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

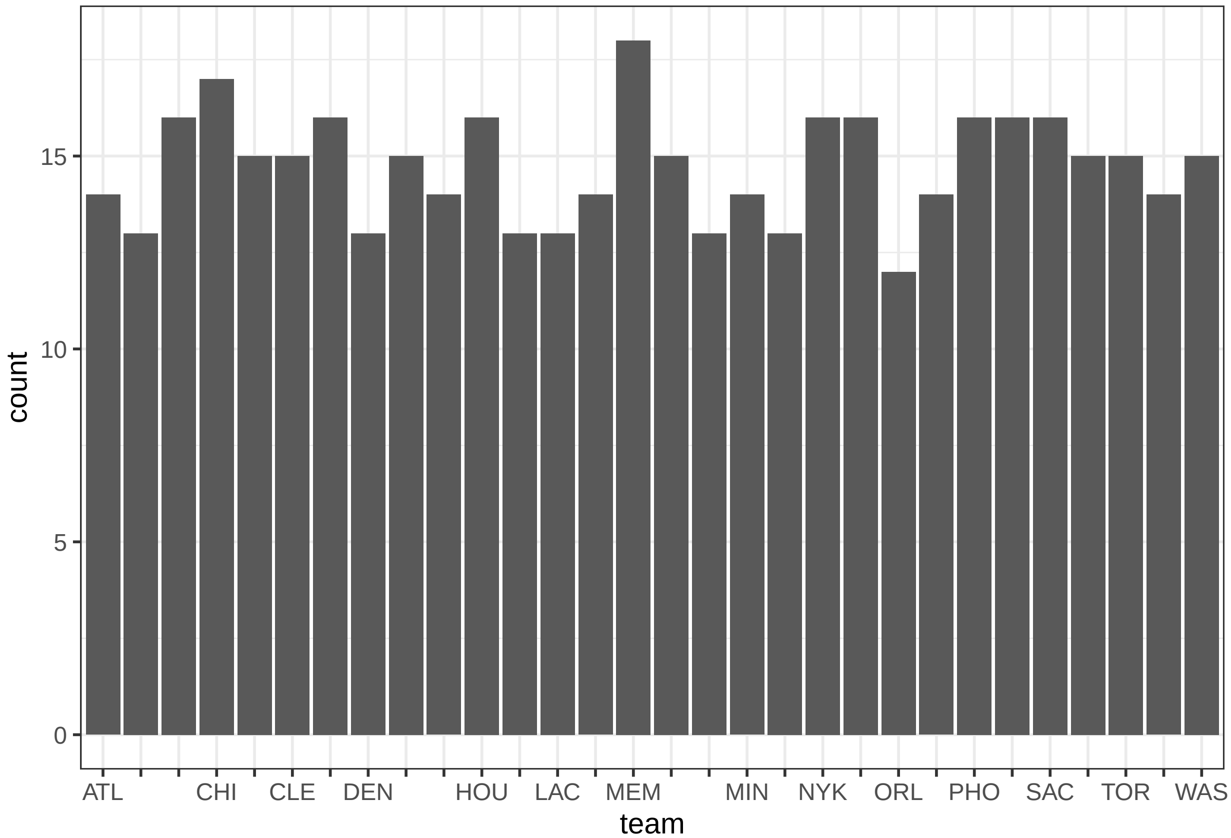
Part 0 - Data Preprocessing

Part 1 - Exploratory Analysis

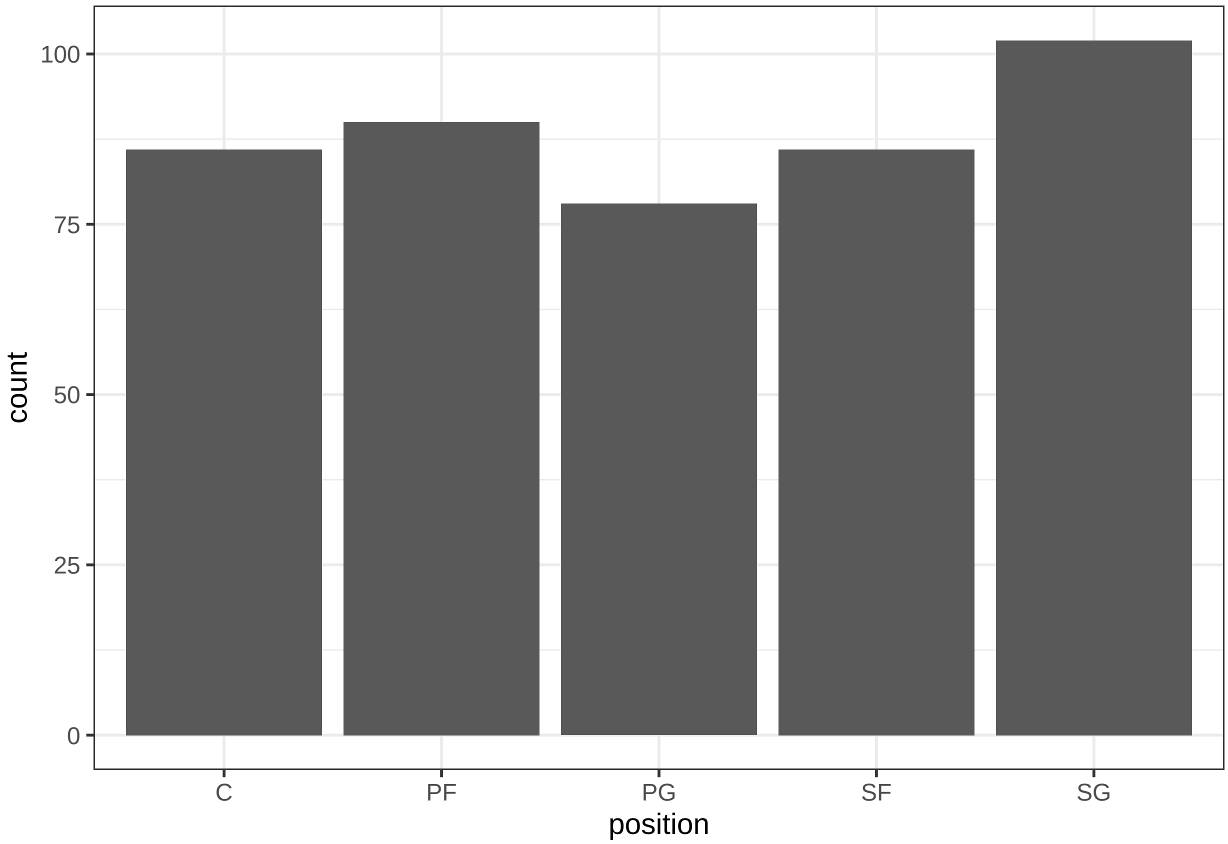
Since `minute` stands for minutes played per game, we will divided variables stands for counts by `minute` to get a rate. These variables includes `field_goal`, `fg_attempt` `x3p`, `x3p_attempt`, `x2p`, `x2p_attempt`, `free_throw`, `ft_attempt`, `offensive_rb` `defenssive_rb`, `total_rb`, `assistance`,`steal`,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

Univariate Analysis

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

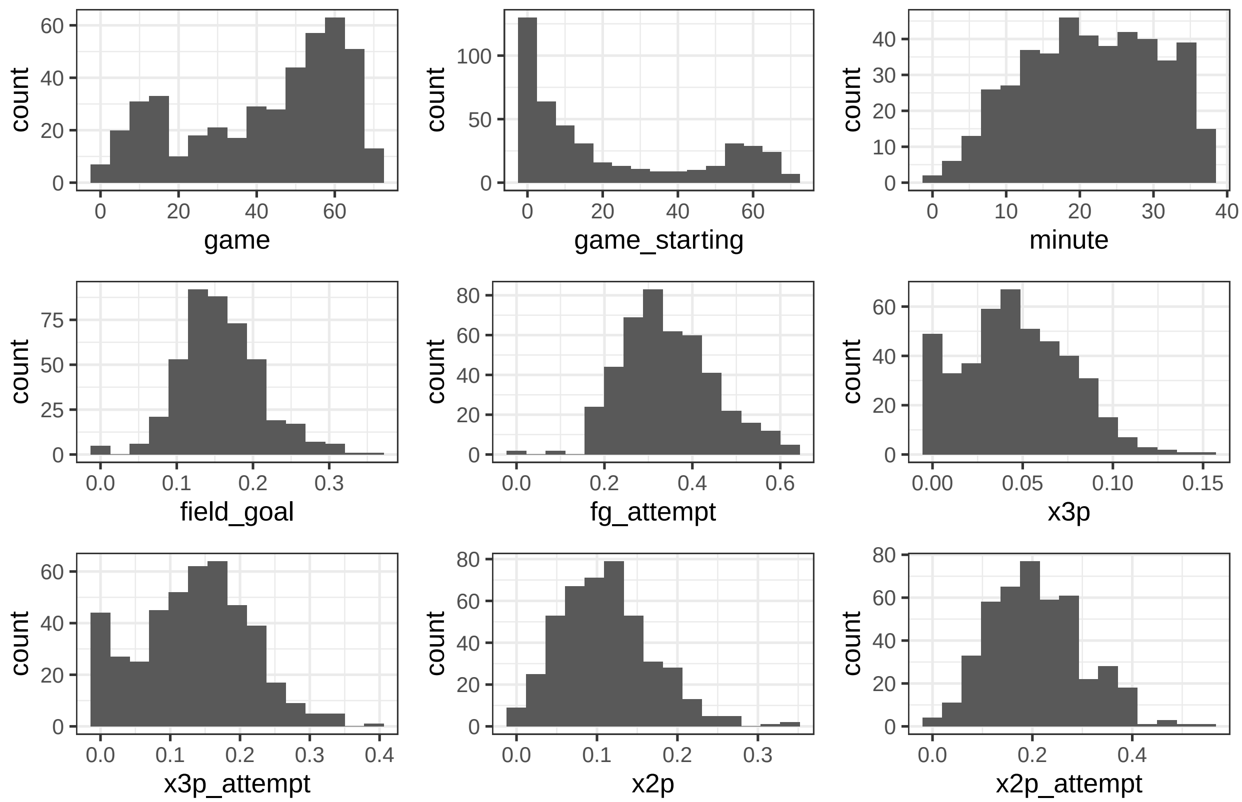


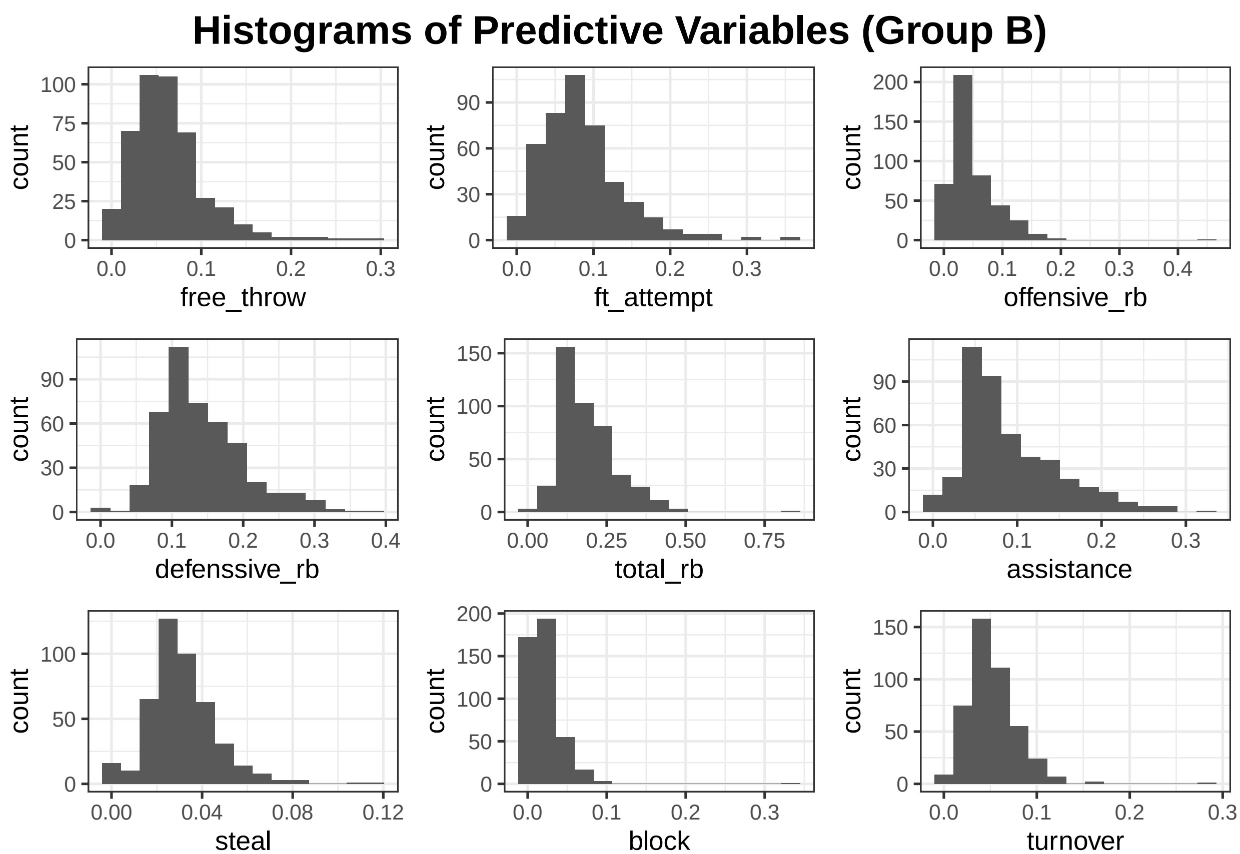
count
0
5
10
15
ATL
CHI
CLE
DEN
HOU
LAC
MEM
MIN
NYK
ORL
PHO
SAC
TOR
WAS
team



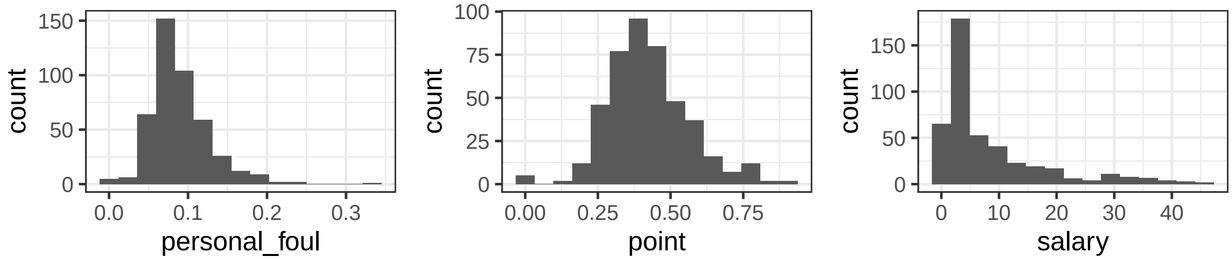
Distributions of other numeric variables.

Histograms of Predictive Variables (Group A)

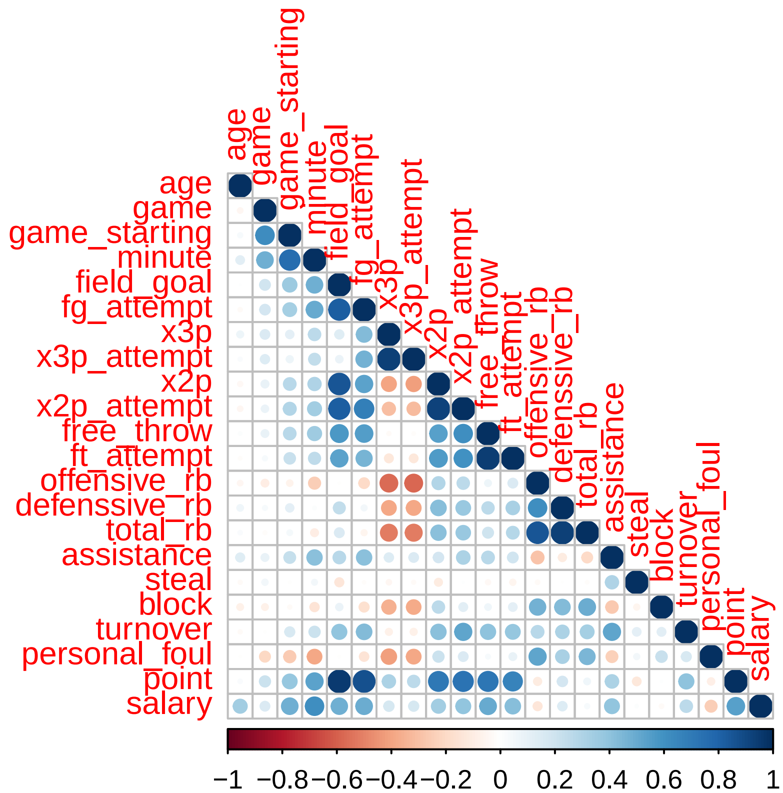



Histograms of Predictive Variables (Group B)
count
free_throw
ft_attempt
offensive_rb
defenssive_rb
total_rb
assistance
steal
block
turnover

Histograms of Predictive Variables (Group C)

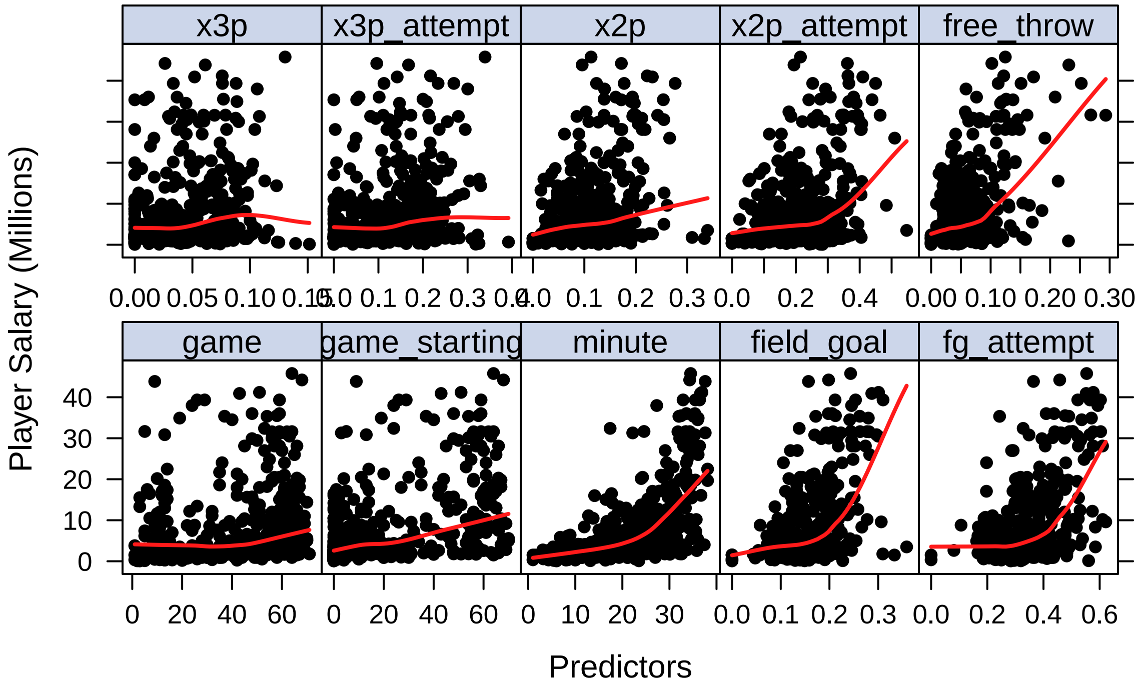


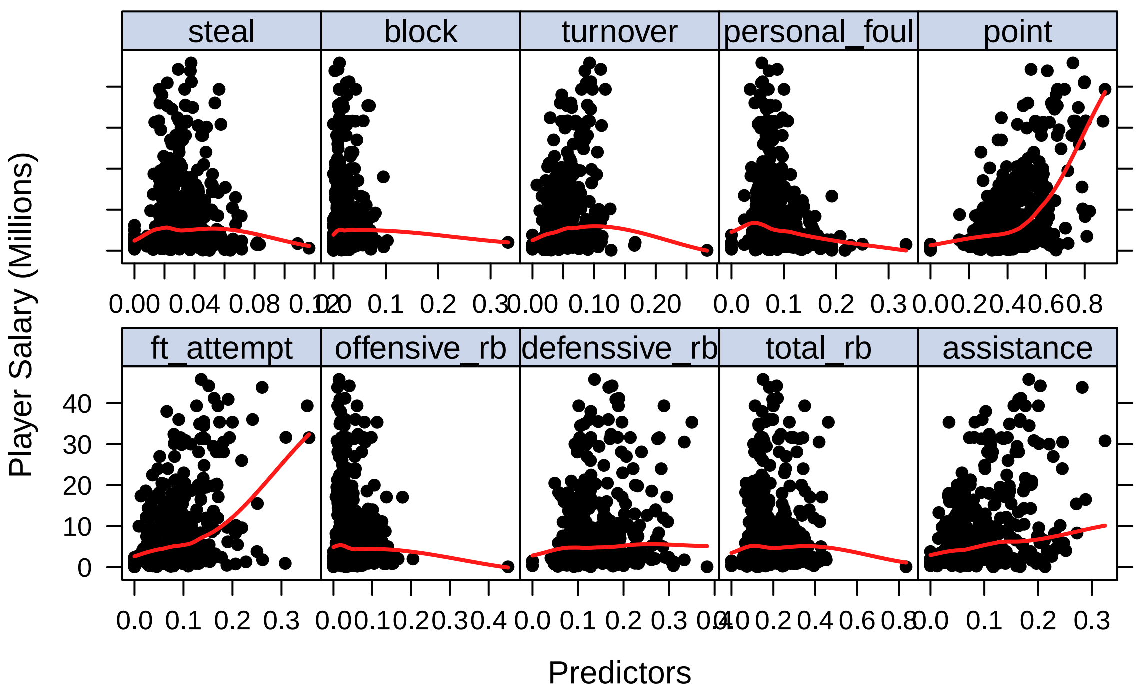
Correlation Analysis



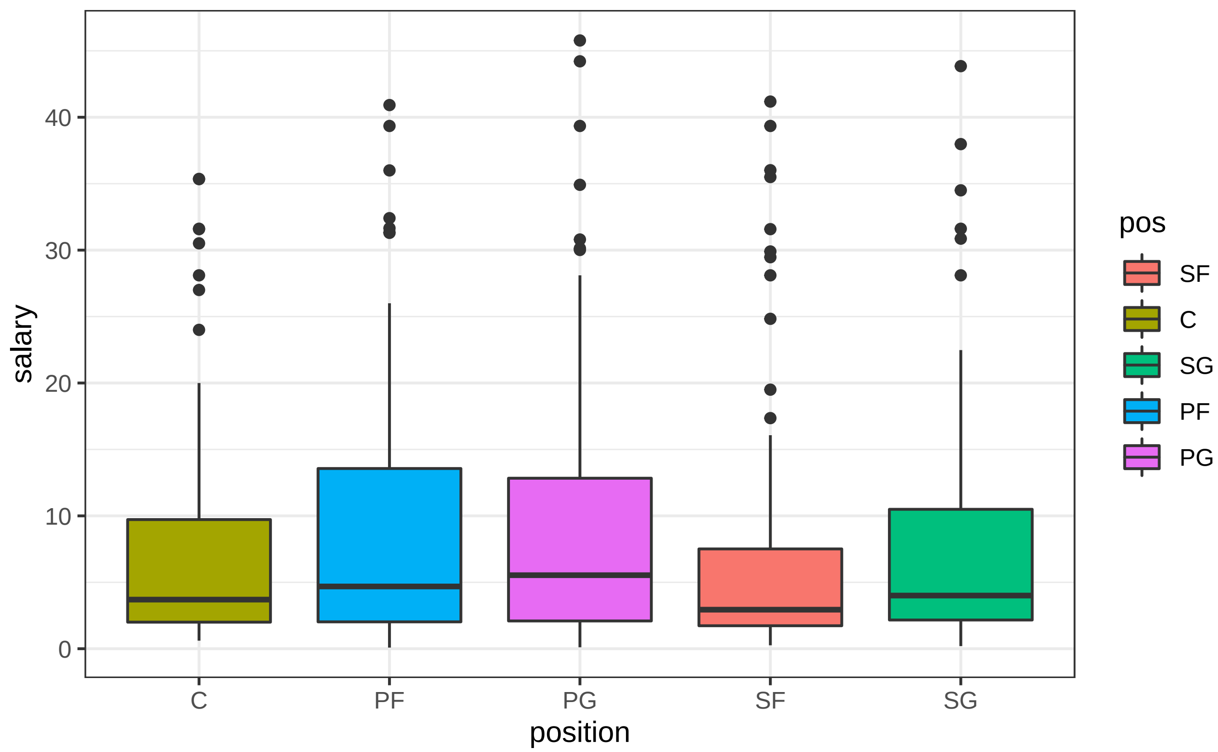
Analyzing trends in data

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.

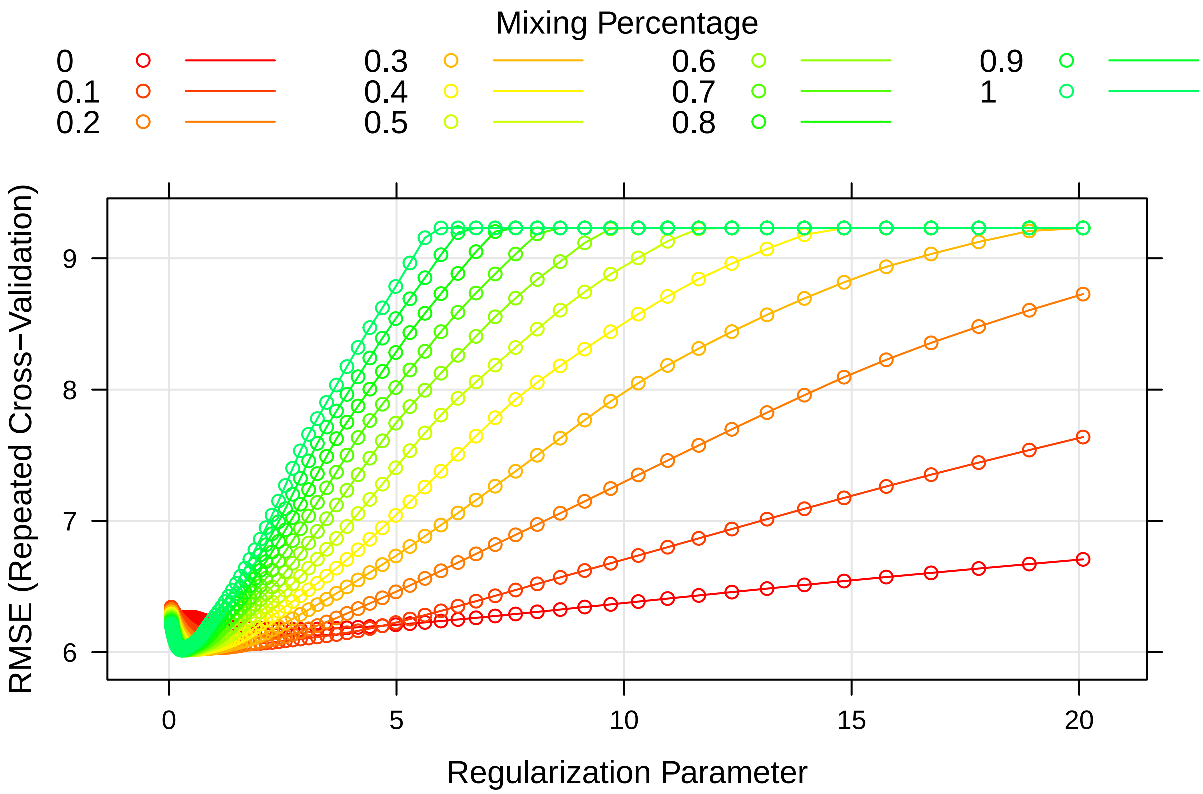

salary
40
30
20
10
0
C
PF
PG
SF
SG
position
pos
SF
C
SG
PF
PG

Models

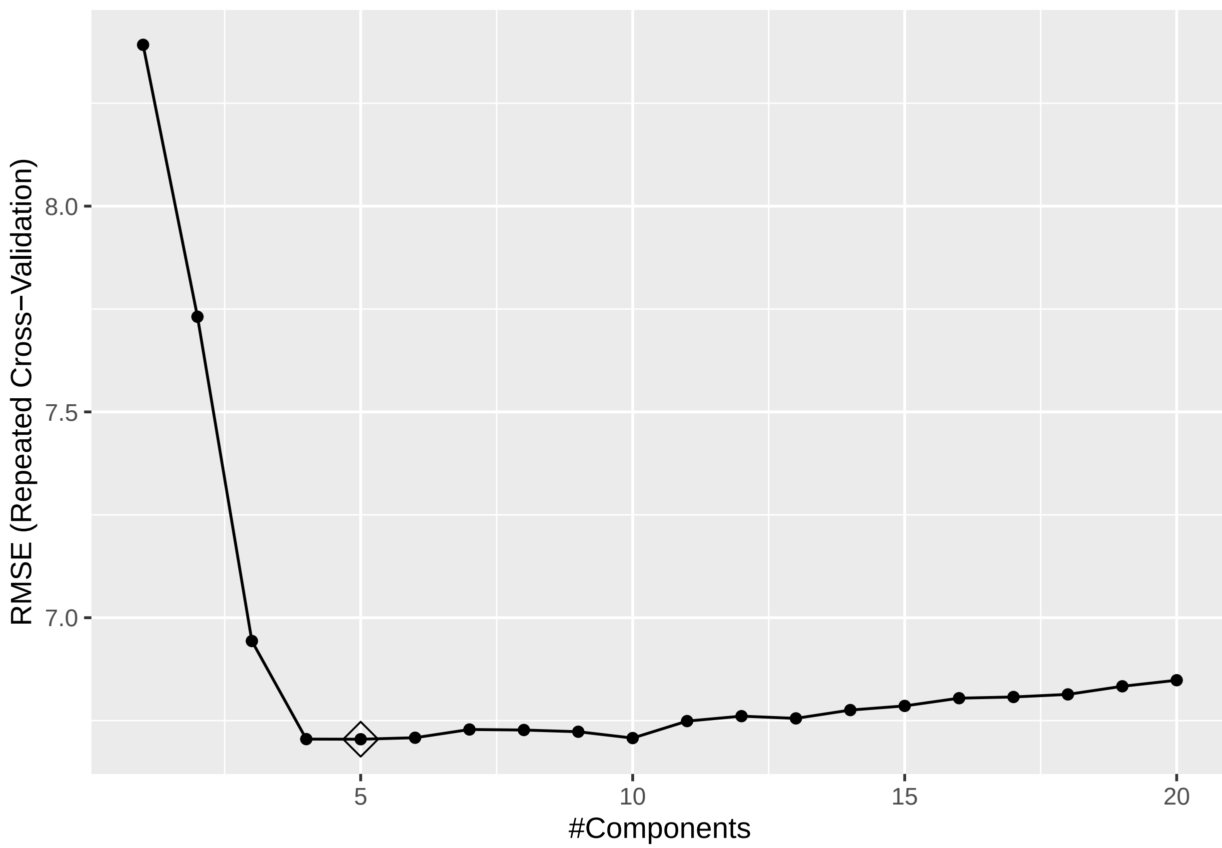
Part 1 Linear regression

(a) Standard Least-Squared

(b) Elastic Net (including lasso/ridge)



###(c) Principle Component Regression



Part 2 Generalized Linear Regression

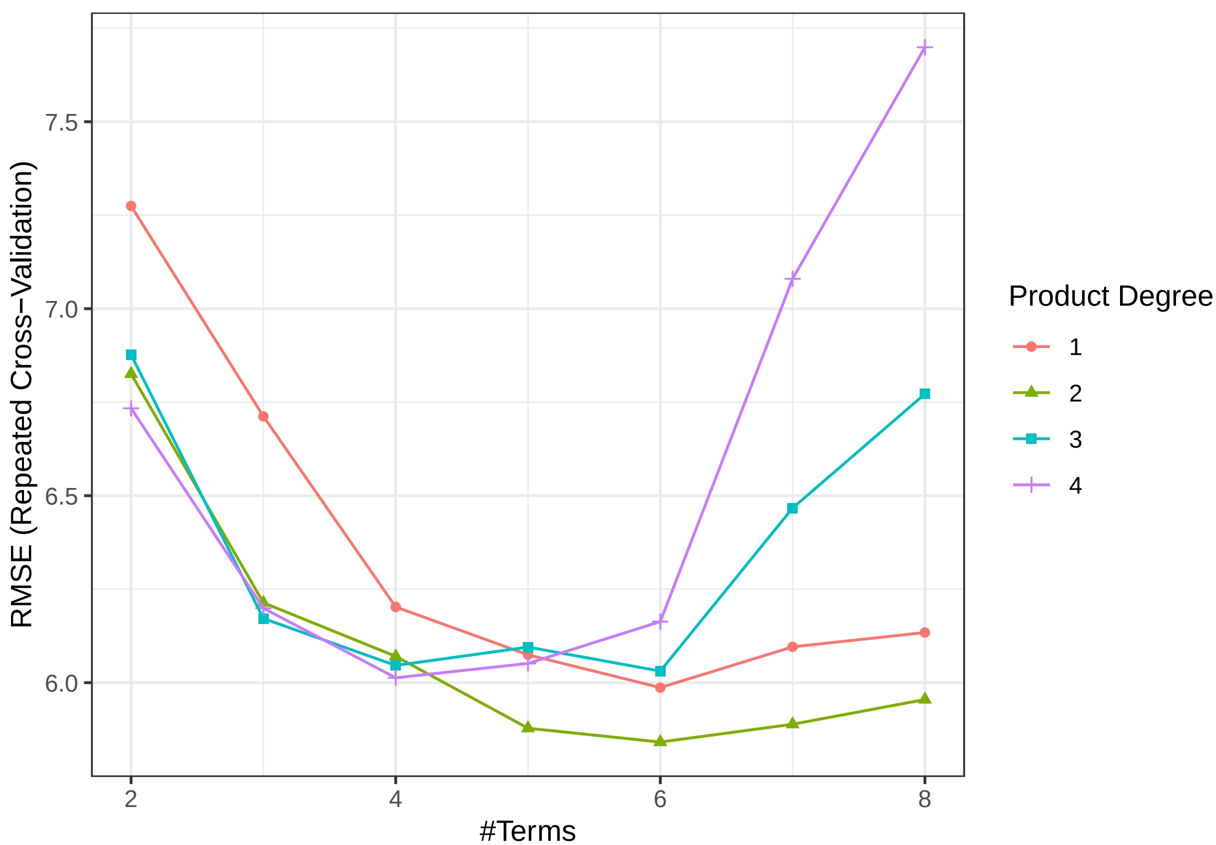
(a) GAM

```
##
## 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)    8.151      0.301   27.08   <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)           4.722  5.775 14.002  < 2e-16 ***
## s(game)          1.000  1.000  4.324 0.038422 *
## s(game_starting) 1.532  1.883 23.181  < 2e-16 ***
## s(free_throw)    7.542  8.452  2.095 0.022370 *
```

```
## s(ft_attempt)     2.098  2.759  0.603 0.485917
## s(defenssive_rb) 1.330  1.585  2.465 0.065744 .
## s(assistance)    1.114  1.217 17.575 2.90e-05 ***
## s(block)         1.000  1.000  0.009 0.923298
## s(personal_foul) 7.693  8.529  3.699 0.000214 ***
## s(point)         3.351  4.242  6.044 7.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.652   Deviance explained = 68.5%
## GCV = 33.514  Scale est. = 30.265    n = 334
```

(b) MARS

```
##    nprune degree
## 12      6      2
```




```
## [1] 40.30051
```

Table 1: RMSE of Different Models

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
LeastSquare	4.92	5.92	6.41	6.44	6.89	9.04	0

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
ElasticNet	4.36	5.44	5.88	6.02	6.59	8.22	0
PCR	4.07	6.10	6.78	6.70	7.41	8.80	0
MARS	4.04	5.38	5.82	5.84	6.45	8.25	0

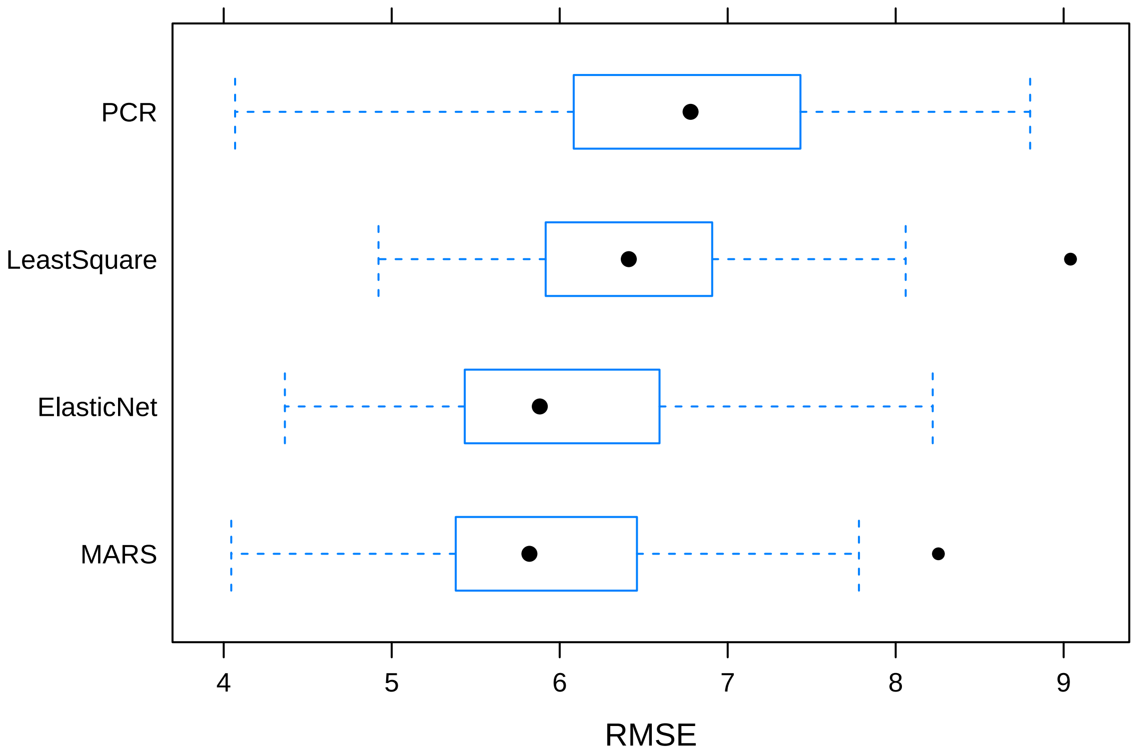


Table 2: RMSE of Different Models on Test Set

	Linear	ElasticNet	PCR	GAM	MARS
RMSE	7.25	7.21	7.34	7.19	6.35