p8106 - Final Project - NBA Players Salary Prediction

Mingkuan Xu, Mengfan Luo, Yiqun Jin

5/6/2022

Introduction

Describe your data set. Provide proper motivation for your work.

What questions are you trying to answer? How did you prepare and clean the data?

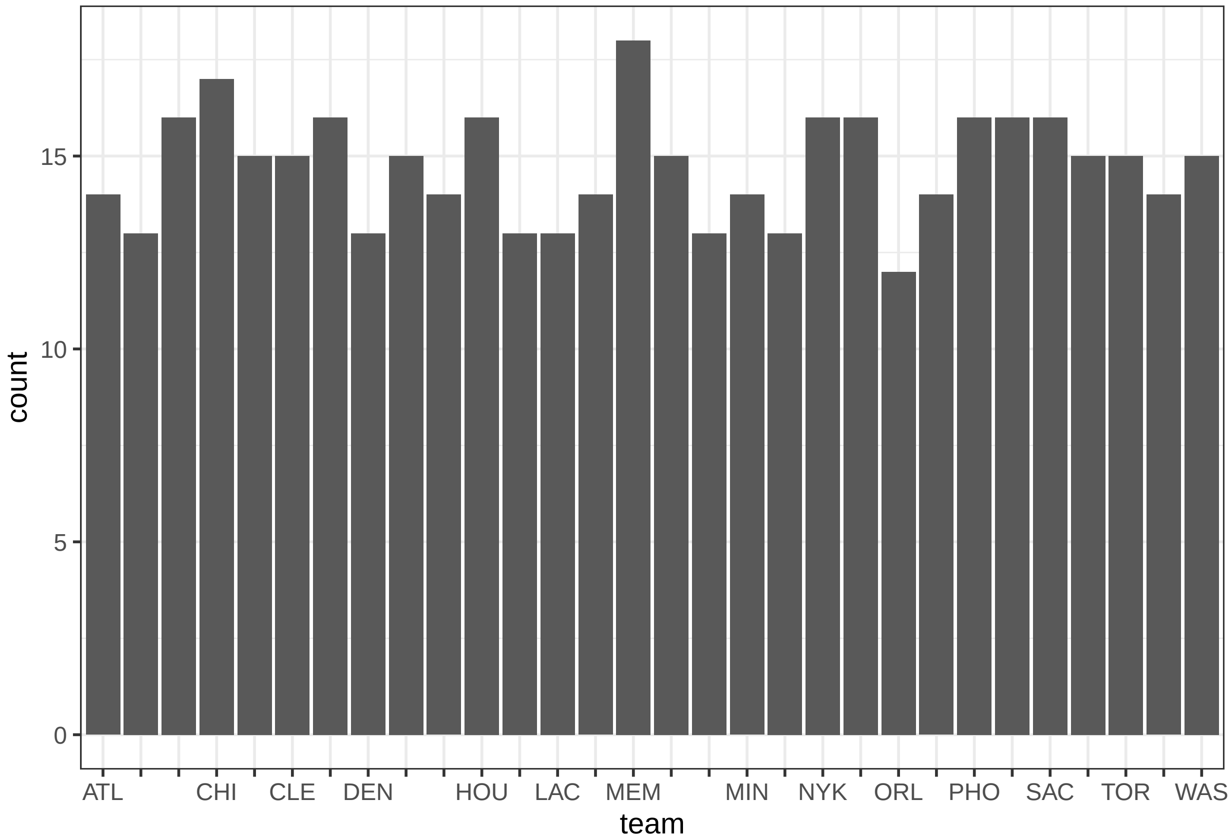
Data Preprocessing

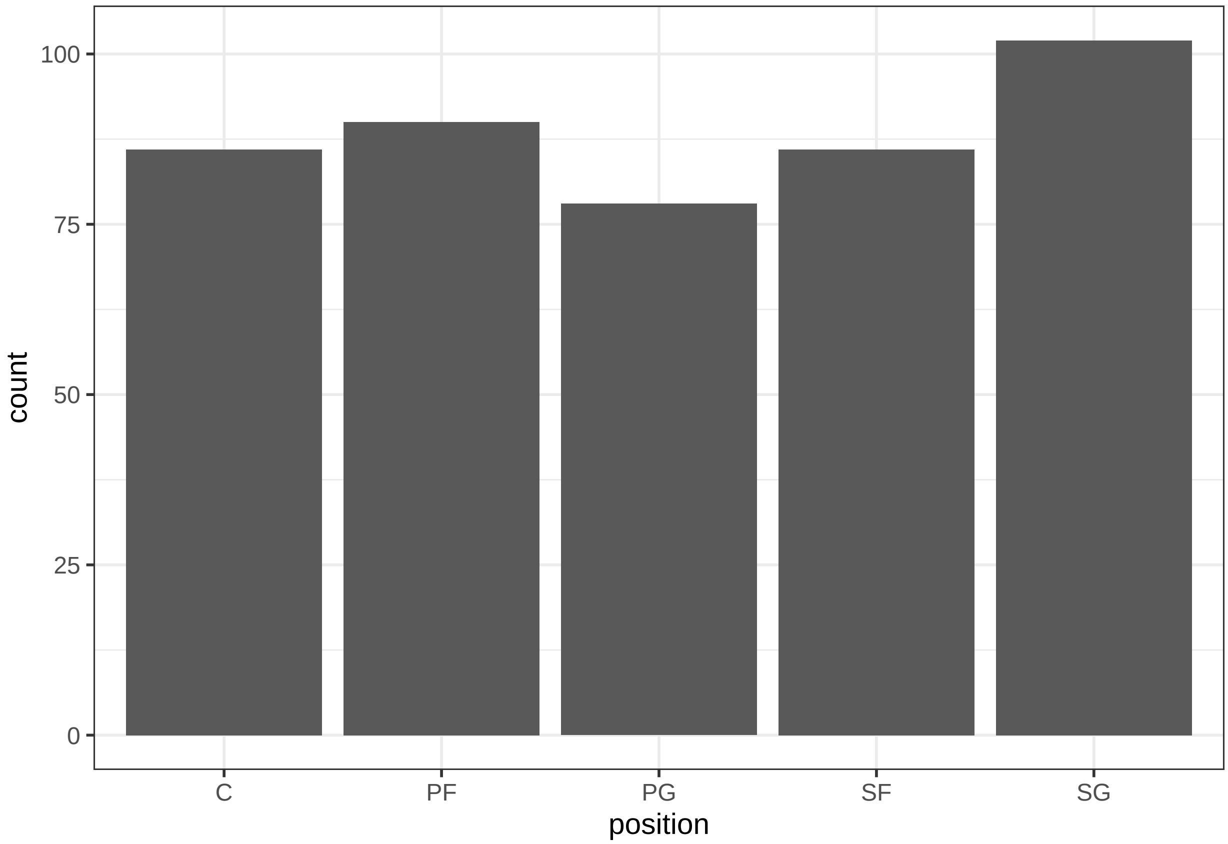
Exploratory analysis/visualization

Since `minute` stands for minutes played per game, we will divided variables stands for counts by `minute` to get a rate. These variables includes `field_goal`, `fg_attempt x3p`, `x3p_attempt`, `x2p`, `x2p_attempt`, `free_throw`, `ft_attempt`, `offensive_rb defenssive_rb`, `total_rb`, `assistance`,`steal`, `block`, `turnover`, `personal_foul` and `point`.

Univariate Analysis

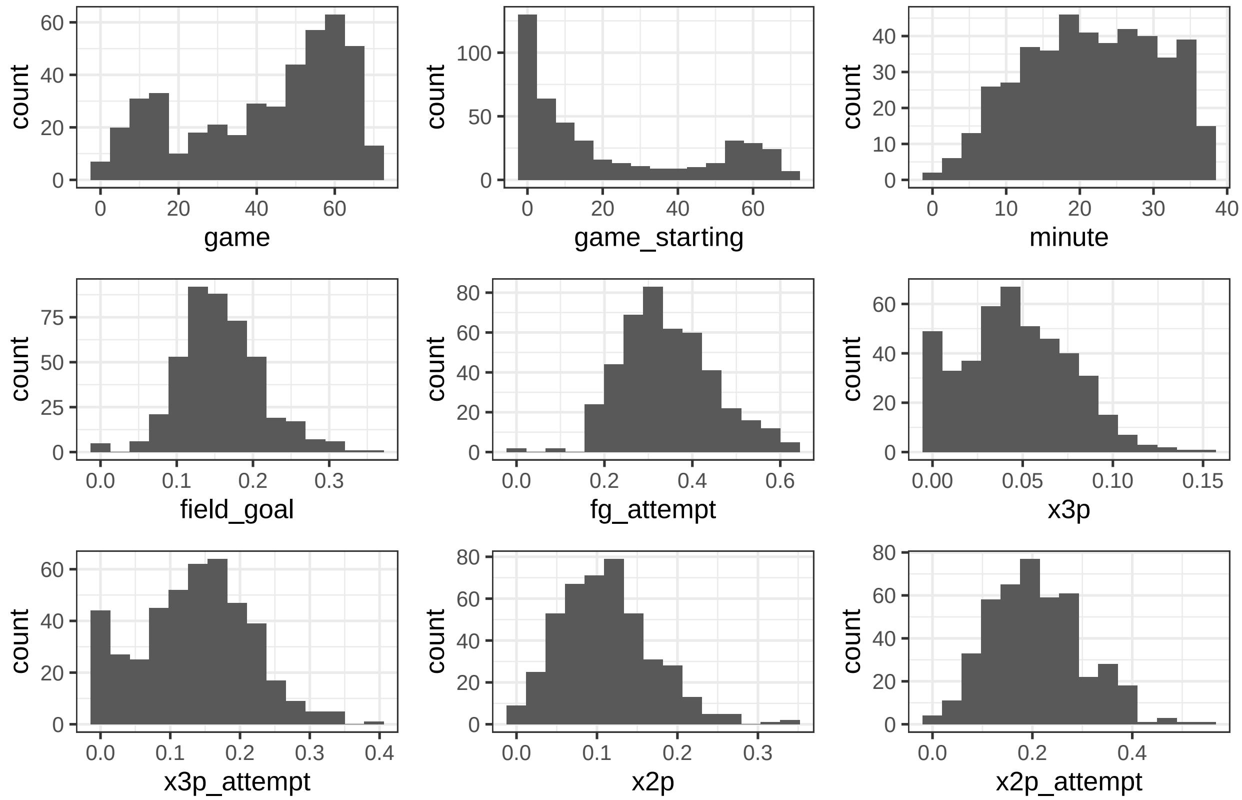
Distributions of the two categorical variables, `team` and `position`.

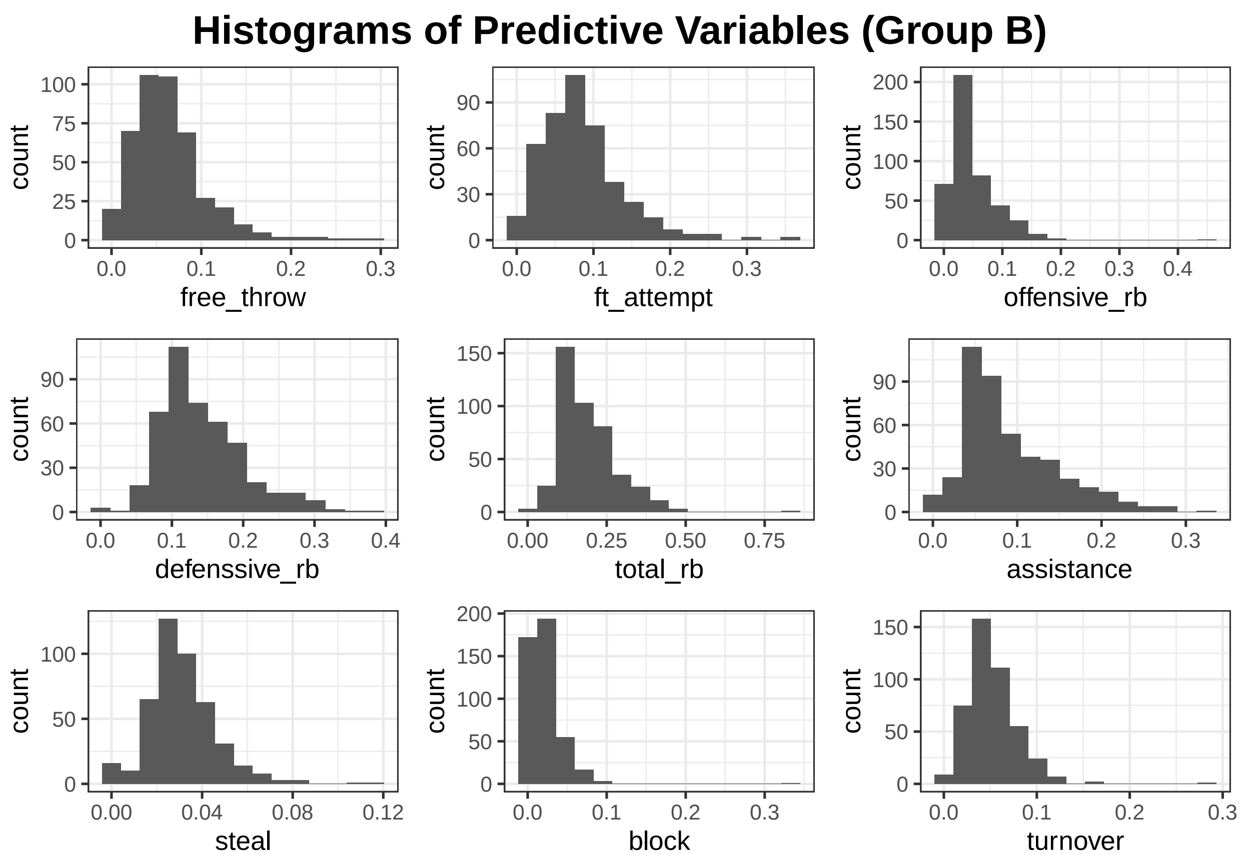

count
15
10
5
0
ATL
CHI
CLE
DEN
HOU
LAC
MEM
MIN
NYK
ORL
PHO
SAC
TOR
WAS
team

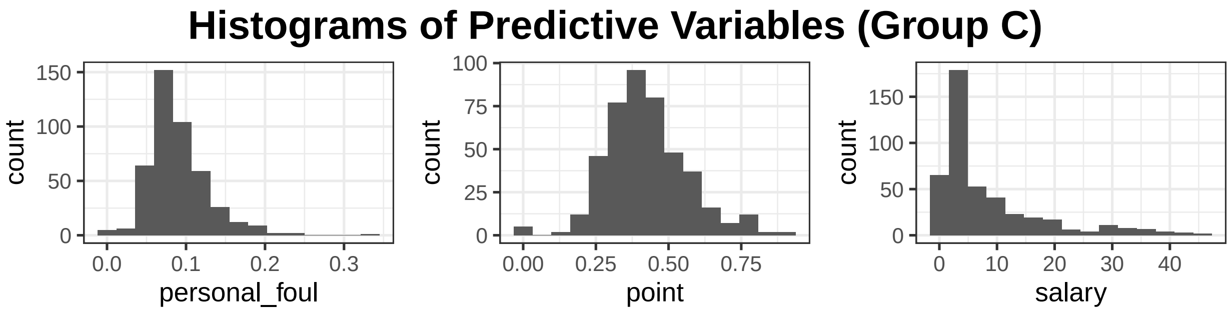


Distributions of other numeric variables.

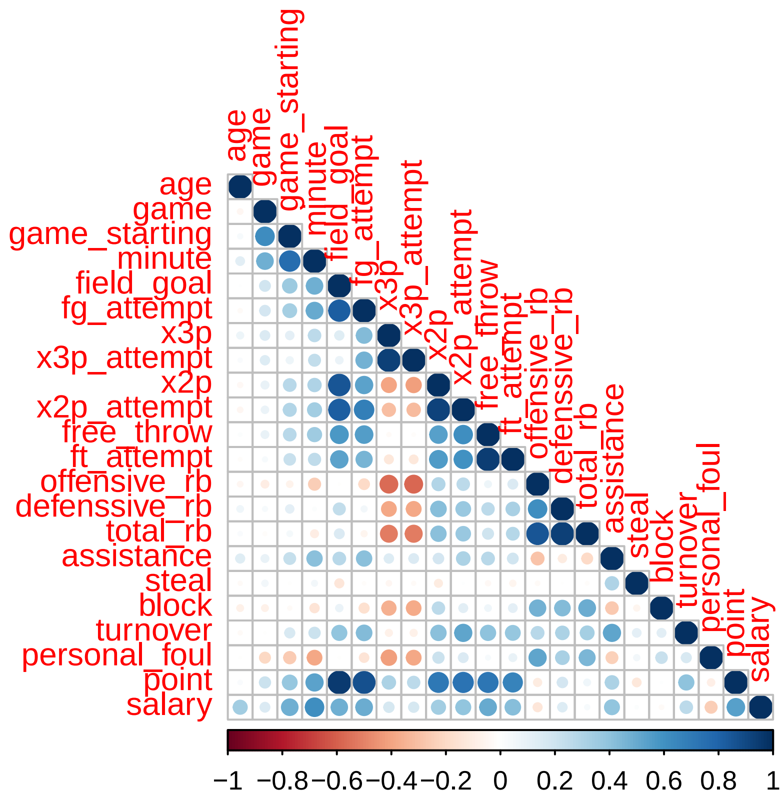
Histograms of Predictive Variables (Group A)




Histograms of Predictive Variables (Group B)
count
free_throw
ft_attempt
offensive_rb
defenssive_rb
total_rb
assistance
steal
block
turnover

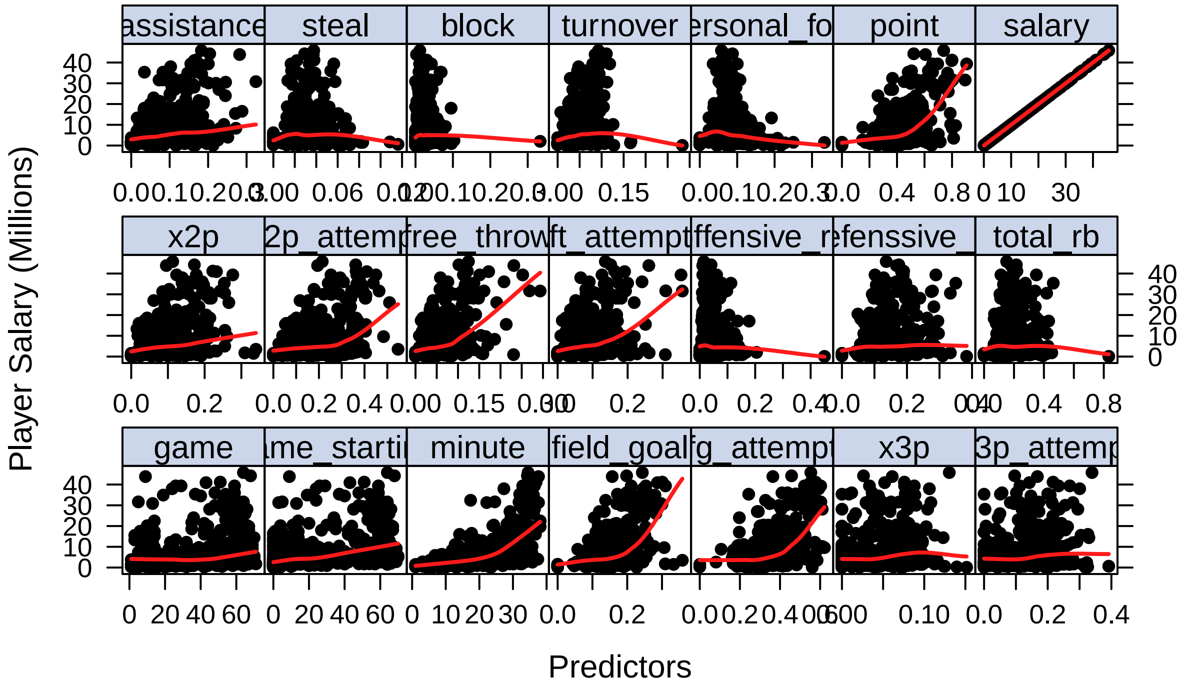

Histograms of Predictive Variables (Group C)
count
150
100
50
0
0.0
0.1
0.2
0.3
personal_foul
count
100
75
50
25
0
0.00
0.25
0.50
0.75
point
count
150
100
50
0
0
10
20
30
40
salary

Correlation Analysis

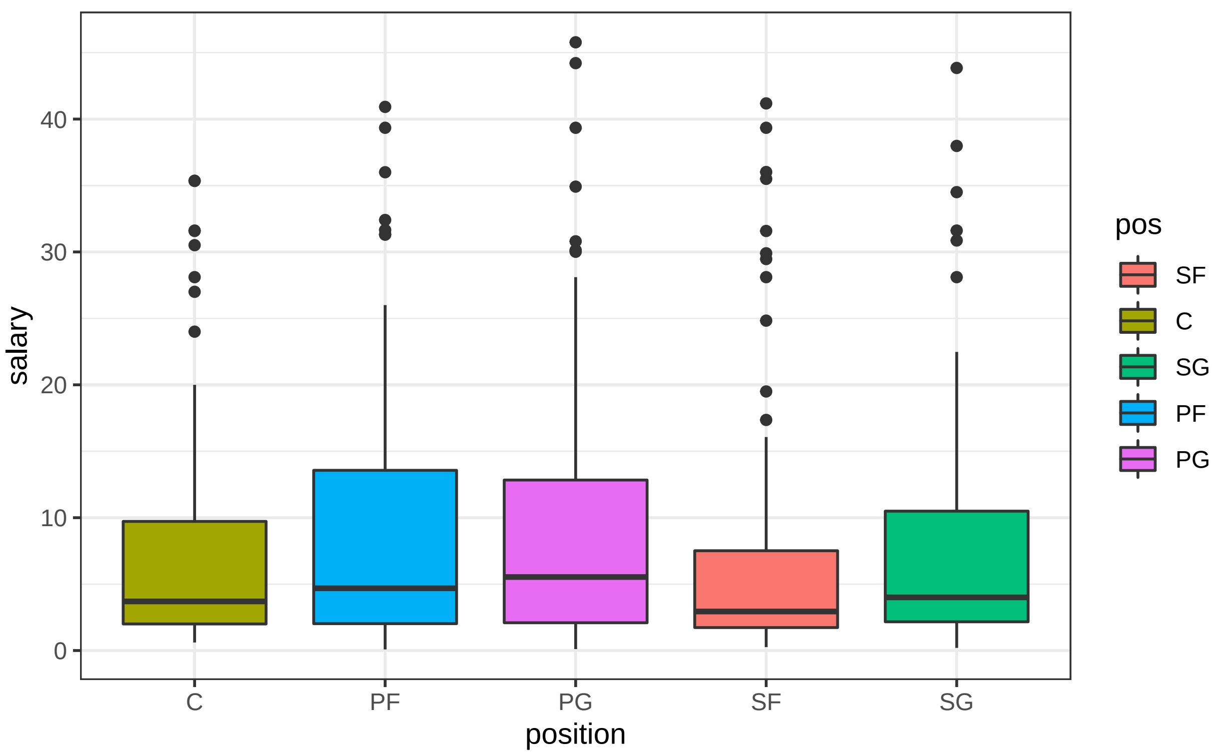


Analyzing trends in data

From numeric variables, we found that `stl`,`x3p`, `age`,`gs` seem to have some non-linear trends.



From categorical variable `position`, extremely high values in salary show in all positions and some teams.

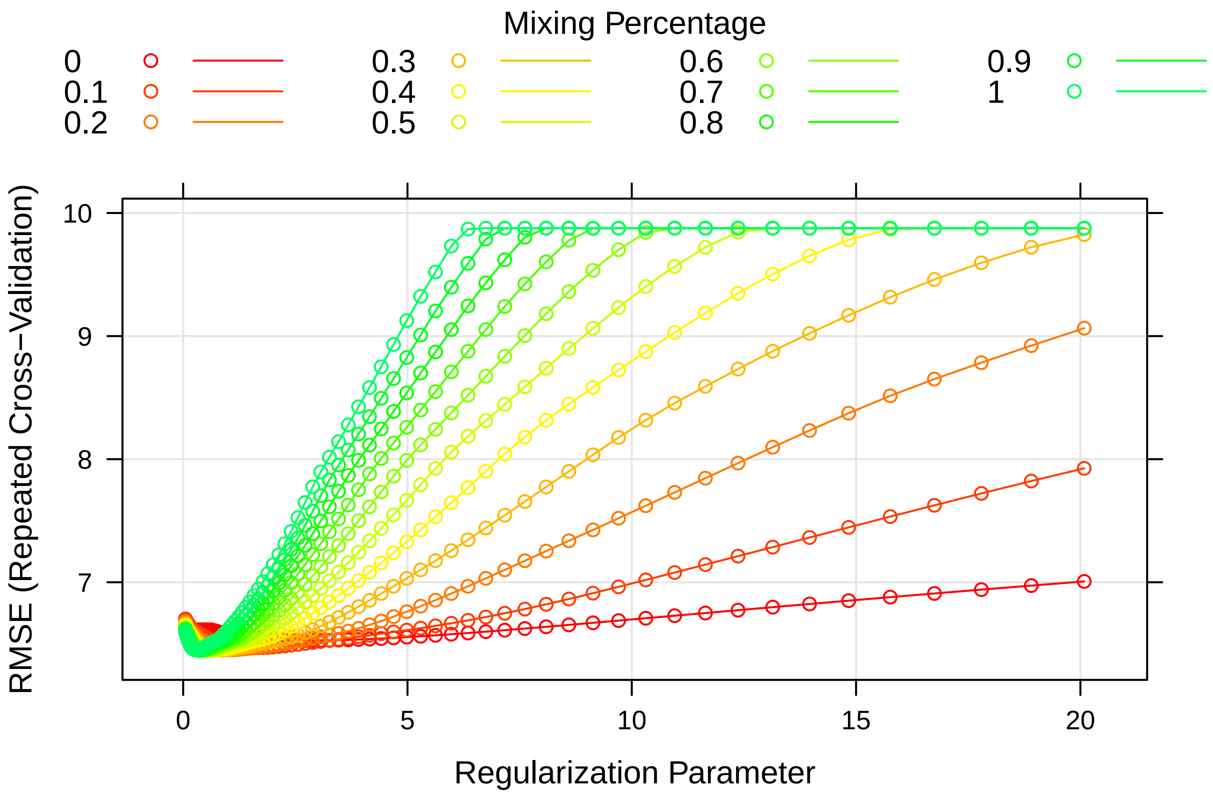


Models

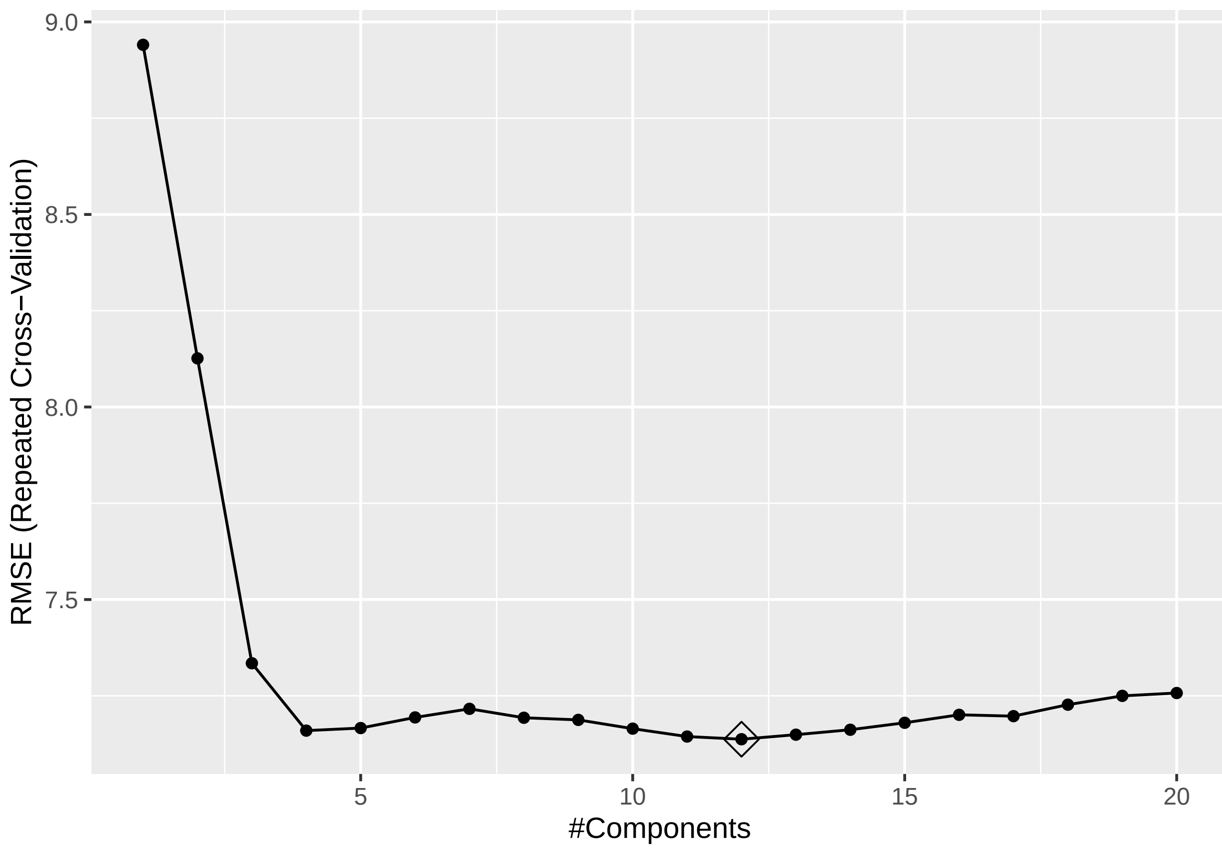
Part 1 Linear regression

(a) Standard Least-Squared

(b) Elastic Net (including lasso/ridge)



###(c) Principle Component Regression



Part 2 Generalized Linear Regression

(a) GAM

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## salary ~ s(age) + s(game) + s(game_starting) + s(free_throw) +
##     s(ft_attempt) + s(defenssive_rb) + s(assistance) + s(block) +
##     s(personal_foul) + s(point)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.5293     0.2958   28.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                    edf Ref.df      F  p-value
## s(age)           4.414  5.455 16.961  < 2e-16 ***
## s(game)          1.695  2.101  4.623  0.00973 **
## s(game_starting) 1.482  1.805 25.494  < 2e-16 ***
## s(free_throw)    8.147  8.791  3.083  0.00538 **
```

```
## s(ft_attempt)    1.000  1.000  0.155  0.69382
## s(defenssive_rb) 1.000  1.000  1.680  0.19591
## s(assistance)    1.000  1.000 18.244 2.58e-05 ***
## s(block)         1.000  1.000  2.758  0.09777 .
## s(personal_foul) 6.851  7.891  5.172 6.56e-06 ***
## s(point)         6.152  7.361  5.415 5.90e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.69   Deviance explained = 71.8%
## GCV = 34.237  Scale est. = 30.974    n = 354
```

(b) MARS

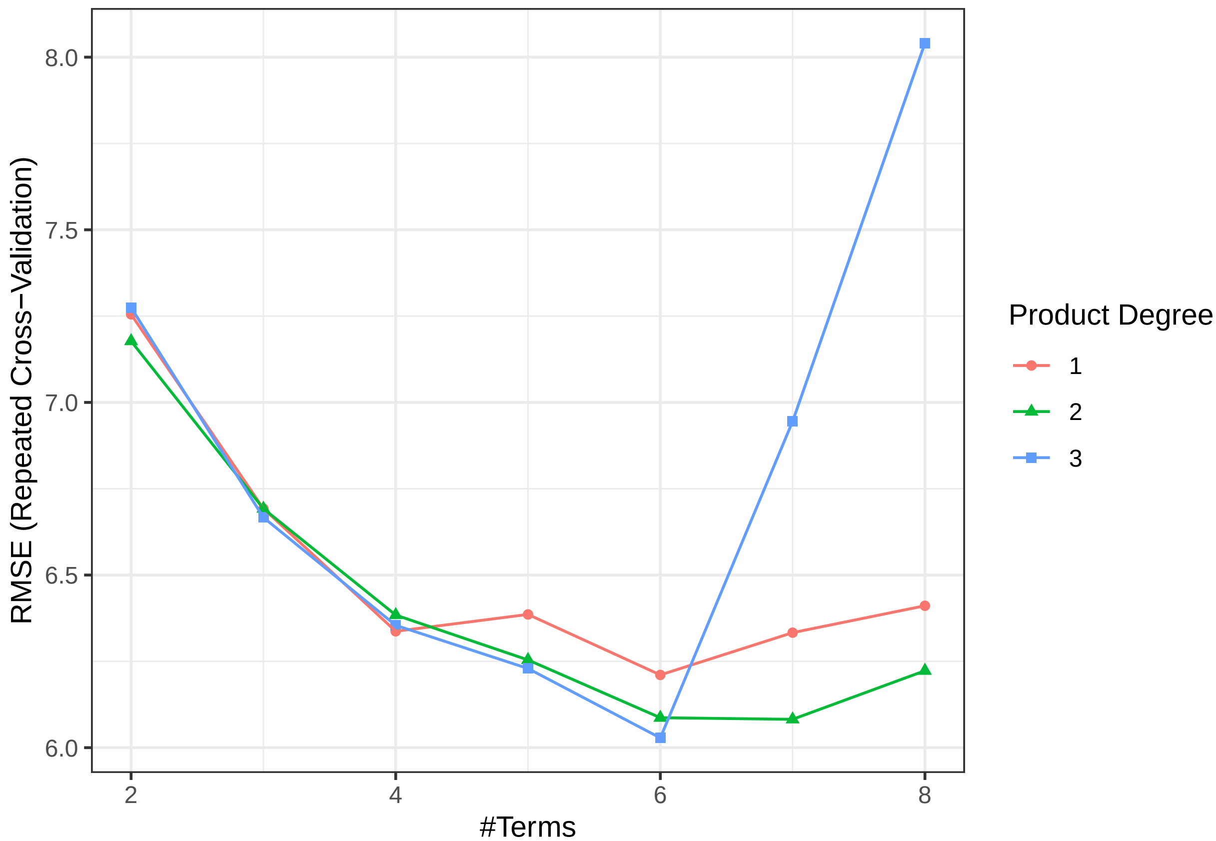


Table 1: Table 1: RMSE of Different Models

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
LeastSquare	4.41	6.12	6.85	6.79	7.46	8.75	0
ElasticNet	4.57	5.95	6.37	6.45	7.06	8.55	0
PCR	5.17	6.24	7.17	7.14	7.87	9.34	0
MARS	4.05	5.25	5.89	6.03	6.71	8.74	0

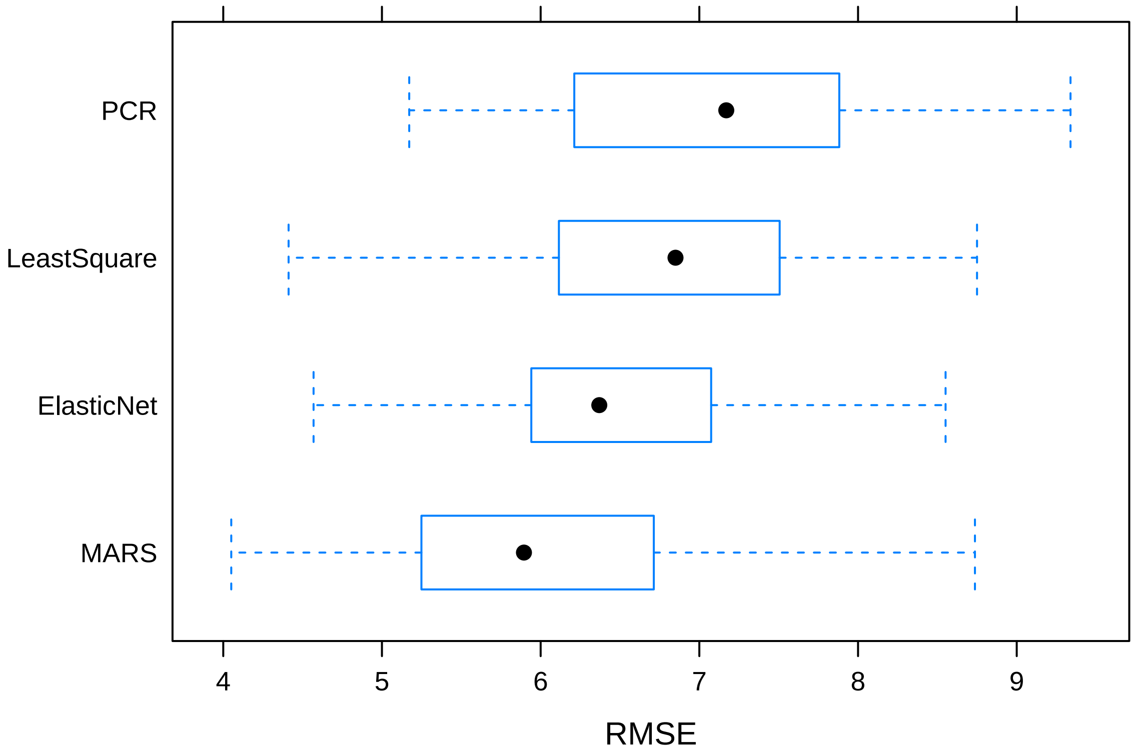


Table 2: Table 2: RMSE of Different Models on Test Set

	Linear	ElasticNet	PCR	GAM	MARS
RMSE	6.66	6.04	5.46	6.84	5.16