파이썬 라이브러리를 활용한 데이터 분석

툴 스터디 2주차

지난주!

- 타이타닉 데이터 살펴보기
- “Survived”, “Pclass” 데이터 타입 변환: int -> object
- 결측 값 확인: info(), isnull().sum()
- 결측 값이 많은 열(Cabin)과, Embarked 데이터 중 결측 값이 존재하는 행 제거!

EDA란?

- Exploratory Data Analysis (탐색적 데이터 분석)

- 데이터를 '다양한 각도'에서 관찰하고 이해하는 과정

- 본격적으로 데이터를 분석하기 전, '그래프'나 '통계적인 방법'을 활용하여 데이터를 직관적으로 파악하고, 이해하는 전체적인 과정을 의미합니다.

EDA를 거쳐야 하는 이유

- 데이터의 '분포'와 '값'을 검토하여, 데이터가 어떤 현상을 표현하고 있는지 이해할 수 있습니다.
- 데이터가 지니고 있는 '잠재적인' 문제들을 발견할 수 있습니다.
- 본격적으로 데이터를 분석하기 전에, 해당 데이터의 어떤 부분을, 어떻게 수집해서, 어떤 부분에 초점을 두고 바라보아야 할 지 파악할 수 있습니다.
- 다양한 각도에서 살펴보면서 다양한 패턴을 파악하고, 이를 바탕으로 가설을 세울 수 있고, 가설을 수정할 수 있게 됩니다.

EDA 과정

- **1. 시각화를 활용하기**

  데이터를 '전체적으로' 살펴보기: 데이터에 이상치, 결측 값이 있는가?

- **2. 통계 지표를 활용하기**

  데이터의 '개별 속성값 ' 을 관찰하기: 각 속성 값이 예측 범위 내의 분포를 갖는지 확인하기,

  그렇지 않다면, 이유가 무엇인지 확인하기

- **3. 상관관계 파악하기**

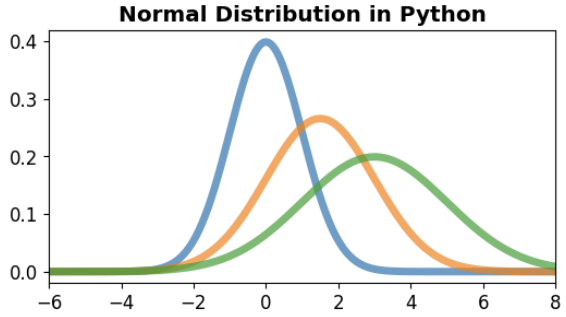
  어떤 변수(속성) 간의 관계를 집중적으로 관찰해야 하는가?

  이 관계를 분석하기 위한 최적의 방법은 무엇인가?

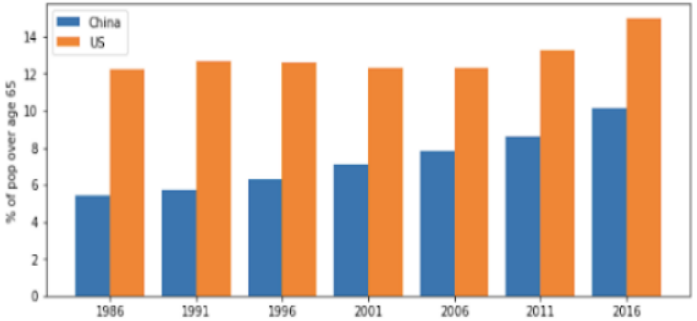
  Ex) 타이타닉 데이터에서 "Survived"속성과 "Age" 속성 간의 상관관계를 파악하기

1. 시각화를 활용하기

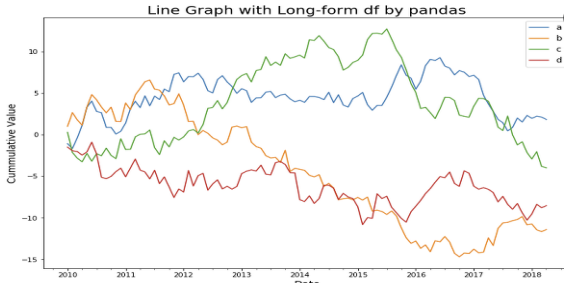
- **확률 밀도 함수** – 연속형 데이터



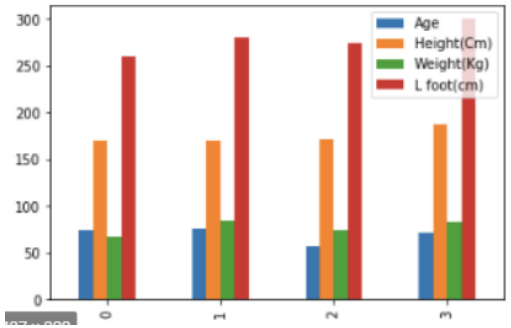
- **막대 그래프** – 범주형 데이터,명목형 데이터



- **시계열 차트**

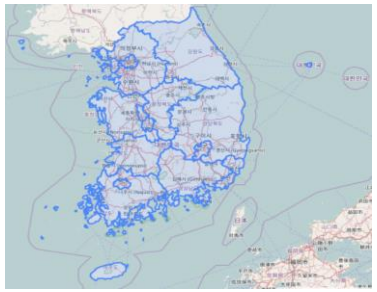


히스토그램 – 연속형 데이터



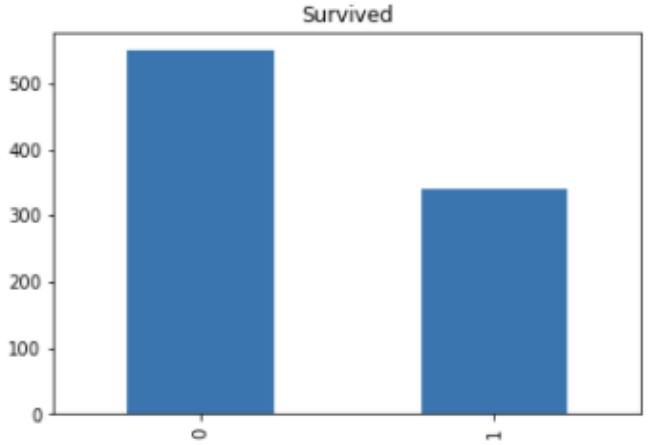
워드 클라우드

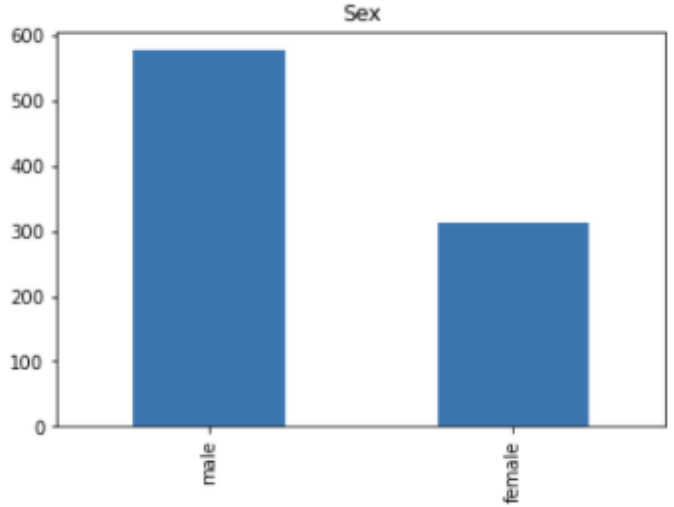
지도

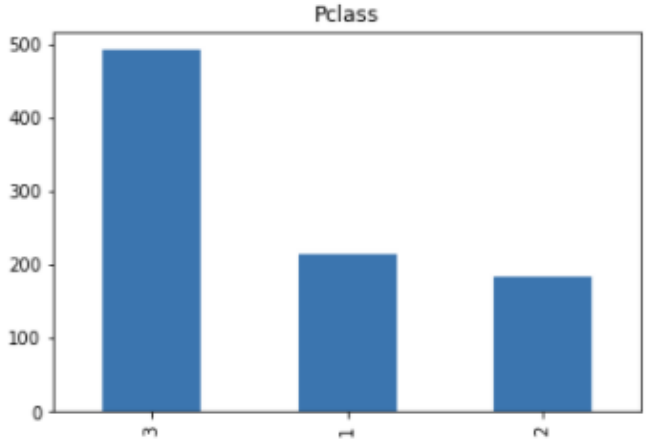


1. 시각화를 활용하기

- 타이타닉의 열 중 object 타입인 “Survived”, “Pclass”, “Name”, “Sex”, “Ticket”, “Embarked” 데이터를 막대 그래프로 시각화하기

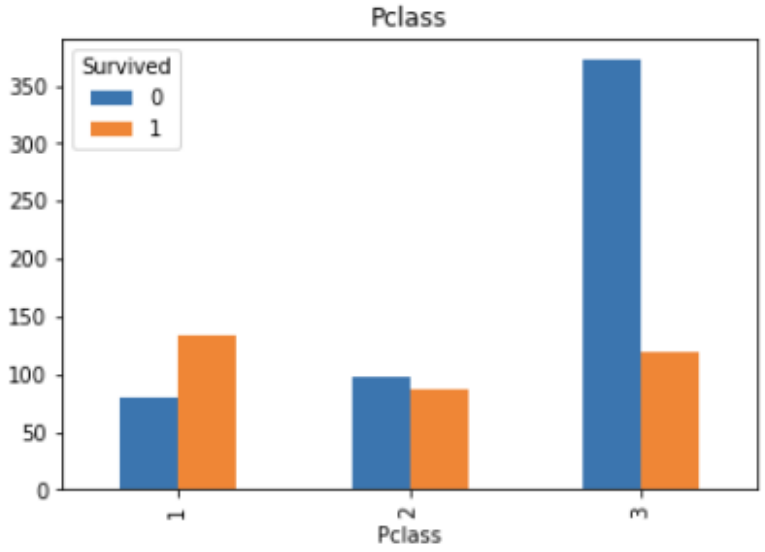


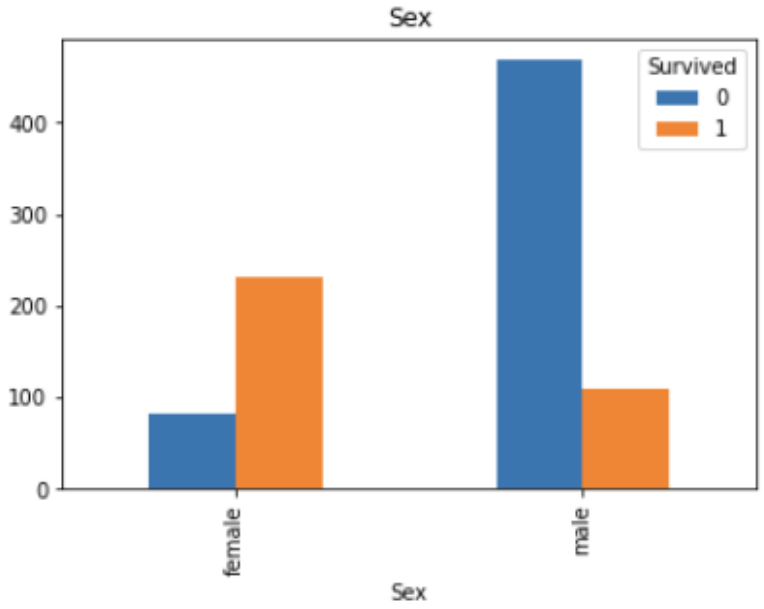


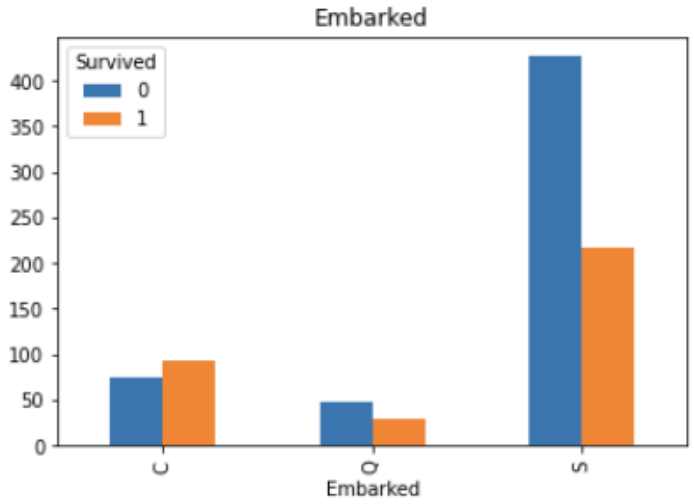


1. 시각화를 활용하기

- “Pclass”, “Sex”, “Embarked” 별로 “Survived” 데이터 수를 막대 그래프로 시각화하기







2. 통계 지표를 활용하기

- **1. 데이터의 '중심' 부분에 대한 정보를 알 수 있는 통계 지표들**

- **1) 평균(mean): 데이터 값을 데이터 표본 수로 나눔**
- **2) 중앙값(median): 전체 데이터 중 가장 중앙에 있는 값**
- **3) 최빈값(mode): 데이터에서 가장 자주 나오는 값**

- **주의해야 할 점: 평균-이상치의 영향을 받음 / 중앙값-이상치의 영향을 받지 않음**
- Ex) 회사 직원들의 연봉 평균: 대부분 연봉 중앙값보다 높은 값이 나옴, 왜냐하면 소수의 고액 연봉자가 전체 회사원들의 연봉 평균을 끌어 올리는 경우가 있기 때문.

2. 통계 지표를 활용하기

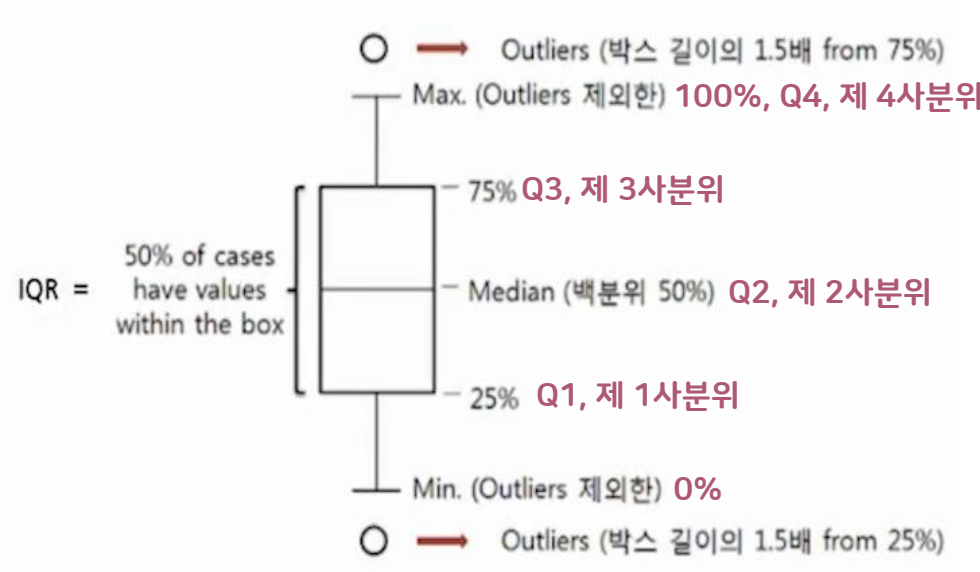
- 2. 데이터가 '얼마나 퍼져 있는지' 나타내는 지표 (데이터의 분산을 알려주는 지표)

- 1) 범위(range): 데이터의 최댓값(max)와 최솟값(min)의 차이

- 2) 분산(variance)과 표준편차(Standard deviation)

2. 통계 지표를 활용하기

- 3) 박스플롯(Boxplot): 데이터 집합의 '범위'와 '중앙값', '이상치 존재 여부'를 확인 가능



- 사분위(Quantile): 100%의 데이터를 4등분한 기준,

  100%, 75%, 50%, 25% 지점에서 분할하여 데이터를 4등분함

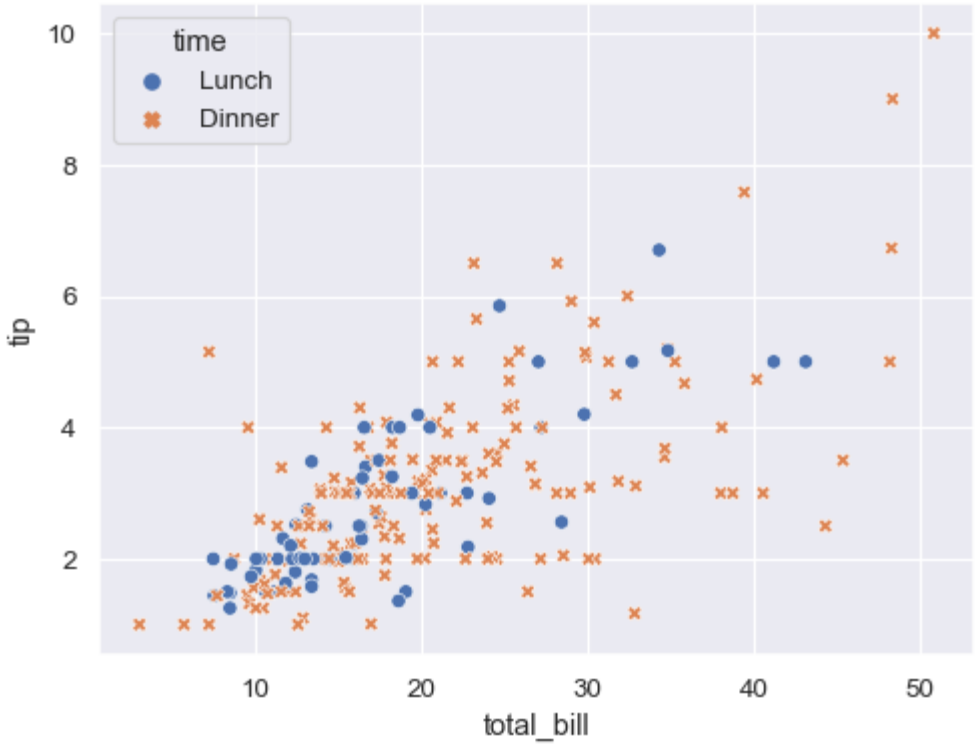
- IQR(Interquartile Range): Q3(75%지점) – Q1(25%지점)

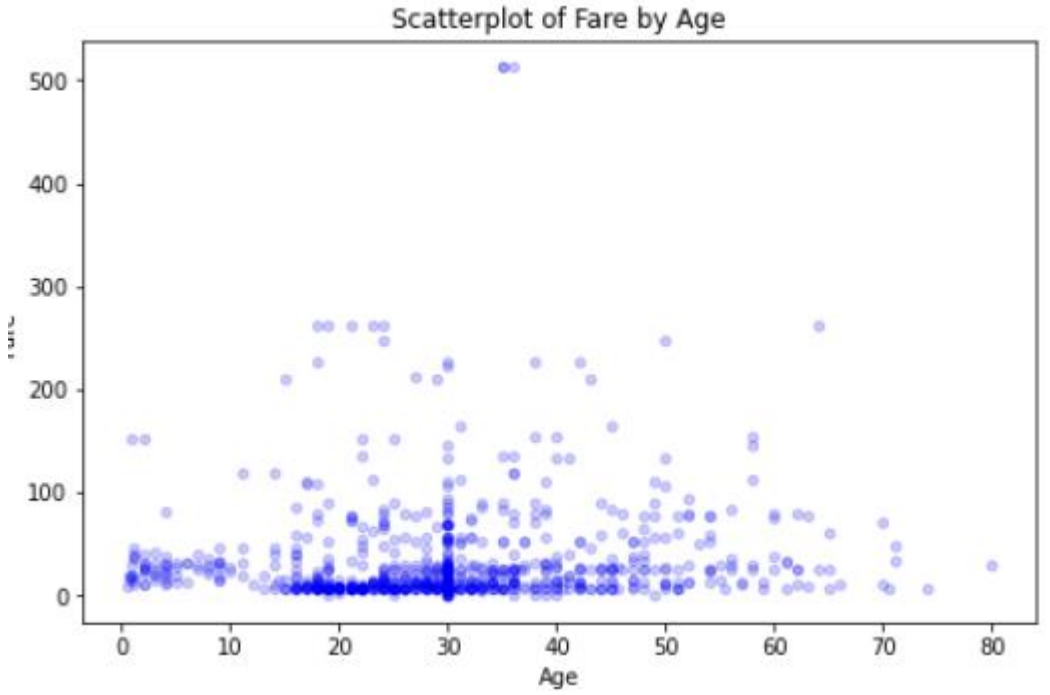
2. 통계 지표를 활용하기

- **4) 산점도(scatter plot):**

  데이터 값에 따라 점을 찍어 전체 데이터의 분포를 확인할 수 있는 차트,

  이상치도 함께 확인 가능





3. 상관관계 파악하기

- 상관관계(correlation)란? 변수들 간의 '관계'를 의미합니다.

  두개 이상의 변수 사이의 관계에서, 한 변수가 변화할 때 다른 변수가 어떻게 변화하는지 파악해야 합니다.

- 상관계수: 변수들의 상관관계의 정도를 의미합니다.

상관계수	상관관계 강도
±0.9 이상	매우 높은 상관관계
±0.7~±0.9	높은 상관관계
±0.4~±0.7	다소 높은 상관관계
±0.2~±0.4	낮은 상관관계
±0.2 미만	상관관계가 거의 없음

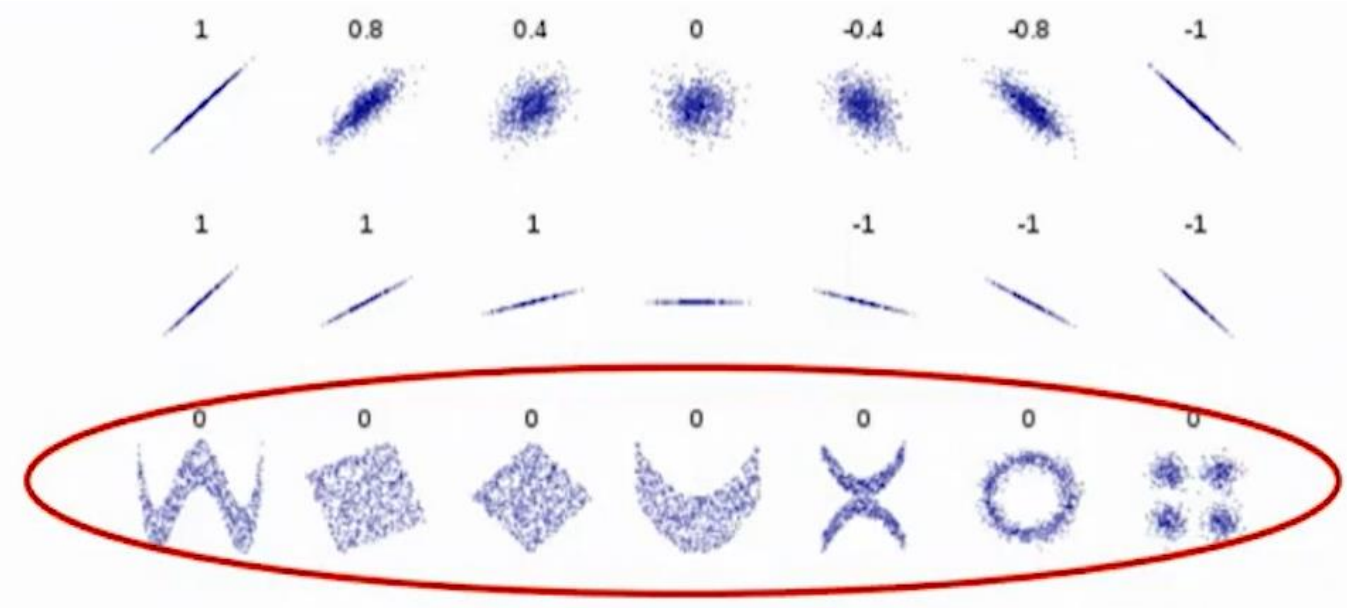
3. 상관관계 파악하기

- 주의할 점

  상관관계 != 인과관계 (원인과 결과)

  상관계수가 0이라고 해서 변수들 간에 상관관계가 '없다'라고 단정짓기는 어려움

  따라서 아까 그려본 산점도(Scatter plot)를 통해 그래프 모양을 확인해야 함!!



3. 상관관계 파악하기

- Seaborn 라이브러리의 Heatmap 그래프를 이용해 변수 간의 상관관계 시각화하기
- Anaconda prompt에서 ‘pip install seaborn’ 입력해 설치 가능