지하철 혼잡도 예측 및 사용자 분산 서비스

2조- 트랜짓 인사이츠 (Transit Insights)

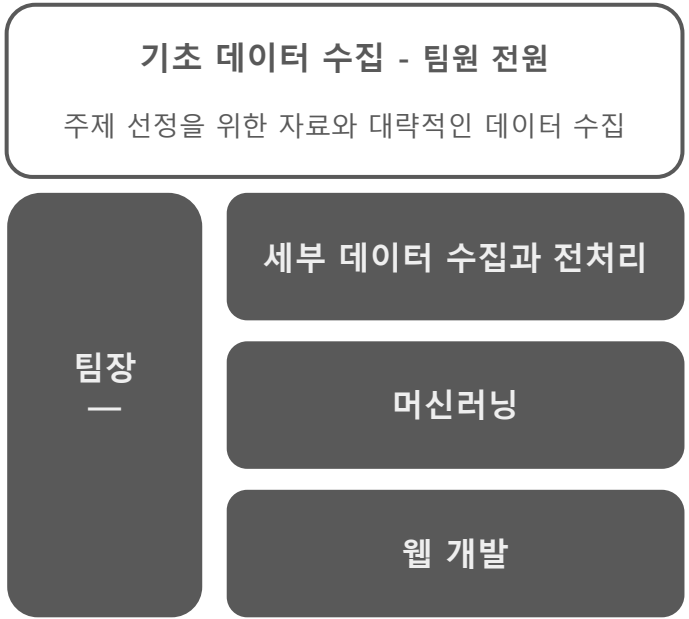
multicampus

데이터 분석 & 엔지니어 취업캠프 28회차

목차

1. 프로젝트 개요

역할	세부 역할	개인 블로그 주소
팀장	프로젝트 총괄, 전반작업	https://mizima-data.tistory.com/
팀원 1	웹개발	https://kimec995.github.io/
팀원 2	데이터 수집과 전처리	https://data-analytics-nayoonee.tistory.com/
팀원 3	데이터 수집과 전처리	https://datasoling.tistory.com/
팀원 4	데이터 수집과 전처리	https://sohyeon-choi.tistory.com/
팀원 5	머신러닝	https://sim-ds.tistory.com/



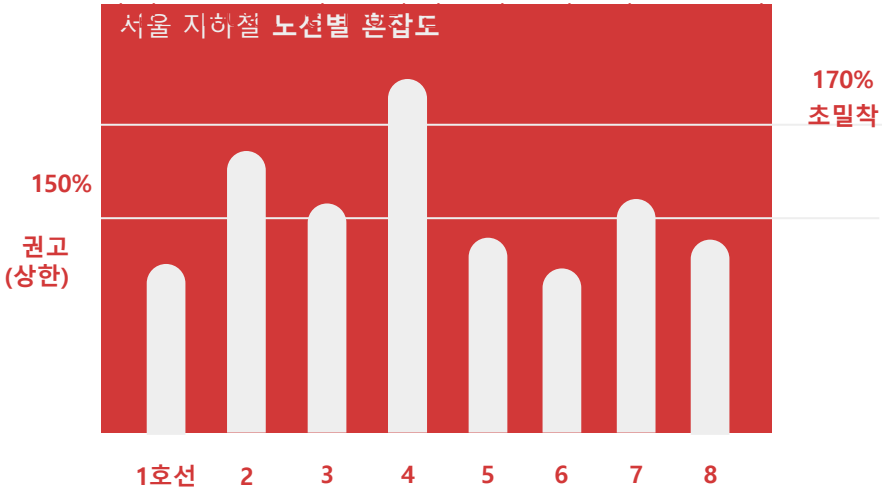
1. 프로젝트 개요

일	월	화	수	목	금	토
					09/1	2
					웹: 구성(논리)	
					기획	
3	4	5	6	7	8	9
기획				데이터 전처리		
웹: 구성(논리) + 대시보드(논리)						
데이터 수집						
10	11	12	13	14	15	16
데이터 전처리					분석 / 모델평가	
웹: 수정						
EDA / 시각화						
17	18	19	20	21	22	23
분석 / 모델평가						
웹: CSS, 배포, 머신러닝 적용, 대시보드 수정						
EDA / 시각화 / 대시보드 작성						
24	25	26	27	28	29	30
웹: 마무리		발표!				
전체 마무리						

(2022년 연간 승하차 인원 수)

서울 지하철 이용객 28억명시대,

지하철 내 혼잡 및 안전성 문제 심각화



기존 서비스와의 차별성

티맵 대중교통

- 일부 열차의 정보만 서비스 가능
- 현재 서비스 중단

2호선 일부열차

- 최신 열차 내부에서만 확인 가능
- 서비스 제공 시점의 정보만 제공

- 서울교통공사 1호선~8호선 **노선 정보 제공**
- **사용자의 조건에 따른 혼잡도** 정보 제공
- 지하철 탑승 전 **미리 혼잡도 예측 가능**

기대효과

- 안전성 향상
- 편의성 향상
- 정책 대책
- 편익산출

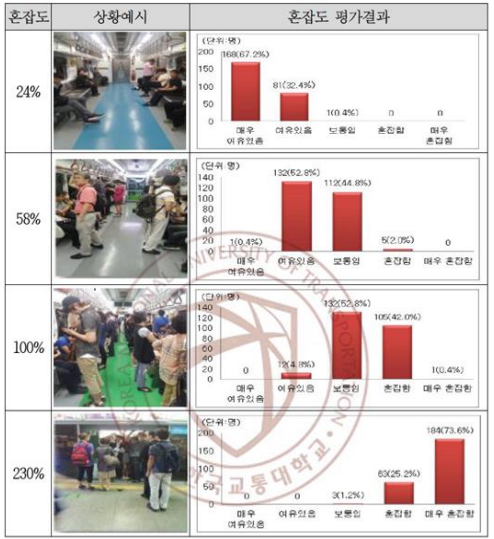
혼잡도 100%

지하철 인원 밀집도를 의미

1 량 당 정원 160명

국도교통부 예규 '도시철도의 건설과 지원에 관한 기준'

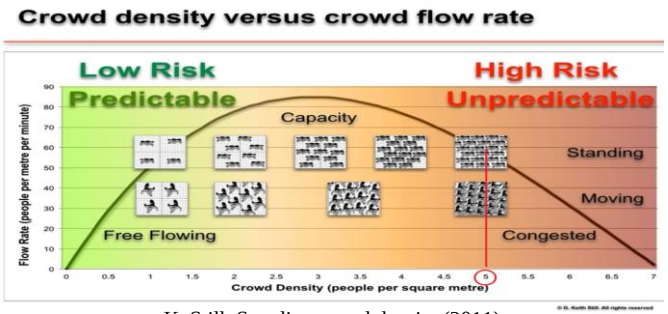
부록참고(1): 기준



출처: 경기개발연구원, 빈미영, 교통복지를 고려한 철도사업의 타당성 평가방안 연구, 2012

부록참고(2): 계산식

좌석 정원	차량 입석 면적 평균	군중 밀집 임계점	위험 혼잡도 (주의단계)
54 명	32.56 m^2	5 명/m^2	135.5 %



K. Still, Standing crowd density (2011)

데이터 탐색

데이터 수집 | 전처리 / EDA

데이터 수집 개요 부록참고

분류	생성 변수	데이터 명	출처	비고
승객	*시간별 혼잡도*	서울교통공사 역별 시간대별 혼잡도	*https://data.seoul.go.kr/dataList/OA-12928/F/1/datasetView.do*	*서울 열린데이터 OpenAPI*
승객	*시간별 승/하차 인구*	서울교통공사 연도별 일별 역별 시간대별 승하차인원(1_8호선)	*https://data.seoul.go.kr/dataList/OA-12252/S/1/datasetView.do*	*서울 열린데이터 OpenAPI*
승객	*월별 승/하차 인구*	서울시 지하철 호선별 역별 시간대별 지하철 인원 정보	*https://data.seoul.go.kr/dataList/OA-12252/S/1/datasetView.do*	*서울 열린데이터 OpenAPI*
승객	*월별 승/하차 인구*	서울교통공사 월별 승하차인원	*https://www.data.go.kr/data/15044249/fileData.do*	*공공데이터 openAPI*
기상	*일일 기상 조건*	연도별 일일 체감온도, 기온, 강수량	*기상청*	*csv 파일*
시간	*지하철 시간 데이터*	서울시 지하철 실시간 도착 정보	*https://data.seoul.go.kr/dataList/OA-12764/F/1/datasetView.do*	*서울 열린데이터 OpenAPI*
시간	*21~23 대한민국 공휴일*	공휴일 21년~23년	*https://www.data.go.kr/tcs/dss/selectApiDataDetailView.do?publicDataPk=15012690*	*공공데이터 openAPI*

데이터 전처리 개요

Raw Data (기상)

· **2021~2023년 기상조건**

Raw Data (승객)

· **역+시간대 혼잡도**
(2011~2022년, 홀수년도)
· **호선+역+시간대 승하차인원**
(2015~2019, 2022년)
· **연도+일+역+시간대 승하차인원**
(2022년)

Raw Data (시간)

· **2021~2023년 공휴일**

결측치 처리

· 운행하지 않는 시간대의 인원: 0으로 처리

데이터 필터링

· 인천 7호선 데이터 삭제
· 연도별 상이한 역명 : 통일(23년)
· 날씨 데이터 삭제

시간 데이터 처리

· 공휴일은 일요일로 처리

Preprocessing Data

연도별 혼잡도 데이터

승차인원 데이터 파악

데이터 가공

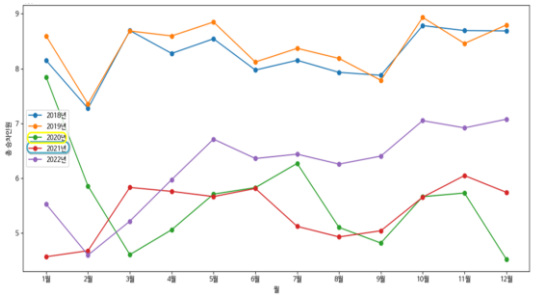
최종 혼잡도

데이터 선정 및 기간 설정

◦ **코로나 관련 정부 지침(사회적 거리두기)**으로 2020~2021년승차인원 패턴이 이전과 크게 동떨어진 것으로 판단

⇒ 역+시간대 혼잡도 데이터의 기간 중 **2020년, 2021년은 전처리 대상에서 제외**

*부록 참고: 사회적 거리두기 및 정부 지침



연도별 1호선 승차인원 추이 라인그래프

결측치 처리

◦ **운행하지 않는 시간대**의 혼잡도 NaN

⇒ 모두 **0**으로 처리

```
df_13.isna().sum()
```

요일구분	0
호선	0
역명	0
상하구분	0
05시~06시	0
06시~07시	0
07시~08시	0
08시~09시	0
09시~10시	0
10시~11시	0
11시~12시	0
12시~13시	0
13시~14시	0
14시~15시	0
15시~16시	0
16시~17시	0
17시~18시	0
18시~19시	0
19시~20시	0
20시~21시	0
21시~22시	0
22시~23시	0
23시~24시	0
24시~01시	113

dtype: int64

```
df_13.isna().sum()
```

요일구분	0
호선	0
역명	0
상하구분	0
05시~06시	0
06시~07시	0
07시~08시	0
08시~09시	0
09시~10시	0
10시~11시	0
11시~12시	0
12시~13시	0
13시~14시	0
14시~15시	0
15시~16시	0
16시~17시	0
17시~18시	0
18시~19시	0
19시~20시	0
20시~21시	0
21시~22시	0
22시~23시	0
23시~24시	0
24시~01시	0

dtype: int64

(좌) 결측치 처리 전 (우) 결측치 처리 후

데이터 필터링

◦ **관할 공사 기준 데이터**

서울교통공사 관할이 아닌 역이 있어 최종 혼잡도 산출에 오류 발생
따라서 **서울교통공사 관할 호선과 역으로 통일**

◦ **모든 역 명 최신화**

일부 지하철 역명의 변경으로 인해 최종 혼잡도 산출에 오류 발생 (역 표기 변경 포함)

ex. 성신여대 → 성신여대입구,
신천역 → 잠실새내역

◦ **오타 수정**

2015년 혼잡도 역명 중

'무악제' → '무악재' 로 변경

혼잡도 계산 방식

연도별 혼잡도 데이터

·평일, 토요일, 일요일로 구분
→ 요일별 데이터가 제공되지 않음

·시간대별 혼잡도는 연 평균 혼잡도
→ 월별 데이터가 제공되지 않음

승차인원 데이터 파악

·충무로역(3, 4호선), 연신내역(3, 6호선), 까치산역(2, 5호선)은 개찰구를 공동으로 사용
·6호선 신내역은 모든 시간대 인원이 0으로 되어 있음
·6호선 순환선은 단방향만 존재

데이터 가공

·3호선 충무로역 승객수
= 총 승객수 × $\frac{\text{(3호선 평일혼잡도 합)}}{\text{(3, 4호선 평일혼잡도합)}}$
·신내역(6호선 마지막역)의 혼잡도는 직전역인 봉화산역의 요일 비율값으로 대체

최종 혼잡도

■ 월별 비율

$$\#\text{월의 비율} = \frac{\#\text{월 평균 승차인원수}}{\text{월 전체 합 평균} / \ 12}$$

■ 요일별 비율

$$\#\text{요일의 비율} = \frac{\#\text{요일 평균 승차인원수}}{\text{요일 전체 합 평균} / \ 7}$$

■ 최종 혼잡도

연평균혼잡도 × 월별비율 × 요일별비율

순번	데이터 명	사용한 데이터	비고
1	**역별 월별 비율**	서울시 지하철 호선별 역별 시간대별 승하차 인원 정보(2015~2019,22)	-
2	**역별 요일별 비율**	서울교통공사 2022년 일별 역별 시간대별 승하차인원(1~8호선)	-
3	**호선별 최종 혼잡도**	⋆ 역별 시간대별 혼잡도 ⋆ 역별 월별 비율 ⋆ 역별 요일별 비율	⋆ 칼럼 영어로 정리 ⋆ 30분단위, 1시간 단위로 있던 시간대를 1시간 단위로 통일 ⋆ 호선별 데이터 분할

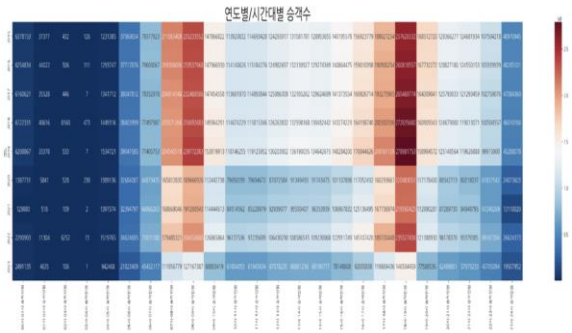
	YEAR	MONTH	DAY	LINE	STATION	DIRECTION	TIME_00	CONGESTION
0	2011	1	금	1	동대문	상선	TIME_05	9.1
1	2011	1	금	1	동대문	상선	TIME_06	12.2
2	2011	1	금	1	동대문	상선	TIME_07	23.3
3	2011	1	금	1	동대문	상선	TIME_08	44.6
4	2011	1	금	1	동대문	상선	TIME_09	39.6

부록참고

1호선 최종 혼잡도 산출물

‘시간대(24시)’, ‘월’, ‘요일’ 은 유의미한 변수라고 판단

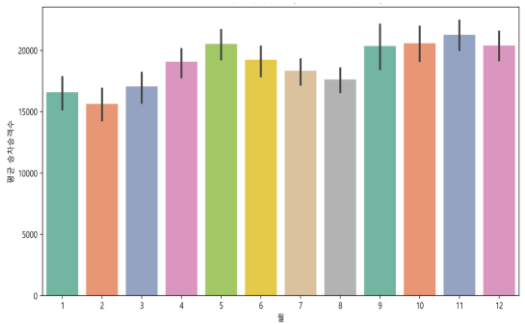
시간대 : 24시



2015~2023년 연도별 시간대별 평균 승차인원 히트맵

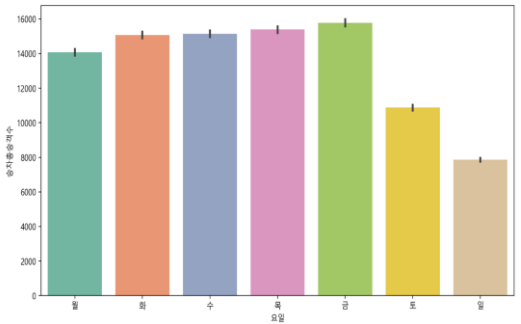
- 07~09시, 17~19시 출퇴근시간 승차인원증가

월별



- 봄/가을 증가, 여름/겨울 감소
- 계절별로 주기성을 보임.

요일 : 평일, 주말

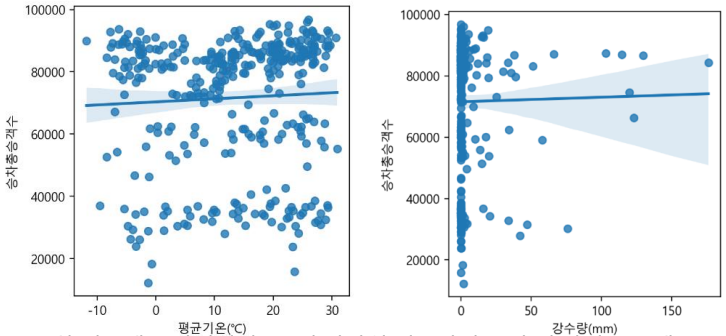


(좌) 4호선 성신여대입구역(연평균 혼잡도 최댓값: 8시 185.5) 기준 월별 / (우) 요일별 평균 승차 총 승객수

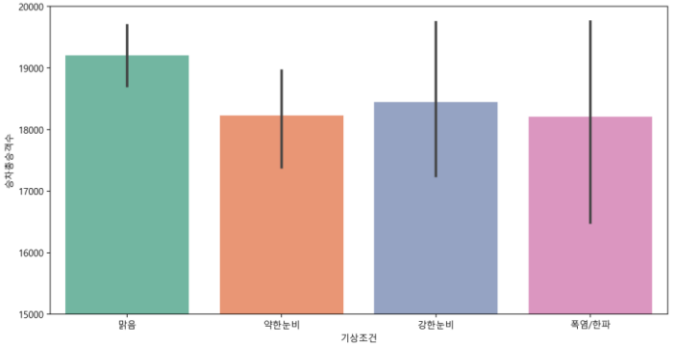
- 주중 > 주말
- 월→금 평균 승차 승객 수 증가

‘기상상황’ 은 변수에서 제외

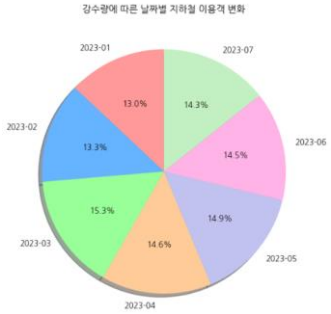
- 강수량은 0에 다수의 값이 몰려 있음.
- 평균기온, 강수량과 승차총승객수 사이의 관련성이 보이지 않음.
- 기상 조건 별 승차총승객수에 큰 차이가 나타나지 않으며 특히 강한 눈비, 폭염/한파 조건에서 오차범위가 크게 나타남.
- 1~7월 강수량의 차이에 비해 지하철 이용객 수의 변화가 크지 않음.



승차 이용객수 상위역인 2호선 강남역 기준날씨 조건 별 산점도 그래프



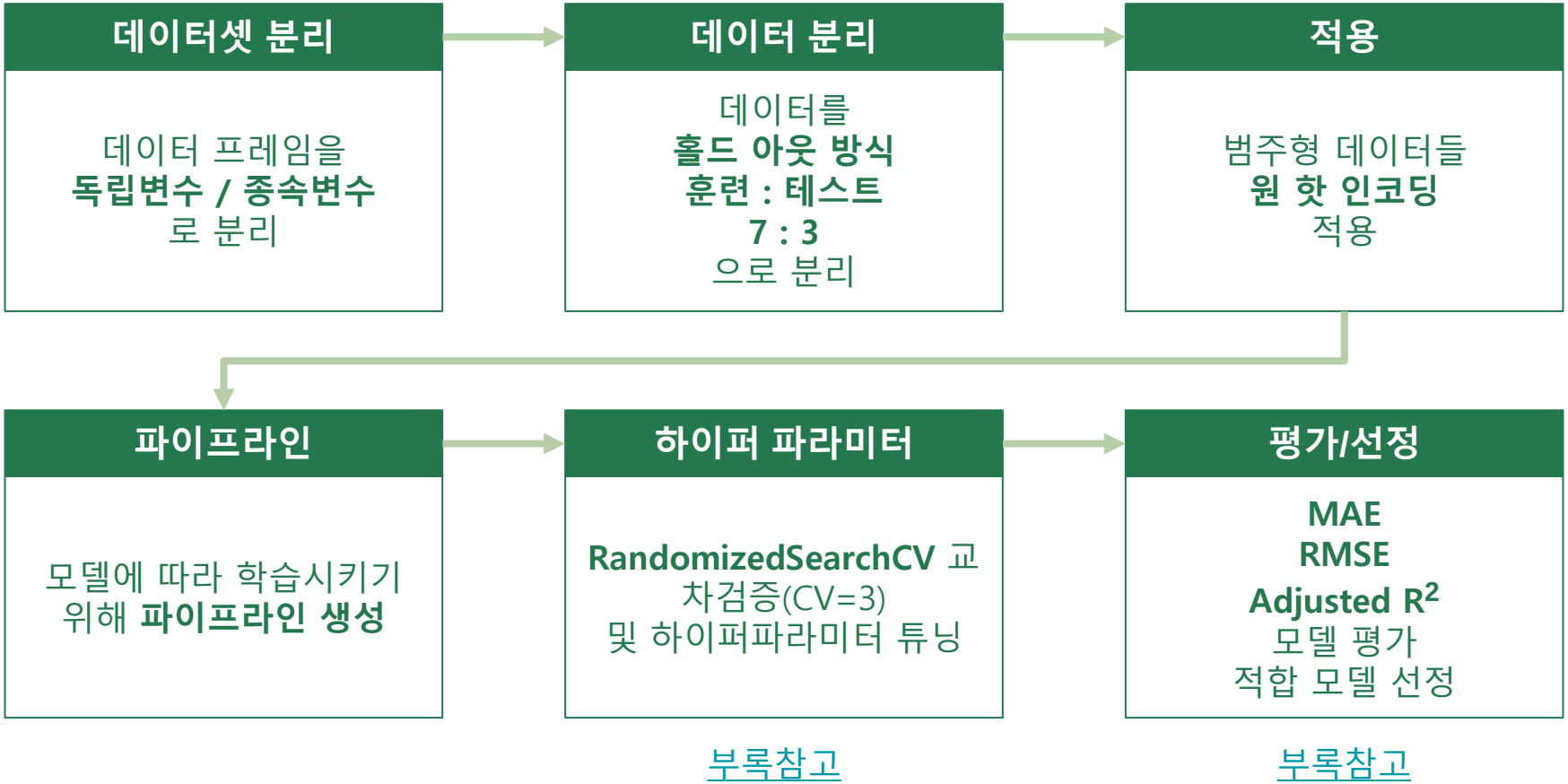
(좌) 기상조건별 지하철 이용객 수 비교 막대그래프



(우) 날짜별 강수량에 따른 지하철 이용객 수 파이 차트

모델개발 + 예측/평가

모델 학습 방식 | 최종 모델 | 모델 선정 근거 | 호선별 평가지표


데이터셋 분리
데이터 프레임을
독립변수 / 종속변수
로 분리
데이터 분리
데이터를
홀드 아웃 방식
훈련 : 테스트
7 : 3
으로 분리
적용
범주형 데이터들
원 핫 인코딩
적용
파이프라인
모델에 따라 학습시키기
위해 파이프라인 생성
하이퍼 파라미터
RandomizedSearchCV 교
차검증(CV=3)
및 하이퍼파라미터 튜닝
평가/선정
MAE
RMSE
Adjusted R2
모델 평가
적합 모델 선정

부록참고

부록참고

XGBRegressor

- 평가지표 + 피처 중요도 ⇒ XGBRegressor가 합리적.
- 서비스 적용 시 학습 완료 된 모델을 '.pkl' 파일로 사용함.
- **위험예측모델**을 만든다는 점.

⇒ **학습 시간과 관계없이 성능이 좋은 모델을 사용하기로 결정**

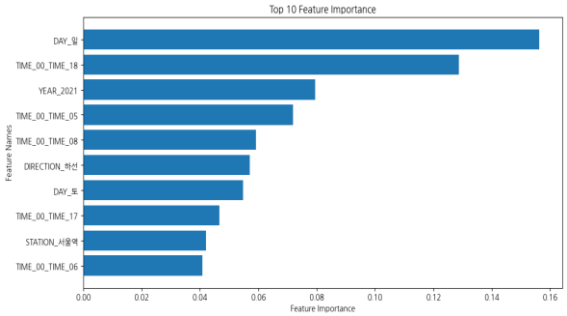
RandomForestRegressor

많은 데이터로 인해

학습에 많은 시간이 걸리지만

성능 향상은 적음

피처 중요도 Top 10



평가 지표 & 예측값 예시

```
Test 성능 지표:
Mean Absolute Error (MAE): 8.053420621977951
Root Mean Squared Error (RMSE): 11.614258229660223
Adjusted R-squared (Adjusted R2): 0.6806416832458583
```

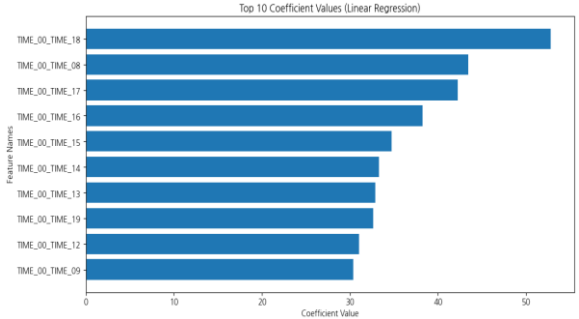
RandomForestRegressor

```
[ ] print_congestion_predictions('서울역', '동묘앞', 2023, 3, '월', 1, 'TIME_18', list_data)

    역: 서울역, 예측 혼잡도: 54.19
    역: 시청, 예측 혼잡도: 87.49
    역: 종각, 예측 혼잡도: 92.90
    역: 종로3가, 예측 혼잡도: 96.24
    역: 종로5가, 예측 혼잡도: 109.83
    역: 동대문, 예측 혼잡도: 92.90
    역: 신설동, 예측 혼잡도: 92.92
    역: 제기동, 예측 혼잡도: 92.81
    역: 청량리, 예측 혼잡도: 92.57
    1호선
    평균 혼잡도: 90.20
```

LinearRegression

범주형 변수를 사용하여

설명력이 떨어짐




```
Test 성능 지표:
Mean Absolute Error (MAE): 9.592732263217238
Root Mean Squared Error (RMSE): 12.998283027824531
Adjusted R-squared (Adjusted R2): 0.5999933114891132
```

LinearRegression

```
[ ] print_congestion_predictions('서울역', '동묘앞', 2023, 3, '월', 1, 'TIME_18', list_data)

    역: 서울역, 예측 혼잡도: 66.93
    역: 시청, 예측 혼잡도: 67.42
    역: 종각, 예측 혼잡도: 68.38
    역: 종로3가, 예측 혼잡도: 69.50
    역: 종로5가, 예측 혼잡도: 70.19
    역: 동대문, 예측 혼잡도: 68.64
    역: 신설동, 예측 혼잡도: 66.80
    역: 제기동, 예측 혼잡도: 64.68
    역: 청량리, 예측 혼잡도: 62.27
    1호선
    평균 혼잡도: 67.20
```

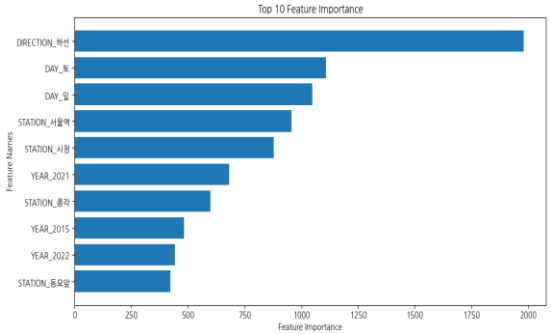
LGBMRegressor

XGBRegressor보다

성능은 낮지만

학습 시간에 있어서 유리

피처 중요도 Top 10



평가 지표 & 예측값 예시

```
Test 성능 지표:
Mean Absolute Error (MAE): 3.4350431853275323
Root Mean Squared Error (RMSE): 5.103735724115479
Adjusted R-squared (Adjusted R2): 0.9383303290485041
```

LGBMRegressor

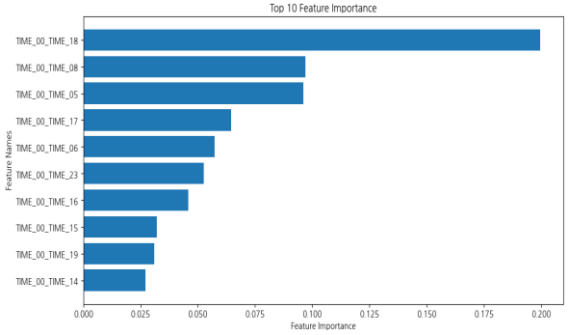
```
[ ] print_congestion_predictions('서울역', '동묘앞', 2023, 3, '월', 1, 'TIME_18', list_data)

    역: 서울역, 예측 혼잡도: 64.84
    역: 시청, 예측 혼잡도: 82.68
    역: 종각, 예측 혼잡도: 93.70
    역: 종로3가, 예측 혼잡도: 95.70
    역: 종로5가, 예측 혼잡도: 101.20
    역: 동대문, 예측 혼잡도: 89.15
    역: 신설동, 예측 혼잡도: 90.36
    역: 제기동, 예측 혼잡도: 83.85
    역: 청량리, 예측 혼잡도: 82.76
    4호선
    평균 혼잡도: 87.14
```

XGBRegressor

4가지 모델 중

가장 성능이 좋으나,

학습 시간이 오래걸림

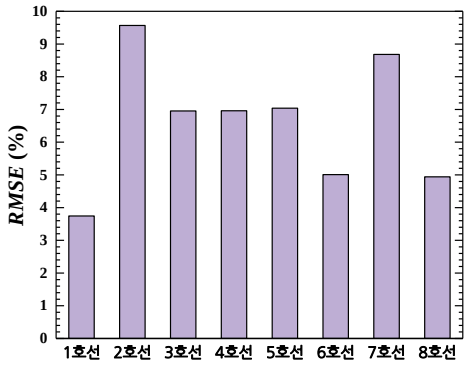


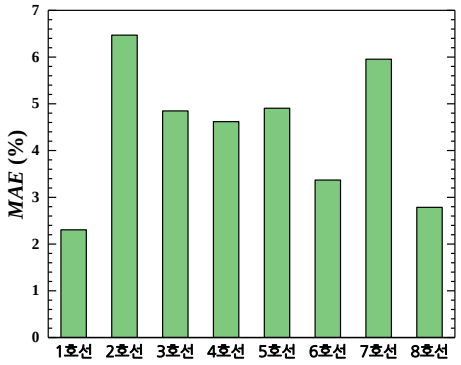

```
Test 성능 지표:
Mean Absolute Error (MAE): 2.305130711934932
Root Mean Squared Error (RMSE): 3.74522863563344
Adjusted R-squared (Adjusted R2): 0.9667912946014372
```

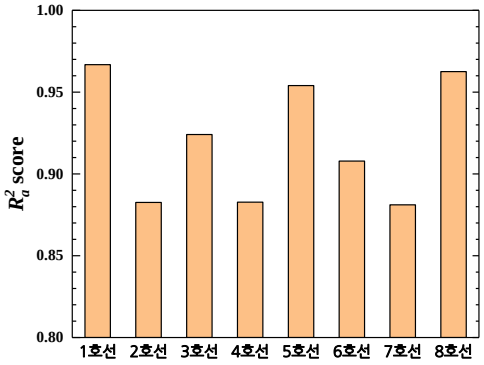
XGBRegressor

```
[ ] print_congestion_predictions('서울역', '동묘앞', 2023, 3, '월', 1, 'TIME_18', list_data)

    역: 서울역, 예측 혼잡도: 63.41
    역: 시청, 예측 혼잡도: 84.20
    역: 종각, 예측 혼잡도: 95.89
    역: 종로3가, 예측 혼잡도: 100.25
    역: 종로5가, 예측 혼잡도: 106.14
    역: 동대문, 예측 혼잡도: 91.73
    역: 신설동, 예측 혼잡도: 93.60
    역: 제기동, 예측 혼잡도: 89.71
    역: 청량리, 예측 혼잡도: 84.17
    4호선
    평균 혼잡도: 89.90
```

호선별 평가지표 정리 (XGBRegressor) 부록참고







RMSE

3.7~9.6%

MAE

2.3~6.5%

Adjusted R^2

0.88 이상

웹 서비스

영상 | 앱 개요 | 전체 정리

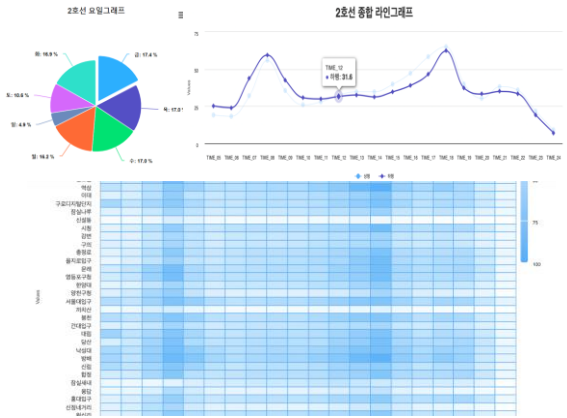
영상 링크

서비스 링크

부록참고

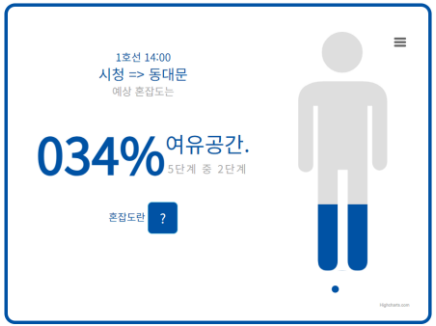
대시보드 - 분석 + 전처리

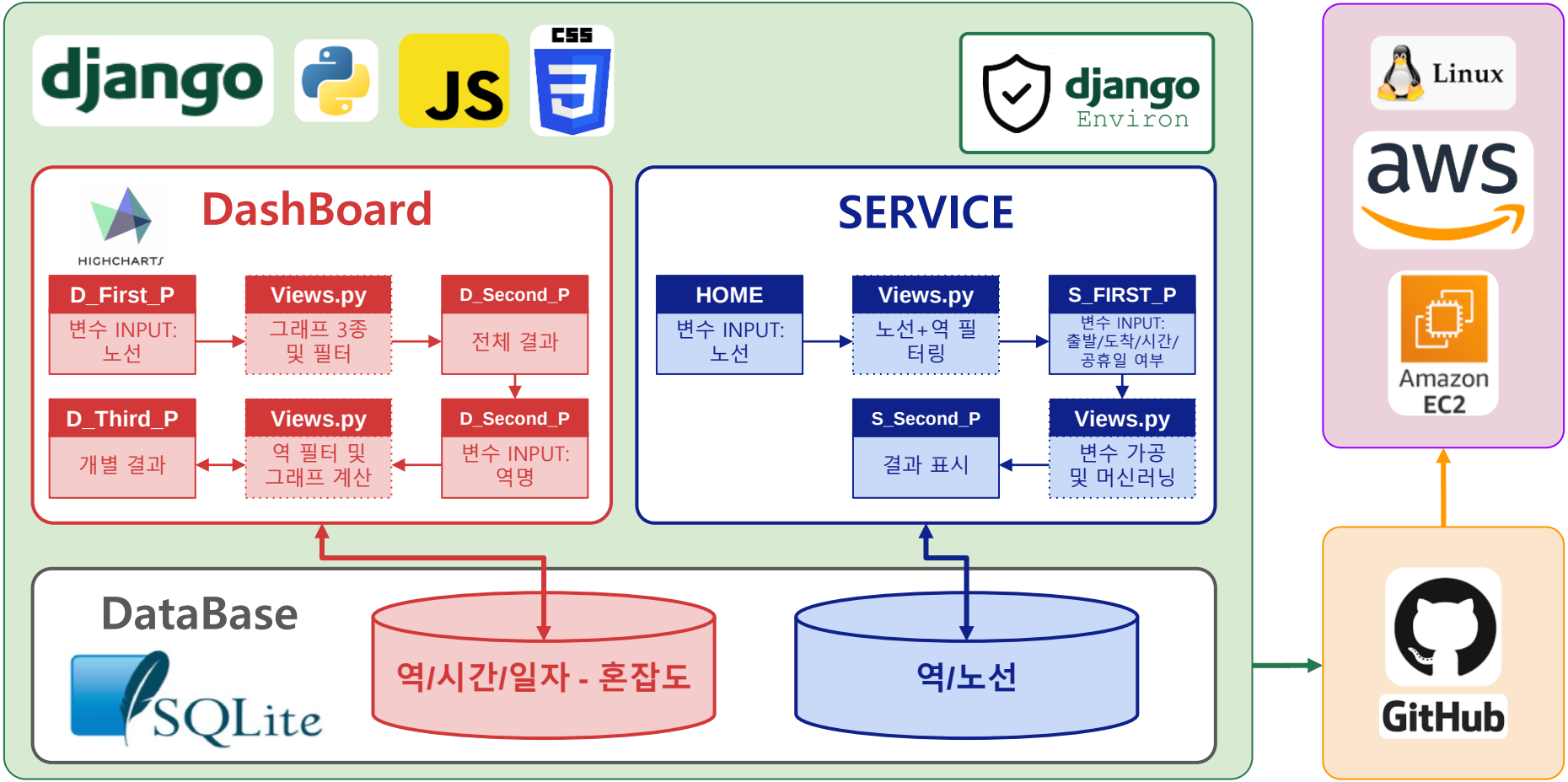
1. 데이터 전처리
2. 시각화
- 파이: 요일
- 선: 시간
- 히트맵: 시간 + 역



서비스 - 머신러닝

1. pkl 파일로 저장: 학습데이터
2. 변수 삽입: 출발/도착역, 시간,
3. 전체 혼잡도의 평균 계산
4. 별도의 로직 구현: 특이호선




django
JS
CSS
django Environ
DashBoard
HIGHCHARTS
D_First_P
변수 INPUT: 노선
Views.py
그래프 3종 및 필터
D_Second_P
전체 결과
D_Third_P
개별 결과
Views.py
역 필터 및 그래프 계산
D_Second_P
변수 INPUT: 역명
SERVICE
HOME
변수 INPUT: 노선
Views.py
노선+역 필터링
S_FIRST_P
변수 INPUT: 출발/도착/시간/공휴일 여부
S_Second_P
결과 표시
Views.py
변수 가공 및 머신러닝
DataBase
SQLite
역/시간/일자 - 혼잡도
역/노선
Linux
aws
Amazon EC2
GitHub

DOCUMENT

한계점 및 개선사항, 기대효과 | 참고문헌 | 부록

한계점 및 개선사항, 기대효과

한계점 1	서울교통공사가 제공하는 데이터만 사용(1~8호선)
개선 사항 1	서울시 노선도의 다른 운영사에 혼잡도 데이터 제공 요청하여 가용 노선을 늘림
한계점 2	실시간 지하철 혼잡도 데이터를 사용하지 못함 - 과거 데이터로만 예측하므로 실제 상황과의 차이가 나타날 것
개선 사항 2	SKT 등 통신 빅데이터와 융합하여 데이터를 보완하여 실시간 혼잡도를 적용
한계점 3	COVID-19 이후 최신 혼잡도 데이터가 2022년만 존재 연도별 데이터 불충분으로 데이터 가공과정에서 오차 가능성
개선 사항 3	혼잡도 비율 산출 과정에서 2020년, 2021년을 제외, 추후 데이터 누적으로 정확도 향상
기대 효과	승객들의 만족도 향상 / 다른 분야에서의 데이터 활용 기대(교통 인프라, 도시 계획 등)

보도자료

·대한민국 정책브리핑 https://www.korea.kr/main.do
·중앙일보 https://www.joongang.co.kr/
·연합뉴스 https://www.yna.co.kr/
·JTBC뉴스 https://bit.ly/455i1kX
·현대로템 https://blog.hyundai-rotem.co.kr/691

혼잡도 참고 자료

·국토교통부 공식 네이버 포스트
https://post.naver.com/my.naver?memberNo=5113437
·서울시 도시기반시설본부 도시철도계획부,
「도시철도 관련 법령 및 규정 자료집」, 2020, p.149
·Keith Still 교수, 군중밀집임계점
https://www.gkstill.com/Support/crowd-density/CrowdDensity-1.html

논문 및 간행물

·최상기, 이종호 and 오승훈(2013), 기상조건이 대중교통수요에 미치는 영향에 관한 연구,
대한토목학회논문집(국문), 33(6), 2447-2453
·주재영(2015), 도시철도 혼잡도 기준 개선방안 연구, 한국교통대학교 교통대학원 석사학위논문
·김승준(2016), 서울시 지하철의 혼잡비용 산정과 정책적 활용방안, 서울 연구원 정책리포트 208
·최정윤, 원민수(2020), 기상에 따른 대중교통 이용변화의 영향도 분석, 한국교통학회 학술대회지, 286~287
·통계청 연구기획실(2023), 코로나19 확산에 따른 도시철도의 통행량 변화, KOSTAT 통계플러스 2023년 여름호, 6-17

Q&A

부록

혼잡도별 상태도

지하철 혼잡도 계산식

데이터 기간 설정 근거 자료

최종 혼잡도 계산

데이터 정의서

머신러닝 모델 평가지표

하이퍼파라미터 튜닝

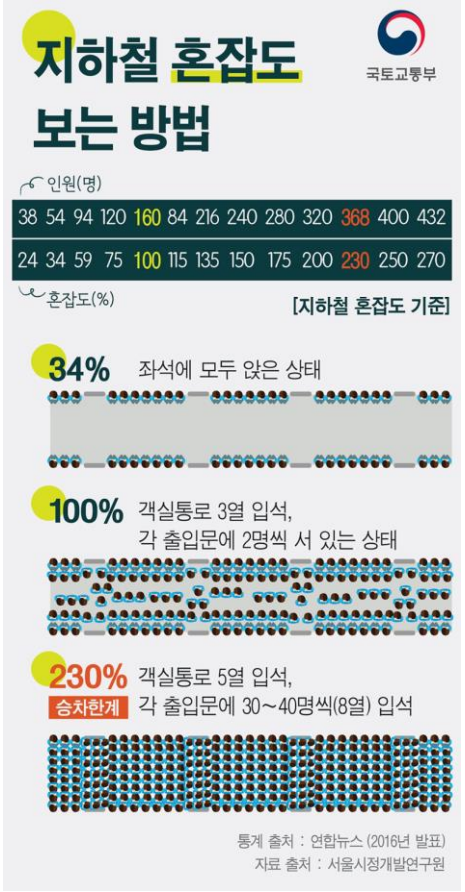
XGBRegressor 호선별 평가지표(상세값)

혼잡도별 상태도 돌아가기

표 9. 지하철 차량내부 혼잡도 조건표 재구성

혼잡도	재차인원	차내상태 설명 ① 입석승객 대상, ② 좌석승객 대상	혼잡 상태도
50%	80명	① 좌석에 모두 착석하고 간간이 서 있음 ② 앞의 시야가 트임	
100%	160명	① 여유롭게 서 있음 ② 앞에 사람들이 서 있어서 시야가 다소 막힘	
125%	200명	① 지나갈 때 사람과 부딪치게 되는 다소 혼잡한 상태 ② 앞에 사람들이 많이 서 있어서 시야가 막힘	
150%	240명	① 출입문 주변이 혼잡하고 서로 어깨가 밀착됨 ② 앞에 서 있는 사람들이 밀치기도 하여 불쾌감을 느끼기도 함	
175%	280명	① 출입문 주변이 매우 혼잡하고 서로 몸이 밀착되어 팔을 들 수 없음 ② 앞에 서 있는 사람들과 무릎이 닿기도 하여 불쾌함	
200%	320명	① 출입문 주변이 매우 혼잡하고 서로 몸과 얼굴이 밀착되어 숨이 막힘 ② 서 있는 사람들이 심하게 밀려 발이 밟히기도 하고 '악' 소리가 나면서 소란스러움	

김승준, 서울시 지하철의 혼잡비용 산정과 정책적 활용방안, *서울연구원 정책리포트* 208, p. 12 (2016)



지하철 혼잡도 계산식 - 서울시 돌아가기

$$(\text{수송용량})_{l,d,s,t} = (\text{운행횟수})_{l,d,s,t} \times (\text{차량편성})_{l} \times (\text{차량정원})_{l}$$

$$(\text{차내혼잡도})_{l,d,s,t} = 100 \times \frac{(\text{재차인원})_{l,d,s,t}}{(\text{수송용량})_{l,d,s,t}}$$

수송용량 : 시간당 인원수
운행횟수 : 시간당 횟수
차량편성 : 량
차량정원 : 1량당 인원수

차내 혼잡도 : %
재차인원 : 시간당 인원수

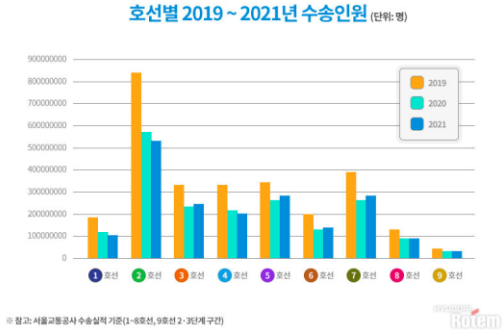
l : 지하철 노선
d : 방향(상행/하행 또는 내선/외선)
s : 지하철역 간 구간
t : 시간대(지하철 운행시간을 1시간 단위로 구분: 05~25시(익일 1시))

김승준, 서울시 지하철의 혼잡비용 산정과 정책 활용, *서울연구원*, p. 71 - 73 (2014)

데이터 기간 선정 근거 (1) 돌아가기

[코로나19에 따른 지하철 통행량 변화]

사회적 거리두기 : 2020년 5월 6일 시작, 2020년 11월 7일에 5단계 지침이 적용되었다가 2022년 4월 18일 전면 해제



호선별 2019 ~ 2021년 수송인원 변화 (현대로템)

도시철도 월간 통행량 변화 (KOSTAT 통계플러스 2023년 여름호)

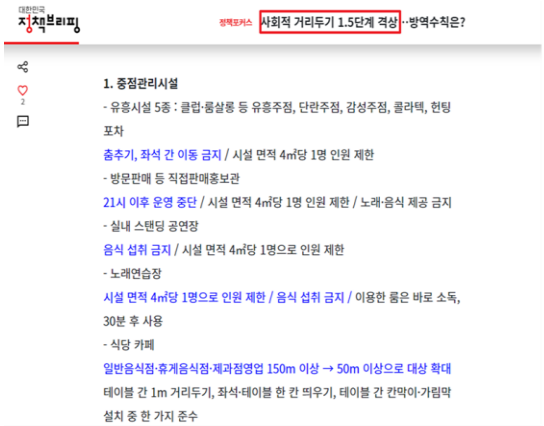
코로나 이후(2020, 2021년)에 전 호선의 수송인원이 크게 감소 (*서울교통공사 운영 구간 기준

- 2022년 : 약 28억 2,625만 명
- 2021년 : **약 19억 9,935만 명**
- 2020년 : **약 19억 7,912만 명**

코로나19 발생 이후 2020년 3월에 가장 크게 감소하여 이전에 비해 약 **43% 감소**, 그 이후 2021년까지도 통행량이 감소되다가 2022년에 이르러서 완만하게 복원되는 상황.

데이터 기간 선정 근거 (2)

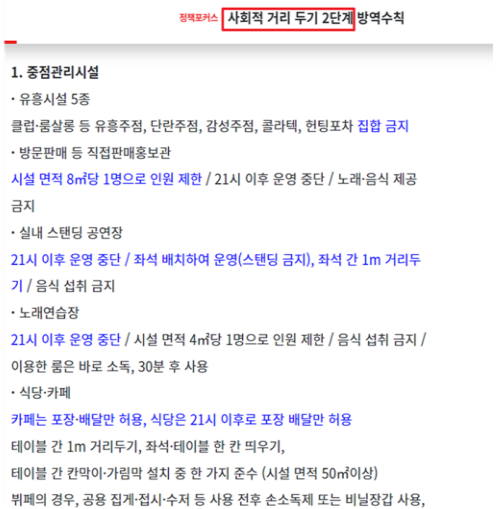
[사회적 거리두기 방역수칙]



대한민국 정책브리핑

정책포커스 사회적 거리두기 1.5단계 격상‥방역수칙은?

1. 중점관리시설
- 유흥시설 5종 : 클럽·룸살롱 등 유흥주점, 단란주점, 감성주점, 콜라텍, 헌팅포차
춤추기, 좌석 간 이동 금지 / 시설 면적 4㎡당 1명 인원 제한
- 방문판매 등 직접판매홍보관
21시 이후 운영 중단 / 시설 면적 4㎡당 1명 인원 제한 / 노래·음식 제공 금지
- 실내 스탠딩 공연장
음식 섭취 금지 / 시설 면적 4㎡당 1명으로 인원 제한
- 노래연습장
시설 면적 4㎡당 1명으로 인원 제한 / 음식 섭취 금지 / 이용한 룸은 바로 소독, 30분 후 사용
- 식당 카페
일반음식점·휴게음식점·제과점영업 150m 이상 → 50m 이상으로 대상 확대
테이블 간 1m 거리두기, 좌석·테이블 한 칸 띄우기, 테이블 간 칸막이·가림막 설치 중 한 가지 준수



정책포커스 사회적 거리 두기 2단계 방역수칙

1. 중점관리시설
· 유흥시설 5종
클럽·룸살롱 등 유흥주점, 단란주점, 감성주점, 콜라텍, 헌팅포차 집합 금지
· 방문판매 등 직접판매홍보관
시설 면적 8㎡당 1명으로 인원 제한 / 21시 이후 운영 중단 / 노래·음식 제공 금지
· 실내 스탠딩 공연장
21시 이후 운영 중단 / 좌석 배치하여 운영(스탠딩 금지), 좌석 간 1m 거리두기 / 음식 섭취 금지
· 노래연습장
21시 이후 운영 중단 / 시설 면적 4㎡당 1명으로 인원 제한 / 음식 섭취 금지 / 이용한 룸은 바로 소독, 30분 후 사용
· 식당·카페
카페는 포장·배달만 허용, 식당은 21시 이후로 포장 배달만 허용
테이블 간 1m 거리두기, 좌석·테이블 한 칸 띄우기,
테이블 간 칸막이·가림막 설치 중 한 가지 준수 (시설 면적 50㎡이상)
뷔페의 경우, 공용 집게·접시·수저 등 사용 전후 손소독제 또는 비닐장갑 사용,

[관련 표제]

'코로나바이러스감염증-19' 대응
수도권 거리두기 2.5단계·비수도권 2단계로 격상…8일부터 3주간 적용
수도권 밤 9시 이후 영화관·대형마트 중단…결혼식장 50인 미만으로
등교 인원은 3분의 1 수준 유지…종교활동 비대면 원칙 속 20명 이내

'코로나바이러스감염증-19' 대응
현행 거리두기 2주 연장…설 연휴 직계가족도 5인 이상 모임 금지
다음달 14일까지 수도권 2.5단계·비수도권 2단계 유지

오늘부터 서울지하철 밤10시 이후 20% 단축 운행
중앙일보 | 입력 2020.11.27 10:30
뉴스홈 최신기사

12일부터 수도권 전철 밤 10시 이후 운행 감축‥막차시간 빨라져
송고시간 | 2021-07-09 18:19
요약
멈춤 기간' 관련 대책 중 하나로 시민들의 이른 귀가를 유도하기 위한 조치다.

데이터 정의서 (1) 돌아가기

번호	테이블 명
1	혼잡도(11, 13, 15, 17, 19, 21, 22년도)
2	서울교통공사 일별 역별 시간대별 승하차인원(1~8호선)(08년~22년)
3	공휴일 21년~23년, 호선별_역코드, 서울시 지하철역 정보 검색 (역명)
4	station_name / transfer_station_name
5	서울교통공사 월별 승하차인원
6	호선별역별_승하차정보(2021-2023)
7	서울_기상조건_2021, 월평균상대습도(2021-2023), 체감온도(2021-2023)
8	최종혼잡도(영문)
9	역별요일비율, 역별월별비율

데이터 정의서 (2)

번호	데이터 셋	테이블	작성일자	열 이름	데이터 타입	NULL
1	Raw_Data	1	2023.09.22.	요일구분	OBJECT	비허용
2		1,3,4		호선	OBJECT	비허용
3		1		역명	OBJECT	비허용
4		1		상하구분	OBJECT	비허용
5		1		05시~06시(이후 시간대 동일)	INTEGER	비허용
6		2		연번	INTEGER	비허용
7		2		수송일자	OBJECT	비허용
8		2		호선	INTEGER	비허용
9		2		고유역번호(외부역코드)	OBJECT	비허용
10		2		역명	OBJECT	비허용
11		2		승하차구분	OBJECT	비허용
12		2		06시이전(이후 시간대 동일)	INTEGER	비허용
13		5		사용월	INTEGER	비허용
14		5		호선명	OBJECT	비허용
15		5,6		지하철역	OBJECT	비허용

데이터 정의서 (3)

16	Raw_Data	5	2023.09.22	04시-05시 승차인원(이후 시간대 동일)	INTEGER	비허용
17		3,4		dateName	OBJECT	비허용
18		3,4		locdate	INTEGER	비허용
19		3,4		전철역코드	INTEGER	비허용
20		3,4		전철역명	OBJECT	비허용
21		3,4		외부코드	OBJECT	비허용
22		3,4		STATN_NM(호선이름)	OBJECT	비허용
23		6		사용일자	OBJECT	비허용
24		6		연	INTEGER	비허용
25		6		월	INTEGER	비허용
26		6		일	INTEGER	비허용
27		6		요일	OBJECT	비허용
28		6		호선명	OBJECT	비허용
29		6		승차총승객수	FLOAT	비허용
30		6		하차총승객수	FLOAT	비허용

데이터 정의서 (4)

31	Raw_Data	7	2023.09.22	날짜	OBJECT	비허용
32		7		지점	FLOAT	0으로
33		7		평균기온(℃)	FLOAT	0으로
34		7		최저기온(℃)	FLOAT	0으로
35		7		최고기온(℃)	FLOAT	0으로
36		7		강수량(mm)	FLOAT	0으로
37		7		일시	OBJECT	비허용
38		7		평균상대습도(%)	INTEGER	비허용
39		7		최소상대습도(%)	INTEGER	비허용
40		7		일자	OBJECT	비허용
41		7		기온(°C)	FLOAT	0으로
42		7		풍속(km/h)	FLOAT	0으로
43		7		체감온도(°C)	FLOAT	0으로

데이터 정의서 (5)

44	PreProcessing_Data	8	2023.09.22.	YEAR	INTEGER	비허용
45		8		MONTH	INTEGER	비허용
46		8		DAY	OBJECT	비허용
47		8		LINE	INTEGER	비허용
48		8		STATION	OBJECT	비허용
49		8		DIRECTION	OBJECT	비허용
50		8		TIME_00	OBJECT	비허용
51		8		CONGESTION	FLOAT	비허용
52		9		호선	INTEGER	비허용
53		9		역명	OBJECT	비허용
54		9		요일	OBJECT	비허용
55		9		비율	FLOAT	비허용
56		9		호선 명	OBJECT	비허용
57		9		지하철 역	OBJECT	비허용
58		9		사용월	INTEGER	비허용

데이터 정의서- DataBase 돌아가기

작성자: 김은채 | SQLite3					
번호	데이터 셋	DB_Name	작성일자	열 이름	데이터 타입
1	PreProcessing_Data	DashBoard_vi_csv_no 1~8,21,22,51	23.09.19	id	IntegerField
2				MONTH	IntegerField
3				DAY	CharField
4				STATION	CharField
5				DIRECTION	CharField
6				TIME_05~24	FloatField
7		FirstApp_all_ln_hs_f		SW_ID	IntegerField
8				LINE	IntegerField
9				STATION	CharField

머신러닝 모델 평가지표(1) 돌아가기

❏ MAE (Mean Absolute Error)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

1. 실제값과 예측값의 차이들을 평균으로 표시
2. 직관적으로 편차를 알 수 있다.
3. 잔차의 부호를 알 수 없다.
4. 미분 불가능한 점이 있다.

머신러닝 모델 평가지표(2) 돌아가기

❑ RMSE (Root Mean Square Error)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

1. MSE 값에 제곱근을 취하여 종속변수와의 단위를 맞춰준다.
2. 잔차가 큰 부분에 더 큰 패널티를 주어 MAE보다 학습에 유리
3. 단위는 같으나 주로 MAE보다 큰 값을 가져, 직관적으로 편차를 비교하기 힘들다.
4. 잔차의 부호를 알 수 없다.

머신러닝 모델 평가지표(3) 돌아가기

❑ Adjusted R^2

$$R_a^2 = 1 - \frac{(n-1)(1-R^2)}{n-p-1}$$

1. 적합도 검정을 위한 지표로 1에 가까울 수록 설명력이 좋음
2. 상관성이 있는 특성변수의 개수가 늘어나면 필연적으로 R^2 값이 늘어나기 때문에 변수 개수에 따른 패널티 값으로 보정해줌

하이퍼파라미터 튜닝 (1) RandomForestRegressor

돌아가기

- n_estimator: **100** (트리의 개수)
- max_depth: 6, 8, 10, **12** (각 트리의 최대 깊이) - 과적합 방지용
- min_samples_split: 2, **5**, 10 (최소 노드 분할 개수)

하이퍼파라미터 튜닝 (2) LGBMRegressor

- learning_rate: 0.01, 0.05, **0.1** (학습률)
- n_estimators: **500** (트리의 개수)
- max_depth: 4, **6**, 8 (각 트리의 최대 깊이)
- colsample_bytree: 0.7, 0.8, **0.9**, 1.0 (열 샘플링 비율)
- subsample: 0.7, **0.8**, 0.9, 1.0 (행 샘플링 비율)
- min_child_samples: **20**, 50, 100 (리프 노드의 최소 데이터 수)
- reg_alpha: 0, **0.01**, 0.1 (L1 규제 강도)
- reg_lambda: 0, **0.01**, 0.1 (L2 규제 강도)

하이퍼파라미터 튜닝 (3) XGBRegressor

돌아가기

- learning_rate: 0.01, 0.05, 0.1, 0.3, **0.5** (학습률)
- n_estimators: 50, 100, **150**, 200, 250 (트리의 개수)
- max_depth: 3, 4, 5, 6, **7**, 8 (각 트리의 최대 깊이)
- min_child_weight: **1**, 5, 10 (리프 노드의 최소 가중치 합)
- colsample_bytree: 0.5, 0.6, 0.7, 0.8, **0.9** (열 샘플링 비율)
- subsample: 0.5, 0.6, 0.7, 0.8, **0.9** (행 샘플링 비율)
- gamma: **0**, 0.1, 0.2, 0.3, 0.4 (분할 최소 이득) - Gain 값이 gamma값 이상일 때만 분할

XGBRegressor 호선별 평가지표(상세값) 돌아가기

	1호선	2호선	3호선	4호선	5호선	6호선	7호선	8호선
MAE	2.3051307 11934932	6.4703175 73801609	4.8487172 79813802	4.6176506 97623419	4.9049185 04183993 6	3.3684238 79204473	5.9548073 26621816	2.7845271 6337553
RMSE	3.7452286 3563344	9.5683861 15553132	6.9540029 0658048	6.9606261 92862513	7.0403112 36860753	5.0095577 60212106	8.6825997 4873913	4.9416186 54246216
R2	0.9667912 94601437 2	0.8825795 15184846 7	0.9240798 22065628 7	0.8827364 54075562 6	0.9540203 83460663 8	0.9078869 14843878 6	0.8811254 57166286 7	0.9625757 21150904 3