PNU SW학습공동체 중간보고서

Github로 파일업로드하여 제출

1. 프로젝트 소개

가. 배경 및 필요성

사람들은 날씨나 요일, 일정에 따라 외출 패턴이 달라진다. 출퇴근 시간대는 물론, 비가 오거나 더운 날, 방학이나 연휴가 있는 시기에는 지하철을 이용하는 사람 수가 크게 변한다.

비록 현재 지하철이 출퇴근, 평시, 야간 시간대에 따라 배차 간격에 차이가 있지만 기상 상황, 수요 급증 등 예외 상황에 대해 변화를 잘 반영하지 못하고 고정된 배차 간격으로 운행되는 경우가 많다. 결과적으로 이러한 상황은 승객의 불편을 초래할 뿐 아니라, 교통 자원의 비효율적 운영으로도 이어질 수 있다.

이런 문제를 해결하기 위해, 우리는 날씨와 사회적 요인을 바탕으로 지하철 승하차 인원을 예측하고, 이를 통해 날짜별·시간대별로 배차를 유연하게 조정할 수 있는 전략을 제안하려 한다.

나. 개발목표 및 주요내용, 세부내용 등

이번 프로젝트의 목표는 날씨와 요일, 기상특보, 주요 일정(방학·공휴일 등) 같은 외부 요인들이 지하철 이용에 어떤 영향을 미치는지 분석하고, 이를 예측 모델로 만드는 것이다.

예측된 정보를 기반으로 어떤 날 어떤 시간대에 지하철이 붐빌지를 미리 파악하고, 지하철을 더 자주 보내거나 덜 보내는 식의 탄력적 배차 조정 전략을 함께 제안하고자 한다.

주요 분석 내용

- **기본 정보 기반 분석**: 날짜, 시간대, 요일에 따른 이용객 변화 양상
- **날씨 요인 분석**: 기온, 강수, 미세먼지, 기상특보(폭염, 폭설, 폭우, 태풍)
- **사회적 요인 분석**: 학원가, 대학로 주변, 오피스 밀집 지역, 관광지 주변 역
- **시즌 요소 고려**: 방학, 연휴, 꽃놀이 시즌 등 월별 특성 고려
- **예측 모델 개발**: 날씨·요일 등 외부 요인 기반으로 승하차 인원 예측

다. 사회적가치 도입 계획 등

- **출퇴근·등하교 시간대 혼잡 완화**: 예측 정보를 활용해 혼잡 예상 시간에 지하철을 더 자주 운행할 수 있음
- **시민 편의 향상**: 날씨나 일정 변화에도 쾌적한 대중교통 이용 가능
- **스마트 교통 운영 기반**: 실제 데이터 기반으로 배차를 유동적으로 조정하는 시범 모델

- **공공데이터 활용 가치 확대**: 지하철, 기상, 관광 데이터 등 다양한 공공데이터를 융합해 활용

2. 상세설계

가. 시스템 구성도, 사용기술 등

1. **기본 데이터 처리 및 전처리**
   - pandas : 표 형식 데이터 처리 및 탐색
   - numpy : 수치 연산 지원
2. **위·경도 변환**
   - 기존 데이터셋에 포함된 stn(기상청 관측소 번호)을 기반으로, 기상청에서 제공하는 AWS 관측소 위치 정보와 매핑하여 각 지하철역에 대응되는 위도, 경도 정보를 확보하였음.
   - 이후 확보된 위도·경도 정보를 활용하여 카카오 API를 통해 지번 주소로 역변환하였음.
   - 향후 인구, 소득, 인프라 등 외부 공공데이터와의 연계에 활용할 수 있도록 변환된 주소를 통해‘시/군/구’ 행정단위를 추출하여 따로 저장하였음.
3. **시각화**
   - matplotlib.pyplot : 기초 시각화 도구
   - seaborn : 통계적 시각화 강화
   - plotly : 대화형 시각화
4. **날짜·시간 처리**
   - datetime, pandas.to_datetime : 시간 데이터 정제 및 분해
5. **외부 요인 분석용 라이브러리**
   - geopandas / folium : 역별 위치 데이터 및 밀집 지역 시각화
6. **모델링 및 예측 (머신러닝)**
   - scikit-learn : 예측 모델 생성 및 평가
   - xgboost, lightgbm : 고성능 예측 모델 (시계열적 특성 있는 데이터에 적합)

3. 개발결과

가. 전체시스템 흐름도, 기능설명, 기능명세서, 디렉토리 구조 등

1) 분석 흐름

데이터 수집 → 전처리 (시간/사회/기후 요인 기준) → 변수 선택 및 정리

→ 상관관계 분석 → 시각화 → 중간 결과 확인

2) 전처리 및 분석 내용 요약

분석 항목	설명
시간별 요인	월/일/시간대/공휴일/출근시간/요인 등으로 혼잡도 변화 분석
사회적 요인	대학가, 관광지, 중심가, 환승역, 학원가, 업무 지구 등 영향 확인
기후 요인	미세먼지, 강수량, 기온 등과 혼잡도의 상관관계 분석
전처리 방식	결측치 처리, 포맷 통일, 이상치 제거, 변수 통합 등
분석 도구	Python (pandas, seaborn, matplotlib 등) 사용
상관관계 분석	시각화 기반 해석 수행

3) 중간 데이터셋 전처리 결과

- **시간 별 결과**

- 요일 그룹별 비교

  데이터를 '월목'(평일)과 '금일'(주말 및 금요일)로 나누어 비교한 결과, 월~목 그룹의 지하철 이용객 수가 확연히 더 많은 경향을 보였다. 이는 출근 및 등하교 목적의 이동 수요가 주요 원인으로 분석된다.

- 시간대별 이용 패턴

  특정 역(예: 강남, 서울역, 이대 등)을 기준으로 시간대별 승하차 인원을 비교한 결과,

  오전 출근·등교 시간대(07~09시)에는 하차 인원이 승차 인원보다 많았고,
  저녁 퇴근·하교 시간대(17~19시)에는 승차 인원이 하차 인원보다 많았다.

이는 대중교통의 대표적인 출퇴근 흐름을 잘 반영하고 있으며, 특정 시간대에 집중되는 승객 수요를 탄력적으로 관리할 필요성을 시사한다.

- 월별 이용 추이

  봄과 여름 데이터를 비교 분석했을 때, 월별 이용자 수 자체는 변동이 있었지만 날짜별·요일별 분석만큼 명확한 차이를 보이지는 않았다. 이는 월별 이용량보다는 요일과 시간대의 영향이 더욱 크다는 점을 시사한다.

- 공휴일 vs 비공휴일 비교

  공휴일에는 비공휴일(평일)에 비해 지하철 총 이용객 수가 약 20% 감소하는 경향이 나타났다. 이는 주로 출퇴근 목적의 이동이 줄어든 데 따른 것으로 해석되며, 예측 모델에서 공휴일 여부를 중요한 변수로 고려할 수 있음을 보여준다.

- 연도별 비교

  2021년과 2023년의 지하철 혼잡도를 학기 중(Semester)과 방학(Vacation) 기간으로 나누어 비교한 결과, 2023년이 2021년에 비해 전반적으로 혼잡도가 더 높게 나타났다. 이는 코로나19 방역 조치 완화 이후 대중교통 이용 수요가 회복되었음을 보여주는 지표로 해석할 수 있다.

2021년은 사회적 거리두기 강화, 재택근무, 온라인 수업 등의 영향으로 지하철 수요가 일시적으로 감소했던 시기였다. 반면 2023년은 방역 규제 대부분이 해제되며 일상 생활이 회복 국면에 들어섰고, 이에 따라 혼잡 수준 역시 상승하는 경향을 보였다. 이는 특히 상위 사분위수 이상 혼잡도 분포와 이상치(outlier)의 수가 크게 늘어난 점에서도 확인할 수 있었다.

또한 두 연도 모두에서 학기 중의 혼잡도가 방학 기간보다 확연히 높은 경향을 보였다. 이는 학생들의 통학 및 직장인의 출퇴근 수요가 집중되는 학기 중에 대중교통 이용이 더 활발하게 이뤄지기 때문이며, 방학 기간에는 상대적으로 혼잡도가 완화되는 모습을 나타냈다.

이러한 결과는 지하철 승객 수요가 요일 및 시간대, 공휴일 여부 등 특정 패턴에 따라 반복적으로 나타나는 구조를 가지고 있음을 보여주며, 향후 예측 모델의 신뢰성을 높이는 데 중요한 근거로 활용될 수 있다.

- **사회적 요인에 따른 지하철 혼잡도 분석 결과**

- 지역 유형별 혼잡도 비교

  지하철 혼잡도는 외국인 관광지, 도심 상업지구, 환승역, 학원가, 업무지구 등 다양한 지역 유형에서 차이를 보였다. 전체 평균 혼잡도는 39.2%였으며, 외국인 관광지(48.5%)와 환승역(45.3%), 업무지구(44.0%)의 혼잡도가 특히 높게 나타났다. 이는 해당 지역들이 관광, 학업, 업무 등 뚜렷한 목적을 지닌 유동 인구가 많은 공간이라는 점에서 기인한 것으로 분석된다. 반면, 대학 인근역(42.1%)과 학원가(40.2%)는 상대적으로 평균에 근접한 수치를 기록하였다.

- 주요 지역별 혼잡도 비교

  이태원, 홍대, 노량진, 대치역과 같이 혼잡할 것으로 예상되었던 특정 지역들의 혼잡도를 실제 데이터와 비교한 결과, 예상과 달리 전체 평균 혼잡도 대비 뚜렷하게 높은 수치는 나타나지 않았다. 이는 사회적 인식이나 선입견이 실제 이용 행태와 다를 수 있음을 보여주는 사례로, 특정 지역의 혼잡도를 단순히 이미지나 평판에 의존해 예측하는 것의 한계를 시사한다.

<3개년 평균 사회적 요인 별 혼잡도 비교>

구분	전체 평균	대학 인근역	관광지	도심 중심가	환승	학원가	업무 지구
혼잡도(%)	39.2%	42.1%	48.5%	41.9%	45.3%	40.2%	44.0%

- 혼잡도 영향 요인별 분석

  사회적 특성에 따른 혼잡도 변화를 종합적으로 분석한 결과, 대학가를 제외한 대부분의 요인이 지하철 혼잡도에 유의미한 영향을 미치는 것으로 나타났다. 특히

외국인 관광지 및 환승역과 같은 공간은 이용 목적이 뚜렷하고 인구 밀집도가 높은 경향을 보여, 지하철 수요 예측 시 주요 변수로 고려될 필요가 있다.

- **기후 요인에 따른 지하철 혼잡도 분석 결과**

- 미세먼지 농도에 따른 승하차 인원 비교

미세먼지 농도가 '좋음'일 때 지하철 이용량이 가장 높았으며, '나쁨'일 때는 평균적으로 약 15% 감소하는 경향을 보였습니다. 이는 대기오염에 대한 우려로 외출을 자제하는 경향이 반영된 결과로 보입니다.

4. 설치 및 사용방법

1) 분석 환경

- 언어 및 라이브러리
  - Python 3.10
- 주요 라이브러리
  - pandas, numpy – 데이터 처리
  - matplotlib, seaborn – 시각화
- 분석 환경
  - Google Colab 또는 Jupyter Notebook 사용
  - OS: Windows 10 / macOS / Ubuntu (개발자 환경에 따라 다름)

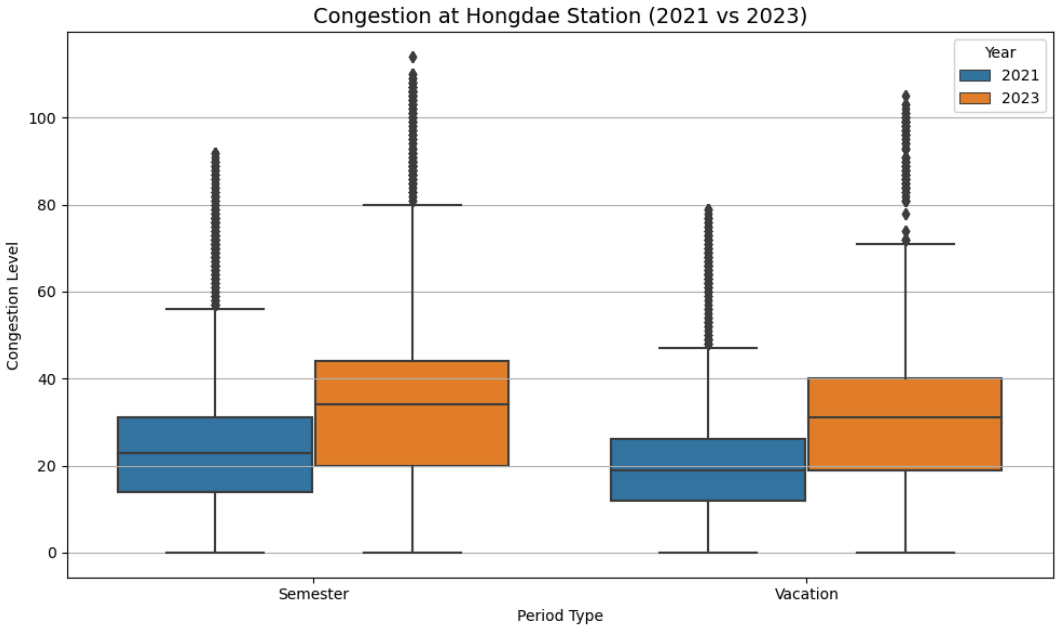
2) 실행 방법

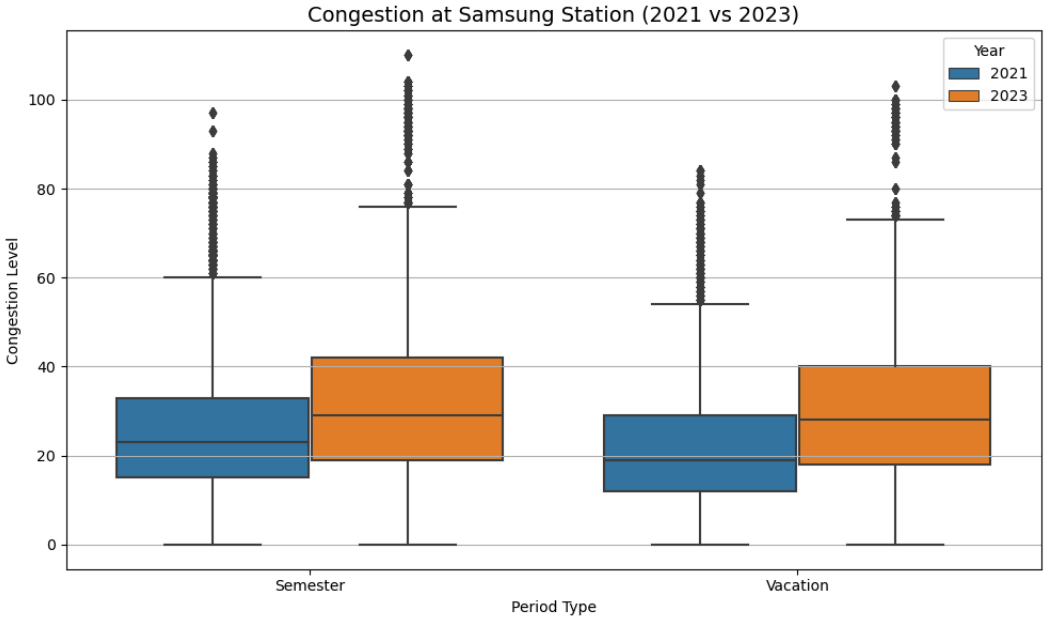
- data/ 폴더에 전처리된 CSV 파일을 위치시킨 후,
- notebooks/ 디렉토리의 분석 노트북(*.ipynb) 파일을 열어 셀을 순차적으로 실행
- 결과 시각화 및 해석은 노트북 내부에서 바로 확인 가능

3) 주의사항

- 데이터 파일 경로가 로컬 환경에 따라 달라질 수 있으므로, 노트북 상의 경로 설정 부분을 사용자 환경에 맞게 수정 필요.

5. 소개 영상 또는 시연영상





※ 위 분석은 전체 역 중 일부 대표적인 역만을 대상으로 한 것이며, 개별 역의 지역적 특성과 유동 인구 특성에 따라 나타난 결과로 해석할 수 있다. 향후 분석 범위를 전체 역으로 확장하면 보다 일반화된 패턴을 도출할 수 있을 것이다.

우선 서울 주요 지하철역 중 두 곳(삼성역, 홍대입구역)을 선정하여, 2021년과 2023년 간의 지하철 혼잡도를 비교하고, 각 연도 내에서 학기 중(Semester)과 방학(Vacation) 기간의 차이를 분석하였다.

두 역 모두에서 공통적으로 나타난 주요 경향은 다음과 같다:

1. **2023년 > 2021년**
    - 혼잡도는 2023년이 2021년에 비해 전반적으로 높게 나타났다.
    - 박스플롯의 중앙값(median), 상위 사분위수(3rd quartile), 이상치(outliers) 등이 모두 2023년에 더 높은 수준을 보였다.
    - 이는 코로나19 거리두기 완화 이후 대중교통 수요가 회복된 영향으로 해석할 수 있다.

2. **학기 중 > 방학**
    - 두 연도 모두에서 학기 중 혼잡도가 방학 기간보다 높았으며, 이는 출퇴근 및 등하교 수요 증가에 따른 결과로 보인다.
    - 특히 삼성역과 홍대입구역 모두에서 이 경향이 더욱 뚜렷하게 나타났다.

역별 특징은 다음과 같다:

- **삼성역 (Samsung Station)**
  2023년 학기 중에는 혼잡도 중앙값이 약 30, 상위 사분위수는 40 이상으로 높은 혼잡도를 보였고, 특히 2021년 방학과 비교했을 때 2023년 방학조차도 뚜렷한 증가를 보였다. 이상치의 개수와 범위도 2023년에 많아져, 특정 일자에 매우 높은 이용률이 있었음을 시사한다. 이는 삼성역이 위치한 업무지구 및 대형 전시장(예: 코엑스) 수요 회복과 밀접한 관련이 있다.

- **홍대입구역 (Hongdae Station)**
  2023년 학기 중 혼잡도 중앙값은 약 35 이상으로 삼성역보다도 높은 수준을 보였다. 2023년 방학 기간조차도 2021년 학기 중보다 높은 혼잡도를 나타냈으며, 이상치 역시 더 자주 관찰되었다. 이는 홍대입구역의 특수성(대학가, 외국인 관광지, 유흥 상권 중심)으로 인해 방학 중에도 지속적인 유동 인구가 존재했으며, 코로나 이후 관광객 유입 증가의 영향을 반영한 것으로 보인다.

6. 팀 소개 (소속, 구성원별 역할)

소속 : 의생명융합공학부 데이터사이언스 전공

김서현, 국서연: 월별, 일별, 시간별, 기후(미세먼지) 등 시간적 요인에 데이터셋 전처리 및 확인

김예지, 변예서 : 데이터 전처리를 통한 사회적 요인에 의한 지하철 혼잡도 유의미 여부 확인

정유민, 안소연 : 날씨 요인에 따른 데이터셋 전처리 및 지하철 혼잡도와 상관관계 존재 여부 확인

7. 참여후기

이번 기상청 공모전에 SW 학습 공동체 팀으로 참여하게 되면서, 실질적인 데이터 분석 경험을 쌓을 수 있는 뜻깊은 기회가 되었습니다. 현재까지 우리는 지하철 혼잡도와 날씨(기온, 강수량, 미세먼지 등) 사이의 상관관계를 파악하기 위해 다양한 방향에서 데이터를 탐색하고 전처리하는 과정을 진행해 왔습니다.

특히 월별, 일별, 시간대별로 데이터를 나누어 시간적 요인이 혼잡도에 어떤 영향을 미치는지 분석하였고, 미세먼지와 같은 기후 요인이 실제로 혼잡도 변화에 유의미한 영향을 미치는지에 대해서도 검토하였습니다. 이를 통해 지하철 혼잡도는 단순히 시간대에 따라 달라지는 것뿐만 아니라, 사회적·환경적 요인들과도 복합적인 관계가 있다는 점을 실감할 수 있었습니다.

데이터셋 전처리는 생각보다 많은 시간과 노력이 요구되었지만, 팀원들과 함께 역할을 나누고 문제를 해결해 나가는 과정에서 협업의 중요성과 데이터 분석의 실제적인 어려움을 체감할 수 있었습니다. 아직은 상관관계 분석을 완료한 단계이지만, 앞으로 이를 바탕으로 보다 정교한 인사이트 도출 및 시각화 작업을 진행할 예정입니다.

공모전을 준비하면서 단순한 분석을 넘어서 실생활과 밀접한 데이터를 다룬다는 점에서 책임감도 느꼈고, 사회적 가치를 고려한 데이터 해석의 중요성도 다시금 느낄 수 있었습니다. 앞으로 남은 기간 동안도 팀원들과 협력하여 의미 있는 결과를 도출하고, 데이터 분석 실무 역량을 한층 더 키울 수 있도록 노력할 것 입니다.

8. 참고문헌 및 출처

- 공모전 주최측 데이터 셋(기상청)
- 서울교통공사_역별 시간대별 승하차인원 (서울 열린데이터 광장)

- 서울시 일별 평균 대기오염도 정보 (서울 열린데이터 광장)