 # 1

US Crime 데이터는 1985년 미국의 각 주에서 발생한 범죄의 수를 나타낸 데이터이다. Murder, Robbery, Assault, Burglary, Larceny, Auto theft 이렇게 총 7개의 범죄가 기록되어 있다. 그리고 해당 주가 속해 있는 지역(Northeast, Midwest, South, West)을 나타내는 열도 있다. Division을 나타내는 열은 삭제하였다. [그림1]을 보면 수치형 변수들 사이 다중공산성이 존재한다는 것을 알 수 있다. 따라서 변수를 줄여 차원을 축소하고 변수들 사이의 correlation을 없애고자 PCA를 진행한다. PCA는 unsupervised learning이다. 해당하는 response가 없으므로 9개의 변수($X_1, X_2, \cdots, X_9$)만을 가지고 분석한다.

PCA를 진행할 때 모든 데이터를 표준화시킨다. 모든 숫자형 변수가 평균이 0이고 표준편차가 1이 되도록 scale을 맞춰준 것이다. PCA에서 얻어낸 principal component(PC)는 9개의 변수($X_1, X_2, \cdots, X_9$)의 선형 결합이다. PC1은 가장 큰 분산을 가지고 있어 전체 데이터 variability의 약 50%를 설명한다. 그리고 PC2는 PC1를 제외하고 그 다음으로 가장 큰 분산을 가지고 있으며 PC1과 uncorrelated한 관계이다. [그림2]에 따르면 이 데이터에서 PC1와 PC2 두 개의 components만으로도 전체 variability 중 60%에 대한 정보를 담을 수 있다.

[그림2]의 scree plot을 통해 차원을 축소시킬 때 사용하는 PC의 개수를 결정할 수 있다. 보통 scree plot의 elbow 지점의 개수를 선택한다. 이 경우에는 elbow 지점이 2이기 때문에 PC1과 PC2만을 선택할 수 있다. 하지만 PC1과 PC2를 선택하면 약 65%의 전체 variability를 설명할 수 있는데 PC1, PC2, PC3을 선택하면 전체 variability의 약 80%를 설명할 수 있다. 뿐만 아니라 [그림3]의 PC2 vs PC3 그래프를 보면 PC1 vs PC2 그래프보다 clustering이 더 잘 되는 것을 확인할 수 있다. 그렇기 때문에 나는 PC3까지 포함시켜 총 3개의 PC로 차원을 축소하였다.

차원을 축소한 후 [그림3]과 [그림4]로 해석을 한다. [그림4]에 의하면 PC2가 7개의 범죄를 잘 구분 짓는다. Murder, Assault, Rape와는 음의 상관관계를 갖고 Auto theft, Larceny, Burglary, Robbery와는 양의 상관관계를 갖는다. 그리고 [그림3]을 보면 PC2의 값이 작은 부분은 South 지역의 주가 대부분 차지하고 있다. 이를 통해 South 지역의 주는 다른 주보다 Murder, Assault, Rape가 자주 일어난다고 할 수 있다. 또한 [그림4]를 보면 PC3은 land.area를 잘 나타내고 있다. 그리고 [그림3]에서 알레스카 주는 다른 주에 비해서 PC3이 매우 크다. 즉 알래스카 주는 다른 주에 비해서 면적이 넓다는 것을 의미한다. [그림4]에서 변수들이 원에 가까울수록 두 PC로 잘 설명되는 변수이다. [그림4]의 첫 번째 그래프를 보면 다른 변수들과는 다르게 land.area가 잘 설명되지 않고 있다. 즉 land.area를 제외하면 PC1과 PC2만으로도 나머지 변수들이 잘 설명될 것이라고 예상할 수 있다.

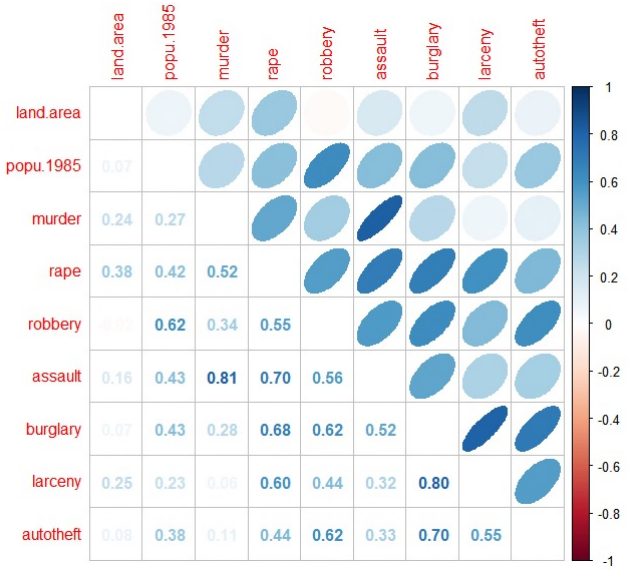
따라서 land.area를 제외하고 다시 PCA를 진행하였다. 실제로 [그림5]에서 PC1의 분산이 land.area를 포함했을 때보다 커져서 더 많은 총 variability를 설명한다. 따라서 PC1와 PC2만으로도 전체 variability의 약 75%를 설명한다. 그리고 [그림5]의 loading plot을 보면 변수들이 원 주변에 위치하여 PC1와 PC2로 변수들에 대한 정보를 잘 담고 있다고 확인할 수 있다.

 # 2

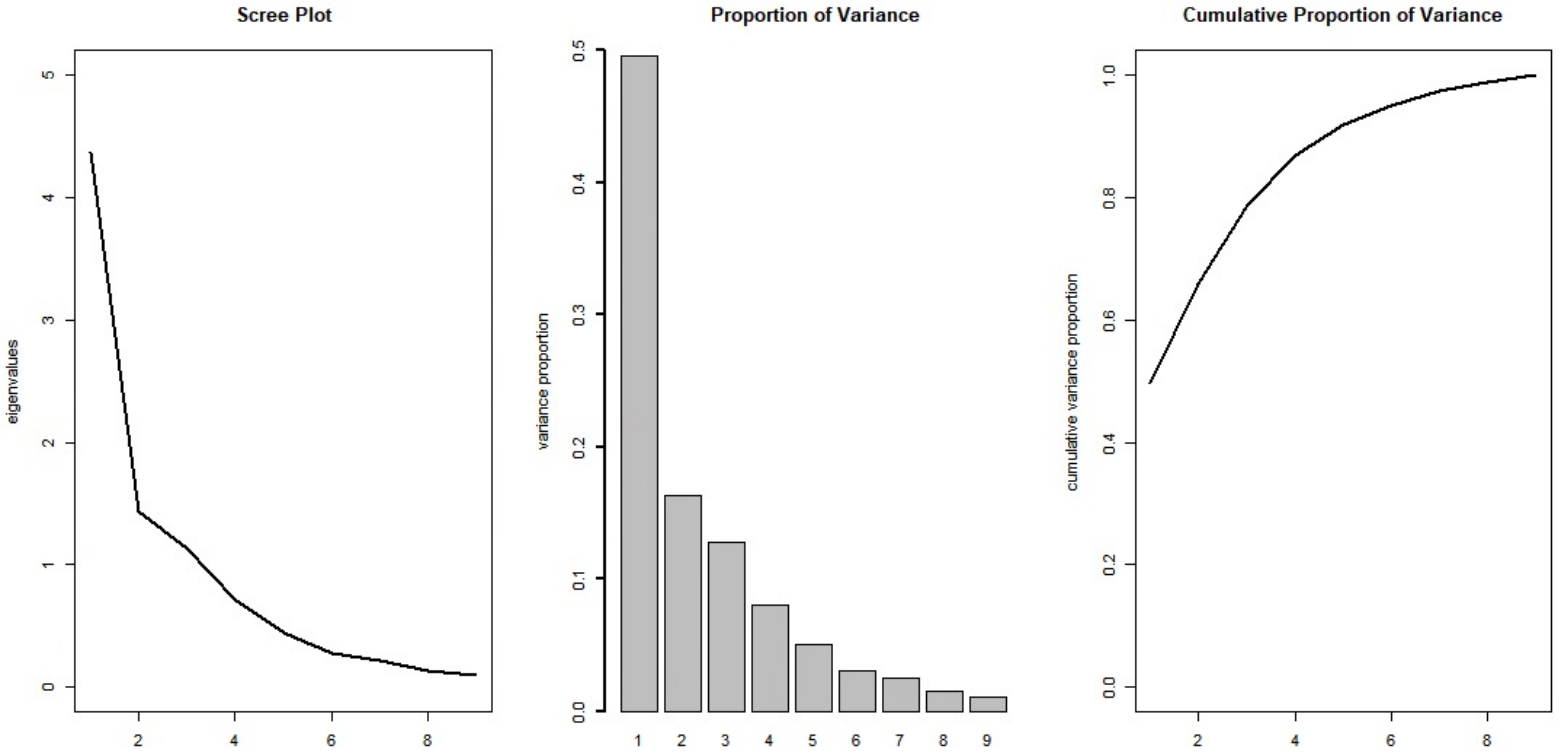
비슷하게 US health 데이터를 가지고 PCA를 진행한다. US health 데이터는 미국의 주마다 사망 원인별 수를 기록한 데이터이다. Accident, Cardiovascular, Cancer, Pulmonary, Pneumonia Flu, Diabetes, Liver 이렇게 7개의 사망 원인이 있다. 그리고 각 주의 의사 수와 병원 수에 관한 정보도 있다. 또한 문제 1번의 데이터와 비슷하게 해당 주가 속해 있는 지역(Northeast, Midwest, South, West)에 대한 변수도 추가되어 있다. 숫자형 변수들은 서로 강한 상관관계를 가지고 있어 다중공산성 문제가 존재한다. 따라서 PCA를 통해 차원을 축소하고 다중공산성 문제를 해결하고자 한다.

먼저 데이터를 표준화시킨다. 즉 모든 숫자형 변수가 평균이 0이고 표준편차가 1이 되도록 scale을 조정해준다. 그리고 spectral decomposition을 통해 principal component(PC)를 구한다. [그림6]는 각 PC가 총 variability의 얼만큼을 설명하고 있는지 나타내고 있다. [그림6]의 scree plot을 보면 elbow가 세 번째 PC에 위치한다. 따라서 PC1, PC2, PC3으로 차원을 축소한다. 이 세 개의 PC는 총 variability의 약 75%를 설명한다.

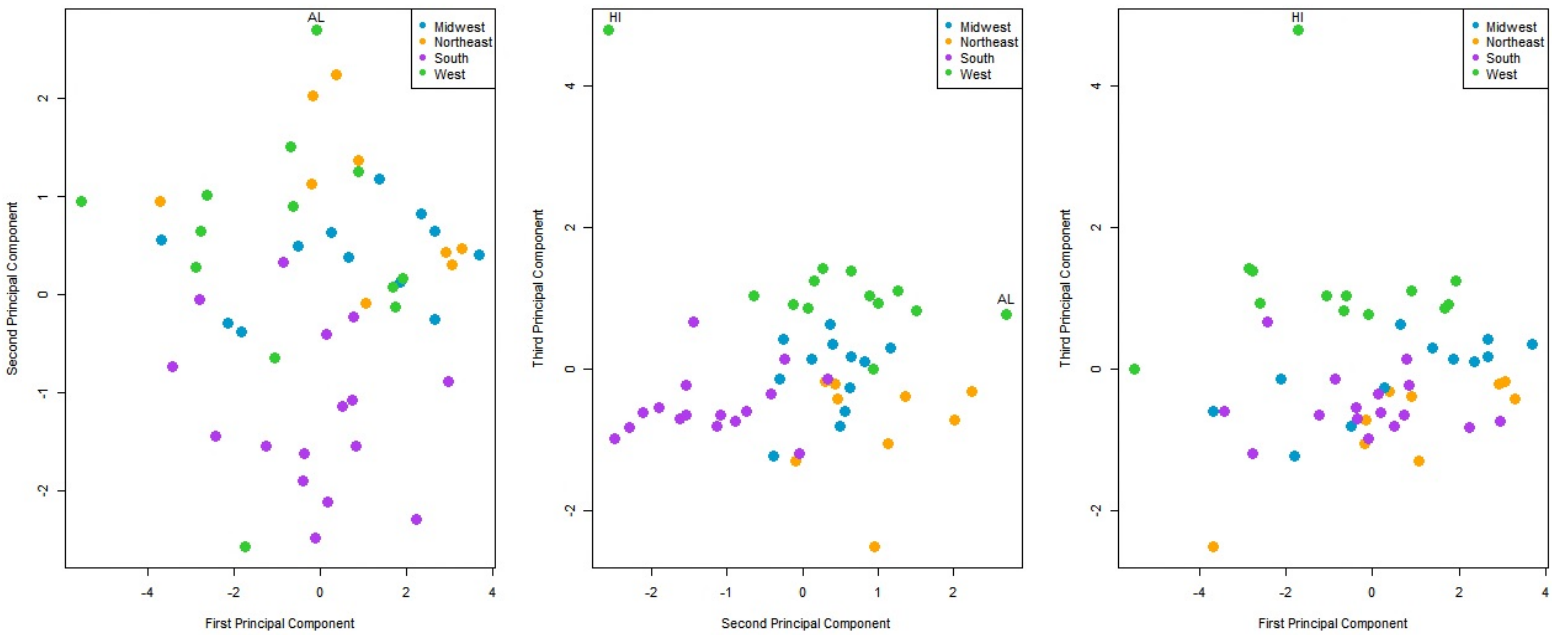
이렇게 차원을 축소한 후 [그림7]과 [그림8]로 해석을 한다. [그림8]에서 PC1 vs PC2 그래프를 보면 PC1은 Accident을 잘 구분해며 Accident와 양의 상관관계를 갖는다. 따라서 [그림7]에서 큰 PC1 값을 가지는 알레스카 주는 다른 주에 비해 Accident로 인한 죽음이 많이 발생한다. PC2는 의사의 수, 병원의 수, 인구 수와 같은 변수와 음의 상관관계를 갖는다. 이 변수들은 죽음의 원인과는 관련이 없고 주의 성질과 관련이 있는 변수이다. 따라서 [그림7]에서 매우 작은 PC2 값을 갖는 캘리포니아 주와 뉴욕 주는 의사의 수가 많고, 병원이 많으며, 전체 인구 수 또한 많다는 것을 알 수 있다. 반대로 매우 큰 값을 갖는 알레스카 주는 의사 수가 적고, 병원이 별로 없으며 전체 인구 수가 많지 않다. 즉 PC2는 죽음의 종류보다는 주의 성질을 나타내는 정보를 담고 있으며 PC1와 PC3은 죽음의 원인에 관한다고 할 수 있다.


land.area
popu.1985
murder
rape
robbery
assault
burglary
larceny
autotheft
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

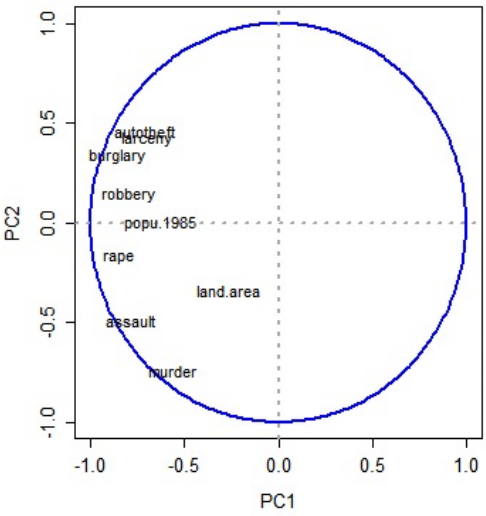
[그림1]

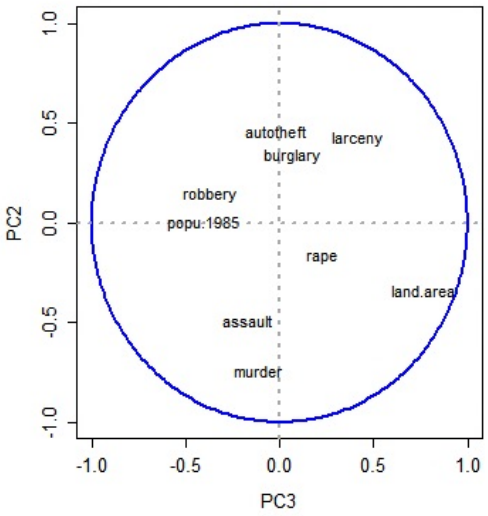

Scree Plot
eigenvalues
Proportion of Variance
variance proportion
Cumulative Proportion of Variance
cumulative variance proportion

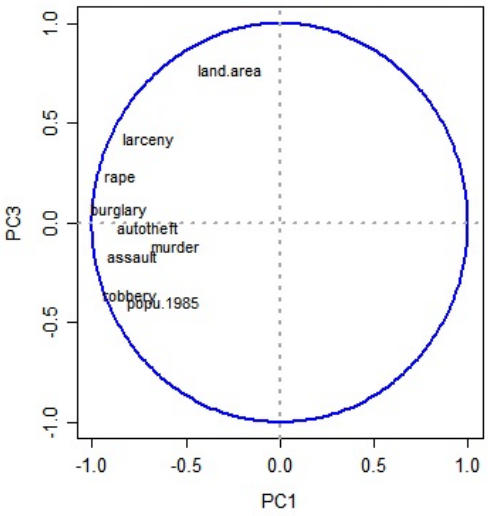
[그림2]


AL
HI
Midwest
Northeast
South
West
Second Principal Component
First Principal Component
Third Principal Component

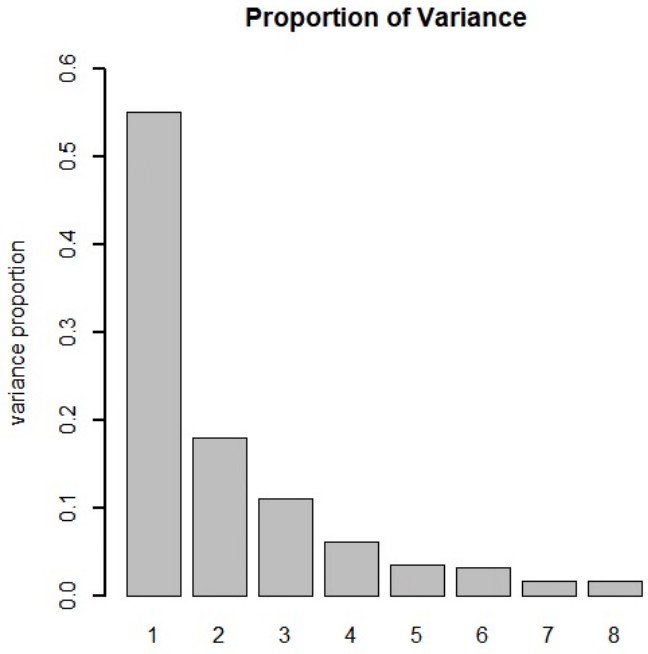
[그림3]

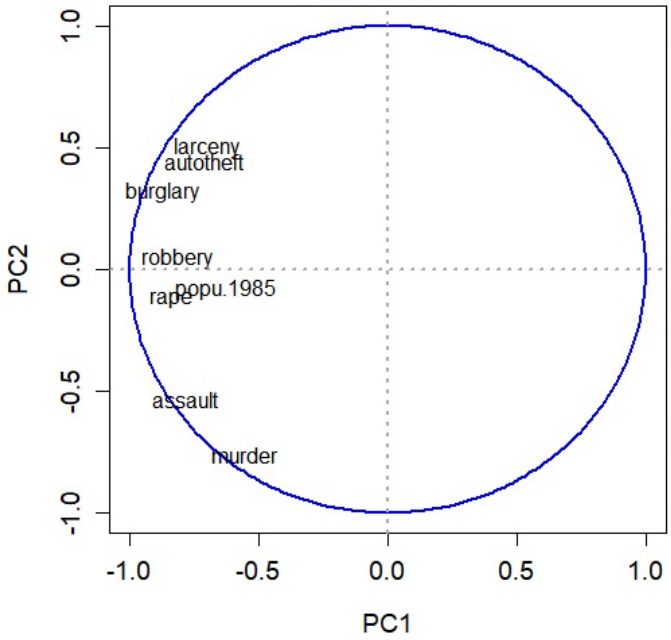

PC2
PC1
autotheft
larceny
burglary
robbery
popu.1985
rape
land.area
assault
murder


PC2
PC3
autotheft
larceny
burglary
robbery
popu.1985
rape
land.area
assault
murder

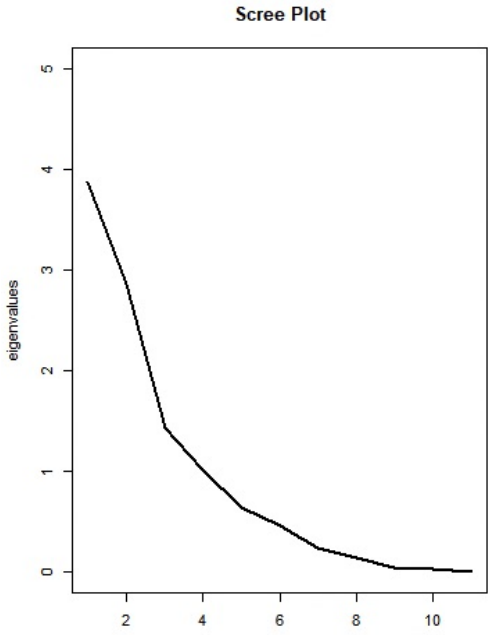

PC3
PC1
land.area
larceny
rape
burglary
autotheft
murder
assault
robbery
popu.1985

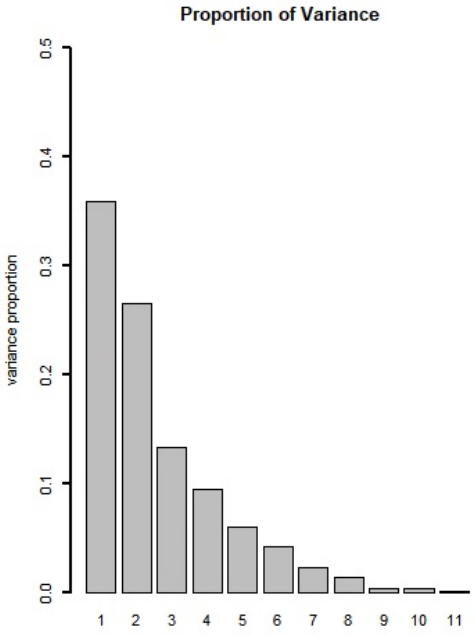
[그림4]

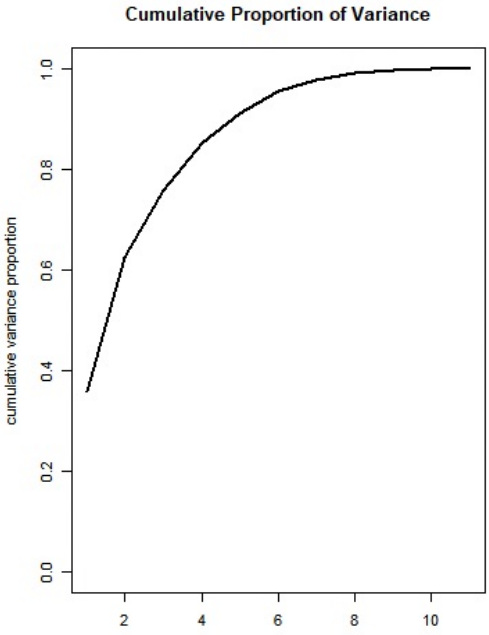

Proportion of Variance
variance proportion


PC2
PC1
larceny
autotheft
burglary
robbery
popu.1985
rape
assault
murder

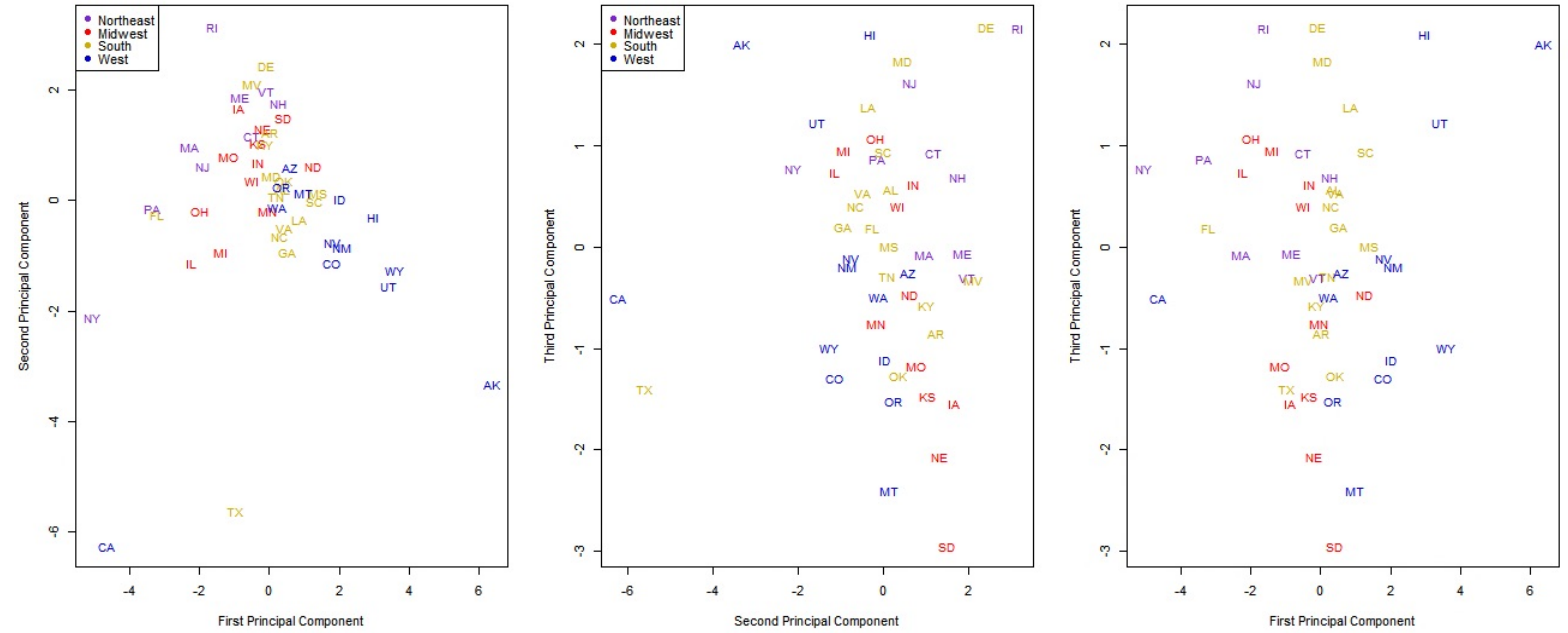
[그림5]


Scree Plot
eigenvalues

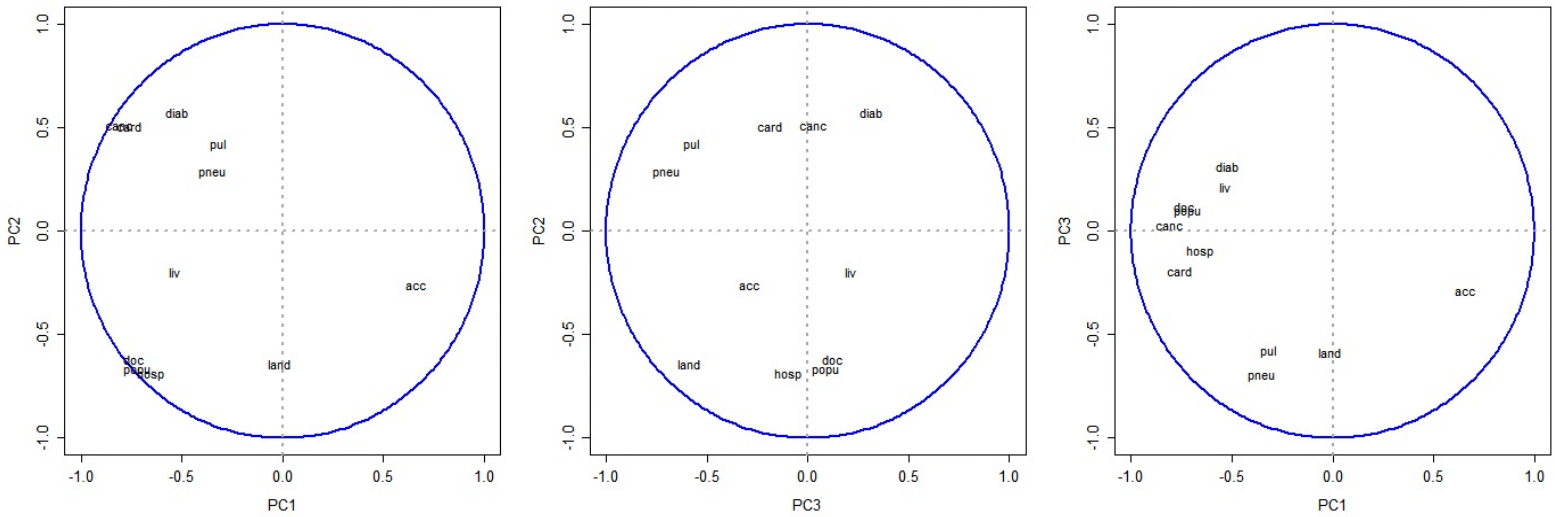

Proportion of Variance
variance proportion


Cumulative Proportion of Variance
cumulative variance proportion

[그림6]


Northeast
Midwest
South
West
Second Principal Component
First Principal Component
Third Principal Component
Second Principal Component
Third Principal Component
First Principal Component

[그림7]


PC2
PC1
diab
pul
pneu
liv
acc
land
PC2
PC3
card
canc
diab
pul
pneu
acc
liv
land
hosp
doc
popu
PC3
PC1
diab
liv
canc
hosp
card
acc
pul
land
pneu

[그림8]