TRINITY COLLEGE DUBLIN
COMPUTER SCIENCE - DATA SCIENCE

CS7DS3 – APPLIED STATISTICAL MODELLING
FINAL ASSIGNMENT

SUBMITTED BY
KAAVIYA KARUNANIDHI
21331515

Introduction:

1. A chess dataset is analysed using statistical methods which contains records of player performance across all matches played in the chess world championship from 1866 to 2021 tournament. Player performance is rated in terms of ACPL (average centipawn loss), the ratings performed by specialist chess engines. Lower ACPL values indicates better performance and a loss of 100 centipawns can be interpreted as losing a pawn without compensation and moves may also be penalised based on poor positional play.

2. Linear Regression model is built to show the performance of players have improved over time or not, and how their performance has been influenced by the development of chess engines, which have surpassed the best human players since the famous Kasparov vs. Deep Blue challenge in 1996.

The statistical modelling of the dataset is done using R and the R Script is given in the github link - https://github.com/Karunank/Applied-Statistical-Modelling

Solution 1 a).

Handling of Data:

- Since we are evaluating only the white pieces, only some columns of the dataset is considered – White.Player, White.ACPL and White.Player_ID
- Chess Analysis Dataset is loaded in R and data cleaning is performed in order to find any missing data or data outliers
- We are considering only two players - Viswanathan Anand, and Magnus Carlsen in our case, so the White.Player column is filtered to only for these two players in order to compare their performance
- The filtered data frame is used for the analysis which is shown below in detail

Analysis:

- In order to understand the data in a better way, the summary of the filtered data frame containing only two players is shown below:

```
> summary(df3)
              White.Player    White.ACPL      White.Num.Moves White.Player_ID
 Anand, Viswanathan :46    Min.   : 3.385   Min.   :15.00   Min.   :2.00
 Carlsen, Magnus    :28    1st Qu.: 8.284   1st Qu.:28.25   1st Qu.:2.00
 Alekhine, Alexander: 0    Median :11.338   Median :38.50   Median :2.00
 Aronian,L          : 0    Mean   :12.849   Mean   :42.69   Mean   :4.27
 Bogoljubow, Efim   : 0    3rd Qu.:15.480   3rd Qu.:55.00   3rd Qu.:8.00
 Botvinnik, Mikhail : 0    Max.   :44.963   Max.   :99.00   Max.   :8.00
 (Other)            : 0
>
```

- From the obtained summary of the data, White.ACPL follows a skewed distribution with minimum value of 3.385 and maximum of 44.963 and the median is 11.338
- A box plot is shown below in Figure 1 for the two players with jittered data on their White.ACPL

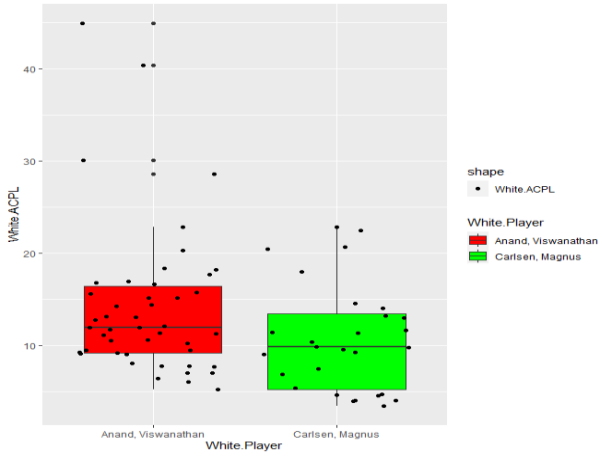

Figure 1. Box plot for White ACPL and White Player for two groups

- In Figure 1, it can be seen that the Anand, Viswanathan is represented by red colour area and the Carlsen, Magnus is represented by green colour
- The data is positively skewed for Viswanathan ACPL as the median is closer to the bottom quartile (Q1) and negatively skewed for Magnus ACPL as the median is closer to the upper quartile (Q3)
- Summary of data with **Carlsen, Magnus** is given below:

```
> summary(df3_CM)
             White.Player    White.ACPL      White.Num.Moves White.Player_ID
 Carlsen, Magnus    :28    Min.   : 3.385   Min.   :15.00   Min.   :8
 Alekhine, Alexander: 0    1st Qu.: 5.197   1st Qu.:36.25   1st Qu.:8
 Anand, Viswanathan : 0    Median : 9.801   Median :47.00   Median :8
 Aronian,L          : 0    Mean   :10.703   Mean   :50.00   Mean   :8
 Bogoljubow, Efim   : 0    3rd Qu.:13.403   3rd Qu.:57.00   3rd Qu.:8
 Botvinnik, Mikhail : 0    Max.   :22.843   Max.   :99.00   Max.   :8
 (Other)            : 0
>
```

- Summary of data with **Anand, Viswanathan** is given below:

```
> summary(df3_AV)
             White.Player    White.ACPL      White.Num.Moves White.Player_ID
 Anand, Viswanathan :46    Min.   : 5.156   Min.   :15.00   Min.   :2
 Alekhine, Alexander: 0    1st Qu.: 9.177   1st Qu.:24.50   1st Qu.:2
 Aronian,L          : 0    Median :11.907   Median :33.50   Median :2
 Bogoljubow, Efim   : 0    Mean   :14.155   Mean   :38.24   Mean   :2
 Botvinnik, Mikhail : 0    3rd Qu.:16.374   3rd Qu.:48.50   3rd Qu.:2
 Bronstein, David I : 0    Max.   :44.963   Max.   :82.00   Max.   :2
 (Other)            : 0
```

- From the above data summary, the average White ACPL for Anand, Viswanathan is 14.15 and for Carlsen, Magnus is 10.7 and the median of white ACPL for Anand, Viswanathan and Carlsen, Magnus is 11.907 and 9.801 respectively and the standard deviation for Viswanathan and Magnus is 8.159567 and 5.879884 respectively
- Since the number of records between two groups are different we cannot conclude the performance with the mean difference, so we conduct a t-test for the above dataset in order to reject or accept null hypothesis
- Two Sample T-test is conducted on the dataset containing two players in order to reject/accept the null hypothesis which is given below:

```
> #t test between two groups
> t.test(White.ACPL ~ White.Player, data=df3, var.equal = FALSE)

        Welch Two Sample t-test

data:  White.ACPL by White.Player
t = 2.1078, df = 69.829, p-value = 0.03864
alternative hypothesis: true difference in means between group Anand, Viswanathan and group Carlsen, Magnus is not equal to 0
95 percent confidence interval:
 0.1854561 6.7183767
sample estimates:
mean in group Anand, Viswanathan    mean in group Carlsen, Magnus
                        14.15502                         10.70310
```

- From the above t-test we can infer that the null hypothesis is rejected and there is 95% confidence interval for the true difference in the mean between two players is [0.1854561, 6.7183767]

- In order to explicitly model the difference between two mean scores of each player we are using Bayesian model to compare two players (cmp_two_players)
- Gibbs sampling is used in which each parameter is proposed separately and the conditional posterior distribution is used as its proposal and the acceptance proposal here is always equal to 1. In order to make the predictions on unobserved data Gibbs sampler is used to predict the probability of who is performing better than the other.

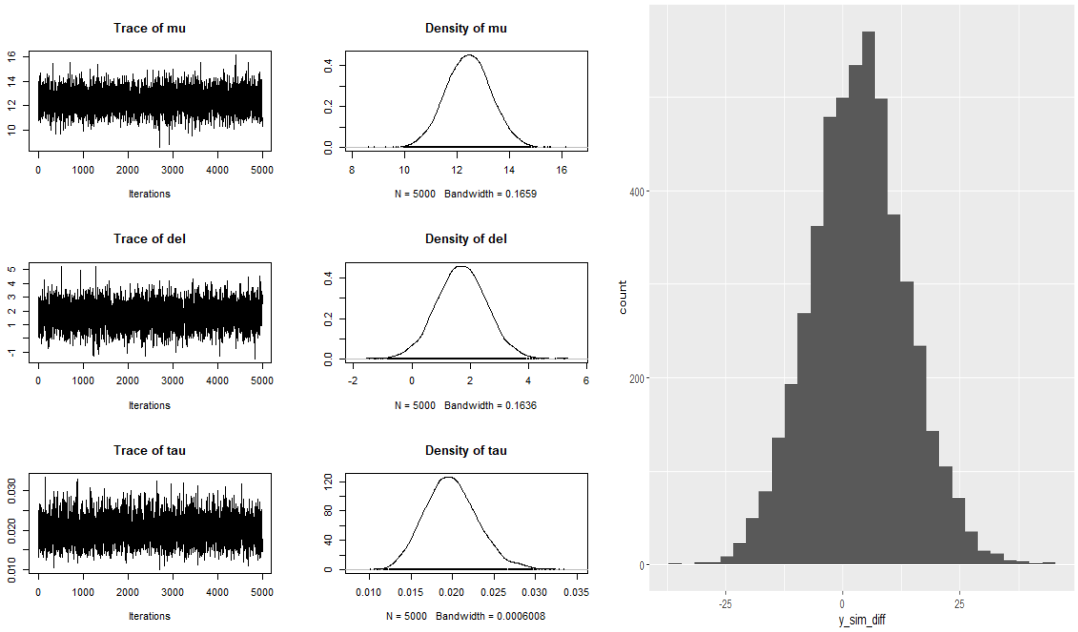

Figure 2. Basic Properties of Posterior Distribution

- By default the function sets hyperparameters μ0=50 and σ0=20, i.e., τ0=1/400 for μ. The hyperparameters for d are set to be δ0=0 and γ0=1/400. For τ, the default is a0=1 and b0=50
- The default hyperparameter used for Gibbs sampling – mu, del, tau
  Here mu = mean of the two defined groups which is equal to 12.8,
  tau = precision 0.0177
  b0 = 0.2233
  a0 = 2.858
- After generating samples from Gibbs Method, MCMC fit method is used to create plots and to analyse the convergence of the parameters (del, mu and tau) which is shown in Figure 2. This shows the normal distribution of posterior means where the maximum probability density of ACPL points is near ~ 12.9
- The Gibbs Sampler performance for comparing two groups is shown in Table 1 where Dependence factor and Burn in period is less and has a value closer to 1 indicating the Gibbs sampler performance is satisfactory

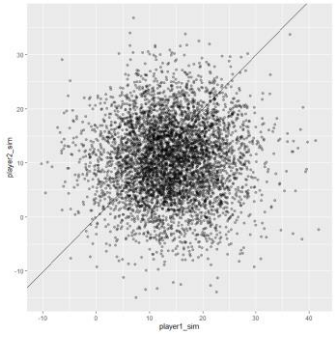

Figure 3. Compare probability between two players

```
> raftery.diag(as.mcmc(fit))

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

     Burn-in  Total Lower bound  Dependence
     (M)      (N)   (Nmin)       factor (I)
 mu  2        3805  3746         1.02
 del 2        3803  3746         1.02
 tau 2        3803  3746         1.02
```

```
> apply(fit, 2, mean)
         mu         del         tau
12.42684817  1.71175049  0.01989296
> apply(fit, 2, sd)
         mu         del         tau
0.859637326 0.847800143 0.003179279
> mean(1/sqrt(fit[, 3]))
[1] 7.158603
> sd(1/sqrt(fit[, 3]))
[1] 0.5784523
```

```
> mean(player1_sim > player2_sim)
[1] 0.6368
```

Table 1. Gibbs Sampler Performance when comparing Two groups

Conclusion:

From the above analysis and data summaries, we can conclude that the difference in average performance values of both the payers is equal to 0.6368 and therefore there is 0.6368 probability that Carlsen, Magnus performs better than Anand, Viswanathan.

Solution 1 b).

Analysis:

- Instead of only two players which we used in the last part, here we use all the White players from the analysis dataset
- In order to compare all the White Players, summary of the data frame with all white players is shown below:

```
> summary(df2)
             White.Player    White.ACPL      White.Num.Moves White.Player_ID
 Kasparov, Gary     : 99   Min.   :  2.35   Min.   : 9.00   Min.   : 1.0
 Karpov, Anatoly    : 97   1st Qu.: 10.89   1st Qu.:31.00   1st Qu.: 7.0
 Botvinnik, Mikhail : 88   Median : 16.36   Median :40.00   Median :18.0
 Alekhine, Alexander: 70   Mean   : 20.23   Mean   :42.27   Mean   :17.8
 Steinitz, William  : 57   3rd Qu.: 26.23   3rd Qu.:51.00   3rd Qu.:27.0
 Lasker, Emanuel    : 56   Max.   :114.20   Max.   :99.00   Max.   :39.0
 (Other)            :565
```

- The box plot for all the White Players and their ACPL mean values are shown below in Figure 4. From the box plot, it can be seen that Karjakin, Sergey has the lowest ACPL mean value which is 7.427 and Chigorin, Mikhail has the highest mean ACPL of 45.9
- Figure 5 shows the count of each White players who played matches in different years where we can see that Kasparo, Gary has played a large number of matches in total and Schlechter, Carl has played only during the year 1910 who played the least number of matches among other players
- Figure 6 shows the histogram of ACPL values among different players with the highest number of players having ACPL values in the range 10-20 and Figure 7 shows the scatter plot where ACPL points shifts towards mean when sample size is increased
- Here the sample of all the players are small in size so predicting the probability of better players is difficult
- Gibbs sampling technique using MCMC (Markov Chain Monte Carlo) method is used to find the marginal distribution of all the players by simulating the posterior parameters which is derived from joint probability distribution
- The default hyperparameter used for Gibbs sampling – mu, del, tau
  Here mu = mean of the all defined groups which is equal to 20.23,
  tau = precision 0.0056
  b0 = 0.114

a0 = 2.306

- Sorted mean ACPL for white players is shown below in Figure 8 where we can see a linear relationship between the white players and their ACPL values
- The Gibbs Sampler performance for comparing multiple players is shown in Table 2 where Dependence factor and Burn in period is less and has a value closer to 1 indicating the Gibbs sampler performance is satisfactory

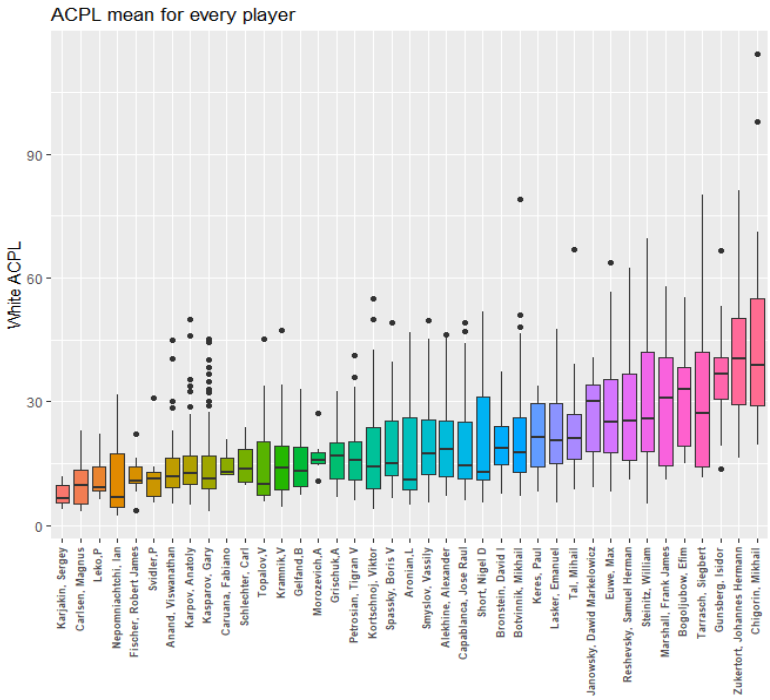

Figure 4. Box plot for White ACPL and every White Player

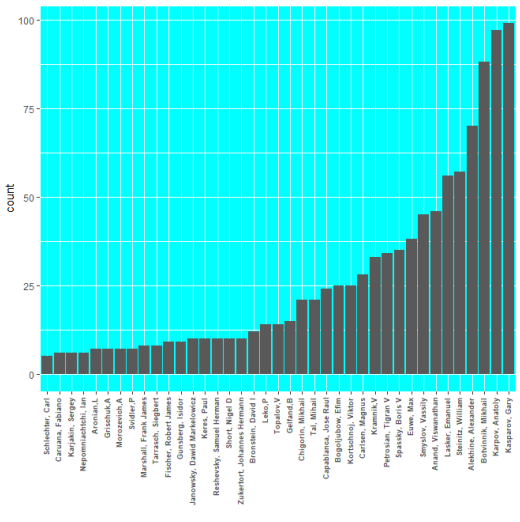

Figure 5. Matches played by White players and their count

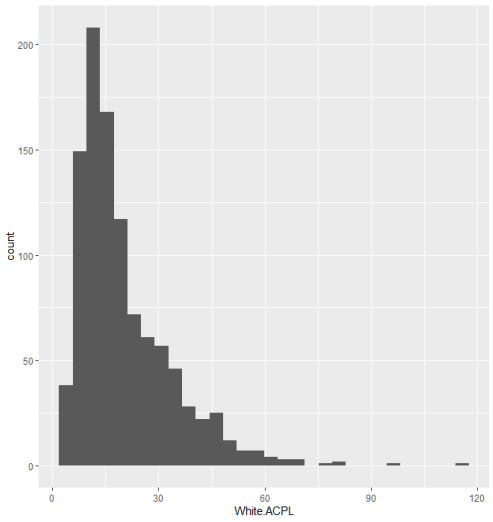

Figure 6. Range and Frequency of White ACPL among Players

```
> apply(fit$params, 2, mean)
        mu       tau_w       tau_b
21.02573985  0.00764793  0.01597045
> apply(fit$params, 2, sd)
          mu        tau_w        tau_b
1.3823230626 0.0003395634 0.0079730448
> mean(1/sqrt(fit$params[, 3]))
[1] 8.152083
> sd(1/sqrt(fit$params[, 3]))
[1] 1.100807
```

```
> raftery.diag(as.mcmc(fit$params))

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

       Burn-in  Total Lower bound  Dependence
       (M)      (N)   (Nmin)       factor (I)
mu     2        3995  3746         1.07
tau_w  2        3837  3746         1.02
tau_b  3        4129  3746         1.10
```

Table 2. Gibbs Sampler Performance when comparing M groups

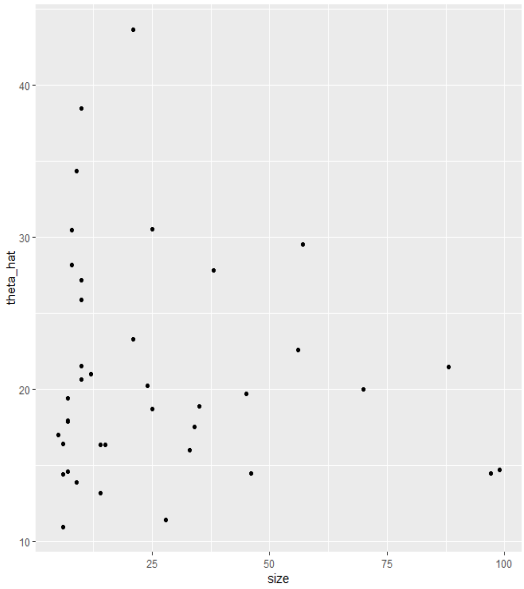

Figure 7. Range and Frequency of White ACPL among Players

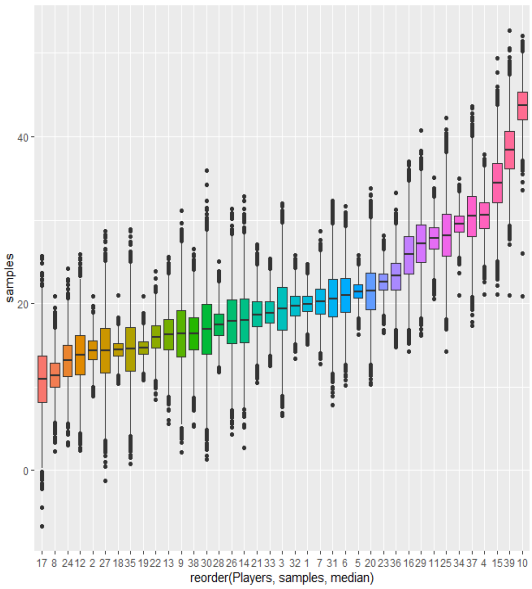

Figure 8. Sorted mean ACPL for White Players from generated samples

Conclusion:

From the above analysis, it is clear that some group of players always perform well when compared to some set of players as shown below:

Players performing in a superior manner – **(i) Karjakin, Sergey (ii) Carlsen, Magnus (iii) Leko,P (iv) Fischer, Robert James (v) Anand, Viswanathan (vi) Nepomniachtchi, Ian (vii) Karpov, Anatoly**

Players performing in an inferior manner – **(i) Marshall, Frank James (ii) Steinitz, William (iii) Tarrasch, Siegbert (iv) Bogoljubow, Efim (v) Gunsberg, Isidor (vi) Zukertort, Johannes Hermann (vii) Chigorin, Mikhail**

Solution 2a)

Linear Regression is used to model the players performance over the next 10 years in order to find out how their performance is influenced by chess engines.

Data Analysis:

- Data frame consisting of columns from Year to PreDeepBlue is initially used in order to select the features to perform linear regression model. The data frame is scaled for all the numeric variables in the dataset and the PreDeepBlue columns values containing True is changed to 1 and containing False is changed to 0 in order to make the model learn and understand the problem easily. Summary of the scaled chess data is shown below:

```
> summary(scaled_chess)
      Year.V1          Game.Number.V1       White.ACPL.V1     White.Num.Moves.V1     Black.ACPL.V1
 Min.   :-1.9303660   Min.   :-1.216618   Min.   :-1.342242   Min.   :-2.022610   Min.   :-1.326962
 1st Qu.:-0.6733150   1st Qu.:-0.726284   1st Qu.:-0.701065   1st Qu.:-0.685004   1st Qu.:-0.778155
 Median : 0.0599647   Median :-0.235950   Median :-0.290634   Median :-0.137802   Median :-0.264104
 Mean   : 0.0000000   Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.000000
 3rd Qu.: 0.7146787   3rd Qu.: 0.450518   3rd Qu.: 0.450786   3rd Qu.: 0.531001   3rd Qu.: 0.565814
 Max.   : 1.6050899   Max.   : 4.177057   Max.   : 7.055566   Max.   : 3.449412   Max.   : 6.259480
 Black.Num.Moves.V1   Combined.ACPL.V1    Black.Player_ID.V1  White.Player_ID.V1   PreDeepBlue
 Min.   :-2.049181   Min.   :-1.449275   Min.   :-1.4962121   Min.   :-1.4986281   Min.   :0.000
 1st Qu.:-0.712382   1st Qu.:-0.788870   1st Qu.:-0.9593238   1st Qu.:-0.9633235   1st Qu.:1.000
 Median :-0.165510   Median :-0.174465   Median : 0.0249716   Median : 0.0180683   Median :1.000
 Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.0000000   Mean   : 0.0000000   Mean   :0.811
 3rd Qu.: 0.563653   3rd Qu.: 0.568910   3rd Qu.: 0.7631930   3rd Qu.: 0.8210252   3rd Qu.:1.000
 Max.   : 3.480305   Max.   : 6.718510   Max.   : 1.9040808   Max.   : 1.8916344   Max.   :1.000
```

- The correlation matrix is used to find the relation between other variables as shown below in order to understand the relationship between features for our linear regression model.

```
> correlationMatrix
                      Year  Game.Number  White.ACPL White.Num.Moves  Black.ACPL Black.Num.Moves Combined.ACPL
Year            1.00000000  0.07704037 -0.43942412    -0.08016230 -0.41935087     -0.07958958   -0.47589201
Game.Number     0.07704038  1.00000000 -0.03897849    -0.03434445 -0.03494423     -0.03487071   -0.04089610
White.ACPL     -0.43942412 -0.03897849  1.00000000     0.09756079  0.62307924      0.10146867    0.88938993
White.Num.Moves -0.08016230 -0.03434450  0.09756079     1.00000000  0.07208488      0.99954129    0.09337223
Black.ACPL     -0.41935087 -0.03494230  0.62307924     0.07208488  1.00000000      0.06261156    0.91172377
Black.Num.Moves -0.07958958 -0.03487013  0.10146867     0.99954129  0.06261156      1.00000000    0.08988783
Combined.ACPL  -0.47589201 -0.04089610  0.88938993     0.09337223  0.91172377      0.08988783    1.00000000
Black.Player_ID -0.11945020 -0.04668510  0.03174979    -0.02348824  0.09227863      0.03183993    0.08630953
White.Player_ID -0.11216118 -0.04595398  0.05148198    -0.02348824  0.05485891     -0.02303981    0.05910272
PreDeepBlue     -0.63022634  0.00951311  0.23466078     0.01565349  0.20040891      0.01612349    0.24038132
                Black.Player_ID White.Player_ID  PreDeepBlue
Year                -0.11945020     -0.11216118 -0.630226341
Game.Number         -0.04668510     -0.04595398  0.009513118
White.ACPL           0.03174979      0.05148198  0.234660782
White.Num.Moves      0.03174979     -0.02348824  0.015653493
Black.ACPL           0.09227863      0.05485891  0.200408905
Black.Num.Moves      0.03183993     -0.02303981  0.016123488
Combined.ACPL        0.08630953      0.05910272  0.240381322
Black.Player_ID      1.00000000      0.10616530  0.082119600
White.Player_ID      0.10616530      1.00000000  0.081198472
PreDeepBlue          0.08211960      0.08119847  1.000000000
```

- To find the distribution of the variables we are using histogram to plot the frequencies of combined ACPL as shown below in Figure 9. Pairs plot is used in Figure 10 to produce matrix of scatter plots between feature variables and Combined ACPL values and Figure 11 shows the scatter plot between Year and Combined ACPL from which we can infer that the combined ACPL values decreases as years increase. Figure 12 shows the box plot for the same in order to spot any outliers

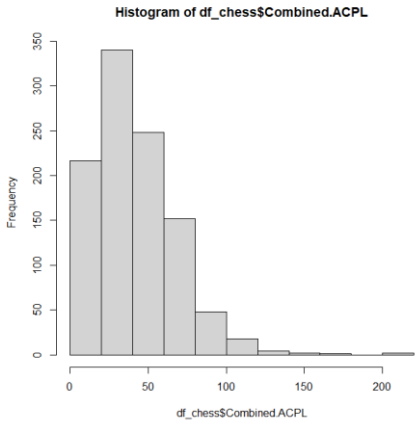

Figure 9. Historgram of Combined ACPL

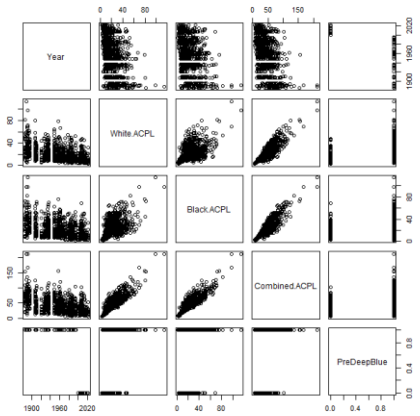

Figure 10. Correlation Plot

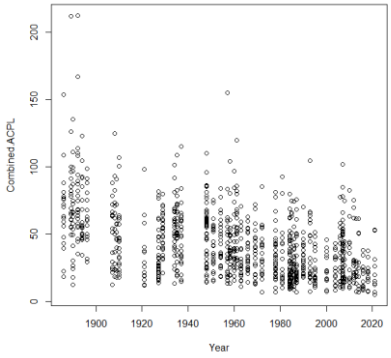

Figure 11. Scatter Plot Year vs Combined ACPL

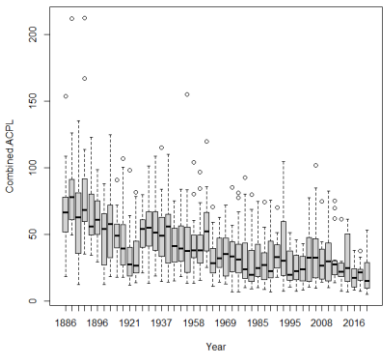

Figure 12. Box Plot of Year and Combined ACPL

- Step AIC method is used to find the best combination of features for the linear regression model which is shown below in Table 2.

```
> step_AIC_forward

Call:
lm(formula = Combined.ACPL ~ Year + PreDeepBlue, data = df_chess)

Coefficients:
(Intercept)         Year   PreDeepBlue
   747.5084      -0.3574       -6.3949
```

Table 2. Step AIC forward for the chess dataset

- Using "Year" and "PreDeepBlue" as feature variables and "Combined.ACPL" as target variable, the linear regression is performed and the summary of the linear regression model is shown below:

```
> summary(lm1)

Call:
lm(formula = Combined.ACPL ~ ., data = df_chess)

Residuals:
    Min      1Q  Median      3Q     Max
-53.513 -15.182  -3.309  11.954 147.370

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 747.737835  47.181025  15.848  < 2e-16 ***
Year         -0.357523   0.023509 -15.208  < 2e-16 ***
Game.Number   0.003778   0.068352   0.055  0.95593
PreDeepBlue  -6.404359   2.285264  -2.802  0.00517 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.25 on 1028 degrees of freedom
Multiple R-squared:  0.2324,    Adjusted R-squared:  0.2301
F-statistic: 103.7 on 3 and 1028 DF,  p-value: < 2.2e-16
```

Table 3. Summary of the Linear Regression model

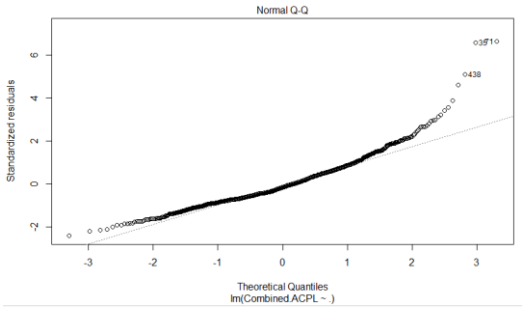

Figure 13. Linear regression model plot

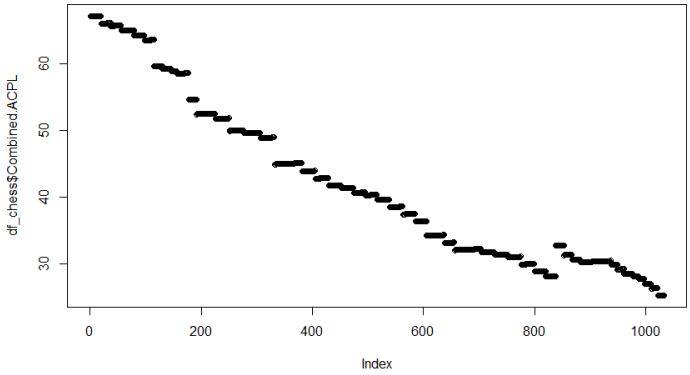

Figure 14. Linear regression Prediction plot

- From the above summary of linear regression model in Table 3, the feature variable "Year" has a coefficient value of -0.3574 with respect to the target variable "Combined.ACPL" indicating that the Combined ACPL values decreases as the years increase and also the model is fit good with the R squared value of 0.2301 which can also be seen from Figure 13. Figure 14 shows the prediction plot for combined ACPL values and it can be seen that the values are decreasing over the next few years

Conclusion:

From above analysis using Linear Regression model, it is clear that the Combined ACPL decreases over the next ten years and thus the player's performance increases.

Solution 2 b)

Here we develop a linear model with two variables with target variable as Combined ACPL and feature variable as PreDeepBlue as shown below:

```
> lm2 <-lm(Combined.ACPL~PreDeepBlue, df_chess)
> summary(lm2)

Call:
lm(formula = Combined.ACPL ~ PreDeepBlue, data = df_chess)

Residuals:
    Min      1Q  Median      3Q     Max
-38.226 -18.295  -4.126  14.171 167.433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   29.329      1.764  16.630  < 2e-16 ***
PreDeepBlue   15.564      1.958   7.948 4.96e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.63 on 1030 degrees of freedom
Multiple R-squared:  0.05778,   Adjusted R-squared:  0.05687
F-statistic: 63.17 on 1 and 1030 DF,  p-value: 4.961e-15
```

From the above summary, we can see that the linear model does not fit properly with the mentioned data as the R-Squared value is very less with only 5% of the data variability. Also, the correlation between the combined ACPL and PreDeepBlue is very less which is **0.2403813** showing that there is no relationship between player’s performance and use of chess engines by players after the year 1996.