<변수 정리>

* Data_description.txt 파일과 실제 데이터가 다른 것이 일부 존재

Variable: 80
Observation: 1460

Variable Name	Explanation	Type	#Categories	Min-Max	#Missing(%)
MSSubClass	건물 유형	Categorical	16	-	-
MSZoning	용도지역 분류	Categorical	5	-	-
LotFrontage	건물과 인접 거리의 직선 길이(feet)	Numerical-continuous	-	21-313	259(17.7%)
LotArea	건물 부지 면적	Numerical-continuous	-	1300-215245	-
Street	건물과 인접한 도로 유형	Categorical	2	-	-
Alley	건물과 인접한 골목 유형	Categorical	2	-	1369(93.8%)
LotShape	건물의 일반적인 모양	Categorical	4	-	-
LandContour	토지의 평탄정도	Categorical	4	-	-
Utilities	사용가능한 공익사업(수도전기가스)	Categorical	4	-	-
LotConfig	건물 부지 구조	Categorical	5	-	-
LandSlope	토지 경사	Categorical	3	-	-
Neighborhood	'Ames city' 내의 위치	Categorical	25	-	-
Condition1	교통 접근성1	Categorical	9	-	-
Condition2	교통 접근성2(2개 이상 case)	Categorical	9	-	-
BldgType	주택 형태	Categorical	5	-	-
HouseStyle	주택 스타일	Categorical	8	-	-
OverallQual	주택의 전반적인 자재 및 마감 등급	Categorical-ordinal	10	-	-
OverallCond	주택의 전반적인 컨디션 등급	Categorical-ordinal	10	-	-
YearBuilt	시공일자	Numerical-continuous	-	1872-2010	-
YearRemodAdd	리모델링 일자(없으면 시공일과 동일)	Numerical-continuous	-	1950-2010	-
RoofStyle	지붕 형태	Categorical	6	-	-
RoofMatl	지붕 소재	Categorical	8	-	-

	variable	n_miss	pct_miss
1	PoolQC	1453	99.52054795
2	MiscFeature	1406	96.30136986
3	Alley	1369	93.76712329
4	Fence	1179	80.75342466
5	FireplaceQu	690	47.26027397
6	LotFrontage	259	17.73972603
7	GarageType	81	5.54794521
8	GarageYrBlt	81	5.54794521
9	GarageFinish	81	5.54794521
10	GarageQual	81	5.54794521
11	GarageCond	81	5.54794521
12	BsmtExposure	38	2.60273973
13	BsmtFinType2	38	2.60273973
14	BsmtQual	37	2.53424658
15	BsmtCond	37	2.53424658
16	BsmtFinType1	37	2.53424658
17	MasVnrType	8	0.54794521
18	MasVnrArea	8	0.54794521
19	Electrical	1	0.06849315

* 결측치 수 및 비율

상관관계가 높아 보이지만 그렇지 않음

```
> cor(data$OverallCond, data$OverallQual, method="spearman")
[1] -0.1775207
```

Variable Name	Explanation	Type	#Categories	Min-Max	#Missing(%)
Exterior1st	외부벽 소재1	Categorical	17	-	-
Exterior2nd	외부벽 소재2(2개 이상 case)	Categorical	17	-	-
MasVnrType	석조 베니어벽 소재	Categorical	5	-	8(0.5%)
MasVnrArea	석조 베니어벽 면적	Numerical-continuous	-	0-1600	8(0.5%)
ExterQual	건물 외부 재료 품질	Categorical-ordinal	5	-	-
ExterCond	건물 외부 재료 현재 상태	Categorical-ordinal	5	-	-
Foundation	건물 기초공사 자재	Categorical	6	-	-
BsmtQual	지하실 높이	Categorical-ordinal	5	-	37(2.5%)
BsmtCond	지하실 상태	Categorical-ordinal	5	-	37(2.5%)
BsmtExposure	Garden level wall 노출정도	Categorical-ordinal	4	-	38(2.6%)
BsmtFinType1	공사한 지하실 등급1	Categorical-ordinal	6	-	37(2.5%)
BsmtFinSF1	공사한 지하실 면적1	Numerical-continuous	-	0-5644	-
BsmtFinType2	공사한 지하실 등급2(2개 이상 case)	Categorical-ordinal	6	-	38(2.6%)
BsmtFinSF2	공사한 지하실 면적2	Numerical-continuous	-	0-1474	-
BsmtUnfSF	공사하지 않은 지하실 면적	Numerical-continuous	-	0-2336	-
TotalBsmtSF	전체 지하실 면적	Numerical-continuous	-	0-6110	-
Heating	난방 형태	Categorical	6	-	-
HeatingQC	난방 품질	Categorical-ordinal	5	-	-
CentralAir	에어컨 중앙제어	Categorical	2	-	-
Electrical	전기시스템	Categorical	5	-	1(0.1%)
1stFlrSF	1층 면적	Numerical-continuous	-	334-4692	-
2ndFlrSF	2층 면적	Numerical-continuous	-	0-2065	-
LowQualFinSF	저품질 면적(모든 층)	Numerical-continuous	-	0-572	-
GrLivArea	(지상) 거주 면적	Numerical-continuous	-	334-5642	-
BsmtFullBath	(지하) full bathroom 수	Numerical-discrete	-	0-3	-
BsmtHalfBath	(지하) half bathroom 수	Numerical-discrete	-	0-2	-
FullBath	(지상) full bathroom 수	Numerical-discrete	-	0-3	-
HalfBath	(지상) half bathroom 수	Numerical-discrete	-	0-2	-
BedroomAbvGr	(지상) 침실(지하 미포함) 수	Numerical-discrete	-	0-8	-

대부분 석조 베니어벽을 가지고 있음

등급1에서는 결측치가 37개,
등급2에서는 결측치가 38개여서 확인해보니,
관측치 하나가 BsmtFinSF2는479이지만 BsmtFinType2가 NA인 경우였음.
BsmtFinType2가 누락된 것으로 확인

BsmtFinSF1	BsmtFinType2	BsmtFinSF2	BsmtUnfSF	TotalB
0	NA	0	0	
0	NA	0	0	
1124	NA	479	1603	
0	NA	0	0	

또한
BsmtFinSF1+BsmtFinSF2+BsmtUnfSF
=TotalBsmtSF 관계를 가지고 있음

```
> sum(tmp$BsmtArea!=tmp$TotalBsmtSF)
[1] 0
```

1층 면적+2층 면적 != 거주면적 (26개)

```
> sum(!(tmp$SF==tmp$GrLivArea))
[1] 26
```

Variable Name	Explanation	Type	#Categories	Min-Max	#Missing(%)
KitchenAbvGr	(지상) 부엌 수	Numerical-discrete	-	0-3	-
KitchenQual	부엌 품질	Categorical-ordinal	5	-	-
TotRmsAbvGrd	(지상) 전체 방 수(bathroom 미포함)	Numerical-discrete	-	2-14	-
Functional	Home functionality	Categorical	8	-	-
Fireplaces	벽난로 수	Numerical-discrete	-	0-3	-
FireplaceQu	벽난로 품질	Categorical-ordinal	5	-	690(47.3%)
GarageType	차고 위치	Categorical	6	-	81(5.5%)
GarageYrBlt	차고 건축일	Numerical-continuous	-	1900-2010	81(5.5%)
GarageFinish	차고 완성도	Categorical-ordinal	3	-	81(5.5%)
GarageCars	차고 수용 가능 차량 수	Numerical-discrete	-	0-4	-
GarageArea	차고 면적	Numerical-continuous	-	0-1418	-
GarageQual	차고 품질	Categorical-ordinal	5	-	81(5.5%)
GarageCond	차고 상태	Categorical-ordinal	5	-	81(5.5%)
PavedDrive	진입로 포장상태	Categorical	3	-	-
WoodDeckSF	나무 데크 면적	Numerical-continuous	-	0-857	-
OpenPorchSF	개방된 현관 앞 공간 면적	Numerical-continuous	-	0-547	-
EnclosedPorch	폐쇄된 현관 앞 공간 면적	Numerical-continuous	-	0-552	-
3SsnPorch	3계절 현관 넓이	Numerical-continuous	-	0-508	-
ScreenPorch	유리로 둘러싸인 현관 면적	Numerical-continuous	-	0-480	-
PoolArea	수영장 면적	Numerical-continuous	-	0-738	-
PoolQC	수영장 품질	Categorical-ordinal	4	-	1453(99.5%)
Fence	울타리 품질	Categorical-ordinal	4	-	1179(80.8%)
MiscFeature	기타 특징	Categorical	5	-	1406(96.3%)
MiscVal	기타 특징 가치(달러)	Numerical-continuous	-	0-15500	-
MoSold	매각일자(월)	Numerical-continuous	-	1-12	-
YrSold	매각일자(년)	Numerical-continuous	-	2006-2010	-
SaleType	판매 형태	Categorical	9	-	-
SaleCondition	판매 조건	Categorical	6	-	-
SalePrice	판매 가격	Numerical-continuous	-	34900-755000	-

결측치는 차고가 없는 집

4개 모두 0인 관측치는 458개

```
> sum(tmp$flg==0)
[1] 458
```

수영장을 가진 집이 거의 없음

기타 특징을 가진 집은 별로 없음

이 데이터는 건물의 위치(환경적, 기능적 등)부터 건물 내/외부 정보가 상세하게 있다.

변수의 수는 많지만 모든 변수가 서로 다른 정보를 가진 것이 아니고, 일부 변수들은 동일한 공간에 대한 정보를 가지고 있다. 예를 들어, Garage_ 변수들은 차고에 대한 정보를 공통적으로 담고 있다.

즉, 변수들의 특성은 크게 2가지로 나눌 수 있는데, 건물이 존재한다면 필수적으로 있어야 할 변수(건물 유형, 시공일자 등)와 필수는 아니지만 건물의 옵션으로 생각할 수 있는 변수(지하실, 차고 등)이다.

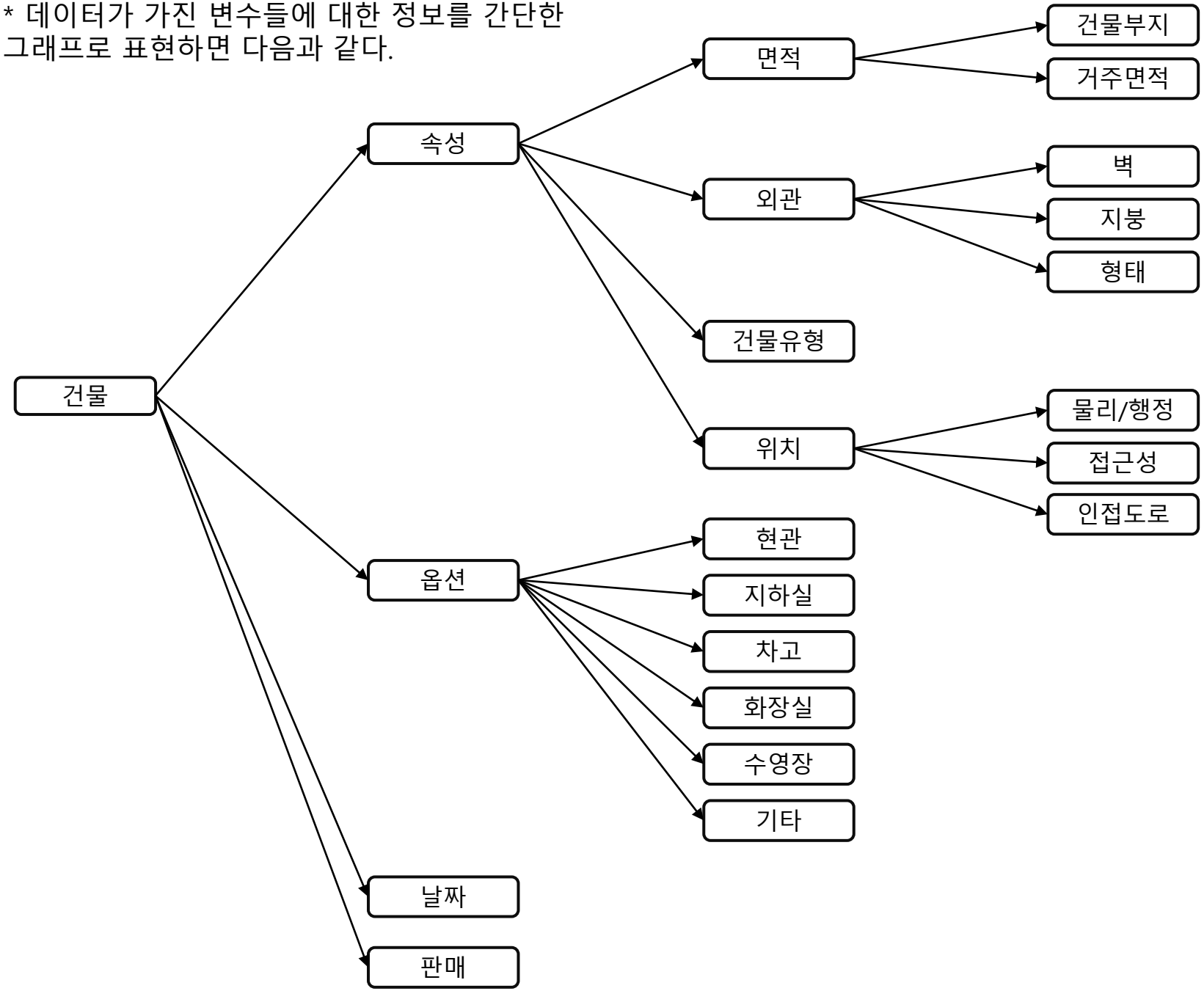
일부 범주형 변수들은 범주가 상당히 많다. 축소할 필요가 있어 보인다.

종속변수로 가져갈 만한 변수는 'SalePrice', 'LotArea', 'MSZoning', 'BedroomAbvGr', 'GrLivArea', 'TotRmsAbvGrd'

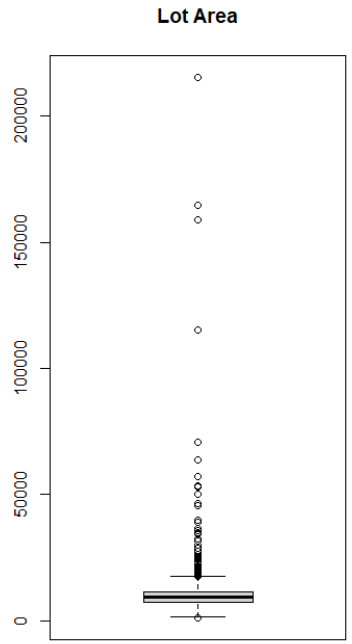
'SalePrice': 판매가격 예측
'LotArea': 건물 부지 면적 예측
'BedroomAbvGr', 'TotRmsAbvGrd': 침실 및 방 개수 예측
'GrLivArea': 거주 면적 예측(1층, 2층 면적 변수 삭제하고)
'MSZoning': 용도지역 분류

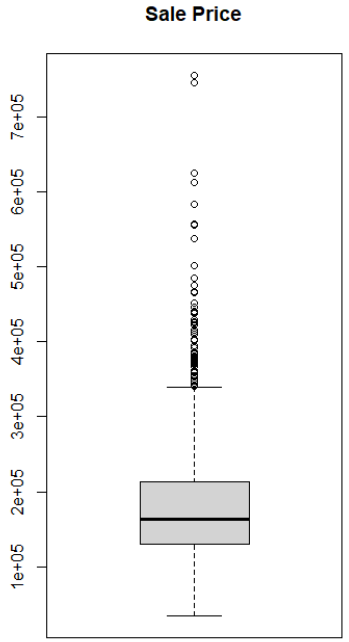
또는 리모델링한 건물과 그렇지 않은 건물의 가격차이가 존재하는지?

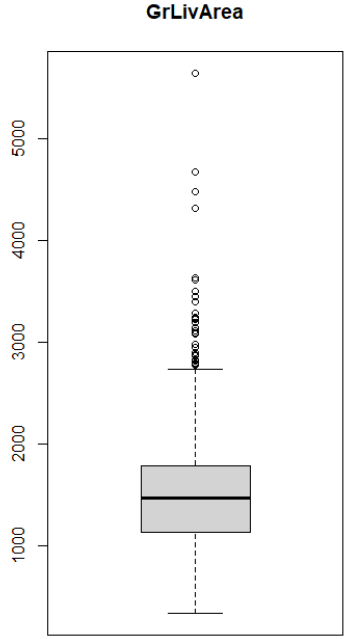
* 데이터가 가진 변수들에 대한 정보를 간단한 그래프로 표현하면 다음과 같다.

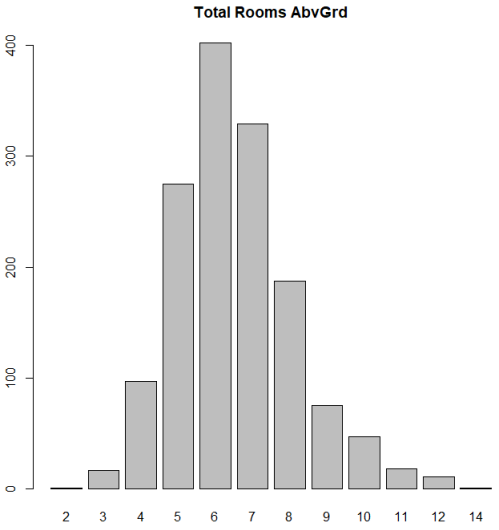


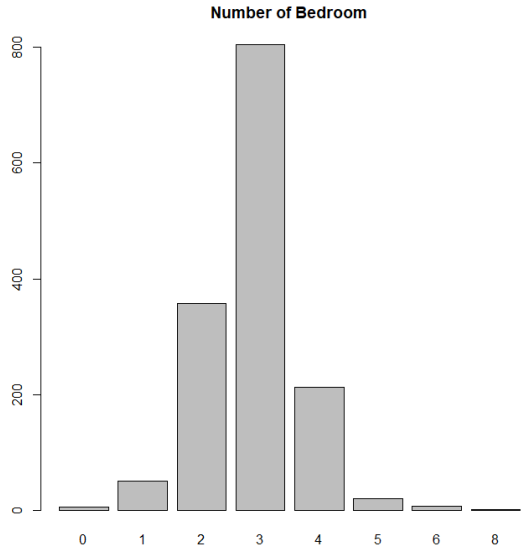
* 분석 해 볼만하다고 언급했던 변수들의 boxplot 및 빈도 그래프

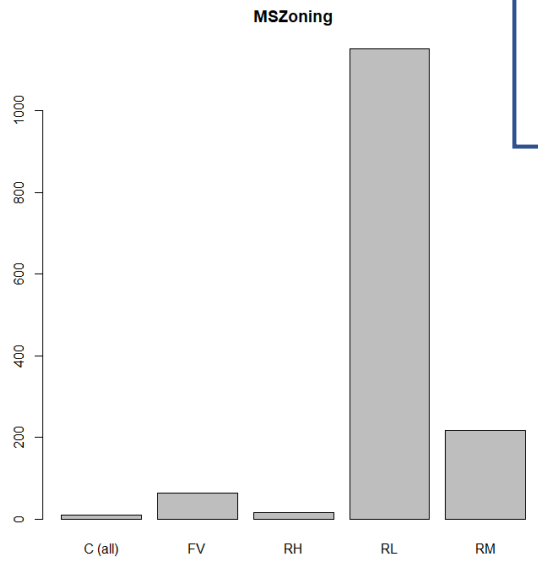




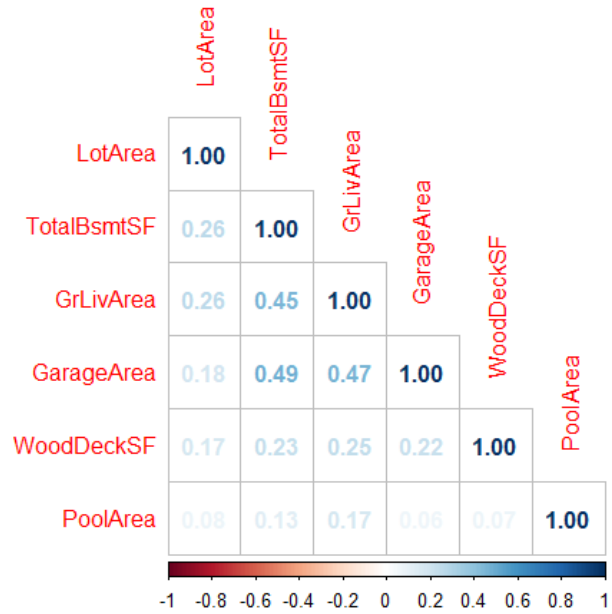








* imbalance함

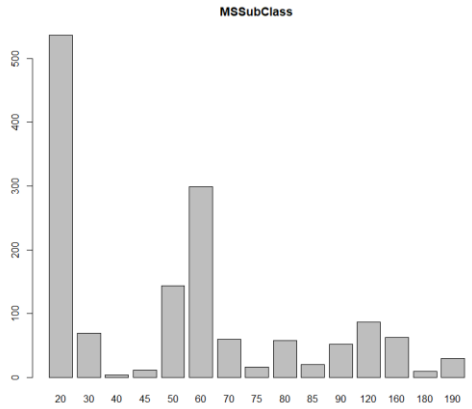


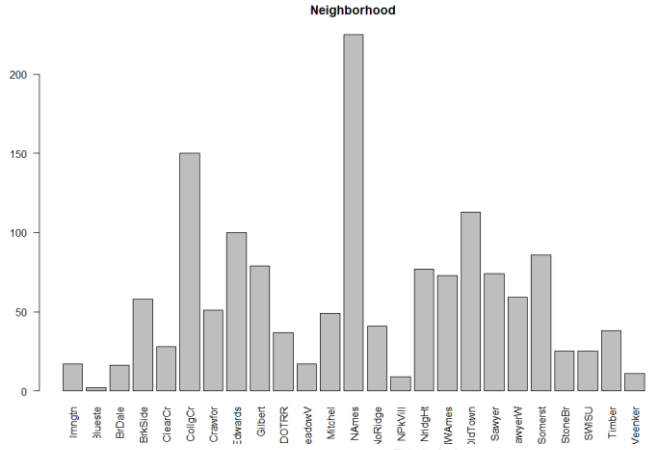
- 면적 변수들 간의 상관계수

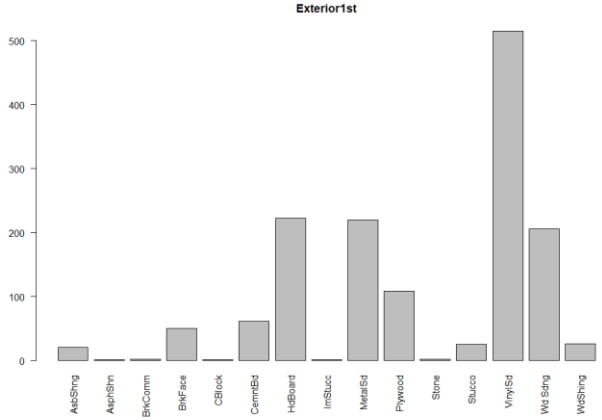
차고 면적 – 전체 층 면적 – 건물 부지 면적 세 변수 사이에 양의 상관관계가 조금 보인다.

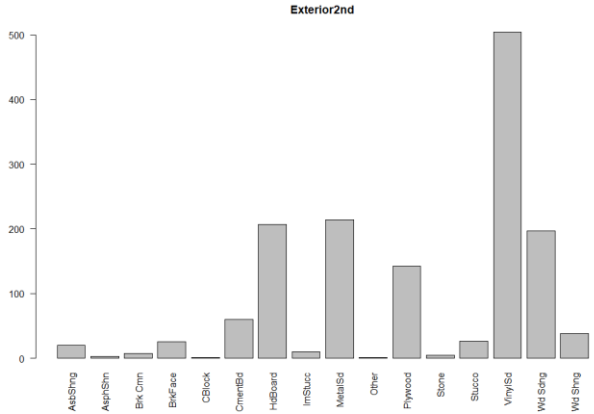
수영장 면적은 사실 99.5%가 0이기 때문에 의미는 없다.

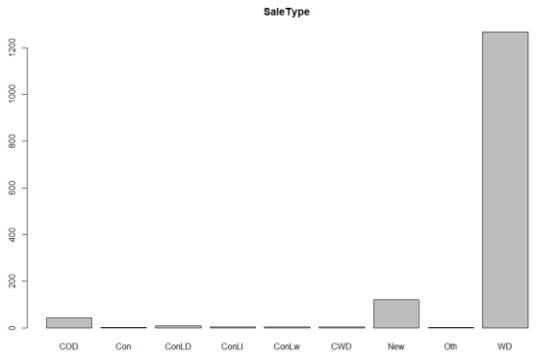
* 범주 개수가 많은 데이터들의 범주별 빈도수(기준: 명목형 중 6개 이상)

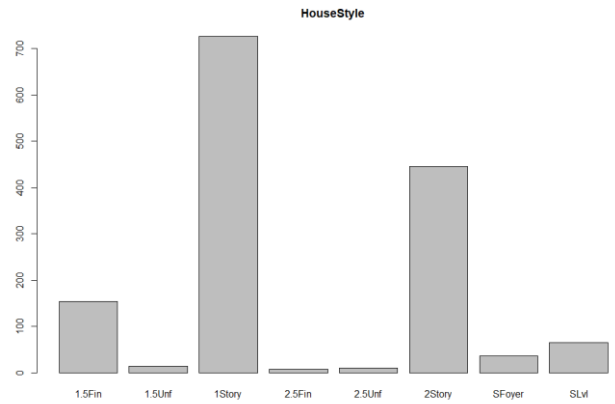


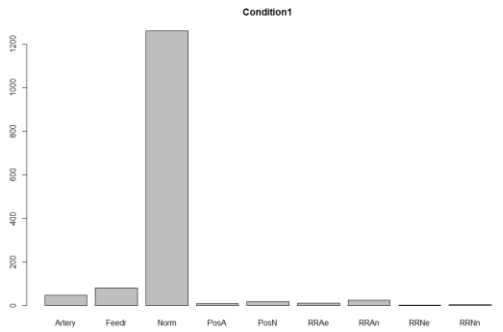


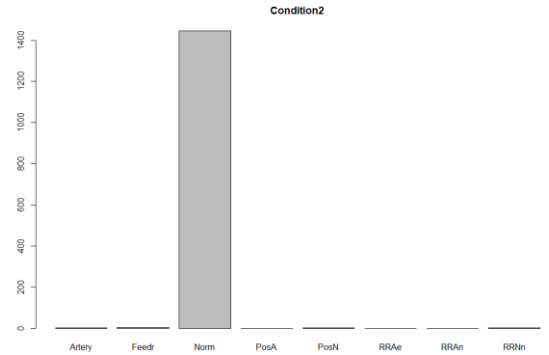


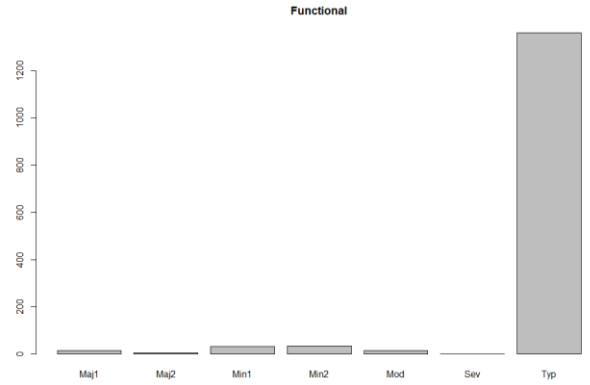


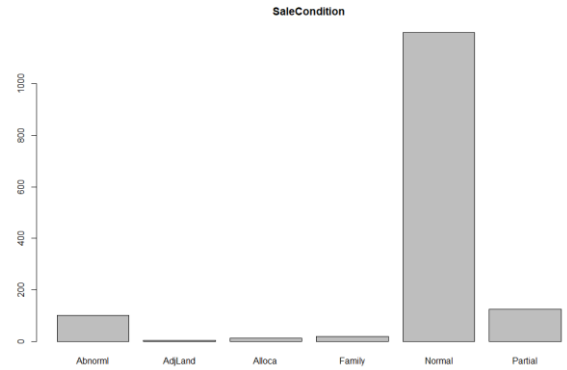








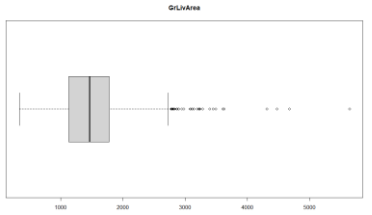


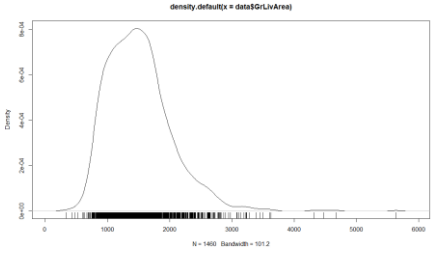


Exterior1 = Exterior2 인 것이 1245개이다.(중복정보)
Condition1 = Condition2 인 것이 1265개이다.(중복정보)

- Neighborhood와 MSSubClass 변수를 제외하고 나머지 변수에는 거의 존재하지 않는 범주가 많다. 이들을 제거하거나 비슷한 성질을 가진 범주끼리 묶어서 범주의 수를 줄이는 것이 좋을 것 같다.

주제 < **주택의 어떤 요소가 거주면적과 밀접한 관련이 있을까?** >





주제에서 이야기하는 거주면적은 data set에서 'GrLivArea' 변수를 말한다.

우선 'GrLivArea' 변수는 테이블 정의서에서 'Above grade (ground) living area square feet' 로 정의되어 있다.

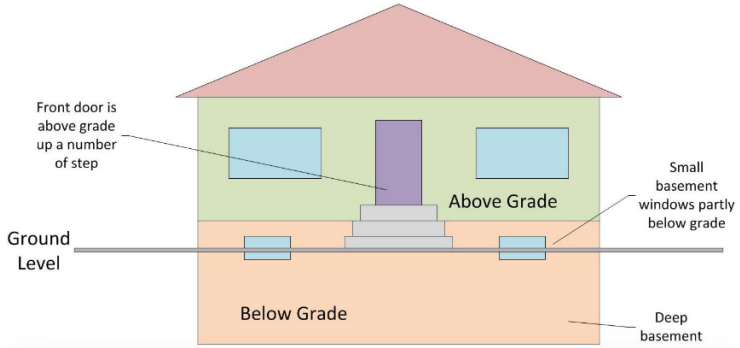
Living area square feet는 말 그대로 거주(생활) 가능한 2차원 공간에 대한 면적(단위: square feet)이고,

Above grade (ground) 부분이 건축 관련 도메인이 있지 않은 나에게는 생소한 단어이다.

구글에서 'above grade'가 무엇을 의미하는지 찾아보았다. 오른쪽은 정의이고, 아래는 이해를 돕기 위한 시각자료이다.
(출처: https://www.gimme-shelter.com/what-is-a-bungalow-50099/
https://www.clearcapital.com/resources/glossary-of-terms/above-grade-square-feet/)

Definition:

Above Grade Square Feet is the term referring to all living square feet in a home that is above the ground. It does not include basements even if the basement is a finished walkout or daylight basement. Lower level rooms are counted in appraisals and BPOs as below grade square feet, and are usually valued at a lower price per square foot. In BPOs, only above grade square feet is counted in the living square feet.



Ground level에 정확히 걸쳐 있는 상태의 'On Grade' 용어도 있지만, data에서는 above만을 정의하고 있기에 넘어가기로 한다.
(개인적인 생각으로는 ground라는 단어가 on grade의 의미를 내포하는 것 같기도 하다.)

결론적으로 'Above grade (ground) living area square feet'는 지면 위 거주 가능한 공간에 대한 면적을 의미한다. (not basement)

데이터에는 nonnegative integer 형태로 들어가 있다.

```
> typeof(data$GrLivArea)
[1] "integer"
```

건물에는 여러 종류가 있겠지만 이 data set에서 의미하는 건물은 주택으로 한정된 것으로 보인다.

차고, 지하실부터 지붕, 현관 등 주택에 필수적일 법 한 변수들이 있고,

오른쪽 사진들은 'MSSubClass' 변수 안에 있는 범주들 중 일부의 사진인데

이 역시 주택의 모양을 하고 있다.

Split Foyer

Duplex

주택의 구성요소 중 가장 중요한 요소 하나를 꼽으라면 거주면적을 꼽을 것이다.

아무리 건물이 크고 화려하더라도 내부에 생활 공간이 좁다면 그렇게 답답할 수가 없을 것이다.

물리적으로, 거주면적은 건물의 크기 즉, 내부 전체 면적에 dependent 할 것이고 (거주면적 > 내부 전체 면적: realistically impossible)

건물 내부의 공간을 차지하는 다른 구성 요소(ex 화장실, 창고, 보일러실 등)가 많으면 그만큼 실제 생활할 수 있는 면적은 줄어들 것이다.

한편, 우리가 일반적으로 (관리가 잘 된 집) -> (관리할 곳이 많지 않음) -> (작은 집) -> (작은 거주면적) ,

(지하실이 넓은 집) -> (큰 건물) -> (넓은 거주면적) 또는 (높은 퀄리티의 건축 자재 사용) -> (소득 수준 높음) -> (큰 집) -> (넓은 거주면적)

과 같은 흐름으로 생각하는 것은 지극히 합리적이라고 보여진다.

화살표로 방향을 표시함으로써 인과관계를 나타낸 것은 아니고, 중간의 과정을 생략하여 (관리가 잘 된 집) – (작은 거주면적) ,

(지하실이 넓은 집) – (넓은 거주면적) 과 같은 관계가 있는지, 어떤 변수가 거주면적을 잘 설명할 수 있는 변수인지 분석해보려고 한다.

데이터에 거주면적과 매우 밀접한 변수가 있는데, '1stFlrSF'와 '2ndFlrSF'이다.

이 두 변수는 각각 1층 면적, 2층 면적을 의미하는데 기존에 확인한 바로는 '1층 면적+2층 면적 = 거주면적'이 되는 관측치가 총 1434개이다.

너무 직접적으로 거주면적을 설명하고 있고, 이 둘을 들고 간다면 모델에서 영향력이 너무 강해 분석의 취지와 맞지 않는다고 판단하여

두 변수는 제거하고 모델을 만들 것이다.

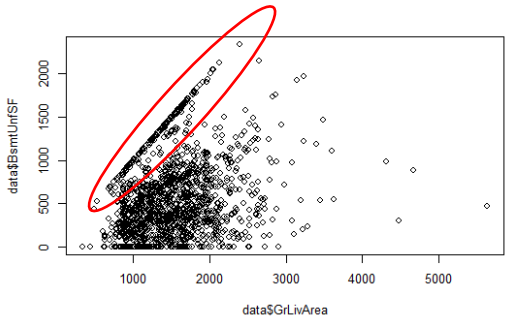
한 가지 이상한 점이 'BsmtUnfSF' 와 'GrLivArea' 의 값이 같은 경우가 96개 있다.

```
> sum(data$BsmtUnfSF==data$GrLivArea)
[1] 96
```

'BsmtUnfSF' 변수는 테이블 정의서에 의하면 Unfinished square feet of basement area로 공사가 끝나지 않은 지하실 면적인데,

왜 두 값이 같은지는 잘 모르겠다. (오른쪽 plot에서 빨간 부분이 이상하다고 판단되어 확인해보았다.)

삭제하기에는 개수가 많은 것도 아니고 정보가 아깝다고 생각해서 전처리 후 사용할 계획이다.



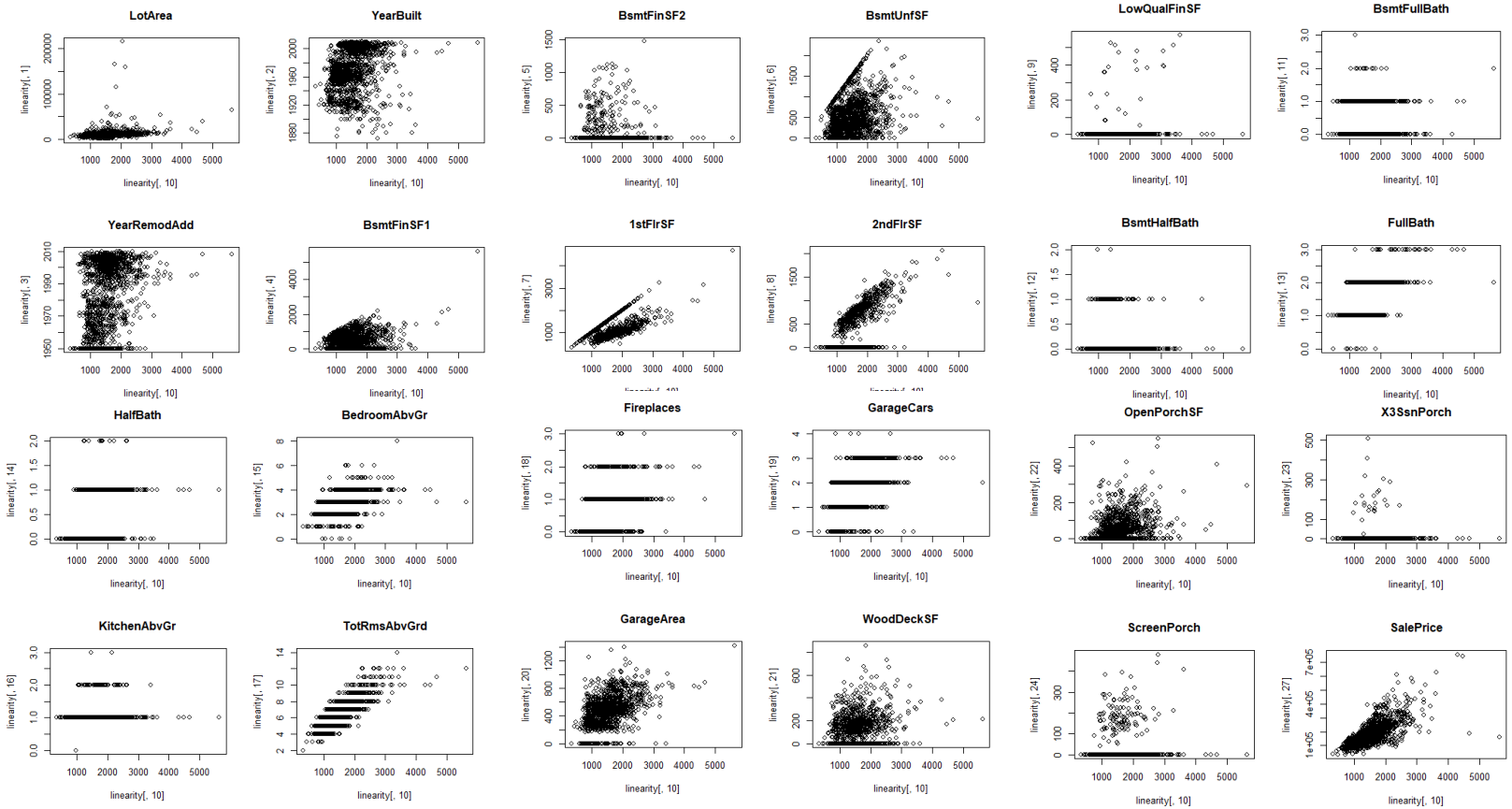
Method1. Regression

어떤 변수가 다른 변수에 영향을 주는지 알 수 있는 대표적인 방법으로 회귀분석이 있다.

GrLivArea 변수를 종속변수로 두고, 나머지 변수를 독립변수로 두어 다중회귀분석을 하는 방법이다.

다만 회귀모형에는 몇 가지 가정들이 있고, 이를 만족하는지 확인해 보아야 한다.

아래는 종속변수인 GrLivArea를 가로축에, numerical 변수들을 세로축으로 놓은 plot들이다. (linearity 확인)



이외에 범주형 변수까지 포함한다면 변수의 개수가 매우 많기 때문에 비슷한 속성(의미)을 가진 변수들을 묶어 PCA를 하거나 합칠 수 있으면 합칠 계획이다.

아직 전처리를 하지 않은 상태이기에 missing, outlier가 있고, 이분산(heteroskedasticity)이 일부 존재할 가능성을 확인할 수 있다.

몇몇 변수는 선형성이 없어 보이기도 하나, 전처리 이후 다시 체크해서 해결하는 방향으로 계획한다.

Method2. Random Forest

Random Forest는 트리모형을 합쳐 만든 앙상블 모형이다.

블랙박스 모형이고, 여러 모형을 합쳤기 때문에 해석이 쉽지 않다.

그러나 정확도 향상에 얼마나 기여를 했는지 알 수 있는 Feature importance 지표가 나오는데,

이를 통해 모형에서 어떤 변수가 종속변수 예측에 얼마나 영향력을 가졌는지 알 수 있고,

영향력이 클수록, 종속변수에 대한 설명력이 크다고 볼 수 있다.

Method3. XGBoost

XGBoost도 트리모형 기반 앙상블 모형인데, Random Forest는 bagging 방식인 반면 XGBoost는 boosting 방식으로 앙상블 방식이 다르다.

XGBoost는 분기가 나누어지면서 변수별로 Gain 점수를 얻는데, 이를 총합하면 Feature importance가 나온다.

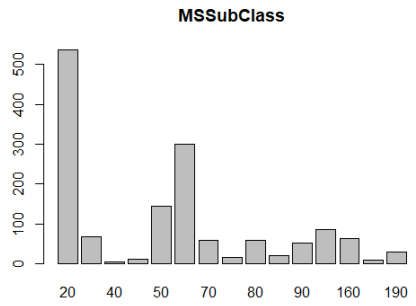
Feature importance는 어떤 변수로 분기를 나눌 때 이게 전체적인 모형 성능에 얼마나 큰 영향을 주는지 체크할 수 있게 된다.

이를 이용하면 모형에서 어떤 변수가 중요도가 높은지, 다시 말하면 영향력이 높은지 알 수 있게 된다.

위 두 모형은 회귀모형에서처럼 계수를 비교하여 독립변수가 종속변수에 직접적으로 어떤 영향을 미치는지는 알 수 없지만,

영향성의 양(quantity)으로써 어떤 변수가 거주면적에 깊은 관여를 하고 있는지는 알 수 있을 것이라 생각한다.

<Data preprocessing>



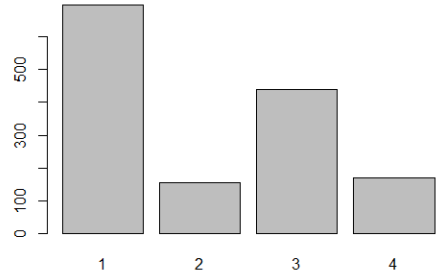
20	1-STORY 1946 & NEWER ALL STYLES
30	1-STORY 1945 & OLDER
40	1-STORY W/FINISHED ATTIC ALL AGES
45	1-1/2 STORY - UNFINISHED ALL AGES
50	1-1/2 STORY FINISHED ALL AGES
60	2-STORY 1946 & NEWER
70	2-STORY 1945 & OLDER
75	2-1/2 STORY ALL AGES
80	SPLIT OR MULTI-LEVEL
85	SPLIT FOYER
90	DUPLEX - ALL STYLES AND AGES
120	1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150	1-1/2 STORY PUD - ALL AGES
160	2-STORY PUD - 1946 & NEWER
180	PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190	2 FAMILY CONVERSION - ALL STYLES AND AGES

MSSubClass 변수: 건물 유형

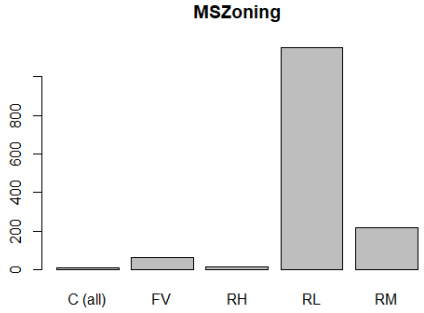
범주가 16개로 매우 많다.

데이터 정의서의 Story를 기준으로 묶어서

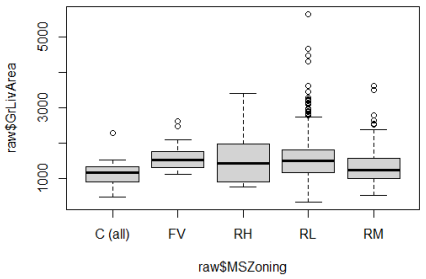
1-story 계열은 1 , 1-1/2 story 계열은 2 ,
2-story 계열은 3 , 나머지는 4



전처리 후 빈도수
1: 696 / 2: 156 / 3: 438 / 4: 170



A	Agriculture
C	Commercial
FV	Floating Village Residential
I	Industrial
RH	Residential High Density
RL	Residential Low Density
RP	Residential Low Density Park
RM	Residential Medium Density



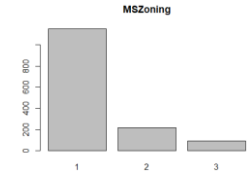
MSZoning 변수: 용도지역 분류

범주가 많진 않지만 RL 범주에 치우쳐져 있다.

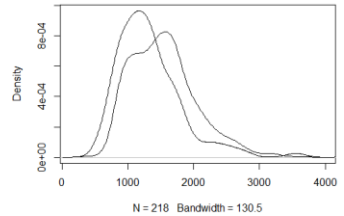
그나마 RM에 데이터가 조금 있다.

RL과 RM에 대한 등분산성 검정 후 t-test 결과
유의미한 차이가 있는 것으로 나왔다.

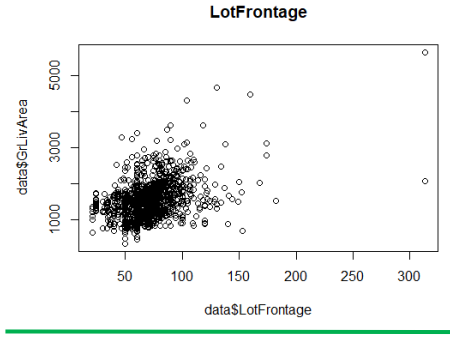
RL을 1 , RM을 2 , 나머지는 3



전처리 후 빈도수
1: 1151 / 2: 218 / 3: 91



RL과 RM 분포



LotFrontage: 건물에서 인접 거리까지 직선 길이

결측치가 259개가 있다. 18%정도로 무시하긴 힘든 비율이기에,

기존의 데이터에서 값을 random sampling 해서 그 값을 결측치에 할당하였다.

(데이터가 어느 정도 군집해 있기에 랜덤 값을 할당함)

Street: 건물과 인접한 도로 유형

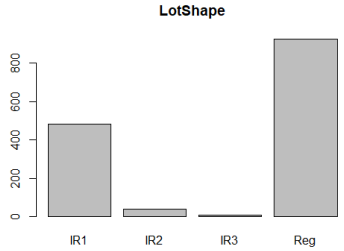
Pave범주가 극단적으로 많기에 변수 삭제

(구분의 의미가 없음)

```
> sum(is.na(data$Alley))
[1] 1369
```

Alley: 건물과 인접한 골목 유형

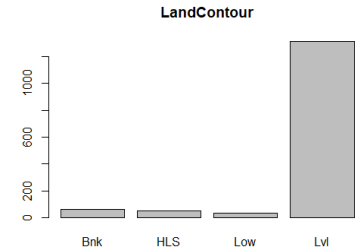
Missing이 약 94%로 변수 삭제



Reg	Regular
IR1	Slightly irregular
IR2	Moderately Irregular
IR3	Irregular

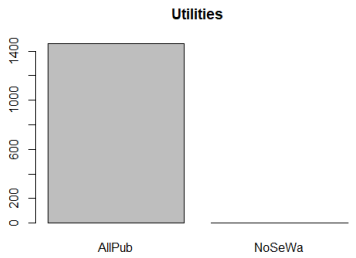
LotShape: 건물의 모양

IR2, IR3는 거의 없음. 따라서 IR1, 2, 3을 1로 묶고,
Reg를 0으로 묶음



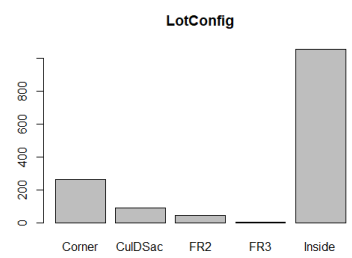
LandContour: 토지 평탄정도

평탄함(Lvl): 0,
평탄하지 않음(나머지): 1



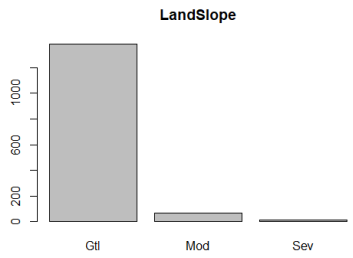
Utilities: 수도전기가스

거의 AllPub. 변수 삭제



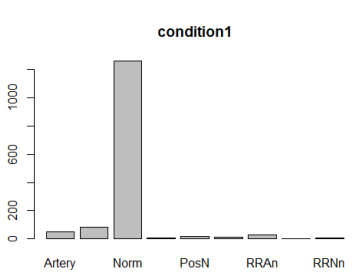
LotConfig: 건물 부지 구조

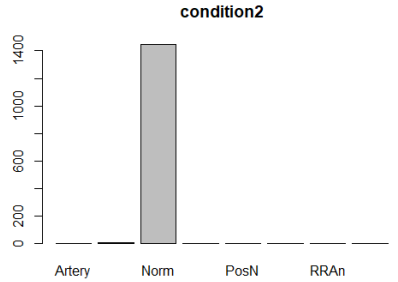
Inside: 1 , corner: 2, 나머지 3



LandSlope: 토지경사

Gtl: 0 , 나머지 1

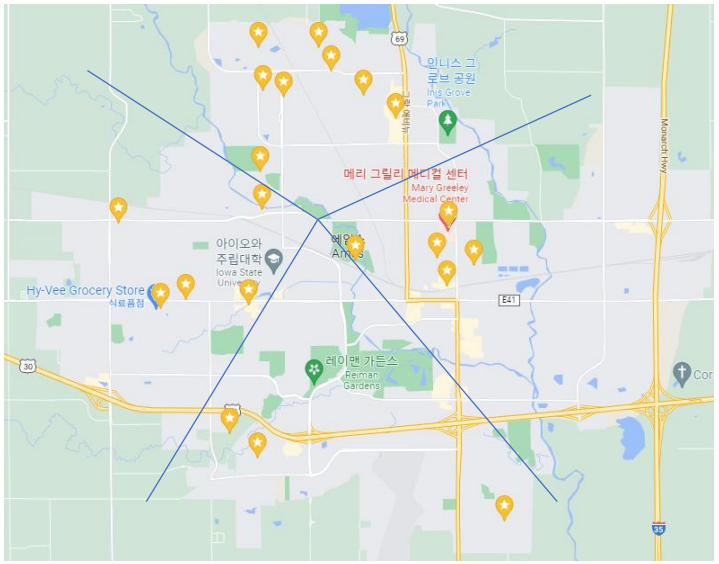





```
   0    1
1260  200
```

Condition: 교통접근성

Condition2의 경우는 2개 이상인 case 라고 하지만 1도 norm이 가장 많고 2도 norm이 가장 많다. 이는 중복으로 데이터가 들어간 것이기에 norm 외에 다른 범주가 거의 없는 condition2 변수는 제거하고, condition1도 norm이 0, 나머지는 1로 넣는다. (norm이 아닌 나머지는 교통편의성과 매우 가까운 범주들임)

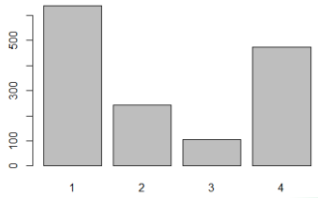


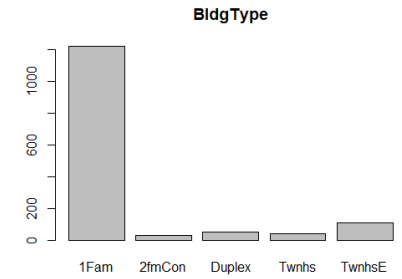
Neighborhood: Ames city 내 위치

미국 아이오와 주의 한 도시이고, 데이터에 나와 있는 범주를 구글맵에 매핑한 그림이 왼쪽 그림이다.

지형, 건물 밀집도, 교통을 고려하여 동서남북 4방향으로 구역을 나누었다.

북: 1 , 동: 2, 남: 3, 서: 4





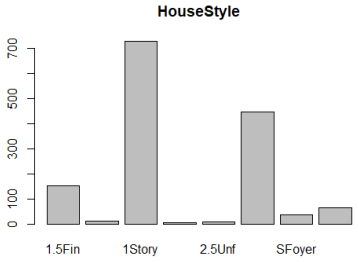

```
   1    2
1251  209
```

```
1Fam   Single-family Detached
2FmCon      Two-family Conversion; originally built as one-family dwelling
Duplx  Duplex
TwnhsE      Townhouse End Unit
TwnhsI Townhouse Inside Unit
```

BldgType: 주택형태

데이터 정의서를 보면 duplex, townhouse의 경우는 2개 이상의 가구가 붙어있도록 크게 지은 건물들이고, 1Fam과 2FmCon은 한 가족이 살 수 있도록 한 주택 형태이다.

1Fam과 2FmCon을 1, 나머지를 2로 묶는다.

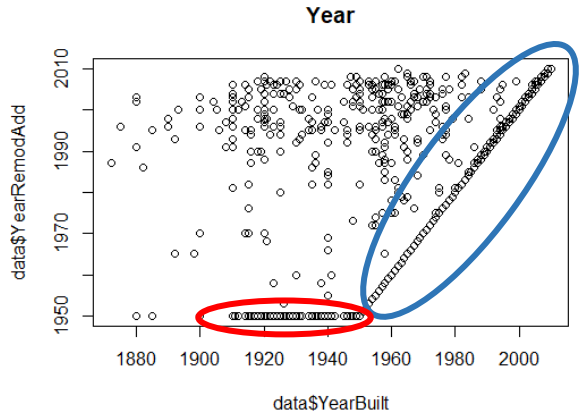



```
1Story One story
1.5Fin  One and one-half story: 2nd level finished
1.5Unf One and one-half story: 2nd level unfinished
2Story Two story
2.5Fin  Two and one-half story: 2nd level finished
2.5Unf Two and one-half story: 2nd level unfinished
SFoyer Split Foyer
SLvl    Split Level
```

HouseStyle: 주택 스타일

One story~ 는 1
Two story~ 는 2
Split는 3

```
   1    2    3
894  464  102
```



YearBuilt: 시공연도 / YearRemodAdd: 리모델링 연도(없으면 시공일과 동일)

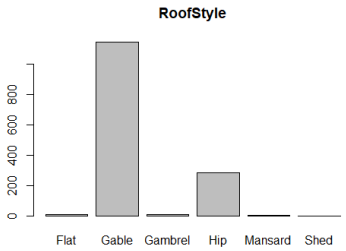
왼쪽 plot은 가로축에 시공연도, 세로축에 리모델링 연도를 놓고 그린 plot이다. 리모델링하지 않았으면 시공연도와 동일해야 하므로 파란 부분처럼 linear하게 찍히는 것은 정상적이다. 하지만 빨간 부분을 보면 1950년 수십 가구가 동시에 리모델링한 듯이 비정상적으로 보인다. 본인 생각에는 1950년대 이전에 리모델링 된 가구는 그냥 1950년으로 놓은 것 같다.

문제는 1920년에 시공을 하고, 1940년에 리모델링을 한 가구는 1920/1950으로 데이터가 나오겠지만, 1920년에 시공을 하고, 리모델링하지 않은 가구도 1920/1950으로 나올 것이다. 즉, 1950년 이전에 짓고, 리모델링을 1950년 이전에 했다면 리모델링 했는지 안 했는지 알 수가 없게 된다.

이런 문제가 있는 데이터는 총 178개이다. Plot을 보면 1990년대부터 리모델링을 활발하게 하는 것으로 보이기도 하고, 이런 추세로 보았을 때 1950년 이후 리모델링하지 않은 집들은 여전히 리모델링하지 않았을 가능성이 높아 보인다.

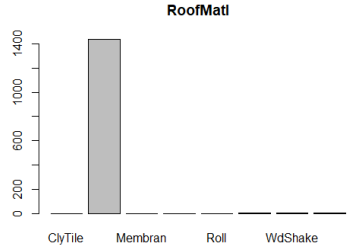
따라서 알 수 없는 집들은 리모델링하지 않은 것으로 가정한다.

리모델링 안함: 0 , 리모델링 함: 1 (새로운 변수 Remod, 기존 변수 2개는 제거)



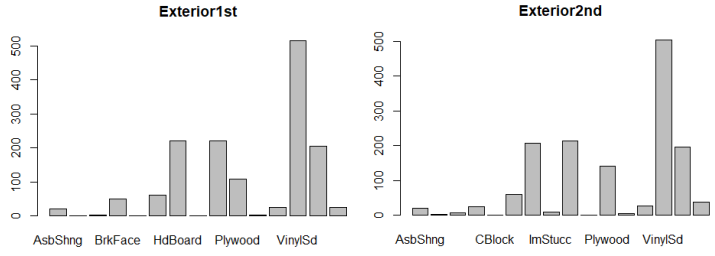
RoofStyle: 지붕형태

Gable은 0, 나머지는 1



RoofMatl: 지붕소재

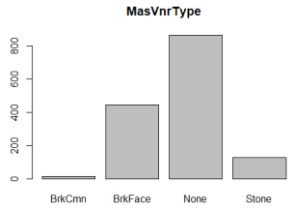
CompShg 범주 이외에는 거의 없음. -> 변수 제거




```
> sum(data$Exterior1st == data$Exterior2nd)
[1] 1245
```

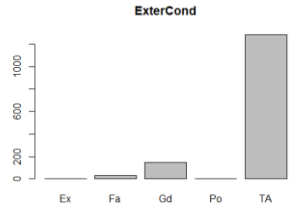
Exterior: 외부벽 소재

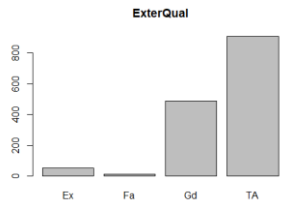
이 역시 condition과 마찬가지로 1과 2가 동일한 데이터가 매우 많다. 1245개. 왼쪽의 히스토그램을 보면 범주별 비율이 거의 동일한 것으로 보이긴 하지만, 다른 벽을 가진 215개 데이터를 무시하는 것은 좋지 않을 것 같다.
따라서 exterior1st 변수는 살리고, exterior2nd 변수는 제거하며, Exterior_d 변수를 하나 추가하여 동일한 벽이면 0, 다른 벽이면 1로 둔다.
또한 exterior1st변수는 Exterior로 이름을 바꾸고,
상위 4개의 범주 VinylSd: 1, HdBoard: 2, MetalSd: 3, Wd Sdng: 4, 나머지는 5로 한다.



MasVnrType: 석조 베니어벽 소재

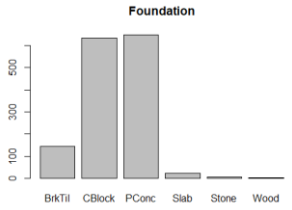
None이 베니어벽이 없는 경우인데 800개 이상으로 절반이 넘는다. 또한 stone이나 brkcmn은 수가 적으므로 none인 경우 0, 베니어벽이 있는 경우는 1로 가져간다.
MasVnrArea(석조 베니어벽 면적)와 동일하게 결측치가 8개가 있는데 결측치는 없는 것으로 가정하고 0으로 둔다.





ExterCond: 건물 외부 자재 현재 상태,
ExterQual: 건물 외부 자재 품질

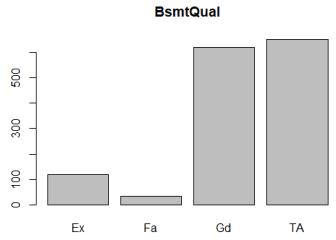
TA는 평균적인 상태/품질이고 이보다 좋으면 Gd, Ex이고 이보다 나쁘면 Fa, Po인데, 3가지로 나눈다면 나쁨/보통/좋음 으로 나뉜다. ExterCond는 극단에 있는 범주 Ex와 Po가 거의 없다. 이렇게 되면 사실 평균적인 수치에 몰려 있는 것이고, 나눈다해도 의미가 없어보인다. ExterCond 제거.
ExterQual의 경우는 Fa, Po가 거의 없고, 전부 평균이상인 상태이다. 따라서 TA, Fa는 0, 평균 이상인 Gd와 Ex는 1로 가져간다.



Foundation: 건물 기초공사 자재

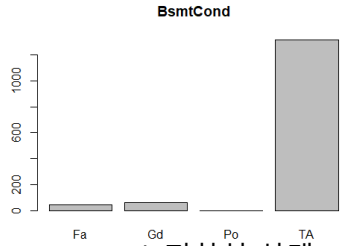
Pconc: 1, Cblock: 2, 나머지는 3

Bsmt~는 지하실에 대한 변수들인데 결측치가 37, 38개가 있음. 이 관측치들의 경우는 지하실이 없는 것으로 보여짐 (지하실면적이 0). 지하실이 있는 것과 없는 것은 집의 전체적인 크기와도 관련이 있기에 이 차이가 유의할 수도 있으나 지하실이 없는 집이 있는 집에 비해 매우 작은 관측치여서 모형에서 의미가 있다고 해도 실제적으로 유의한지에 대해 논의하기에는 다소 에러가 있지 않을까 판단하였음.
따라서 관측치를 제거.



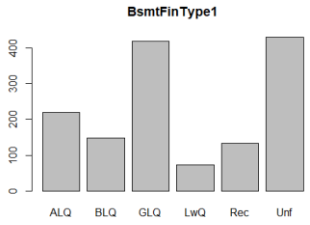
BsmtQual: 지하실 높이

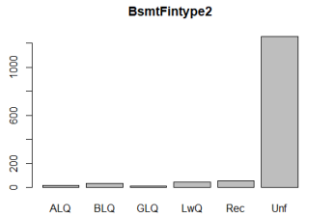
TA, Fa는 0, Ex, Gd는 1



BsmtCond: 지하실 상태

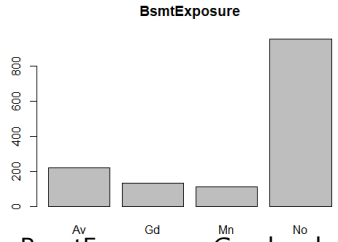
Fa-TA-Gd는 한 단계씩의 차이이고 주관이 들어간 데이터기에 이를 구분하는 것은 낭비가 아닌가 싶음. 변수 제거





BsmtFinType: 공사한 지하실 등급

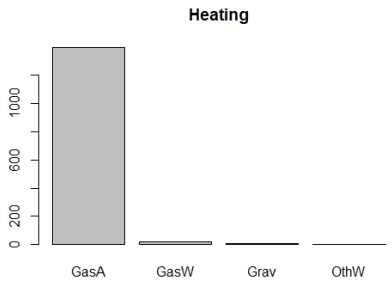
Type2에서는 unf가 1255개이다. 이 말은 지하실이 1개인 집이 1255개인 것과 같다. Type2는 변수를 제거하고, 차라리 지하실이 2개인 집을 따로 구분하는 것이 좋을 것 같다. 새로운 변수 BsmtFinType_d를 하나 추가하여 지하실이 1개면 0, 2개 이상이면 1로 둔다.
Type1에서는 테이블 정의서에 따라 living quarters면 1, rec room이면 2, 나머지는 0으로 둔다.



BsmtExposure: Garden level wall 노출 정도

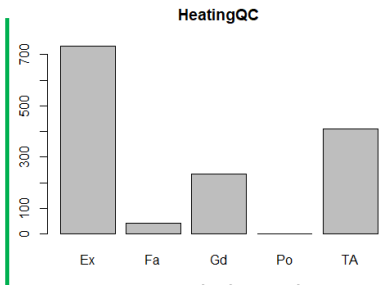
No를 0, 나머지는 1

지하실 면적은 BsmtFinSF1과 BsmtFinSF2, TotalBsmtSF 3개의 변수로 있는데, BsmtFinSF1+BsmtFinSF2=TotalBsmtSF인 관계가 성립하므로, TotalBsmtSF 변수 하나만 들고가고 나머지는 삭제한다.



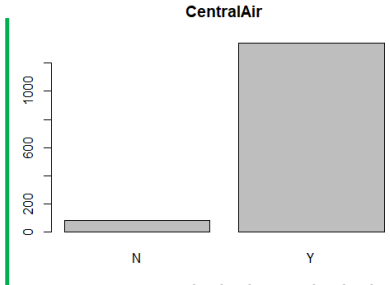
Heating: 난방 형태

GasA 가 1395개로 대다수를 차지하고 있기에 다른 형태와 구분하는 것은 의미가 없어보임.
-> 변수 제거



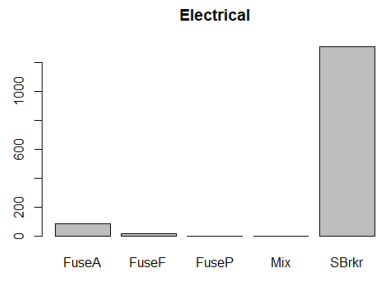
HeatingQC: 난방 품질

Po, Fa, TA를 0, Gd를 1, Ex를 2로 한다.



CentralAir: 에어컨 중앙제어

Y를 1, N을 0



Electrical: 전기시스템

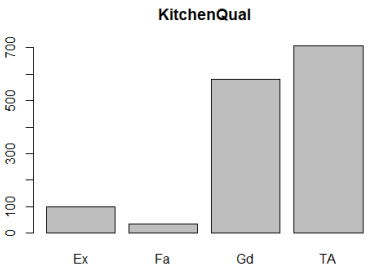
결측치 1개는 제거

SBrkr은 0, 나머지는 1

1stFlrSF는 1층 면적, 2ndFlrSF는 2층 면적, LowQualFinSF는 저품질 면적이다. 이 셋을 더하면 GrLivArea 값이 나온다. 따라서 그대로 쓰지 않고, 전체 면적에서 1층이 차지하는 비율, 2층이 차지하는 비율, 저품질 면적이 차지하는 비율로 나누어 직접적인 수치가 아닌 1, 2층, 저품질의 비율 정보로 가져간다.
새로운 변수 1stR, 2ndR, LowR을 만들어 전체 면적에서 이 셋이 각각 차지하는 비율 값을 넣는다. 기존에 있던 변수는 삭제
(LowR 변수는 삭제하였음. 1420개의 관측치 중 1395개가 0)

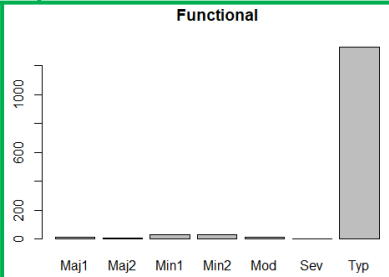
FireplaceQu: 벽난로 품질

결측치가 690개로 거의 절반이다. 벽난로 품질보다는 벽난로 수만 가져가는 것이 좋을 것 같다. 변수 제거



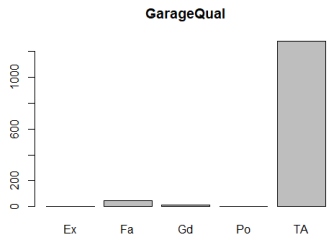
KitchenQual: 부엌 품질

Fa, TA는 0, Gd는 1, Ex는 2



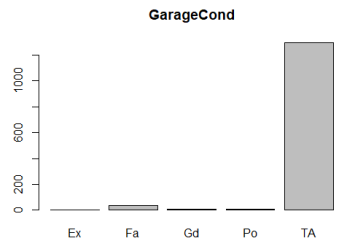
Functional: home functionality
구분해주기엔 다른 범주들의 수가 너무 작다. 변수 제거

Garage는 차고 관련 변수이다. 결측치는 81개로 이들은 차고가 없는 집인 것으로 보인다(차고 면적 0).
차고가 있고 없고는 전체적인 집 크기와도 관련이 있기에 정보를 완전히 무시하기에는 힘들어 보인다.
면적과 같은 수치형 변수는 어차피 0으로 들어가 있기에, 범주형 변수만 최빈범주로 대치하여 결측치를 처리하였다.
건축일의 경우는 시공연도와 동일하다고 가정하고 결측치를 처리하였다(시공연도와 차고 건축일이 같은 경우는 1379개 관측치 중 1089개).



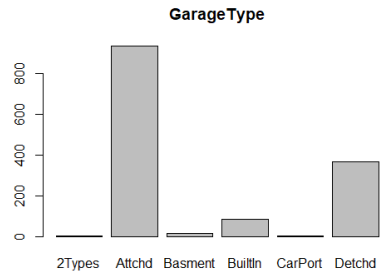
GarageQual: 차고 품질

구분의 의미가 없어보인다.
변수 제거



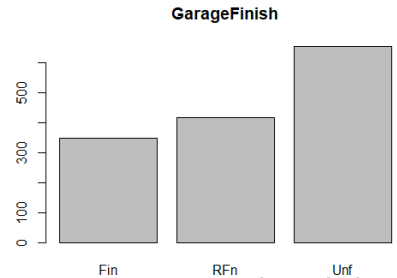
GarageCond: 차고 상태

구분의 의미가 없어보인다.
변수 제거



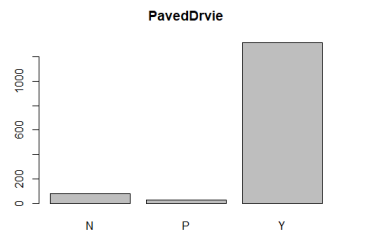
GarageType: 차고 위치

Attchd와 BuiltIn은 0, 나머지는 1



GarageFinish: 차고 완성도

Unf: 0, 나머지 1



PavedDrive: 진입로 포장상태

데이터 정의서에 의해
P, Y: 1, N: 0

Porch~는 현관 관련 공간 면적변수이다. 총 4가지가 있는데, 대부분 1~2개를 가지고 있다. 이 4개를 전부 쓰는 것보단 4개 변수의 평균값으로 대체해서 하나로 가져가는 것이 좋을 것 같다.
새로운 변수 PorchSF를 만들어 4개 변수의 평균을 넣고 나머지는 삭제

Pool~은 수영장 관련 변수인데, 결측치가 99.5%로 거의 모두이다.
따라서 품질 변수인 PoolQC는 삭제하고, 면적같은 경우는 어차피 없는 경우 0이기 때문에 일단은 들고 간다.

Misc~은 기타 특징들(데이터에 없는 또다른 옵션)인데, 결측치가 96.3%이다.
이 역시 옵션의 종류를 구별하는 MiscFeature는 제거하고 이 가치에 대한 정보를 가지고 있는 MiscVal은 들고 간다. 어차피 없는 경우 0이기 때문이다.

매각일과 관련한 변수는 사실 집의 특징과 전혀 상관이 없기에 삭제하고, 판매 형태, 판매 조건 또한 삭제한다.
판매 가격은 집의 크기와도 상관이 있을 것 같으므로 들고 간다.

<Regression> (범주형 변수들은 caret패키지로 one-hot encoding하였음)

```
Call:
lm(formula = GrLivArea ~ ., data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-640.34  -94.54  -12.42   79.32 1260.20

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.515e+03  6.422e+02   5.473 5.28e-08 ***
MSSubClass1    7.372e+01  3.772e+01   1.954 0.050859 .
MSSubClass2    8.496e+01  4.168e+01   2.038 0.041703 *
MSSubClass3    3.965e+01  4.877e+01   0.813 0.416289
MSZoning1      2.398e+01  2.143e+01   1.119 0.263385
MSZoning2      4.887e+01  2.465e+01   1.983 0.047553 *
LotFrontage    6.247e-01  2.405e-01   2.598 0.009488 **
LotArea       -9.354e-05  5.379e-04  -0.174 0.861986
LotShape       8.083e+00  1.076e+01   0.751 0.452837
LandContour    2.311e+01  1.864e+01   1.239 0.215421
LotConfig1     9.903e+00  1.644e+01   0.602 0.547135
LotConfig2     3.858e+01  1.892e+01   2.039 0.041667 *
LandSlope      7.560e+01  2.488e+01   3.039 0.002421 **
Neighborhood1 -1.676e+01  1.218e+01  -1.376 0.168954
Neighborhood2 -1.519e+00  1.603e+01  -0.095 0.924541
Neighborhood3 -3.017e+01  1.973e+01  -1.529 0.126522
Condition1     2.985e+01  1.391e+01   2.145 0.032091 *
BldgType1     -4.803e+01  1.841e+01  -2.609 0.009178 **
HouseStyle1   -1.605e+02  3.958e+01  -4.055 5.29e-05 ***
HouseStyle2   -1.011e+02  5.482e+01  -1.844 0.065471 .
OverallQual   -9.104e-01  6.715e+00  -0.136 0.892186
OverallCond   -7.231e+00  5.222e+00  -1.385 0.166388
Exterior0     -1.423e+01  1.227e+01  -1.160 0.246116
MasVnrType    -2.938e+01  1.404e+01  -2.093 0.036496 *
MasVnrArea     9.759e-02  3.807e-02   2.564 0.010466 *
ExterQual      2.664e+00  1.647e+01   0.162 0.871507
Foundation1   -8.003e+01  2.133e+01  -3.751 0.000183 ***
Foundation2   -4.103e+01  1.921e+01  -2.136 0.032859 *
BsmtQual      -1.102e+01  1.573e+01  -0.701 0.483655
BsmtExposure  -1.998e+01  1.176e+01  -1.699 0.089611 .
BsmtFinType10  8.093e-01  1.902e+01   0.043 0.966063
BsmtFinType11 -1.161e+01  1.746e+01  -0.665 0.506006
BsmtUnfSF     -3.382e-02  1.859e-02  -1.820 0.069031 .
TotalBsmtSF    4.944e-01  2.464e-02  20.063  < 2e-16 ***
HeatingQC0    -1.118e+01  1.368e+01  -0.817 0.414101
HeatingQC1    -9.887e+00  1.421e+01  -0.696 0.486554
CentralAir    -4.064e+01  2.379e+01  -1.708 0.087773 .
Electrical     2.840e+01  1.894e+01   1.500 0.133969
BsmtFullBath  -1.805e+01  1.333e+01  -1.354 0.175958
BsmtHalfBath  -1.182e+01  2.027e+01  -0.583 0.559699
FullBath       1.063e+02  1.425e+01   7.457 1.57e-13 ***
HalfBath       4.805e+01  1.355e+01   3.546 0.000405 ***
BedroomAbvGr   2.437e+01  9.049e+00   2.693 0.007172 **
KitchenAbvGr   8.832e+00  3.670e+01   0.241 0.809883
KitchenQual0   5.707e+01  2.534e+01   2.252 0.024460 *
KitchenQual1   8.051e+01  2.172e+01   3.707 0.000218 ***
TotRmsAbvGrd   9.488e+01  5.764e+00  16.462  < 2e-16 ***
Fireplaces     4.783e+01  8.802e+00   5.434 6.52e-08 ***
GarageType    -2.297e+01  1.426e+01  -1.611 0.107320
GarageYrBlt   -1.166e+00  3.026e-01  -3.853 0.000122 ***
GarageFinish  -2.116e+01  1.356e+01  -1.560 0.118896
GarageCars    -5.384e+01  1.469e+01  -3.666 0.000256 ***
GarageArea     2.727e-01  5.056e-02   5.395 8.09e-08 ***
PavedDrive     5.202e+00  2.233e+01   0.233 0.815788
WoodDeckSF     9.834e-02  3.990e-02   2.465 0.013822 *
PoolArea       4.933e-01  1.155e-01   4.270 2.09e-05 ***
MiscVal       -2.958e-03  9.317e-03  -0.317 0.750917
SalePrice      1.279e-03  1.391e-04   9.196  < 2e-16 ***
Remod          5.546e+01  1.042e+01   5.322 1.20e-07 ***
Exterior_d     2.379e+01  1.354e+01   1.757 0.079163 .
BsmtFinType_d  8.258e+00  1.548e+01   0.533 0.593832
X1stR         -1.439e+03  1.715e+02  -8.388  < 2e-16 ***
X2ndR         -5.952e+02  1.668e+02  -3.567 0.000373 ***
PorchSF        5.640e-01  1.879e-01   3.001 0.002736 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 168.4 on 1356 degrees of freedom
Multiple R-squared:  0.9024,    Adjusted R-squared:  0.8978
F-statistic: 198.9 on 63 and 1356 DF,  p-value: < 2.2e-16
```

```
> vif(model_1)
  MSSubClass1   MSSubClass2   MSSubClass3     MSZoning1
    17.785989     8.367030     25.352349      3.872277
  MSZoning2   LotFrontage       LotArea      LotShape
     3.894873     1.658986     1.479239      1.355995
 LandContour    LotConfig1    LotConfig2     LandSlope
     1.596525     2.737334     2.658775      1.590193
Neighborhood1 Neighborhood2 Neighborhood3    Condition1
     1.830740     1.808152     1.288600      1.134304
   BldgType1   HouseStyle1   HouseStyle2   OverallQual
     2.010994    18.732497    32.940640      4.171925
 OverallCond     Exterior0    MasVnrType    MasVnrArea
     1.682890     1.286214     2.386168      2.402932
  ExterQual   Foundation1   Foundation2      BsmtQual
     3.194149     5.647331     4.554588      3.096791
BsmtExposure BsmtFinType10 BsmtFinType11    BsmtUnfSF
     1.533630     4.141554     3.775259      3.308392
 TotalBsmtSF    HeatingQC0    HeatingQC1    CentralAir
     4.999894     2.038985     1.391619      1.524832
  Electrical  BsmtFullBath  BsmtHalfBath    FullBath
     1.305879     2.414465     1.203362      3.091555
    HalfBath  BedroomAbvGr  KitchenAbvGr  KitchenQual0
     2.348564     2.692443     2.666148      8.027041
KitchenQual1  TotRmsAbvGrd   Fireplaces    GarageType
     5.707650     4.349194     1.617890      2.063380
 GarageYrBlt  GarageFinish   GarageCars    GarageArea
     2.900807     2.287956     5.977648      5.811025
  PavedDrive   WoodDeckSF     PoolArea       MiscVal
     1.327668     1.269145     1.109228      1.058241
   SalePrice       Remod   Exterior_d BsmtFinType_d
     6.106925     1.264680     1.139138      1.239476
      X1stR        X2ndR       PorchSF
    75.553452    70.713395     1.231545
```

먼저 1층면적과 2층면적의 VIF 값은 70 이상으로 매우 높은 수치이고 p-value도 낮다. 둘을 더하면 거의 1이니 다중공선성이 있는 것은 당연하다. 2층면적 비율을 없애고, 1층면적 비율만 가져가는 것으로 한다.

그 다음으로 MSSubClass 범주들이 모두 VIF가 매우 높다. P-value도 유의하지 않다고 나오고, 모형에 적절치 않은 변수로 보인다. MSSubClass는 전부 제거한다.

HouseStyle의 범주도 VIF가 매우 높다. MSSubClass와 HouseStyle을 데이터 정의서에서 보면, 범주가 거의 유사하다. 이때문에 두 변수 사이에 다중공선성이 발생했고, housestyle이 모형 설명에 있어 조금 더 적합한 변수로 나온 것이 아닌가 싶다. 우선 MSSubClass를 제거하고 다시 모형을 적합시켜 VIF를 확인한다.

```
Call:
lm(formula = GrLivArea ~ ., data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-648.09  -94.54  -14.63   80.78 1251.63

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.774e+03  5.966e+02   4.651 3.63e-06 ***
MSZoning1      1.912e+01  2.148e+01   0.890 0.373548
MSZoning2      4.985e+01  2.471e+01   2.017 0.043884 *
LotFrontage    6.367e-01  2.413e-01   2.638 0.008425 **
LotArea       -2.152e-04  5.378e-04  -0.400 0.689130
LotShape       8.762e+00  1.081e+01   0.811 0.417563
LandContour    2.414e+01  1.864e+01   1.295 0.195648
LotConfig1     9.131e+00  1.649e+01   0.554 0.579933
LotConfig2     3.812e+01  1.899e+01   2.008 0.044869 *
LandSlope      7.743e+01  2.498e+01   3.099 0.001979 **
Neighborhood1 -1.758e+01  1.223e+01  -1.438 0.150754
Neighborhood2 -4.884e+00  1.604e+01  -0.304 0.760821
Neighborhood3 -3.165e+01  1.982e+01  -1.597 0.110484
Condition1     2.995e+01  1.397e+01   2.144 0.032197 *
BldgType1     -4.134e+01  1.820e+01  -2.272 0.023250 *
HouseStyle1   -9.217e+01  2.124e+01  -4.339 1.53e-05 ***
HouseStyle2   -8.052e+01  2.671e+01  -3.015 0.002620 **
OverallQual   -6.649e-01  6.733e+00  -0.099 0.921357
OverallCond   -7.427e+00  5.223e+00  -1.422 0.155275
Exterior0     -1.396e+01  1.220e+01  -1.144 0.252684
MasVnrType    -3.086e+01  1.409e+01  -2.190 0.028678 *
MasVnrArea     9.718e-02  3.821e-02   2.544 0.011085 *
ExterQual      4.660e+00  1.653e+01   0.282 0.778040
Foundation1   -8.096e+01  2.137e+01  -3.789 0.000158 ***
Foundation2   -4.328e+01  1.915e+01  -2.260 0.023973 *
BsmtQual      -1.553e+01  1.569e+01  -0.990 0.322525
BsmtExposure  -2.261e+01  1.173e+01  -1.929 0.053992 .
BsmtFinType10  1.290e-01  1.907e+01   0.007 0.994604
BsmtFinType11 -1.259e+01  1.753e+01  -0.718 0.472594
BsmtUnfSF     -3.318e-02  1.867e-02  -1.777 0.075803 .
TotalBsmtSF    4.998e-01  2.451e-02  20.390  < 2e-16 ***
HeatingQC0    -8.798e+00  1.372e+01  -0.641 0.521377
HeatingQC1    -8.436e+00  1.424e+01  -0.592 0.553674
CentralAir    -3.609e+01  2.363e+01  -1.527 0.126951
Electrical     2.969e+01  1.900e+01   1.563 0.118275
BsmtFullBath  -1.905e+01  1.335e+01  -1.427 0.153780
BsmtHalfBath  -1.124e+01  2.034e+01  -0.553 0.580639
FullBath       1.030e+02  1.428e+01   7.213 9.04e-13 ***
HalfBath       4.369e+01  1.352e+01   3.231 0.001261 **
BedroomAbvGr   1.954e+01  8.884e+00   2.199 0.028040 *
KitchenAbvGr  -3.471e+01  2.868e+01  -1.210 0.226298
KitchenQual0   5.763e+01  2.544e+01   2.266 0.023631 *
KitchenQual1   7.997e+01  2.177e+01   3.673 0.000249 ***
TotRmsAbvGrd   9.719e+01  5.757e+00  16.884  < 2e-16 ***
Fireplaces     4.805e+01  8.836e+00   5.438 6.38e-08 ***
GarageType    -2.136e+01  1.428e+01  -1.496 0.134833
GarageYrBlt   -1.044e+00  3.008e-01  -3.471 0.000534 ***
GarageFinish  -2.122e+01  1.356e+01  -1.565 0.117819
GarageCars    -5.324e+01  1.473e+01  -3.614 0.000313 ***
GarageArea     2.629e-01  5.070e-02   5.185 2.49e-07 ***
PavedDrive     6.127e+00  2.240e+01   0.273 0.784537
WoodDeckSF     9.801e-02  4.008e-02   2.445 0.014593 *
PoolArea       5.027e-01  1.160e-01   4.335 1.57e-05 ***
MiscVal       -3.924e-03  9.347e-03  -0.420 0.674674
SalePrice      1.270e-03  1.397e-04   9.093  < 2e-16 ***
Remod          5.671e+01  1.046e+01   5.424 6.89e-08 ***
Exterior_d     2.282e+01  1.357e+01   1.682 0.092827 .
BsmtFinType_d  1.032e+01  1.551e+01   0.665 0.506082
X1stR         -8.919e+02  5.134e+01 -17.372  < 2e-16 ***
PorchSF        5.774e-01  1.884e-01   3.066 0.002216 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 169.1 on 1360 degrees of freedom
Multiple R-squared:  0.9012,    Adjusted R-squared:  0.8969
F-statistic: 210.2 on 59 and 1360 DF,  p-value: < 2.2e-16
```

```
> vif(model_2)
    MSZoning1     MSZoning2   LotFrontage       LotArea
     3.851756      3.879667      1.655417      1.464873
     LotShape   LandContour    LotConfig1    LotConfig2
     1.353760      1.581849      2.727926      2.652922
    LandSlope Neighborhood1 Neighborhood2 Neighborhood3
     1.588767      1.829108      1.793753      1.287742
   Condition1     BldgType1   HouseStyle1   HouseStyle2
     1.132095      1.947010      5.345598      7.746254
  OverallQual   OverallCond     Exterior0    MasVnrType
     4.155341      1.667898      1.260939      2.382156
  MasVnrArea     ExterQual   Foundation1   Foundation2
     2.398369      3.188977      5.612546      4.485936
    BsmtQual  BsmtExposure BsmtFinType10 BsmtFinType11
     3.050382      1.509596      4.126593      3.769981
   BsmtUnfSF   TotalBsmtSF    HeatingQC0    HeatingQC1
     3.307509      4.901585      2.031130      1.385345
  CentralAir    Electrical  BsmtFullBath  BsmtHalfBath
     1.490903      1.301178      2.400898      1.201333
    FullBath      HalfBath  BedroomAbvGr  KitchenAbvGr
     3.075591      2.316832      2.571103      1.612486
 KitchenQual0  KitchenQual1  TotRmsAbvGrd    Fireplaces
     8.013888      5.683091      4.298200      1.615287
  GarageType   GarageYrBlt  GarageFinish    GarageCars
     2.050922      2.840068      2.267797      5.958295
  GarageArea    PavedDrive    WoodDeckSF     PoolArea
     5.790689      1.324598      1.269051      1.107275
     MiscVal     SalePrice         Remod    Exterior_d
     1.055311      6.104384      1.261575      1.133236
BsmtFinType_d        X1stR      PorchSF
     1.232721      6.704613      1.225931
```

다행히 HouseStyle과 1층 면적 비율의 VIF가 상당히 많이 줄어들었다.
HouseStyle과 MSSubClass 간에 다중공선성이 크게 있던 것이 맞는 것 같다.
줄어들었지만 그래도 조금은 높은 편이며, KitchenQual의 경우도 VIF가 높은 편이다.
기준 VIF를 10으로 잡는다면 그래도 나쁘지 않은 경우이니 변수는 이대로 들고 간다.

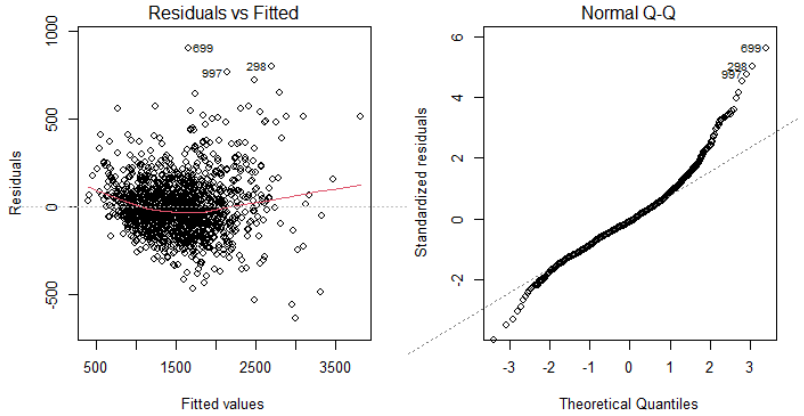
다음으로 stepwise 방법을 이용해 모형에 적합한 변수가 무엇인지 찾아보았다.

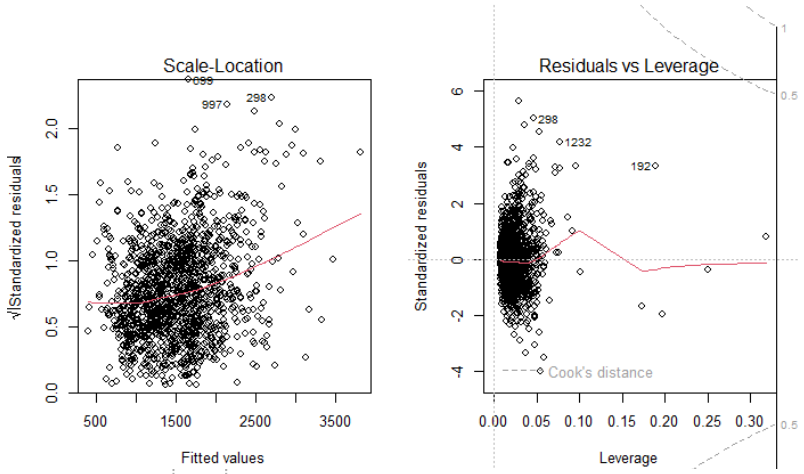
```
Residuals:
    Min      1Q  Median      3Q     Max
-638.30  -93.40  -16.05   79.45 1282.93

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.569e+03  5.421e+02   4.738 2.38e-06 ***
MSZoning2      2.974e+01  1.552e+01   1.916 0.055551 .
LotFrontage    6.843e-01  2.342e-01   2.921 0.003541 **
LotConfig2     2.817e+01  1.208e+01   2.333 0.019805 *
LandSlope      9.137e+01  2.101e+01   4.349 1.47e-05 ***
Neighborhood1 -1.870e+01  1.082e+01  -1.728 0.084152 .
Neighborhood3 -2.754e+01  1.903e+01  -1.447 0.148126
Condition1     2.845e+01  1.376e+01   2.067 0.038952 *
BldgType1     -3.148e+01  1.652e+01  -1.905 0.056967 .
HouseStyle1   -8.085e+01  2.046e+01  -3.951 8.19e-05 ***
HouseStyle2   -7.403e+01  2.633e+01  -2.811 0.005001 **
MasVnrType    -2.732e+01  1.355e+01  -2.016 0.043954 *
MasVnrArea     1.002e-01  3.757e-02   2.668 0.007711 **
Foundation1   -7.414e+01  2.037e+01  -3.640 0.000283 ***
Foundation2   -3.643e+01  1.807e+01  -2.016 0.043997 *
BsmtExposure  -2.344e+01  1.151e+01  -2.036 0.041967 *
BsmtUnfSF     -2.687e-02  1.541e-02  -1.743 0.081535 .
TotalBsmtSF    5.000e-01  2.313e-02  21.622  < 2e-16 ***
Electrical     3.241e+01  1.857e+01   1.746 0.081114 .
BsmtFullBath  -1.998e+01  1.219e+01  -1.640 0.101265
FullBath       9.654e+01  1.349e+01   7.157 1.34e-12 ***
HalfBath       4.366e+01  1.319e+01   3.309 0.000959 ***
BedroomAbvGr   2.038e+01  8.580e+00   2.376 0.017651 *
KitchenQual0   5.372e+01  2.431e+01   2.210 0.027301 *
KitchenQual1   7.500e+01  2.139e+01   3.506 0.000469 ***
TotRmsAbvGrd   9.747e+01  5.401e+00  18.046  < 2e-16 ***
Fireplaces     5.113e+01  8.447e+00   6.053 1.83e-09 ***
GarageYrBlt   -9.925e-01  2.773e-01  -3.579 0.000357 ***
GarageCars    -5.622e+01  1.414e+01  -3.976 7.36e-05 ***
GarageArea     2.513e-01  4.846e-02   5.186 2.47e-07 ***
WoodDeckSF     1.006e-01  3.907e-02   2.576 0.010108 *
PoolArea       5.150e-01  1.149e-01   4.483 7.97e-06 ***
SalePrice      1.269e-03  1.257e-04  10.091  < 2e-16 ***
Remod          5.225e+01  9.869e+00   5.294 1.39e-07 ***
Exterior_d     2.774e+01  1.328e+01   2.089 0.036848 *
X1stR         -8.904e+02  4.981e+01 -17.875  < 2e-16 ***
PorchSF        5.656e-01  1.854e-01   3.051 0.002327 **
CentralAir    -3.881e+01  2.224e+01  -1.745 0.081189 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 168.7 on 1382 degrees of freedom
Multiple R-squared:    0.9,     Adjusted R-squared:  0.8974
F-statistic: 336.3 on 37 and 1382 DF,  p-value: < 2.2e-16
```

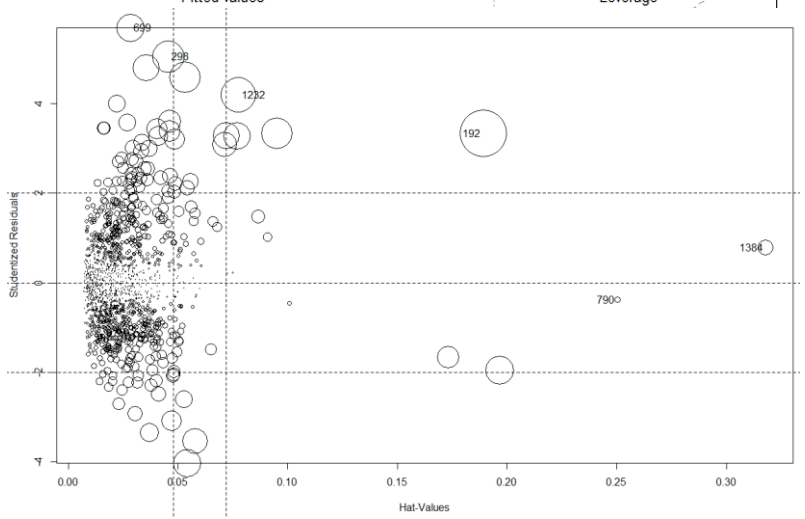
P-value가 매우 낮은 변수는
1층 면적 비율, 집의 가격,
전체 방의 수, 지하실 전체 면적이다.

이외에 땅의 경사도, 주택 형태, 건물 기초공사 자재, 화장실의 수, 벽난로 수, 차고, 수영장, 리모델링 변수가 낮게 나왔다.

여전히 변수가 많지만, 초기 모델에 비해서는 많이 줄어든 모습이다.








```
        Shapiro-Wilk normality test

data:  residuals(model_3)
W = 0.9574, p-value < 2.2e-16
```

```
> durbinWatsonTest(residuals(model_3))
[1] 1.962273
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 225.785, Df = 1, p = < 2.22e-16
```

Durbin-Watson 통계량이 2에 가까우므로 오차항은 독립성을 만족한다고 볼 수 있다.

그러나 정규성과 등분산성을 만족한다고 보기는 어렵다.

QQ-plot도 그렇고, 왼쪽 가장 아래 그림을 보면 이상치와 영향력이 큰 값들이 상당히 많다.

회귀분석의 기본가정들을 만족하지 않기에 본 모형은 적합한 모형이 아니다.

다른 대안으로 Lasso나 elasticnet을 생각하고 해보려 했으나 아직 숙달되지 않은 부분이어서 결과를 내지는 못했다.

그러나 단순히 p-value와 회귀계수만을 본다면

1층 면적 비율 , 집의 가격 , 전체 방 개수 , 지하실 전체 면적 정도가 종속변수인 거주면적에 영향을 줄 수 있을 것으로 보인다.

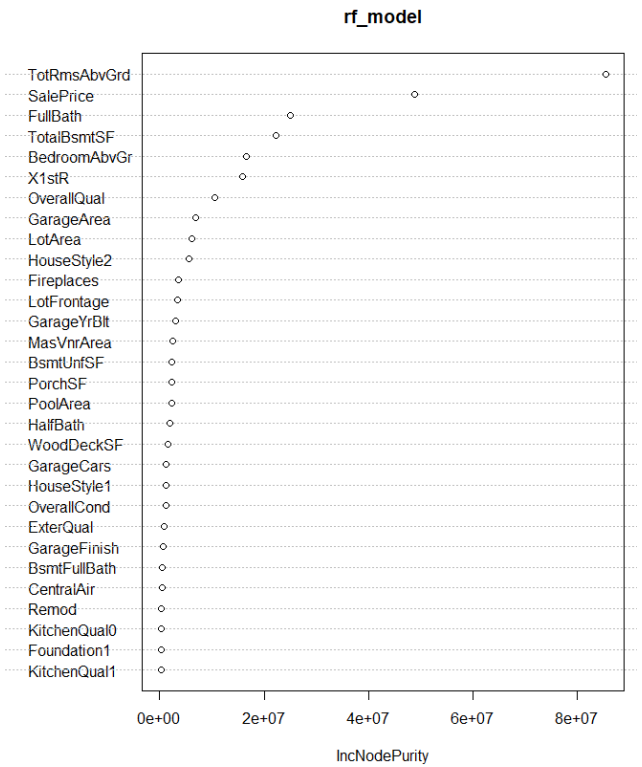
1층 면적 비율의 계수가 음수로 나타났다. 1층 면적 비율이 늘어날수록 2층 면적 비율이 줄어들게 되므로 1층 면적 비율이 높아질수록 거주면적이 작아진다는 것을 의미한다.

가격은 계수가 양수이므로 비싼 집일 수록 거주면적이 늘어난다.

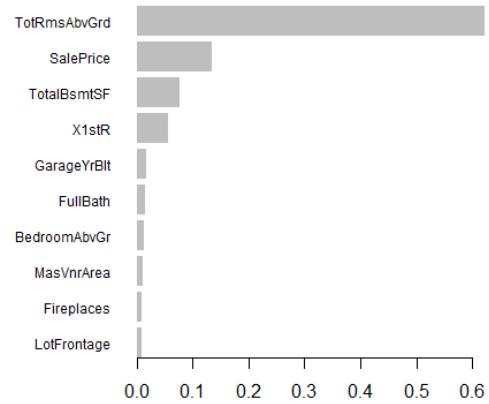
전체 방 개수의 계수도 양수이며 방이 많을 수록 거주면적이 늘어나고, 지하실 전체 면적 계수도 양수로 지하실이 크면 클수록 거주면적도 커진다.

이 밖에도 벽난로 수가 많을수록, 화장실 수가 많을수록 거주면적도 커지는 양의 관계를 가지고 있다.

<Random Forest / XGBoost>



<Random Forest feature importance>

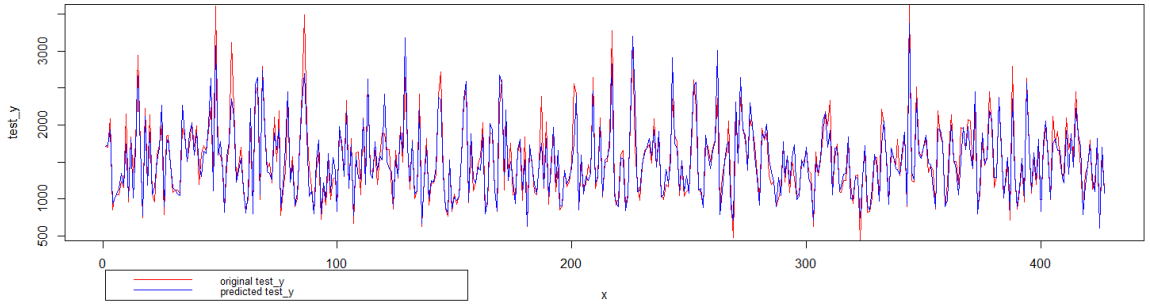


<XGBoost feature importance 상위 10>

```
Call:
 randomForest(formula = GrLivArea ~ ., data = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 19

          Mean of squared residuals: 36313.46
                    % Var explained: 87.23
```

```
> sqrt(36313.46)
[1] 190.5609
```

<Random Forest RMSE>



<XGBoost actual and predicted data Visualization>

좌측은 Random Forest의 feature importance 이고, 우측은 XGBoost의 feature importance 이다.

둘 다 가장 importance 가 큰 변수는 TotRmsAbvGrd 즉, 전체 방 개수(화장실 미포함)이다.

그리고 SalePrice(판매 가격)가 그 다음으로 두 모형 다 동일하다.

3번째부터는 순위가 조금씩 다르지만 두 그래프를 비교해보면 1층 면적 비율, Full Bathroom 수, 지하실 전체 면적, 침실의 수 등 상위에 랭크 되어 있는 변수들이 비슷비슷하다.

```
> caret::RMSE(test_y, pred_y) #RMSE
[1] 172.8196
```

<XGBoost RMSE>

<Conclusion>

회귀모형의 경우는 모형에 대한 기본 가정을 충족시키지 못하였기에 모형 그 자체로서는 적절치 못한 모형이지만, 단순히 종속변수와 독립변수와의 관계만을 바라보았을 때 의미 있어 보이는 변수가 몇 가지 있었다. (1층 면적 비율 / 집의 가격 / 전체 방 개수 / 지하실 전체 면적)

Random Forest의 경우는 기본적으로 세팅되어 있는 hyperparameter로 학습을 시켰다.
두 모형 다 전체 방 개수 / 집의 가격 / 1층 면적 비율 / 지하실 전체 면적이 상위에 랭크되었다.

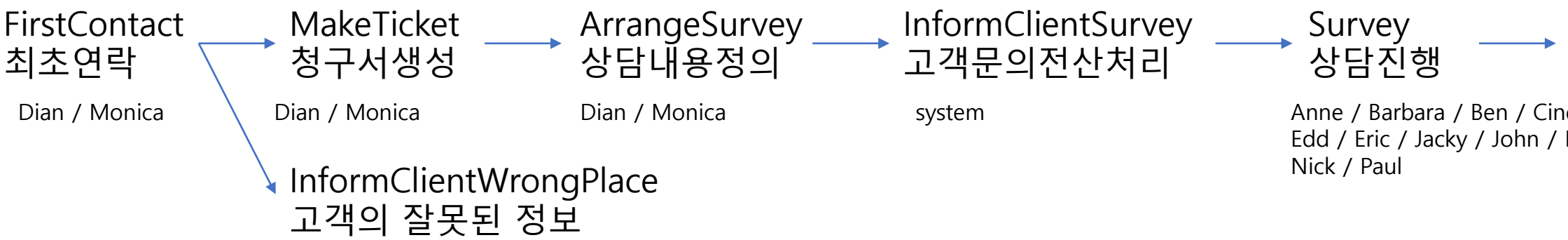
회귀모형과 Random Forest, XGBoost 모두 공통적으로 방의 개수, 집의 가격, 지하실 면적, 1층 면적 비율이 거주면적과 밀접한 변수라는 것을 가리키고 있다.

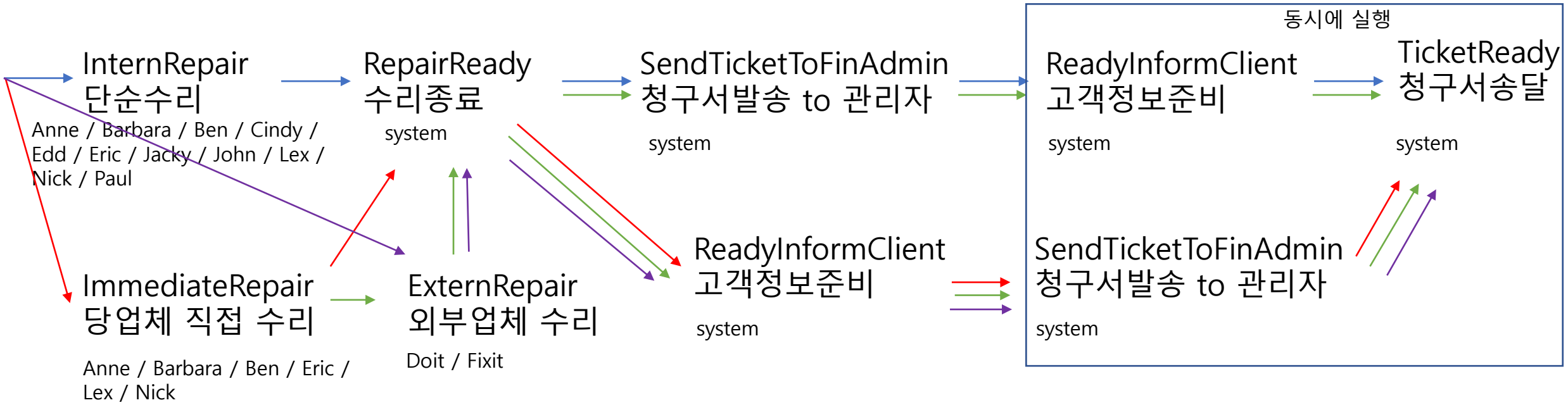
역시 집 전반적인 크기와 직접적으로 관련되어 있는 변수들이 뽑혔고, 이 다음으로 집이 크면 들어가는 옵션도 많으니 벽난로의 수도 거주면적과 관련되어 있다는 것을 알 수 있다.

특별한 결과가 나오지는 않았지만, 상식적으로 큰 집이, 비싼 집이 거주면적이 넓다고 생각했던 것들이 통계적으로도 맞는 말이라는 것을 어느 정도 증명한 것 같다.

아쉬웠던 점이 많은데, 그 중에서도 집의 전반적인 퀄리티, 세부적인 퀄리티 등 물리적 요소가 아닌 변수들이 모형에 안 맞는 것인지 전처리를 잘 못한 것인지 눈에 띄지 않았던 점이다. 회귀모형을 더 잘 적합시킬 수 있었을 것 같은 느낌도 든다. 또 전처리를 하면서 변수를 많이 솎아내지 못한 것도 아쉽다.

<업무 프로세스>


FirstContact
최초연락
Dian / Monica
MakeTicket
청구서생성
Dian / Monica
InformClientWrongPlace
고객의 잘못된 정보
ArrangeSurvey
상담내용정의
Dian / Monica
InformClientSurvey
고객문의전산처리
system
Survey
상담진행
Anne / Barbara / Ben / Cindy /
Edd / Eric / Jacky / John / Lex /
Nick / Paul


InternRepair
단순수리
Anne / Barbara / Ben / Cindy /
Edd / Eric / Jacky / John / Lex /
Nick / Paul
RepairReady
수리종료
system
SendTicketToFinAdmin
청구서발송 to 관리자
system
동시에 실행
ReadyInformClient
고객정보준비
system
TicketReady
청구서송달
system
ImmediateRepair
당업체 직접 수리
Anne / Barbara / Ben / Eric /
Lex / Nick
ExternRepair
외부업체 수리
Doit / Fixit
ReadyInformClient
고객정보준비
system
SendTicketToFinAdmin
청구서발송 to 관리자
system

데이터는 case별로 누가 어떤 업무를 몇 시에 어떻게 처리했는지 순차적으로 쌓인다.

따라서 원본 데이터인 source 데이터셋은 case – time 순서로 정렬되어 있다.

이 회사는 먼저 Letter/Personal/Phone/Web 을 통해 고객으로부터 상담요청을 받는다.

이후 고객의 정보를 파악하는데, 고객의 정보가 일치하면 다음 프로세스로 진행되고 일치하지 않으면 프로세스를 중단하게 된다.

그 다음으로는 상담을 진행하고, 단순 수리인지 업체 직접 수리인지 판단한다. 업체 직접 수리로 충분치 않으면 외부업체를 부르게 된다.

수리가 종료되면 관리자에게 청구서를 발송한 후 고객 정보와 함께 고객에게 청구서를 송달하거나
고객정보를 우선 준비하고, 관리자에게 청구서를 발송하면서 고객에게 청구서를 송달한다.

이 회사의 본 상담 진행 전 상담 요청을 받는 사람은 Dian / Monica 두 명이고,
본 상담은 Ann / Barbara / Ben / Cindy / Edd / Eric / Jacky / John / Lex / Nick / Paul 11명이다.

이후 수리를 진행할 때 단순 수리업무의 경우에는 본 상담을 맡았던 11명 인원 전부 가능하지만,
직접 수리업무의 경우에는 Anne / Barbara / Ben / Eric / Lex / Nick 6명만 가능하다.

여기서 의문이 드는 점은 단순 수리 업무의 경우에 상담 일시와 격차가 조금 나지만, 당 업체 직접 수리의 경우 상담 직후 진행된다.

개인적인 생각으로는 둘이 바뀐 것이 아닌가 싶다. 단순 수리의 경우 유선 상으로 고객에게 설명하며 고객이 직접 고칠 수 있게 유도할 수 있지만, 당 업체 직접 수리의 경우는 수리 기사가 고객이 원하는 장소로 방문하여 고치는 것이 맞지 않나 싶다.

두번째로는 단순 수리 업무를 나타내는 internRepair가 같은 case에 2번 나타나는 경우도 있다. 다른 직원으로 교체 후 수리하는 경우와 동일한 직원이 2번 나타나는 경우이다.

세번째로는 고객으로부터 상담요청을 받고 고객 정보를 확인하는데 여기서 정보가 일치하지 않으면 프로세스를 중단한다는 점이다. 고객 정보를 어떻게 수집하는지는 이 데이터로부터 알 수 없지만 고객이 직접 정보를 입력하는 것이 보편적인 것이라 생각된다. 그런 점에서 고객이 실수로 정보를 잘못 적었을 경우 이에 대한 시정을 요청하는 시스템이 없는 듯 하다.

네번째로는 수리를 하는데 있어 외부 업체를 이용하는 경우이다. 외부 업체 수리를 하게 되는 경우는 바로 외부 업체를 부르기도 하지만 당업체 직접 수리 후 수리가 힘든 상황이면 외부 업체를 부르는 것 같은데, 상담 단계에서 확실하게 문제를 파악해 처음부터 외부 업체를 통해 해결하는 방식이 더 효율적이지 않을까 싶다.

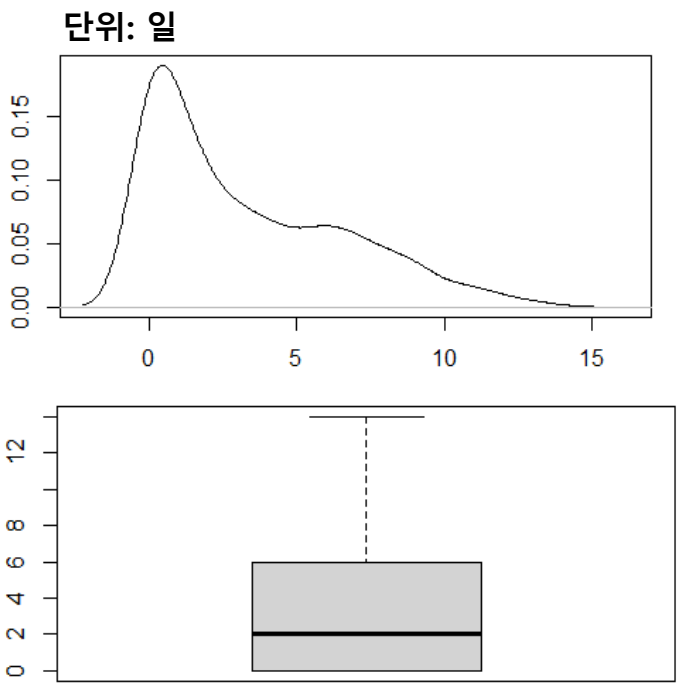
다섯번째로는 단순 수리와 직접 수리 파트를 나누는 것이 좋을 것 같다. 일부 인력은 단순 수리만 가능하여, 상담 이후 단순 수리가 아닌 경우에는 다른 직원으로 업무를 넘기게 된다. 이 과정이 비효율적인 것 같다. 최초 연락 과정에서 판단을 하여 수리 수준이 맞는 직원에게 바로 업무가 갈 수 있도록 분업하는 것이 좋아 보인다. (소규모 업체인 경우 인력이 얼마 없기에 분업하지 않았을 수도 있다고 생각한다.)

마지막으로 수리 종료 이후 프로세스가 확실치 않은 것 같다. 어떤 case는 관리자에게 청구서를 발송하여 결제를 받은 이후 고객에게 최종 청구서를 송달하게 되고, 또다른 case는 관리자와 고객에게 동시에 청구서를 송달하게 된다.

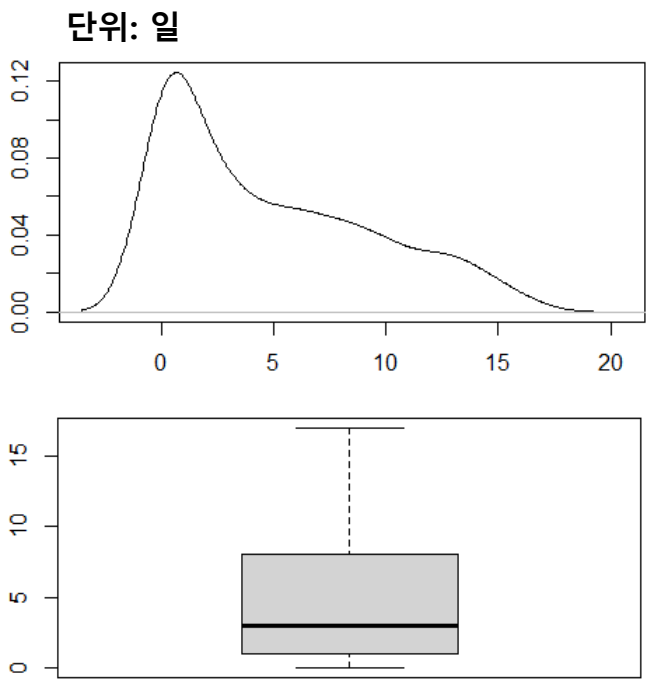
원본 데이터에서 target 데이터로 넘어가면서 objectKey 변수가 사라졌다. 이 변수는 RepairType 변수와 pair 되어 있다.

그리고 수리 직원이 2명 이상인 경우에 target 데이터에는 수리 시작 시점과 끝 시점만 표기되기에 언제 교체 되었는지 알 수 없다.

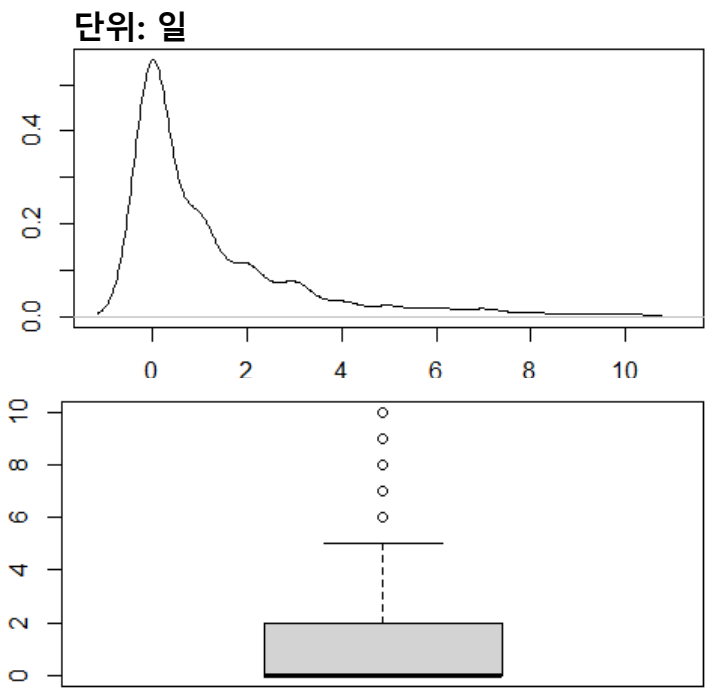
또한 원본 데이터에는 RepairReady로 수리가 완전히 끝난 시점이 표기되는데 target 데이터에는 외부업체를 쓴 경우 외부 업체가 수리를 시작한 시점밖에 나와 있지 않아 전체 수리 시간을 알 수 없다.



FirstContact 이후 Survey 까지 걸리는 시간 분포



FirstContact 이후 InternRepair 까지 걸리는 시간 분포

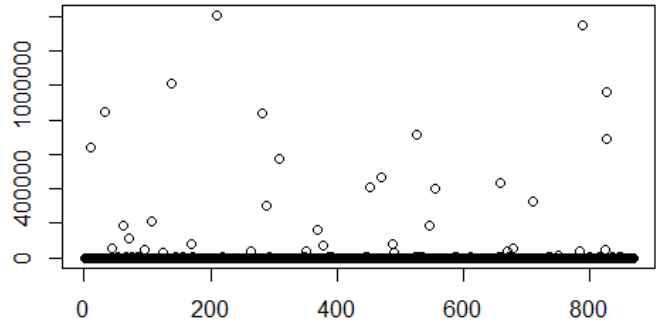


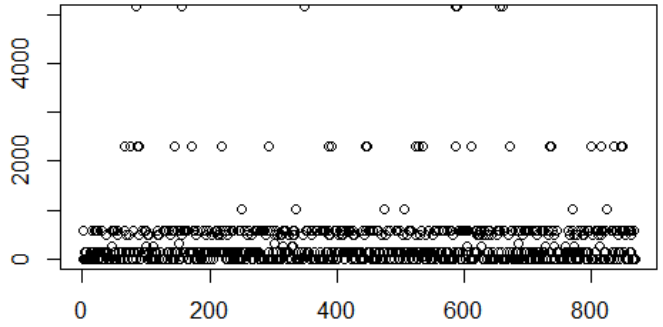
Survey 이후 InternRepair 까지 걸리는 시간 분포

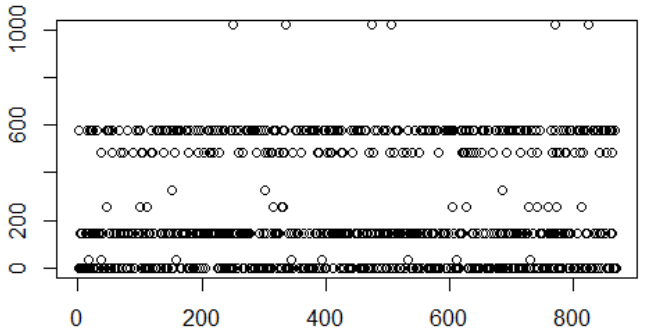
서비스업 치고는 고객의 첫 컴플레인 이후 수리 완료 시점까지 느린 편인 것 같다.

FirstContact 이후 Survey까지는 최대 14일, InternRepair까지는 최대 17일이 걸린다. 두 분포가 비슷한 것을 보아 첫 컴플레인 이후 수리 완료 시점까지 걸리는 시간에 가장 영향을 많이 주는 요소는 FirstContact 이후 Survey까지 시간인 것 같다. 이 시간을 단축해야 고객 만족도가 높아질 것으로 보인다.

따라서 FirstContact 이후 Survey까지 프로세스가 빠르게 이루어지도록 하는 것이 좋을 것 같다. Survey와 Repair하는 인원이 동일하기 때문에 인력이 부족하여 그런 것으로 생각된다.







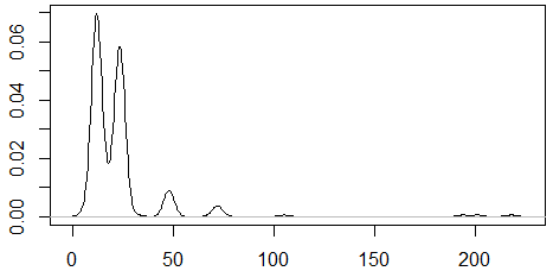
좌측 plot은 고객에게 말한 예상 수리 시간과 실제 수리 시간의 차이 제곱을 나타낸 plot이다.
(외부 업체 수리의 경우 제외)

가장 아래는 차이 제곱이 0~1000 사이이고 중간은 0~5000, 맨 위는 0~max 이다.

868개의 관측치 중 266개가 0이다. 예상 시간과 실제 수리 시간이 동일한 경우는 266건이라는 것.

그리고 예상 시간보다 빠르게 끝난 경우는 315건이다.

나머지 287건은 예상시간보다 늦게 끝났으며 분포는 오른쪽과 같다.(단위: 분)



대부분은 초과 1시간 이내에 수리가 끝났다.

예상 시간보다 수리 시간이 더 걸릴 확률은 무려 33%나 된다.

예상 시간보다 빠르게 끝나는 것은 크게 문제 되지 않지만 느리게 끝나는 빈도가 많다면 문제가 있다고 생각한다.

예상 시간을 잘 예측할 수 있게 도와주는 모델이 필요한 것 같다.