[Spring 2023] DATA MINING THEORY AND APPLICATION (IIE 4102)

Practice-HW3
Clustering Analysis

논문참조

강세정, 김규현, 신재욱, 장현우

Introduction

01. Research Objective

- 다양한 Hierarchical Clustering과 Partitional Clustering을 모두 사용하여 어떠한 방법론이 날씨 데이터들을 더 우수하게 군집화하는지 비교해보고자 함
- 거리가 가까운 날씨 데이터들 간에 어떠한 유사한 특징이 있는지 파악해보고자 함

02. Data

데이터는 기상청(ASOS)에서 서울 지역에 대해 직접 수집함

- 데이터 기간 : 10년 (2013.03.29~2023.03.28)
- National Geographic의 '날씨' 정의에 기반해 변수들을 선택함

X(Numerical Variable)

평균기온 (°C)	일강수량 (mm)	평균 풍속 (m/s)	평균 상대 습도(%)	평균 현지 기압(hPa)	평균 전운량 (1/10)

일강수량, 평균 풍속, 평균 현지기압에 결측치 존재

- 일강수량 : 0으로 채움 [`data['일강수량'].fillna(0)`]
- 평균 풍속, 평균 현지기압 : 보간법(interpolation) 사용 [`data.interpolate()`]

IQR 기준으로 결측치(outlier) 제거

- (1Q - 1.5 * IQR, 3Q + 1.5 * IQR) 범위 밖의 값을 가지는 row 제거
- (3652, 6) → (3536, 6)

① `data.head()`

	평균기온	일강수량	평균 풍속	평균 상대습도	평균 현지기압	평균 전운량
0	5.4	0.0	3.4	28.6	1010.4	1.9
1	4.9	0.2	2.3	43.1	1011.9	7.8
2	6.3	0.0	2.7	46.4	1012.7	0.0
3	9.2	0.5	2.2	44.5	1009.3	7.3
4	7.2	8.5	2.7	86.1	1002.7	7.6

② `data.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3652 entries, 0 to 3651
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   평균기온      3652 non-null   float64
 1   일강수량      1421 non-null   float64
 2   평균 풍속     3649 non-null   float64
 3   평균 상대습도   3652 non-null   float64
 4   평균 현지기압   3651 non-null   float64
 5   평균 전운량    3652 non-null   float64
dtypes: float64(6)
memory usage: 171.3 KB
```

결측치 제거 이후 `data.info()`

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3536 entries, 0 to 3651
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   평균기온      3536 non-null   float64
 1   일강수량      3536 non-null   float64
 2   평균 풍속     3536 non-null   float64
 3   평균 상대습도   3536 non-null   float64
 4   평균 현지기압   3536 non-null   float64
 5   평균 전운량    3536 non-null   float64
dtypes: float64(6)
memory usage: 193.4 KB
```

③ `data.describe()`

	평균기온	일강수량	평균 풍속	평균 상대습도	평균 현지기압	평균 전운량
count	3536.000000	3536.000000	3536.000000	3536.000000	3536.000000	3536.000000
mean	13.507749	3.177489	2.241502	60.838660	1006.238009	4.765639
std	10.511083	12.241339	0.659321	14.769925	7.920814	3.078080
min	-14.900000	0.000000	0.600000	21.800000	981.000000	0.000000
25%	4.600000	0.000000	1.800000	49.900000	999.800000	2.100000
50%	14.550000	0.000000	2.100000	60.800000	1006.500000	4.800000
75%	22.900000	0.300000	2.700000	70.925000	1012.500000	7.400000
max	33.700000	176.200000	4.000000	99.800000	1026.800000	10.000000

StandardScaling 이후 `data.describe()`

	평균기온	일강수량	평균 풍속	평균 상대습도	평균 현지기압	평균 전운량
count	3.536000e+03	3.536000e+03	3.536000e+03	3.536000e+03	3.536000e+03	3.536000e+03
mean	-3.617016e-17	-6.832142e-17	4.822688e-17	-1.366428e-16	1.851510e-14	-2.009453e-17
std	1.000141e+00	1.000141e+00	1.000141e+00	1.000141e+00	1.000141e+00	1.000141e+00
min	-2.703029e+00	-2.596071e-01	-2.490039e+00	-2.643492e+00	-3.186740e+00	-1.548469e+00
25%	-8.475823e-01	-2.596071e-01	-6.697259e-01	-7.407083e-01	-8.129113e-01	-8.661295e-01
50%	9.917137e-02	-2.596071e-01	-2.146478e-01	-2.617818e-03	3.308094e-02	1.116466e-02
75%	8.936833e-01	-2.350965e-01	6.955085e-01	6.829938e-01	7.906860e-01	8.559664e-01
max	1.921315e+00	1.413628e+01	2.667514e+00	2.638257e+00	2.596311e+00	1.700768e+00

Body

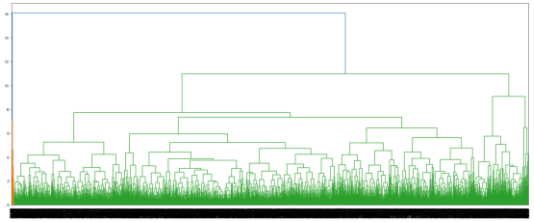
Contents

Clustering Analysis

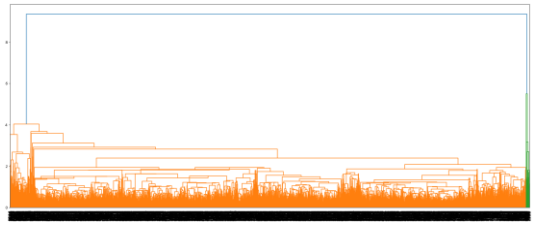
Hierarchical Clustering

01. 관측치들 간 거리 계산 방식 선정 : 개별 클러스터의 실루엣 계수 평균값의 편차가 작도록 threshold 설정

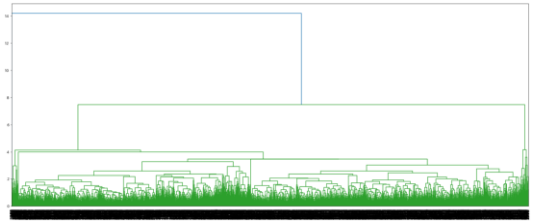
① Complete : Silhouette Average Score 0.226



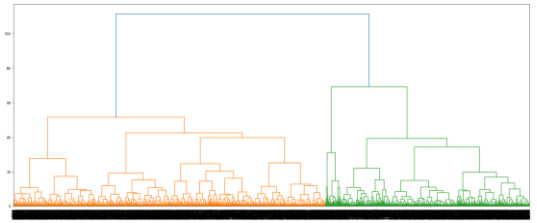
② Centroid : Silhouette Average Score 0.022



③ Average : Silhouette Average Score 0.182



④ Ward : Silhouette Average Score 0.179



군집별 평균 silhouette score 값의 편차가 작은 Ward, Complete 방식 선정

▶ 군집별 평균 silhouette_score 값 / 군집별 데이터 수

```
hc_cluster
1    0.306824
2    0.055304
3    0.304332
4    0.254853
Name: silhouette_coeff, dtype: float64
표준편차: 0.11911213167712349
```

```
2    1566
4    1286
1     573
3     111
dtype: int64
```

Ward 방식

```
hc_cluster
1    0.403099
2    0.211975
3    0.327468
4    0.415107
Name: silhouette_coeff, dtype: float64
표준편차: 0.09339620399460181
```

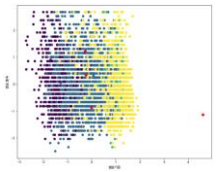
```
2    3153
3     328
4      38
1      17
dtype: int64
```

Complete 방식

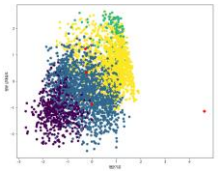
02. EDA / Interpretation

① Ward

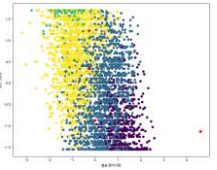
평균 기온 - 평균 풍속



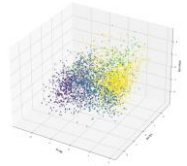
평균 기온 - 평균 상대습도



평균 현지기압 - 평균 전운량



평균 기온 - 평균 풍속 - 평균 상대습도



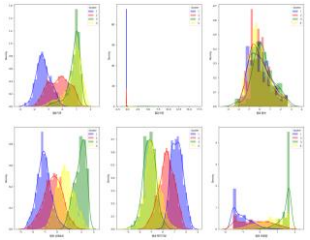
군집별 평균

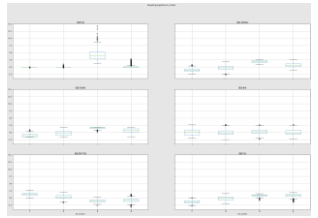
hc_cluster	평균기온	일강수량	평균 풍속	평균 상대습도	평균 현지기압	평균 전운량	silhouette_coeff
1	-1.341538	-0.257939	0.037732	-1.060858	1.231961	-1.018271	0.306824
2	-0.231883	-0.236145	-0.027873	-0.317785	0.340211	-0.252696	0.055304
3	0.833078	4.601439	0.318327	1.896140	-1.128353	1.464540	0.304332
4	0.808210	0.005321	-0.010346	0.695997	-0.865814	0.635013	0.254853

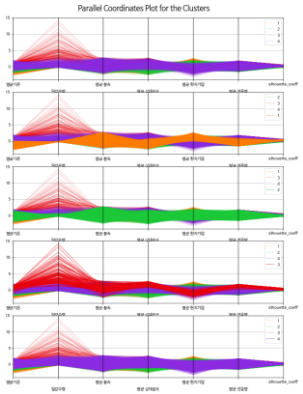
군집별 표준편차

hc_cluster	평균기온	일강수량	평균 풍속	평균 상대습도	평균 현지기압	평균 전운량	silhouette_coeff
1	0.497139	0.014425	1.020855	0.573219	0.501741	0.519271	0.106167
2	0.741035	0.086646	0.986619	0.764795	0.666116	0.867639	0.116672
3	0.435673	2.552235	0.995108	0.455473	0.522917	0.338407	0.202608
4	0.612362	0.508210	1.003554	0.674607	0.630762	0.758999	0.135154

- cluster1 : 기온이 매우 낮고, 상대습도와 전운량도 매우 낮음. 비가 거의 오지 않음 → 건조한 겨울 날씨
- cluster2 : 비는 거의 오지 않았으나 나머지 모든 값이 0에서 크게 벗어나지 않는 수치를 보임. 기온이 다소 낮고 기압이 다소 높음 → 건조한 가을 날씨
- cluster3 : 기온이 높고, 강수량과 상대습도 및 전운량이 매우 높음 → 습하고 비오는 여름 날씨
- cluster4: cluster3과 유사하지만 비가 별로 오지 않았음 → 비오기 전후 여름 날씨







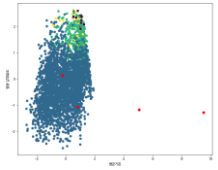
Hierarchical Clustering(CONT.)

02. EDA / Interpretation(CONT.)

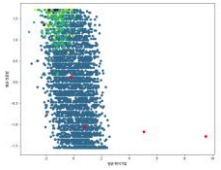
② Complete

평균 기온 - 평균 풍속

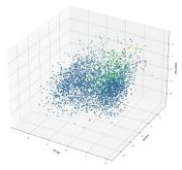
평균 기온 - 평균 상대습도



평균 현지기압 - 평균 전운량



평균 기온 - 평균 풍속 - 평균 상대습도



군집별 평균

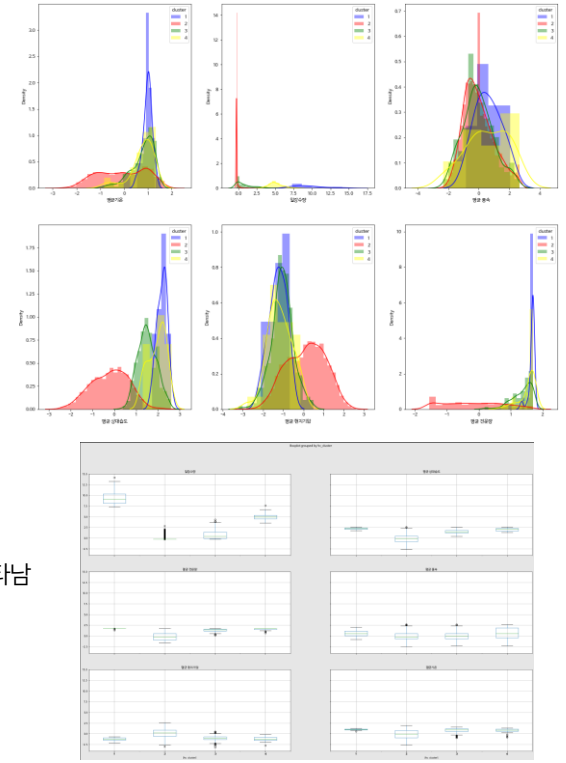
hc_cluster	평균기온	일강수량	평균 풍속	평균 상대습도	평균 현지기압	평균 전운량	silhouette_coeff
1	-1.341538	-0.257939	0.037732	-1.060858	1.231961	-1.018271	0.306824
2	-0.231883	-0.236145	-0.027873	-0.317785	0.340211	-0.252696	0.055304
3	0.833078	4.601439	0.318327	1.896140	-1.128353	1.464540	0.304332
4	0.808210	0.005321	-0.010346	0.695997	-0.865814	0.635013	0.254853

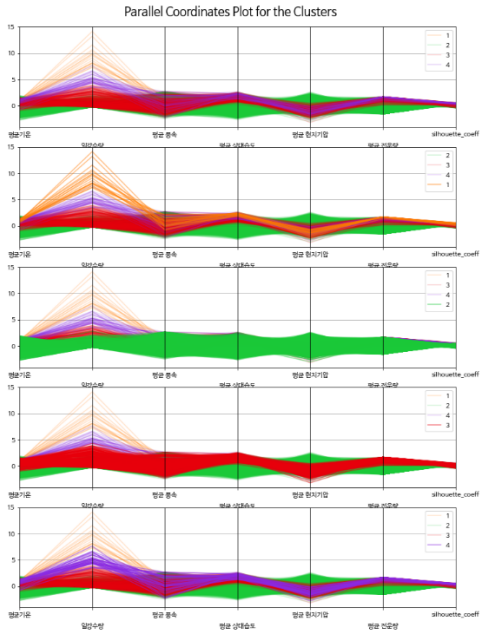
군집별 표준편차

hc_cluster	평균기온	일강수량	평균 풍속	평균 상대습도	평균 현지기압	평균 전운량	silhouette_coeff
1	0.497139	0.014425	1.020855	0.573219	0.501741	0.519271	0.106167
2	0.741035	0.086646	0.986619	0.764795	0.666116	0.867639	0.116672
3	0.435673	2.552235	0.995108	0.455473	0.522917	0.338407	0.202608
4	0.612362	0.508210	1.003554	0.674607	0.630762	0.758999	0.135154

- 전체 표본의 평균이 0이 되어야 하므로 표의 값에 의하면 압도적으로 많은 데이터가 cluster 2에 속해야 하고, 실제로도 그렇게 나타남
- 따라서 cluster 2는 특징 설명이 제한되며 1, 3, 4에 대해서만 특징 추론 가능
- cluster 1과 4는 모든 경향성이 유사하나 cluster 1이 더욱 극단적인 값들을 가짐

- **cluster 1** : 비오고 습한 한여름 날씨
- **cluster 3** : cluster 1과 4에 비해서는 기온이 다소 낮으며 강수량이 높지 않음 → 습한 봄가을 날씨
- **cluster 4** : 비오고 습한 여름 날씨





▶ 계층적 클러스터링의 경우, Ward 방식으로 관측치들 간 거리를 계산하는 것으로 결정

이후에 Partitional Clustering과 비교

Body

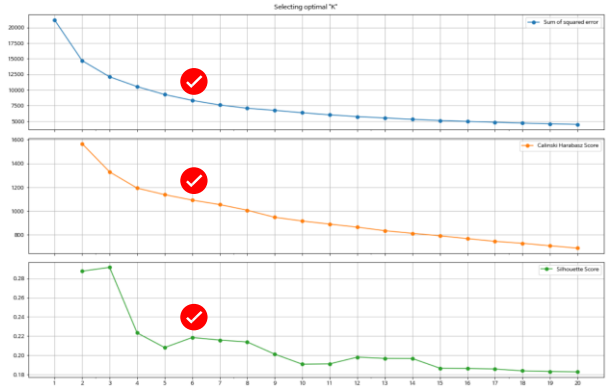
Contents

Clustering Analysis

Partitional Clustering

01. 군집 개수 결정

SSE, calinski harabasz score, silhouette score를 고려해 6으로 결정



02. K-means Clustering

군집별 centroid 좌표값

	평균기온	일강수량	평균 풍속	평균 상대습도	평균 현지기압	평균 전운량
0	0.591840	-0.250977	-0.022115	-0.583067	-0.316184	-0.691171
1	0.235707	0.130939	1.292674	0.703508	-0.655242	0.812549
2	-0.729474	-0.225823	-0.674428	-0.004291	0.835885	0.099950
3	0.972875	-0.019733	-0.549066	0.757095	-0.909000	0.758778
4	0.805164	4.931109	0.353817	1.923832	-1.092103	1.482840
5	-1.195199	-0.254253	0.516300	-1.002442	1.027414	-1.018255

기존 데이터 + cluster

	평균기온	일강수량	평균 풍속	평균 상대습도	평균 현지기압	평균 전운량	cluster
0	-0.771461	-0.259607	1.757358	-2.183032	0.525524	-0.931114	5
1	-0.819037	-0.243267	0.088738	-1.201168	0.714925	0.985936	2
2	-0.685825	-0.259607	0.695509	-0.977710	0.815939	-1.548469	5
3	-0.409887	-0.218756	-0.062955	-1.106368	0.386630	0.823474	2
4	-0.600189	0.434860	0.695509	1.710565	-0.446736	0.920951	1
...	...	...	...	...	...	...	...
3531	-0.095888	-0.259607	0.088738	-1.627771	-0.232081	1.050921	0
3532	-0.095888	-0.259607	0.088738	-0.063561	0.121468	-0.216282	0
3533	-0.390857	-0.259607	2.060743	-1.817373	0.070961	-1.288530	5
3534	-0.590674	-0.259607	0.088738	-1.086053	0.626538	-1.191053	5
3535	-0.409887	-0.259607	-0.214648	-0.510478	0.866446	-0.606190	2

군집별 데이터 개수

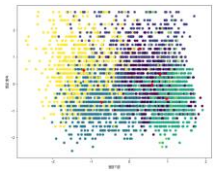
2	786
3	786
0	729
5	679
1	457
4	99

03. Evaluation

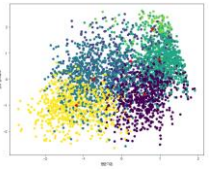
① 평균 silhouette score, 군집별 silhouette score 확인

② 군집 결과 시각화

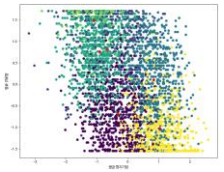
```
Average Silhouette Score:0.456

k_means_cluster
0    0.446444
1    0.301433
2    0.423431
3    0.475140
4    0.369637
5    0.600282
Name: k_silhouette_coeff, dtype: float64
표준편차: 0.10129124855096026
```



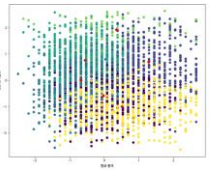
평균 기온 – 평균 풍속



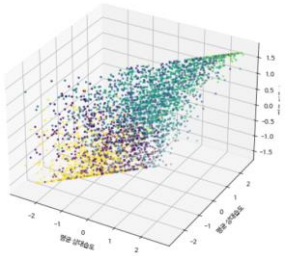
평균 기온 – 평균 상대습도



평균 현지기압 – 평균 전운량



평균 풍속 – 평균 상대습도



평균 기온 – 평균 풍속 – 평균 상대습도

04. EDA/Interpretation

군집별 평균

k_means_cluster	평균기온	일강수량	평균 풍속	평균 상대습도	평균 현지기압	평균 전운량	cluster	k_silhouette_coeff
0	0.593272	-0.250989	-0.022379	-0.582577	-0.317472	-0.690385	0.0	0.446444
1	0.238161	0.131257	1.294977	0.703038	-0.656666	0.812738	1.0	0.301433
2	-0.729672	-0.225554	-0.673264	-0.003117	0.834815	0.100746	2.0	0.423431
3	0.972032	-0.019428	-0.549491	0.758346	-0.908560	0.759895	3.0	0.475140
4	0.805164	4.931109	0.353817	1.923832	-1.092103	1.482840	4.0	0.369637
5	-1.195199	-0.254253	0.516300	-1.002442	1.027414	-1.018255	5.0	0.600282

군집별 표준편차

k_means_cluster	평균기온	일강수량	평균 풍속	평균 상대습도	평균 현지기압	평균 전운량	cluster	k_silhouette_coeff
0	0.486617	0.075140	0.761143	0.654974	0.579369	0.645533	0.0	0.092822
1	0.749448	0.595983	0.692122	0.739553	0.769205	0.679346	0.0	0.098109
2	0.549628	0.134836	0.642902	0.738083	0.500183	0.789422	0.0	0.104006
3	0.427181	0.496910	0.584156	0.629009	0.576952	0.609181	0.0	0.101862
4	0.456518	2.507175	1.034144	0.450827	0.544162	0.332438	0.0	0.157579
5	0.581044	0.028578	0.975306	0.597545	0.586415	0.540348	0.0	0.059316

- Cluster0 : 기온이 높은 편이며 비가 거의 오지 않고 습도와 전운량이 낮음 → 맑고 따뜻한 날
- Cluster1 : 습도와 전운량이 높으며 풍속이 매우 높음. 적은 양의 비가 내림 → 흐리고 바람부는 날
- Cluster2 : 기온이 낮은 편이며 바람이 전혀 불지 않고 기압이 매우 높음 → 선선한 날
- Cluster3 : 기온이 매우 높고 기압은 매우 낮으며 전운량이 매우 높음. 그러나 비는 별로 오지 않음 → 덥고 습한 날
- Cluster4 : Cluster3과 유사하지만 비가 매우 많이 옴 → 비오고 습한 날
- Cluster5 : 기온이 매우 낮으나 비가 전혀 오지 않고 상대습도가 매우 낮으며 전운량이 매우 적음 → 춥고 맑은 날

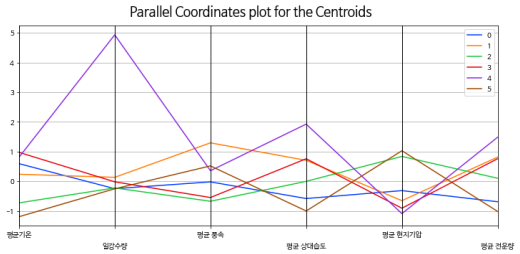
Body / Conclusion

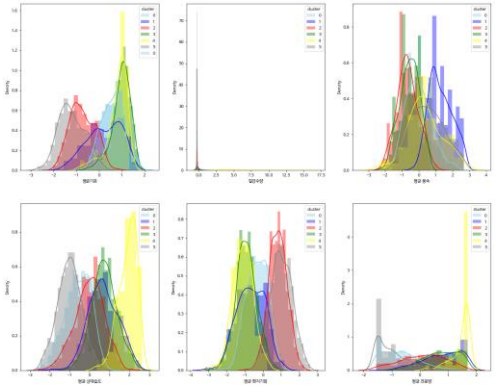
Contents

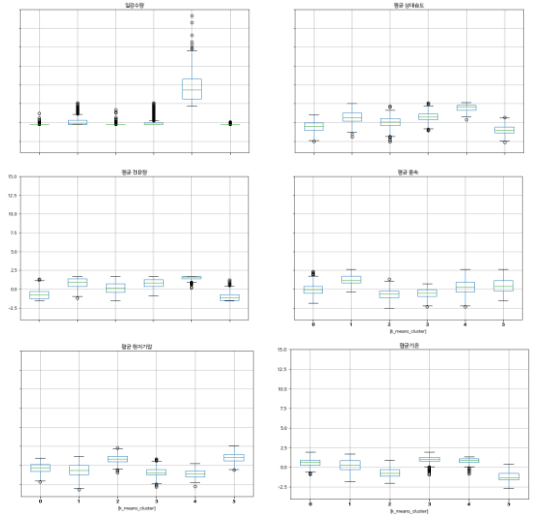
Clustering Analysis

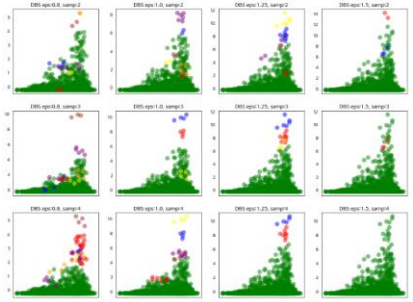
Partitional Clustering(CONT.)

04. EDA/Interpretation(CONT.)









cf) Density-based Clustering(DBSCAN)

Conclusion

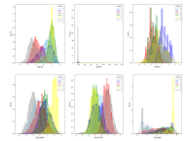
01. Result

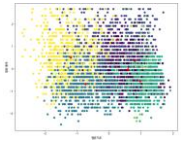
수집한 날씨 데이터에 가장 어울리는 Clustering Method = K-means Clustering

Average Silhouette Score 값이 0.456으로 clustering method 중 가장 높으며, 군집 간의 편차도 적음

이는 시각화 결과에서도 확인할 수 있음

각 군집에 해당하는 데이터 수의 편차 역시 K-means Clustering 방식에서 가장 작음





수집한 날씨 데이터의 Clustering 결과

K-means Clustering을 활용한 결과, 아래와 같이 최근 10년간의 서울 날씨를 군집화 할 수 있음

- Cluster0 : 기온이 높은 편이며 비가 거의 오지 않고 습도와 전운량이 낮음 → 맑고 따뜻한 날
- Cluster1 : 습도와 전운량이 높으며 풍속이 매우 높음. 적은 양의 비가 내림 → 흐리고 바람부는 날
- Cluster2 : 기온이 낮은 편이며 바람이 전혀 불지 않고 기압이 매우 높음 → 선선한 날
- Cluster3 : 기온이 매우 높고 기압은 매우 낮으며 전운량이 매우 높음. 그러나 비는 별로 오지 않음 → 덥고 습한 날
- Cluster4 : Cluster3과 유사하지만 비가 매우 많이 옴 → 비오고 습한 날
- Cluster5 : 기온이 매우 낮으나 비가 전혀 오지 않고 상대습도가 매우 낮으며 전운량이 매우 적음 → 춥고 맑은 날

02. Limitation

(1) Data 관련 한계

→ 이상치에 민감한 군집화 알고리즘이기에 사전에 이상치를 제거해주는데, 수집한 데이터 중 "강수량" 데이터가 대부분 이상치로 인식됨

(2) Clustering 관련 한계(k-means clustering 알고리즘)

- 클러스터 중심의 초기 배치에 민감함. 즉, 결과가 최적이 아닐 수 있음
- 군집이 구형이며 크기가 같다고 가정한 채 진행됨
- threshold, 클러스터 개수 k값을 입력 파라미터로 지정해주었는데, 그 값의 변화에 따라 결과가 크게 달라짐

03. Development

(1) Data 관련 개선점

→ 대부분 이상치로서 존재하는 강우량 데이터를 적절히 대체할 수 있는, 증발산량 등의 변수를 적절히 활용할 수 있으리라 생각함

(2) Clustering 관련 개선점

→ 위의 한계점을 보완하기 위해, 계층적 또는 밀도 기반 클러스터링과 같은 보다 정교한 클러스터링 기술을 통해 보다 우수한 군집화를 진행할 수 있음

Data Source

[데이터 출처]

(1) 기상청(2013.03.29~2023.03.28, 10년 치)
- 일별 종관기상관측(ASOS) 자료 https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36

[참고 문헌]

(1) WEATHER
- https://education.nationalgeographic.org/resource/resource-library-weather/