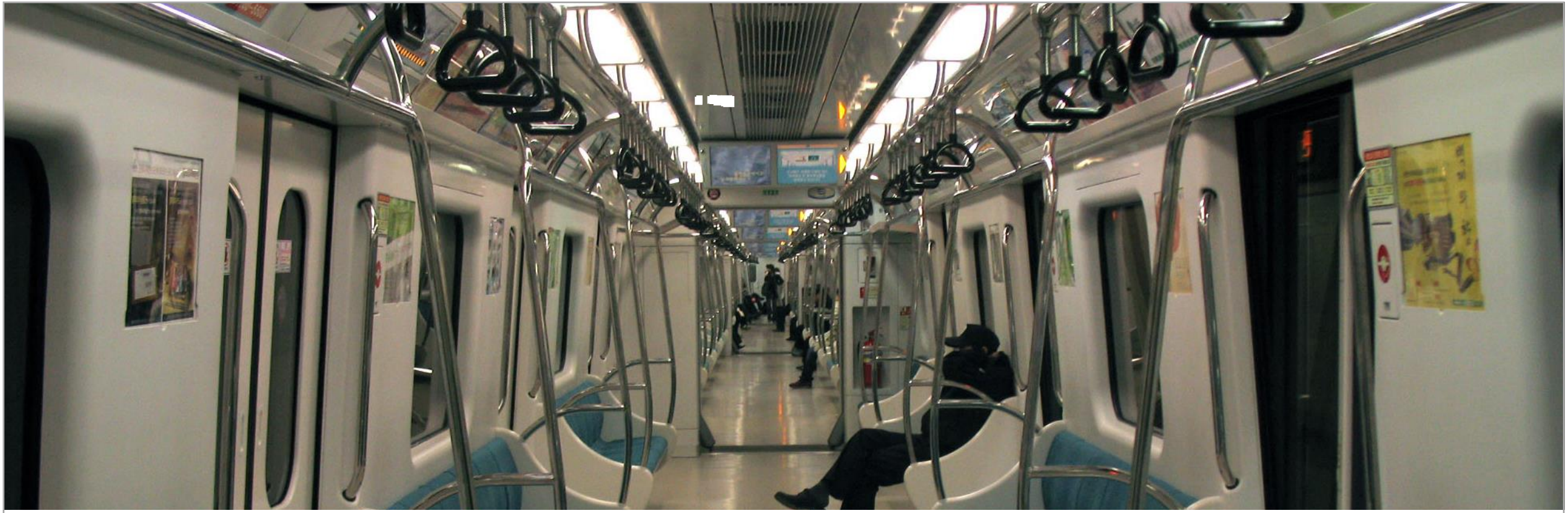


『2021 빅데이터 분석 및 아이디어 공모전』

공모명: 대전광역시 지하철 이용자 수 분석을 통한 지하철 활성화 방안 제안

팀명	Be Big KU (대표: 이가영)		지원부분	분석
소속	고려대학교			
연락처	휴대폰	010-9124-1510	E-mail	gyleeheyum3@naver.com

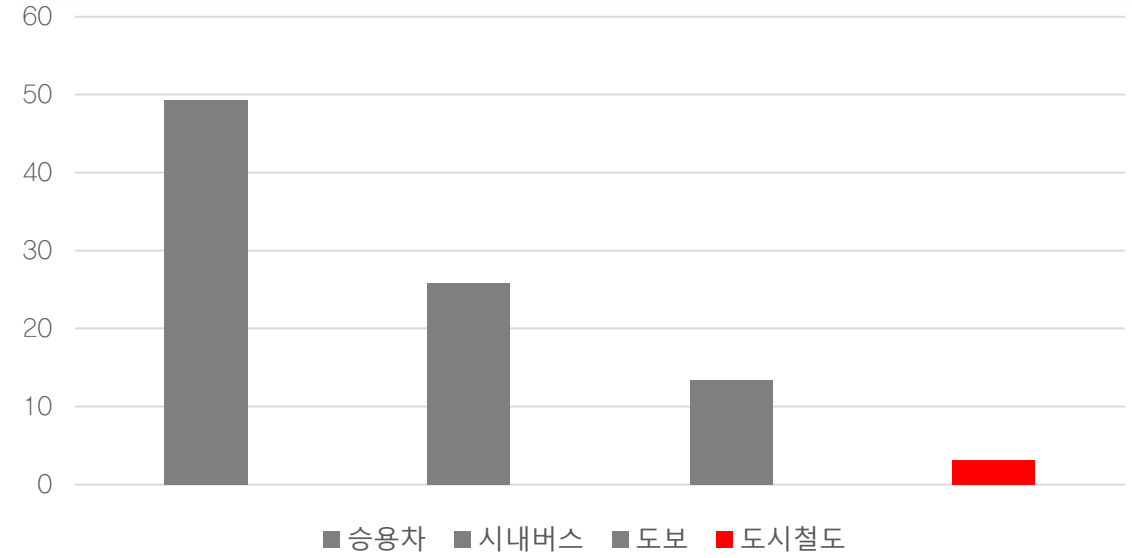
목차

01 상황 분석 및 문제점 도출

대전광역시 지하철은 그 이용률이 매우 낮고, 이로 인해 지속적인 당기순손실이 발생하는 등의 문제가 생김

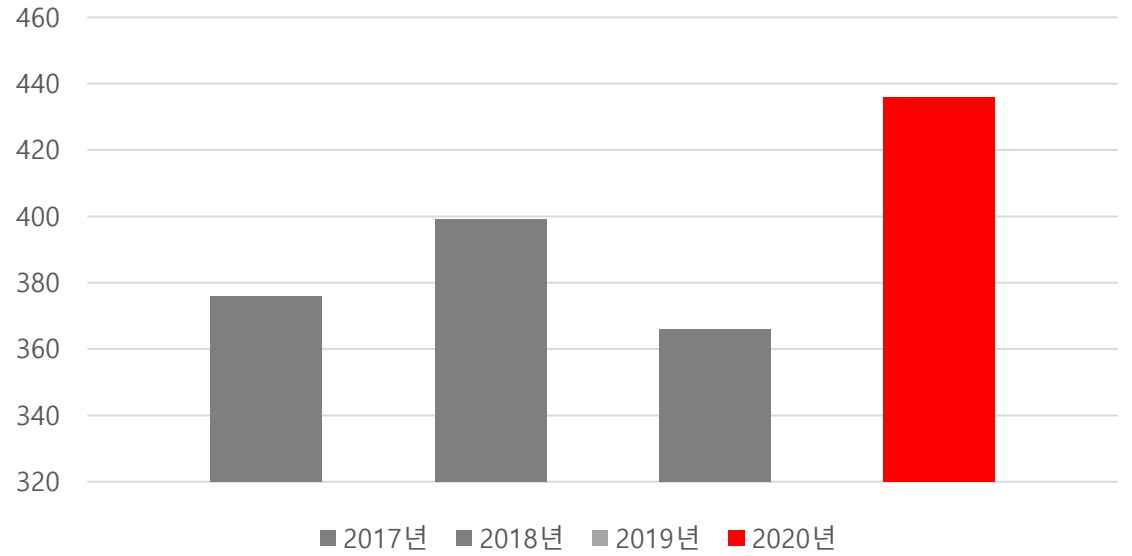
- 대전광역시에선 통근, 통학시 주로 승용차와 시내버스를 사용함, 지하철 이용률은 3.1%로 도보에도 미치지 못함
- 대전도시철도공사는 지난해(2020년) 436억원의 당기순손실을 기록함
- 사회적 거리두기와 재택근무 활성화 등 코로나19으로 인해 수송 인원이 큰 폭으로 줄어들며 상황이 악화됨

2019 대전광역시 통근 · 통학시 이용하는 교통수단 (단위:%)



출처: 2019 대전광역시 사회조사

연간 대전도시철도공사 당기순손실액 (단위: 억원)



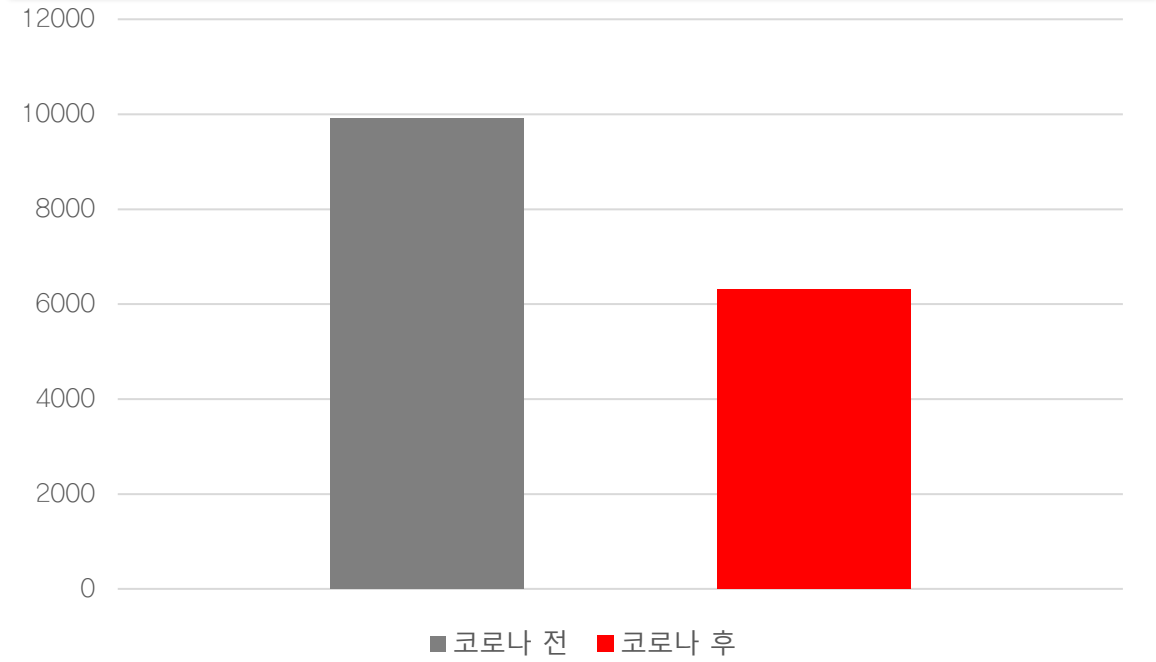
출처: 대전도시철도공사

대전시민들의 지하철 이용률 증대와 효율적인 운영을 위해 **지하철 개선방안 도출**이 필요함

02 데이터 설명

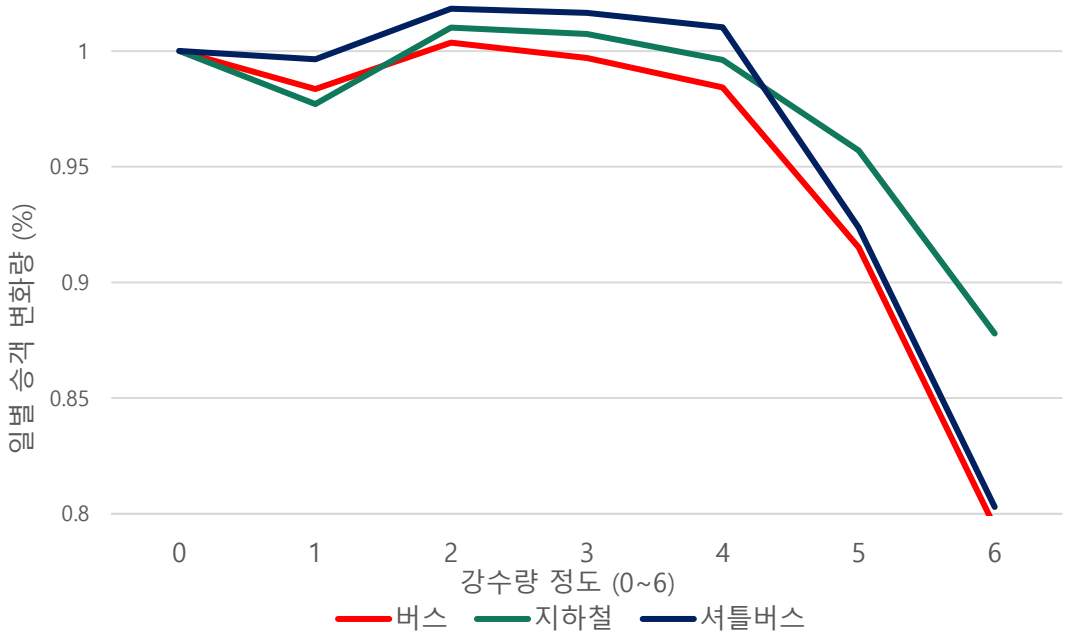
본격적인 분석에 앞서 EDA와 선행연구 조사를 통해 데이터를 파악함

대전광역시 지하철역 별 이용자 수 평균



출처: 대전시공공데이터 대전광역시도시철도공사 제공

강수량과 대중교통 이용량의 관계



출처: SUR 모형을 이용한 강수량과 대중교통 승객 수간 관계 분석

코로나 발병 전후의 지하철역 별 평균 이용자 수 차이

- 정확한 분석을 위해 코로나 전과 후의 이용자 수 예측 모델을 각각 만들어야 할 필요성 확인

강수량과 대중교통 이용량의 유효한 관계를 분석한 선행연구 존재

- 일우량, 장마기간, 적설량, 최저기온 및 최고기온 데이터 추가
- 학기 및 휴일 변수 추가

02 데이터 설명

분석에 사용될 전체 데이터는 공통변수와 코로나 발병여부를 기준으로 추가한 개별변수로 이루어짐

대전광역시 대중교통 통행목적 (기준:%)

출퇴근	여가	쇼핑	등하교	업무	학원	기타
29.8	22.1	16.0	14.4	13.3	2.7	1.7

출처: 국토교통부 2019, 대중교통현황조사 중 대전시 대중교통 통행목적

대중교통 통행목적 중 높은 비율을 보인 항목들을 분석에 반영

- 여가 → 영화관 등
- 쇼핑 → 백화점, 올리브영, 대형마트 등
- 등하교 → 학교 (초등학교, 중학교, 고등학교, 대학교) 등

이외에도 창의적인 분석을 위해 변수를 추가

- 유동인구와 발달상권의 간접지표 → 스타벅스, 프랜차이즈 술집 등
- 연령대별 대중교통 이용 비율 고려 → 노령화지수 등
- 대중교통 방문 예상장소 → 병원, 은행 등

데이터 집계방식

- 반경 500m 기준 (단, 병원은 1km, 버스정류장은 100m로 집계)
- 인구수, 노령화지수: 지하철역이 속한 행정동의 2020년 12월 기준
- 백화점, 영화관은 유무에 따라 이진변수로 추가

분석에 사용될 데이터 수집

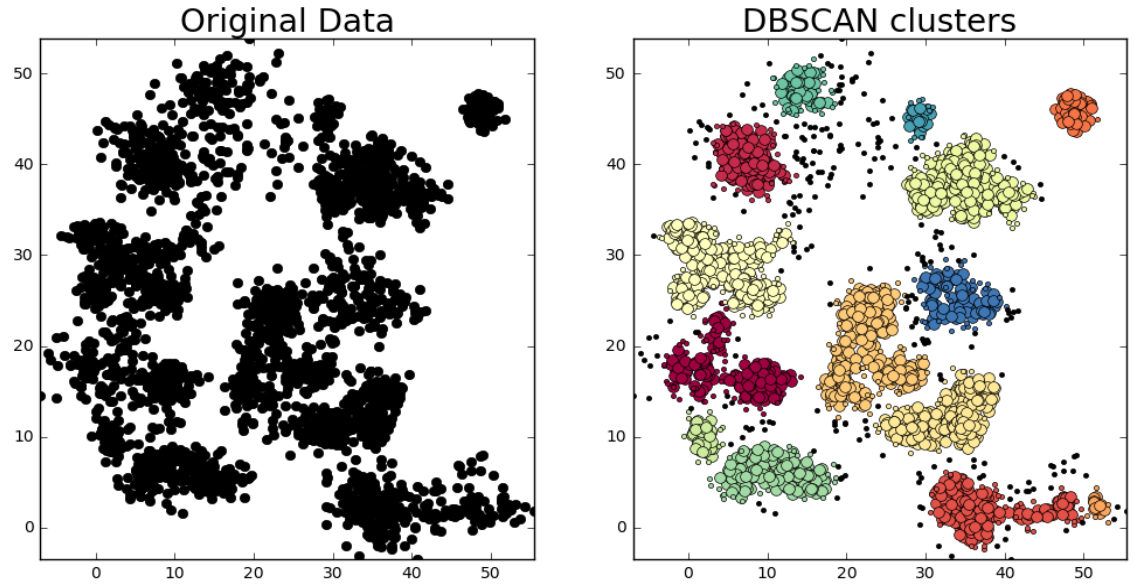
	코로나 이전	코로나 이후
공통	이용자 수, 출구, 버스정류장, 백화점, 학교, 노령화지수, 스타벅스, 올리브영, 영화관, 은행, 술집, 일우량, 휴일, 최저기온, 최고기온, 날짜, 병원, 대형마트, 적설량, 도서관, 버스터미널, 인구수 지하철역: 갈마, 갑천, 구암, 노은, 대동, 대전, 반석, 서대전네거리, 시청, 신흥, 오룡, 용문, 월드컵경기장, 월평, 유성온천, 정부청사, 중구청, 중앙로, 지족, 탄방, 판암, 현충원	
개별	학기여부, 장마기간	확진자 수
수집기간	2017년 1월 ~ 2019년 12월	2020년 2월 ~ 2021년 3월

03 군집분석 (HDBSCAN)

심도 있는 분석을 위해 군집 분석을 실시했고, 그 결과 총 4개의 군집으로 분류됨

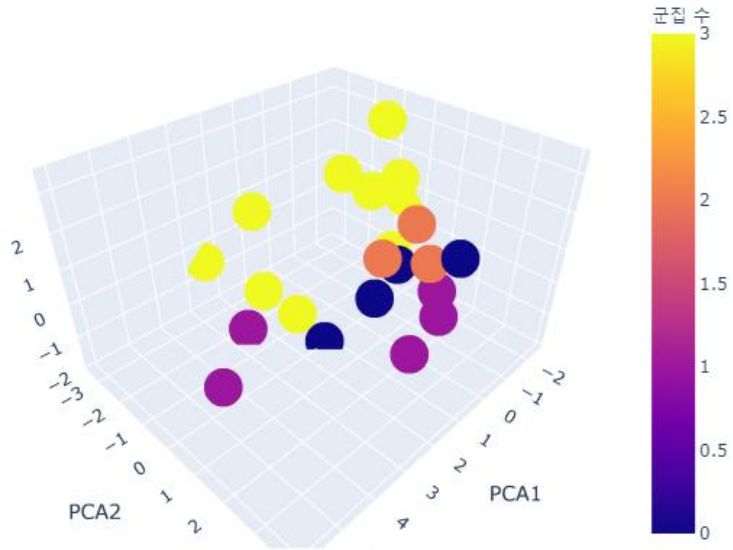
- 기본모델은 DBSCAN(Density based clustering of applications with the noise)은 데이터의 공간을 다루는 알고리즘
- 전체 공간에서 데이터가 가장 밀집된 영역을 찾고, 미리 설정한 거리에 따라 군집 형태가 갈라짐
- 분석 대상의 데이터는 변수가 20개 이상인 고차원 데이터이고 비선형임, DBSCAN은 이런 데이터 분류에 적합하지 않음
- HDBSCAN은 DBSCAN의 단점을 보완한 기법으로 고차원 데이터 분석에 용이하고, 특히 비선형 데이터 분포에서 그 성능이 좋음

DBSCAN Algorithm



출처: Github//chriswernst

대전광역시 지하철역 데이터 군집분석 결과



출처: 자체분석 결과

03 군집분석 (HDBSCAN)

군집별 특성을 보면 HDSCAN을 통한 군집 분류가 유의미한 것을 알 수 있음

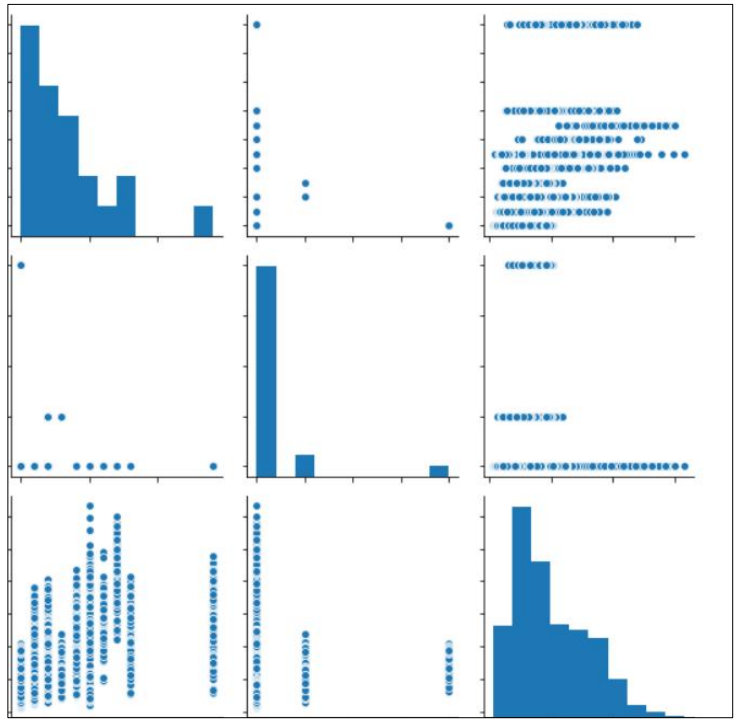
- 군집1: 노령화지수가 가장 낮았고(51.06) 실제로 코로나 이후 이용자 수의 감소가 가장 적었음
- 군집2: 발달상권의 지표인 스타벅스가 평균적으로 가장 많음(1.8개), 실제로 이용자 수가 가장 많았음
- 군집3: 버스정류장의 수(2.66개)가 가장 적고, 발달상권 지표인 스타벅스도 가장 적음 (0.33개), 실제로 이용자 수가 가장 적었음
- 군집4: 노령화지수가 가장 높았고(293.19) 실제로 코로나 이후 이용자 수의 감소가 가장 많았음

		변수 (평균값)				코로나 이전	코로나 이후	
	지하철역	인구수 (명)	노령화지수	버스정류장(개)	스타벅스 (개)	이용자 수 (명)	이용자 수 (명)	증감율
군집 1	노은, 반석, 월드컵경기장, 지족	41852	**51.06**	3.25	0.75	8251	5555	**-32.67 %**
군집 2	구암, 시청, 유성온천, 정부청사, 현충원	37675	98.96	4.6	**1.8**	**11529**	7467	-35.23 %
군집 3	갈마, 갑천, 월평	23807	128.72	**2.66**	**0.33**	**5271**	**3378**	-35.91 %
군집 4	대동, 대전, 서대전네거리, 신흥, 오룡, 용문, 중구청, 중앙로, 탄방, 판암	7532	**293.19**	4.4	0.9	11177	6908	**-38.19 %**

04 요인분석을 위한 회귀분석

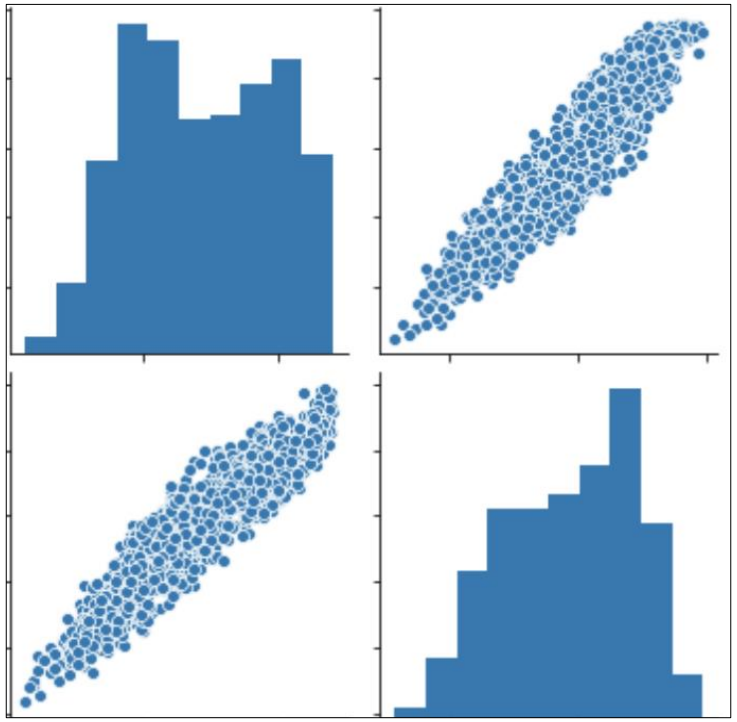
지하철 사용자 수에 유효한 영향을 미치는 변수를 알아보기 위해 회귀분석을 실시함
사용한 회귀분석 모델: OLS, Ridge, Lasso, Elastic Net

데이터 전처리



선형성 및 정규성 확인

- 정규성을 띄지 않는 데이터의 로그 변환



다중공선성 확인 및 변수 선택

- VIF 및 시각화를 통해 다중 공선성 확인
- 날짜, 병원, 대형마트, 적설량, 도서관, 버스터미널, 인구수, 장마기간 변수 삭제

모델 선택을 위한 데이터 적합

- 변수 선택이 끝난 뒤 4개의 모델(OLS, Ridge, Lasso, Elastic Net)에 데이터 적합
- Ridge 회귀, Lasso 회귀, Elastic Net 적합 결과 해당 모델들은 정보의 손실이 많이 이루어짐

모델 선택

- 분석에 필요한 데이터 보존을 위해 기본적인 OLS 회귀 모델을 선택함

04 요인분석을 위한 회귀분석

회귀분석 결과 실제로 시민들의 삶과 밀접한 변수들이 유의미한 결과를 보여줌

회귀분석 결과

	OSL 회귀계수 값
백화점	0.0830
술집	0.0577
버스정류장	0.0448
스타벅스	0.0405
출구	0.0228
학기여부	0.0019
휴일	−0.0145

회귀분석 결과 해석

- 회귀분석 결과 유효한 변수로 좌측 7개의 변수가 선정됨
- 일별 특징 변수로 학기에 지하철 이용 빈도가 많으며 휴일인 경우에는 지하철 이용자 수가 감소하는 점을 알 수 있음
- 지하철역별 특징 변수로 백화점, 술집, 버스정류장, 스타벅스 변수가 유효한 것으로 나타남. 해당 지표가 많을수록 지하철 이용자 수가 증가함을 알 수 있음
- 이용자 수는 지하철 출구와 양의 상관관계가 있음, 다만 지하철역의 출구는 추가적인 설치나 조정이 불가능하다고 판단해 해석에서 제외함

기대효과

- 회귀분석 결과 실제로 시민들의 삶과 밀접한 변수(백화점, 술집, 버스정류장, 스타벅스 등)가 유의미한 결과를 보여줌
- 이후 시민생활기반의 데이터를 추가적으로 활용하면 분석의 예측력을 높일 수 있을 것으로 예상됨

05 예측모델: Random Forest, LGBM, XGBoost

이상치와 노이즈에 큰 영향을 받지 않고, 모델의 해석력이 높은 트리(Tree)기반의 모델을 선정함

모델설명	Random Forest	LGBM	XGBoost
	➢ 예측의 변동성이 작음 ➢ 과적합(Overfitting)을 방지할 수 있음 ➢ 변수의 중요도를 파악할 수 있음	➢ Gradient Boosting 방법 중 한가지 ➢ 빠르고 높은 효율의 모델 Training ➢ 다른 Boosting 알고리즘보다 정확도가 높음 ➢ 큰 데이터도 빠르게 처리할 수 있음	➢ Gradient Boosting 방법 중 한가지 ➢ Decision Tree 기반 앙상블 학습 중 Boosting 기반 ➢ 과적합(Overfitting)을 방지할 수 있음 ➢ 비교적 높은 성능을 보여줌

모델 구성 방법

➢ 각 지하철역을 설명변수로 넣은 후 원-핫 인코딩(One-Hot Encoding) 처리를 함
➢ 지하철역 이용자 수를 예측할 때, 각 지하철 역이 갖는 가중치를 통해 지하철역의 영향을 효율적으로 파악하기 위한 목적

06 모델별 성능비교

이상치에 강건하고, Under Estimation에 큰 패널티를 부여하는 장점이 있는 RMLSE를 평가지표로 선택함
값이 작을수록 예측 값과 실제 값 사이의 차이가 적다는 뜻으로, 회귀성능이 좋다고 해석할 수 있음

	코로나 전					코로나 후				
	전체	군집1	군집2	군집3	군집4	전체	군집1	군집2	군집3	군집4
Random Forest	0.1571	0.1678	0.1523	0.135	0.1489	0.2458	0.2463	0.2649	0.2285	0.2389
LGBM	0.1583	0.1562	0.1714	0.1283	0.1486	0.2003	0.2013	0.2233	0.1929	0.1907
XGBoost	0.1530	0.1522	0.1694	0.1274	0.1444	0.2068	0.2197	0.2473	0.2067	0.2222

전반적으로 코로나 이전에는 XGBoost가 좋은 성능을 보여줬고, 코로나 이후에서는 모든 분석에서 LGBM이 가장 좋은 성능을 보여줬음
10개의 분석별로 가장 좋은 성능을 보인 모델을 별도로 표시하여 향후 논의에 활용하였음

군집별 지하철역

- 군집1: 노은, 반석, 월드컵경기장, 지족
- 군집2: 구암, 시청, 유성온천, 정부청사, 현충원
- 군집3: 갈마, 감천, 월평
- 군집4: 대동, 대전, 서대전네거리, 신흥, 오룡, 용문, 중구청, 중앙로, 탄방, 판암

07 결과해석 및 시사점 도출

코로나 전/후 전체 모델의 Feature Importance 분석을 통해 지하철 이용자 수 예측에 유의미한 요인을 분석함

코로나 전 (XGBoost)		코로나 후 (LGBM)	
Feature	Importance	Feature	Importance
스타벅스	0.1056	확진자 수	0.2340
버스정류장	0.0965	최저기온	0.2069
영화관	0.0873	최고기온	0.1353
백화점	0.0773	출구	0.0774
출구	0.0749	휴일	0.0723
대전	0.0561	버스정류장	0.0537
은행	0.0426	일우량	0.0490

결과해석

코로나 전 (XGBoost)

- 스타벅스 변수가 가장 높은 요인으로 나옴
- 또한 영화관, 백화점과 같은 변수들도 이용자 수 예측에 유효하게 나타남

코로나 후 (LGBM)

- 코로나 확진자 수가 지하철 이용자 수에 가장 큰 영향을 줌
- 기온 및 일우량과 같은 날씨 요인도 이용자수에 유효한 영향을 주고 있음

시사점 도출

- 코로나 전/후 모델 모두 지하철역 주변 버스정류장의 개수가 지하철 이용자 수 예측에 큰 영향을 주고 있음
- 분석 전 추가했던 시민생활과 밀접한 변수들(스타벅스, 영화관)이 실제 지하철 이용자 수를 예측하는데 유효한 역할을 함을 파악함

07 결과해석 및 시사점 도출

코로나 전/후 모델의 군집별 Feature Importance 분석을 통해 지하철 이용자 수 예측에 유의미한 요인을 분석함

코로나 전 군집별 분석								코로나 후 군집별 분석							
XGBoost		Random Forest		XGBoost				LGBM							
군집1		군집2		군집3		군집4		군집1		군집2		군집3		군집4	
반석	0.2393	은행	0.1669	학교	0.1991	영화관	0.2188	최저기온	0.2883	최저기온	0.2853	최저기온	0.3151	최저기온	0.2700
술집	0.1716	휴일	0.1425	갑천	0.1760	대전	0.1857	확진자수	0.2648	확진자수	0.2463	최고기온	0.2439	확진자수	0.2354
휴일	0.1289	학교	0.1047	버스 정류장	0.1702	올리브영	0.0935	최고기온	0.2120	최고기온	0.2168	확진자수	0.2346	최고기온	0.1698
올리브영	0.1132	스타벅스	0.0983	출구	0.1696	노령화 지수	0.0867	일우량	0.0687	일우량	0.0689	일우량	0.0735	휴일	0.0765
노은	0.097001	출구	0.0794	은행	0.1334	출구	0.0519	휴일	0.0547	출구	0.0559	휴일	0.0534	일우량	0.0703
영화관	0.0750	현충원	0.0622	올리브영	0.0483	은행	0.0454	출구	0.0354	휴일	0.0503	학교	0.0360	노령화 지수	0.0460
은행	0.053	영화관	0.0499	노령화 지수	0.0270	신흥	0.0405	버스 정류장	0.0337	버스 정류장	0.0383	출구	0.0273	버스 정류장	0.0392

결과해석 및 시사점 도출

- 도심과 신도시의 특성이 나타나는 군집2에서는 은행과, 스타벅스가 유효한 지표로 분석되고 있음
- 중구와 동구의 지하철역들이 묶인 군집4에서는 노령화지수가 코로나 전/후 모두 영향력 있는 요인으로 분석됨, 따라서 고령 인구에 적합한 지하철 서비스를 제공하는 것이 이용하는 것이 이용률을 증가시키고 효율적인 운용에 도움이 될 것으로 예상됨
- 유성구의 주거지구로 묶인 군집1은 술집, 올리브영, 영화관, 은행과 같은 생활 및 여가 시설 요인들이 유효하게 나타남
- 코로나 후 LGBM으로 분석한 결과 대전시 차원에서 통제가능한 변수는 버스정류장, 학교 등 소수에 그침

08 기대효과

08 기대효과

1. 대전광역시 스마트도시 비전: 데이터 시티 대전 (Data City Daejeon)에 맞춰 시민의 삶을 긍정적으로 변화시킴

기대효과

- **〈대전광역시 스마트도시기본계획 보고서〉**에 따르면, 대전시는 **교통량 데이터, 대중교통 환승데이터**를 활용하여 대중교통 **수송분담률 증가, 교통량 저감** 등의 성과창출을 기대하고 있음
- 본 분석에서는 대전시가 가진 공공데이터를 확장시켜 지하철 이용자 수와 더불어 **일별 특징 데이터(일우량, 기온, 휴일 등)**, 지하철역별 **특징 데이터(노령화지수, 버스정류장, 백화점 개수 등)**도 함께 고려함
- 이전까지 **대전시가 고려하지 않았던 실제 시민들의 삶과 밀접한 관련이 있는 새로운 지표들을 사용하여** 기존 방법과는 차별화되고 예측도가 더욱 높은 방안을 제시함
- 이러한 새로운 지표들을 참고하여 이후 트램역 선정 또는 버스정류장 노선 등 **신규 대중교통 편성에 이용할** 수 있을 것이라 기대됨

도시목표	Change: 시민 삶의 변화			
데이터 목표	데이터 생산: 보다 필요한 데이터 생산			
추진전략	모두에게 열려있는 행정	**편리하게 이용하는 교통**	신속하게 대응하는 안전	쾌적하게 관리되는 환경
관련분야	행정 서비스	**교통 서비스**	안전 서비스	환경 서비스
서비스	공간공유	**대중교통연계환승**	지능형방범	미세먼지관제
	와이파이 공유	**타슈 및 전기자전거 공유**	전기화재예방	스마트관망
	온통대전	**주차공유**	무인드론 안전망	음식물쓰레기제로
	빅데이터 플랫폼	**교통흐름 최적화**	재난예경보	에너지다이어트 불법쓰레기예방
	마을단위스마트포털	**도로 인프라 유지관리**	안심귀가	쓰레기재활용 교육-체험
	타임뱅크	**교통약자 승차지원**	E-call	시설물통합관리 드론기반빈집관리
도시관점 KPI	리빙랩 건수	**대중교통 수송분담률**	검거율 상승	쓰레기저감
	공간공유 활용건수	**교통량 저감**	골든타임내 대응건수	미세먼지 저감량
	전자화폐 이용액	**공유주차장 수**	재난예경보 시간단축	에너지 저감량
데이터 및 스마트관점 KPI	시민의견 비정형데이터	**교통량 데이터**	안전관리 영상데이터	에너지 데이터
	공유공간 데이터	**대중교통 환승데이터**	유동인구 데이터	미세먼지 데이터

출처: 대전광역시 스마트도시기본계획 요약보고서

08 기대효과

2. 이용자수 예측 모델을 통해 지하철 시설물 관리 및 방역 활동을 수월하게 하고, 신설될 트램 계획 수립을 도움

기대효과

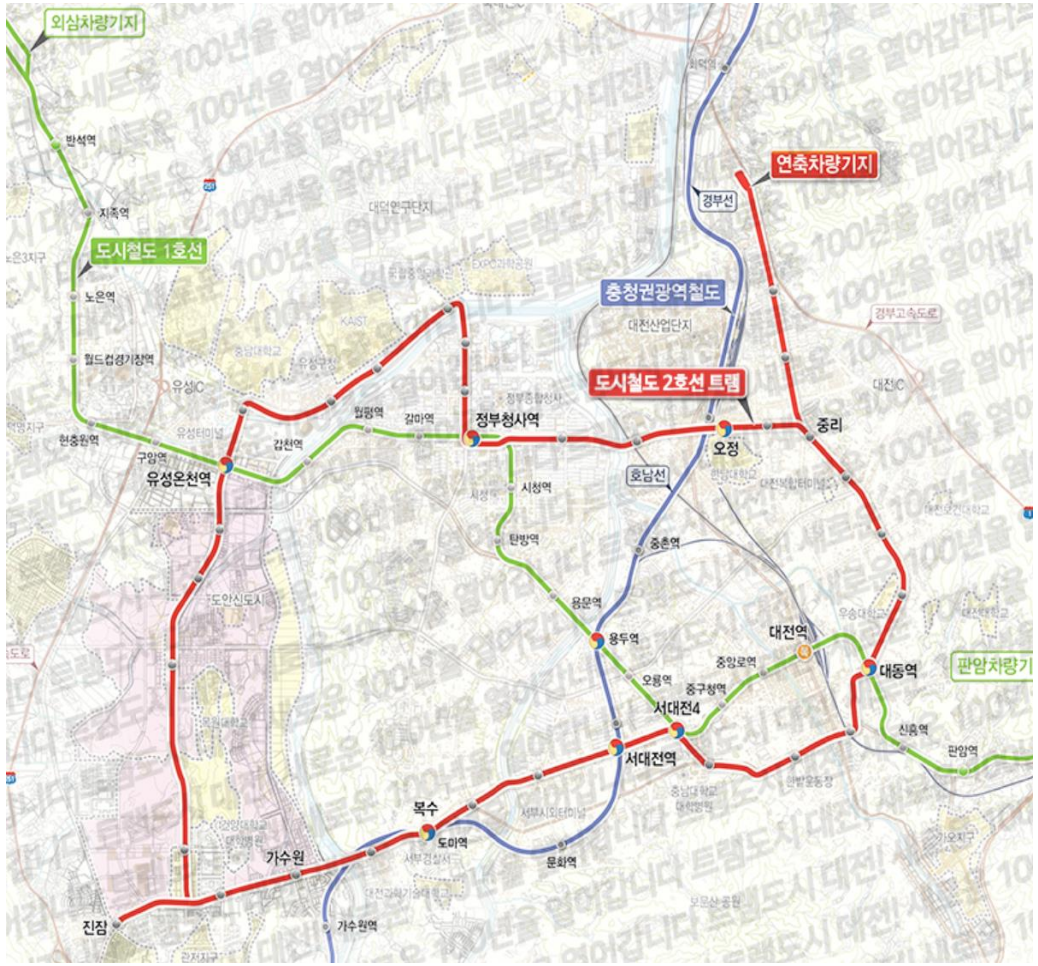
1) 지하철 시설물 관리 및 코로나 방역

- 본 분석에 의하면 **기온과 일우량의 변수 중요도가** 높게 나타남, 이를 통해 시민들의 지하철 이용에 날씨가 큰 영향을 끼치는 것을 알 수 있었음
- 따라서 일기예보에 따른 이용자수 예측모델을 통해 지하철 배치 등 효율적인 운행조절이 가능함
- 또한 코로나 이후의 모델을 활용하여 일별 이용자수가 많을 것으로 예측되는 날에 **지하철내 방역을 강화**하거나, **시설물 관리 인력을 추가로 보충**하는 등의 효율적인 지하철 운영이 가능함

2) 신설될 트램과의 연계성

- 트램 노선계획을 보면 지하철과 트램이 공유하는 역이 있음, 따라서 후에 개통될 **트램의 배차간격 계획 수립에 모델이 활용**될 수 있음
- 또한 본 분석에서는 **대형마트, 백화점 등의 요소를 반영**했기 때문에 추후 트램역을 신설할 때 고려할 수 있음

대전시 트램 노선소개

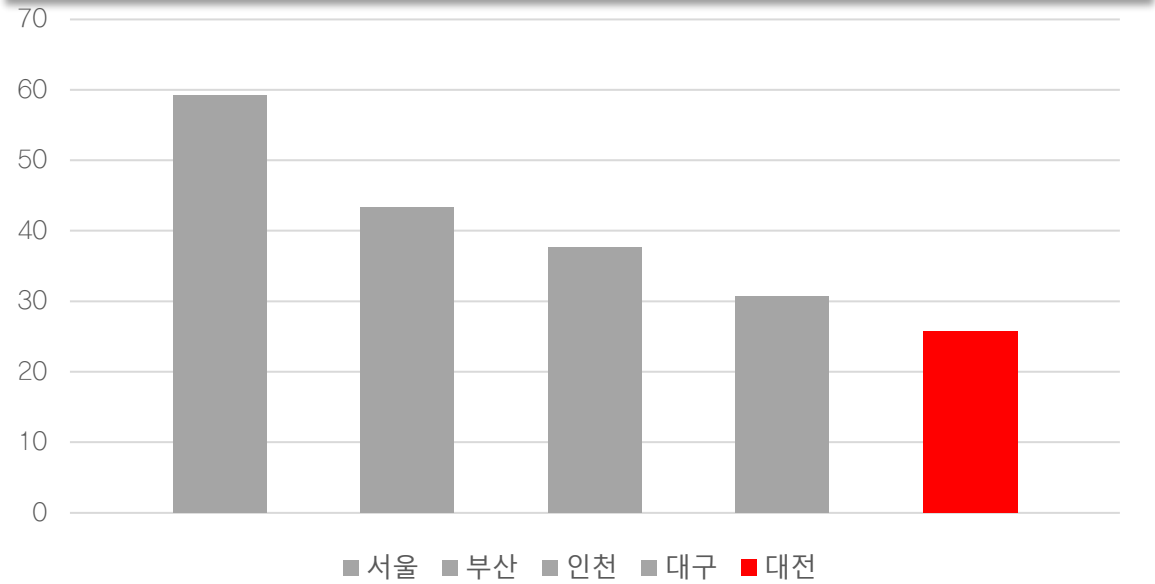


출처: 대전광역시 스마트도시기본계획 요약보고서

09 기대효과

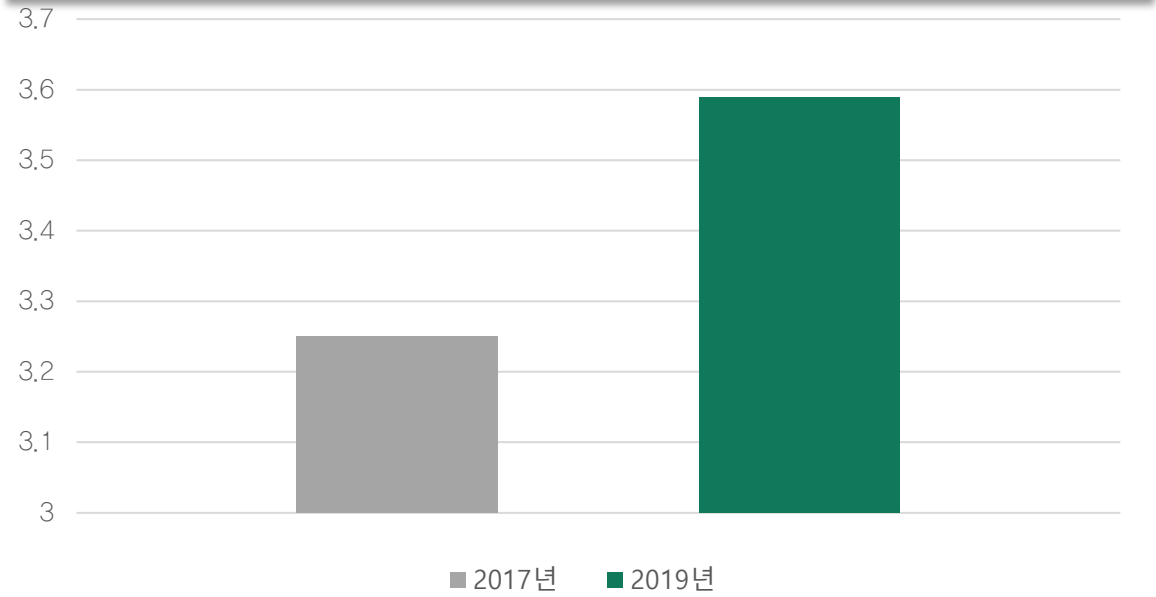
3. 버스정류장과 버스노선의 효율적 증설을 통해 지하철을 활성화시킴

2018년 통행량 대비 대중교통 분담률 (단위: %)



출처: 2018년 국토교통위원회 국정감사 〈대전시 통행량 대비 대중교통 분담률〉

대전광역시 지하철 만족도 변화



출처: 2019년 대전광역시 시회조사

기대효과

- 대전시의 대중교통 분담률은 2018년 기준 25.7%로 다른 도시에 비해 상당히 낮음, 이는 교통지체의 증가로 연결될 수 있음
- 사실 지하철에 대한 만족도는 3.59점으로 높으나, 접근성이 떨어져 시민들이 활발하게 사용하지 않았음
- 모델 분석 결과에 따르면 **지하철 이용자수와 지하철역 근방 100m 이내의 버스정류장 수 사이에는 상당한 연관성**이 있었음
- 따라서 지하철역 인근 버스정류장의 증설이나 버스 노선을 증설하면 **지하철 활성화가 가능**할 것으로 예상됨

09 후속 연구방향 제안

본 분석에서 미처 다루지 못한 점들은 다음과 같다, 이를 보완하면 대전광역시에 보다 도움될 것으로 예상함

추가적인 데이터수집을 통한 예측모델 정교화

- 시간대별 이용자수, 시간대별 날씨 관련 데이터를 수집한다면 시간대별 이용자수도 예측도 가능할 것으로 기대됨
- 연령대별로 상이한 지하철 이용률을 반영하고자 했지만, 데이터 수집의 한계로 간접 지표인 노령화지수를 사용함. 향후 대전시 교통카드의 일별 데이터를 이용한다면 지하철 이용자수의 인구 구성비를 반영할 수 있을 것으로 기대됨

버스정류장 및 대중교통 데이터 활용

- 대전시 지하철은 다른 지자체 지하철에 비해 비교적 규모가 작아 다각적인 분석 및 예측모델 성능 확인에 어려움이 있었음
- 대전도시철도 2호선(트램)이 개통된 이후, 본 연구의 범위를 트램까지 포함시켜 환승역 이용자 수, 효율적인 환승역 배치 등의 측면으로도 분석을 진행할 수 있을 것으로 기대됨

트램노선을 활용한 다각적인 분석

- 본 분석결과는 단일 지하철 호선만을 고려한 분석임
- 대전광역시의 각 버스정류장 및 버스노선별 이용자 수 데이터와 연계해 버스노선 증설, 버스–지하철 환승 최적화 등에 활용한다면 데이터 시티 대전의 일환으로 교통흐름 최적화에 기여할 수 있을 것으로 예상됨

참고문헌 및 데이터 출처

참고문헌

김흥순, 김영덕, 원윤재, 신은하. (2020). 스타벅스 입지의 공간적 효과에 관한 연구 – 2016년 개점된 서울시내 점포를 대상으로 –. , 54(1), 77–89.
신강원, 최기주. (2014). SUR 모형을 이용한 강수량과 대중교통 승객 수간 관계 분석. 대한교통학회지, 32(2), 83–92.
조수진, 김보경, 김나현, 송종우. (2019). 데이터마이닝 기법을 이용한 서울시 지하철역 승차인원 예측. 응용통계연구, 32(1), 111–128.

데이터 출처

https://www.daejeon.go.kr/sta/StaStatisticsFldView.do?ntatcSeq=1359379876&menuSeq=&colmn1Cont=C0201&colmn2Cont=C020101&boardId=normal_0009&pageIndex=1#
https://data.kma.go.kr/climate/StatisticsDivision/selectStatisticsDivision.do?pgmNo=158
https://data.kma.go.kr/stcs/grnd/grndRnDay.do?pgmNo=156
http://www.kric.go.kr/jsp/industry/rss/citystapassList.jsp?q_org_cd=A010010025&q_fdate=2020
https://data.go.kr/data/15043917/fileData.do

E.O.D

대전광역시
DAEJEON METROPOLITAN CITY