|  |  |
| --- | --- |
| 교육 제목 | **Seaborn 라이브러리 및 데이터 전처리** |
| 교육 일시 | 2021년 11월 10일 수요일 |
| 교육 장소 | 영우글로벌러닝 2층 |
| **교육 내용** | |
| 오전 | 1. Seaborn 라이브러리 2. Matplotlib의 기능과 스타일을 확장한 파이썬 시각화 도구 3. 예, 4. 회귀선이 있는 산점도  * Fig = plt.figure(figsize=(15,5)) : 화면 사이즈 객체 생성 * Ax1=fig.add\_subplot(1,2,1)  화면에 보여지는 그래프 개체1 생성 2행 1열 중 첫번째 * Ax2=fig.add\_subplot(1,2,2)  화면에 보여지는 그래프 개체1 생성 2행 1열 중 두번째 * Sns.regplot(x=’age’, #x축 y=’fare’, #y축 data=titanic, #데이터 ax=ax1) #1번째 그래프 * Plt.show()  1. 히스토그램 /커널 밀도 그래프 (kernel density graph)    1. Displot()    2. 하나의 변수 데이터의 분포를 확인할 때 2. 히트맵 heatmap()    1. 2개의 범부형 변수를 각각 x, y축에 넣고 데이터를 매트릭스 형태로 분류 3. 범주형 데이터의 산포도 4. Stripplot() & swampplot() 5. 막대 그래프 barplot() 6. 빈도 그래프 countplot() 7. 박스 플롯/ 바이올린 그래프 : Boxplot(), violinplot() 8. 조건을 적용하여 화면을 그리드로 분할하기 |
| 오후 | 1. Folium 라이브러리 – 지도 활용 2. 데이터 전처리 과정    1. 결측치 처리    2. 중복데이터 처리    3. 데이터형 변환 (구간 분할, np.histogram(), pd.cut())    4. 원핫인코딩 (더미변수, sklearn.preprocessing)    5. 정규화 3. 결측치 처리    * 1. 누락데이터 확인      2. Info(), value\_counts(), isnull(), notnull(),      3. 누락데이터 제거         1. Dropna(axis=1, thresh =500) 🡪 NaN 값을 500개 이상 갖는 모든 열을 삭제         2. Df.dropna(subset=[‘age’, how =any, axis=0] 🡪 age열의 행 중에서 NaN 값이 있는 모든 행을 제거      4. 누락데이터 치환      5. 누락데이터가 NaN으로 표시되지 않는 경우         1. Import numpy as np         2. Df.replace(‘?’, np.nan, inplace= True)    1. 중복데이터 처리       1. 중복 데이터 확인 duplicated(),    2. 데이터 표준화       1. 데이터 포맷의 일관화       2. 단위 환산       3. 자료형 변환    3. 데이터형 변환 (구간 분할, np.histogram(), pd.cut())    4. 범주형 데이터 처리       1. 원핫인코딩 (더미변수, sklearn.preprocessing)    5. 정규화 normalization       1. 각 열(변수)에 속하는 데이터 값을 동일한 크기 기준으로 나눈 비율로 나타내는 것 4. 시계열 데이터 5. Time series data 6. 주식, 금융 데이터 7. Timestamp(), Period |