2021-1 YDMS 1주차 과제

<1주차 주제 : 데이터마이닝 프로세스 개요 및 데이터 시각화>

김하은 2019251034

1. EDA를 하는 이유에 대해서

EDA는 '데이터 탐색 과정'으로, 데이터 분석 시 가장 먼저 하게 되는 과정이다. EDA는 가설 검정이나 문제해결을 위한 과정이 아니라 전반적인 데이터 흐름을 파악하고 특이 패턴을 찾아내는 과정이다. 따라서 주어진 데이터를 편향되지 않은 시선으로 분석할 수 있는데, 이는 효과적인 전처리를 돕는다. 또한 탐색 과정에서 이상치나 결측치를 탐지하여 정제된 데이터를 만들어낼 수 있게 한다.

무엇보다 데이터의 흐름을 파악하고, 변수 간 관계를 파악할 수 있다는 점은 EDA의 큰 장점이다. 이를 바탕으로 기존의 가설을 수정하기도 하고 새로운 가설을 세우기도 한다. 이러한 과정은 이후 모델 성능을 향상시키기도 하며 분석의 진전을 돕기 때문에 EDA는 데이터 분석의 중요한 과정이라 할 수 있다.

2. '데이터 시각화'에 대한 자신의 생각 정리해오기

'데이터 시각화'란 데이터 탐색 시 그래프 등을 이용하는 것이다. 데이터 시각화를 통해 데이터의 분포 양상, 흐름 등을 파악할 수 있으며, 이는 보다 직관적인 판단을 돕는다. 데이터 시각화 과정에서 주의해야할 점은 변수 파악이라 생각한다. 변수에 대한 이해가 낮은 상태로 시각화를 진행하게 되면 제대로 된 그래프를 얻을 수 없으며 데이터를 오독할 수 있기 때문이다.

따라서 데이터 시각화 진행 시 첫 번째로 해야 하는 과정은 변수 분류이다. 자료는 크게 범주형 자료와 수치형 자료로 나누어진다. 따라서 주어진 데이터의 변수를 범주형 변수와 수치형 변수로 분류해야하고, 각 형태에 맞는 그래프를 그려 시각화해야 한다. 범주형 변수의 경우 막대그래프, 원그래프, 파레토차트 등을 이용한다. 반면 수치형 변수의 경우 히스토그램, 박스플롯 등을 이용하여 데이터의 분포를 파악한다. 또한 산점도를 이용하여 변수 간 상관관계와 이상치를 파악한다.

데이터 시각화를 하게 되면, 데이터를 탐색하기 용이해지며 결과를 효과적으로 전달할 수 있다. 또한 이상치, 결측치 등을 찾아줌으로써 데이터 정제에도 도움을 준다. 마지막으로 변수 간 기본 패턴과 가설 설정까지 도움을 주기 때문에 데이터 시각화는 분석 시 꼭 필요한 단계라고 생각한다.

3. 지도학습&비지도학습에 대해 생각하고 사례에 따른 학습 종류 생각해보기

우선 지도학습은 '예측변수'가 존재하는 경우로, 특정 결과값을 예측 및 추정하기 위해 모델링하는 방법이다. 이때 반응변수와 예측변수와의 관계를 통해 모델링하게 되는데, 모델링 시 이용되는 데이터를 '학습데이터'라고 부른다. 이때 분류/알고리즘으로 예측변수와 종속변수 사이의 관계를 학습하여 모델이 구축된다. 이 과정을 통해 모델이 구축되면, 모델의 성능을 평가하게 되고, 다른 모델과 비교 끝에 선정된다. 이렇게 모델이 선정되면, 미래 데이터의 종속

변수의 값을 예측할 수 있게 된다. 이러한 지도학습에는 대표적으로 단순 선형 회귀분석이 있다. 단순 선형회귀분석은 반응변수와 예측변수의 관계를 선형식으로 도출하는 것으로, 반응변수의 값을 예측하는데 사용된다.

반면 비지도학습은 주로 예측변수가 존재하지 않는 경우에 사용된다. 따라서 비지도학습은 결과값을 예측하기 위한 모델링 과정이 아닌 데이터 속 정보를 찾아내기 위한 방법이다. 주로 데이터 내 관계와 구조를 찾고, 비슷한 관측치를 군집하며, 차원을 축소하는데 이용된다. 대표적인 비지도학습에는 군집분석이 있다. 군집분석은 데이터 관측값들의 유사성을 중심으로 패턴을 분석하는 것이다.

두 방법이 따로 쓰이는 것만은 아니고 종종 같이 사용되기도 한다. 예를 들어 하나의 데이터를 비지도 학습을 통해 연관성을 찾고, 이후 지도학습으로 예측변수에 대한 값을 추정하기도 한다.

4. 변수 TYPE에 대해 정리하기 (EX. 범주형, 연속형, 명목형, 이산형, 순서형 등)

범주형 자료 (질적 자료)	**명목형 변수**	성별, 혈액형 같이 특별한 순위가 없는 경우
	순서형 변수	학력 같이 큰 값, 작은 값으로 표현할 수 있는 경우
수치형 자료 (양적 자료)	**연속형 변수**	키, 몸무게처럼 연속적인 모든 값을 가질 수 있는 경우
	이산형 변수	가족 수처럼 값 사이에 간격이 존재하는 경우. 이산적인 값을 가진 경우

5. 변수변환의 필요성과 방법에 대해 생각해오기

변수변환이란 필요시 데이터 변수를 변환시키는 것이다. 변수의 형태를 바꿀 수도 있고, 변수의 값을 변환시킬 수도 있다. 가령 세금과 같이 범위는 크고 데이터들의 분포는 일정하지 않은 경우, 세금을 일정 구간으로 나누어 'Low', 'Mid', 'High' 과 같이 범주화 시킬 수 있다. 이것은 변수의 형태를 바꾸는 변수변환 방식이다. (수치형 변수 → 범주형 변수) 반면, 회귀분석 시, 변수의 형태를 $y = x^2$에 $\sqrt{\ }$를 적용하여 $y' = x$로 변환시켜 값을 구하는 것은 변수의 값을 변환시킨 변수변환 방식이다.

변수변환이 필요한 이유는 데이터 분석을 돕기 때문이다. 컴퓨터가 처리하는 것도 도와줄 뿐만 아니라 우리가 데이터를 해석할 때에도 도움을 준다. 이때 각 변수들의 의미를 정확히 알고 진행해야하며, 이와 같은 과정이 모델의 성능을 떨어뜨릴 수 있으므로 주의해야한다.

6. 타켓 변수란?

타겟 변수란 값을 모형화하고, 다른 변수를 이용해 예측해야하는 변수이다. 따라서 주어진 데이터에 타겟 변수가 존재한다면, 최종적으로 그 변수에 관한 정보를 얻는 것이 분석의 목적이 될 것이다. 이는 선형 회귀 분석의 종속변수와 유사하다.

7. 주어진 Boston housing data set을 이용하여 데이터 탐색을 진행하고 탐색에 따른 자신의 견해 작성하기.

데이터 탐색에서 가장 중요하다고 생각하는 것은 변수에 대한 이해이다. 따라서 가장 먼저 주어진 boston_houseprice의 각 변수별 형태와 설명을 함께 보며 변수에 대해 살펴보았다.

우선 변수를 크게 범주형과 수치형으로 나누어 살펴보았고, 범주형 변수는 명목형 변수와 순서형 변수로 다시 구별해주었다. 수치형 변수를 살피기 전, 비율과 관련된 변수가 상황에 따라 다른 단어로 설명되고 있음을 보았고, 이것이 데이터 수치와 관련되어 있다고 판단하였기 때문에 단어에 대한 학습을 먼저 진행하였다.

초점을 둔 단어는 'RATE, RATIO, PROPORTION, PERCENT' 이다. 비슷한 흐름임에도 서로 다른 상황에 맞춰진 단어임을 알 수 있다. 따라서 단어에 따라 데이터에 대한 시각이 달라질 수 있다. 특히 이상치에 관한 시각이 해당될 것이다.

'RATE, RATIO, PROPORTION, PERCENT' 대한 설명은 다음과 같다.

RATE	특정기간 동안 발생된 사건으로 보통 천분율(‰)로 표현 예) 1년 동안의 이혼 건수
RATIO	분모와 분자는 독립적인 관계 (서로 다른 범주) 예) 성비, 인구밀도(인구/면적)
PROPORTION	비의 특수한 형태로, 분자가 분모에 포함되어있음 예) 재학생 중 남학생의 비율 (남/남+여)
PERCENT	전체 수량을 100으로 하여, 해당 수량을 파악하는 것. (%)

위에서 알아본 것을 바탕으로 변수 설명을 작성하였고, RATE는 관념에 따라 천분율이라 보았고, PROPORTION의 경우 변수의 값들을 고려하여 PERCENT의 의미로 작성된 것으로 판단하였다. 이를 바탕으로 정리한 것은 아래와 같다.

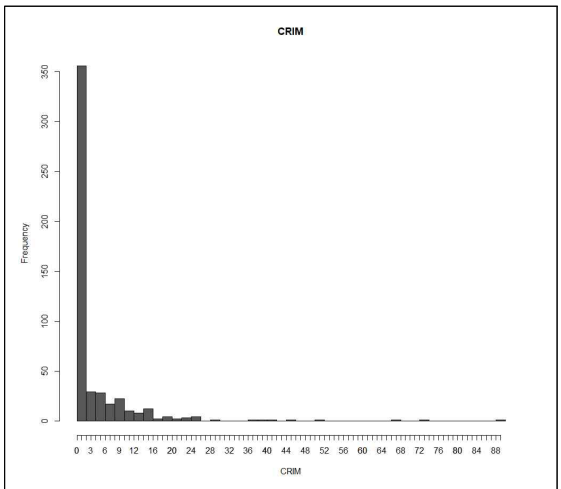
```
> summary(boston)
      CRIM                ZN              INDUS            CHAS              NOX                RM              AGE        
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90  
 1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00  
      DIS              RAD              TAX            PTRATIO            B              LSTAT            MEDV      
 Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00  
 1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02  
 Median : 3.207   Median : 5.000   Median :330.0   Median :19.05   Median :391.44   Median :11.36   Median :21.20  
 Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65   Mean   :22.53  
 3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00  
 Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00  
```

순서	변수명	변수 설명	
1	▶CRIM	수치형	자치시(town)별 1인당 범죄율 (RATE-천분율‰)
2	▶ZN	수치형	25,000평방피트를 초과하는 거주 지역 비율 (PROPORTION<1)
3	▶INDUS	수치형	자치시(town)별 비상업지역의 비율 (acres 단위) (PROPORTION<1)
4	▶CHAS	범주형	찰스 강에 관한 변수 [명목형 변수] 1=강과 인접 0=강과 비인접
5	▶NOX	수치형	10ppm당 일산화질소
6	▶RM	수치형	가구당 평균 방의 개수
7	▶AGE	수치형	1940년 이전에 건축된 소유주택의 비율 (PROPORTION<1)
8	▶DIS	수치형	보스턴 5대 상업지구와의 거리
9	▶RAD	범주형	방사형 고속도로까지의 접근성 지수 [순서형 변수]
10	▶TAX	수치형	10,000달러당 종합재산세율 (RATE-천분율‰)
11	▶PTRATIO	수치형	자치시(town)별 $\frac{학생}{교사}$ 비율 (학생 대 교사의 비율) (RATIO-독립변수끼리)
12	▶B	수치형	1000(Bk-0.63)^2 여기서 Bk는 도시별 흑인의 비율
13	▶LSTAT	수치형	저소득층 비율 (%-백분율)
14	▶MEDV **(타겟 변수)**	수치형	소유자 주택 가격의 중앙값 ($1,000단위)

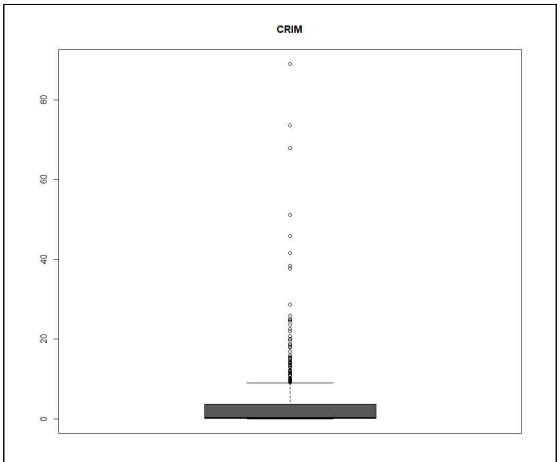
#CRIM

CRIM은 도시별 1인당 범죄'율'을 나타내는 변수인데, rate로 계산된 변수이므로 통상 쓰이는 관념에 맞게 천분율 값으로 보았다. 따라서 데이터의 기준을 1,000명으로 잡고 해석하였다. 그래프는 연속형 변수에 맞게 히스토그램으로 시각화하였다.

전체적인 데이터 분포는 'CRIM=0' 근방에 가장 많이 몰려있으며 범죄율이 높은 수준으로 갈수록 거의 분포하지 않았다. 천분율을 바탕으로 해석하면, 범죄가 발생하지 않는 town이 가장 많았으며 자료의 약 58%를 차지하였다. (CRIM의 자료를 반올림하여 테이블을 만들었으며 이를 바탕으로 대략적인 도수를 파악했다.) 또한 범죄율이 가장 높은 경우는 인구 1,000명당 88명이 노출된 경우였다. 이를 통해 보스턴의 대부분 town들은 범죄율이 낮은편이며, 특정 town들만 범죄율이 높다는 것을 알 수 있다.



hist(boston$CRIM,col='grey34',main='1000명당 범죄발생 횟수',xlab='CRIM',breaks=35)+axis(side=1,at=seq(0,89,1))



boxplot(boston$CRIM,col='grey34',main='CRIM')

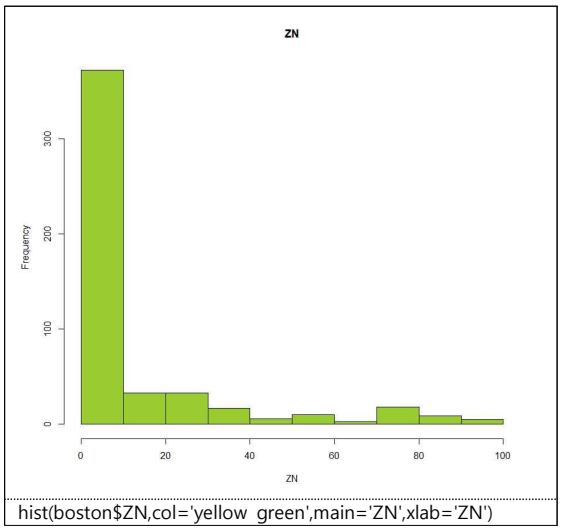
0	294
1	56
2	18
3	8
4	15
5	13
6	12
7	9
8	11
9	9
25	2
26	1
29	1
38	2
42	1
46	1
51	1
68	1
74	1
89	1

crim<-table(round(boston$CRIM,0))
View(crim)

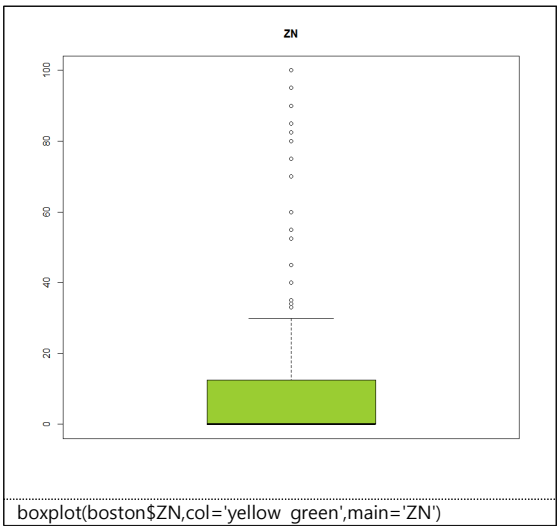
<반올림 테이블 도수분포표>

#ZN

ZN은 25,000평방피트를 초과하는 거주지역의 비율을 나타내는 변수로 연속형 변수에 해당한다. 이에 맞게 히스토그램과 상자그림을 그려 시각화하였다. PROPORTION 에 관한 변수이며 전체를 100으로 둔 percent 개념으로 접근하여 해석하였다. 대부분의 town에선 실거주지가 25,000평방피트를 초과하는 비율이 낮았지만, 특정 town은 실거주지가 25,000평방피트를 초과하는 비율이 매우 높게 형성되어있음을 알 수 있다.



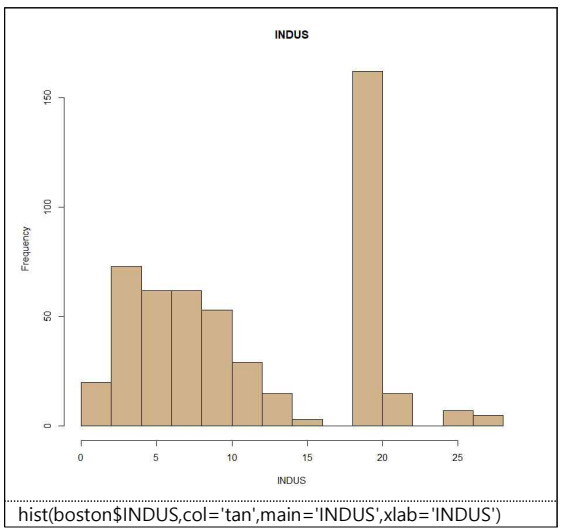
hist(boston$ZN,col='yellow green',main='ZN',xlab='ZN')



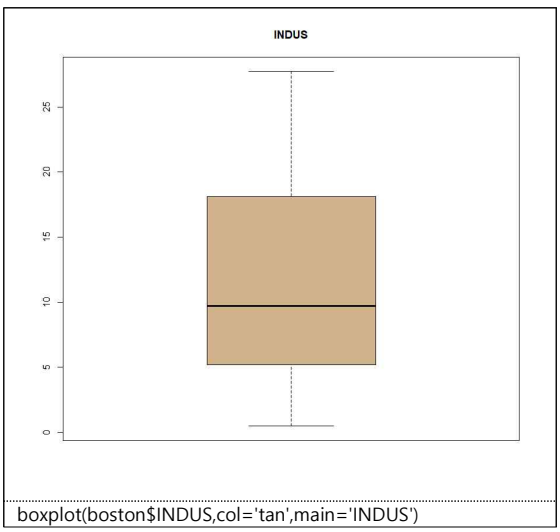
boxplot(boston$ZN,col='yellow green',main='ZN')

#INDUS

INDUS 변수는 비상업지대의 토지비율로, 쉽게 말하면 상권가가 갖춰지지 않은 구역의 비율을 나타내는 변수이다. INDUS는 연속형 변수이므로 히스토그램과 상자그림으로 시각화하였다. 또한 PROPORTION 에 관한 변수이므로 전체를 100으로 둔 percent 개념으로 접근하여 해석하였다. 모든 데이터는 30% 내에서 존재하고 있는데, 이는 보스턴 town 별 비상업지대 토지비율이 30%를 넘지 않는다고 해석할 수 있다. 보스턴의 경우 미국 내 도시별 1인당 GDP 순위에서 1위에 해당하는 도시인만큼 미국 내에서 경제 수준이 우수한 도시이므로, 이를 고려하면 상권가의 비중이 낮을 수 없다는 판단이다.



hist(boston$INDUS,col='tan',main='INDUS',xlab='INDUS')

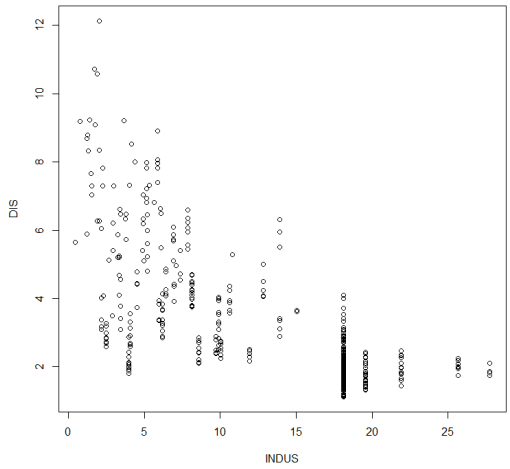


boxplot(boston$INDUS,col='tan',main='INDUS')

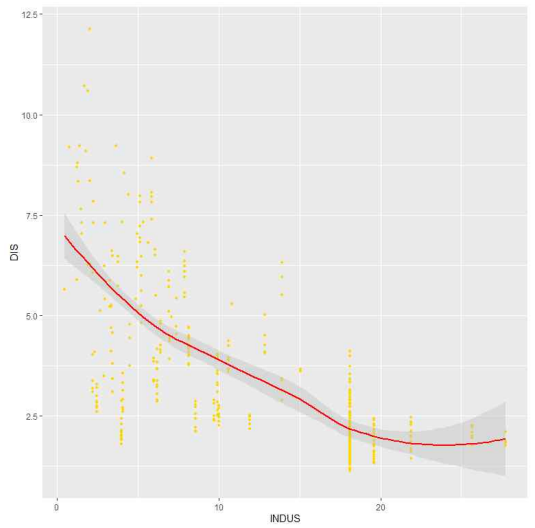
한편 중심가일수록 상권가가 많이 몰려있다는 가정을 바탕으로, 중심가와 상권가를 연결하여 생각해보았다. 따라서 비상업지대의 비율이 낮을수록 보스턴 중심가에 가까운 house 라 추측하였고, 특히 'INDUS=0'인 구간은 보스턴 중심가에 위치한 house 일 것이라 추측하였다. 이때 보스턴 중심가는 보스턴 5대 상업지대와 연관 있을 것이라 판단하였고, 'INDUS=0' 구간은 DIS의 데이터가 높게 형성될 것이란 생각을 하게 되었다. 가정이 맞는지 판단하기 위해 INDUS와 DIS의 상관도를 살펴보았다.

가정과는 달리, 비상업지대의 비율이 올라갈수록 보스턴 5대 상업지대와 가까워지는 형태를 보였다. 상관계수는 -0.71로 매우 높은 음의 상관관계를 보였다.

[INDUS] 와 [DIS]의 관계 : 비상업지대 비율이 올라갈수록 상업지대로부터 가까워짐



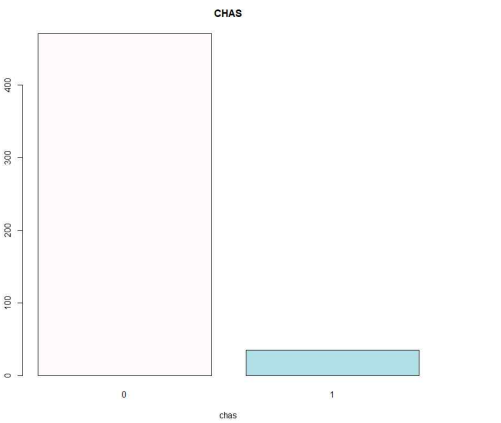
plot(boston$NOX~boston$INDUS,xlab='INDUS',ylab='NOX')



ggplot(data=boston,aes(x=INDUS,y=NOX))+geom_point(size=1, color="gold")+stat_smooth(color='red', fill='grey')

#CHAS

CHAS 변수는 찰스강을 기준으로 나누어진 변수로, 강에 인접한 house는 '1'로, 강에 인접하지 않은 house는 '0'으로 나눈 변수이다. 강에 인접하지 않은 house들이 대부분임을 알 수 있다.




```
> table(boston$CHAS)

  0   1
471  35
```

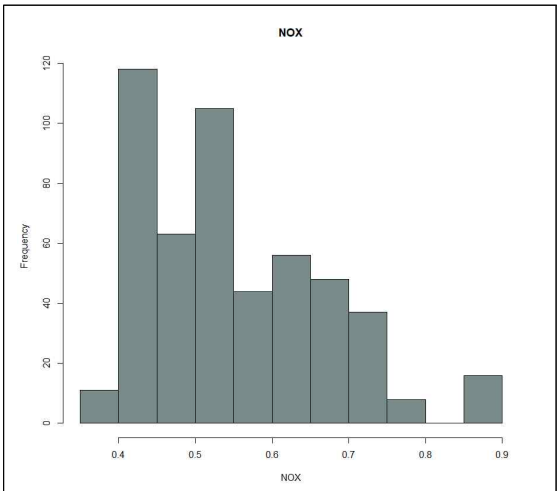
chas<-table(boston$CHAS)
barplot(chas,xlab='chas',main='CHAS',col=c('snow','powder blue'))

#NOX

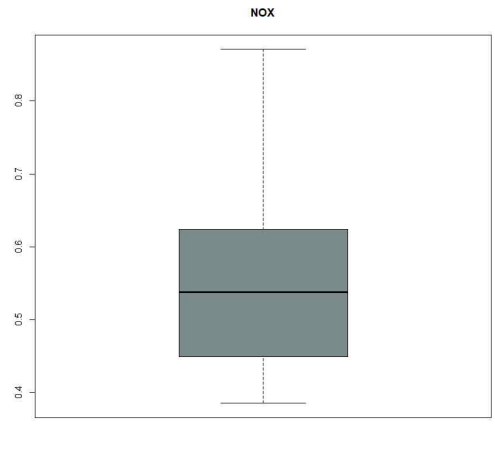
NOX는 10ppm당 농축 일산화탄소의 농도를 나타내는 변수이다. 질소산화물은 대기오염의 주원인으로 오존층 파괴의 주범이기도 하다. 질소산화물은 주로 자동차 배기가스나 공장지대로부터 다량 방출된다.

한편 일산화질소의 농도는 연속형 변수이므로 히스토그램을 그려보았는데, NOX 변수 자체만으론 일정한 패턴을 가지고 있지 않았다. 따라서 연속형 변수들과의 상관계수를 계산하여 알아보았고, INDUS변수와 DIS 변수가 가장 밀접한 관계를 보였다. 비상권 지역일수록 일산화질소의 농도가 더 높았으며 보스턴 5대 상업지대와 가까워질수록 일산화질소의 농도는 높았다.

따라서 보스턴 5대 상업지대는 유동인구가 많을 것으로 추측되며 이는 자동차 배기가스로부터의 일산화질소라고 추측하였고, 상권이 발달되지 않은 지역은 공장지대라 생각하여 공장지대로 인한 고농도 일산화질소라고 추측하였다.

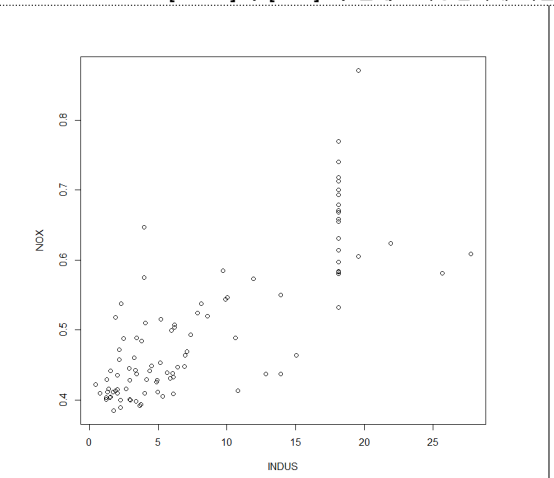


hist(boston$NOX,col='light cyan4',main='NOX',xlab='NOX')

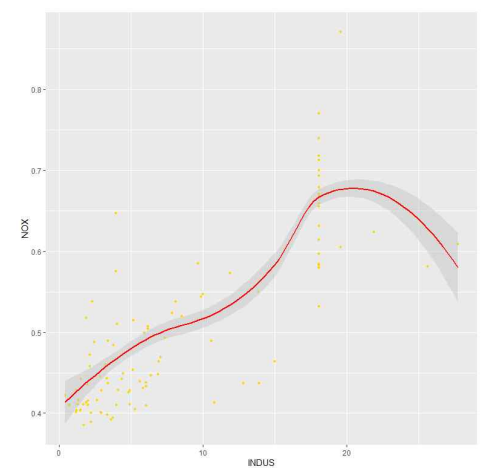


boxplot(boston$NOX,col='light cyan4',main='NOX')

[INDUS] 와 [NOX] 의 관계 : 비상업지대 비율이 올라갈수록 일산화질소의 농도가 올라감

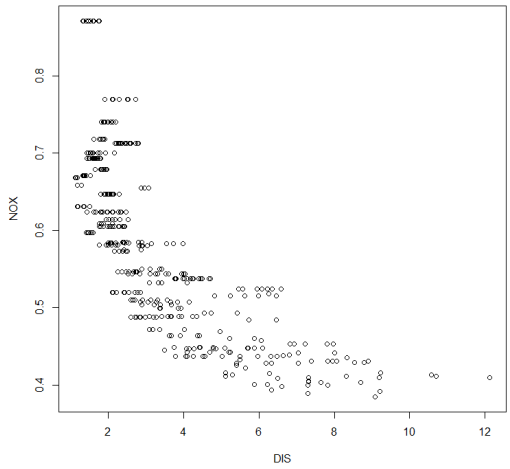


plot(boston$NOX~boston$INDUS,xlab='INDUS',ylab='NOX')

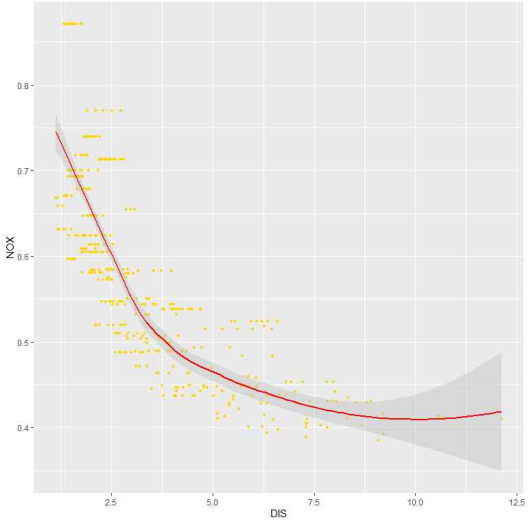


ggplot(data=boston,aes(x=INDUS,y=NOX))+geom_point(size=1, color="gold")+stat_smooth(color='red', fill='grey')

[DIS] 와 [NOX] 의 관계 : 보스턴 5대 상업지대와 멀어질수록 일산화질소의 농도는 낮아짐



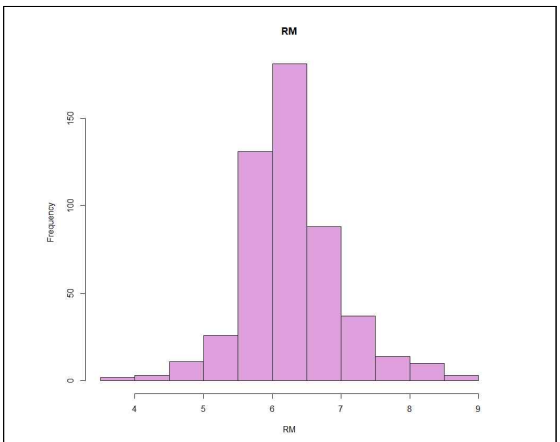
plot(boston$NOX~boston$DIS,xlab='DIS',ylab='NOX')



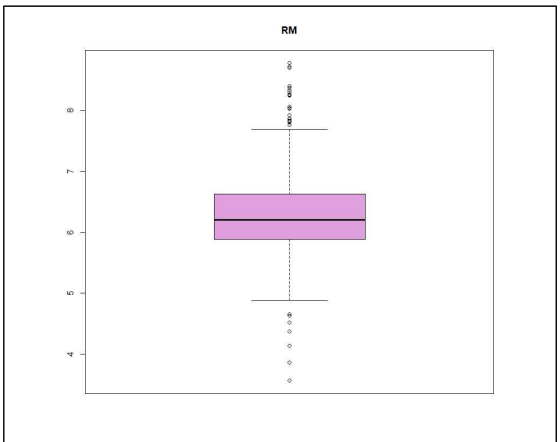
ggplot(data=boston,aes(x=DIS,y=NOX))+geom_point(size=1, color="gold")+stat_smooth(color='red', fill='grey')

#RM

RM은 가구당 평균 방의 개수로, 정규분포와 비슷한 형태로 분포하고 있다. 평균 방의 개수는 6개이며, 최소 3개에서 6개까지 존재한다.



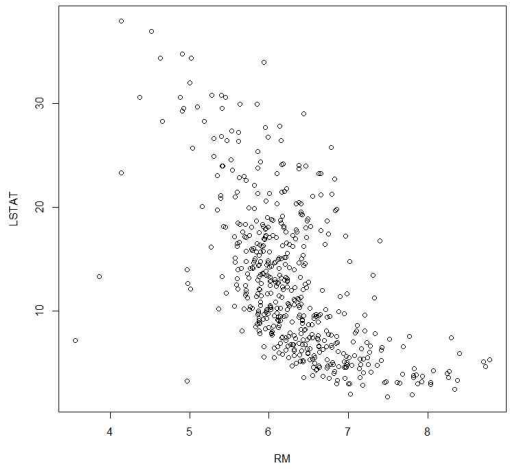
hist(boston$RM,col='plum',main='RM',xlab='RM')



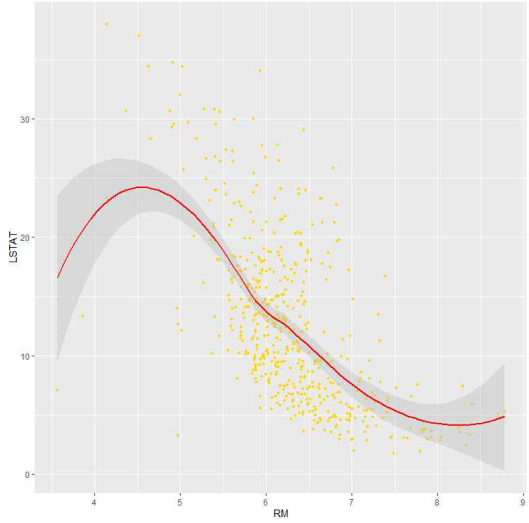
boxplot(boston$RM,col='plum',main='RM')

한편 타겟 변수를 제외한 다른 이산형 변수와의 상관계수를 비교해본 결과, LSTAT 변수가 RM 변수에 가장 밀접한 관계가 있었다. LSTAT 변수는 저소득층 비율이므로 거주지는 상대적으로 저렴해야하고, 방의 개수가 많을수록 집값은 높게 형성된다. 따라서 house의 방의 개수가 많아질수록 저소득층의 비율은 낮아진다고 해석할 수 있다.

[RM]와 [LSTAT]의 관계 : 방의 개수가 많아질수록 저소득층 비율은 낮아짐



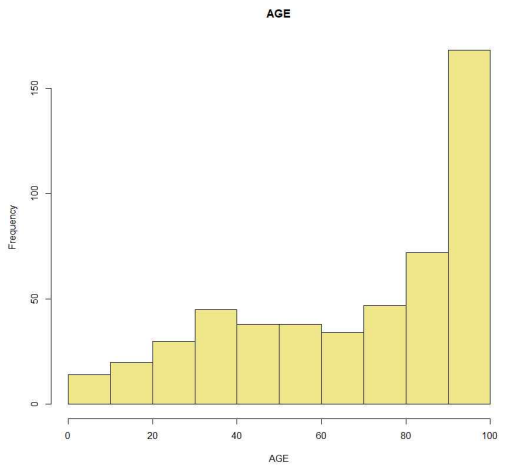
plot(boston$LSTAT~boston$RM,xlab='RM',ylab='LSTAT')



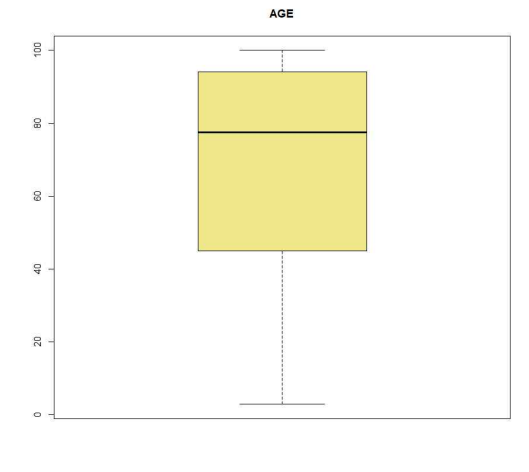
ggplot(data=boston,aes(x=RM,y=LSTAT))+geom_point(size=1, color="gold")+stat_smooth(color='red', fill='grey')

#AGE

AGE는 1940년 이전에 건축된 소유주택의 비율을 나타낸 변수이다. PROPORTION 에 관한 변수이며 전체를 100으로 둔 percent 개념으로 접근하여 해석하였다. 따라서 이에 맞게 데이터는 최소 0부터 100까지 연속형으로 존재하였고, 이에 맞게 히스토그램과 상자그림을 그려 시각화하였다. 평균은 68이며 대부분의 건물들이 1940년 이전에 건축되었음을 알 수 있다.



hist(boston$AGE,col='khaki',main='AGE',xlab='AGE')



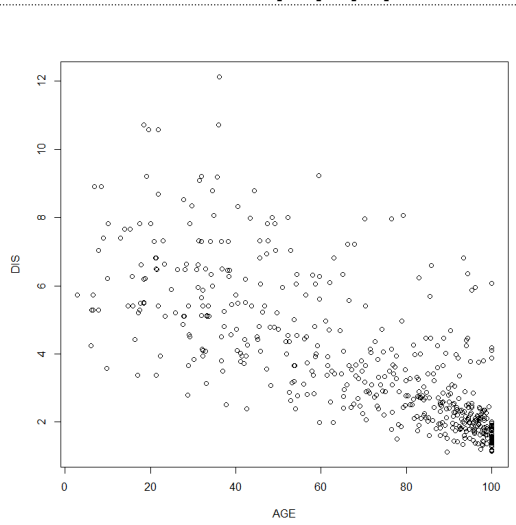
boxplot(boston$AGE,col='khaki',main='AGE')

다른 연속형 변수와의 상관관계를 따져볼 때 가장 연관성 있는 변수는 의외로 DIS (보스턴 5대 상업지구와의 거리) 변수였다. 상관계수는 -0.75로 강한 음의 상관관계를 보였다. 해석해보면 상업지구와 가까울수록 1940년 이전에 건축된 비율이 올라감을 알 수 있다.

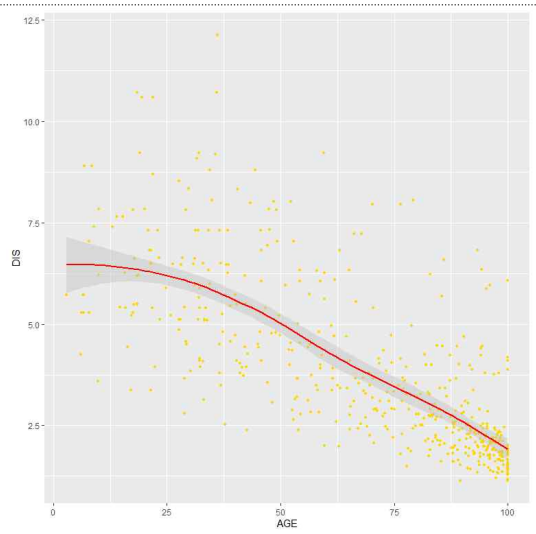
또한 앞서 언급했듯이 DIS와 NOX는 높은 상관관계를 보이고 있다. 흥미로운 점은 AGE도 NOX와 강한 상관관계를 보인다는 것인데 0.73으로 강한 양의 상관관계이다. 오래된 건축물이

많을수록 일산화질소의 농도가 높음을 알 수 있다. 따라서 DIS, NOX, AGE는 서로 밀접한 관계를 이루고 있음을 알 수 있다. 정리하자면, 보스턴 5대 상업지대일수록 일산화질소의 농도가 짙으며 이곳은 1940년대 이전에 건축된 건물들이 많다.

[AGE]와 [DIS]의 관계 : 오래된 건축물일수록 상업지대와 가깝다



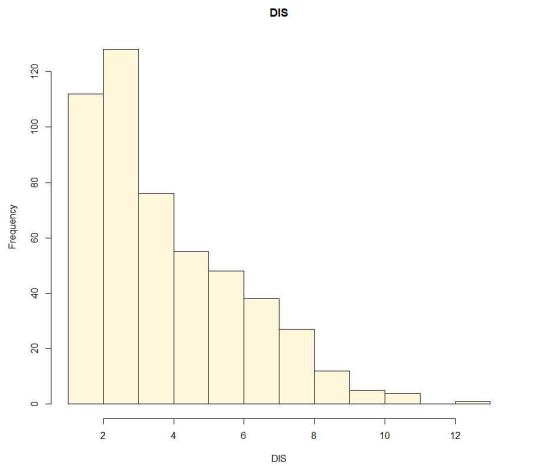
plot(boston$DIS~boston$AGE,xlab='AGE',ylab='DIS')



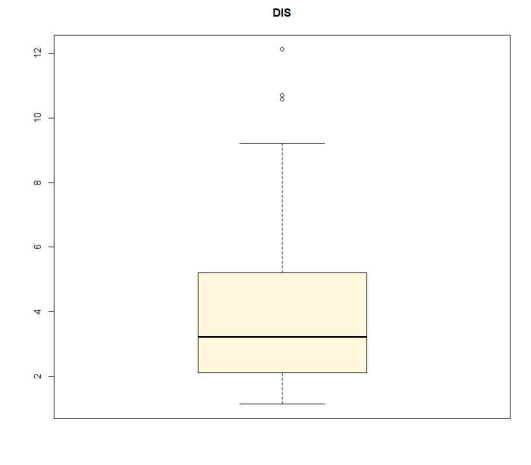
ggplot(data=boston,aes(x=AGE,y=DIS))+geom_point(size=1, color="gold")+stat_smooth(color='red', fill='grey')

#DIS

DIS는 보스턴 5대 상업지대와의 거리를 나타내는 변수이다. 연속형 변수에 맞게 히스토그램과 상자그림을 그려 시각화하였다. 그래프는 우향 왜곡 분포를 띠고 있으며 평균은 3.8이다. 대부분의 house들이 보스턴 5대 상업지대와 가까이 위치하고 있음을 알 수 있다.



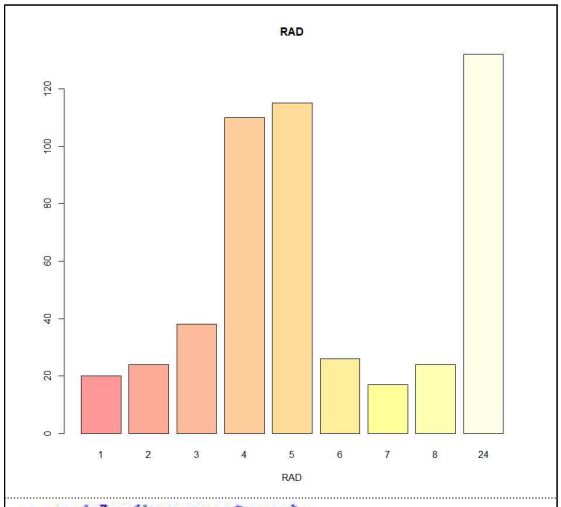
hist(boston$DIS,col='cornsilk',main='DIS',xlab='DIS')



boxplot(boston$DIS,col='cornsilk',main='DIS')

#RAD

RAD는 고속도로까지의 접근성을 지수(quotient)로 표현한 것이다. 접근성을 기준으로 그룹화하였다고 생각하여 범주형 변수 중 순서형 변수라고 판단하였다. 따라서 막대그래프를 그려 시각화하였다. 가장 두들어지는 그룹은 '3', '4' ,'24' 이다.



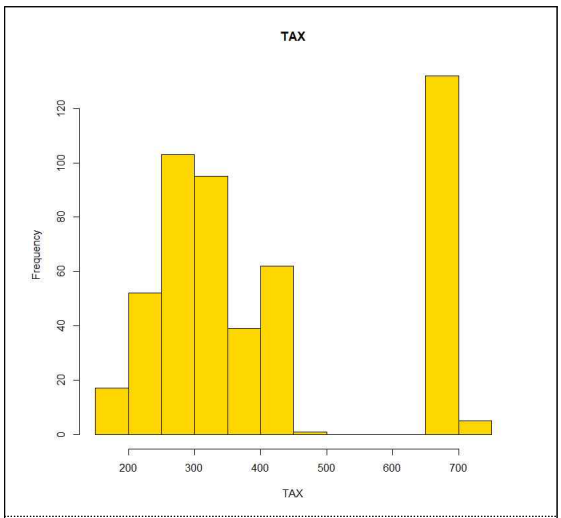

```
> table(boston$RAD)

  1   2   3   4   5   6   7   8  24
 20  24  38 110 115  26  17  24 132
```

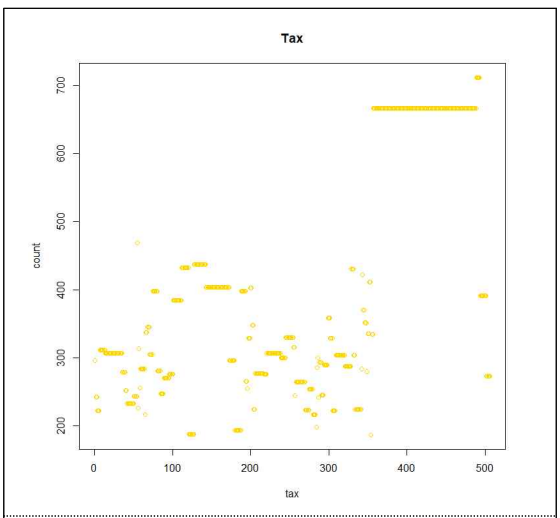
```
rad<-table(boston$RAD)
barplot(rad,xlab='RAD',main='RAD',col=heat.colors(9,alpha=0.4))
```

#TAX

TAX는 10,000달러당 종합재산세율로, 세금이 얼마나 부과되는지에 관한 변수이다. 구간이 정해진 변수가 아니기에 연속형 변수로 판단하였고 히스토그램를 그려 시각화하였다. 히스토그램을 보고 값의 분포파악을 더 직관적으로 하기 위해 상자그림이 아닌 산점도로 시각화하였다. TAX는 187에서 711까지 분포하고 있었으며 대부분은 500 이하에 분포하고 있었으나, 700 근방에도 몰려있음을 확인하였다.



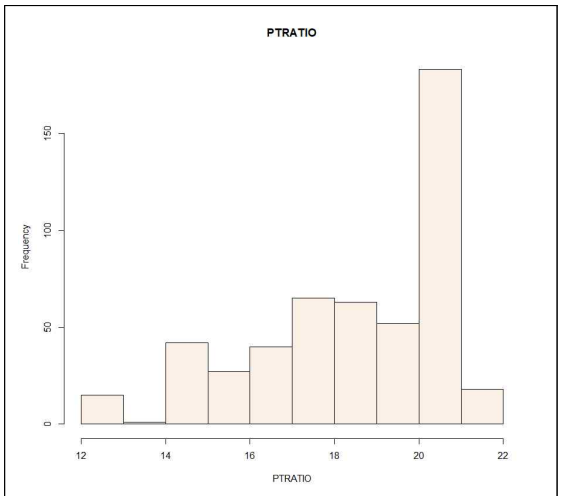

```
hist(boston$TAX,col='gold',main='TAX',xlab='TAX')
```



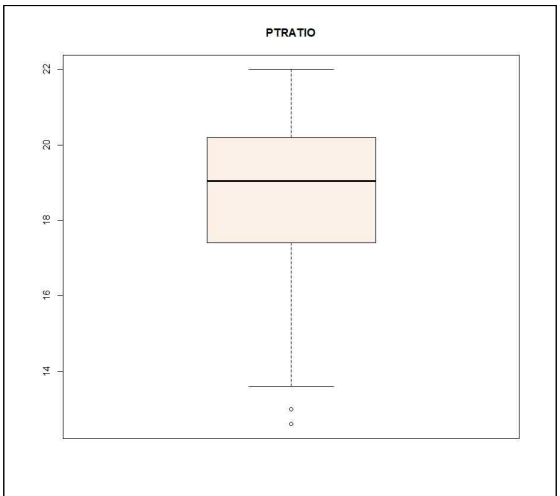

```
plot(boston$TAX,main='Tax',xlab='tax',ylab='count',col='gold')
```

#PTRATIO

PTRATIO는 town별 학생 대 교사의 비율을 나타낸 변수이다. 연속형 변수에 맞게 히스로그램과 상자그림을 그려 시각화하였다. 그래프는 좌향 왜곡 분포이며 20과 21 사이에 가장 많은 데이터들이 몰려있음을 알 수 있다. PTRATIO의 평균은 18.45로, 평균적으로 교사 한 명당 학생 18명임을 알 수 있다.



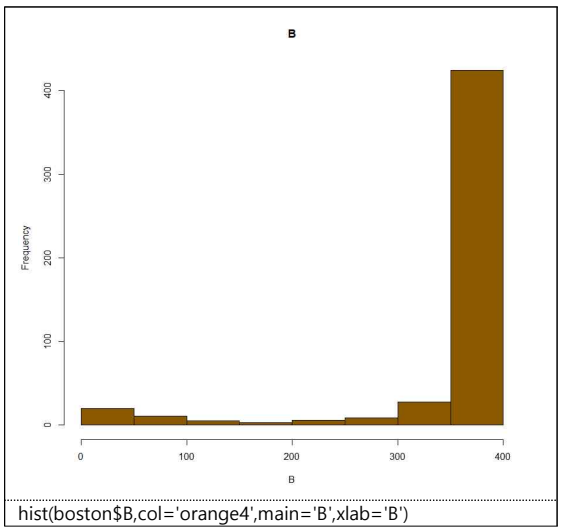

```
hist(boston$PTRATIO,col='linen',main='PTRATIO',xlab='PTRATIO')
```



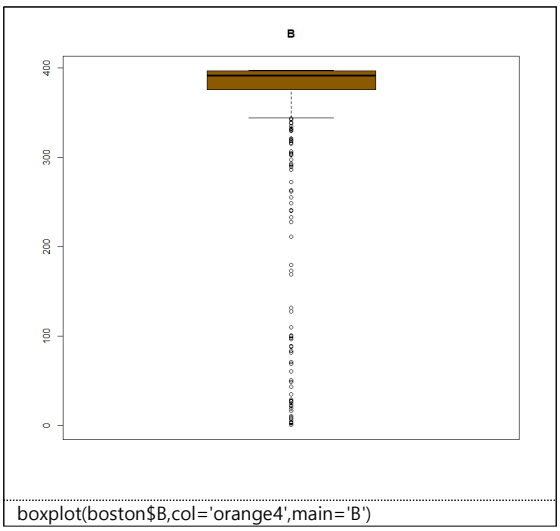

```
boxplot(boston$PTRATIO,col='linen',main='PTRATIO')
```

#B

B는 도시별 흑인의 비율을 $1000(\text{흑인의 비율}-0.63)^2$ 에 대입하여 얻은 변수이다. 흑인의 비율은 0이상의 유리수이다. B는 흑인의 비율이 0.63 일 때, 0으로 최솟값이고, 흑인의 비율이 0 일 때, 396.9로 최댓값이다. 하지만 0.37부터 1까지의 함수값은 대칭으로 이루어져 있기 때문에 정확한 판단은 할 수 없다. 정확한 것은 $136.6 < B < 396.9$ 인 경우는 흑인의 비율(Bk)은 $0.26 > Bk > 0$ 임은 알 수 있다. 이를 바탕으로 보면 대부분의 값들은
$136.6 < B < 396.9$ 사이에서 이루어져 있으므로 흑인의 비율은 대부분 0.26 이하에 존재할 것이다. 따라서 대부분의 town에 존재하는 흑인의 비율은 높지 않다고 판단하였다.



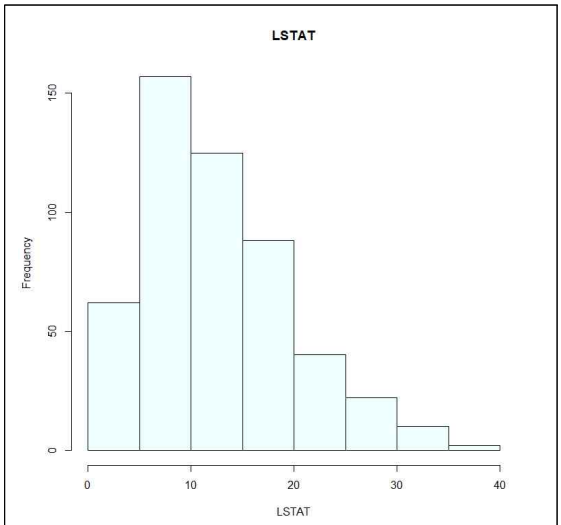

```
hist(boston$B,col='orange4',main='B',xlab='B')
```



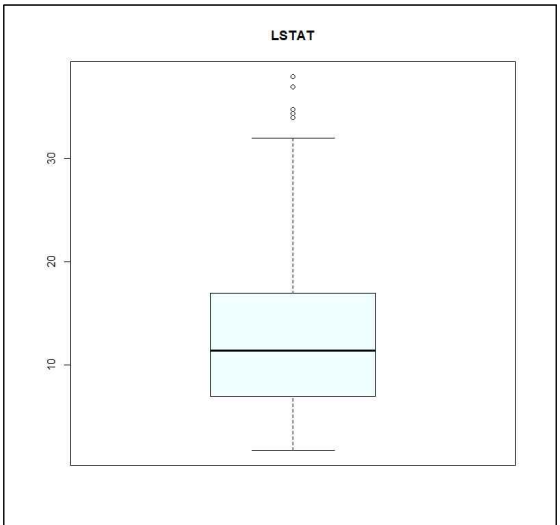

```
boxplot(boston$B,col='orange4',main='B')
```

#LSTAT

LSTAT는 저소득층의 비율을 나타내는 변수로 연속형 변수이다. 따라서 히스토그램과 상자그림을 이용하여 시각화하였다. 그래프는 우향 왜곡 분포를 띠고 있으며 10 근방에 가장 많이 몰려있다. LSTAT의 평균은 12.65 이다.



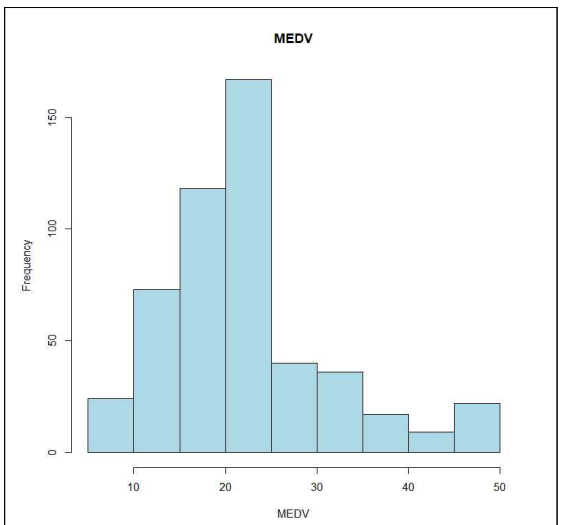

```
hist(boston$LSTAT,col='azure',main='LSTAT',xlab='LSTAT')
```



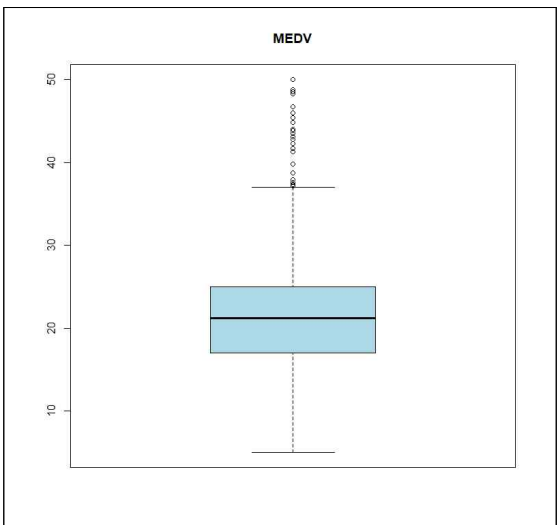

```
boxplot(boston$LSTAT,col='azure',main='LSTAT')
```

#MEDV

MEDV는 타겟변수로, 소유자 주택 가격의 중앙값을 나타내는 변수이다. 연속형 변수이므로 히스토그램과 상자그림을 그려 시각화하였다. MEDV의 평균은 22.53 으로, $1,000단위임을 고려하면 $22,530임을 알 수 있다. 또한 최솟값은 5 ($5,000) 이며 최댓값은 50 ($50,000) 이다.




```
hist(boston$MEDV,col='lightblue',main='MEDV',xlab='MEDV')
```




```
boxplot(boston$MEDV,col='lightblue',main='MEDV')
```

#연속형 변수들 사이의 상관관계

다음은 상관계수를 구하여 표현한 것이고, 이를 직관적으로 판단하기 위해 시각화한 그림이다. 푸른 계열로 갈수록 양의 상관관계를 뜻하며, 반대로 붉은 계열은 음의 상관관계를 뜻한다. 또한 원의 크기가 클수록 상관관계가 강하다는 것이다.

한편 이를 통해 연속형 변수 간 상관관계를 따져보았는데, 타겟 변수를 제외한 변수 간 비교는 앞서 언급하였으므로 대략적인 그래프만 다시 표기하였다.

<연속형 변수만 모아놓은 테이블>

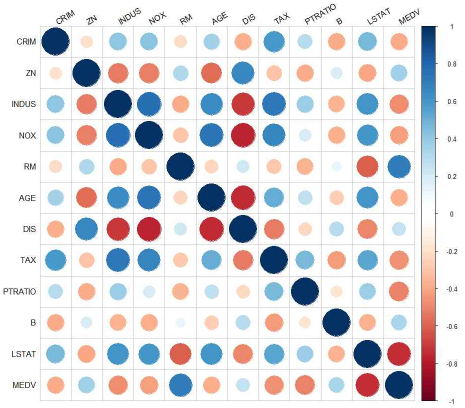
	CRIM	ZN	INDUS	NOX	RM	AGE	DIS	TAX	PTRATIO	B	LSTAT	MEDV
1	0.00632	18.0	2.31	0.5380	6.575	65.2	4.0900	296	15.3	396.90	4.98	24.0
2	0.02731	0.0	7.07	0.4690	6.421	78.9	4.9671	242	17.8	396.90	9.14	21.6
3	0.02729	0.0	7.07	0.4690	7.185	61.1	4.9671	242	17.8	392.83	4.03	34.7
4	0.03237	0.0	2.18	0.4580	6.998	45.8	6.0622	222	18.7	394.63	2.94	33.4
5	0.06905	0.0	2.18	0.4580	7.147	54.2	6.0622	222	18.7	396.90	5.33	36.2
6	0.02985	0.0	2.18	0.4580	6.430	58.7	6.0622	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0.5240	6.012	66.6	5.5605	311	15.2	395.60	12.43	22.9
8	0.14455	12.5	7.87	0.5240	6.172	96.1	5.9505	311	15.2	396.90	19.15	27.1
9	0.21124	12.5	7.87	0.5240	5.631	100.0	6.0821	311	15.2	386.63	29.93	16.5
10	0.17004	12.5	7.87	0.5240	6.004	85.9	6.5921	311	15.2	386.71	17.10	18.9
11	0.22489	12.5	7.87	0.5240	6.377	94.3	6.3467	311	15.2	392.52	20.45	15.0
12	0.11747	12.5	7.87	0.5240	6.009	82.9	6.2267	311	15.2	396.90	13.27	18.9
13	0.09378	12.5	7.87	0.5240	5.889	39.0	5.4509	311	15.2	390.50	15.71	21.7
14	0.62976	0.0	8.14	0.5380	5.949	61.8	4.7075	307	21.0	396.90	8.26	20.4
15	0.63796	0.0	8.14	0.5380	6.096	84.5	4.4619	307	21.0	380.02	10.26	18.2
16	0.62739	0.0	8.14	0.5380	5.834	56.5	4.4986	307	21.0	395.62	8.47	19.9
17	1.05393	0.0	8.14	0.5380	5.935	29.3	4.4986	307	21.0	386.85	6.58	23.1
18	0.78420	0.0	8.14	0.5380	5.990	81.7	4.2579	307	21.0	386.75	14.67	17.5
19	0.80271	0.0	8.14	0.5380	5.456	36.6	3.7965	307	21.0	288.99	11.69	20.2
20	0.72580	0.0	8.14	0.5380	5.727	69.5	3.7965	307	21.0	390.95	11.28	18.2
21	1.25179	0.0	8.14	0.5380	5.570	98.1	3.7979	307	21.0	376.57	21.02	13.6
22	0.85204	0.0	8.14	0.5380	5.965	89.2	4.0123	307	21.0	392.53	13.83	19.6
23	1.23247	0.0	8.14	0.5380	6.142	91.7	3.9769	307	21.0	396.90	18.72	15.2
24	0.98843	0.0	8.14	0.5380	5.813	100.0	4.0952	307	21.0	394.54	19.88	14.5
25	0.75026	0.0	8.14	0.5380	5.924	94.1	4.3996	307	21.0	394.33	16.30	15.6
26	0.84054	0.0	8.14	0.5380	5.599	85.7	4.4546	307	21.0	303.42	16.51	13.9
27	0.67191	0.0	8.14	0.5380	5.813	90.3	4.6820	307	21.0	376.88	14.81	16.6
28	0.95577	0.0	8.14	0.5380	6.047	88.8	4.4534	307	21.0	306.38	17.28	14.8
29	0.77299	0.0	8.14	0.5380	6.495	94.4	4.4547	307	21.0	387.94	12.80	18.4
30	1.00245	0.0	8.14	0.5380	6.674	87.3	4.2390	307	21.0	380.23	11.98	21.0

```
boston%>%select(CRIM,ZN,INDUS,NOX,RM,AGE,DIS,TAX
,PTRATIO,B,LSTAT,MEDV)->conti
View(conti)
```

*결과값이 너무 많아서 뒷부분은 생략함.

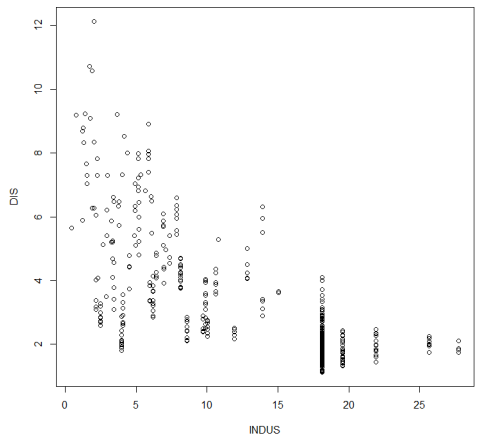
```
> round(cor(conti2),2)->a;a
        CRIM    ZN INDUS   NOX    RM   AGE   DIS   TAX PTRATIO     B LSTAT  MEDV
CRIM    1.00 -0.20  0.41  0.42 -0.22  0.35 -0.38  0.58    0.29 -0.39  0.46 -0.39
ZN     -0.20  1.00 -0.53 -0.52  0.31 -0.57  0.66 -0.31   -0.39  0.18 -0.41  0.36
INDUS   0.41 -0.53  1.00  0.76 -0.39  0.64 -0.71  0.72    0.38 -0.36  0.60 -0.48
NOX     0.42 -0.52  0.76  1.00 -0.30  0.73 -0.77  0.67    0.19 -0.38  0.59 -0.43
RM     -0.22  0.31 -0.39 -0.30  1.00 -0.24  0.21 -0.29   -0.36  0.13 -0.61  0.70
AGE     0.35 -0.57  0.64  0.73 -0.24  1.00 -0.75  0.51    0.26 -0.27  0.60 -0.38
DIS    -0.38  0.66 -0.71 -0.77  0.21 -0.75  1.00 -0.53   -0.23  0.29 -0.50  0.25
TAX     0.58 -0.31  0.72  0.67 -0.29  0.51 -0.53  1.00    0.46 -0.44  0.54 -0.47
PTRATIO 0.29 -0.39  0.38  0.19 -0.36  0.26 -0.23  0.46    1.00 -0.18  0.37 -0.51
B      -0.39  0.18 -0.36 -0.38  0.13 -0.27  0.29 -0.44   -0.18  1.00 -0.37  0.33
LSTAT   0.46 -0.41  0.60  0.59 -0.61  0.60 -0.50  0.54    0.37 -0.37  1.00 -0.74
MEDV   -0.39  0.36 -0.48 -0.43  0.70 -0.38  0.25 -0.47   -0.51  0.33 -0.74  1.00
```

conti2<-na.omit(conti); cor(conti2)
round(cor(conti2),2)->a;a

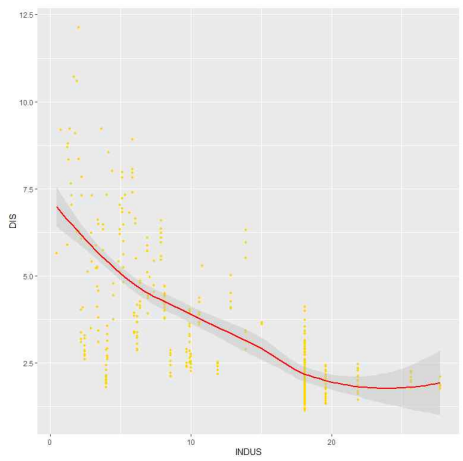


corrplot(a,method="circle",tl.col="black",tl.srt=30)

[INDUS] 와 [DIS]의 관계 : 비상업지대 비율이 올라갈수록 상업지대로부터 가까워짐

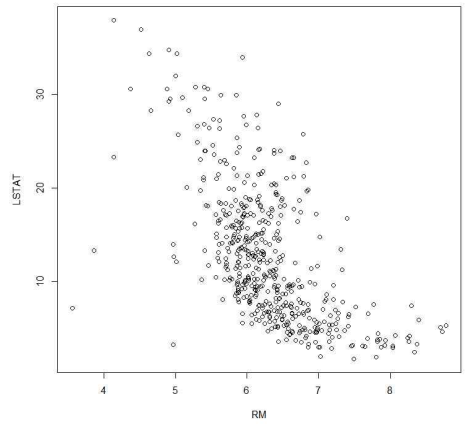


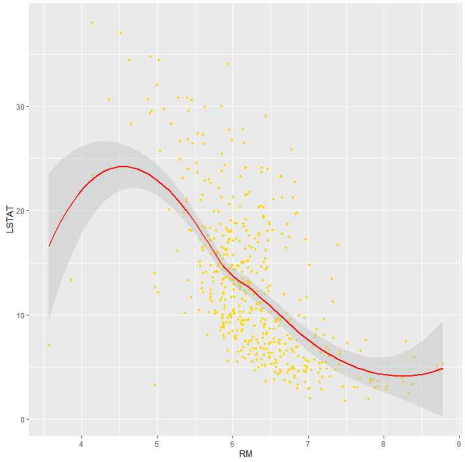
plot(boston$NOX~boston$INDUS,xlab='INDUS',ylab='NOX')



ggplot(data=boston,aes(x=INDUS,y=NOX))+geom_point(size=1, color="gold")+stat_smooth(color='red', fill='grey')

[RM]와 [LSTAT]의 관계 : 방의 개수가 많아질수록 저소득층 비율은 낮아짐

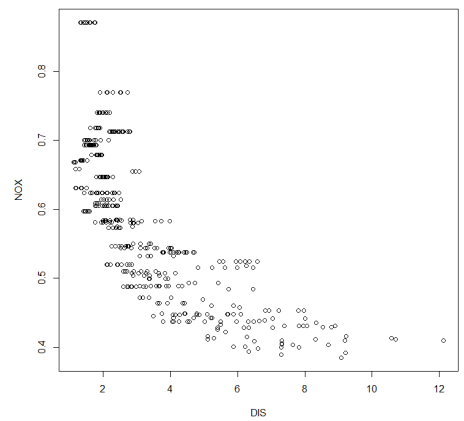


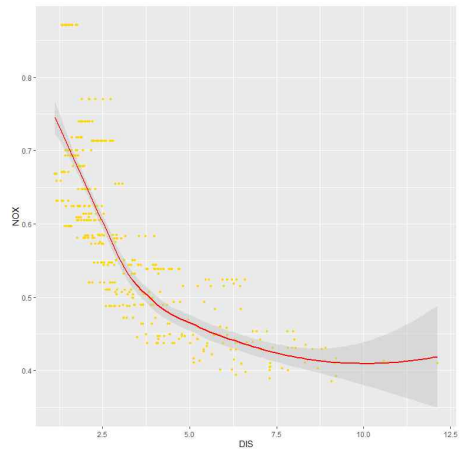


plot(boston$LSTAT~boston$RM,xlab='RM',ylab='LSTAT')

ggplot(data=boston,aes(x=RM,y=LSTAT))+geom_point(size=1, color="gold")+stat_smooth(color='red', fill='grey')

[DIS] 와 [NOX] 의 관계 : 보스턴 5대 상업지대와 멀어질수록 일산화질소의 농도는 낮아짐

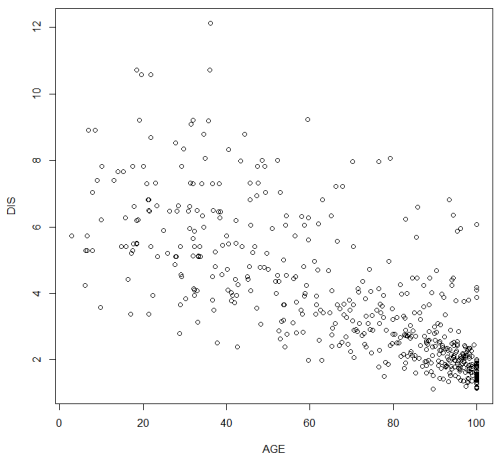


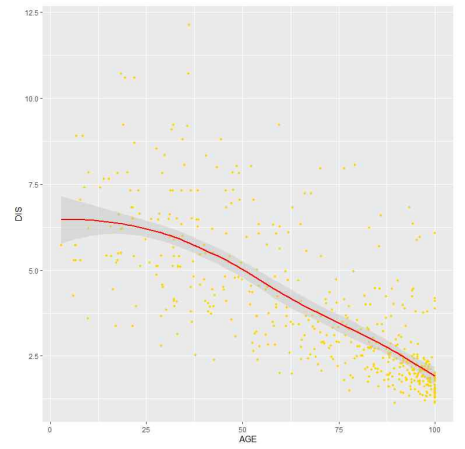


plot(boston$NOX~boston$DIS,xlab='DIS',ylab='NOX')

ggplot(data=boston,aes(x=DIS,y=NOX))+geom_point(size=1, color="gold")+stat_smooth(color='red', fill='grey')

[AGE]와 [DIS]의 관계 : 오래된 건축물일수록 상업지대와 가깝다





plot(boston$DIS~boston$AGE,xlab='AGE',ylab='DIS')

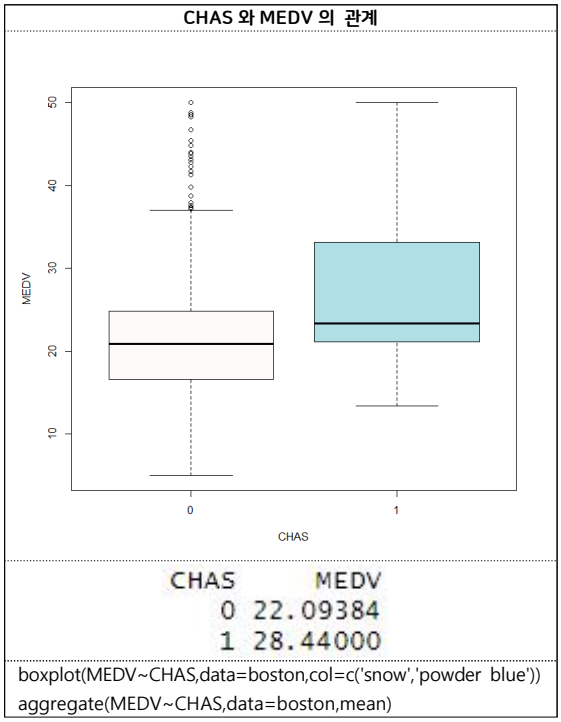
ggplot(data=boston,aes(x=AGE,y=DIS))+geom_point(size=1, color="gold")+stat_smooth(color='red', fill='grey')

#Target Variable (MEDV) 와 변수 사이의 관계

지금까지 각 변수 별 형태를 살펴보았다. 결국 타겟변수는 MEDV이므로 각 변수와 MEDV 간의 관계를 살펴볼 것이다. 이때 연속형 변수와 타겟 변수 간의 관계는 앞서 보인 연속형 변수의 상관관계를 바탕으로 진행하였다. 상관계수가 |0.5|이상인 변수만 유의하게 보았고, 따라서 연속형 변수는 RM, PTRATIO, LSTAT 만 타겟변수와 비교하였다.

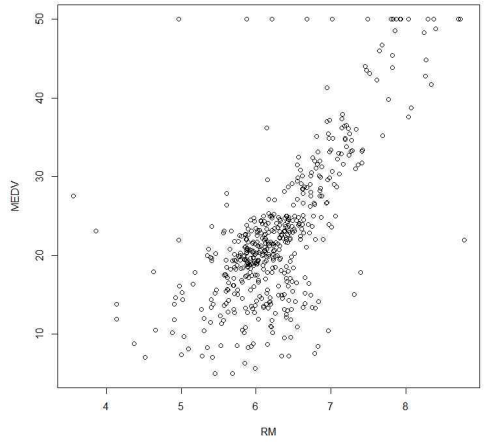
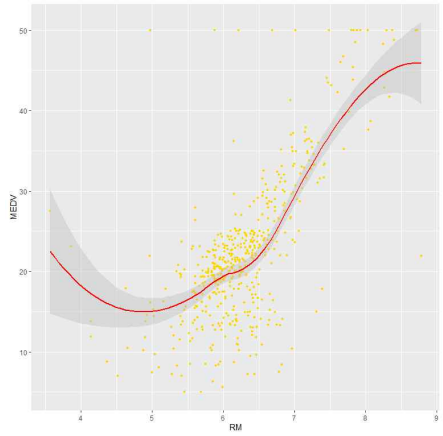
#CHAS 와 MEDV

다음은 강가 주변 위치에 따른 house 의 가격을 나타내는 그래프이다. 앞서 한강을 예시로 들며 강가 주변의 집값이 더 높을 것이라 추측하였는데, 그에 맞는 그래프 모양이 나옴을 확인할 수 있다. 찰스강 경계에 위치한 house 일수록 집값이 높게 형성되어 있음을 알 수 있다.



#RM 과 MEDV

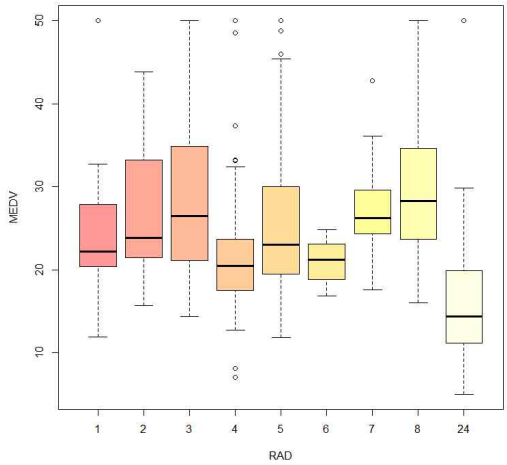
다음은 가구 당 평균 방의 개수에 따른 house 의 가격을 나타내는 그래프이다. 두 변수의 상관계수는 0.7 로 강한 양의 상관관계를 띠고 있다. 방의 개수가 많을수록 house 의 가격도 올라감을 알 수 있다.

RM 과 MEDV 의 관계	
	
plot(boston$MEDV~boston$RM,xlab='RM',ylab='MEDV')	ggplot(data=boston,aes(x=RM,y=MEDV))+geom_point(size=1, color="gold")+stat_smooth(color='red', fill='grey')

#RAD 와 MEDV

다음은 고속도로와의 접근성에 따른 house 의 가격을 나타내는 그래프이다. 접근성 지수가 10 이하인 경우 house 의 가격은 비슷하였으나, 접근성 지수가 24인 경우는 약 절반가량 줄어든 가격을 볼 수 있다. 따라서 접근성 지수가 10 이하인 경우는 house 의 가격과 관련이 낮지만, 24 이상인 경우 house 의 가격이 떨어짐을 알 수 있다. house 에서 고속도로까지 일정 수준 이상 떨어진 경우엔 집값이 낮게 형성된다고 해석할 수 있다.

RAD 와 MEDV 의 관계



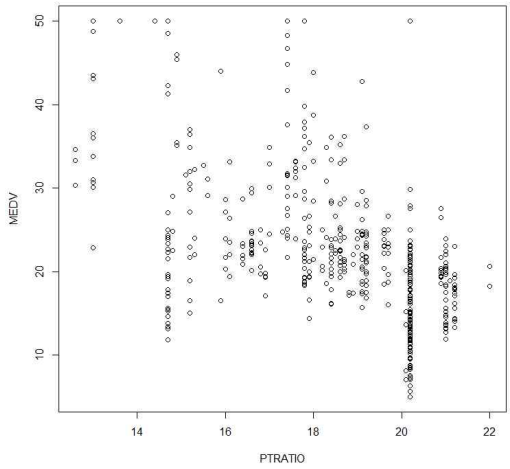
RAD	MEDV
1	24.36500
2	26.83333
3	27.92895
4	21.38727
5	25.70696
6	20.97692
7	27.10588
8	30.35833
24	16.40379

```
boxplot(MEDV~RAD,data=boston,col=heat.colors(9,alpha=0.4))
aggregate(MEDV~RAD,data=boston,mean)
```

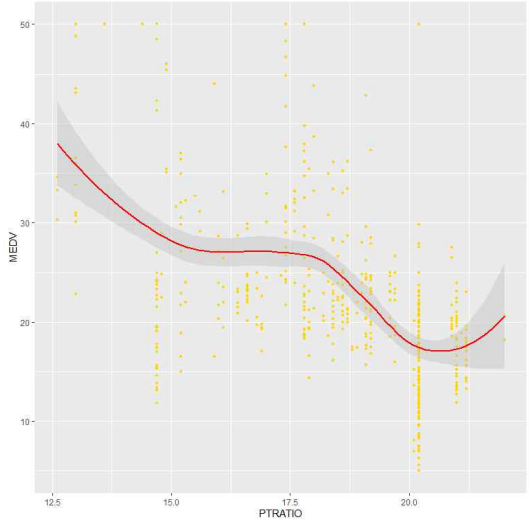
#PTRATIO 와 MEDV

다음은 학생 대 교사의 비율에 따른 house 의 가격을 나타내는 그래프이다. 두 변수의 상관계수는 -0.51 로 유의미한 음의 상관계수를 보인다. 교사가 감당해야할 학생의 수가 많아지는 곳일수록 house 의 가격은 낮아진다.

PTRATIO 와 MEDV 의 관계



plot(boston$MEDV~boston$B,xlab='B',ylab='MEDV')

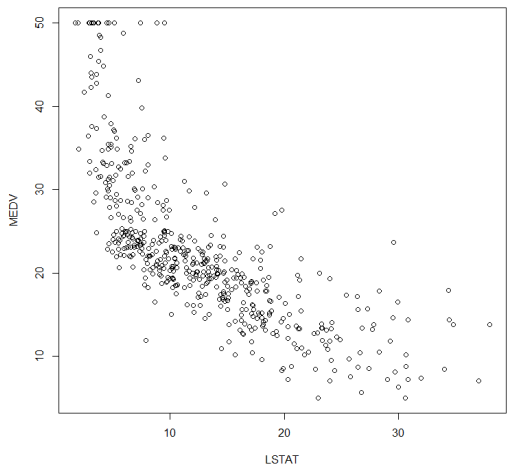


ggplot(data=boston,aes(x=B,y=MEDV))+geom_point(size=1, color="gold")+stat_smooth(color='red', fill='grey')

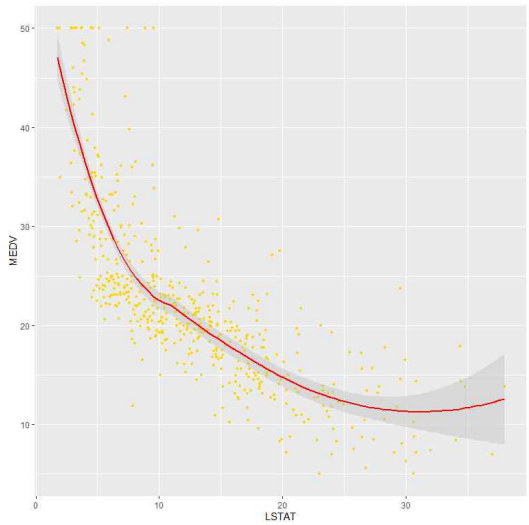
 # LSTAT와 MEDV

다음은 저소득층 비율에 따른 house 의 가격을 나타내는 그래프이다. 두 변수의 상관계수는 -0.74 로 강한 음의 상관관계를 보인다. 저소득층의 비율이 높아질수록 house 의 가격은 낮아짐을 알 수 있다. 저소득층에 해당하는 거주민들은 가격이 높은 house에 살 경제력이 갖춰지지 않았고 따라서 그들이 선호하는 지역은 낮은 가격의 house 일 것이라 해석된다.

LSTAT 와 MEDV 의 관계



plot(boston$MEDV~boston$LSTAT,xlab='LSTAT',ylab='MEDV')



ggplot(data=boston,aes(x=LSTAT,y=MEDV))+geom_point(size=1, color="gold")+stat_smooth(color='red', fill='grey')

#결론

결론적으로 House 는 찰스강 경계에 위치할수록, 방의 개수가 많을수록, 고속도로의 접근성이 뛰어날수록, 학생 대 교사의 비율이 낮을수록, 저소득층의 비율이 낮을수록 가격이 높게 형성될 것이다.