R Notebook

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Cmd+Option+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Cmd+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)
library(dplyr)
library(fs)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
NHLALL = read.csv("NHLMega.csv",header=TRUE)
head(NHLALL)
```

```
##   Rk GP  W  L W.L.   G GA DIFF CQF.GP CQF.OT.GP CQF.W.L. CQF.RS.W.L.
## 1  1 36 25  9 0.69 107 75   28      4         1     1.00        0.60
## 2  2 39 27 12 0.69 129 84   45      6         1     0.67        0.75
## 3  3 28 16 12 0.57  68 65    3      6         1     0.67        0.50
## 4  4 31 17 14 0.55  83 84   -1      6         2     0.67        0.50
## 5  5 23 13 10 0.57  55 56   -1      4         2     1.00        0.00
## 6  6 27 17 10 0.63  80 54   26      6         2     0.67        0.00
##   CSF.GP CSF.OT.GP CSF.W.L. CSF.RS.W.L. CF.GP CF.OT.GP CF.W.L. CF.RS.W.L.
## 1      7         2     0.57        0.50     5        3     0.8        0.5
## 2      7         2     0.57        0.67     5        0     0.8        0.6
## 3      4         0     1.00        1.00     5        3     0.2        0.5
## 4      7         3     0.57        0.75     5        0     0.2        0.4
## 5      7         2     0.43        0.33     0        0     0.0        0.0
```

```
## 6      7         3     0.43        0.25     0        0     0.0        0.0
##   F.GP F.OT.GP F.W.L. F.RS.W.L. ELITE.F ELITE.D ELITE.GK Champs Year AvAge
## 1    7       0  0.571         1    4.75    3.75     0.75      8 2001  28.4
## 2    7       0  0.429         1    4.50    5.00     0.75      7 2001  28.5
## 3    0       0  0.000         0    3.75    1.50     0.75      6 2001  28.5
## 4    0       0  0.000         0    3.75    2.25     0.75      5 2001  28.0
## 5    0       0  0.000         0    0.00    4.50     0.75      4 2001  29.4
## 6    0       0  0.000         0    1.00    4.00     1.00      3 2001  28.1
##   Exp
## 1   6
## 2   6
## 3   3
## 4   4
## 5   5
## 6   5
```

```
train=read.csv("Training.csv",header=TRUE)
summary(train)
```

```
##        Rk               GP              W               L
##  Min.   : 1.000   Min.   :17.00   Min.   : 6.0   Min.   : 5.00
##  1st Qu.: 5.000   1st Qu.:24.00   1st Qu.:12.0   1st Qu.: 9.00
##  Median : 9.000   Median :26.00   Median :15.0   Median :11.00
##  Mean   : 8.781   Mean   :27.99   Mean   :15.6   Mean   :10.95
##  3rd Qu.:12.250   3rd Qu.:33.00   3rd Qu.:19.0   3rd Qu.:13.00
##  Max.   :16.000   Max.   :45.00   Max.   :30.0   Max.   :19.00
##  NA's   :779      NA's   :779     NA's   :779    NA's   :779
##       W.L.               G               GA              DIFF
##  Min.   :0.3000   Min.   : 31.00   Min.   : 33.00   Min.   :-24.000
##  1st Qu.:0.5000   1st Qu.: 60.00   1st Qu.: 57.00   1st Qu.: -2.000
##  Median :0.5600   Median : 73.50   Median : 70.00   Median :  6.000
##  Mean   :0.5503   Mean   : 78.30   Mean   : 72.09   Mean   :  6.227
##  3rd Qu.:0.6062   3rd Qu.: 92.75   3rd Qu.: 83.00   3rd Qu.: 14.000
##  Max.   :0.7830   Max.   :149.00   Max.   :125.00   Max.   : 46.000
##  NA's   :779      NA's   :779      NA's   :779      NA's   :779
##      CQF.GP         CQF.OT.GP        CQF.W.L.       CQF.RS.W.L.
##  Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:5.000   1st Qu.:1.000   1st Qu.:0.3300   1st Qu.:0.250
##  Median :6.000   Median :1.000   Median :0.4300   Median :0.500
##  Mean   :5.781   Mean   :1.344   Mean   :0.4755   Mean   :0.487
##  3rd Qu.:6.250   3rd Qu.:2.000   3rd Qu.:0.6670   3rd Qu.:0.690
##  Max.   :7.000   Max.   :5.000   Max.   :1.0000   Max.   :1.000
##  NA's   :779     NA's   :779     NA's   :779      NA's   :779
##      CSF.GP         CSF.OT.GP         CSF.W.L.       CSF.RS.W.L.
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :  0.000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  0.000
##  Median :0.000   Median :0.0000   Median :0.0000   Median :  0.000
##  Mean   :2.695   Mean   :0.6094   Mean   :0.2385   Mean   :  1.026
##  3rd Qu.:6.000   3rd Qu.:1.0000   3rd Qu.:0.4290   3rd Qu.:  0.500
##  Max.   :7.000   Max.   :3.0000   Max.   :1.0000   Max.   :100.000
##  NA's   :779     NA's   :779      NA's   :779      NA's   :779
##      CF.GP          CF.OT.GP         CF.W.L.         CF.RS.W.L.
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000   Median :0.0000   Median :0.0000
```

```
##  Mean   :1.305   Mean   :0.2969   Mean   :0.1168   Mean   :0.1261
##  3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :7.000   Max.   :3.0000   Max.   :1.0000   Max.   :1.0000
##  NA's   :779     NA's   :779      NA's   :779      NA's   :779
##       F.GP          F.OT.GP          F.W.L.         F.RS.W.L.
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :  0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  0.0000
##  Median :0.000   Median :0.0000   Median :0.0000   Median :  0.0000
##  Mean   :0.625   Mean   :0.1094   Mean   :0.0531   Mean   :  0.8086
##  3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:  0.0000
##  Max.   :7.000   Max.   :3.0000   Max.   :0.8000   Max.   :100.0000
##  NA's   :779     NA's   :779      NA's   :779      NA's   :779
##     ELITE.F          ELITE.D          ELITE.GK         Champs
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.00
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.00
##  Median :0.0000   Median :0.250   Median :0.2500   Median :0.00
##  Mean   :0.8012   Mean   :1.285   Mean   :0.4473   Mean   :2.07
##  3rd Qu.:0.5625   3rd Qu.:2.250   3rd Qu.:0.7500   3rd Qu.:4.00
##  Max.   :5.7500   Max.   :5.000   Max.   :1.5000   Max.   :8.00
##  NA's   :779      NA's   :779     NA's   :779      NA's   :779
##       Year          AvAge            Exp
##  Min.   :2001   Min.   :25.90   Min.   :0.000
##  1st Qu.:2006   1st Qu.:27.40   1st Qu.:1.000
##  Median :2010   Median :28.10   Median :3.000
##  Mean   :2009   Mean   :28.28   Mean   :2.828
##  3rd Qu.:2013   3rd Qu.:29.00   3rd Qu.:4.000
##  Max.   :2017   Max.   :32.10   Max.   :9.000
##  NA's   :779    NA's   :779     NA's   :779
```

```
attach(train)
```

#CORRELATION PLOT

```
knitr::opts_chunk$set(echo = TRUE)
mat=cor(NHLALL)
require(corrplot)
```

```
## Loading required package: corrplot
```

```
## corrplot 0.84 loaded
```

```
png(height=1200, width=1500, pointsize=15, file="NHLoverlap.png")
corrplot(mat, method = "color", addCoef.col="grey", order = "AOE")
```

From the correlation plot above, we see that NHL Champions has a relatively strong correlation with F.W.L.(Finals Win Percentage), as expected because the team with the highest Finals Win Percentage will be the NHL Champions. It is also important to notice these pairs of variables with a very high correlation between the independent variables, we cannot overlook them: F.W.L-0.93, GP-0.91, F.GP(Number of Final Games PLayed) and G-0.89(Goals Scored), ELITE.F-0.83,ELITE.D-0.81 and ELITE.GK-0.49. We may need to drop some of these variables when creating the model.

```
train %>% mutate(Champs2=case_when(
  Champs == 8 ~ 3,
  between(Champs, 6, 7) ~ 2,
  between(Champs, 2, 5) ~ 1,
  TRUE ~ 0
)) -> train2
```

```
head(train2)
```

```
##   Rk GP  W  L  W.L.   G  GA DIFF CQF.GP CQF.OT.GP CQF.W.L. CQF.RS.W.L.
## 1  6 25 14 11 0.560  69  54   15      6         0    0.670       0.750
## 2  2 45 23 19 0.511 132 125    7      5         1    0.800       0.833
## 3  9 20 10 10 0.500  43  51   -8      7         1    0.429       0.170
## 4 16 23  7 14 0.300  56  74  -18      4         0    0.000       0.750
## 5 12 26 13 11 0.500  64  69   -5      6         1    0.330       0.670
## 6  4 37 23 14 0.620 120  95   25      7         1    0.570       0.500
##   CSF.GP CSF.OT.GP CSF.W.L. CSF.RS.W.L. CF.GP CF.OT.GP CF.W.L. CF.RS.W.L.
## 1      6         2     0.33        0.60     0        0     0.0       0.00
## 2      7         2     0.57        0.50     5        0     0.8       0.50
## 3      0         0     0.00        0.00     0        0     0.0       0.00
## 4      0         0     0.00        0.00     0        0     0.0       0.00
## 5      0         0     0.00        0.00     0        0     0.0       0.00
## 6      4         1     1.00        0.75     5        0     0.2       0.25
##   F.GP F.OT.GP F.W.L. F.RS.W.L. ELITE.F ELITE.D ELITE.GK Champs Year AvAge
## 1    0       0  0.000         0     0.0    3.75     0.75      3 2003  30.0
## 2    6       2  0.333         1     3.0    0.25     1.50      7 2010  27.7
## 3    0       0  0.000         0     0.0    0.00     0.00      0 2004  27.5
## 4    0       0  0.000         0     0.0    0.00     0.75      0 2008  28.5
## 5    0       0  0.000         0     0.0    0.00     0.00      0 2009  28.7
## 6    0       0  0.000         0     2.5    4.25     0.75      5 2006  28.2
##   Exp Champs2
## 1   2       1
## 2   4       2
## 3   5       0
## 4   6       0
## 5   4       0
## 6   4       1
```

```
half=polr(ordered(Champs2) ~ ELITE.D + ELITE.F + CQF.W.L., data = train2, Hess = TRUE)
summary(half)
```

```
## Call:
## polr(formula = ordered(Champs2) ~ ELITE.D + ELITE.F + CQF.W.L.,
##     data = train2, Hess = TRUE)
##
## Coefficients:
##           Value Std. Error t value
## ELITE.D  0.5195     0.2285   2.273
## ELITE.F  1.1776     0.2741   4.296
## CQF.W.L. 7.4846     1.6496   4.537
##
## Intercepts:
##     Value   Std. Error t value
## 0|1  5.3465  0.9204     5.8089
## 1|2  9.6942  1.3218     7.3339
## 2|3 13.1031  1.6817     7.7917
##
## Residual Deviance: 117.0494
## AIC: 129.0494
## (779 observations deleted due to missingness)
```

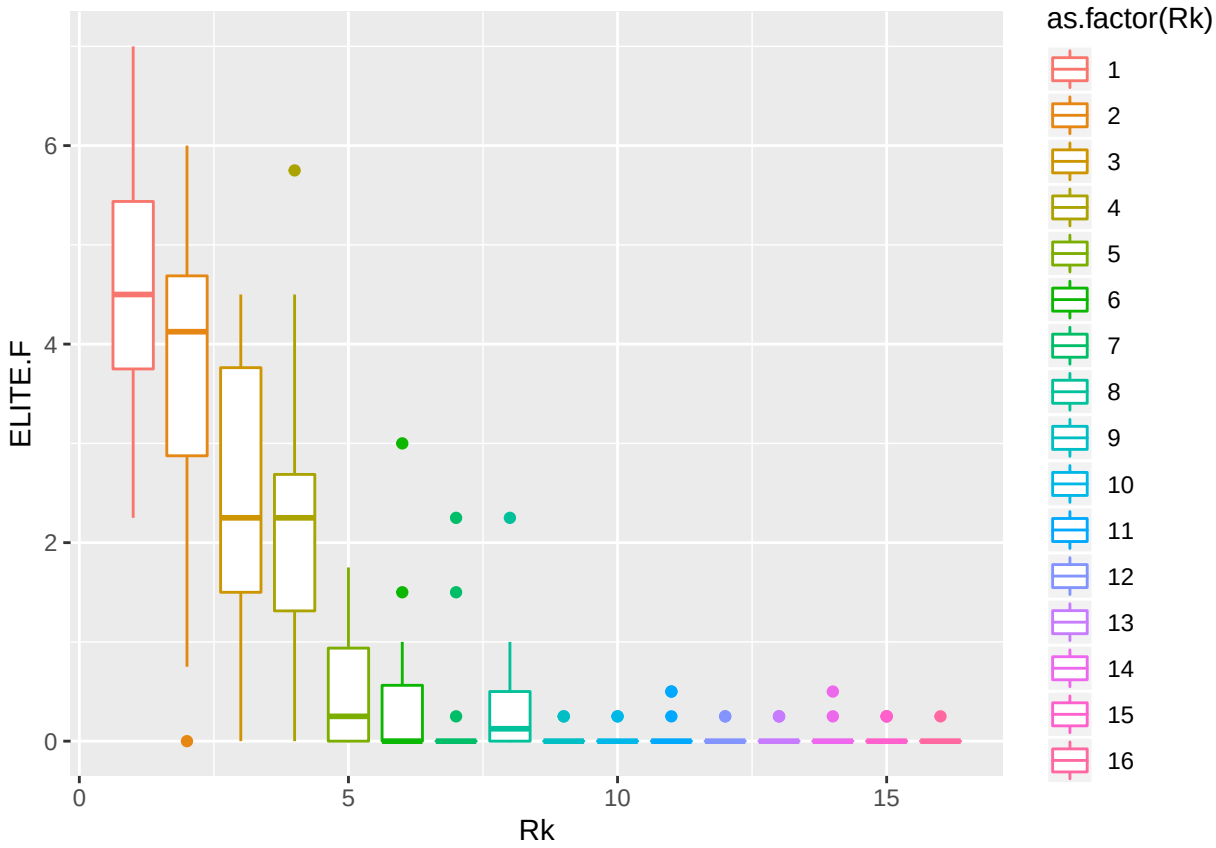
```
drop1(half, test="Chisq")
```

```
## Single term deletions
##
## Model:
## ordered(Champs2) ~ ELITE.D + ELITE.F + CQF.W.L.
##          Df    AIC     LRT  Pr(>Chi)
## <none>      129.05
## ELITE.D   1 132.44  5.3864   0.02029 *
## ELITE.F   1 153.66 26.6156 2.482e-07 ***
## CQF.W.L.  1 154.68 27.6269 1.471e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For this dataset we looked at 16 teams from 16 different NHL seasons, so in total we have a dataset of size 256. Hence when we create our Training data set we randomly would choose 50% of the data from the original data set. We used the logistic regression to get a better understanding of the model to determine which variables has an influence on the model. After performing backwards elimnation and using our intution several variables were considered innsignificant and was dropped. Variables such as F.W.L(Finals Win Percentage) was dropped based off our intution because it is very evident that if you make it to the finals you have a higher chance to win the Championship. The variable DIFF(Goal Differnce) was dropped because it was insignificant to the model and this can be further proven from the scartter plots below. When we observe both scatter plots we can see the better players a team has the higher the Goal Difference. If where to look at the scatter plot observing the number of Elite Forwards has, the more goals will be scored which will result in the teams having a higher Goal Difference. If we were to also observe the regression line it clearly indicates as the number of Elite Forward increase, the higher the Goal Difference. The same can be said about the scatter plot which observes ELITE.D (Elite Defenders, the more Elite Defenders fewer goals will be scored against a team which will result in the teams having a higher Goal Difference. If we were to also observe the regression line it clearly indicates as the number of Elite Defenders increase, the higher the Goal Difference. Due to these two scatter plots we've decided that DIFF was insignificant variable and have decided to drop it.Our final model is polr(formula = ordered(Champs2) ~ ELITE.D + ELITE.F + CQF.W.L. We have discovered from this model that ELITE.F(Elite Forwards) are more significant then ELITE.D(Elite Defenders) in winning championships. From this we can acknowledge that it is important to have both great defenders and forwards, however the more Elite Forwards your team has the higher the chance you have of winning the NHL Stanley Cup Championship. This can be further be proven from observing the box-plots below. As we can see the best team Rk-1 has the most number of Elite Forwards. Where as for Elite Defenders we can see the best team Rk-1 has alot of Elite Defender yet they do not have the most. Hence, Elite Forwards are more significant then Elite Defenders in winning the NHL Stanley Cup Championship. It is also important to point out another significant variable CQF.W.L. From the model and box-plot below it shows it is very important to have a high Conference Quarter Final Win percentage yet you do not need to have the highest.

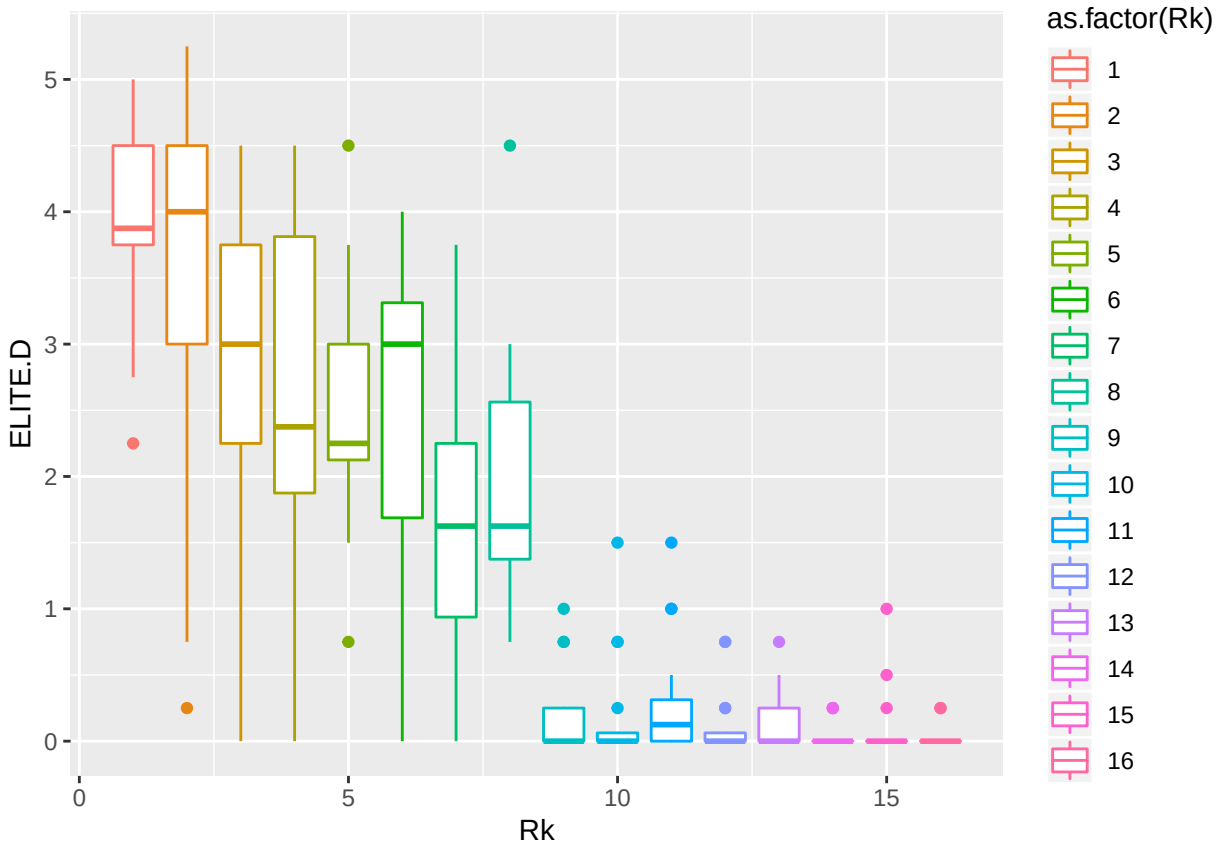
PLOTS

```
ggplot(NHLALL,aes(x=Rk,y=ELITE.F,colour=as.factor(Rk)))+geom_boxplot()
```



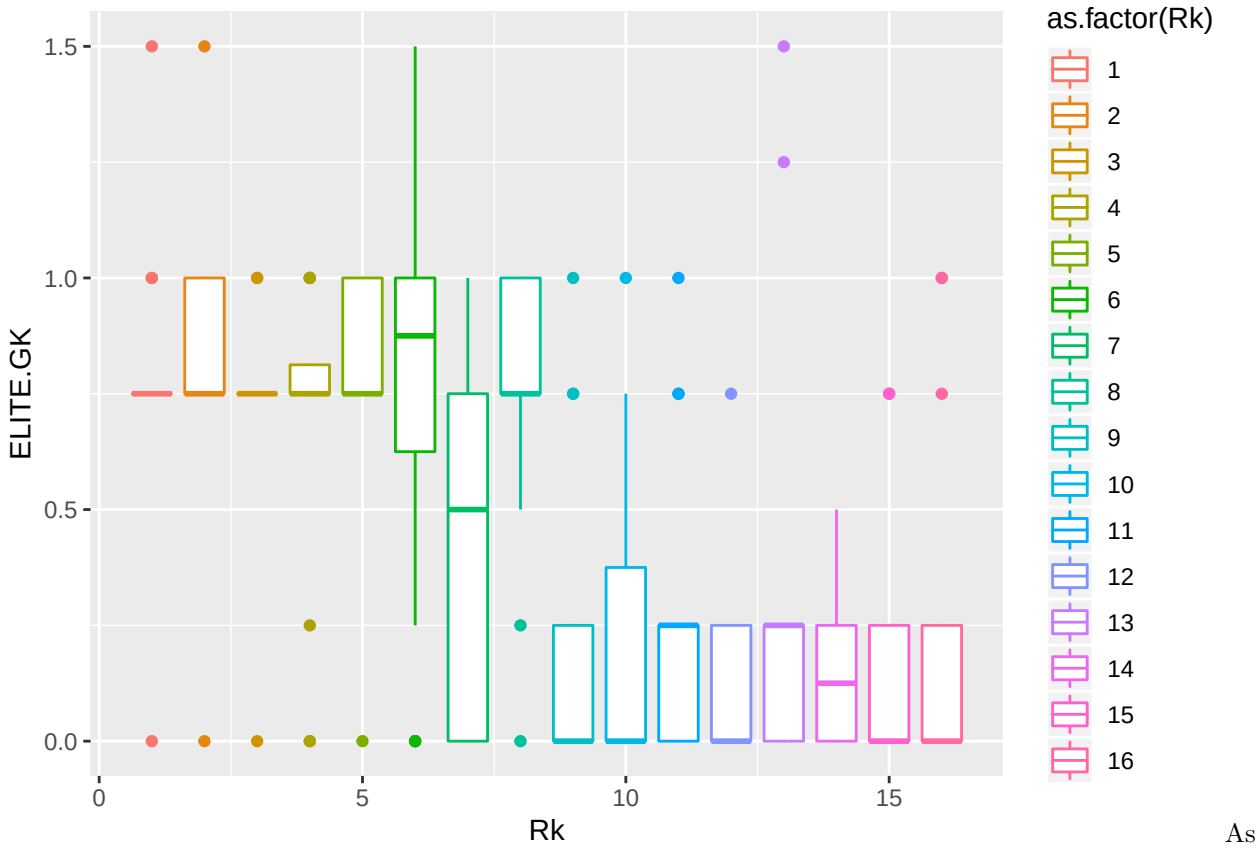
As we see in the boxplot, the x-axis is the Variable Rk where it ranges from Rk-1 which is the team that won the Championship till Rk-16, the team that did the worst. The y-axis indicates the number of Elite forawrds each team has and it is evident t hat there are outliers in teams ranked 2,4,6,7,8,9,10,11,12,13,14,15,and 16. It is also important to note that the teams with the most number of Elite Forwards tend to be closer in winning the championship. In this boxplot it cleary indicated that the team Rk-1(who won the championship), has the highest number of Elite Forwards on their team. It is also important to point out that Rk-8 has more Elite forwards then Rk-7, but Rk-7 still finishes one spot above Rk-8.

```
ggplot(NHLALL,aes(x=Rk,y=ELITE.D,colour=as.factor(Rk)))+geom_boxplot()
```



As we see in the boxplot, the x-axis is the Variable Rk where it ranges from Rk-1 which is the team that won the Championship till Rk-16, the team that did the worst. The y-axis indicates the number of Elite Defenders each team has and it is evident that there are outliers in all the teams except teams Rk-3,Rk-4, Rk-6 and Rk-7. It is also important to note that the teams with the most number of Elite Defenders tend to be closer in winning the championship. However, in this boxplot it cleary indicated that the team Rk-1(who won the championship), does not have the highest number of Elite Defenders on their team. Rk-2 has a higher variability then Rk-1 (who won the championship).This indicates that it is important to have a very good number of Elite Defenders more than 14 other teams to win the championship, however you do not need to have the most.

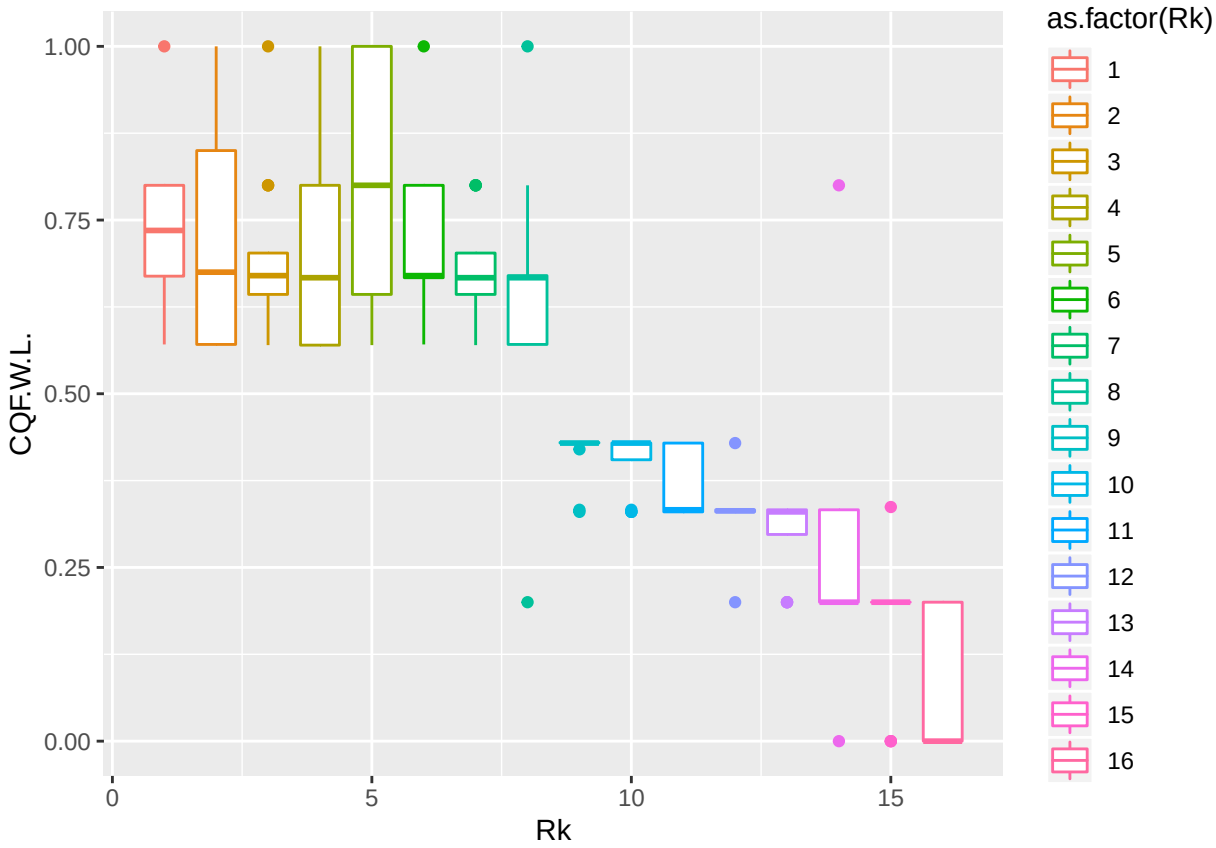
```
ggplot(NHLALL,aes(x=Rk,y=ELITE.GK,colour=as.factor(Rk)))+geom_boxplot()
```



As we see in the boxplot, the x-axis is the Variable Rk where it ranges from Rk-1 which is the team that won the Championship till Rk-16, the team that did the worst. The y-axis indicates the number of Elite Goalies each team has and it is evident that there are outliers in all the teams It is also important to note that the teams with the most number of Elite Goalies tend to be closer in winning the championship. However, in this boxplot it cleary indicated that the team Rk-1(who won the championship), does not have the highest number of Elite Goalies on their team. Rk-2,Rk-4,Rk-5,Rk-6 and Rk-7 has a higher variability then Rk-1 (who won the championship).This indicates that it is important to have a good number of Elite Goalies to win the championship, however you do not need to have the most.

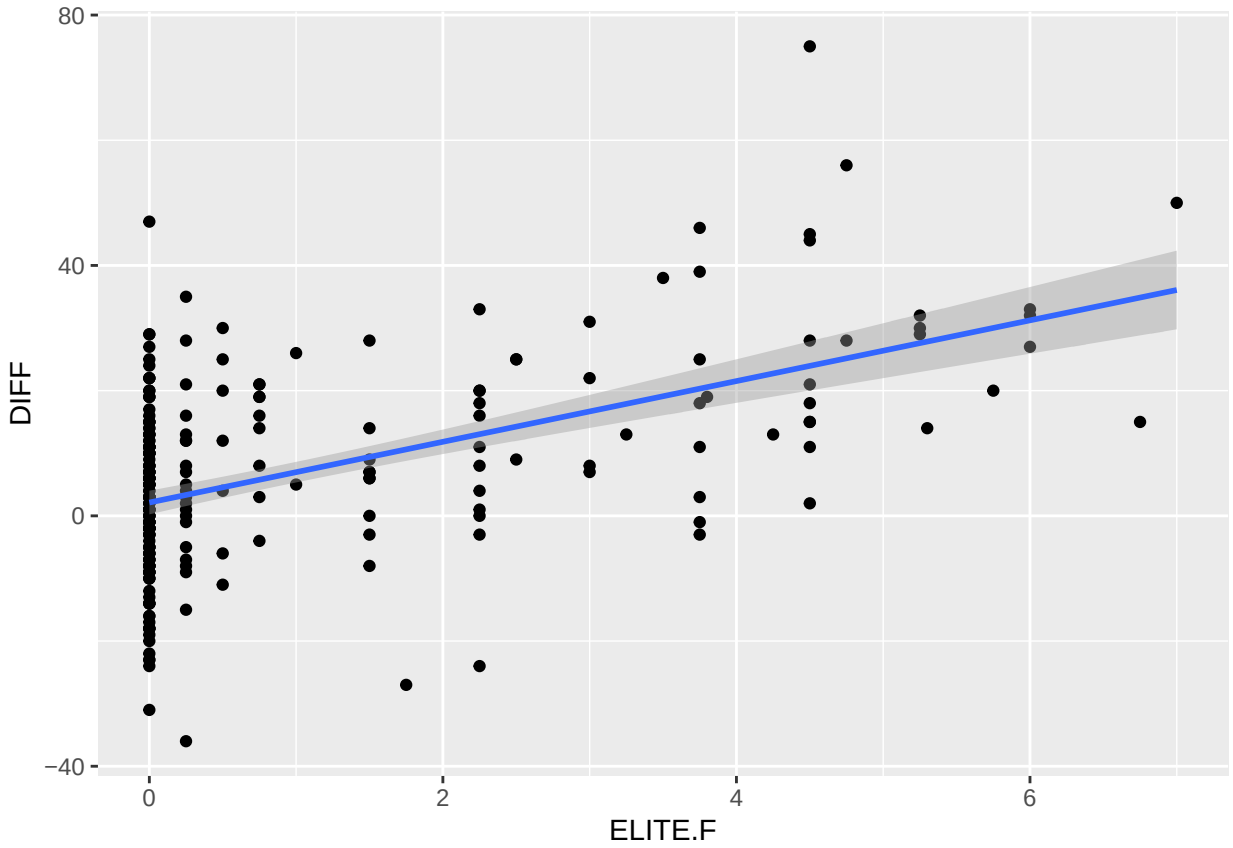
From observing the following 3 box-plots above we can see that all three positions are important. However, ELITE.F (Elite Forwards) are what wins championships. When we look at the box-plot which observes ELITE.GK (Elite Goalkeepers) it indicates that it is important to have a good number of Elite Goalies to win the championship primarily to get you into the playoffs, however you do not need to have the most. As we can see teams ranked Rk-2,Rk-4,Rk-5,Rk-6 and Rk-7 has a higher a better Goalkeeper then Rk-1 (who won the championship). Now when we focus on ELITE.D (Elite Defenders) it indicates that it is important to have a very good number of Elite Defenders more than 14 other teams to win the championship, however you do not need to have the most. Having Elite Defenders is cruicial in winning the NHL Championship but it is not the most important factor when observing the box-plot. Now the most significant position in winning the NHL Championship is the number of ELITE.F(Elite Forwards) a team has. Usually the team that won the chammpionship has the most number of Elite Forwards of that season compared to the other 15 teams in the playoffs. Our model also indicates this as well.

```
ggplot(NHLALL,aes(x=Rk,y=CQF.W.L.,colour=as.factor(Rk)))+geom_boxplot()
```



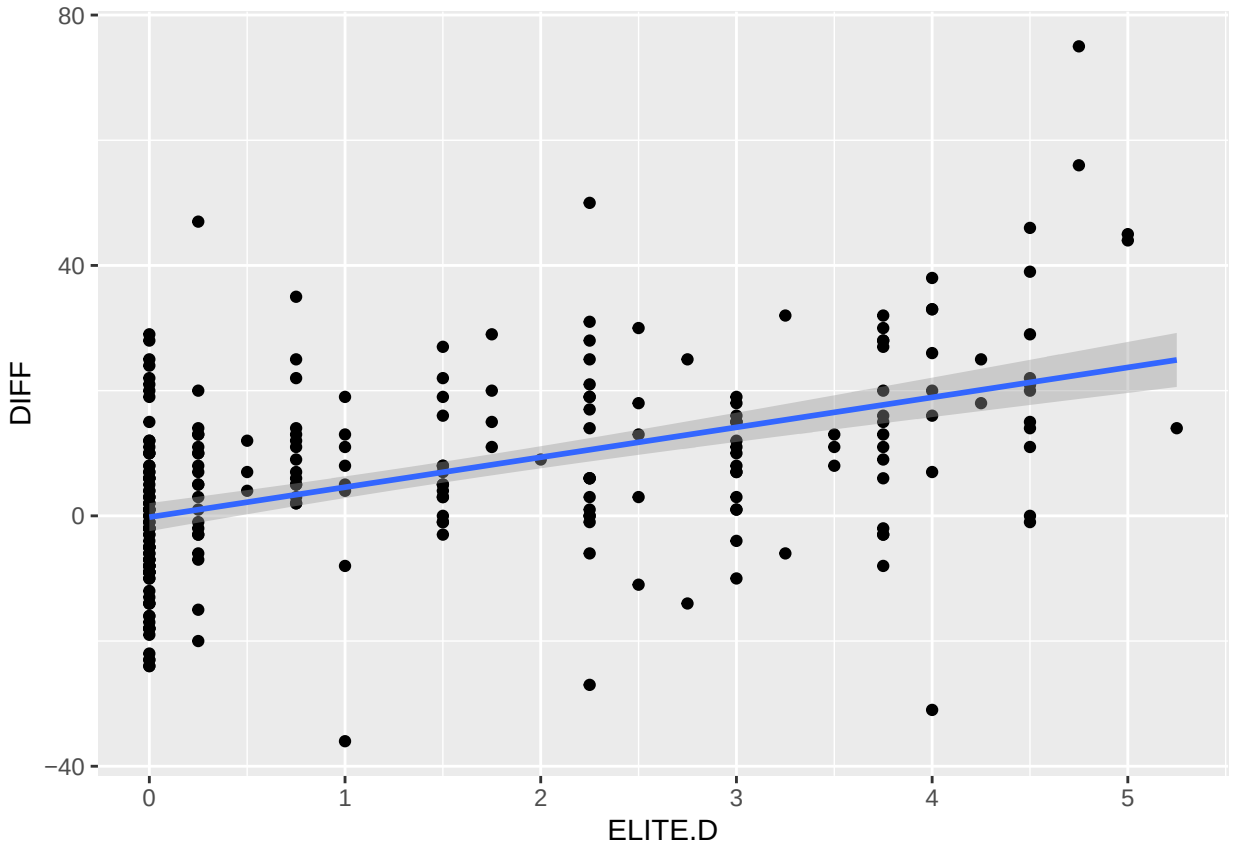
As we see in the boxplot, the x-axis is the Variable Rk where it ranges from Rk-1 which is the team that won the Championship till Rk-16, the team that did the worst. The y-axis indicates the win percentage for each team in Conference Quater Finals. It is evident that there are outliers in all the teams rankings excepts teams ranked, Rk-4,Rk-5,Rk-11 and Rk-16. Based on intution its very evident that the teams ranked from Rk-1 till Rk-8 will have a higher win percentage than teams ranked from Rk-9 till Rk-16. When foccusing on the top 8 teams, the team ranked 1st doesnt have the highest spread of win percentage. Infact teams ranked 2nd and 5th(The Highest) have a higher Conference Quarter Final win percentage then 1st ranked team.

```
ggplot(NHLALL,aes(x=ELITE.F,y=DIFF))+geom_point()+geom_smooth(method="lm")
```



From observing the scatter plot which is showing the effect ELITE Forwards have on Goal Difference, it shows the obvious. The more Elite Forwards the more goals will be scored which will result in the teams having a higher Goal Difference. If we were to also observe the regression line it clearly indicates as the number of Elite Forward increase, the higher the Goal Difference.

```
ggplot(NHLALL,aes(x=ELITE.D,y=DIFF))+geom_point()+geom_smooth(method="lm")
```



From observing the scatter plot which is showing the effect ELITE DEFENDERS have on Goal Difference, it shows the obvious. The more Elite Defenders fewer goals will be scored against a team which will result in the teams having a higher Goal Difference. If we were to also observe the regression line it clearly indicates as the number of Elite Defenders increase, the higher the Goal Difference.

When we observe both scatter plots we can see the better players a team has the higher the Goal Difference. If where to look at the scatter plot observing the number of Elite Forwards has, the more goals will be scored which will result in the teams having a higher Goal Difference. If we were to also observe the regression line it clearly indicates as the number of Elite Forward increase, the higher the Goal Difference. The same can be said about the scatter plot which observes ELITE.D (Elite Defenders, the more Elite Defenders fewer goals will be scored against a team which will result in the teams having a higher Goal Difference. If we were to also observe the regression line it clearly indicates as the number of Elite Defenders increase, the higher the Goal Difference. However if you look at the regression line for both, the Elite Forwards have a higherslope, this means Elite Forwards are more important in providing a bigger Goal Difference. Hence this further supports our model such that Elite Forwards are more valuable then Elite Defenders.

Champs~Rk+GP+W+L+W.L.+G+GA+DIFF+CQF.GP+CQF.OT.GP+CQF.W.L.+CQF.RS.W.L.+
CSF.GP+CSF.OT.GP+CSF.W.L.+CSF.RS.W.L.+CF.GP+CF.OT.GP+CF.W.L.+CF.RS.W.L.+F.GP+F.OT.GP+F.W.L.+