AI데이터 플랫폼을 활용한
빅데이터 분석 전문가 과정

지하철 승차인원 예측

2차 프로젝트

지하철 승차인원 예측

목차

- 프로젝트 개요 및 목적
- 분석과정 및 결과
- 활용방안 및 기대효과

프로젝트 개요 및 목적

- 개요 :

2021년도 서울지하철 연간 이용객 수 약 26억 명, 하루 평균 이용객 약 700만 명 이상. 많은 인원이 이용하는 지하철을 혼잡한 때를 피해 이용하기 위해 승차인원 예측해보고자 함.

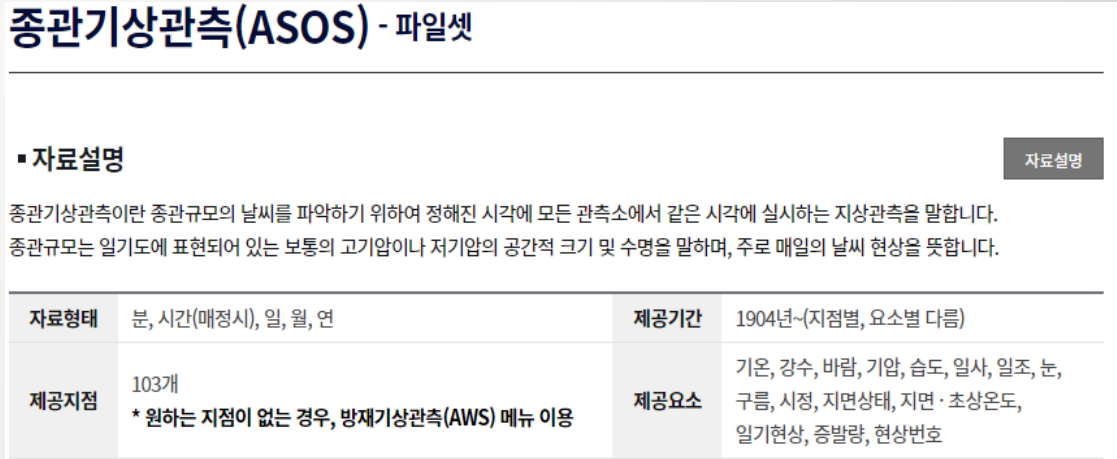
- 목적

탑승할 역에서 실시간 승차인원을 예측하여 혼잡도를 예측.

분석과정 및 결과

지하철 승차인원에 영향을 미칠 수 있는 요인 선정

지하철 승차인원 데이터
- 출처 : 서울 열린데이터 광장



기상 데이터 (기온, 풍속, 강수량)
- 출처 : 기상청 기상자료개방포털

데이터수집 → 전처리 → EDA → 변수 선택 → 모델 선택 → 모델 최적화 → 모델평가

분석과정 및 결과

데이터 불러오기

```
# 수집데이터 불러오기(지하철 승하차 인원)
subway_2017_list = pd.read_csv('E:/Git/data/Academy_Second_Project/passenger_data(2016-2020)
subway_2018_list = pd.read_csv('E:/Git/data/Academy_Second_Project/passenger_data(2016-2020)
subway_2019_list = pd.read_csv('E:/Git/data/Academy_Second_Project/passenger_data(2016-2020)
subway_2020_list = pd.read_csv('E:/Git/data/Academy_Second_Project/passenger_data(2016-2020)
```

✓ 2.3s

```
print(subway_2017_list.shape)
print(subway_2018_list.shape)
print(subway_2019_list.shape)
print(subway_2020_list.shape)
```

✓ 0.8s

```
(200751, 27)
(200751, 26)
(200805, 26)
(202280, 25)
```

```
print(subway_2017_list.columns)
print(subway_2018_list.columns)
print(subway_2019_list.columns)
print(subway_2020_list.columns)
```

✓ 0.2s

```
Index(['Unnamed: 0', 'Unnamed: 1', 'Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4',
       'Unnamed: 5', 'Unnamed: 6', 'Unnamed: 7', 'Unnamed: 8', 'Unnamed: 9',
       'Unnamed: 10', 'Unnamed: 11', 'Unnamed: 12', 'Unnamed: 13',
       'Unnamed: 14', 'Unnamed: 15', 'Unnamed: 16', 'Unnamed: 17',
       'Unnamed: 18', 'Unnamed: 19', 'Unnamed: 20', 'Unnamed: 21',
       'Unnamed: 22', 'Unnamed: 23', 'Unnamed: 24', 'Unnamed: 25', '(단위:명)'],
      dtype='object')
```

2017 ~ 2020 지하철 승차인원 데이터 불러오기

```
# 2017 ~ 2020 지하철 승차인원 데이터 합산
subway_total = subway_2017_list
subway_total = subway_total.append(subway_2018_list)
subway_total = subway_total.append(subway_2019_list)
subway_total = subway_total.append(subway_2020_list)

subway_total.head(5)
```

✓ 0.3s

	날짜	호선	역번호	역명	구분	06:00 이전	06:00 ~ 07:00	07:00 ~ 08:00	08:00 ~ 09:00	09:00 ~ 10:00
1	2017-01-01	1	150	서울역	승차	470	286	397	786	1,421
2	2017-01-01	1	150	서울역	하차	278	880	859	964	1,407
3	2017-01-01	1	151	시청	승차	204	105	112	162	288
4	2017-01-01	1	151	시청	하차	73	203	314	483	669
5	2017-01-01	1	152	종각	승차	791	390	245	270	323

5 rows × 25 columns

Column명 수정 후 데이터 합산

```
weather_2020_list = pd.read_csv('E:/Git/data/Academy_Second_Project/weather_data(

# 기상데이터 불러오기
weather_2020_edit1 = weather_2020_list

# 1차 가공 데이터 (기간)
weather_2020_edit1
```

[25]

	지점	지점명	일시	기온(℃)	풍향(deg)	풍속(m/s)	강수량(mm)
0	116	관악(레)	2020-01-01 01:00	-7.7	251.2	4.8	0.0
1	116	관악(레)	2020-01-01 02:00	-7.5	265.6	5.2	0.0
2	116	관악(레)	2020-01-01 03:00	-7.0	282.3	3.4	0.0
3	116	관악(레)	2020-01-01 04:00	-6.6	277.2	3.8	0.0
4	116	관악(레)	2020-01-01 05:00	-6.1	269.0	4.9	0.0
...	...	...	...	...	...	...	...
244651	889	현충원	2020-12-30 20:00	-10.8	281.8	3.2	0.0
244652	889	현충원	2020-12-30 21:00	-11.5	289.2	2.8	0.0
244653	889	현충원	2020-12-30 22:00	-11.6	280.8	1.7	0.0
244654	889	현충원	2020-12-30 23:00	-11.5	277.4	2.4	0.0
244655	889	현충원	2020-12-31 00:00	-11.3	316.8	1.6	0.0

244656 rows × 7 columns

2020년도 월별 시간별 기상 데이터 (기온, 풍속, 강수량)

데이터수집 | 전처리 | EDA | 변수 선택 | 모델 선택 | 모델 최적화 | 모델평가

분석과정 및 결과

승차인원 데이터 전처리

```
# 2차 가공데이터 (시간별 승차인원 데이터로 형태 변경)
sub_2020_edit2 = pd.DataFrame()

# 시간별 승차인원 데이터
for i in range(len(sub_2020_edit1)) :
    a = sub_2020_edit1.iloc[i]
    a = pd.DataFrame(a)
    a = a.transpose()
    b = pd.melt(a, id_vars=['날짜','호선','역명','역번호','구분'])
    sub_2020_edit2 = sub_2020_edit2.append(b)

sub_2020_edit2.head()
```

날짜 & 시간 데이터 수정

```
# 5차 가공 데이터 (데이터 타입 변경)
sub_2020_edit5 = sub_2020_edit4
# 날짜 columns 타입 변경
sub_2020_edit5['Date'] = sub_2020_edit5['Date'].astype(str)
# 일시 columns 추가 (일시 : 날짜 + 시간)
sub_2020_edit5['Date'] = sub_2020_edit5['Date'] +' '+ sub_2020_edit5['Time']

# Columns 수정
sub_2020_edit5 = sub_2020_edit5[['Date','Line','Location','Sub_station','Passer
sub_2020_edit5 = sub_2020_edit5.reset_index(drop=True)

# 일시 columns 타입 변경
sub_2020_edit5['Date'] = pd.to_datetime(sub_2020_edit5['Date'])

sub_2020_edit5
```

✓ 0.2s

	Date	Line	Location	Sub_station	Passenger_num
0	2020-01-01 06:00:00	2호선	중구	시청	46
1	2020-01-01 07:00:00	2호선	중구	시청	73
2	2020-01-01 08:00:00	2호선	중구	시청	75
3	2020-01-01 09:00:00	2호선	중구	시청	99
4	2020-01-01 10:00:00	2호선	중구	시청	187
...	...	...	...	...	...
365995	2020-12-31 21:00:00	2호선	동대문구	용두(동대문구청)	43
365996	2020-12-31 22:00:00	2호선	동대문구	용두(동대문구청)	33
365997	2020-12-31 23:00:00	2호선	동대문구	용두(동대문구청)	13
365998	2020-12-31 00:00:00	2호선	동대문구	용두(동대문구청)	3
365999	2020-12-31 01:00:00	2호선	동대문구	용두(동대문구청)	0

366000 rows × 5 columns

시간 데이터 수정

```
# 3차 가공 데이터에 위치 정보 생성
location = []
for i in sub_2020_edit3['역명'] :
    location.append(sub_location[i])
sub_2020_edit3['위치'] = location
```

✓ 0.1s

위치 데이터 생성

```
# 4차 가공 데이터 (Column 편집)
sub_2020_edit4 = sub_2020_edit3[['날짜','variable','호선','위치','역명','value']]
sub_2020_edit4 = sub_2020_edit4.rename(columns = {'날짜' : 'Date',
                                                  'variable' : 'Time',
                                                  '호선' : 'Line',
                                                  '위치' : 'Location',
                                                  '역명' : 'Sub_station',
                                                  'value' : 'Passenger_num'})
sub_2020_edit4 = sub_2020_edit4.reset_index(drop=True)

sub_2020_edit4.head(2)
```

✓ 0.8s

	Date	Time	Line	Location	Sub_station	Passenger_num
0	2020-01-01	06:00	2호선	중구	시청	46
1	2020-01-01	07:00	2호선	중구	시청	73

Column명 수정

분석과정 및 결과

기상 데이터 전처리

```
# null값 확인
print(weather_2020_edit1.isnull().sum())
print('----------------------')

# 2차 가공 데이터 (null값 치환)
weather_2020_edit2 = weather_2020_edit1

# 풍향 column 사용 안함 (null값 무시)
# 강수량 null값은 0으로 대체
weather_2020_edit2['강수량(mm)'] = np.where(pd.notnull(weather_2020_ed
# 기온과 풍속의 null값은 위/아래값으로 대체
weather_2020_edit2 = weather_2020_edit2.fillna(method = 'ffill')
# null값 재확인
print(weather_2020_edit2.isnull().sum())
```

```
지점           0
지점명          0
일시           0
기온(°C)      591
풍향(deg)     812
풍속(m/s)     760
강수량(mm)      0
dtype: int64
----------------------
지점           0
지점명          0
일시           0
기온(°C)       0
풍향(deg)      0
풍속(m/s)      0
강수량(mm)     0
dtype: int64
```

이상치 처리

```
# 3차 가공 데이터 (columns 수정)
weather_2020_edit3 = weather_2020_edit2
weather_2020_edit3 = weather_2020_edit3[['일시','지점명','기온(°C)','풍속(m/s)','강수

weather_2020_edit3['일시'] = pd.to_datetime(weather_2020_edit3['일시'])
weather_2020_edit3 = weather_2020_edit3.rename(columns = {'날짜' : 'Date',
                                                          '일시' : 'Date',
                                                          '지점명' : 'Location',
                                                          '기온(°C)' : 'Temp',
                                                          '풍속(m/s)' : 'Wind',
                                                          '강수량(mm)' : 'Rain'})
weather_2020_edit3 = weather_2020_edit3.reset_index(drop=True)

weather_2020_edit3
```

	Date	Location	Temp	Wind	Rain
0	2020-01-01 01:00:00	관악(레)	-7.7	4.8	0.0
1	2020-01-01 02:00:00	관악(레)	-7.5	5.2	0.0
2	2020-01-01 03:00:00	관악(레)	-7.0	3.4	0.0
3	2020-01-01 04:00:00	관악(레)	-6.6	3.8	0.0
4	2020-01-01 05:00:00	관악(레)	-6.1	4.9	0.0
...	...	...	...	...	...

Column명 수정

분석과정 및 결과

데이터 합치기

```
total_data = pd.merge(sub_2020_edit5, weather_2020_edit4, how='left', on=None )

total_data
```

✓ 0.1s

	Date	Line	Location	Sub_station	Passenger_num	Temp	Wind	Rain
0	2020-01-01 06:00:00	2호선	중구	시청	46	-6.7	1.5	0.0
1	2020-01-01 07:00:00	2호선	중구	시청	73	-6.6	2.5	0.0
2	2020-01-01 08:00:00	2호선	중구	시청	75	-6.2	2.1	0.0
3	2020-01-01 09:00:00	2호선	중구	시청	99	-6.1	2.2	0.0
4	2020-01-01 10:00:00	2호선	중구	시청	187	-6.2	2.3	0.0
...	...	...	...	...	...	...	...	...
365995	2020-12-31 21:00:00	2호선	동대문구	용두(동대문구청)	43	NaN	NaN	NaN
365996	2020-12-31 22:00:00	2호선	동대문구	용두(동대문구청)	33	NaN	NaN	NaN
365997	2020-12-31 23:00:00	2호선	동대문구	용두(동대문구청)	13	NaN	NaN	NaN
365998	2020-12-31 00:00:00	2호선	동대문구	용두(동대문구청)	3	-10.9	1.0	0.0
365999	2020-12-31 01:00:00	2호선	동대문구	용두(동대문구청)	0	NaN	NaN	NaN

366000 rows × 8 columns

지하철 승차인원 데이터와 기상데이터

```
total_data.dropna(axis = 0, inplace = True)
total_data
```

✓ 0.1s

	Date	Line	Location	Sub_station	Passenger_num	Temp	Wind	Rain
0	2020-01-01 06:00:00	2호선	중구	시청	46	-6.7	1.5	0.0
1	2020-01-01 07:00:00	2호선	중구	시청	73	-6.6	2.5	0.0
2	2020-01-01 08:00:00	2호선	중구	시청	75	-6.2	2.1	0.0
3	2020-01-01 09:00:00	2호선	중구	시청	99	-6.1	2.2	0.0
4	2020-01-01 10:00:00	2호선	중구	시청	187	-6.2	2.3	0.0
...	...	...	...	...	...	...	...	...
365918	2020-12-31 00:00:00	2호선	동대문구	신설동	10	-10.9	1.0	0.0
365938	2020-12-31 00:00:00	2호선	구로구	도림천	3	-11.6	2.9	0.0
365958	2020-12-31 00:00:00	2호선	양천구	양천구청	19	-11.3	1.5	0.0
365978	2020-12-31 00:00:00	2호선	양천구	신정네거리	9	-11.3	1.5	0.0
365998	2020-12-31 00:00:00	2호선	동대문구	용두(동대문구청)	3	-10.9	1.0	0.0

364216 rows × 8 columns

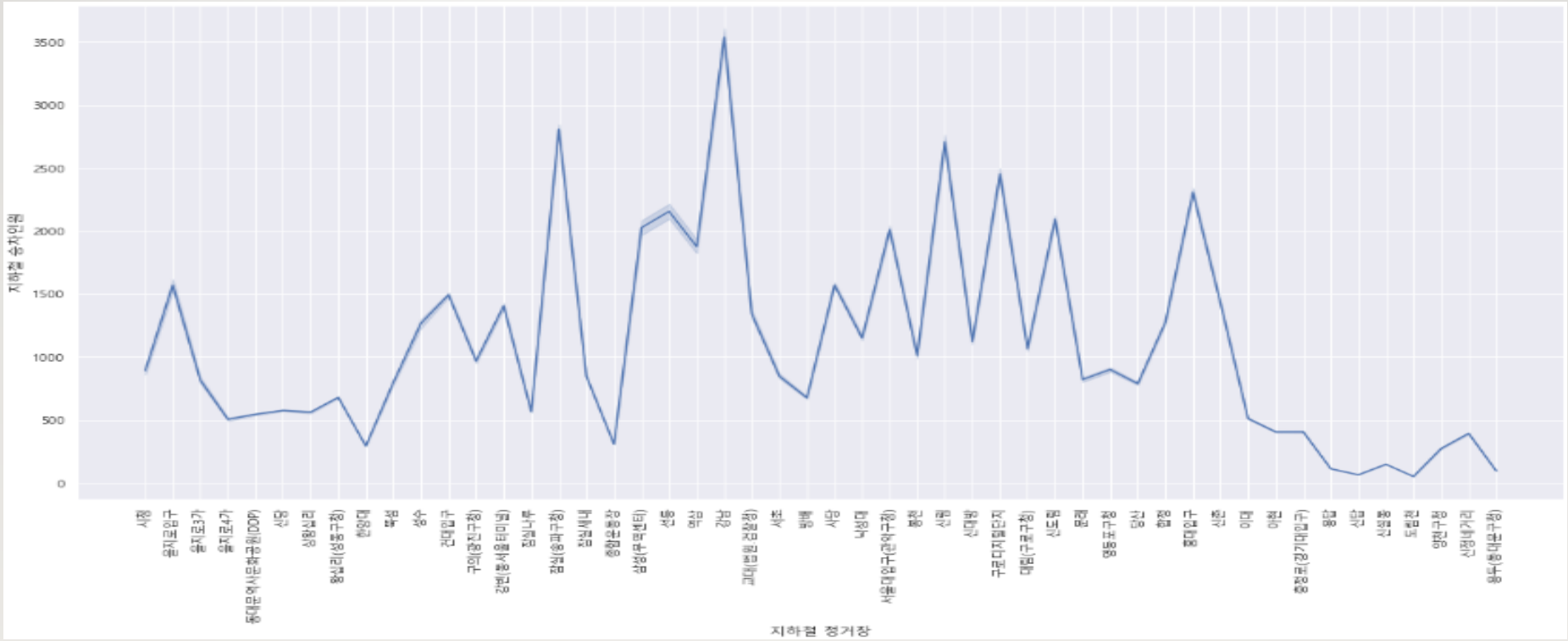
```
total_data.isnull().sum()
```

✓ 0.1s

```
Date           0
Line           0
Location       0
Sub_station    0
Passenger_num  0
Temp           0
Wind           0
Rain           0
dtype: int64
```

이상치 확인 및 제거

데이터 시각화 및 EDA



2020년도 2호선 지하철 이용객 추세 => 역마다 이용객 편차가 큼