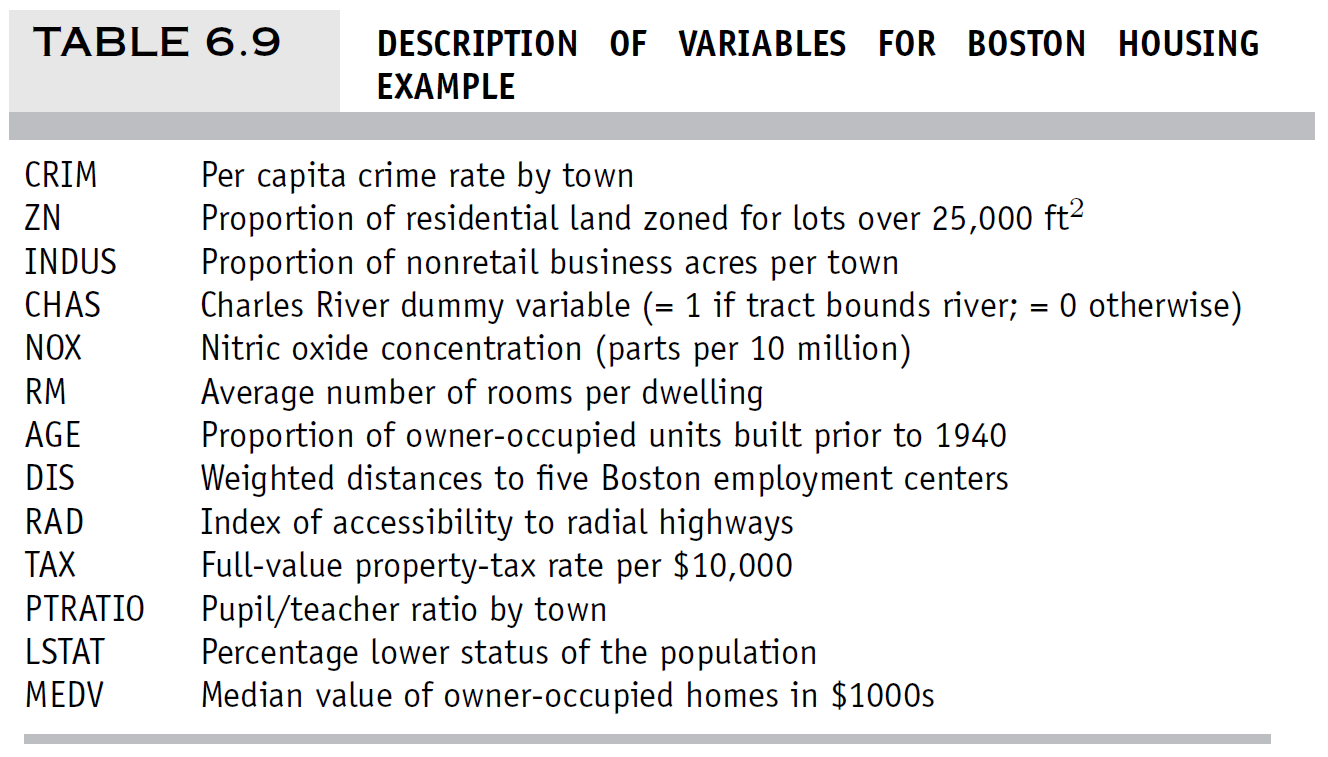
**[여덟 번째 과제]**

**IT응용시스템공학과**

**1494053**

**김희택**

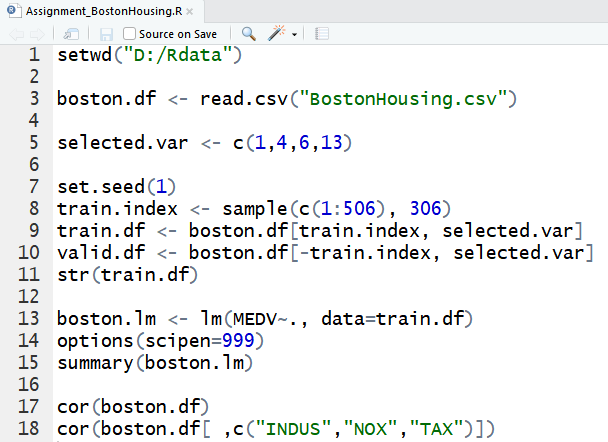
**6.1 보스턴 주택가격 예측 BostonHousing.csv 파일은 미국 인구통계조사국에 의해 수집된 매사추세츠주 보스턴 지역의 주택정보를 담고 있다. 이 데이터세트는 보스턴 지역의 506개의 인구통계조사 주택단지 정보를 포함한다. 데이터 수집의 목표는 범죄율, 공해, 방의 개수에 대한 정보에 근거하여 새로운 주택단지의 주택가격의 중앙값(median)을 예측하는 것이다. 데이터세트는 13개의 예측변수와 반응(목표)변수인 주택가격의 중앙값(MEDV)을 포함한다. <표6.9>는 예측변수들과 반응변수를 설명하고 있다.**

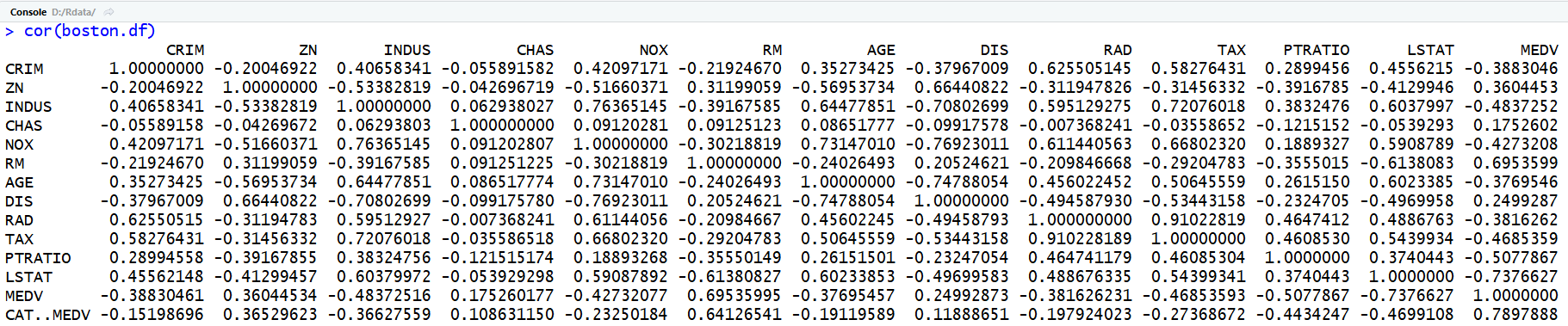


**d. 예측 변수의 수를 감소하시오.**

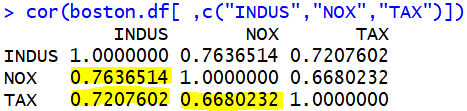
1. **13개의 예측변수들 중에서 어떠한 예측변수들이 동일한 것을 측정할 것 같은가?**

**INDUS, NOX, TAX 간의 관계에 대하여 논하시오.**





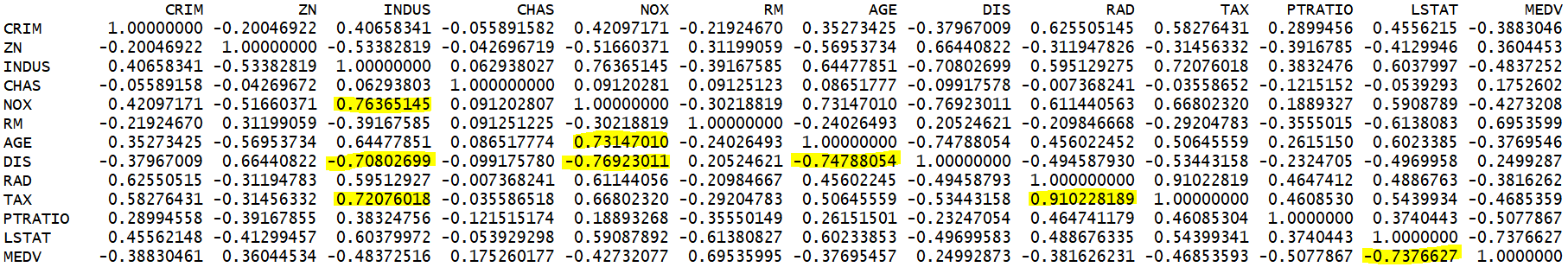
* 위 사진은 boston.df에 있는 모든 변수 사이에 상관계수를 구한 결과입니다.
* 전체 수치를 확인하면, **TAX와 RAD사이의 상관계수가 0.910228189로 가장 큰 값인 것을 알 수 있습니다.**
* 즉 **TAX와 RAD** 둘 사이에 밀접한 관계가 존재하여, 13개의 예측변수들 중 동일한 것을 측정할 것 같다고 말할 수 있습니다.



* 위 사진은 INDUS, NOX, TAX 사이에 상관계수를 구한 결과입니다.
* 즉 시별 비상업지역의 비율**(INDUS)**이 클수록 10ppm당 일산화질소**(NOX)**와 10,000달러당 종합재산세율**(TAX)**이 높은 것을 알 수 있습니다.

1. **12개의 수치형 예측변수들에 대한 상관관계 표를 계산하고 높은 상관관계를 갖는 변수의 쌍을 찾으시오. 이들은 잠재적으로 중복성을 가지며 다중공선성의 원인이 될 수 있다.**

**상관관계 표를보고 어떤 변수들을 제거할지 고르시오.**



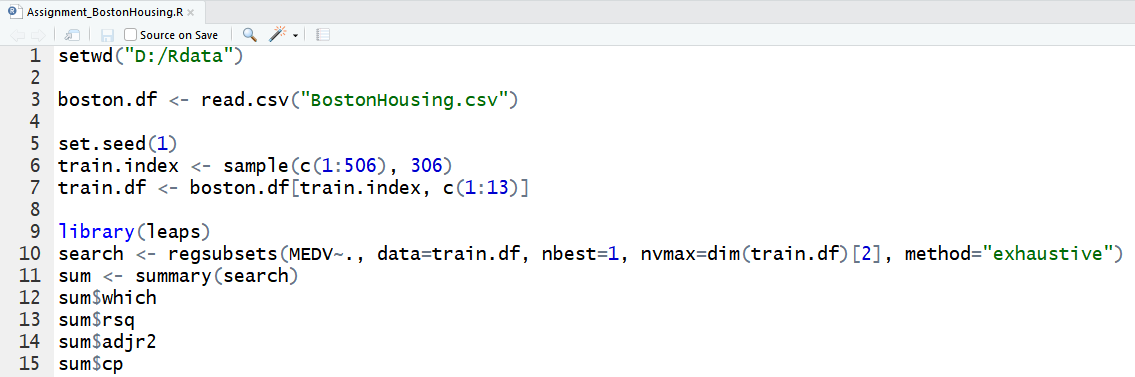
* 높은 상관관계를 갖는 변수의 쌍은 다음과 같습니다.

**2. (NOX, INDUS)** / (DIS, INDUS) / (TAX, INDUS) / (AGE, NOX) / (DIS, NOX) / (DIS, AGE) / 1. **(TAX, RAD)** / (MEDV, LSTAT)

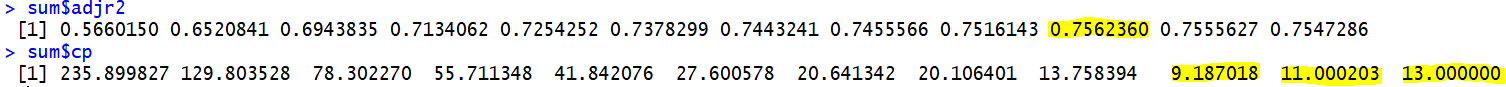
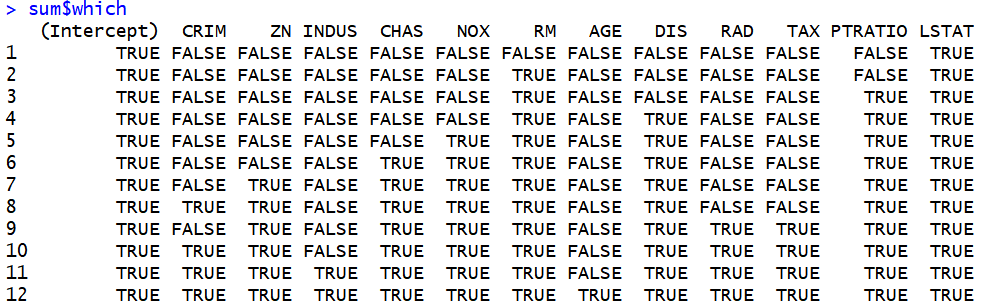
* ~~위 쌍을 이루는 각 예측변수들의 횟수와 상대적으로 큰 값의 상관관계 수치를 고려해보면,~~ **~~INDUS, NOX, AGE~~**~~를 제거할 수 있다고 생각합니다.~~ 둘 중에 하나만 없애는 것이다. 둘 다 없애면 안 된다. 위 진하게 된 쌍의 각 하나씩 아무거나 제거해야 됨. 예를 들어 1에서 TAX, 2에서 NOX를 제거하도록 해보자.

1. **나머지 예측변수들을 감소시키기 위해서 전역 탐색방법을 사용하시오**

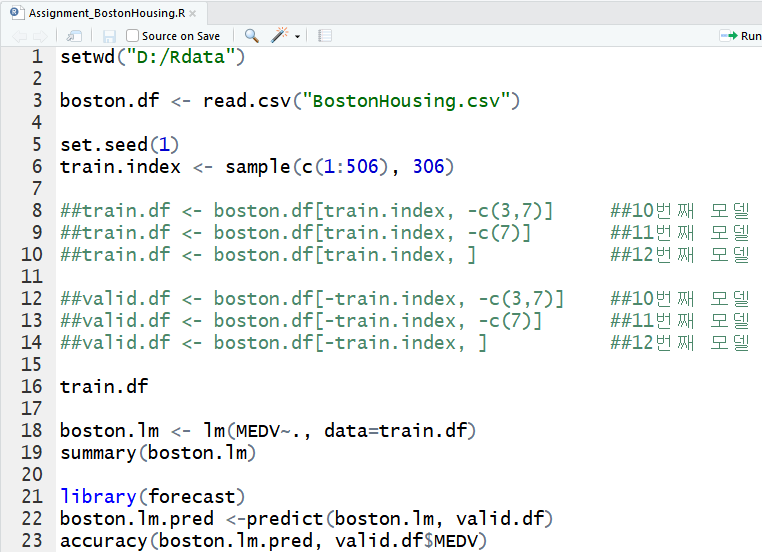
**첫째, 성능이 좋은 상위 3개 모델을 선택하시오. 그 다음으로 이 모델들을 학습 세트에 대하여 각각 구축한 후, 검증 세트에 대하여 예측정확도를 비교하시오. RMSE, 평균오차, 리프트 도표를 사용하여 비교하시오. 마지막으로 가장 좋은 모형을 선택하고 이에 대하여 기술하시오.**



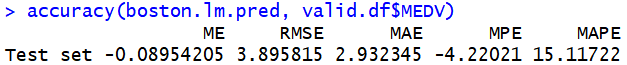
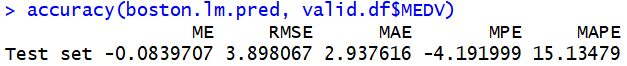
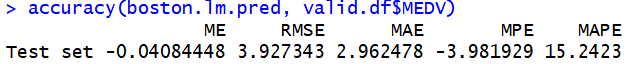
* 나머지 예측변수들을 감소시키기 위해서 전역 탐색방법을 사용하였습니다.



* 위 결과를 확인하면 는 10개의 예측변수를 사용할 때까지 증가하다가 그 후 안정되는 것을 알 수 있습니다. 또한, p가 작으면서 Mallow’s 일 때 좋은 모델이므로, 10, 11, 12번째 모델을 포함하여 상위 3개 모델을 선택할 수 있습니다. 변수가 하나 더 증가하여 복잡해지기는 하지만 설명력은 감소한다(10을 기준으로 뒤 부분). 즉 10을 기준으로 그 전 8, 9, 10번째 모델이 상위 모델 가능성이 있다.



* 위 10, 11, 12번째 모델을 학습 세트에 대하여 각 각 구축한 후, 검증 세트에 대하여 예측정확도를 비교하였습니다. 아래는 10, 11, 12번째 예측정확도 결과입니다.

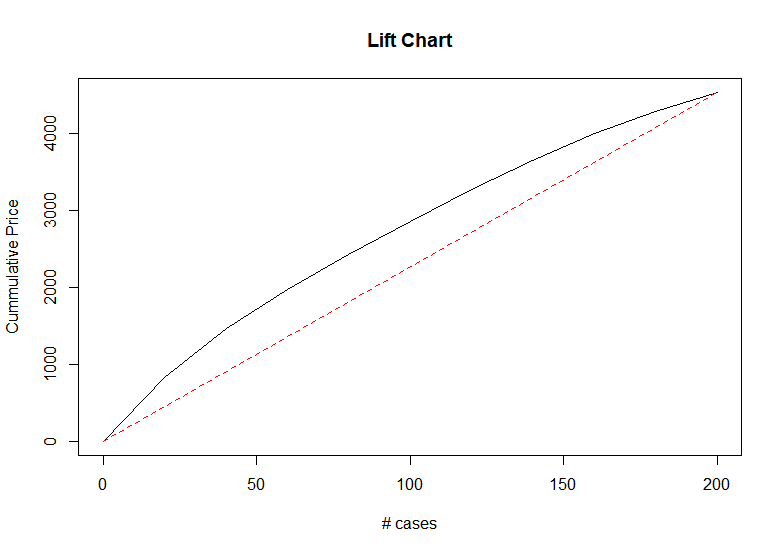
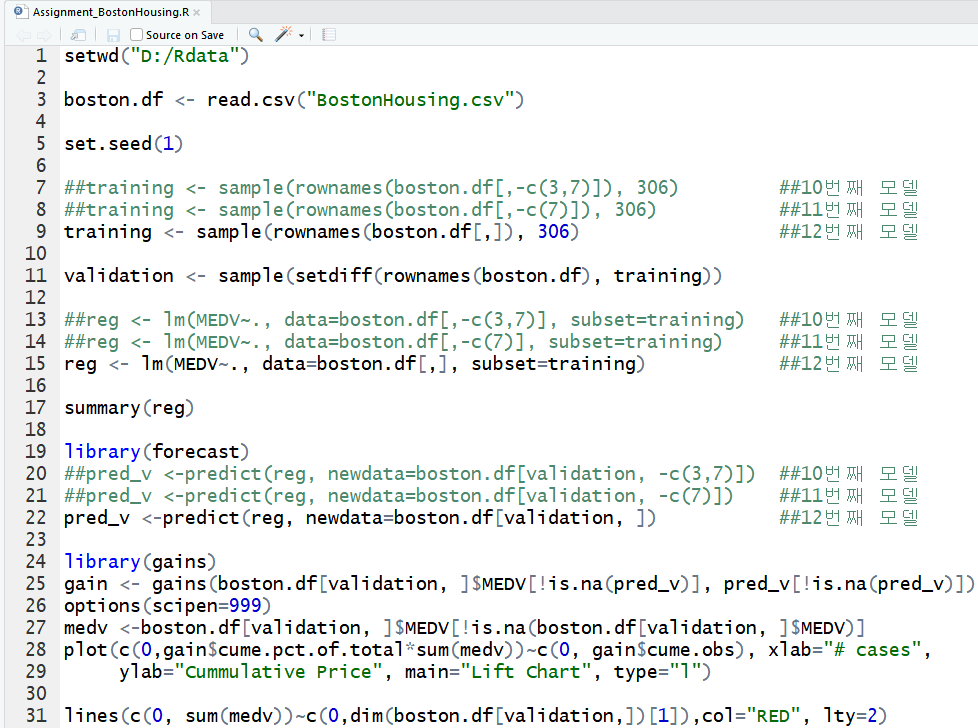


12번째 모델

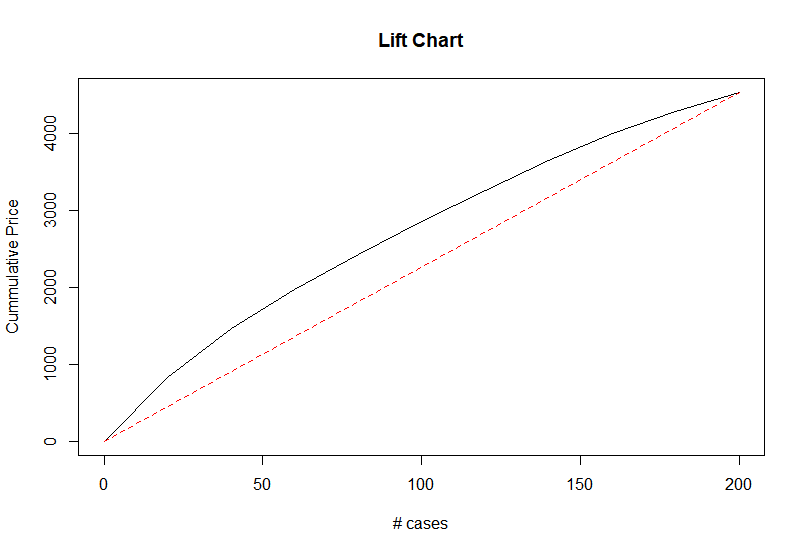
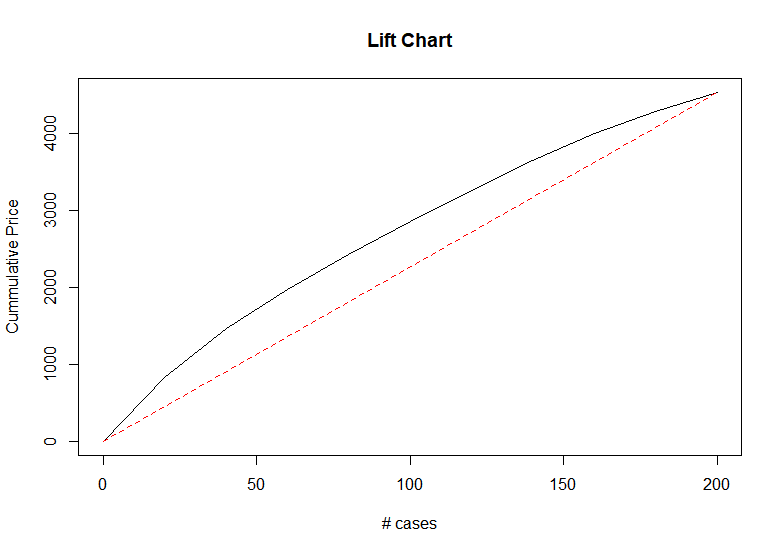
11번째 모델

10번째 모델

* 10번째 모델은 평균오차(ME)가 가장 작고 RMSE는 큰 값을 가지며, 반면에 12번째 모델은 평균오차(ME)가 가장 크고 RMSE는 작은 값을 가집니다.



10번째 모델



12번째 모델

11번째 모델

* 다음은 10, 11, 12번째 모델의 리프트도표 결과입니다. 육안으로 리프트 도표 차이를 확인하는 것은 어렵습니다.
* 따라서 평균오차가 가장 작은 10번째 모델이 가장 좋은 모형이며, 사용되는 예측 변수는 CRIM, ZN, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO, LSTAT 입니다. 굳이 평균오차가 아닌 RMSE를 기준으로 가장 작은 다 집어넣은 12번째 모델이 가장 좋은 모형이다.