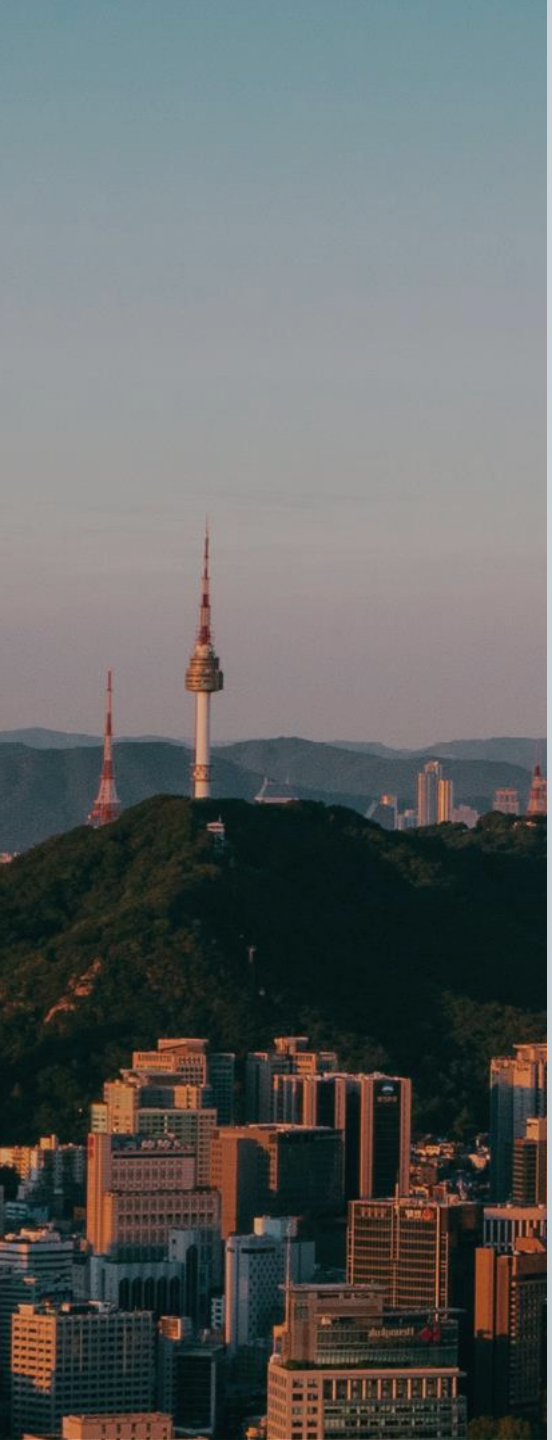
혼자 사는 청년들을 위한 서울시 살 곳 추천

I·SEOUL·U

4팀 데이터마이닝
박정현 김지민 노정화 염예빈 전규리



Contents

01 주제 선정

001 주제 선정 배경

002 문제 정의

02 전처리 및 EDA

: 사용 데이터 목록 소개

03 클러스터링

001 데이터 시각화

002 클러스터링

04 다음 주 예고

01 문제 상황 : 청년 주거 문제의 세 가지 양상

1. 과도한 주거비 부담

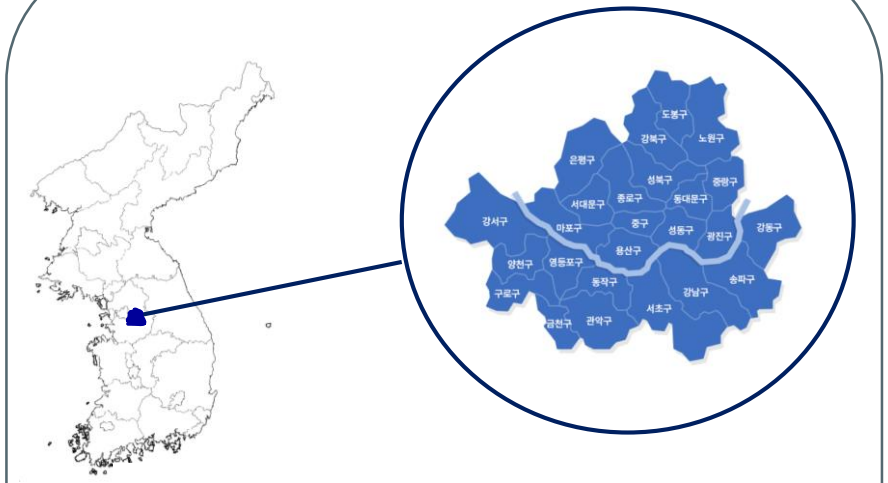
2. 열악한 주거 환경

3. 공공의 역할 부재

+

" 서울시의 특수성 "

서울시의 특수성?



서울특별시 전체 969만 9,232명

'20.09, KOSIS (행정안전부, 주민등록인구현황)

과도한 인구 밀집과 수위도시의 종주도시화

출처: 참여연대 복지동향 [청년 주거 문제 실태와 현 공공임대주택 정책의 한계]

01 주제 선정
1.1 주제선정배경
1.2 문제 정의
02 전처리 및 EDA
03 클러스터링
04 다음주 예고

01 문제 상황 : 청년 주거 문제의 세 가지 양상

1. 과도한 주거비 부담
2. 열악한 주거 환경
3. 공공의 역할 부재

+

" 서울시의 특수성 "

인구포화 및 집값 상승 현상 심각
특히 청년 주거가 뜨거운 감자

서울시의 특수성?

서울특별시 전체 969만 9,232명

'20.09, KOSIS (행정안전부, 주민등록인구현황)

이 중 2030세대의 인구

308만 2,487명

(서울시 인구의 30% 이상)

출처: 참여연대 복지동향 [청년 주거 문제 실태와 현 공공임대주택 정책의 한계]

01 주제 선정
1.1 주제선정배경
1.2 문제 정의
02 전처리 및 EDA
03 클러스터링
04 다음주 예고

01 기존 부동산 앱에서 제공하는 정보

매물 정보

주차	불가능
엘리베이터	없음
입주가능일	즉시 입주
관리비	3만원 관리비 외 사용 따라 부과 전기, 가스, 티비
구조	오픈형 원룸(욕실 1개)
면적(전용)	19.83m²
방향	남향
준공날짜	2012-01-25
층/건물층수	3층/3층
건물종류	다중주택
주소	종로구 명륜3가 64-3

면적, 관리비, 구조,
월세, 보증금, 방향,
준공날짜, 엘리베이터 여부,
관리비, 주차 가능 여부
등등

" 주거지의 **내부적 요소에 한정**됨 "

01 주제 선정
1.1 주제선정배경
1.2 문제 정의
• 분석 목적
• 대상 및 주제 선정
02 전처리 및 EDA
03 클러스터링
04 다음주 예고

01 주거지 선택 시 총 고려사항

내부적 요소

주거지 내부적 요소

- 월세, 전세, 보증금
- 층수, 방향, 방개수
- 기본옵션
- 낙후정도

+

외부적 요소

주거지 외부적 요소

- 주거지역의 상권 발달 정도
- 주거지역의 치안
- 주변 교통 인프라
- 의료 인프라

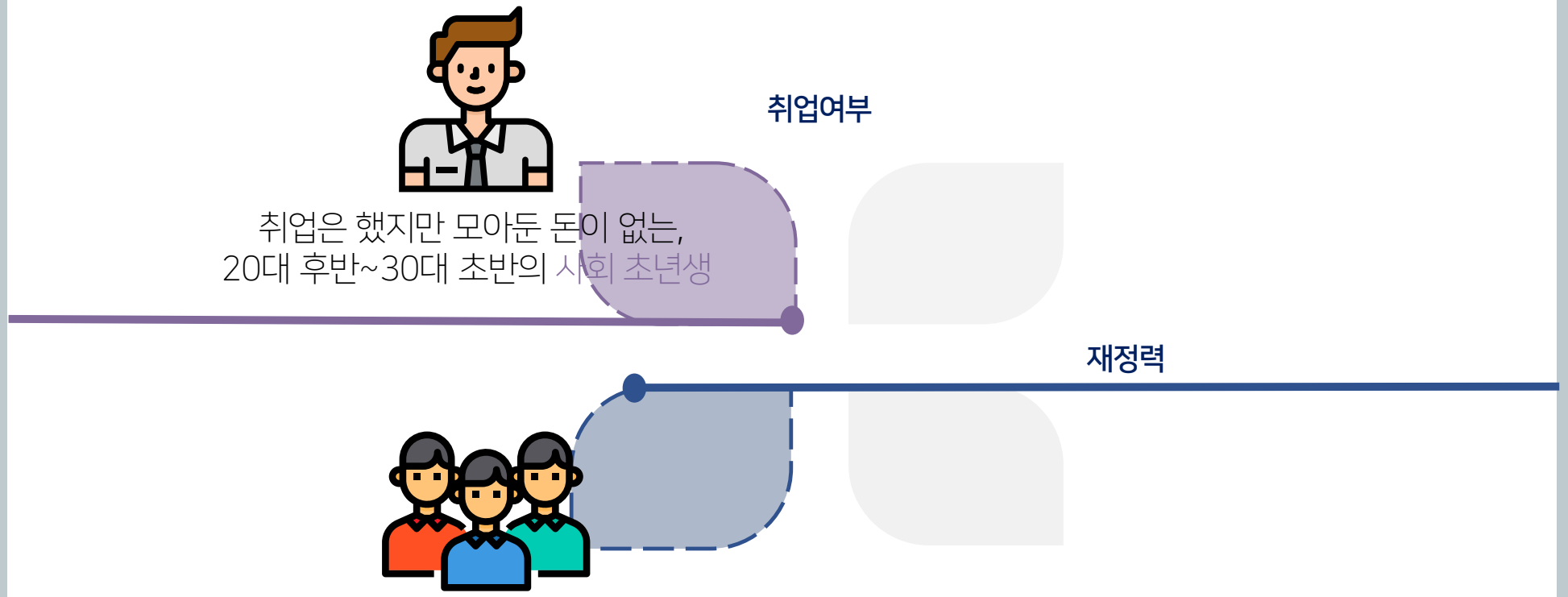
개인의 선호에 따라

조정 가능

01 주제 선정
1.1 주제선정배경
1.2 문제 정의
• 분석 목적
• 대상 및 주제 선정
02 전처리 및 EDA
03 클러스터링
04 다음주 예고

01 대상 및 주제 선정

추천 대상 재정의 : 원하는 집을 사기에 돈이 충분하지 않은 사람들



01 주제 선정
1.1 주제선정배경
1.2 문제 정의
• 분석 목적
• 대상 및 주제 선정
02 전처리 및 EDA
03 클러스터링
04 다음주 예고

02 유동인구 데이터 소개

[서울특별시 행정동별 유동인구 데이터]

2020년 1분기

gu	dong	총 합계
종로구	사직동	9,841
종로구	삼청동	2,993
종로구	부암동	10,597
종로구	평창동	18,768
종로구	무악동	8,767

2020년 2분기

gu	dong	총 합계
종로구	사직동	9,787
종로구	삼청동	2,973
종로구	부암동	10,421
종로구	평창동	18,696
종로구	무악동	8,691

01 주제 선정
02 전처리 및 EDA
2.1 유동인구 데이터
2.2 CCTV 데이터
2.3 카드 사용 데이터
2.4 범죄 데이터
2.5 도시위험도 데이터
2.6 전월세 데이터
2.7 역 · 정류장 데이터
03 클러스터링
04 다음주 예고

02 유동인구 데이터 분석

[서울특별시 행정동별
주민등록인구 수 대비 유동인구 수 평균 데이터]

time	dong_code	gu	dong	mean
0	11110515	종로구	청운효자동	0.00[illegible]289466
0	11110530	종로구	사직동	0.009574944
0	11110540	종로구	삼청동	0.004697922
0	11110550	종로구	부암동	0.007138906
0	11110560	종로구	평창동	0.004362032

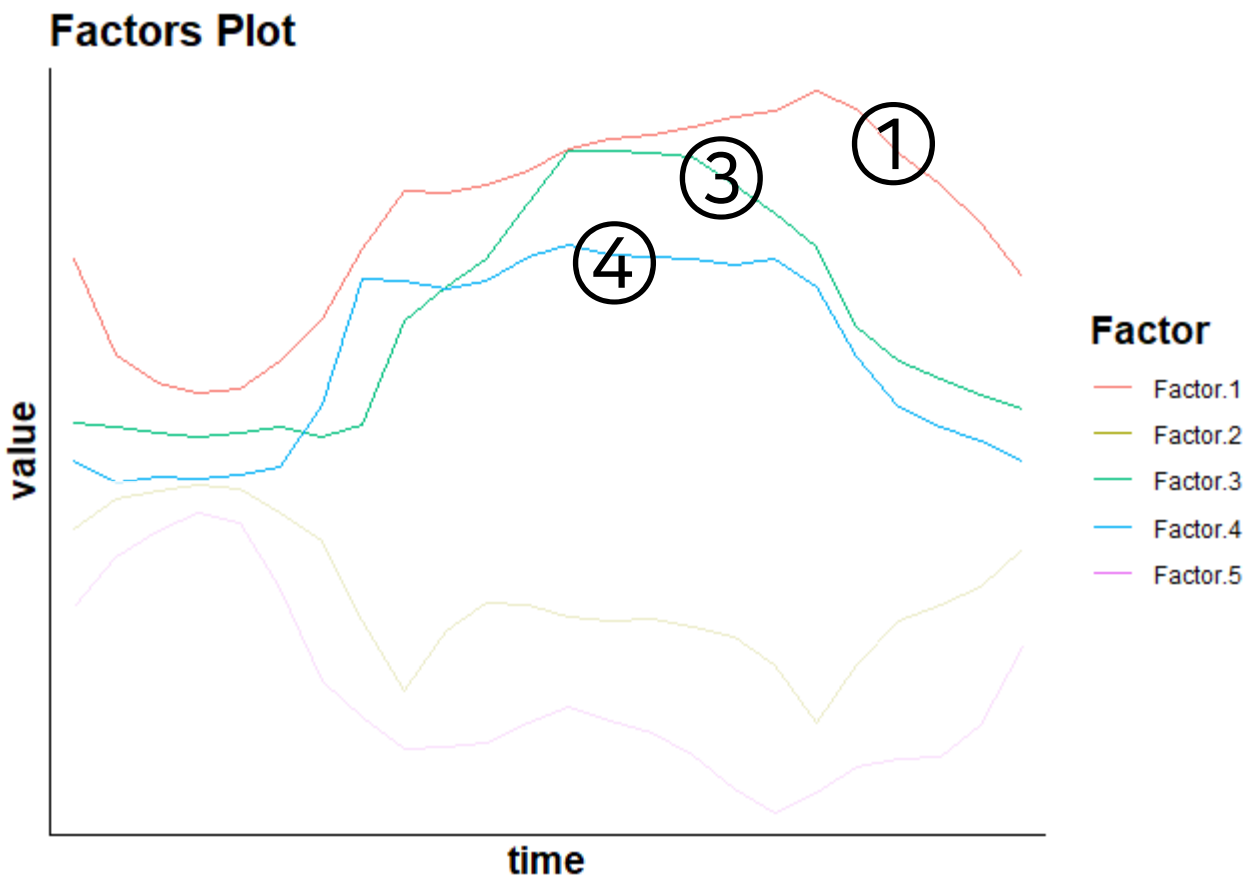
24 (0시 – 23시) X 425 (행정동 개수) = 행 10,200개

'요인분석'

01 주제 선정
02 전처리 및 EDA
2.1 유동인구 데이터
2.2 CCTV 데이터
2.3 카드 사용 데이터
2.4 범죄 데이터
2.5 도시위험도 데이터
2.6 전월세 데이터
2.7 역 · 정류장 데이터
03 클러스터링
04 다음주 예고

02 유동인구 데이터 분석

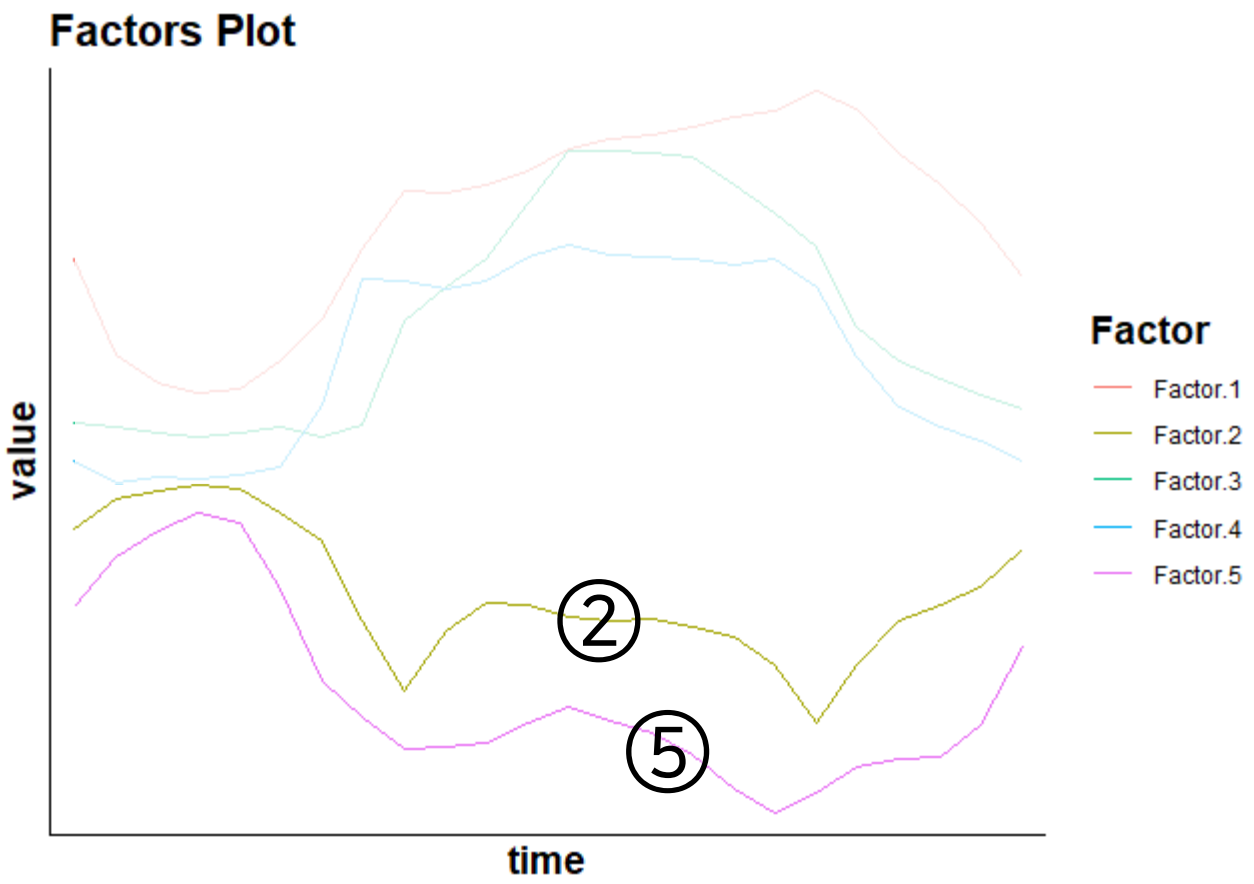
```
mv_fa <- estTSF.ML(joined_ts[,-c(1,sample(2:426,405))], 5, normalize = TRUE)
```



01 주제 선정
02 전처리 및 EDA
2.1 유동인구 데이터
2.2 CCTV 데이터
2.3 카드 사용 데이터
2.4 범죄 데이터
2.5 도시위험도 데이터
2.6 전월세 데이터
2.7 역 · 정류장 데이터
03 클러스터링
04 다음주 예고

02 유동인구 데이터 분석

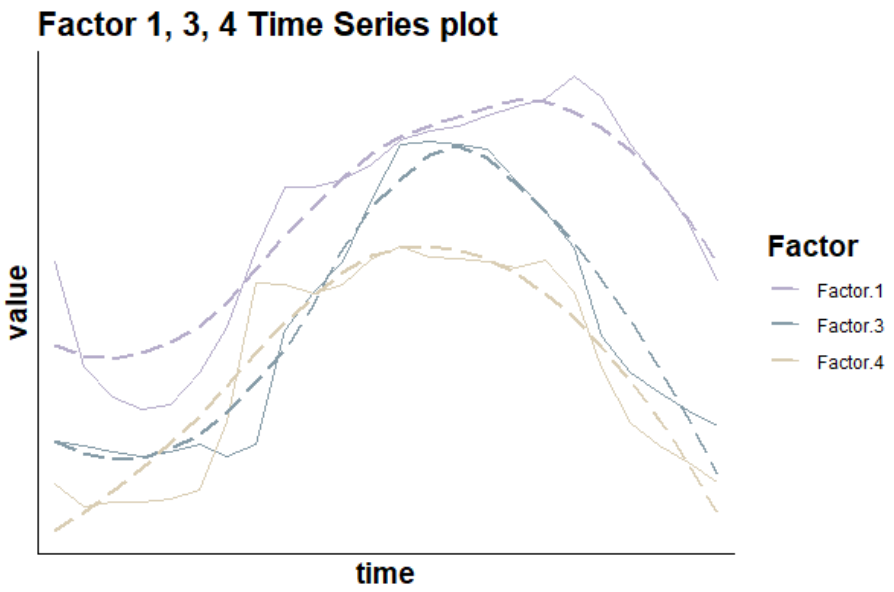
```
mv_fa <- estTSF.ML(joined_ts[,-c(1,sample(2:426,405))], 5, normalize = TRUE)
```

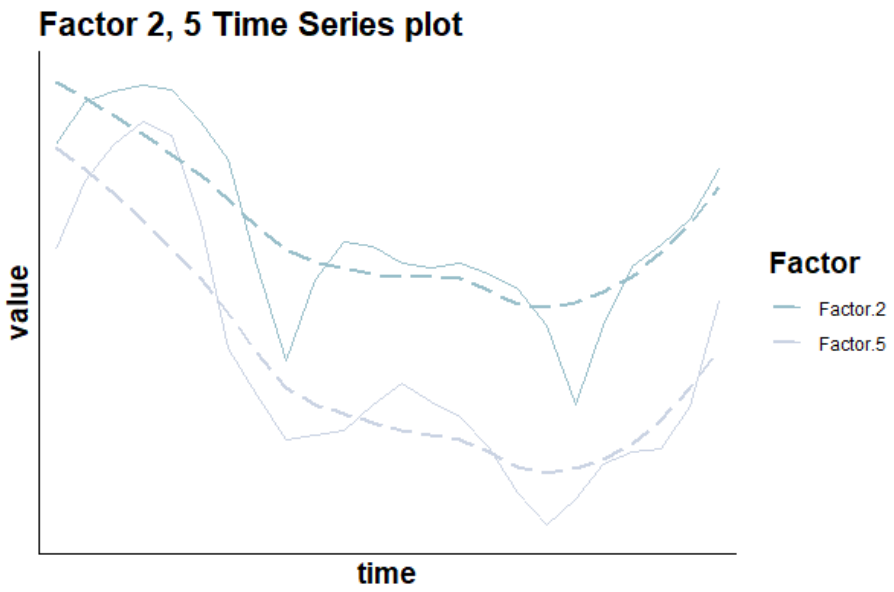


01 주제 선정
02 전처리 및 EDA
2.1 유동인구 데이터
2.2 CCTV 데이터
2.3 카드 사용 데이터
2.4 범죄 데이터
2.5 도시위험도 데이터
2.6 전월세 데이터
2.7 역 · 정류장 데이터
03 클러스터링
04 다음주 예고

02 유동인구 데이터 분석

각 군집에서 나타나는 추세를 그려보면,





01 주제 선정
02 전처리 및 EDA
2.1 유동인구 데이터
2.2 CCTV 데이터
2.3 카드 사용 데이터
2.4 범죄 데이터
2.5 도시위험도 데이터
2.6 전월세 데이터
2.7 역 · 정류장 데이터
03 클러스터링
04 다음주 예고

02 서울시 자치구별 CCTV 데이터

데이터 전처리

완성된 데이터셋

도로명주소	카메라대수	행정동
서울특별시 강남구 선릉로 7	3	개포1동
서울특별시 강남구 언주로 330	4	개포2동
서울특별시 강남구 개포로 310	4	개포2동
서울특별시 강남구 선릉로 103	3	개포1도

행정동별로 group_by

행정동별 카메라대수의 합

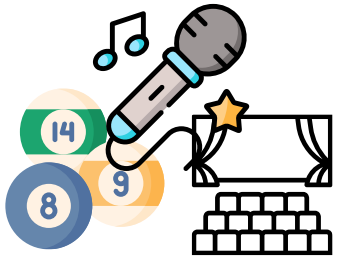
행정동	카메라대수
개포1동	55
개포2동	49

01 주제 선정
02 전처리 및 EDA
2.1 유동인구 데이터
2.2 CCTV 데이터
2.3 카드 사용 데이터
2.4 범죄 데이터
2.5 도시위험도 데이터
2.6 전월세 데이터
2.7 역 · 정류장 데이터
03 클러스터링
04 다음주 예고

02 업종별 카드결제 데이터

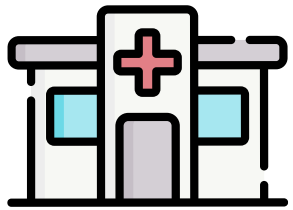
데이터 소개 : BC카드 업종별 카드결제 데이터

"동네의 문화 발달 지수, 의료 인프라 지수, 식문화 인프라 지수를 고려할 수 있는 방법은 없을까?"



노래방, 당구장, 영화관 매출

문화 지수



각종 의원 매출

의료 지수

음식점, 카페, 대형마트 매출

식문화 지수

01 주제 선정
2.1 유동인구 데이터
2.2 CCTV 데이터
2.3 카드 사용 데이터
2.4 범죄 데이터
2.5 도시위험도 데이터
2.6 전월세 데이터
2.7 역 · 정류장 데이터
03 클러스터링
04 다음주 예고

02 범죄 데이터 분석

5대 범죄의 행정구별 차이가 유효한가?

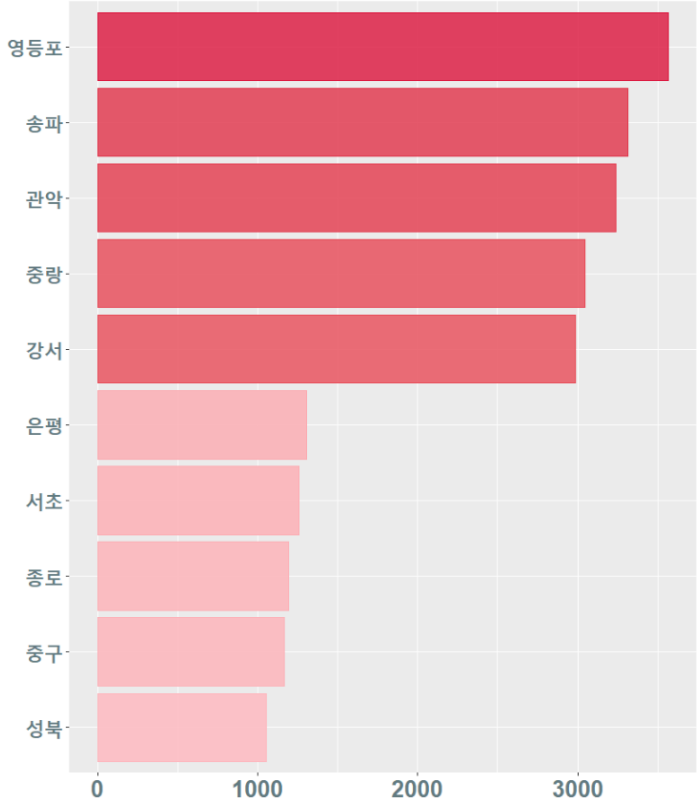
*범죄, 살인, 강간, 강도, 절도의 **10년치 발생건수를 통해 통계분석**을 해보자*

- **시각화** — 범죄 건수 상하위 5개지역의 차이를 보임
- **ANOVA 테스트** — 25개 행정구역간 범죄발생이 동일하다
- **Scheffe 다중검정** — 범죄 건수가 비슷한 행정구별로 그룹화

01 주제 선정
02 전처리 및 EDA
2.1 유동인구 데이터
2.2 CCTV 데이터
2.3 카드 사용 데이터
2.4 범죄 데이터
2.5 도시위험도 데이터
2.6 전월세 데이터
2.7 역 · 정류장 데이터
03 클러스터링
04 다음주 예고

02 범죄 데이터 분석 - 폭력

시각화



25개 행정구중

폭력발생건수가 가장 많은

상위 5개 행정구와

가장 적은 **하위 5개 행정구**를 추출

행정구별 평균 폭력건수

낮음 높음

01 주제 선정
02 전처리 및 EDA
2.1 유동인구 데이터
2.2 CCTV 데이터
2.3 카드 사용 데이터
2.4 범죄 데이터
2.5 도시위험도 데이터
2.6 전월세 데이터
2.7 역 · 정류장 데이터
03 클러스터링
04 다음주 예고

02 범죄 데이터 분석 - 폭력

ANOVA 테스트

```
             Df Sum Sq Mean Sq F value Pr(>F)
구분         24   2966   123.6   12.36 <2e-16 ***
Residuals   285   2849    10.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

귀무가설 : 25개 행정구의 폭력범죄발생 건수가 동일하다

P-value < 0.05

"25개 행정구의 폭력범죄 발생 건수는 동일하지 않다"

01 주제 선정
02 전처리 및 EDA
2.1 유동인구 데이터
2.2 CCTV 데이터
2.3 카드 사용 데이터
2.4 범죄 데이터
2.5 도시위험도 데이터
2.6 전월세 데이터
2.7 역 · 정류장 데이터
03 클러스터링
04 다음주 예고

02 범죄 데이터 분석 - 폭력

다중분류 분석 (Scheffe test)

25개 행정구역은
폭력발생건수에 따라
18개의 그룹으로 나뉘어진다

행정구역간 폭력발생 양상이
상당히 다양하다는 것을 알 수 있다

Ex. 영등포구가 송파구보다 위험한 정도가
송파구가 관악구보다 위험한 정도보다 월등히 높다

	건수	groups
영등포	3565.70	a
송파	3316.20	ab
관악	3238.90	ab
중랑	3044.00	abc
강서	2986.70	abcd
구로	2955.00	abcde
마포	2720.80	abcdef
강동	2644.10	bcdef
노원	2600.50	bcdef
광진	2512.10	bcdef
강북	2478.20	bcdefg
동대문	2465.80	bcdefg
양천	2254.70	cdefgh
강남	2208.30	defgh
용산	2163.10	defgh
금천	2076.90	efghi
서대문	1897.10	fghij
동작	1835.50	fghijk
성동	1590.50	ghijk
도봉	1506.90	hijk
은평	1304.55	ijk
서초	1257.05	jk
종로	1190.85	jk
중구	1166.50	jk
성북	1053.15	k

01 주제 선정
02 전처리 및 EDA
2.1 유동인구 데이터
2.2 CCTV 데이터
2.3 카드 사용 데이터
2.4 범죄 데이터
2.5 도시위험도 데이터
2.6 전월세 데이터
2.7 역 · 정류장 데이터
03 클러스터링
04 다음주 예고

02 범죄 데이터 분석

행정동 기준 맞추기

행정구	절도	폭력	살인	강간	강도
강남구	1956.75	2273.778	5.944	236	21.916

행정구 기준

(25 rows, 6 columns)

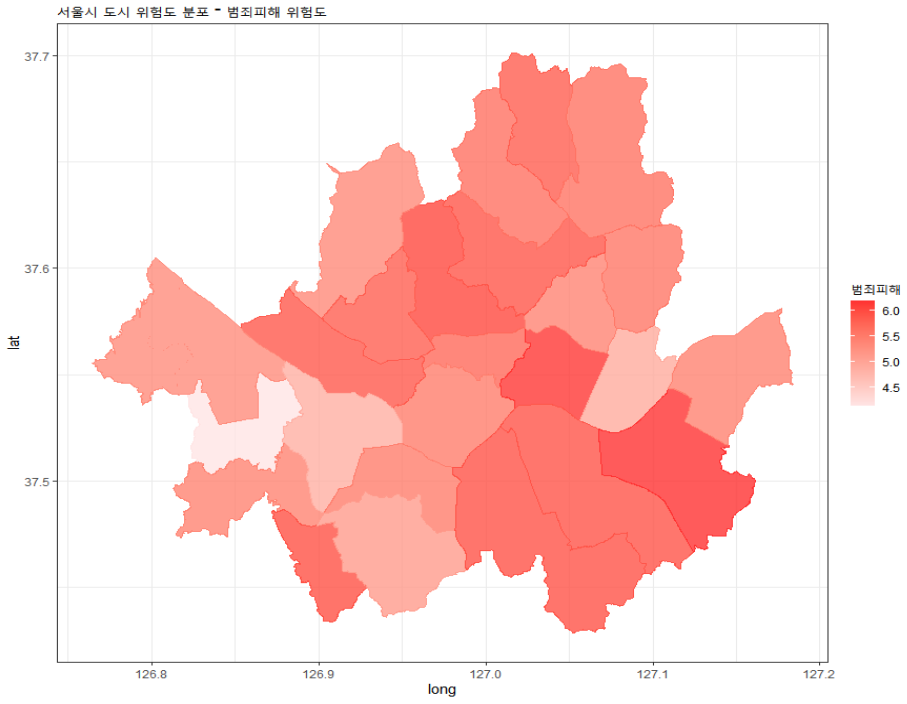
행정구	행정동	절도	폭력	살인	강간	강도
강남구	대치1동	1956.75	2273.778	5.944	236	21.916
강남구	논현2동	1956.75	2273.778	5.944	236	21.916
강남구	논현1동	1956.75	2273.778	5.944	236	21.916
강남구	신사동	1956.75	2273.778	5.944	236	21.916

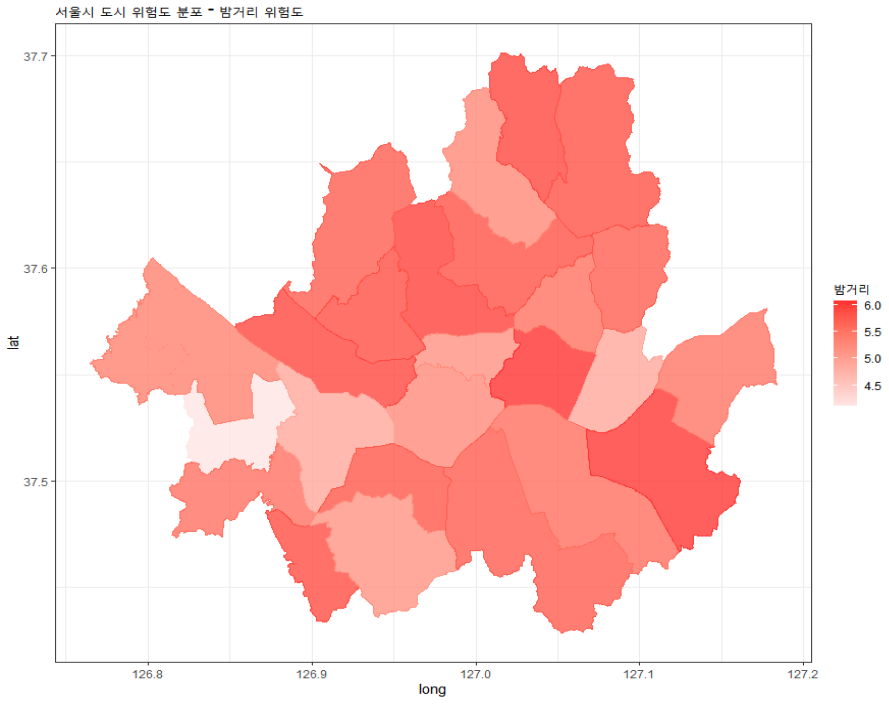
행정동 추가

(425 rows, 7colums)

01 주제 선정
02 전처리 및 EDA
2.1 유동인구 데이터
2.2 CCTV 데이터
2.3 카드 사용 데이터
2.4 범죄 데이터
2.5 도시위험도 데이터
2.6 전월세 데이터
2.7 역 · 정류장 데이터
03 클러스터링
04 다음주 예고

02 도시 위험도 구별 시각화 – 범죄피해와 밤거리 위험도

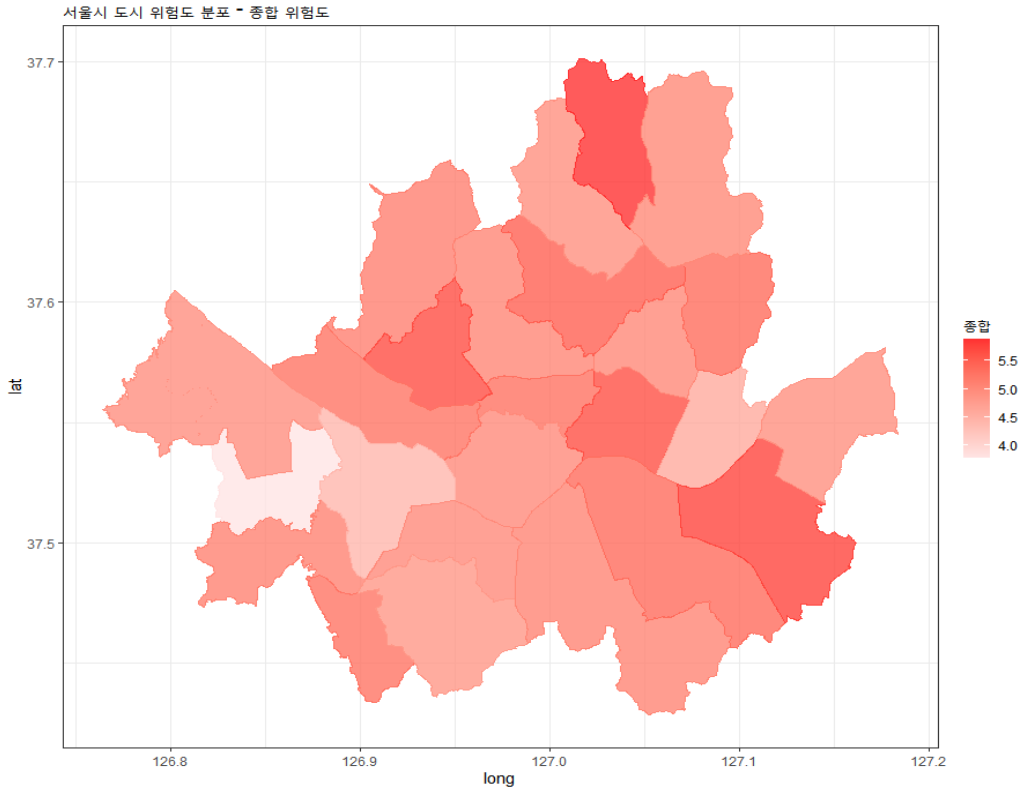




각 구의 **보안과 치안** 면에서 비슷한 항목들이기 때문에 양상이 비슷하게 나타남.

01 주제 선정
02 전처리 및 EDA
2.1 유동인구 데이터
2.2 CCTV 데이터
2.3 카드 사용 데이터
2.4 범죄 데이터
2.5 도시위험도 데이터
2.6 전월세 데이터
2.7 역 · 정류장 데이터
03 클러스터링
04 다음주 예고

02 도시 위험도 구별 시각화 – 종합 위험도



- 종합 위험도 상위 5개구
  : 도봉구, 송파구, 서대문구, 성동구, 성북구
- 종합 위험도 하위 5개구
  : 강북구, 관악구,광진구, 영등포구, 양천구
- 동별 데이터가 아닌 **구별 데이터**
  → 활용 방법 추후 논의 예정

01 주제 선정
02 전처리 및 EDA
2.1 유동인구 데이터
2.2 CCTV 데이터
2.3 카드 사용 데이터
2.4 범죄 데이터
2.5 도시위험도 데이터
2.6 전월세 데이터
2.7 역・정류장 데이터
03 클러스터링
04 다음주 예고

02 역-정류장 데이터 전처리

[서울시 버스정류장 위치 정보 데이터]

정류소번호	정류소명	X좌표	Y좌표
100	종로2가사거리	126.9877	37.56977
	벽원	126.9966	37.57918
		126.9983	37.58267
		126.9876	37.56858

역지오코딩(Reverse Geo-coding)

: 좌표를 주소로 바꾸는 것

X,Y 좌표(경도,위도) > 카카오 Map API > 주소 > 행정동

01 주제 선정

02 전처리 및 EDA

2.1 유동인구 데이터

2.2 CCTV 데이터

2.3 카드 사용 데이터

2.4 범죄 데이터

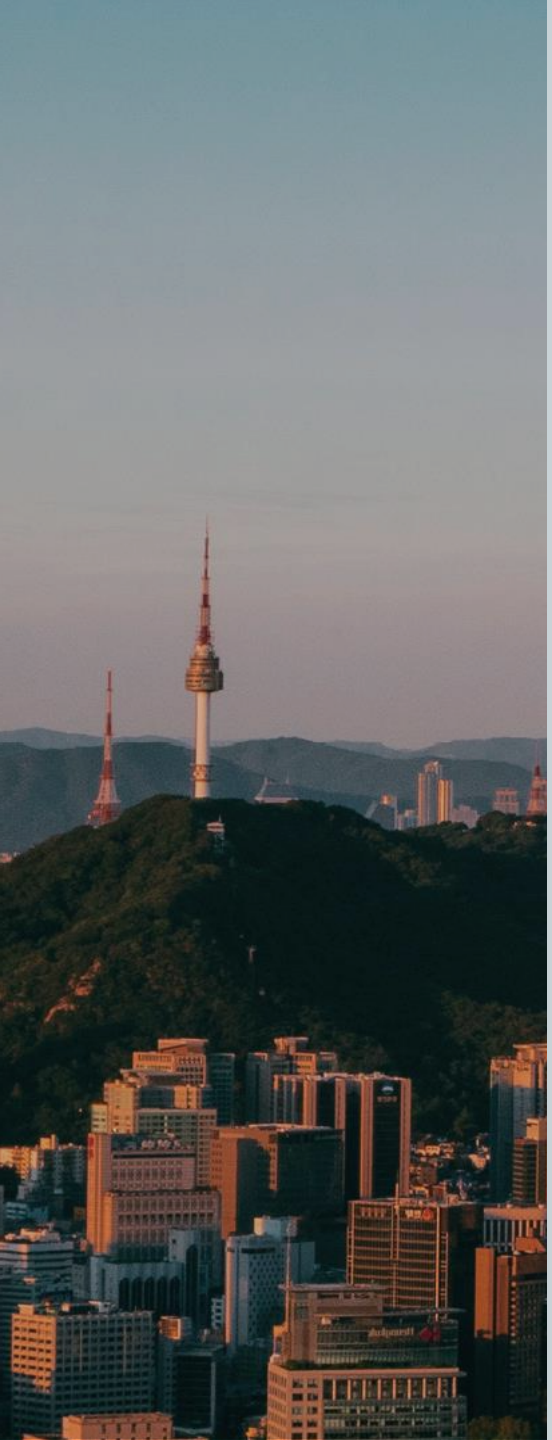
2.5 도시위험도 데이터

2.6 전월세 데이터

2.7 역 · 정류장 데이터

03 클러스터링

04 다음주 예고



클러스터링

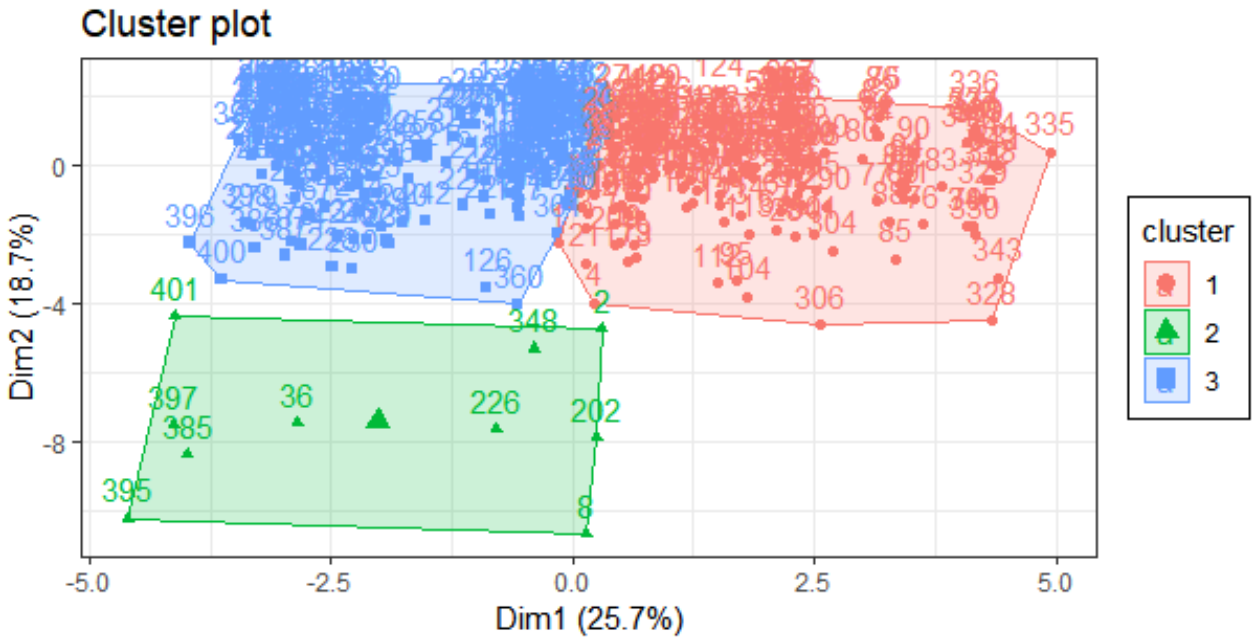
3

3.1 모델링 예시

3.2 K-means 클러스터링

03 K-means Clustering

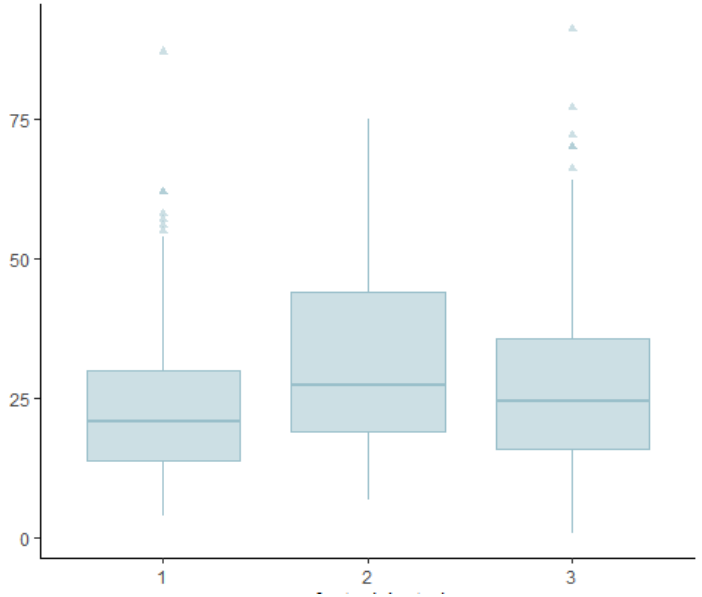
Cluster plot: 3개의 집단으로 클러스터링한 결과

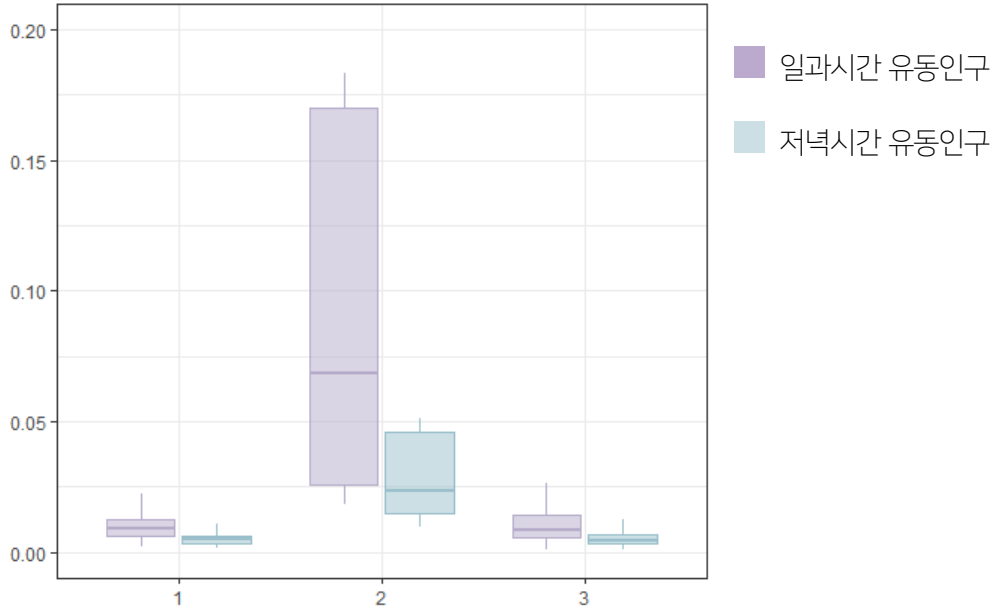


01 주제 선정
02 전처리 및 EDA
03 클러스터링
3.1 K-means 모델링
3.2 결과 해석
04 다음주 예고

03 K-means Clustering : 결과해석

시간대별 유동인구, 버스정류장 수 : 라이프스타일, 교통 인프라





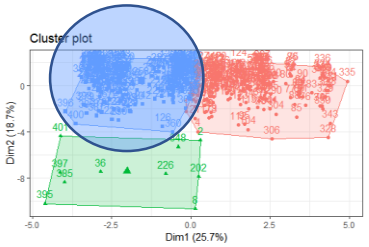
클러스터 2가 다른 클러스터들보다
주민등록인구 대비 낮 유동인구, 밤 유동인구가 월등히 많음+ 버스정류장 수 많음

→ 교통 중심지일 것으로 예상

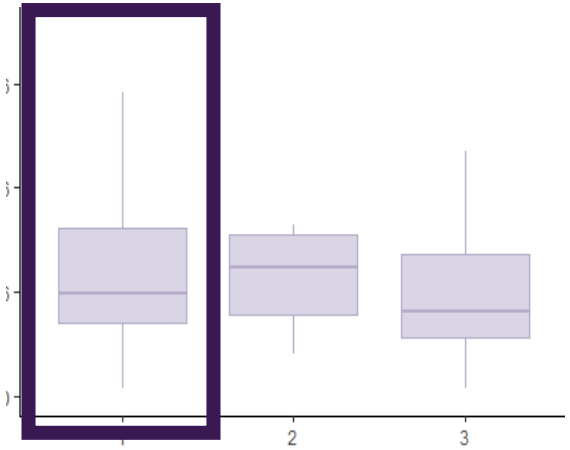
01 주제 선정
02 전처리 및 EDA
03 클러스터링
3.1 K-means 모델링
3.2 결과 해석
04 다음주 예고

03 결과해석

클러스터 1: 상업 지구 혹은 서울 외곽 지역

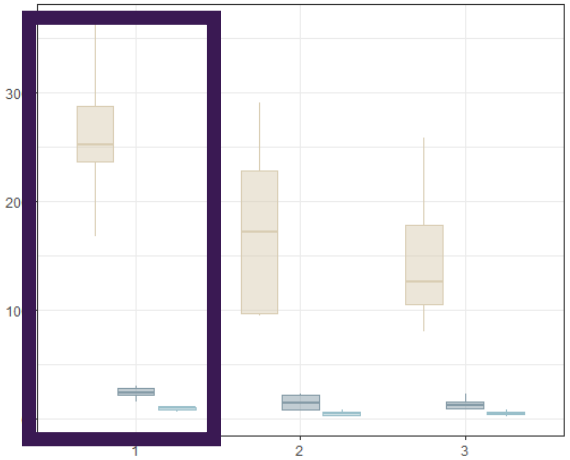


행정동별
매출액 평균



행정동별 매출액 평균의
분포가 가장 넓고,
다른 클러스터에 비해 높음

절도, 폭력
강도 발생
건수 분포

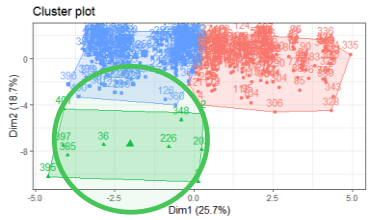


세 죄종 모두에서
다른 클러스터에 비해 높음
→치안이 불안정한 지역

01 주제 선정
02 전처리 및 EDA
03 클러스터링
3.1 K-means 모델링
3.2 결과 해석
04 다음주 예고

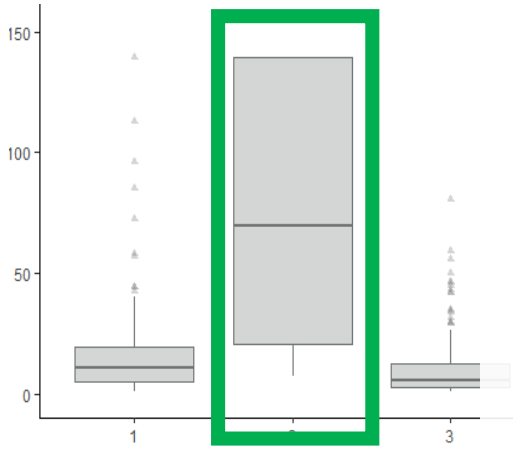
03 결과해석

클러스터 2: 교통의 중심지, 핫플레이스

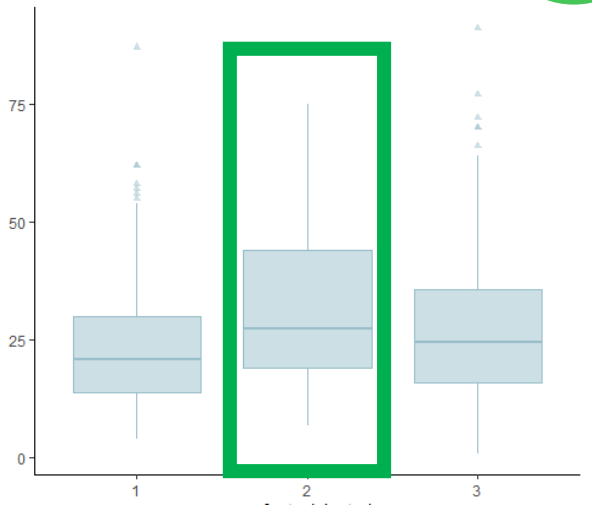


일과시간 유동인구

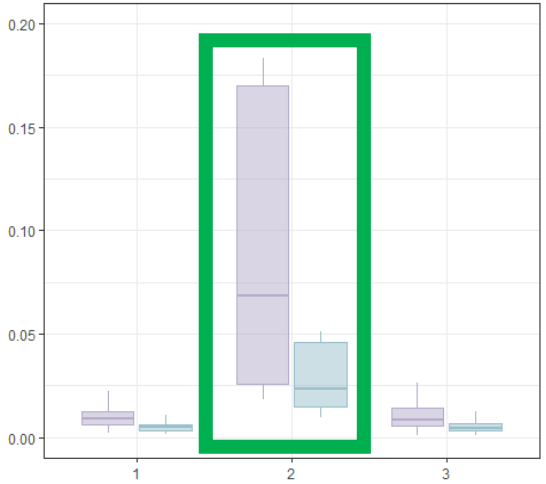
저녁시간 유동인구



행정동별
문화 지수



행정동별
버스정류장 불포



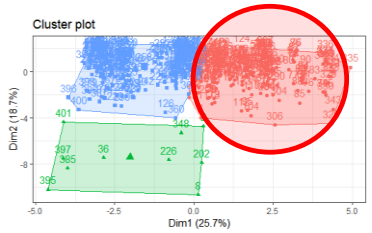
행정동별
낮, 밤 유동인구
평균

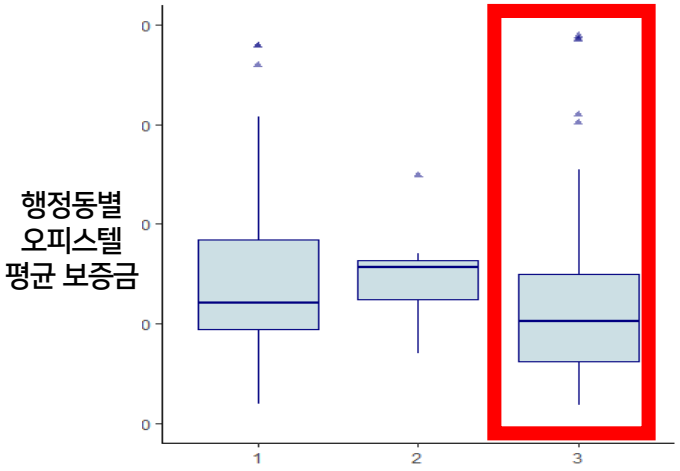
클러스터 2가 다른 클러스터들보다
주민등록인구 대비 낮, 밤 유동인구가, 버스정류장 수 많음+ 문화 지수 높음
→ 교통 중심지일 것으로 예상

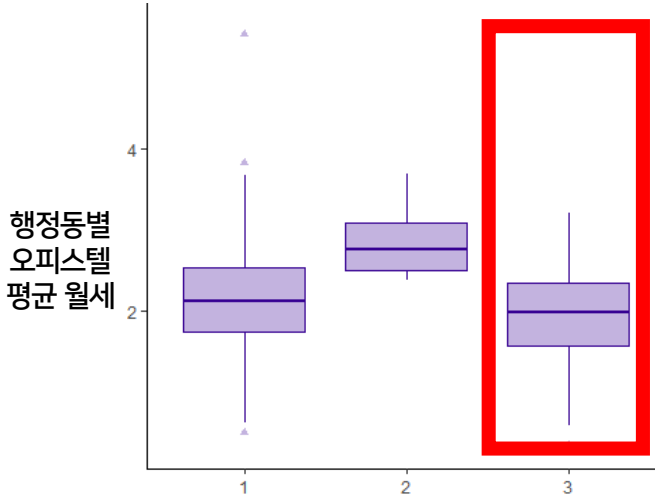
01 주제 선정
02 전처리 및 EDA
03 클러스터링
3.1 K-means 모델링
3.2 결과 해석
04 다음주 예고

03 결과해석

클러스터 3: 전형적인 주거 구역







클러스터 3이 다른 클러스터들보다
집값 시세가 높지 않고 문화 지표, 유동인구 지표 모두에서 양호
→ 사용자에게 적절한 주거 구역!

01 주제 선정
02 전처리 및 EDA
03 클러스터링
3.1 K-means 모델링
3.2 결과 해석
04 다음주 예고