



상권 데이터 분석

DS팀 차수빈 여채운 조혜빈 황선경

목차

#01 주제 소개 & 데이터 출처

#02 데이터 전처리

#03 EDA

#04 상관분석

#05 군집분석



01 주제 소개



#1.1 주제 소개

주제 : 서울시 음식점 상권 데이터 분석



“서울시의 음식점의 분포는
어떤 특징을 가지고 있을까?”

“특정 업종이 특정 행정구에 많다면,
그 이유는 무엇일까?”

“모든 조건이 동일할 때,
어떤 행정구에 음식점을 차리는 것이 유리할까?”



#1.1 주제 소개

주제 : 서울시 음식점 상권 데이터 분석

- 서울시 행정구에 따른 음식점의 분포 양상 분석
- 상관 분석을 통해 업종 별 업소 수에 영향을 미치는 여러가지 요소 파악

요소 : 유동인구(나이별), 인구 소득, 대학 정보, 지하철역분포, 주민등록인구, 초중고등학교, 상권변화지표

-> 총 **7개**의 요소가 음식점 상권에 영향을 미친다고 가정

- 군집 분석을 통해 성격이 비슷한 행정구 군집화

=> 음식점을 창업하고자 하는 사람들이 **창업 지역, 음식 업종**을 선택할 때 도움을 줄 수 있음

#1.2 데이터 출처

#1 **상권정보** 데이터 : <https://www.data.go.kr/data/15083033/fileData.do>

#2 생활인구 데이터: <https://data.seoul.go.kr/dataList/0A-14991/S/1/datasetView.do>

#3 대학정보 데이터: <https://data.seoul.go.kr/dataList/0A-12974/S/1/datasetView.do>

#4 초중고학교정보 데이터: <https://www.data.go.kr/data/15099519/fileData.do>

#5 소득정보 데이터: https://www.bigdata-environment.kr/user/data_market

#6 상권변화지표 데이터: <https://data.seoul.go.kr/dataList/0A-15575/S/1/datasetView.do#>

#7 주민등록인구 데이터: <https://data.seoul.go.kr/dataList/10727/S/2/datasetView.do>

#8 지하철역사정보 데이터: https://data.kric.go.kr/rips/M_01_01/detail.do?id=32

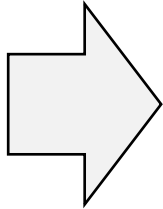
02 데이터 전처리



#2.1 데이터 전처리

서울시 상가정보 데이터
총 39개의 변수
195356개의 데이터

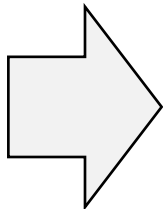
상가업소번호
상호명
지점명
상권업종 대분류명
...
도로명주소
위도



“음식” 상가만 선택
시군구명/행정동명/상호명/중분류명

시군구명	행정동명	상호명	상권업종중분류명
도봉구	방학2동	진미왕족발	한식
강남구	대치4동	죽이야기	한식
서초구	반포3동	터미널포차	유흥주점
마포구	서교동	브루브로스커피	커피점/카페
종로구	종로1.2.3.4가동	커피스미스	커피점/카페

행정구/행정동 단위로 집계한(합계/평균) 최종 데이터
총 36개의 변수
426개의 데이터



시군구명	행정동명	갈비/삼겹살	곱창/양구이전문	기사식당
강남구	개포1동	0	0	0
강남구	개포2동	5	1	0
강남구	개포4동	3	2	0
강남구	논현1동	32	11	0
강남구	논현2동	34	8	0

이와 같은 방식으로 총 8개의 데이터를 전처리 후 하나의 데이터 셋으로 병합

#2.2 최종 데이터

① 행정동 ver : 49개의 변수(행정구 + 행정동 + 10가지 업종 + 요소 37개) & 426개의 데이터(행정동 개수)

	행정구	행정동	갈비/삼겹살	닭/오리 요리	분식	양식	유흥주점	일식/수산물	제과제빵떡케익	커피점/카페	...	폐업점포영업개월	10대 미만	10대	20 ~ 30대	40 ~ 50대	60대	70대 이상	총인구	역(전체)	역(환승역)
0	종로구	청운호자동	4	4	27	34	11	9	12	50	...	62.0	703	1134	3081	4020	1419	1530	11887	0	0
1	종로구	사직동	13	14	51	66	37	41	24	131	...	62.0	532	681	2643	3037	1212	1274	9379	2	0
2	종로구	삼청동	1	0	17	42	10	9	7	95	...	58.0	121	218	646	814	432	428	2659	1	0
3	종로구	부암동	1	3	12	25	6	8	13	54	...	62.0	453	844	2575	3184	1311	1223	9590	0	0

② 행정구 ver : 71개의 변수(행정구 + 34가지 업종 + 요소 37개) & 25개의 데이터(행정구 개수)

	행정구	갈비/삼겹살	곱창/양구이전문	기사식당	기타고기 요리	냉면집	닭/오리 요리	돌솥/비빔밥 전문점	두부요리 전문	버섯전문점	...	폐업점포영업개월	10대 미만	10대	20 ~ 30대	40 ~ 50대	60대	70대 이상	총인구	역(전체)	역(환승역)
0	강남구	301	79	3	65	37	340	7	3	0	...	51	33039	60155	147884	180804	61167	51939	534988	33	18
1	강동구	174	56	1	25	29	270	2	3	1	...	52	33211	39020	130022	147455	67493	47289	464490	12	2
2	강북구	102	56	4	22	13	213	4	3	0	...	52	13243	20362	78696	94561	47245	44515	298622	11	0
3	강서구	188	55	2	36	27	328	4	5	0	...	52	34702	43165	184028	173825	79738	60355	575813	23	8

3. EDA



#3.1 데이터 타입 파악

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 426 entries, 0 to 425
```

```
Data columns (total 72 columns):
```

```
# Column Non-Null Count Dtype
```

```
0 행정구 426 non-null object
1 행정동 426 non-null object
2 갈비/삼겹살 426 non-null int64
3 곱창/양구이전문 426 non-null int64
4 기사식당 426 non-null int64
5 기타고기요리 426 non-null int64
6 냉면집 426 non-null int64
7 닭/오리요리 426 non-null int64
8 돌솥/비빔밥전문점 426 non-null int64
9 두부요리전문 426 non-null int64
10 버섯전문점 426 non-null int64
11 별식/퓨전요리 426 non-null int64
12 보리밥전문 426 non-null int64
13 부대찌개/석어찌개 426 non-null int64
14 부페 426 non-null int64
15 분식 426 non-null int64
16 설렁탕집 426 non-null int64
17 순두부전문 426 non-null int64
18 찜밥전문 426 non-null int64
```

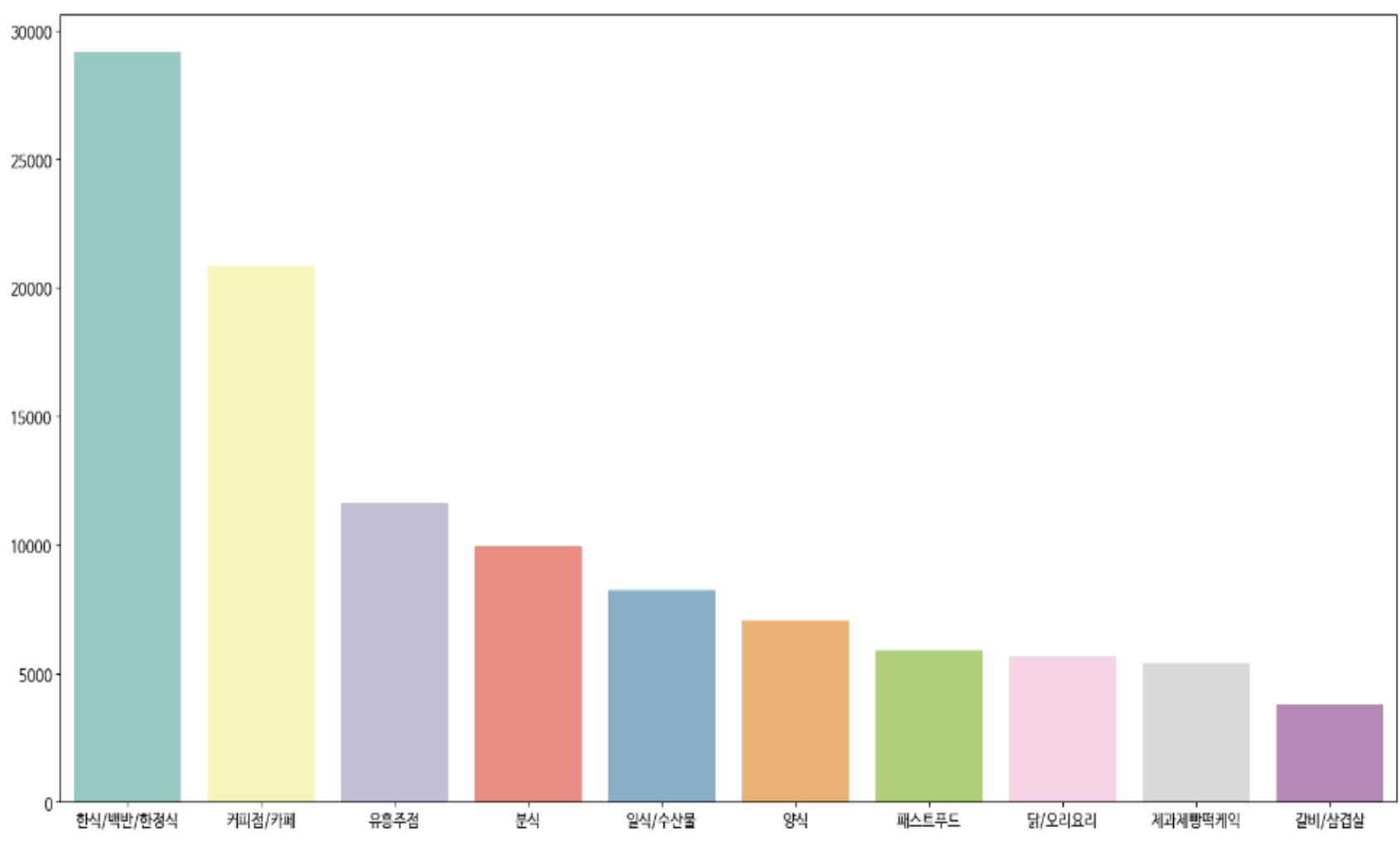
```
19 양식 426 non-null int64
20 유흥주점 426 non-null int64
21 음식배달서비스 426 non-null int64
22 일식/수산물 426 non-null int64
23 재첩국전문 426 non-null int64
24 제과제빵떡케익 426 non-null int64
25 족발/보쌈전문 426 non-null int64
26 중식 426 non-null int64
27 추어탕전문 426 non-null int64
28 커피점/카페 426 non-null int64
29 파전전문 426 non-null int64
30 패스트푸드 426 non-null int64
31 한식/백반/한정식 426 non-null int64
32 한정식전문 426 non-null int64
33 해장국/감자탕 426 non-null int64
34 황태전문 426 non-null int64
35 총생활인구수_평일 426 non-null float64
36 10대 미만 생활인구_평일 426 non-null float64
37 10대 생활인구_평일 426 non-null float64
38 20~30대 생활인구_평일 426 non-null float64
39 40~50대 생활인구_평일 426 non-null float64
40 60대 생활인구_평일 426 non-null float64
41 70대 이상 생활인구_평일 426 non-null float64
42 총생활인구수_주말 426 non-null float64
43 10대 미만 생활인구_주말 426 non-null float64
44 10대 생활인구_주말 426 non-null float64
```

```
45 20~30대 생활인구_수말 426 non-null float64
46 40~50대 생활인구_주말 426 non-null float64
47 60대 생활인구_주말 426 non-null float64
48 70대 이상 생활인구_주말 426 non-null float64
49 대학교(전체) 426 non-null int64
50 일반대학 426 non-null int64
51 온라인대학 426 non-null int64
52 초중고(전체) 426 non-null int64
53 초등학교 426 non-null int64
54 중학교 426 non-null int64
55 고등학교 426 non-null int64
56 총인구수 426 non-null int64
57 총가구수 426 non-null int64
58 가구당인구수 426 non-null float64
59 평균소득금액 426 non-null float64
60 상권변화지표 425 non-null object
61 운영점포영업개월 425 non-null float64
62 폐업점포영업개월 425 non-null float64
63 10대 미만 426 non-null int64
64 10대 426 non-null int64
65 20 ~ 30대 426 non-null int64
66 40 ~ 50대 426 non-null int64
67 60대 426 non-null int64
68 70대 이상 426 non-null int64
69 총인구 426 non-null int64
70 역(전체) 426 non-null int64
71 역(환승역) 426 non-null int64
dtypes: float64(18), int64(51), object(3)
memory usage: 239.8+ KB
```

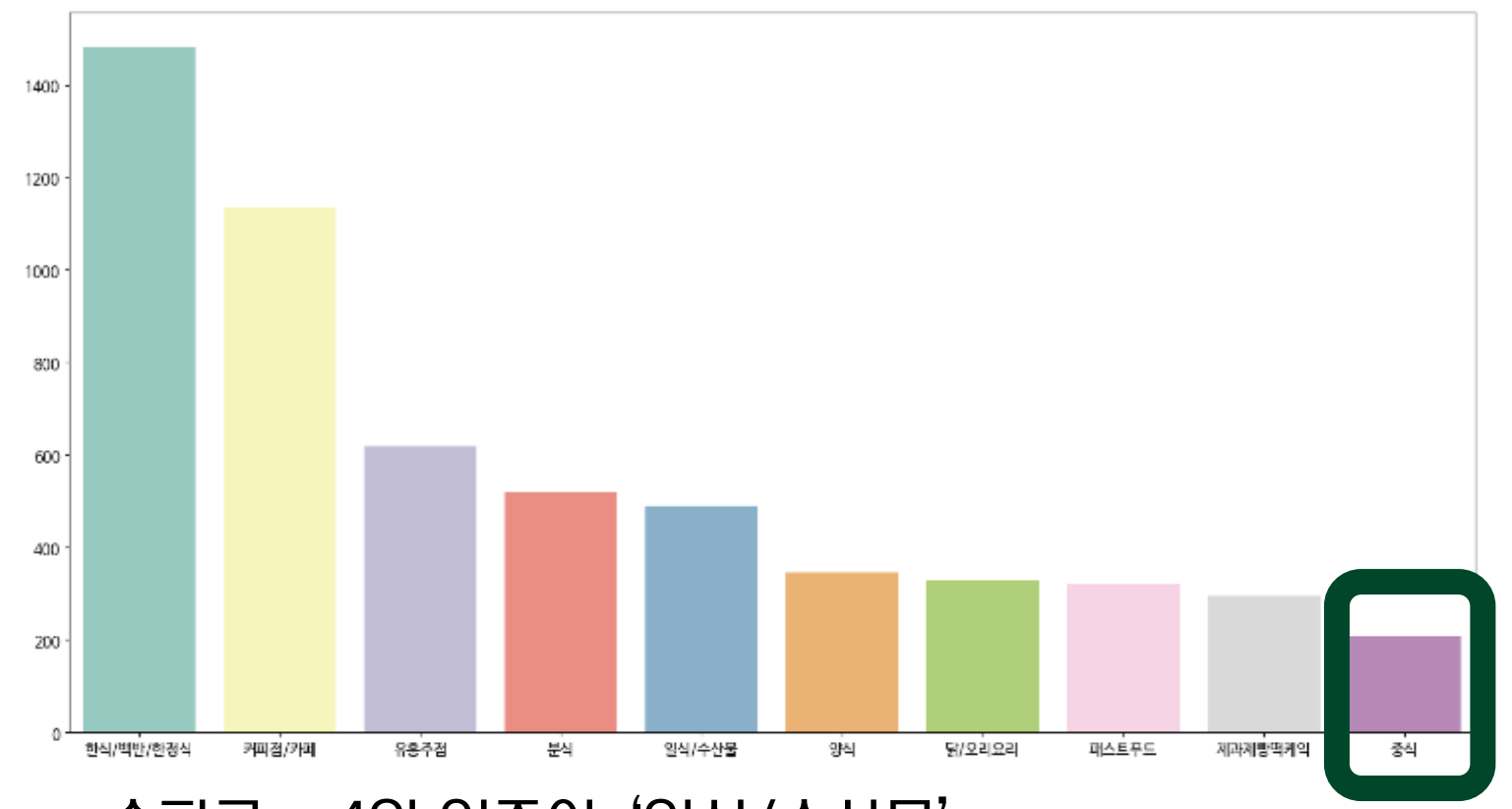
- Null 값 존재 x
- 행정구/행정동: 문자형 데이터 (object)
- 나머지: 숫자형 데이터 (int, float)

#3.2 상위 10개 업종

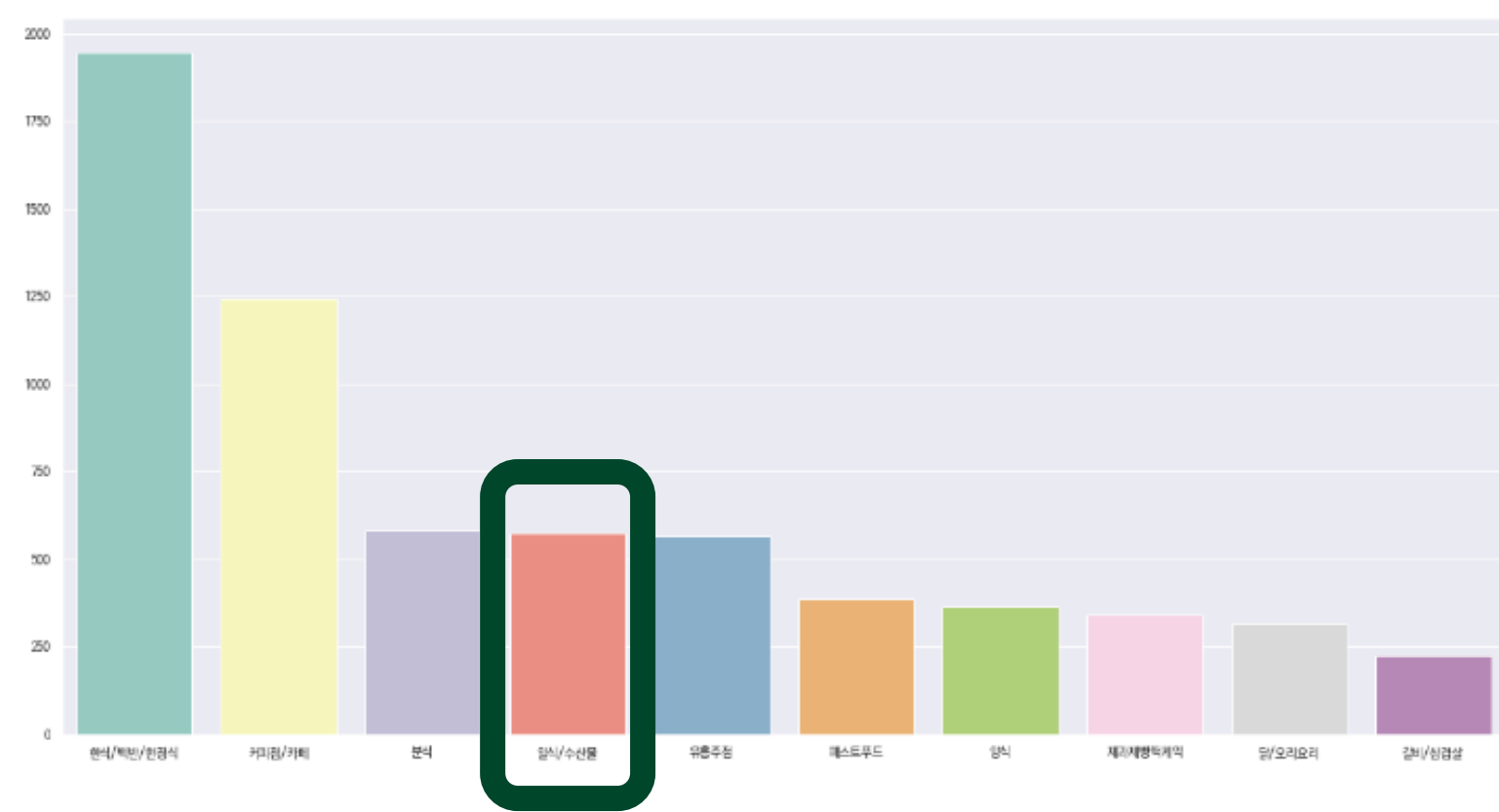
▼ 서울시 전체 업종 분포



▼ 강남구 - 10위 업종이 '중식'



▼ 송파구 - 4위 업종이 '일식/수산물'



- 각 행정구 별로 업종 분포가 대체적으로 서울시 전체 분포와 비슷한 양상을 보임

➔ 서울시 전체 상위 10개 업종에 대한 분석을 수행

#3.3 왜도/스케일 차이

1. 왜도(skew) 확인

- 전체 왜곡 정도: 1.018088

- 각 변수 별 왜도

재첩국전문	4.694855
부대찌개/섞어찌개	2.965599
양식	2.595050
일식/수산물	2.416808
평균소득금액	2.257700
...	
총인구	0.055246
두부요리전문	0.050104
60대	0.039582
40 ~ 50대	0.026885
총가구수	0.011622
Length: 69, dtype: float64	

- 대부분의 feature들이 왜곡된 분포를 가지고 있음을 확인할 수 있음

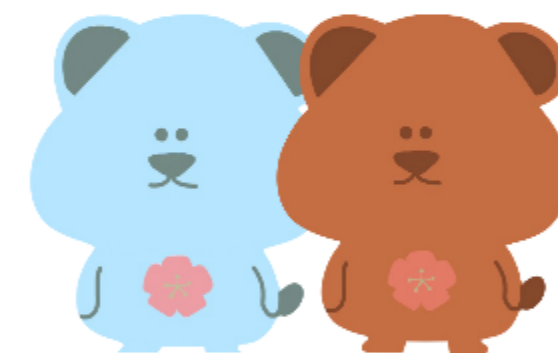
2. df.describe()을 통한 스케일 차이 확인

	갈비/삼겹살	곱창/양구이전문	기사식당	기타고기요리	냉면집	닭/오리요리	돌솥/비빔밥전문점	두부요리전문	버섯전문점	별식/퓨전요리	...
count	25.000000	25.000000	25.000000	25.000000	25.000000	25.000000	25.000000	25.000000	25.000000	25.000000	...
mean	151.080000	47.720000	2.560000	26.12000	19.400000	227.040000	2.920000	3.680000	0.320000	104.680000	...
std	47.416523	15.504085	2.501333	11.21502	6.244998	56.018211	1.525341	2.014944	0.556776	34.985378	...
min	95.000000	25.000000	0.000000	13.00000	11.000000	128.000000	0.000000	0.000000	0.000000	57.000000	...
25%	116.000000	35.000000	1.000000	20.00000	15.000000	187.000000	2.000000	3.000000	0.000000	82.000000	...
50%	142.000000	45.000000	2.000000	23.00000	19.000000	215.000000	3.000000	4.000000	0.000000	95.000000	...
75%	174.000000	56.000000	3.000000	31.00000	22.000000	259.000000	4.000000	5.000000	1.000000	118.000000	...
max	301.000000	80.000000	9.000000	65.00000	37.000000	340.000000	7.000000	8.000000	2.000000	224.000000	...
...											
	10대 미만	10대	20 ~ 30대	40 ~ 50대	60대	70대 이상	총인구	역(전체)	역(환승역)		
	25.000000	25.000000	25.000000	25.000000	25.000000	25.000000	25.000000	25.000000	25.000000		
	12627.440000	30468.000000	117881.280000	121178.440000	52243.640000	42886.240000	387285.040000	15.840000	7.040000		
	9279.298842	14173.093049	39901.123404	40911.117016	16830.810388	12401.649794	125804.872677	6.786506	5.012318		
	6520.000000	6938.000000	43092.000000	39138.000000	18495.000000	16581.000000	130764.000000	4.000000	0.000000		
	17359.000000	23372.000000	92681.000000	95409.000000	41630.000000	37473.000000	314884.000000	11.000000	4.000000		
	12140.000000	27057.000000	119016.000000	122383.000000	51451.000000	44392.000000	392628.000000	15.000000	6.000000		
	17745.000000	39020.000000	132113.000000	147455.000000	61301.000000	51351.000000	464490.000000	21.000000	10.000000		
	17196.000000	60155.000000	204193.000000	212357.000000	88693.000000	62360.000000	665847.000000	33.000000	18.000000		

- Feature간의 스케일 차이가 큼

→ 군집분석 시 데이터 변환과 데이터 스케일링을 진행

4. 상관 분석



#4.1 업종&변수 간 상관 계수

▼ 상관계수의 절댓값이 0.3 이상인 변수만 추출

	업종	변수	상관계수
0	갈비/삼겹살	20~30대 생활인구_평일	0.743052
1	갈비/삼겹살	20~30대 생활인구_주말	0.742392
2	갈비/삼겹살	총생활인구수_평일	0.663073
3	갈비/삼겹살	40~50대 생활인구_평일	0.635479
4	갈비/삼겹살	총생활인구수_주말	0.608314
...
160	한식/백반/한정식	역(환승역)	0.352641
161	한식/백반/한정식	총가구수	0.335675
162	한식/백반/한정식	70대 이상 생활인구_주말	0.330892
163	한식/백반/한정식	평균소득금액	0.323504
164	한식/백반/한정식	가구당인구수	-0.388263

→ 대체로 **생활 인구 데이터**의 변수들이
각 업종 개수와 높은 상관 관계를 가짐

▼ 행정구 & 업종 별 상관 계수가 높은 변수 개수

	갈비/삼겹살	양식	한식/백반/한정식	...	일식/수산물
강남구	21/36	17/36	21/36		20/36
...					
마포구	22/36	22/36	23/36		22/36

```
result.groupby('업종')['변수'].count()
```

업종
갈비/삼겹살 14
닭/오리요리 20
분식 19
양식 10
유흥주점 15
일식/수산물 17
제과제빵떡케익 19
커피점/카페 15
패스트푸드 20
한식/백반/한정식 16
Name: 변수, dtype: int64

→ 총 36개의 변수 중
해당 개수만큼 추출됨

#4.2 heatmap

▼ 업종&변수 간의 상관 관계를 나타내는 heatmap (강남구)



변수들 간의 상관도가 높음
→ PCA로 변수들의 차원을 축소하여 군집 분석 수행

5. 군집분석



#5.1 행정구 단위 분석

- 상위 10개 업종에 대한 군집분석 수행
 - 갈비/삼겹살, 닭/오리요리, 분식, 양식, 유흥주점, 일식/수산물, 제과/빵/떡/케익, 커피점/카페, 패스트푸드, 한식/백반/한정식
 - 군집화 방법: K-Means, 평균이동. 병합군집, GMM, 베이지가우시안, DBSCAN
- 이중 군집화가 비교적 잘 된 **한식/백반/한정식**과 **일식/수산물**에 대해 다양한 방법으로 군집화 재수행
 - i) 변수 선택: 전체 변수 중 상관계수가 높은 10개 변수 선택 vs 모든 변수를 PCA로 차원 축소 한 후 이용
 - ii) 데이터 변환: 로그변환, Box-Cox 변환
 - iii) 데이터 스케일링: 표준화(StandardScaler), 정규화(MinMaxScaler), 표준정규화(RobustScaler)

Ver 1	Ver 2	Ver 3	Ver 4	Ver 5	Ver 6	Ver 7	Ver 8
상위 10개 변수	상위 10개 변수	전체 변수 PCA	전체 변수 PCA	전체 변수 PCA	전체 변수 PCA	전체 변수 PCA	전체 변수 PCA
로그 변환	Box-Cox	로그 변환	Box-Cox	로그 변환	Box-Cox	로그변환	Box-Cox
표준화	표준화	표준화	표준화	정규화	정규화	표준정규화	표준정규화

#5.1 행정구 단위 분석

- 군집 평가 지표

i) 실루엣 계수

- 각 데이터 별로 해당 데이터가 속한 군내의 유사도와 인접한 군의 유사도를 비교하는 지표
- -1 부터 1 사이의 값을 가지며 1에 가까울수록 최적화된 군집

- a 를 자료점과 그 자료점과 같은 그룹의 나머지 자료간의 평균 거리,
- b 를 자료점과 그 자료점이 속하지 않은 가장 가까운 그룹의 자료간의 평균 거리,

$$|s| = \frac{b - a}{\max(a, b)}$$

ii) DBI(Davies Bouldin Index)

- 군집 내에서의 분포와 비교하여 다른 군집 간의 분리 정도의 비율로 계산되는 값
- 모든 두 개의 군집 쌍에 대해 각 군집의 크기의 합을 각 군집의 중심 간 거리로 나눈 값으로 표현되는 함수
- 값이 작을수록 최적화된 군집

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$
$$DB = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} R_{ij}$$

C_i : 특정 군집

C_j : C_i 와 가장 유사한 군집

s_i : C_i 의 평균 크기 (군집 내에 속한 데이터와 군집 중심 간의 평균 거리)

d_{ij} : C_i, C_j 간의 중심 거리

#5.1 행정구 단위 분석

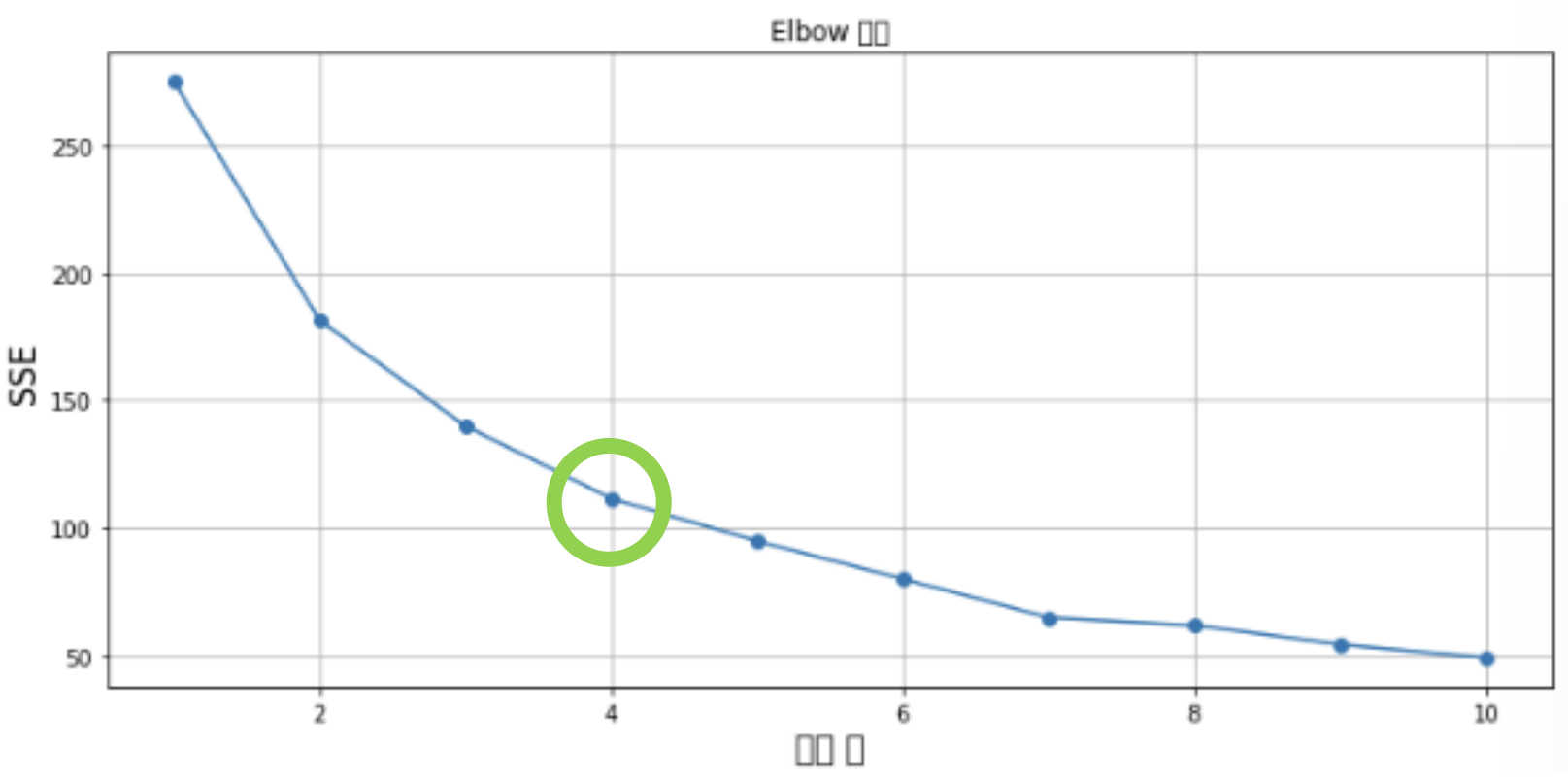
- 한식/백반/한정식 – 실루엣 계수, DBI

	Ver 1		Ver 2		Ver 3		Ver 4		Ver 5		Ver 6		Ver 7		Ver 8	
GMM	-	1.097	-	0.899	-	0.874	-	1.23	-	0.956	-	0.866	-	0.924	-	0.913
베이지스 가우시안	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DBSCAN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
평균 이동	0.278	0.927	0.358	0.994	0.4	0.371	0.385	0.753	0.386	0.689	0.375	0.782	0.428	0.691	0.421	0.682
K-Means	0.226	1.044	0.357	0.819	0.302	1.806	0.293	1.112	0.243	0.830	0.248	0.807	0.328	1.011	0.352	0.965
병합 군집	0.205	1.075	0.224	1.036	0.249	1.086	0.219	1.065	0.199	0.836	0.244	0.832	0.282	0.905	0.269	1.014

-> 대략적인 지표들과 시각화 결과를 고려했을 때 Ver 2(상위 10개 변수/Box-Cox변환/표준화)의 K-Means에서 최적의 군집화 결과를 보임

#5.1 행정구 단위 분석

- 한식/백반/한정식 – Kmeans 군집화

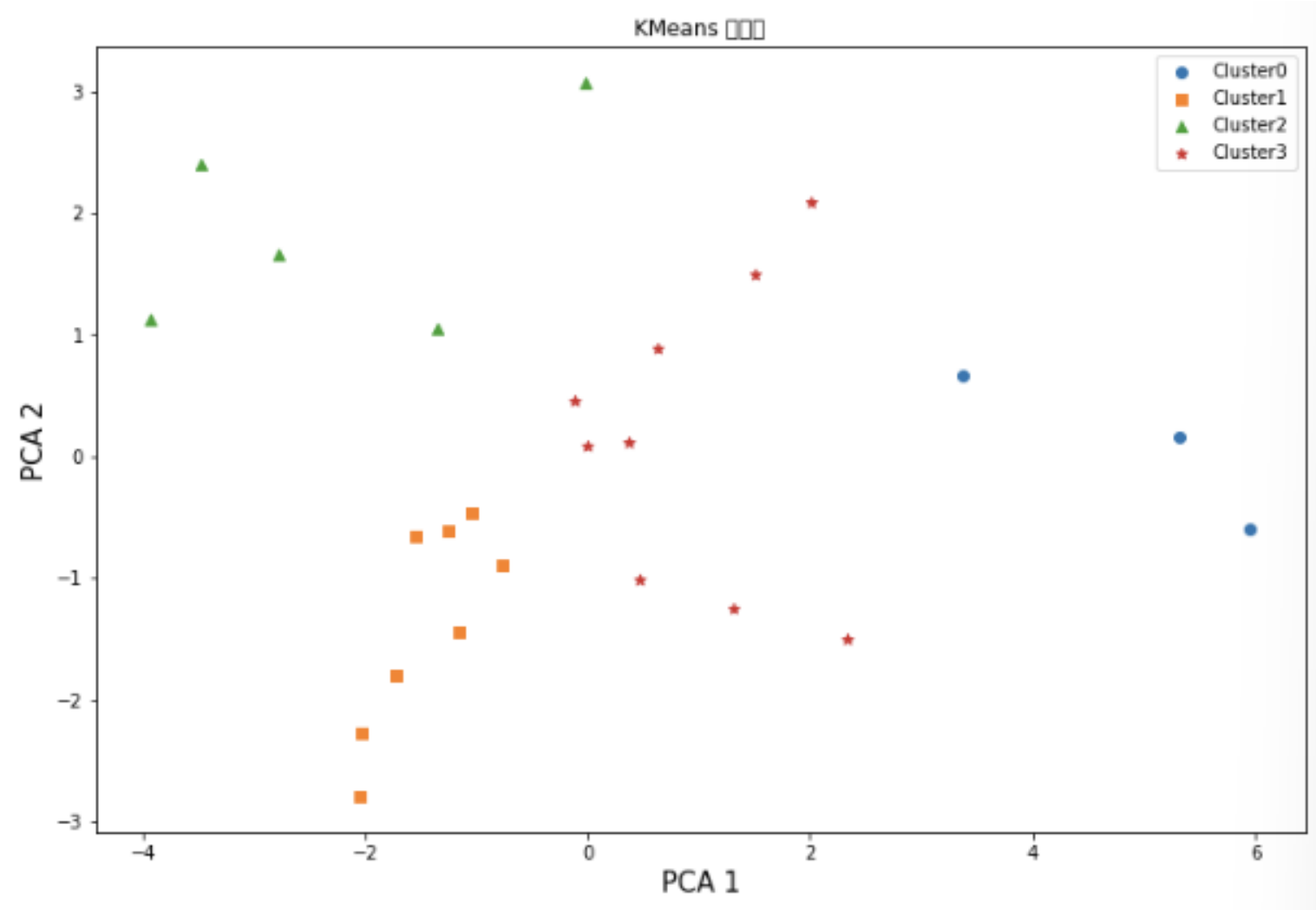


→ k = 4일 때 급격한 거리의 변화가 일어난 것으로 보아 클러스터 수로 4개가 적절해 보인다.

▲ Elbow Method

```
cluster
0      3
1      8
2      5
3      9
Name: 행정구, dtype: int64
```

→ 군집별 행정구 개수 확인



▲ 군집화 결과 시각화

#5.1 행정구 단위 분석

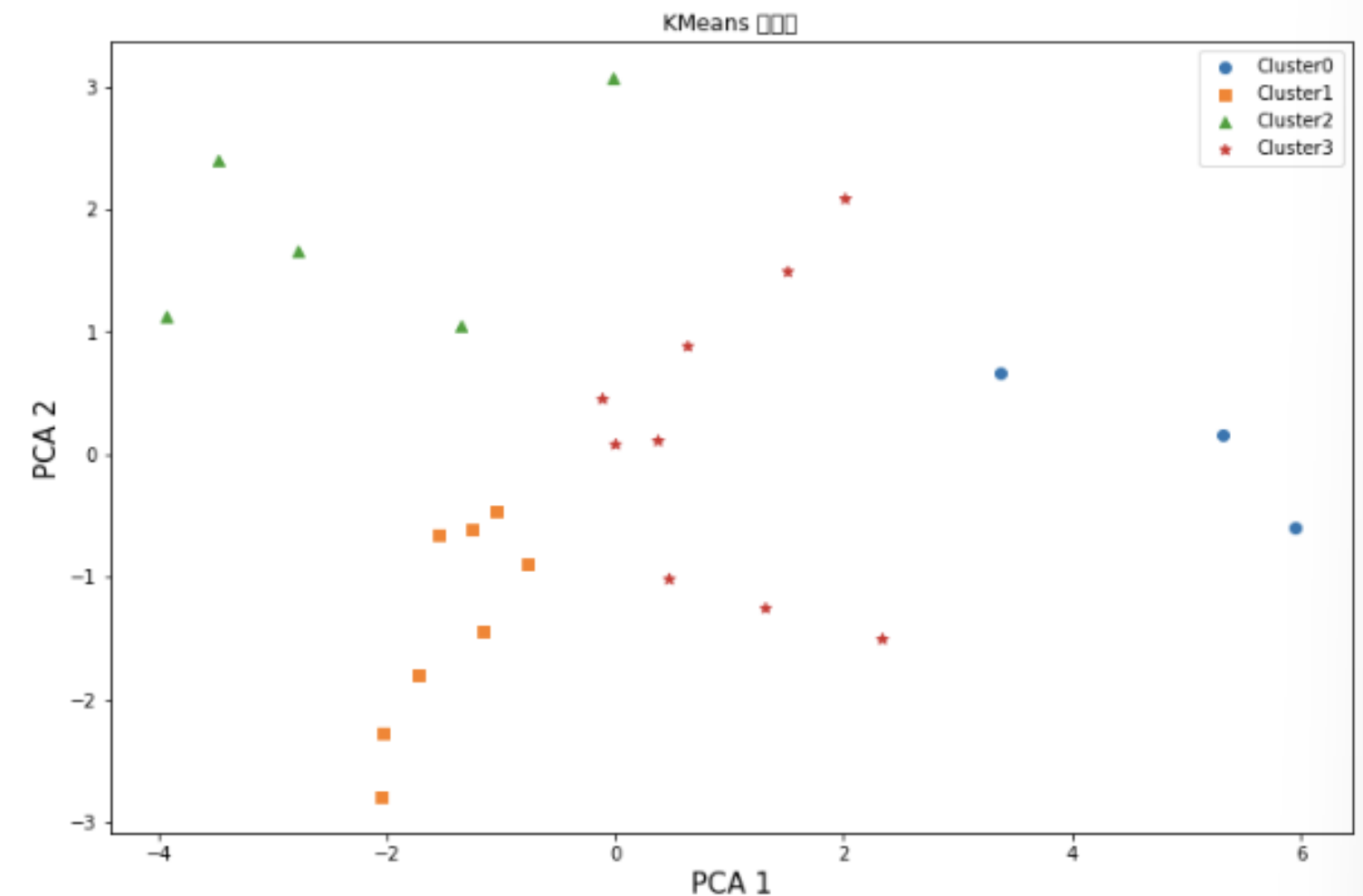
- 한식/백반/한정식 – Kmeans 군집화

▼ 군집화 결과

- 군집0: [강남구, 강서구, 송파구]
- 군집1: [광진구, 동대문구, 동작구, 서대문구, 성동구, 용산구, 종로구, 중구]
- 군집2: [강북구, 금천구, 도봉구, 양천구, 중랑구]
- 군집3: [강동구, 관악구, 구로구, 노원구, 마포구, 서초구, 성북구, 영등포구, 은평구]

▼ 군집 별 특징

- 군집0
 - 전체 지하철역 수가 가장 많은 지역들이다.
- 군집1
 - 10대 미만, 10대 주민등록인구가 적은 지역들이다.
- 군집2
 - 평균소득금액이 비교적 낮은 지역들이다.
- 군집3
 - 40~50대 평일 생활인구가 많은 지역들이다.



#5.1 행정구 단위 분석

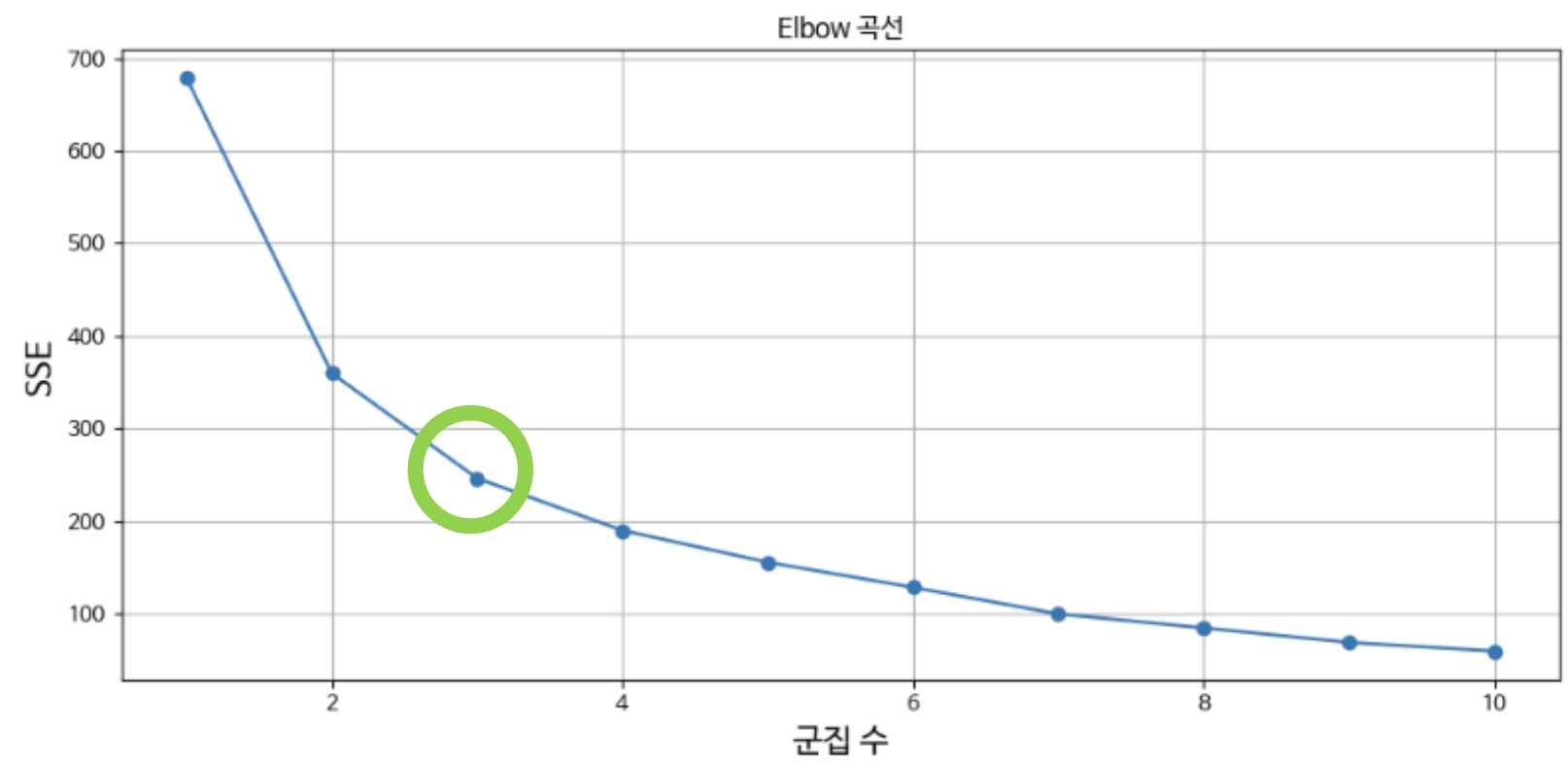
▼ 일식/수산물 – 실루엣 계수, DBI

	Ver 1		Ver 2		Ver 3		Ver 4		Ver 5		Ver 6		Ver 7		Ver 8	
GMM	-	1.110	-	1.025	-	1.116	-	1.198	-	0.954	-	1.098	-	1.077	-	1.025
베이지스 가우시안	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DBSCAN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
평균 이동	0.377	0.912	0.349	0.812	0.403	0.726	0.386	0.845	0.389	0.685	0.375	0.779	0.433	0.685	0.423	0.677
K-Means	0.239	1.309	0.344	0.990	0.289	1.084	0.386	1.845	0.302	0.969	0.256	0.976	0.324	0.985	0.282	0.974
병합 군집	0.186	1.139	0.246	1.116	0.256	1.072	0.220	1.060	0.246	0.826	0.251	0.804	0.292	0.932	0.323	0.952

-> 대략적인 지표들과 시각화 결과를 고려했을 때 Ver 7(전체 변수 PCA/로그변환/표준정규화)의 K-Means에서 최적의 군집화 결과를 보임

#5.1 행정구 단위 분석

- 일식/수산물 – Kmeans 군집화

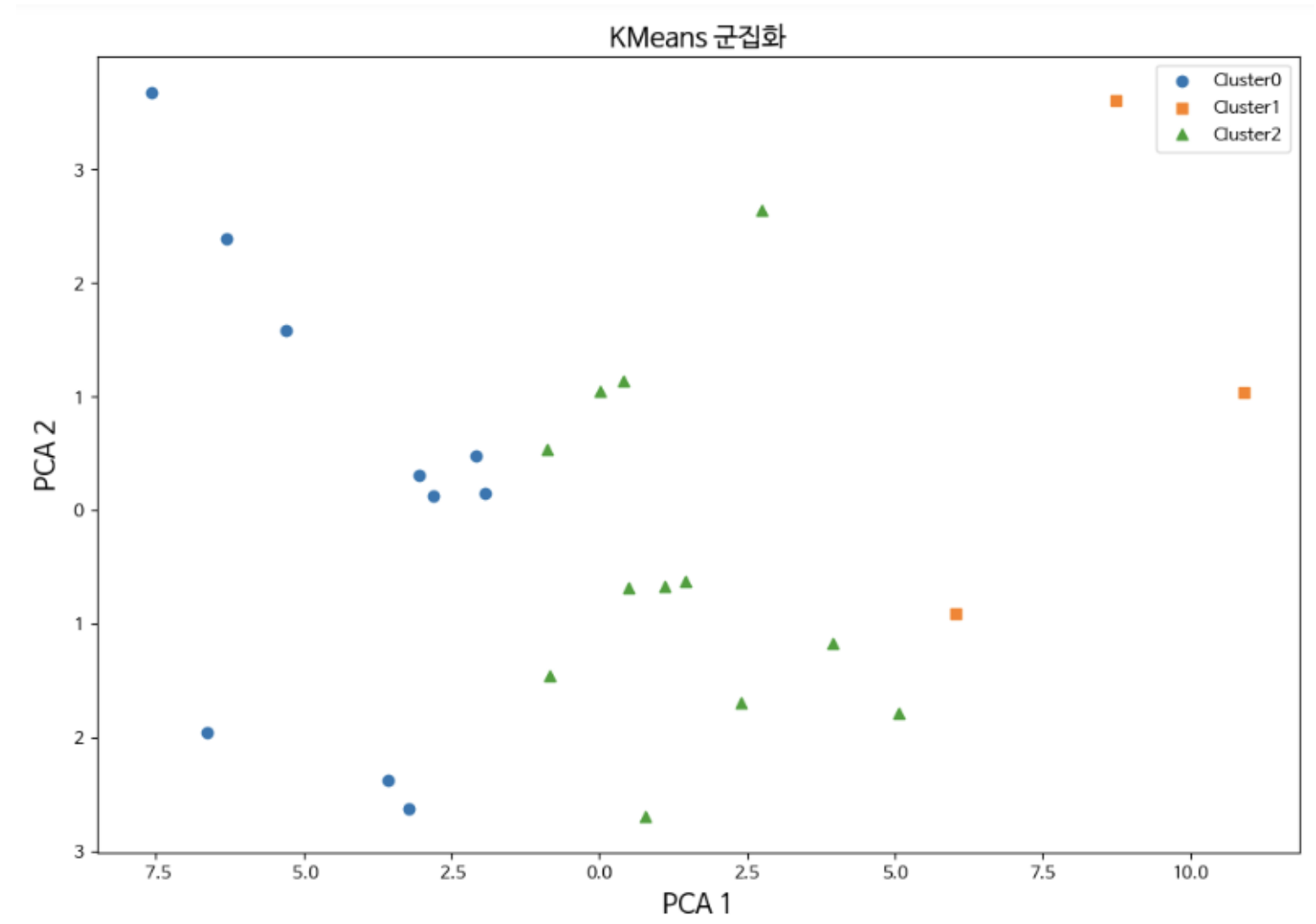


→ k = 3일 때 급격한 거리의 변화가 일어난 것으로 보아 클러스터 수로 3개가 적절해 보인다.

▲ Elbow Method

```
cluster
0    10
1     3
2    12
Name: 행정구, dtype: int64
```

→ 군집별 행정구 개수 확인



▲ 군집화 결과 시각화

#5.1 행정구 단위 분석

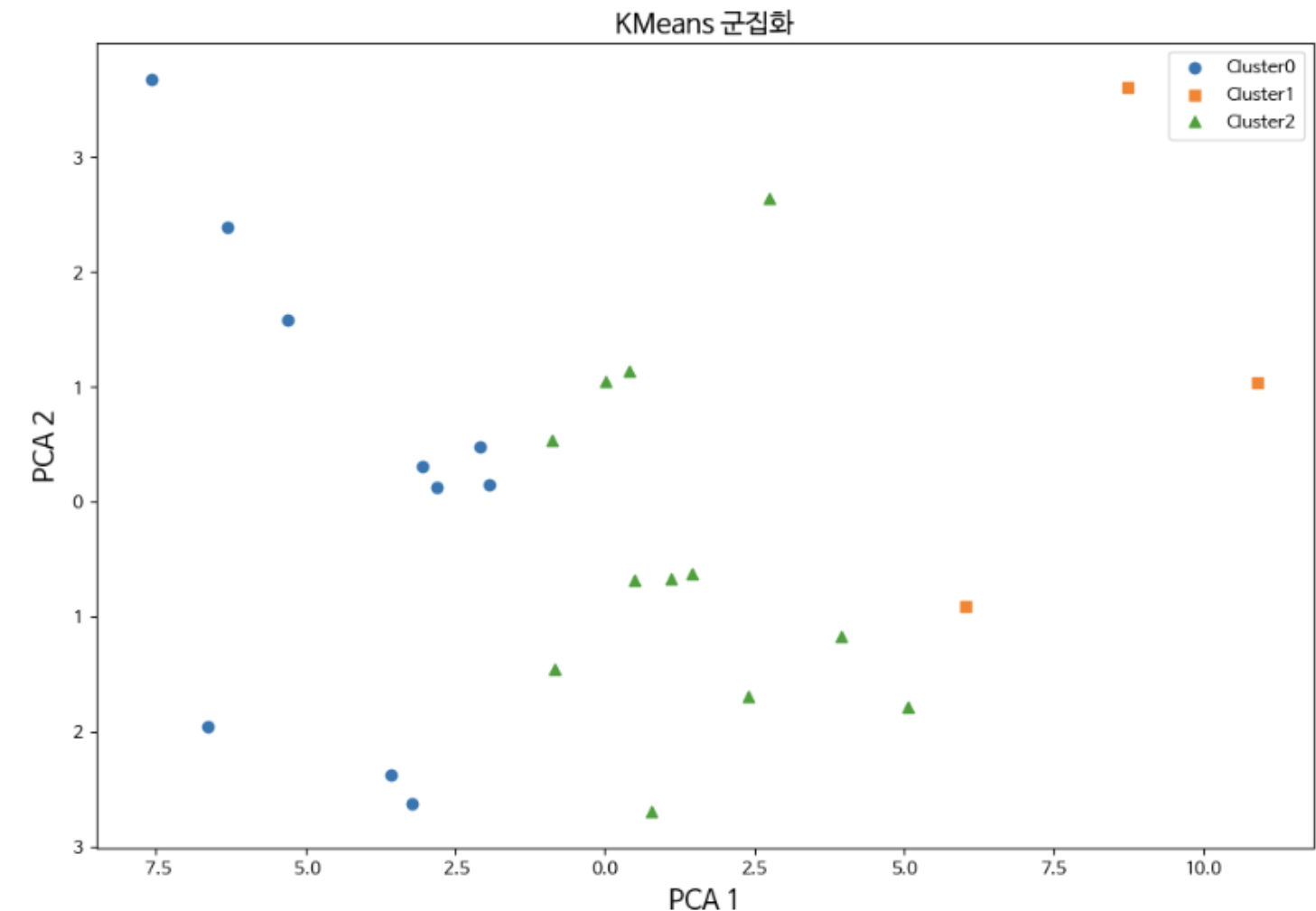
- 일식/수산물 – Kmeans 군집화

▼ 군집화 결과

- 군집0: [강북구, 광진구, 금천구, 도봉구, 동대문구, '서대문구', '성동구', '용산구', '종로구', '중구']
- 군집1: ['강남구', '강서구', '송파구']
- 군집2: ['강동구', '관악구', '구로구', '노원구', '동작구', '마포구', '서초구', '성북구', '양천구', '영등포구', '은평구', '중랑구']

▼ 군집 별 특징

- 군집0
 - 총생활인구수가 비교적 적은 지역들이다.
- 군집1
 - 전체 지하철역 수가 가장 많은 지역들이다.
- 군집2
 - 총생활인구수가 비교적 많은 지역들이다.
 - 평일과 주말 모두 10대 생활인구가 비교적 많은 지역들이다.



#5.2 행정동 단위 분석

- 한식/백반/한정식 – Kmeans 군집화

▼ 군집화 결과

```
cluster
0      126
1      115
2      113
3       72
Name: 행정동, dtype: int64
```

→ 군집별 행정구 개수 확인

ex> 군집3 내의 지역 분포

- 금천구, 용산구, 종로구, 중구 내 행정동들의 과반수 이상이 군집3에 속함
- 같은 행정구 내의 행정동들이 각각 다른 군집으로 군집화됨을 확인할 수 있음

행정구	행정동	행정구	행정동	행정구	행정동	행정구	행정동	행정구	행정동
강남구	1/22	광진구	-	동대문구	-	성동구	3/17	용산구	8/16
강동구	1/19	구로구	1/16	동작구	2/15	성북구	5/20	은평구	2/16
강북구	2/13	금천구	5/10	마포구	-	송파구	3/27	종로구	11/17
강서구	3/20	노원구	1/19	서대문구	2/14	양천구	4/18	중구	8/15
관악구	2/21	도봉구	2/14	서초구	1/18	영등포구	3/18	중랑구	2/16

#5.2 행정동 단위 분석

- 일식/수산물 – Kmeans 군집화

▼ 군집화 결과

```
cluster
0    121
1    218
2     87
Name: 행정동, dtype: int64
```

→ 군집별 행정구 개수 확인

ex> 군집2 내의 지역 분포

- 용산구, 종로구, 중구 내 행정동들의 과반수 이상이 군집2에 속함
- 같은 행정구 내의 행정동들이 각각 다른 군집으로 군집화됨을 확인할 수 있음

행정구	행정동	행정구	행정동	행정구	행정동	행정구	행정동	행정구	행정동
강남구	1/22	광진구	2/15	동대문구	1/14	성동구	6/17	용산구	11/16
강동구	1/19	구로구	1/16	동작구	3/15	성북구	3/20	은평구	1/16
강북구	1/13	금천구	3/10	마포구	-	송파구	4/27	종로구	13/17
강서구	3/20	노원구	-	서대문구	2/14	양천구	4/18	중구	15/15
관악구	4/21	도봉구	1/14	서초구	1/18	영등포구	4/18	중랑구	2/16

THANK YOU

