머신러닝20220120

**Basic Data Exploration**

Using Pandas to Get Familiar With Your Data

머신러닝 프로젝트의 첫 번째 단계는 데이터를 숙지하는 것.

pandas library를 이용하면 됨. pandas는 과학자들이 데이터를 탐색하고 조작하기 위해 사용하는 주요 도구이다. 대부분의 사람들은 코드에서 pandas를 pd로 줄인다. 우리는 명령을 사용하여 이것을 한다.

import pandas as pd

pandas library의 가장 중요한 부분은 데이터프레임. DataFrame은 테이블로 생각할 수 있는 유형의 데이터를 보관. 이것은 Excel의 시트 또는 SQL 데이터베이스의 테이블과 유사.  
  
pandas는 여러분이 이런 종류의 데이터로 하고 싶어하는 대부분의 것들을 위한 강력한 방법들을 가지고 있다.  
  
예를 들어, 우리는 호주 멜버른의 집값에 대한 데이터를 살펴볼 것. 실습에서는 동일한 프로세스를 아이오와에서 집값이 책정된 새 데이터 세트에 적용.  
  
예제(Melbourne) 데이터는 파일 경로 ../input/melbourne-housing-snapshot/melb\_data.csv에 있다.  
  
다음 명령을 사용하여 데이터를 로드하고 탐색.

*# save filepath to variable for easier access*

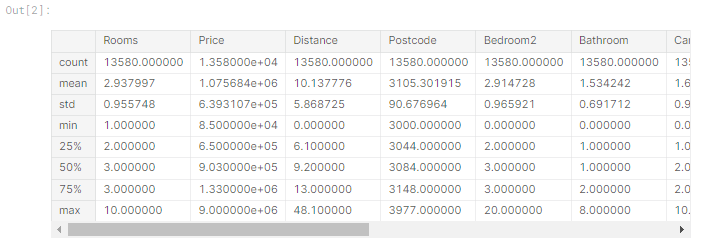
melbourne\_file\_path = '../input/melbourne-housing-snapshot/melb\_data.csv'

*# read the data and store data in DataFrame titled melbourne\_data*

melbourne\_data = pd.read\_csv(melbourne\_file\_path)

*# print a summary of the data in Melbourne data*

melbourne\_data.describe()



데이터 설명 해석  
결과에는 원래 데이터 집합의 각 열에 대해 8개의 숫자가 표시. 첫 번째 숫자인 카운트는 결측값이 아닌 행의 수를 나타냄.  
  
결측값은 여러 가지 이유로 발생. 예를 들어, 1개의 침실 주택을 조사할 때 2번째 침실의 크기는 수집되지 않음. 누락된 데이터에 대한 주제로 다시 돌아감.  
  
두 번째 값은 평균이다. 그 아래 std는 표준 편차이며, 값이 수적으로 얼마나 퍼져 있는지 측정함.  
  
최소값 25%, 50%, 75% 및 최대값을 해석하려면 각 열을 가장 낮은 값에서 가장 높은 값으로 정렬한다고 가정. 첫 번째(가장 작은) 값은 최소값. 목록의 4분의 1로 넘어가면 값의 25%보다 크고 값의 75%보다 작은 숫자가 나옴. 이는 25%의 값. 50번째와 75번째 백분위수는 유사하게 정의되며, 최대값이 가장 큰 숫자이다.