|  |  |
| --- | --- |
| 교육 제목 | 판다스 데이터 분석 |
| 교육 일시 | 20210928 화요일 |
| 교육 장소 | 영우글로벌러닝 |
| **교육 내용** | |
| 오전 | import pandas as pd  df = pd.read\_csv("C:/Users/cwh/myreposit/jpynb/auto-mpg.csv", header=None)  df.columns = ['mpg','cylinders','displacement','horsepower','welght','acceleration','model\_year','origin','name']  print(df.head(3))  print("\n")  print(df.tail())  print(df.shape) # 데이터프레임의 차원을 보여줌  #데이터프레임의 행의 갯수와 열의 갯수를 출력  print("rows : {}, columns : {}".format(df.shape[0],df.shape[1]))  print('\n')  df.info()  print(df.shape) #데이터프레임의 차원을 보여줌  #데이터 프레임의 행의 갯수와 열의 갯수를 출력  print()  print(df.describe()) # df.describe() : return dat type : DataFrame  df\_desc = df.describe().loc[['count','std'],['mpg','welght']]  df\_desc  print(df.count()) # 전체 컬럼들의 갯수  df['mpg'].value\_counts() # 같은 값들의 갯수  df['origin'].value\_counts()  df['name'].value\_counts()  # 판다스 내장 그래프 활용  df = pd.read\_excel("./남북한발전전력량.xlsx")  df\_ns = df.iloc[[0,5],3:]  df\_ns.index = ['South','North']  print(df\_ns.columns)  df\_ns.columns = df\_ns.columns.map(int) # 컬럼명을 int로 변경  df\_ns.columns  #데이터 전처리 : 컬럼의 type 변경, 원하는 정보만 추출, 데이터 가공, Nan 데이터 처리...... |
| 오후 | import pandas as pd  import matplotlib.pyplot as plt  # plt.rcParams['font.family'] = 'malgun.ttf'  plt.rcParams['font.family'] = 'Malgun Gothic'  plt.rcParams['axes.unicode\_minus'] = False  df = pd.read\_excel("./0930pandas/시도별 전출입 인구수.xlsx",header=0)  df = df.fillna(method='ffill') # 전출이 누락된 테이터를 앞 데이터로 채움  df.head()  # 서울에서 다른 지역으로 이동한 데이터만 추출하여 정리  mask = (df['전출지별'] == '서울특별시') & (df['전입지별'] != '서울특별시')  df\_seoul = df[mask] # mask를 이용하여 True인 데이터만 추출  df\_seoul = df\_seoul.drop(['전출지별'], axis=1) # 전출지별 컬럼 삭제  df\_seoul.rename({'전입지별':'전입지'}, axis=1, inplace=True) # 컬럼명 변경  df\_seoul.set\_index('전입지', inplace=True) # '전입지' 컬럼을 인덱스로 설정  df\_seoul  # 부산에서 다른지역으로 전출한 데이터만 추출해서 df\_busan으로 저장  # 불필요한 컬럼 제거, 인덱스는 '전입지별' -> '전입지'로 바꾸어서 이덱스 설정  mask = (df['전출지별'] == '부산광역시') & (df['전입지별'] != '부산광역시')  df\_busan = df[mask]  df\_busan = df\_busan.drop(['전출지별'], axis=1)  df\_busan.rename({'전입지별':'전입지'}, axis=1, inplace=True)  df\_busan.set\_index('전입지', inplace=True)  print(df\_busan)  df\_busan\_t = df\_busan.T # 데이터프레임의 행과 열을 바꿈  df\_busan\_t.plot(kind='line')  df\_1 = df\_seoul.loc['경기도']  plt.plot(df\_1.index, df\_1.values)  plt.style.use('seaborn-deep')  plt.title("서울에서 경기도로 전출한 인구수", size=20)  plt.xlabel("연도") # x축 라벨  plt.ylabel("인구수",rotation=0) # y축 라벨  plt.xticks(rotation=120) # x라벨의 기울기  plt.legend(labels=["서울 -> 경기"])  plt.show()  plt.style.available #스타일 리스트  import pandas as pd  import matplotlib.pyplot as plt  mask = (df['전출지별'] == '서울특별시') & (df['전입지별'] != '서울특별시')  df\_seoul = df[mask]  df\_seoul = df\_seoul.drop(['전출지별'], axis=1)  df\_seoul.rename({'전입지별':'전입지'}, axis=1, inplace=True)  df\_seoul.set\_index('전입지', inplace=True)  # 서울에서 '충청남도','경상북도', '강원도', '전라남도'로 이동한 인구 데이터 값만 선택  col\_years = list(map(str, range(1970, 2018)))  df\_4 = df\_seoul.loc[['충청남도','경상북도', '강원도', '전라남도'], col\_years]  # 라이브러리 불러오기  import pandas as pd  import matplotlib.pyplot as plt  # matplotlib 한글 폰트 오류 문제 해결  # from matplotlib import font\_manager, rc  # font\_path = "./malgun.ttf" #폰트파일의 위치  # font\_name = font\_manager.FontProperties(fname=font\_path).get\_name()  # rc('font', family=font\_name)  plt.style.use('ggplot') # 스타일 서식 지정  plt.rcParams['axes.unicode\_minus']=False # 마이너스 부호 출력 설정  # Excel 데이터를 데이터프레임 변환  df = pd.read\_excel('./남북한발전전력량.xlsx', engine= 'openpyxl', convert\_float=True)  df = df.loc[5:9]  df.drop('전력량 (억㎾h)', axis='columns', inplace=True)  df.set\_index('발전 전력별', inplace=True)  df = df.T  # 증감율(변동률) 계산  df = df.rename(columns={'합계':'총발전량'})  df['총발전량 - 1년'] = df['총발전량'].shift(1)  df['증감율'] = ((df['총발전량'] / df['총발전량 - 1년']) - 1) \* 100  # 2축 그래프 그리기  ax1 = df[['수력','화력']].plot(kind='bar', figsize=(20, 10), width=0.7, stacked=True)  ax2 = ax1.twinx()  ax2.plot(df.index, df.증감율, ls='--', marker='o', markersize=20,  color='green', label='전년대비 증감율(%)')  ax1.set\_ylim(0, 500)  ax2.set\_ylim(-50, 50)  ax1.set\_xlabel('연도', size=20)  ax1.set\_ylabel('발전량(억 KWh)')  ax2.set\_ylabel('전년 대비 증감율(%)')  plt.title('북한 전력 발전량 (1990 ~ 2016)', size=30)  ax1.legend(loc='upper left')  plt.show()  # 라이브러리 불러오기  import matplotlib.pyplot as plt  import seaborn as sns    # Seaborn 제공 데이터셋 가져오기  titanic = sns.load\_dataset('titanic')    # 스타일 테마 설정 (5가지: darkgrid, whitegrid, dark, white, ticks)  sns.set\_style('whitegrid')  # 그래프 객체 생성 (figure에 3개의 서브 플롯을 생성)  fig = plt.figure(figsize=(15, 5))  ax1 = fig.add\_subplot(1, 3, 1)  ax2 = fig.add\_subplot(1, 3, 2)  ax3 = fig.add\_subplot(1, 3, 3)    # x축, y축에 변수 할당  sns.barplot(x='sex', y='survived', data=titanic, ax=ax1)  # x축, y축에 변수 할당하고 hue 옵션 추가  sns.barplot(x='sex', y='survived', hue='class', data=titanic, ax=ax2)  # x축, y축에 변수 할당하고 hue 옵션을 추가하여 누적 출력  sns.barplot(x='sex', y='survived', hue='class', dodge=False, data=titanic, ax=ax3)  # 차트 제목 표시  ax1.set\_title('titanic survived - sex')  ax2.set\_title('titanic survived - sex/class')  ax3.set\_title('titanic survived - sex/class(stacked)')  plt.show()  import seaborn as sns    # Seaborn 제공 데이터셋 가져오기  titanic = sns.load\_dataset('titanic')    # 스타일 테마 설정 (5가지: darkgrid, whitegrid, dark, white, ticks)  sns.set\_style('darkgrid')  # 그래프 객체 생성 (figure에 2개의 서브 플롯을 생성)  fig = plt.figure(figsize=(15, 5))  ax1 = fig.add\_subplot(1, 2, 1)  ax2 = fig.add\_subplot(1, 2, 2)    # 그래프 그리기 - 선형회귀선 표시(fit\_reg=True)  sns.regplot(x='age', #x축 변수  y='fare', #y축 변수  data=titanic, #데이터  ax=ax1) #axe 객체 - 1번째 그래프  # 그래프 그리기 - 선형회귀선 미표시(fit\_reg=False)  sns.regplot(x='age', #x축 변수  y='fare', #y축 변수  data=titanic, #데이터  ax=ax2, #axe 객체 - 2번째 그래프  fit\_reg=False) #회귀선 미표시  plt.show() |