

2021 인공지능 소수전공

13~16차시: Matplotlib

2021.07.22 18:30~22:15

Seokhwan Yang

- 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 시각화는

- **스토리텔링**이다. 사람들은 눈으로 본 것을 더 빨리 체득하므로 **데이터 시각화는 목적이 있는 스토리텔링**이라고 할 수 있다.
- 데이터를 더 **이해하기 쉬운 형식으로** 조정하고 **추세와 이상 값을 강조함**으로써 스토리텔링을 돕는다.

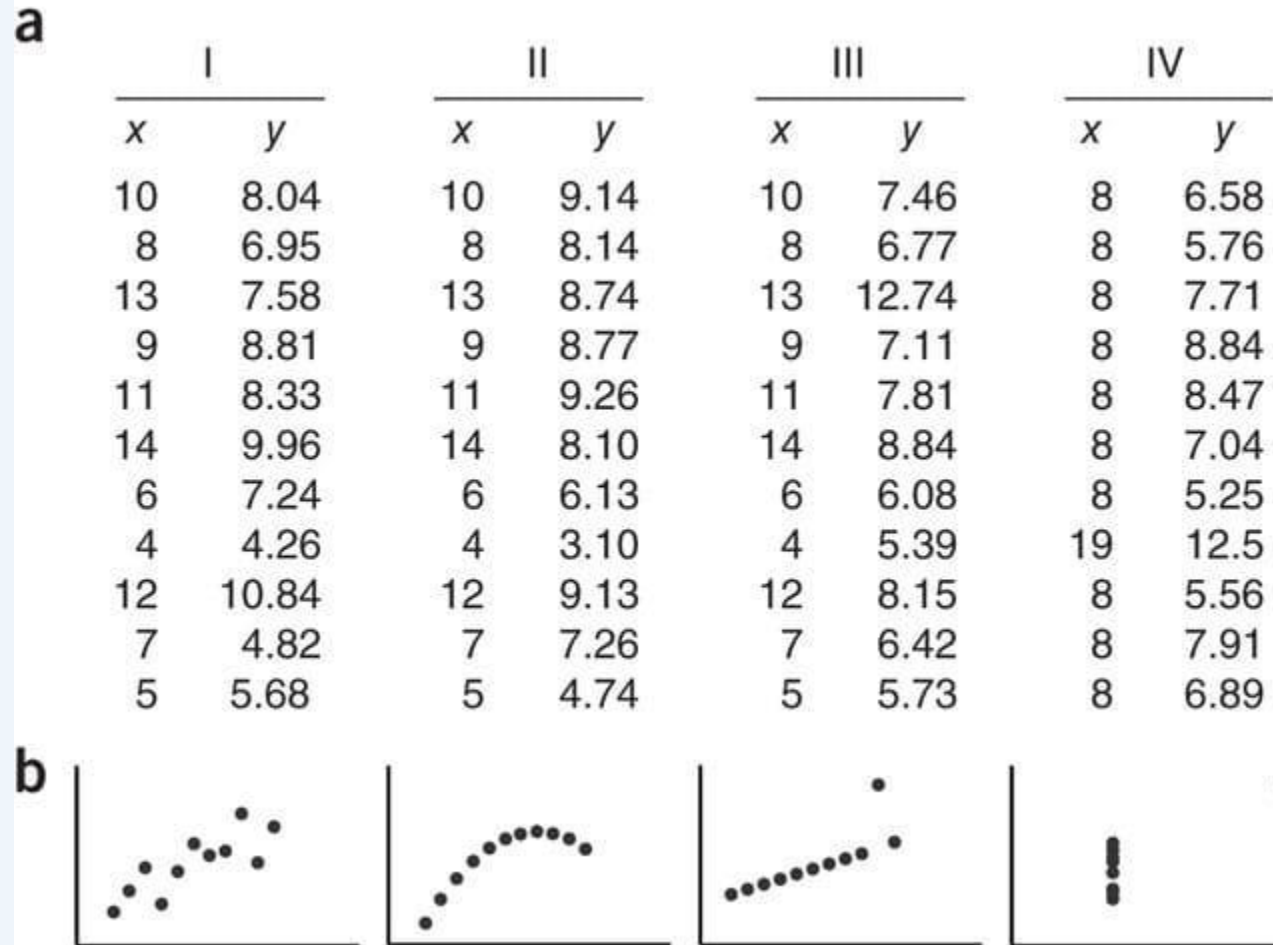
- 모든 직업 영역에서 데이터 시각화가 중요한 이유
 - STEM 분야(과학, 기술, 공학, 수학)는 데이터를 이해함으로써 이익을 얻는다.
 - 정부 분야, 재무, 마케팅, 역사, 소비재, 서비스 산업, 교육, 스포츠 등에서도 데이터의 중요성은 갈수록 높아지고 있다.
 - 데이터가 가진 의미를 시각적으로 더 잘 전달할 수 있다면 그 정보를 보다 더 효과적으로 활용할 수 있다.
 - 현대 직업 세계에서는 창조적인 스토리텔링과 기술 분석의 영역을 아우르는 것이 중요하며 데이터 시각화는 분석과 스토리텔링을 이어주는 위치에 있다.

- 데이터 시각화의 필요성

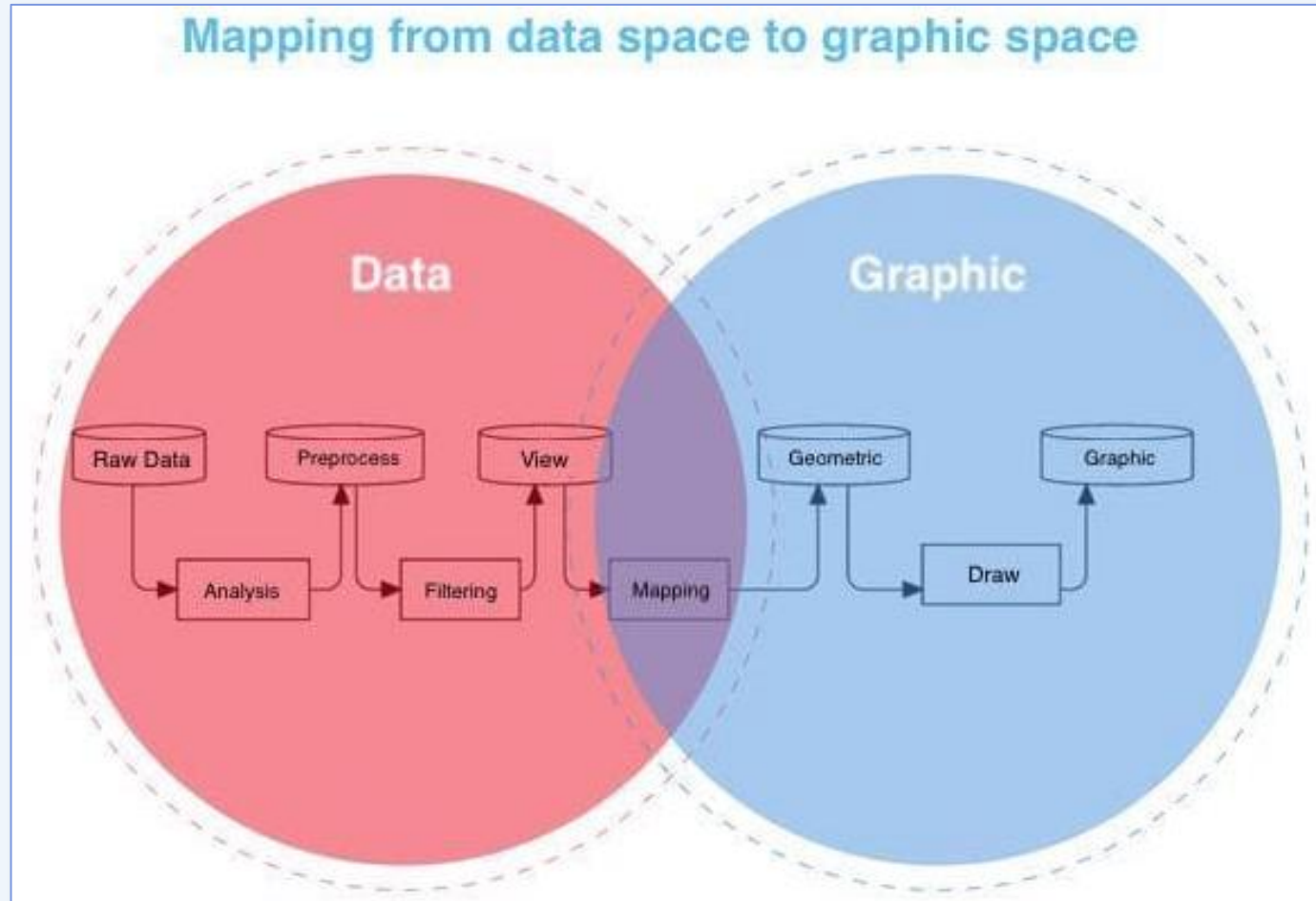
- 인간은 시력을 통해 얻는 정보량은 다른 기관의 정보보다 훨씬 많음
- 지나치게 많은 데이터로 인해 이를 관리하고 이해하는 어려움이 계속해서 증가
- 대부분의 사람들은 통계 데이터에 대해 잘 알지 못하며, 기본적인 통계 방법(평균, 중위수, 범위 등)은 인간의 인지적 성격과 맞지 않음
- 통계 방법에 따라 규칙을 보는 것은 어렵지만, 데이터가 시각화되면 규칙은 매우 명확히 인지 가능(예: 안스콤비의 4종주)

데이터 시각화(Data Visualization)

- 안스콤비의 4중주(Anscombe's quartet)



- 데이터 시각화는 데이터 공간에서 그래픽 공간으로의 매핑이다



전형적인 시각적 구현 절차

1. 데이터를 처리하고 필터링
2. 표현 가능한 시각적 형태로 변환
3. 사용자가 볼 수 있는 보기로 렌더링

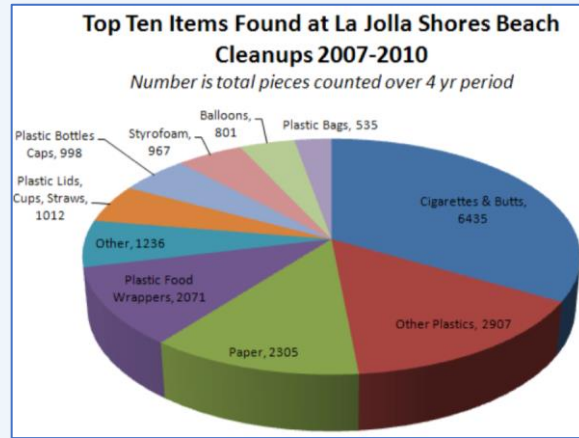
- 데이터 시각화에서 요구되는 기술

- 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

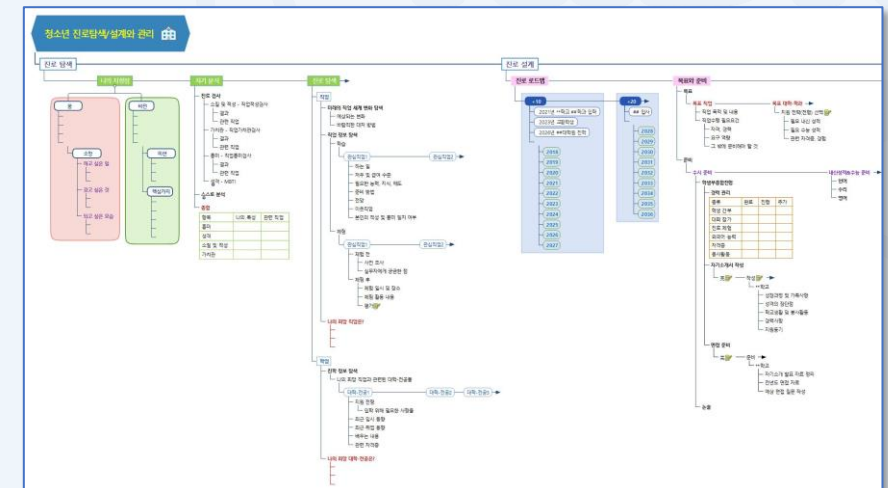
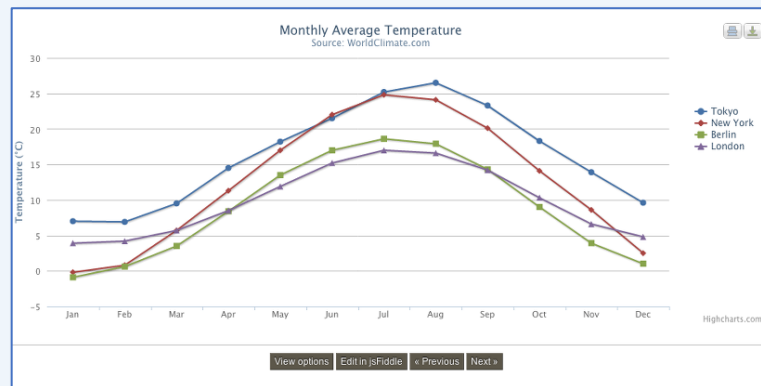
데이터 시각화의 유형

• 널리 사용되는 데이터 시각화의 일반적인 유형

- 차트
- 테이블
- 그래프
- 맵
- 인포그래픽
- 대시보드

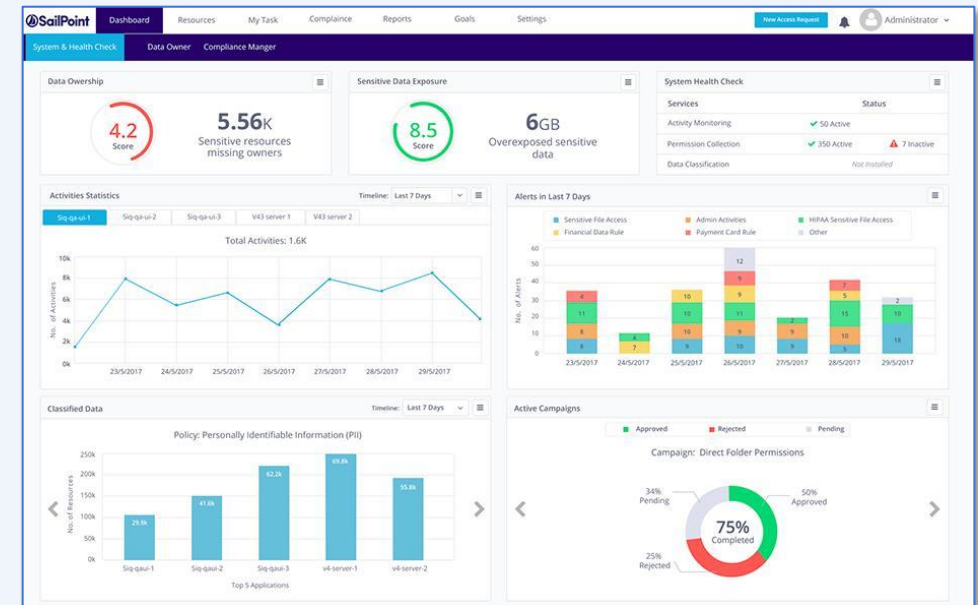
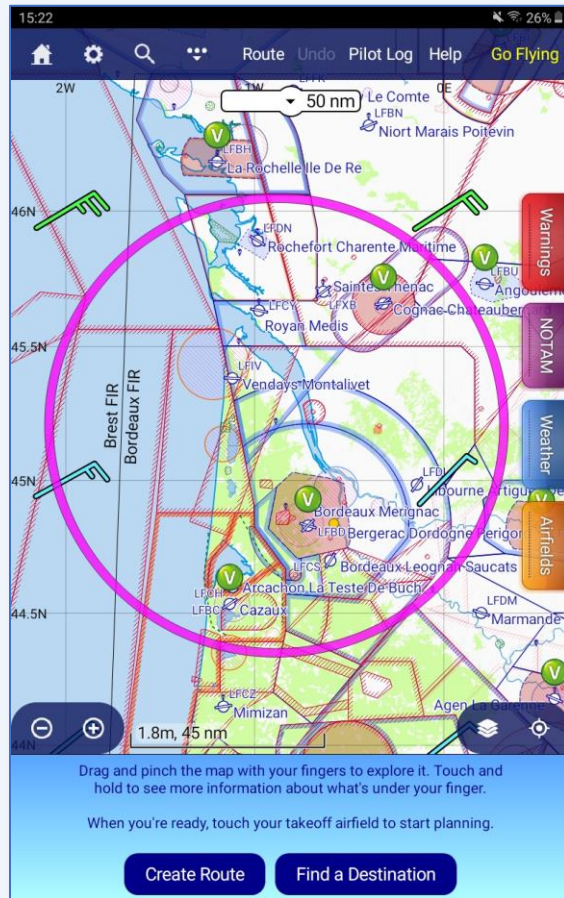


	A	B	C	D	E	F	G	H	I
1	휴대폰 제조업체 매출현황								
2									
3									단위:백만원
4	브랜드명	제조업체	생산지	최신모델	2007년	2008년	2009년	2010년	평균매출액
5	스카이	KS	미국	IM-7100	20,000	24,000	30,000	34,600	27,150
6	호르미	NIC	일본	MS-150	16,000	18,600	20,000	24,000	19,650
7	멀티규	NIC	일본	SCP-A011	12,000	16,000	19,000	18,500	16,375
8	큐텔	GLS	한국	S2	18,600	23,500	26,400	28,800	24,325
9	레디안	GLS	한국	SD2100	21,000	30,000	32,000	41,000	31,000
10	애드를	SAMS	한국	E-170	35,000	42,000	56,000	66,400	49,850
11	클맨	SAMS	한국	E-2500	-	52,000	26,000	28,400	26,600
12	스카이	KS	한국	IM-8100	-	24,000	26,000	34,000	21,000
13	호르미	NIC	일본	SCP-A012	12,000	16,000	19,000	18,500	16,375
14	큐텔	GLS	한국	S3	23,500	32,400	26,400	23,400	26,425
15	레디안	GLS	한국	SD2101	32,100	21,000	32,000	32,000	29,275
16	애드를	SAMS	한국	E-171	43,000	45,200	32,100	25,000	36,325
17	클맨	SAMS	미국	E-2600	-	32,600	26,000	15,000	18,400
18									



데이터 시각화의 유형

- 널리 사용되는 데이터 시각화의 일반적인 유형



• 데이터 시각화의 구체적인 예

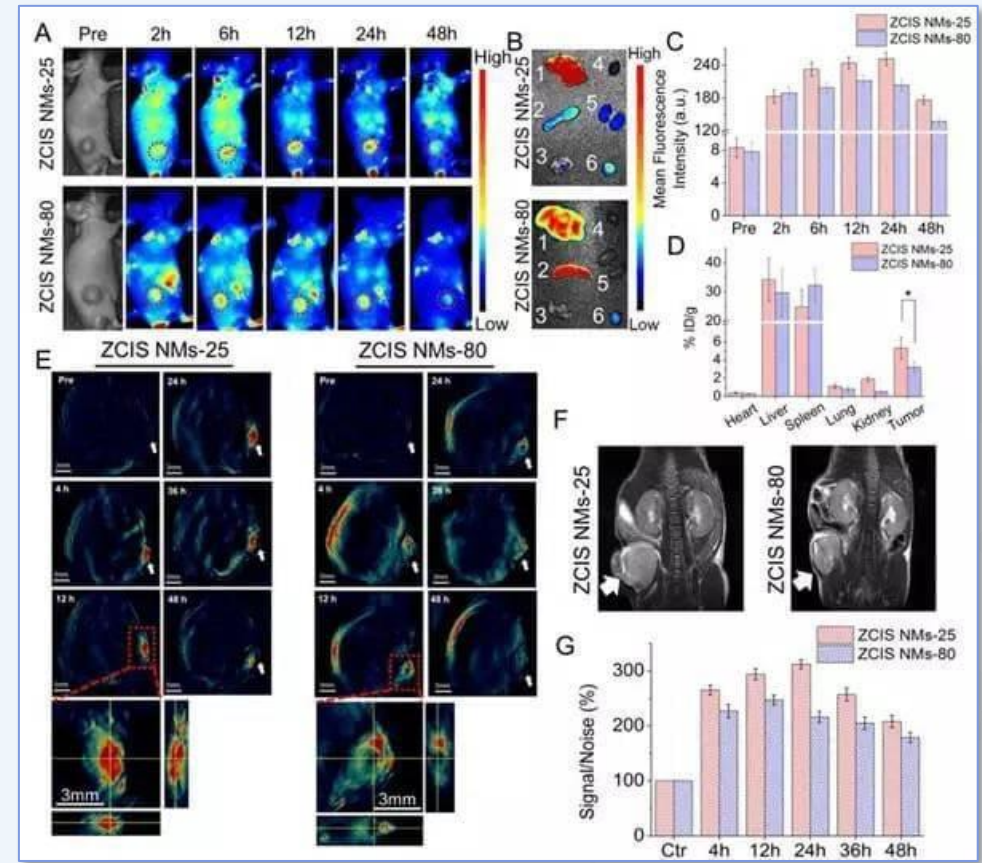
- 영역 차트
- 막대 차트
- 상자-수염 차트
- 버블 클라우드
- 불릿 그래프
- 카토그램
- 원 뷰
- 점 분포 맵
- 간트 차트
- 히트 맵
- 하이라이트 테이블
- 히스토그램
- 행렬
- 네트워크
- 극좌표형 영역(Polar Area)
- 방사형 트리
- 분산형 차트(2D / 3D)
- 스트림 그래프
- 텍스트 테이블
- 타임라인
- 트리 맵
- 썰기형 누적 그래프
(Wedge Stack Graph)
- 워드 클라우드
- 대시보드를 통한 모든 유형의 조합 등

- 데이터 시각화의 3가지 분류
 - 과학적 시각화
 - 정보 시각화
 - 시각화 분석



- 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



- 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

- 시각적 분석 (Visual Analytics)

- 과학적 시각화와 정보 시각화의 발전과 함께 진화한 새로운 분야
- 대화형 시각화 인터페이스를 통한 분석 추론을 강조



- **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

- 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



- 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()`

• 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

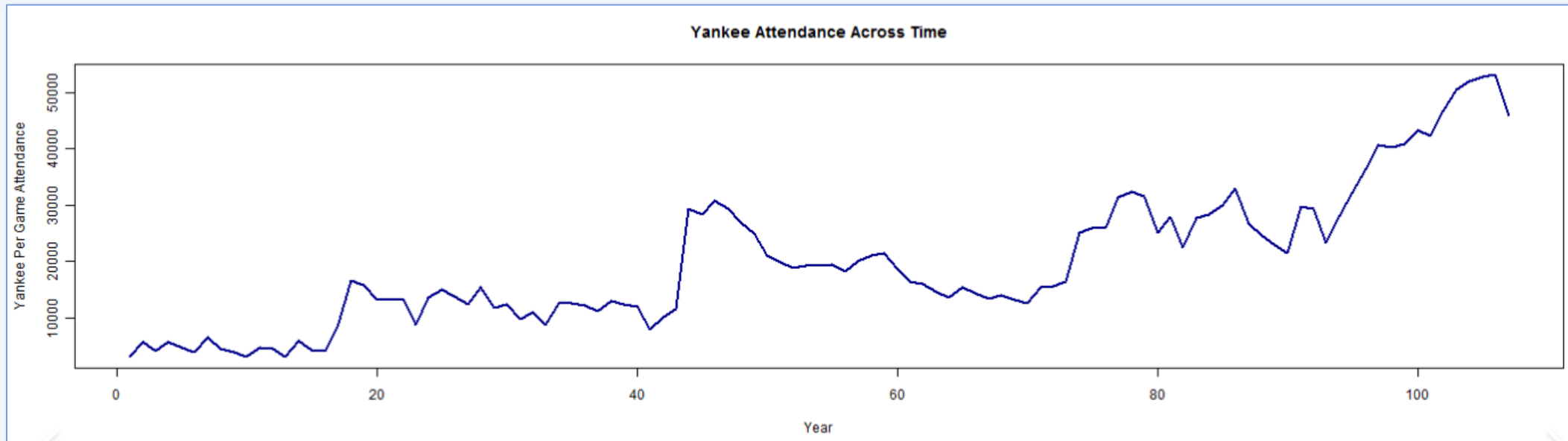
Color

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

- 선 그래프 (Line Plot)



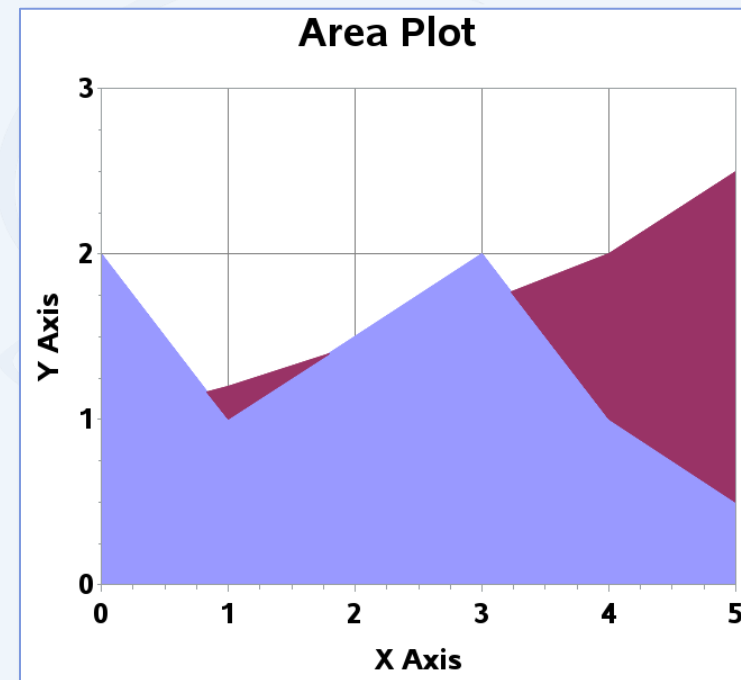
- 면적 그래프(Area Plot)

- 선 그래프를 확장한 개념

- 각 열의 패턴과 함께 열 전체의 합계가 어떻게 변하는지 파악할 수 있음

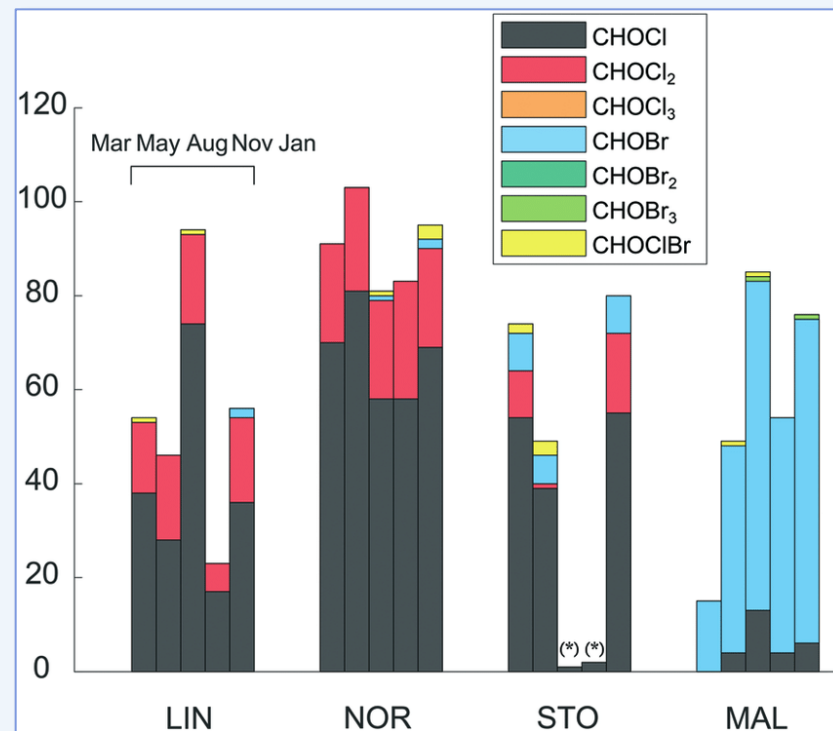
- 기본 사용법

- 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)



• 막대 그래프 (Bar Plot)

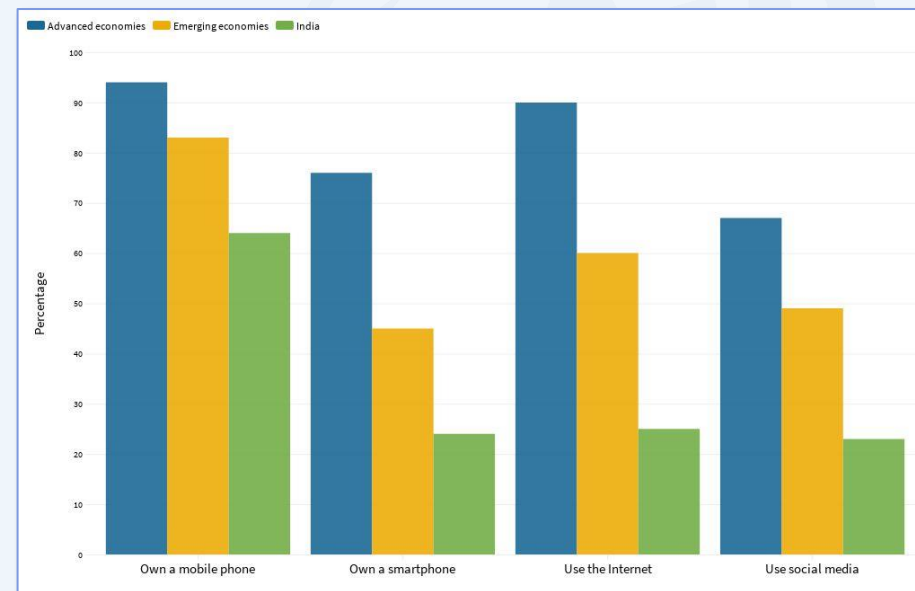
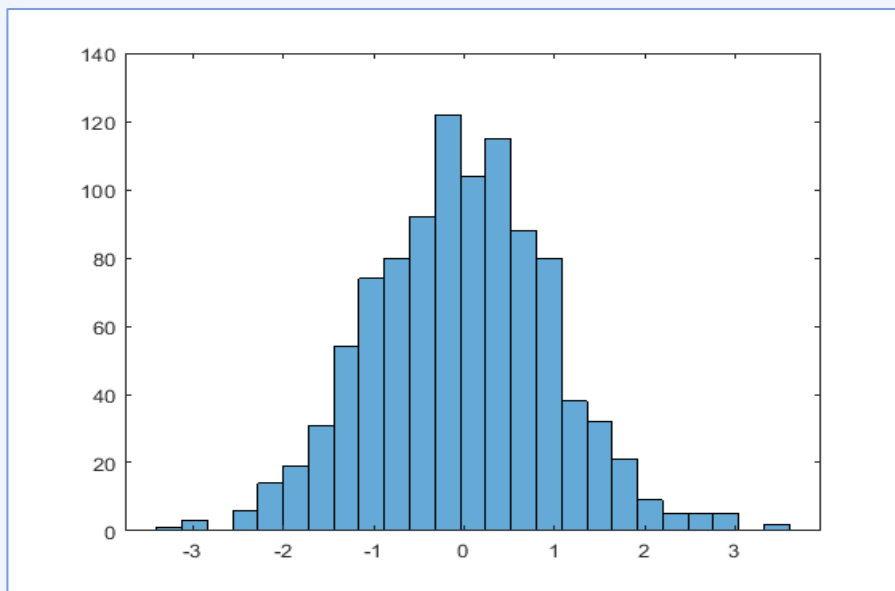
- 데이터 값의 크기에 비례하여 높이를 가지는 직사각형 막대로 표현
- 세로형 막대 그래프는 시계열 데이터를 표현하는데 적합
- 가로형 막대 그래프는 각 변수 사이의 값의 크기 차이를 설명하는데 적합



- 히스토그램 (Histogram)

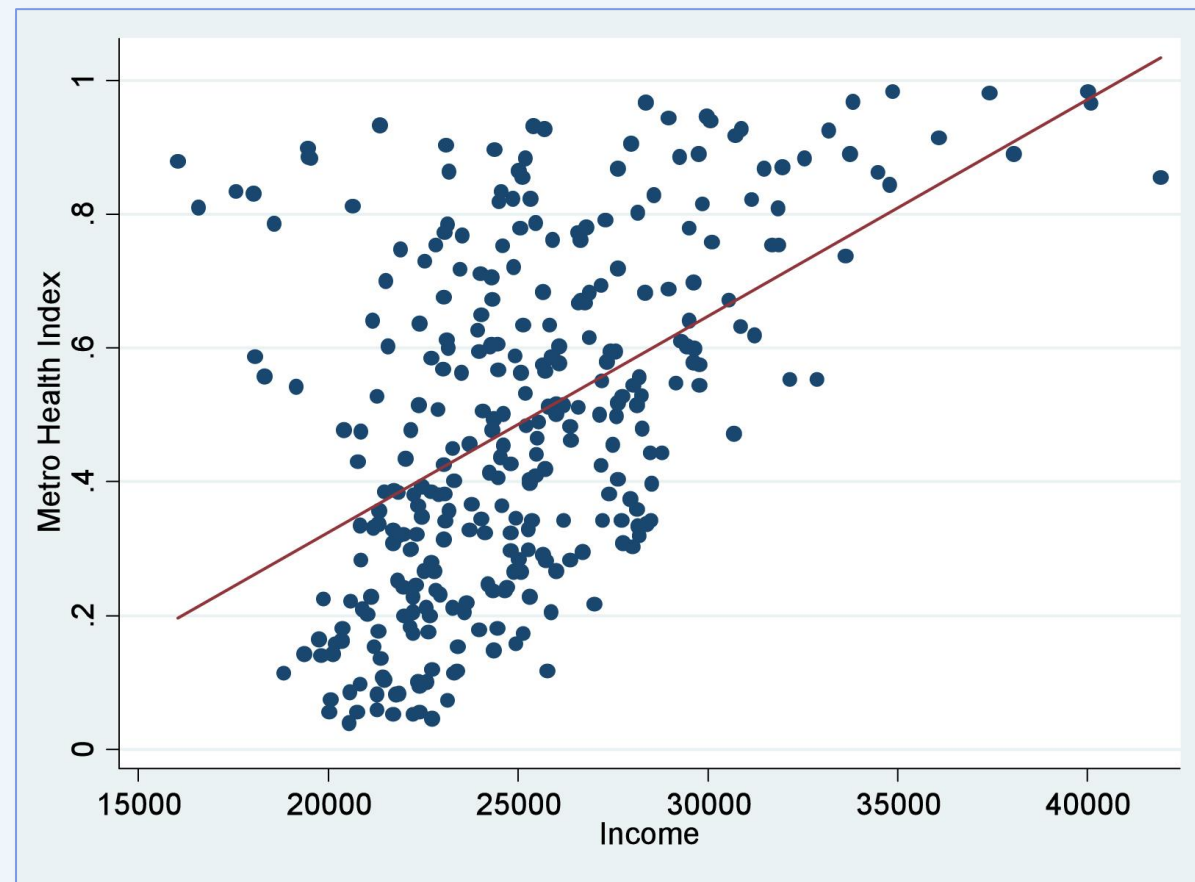
- 변수가 하나인 단변수 데이터에 대한 빈도수를 표현

- x축: 같은 크기의 여러 구간, 계급 구간
 - y축: 각 구간에 속하는 데이터 값의 개수(빈도)



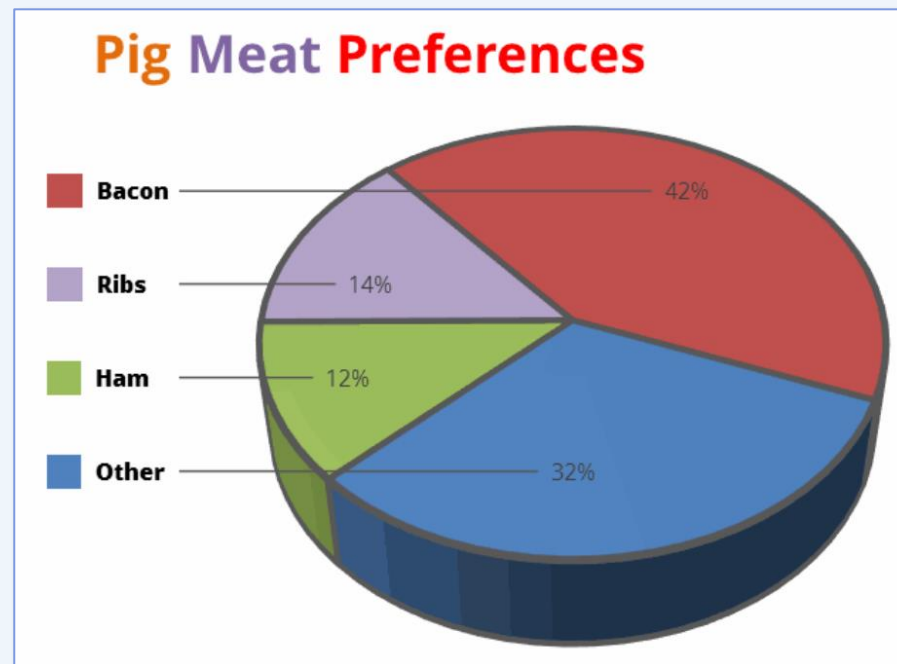
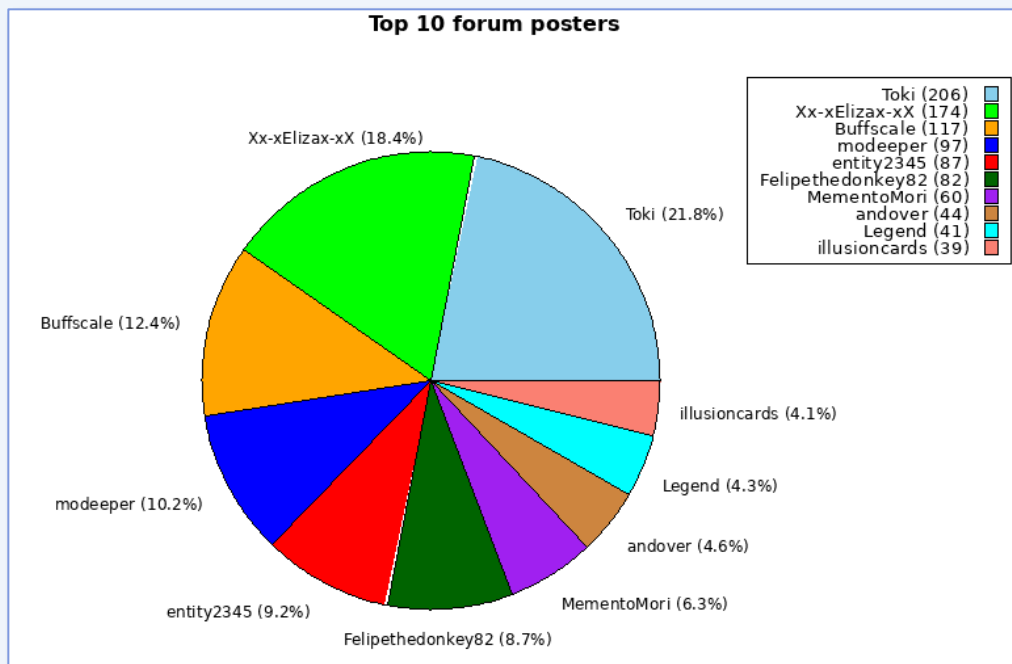
- 산점도 (Scatter Plot)

- 분산 그래프
- 서로 다른 두 변수 사이의 관계를 나타냄



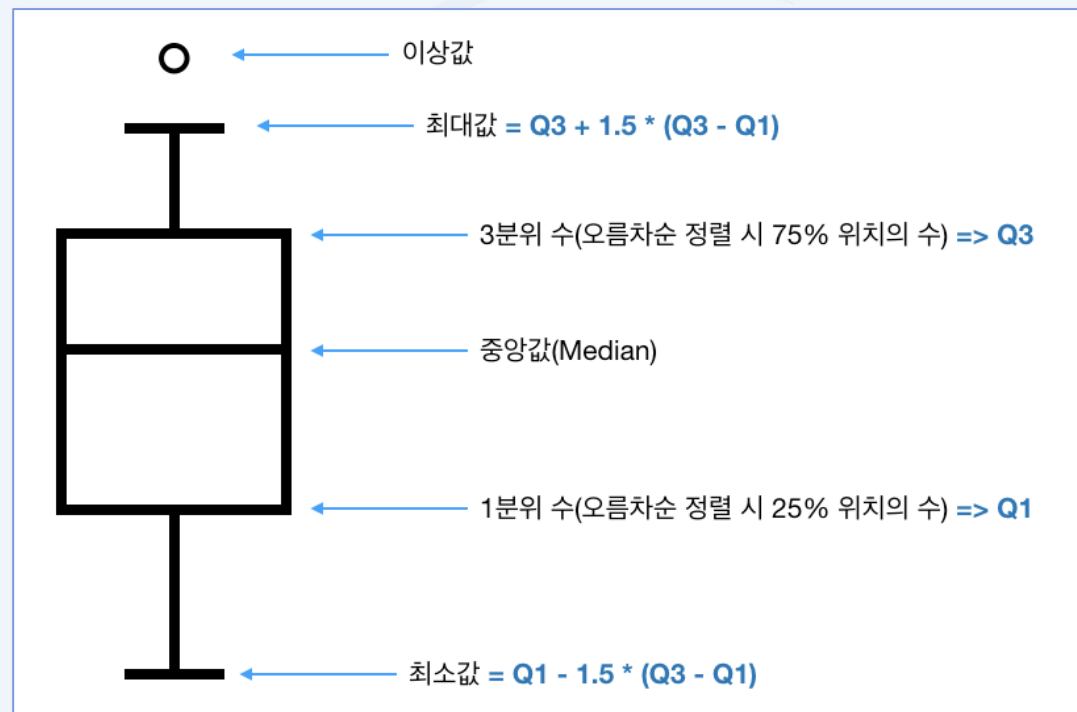
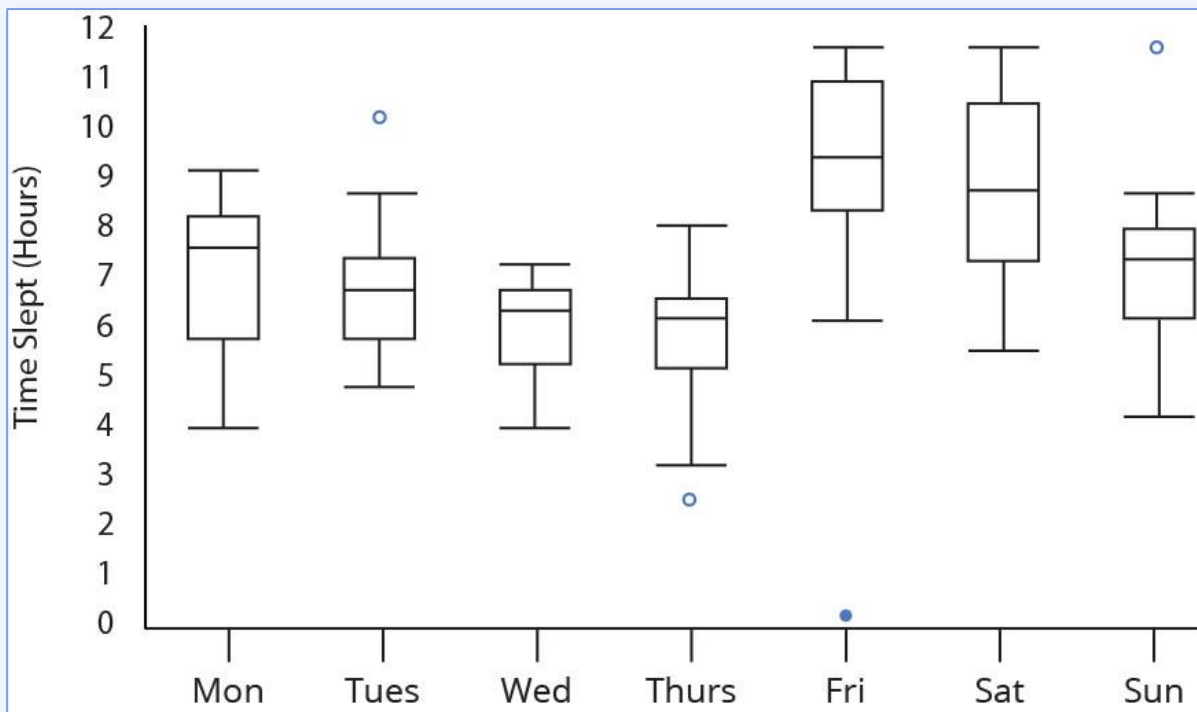
- 파이 차트 (Pie Chart)

- 원을 파이 조각처럼 나누어서 표현



• 박스 플롯 (Box Plot)

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



- 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')
```

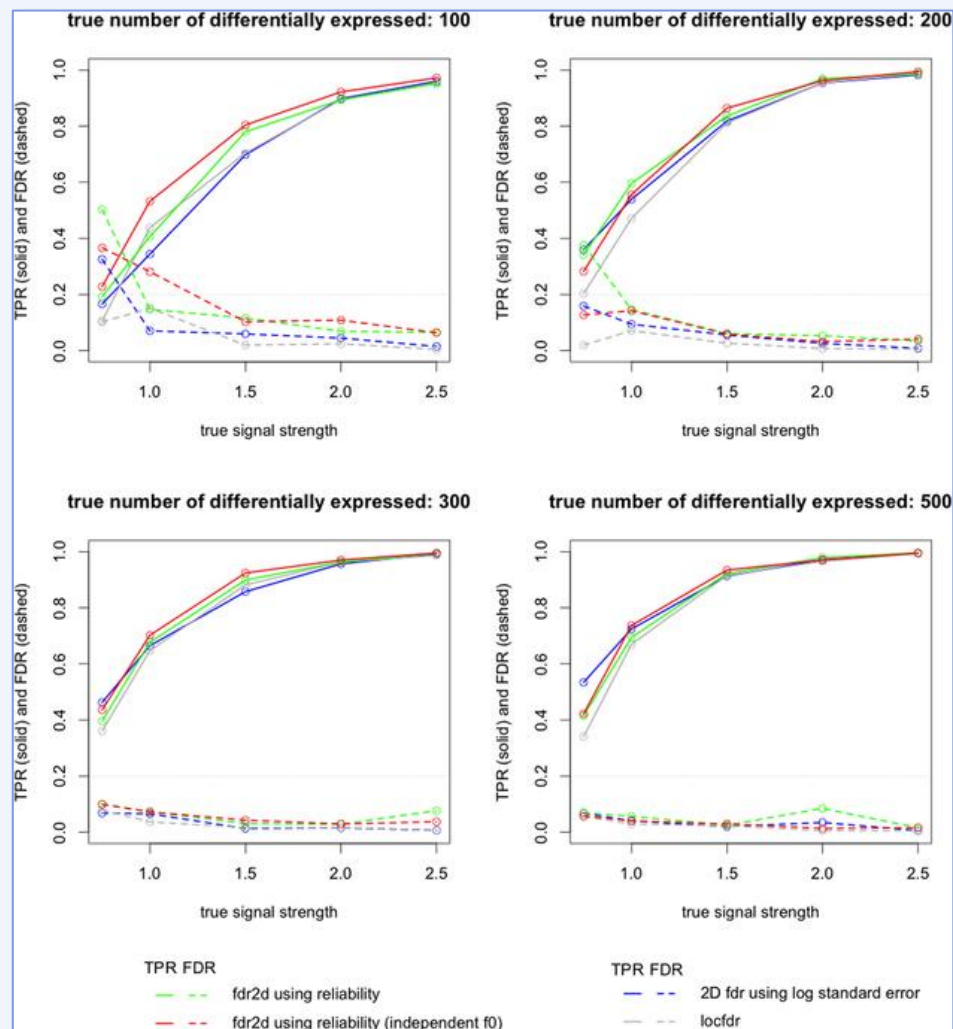
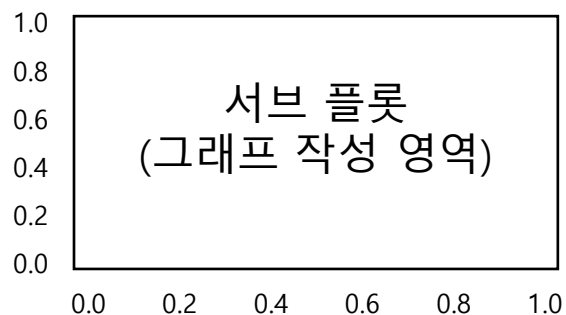
```
plt.imshow(img1)
```

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

- 화면 분할 (Sub Plot)

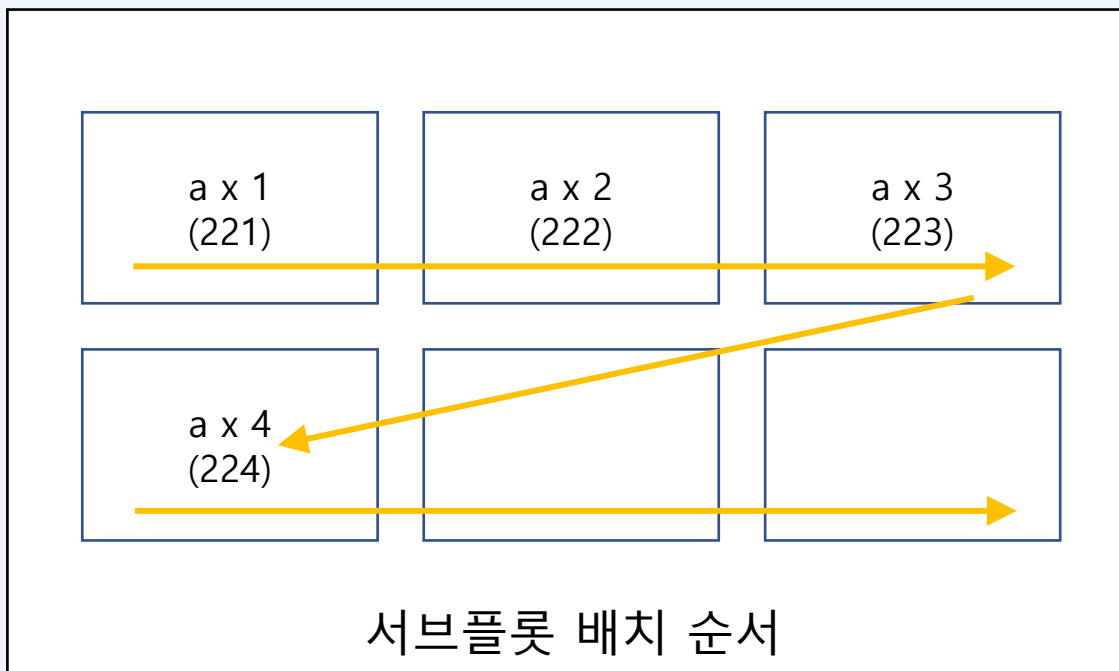
- 여러 개의 그래프를 한 화면에 표시하기 위하여 화면을 특정 영역으로 분할하여 각 그래프를 배치, 표시하는 기능

피겨 (서브 플롯 작성 영역)



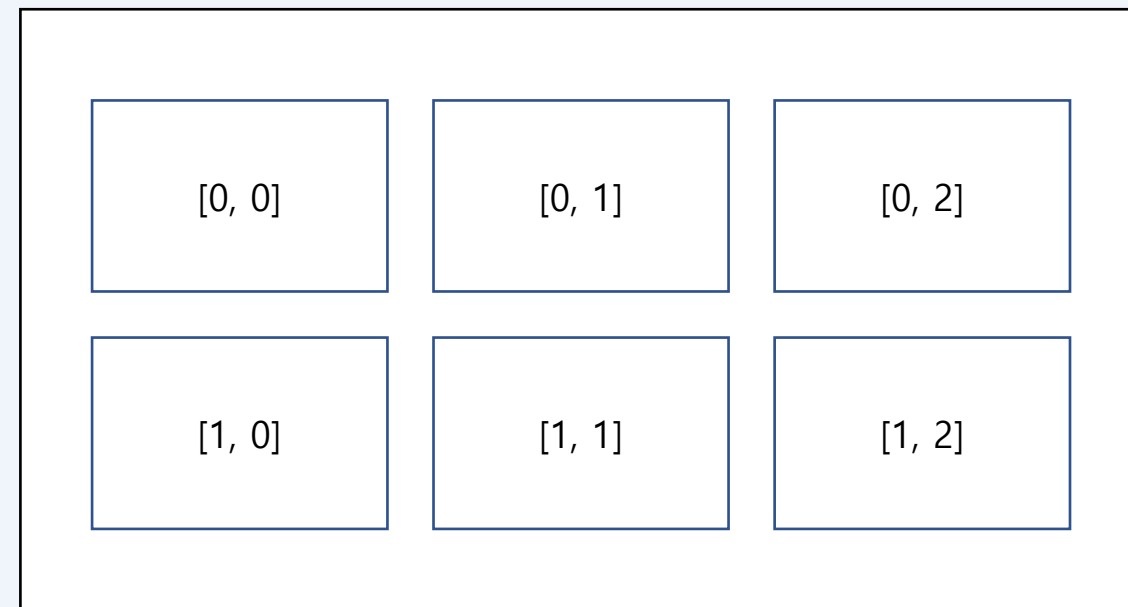
• 화면 분할 (Sub Plot)

add_subplot() 메서드로 서브 플롯 배치 시



subplots() 메서드 사용 시

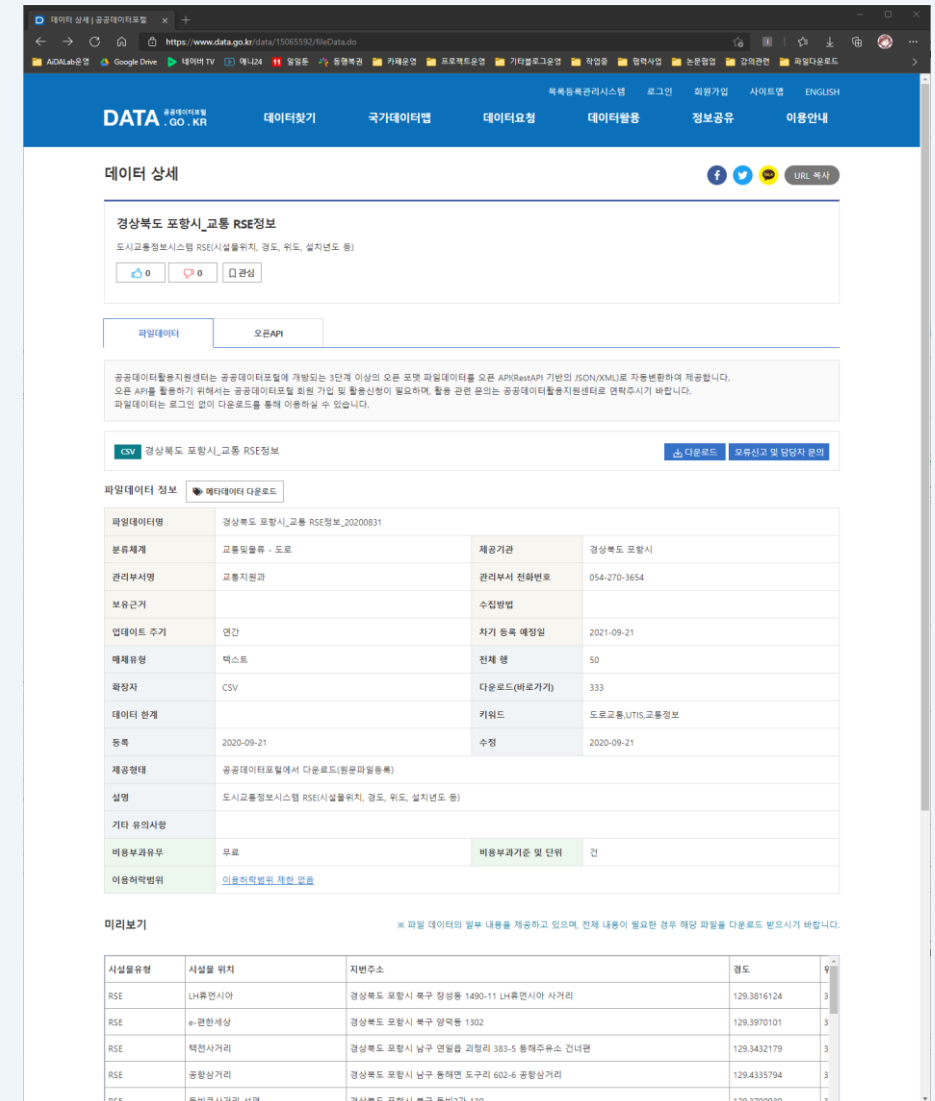
- 피겨 생성, 서브 플롯 배치 동시 처리
- 행렬처럼 접근



- 공공 데이터 포털

- <https://www.data.go.kr/index.do>

- 공공기관이 생성 또는 취득하여 관리하고 있는 공공데이터를 한 곳에서 제공하는 통합 창구



The screenshot displays the '데이터 상세' (Data Detail) page on the Data.go.kr portal. The page title is '경상북도 포항시_교통 RSE정보' (Gyeongsangbuk-do Pohang-si Transportation RSE Information). Below the title, there are buttons for '다운로드' (Download) and '관심' (Interest). The page is divided into two tabs: '파일데이터' (File Data) and '오픈API' (Open API). The '파일데이터' tab is selected, showing a table with details about the data file. The table includes columns for '파일데이터명' (File Data Name), '분류체계' (Classification System), '관리부서명' (Managing Department Name), '보유근거' (Basis for Possession), '업데이트 주기' (Update Cycle), '데이터 유형' (Data Type), '확장자' (Extension), '데이터 연계' (Data Linkage), '등록' (Registration), '제공형태' (Provision Form), '설명' (Description), '기타 유의사항' (Other Notes), '비밀유지유무' (Whether to Maintain Confidentiality), and '이용허락범위' (Scope of Use Permission). The table shows that the data is a CSV file, managed by the Pohang City Transportation RSE Information, and is available for download. Below the table, there is a '미리보기' (Preview) section showing a sample of the data, including columns for '시설물유형' (Facility Type), '시설물 위치' (Facility Location), '지번주소' (Address), '경도' (Longitude), and '위도' (Latitude).

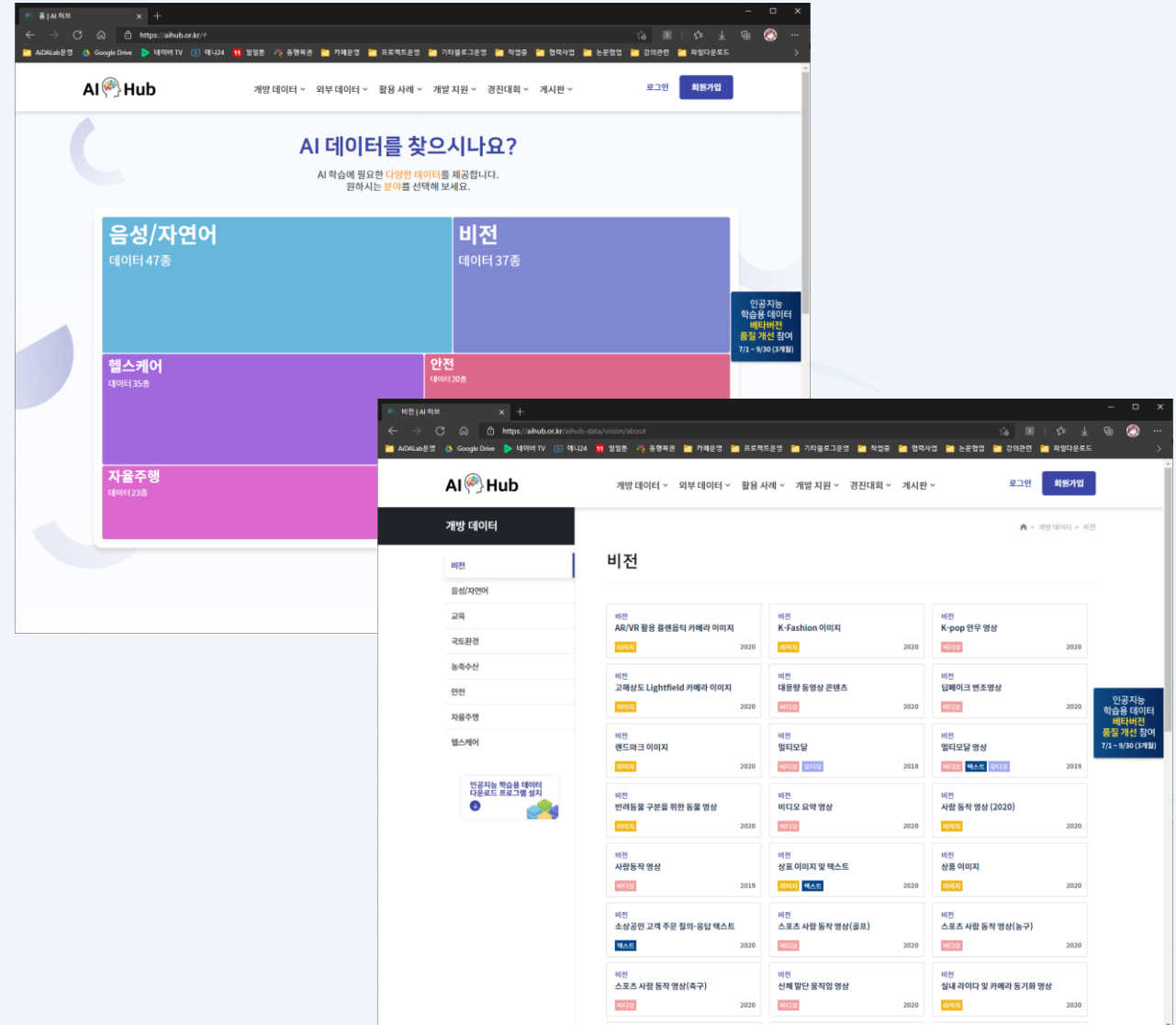
파일데이터명	분류체계	관리부서명	보유근거	업데이트 주기	데이터 유형	확장자	데이터 연계	등록	제공형태	설명	기타 유의사항	비밀유지유무	이용허락범위
경상북도 포항시_교통 RSE정보_20200831	교통정보 - 도로	교통정보과	수집방법	연간	텍스트	CSV	경상북도 포항시	2020-09-21	공공데이터포털에서 다운로드(원문파일제공)	도시교통정보시스템 RSE(시설물위치, 경로, 위도, 설치년도 등)		무로	이용허락범위, 검색, 열람

시설물유형	시설물 위치	지번주소	경도	위도
RSE	LH휴먼시아	경상북도 포항시 북구 장성동 1490-11 LH휴먼시아 사거리	129.3816124	3
RSE	e-편한세상	경상북도 포항시 북구 양곡동 1302	129.3970101	3
RSE	택전사거리	경상북도 포항시 남구 연일읍 괴정리 383-5 통해주유소 건너편	129.3432179	3
RSE	공항사거리	경상북도 포항시 남구 흥해면 도구리 602-6 공항사거리	129.4335794	3
RSE	충빈큰사거리, 선관	경상북도 포항시 북구 흥해면 139	129.3700939	3

- AI Hub

- <https://aihub.or.kr/>

- AI 기술 및 제품·서비스 개발에 필요한 AI 인프라(AI 데이터, AI SW API, 컴퓨팅 자원)를 지원함으로써 누구나 활용하고 참여하는 AI 통합 플랫폼



- 그 외, 오픈 데이터를 얻을 수 있는 곳
 - ETRI 공공 인공지능 오픈 API, Data (<https://aiopen.etri.re.kr/>)
 - Kaggle (<https://www.kaggle.com/>)
 - Tensorflow Datasets: a collection of read-to-use datasets (<https://tensorflow.google.cn/datasets>)
 - COCO (Common Objects in Context) (<https://cocodataset.org/>)
 - 다양한 기관에서 데이터 분석, AI 학습을 위한 오픈 데이터를 제공함