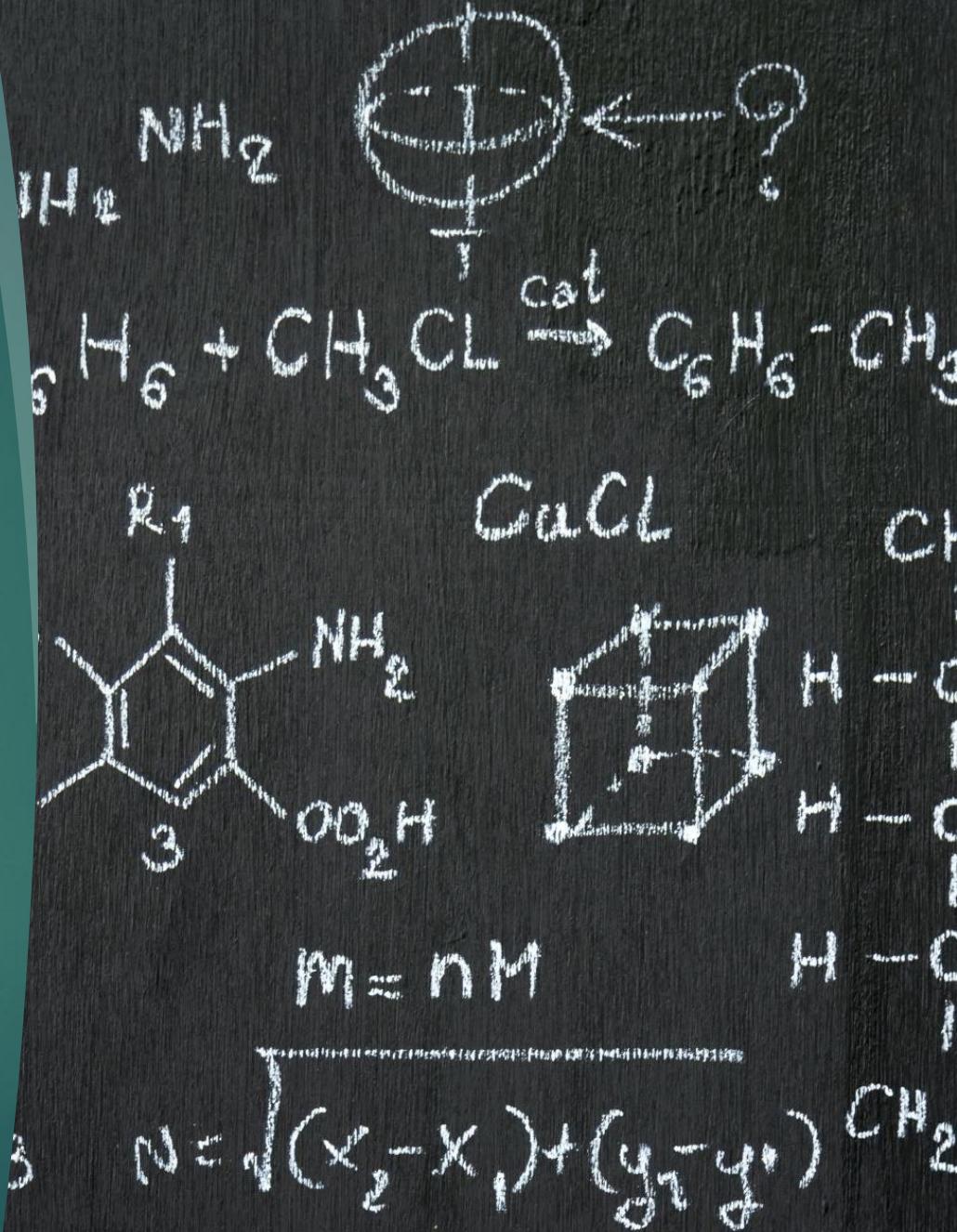


첫 번째 프로젝트 발표

AI+X 선도 인재 양성 기초 프로젝트

- ▶ 팀장: 강민준
- ▶ 팀원: 이제연, 이선역, 김유빈, 이현준
- ▶ 발표: 이제연, 이선역(2명)
- ▶ PPT 제작: 강민준(1명)
- ▶ 결과 시각화 및 검증 및 코드 정리: 이현준, 김유빈(각각 1명씩 담당)



목차

지하철 이용승객 분석 EDA 숙제

✓ 지하철 이용승객 EDA분석시 질문 리스트에 대한 정답과 실습파일 제출

- > Q) 2019.01~06중에 언제 지하철을 가장 많이 이용했을까? (기준: 승하자총승객수)
- > Q. 가설) 1월~6월중에 5월에 지하철 승객수가 많다? (기준: 승하자총승객수)
- > Q. 가설) 요일중에서 목요일에 지하철 승객수가 많다? (기준: 승하자총승객수)
- > Q) 연월 각각에 대해 일자별(월일별) 승하차총승객수 그래프 그려볼까요?(pointplot)
- > Q) 가장 승객이 많이 탄는 승차역은?
- > Q) 노선별로 역별/요일별 승차승객수를 비교해볼수 있을까? (1~9호선, 역별/요일별 heatmap)
- > Q) 1호선에서 가장 하자를 많이 하는 역은? (groupby)
- > Q) 2호선중에서 어느 역에서 승차가 가장 많이 발생할까? (Folium역 표시)



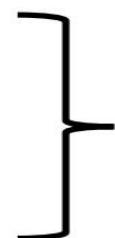
- 연속형 데이터 통계 (모수, 비모수)

해당 문제들의 답을 통계적으로 해석

- 상관분석

- 문제 외 아이디어 검증

- 외부 데이터 연동



소개하고 싶은 내용 몇가지 선별

가설 설정

- 귀무가설: 5월의 평균 승하자총승객수 \leq 다른 달의 평균 승하자총승객수
 - 대립가설: 5월의 평균 승하자총승객수 > 다른 달의 평균 승하자총승객수
- => 단측검정 t-test 시행

"5월이 더 많을 것이다"라는 한쪽 방향으로만 관심

=> 한쪽 방향으로만 차이를 검정할 때 → 단측검정(one-tailed test) 사용

가설 검증 1)

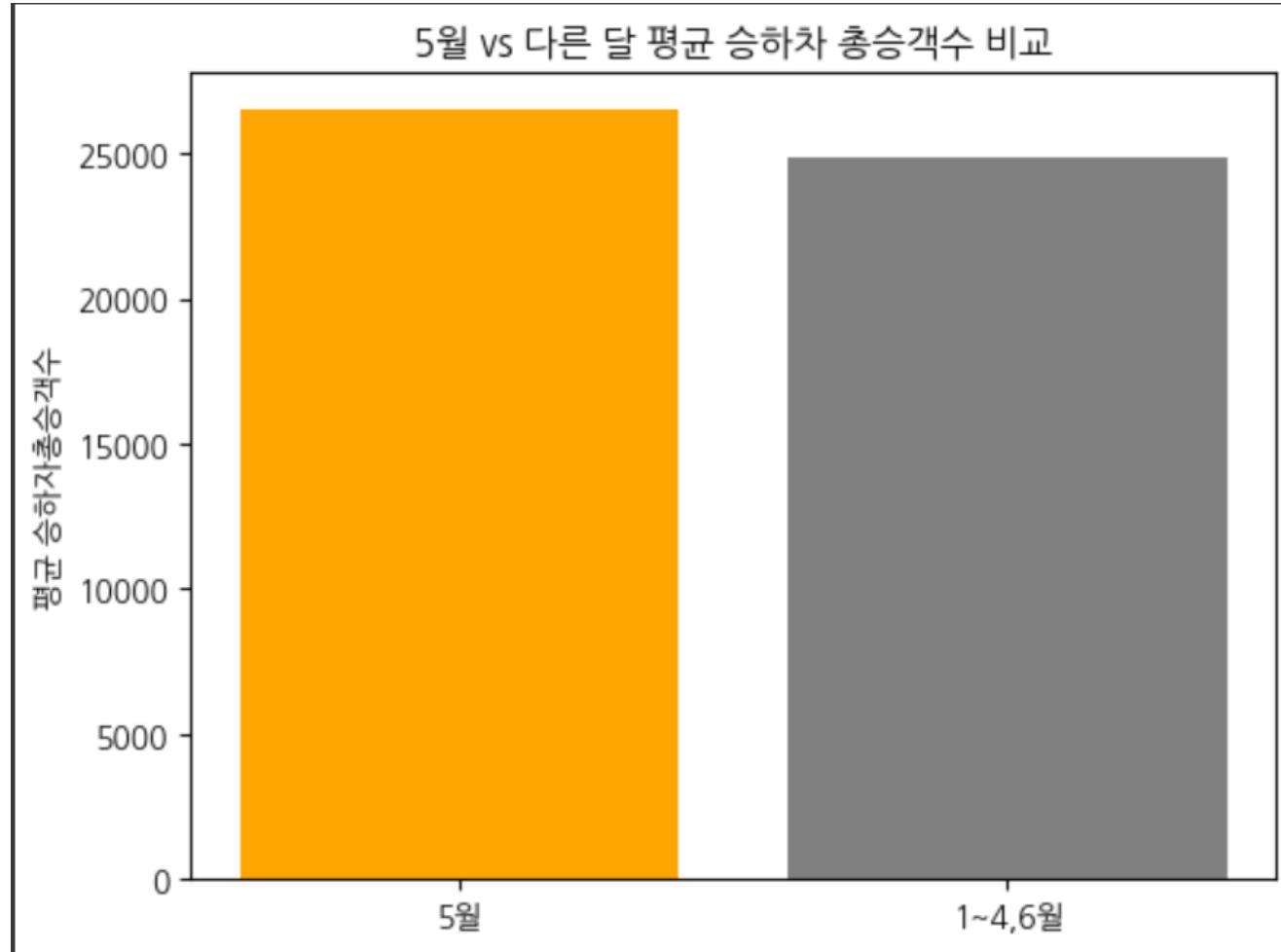


t통계량=7.640, p-value=0.0000000000001125886

p-value < 0.05 → 귀무가설 기각 → “5월의 승하차 인원이 유의하게 많다”

가설 검증 결과

가설 검증 그래프 1)



가설 설정

대립 가설 : 요일 중 목요일에 지하철 승객 수가 많다? (차이가 있을 것이다)

귀무 가설 : 차이가 없을 것이다.

=> Kruskal-Wallis 비모수 검정 진행

가설 검증 2)

가설 검증 결과 2)

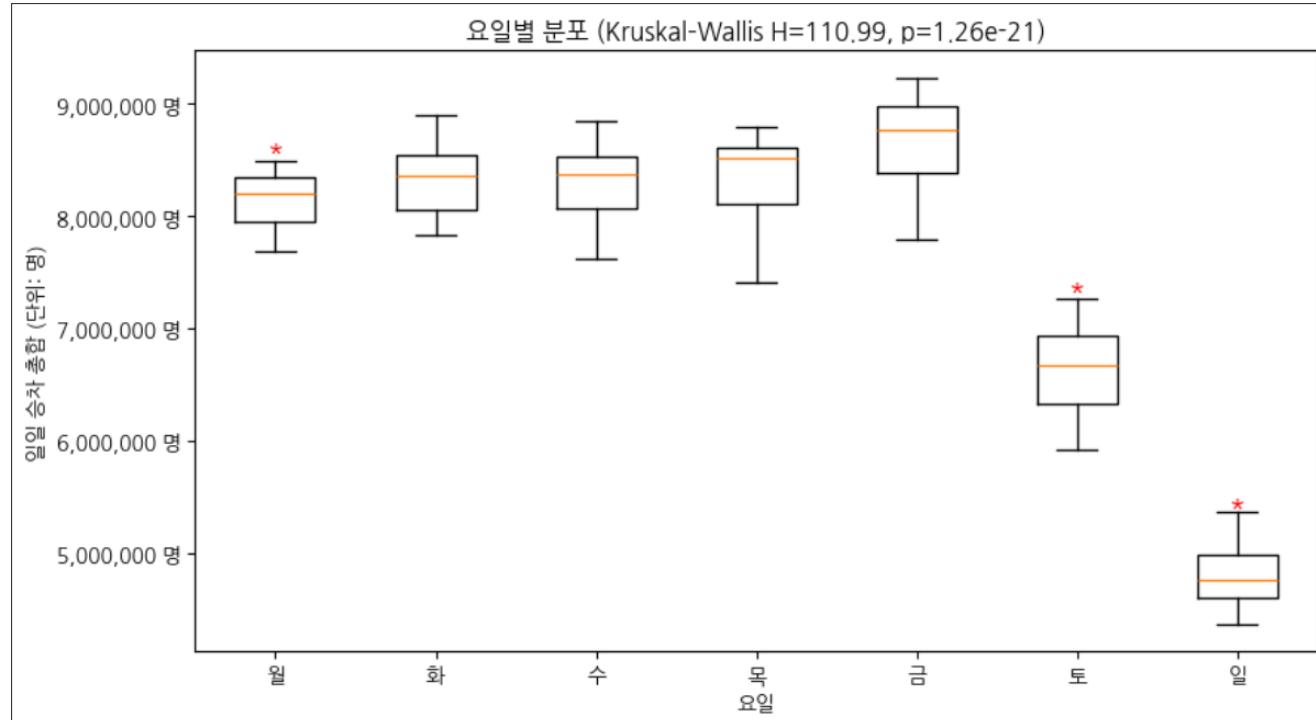
➡ [주중 단축 Mann-Whitney] U=2786.0, p=0.0009149

요일	count	mean	median	std
월	25	7.844900e+06	8199981.0	1.186185e+06
화	26	7.937223e+06	8355661.0	1.441557e+06
수	26	8.113369e+06	8375562.5	9.927739e+05
목	26	8.259079e+06	8509755.5	7.048665e+05
금	26	8.579001e+06	8771226.0	6.627577e+05
토	26	6.593336e+06	6664431.0	4.876941e+05
일	26	4.775233e+06	4756814.0	3.929390e+05

U = 2786 -> 목요일 data들이 더 높은 값을 차지

p-value < 0.05 → 귀무가설 기각 → “목요일의 승객 수가 목요일이 아닌 날들보다 유의하게 많다”

가설 검증 그래프 2)



2019.01~06 중 언제 지하철을 가장 많이 이용했을까?

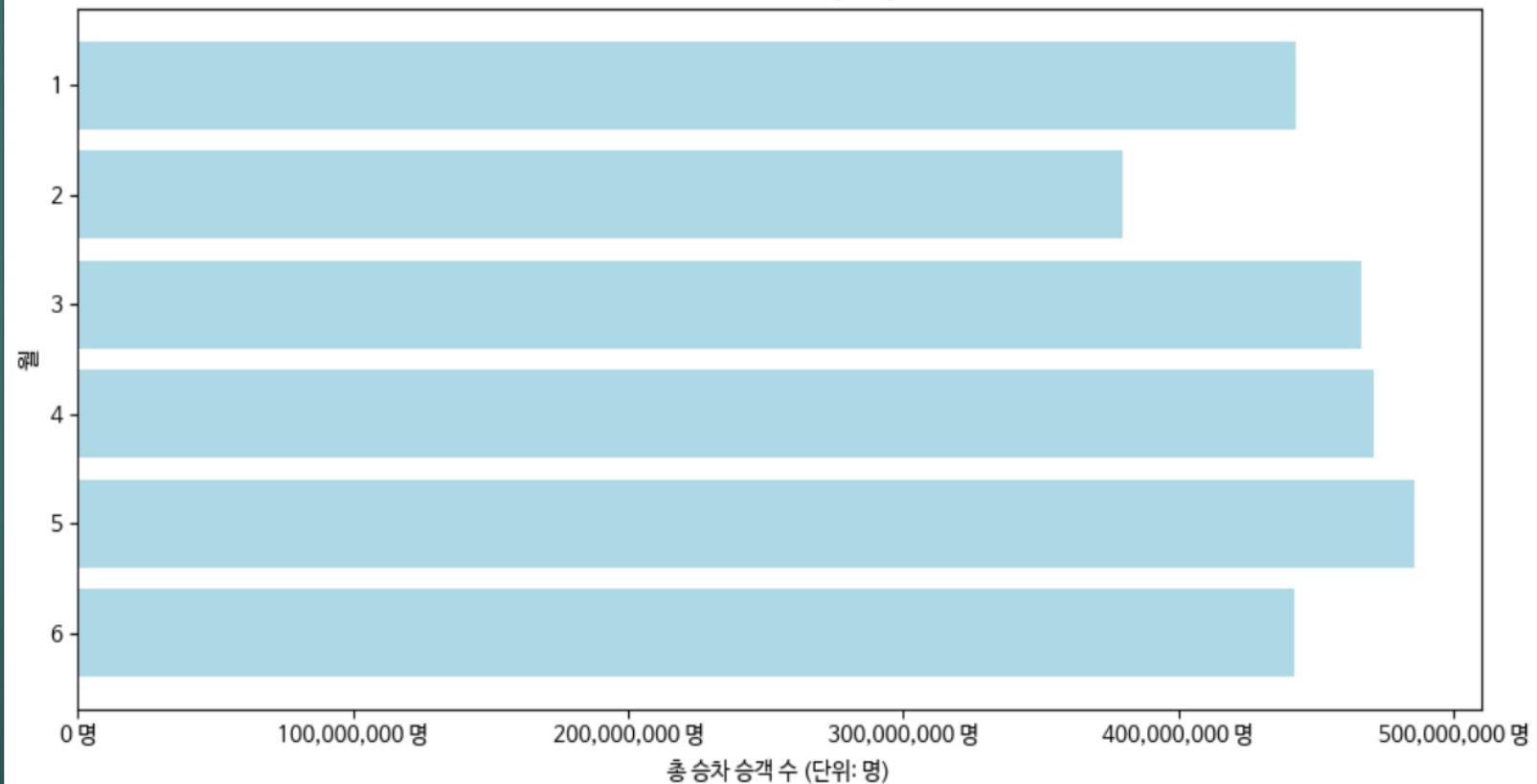
▶ 2019년 월별 지하철 총 이용객수

월

1.0	442746389.0
2.0	379836010.0
3.0	466692826.0
4.0	470934348.0
5.0	485718557.0
6.0	442210635.0

Name: 총_이용객수, dtype: float64

2019년 월별 지하철 총 이용객수

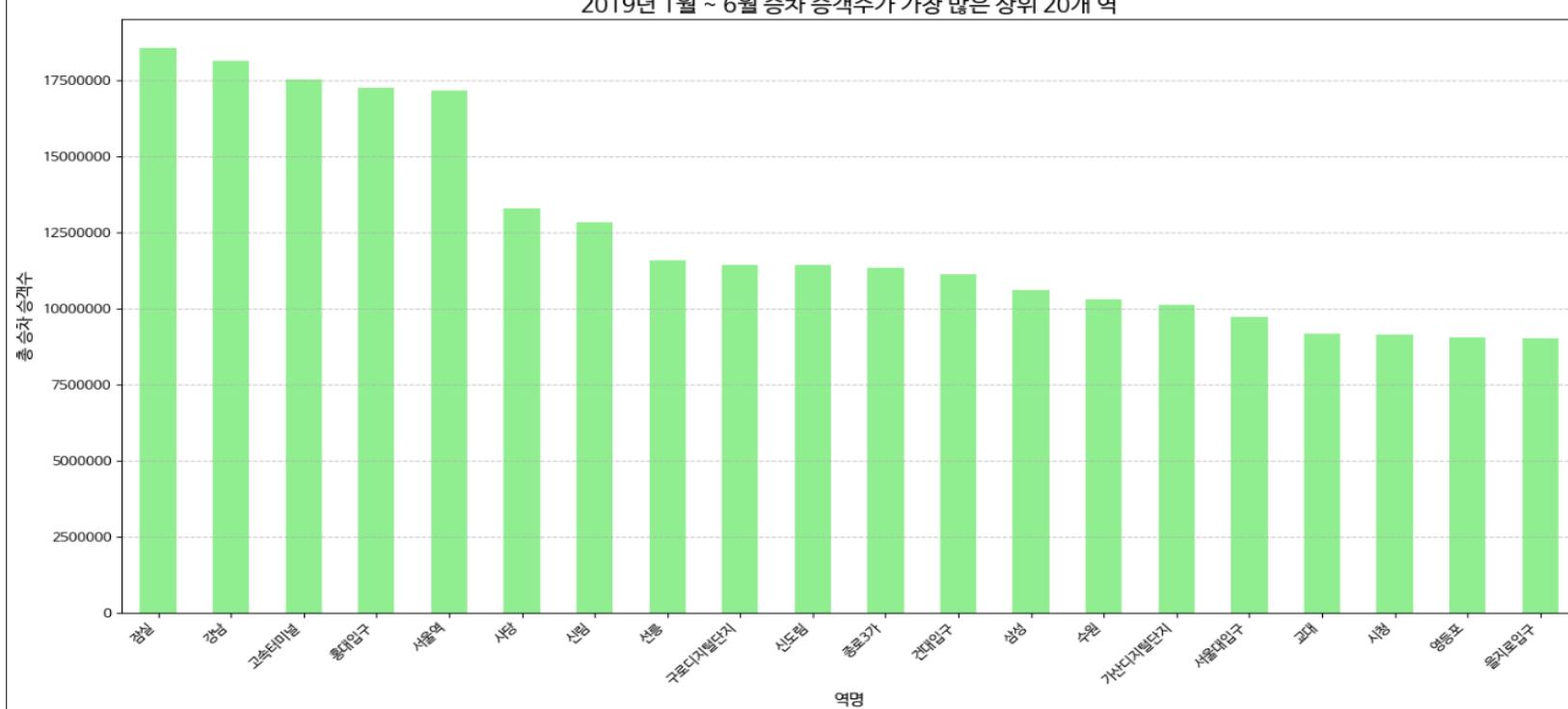




역명	승객수
잠실	18574323.0
강남	18148024.0
고속터미널	17541287.0
홍대입구	17270084.0
서울역	17165598.0
사당	13294251.0
신림	12831374.0
선릉	11582155.0
구로디지털단지	11421335.0
신도림	11420882.0
종로3가	11347625.0
건대입구	11123655.0
상봉	10611401.0
수원	10311002.0
가산디지털단지	10111317.0
서울대입구	9712111.0
교대	9156484.0
시청	9133667.0
영등포	9042814.0
을지로입구	9012670.0

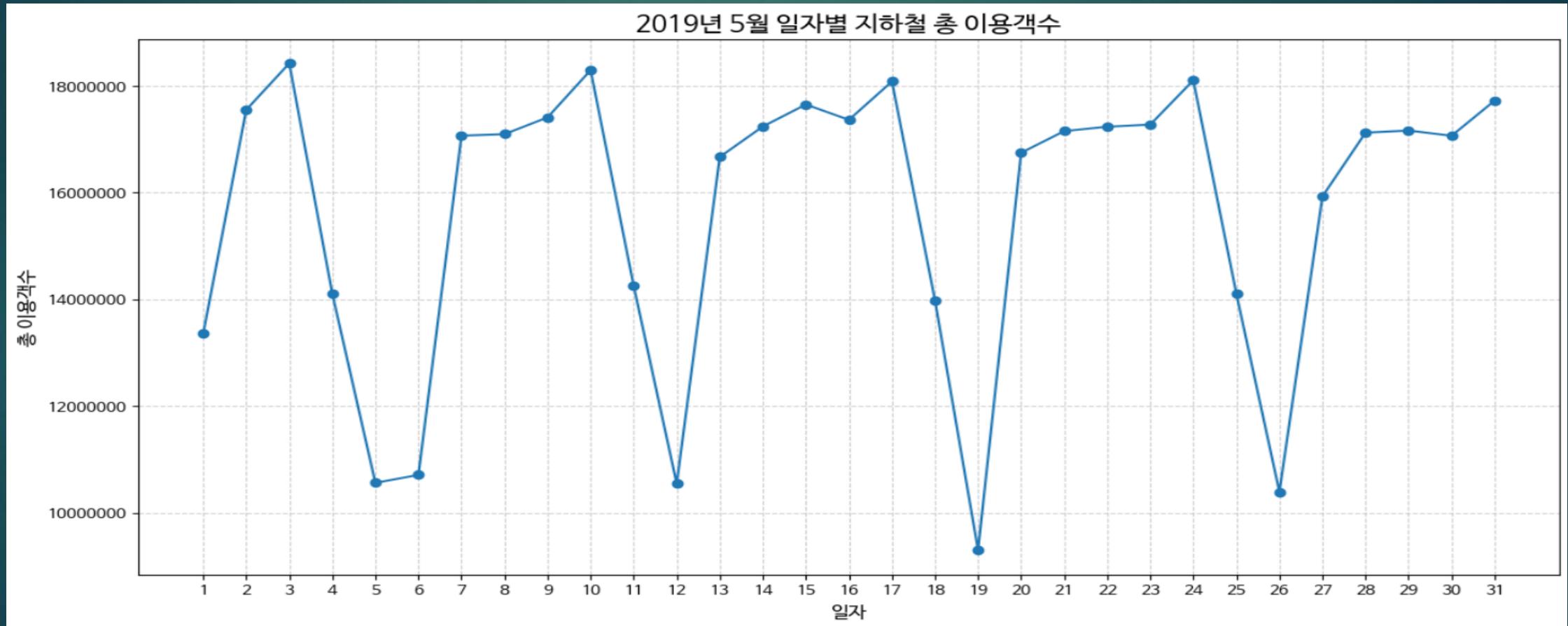
Name: 승차총승객수, dtype: float64

2019년 1월 ~ 6월 승차 승객수가 가장 많은 상위 20개 역

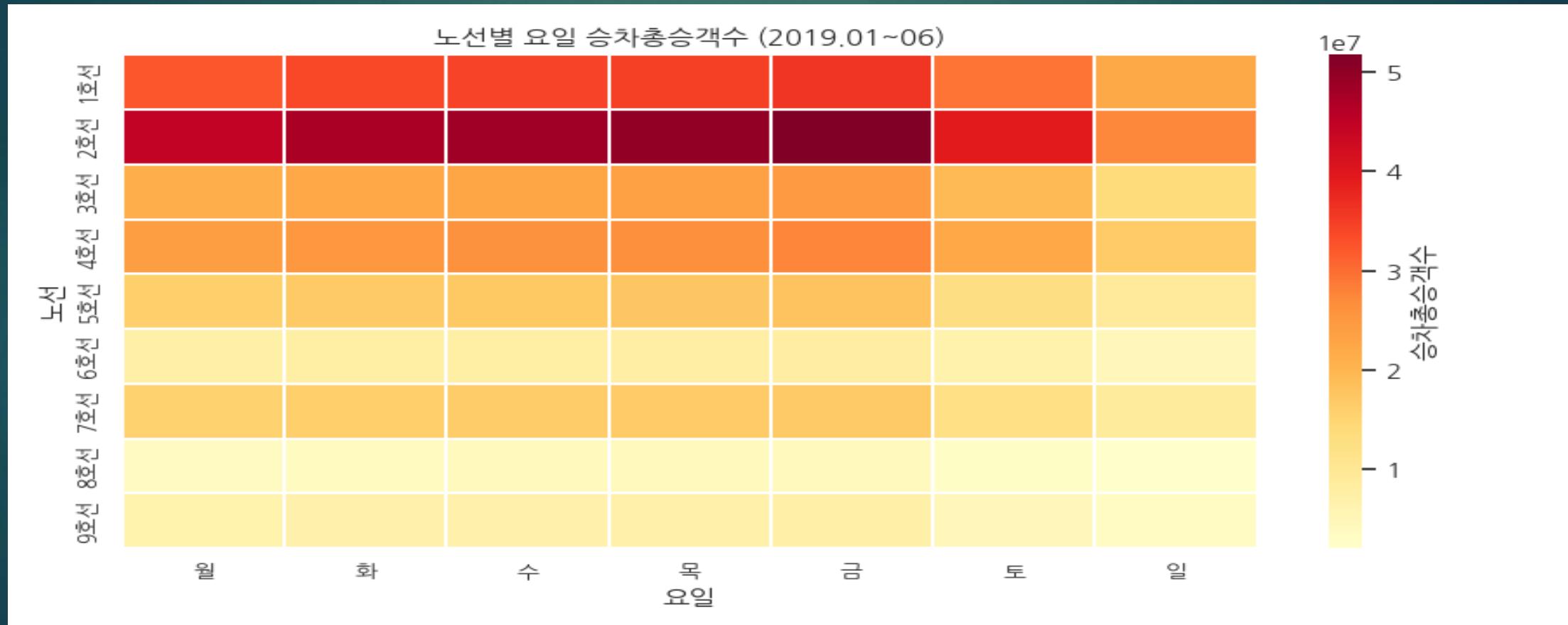


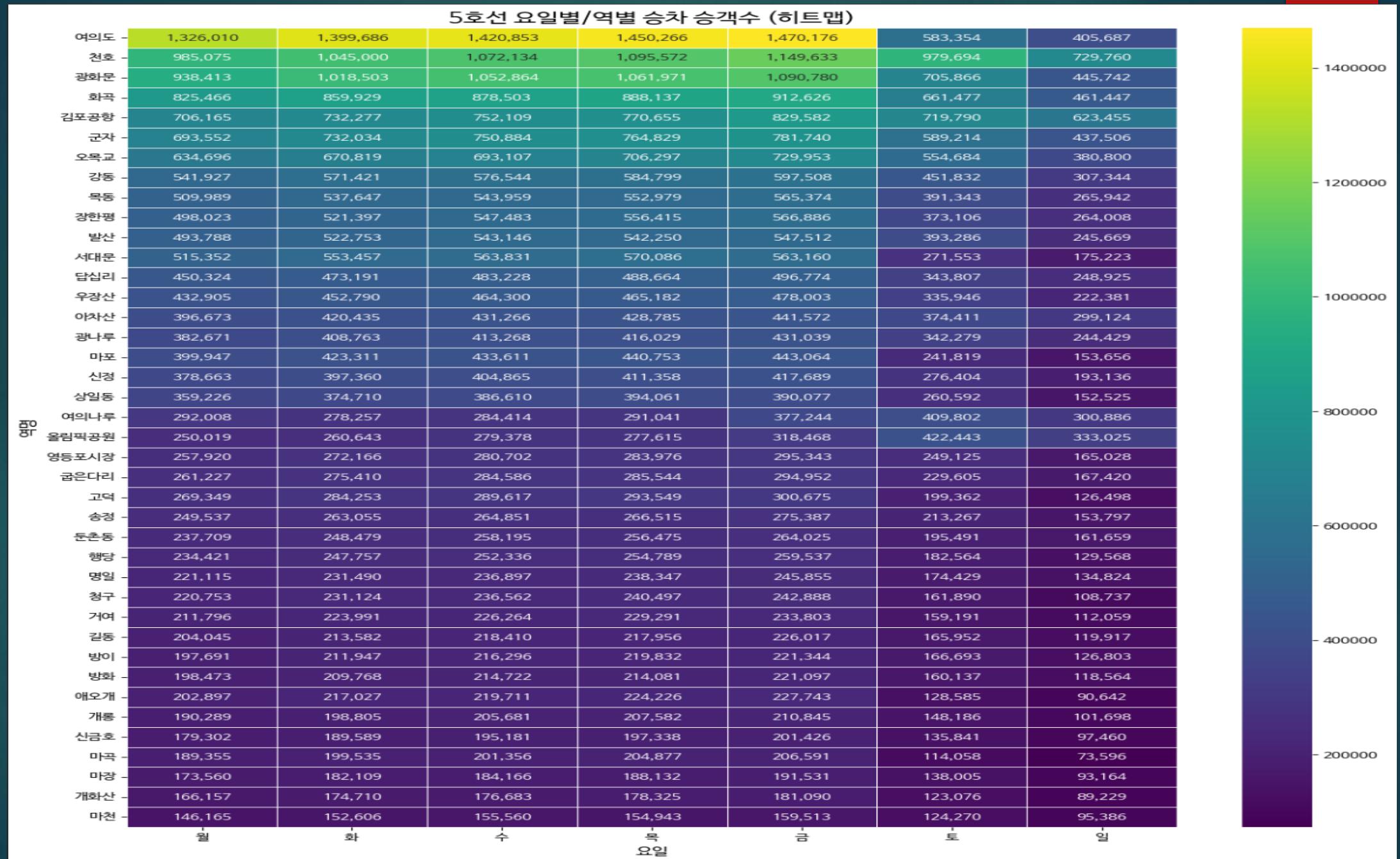
최다 승차 승객 분석 결과

연월 각각에 대해 일자별(월일별) 승하차 총 승객 수 그래프



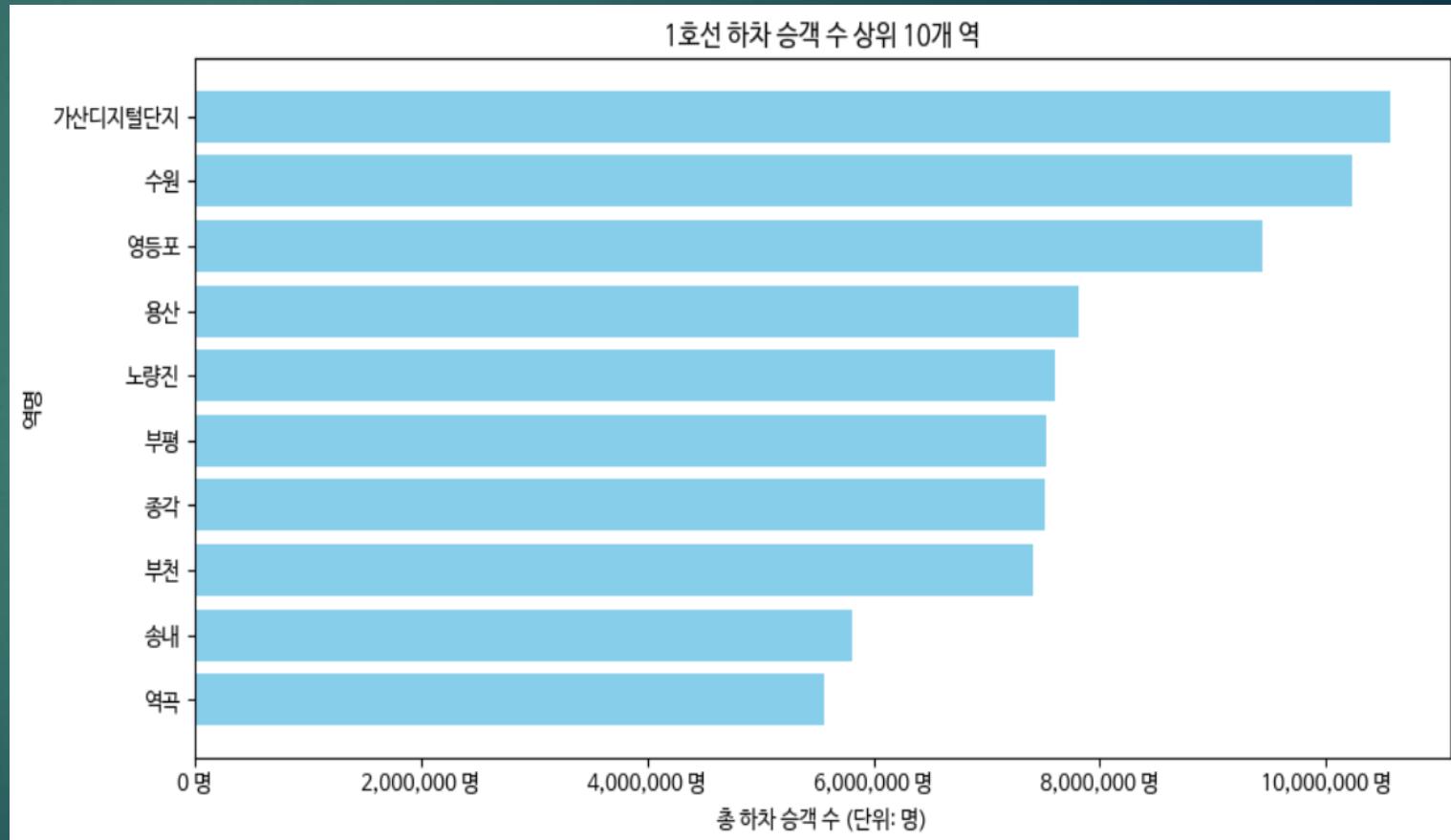
노선별로 역별/요일별 승차승객수 비교(히트맵)





1호선에서 가장 하차를 많이 하는 역은?(groupby)

1호선 하차 승객수가 가장 많은 상위 10개 역:
역명
가산디지털단지 10571381.0
수원 10226609.0
영등포 9432067.0
용산 7817685.0
노량진 7603258.0
부평 7531774.0
종각 7517515.0
부천 7412269.0
송내 5807791.0
역곡 5559994.0
Name: 하차총승객수, dtype: float64

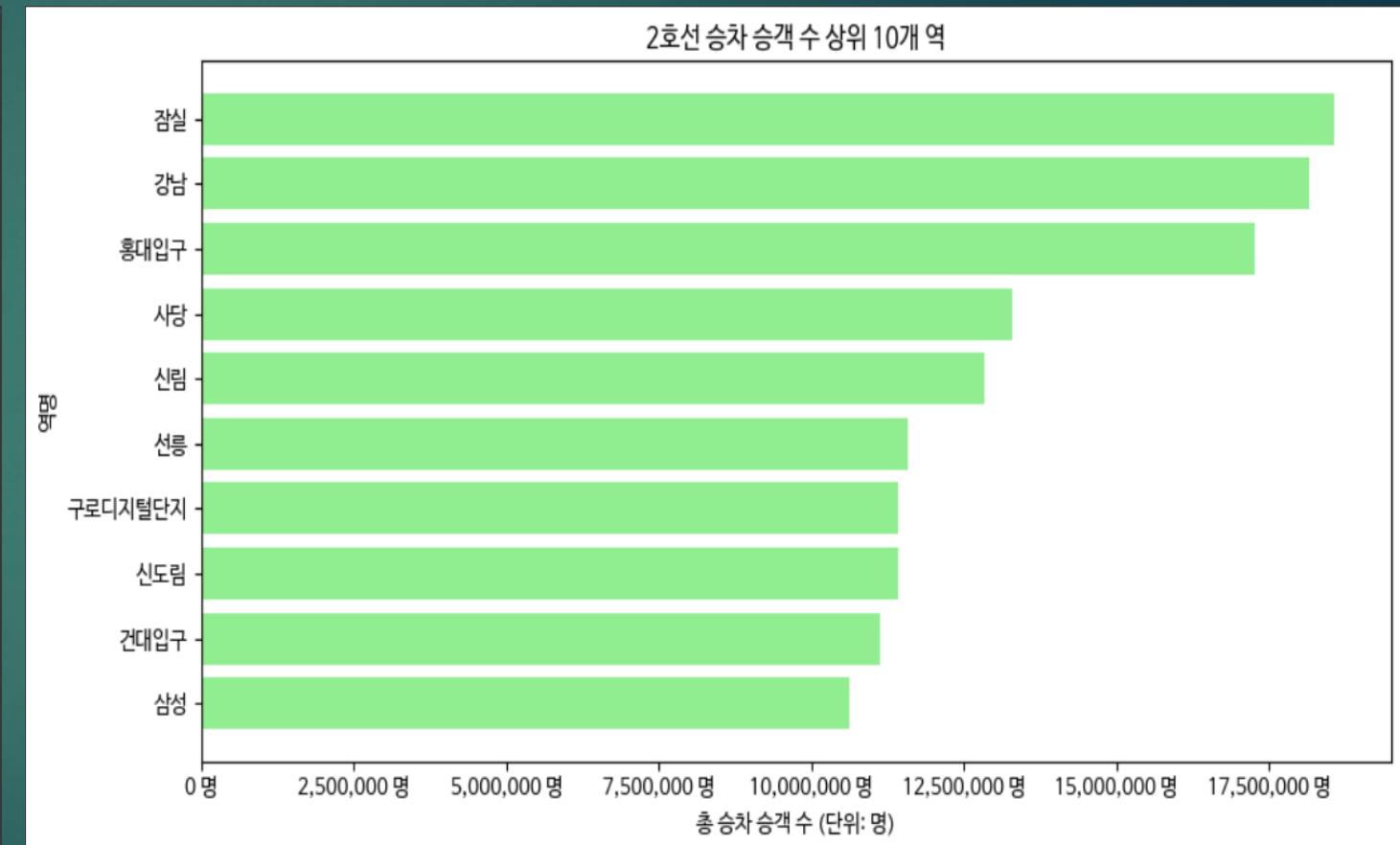


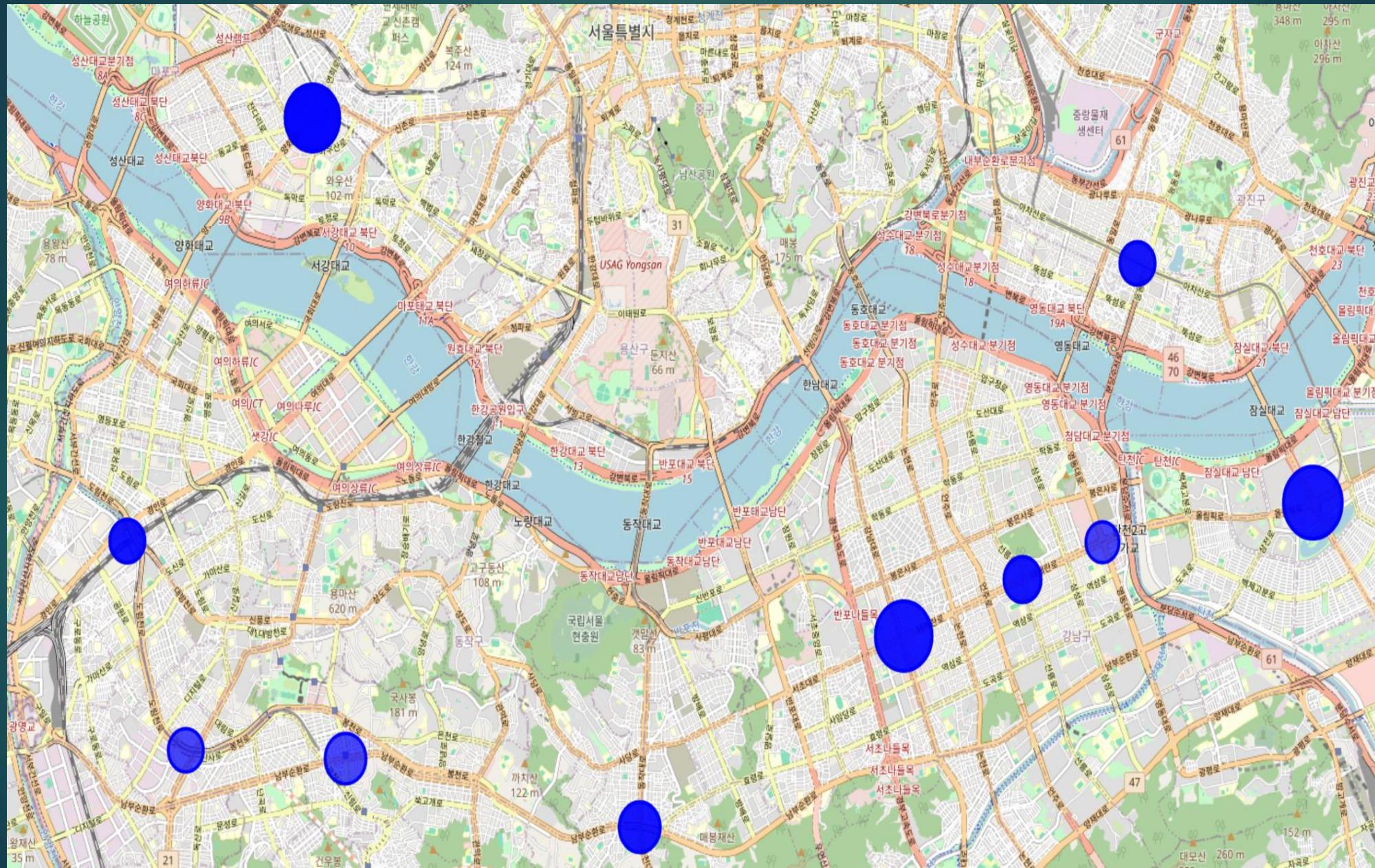
2호선 중에서 어느 역에서 승차가 가장 많아 발생하는가?

2호선 승차 승객수가 가장 많은 상위 10개 역 :

역명	승객수
잠실	18574323.0
강남	18148024.0
홍대입구	17270084.0
사당	13294251.0
신림	12831374.0
선릉	11582155.0
구로디지털단지	11421335.0
신도림	11420882.0
건대입구	11123655.0
삼성	10611401.0

Name: 승차총승객수, dtype: float64

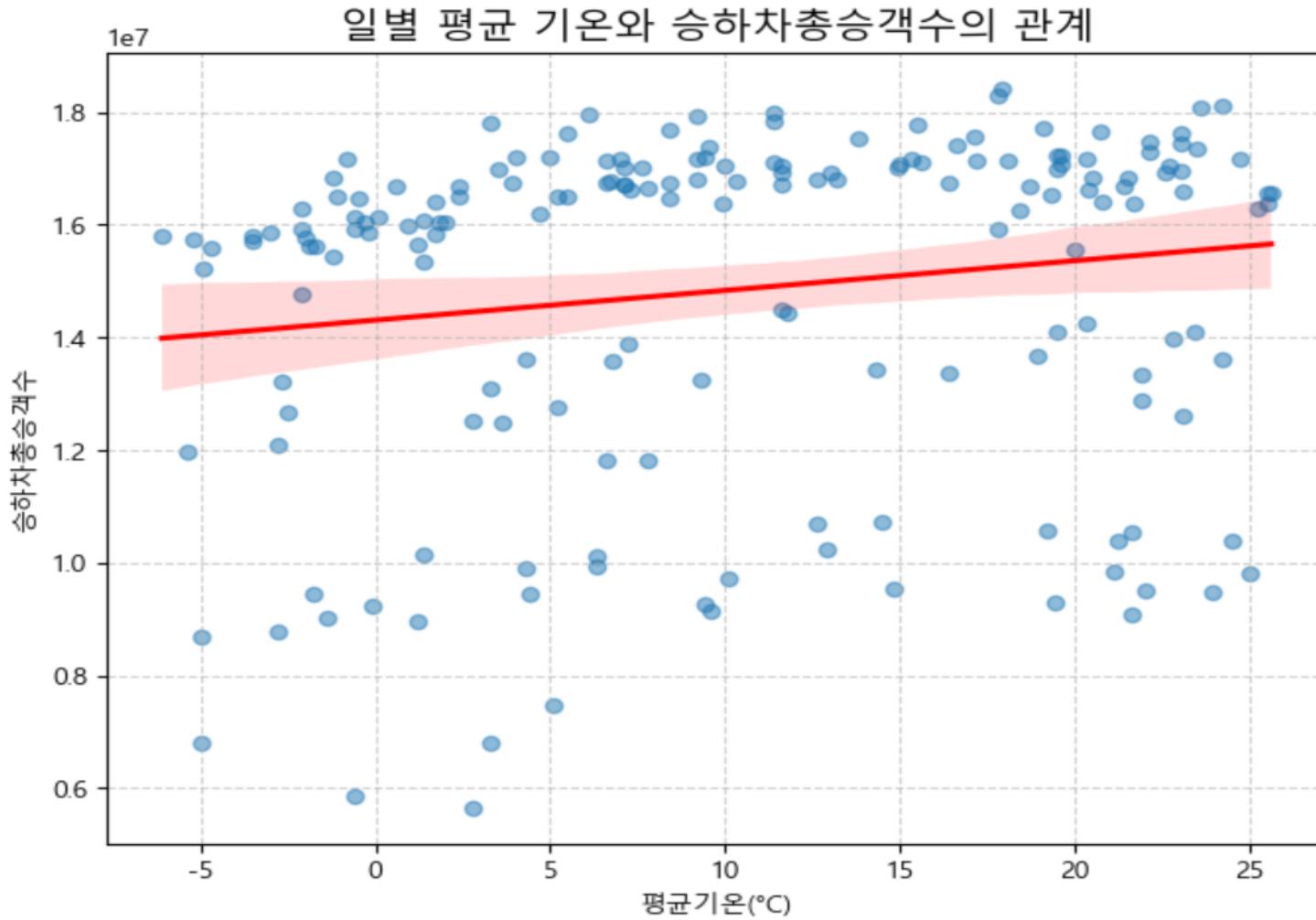




상관 분석

▶ '상관분석, 일별 평균 기온과 승하차총승객수의 관계'

--- 피어슨 상관 분석 결과 ---
- 피어슨 상관계수(r): 0.1590
- P-value: 0.0326



상관 분석
일별 평균 기온과
승하차총승객수의
관계

문제 외 아이디어 검정

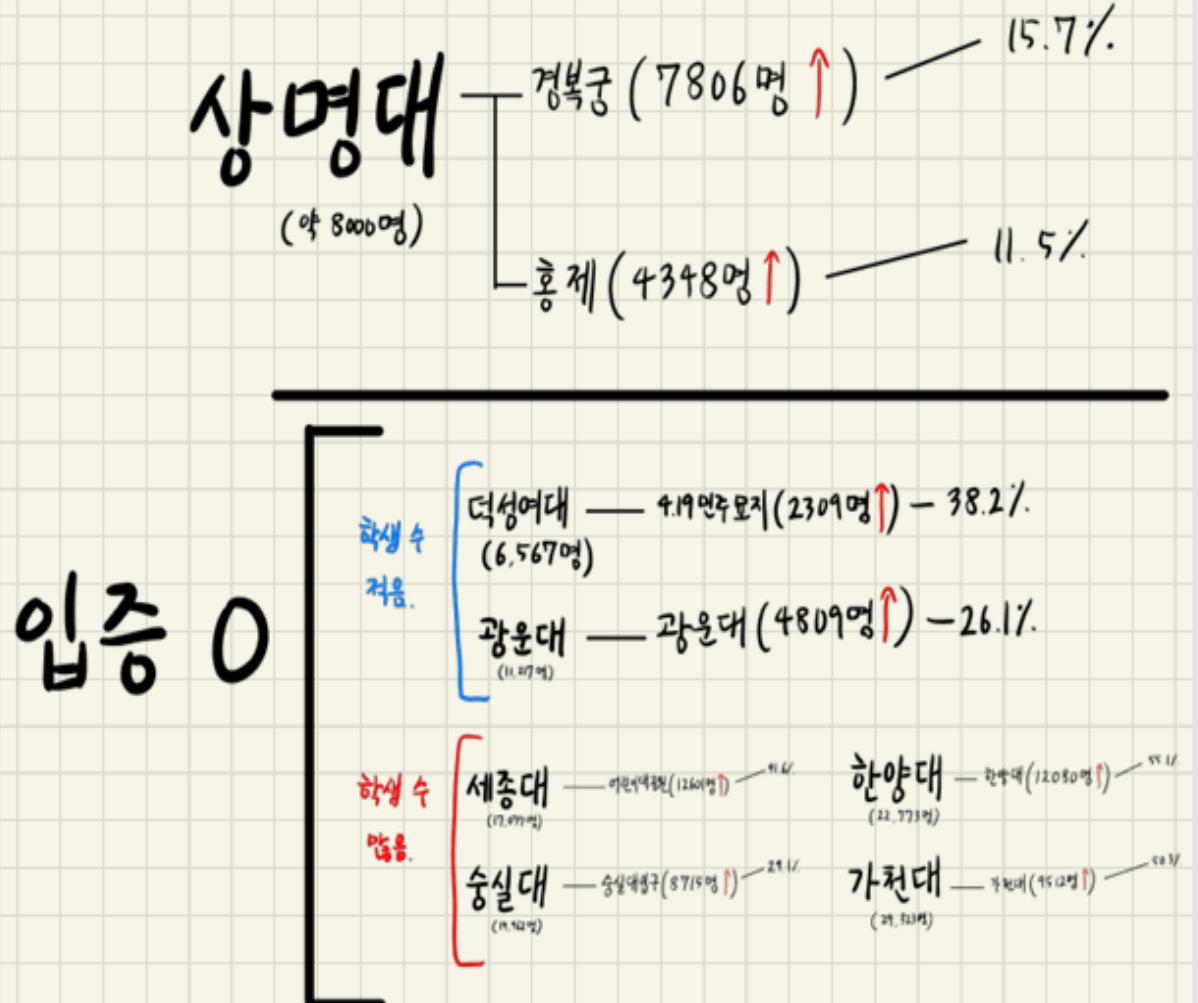
문제: 상명대학교 주변에 인접한 지하철 역이 없어서 학생들이 등교를 하는 방법이 매우 제한적이다

문제 원인을 ‘다른 서울권 대학보다 지하철역이 멀어서 지하철 이용률이 낮다고 설정.

문제 원인을 검증하기 위해 “상명대 학생들이 다른 서울 대학의 학생들에 비해 지하철 이용 빈도수가 적다”라는 가설을 설정함.

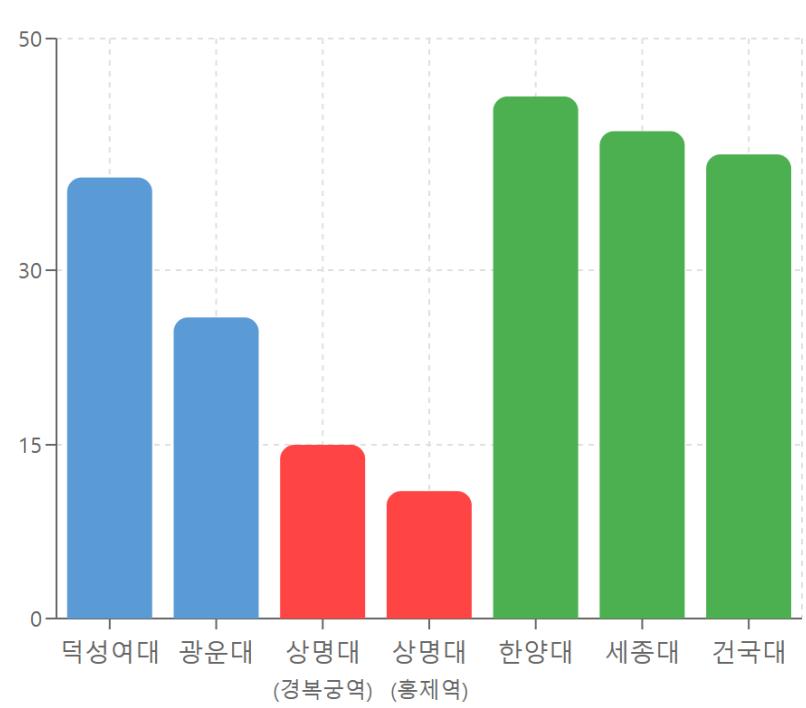
상명대와 다른 서울권 대학의 지하철 이용 빈도수를 비교함.

역 조사



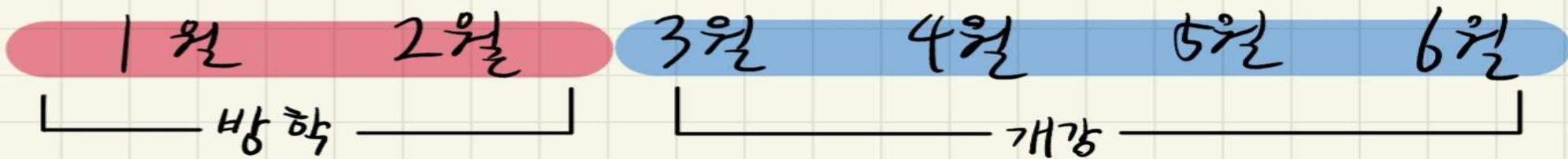
대학별 지하철역 접근성 분석

학생 수 대비 역 이용객 비율 비교



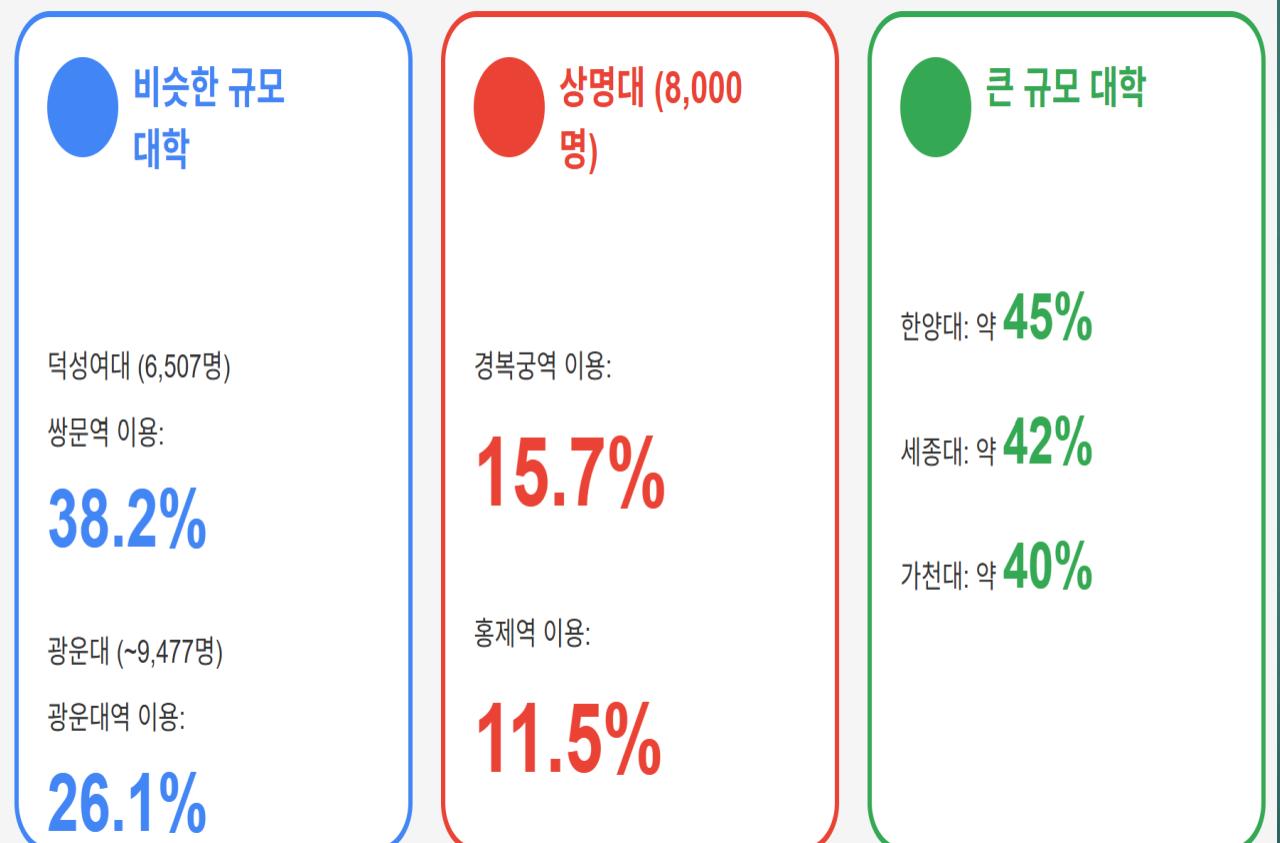
사용한 계산 방법

계산 방법



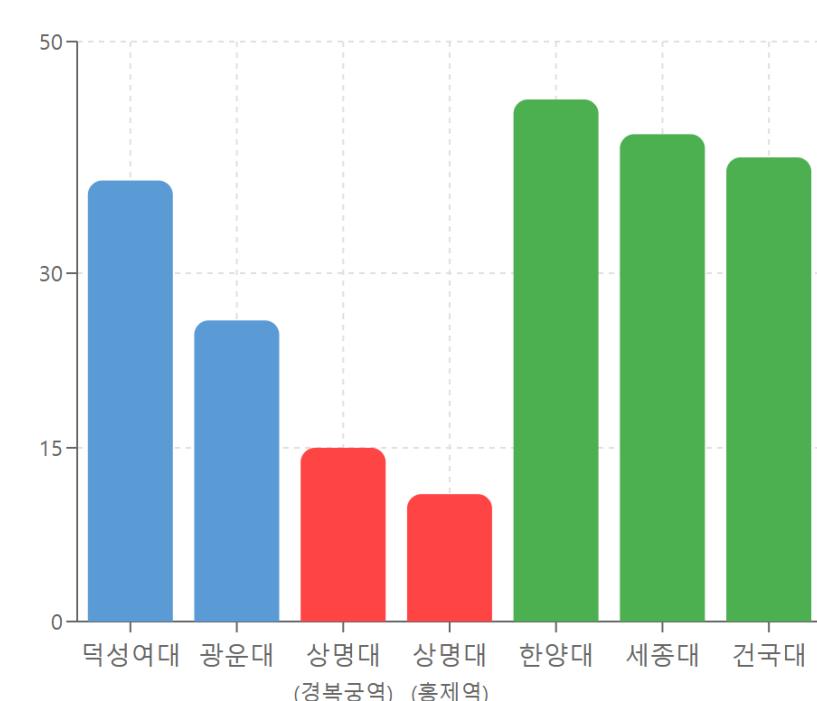
$\frac{(\text{"개강시기 승하차 수"} \text{ 평균}) - (\text{"방학시기 승하차 수"} \text{ 평균})}{(\text{"전체 승하차 수"} \text{ 평균})}$

분석 결과



대학별 지하철역 접근성 분석

학생 수 대비 역 이용객 비율 비교



결과 해석

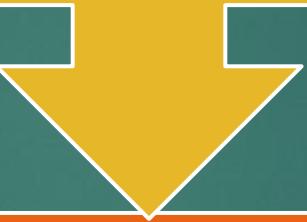
❖ 핵심 결론

- ✓ 비슷한 규모 대학들은 **26.1~38.2%**의 높은 역 이용률을 보임
 - ✓ 큰 규모 대학들은 전용 역으로 **40~45%**의 최고 접근성 확보
 - ✓ 상명대는 두 역을 합쳐도 **27.2%**로 여전히 낮은 수준
- ⇒ **상명대 인근 신규 역 개설의 필요성 명확**



결론

서울권 대학 중 지하철 역을 가지고 있는 대학의 지하철 이용률이 상명대에 비해 상대적으로 높다.



상명대 인근에 지하철역을 개설하면 지금보다
상대적으로 지하철을 더 많이 이용할 것이다.

지하철을 많이 이용하면, 등하교의 혼잡도가 줄어들
것으로 예상된다.

연구 기대효과, 향후계획

1) 상명대 학생들의 역 이용률을 시각화해서 분석하는 과정에서 타 대학 학생들과의 역 이용률 차이를 비교함.



2) 이를 통해 상명대 근처 역 개설의 필요성을 구체적으로 전달했음.



4) 버스 승객수 데이터, 지역별 인구 데이터, 지하철역 위치 데이터 등을 활용해 제안한 주장을 더 유의미하게 발전할 계획임.



3) 그러나 지하철역은 하나의 정보만 가지고 결정되는 것은 아니고 다양한 데이터가 필요함.



감사합니다

Thank you