PCA_1210_데이터마이닝 팀 프로젝트

#1 성능 비교: 훈련 데이터 70%로 했을 때

Show perfomance scores

Model	MSE	RMSE	MAE	MAPE	R2
kNN	80.837	8.991	6.876	0.107	0.068
SVM	38.759	6.226	5.186	0.075	0.553
Linear Regression	16.110	4.014	2.988	0.045	0.814

PCA 하기 전 성능

Show perfomance scores

Model	MSE	RMSE	MAE	MAPE	R2
kNN (1)	25.220	5.022	3.672	0.056	0.709
Linear Regression (1)	32.984	5.743	4.625	0.071	0.620
SVM (1)	135.061	11.622	10.484	0.148	-0.557

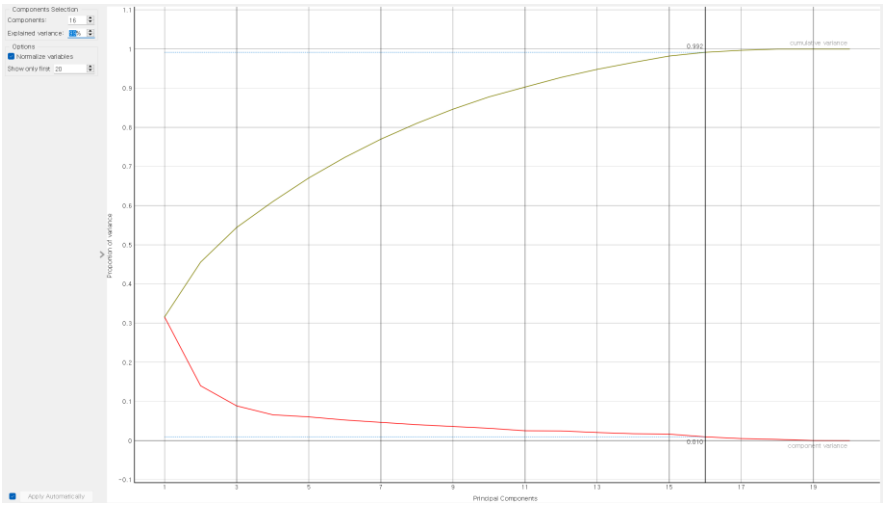
PCA 한 후 성능(주성분 2개)

Show perfomance scores

Model	MSE	RMSE	MAE	MAPE	R2
kNN (1)	7.917	2.814	1.898	0.029	0.909
Linear Regression (1)	16.974	4.120	3.043	0.046	0.804
SVM (1)	39.575	6.291	5.214	0.077	0.544

(주성분 16개)

#2 PCA 실행



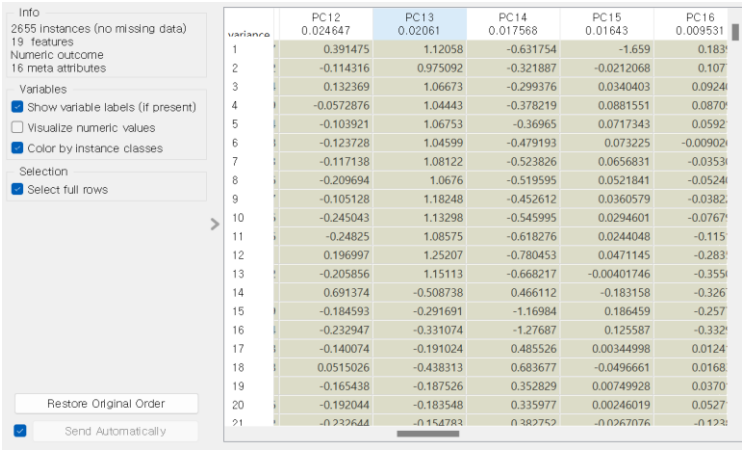
기울기가 급격하게 낮아지는 부분의 수치로 주성분의 개수를 정한다 = 2

위에 따르면 주성분 2개지만, 설명력 99%일 때의 주성분을 보겠음

Scree plot

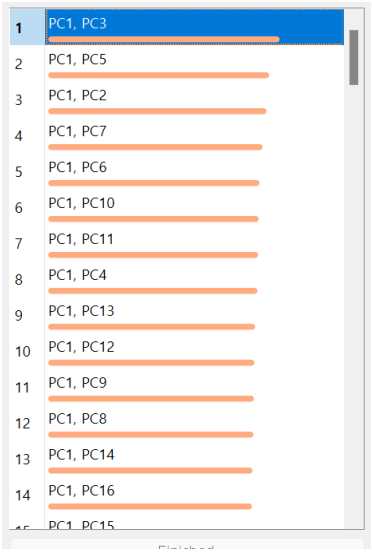
붉은색 선분은 주성분의 각 개수 마다의 변동의 비율값이고, 초록색 선분은 주성분의 개수가 증가할 때의 변동의 누적 비율이다. X 축은 주성분들의 개수이고 y축은 변동의 비율 값이다. 이 도형에서 y축은 계산된 PCA의 값이 갖는 원 데이터세트에 대한 설명력이라고 보면 되며 값이 0.8,

즉 80% 이상의 설명력을 가진 때의 x 축 값을 보면 된다. 위의 그림에서는 PC(주성분)의 개수가 1일 때 0.3(충분한 설명력 보장X)이고 주성분의 개수가 16일 때 99%로 거의 100%의 설명력을 가짐을 알 수 있다.

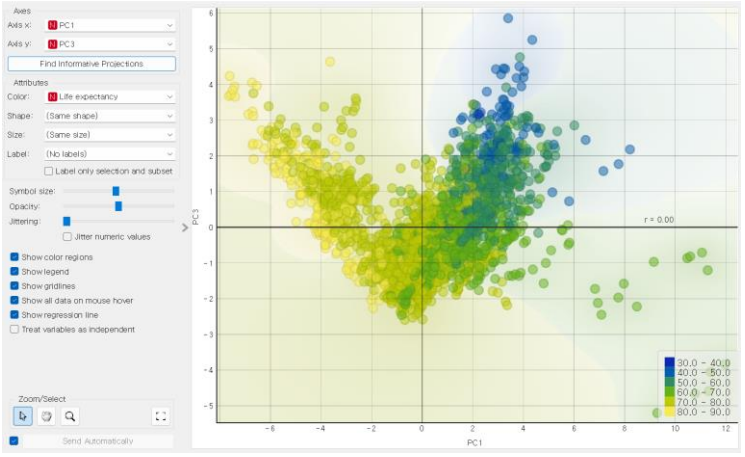


19개 feature에서 16개의 PC 생성

여기서는 PC1 즉 첫번째 주성분이 분산을 0.31을 보존했으므로 설명력이 31%라고 볼 수 있다. PC2는 두번째 주성분으로써 분산이 0.13 정도로써 PC1보다는 설명력이 높지 않음을 알 수 있다.

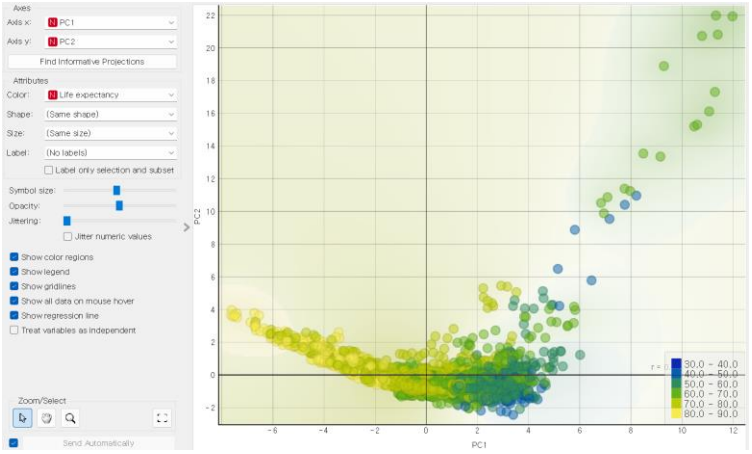


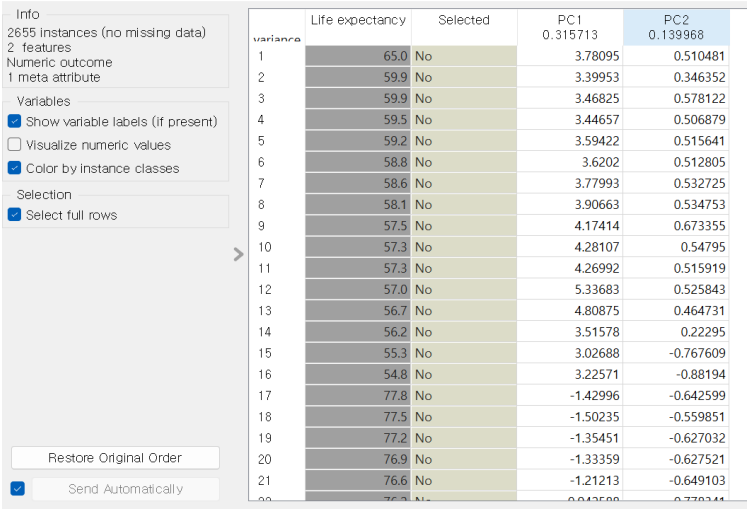
informative projections 순위



PC1과 PC3을 x,y축

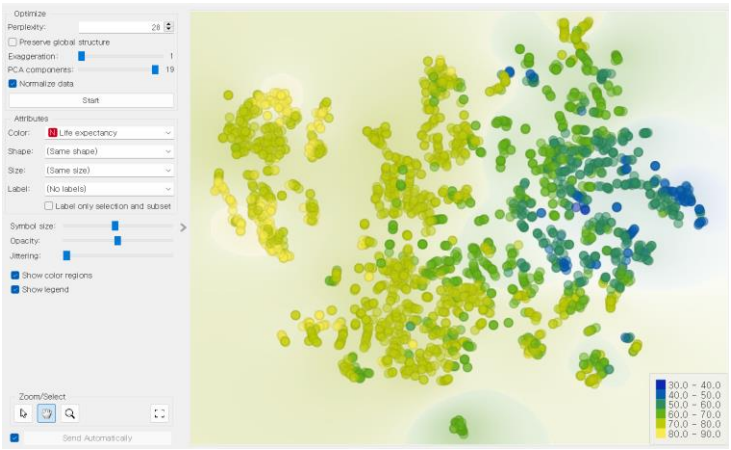
IF 주성분의 개수를 2개로 하였을 때





	Life expectancy	Selected	PC1 0.315713	PC2 0.139968
1	65.0	No	3.78095	0.510481
2	59.9	No	3.39953	0.346352
3	59.9	No	3.46825	0.578122
4	59.5	No	3.44657	0.506879
5	59.2	No	3.59422	0.515641
6	58.8	No	3.6202	0.512805
7	58.6	No	3.77993	0.532725
8	58.1	No	3.90663	0.534753
9	57.5	No	4.17414	0.673355
10	57.3	No	4.28107	0.54795
11	57.3	No	4.26992	0.515919
12	57.0	No	5.33683	0.525843
13	56.7	No	4.80875	0.464731
14	56.2	No	3.51578	0.22295
15	55.3	No	3.02688	-0.767609
16	54.8	No	3.22571	-0.88194
17	77.8	No	-1.42996	-0.642599
18	77.5	No	-1.50235	-0.559851
19	77.2	No	-1.35451	-0.627032
20	76.9	No	-1.33359	-0.627521
21	76.6	No	-1.21213	-0.649103

PC1, PC2 설명력 변화 없음



t-SNE

t-distributed stochastic neighbor embedding: 높은 차원의 복잡한 데이터를 2차원에 차원 축소하는 방법, 낮은 차원 공간의 시각화에 주로 사용하며 차원 축소할 때는 비슷한 구조끼리 데이터를 정리한 상태이므로 데이터 구조를 이해하는 데 도움을 준다. 매니폴드(Manifold) 학습의 하나로 복잡한 데이터의 시각화가 목적이다. 높은 차원의 데이터를 2차원 또는 3차원으로 축소시켜 시각화 합니다. 높은 차원 공간에서 비슷한 데이터 구조는 낮은 차원 공간에서 가깝게 대응하며, 비슷하지 않은 데이터 구조는 멀리 떨어져 대응된다.