|  |  |
| --- | --- |
| 교육 제목 | **Pandas 기본 및 활용** |
| 교육 일시 | 20211110 09:00 ~ 18:00 |
| 교육 장소 | 자택 온라인 수업 |
| **교육 내용** | |
| 오전 | # 파일명의 프로그램 코드를 현재의 프로그램에 삽입  # %load 파일명  import os  os.getcwd()  # % load '/Users/dongjinkim/OneDrive/2. Private/AI lecture\_202110/coding\_python/Anaconda\_python/파일명'  import matplotlib  import matplotlib.pyplot as plt  import pandas as pd  import numpy as np  import seabon as sns  matplotlib.rcParams['font.']  # age 컬럼의 Nan 처리  # 누락 데이터가 400개 이상인 컬럼을 삭제 -> 데이터 분석에 영향이 없는 컬럼  df\_t = df.dropna(axis=1, thresh=400)  df\_t.head()  # 1. 제거하는 방법  df\_age = df.dropna(subset = ['age'], how = 'any', axis=0)  print(len(df), len(df\_age))  # 2. 치환하는 방법 : 평균  mean\_age = df['age'].fillna(np.floor(df['age'].mean(axis=0)))  # mean\_age = df['age'].fillna(0)  mean\_age.head(10) |
| 오후 | # 중복 제거  df2 = df.copy()  df2 = df2.drop\_duplicate()  df2  # 컬럼 기준으로 중복 제거  df3 = df.copy()  df3= df3.drop.duplicates(subset=['c2','c3'])  df3  # horsepower 자료에 Nan 확인 -> Nan 없애고 시작  df\_mpg.dropna(subset = ['horsepower'], axis=0, inplace=True)  df\_mpg.info()  # 2. 자료형 변환 : 범주형 데이터를 확인 변형 astype('int32', 'category')  print(df\_mpg['origin'].dtypes)  df\_mpg['origin'] = df\_mpg['origin'].astype('category')  print(df\_mpg['origin'].dtypes)  df\_1["horsepower"].unique()  df\_1['horsepower'].replace('?', np.nan, inplace=True)  df\_1.dropna(shbset=['horsepower'], axis = 0, inplace=True)  df\_1['horsepower']=df\_1['horsepower'].astype('float')  # 데이터 전처리  # 1. 결측치 처리  # 2. 데이터형 변환  # 3. 범주형으로 변경 (구간 분할, np.histogram(), pd.cut())  # 4. 원한인코딩 (더미변수, skjearn.preprocessing)  # 5. 정규화  # df.reset\_index(inplace=True)  df['New\_Data'].dt.year[0]  df['New\_Data'].dt.month[0]  df['New\_Data'].dt.day[0] |