AIVLE

kt

KT AIVLE School

서울시 생활정보데이터 기반 대중교통 수요 분석

AI 4반 12조

AIVLE

Let's make it possible

가설 수립

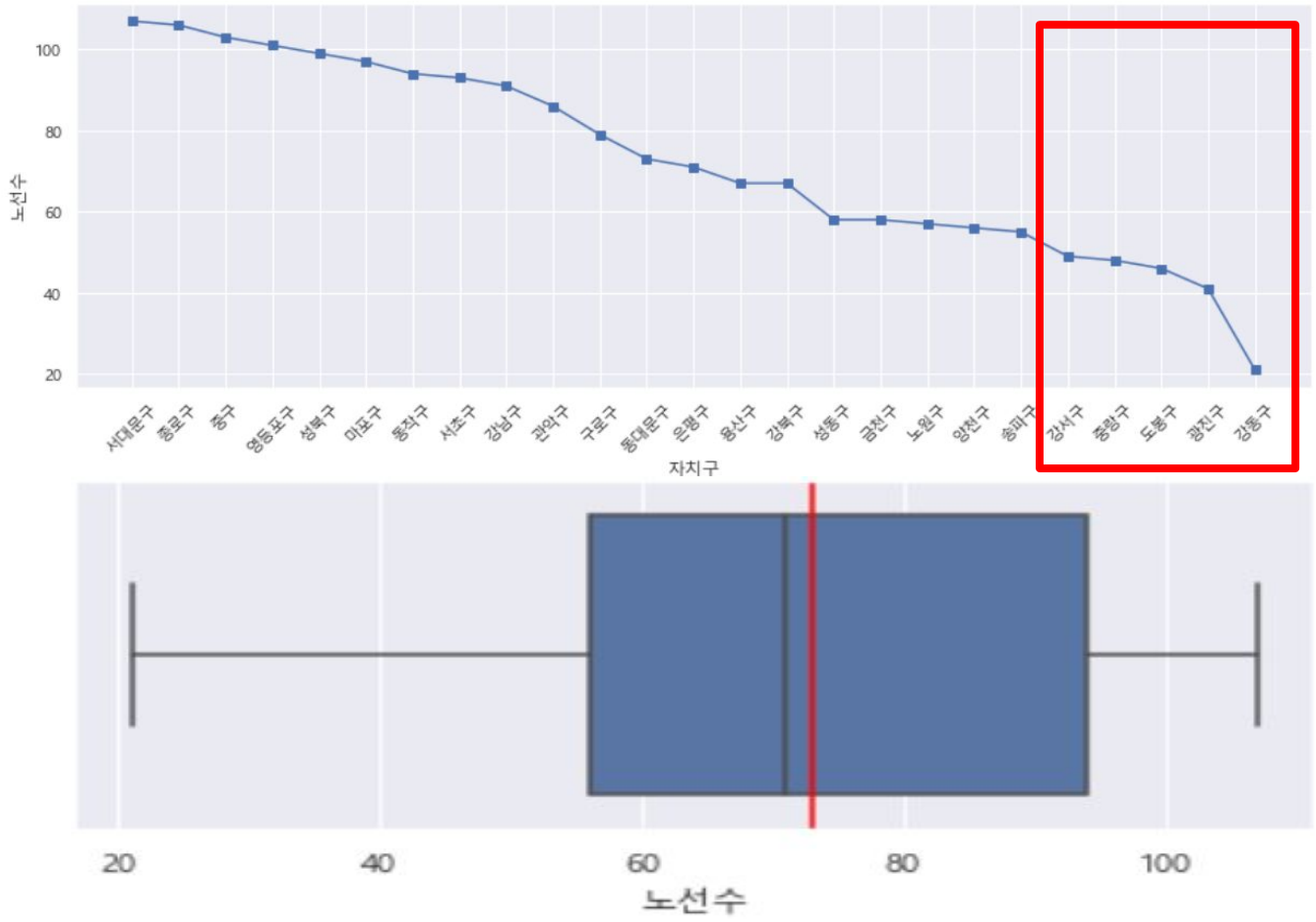
- **가설 1**
  - 노선수가 적은 자치구일수록 평균 이동시간이 높을 것이다.

- **가설 2**
  - 노선당이용자수가 많은 자치구가 평균이동시간이 높을 것이다.

- **가설 3**
  - 정류장수와 하차총승객수와 관련이 있다.

단변량 분석

✔ 노선수 분포



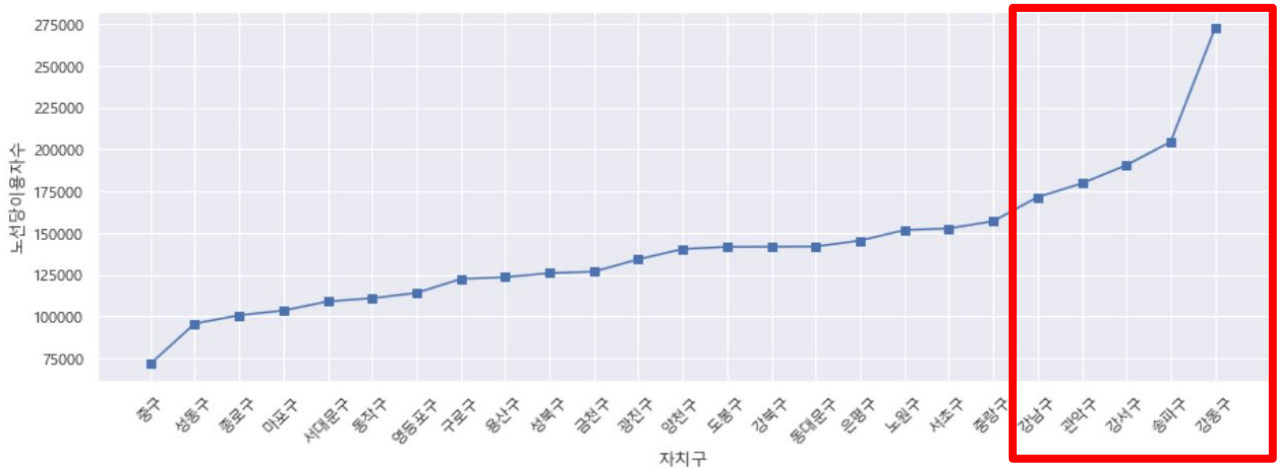
노선수 적은 자치구 TOP 5

1. **강동구**
2. 광진구
3. 도봉구
4. 중랑구
5. 강서구

- 서울의 50%의 자치구가 71개 이상의 노선을 가지고 있다.
- 평균값과 중앙값이 거의 일치하는 그래프이다.

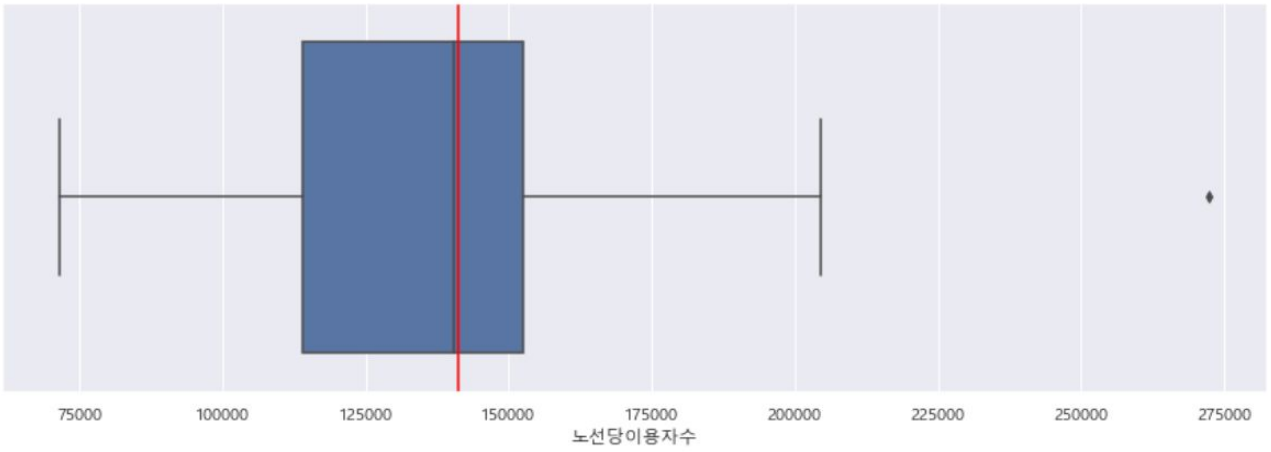
단변량 분석

✔ 노선당이용자수 분포 (노선당이용자수 : 4월간 한 노선당 이용자수)



노선당이용자수가 많은 자치구 TOP 5

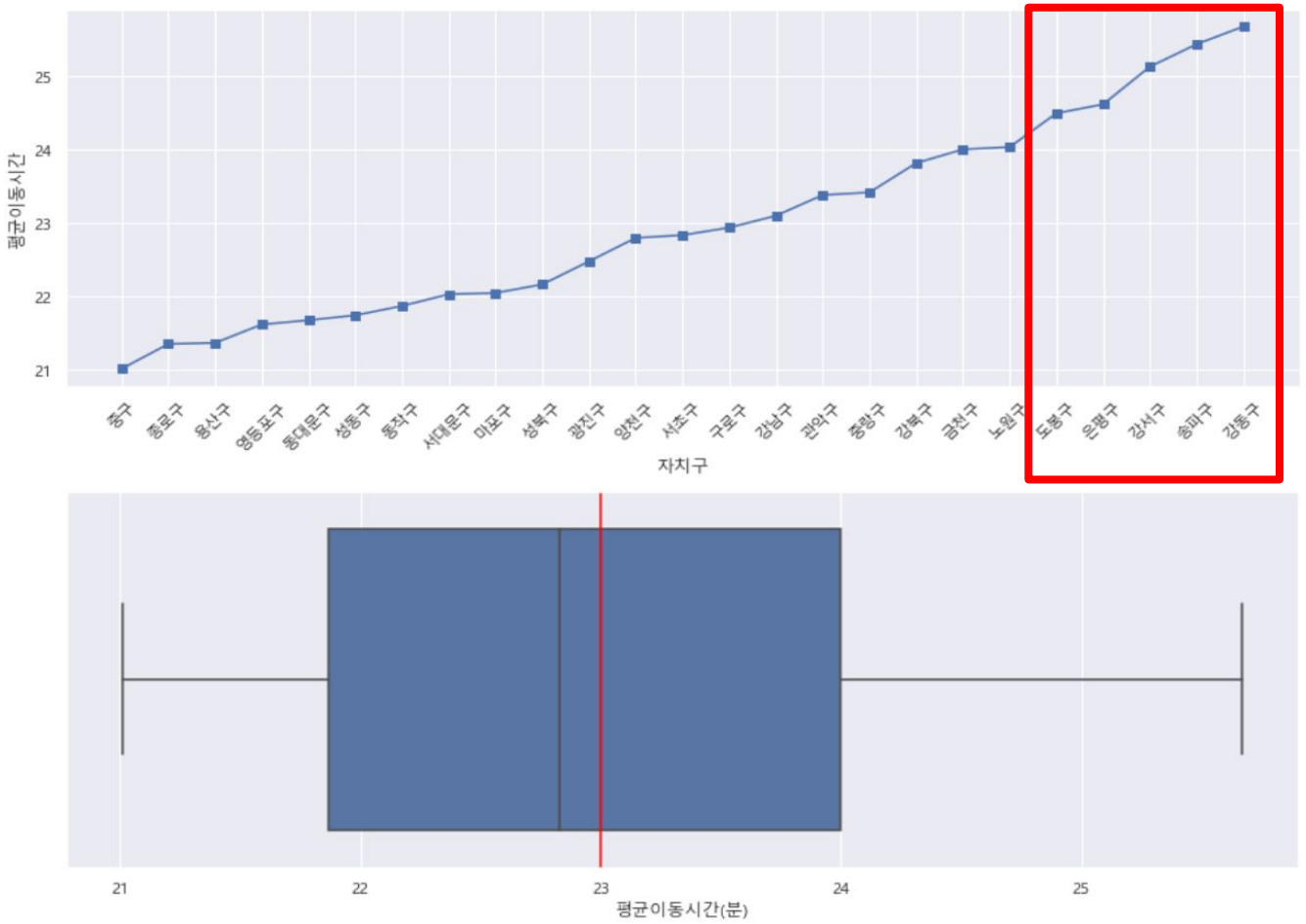
1. **강동구**
2. 송파구
3. 강서구
4. 관악구
5. 강남구



- 노선당이용자수의 50%가 약 140,363 이하이다.
- 평균값과 중앙값이 거의 일치하는 그래프이다.

단변량 분석

✔ 평균이동시간 분포



평균이동시간이 긴 자치구 TOP 5

1. **강동구**
2. 송파구
3. 강서구
4. 은평구
5. 도봉구

- 평균이동시간의 50%가 약 23분 이상이다.
- 평균값이 중앙값의 오른쪽에 위치하고, 오른쪽 꼬리가 긴 그래프이다.

단변량 분석

✔ 하차총승객 분포

하차총승객이 많은 정류장의 자치구 TOP 5

1. **관악구★**
2. 강남구
3. 서초구
4. 성북구
5. 서대문구

출처:https://news.nate.com/view/20210204n65792

서울 관악구, 지역상권 살리기 사업 팔 걷어 붙였다

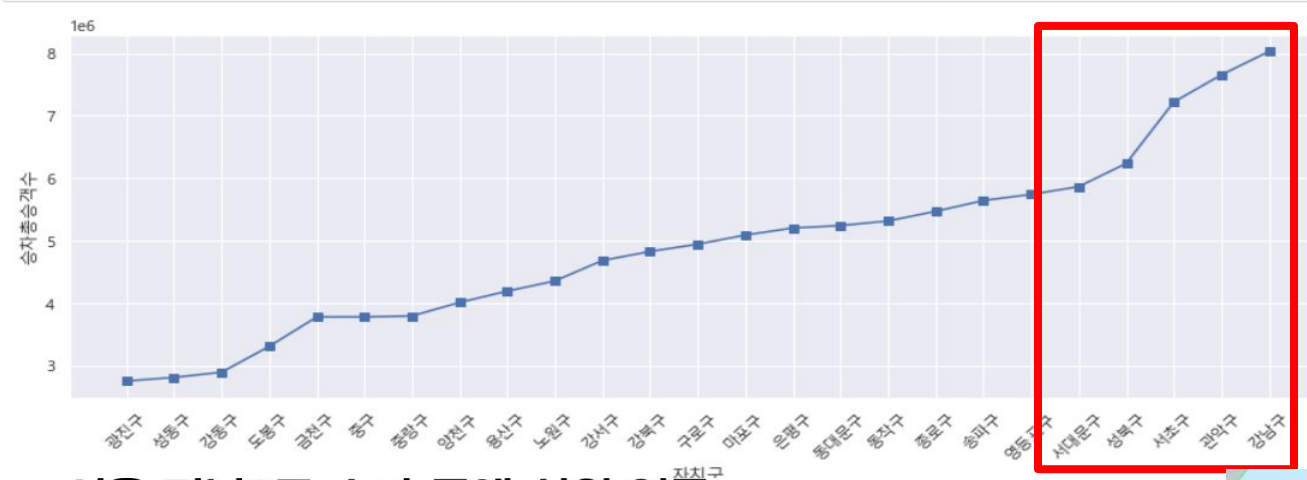
서울 관악구가 코로나19로 침체된 지역 골목상권에 활력을 불어넣기 위해 총력을 기울이고 있다.

2021년도 기사내용

출처:위키백과

단변량 분석

✔ 승차총승객 분포



승차총승객이 많은 정류장의 자치구 TOP 5

1. **강남구 ★**
2. 관악구
3. 서초구
4. 성북구
5. 서대문구

서울 강남3구 소비 금액 상위 업종

(단위 : 억 원, %)

유통	5193(25.1)
의료	3683(17.8)
요식 · 유흥	3136(15.1)
가정생활 · 서비스	2316(11.2)
교육 · 학원	1541(7.4)
자동차	1253(6.1)
스포츠 · 문화 · 레저	812(3.9)
의류 · 잡화	738(3.6)

※신한카드 2021년 12월 사용 실적(전자상거래 업종 제외)
자료 : 서울시 빅데이터 캠퍼스

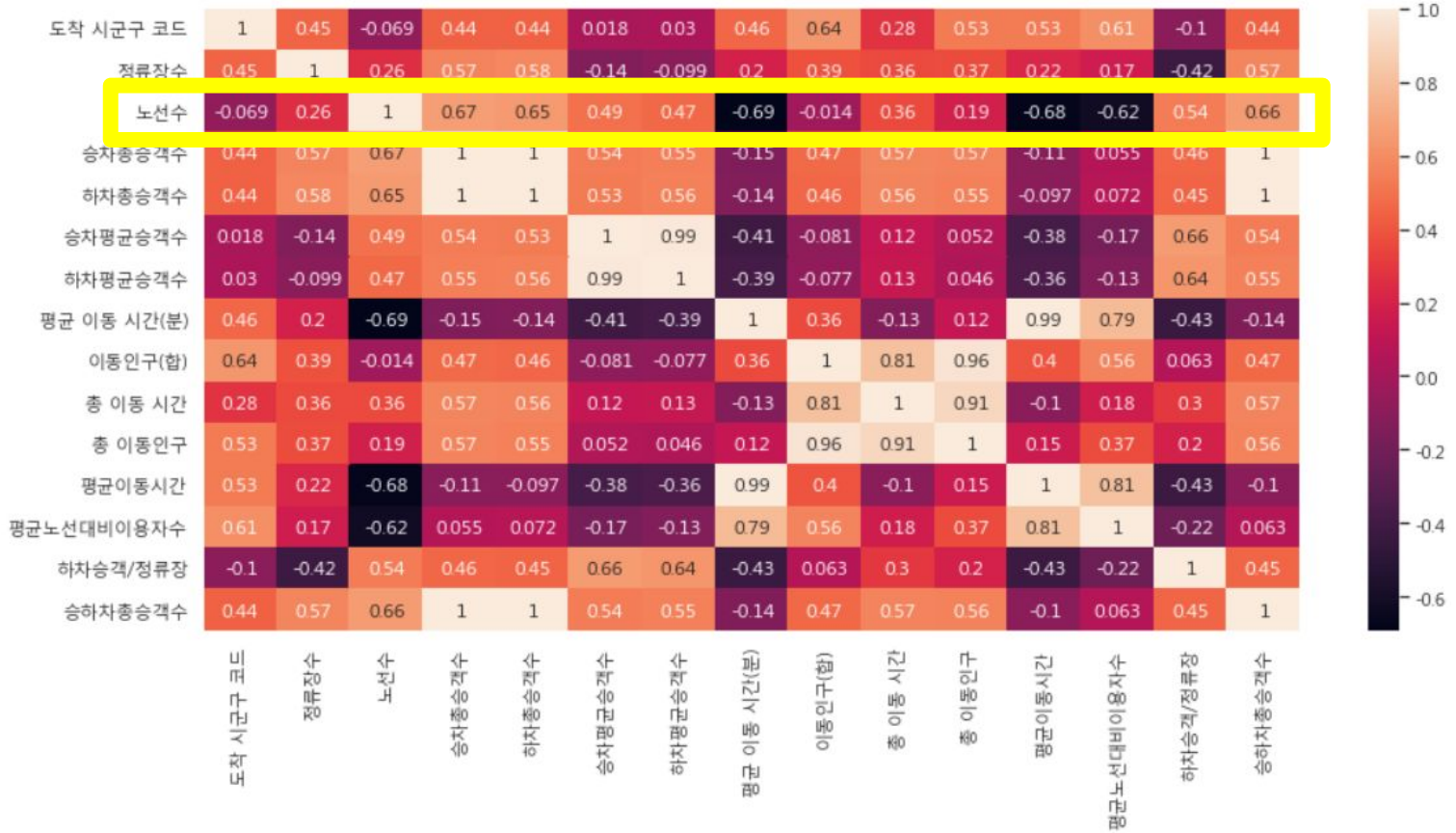
서울시민, 강남3구서 車 구입비 53% 썼다

국내 최대 소비 중심지인 강남 3구(강남·서초·송파구)의 자동차 및 관련 용품·서비스 소비 금액이 서울 전체의 절반 이상인 53%로 나타났다.

출처:위키백과

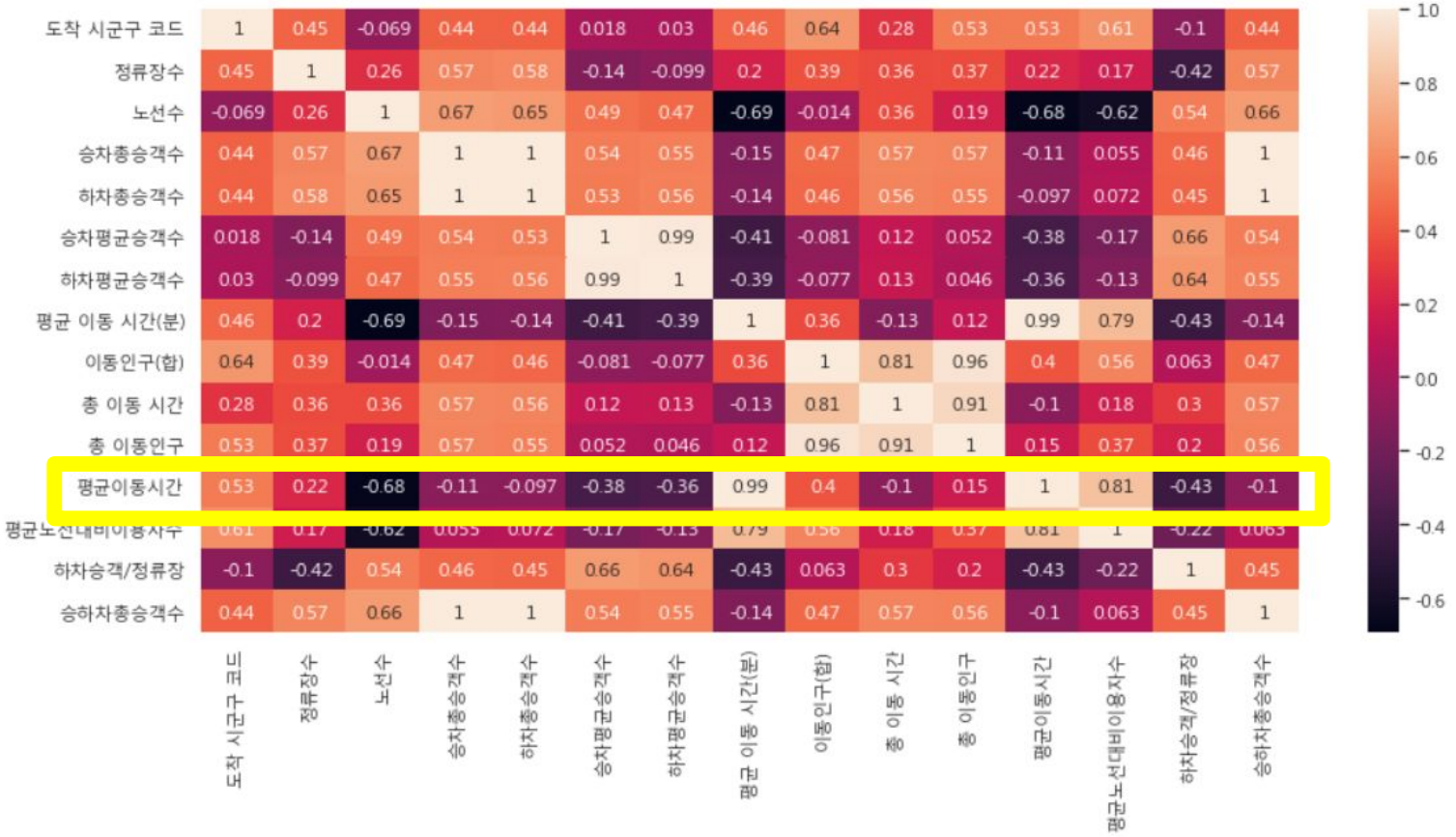
출처:https://m.sedaily.com/NewsView/26288Q7JAS#cb

이변량 분석 - 관계 정리 (Y = 노선수)



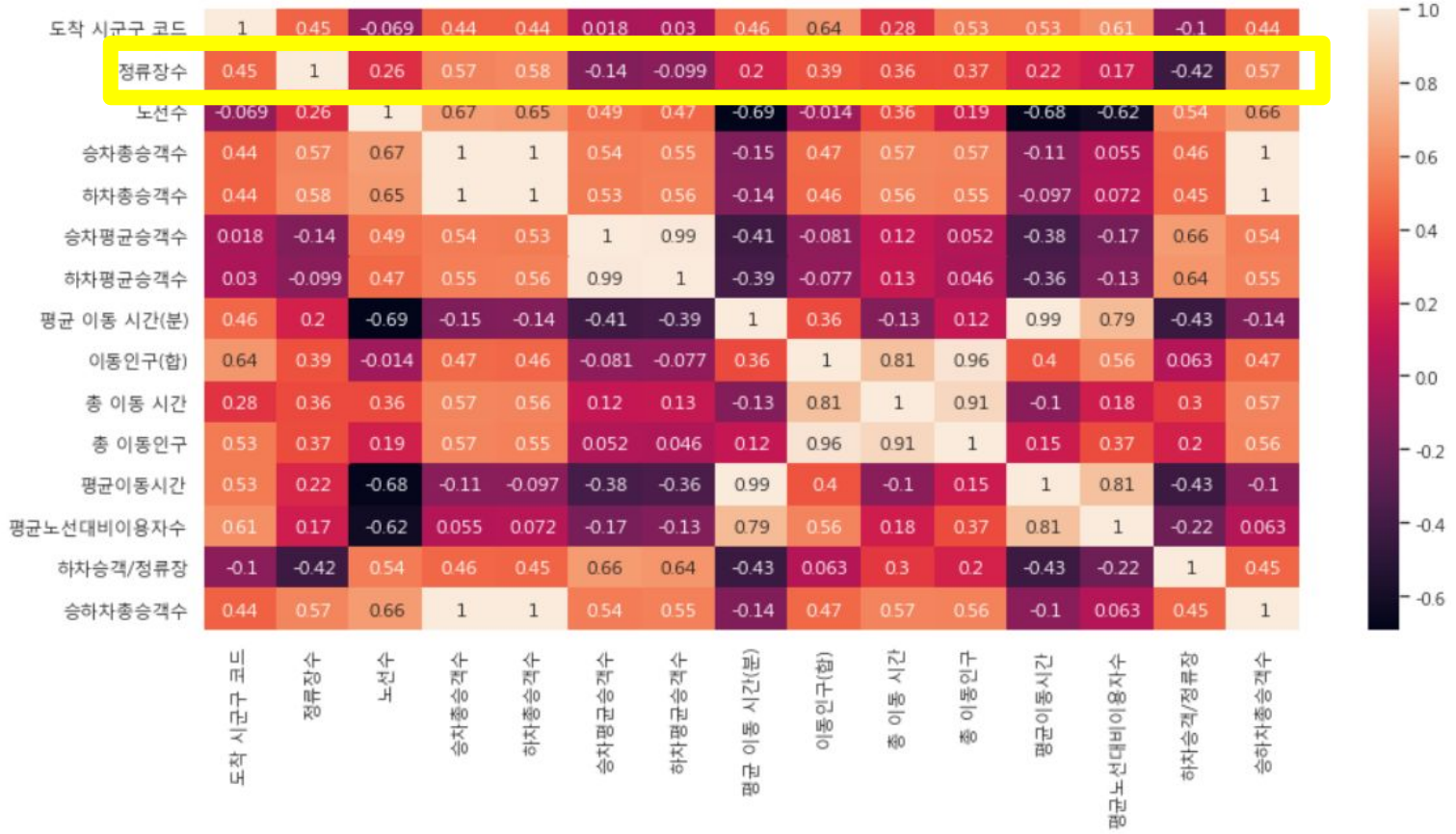
- **강한** 관계의 x : 승차총승객수, 하차총승객수, 평균이동시간, 평균노선대비이용자수, 승하차총승객수
- **약한** 관계의 x : 정류장수, 이동 인구(합), 총 이동 인구

이변량 분석 - 관계 정리 (Y = 평균 이동 시간)



- **강한** 관계의 x : 노선수, 평균노선대비이용자수
- **약한** 관계의 x : 승차총승객수, 하차총승객수, 총 이동 시간, 총 이동인구, 승하차총승객수

이변량 분석 - 관계 정리 (Y = 정류장수)



- **강한** 관계의 x : 승차총승객수, 하차총승객수, 승하차총승객수
- **약한** 관계의 x : 노선수, 승차평균승객수, 하차평균승객수, 평균 이동 시간, 평균노선대비이용자수

이변량 분석 - 노선수 평균이동시간 분석

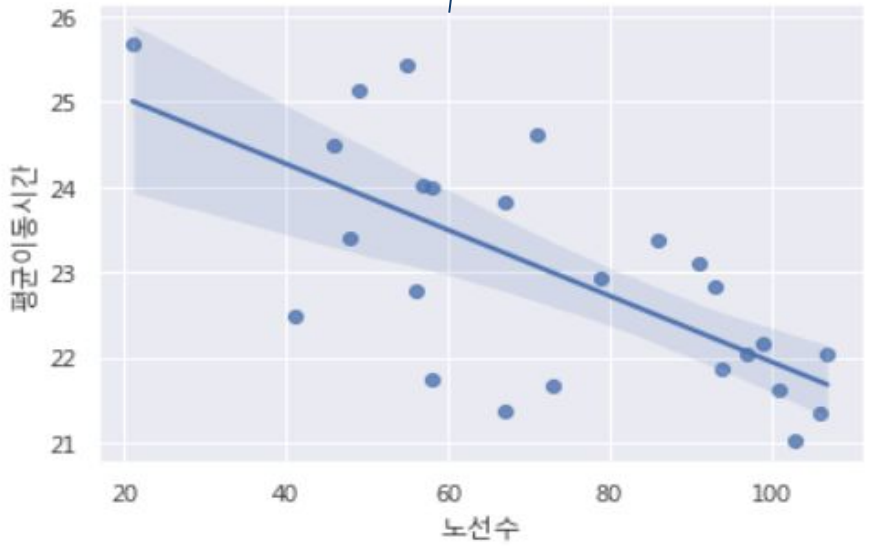
▪ 상관 분석(상관 계수 및 p-value)

(-0.6786000547455109, 0.00019237294667152178)

강한 음의 상관관계

p-value = 0.0192%

▪ 상관 분석(시각화)



노선수와 평균이동시간과의 관계는 67%이상의 강한 음의 상관관계를 가짐

⇒ 노선수가 많을수록 평균이동시간이 감소

이변량 분석 - 노선당이용자수 평균이동시간 분석

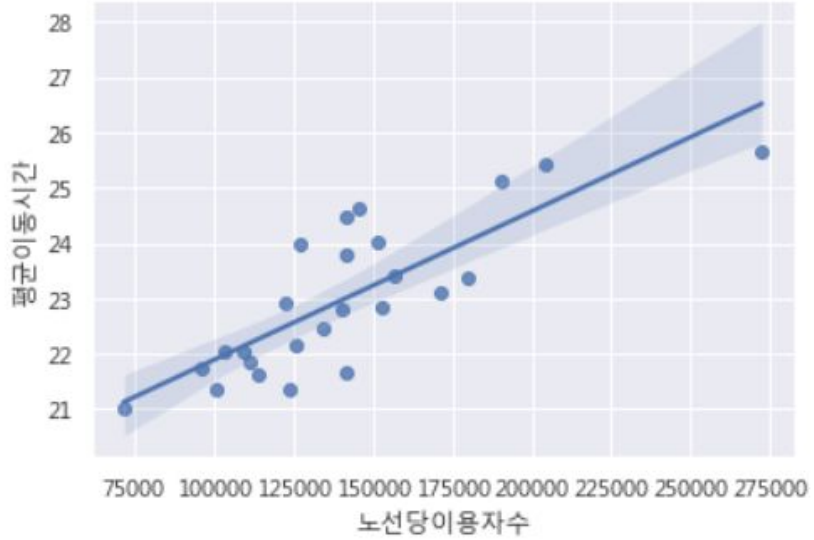
▪ 상관 분석(상관 계수 및 p-value)

(0.8104768922168128, 9.050536906896215e-07)

강한 양의 상관관계

p-value = 0.00009%

▪ 상관 분석(시각화)



노선당이용자수와 평균이동시간과의 관계는 81% 이상의 매우 강한 상관관계를 보임

⇒ 노선당이용자수가 높을수록 평균 이동시간이 증가

이변량 분석 - 정류장수 하차총승객수 분석

- 상관 분석(상관 계수 및 p-value)

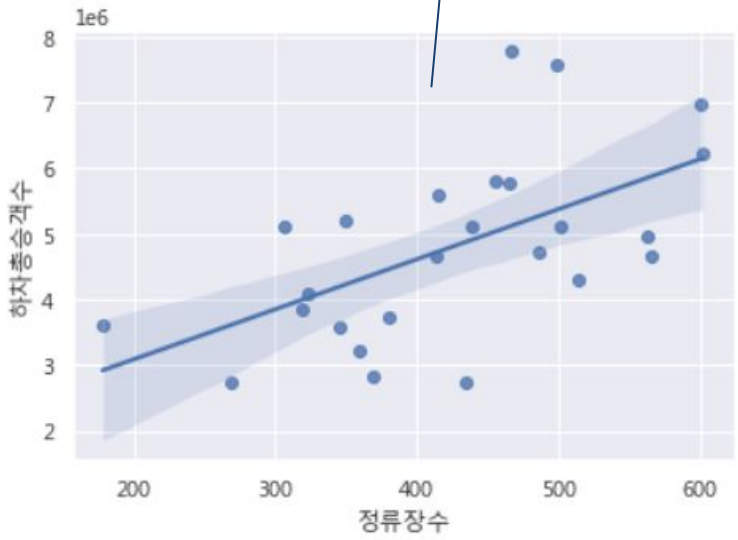
(0.5784790149467576, 0.0024518332034538645)

양의 상관관계

p-value = 0.245%

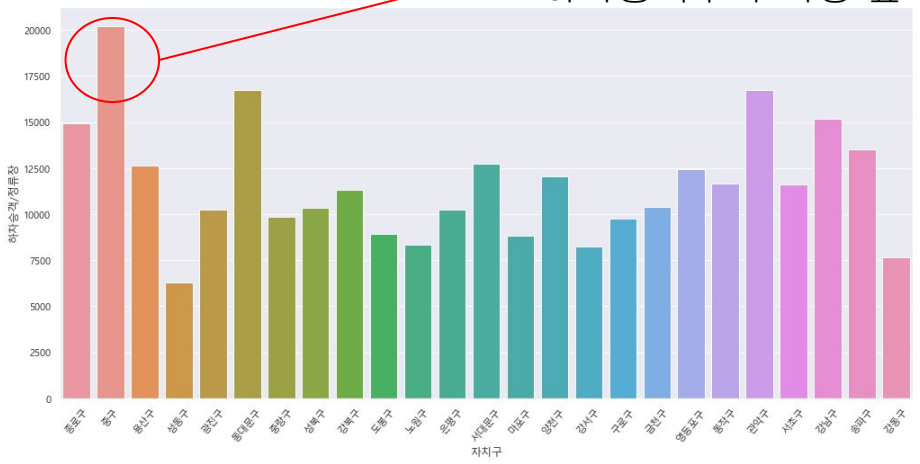
중구: 20219(최대)
성동구: 6314(최소)
단위 (하차승객수 / 정류장수)

- 상관 분석(시각화)



- 정류장 대비 하차승객수

중구에서 정류장 대비
하차승객수가 가장 높다



가설 검증 과정

변수 및 가설 설정 배경

팀원들의 각각의 가설을 종합했을 때 아래와 같이 분석한 케이스가 많았다.

- **정류장 수 - 이동인구수(총승차고객수, 총하차고객수)**
- **노선수 - 평균이동 시간**
- **노선수 - 승하차승객수**

이를 바탕으로 노선수와 정류장 수 중 어떤 변수가 더 상관관계가 있는지 확인한 결과
정류장수 ⇔ 이동인구수 대비 노선수 ⇔ 이동인구수 와 상관계수가 더 큰 것을 발견하여
노선수를 변수로 채택하였다.
또한 실제로 모든 정류장에 버스가 정차하는 것은 아니기 때문에 노선수가 상대적으로 실제
운행하는 버스 수와 좀 더 밀접한 관련이 있다고 추론하였다.

가설 검증 과정

노선수가 적은 구를 찾기위해 단변량 분석을 시행하였다.

강동구가 가장 적은 노선을 운행하고 있으며 그 뒤로 광진구, 도봉구가 뒤를 이었다.

평균 이동시간이 가장 많은 구를 찾기 위해 단변량 분석을 시행하였다.

강동구가 가장 많은 이동시간을 필요로 하였고 송파구, 강서구 등이 그 뒤를 이어 시간소요가 많았다.

하차총승객이 가장 많은 구를 찾기 위해 단변량 분석을 시행하였다.

관악구가 하차총승객이 가장 많았고 강남구, 서초구 등이 그 뒤를 이어 하차총승객이 많았다.

승차총승객이 가장 많은 구를 찾기 위해 단변량 분석을 시행하였다.

강남구가 승차총승객이 가장 많았고 관악구, 서초구 등이 그 뒤를 이어 승차총승객이 많았다.

단변량 분석 과정에서 도출한 내용

앞에서 가장 많이 언급된 변수들을 가지고 단변량 분석을 실시했을 때 노선수가 가장 적은 구,

평균이동시간이 가장 높은 구는 강동구로, 승하차인원 분석시에 관악구, 강남구, 서초구가 의미있는 구로

집계되었다.

승하차 인원에 대한 부분은 이변량 분석에서 **노선당이용자수**((총승차승객수+총하차승객수)/노선수) 이라는

새로운 변수에 포함하여 추가로 검증해보기로 하였다.

가설 검증 과정

가설 -1 **노선수가 적은 자치구일수록 평균 이동시간이 높을 것이다.**

이변량 분석을 통해 노선수와 평균 이동시간사이에

(-0.6786000547455109, 0.00019237294667152178)

강한 음의 상관계수를 가졌으며 이를통해 유의적인 영향을 미치는 것으로 나타났다.

가설 -2 **노선당이용자수 평균이동시간 분석**

이변량 분석을 통해 노선당이용자수((총승차승객수+총하차승객수)/노선수)과 평균 이동시간사이에
(0.8104768922168128, 9.050536906896215e-07)

강한 양의 상관계수를 가졌으며 이를통해 유의적인 영향을 미치는 것으로 나타났다.

가설 -3 **정류장수 하차총승객수 분석**

이변량 분석을 통해 정류장수와 하차총승객수사이에
(0.5784790149467576, 0.0024518332034538645)

양의 상관계수를 가졌으며 이를통해 유의적인 영향을 미치는 것으로 나타났다.

세개의 가설 모두 유의미한 값을 가졌으며, 상대적으로 가설3이 약한 상관관계를 가지는 것으로 확인되었다.

가설2
노선당이용자수
|
평균시간
(≒0.81)

>

가설1
노선수
|
평균이동시간
(≒-0.67)

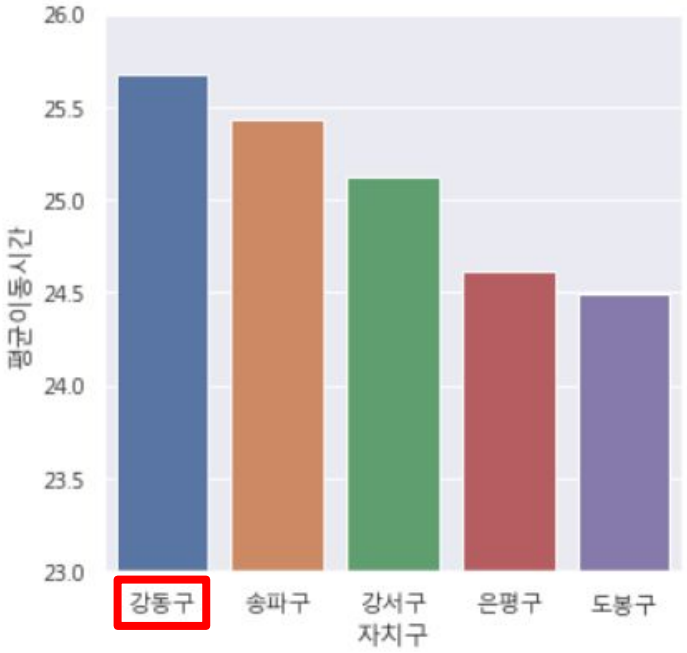
>

가설3
정류장수
|
하차총승객수
(≒0.57)

결 론

상관분석 결과에 따라 최종적으로 가설 1과 가설 2를 결론에 반영했다.

가설 1에 따르면 **노선 수**가 적을수록 **평균 이동시간**이 늘어난다.
가설 2에 따르면 **노선 당 이용자 수**가 많을수록 **평균 이동 시간**이 늘어난다.



결 론

노선 당 이용자 수가 가장 많고, 평균 이동시간이 가장 높은 구는 **강동구**이다.

따라서 강동구에 버스 노선을 추가하는 것이 필요하다.

AIVLE

AIVLE

Let’s make it possible