주제분석 3주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 주제분석 3주차 패키지 문제의 조건 및 힌트는 R을 기준으로 하지만, R을 사용해도 무방합니다.
- 제출형식은 HTML, PDF 모두 가능합니다. **.ipynb 이나 .R 등의 소스코드 파일은 불가능합니다.** 파일은 psat2009@naver.com으로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는 시간 후에 하게 되며, 랜덤으로 4시 00분에 발표됩니다.
- 제출기한은 **5/11 자정까지** 이고, 지각 시 벌금 5000원, 미제출시 10000원입니다. 패키지 무단 미제출 2회 시 퇴출이니 유의해주세요.

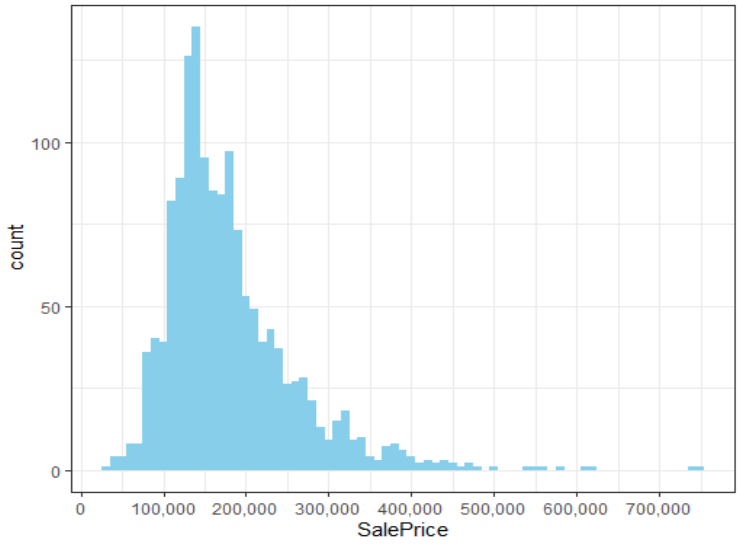
Chapter 1 EDA & 결측치 보간

이번 데이터는 집값을 예측하는 데이터로 각 변수에 대한 추가 설명은 'data_description.txt' 파일에 있습니다. 데이터에는 많은 수의 변수가 있고, 변수 유형 또한 다양해 보입니다. 이번 패키지에서 전체적인 분석을 진행하긴 하지만, 어떻게 변수를 처리하고, 파생변수들을 생성해볼지에 집중하여 진행하겠습니다. 또한 이번 데이터에서는 집값에 영향을 주는 요소들을 직관적으로 생각해볼 수 있기 때문에 관계를 생각하며 진행해주세요.

문제 1. 'train.csv' 파일을 불러온 후 데이터 구조를 파악해주세요.

문제 2. Test 데이터의 'ID' 칼럼만 따로 변수에 저장하고 train, test 데이터에서 'ID' 칼럼을 삭제해주세요.

문제 3. train data의 'SalesPrice' 분포를 시각화해주세요. 색은 'skyblue', 테마는 'theme_bw'를 사용했습니다.



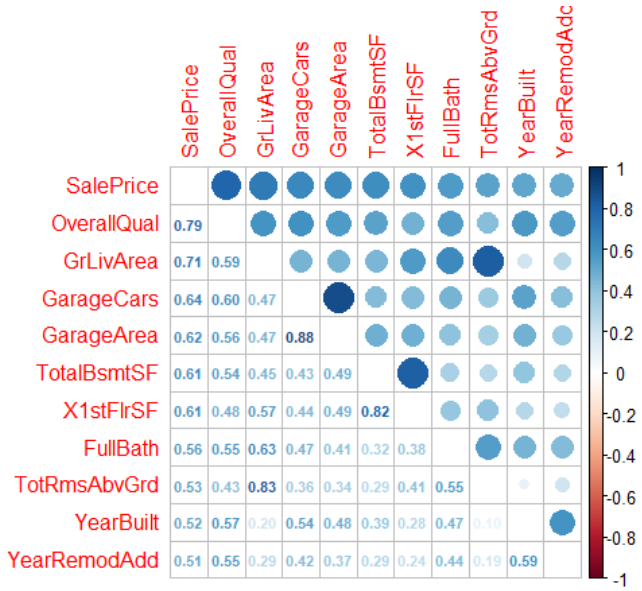
문제 4. 'SalePrice'와 numeric변수들의 상관관계를 알아보겠습니다.

문제 4-1. numeric 변수들만 골라 'SalePrice'와의 상관관계 절댓값을 구해주세요.

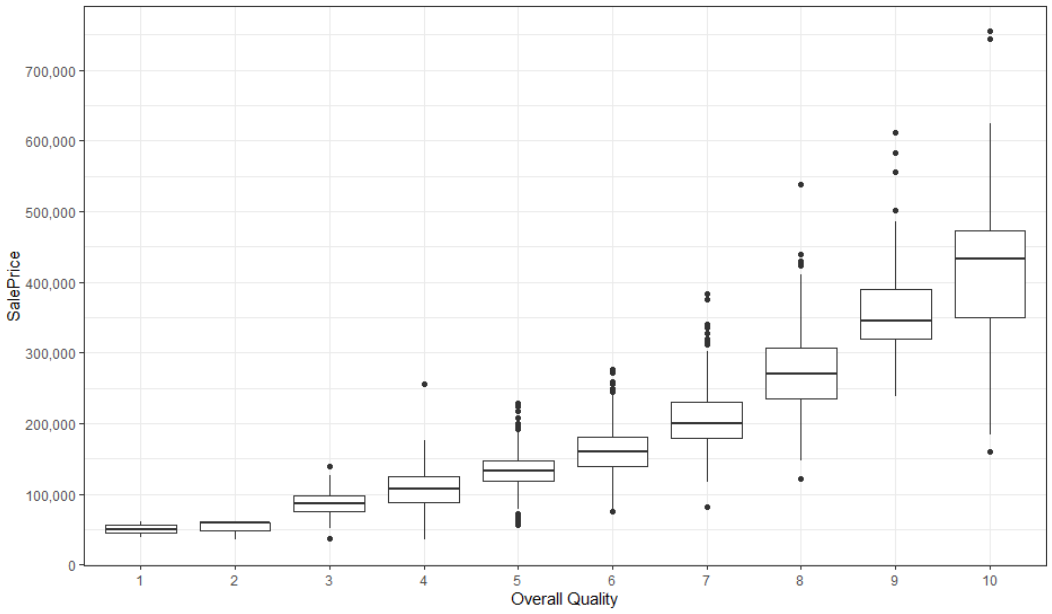
(이때, 추후에 사용하기 위해 numeric 변수들의 변수명을 담은 'numericVarNames' 변수를 따로 저장해주세요.)

문제 4-2. 내림차순으로 정렬 후 값이 0.5 이상인 것들만 선택해주세요.

문제 4-3. 다음과 같이 시각화해주세요.



문제 5. 'OverallQual' 변수와 'SalePrice'와의 관계를 다음과 같이 시각화하여 어떤 관계가 있는지 알아보겠습니다.



문제 6. 결측치 대체와 데이터 유형 변환을 해보겠습니다. (문제에서는 유형 변환을 바로 진행했지만, 해당 변수가 순서형인지, 아닌지를 직관적으로 알 수 없는 변수는 개별 변수들의 상관관계나 EDA를 통해 논리적으로 판단하여 설정해야 합니다.)

문제 6-1. 결측치가 각 변수에 몇 개씩 존재하는지 확인해주세요.

문제 6-2. 'PoolQC', 'FireplaceQU', 'GarageType ', 'GarageFinish', 'GarageQual', 'GarageCond', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2', 'MasVnrType' 의 세부사항에 따르면 NA값은 해당 항목이 집에 없다는 뜻이며 모두 순서형 변수입니다. 결측치가 있다면 'None'으로 바꾼 후 각 칼럼의 등급별로 숫자로 변환하고 정수형으로 설정해주세요.

Example. PoolQC

```
Ex   Excellent            <- 5
Gd   Good                 <- 4
TA   Average/Typical      <- 3
Fa   Fair                 <- 2
NA   No Pool              <- 1
```

문제 6-3. 'MiscFeature', 'Alley', 'Fence' 의 결측치를 'None'으로 바꾼 후 factor로 변환해주세요.

문제 6-4. 'LotFrontage'의 결측치는 'Neighborhood' 별 중위값으로 대체하여 정수형으로 설정해주세요.

문제 6-5. 'GarageYtBlt' 결측치는 같은 행의 'YearBuilt' 값으로 대체해주세요.

문제 6-6. 'MasVnrArea' 결측치는 0 으로 대체해주세요.

문제 6-7. 'Electrical' 결측치는 최빈도값으로 대체해주세요.

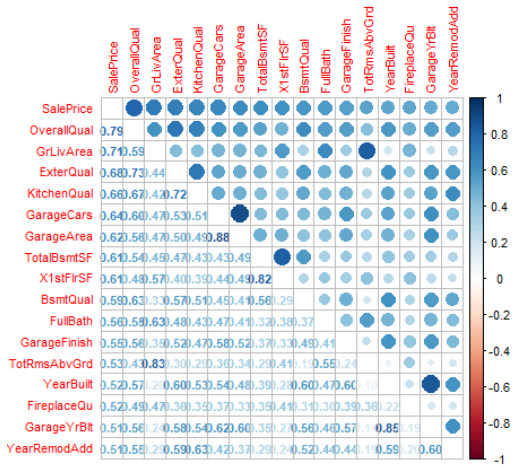
문제 7. 결측치가 없는 변수도 확인해보겠습니다

문제 7-1. 'Utilities'변수는 삭제해주세요. 모든 값이 일치합니다.

문제 7-2. 'Exterior2nd', 'Exterior1st', 'MSZoning', 'KitchenQual', 'Foundation', 'Heating', 'RoofStyle', 'RoofMatl', 'LandContour', 'BldgType', 'HouseStyle', 'Neighborhood', 'Condition1', 'Condition2', 'SaleType', 'SaleCondition', 'LotConfig', 'MSSubClass', 'MoSold' 은 순서형이 아니기 때문에 factor로 변환해주세요.

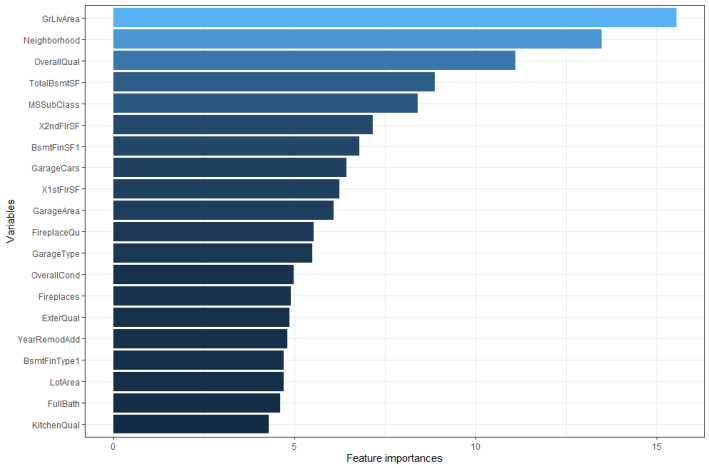
문제 7-3. 'LotShape', 'ExterCond', 'ExterQual', 'Functional', 'HeatingQC', 'CentralAir', 'LandSlope', 'Street', 'PavedDrive', 'KitchenQual' 는 순서형 변수이기 때문에 각 칼럼의 등급별로 숫자로 변환하고 정수형으로 설정해주세요.

문제 8. '문제 4' 를 다시 진행하여 다음과 같은 numeric변수들의 상관관계 plot을 그려주세요.

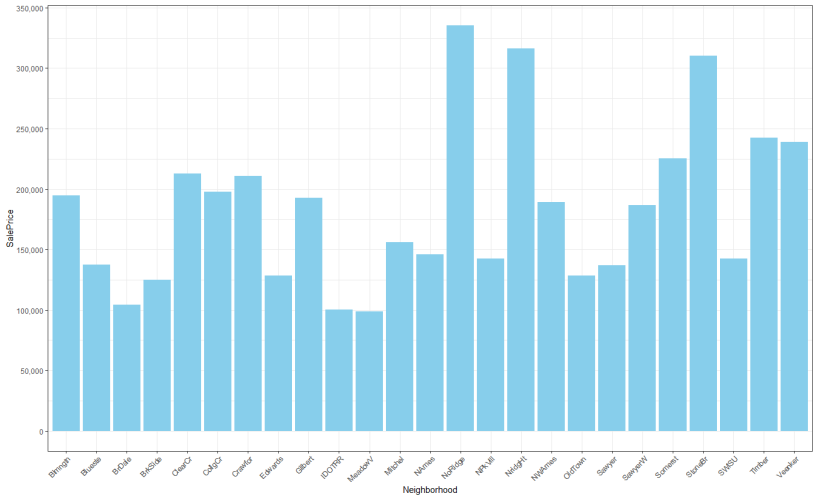


문제 9. randomforest를 통해 중요변수를 알아보겠습니다.(ntree = 100, importance = TRUE 로 설정)

문제 9-1. 중요도 상위 20개를 시각화해주세요



문제 9-2. 중요도 순위가 높은 변수들 중에서 예시로 'Neighborhood' 변수와 'SalePrice' 시각화를 통해 유의한 차이가 있는지 확인해보겠습니다.



Chapter 2 FE

이번 데이터에는 다양한 변수들이 있기 때문에 여러 시도를 통해 파생 변수를 생성해볼 수 있습니다.
파생변수를 만든 후 y 변수와 유의미한 관계가 존재하는지 확인해보겠습니다.

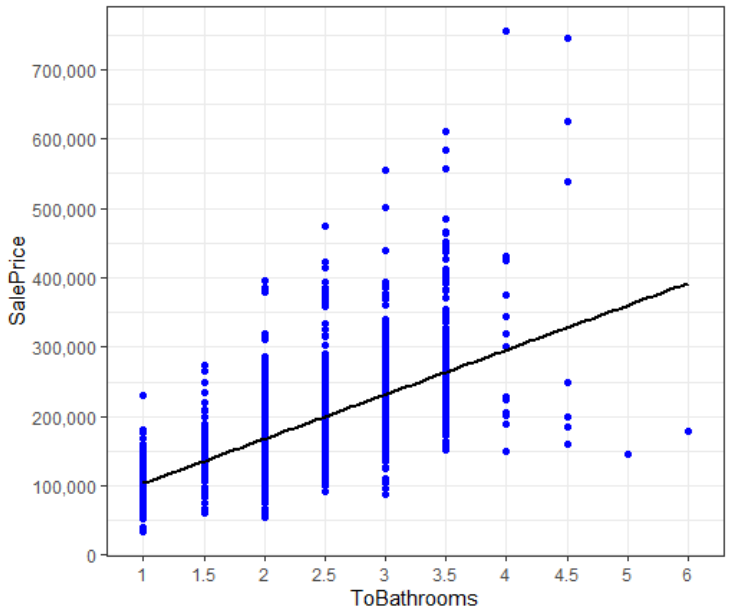
문제1. Bath 관련 4가지 개별적 변수는 큰 영향이 없어 보이지만 그 변수들을 사용하여 유의한 영향을 주는 새로운 변수를 만들어 보겠습니다.

문제1-1. 총 Bathroom의 개수 정보를 담은 'TotBathrooms' 변수를 추가해주세요.

HINT1. fullbath는 1, halfbath는 0.5를 곱하여 총 합산 값을 구하면 됩니다.

문제1-2. 새로 만든 'TotBathrooms' 변수와 'SalePrice'와의 관계를 시각화해주세요.

HINT1. 추세선은 geom_smooth()를 사용합니다.(method= 'lm')

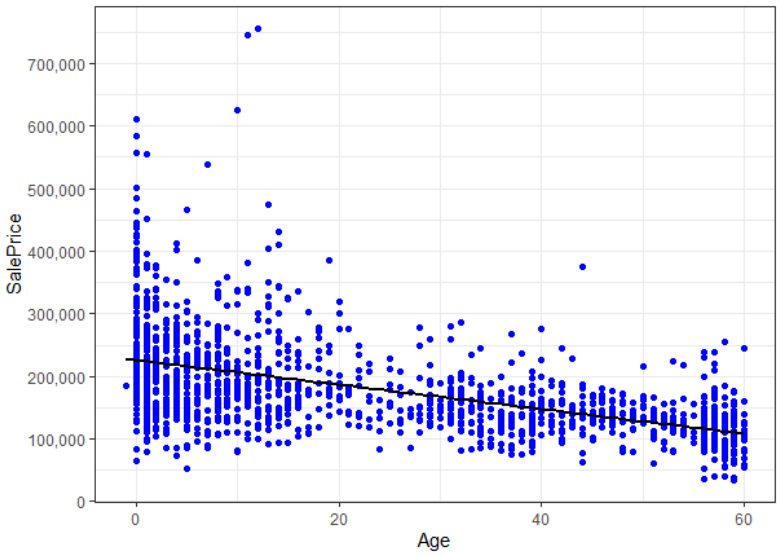


문제2. Year과 관련된 3개의 변수로 새로운 변수를 만들어 보겠습니다.

문제2-1. 'YearBuilt'와 'YearRemodAdd'가 같다면 0(리모델링을 하지 않음), 다르면 1(리모델링 함) 값을 갖는 'Remod' 변수를 추가해주세요.

문제2-2. 'YrSold'에서 'YearRemodAdd'을 뺀 값으로 'Age' 변수를 추가해주세요.

문제2-3. 'SalePrice'와 'Age'의 관계를 시각화해주세요.

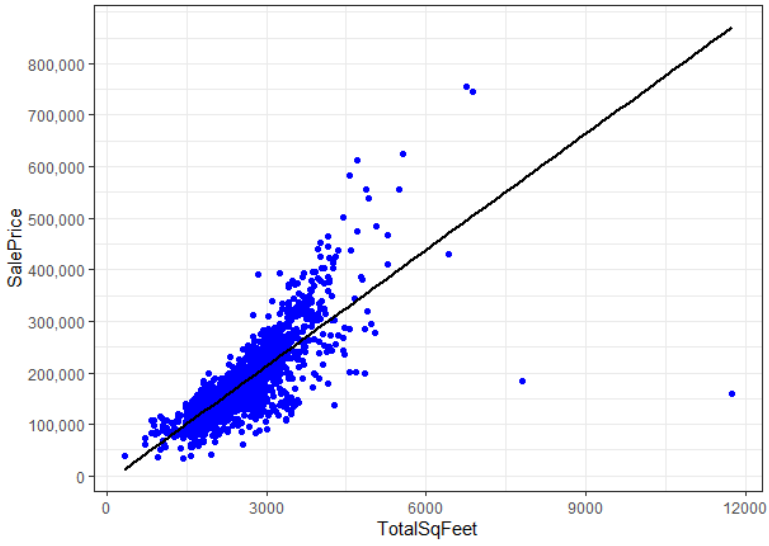


문제2-4. 'YrSold' 와 'YearBuilt'가 같으면 1, 다르면 0의 값을 갖는 'IsNew' 변수를 추가해주세요.

문제2-5. 'YrSold' 변수로 더 이상 연산을 하지 않기 때문에 'YrSold'를 factor로 변환해주세요.

문제3. 지상과 지하 공간을 모두 합쳐 총 공간 변수를 만들겠습니다. 'GrLivArea'와 'TotalBsmtSF'를 더한 'TotalSqFeet' 변수를 추가해주세요.

문제3-1. 'TotalSqFeet'와 'SalePrice' 관계를 시각화해주세요.



문제4. 파생변수 생성이 끝난 후 필요 없는 변수는 삭제해야 합니다. 서로 상관 관계가 높은 두 변수 중 'SalePrice'와 상관 계수가 더 낮은 변수인 'YearRemodAdd', 'GarageYrBlt', 'GarageArea', 'GarageCond', 'TotalBsmtSF', 'TotalRmsAbvGrd', 'BsmtFinSF1' 변수들을 삭제해주세요.

Chapter 3 Modeling

모델링을 하여 예측을 해보겠습니다. 다양한 모델을 시도해 본 후, 파라미터 튜닝, 앙상블 등 추가적인 성능 향상을 위한 방법이 있지만 이번에는 간단한 모델링까지만 해보겠습니다. 그 전에, 모델링을 하기 위해 필요한 변수처리 또한 진행해보겠습니다.

문제1. 수치형 독립 변수를 조정하고 범주형 변수를 더미변수로 변환하기 위해 데이터프레임을 각각 분할하겠습니다.

문제1-1. 챕터1 문제 4-1에서 정의한 numericVarNames 에서 'MSSubClass', 'MoSold', 'YrSold', 'SalePrice', 'OverallQual', 'OverallCond' 를 빼고 'Age', 'TotBathrooms', 'TotalSqFeet' 을 더해주세요.

문제1-2. train 데이터에서 numericVarNames에 존재하는 칼럼들만 따로 다른 numericDF 데이터프레임으로 저장해주세요.

문제1-3. train 데이터에서 numericVarNames에 존재하지 않는 칼럼들만 따로 다른 factorDF 데이터프레임으로 저장해주세요.

문제1-4. factorDF에서 'SalesPrice'를 제거해주세요.

문제2. numericDF에서 왜곡이 큰 변수들을 조정해보겠습니다.

문제2-1. numericDF의 칼럼들의 왜도 절대값을 구하여 0.8 이상인 칼럼들은 로그 변환을 해주세요.

HINT1. abs(), skew() 함수를 사용하여 왜도를 구할 수 있습니다.

HINT2. log1p() 함수를 사용하여 변환을 할 수 있습니다.

문제3. numericDF 변수들 간에 range의 차이가 과도하게 큰 변수들이 많아 스케일링을 진행하겠습니다. caret 패키지의 preProcess() 함수를 사용하여 스케일링을 진행해주세요.

HINT1. method 는 'center', 'scale'로 설정하여 standardization 방식으로 진행해주세요.

문제4. factorDF에 model.matrix()를 사용하여 one-hot encoding 을 진행해주세요.

문제4-1. 칼럼별(더미변수) 합계를 구하여 10보다 작은 칼럼은 삭제해주세요.

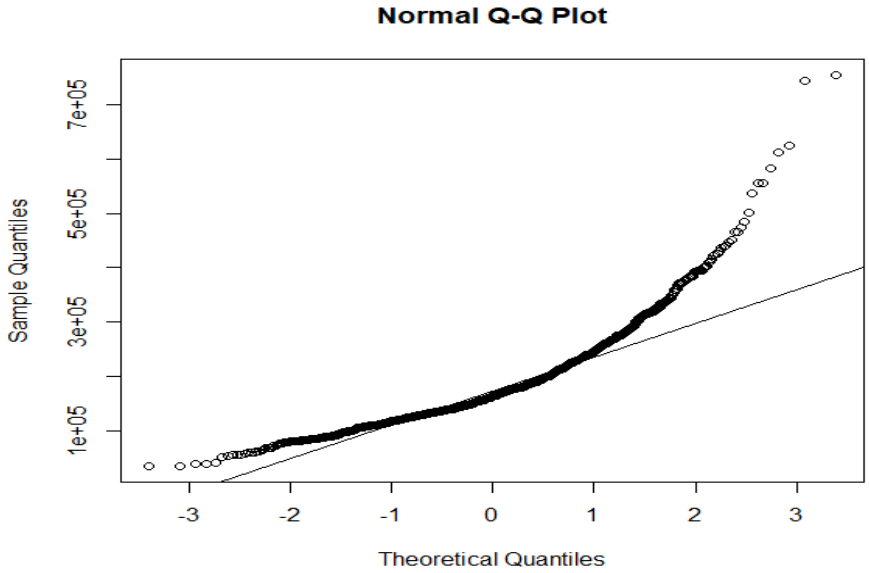
문제5. 두 데이터프레임(문제3, 문제 4)을 합쳐주세요.

문제6. 'SalePrice'의 왜도를 조정해보겠습니다.

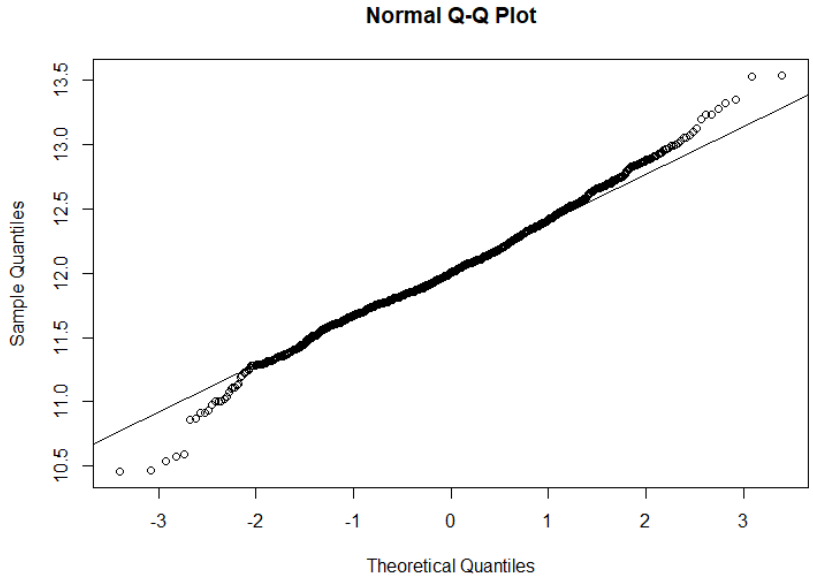
문제6-1. train 데이터에서 'SalePrice'의 왜도를 확인해주세요.

문제6-2. 'Saleprice'의 normal quantile-quantile plot을 그려주세요.

HINT1. qqnorm, qqline 함수를 사용하시면 됩니다.



문제6-3. 'Saleprice'에 로그를 취한 후 왜도와 qq-plot을 비교해주세요.



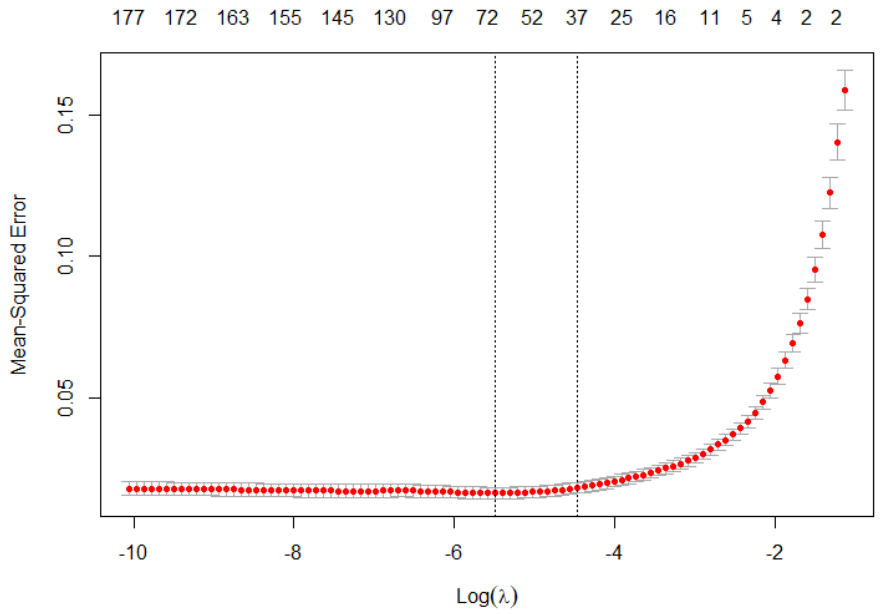
문제6-4. 조정한 'Saleprice'를 label 변수에 저장해주세요.

문제7. Lasso regression 모델을 학습시키겠습니다.

문제7-1. 적절한 람다값을 찾기 위해 교차검증을 진행해주세요.

HINT1. cv.glmnet() 을 사용하시면 됩니다.

문제7-2. 람다값에 따라 MSE가 어떻게 변화하는지 확인해보세요.



문제7-3. lambda.min을 통해 최적의 람다값이 무엇인지 확인해주세요.

문제7-4. coef() 함수를 사용하여 어떤 변수들을 사용했는지 구체적인 회귀계수들의 값을 확인할 수 있습니다.

HINT1. 문제7-3에서 구한 최적의 람다값으로 설정해주세요.