Types of Data

- Nomial Data(명목척도): 순서 X ex) 성별(남자, 여자)/인종(한국인, 미국인, 중국인)
- Ordial Data(서열척도): 순서가 있지만 간격이 같지 않음 ex) 만족도(매우 만족, 만족, 중립, 불만족, 매우 불만족)/사회적지위(높음, 중간, 낮음)
- Interval Data(동간척도): 순서가 있고 간격이 같음, true zero point 없음 ex) 온도-0도라고 해서 온도가 존재하지 않는 것은 아님
- Ratio Data(비율척도): 동간척도와 같으나 true zero point 있음 ex) 월급-0원은 월급이 실제로 없는 것

Multivariate Data의 종류

- Survey Data ex) 성별, 반려동물 선호도, 만족도 조사 데이터
- Health Data ex) 나이, 성별, 혈압, 질병 데이터
- Financial Data ex) 기업의 수익과 산업분야 데이터
- Environmental Data ex) 지역에 따른 월별평균온도, 강수량, 초목

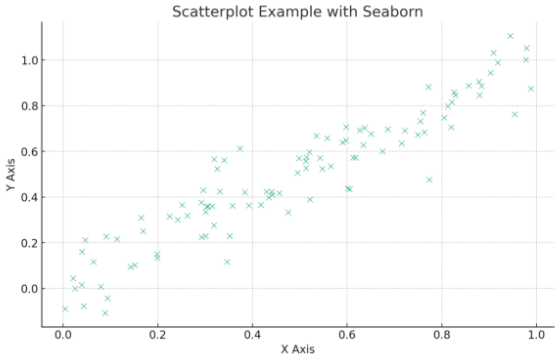
Visualization 하기

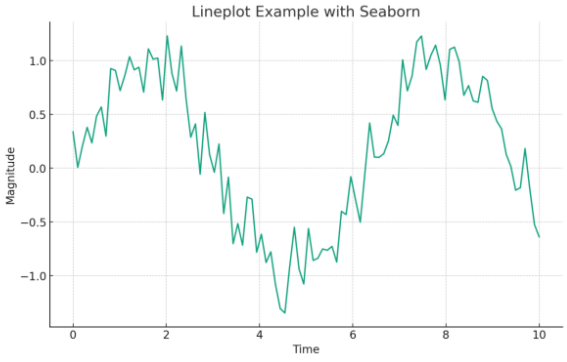
관련 파이썬 라이브러리

- matplotlib: 가장 기본적인 시각화 라이브러리
- seaborn: matplotlib보다 더 고수준의 인터페이스를 제공
- plotly: 고급 시각화를 위한 라이브러리
- pandas: 데이터 프레임이라는 표 형식의 데이터 구조를 제공
- numpy: 배열, 행렬 및 수학 함수 계산을 위한 기본적인 패키지

Seborn의 함수

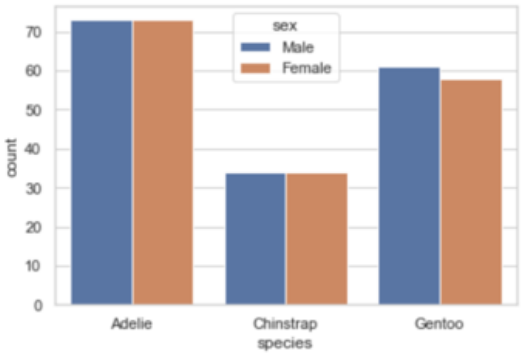
- **Relational Plots: 변수 사이의 관계를 나타냄**
    - scatterplot(), lineplot()

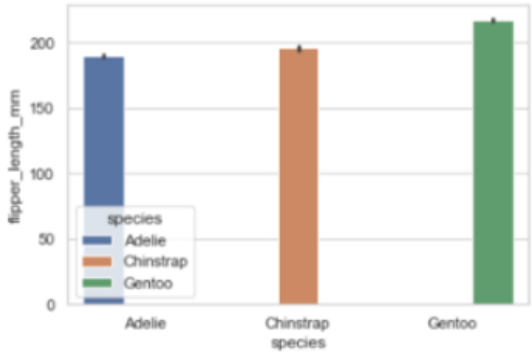


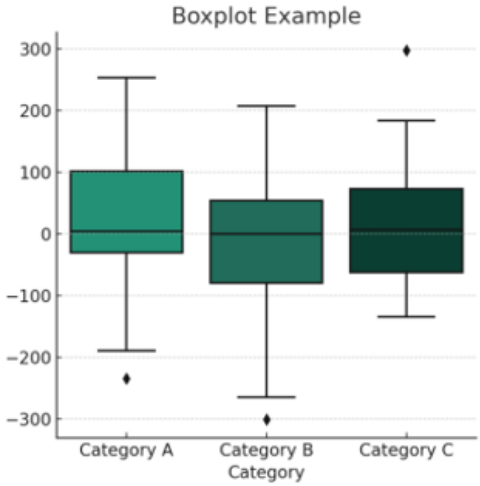


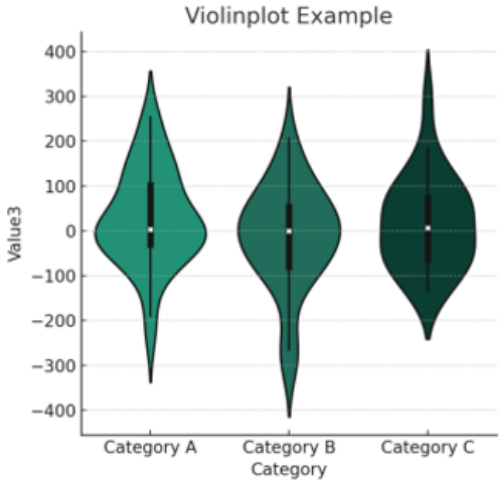
- **Categorial Plots: 범주를 나타냄**
    - catplot(): figure-level 함수로, 여러 개의 그래프를 생성 및 관리(axes-level 함수는 단일그래프)
        - countplot(): 일반적인 막대그래프. 데이터의 수치값을 찍어줌
        - barplot(): 범주와 연속형 변수의 관계를 볼 때 사용. 평균값을 범주별로 찍어주고, error bar를 이용해 편차를 나타냄

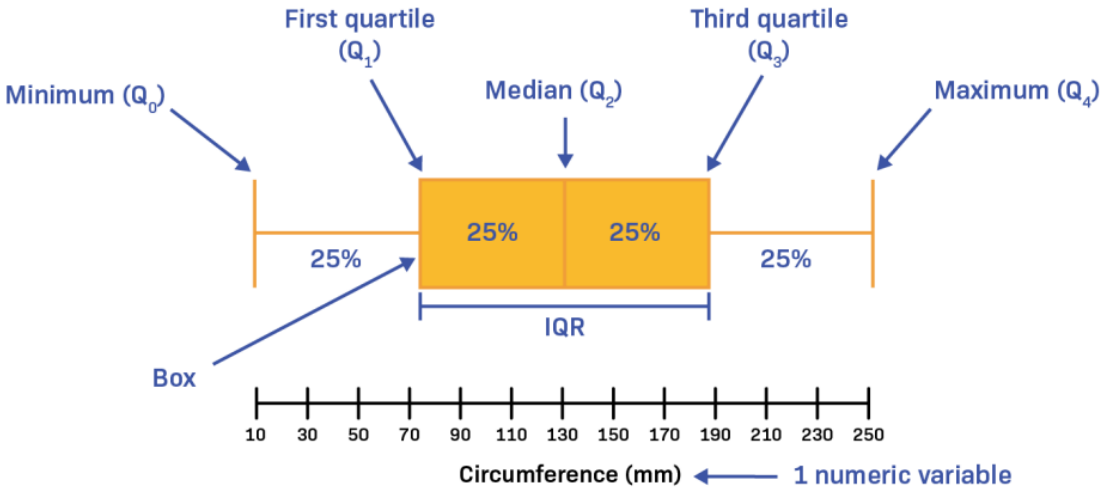
- box plot(): 범주에 해당하는 연속형 변수의 분포를 중앙값(중앙선), 사분위수(상자), 이상치(점)를 포함하여 상자 그림으로 나타냄
- violin plot(): 데이터의 분포를 커널 밀도 추정(KDE, 확률밀도함수)으로 표현하며, 상자 그림보다 밀도를 잘 알려주는 대신 기준선이 없어서 한 눈에 보기는 힘듦

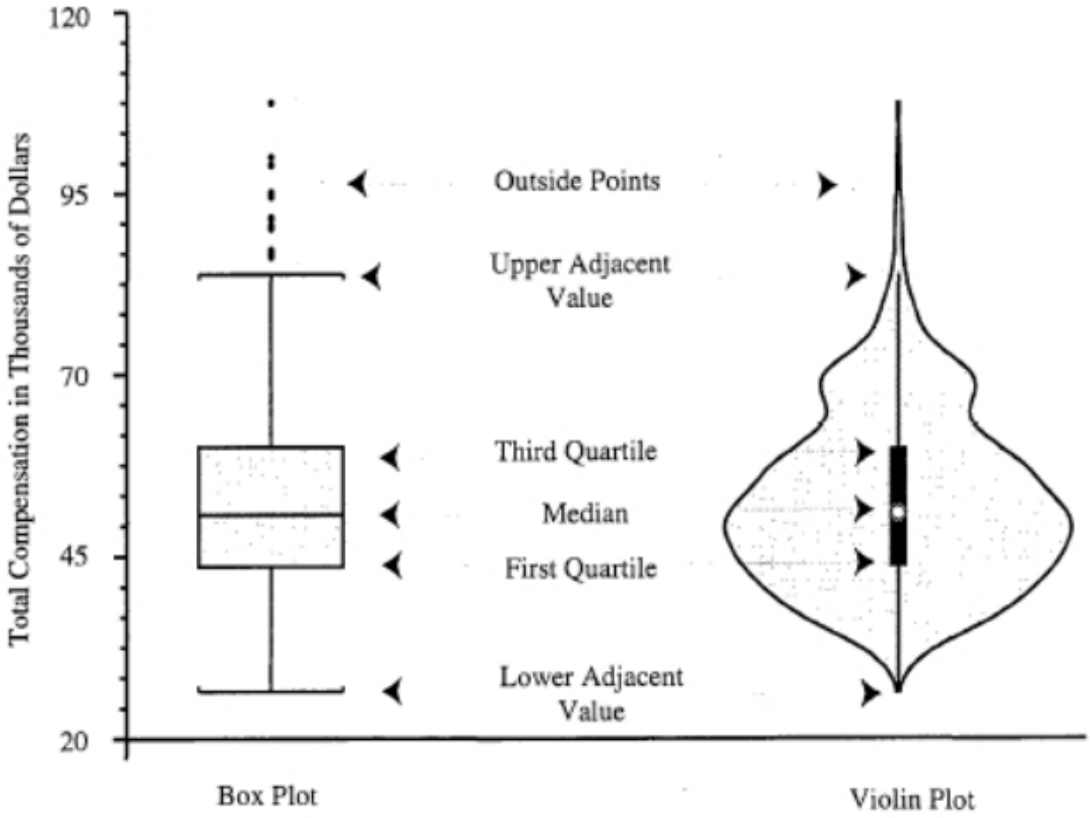






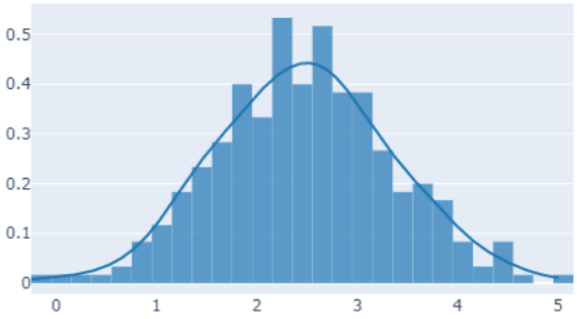




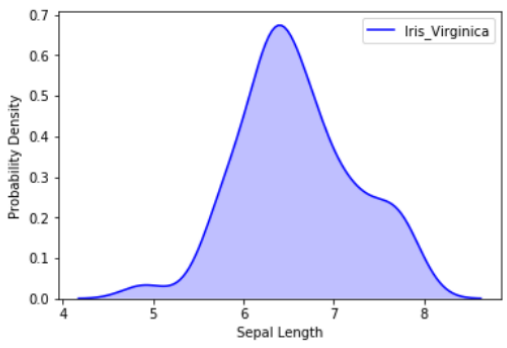


- **Distribution Plots: 분포를 나타냄**
  - distplot()/histplot(): 단변량데이터의 분포를 히스토그램으로 나타내고 KDE를 그려줌
  - kdeplot(): 단변량데이터의 KDE를 그려줌
  - jointplot(): 산점도를 기본으로 표시하고 x, y축에 각 변수에 대한 히스토그램을 동시에 보여준다.

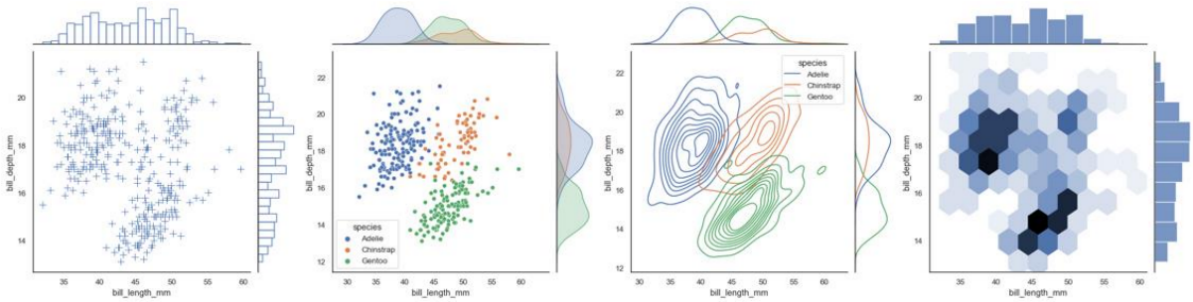
distplot()/histplot()



Kdeplot()

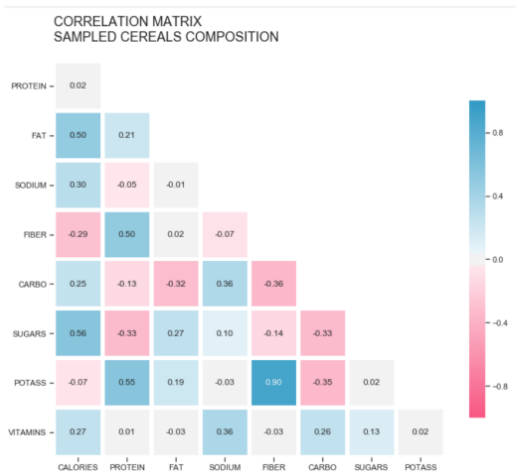


jointplot()

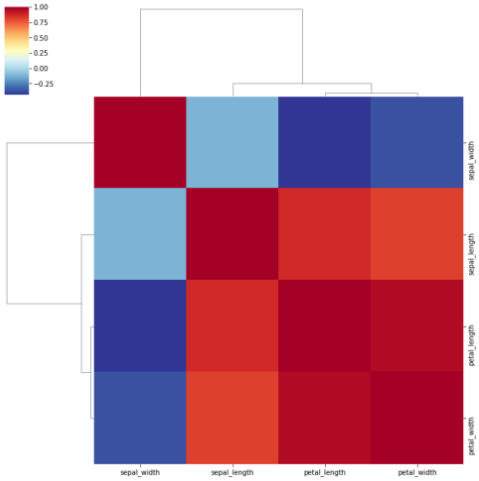


- **Matrix Plots: 데이터를 행렬 형태로 표현**
  - heatmap(): 행과 열로 구성된 데이터의 값을 색상으로 나타내 상관관계를 보여줌
  - clustermap(): 데이터의 상관 관계에 따라 행렬의 행과 열을 클러스터링하고, 히트맵 형태로 시각화

heatmap()

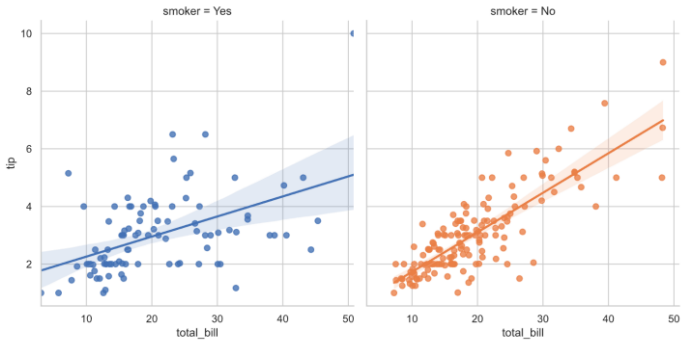


clustermap()

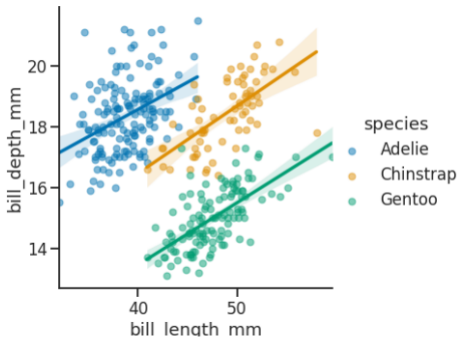


- **Regression Plots: 변수 간 '선형'관계를 보여줌**
  - regplot(): 산점도와 함께 선형 회귀 직선을 그려줌
  - lmplot(): regplot()의 고수준 버전, hue로 구분짓기 가능

regplot()



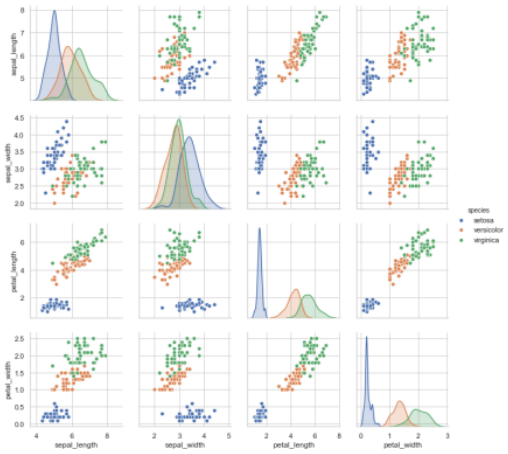
lmplot()



- **Multi-Plot Grids: 데이터의 서로 다른 부분집합을 비교하는 데 유용**

- pairplot(): 인자로 전달되는 데이터프레임의 열(변수)을 두 개씩 짝 지을 수 있는 모든 조합에 대해 관계 그래프를 그림 -> 3개의 열이라면 3행 x 3열의 크기로 9개 생성. 같은 변수끼리 짝을 이루는 대각선 방향으로는 히스토그램을 그리고, 서로 다른 변수 간에는 산점도를 그린다.
- FacetGrid, PairGrid, JointGrid: 대각선을 기준으로 각각 다른 시각화 툴을 적용 가능하도록 pairplot()을 개선한 클래스

pairplot()



pairgrid()