AIchemist Train Session

CHAP 07 군집화

군집화 실습 목차

01.

데이터 수집 프로세스

공모전 소개

○ 공모개요

빅데이터 활용 미래 사회문제 해결 아이디어 해커톤을 개최하오니 많은 관심과 참여 부탁드립니다.

○ 공모주제

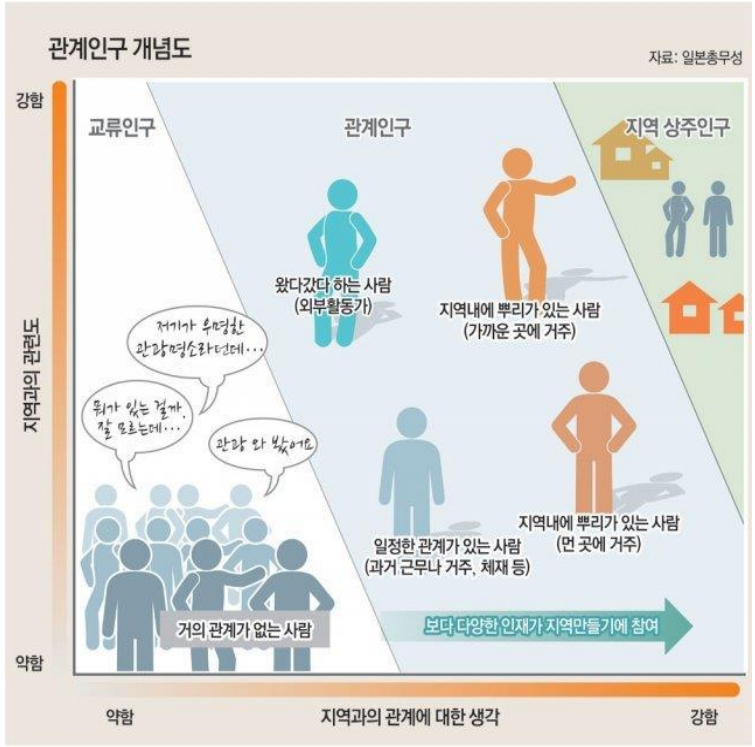
환경, 질병/재난, 도시, 직업 관련 미래 사회문제 예측 및 해결방안 제시

- 환경 | 기상, 기후, 해양, 지구온난화 등
- 질병, 재난 | 질병, 의료서비스, 재난재해 등
- 도시 | 교통, 주택, 보안 등
- 직업 | 신산업 직업전망, 산재예방, 장애인고용 등

○ 공공데이터포털 활용 (DATA.GO.KR)

*타 데이터 및 민간데이터 사용가능

참가 주제 소개



지역 위기를 극복할 방법, 관계 인구 확충

📌 주제

지역 인구 소멸 위기에 대응해
지역 인구 창출을 위한 관광객 패턴
분석 및 '잠재 관광 특구' 중심 사업화

데이터 수집 프로세스

1. 문제 정의

'어떤 데이터로써 이 문제를 해결할 수 있을까?'

2. 데이터 수집

- 공개 데이터셋 활용
- 자체 데이터 수집 (스크래핑 & 크롤링)

3. 데이터 정제 (전처리)

① 데이터 형식 맞추기

- 날짜 표시 형식, 금액 표시 형식 등 모든 데이터를 **일관된 포맷**으로 정리 !

② **Null값** 처리

- 근사값, 평균값, 최빈값 등

4. 연관 데이터 추가해 최종 데이터셋 취합

- *날짜 데이터에 요일, 계절, 날씨 등을 추가*

데이터 수집처

Garbage in, garbage out

데이터의 품질이 높을수록

모델의 성능도 더 우수해진다 !

[국내]

- AI 팩토리 : http://aifactory.space
- 공공데이터포털 : https://www.data.go.kr/datasetsearch
- AI허브 : http://www.aihub.or.kr
- 데이콘 : https://dacon.io
- 보건의료빅데이터개방시스템 : https://opendata.hira.or.kr

[국외]

- 캐글 : https://www.kaggle.com/datasets
- 구글 : https://toolbox.google.com/datasetsearch
- 레딧 : https://www.reddit.com/r/datasets/
- UCI : https://archive.ics.uci.edu/ml/

(오프라인) 데이터 안심구역 - 미개방 데이터, 다양한 분석 툴 제공

https://dsz.kdata.or.kr/svc/main/main.do

은솜이와 세은이의 여정

1. 미개방 빅데이터 사용 - 데이터 안심구역 센터

은솜이와 세은이의 여정

2. 웹 크롤링 - 한국관광 데이터랩

***크롤링**

: 웹 페이지를 그대로 가져와서 거기서 데이터를 추출해 내는 행위.

필요한 데이터가 있는 웹페이지의 구조를 분석하고 파악하여 긁어옵니다

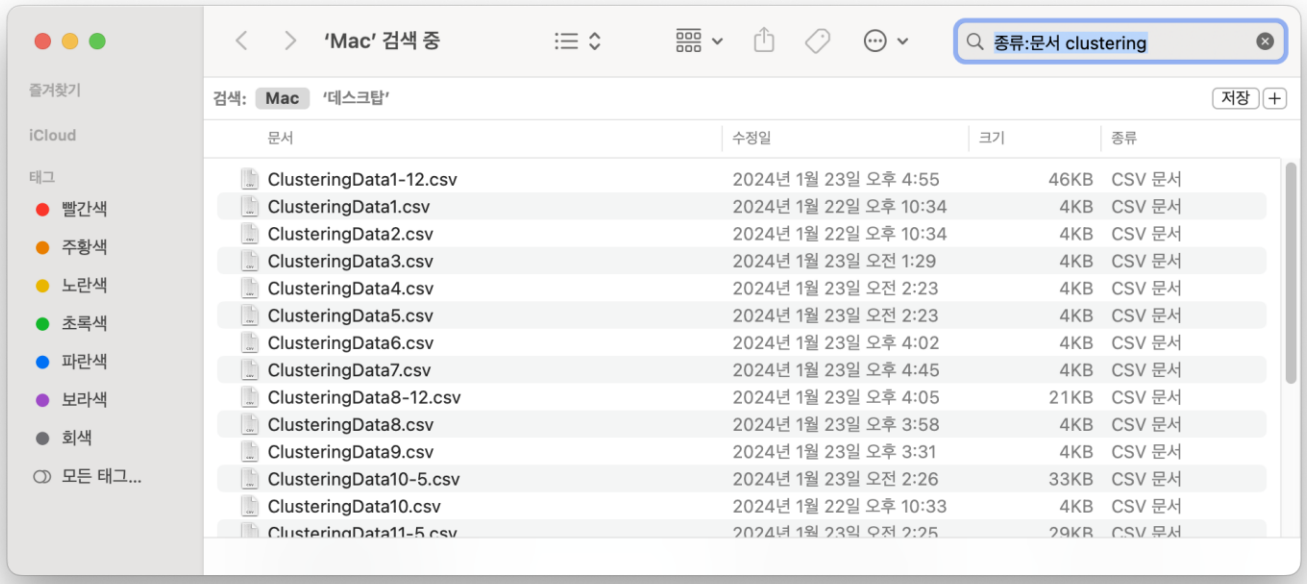
	metrocode		food	shopping	express	leisure	hostel	tour
2			식음료업	쇼핑업	운송업	여가서비스	숙박업	여행업
3								
4	부산	강서구	32.4	26.7	36.3	3	1.5	
5		금정구	57.8	31.1		9.5	1.6	
6		기장군	30.5	58.9		5.2	5.4	
7		남구	60.9	28.7		7.2	3.2	
8		동구	56.7	27.8	0.4	10.7	3.9	0.4
9		동래구	51	42		2.2	4.8	
10		부산진구	45.8	49.7		2.1	2.4	0.1
11		북구	18.7	78.8		1.3	1.2	
12		사상구	51.2	43.9		2.5	2.4	
13		사하구	64.9	28.8		3.6	2.7	
14		서구	78.3	4.3		10.8	6.5	0.1
15		수영구	70.4	26		1.3	2.3	
16		연제구	41.7	55.2		1.7	1.4	
17		영도구	77.2	13.8	0.4	2.9	5.8	
18		중구	49.2	47.4		0.9	2.1	
19		해운대구	26.5	61.5	0.3	1.9	9.9	
20								
21	대구	군위군	67.7	4.3		26.6	1.4	
22		남구	83.6	10.9	0.1	3.3	2.1	
23		달서구	48.9	40.6		7.9	2.5	
24		달성군	78.6	14		5.8	1.7	

02.

관광 데이터 Cluster

관광 데이터 (Clustering Data)

사용 데이터



관광 데이터 (Clustering Data)

1. Data Import

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings("ignore", category=RuntimeWarning)

df = pd.read_csv("./ClusteringData1-12.csv", encoding='cp949')
df.head()
```

Out[1]:

	metrocode	Unnamed: 1	citycode	age0	age1	age2	age3	male	female	food	shopping	express	leisure	hostel	tour	month
0	부산광역시	강서구	2644.0	21.8	41.9	32.0	4.2	65.0	34.9	2.4	26.7	36.3	3.0	1.5	0.0	1
1	부산광역시	금정구	2641.0	23.5	37.0	33.9	5.6	58.5	41.5	57.8	1.1	0.0	9.5	1.6	0.0	1
2	부산광역시	기장군	2671.0	20.8	40.0	32.9	6.2	56.7	43.2	30.5	58.9	0.0	5.2	5.4	0.0	1
3	부산광역시	남구	2629.0	21.9	34.9	36.8	6.4	60.0	40.0	60.9	28.7	0.0	7.2	3.2	0.0	1
4	부산광역시	동구	262.0	26.0	36.3	33.7	3.9	59.4	40.5	56.7	27.8	0.4	10.7	3.9	0.4	1

In [4]:
```
df.drop('Unnamed: 1', axis=1, inplace=True)
df.drop('month', axis=1, inplace=True)
df.drop('metrocode', axis=1, inplace=True)
df.head()
```

Out[4]:

	age0	age1	age2	age3	male	food	shopping	express	leisure	hostel	tour
0	21.8	41.9	32.0	4.2	65.0	2.4	26.7	36.3	3.0	1.5	0.0
1	23.5	37.0	33.9	5.6	58.5	57.8	1.1	0.0	9.5	1.6	0.0
2	20.8	40.0	32.9	6.2	56.7	30.5	58.9	0.0	5.2	5.4	0.0
3	21.9	34.9	36.8	6.4	60.0	60.9	28.7	0.0	7.2	3.2	0.0
4	26.0	36.3	33.7	3.9	59.4	56.7	27.8	0.4	10.7	3.9	0.4

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 612 entries, 0 to 611
Data columns (total 11 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age0      612 non-null    float64
 1   age1      612 non-null    float64
 2   age2      612 non-null    float64
 3   age3      612 non-null    float64
 4   male      612 non-null    float64
 5   food      612 non-null    float64
 6   shopping  612 non-null    float64
 7   express   612 non-null    float64
 8   leisure   612 non-null    float64
 9   hostel    612 non-null    float64
 10  tour      612 non-null    float64
dtypes: float64(11)
memory usage: 52.7 KB
```

관광 데이터 (Clustering Data)

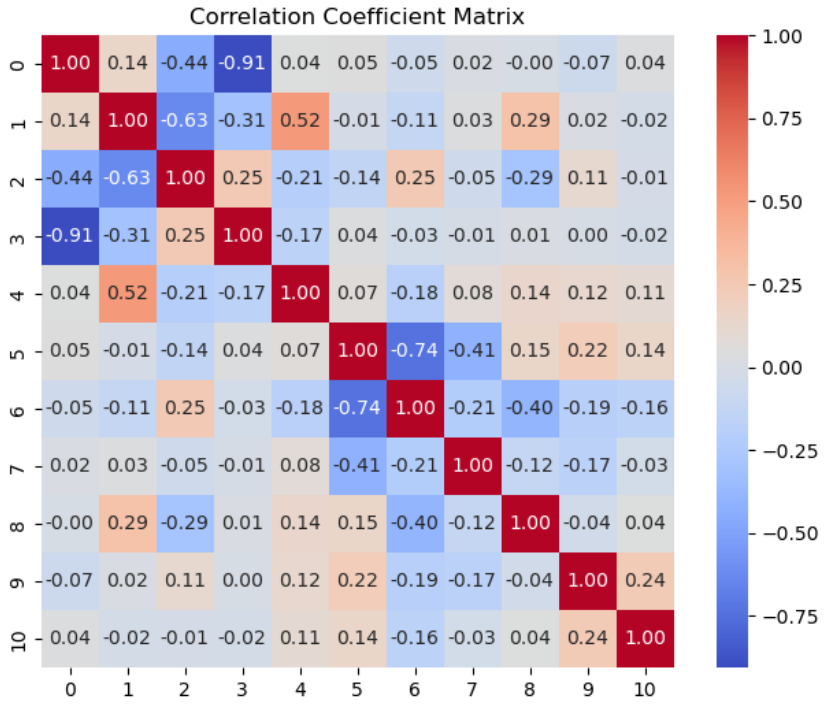
2. Correlation

correlation

In [6]:
```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# 상관계수 행렬 계산
correlation_matrix = np.corrcoef(df, rowvar=False)
```

In [7]:
```
# 히트맵 출력
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", square=True)
plt.title('Correlation Coefficient Matrix')
plt.show()
```

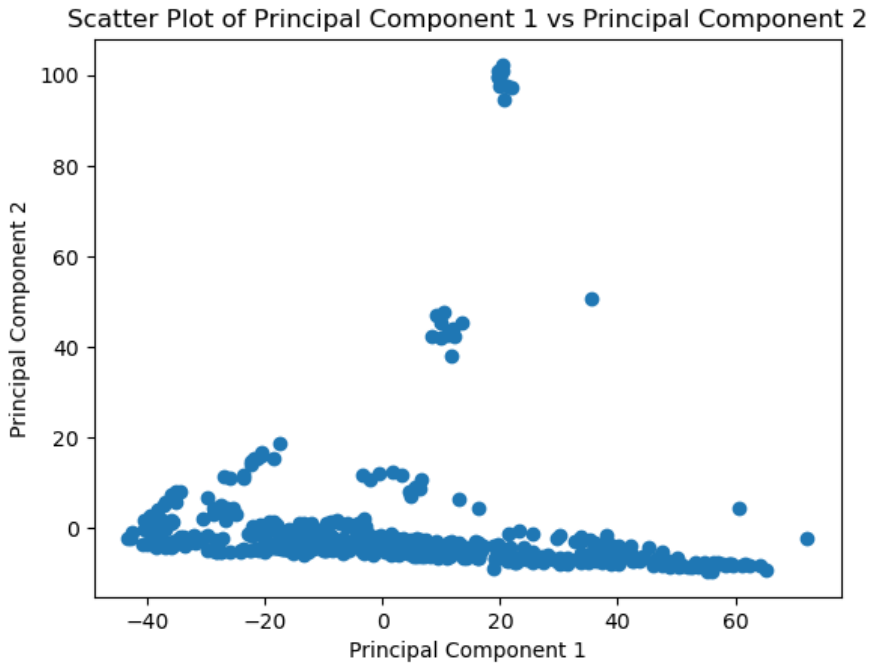


Age0과 age3는 반대되는 상관관계 보임

관광 데이터 (Clustering Data)

3. Scaling

스케일링 전에 전체 데이터 분포 확인



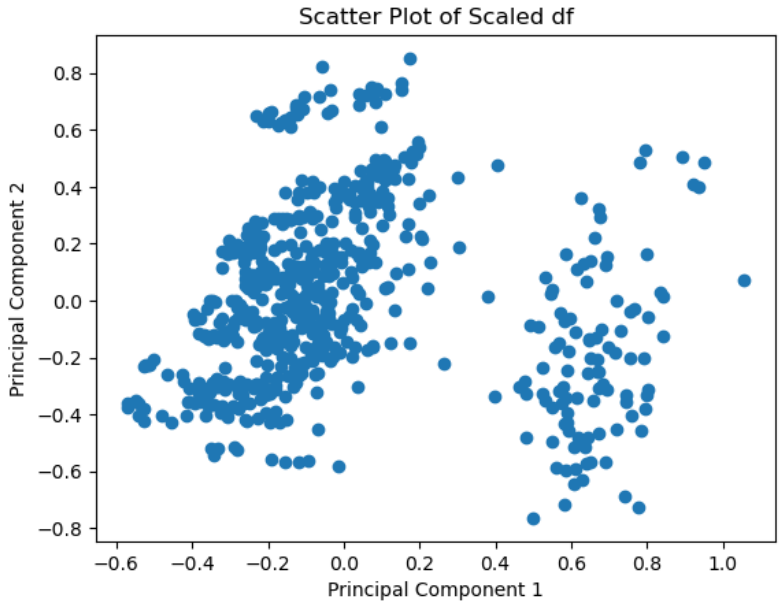

```python
import pandas as pd
from sklearn.preprocessing import StandardScaler, MinMaxScaler

features = ['age0', 'age1', 'age2', 'age3', 'male', 'food', 'shopping','express', 'leisure', 'hostel', 'tour']

# 표준화 (Standardization)
scaler = StandardScaler()
df = scaler.fit_transform(df)
df = pd.DataFrame(scaler.fit_transform(df), columns=features)

# 정규화 (Normalization)
scaler = MinMaxScaler()
df = scaler.fit_transform(df)
df = pd.DataFrame(scaler.fit_transform(df), columns=features)

# 결과 출력
df.head()
```



관광 데이터 (Clustering Data)

4. PCA

2차원으로 시각화 위해 차원축소 n_components = 2

PCA

df_p는 2차원 차원축소한 2개의 칼럼

```
from sklearn.decomposition import PCA # sklearn 라이브러리의 PCA를 import한다

pca = PCA(n_components = 2)           # 2차원으로 시각화를 진행할 것이므로 2개로 설정한다.
pca.fit(df)
df_p = pca.transform(df)
```

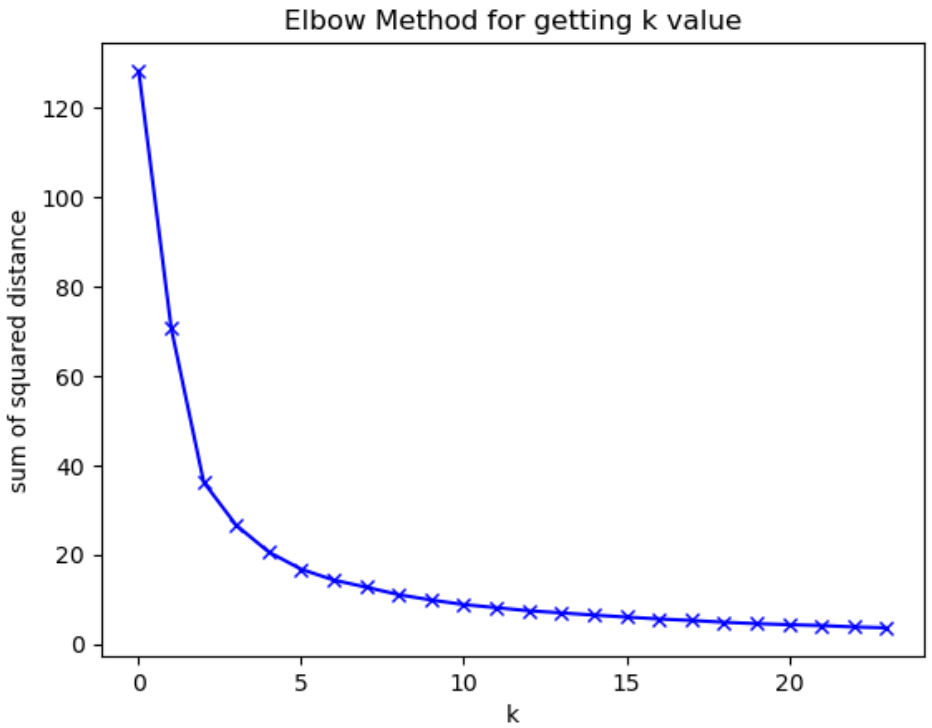
```
df_p = pd.DataFrame(df_p, columns = ['PC1','PC2']) #PCA진행 한 두 개의 값을 column으로 데이터프레임화 시킨다.
df_p.head()
```

	PC1	PC2
0	-0.112863	0.423952
1	-0.230673	-0.210161
2	0.002921	0.374251
3	-0.081792	-0.034416
4	-0.276039	-0.016742

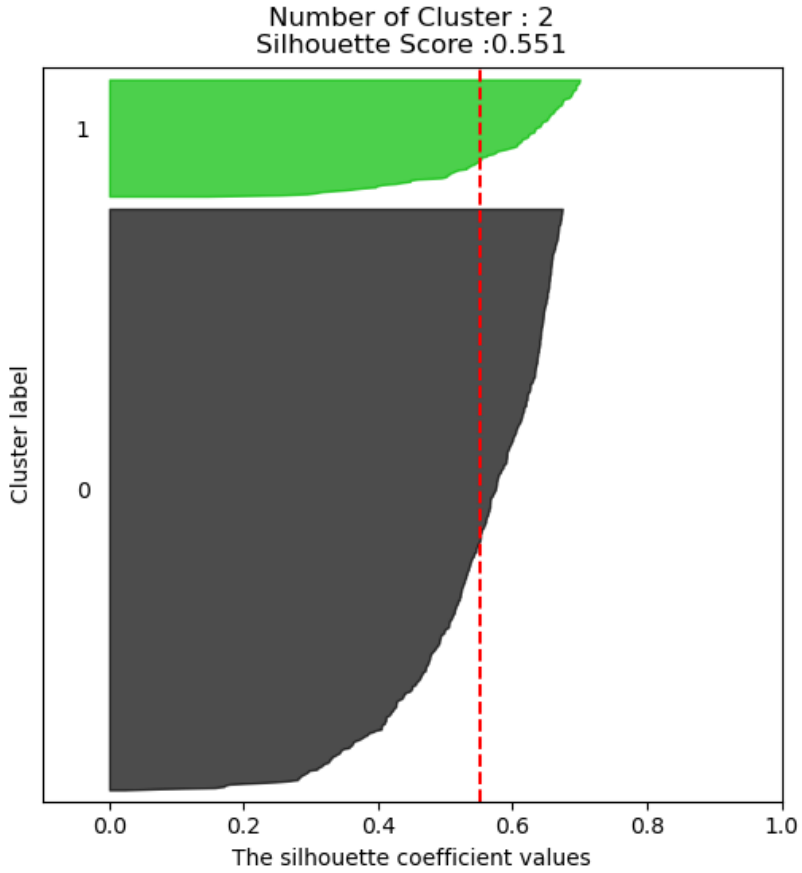
관광 데이터 (Clustering Data)

5. Finding N_clusters

- Elbow 기법으로 n_clusters 개수 정하기



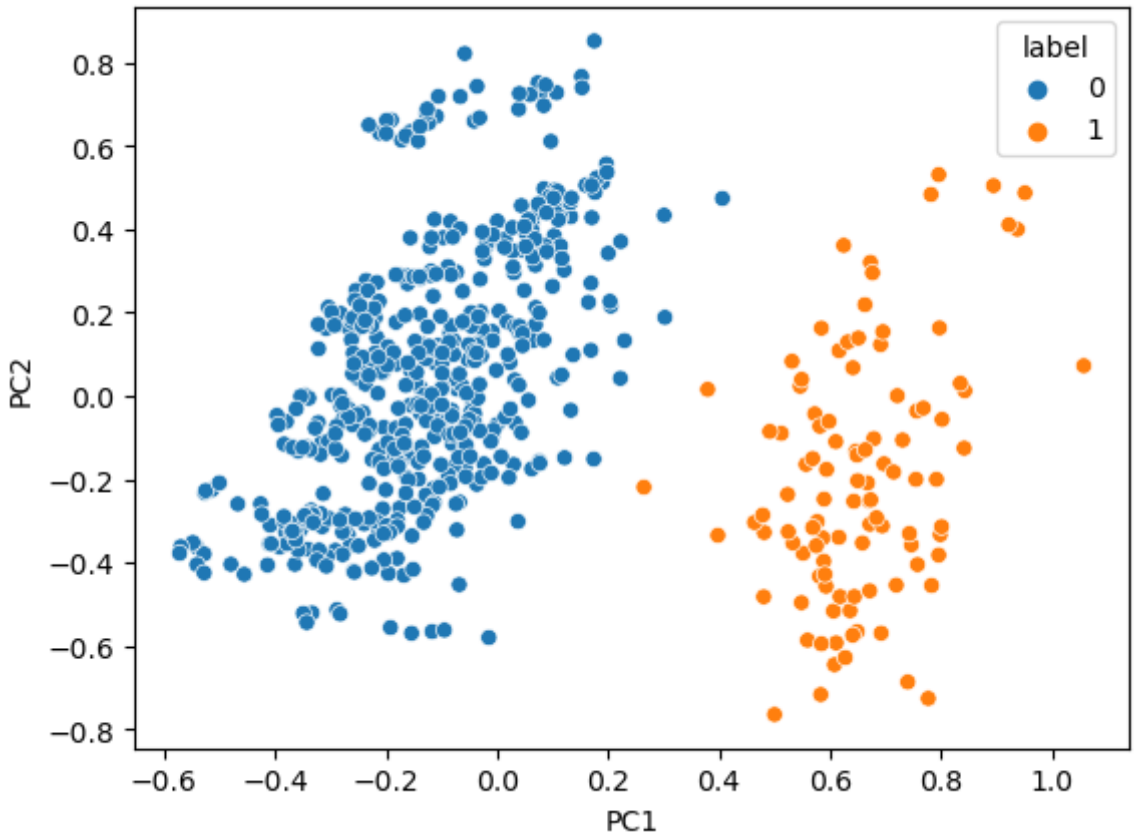
- silhouette 기법으로 n_clusters 개수 정하기



-> N=2

관광 데이터 (Clustering Data)

- **군집화 시각화 결과**

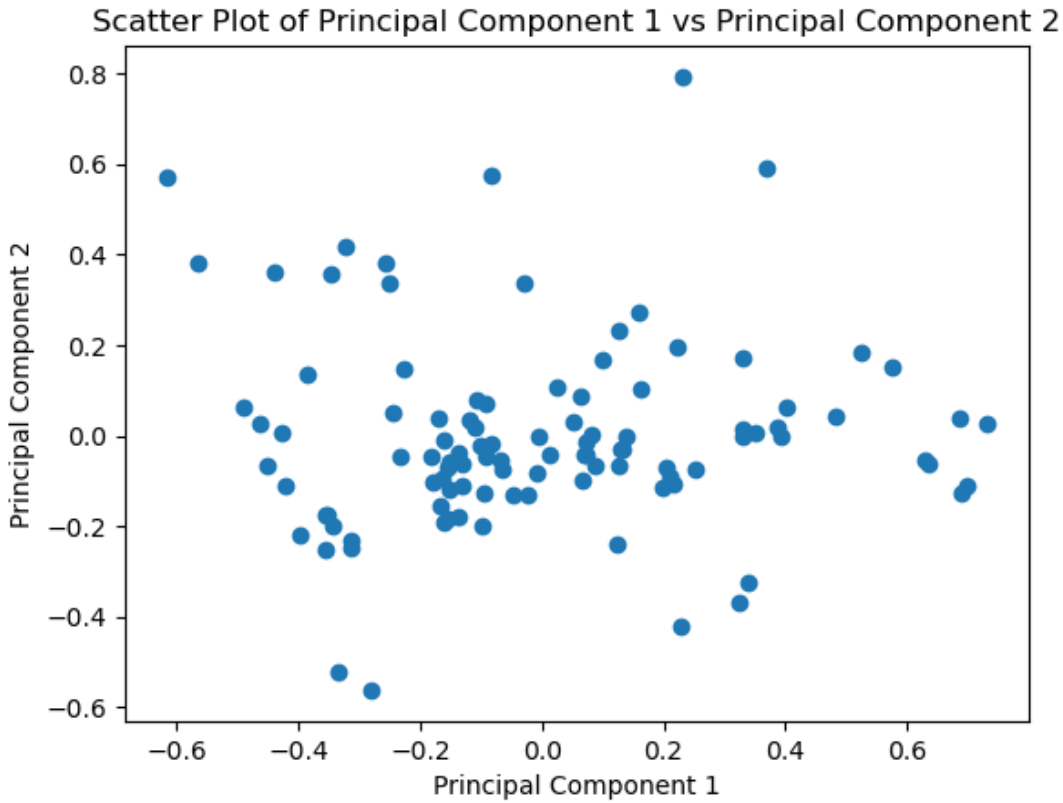


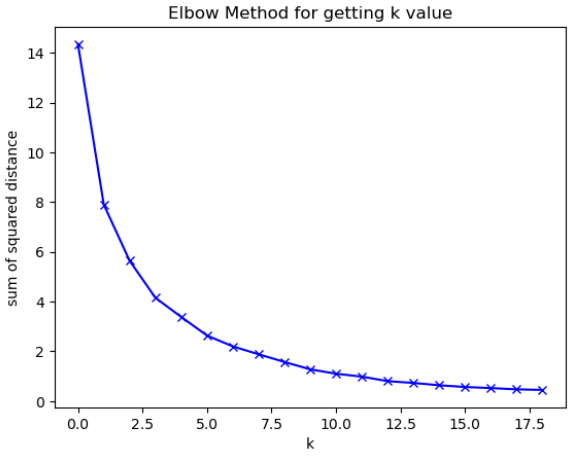
label	age0	age1	age2	age3	male	food	shopping	express	leisure	hostel	tour	PC1	PC2
0	0.655233	0.756086	0.515466	0.101924	0.789530	0.622532	0.375108	0.031114	0.105636	0.267417	0.020629	-0.132213	0.039621
1	0.081821	0.719488	0.597022	0.693019	0.772957	0.643416	0.347592	0.031033	0.117635	0.272446	0.020874	0.653365	-0.195797

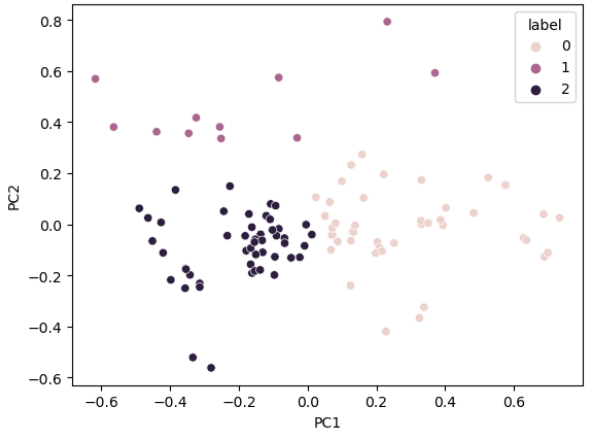
Cluster_mean을 살펴보았을 때 age를 기준으로 구분된 것 확인 가능

관광 데이터 (Clustering Data)

- **군집 0에 대해 군집화**







Process 반복...

관광 데이터 (Clustering Data)

- n=0

df_label_1_0

	age0	age1	age2	age3	male	food	shopping	express	leisure	hostel	tour	PC1	PC2	label	metrocode	Unnamed: 1	month
1	0.010344827586206800	0.625935162094763	0.8206278026905830	0.728937728937729	0.90625	0.3744239631336410	0.23671497584541100	0.43396226415094300	0.072088724584103	0.15929203539823000	0.0	0.1245067410358070	-0.24023578361009000	0	부산광역시	강서구	8
3	0.09655172413793100	0.658354114713217	0.7443946188340810	0.652014652014652	0.7250000000000000	0.41705069124424000	0.5253623188405800	0.0	0.1922365988909430	0.6017699115044250	0.0	0.15829673352234900	0.273759812297852	0	부산광역시	기장군	8
7	0.08620689655172410	0.892768079800499	0.2242152466367710	0.743589743589744	0.7062500000000000	0.6359447004608300	0.4251207729468600	0.0	0.055452865064695	0.29203539823008800	0.1	0.0990204369105066	0.168425552542680	0	부산광역시	부산진구	8
8	0.04137931034482760	0.640897755610973	0.6143497757847530	0.84981684981685	0.8208333333333330	0.2615207373271890	0.856280193236715	0.0	0.0388170055452865	0.1061946902654870	0.0	0.627875765699087	-0.05344785686004070	0	부산광역시	북구	8
13	0.07241379310344820	0.705735660847880	0.4798206278026900	0.8205128205128210	0.7270833333333330	0.5817972350230420	0.5120772946859900	0.003329633740288570	0.04251386321626620	0.1327433628318580	0.0	0.20254326569448200	-0.06932927721649770	0	부산광역시	연제구	8
15	0.03448275862068960	0.700748129675811	0.4618834080717490	0.882783882783883	0.7229166666666670	0.6589861751152070	0.42028985507246400	0.001109877913429530	0.03142329020332720	0.2831858407079650	0.0	0.05021627722971220	0.03253283802147500	0	부산광역시	중구	8
19	0.09999999999999990	0.733167082294264	0.5470852017937220	0.7032967032967040	0.7645833333333330	0.6129032258064520	0.40096618357487900	0.0	0.1534195933456560	0.2035398230088500	0.0	0.07290147497445440	-0.04068725014081910	0	대구광역시	달서구	8
21	0.07931034482758610	0.748129675810474	0.6233183856502240	0.6373626373626370	0.7395833333333330	0.22465437788018400	0.8840579710144930	0.002219755826859050	0.04621072088724580	0.1415929203539820	0.0	0.685819361970235	0.03925372654253240	0	대구광역시	동구	8
22	0.06896551724137930	0.715710723192020	0.6457399103139010	0.6739926739926740	0.7937500000000000	0.6221198156682030	0.4227053140096620	0.001109877913429530	0.10536044362292100	0.17699115044247800	0.0	0.08700923698988750	-0.06762661541008420	0	대구광역시	북구	8

- n=1

df_label_1_1

	age0	age1	age2	age3	male	food	shopping	express	leisure	hostel	tour	PC1	PC2	label	metrocode	Unnamed: 1	month
14	0.06896551724137930	0.730673316708229	0.515695067264574	0.761904761904762	0.8145833333333330	0.8963133640553000	0.09661835748792280	0.005549389567147610	0.044362292051756000	0.7345132743362830	0.0	-0.43939781482007600	0.36223022936136300	1	부산광역시	영도구	8
16	0.13448275862069000	0.900249376558604	0.4035874439461880	0.5347985347985350	0.6479166666666670	0.368663594470046	0.6099033816425120	0.005549389567147610	0.05175600739371530	0.9823008849557520	0.0	0.2312765809736470	0.7929962328410300	1	부산광역시	해운대구	8
24	0.089655172413793	0.745635910224439	0.5919282511210760	0.6483516483516490	0.7395833333333330	0.7430875576036870	0.21739130434782600	0.0	0.12199630314232900	0.6814159292035400	0.0	-0.2518564224511310	0.33583118410505900	1	대구광역시	수성구	8

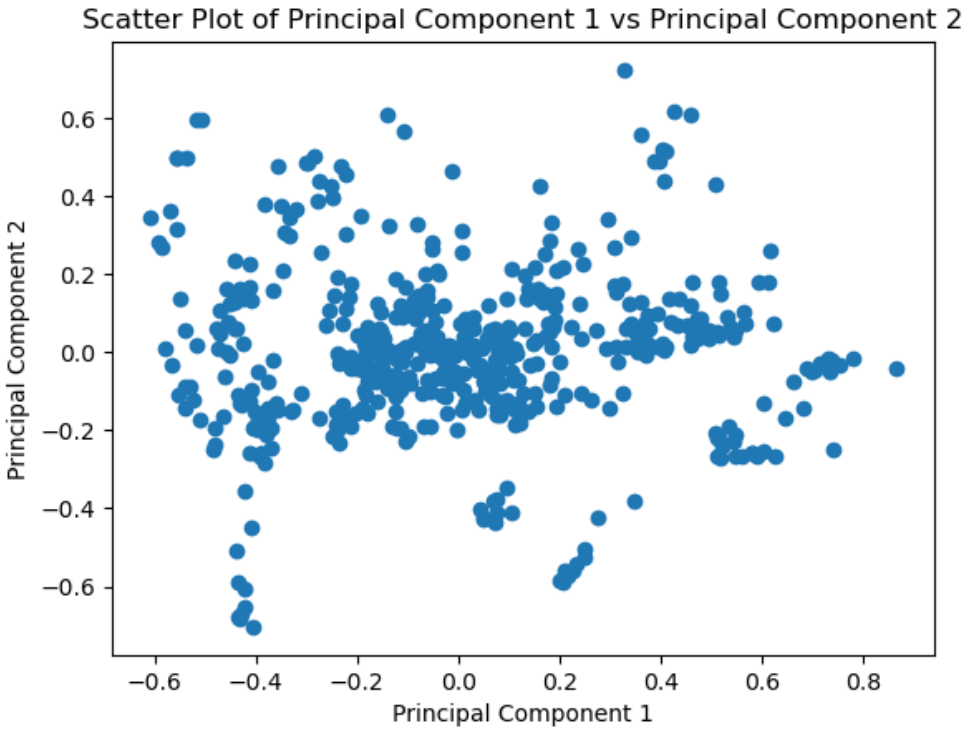
- n=2

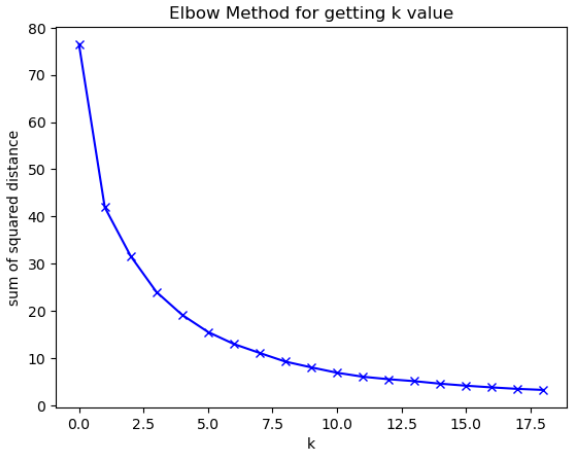
df_label_1_2

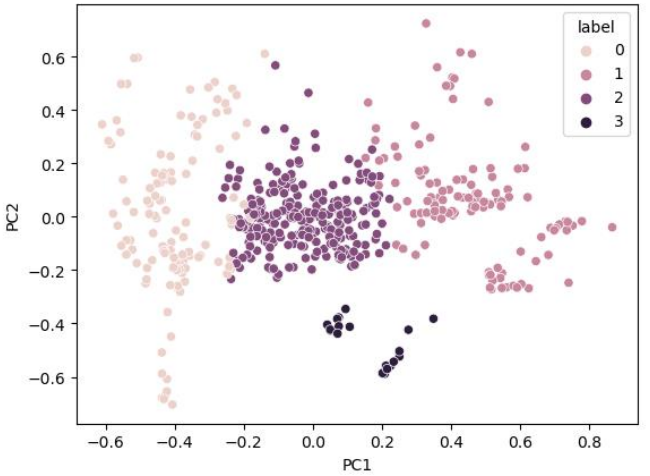
	age0	age1	age2	age3	male	food	shopping	express	leisure	hostel	tour	PC1	PC2	label	metrocode	Unnamed: 1	month
0	0.3793103448275860	0.638403990024938	0.6995515695067270	0.413919413919140	0.8125	0.7235023041474650	0.22584541062801900	0.0	0.24399260628465800	0.19469026548672600	0.0	-0.1309378727728090	-0.10984084882348500	2	세종특별자치시	세종특별자치시	3
2	0.07586206896551720	0.690773067331671	0.5605381165919280	0.7655677655677770	0.7604166666666670	0.7258064516129030	0.251207729468599	0.0	0.2199630314232900	0.1061946902654870	0.0	-0.09694152527954840	-0.19857767467533000	2	부산광역시	금정구	8
4	0.089655172413793	0.783042394014963	0.4573991031390130	0.7106227106227110	0.8187500000000000	0.73963133640553	0.2572463768115940	0.0	0.1497227356746770	0.29203539823008800	0.0	-0.12066948919301500	0.03341752832149720	2	부산광역시	남구	8
5	0.02758620689655160	0.720698254364090	0.4977578475336320	0.8424908424908430	0.7687500000000000	0.7338709677419360	0.213768115942029	0.005549389567147610	0.17190388170055500	0.4690265486725660	0.20000000000000000	-0.22614339092146400	0.1492145777466840	2	부산광역시	동구	8
6	0.07241379310344820	0.600997506234414	0.5426008968609870	0.9304029304029310	0.7458333333333330	0.7131336405529950	0.32608695652173900	0.0	0.05730129390018480	0.42477876106194700	0.0	-0.10768917855428400	0.07960763024635480	2	부산광역시	동래구	8
9	0.02068965517241370	0.700748129675811	0.6098654708520180	0.7765567765567770	0.8645833333333330	0.6808755760368660	0.38647342995169100	0.001109877913429530	0.05914972273567470	0.2212389380530970	0.0	0.012028280118276900	-0.04044200983868770	2	부산광역시	사상구	8
10	0.07931034482758610	0.713216957605985	0.5246636771300450	0.7802197802197800	0.8416666666666670	0.7845622119815670	0.2632850241545890	0.0	0.07948243992606280	0.23893805309734500	0.0	-0.13713193286340300	-0.03928441725887330	2	부산광역시	사하구	8

관광 데이터 (Clustering Data)

- **군집 1에 대해 군집화**







Process 반복...

관광 데이터 (Clustering Data)

- 총 6개의 군집으로 군집화

군집 0에 대한 i=0 (n=6)

```
# 'label'이 0인 행들만 추출하여 새로운 데이터프레임 생성
df_label_0_0 = final_df_0[final_df_0['label'] == 0].copy()

df_label_0_0.head(5)
```

	age0	age1	age2	age3	male	food	shopping	express	leisure	hostel	tour	PC1	PC2	label	metrocode	Unname
13	0.724138	0.750623	0.457399	0.087912	0.825000	0.866359	0.153382	0.004440	0.053604	0.513274	0.0	-0.366826	0.160415	0	부산광역시	영도
23	0.548276	0.740648	0.600897	0.157509	0.750000	0.688940	0.288647	0.000000	0.116451	0.610619	0.0	-0.137749	0.325498	0	대구광역시	수성
33	0.486207	0.683292	0.753363	0.183150	0.889583	0.995392	0.014493	0.001110	0.001848	0.796460	0.1	-0.507073	0.596065	0	인천광역시	옹진
34	0.465517	0.788030	0.708520	0.098901	0.827083	0.571429	0.270531	0.106548	0.088725	0.929204	0.0	-0.108681	0.567215	0	인천광역시	중
40	0.572414	0.815461	0.538117	0.084249	0.847917	0.949309	0.092995	0.000000	0.036969	0.415929	0.0	-0.410898	0.166261	0	대전광역시	대덕

```
df_label_0_0.to_csv("df_label_0_0.csv", index = True)
```

군집 0에 대한 i=1 (n=6)

```
# 'label'이 0인 행들만 추출하여 새로운 데이터프레임 생성
df_label_0_1 = final_df_0[final_df_0['label'] == 1].copy()

df_label_0_1.head(5)
```

	age0	age1	age2	age3	male	food	shopping	express	leisure	hostel	tour	PC1	PC2	label	metrocode	Unnamed
3	0.624138	0.705736	0.614350	0.128205	0.808333	0.678571	0.333333	0.00000	0.133087	0.283186	0.0	-0.044983	0.079459	1	부산광역시	남구
4	0.765517	0.740648	0.475336	0.036630	0.795833	0.630184	0.322464	0.00444	0.197782	0.345133	0.2	-0.103405	0.001938	1	부산광역시	동구
5	0.796552	0.723192	0.385650	0.113553	0.741667	0.564516	0.493961	0.00000	0.040665	0.424779	0.0	0.061054	0.051389	1	부산광역시	동래구
8	0.672414	0.780549	0.493274	0.069597	0.864583	0.566820	0.516908	0.00000	0.046211	0.212389	0.0	0.142164	-0.032172	1	부산광역시	사상구
9	0.668966	0.715711	0.538117	0.124542	0.825000	0.724654	0.334541	0.00000	0.066543	0.238938	0.0	-0.075185	0.020101	1	부산광역시	사하구

군집 0에 대한 i=2 (n=6)

```
In [442]: # 'label'이 0인 행들만 추출하여 새로운 데이터프레임 생성
          df_label_0_2 = final_df_0[final_df_0['label'] == 2].copy()

          df_label_0_2.head(5)
```

Out[442]:

	age0	age1	age2	age3	male	food	shopping	express	leisure	hostel	tour	PC1	PC2	label	metrocode	Unname
1	0.679310	0.758105	0.484305	0.098901	0.777083	0.642857	0.000000	0.00000	0.175601	0.141593	0.00	-0.273281	-0.169257	2	부산광역시	금정
10	0.896552	0.730673	0.313901	0.047619	0.787500	0.879032	0.038647	0.00000	0.199630	0.575221	0.05	-0.539442	0.057566	2	부산광역시	서
16	0.596552	0.952618	0.201794	0.124542	0.797917	0.756912	0.038647	0.00000	0.491682	0.123894	0.00	-0.420690	-0.357455	2	대구광역시	군위
17	0.758621	0.740648	0.452915	0.065934	0.731250	0.940092	0.118357	0.00111	0.060998	0.185841	0.00	-0.374682	-0.076239	2	대구광역시	남
19	0.596552	0.827930	0.412556	0.139194	0.818750	0.882488	0.155797	0.00000	0.107209	0.150442	0.00	-0.310784	-0.105416	2	대구광역시	달성

```
In [443]: df_label_0_2.to_csv("df_label_0_2.csv", index = True)
```

군집 0에 대한 i=3 (n=6)

```
In [444]: # 'label'이 0인 행들만 추출하여 새로운 데이터프레임 생성
          df_label_0_3 = final_df_0[final_df_0['label'] == 3].copy()

          df_label_0_3.head(5)
```

Out[444]:

	age0	age1	age2	age3	male	food	shopping	express	leisure	hostel	tour	PC1	PC2	label	metrocode	Unnam
0	0.620690	0.880299	0.399103	0.047619	0.912500	0.004608	0.309179	0.402886	0.055453	0.132743	0.0	0.348464	-0.382045	3	부산광역시	강
41	0.648276	0.718204	0.565022	0.124542	0.791667	0.054147	0.068841	0.955605	0.001848	0.026549	0.0	0.249773	-0.523821	3	대전광역시	
51	0.634483	0.882793	0.367713	0.054945	0.914583	0.355991	0.266908	0.429523	0.064695	0.141593	0.0	0.094315	-0.345586	3	부산광역시	강
92	0.624138	0.700748	0.587444	0.157509	0.800000	0.059908	0.061594	0.956715	0.001848	0.026549	0.0	0.249918	-0.503619	3	대전광역시	
102	0.655172	0.900249	0.340807	0.029304	0.952083	0.360599	0.257246	0.432852	0.070240	0.132743	0.0	0.074679	-0.377292	3	부산광역시	강

군집 0에 대한 i=4 (n=6)

```
In [446]: df_label_0_4 = final_df_0[final_df_0['label'] == 4].copy()

          df_label_0_4.head(5)
```

Out[446]:

	age0	age1	age2	age3	male	food	shopping	express	leisure	hostel	tour	PC1	PC2	label	metrocode	Unnamed
2	0.586207	0.832918	0.439462	0.120879	0.739583	0.328341	0.698068	0.00000	0.096118	0.477876	0.00	0.358403	0.128153	4	부산광역시	기장
6	0.662069	0.571072	0.798206	0.135531	0.706250	0.504608	0.586957	0.00000	0.038817	0.212389	0.05	0.312187	0.154968	4	부산광역시	부산진
12	0.741379	0.700748	0.461883	0.139194	0.720833	0.457373	0.653382	0.00000	0.031423	0.123894	0.00	0.323929	-0.105684	4	부산광역시	연제
15	0.524138	0.663342	0.753363	0.190476	0.664583	0.282258	0.729469	0.00333	0.035120	0.876106	0.00	0.426197	0.616304	4	부산광역시	해운대
21	0.565517	0.775561	0.556054	0.142857	0.785417	0.529954	0.522947	0.00000	0.107209	0.159292	0.00	0.199904	-0.025408	4	대구광역시	북

```
In [447]: df_label_0_4.to_csv("df_label_0_4.csv", index = True)
```

군집 0에 대한 i=5 (n=6)

```
In [448]: df_label_0_5 = final_df_0[final_df_0['label'] == 5].copy()

          df_label_0_5.head(5)
```

Out[448]:

	age0	age1	age2	age3	male	food	shopping	express	leisure	hostel	tour	PC1	PC2	label	metrocode	Unnamed
7	0.717241	0.758105	0.434978	0.095238	0.812500	0.192396	0.938406	0.00000	0.024030	0.106195	0.00	0.680832	-0.143222	5	부산광역시	북
20	0.548276	0.763092	0.618834	0.124542	0.741667	0.000000	0.952899	0.00222	0.027726	0.115044	0.00	0.865489	-0.039540	5	대구광역시	동
22	0.710345	0.822943	0.363229	0.069597	0.812500	0.437788	0.655797	0.00222	0.046211	0.168142	0.00	0.297114	-0.142855	5	대구광역시	서
28	0.696552	0.845387	0.349776	0.062271	0.875000	0.158986	1.000000	0.00000	0.000000	0.017699	0.05	0.739341	-0.247328	5	인천광역시	동
29	0.579310	0.735661	0.681614	0.080586	0.808333	0.005760	0.765700	0.00000	0.024030	0.150442	0.00	0.729956	-0.017510	5	인천광역시	미추홀

03.

데이터 수집 실습

데이터 수집 프로세스 실습

앞서 소개한 데이터 수집 프로세스에 따라, 아래 3단계의 실습을 진행해봅시다 !

공모전에 본격 돌입하기 전 예행 연습 단계라고 생각해 보시고, 편하게 연습해보세요 ~

각자 문제를 정의하고 적합한 데이터를 서칭 및 수집해보는 과정은 공모전 준비에 많은 도움이 될 거예요 :D

1. 문제 정의 - 해결하고자 하는 문제를 명확히 정의하는 단계

해결하고자 하는 문제가 무엇인지, 이를 위해 필요한 데이터가 무엇일지, 문제 해결의 목표는 무엇일지 고민해보세요 !

예시:

- 문제: 지역 내 교통사고 발생률을 줄이기 위한 방안을 찾고자 함
- 필요 데이터: 교통사고 기록, 도로 정보, 날씨 데이터 등
- 목표: 교통사고 다발 지역 파악 및 사고 요인 분석

데이터 수집 프로세스 실습

2. 적합한 데이터셋 선정 *데이터셋 수집은 7번 슬라이드 참고 :)*

추가로 고려해야 할 조건은 아래와 같습니다.

- **데이터의 정확성 및 신뢰성** (출처 등을 기반으로 믿을만한 데이터가 맞는지 판단해보아야 합니다)
- **데이터의 최신성** (주제에 따라 너무 오래된 자료는 문제가 될 수 있어요 !)
- **데이터 형식 및 구조**

3. 주제 선정 – 수집한 데이터를 분석해 문제 해결 솔루션 짜기

- 모델링 : 데이터셋을 기반으로 문제 해결을 위한 **모델 설계** (예: 회귀 분석, 분류 모델 등)
- 분석 결과 해석 및 **해결 솔루션 제안**

시간 관계상, 3단계는 가상의 계획, 가설을 세우는 정도로 가볍게 진행해주세요 !
*이번 실습에서 가장 중요한 부분은 **2단계의 데이터 수집을 실습**해보는 것이랍니다 :)*

데이터 수집 프로세스 실습

[실습 예시]

1. **문제 정의**

   "지역 내 교통사고 발생률을 줄이기 위한 방안을 찾고자 함"

2. **적합한 데이터셋 선정**

   행정안전부 교통사고 기록 통계, 도로교통공사 고속국도 데이터, 기상청 날씨 정보 등 수집

3. **주제 선정 (간략하게...)**

   특정 지역의 교통사고 다발 원인 분석 -> 날씨 등의 요인에 따른 교통사고 발생 빈도 예측 회귀 모델 설계

   -> 주요 사고 원인 도출 및 사고 예방을 위한 정책 제안

수고하셨습니다