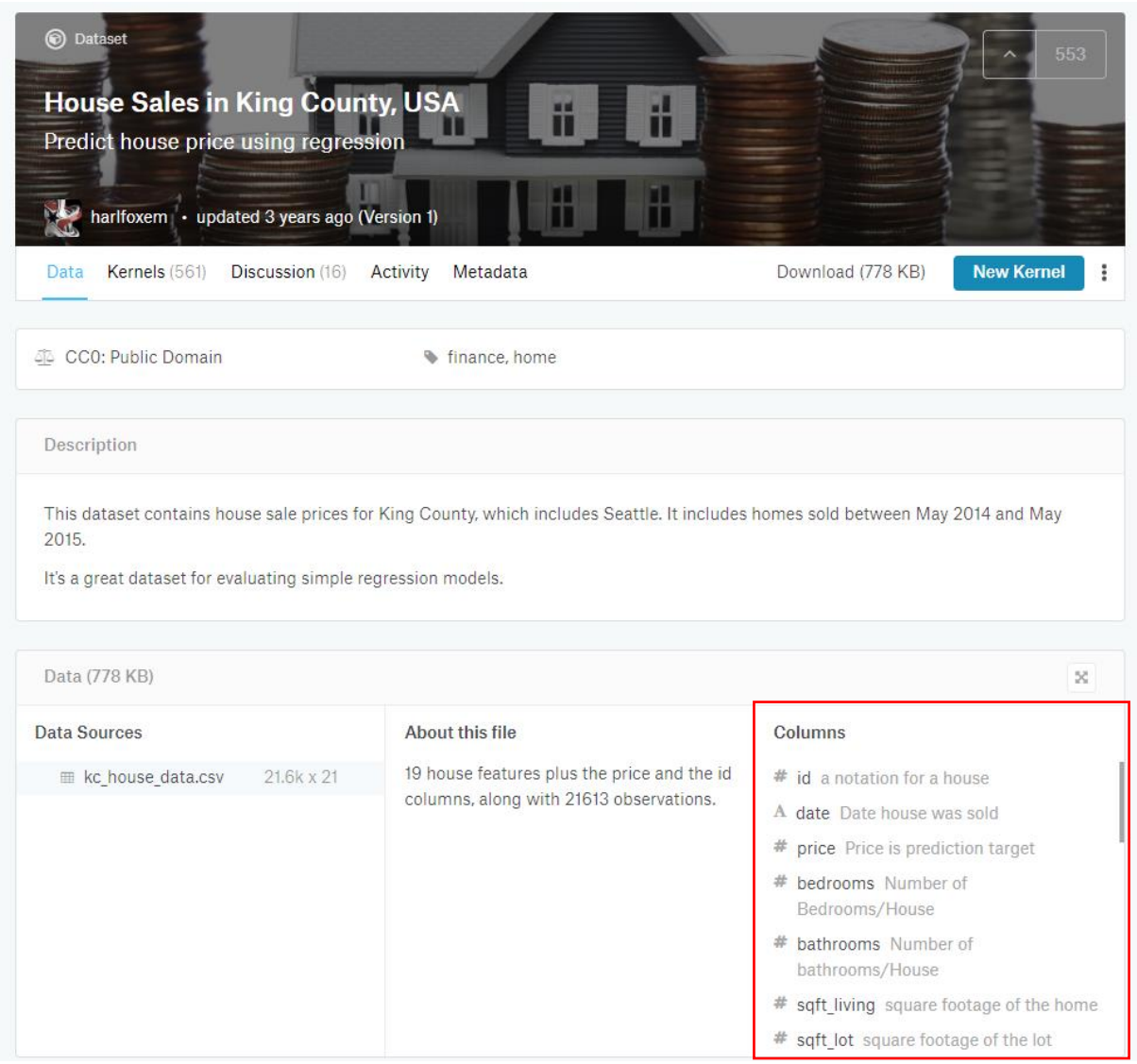
Multivariate Data Analysis Assignment #1

Multivariate Linear Regression (MLR)

Dataset: House Sales in King County, USA, kc_house_data.csv

(https://www.kaggle.com/harlfoxem/housesalesprediction)

해당 데이터셋은 미국 King Coynty에서 2014년 5월부터 2015년 5월까지 거래된 주택들에 대한 정보 및 가격이 포함되어 있다. 각 변수에 대한 설명은 제공된 URL에서의 Column 항목을 통해 확인할 수 있다. 이 중 세 번째 항목인 price가 MLR 모형의 target variable이다.



[Q1] MLR 모형 구축을 위해 필요하지 않은 변수는 어떤 것들이 있는가? 왜 그렇게 생각하는가?

다음 물음에 대해서는 [Q1]에서 선택한 변수들은 제외하고 답변하시오.

[Q2] 개별 입력 변수들에 대하여 각각 다음과 같은 단변량 통계량을 계산하고 Box plot을 도시하시오: Mean, Standard deviation, Skewness, Kurtosis. 전체 변수 중에서 정규분포를 따른다고 할 수 있는 변수들은 몇 개인가? 정규분포를 따른다고 가정한 근거는 무엇인가?

[Q3] [Q2]의 Box plot을 근거로 각 변수들에 대한 이상치(너무 값이 크거나 작은 객체) 조건을 정의하고, 해당하는 객체들을 데이터셋에서 제거해 보시오.

다음 각 물음에 대해서는 [Q3]에서 제거된 객체들을 제외시킨 데이터프레임을 사용하여 답하시오.

[Q4] 가능한 모든 두 쌍의 입력 변수 조합에 대한 산점도(scatter plot) 및 Correlation plot (hint: "corrplot" 패키지의 corrplot( ) 함수 사용) 상관관계를 계산해 보시오. 어떤 두 조합의 변수들이 서로 강한 상관관계가 있다고 할 수 있는가?

[Q5] 전체 데이터셋을 70%의 학습 데이터와 30%의 테스트 데이터로 분할한 후 모든 변수를 사용하여 MLR 모델을 학습해 보시오. Adjusted R^2값을 통해 데이터의 선형성(linearity)을 판단해 보시오. Residual plot과 Q-Q Plot을 도시하고 Ordinary Least Square 방식의 Solution이 만족해야 하는 가정들이 만족될만한 수준인지 정성적으로 판단해 보시오.

[Q6] 유의수준 0.01에서 모형 구축에 통계적으로 유의미한 변수들은 어떤 것들이 있는가? 해당 변수들은 Price와 양/음 중에서 어떤 상관관계를 갖고 있는가?

[Q7] Test 데이터셋에 대하여 MAE, MAPE, RMSE를 계산하고 그에 대한 해석을 해 보시오.

[Q8] 만약 7개의 입력 변수만을 사용하여 모델을 구축해야 할 경우 어떤 변수들을 선택하겠는가? [Q4]와 [Q6]의 답변을 바탕으로 본인이 선택한 변수들에 대한 근거를 제시하시오.

[Q9] [Q8]에서 선택한 변수들만을 사용하여 MLR 모형을 다시 학습하고 Adjusted R^2, Test 데이터셋에 대한 MAE, MAPE, RMSE를 산출한 뒤, 두 모형(모든 변수 사용 vs. 7개 변수 선택)을 비교해 보시오.

[Extra Question] 이 외 해당 데이터셋을 통해 MLR 관점에서 가능한 추가적인 분석을 웹에서 검색해서 수행하고 그 결과를 해석해 보시오.