- 분석 툴은 R입니다. 문제의 조건 및 힌트는 R을 기준으로 하지만, 아직 R에 익숙하지 않은 경우 Python을 사용해도 괜찮습니다.
- 제출형식은 HTML, PDF 모두 가능합니다. .ipynb 이나 .R 등의 소스코드 파일은 불가능합니다. 파일은 psat2009@naver.com으로 보내주세요.
- 제출기한은 2월 29일 목요일 자정까지 입니다. 패키지와 마찬가지로 무단 미제출 2회 시 퇴출이니 유의해주세요.

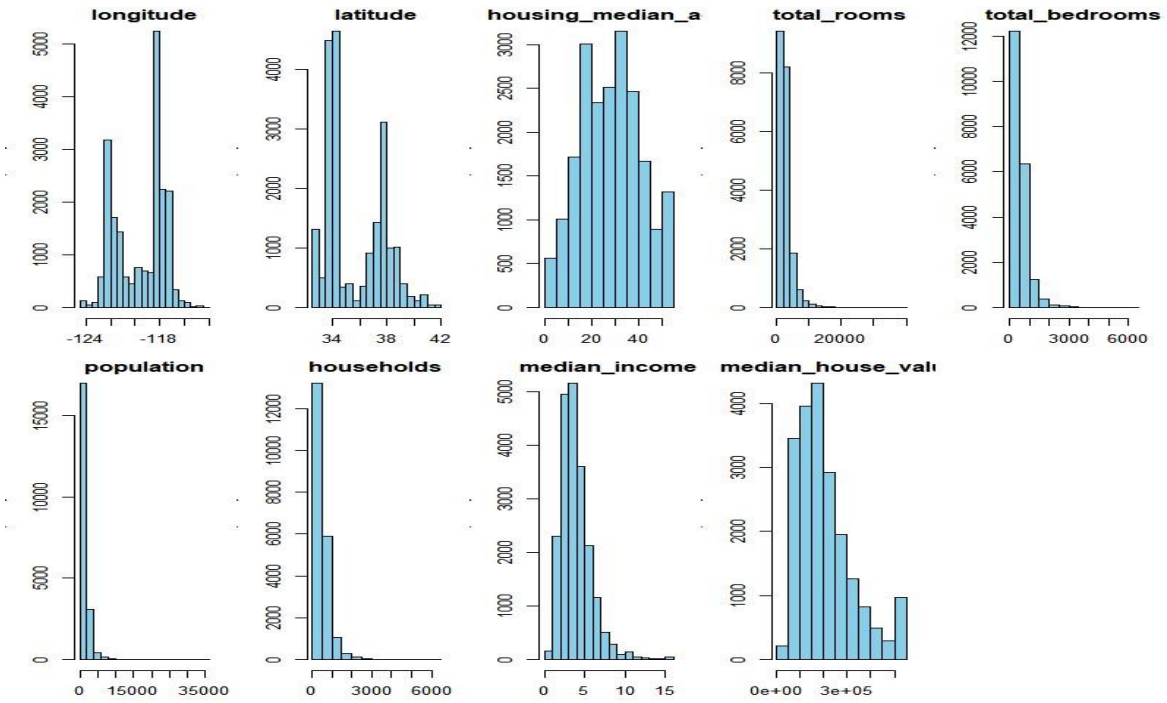
Chapter 1 : PCA와 EDA, Modeling 기초

3주차 전반부에는 고유값 분해와 PCA에 대해 배워보았습니다. PCA는 다차원의 데이터에서 분산을 최대로 보존하며 차원을 축소하는 축을 찾는 기법이라고 할 수 있습니다. 지금까지 배운 내용이 모두 PCA를 다루기 위해서라고 해도 무방할 정도로 많은 개념이 PCA에 응용되었는데요, 이러한 PCA를 이용해 직접 데이터 분석의 과정을 수행해보며, PCA는 물론 분석 과정에도 익숙해지는 시간이 될 수 있었으면 좋겠습니다.
앞으로의 패키지 과제가 이번 Chapter1과 비슷한 흐름으로 진행될 예정입니다. 아직 R이나 이런 분석 흐름에 익숙하지 않으신 학회원께서는 할 수 있는 부분만 해서 보내주셔도 됩니다!

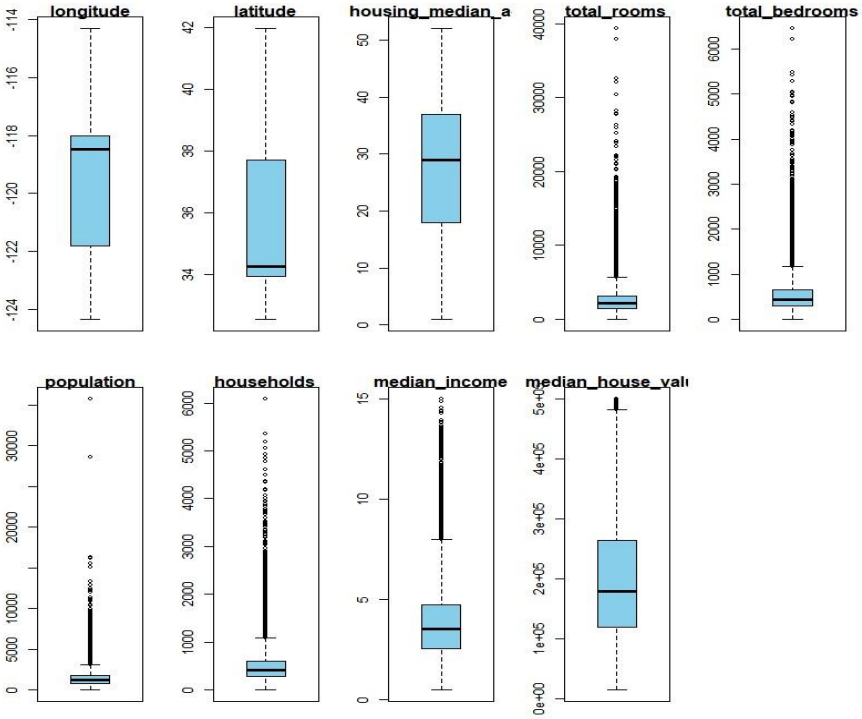
문제1. PCA의 목적과 작동 원리에 대해 설명해주세요.

문제2. PCA에서 적절한 주성분의 수를 결정하는 방법을 찾아본 후, 작성해주세요.

문제 3. housing.csv 데이터를 R 로 불러와주세요. data 의 구조를 자유롭게 파악하고, 수치형 변수만 남겨주세요. 또 아래와 같이 시각화를 진행해본 후 그 특징을 설명해주세요. (형식은 자유입니다)
(HINT) R 에서는 데이터 구조 파악을 위해 head, tail, summary, glimpse, str 등 다양한 함수가 쓰입니다. 저는 히스토그램을 그리기 위해 hist() 함수를 사용했습니다.

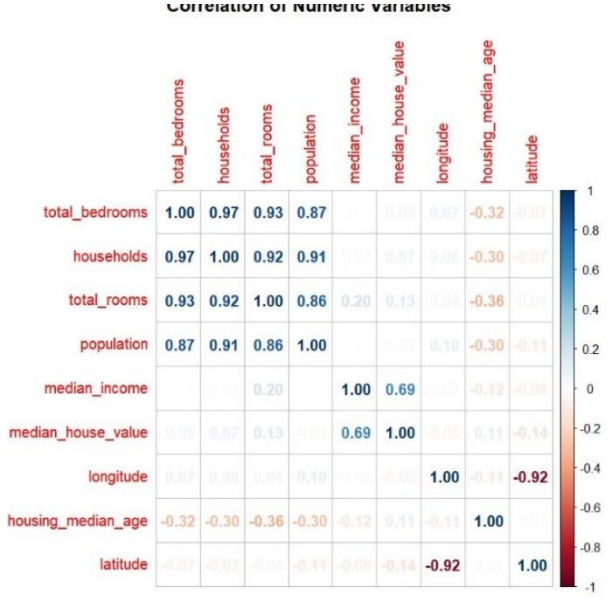


문제 4. 변수 별로 박스 플랏을 그려본 후, 데이터의 특징에 대해 추측해주세요. 이상치가 PCA 에 미치는 영향에 대해 추측해본 후 작성해주세요. (형식은 자유입니다)
(HINT) 저는 박스 플랏을 그리기 위해 boxplot() 함수를 사용했습니다.



문제 5. 데이터 내에 결측치(NA)가 있는지 간단히 파악해보세요. 위의 시각화를 통해 파악한 데이터의 특징을 바탕으로, 결측치를 적절한 값으로 대체해주세요. (정답은 없습니다. 그렇게 대체한 이유도 적어주세요)

문제 6. 데이터에서 수치형 변수들만 사용하여, 수치형 변수들 간의 상관관계를 다음과 같은 상관관계 Plot 을 통해 확인해주세요. 그리고 그 결과에 대해서 간단히 해석해주세요.



(HINT) corrplot 패키지를 활용하면, 상관계수 행렬로 쉽게 상관관계 plot 을 그릴 수 있습니다.

문제 7. median_house_value 를 y 변수로 두어 일반 선형회귀모형을 적합해보고, 결과를 해석해주세요.

문제 8. 이제 PCA 를 진행해보겠습니다.

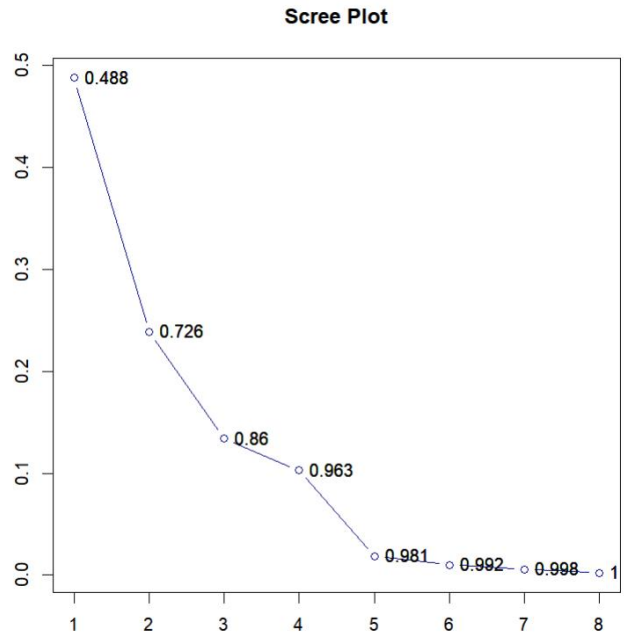
문제 8-1. 데이터를 표준화하여 data_scaled 데이터를 만들어주세요. 데이터를 centering만 했을 때와 표준화까지 진행했을 때 어떤 차이가 발생하는지 설명해주세요.

문제 8-2. data_scaled의 공분산 행렬 cov를 고유값 분해한 후 고유값의 크기 순으로 고유값과 고유벡터를 재정렬 해주세요. 공분산 행렬의 고유값이 갖는 의미에 대해 설명해주세요.

(HINT) cov(), eigen(), order() 등의 함수를 사용해주시면 됩니다.

문제 8-3. 주성분의 개수를 정하기 위해, 각 주성분이 설명하는 분산의 비율을 구하고 Scree Plot을 그려주세요. 이를 바탕으로 적절한 주성분의 개수를 결정해주세요. (정답은 없습니다)

(BONUS) 누적비율도 포인트 옆에 함께 표시해주세요



(HINT) Scree Plot이란 각 주성분(PCA)에 해당하는 고유값의 % 비율을 차트로 그려낸 것을 의미합니다. 분산 설명 비율은 교안을 참고해주세요. 누적비율은 분산 설명 비율의 누적합을 계산한 후, text()를 통해 표시하면 됩니다.

문제 9. 선택한 주성분의 개수로 새로운 데이터셋을 만들어주세요. 즉, 선택한 주성분에 data_scaled 를 projection 하여 data_pca 를 만들어주세요.

(HINT) R 에 내적계산을 위한 함수로 '%*%'가 있습니다.

문제 10. data_pca 를 X(설명변수), median_house_value 를 y(종속변수)로 설정하여 선형회귀모델을 적합한 후, 결과를 파악해주세요. PCA 를 이용한 선형 회귀는 분석 결과를 해석하기에 적절한가요?

문제 11. R 의 pca 패키지를 불러와, data_scaled 에 대해 같은 개수의 주성분 개수를 선택하여 pca 를 수행해주세요. 이후 선형 회귀 모델을 적합해주세요. 같은 결과가 나오는지 확인해주세요.

문제 12. 위의 분석 과정에서, 아쉬운 점이나 발전시킬 사항이 있었다면 이를 자유롭게 언급해주세요.

Chapter 2: SVD

Chapter2는 3주차 후반부에서 배운 SVD의 응용 사례를 직접 구현해보는 시간을 갖도록 하겠습니다. SVD, 특히 truncated SVD의 특징을 잘 생각해보며 과제를 가벼운 마음으로 진행해주시면 좋을 것 같습니다. 모두 한 달 동안, 정말정말 고생 많으셨습니다!!

문제1. 'imager' 라이브러리를 불러와주세요. picture.jpeg를 불러와 image에 저장해주세요.

문제2. 행렬화를 위해 사진을 흑백으로 변경해주세요.

```
image<- grayscale(image)
im.mat <- as.matrix(image)
plot(image)
```

를 시행해주시면 됩니다.

문제3. im.mat에 대해 svd()함수로 svd를 시행해 im_svd에 저장해주세요. im_svd의 데이터 구조를 str()로 파악해주세요.

문제4. 아래 코드를 통해, svd를 통해 압축한 새로운 이미지를 표현해보겠습니다. 아래 코드에서 특이값을 조금씩 바꾸어가며 결과를 확인해보세요. 어떤 차이가 발생합니까?

```
U=svd$u[,1:50]
D = diag(svd$d[1:50])
V = svd$v[,1:50]
im_part = U%*%D%*%t(V)
im_new <- as.cimg(im_part)
plot(im_new)
```