A조 보고서

주제:선수 데이터를 통한 월드컵 결과 예측

조장:강민석

조원:김도균,이승준,한수용,황경민

목차

IV. 결론

1) 결론

2)추가 활용 계획

I. 프로젝트 개요

1.1 프로젝트 배경

- 축구 산업은 스포츠 산업중 가장 큰 분야 중 하나이며 유럽 축구시장은 규모가 약 30조원이 넘는다. 그중 1위는 프리미어리그(EPL) 이고 규모는 약 7조 2천억 이상이다. 그만큼 EPL은 수 많은 축구 선수들과 축구 팬들 사이에선 전 세계에서 가장 수준이 높고 경쟁력이 뛰어난 리그로 알려져있다.

Country coefficients

2018/19 2019/20 2020/21 2021/22

Pos	Country	17/18	18/19	19/20	20/21	21/22	Pts	Clubs
1 =	England	20.071	22.642	18.571	24.357	21.000	106.641	7
2 =	Spain	19.714	19.571	18.928	19.500	18.428	96.141	7
3 =	Italy	17.333	12.642	14.928	16.285	15.714	76.902	7
4 =	Germany	9.857	15.214	18.714	15.214	16.214	75.213	1/7
5 =	France	11.500	10.583	11.666	7.916	18.416	60.081	6

(참고 : EPL이 왜 세계 최고의 리그인가 : How association club coefficients are calculated | Country coefficients | UEFA Coefficients | UEFA.com

→ 세계적으로 가장 큰 클럽 대회라고 할 수 있는 UEFA 에서의 승리에 따른 계수들에 의해 계산되었고, 실제로 EPL이 다른 리그들을 제치고 1위하였다.)

- 최근 기술의 발전으로 인해 이 역사깊은 리그에서 다양한 데이터 수집이 가능해 지면서 축구를 더욱 재밌게 즐길 수 있는 요소 중 하나인 "승부예측"에 대한 관심이 뜨거워 지고 있다.
- 또한 손흥민 선수가 21~22 시즌 프리미어리그 득점왕을 기록하면서 대한민국은 다시 한번 축구에 대한 관심이 커지게 되었다. 동시에 올해 11월에는 4년에 한번 있는 세계의 축제인 월드컵이 카타르에서 열리게 되어 축구에 대한 관심을 더욱 가열 될 예정이다.

1.2 프로젝트 목적

- 축구계를 대표하고 최고의 선수들이 모이는 리그 EPL 데이터를 분석하여 경기 결과에 영향을 미치는 경기 외부적 , 내부적 변수를 해석하여 예측에 유효한 변수를 탐색해본다.
- 위에 과정을 통해 변수를 조정하여 실제 월드컵 축구 조별 예선 통과 국가 팀 예측과 최종적으로 프로젝트의 목적인 최종 월드컵 우승 결과팀을 예측하는 모델을 만든다.

1.3 프로젝트 작업환경

작업환경(Workspace environment)	
운영체제	Windows 10 Pro x64

분석도구	Python_3.9, R_3.6.1 , Jupyter Notebook
전처리	Excel 2020 , Notepad++

II. 데이터수집 및 분석

2.1. 데이터 수집

- 우선 EPL 데이터를 알아야 하므로 Kaggle에서 제공하는 EPL 프리미어리그 결과 데이터를 사용하였다.
- 선수 공격, 수비 데이터를 알아보기 위하여 Fifa22 게임 캐릭터 능력치 데이터를 활용하였다.
- FIFA 제작사는 실제로 약 6000명에 달하는 일반인 경기리뷰어 참가자들, 400명 정도의 프리랜서 축구선수 스카우터들과 이를 지휘하는 30명의 피파 종합점수 담당 프로듀서들이 모여 Overall 점수를 조율하기 때문에 믿을 만한 점수이다.

(참고 : http://fifa-talentscout.ea.com/ → 실제 일반인 경기리뷰어들을 모집하는 사이트이고, 소개글을 통해 EA Scouter의 규모를 알 수 있었다)

- 크롤링을 통하여 EPL 각 팀의 선수들 수, 출전 경기 수를 수집하였다.
- 추가적인 크롤링을 통해 조별 예선 출전 국가의 베스트 11 선수명단을 수집하였다.

2.2. 데이터 전처리

1) Kaggle에서 가져온 데이터 파일을 정제

- 데이터 분석에 불필요한 column을 삭제
- (예: sofifa_id, 선수_사진_url)
- 이름에 스페인어, 일본어, 중국어와 같이 인코딩이 다른 언어가 깨져 물음표로 나오거나 이상한 문자로 나옴 -> 이를 문제 없이 사용하기 위해 문자 수정 (예 : é -> e)
- 수정한 데이터를 Python을 사용하여 각 EPL 팀별 베스트 11 선수 명단을 정리 (예 : Arsenal.csv, Tottenham.csv …..)

2) EPL 경기 별 데이터 : EPL 공식 사이트에 저장된 Standings, Stats의 테이블 데이터를 크롤링 하여, 상황에 따른 독립변수와 종속변수의 상관관계 및 회귀분석에 사용하기 위해 csv 파일을 추가 생성 (예: A vs B의 EPL 경기의 세부 결과 → 홈팀?/어웨이팀? , 최종 점수차, 슈팅 개수, 유효슈팅 개수, 파울 개수, 코너킥 수, ….)

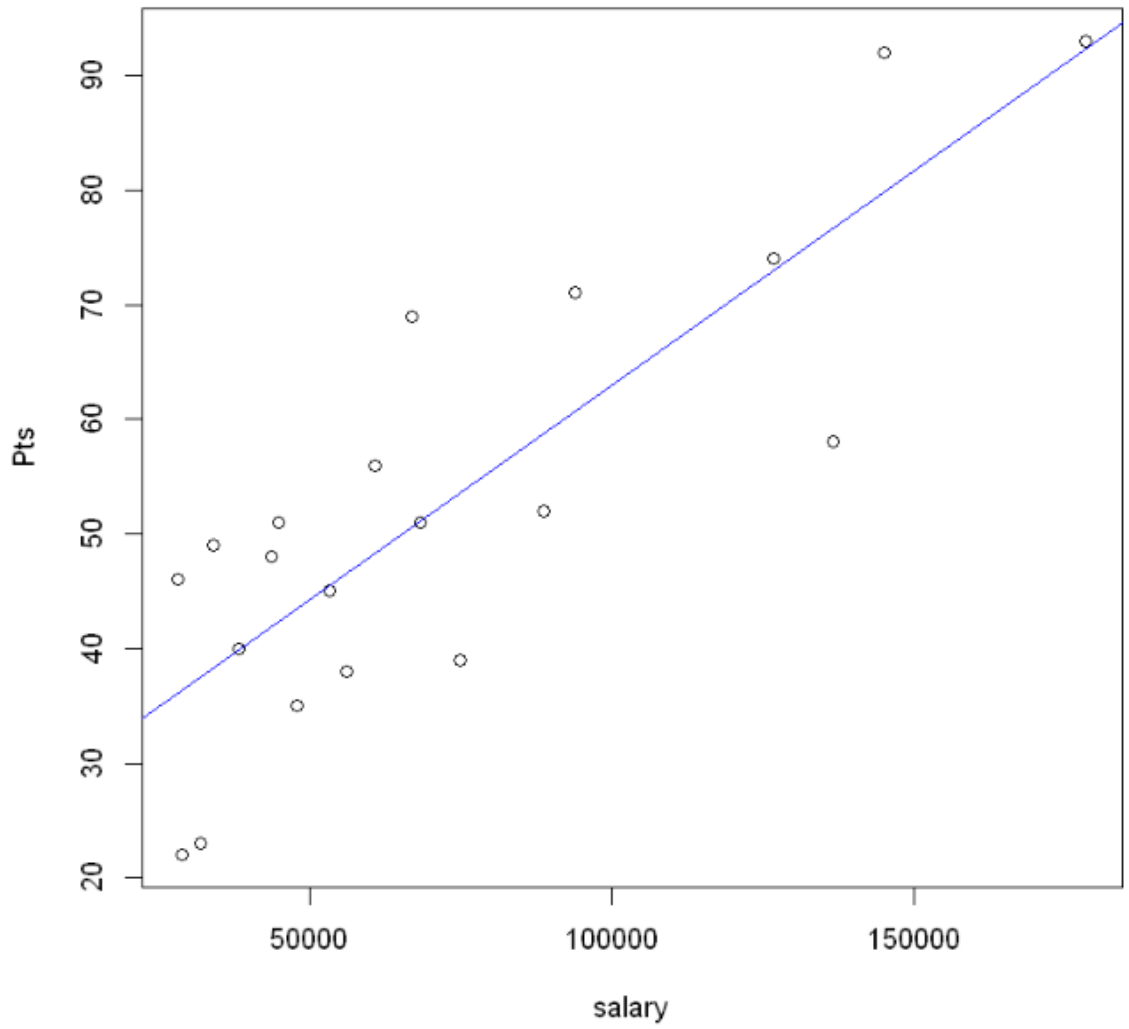
3) 상관관계/회귀분석을 파악하기 위해 R을 사용하여 추가적인 파생 변수를 만들고 계산하였다. (예: 특정 팀의 공격수의 FIFA 평균 종합점수, 수비수의 평균 종합점수, …)

4) 월드컵 조별 예선을 통과한 나라 별 베스트 11 명단:

최종목적인 월드컵 출전 국가들의 승리예측을 위한 선발명단에 포함된 선수들의 최종적인 점수 지표(EPL 데이터를 활용)를 포함하여 이를 예측모델에 사용한다.

2.3 데이터 시각화

1) 연봉과 승점의 상관관계

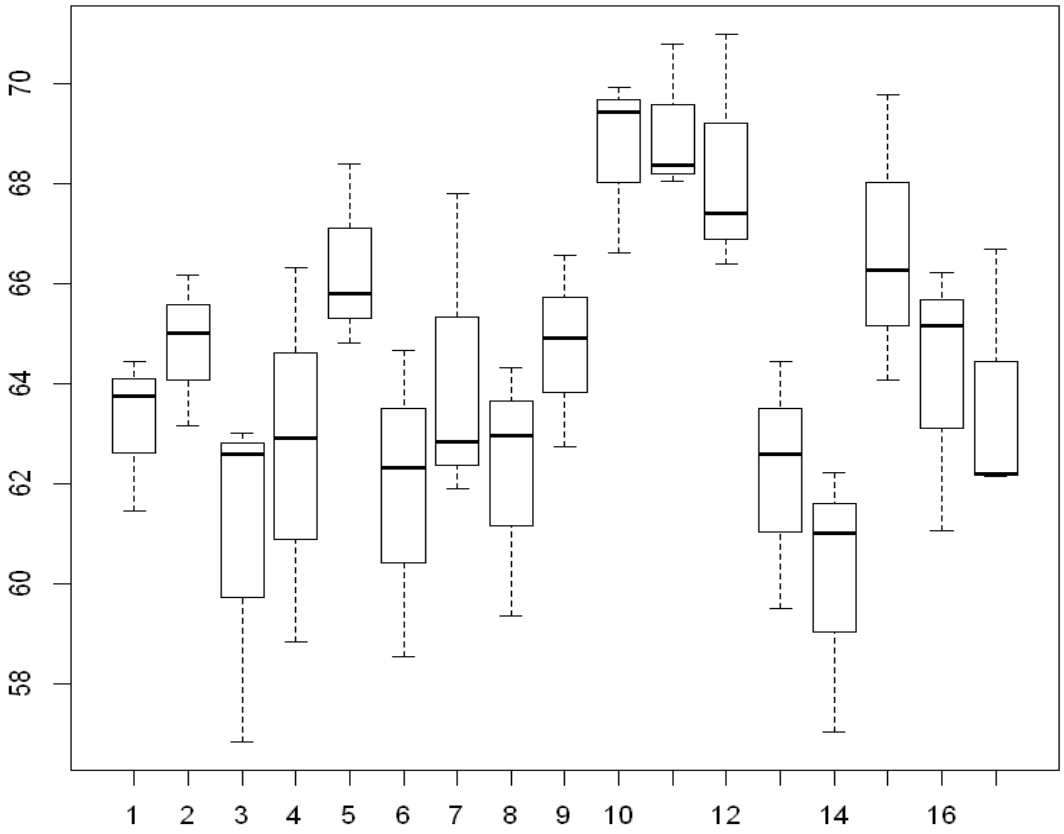


< 그림 1> 연봉과 승점의 선형회귀분석 그래프

- 연봉과 승점에 대한 상관계수를 통하여 연관성을 확인하였다.(피어슨 상관계수:0.83556)
- 선형 회귀분석을 통하여 선수들의 연봉과 승점에 대한 상관관계를 분석하였다.
- 선수들의 연봉은: 선수들의 평균연봉으로 구하였다.
- p-value: 4.52e-06, R-squared: 0.6982, F-statistic: 41.64

- 결론 적으로 약 70퍼센트의 정확도를 가지는 것을 확인 할 수 있었다.

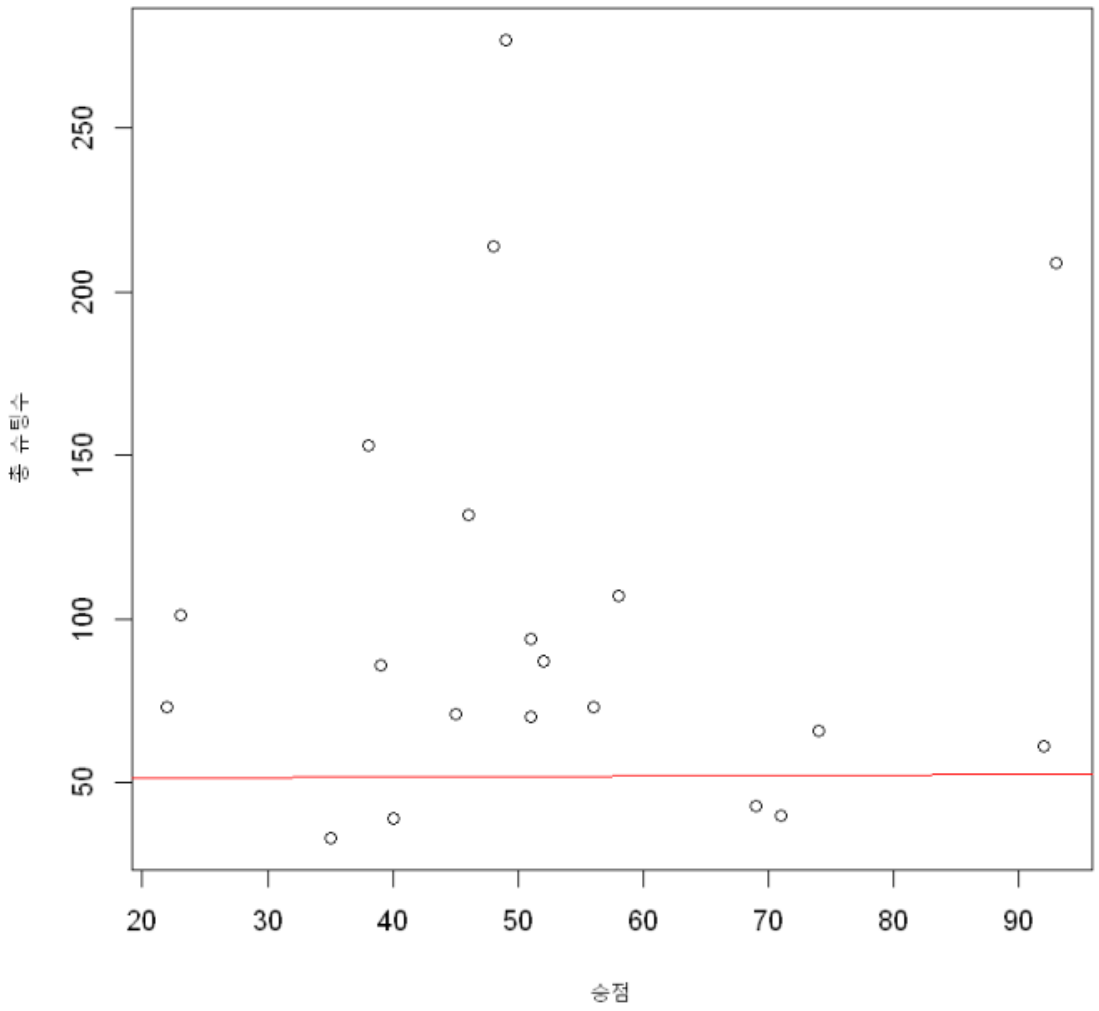
2) 경기 승리 연관성 분석



Aston_Villa,Leeds_Unite,Brighton_and_Hove_Albion,Chelsea,Crystal_Palace,Everton,,Leicester_City,Brentford,Tottenham_Hotspur,Liverpool,Manchester_City,Manchester_United,Newcastle_United,Southampton,,Arsenal,West_Ham_United,Wolverhampton_Wanderers

팀별 선수 능력치에 따른 boxplot (DF,MF,FW의 평균치)

3) 슈팅과 승점 연관성



< 그림 3> 슈팅 수 – 승점 연관성

- 상관분석 결과

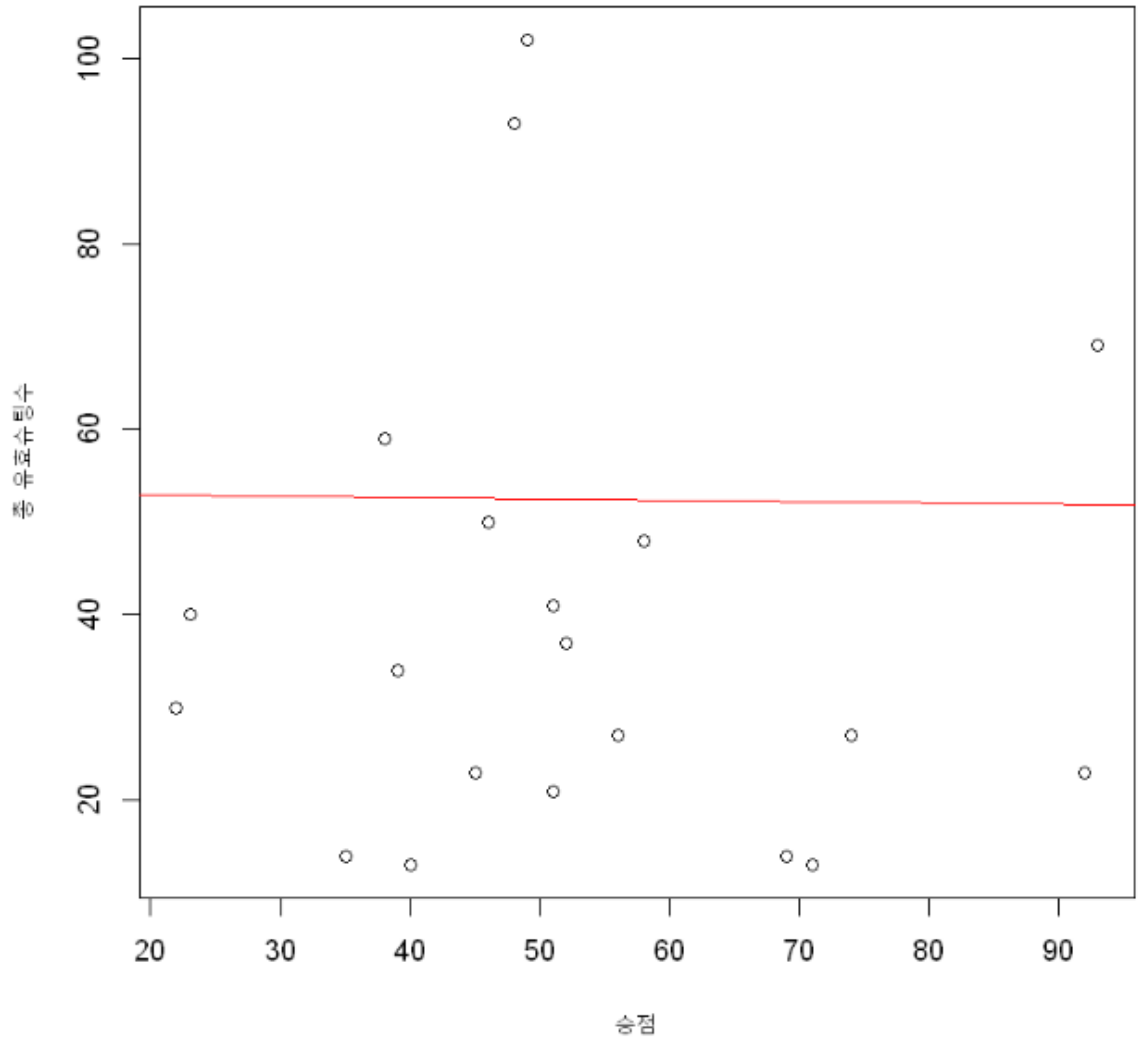
  상관계수 = 0.046 으로 낮은 상관 관계가 있음을 확인.

- 회귀분석 결과

  p-value: = 0.8454, R-squared: -0.05327, F-statistic: 0.03914

  R-squared 값에 의해 신뢰성이 없다고 판단.

4) 유효슈팅과 승점 연관성



< 그림 4> 유효슈팅 수 – 승점 연관성

- **상관분석 결과**

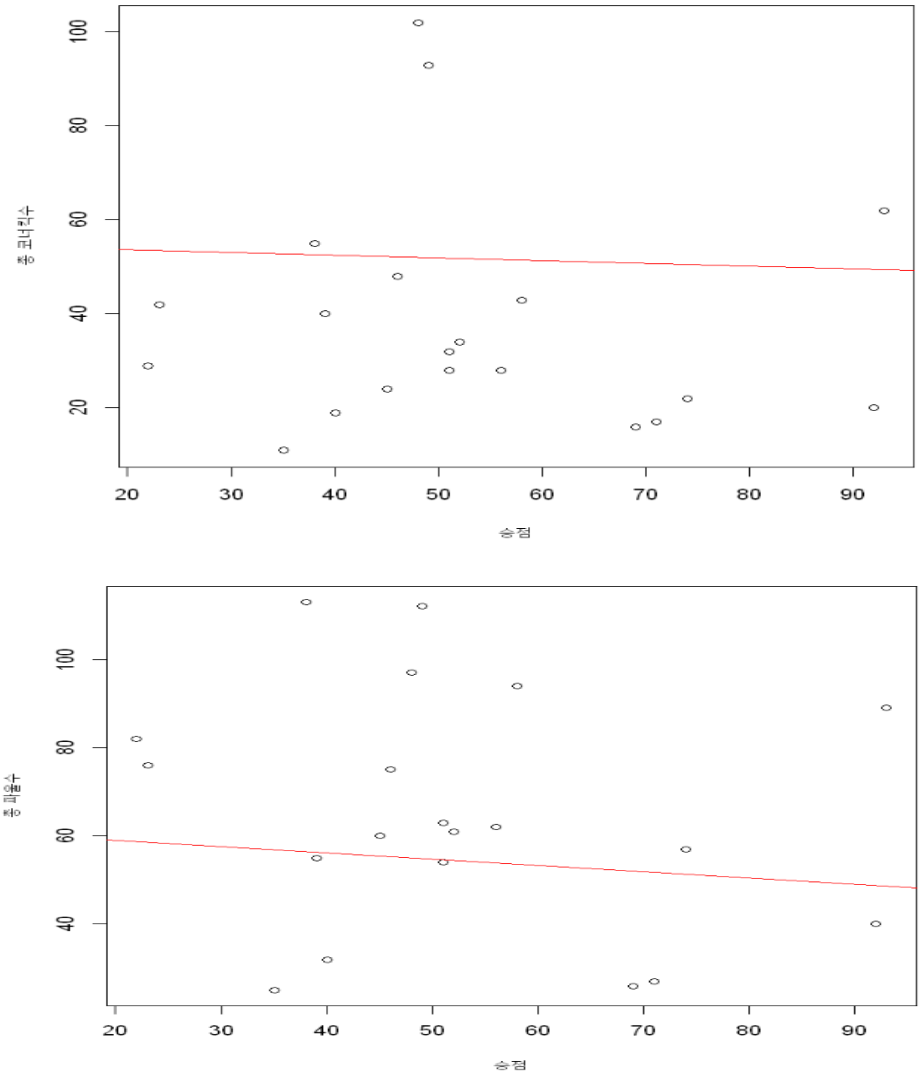
  상관계수 = −0.0177 으로 낮은 상관 관계가 있음을 확인.

- **회귀분석 결과**

  p-value: = 0.9407, R-squared: -0.05522, F-statistic: 0.005693

R-squared 값에 의해 신뢰성이 없다고 판단.

5) 파울 및 코너킥 승점 연관성



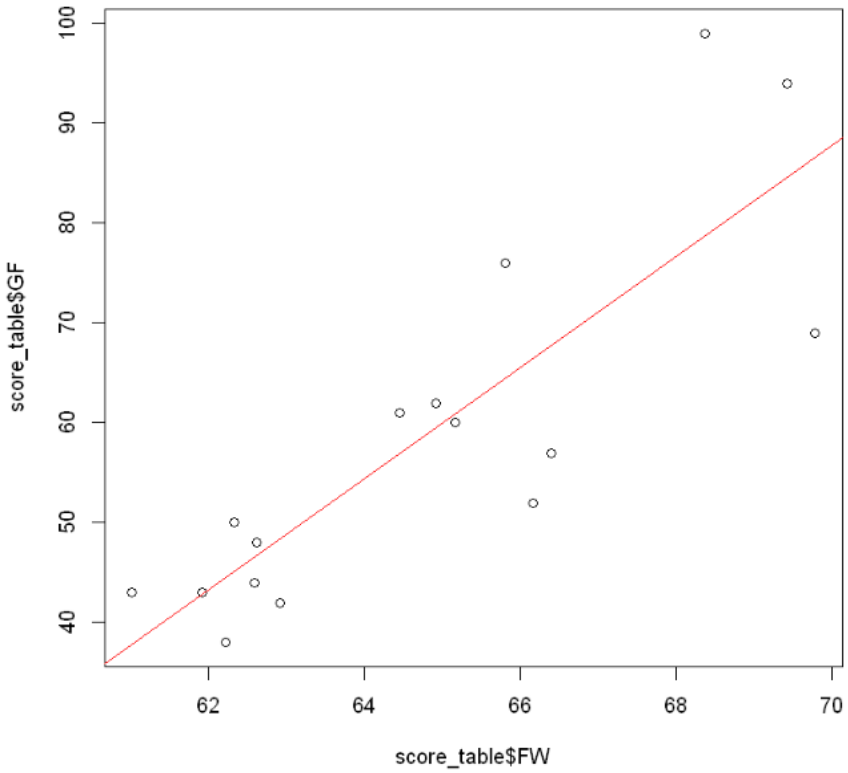
< 그림 4> 유효슈팅 수 – 승점 연관성

- 상관분석 및 회귀분석 결과

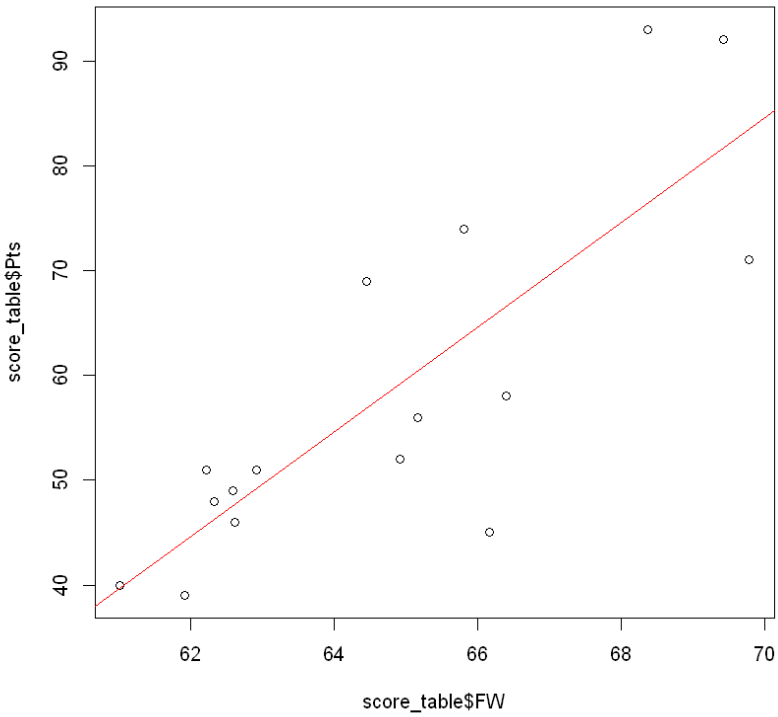
각각 상관계수 = −0.074 , −0.1989 으로 낮은 상관관계가 있음을 확인.

6) 경기 득점 연관성

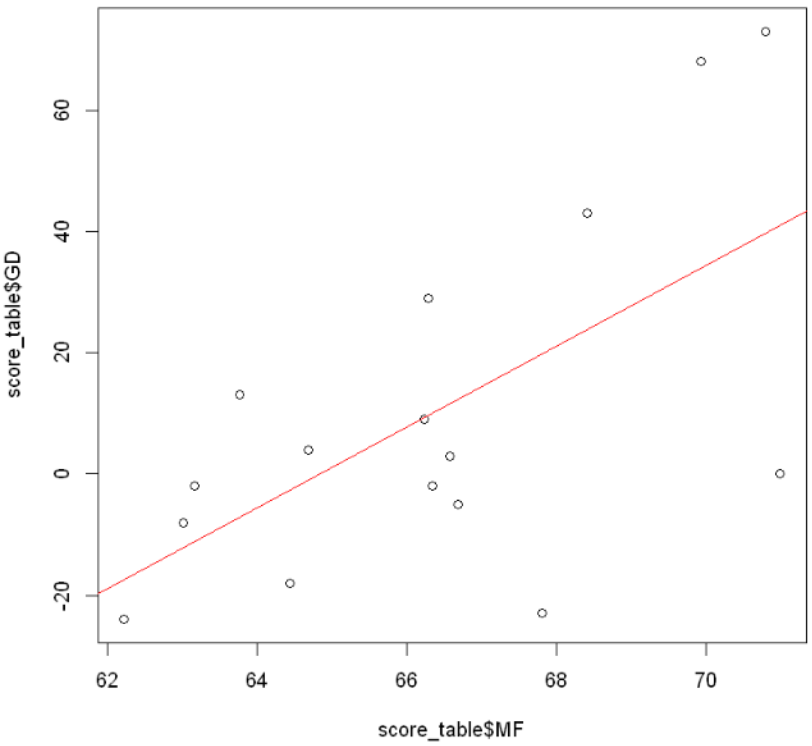
- 공격수와 골수와의 연관성



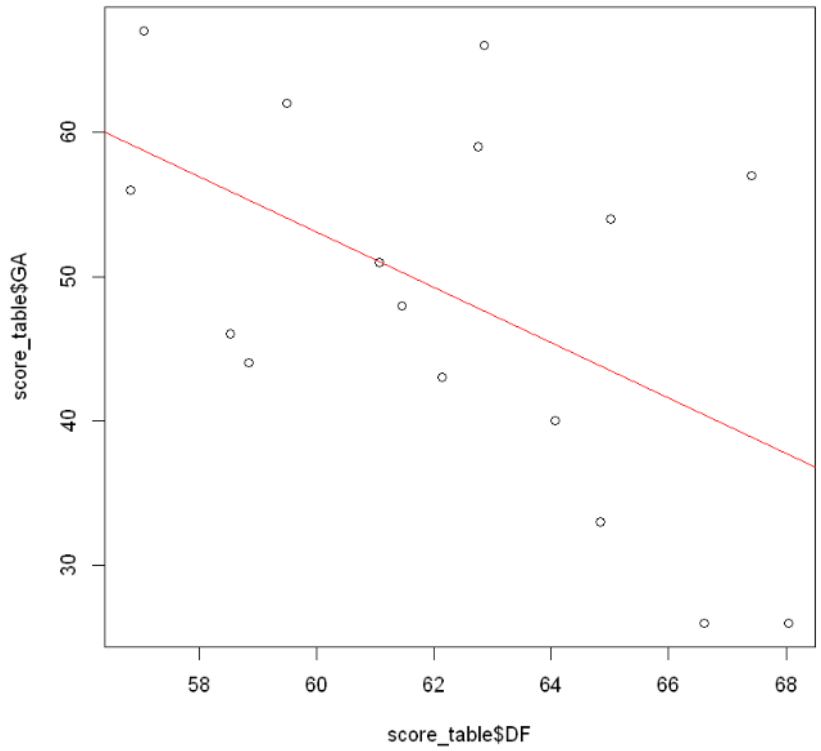
- 공격수와 총 득점과의 연관성



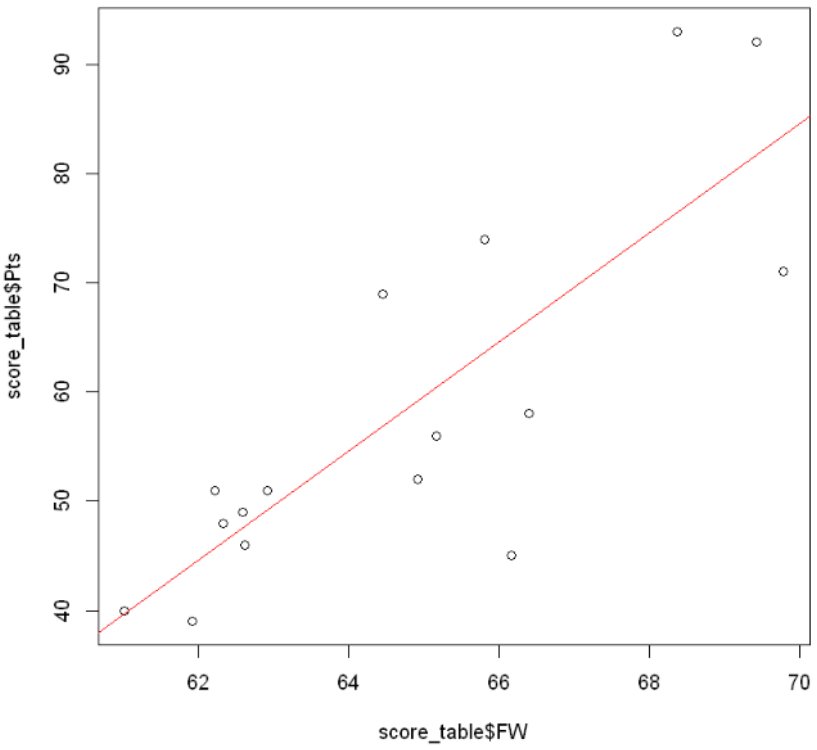
- 미드필더와 총 득점과의 연관성



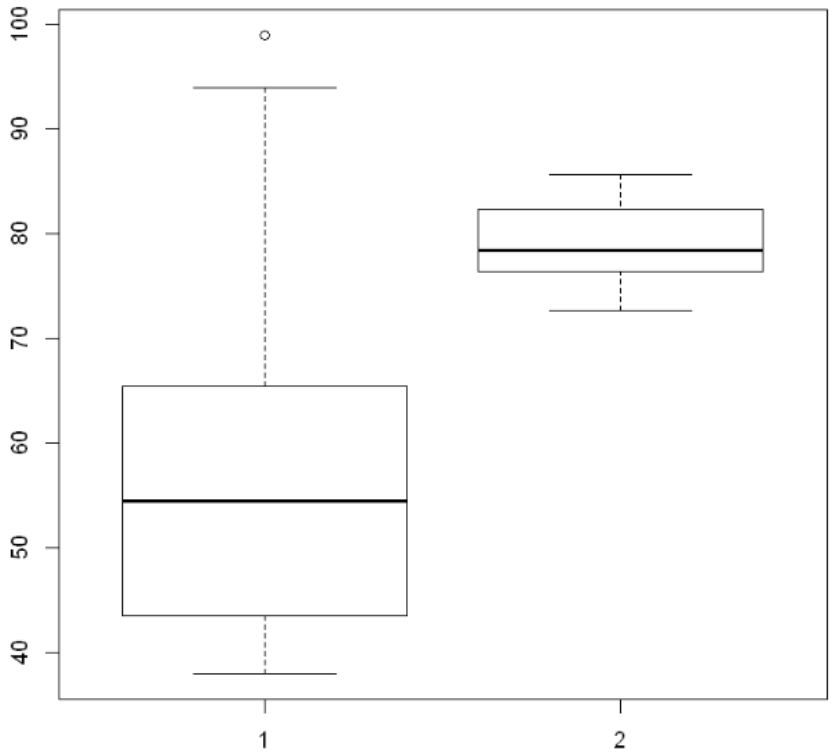
- 수비수와 총 실점과의 연관성



- 공격수와 총 득실점과의 연관성



- overall value 의 값을 통하여 연관성 확인



overall_score_table$GF, overall_score_table$FW

data: overall_score_table$FW and overall_score_table$GF

t = 5.4907, df = 14, p-value = 7.954e-05

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.5601007 0.9378838

sample estimates:

cor

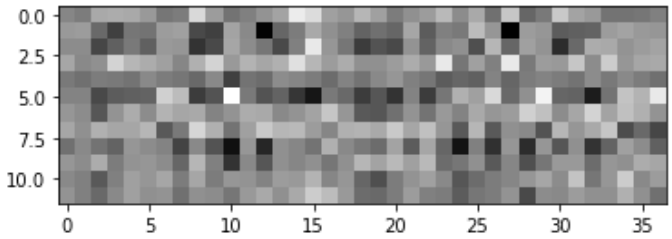
0.8263689

위 항목 모두 각 요인과 승점 간의 관계와 같이 상관계수가 낮기 때문에 경기에 영향을 주는 요인이라고 판단하기 힘듦.

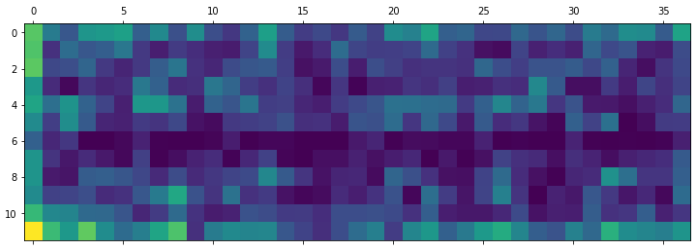
III. 모델링과 경기 예측

위와 같은 연관성을 바탕으로, 과거의 경기들로 선수들의 세부 능력치들과 그 결과의 연관성을 적절히 계산한다면 앞으로의 경기 결과 역시 예측이 가능하다고 판단하고 모델을 설계하였다.

1. 홈 팀과 원정 팀의 최다 출전 선수들의 세부 능력치들의 차이를 비교하였다.



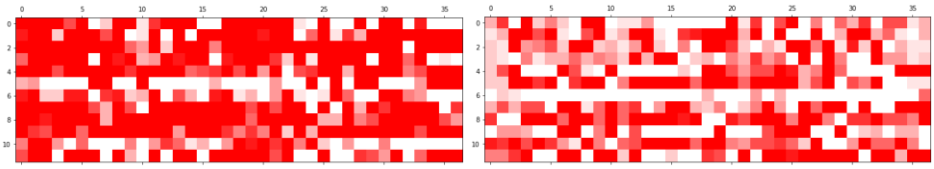
2. 이를 홈 팀 기준 경기 결과 "승/무/패" 에 따라서 각 포지션에서 어떤 능력치 값의 차이가 결과에 영향을 끼치는 지 머신러닝을 통해 파악하였다.



다음과 같이 경기 결과에는 최전방 공격수와 수비수의 능력치들이 미드필더 들의 세부 능력치보다 크게 영향을 끼치는 것을 파악하였다.

3. 승리한 경기와 패배한 경기에 대해 각각 가중치와 능력치 차이를 곱해본 결과이다. 공격수와 수비수의 능력치

차이가 클수록 승리할 확률이 올라가는 것을 확인할 수 있다.



[승리로 예측할 경우]　　*[패배로 예측하는 경우]*

4. 위와 같은 결과를 바탕으로 지난 2021/22 시즌 EPL의 결과를 시험해본 결과이다. 결과적으로 최종 모델을 53%의 확률을 가지고 승/무/패를 예측할 수 있었다.

```
1.  ['Leicester City', 1, 2, 'Everton', '-1']
    [0. 0. 1.]
    array([[0.77302426, 0.06651089, 0.16046482]], dtype=float32)

2.  ['Newcastle United', 2, 0, 'Arsenal', 2]
    [1. 0. 0.]
    array([[0.14248694, 0.1474754 , 0.71003765]], dtype=float32)

3.  ['Manchester City', 3, 2, 'Aston Villa', 1]
    [1. 0. 0.]
    array([[0.9026154 , 0.04202256, 0.05536198]], dtype=float32)
```

[홈 팀 , 1 : 2 , 어웨이 팀, 승/무/패 실제 결과]

[승. 무. 패 예측 결과]

[승. 무. 패 확률]

두 번째 예측 결과는 11위 팀 NewCastle 과 5위 팀 Arsenal 의 경기를 Newcastle의 [승리 14.2%, 무승부 14.7%, 패배 71.0%] 로 예측하였고, 실제 결과와는 다르지만 각 팀의 최종 결과를 보면 매우 합리적인 추측인 것을 확인할 수 있다.

5. 합리적인 수준으로 예측하는 모델을 통해 2022년 카타르 월드컵의 각 나라의 경기 결과를 예측해보고자 한다.

IV. 결론

- 축구 경기 예측을 위해선 축구 선수 개개인의 능력 보단 11명의 축구 선수들의 높은 평균의 종합 능력이 경기 결과에 큰 영향을 끼친다

- 데이터로 사용했던 팀의 각 포지션 별 평균 FIFA 능력치가 상대에 비해 크다면 실제로 예측결과 승리할 확률이 높았다. 이를 통해 월드컵 경기 예측의 가능하다는 것을 알게되었다.

IV. 추가 활용 계획

- EPL 뿐만 아니라 다른 리그의 선수 데이터를 종합하여, 최종적으로 월드컵 뿐만 아니라 각종 컵 대회, 대륙 간 대항전(UEFA/Champions League ,

UEFA/Conference League , etc …) 등의 결과를 예측 모델 기능 설계.

- 일반적인 합법 도박 사이트와 예측사이트의 실제 사용 모델의 정확성에 버금가는 (위의 사이트들은 평균 30% 중후반의 정확성을 가진다고 함) 모델을 만들었지만, 추가적으로 프로젝트 기간내에 활용 하지 못했던 추가적인 독립변수들 (선수의 당일 컨디션, 날씨, 경기 관중 수, …) 을 사용하여 더욱 정밀하고 치밀한 예측 모델 추가 보완.