PCA / FA

DASS(Depression Anxiety Stress Scales Responses) 분석

데이터마이닝이론및응용
에이쁠원하조

2020147024 김우영
2020251009 김혜리
2020195053 노가원
2020147018 조윤영

목차

- 연구 목적 및 데이터
- **PCA**
- **FA**
- 연구의 한계점 및 개선안

1. 연구 목적 및 데이터 (Depression Anxiety Stress Scales Responses | Kaggle)

- 연구 목적: Depression Anxiety Stress Scales (DASS) dataset은 우울증, 불안, 스트레스에 대한 설문조사 데이터이다. DASS data를 PCA/FA를 이용해 분석함으로써 설문 문항들 간의 관계를 확인하고 정리해보고자 한다. 이를 통해 설문 문항 데이터의 구조를 파악할 수 있을 것이다.
- 데이터 프레임 재구성: Original dataset 중 우리가 필요한 데이터는 "각 설문 문항에 대한 응답자들의 답변(=A)" 데이터이므로 A를 제외한 데이터는 제외하고 데이터 프레임을 재구성하였다.

설문 문항 42개의 내용은 아래와 같다.

Q1 I found myself getting upset by quite trivial things.
Q2 I was aware of dryness of my mouth.
Q3 I couldn't seem to experience any positive feeling at all.
Q4 I experienced breathing difficulty (eg, excessively rapid breathing, breathlessness in the absence of physical exertion).
Q5 I just couldn't seem to get going.
Q6 I tended to over-react to situations.
Q7 I had a feeling of shakiness (eg, legs going to give way).
Q8 I found it difficult to relax.
Q9 I found myself in situations that made me so anxious I was most relieved when they ended.
Q10 I felt that I had nothing to look forward to.
Q11 I found myself getting upset rather easily.
Q12 I felt that I was using a lot of nervous energy.
Q13 I felt sad and depressed.
Q14 I found myself getting impatient when I was delayed in any way (eg, elevators, traffic lights, being kept waiting).
Q15 I had a feeling of faintness.
Q16 I felt that I had lost interest in just about everything.
Q17 I felt I wasn't worth much as a person.
Q18 I felt that I was rather touchy.
Q19 I perspired noticeably (eg, hands sweaty) in the absence of high temperatures or physical exertion.
Q20 I felt scared without any good reason.
Q21 I felt that life wasn't worthwhile.
Q22 I found it hard to wind down.
Q23 I had difficulty in swallowing.
Q24 I couldn't seem to get any enjoyment out of the things I did.
Q25 I was aware of the action of my heart in the absence of physical exertion (eg, sense of heart rate increase, heart missing a beat).
Q26 I felt down-hearted and blue.
Q27 I found that I was very irritable.
Q28 I felt I was close to panic.
Q29 I found it hard to calm down after something upset me.
Q30 I feared that I would be thrown by some trivial but unfamiliar task.
Q31 I was unable to become enthusiastic about anything.
Q32 I found it difficult to tolerate interruptions to what I was doing.
Q33 I was in a state of nervous tension.
Q34 I felt I was pretty worthless.
Q35 I was intolerant of anything that kept me from getting on with what I was doing.
Q36 I felt terrified.
Q37 I could see nothing in the future to be hopeful about.
Q38 I felt that life was meaningless.
Q39 I found myself getting agitated.
Q40 I was worried about situations in which I might panic and make a fool of myself.
Q41 I experienced trembling (eg, in the hands).
Q42 I found it difficult to work up the initiative to do things.

2. 데이터 및 변수 설명 - 기초통계량

Q1A~Q42A는 각각 질문 Q1~Q42에 대한 응답 데이터들의 모음이며,
본 dataset은 총 39775명의 응답으로 이루어져있다.
답변은 {1, 2, 3, 4} 중 하나로 이루어져있다. (1 = Did not apply to me at all
2 = Applied to me to some degree, or some of the time
3 = Applied to me to a considerable degree, or a good part of the time
4 = Applied to me very much, or most of the time)

아래는 기초통계량이다.

	Q1A	Q2A	Q3A	Q4A	Q5A	Q6A	Q7A	Q8A	Q9A	Q10A	...
count	39775.000000	39775.000000	39775.000000	39775.000000	39775.000000	39775.000000	39775.000000	39775.000000	39775.000000	39775.000000	...
mean	2.619485	2.172269	2.226097	1.950170	2.521458	2.540214	1.924928	2.480427	2.669591	2.447316	...
std	1.032117	1.111563	1.038526	1.042218	1.069908	1.049672	1.033528	1.052436	1.067866	1.139350	...
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...
25%	2.000000	1.000000	1.000000	1.000000	2.000000	2.000000	1.000000	2.000000	2.000000	1.000000	...
50%	3.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	3.000000	2.000000	...
75%	4.000000	3.000000	3.000000	3.000000	3.000000	3.000000	3.000000	3.000000	4.000000	4.000000	...
max	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	...

8 rows × 42 columns

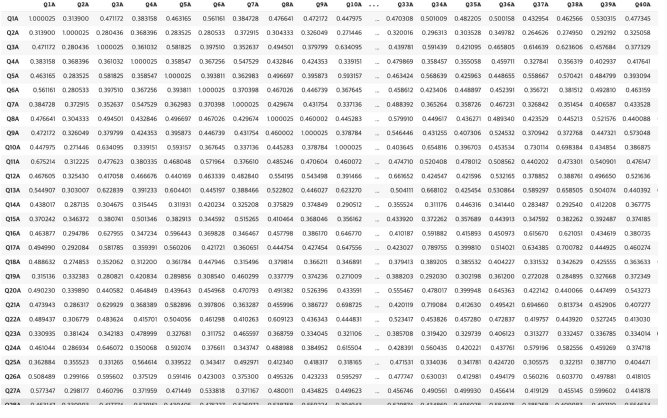
3. 표준화

우리가 다루는 데이터에서는 변수 간 스케일 차이가 없으므로 표준화를 하지 않아도 무리가 없지만, 해석의 편의성을 위해 표준화를 진행하였다.

오른쪽 표는 표준화 후에 구한 공분산 행렬이다.

```
data_scale.mean(axis=0)
```

```
data_scale.cov()
```



연구 목적 및 데이터

1. PCA 주성분 개수 결정

PCA(1)

1) eigen value와 scree plot을 통해 결정

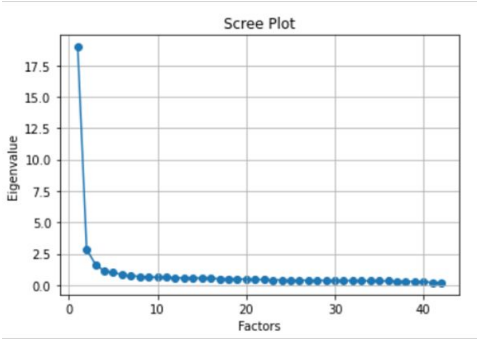
고유값이 큰 순서대로 주성분을 나열했을 때 **eigen value**가 **1** 이상인 주성분이 **4**개인 것을 확인 할 수 있다.
또한 **scree plot**을 확인 했을 때, 주성분이 **4**개인 경우부터 기울기가 완만해지는 것을 볼 수 있다.

```
[ ]  # Eigen value
     fa = FactorAnalyzer(n_factors=data_scale.shape[1], rotation=None) # rotation:
     fa.fit(data_scale)

     ev, v = fa.get_eigenvalues()
     print('eigen value:', ev,'₩n', "number of eigen value:", len(ev))

     print('적합한 PC 개수:', sum(ev >= 1))
```

```
eigen value: [18.96485725  2.86014105  1.64011679  1.10761201  0.98976351  0.83703621
  0.73872579  0.69003303  0.64952984  0.61290793  0.60540206  0.58962406
  0.5649372   0.5566493   0.54707564  0.53019429  0.50828819  0.50801573
  0.49759778  0.45818426  0.44354743  0.42303039  0.41409666  0.41005221
  0.39921098  0.38917797  0.38056682  0.37725025  0.36955681  0.36319195
  0.36065407  0.34404278  0.33983758  0.33344802  0.32765883  0.32566775
  0.31353891  0.30999681  0.26519283  0.2593666   0.20920004  0.18502242]
number of eigen value: 42
적합한 PC 개수: 4
```



2) 주성분 개수에 따른 분산 설명력 확인

주성분의 누적 분산 비율 결과를 보면 주성분이 **4**개인 경우 총분산의 약 **59%**를 설명한다. **80%** 이상의 분산을 설명하기 위해서는 **19**개의 주성분이 필요하지만 차원을 충분히 축소하기 위해 고유값의 크기를 기준으로 **4**개의 주성분을 선택하기로 결정했다.

```
누적 분산 설명 비율:
 [0.45154422 0.51964282 0.55869322 0.58506493 0.60863073 0.62856016
 0.64614887 0.66257823 0.67804323 0.69263627 0.70705061 0.72108927
 0.73454016 0.74779371 0.76081933 0.773443   0.7855451  0.79764071
 0.80948828 0.82039743 0.83095808 0.84103023 0.85088968 0.86065282
 0.87015785 0.87942399 0.8884851  0.89746725 0.90626622 0.91491365
 0.92350065 0.93169215 0.93978352 0.94772276 0.95552416 0.96327815
 0.97074336 0.97812424 0.98443836 0.99061375 0.9955947  1.        ]
```

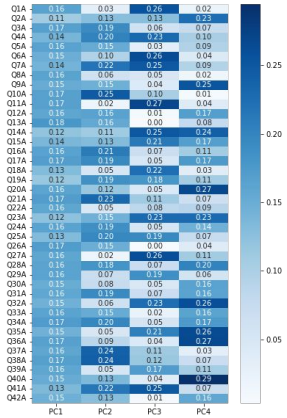
2. 4개의 주성분으로 진행한 PCA 결과

PCA(2)

1) 주성분 변수의 계수

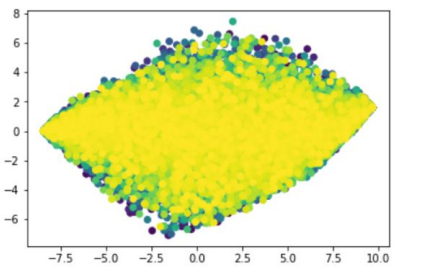
PCA 결과 주성분에 따른 변수 별 계수는 아래와 같고 누적 분산 설명율은 총 58.5%로 나타났다.
아래의 표를 바탕으로 계수의 절대값 히트맵을 그렸다. 세로로 해석했을 때 PC1 축의 색이 평균적으로 진한 것을 보아 모든 변수가 PC1에 영향을 주는 것을 확인할 수 있다. 그리고 PC2,3,4는 주성분이 일부 변수에 강한 영향을 받는 것을 확인할 수 있다. 가로로 분석을 했을 때는 각 변수가 가장 영향을 많이 주는 주성분을 확인할 수 있었다.

	PC1	PC2	PC3	PC4
Q1A	0.163876	0.030918	-0.255661	-0.021288
Q2A	0.108999	0.128054	0.127555	0.225015
Q3A	0.166012	-0.191221	0.060063	0.074447
Q4A	0.140552	0.201424	0.230056	0.102089
Q5A	0.164842	-0.152329	0.029933	0.094806
Q6A	0.148111	0.095267	-0.263481	-0.042198
Q7A	0.141987	0.218215	0.248558	0.093290
Q8A	0.162221	0.060714	-0.048781	0.015003
Q9A	0.151354	0.146318	-0.041687	-0.251523
Q10A	0.165335	-0.245823	0.102031	-0.005894
Q11A	0.165916	0.017517	-0.269113	-0.039375
Q12A	0.159218	0.164489	0.009049	-0.169135
Q13A	0.178275	-0.155660	0.003468	-0.076837
Q14A	0.121673	0.110884	-0.249854	0.240654
Q15A	0.136833	0.129800	0.212643	0.168702
Q16A	0.164493	-0.210295	0.068084	0.107048
Q17A	0.170059	-0.193916	0.049024	-0.174192
Q18A	0.131296	0.051190	-0.223744	0.034627
Q19A	0.115555	0.186100	0.179536	0.107122
Q20A	0.161160	0.116319	0.045195	-0.268013
Q21A	0.171436	-0.231577	0.111234	-0.066832
Q22A	0.159882	0.045641	-0.084192	0.090339
Q23A	0.123570	0.145231	0.229631	0.229311
Q24A	0.163348	-0.187341	0.050596	0.144841
Q25A	0.133045	0.196605	0.194113	0.073773
Q26A	0.170717	-0.148059	0.004535	-0.039834
Q27A	0.159832	0.017937	-0.260531	0.108988
Q28A	0.161734	0.181278	0.068265	-0.201811
Q29A	0.163026	0.070456	-0.187571	-0.055999
Q30A	0.150279	0.075069	-0.046375	-0.164767
Q31A	0.159142	-0.193036	0.068415	0.163737
Q32A	0.145444	0.060466	-0.225109	0.261075
Q33A	0.162751	0.151626	0.022643	-0.162198
Q34A	0.171966	-0.197766	0.052531	-0.174675
Q35A	0.145188	0.048173	-0.205003	0.264189
Q36A	0.165214	0.085803	0.043401	-0.265005
Q37A	0.160051	-0.237858	0.111865	-0.031242
Q38A	0.167753	-0.240922	0.115670	-0.068679
Q39A	0.159708	0.051209	-0.172827	0.106322
Q40A	0.151535	0.134210	-0.038030	-0.294478
Q41A	0.134944	0.215072	0.248073	0.069875
Q42A	0.146361	-0.127546	0.007876	0.157140



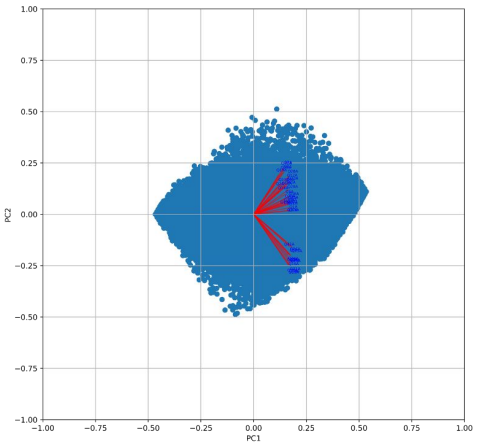
2) Score Plot

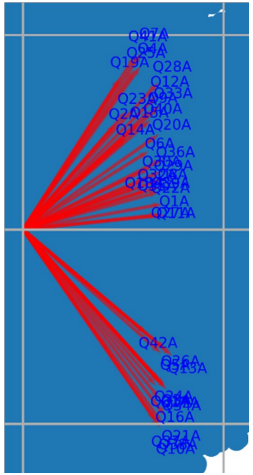
데이터가 0주위로 무작위로 배치되어 정규 분포에 근사하는 것을 확인할 수 있고, 이상치도 관측되지 않는다.



3) Biplot

Biplot은 아래와 같이 도출되었다. Loading plot에서 화살표가 서로 가까운 변수의 그룹들을 확인하였고 가까운 변수 사이의 상관관계를 예측해볼 수 있다, 이를 통해 설문의 항목이 몇가지 주제로 분류될 수 있다고 생각된다. 또한 Q27A, Q11A 변수가 PC1 축에 거의 평행하여 PC1에 강하게 영향을 주고 있는 것을 확인할 수 있다. PC2의 경우, PC2축에 평행에 가까운 기울기를 가지는 변수가 없는 것을 관찰할 수 있는데, 이것은 PCA 히트맵에서 확인했듯이 모든 변수들이 PC1에 영향을 주기 때문에 오직 PC2에만 강하게 영향을 주는 변수는 없음을 보여준다.

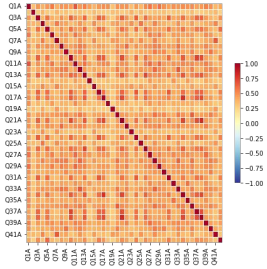




FA

FA

1)



변수들간에 강한 상관관계가 보임으로 몇가지 factor로 묶일 수 있을 것

```
from factor_analyzer.factor_analyzer import calculate_kmo
kmo_all, kmo_model = calculate_kmo(data_scale)
kmo_model # 0.8 이상이므로 꽤 좋음
```

```
/usr/local/lib/python3.9/dist-packages/factor_analyzer/utils.py:244: UserWarning: The inverse of the variance-cova
  warnings.warn(
0.9851280484409755
```

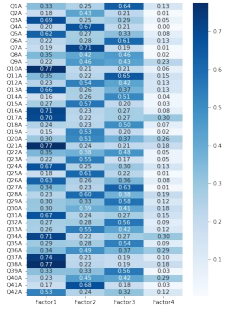
KMO Test) 진행 결과 0.9 이상으로, 변수들의 선정이 적절하다고 볼 수 있음

```
[31] from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
     chi_square_value, p_value = calculate_bartlett_sphericity(data_scale)
     chi_square_value, p_value # p-value < 0.05 --> 귀무가설 기각 --> 요인분석 모델 사용 가능

     (1092141.0647031434, 0.0)
```

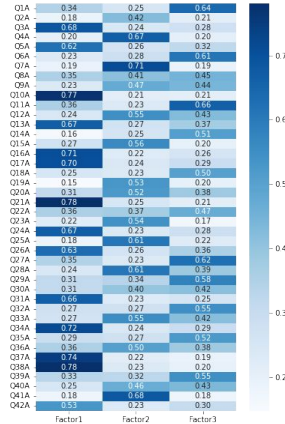
Bartlett Test) 귀무가설 기각으로 요인분석 모델 사용 가능

2)



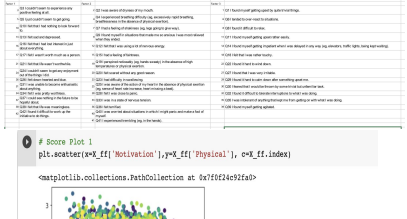
절대값 히트맵을 통해 3가지 factor (factor 1,2,3)를 선택함

3) varimax



factor1:Q3,Q5,Q10,Q13,Q16,Q17,Q21,Q24,Q26,Q31,Q34,Q37,Q38,Q42
factor2:Q2,Q4,Q7,Q9,Q12,Q15,Q19,Q20,Q23,Q25,Q28,Q33,Q36,Q40,Q41
factor3:Q1,Q6,Q8,Q11,Q14,Q18,Q22,Q27,Q29,Q30,Q32,Q35,Q39

factor 1에 해당하는 질문들은 대부분 motivation과 관련되어 있고, factor 2에 해당하는 질문들은 physical health, factor 3에 해당하는 질문들은 emotional states와 연관이 있는 것을 파악함

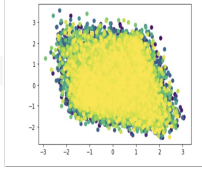


데이터및 각 Factor 들의 공통 요인들을 참고하여 Factors의 naming 진행

Factor 1: motivation
Factor 2: physical
Factor 3: emotional

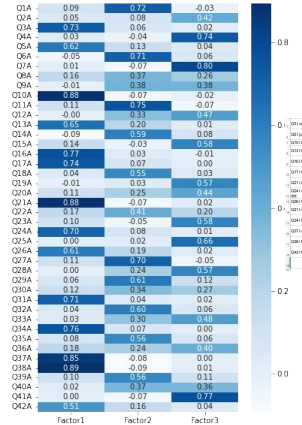
```
# Score Plot 1
plt.scatter(x=X_ff['Motivation'],y=X_ff['Physical'], c=X_ff.index)
```

```
<matplotlib.collections.PathCollection at 0x7f0f24c92fa0>
```



score plot에서 유사한 형태를 보이는 변수간에 상관관계가 있음을 예측해볼 수 있다.

4) oblimin

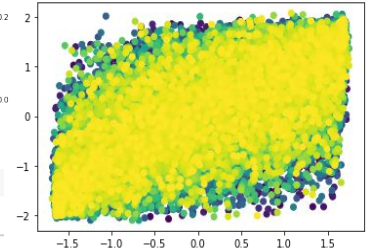


factor1: Q3,Q5,Q10,Q13,Q16,Q17,Q21,Q24,Q26,Q31,Q34,Q37,Q38,Q42
factor2:Q1,Q6,Q8,Q9,Q11,Q14,Q18,Q22,Q27,Q29,Q30,Q32,Q35,Q39,Q40
factor3:Q2,Q4,Q7,Q9,Q12,Q15,Q19,Q20,Q23,Q25,Q28,Q33,Q36,Q41

factor 1은 varimax와 동일, factor 2는 discouraged, factor 3는 anxiety와 연관

데이터및 각 Factor 들의 공통 요인들을 참고하여 Factors의 naming 진행

Factor 1: motivation
Factor 2: physical
Factor 3: anxiety



score plot에서 유사한 형태를 보이는 변수간에 상관관계가 있음을 예측해볼 수 있다.

연구의 한계점 및 개선안

PCA,FA진행시 연구자의 주관적인 의견에 따라 결과를 작위적으로 조절할 수 있고 해석의 어려움이 존재하기도 한다. 고차원의 데이터를 저차원의 데이터로 차원 축소시에 정보의 손실이 발생할 수 있다. PCA의 경우 몇개의 주성분을 이용할지에 대한 결정은 명확한 정답이 없고, 오로지 연구자의 주관적인 판단으로 결정되는 것이기 때문에 신뢰성이 떨어질 수도 있다. FA의 경우에도 회전방법론에 따라 각 factor에 해당하는 변수의 구성이 달라지기도 하므로 어떤 방법론을 선택하는지 또한 연구자의 주관적인 판단에 맡길 수 밖에 없다.

Reference

LUCAS GREENWELL. “Data Set.” LUCAS GREENWELL.