2021-03-18

[데이터 시각화(Visualization)]

{1. 시각화 }

* 시각화의 목적은 데이터를 탐색하거나 마이닝 결과를 전달하는데 있다.

<주요 시각화 방식>

* 히스토그램: 수치형 데이터의 분포를 확인하기에 유용. 데이터가 어디에 몰려 있는지 등을 알 수 있음.
* 막대/파이 차트: 수치가 아닌 카테고리와 같은 명목형 데이터를 확인할 때 유용. 단, 카테고리가 많으면 시각화에 방해가 되므로 Top N을 뽑아서 시각화 한다. 파이 차트는 파이의 크기가 잘 구분이 되지 않으므로 주의.
* 산점도/직선 그래프 (Scatter/Line plots)

두 변수의 관계를 확인할 때 유용. 예시를 보면 평수와 가격에 따른 두 변수의 관계를 나타내고 있음.

* 시계열 그래프 (Time series plot)

시간에 따라 변수가 변화하는 트렌드를 분석하기 위한 용도

* 관계 맵

복잡한 가정을 수립할 때 데이터의 관계를 시각화 하는 방법, 컬럼 별로 다른 색으로 표시 됨. 선의 두께는 두 컬럼 사이에 관계의 중요도를 나타냄. 연구 분야를 선정하거나, 약의 용도를 예측할 때 사용.

* 히트 맵

색으로 데이터의 빈도나 트렌드를 파악. 예시는 히트 맵을 이용하여 레벤슈타인 거리를 표현함. 변수 간의 상관관계를 확인할 때 사용.

* 지도

지역에 따른 변수의 변화를 파악할 수 있음.

* 단어 클라우드

단어의 빈도가 클수록 글자의 크기가 커짐.

* 3D 그래프, 고차원 그래프

고차원의 데이터는 차원을 축소해서 시각화함. 3D로 표현하면 좀 더 interactive 함.

{2. Matplotlib }

* Matplotlib 는 가장 많이 사용되고 있는 시각화 라이브러리임.

<용어>

* Figure: 그래프 전체 그림
* Axes: 그래프가 그려지는 좌표 평면이자 subplot의 단위
* X axis, Y axis: X축 Y축
* Tick: 눈금 (Major 눈금, Minor 눈금)
* Spines: 그래프 테두리
* Line: 선 그래프의 선
* Markers: 그래프 상의 점의 형태
* Grid: 그래프 격자
* Title: 그래프 제목
* Label: 각 축이나 눈금 등에 붙이는 문자
* Legend: 범례 (그래프를 구분하는 이름 또는 설명 목록)

<기본 그래프>

Ex)

Import matplotlib.pyplot as plt

Plt.plot([1,2,3,4])

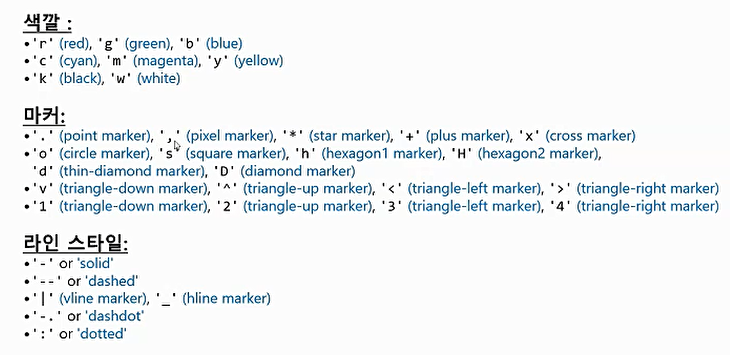
Plt.ylabel(‘some numbers’)

Plt.show()

* Plot의 매개변수로 한 가지 리스트만 주면 y 좌표로 인식하고, x 좌표는 자동적으로 [0,1,2,3…] 순으로 생성된다.
* Plot 함수의 리스트가 2개면 x값과 y값으로 간주

<스타일 추가>

* .plot 안의 매개변수로 전달해주면 된다.
* 색깔은 꼭 밑의 예시에 있는 색이 아니라 rgb hex 코드로도 가능하다.



<가로 세로 축 범위>

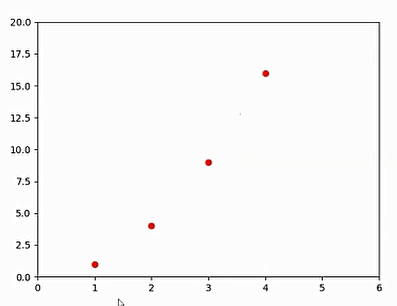
Ex)

Plt.plot([1,2,3,4], [1,4,9,16], ‘ro’)

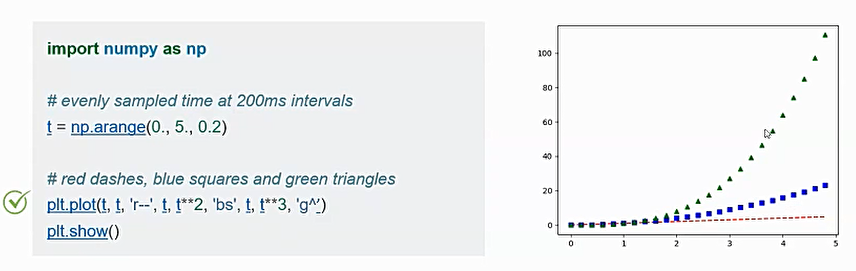
Plt.axis([0,6,0,20])

Plt.show()

* Axis : [xmin, xmax, ymin, ymax] 를 의미.
* ‘ro’는 red, marker를 의미.

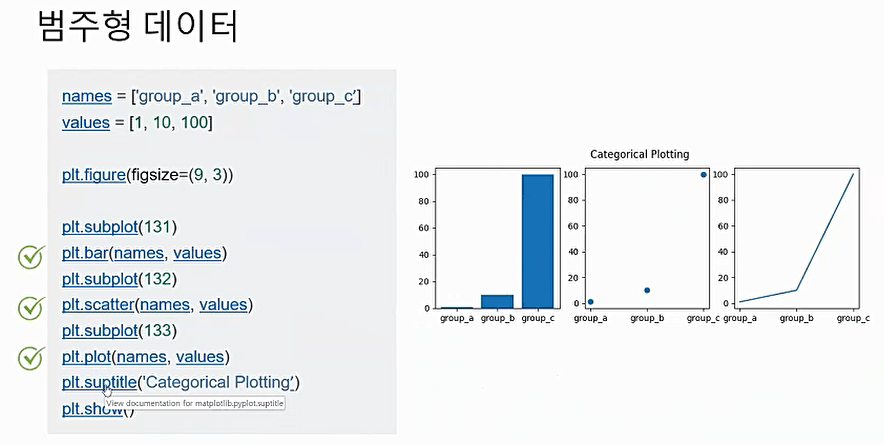


<스타일 추가 +>



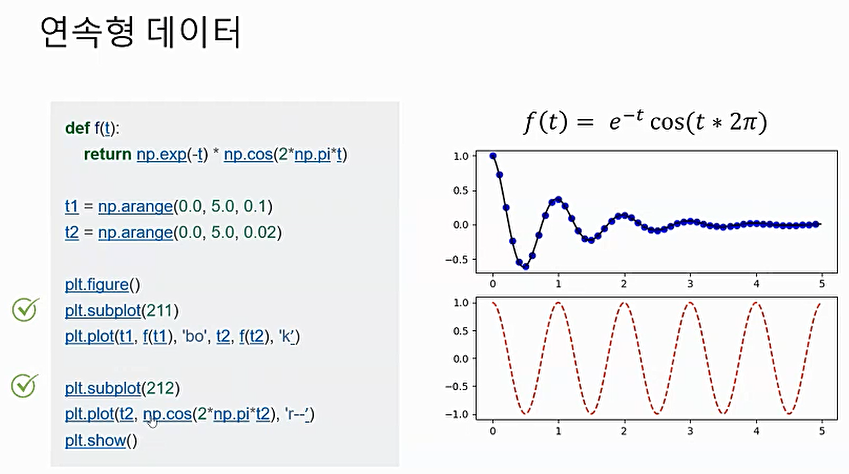
* Numpy는 아직 배우기 전. 일단 리스트로 알고 있어도 무방하다.
* Np.arange(0. , 5., 0.2) = 0에서 5까지의 범위를 0.2 간격으로 나눔. 총 25개의 원소를 가진 리스트
* 한 plot 안에 여러가지 그래프를 그리기 위해선 위와 같이 하나의 plot 안에 여러가지 인자를 주어도 되고, plot을 3번 불러도 된다.

<범주형 데이터>



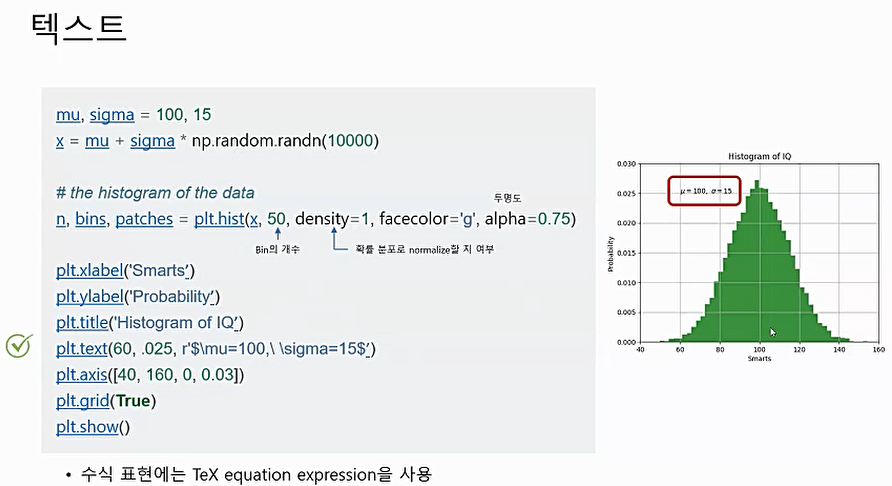
* 범주형 그래프는 위에서 말했듯 주로 막대/파이 그래프로 그리는데 예시와 같이 산포도나 직선 그래프로도 그릴 수 있다.
* Subplot(131) == subplot(1,3,1) 과 같다. 하나의 row에 3개의 그래프가 있음.

<연속형 데이터>



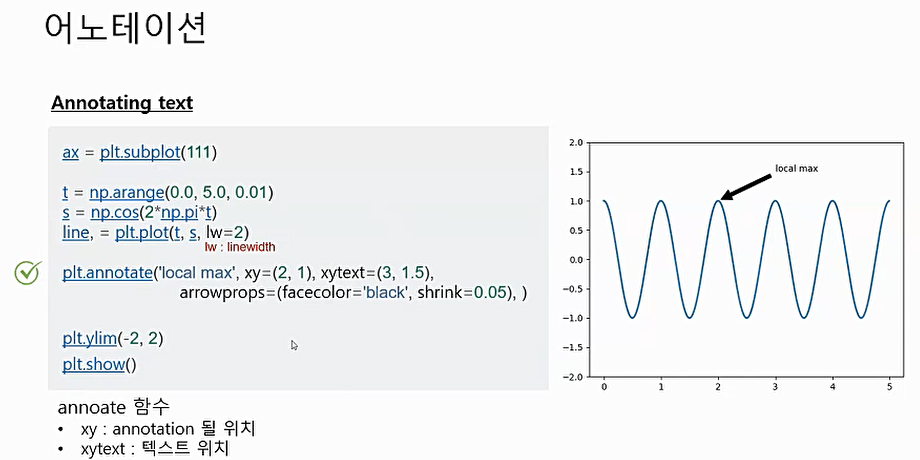
* 수치형 그래프를 이와 같이 나타낸다. 예시와 같이 식이 주어진다면 위에 함수에 대해 코드를 작성하면 된다. 범주형 그래프와 다른 점은 x 좌표 리스트에 카테고리가 아닌 수치 리스트를 넣어줬다는 차이점이 있다.

<텍스트>



* Mu 변수는 평균을 의미, sigma 변수는 표준편차를 의미. 평균이 100이고, 표준편차가 15인 랜덤한 수 10000개를 생성해라는 의미. Bin은 막대 그래프들의 간격, destiny = 1은 원래는 히스토그램의 각 y축은 count 개수지만 destiny=1을 하면 확률분포로 일반화한다.
* 텍스트는 좌표를 표시해주면 된다.

<어노테이션>



* 어노테이션도 텍스트이긴 하지만 일반 텍스트와 다른 점은 특정 포인트에 대해 텍스트를 보여줄 때 사용. Xy: annotation 될 위치, xytext: 텍스트의 위치 그리고 arrowprops 라는 매개변수를 이용하여 화살표로 가리킬 수도 있다. Shrink는 화살표 촉의 굵기를 의미.
* Ylim은 y 좌표만 따로 범위에 limitation을 준 것임.

<이미지>

* Mpimg를 이용하여 imread 사용하여서 plot에 이미지를 넣어줄 수 있다.

{3. 예시 }

<연도 별 GDP>

{4. 시각화 과제 }

{5. 판다스 }

<테이블 형태의 자료 구조>

* 데이터 프레임은 테이블 형태의 데이터를 표현하며 ‘시리즈의 시리즈 구조’ 이다.
* 시리즈란 딕셔너리라고 생각하면 된다. 키와 값으로 이루어져 있는 것.

<k-최근접 이웃>