**기계학습 및 데이터마이닝 과제1**

이름: 김도운

학과: 사이버보안학과

학번: 201520893

**문제 1. One-hot encoding 및 결측값 처리하기**

****

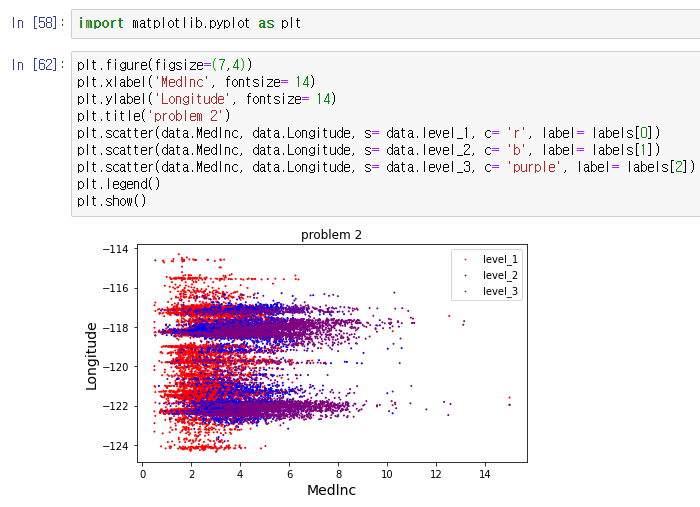
먼저 매개변수로 전달받은 numeric변수의 데이터를 갖는 target을 cut 함수를 이용해 각 범위마다 각 label을 가지는 category 변수로 바꾸었다.

범위 밖의 결측값 제거를 위해 dropna()함수를 이용했다. 결측값이 나타난 행을 제거하면 되기 때문에 따로 매개변수로 설정해준 것은 없다.

범주화 된 target 데이터를 각 범위별로 0과 1로 나타내기 위해 get\_dummies()함수를 사용했다.

마지막으로 target 데이터는 현재 전체 데이터가 아닌 열이 3개인 데이터이다. 따라서 전체 데이터 df와 target을 합쳐야 되는데, 초반에 join함수를 사용해 시도했지만 결측값이 다시 복구되는 현상으로 인해 concat함수를 사용했다. 여기서 두개의 데이터프레임 df, target을 옆으로 합치기 위해 axis=1로 설정했고, join=”inner”로 설정함으로써 교집합으로 합쳤다. Outer로 설정하면 합집합으로 합치기 때문에 df 데이터프레임으로 인해 다시 결측값이 생기게 된다.

**문제 2. 문제 1에서 one-hot 인코딩 기반으로 생성된 dataframe을 바탕으로 scatter plot을 만들고, 이를 기반으로 얻을 수 있는 사실을 서술하시오.**

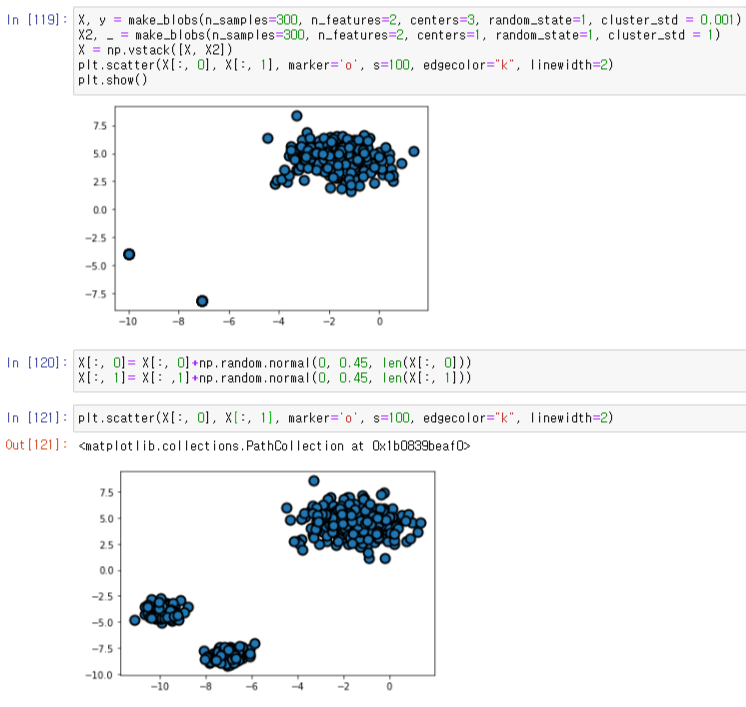
****

데이터 시각화 기법인 scatter plot으로 나타내면 시각적으로 받아들이기 편하다. 원하는 x축 y축을 설정하고 축에 따른 데이터들의 분포도를 한 눈에 볼 수 있다. x, y축 데이터를 각각 소득의 중앙값과 행정 구역의 경도로 설정했고, 각 구간별로 보기 위해 s 인자를 level 데이터로 설정했다.

이를 통해 알 수 있는 사실은 Medlnc(행정 구영 내 소득의 중앙값)이 비교적 낮은 구간에 경도에 상관없이 전반적으로 빨간색 점들이 많은 것을 보아 Target(행정구역 내 주택 가격의 중앙값)이 낮다는 것을 알 수 있다. 또한 경도가 -118, -122인 곳은 소득 중앙값에 상관없이 보라색 점들이 많은 것을 보아 주택 가격이 높다는 것을 알 수 있다. 파란 점들의 분포를 보면 경도가 -118, -122 근처인 부근의 주택 가격이 중간 정도 한다는 것도 알 수 있다.

정리하자면 낮은 소득의 경우 경도에 관계없이 주택가격이 낮고, 특정 위치에서는 소득에 관계없이 주택 가격이 중간 이상인 것을 알 수 있다.

**문제 3. jittering을 주고 이러한 jittering이 필요한 경우에 대해서 서술하시오.**

jittering 효과를 주기 위해 X의 각 열에 난수를 더해주었다. 난수를 생성하는 np.random.uniform함수 등 다른 함수들을 이용해 난수들을 더해줄 수도 있었지만, 모양이 너무 네모 반듯하게 나와 정규분포의 난수를 생성하는 normal함수를 이용했다.

이러한 jittering 효과는 여러 데이터들이 같은 값을 가짐으로써 그래프 상에서 하나의 점으로 나타나는 필요하다. 이렇게 하나의 점으로 표현되는 경우, 그래프로부터 데이터의 정확한 정보를 얻을 수 없다. 따라서 같은 값을 가지더라도 약간의 오차를 주어서 위와 같이 살짝 겹치게 보이도록 만들어 데이터를 쉽게 파악하고 정확하게 전달할 수 있다.