**교육일지**

**교육 제목 : 데이터 프레임 구조, 시각화 도구(matpolotlib, seaborn, folium)**

**교육 장소 : YGL C6 강의실**

**교육 일시 : 2021/09/29**

**오전 교육**

**데이터 프레임 구조**

**'mpg'( 연비) - 연속의 값**

**'cylinders'(실린더 수) - 이산값 예(3,4,6,8)**

**'displacement'(배기량) - 연속값**

**'horsepower'(출력) - 연속값**

**'weight'(차중) - 연속값**

**'acceleration'(가속 능력) - 연속값**

**'model\_year'(출시년도) - 이산값 예(70,71,80)**

**'origin'(제조국) - 이산값 예( 1:usa, 2:eu)**

**'name'(모델명) - 문자열**

**import pandas as pd**

**df = pd.read\_csv('C:/Users/ADMIN/Desktop/python\_src/auto-mpg.csv', header=None)**

**df.columns = ['mpg','cylinders','displacement',**

**'horsepower', ' weight', 'acceleration',**

**'model\_year', 'origin', 'name']**

**print(df.shape) # 데이터프레임의 차원을 보여줌**

**# 데이터프레임의 행과 열의 갯수를 출력**

**print("row : {}, columns : {}".format(df.shape[0],9df.shape[1]))**

**df.info() # 데이터의 정보 : 컬럼의 갯수, 각 컬럼의 row갯수 , 데이터 타입**

**print(df.count()) # 전체 칼럼들의 갯수**

**df['origin'].value\_counts() # 같은 값들의 갯수를 출력**

**#df.value\_counts()**

**# 평균값 df.mean() : 전체 컬럼들의 평균**

**#df["컬럼명"] : 해당 컬럼의 평균값**

**통계함수 활용 (# mean, median, min, max, std, corr(열 이름 리스트)**

**df.mean() df["mpg"].mean()**

**df.std() df["mpg"].std()**

**df[["mpg","weight"]].corr()**

**판다스 내장 그래프 도구 활용**

**line(선그래프), bar(수직막대그래프),barth(수평막대그래프),his(히스토그램),**

**box(박스플랜),kde(커널밀도그래프),area(면적그래프),pie(파이그래프),**

**scatter(산점도 그래프),hexbin(고밀도산점도그래프)**

**# 판다스 내장 그래프 활용**

**df= pd.read\_excel("C:/Users/ADMIN/Desktop/python\_src/남북한발전전력량.xlsx")**

**df\_ns= df.iloc[[0,5], 3:]**

**df\_ns.index = ['South','North']**

**df\_ns.columns = df\_ns.columns.map(int) #컬럼명은 int변경**

**print(df\_ns.head())**

**df\_ns.columns**

**# 데이터 전처리 : 컬럼의 데이터 타입 변경, 원하는 정보만 추출**

**#데이터 가공 .NaN 데이터 처리....**

**#df\_ns.plot()**

**df\_ns\_t = df\_ns.T # 데이터프레임의 행과 열을 바꿈**

**df\_ns\_t.plot(kind='line')**

**df\_ns\_t.plot(kind='bar')**

**df\_ns\_t.plot(kind='hist')**

**df\_mpg = pd.read\_csv("./auto-mpg.csv", header=None)**

**df\_mpg.columns = ['mpg','cylinders','displacement',**

**'horsepower','weight', 'acceleration',**

**'model\_year','origin','name']**

**df\_mpg[['mpg','cylinders']].plot(kind='box')**

**df\_mpg.plot(x="weight", y="mpg", kind="scatter")**

**matplotlib**

**import pandas as pd**

**import matplotlib.pyplot as plt**

**plt.rcParams['font.family'] = 'Malgun Gothic'**

**plt.rcParams['axes.unicode\_minus'] = False**

**df = pd.read\_excel("C:/Users/ADMIN/Desktop/python\_src/시도별 전출입 인구**

**#데이터 전처리 작업**

**df = df.fillna(method='ffill') # 전출지 누락된 데이터를 앞 데이터로 채움**

**df.head()**

**#서울에서 다른 지역으로 이동한 데이터만 추출하여서 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**

**df\_seoul.set\_index('전입지', inplace=True) # '전입지 칼럼을 인덱스로 설정'**

**#df\_seoul**

**df\_seoul.plot()**

**seaborn**

**# 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()**