*kc_house_data*는 미3국 king county의 집값과 그 집의 특성에 대한 데이터이다. 이 데이터로 집값을 예측하는 회귀 분석 모델을 만들 수 있다. 데이터 전처리를 한 후 집값을 예상할 수 있는 변수들은 다음과 같다.

Bedrooms: 방의 개수

Bathrooms: 화장실의 개수

Floors: 층수

Waterfront: 우리나라의 한강뷰와 같이 강이나 바다가 보이는지 여부를 나타내는 변수

View: 집을 구하는 사람이 조회한 기록이 있으면 1을 나타냄

Condition: 전체적인 방의 상태

Grade: king county에 의해 산정된 점수

Yr_built: 집이 지어진 연도

Zipcode: 우편번호

Lat: 위도

Long: 경도

Year/month/day: 각각 집이 팔린 연도, 달, 날짜

Log_price: 집이 팔린 가격

log_sqft_living, log_sqft_lot, log_sqft_above: 면적

log_sqft_living15, log_sqft_lot15: 주변 15개 집의 평균 면적

Renovated: 리모델링의 여부

Basement: 지하실의 여부

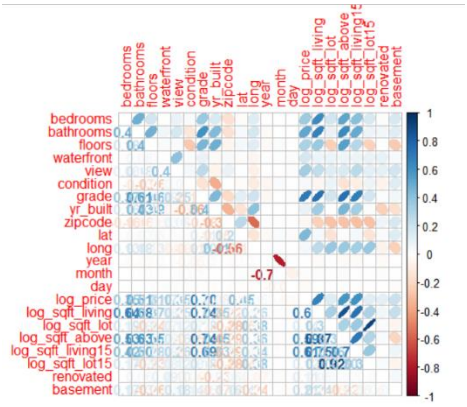
이러한 변수들을 가지고 log_price를 예측하려고 하는데 변수가 너무 많다. [그림1]을 보면 변수들 사이의 correlation이 존재한다. 특히 집의 넓이를 나타내는 변수들끼리의 상관관계가 높다. 따라서 FA를 통해 변수의 개수를 줄이고자 한다.

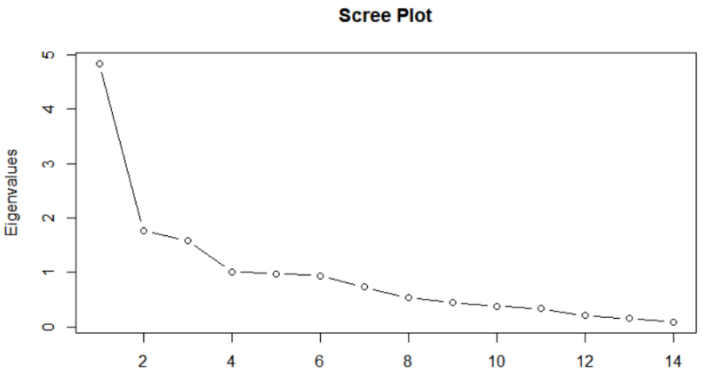
FA는 범주형 변수에 대해서는 적용할 수 없으니 범주형 변수를 제외한 데이터로 FA를 한다. 먼저 요인의 개수를 결정한다. [그림2]는 eigenvalues가 기록되어 있는 scree plot이다. 이 scree plot을 보면 elbow 지점이 4이다. 따라서 4개의 요인으로 변수를 설명하기로 한다. 요인의 개수를 정한 이후 본격적으로 FA를 진행한다. 이때 모두 표준화한 변수들을 사용한다. [그림3]은 maximum likelihood method로 FA를 한 결과이다. 그런데 Factor4를 보면 절댓값이 0.5가 넘는

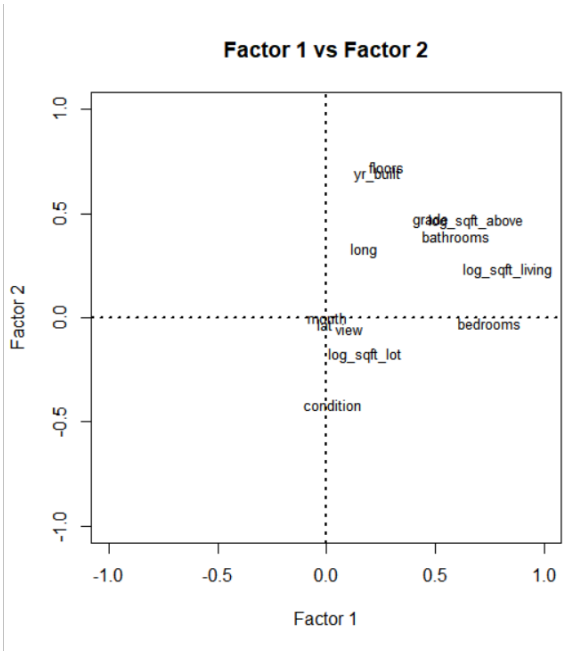
factor loading가 없다. 따라서 scree plot에서 elbow 지점이 4였지만 3개의 요인으로 FA를 다시 한다. 그 결과는 [그림4]이다. Factor1은 log_sqft_living, bedrooms, bathrooms, log_sqft_above가 1에 가까운 값이 나온다. 즉 factor1은 집의 넓이라고 할 수 있다. Factor2는 floors와 yr_built가 1에 가까운 값을 가진다. 따라서 factor2는 집이 지어진 연도와 층수와 관련된 조건이라고 할 수 있다. Condition이 음의 값을 가지는 것은 이와 맥락이 맞는다. 집이 오래전에 지어졌을수록 condition이 나쁠 확률이 높기 때문이다. 마지막으로 factor3은 long와 log_sqft_lot이 비교적 큰 값을 갖는다. 따라서 factor3은 경도와 차고의 면적과 관련된 조건이라고 할 수 있다.

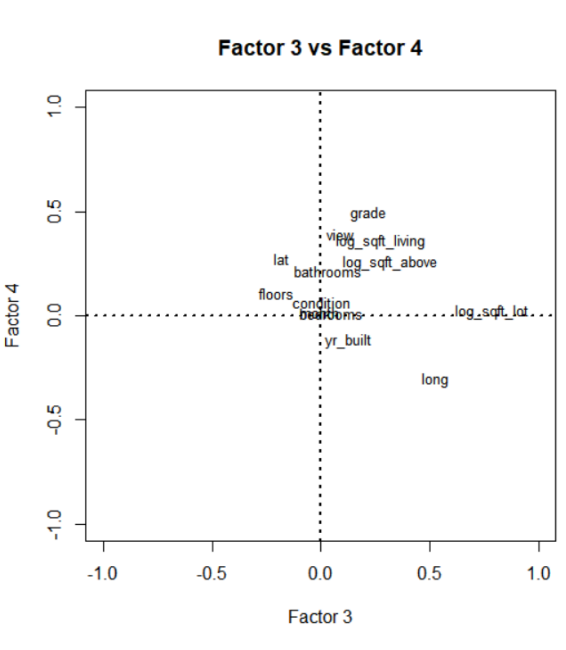
한편 communality를 보면 view, condition, lat, month 변수에 대해서 communality가 매우 작게 나온다. Communality란 공통적인 요인으로 분해되는 정도를 나타내는 지표이기 때문에 이 세 변수를 빼고 해석하는 것이 좋을 것 같다.

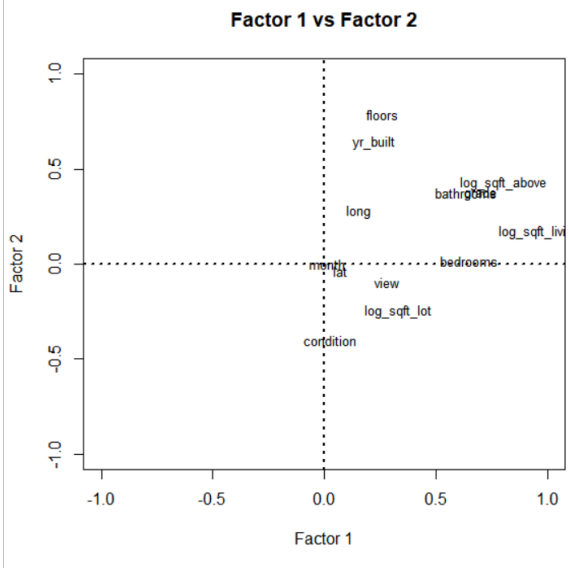
또한 이 변수들로 집값을 잘 예측할 수 있는지 확인하기 위해서 LDA를 했다. Log_price의 평균을 기준점으로 한다. log_price보다 크면 비싼 집이라고 하여 1이고, log_price보다 작으면 싼 집이라고 하여 0인 변수(*high*)를 만든다. 그리고 변수들로 비싼 집인지 아닌지를 판별할 수 있는지 LDA(판별분석)을 하였다. 판별분석이란 우리가 관심 있는 개체가 어떤 집단에 속하는지 구분하는 함수에 대한 분석 기법으로 선형판별함수와 비선형판별함수가 있다. 여기서는 선형판별함수를 사용한다. 선형판별함수는 비선형판별함수와는 달리 첫 번째 그룹의 공분산 행렬과 두 번째 그룹의 공분산 행렬이 같다고 가정한다. 지금 변수들이 집값을 잘 예측할 수 있는지 대략적으로 확인해보기 위해서 LDA를 하는 것이므로 하나의 데이터에서 임의의 기준점을 정해서 두 개의 데이터로 나누었다. 하나의 데이터에서 분리하였으므로 두 그룹의 공분산 행렬이 같다고 가정하는 것이 좋을 것 같다. 첫 번째 그룹의 공분산 행렬과 두 번째 그룹의 공분산 행렬을 통해서 $\boldsymbol{W}$ 행렬을 구한다. 그리고 이 $\boldsymbol{W}$ 행렬을 통해서 집단 간 분산과 집단 내 분산의 비율을 최대화하는 벡터 $\boldsymbol{a}$를 구한다. 그렇다면 이제 판별함수를 만들 수 있다. [그림5]는 판별분석의 결과를 나타내는 confusion matrix이다. 집값이 비싼데 싸다고 판별한 경우와 집값이 싼데 비싸다고 판별한 경우의 수는 비슷하다. 그리고 misclassification 비율은 0.158 정도가 나온다. [그림6]은 집값이 비싼 경우의 관측치와 싼 경우의 관측치를 projection했을 때 그래프이다. 두 그룹의 평균은 서로 떨어져 있으며 겹치는 부분이 많지 않다. Error rate도 약 15%로 높지 않기 때문에 위 변수들은 집값이 높고 낮음을 예측하는 데 적절한 변수들이라고 결론을 내릴 수 있다.

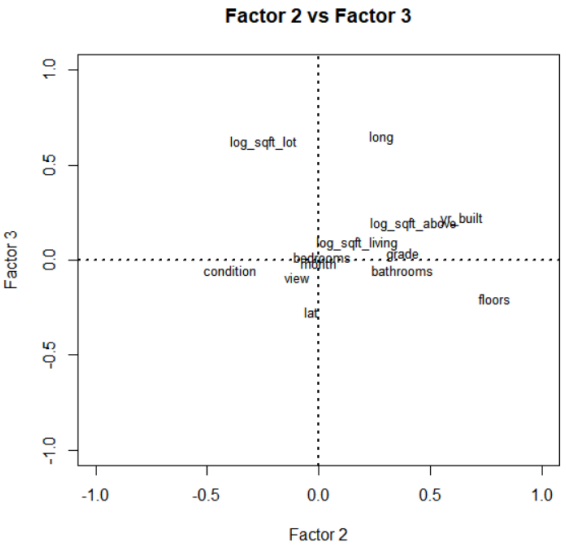

bedrooms
bathrooms
floors
waterfront
view
condition
grade
yr_built
zipcode
lat
long
year
month
day
log_price
log_sqft_living
log_sqft_lot
log_sqft_above
log_sqft_living15
log_sqft_lot15
renovated
basement


Scree Plot
Eigenvalues


Factor 1 vs Factor 2
Factor 1
Factor 2
yr_built
floors
grade
log_sqft_above
bathrooms
long
log_sqft_living
month
lat
view
bedrooms
log_sqft_lot
condition


Factor 3 vs Factor 4
Factor 3
Factor 4
grade
view
log_sqft_living
lat
log_sqft_above
bathrooms
floors
condition
log_sqft_lot
yr_built
long


Factor 1 vs Factor 2
Factor 1
Factor 2
floors
yr_built
log_sqft_above
long
log_sqft_living
month
lat
view
bedrooms
log_sqft_lot
condition


Factor 2 vs Factor 3
Factor 2
Factor 3
log_sqft_lot
long
log_sqft_above
yr_built
log_sqft_living
condition
view
bathrooms
floors
lat

	Group1	Group2
Predicted Group1	8604	1674
Predicted Group2	1742	9592