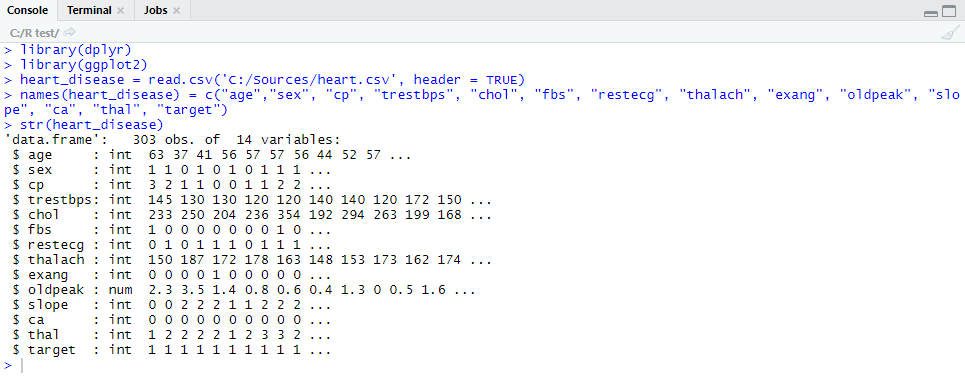
**데이터 마이닝 프로젝트**

**데이터 마이닝(화123) 정경용 교수님**

**201511816 박지훈**

* **주제 : Heart Disease 데이터를 이용한 심장병 유무 분류 예측**

프로젝트에 활용된 데이터는 uci repository에 존재하는 Heart Disease 데이터를 사용하게 되었습니다. (<https://archive.ics.uci.edu/ml/datasets/Heart+Disease>)



uci repository에서 Heart Disease 데이터를 다운받았습니다. csv파일로 바꾸어 R studio에서 read.csv()를 이용해 데이터를 불러왔습니다. csv파일로 바뀌면서 인코딩에 문제가 생겨 names()를 이용해 속성의 이름을 바꿔주었습니다.

데이터의 구조입니다. 총 303개의 행과 14개의 속성을 가지고 있습니다.

1. age : 나이

2. sex : 성별 (0=여자 / 1=남자)

3. cp : 가슴 통증 타입 (1=전형적인 협심증 / 2=비정형 협심증 / 3=비혈관 통증 / 4=무증상)

\* 협심증이란? 심장에 혈액을 공급하는 혈관이 좁아지거나 막혀 심장근육이 필요로 하는 것보다 부족한 양의 산소가 심근에 공급되기 때문에 발생하는 증상

4. trestbps : 휴면 혈압 (입원 시 mm Hg 단위)

5. chol : 혈청 콜레스테롤 (mg/dl)

6. fbs : 당 혈당 수치 > 120mg/dl 인 경우 (0=false / 1=true)

7. restecg : 휴식 중일 때 심전도 결과 (0=정상 / 1=T파 반전 및 ST파 상승 / 2=Estes의 기준에 의해 발생가능하거나 확실한 좌심실 비대증을 표현)

8. thalach : 최대 심박수

9. exang : 운동 유도 협심증 (0=no / 1=yes)

10. oldpeak : 휴식과 관련된 운동으로 인한 ST 우울증

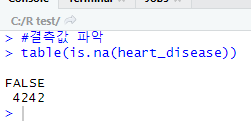
11. slope : peak 운동의 ST 세그먼트의 기울기 (1=upsloping / 2=flat / 3=downsloping)

12. ca : flourosopy에 의해 색칠된 주요 혈관 수 (0-3)

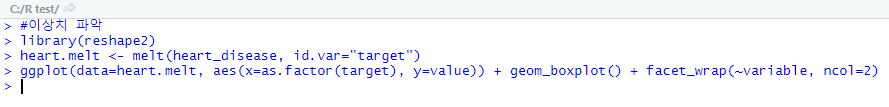
13. thal : (3=정상 / 6=치료불가 결함 / 7=치료 가능한 결함)

14. target : 심장질환 진단 (=혈관질환 상태) (0=심장병 환자X / 1=심장병 환자 맞음)

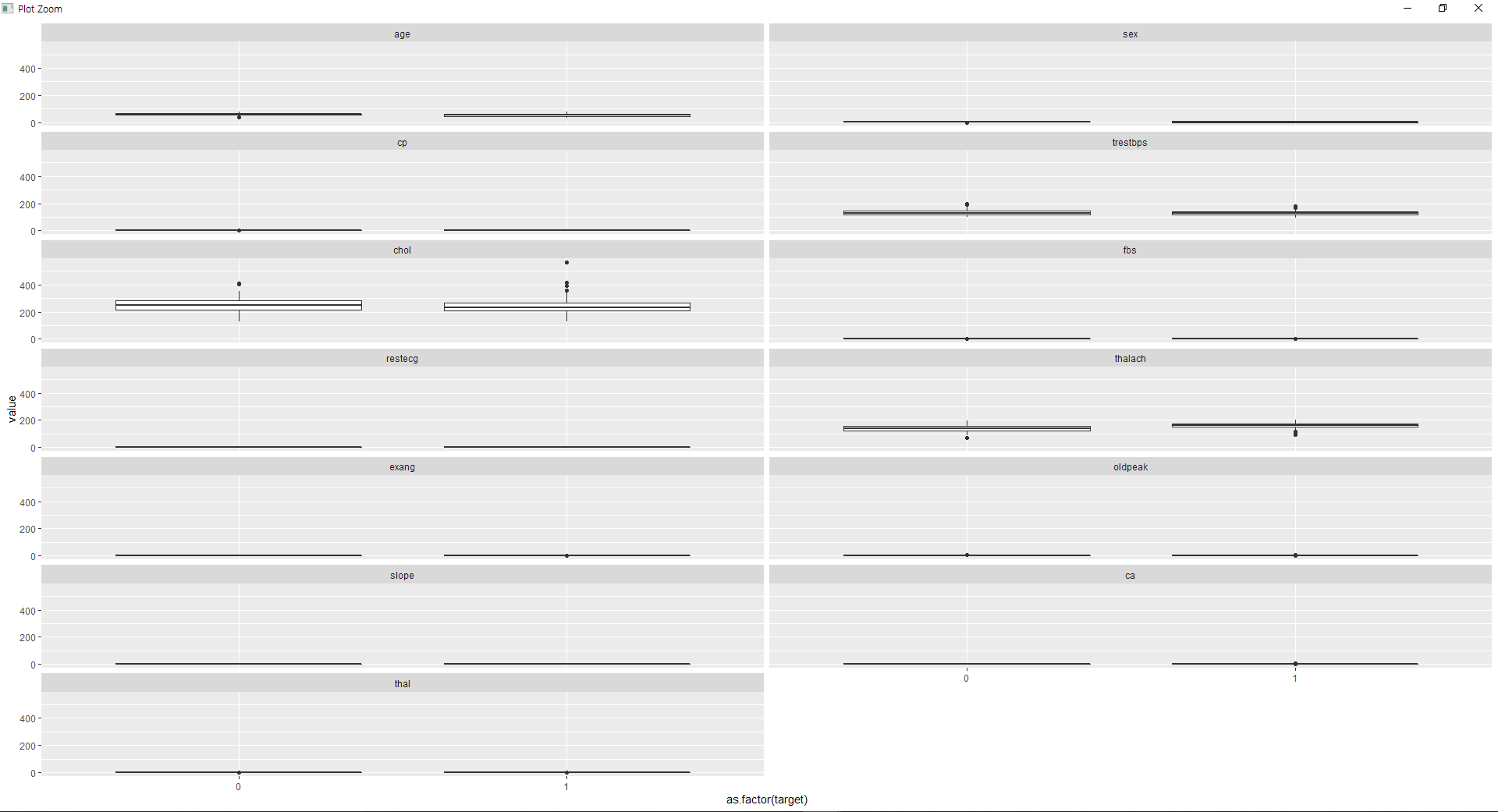
-> 데이터의 속성을 파악한 결과 14번째 변수인 **target**을 반응 변수(종속 변수)로 설정하였습니다.



결측값(missing value)을 파악한 결과 결측값이 없는 것으로 나타났습니다. 결측값을 제거하는 과정을 따로 진행하지 않았습니다.



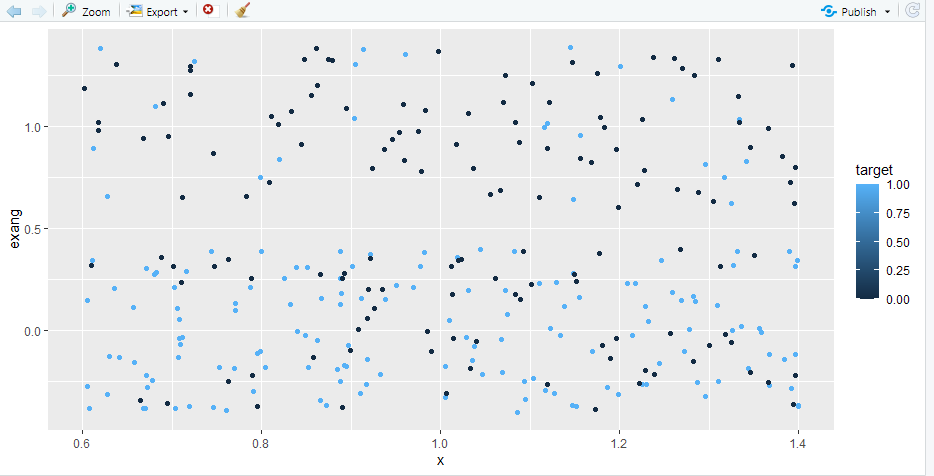
이상값을 파악하기 위해 데이터의 행과 열을 바꿔주는 melt()를 이용해 식별자 속성을 target을 지정해주고, 데이터의 구조를 변경하였습니다.



이상값을 시각화 하기 위해 boxplot()을 이용하였습니다. chol 부분에서 차이가 큰 이상값을 발견할 수 있었으나 반응 변수에 엄청난 영향을 미치지 않을 것으로 예상되어 이상값 처리하는 과정을 진행하지 않았습니다.

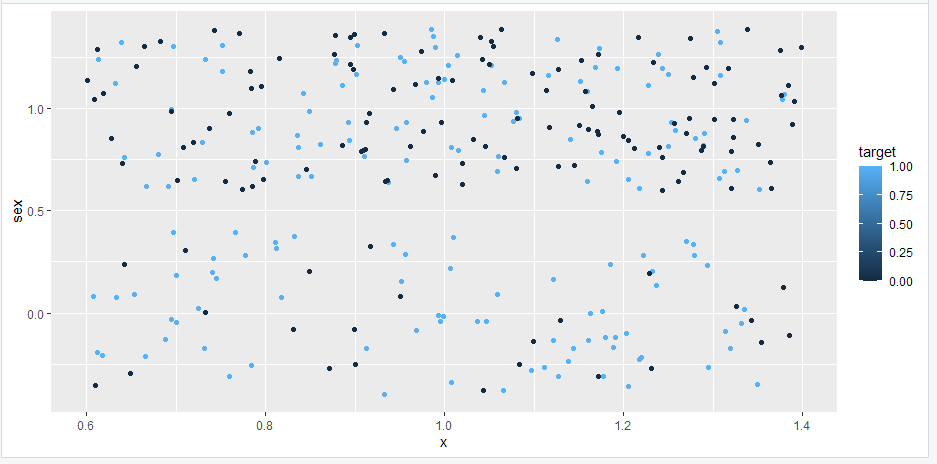
데이터를 시각적으로 직접 확인하고, 데이터를 선형적으로 처리할 수 있는지 아닌지에 대해 알기 위해 데이터의 시각화(산점도)를 진행하였습니다.





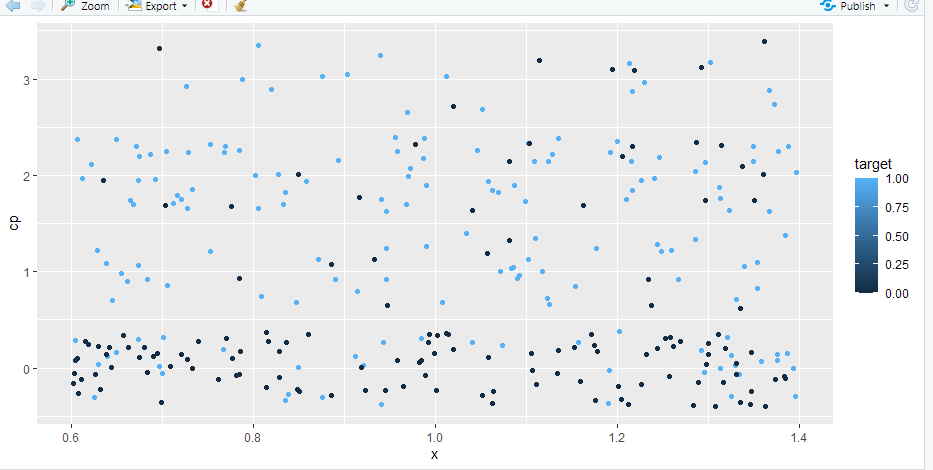
운동 유도 협심증(exang)이 없을수록(정상일수록) 심장병 환자가 많은 것을 볼 수 있습니다. exang이 올라갈수록 심장병 환자도 조금씩 존재합니다.





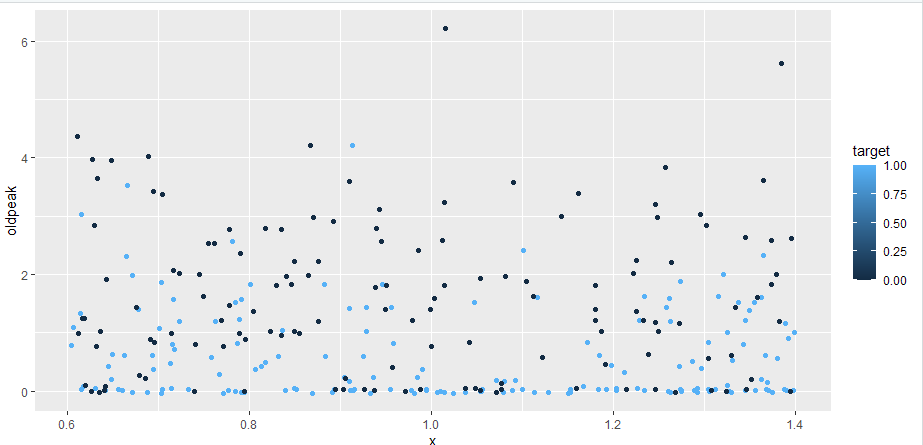
남성일 경우 심장병 환자가 좀 더 많은 경우를 볼 수 있습니다. 하지만 여성도 전체 여성중에 정상인 여성보다 심장병 환자인 여성이 더 많은 것을 볼 수 있습니다.





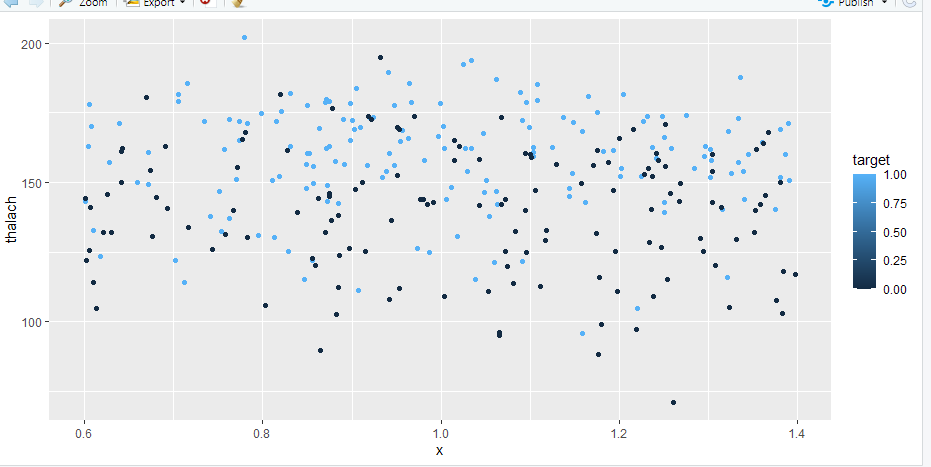
cp(가슴 통증)이 1, 2, 3일 경우 심장병 환자가 더 많은 것으로 파악됩니다. 즉 가슴 통증이 없는 경우(4) 심장병 환자가 더 적은 것으로 파악됩니다.





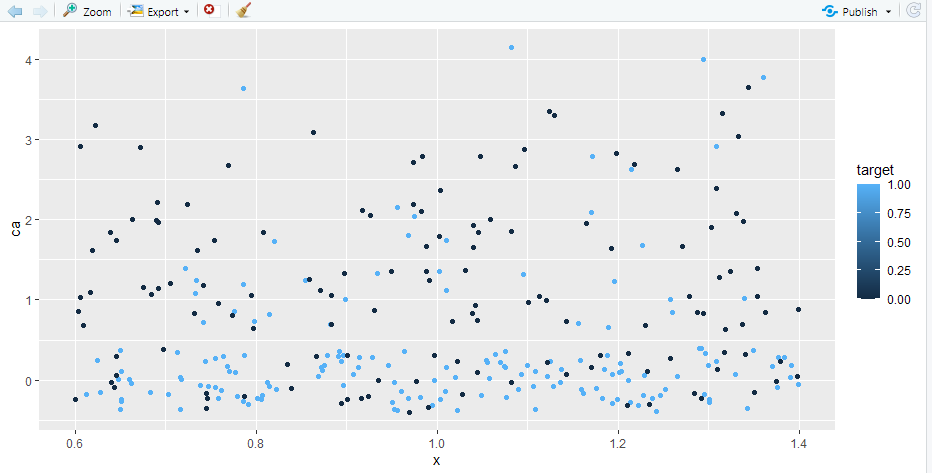
휴식과 관련된 ST우울증값이 낮은 경우에 심장병 환자가 많은 것으로 파악됩니다.





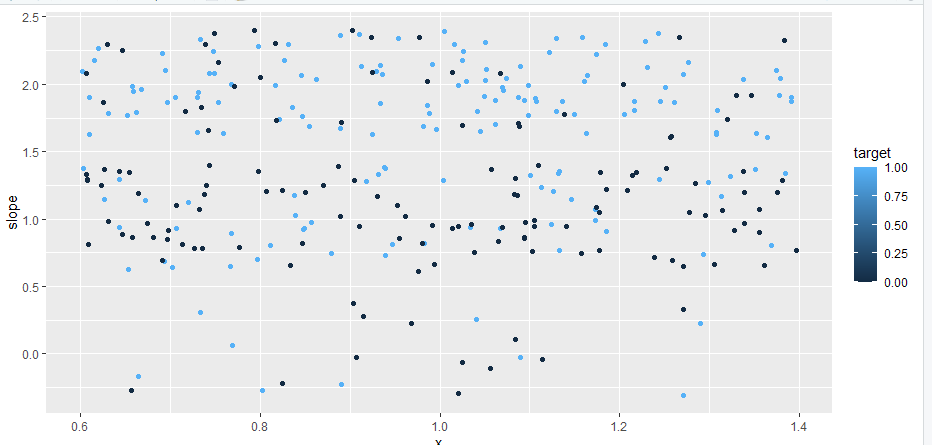
최대 심박수(thalach)가 높을수록 심장병 환자가 많은 것을 파악됩니다.





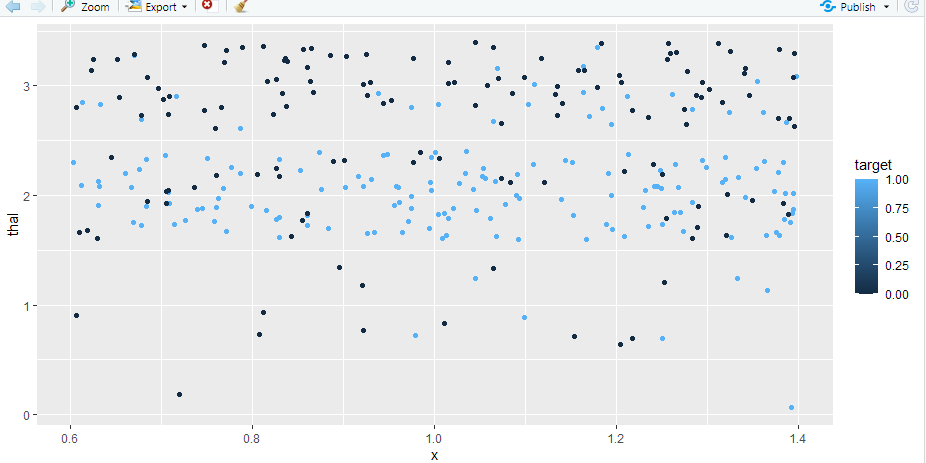
flourosopy에 의해 색칠된 주요 혈관 수가 작을수록 심장병 환자가 많은 것으로 파악됩니다.





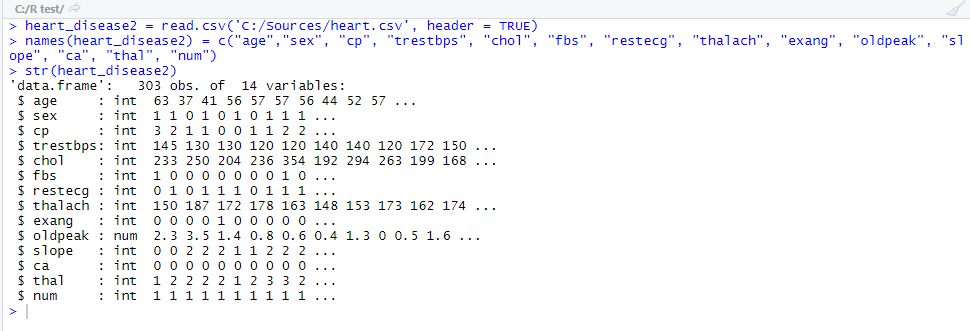
peak 운동의 ST 세그먼트의 기울기가 높을수록 (downsloping일수록) 심장병 환자가 많은 것으로 파악됩니다. 하지만 flat하거나 upsloping에서도 심장병 환자가 조금 많은 것으로 파악됩니다.





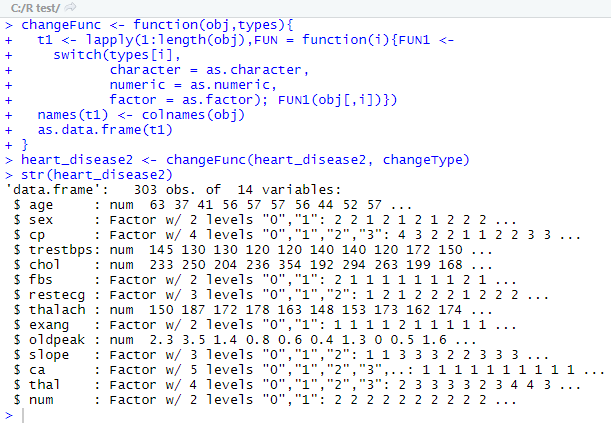
thal값이 2일 때 심장병 환자가 많은 것을 파악할 수 있습니다.

심장병의 유무가 성별에 영향을 받고 싶은지 더 정확하게 알고 싶어 시각화를 해보았습니다.

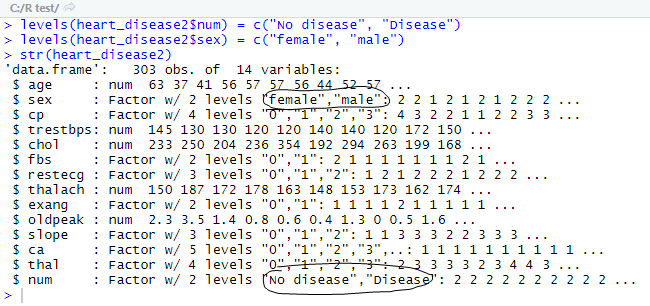


기존의 heart\_disease를 train/test 데이터로 활용하기 위해 변경하지 않고 heart\_disease2라는 데이터셋을 하나 더 생성하였습니다. heart\_disease2은 oldpeak속성만 numeric형태이고, 나머지 속성은 전부 int형입니다. heart\_disease2의 데이터 형 변환을 처리하는 작업을 진행하였습니다.





changeFunc이라는 함수를 만들었습니다. types라는 매개변수에 character가 들어오면 character형으로 바꿔주고 numeric이 들어오면 numeric형으로 변환해주는 함수를 생성하였습니다. factor형도 이하동문입니다. 이렇게 형 변환되는 횟수는 데이터의 속성의 전체 개수만큼 수행하게 됩니다. heart\_disease2의 데이터 구조를 보시게 되면 데이터 형이 변환됨을 알 수 있습니다.



sex의 0, 1이었던 범주를 female, male로 바꿔주었고, num(=target)또한 No disease, Disease로 바꾸어 주었습니다. 시각화를 통한 성별과 심장병의 유무의 상관관계만을 파악하는 것이 목적이기 때문에 sex, num(=target)의 범주형 값만 변환하였습니다.



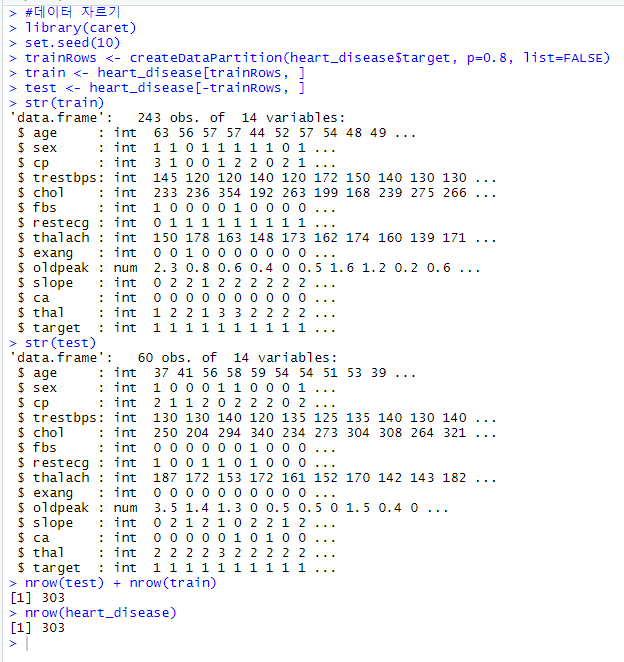


구글링을 통해 mosaicplot()을 알 수 있었고, 범주형 데이터를 표현하는데 적합한 그래프라고 하여서 적용하였습니다. sex ~ num(=target) 형태의 식을 사용하고 색상을 적용하여 x축에는 성별, y축에는 심장병 유무를 지정하였습니다.

남성의 경우 심장병 환자인 남성이 정상인 남성보다 더 적으나, 여성의 경우 정상인 여성보다 심장병 환자인 여성이 더 많은 것으로 파악되는 것을 볼 수 있습니다. (데이터 내에서 전체 남성의 수가 여성의 수보다 많습니다.)

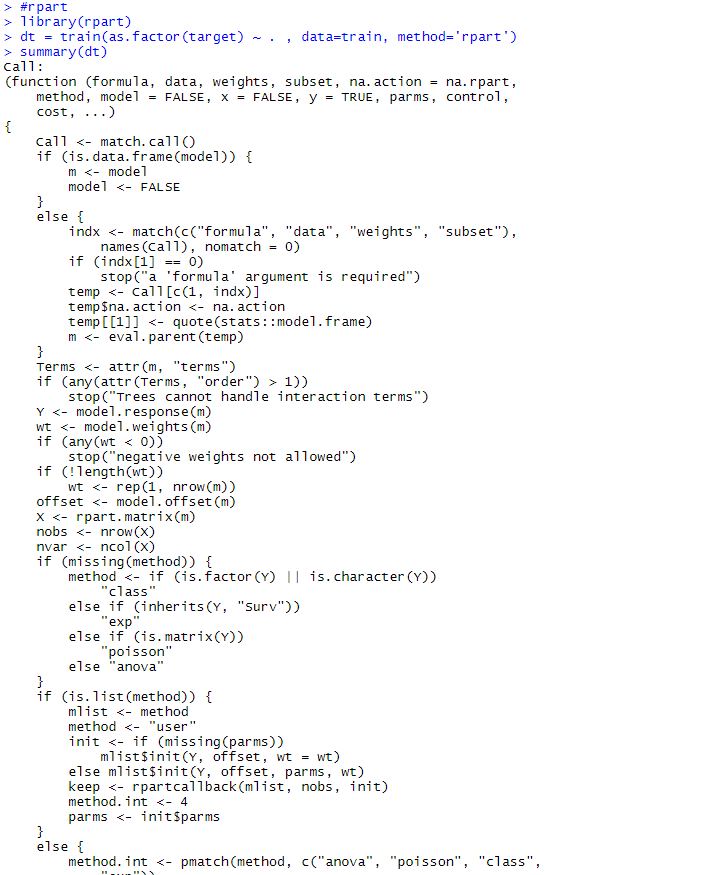
* 데이터의 시각화를 통해 반응 변수인 target을 제외한 모든 속성을 모델링의 설명 변수(독립 변수)로 활용하기로 판단하였습니다.

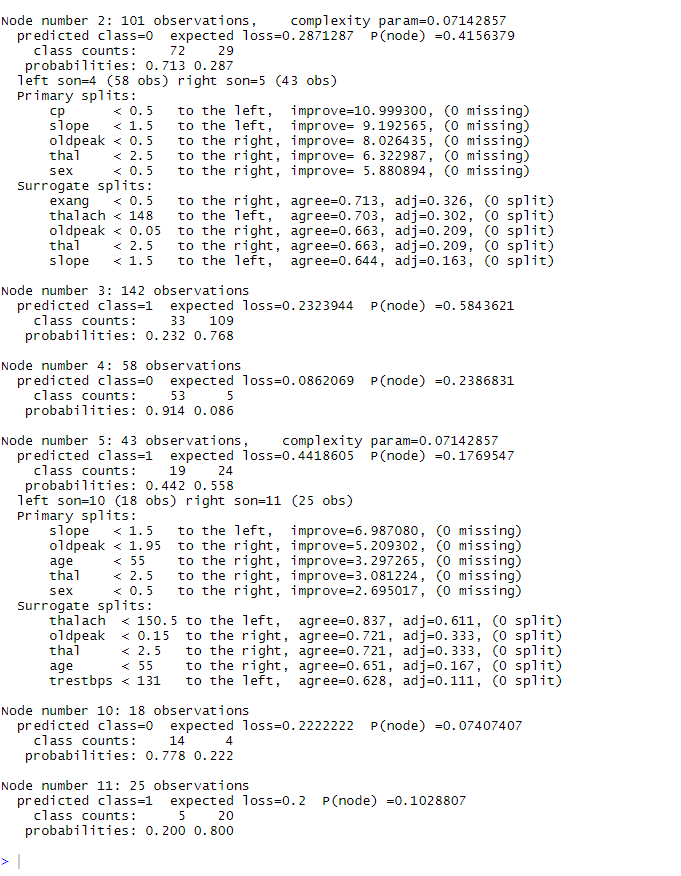
마지막으로 데이터를 활용하여 모델링을 구현하겠습니다. 먼저 데이터를 train데이터와 test데이터를 나누는 과정을 진행하였습니다. 비율은 train : test = 8:2로 나누었습니다.

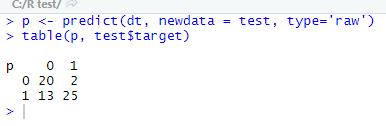


train데이터의 개수는 243개, test데이터의 개수는 60개로 8대2로 데이터를 대략적인 수치로 나누었습니다.

* 결정 트리(Decision Tree)로 모델링을 진행하였습니다.

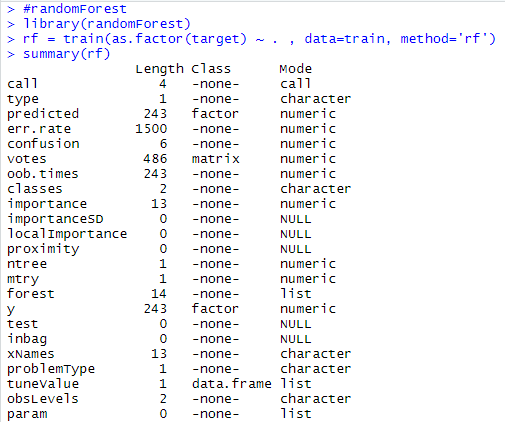


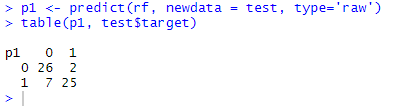




(20+25) / 60 = 0.75로 75%의 정확도를 보여줍니다.

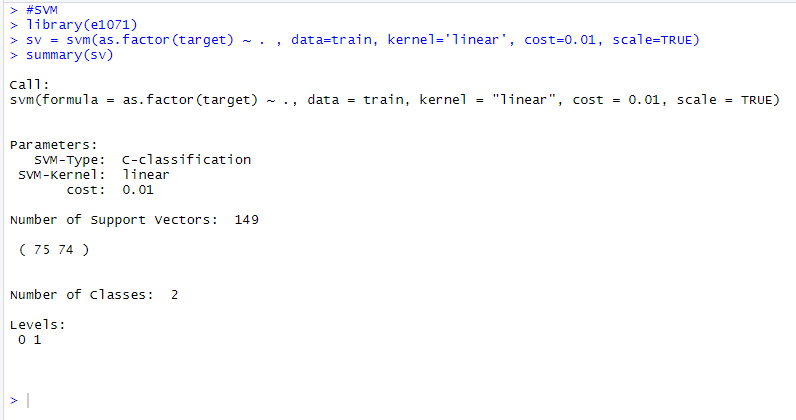
* randomForest로 모델링을 진행하였습니다.

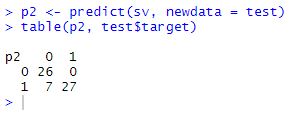




(26+25) / 60 = 0.85로 85%의 정확도를 보여줍니다.

* SVM으로 모델링을 진행하였습니다.

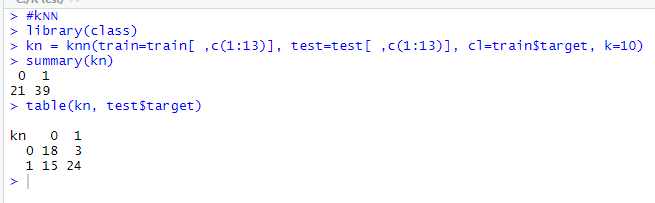




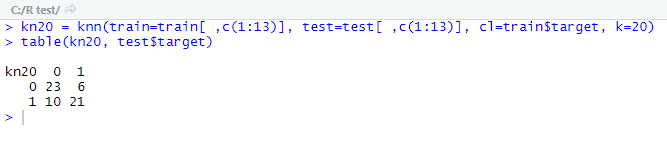
(26+27) / 60 = 0.96363…로 96%의 정확도를 보여줍니다.

* k-NN으로 모델링을 진행하였습니다.

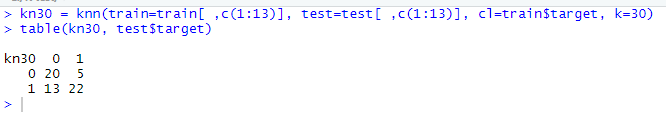
k-NN 알고리즘은 k값에 따라 분류 결과가 민감해질 수 있으므로 다양한 모델링을 해보았습니다. 또한 데이터 형에 영향을 받을 수 있으므로 2가지의 데이터에 대해서 k-NN으로 모델링을 해보았습니다. (k=1인 경우에는 과잉적합(overfitting)의 가능성이 있어 k=1로 설정하지 않았습니다.)



(18+24) / 60 = 0.7로 70%의 정확성을 보여줍니다.



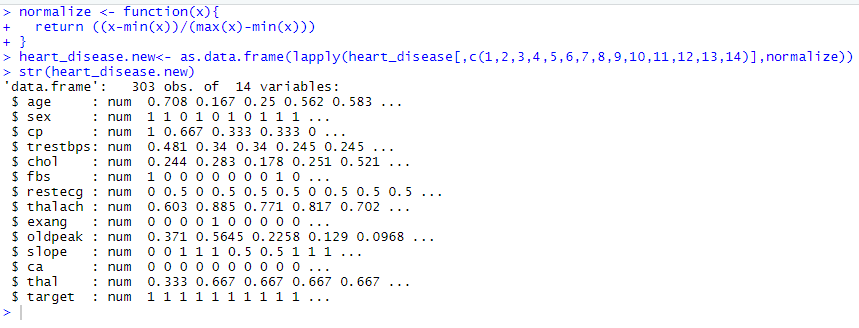
(23+21) / 60 = 0.733…로 73%의 정확성을 보여줍니다.



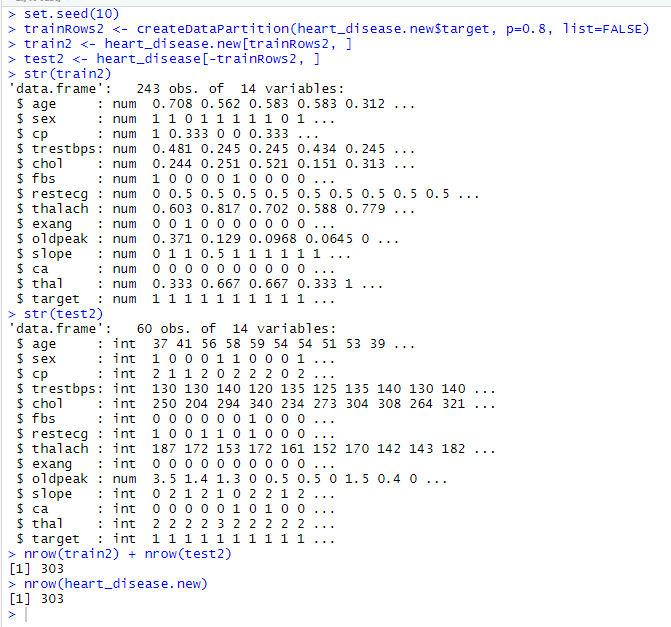
(20+22) / 60 = 0.7로 70%의 정확성을 보여줍니다.

위의 결과에 알 수 있듯이 k의 적절한 값은 20이라고 판단하였습니다. k=20이라 하고 데이터 형의 변환이 된 정규화된 데이터 셋에서 k-NN의 모델링을 진행하였습니다.

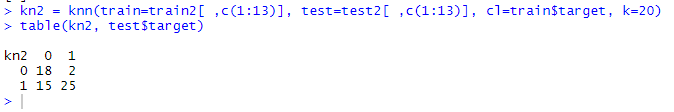
heart\_disease의 데이터 셋을 정규화 하였습니다.



heart\_disease.new에 heart\_disease 데이터 셋의 데이터 형 변환을 하였습니다. 데이터의 구조를 보면 모든 속성이 numeric형으로 바뀐 것을 알 수 있습니다.

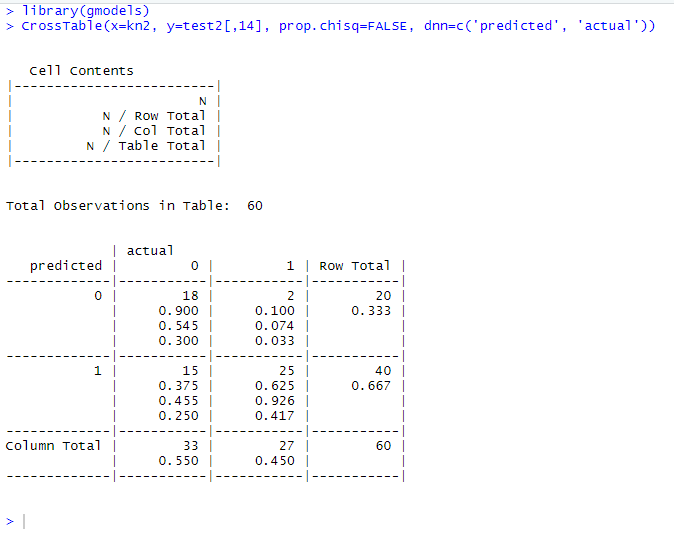


정규화된 heart\_disease.new 데이터셋을 train : test = 8:2로 나누었습니다.



k=20일 때 정규화된 데이터 셋에 좋은 일반화 능력을 가지고 있는 k-NN 모델링이라는 것을 알 수 있습니다.

(18+25) / 60 = 0.71666…로 71.6%의 정확도를 보여줍니다.



gmodel의 라이브러리의 CrossTable()을 활용해서 k-NN 모델링의 교차 분할표를 출력했습니다.

표 결과는 위에서 아래 순서대로 일반 횟수, 카이 제곱(기대치 비율), 행을 기준으로 비율 값(가로로 읽음), 컬림을 기준으로 비율 값(세로로 읽음), 전체를 기준으로 비율 값을 표현하고 있습니다.