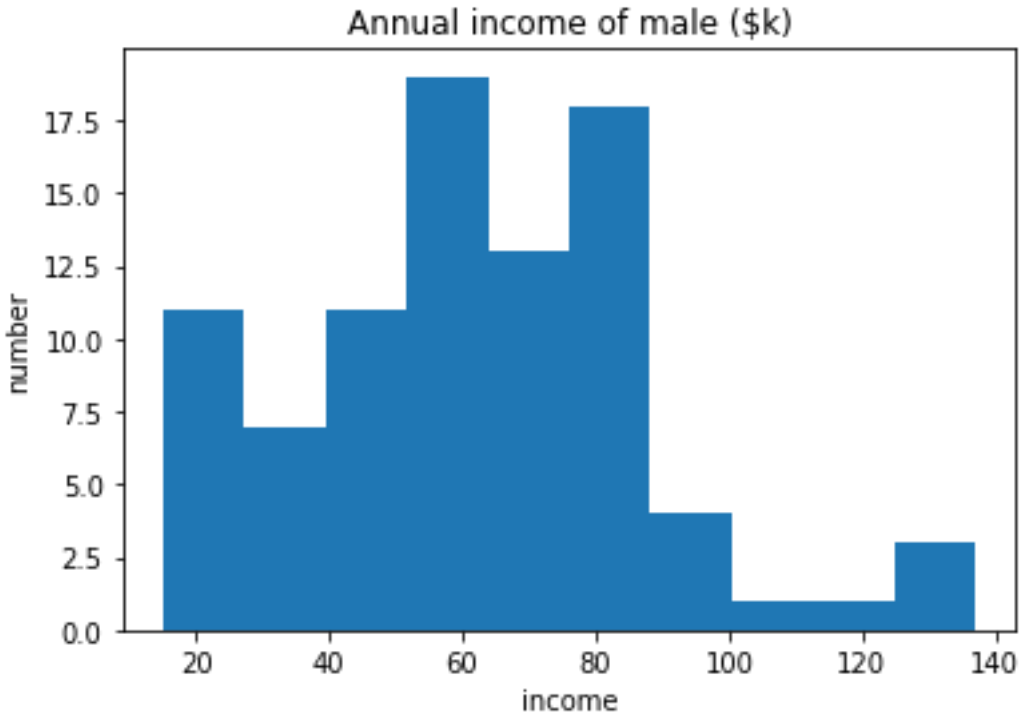
SPDS HW#7 보고서

2020-23844

김장호

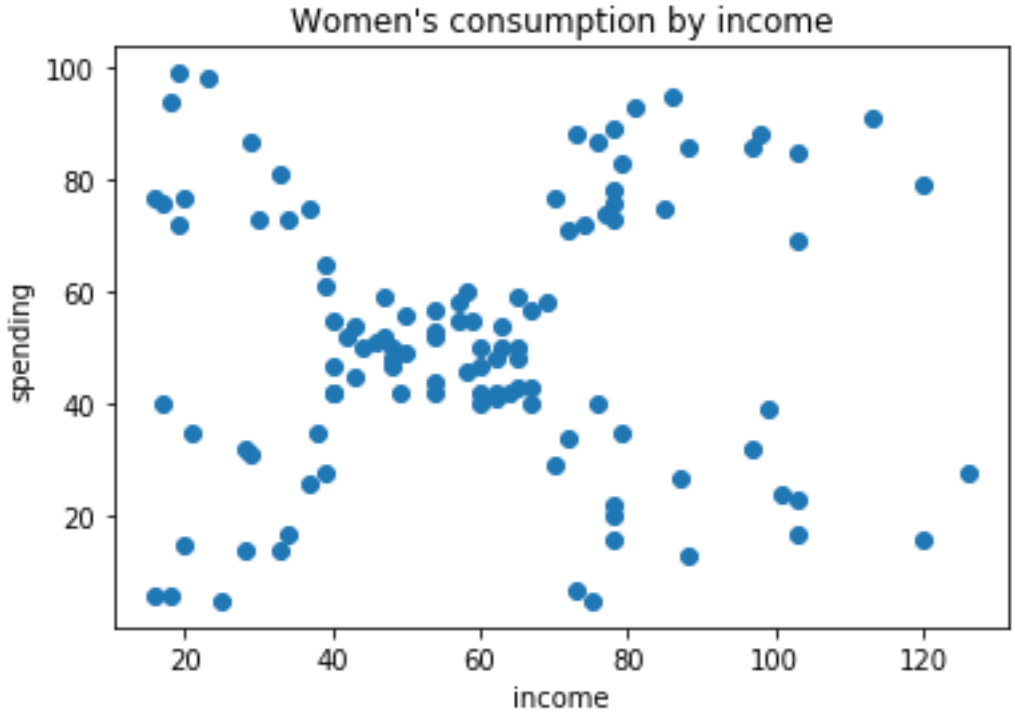
이번 과제에서는 numpy, scipy 그리고 matplotlib 모듈 외에 dataset.csv파일을 읽기 위해 pandas 라이브러리를 사용하였다.

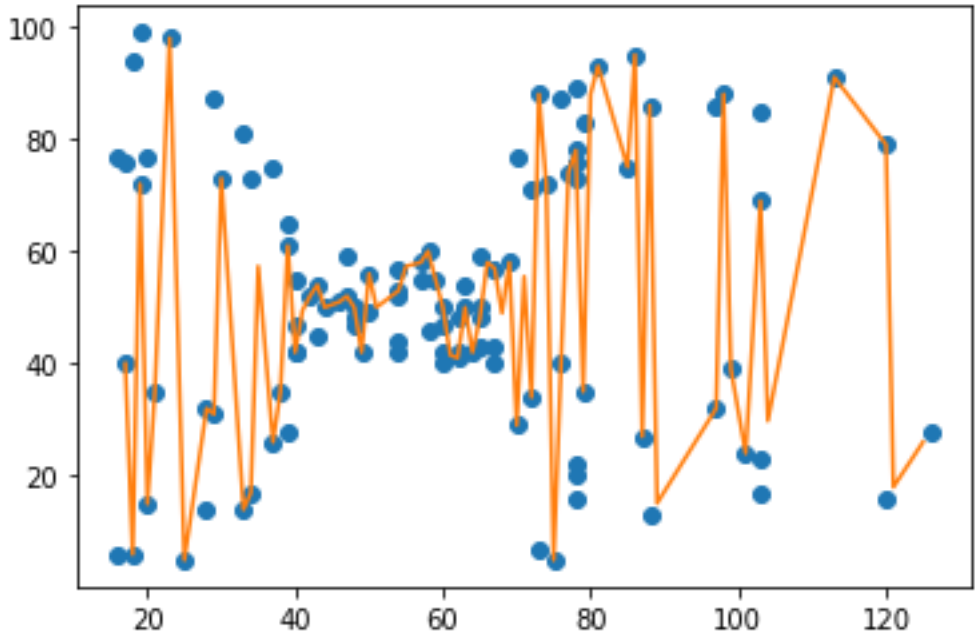
1. 남성 고객의 수입분포에 대한 히스토그램



우선 불러온 csv파일에서 성별이 남성인 행들만을 가지고 별개의 data_male 데이터프레임을 만든다. 그 이후 4열(인덱스는 3)의 수입들을 가지고 matplotlib 모듈의 hist() 기능을 통해 히스토그램으로 시각화하였다.

2. 여성고객의 수입 수준에 따른 소비와 1-D interpolation

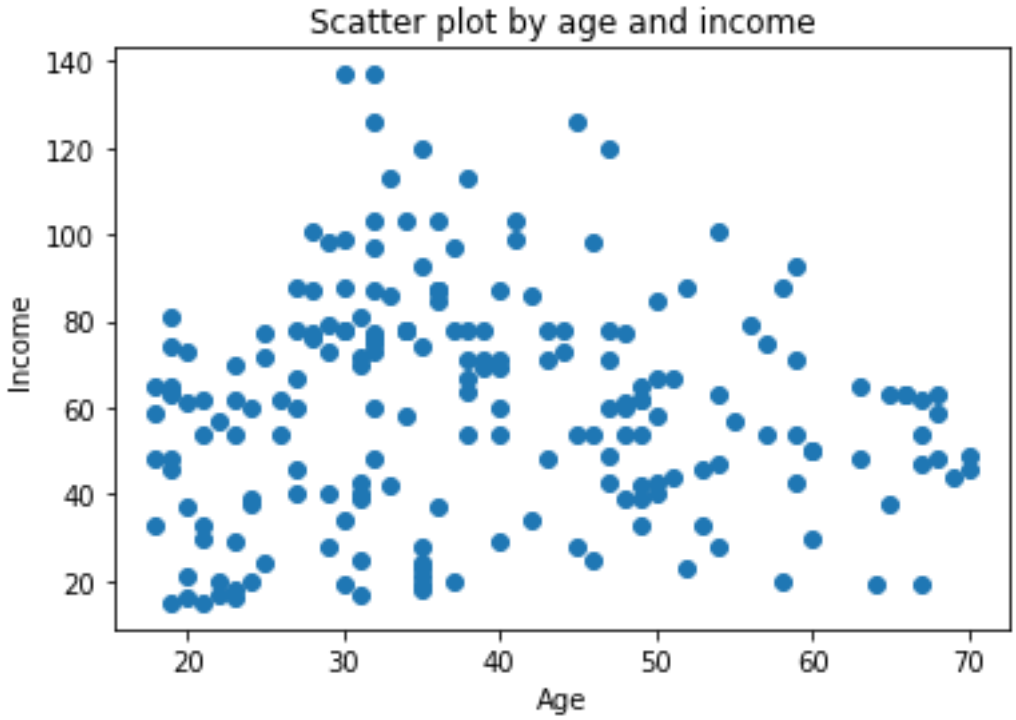


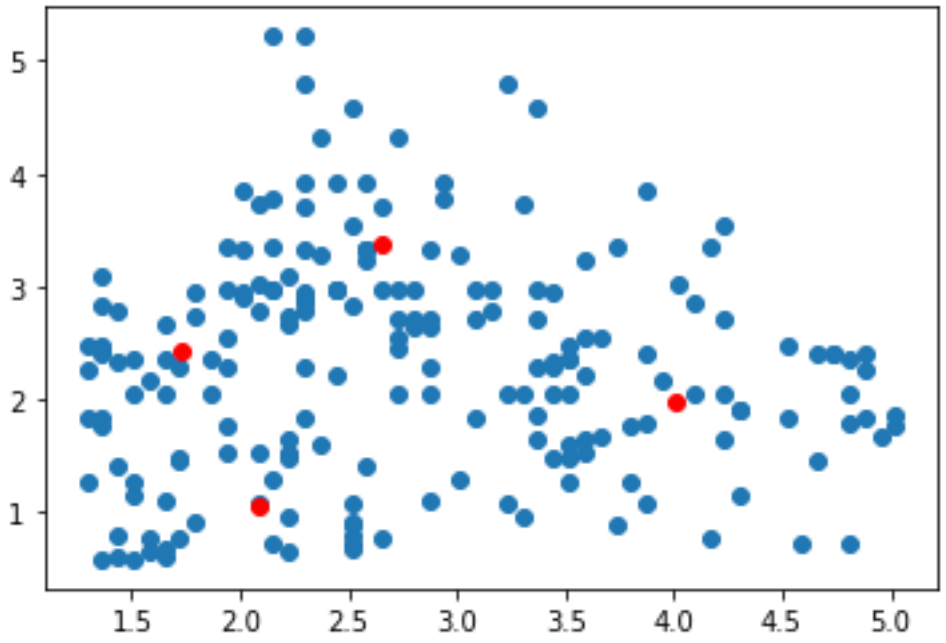


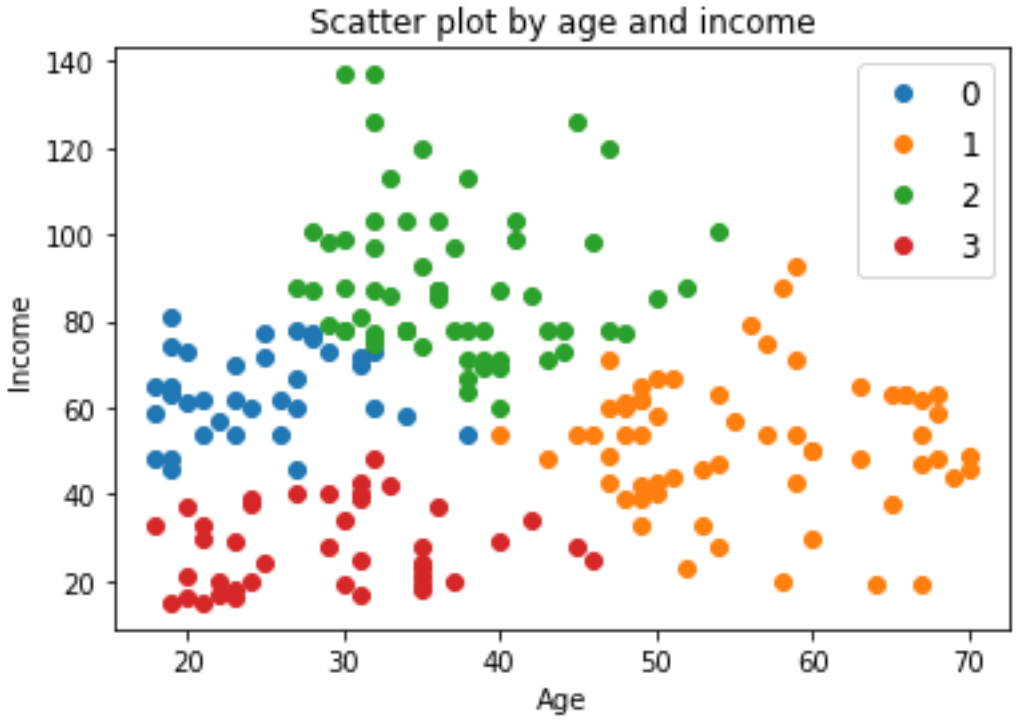
1번과 마찬가지로 여성만을 골라 data_female 데이터 프레임을 만든 후, 4열의 수입을 x로, 5열의 소비지수를 y로 뽑는다. X, Y를 각축으로 하는 산포도를 plt.scatterplot()으로 만든다. 이후 scipy의 interpolate 모듈을 통해 1-D interpolation을 진행하여 x에 대한 y의 함수를 추론하고 이를 plotting을 통해 기

존 산포도에 추가한다.

3. 전체고객의 나이, 수입에 대한 산포도와 K means clustering 결과







기존 전체 데이터에서 나이와 수입에 해당하는 열을 추출하여 산포도를 그린다. 이 두열을 합친 부분을 데이터에서 뽑아와 numpy array로 바꿔주고 scipy cluster모듈의 white 함수를 사용하여 각 변수를 정규화시킨다.

그리고 kmeans 함수를 통해 이 표준화된 값들을 4개의 그룹으로 군집화한다. 6번의 반복 계산을 통해 이 군집화의 기준이 되는 distortion이 가장 작아지는 (Euclidean distance의 합이 가장 작아지는) 4개의 기준점을 빨갛게 표시했다. (codebook이 이 네 점의 위치이다) vq 함수는 그리고 이 codebook의 네 점을 기준으로 정규화된 데이터를 라벨링한다(0,1,2,3).

이 라벨들을 기존데이터에 붙이고 matplotlib을 통해 그룹별 색깔을 달리하여 산포도를 그린다.