Exploring NBA Basketball Data

Group 12: Meghana Kantharaj, Mark Lerret,
Mehul Sharma, Brian Trippi

Research Questions

- What are the most important statistics in predicting success in the NBA?
- How much do individual player statistics contribute to a teams win rate?
- How are physical traits related to basketball performance?

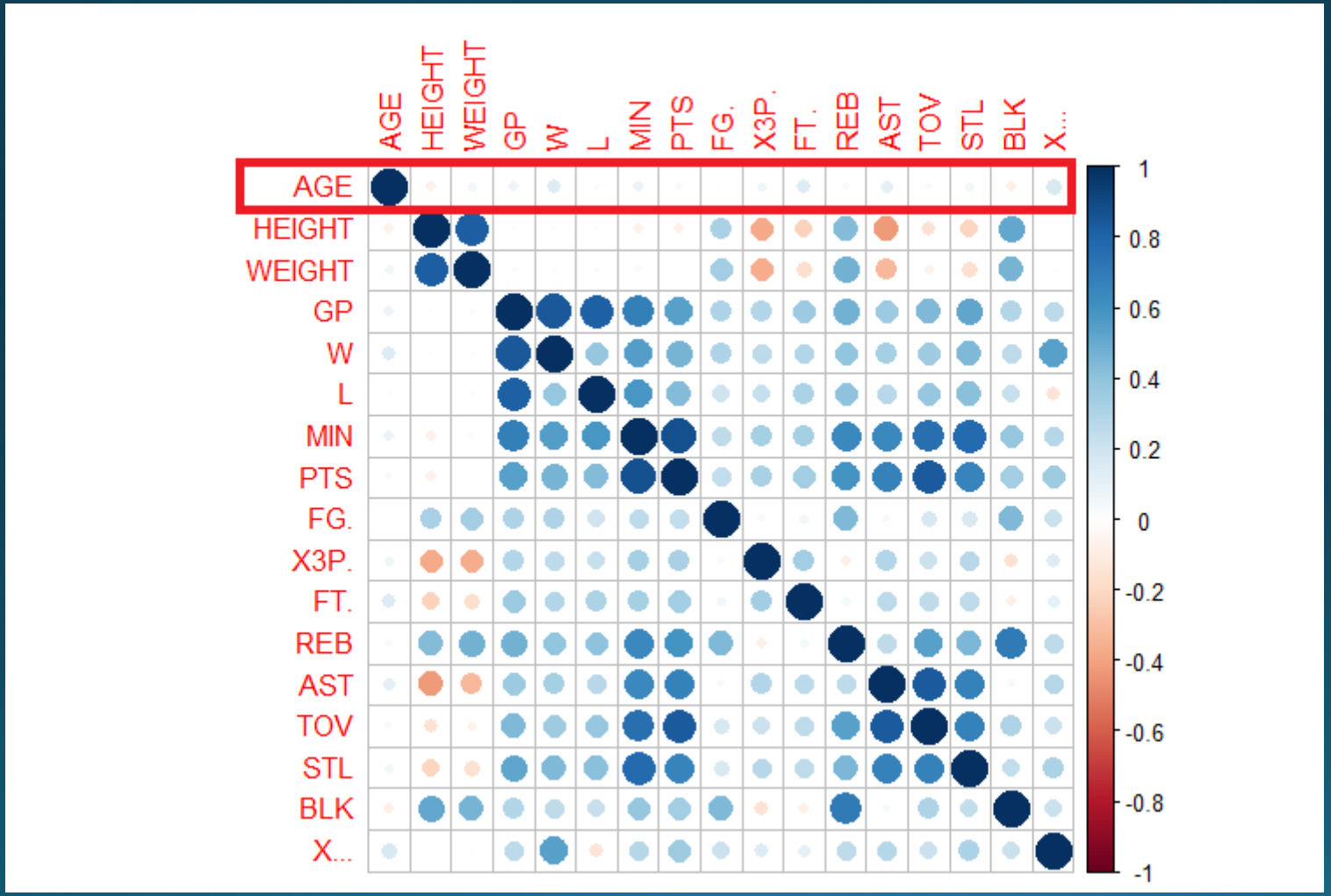
Why Should We Care?

- Professional sports are a big industry and statistics plays an important role in decision making
- Basketball is a proving ground for new statistical methods and is an area undergoing lots of innovation
- Stats.nba.com is a great resource for basketball statistics

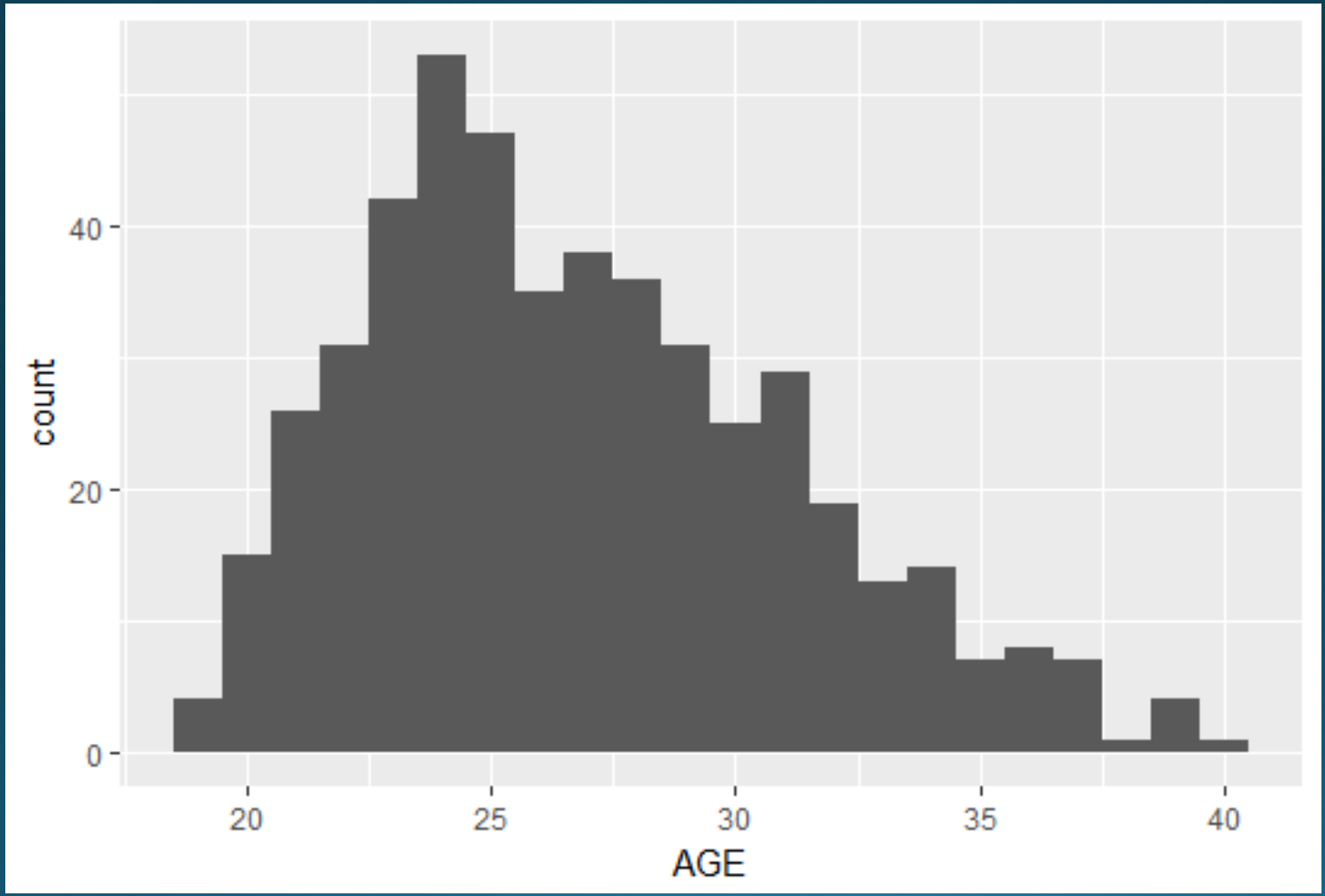
Description of the Dataset

- Data taken from the 2016-2017 season
- 32 Variables
- Statistics for 486 Players

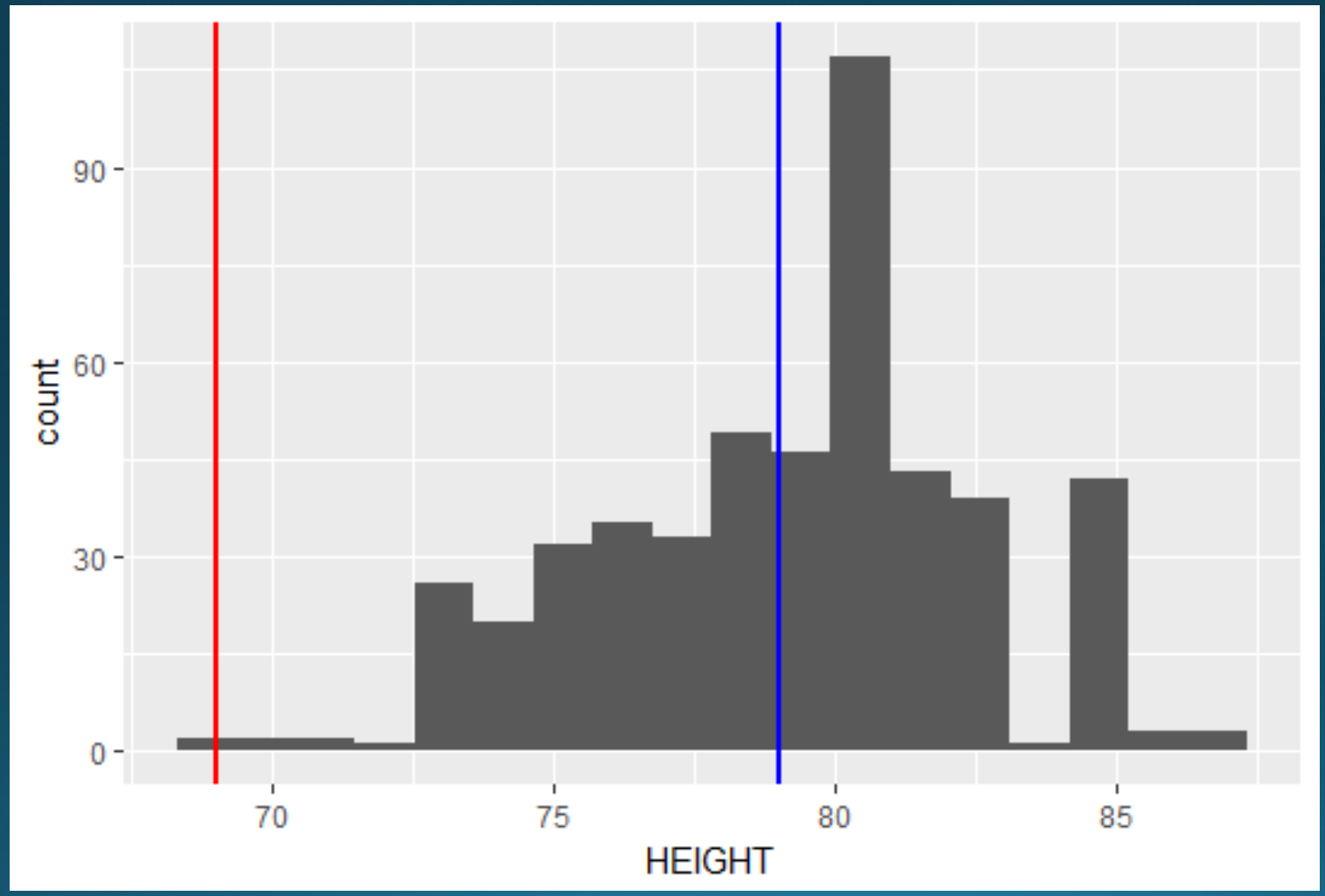
Correlation Matrix



Distribution of NBA Player Ages

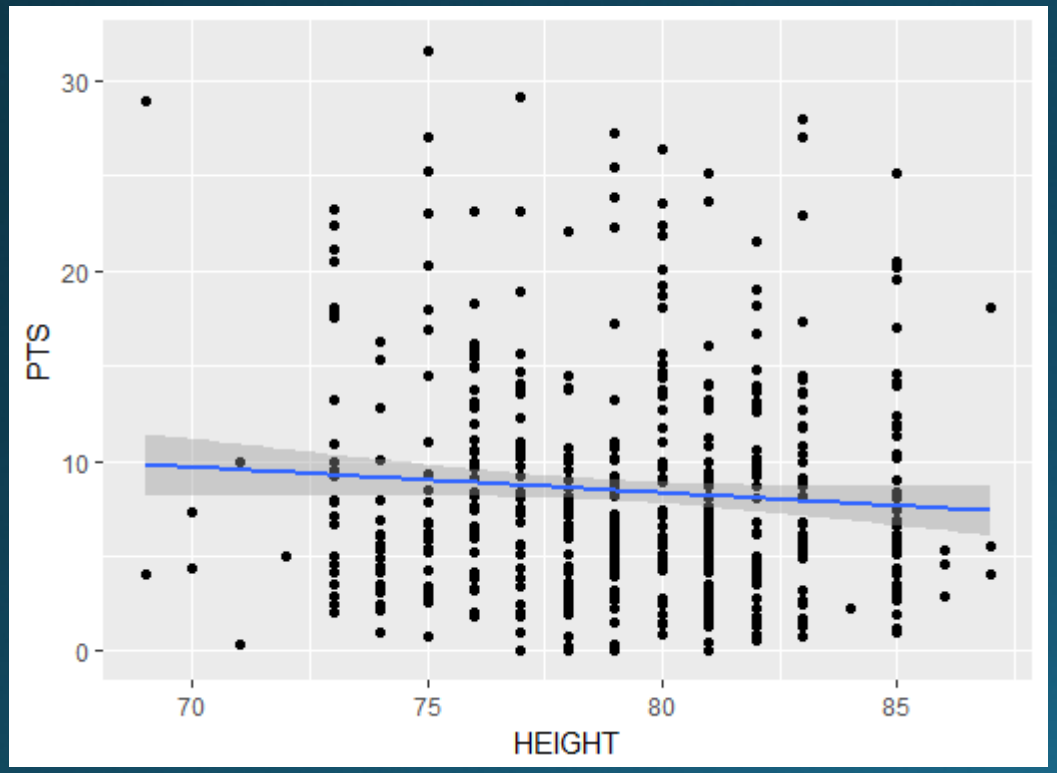


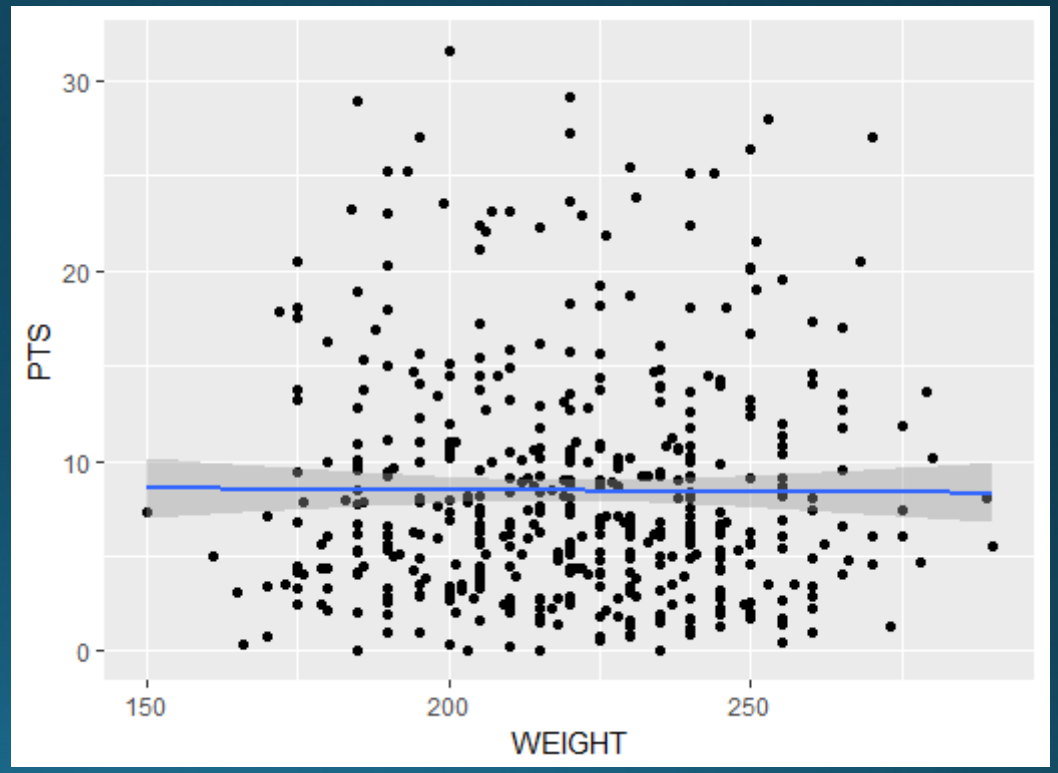
Distribution of NBA Player Heights



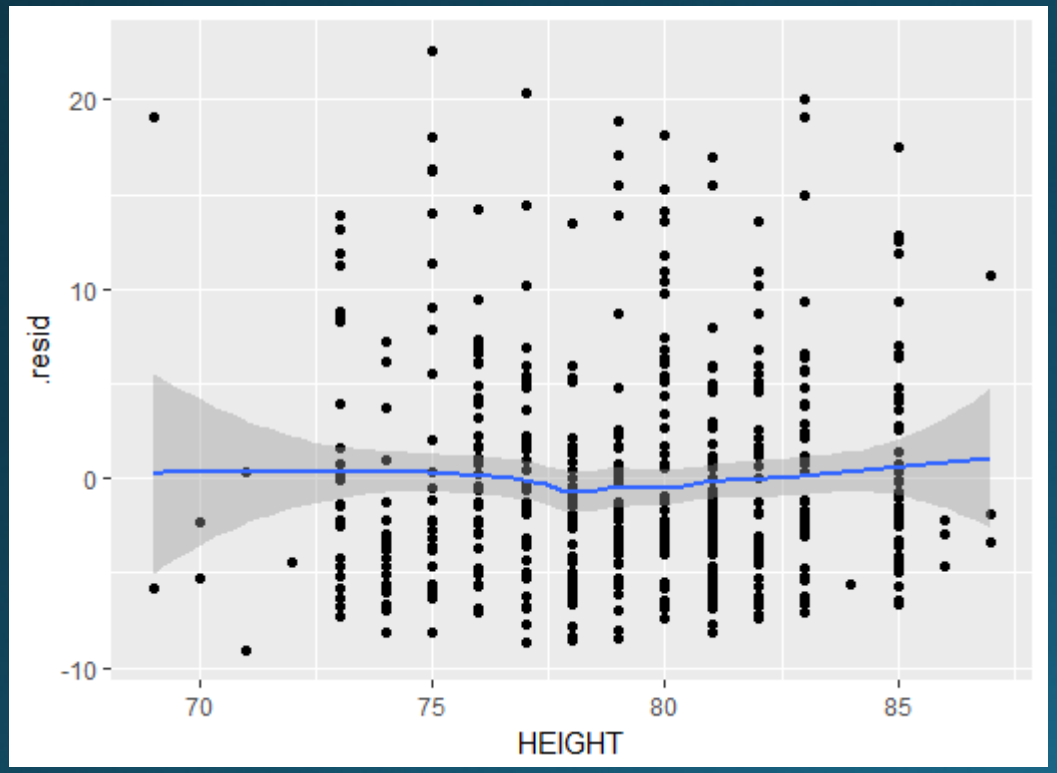
THE DATA

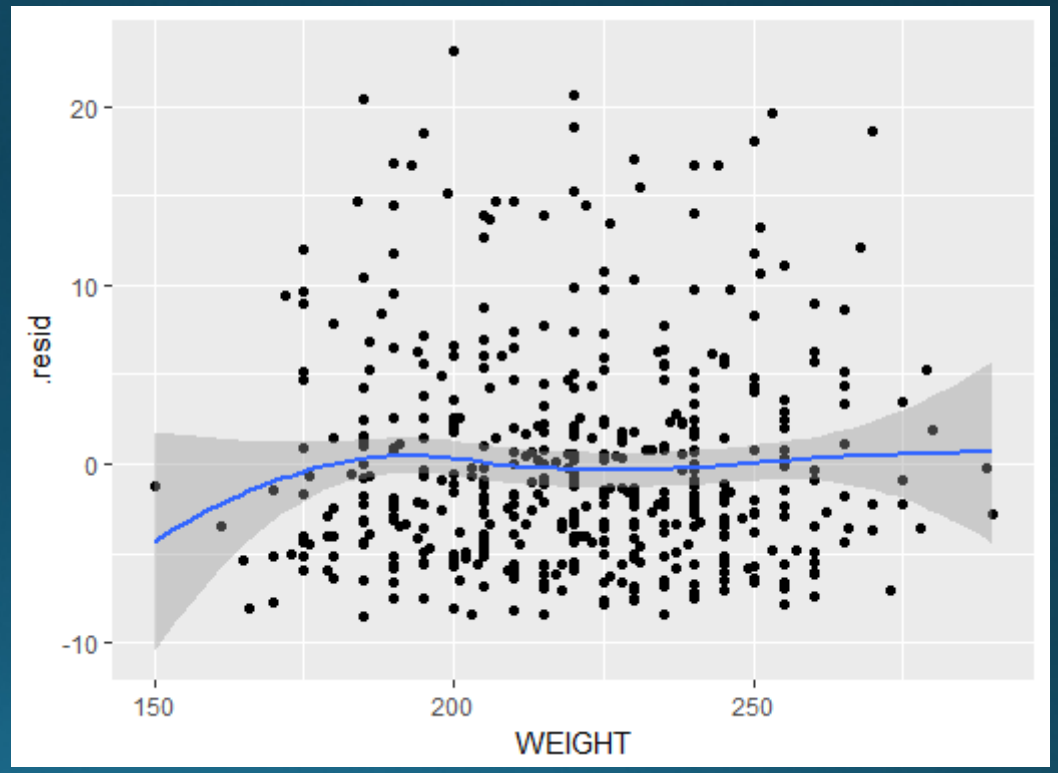
Points Per Game



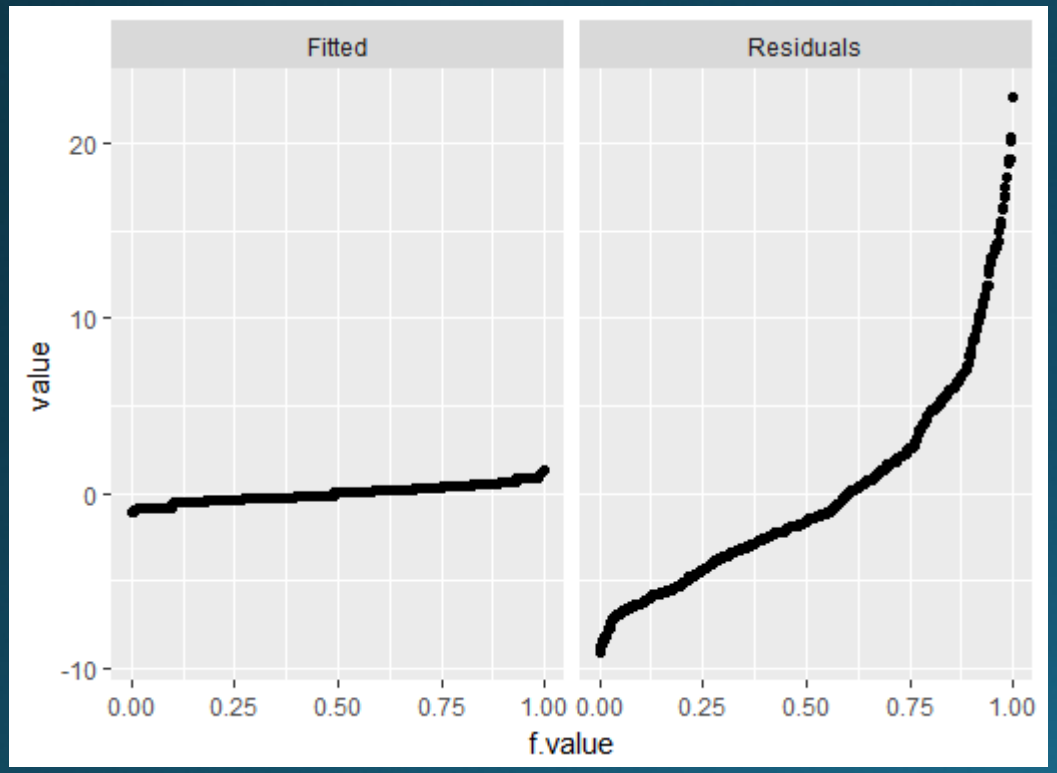


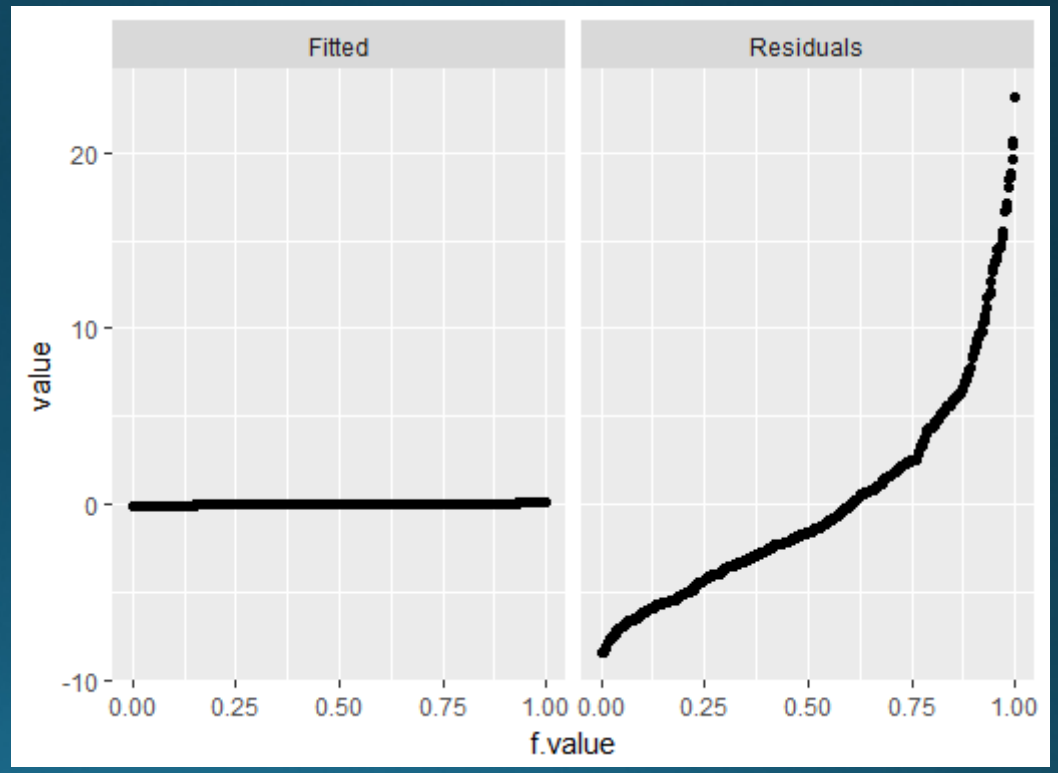
Residual Plots (PPG)



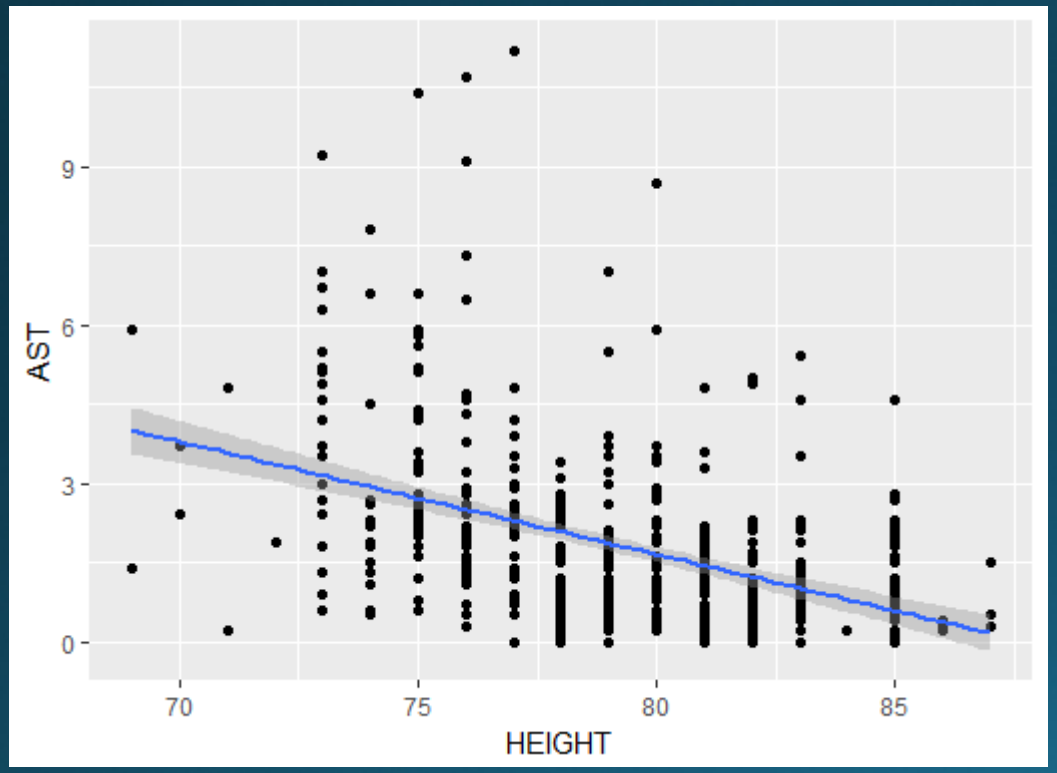


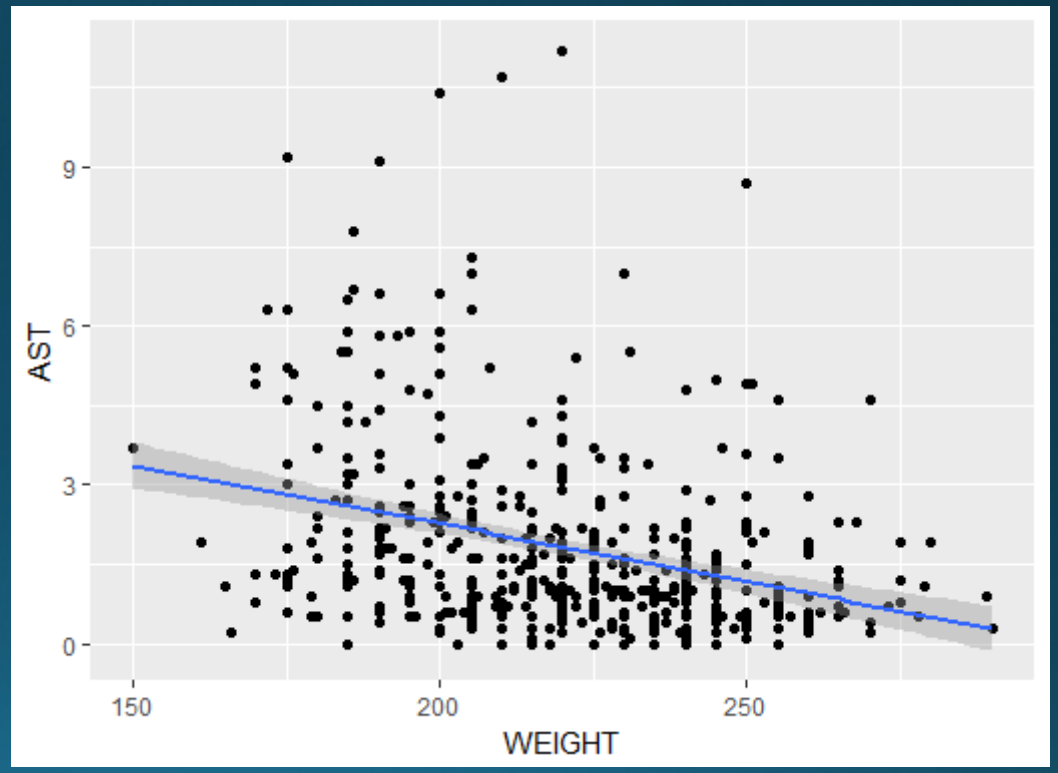
R-F Plots (PPG)


Fitted
Residuals
value
20
10
0
-10
0.00
0.25
0.50
0.75
1.00
f.value

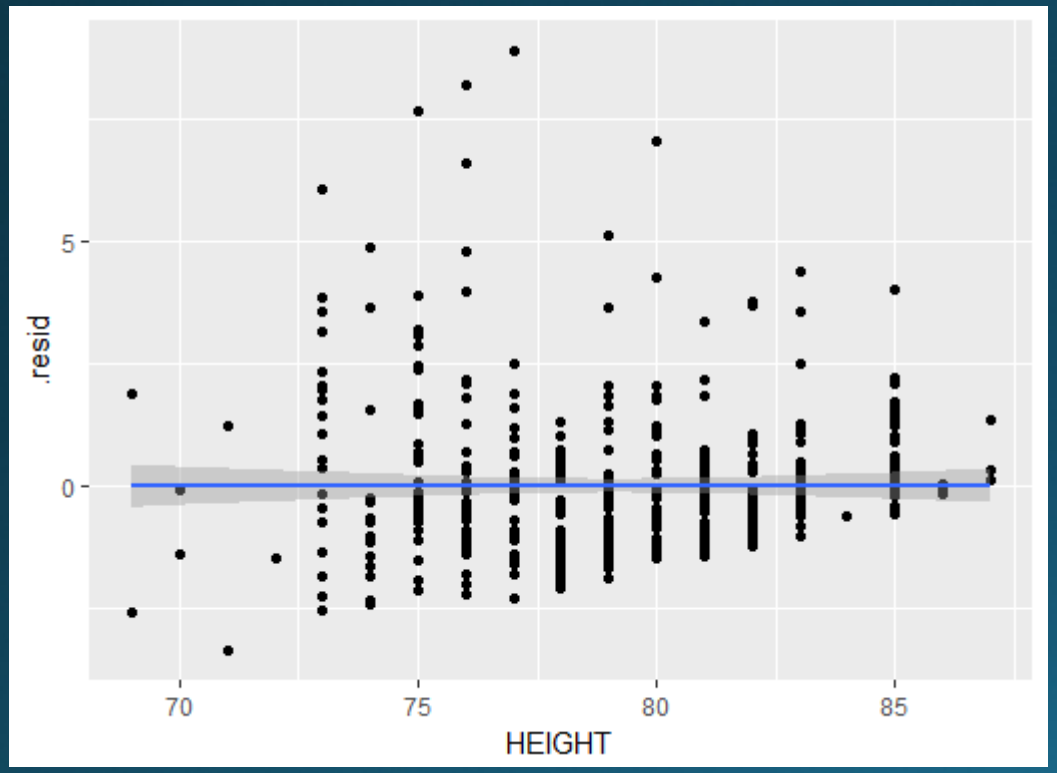

Fitted
Residuals
value
20
10
0
-10
0.00
0.25
0.50
0.75
1.00
f.value

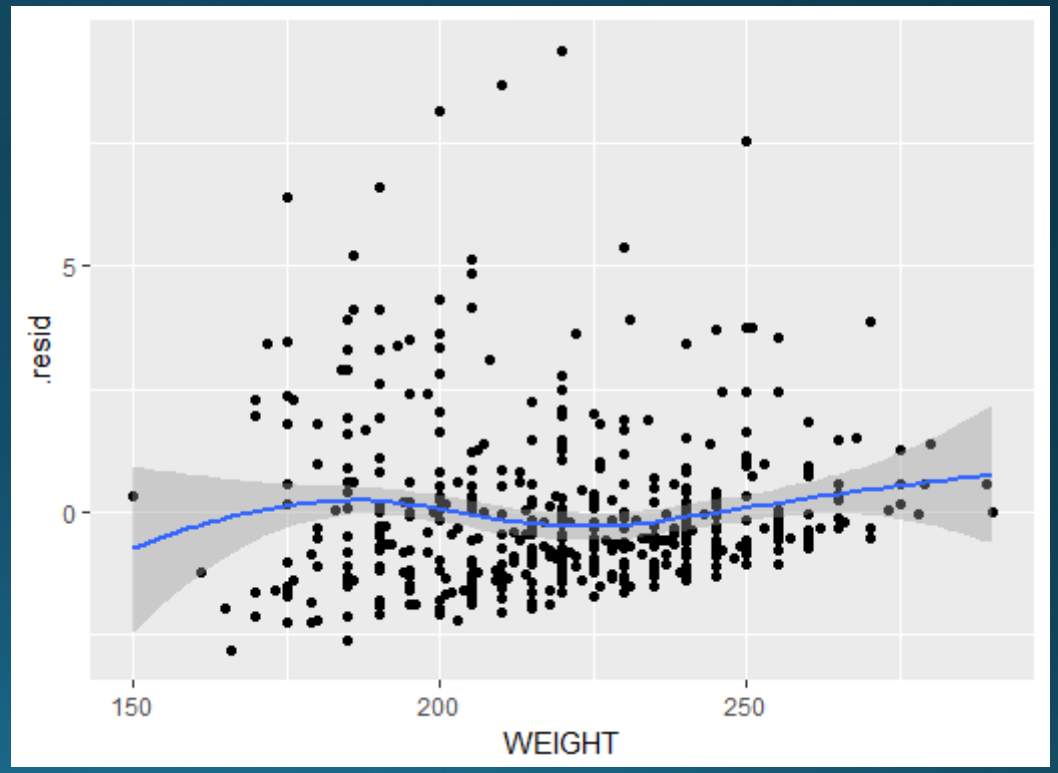
Assists Per Game



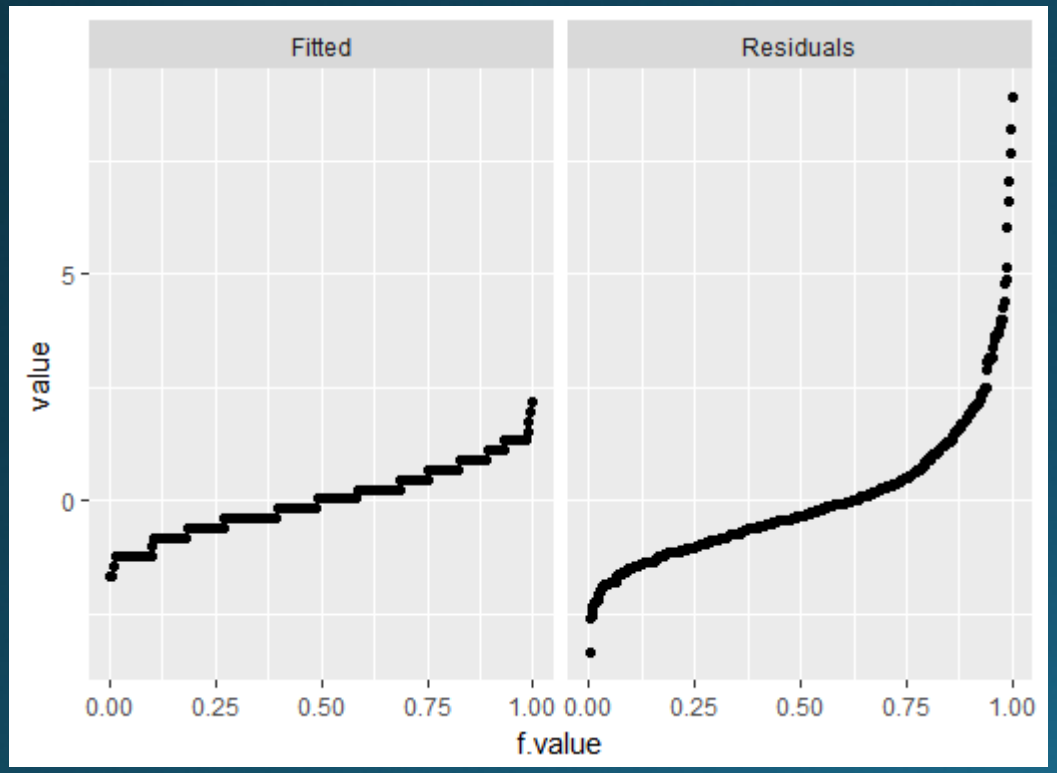


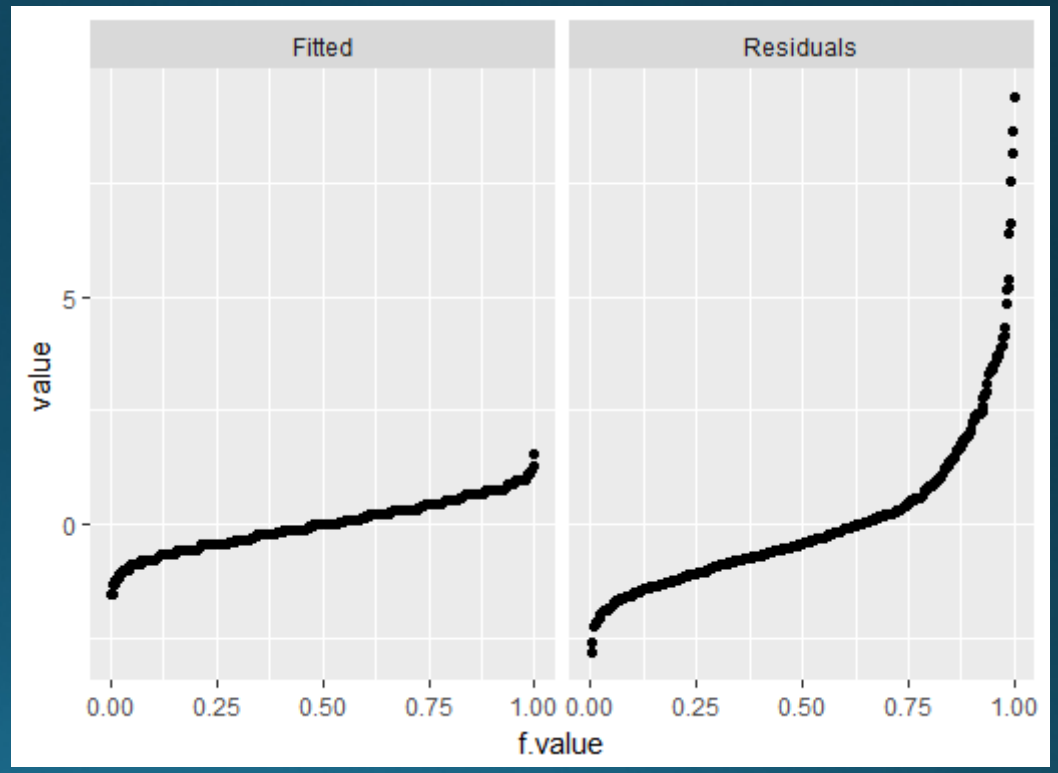
Residual Plots (APG)



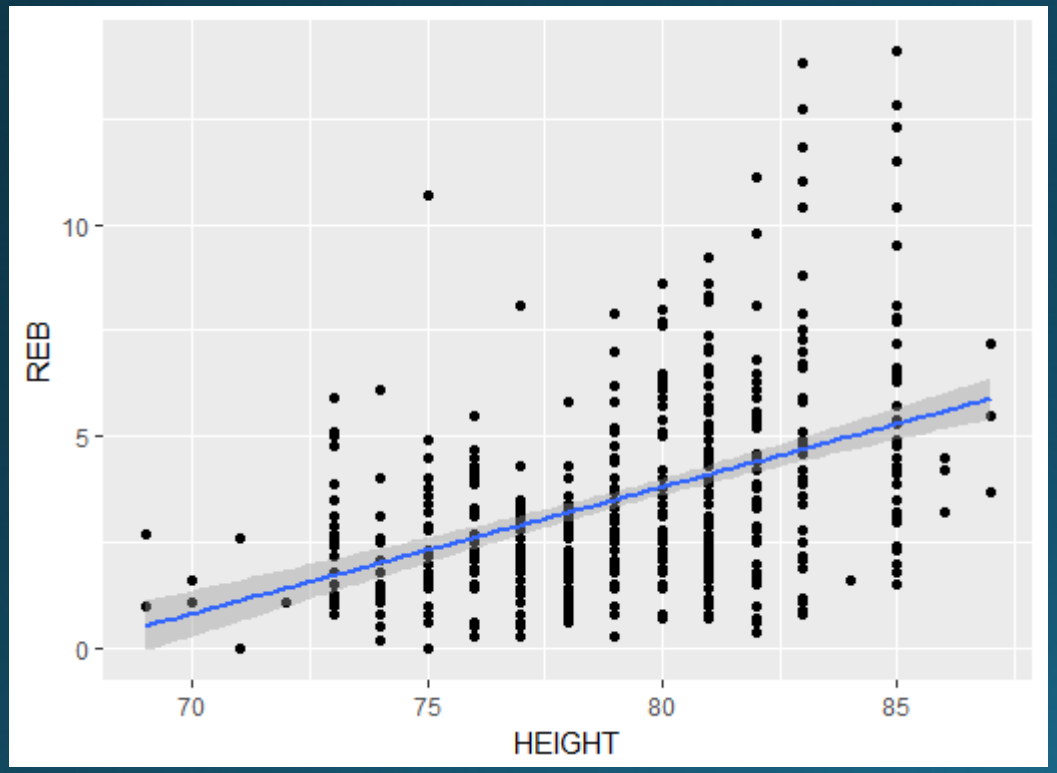


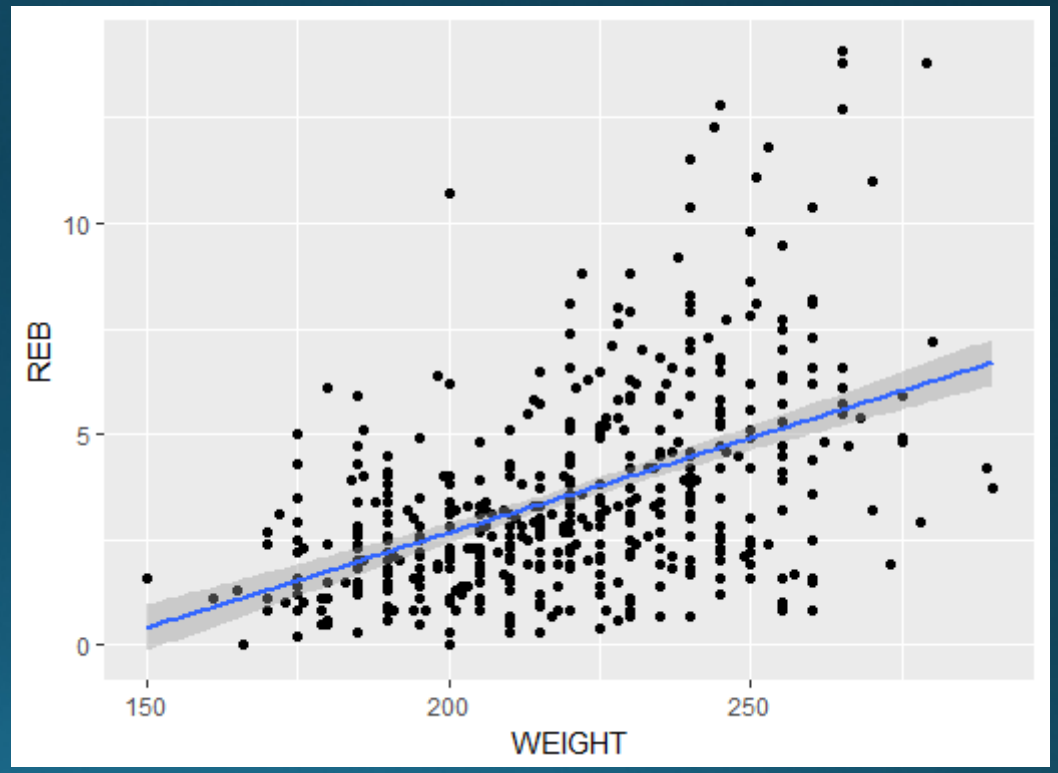
R-F Plots (APG)


Fitted
Residuals
value
5
0
0.00
0.25
0.50
0.75
1.00
f.value

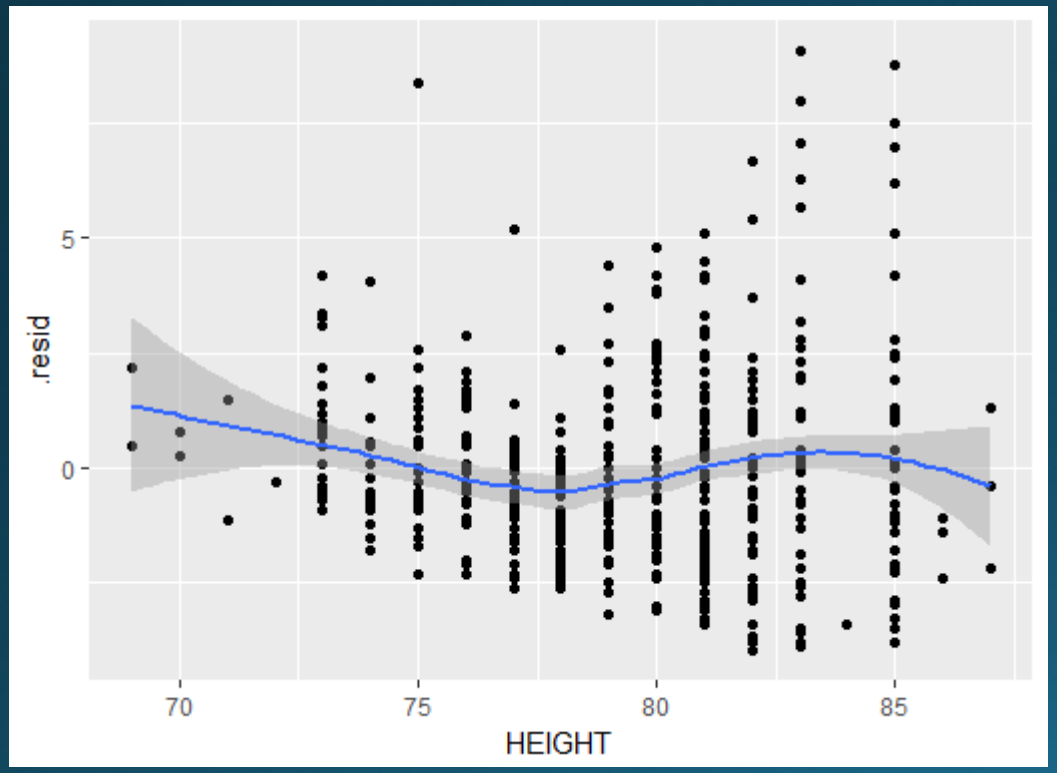

Fitted
Residuals
value
5
0
0.00
0.25
0.50
0.75
1.00
f.value

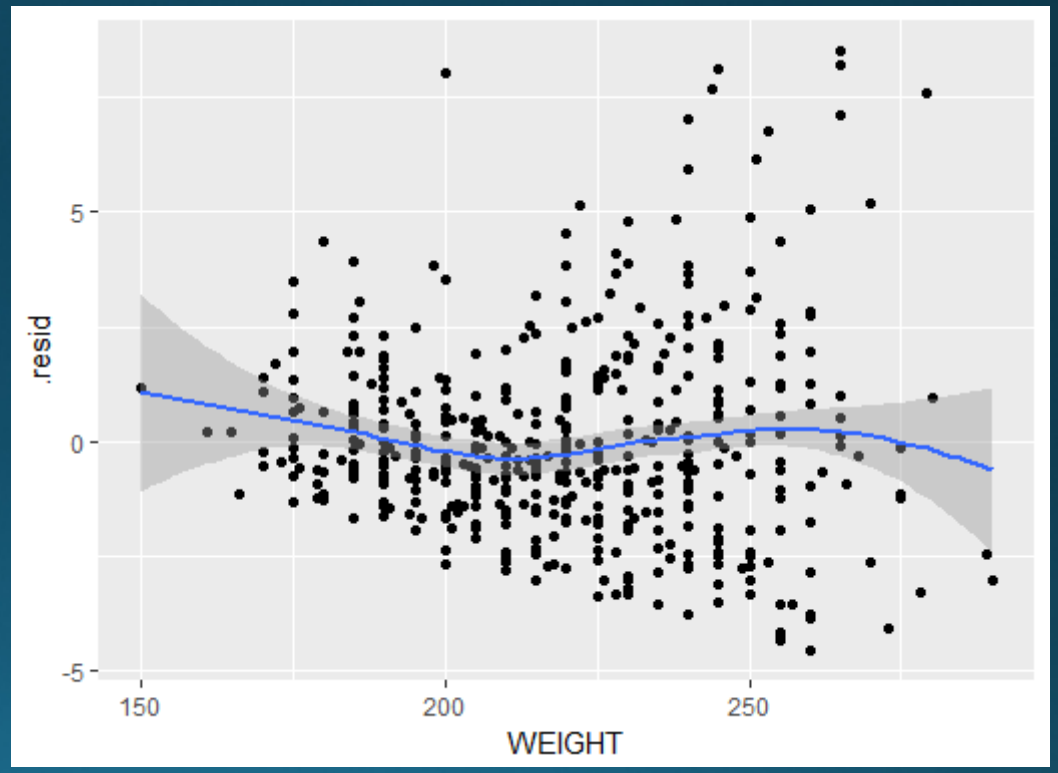
Rebounds Per Game



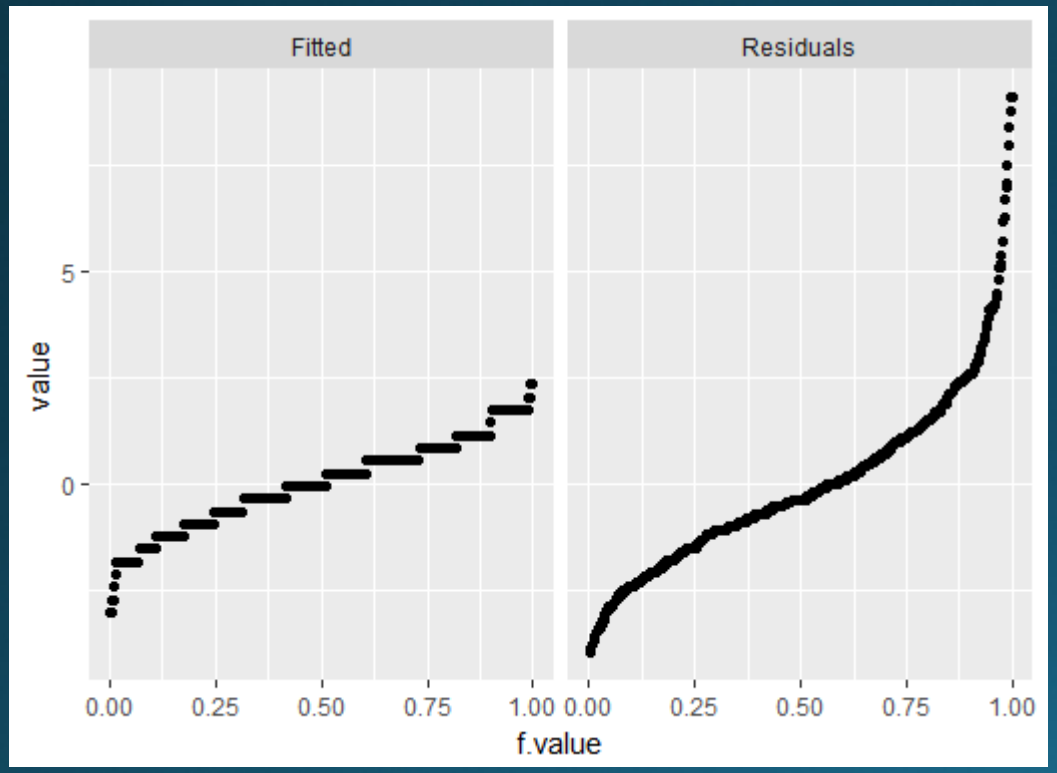


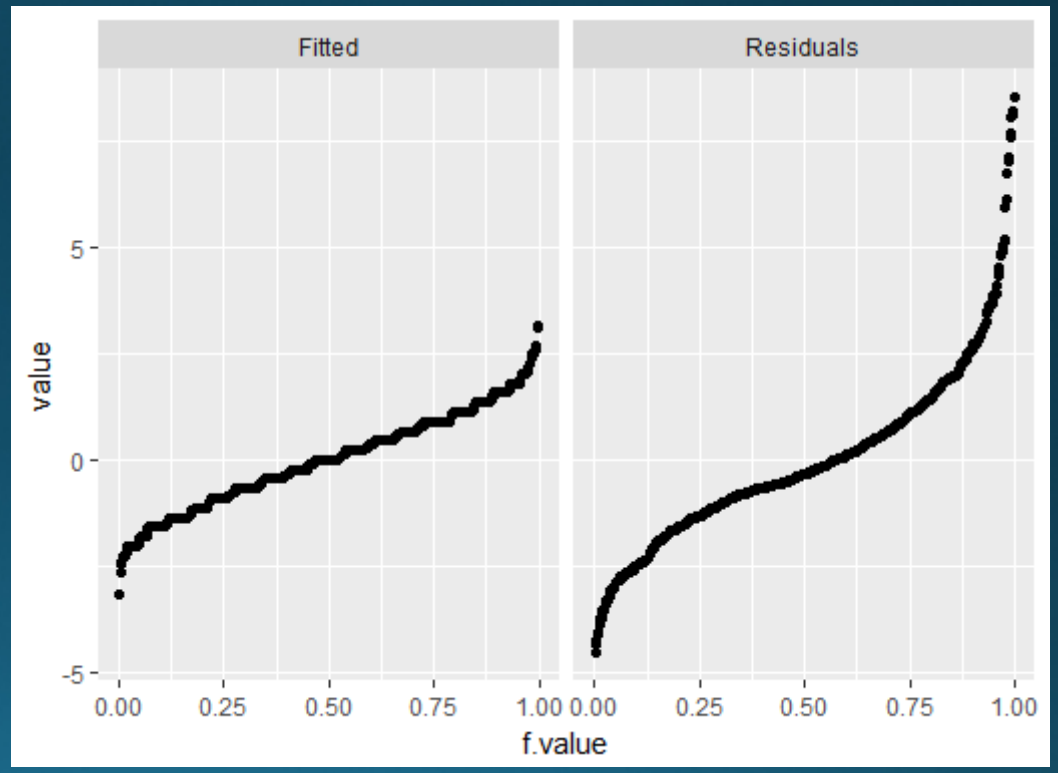
Residual Plots (RPG)



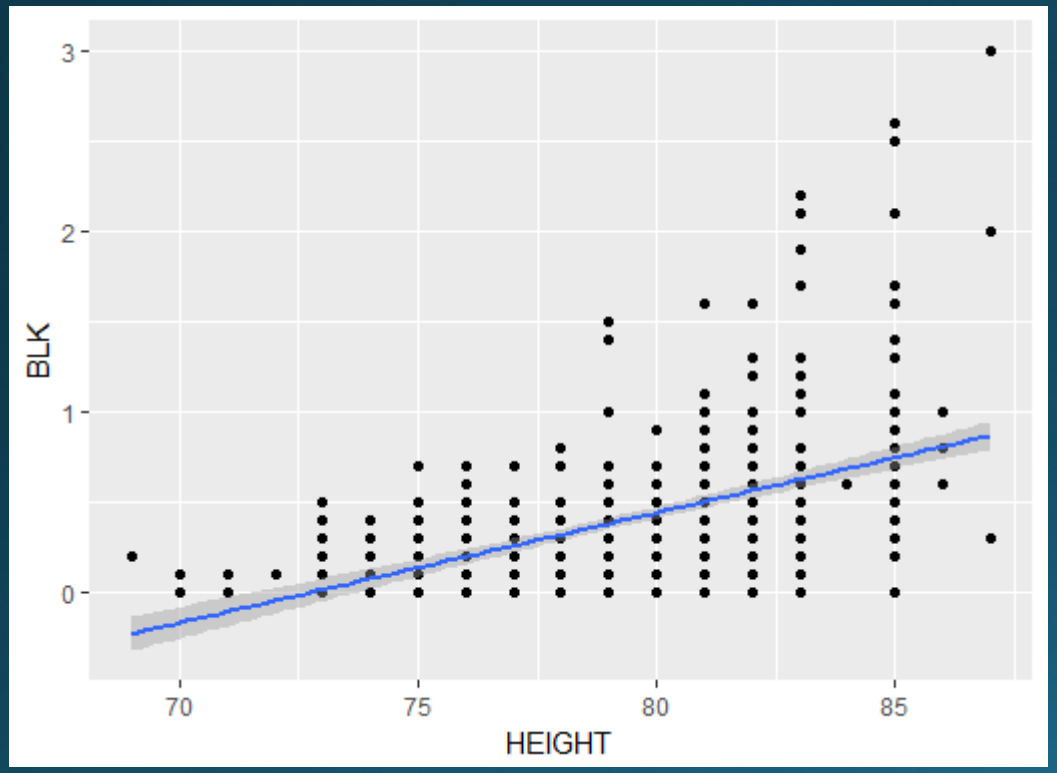


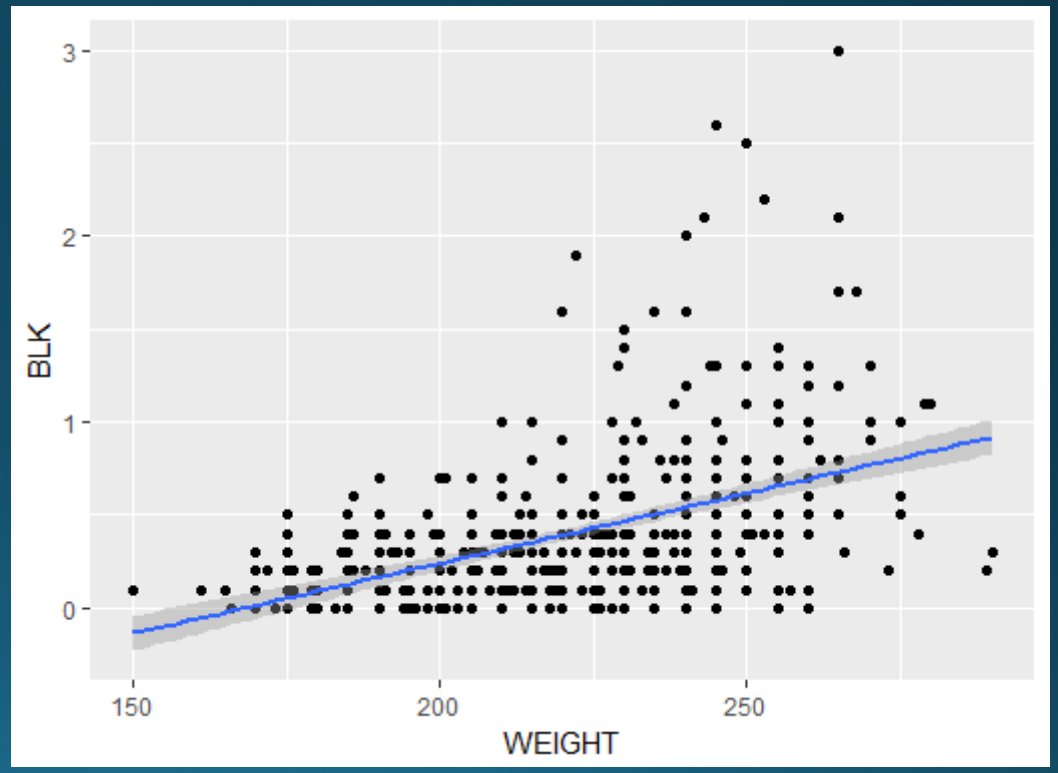
R-F Plots (RPG)


Fitted
Residuals
value
5
0
0.00
0.25
0.50
0.75
1.00
f.value

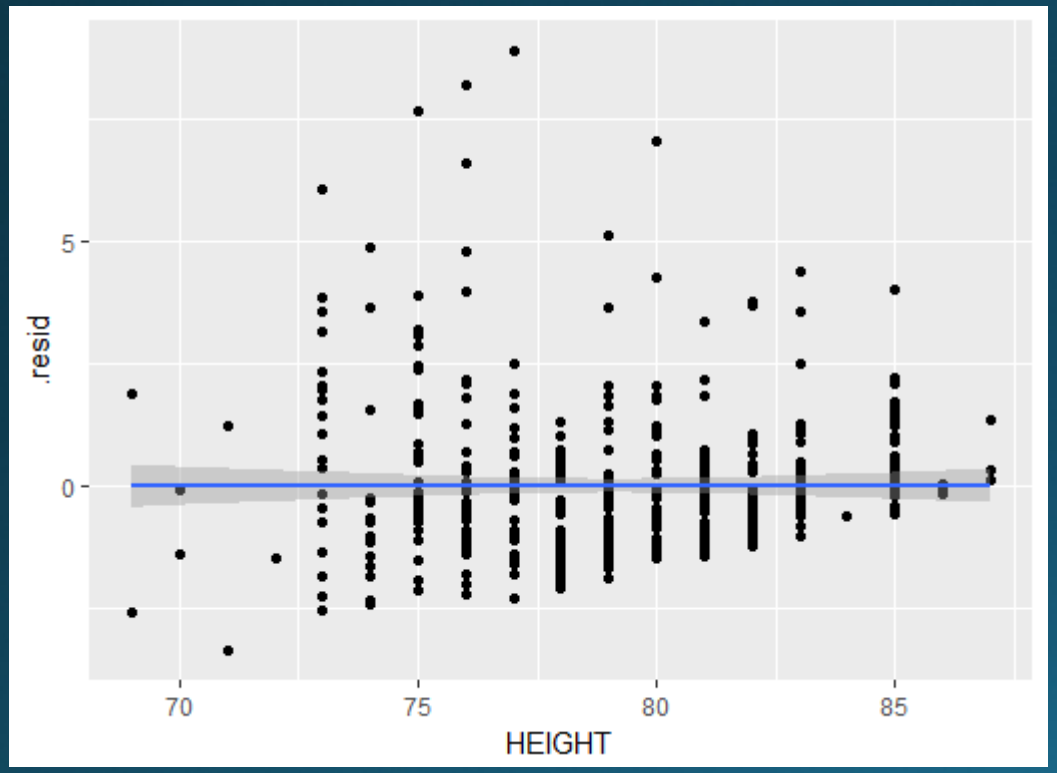

Fitted
Residuals
value
5
0
-5
0.00
0.25
0.50
0.75
1.00
f.value

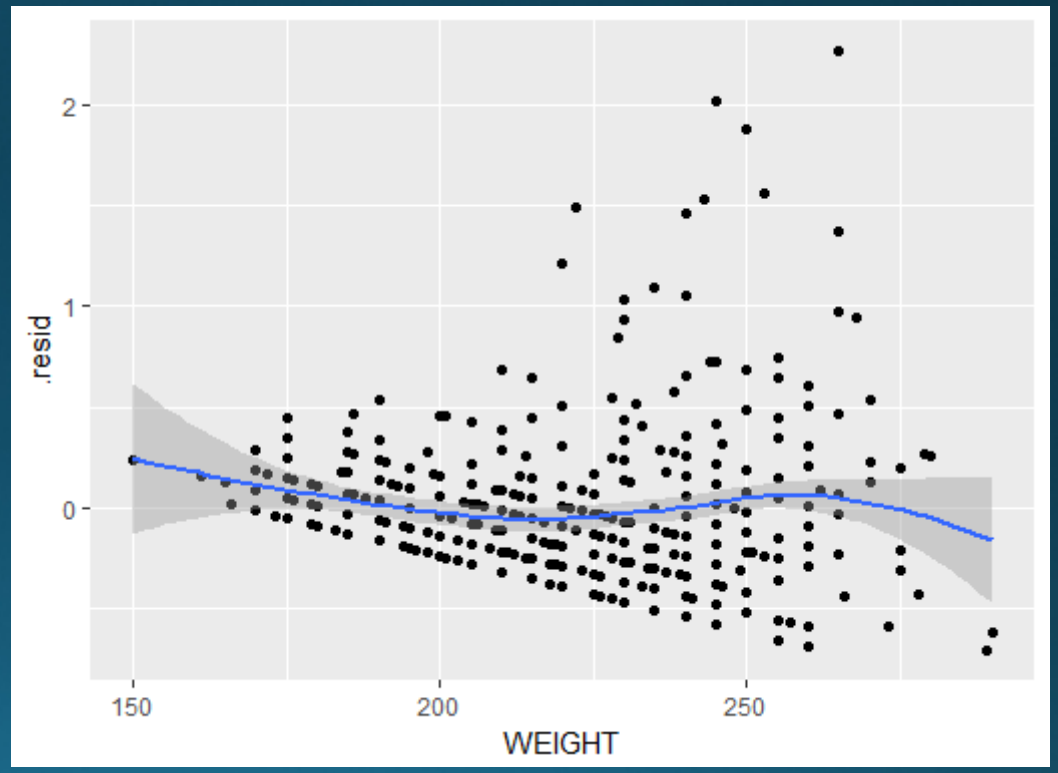
Blocks Per Game



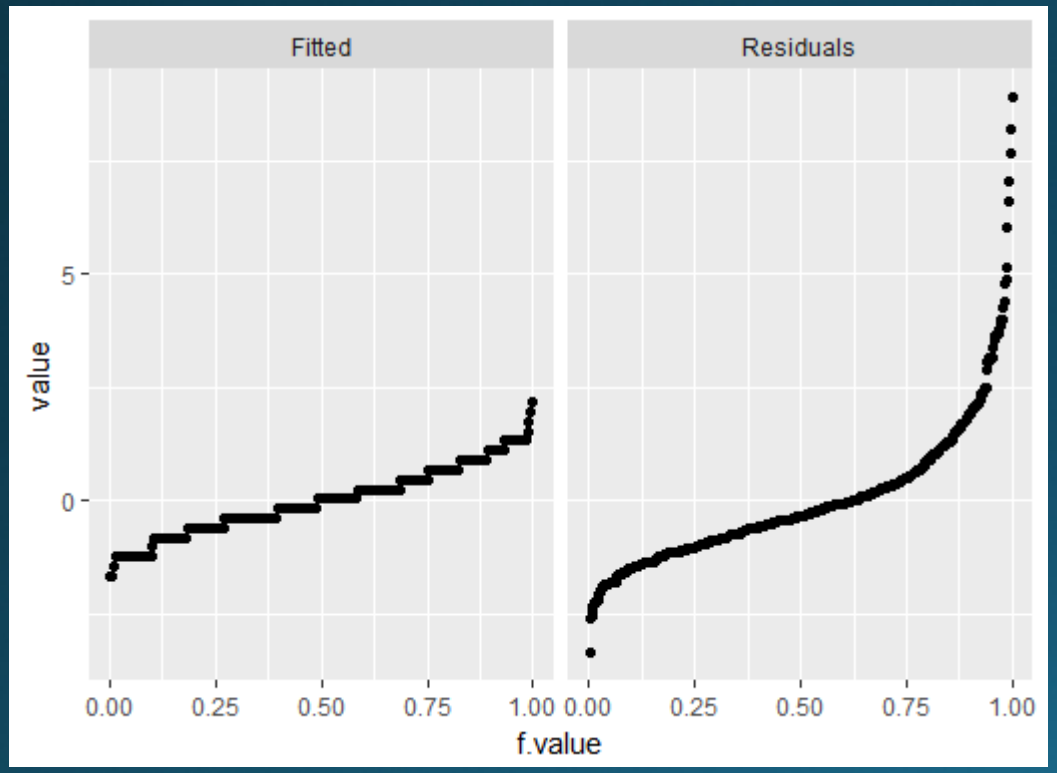


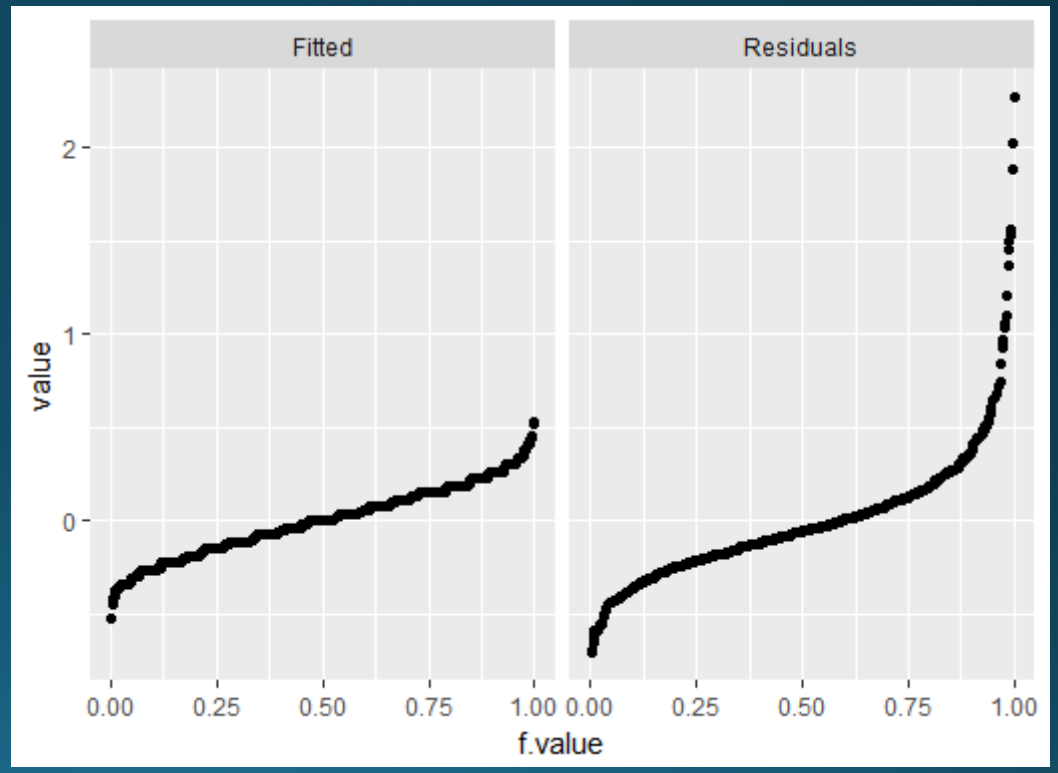
Residual Plots (BPG)





R-F Plots (BPG)


Fitted
Residuals
value
5
0
0.00
0.25
0.50
0.75
1.00
f.value


Fitted
Residuals
value
2
1
0
0.00
0.25
0.50
0.75
1.00
f.value

HOW TO EVALUATE PLAYER PERFORMANCE?

WELL KNOWN METRICS-

-PER (Player Efficiency Rating)

-PIE (Player Impact Estimate)

-EFF (efficiency)

PROPOSED METRIC-

-W/GP (Wins/Games Played)

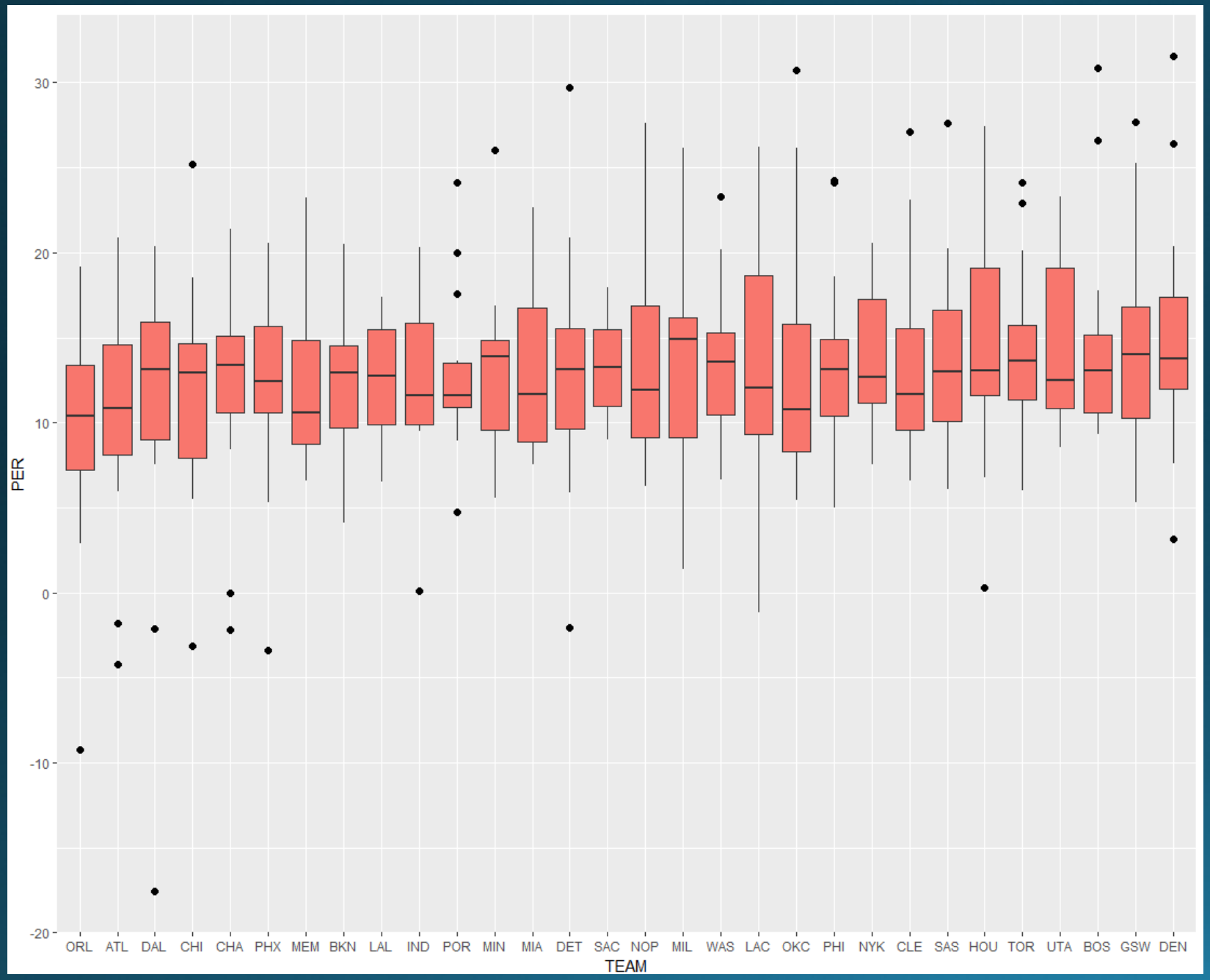
Variables stats-

```
   PLAYER              TEAM          AGE             HEIGHT          WEIGHT          GP              W
Length:486         DAL    : 19   Min.   :19.00   Min.   :69.00   Min.   :150.0   Min.   : 1.00   Min.   : 0.00
Class :character   CLE    : 19   1st Qu.:24.00   1st Qu.:77.00   1st Qu.:200.0   1st Qu.:35.25   1st Qu.:15.00
Mode  :character   ORL    : 18   Median :26.00   Median :79.00   Median :220.0   Median :62.50   Median :27.00
                   ATL    : 18   Mean   :26.85   Mean   :79.19   Mean   :220.1   Mean   :53.78   Mean   :26.87
                   BKN    : 18   3rd Qu.:30.00   3rd Qu.:82.00   3rd Qu.:240.0   3rd Qu.:75.00   3rd Qu.:38.00
                   CHA    : 17   Max.   :40.00   Max.   :87.00   Max.   :290.0   Max.   :82.00   Max.   :66.00
                   (Other):377
       L                MIN              PTS              FP              PER            W_by_GP
Min.   : 1.00   Min.   : 0.80   Min.   : 0.000   Min.   : 0.400   Min.   :-17.550   Min.   :0.0000
1st Qu.:16.00   1st Qu.:12.72   1st Qu.: 4.125   1st Qu.: 9.225   1st Qu.:  9.803   1st Qu.:0.3750
Median :28.00   Median :19.10   Median : 6.800   Median :15.500   Median : 12.865   Median :0.4878
Mean   :26.91   Mean   :19.90   Mean   : 8.427   Mean   :17.398   Mean   : 13.070   Mean   :0.4806
3rd Qu.:37.00   3rd Qu.:27.00   3rd Qu.:10.975   3rd Qu.:23.125   3rd Qu.: 15.875   3rd Qu.:0.6000
Max.   :59.00   Max.   :37.80   Max.   :31.600   Max.   :60.600   Max.   : 31.530   Max.   :0.8889
```

OBSERVATIONS-

- HEIGHT AND WEIGHT
- POINTS AND FANTASY
- AGE
- PER AND W_by_GP

HOW DO INDIVIDUAL PLAYER STATS INFLUENCE THE TEAM?

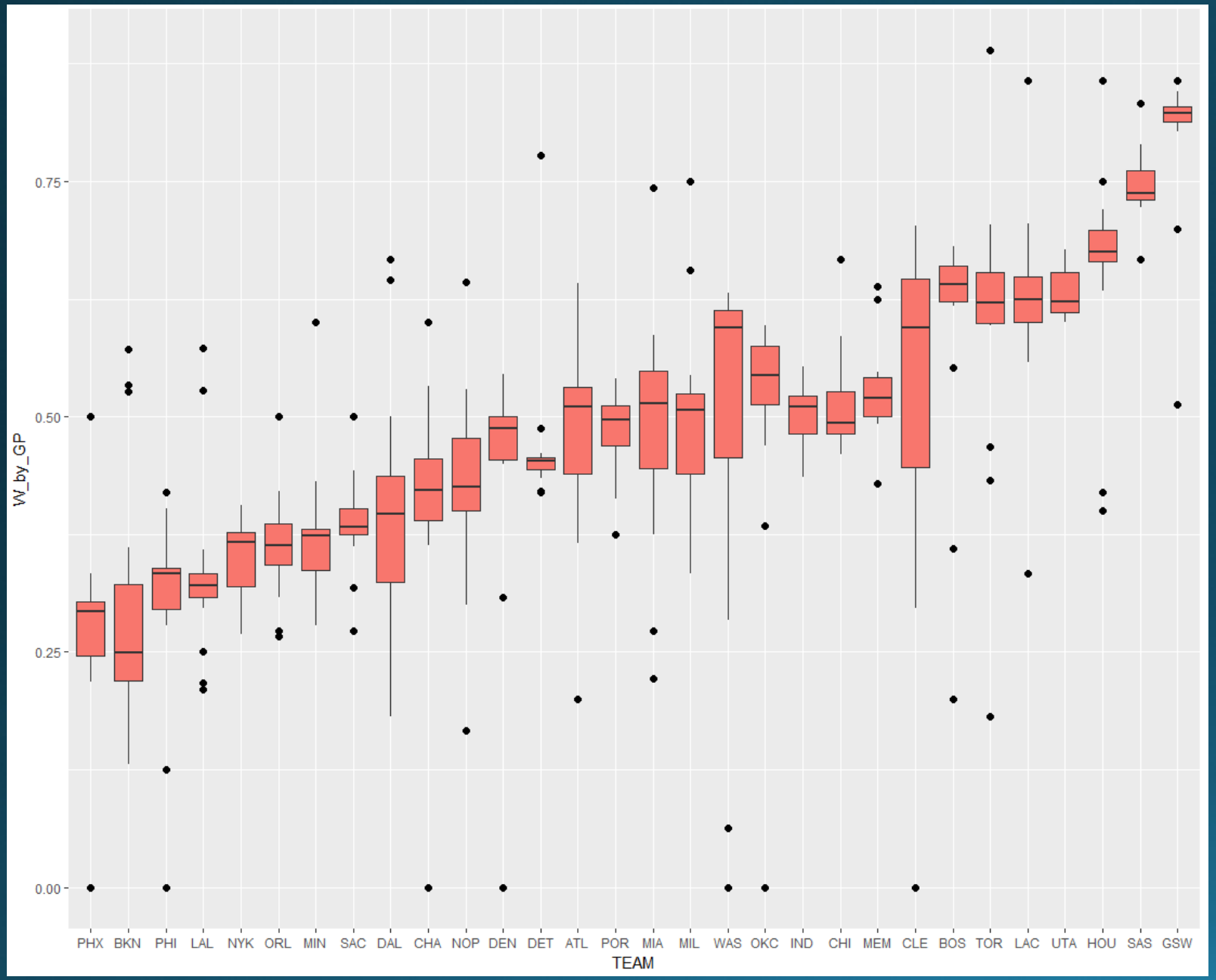


MEAN PER ACROSS TEAMS

Many outliers

Mean and median of team PER

Teams have comparable PER

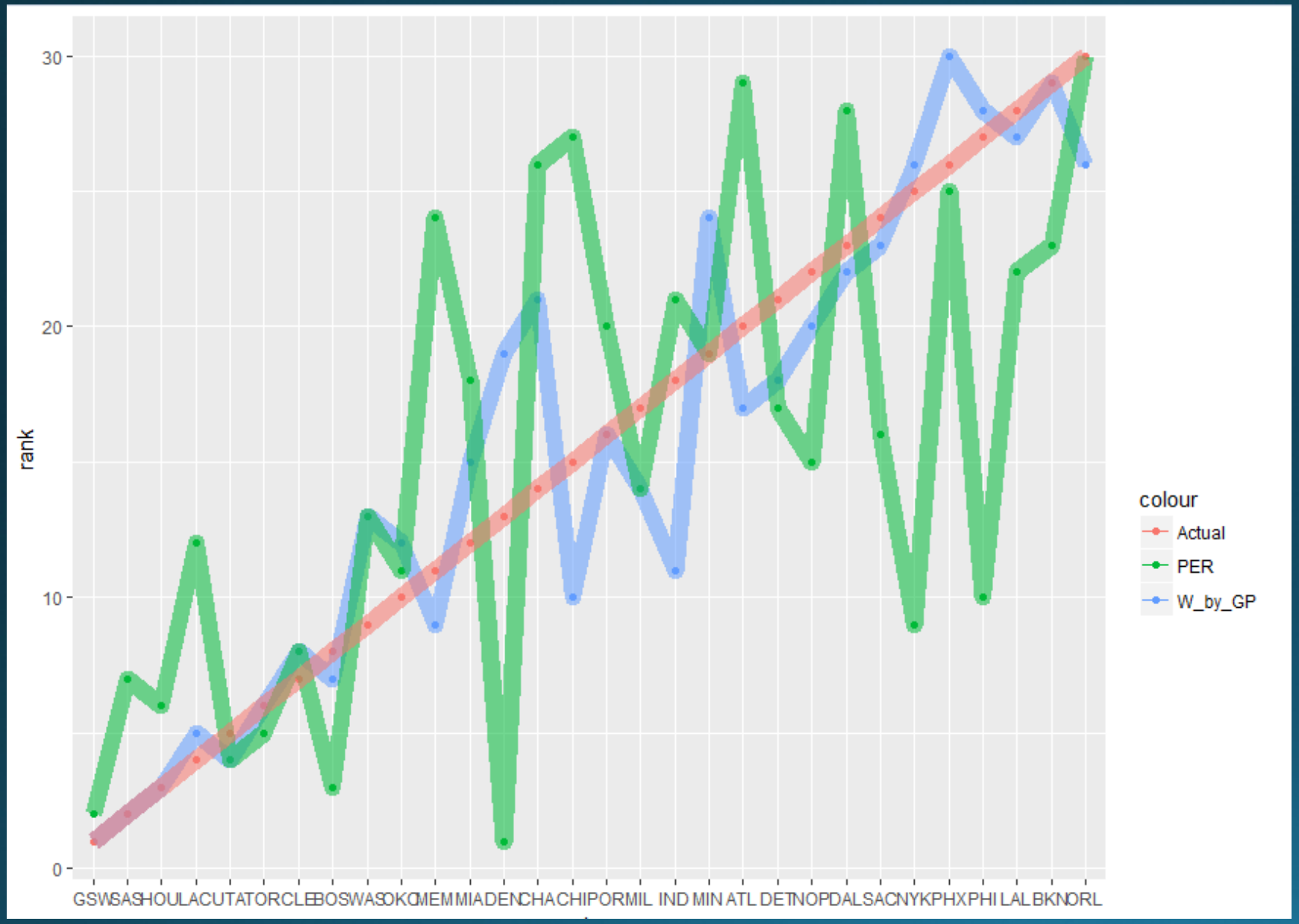


MEAN W_by_GP ACROSS TEAMS

Many outliers

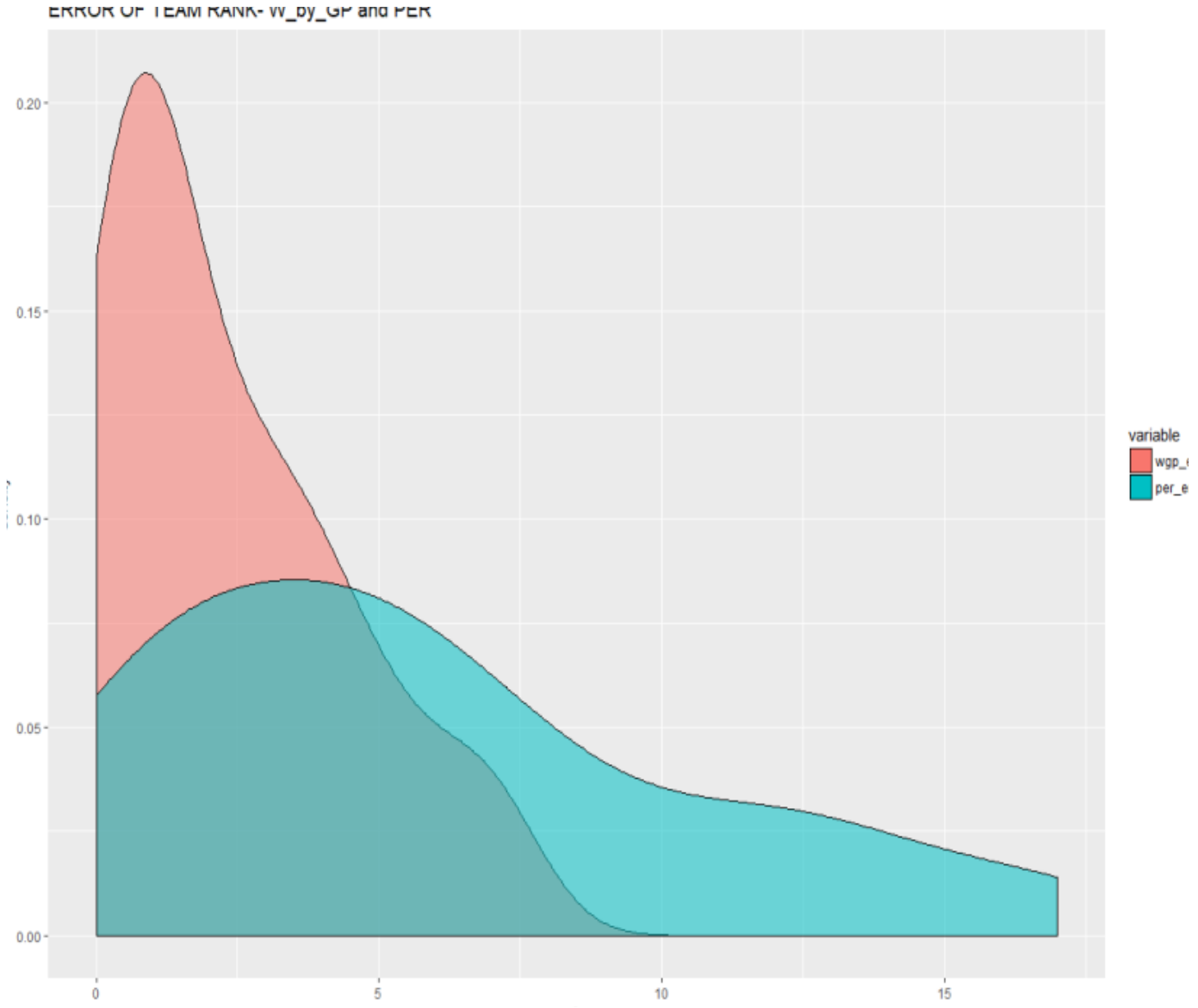
Mean not indicative of median W_by_GP

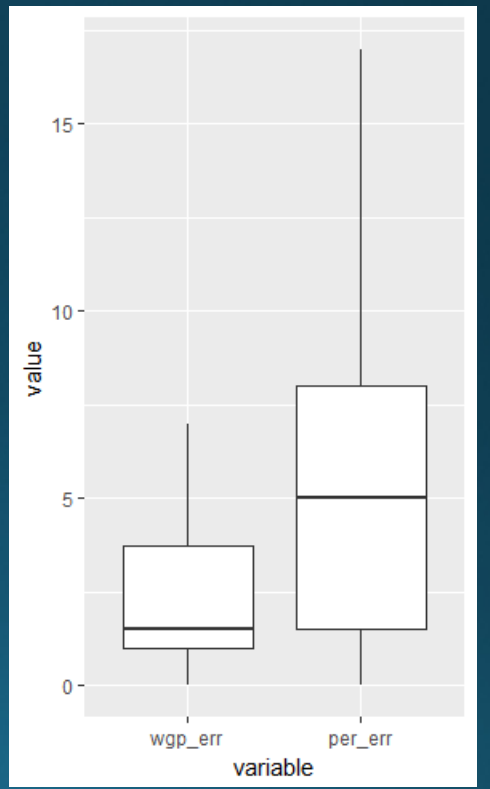
Teams have distinguishable ranges of W_by_GP



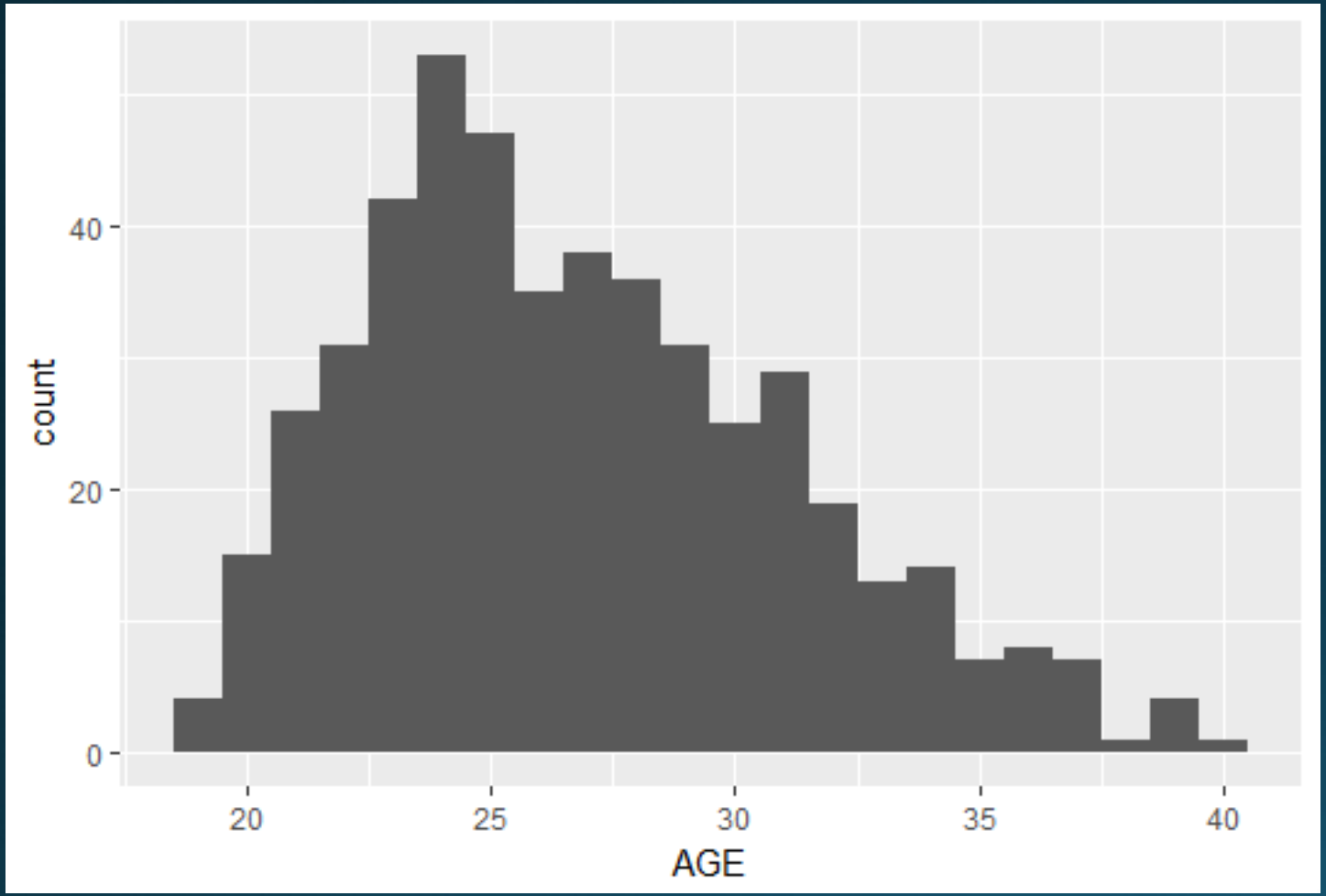
PLOTTING TEAM RANK USING PER AND W_by_GP

ERROR OF TEAM RANK PREDICTION



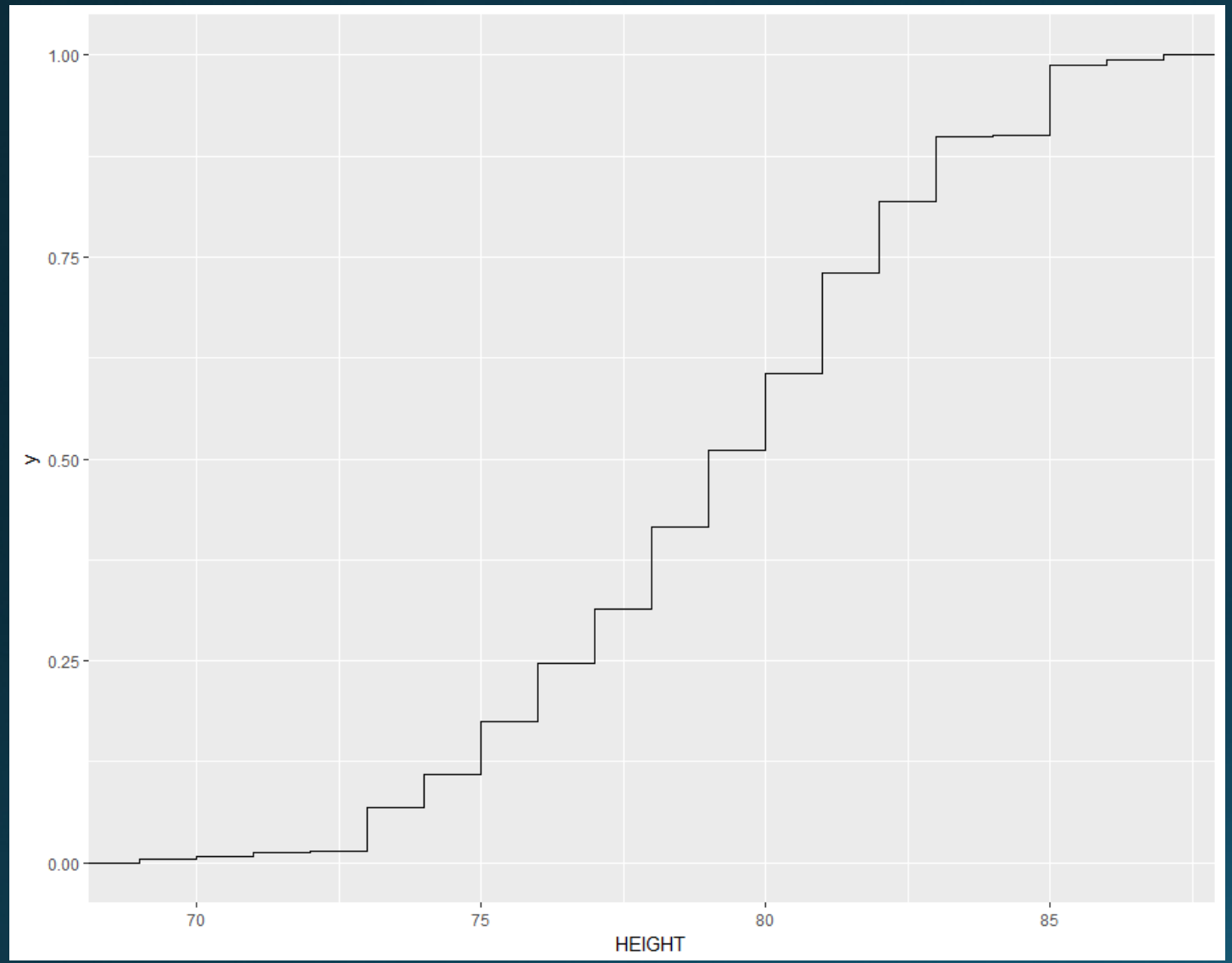


HOW DO PHYSICAL TRAITS INFLUENCE PLAYER PERFORMANCE?



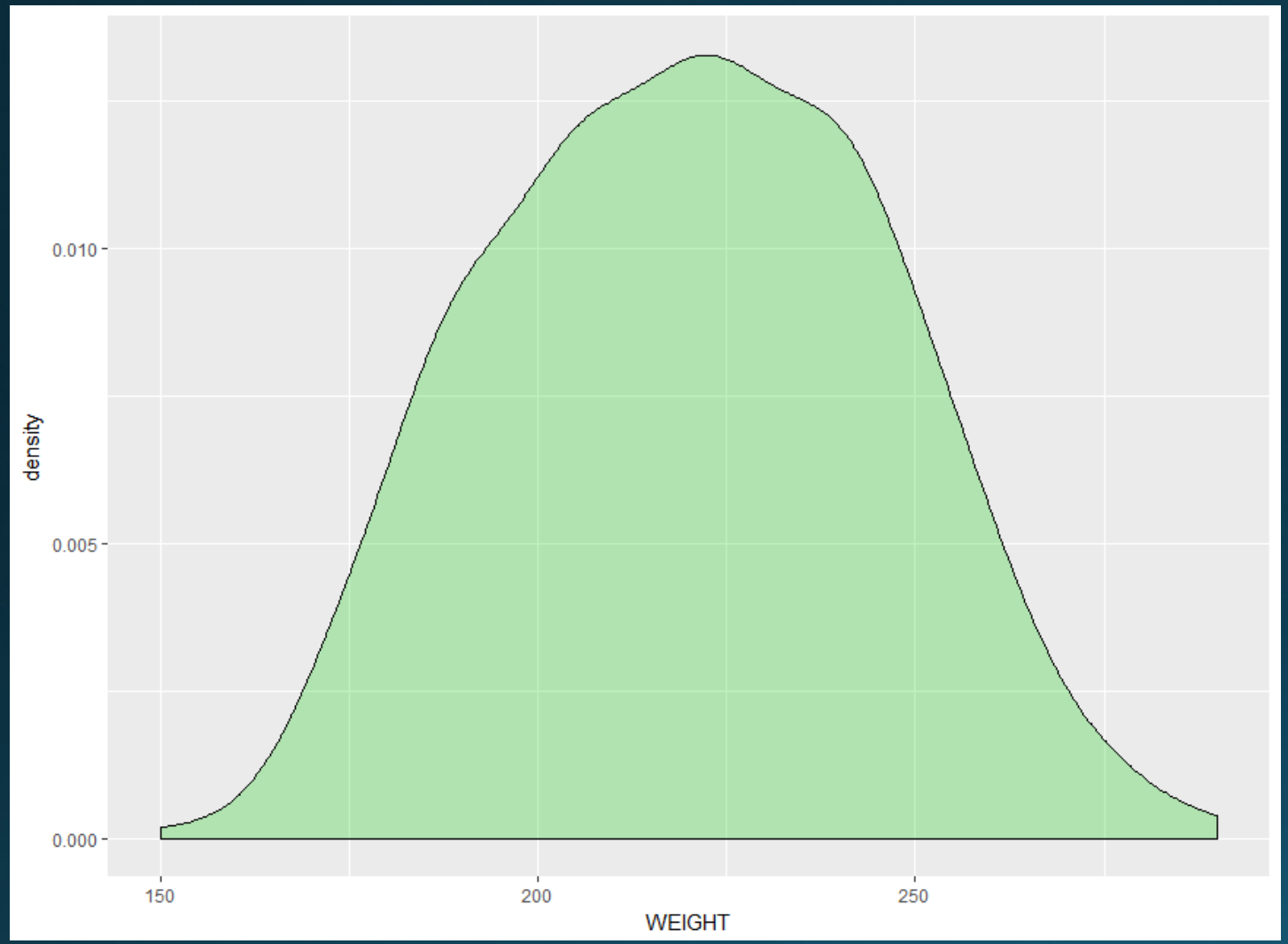
VARIABLES

- AGE
- HEIGHT
- WEIGHT



VARIABLES

- AGE
- HEIGHT
- WEIGHT

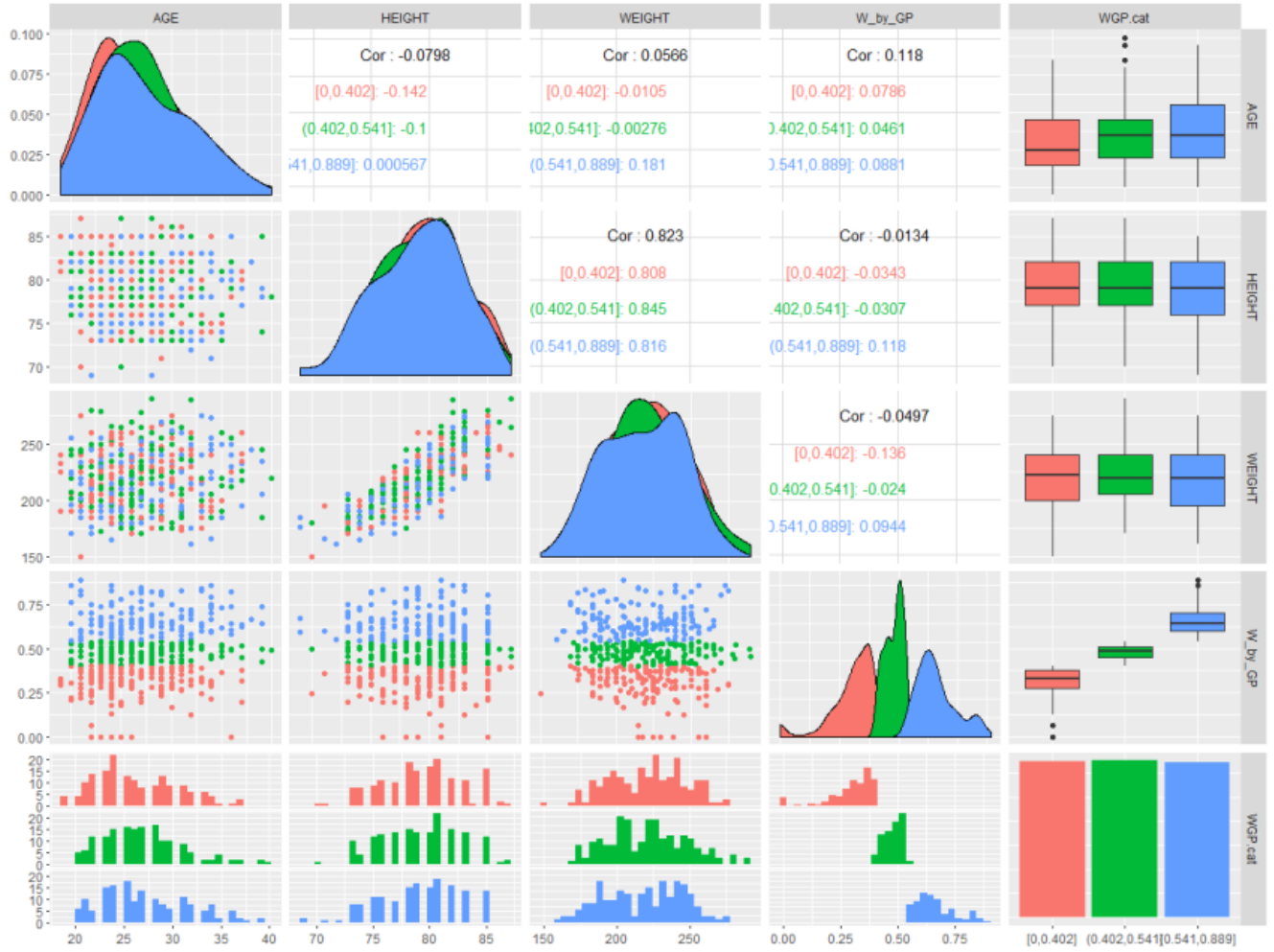


VARIABLES

- AGE
- HEIGHT
- WEIGHT

CORRELATION

- W_by_GP
- No influence



CAN WE PREDICT THE WIN PERCENTAGE OF A PLAYER?

Running the Models

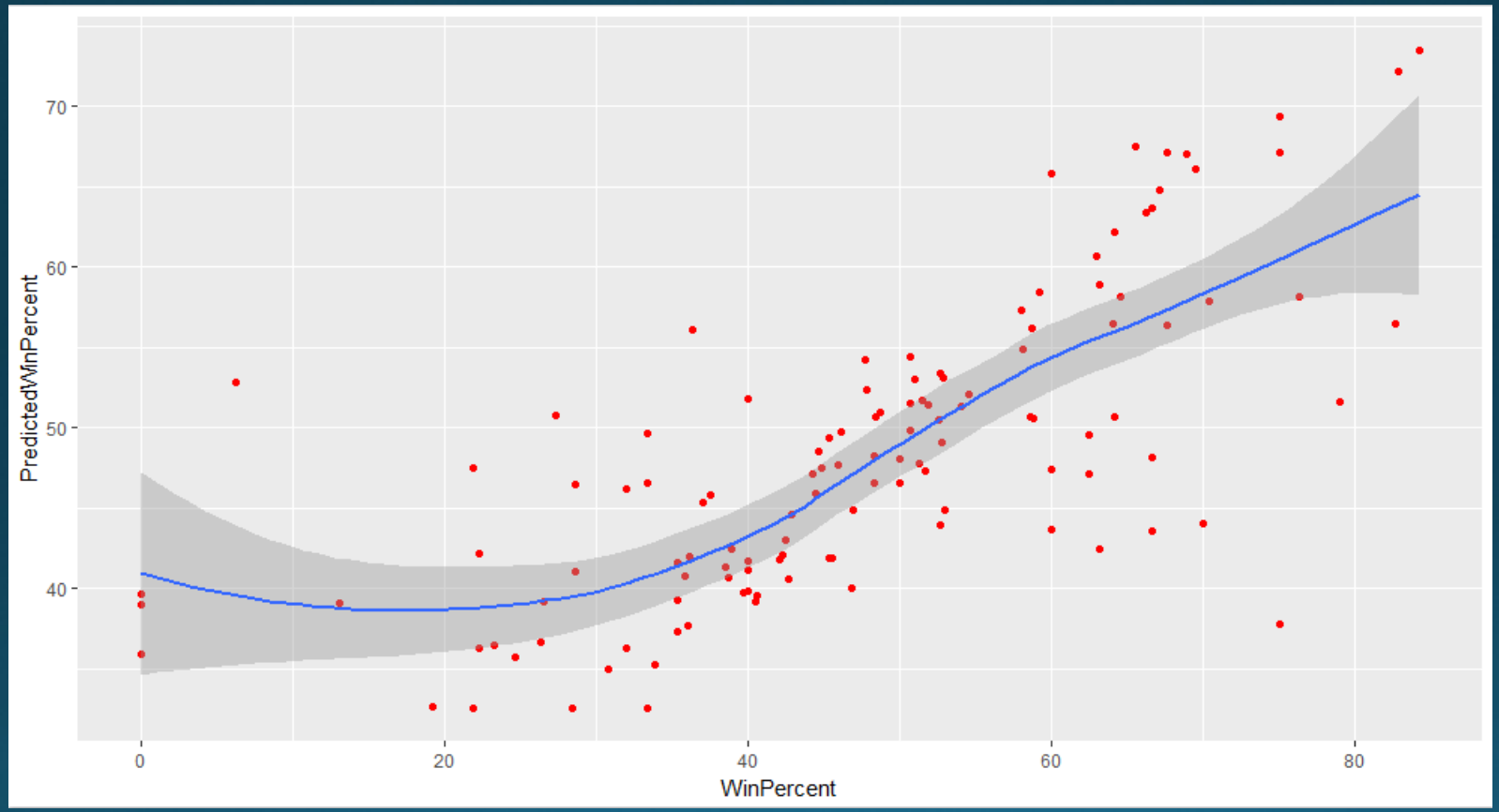
Due to high correlation, the following variables were removed:

1. FGA
2. FTA
3. FP
4. DREB
5. 3PA

Also, Names and Teams of the players were removed from the dataset.

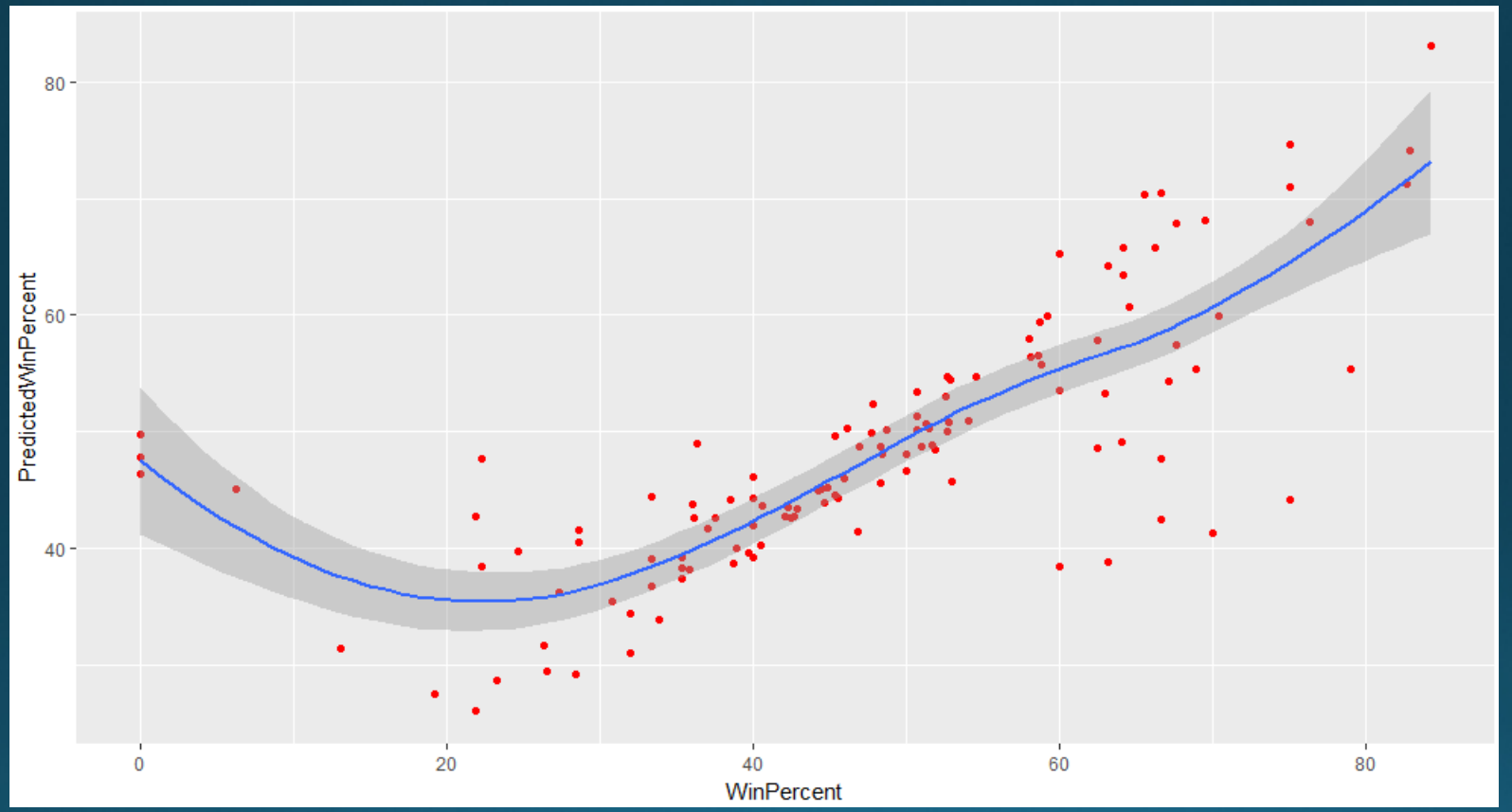
The dependent variable in the dataset is Percentage of Wins (No. of Wins/No. of Games Played) of a player

Random Forest



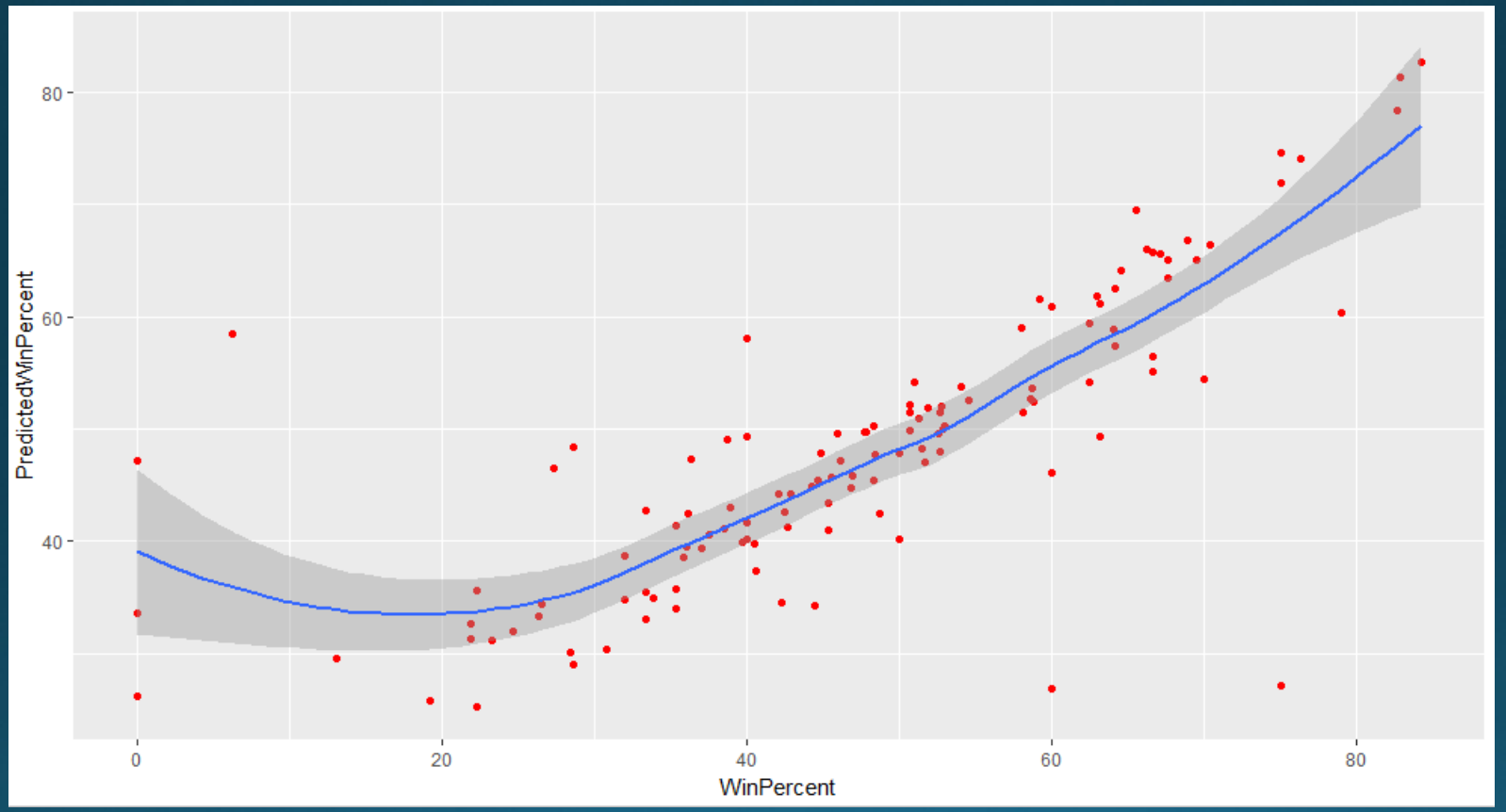
RMSE: 12.6

SVM Regression



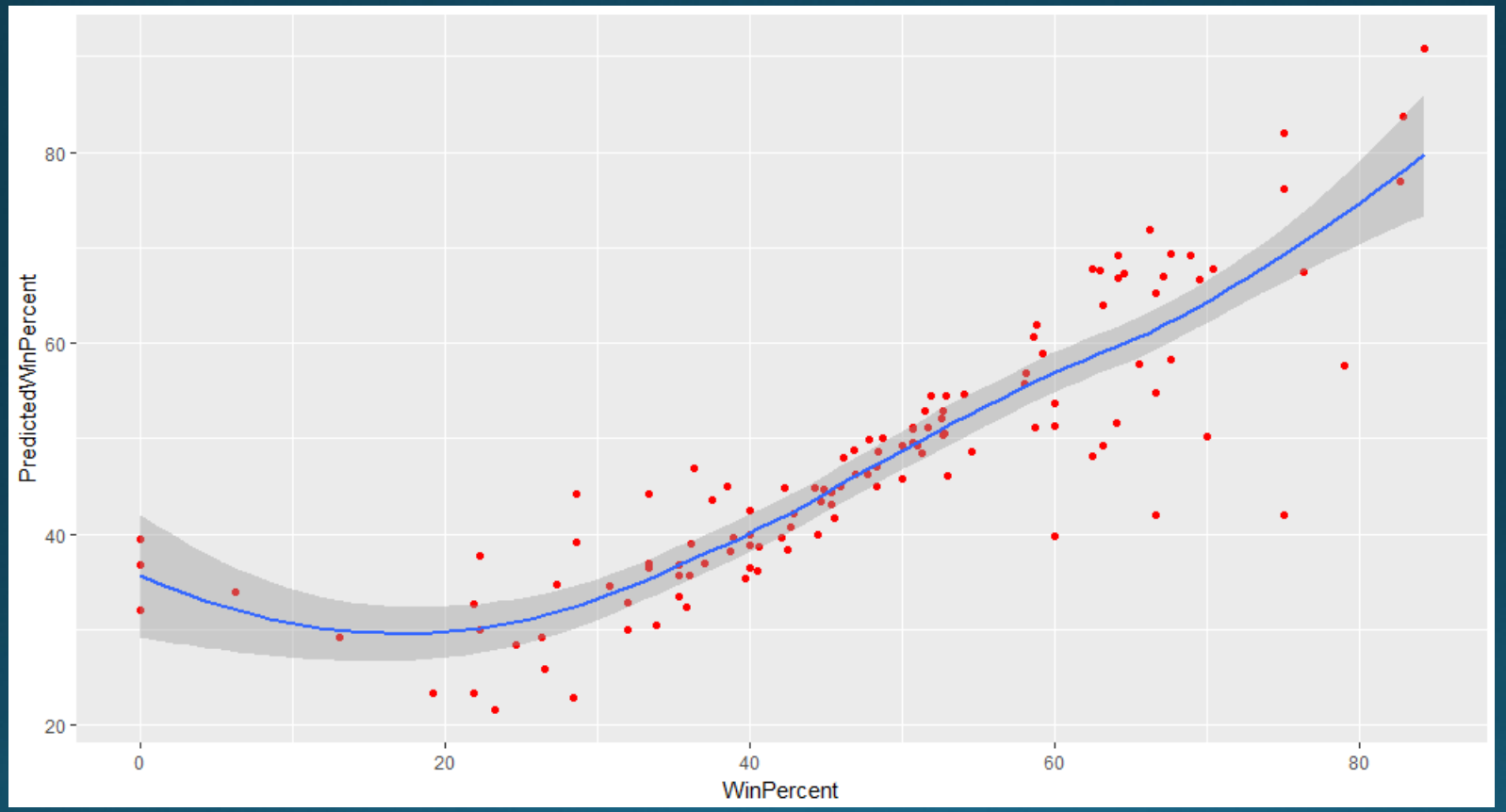
RMSE: 12.02

XGBoost



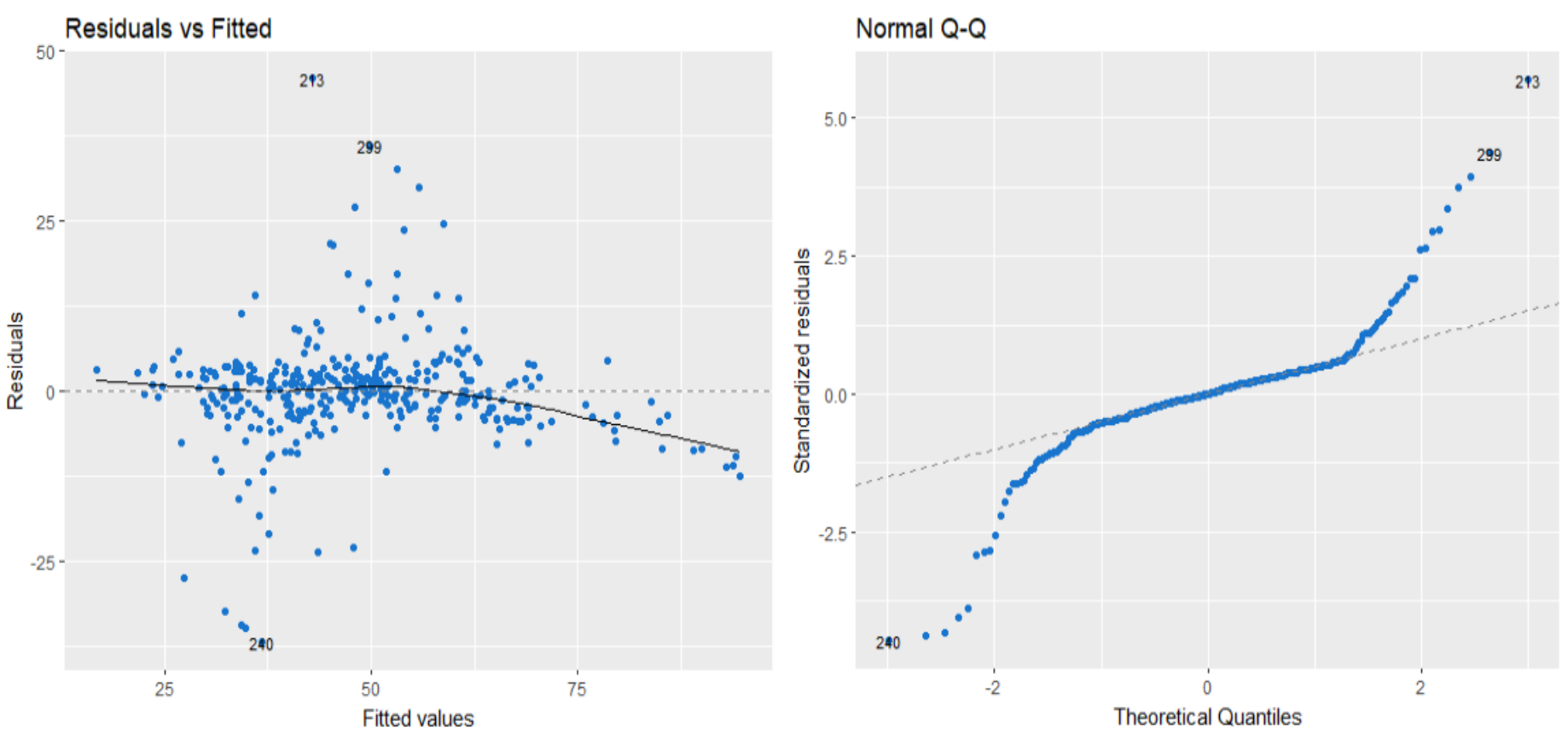
RMSE: 11.04

Linear Model



RMSE: 9.45

Residuals vs Fitted and QQ Plot



R-Squared: 0.74

Prediction Values of Test Data

	WinPercent	PredictedWinPercent
1	40.00000	42.51925
2	67.64706	69.28313
3	50.00000	45.70681
4	66.66667	54.73526
5	84.21053	90.87151
6	52.77778	50.50057
7	13.04348	29.14509
8	27.27273	34.79007
9	51.02041	49.28662
10	40.00000	36.48709
11	45.33333	44.33261
12	58.10811	56.81520
13	44.26230	44.89159
14	65.57377	57.74780
15	28.39506	22.90252
16	60.00000	51.29090
17	26.31579	29.26231
18	52.94118	46.11930

FUTURE WORK

- Additional seasons
- College basketball
- “second generation statistics” from motion capture technology

QUESTIONS?