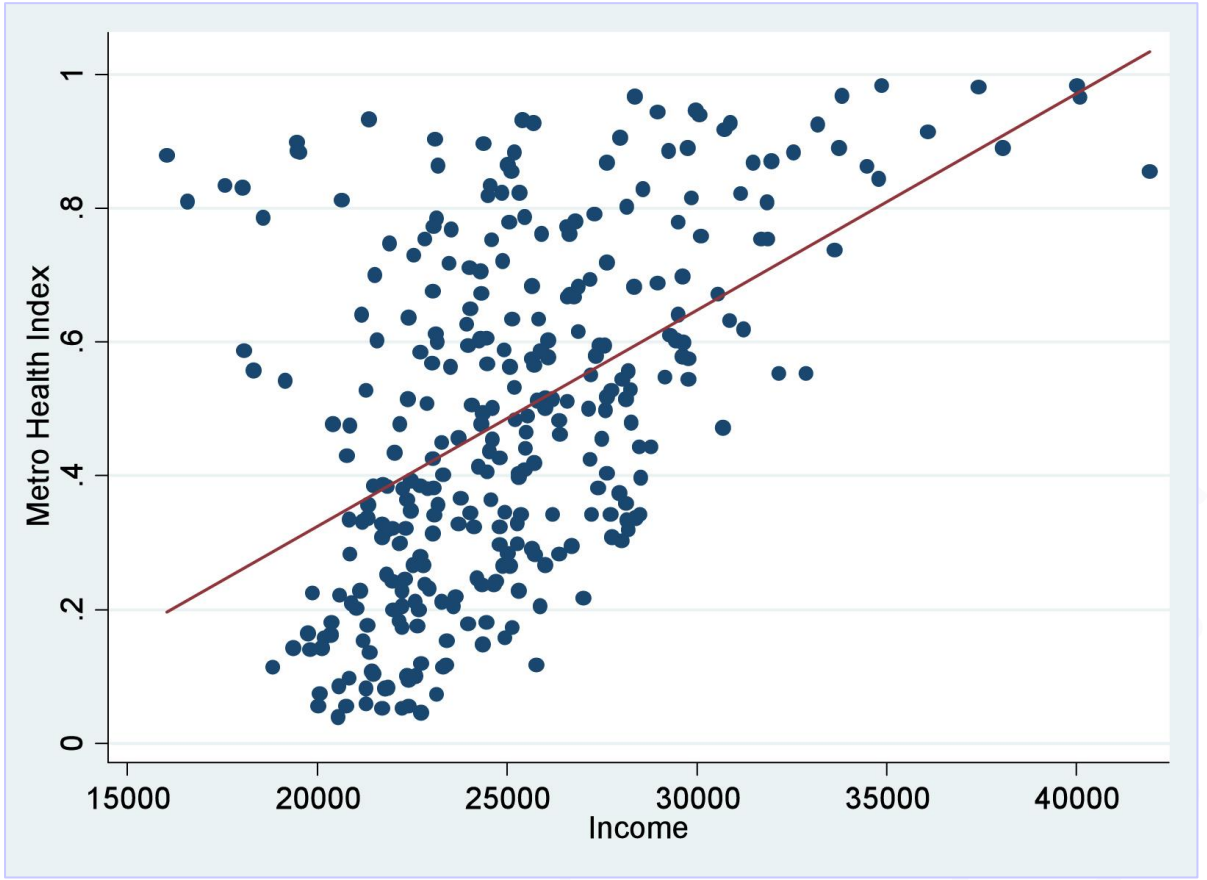
- **변수가 하나인 단변수 데이터에 대한 빈도수를 표현**
  - x축: 같은 크기의 여러 구간, 계급 구간
  - y축: 각 구간에 속하는 데이터 값의 개수(빈도)





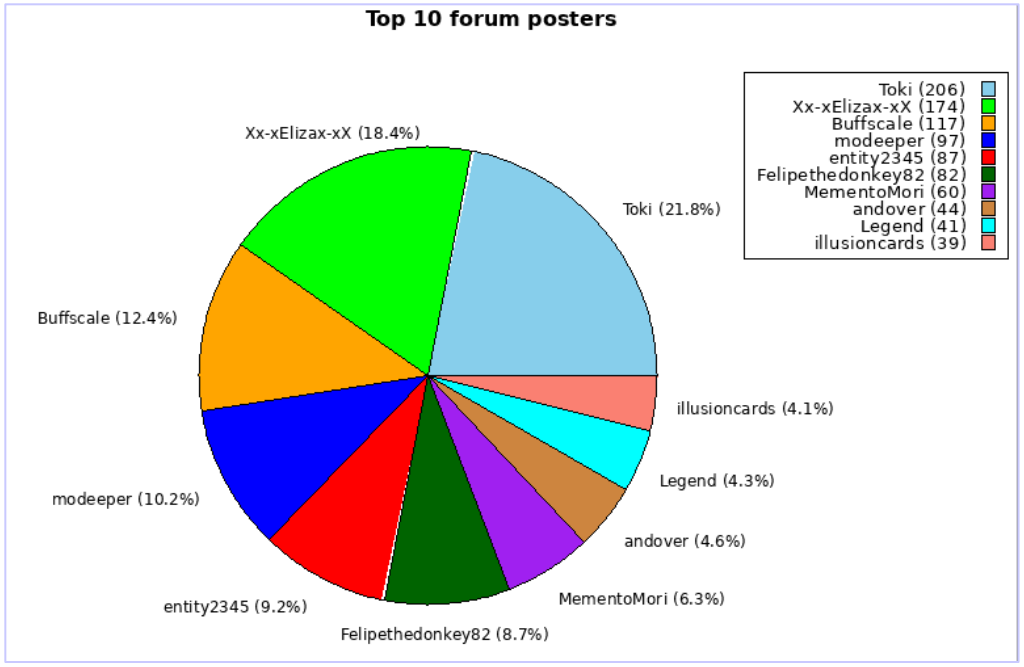
Matplotlib 활용 기초

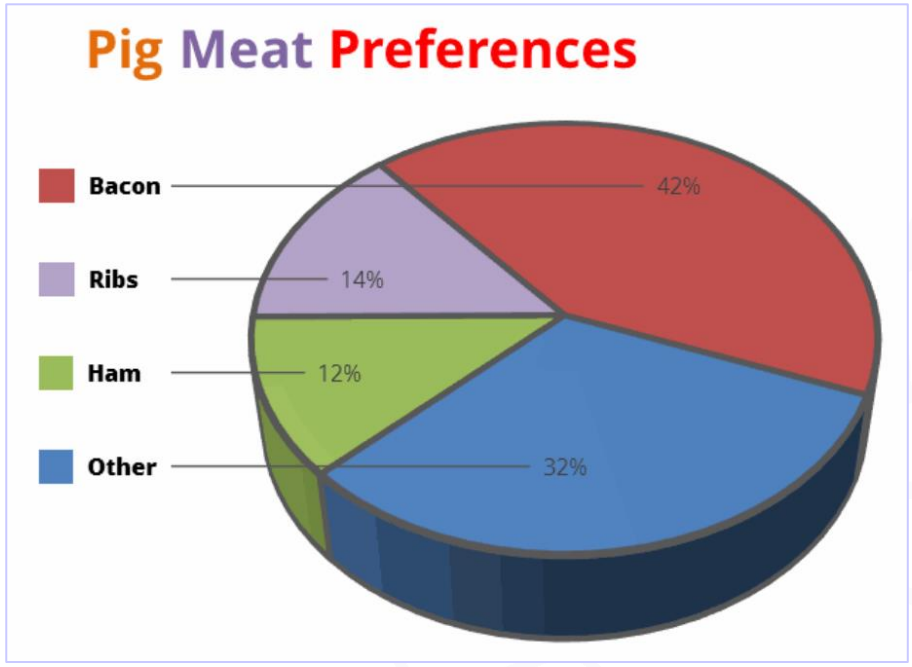
- 산점도 (Scatter Plot)
  - 분산 그래프
  - 서로 다른 두 변수 사이의 관계를 나타냄



Matplotlib 활용 기초

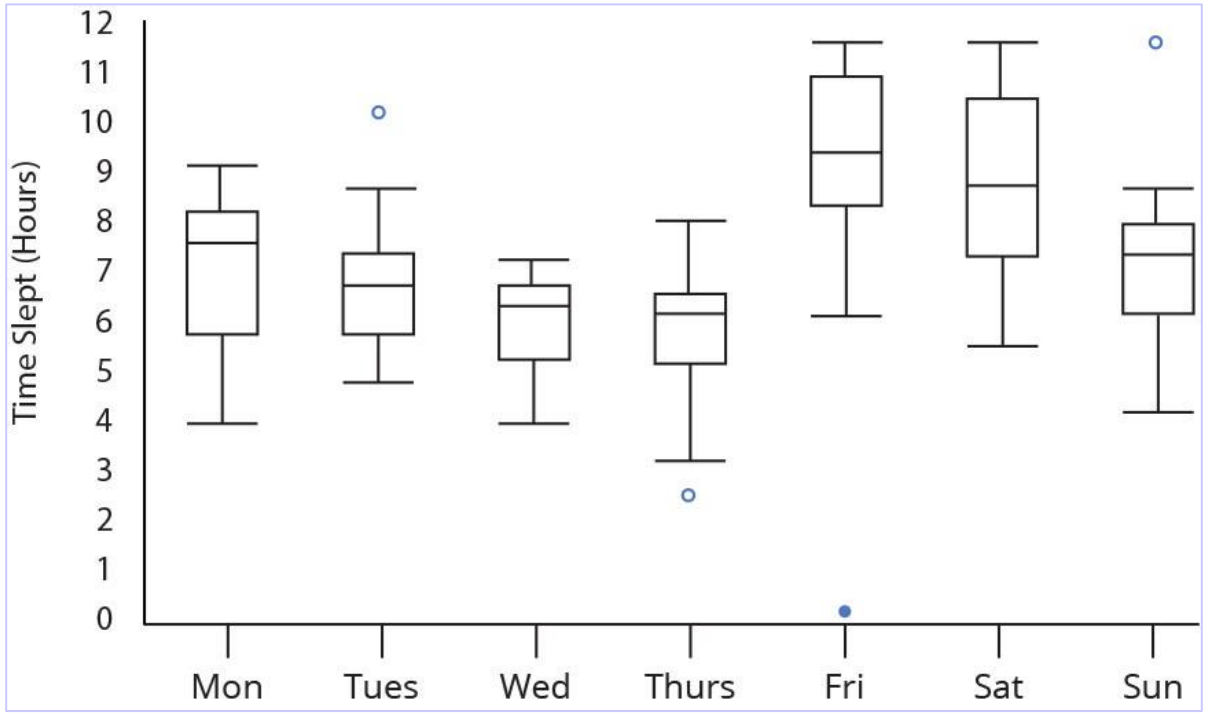
- 파이 차트 (Pie Chart)
  - 원을 파이 조각처럼 나누어서 표현

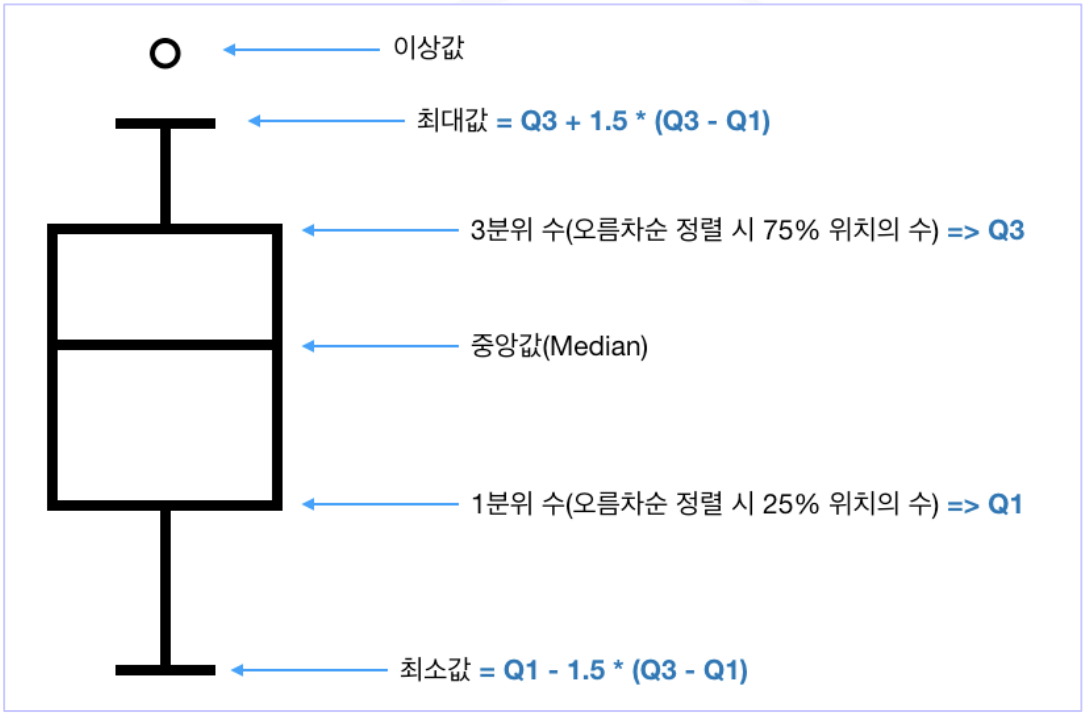




Matplotlib 활용 기초

- 박스 플롯 (Box Plot)
  - 범주형 데이터의 분포(특히 데이터의 불균형)를 파악하는데 적합
  - 5개의 통계 지표(최소값, 1분위값, 중간값, 3분위값, 최대값)를 제공





Matplotlib 활용 기초

- 이미지 출력
  - 2D 이미지
    - 2D Array로 표현되는 이미지
    - 기본 사용법
      - plt.imread()로 이미지를 로드하고 ndarray로 저장
      - plt.imshow()로 내용 확인

```
img1 = plt.imread('c:/data/icecream.jpg')
```

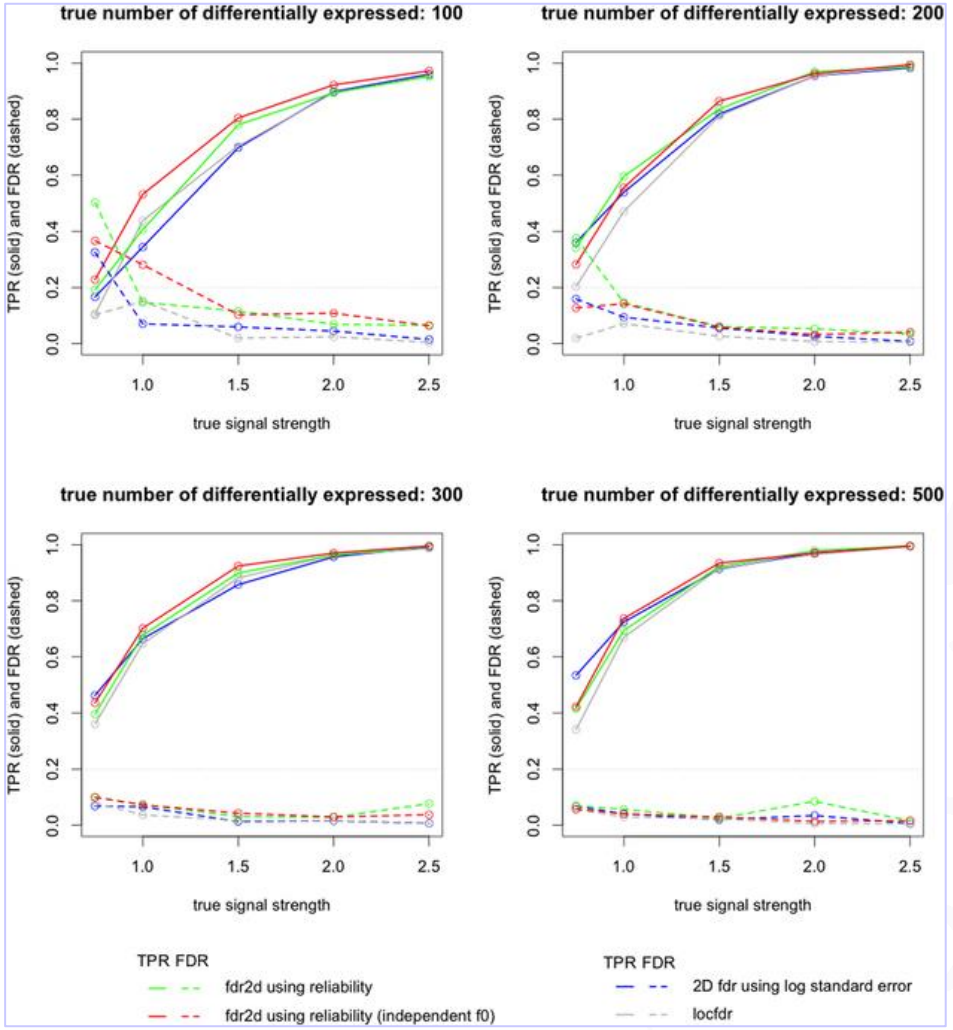
```
plt.imshow(img1)
```

```
plt.imshow(img1[:,:,0], cmap="Reds")
plt.show()
```

Matplotlib 활용 기초

- **화면 분할 (Sub Plot)**
  - **여러 개의 그래프를 한 화면에 표시하기 위하여 화면을 특정 영역으로 분할하여 각 그래프를 배치, 표시하는 기능**

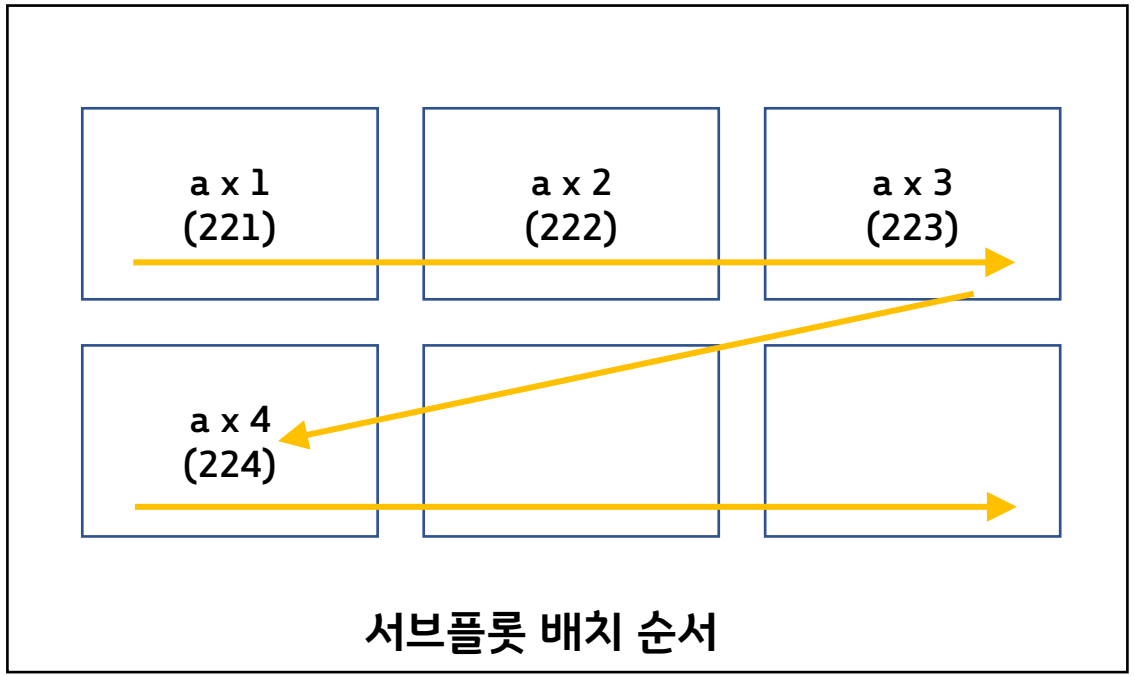
서브 플롯
(그래프 작성 영역)



Matplotlib 활용 기초

- **화면 분할 (Sub Plot)**

add_subplot() 메서드로 서브 플롯 배치 시



subplots() 메서드 사용 시
→ 피겨 생성, 서브 플롯 배치 동시 처리
→ 행렬처럼 접근