Reyeme Affiar

2023-2학기 통계자료분석 1조

2023-12-06

## 1. 데이터 정리

### 1-1. 결측값

주어진 데이터는 456개의 관측치와 17개의 변수로 구성되어 있다. 이 중 결측값 9개를 확인했다.

# 결측값 존재 확인  
  
sum(is.na(data1))

[1] 9

# 결측값 있는 행 확인  
  
which(rowSums(is.na(data1)) > 0)

[1] 27 181 248 254 350 368 383 454 455

결측값 중 한개를 제외하곤 모두 UNIT 열 값이었다. UNIT는 건물내 특정 주거공간을 나타내는 호수(ex 101호, 102호)이다. Reyem Affiar가 구매하려는 건물의 UNIT이 알려져 있지 않았기 때문에 Reyem Affiar가 구매하려는 건물 가격을 추정하는데에 UNIT 변수는 사용하지 않기로 결정했다.

나머지 한개의 결측값은 First Date에서 나타났으며, 254행에 있다. FirstDate열은 차후 HousePriceIndex 변수의 각 행의 값를 정하는 기준으로 사용한다. 때문에 First Date 값이 누락된 254행을 제거하기로 결정했다.

### 1-2. 변수 선택

#### 1) Area

건물 가격을 예측함에 있어서 지역변수는 매우 중요하다고 판단한다. 같은 도시라고 하여도 그 지역, 입지에 따라 건물 가격의 평균이 달라진다. 예를 들어, 강남구의 건물과 금천구의 건물은 같은 서울일지라도 크게 다르다. Cambridge 역시 마찬가지 행태를 보일 것이라 예상한다.

#### 2) FirstDate

FirstDate는 SalePrice와 높은 상관관계를 갖는 Price1이 정해진 시기이다. 건물 가격을 결정하는데 시점은 중요한 의미를 갖는다. 시점에 따라 부동산 시장의 상황은 크게 변화할 수 있다. 예를 들어, 현재와 같은 고금리 상황에서는 대출이자 부담으로 인해 부동산 시장이 크게 꺾이는 시점이 되기도 한다. 때문에, 건물 가격을 예측하는데 시점변수는 매우 중요하다. 그러나, FirstDate를 독립변수로 하고, SalePrice를 종속변수로 하여 분산분석 결과를 사용하는 것보다는, FRED에서 제공하는 1989년에서 1994년 Cambridge 지역 주택가격 지수를 사용하는 것이 건물가격을 추정하는데 더 나은 정확도를 제공할 것이라고 판단했다.

#### 3) HousePriceIndex

FirstDate는 1993-07-17 와 같은 형식으로 된 데이터이다. 여기에서 연월만 추출하여 다시 분기로 대체한 데이터를 사용했다. 예를 들어, 1993-07-17 데이터는 1993-3Q 와 같이 ‘연 - 분기’ 형식으로 대체했다.

이와 같이 분기별로 나눈 이유는 FRED에서 제공하는 “All-Transactions House Price Index for Cambridge-Newton-Framingham, MA”(https://fred.stlouisfed.org/series/ATNHPIUS15764Q) 자료가 분기별로 값을 제공하기 때문이다.

위 자료는 1995년 1분기의 가격 지수를 100 기준으로 한다. 자료에 따르면, 1989년 1분기(데이터에서 1920년을 제외한 가장 오래된 관측치)에서 1994년 2분기(Reyme Affiar가 건물 매수를 고려하는 시점) 사이 중, 1989년 4분기의 가격지수가 108.92로 가장 높았으며, 1992년 2분기의 가격지수가 96.83으로 가장 낮다.

FRED에서 제공하는 House Price Index 분기별 자료와 각 관측치의 ’연-분기’를 대조하여 HousePriceIndex 변수를 만들었다. 예를 들어, 1993-3Q 변수값을 갖는 관측치는 HousePriceIndex 변수값으로 98.70을 갖게 된다.

#### 4) Days

첫 호가를 올린 시점부터 판매 시점까지의 기간이 길어질수록 SalePrice가 떨어질 것이라 판단했다. 부동산 시장은 보통 매수자가 있기 마련이라 대부분의 경우 적절한 기간이 지나면 거래가 체결된다. 그러나, 건물의 호가가 시장이 생각하는 가격에 비해 높거나, 부동산 시장 상황 악화로 매수자가 줄어드는 경우 거래완료까지 기간이 늘어날 수 있다. 이 경우 매도자는 호가를 낮춰 거래를 성사시키려 하기 때문에, Days가 커질수록, SalePrice는 떨어진다.

#### 5) Interior

내부면적으로 공용면적을 제외한 수치이다. 내부면적은 SalePrice를 결정하는데 중요한 변수라 판단한다. 내부면적이 커질 수록 SalePrice가 커질 것이다. 그러나, Interior 변수는 공용면적을 포함하고 있지 않기 때문에, 공용면적을 나타내는 다른 변수가 필요하다.

#### 6) Condo

매달 관리비를 나타내는 변수다. 관리비는 utility(가스비, 수도세, 전기세 등)와 공용공간 관리비를 포함한다. 예를 들어, 대부분의 오피스텔 관리비에 포함된 복도, 엘리베이터, 분리수거장 등 공용공간에 대한 청소용역비가 Condo에 해당할 것이다. Condo는 공용면적이 늘어날 수록 비례해서 늘어날 것이라 추측할 수 있다.

Condo가 클수록 헬스장, 수영장등을 포함한 공용면적이 넓은 고급 주거공간일 수 있기 때문에, SalePrice가 커질 것이라 판단했다.

#### 7) Rooms

방의 수는 Interior(내부 면적)과 높은 상관관계를 갖고 있기 때문에 두 변수를 사용할 경우, 다중공선성이 일어날 수 있다. 하지만 내부면적이 같은 경우에도 방의 수가 너무 많다면, 선호도가 떨어져 SalePrice가 낮아지는 경우가 생길 수 있다. 즉, Rooms는 내부면적과 관련이 있지만, 건물의 퀄리티(설계 수준)과 관련이 있기 때문에 SalePrice를 결정하는 요인으로 판단했다.

#### 8) Bed

침실의 수는 방의 수와 높은 상관관계를 갖기 때문에 역시 다중공선성을 조심해야 하는 변수이다. 그러나, 방의 수로만 설명할 수 없는 지점을 침실의 수 변수와 함께 설명할 수 있다. 예를 들어, 방의 수가 동일한 건물에서 침실의 수가 너무 많다면, 작은 침실이 많은 구조도를 갖고 있을 것이다. 작은 침실은 보통 선호도가 떨어져 SalePrice가 낮아지는 경우가 생긴다. 즉, Bed 역시 건물의 퀄리티(설계 수준)와도 관계가 되기 때문에 SalePrice를 결정하는 요인으로 판단했다.

#### 9) Bath

화장실의 수는 방의 수와 높은 상관관계를 갖기 때문에 역시 다중공선성을 조심해야 하는 변수이다. 그러나, 침실의 수로만 설명할 수 없는 지점을 화장실의 수 변수와 함께 설명할 수 있다. 예를 들어, 침실의 수가 3개 이상이어서 4인 이상 가구가 거주할 것으로 예상하는 건물에, 화장실의 수가 1개인 건물과 2개인 건물에 대한 선호도는 다를 것이다.(보통 화장실이 2개인 건물을 선호할 것이다.) 이로 인해 SalePrice에 차이가 생길 수 있다. 즉, Bath 역시 건물의 퀄리티(설계 수준)와도 관계가 되기 때문에 SalePrice를 결정하는 요인으로 판단했다.

#### 10) Tax

Cambridge가 위치한 메사추세츠 주에서는 부동산에 부과하는 세금(재산세)를 공시가(Assesed value)를 기준으로 부과한다. 공시가는 감정평가사(Professional appraiser)에 의해 정해지는데, 감정평가사는 아래와 같은 기준으로 공시가를 정한다.

1. 건물 퀄리티
2. 방 상태
3. 동네
4. 평수
5. 침실수
6. 화장실 수

Tax는 SalePrice를 결정하는 많은 요인들이 반영되어 있는 변수이기 때문에, 다른 변수들과의 다중공선성을 조심해야 한다. 그러나 건물 퀄리티, 방 상태와 같이 다른 변수들이 충분히 설명하지 못하는 부분을 반영하고 있기 때문에 사용하기로 판단했다.

#### 11) RC

RC(Rent Control Status)는 임대료 규제가 있는지에 대한 여부를 보여주는 변수이다. 값이 “1”이면 건물주가 임대를 할 때, 규제에 의해 임대료를 일정 수준 이상 올릴 수 없다. 때문에 임대료 규제가 존재한다면, SalePrice가 낮아질 것이라 추측한다.

## outlier 제거

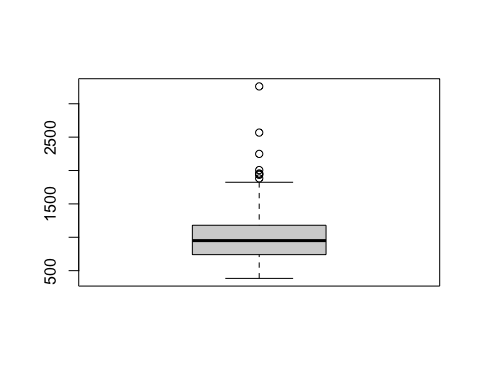
### 1. 1920년 데이터 제거

주어진 데이터는 주로 1988년 ~ 1994년에 이뤄진 매매에 대한 자료이다. 그런데, 1920년에 이뤄진 거래에 대한 관측치가 1개 존재한다. 이 자료로 인해 회귀분석 결과가 크게 왜곡될 수 있기 때문에, 제거하였다.

# YearF(FirstDate에서 연도만 추출) 변수에서 1920년 매매 자료 확인  
outliers\_YearF = data4$Days[data4$YearF=="1920"]  
outliers\_YearF

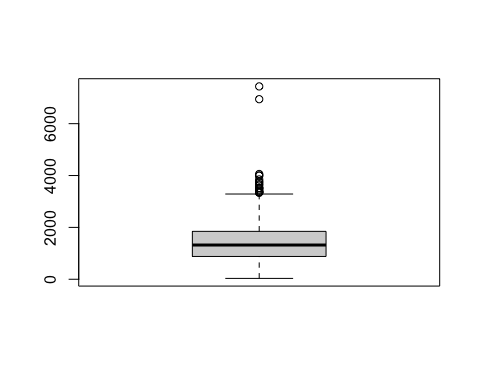
[1] 113

### 2. 너무 크거나 작은 내부면적(Interior)을 갖는 관측치 제거



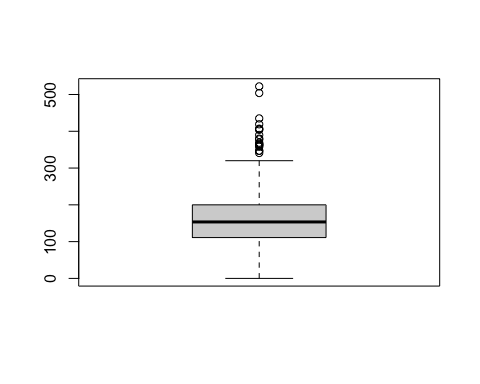
Interior 변수값에서 IQR 1.5배를 상한선, 하한선으로 하여 밖에 위치한 데이터는 극단값으로 보고 제거했다. 그 결과, 8개의 관측치를 제거했다.

### 3. 너무 크거나 작은 Tax를 갖는 관측치 제거



Tax 변수값에서 IQR 1.5배를 상한선, 하한선으로 하여 밖에 위치한 데이터는 극단값으로 보고 제거했다. 그 결과, 20개의 관측치를 제거했다.

### 4. 너무 크거나 작은 Condo를 갖는 관측치 제거



Tax 변수값에서 IQR 1.5배를 상한선, 하한선으로 하여 밖에 위치한 데이터는 극단값으로 보고 제거했다. 그 결과, 19개의 관측치를 제거했다.

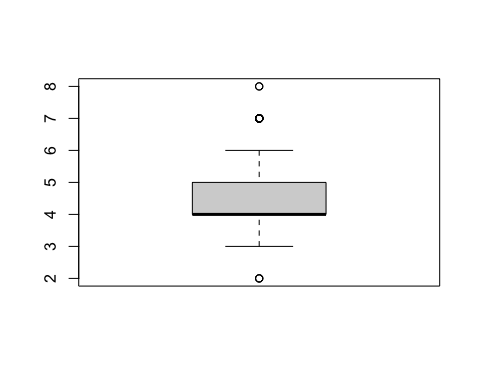
# Condo변수에서 1.5 IQR을 넘는 극단값 확인  
  
  
Lo = quantile(data7$Condo,0.25)-1.5\*IQR(data7$Condo)  
Up = quantile(data7$Condo,0.75)+1.5\*IQR(data7$Condo)  
outliers\_Condo =data7$Condo[(data7$Condo<Lo) | (data7$Condo>Up)]  
outliers\_Condo

[1] 380 366 390 347 341 378 363 358 364 347 435 419 522 406 504 369 404 361 407

# Condo변수 극단값의 인덱스 추출 및 제거(dataB)  
  
  
outliers\_Condo\_index = which((data7$Condo<Lo) | (data7$Condo>Up))  
data8 = data7[-outliers\_Condo\_index,]

### 5. 너무 많거나 적은 Rooms를 갖는 관측치 제거

# Rooms 변수의 boxplot 확인  
  
boxplot(data8$Rooms,range = 1.5)



Rooms 변수값에서 IQR 1.5배를 상한선, 하한선으로 하여 밖에 위치한 데이터는 극단값으로 보고 제거했다. 그 결과, 20개의 관측치를 제거했다.

# Rooms변수에서 1.5 IQR을 넘는 극단값 확인  
Lo = quantile(data8$Rooms,0.25)-1.5\*IQR(data8$Rooms)  
Up = quantile(data8$Rooms,0.75)+1.5\*IQR(data8$Rooms)  
outliers\_Rooms =data8$Rooms[(data8$Rooms<Lo) | (data8$Rooms>Up)]  
outliers\_Rooms

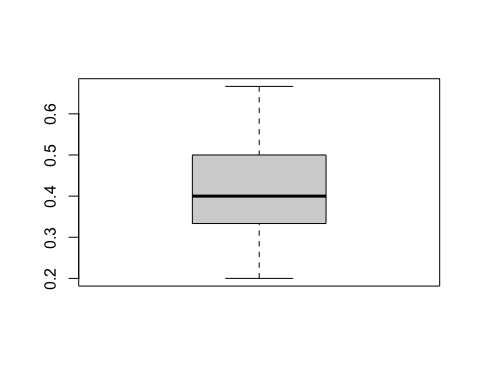
[1] 8 7 7 7 7 7 8 7 7 7 2 7 7 7 7 7 2 7 2 7

# Rooms 변수 극단값의 인덱스 추출 및 제거  
outliers\_Rooms\_index = which((data8$Rooms<Lo) | (data8$Rooms>Up))  
data9 = data8[-outliers\_Rooms\_index,]

### 6. 너무 많거나 적은 BED/Rooms(= BedPerR 변수)를 갖는 관측치 제거

방의 수 대비 너무 많거나 적은 침실을 갖고 있는 건물은 극단값에 해당한다고 보았다.

# BedPerR 변수의 boxplot 확인  
   
boxplot(data9$BedPerR, range = 1.5)

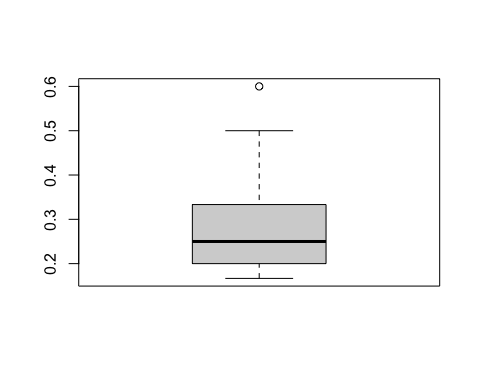


Box plot으로 확인한 결과, 극단값이 보이지 않기 때문에 관측치를 제거하지 않기로 했다.

### 7. 너무 많거나 적은 Bath/Rooms (= BathPerR변수)를 갖는 관측치 제거

방의 수 대비 너무 많거나 적은 화장실을 갖고 있는 건물은 극단값에 해당한다고 보았다.

# BathPerR 변수의 boxplot 확인  
  
boxplot(data9$BathPerR,range = 1.5)

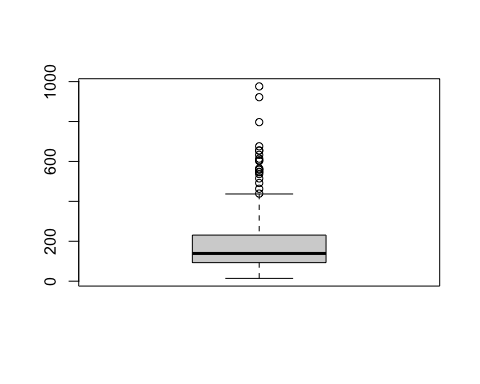


BathPerR 변수값에서 IQR 1.5배를 상한선, 하한선으로 하여 밖에 위치한 데이터는 극단값으로 보고 제거했다. 그 결과, 1개의 관측치를 제거했다.

해당 관측치는 Bath/Rooms = 0.6으로 방 5개 중 화장실이 3개, 침실이 2개인 구조를 갖고있는 건물이다.

### 8. Days가 너무 긴 혹은 짧은 관측치 제거

# Days변수의 boxplot 확인  
boxplot(dataA$Days, range = 1.5)



Days 변수값에서 IQR 1.5배를 상한선, 하한선으로 하여 밖에 위치한 데이터는 극단값으로 보고 제거했다. 그 결과, 21개의 관측치를 제거했다.

21개 제거

### 9. Area 수준 중 관측치 10개 미만인 수준 제거

Area 수준 중 관측치가 10개 미만인 수준이 있으면 SalesPrice를 종속변수로 하고 Area를 독립변수로 하는 분산분석의 왜곡된 결과를 초래할 수 있다. 때문에, 10개 미만의 관측치를 갖는 수준은 제거한다.

# Area factor의 수준별 빈도수 확인  
table(dataB$Area)

AG AH C CP E FP HS K M N NW PS RA RS SH W   
10 31 2 20 15 2 69 15 94 13 5 14 33 14 5 23

C,FP,NW,SH 수준의 관측치가 10개 미만으로 나타났기 때문에, 해당 관측치는 모두 제거하고, C,FP,NW,SH 지역은 분산분석 및 회귀분석에서 고려하지 않는다.

총 14개의 관측치가 제거되었다.

### 10. RC 0,1 갯수 확인 (한쪽이 너무 적으면 제거)

# RC factor의 수준별 빈도수 확인  
  
table(dataC$RC)

0 1   
113 238

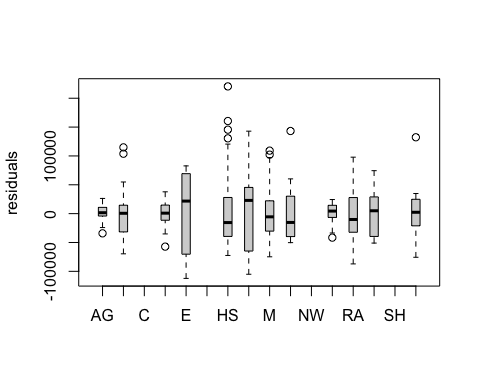
둘다 100개 이상의 데이터를 갖고 있으므로 제거하지 않았다.

### 11. SalePrice ~ Area 분산분석 모형 극단값 제거

SalePrice ~ Area 분산분석 모형(M1)에서 각 Area 수준별 잔차가 IQR 1.5배 상한선, 하한선 밖에 위치한 극단값을 갖는 관측치를 제거한다.

그 결과 13개 관측치가 제거 되었다.

# M1(dataA) outlier 확인(boxplot IQR1.5)  
bp =boxplot(ri ~ dataC$Area, xlab = "", ylab = "residuals", range= 1.5, boxwex = 0.4) # boxwex는 상자폭을 0.4배 줄이는 옵션



### outlier 제거 결과

총 338개 관측치가 남았다. 이는 주어진 456개 관측치 중 118개의 관측치를 극단값으로 판단하여 제거한 결과이며, 약 25%의 관측치를 극단값으로 판단한 것이다.

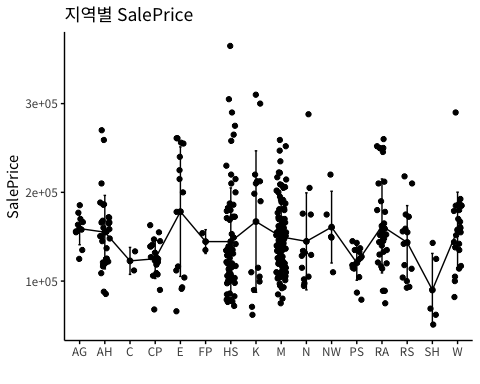
## SalePrice ~ Area 분산분석

# 분산분석 M2(SalePrice ~ Area)  
  
M2 = aov(SalePrice ~ Area, dataD)  
anova(M2)

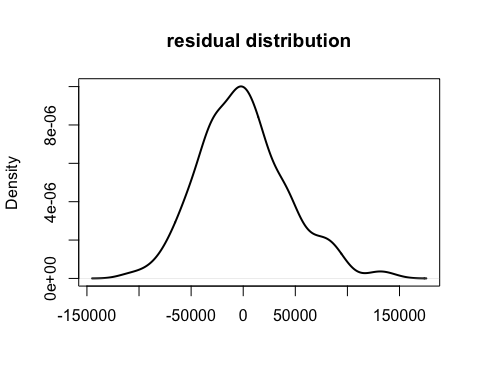
Analysis of Variance Table  
  
Response: SalePrice  
 Df Sum Sq Mean Sq F value Pr(>F)   
Area 11 5.7672e+10 5242927264 2.7916 0.001736 \*\*  
Residuals 326 6.1225e+11 1878077747   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

p값이 0.001736 \*\*으로 Area에 따른 SalePrice의 평균이 같다는 귀무가설을 기각했다.

### 등분산성 확인



그래프상에서 Area별로 SalePrice의 분포가 크게 다른 것으로 보인다. 때문에, 등분산성을 만족하지 않는 것으로 예상할 수 있다.



잔차의 분포는 상당히 정규성을 띄는 분포로 보인다. 그러므로 bartlett test를 이용하여 등분산성을 검정한다.

# BartlettTest - 등분산성 검정  
  
bartlett.test(SalePrice ~ Area, dataD)

Bartlett test of homogeneity of variances  
  
data: SalePrice by Area  
Bartlett's K-squared = 80.965, df = 11, p-value = 9.6e-13

p값이(9.6e-13)으로 ’등분산성을 만족한다’는 귀무가설을 기각하고 있기 때문에 등분산성을 갖고 있다고 볼 수 없다. 때문에 비모수검정을 진행한다.

### 비모수검정

# 비모수검정 kruskal.test  
  
kruskal.test(SalePrice ~ Area, dataD)

Kruskal-Wallis rank sum test  
  
data: SalePrice by Area  
Kruskal-Wallis chi-squared = 25.193, df = 11, p-value = 0.008542

p-value(p-value= 0.008542)가 매우 작으므로 그룹간 차이가 없다는 귀무가설을 기각하고 있다. 때문에 그룹간 차이가 존재하는 것으로 판단할 수 있다.

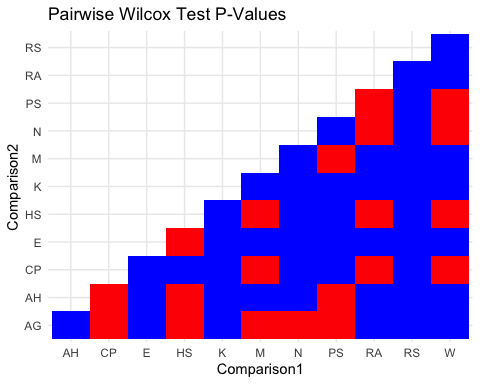
#### 사후검정, 다중비교

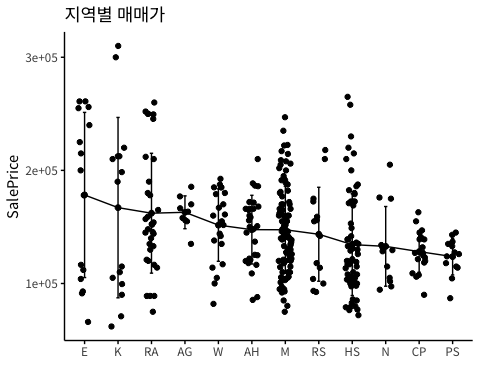
작은 차이라도 찾아내기 위해 fisher lsd 검정과 같이 개별 유의수준을 통제하는 방식으로 wilcoxon 검정을 진행했다.

# dataD Wilcoxon 검정 - 다중 비교 사후검정 (순위기반)  
  
pairwise.wilcox.test(dataD$SalePrice, dataD$Area, p.adjust.method = "none")

Pairwise comparisons using Wilcoxon rank sum test with continuity correction   
  
data: dataD$SalePrice and dataD$Area   
  
 AG AH CP E HS K M N PS   
AH 0.18045 - - - - - - - -   
CP 0.00025 0.01819 - - - - - - -   
E 0.47418 0.21108 0.17606 - - - - - -   
HS 0.01647 0.03398 0.64949 0.06443 - - - - -   
K 0.81144 0.56059 0.49869 0.48063 0.27523 - - - -   
M 0.09993 0.71107 0.05075 0.16020 0.01442 0.59960 - - -   
N 0.02782 0.20220 0.90313 0.17957 0.97199 0.39294 0.17272 - -   
PS 0.00027 0.01094 0.55192 0.19705 0.98395 0.47510 0.03002 0.91336 -   
RA 0.34971 0.57246 0.01148 0.47651 0.00642 0.92909 0.25434 0.06273 0.00582  
RS 0.17540 0.59511 0.40203 0.20552 0.42969 0.54109 0.68477 0.68067 0.30811  
W 0.54206 0.63439 0.01205 0.25871 0.02499 0.51579 0.44654 0.07731 0.00734  
 RA RS   
AH - -   
CP - -   
E - -   
HS - -   
K - -   
M - -   
N - -   
PS - -   
RA - -   
RS 0.38293 -   
W 0.88387 0.39862  
  
P value adjustment method: none

위 wilcoxon 검정 결과를 기반으로 두 변수간 p값이 0.1 이하인 경우 붉은 타일이 나타나는 타일플롯을 그렸다.





### Area 분산분석 결과

위 분산분석결과를 통해 SalePrice 평균 차이가 나지 않는 지역을 묶고, SalePrice 평균 차이가 나는 지역을 나눈 새로운 Zone 변수를 만들었다.

1Zone = “E”,“K”,“RA”,“AG”  
2Zone = “W”,“AH”, “M”,“RS”  
3Zone = “HS”,“N”,“CP”,“PS”

## 다중회귀분석

### 다중 회귀 모형 - 단계별 선택법

모든 변수(Interior,Zone,HousePriceIndex,Days,Rooms,Bed,Bath,Condo,Tax,RC)를 독립변수로 하고 모든 상호작용을 고려한 Full model(SalePrice ~ Interior\*Zone\*HousePrice\*Index\*Days\*Rooms\*Bed\*Bath\*Condo\*Tax\*RC)은 다중공선성 문제가 발생하는 것으로 판단하여, 상호작용을 고려하지 않는 model에서 시작하여 단계별 선택법을 통해 다중회귀 모형을 결정하기로 했다.

# 상호작용을 고려하지 않은 다중회귀 모형  
  
LM2 = lm(SalePrice ~ Interior + Zone + HousePriceIndex + Days + Rooms + Bed + Bath + Condo + Tax + RC, dataG)

# 상호작용을 고려하지 않은 다중회귀 모형의 MSE확인  
  
F2 = anova(LM2)  
F2[11,3] # MSE : 715,208,196

[1] 715208196

단계별 선택법의 결과는 아래와 같다.

# 다중회귀 모형의 model selection - 단계별 선택법  
  
step = stepAIC(LM2, scope = list(upper = ~ Interior\*Zone\*HousePriceIndex\*Days \* Rooms \* Bed \* Bath \* Condo \* Tax \* RC, lower = ~1),direction="both",trace = FALSE )

# 다중회귀 모형의 model selection 단계별 선택법 결과  
  
step$anova

Stepwise Model Path   
Analysis of Deviance Table  
  
Initial Model:  
SalePrice ~ Interior + Zone + HousePriceIndex + Days + Rooms +   
 Bed + Bath + Condo + Tax + RC  
  
Final Model:  
SalePrice ~ Interior + Zone + HousePriceIndex + Days + Rooms +   
 Bed + Bath + Condo + Tax + RC + Zone:Tax + Rooms:Bath + Interior:Zone +   
 Zone:Bath + Condo:RC + Days:Bath + Bed:RC + Interior:Rooms +   
 Interior:Tax + Rooms:Bed + Interior:Zone:Tax  
  
 Step Df Deviance Resid. Df Resid. Dev AIC  
1 326 233157871982 6902.954  
2 + Zone:Tax 2 12260170046 324 220897701935 6888.697  
3 + Rooms:Bath 1 8782577705 323 212115124230 6876.984  
4 + Interior:Zone 2 6206866278 321 205908257953 6870.946  
5 + Zone:Bath 2 5495005933 319 200413252020 6865.803  
6 + Condo:RC 1 5484529011 318 194928723008 6858.425  
7 + Days:Bath 1 3129978980 317 191798744028 6854.953  
8 + Bed:RC 1 2218764051 316 189579979978 6853.020  
9 + Interior:Condo 1 2498672563 315 187081307415 6850.536  
10 + Interior:Rooms 1 1754376756 314 185326930659 6849.351  
11 + Interior:Tax 1 2319471015 313 183007459644 6847.094  
12 + Interior:Zone:Tax 2 3517615143 311 179489844501 6844.534  
13 + Rooms:Bed 1 1987406171 310 177502438330 6842.771  
14 - Interior:Condo 1 218408414 311 177720846743 6841.187

단계별 선택법에 의한 다중회귀 모형의 MSE는 아래와 같다.

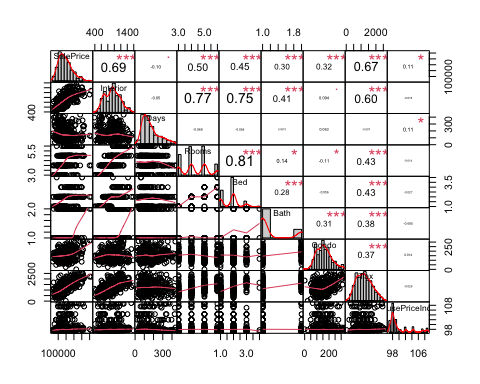
[1] 571449668

상호작용항을 고려하지 않은 모형의 MSE(715,208,196)에 비해 크게 감소한 것을 확인할 수 있다.

## 다중공선성 및 영향점 확인

### 다중회귀모형에 사용한 변수간 상관관계 확인(수치형 변수만)

chart.Correlation(dataG\_N)



Interior - Rooms  
Interior - Bed  
Interior - Bath  
Interior - Tax  
Rooms - Bed

위 변수간에 상관관계가 특히 높다.

다중회귀 모형의 회귀계수와 분산분석을 확인해보면, Bed와 Bath 변수의 회귀계수 부호, 통계적 유의성에 문제가 보인다. 그러나, Bed 혹은 Bath와 Bed 혹은 Bath가 속한 상호작용을 제거할 경우, MSE가 지속적으로 증가하여 예측력이 떨어지는 것을 확인하였다. 때문에 다중공선성 문제보단 영향점을 제거하는데 집중했다.

### 잔차 분석

gvlma(LM3)

Call:  
lm(formula = SalePrice ~ Interior + Zone + HousePriceIndex +   
 Days + Rooms + Bed + Bath + Condo + Tax + RC + Zone:Tax +   
 Rooms:Bath + Interior:Zone + Zone:Bath + Condo:RC + Days:Bath +   
 Bed:RC + Interior:Rooms + Interior:Tax + Rooms:Bed + Interior:Zone:Tax,   
 data = dataG)  
  
Coefficients:  
 (Intercept) Interior Zone2Zone   
 -3.072e+04 1.750e+02 -8.728e+04   
 Zone3Zone HousePriceIndex Days   
 -4.097e+04 1.920e+03 -1.367e+02   
 Rooms Bed Bath   
 1.022e+03 -4.829e+04 -1.414e+05   
 Condo Tax RC1   
 1.357e+02 -4.901e+01 -7.290e+03   
 Zone2Zone:Tax Zone3Zone:Tax Rooms:Bath   
 4.774e+01 5.768e+01 2.194e+04   
 Interior:Zone2Zone Interior:Zone3Zone Zone2Zone:Bath   
 7.745e+01 -4.145e+00 3.682e+04   
 Zone3Zone:Bath Condo:RC1 Days:Bath   
 2.826e+04 -1.017e+02 8.035e+01   
 Bed:RC1 Interior:Rooms Interior:Tax   
 1.464e+04 -3.597e+01 9.150e-02   
 Rooms:Bed Interior:Zone2Zone:Tax Interior:Zone3Zone:Tax   
 6.370e+03 -7.827e-02 -6.139e-02   
  
  
ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS  
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:  
Level of Significance = 0.05   
  
Call:  
 gvlma(x = LM3)   
  
 Value p-value Decision  
Global Stat 367.8211 0.00000 Assumptions NOT satisfied!  
Skewness 7.5930 0.00586 Assumptions NOT satisfied!  
Kurtosis 356.2798 0.00000 Assumptions NOT satisfied!  
Link Function 0.2561 0.61283 Assumptions acceptable.  
Heteroscedasticity 3.6923 0.05466 Assumptions acceptable.

gvlma 결과 영향점 있을 수 있다고 판단했다.

### 영향점 확인

외적 스튜던트화된 잔차가 1.5이상인 경우를 영향점으로 봤다. 총 28개의 관측치가 해당되었으며 모두 제거했다.

AddressNo AddressName Street Unit Area AreaCode Price1 Price2 SalePrice  
7 43 Antrim St 3 M 9 194000 194000 189000  
19 10 Avon St 2 AH 2 197500 169500 166000  
27 41 Bowdoin St 43 RA 13 203800 169900 152000  
28 246 Brattle St 41 W 16 179000 165000 144000  
71 20 Chestnut St 403 CP 4 145000 138000 132000  
79 5 Chilton St 3 M 9 249000 249000 235000  
100 River Court <NA> 606 K 8 339000 339000 310000  
101 River Court <NA> 216 K 8 225000 225000 212500  
102 River Court <NA> 701 K 8 315000 315000 300000  
123 170 Gore St 314 E 5 99000 99000 91250  
154 280 Harvard St LD M 9 145000 105000 97000  
158 352 Harvard St 1D HS 7 229000 229000 220000  
174 12 Inman St 51 M 9 149000 119000 106000  
175 12 Inman St 21 M 9 139900 113000 105000  
224 1580 Mass Ave 4D HS 7 179900 179900 172500  
225 1600 Mass Ave 808 HS 7 249900 229900 210000  
232 1580 Mass Ave 2C HS 7 209000 209000 200000  
239 992 Memorial Dr 505 HS 7 395000 298000 265000  
243 205 Mt Auburn 3B W 16 197000 192000 185000  
270 12 Remington St 105 HS 7 129000 129000 121500  
284 32 Shepard St 43 RA 13 159000 159000 159000  
291 34 Soden St 3 RS 14 163500 101000 92500  
299 218 Thorndike St 104B K 8 225000 119000 105000  
302 218 Thorndike St 103B K 8 69900 69900 62000  
303 217 Thorndike St 205 K 8 119000 119000 115000  
304 217 Thorndike St 106A K 8 115000 100000 71000  
314 275 Walden St 2R W 16 125000 125000 114000  
317 12 Ware St 12B HS 7 119900 104900 85000  
 YearF MonthF DateF CloseDate Days Interior Bed Bath Rooms Condo Tax RC  
7 1993 05 18 1993-08-06 80 1058 3 1 5 163 1682 0  
19 1993 07 31 1993-12-15 137 730 1 1 4 170 1011 1  
27 1993 07 17 1993-09-30 75 900 2 1 5 182 2243 1  
28 1990 09 07 1991-02-01 147 1170 2 1 6 277 1929 1  
71 1993 04 24 1993-09-01 130 931 2 2 5 204 1825 0  
79 1992 04 18 1992-09-04 139 1266 3 2 6 193 1268 0  
100 1993 02 03 1993-06-04 121 1277 2 2 4 320 2197 0  
101 1993 03 27 1993-07-16 111 970 2 2 4 249 1585 0  
102 1993 06 19 1993-10-29 132 1119 2 2 4 319 2209 0  
123 1992 02 01 1992-04-13 72 640 1 1 3 198 679 0  
154 1992 08 18 1993-03-04 198 874 2 2 4 255 1746 0  
158 1993 05 28 1993-07-30 63 887 2 1 5 173 1407 1  
174 1990 10 12 1991-03-22 161 929 2 1 4 143 859 1  
175 1992 12 26 1993-03-19 83 929 2 1 4 140 726 1  
224 1992 05 16 1992-07-29 74 972 2 2 4 156 1208 0  
225 1991 10 26 1992-08-07 286 1060 2 2 4 186 1865 0  
232 1993 09 10 1993-11-10 61 1068 2 2 4 163 1284 0  
239 1990 03 11 1990-12-10 274 1050 2 1 5 284 2086 1  
243 1993 04 24 1993-12-28 248 915 2 1 5 247 1551 1  
270 1993 11 19 1994-02-18 91 915 1 1 5 232 1802 1  
284 1990 04 22 1990-08-21 121 642 2 1 4 107 724 1  
291 1991 11 07 1992-03-16 130 801 2 1 5 88 789 1  
299 1991 02 05 1991-05-21 105 1253 2 2 4 274 2186 0  
302 1993 02 26 1993-07-06 130 574 1 1 3 129 1085 0  
303 1993 05 15 1993-08-06 83 1158 2 2 5 253 2323 0  
304 1993 06 10 1993-09-30 112 841 2 2 4 184 1689 0  
314 1993 10 09 1994-02-22 136 1000 2 1 4 100 1190 1  
317 1990 08 03 1990-12-13 132 895 2 1 4 170 1195 1  
 PrPerI BedPerR BathPerR rstud  
7 178.63894 0.6000000 0.2000000 1.860105  
19 227.39726 0.2500000 0.2500000 1.564319  
27 168.88889 0.4000000 0.2000000 -1.844298  
28 123.07692 0.3333333 0.1666667 -1.765715  
71 141.78303 0.4000000 0.4000000 -1.740420  
79 185.62401 0.5000000 0.3333333 1.928749  
100 242.75646 0.5000000 0.5000000 3.045212  
101 219.07216 0.5000000 0.5000000 3.342003  
102 268.09651 0.5000000 0.5000000 4.132324  
123 142.57812 0.3333333 0.3333333 -1.590053  
154 110.98398 0.5000000 0.5000000 -2.185220  
158 248.02706 0.4000000 0.2000000 2.961382  
174 114.10118 0.5000000 0.2500000 -1.625241  
175 113.02476 0.5000000 0.2500000 -1.599062  
224 177.46914 0.5000000 0.5000000 2.435121  
225 198.11321 0.5000000 0.5000000 2.182729  
232 187.26592 0.5000000 0.5000000 3.069050  
239 252.38095 0.4000000 0.2000000 3.211199  
243 202.18579 0.4000000 0.2000000 1.615376  
270 132.78689 0.2000000 0.2000000 -2.219029  
284 247.66355 0.5000000 0.2500000 1.694350  
291 115.48065 0.4000000 0.2000000 -1.599301  
299 83.79888 0.5000000 0.5000000 -5.969612  
302 108.01394 0.3333333 0.3333333 -1.988622  
303 99.30915 0.4000000 0.4000000 -5.246671  
304 84.42331 0.5000000 0.5000000 -1.743246  
314 114.00000 0.5000000 0.2500000 -1.758449  
317 94.97207 0.5000000 0.2500000 -2.547085

### 영향점 제거 후 다중회귀모형 성능 확인

영향점 제거 후 동일한 모형의 MSE를 확인할 결과 259,626,691로, 영향점 제거 전 MSE의 45% 수준으로 감소했다.

# 영향점 제거 후 MSE 확인   
  
LM3 =lm(SalePrice ~ Interior + Zone + HousePriceIndex + Days + Rooms +   
 Bed + Bath + Condo + Tax + RC + Zone:Tax + Rooms:Bath + Interior:Zone +   
 Zone:Bath + Condo:RC + Days:Bath + Bed:RC + Interior:Rooms +   
 Interior:Tax + Rooms:Bed + Interior:Zone:Tax,dataH)  
F3 = anova(LM3)  
F3[22,3] # MSE = 259,626,691 (이전보다 반으로 줄어듦 571,449,668)

[1] 259626691

영향점 제거 후 각 회귀계수의 통계적 유의성을 검정하는 F값이 증가하여 p값이 크게 감소하였다. 이는 앞서 다중회귀모형의 회귀계수 부호와 통계적 유의성 문제의 원인이 다중공선성이 아닌 영향점이라는 것을 알 수 있다.

# 영향점 제거 후 다중회귀모형 분산분석 확인  
  
F3

Analysis of Variance Table  
  
Response: SalePrice  
 Df Sum Sq Mean Sq F value Pr(>F)   
Interior 1 2.8640e+11 2.8640e+11 1103.1051 < 2.2e-16 \*\*\*  
Zone 2 4.1369e+10 2.0684e+10 79.6694 < 2.2e-16 \*\*\*  
HousePriceIndex 1 1.1726e+10 1.1726e+10 45.1667 9.946e-11 \*\*\*  
Days 1 4.3986e+09 4.3986e+09 16.9421 5.060e-05 \*\*\*  
Rooms 1 1.8155e+09 1.8155e+09 6.9926 0.0086415 \*\*   
Bed 1 3.9586e+09 3.9586e+09 15.2474 0.0001180 \*\*\*  
Bath 1 1.3430e+07 1.3430e+07 0.0517 0.8202441   
Condo 1 1.9153e+10 1.9153e+10 73.7718 5.981e-16 \*\*\*  
Tax 1 3.6231e+10 3.6231e+10 139.5489 < 2.2e-16 \*\*\*  
RC 1 9.7797e+08 9.7797e+08 3.7668 0.0532716 .   
Zone:Tax 2 1.7380e+10 8.6900e+09 33.4713 8.958e-14 \*\*\*  
Rooms:Bath 1 9.1333e+09 9.1333e+09 35.1787 8.754e-09 \*\*\*  
Interior:Zone 2 6.7598e+09 3.3799e+09 13.0183 3.903e-06 \*\*\*  
Zone:Bath 2 3.7627e+09 1.8814e+09 7.2465 0.0008528 \*\*\*  
Condo:RC 1 3.8838e+09 3.8838e+09 14.9593 0.0001364 \*\*\*  
Days:Bath 1 4.0836e+09 4.0836e+09 15.7289 9.267e-05 \*\*\*  
Bed:RC 1 4.3710e+09 4.3710e+09 16.8357 5.335e-05 \*\*\*  
Interior:Rooms 1 1.9870e+09 1.9870e+09 7.6534 0.0060395 \*\*   
Interior:Tax 1 4.9622e+09 4.9622e+09 19.1127 1.733e-05 \*\*\*  
Rooms:Bed 1 4.2460e+08 4.2460e+08 1.6354 0.2020025   
Interior:Zone:Tax 2 7.0722e+09 3.5361e+09 13.6200 2.252e-06 \*\*\*  
Residuals 283 7.3474e+10 2.5963e+08   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#### 회귀계수 확인

회귀계수 확인시 더미변수를 모두 확인하려면 항이 너무 많이지기 때문에 Reyeme Affiar가 고려하는 건물의 Area(M)에 해당하는 2Zone=1, RC=1 인 경우만 고려하여 확인했다.

전체 다중회귀 모형은 다음과 같을 것이다.

##### 1) Interior의 회귀계수

는 2zone에 속한 모든 집에서 0보다 클 것이다. Rooms가 7이상일 경우 이 되는데, 데이터상에서는 Rooms가 7이상인 관측치가 존재하지 않는다.

때문에 Interior가 클수록 SalePrice가 증가한다는 예측과 부합하는 회귀계수이다.

또한 내부면적이 같은 경우에도 방의 수가 너무 많다면, 선호도가 떨어져 SalePrice가 낮아지는 경우가 생길 수 있을것이라는 예측과 관련하여, Rooms와의 상호작용에 의한 항도 예측과 부합한다고 판단했다.

##### 2) Rooms의 회귀계수

침실의 수는 모든집에서 1개 이상이므로, 총합을 계산해 보면 이다. 합당한 회귀계수 부호를 갖고 있다고 판단한다.

##### 3) Bed의 회귀계수

침실의 수가 너무 많으면 SalePrice가 감소할 것이라는 예측에 부합하는 회귀계수 부호를 갖고 있다고 판단했다.

##### 4) Bath의 회귀계수

화장실의 수가 너무 많으면 SalePrice가 감소할 것이라는 예측에 부합하는 회귀계수 부호를 갖고 있다고 판단했다.

##### 5) Condo의 회귀계수

Condo가 커지면, SalePrice도 클 것이라는 예상과 부합한다.

##### 6) Tax의 회귀계수

Interior가 보통 1000을 상회하기 때문에 총합을 계산하면, 이다. Tax가 커지면, SalePrice도 클 것이라는 예상과 부합한다.

##### 7) Days의 회귀계수

Days의 회귀 계수는 해석할 수 없다. 다중공선성 문제나 영향점 제거가 안되었을 가능성이 있다.

##### 8) HousePriceIndex의 회귀계수

HousePriceIndex가 클수록, SalePrice는 클 것이라는 예상과 부합하는 회귀계수 이다..

##### 9) (Zone이 2Zone일 경우)의 회귀계수

Reference Group인 1Zone에 2Zone은 가격이 낮은 지역이므로, 2Zone일 경우, SalePrice가 떨어질 것이라는 예상과 부합한다.

##### 10) (RC= 1 일 경우)의 회귀계수

임대료 규제가 있다면(RC=1) 없는 경우보다 SalePrice가 떨어질 것이다.

### 영향점 제거 후 다중회귀모형을 통한 Reyeme Affiar의 건물 가격 예측

다중회귀모형을 이용한 매매가격에 대한 예측 결과는 아래와 같다.(단, 매매가 5월 4일에 이뤄졌다고 가정)

# 매매가격에 대한 예측  
predict(LM3, newdata = data.frame(Interior = 1040, Zone = "2Zone",HousePriceIndex = 99.38, Days = 24, Rooms = 5, Bed = 2, Bath = 1, Condo = 175 , Tax = 1121 , RC = "1"), interval = "prediction")

fit lwr upr  
1 168540.1 136353.8 200726.4

이 건물의 매매가격은 $168,540 로 추정할 수 있으며, $200,726 이상으로 매매가격이 나타나는 것은 이례적으로 비싼 가격이라 할 수 있다. $136,353이하로 매매가 이루어지기는 어려울 것이다.