AIVLE

kt

KT AIVLE School

1차 미니프로젝트 - 조별 발표

수도권 AI 2반 6조

AIVLE

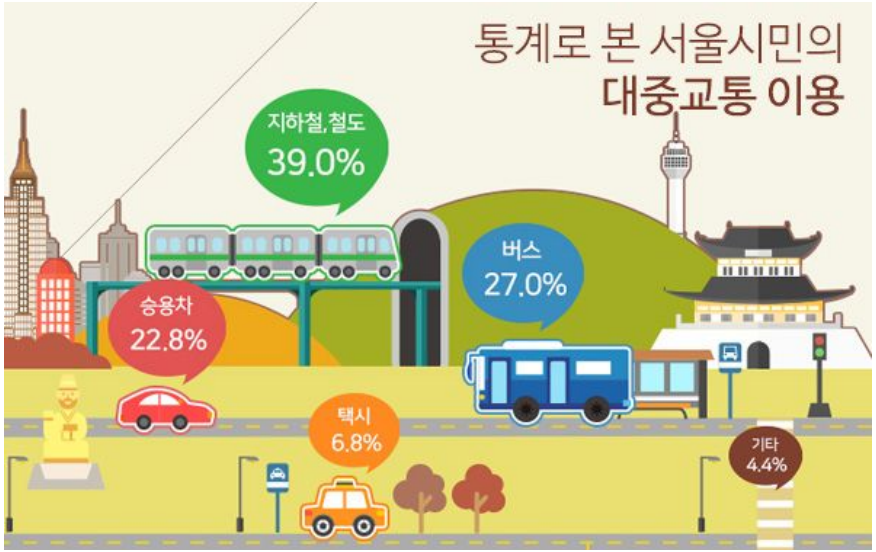
Let's make it possible

목차

분석 목표 설정

서울시민 중 **27.0%**가 버스를 사용해 이동

-> 서울 시민의 삶의 질 향상을 위해 버스 노선 최적화 필요



가설 수립

✔ **가설 1** : 승차총승객수와 하차총승객수는 노선수와 관련이 있을 것이다.

✔ **가설 2** : 승하차총합수와 정류장 수는 관련이 있을 것이다.

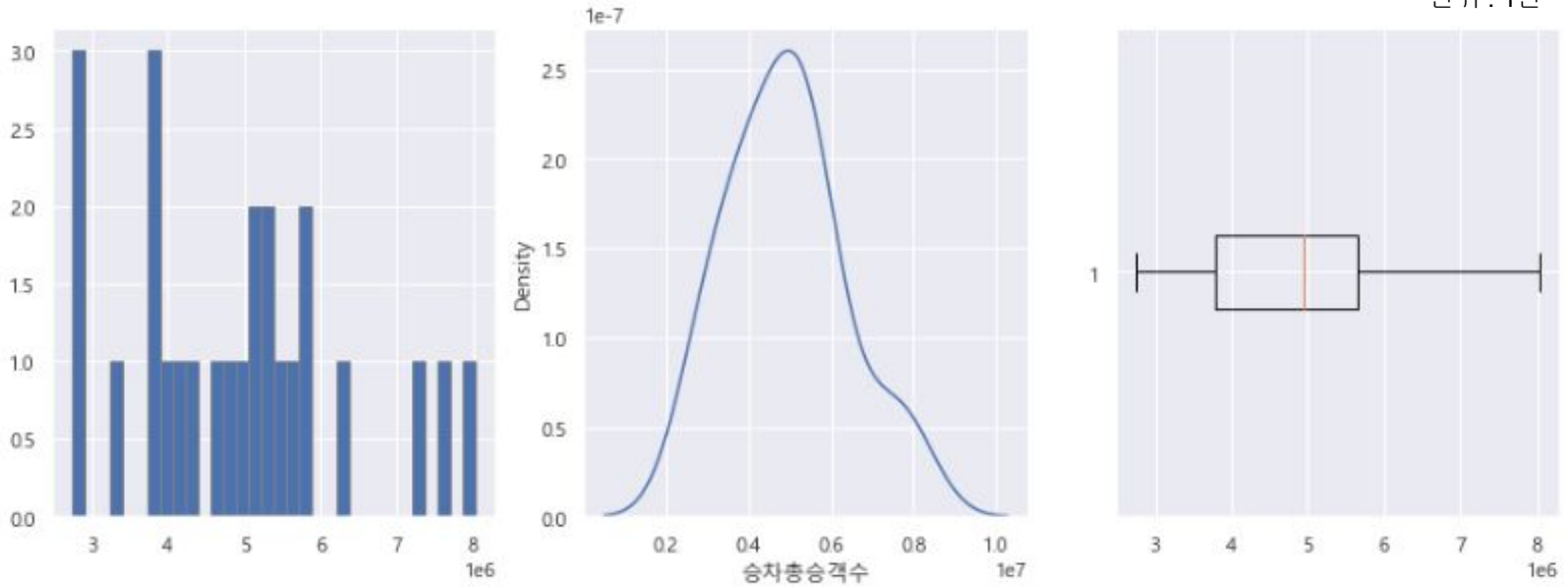
✔ **가설 3** : 정류장 수 대비 노선수가 많으면 평균이동시간이 줄어들 것이다

✔ **가설 4** : 평균 이동시간 대비 승하차승객수가 노선 수와 상관 있을 것이다.

단변량 분석 - 승차총승객수

▪4월 한달간 구 별 승객 수 분포 그래프

단위 : 1만

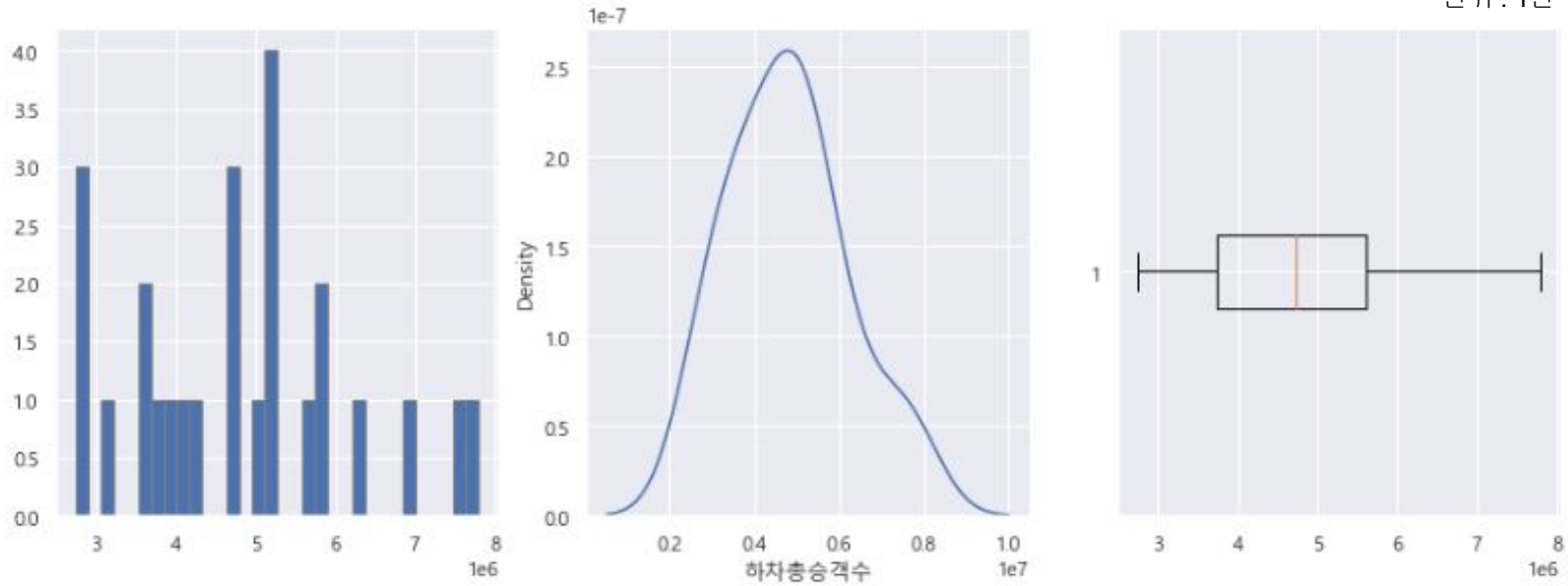


→ 승차승객수가 특히 많은(700만 이상) 몇개의 자치구가 존재함

단변량 분석 - 하차총승객수

▪구 별 하차 승객 수 분포 그래프

단위 : 1만

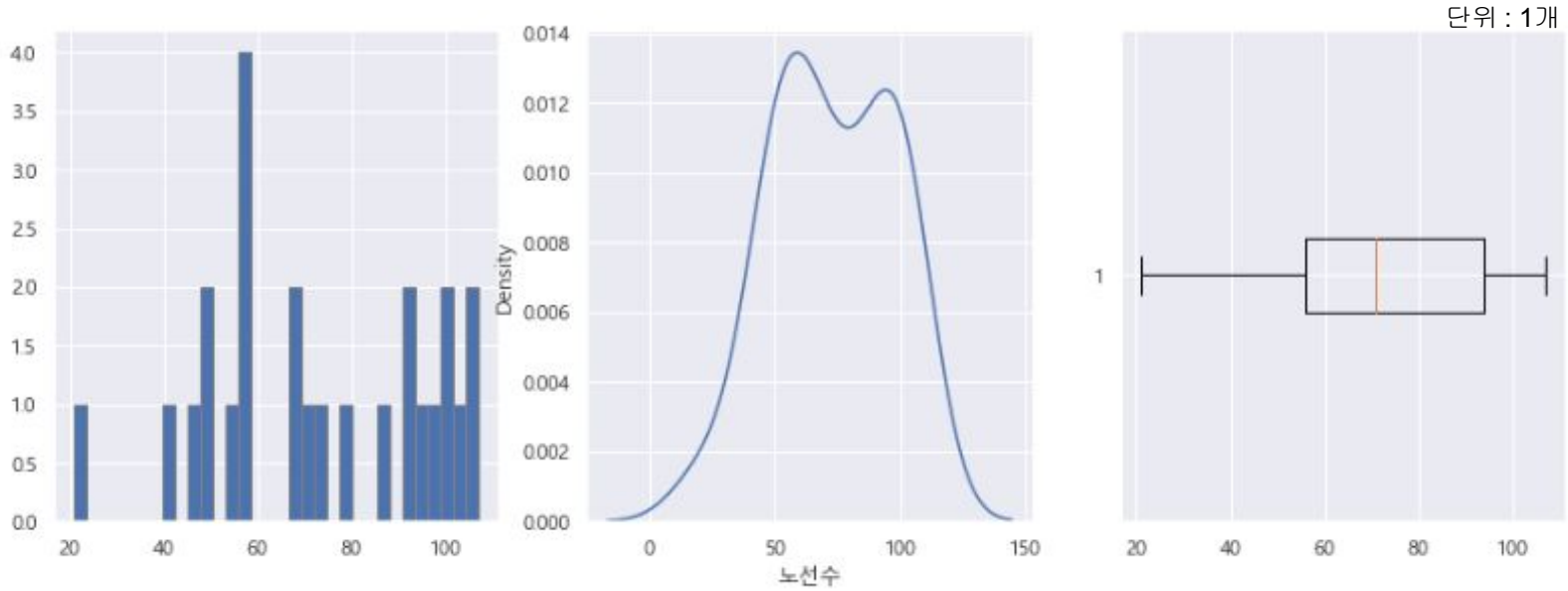


→ 승차승객수와 마찬가지로 하차승객수가 특히 많은 몇개의 자치구가 존재함

단변량 분석 - 노선수

✔ 승차총승객수 ✔ 하차총승객수 ✔ 노선수

▪구 별 노선수 분포 그래프



다른 구들에 비해 노선수가 특히 적은 자치구 존재 → 수요 존재시 버스 노선 증설 대상

이변량 분석

▪ Y(노선수, 정류장수)와의 관계 분류

1)강한 상관관계 존재

* 승차(하차)총승객수(x)와 노선수(y) 사이에 양의 상관관계(+)
* 승하차총합수(x)와 정류장수(y) 사이에 양의 상관관계(+)
* 정류장 수 대비 노선수(x)와 평균이동시간(y) 사이에 음의 상관관계(-)

—> 정류장 수 대비 노선수가 많으면 평균 이동시간이 줄어들기 때문에 노선을 많이 개설하면 교통문제 해소 가능

* 평균 이동시간 대비 승하차총승객수(x)와 노선수(y) 사이에 강한 양의 상관관계(+)

2)약한 상관관계 존재

정류장수 - 노선수 : 양의 상관관계(+)

총 이동인구 - 노선수 : 양의 상관관계(+)

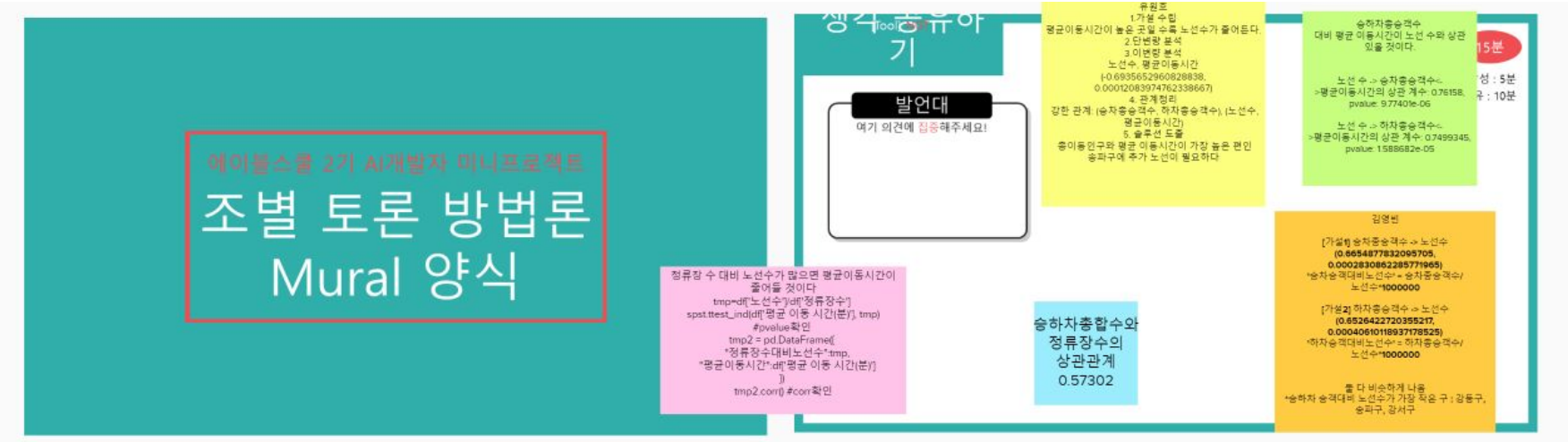
총 이동시간 - 노선수 : 양의 상관관계(+)

기타수단 이동인구 - 정류장수 : 양의 상관관계(+) ← (기타수단 이동인구 = 총 이동인구-승하차총합)

가설 검증 과정

- 조별 토론 방법론 적용 내용을 보여주세요.

  - **MURAL**을 사용하여 포스트잇으로 각자 가설을 공유하며 토론 진행



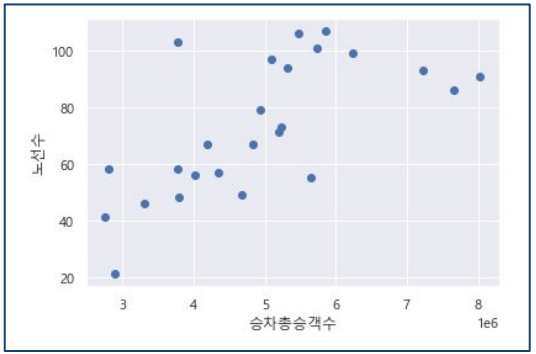
가설 검증 과정

[1] 승차총승객수와 하차총승객수는 노선수와 관련이 있을 것이다.

```
target = '노선수'
feature = '승차총승객수'

# 산점도 시각화
plt.scatter(feature, target, data = df)
plt.xlabel('승차총승객수')
plt.ylabel('노선수')
plt.show()


# 상관계수 분석
spst.pearsonr(df[feature], df[target])
```




```
(0.6654877832095705, 0.0002830862285771965)
```

- 강한 상관계수, 0.05보다 작은 p-value
- 하차총승객수도 거의 동일하게 나옴

'승차승객대비노선수'와 '하차승객대비노선수' 칼럼을 새로 만들어주어서 승하차 승객수 대비 노선수가 가장 적은 구를 찾아준다.

```
pd.options.display.float_format = '{:.2f}'.format
df['승차승객대비노선수'] = df['노선수'] / df['승차총승객수'] *1000000

pd.options.display.float_format = '{:.2f}'.format
df['하차승객대비노선수'] = df['노선수'] / df['하차총승객수'] *1000000

df.sort_values(by=['승차승객대비노선수']).head()
```

	자치구	정류장수	노선수	승차총승객수	하차총승객수	승차평균승객수	하차평균승객수	평균 이동시간	평균 이동인구	총 이동시간	총 이동인구	승차승객대비노선수	하차승객대비노선수
1	강동구	369	21	2890053	2830506	99.16	97.12	25.93	73.49	5082380	14404079.28	7.27	7.42
17	송파구	415	55	5641742	5603488	114.27	113.50	25.74	94.97	6234190	23004946.79	9.75	9.82
3	강서구	566	49	4681083	4652828	88.12	87.59	25.52	65.19	5486630	14016047.01	10.47	10.53
4	관악구	466	86	7655819	7792476	154.75	157.52	23.48	53.06	5150450	11638671.68	11.23	11.04
0	강남구	499	91	8030483	7569213	128.77	121.37	23.42	104.46	6543820	29188819.52	11.33	12.02

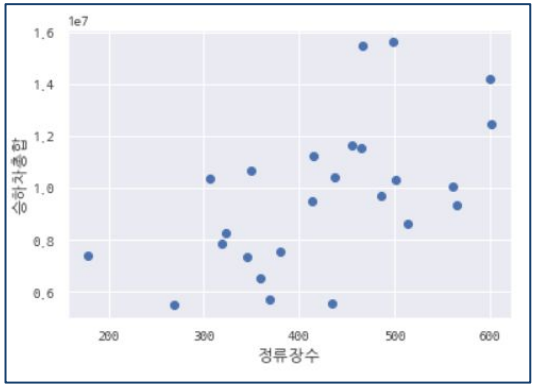
가설 검증 과정

[2] 승하차총합수와 정류장 수는 관련이 있을 것이다.

```
# 칼럼 생성
df['승하차총합'] = df['승차총승객수'] + df['하차총승객수']
df.head(2)

target = '승하차총합'
feature = '정류장수'

# 산점도 시각화
plt.scatter(feature, target, data = df)
plt.xlabel(feature)
plt.ylabel(target)
plt.show()
```




```
# 상관계수 확인
spst.pearsonr(df['승하차총합'], df['정류장수'])
```

```
(0.573022669787874, 0.0027529734757331683)
```

- 강한 상관계수, 0.05보다 작은 p-value

```
[32]: df['정류장수대비 승하차총합'] = df['승하차총합']/df['정류장수']
      df.sort_values(by='정류장수대비 승하차총합', ascending=False).head(3)
```

[32]:

	자치구	정류장수	노선수	승차총승객수	하차총승객수	승차평균승객수	하차평균승객수	평균 이동시간	평균 이동인구	총 이동시간	총 이동인구	승차승객대비노선수	하차승객대비노선수	승하차총합	기타수단 이동인구	정류장수대비 승하차총합
23	중구	178	103	3776675	3598932	121.02	115.32	21.52	46.24	5368440	11533349.41	27.27	28.62	7375607	4157742.41	41435.99
10	동대문구	306	73	5240565	5115379	131.84	128.69	22.15	41.07	5132710	9517475.46	13.93	14.27	10355944	-838468.54	33842.95
4	관악구	466	86	7655819	7792476	154.75	157.52	23.48	53.06	5150450	11638671.68	11.23	11.04	15448295	-3809623.32	33150.85

가설 검증 과정

[3] 정류장 수 대비 노선수가 많으면 평균이동시간이 줄어들 것이다.

```
tmp=df['노선수']/df['정류장수']
spst.pearsonr(df['평균 이동시간'], tmp) #pvalue확인
```

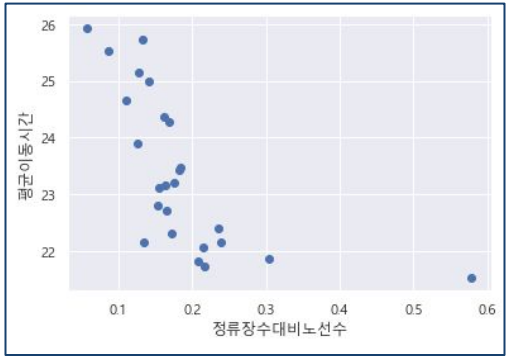
```
(-0.6359705297245506, 0.0006333143239016103)
```

- 강한 상관계수
- 0.05보다 작은 p-value

```
# 상관계수 확인
tmp2 = pd.DataFrame({
    "정류장수대비노선수":tmp,
    "평균이동시간":df['평균 이동시간']
})
tmp2.corr() #corr확인
```

	정류장수대비노선수	평균이동시간
정류장수대비노선수	1.00	-0.64
평균이동시간	-0.64	1.00

```
# 시각화
plt.scatter(tmp, df['평균 이동시간'])
plt.xlabel('정류장수대비노선수')
plt.ylabel('평균이동시간')
plt.show()
```



가설 검증 과정

[4] 평균 이동시간 대비 승하차객수가 노선 수와 상관 있을 것이다.

```
test['승차총승객수<->평균이동시간'] = df['승차총승객수'] // df['평균 이동시간']
test['하차총승객수<->평균이동시간'] = df['하차총승객수'] // df['평균 이동시간']

# 상관계수 확인

temp = spst.pearsonr(test['노선수'], test['승차총승객수<->평균이동시간'])
print('노선 수 -> 승차총승객수<->평균이동시간의 상관 계수: {}, pvalue: {}'.format(temp[0], temp[1]))
temp = spst.pearsonr(test['노선수'], test['하차총승객수<->평균이동시간'])
print('노선 수 -> 하차총승객수<->평균이동시간의 상관 계수: {}, pvalue: {}'.format(temp[0], temp[1]))
```

```
노선 수 -> 승차총승객수<->평균이동시간의 상관 계수: 0.7615827109075217, pvalue: 9.774016432776943e-06
------------------------------
노선 수 -> 하차총승객수<->평균이동시간의 상관 계수: 0.7499345331162306, pvalue: 1.5886827219705295e-05
```

- 강한 상관계수, 0.05보다 작은 p-value

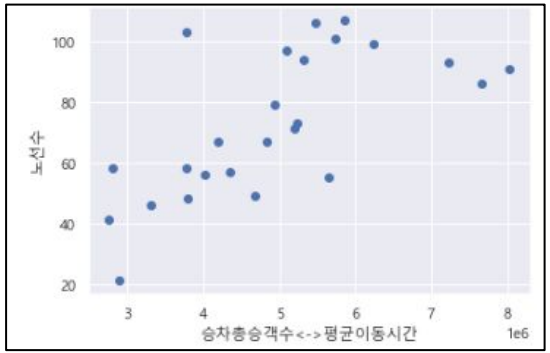
```
# 정렬
total_on = test.sort_values(by='승차총승객수<->평균이동시간', ascending=False)
total_out = test.sort_values(by='하차총승객수<->평균이동시간', ascending=False)
```

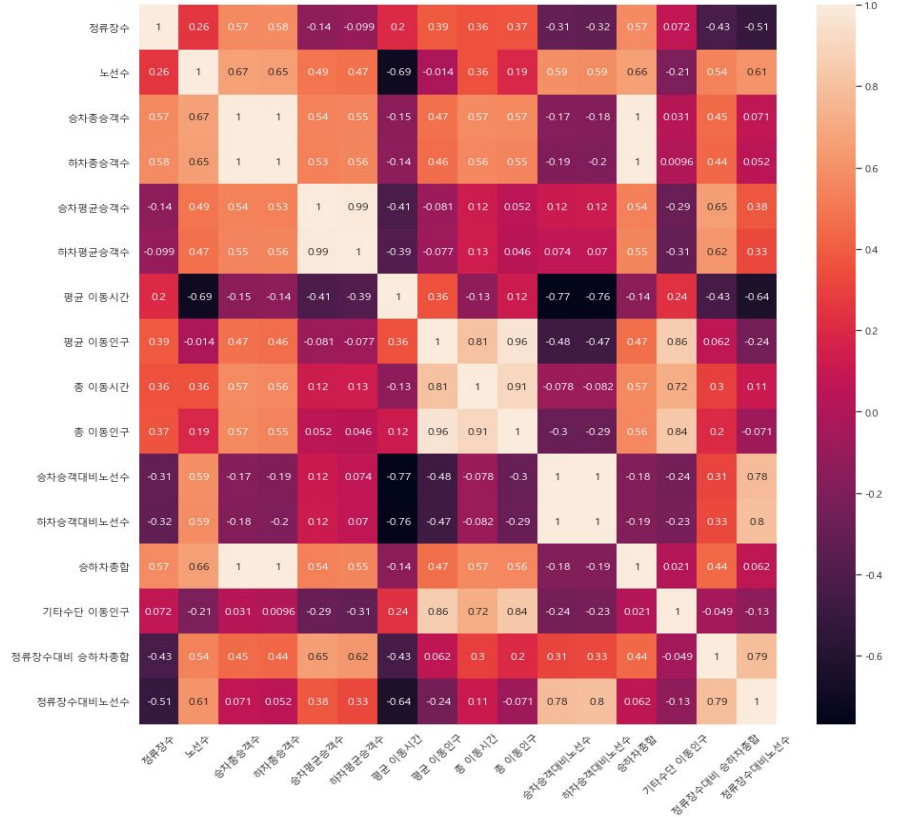
가설 검증 과정

[4] 평균 이동시간 대비 승하차객수가 노선 수와 상관 있을 것이다.

```
# 히트맵 시각화
plt.figure(figsize = (13, 13))
sns.heatmap(total_on.corr(), annot = True)
plt.show()

# 산점도 시각화
plt.scatter(feature, target, data = df)
plt.xlabel('승차총승객수<->평균이동시간')
plt.ylabel('노선수')
plt.show()
```





결론

- 어느 구에 버스 시설의 추가가 가장 필요한가요??

가설1 : 승차총승객수와 하차총승객수는 노선수와 관련이 있을 것이다.

- 승차총승객수가 많은 지역은 많은 노선수를 갖고 있는 것으로 미루어 보아
- **'승차승객대비노선수'**가 가장 작은 구에 필요하다고 생각된다.

==> 강동구, 송파구, 강서구

<'승차승객대비노선수' 기준 오름차순 정렬해서 확인>

```
[60]: df.sort_values(by='승차승객대비노선수').head(3)
```

[60]:

	자치구	정류장수	노선수	승차총승객수	하차총승객수	승차평균승객수	하차평균승객수	평균 이동시간	평균 이동인구	총 이동시간	총 이동인구	승차승객대비노선수	하차승객대비노선수	승하차총합	기타수단 이동인구
1	강동구	369	21	2890053	2830506	99.16	97.12	25.93	73.49	5082380	14404079.28	7.27	7.42	5720559	8683520.28
17	송파구	415	55	5641742	5603488	114.27	113.50	25.74	94.97	6234190	23004946.79	9.75	9.82	11245230	11759716.79
3	강서구	566	49	4681083	4652828	88.12	87.59	25.52	65.19	5486630	14016047.01	10.47	10.53	9333911	4682136.01

결론

가설2 : 승하차총합수와 정류장 수는 관련이 있을 것이다.

- **'정류장수대비승차총합'** 칼럼을 추가
- '정류장수대비승하차총합'이 가장 큰 곳에 정류장 설치 필요

==> 중구, 동대문구, 관악구

<'정류장수대비 승하차총합' 기준 내림차순 정렬해서 확인>

[32]:
```
df['정류장수대비 승하차총합'] = df['승하차총합']/df['정류장수']
df.sort_values(by='정류장수대비 승하차총합', ascending=False).head(3)
```

[32]:

	자치구	정류장수	노선수	승차총승객수	하차총승객수	승차평균승객수	하차평균승객수	평균 이동시간	평균 이동인구	총 이동시간	총 이동인구	승차승객대비노선수	하차승객대비노선수	승하차총합	기타수단 이동인구	정류장수대비 승하차총합
23	중구	178	103	3776675	3598932	121.02	115.32	21.52	46.24	5368440	11533349.41	27.27	28.62	7375607	4157742.41	41435.99
10	동대문구	306	73	5240565	5115379	131.84	128.69	22.15	41.07	5132710	9517475.46	13.93	14.27	10355944	-838468.54	33842.95
4	관악구	466	86	7655819	7792476	154.75	157.52	23.48	53.06	5150450	11638671.68	11.23	11.04	15448295	-3809623.32	33150.85

결론

가설3 : 정류장 수 대비 노선수가 많으면 평균이동시간이 줄어들 것이다

- 정류장 수 대비 노선수가 많아지면 평균 이동시간이 줄어들 것이기 때문에
- 각 자치구들을 '정류장수대비노선수' 컬럼을 기준으로 정렬했을 때,
- **'정류장수대비노선수'**가 가장 적은 구에 노선을 설치해야 한다.

==> 강동구, 강서구, 노원구

*<**'정류장수대비노선수'** 기준 오름차순 정렬해서 확인>*

```
[33]: df['정류장수대비노선수'] = df['노선수']/df['정류장수']
      df.sort_values(by="정류장수대비노선수", ascending=True).head(3)
```

[33]:

	자치구	정류장수	노선수	승차총승객수	하차총승객수	승차평균승객수	하차평균승객수	평균 이동시간	평균 이동인구	총 이동시간	총 이동인구	승차승객대비노선수	하차승객대비노선수	승하차총합	기타수단 이동인구	정류장수대비 승하차총합	정류장수대비노선수
1	강동구	369	21	2890053	2830506	99.16	97.12	25.93	73.49	5082380	14404079.28	7.27	7.42	5720559	8683520.28	15502.87	0.06
3	강서구	566	49	4681083	4652828	88.12	87.59	25.52	65.19	5486630	14016047.01	10.47	10.53	9333911	4682136.01	16491.01	0.09
8	노원구	514	57	4353295	4292724	88.48	87.25	24.66	59.06	5235480	12538744.48	13.09	13.28	8646019	3892725.48	16821.05	0.11

결론

가설4 : 평균 이동시간 대비 승하차총승객수 노선 수와 상관 있을 것이다.

=> 기존 승차 총객수와 상관관계가 유사한 결과, 그러나 기존 노선 평균이동시간간의 -0.69의 상관계수보다 높은 0.74~0.76의 상관계수를 도출

결론)

- 평균이동시간 대비 승하차객수가 높은 곳은 수요가 높다고 판단되어 해당 지역에 우선적으로 노선을 투입해야한다.
- 그러므로 **평균이동시간 대비 승차객** 기준 **강남구**, **평균이동시간 대비 하차객** 기준 가장 수요가 높은 **관악구**에 노선을 증설해야 한다.

<내림차순 정렬해서 확인>

```
[66]: total_on.head()
```

[66]:

	자치구	정류장수	노선수	승차총승객수	하차총승객수	승차평균승객수	하차평균승객수	평균 이동시간	평균 이동인구	총 이동시간	총 이동인구	승차승객대비노선수	하차승객대비노선수	승하차총합	승차총승객수<->평균이동시간	하차총승객수<->평균이동시간
0	강남구	499	91	8030483	7569213	128.77	121.37	23.42	104.46	6543820	29188819.52	11.33	12.02	15599696	342907.00	323211.00
4	관악구	466	86	7655819	7792476	154.75	157.52	23.48	53.06	5150450	11638671.68	11.23	11.04	15448295	326046.00	331866.00
14	서초구	600	93	7221330	6977950	126.69	122.42	23.12	72.54	5944440	18651195.64	12.88	13.33	14199280	312345.00	301818.00
16	성북구	602	99	6236424	6231238	123.25	123.15	22.72	48.97	5232500	11278789.10	15.87	15.89	12467662	274534.00	274306.00
19	영등포구	465	101	5739875	5763211	120.34	121.25	21.75	61.76	5525180	15693027.00	17.60	17.46	11523086	263956.00	265948.00

```
[67]: total_out.head()
```

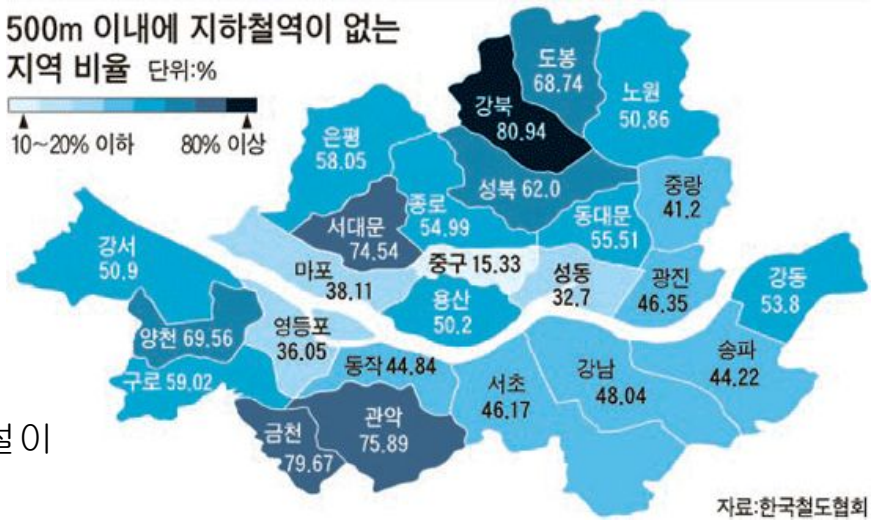
[67]:

	자치구	정류장수	노선수	승차총승객수	하차총승객수	승차평균승객수	하차평균승객수	평균 이동시간	평균 이동인구	총 이동시간	총 이동인구	승차승객대비노선수	하차승객대비노선수	승하차총합	승차총승객수<->평균이동시간	하차총승객수<->평균이동시간
4	관악구	466	86	7655819	7792476	154.75	157.52	23.48	53.06	5150450	11638671.68	11.23	11.04	15448295	326046.00	331866.00
0	강남구	499	91	8030483	7569213	128.77	121.37	23.42	104.46	6543820	29188819.52	11.33	12.02	15599696	342907.00	323211.00
14	서초구	600	93	7221330	6977950	126.69	122.42	23.12	72.54	5944440	18651195.64	12.88	13.33	14199280	312345.00	301818.00
16	성북구	602	99	6236424	6231238	123.25	123.15	22.72	48.97	5232500	11278789.10	15.87	15.89	12467662	274534.00	274306.00
19	영등포구	465	101	5739875	5763211	120.34	121.25	21.75	61.76	5525180	15693027.00	17.60	17.46	11523086	263956.00	265948.00

최종결론

- 각 가설 별 노선의 추가 개설이 필요한 상위 3개 구

가설1) 강동구, 송파구, 강서구
가설2) 중구, 동대문구, 관악구
가설3) 강동구, 강서구, 노원구
가설4) 강남구, 관악구, 서초구



- 강동구, 강서구, 관악구 등에 버스노선의 추가 개설이 필요하다.
- 위의 자치구들을 경로로 두는 노선이 필요해보인다.
- 지하철역이 없는 지역을 고려하는 것도 필요해보인다.

AIVLE

AIVLE

Let’s make it possible