2025년 10월 서울시 지하철 호선별(1호선~9호선), 역별 승하차 인원 정보 및 예측

202108126 조민재

프로젝트 범위

- 데이터 수집,크롤링 -> DATA.GO.KR(공공데이터 포털) 지하철/버스 승하차 데이터 https://data.seoul.go.kr/](https://www.data.go.kr/data/15134550/fileData.do#tab-layer-file)
- 데이터 저장,추출 -> CSV + MYSQL
- 데이터 가공/정제 -> Pandas
- 데이터 분석 -> python + Scikit-learn
- 데이터 시각화 -> matplotlib

데이터 수집 과정

- **데이터 출처**

- 서울 열린데이터 광장(서울시 공공데이터 포털)에서 제공하는 「서울시 지하철 호선별 역별 승하차 인원 정보」 **월별 자료 중 2025년 10월 데이터** (CARD_SUBWAY_MONTH_202510.csv) **를 사용**

- 데이터는 호선별 · 역별 · 일별 승차/하차 인원**이 집계된 형태로 제공**

데이터 전처리 과정

1. 컬럼 구조

-사용일자: 승.하차 인원이 집계된 날짜 (문자열 형식, 예: "20251001")

-노선명: 지하철 노선명 (예: "1호선", "2호선")

-역명: 지하철역 명칭 (예: "강남", "홍대입구")

-승차총승객수: 해당 날짜.역.노선의 **승차 인원 합계**

-하차총승객수: 해당 날짜.역.노선의 **하차 인원 합계**

데이터 전처리 과정

2. 날짜 형식 변환

-초기 데이터의 사용일자는 문자열이므로, 시계열 분석을 위해 datetime 타입으로 변환
- 혹시 모를 공백/기타 문자를 제거하기 위해 **숫자만 추출 후 8자리 날짜로 인식**하는 방식을 사용

```
date_str = (
    df['사용일자']
      .astype(str)
      .str.replace(r'₩D', '', regex=True)  # 숫자가 아닌 것 제거
      .str.slice(0, 8)                     # 앞 8자리: 20251001
)

df['사용일자'] = pd.to_datetime(date_str, format='%Y%m%d', errors='coerce')
```

데이터 전처리 과정

3. **결측치 제거**

- 분석의 기본 단위가 **(날짜, 노선, 역)** 이므로, 이 중 어느 하나라도 없는 행은 제거
- 사용일자가 결측인 행을 제거

```
df_clean = df.dropna(subset=['사용일자'])
print("정제 후 df 크기:", df_clean.shape)
print(df_clean.head())
```

데이터 전처리 과정

4. 파생변수 생성: 총 이용객 수

- 승차와 하차를 합산한 **하루 총 이용객수** 파생 변수 생성
- 날짜별 전체 이용량, 노선별/역별 TOP10, 11월 예측 등 다양한 분석에 사용

```
# 2) 총 이용객수
df['총이용객수'] = df['승차총승객수'] + df['하차총승객수']
```

데이터 변환 과정

1. 날짜별 전체 이용객수 집계

- 2025년 10월 한 달 동안의 **날짜별 전체 승차/하차/총 이용객수**

```
daily = (
    df_clean.groupby('사용일자')[['승차총승객수','하차총승객수','총이용객수']]
            .sum()
            .reset_index()
)
```

데이터 변환 과정

2. 노선별 전체 이용객수 집계

- 노선별 추세를 분석하기 위해 **(날짜, 노선)** 단위로 총 이용객수를 집계

```
# 노선별 날짜별 총 이용객수 집계
line_daily = (
    df_1to9.groupby(['사용일자', '노선명'])['총이용객수']
    .sum()
    .reset_index()
)
```

데이터 변환 과정

3. 역별 전체 이용객수 집계

- 어느 역이 가장 많이 이용되는지 확인하기 위해 **역 단위 합계**를 계산

```
# 노선별 날짜별 총 이용객수 집계
line_daily = (
    df_1to9.groupby(['사용일자', '노선명'])['총이용객수']
    .sum()
    .reset_index()
)
```

데이터 변환 과정

4. 2025년 11월 승·하차 인원 예측

-각 (노선, 역) 조합에 대해 과거 10월 데이터의 `day_idx`(x)와 승차/하차 인원(y)을 사용해 **1차 회귀(추세선)** 를 구하거나,
데이터가 부족하거나 변동이 거의 없을 경우 **평균값**으로 예측치를 설정

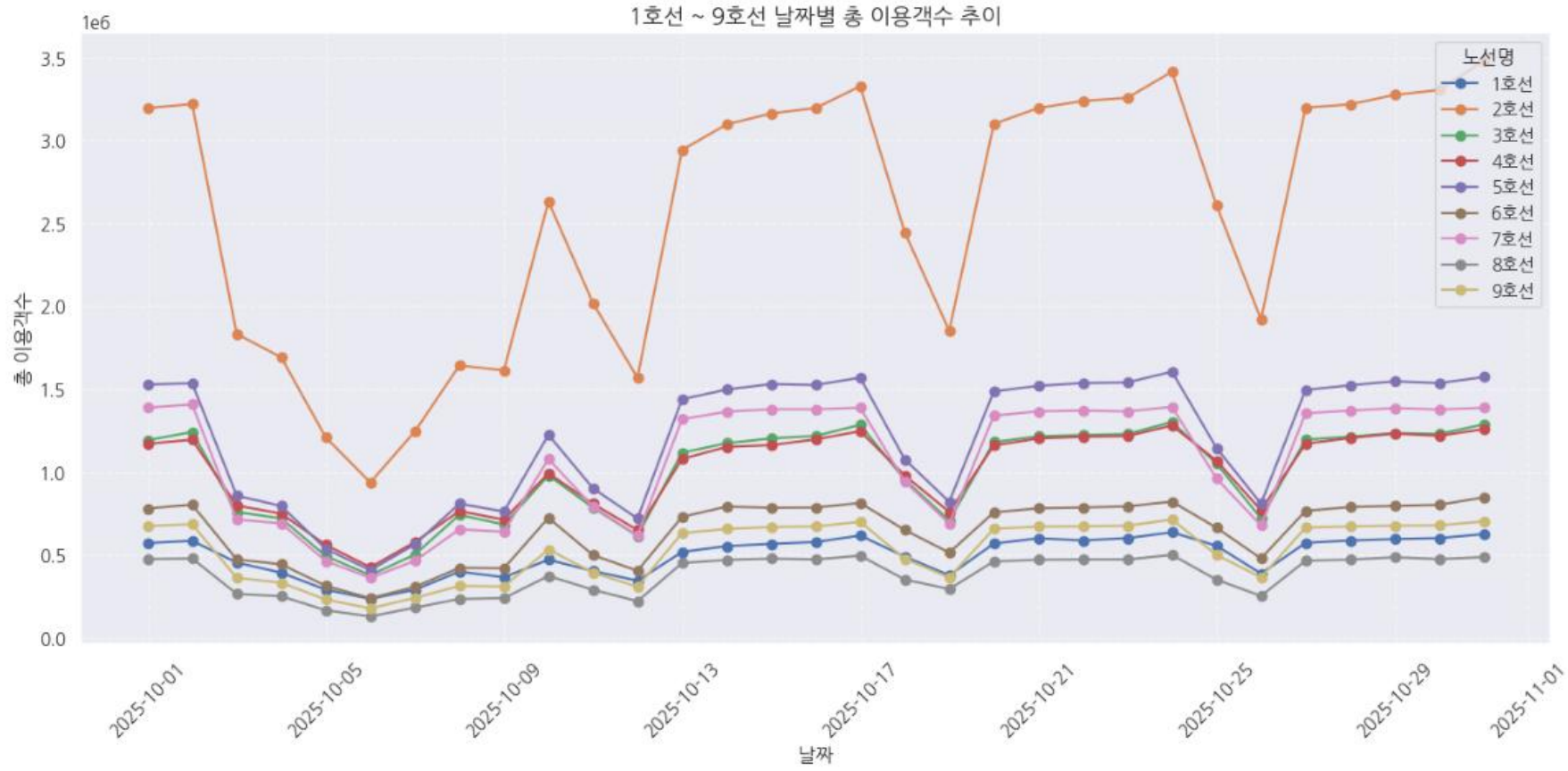
```
pred_rows = []

for (line, station), g in station_daily.groupby(['노선명', '역명']):
    x = g['day_idx'].values.astype(float)

    # 승차/하차 각각에 대해 예측
    for col in ['승차총승객수', '하차총승객수']:
        y = g[col].values.astype(float)

        if len(g) >= 3 and np.std(y) > 0:
            # 데이터가 어느 정도 있으면 1차 회귀 (y = a*x + b)
            a, b = np.polyfit(x, y, 1)
            y_future = a * future_idx + b
        else:
            # 데이터가 너무 적거나 변화가 거의 없으면 평균값 사용
            y_future = np.full_like(future_idx, y.mean())

        # 음수 예측치 방지
        y_future = np.where(y_future < 0, 0, y_future)

        # 결과 쌓기
        for d, idx_val, y_hat in zip(future_dates, future_idx, y_future):
            pred_rows.append({
                '사용일자': d,
                '노선명': line,
                '역명': station,
                '지표': col,          # 승차/하차 구분
                '예측값': float(y_hat)
            })
```

날짜별 지하철 총 이용객수 추세

```
NaN 개수:
사용일자      0
승차총승객수    0
하차총승객수    0
총이용객수     0
dtype: int64
형 변환 후 dtypes:
사용일자      datetime64[ns]
노선명       string[python]
역명         string[python]
승차총승객수              int64
하차총승객수              int64
등록일자                 int64
총이용객수                int64
dtype: object
정제 후 df 크기: (19127, 7)
        사용일자  노선명    역명  승차총승객수  하차총승객수      등록일자  총이용객수
0 2025-10-01  7호선    중계   18030   16820  20251004  34850
1 2025-10-01  1호선   종로3가   27337   24307  20251004  51644
2 2025-10-01  1호선   종로5가   25140   24649  20251004  49789
3 2025-10-01  1호선    동대문   12569   11849  20251004  24418
4 2025-10-01  1호선    신설동   16055   15400  20251004  31455
daily head:
        사용일자    승차총승객수    하차총승객수      총이용객수
0 2025-10-01  8140983  8109833  16250816
1 2025-10-02  8324026  8288301  16612327
2 2025-10-03  4985484  4958004   9943488
3 2025-10-04  4693101  4665185   9358286
4 2025-10-05  3342252  3318936   6661188
```

노선별 날짜별 총 이용객수 추이

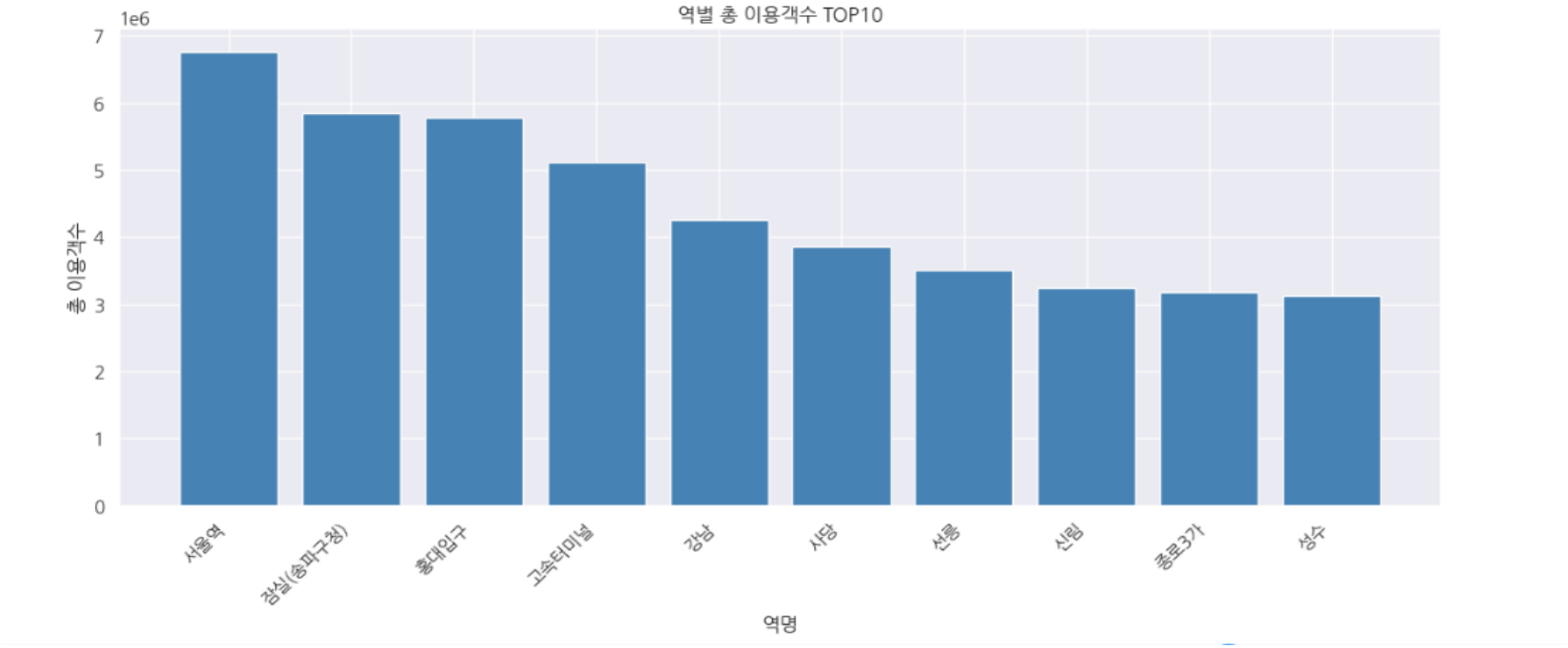
```
필터링된 노선 목록: <StringArray>
['7호선', '1호선', '2호선', '3호선', '4호선', '5호선', '6호선', '8호선', '9호선']
Length: 9, dtype: string
```



역별 이용객수 TOP10

역별 이용객수 TOP10

	역명	총이용객수
251	서울역	6758534
428	잠실(송파구청)	5843573
517	홍대입구	5785338
35	고속터미널	5119536
11	강남	4261408
218	사당	3855676
261	선릉	3506428
309	신림	3241352
445	종로3가	3169350
267	성수	3128438



2025년 11월 서울시 승하차 인원 예측

```
원본 station_daily 예시:
        사용일자  노선명   역명  승차총승객수  하차총승객수  day_idx
0 2025-10-01  1호선  동대문   12569   11849        0
1 2025-10-01  1호선  동묘앞    9759   10188        0
2 2025-10-01  1호선  서울역   73447   73541        0
3 2025-10-01  1호선   시청   31346   31054        0
4 2025-10-01  1호선  신설동   16055   15400        0
=== 2025년 11월 역별/노선별 승하차 인원 예측 예시 (상위 10개) ===
지표       사용일자  노선명             역명        승차총승객수        하차총승객수       총이용객수_예측
0  2025-11-01  1호선            동대문  13859.967742  13330.664516   27190.632258
1  2025-11-01  1호선            동묘앞  11573.251613  11907.612903   23480.864516
2  2025-11-01  1호선            서울역  87697.083871  81507.541935  169204.625806
3  2025-11-01  1호선             시청  33022.058065  33233.619355   66255.677419
4  2025-11-01  1호선            신설동  16766.754839  16205.154839   32971.909677
5  2025-11-01  1호선            제기동  18292.883871  18578.916129   36871.800000
6  2025-11-01  1호선             종각  46502.619355  45609.412903   92112.032258
7  2025-11-01  1호선           종로3가  30003.348387  26718.135484   56721.483871
8  2025-11-01  1호선           종로5가  27932.638710  27544.064516   55476.703226
9  2025-11-01  1호선  청량리(서울시립대입구)  25766.806452  25685.400000   51452.206452
```

역/노선별로 간단한 1차 추세선과 평균을 기반으로 예측