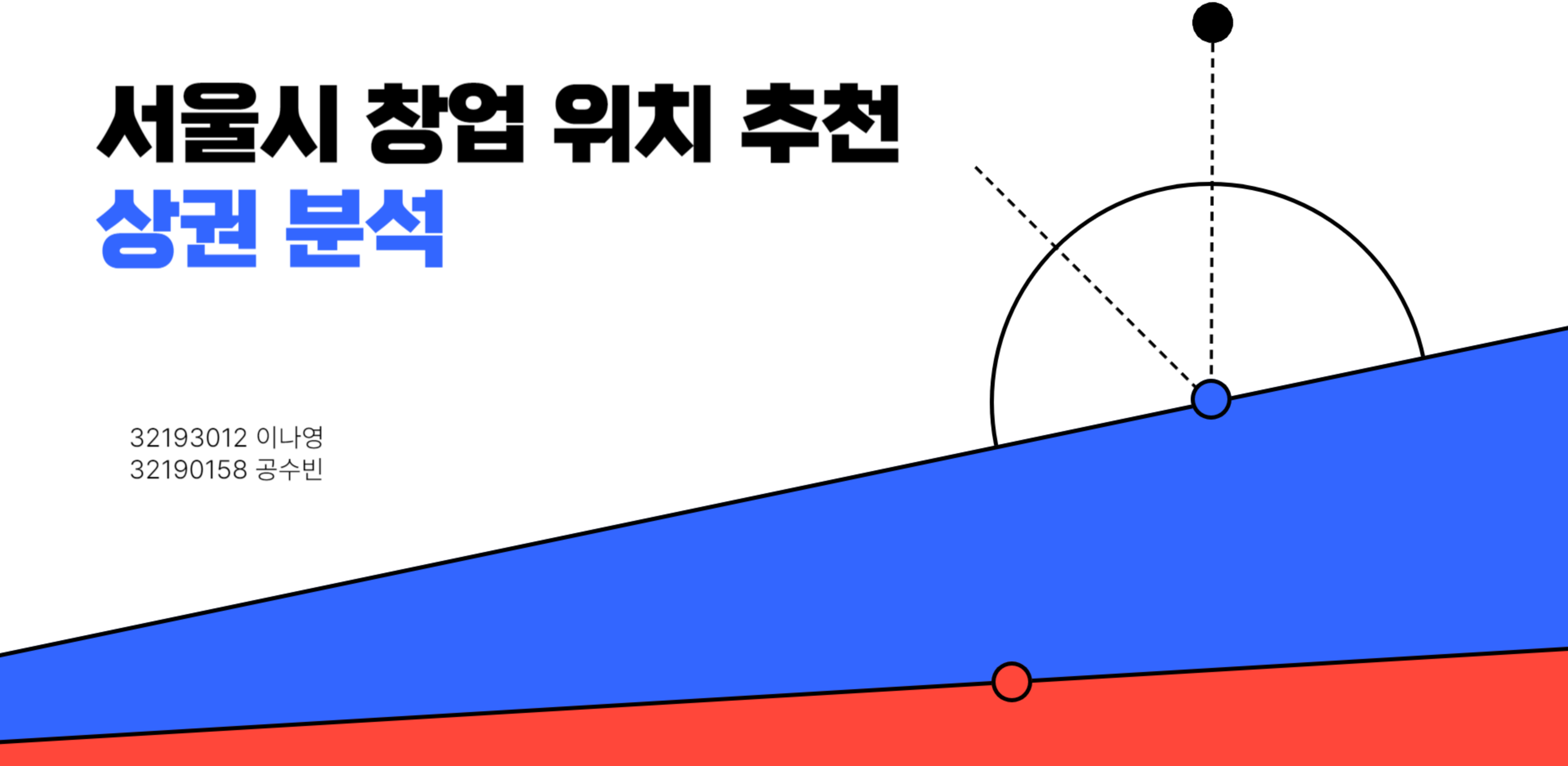
# 서울시 창업 위치 추천
## 상권 분석

32193012 이나영
32190158 공수빈



# 목차

# 최종 데이터 생성

**< 상권별 데이터 >**

최종 업종별 데이터는 업종별, 분기별로 나누어진 상태이기 때문에 업종, 분기를 나타내는 열이 없어서 각 데이터에 업종, 분기별로 번호를 부여하여 열을 생성하였음.

한식 : 0 / 양식 : 1 / 일식 : 2 / 중식 : 3 / 분식 : 4 / 제과점 : 5 /
카페 : 6 / 치킨 : 7 / 호프 : 8 / 패스트푸드 : 9

+ 행정동 코드 매핑
상권영역 데이터를 바탕으로 각 상권코드에 해당하는 행정동 코드를 매핑

# 최종데이터 - 상권 데이터

## <서울시 1분기 상권 데이터>

| | 상권코드 | 분기당매출건수 | 주중매출건수 | 주말매출건수 | 월요일매출건수 | 화요일매출건수 | 수요일매출건수 | ... | 11시~14시 유동인구 | 14시~17시 유동인구 | 17시~21시 유동인구 | 21시~24시 유동인구 | 업종코드 | 분기코드 | 행정동코드 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1001496 | 203237 | 157315 | 45922 | 30058 | 29882 | 32169 | ... | 848659 | 1026131 | 1815123 | 516101 | 0 | 1 | 11680580 |
| **1** | 1001495 | 432132 | 289385 | 142747 | 53388 | 57594 | 59754 | ... | 105810 | 125122 | 211767 | 75950 | 0 | 1 | 11710566 |
| **2** | 1001494 | 829999 | 621779 | 208220 | 121189 | 121012 | 122396 | ... | 3699463 | 4091380 | 5320223 | 1428696 | 0 | 1 | 11110615 |
| **3** | 1001493 | 262147 | 211872 | 50275 | 40287 | 40923 | 42988 | ... | 1994525 | 2254692 | 2899096 | 1031914 | 0 | 1 | 11140590 |
| **4** | 1001492 | 1242098 | 1080132 | 161966 | 212992 | 214132 | 207575 | ... | 2765755 | 3333748 | 6395073 | 1642232 | 0 | 1 | 11140520 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **8572** | 2110019 | 2601 | 2303 | 298 | 461 | 324 | 447 | ... | 1795305 | 1996646 | 2662794 | 850078 | 9 | 1 | 11110615 |
| **8573** | 2110018 | 3710 | 2399 | 1311 | 325 | 661 | 533 | ... | 39489 | 47339 | 51326 | 10522 | 9 | 1 | 11110600 |
| **8574** | 2110010 | 2171 | 1579 | 592 | 211 | 270 | 326 | ... | 85620 | 105926 | 127250 | 45263 | 9 | 1 | 11110560 |
| **8575** | 2110005 | 1056 | 702 | 354 | 119 | 160 | 119 | ... | 171336 | 214601 | 301124 | 106398 | 9 | 1 | 11110550 |
| **8576** | 2110002 | 2929 | 1964 | 965 | 280 | 438 | 351 | ... | 445539 | 497466 | 616231 | 219220 | 9 | 1 | 11110570 |

# 클러스터링 - 업종별 데이터

<업종별 데이터의 목적>

수많은 업종별 데이터들을 유사성을 띄는 데이터끼리 묶어 공통점을 이용해 사용자에게 위치를 추천하기 위함.

<비지도 학습>

- 정답이 없는 데이터
- 비슷한 특징까리 군집화하여 결과를 예측하는 방법
- 라벨링이 되어 있지 않은 데이터로부터 패턴을 찾아내는 것.

# 클러스터링 - 업종별 데이터

**<클러스터링>**

K-Means

- 센트로이드(클러스터 중심)를 랜덤하게 위치시키기 때문에 매번 결과가 달라질 수도 있음.
- 한 번에 k개의 센트로이드 랜덤하게 생성하기 때문에, 각 센트로이드 사이의 거리가 짧으면 분류가 제대로 이루어지지 않을 수 있음.

K-Means++

- 센트로이드를 한 번에 k개 모두 생성하는 것이 아니라, 센트로이드 사이의 거리를 최대한 멀리 위치시키는 방향으로 1개씩 총 k번 반복하여 k개의 클러스터를 만들어냄.

# 클러스터링 - 업종별 데이터

## 1. 피처 수 줄이기

- 피처가 너무 많으면 클러스터링이 잘 되지 않고 시각화도 어려움.

- 9개의 핵심 피처만 남긴 뒤 클러스터링 진행

→ 주중매출건수, 주말매출건수, 주중매출금액, 주말매출금액, 총생활인구수, 총직장인구수, 집객시설수, 유동인구, 점포수

※유동인구 : 모든 시간대 유동인구를 더해 총유동인구로.

| 상권코드 | 주중매출건수 | 주말매출건수 | 주중매출금액 | 주말매출금액 | 점포수 | 총생활인구수 | 총직장인구수 | 집객시설수 | 유동인구 |
|---|---|---|---|---|---|---|---|---|---|
| **1001496** | 157315 | 45922 | 2279474959 | 567598046 | 42 | 92815 | 15904.0 | 34.0 | 6131908 |
| **1001495** | 289385 | 142747 | 14046102369 | 7492124699 | 254 | 3221641 | 24375.0 | 149.0 | 758764 |
| **1001494** | 621779 | 208220 | 19395312772 | 7034868469 | 461 | 3610698 | 32935.0 | 243.0 | 19041503 |
| **1001493** | 211872 | 50275 | 5321613705 | 1450050074 | 158 | 2996402 | 6054.0 | 149.0 | 11263125 |
| **1001492** | 1080132 | 161966 | 42065368058 | 7300871543 | 602 | 3657861 | 104830.0 | 426.0 | 20529146 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **2110005** | 8578 | 3845 | 254664464 | 152448344 | 8 | 173118 | 1006.0 | 15.0 | 1106926 |
| **2110004** | 7304 | 4109 | 251498517 | 181437487 | 5 | 369100 | 32.0 | 8.0 | 3770350 |
| **2110003** | 4288 | 828 | 108186166 | 28772667 | 7 | 403155 | 485.0 | 15.0 | 1181315 |
| **2110002** | 12895 | 2561 | 265953724 | 80898347 | 8 | 302828 | 475.0 | 18.0 | 2599054 |
| **2110001** | 16834 | 9555 | 391217096 | 290592977 | 11 | 148832 | 1066.0 | 12.0 | 446729 |

# 클러스터링 - 업종별 데이터

## 2. 스케일링

- 피처 간 값의 범위가 매우 다르기 때문에 StandardScaler 사용해서 정규화 진행

피처 스케일링 : 서로 다른 피처간에 척도를 동일하게 해주는 것.

ex) 키, 몸무게 값들의 평균을 0으로 맞춰줌.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 2.109031 | 1.554819 | 1.375848 | 1.165310 | 1.021213 | -0.996372 | 0.998934 | -0.113582 | 0.240682 |
| **1** | 0.890374 | 1.423076 | 0.952943 | 1.407848 | 2.653888 | 2.394062 | 1.764770 | 2.235300 | -0.853929 |
| **2** | 1.515592 | 0.460707 | 0.698633 | 0.221138 | 1.254452 | 2.815648 | 2.538652 | 4.155257 | 2.870610 |
| **3** | 0.836175 | 1.291639 | 0.751790 | 1.078261 | 1.394396 | 2.149990 | 0.108428 | 2.235300 | 1.286008 |
| **4** | 2.319568 | 1.132861 | 3.194343 | 1.871639 | 3.353606 | 2.866755 | 9.038444 | 7.893043 | 3.173671 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **430** | -0.512622 | -0.361061 | -0.413123 | -0.314153 | -0.518167 | -1.052420 | 0.247202 | -0.011457 | -0.963991 |
| **431** | -0.397337 | -0.276363 | -0.095086 | 0.034729 | -0.518167 | -1.048140 | -0.386640 | -0.808034 | -0.760648 |
| **432** | -0.203186 | -0.293547 | 0.303029 | -0.201097 | -0.144984 | -0.496058 | -0.342973 | -0.583359 | -0.561158 |
| **433** | -0.429973 | -0.217862 | -0.339004 | -0.171988 | -0.238279 | -0.944623 | -0.353008 | -0.746759 | -0.931055 |
| **434** | -0.324907 | -0.070692 | -0.027466 | 0.292772 | -0.005040 | 0.112200 | -0.388990 | -0.583359 | -0.874992 |

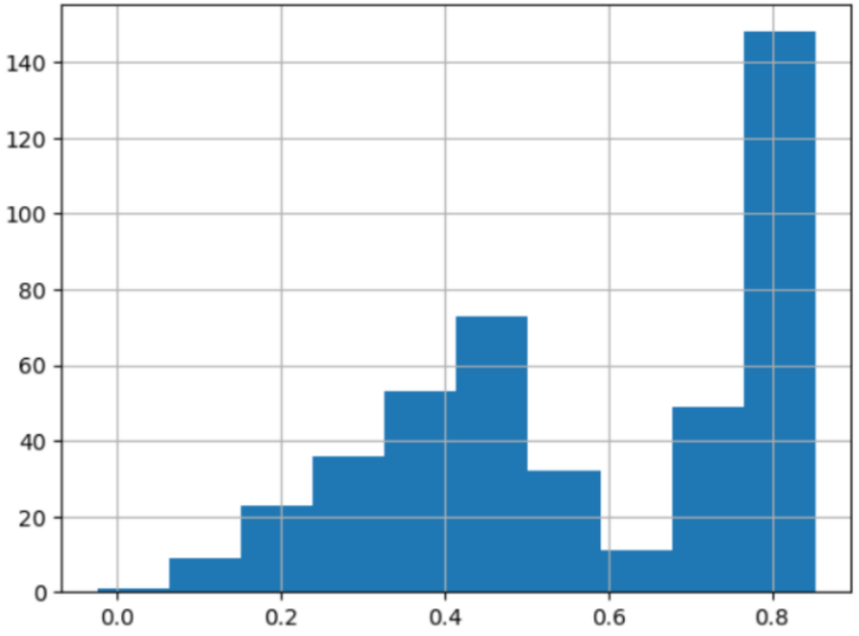
# 클러스터링 - 업종별 데이터

### 3. 클러스터 개수 정하기

최적의 클러스터 개수를 찾기 위해 실루엣 계수 이용.

실루엣 계수 : 개별 데이터가 할당된 군집 내 데이터와 얼마나 가깝게 군집화 되어있는지, 다른 군집에 있는 데이터와 얼마나 멀리 분리되어 있는지를 수치로 나타냄.

```
평균 :  0.5713273841626 96
그룹별 평균 :  cluster
0    0.368741
1    0.422194
2    0.271765
3    0.292083
4    0.847745
5    0.771768
6    0.288490
7    0.095974
8    0.362659
```



# 클러스터링 - 업종별 데이터

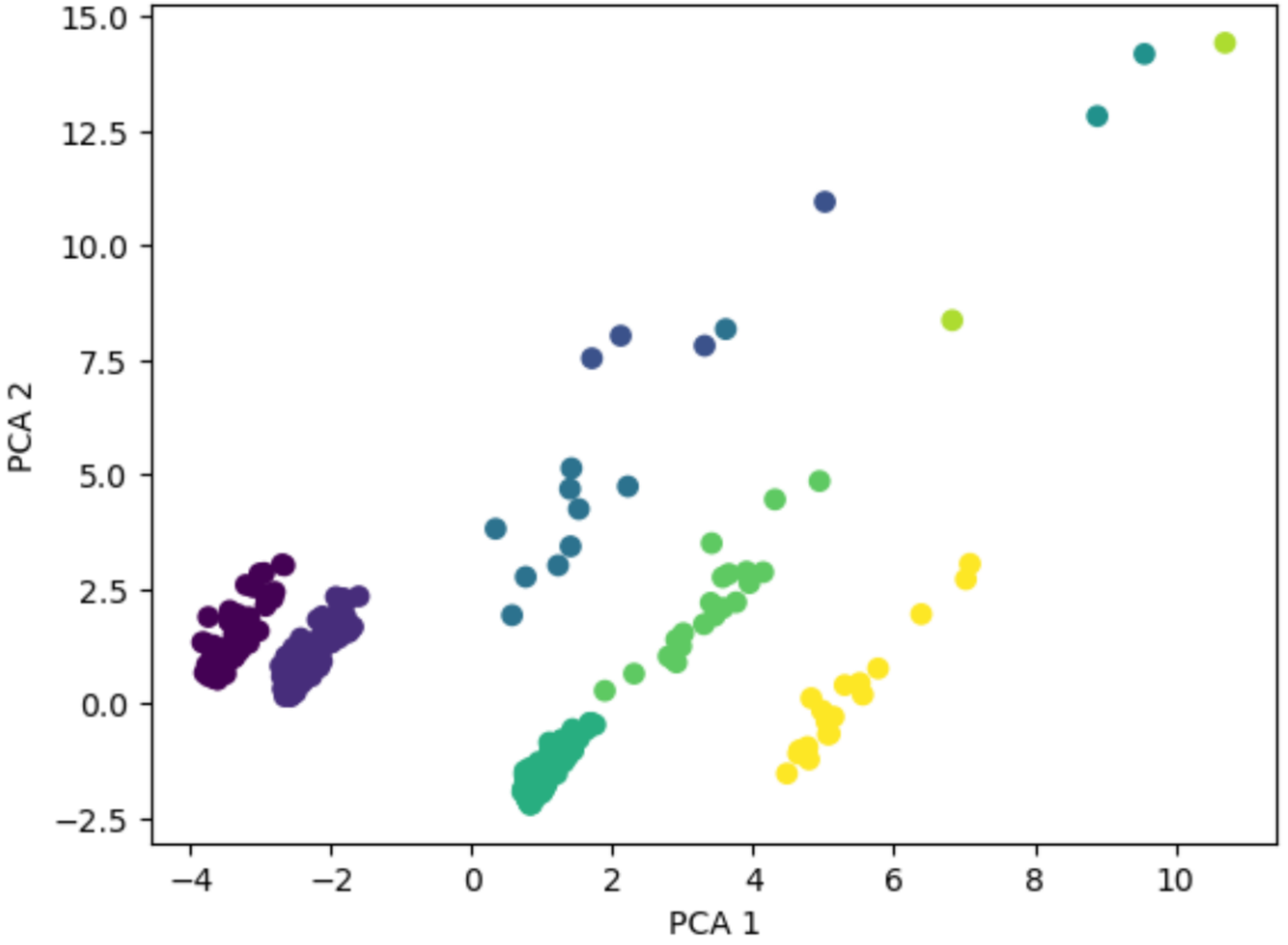
## 4. 클러스터링 확인

클러스터링이 잘 되었는지 확인하기 위해 클러스터링 된 데이터를 PCA로 차원축소한 뒤 시각화해서 확인한다.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | cluster | silhouette_coeff | pca_x | pca_y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2.109031 | 1.554819 | 1.375848 | 1.165310 | 1.021213 | -0.996372 | 0.998934 | -0.113582 | 0.240682 | 8 | 0.460956 | 5.514768 | 0.443659 |
| **1** | 0.890374 | 1.423076 | 0.952943 | 1.407848 | 2.653888 | 2.394062 | 1.764770 | 2.235300 | -0.853929 | 6 | 0.226150 | 4.142063 | 2.859653 |
| **2** | 1.515592 | 0.460707 | 0.698633 | 0.221138 | 1.254452 | 2.815648 | 2.538652 | 4.155257 | 2.870610 | 6 | 0.297997 | 3.410167 | 3.492774 |

# 클러스터링 - 업종별 데이터

## 4. 클러스터링 확인



# 감사합니다