Coding Homework1

학과 : 소프트웨어학과

학번 : 201820742

이름 : 고예준

[문제1] one-hot encoding 및 결측값 처리하기

<make\_onehot함수>

스크린샷이(가) 표시된 사진

자동 생성된 설명

<생성된 dataframe>

컴퓨터이(가) 표시된 사진

자동 생성된 설명

대형, 컴퓨터이(가) 표시된 사진

자동 생성된 설명

[문제2]

<문제 1에서 one-hot 인코딩 기반으로 하여 만들어진 dataframe 바탕으로 만든 scatter plot>

스크린샷이(가) 표시된 사진

자동 생성된 설명

<이를 기반으로 알 수 있는 내용>

우리는 캘리포니아의 각 행정 구역 내 주택의 중앙값으로 저렴한 편의 주택, 중간 주택, 비싼 가격의 주택의 전체적인 특성에 대해 알기 위하여 1번의 dataFrame을 완성하였다. 이를 기반으로 만든 scatter plot의 결과를 살펴보면 다음과 같다.

빨간색 동그라미인 저렴한 편의 주택(level\_1)은 행정구역 내 소득의 값인 MedInc값(x값)이 (0~4) 값에 분포되어 있어 상대적으로 왼쪽에 분포되어 있다. 이를 통해, 저렴한 편의 주택에서는 행정 구역 내 소득의 중앙값이 낮은 곳에 위치한다고 할 수 있다. Longitude의 값(y축 값)은 -114도부터 -124도까지 많이 퍼져있는 결과를 통하여 저렴한 편의 주택은 다양한 경도에 분포되어 있음을 알 수 있다.

파란색 동그라미인 중간 주택(level\_2)는 행정구역 내 소득의 값인 MedInc값(x값)이 (3~6) 값에 분포되어 있어 상대적으로 중간에 분포되어 있다. 이를 통해, 중간 주택에서는 행정 구역내 소득의 중앙값이 보통인 곳에 주로 분포한다고 할 수 있다. 중간 주택은 Longitude의 값(y축 값)이 약 -116도부터 약 -118도까지, 약 -120도부터 약 -123도의 경도에 많은 중간 주택이 있음을 알 수 있다.

보라색 동그라미인 비싼 주택(level\_3)는 행정구역 내 소득의 값인 MedInc 값(x값)이 (3-9) 값에 분포되어 있어 행정 구역 내 소득이 보통인 곳에서 높은 편인 곳에 비싼 주택이 주로 있다는 것을 알 수 있다. 또한, longitude의 값(y축 값)은 약-117도에서부터약 -119도까지, 그리고 약 -122도 근처에 비싼 주택이 주로 분포되어 있음을 알 수 있다.

정리하자면, 저렴한 편의 주택은 행정 구역 내 소득의 중앙값이 낮은 곳에 주로 위치하고, 다양한 경도에 분포되어 있고, 중간 주택은 행정 구역 내 소득의 중앙값이 저렴한 편의 주택과 비싼 주택 사이인 행정구역에 위치하고, 주로 약 -116도부터 약 -118도까지, 약 -120도부터 약 -123도의 경도에 위치하고 있다. 마지막으로, 비싼 주택은 행정 구역 내 소득의 중앙값이 상대적으로 보통에서 높은 편인 곳인 행정구역에 주로 위치하고, 약 -117도에서부터 약 -119도까지, 그리고 약 -122도 근처에 비싼 주택이 주로 분포되어 있다.

[문제3]

<데이터 X를 jittering 준 결과>

스크린샷이(가) 표시된 사진

자동 생성된 설명

<jittering이 필요한 경우>

데이터가 몰려있는 경우, 데이터를 더 풍부하게 하기 위하여 jittering이 필요하다. Jittering은 정수형 데이터인 경우 또는, 데이터가 많이 겹치는 경우에 많이 이용한다. 이는 데이터가 겹쳐져 그려지는 것을 방지하기 위하여 임의의 noise를 주는 기법이다.

따라서, Jittering은 눈에 더 잘 들어오게 하고, 데이터를 풍부하게 한다는 장점이 있다.