

서울 자치구별 사업체의 특성분석

구로구와 강남구를 중심으로



주제 변경 사유

- 기존 분석배경 및 계획은 SNS 위험성 및 규제 필요성 인식 증대에 따른 규제 방안 도출이었습니다.
- 자료조사 과정에서 메타의 청소년 보호조치인 'Teens Accounts'를 25년 1월부터 한국에 적용한다고 11월에 발표하였습니다.
- 청소년 계정 적용 대상인지를 파악하기 위해 인공지능(AI) 기술 등을 활용한 성인 판별 시스템을 이용해 규제 대상을 정의하고 제한조치를 내리는 방법을 이용한다는 내용이 포함되어있었는데, 이는 제가 생각했던 내용이지만 더 정확하고 수준높은 결과물이 이미 나온 것이라고 생각했습니다.
- 자체적으로 회사에서 생산되는 데이터로 더 상세한 분석을 통해 제한 방안을 도출한 것을 보고 제가 하는 분석 내용이 이 내용보다 뛰어날 수 없다고 판단했고, 이로 인해 분석 내용에 대한 흥미가 감소하여 주제를 변경하게 되었습니다.

01

주제 소개

선정배경, 주요질문, 분석 개요

02

수집 데이터

데이터 상세

03

데이터 전처리

군집 분석 데이터셋 구축

04

EDA

주요 질문에 대한 분석

05

군집 분석

자치구별 사업체 특성, 소득,
사회적활동성 군집화

06

결론

결론, 한계



구로디지털단지



강남

● 주제선정 배경

01

"직장을 구로 쪽으로 가는 것과 강남 쪽으로 가는 것은 차이가 있다",
"강남구 사무실 하나 월세면 구로구에서 한 층을 쓸 수 있다" 라고
말씀하신 교수님의 의견이 발단

02

'지역별 사업체의 특성에 어떤 차이가 있을까, 있다면 어느정도로 차이가 날까'가
궁금해짐

03

데이터를 통해 교수님의 의견에 대해 확인해보고,
자치구별 지역 특성 유형을 나눠보자 분석 진행

● 주요 질문

01 구로구와 강남구의 IT직종 사람들은 얼마나 될까?



01- 1

정보 통신업 종사자수

02 해당 지역에서 근무하는 사람들은 어디에 거주할까?



02- 1

출근 목적 이동자 거주지

03 구로구와 강남구의 직장인 소득은 어떠할까?



03- 1

소득 구간별 주민 비율

04 실제로 평당 임대료 차이가 클까?



04- 1

층별 임대료

04- 2

강남 사무실 - 구로 한 층 임대 가능 여부

● 분석 개요

분석 기획

주요질문을 통한 EDA 과제 도출

- 지역별 산업 비중, 종사자수
- 직장 지역과 거주지 분석
- 소득 구간 비율 및 평균 소득
- 총별 임대료 분석, 임대 가능여부 분석

지역별 유형 분류를 위한 군집분석

- 분석 목표: 지역별 유형 분석을 통해 구로구와 강남구의 실제적 차이 분석 및 직장 선택에 따른 향후 생활특성 유추
- 주요 특성
 - 지역 거주민의 사회적 활동성
 - 지역 거주민의 경제적 특성
 - 지역 사업체의 특성

데이터 수집

공공데이터포털

- 대분류/업종별 사업체 현황
(사업체수/종사자수)

한국부동산원

- 상업용 부동산 총별 임대료

서울 열린데이터광장

- 서울 시민생활 데이터
- 서울 생활이동 데이터
- 서울시 건축물대장 총별개요
- 서울시 상권분석서비스
(소득소비-자치구)

문화빅데이터 플랫폼

- 전국 시군구단위 소득 구간별 주민비율

데이터 전처리

생활이동 데이터 전처리

- 24년 9월 일별 데이터 모두 결합
- 행정동명 결합
- 서울 내 이동만 필터링
- 출발-도착시간 데이터 형식 통일
- 행정동 기준 자치구명 매핑
- 자치구 기준으로 집계
- 자치구 중심점 위경도 매핑

상업용부동산 총별 임대료 전처리

- 상권별로 분류된 데이터를 자치구별로 재분류
- 결측값 처리 및 대체
- 기준 임대료 단위(천원/m²)를 평당 임대료로 변환
- 피벗 형태의 통계테이블을 tidy 형태로 변환

군집분석용 데이터셋 구축

데이터 분석

EDA 과제

- 지역별 산업 비중 파이차트
- 종사자 수 라인차트
- 출근 목적 OD 네트워크 차트
- OD 히트맵
- 자치구 소득 구간 비율 스택형 바차트
- 구로, 강남 소득구간 그룹형 바차트
- 총별 임대료 그룹형 바차트

군집 분석 과제

- 시민생활 데이터에서 사회적 활동성 관련 특성 도출
 - 소통빈도: 통화, 문자 빈도
 - 네트워크 크기: 통화, 문자 인원
 - 휴일 이동횟수, 휴일 이동거리
 - 동일 가중치 적용해 통합지표 산출
- 소득데이터를 경제적 특성으로 이용
- 사업체현황에서 사업체 특성 도출
 - 사업체/종사자수, 업종 비율, 산업집중도

● 데이터 상세

데이터 세부 정보	출처	크기	주요 속성	시간적 범위	파일 확장자
서울 시민생활데이터	열린데이터광장	8.9MB	행정동, 성별, 연령대, 인구수, 1인가구수, 평균 통화/문자량 & 대상자수, 휴일이동횟수, 이동거리	2024년 9월	xlsx
서울시 상권분석서비스 (소득소비-자치구)	열린데이터광장	102KB	기준_년_분기_코드, 월평균 소득, 총지출, 지출 산업별 지출액	2019년 1분기 ~ 2024년 2분기	csv
상업용부동산 임대료	한국부동산원	149KB	연도/분기별_층별_임대료, 연도/분기별_효용비율	22년 1분기 ~ 24년 1분기	csv
사업체 현황	공공데이터포털	10KB이내	사업체수, 종사자수, 업종별 종사자수	2022년	csv
서울 생활이동 데이터	열린데이터광장	1.7GB	출발 행정동, 도착 행정동, 출발시간, 도착시간, 이동 목적, 이동거리, 이동시간	2024년 9월	csv

● 군집분석 데이터셋 구축

사회적 활동성 지표 정의

- 소통빈도: 통화, 문자 빈도
- 네트워크 크기: 통화, 문자 인원
- 휴일 이동횟수, 휴일 이동거리

경제적 특성

- 지역별 평균 소득

지역별 사업체 특성

- 지역별 사업체 & 종사자 수
- 평당 임대료
- 산업집중도(HHI)

시민생활 데이터에서
특성값 도출

시군구별 소득 데이터에서
특성값 도출

사업체 현황, 상업용 부동산
임대료 데이터에서 특성값 도출

JOIN
군집분석 데이터셋 구축

왜 이 특성을 선택하였는가?

사회적 활동성 지표 정의

- 소통수
- 네트워크 크기: 통화, 문자, 인원
- 휴일 이동횟수, 휴일 이동거리

경제적 특성

사업체의 특성이 어떠한 경우에 사회적 활동성이 높게 나타나는지를 확인해보기 위해.

- 지역별 평균 소득

지역별 사업체 특성

- 평당 임대료
- 산업집중도(HHI)

경제적 특성

사업체의 특성이 어떠한 경우에 경제력이 높고 낮은지를 확인해보기 위해.

시민생활 데이터에서
특성값 도출

시군구별 소득 데이터에서
특성값 도출

사업체 현황, 상업용 부동산
임대료 데이터에서 특성값 도출

JOIN

군집분석 데이터셋 구축

● 시민생활 데이터 전처리

시민생활 데이터

시군구별 소득 데이터

사업체 현황, 상업용 부동산

```

import pandas as pd
# 서울 시민 생활 데이터 로드
life_df = pd.read_excel('2024.9월_29개 통신정보.xlsx')
# 사회적 활동성 지표에 필요한 컬럼 추출
life_df = life_df[['자치구', '행정동', '성별', '연령대', '총인구수', '1인가구수', '평균 통화량', '평균 문자량', '평균 통화대상자 수', '평균 문자대상자 수',
'주간상주지 변경횟수 평균', '휴일 총 이동 횟수 평균', '휴일 총 이동 거리 합계']]

# 대출 신입 평균 연령대부터 정년퇴임 나이까지 필터링
life_df = life_df[(life_df['연령대'] >= 30) & (life_df['연령대'] <= 60)]


# 자치구별 집계
grouped_gu = life_df.groupby('자치구').agg({
    '총인구수': 'sum',
    '평균 통화량': 'mean',
    '평균 문자량': 'mean',
    '평균 통화대상자 수': 'mean',
    '평균 문자대상자 수': 'mean',
    '주간상주지 변경횟수 평균': 'mean',
    '휴일 총 이동 횟수 평균': 'mean',
    '휴일 총 이동 거리 합계': 'mean'
}).reset_index()

grouped_gu

```

```

1 # 사회 활동성 통합 지표 생성 - 가중치 등급하게 적용
2 grouped_gu['소통 빈도'] = (grouped_gu['평균 통화량'] + grouped_gu['평균 문자량']) / 2
3 grouped_gu['네트워크 크기'] = (grouped_gu['평균 통화대상자 수'] + grouped_gu['평균 문자대상자 수']) / 2
4
5 # 통합 지표 계산
6 grouped_gu['통합 활동성 지표'] = (grouped_gu['소통 빈도'] + grouped_gu['네트워크 크기']
+ grouped_gu['휴일 총 이동 횟수 평균'] + grouped_gu['휴일 총 이동 거리 합계']) / 4
7
8
9 social_df = grouped_gu[['자치구', '소통 빈도', '네트워크 크기', '휴일 총 이동 횟수 평균', '휴일 총 이동 거리 합계', '통합 활동성 지표']]

```

1. 사회적 활동성 지표 산출에 필요한 컬럼 추출
2. 대출신입 평균 연령대인 30대부터 정년퇴임 나이인 60세 이하까지 필터링
3. 자치구 기준 집계
4. 소통 빈도, 네트워크 크기, 통합 활동성 지표 산출

● 소득 데이터 전처리

시민생활 데이터

시군구별 소득 데이터

사업체 현황, 상업용 부동산

```
1 # 소득데이터 로드
2 import pandas as pd
3 wage_df = pd.read_csv('서울시 상권분석서비스(소득소비-자치구).csv', encoding='ANSI').loc[:, '기준_년분기_코드':'지출_총금액']
4 # 가장 최근 기준 평균 소득 추출하기
5 wage_df = wage_df[wage_df['기준_년분기_코드'] == 20242].loc[:, '행정동_코드_명':'월_평균_소득_금액'].reset_index(drop=True)
6 wage_df = wage_df.rename(columns = {'행정동_코드_명':'자치구'})
7 # social_df와 join
8 df = pd.merge(wage_df, social_df, on = '자치구', how='left')
9 wage_df
```

1. 기준 년 분기를 가장 최신인 2024년 2분기로 필터링하고 월평균소득만 추출
2. 시민 생활 데이터에 자치구명을 기준으로 left join

● 상업용 부동산, 사업체 현황 데이터 전처리

시민생활 데이터

시군구별 소득 데이터

상업용 부동산, 사업체 현황

1. 상권으로 구분되어 있는 임대료 데이터에 자치구를 매핑시켜 자치구 컬럼 추가
2. 자치구 단위로 평당 임대료 집계
3. 도봉구와 서대문구 지역은 해당 데이터에 없어 인근지역 기준 평균값으로 대체
 - 도봉구: 강북구와 노원구의 임대료 평균
 - 서대문구: 종로구와 마포구의 임대료 평균
4. 자치구 기준 임대료 데이터 left join
5. 사업체 수와 종사자 수 left join
6. 업종대분류별 사업체수 데이터에서 산업 집중도(HHI) 산출
 - HHI란?
 - Herfindahl-Hirschman Index(허핀달-허쉬만 지수), 0과 1사이의 값
 - 시장 또는 지역 내 특정 산업이 얼마나 집중되어있는지 평가하는데 사용
 - 높을 수록 특정 산업 집중도 높고, 낮을수록 산업이 고루게 분포

$$HHI = \sum_{i=1}^N s_i^2$$

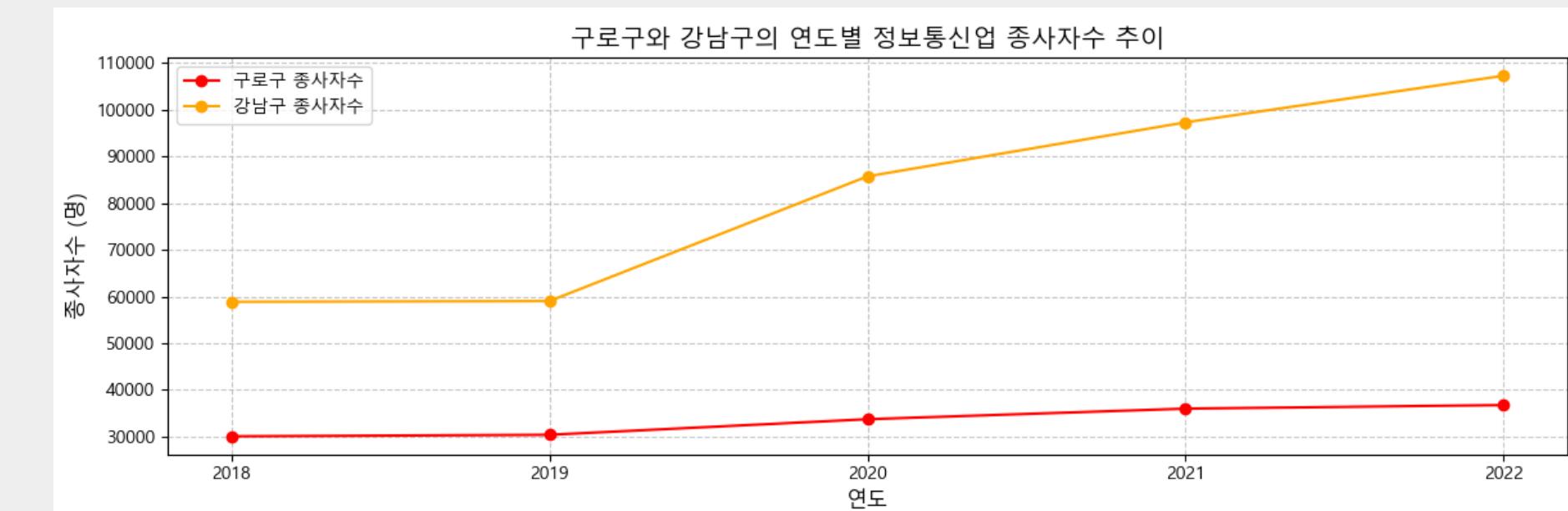
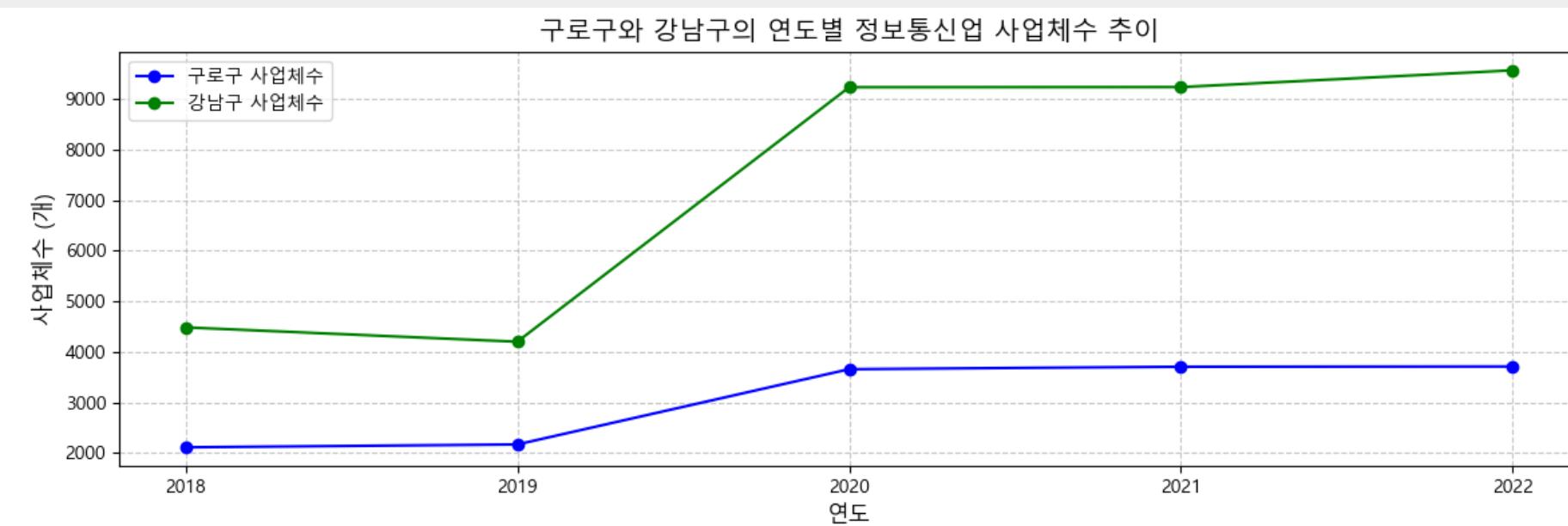
N: 시장(또는 지역) 내 존재하는 기업 또는 산업의 총 개수
s_i: 각 기업(또는 산업)의 시장 점유율(또는 비율)

2019년도와 2020년도 사이에 강남구의 사업체수와 종사자수가 거의 2배에 가깝게 대폭 상승하였지만 구로구는 소폭 상승하였습니다.

1. IT업체 수, 종사자 수

구로구와 강남구의 사업체수는 2020년도를 기점으로 3배의 차이가 되었으며, 강남구의 종사자수는 꾸준히 상승하는 모습을 보이고 있습니다.

이러한 결과를 볼때, 사업을 운영하는 사람과 종사자에게 모두 강남의 선호도가 높음을 확인하였습니다.

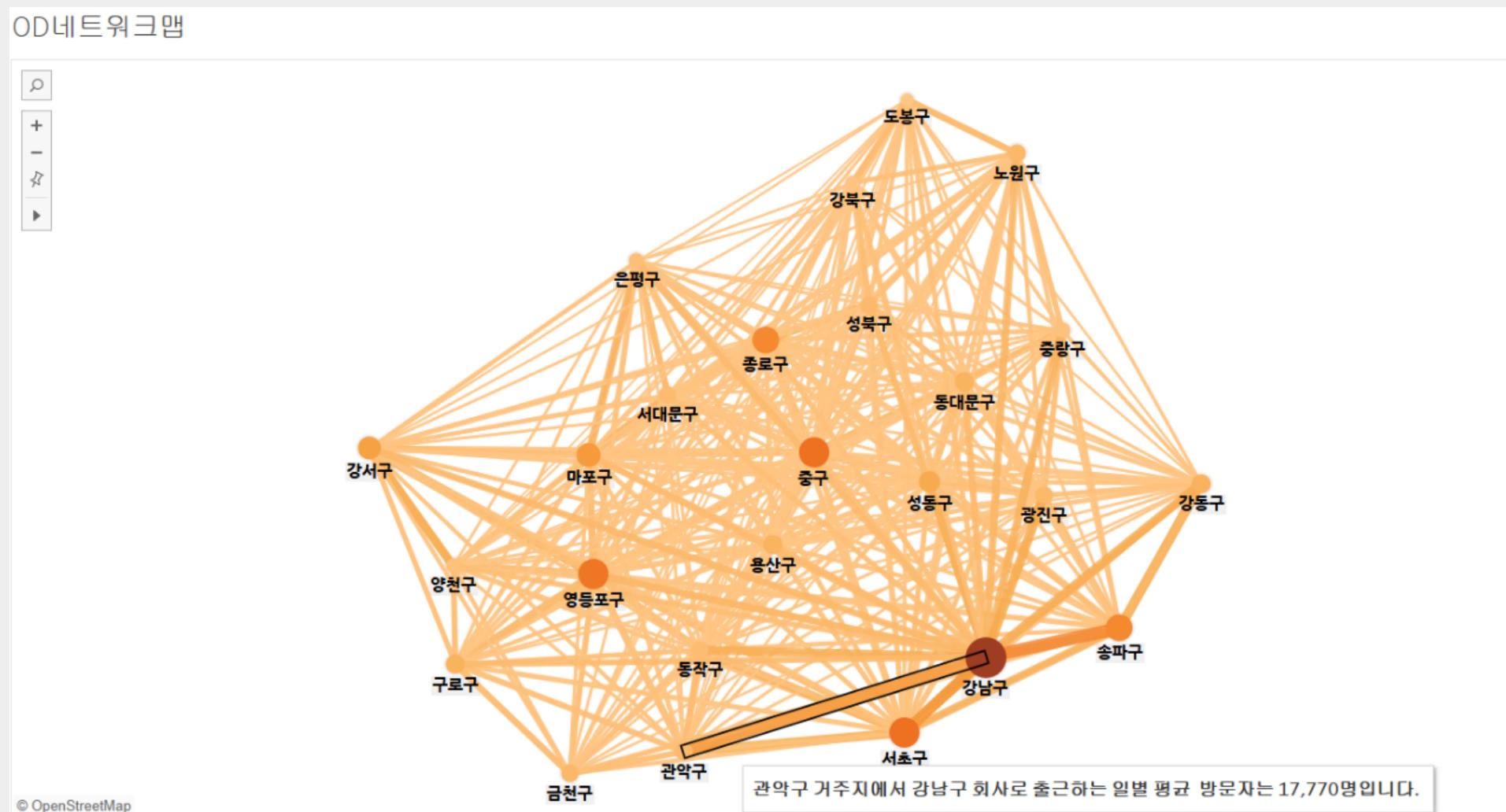


2. 직장지역, 거주지역

OD네트워크 차트를 먼저 보면 직장지역은 강남구, 서초구, 종로구, 중구, 영등포구 등이 상위임을 확인할 수 있습니다.

출발지 도착지간 히트맵은 행이 출발지, 열이 도착지로 구성이 되어있는데 대각선 줄이 빨갛게 나타나는 것은 본인이 거주하는 지역 내에서 근무하는 사람이 많다는 것을 의미합니다.

강남구 종사자는 서초, 송파, 관악에 주로 거주하며 구로구는 관악, 양천, 영등포가 많았습니다.



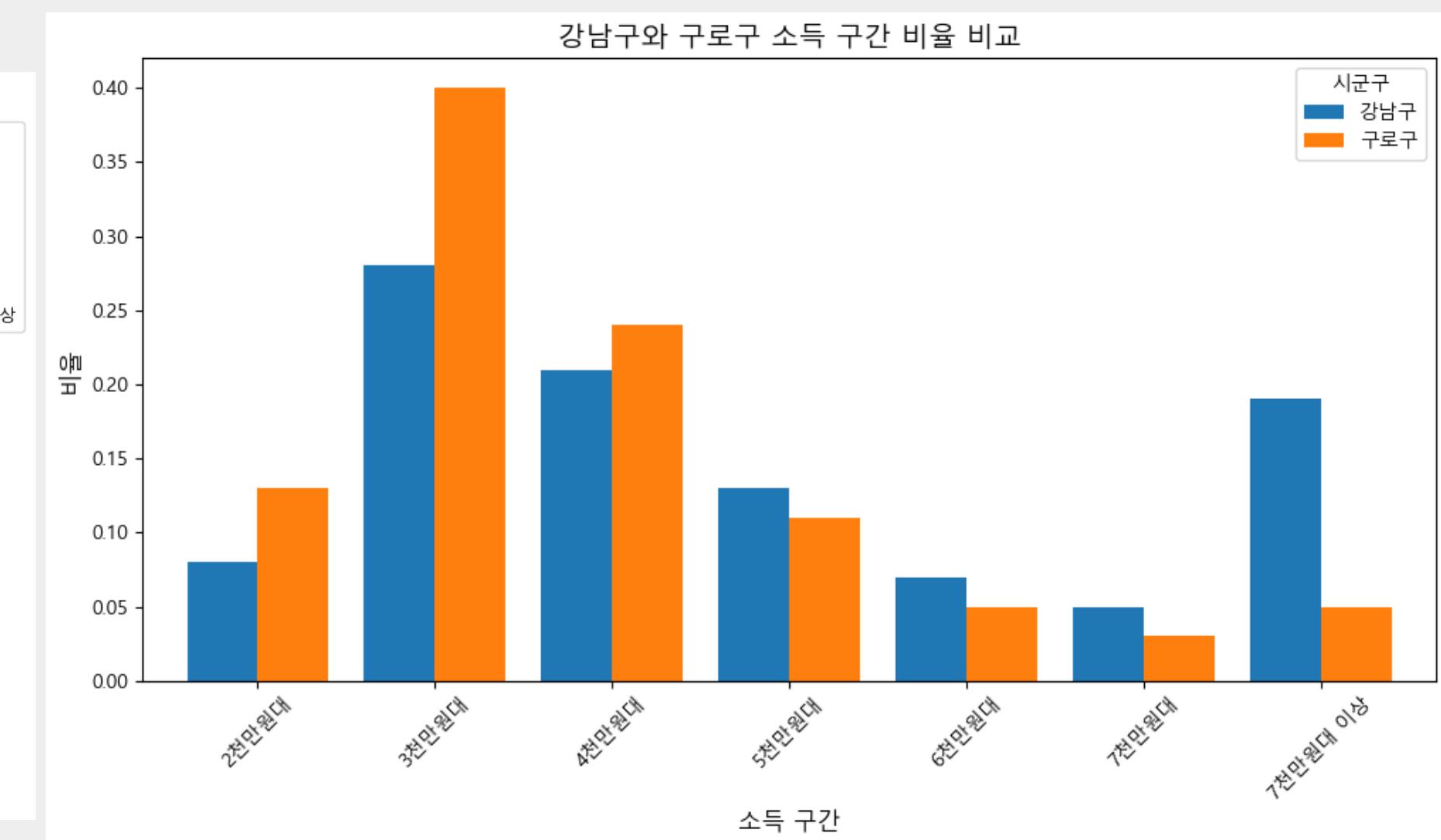
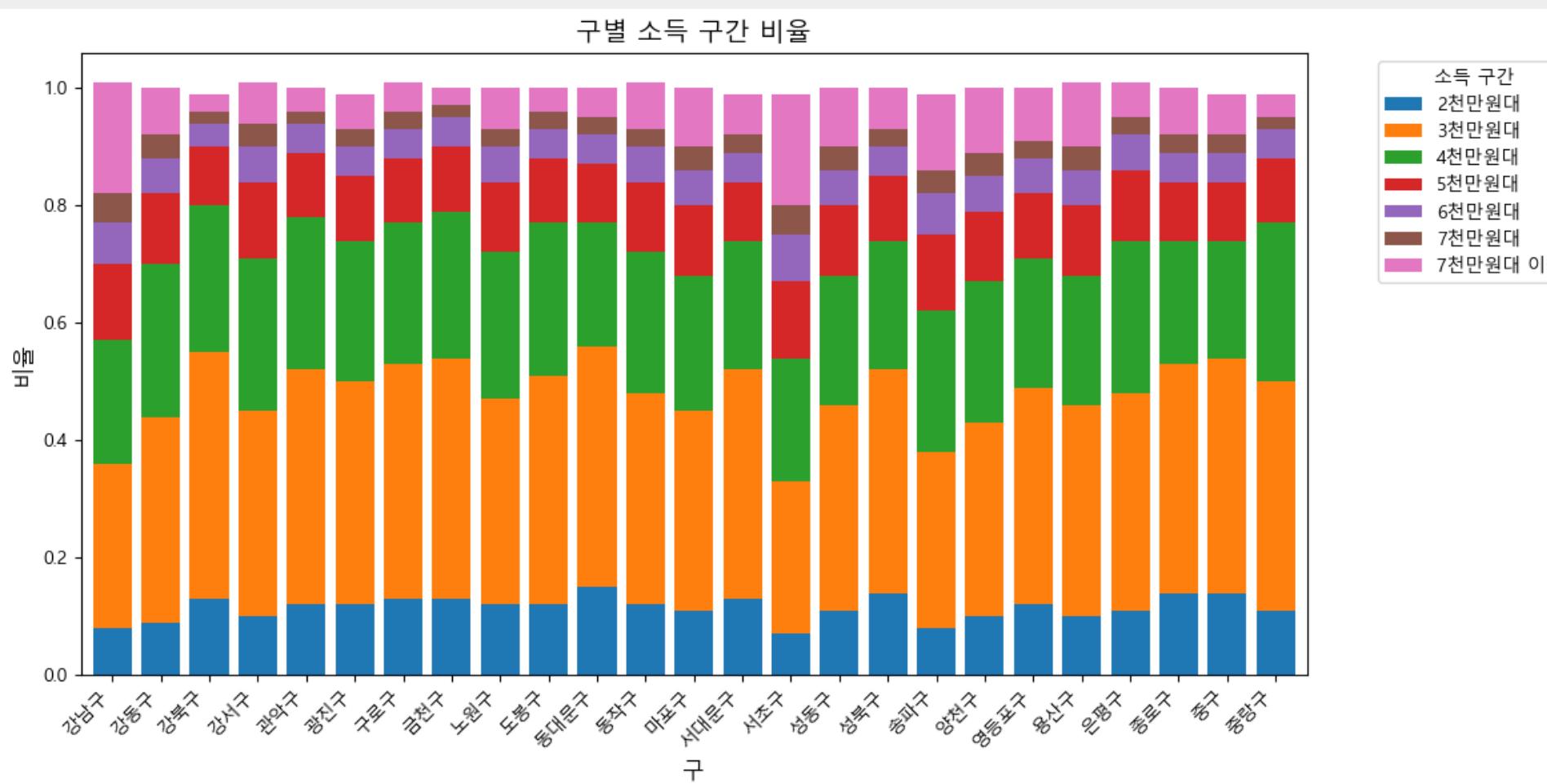
출발지-도착지 방문자수 히트맵	
강남구 - 61891.3	1802.5
강동구 - 11215.3	34169.9
강북구 - 2972.7	349.7
강서구 - 6209.6	323.4
관악구 - 17769.5	570.8
광진구 - 12197.5	2245.7
구로구 - 5181.5	199.0
금천구 - 2874.9	138.9
노원구 - 7487.1	281.7
도봉구 - 3228.2	427.9
동대문구 - 6091.9	1081.0
동작구 - 10438.1	460.6
마포구 - 4919.0	346.1
서대문구 - 3515.4	275.7
서초구 - 18262.1	666.9
성동구 - 9378.3	874.9
성북구 - 5330.3	583.9
송파구 - 24431.1	6861.2
양천구 - 3821.6	232.0
영등포구 - 6666.9	344.6
용산구 - 4428.0	241.1
은평구 - 5507.8	420.6
종로구 - 2308.4	216.9
중구 - 3105.8	265.8
종량구 - 8910.6	1389.9
강남구	1802.5
강동구	34169.9
강북구	323.4
강서구	570.8
관악구	2245.7
광진구	199.0
구로구	138.9
금천구	281.7
노원구	7487.1
도봉구	427.9
동대문구	1081.0
동작구	460.6
마포구	346.1
서대문구	275.7
서초구	18262.1
성동구	874.9
성북구	583.9
송파구	6861.2
양천구	232.0
영등포구	344.6
용산구	241.1
은평구	5507.8
종로구	216.9
중구	3105.8
종량구	1389.9

● 3. 소득구간비율, 평균소득

좌측 스택형 막대그래프를 보면 강남구 종사자가 많은 송파, 서초, 관악등은 3천만원대 소득이 가장 많긴 하나 송파는 4-5천만원대 주민비중이 높고, 서초구는 7천만원대 이상 소득인 비율이 상대적으로 아주 높습니다.

구로구 종사자가 많은 양천, 영등포는 전반적으로 3천, 4천, 5천만원대 소득 주민 비중이 높았습니다.

강남구와 구로구의 그룹형 바차트를 보면 4천만원대까지는 구로구의 비중이 높다가 5천만원대 이상 부터는 강남구의 비중이 더 높게 나타나는 것을 확인 할 수 있고 강남구는 7천만원대 이상의 주민 비중이 자치구 중 가장 높았고 구로구와도 큰 차이를 보였습니다.

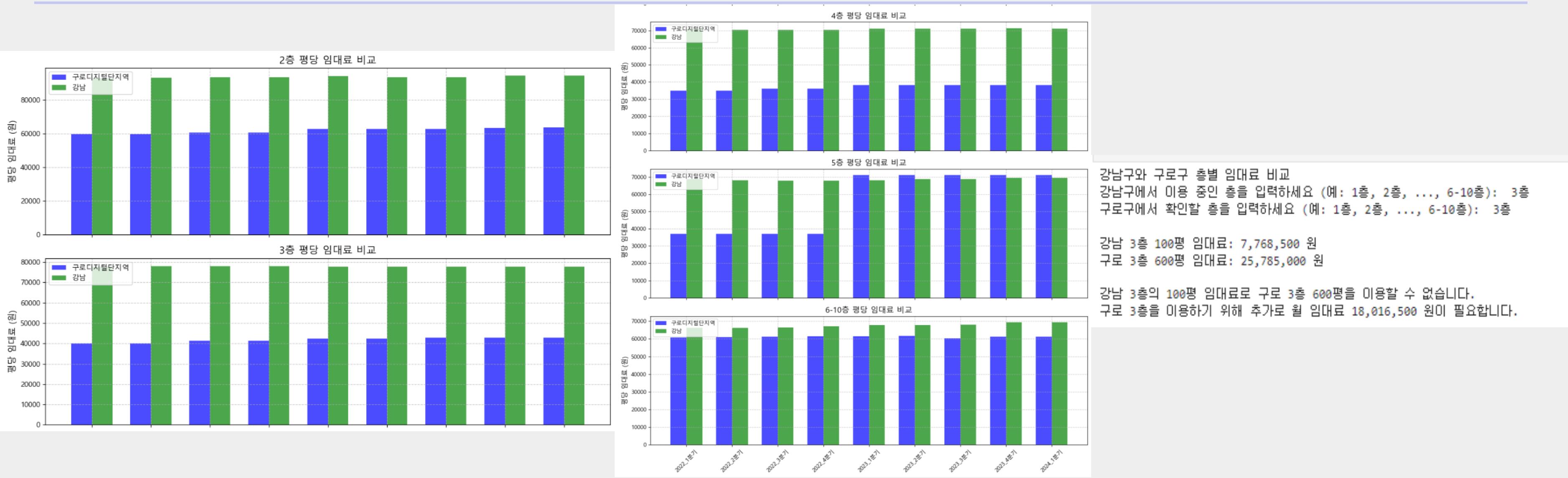


4. 평당 임대료

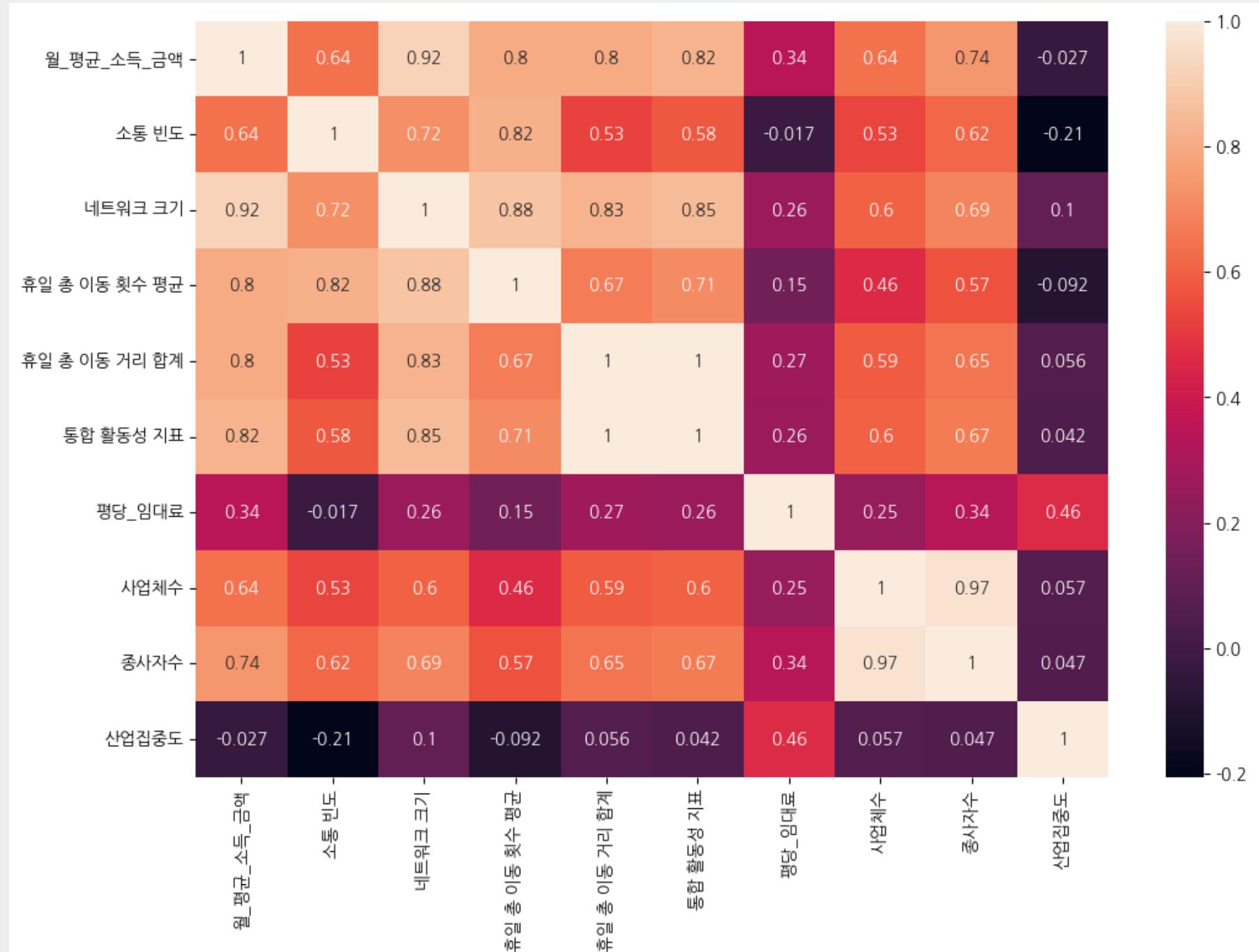
층별 평당 임대료의 경우 전반적으로 강남구의 임대료가 더 높았고 지하 1층부터 4층까지 구로구와 강남구의 임대료 차이는 꽤 크게 나타났습니다.

한 편 특이한 점은 5층의 평당 임대료가 23년도부터 강남구를 역전하는 모습을 보였는데 이는 추후 원인분석을 진행할 가치가 있어보인다고 생각하였습니다.

실제로 교수님이 말씀하셨던 강남구 사무실 임대료로 구로구의 한층을 임대하는게 가능한지 확인하는 간단한 프로그램을 만들어서 확인해봤는데 분석 결과 이는 불가능할 것으로 보입니다.



● 상관관계 분석

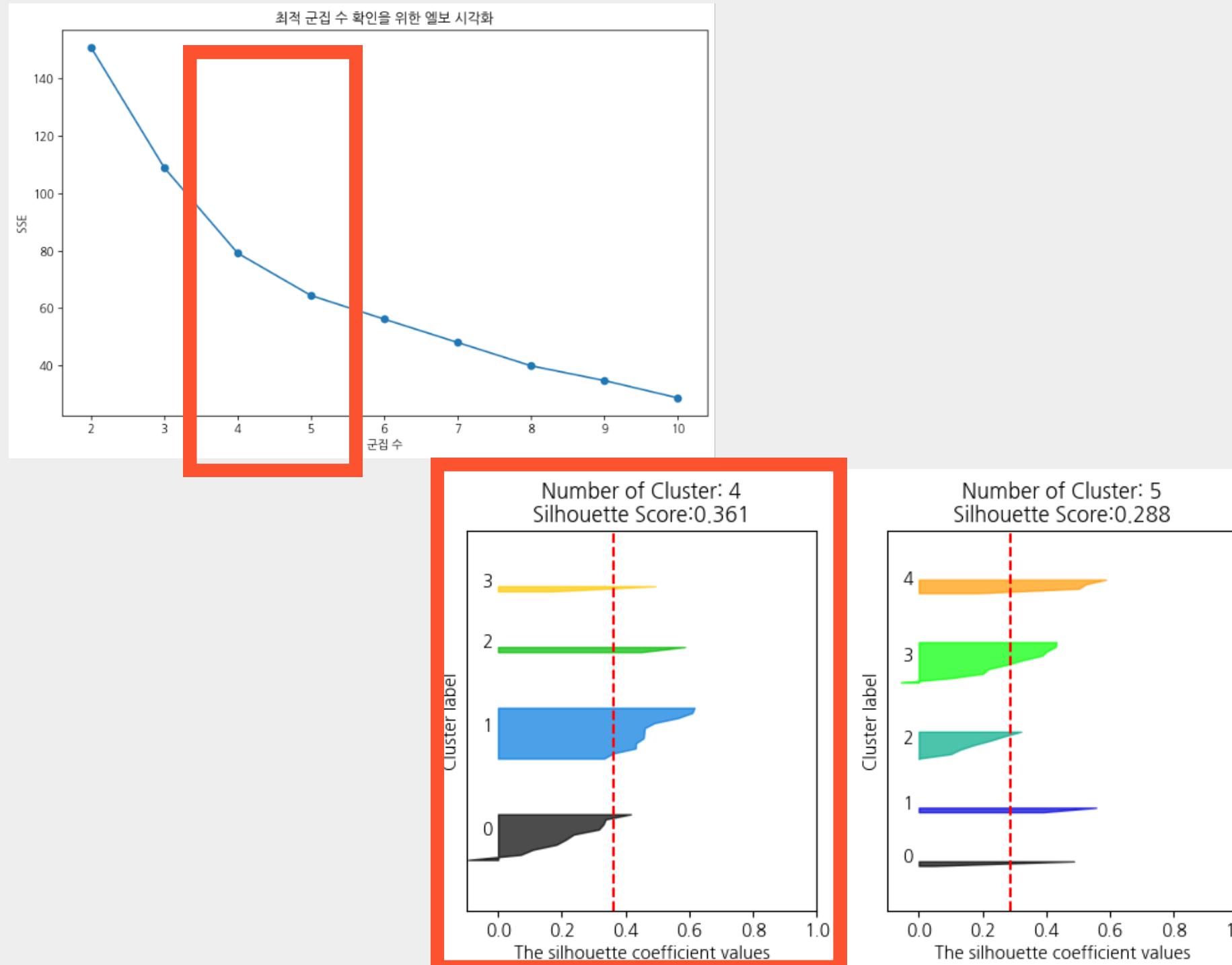


상관관계 히트맵

사회적활동성 – 경제적 특성 – 사업체 특성간의 관계를 위주로

- 평당 임대료 – 월평균 소득금액 약한 양의 상관관계
- 사업체 수와 종사자 수 – 사회적 활동성 양의 상관관계
- 소득과 사회적 활동성이 강한 양의 상관관계를 가짐.

● 최종 군집화 모델 개요

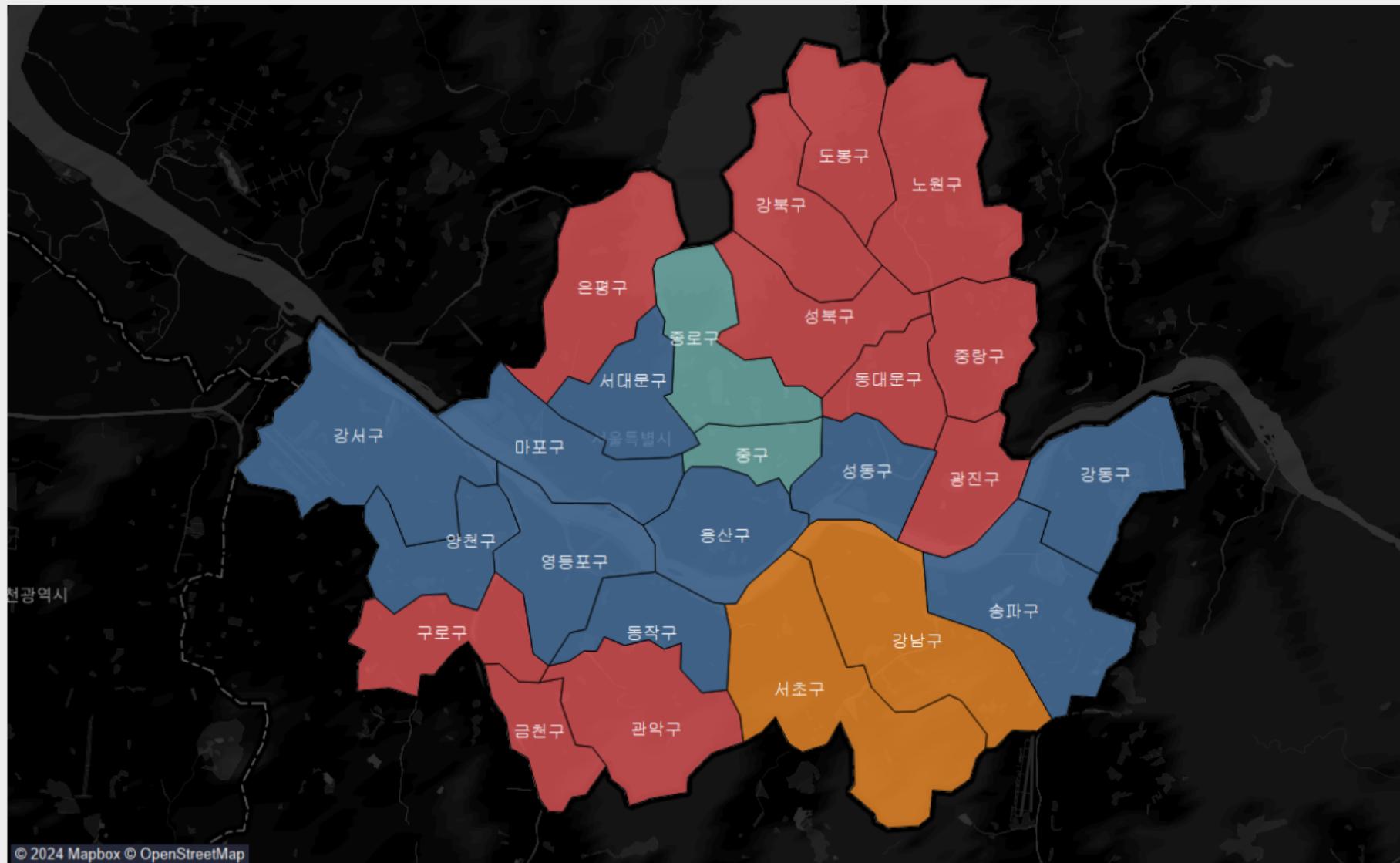


K-means

K-means와 계층적 군집화 알고리즘을 비교해보았으나
PCA 적용 후 K-means 알고리즘을 이용하는 것이
군집화 성능이 가장 높게 나타남

- 군집 분류대상: 자치구
- 실루엣계수: 0.361
- 최적 군집수 : 4

● 군집 분석 결과



클러스터 구분 맵차트

4개의 군집은 각각 균형 지역, 고소득/활성화형 지역, 저소득/저활성 지역, 특정산업 몰입형 지역으로 구분됨

- 구로구는 빨간색 지역으로, 저소득/저활성지역에 속함
- 강남구는 주황색 지역으로, 고소득/활성화 지역에 속함
- 그 밖에 파란색과 청록색은 각각 균형 지역, 특정산업 몰입형 지역으로 구분됨

● 결론 및 한계

EDA

EDA 결과 전반적으로 강남구에 대한 선호도가 높긴 하였음. 다만 강남구 사무실 임대료로 구로구 한총 임대는 불가능

Clustering

사업체 특성과 사회적 활동성, 경제적 특성으로 클러스터링 해본 결과 서울 자치구 내에서는 전반적으로 직장 선택에 있어 아쉬운 부분이 드러나는 지역이었고, 강남구와 정반대되는 군집이었음.

CONCLUSION

직관적으로도 인식하고 있었지만 데이터를 통한 검증에서도 강남구는 여러 특성을 고려했을 때 구로구에 비해 사람들의 선호도가 높은 지역임은 분명함을 확인하였음. 교수님의 표현에 과장이 있었으나 실제적 차이는 존재함

한계

해당 주제를 선정하고 데이터를 수집하는데 많은 어려움을 겪었고 원하는 분석과 비교 그림을 온전히 실현하지 못하였음. 또한 분석 과정에서 주관적인 의견이 개입되는 부분을 피할 수 없었음.



감사합니다.

참고 문헌

- 김필종. (2019). 의원, 치과의원, 한의원에 대한 산업집중도 분석. 한국산학기술학회 논문지, 20(4), 401-408.
- 장훈, 송민경. (2010). 군집분석을 이용한 수도권 도시의 유형화에 관한 연구. 대한공간정보학회지, 18(1), 83-88.
- 전혜원, (2024.10.09), 시사IN
<https://www.sisain.co.kr/news/articleView.html?idxno=54123>
- OPENAI. (2024). ChatGPT (40버전)[LLM].
<https://chat.openai.com>