**배운 내용 정리**

df = pd.read\_csv("파일 위치", low\_memory=False)

가져온 데이터파일을 판다스로 인식해 df(데이터 프레임 약칭)에 담는다.

low\_memory=False 다른 타입이 섞여 있어 써주어야 한다.

df.shape 는 데이터 프레임의 행과 열의 수를 보여주고 (행, 열)로 표시된다.

Shift+Tap 을 했을 때 docstring을 볼 수 있다. Shift+Tap+Tap은 docstring을 크게 볼 수 있다.

df.head() 는 데이터 프레임의 앞부분을 보여주고

기본값은 다섯 개만 보여주지만 괄호 안에 수를 넣으면 그 수만큼 보여준다.

df.tail() 은 데이터 프레임의 앞부분을 보여주고

기본값은 다섯 개만 보여주지만 괄호 안에 수를 넣으면 그 수만큼 보여준다.

df.info() 로 데이터 요약을 볼 수 있다.

* 첫번째는 컬럼명을 볼 수 있다
* 두번째는 각 컬럼의 수를 볼 수 있다.

다른 컬럼과 숫자 차이가 나면 결측치가 들어가 있다는 걸 알 수 있다.

* 세번째는 nun-null 타입을 볼 수 있다.
* 네번째는 데이터의 타입을 볼 수 있다.
* 마지막줄에는 메모리 사용량이 적혀있다.

df.columns() 로 컬럼명만 출력할 수 있다.

df.dtypes() 로 데이터 타입만 출력 할 수 있다.

df.isnull() 로 결측치를 볼 수 있다.

null값은 True 로 표시 된다.

df.isnull().sum() 을 하면 파이썬에서는 True==1 이기 때문에 sum으로 결측치를 셀 수 있다.

df.plot.barh() 자료를 가로로 된 막대그래프로 보여준다.

* 괄호 안에 figsize=(x,y)를 입력하면 x 만큼 가로로 늘어나고 y 만큼 세로로 늘어난다.
* 괄호 안에 rot=수 넣으면 입력한 수만큼 글씨가 돌아간다.
* 괄호 안에 grid=True 를 입력하면 격자가 표시된다.

df.reset\_index 를 하면 자료를 데이터 프레임의 형태로 만든다

df.columns = [“ ”, “ ”] 에서 대괄호 안에 스트링을 넣으면 컬럼명을 지정할 수 있다.

df.sort\_values(by=" ", ascending=]

* by= 뒤에 컬럼명을 입력하면 컬럼을 지정 할 수 있다
* ascending= 뒤에 True 를 입력하면 오름차순 False 를 입력하면 내림차순이 된다.

df[" "] 로 컬럼을 불러 올 수 있디.

NaN 로 표시되는 것은 Not a Number 의 약자로 결측치를 의미한다.

df[" "].tolist() 컬럼 하나를 리스트로 만들어준다.

df.drops() 로 원하는 행이나 열을 지울 수 있다.

* 첫 부분에 지우고 싶은 위치를 선정해준다.
* axis= 뒤에 0을 넣으면 행을 삭제하고 1을 넣으면 열을 삭제한다.

df[" "].mean() 지정한 컬럼의 평균값을 알 수 있다.

df[" "].median() 지정한 컬럼의 중간 값을 알 수 있다.

df[" "].max() 지정한 컬럼의 최대값을 알 수 있다.

df[" "].min() 지정한 컬럼의 최소값을 알 수 있다.

df[" "]. describe() 지정한 컬럼의 데이터 요약값을 볼 수 있다

* 25%는 앞에서 1/4 되는 값이고, 1사분위수라 부릅니다.
* 50%는 중앙값이고, 2사분위수라고도 합니다.
* 75%는 앞에서 3/4 번째 되는 수고, 3사분위수라 부릅니다.
* mean, min, max는 각각 평균, 최소값, 최대값입니다.

df[[" ", " "]].describe()

* 요악값을 시리즈가 아닌 데이터 프레임의 형태로 만든다.
* 2개 이상의 데이터를 출력하려면 컬럼을 리스트 형태로 넣어야한다.

df.describe(include=" ")

* 따옴표 안에 number 가 들어가면 숫자로 된 데이터만 요약
* 따옴표 안에 object 가 들어가면 문자열 데이터만 요약
  + count 는 데이터의 개수를 세준다.
  + unique 는 중복을 제거한 값을 보여준다.
  + top 은 가장 높은 빈도로 저장된 데이터, freq는 그 데이터의 빈도수를 표시한다.

df[" "].unique() 컬럼의 중복을 제거한 값을 보여준다.

df[" "].nunique() unique의 수를 보여준다.

df[" "].value\_counts() 카테고리 형태로 데이터 개수를 세준다.

괄호 안에 normalize=True 를 넣으면 비율을 구해준다.

df.plot.pie() 자료의 파이 그래프를 보여준다.

하지만 서로의 크기 차이를 확연히 보기 힘들기 때문에 그래프 해석에 오류가 생길 수 있어 잘 사용하지 않는다.

sns.countplot(data=df, x or y=" ") seaborn 으로 그래프를 그린다.

* seaborn 에서 색을 부여하기 때문에 미관상 보기 좋다.
* data=df 데이터를 df 로 지정한다.
* x를 하면 세로 막대 그래프가 나오고 y 를 하면 가로 막대 그래프가 나온다.
* 큰 따옴표 안에 컬럼명을 적으면 그 컬럼의 그래프를 그란다.

df[df[" "] == " "].copy() 원하는 데이터를 지정해서 가져온다.

변수에 지정할 때 df[“ ”].copy()를 하지 않으면 같은 데이터를 쓰기 때문에 불편 할 수 있다.

df.loc[df[" "] == " ", " "].value\_counts() df.loc 는 행과 열을 함께 가져올 수 있게 한다.

df[(df[" "] == " ") & (df[" "] == " ")] 판다스에서는 and 연산자가 & 이다

데이터를 가져올 땐 우선순위에 의해 괄호를 꼭 쳐 주어야 한다.