|  |
| --- |
| **SSAC**  **파이썬을 활용한 비즈니스 빅데이터 분석가 양성 과정**  **(TEMPLATE) 자가학습 워크시트** |

|  |  |
| --- | --- |
| **이름** | 백승재 |
| **날짜** | 2021-10-31 |
| **온라인 콘텐츠(인강)**  **학습 범위** | * 직장인을 위한 Python : 전처리02.12 ~03.06, Part1: 시각화 라이브러리 ch01\_01~03 밑그림 그리기(단일그래프, 다중그래프) |
| **복습 노트 작성 (자유 양식)** | Ch 02-전처리-12.원핫인코딩의 개념과 get\_dummies  Ch 02-전처리-13~14.부동산 데이터로 데이터 분석 실습  - 전처리 : 공백 제거 및 처리-> NAN값 처리 -> , 대체 -> -처리 -> NAN값처리  - 문자 타입 변경 : object -> int  - 피벗 테이블 생성-  Ch 03.Pandas 시각화-01.데이터 시각화의 목적  - 가장 직관적인 시각적인 정보를 얻기 위함  - 탐색적 데이터 분석(EDA) : 다양한 각도에서 관찰하고 이해하는 과정  Ch 03.Pandas 시각화-02.colab 한글폰트 깨짐현상 해결(시각화)~ 03.Plot,line,bar,histogram  Step1: df 인덱싱  Step2: df 인덱싱.plot(kind=””)  Ch 03.Pandas 시각화-04. Kde,hexbin  - kde (히스토그램과 유사하게 밀도를 보여주는 그래프): df[“분양가”].plot(kind = “kde”)  - hexbin(고밀도 산점도 그래프) : df.plot(kind=’hexbin’, x = ‘분양가’, y=’연도’, gridsize = 20)  Ch 03.Pandas 시각화-05.boxplot: outlier 감지할 때 많이 사용, IQR(사분위 범위), box max,min df\_seoul["분양가"].plot(kind="box")  Ch 03.Pandas 시각화-06.are,pie,scatter  Area (line에 하한 음영처리) : df.groupby('월')['분양가'].count().plot(kind='area')  Pie (데이터의 비율 확인): df.groupby("연도")["분양가"].count().plot(kind="pie")  Scatter(numeric, 데이터 분포도) df.plot(kind="scatter",x="월",y="분양가")  Part1: 01~03 단일 그래프, 다중 그래프  plt.figure() : 새로운 그래프 canvas를 생성  plt.show() : 그래프를 보여줌 |
| 퀴즈 제출 | Q1. 다음은 df\_seoul[“분양가”]를 boxplot으로 시각화한 것이다. 그래프 해석에 대해 적절하지 않은 것은?    1) box plot 은 다음 코드에 출력물에 해당한다.  df\_seoul["분양가"].plot(kind="box")  2) 박스 플럿의 min값은 1Q - IQR이다.  3) Df\_seoul의 max값은 박스 플럿의 max와 같다.  4) 사분위범위는 3Q-1Q이다.  정답 3 : DF의 max값은 박스 플럿의 max(상한치)와 다르다.  Max: 2020  박스플럿의 max : 3Q+IQR = 9550.375  Q2.random함수를 이용하여 만든 두개의 표준 정규 분포(range : -1~1)를 이루는 array를 육각형의 분포로 시각화한 것은? (n = 10000)  1) df = pd.DataFrame({'x': np.random.randn(n),'y': np.random.randn(n)})  df.plot(kind="kde")  2) df = pd.DataFrame({'x': np.random.rand(n),'y': np.random.rand(n)})  df.plot.hexbin(x='x', y='y', gridsize=20)  3) df = pd.DataFrame({'x': np.random.randn(n),'y': np.random.randn(n)})  df.plot.hexbin(x='x', y='y', gridsize=20)  4) df = pd.DataFrame({'x': np.random.randn(n),'y': np.random.randn(n)})  df.plot.hist()  정답 : 3  -1<x,y<1사이의 표준 정규분포의 난수로 array-> randn함수를 사용  Array를 육각형의 분포로 시각화 -> hexbin plot사용    Q3. 다음 코드에 대한 설명으로 적절한 것은?  data = np.arange(1, 51)  plt.plot(data)  data2 = np.arange(51, 101)  plt.figure()  plt.plot(data2)  plt.show()  1) 다중 그래프로 2개의 그래프가 1개의 canvas에 표시된다.  2) plt.figure() : 새로운 캔버스에 data2를 출력한다.  3) plt.show()는 그래프를 출력하는 함수로 해당 코드가 없으면 그래프가 출력이 안된다.  4) data1과 dat2에 대해 곡선 형태의 그래프가 출력된다.  정답 : 2  1) 다중 그래프가 개별 canvas에 출력된다.  3) show함수가 없어도 그래프가 출력된다.  4) 직선 형태의 그래프가 출력된다. |
| **교육생 사진 (교육생 모습 + 교육 화면)** |  |
| **화면 캡쳐**  **(교육 화면 + 교육일시)** |  |