버스탑승
영향 요인 분석

버스탑승 영향 요인 분석 및 개선방안 제시

김종윤,남현원,이석진,정용훈

목차

Part 1

분석 개요 | 배경, 필요성

서울시의 '도로'의 증가보다 도로 위의 '차량'이 더 빠르게 증가해 왔고, 도로평균 속도가 약 17km로 서울시의 교통체증 문제가 여전한 가운데 해결되고 있지 않음

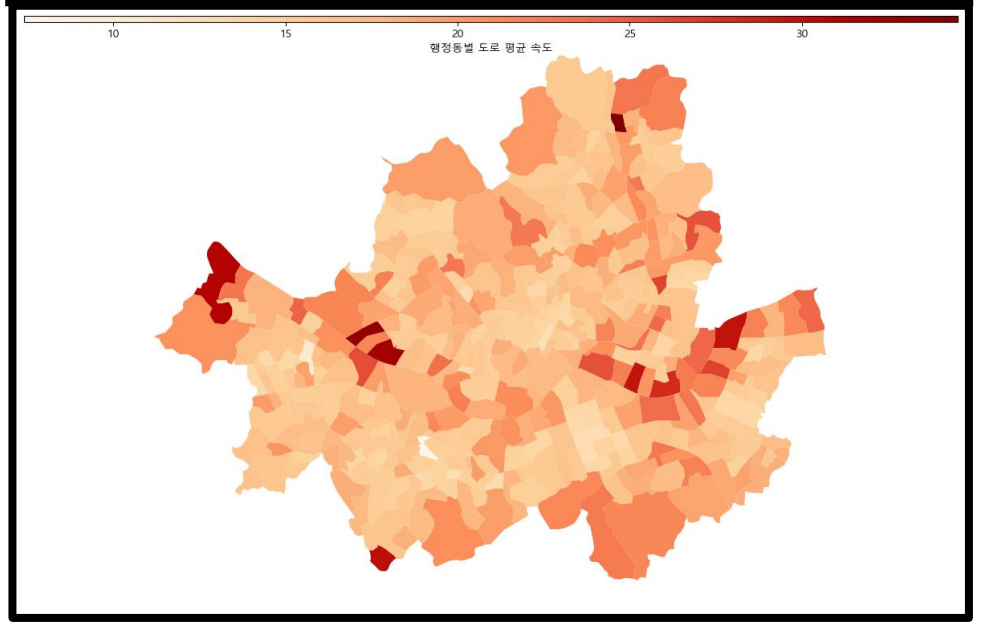
배경

- ✓ 서울시는 교통체증 문제 해결을 위해 다양한 정책 활용과 많은 노력을 기울였지만 해결되고 있지는 않음
- ✓ 서울시의 도로 평균 속도는 평일 평균 시속 17.5km, 전체 평균 시속은 17.3km으로 서울 시내 교통체증이 심하다는 것을 알 수 있음
- ✓ 자동차 등록대수는 2013년 이후로 계속 증가하고 있는 것을 알 수 있음

기존 문제점

- ✓ 2023년은 차량등록대수가 소폭 줄긴 했지만, 지속적인 수도권 쏠림 현상으로 인한 20-40대 증가 → 실질적인 자동차 이용 인구수는 증가
- ✓ 자동차는 많아지는데 도로의 증가의 더딘 속도로 인해 인프라가 자동차 수 커버를 못하고 있음

서울시 행정동별 도로 평균 속도



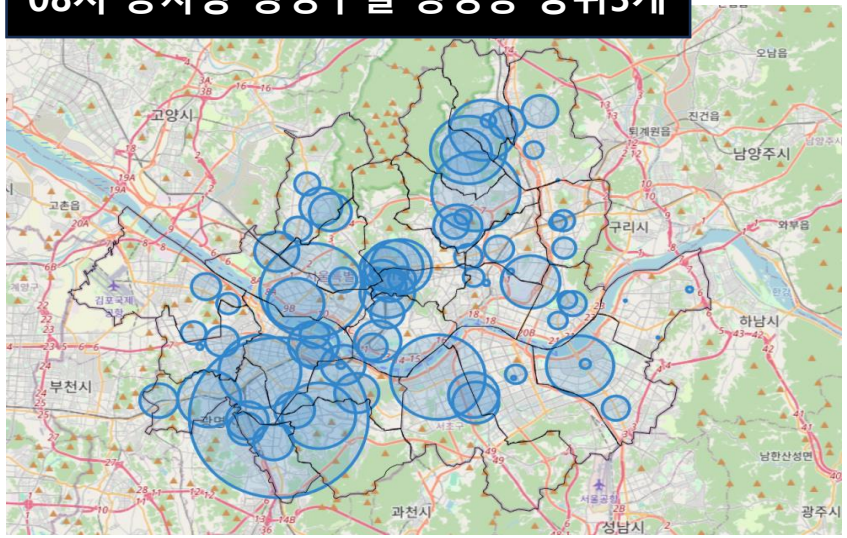
- ✓ 서울시 시내버스의 이용률을 높여 서울교통체증 문제 해결 목표
- ✓ 서울 시내버스 탑승 영향 요인 분석을 통해 해결방안 모색

Part 2

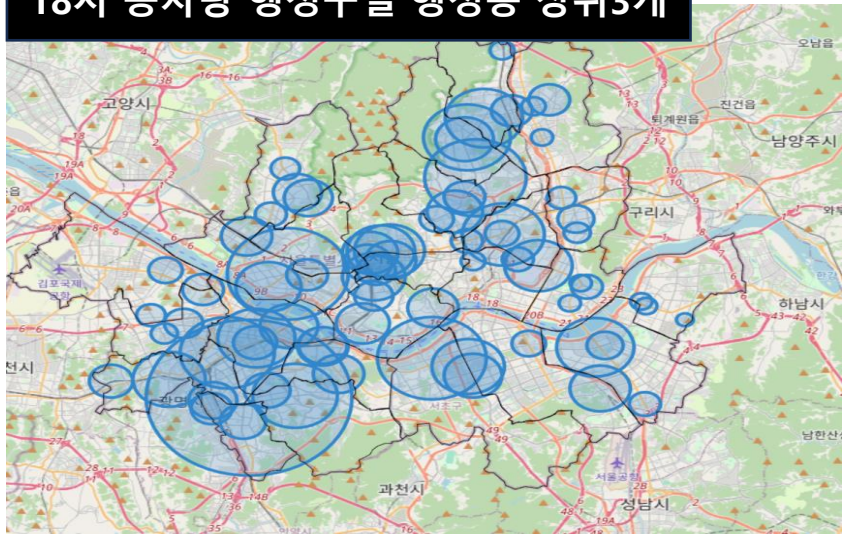
데이터 수집 | 시내버스 탑승 요인 고려

행정구역별 버스 승하차량을 통해서 승하차량이 많은 지역과 적은 지역의 특징 확인

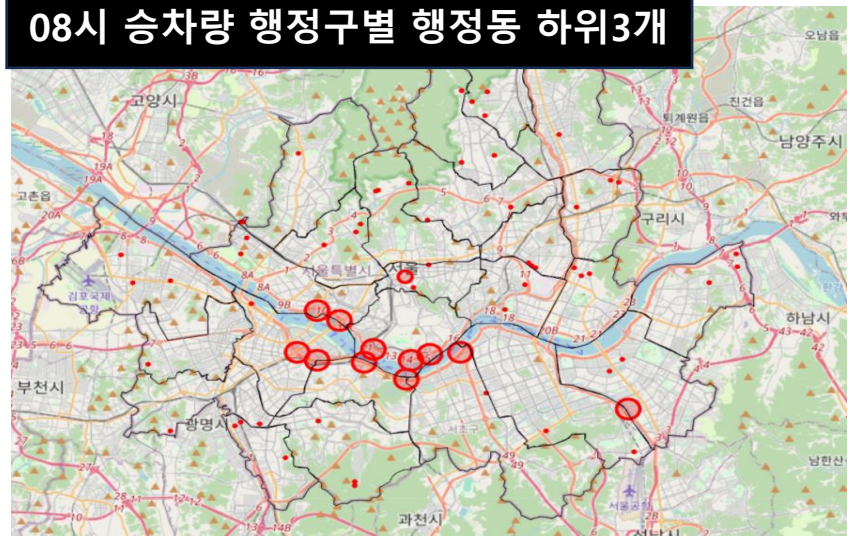
08시 승차량 행정구별 행정동 상위3개



18시 승차량 행정구별 행정동 상위3개



08시 승차량 행정구별 행정동 하위3개



18시 승차량 행정구별 행정동 하위3개

출퇴근시간 행정구별
행정동 버스 승차량
상,하위 3개를 확인

각 지역의 특성을
확인하여 버스 탑승
영향 요인 고려

Part 2

데이터 수집 | 시내버스 탑승 요인 고려

✓ **기존 문제점과 앞서 확인한 지역 특징을 토대로 시내버스 탑승 요인 고려→ 인구, 사업체, 자동차, 교통체증 정도, 교통 인프라 고려 데이터 수집**

인구

출퇴근시간 교통체증 심화 : 30~40대 인구비율
학교 등교, 문화활동이 참여가 활발 : 10, 20대 인구 비율
노약자의 운전면허증 반납 : 노약자 인구 비율
인구가 거주하는 곳 : 집 개수

출퇴근

주간 상주지 : 사업체 수
사업체 종사하는 인원수 확인 : 종사자 수
주 종사자 : 30~40대 인구 비율
야간 상주지 : 집 개수

교통체증, 자동차

주 자동차 이용 인구 : 30~40대 인구비율
자동차 모는 20대 증가 : 20대 인구 비율
교통체증 정도 파악 : 도로 평균 속도
소득에 따른 자동차 소유 : 소득 수준, 자동차 등록 대수

교통 인프라

행정구역별 교통 인프라 확인
→ 버스, 지하철 정류장 개수
→ 시내버스노선 개수

15개의 변수

인구수	10대 인구 비율	20대 인구 비율
30, 40대 인구 비율	노약자 인구 비율	사업체수
종사자수	집 개수	소득 수준
도로 평균 속도	1인당 자동차 등록 대수	버스 정류장 개수
지하철역 개수	시내버스 노선 개수	버스 이용 인원

Part 2

데이터수집 & 전처리 | 분석로드맵

데이터 수집

- 서울시 주민등록인구
- 서울시 사업체 현황
- 서울시 주택 통계
- 서울시 상권분석서비스
- 서울시 연령별 통계
- 서울시 지하철역 정보
- 서울시 정류장별 시간대별 승하차 인원정보
- 서울시 시내버스 정류장 현황
- 서울시 자동차 등록대수
- 서울시 읍면동 정보
- 서울시 혼잡시 평균속도

데이터 전처리

- 데이터 추출
- 데이터 병합
- 결측치 처리
- 이상치 처리
- 행정동 코드 결합
- 위경도 데이터 결합

분석 및 시각화

- 변수 상관계수 분석
- 후진제거법을 이용한 다중회귀 분석
- 다중 공선성 검정
- 잔차의 정규성, 독립성, 등분산성 분석
- 선택된 변수 분석 및 시각화
- 회귀 분석 모델 검정
- 회귀식을 통한 최적의 노선 결정 기준

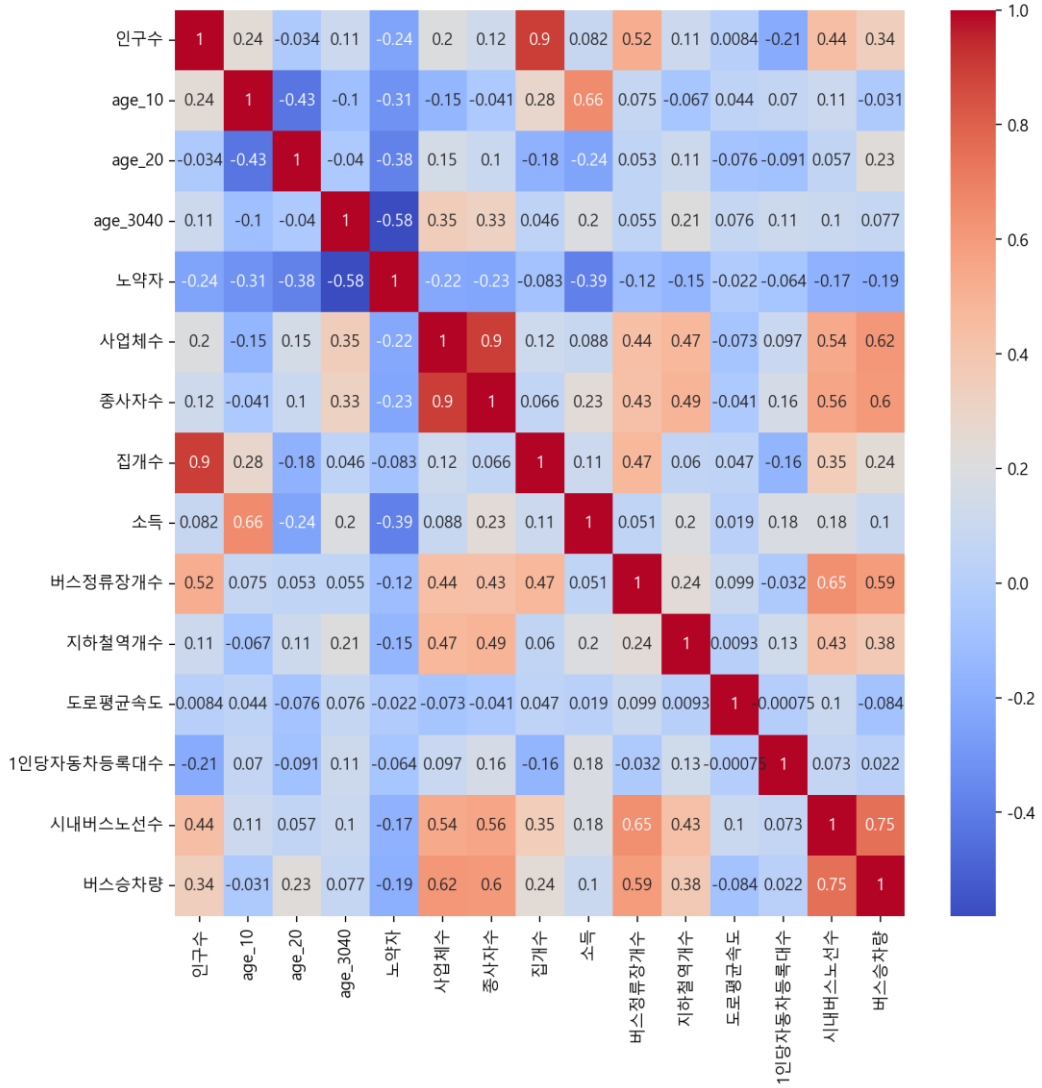
인사이트 도출

- 회귀식을 이용한 최적 노선 결정
- 최적 노선 타당성 검증 가능성 발견
- 기대효과, 활용방안 도출

Part 2

데이터 전처리 | 상관관계 확인

Heat Map



수집한 15개의 변수 상관관계 시각화

- ✓ 변수들 간의 강한 상관관계 : 다중 회귀 모델링에서 다중공선성의 가능성을 나타낼 수 있음
- ✓ 더 정확한 예측을 위해 어떤 변수를 포함시킬지, 제외시킬지 고려
- ✓ Heat Map : 더욱 직관적이고 시각적으로 표현

Correlation Analysis

- ✓ 두 개 변수 사이에 존재하는 상호 연관성 존재와 그 강도를 측정
- ✓ 1과 -1사이의 값을 가짐
- ✓ 1에 가깝고 빨간색이 짙을 수록 상관관계가 크다는 것을 의미

데이터 분석 | 단계적 회귀

Stepwise Regression

통계적인 회귀 분석에서 결과값에 유의(有意)한 영향을 미치는 독립 변수의 항을 결정하는 방법.

***다중회귀분석: 독립 변수(설명 변수)가 2개 이상인 경우를 분석 대상으로 하는 회귀 분석 방법 중 하나**

OLS Regression Results

Dep. Variable:	버스 승차량	R-squared:	0.680
Model:	OLS	Adj. R-squared:	0.675
Method:	Least Squares	F-statistic:	148.0
Date:	Sat, 27 Jan 2024	Prob (F-statistic):	4.05e-100
Time:	22:24:44	Log-Likelihood:	-6652.0
No. Observations:	425	AIC:	1.332e+04
Df Residuals:	418	BIC:	1.335e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.593e+06	7.42e+04	48.394	0.000	3.45e+06	3.74e+06
20대 비율	3.724e+05	7.56e+04	4.924	0.000	2.24e+05	5.21e+05
3040 비율	-1.632e+05	8.11e+04	-2.013	0.045	-3.23e+05	-3802.585
사업체수	6.943e+05	9.85e+04	7.050	0.000	5.01e+05	8.88e+05
버스정류장 개수	3.931e+05	9.87e+04	3.985	0.000	1.99e+05	5.87e+05
도로 평균 속도	-3.188e+05	7.65e+04	-4.169	0.000	-4.69e+05	-1.69e+05
시내버스 노선수	1.417e+06	1.06e+05	13.398	0.000	1.21e+06	1.62e+06

Omnibus:	77.787	Durbin-Watson:	1.740
Prob(Omnibus):	0.000	Jarque-Bera (JB):	250.210
Skew:	0.819	Prob(JB):	4.65e-55

...

후진 제거법

모든 독립 변수를 다중 회귀 모형에 포함하여 회귀 분석 한 다음 적정한 회귀 모형이 구성될 때까지 가장 중요하지 않다고 생각되는 변수부터 차례로 제거

A + B + C +D → (변수 제거) → A + B + C

변수 제거
→ VIF수치, P-value 값을 확인하여 차례로 변수 제거
제거한 변수
→ 인구수, age_10, 노약자, 종사자수, 집개수, 소득, 지하철 역 개수, 1인당 자동차 등록대수

R-squared : 0.675
개선점이 필요하지만 결과값이 유의미하다고 판단

Part 3

데이터 분석 | 모델 검정1 – 독립성, 다중공선성

잔차의 독립성 확인

Omnibus:	77.787	Durbin-Watson:	1.740
Prob(Omnibus):	0.000	Jarque-Bera (JB):	250.210
Skew:	0.819	Prob(JB):	4.65e-55
Kurtosis:	6.383	Cond. No.	2.59

Durbin-Waston	1.740

✓ 더빈-왓슨 테스트 : 회귀분석 후 잔차의 독립성을 확인
✓ 더빈-왓슨 통계량
2에 가까울수록 오차항의 자기 상관이 없음
1.5 ~ 2.5 이면 정상이라 판단

Durbin-Waston
1.740 : 1.5 ~ 2.5사이이므로 정상

다중공선성 문제 확인

VIF Factor	Features
1.038173	20대 비율
1.192742	30,40대 비율
1.759459	사업체수
1.765904	버스정류장 개수
1.061142	도로 평균 속도
2.028490	시내버스 노선수

✓ VIF : 독립 변수간 상관관계가 있는지 측정하는 척도
✓ VIF 수치 : 값 작을수록 다중공선성 정도가 작을 것을 의미
✓ 판단기준 : 10이 넘으면 다중공선성이 있다고 판단

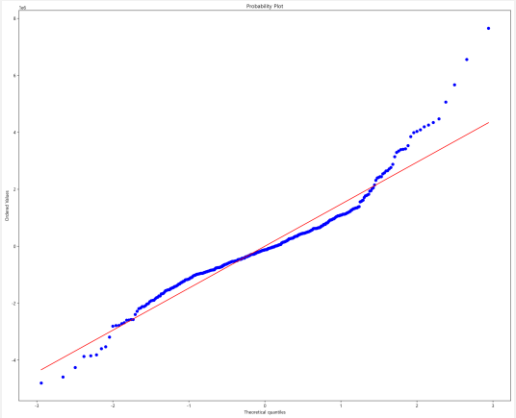
➢ 선택한 변수들의 VIF 수치 10이하
선택한 변수들의 상관관계가 적음

다중공선성이 적절히 제어되어 회귀모형이 신뢰성을 가질 가능성이 높음

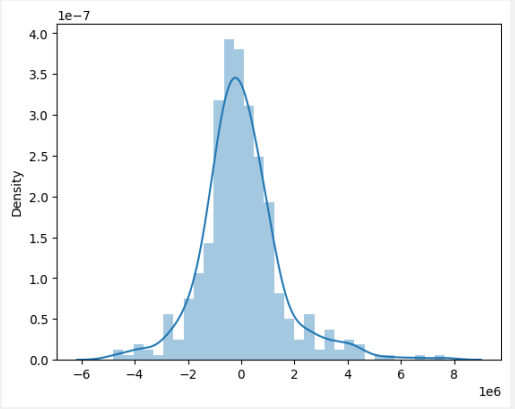
Part 3

데이터 분석 | 모델 검정2 – 정규성, 등분산성

정규성 확인



Q-Q plot

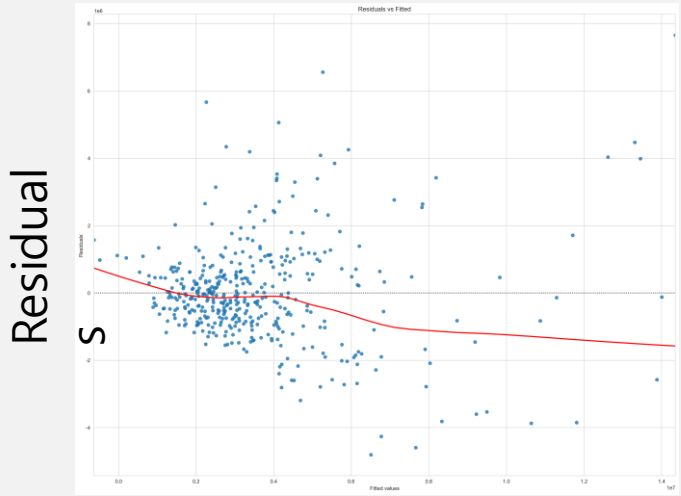


잔차의 히스토그램

- ✓ Q-Q plot : 시각적인 정규성 검정 방법 사용
- ✓ 확인 방법 : 점들이 45도 각도의 직선에 밀접한지 확인

- ✓ 잔차의 히스토그램 : 시각적인 정규성 검정 방법 사용
- ✓ 확인 방법 : 종모양에 가까우면 정규성 위배되지 않는다고 판단

등분산성 확인

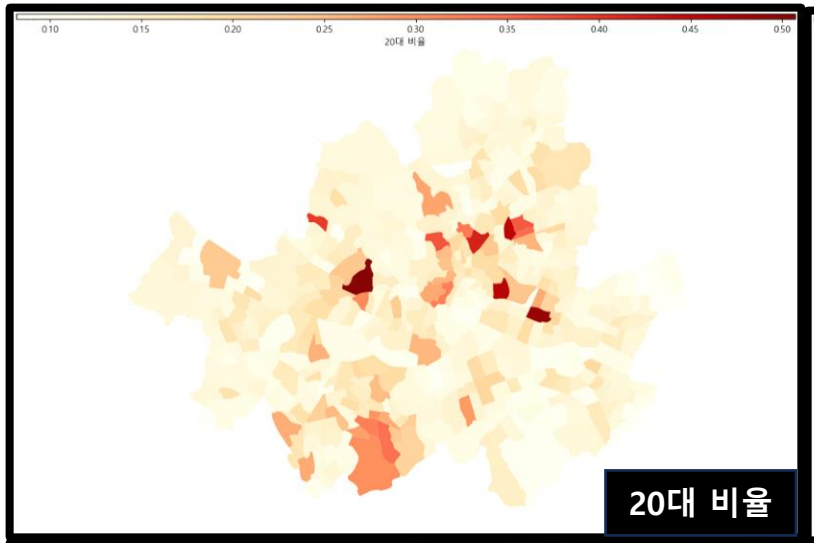


- ✓ Yhat 값에 따른 잔차가 무작위로 분포 됨
- ✓ 잔차의 등분산성이 만족함

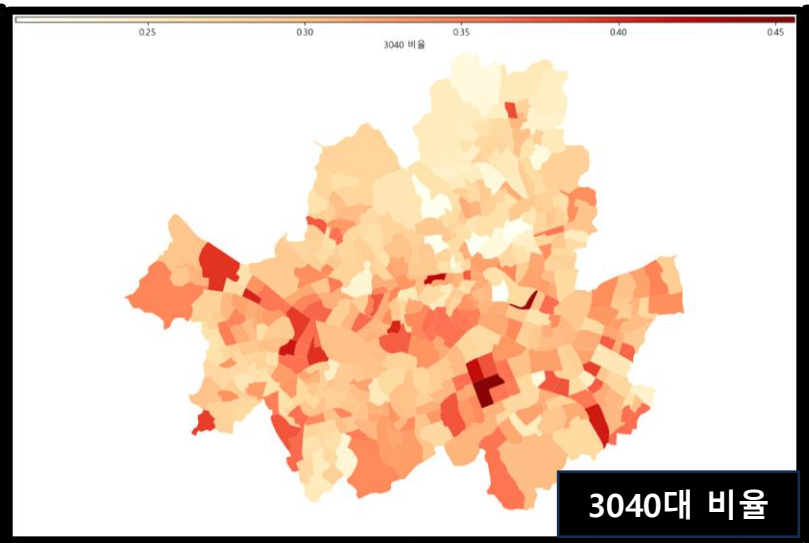
정규성, 등분산성이 적절히 제어되어 회귀모형이 신뢰성을 가질 가능성이 높음

Part 3

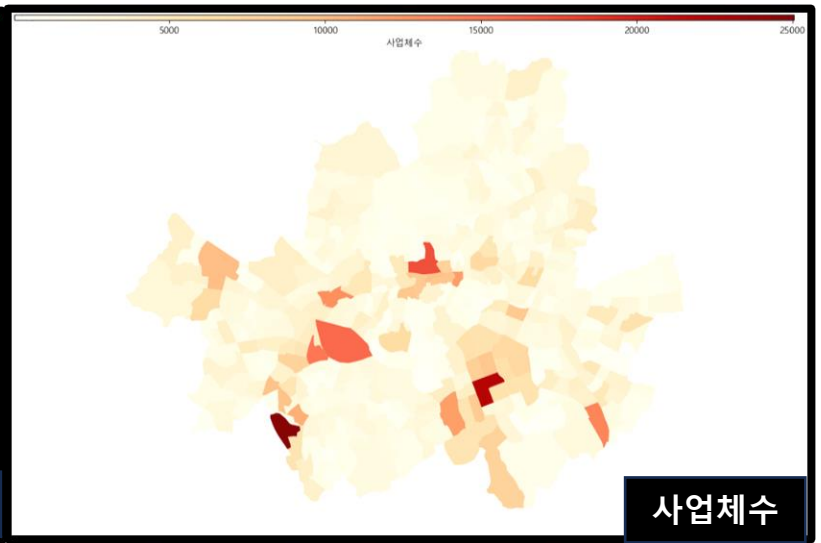
데이터 분석 | 변수 확인



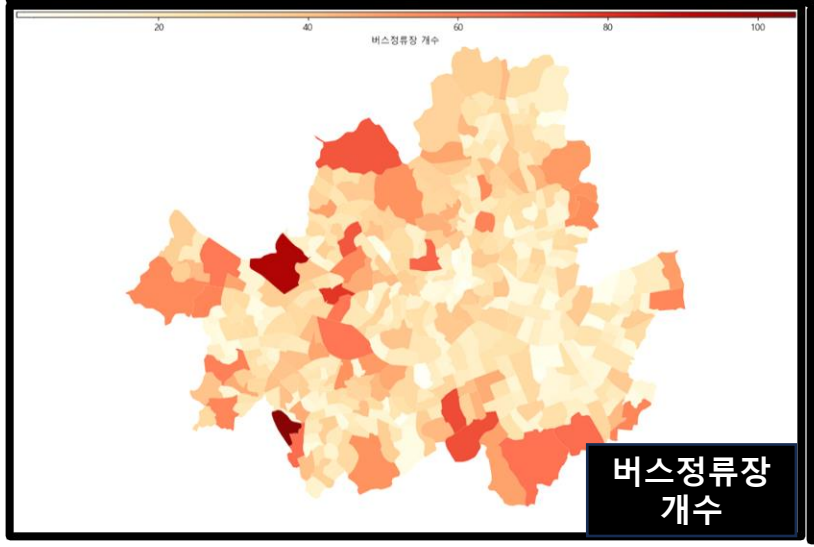
20대 비율



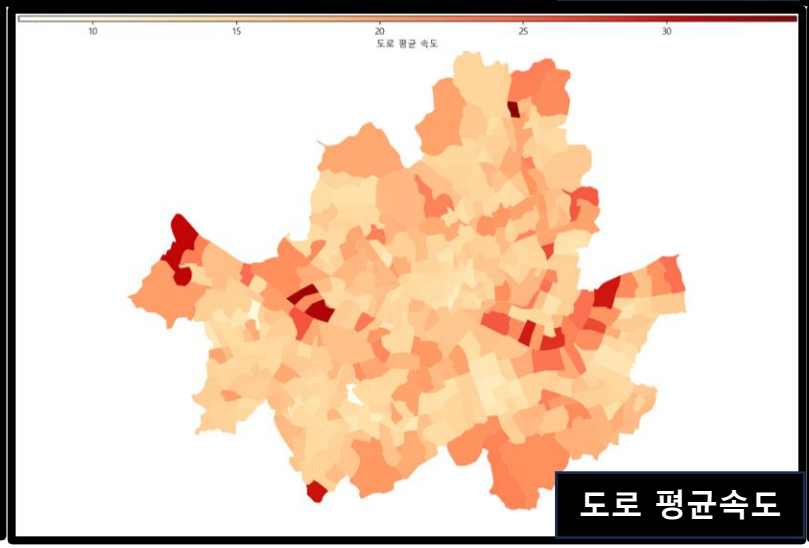
3040대 비율



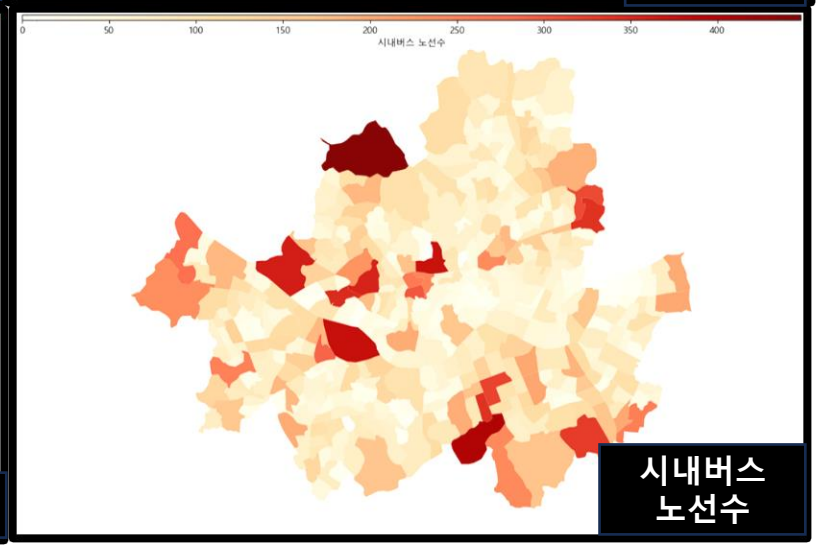
사업체수



버스정류장 개수



도로 평균속도



시내버스 노선수

데이터 분석 | 결과 : 회귀분석 – 모델 검정

다중회귀분석

독립 변수(설명 변수)가 2개 이상인 경우를 분석 대상으로 하는 회귀 분석 방법 중 하나

사용 변수	변수		설명
종속 변수	Y		버스탑승 인원 수
설명 변수	인구	x1	20대 비율
		x2	30, 40대 비율
		x3	사업체 수
	교통	x4	버스정류장 개수
		x5	도로 평균 속도
		x6	시내버스 노선 수

값마다 범위가 다름 => Standard Scaling을 적용해 분석을 용이하게
다중공선성 문제를 위해 VIF 기법을 활용하여 변수 제거

F-statistic: 148
P-value: 4.05e-100
R-Square: 0.680
Adj. R-squared: 0.675

P-value: 4.05e-100으로
모형은 통계적으로 유의함

수정된 결정계수가
전체의 67.5%를 설명

데이터 분석 | 결과 : 회귀분석 – 계수 검정

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.593e+06	7.42e+04	48.394	0.000	3.45e+06	3.74e+06
20대 비율	3.724e+05	7.56e+04	4.924	0.000	2.24e+05	5.21e+05
3040 비율	-1.632e+05	8.11e+04	-2.013	0.045	-3.23e+05	-3802.585
사업체수	6.943e+05	9.85e+04	7.050	0.000	5.01e+05	8.88e+05
버스정류장 개수	3.931e+05	9.87e+04	3.985	0.000	1.99e+05	5.87e+05
도로 평균 속도	-3.188e+05	7.65e+04	-4.169	0.000	-4.69e+05	-1.69e+05
시내버스 노선수	1.417e+06	1.06e+05	13.398	0.000	1.21e+06	1.62e+06

P-값

모든 변수의 p-value값이 0.05미만이므로 유의한 변수라 판단

Coef값

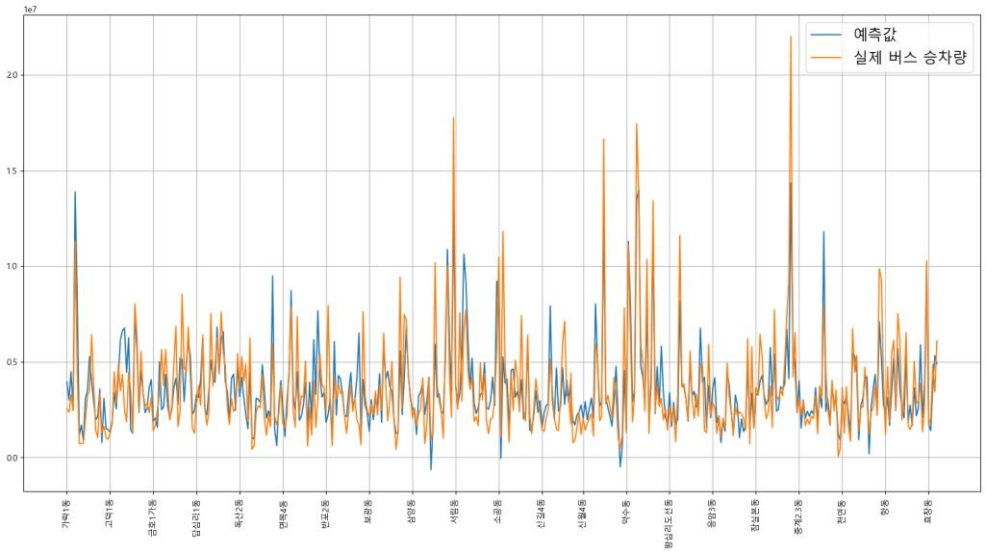
20대 비율, 사업체수, 버스정류장 개수, 시내버스 노선 수는 양(+) 요인

30, 40대 비율, 도로 평균 속도는 음(-) 요인

Part 4

결론 | 기대효과

결론



	행정구	행정동	20대 비율	3040 비율	사업체수	버스정류장 개수	도로 평균 속도	시내버스 노선수	예측값	버스 승차량	차이
142	강서구	방화2동	-0.238598	-0.033970	0.389894	0.635280	4.174134	2.686595	6504594	1698040	4806554
122	송파구	문정2동	0.212834	2.776801	4.035406	0.368751	0.472628	1.162957	7662059	3070248	4591811
28	강서구	공항동	-0.402017	1.057812	0.015564	2.034558	0.961276	2.117011	6773374	2513997	4259377
193	서초구	서초2동	-0.248124	0.974102	2.189026	1.234970	-0.547424	3.612170	10638002	6770853	3867149
369	은평구	진관동	-0.453819	-0.467008	0.125079	2.967409	0.750547	5.150047	11809425	7960072	3849353

	행정구	행정동	20대 비율	3040 비율	사업체수	버스정류장 개수	도로 평균 속도	시내버스 노선수	예측값	버스 승차량	차이
353	종로구	종로1.2.3.4가동	-0.029691	-0.133354	5.307471	2.767512	-0.425262	4.124795	14354784	22008316	-7653532
188	마포구	서교동	0.910841	0.341267	3.746417	3.500467	-0.617667	3.711847	13308377	17781562	-4473185
278	영등포구	여의동	-0.705792	0.075659	4.814468	2.434351	0.216088	4.167514	13451656	17444446	-3992790
262	서대문구	신촌동	5.854066	-1.576255	0.908170	1.368235	0.066440	3.840003	12616917	16653153	-4036236
279	강남구	역삼1동	0.924623	3.948565	7.153486	1.368235	-1.420881	3.355857	14003814	13882717	121097

회귀모델 예측결과 실제결과와 차이가 나는 동 선정 -> 강서구 방화2동, 종로 1,2,3,4가동 선정

결론 | 기대효과

기대효과

선정된 동의 정거장들간의 버스이동시간, 자차이동시간을 naver map API와 tmap API를 이용해 비교(Github의 bus_car_output.txt 참고)

버스이동시간과 자차이동시간의 차이가 큰 곳은 버스 이용이 어려운 곳, 이 구간만 가는 노선이 존재하면 이동시간 단축 가능.

시간 관계상 일부 역만 선택해 수행해보았지만, 모든 역에서 시간을 측정해 버스 이용이 불편한 구역을 선정해 효율적인 노선 추천 및 개설 시스템 구축 가능할 듯.

```
start :  ('126.8029604', '37.5713698')
goal :  ('126.9773838', '37.5723915')
67.46666666666667
find!==========================================
start :  서울특별시 강서구 하늘길 38
goal :  서울특별시 종로구 세종로 84-8
car_duration :  43.65688333333333
bus_duration :  67.46666666666667
================================================


start :  ('126.8029604', '37.5713698')
goal :  ('126.986903', '37.57044')
74.63333333333334
find!==========================================
start :  서울특별시 강서구 하늘길 38
goal :  서울특별시 종로구 종로2가
car_duration :  47.26545
bus_duration :  74.63333333333334
================================================


start :  ('126.8029604', '37.5713698')
goal :  ('126.98595', '37.575067')
75.48333333333333
find!==========================================
start :  서울특별시 강서구 하늘길 38
goal :  서울특별시 종로구 경운동
car_duration :  43.60755
bus_duration :  75.48333333333333
================================================
```

결론 | 한계 및 시사점

한계점

- 상권 분석, 지역의 특수한 행사와 같은 외부변수를 고려 하면 더 정확한 결과를 기대할 수 있을 것

- 시간이 더 주어진다면 특정한 정거장을 어이주는 노선을 추가 또는 개선하는 시스템을 설계해 볼 수 있을 것이라 생각됨

시사점

- 회귀분석의 결과를 통해 수요가 부족한 지역을 파악할 수 있는 기준을 제시함

- 정거장을 이어주는 노선을 추가할 수 있는 정량적인 기준을 제시함

Part 4

결론 | 데이터 출처

데이터명	데이터 출처	분석 내용	수집 범위
서울시 주민등록인구 (동별) 통계	서울시 열린 데이터 광장	서울시 행정동별 인구수 분석	2023년
서울시 사업체현황 (산업대분류별/동별) 통계		서울시 행정동별 사업체 수, 종사자 수 분석	2023년
서울시 주택종류별 주택 (동별) 통계		서울시 행정동별 거주지 수 분석	2023년
서울시 상권분석서비스(소득소비-행정동)		서울시 행정동별 평균 소득 분석	2023년 3분기
서울시 주민등록인구 (연령별/구별) 통계		서울시 인구별 비율 분석	2023년 4분기
서울시 역사마스터 정보		서울시 지하철역 개수 분석	2024년
서울시 버스노선별 정류장별 시간대별 승하차 인원 정보		서울시 행정동별 승차량 분석	2023년
서울시 시내버스 정류소 현황		서울시 행정동별 시내버스 노선 수 분석	2023년
서울시 행정동별 자동차 등록대수 현황		1인당 자동차 등록대수 분석	2023년
서울시 읍면동마스터 정보		데이터프레임 병합 목적으로 이용	2024년
2021년_혼잡시평균속도_행정구역_읍면동단위	국가 데이터 오픈 마켓	서울시 행정동별 도로 혼잡도 분석	2021년

Part 5

부록

-깃허브

https://github.com/hun9008/ideaton_mireaSW.git