# Football Player Performance Analysis and Prediction

Haohan Cui
Tianfeng Liao
Yan Shao

SUSTech

December 17, 2024

# Outline



南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

# Outline

# Clear Data

```
mis <- is.na(data1)
misss <- is.na(data2)
all(names(data1) == names(data2))
```

```
## Warning in names(data1) == names(data2): 长的对象长度不是短的对象长度的整倍数
```

```
## [1] FALSE
```

```
data2$综合能力 <- NA
all(names(data1) == names(data2))
```

```
## [1] TRUE
```

Figure: Check Data

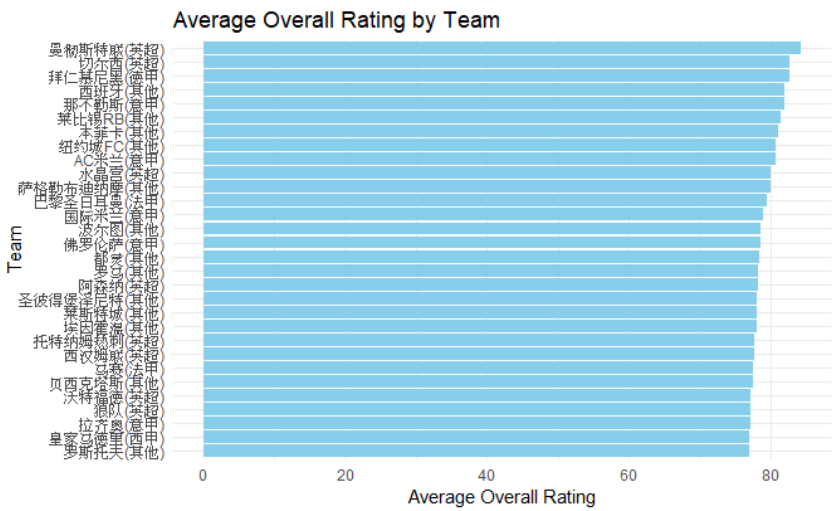
```
sorted_data <- cmobi %>%
  arrange(姓名, 年龄)
head(sorted_data)
```

```
## # A tibble: 6 × 47
##   姓名      年龄  身高  体重 俱乐部 位置  惯用脚 逆足能力 花式技巧 国际声誉  传中
##   <chr>   <dbl> <dbl> <dbl> <chr>  <chr> <chr>     <dbl>    <dbl>    <dbl> <dbl>
## 1 Aarón …    19   188.   159 马拉加 替补  右脚          3        1        1    20
## 2 Aarón …    20   188.   159 马拉加 替补  右脚          3        1        1    20
## 3 Aarón …    21   183.   159 格拉…  替补  右脚          3        1        1    20
## 4 Aarón …    22   185.   157 格拉…  主力… 右脚          3        1        1    20
## 5 Aarón …    23   185.   157 格拉…  主力… 右脚          3        1        1    20
## 6 Aarón …    24   185.   157 格拉…  主力… 右脚          3        1        1    20
## # i 36 more variables: 射术 <dbl>, 头球精度 <dbl>, 短传 <dbl>, 凌空 <dbl>,
## #   盘带 <dbl>, 弧线 <dbl>, 任意球精度 <dbl>, 长传 <dbl>, 控球 <dbl>,
## #   加速 <dbl>, 速度 <dbl>, 敏捷 <dbl>, 移动反应 <dbl>, 平衡 <dbl>,
## #   射门力量 <dbl>, 弹跳 <dbl>, 体能 <dbl>, 强壮 <dbl>, 远射 <dbl>,
## #   侵略性 <dbl>, 拦截意识 <dbl>, 跑位 <dbl>, 视野 <dbl>, 点球 <dbl>,
## #   沉着 <dbl>, 盯人 <dbl>, 抢断 <dbl>, 铲球 <dbl>, 鱼跃 <dbl>, 手形 <dbl>,
## #   开球 <dbl>, 站位 <dbl>, 守门反应 <dbl>, 联赛 <chr>, 球队位置 <chr>, …
```

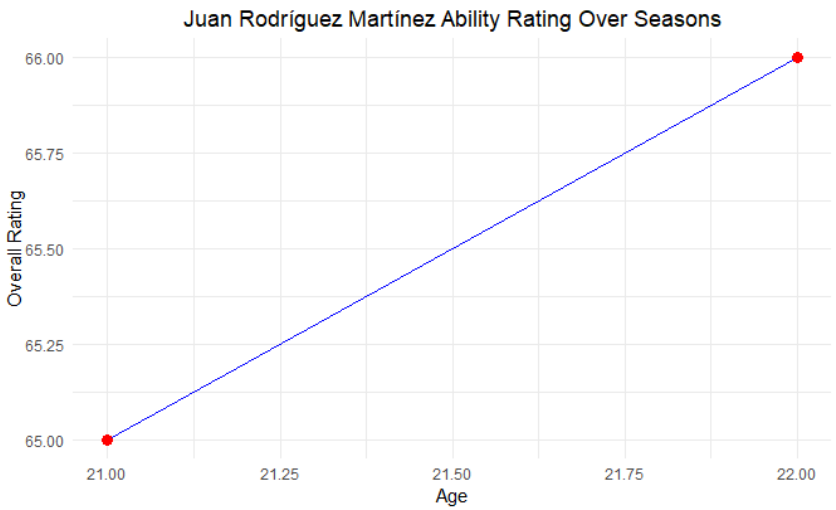
Figure: Combine Two Datasets

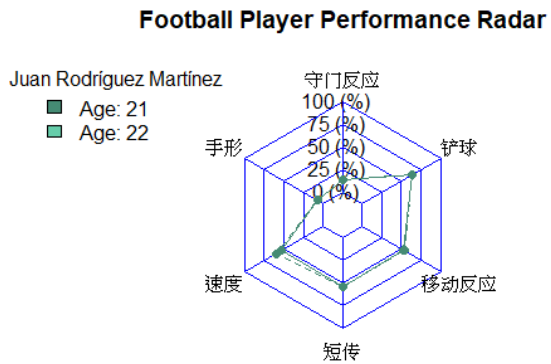
# Football Team Overall Ability

**Overview:**
We selected the top 30 teams with average comprehensive strength rankings for visualization.



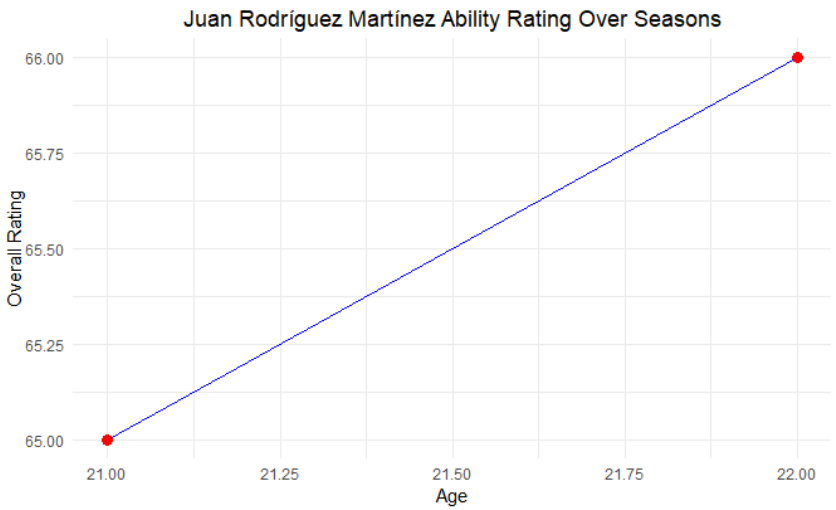
# Player Visual Chart

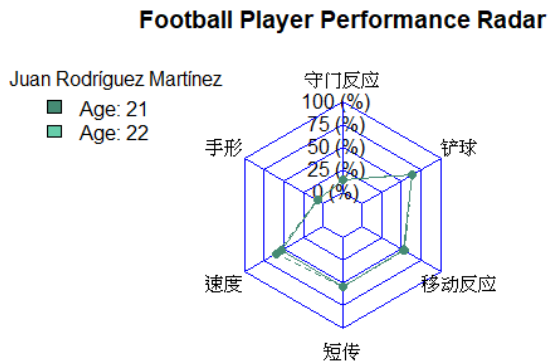



## Key Insights:

# Player Visual Chart



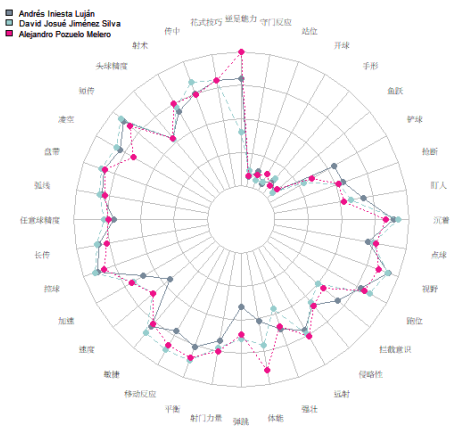




**Key Insights:**

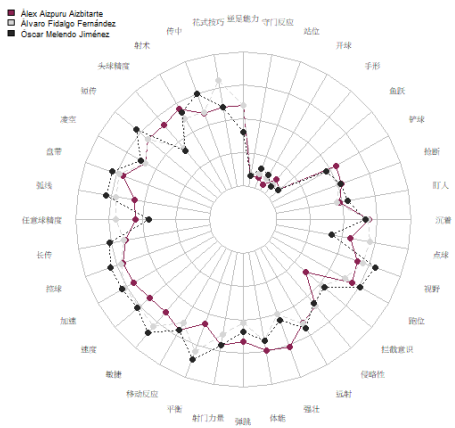
- Two charts showing the changes in various abilities of a player as his age.

# Another View

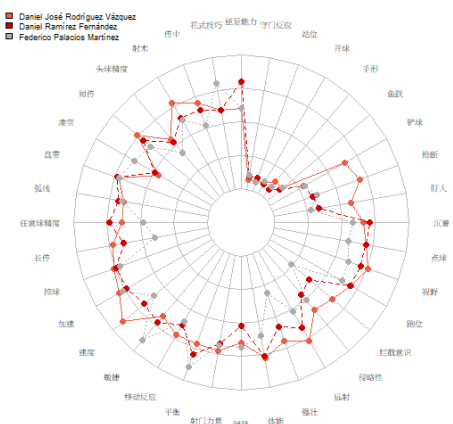
**Football Player Performance Radar in the same position**



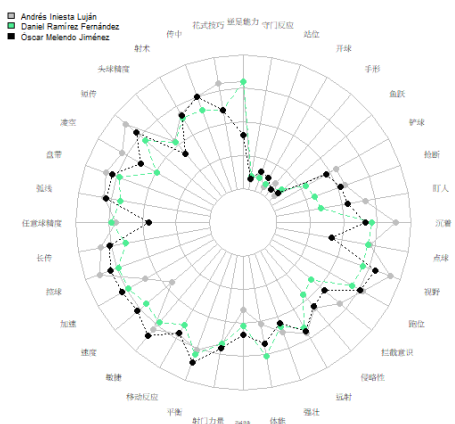

**Football Player Performance Radar in the same position**




**Football Player Performance Radar in the same position**




**Football Player Performance Radar in the same position**



# Another View

```
new <- golaas[, -c(45,46,47,2,3,4,5,6,7,10,1)]
p_values <- numeric(ncol(new))
for (i in 2:ncol(new)) {
  test_result <- t.test(new[[i]], mu = mean(new[[i]]))
  p_values[i] <- test_result$p.value
}
names(p_values) <- colnames(new)
significant_columns <- names(p_values[p_values < 0.5])
print("Columns with p-value < 0.5:")
```

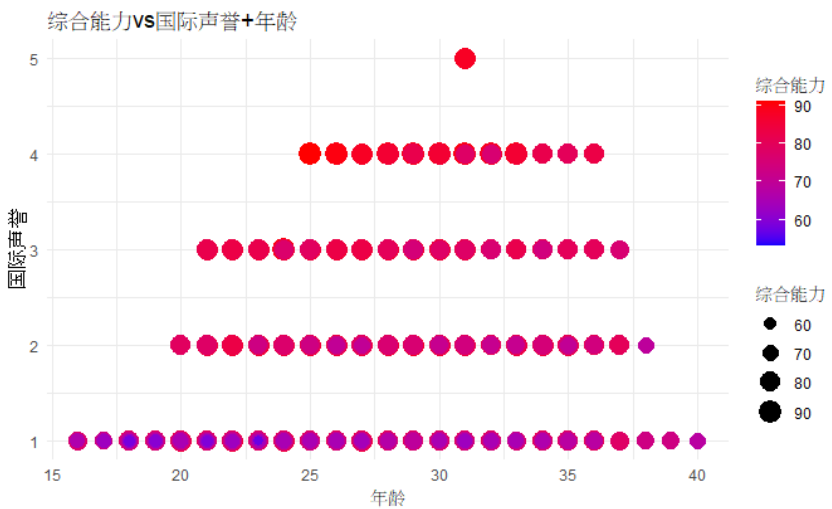
```
## [1] "Columns with p-value < 0.5:"
```

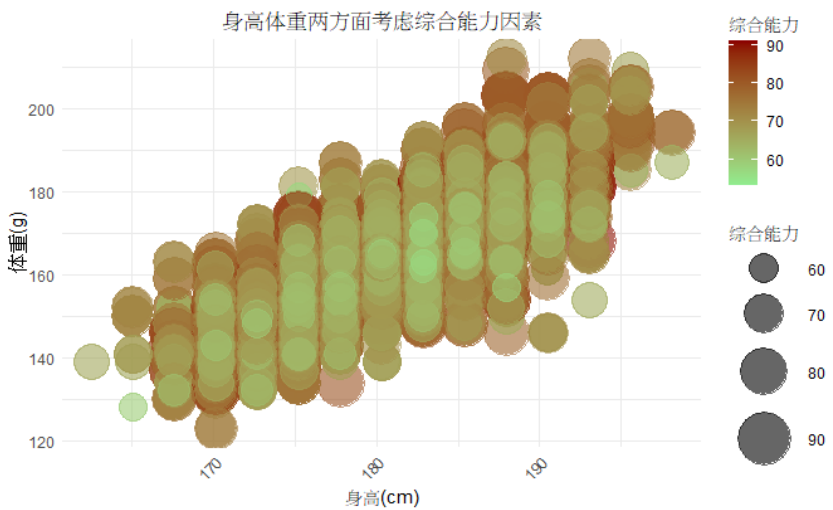
```
print(significant_columns)
```

```
## [1] "逆足能力"
```

Figure: T test for some guesses

# About Comprehensive Ability

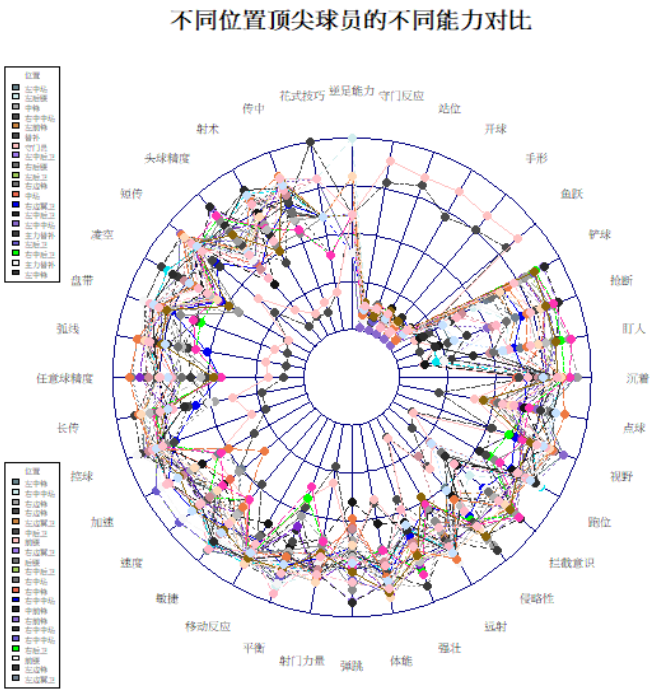


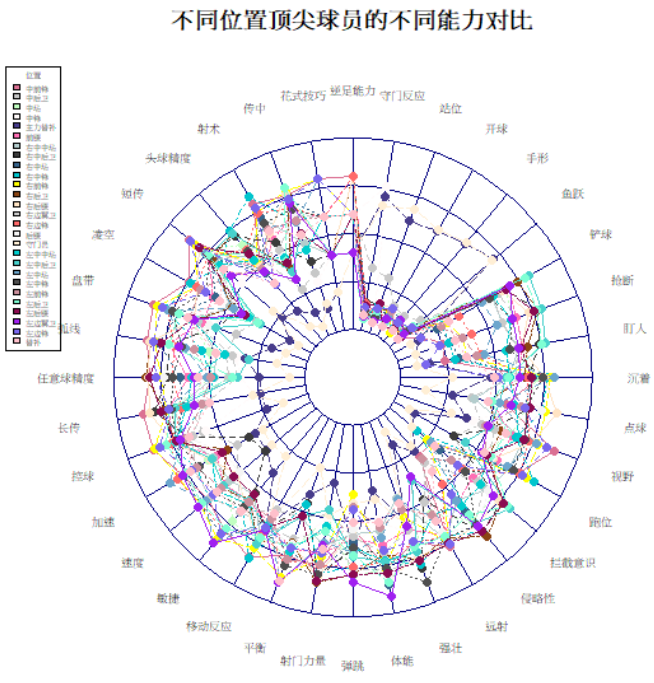




**Two View:Reputation and Age v.s Height and Weight**

# Comparison of Top Players' in Different Positions

**Summary:**




南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

# Outline

# Dealing Datasets

```
##                      PC1
## 年龄          -0.019488933
## 逆足能力      -0.067403986
## 花式技巧      -0.176517608
## 传中          -0.198640942
## 射术          -0.181021523
## 头球精度      -0.154289287
## 短传          -0.212485468
## 凌空          -0.177365149
## 盘带          -0.214369774
## 弧线          -0.197948001
## 任意球精度    -0.186189963
## 长传          -0.184790511
## 控球          -0.222548641
## 加速          -0.150277626
## 速度          -0.139751516
## 敏捷          -0.145356941
## 移动反应      -0.083694003
## 平衡          -0.132896022
## 射门力量      -0.183372317
## 弹跳          -0.011023951
## 体能          -0.172088676
## 强壮          -0.015827332
## 远射          -0.193927459
## 侵略性        -0.137278484
## 拦截意识      -0.111706749
## 跑位          -0.204577641
## 视野          -0.164452256
## 点球          -0.179307668
```

```
head(df)
```

```
## # A tibble: 6 × 39
##    年龄 逆足能力 花式技巧  传中  射术 头球精度  短传  凌空  盘带  弧线
##   <dbl>    <dbl>    <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl>
## 1    31        4        4    79    73       54    92    74    90    80
## 2    24        3        1    17    13       21    31    13    13    21
## 3    29        3        3    74    60       86    76    66    57    73
## 4    26        3        3    62    60       70    89    44    74    66
## 5    28        3        3    86    76       74    92    81    81    77
## 6    29        2        4    85    72       54    92    80    87    83
## # i 29 more variables: 任意球精度 <dbl>, 长传 <dbl>, 控球 <dbl>, 加速 <dbl>,
## #   速度 <dbl>, 敏捷 <dbl>, 移动反应 <dbl>, 平衡 <dbl>, 射门力量 <dbl>,
## #   弹跳 <dbl>, 体能 <dbl>, 强壮 <dbl>, 远射 <dbl>, 侵略性 <dbl>,
## #   拦截意识 <dbl>, 跑位 <dbl>, 视野 <dbl>, 点球 <dbl>, 沉着 <dbl>, 盯人 <dbl>,
## #   抢断 <dbl>, 铲球 <dbl>, 鱼跃 <dbl>, 手形 <dbl>, 开球 <dbl>, 站位 <dbl>,
## #   守门反应 <dbl>, 综合能力 <dbl>, BMI <dbl>
```

Correction Data

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     4.3405 2.3446 1.6758 1.4285 1.25143 0.95312 0.94147
## Proportion of Variance 0.4958 0.1447 0.0739 0.0537 0.04121 0.02391 0.02333
## Cumulative Proportion  0.4958 0.6405 0.7144 0.7681 0.80928 0.83318 0.85651
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.79022 0.74293 0.68724 0.60525 0.56395 0.53623 0.50804
## Proportion of Variance 0.01643 0.01452 0.01243 0.00964 0.00837 0.00757 0.00679
## Cumulative Proportion  0.87294 0.88747 0.89990 0.90954 0.91790 0.92547 0.93226
```

Part of Results

# Choose Important Factors

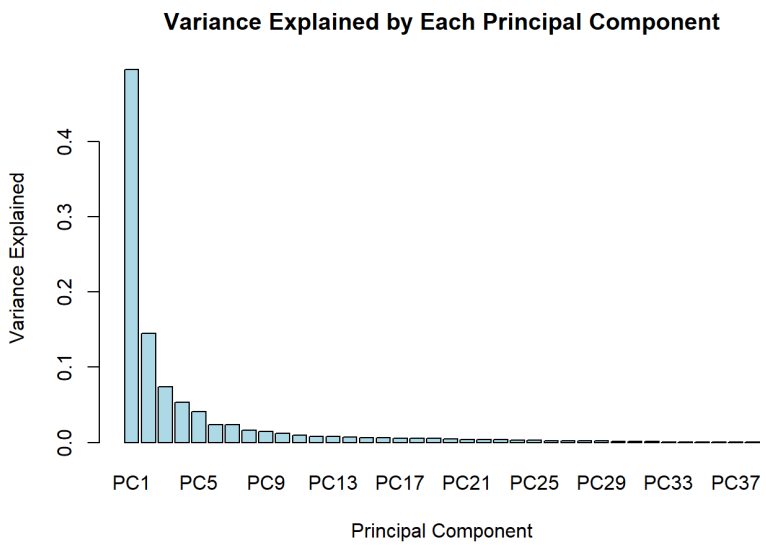



Figure: Propotion of Variance

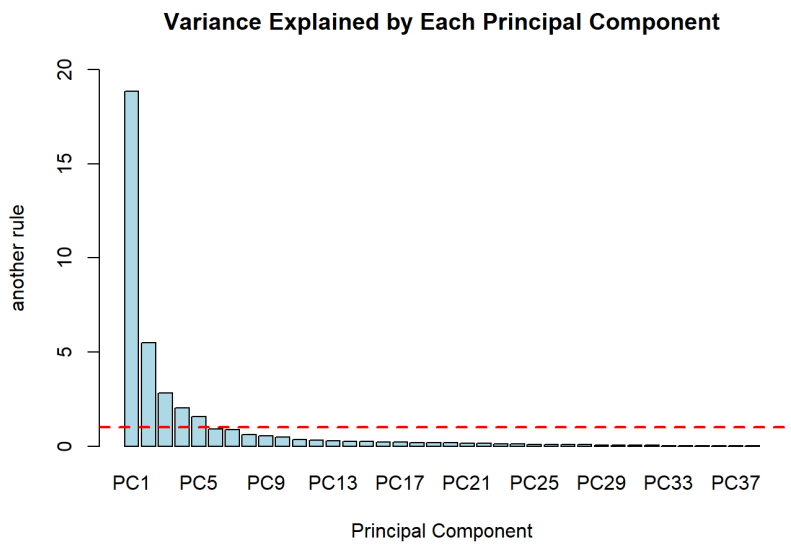



Figure: Standard Deviation

# PCA Scores Plot

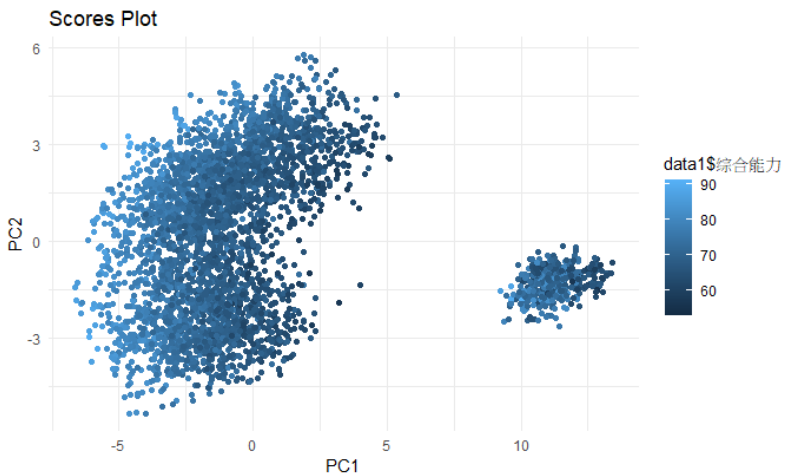



Figure: Scores Plot of PCA1&2

# PCA Scores Plot

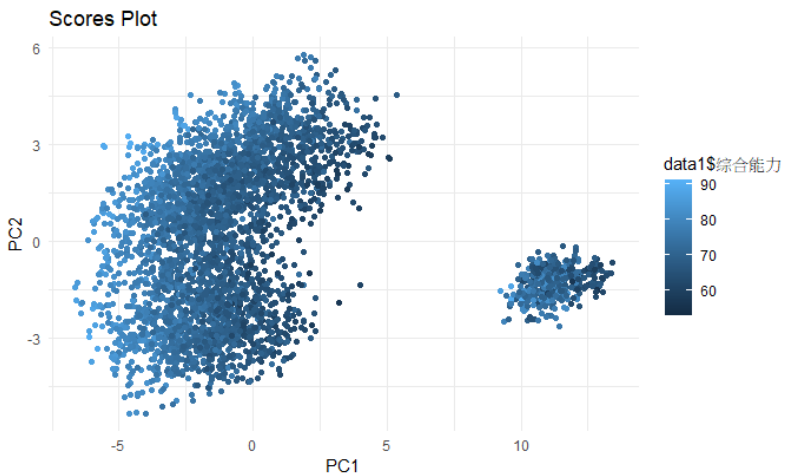



Figure: Scores Plot of PCA1&2

- Left Cluster: Points are sparse over a wider range, covering a broad spectrum of PC1 and PC2 values.

# PCA Scores Plot

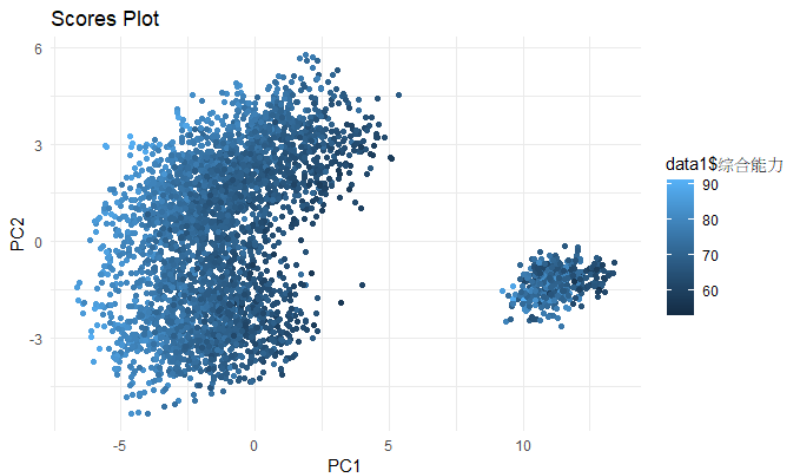



Figure: Scores Plot of PCA1&2

- Left Cluster: Points are sparse over a wider range, covering a broad spectrum of PC1 and PC2 values.
- Right Cluster: Points are more concentrated, primarily around high PC1 values (approximately 10) and a narrow range for PC2 (close to 0). This separation may indicate underlying categories or groups in the data

# Result

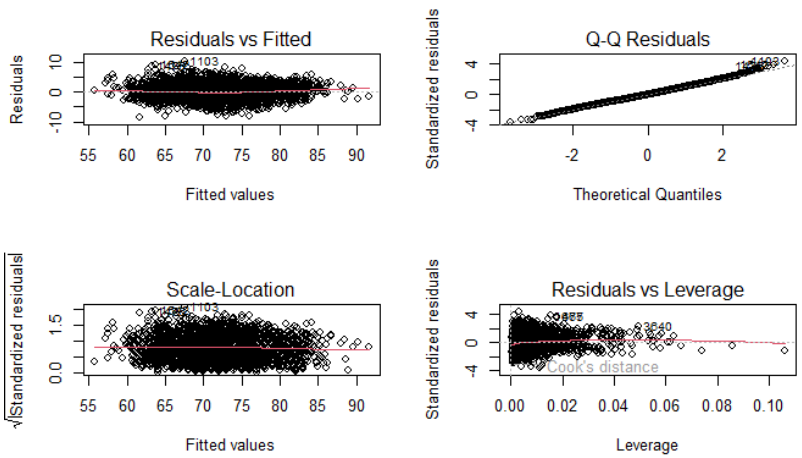



Figure: First Trial

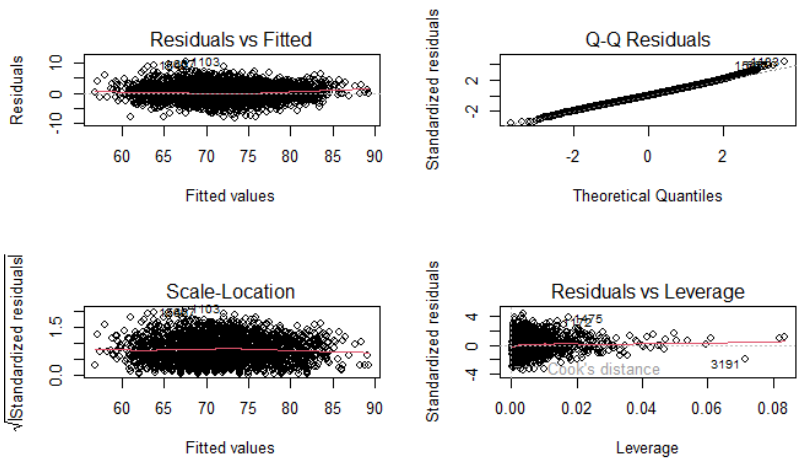



Figure: Final Trial

# Charts

```
##   [1] 3 1 2 2 3 3 3 3 2 3 3 2 3 3 3 3 3 1 2 3 2 3 3 3 3 3 3 2 1 2 3 1 2 1 2 2 3
##  [38] 3 3 3 2 2 2 2 3 1 2 3 1 3 2 3 3 2 3 3 2 3 1 3 2 2 3 3 2 2 1 3 1 3 3 3 3 3
##  [75] 3 3 3 3 2 3 2 3 2 1 2 3 3 3 3 2 3 2 3 2 2 2 2 3 2 3 3 2 3 2 3 1 1 3 3 3 1
## [112] 3 2 3 2 2 3 2 2 3 3 1 2 2 2 1 2 3 2 2 2 3 3 3 3 3 2 2 2 3 3 3 3 3 2 2 1 2
## [149] 3 2 1 3 2 3 3 3 2 3 1 2 3 3 2 1 3 3 3 1 2 3 3 3 3 2 2 3 2 3 3 2 2 2 2 2 2
## [186] 2 2 3 3 1 2 3 2 2 3 3 1 3 2 2 2 1 2 2 2 2 2 2 3 3 3 1 3 2 2 3 2 2 3 1 3 2
## [223] 1 3 3 1 3 1 2 2 3 3 2 1 3 3 3 2 2 3 3 3 2 2 2 2 3 3 2 3 3 3 2 3 2 2 3 3 2
## [260] 2 3 3 2 2 2 3 2 3 2 2 2 3 3 3 1 1 3 3 2 2 2 3 2 3 2 2 3 2 2 2 2 3 2 2 2 2
## [297] 2 2 2 3 2 3 3 2 3 3 2 2 1 1 3 2 2 3 3 3 3 2 3 3 3 3 3 3 1 3 3 3 2 1 2 2 3
## [334] 1 2 3 2 2 2 3 2 3 3 2 2 3 2 3 3 3 2 3 2 2 3 3 2 2 3 3 2 3 1 2 3 3 3 2 2 2 3
## [371] 3 2 3 1 2 2 3 2 2 3 3 2 2 1 3 3 2 1 3 2 2 2 1 2 1 3 3 2 3 2 2 3 3 2 3 3 3
## [408] 3 2 2 2 2 1 3 2 3 3 3 2 3 3 3 2 3 2 3 2 1 2 3 2 2 1 3 3 2 2 3 3 2 3 3 2 2
## [445] 3 3 2 2 3 2 3 2 3 3 3 2 2 2 3 3 3 2 2 3 2 2 2 3 1 2 2 2 2 2 2 2 1 2 2 1
## [482] 3 3 3 3 3 3 3 2 2 2 2 2 2 2 1 2 3 3 1 3 3 2 1 2 2 2 2 3 2 3 3 3 3 2 2 3 2
## [519] 2 3 1 2 2 3 2 2 3 2 3 2 2 2 3 3 2 3 3 2 2 3 2 3 2 3 3 3 3 3 3 2 1 3 1 2 3
## [556] 2 3 1 2 2 2 2 2 3 2 3 2 2 2 2 2 2 3 2 3 2 3 3 3 2 3 3 3 2 2 2 3 2 1 3 3 2
## [593] 3 1 3 3 3 2 1 2 1 2 2 2 3 3 3 2 2 2 3 2 2 1 2 3 2 3 3 3 1 2 2 3 3 2 3 2 2
```

Figure: K-mean Result

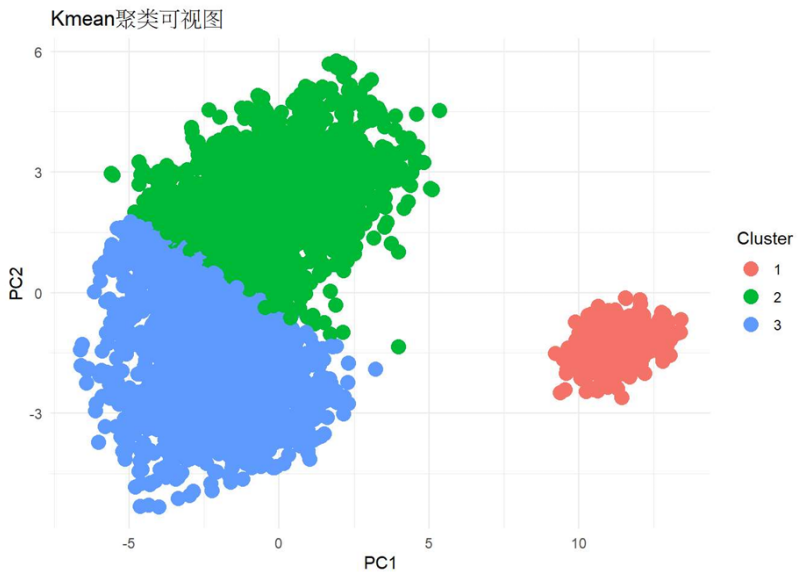



Figure: K-mean Chart

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

# Result

```
> head(df_reordered)
# A tibble: 6 × 40
  Cluster  年龄 逆足能力 花式技巧  传中  射术 头球精度  短传  凌空  盘带  弧线 任意球精度  长传
    <int> <dbl>    <dbl>    <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl>      <dbl> <dbl>
1       2    31        4        4    79    73       54    92    74    90    80         70    86
2       1    24        3        1    17    13       21    31    13    13    21         19    32
3       3    29        3        3    74    60       86    76    66    57    73         68    70
4       3    26        3        3    62    60       70    89    44    74    66         68    79
5       2    28        3        3    86    76       74    92    81    81    77         81    91
6       2    29        2        4    85    72       54    92    80    87    83         77    85
# ℹ 27 more variables: 控球 <dbl>, 加速 <dbl>, 速度 <dbl>, 敏捷 <dbl>, 移动反应 <dbl>,
#   平衡 <dbl>, 射门力量 <dbl>, 弹跳 <dbl>, 体能 <dbl>, 强壮 <dbl>, 远射 <dbl>, 侵略性 <dbl>,
#   拦截意识 <dbl>, 跑位 <dbl>, 视野 <dbl>, 点球 <dbl>, 沉着 <dbl>, 盯人 <dbl>, 抢断 <dbl>,
#   铲球 <dbl>, 鱼跃 <dbl>, 手形 <dbl>, 开球 <dbl>, 站位 <dbl>, 守门反应 <dbl>,
#   综合能力 <dbl>, BMI <dbl>
```

Figure: Completed Data

# Outline



部分数据参考来源：FC 25

最后更新时间：2024-10-18

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.976 on 3922 degrees of freedom
Multiple R-squared:  0.7331,    Adjusted R-squared:  0.7305
F-statistic: 283.5 on 38 and 3922 DF,  p-value: < 2.2e-16
```

Figure: model_0

# Stepwise Prediction

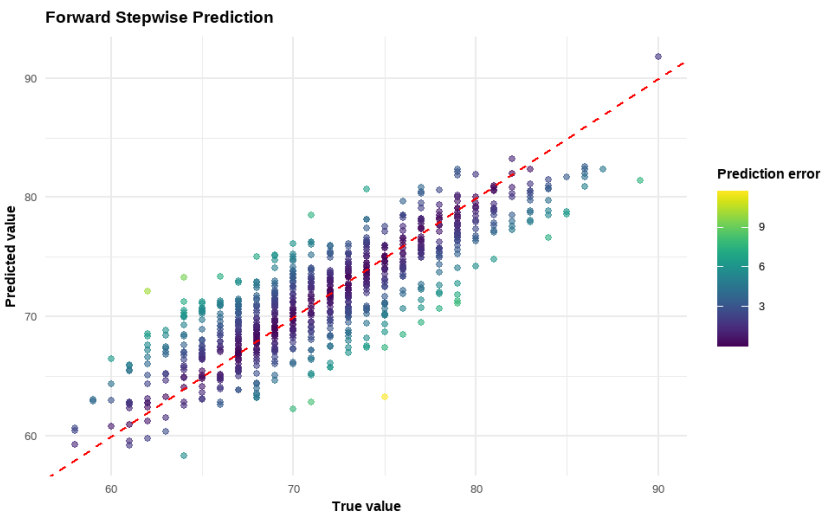



Figure: Forward

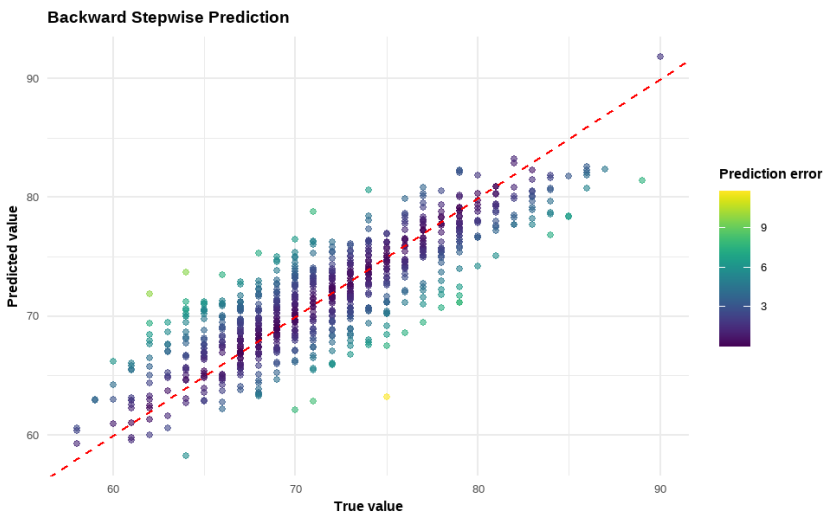



Figure: Backward

# Stepwise Method

Forward and backward stepwise direction choose 24 and 26 variables respectively.

```
========== Forward Stepwise 选择的变量 ==========
综合能力 ~ 移动反应 + 长传 + 站位 + 控球 + 强壮 +
    速度 + 短传 + 鱼跃 + 头球精度 + 手形 + 沉着 +
    跑位 + 年龄 + 体能 + 花式技巧 + 侵略性 + 任意球精度 +
    开球 + 加速 + 凌空 + 点球 + BMI + 守门反应 +
    平衡

========== Backward Stepwise 选择的变量 ==========
综合能力 ~ 年龄 + 花式技巧 + 头球精度 + 短传 +
    凌空 + 任意球精度 + 长传 + 控球 + 加速 + 速度 +
    移动反应 + 平衡 + 体能 + 强壮 + 侵略性 + 拦截意识 +
    跑位 + 点球 + 沉着 + 抢断 + 鱼跃 + 手形 + 开球 +
    站位 + 守门反应 + BMI
```

Figure: Results

# LASSO

All coefficients are larger than 0! Use top 25 variables to construct a linear model.

# Comparison

Model validation is performed using k-fold cross-validation, implemented with the `crossval()` function from the `bootstrap` package.

| **Model** | **CV RMSE** | **AIC** | **$R^2$** |
|---|---|---|---|
| Forward | 2.96 | 15926.95 | 0.7361 |
| Backward | 2.95 | 15914.89 | 0.7374 |
| LASSO | 8.92 | 15917.33 | 0.7370 |

- But for the true LASSO model, the AIC is about 6929.927.

# Comparison

Model validation is performed using k-fold cross-validation, implemented with the `crossval()` function from the `bootstrap` package.

| **Model** | **CV RMSE** | **AIC** | **$R^2$** |
|---|---|---|---|
| Forward | 2.96 | 15926.95 | 0.7361 |
| Backward | 2.95 | 15914.89 | 0.7374 |
| LASSO | 8.92 | 15917.33 | 0.7370 |

- But for the true LASSO model, the AIC is about 6929.927.
- Backward result is chosen.

# Normality, independence, linearity, homoscedasticity

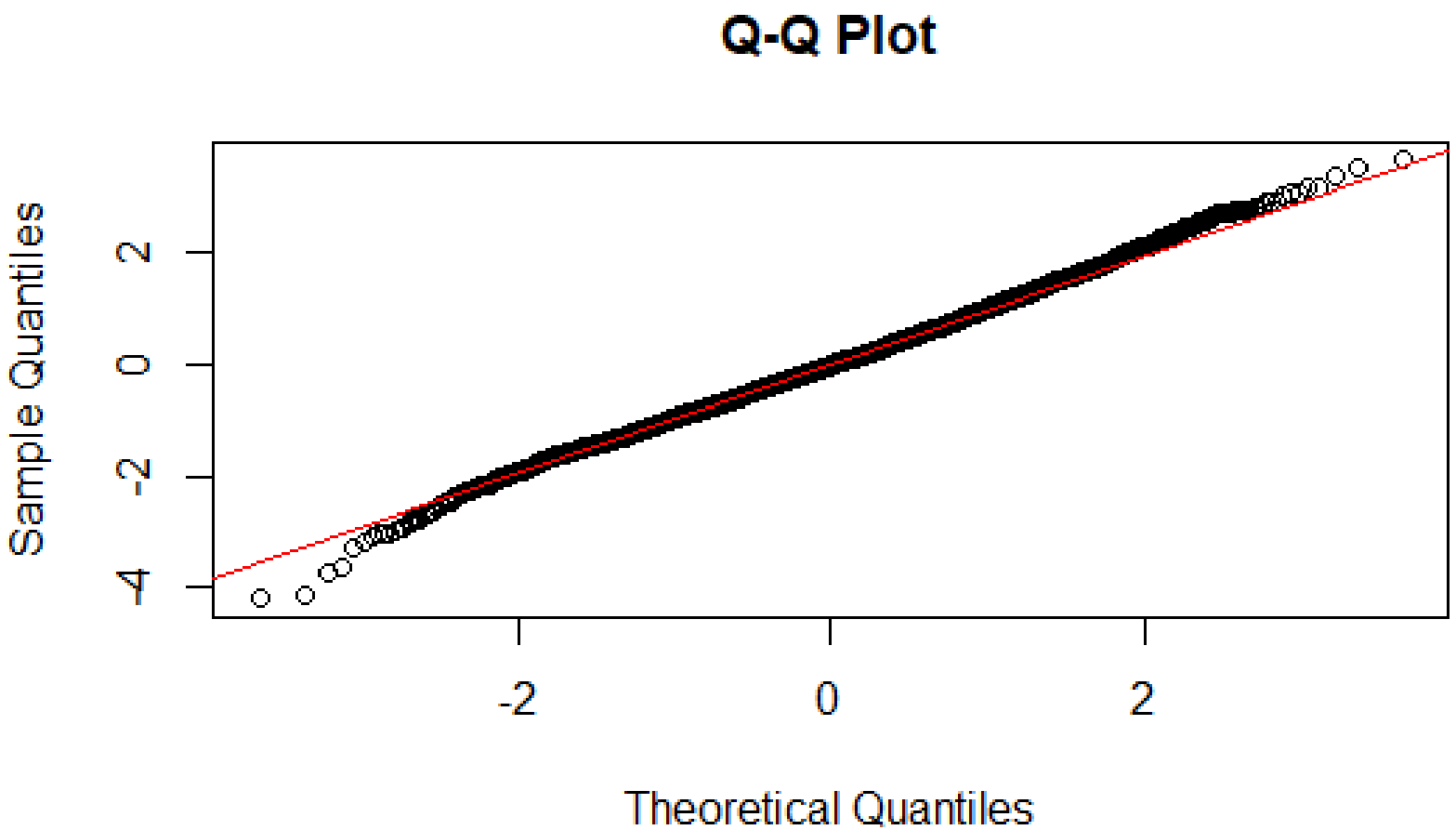
```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  model_back$residuals
D = 0.24038, p-value < 2.2e-16
alternative hypothesis: two-sided

 lag Autocorrelation D-W Statistic p-value
   1       0.1168295      1.765556       0
 Alternative hypothesis: rho != 0
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 16.75128, Df = 1, p = 4.2614e-05
```

# Normality, independence, linearity, homoscedasticity

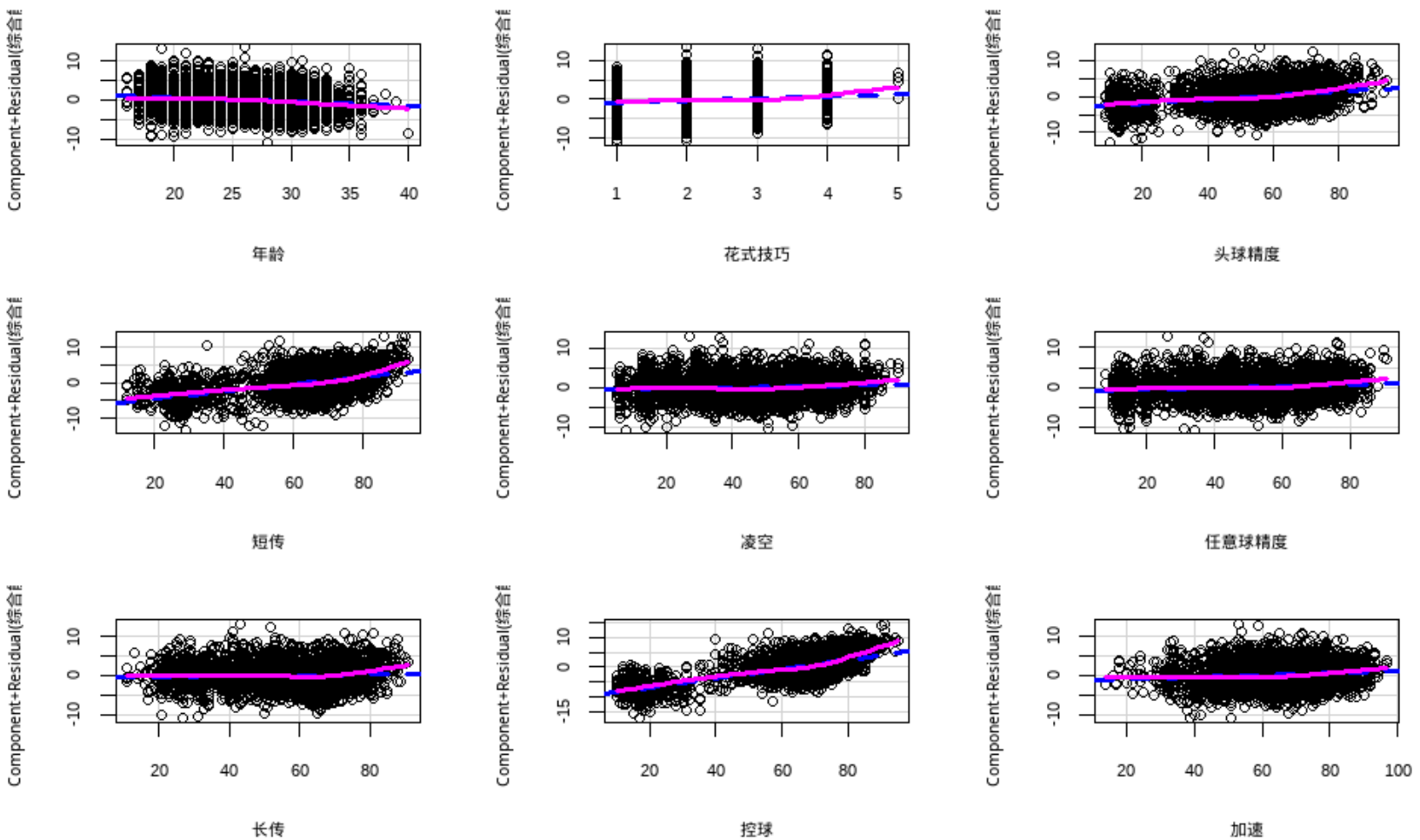
| | Value <dbl> | p-value <dbl> | Decision <chr> |
|---|---|---|---|
| Global Stat | 262.88325 | 0.000000e+00 | Assumptions NOT satisfied! |
| Skewness | 38.28176 | 6.123237e-10 | Assumptions NOT satisfied! |
| Kurtosis | 31.52233 | 1.971601e-08 | Assumptions NOT satisfied! |
| Link Function | 154.06471 | 0.000000e+00 | Assumptions NOT satisfied! |
| Heteroscedasticity | 39.01445 | 4.206793e-10 | Assumptions NOT satisfied! |

Figure: gvlma()

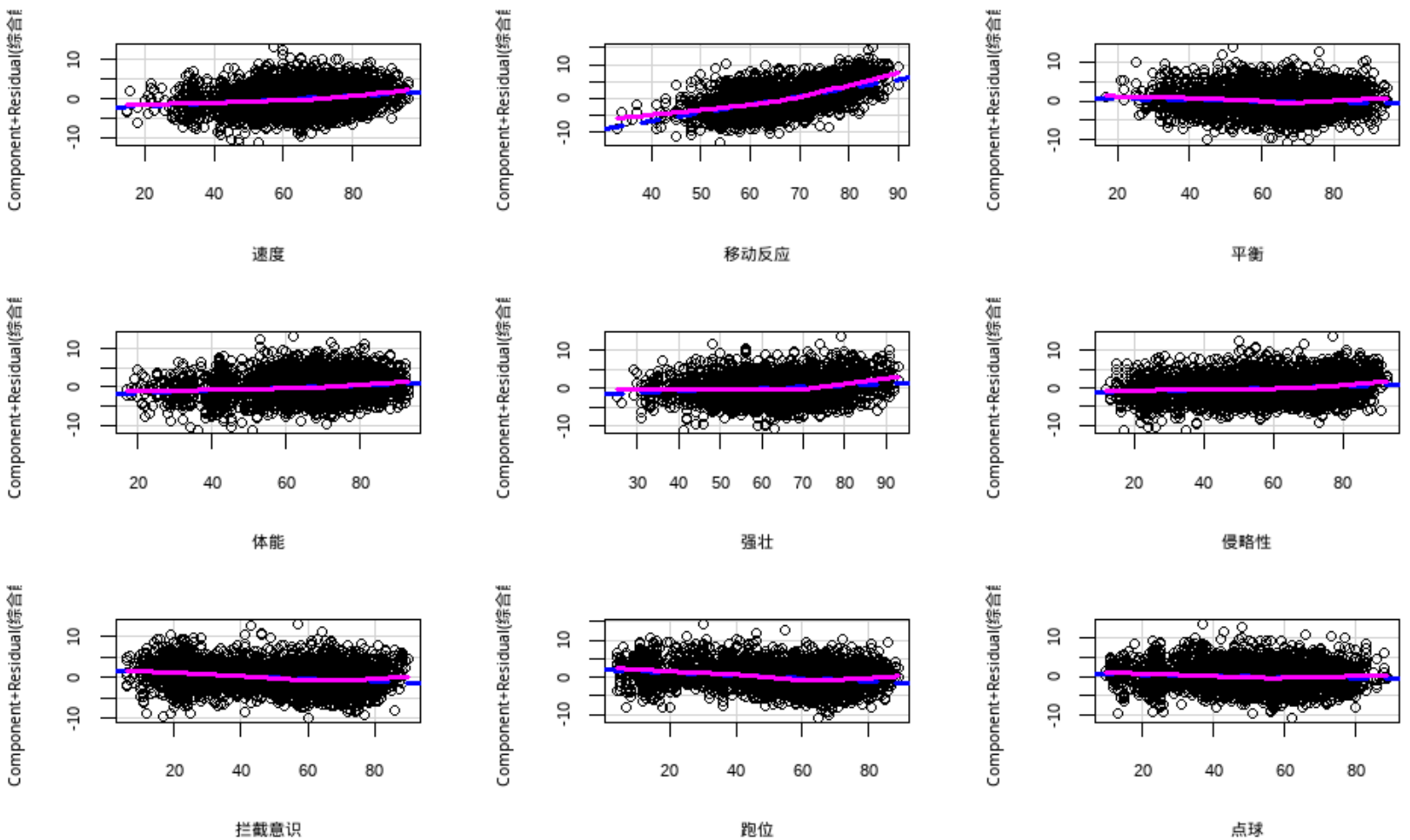
# Normality, Independence, Linearity, Homoscedasticity



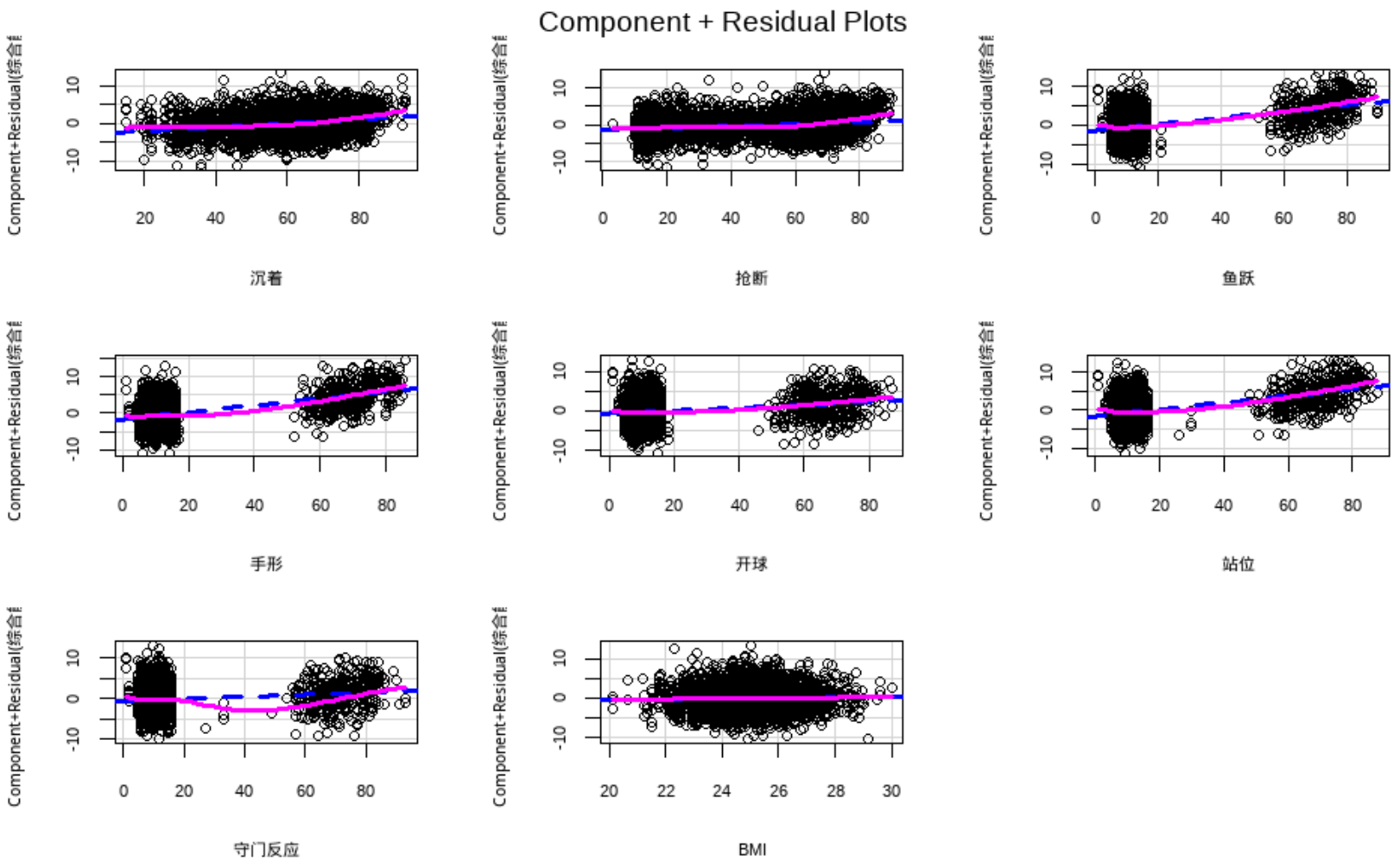
# Normality, Independence, Linearity, Homoscedasticity



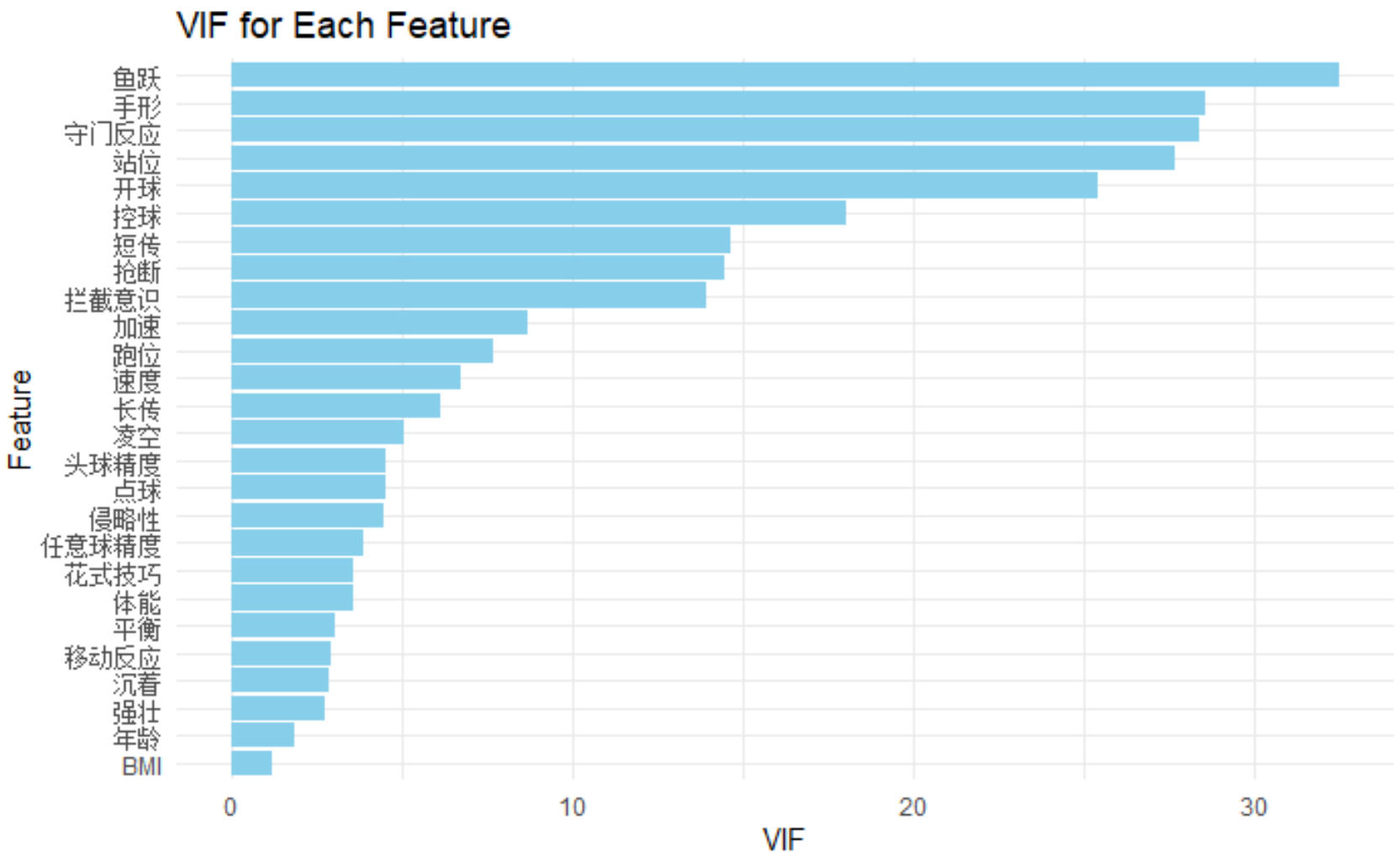
# Normality, Independence, Linearity, Homoscedasticity



# Normality, Independence, Linearity, Homoscedasticity



# Multicollinearity



# Multicollinearity

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.447770   1.195069  26.315  < 2e-16 ***
年龄         -0.046299   0.016057  -2.883 0.003955 **
花式技巧      0.058224   0.124657   0.467 0.640475
头球精度     -0.029205   0.005262  -5.550 3.04e-08 ***
凌空          0.025643   0.006689   3.834 0.000128 ***
任意球精度    0.015030   0.005961   2.522 0.011721 *
长传          0.053210   0.006002   8.865  < 2e-16 ***
加速          0.035937   0.011987   2.998 0.002734 **
速度          0.041663   0.010734   3.881 0.000106 ***
移动反应      0.449739   0.008559  52.546  < 2e-16 ***
平衡         -0.028778   0.007272  -3.958 7.70e-05 ***
体能          0.010974   0.006981   1.572 0.116059
强壮          0.063416   0.007847   8.082 8.40e-16 ***
侵略性       -0.023459   0.005670  -4.138 3.58e-05 ***
跑位         -0.058530   0.007007  -8.353  < 2e-16 ***
点球         -0.019199   0.007189  -2.671 0.007598 **
沉着          0.063479   0.007291   8.707  < 2e-16 ***
BMI           0.056741   0.045580   1.245 0.213257
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.476 on 3943 degrees of freedom
Multiple R-squared:  0.6339,    Adjusted R-squared:  0.6324
F-statistic: 401.7 on 17 and 3943 DF,  p-value: < 2.2e-16
```

Figure: Model_reduced

# Multicollinearity

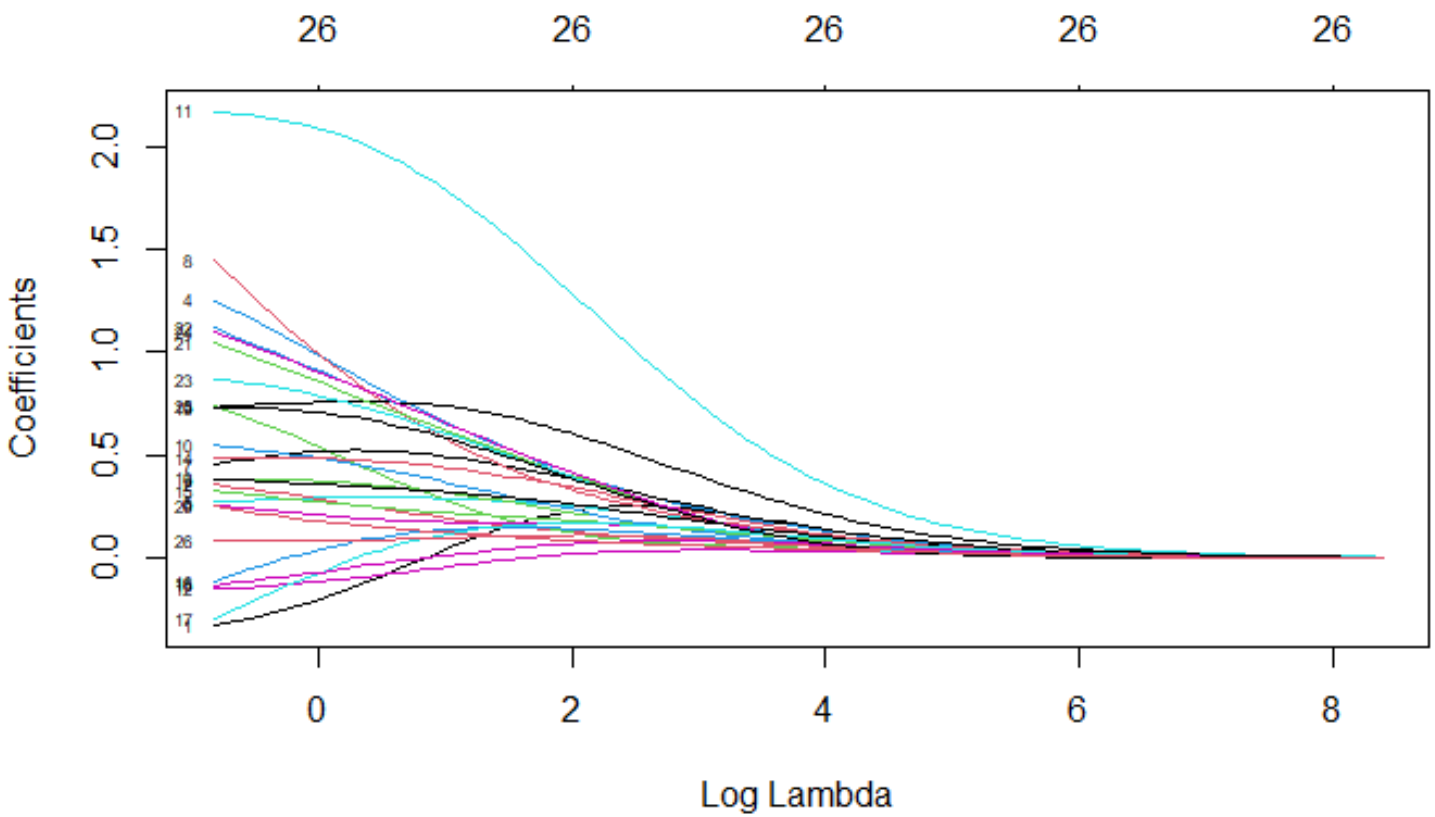



Figure: Ridge Regression

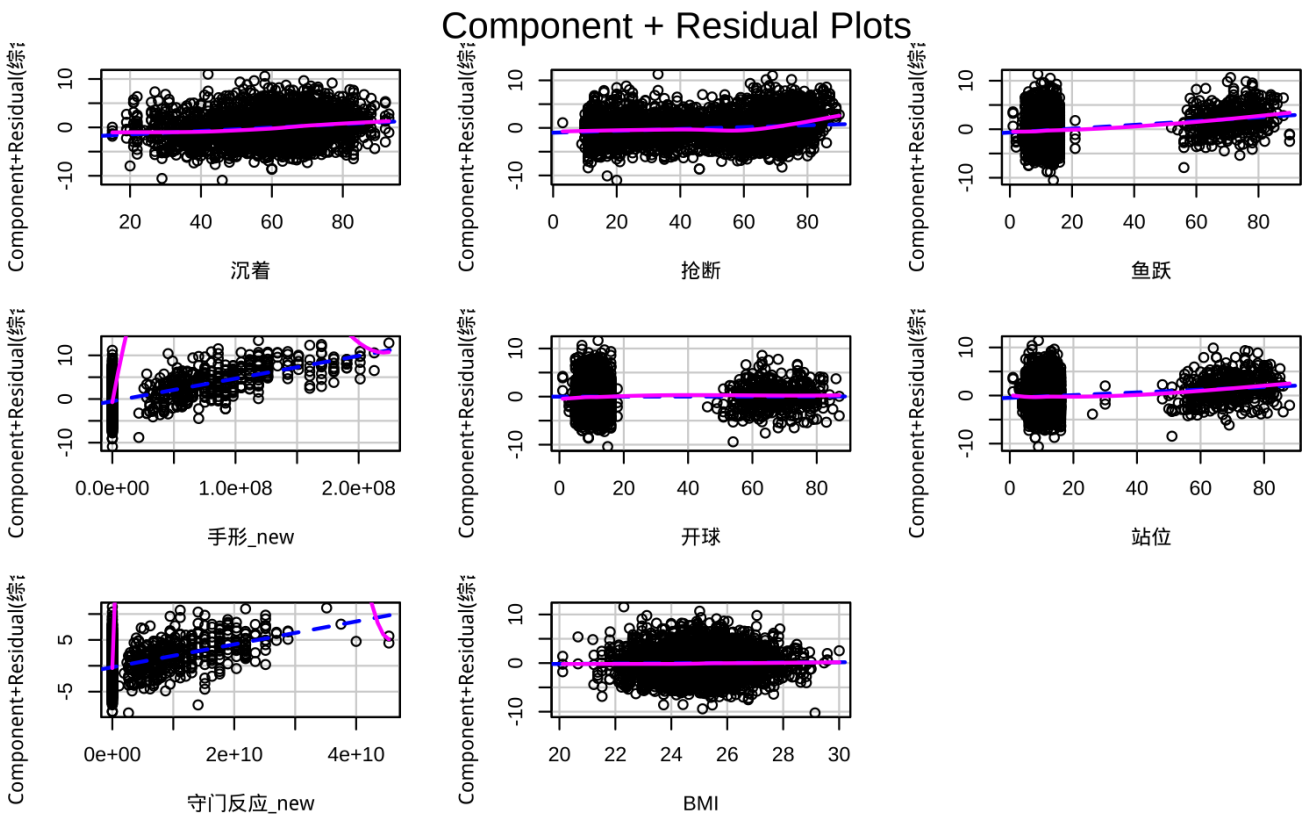
# model2 and model3, by boxTidwell

6 variables are transformed.

```
          MLE of lambda Score Statistic (t)  Pr(>|t|)
头球精度         3.8343                1.9107   0.05611 .
短传             2.5087                4.5316 6.027e-06 ***
移动反应         1.0156                5.9078 3.759e-09 ***
控球             5.2311                6.7880 1.306e-11 ***
守门反应         5.8016                4.6637 3.209e-06 ***
手形             4.6627               10.4049 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

iterations =  26

Score test for null hypothesis that all lambdas = 1:
F = 78.502, df = 6 and 3948, Pr(>F) = < 2.2e-16
```

Figure: boxTidwell

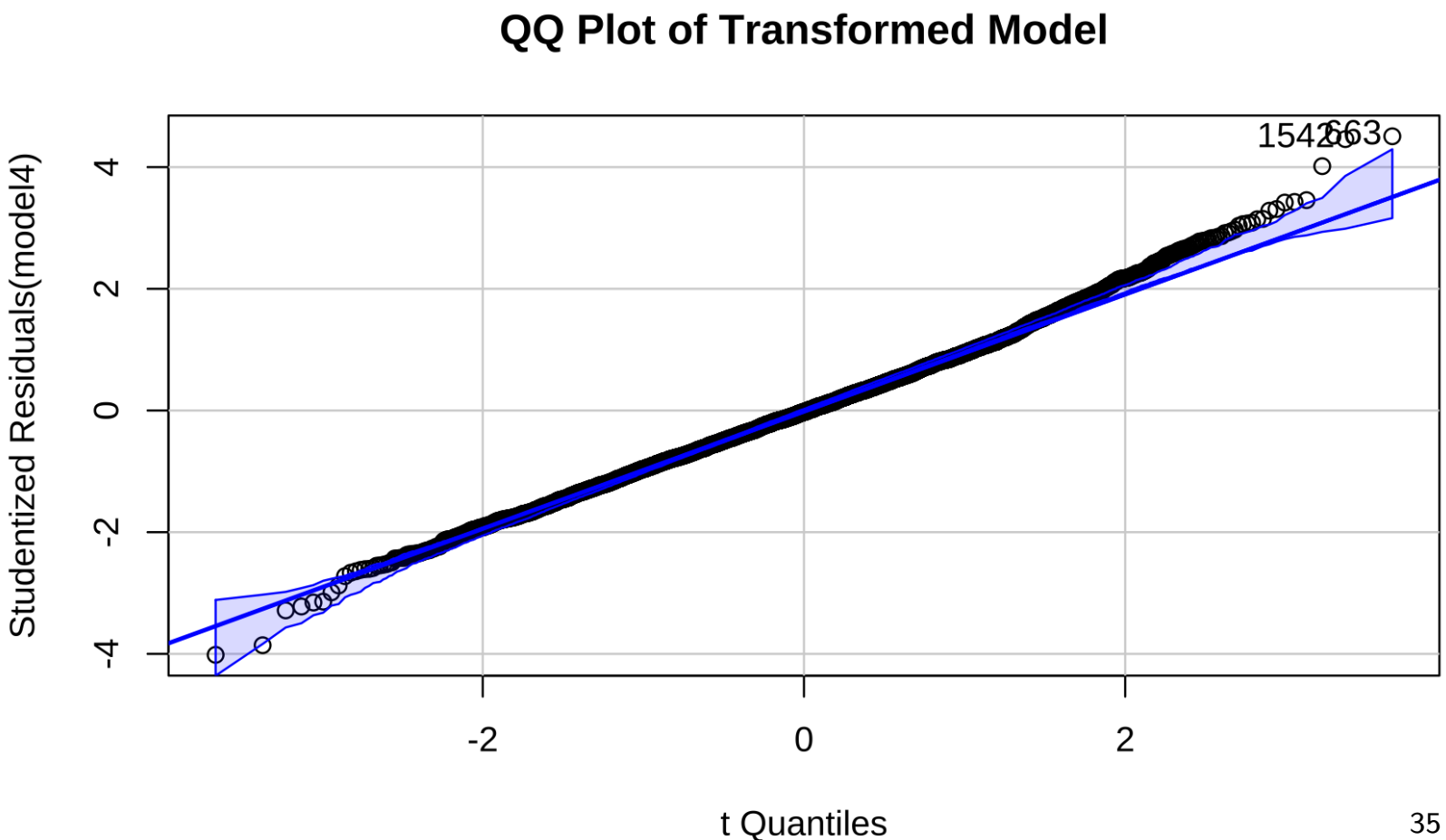
# model2 and model3, by `boxTidwell`




model_3: 2 variables (short passes and moving reaction)are transformed

# model4 by powerTransform

$\lambda = 0.5$



# model5 by boxcox

- $\lambda_5 = 0.3434$

.

# model5 by `boxcox`

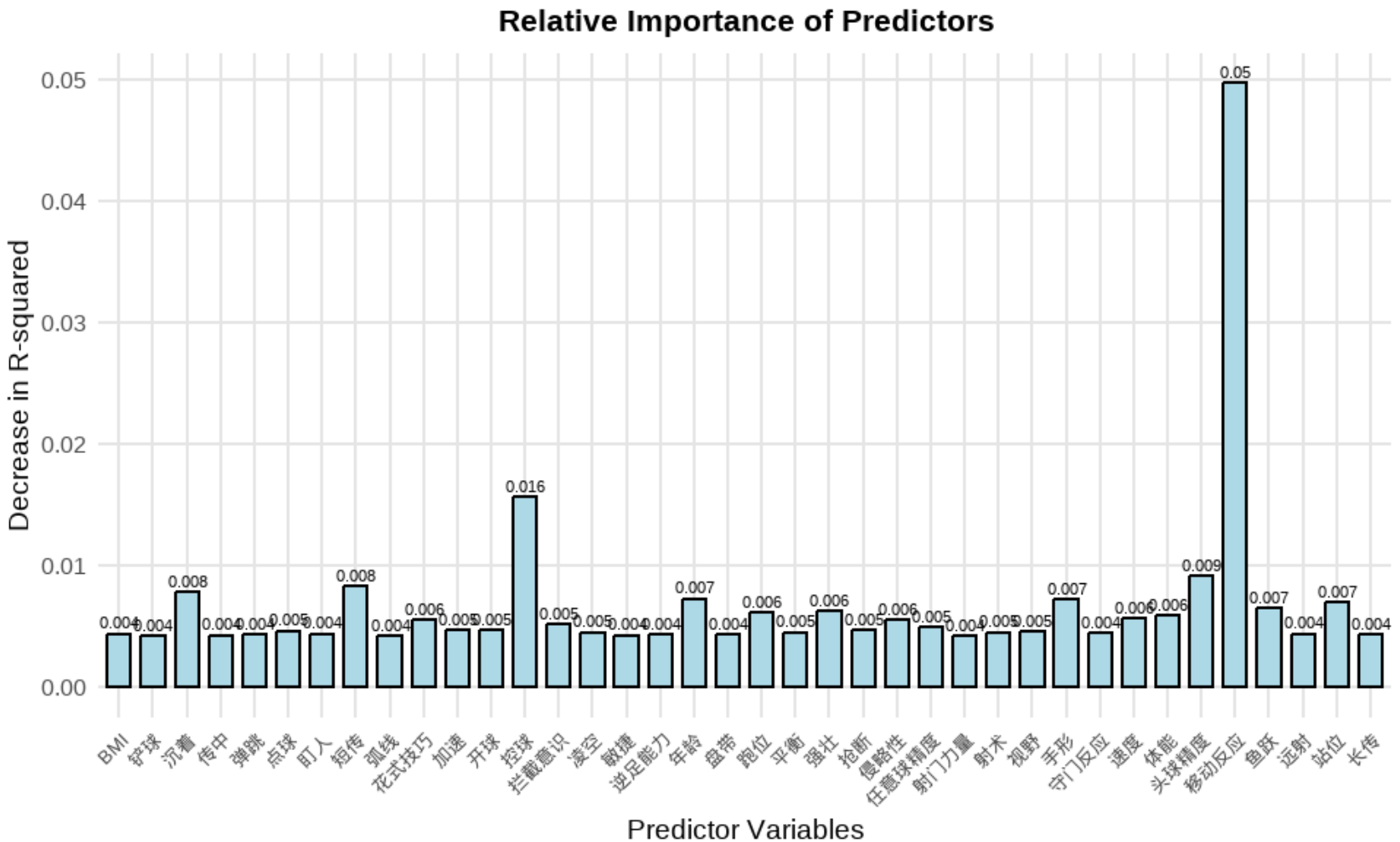
- $\lambda_5 = 0.3434$
- Non-constant Variance Score Test
  Variance formula: `fitted.values`
  Chisquare = **72.35854**, Df = **1**, p = **2.22e-16**

.

# Model Comparison

| Model | CV RMSE | $R^2$ |
|---|---|---|
| model_0 | 2.984079 | 0.7331283 |
| model_back | 2.989138 | 0.7319983 |
| model2 | 2.989289 | 0.7822480 |
| model3 | 2.990642 | 0.7434827 |
| model4 | 2.990967 | 0.7426850 |
| model5 | 2.998563 | 0.7422501 |

# Relative Importance



# Influential Observations

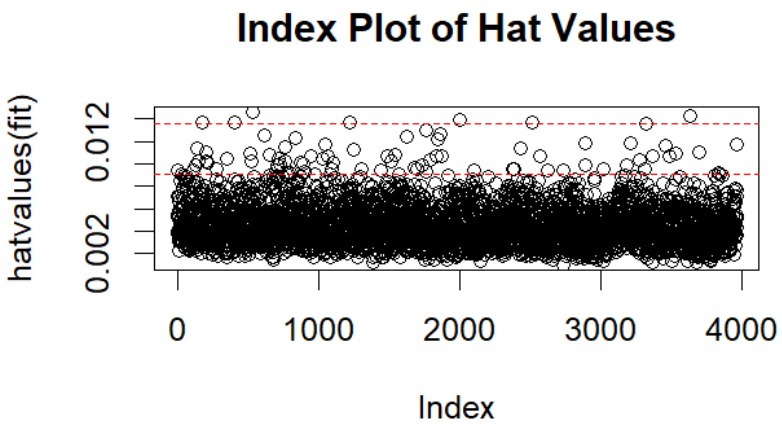



Figure: Hat values

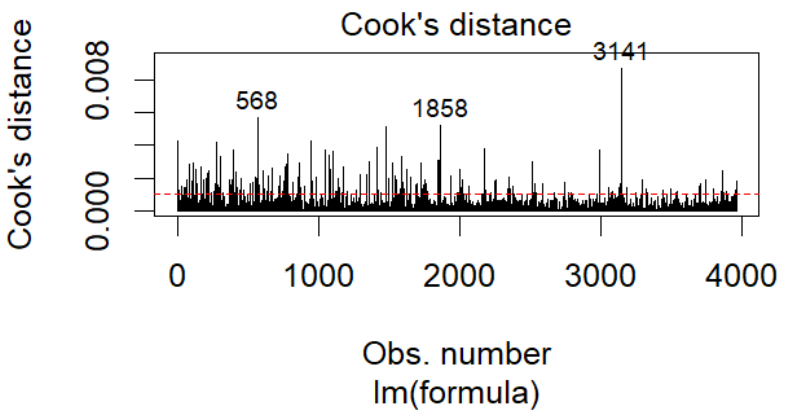



Figure: Cook's distance

# Outline

Since the assumption of normality is violated, we use bootstrap to calculate some statistics and generate confidence intervals.

- We set the repeat times as 1000 and perform bootstrap simulation to gain statistics like $R^2$ and linear coefficients.

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = df, statistic = rsq, R = 1000, formula = model_formula)


Bootstrap Statistics :
     original      bias    std. error
t1* 0.7319983 0.001826182  0.00785419
```

Figure: Bootstrap for $R^2$

| | Original_Stat <dbl> | Bias <dbl> | Std_Error <dbl> |
|---|---|---|---|
| (Intercept) | 17.48793209 | 0 | 1.129379812 |
| 年龄 | -0.09515048 | 0 | 0.014479269 |
| 花式技巧 | 0.50633994 | 0 | 0.109022096 |
| 头球精度 | 0.05547310 | 0 | 0.005619050 |
| 短传 | 0.09882375 | 0 | 0.013546714 |
| 凌空 | 0.01379898 | 0 | 0.005914863 |
| 任意球精度 | 0.01796274 | 0 | 0.004935751 |
| 长传 | 0.01136592 | 0 | 0.007539770 |
| 控球 | 0.15985436 | 0 | 0.012344618 |
| 加速 | 0.02463873 | 0 | 0.011107542 |

Figure: Bootstrap for Linear Coefficients

The bias is small enough to ignore.

# Outline

# Data Preparation

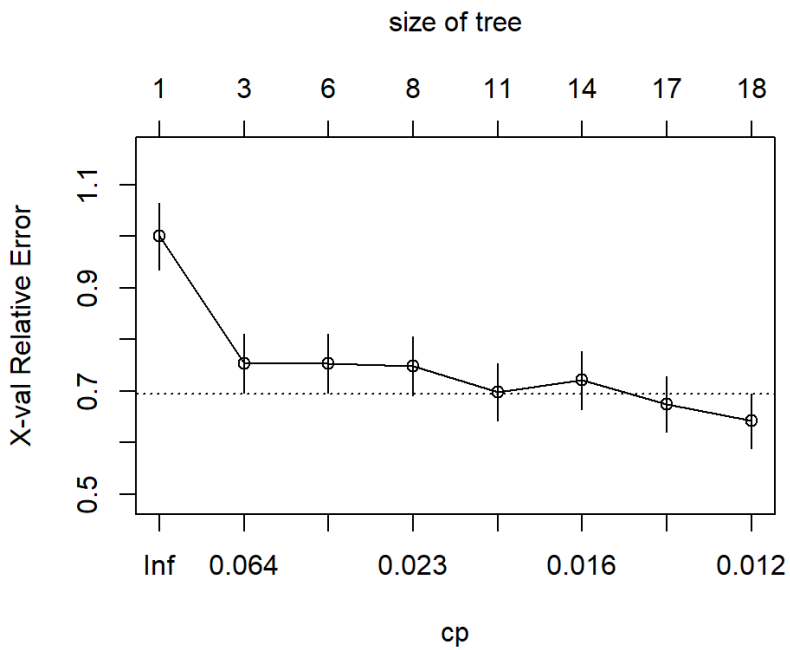
- Removed irrelevant columns (e.g., player names).

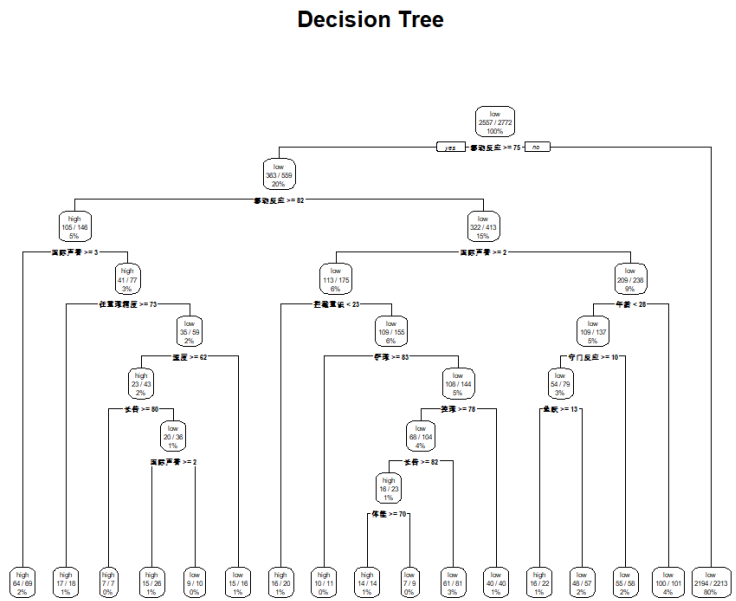
# Data Preparation

- Removed irrelevant columns (e.g., player names).
- Binary classification target created (scores bigger than 76).

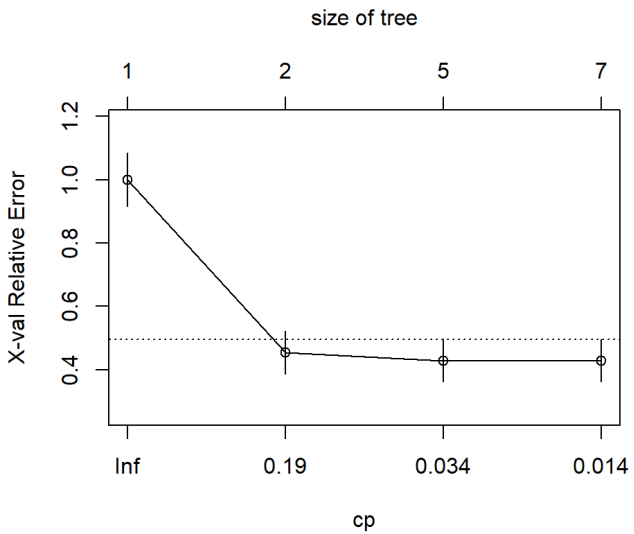
# Classical Decision Tree - Overall

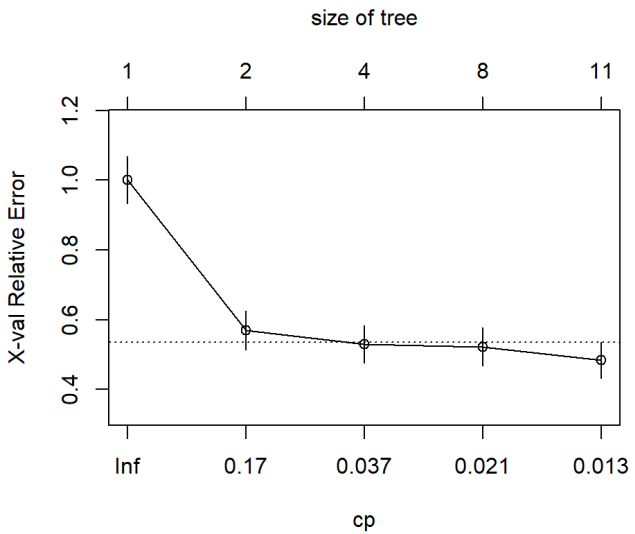
The complexity parameter table was analyzed. Selected a tree with 16 splits based on cross-validation error.



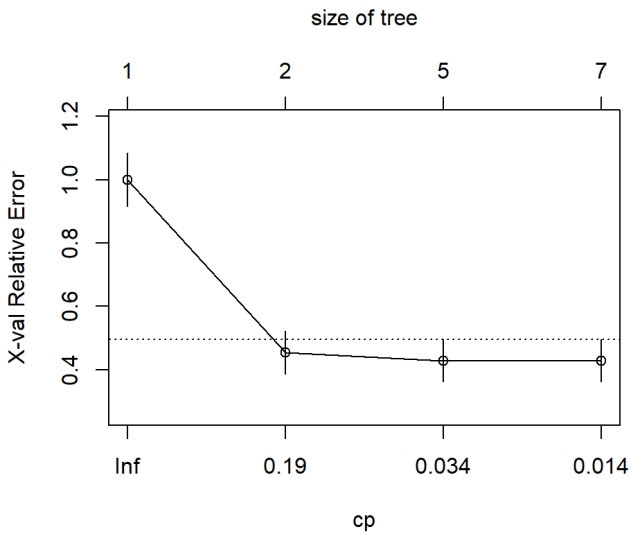



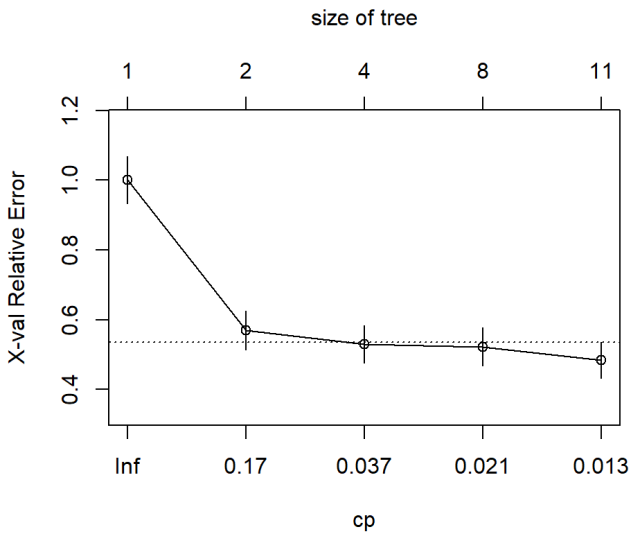
# Classical Decision Tree

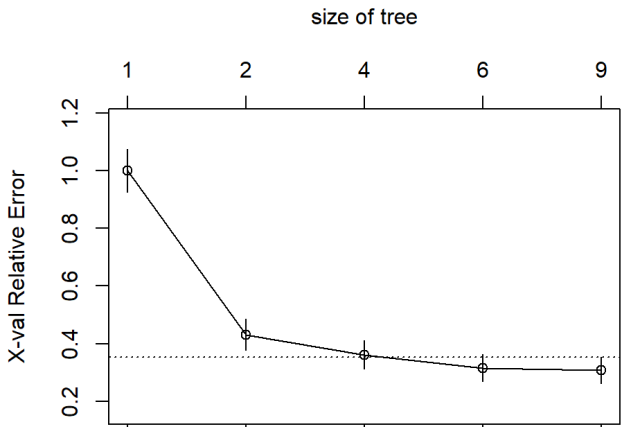


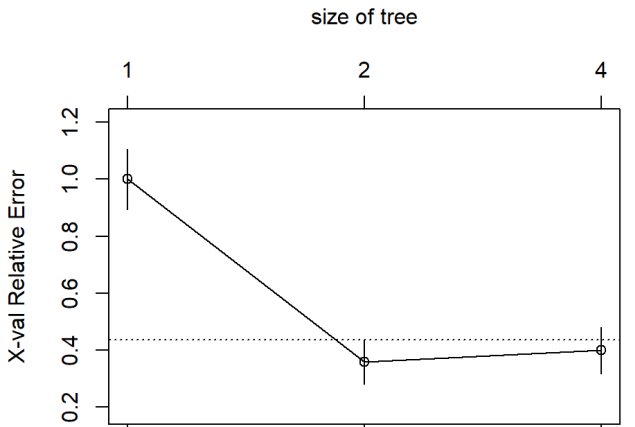



# Classical Decision Tree

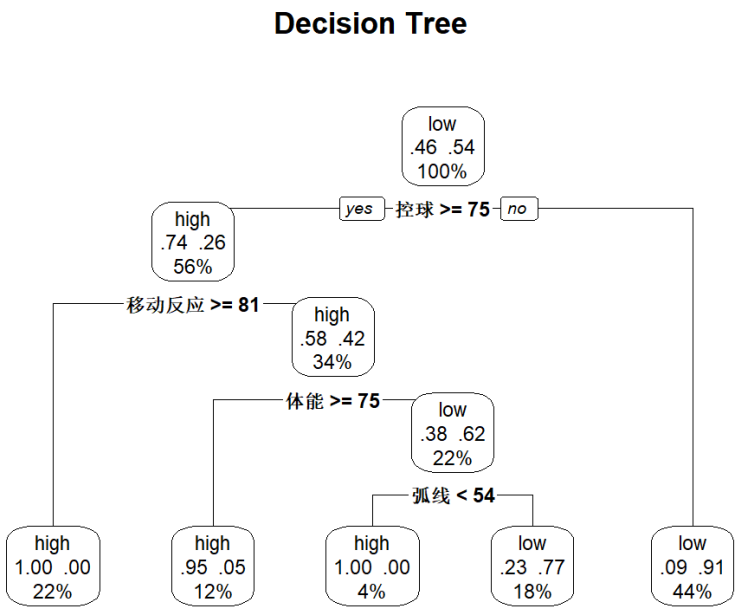


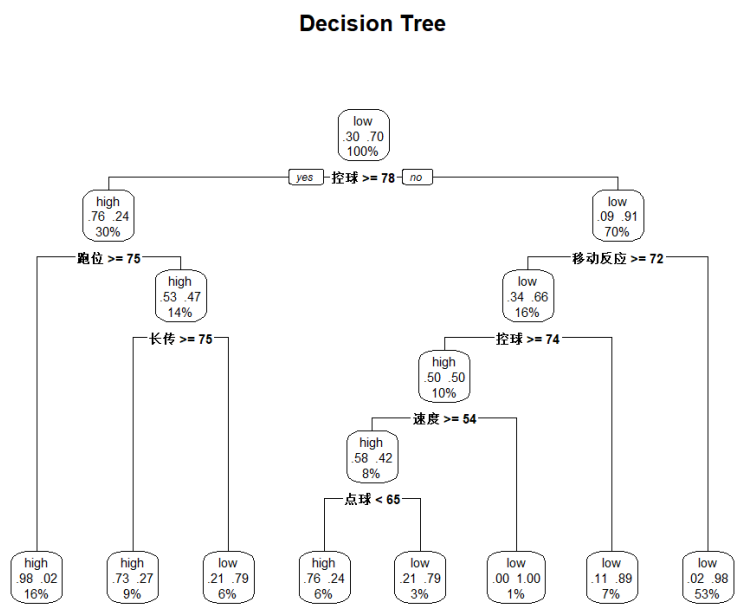



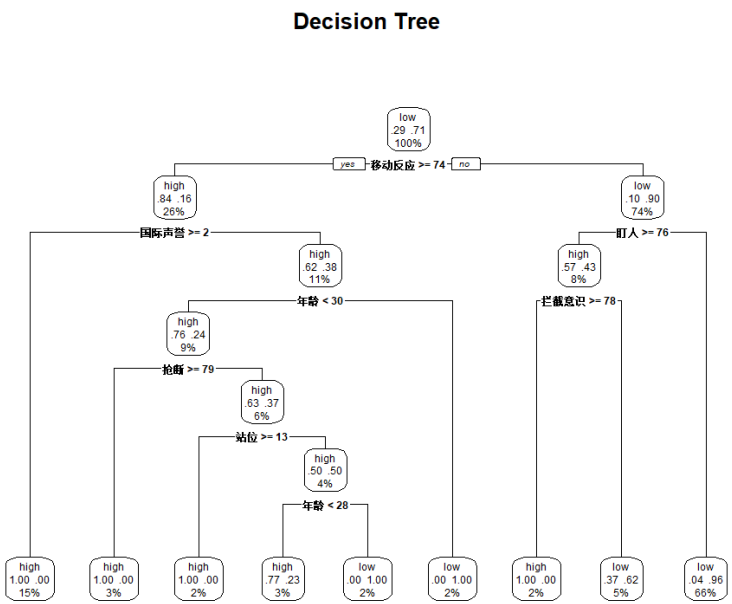



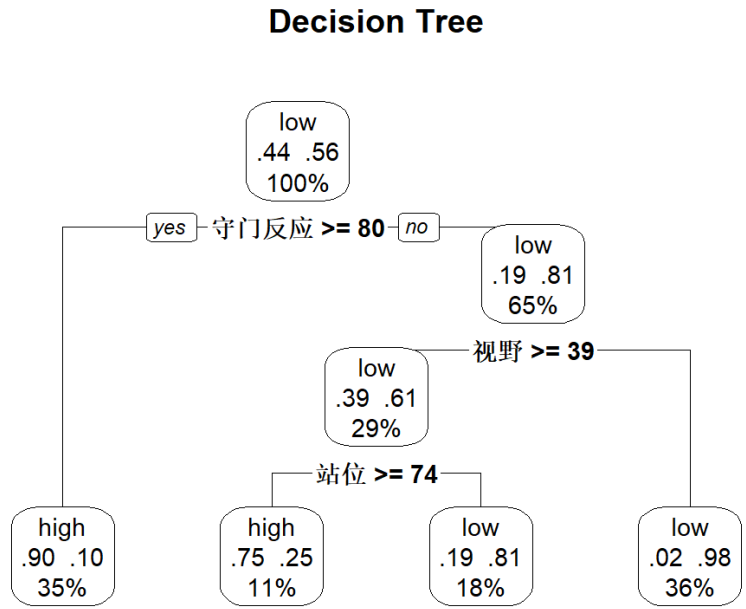
# Classical Decision Tree - Front Mid



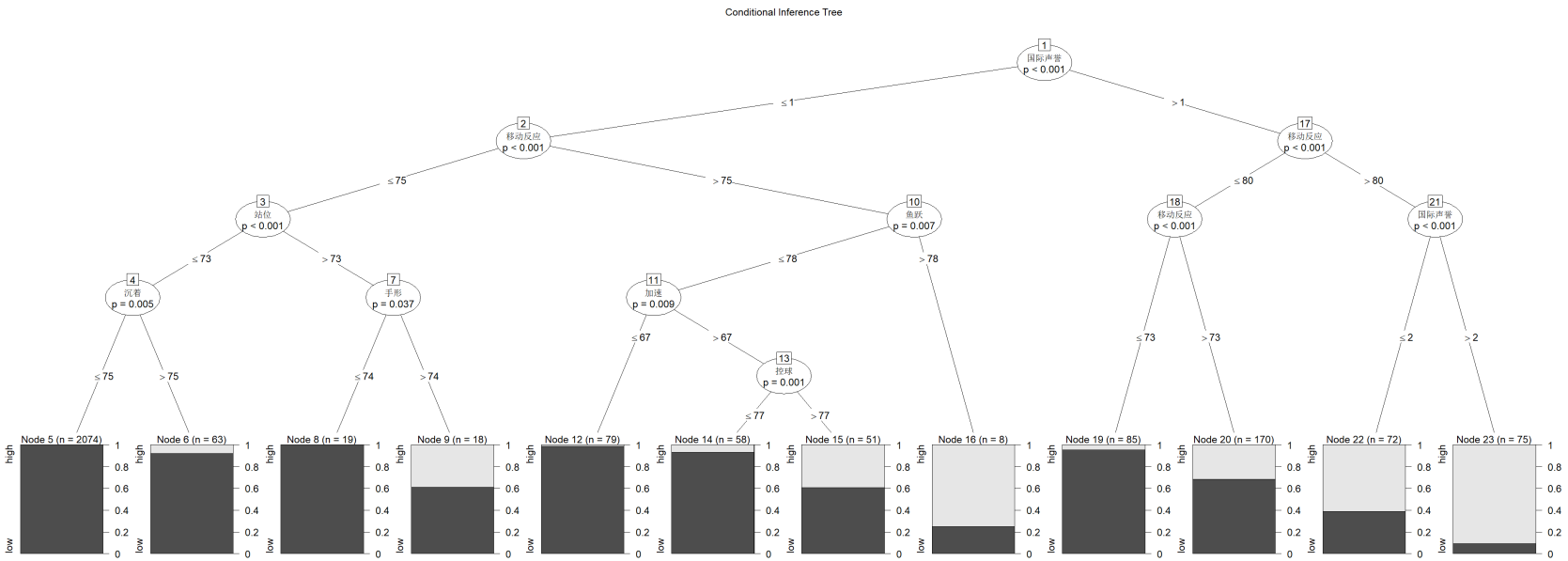



# Classical Decision Tree - Back Goalkeeper

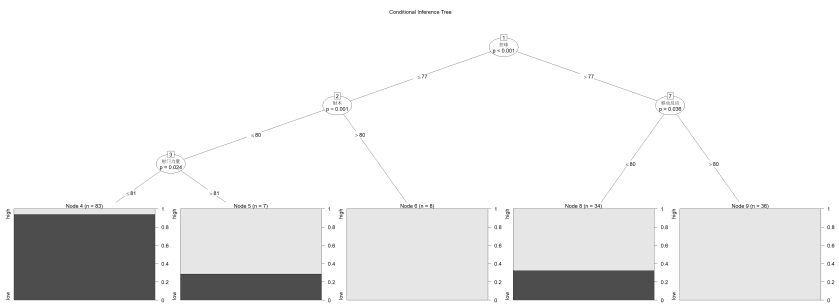


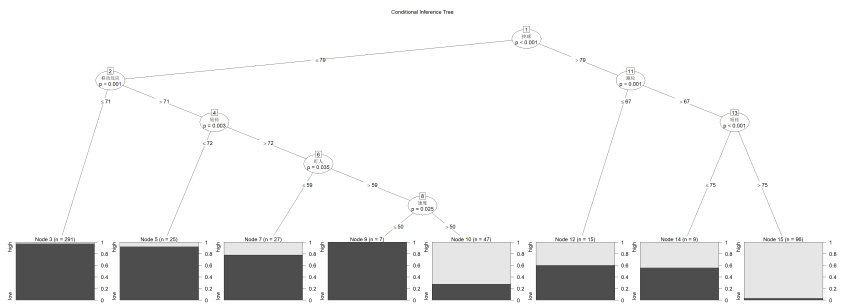



# Conditional Inference Trees - Overall

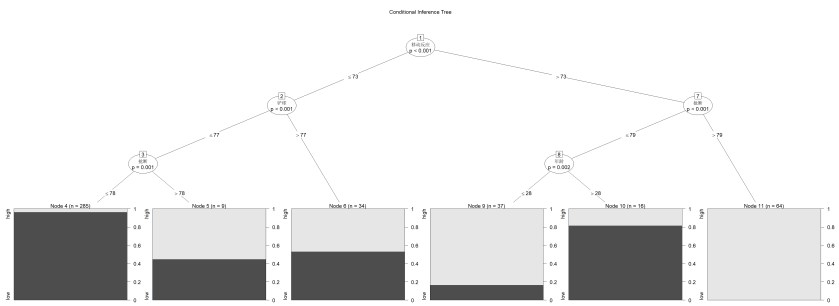


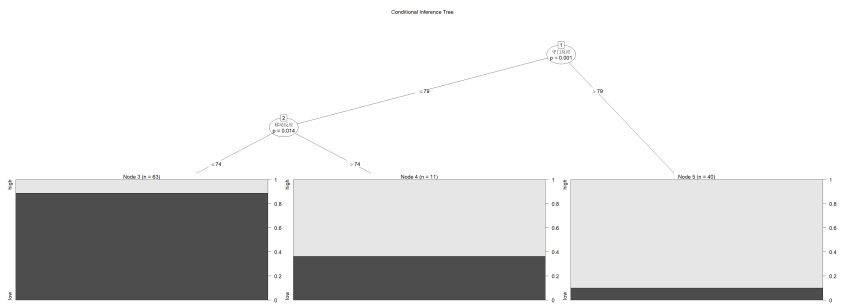
# Conditional Inference Trees - Front Mid



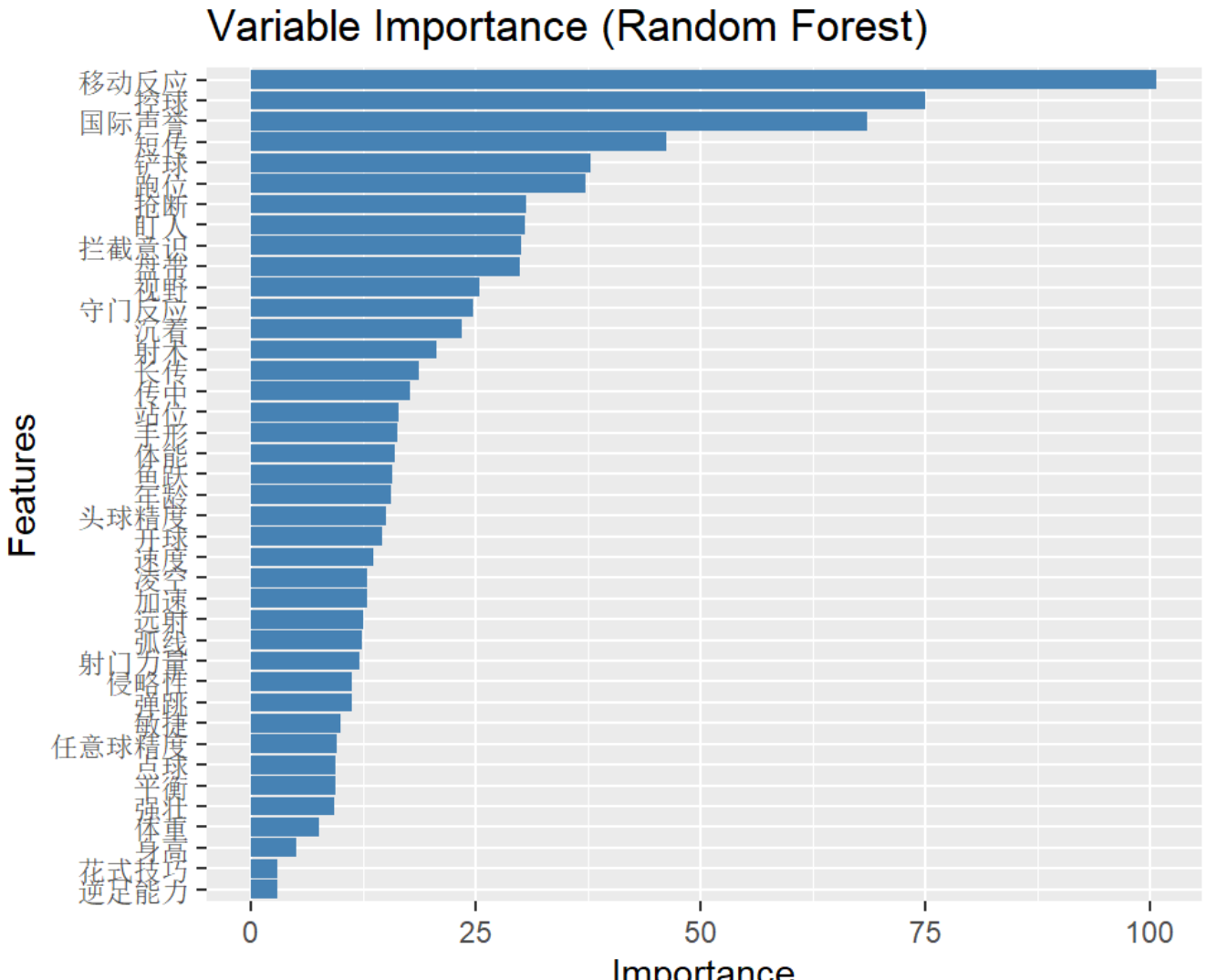


# Conditional Inference Trees - Back Goalkeeper

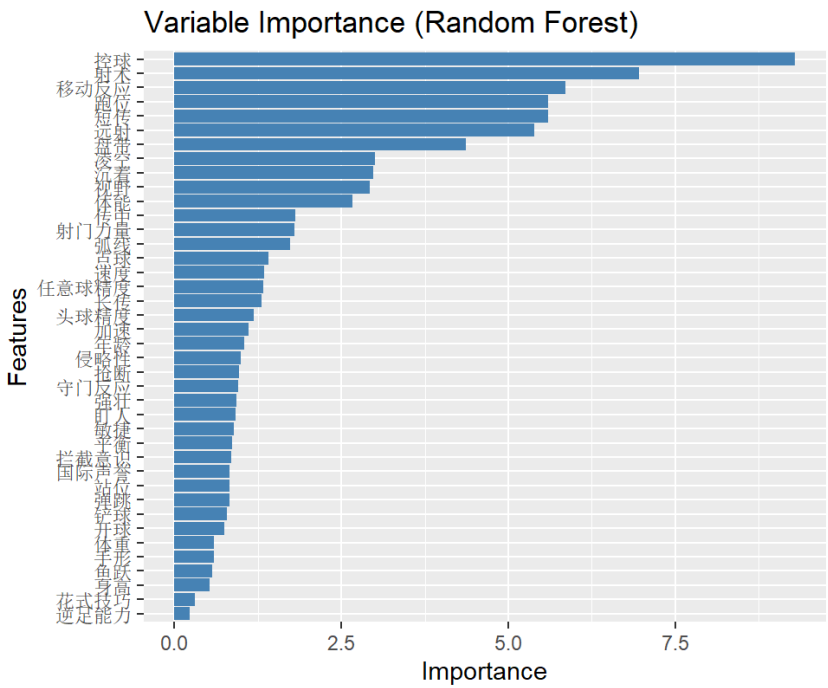


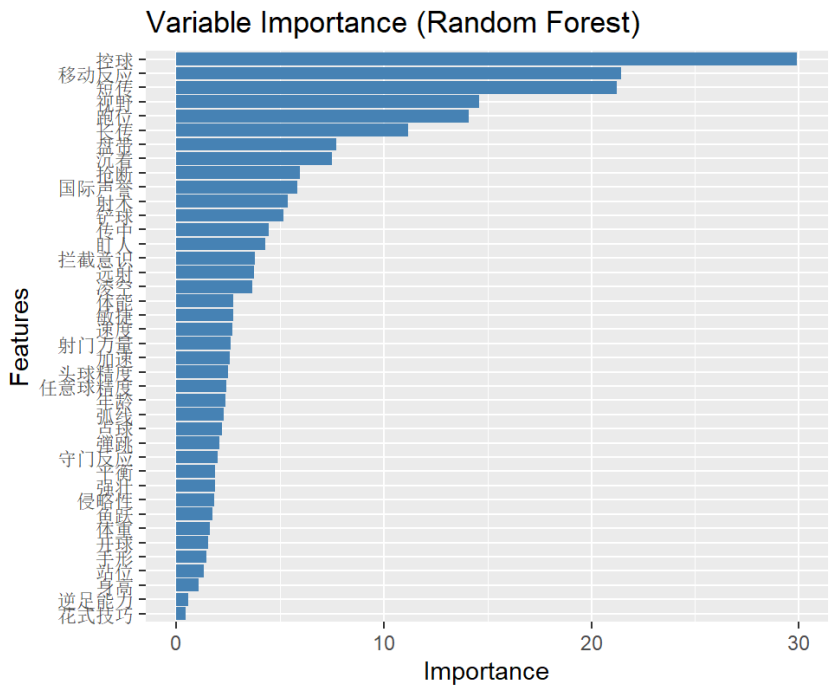


# Importance - Overall

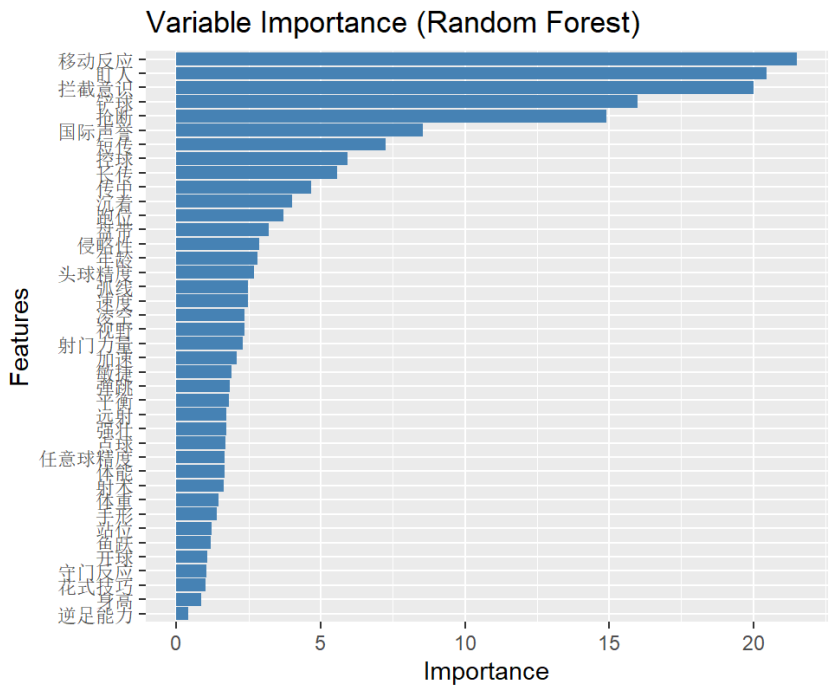


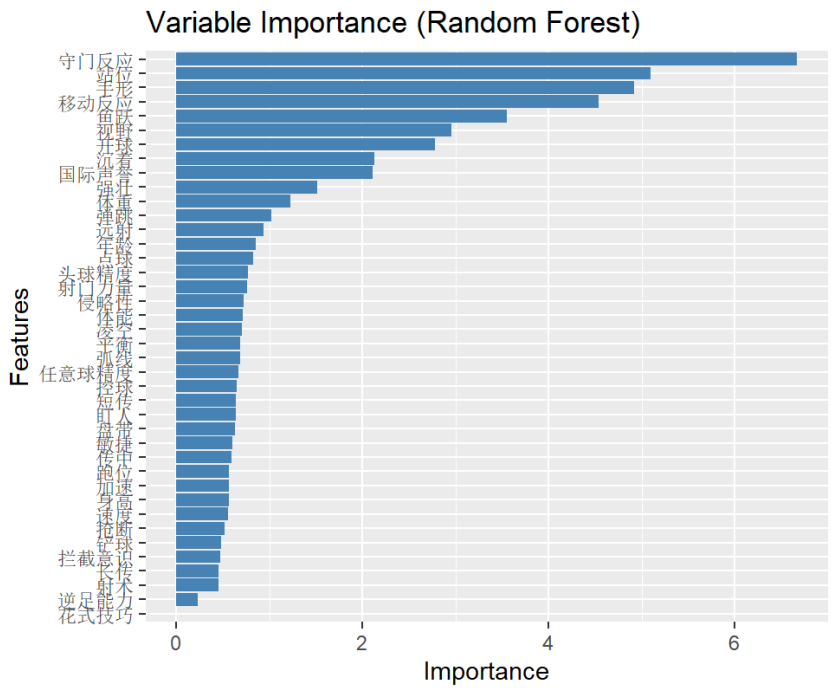
# Importance - Front Mid



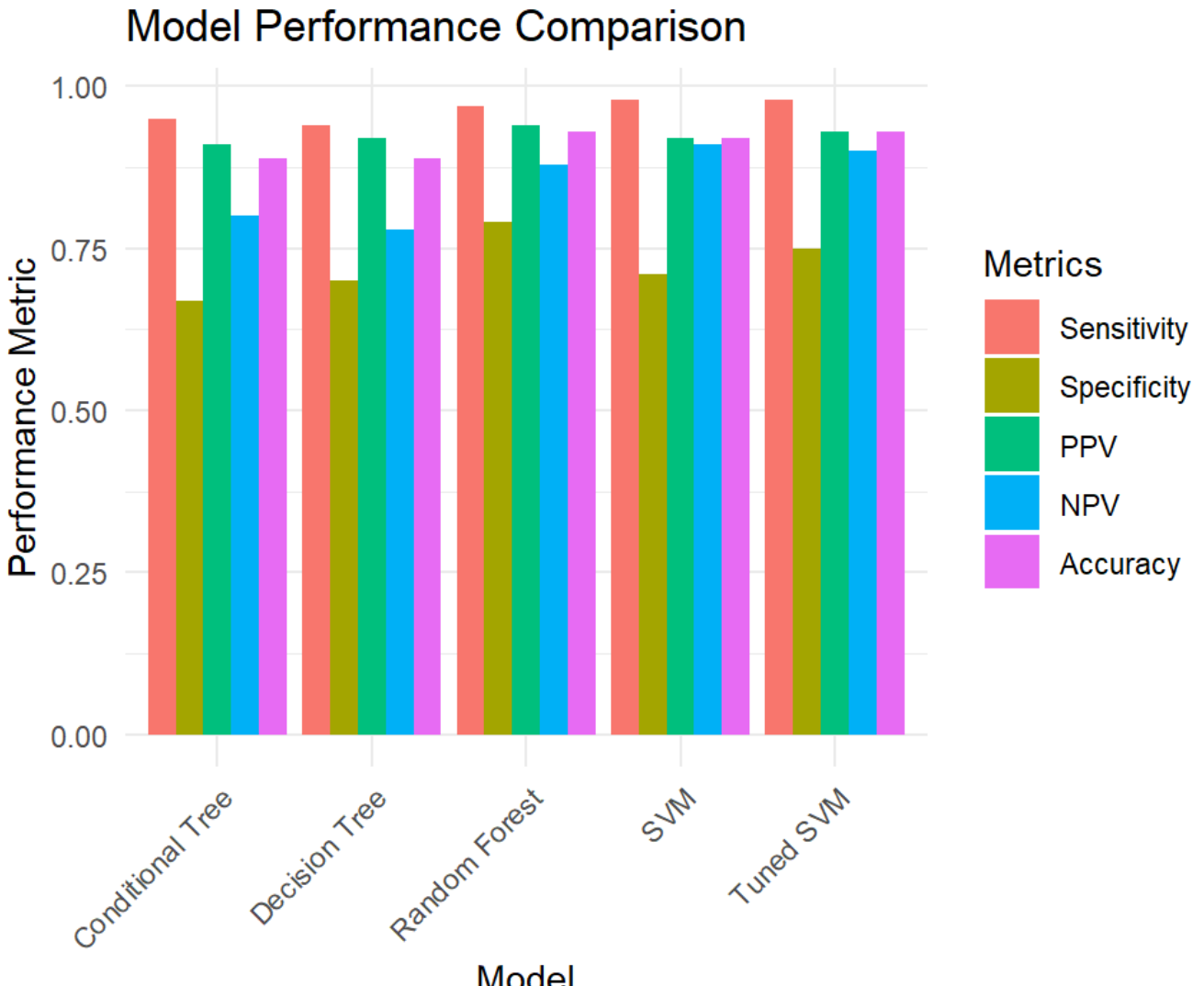



# Importance - Back Goalkeeper

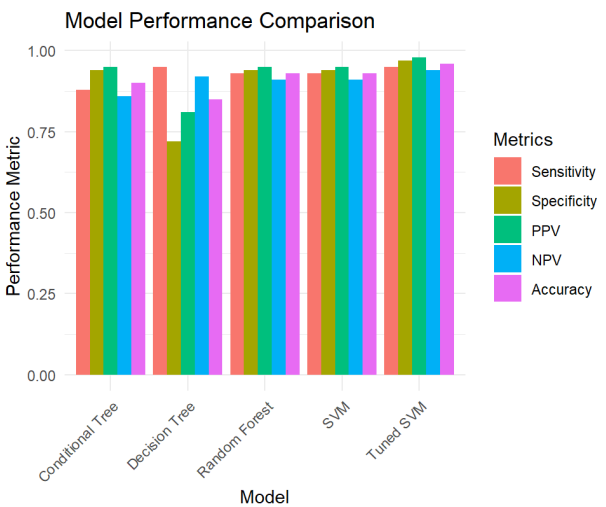


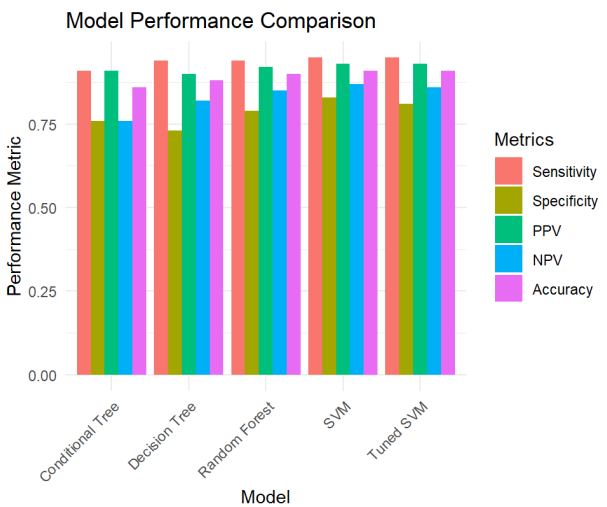



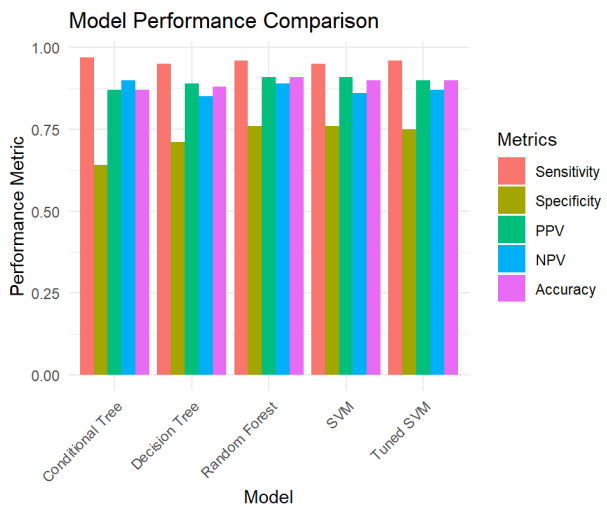
# Model Choose - Overall



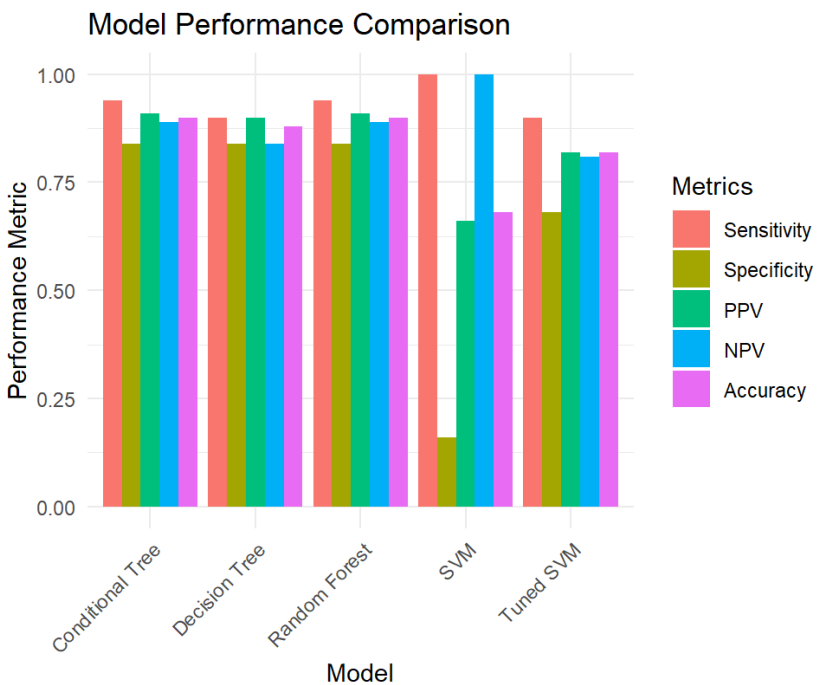
# Model Choose - Front Mid Back



# Model Choose - Goalkeeper




# Thank you

Thank you for listening!