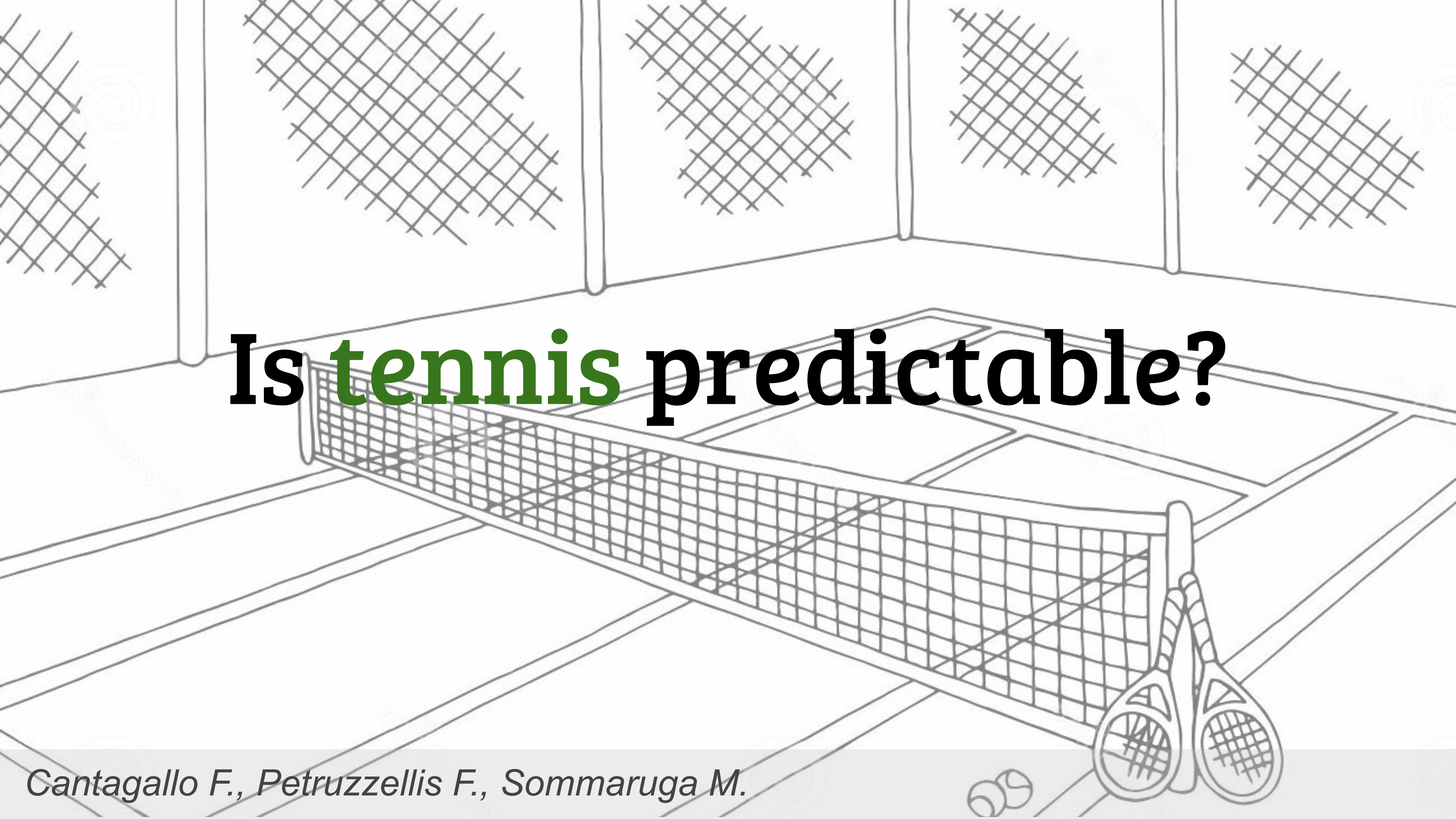
Is tennis predictable?



Cantagallo F., Petruzzellis F., Sommaruga M.

Overview

- Dataset introduction
- EDA
    - Features distribution
    - PCA
- Modeling Total Points Won with Multiple Linear Regression
- Modeling Match Result with Logistic Regression & k-NN
- Interpretation of results: is tennis predictable?
- Technical Appendix

Tennis Major Tournaments Dataset

- The data were downloaded from UCI Machine Learning repository
- The dataset was originally composed of tables with the same structure containing single tournament's statistics, also divided by gender. We merged these matrices in a single dataset.
- The dataset has originally 42 attributes and 943 instances, describing male and female tennis matches (i.e. statistical units) which were played in 2013 in major world tournaments
- Each row contains information about the performance of both players

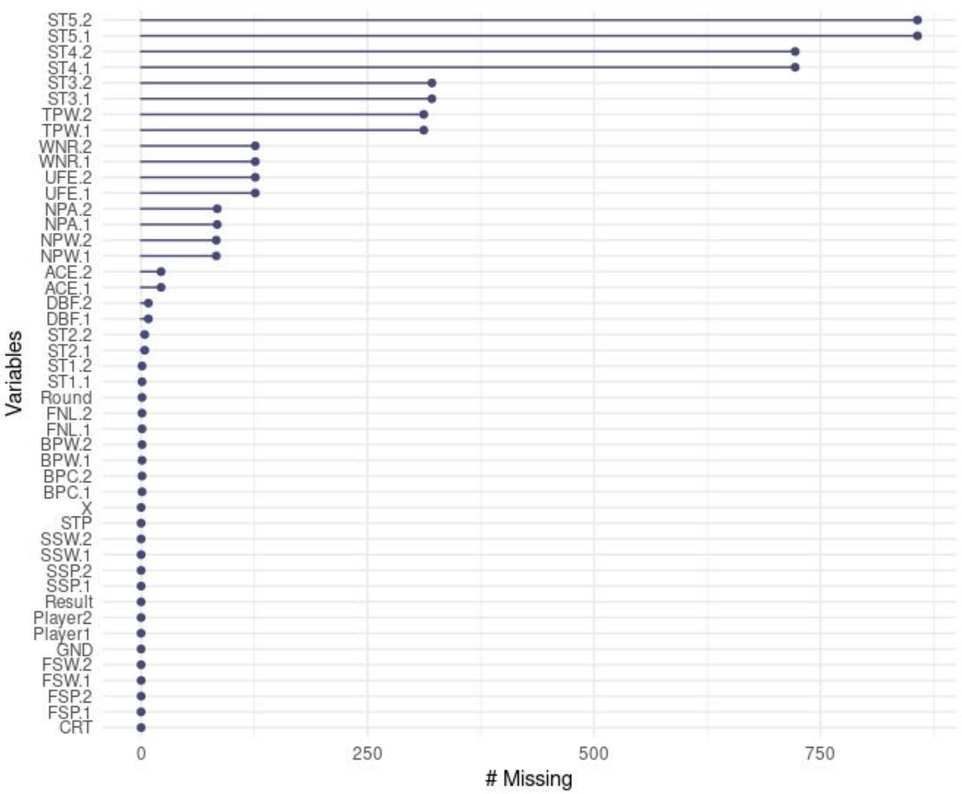
Tennis Major Tournaments Dataset

- The data were downloaded from UCI Machine Learning repository
- The dataset was originally composed of tables with the same structure containing single tournament's statistics, also divided by gender. We merged these matrices in a single dataset.
- The dataset has originally 42 attributes and 943 instances, describing male and female tennis matches (i.e. statistical units) which were played in 2013 in major world tournaments
- Each row contains information about the performance of both players

Match	Round	Result	FNL.1	FNL.2	FSP.1	FSW.1	SSW.1	ACE.1	DBF.1	WNR.1	UFE.1	BPC.1	BPW.1	NPA.1	NPW.1	TPW.1
Serena Williams/Ashleigh Barty	1	1	2	0	59	20	8	6	2	31	17	10	5	11	10	58
Heather Watson/Daniela Hantuchova	1	0	1	2	61	41	19	8	3	27	45	7	4	13	10	88
Samantha Stosur/Klara Zakopalova	1	1	2	0	65	28	11	6	1	19	18	10	7	10	7	74
Tsvetana Pironkova/Silvia Soler-Espinosa	1	1	2	0	62	28	12	5	0	30	21	5	3	7	4	68
Annika Beck/Petra Martic	1	1	2	0	67	18	8	0	0	8	10	11	6	3	3	52
Kiki Bertens/Ana Ivanovic	1	0	0	2	61	15	11	2	4	23	35	11	6	16	10	61

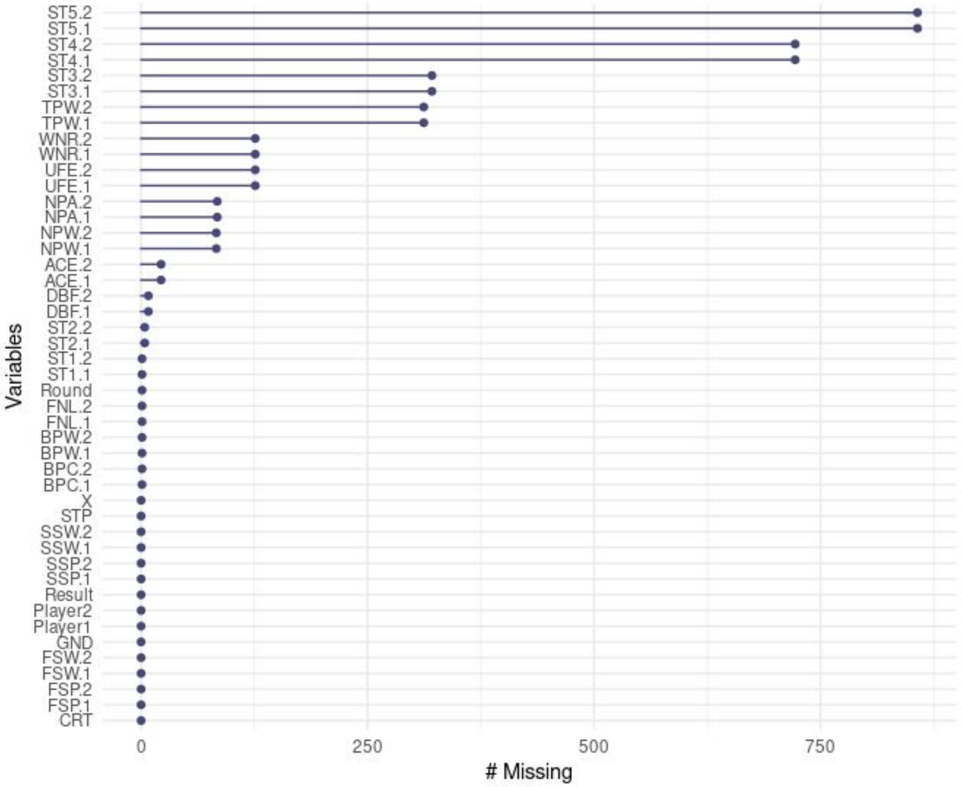
(only first player's variables are displayed)

Data cleaning and filtering



- We removed STx columns which contained many NAs by construction
- We then removed any row containing NAs for any other feature (we checked that the numerosity was still relevant).
- We manually added the features gender (GND), number of sets played (STP) and kind of court (CRT).

Data cleaning and filtering

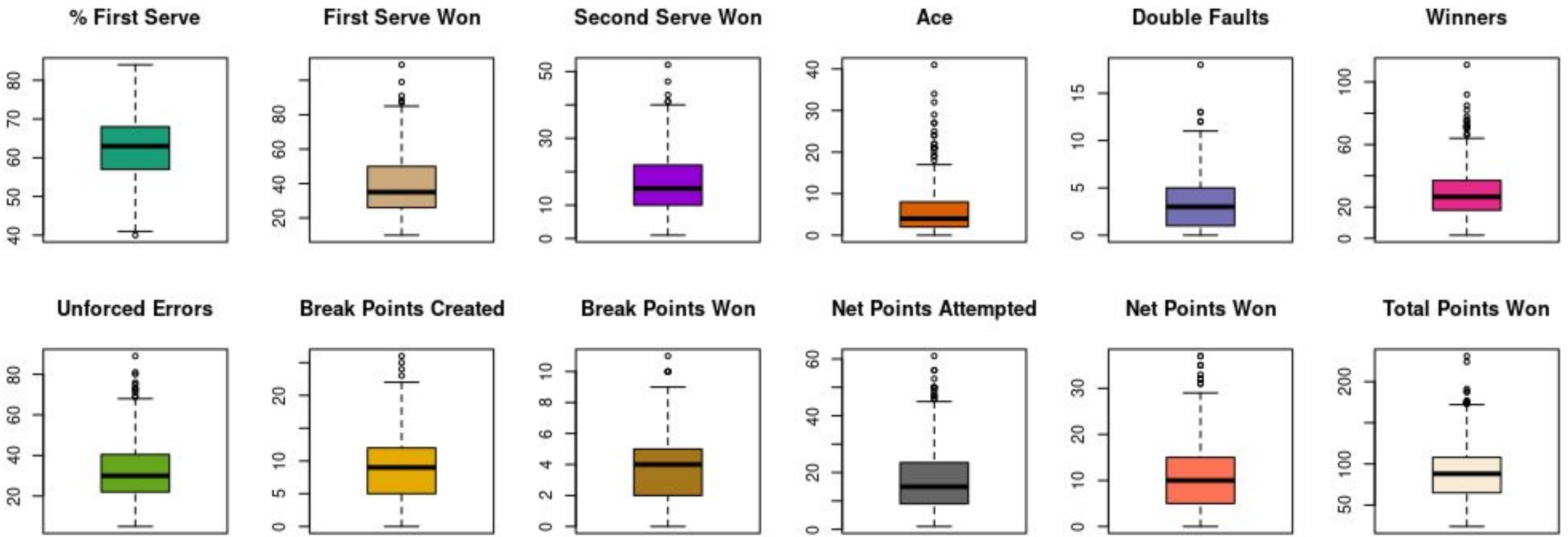


- We removed SSP features, which carried the same information as FSP.
- We also removed FNL.1 and FNL.2 variables which were not useful to us.
- Wimbledon data were missing TPW features, so they were excluded.
- Final dimensionality: 436 instances and 32 features, with balanced split w.r.t. the match winner.

Exploratory data analysis

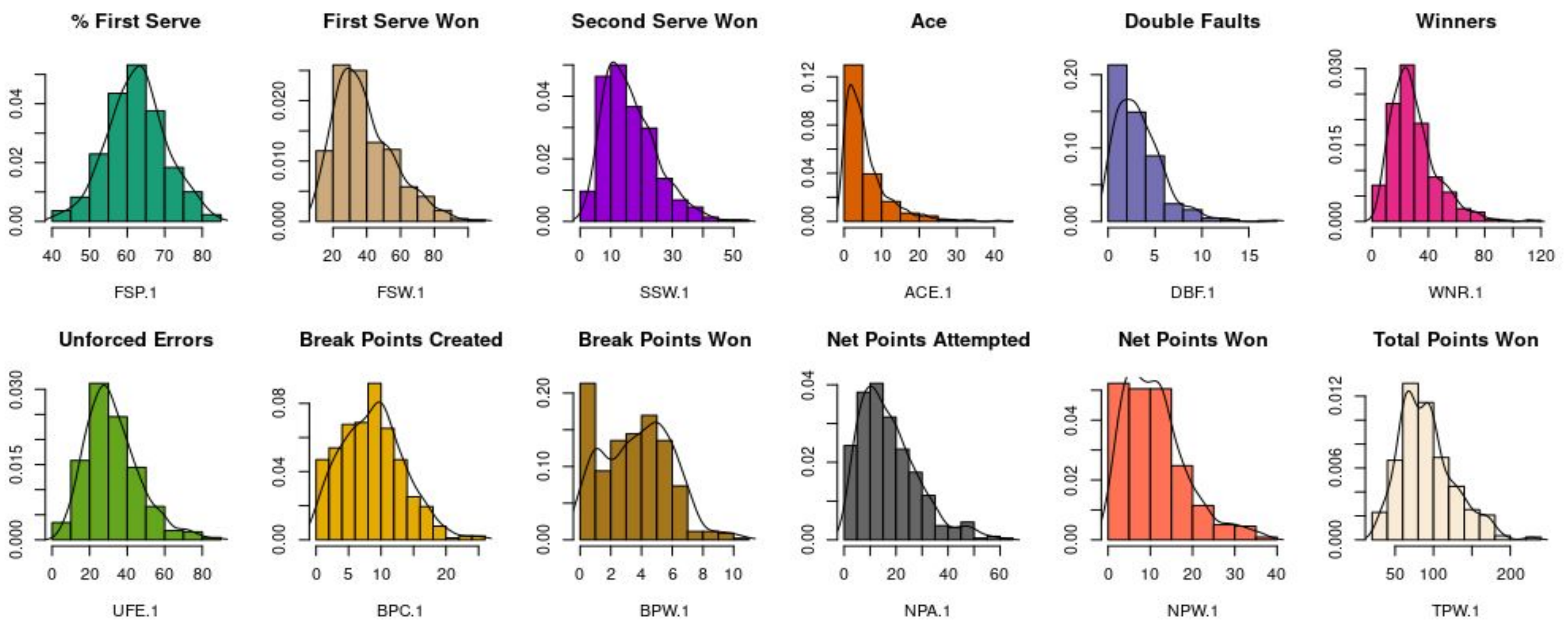
- Features distributions
- Assessing normality
- Principal Components Analysis

Features distributions: boxplots



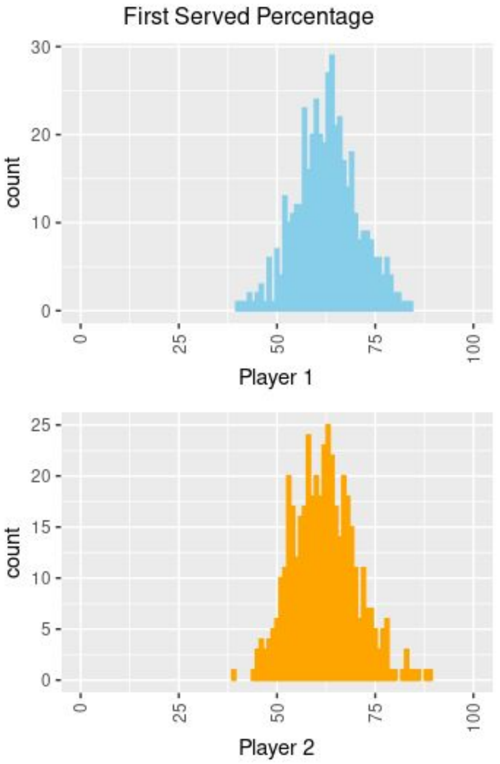
(here we show only player 1 features)

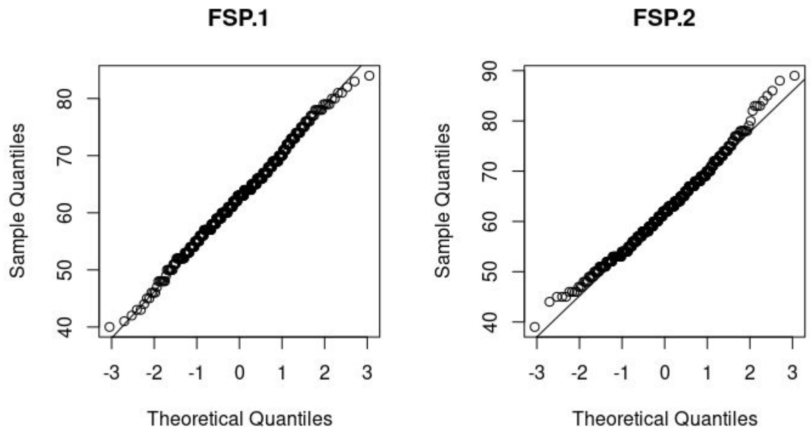
Features distributions: histograms



(here we show only player 1 features)

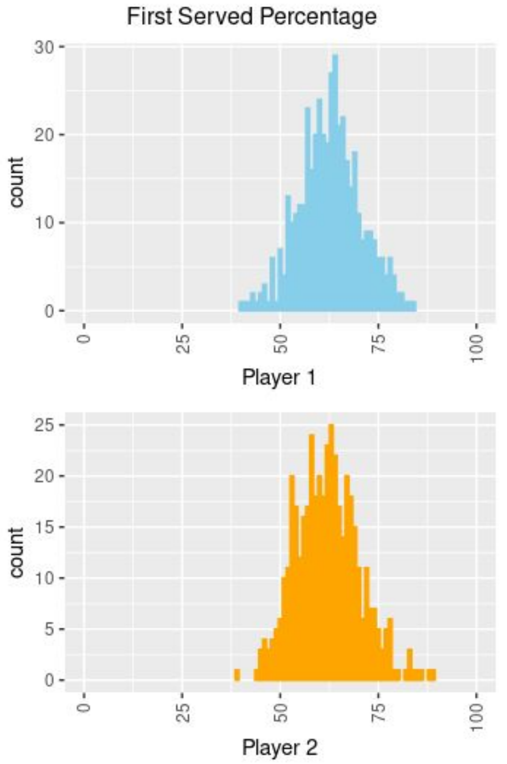
Exploring normality: FSP

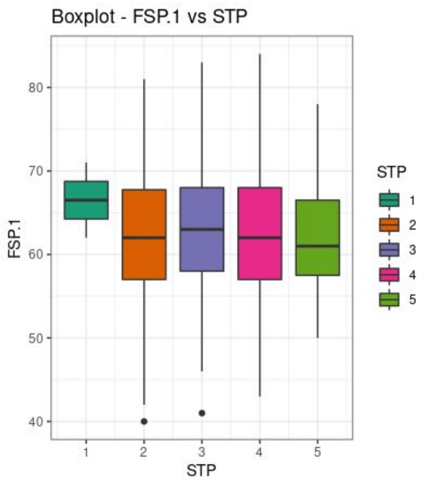


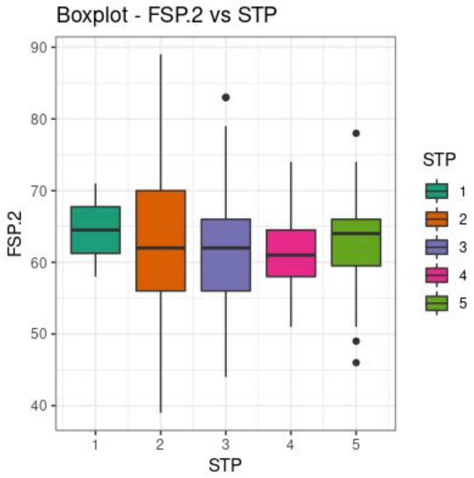


We can infer from both plots that First Served Percentage is normally distributed for both players

Exploring normality: FSP

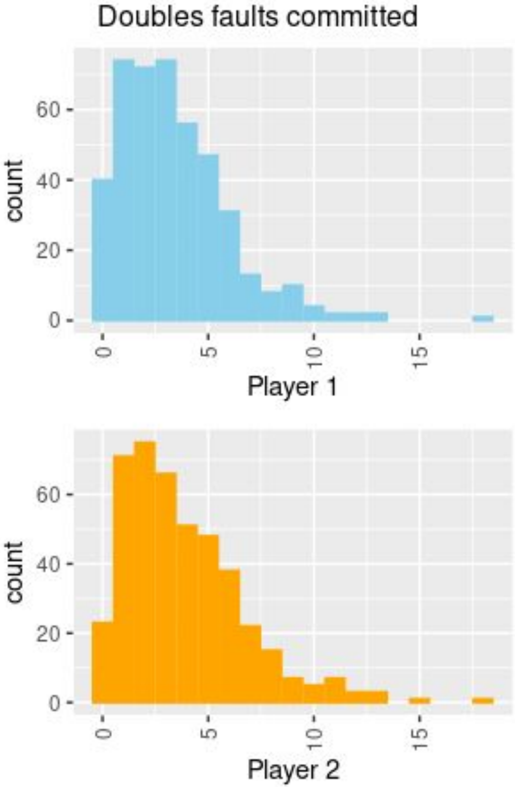


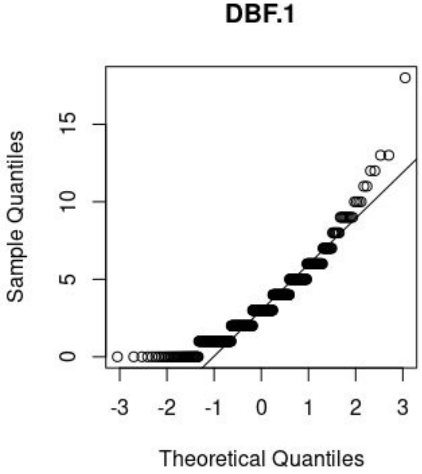


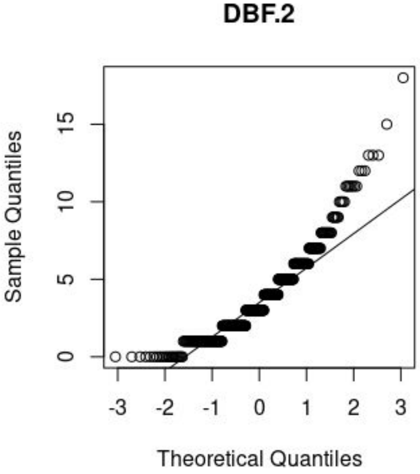


Also, we can see from the boxplot generated conditioning on the number of sets played that normality is stable across values of this feature

Exploring normality: DBF

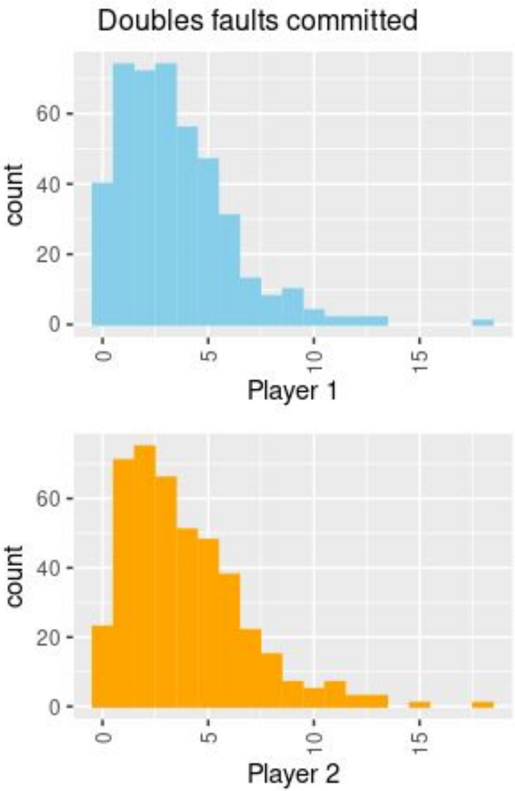


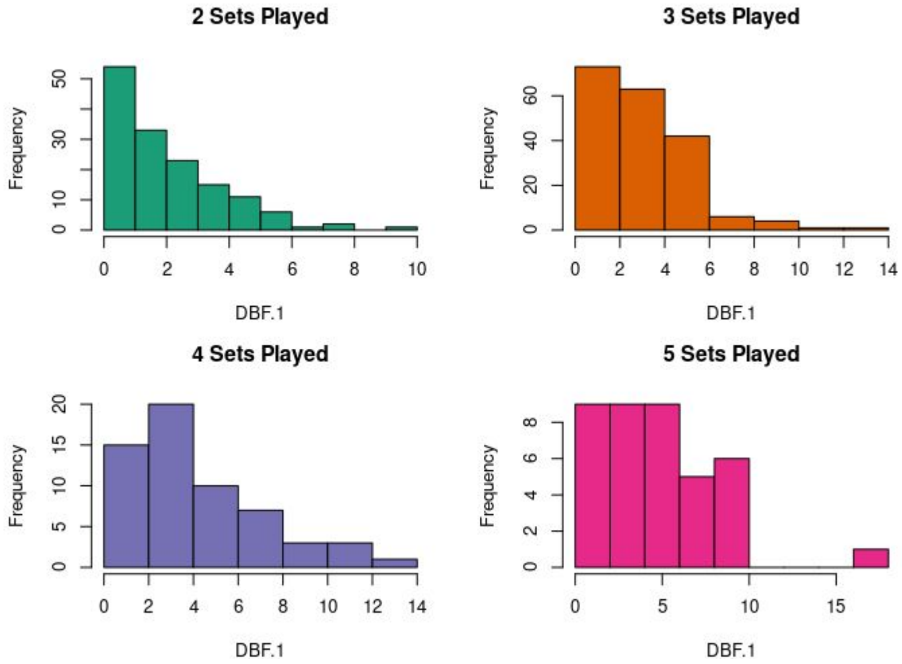




On the contrary, many other features like Double Faults Committed are not normally distributed.

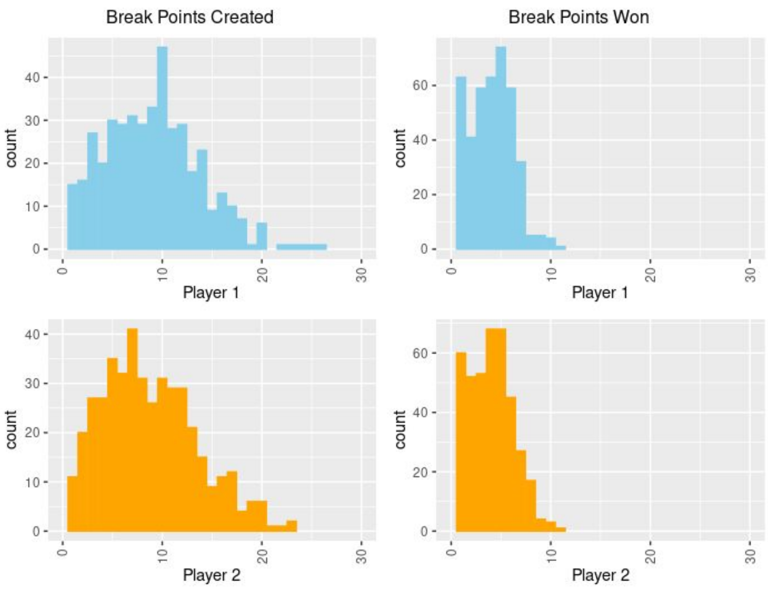
Exploring normality: DBF

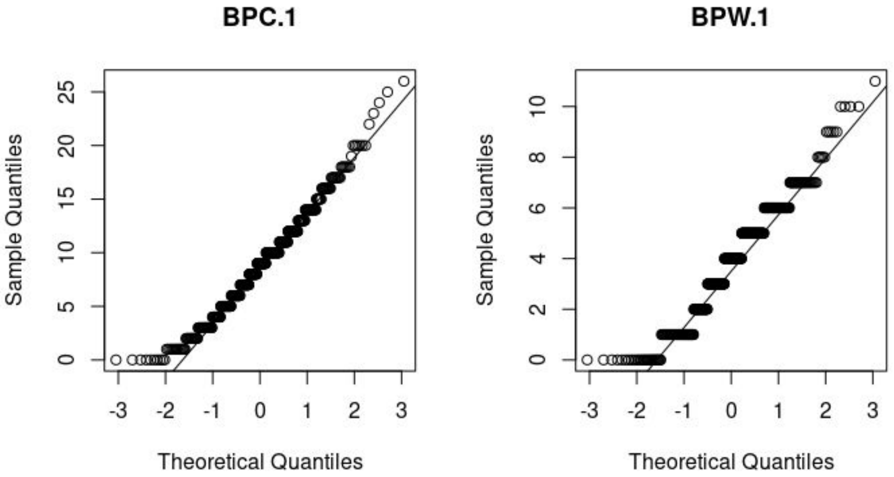




In this case, we see that the feature is consistently not normally distributed across values of STP

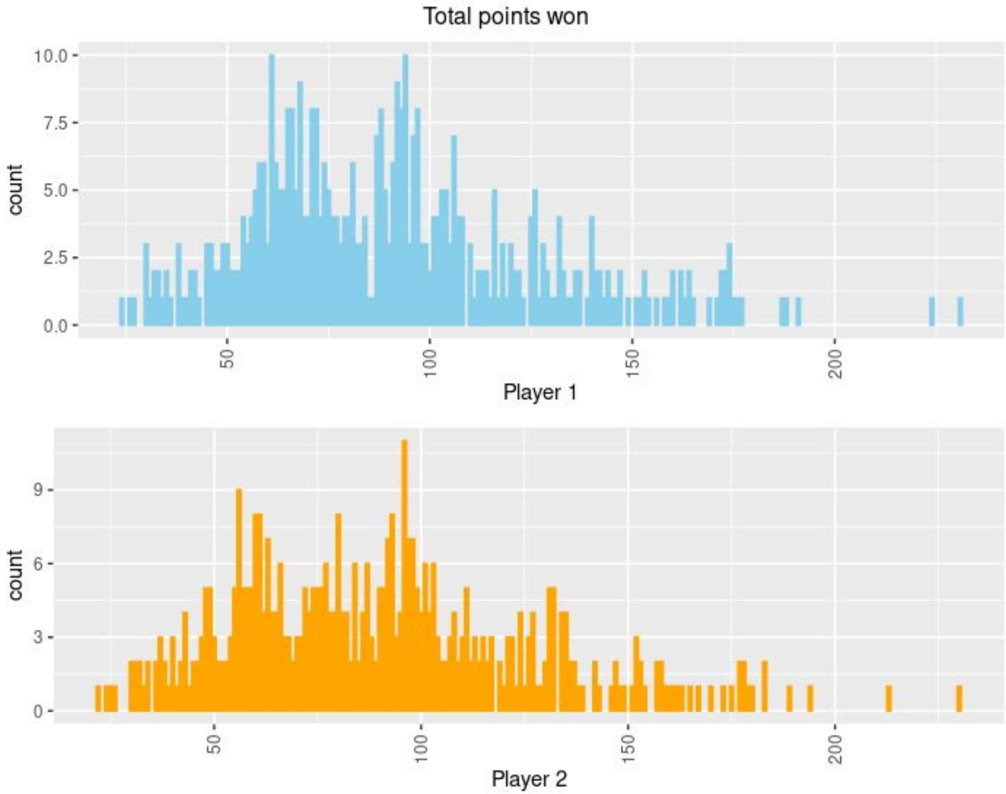
Exploring normality: BPC & BPW





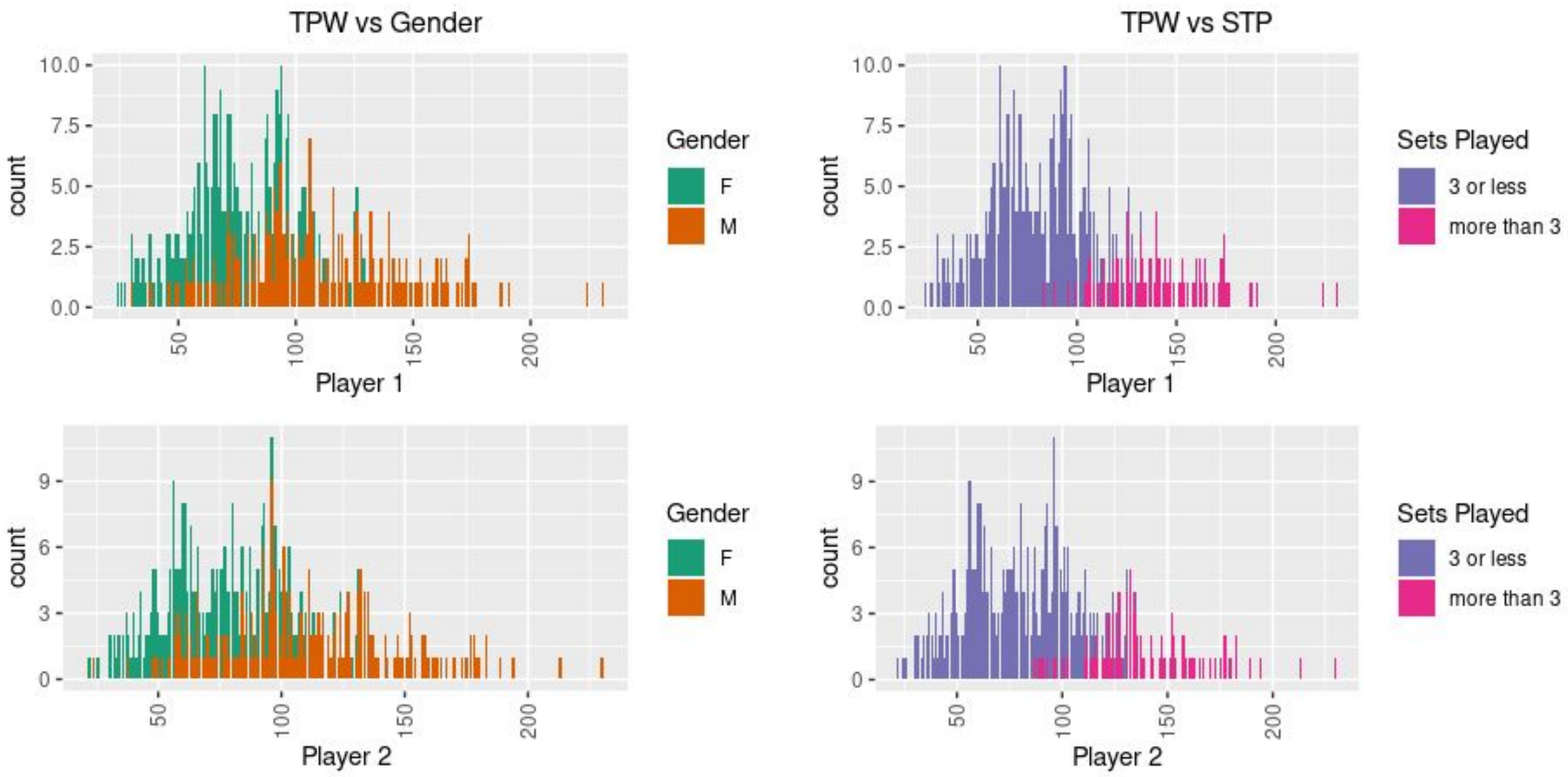
Also features Break Points Created and Break Points Won are not normally distributed.

TPW skewness



The number of Total Points Won by each player has very right-skewed distribution. We further investigate its shape by conditioning it on other possibly relevant variables.

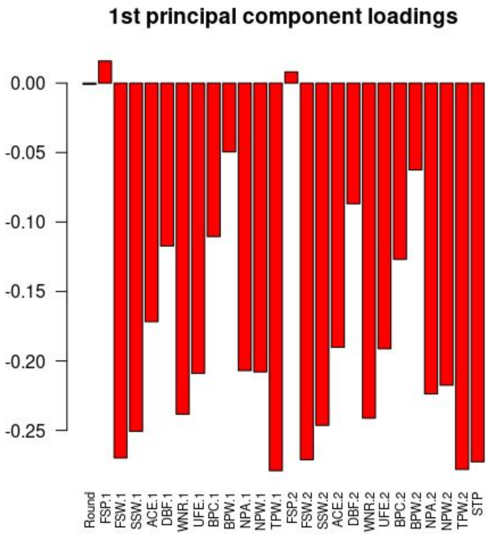
TPW skewness



Principal Components Analysis

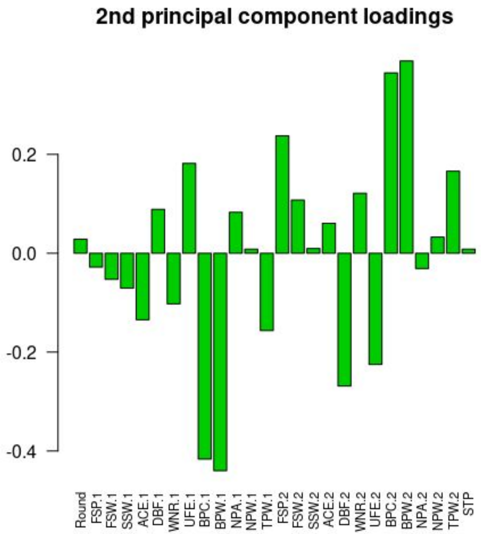
PCA is a dimensionality reduction technique which identifies the *components* that explain the greatest proportion of variance. Components are defined as linear combinations of the features in the original data matrix.

We applied PCA on the standardized data matrix.



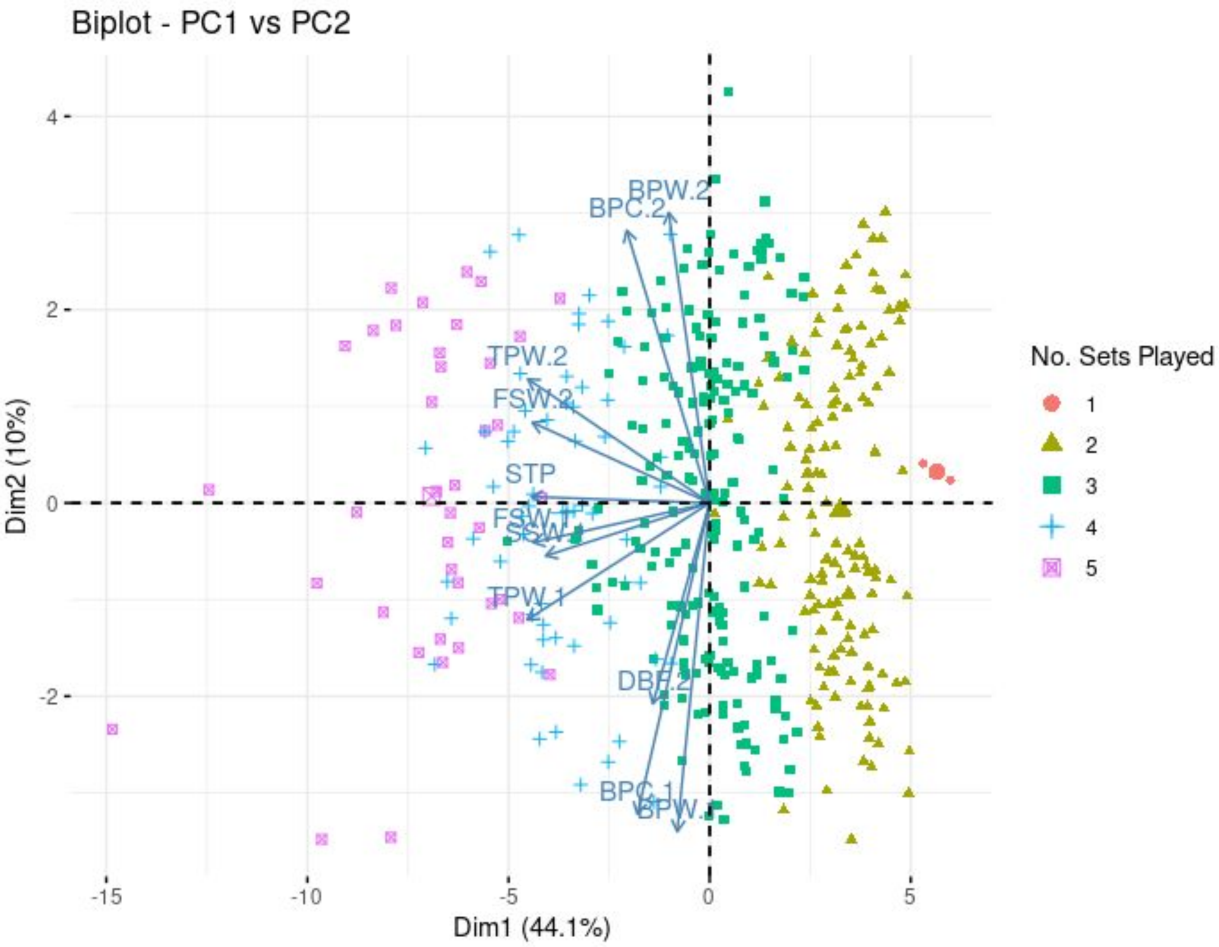
Fraction of total variance explained by the first component

44,1%

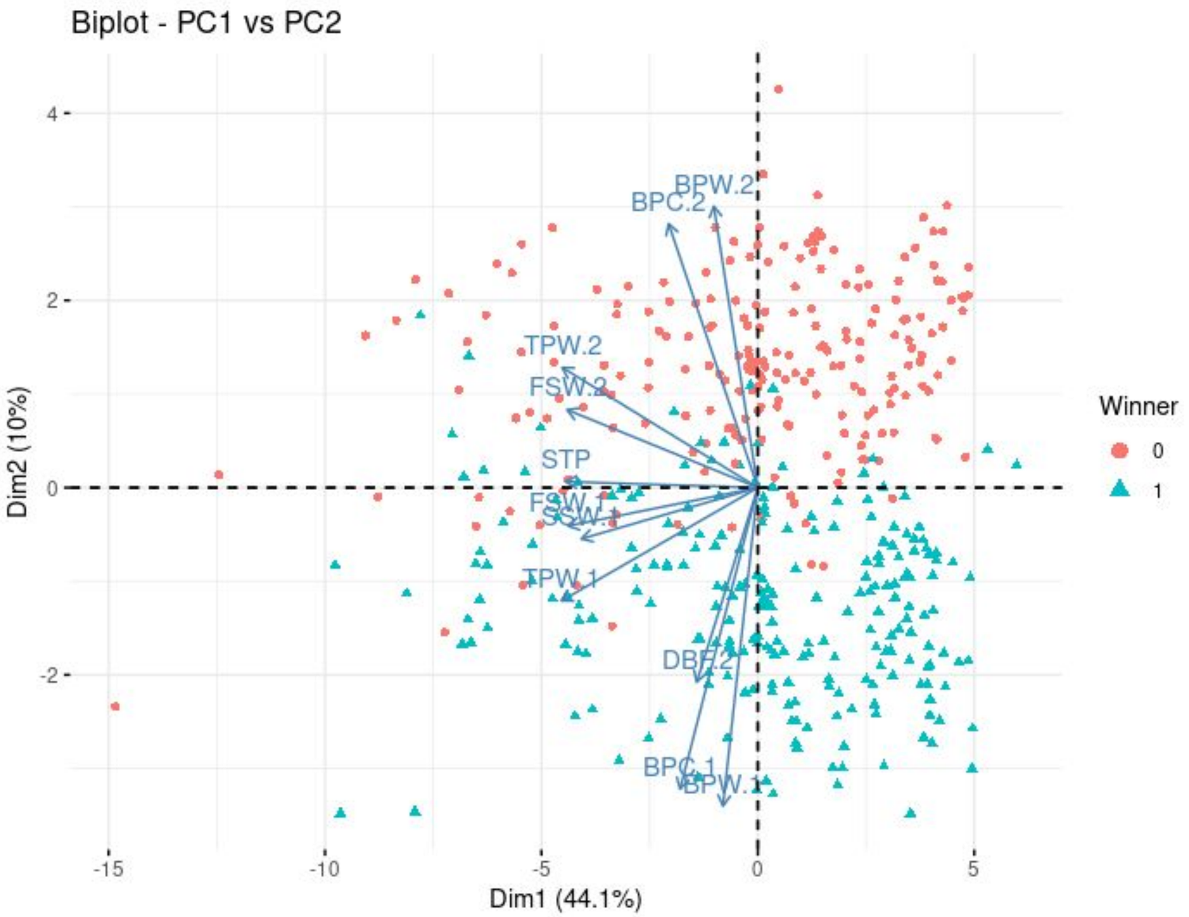


Fraction of total variance explained by the second component

10,0%



The **first component** is strictly related to the duration of the match. Indeed, the dataset instances projected on the plane spanned by PC1 and PC2 are well separated w.r.t. the number of sets played.



The **second component** is instead related to the match outcome. In this case, the projection of the instances are well separated along the y-axis w.r.t. the winner of the match. Indeed, two of the main features in the 2nd PC are BPC and BPW.

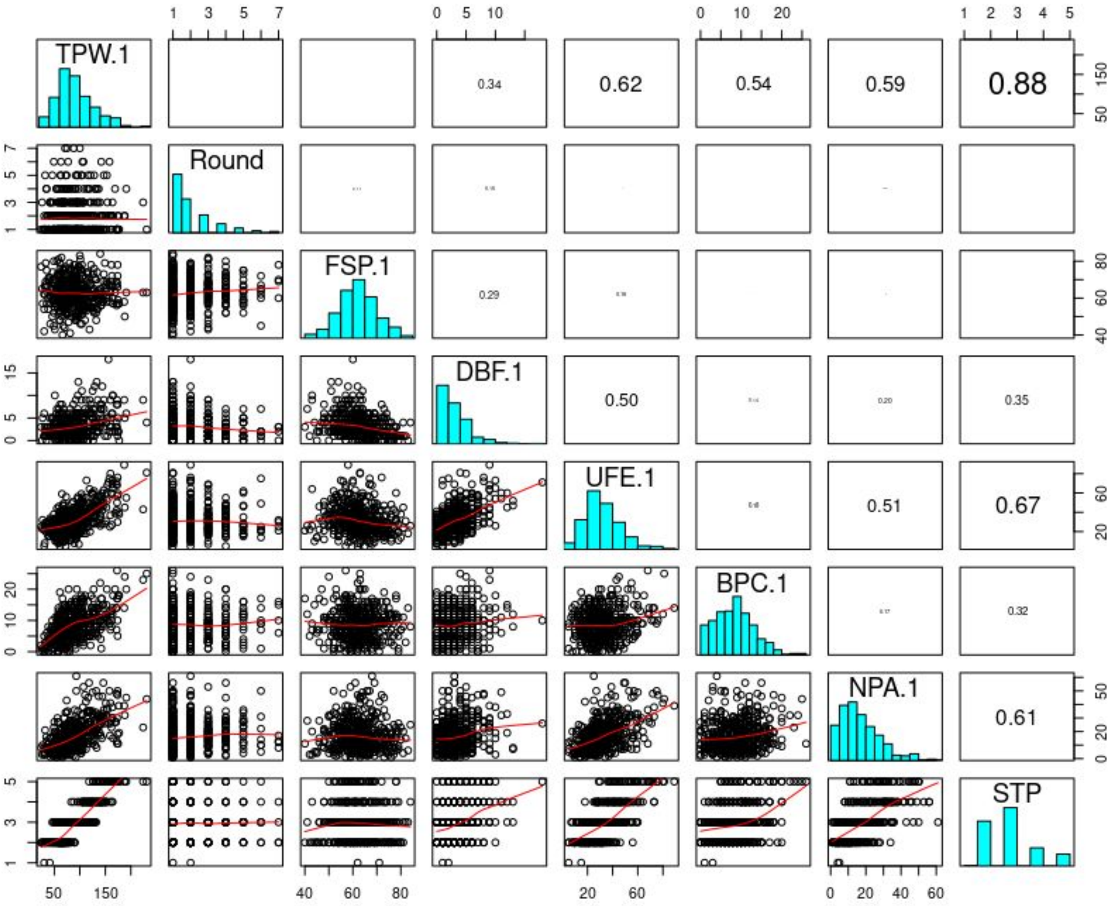
Model data

- Simple Linear Regression
- Multiple Linear Regression
- Logistic Regression
- k-NN Classifier

Pairs plots

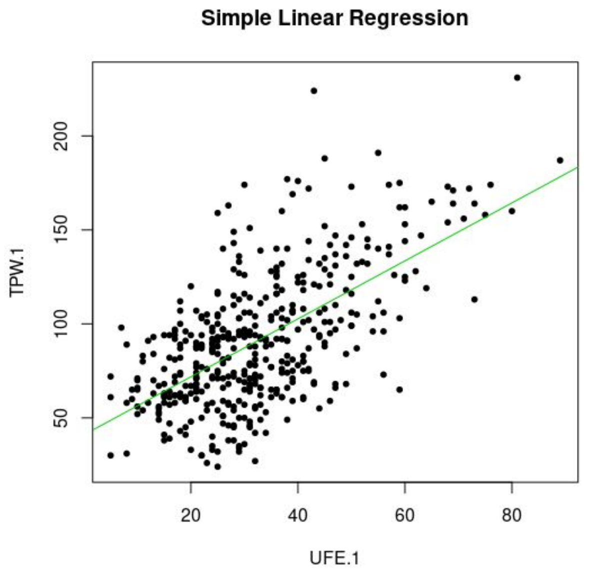
Looking at this plot we notice that some features are linearly correlated with TPW.1.

STP is naturally correlated with it, since it describes the match duration. Other correlated variables are Unforced Errors (UFE), Break Points Created (BPC) and Net Points Attempted (NPA).



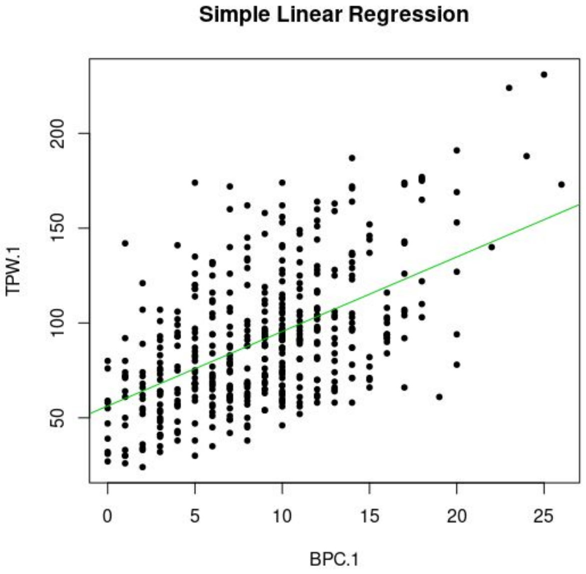
Simple Linear Regression

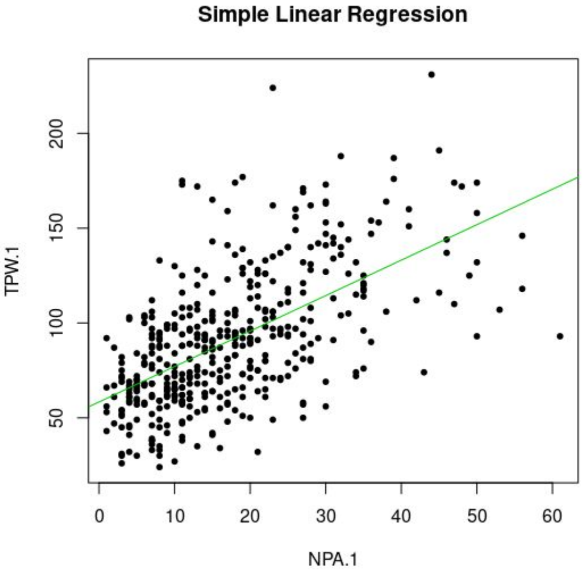
As a first trial, we attempted to model the number of Total Points Won by player 1 (TPW.1) using a simple linear regression model. We used as predictors the features with greatest correlation with the target variable. For none of these predictors we obtained satisfactory results.




```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.97707    3.30581   12.39   <2e-16 ***
UFE.1        1.54244    0.09318   16.55   <2e-16 ***

Residual standard error: 27.9 on 434 degrees of freedom
Multiple R-squared:  0.387,     Adjusted R-squared:  0.3856
F-statistic:   274 on 1 and 434 DF,  p-value: < 2.2e-16
```

Simple Linear Regression





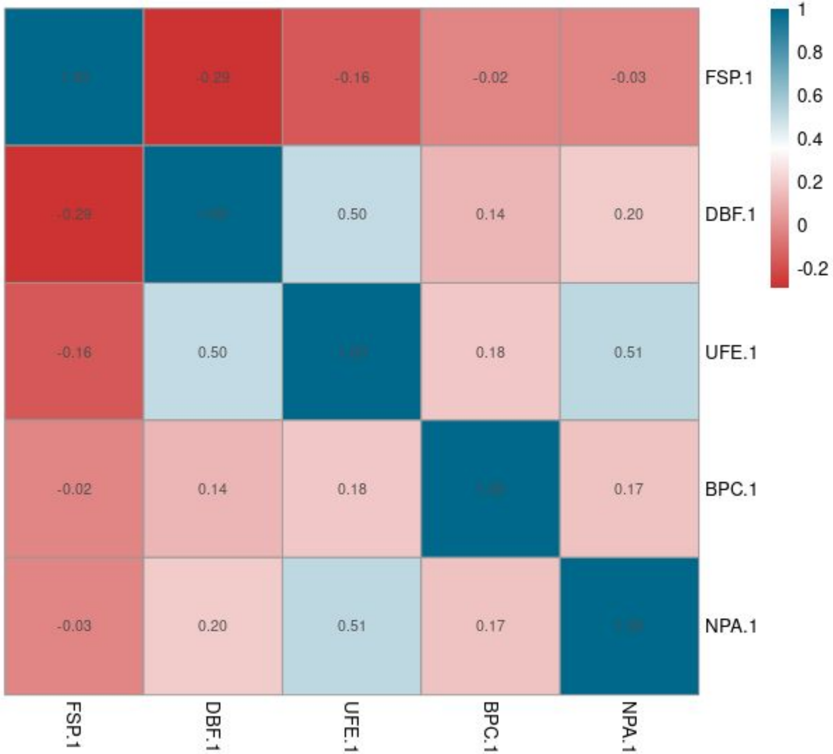

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  56.2591     2.9864   18.84   <2e-16 ***
BPC.1         3.9298     0.2957   13.29   <2e-16 ***

Residual standard error: 30.05 on 434 degrees of freedom
Multiple R-squared:  0.2892,     Adjusted R-squared:  0.2876
F-statistic: 176.6 on 1 and 434 DF,  p-value: < 2.2e-16
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  58.3698     2.5352   23.02   <2e-16 ***
NPA.1         1.8698     0.1219   15.33   <2e-16 ***

Residual standard error: 28.7 on 434 degrees of freedom
Multiple R-squared:  0.3514,     Adjusted R-squared:  0.3499
F-statistic: 235.1 on 1 and 434 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression

We tried then fitting a Multiple Linear Regression to predict TPW.1, using only features related to Player 1 which are not part of the target variable.

```
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -16.69523    9.08203  -1.838  0.06671 .  
FSP.1         0.48704    0.13148   3.704  0.00024 ***
DBF.1         0.93859    0.46153   2.034  0.04260 *  
UFE.1         0.90649    0.09182   9.873  < 2e-16 ***
BPC.1         3.01315    0.20785  14.497  < 2e-16 ***
NPA.1         1.02974    0.10321   9.978  < 2e-16 ***

Residual standard error: 20.66 on 430 degrees of freedom
Multiple R-squared:  0.667,	Adjusted R-squared:  0.6631 
F-statistic: 172.3 on 5 and 430 DF,  p-value: < 2.2e-16
```



A constraint based algorithm for feature selection

- UFE.1

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2230     9.8762  -0.023   0.9820
FSP.1         0.4294     0.1453   2.955   0.0033 **
DBF.1         2.9762     0.4567   6.517 2.01e-10 ***
BPC.1         3.1473     0.2294  13.717  < 2e-16 ***
NPA.1         1.5189     0.1002  15.166  < 2e-16 ***

Residual standard error: 22.86 on 431 degrees of freedom
Multiple R-squared:  0.5915,     Adjusted R-squared:  0.5877
F-statistic:   156 on 4 and 431 DF,  p-value: < 2.2e-16
```

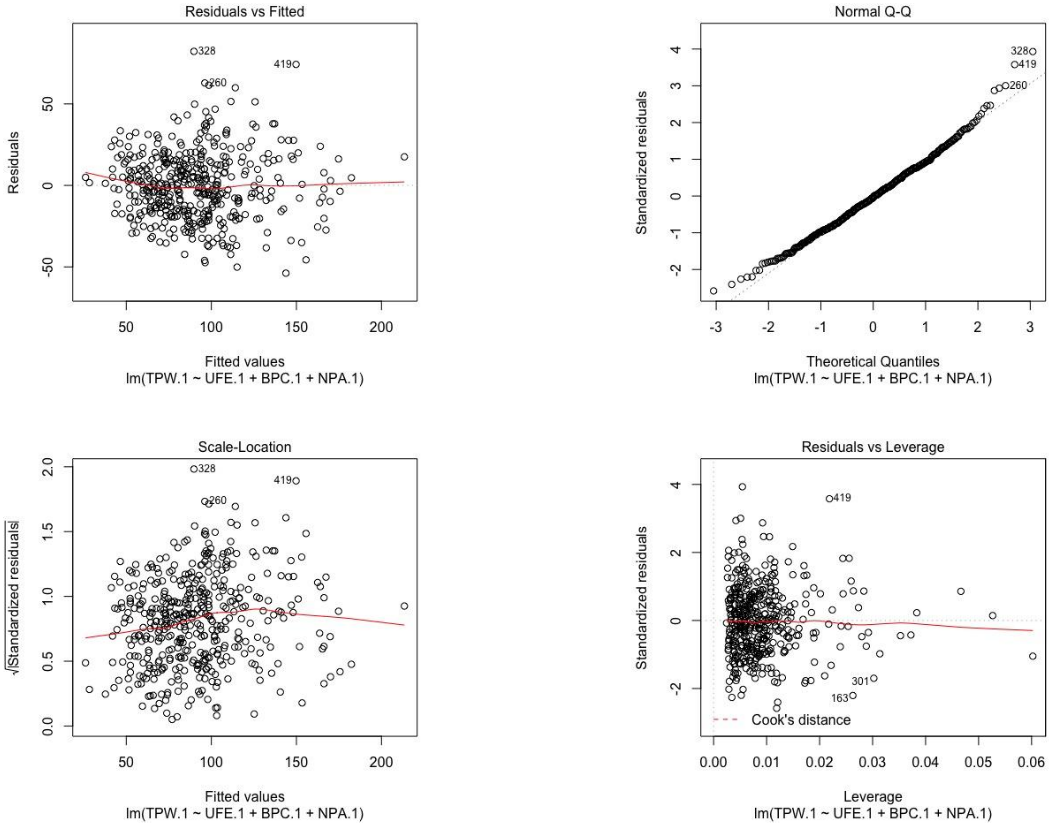
- DBF.1

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.25575    8.84778  -1.385  0.16671
FSP.1         0.42345    0.12817   3.304  0.00103 **
UFE.1         0.98999    0.08242  12.011  < 2e-16 ***
BPC.1         3.04097    0.20815  14.609  < 2e-16 ***
NPA.1         1.01440    0.10330   9.820  < 2e-16 ***

Residual standard error: 20.74 on 431 degrees of freedom
Multiple R-squared:  0.6638,     Adjusted R-squared:  0.6607
F-statistic: 212.7 on 4 and 431 DF,  p-value: < 2.2e-16
```

- FSP.1

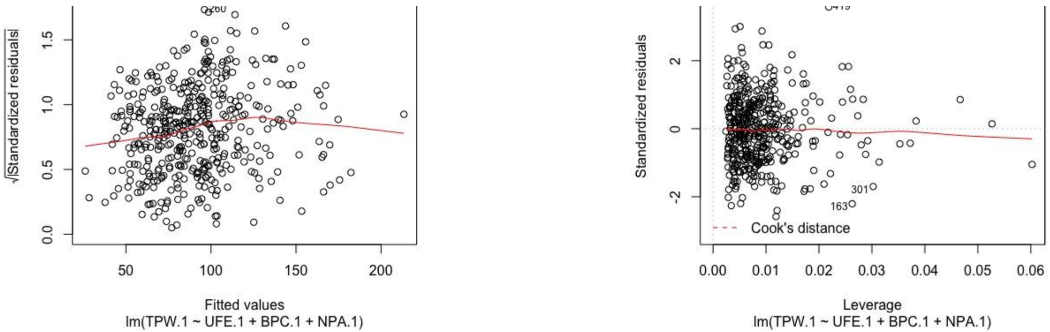
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.4037     2.8953   5.320 1.67e-07 ***
UFE.1         0.9427     0.0821  11.483  < 2e-16 ***
BPC.1         3.0441     0.2105  14.459  < 2e-16 ***
NPA.1         1.0361     0.1043   9.937  < 2e-16 ***

Residual standard error: 20.97 on 432 degrees of freedom
Multiple R-squared:  0.6553,     Adjusted R-squared:  0.6529
F-statistic: 273.7 on 3 and 432 DF,  p-value: < 2.2e-16
```

..and the assumptions of the model?



..and the assumptions of the model?

Verified!



What happens without outliers and HL points?

```
Residual standard error: 20.97 on 432 degrees of freedom
Multiple R-squared:  0.6553, Adjusted R-squared:  0.6529
F-statistic: 273.7 on 3 and 432 DF,  p-value: < 2.2e-16
```

without outliers

```
Residual standard error: 19.71 on 427 degrees of freedom
Multiple R-squared:  0.6748, Adjusted R-squared:  0.6726
F-statistic: 295.4 on 3 and 427 DF,  p-value: < 2.2e-16
```

without high-leverage points

```
Residual standard error: 20.6 on 397 degrees of freedom
Multiple R-squared:  0.5674, Adjusted R-squared:  0.5641
F-statistic: 173.5 on 3 and 397 DF,  p-value: < 2.2e-16
```

What happens without outliers and HL points?

```
Residual standard error: 20.97 on 432 degrees of freedom
Multiple R-squared:  0.6553, Adjusted R-squared:  0.6529
F-statistic: 273.7 on 3 and 432 DF,  p-value: < 2.2e-16
```

without outliers

```
Residual standard error: 19.71 on 427 degrees of freedom
Multiple R-squared:  0.6748, Adjusted R-squared:  0.6726
F-statistic: 295.4 on 3 and 427 DF,  p-value: < 2.2e-16
```

without high-leverage points

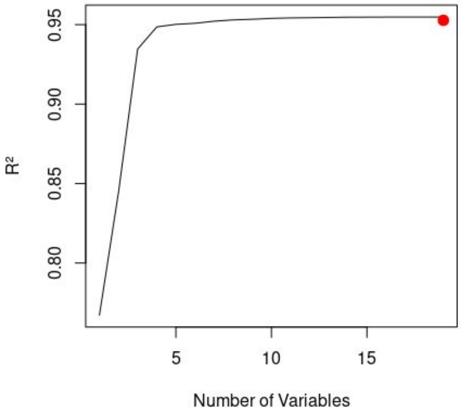
```
Residual standard erro                97 degrees of freedom
Multiple R-squared:  0.              usted R-squared:  0.5641
F-statistic: 173.5 on                F,  p-value: < 2.2e-16
```

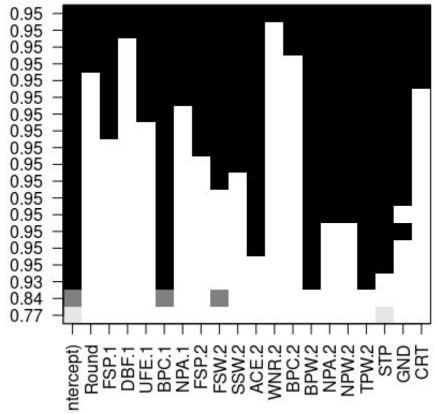
Multiple Linear Regression

Afterwards, we decided to model TPW.1 using as predictors also the variables which describe the points attempted and won by player 2.

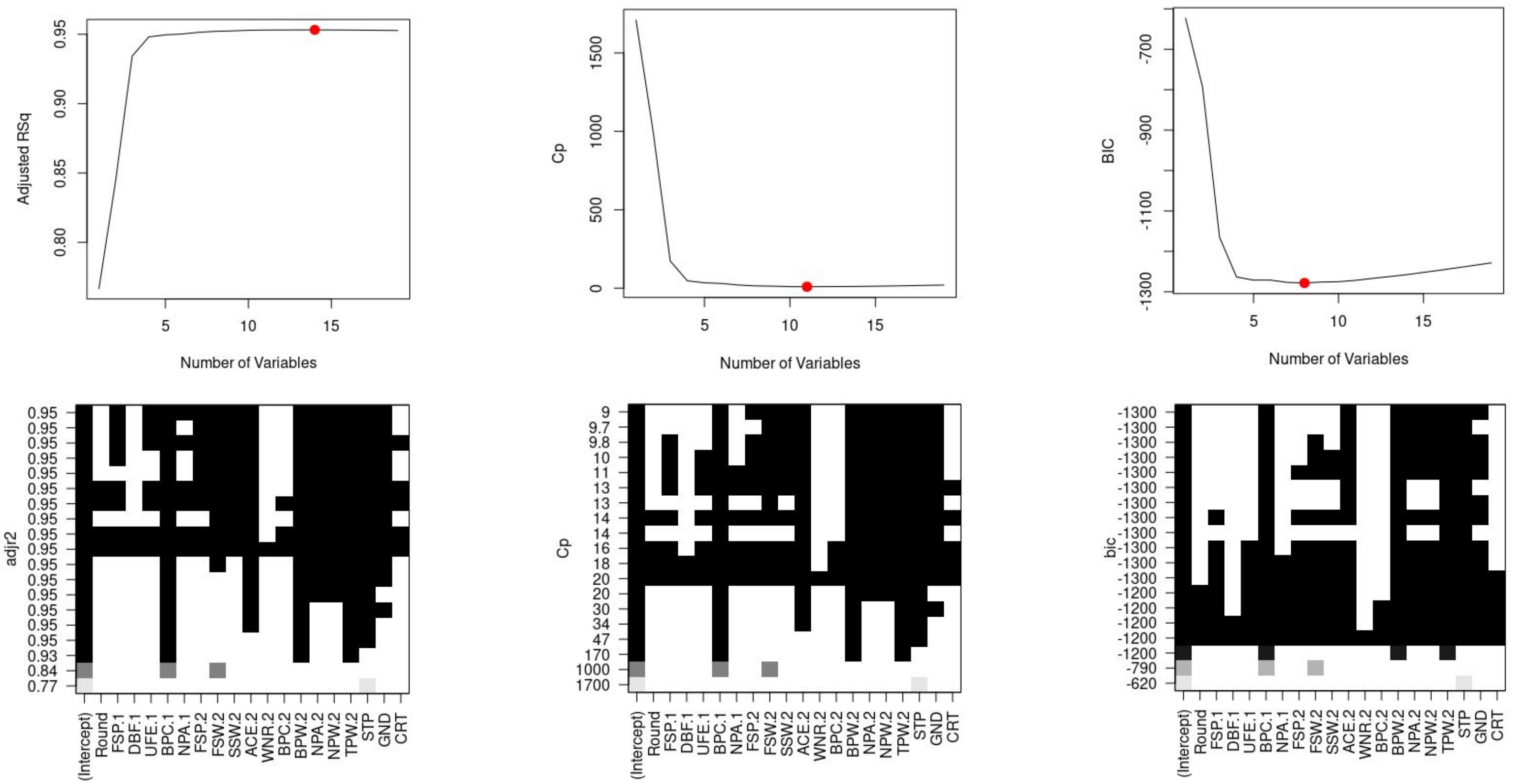
In this case, we performed a best subset selection to reduce the number of predictors.

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.123605   5.814374   0.193 0.846861
Round        -0.185140   0.296366  -0.625 0.532509
FSP.1         0.072215   0.050879   1.419 0.156551
DBF.1         0.024613   0.183899   0.134 0.893593
                           ...
STP          10.828757   1.052408  10.290  < 2e-16 ***
GND          -2.338606   1.167402  -2.003 0.045798 *
CRT          -0.271521   0.420857  -0.645 0.519178

Residual standard error: 7.741 on 416 degrees of freedom
Multiple R-squared:  0.9548,     Adjusted R-squared:  0.9527
F-statistic: 462.3 on 19 and 416 DF,  p-value: < 2.2e-16
```





Multiple Linear Regression

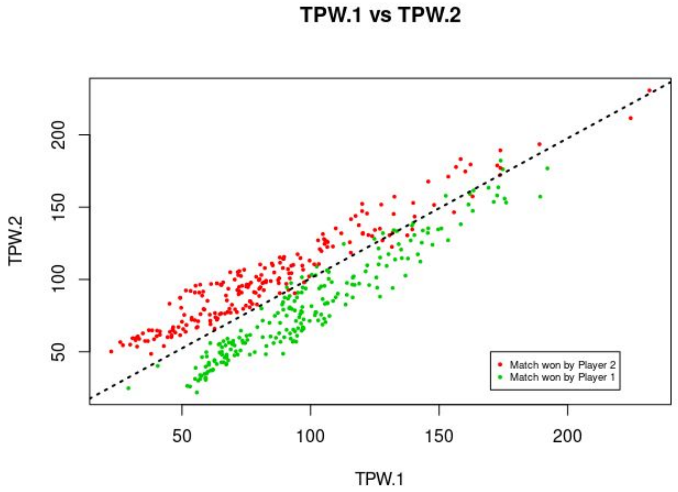


Logistic Regression

We also tried to predict the result of each match using Logistic Regression models.

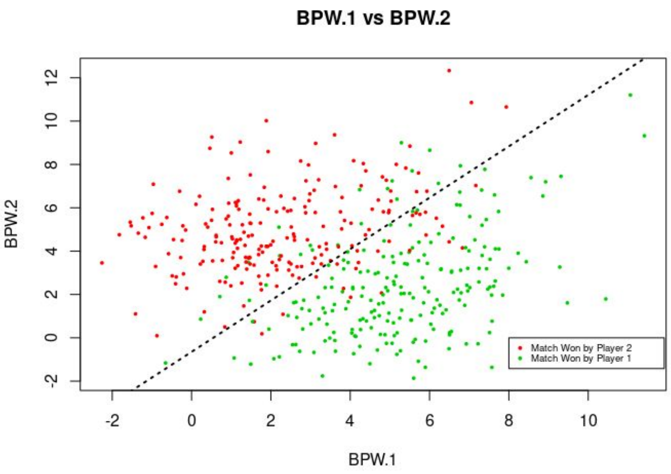
As a first step, we investigated the impact of some pairs of corresponding variables on the outcome of the matches.

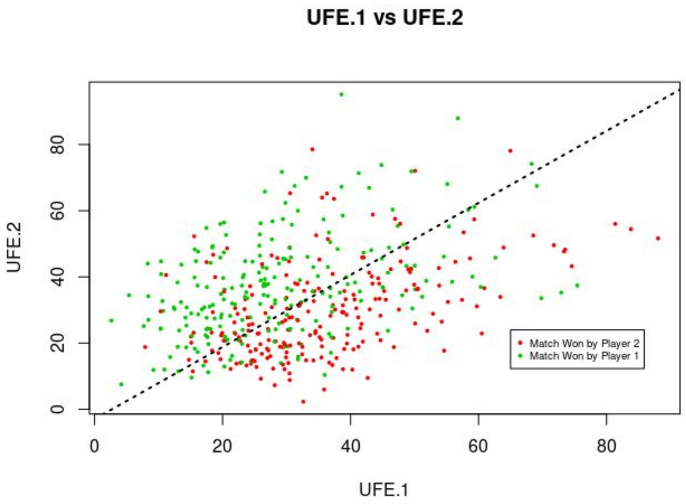
We notice that the results are consistent with what we observed analysing the PCA.




```
Accuracy = 0.94

Precision = 0.95

Recall = 0.94
```

```
pred.TPW    0    1

       0  202   14

       1   11  209
```



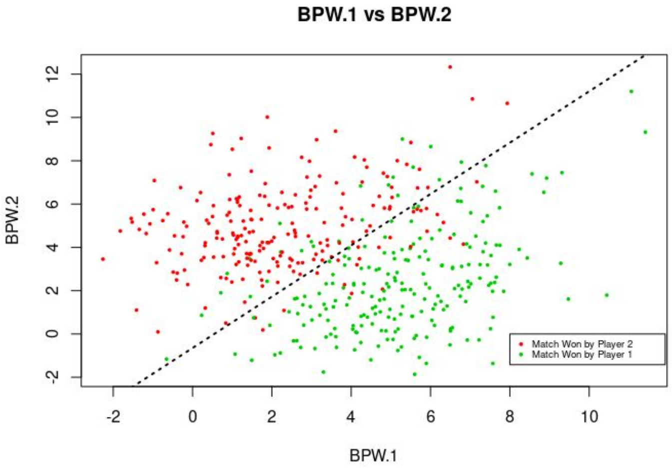


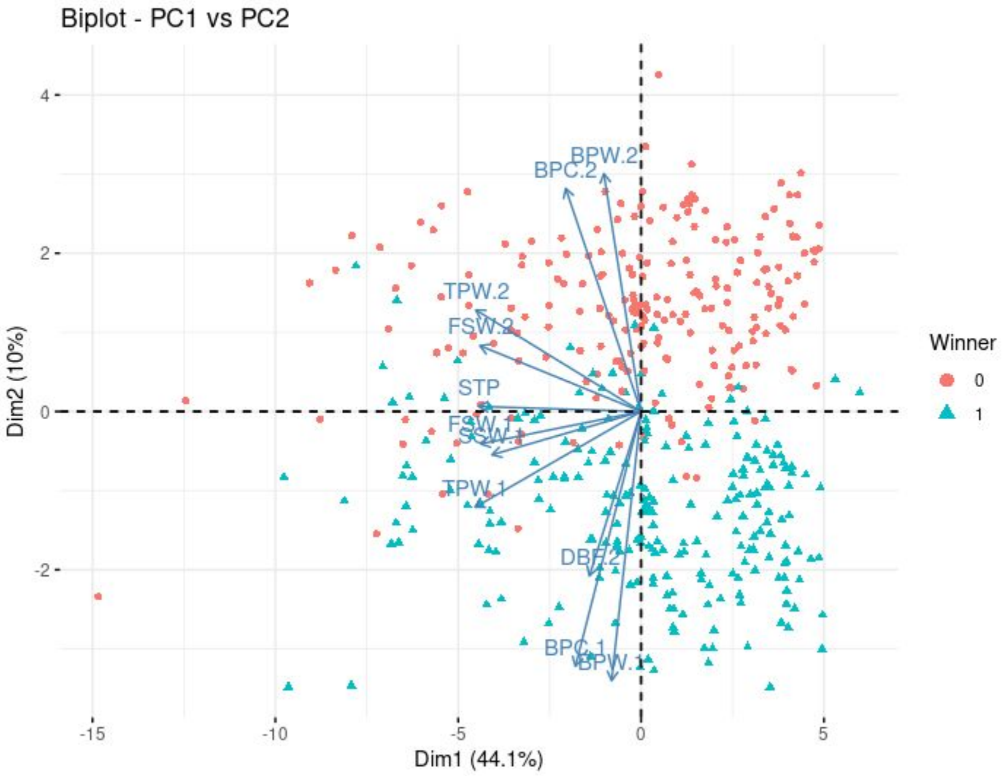

```
Accuracy = 0.94

Precision = 0.95

Recall = 0.94
```

pred.BPW	0	1
0	201	14
1	12	209

```
Accuracy = 0.70

Precision = 0.71

Recall = 0.70
```

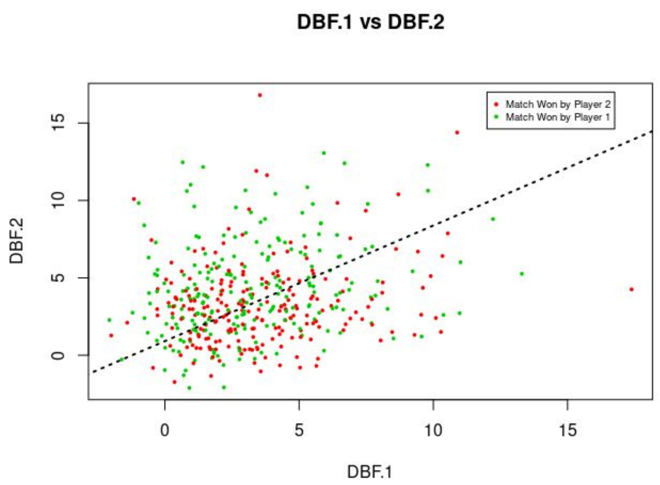
pred.UFE	0	1
0	150	67
1	63	156

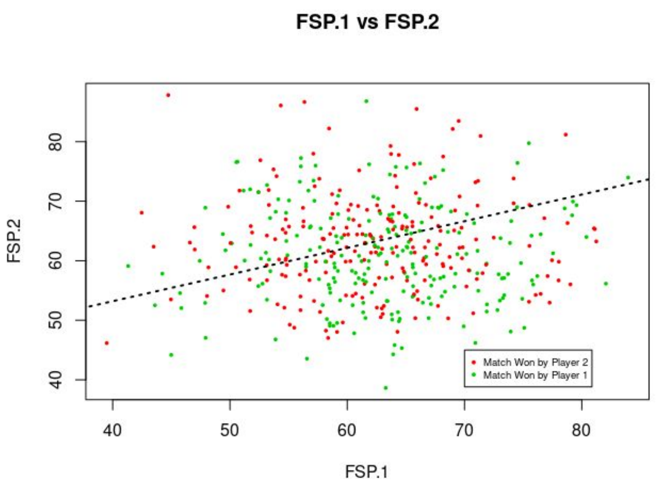





```
Accuracy = 0.94

Precision = 0.95

Recall = 0.94
```

pred.BPW	0	1
0	201	14
1	12	209



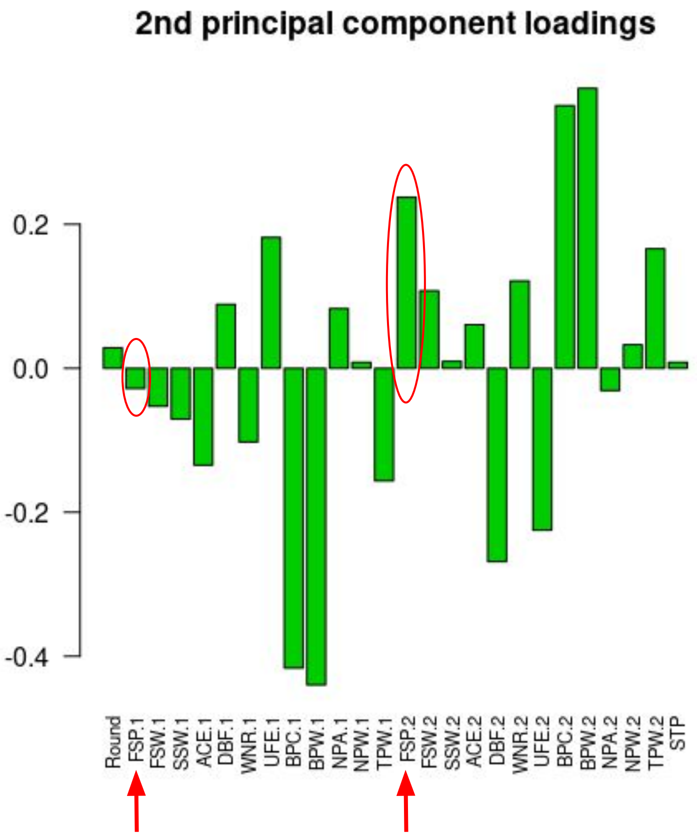


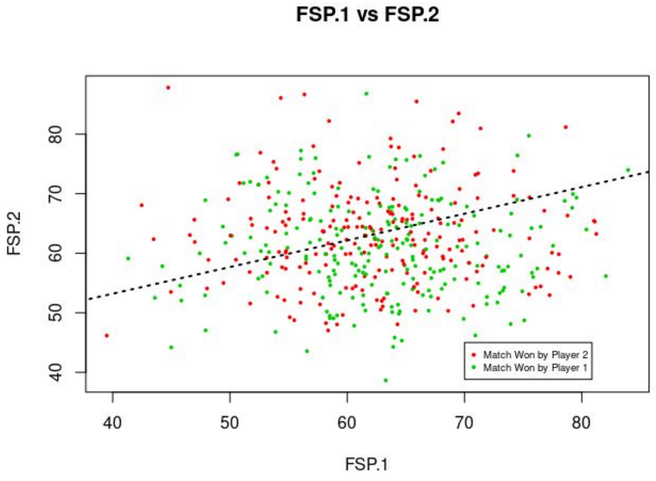

```
Accuracy = 0.57

Precision = 0.58

Recall = 0.58
```

pred.DBF	0	1
0	117	93
1	96	130

```
Accuracy = 0.57

Precision = 0.57

Recall = 0.64
```

pred.FSP	0	1
0	104	80
1	109	143





Accuracy = 0.57

Precision = 0.57

Recall = 0.64

pred.FSP	0	1
0	104	80
1	109	143

Logistic Regression

After exploring the impact of these features, we decided to model the result of the match using all variables but the number of sets won by each player (FNL.1, FNL.2).

This time, we performed selected variables using a greedy approach, i.e. performing a **Backward Stepwise Selection**.

Starting from the full model (28 covariates) with AIC equal to 121.71, the method selected a model with AIC equal to 92.54 (8 covariates).

Full model

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.285348   9.886497  -0.130  0.89656
Round        0.125183   0.345923   0.362  0.71744
...
STP         -1.282637   1.136163  -1.129  0.25893

    Null deviance: 604.195  on 435  degrees of freedom
Residual deviance:  63.712  on 407  degrees of freedom
AIC: 121.71
```

Reduced model

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.98798    1.08665  -0.909 0.363243
SSW.1       -0.13031    0.06019  -2.165 0.030375 *
BPW.1        1.52846    0.43600   3.506 0.000456 ***
NPA.1       -0.25318    0.09404  -2.692 0.007095 **
NPW.1        0.32892    0.14001   2.349 0.018813 *
TPW.1        0.24892    0.06293   3.955 7.64e-05 ***
FSW.2       -0.30203    0.07567  -3.992 6.56e-05 ***
SSW.2       -0.37709    0.10220  -3.690 0.000225 ***
BPW.2       -1.75869    0.33194  -5.298 1.17e-07 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 604.195  on 435  degrees of freedom
Residual deviance:  74.544  on 427  degrees of freedom
AIC: 92.544
```

The effect on the deviance of BPW.2

Once found the best logistic model, the removal of other variables, such as BPW.2, let the deviance rise dramatically.

```
                Reduced with BPW.2
Null deviance: 604.195  on 435  degrees of freedom
Residual deviance:  74.544  on 427  degrees of freedom
AIC: 92.544
```

```
                Reduced model without BPW.2
Null deviance: 604.19  on 435  degrees of freedom
Residual deviance: 159.42  on 428  degrees of freedom
AIC: 175.42
```

Then, the accuracy of the best reduced model was evaluated with **LOOCV**.

```
Accuracy = 0.97

Precision = 0.96

Recall = 0.97
```

```
Pred    0   1

   0  205   6

   1    8 217
```

k-Nearest Neighbors

Finally, we attempted a modeling of the match result using k-NN. Again, we used as predictors all variables but the number of sets won by each player (FNL.1, FNL.2).

We evaluated the test-set accuracy of the model performing a **LOOCV**, searching also (not exhaustively) the best value of the parameter *k*.

Taking into account the model complexity as well as the mean accuracy value, we can select the 25-NN model as the optimal one.

	acc
25	0.9266055
97	0.9266055
28	0.9243119
31	0.9243119
34	0.9220183
37	0.9220183
22	0.9197248
100	0.9197248
103	0.9197248
106	0.9197248

```
Accuracy = 0.93

Precision = 0.94

Recall = 0.92
```

Pred	0	1
0	199	18
1	14	205

k-Nearest Neighbors

We tried also to apply kNN **standardized** and **normalized** variables, since the scale of the values can influence the result of this algorithm based on the computation of distance between samples.

Indeed, both standardizing and normalizing data we obtained an improvement in the classification accuracy (respectively, ~ 96% and ~ 94%) and different values for the best selected *k* parameter.

	acc
7	0.9610092
73	0.9564220
64	0.9541284
70	0.9541284
79	0.9541284

Standardized

Pred Std	0	1
0	206	10
1	7	213

	acc
34	0.9426606
28	0.9357798
31	0.9357798
37	0.9357798
40	0.9311927

Normalised

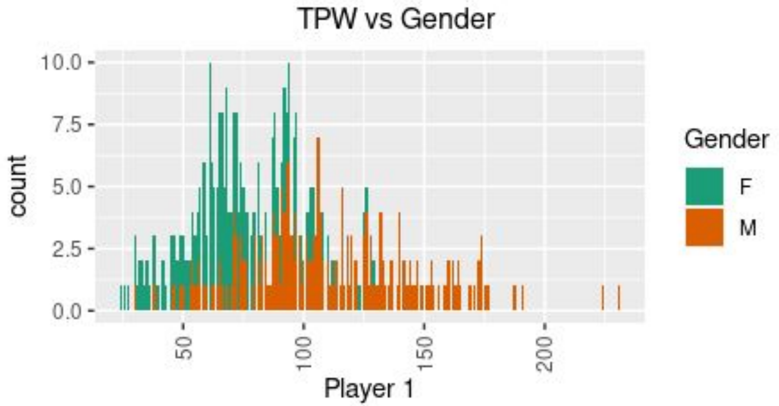
Pred Norm	0	1
0	203	15
1	10	208

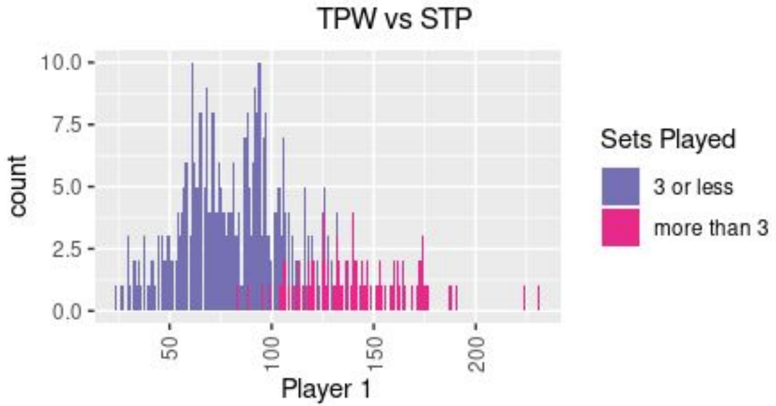
Interpreting data

- Gender differences in tennis
- Different grounds and rounds effect on performances

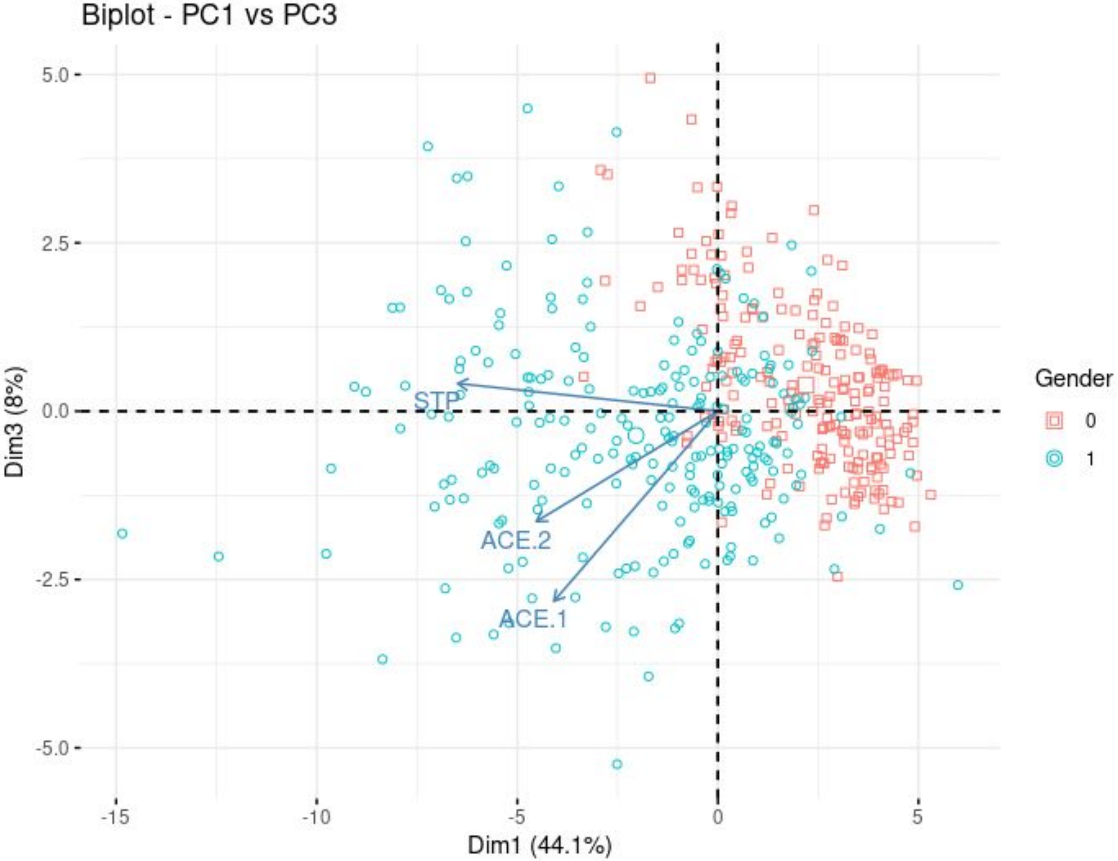
Men’s matches are longer

By the official rules of major tennis tournaments, men play longer matches. Indeed, they play at the best of five sets, while women play at the best of three. Hence, men make generally more points than women in a match.





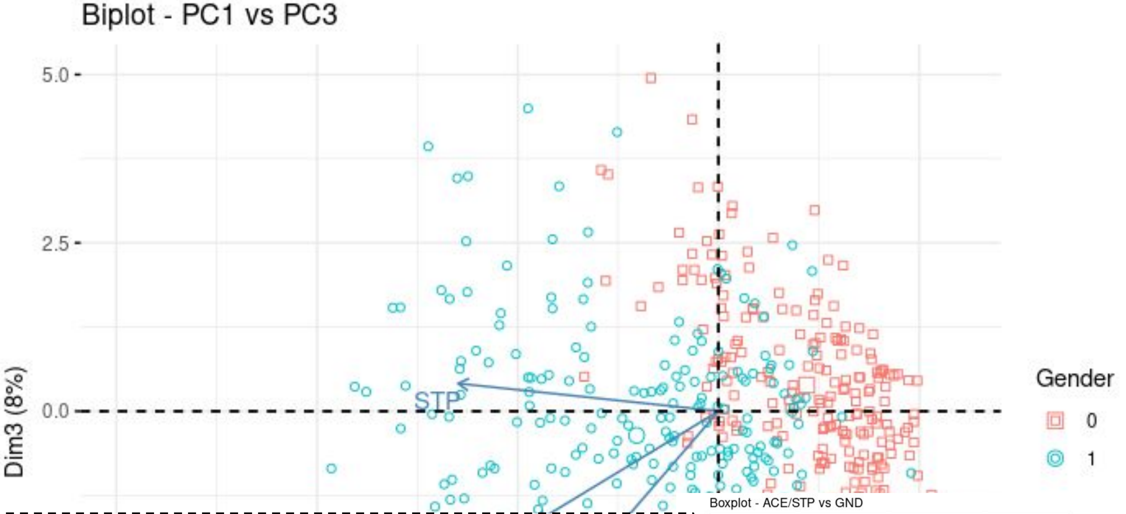
Men score more aces



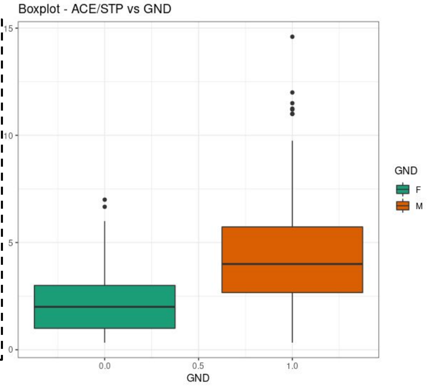
Looking at the **third component** obtained in the PCA, we can gain an interesting insight about gender differences in the style of play.

Some of the features that have more weight in the 3rd component are ACE.x. In fact, the nice separation of points is due to the higher rate of aces in men's matches than in women's.

Men score more aces...



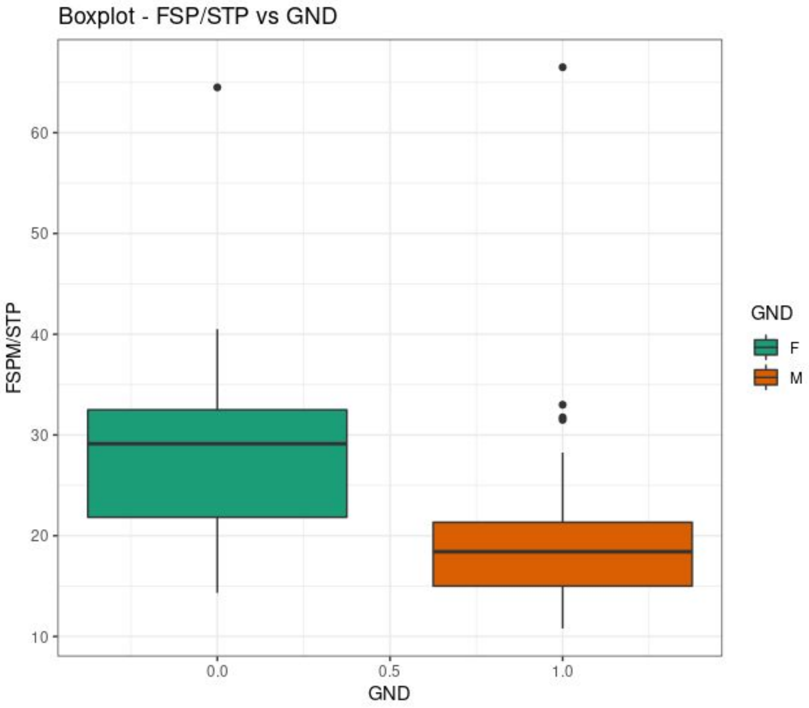

```
Welch Two Sample t-test (equal.var=FALSE)

data:  ACEM/STP by GND
t = -12.066, df = 365, p-value < 2.2e-16
alternative hypothesis: true difference in
means is less than 0
95 percent confidence interval:
      -Inf -2.010369
sample estimates:
mean in group 0 mean in group 1
       2.140476        4.469100
```

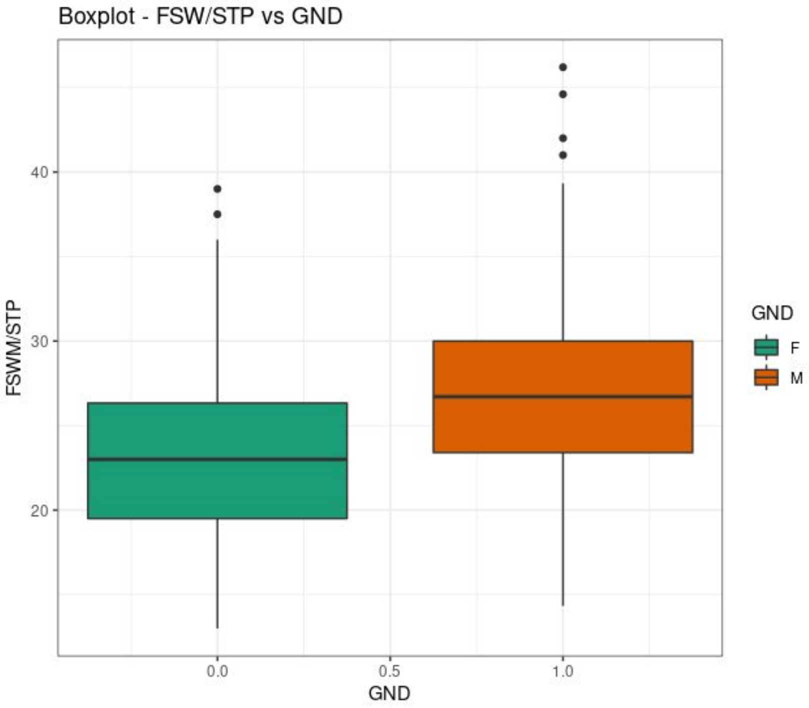


Looking at the **third component** obtained in the PCA, we can gain an interesting insight about gender differences in the style of play.

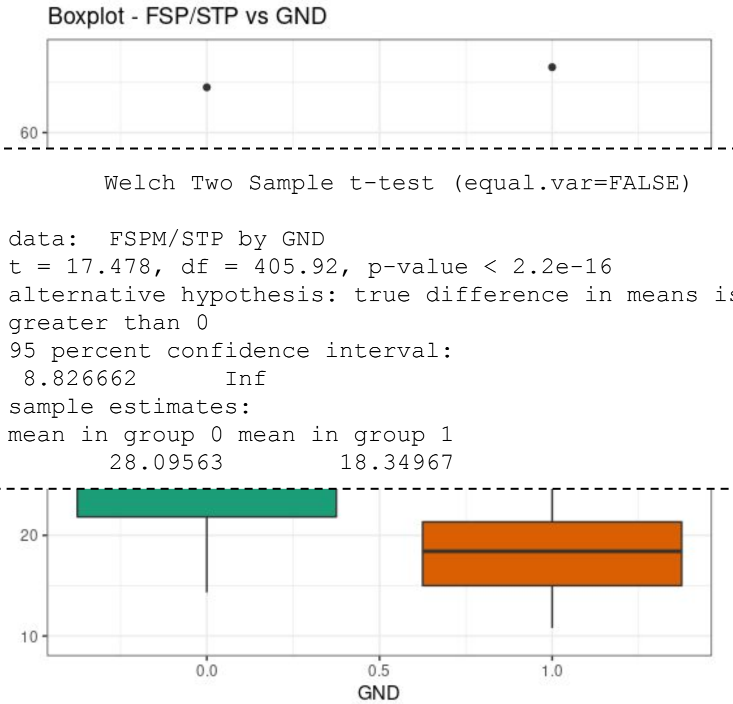
Some of the features that have more weight in the 3rd component are ACE.x. In fact, the nice separation of points is due to the higher rate of aces in men’s matches than in women’s.

…and they also force first serve more

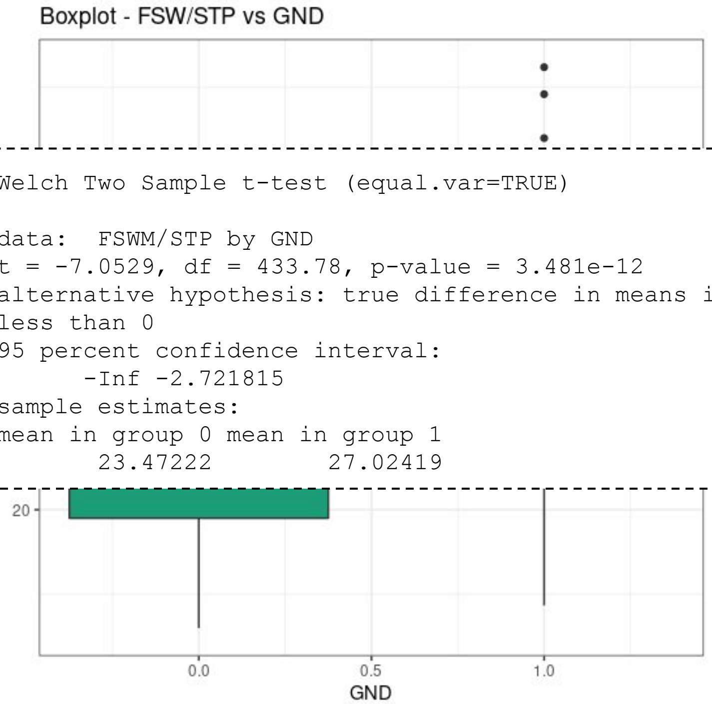




…and they also force first serve more

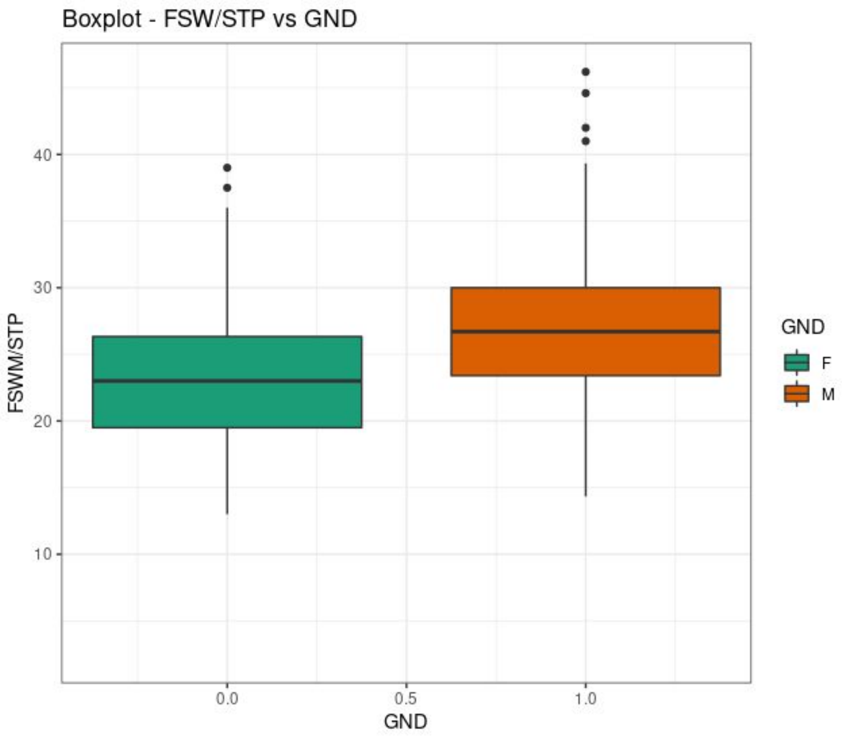


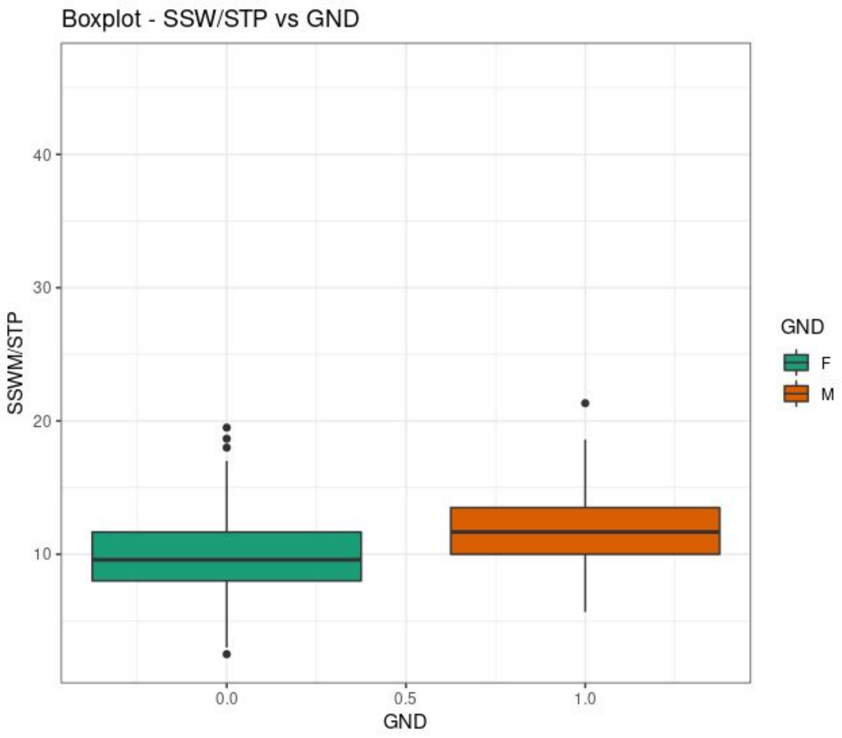

```
      Welch Two Sample t-test (equal.var=FALSE)

data:  FSPM/STP by GND
t = 17.478, df = 405.92, p-value < 2.2e-16
alternative hypothesis: true difference in means is
greater than 0
95 percent confidence interval:
 8.826662      Inf
sample estimates:
mean in group 0 mean in group 1
       28.09563        18.34967
```



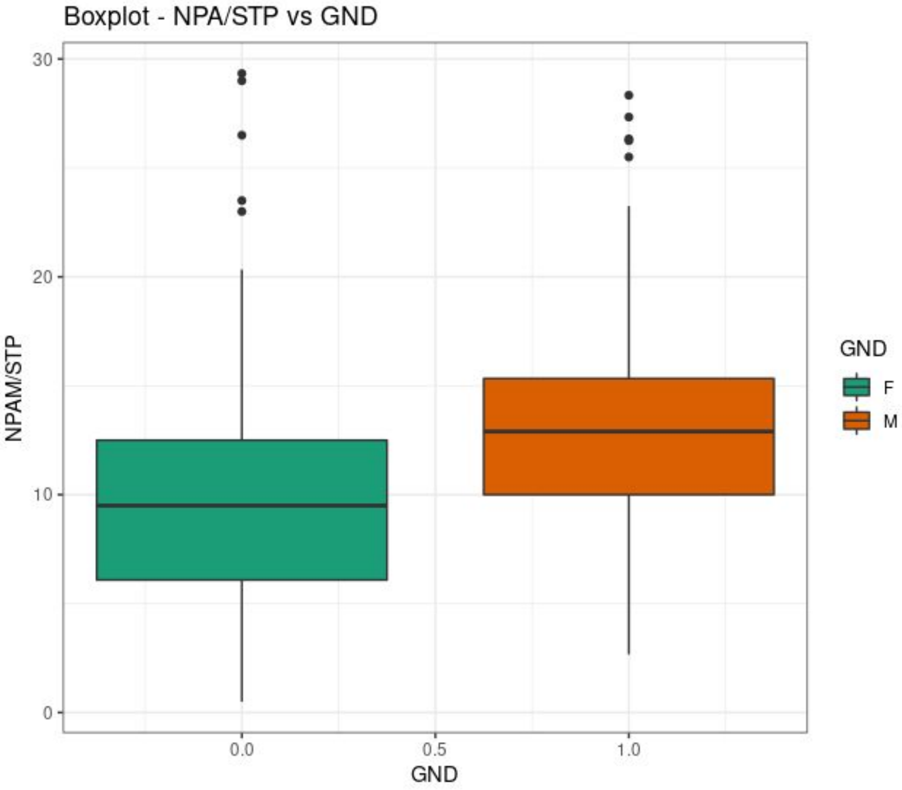

```
Welch Two Sample t-test (equal.var=TRUE)

data:  FSWM/STP by GND
t = -7.0529, df = 433.78, p-value = 3.481e-12
alternative hypothesis: true difference in means is
less than 0
95 percent confidence interval:
      -Inf -2.721815
sample estimates:
mean in group 0 mean in group 1
       23.47222        27.02419
```

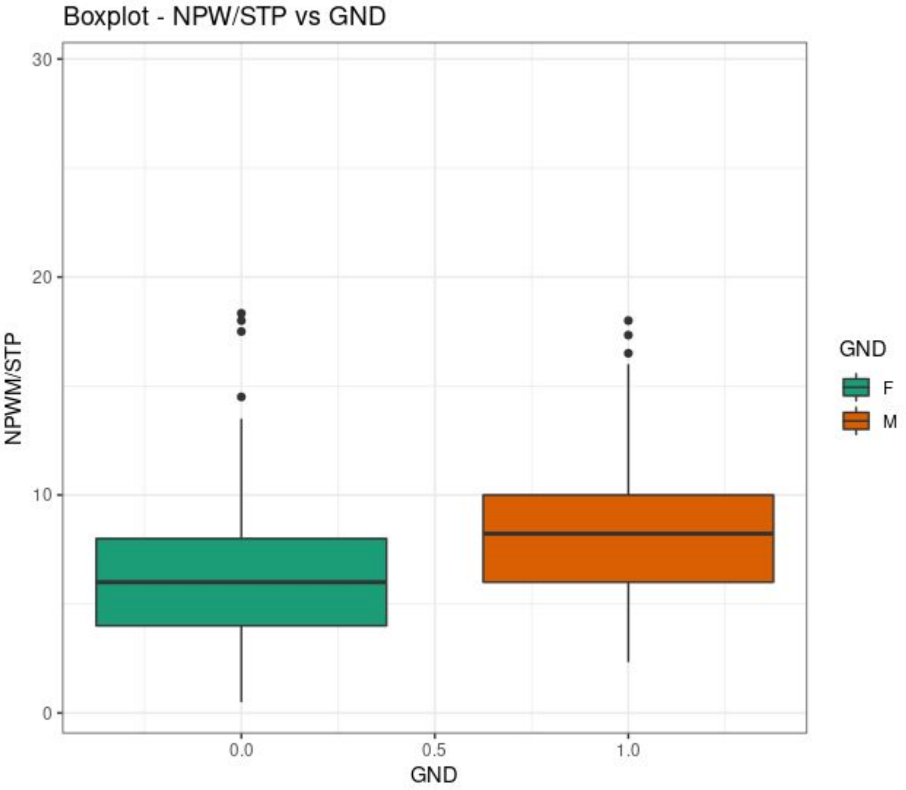
Second serve points are harder to win!



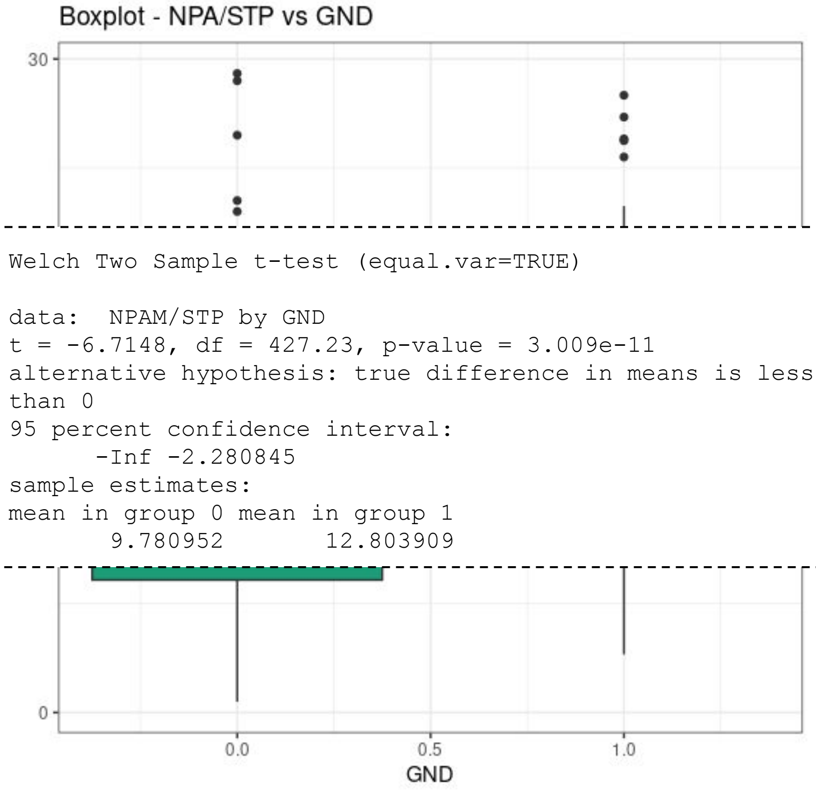


Men attempt (and win) more net points

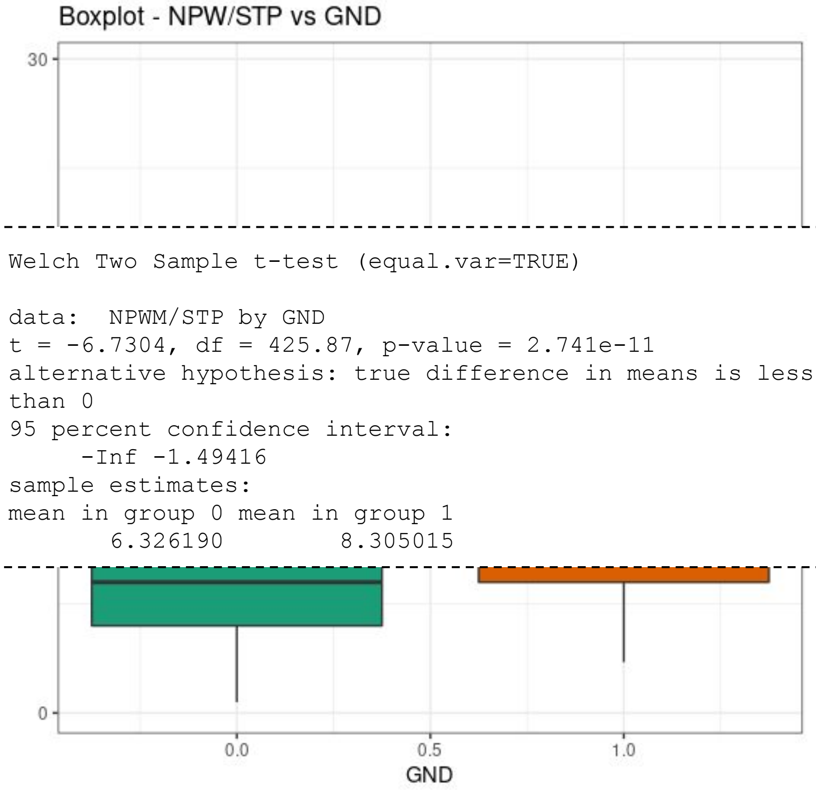




Men attempt (and win) more net points

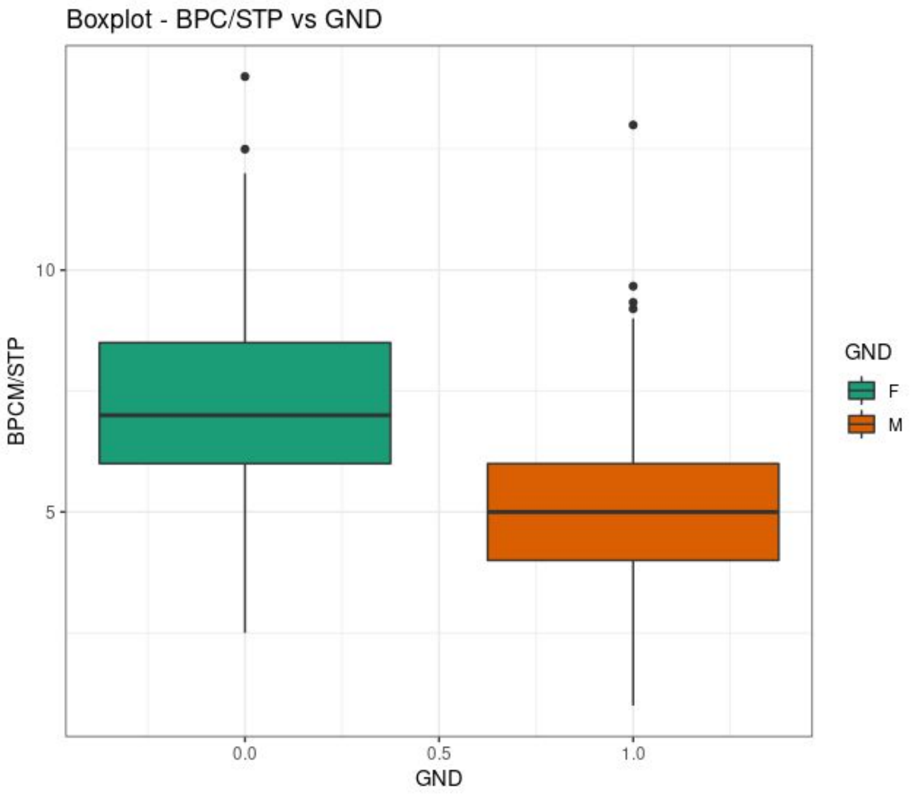



```
Welch Two Sample t-test (equal.var=TRUE)

data:  NPAM/STP by GND
t = -6.7148, df = 427.23, p-value = 3.009e-11
alternative hypothesis: true difference in means is less
than 0
95 percent confidence interval:
      -Inf -2.280845
sample estimates:
mean in group 0 mean in group 1
       9.780952       12.803909
```

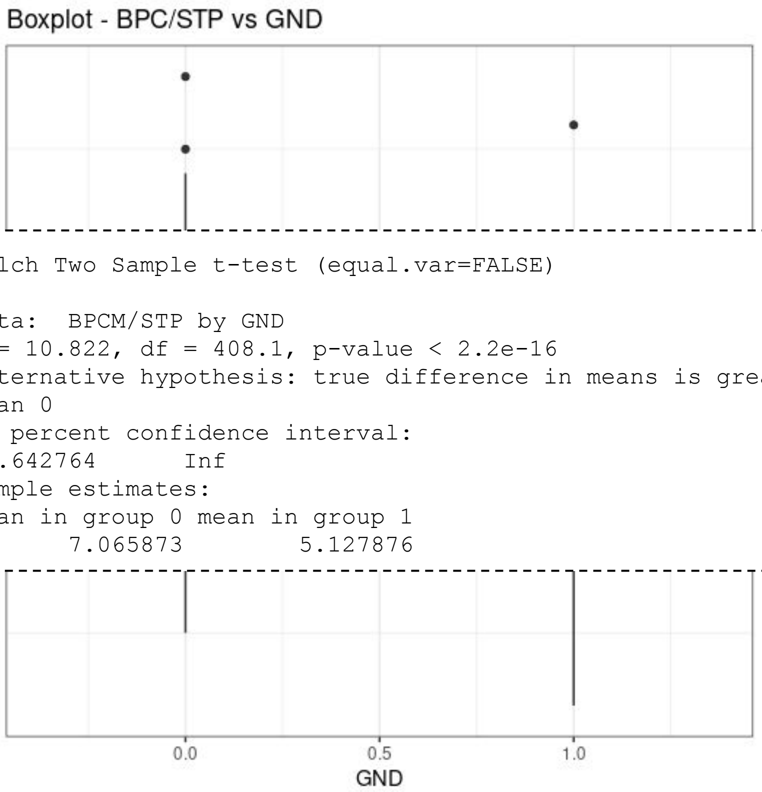



```
Welch Two Sample t-test (equal.var=TRUE)

data:  NPWM/STP by GND
t = -6.7304, df = 425.87, p-value = 2.741e-11
alternative hypothesis: true difference in means is less
than 0
95 percent confidence interval:
     -Inf -1.49416
sample estimates:
mean in group 0 mean in group 1
       6.326190        8.305015
```

Women break more easily

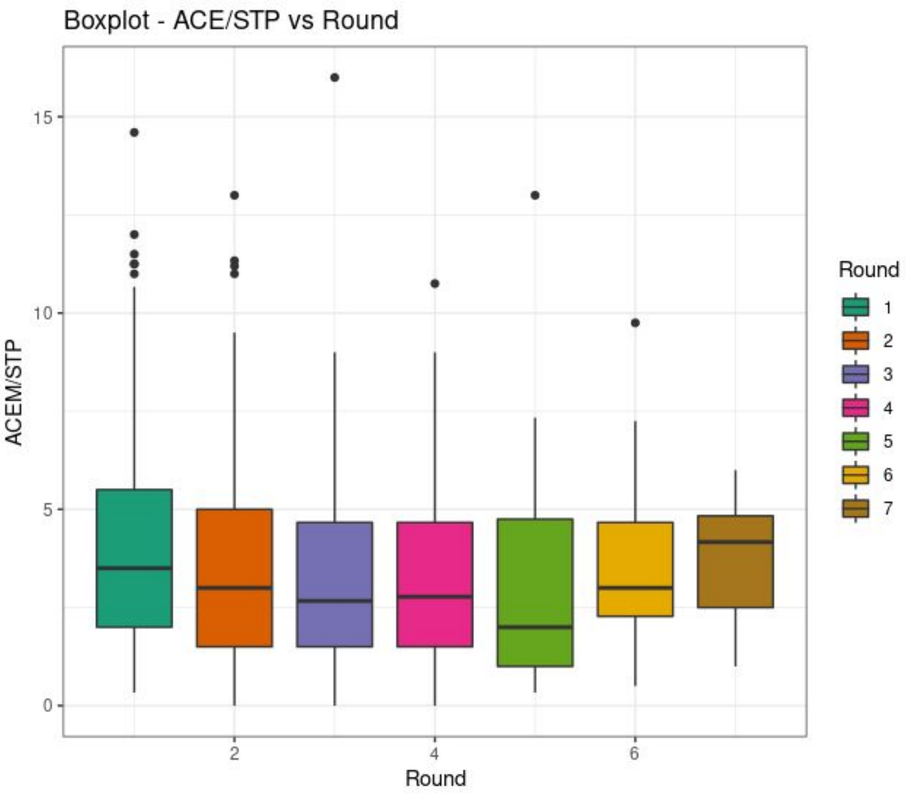


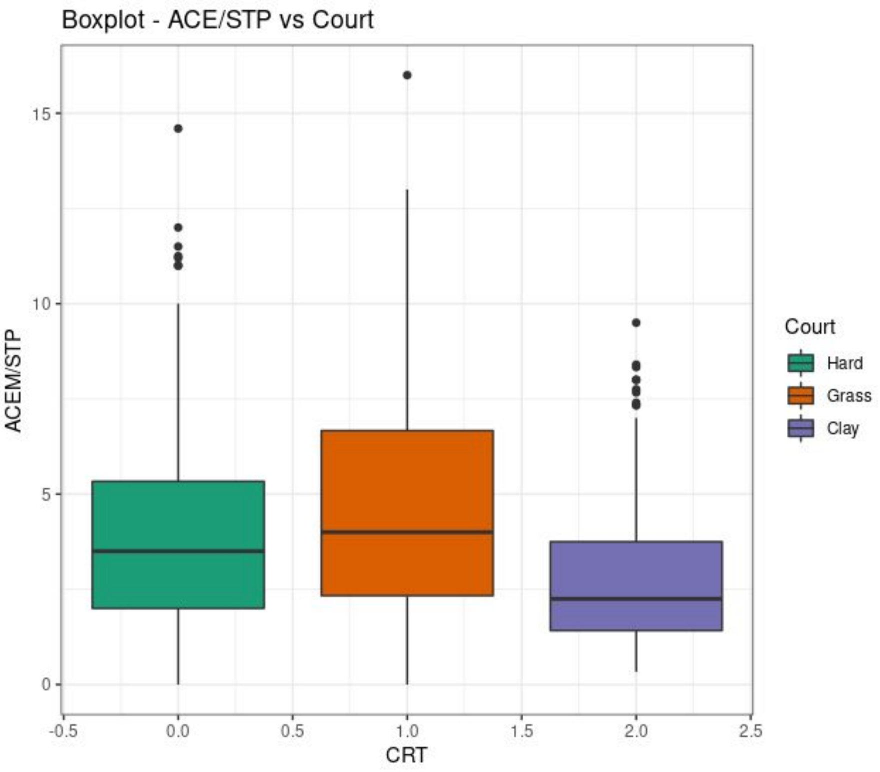
Women break more easily



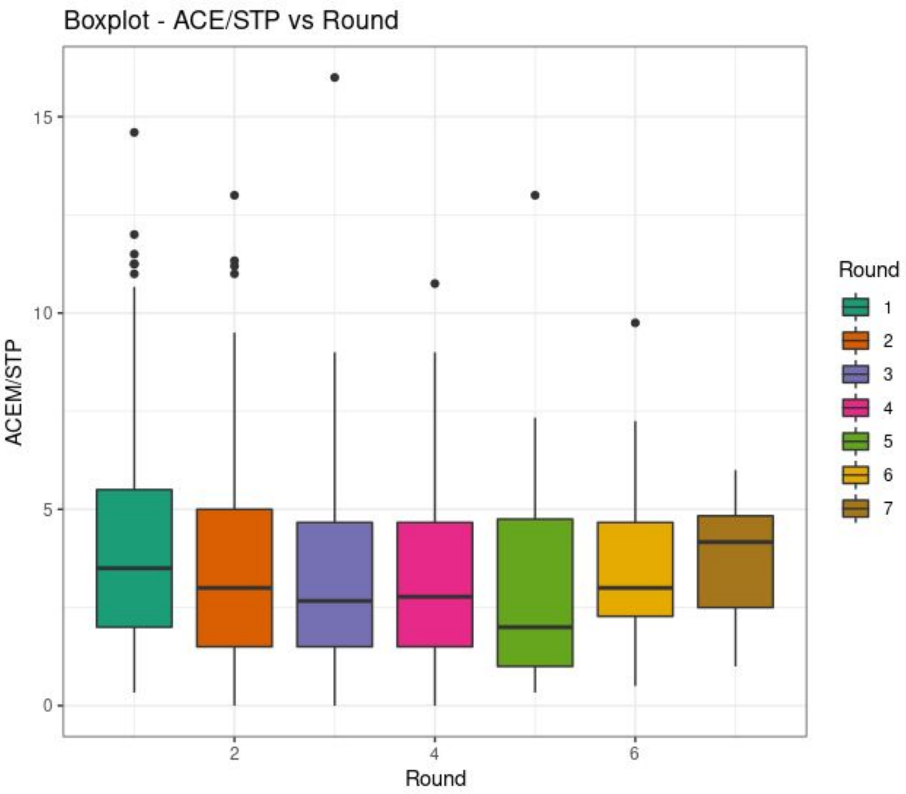

```
Welch Two Sample t-test (equal.var=FALSE)

data:  BPCM/STP by GND
t = 10.822, df = 408.1, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater
than 0
95 percent confidence interval:
 1.642764      Inf
sample estimates:
mean in group 0 mean in group 1
       7.065873        5.127876
```

Do tournament rounds and court kinds matter?





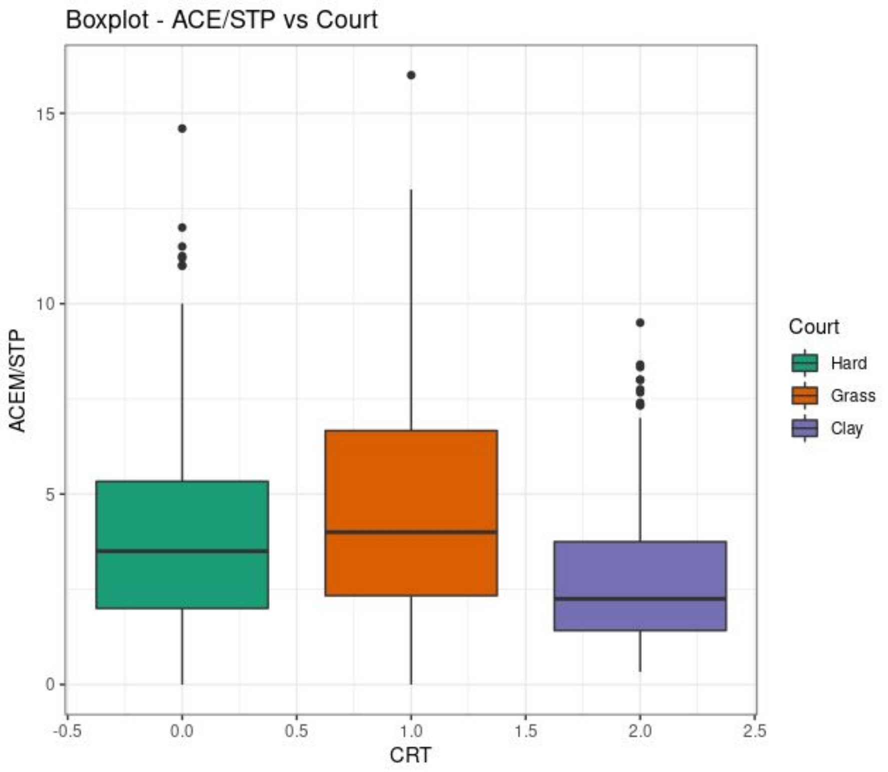
Do tournament rounds and court kinds matter?



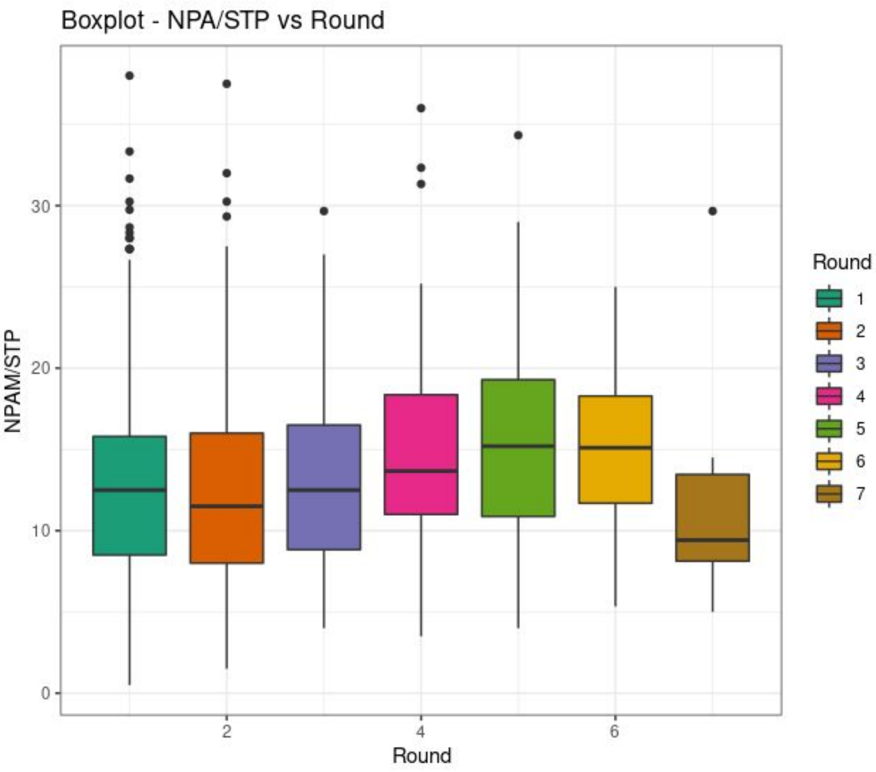

```
                    Shapiro-Wilk normality test

data:  selected_round$ACEM/selected_round$STP
Round  1
W = 0.94113, p-value = 1.731e-13

Round  2
W = 0.92055, p-value = 1.026e-08

Round  3
W = 0.86441, p-value = 2.335e-08

Round  4
W = 0.91662, p-value = 0.001116

Round  5
W = 0.80108, p-value = 0.0001429

Round  6
W = 0.89061, p-value = 0.08248

Round  7
W = 0.9462, p-value = 0.7095


        Bartlett test of homogeneity of variances

data:  ACEM/STP by Round
Bartlett's K-squared = 6.2725, df = 6, p-value = 0.3934
```

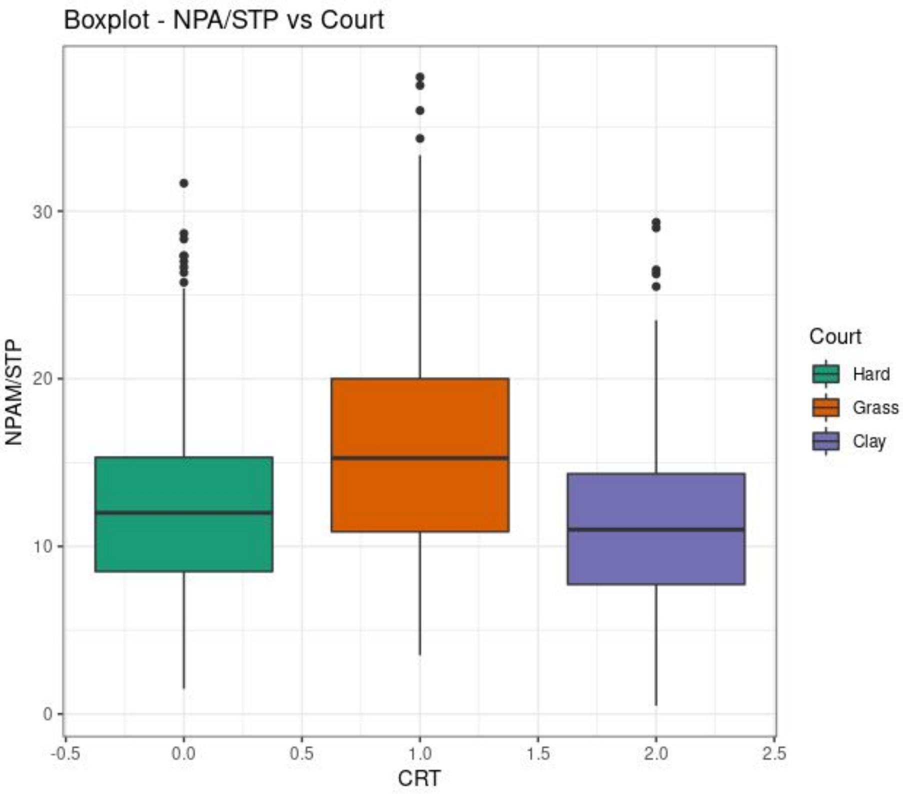
Do tournament rounds and court kinds matter?

```
                  Shapiro-Wilk normality test

data:  selected_crt$ACEM/selected_crt$STP
Court  0
W = 0.94572, p-value = 9.29e-12

Court  1
W = 0.94974, p-value = 3.334e-07

Court  2
W = 0.90331, p-value = 2.735e-11


         Bartlett test of homogeneity of variances

data:  ACEM/STP by CRT
Bartlett's K-squared = 42.251, df = 2, p-value = 6.687e-10
```

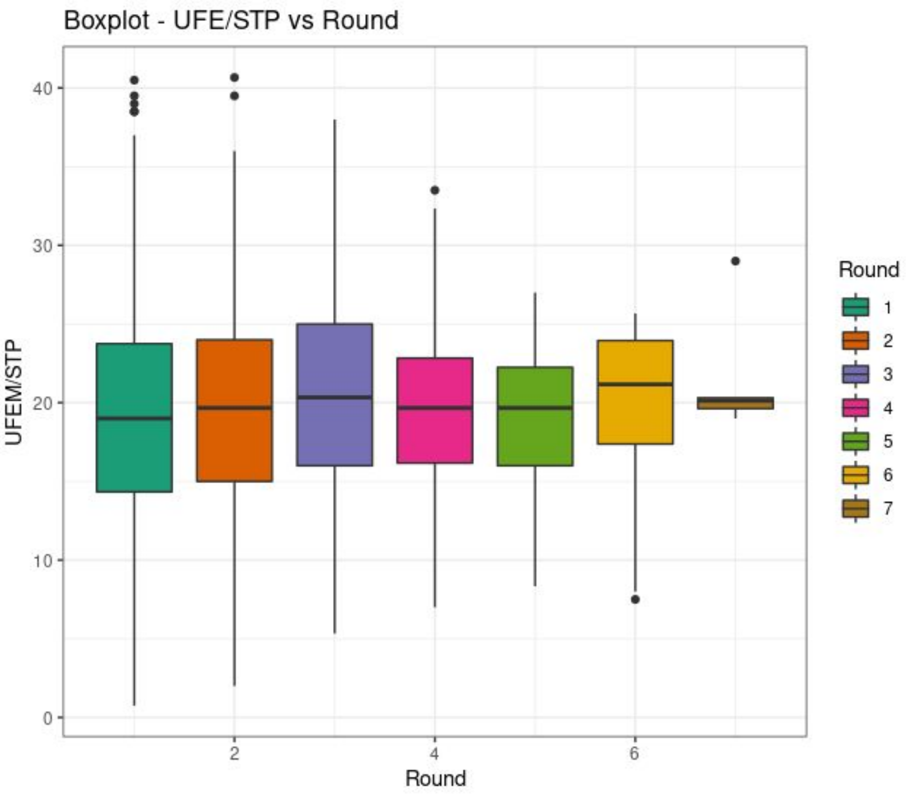


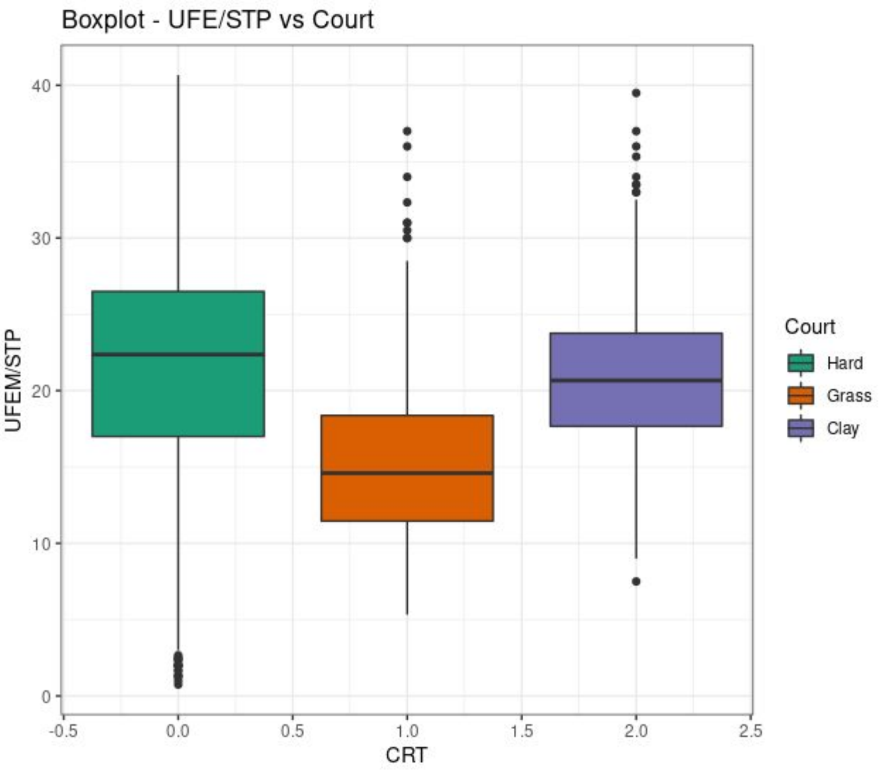
Do tournament rounds and court kinds matter?



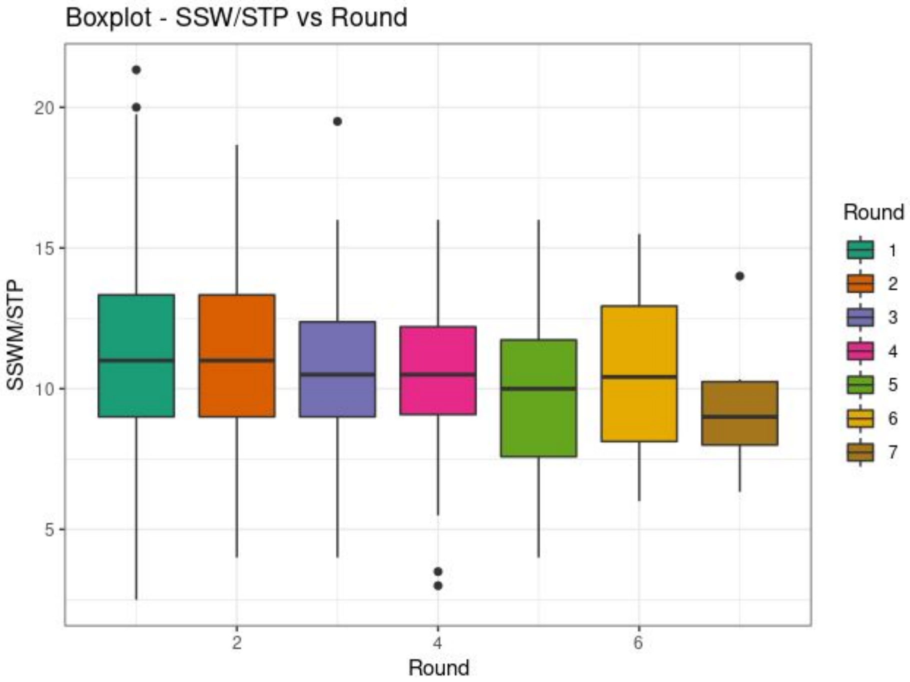


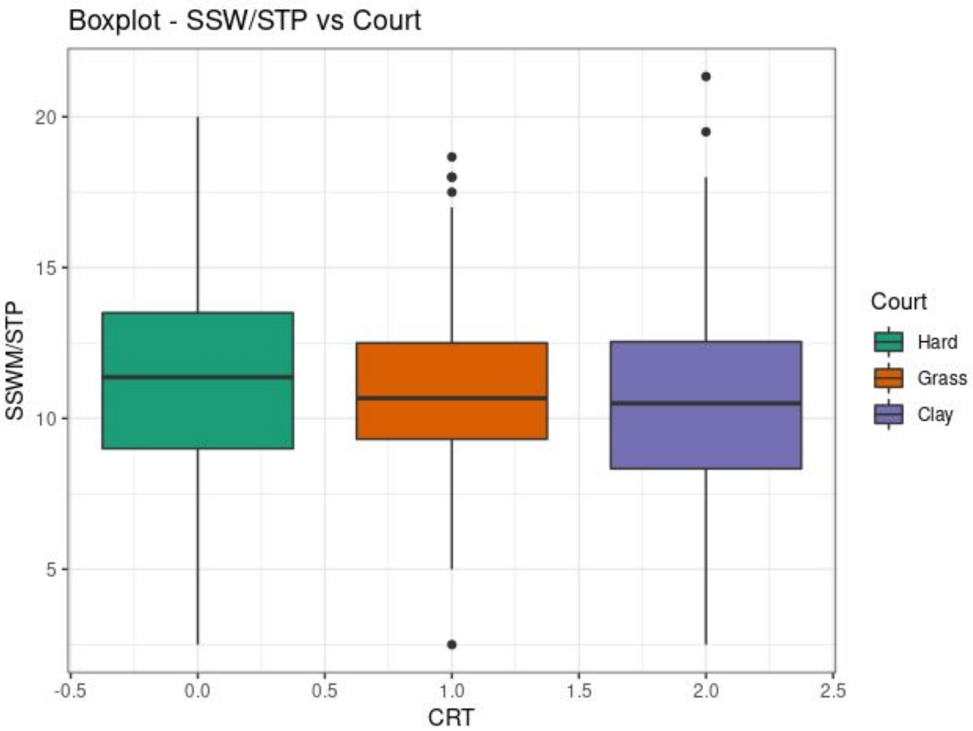
Do tournament rounds and court kinds matter?





Do tournament rounds and court kinds matter?





Technical Appendix

- Principal Components Analysis
- Linear Regression
- Logistic Regression
- k-Nearest Neighbors

Technical Appendix - Principal Components Analysis

Given a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, containing the realizations of a random vector $\boldsymbol{X}$, Principal Components Analysis is the process of finding a vector $\boldsymbol{a} \in \mathbb{R}^{m}$ s.t. the **variance** of $A\boldsymbol{a}$ is maximised.

The projected matrix is called a *principal component*.

$$var[X] = \Sigma \ \Rightarrow var[A\boldsymbol{a}] = \boldsymbol{a}^{\mathrm{T}}\Sigma\, \boldsymbol{a}$$

The solution to the maximisation problem is the eigenvalue λ_k corresponding to the eigenvector $\boldsymbol{a}_k$ along which the projection is performed.

$$\max_{\boldsymbol{a}_k} \boldsymbol{a}_k^{\mathrm{T}}\Sigma\, \boldsymbol{a}_k = \lambda_k \quad \|\boldsymbol{a}_k\| = 1,\ a_k \perp a_j\ \forall j \neq k$$

$tr(\boldsymbol{\Sigma})$ is called the total variance of $\boldsymbol{A}$. Hence, the ratio between λ_k and the trace gives the fraction of total variance explained by the k-th component.

$$\frac{\sum_i^k \lambda_i}{tr(\Sigma)}$$

Technical Appendix - Linear Regression

In simple and multiple linear regression, we model one dependent variable $\mathbf{Y}$ as the linear combination of one or more predictors $\boldsymbol{x}_1$, …, $\boldsymbol{x}_n$ plus an intercept $\boldsymbol{\beta}_0$ and a random error ε.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_n x_{in} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2),\ \forall i$$

The parameters of the model are fitted minimizing the Residual Sum of Squares error measure.

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

The R^2 is a measure of fitting of the model to the training data. It captures how well the predictors are able to explain the variability contained in the target values.

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

Technical Appendix - Linear Regression

To compare multiple linear regression models that involve different predictors, **indirect methods** can be used to approximate the test-set error of the models by adjusting the training error in different ways.

C_p and AIC share the same structure, i.e. a first part depending on RSS and a second part accounting for the model complexity.

Since BIC replaces the penalisation 2*d* of the AIC with a log(*n*)*d*, $\forall n > 7$ the BIC statistics generally places a heavier penalty on models with many variables and hence the results in the selection of smaller models than AIC and C_p do.

Assuming d is the number of parameters:

$$\bar{R}^2 = 1 - \frac{\mathrm{RSS}/(n-d-1)}{\mathrm{TSS}/(n-1)}$$

$$C_p = \frac{1}{n}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$$

$$\mathrm{AIC} = n \log\left(\frac{\mathrm{RSS}}{n}\right) + 2d$$

(assuming Gaussian errors)

$$\mathrm{BIC} = n \log\left(\frac{\mathrm{RSS}}{n}\right) + \log(n)\, d$$

(assuming Gaussian errors)

Technical Appendix - Logistic Regression

In logistic regression, we assume that the dependent variable we want to model follows a Bernoulli distribution. We then attempt to model the π parameter describing the distribution.

$$Y_i \sim \mathrm{Ber}(\pi_i)$$

To adapt the linear form used in (multiple) linear regression to this modeling framework, a ***link function*** is used (logit function).

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_n x_{in}$$

The parameters of the model are estimated maximizing the *log-likelihood*.

$$\ell(\beta; y) = \sum_{i=1}^{n} \left(y_i \log(\pi_i) + (1-y_i) \log(1-\pi_i) \right)$$

Technical Appendix - k-Nearest Neighbors

In k-Nearest Neighbors, a non-parametric approach, no assumption is made on the form of the function used to model the response variable.

The classification process is based on a **distance measure** computed between the target instance and all other instances in the dataset. Then, the majority label among the *k* points with minimal distance is chosen as predicted target label.

Algorithm 1 KNN algorithm

Input: $\mathbf{x}, S, d$
Output: class of $\mathbf{x}$
for $(\mathbf{x}', l') \in S$ **do**
  Compute the distance $\mathrm{d}(\mathbf{x}', \mathbf{x})$
end for
Sort the $|S|$ distances by increasing order
Count the number of occurrences of each class l_j among the k nearest neighbors
Assign to $\mathbf{x}$ the most frequent class

$$d(\boldsymbol{x}', \boldsymbol{x}) = \sqrt{\sum_{i=1}^{m} \left(x'_i - x_i \right)^2}$$

Thank you for your attention!