금융인을 위한

통계와 데이터 분석 입문

데이터를 시각화하기
: 수치형 변수

한국금융연수원
KOREA BANKING INSTITUTE

학습 내용

1. 데이터 시각화의 목적
2. 수치형 변수의 시각화에 많이 활용되는 그래프의 종류
   - 하나의 변수 : 히스토그램(histogram), 상자그림(box plot)
   - 여러 변수 간의 관계 : 산점도(scatter plot),
     산점도 행렬(scatter plot matrix)
3. (실습) 파이썬을 활용한 수치형 자료의 시각화
   - 아파트 실거래가 데이터
   - 엔스콤 사인방 데이터

데이터 시각화의 목적

- 많은 양의 데이터를 한 눈에 파악할 수 있고 보다 쉽게 데이터 인사이트를 찾을 수 있음
- 수준 높은 통계 지식에 대한 설명 없이도 데이터에 대한 정보를 효과적으로 전달할 수 있음
- 단순하게 기술통계량만을 보는 것은 잘못된 해석으로 이어질 수 있는데 시각화가 해석의 오류를 줄여줄 수 있음

예

동일한 기술통계량을 가지고 있지만 분포가 매우 다른 4개의 데이터셋을 포함한 **앤스콤 사인방 (anscombe's quartet)**

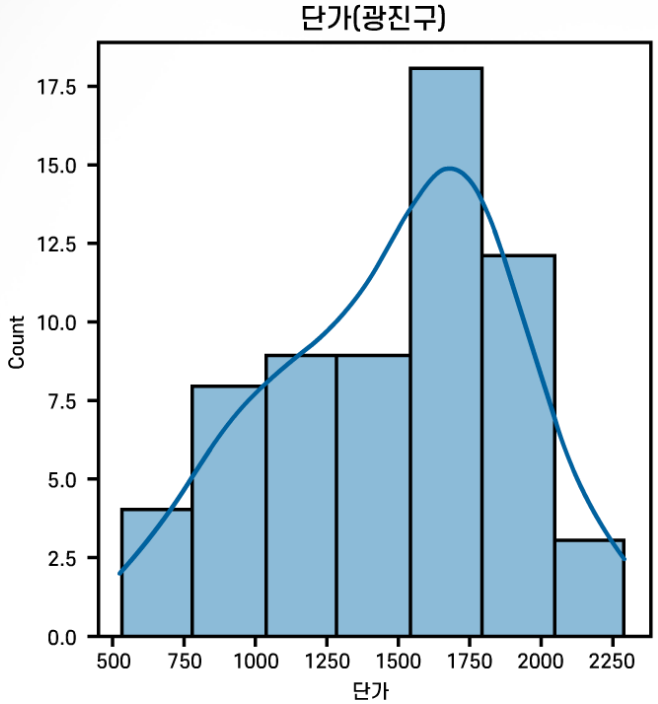
히스토그램(histogram)

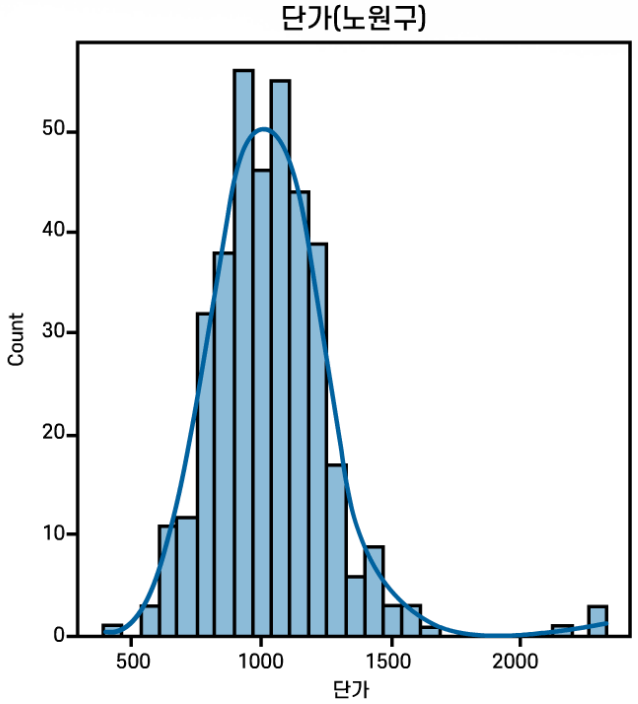
히스토그램

연속인 수치형 변수를 겹치지 않는 **일정한 구간**으로 나누고, x축은 **구간**, y축은 그 구간에 속한 **관측치들의 갯수**로 하여 그래프로 표현한 것

- 필요에 따라 x, y 축을 반대로 할 수 있음
- 구간의 폭을 조정하면 히스토그램의 모양이 바뀔 수 있음

히스토그램(histogram)





상자그림(box plot)

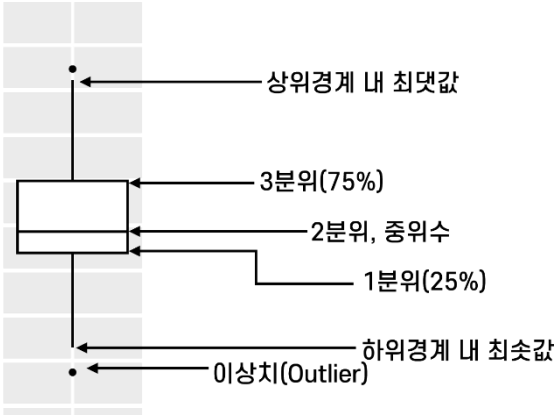
상자그림

수치형 자료로부터 얻어낸 **5가지 요약수치 (five-number summary)**를 가지고 그린 그림

- 상자-수염그림(box-and-whisker plot)이라고도 불림
- 5가지 요약수치
  - 최솟값, 제1사분위수(Q_1), 제2사분위수(Q_2), 제3사분위수(Q_3), 최댓값

상자그림(box plot)

- 상자그림 그리는 법
  - Q_1, Q_2, Q_3 값들을 표시하여 상자를 그림
  - 상자에서 최솟값, 최댓값까지를 선으로 연결함
    - [Q_1 - 1.5 x IQR, Q_3 + 1.5 x IQR] 를 벗어나는 **이상치**가 있으면 그 값을 **제외**한 상태에서 최솟값, 최댓값을 활용함
    - 이상치는 점이나 별로 표기함



산점도 (scatter plot)

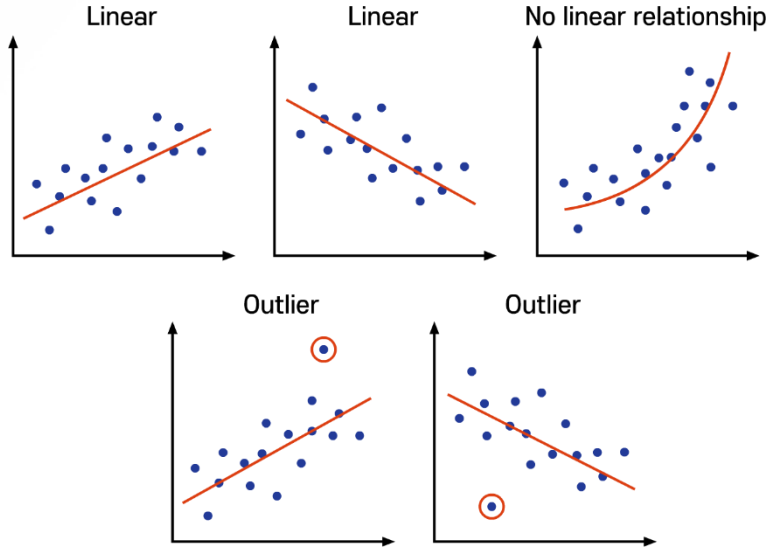
산점도

두 개의 수치형 변수간 관계 탐색을 위해
하나의 변수를 x축, 다른 변수를 y축으로 하고
관측치들을 점으로 표현하여 그린 그래프

- 두 변수의 선형/비선형 등의 관계를 알아볼 수 있고
  특별한 패턴이 있는 경우 거기에서 벗어나는 관측치도
  알아볼 수 있음

수치형 변수 탐색을 위한 그래프

산점도(scatter plot)



- 데이터에 관계를 알아보고 싶은 변수가 여러 개 포함되어 있는 경우 산점도 행렬(scatter plot matrix)을 통해 알아보기도 함

파이썬의 대표적인 시각화 패키지

matplotlib

- matplotlib의 모듈 중 pyplot을 사용하여 다양한 그래프 생성 가능함
- http://matplotlib.org/

seaborn(주로 사용할 예정)

- matplotlib에 기반한 시각화 패키지
- 쉬운 문법을 가지고 있어 사용에 편리하고 데이터 분석에 용이하다는 장점이 있음
- https://seaborn.pydata.org/

학습정리

- 데이터 시각화의 목적은 자료의 탐색 및 전달
- 수치형 변수 하나에 관한 시각화에 히스토그램(histogram), 상자그림(box plot)이 많이 활용됨
- 여러 개의 수치형 변수 간의 관계의 시각화에 산점도(scatter plot), 산점도 행렬(scattter plot matrix)이 활용됨
- 파이썬을 활용하여 수치형 변수들을 시각화할 수 있음