22-02-21

EDA

* data.shape : 테이블의 행, 열 수
* 컬럼 이름을 바탕으로 정수 인덱스(위치) 정보 : data.columns.get\_loc("컬럼") 🡪컬럼 수가 여러 개 일 때 : cols = [“컬럼1”, “컬럼2”, ..] idx\_cols = [data.columns.get\_loc(col) for col in cols]
* 기존 데이터 한 컬럼에 타입 변경 : data["컬럼"] = data["컬럼"].astype("int64")
* data.describe( ) 은 문자로 된 컬럼은 처리 안 됨
* 컬럼의 종류 와 수 : data["컬럼"].value\_counts() 🡪 비율구하기 : len(data)로 나누기 혹은 data["컬럼"].value\_counts(normalize = True)
* data.sort\_values(by=["컬럼1", “컬럼2], ascending= [False, True] ) : 컬럼1을 기준으로 내림차순 정렬, 그 다음 기준인 컬럼2은 오름차순 정렬
* data[data["컬럼"] == 1] ["컬럼2"].mean( ) : 컬럼이1인 컬럼2에 대한 데이터의 평균 구하기
* data[data["컬럼"].apply(lambda x : x[0] == "5") ] : 데이터의 컬럼중에서 첫자리가 5인 데이터 불러오기

1. 조건이 1개 일 때 labmda x : 값1 if 조건1 else 엘스값
2. 조건이 2개 일 때 lambda x : 값1 if 조건1 else ( 값2 if 조건2 else 엘스값2)

* d = {"Yes":True, "No":False} data = data.replace( {"컬럼": d} ) : 컬럼의 Yes를 True로 No를 False로 바꾸기
* data = data.drop(['컬럼1', '컬럼2'], axis = 1) 컬럼1과 컬럼2을 지우기 🡪 drop는 업데이트 되지 않아서 할당 필요함
* Seaborn을 위해 필요한 패기지들 import matplotlib.pyplot as plt import seaborn as sns
* cols = [“컬럼1”, “컬럼2”] data[cols].hist(figsize = (10,4)) : 컬럼1과 컬럼2에 대한 히스토그램을 1판에 2개 그리기
* data[cols].plot(kind = “density”, subplots = True, sharex = True) 🡪 density : 밀도함수 , subplot : 분리해서 그리기, sharex : x축 같은값으로 나타내기 (sharey도 가능)
* 히스토그램, 밀도함수 같이 나타내기 : sns.distplot(data["컬럼"]) or sns.displot( data = data, x = “컬럼”, kde = True)
* boxplot 생성 : sns.boxplot(data = data, x = “컬럼”)