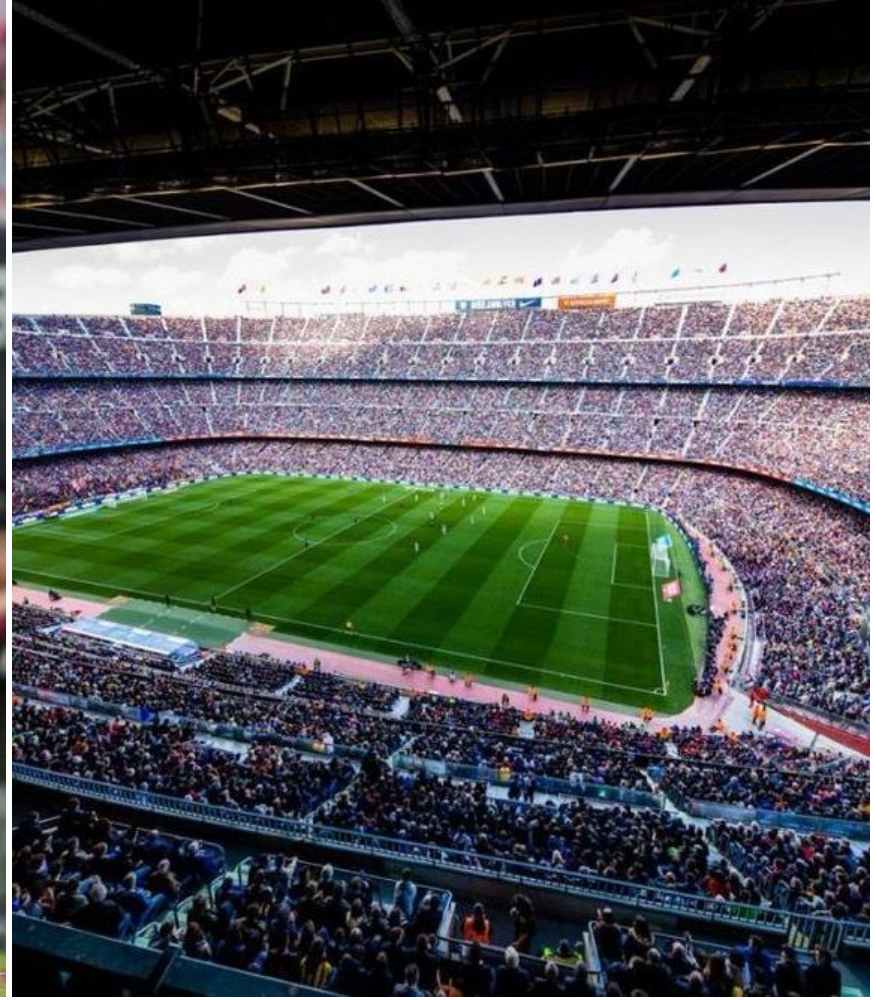




머신러닝 기반 프로스포츠 관중수 예측모델 개발 및 분석

태양의눈 and (이종범)

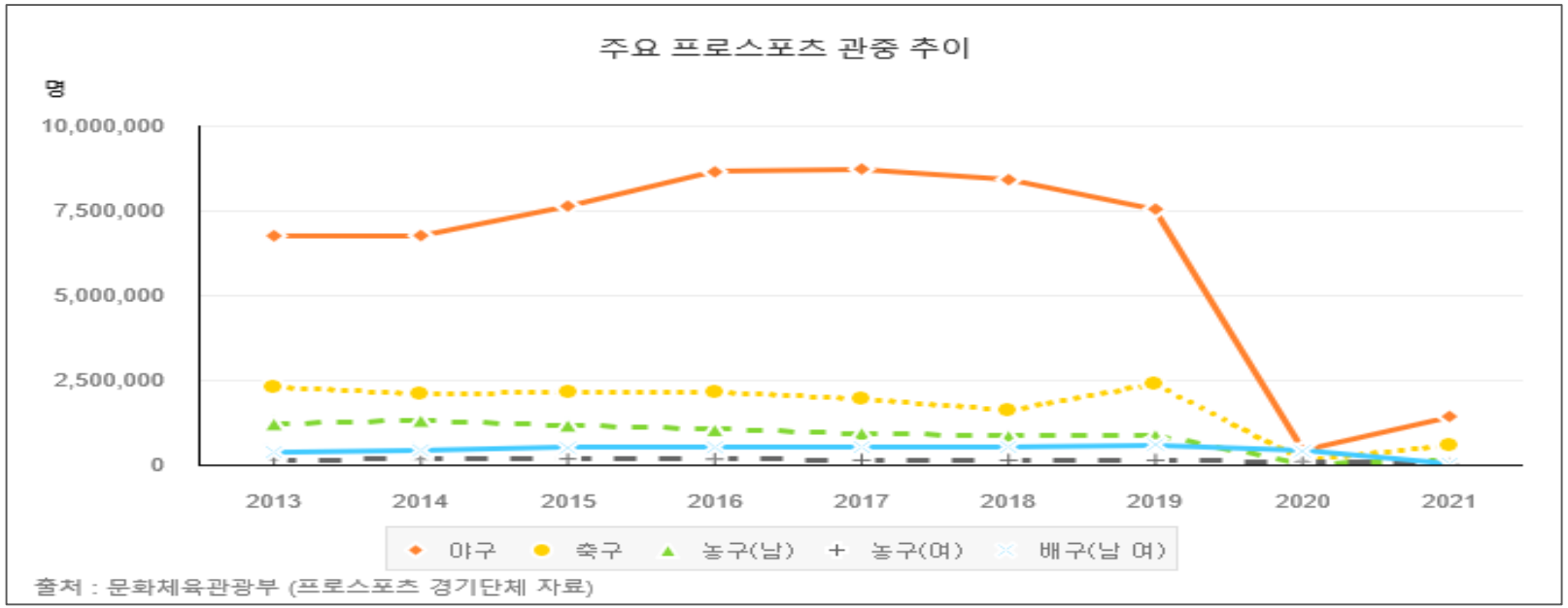
목차

I. 프로젝트 배경

1. 필요성 및 목적

- 기존 통계중심 예측에서 벗어나, 인공지능 기술(선형회귀분석, 랜덤포레스트)을 통한 프로 스프츠 종목의 일별 관중수 예측을 하고자 한다.

- 코로나 19 발생으로 인한 무관중 경기를 감안하여 코로나 이전과 이후의 관중 모델을 예측해, 향후 프로스포츠 관중 증가를 위한 전략을 세우는데 목적으로 한다.



I. 프로젝트 배경

2. 개발 환경

Python 3.10.4

Google Colab

Numpy

Scikit-learn

seaborn

Matplotlib

Ⅱ. 프로젝트 수행

1. 데이터 수집

- **기상청 : 기상자료개방포털 동네날씨예보 데이터 2018~2022년도 지역별 평균기온, 날씨상태 자료 수집**
(https://data.kma.go.kr/data/rmt/rmtList.do?code=400&pgmNo=570)

- **한국프로스포츠협회 :** 2018년~2022년 7월 프로스포츠 정보광장 포털에서 **일별 전체관중현황 수집**
(http://data.prosports.or.kr/spectator/m0201/search)

- 경기장 수용인원 수집을 위해 **구글 검색** 활용

- **네이버 스포츠뉴스**: 2018~2022 각 종목별 **경기결과 및 시간대, 일별 전체관중현황에서 누락된 경기 데이터 수집**
(https://sports.news.naver.com/kfootball/schedule/index)

Ⅱ. 프로젝트 수행

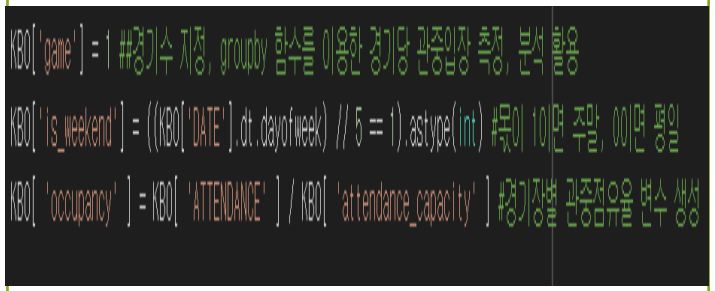
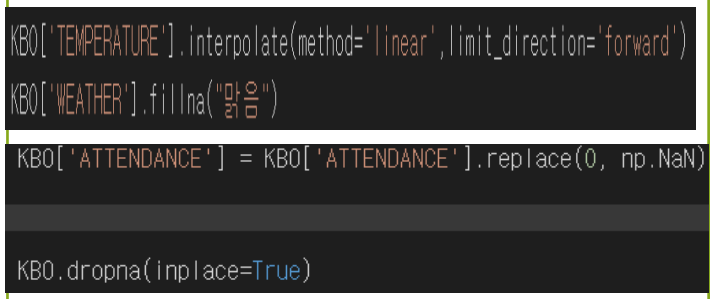
2. RAW DATA 전처리(1)

	전처리 내용	코드
날짜, 문자열 변환	1.원 데이터에서 연-월-일로 구분된 컬럼을 합친 후, **날짜형** 변환 2.기온 데이터의 ℃를 문자열에서 정수로 변환	`WOMAN_KOVO['DATE'] = pd.to_datetime(WOMAN_KOVO['DATE'])` `WOMAN_KOVO['TEMPERATURE'] = WOMAN_KOVO['TEMPERATURE'].astype(float)`
특수문자	1. 기온 컬럼을 **실수형으로 변환**하기 위해 문자열 제거	`KOVO['TEMPERATURE'] = KOVO['TEMPERATURE'].str.replace(pat=r'[℃]', repl=r' ', regex=True) # replace multiple spaces with a single space(공백 없애)`

가변수 생성	1.범주형 컬럼을 머신러닝 모델 구축을 위해 원핫인코딩 및 라벨인코딩 진행 후 기존 컬럼과 병합	`team_dummies=pd.get_dummies(KLEAGUE['hometeam'])`
		`weather_dummies=pd.get_dummies(KLEAGUE['WEATHER'])`
		`KLEAGUE=pd.concat([KLEAGUE,team_dummies,weather_dummies], axis=1)`

II. 프로젝트 수행

2. RAW DATA 전처리(2)

	전처리 내용	코드
파생변수 추가	1.각 경기장 수용인원 컬럼 추가 후, 일일 관중 입장 수를 나누어 관중 점유율 컬럼 생성 2. Pandas의 Groupby 함수를 이용해 경기당 관중 데이터 칼럼을 생성하기 위한 경기수 칼럼 생성 3. 날짜형 데이터를 주말, 평일 구분 컬럼 생성	
결측치 제거	1.기온, 날씨의 결측치 확인 후, 평균값 변환 및 기상청 날씨 데이터를 활용해 일일 평균기온, 날씨 입력 2. 2020-2021년 무관중 경기 결측치 변환 후 제거	

II. 프로젝트 수행

3. 예측 분석 (EDA)

1) 현황 _ 스포츠 종목별 전체 관중수

- 2018~2022년도 전체 총 관중수 분석
- 종목별로 전체 관중수 분석
- 코로나 이전과 이후 관중수 비교 분석

```
#2018~2022.7월 종목별 전체관중 현황 EDA

figure, ((ax1,ax2,ax3),(ax4,ax5,ax6)) = plt.subplots(nrows=2, ncols=3)
figure.set_size_inches(23.5,20)

sns.barplot(data=KBO, x="year", y="ATTENDANCE", estimator=sum, ax=ax1)
sns.barplot(data=KLEAGUE, x="year", y="ATTENDANCE", estimator=sum, ax=ax2)
sns.barplot(data=KBL, x="SEASON", y="ATTENDANCE", estimator=sum, ax=ax3)
sns.barplot(data=WOMAN_KOVO, x="SEASON", y="ATTENDANCE",estimator=sum,ax=ax4)
sns.barplot(data=MAN_KOVO, x="SEASON", y="ATTENDANCE",estimator=sum,ax=ax5)
sns.barplot(data=WKBL, x="SEASON", y="ATTENDANCE",estimator=sum,ax=ax6)

ax1.set(xlabel='연도',ylabel='관중수', title="프로야구 2018-2022 총 관중수")
ax2.set(xlabel='연도',ylabel='관중수',title="프로축구 2018-2022 총 관중수")
ax3.set(xlabel='시즌',ylabel='관중수',title="프로농구 2018/19-2021/2022시즌 총 관중수")
ax4.set(xlabel='시즌',ylabel='관중수', title="여자프로배구 2018/19-2021/22시즌 총 관중수")
ax5.set(xlabel='시즌',ylabel='관중수', title="남자프로배구 2018/19-2021/22시즌 총 관중수")
ax6.set(xlabel='시즌',ylabel='관중수', title="여자프로농구 2018/19-2021/22시즌 총 관중수")
```

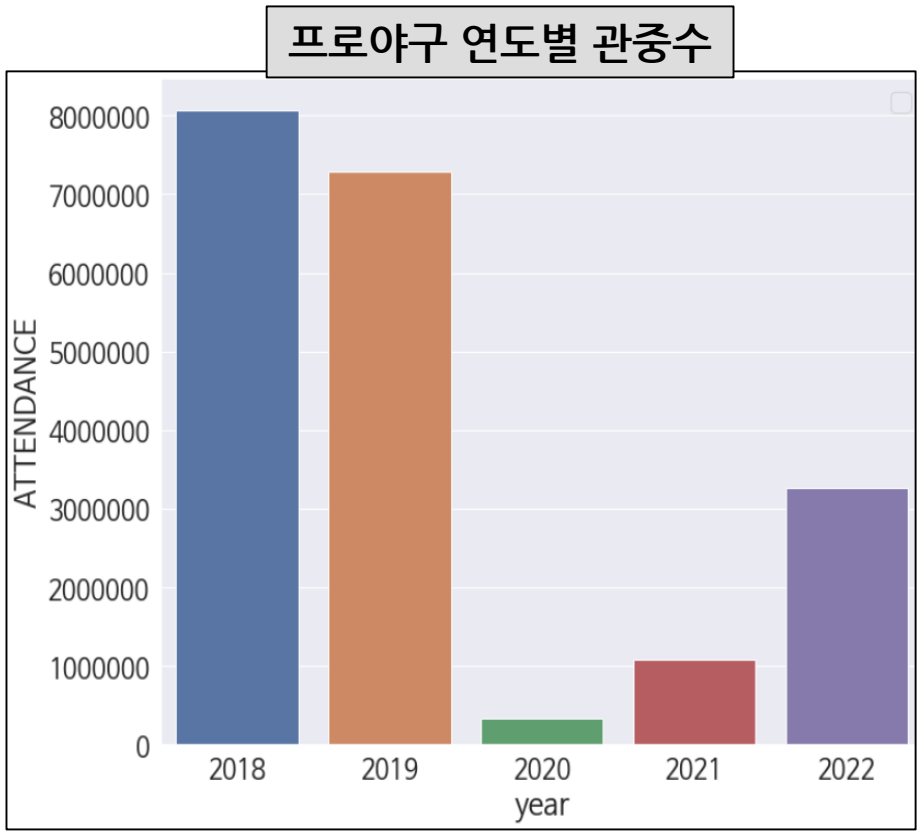
[분석결과]

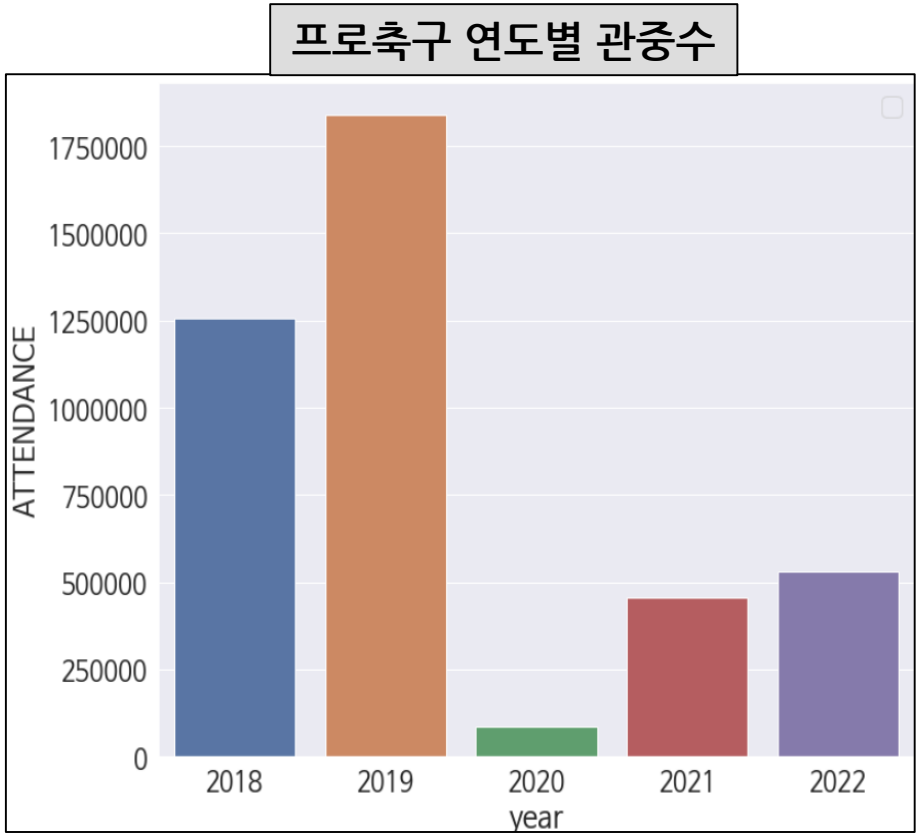
- 프로야구는 2018년 800만대를 돌파한 후, 지속적으로 감소 추세
- 프로축구(K리그) 역시 2019년 180만명을 돌파한 후 코로나 19 이슈와 맞물려, 관중입장 제한으로 인한 감소 추세
- 프로농구는 2018-2019 시즌 80만명대 관중 동원을 보여준 후, 2020-2021 시즌 코로나 19 여파로 10만명대로 낮아짐
- 여자프로농구는 2020-2021년 정규시즌 전경기 무관중 진행으로 해당 시즌 관중수 0을 기록
- 프로배구 남자부, 여자부 관중의 경우, 2021-22 시즌에 여자부 전체관중이 남자부 전체관중을 넘어섬
- 전체 종목 모두 코로나 19로 인해 2020년 경기들의 관중 수가 감소했음

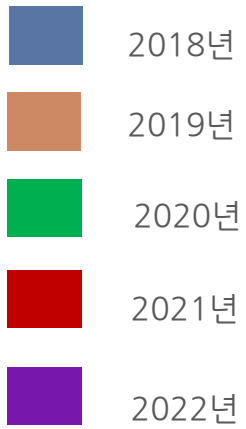
Ⅱ. 프로젝트 수행

3. 예측 분석 (EDA)

2) 시각화 _ 스포츠 종목별 전체 관중수(1)



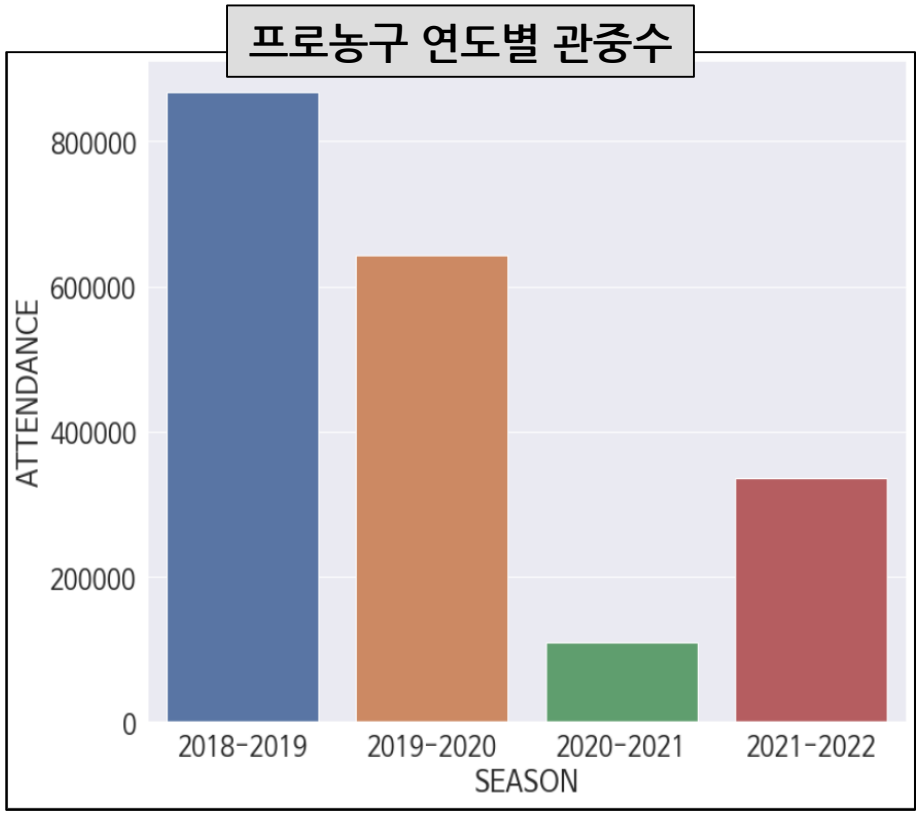


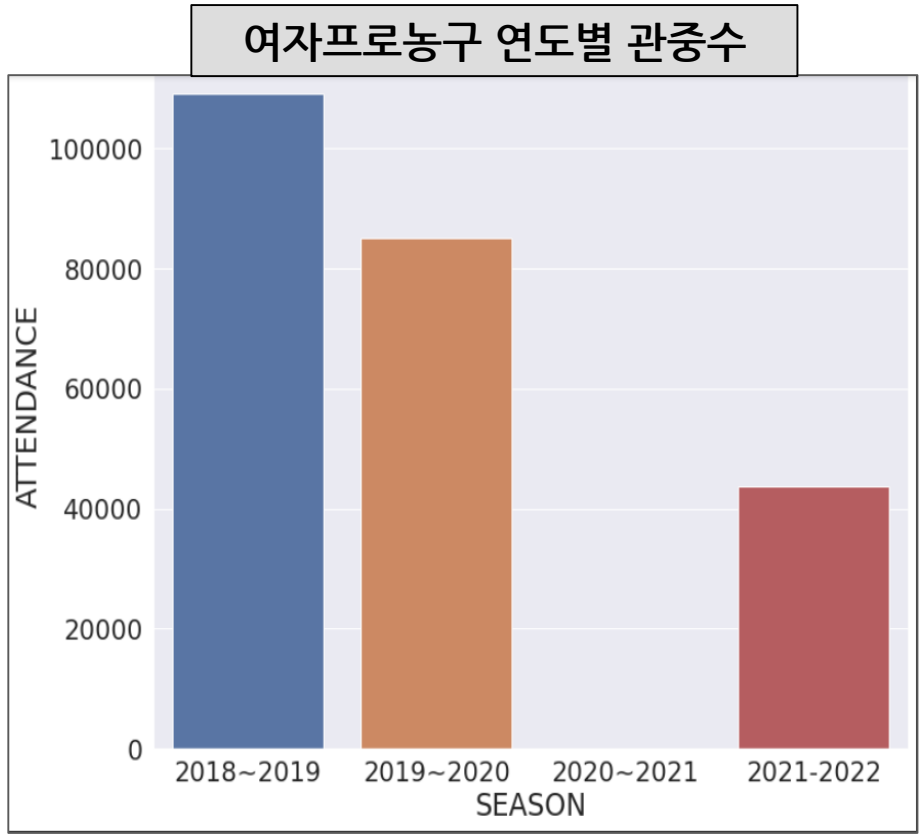


Ⅱ. 프로젝트 수행

3. 예측 분석 (EDA)

3) 시각화 _ 스포츠 종목별 전체 관중수(2)



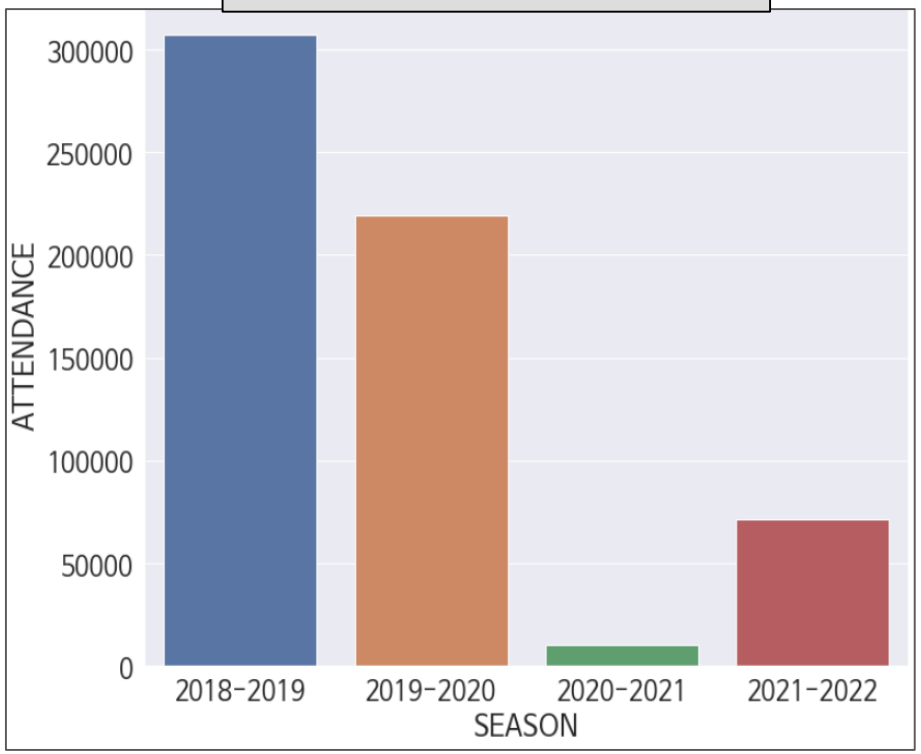


Ⅱ. 프로젝트 수행

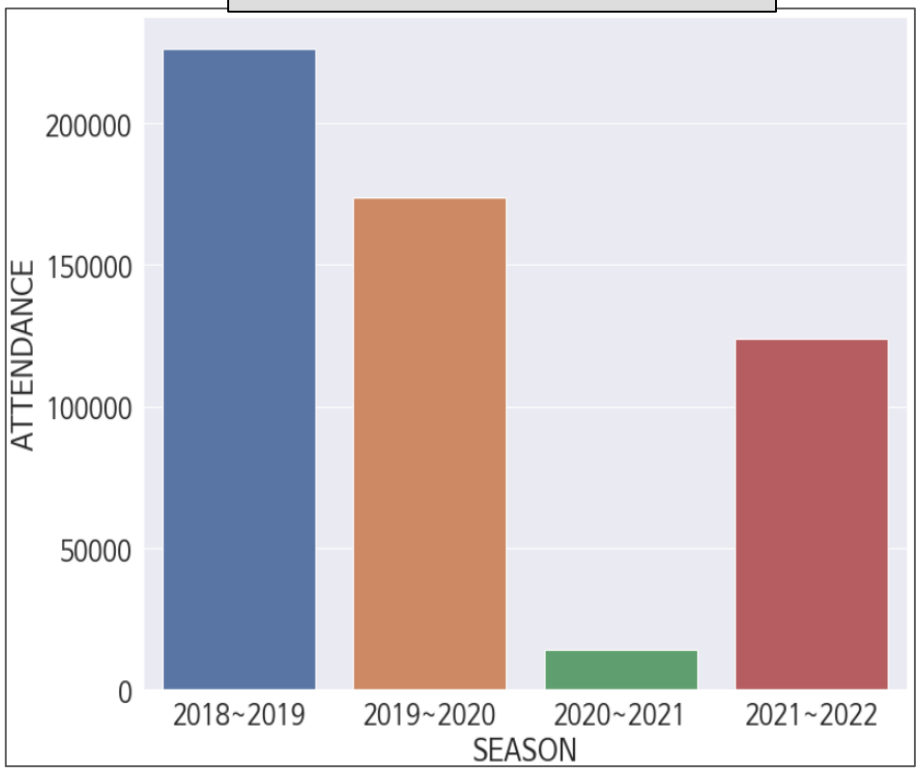
3. 예측 분석 (EDA)

4) 시각화 _ 스포츠 종목별 전체 관중수(3)

남자프로배구 연도별 관중수



여자프로배구 연도별 관중수



II. 프로젝트 수행

3. 예측 분석 (EDA)

5) 현황_요일별 스포츠 종목별 관중수

- 각 종목별로 요일별 입장 수 분석

[분석결과]

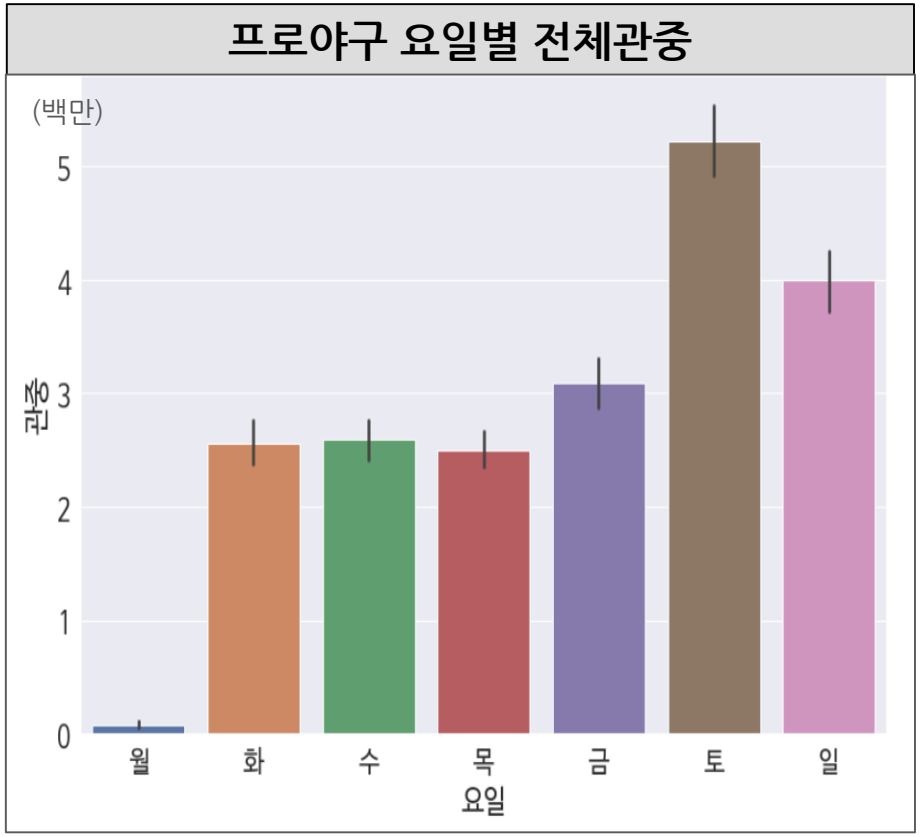
- 모든 종목들이 토, 일요일 관중이 타 요일보다 많았음
- 매일 진행되는 종목일수록
- 여자프로농구의 경우 월요일 입장이 토, 일요일 관중과 비슷한 추세를 보임

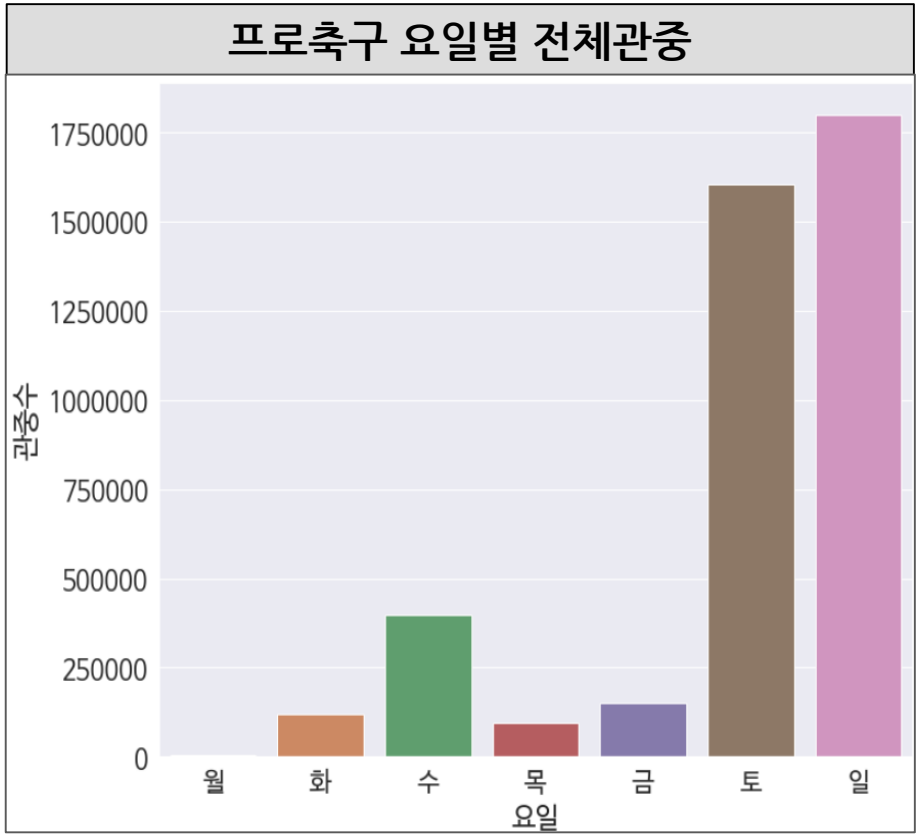
```
[42] #종합적인 요일별 관중입장
     figure, ((ax1,ax2), (ax3,ax4), (ax5,ax6)) = plt.subplots(nrows=3, ncols=2)
     figure.set_size_inches(30,30)

     sns.barplot(data=KBO, x="dayofweek", y="ATTENDANCE", ax=ax1)
     sns.barplot(data=KLEAGUE, x="dayofweek", y="ATTENDANCE", ax=ax2)
     sns.barplot(data=KBL, x="dayofweek", y="ATTENDANCE", ax=ax3)
     sns.barplot(data=WOMAN_KOVO, x="dayofweek", y="ATTENDANCE",ax=ax4)
     sns.barplot(data=MAN_KOVO, x="dayofweek", y="ATTENDANCE",ax=ax5)
     sns.barplot(data=WKBL, x="dayofweek", y="ATTENDANCE",ax=ax6)

     ax1.set(xlabel='요일',ylabel='관중', title="프로야구 요일별 평균관중",)
     ax2.set(xlabel='요일',ylabel='관중',title="프로축구 요일별 평균관중")
     ax3.set(xlabel='요일',ylabel='관중',title="프로농구 요일별 평균관중")
     ax4.set(xlabel='요일',ylabel='관중', title="프로배구(여자) 요일별 평균관중")
     ax5.set(xlabel='요일',ylabel='관중', title="프로배구(남자) 요일별 평균관중")
     ax6.set(xlabel='요일',ylabel='관중', title="프로농구(여자) 요일별 평균관중")
```

Ⅱ. 프로젝트 수행

3. 예측 분석 (EDA)

6) 시각화_요일별 스포츠 종목별 관중수(1)

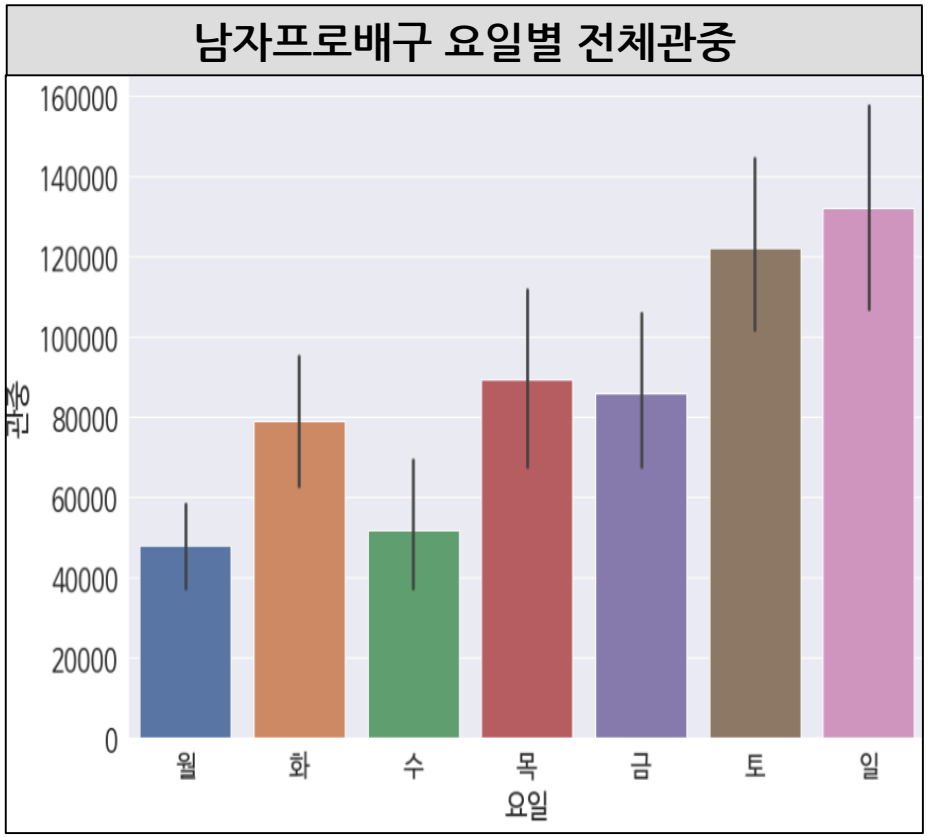


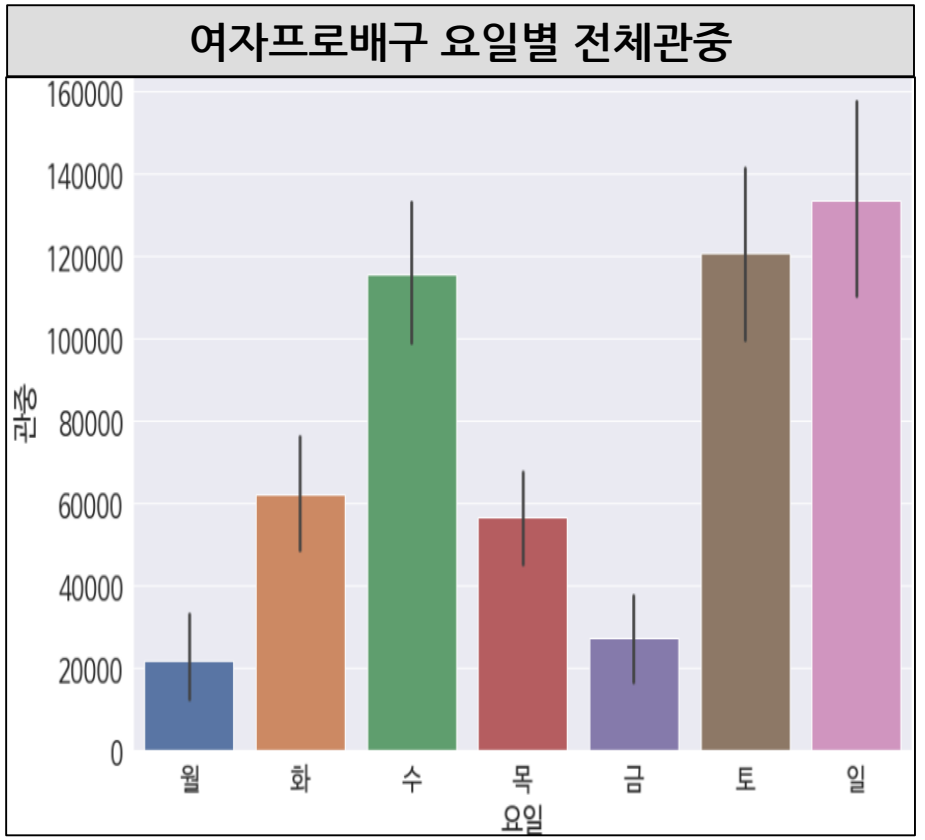


II. 프로젝트 수행

3. 예측 분석 (EDA)

7) 시각화_요일별 스포츠 종목별 관중수(2)

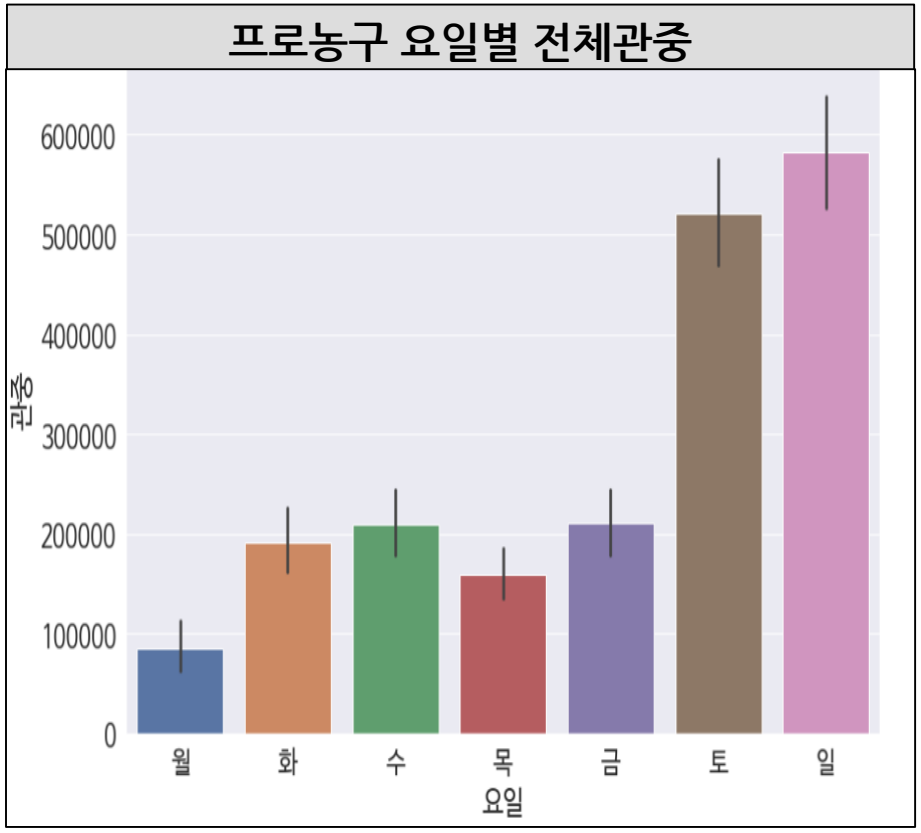


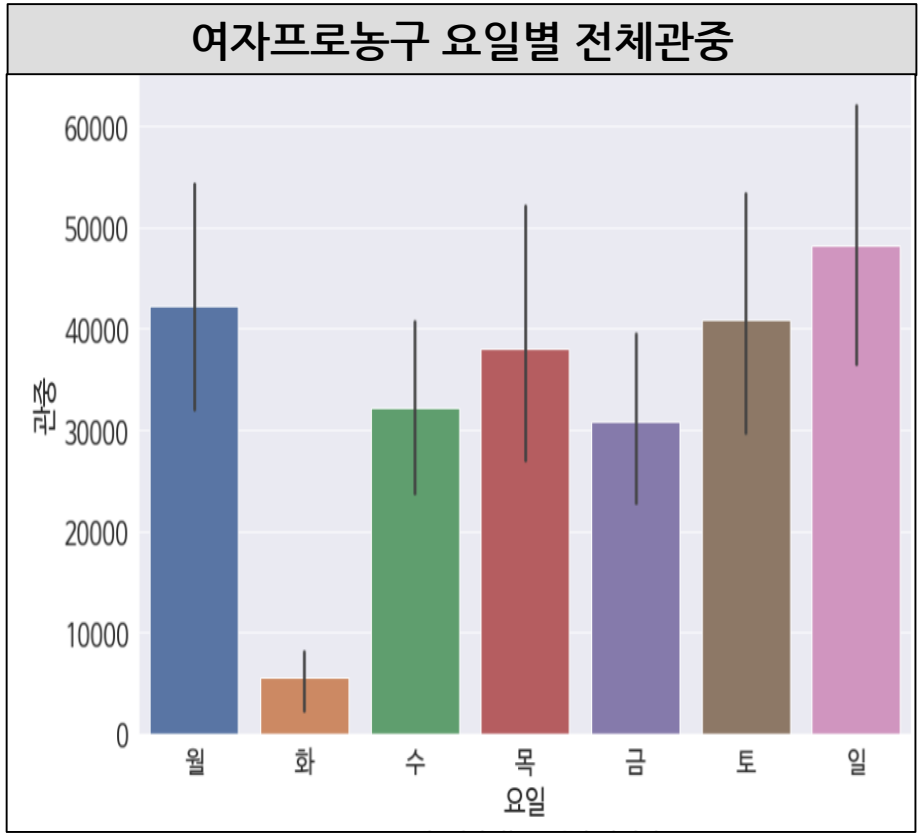


II. 프로젝트 수행

3. 예측 분석 (EDA)

8) 시각화_요일별 스포츠 종목별 관중수(3)





Ⅱ. 프로젝트 수행

3. 예측 분석 (EDA)

9) 현황_프로축구, 프로야구 관중 상관관계

- 2018년에서 2022년까지 프로야구와 프로축구의 관중 상관관계를 히트맵으로 시각화

[분석 결과]

- 프로야구의 관중 상관관계는 관중점유율, 주말/평일 분류 변수, 요일순으로 상관관계가 높았음

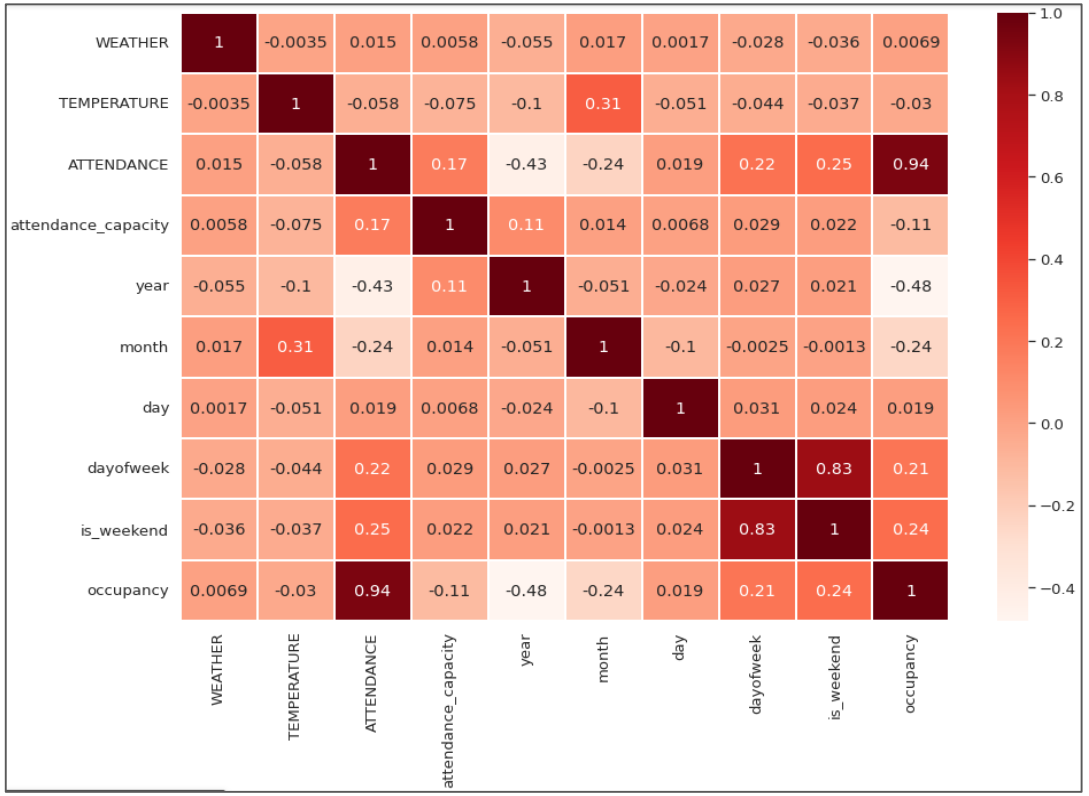
- 프로축구의 관중 상관관계는 관중점유율, 경기장 수용인원, 요일 순으로 상관관계가 높았음

II. 프로젝트 수행

3. 예측 분석 (EDA)

10) 시각화_프로축구, 프로야구 관중 상관관계

- 프로야구 관중 상관관계: 관중점유율>평일/주말구분>요일(관중수가 많을수록 경기 당 매진이 증가)
- 프로축구 관중 상관관계: 관중점유율> 경기장 수용인원>요일



프로야구 관중 상관관계 히트맵

프로축구 관중 상관관계 히트맵

II. 프로젝트 수행

3. 예측 분석 (EDA)

11) 현황_ 프로야구와 프로축구 기온과 관중 입장 상관관계 분석

- 대표적인 프로스포츠이자 야외 종목인 프로야구, 프로축구에서 기온과 관중 입장 상관관계 분석

[분석 결과]

- 기온이 높을수록, 관중 수 밀집도 높음
- 기온과 관중 수의 상관관계는 두 종목 모두 약한 상관관계를 보여줌

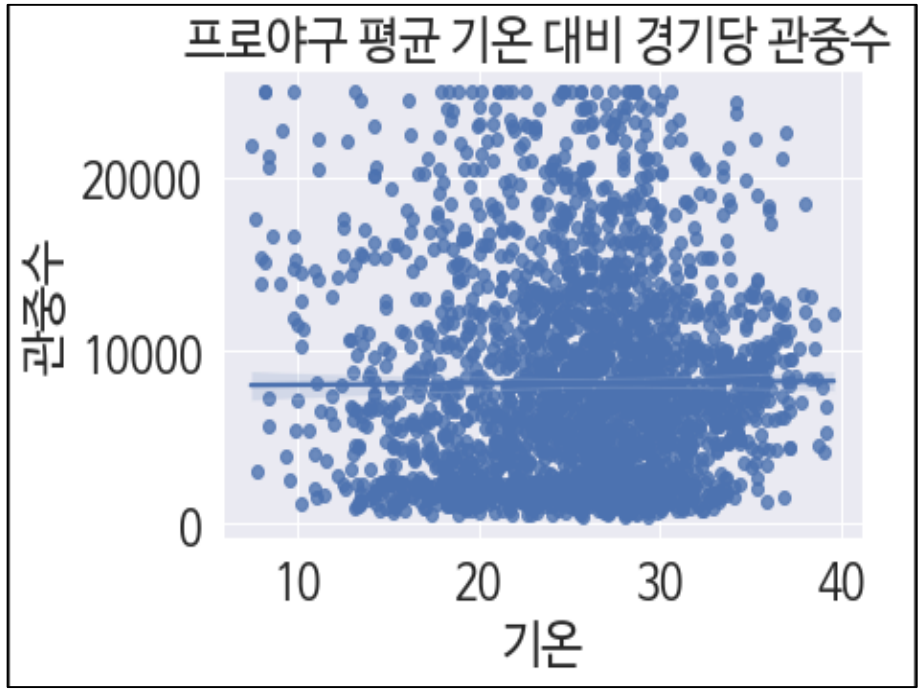
```
[87]  #기온과 관중수간 선형분석
      plt.rc('font',family='NanumBarunGothic')
      ax = sns.regplot(x="TEMPERATURE", y="ATTENDANCE", data = KBO)
      sns.set(font_scale=1.5)


      plt.xlabel('기온')
      plt.ylabel('관중수')
```

II. 프로젝트 수행

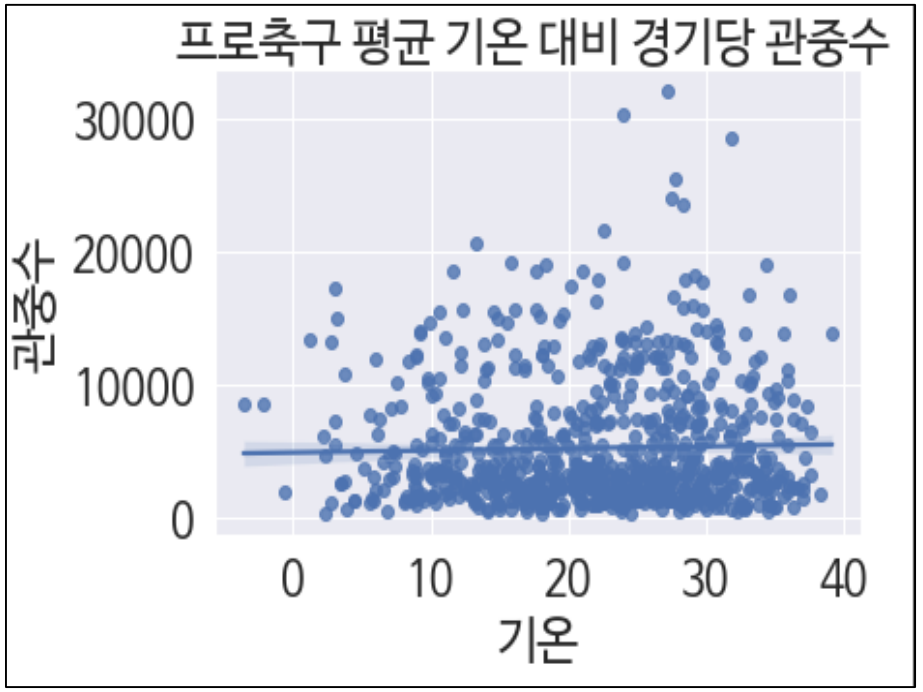
3. 예측 분석 (EDA)

12) 시각화_ 프로야구와 프로축구 기온과 관중 입장 상관관계 분석

- 기온이 높을수록, 관중 수 밀집도 높음
- 기온과 관중 수의 상관관계는 약한 상관관계를 보여줌



프로야구 경기당 관중과 기온 간
상관관계 분석



프로축구 경기당 관중과 기온 간
상관관계 분석

III. 모델링

1. 모델 개발

1) 프로야구 관중수 예측 선형회귀모델 개발

- 2018~2022년 일일 관중수를 테스트세트와 훈련세트 분리(7:3)
- 평가지표를 MAE(평균절대오차)와 RMSE(평균 제곱근 오차)로 관중 예측모델 평가.

```
#선형회귀 코드 입력 및 테스트, 실행 모델
from sklearn.linear_model import LinearRegression, Lasso ,Ridge
from sklearn.preprocessing import scale
from sklearn.model_selection import train_test_split #훈련세트 분리

y=KBO['attendance_per_game'] ## 종속변수= 경기당 관중

x_b=KBO[['year','month','day','dayofweek','두산']] ## 독립변수= 홈팀, 기온, 경기장 수용인원, 연월일, 주말 및 평일 가변수

from sklearn.preprocessing import scale
from sklearn.metrics import mean_absolute_error as mae # mae 코드 호출

X = scale(x_b) # 변수를 일반적인 크기로 조정
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state = 5) ##테스트셋, 훈련셋 70:30으로 분리

linreg_scale = LinearRegression()

linreg_scale.fit(X_train, y_train)

preds = linreg_scale.predict(X_test)
print(preds) ## 예측값 출력
```

프로야구 선형회귀모델 적용 코드

```
[ 7039.14417014  9616.83243325  6865.18877087   755.58649901
 12399.84176534  8912.65943885  5530.30816292  5940.62746974
  6016.91253805  6846.92641295  8476.77917075  5940.62746974
  8304.12995009  2295.97589228  8723.63680183 13779.64177189
  5648.73448005 14660.26379381  1344.9452668  12444.24738839
  9450.17563735 10468.30378622 10346.56638893 10884.68727415
  8598.70428438  9818.95838207  6625.62974316  8594.60080116
  6698.71969132 13491.85226542  5463.02292318 11666.19204278
  7499.27922092  4482.70328589 12175.34997539  2363.07341567
  3291.73820002  6305.49444758  3247.00651776 13008.3712541
 10712.03805349  6241.404328    9741.7649507   1975.8405967
```

프로야구 선형회귀모델 예측 결과

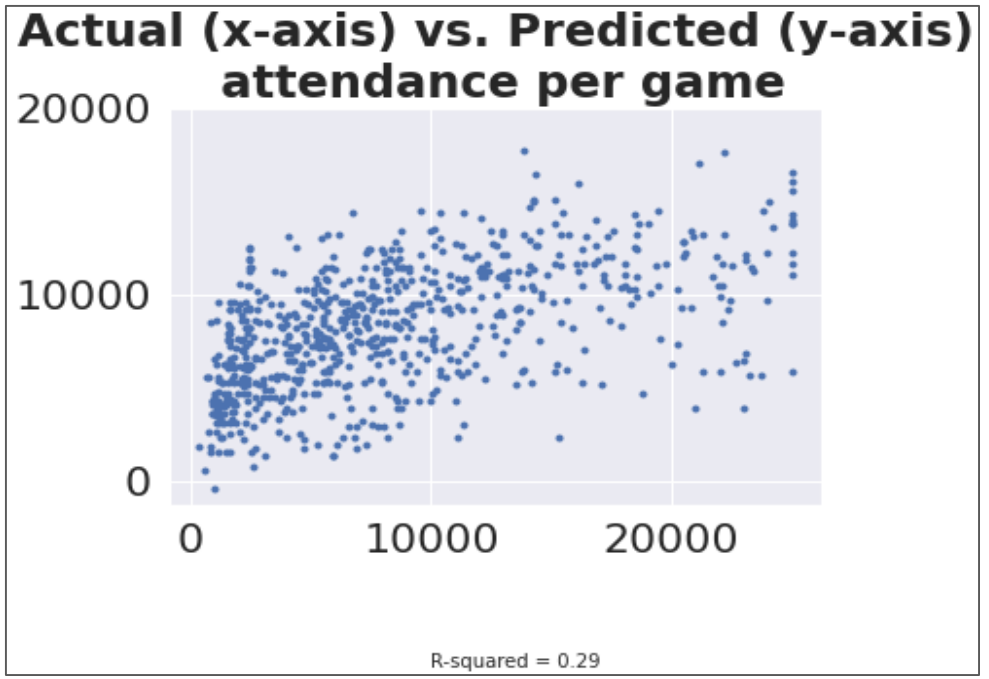
III. 모델링

1. 모델 개발

2) 프로야구 관중수 예측 선형회귀모델 개발 모델 평가 및 실측값과 예측값 비교

1. 프로야구 관중예측모델을 MAE, RMSE로 평가했을 때, RMSE, MAE 모두 평가값의 오차가 심했음
2. R-square 값으로 관중 예측값과 실제 관중 차이를 산점도로 표현했을 때, 30%대의 정확성을 보임
3. 해당 모델의 성능이 좋지 않음을 확인

모델 평가지표	평가값
MAE(평균 절대오차)	3999.30
RMSE (평균제곱오차제곱근)	5200.28
R-square(결정계수)	0.29



프로야구 관중수 예측값과 실측값 비교 산점도

III. 모델링

1. 모델 개발

3) 프로축구 관중수 예측 선형회귀모델 개발

- 2018~2022년 일일 관중수를 테스트세트와 훈련세트 분리(7:3)
- 평가지표를 MAE(평균절대오차)와 RMSE(평균 제곱근 오차)로 관중 예측모델 평가
- 모델 평가값의 성능 향상을 위해 경기당 관중점유율을 독립 변수로 적용

```
from sklearn.preprocessing import scale
from sklearn.metrics import mean_absolute_error as mae # mae 코드 호출

X = scale(x) # 변수를 일반적인 크기로 조정
X_train, X_test, y_1_train, y_1_test = train_test_split(X, y_1, test_size=0.30, random_state = 15) ##테스트셋, 훈련셋 70:30으로 분리

linreg_scale = LinearRegression()

linreg_scale.fit(X_train, y_1_train)

preds = linreg_scale.predict(X_test)
mae_linreg_s = mae(y_1_test, preds)
print(preds)
print('MAE (Mean Absolute Error)를 통한 프로축구 관중 예측 선형회귀모델 값: %.2f' %mae_linreg_s) ## MAE (Mean Absolute Error)를 통한 평
x=KLEAGUE[['year','month','dayofweek','is_weekend','전북','occupancy']]
```

프로축구 선형회귀모델 적용 코드

```
2704.95185508  2264.46313526 12957.85514618  1687.60653245
2328.00278509  6300.23797658  8792.62640408  5973.44330611
1563.36878924  3369.85903394  7615.48930466  4727.70721368
4013.88556601  2850.04875136  3444.91710854  3108.36038255
3440.18491657  1950.42875674  4189.70930703  2686.91587735
7306.33199186  2385.28796504  3153.97223426 11298.1235459
4170.40599544  5981.91781098  2243.79783874  4789.22721604
2922.69531827  2434.53199539  4397.53426641  8458.78758565
1707.80045477  6867.9556032  13110.76322231  7177.10525128
4233.03204456  5078.12199106 10249.38985143  3694.06416238
2930.07490522  2796.72496188  1718.42222395  3329.06168254
2230.75965957  3279.32978787  3153.77508107  5816.25366087
2231.57440125 12014.21307751  7618.47365271  5012.85238128
7568.78661726  4947.36319348  5164.69014156  1147.0855969
```

프로축구 선형회귀모델 예측 결과

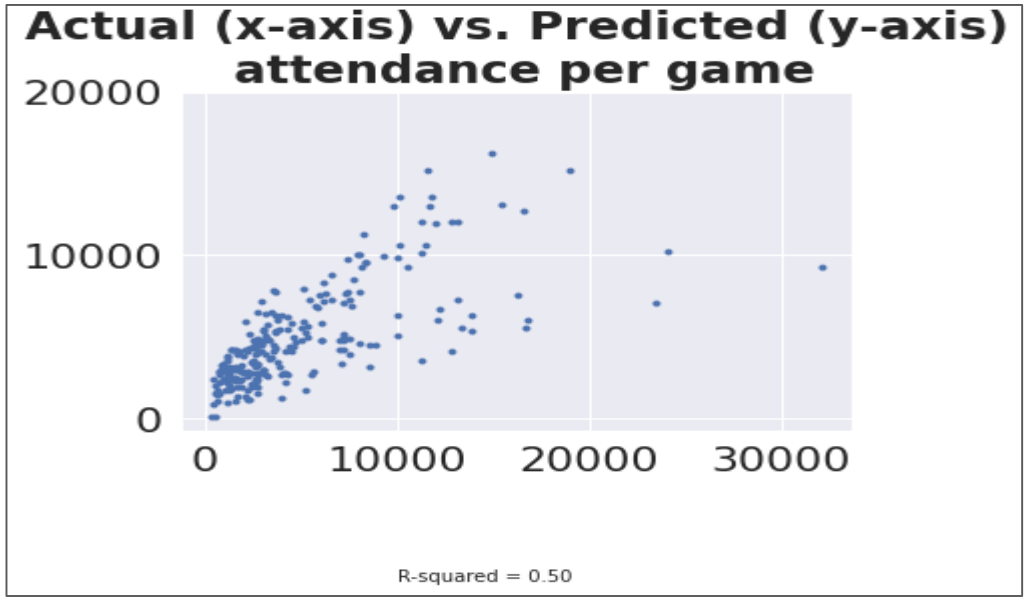
III. 모델링

1. 모델 개발

4) 프로축구 관중수 예측 선형회귀모델 개발 모델 평가 및 실측값과 예측값 비교

1. 프로축구 관중예측모델을 MAE, RMSE로 평가했을 때, RMSE, MAE 모두 평가값의 오차가 심함
2. R-square 값으로 관중 예측값과 실제 관중 차이를 산점도로 표현했을 때, 50%대의 정확성을 보임
3. 산점도를 통해 확인했을 때 관중예측과 실제 인원이 0에 수렴하는 밀집

모델 평가지표	평가값
MAE(평균 절대오차)	1995.73
RMSE (평균제곱오차제곱근)	3214.83
R-square(결정계수)	0.5



프로축구 관중수 예측값과 실측값 비교 산점도

III. 모델링

1. 모델 개발

5) 여자프로배구 관중수 예측 선형회귀모델 개발

- 2018~2022년 일일 관중수를 테스트세트와 훈련세트 분리(7:3)
- 평가지표를 MAE(평균절대오차)와 RMSE(평균 제곱근 오차)로 관중 예측모델 평가
- 여자프로배구의 경기수가 적은 편임을 감안했을 때 모델 분리가 원활

```
## 독립변수
X_2=WOMAN_KOVO[['year','month','day','dayofweek','GS칼텍스']]

#선형회귀 코드 입력 및 테스트, 실행 모델,  MAE 평가값 적용
from sklearn.linear_model import LinearRegression, Lasso ,Ridge
from sklearn.model_selection import train_test_split #훈련세트 분리
from sklearn.metrics import mean_absolute_error as mae # mae 코드 호출

x_2=scale(X_2)

x_2_train, x_2_test, y_2_train, y_2_test = train_test_split(x_2, y_2, test_size=0.30, random_state = 15)

linreg_scale = LinearRegression()

linreg_scale.fit(x_2_train, y_2_train)

preds = linreg_scale.predict(x_2_test)
```

여자프로배구 선형회귀모델 적용 코드

```
820.57418675 1331.70495536 1628.42359566
703.15377279 2726.65446579 2026.25109928
811.83243894 1716.09466542 2555.33384086
425.68040795 2699.31004605 2264.35725425
973.93646475 2834.58858118 2726.05791355
456.51437737 2128.56011246 1373.46137443
727.82023448  930.69165517 2626.0186712
860.4315797  1483.85879711 2012.57888941
900.64508989 1815.49783627 1985.23446968
796.92107329 2201.85148731 2273.5265662
729.33242384 2895.15194342 3263.54566261
218.63610667 1649.67831669 2443.8489837
801.8256264  1387.1335843  2012.57888941
```

여자프로배구 선형회귀모델 예측 결과

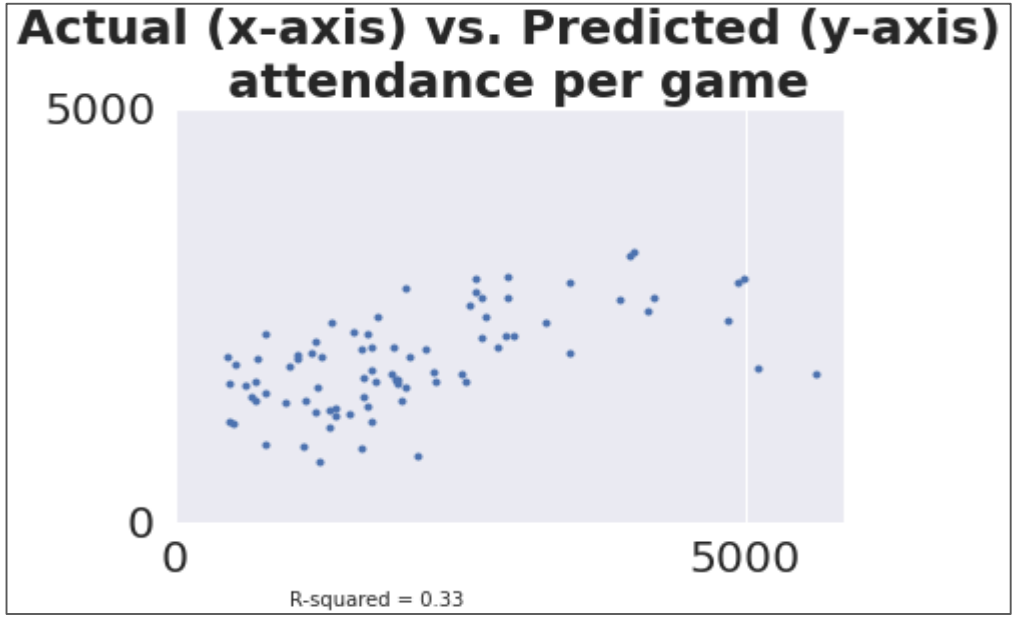
III. 모델링

1. 모델 개발

6) 여자프로배구 관중수 예측 선형회귀모델 개발 모델 평가 및 실측값과 예측값 비교

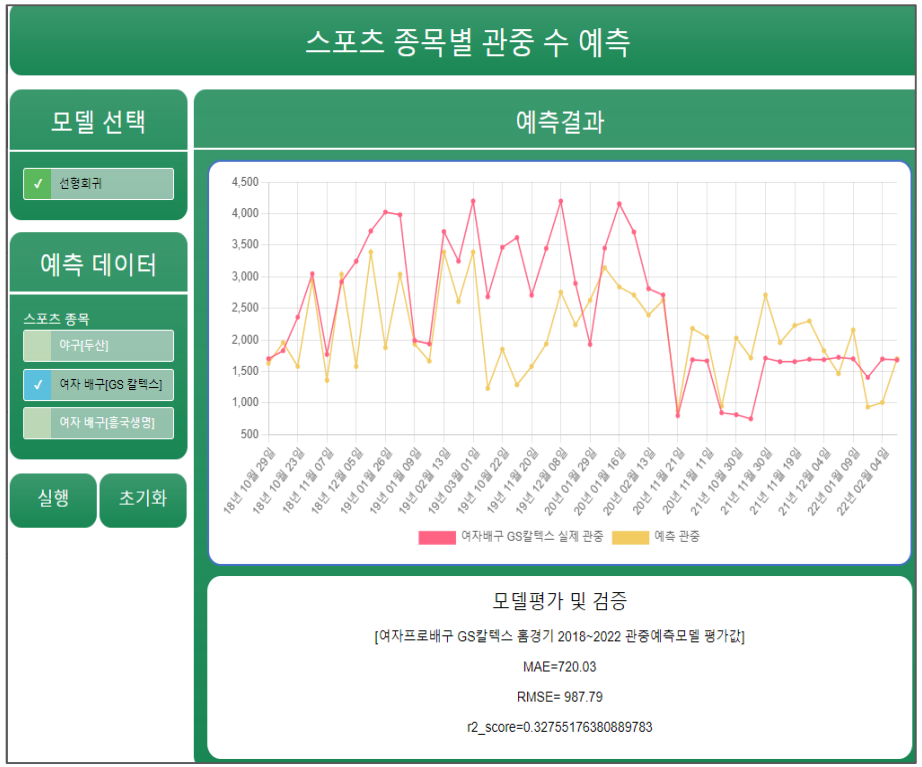
여자프로배구 관중예측모델을 MAE, RMSE로 평가했을 때,
타 종목 모델에 비해 가장 성능이 좋았음.

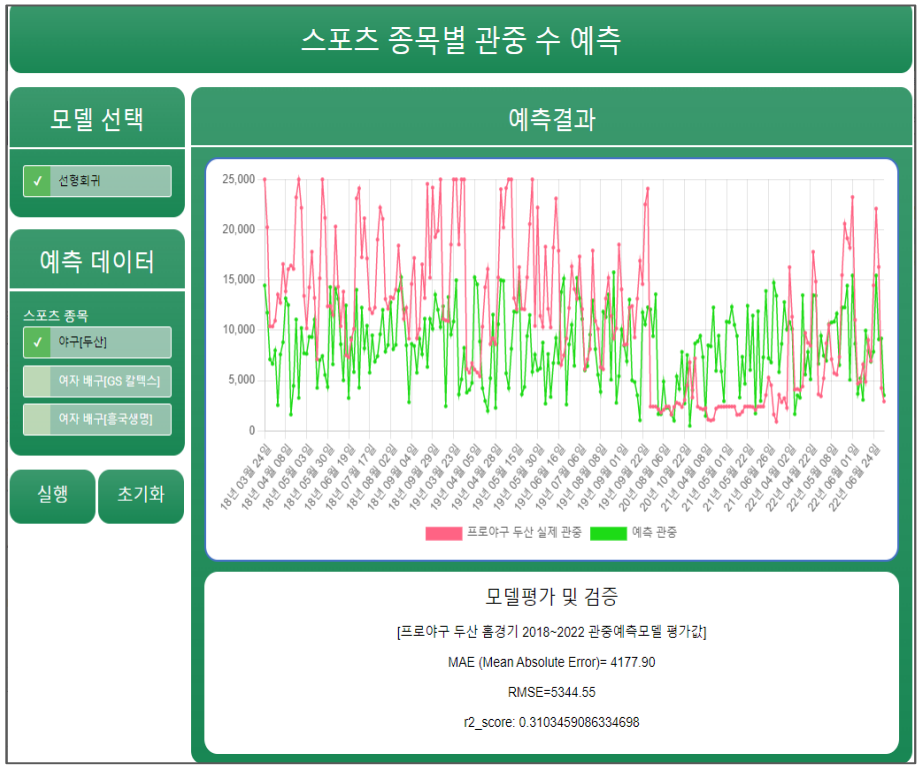
모델 평가지표	평가값
MAE(평균 절대오차)	293.37
RMSE (평균제곱오차제곱근)	988.16
R-square(결정계수)	0.33



여자프로배구 관중수 예측값과 실측값 비교 산점도

Ⅳ. 모델 예측 결과





- Django 웹 프레임을 실행해 예측값과 실제 관중값 차이 시각화 진행
- 코로나 19로 인한 무관중 경기를 제외하고, 실제 관중수와 예측 관중수를 비교했을 때 편차가 심했음

V. 결론

- 변수 중요도를 활용해 관중 수에 영향을 미치는 요인에 대한 정량적 시도를 통해 홈팀 및 요일 변수, 경기장 수용인원을 일일관중으로 나눈 관중점유율이 일일 관중 입장에 영향을 끼침

- 코로나 19로 인한 무관중 변수를 제외하고, 관중 예측모델을 구축했을 때 실제 관중입장과 예측 관중 값의 편차가 심했음

- 모델 성능을 향상 시키기 위해선 경기 결과에 따른 지표와 주간, 야간경기 여부 등의 변수 추가를 통한 지속적인 모형을 구축함으로 예측의 신뢰성을 확보할 필요가 있음