**주제: “2KNBA 게임에서 선수들의 rankings(rating)이 선수들의 기록(stats)를 얼마나 반영하는가?”**

해당 MLR 모델을 구축하기 위해 nba2k 게임의 선수들 rating에 대한 데이터셋을 사용하기로 결정하였다. 사실 해당 모델을 선정한 이유 중 하나는 평상시에도 nba게임을 즐겨하는데 해당 선수들의 rating이 해당 시즌의 활약 및 기록들과 얼마나 관계 있을지 궁금했기 때문이다. 특히 이 중에서 MLR이 적용하기 위해 2014시즌부터 2020시즌까지 선수들의 시즌 평균 기록들을 포함하는 데이터셋을 선정했다. 위 데이터셋은 총 2412개의 관측치와 31개의 변수들이 존재하고, rating이 해당 선수의 performance를 나타내는 평균 기록의 영향과 관련성이 있을 거라 판단하였다.

**다운로드 링크:** [**https://www.kaggle.com/willyiamyu/nba-2k-ratings-with-real-nba-stats**](https://www.kaggle.com/willyiamyu/nba-2k-ratings-with-real-nba-stats)

**테이블이(가) 표시된 사진

자동 생성된 설명**

데이터셋의 종속변수 rating은 nba2k 게임에서 선수의 능력 치를 종합적으로 표현하는 수치이다. 즉 해당 선수가 얼마나 농구를 잘하는지에 대한 평가라고 볼 수도 있다. 물론 실력이라는 것이 기록된 것으로만 표현되지는 않는다. 그러나 그것을 가장 직관적/수치적으로 보여주는 것은 분명 선수의 평균 기록이라고 생각하기 때문에 선형관계가 있다고 생각한다.

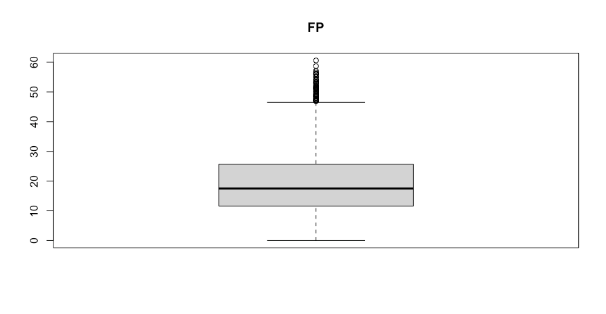
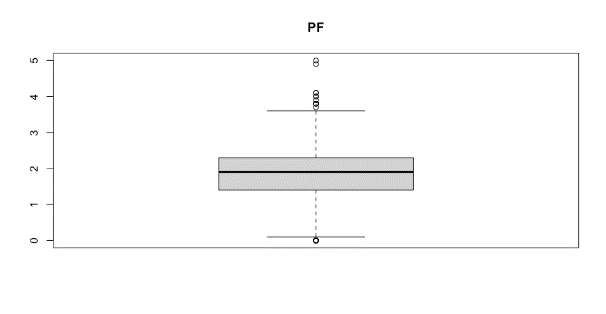
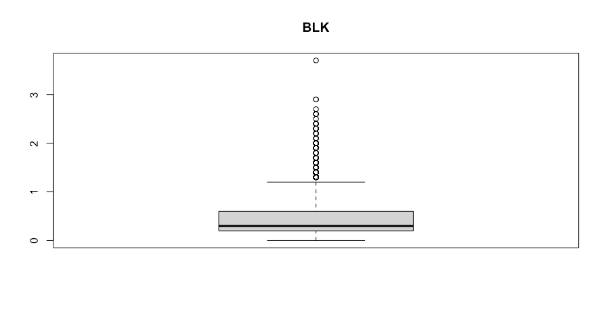
위 데이터 셋은 31개의 선수 기록을 포함한다. 이 중 가장 종속변수 rating과 가장 상관관계가 높을 것으로 생각하는 변수들은 PTS,REB그리고 AST 이다. PTS는 평균 득점, REB는 평균 리바운드, AST는 평균 어시스트로 농구 기록들 중 가장 중요한 세가지 요소이기 때문에 위 변수들이 rating과 가장 높은 상관관계를 갖을 것으로 예상한다.

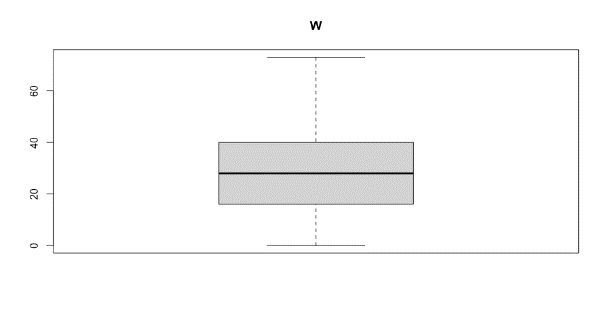
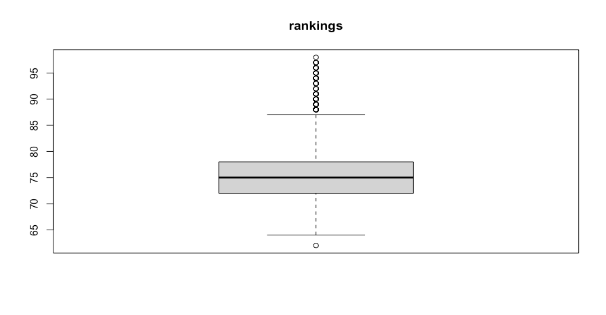
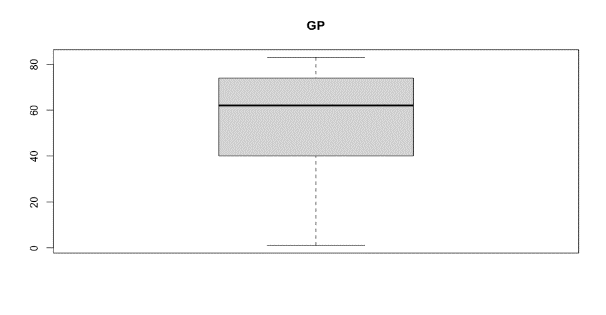
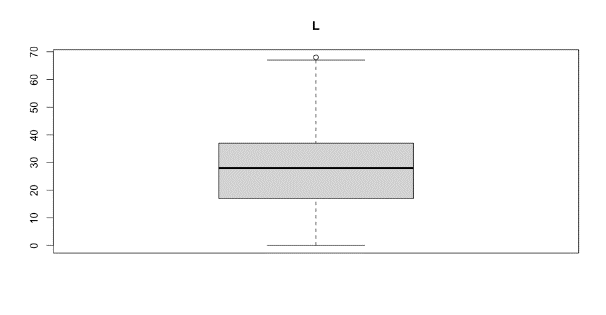
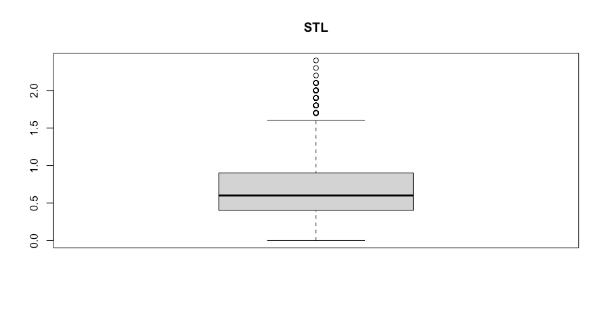
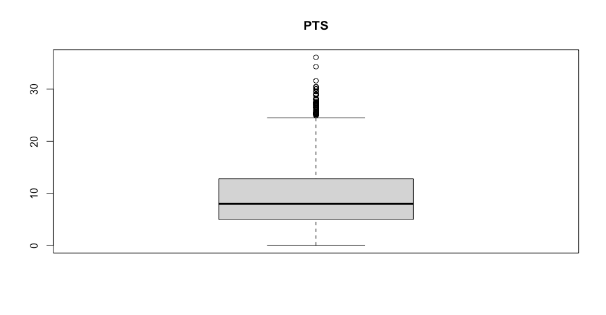
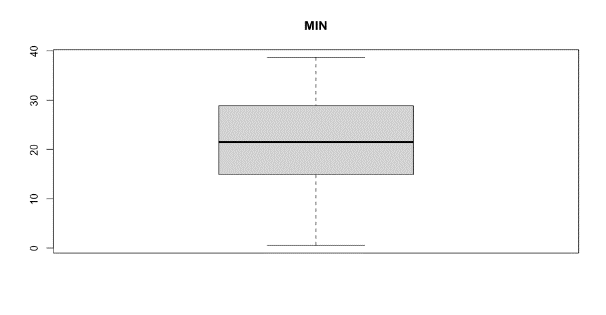
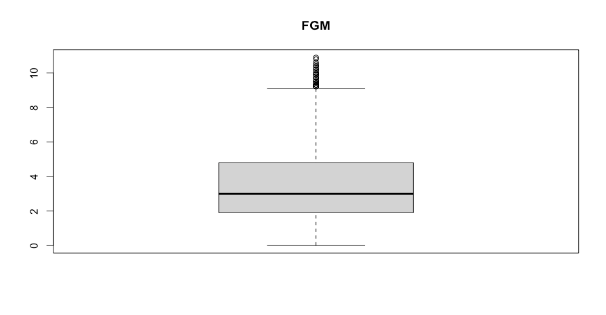
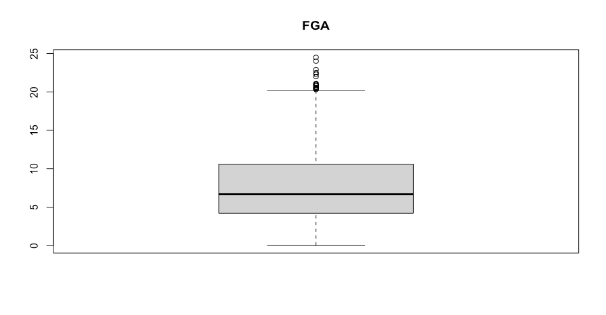
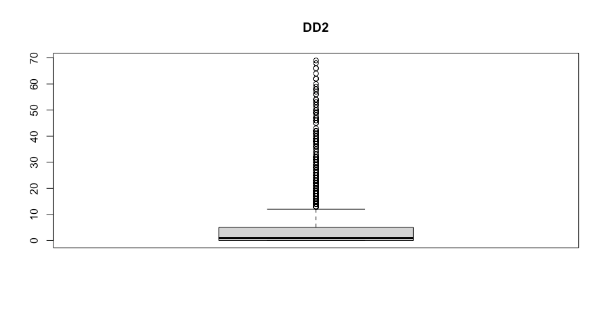
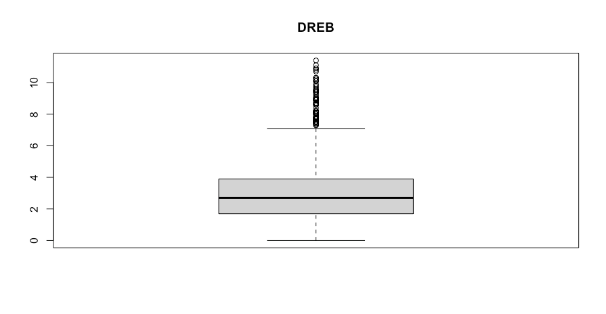
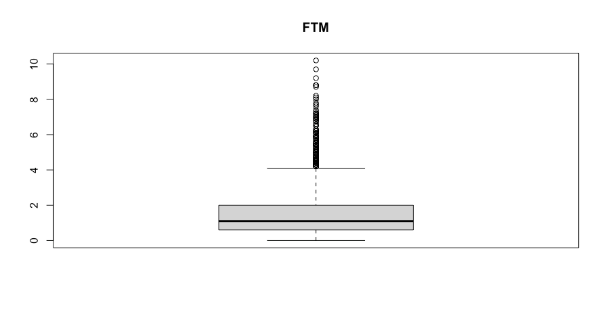
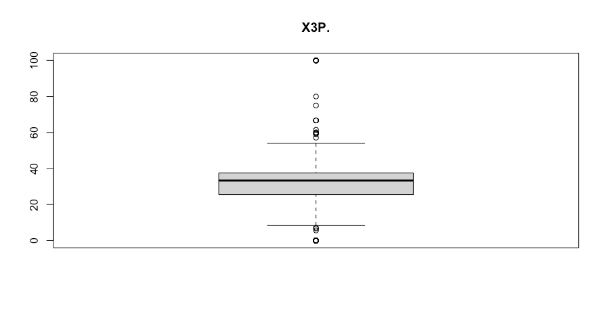
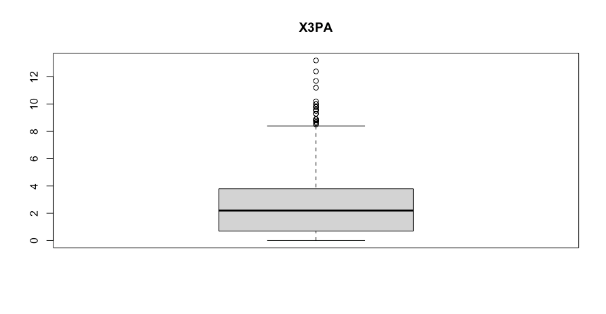
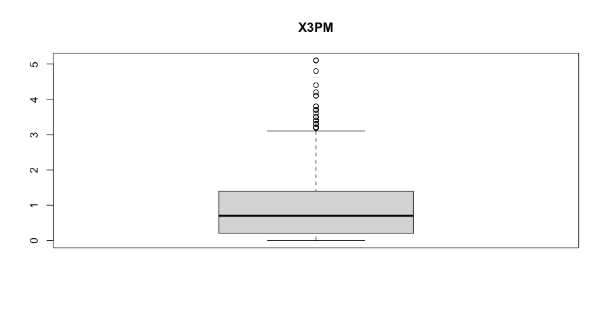
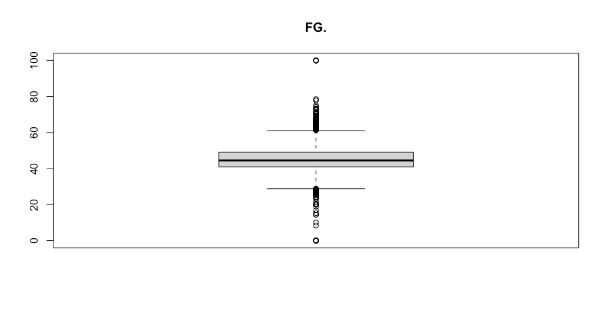
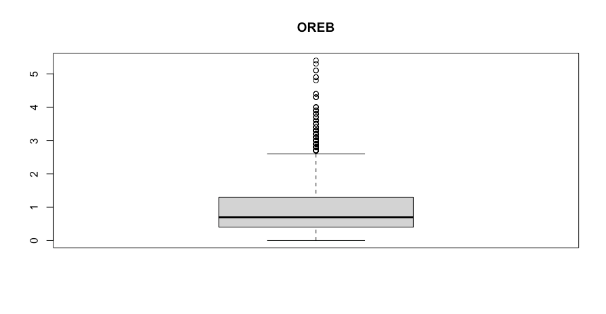
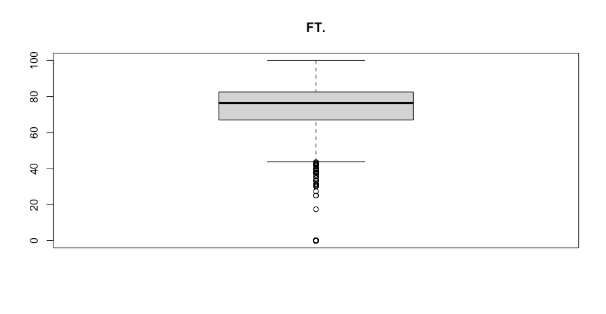
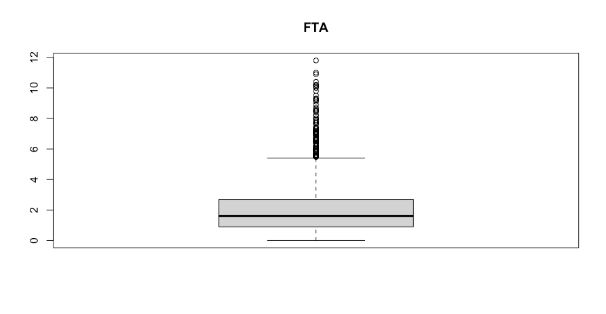
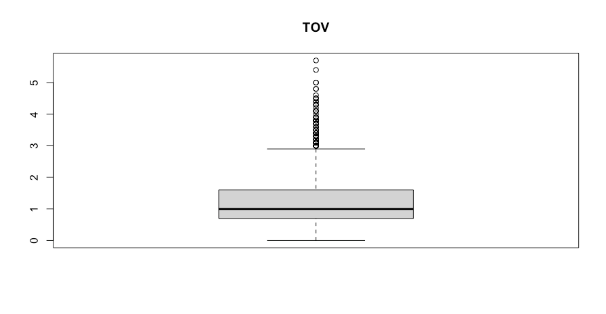
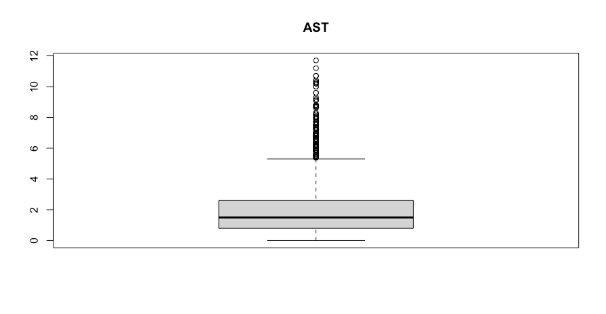
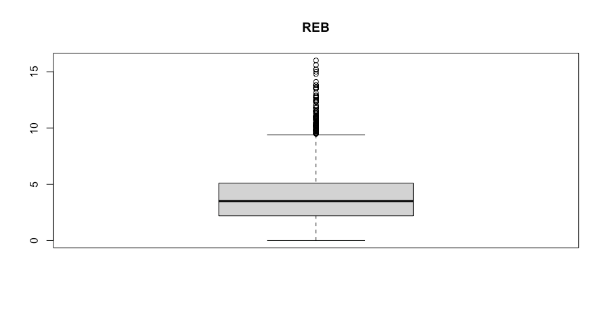
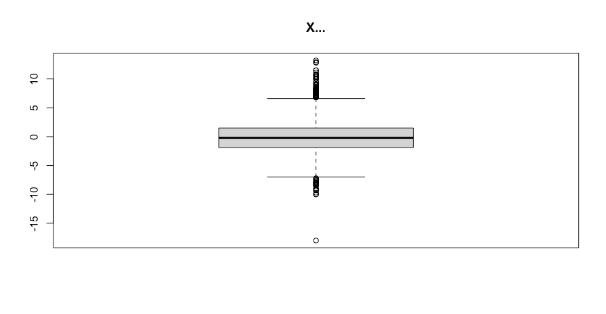
MLR모델 설계를 제외 해야한다고 판단되는 변수들은 player, team, age, season 이다. 해당 선수가 어느 팀의 소속인지, 선수의 나이 그리고 해당 선수가 뛴 시즌이 언제인지는 rating과는 무관할 것이라고 판단하기 때문이다.

테이블이(가) 표시된 사진

자동 생성된 설명

위 표는 describe함수를 통해 단변량분석을 실행한 결과이다. 이상적으로 정규성을 띄는 변수들은 skewness=0 그리고 kurtosis=3에 가까운 값을 가져야한다. 그러나 사실 현실세계에서 그러한 값을 갖고 정규성을 띄는 데이터를 찾는 건 사실상 불가능에 가깝다는 점을 감안해야한다. 결론적으로 이러한 사항들을 고려하면 FTM, BLK, DD2, TD3 변수들을 제외하면 나머지는 어느정도 정규성을 갖는다고 가정할 수 있겠다. 아래 그려진 boxplot들을 보았을 때도 해당 4개의 변수들의 boxplot의 생김새가 비교적 정규성이 떨어짐을 볼 수 있다.

테이블이(가) 표시된 사진

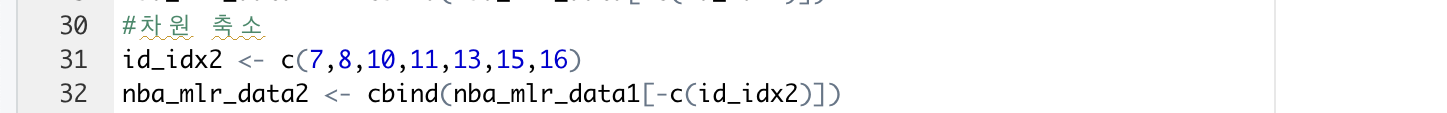
자동 생성된 설명

위 4개 변수들의 인덱스 번호는 FTM=13, BLK=22, DD2=25, TD3=26이며, FT, BLK, DD2, TD3를 제거하는 R코드는 아래와 같다. 

위에서 이미 변수들을 줄였음에도, 해당 데이터셋의 변수들이 23개로 워낙 많아, correlation과 scatter plot이 그려지지 않고 너무 작아 식별이 불가능하다. 따라서 변수들을 줄일 필요가 있다. 줄일 변수들은 서로 상관관계가 높고 줄여도 된다고 판단하는 것들을 줄였다. 제외 시킨 변수들은 아래와 같다:

1. FG%=FGM/FGA 이기때문에 FGM과 FGA는 FG%로 표현 가능하다 판단하여 FGM과 FGA를 제외시킨다.
2. X3P%=X3PM/X3PA이기때문에 X3PM과 X3PA는 X3P%로 표현 가능해서 X3PM과 X3PA를 제외시킨다.
3. FT%=FTM/FTA 이며 FT%에 FTM과 FTA가 포함된 것이다. 이미 Q4에서 FTM는 제외했기 때문에 FTA만 제거한다.
4. REB=OREB+DREB의 식으로 성립된다. 즉 REB라는 수치 안에 OREB와 DREB는 내제되어 있어서 이 두 변수를 소거한다.

아래는 해당 변수들을 MLR을 위한 데이터셋에서 삭제를 시켜주는 코드이다.

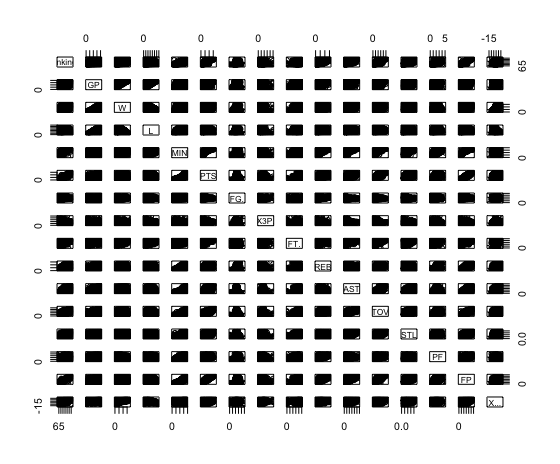


위 코드를 통해 변수들을 삭제 시켜주고, 아래 코드로 correlation과 scatter plot 을 그려준다.

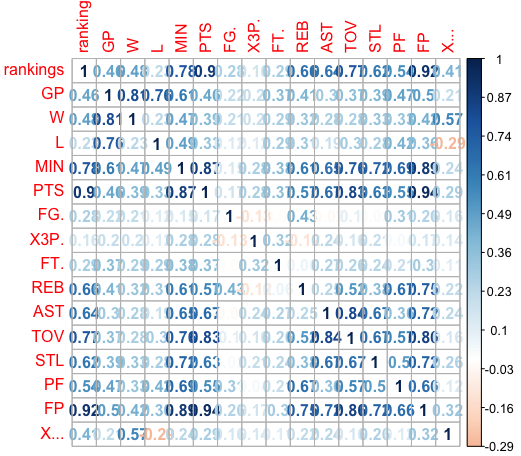
텍스트이(가) 표시된 사진

자동 생성된 설명

**Scatter plot**

****

**Correlation**

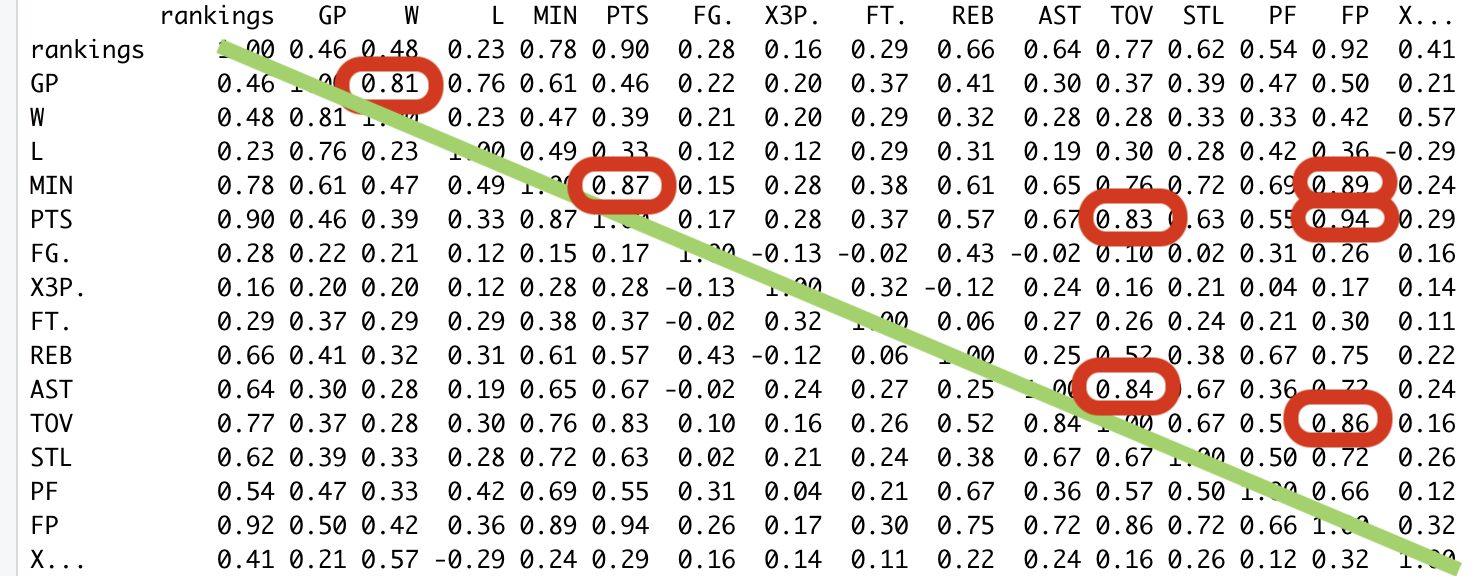
****

해당 그래프들이 변수를 줄였음에도 식별하기 힘들어 correlation함수를 roundup 하여 correlation을 직접적으로 비교한 결과는 아래와 같다.

텍스트이(가) 표시된 사진

자동 생성된 설명

이 코드를 통해 correlation 값을 소수점 2자리로 반올림 한다. 그리고 해당 코드의 결과는 아래와 같다.



위 표에서 빨간 윤곽으로 덮은 수치들이 서로 상관관계가 높은 변수들이다. (기준은 0.8이상이 높은 것으로 잡았다). 이 중 MIN과 PTS의 경우 출전시간이 길수록 득점이 높을 수 있기 때문에 0.87이라는 높은 상관계수를 갖는다.

**[Q6] 전체 데이터셋을 70%의 학습 데이터와 30%의 테스트 데이터로 분할한 후 모든 변수를 사용하여 MLR 모델을 학습해 보시오. Adjusted R2값을 통해 데이터의 선형성(linearity)을 판단해 보시오. Residual plot과 Q-QPlot을 도시하고 OrdinaryLeastSquare 방식의 Solution이 만족해야 하는 가정들이 만족될 만한 수준인지 정성적으로 판단해 보시오.**

아래 R코드를 통해 학습데이터와 테스트 데이터로 전체 데이터셋을 나누고, plot과 summary를 통해 MLR모델을 학습시켜 보았다.

텍스트이(가) 표시된 사진

자동 생성된 설명

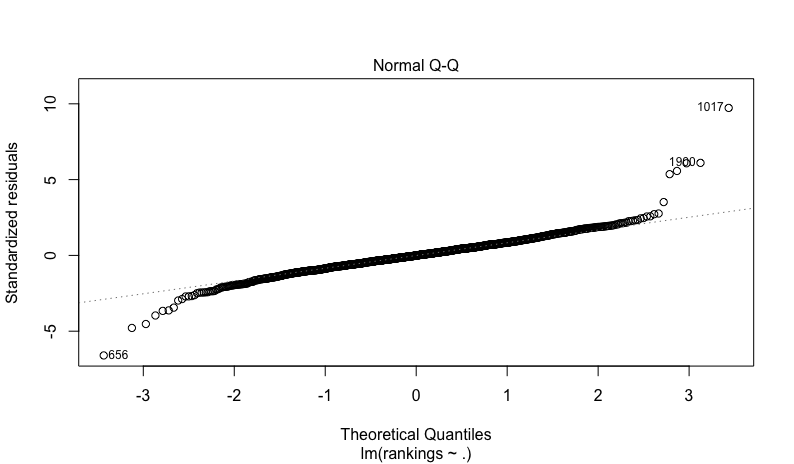
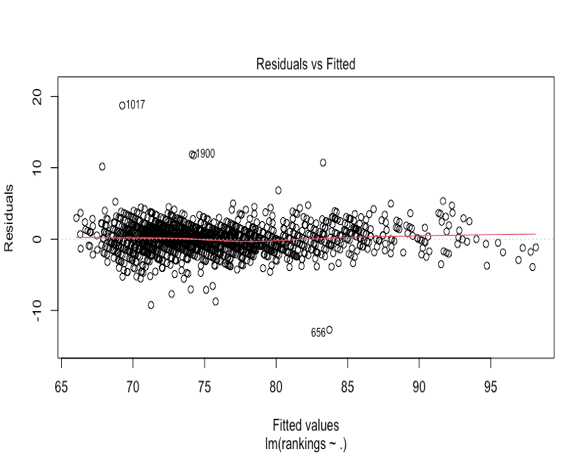
**Summary 결과**

텍스트이(가) 표시된 사진

자동 생성된 설명

Summary 결과를 통해 해당 모델의 adjusted R-squared값이 0.8873임을 볼 수 있다. 이는 실제 데이터임을 감안하면 상당히 높은 값인데, 이는 설명변수와 종속변수간의 선형관계가 성립함을 나타낸다.

**Plot결과**



Plot의 결과로 residual plot과 Q-Q plot을 볼 수 있다. 선형회귀분석의 중요한 가정 중 하나는 잔차의 정규성과 등분산성이다. 이때 등분산성을 우리는 homoskedasticity라 칭하는데, 위 residual plot을 보면 residual이 x값이 증가함에 따라 어떤 패턴을 갖거나 증가하지 않고 일정하게 분포됨을 볼 수 있다. 또한 추정선(빨강)이 평평하다. Q-Q plot에서는 -2와 2 구간을 지나 조금씩 추정선을 벗어나기 때문에 정규성을 띈다고 가정할 수 있겠다.

텍스트이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명

왼쪽에 있는 R 코드 중 summary(mlr\_nba)를 통해 우측에 있는 결과를 얻게 되었다. 결과들 중 보래색 테두리로 둘러싼 변수들이 유의미한 변수들이다. 이 변수들은 모두 ‘\*’이 2개 이상이다. ‘\*’이 두개 이상 있다는 것은 해당 변수가 유의미한 것을 의미하며 이때 이 변수들의 p-value값이 0.01보다 작다는 것이다. P-value 값이 낮으면 우리는 귀무가설(해당 변수는 종속변수 y에 영향을 끼치지 않다)를 기각하기에 대립가설(해당 변수는 종속변수에 y에 영향을 준다)이 성립된다. 따라서 위 변수들 중 intercept를 제외한 GP, W, MIN,PTS,FG%,X3P,FP,그리고 X..은 유의미한 변수들이다. 설명변수와 종속변수간의 관계가 양인지 음인지 판단을 위해서는 estimate 열에 해당되는 값들을 보면 되는데, 이때 이 값들은 MLR모델에서 해당 변수가 갖는 계수 값을 나타낸다. PTS, FG, FP 그리고 X.. 은 양의 관계, GP,MIN, X3P는 음의 관계를 갖는다.

텍스트이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명

R코드를 통해 해당 모델의 RMSE, MAE, MAPE를 구할 수 있었다. 이때 세 값들은 RMSE=1.96, MAE=1.495, 그리고 MAPE는 1.989로 나타났다. MAE=1.495는 게임 상 선수들의 ranking이 평균적으로 1.49 차이 나며, MAPE는 오차가 1.98임을 보여준다.

만약 해당 모델을 재설계 한다면 P-value값들이 높게 나타난 변수들 GP, W , MIN, PTS, FG, X3P%, FP 그리고 +/- 를 포함시키고 비록 P-value는 낮지만 농구선수의 중요한 기록인 AST와 REB를 포함하도록 하겠다.

**R-code**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

해당 코드를 통해 필요없는 변수들을 제거하였다. 이때 종속변수(rankings)를 제외한 나머지 변수들 GP,W,MIN, PTS, FG, X3P%, FP +/-,AST, REB를 포함하였고 총 32개의 변수에서 11개의 변수로 축소하였다.

**변수 제거 전 변수 제거 후**

테이블이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명

좌측의 결과는 변수들을 제거 전, 즉 총 16개의 변수를 포함한 데이터, 우측은 변수 제거 후 총 11개의 데이터만 남았을 때의 결과이다. 이때 제거 전 adjusted R-squared는 0.8873 그리고 제거 후는 0.8869로 제거 전과 후의 차이가 미미함을 볼 수 있다. Adjusted R-square 값이 감소한 이유는 몇몇 변수들을 제거하여 설명력을 잃었기 때문인데, 그럼에도 얼마 줄어들지 않은 것은 해당 MLR모델의 중요한 변수들을 남아있기 때문이다

텍스트이(가) 표시된 사진

자동 생성된 설명

변수 제거 전인 before 와 제거 후 after의 RMSE, MAE, MAPE를 보여준다. 이때 변수를 제거 후 RMSE, MAE 그리고 MAPE가 모두 미미하게 나마 증가하였는데 이는 앞서 변수들을 제거하며 모델의 설명력이 떨어져서 오차들이 증가하였기 때문이다.

추가적인 다변량분석을 위해 신뢰구간 “Confidence Interval”을 구해보았다. 각 변수들의 신뢰구간을 구하기 위한 R코드는 아래와 같다.

텍스트이(가) 표시된 사진

자동 생성된 설명

해당 코드를 통해 아래와 같은 결과를 얻을 수 있었다.

텍스트이(가) 표시된 사진

자동 생성된 설명

이때 alpha=0.05로 설정하고 해당 모델의 신뢰구간을 구하였다. 이때 중요한 점은 변수들의 신뢰구간(2.5%와 97.5% 사이)이 ‘0’의 값을 포함 하느냐의 여부이다. 이때 모든 신뢰구간이 0을 포함하지 않음을 볼 수 있다. 신뢰구간이 0 을 포함하지 않음으로 모든 변수의 계수는 ‘0’ 이라는 귀무가설을 기각함과 동시에 해당 11개의 변수들이 모두 종속변수 rankings에 영향을 끼침을 알 수 있다.