프로젝트 최종 보고서

60182180 안성표

60171759 장동혁

60170911 조성빈

* 1. 주제 선정 – 이유

2014년 소치 올림픽(김연아 은메달 논란), 2022년 베이징 올림픽(황대헌 실격 논란), 2022년 카타르 월드컵(추가시간 논란)과 같이 다양한 스포츠 분야에서 심판의 판정이 경기 및 결과를 직접적으로 결정짓고 논란을 일으켜왔다. 그 중에서도 특히 2014년 소치 올림픽과 2022년 베이징 올림픽은 홈팀인 러시아와 중국이 매우 유리한 판정을 받았다는 지배적인 의견이 있었다. 따라서 우리 조는 스포츠에서 반복되는 이러한 공정하지 않은 판정이 우연인지, 아니면 정말 심판들이 홈/어웨이에 따라 공정하지 않은 판정을 내리는지를 알아보기로 하였다. 그리고 이를 위해 심판의 권한이 절대적으로 크고 심판의 판정 하나하나가 객관적인 지표로 확인될 수 있는 야구를 통해 분석해 보았다.

* 1. 주제 선정 – 가설 설정

H0(귀무가설) : 심판들은 홈/어웨이 구분 없이 공정한 판정을 내린다.

H1(대립가설) : 심판들은 홈/어웨이에 따라 공정한 판정을 내리지 않는다.

2. 데이터 수집

2015년부터 2022년 월드시리즈까지 진행된 18000여개의 야구 경기에 대한 데이터

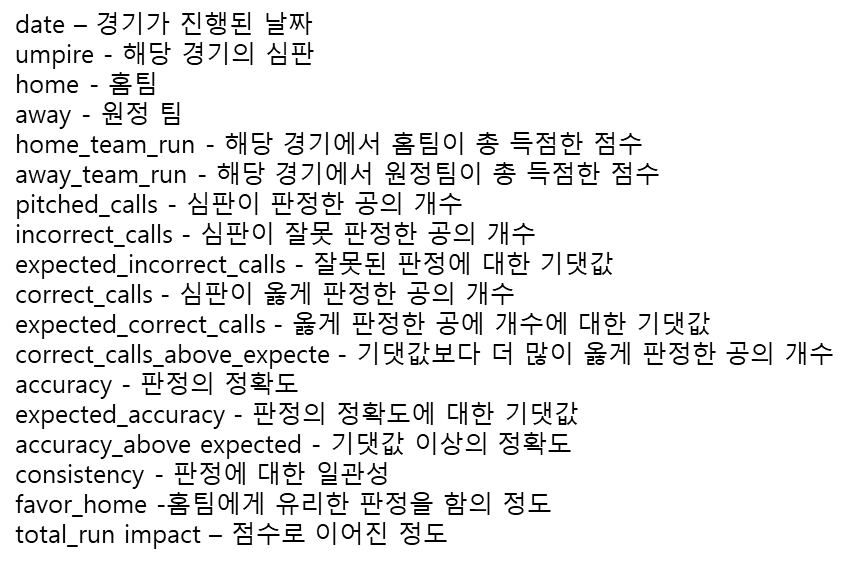
(출처 : [https://www.kaggle.com/datasets/mattop/mlb-baseball-umpire-scorecards-테이블이(가) 표시된 사진

자동 생성된 설명2015-2022](https://www.kaggle.com/datasets/mattop/mlb-baseball-umpire-scorecards-#2015-2022) )

테이블이(가) 표시된 사진

자동 생성된 설명

- 데이터 column 설명



이중에서 우리는 공정하지 않은 판정이 우연인지 아닌지를 분석하기 위해 홈팀에게 내려진 유리한 판정의 정도를 나타내는 ‘favor\_home’ column을 사용할 것이다.

테이블이(가) 표시된 사진

자동 생성된 설명3. 데이터 정제

해당 데이터에서 결측치인 ND를 정제해야 한다. ND를 포함하는 행의 개수테이블이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명는 120개의 row

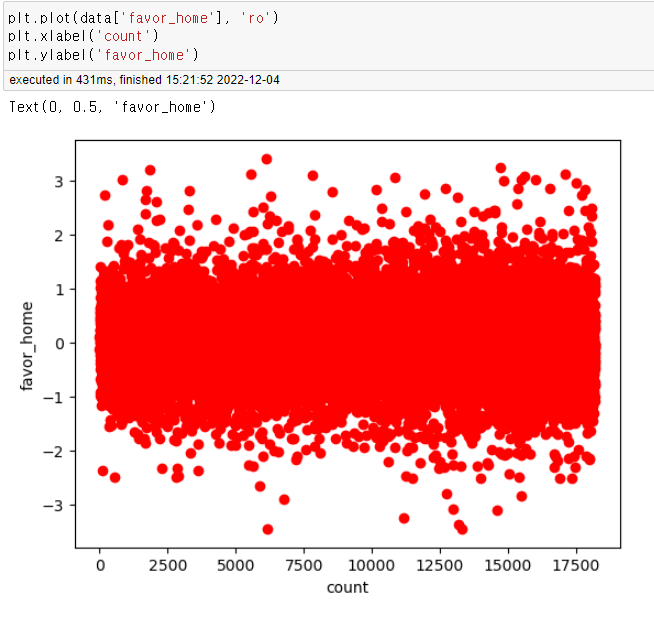
결측치를 제거하여 원래 18213개의 row에서 120개를 제외한 18093개의 row를 가진 데이터셋으로 변경됨

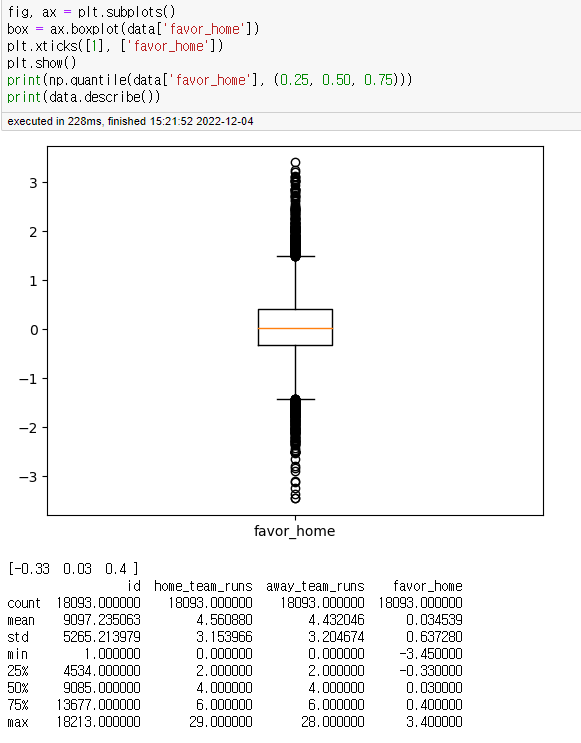
테이블이(가) 표시된 사진

자동 생성된 설명현재 데이터의 데이터타입. 우리가 다룰 favor\_home은 현재 object타테이블이(가) 표시된 사진

자동 생성된 설명입이므로 사용하기 위해 float타입으로 변경해준다.

4. EDA

(1) favor\_home의 분포를 나타내는 plot

(2) favor\_home의 Boxplot

favor\_home의 Boxplot을 통해 Outlier가 많이 존재함을 알 수 있음

텍스트이(가) 표시된 사진

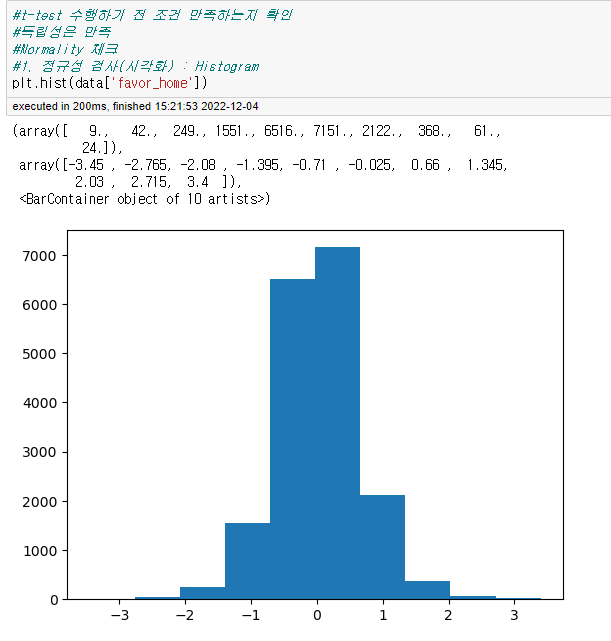
자동 생성된 설명Outlier가 몇 개인지 알기 위한 과정

텍스트이(가) 표시된 사진

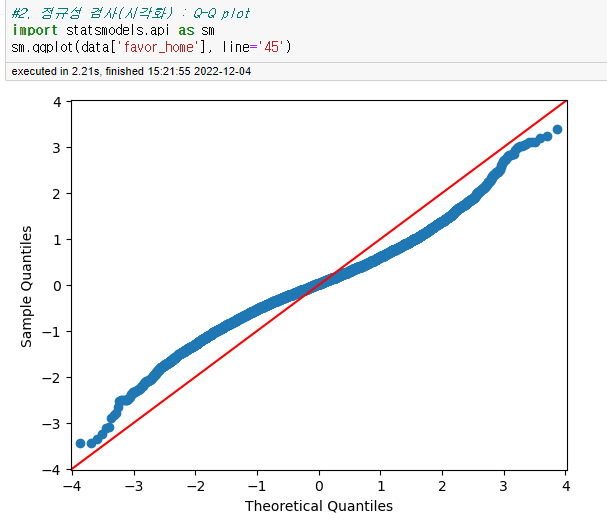
자동 생성된 설명Boxplot에서 윗부분의 Outlier는 305개, 아랫부분의 Outlier는 272개로 총 577개의 Outlier가 존재한다는 것을 알았다. 그러나 이 Outlier는 우리의 분석에서 매우 중요한 데이터이자 지표이므로 삭제하거나 처리하지 않고 그대로 활용할 것이다.

우리의 분석을 위한 t-test를 수행하기 전, 이를 위한 조건을 만족하는지 확인해야 한다. 독립성은 만족하므로 정규성을 만족하는지 확인한다.

5. 정규성 검사

(1) 시각화 - 히스토그램

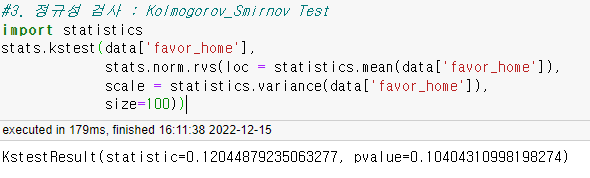
히스토그램을 통해 favor\_home이 정규성을 따르는 것으로 판단됨.

(2) 시각화 – Q-Q plot

Q-Q plot을 통해서도 favor\_home이 정규성을 따르는 것으로 판단됨.

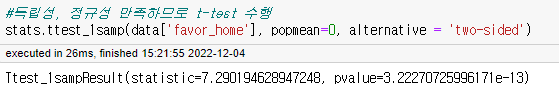
(3) 통계적 기법 - Kolmogorov-Smirnov Test

데이터의 수가 5000개 이상일 때 Shapiro test는 결과를 보장하지 못하므로 대신 ks test를 진행하였다.



p-value가 0.05보다 크므로 favor\_home이 정규성을 따른다고 판단됨.

따라서 데이터가 t-test를 수행하기 위한 조건을 만족하므로 t-test를 수행한다.

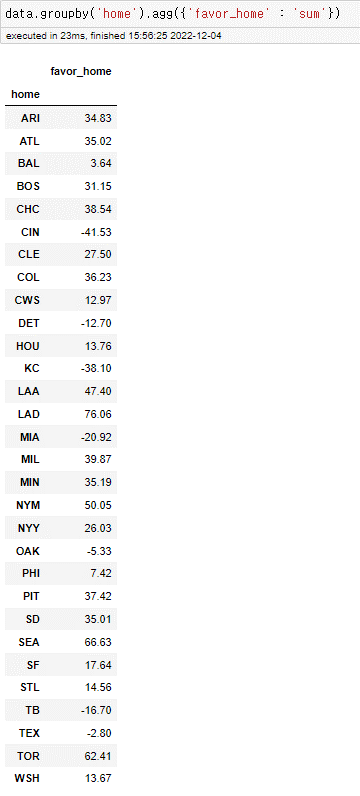
6. 가설검정 t-test 수행

t-test 수행결과 p값이 0.05보다 매우 작으므로 H0(귀무가설)을 기각하고 H1(대립가설)을 채택하여, 심판들은 홈/어웨이에 따라 공정한 판정을 내리지 않는다는 결론을 도출하였다.

앞서 우리는 t-test 수행을 통해 심판들은 홈/어웨이에 따라 공정한 판정을 내리지 않는다는 결론을 도출하였다. 그리고 이 결론을 바탕으로 이번엔 어떤 팀이 공정하지 않은 판정으로 인해 가장 많이 이득을 봤고, 또 어떤 팀이 가장 손해를 봤는지 궁금해져 간단히 확인해보았다.

7. 데이터 그룹화 – groupby





결과 로스앤젤레스 다저스(LAD)가 공정하지 않은 판정으로 가장 많은 이득을 봤고, 신시내티 레즈(CIN)이 가장 손해를 봤음을 알 수 있었다.

8. Regression – Logistic Regression

t-test를 통해 심판이 홈/어웨이에 따라 공정하지 않은 판정을 내린다는 것을 알 수 있었다. 그렇다면 이 공정하지 않은 판정이 경기의 결과인 승리/패배와 관련이 있는지, 즉 공정하지 않은 판정으로 승리/패배를 예측할 수 있는지 분석하였다. 그러나 승리/패배는 1/0처럼 continuous variable이 아니라 categorical variable이므로 수업시간에 배운 선형회귀분석 방법은 적용하지 못한다. 따라서 독립변수가 continuous variable이고 종속변수가 categorical variable일 때 적합한 방법인 로지스틱 회귀(logistic regression)을 진행하였다.

테이블이(가) 표시된 사진

자동 생성된 설명

‘home\_team\_runs’ column에서 ‘away\_team\_runs’ column을 뺀 값이 양수이면 1, 음수이면 0으로 설정하여 홈 팀의 승리 여부 column을 만들어 1과 0의 이산값으로 채워 넣었다. 그리고 favor\_home에 따른 홈 팀의 승리 여부를 예측하는 모델을 만들고, favor\_home이 홈 팀의 승리에 얼마나 영향을 미치는지 분석하였다.

텍스트이(가) 표시된 사진

자동 생성된 설명

독립변수와 종속변수를 분리하고, 학습 세트와 평가 세트를 분리한 뒤 logistic regression을 진행하였다.

텍스트이(가) 표시된 사진

자동 생성된 설명

favor\_home에 따른 홈 팀의 승리 여부를 예측하는 모델의 정확도는 53%가 나왔다. 53%가 그렇게 높은 정확도는 아니지만, 예측하는데 도움이 될 수 있다고 판단하여 분석을 계속 진행하였다. Feature의 coefficient, 즉 계수를 확인해 보니 0.13으로 아주 작은 양수가 나왔다. 이를 통해 favor\_home이 팀의 승리에 영향을 끼치는 것은 맞으나, 그 값이 매우 작으므로 크게 영향을 미치지 않고 favor\_home으로 승리여부를 예측하는 것은 힘들다는 결론을 도출하였다.

9. 분석 결론

(1) t-test 결과 : 심판들은 홈/어웨이에 따라 공정한 판정을 내리지 않는다는 결론을 도출하였다.

(2) groupby 결과 : 로스앤젤레스 다저스(LAD)가 공정하지 않은 판정으로 가장 많은 이득을 봤고, 신시내티 레즈(CIN)이 가장 손해를 봤음을 알 수 있었다.

(3) regression 결과 : favor\_home이 팀의 승리에 영향을 끼치는 것은 맞으나, 그 값이 매우 작으므로 크게 영향을 미치지 않고 favor\_home으로 승리여부를 예측하는 것은 힘들다는 결론을 도출하였다.

10. 프로젝트 github 주소

<https://github.com/tiger9803/SDA.git>

이름 초성으로 각 branch를 만들어 각자의 branch에 파일 업로드 및 commit을 달았다.

branch SB : 60170911 조성빈

branch DH : 60171759 장동혁

branch SP : 60182180 안성표