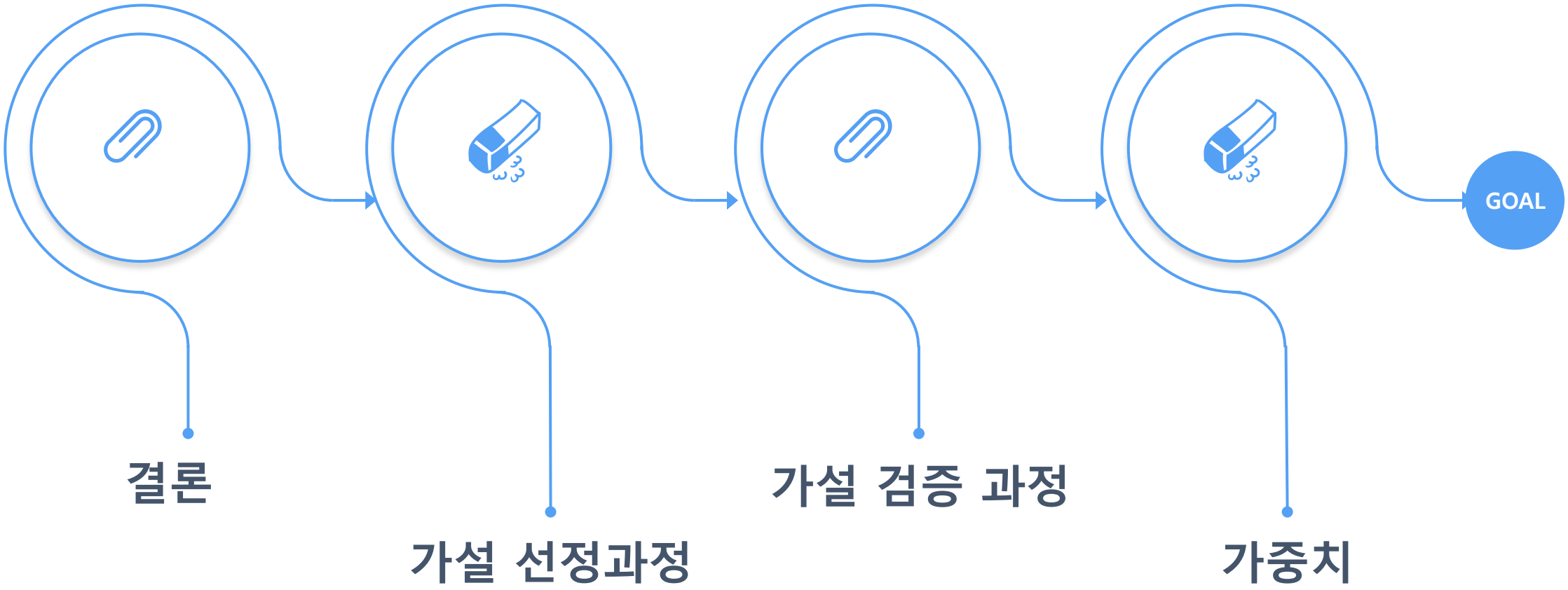
정류장&노선수 추가 설치 자치구 찾기

(데이터 분석 및 인사이트 도출)

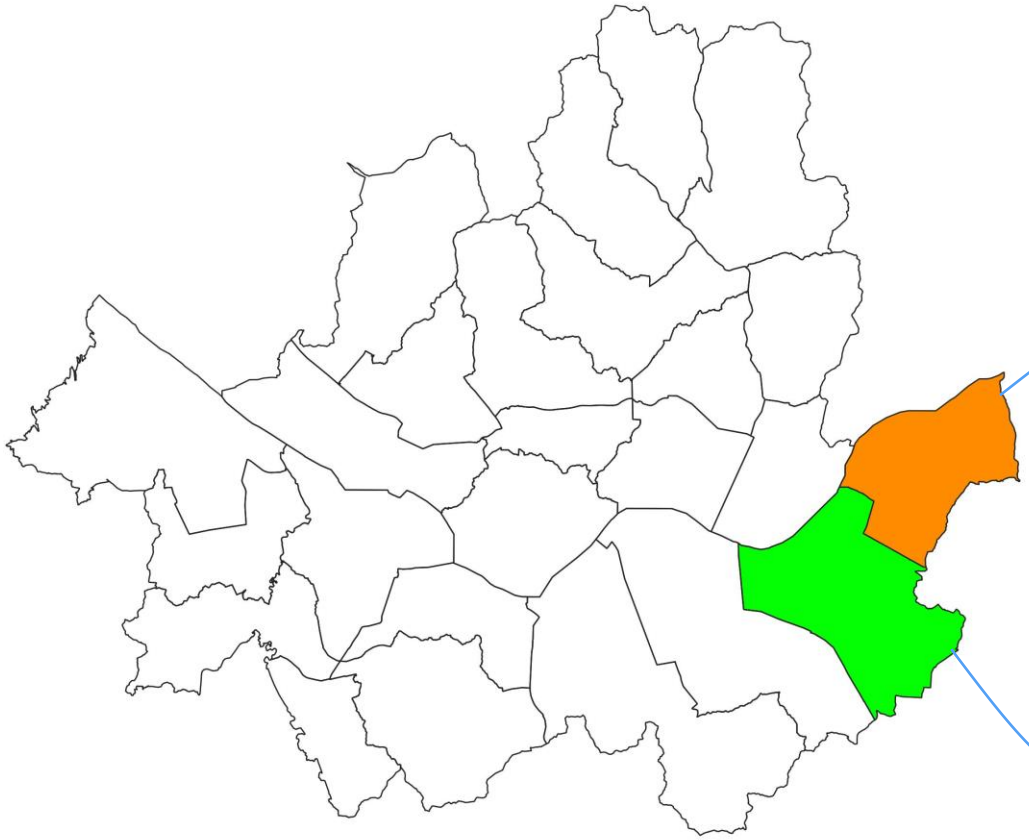
팀명 : 아리아 붐바야 틀어조

팀원 : 201810866 융합전자공학과 김 건
201810892 융합전자공학과 오세욱
201810871 융합전자공학과 문성준
201810930 컴퓨터과학과 김현재
202010139 공간환경학부 이현수

목차



최종 결론- 송파구에 정류장 설치, 강동구에 노선 설치 필요



강동구

송파구

최종 결론 – 송파구에 정류장 추가 설치,
강동구에 노선수 추가 설치가 필요

가설 선정 과정

전제 조건 : 결측치 처리 & 그룹화 기준

'1.2 seoul_moving_month_4'

결측치 처리 |
'1.2 seoul_moving_month_4' 파일의 이동인구 (합) 컬럼 *을 1과 3의 중간값인 2로 변경

```
seoul_moving.loc[seoul_moving['이동인구(합)'] == '*'] / 결측치 확인
seoul_moving = seoul_moving.replace({'이동인구(합)':'*'},'2') /2로 변경
seoul_moving = seoul_moving.astype({'이동인구(합)':'float'}) /자료형 변경
```

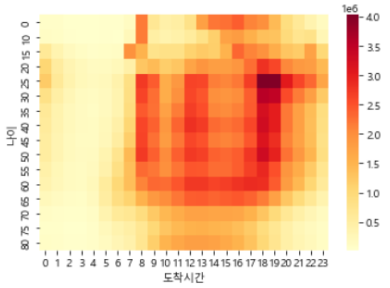
그룹화 기준 |
- 출발 시군구 코드 기준 / 도착 시군구 코드 기준으로 나누어 정류장 수와 노선수 관계
- 이동인구(합)의 평균과 합을 둘 다 확인

```
work_ch = da.groupby(by=["도착 시군구 코드"], as_index=False)['이동인구(합)'].mean()
work_d = da.groupby(by=["출발 시군구 코드"], as_index=False)['이동인구(합)'].mean()
work_ch1 = da.groupby(by=["도착 시군구 코드"], as_index=False)['이동인구(합)'].sum()
work_d1 = da.groupby(by=["출발 시군구 코드"], as_index=False)['이동인구(합)'].sum()
```

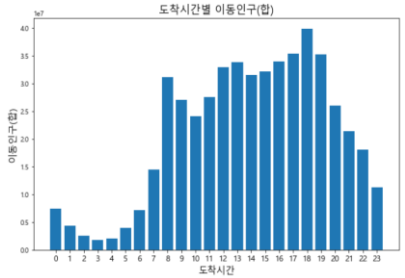
가설 1 & 2 선정 과정

- 1.2 이동인구 합의 수치를 기준으로 다양한 변수의 관계를 시각화

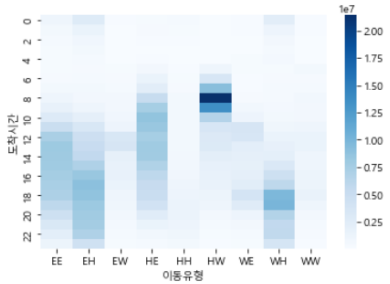
20대의 도착시간 18시인
이동인구(합)이 가장 높은 값을 가짐
또한 7시와 18시 주변 시간의
이동인구(합)이 전체적으로
높은 값을 보임
10대는 평일 9~15시 사이에
학교를 가기 때문에 이동인구가 낮게 찍힘



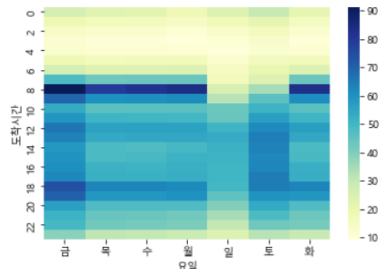
도착시간이 7-19시 사이인
이동인구(합)이 비교적 큰 값을 가졌고,
20시 이후에는 작은 값을 가짐



이동유형이 HW이고
도착시간이 8시 주변의 데이터는
이동인구(합)이 큰 값을 가짐
또한 이동유형 WH이고
도착시간 18시 주변의 데이터도
비교적 큰 값을 가짐



도착시간 8시 주변에
이동인구(합)은 큰 값을 가졌고 ,
18시에도 비교적 큰 값을 가짐



가설 1

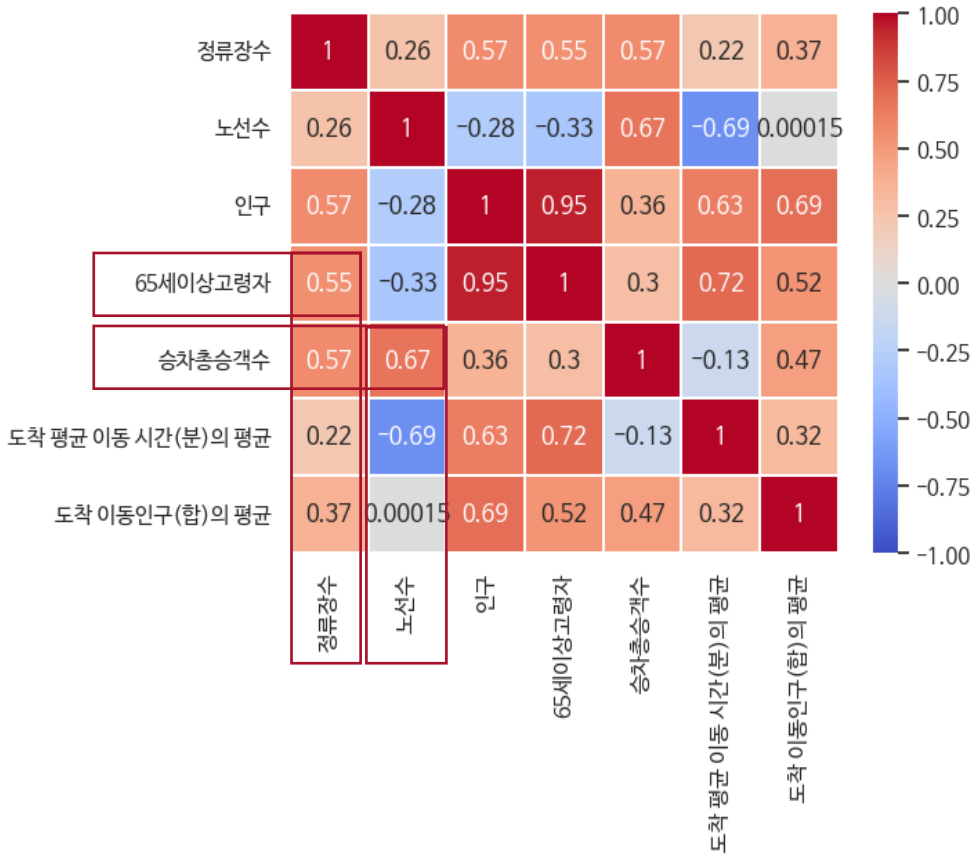
10대의 이동인구 ⇔ 정류장 수

가설 2

직장인 출퇴근 시간 유동인구의 수 ⇔ 정류장 수

가설 3 & 4 선정 과정
- 1.1 & 1.3의 다양한 변수의 관계를 시각화

'1.1 BUM_STATION_BOARDING_MONTH_202204' & '1.3 seoul_people'에서
비슷한 결과를 내는 요소들을 제외하고 히트맵으로 변수들 간 상관관계 나타냄



가설 3

65세 이상 고령자 수 ⇔ 정류장 수

가설 4

승 하차 총 승객 수 ⇔ 노선 수

선정한 가설

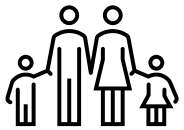
가설 1 | H1. 10대의 이동인구와 정류장 수는 관련이 있다.

가설 2 | H1-1. 직장인 출근 시간 유동인구의 수와 정류장 수는 관련이 있다.

H1-2. 직장인 퇴근 시간 유동인구의 수와 정류장 수는 관련이 있다.

='20대~60대'

가설 3 | H1. 65세 이상 고령자 수는 정류장 수와 관련이 있다.



가설 4 | H1. 승 하차 총 승객수는 노선수와 관련이 있다.

가설 1

대립가설(H1): 10대의 이동인구와 정류장 수는 관련이 있다.

<대립가설> h1: 10대 이동 인구는 버스 정류장 수와 관련이 있다.

1. 나이(10대)로 필터링 – (학생의 나이)

2.월요일부터 금요일에서 9시~15시 사이의 데이터를 제거하는 필터링 - (일반적으로 학교에 있을 시간으로 간주)

10대 이동인구 합과 정류장 수는 상관관계가 있음

- H1 : **10대의 이동인구와 정류장 수는 관련이 있다.**

채택

중간 정도의 상관관계가 있음

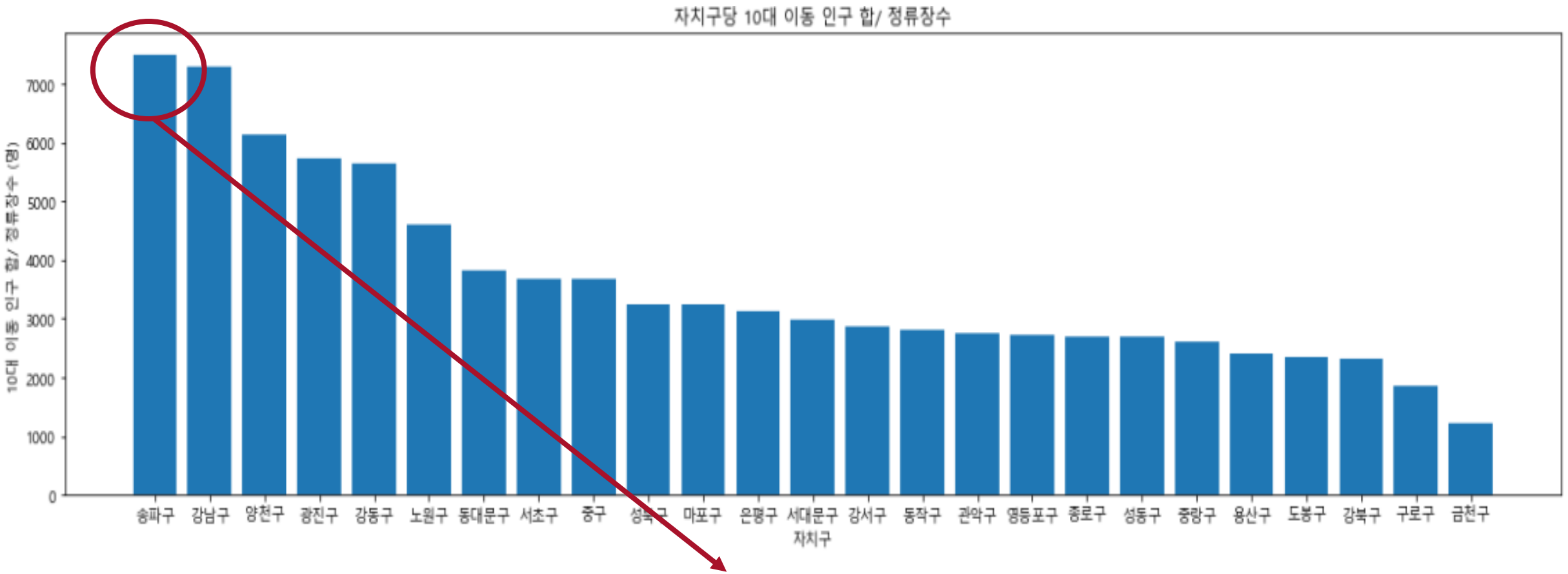
가설	r(상관계수)	p
H1	.44	.003(<.05)

```
spst.pearsonr(ten_age_sum['10대 이동 인구 합'], ten_age_sum['정류장수'])
(0.4461223959520076, 0.0253970152127875)
```

Evans JD. (1996)

Cohen, J.(1988). Statistical power analysis for the behavioral sciences

결론 : 송파구에 정류장 추가 설치 필요



- 정류장 당 10대의 이동인구 합(명)이 가장 많은 구는 송파구

가설 2

대립가설(H1-1) : 직장인 출근 시간 유동인구의 수와 정류장 수는 관련이 있다.

(H1-2) : 직장인 퇴근 시간 유동인구의 수와 정류장 수는 관련이 있다.

<대립가설> H1-1 : 직장인 출근 시간 유동인구의 수와 정류장 수는 관련이 있다.

H1-2 : 직장인 퇴근 시간 유동인구의 수와 정류장 수는 관련이 있다.

1. 나이(20대-60대)로 필터링 – (직장인의 나이)

2. 출근과 퇴근의 시간을 정의해서 필터링(출근 6~10 퇴근 17~21)

3. 출근과 퇴근을 나누어 이동 유형을 필터링 (HW[야간에서 주간 상주지 – 출근] ,

WH[주간에서 야간 상주지 – 퇴근])

3. 평일로 필터링 (보통 평일에 출근)

직장인 출/퇴근 시간 유동인구의 수와 정류장 수는 상관관계가 있음

- H1-1 : 직장인 출근 시간 유동인구의 수와 정류장 수는 관련이 있다.
- H1-2 : 직장인 퇴근 시간 유동인구의 수와 정류장 수는 관련이 있다.

채택

중간 정도의 상관관계가 있음

가설	r(상관계수)	p
H1-1(출근)	.53	.006(<.05)
H1-2(퇴근)	.52	.007(<.05)

-> H1-1(출근)

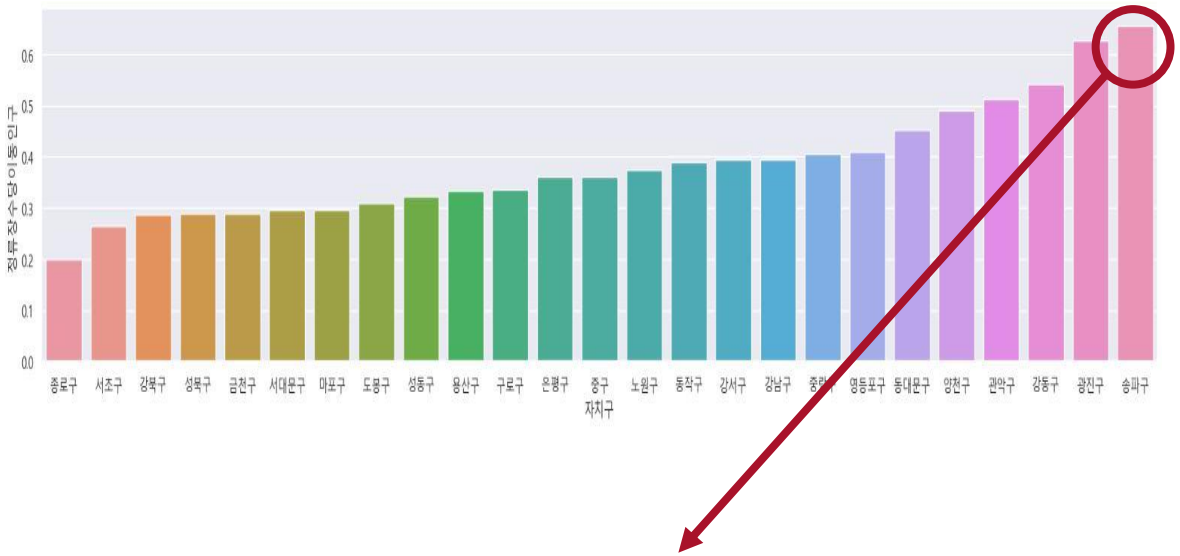
-> H1-2(퇴근)

Evans JD. (1996)

Cohen, J.(1988). Statistical power analysis for the behavioral sciences

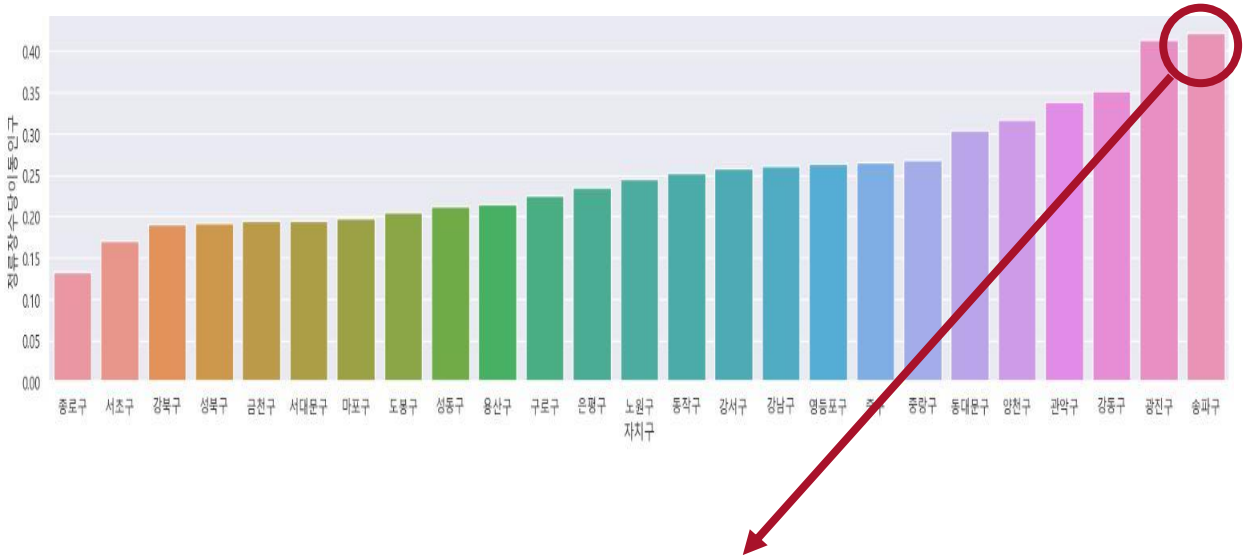
결론 : 송파구에 정류장 추가 설치 필요

- 출근(H1-1),
  정류장 수 당 이동인구(합)이 가장 큰 구



정류장 수 당 이동인구(합)이
가장 큰 구는 송파구

- 퇴근(H1-1),
  정류장 수 당 이동인구(합)이 가장 큰 구



정류장 수 당 이동인구(합)이
가장 큰 구는 송파구

가설 3

대립가설(H1): 65세 이상 고령자 수는 정류장 수와 관련이 있다.

<대립가설> h1: 65세 이상 고령자 수는 정류장 수와 관련이 있다.

1. 두 개 변수의 상관관계와 p-value 값을 확인한다.

2. 정류장 수와 65세 이상 고령자 수의 관계를 확인하기 위해 65세 이상 고령자 수를 정류장 수로 나누어 확인해 보았다.

```
only_65['정류장당 65세이상고령자수'] = only_65['65세이상고령자'] / only_65['정류장수']

only_65.to_csv('only_65.csv', index=False)

df = only[['자치구', '정류장수', '정류장당 65세이상고령자수']]
```

65세 이상 고령자 수와 정류장 수는 상관관계 있음

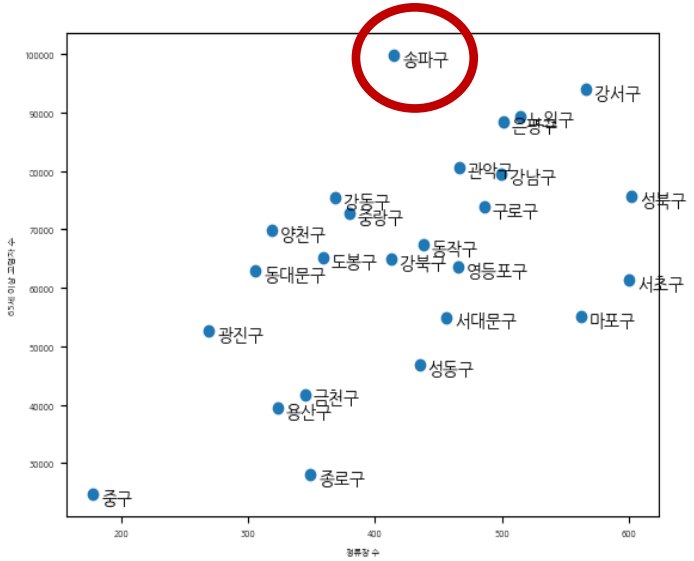
- H1 : 65세 이상 고령자 수와 정류장 수는 관련이 있다.

채택

중간 정도의 상관관계가 있음

가설	r(상관계수)	p
H1	.55	.004(<.05)

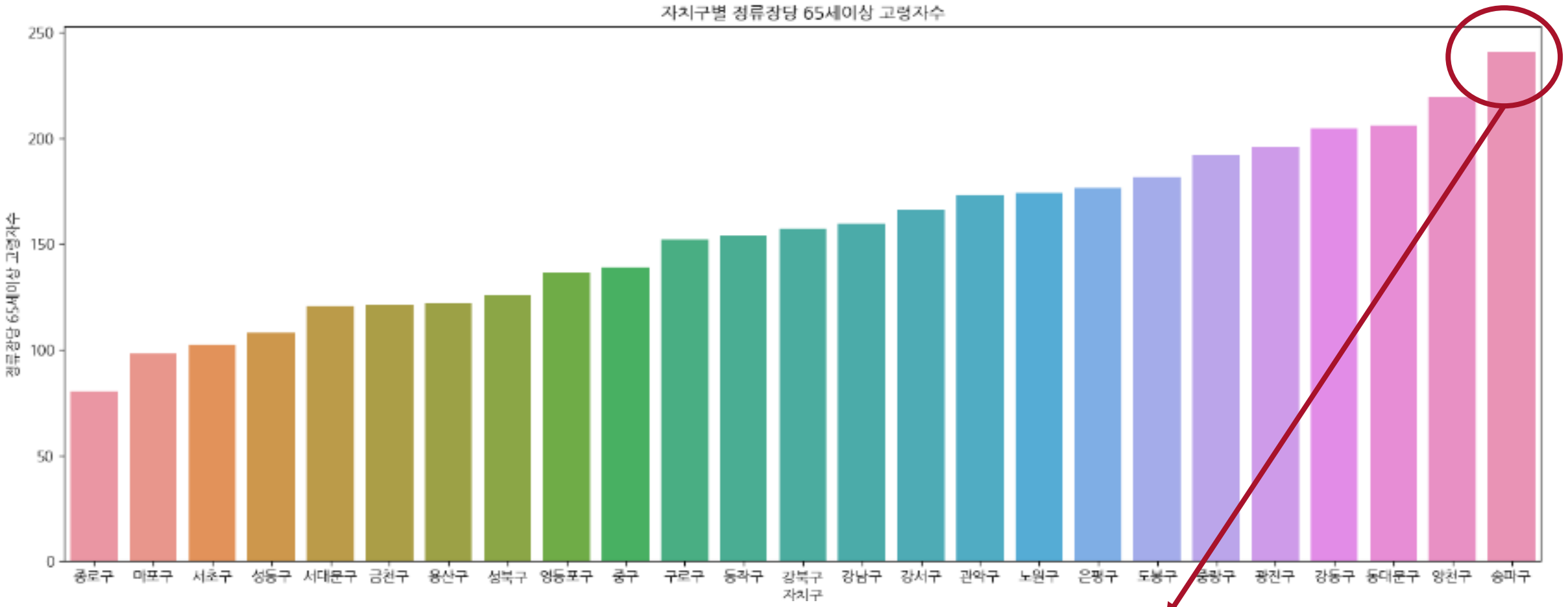
```
import scipy.stats as spst
spst.pearsonr(only_65['65세이상고령자'],only_65['정류장수'])

PearsonRResult(statistic=0.5544251917556622, pvalue=0.0040271348895971960)
```



Evans JD. (1996)

Cohen, J.(1988). Statistical power analysis for the behavioral sciences

결론 : 송파구에 정류장 추가 설치 필요



- 정류장 당 65세 이상 고령자(명)가 가장 많은 구는 송파구

가설 4

대립가설(H1): 승 하차 총 승객수는 노선수와 관련이 있다.

<대립가설> H1: 총 승객수와 노선 수는 관련이 있다.

1. 승차 총 승객수와 하차 총 승객수는 매우 강한 상관관계가 있으므로 합했다.
2. 노선 수와 총 승객수의 관계를 확인하기 위해 총 승객수를 노선수로 나누어 확인해보았다.

```
seoul_sum['총 승객수'] = seoul_sum['승차총승객수'] + seoul_sum['하차총승객수']
seoul_sum['노선당 승객수'] = seoul_sum['총 승객수'] / seoul_sum['노선수']
```

승하차 총 승객수와 노선수는 상관관계가 있음

- H1 : 승하차 총 승객수는 노선수, 정류장수와 관련이 있을 것이다.

채택

강한 정도의 상관관계가 있음

가설	r(상관계수)	p
H1	.66	.003(<.05)

```
import scipy.stats as spst
spst.pearsonr(seoul_sum['노선수'],seoul_sum['총 승객수'])
```

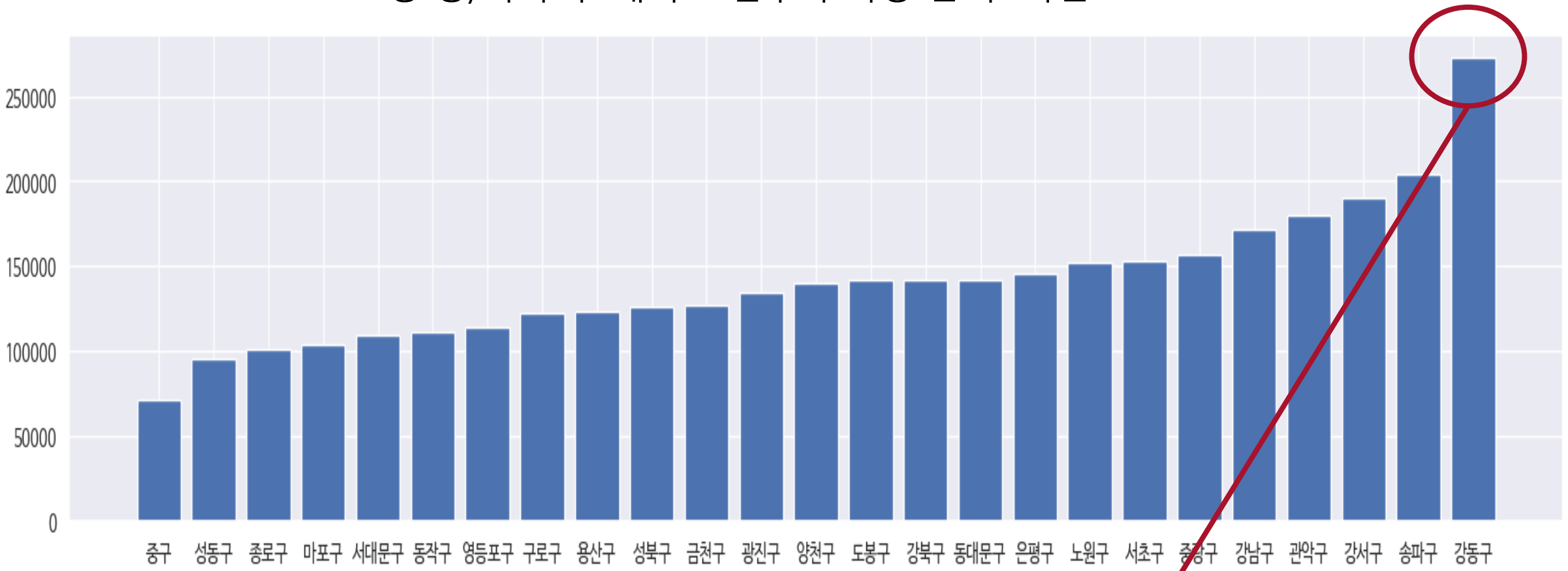
```
PearsonRResult(statistic=0.6596714851091889, pvalue=0.0003340267759449194)
```

Evans JD. (1996)

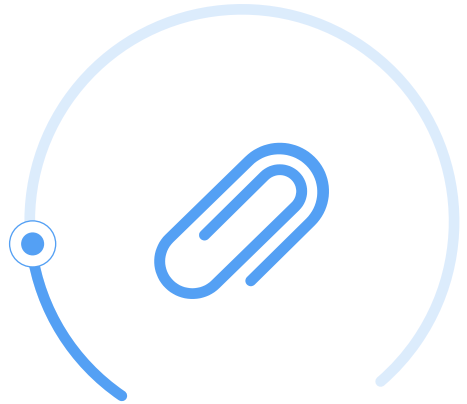
Cohen, J.(1988). Statistical power analysis for the behavioral sciences

결론 : 강동구에 노선수 추가 설치 필요

-총 승/하차 수 대비 노선수가 가장 큰 구 확인



- 승/하차 수 대비 노선수가 가장 큰 구는 강동구



최종 결론

결론 총합

가설1.

정류장수 당 10대 유동인구 합의 평균

송파구

가설2.

정류장수 당 직장인 유동인구 합의 평균 (출근/퇴근)

송파구

가설3.

정류장수 당 65세 이상 노인수

송파구

가설4.

노선수 당 총 승객 수(승차,하차)

강동구

가중치 결정

•최종 결론을 정하는 가중치 결정

정류장

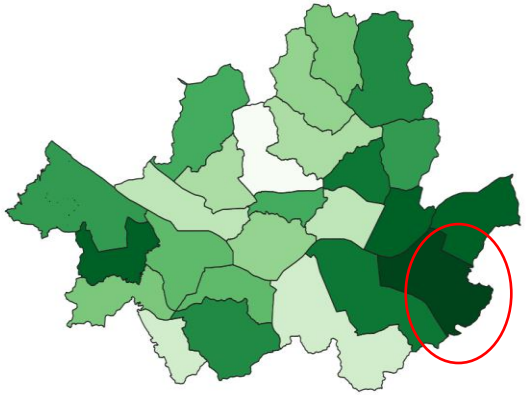
	자치구	10대 이동 인구 합	65세이상인구	출근	퇴근	총합
0	송파구	0.44	0.55	0.53	0.52	2.04
1	광진구	0.34	0.45	0.50	0.51	1.80
2	강동구	0.33	0.47	0.43	0.44	1.67
3	양천구	0.36	0.51	0.39	0.39	1.65
4	강남구	0.43	0.37	0.32	0.32	1.44

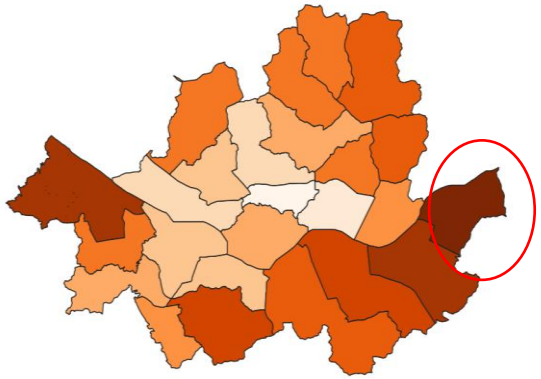
노선

	자치구	총승객수
0	강동구	0.66
1	송파구	0.50
2	강서구	0.46
3	관악구	0.44
4	강남구	0.42

1. 각 가설에서 도출된 '구별 ()값을 정류장수/노선수로 나눈 값'을 해당 열의 최댓값으로 나눠 모든 항목의 최댓값이 1이 되도록 함
2. 그 값에 각 가설에서 도출된 상관계수 값을 가중치로 곱함
3. 모든 가설에서 도출된 값의 총합으로 순위 결정

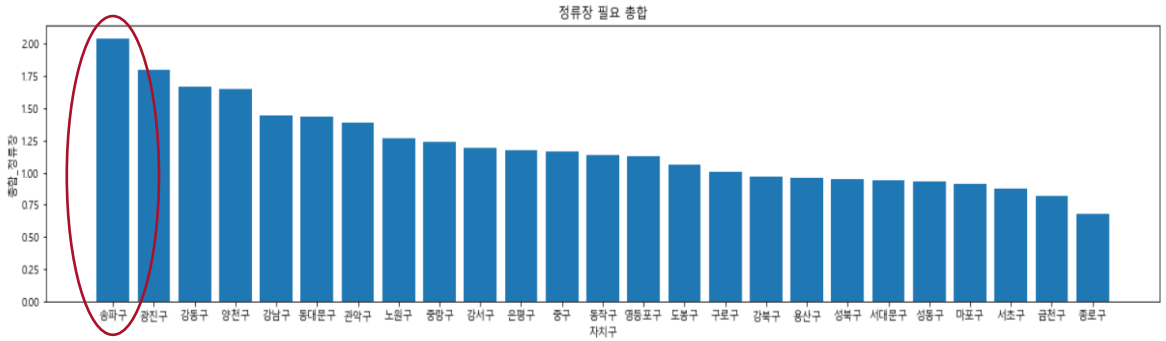
최종 결론

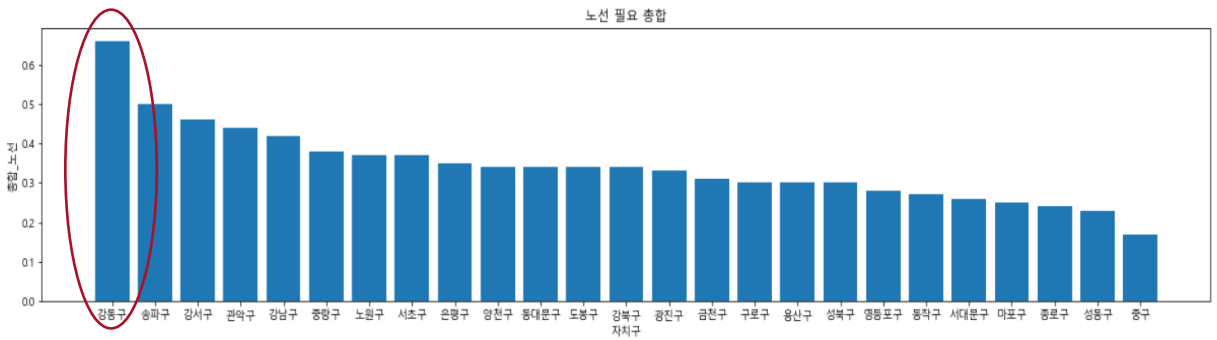




송파구에 정류장 추가 설치가 필요

강동구에 노선수 추가 설치가 필요





Appendix (부록) 중간 발표때!

부록

대립가설(H1): 버스이용밀도는 노선수, 정류장수와 관련이 있다.

버스이용밀도와 노선 수는 상관관계가 있음

•외부 데이터인 서울시 '면적' 데이터를 활용

- H1 : 버스이용밀도와 노선 수는 관련이 있다.
  ='승 하차 총 승객 수/ 면적'

채택

강한 정도의 상관관계가 있음

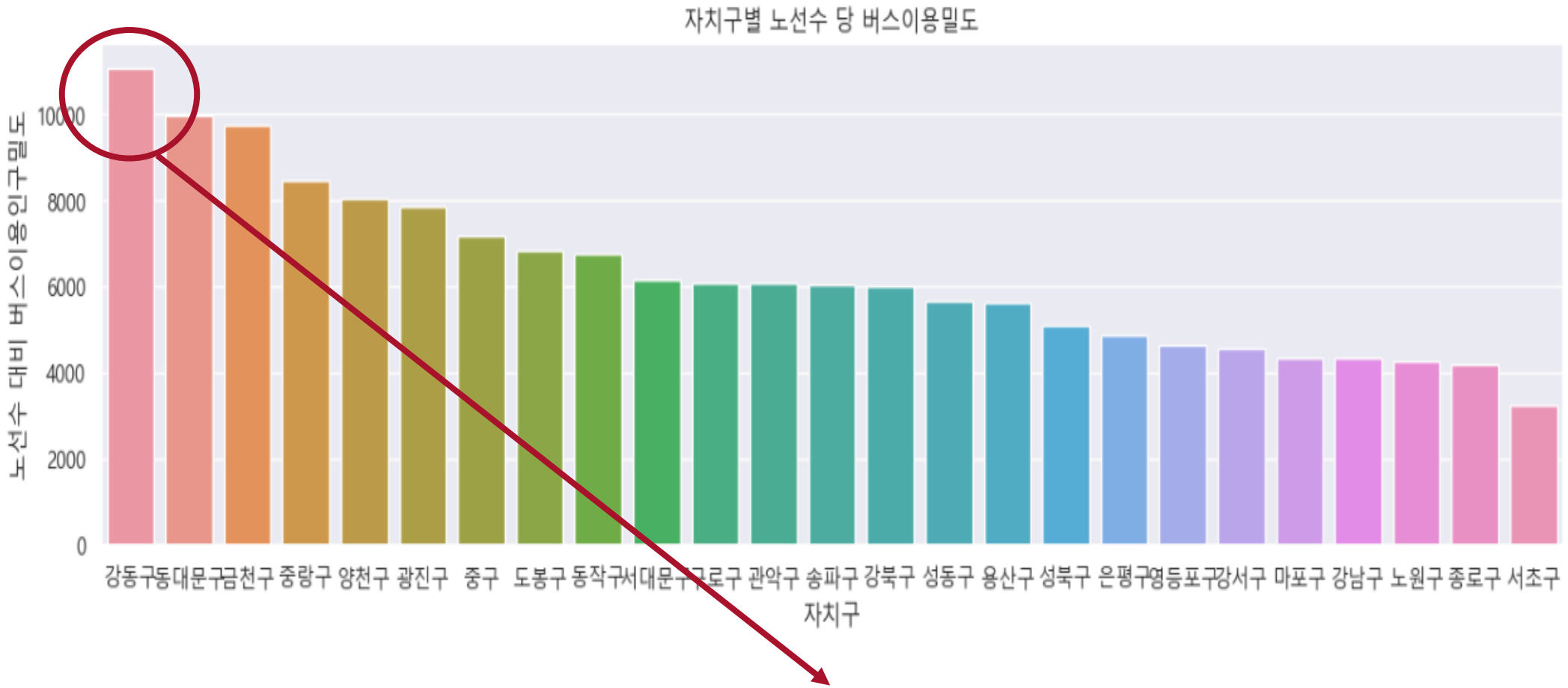
가설	r(상관계수)	p
H1	.60	.001(<.05)

```
import scipy.stats as spst
spst.pearsonr(result4['버스이용인구밀도'],result4['노선수'])

PearsonRResult(statistic=0.6028775215702739, pvalue=0.0014239989480590112)
```

Evans JD. (1996)

Cohen, J.(1988). Statistical power analysis for the behavioral sciences

결론 : 강동구에 노선수 추가 설치 필요



-버스이용밀도가 가장 높은 구는 강동구

감사합니다