[0117]\_Visualizaion\_9기\_조세린\_요약과제제출

1.Visualizaion의 필요성 => 효과적인 내용 전달을 위해

- 데이터의 직관적인 이해

- 데이터의 트렌드 파악

- 변수 간 관계 파악

2. 데이터의 종류

-수치형 (양적 데이터) : 연속형 데이터, 이산형 데이터 / 범주형 데이터 (질적 데이터) : 순위형 데이터, 명목형 데이터

- 고려하는 변수의 개수에 따라 -> 일변량 (고려하는 변수 1개) , 다변량 (고려하는 변수 2개 이상)

3. 시각화 패키지

-matplotlib

-pyplot (matplolib의 서브패키지)

-seaborn (matplotlib기반, 다양한 색상 테마와 통계용 차트 기능 추가)

4. plot 종류

|  |  |  |
| --- | --- | --- |
| countplot | 카테고리별로 데이터가 얼마나 있는지 표시 |  |
| histogram | 수치형 변수의 구간별 빈도 표현 |  |
| barchart | 각 카테고리 별 수치형 데이터 표현 |  |
| boxplot | 카테고리 값별 수치형 변수의 분포 표시 : Q3, Q1, 중앙값, 아웃라이어 표시 |  |
| piechart | 전체에 대한 각 카테고리의 비율 표시 |  |
| scatterplot | 두 실수 데이터 집합의 상관관계 시각화 |  |
| trendline | 데이터들 간에 존재하는 추세선 표시 |  |
| line plot/point plot | 특정 데이터를 x, y로 표시하여 관계 확인 가능 -> 수치형 지표들 간의 경향/시계열 데이터를 파악할 때 유용 |  |
| heatmap | x축과 y축에 해당하는 2개의 범주형 자료의 계급(class)별로 집계한 연속형 자료 표시(집계한 값에 비례하여 색깔을 다르게 하여 시각화) |  |

5. 기타

<word cloud>

-텍스트 데이터 전처리

1) text cleaning : list comprehension을 이용해 빈 줄, 문단의 양 끝단의 공백 제거->다시 하나의 string으로 묶는다

2) 형태소 분석

3) 불용어 제거

4) 빈도분석